XXII

НАЦИОНАЛЬНАЯ КОНФЕРЕНЦИЯ ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ С МЕЖДУНАРОДНЫМ УЧАСТИЕМ

КИИ-2025

ТРУДЫ КОНФЕРЕНЦИИ

Том 1

6-10 октября 2025 г.
Санкт-Петербург

XXII

НАЦИОНАЛЬНАЯ КОНФЕРЕНЦИЯ ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ С МЕЖДУНАРОДНЫМ УЧАСТИЕМ

КИИ-2025

ТРУДЫ КОНФЕРЕНЦИИ

Том 1

6-10 октября 2025 г.
Санкт-Петербург,
Национальный исследовательский университет ИТМО

Российская ассоциация искусственного интеллекта

Федеральный исследовательский центр «Информатика и управление» РАН

Национальный исследовательский университет ИТМО

Санкт-Петербургский Федеральный исследовательский центр РАН

Санкт-Петербург
СПб ФИЦ РАН
2025

УДК 004.8+004.89+004.82+004.032.26(045)+004.9
ББК 32.813
Д 22

Организаторы конференции:

Российская ассоциация искусственного интеллекта

Федеральный исследовательский центр «Информатика и управление» РАН

Национальный исследовательский университет ИТМО

Санкт-Петербургский Федеральный исследовательский центр РАН

Д 22 Двадцать вторая Национальная конференция по искусственному интеллекту с международным участием, КИИ-2025 (Санкт-Петербург, 6-10 октября 2025 г.). Труды конференции. В 3-х томах. Т. 1. – СПб: Изд-во СПб ФИЦ РАН, 2025. – 377 с.

ISBN 978-5-6052274-4-1

Двадцать вторая Национальная конференция по искусственному интеллекту с международным участием КИИ-2025 продолжает традицию советских (российских) конференций, организуемых Российской ассоциацией искусственного интеллекта.

В первом томе трудов публикуются пленарные доклады и доклады участников конференции, представленные на следующих секциях:

Секция 1 «Инженерия знаний»,

Секция 2 «Интеллектуальный анализ данных»,

Секция 3 «Моделирование рассуждений»,

Секция 4 «Интеллектуальный анализ текстов, большие языковые модели».

ББК 32.813

Рецензенты: академик РАН, ИПУ РАН *С.Н. Васильев,*
д.т.н., ФИЦ ИУ РАН *О.Г. Григорьев*

ISBN 978-5-6052274-4-1

ВАЛЕРИЙ БОРИСОВИЧ ТАРАСОВ
16 февраля 1955 – 22 июля 2021

В этом году конференция посвящена 70-летию со дня рождения Валерия Борисовича Тарасова – выдающегося ученого в области искусственного интеллекта, ведущего отечественного специалиста по направлениям семиотического моделирования, многоагентных систем, нечетких систем, мягких вычислений и когнитивных измерений.

СОПРЕДСЕДАТЕЛИ КОНФЕРЕНЦИИ

Соколов И.А., акад. РАН, ФИЦ ИУ РАН, Москва
Васильев В.Н., член-корр. РАН, НИУ ИТМО, Санкт-Петербург
Ронжин А.Л., д.т.н., проф. РАН, СПб ФИЦ РАН, Санкт-Петербург

ПРОГРАММНЫЙ КОМИТЕТ КОНФЕРЕНЦИИ

Сопредседатели Программного комитета
Кобринский Б.А., д.м.н., проф., ФИЦ ИУ РАН, Москва
Котенко И.В., д.т.н., проф., СПб ФИЦ РАН, Санкт-Петербург

Заместители председателя Программного комитета
Грибова В.В., член-корр. РАН, ИАПУ ДВО РАН, Владивосток
Забежайло М.И., д.ф.-м.н., проф., ФИЦ ИУ РАН, Москва

Ответственный секретарь Программного комитета
Подвесовский А.Г., к.т.н., доц., БГТУ, Брянск

Члены программного комитета
Аверкин А.Н., к.ф.-м.н., доц., ФИЦ ИУ РАН, Москва
Афанасьева Т.В., д.т.н., РЭУ им. Г.В. Плеханова, Москва
Бобцов А.А., д.т.н., проф., НИУ ИТМО, Санкт-Петербург
Болодурина И.П., д.т.н., проф., Оренбургский ГУ, Оренбург
Боргест Н.М., к.т.н., доц., Самарский НИУ им. акад. С.П. Королева, Самара
Борисов В.В., д.т.н., проф., филиал НИУ МЭИ, Смоленск
Бухановский А.В., д.т.н., проф., НИУ ИТМО, Санкт-Петербург
Васильев С.Н., акад. РАН, ИПУ РАН, Москва
Визильтер Ю.В., д.ф.-м.н., проф., ГосНИИАС, Москва
*Виноградов Д.*В., д.ф.-м.н., ФИЦ ИУ РАН, Москва
Гаврилова Т.А., д.т.н., проф., СПбГУ, Санкт-Петербург
Гладков Л.А., к.т.н., доц., ИКТИБ ЮФУ, Таганрог
Городецкий В.И., д.т.н., проф., АО «Эврика», Санкт-Петербург
Еремеев А.П., д.т.н., проф., НИУ «МЭИ», Москва
Желтов С.Ю., акад. РАН, ГосНИИАС, Москва
Загорулько Ю.А., к.т.н., ИСИ СО РАН, Новосибирск
Ильин В.А., д.ф.-м.н., проф., НИУ ИТМО, Санкт-Петербург
Калюжная А.В., к.т.н., доц., НИУ ИТМО, Санкт-Петербург
Карпов А.А., д.т.н., проф., СПб ФИЦ РАН, Санкт-Петербург
Ковалев С.М., д.т.н., проф., РГУПС, Ростов-на-Дону
Колесников А.В., д.т.н., проф., БФУ, Калининград
Колоденкова А.Е., д.т.н., доц., Самарский НИУ, Самара
Кузнецов О.П., д.т.н., проф., ИПУ РАН, Москва
Кузнецов С.О., д.ф.-м.н., проф., НИУ ВШЭ, Москва
Лебедев О.Б., д.т.н., доц., ВАГШ ВС РФ, Москва

Лукашевич Н.В., д.т.н., проф., МГУ им. М.В. Ломоносова
*Макаров Д.*А., к.т.н., ФИЦ ИУ РАН, Москва
Мисник А.Е., к.т.н., доц. БРУ, Республика Беларусь, Могилев
Михеенкова М.А., д.т.н., проф., ФИЦ ИУ РАН, Москва
Мошкин В.С., к.т.н., доц., УлГТУ, Ульяновск
Насонов Д.А., к.т.н., доц., проф., НИУ ИТМО, Санкт-Петербург
Пальчунов Д.Е., д.ф.-м.н., доц., ИМ СО РАН, Новосибирск
Палюх Б.В., д.т.н., проф., ТвГТУ, Тверь
Панов А.И., д.ф.-м.н., доц., МФТИ, Москва
Редько В.Г., д.ф.-м.н., проф., НИИСИ РАН, Москва
Ройзензон Г.В., к.т.н., доц., ФИЦ ИУ РАН, Москва
Рыбина Г.В., д.т.н., проф., НИЯУ МИФИ, Москва
Смирнов И.В., д.ф.-м.н., доц., ФИЦ ИУ РАН, Москва
Стефанюк В.Л., д.т.н., проф., ИППИ РАН, Москва
Сулейманов Д.Ш., акад. АН РТ, ИПС АН РТ, Казань
Тельнов Ю.Ф., д.э.н., проф., РЭУ, Москва
Уткин Л.В., д.т.н., проф., СПбПУ, Санкт-Петербург
Финн В.К., д.т.н., проф., ФИЦ ИУ РАН, Москва
Хачумов В.М., д.т.н., проф., ФИЦ ИУ РАН, Москва
Шалфеева Е.А., д.т.н., ИАПУ ДВО РАН, Владивосток
Яковлев К.С., к.ф.-м.н., ФИЦ ИУ РАН, Москва

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ КОНФЕРЕНЦИИ

Сопредседатели организационного комитета

Борисов В.В., д.т.н., проф., филиал НИУ МЭИ, Смоленск
Заколдаев Д.А., к.т.н., доц., НИУ ИТМО, Санкт-Петербург

Члены организационного комитета

Благосклонов Н.А., ФИЦ ИУ РАН, Москва
Волошина Н.В., к.т.н., доц., НИУ ИТМО, Санкт-Петербург
Воробьева А.А., к.т.н., доц., НИУ ИТМО, Санкт-Петербург
Давыдов В.В., к.т.н., НИУ ИТМО, Санкт-Петербург
Десницкий В.А., к.т.н., доц., СПб ФИЦ РАН, Санкт-Петербург
Левшун Д.С., к.т.н., СПб ФИЦ РАН, Санкт-Петербург
Кириллова Е.А., д.э.н., доц., филиал НИУ «МЭИ», Смоленск
Попов И.Ю., к.т.н., доц., НИУ ИТМО, Санкт-Петербург
Синявский Ю.В., к.т.н., филиал НИУ «МЭИ», Смоленск
Солопов Р.В., к.т.н., доц., филиал НИУ «МЭИ», Смоленск
Чечулин А.А., к.т.н., доц., СПб ФИЦ РАН, Санкт-Петербург

ИНФОРМАЦИОННАЯ ПОДДЕРЖКА КОНФЕРЕНЦИИ

ООО «Лаборатория информационных технологий», Смоленск

ПРЕДИСЛОВИЕ

Двадцать вторая Национальная конференция по искусственному интеллекту с международным участием КИИ-2025 продолжает традицию советских (российских) конференций, организуемых Российской ассоциацией искусственного интеллекта (РАИИ).

Конференция посвящена памяти Валерия Борисовича Тарасова, 70-летие со дня рождения которого отмечается в этом году – выдающегося ученого в области искусственного интеллекта, который внес существенный вклад в развитие методов семиотического моделирования, в теорию агентов и многоагентных систем, теорию нечетких систем, мягких вычислений и когнитивных измерений. Валерий Борисович Тарасов был участником Учредительного съезда Советской ассоциации искусственного интеллекта (ныне РАИИ). С 1992 по 2000 год являлся вице-президентом РАИИ, с 2000 года – членом Научного совета РАИИ. Он был одним из основателей Российской ассоциации нечетких систем (ныне Ассоциации нечетких систем и мягких вычислений).

Федеральный проект «Искусственный интеллект» направлен на решение междисциплинарных проблем, сформулированных в Национальной стратегии развития искусственного интеллекта, включая исследования как в области фундаментальной и прикладной науки, так и в образовательной сфере. Эти проблемы постоянно находятся в фокусе внимания членов РАИИ. Их отражение модно видель в докладах конференции, организаторами которой являются *Российская ассоциация искусственного интеллекта*, *Федеральный исследовательский центр «Информатика и управление» РАН*, *Национальный исследовательский университет ИТМО*, *Санкт-Петербургский Федеральный исследовательский центр РАН*.

Тематика конференции охватывает следующие основные направления искусственного интеллекта: инженерия знаний; интеллектуальный анализ данных; моделирование рассуждений; интеллектуальный анализ текстов, большие языковые модели; нечеткие модели, мягкие измерения и вычисления, биоинспирированные методы; интеллектуальные агенты, роботы, компьютерное зрение; интеллектуальное управление и поддержка принятия решений; машинное обучение, нейросетевые методы, нейроинформатика; инструментальные средства и технологии проектирования интеллектуальных систем; прикладные интеллектуальные системы.

Представлены пленарные доклады видных ученых и специалистов в области искусственного интеллекта. Из поданных на конференцию 138 секционных докладов Программным комитетом после рецензирования отобраны 102 доклада.

В конференции принимают участие более 280 ученых, исследователей и специалистов из академических институтов, университетов, отраслевых институтов и организаций Москвы, Санкт-Петербурга, Белгорода, Брянска, Владивостока, Донецка, Иркутска, Казани, Калининграда, Красноярского края, Махачкалы, Московской области, Новосибирска, Пермского края, Ростова-на-Дону, Самары, Смоленска, Таганрога, Твери, Ульяновска, Уфы, Ханты-Мансийска, Челябинска, Ярославля, Могилёва (Республика Беларусь), Суйчжоу, (Китай), Ташкента (Узбекистан), Хайдарабада (Индия).

Программный комитет КИИ-2025

ПЛЕНАРНЫЕ ДОКЛАДЫ

УДК 004.8
doi: 10.15622/rcai.2025.001

ПРИОРИТЕТНЫЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ И КЛЮЧЕВЫЕ ТЕНДЕНЦИИ РАЗВИТИЯ ТЕХНОЛОГИЙ ИИ

Ю.В. Визильтер (*viz@gosniias.ru*)

ФАУ «ГосНИИАС», Москва

Представлены тенденции развития методов и технологий ИИ на текущем этапе (2020-2025), сгруппированные по нескольким приоритетным направлениям исследований в сфере ИИ. Кратко описаны основные тенденции и результаты по следующим ключевым направлениям и поднаправлениям: LLM и другие модели для символьных данных, диффузионные и другие модели для несимвольных данных, мультимодальные модели, методы переноса знаний с адаптацией моделей, аугментация LLM без адаптации моделей, обучение с подкреплением, агентные и мультиагентные системы, элементы общего ИИ (AGI).

Ключевые слова: Artificial Intelligence, Machine Learning, Computer Vision, NLP, LLM, Generative AI, RL, General AI.

Введение

Начало нынешнего периода развития технологий искусственного интеллекта (ИИ) принято отсчитывать от 2011 г. В рамках данного периода можно условно выделить три этапа. На первом этапе (2011-2016) в центре внимания находились сверточные нейронные сети (CNN) [Krizhevsky et al., 2017]. На втором этапе (2016-2020) получил широкое распростране-

ние ряд новых подходов, таких как GAN для синтеза изображений [Goodfellow et al., 2014], архитектуры типа Transformer [Vaswani et al., 2017], GPT [Radford et al., 2018], BERT [Devlin et al., 2018], на основе т.н. «модулей внимания» (Attention [Vaswani et al., 2017]) для задач NLP, обучение с подкреплением [Mnih et al., 2015], [Schulman et al., 2017], [Badia et al., 2020]), графовые модели (Graph CNN [Zhang et al., 2019]), автоматическое обучение (Auto-ML, NAS, HPO [Xin et al., 2021]).

Ниже будут представлены основные тенденции развития технологий ИИ на текущем, третьем этапе (2020-2025), сгруппированные по ряду приоритетных направлений и поднаправлений фундаментальных и поисковых исследований в сфере ИИ.

Приоритетные направления фундаментальных и поисковых исследований в сфере искусственного интеллекта

В рамках Форсайта по приоритетным направлениям фундаментальных/поисковых исследований в сфере искусственного интеллекта [AI Foresight, 2024], подготовленного в 2024 г. группой ведущих российских экспертов по инициативе Аналитического центра при правительстве РФ, Сбера и НИУ ВШЭ, были выделены следующие основные приоритетные направления исследований.

Н1. Архитектуры, алгоритмы машинного обучения (МО), оптимизация и математика, в том числе: разработка новых алгоритмов МО, поиск архитектур глубоких сетей, ускорения вычислений, распределенное и федеративное обучение.

Н2. Вычисления для ИИ, в том числе: разработка вычислителей для ИИ (квантовые, фотонные, нейроморфные, Edge), разработка АПК для ИИ, фреймворки для МО и ИИ;

Н3. Данные для ИИ, в том числе: создание бенчмарков для оценки ИИ, создание и аугментация данных (синтетика, зашумление), сохранение конфиденциальности данных.

Н4. Фундаментальные и генеративные модели, в том числе: LLM и др. модели для символьных данных, диффузионные и др. модели для не символьных данных, мультимодальные LLM модели, методы переноса знаний, аугментация LLM без адаптации моделей.

Н5. Безопасность, доверие и объяснимость, в том числе: выравнивание ценностей (Alignment), объяснимость работы ИИ, обеспечение безопасности разработки и эксплуатации ИИ, обеспечение защиты от результатов использования ИИ.

Н6. ИИ для узких задач (Narrow Al), в том числе: CV (компьютерное зрение), NLP (обработка естественного языка), прочие технологии узкого ИИ.

Н7. Управление, принятие решений, агентные/мультиагентные системы, в том числе: разработка алгоритмов обучения с подкреплением (RL), агентные системы, мультиагентные системы.

Н8. Элементы общего ИИ (AGI), в том числе: рассуждения и рефлексия, гибридный ИИ (Symbolic Al), воплощенный ИИ (Embodiment), моделирование мозга и психики.

Н9. Взаимодействие человека и машины, в том числе: технические средства человеко- машинного взаимодействия, методы и алгоритмы взаимодействия с человеком, способы человеко-машинной интеграции.

Далее будут кратко описаны основные тенденции и результаты по направлениям Н4, Н7 и Н8, которые представляются ключевыми в том смысле, что развитие остальных направлений в значительной степени связано именно с проблемами и достижениями в этих областях.

Н4. Фундаментальные и генеративные модели

Н4.1. LLM и другие модели для символьных данных

Большие языковые и фундаментальные модели. В 2020 г. трансформер GPT-3 [Brown et al., 2020] стал первой из класса больших языковых моделей (Large Language Models, LLM) с многими миллиардами параметров. В 2021 г. было предложено понятие *фундаментальных моделей* (Foundation Models, FM) [Rishi et al., 2021], предобученных на тком количестве данных, что они далее не требуют или требуют минимального дообучения. В 2022 г. создан ChatGPT, который сделал работу с LLM полезной и удобной за счет дообучения LLM методом RLHF (Reinforcement Learning from Human Feedback) [Ouyang et al., 2022].

Повышение вычислительной эффективности LLM. Предложен целый ряд подходов, основанных на квантовании и прореживании (прунинг) весов моделей; дистилляции в модели меньшего размера; быстром предсказании и последующей проверке результатов генерации (speculative decoding); оптимизации KV-кэша в модулях внимания; смеси экспертов (Mixture of Experts, MoE) и др. [Wan et al., 2024]. Многие из этих приемов были использованы и развиты в работах DeepSeek [DeepSeek-V2], [DeepSeek-V3], что позволило резко снизить стоимость работы публично доступных LLM.

Развитие трансформерных архитектур. В последние годы архитектура LLM достаточно активно эволюционировала. Были предложены, в частности, такие архитектурные элементы как Multi-Head Latent Attention (MLA), Grouped-Query Attention (GQA), RMSNorm, Post-Norm, QK-Norm, sliding window attention, MatFormer, Per-Layer Embedding (PLE), RoPE (Rotational

Positional Embeddings) и NoPE (No Positional Embeddings), SwiGLU и др. Хороший русскоязычный обзор современных трансформерных архитектур LLM со ссылками можно найти в [LLM Architecture Evolution, 2025].

Альтернативные архитектуры. Предложен целый ряд как альтернативных трансформерам, так и гибридных архитектур – Retentive Network [Sun et al., 2023], RWKV [Peng et al., 2023], State Space Models [Gu et al., 2021], Mamba [Gu et al., 2023], Jamba [Lieber O. et al., 2024], Hyena [Poli et al., 2023]. В работе 2024 г. Titans [Behrouz et al., 2024] предложен подход к созданию модулей обработки информации для больших моделей, основанный на понимании процесса обучения как запоминания во время исполнения (Test-Time Memorization). Этот подход был далее обобщен и развит для создания целого спектра различных альтернативных модулей LLM [Behrouz et al., 2025], [Wang et al., 2025]. Делаются также попытки разработать принципиальные альтернативы не только трансформерам, но и традиционным нейросетевым архитектурам [Halverson J. et al., 2024].

Concept Models. Возможное направление совершенствования архитектуры LLM для достижения более абстрактного представления знаний показывает подход Large Concept Models [Barrault L. et al, 2024], в котором уровень «концепций» и уровень реализующих их «слов» архитектурно разделены.

Текстовые диффузионные модели. Разработаны диффузионные и потоковые генеративные модели для символьных данных, предполагающие отход от авторегрессионной модели трансформаров для последовательной генерации токенов. Текстовая диффузия реализует модель итеративного «восстановления» генерируемого текста из массива маскированных токенов [Shi et al., 2024]. В 2025 году текстовая диффузионная модель впервые показала результаты, сравнимые с результатами LLM того же размера [Nie et al., 2025]. Модели типа GFlowNets [Bengio et al., 2021] позволяют генерировать объекты, обладающие структурой. Диффузионные модели и GFlowNets могут использоваться, в частности, для создания цепочек рассуждений [Ye et al., 2024], [Takase et al., 2024], [Ho et al., 2024].

Н4.2. Диффузионные и другие модели для несимвольных данных

Переход от генеративно-состязательных сетей к диффузионым моделям. В 2020 г. за первенство среди генеративных моделей с GAN боролись также вариационные автоэнкодеры (Variational Auto-Encoder, VAE [Kingma et al., 2013], [Child, 2020]) и диффузионные модели (Diffusion Models [Ho et al., 2020], [Song et al.]). Сегодня диффузионные модели используются в большинстве работ и приложений [Ling et al., 2022]. С их помощью (в сочетании с трансформерами) решаются задачи гене-

рации изображений по текстовому описанию (DALL-E 2 [Aditya et al., 2022], DALL-E 3 [Improving Image Generation, 2024], Stable Diffusion 3 [Esser et al., 2024]), генерации видео по текстовому описанию (Imagen Video [Ho et al., 2022], OpenAI SORA [Brooks et al., 2024]). В 2025 г. DeepMind представлена система Veo3 [DeepMind VEO, 2025], демонстрирующая возможности симуляция достоверного физического поведения объектов, веществ и персонажей, а также синхронной генерации видеоряда и звука (звуковые эффекты, фоновая музыка, диалоги).

Обучаемые 3D модели трехмерных сцен позволили соединить технологии 3D рендеринга сцен с машинным обучением. В 2021 г. был предложен метод NeRF (neural radiance fields, [Brooks et al., 2024]), который начал широко использоваться не только в задачах генерации сцен, но и в задачах компьютерного зрения. В 2023 г. был предложен альтернативный подход 3D Gaussian Splatting (3DGS) [Kerbl et al., 2023], который обучается гораздо быстрее, работает в реальном времени и обеспечивает сравнимое или лучшее качество рендеринга по сравнению с NeRF [Chen et al., 2024]. Для генерации 3D сцен по текстовым запросам (text-to-3D) в настоящее время используется 3D GS в соединении с LLM (GALA3D, 2024) [Zhou et al., 2024]. В работе [Zielonka et al., 2023] модель 3D GS также используется для генерации и рендернига в реальном времени реалистичных 3D аватар (подвижных моделей тела) людей.

H4.3 Мультимодальные LLM-модели

Возможность создания мультимодальных LLM-моделей была обеспечена, в первую очередь, за счет создания *фундаментальных моделей для задач зрения*: Segment Anything (SAM) [Kirillov et al., 2023] для семантической сегментации, DINO [Liu et al., 2023], DINOv2 [Ouqab et al. 2024] для обнаружения объектов с открытым списком классов и др. Фундаментальные модели для задач зрения могут быть также построены на основе сверточных сетей (YOLO-World, [Cheng et al., 2024]), что вполне обосновано, поскольку лучшие практические решения по обнаружению объектов для бортовых приложений основаны сегодня на гибридных архитектурах сверточных сетей с модулями внимания (YOLOv12, [Tian et al., 2025]).

Мультимодальные LLM-модели (MLLM) [Yin et al., 2024] используются для решения задач, требующих не только анализа изображения, но и понимания контекста, то есть, анализа на уровне модели мира. При этом, как показано в [Wang et al., 2024], эффективная MLLM может быть моделью LLM среднего или небольшого размера, соединенной с фундаментальной моделью для зрения при помощи адаптера.

H4.4. Методы переноса знаний с адаптацией моделей

Для повышения вычислительной эффективности обучения LLM были предложены методы т.н. параметрически эффективной настройки (PEFT), позволяющие обучать лишь небольшое подмножество параметров предварительно обученной модели [Wu et al., 2024]. К таким методам относятся, в частности, методы низкоранговой адаптации и спектральной переметризации.

Low-Rank адаптеры. Низкоранговая декомпозиция матриц весов и запросов (prompts) для уменьшения числа обучаемых параметров. Это такие уже широко используемые методы как LoRA [Hu et al., 2022] и LLM-адаптеры [Hu et al., 2023]. Работы, развивающие данное направление: LoRA² [Zhang et al., 2024], Low-Rank Prompt Adaptation [Jain et al., 2024], RankAdaptor [Zhou et al., 2024]. Перспективным является также использование дистиллированных моделей в комбинации с LoRA [DeepSeek-AI, 2025].

Спектральная параметризация предполагает использование для тех же целей спектрального разложения весов и адаптации моделей в спектральном пространстве: Spectral Adapter [Zhang et al. 2024], LaMDA [Azizi et al., 2024].

H4.5. Аугментация LLM без адаптации моделей

В 2023 г. модели GPT-4 [Achiam et al., 2023], на порядок превосходящей GPT-3 по размеру, удалось показать качество ответов на запросы на уровне людей-профессионалов. С этого момента в фокусе исследования оказались способы адаптации LLM без дообучения (изменения) весов модели. Ключевыми технологиями здесь являются: инженерия запросов (Prompt Engineering, PE) [Sahoo et al., 2024], [Your Guide to Generative AI, 2023], [Prompt Engineering Guide, 2023], включая работу с базами документов Retrieval-Augmented Generation (RAG) [Lewis et al., 2020], [Gao et al., 2024], контекстное обучение (in-Context Learning, iCL) [Dong et al., 2023], логические рассуждения в LLM [Sahoo et al., 2024], [Wang et al., 2022], [Yao et al., 2023], [Yao et al., 2023], создание и использование генеративных агентов (GA).

H7. Управление, принятие решений, агентные/мультиагентные системы

H7.1. Разработка алгоритмов RL (обучения с подкреплением)

Обучение с открытым списком виртуальных сред и целевых задач для приобретения когнитивного поведения. Авторы работы Open-Ended Learning (2021) [Adam et al., 2021] показали, что, если построить вселенную игровых задач и последовательно обучать ИИ-агентов играть в эти игры, то с каждой новой игрой они будут достигать лучших результатов в этой вселенной и за ее пределами за счет овладения навыками когнитивного поведения.

Универсальные агенты для робототехники на основе трансформеров и LLM. В работе GATO (2022) [Scott et al., 2022] был представлен универсальный агент-трансформер, способный играть в игры Atari, давать текстовое описание изображений, вести языковой чат, управлять рукой робота, переставляющего блоки и т.п. В [Bousmalis et al., 2023] описана Vision-Language-Action (VLA) модель RoboCat для управления манипуляторами, обученная методом открытого обучения в реальном мире. VLA модель RT-2 [Brohan et al., 2023] на основе LLM сочетает управление роботом с рассуждениями на основе chain-of-thought. В 2023-2025 гг. создан ряд фундаментальных VLA-моделей: для самообучения физических роботов AutoRT [Brohan et al., 2023], для мультимодальной навигации Uni-NaVid [Zhang et al., 2024], для общего управления роботами π0 [Black et al., 2024].

Совместное использование RL и LLM является одним из основных трендов современного машинного обучения. При этом лучшее качество результатов достигается как в случае использования LLM для реализации RL, так и в случае использования RL для обучения LLM [Pternea et al., 2024].

Использование LLM при реализации RL предполагает такие схемы как вербальное, контекстное и символьное обучение с подкреплением. При вербальном обучении с подкреплением LLM генерирует словесную саморефлексию, чтобы обеспечить детальную и конкретную обратную связь, и затем сохраняет ее в памяти (контексте) действующего агента [Shinn et al., 2023]. Количество потребных для обучения попыток при переходе от градиентного к вербальному RL может сократиться на несколько порядков. В работе [Laskin et al., 2022] предложен метод Algorithm Distillation (AD), реализующий контекстное обучение с подкреплением на основе сбора историй обучения для отдельных алгоритмов RL. Пример символьного RL показан в работе Eureka [Ma et al., 2023], где демонстрируется возможность автоматического подбора критерия оптимизации для RL с использованием LLM и рефлексии.

Использование RL при обучении LLM направлено на решение проблемы преодоления ограничений существующей обучающей выборки. Метод обучения с подкреплением на основе генеративного пополнения обучающей выборки предполагает итеративное использование этапов генерации синтетических данных предобученной LLM, фильтрации полученных синтетических данных и дообучения LLM на отфильтрованных синтетических данных. В случае задач, связанных с математикой или программированием, проверка корректности сгенерированных решений может быть реализована автоматически, что повышает производительность метода. Таким способом DeepMind были обучены модели для автоматического программирования (AlphaCode 2 [AlphaCode 2, 2023]) и автоматического решения математических задач в области геометрии (AlphaGeometry [Trinh et al., 2024], AlphaGeometry2 [Chervonyi et al., 2025]) и комбинато-

рики (FunSearch [Romera-Paredes et al., 2024]). Данный подход также активно использовался DeepSeek при автоматическом обучении модели-«мыслителя» DeepSeek-R1 навыкам логики, математики и программирования [Guo et al., 2025].

Н7.2. Агентные системы. Н7.3. Мультиагентные системы

Генеративные агенты. В 2022-24 гг. активно развивались технологии создания LLM-агентов [LLM Powered Autonomous Agents, 2023], ключевыми компонентами которых являются: декомпозиция и планирование задач, а также использование инструментов. Агенты также могут использовать самокритику и саморефлексию, чтобы учиться на ошибках и совершенствовать свои способы решения. В 2025 году главным фокусом внимания становятся фундаментальные автономные агенты. Современный обзор фундаментальных LLM-агентов можно найти в [Liu et al., 2025].

Model Context Protocol (MCP). В технологическом плане важным новым инструментом создания LLM-агентов стал предложенный компанией Anthropic в ноябре 2024 г. протокол MCP – фреймворк с открытым исходным кодом для стандартизации взаимодействия и обмена данными между моделями ИИ и внешними системами, источниками данных и инструментами. В настоящее время MCP поддерживается всеми основными разработчиками моделей и сервисов ИИ на основе LLM. В репозитории MCP [Model Context Protocol, 2025] доступны реализации MCP-серверов для интерфейса со средами разработки (Python, TypeScript, Java, C#) а также с популярными корпоративными системами (Google Диск, Slack, GitHub, PostgreSQL и др.). Разработчики приложений на основе LLM-агентов могут создавать собственные MCP-серверы.

Многоагентные системы. Предложены различные схемы построения мультиагентных систем с разными сценариями взаимодействия между агентами для принятия групповых решений: конкуренция, координация, кооперация (ReConcile [Hao et al., 2023], Socratic AI [Boiko et al., 2023]). Многоагентные системы могут приобретать новые знания. В работе ChatLLM Network [Hao et al., 2023] предложена структура сети LLM-агентов, в которой организуется процесс прямого и обратного распространения текстовых сообщений для контекстного обучения коллектива LLM-агентов. В другой схеме Socratic Models (SM) [Andy et al., 2022] предложен способ построения расширяемого коллектива бимодальных ИИ-агентов, который может функционировать в условиях «открытого» списка задач и типов входных данных, для работы с которыми добавляются новые агенты.

Агенты для помощи в научных исследованиях. Для решения конкретных научных задач предложены, в частности, агенты в области органического синтеза, создания лекарств и дизайна материалов ([Bran et al., 2023], [Boiko et al., 2023]). Для общей реализации научного метода в работе AI Co-

scientist [AI Co-scientist, 2025] реализована система помощи ученым, включающая специализированных агентов (Generation, Reflection, Ranking, Evolution, Proximity, Meta-review), которые итеративно генерируют, оценивают и уточняют гипотезы с помощью автоматической обратной связи. В работе AI Scientist [Lu et al., 2024] был впервые реализован полный цикл научного исследования в области ML. В 2025 г. статья AI Scientist v2 прошла слепое рецензирование на воркшоп ICLR-2025 (конференция класса A* в области ИИ). OpenAI предлагает услуги ИИ-агента уровня кандидата наук (PhD-Level Agent) [OpenAI PhD-level agents, 2025]. Современное состояние области использования ИИ для научных исследований можно найти в обзоре AI4Research [Chen et al., 2025].

H8. Элементы AGI

Общий ИИ (Artificial general intelligence, AGI) подразумевает способность ИИ-систем выполнять множество задач и включает такие навыки как организация рассуждений, формирование и использование модели мира и модели себя, рефлексия и самокритика, целеполагание и планирование. Агенты Auto-GPT, BabyAGI, BabyBeeAGI [AutoGPT: build & use AI agents, 2023] реализуют модель AGI посредством циклического вызова LLM до тех пор, пока агент не достигнет поставленной цели, генерируя новые подзадачи. AGI-концепция «LLM как операционная система, естественный язык как язык программирования» (см. [LLM OS Experiments, 2023], [Ge et al., 2023], [Wu et al., 2024]) подразумевает создание LLM-агента для управления компьютером (Computer-Using Agent). В 2025 г. эта концепция получила коммерческую реализацию в виде Operator от OpenAI [Computer-Using Agent, 2025].

Рассуждающие LLM. В 2024 г. появился новый класс LLM, позиционируемых как «рассуждающие» (reasoning) [Kumar et al., 2025]. Первой такой моделью стала OpenAI o1 [Learning to Reason with LLMs, 2024]. Прирост качества происходит за счёт скрытых рассуждений LLM перед ответом (inference-time scaling). Официальные советы по промпт-инжинирингу для o1 рекомендовали делать запросы простыми и короткими, а также избегать запросов цепочек рассуждений. В начале 2025 появилась первая публично доступная рассуждающая модель DeepSeek-R1 [Guo et al., 2025], обученная с помощью крупномасштабного обучения с подкреплением (RL) на значительно меньших вычислительных ресурсах, причем была достигнута производительность, сопоставимая с OpenAI o1. В работе [Li et al., 2025] была представлена парадигма многомодальных рассуждений, чередующих текстовые и визуальные шаги, которая улучшает как качество, так и интерпретируемость рассуждений за счет их визуализации. Модели OpenAI o3 и o4-mini также получили дополнитель-

ную способность "мыслить образами" [OpenAI O3 and O4 Mini, 2024]. В работе [Ma et al., 2025] идея визуальных рассуждений была реализована для диффузионных моделей.

Использование рассуждающих моделей в задачах автоматического доказательства теорем привело в последние годы к значительному прогрессу. Если ранее использовался подход, основанный на последовательном пошаговом прогнозировании доказательства, то теперь непосредственно генерируется целое доказательство: DeepSeek-Prover [Xin et al., 2024], Goedel-Prover [Lin et al., 2025]. В работе Kimina-Prover Preview [Wang et al., 2025] предложена схема совместного использования формальных и неформальных рассуждений, а также обучения с подкреплением для математических рассуждений «в стиле человека» (с анализом частных случаев и т.п.), что позволяет решить большее количество сложных задач, например, в области комбинаторики [Liu et al., 2025].

Перспективный подход к обучению рассуждающих моделей предложен в работе Absolute Zero [Zhao et al., 2025], где способы рассуждения выучиваются обучением с подкреплением в открытой вселенной задач, вообще без участия человека даже на этапе постановки задач: задачи также генерируются автоматически при помощи RL.

Были выявлены и проблемы, связанные с рассуждающими моделями, такие как избыточные рассуждения (overthinking, [Kumar et al., 2025], [Cuadron et al., 2025]) и «нечестность» объяснения рассуждений [Attribution Graphs in Biology, 2025]). Таким образом, полностью положиться на рассуждающие LLM, исключив инженерию запросов, гибридные и «прозрачные» методы ИИ, пока невозможно.

Моделирование процессов человеческой психики также является важным направлением при создании генеративных агентов и элементов AGI. Это такие исследования как эмоциональный ИИ, создание двойников личности, моделирование самосознания как нарратива личной истории (эгоцентрический storytelling), которые отрабатываются, в частности, в задачах автоматического литературного творчества. В работе WhatELSE [Lu et al., 2025] представлена интерактивная система создания повествований, использующая ИИ для развития повествовательных пространств и генерации разнообразных сюжетов на основе примеров историй. В работе [Gurung et al., 2025] демонстрируется концепция нарративного ИИ (Narrative AI), а также совершенствование генерации длинных историй с помощью обучения с подкреплением (RL) для улучшения рассуждений в LLM. Подобный подход далее может быть распространен на создание диалоговых систем, профессиональных ассистентов, поддержку принятия решений в сложных длительных процессах и другие приложения ИИ, которые необходимо выполнять «в стиле человека».

Заключение

Возможность практического использования описанных результатов и технологий в значительной степени определяется набором и качеством доступных в каждый момент времени лучших моделей LLM и MLLM. По состоянию на август 2025 года в качестве лидирующих моделей можно выделить: Claude 4 (Opus 4 и Sonnet 4) [Claude-4, 2025] от Anthropic, Grok 4 [Grok-4, 2025] от xAI, Llama 4 (Maverick и Scout) [Llama4, 2025], Gemini 2.5 (Flash и Pro) [Gemini-2-5, 2025] от Google, Gemma 3 [Gemma-3, 2025] и Gemma 3n [Gemma-3n, 2025]) от Google DeepMind, Mistral 3 [Mistral.AI Models, 2025] (Mistrall Small 3.2 [Mistral.AI Small-3-1, 2025], Magistral Small [Mistral.AI Magistral, 2025], Devstrall Small [Mistral.AI Devstral, 2025]), Qwen3 (think и instruct) [Qwen3, 2025], DeepSeek-V3.1 [DeepSeek-V3.1, 2025], GPT-5 [OpenAI GPT-5/, 2025] и GPT-OSS [OpenAI GPT-OSS/, 2025] (120B и 20B) от OpenAI.

В качестве общих тенденций, отличающих современное поколение LLM и MLLM, можно отметить следующие:

- Разделение LLM на thinking и instruction модели (для сложных рассуждений и простых задач соответственно). Например, GPT-5 представляет собой уже не одну LLM, а набор моделей и маршрутизатор, распределяющий задачи между моделями.
- Модели сразу формируются и обучаются как агенты, в частности, адаптированные для использования инструментов и программирования.
- Использование архитектур типа MoE для ускорения на этапе выполнения.
- Увеличение длины контекстного окна.
- Параллельное исследование нескольких гипотез при рассуждениях.
- Интенсивное использование RL и синтетических данных на этапе обучения. Например, командой Qwen разработан алгоритм GSPO (Group Sequence Policy Optimization) [Zheng et al., 2025], призванный заменить популярный алгоритм GRPO при больших размерах и разряженности (MoE) обучаемых моделей.

Следует ожидать, что в 2026 году мы увидим не менее серьезные продвижения как в функциональности LLM, так и в методах их обучения и использования.

Список литературы

[Achiam et al., 2023] Achiam J. et al. Gpt-4 technical report // arXiv preprint arXiv:2303.08774. – 2023.

[Adam et al., 2021] Adam S., et al. Open-Ended Learning Leads to Generally Capable. Agents // arXiv:2107.12808. – 2021.

[Aditya et al., 2022] Aditya R., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents // arXiv:2204.06125. – 2022.
[Alphacode 2, 2023] AlphaCode Team G. Alphacode 2 technical report. – Technical report. – URL https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2_Tech_Report. pdf, 2023.
[Andy et al., 2022] Andy Z., et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language // arXiv:2204.00598. – 2022.
[Azizi et al., 2024] Azizi S., et al. LaMDA: Large Model Fine-Tuning via Spectrally Decomposed Low-Dimensional Adaptation // arXiv: 2406.12832v1. – 2024.
[Badia et al., 2020] Badia A.P., et al. Agent57: Outperforming the atari human benchmark // arXiv preprint arXiv:2003.13350. – 2020.
[Barrault et al, 2024] Barrault L. et al. Large Concept Models: Language Modeling in a Sentence Representation Space // arXiv preprint arXiv:2412.08821. – 2024.
[**Behrouz et al., 2024**] Behrouz A. et al. Titans: Learning to Memorize at Test Time // arXiv:2501.00663v1. – 2024.
[**Behrouz et al., 2025**] It's All Connected: A Journey Through Test-Time Memorization, Attentional Bias, Retention, and Online Optimization // arXiv preprint arXiv: 2504.13173. – 2025.
[Bengio et al., 2021] Bengio E. et al. Flow network based generative models for non-iterative diverse candidate generation // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 27381-27394.
[Black et al., 2024] Black K. et al. π0: A vision-language-action flow model for general robot control. – 2024. – URL https://arxiv. org/abs/2410.24164.
[Boiko et al., 2023] Boiko D.A., MacKnight R., Gomes G. Emergent autonomous scientific research capabilities of large language models // arXiv preprint arXiv:2304.05332. – 2023.
[Bousmalis et al., 2023] Bousmalis K. et al. Robocat: A self-improving foundation agent for robotic manipulation // arXiv preprint arXiv:2306.11706. – 2023.
[Bran et al., 2023] Bran A.M. et al. Chemcrow: Augmenting large-language models with chemistry tools // arXiv preprint arXiv:2304.05376. – 2023.
[Brohan et al., 2023] Brohan A. et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control // arXiv preprint arXiv:2307.15818. – 2023.
[Brooks et al., 2024] Brooks T. et al. Video generation models as world simulators. – 2024.
[Brown et al., 2020] Brown T. et al. Language models are few-shot learners // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 1877-1901.
[Chen et al., 2024] Chen G., Wang W. A survey on 3d gaussian splatting // arXiv preprint arXiv:2401.03890. – 2024.
[Chen et al., 2025] Chen Q. et al. AI4Research: A Survey of Artificial Intelligence for Scientific Research // arXiv preprint arXiv: 2507.01903. – 2025.
[Cheng et al., 2024] Cheng T. et al. Yolo-world: Real-time open-vocabulary object detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2024. – P. 16901-16911.
[Chervonyi et al., 2025] Chervonyi Y. et al. Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2 // arXiv preprint arXiv:2502.03544. – 2025.

[Child, 2020] Child R. Very deep vaes generalize autoregressive models and can outperform them on images // arXiv preprint arXiv:2011.10650. – 2020.
[Cuadron et al., 2025] Cuadron A. et al. The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks // arXiv preprint arXiv:2502.08235. – 2025.
[DeepSeek-AI, 2025] DeepSeek-AI Utilizing the Distilled Model from DeepSeek-R1 for Efficient Fine-Tuning with LoRA and Chain-of-Thought Datasets // arXiv:2406.15734v2. – 2025.
[Dong et al., 2023] Dong Q. et al. A survey on in-context learning // arXiv preprint arXiv:2301.00234. – 2023.
[Esser et al., 2024] Esser P. et al. Scaling rectified flow transformers for high-resolution image synthesis // arXiv preprint arXiv:2403.03206. – 2024.
[Gao et al., 2024] Gao Y. et al. Retrieval-augmented generation for large language models: A survey // arXiv preprint arXiv:2312.10997. – 2024.
[Ge et al., 2023] Ge Y. et al. Llm as os (llmao), agents as apps: Envisioning aios, agents and the aios-agent ecosystem // arXiv preprint arXiv:2312.03815. – 2023.
[Gu et al., 2021] Gu A. et al. Efficiently modeling long sequences with structured state spaces // arXiv preprint arXiv:2111.00396. – 2021.
[Gu et al., 2023] Gu A. et al. Mamba: Linear-time sequence modeling with selective state spaces // arXiv preprint arXiv:2312.00752. – 2023.
[Guo et al., 2025] Guo D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning // arXiv preprint arXiv:2501.12948. – 2025.
[Gurung et al., 2025] Gurung A. et al. Learning to Reason for Long-Form Story Generation // arXiv preprint arXiv:2503.22828. – 2025.
[Halverson et al., 2024] Halverson J. et al. KAN: Kolmogorov–Arnold Networks // arXiv preprint arXiv:2404.19756v1 – 2024.
[Hao et al., 2023] Hao R. et al. Chatllm network: More brains, more intelligence // arXiv preprint arXiv:2304.12998. – 2023.
[Ho et al., 2020] Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 6840-6851.
[Ho et al., 2022] Ho J. et al. Imagen video: High definition video generation with diffusion models // arXiv preprint arXiv:2210.02303. – 2022.
[Ho et al., 2024] Ho M. et al. Proof Flow: Preliminary Study on Generative Flow Network Language Model Tuning for Formal Reasoning // arXiv preprint arXiv:2410.13224. – 2024.
[Hu et al., 2022] Hu E.J. et al. Lora: Low-rank adaptation of large language models // ICLR. – 2022. – Vol. 1, No. 2. – P. 3.
[Hu et al., 2023] Hu Z. et al. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models // arXiv preprint arXiv:2304.01933. – 2023. MLA.
[Jain et al., 2024] Jain A. et al. Prompt Tuning Strikes Back: Customizing Foundation Models with Low-Rank Prompt Adaptation // arXiv:2405.15282v1. – 2024.
[Kerbl et al., 2023] Kerbl B. et al. 3d gaussian splatting for real-time radiance field rendering // ACM Transactions on Graphics. – 2023. – Vol. 42, No. 4. – P. 1-14.
[Kingma et al., 2013] Kingma D.P., Welling M. Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. – 2013.

[Kirillov et al., 2023] Kirillov A. et al. Segment anything // Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2023. – P. 4015-4026.
[Kumar et al., 2025] Kumar A. et al. OverThink: Slowdown Attacks on Reasoning LLMs // arXiv preprint arXiv:2502.02542. – 2025.
[Kumar et al., 2025] Kumar K. et al. Llm post-training: A deep dive into reasoning large language models // arXiv preprint arXiv:2502.21321. – 2025.
[Laskin et al., 2022] Laskin M. et al. In-context reinforcement learning with algorithm distillation // arXiv preprint arXiv:2210.14215. – 2022.
[Lewis et al., 2020] Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459-9474.
[Li et al., 2025] Li C. et al. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought // arXiv preprint arXiv:2501.07542. – 2025.
[Lieber et al., 2024] Lieber O. et al. Jamba: A hybrid transformer-mamba language model // arXiv preprint arXiv:2403.19887. – 2024.
[Lin et al., 2025] Lin et al. Goedel-Prover: A Frontier Model for Open-Source Automated Theorem Proving // arXiv preprint arXiv: 2502.07640. – 2025.
[Ling et al., 2022] Ling Y., et al. Diffusion Models: A Comprehensive Survey of Methods and Applications // arXiv:2209.00796. – 2022.
[Liu et al., 2025] Liu B. et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems // arXiv preprint arXiv:2504.01990. – 2025. MLA.
[Liu et al., 2025] Liu J. et al. CombiBench: Benchmarking LLM Capability for Combinatorial Mathematics // arXiv:2505.03171v1. – 2025.
[Liu et al., 2023] Liu S. et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection // arXiv preprint arXiv:2303.05499. – 2023.
[Lu et al., 2024] Lu C. et al. The AI scientist: Towards fully automated open-ended scientific discovery // arXiv preprint arXiv:2408.06292. – 2024.
[Lu et al., 2025] Lu Z. et al. WhatELSE: Shaping Narrative Spaces at Configurable Level of Abstraction for AI-bridged Interactive Storytelling // arXiv preprint arXiv:2502.18641. – 2025.
[Ma et al., 2025] Ma N. et al. Inference-time scaling for diffusion models beyond scaling denoising steps // arXiv preprint arXiv:2501.09732. – 2025.
[Ma et al., 2023] Ma Y.J. et al. Eureka: Human-level reward design via coding large language models // arXiv preprint arXiv:2310.12931. – 2023.
[Mnih et al., 2015] Mnih V., et al. Human-level control through deep reinforcement learning // Nature. 2015.
[Nie et al., 2025] Nie S. et al. Large language diffusion models // arXiv preprint arXiv:2502.09992. – 2025.
[Ouyang et al., 2022] Ouyang L., et al. Training language models to follow instructions with human feedback // Advances in Neural Information Processing Systems. – 2022. – Vol. 35. – P. 27730-27744.
[Peng et al., 2023] Peng B. et al. Rwkv: Reinventing rnns for the transformer era // arXiv preprint arXiv:2305.13048. – 2023.

[Poli et al., 2023] Poli M. et al. Hyena hierarchy: Towards larger convolutional language models // International Conference on Machine Learning. – PMLR, 2023. – P. 28043-28078.
[Pternea et al., 2024] Pternea M. et al. The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models // Journal of Artificial Intelligence Research. – 2024. – Vol. 80. – P. 1525-1573.
[Rishi et al., 2021] Rishi B., et al. On the Opportunities and Risks of Foundation Models // arXiv:2108.07258. – 2021.
[Romera-Paredes et al., 2024] Romera-Paredes B. et al. Mathematical discoveries from program search with large language models // Nature. – 2024. – Vol. 625, No. 7995. – P. 468-475.
[Sahoo et al., 2024] Sahoo P. et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications // arXiv preprint arXiv:2402.07927. – 2024.
[Schulman et al., 2017] Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms // arXiv preprint arXiv:1707.06347. – 2017.
[Scott et al., 2022] Scott R. et al. A Generalist Agent // arXiv:2205.06175. – 2022.
[Shi et al., 2024] Shi J. et al. Simplified and generalized masked diffusion for discrete data // Advances in neural information processing systems. – 2024. – Vol. 37. – P. 103131-103167.
[Shinn et al., 2023] Shinn N., Labash B., Gopinath A. Reflexion: an autonomous agent with dynamic memory and self-reflection // arXiv preprint arXiv:2303.11366. – 2023.
[Sun et al., 2023] Sun Y. et al. Retentive network: A successor to transformer for large language models // arXiv preprint arXiv:2307.08621. – 2023.
[Takase et al., 2024] Takase R. et al. GFlowNet Fine-tuning for Diverse Correct Solutions in Mathematical Reasoning Tasks // arXiv preprint arXiv:2410.20147. – 2024.
[Tian et al., 2025] Tian Y. et al. Yolov12: Attention-centric real-time object detectors // arXiv preprint arXiv:2502.12524. – 2025.
[Trinh et al., 2024] Trinh T.H. et al. Solving olympiad geometry without human demonstrations //Nature. – 2024. – Vol. 625, No. 7995. – P. 476-482.
[Vaswani et al., 2017] Vaswani A. et al. Attention is all you need // Advances in neural information processing systems. – 2017. – Vol. 30.
[Wan et al., 2024] Wan Z. et al. Efficient large language models: A survey // arXiv preprint arXiv:2312.03863. – 2024.
[Wang et al., 2022] Wang X. et al. Self-consistency improves chain of thought reasoning in language models // arXiv preprint arXiv:2203.11171. – 2022.
[Wang et al., 2025] Wang et al. Kimina-Prover Preview: Towards Large Formal Reasoning Models with Reinforcement Learning // arXiv preprint arXiv:2504.11354. – 2025.
[Wang et al., 2025] Wang K.A. et al. Test-time regression: a unifying framework for designing sequence models with associative memory // arXiv preprint arXiv: 2501.12352. – 2025.
[Wang et al., 2024] Wang W. et al. CogVLM: Visual expert for pretrained language models //Advances in Neural Information Processing Systems. – 2024. – Vol. 37. – P. 121475-121499.

[Wu et al., 2024] [Wu L. et al., 2024] Wan Z. et al. Efficient large language models: A survey // arXiv preprint arXiv:2312.03863. – 2024.
[Wu et al., 2024] Wu Z. et al. Os-copilot: Towards generalist computer agents with self-improvement // arXiv preprint arXiv:2402.07456. – 2024.
[Xin et al., 2021] Xin H., Kaiyong Z., Xiaowen C. AutoML: A Survey of the State-of-the-Art // arXiv:1908.00709v6. – 2021.
[Xin et al., 2024] Xin et al. DeepSeek-Prover: Advancing Theorem Proving in LLMs through Large-Scale Synthetic Data // arXiv preprint arXiv: 2405.14333. – 2024.
[Yao et al., 2023] Yao S. et al. React: Synergizing reasoning and acting in language models // arXiv preprint arXiv:2210.03629. – 2023.
[Yao et al., 2023] Yao Y., Li Z., Zhao H. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models // arXiv preprint arXiv:2305.16582. – 2023.
[Ye et al., 2024] Ye J. et al. Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models // arXiv preprint arXiv:2402.07754. – 2024.
[Yin et al., 2024] Yin S. et al. A survey on multimodal large language models // arXiv preprint arXiv:2306.13549. – 2024.
[Zhang et al., 2019] Zhang S., Tong H., Xu J., Maciejewski R. Graph convolutional networks: a comprehensive review // Comput Soc Netw 6. – 2019. – No. 11.
[Zhang et al. 2024] Zhang F., Pilanci M. Spectral Adapter: Fine-Tuning in Spectral Space // arXiv:2405.13952 – 2024.
[Zhang et al., 2024] Zhang J.-C. et al. LoRA²: Multi-Scale Low-Rank Approximations for Fine-Tuning Large Language Models // arXiv:2408.06854v1. – 2024.
[Zhao et al., 2025] Zhao A. et al. Absolute Zero: Reinforced Self-play Reasoning with Zero Data // arXiv preprint arXiv:2505.03335. – 2025.
[Zheng et al., 2025] Group Sequence Policy Optimization Group sequence policy optimization // arXiv preprint arXiv:2507.18071. – 2025.
[Zhou et al., 2024] Zhou C., et al. RankAdaptor: Hierarchical Rank Allocation for Efficient Fine-Tuning Pruned LLMs via Performance Model // arXiv:2406.15734v2. – 2024.
[Zhou et al., 2024] Zhou X. et al. GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting // arXiv preprint arXiv:2402.07207. – 2024.
[Zielonka et al., 2023] Zielonka W. et al. Drivable 3d gaussian avatars // arXiv preprint arXiv:2311.08581. – 2023.

Электронные ресурсы

[AI Co-scientist, 2025] Accelerating scientific breakthroughs with an AI co-scientist // google. – URL: https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/ (дата обращения: 31.08.2025).
[AI Foresight, 2024] Дмитрий Чернышенко провёл стратегическую форсайт-сессию по фундаментальным исследованиям в сфере искусственного интеллекта // Официальный сайт Правительства Российской Федерации. – URL: http://government.ru/news/51726/ (дата обращения: 31.08.2025).

[Attribution Graphs in Biology, 2025] Attribution Graphs in Biology // Transformer Circuits Blog. – URL: https://transformer-circuits.pub/2025/attribution-graphs/biology.html (дата обращения: 28.04.2025).
[AutoGPT: build & use AI agents, 2023] AutoGPT: build & use AI agents // GitHub. – URL: https://github.com/Significant-Gravitas/AutoGPT (дата обращения: 10.12.2023).
[Claude-4, 2025] www.anthropic.com. – URL: https://www.anthropic.com/news/claude-4 (дата обращения: 31.08.2025).
[Computer-Using Agent, 2025] Computer-Using Agent: a universal interface for AI to interact with the digital world // OpenAI – URL: https://openai.com/index/computer-using-agent/ (дата обращения: 21.04.2025).
[Deepmind VEO, 2025] deepmind.google. – URL: https://deepmind.google/models/veo/ (дата обращения: 31.08.2025).
[DeepSeek, 2025] DeepSeek // Официальный сайт. – URL: https://www.deepseek.com/ (дата обращения: 24.04.2025).
[DeepSeek-V3.1, 2025] DeepSeek-V3.1 Release // deepseek.com. – URL: https://api-docs.deepseek.com/news/news250821 (дата обращения: 31.08.2025).
[Gemini-2-5, 2025] Try Deep Think in the Gemini app // blog.google. – URL: https://blog.google/products/gemini/gemini-2-5-deep-think/ (дата обращения: 31.08.2025).
[Gemma-3, 2025] huggingface.co. – URL: https://huggingface.co/google/gemma-3-12b-it (дата обращения: 31.08.2025).
[Gemma-3n, 2025] huggingface.co. – URL: https://huggingface.co/google/gemma-3n-E4B-it (дата обращения: 31.08.2025).
[Grok-4, 2025] x.ai. – URL: https://x.ai/news/grok-4 (дата обращения: 31.08.2025).
[Improving Image Generation, 2024] Improving Image Generation with Better Captions // Semantic Scholar. – URL: https://www.semanticscholar.org/paper/Improving-Image-Generation-with-Better-Captions-Betker-Goh/cfee1826dd4743eab44c6e27a0cc5970effa4d80 (дата обращения: 21.02.2024).
[Learning to Reason with LLMs, 2024] Learning to Reason with LLMs // OpenAI. – URL: https://openai.com/index/learning-to-reason-with-llms/ (дата обращения: 16.04.2025).
[Llama4, 2025] Welcome Llama 4 Maverick & Scout on Hugging Face // huggingface.co. – URL: https://huggingface.co/blog/llama4-release (дата обращения: 31.08.2025).
[LLM Architecture Evolution, 2025] Эволюция архитектур больших языковых моделей: от GPT-2 к современным решениям // habr.com. – URL: https://habr.com/ru/articles/931382/ (дата обращения: 31.08.2025).
[LLM OS Experiments, 2023] LLM OS Experiments // LLM-OS.net. URL: http://llm-os.net/ (дата обращения: 01.12.2023).
[LLM Powered Autonomous Agents, 2023] LLM Powered Autonomous Agents // LilianWeng. – URL: https://lilianweng.github.io/posts/2023-06-23-agent/ (дата обращения: 11.12.2023).
[Mistral.AI Devstral, 2025] Upgrading agentic coding capabilities with the new Devstral models // mistral.ai. – URL: https://mistral.ai/news/devstral-2507 (дата обращения: 31.08.2025).

[Mistral.AI Magistral, 2025] Magistral // mistral.ai. – URL: https://mistral.ai/news/magistral (дата обращения: 31.08.2025).
[Mistral.AI Models, 2025] huggingface.co. – URL: https://huggingface.co/mistralai/models (дата обращения: 31.08.2025).
[Mistral.AI Small-3-1, 2025] mistral.ai. – URL: https://mistral.ai/news/mistral-small-3-1 (дата обращения: 31.08.2025).
[Model Context Protocol, 2025] Model Context Protocol // github.com. – URL: https://github.com/modelcontextprotocol (дата обращения: 31.08.2025).
[OpenAI GPT-5, 2025] Introducing GPT-5 // openai.com. – URL: https://openai.com/index/introducing-gpt-5/ (дата обращения: 31.08.2025).
[OpenAI GPT-OSS, 2025] Introducing gpt-oss // openai.com. – URL: https://openai.com/index/introducing-gpt-oss/ (дата обращения: 31.08.2025).
[OpenAI O3 and O4 Mini, 2024] OpenAI. Introducing O3 and O4 Mini // OpenAI. – URL: https://openai.com/index/introducing-o3-and-o4-mini/ (дата обращения: 28.04.2025).
[OpenAI PhD-level agents, 2025] OpenAI plots charging $20,000 a month for PhD-level agents // The Information. – URL: https://www.theinformation.com/articles/openai-plots-charging-20-000-a-month-for-phd-level-agents (дата обращения: 21.04.2025).
[Prompt Engineering Guide, 2023] Prompt Engineering Guide // Prompt Engineering Guide. URL: https://www.promptingguide.ai/ (дата обращения: 16.12.2023).
[Qwen3, 2025] Qwen3TechnicalReport // arxiv.org. – URL: https://arxiv.org/pdf/2505.09388 (дата обращения: 31.08.2025).
[Your Guide to Generative AI, 2023] Your Guide to Generative AI // Learn Prompting. – URL: https://learnprompting.org/ (дата обращения: 19.12.2023).

УДК 004.8
doi: 10.15622/rcai.2025.002

СИТУАЦИОННОЕ УПРАВЛЕНИЕ: МОДИФИКАЦИЯ РЕШЕНИЯ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ

Б.А. Кобринский (*kba_05@mail.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В работе рассматривается модифицированный вариант ситуационного управления, предполагающий, взамен решения на основе случайного выбора, передачу управления гибридной интеллектуальной системе (ГИС). Представлена структура интегрированной системы, состоящей из технологии ситуационного управления и ГИС, включающей базу правил, библиотеку прецедентов и нейросеть. Рассмотрена возможность получения пользователем контекстно-зависимых изображений в процессе принятия решения.

Ключевые слова: ситуационное управление, гибридная интеллектуальная система, интеграция подходов, ситуации неопределенности, критические инфраструктуры.

Введение

Разработанный Д.А. Поспеловым метод ситуационного управления [Поспелов, 1971], [Поспелов, 1986], [Astafiev et al., 1986] позволил не только резко повысить эффективность прикладных решений, но и серьезно продвинуться в развитии методологии и расширении инструментального «арсенала» систем управления. Основой методов построения систем управления явились семиотические модели представления объектов управления и описания процедур управления [Pospelov et al., 1995], [Pospelov, 1996]. В их основу была положена идея о том, что любая ситуация, которая может возникнуть в физическом мире, может быть описана через конечное число базовых отношений, из которых при необходимости могут быть порождены производные отношения.

Развитие теории ситуационного управления продолжается [Jakobson et al., 2007], [Madarász et al., 2009], [Gorodetskiy, 2020], [Kovalenko, 2022]. Нечеткое ситуационное управление позволяет вырабатывать управляющие

решения в соответствии с выбранной стратегией управления и учитывать специфику системы благодаря композиционной модели [Борисов и др., 2021]. Предложены архитектурные решения построения конвергентных систем на основе знаний, включающих иерархию моделей, описывающих различные аспекты целевой системы ситуационного управления [Kovalenko, 2022]. В качестве варианта решения в тех случаях, когда коррелятор системы ситуационного управления не может осуществить выбор, взамен перехода управления к блоку случайного выбора, предложено передавать решение гибридной интеллектуальной системе, включающей базу правил и искусственную нейронную сеть (для распознавания изображений) [Кобринский, 2024]. Развитию данного подхода посвящена настоящая статья. Это крайне важно для критических инфраструктур – авиакосмической отрасли, оборонной сферы, здравоохранению и др.

Проблемы в управлении диагностическим и лечебным процессом

Ситуационное исчисление, предназначенное для представления и рассуждений о динамических мирах и эффектах действий во времени, разработанное Джоном Маккарти, использует ситуации (отражающие текущее состояние мира), действия (изменяющие ситуацию) и флюенты (предикаты, истинностные значения которых могут меняться со временем) для моделирования эволюции мира [McCarthy, 2002]. Эволюция организма в процессе жизни человека и в течении патологического процесса включает множество медленно текущих процессов, что не исключает резких изменений в состоянии человека (под действием внутренних или экофакторов), которые могут представлять из себя опасные ситуации. При этом отдельные факторы могут существенно изменяться во времени. Необходимым является выявление предвестников предкритических и критических состояний, только часть из которых являются сигналами с датчиков (киберфизических систем).

Важно выявить и оценить будущие целевые желательные и нежелательные состояния объекта управления, которым является организм пациента, и наиболее существенные факторы, влияющие на переход пациента из одного состояния в другое. Значительная неопределенность в поведении многофакторной системы организма со сложной системой управляющих переходов, сопровождающаяся нечеткостью и неполнотой данных, создает серьезные проблемы в принятии решений как врачами, так и интеллектуальными системами.

Система управления с параметрической неопределенностью [Sugiki et al., 2006] может соответствовать высокой изменчивости параметров организма. Подчиненное управление, применяемое в многоконтурных

системах для случая параметрической неопределенности объекта, позволяет получить желаемое качество регулирования не только системы в целом, но и каждого контура [Опейко, 2015]. Это важно в управлении взаимосвязанными системами организма. Квазиинвариантное управление возможно в случаях, когда неизвестны не только параметры управляющей функции, но и параметры самого объекта управления, в условиях априорной параметрической неопределённости и при наличии внешнего воздействия [Гельфер и др., 2013]. В медицине это имеет место в различных ситуациях, например, при лабильной форме артериальной гипертонии, в период резких изменений солнечной активности. Распознавание многообразных патологических ситуаций и процесс лечения, обеспечивающий регресс болезни, являются важнейшей проблемой в медицине. В решении этой проблемы существенную роль может сыграть интеллектуальная система управления.

Управление медико-технологическим процессом осложняется неравновесным состоянием многочисленных параметров организма больного человека. Особенности управления в такой ситуации сводятся к следующему [Kobrinskii, 2024]:

- достижение ближайшей цели (подцели общей целевой ситуации) с учетом условий безопасности пациента, определяемых уровнем риска применяемых методов обследования и лечения;
- преодоление неопределенности данных о состоянии организма путем учета и анализа второстепенных параметров;
- оперативный совокупный интеллектуальный анализ изменений параметров, получаемых с киберфизических систем (КФС) и дискретной информации из других источников;
- уточнение прогноза течения болезни, выбор методов дополнительного обследования и лечения;
- планирование действий (исследований, манипуляций, методов терапевтического и хирургического лечения) на основе возможных траекторий принятия решений в системе управления;
- выбор последовательности управляющих решений по переходу в целевую ситуацию в зависимости от выявленных точек перехода организма (и отдельных его подсистем) из одного фазового состояния в другое [Кобринский, 2023];
- адаптации управления к изменению структуры и параметров объекта (системных и внешних факторов) в процессе функционирования системы.

Подходы к управлению медико-технологическим процессом

В процессе мультипараметрического мониторинга от киберфизических систем контроля жизненно важных функций организма может постоянно поступать значительный объем данных [Saeed et al., 2011],

[Davoudi et al., 2019]. В этом смысле медико-технологическому процессу можно сопоставить сложный технологический производственный процесс. В статье [Кулинич, 2016] представлены подходы к принятию решений в плохо определенных и слабоструктурированных ситуациях, представленных в лингвистическом виде в форме экспертных знаний и субъективных оценок. Рассмотрены ситуационный, когнитивный и семиотический подходы к поддержке принятия решений в таких случаях. На оперативном уровне, при оказании медицинской помощи должны учитываться логистические проблемы, управление данными и алгоритмическое управление [Shung et al., 2021].

Аналогом того, что необходимо в управлении диагностическим и лечебным процессом является сложная техническая система (СТС), характеризующаяся многокомпонентностью, большим числом количественно-качественных параметров, нелинейностью отношений, неполнотой информации, разнообразием воздействий внутренних и внешних факторов, рисками возникновения опасных ситуаций и катастрофичностью их последствий [Борисов и др., 2021]. Это позволяет учитывать специфику ситуационного управления в зависимости от текущего состояния системы в условиях нечетких признаков, нечетких ситуаций и нечетких управляющих решений, множества нечетких управляющих переходов между нечеткими ситуациями. И множество маршрутов между различными идентифицированными текущими и целевыми нечеткими ситуациями. Все перечисленное имеет место в медицинской предметной области и создает серьезные трудности в представлении этих проблем и их реализации в форме управляющей системы.

Контур управления здоровьем человека

В архитектуре ситуационной системы управления сложным поливариантным лечебно-диагностическим процессом необходима гармонизация различных компонент, соответствующих группам бизнес-процессов на разных этапах клинических путей. Это определяется наличием ряда развилок и траекторий в соответствии с текущим состоянием пациента, что определяется состоянием систем организма и оказываемыми на них лечебно-профилактическими воздействиями.

В контуре управления здоровьем человека находится врач. В соответствии с этим сформулируем, в соответствии с [Поспелов, 1986], следующие определения [Kobrinskii, 2024].

Определение 1. Будем называть текущей ситуацией на объекте управления (процесс диагностики и/или лечения конкретного человека) совокупность всех доступных сведений об объекте управления и его функционировании в данный момент времени.

Определение 2. Будем называть прогностической ситуацией совокупность возможных траекторий течения патологического процесса, включая осложнения (отклонения в работе разных систем организма), являющиеся следствием негативного развития болезни или побочным результатом применения лекарственных средств.

Определение 3. Будем называть полной ситуацией совокупность, состоящую из текущей ситуации, знаний о состоянии системы (организма) в данный момент времени и знаний о методах управления (теоретически возможных профилактических и лечебных воздействиях), которые должны быть выданы врачу-пользователю в форме возможных решений. Таким образом, если на объекте управления сложилась ситуация *Qj*, при которой состояние системы и технологическая схема управления, определяемые *Si*, допускают использование воздействия *Uk* (при условии безопасности применяемых методов для конкретного пациента с учетом его состояния), то оно применяется, и текущую ситуацию *Qj* предлагается, с использованием логико-трансформационных правил, преобразовать в новую ситуацию *Qi*.

Принципы построения модифицированной системы ситуационного управления

Методы ситуационного управления должны функционировать как ансамбль с передачей управления.

Рассмотрим систему ситуационного управления при условии, что корреляtор не способен осуществить выбор. В этом случае, взамен передачи управления блоку случайного выбора, передадим решение гибридной интеллектуальной системе (ГИС). С этого момента единая база знаний будет применяться для управления и поддержки решений, учитывая, что для формирования классов ситуаций и правил используются знания экспертов проблемной области. В то время как текущая информация из электронных медицинских карт (ЭМК) пациентов, включая диагностически значимые изображения, будет поступать в рабочую область системы.

Архитектура ГИС может включать базу знаний, библиотеку прецедентов и искусственную нейросеть для распознавания изображений. Результаты распознавания изображений будут передаваться в базу знаний и обеспечивать возможность комплексного анализа ситуаций с использованием всех имеющихся данных. Развитие искусственного интеллекта позволяет также сделать вывод о возможности интеграции в базе знаний вербализованных данных и результатов распознавания визуальных данных (images) нейросетью. Это отвечает представлению Д.А. Поспелова о том, что внедрение в системы ситуационного управления процедур, позволяющих работать с видеообразами ситуаций, сулит качественный ска-

чок на всех шагах процесса выработки управляющих решений в интегрированных интеллектуальных управляющих системах [Поспелов, 1995]. Повышение распознавания уникальных (нетипичных) случаев будет реализовано за счет поиска сходных описаний больных в библиотеке прецедентов [Khan et al., 2019], [Грибова и др., 2023].

На рис. 1 представлена схема, демонстрирующая передачу решения от системы ситуационного управления к гибридной интеллектуальной системе.

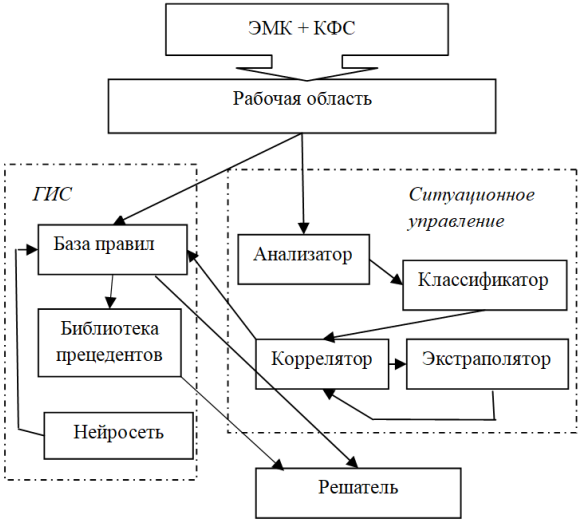


Рис. 1. Интегрированная с электронной медицинской картой интеллектуальная управляющая система

Таким образом, в сложных случаях принятие диагностических решений может осуществляться последовательно с использованием ситуационного управления, базы правил, аргументации на прецедентах и распознавания изображений на нейросети.

Развитие человеко-машинного взаимодействия в процессе принятия решений позволяет рассмотреть ситуацию использования сходных изображений в сложных ситуациях. В [Milov et al., 2022] рассматривается проблема пересмотра трансформации системы управления в направлении присутствия человека как в объекте управления, так и в контуре управления. Исходя из этого, в базе знаний ГИС, наряду с лингвистическими правилами, могут присутствовать ряды соответствующих образов [Кобринский, 2022]. Они могут предоставляться пользователю системы как дополнительная информация в контексте с предлагаемым решением.

Заключение

Нечеткое ситуационное управление в принятии решений в неполностью определенных ситуациях (например, на этапе предварительной диагностики) крайне актуально как в медицине, так и в технологических процессах.

Ансамблевая концепция управления медико-технологическим процессом, с передачей решения гибридной интеллектуальной системе в условиях невозможности в технологии ситуационного управления другого выбора, кроме случайного, позволит повысить эффективность принятия решений в характеризующейся нечеткостью, неопределенностью и недоопределенностью данных и ситуаций медицинской предметной области.

Список литературы

[Борисов и др., 2021] Борисов В.В., Авраменко Д.Ю. Нечеткое ситуационное управление сложными системами на основе их композиционного гибридного моделирования // Системы управления, связи и безопасности. – 2021. – № 3. – С. 207-237. – doi: 10.24412/2410-9916-2021-3-207-237.

[Гельфер и др., 2013] Гельфер И.С., Котельников И.В., Теклина Л.Г. Синтез системы управления с эталонной моделью в условиях параметрической неопределенности объекта управления и наличия внешнего возмущения // Вестник Нижегородского университета им. Н.И. Лобачевского. Серия: Математическое моделирование. Оптимальное управление. – 2013. – № 4(1). – С. 204-207.

[Грибова и др., 2023] Грибова В.В., Ковалев Р.И., Окунь Д.Б. Система назначения персонифицированного лечения по аналогии на основе гибридного способа извлечения прецедентов // Программные продукты и системы. – 2023. – Т. 36, № 3. – С. 486-492. – doi: 10.15827/0236-235X.142.486-49.

[Кобринский, 2022] Кобринский Б.А. Образы в системах искусственного интеллекта: поиски и перспективы // В: Всероссийская конференция «Поспеловские чтения: искусственный интеллект – проблемы и перспективы», Поспеловские чтения-2022 (Москва, 19-20 декабря 2022 г.): Труды конф. – М.: Изд-во ФИЦ ИУ РАН, 2022. – С. 32-42.

[Кобринский, 2023] Кобринский Б.А. О моделировании переходных состояний организма // Вестник Тверского государственного технического университета. Серия «Технические науки». – 2023. – № 1(17). – С. 79-86. – doi: 10.46573/2658-5030-2023-1-79-86.

[Кобринский, 2024] Кобринский Б.А. Ситуационное управление и поддержка на этапах медико-технологического процесса // В: Гибридные и синергетические интеллектуальные системы: Сборник статей VII Всероссийской Поспеловской конференции [Электронный ресурс]: научное электронное издание / А.В. Колесников, отв. ред. – Калининград, Санкт-Петербург: Изд-во РХГА, 2024. – С. 22-31.

[Кулинич, 2016] Кулинич А.А. Ситуационный, когнитивный и семиотический подходы к принятию решений в организациях // Открытое образование. – 2016. – № 6. – С. 9-17. – https://doi.org/10.21686/1818-4243-2016-6-9-17.

[Опейко, 2015] Опейко О.Ф. Подчиненное управление объектом с параметрической неопределенностью // Системный анализ и прикладная информатика. – 2015. – № 3. – С. 21-24.

[Поспелов, 1971] Поспелов Д.А. Принципы ситуационного управления // Известия АН СССР, Техническая кибернетика. – 1971. – № 2. – С. 10-17.

[Поспелов, 1986] Поспелов Д.А. Ситуационное управление: теория и практика. – М.: Наука, 1986.

[Поспелов, 1995] Поспелов Д.А. Ситуационное управление: новый виток развития // Известия РАН. Теория и системы управления. – 1995. – № 5. – С. 152-159.

[Astafiev et al., 1986] Astafiev V.I., Gorsky Y.M., Pospelov D.A. Contradictions in the control of large systems // Computers and Artificial Intelligence. – 1986. – Vol. 5. – P. 89-102.

[Davoudi et al., 2019] Davoudi, A., Malhotra, K.R., Shickel, B. et al. Intelligent ICU for Autonomous Patient Monitoring Using Pervasive Sensing and Deep Learning // Scientific Reports. 2019. Vol.9. Article number: 8020 doi.org/10.1038/s41598-019-44004-w.

[Gorodetskiy, 2020] Gorodetskiy A.E. The Principles of Situational Control SEMS Group // In: Smart Electromechanical Systems. Studies in Systems, Decision and Control. Vol 261 / Gorodetskiy, A., Tarasova, I., eds. – Cham: Springer, 2020. – doi.org/10.1007/978-3-030-32710-1_1.

[Jakobson et al., 2007] Jakobson G., Buford J., Lewis L. Situation Management: Basic Concepts and Approaches // In: Information Fusion and Geographic Information Systems. Lecture Notes in Geoinformation and Cartography / Popovich V.V., Schrenk M., Korolenko K.V., eds. – Berlin, Heidelberg: Springer, 2007. – P. 18-33.

[Khan et al., 2019] Khan M.J., Hayat H., Awan I. Hybrid case-base maintenance approach for modeling large scale case-based reasoning systems // Hum. Cent. Comput. Inf. Sci. – 2019. – Vol. 9. – Article number:9. – doi.org/10.1186/s13673-019-0171-z.

[Kobrinskii, 2024] Kobrinskii B. Fuzzy Situational Control at the Stages of the Medical-and-Technological Process: Problems and Possible Solutions // Proceedings of the Eighth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’24). Vol. 1. Lecture Notes in Networks and Systems. – Vol. 1209. – P. 312-323. – doi: 10.1007/978-3-031-77688-5_30.

[Kovalenko, 2022] Kovalenko O. Knowledge Driven Cyber-Convergent Systems Based on Situational Agents // In: 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT). – IEEE, 2022. – P. 243-246.

[Madarász et al., 2009] Madarász L., Andoga R., Fozo L., Lazar T. Situational Control, Modeling and Diagnostics of Large Scale Systems // In: Towards Intelligent Engineering and Information Technology. Studies in Computational Intelligence. – Vol. 243 / I.J. Rudas, J.odor, J. Kacprzyk, eds. – Berlin, Heidelberg: Springer, 2009. – P. 153-164. – doi.org/10.1007/978-3-642-03737-5_11.

[McCarthy, 2002] McCarthy J. Actions and other events in situation calculus // Proc. of Proceedings of the Eighth Intern. Conf. on Principles of Knowledge Representation and Reasoning (KR-2002) / D. Fensel et al., eds. – San Francisco: Morgan Kaufmann Publ. 2002. – P. 615-628.

[Milov et al., 2022] Milov O., Khvostenko V., Voropay N. et al. Situational Control of Cyber Security in Socio-Cyber-Physical Systems // In: 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). 09-11 June 2022, Ankara, Turkey. – IEEE, 2022. – doi: 10.1109/HORA55278.2022.9800049.

[Pospelov et al., 1995] Pospelov D.A., Ehrlich A.I., Osipov G.S. Semiotic Modeling and Situation Control // In: Proceedings of 1995 ISIC Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems / J. Albus, A. Meystel, D. Pospelov, T. Reader, eds. – AdRem, Cynwyd, 1995. – P. 127-129.

[Pospelov, 1996] Pospelov D.A. Situation Control: an Overview // In: Proceedings of Workshop on Russian Situation Control and Cybernetic / R.J. Strohn, ed. – Battelle, Columbus, 1996. – P. 7-37.

[Saeed et al., 2011] Saeed M., Villarroel M., Reisner A.T. et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database // Crit Care Med. – 2011. – Vol. 39, No. 5. – P. 952-960. – doi: 10.1097/CCM.0b013e31820a92c6.

[Shung et al., 2021] Shung D.L, Sung J.J.Y. Challenges of developing artificial intelligence-assisted tools for clinical medicine // J. Gastroenterol. Hepatol. 2021. – Vol. 36, No. 2. – P. 295-298. – doi: 10.1111/jgh.15378.

[Sugiki et al., 2006] Sugiki A., Furuta K. Posicast Control Design for Parameter-Uncertain Plants // In: Proceedings of the 45th IEEE Conference on Decision and Control (Diego, CA, USA, 13-15 December 2006). – P. 3192-3197. – doi: 10.1109/CDC.2006.376979.

УДК 004.8

АВТОМАТИЧЕСКОЕ МАШИННОЕ ОБУЧЕНИЕ ДЛЯ БОЛЬШИХ ФУНДАМЕНТАЛЬНЫХ МОДЕЛЕЙ

А.В. Бухановский (*avbukhanovskii@itmo.ru*)

Университет ИТМО, Санкт-Петербург

Интенсивное внедрение современных технологий искусственного интеллекта (ИИ) на основе больших фундаментальных моделей (БФМ) в различных отраслях экономики и социальной сферы требует новых механизмов кастомизации и автоматизации разработки таких решений под конкретную задачу. Если для моделей ИИ, использующих классические методы работы с данными, эта проблема успешно решается за счет использования технологий автоматического машинного обучения (AutoML), то для больших моделей ИИ данный вопрос остается открытым. В докладе предлагается подход к автоматизации конструирования и обучения мультиагентных систем ИИ на основе БФМ, способный учитывать как индивидуальные характеристики агентов, так и топологию их взаимодействия. Для этого вводится класс интеллектуальных мета-агентов, способных динамически объединять различных прикладных агентов на принципах композитного ИИ. Для реализации данного подхода предлагается использовать «лабораторию агентов» – программно-аппаратную среду, предоставляющую агентам возможность доступа к данным и вычислениям, а также эффективного использования различных больших языковых моделей. На основе «лаборатории агентов» допустимо, в том числе, моделировать различные когнитивные эффекты, присущие системам сильного ИИ. Обсуждаемые подходы и технологии будут проиллюстрированы на задачах создания БФМ для нефтегазовой промышленности, градостроительства и поддержки научной деятельности.

Ключевые слова: большие фундаментальные модели, автоматическое машинное обучение, мультиагентная система

УДК 004.8

АДАПТАЦИЯ ПРЕДМЕТНОЙ ОБЛАСТИ И ОБОБЩЕНИЕ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ

А. Кумар (*abhinavkumar@ee.iith.ac.in*)

Индийский технологический институт, Хайдарабад, Индия

Модели глубокого обучения добились успеха в таких областях, как робототехника, медицинская визуализация и автономное вождение. Однако их эффективность часто снижается при сдвиге предметной области, когда статистические свойства данных развертывания отличаются от данных обучения. Этот сдвиг может происходить как ковариационный, априорный, условный или смешанный сдвиг, и может существенно ограничить надежность модели в реальных конвейерах. Для решения этих проблем исследования были сосредоточены на неконтролируемой адаптации предметной области, когда немаркированные целевые данные доступны во время обучения, и обобщении предметной области, когда целевые данные недоступны. В докладе рассматриваются как унимодальные (только зрительные), так и многомодальные (зрительные языковые) задачи, включая адаптацию предметной области с одним источником, несколькими источниками и несколькими целями, а также обобщение предметной области. Также рассматриваются практические приложения, такие как адаптация RGB-подсветки к тепловизионным изображениям для распознавания жестов. В докладе рассматриваются ограничения существующих состязательных и самообучающихся подходов и предлагаются целевые решения для развития методов адаптации предметной области и обобщения предметной области.

Ключевые слова: модели глубокого обучения, адаптации предметной области, обобщение предметной области.

УДК 004.8

РАССУЖДАЮЩИЕ МОДЕЛИ: ЗА И ПРОТИВ

С.И. Николенко (*s.nikolenko@spbu.ru*)

Санкт-Петербургский государственный университет,
Санкт-Петербург

2025 год стал для искусственного интеллекта годом рассуждающих моделей. В октябре 2024-го вышел o1-preview, в январе 2025-го DeepSeek выпустил модель R1, и практически сразу же все ведущие модели стали рассуждающими. В докладе мы поговорим о том, что такое рассуждающие модели, откуда они происходят и как работают. Мы рассмотрим последние результаты о том, насколько рассуждения действительно помогают и насколько они полезны для других целей, в частности для обеспечения безопасности искусственного интеллекта.

Ключевые слова: рассуждающие модели, безопасность систем искусственного интеллекта.

Секция 1 | ИНЖЕНЕРИЯ ЗНАНИЙ

УДК 004.8
doi: 10.15622/rcai.2025.003

КАРТОГРАФИЯ И СЕМАНТИКА НАУЧНОГО ЗНАНИЯ: ПИЛОТНЫЙ ПРОЕКТ

Т.А. Гаврилова (*gavrilova@gsom.spbu.ru*) [A]
В. Шванкин (*v.shvankin@salesai.ru*) [B]
М.В. Кубельский (*m.kubelskiy@gsom.spbu.ru*) [A]
Н.В. Иваникова (*n.ivanikova@spbu.ru*) [C]
В. Луцков (*st098065@student.spbu.ru*) [A]

[A] Высшая школа менеджмента,
Санкт-Петербургский государственный университет,
Санкт-Петербург
[B] ООО “Герофарм”, Санкт-Петербург
[C] Управление научных исследований,
Санкт-Петербургский государственный университет,
Санкт-Петербург

В докладе представлена методика визуализации библиометрических данных исследователей одного из подразделений университета на основе семантики текстов публикаций. С помощью глубокого рассмотрения метаданных публикаций, включающего семантический анализ названий и аннотаций, сформирован прототип интерактивных карт научных интересов и связей между авторами. Такой подход позволяет выявить ключевые направления исследований, междисциплинарные взаимодействия и структуру научного коллектива, что способствует более эффективному управлению научной деятельностью и развитию сотрудничества внутри университета.

Ключевые слова: карты знаний, библиометрия, семантический анализ.

Введение

Интеллектуальный капитал университета, как сумма всех нематериальных активов, включает в себя в первую очередь человеческий капитал, т.е. знания и навыки преподавателей, их публикации и выступления [Полюшкевич, 2018]. Долгое время этот капитал трудно поддавался измерению и какой-либо автоматизации в управлении [García-Carbonell et al., 2021], [Гаврилова и др., 2010]. Анализ библиометрических данных является важным инструментом управления и мониторинга этого актива [Руководство по наукометрии: индикаторы развития науки и технологии, 2021] в наукоемких коллективах [Donthu et al., 2021].

Специальное программное обеспечение, представляющее результаты библиометрического анализа в графическом виде, делает его доступным для широкого круга пользователей (VOSviewer, CiteSpace, SciVal, Bibliometrix / Biblioshiny, eLIBRARY). Визуализация результатов анализа позволяет создавать комплексные карты научного знания с учетом семантики научных текстов. Автоматизированный семантический анализ определенных разделов научных публикаций, таких как, например, название и аннотация, может обеспечить объективное представление об интеллектуальных активах, которыми обладает научно-исследовательское или образовательное учреждение [Chagnon et al., 2024].

В докладе используются и развиваются некоторые из результатов проекта «Методология и технология разработки цифровых карт знаний для учебных и научных коллективов (МЕТАКАРТА)» (грант РНФ 2023-24 № 23-21-00168).

1. Теоретические основы картографии научного знания

Библиометричекий анализ научных публикаций является ценным инструментом для оценки результатов исследований, влияния и тенденций [Donthu et al., 2021], [McAlliste et al, 2022]. Для эффективной передачи сложных данных и взаимосвязей, выявленных с помощью библиометрии, визуальные диаграммы играют решающую роль. Они облегчают идентификацию закономерностей, упрощают коммуникацию и помогают в принятии решений [Дудко и др., 2023].

Идея визуализации сетей цитирования и картирования науки восходит к 60-м годам 20-го века и тесно связана с развитием наукометрии. Один из основоположников наукометрии Д. Прайс предложил на основе анализа сетей цитирования документов определять фронты развития научных исследований [Price, 1965]. В настоящее время при построении библиометрических карт чаще всего используют анализ цитирования и со-встречаемости ключевых слов в документах. Метод ко-цитирования основан на выявлении публикаций, которые совместно цитируются другими авторами [Маршакова, 1973], Small, 1973]. При кластеризации на основе совместного использо-

вания ключевых слов используют автоматическое извлечение ключевых слов из текстов или предоставленные авторами ключевые слова [Van Eck et al., 2014]. Современные подходы в области обработки естественного языка и кластеризации векторных представлений текстов, такие как тематическое моделирование – мощный инструмент, создающий новые возможности для картирования науки [Thijs et al. 2018].

Основными типами библиометрических визуальных диаграмм являются [Subramanyam, 1983], [Судакова и др., 2025]:

- Сети цитирования: визуализируют связи между публикациями по цитированию.
- Сети совместных ключевых слов: раскрывают темы и взаимосвязи.
- Диаграммы влияния цитирования: представляют количество цитирований и импакт-факторы публикаций, позволяя сравнивать высокоцитируемые работы.
- Карты сотрудничества: иллюстрирует отношения на основе соавторства.
- Карты предметных категорий на основе существующих рубрикаторов: иерархически организуют научные публикации и их взаимосвязи по категориям.

Многие авторы обсуждают важность полноты и точности данных, которая может повлиять на достоверность визуализаций [Руководство по наукометрии: индикаторы развития науки и технологии, 2021]. Следует также учитывать субъективность интерпретации и особенности публикационной активности в изучаемой области знания. Кроме того, многие исследователи сталкиваются с ограничениями программного обеспечения [Moral-Munoz et al, 2020].

2. Методология построения карт научного знания

Пилотная разработка библиометрических карт на основе семантического анализа потребовала решения ряда задач в рамках концепции (proof-of-concept), т.е. доказательства жизнеспособность выбранного подхода.

Для этого в рамках исследования были выделены следующие этапы (подробнее описаны в следующих параграфах):

a) Выбор показателей и метрик.
b) Анализ и отбор источников данных.
c) Сбор и описание данных.
d) Разработка прототипа архитектуры программного обеспечения.
e) Анализ и визуализация данных.
f) Создание сценариев использования.

a) Выбор показателей

В работе использовался реестр авторов в привязке к кафедрам. Он содержит ссылки на профили сотрудников на основных библиографических ресурсах. Реестр включает в себя справочную информацию о сотрудниках и содержит наборы полей: *ФИО/Факультет/Должность/Степень/ Elibrary_SPIN/Scopus_AuthorID/WoS_ResearcherID /WoS_link.*

Базовыми показателями, используемыми для анализа результативности ученых, являются количество публикаций и цитирований, индекс Хирша.

При анализе библиометрических данных важно учитывать отличия в характере публикационной активности в различных областях знаний. Например, в общественных и гуманитарных науках влияние книжных, а также не англоязычных публикаций, многие из которых не индексированы в международных наукометрических базах данных, выше, чем в естественных науках [Руководство по наукометрии: индикаторы развития науки и технологии, 2021], [Aksnes et al., 2021]. Таким образом, для увеличения охвата публикаций и более полного анализа следует обращаться в том числе к данным из национальных указателей цитирования, включающих не англоязычные публикации [Sile et al., 2017]. При этом, наряду с публикациями в периодических научных изданиях, целесообразно включать и другие типы публикаций, в частности монографии и сборники материалов конференций.

Для обеспечения возможности построения различных карт знаний были сформулированы минимальные требования к данным. Они представлены в табл. 1.

Таблица 1

Метаданные и показатели

А. Мета-данные автора /участника	Б. Библиометрические показатели по каждому автору/сотруднику	С. Семантически значимая информация для определения областей компетенции
• Фамилия • Имя • Уникальный идентификатор	• Список публикаций • Количество цитирований для каждой публикации • Индекс Хирша и другие имеющиеся агрегированные показатели	• Названия публикаций • Аннотации публикаций

b) Анализ источников данных

В качестве источников данных о результатах научной деятельности сотрудников были рассмотрены следующие ресурсы сети Интернет:

- РИНЦ (https://elibrary.ru/).

- Scopus (https://www.scopus.com/home.uri).
- Web of Science (https:// clarivate.com/).
- Google Scholar (https://scholar.google.com/).

Для выбора источника данных были сформированы следующие критерии: (а) возможность автоматической загрузки данных / (б) свободный доступ / (в) полнота данных / (г) учет публикаций и на английском, и на русском языке.

Библиографические базы данных Scopus, Web of Science и РИНЦ не соответствуют критериям (а) и (б), так как для получения доступа к API системы требуется оплата подписки. В Scopus и Web of Science не англоязычные публикации недостаточно представлены [Aksnes et al., 2021], а значит не соблюдается еще и критерий (г).

Google Scholar, представляющий собой автоматизированный агрегатный сервис, осуществляющий сбор метаданных публикаций путем сканирования ресурсов сети Интернет, обеспечивает достаточно высокую полноту охвата публикаций. При этом данные Google Scholar находятся в свободном доступе с возможностью их автоматической загрузки.

Таким образом, в качестве источника данных был выбран Google Scholar, отвечающий большинству критериев.

c) Разработка прототипа архитектуры программного обеспечения

Для автоматического сбора данных, их анализа и визуализации карт знаний был разработан программный модуль BIB-METR.2 на языке Python, включающий библиотеки для поиска и подключения к прокси Google Scholar, парсинга и обработки данных [Kubelski et al., 2024], семантического анализа и создания графов – scholarly, pandas, PyVis, NetworkX, spaCy, Pickle, SentenceTransformers и другие. Разработка оригинального модуля позволила обеспечить гибкость работы с необходимой структурой данных и создала платформу для дальнейших этапов исследования.

d) Анализ и визуализация данных

BIB-METR.2 реализует систему семантического анализа и визуализации научной активности. Результатом работы модуля стала интерактивная карта знаний, основанная на библиометрических и текстовых данных, агрегированных по кафедрам, научным сотрудникам и тематическим направлениям.

Построенная карта знаний обладает чётко выраженной структурной иерархией, обеспечивая возможность анализа научной активности на нескольких уровнях детализации. Визуализация выполнена с использованием графовой модели (рис. 1), где:

- кафедры представлены в виде прямоугольных узлов синего цвета;
- исследователи (авторы) – в виде жёлтых узлов круглой формы, размер которых зависит от количества публикаций и уровня цитируемости;
- ключевые исследовательские термины – в виде небольших красных узлов.

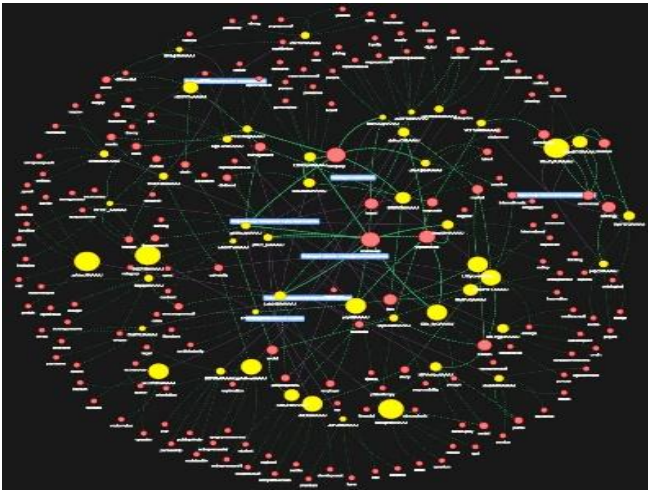


Рис. 1. Исследователи и их научные интересы

Таким образом, достигается высокая степень наглядности, что способствует быстрой идентификации центров научной активности, ключевых участников и направлений научной работы. Кроме того, визуализация позволяет выявлять междисциплинарные взаимодействия и исследовательские пробелы, ранее не поддававшиеся прямой аналитике.

Модуль обеспечивает интерактивную работу с данными, включая фильтрацию, детализацию связей, переход к исходной публикационной информации, что расширяет его аналитический потенциал и делает его применимым как для внутренних нужд вуза, так и для внешних заинтересованных сторон.

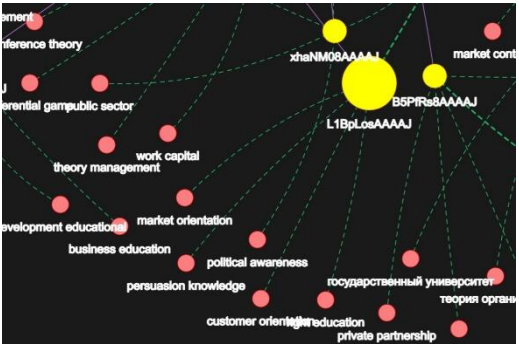


Рис. 2. Увеличенный фрагмент карты знаний

е) О сценариях использования

Ниже представлены сценарии для различных групп пользователей:

е1) Для исследователей система служит инструментом анализа текущих научных трендов и поиска перспективных направлений. Карта позволяет:

- выявить коллег, работающих в смежных тематических областях;
- формировать междисциплинарные научные коллективы;
- оценить собственную публикационную активность и позиционирование в научном сообществе университета.

е2) Руководство вуза получает возможность принимать обоснованные стратегические решения на основе агрегированной информации о:

- ведущих научных направлениях;
- ключевых исследовательских группах;
- динамике публикационной активности и цитируемости.

Формализованные аналитические запросы, доступные в системе:

- Определение авторов, объединяющих несколько кафедр.
- Идентификация направлений с единственным специалистом.
- Выявление направлений с высокой динамикой публикационной активности.

е3) Для сторонних пользователей (бизнеса, государственных структур, фондов) карта знаний выступает в роли инструмента:

- быстрой оценки научного потенциала университета;
- выбора исполнителей под прикладные научные задачи;
- принятия решений о финансировании и развитии партнёрств.

Таким образом, реализация модуля BIB-METR.2 создаёт предпосылки для формирования единой аналитико-коммуникационной платформы, способствующей эффективному взаимодействию между всеми заинтересованными субъектами: научным сообществом, администрацией вуза и внешними партнёрами.

Заключение

В докладе обсуждаются принципы разработки и предварительные результаты пилотного проекта "НАучная ВИзуализациЯ для сотрудничества и развития" (НАВИЯ). Проект ориентирован на создание методологии разработки библиометрических карт знаний кафедр и факультетов университета. В ходе работы над проектом предложен и апробирован подход к обобщению аналитической информации о публикационной активности научно-педагогических работников вуза. Дальнейшие направления развития предлагаемого подхода включают использование прототипа для расширения палитры анализа и углубления семантического анализа для выявления неочевидных связей и ассоциаций.

Основные бенефициары проекта и решаемые задачи:

- Сами члены научного коллектива и отдельные исследователи: выявление пробелов и ниш в исследованиях, поиск партнеров для сотрудничества, оценка эффективности, мониторинг трендов.
- Руководство университетов и научных институтов: принятие решений о финансировании, развитии приоритетных направлений, оценке результативности научных групп.
- Государственные органы и бизнес: определение приоритетных направлений развития науки и технологий, оценка вклада отдельных ученых и организаций в развитие страны.

Список литературы

[Гаврилова и др., 2010] Гаврилова Т.А., Кудрявцев Д.В. Информационные технологии управления знаниями // В книге: Инновационное развитие: экономика, интеллектуальные ресурсы, управление знаниями. – М.: ИНФРА-М, 2009. – С. 500-515.

[Дудко и др., 2023] Дудко В.В., Патаракин, Е.Д. Исследование научных школ университета средствами библиометрического картирования. Территория новых возможностей // Вестник Владивостокского государственного университета экономики и сервиса. – 2023. – Т. 15, № 1. – С. 150-167. – doi: 10.24866/VVSU/2949-1258/2023-1/150-167.

[Маршакова, 1973] Маршакова И.В. Система связей между документами, построенная на основе ссылок: по данным Science Citation Index // Научно-техническая информация. Сер. 2. – 1973. – № 6. – С. 3-8.

[Полюшкевич, 2018] Полюшкевич О.А. Интеллектуальный капитал университета // Alma mater (Вестник высшей школы). – 2018. – № 5. – С. 25-27. – doi: 10.20339/AM.05-18.025.

[Руководство по наукометрии: индикаторы развития науки и технологии, 2021] Акоев М.А., Маркусова В.А., Москалева О.В., Писляков В.В. Руководство по наукометрии: индикаторы развития науки и техники. – М.: Изд-во Уральского университета, 2021. – 358 с. – doi: 10.15826/B978-5-7996-1352-5.0000.

[Судакова и др., 2025] Судакова А.Е., Агарков Г.А. Датасет о наукометрии российских ученых: кейс e. Library // Вопросы образования. – 2025. – № 1. – С. 304-330. – doi: 10.17323/vo-2025-21514.

[**Aksnes et al., 2021**] Aksnes D.W., Sivertsen G. A Criteria-based Assessment of the Coverageof Scopus and Web of Science // Journal of Data and Information Science. – 2019. – Vol. 4(1). – P. 1-21. – doi: 10.2478/jdis-2019-0001.

[**Donthu et al., 2021**] Donthu N., Kumar S., Mukherjee D., Pandey N., and Lim W.M. How to conduct a bibliometric analysis: An overview and guidelines // Journal of business research. – 2021. – No. 133. – P. 285-296. – doi: 10.1016/j.jbusres.2021.04.070.

[García-Carbonell et. al., 2021] García-Carbonell N., Guerrero-Alba F., Martín-Alcázar F., Sánchez-Gardey G. Academic human capital in universities: definition and proposal of a measurement scale. Science and public policy. – 2021. – Vol. 48(6), 877. – doi: 888. 10.1093/scipol/scab062.

[**Kubelskiy et al., 2024**] Kubelskiy M., Kuznetsova A., Leshcheva I., Gavrilova T., Shvankin V. Ontology-Based Approach for Research Activity Mapping // International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russian Federation. – 2024. – P. 334-338. – doi: 10.1109/SIBIRCON63777.2024.10758489.

[McAllister et al., 2022] McAllister J.T., Lennertz L., Mojica Z.A. Mapping a discipline: a guide to using VOSviewer for bibliometric and visual analysis. Science & Technology Libraries. – 2022. – Vol. 41, No. 3. – P. 319-348. – doi: 10.26811/peuradeun.v12i3.1125.

[**Moral-Munoz et al., 2020**] Moral-Muñoz J.A., Herrera-Viedma E., Santisteban-Espejo A., and Cobo M.J. Software tools for conducting bibliometric analysis in science: An up-to-date review // Information Professional. – 2020. – 29(1). – doi: 10.3145/epi.2020.ene.03.

[**Van Eck et al., 2014**] Van Eck N.J., Waltman L. Visualizing bibliometric networks. In Measuring scholarly impact: Methods and practice. – Springer, 2015. – P. 285-320. – doi: 10.1007/978-3-319-10377-8_13.

[**Price, 1965**] Price D.J.D.S. Networks of Scientific Papers // Science. – 1965. – Vol. 149(3683). – P. 510-515. – doi: 10.1126/science.149.3683.510.

[Chagnon et al., 2025] Chagnon E., Pandolfi R., Donatelli J., Ushizima D. Benchmarking topic models on scientific articles using BERTeley // Natural Language Processing Journal. – 2024. – Vol. 6(100044). – doi: 10.1016/j.nlp.2023.100044.

[**Sile et al.**, **2017**] Sile L., Guns R., Sivertsen G., Engels T. European Databases and Repositories for Social Sciences and Humanities Research Output. Report. Antwerp: ECOOM & ENRESSH. – 2017. – doi: 10.6084/m9.figshare.5172322 2017.

[**Small, 1973**] Small H. Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents // Journal of the American Society for Information Science. – 1973. – doi: 10.1002/asi.4630240406.

[**Subramanyam, 1973**] Subramanyam K. Bibliometric studies of research collaboration: A review // Journal of information Science. – 1983. – 6(1). – P. 33-38. – doi: 10.1177/016555158300600.

[**Thijs et al., 2018**] Thiis B., Glanzel W. The contribution of the lexical component in hybrid clustering, the case of four decades of ‘‘Scientometrics’’ // Scientometrics. – 2018. – Vol. 115. – P. 21-33. – doi: 10.1007/s11192-018-2659-0.

УДК 004.822:004.89:004.855
doi: 10.15622/rcai.2025.004

ОРГАНИЗАЦИЯ СОДЕРЖАТЕЛЬНОГО ДОСТУПА К СИСТЕМАТИЗИРОВАННЫМ ЗНАНИЯМ И РЕСУРСАМ ПО МАШИННОМУ ОБУЧЕНИЮ НА ОСНОВЕ ОНТОЛОГИИ

Ю.А. Загорулько (*zagor@iis.nsk.su*) [A,B]
Г.Б. Загорулько (*zagor@iis.nsk.su*) [A,B]
Е.А. Сидорова (*lsidorova@iis.nsk.su*) [A,B]
И.О. Плотникова (*i.plotnikova1@g.nsu.ru*) [B]

[A] Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск
[B] Новосибирский государственный университет, Новосибирск

Несмотря на то, что область машинного обучения (МО) активно развивается, она все еще слабо формализована, а наработанные в ее рамках инструменты и ресурсы недостаточно систематизированы. Это не только удлиняет период вхождения в область МО, но и затрудняет пользователям эффективный выбор необходимых для решения их задач инструментов и ресурсов. Такое положение дел в области МО вызывает необходимость разработки информационно-аналитического интернет-ресурса, который обеспечивал бы систематизацию знаний и информационных ресурсов по МО и содержательный доступ к накопленным в этой области инструментам, моделям, методам и наборам данных. В докладе описывается подход к построению такого ресурса, базирующегося на разработанной авторами онтологии машинного обучения.

Ключевые слова: машинное обучение, онтология, информационно-аналитический интернет-ресурс, паттерны онтологического проектирования.

Введение

В настоящее время все больше людей оказывается вовлечено в область машинного обучения (МО) [Mohri et al., 2018]. Среди них – преподаватели и студенты, осваивающие МО, ученые, использующие МО в своих исследованиях, представители индустрии и органов власти, применяющие ме-

тоды МО для решения своих практических задач. Несмотря на то, что область МО бурно развивается, она до сих пор слабо формализована, а наработанные в ее рамках методы, средства и ресурсы недостаточно систематизированы. Это не только удлиняет период вхождения в область МО, но и затрудняет пользователям эффективный поиск и выбор необходимых для решения их задач методов, моделей, инструментов и наборов данных.

Такое положение дел в области МО диктует потребность в интернет-ресурсе, который обеспечивал бы систематизацию накопленных в этой области научных знаний и данных и поддерживал содержательный доступ к ним. На данный момент в сети Интернет присутствует большое число ресурсов, относящихся к области МО.

Самым известным в России ресурсом такого рода является MachineLearning.ru – русскоязычный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных (ИАД). Ресурс строится по принципам Википедии. Сейчас ресурс содержит около 1100 статей на русском языке, предназначенных как для профессиональных аналитиков, так и для студентов и преподавателей. На нем представлены учебные курсы по МО и ИАД, информация о наиболее важных публикациях, конференциях и семинарах, достижениях научных школ России и стран СНГ в области МО и ИАД.

Этот ресурс может быть полезным для студентов в качестве источника учебных материалов по МО и ИАД, а для специалистов в этих областях – в качестве справочного материала и источника информация о достижениях в этой области. Но доступ к представленным на нем знаниям и данным затруднен из-за отсутствия их четкой систематизации. Кроме того ценность ресурса снижается из-за отсутствия информации о наиболее востребованных алгоритмах, фреймворках и доступа к существующим в свободном доступе наборам данных и обученным моделям.

Из зарубежных ресурсов по МО необходимо отметить, прежде всего, веб-платформы компаний Kaggle и Hugging Face.

На веб-платформе Kaggle (https://www.kaggle.com/) пользователи могут публиковать наборы данных, исследовать и создавать модели, взаимодействовать с другими специалистами по анализу данных и машинному обучению, а также организовывать соревнования по анализу данных и участвовать в них. На платформе размещены наборы открытых данных, предоставляются облачные инструменты для обработки данных и машинного обучения. На этом же ресурсе размещены и реализованы обучающие ресурсы по МО.

Веб-платформа компании Hugging Face (https://huggingface.co/) предоставляет доступ к инструментам и датасетам для создания приложений с использованием машинного обучения. На ней доступны библиотека Transformers, созданная для приложений обработки естественного языка,

и платформа Hugging Face Hub, которая позволяет пользователям обмениваться моделями машинного обучения и наборами данных, а также демонстрировать свою работу. На этой платформе размещено большое количество предварительно обученных моделей, которые поддерживают решение задач в различных модальностях, таких как обработка естественного языка, компьютерное зрение и аудио.

Описанные выше веб-платформы, как и подобные им зарубежные веб-ресурсы, предоставляют ценные данные и инструменты, полезные как для начинающих пользователей, так и для продвинутых разработчиков МО-приложений. Однако они не предоставляют удобного доступа к систематизированным знаниям об области МО и используемым в ней моделям, методам, наборам данных и метрикам.

Для создания интернет-ресурсов, предоставляющих такой доступ, предлагается использовать технологию построения интеллектуальных информационных интернет-ресурсов [Загорулько и др., 2016], базирующуюся на онтологии предметной области. В связи с этим первым шагом на пути создания такого ресурса является разработка онтологии машинного обучения.

1. Основные понятия и термины области машинного обучения

Прежде, чем приступить к описанию разработки требуемой онтологии МО, рассмотрим базовые понятия и термины области знаний МО, которые обязательно должны быть в ней представлены.

Сначала уточним, что под термином «машинное обучение» мы будем понимать область исследования искусственного интеллекта, связанную с разработкой и изучением статистических алгоритмов, которые могут обучаться на данных и обобщать их на данные, которые они раньше не «видели», и, таким образом, выполнять задачи без явных инструкций [Machine Learning, 2025].

Алгоритмы МО обучают модель на основе примеров данных, известных как данные обучения, для того, чтобы делать предсказания или находить решения, не будучи явно запрограммированными на это.

Модель машинного обучения – это тип математической модели, которая после «обучения» на заданном наборе данных может использоваться для прогнозирования или классификации новых данных. В более широком смысле термин «модель» может относиться к нескольким уровням специфичности, от общего класса моделей и связанных с ними алгоритмов обучения до полностью обученной модели со всеми ее настроенными внутренними параметрами.

Данные в МО представляются в виде датасетов, т.е. наборов данных, используемых для обучения моделей машинного обучения [Mohri et al., 2018]. **Датасеты** состоят из примеров (объектов), представленных набором признаков, а также соответствующих им целевых параметров.

Для облегчения поиска необходимых методов и алгоритмов МО они должны быть систематизированы по различным аспектам, например, по типу МО, по типам решаемых задач и т.п.

Наиболее важным аспектом МО является **тип машинного обучения**. В научной литературе выделяется четыре основных типа МО: обучение с учителем, обучение без учителя, обучение с частичным привлечением учителя и обучение с подкреплением [Mohri et al., 2018].

Важным является также деление методов МО по лежащим в их основе математическим моделям и алгоритмам. По этому аспекту методы МО делятся на **классические методы**, в основе которых лежат статистические модели и алгоритмы, и **методы глубокого обучения**, базирующиеся на использовании нейронных сетей.

Еще одним типом МО, который необходимо выделить, является ансамблевое обучение, представляющее собой технику машинного обучения, основанную на совместном использовании нескольких обученных алгоритмов с целью получения лучшей предсказательной эффективности, чем можно было бы получить от каждого алгоритма по отдельности [Rokach, 2010].

Типовыми задачами машинного обучения являются классификация, регрессия, кластеризация, уменьшение размерности и обнаружение аномалий [Mohri et al., 2018]. Для решения каждого типа задач применяются определенные методы МО, которые в свою очередь применяются для решения различных **прикладных задач** – от диагностики заболеваний до анализа и генерации текстов, изображений, аудио, видео и пр.

Важную роль в оценке алгоритмов машинного обучения играют метрики качества, так как они позволяют определить, насколько хорошо модель работает на данных и какие улучшения ей требуются. В связи с этим в онтологии должны быть представлены все используемые на данный момент метрики, в частности, precision (точность), recall (полнота), F-мера (комбинированная мера) и другие.

Для решения своих задач пользователю необходимо получить доступ не только к методам МО и их реализациям, но и к наборам данных и моделям МО, ранее использованным для решения подобных задач. Следовательно, методы МО должны быть связаны с наборами данных и ранее обученными моделями. В свою очередь модели должны быть связаны с наборами данных, на которых они обучались.

Для моделей и методов глубокого обучения важным аспектом является используемая при их реализации архитектура, которая может включать одну или несколько нейронных сетей. В связи с этим в онтологии должны быть представлены такие архитектуры.

Важно представить в онтологии и информацию о публикациях по МО и информационных ресурсах, которые могут содержать ссылки на описание и реализацию методов и датасетов.

Модели и методы МО используются в каких-то приложениях и при этом работают в каком-то окружении (библиотеки, фреймворки, операционные системы и вычислительные устройства). Эту информацию нужно отразить в онтологии, также как и информацию о персонах и организациях, вовлеченных в область машинного обучения, и различных видах деятельности, выполняемых в области МО.

Таким образом, в онтологии должны быть представлены как понятия, специфичные для области МО, такие как метод (алгоритм), модель, задача, набор данных, метрика, нейросетевая архитектура, окружение, приложение, так и понятия, служащие для описания деятельности, выполняемой в любой научной области: персона, организация, деятельность, публикация, информационный ресурс и др.

2. Разработка онтологии машинного обучения

Для того, чтобы онтология была практически полезной, она должна не только представлять все базовые понятия МО и связи между ними, но и содержать описания соответствующих этим понятиям сущностей, т.е. описания конкретных методов, моделей, наборов данных и т.п.

1.1. Обзор онтологий машинного обучения

На данный момент разработано несколько онтологий, так или иначе относящихся к области машинного обучения.

Прежде всего, стоит отметить разработанную консорциумом W3C онтологическую схему ML-Schema [ML Schema, 2016]. Она предоставляет набор классов, свойств и ограничений, которые можно использовать для представления и обмена информацией об алгоритмах интеллектуального анализа данных и машинного обучения, наборах данных и выполняемых с их использованием экспериментов. Однако ML-Schema не удовлетворяет выдвинутым нами выше требованиям к онтологии МО, та как она не дает полного и целостного представления об области МО. В ней нет описания конкретных методов, алгоритмов, датасетов и задач.

Другая онтология из области МО – SML [Kallab et al., 2023]. Сами авторы позиционируют SML как основанную на онтологии модель для описания семантического машинного обучения. SML описывает модели машинного обучения с помощью понятного человеку и машине словаря, чтобы облегчить понимание, оценку и выбор удобной модели МО для использования в данном контексте. Она позволяет представлять и хранить характеристики и рабочие спецификации уже реализованных моделей

машинного обучения (например, используемые ими алгоритмы и обучающие и тестовые наборы данных, результаты их оценки и т.д.)., что должно облегчить и улучшить для пользователя, обладающего ограниченными знаниями в области машинного обучения, выбор модели машинного обучения, наиболее подходящей для данного контекста и решаемой задачи. Таким образом, эта, безусловно, полезная онтология в основном ориентирована на подробное описание уже реализованных моделей МО, но не содержит сведений о многих базовых понятиях МО и описаний соответствующих им конкретных сущностей.

В близкой к МО области – интеллектуальный анализ данных (ИАД) или data mining – также разработан ряд онтологий. Наиболее известная из них – OntoDM [Džeroski et al., 2008] включает определения основных сущностей ИАД, таких как типы данных и наборы данных, задачи ИАД, алгоритмы ИАД и их компоненты (например, функция расстояния), ограничения и т.д.

Другая онтология – Expos´e [Vanschoren et al., 2010] позволяет подробно описывать эксперименты по анализу данных, включая контекст эксперимента, метрики оценки, методы оценки производительности, наборы данных и алгоритмы.

Обе приведенные выше онтологии, хотя и содержат понятия, являющиеся общими для обеих предметных областей (МО и ИАД), но в них не представлены многие базовые понятия МО и описания соответствующих им сущностей.

Авторам известна только одна онтология MLOnto [Braga et al., 2020], которая содержательно описывает область знаний МО, т.е. в этой онтологии представлены не только понятия, но и конкретные сущности из этой области, например, конкретные алгоритмы МО (Linear Regression, Support Vector Machine и т.п.) и фреймворки (Keras, PyTorch и др.), используемые при решении задач методами МО. К сожалению, эта онтология включает неполный набор базовых понятий, необходимых для описания области МО. В частности, в ней не представлены такие важные понятия, как модели и задачи МО, наборы данных и метрики оценки качества работы алгоритмов МО. Кроме того, в этой онтологии конкретные сущности представлены не объектами (индивидами), а классами, что противоречит принципам онтологического моделирования и делает невозможным детальное описание конкретных сущностей.

Приведенный небольшой обзор показывает, что на данный момент нет онтологии, которая могла бы одновременно и систематизировать фундаментальные знания области МО, и содержать подробные описания наработанных в ней методов, моделей, инструментов и наборов данных. В связи с этим было принято решение разработать новую онтологию МО.

1.2. Реализация онтологии машинного обучения

Реализация новой онтологии МО была выполнена в соответствии с методологией, описанной в [Загорулько и др., 2020]. Данная методология предлагает в качестве основы для построения онтологии целевой научной предметной области (НПО) использовать базовую онтологию научных предметных областей, включающую понятия, характерные для большинства научных предметных областей, и понятия, служащие для описания научно-исследовательской деятельности, а также систему паттернов онтологического проектирования (паттернов ОП), предназначенных для описания решений типовых проблем онтологического инжиниринга.

Построение онтологии конкретной НПО сводится к специализации паттернов ОП на эту область, при необходимости – разработке новых, специфичных для рассматриваемой области паттернов, и дальнейшем построении на их основе фрагментов целевой онтологии путем конкретизации базовых, специализированных и специфичных для этой области паттернов.

Рассмотрим пример специализации представленного в базовой онтологии паттерна Метод исследования на область МО.

В верхней части рис. 1 показан паттерн *Метод исследования*, а в нижней части – паттерн *Метод МО,* полученный в результате его специализация. Данные паттерны реализованы как классы на языке OWL. При этом класс *Метод МО* является подклассом *Метода исследования* и наследует все его свойства. В паттерн этого понятия были добавлены новые связи, отражающие его специфику: *использует Набор данных*, *использует Модель МО*. К *Методу МО применяется Метод оценки качества*, который *использует* разные *Метрики*. Для удобства систематизации *методов МО* среди них выделяются *Классические методы* и *Методы глубокого обучения*. В отдельный класс (*Ансамблевый метод*) выделены методы, использующие ансамбли методов МО.

Классические методы дополнительно группируются по *типу МО* (обучение с учителем, обучение без учителя, обучение с частичным привлечением учителя и обучение с подкреплением).

Разрабатываемая онтология содержит и новые понятия, характерные именно для этой области. Для таких понятий, как *Модель МО*, *Набор данных*, *Метрика* и др., были разработаны паттерны ОП «с нуля».

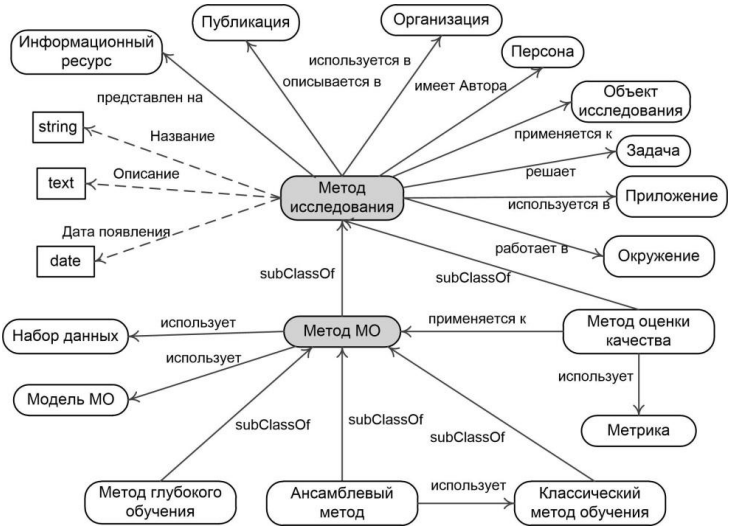


Рис. 1. Специализация паттерна *Метод исследования*

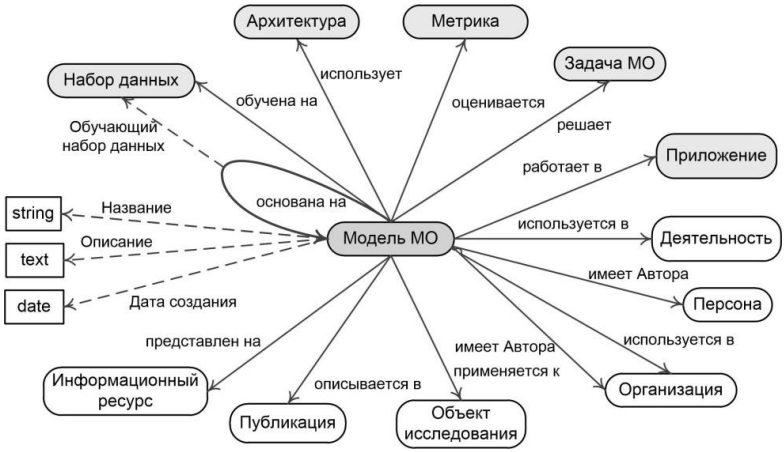


Рис. 2. Паттерн *Модель МО*

На рис. 2 представлен паттерн для описания понятия *Модель МО*. В этом паттерне модель МО связывается с архитектурой, которая используется для реализации модели, с наборами данных, на которых она была обучена, с метриками, используемыми для оценки ее качества, с задачами, для решения которых она предназначена, и с приложениями, в которых она работает. Здесь же могут быть указаны ее авторы и пользователи, а

также публикации о ней и информационные ресурсы, на которых она представлена. Модель МО может основываться на другой модели, т.е. быть получена из нее путем дообучения на каких-то наборах данных. Для представления такой информации вводится атрибутированное отношение "*основана на"*, связывающее две модели и имеющее атрибут "*Обучающий набор данных*". Для этого отношения был разработан структурный логический паттерн.

В связи с наличием огромного количества предварительно обученных нейросетевых моделей в классе *Модель МО* выделяются два подкласса: *Нейросетевая модель* и *Предобученная модель*.

На данный момент существует множество наборов данных, предназначенных для решения задач в различных областях. В связи с этим класс *Набор данных* имеет подклассы: *Набор данных для NLP*, *Набор данных для компьютерного зрения* и *Набор данных общего назначения*.

Рис. 3. Иерархия классов онтологии МО

Базовая онтология НПО и все представленные выше паттерны ОП реализованы на языке OWL. На рис. 3. представлена иерархия классов онтологии МО, построенной в редакторе Protégé.

Включение в онтологию конкретных сущностей выполняется с помощью операции конкретизации паттернов ОП, которая состоит в подстановке в паттерны ОП конкретных значений свойств и добавлении полученных фрагментов в создаваемую онтологию.

3. Реализация ИАИР МО

На основе описанной выше онтологии и упомянутой выше технологии построения интеллектуальных информационных интернет-ресурсов, был реализован ИАИР по машинному обучению (ИАИР МО), который представляет собой доступную через Интернет информационную систему, интегрирующую систематизированные знания, данные и информационные ресурсы из области знаний «Машинное обучение» и обеспечивающую содержательный эффективный доступ к ним.

Рис. 4. Интерфейс информационно-аналитического интернет-ресурса по МО

На рис. 4 показана страница этого ресурса. В левой части страницы представлены понятия онтологии, организованные в иерархию по отношению «общее-частное». При выборе конкретного понятия на странице

отображается поддерево его понятий-потомков и список объектов, соответствующих этому понятию. При выборе какого-либо объекта из этого списка отображается описание его свойств (атрибутов и связей с другими объектами) в табличном или графическом виде. При этом объекты, связанные с данным, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.

Например, на рис. 4 в табличном виде показано описание набора данных Toronto Book Corpus, из которого можно узнать, что этот набор данных собран в 2015 году, состоит из текстов около 7000 книг, включает 985 миллионов слов, использовался в качестве обучающих данных для большой языковой модели BERT и т.д.

На основе онтологии организуется не только навигация по контенту ИАИР МО, но и содержательный поиск. При этом пользователю доступны два вида поиска: простой и расширенный.

Входом для простого поиска является строка, которая ищется по значениям атрибутов всех объектов, содержащихся в контенте. Результатом простого поиска является список рассортированных по классам онтологии объектов, значения атрибутов которых содержат искомую строку.

При расширенном поиске пользователь формулирует запросы через специальный графический интерфейс, управляемый онтологией. Он может выбрать понятие, к которому относится искомый объект, и задать ограничения, которым должны удовлетворять его свойства.

Заключение

В работе описан подход к разработке интеллектуального научного интернет-ресурса, который обеспечивает содержательный доступ к систематизированным знаниям и данным области МО, тем самым помогая пользователям выбирать конкретные инструменты, методы, модели и наборы данных, необходимые для решения их практических задач. В основе данного ресурса лежит разработанная авторами онтология машинного обучения, в которой формализованы и систематизированы как общие знания об области МО, так и накопленные в этой области методы, модели, инструменты и наборы данных.

На данный момент полностью реализован верхний уровень онтологии МО и выполнено ее частичное наполнение конкретными сущностями. На основе данной онтологии построен рабочий прототип ресурса. В ближайших планах – доведение прототипа ресурса до рабочей версии путем добавления в его контент информации обо всех наиболее важных и популярных инструментах и ресурсах из области машинного обучения.

Кроме того, для поддержки актуальности онтологии предполагается дополнить ресурс модулем, реализующим интеграцию ИАИР МО с наиболее значимыми и популярными ресурсами по МО, в частности, с веб-

платформами компаний Kaggle и Hugging Face. Настройка данного модуля на внешние ресурсы будет осуществляться с помощью соответствующих паттернов онтологического проектирования.

Список литературы

[Загорулько и др., 2016] Загорулько Ю.А., Загорулько Г.Б., Боровикова О.И. Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии // Программная инженерия. – 2016. – Т. 7, № 2. – С. 51-60. – doi: 10.17587/prin.7.51-60.

[Загорулько и др., 2020] Загорулько Ю.А., Боровикова О.И. Использование системы разнородных паттернов онтологического проектирования для разработки онтологий научных предметных областей // Программирование. – 2020. – № 4. – С. 27-35. – doi: 10.1134/S0361768820040064.

[Braga et al., 2020] Braga J., Dias J.L.R., Regateiro F. A Machine Learning Ontology. Preprint. October 2020. – doi: 10.31226/osf.io/rc954

[Burkov, 2019] Burkov A. The hundred-page machine learning book. Polen: Andriy Burkov, 2019.

[Džeroski et al., 2008] Džeroski S., Soldatova L., Panov P. OntoDM: An Ontology of Data Mining // In: Proc. 2008 IEEE International Conference on Data Mining Workshops, Pisa, Italy, 2008. – P. 752-760. – doi: 10.1109/ICDMW.2008.62.

[Kallab et al., 2023] Kallab L., Mansour T., Chbeir R. SML: Semantic Machine Learning Model Ontology / In: Proc. Zhang, F., Wang, H., Barhamgi, M., Chen, L., Zhou, R. (eds.) // Web Information Systems Engineering – WISE 2023, LNCS, 2023. – Vol. 14306. – P. 896-911. – Springer, Singapore. – doi: 10.1007/978-981-99-7254-8_70.

[Machine Learning, 2025] Machine Learning. – https://en.wikipedia.org/wiki/Machine_learning#cite_note-1, last accessed 2025/06/10.

[**Mitchell, 1997]** Mitchell T.M. Machine learning. – McGraw-Hill, New York, 1997.

[ML Schema, 2016] ML Schema Core Specification: Release 17 October 2016. – http://ml-schema.github.io/documentation/ML%20Schema.html, last accessed 2024/08/25.

[Mohri et al., 2018] Mohri M., Rostamizadeh A., Talwalkar A. Foundations of machine learning, 2nd edn. – The MIT Press, Cambridge, MA, 2018.

[Rokach, 2010] Rokach L. Ensemble-based classifiers // Artif Intell Rev. – 2010. – Vol. 33. – P. 1-39. – doi: 10.1007/s10462-009-9124-7.

[Vanschoren et al., 2010] Vanschoren J., Soldatova L. Exposé: An ontology for data mining experiments // In: Proc. Int. workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010). – 2010. – P. 31-46.

УДК 620.9:004
doi: 10.15622/rcai.2025.005

МЕТОДЫ ПОСТРОЕНИЯ ЭКОСИСТЕМЫ ЗНАНИЙ НА ПРИМЕРЕ ЭНЕРГЕТИКИ[1]

Л.В. Массель (*massel@isem.irk.ru*)
А.Г. Массель (*amassel@isem.irk.ru*)
В.Р. Кузьмин (*kuzmin_vr@isem.irk.ru*)

Институт систем энергетики имени Л.А. Мелентьева СО РАН, Иркутск

В статье рассмотрен подход к построению экосистемы знаний (на примере энергетики), как инновационного подхода к управлению знаниями. Предлагается архитектура цифровой платформы, как основы экосистемы знаний. Используется опыт авторов, полученный при построении ИТ-инфраструктуры системных исследований в энергетике. Методы построения экосистемы знаний определяются архитектурой цифровой платформы, частично апробированы в предыдущих работах авторов. Предлагается использовать имеющиеся авторские научные прототипы инструментальных средств, которые могут стать основой разрабатываемых сервисов.

Ключевые слова: управление знаниями, экосистема знаний, цифровая платформа, энергетика.

Введение

В условиях стремительной цифровизации современного мира информация приобрела характер ключевого стратегического ресурса, а эффективное управление знаниями стало критическим фактором конкурентоспособности и устойчивого развития. Особую актуальность этот тезис получает в сложных, наукоемких отраслях, например, таких, как энергетика, где объемы данных, необходимых для принятия решений, растут экспоненциально.

Управление знаниями является важной составляющей развития организаций и инструментом, позволяющим его сотрудникам справляться с огромными объемами данных по нескольким причинам: 1) в эпоху ин-

[1] Результаты получены в рамках выполнения проекта по госзаданию ИСЭМ СО РАН FWEU-2021-0007 № АААА-А21-121012090007-7.

формационных технологий (ИТ) компании получили доступ к такому объему информации (внутренней и внешней), что выявление актуальной информации, необходимой для принятия решений, требует значительных усилий; 2) из-за постоянных изменений во внешней среде знания довольно быстро теряют свою актуальность, поэтому компаниям необходимо быстро находить и применять новые знания; 3) все больше компаний осознают, что ошибки и незнание проблемы ссылок могут иметь фатальные последствия [Гаврилова и др., 2017].

В настоящее время всё большее распространение, по мере развития цифровой экономики, получает термин «Цифровая экосистема». Под этим термином понимается сеть взаимосвязанных цифровых технологий, услуг и платформ, которые взаимодействуют друг с другом для создания ценности для потребителей и бизнеса.

Основой цифровой экосистемы является цифровая платформа, под которой понимают систему средств, поддерживающую использование цифровых процессов, ресурсов и сервисов значительным количеством субъектов цифровой экосистемы и обеспечивающую возможность их бесшовного взаимодействия[2]. Создание экосистемы знаний на основе цифровой платформы предполагает дополнительную разработку организационных соглашений, регламентирующих отношения разработчиков и пользователей (бенефициаров) экосистемы знаний.

В статье выполнен анализ современного состояния этой области исследований в России и за рубежом, рассмотрены предлагаемая архитектура цифровой платформы экосистемы знаний (на примере энергетики) и методы ее построения.

1. Экосистема знаний

В зарубежной литературе экосистемы знаний рассматриваются с конца 90-х годов [Shrivastava, 1998], подробный разбор понятия и работ по этой тематике был приведён в статье [Robertson, 2020]. В России вопрос экосистем знаний проработан хуже, чем в зарубежных публикациях, интерес к этой тематике стал проявляться относительно недавно, в статьях еще используется различная терминология для данного термина, как, например, «экосистема управления знаниями» [Кулясова и др., 2019], «экосистема знаний» [Масюк и др., 2022] или «знаниевая экосистема» [Абузярова, 2021].

[2] Решение Высшего Евразийского экономического совета от 11.10.2017 № 12 «Об основных направлениях реализации цифровой повестки Евразийского экономического союза до 2025 года».

Концепция «Экосистемы знаний» представляет собой инновационный подход к управлению знаниями, который направлен на развитие взаимодействия между участниками обмена, упрощение процесса принятия решений, а также на стимулирование инноваций путём эволюции сотрудничества между участниками [Shrivastava, 1998].

Экосистемы знаний основаны на активном вовлечении разработчиков и пользователей в совместное создание, изучение и применение общей базы знаний, что приносит пользу всем участникам. Участие в такой экосистеме позволяет субъектам преобразовывать первоначально полученные знания в новые – для коммерциализации продуктов/услуг или для открытия недоступных в одиночку бизнес-моделей и процессов [Järvi et al., 2018].

Поскольку суть экосистем знаний заключается в коллективном обмене знаниями, знания становятся ключевым инструментом взаимодействия между участниками. Результатом на уровне экосистемы обычно являются исследовательские знания и связанные с ними приложения, создаваемые и изучаемые участниками совместно как общий ресурс. Таким образом, экосистемы знаний можно определить как организации, объединяющие разнородных участников вокруг совместного поиска ценных знаний, при этом сохраняющих свою автономную деятельность за пределами этой экосистемы [Scaringella et al., 2018], [Valkokari, 2015].

В отличие от многих других типов экосистем, экосистемы знаний характеризуются ориентацией организаций-участников на исследования, причем эти исследования носят широкий и фундаментальный характер, что позволяет компаниям и другим акторам адаптировать или модифицировать полученные знания под свои конкретные контексты и потребности.

Участников экосистемы знаний можно разделить на две основные категории по их роли относительно центральной базы знаний: создатели (организации и отдельные лица, активно участвующие в обмене, исследовании и формировании общей базы знаний для совместного использования) и пользователи (бенефициары), чья основная цель – использование этой общей базы знаний для последующих инноваций, выхода на рынок или технологического развития. Важно отметить, что эти роли не являются взаимоисключающими: создатели могут становиться пользователями и наоборот. Различение этих категорий не критично, поскольку каждая вносит уникальный вклад в развитие экосистемы [Trischler et al., 2020].

Таким образом, на основе вышесказанного, можно сделать вывод, что разработка методологии построения экосистем знаний, как нового подхода к управлению знаниями, является актуальной. Стоит отметить, что в результате анализа литературных источников не удалось найти сведений о методах и технологиях, используемых при разработке подобных систем, в статьях зарубежных авторов приводятся общие сведения об экосистемах

знаний и результатах их применения, на вопросах реализации и, тем более, конкретной архитектуры авторы не останавливаются, например, [Robertson, 2020], в которой приводятся ссылки на 80 источников.

Авторы предлагают разработку цифровой платформы, как основы построения экосистемы знаний. Таким образом, методы построения экосистемы знаний определяются методами построения основных компонентов цифровой платформы. Ниже рассмотрена предлагаемая авторами архитектура цифровой платформы экосистемы знаний в энергетике.

2. Архитектура цифровой платформы экосистемы знаний в энергетике

Исходя из определения цифровой платформы, приведенного выше, авторами были выделены следующие основные ресурсы, процессы и сервисы для разработки предлагаемой цифровой платформы (рис. 1).

Ресурсы включают в себя знания (а также модели их представления), математические модели и данные. Данные включают, в свою очередь: структурированные данные (описанные моделями данных), слабоструктурированные и неструктурированные, а также потоковые и датасеты.

Процессы включают в себя три основные группы:

1. Процессы с ресурсами – включают такие процессы, как: получение, обработка, хранение и передача данных; эмуляция и импутация данных для цифровых двойников; извлечение данных и знаний, а также генерацию и обработку знаний (в т.ч. анализ и верификацию данных и знаний).
2. Пользовательские процессы – включают в себя регламентацию работы (авторизация, аутентификация и разграничение доступа) и технологию использования, которая опирается на сценарии использования, поддерживает методологию использования и включает нормативные методики.
3. Сервисные процессы – используются при конвертации данных, интеграции сервисов и мониторинге работы сервисов.

Сервисы в рамках цифровой платформы экосистемы знаний определяются интерфейсом (интерфейс пользователя либо интеллектуальный ассистент с применением больших языковых моделей). Предлагаемые сервисы включают в себя три основных группы:

- Хранилище данных (хранит в себе структурированные данные), Data Lake (слабоструктурированные и неструктурированные данные), сервисы математического моделирования, сервис предсказательной аналитики.
- Хранилище знаний, сервисы семантического моделирования, а также базовые компоненты цифровых двойников.
- Сервисы генерации данных и знаний, также включающие в себя сервис работы с большими языковым моделями.

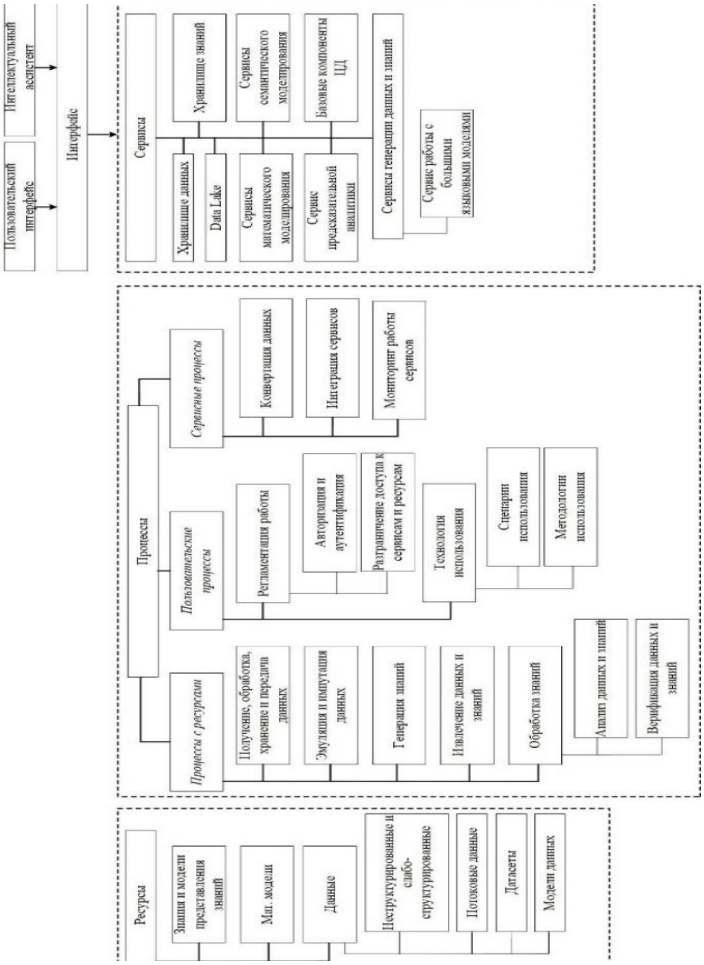


Рис. 1. Архитектура цифровой платформы экосистемы знаний в энергетике

За основу предлагаемой архитектуры цифровой платформы экосистемы знаний (рис. 1) взята разработанная ранее коллективом, который представляют авторы, архитектура ИТ-инфраструктуры системных исследований в энергетике [Массель и др., 2023], [Massel et al., 2024].

Таким образом, отталкиваясь от архитектуры цифровой платформы, основными методами построения экосистемы знаний являются:

- Методы математического и семантического моделирования и предсказательной (предиктивной) аналитики.

- Методы извлечения и представления знаний, методы проектирования хранилища данных и знаний.
- Методы генерации и обработки знаний, в т.ч анализа и верификации данных и знаний.
- Методы построения цифровых двойников энергетических объектов и эмуляции и импутации данных для них.
- Методы интеграции сервисов и мониторинга их работы.
- Методы картирования знаний [Гаврилова и др., 2024].
- Методы разработки интерфейсов, в том числе построения интеллектуальных помощников с использованием больших языковых моделей.

Методы машинного обучения (МО) используются как для прогнозирования погодных характеристик при разработке цифровых двойников возобновляемых источников энергии, так и для проведения расчётов выбросов загрязняющих веществ от объектов энергетики в окружающую среду. При построении цифровых двойников возобновляемых источников энергии для прогнозирования погодных характеристик (освещенность, ветровая обстановка и др.) используется такой метод МО, как LSTM-сеть.

LSTM–сети способны работать с большим количеством данных в сравнении с РНС (рекуррентные нейронные сети) за счет «вентилей», которые понимают долговременное и кратковременное состояние, и вектора состояния, который хранит информацию о состоянии сети на определённом шаге. LSTM–сеть способна выдать точный прогноз, основанный на последней поданной информацией с учетом долгосрочной памяти. В работе используется также такой вид РНС, как управляемые рекуррентные блоки (Gated Recurrent Unit, GRU). Этот вид РНС является «упрощенной» версией LSTM–сети, которая не использует отдельный канал для элемента памяти, а находится в самом векторе состояния. Это позволяет ускорить работу, но, в то же время, теряется точность прогноза [Multi-step wind power forecast, 2019], [Wei et al., 2019].

Для построения модели прогноза погодных характеристик необходимо описать используемые данные (датасет). Данные представляют собой погодные характеристики, которые необходимы для расчета выходных характеристик солнечной электростанции.

Были проведены вычислительные эксперименты с использованием конкретного датасета, содержащего данные о погодных характеристиках в г. Байкальск в размере 88889 записей, сделанных с интервалом в 1 час. Этот датасет содержит столбцы: скорость ветра, температура воздуха, атмосферное давление, прямая солнечная радиация, рассеянная солнечная радиация и суммарная солнечная радиация. Датасет содержит только количественные переменные. Реализация анализа и прогноза выполнена при помощи языка Python с использованием встроенных и сторонних библиотек. Для работы и

построения прогноза применен разведочный анализ данных. Разведочный анализ данных (Exploratory data analysis, EDA) – это процесс исследования и анализа данных с целью выявления закономерностей, паттернов, аномалий и взаимосвязей между переменными [Komorowski et al., 2016]; он включает в себя использование различных методов и инструментов для описательной статистики, визуализации данных и построения графиков, а также применения статистических тестов и моделей для проверки гипотез и извлечения информации из набора данных.

Также разработаны архитектуры отдельных компонентов экосистемы знаний, например, архитектура онтологического портала (приводится на рис. 2).

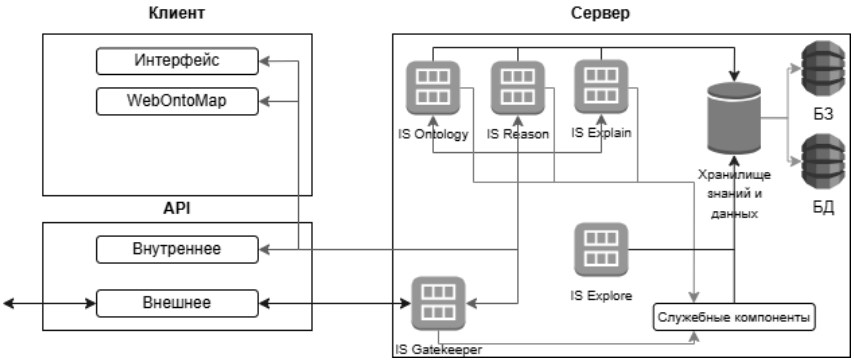


Рис. 2. Архитектура онтологического портала (Онтопортал)

Клиентская часть онтологического портала включает в себя пользовательский интерфейс, а также компонент WebOntoMap для работы с онтологиями. В API выделяется внутреннее (для взаимодействия между компонентами системы) и внешнее (для организации взаимодействия с внешними системами). Серверная часть включает следующие компоненты:

- IS Ontology – подсистема для работы с онтологиями, хранящимися в базе знаний;
- IS Reason – подсистема для организации вывода на онтологиях;
- IS Explain – подсистема для обработки запросов от пользователя на естественном языке и поиска в базе данных и знаний;
- IS Explore – подсистема поиска знаний и данных в открытом доступе для пополнения баз;
- IS Gatekeeper – подсистема для обработки и маршрутизации запросов, отправленных на методы внешнего API;
- база знаний и база данных;
- служебные компоненты.

На рис. 3 приведён пример диаграммы рабочих процессов для роли «бенефициар».

Рис. 3. Диаграмма рабочих процессов для роли «Бенефициар»

Большинство перечисленных методов разработаны и апробированы при разработке ИТ-инфраструктуры системных исследований в энергетике; при разработке экосистемы знаний потребуется их адаптация и/или развитие и доработка.

Кроме того, предлагается использовать в архитектуре цифровой платформы авторские научные прототипы инструментальных средств, которые могут стать основой разрабатываемых сервисов [Массель и др., 2024].

Предлагаемая цифровая платформа экосистемы знаний может быть использована для различных исследований в энергетике, так как платформа обладает гибкой архитектурой, позволяющей подключать дополнительные компоненты в платформу. В частности, цифровая платформа позволяет моделировать четыре уровня энергетической инфраструктуры (агрегат, объект энергетики, энергетическая система и ТЭК страны) и использовать программные средства и технологии для проведения их исследований.

Основным направлением применения цифровой платформы авторы считают интеллектуальную поддержку принятия стратегических решений по развитию энергетики России.

На данный момент основным разработчиком экосистемы знаний является Институт систем энергетики им. Л.А. Мелентьева СО РАН, соорганизаторы – участники разработки: Институт систем информатики СО РАН (Новосибирск), ФИЦ ИУ РАН (Москва), НИИ информационных технологий (Ханты-Мансийск), Институт информационных технологий и анализа данных Иркутского национального исследовательского технического университета. На первом этапе бенефициары, кроме перечисленных разработчиков – представители энергетических организаций и Энергетических институтов ВУЗов России.

Заключение

В статье рассмотрен предлагаемый авторами подход к построению экосистемы знаний (на примере энергетики), как инновационный подход к управлению знаниями. Предлагаемые методы построения экосистемы знаний определяются разработанной коллективом, представляемым авторами, архитектурой цифровой платформы. Рассматриваемые методы разработаны и частично апробированы в предыдущих работах авторов. Предлагается архитектура цифровой платформы, как основа экосистемы знаний, при этом используется опыт авторского коллектива, полученный ранее при построении ИТ-инфраструктуры системных исследований в энергетике. Предлагается использовать имеющиеся авторские научные прототипы инструментальных средств, которые могут стать основой разрабатываемых сервисов. Перечислены организации – участники разработки и возможные бенефициары экосистемы знаний.

Список литературы

[Абузярова, 2021] Абузярова М.И. Знаниевые экосистемы как доминирующий подход формирования новых моделей управления // Экономика, предпринимательство и право. – 2021. – Т. 11, № 12. – С. 2259-2669.

[Гаврилова и др., 2017] Гаврилова Т.А., Алсуфьев А.И., Кокоулина Л.О. Управление знаниями с российским акцентом: победы и поражения // Инновации. – 2017. – № 1(219). –С. 59-69.

[Гаврилова и др., 2024] Гаврилова Т.А., Кузнецова А.В., Алканова О.Н., Гринберг Э.Я. Визуализация компетенций сотрудников с помощью карт знаний // Российский журнал менеджмента. – 2024. – № 22(1). – С. 86-112. – doi.org/10.21638/spbu18.2024.104.

[Кулясова и др., 2019] Кулясова Е.В., Дьяконова М.А. Экосистема управления знаниями в отрасли производства минеральных удобрений // Путеводитель предпринимателя. – 2019. – № 44. – С. 105-116.

[Массель и др., 2023] Массель Л.В., Массель А.Г. Построение экосистемы знаний на основе ИТ-инфраструктуры системных исследований в энергетике // Вестник Югорского университета. – 2023. – № 4. – С. 78-87. – doi: 10.18822/byusu20230478-87].

[Массель и др., 2024] Массель Л.В., Массель А.Г. Инженерия знаний в исследованиях устойчивости энергетических и экологических систем // XXI Национальная конференция по искусственному интеллекту с международным участием, КИИ-2023: Труды конференции в 2-х т. Т. 1. – Смоленск: Принт-экспересс, 2023. – С. 113-122.

[Масюк и др., 2022] Масюк Н.Н., Бушуева М.А., Герасимова А.А. Концепция экосистем в экономике знаний: теоретический базис // Естественно-гуманитарные исследования. – 2022. – № 44(6). – С. 208-212.

[Järvi et al., 2018] Järvi K., Almpanopoulou A., Ritala P. Organization of Knowledge Ecosystems: Prefigurative and Partial Forms // Res. Policy. – 2018. – Vol. 47. – P. 1523-1537.

[Komorowski et al., 2016] Komorowski M., Marshall D.C., Salciccioli J.D., Crutain Y. Exploratory Data Analysis // In: Secondary Analysis of Electronic Health Records. – Springer, Cham, 2016. – P. 185-203. – https://doi.org/10.1007/978-3-319-43742-2_15.

[Massel et al., 2024] Massel L.V., Massel A.G., Pesterev D.V. Knowledge Engineering in Resilience Research of Energy and Ecological System // Pattern Recognition and Image Analysis. – 2024. – Vol. 34, No. 3. – P. 464-469. – https://doi.org/10.1134/S1054661824700214.

[Multi-step wind power forecast, 2019] Han L., Zhang R., Wang X. et.al. Multi-step wind power forecast based on VMD-LSTM // IET renewable power generation. – 2019. – Vol. 13, Iss. 10. – P. 1690-1700. – DOI: 10.1049/iet-rpg.2018.5781.

[Robertson, 2020] Robertson J. Competition in Knowledge Ecosystems: A Theory Elaboration Approach Using a Case Study // Sustainability. – 2020. – Vol. 12(18):7372. – https://doi.org/10.3390/su12187372.

[Scaringella et al., 2018] Scaringella L., Radziwon A. Innovation, Entrepreneurial, Knowledge, and Business Eco-systems: Old Wine in New Bottles? Technol. Forecast. Soc. Chang. 2018. – Vol. 136. – P. 59-87.

[Shrivastava, 1998] Shrivastava P. Knowledge Ecology: Knowledge Ecosystems for Business Education and Training [Электронный ресурс]. – URL: https://web.archive.org/web/20170825081451/http:/www.facstaff.bucknell.edu/shrivast/KnowledgeEcology.html (дата обращения: 20.05.2025).

[Trischler et al., 2020] Trischler J., Johnson M., Kristensson P. A Service Ecosystem Perspective on the Diffusion of Sustainability-Oriented User Innovations // J. Bus. Res. – 2020. – 116. – P. 552-560.

[Valkokari, 2015] Valkokari K. Business, Innovation, and Knowledge Ecosystems: How They Differ and How to Survive and Thrive within Them // Technol. Innov. Manag. Rev. – 2015. – Vol. 5. – P. 17-24.

[Wei et al., 2019] Wei W., Li P. Multi-channel LSTM with different time scales for foreign exchange rate prediction // Proceedings of the International conference on advanced information science and system. – 2019. – No. 28. – P. 1-7. – DOI: 10.1145/3373477.3373693.

УДК 004.8
doi: 10.15622/rcai.2025.006

КЛАССОВАЯ ТИПИЗАЦИЯ УЗЛОВ В МЕТА-АССОЦИАТИВНЫХ ГРАФАХ

А.Е. Мисник (*anton@misnik.by*)

Белорусско-Российский университет,
Республика Беларусь, Могилев

Статья посвящена теоретическому обоснованию классовой типизации узлов в мета-ассоциативных графах – гибридной модели представления знаний, объединяющей графовую структуру и объектно-ориентированные принципы. Описываются формальная структура, иерархические и ассоциативные связи, механизмы наследования и полиморфизма, а также способ хранения схемы знаний внутри самой модели. Показано, как типизация упрощает автоматизацию обработки данных, оптимизирует запросы и повышает адаптивность информационных систем. Отмечена важность унификации хранения и доступа к знаниям, обеспечивающей масштабируемое расширение онтологии без миграции данных. Полученные результаты расширяют теорию метаграфов и служат основой для разработки семантически богатых высокопроизводительных систем управления знаниями.

Ключевые слова: мета-ассоциативные графы, инженерия знаний, онтологии, сложные системы.

Введение

Одним из инструментов инженерии знаний выступают онтологии – формализованные описания понятий предметной области и отношений между ними. Онтологии позволяют создавать общие семантические основы для интеграции данных из разнородных источников, обеспечивая единое понимание информации различными компонентами сложных систем. Такая семантическая унификация необходима, например, для современных кибер-физических систем, где взаимодействуют программные компоненты и физические процессы: единая онтология служит концептуальным интерфейсом, через который различные подсистемы «разговаривают» на одном языке.

Традиционные графовые модели – такие как семантические сети, фреймовые системы, графы знаний на основе RDF/OWL – продемонстрировали возможности их применения, однако, по мере роста сложности моделируемых предметных областей и требований к гибкости знаний, выявляются ограничения этих подходов. RDF-графы оперируют простыми триплетами «субъект–свойство–объект» и не предполагают встроенного механизма для задания процедурных аспектов знаний (таких как поведение объектов или события), что затрудняет моделирование динамики систем [Arpirez et al., 1998]. Объектно-ориентированные базы данных, хотя и близки по духу к онтологиям, не получили широкого распространения в силу недостаточной гибкости стандартных языков запросов и средств онтологического вывода.

Перспективным кажется подход, сочетающий объектно-ориентированное представление (классы, объекты, методы) с графовой природой знаний (узлы и связи между ними). Одной из таких гибридных моделей является мета-ассоциативный граф. Мета-ассоциативные графы были предложены как расширение концепции метаграфов для нужд онтологического инжиниринга. Метаграф в классическом понимании – это обобщение направленного графа и гиперграфа, позволяющее отображать отношения между множествами объектов (например, гиперребро может связывать одновременно несколько вершин) [Basu et al., 2007]. Мета-ассоциативный граф развивает эту идею, вводя единое понятие узла, объединяющее обычную вершину и мета-вершину (вершину, представляющую множество других вершин). Кроме того, каждый узел наделяется внутренней структурой – именем, атрибутами, событиями и методами – что привносит в граф элементы объектно-ориентированного описания.

Актуальность мета-ассоциативных графов обусловлена потребностью в моделях знаний, способных гибко эволюционировать, отражать многоуровневые ассоциации и поддерживать встроенную логику [Bobryakov et al., 2022]. В кибер-физических и других сложных системах нередко возникает ситуация, когда структура знаний должна оперативно перестраиваться при появлении новых типов сущностей или отношений, а инструменты поиска информации должны «понимать» новые взаимосвязи без полной переработки модели. Мета-ассоциативные графы, сочетая графовую гибкость с принципами инкапсуляции и наследования, предлагают решение этой проблемы. Они позволяют описывать сложные системы квази-иерархически: с одной стороны, присутствует иерархия классов и компонентов (дерево или сеть отношений «родитель–потомок»), с другой – допускаются рекурсивные связи и латеральные ассоциации между элементами, выходящие за рамки строгого дерева. Благодаря этому удаётся достичь высокого уровня точности и глубины при моделировании взаимосвязей реального мира.

Мета-ассоциативные графы

Мета-ассоциативный граф – это развитая модель графового представления знаний, в которой каждому узлу приписывается класс (тип), определяющий его структуру и поведение. В традиционных онтологиях (например, OWL) существует разграничение между классами и экземплярами: класс задаёт набор свойств, которыми могут обладать экземпляры. Мета-ассоциативный граф перенимает эту идею и интегрирует её непосредственно в узлы графа, делая классовую типизацию частью самой графовой модели.

Формально узел мета-ассоциативного графа можно представить как кортеж, включающий несколько составляющих: идентификатор, атрибуты, события и методы. Предложено следующее определение узла (N) мета-ассоциативного графа:

$$N = \{ I, \{AS\}, \{EV\}, \{M\} \},$$

где I – имя узла (уникальный идентификатор или метка, однозначно определяющая данный узел); AS – множество ассоциативных атрибутов узла; EV – набор событий, связанных с узлом; M – набор методов узла. Ассоциативные атрибуты (AS) при этом представляют собой совокупность значений и ссылок на другие узлы графа. Формально это можно записать как:

$$AS = \{ V, N \},$$

то есть атрибут узла может содержать либо скалярное значение (число, строку, дату и т.п.), либо ссылку (или несколько ссылок) на другие узлы мета-ассоциативного графа. Такое определение объединяет в одном понятии и традиционные «атрибуты-значения», и «атрибуты-связи» (ассоциации). То есть, граница между данными и связями размывается: атрибут может непосредственно содержать связь на другой объект, превращаясь тем самым в своеобразное ребро, встроенное в узел [Misnik, 2022].

События (EV) – это совокупность определённых ситуаций или изменений состояния, на которые узел способен реагировать. События могут быть разных типов: изменение значения атрибута, появление связи, удаление узла, достижение некоторого условия и т.д. Связав определённые обработчики с такими событиями, система приобретает свойство реактивности – способность запускать определённую логику при наступлении тех или иных условий.

Методы узла (M) – это функциональные возможности, заложенные в объект. Метод можно рассматривать как функцию или процедуру, оперирующую состоянием данного узла (его атрибутами, связями) и, возможно, влияющую на другие узлы. В отличие от чисто декларативных онтологий (OWL, RDF), где знание задаётся только в виде фактов и логических аксиом, наличие методов позволяет включить в модель элементы алгоритмической логики [Gomez-Perez et al., 2004]. Методы могут вызываться

извне (пользовательским кодом или интерфейсом системы) или внутри самой системы – например, как реакция на событие. Наличие методов и событий фактически превращает пассивную модель знаний в активную онтологию, где объекты могут обладать элементами поведения.

Мета-ассоциативный граф можно представить как ориентированный граф, узлы которого являются объектами (в смысле объектно-ориентированного подхода). Дуги (ребра) графа при этом служат для выражения базовых отношений иерархии между узлами – по определению, это ненагруженные (без собственных атрибутов) связи «родитель–потомок». Иерархическими ребрами могут оформляться, в частности, отношения между классами и подклассами, а также между классами и экземплярами, или отношения композиции (часть–целое) между объектами. Важное упрощение мета-ассоциативной модели состоит в том, что все связи унифицируются под общим понятием “иерархия”. Фактически, любая направленная зависимость представляется дугой: более общий объект связывается с более частным или составным объектом как «родитель с потомком». Такой подход охватывает как классические отношения наследования, так и включение/агрегацию. Несмотря на использование единого типа связи, семантическая разница между «классическим» наследованием и композицией сохраняется за счёт смысла узлов: родитель, помеченный как класс, интерпретируется как обобщённое понятие, а родитель, помеченный как сущность-целое, трактуется как агрегирующий объект. Однако формально и то, и другое – просто отношения иерархии в графе.

Следует отметить, что мета-ассоциативные графы предложены именно как развитие теории метаграфов, поэтому важно понимать их связь и отличия. Метаграф – это структура, обобщающая понятие графа, где вершины могут группироваться в метавершины, а рёбра отображают один набор вершин в другой набор. Такой аппарат удобен для моделирования, например, зависимостей «многие-ко-многим» или задач моделирования потоков данных, где каждое действие рассматривается как отображение множества входных документов в множество выходных [Gapanyuk, 2019], [Chernenkiy et al., 2018]. Однако классический метаграф, будучи мощнее обычного графа, всё же остаётся статической конструкцией: в нём нет понятия событий, отсутствует механизм для описания поведения при изменениях структуры. Кроме того, разделение на вершины, мета-вершины и атрибуты в ряде случаев оказывается слишком ригидным, особенно в динамично изменяющихся системах знаний – порой грань между атрибутом и отдельным узлом становится условной. Например, некоторая характеристика объекта (атрибут) может со временем превратиться в самостоятельный объект системы (узел) по мере роста её значимости. В классической модели требовалась бы перестройка: удаление атрибута, введение новой вершины и переопределение связей.

Мета-ассоциативный граф устраняет эти недостатки за счёт унификации понятий вершины и мета-вершины до единого понятия узла и включения в описание узла дополнительных компонент (событий и методов). По сути, любой узел мета-ассоциативного графа может выступать и как отдельный объект с собственными данными, и как контейнер для других узлов. А событие и методы, «встроенные» в узел, предоставляют естественный механизм реагирования на изменения и инкапсуляции поведения. Именно благодаря этому мета-ассоциативные графы ориентированы на использование в онтологическом инжиниринге – они позволяют заложить в модель знаний не только статические связи, но и динамическую логику реакций внешние и внутрисистемные изменения.

Классовая типизация узлов в мета-ассоциативных графах

Классовая типизация узлов напрямую заимствует базовые принципы объектно-ориентированного программирования: инкапсуляцию, наследование и полиморфизм. Узел мета-ассоциативного графа инкапсулирует данные (атрибуты) и связанные с ними методы, образуя автономный объект, самостоятельно управляющий своим состоянием. Наследование проявляется в том, что класс узла может быть организован в иерархию подчинённых классов, перенимающих свойства и поведение базовых классов. Полиморфизм в контексте знаний означает, что разные по типу узлы могут обрабатываться единообразно через интерфейс родительского класса. В онтологии это выражается в возможности задавать запросы или действия на уровне обобщённого класса и применять их к любому подтипу.

Заметим, что класс в мета-ассоциативном графе сам может быть представлен узлом графа (т.е. классы – такие же объекты верхнего уровня, обладающие атрибутами, связями с родительскими классами и т.д.). Это значит, что онтология (схема) системы может быть выражена теми же средствами, что и сами данные. Данный подход соответствует философии метаданных, когда описание данных хранится вместе с данными. Таким образом, мета-ассоциативный граф способен хранить не только факты предметной области, но и свою собственную схему (иерархию классов, описание атрибутов и методов), фактически выступая в роли самодокументируемой модели.

Одно из ключевых достоинств введения классов в графовую модель знаний – это унификация обработки узлов. Имея классовую типизацию, система приобретает способность обрабатывать все узлы одного типа по единым правилам, что упрощает код, повышает надёжность и облегчает сопровождение знаний.

Рассмотрим каким образом классы узлов позволяют унифицировать и автоматизировать работу с данными.

Единый интерфейс и полиморфизм. Когда узлы структурированы по классам, возможно определение общих операций, применимых ко всем узлам данного класса или иерархии классов. Полиморфная унификация повышает гибкость системы: логика работы описывается один раз, а применяется к многим разнотипным объектам.

Повторное использование и модульность. Классовая организация знаний способствует повторному использованию фрагментов онтологии и связанных процедур. Общие свойства (атрибуты) и методы можно вынести в базовые классы. Если несколько категорий объектов имеют нечто общее, нет необходимости дублировать информацию – достаточно сформировать класс-родитель с общими элементами, что соответствует принципу «не повторяйся» (DRY), и считается хорошей практикой проектирования. В инженерии знаний повторное использование проявляется и в том, что целые фрагменты онтологии – классы с поддеревьями – могут служить шаблонами для различных приложений. Такие классы-шаблоны могут быть оформлены как модули или библиотеки знаний и при необходимости подключаться к основному графу.

Автоматизация задач на основе схемы. Когда для узлов определены схемы (наборы атрибутов определённого типа), это открывает возможности для автоматической генерации кода, пользовательских интерфейсов и других вспомогательных средств. Такой подход активно применяется в информационных системах (технологии low-code, no-code), где метаданные о типах сущностей используются для максимального сокращения ручного кодирования. Мета-ассоциативный граф сам хранит метаданные (классы и атрибуты), что даёт возможность встроенным инструментам генерировать интерфейсы и шаблоны обработки.

Унификация хранения и доступа. Классовая типизация подразумевает, что объекты одного класса хранятся и организованы схожим образом. Данный аспект можно использовать для оптимизации хранения данных: например, для каждого класса можно создать специализированную структуру, хранящую все экземпляры этого класса и быстро предоставляющую к ним доступ по основному ключу (имени или ID). В реляционных СУБД подобное разделение по схемам – норма (таблица на класс), но в графовых хранилищах классическая модель RDF не предусматривает физического разбиения данных по типу (все триплеты лежат в общей массе). В мета-ассоциативном графе же мы явно оперируем объектами классов, что позволяет реализовать гибридное хранение: часть данных (например, скалярные атрибуты) хранить в структурированной форме, а ссылки – отдельно как указатели на другие объекты. Такой подход улучшает когерентность данных в памяти: однородные объекты размещаются близко друг к другу, что положительно сказывается на возможностях кэширования и скорости обхода. Кроме того, зная класс объекта, можно

напрямую переходить к операциям, специфичным для этого класса, без многочисленных проверок во время выполнения. По сути, система получает «знание», как обрабатывать объект, глядя на его принадлежность к конкретному классу – далее задействуется соответствующий метод или алгоритм, что аналогично виртуальным методам в объектно-ориентированном программировании, когда вызов процедуры объекта выполняется поразному в зависимости от реального класса объекта.

Управление жизненным циклом через классы. Наличие классов позволяет централизованно контролировать создание, удаление и преобразование объектов. Можно определить методы-конструкторы на уровне класса – процедуры, которые вызываются при создании нового экземпляра. Они могут выполнять дополнительную инициализацию, проверять ограничения или автоматически связывать новый объект с другими компонентами. Аналогично, при удалении объекта класс может предписать каскадное удаление связанных объектов (для отношения часть-целое) или другие изменения в данных. Классы выступают как «супервизоры» экземпляров, обеспечивая стандартные действия по управлению их жизненным циклом, что обеспечивает ещё один уровень унификации: вместо того чтобы каждый раз при добавлении узла выполнять набор однотипных шагов, класс гарантирует выполнение всех необходимых процедур (через конструкторы, события или системные методы).

Перечисленные особенности показывают, что классовая типизация узлов приводит к стандартизации способов обработки объектов. Система знаний с самого начала проектируется структурированной и модульной. В результате, инженеры по знаниям получают более упорядоченную, предсказуемую среду. Добавляя новый класс в систему, они сразу определяют, какие атрибуты он имеет, как взаимодействует с другими классами, какие методы доступны – и все экземпляры этого класса будут вести себя согласно заданному шаблону, что резко контрастирует с неструктурированными подходами (произвольные графы без схемы), где каждый узел может иметь произвольные свойства и заранее неизвестно, как с ним работать. Мета-ассоциативный граф с типизированными узлами вводит «сильную» семантическую типизацию в графы знаний, аналогично тому, как статическая типизация в языках программирования помогает идентифицировать ошибки и упрощать разработку. Классы узлов выполняют роль «каркаса», на котором строится обработка знаний.

Семантические связи и оптимизация запросов

Связи в мета-ассоциативном графе несут определённый смысл (семантику) в контексте предметной области. Благодаря введению классов и строгой организации связей, появляется возможность оптимизировать запросы к графу, опираясь на знание о типах узлов и видах связей между

ними. Мета-ассоциативный граф по умолчанию трактует все рёбра как иерархические (родитель–потомок). Однако семантически эти связи могут представлять различные отношения.

Связь класс–подкласс отражает отношение наследования между классами в онтологии. Такая связь обозначает, что все свойства и методы родительского класса применимы к потомку (стандартное наследование). Семантически это «является разновидностью» (is-a). В графе она помогает быстро идентифицировать, что узел принадлежит более общему класса, и применить к нему соответствующие обобщённые правила. Система может кэшировать для каждого класса список всех подклассов и потомков, чтобы такие запросы выполнялись мгновенно, а не через обход всего графа.

Связь класс–экземпляр - фактически, экземпляр считается потомком своего класса. Это отношение аналогично rdf:type в RDF, но в мета-ассоциативном графе оно выражено просто как ребро от класса к объекту. Семантически – «является экземпляром». Для каждого класса целесообразно автоматически поддерживать индекс или список всех его экземпляров (потомков). Тогда запрос «выбрать все экземпляры класса» может быть выполнен очень быстро, обращением к индексу, вместо обхода всего графа в поисках узлов с определённым атрибутом типа. При добавлении или удалении связей класс–экземпляр индекс обновляется. Это эквивалентно тому, как реляционная база использует индекс по столбцу класса для выборки строк – но тут мы опираемся на семантику графа.

Связь часть–целое (композиция) описывает, что один узел является компонентом другого. В графе это тоже выражается как связь родитель–потомок (целое – родитель, часть – потомок). Семантически – «является частью» (part-of). Зная, что данное отношение иерархическое (а оно является по-умолчанию таковым для мета-ассоциативного графа), можно применять рекурсивные запросы ограниченно. Например, можно заранее вычислять транзитивное замыкание для отношений часть–целое (подобно тому, как в реляционной базе могут храниться материализованные пути в дереве).

Ассоциативные связи (неиерархические) используются когда необходимо отразить отношение, которое не обладает строгой древовидной структурой. В случае мета-ассоциативного графа такую связь можно моделировать через ассоциативный атрибут. Если такие связи часты и критичны, можно также хранить их в индексированной форме: каждому типу ассоциации соответствует таблица индекса. Система, понимая смысл отношения, может выбирать оптимальный способ хранения (в зависимости от преобладающих запросов).

Иерархия задач/процессов – если узлы могут представлять события или процессы, между ними тоже возможны отношения (например, последовательность операций). Их можно отразить либо иерархически (как подзадачи в задаче), либо тоже через ассоциативные поля вида «следую-

щая стадия». Семантика таких связей помогает оптимизировать планирование: но это уже более специфично и выходит за рамки общей модели, поэтому здесь упомянем лишь, что любые предопределённые типы связей можно учитывать при оптимизации.

Когда запрос к базе знаний сформулирован, он обычно оперирует определёнными условиями на свойства узлов и их связи. В мета-ассоциативном графе наличие классов и иерархий позволяет сузить область поиска задействованных данных и тем самым ускорить выполнение запроса. Приведём типичные ситуации.

Поиск по классу. Запрос «найти все объекты класса X с условием ...» – один из самых распространённых в системах знаний. В классической графовой БД без индексации по типу пришлось бы перебрать все узлы или все узлы, помеченные определённым свойством. В случае мета-ассоциативного графа можно сразу взять список его потомков (что можно сделать очень быстро при наличии индекса) и далее отфильтровать по необходимому атрибуту. Если искомый класс имеет подклассы, то по семантике наследования мы знаем, что все они тоже относятся к этому классу, и потому их потомки тоже должны учитываться. Система может автоматически пройти по дереву подклассов (или иметь подготовленный список всех подклассов) и объединить списки их экземпляров.

Использование отношений для ограничения поиска. Предположим, запрос: «найти все объекты класса A, которые связаны с объектом класса B». Даже без явного индекса, сам факт того, что рассматриваются только узлы класса A, уже отсекает все прочие узлы. Очевидно, что в больших графах такой подхлж экономит огромное количество операций.

Предикаты на иерархические связи. Запросы могут требовать учёта иерархии. Запрос сводится к обходу поддерева определенного класса на определённую глубину. Можно оптимизировать такой обход за счёт хранения указателей на родителя у каждого узла (для быстрого подъёма) и/или списка потомков (для быстрого спуска).

Семантические сокращения путей. Будучи осведомлена о типах связей, система может упрощать логический путь запроса. Если бы граф не имел схемы, система могла бы выполнять произвольный поиск путём соединения всех триплетов. Но имея схему запрос может быть трансформирован в две стадии: 1) получить все объекты нужного (по индексу или списку условий); 2) отфильтровать эти объекты по нужному условию. Каждая стадия использует конкретную связь, а не обходит весь граф.

Благодаря описанным возможностям, мета-ассоциативный граф обеспечивает высокую производительность при выполнении типичных запросов инженерии знаний, несмотря на общую сложность модели. Если сравнить с RDF-хранилищем (триплетным), то там любой сложный запрос распадается на множество сочетаний триплетов и требует соединения по

общим узлам [Rasheed et al., 2019]. При большом числе триплетов (миллионы и более) такие соединения могут быть ресурсозатратны, поэтому RDF-движки применяют тяжёлые оптимизации (многоиндексные схемы, сжатие). В случае мета-ассоциативного графа некоторая информация уже находится в самой структуре: класс задаёт область поиска, иерархия задаёт ограничение путей. То есть запросы в мета-ассоциативном графе могут более точно адресовать нужные узлы. Даже смешанные запросы (с условиями на значения и на связи) выигрывают: условие на значение может проверяться только на узлах нужного класса, не затрагивая остальные.

Оценка производительности

Приведём пример количественного бенчмарка. Пусть в графе 100 000 узлов, из них 10 000 узлов класса «Студент». Запрос: «студенты со средним баллом > 80». В триплетной модели нужно просмотреть все узлы с свойством «средний балл», а ещё сначала найти кто является студентом (или хранить предикат rdf:type). Даже при индексе по rdf:type, мы получаем 10 000 кандидатов, потом среди них отфильтровываем по «средний балл»– итого 10 000 проверок. В мета-ассоциативном графе мы обращаемся к узлу класса «Студент», сразу берём 10 000 потомков (или имеем их список) – 10 000 проверок, в целом, кажется, что порядок сложности тот же. Но если запрос сложнее, скажем: «студенты факультета А с «средний балл» > 80». Без семантики пришлось бы связывать студента с факультетом: либо через два-три триплета (студент -> кафедра -> факультет), соединяя результаты. С семантикой: мы берём узел факультета А, собираем всех студентов (потомков по иерархии организационной структуры) – предположим их 500. Затем отфильтровываем по «средний балл» – всего 500 проверок. Это в 20 раз меньше, чем проверять 10 000 узлов, как в ситуации, когда мы бы сначала отобрали всех студентов, а потом проверяли их принадлежность факультету. Экономия растёт с увеличением общего размера графа, если семантические ограничения сильно сужают выборку.

Заключение

Можно констатировать, что мета-ассоциативные графы с классовой типизацией узлов представляют значимый шаг вперёд в эволюции средств инженерии знаний. Они устраняют разрыв между статичными моделями данных и динамическими требованиями приложений, позволяя описывать и немедленно использовать знания в их естественной комплексности. Как и всякая новая парадигма, этот подход требует освоения и доводки, однако первые результаты – как теоретические, так и практические – демонстрируют его жизнеспособность и эффективность. Для достижения описан-

ных преимуществ необходима поддержка индексов и кэшей по ключевым семантическим признакам. Сама по себе графовая модель – это описание, а чтобы ускорить запросы, система хранения должна включать структуры данных: для каждого узла-класса – быстрый доступ к его экземплярам; для каждого узла-объекта – быстрый доступ к компонентам; для каждого именованного ассоциативного атрибута – индекс объектов по значению/ссылке этого атрибута. Создание и поддержание таких индексов требует дополнительных ресурсов, но без этого семантическая информация не будет эффективно использована. В реляционных СУБД администратор сам решает, на какие поля ставить индексы. В мета-ассоциативном графе можно автоматически ставить индексы на все связи, определённые в схеме, либо на те, которые часто используются в запросах (что можно определить динамически, профилировав систему). Такие избыточные хранения (денормализация) оправданы для ускорения, и они контролируемо реализуются на уровне системы.

Список литературы

[Arpirez et al, 1998] Arpirez J., Gomez- Perez A., Lozano A., Pinto S. (ONTO) 2Agent: An ontology – based WWW broker to select Ontologies // Workshop on Applications of ontologies and Problem Solving Methods. ECAI'98.

[Basu et al, 2007] Basu A., Blanning R. Metagraphs and their applications. – Springer, 2007. – 174 p.

[Bobryakov et al., 2022] Bobryakov A.V., Borisov V.V., Misnik A.E. and Prokopenko S.A. Design and Implementation of Information-Analytical and Industrial and Technological Processes in Production Based on Neuro-Fuzzy Petri Nets // 2022 VI International Conference on Information Technologies in Engineering Education (Inforino). – 2022. – P. 1-6. – doi: 10.1109/Inforino53888.2022.9782997.

[Bobryakov et al., 2019] Bobryakov A.V., Borisov V.V., Gavrilov A.I., Tikhonova E.A. Compositional Fuzzy Modeling of Energy- and Resource Saving in Socio-Technical Systems // EAI Endorsed Transactions on Energy Web and Information Technologies. – DOI: 10.4108/eai.12-9-2018.155863.

[Borisov et al., 2021] Borisov V.V., Zakharchenkov K.V., Kutuzov V.V., Misnik A.E. and Prokopenko S.A. Modeling educational processes based on neuro-fuzzy temporal Petri nets // Applied Informatics. – 2021. – Vol. 16, No. 4. – P. 35-47. – DOI: 10.37791 / 2687-0649-2021-16-4-35-47.

[Chernenkiy et al., 2018] Chernenkiy V.M., Gapanyuk Yu.E., Revunkov G.I., Andreev, A.M., Kaganov Yu.T., Dunin I.V. The Principles and the Conceptual Architecture of the Metagraph Storage System / In: Manolopoulos Ya., Stupnikov S. (eds.) // 20th International Conference, DAMDID/RCDL 2018, Revised Selected Papers, CCIS. – Vol. 1003. – P. 73-87.

[Gapanyuk, 2019] Gapanyuk Yu.E. Metagraph Approach to the Information-Analytical Systems Development // In: Proceedings of the 6th International Conference Actual Problems of System and Software Engineering, Moscow, Russia, 2019. – P. 428-439.

[Gomez-Perez et al., 2004] Gomez-Perez A., Fernández-López M., Corcho O. Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web. – 2004.

[Misnik, 2022] Misnik A.E. Ontological Engineering on Metagraphs Basis // 2022 VI International Conference on Information Technologies in Engineering Education (Inforino). – 2022. – P. 1-6. – doi: 10.1109/Inforino53888.2022.9782909.

[Rasheed et al., 2019] Rasheed B., Popov A.Yu. Network graph datastore using DiSc processor // In: Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and ElectronicEngineering, ElConRus. – 2019. – P. 1582-1587.

УДК 007(082)
doi: 10.15622/rcai.2025.007

ТЕОРЕТИКО-КАТЕГОРНЫЙ ПОДХОД К ОБУЧЕНИЮ НА ОСНОВЕ ОБОБЩЕНИЯ

В.Л. Стефанюк (*stefanuk@iitp.ru*)
А.В. Жожикашвили (*zhozhik@iitp.ru*)

Институт проблем передачи информации РАН, Москва

В работе описывается новый подход к машинному обучению – обучение на основе обобщения. В работе излагается математический аппарат, основанный на языке теории категорий, который может быть использован в качестве математического формализма для такого обучения.

Ключевые слова: обучение, обобщение, образец, теория категорий.

Введение

В настоящий момент в Искусственном интеллекте происходит бум, связанный с нейронными сетями. По мнению авторов настоящего доклада он связан не столько с самими нейронными сетями, сколько с тем, что утвердилась идея, состоящая в том, что компьютер может приобрести способность решать интеллектуальные задачи в результате обучения. Сама идея обучающегося компьютера достаточно стара, однако в последние годы многократно возрос доступный объем данных, которые можно использовать для обучения. Это позволило обучающимся системам, в частности, нейросетевым, успешно решать прикладные задачи.

Но нейронные сети являются лишь простым и удобным инструментом для обучения, не единственным, а возможно – и не лучшим. Авторы настоящего доклада мнОго лет занимались продукционными интеллектуальными системами, в частности – экспертными системами. Такие системы имеют целый ряд привлекательных моментов. Упомянем, к примеру, их способность объяснить свои решения, указав правила, на основе которых эти решения были приняты. Нейронные сети таким свойством не обладают, человек не в состоянии понять, каким образом они получили результат. Это усложняет их применение в областях с высокой ценой ошибки, скажем – в медицине.

Однако возможность обучаться, свойственная нейронным сетям, перевешивает все эти недостатки. В классических продукционных сетях обучение не используется. Система продукций закладывается человеком-экспертом. Иногда этим даже занимается особый специалист – инженер по знаниям. Понимая, что этот подход тормозит использование продукционных систем, авторы доклада постоянно стремились создать механизмы автоматического формирования и развития базы знаний в процессе обучения, анализа правил, содержащихся в базе знаний, их модификации и создания новых правил.

К числу алгоритмов обучения, не связанных с нейронными сетями, которые можно использовать и в продукционных системах, мы относим алгоритм обучения на основе обобщения, многие годы развиваемый в нашем коллективе. В свое время авторами доклада была предложена математическая формализация ряда механизмов, связанных с представлением знаний [Стефанюк, 1999], основанная на аппарате теории категорий [Маклейн, 2004]. К их числу относится и механизм обучения на основе обобщения. Такая формализация позволяет формулировать алгоритмы обучения, не связанные буквально с тем, как устроены знания, которые система должна приобрести в результате этого обучения. Предложенная авторами теоретико-категорная технология создания интеллектуальных систем, основанных на знании [Zhozhikashvili, 2023] позволяет унифицировать и программирование таких систем.

1. Продукции, образцы, сопоставлнение и конкретизация

Авторы докладов много лет изучали интеллектуальные компьютерные системы, основанные на правилах, или продукциях – продукционные системы. Именно такие системы были формализованы на теоретико-категорном языке. Все это было подробно описано в наших многочисленных статьях на эту тему. Приведем здесь возможно более короткое изложение.

Продукционная система, решая задачу, совершает целый ряд последовательных шагов. На каждом шаге она находит применимое в этот момент правило и применяет его. Затем она применяет следующее применимое правило и так далее, пока не достигнет поставленной цели. Применение каждой продукции увеличивает количество информации, известной системе. Совокупность информации, известной в некоторый момент, мы называем *ситуацией*. Таким образом, применение продукции меняет текущую ситуацию на новую, добавляя к ней информацию. Задача продукциогной системы – получить ситуацию, являющуюся решением системы. Такая ситуация называется *целевой*. Поскольку под ситуацией мы понимаем информацию, известную системе, под целевой ситуацией разумно понимать информацию, содержащую, среди прочего, то, что система должна определить в результате решения задачи. *Продукцией* мы называ-

ем описание двух ситуаций – ситуации, в которой продукция применима, и ситуации, которая возникает после ее применения. Эти описания ситуаций, составляющие правило, не являются абсолютно детализированными, некоторые моменты, не существенные для применимости данного правила, в них опущены. Такое описание, в котором опущен ряд деталей, мы называем образцом ситуации, или просто *образцом*. Если добавить к образцу эти опущенные детали, образец превратится в ситуацию. Такую операцию добавления опущенных деталей мы называем *конкретизацией* образца, а сами добавляемые детали – *конкретизатором*.

Пусть S – множество всех ситуаций, которые могут возникнуть при работе системы, p – некоторый образец, C – множество всех возможных конкретизаторов, которые могут быть использованы для конкретизации образца p. Выбирая любой конкретизатор $c \in C$ и конкретизируя посредством этого конкретизатора образец p, мы получаем некоторую ситуацию $s \in S$. Таким образом, образец p можно рассматривать как отображение $p: C \longrightarrow S$. Теперь можно уточнить описание продукции, именно, продукцией мы будем называть пару образцов с совпадающим множеством конкретизаторов, т.е. пару отображений $p: C \longrightarrow S$, $q: C \longrightarrow S$. Если $s \in S$ – ситуация, то применении такой продукции к ситуации s дает ситуацию $t \in S$ тогда и только тогда, когда $\exists c \in C: p(c) = s, q(c) = t$. Отметим, что результат применения продукции в общем случае определен неоднозначно, так как выбор элемента $c \in C$ по условию $p(c) = s$ может не быть однозначным.

В сложных задачах в качестве множеств ситуаций и множеств конкретизаторов могут использоваться различные множества, причем одно и тоже множество может в одном случае выступать как множество ситуаций, в другом – как множество конкретизаторов, поэтому, если мы хотим использовать приведенный выше формальный подход в образцам, мы должны указать, какие множества могут быть множествами ситуаций/конкретизаторов и какие отображения между ними могут выступать в качестве образцов. При некоторых естественных условиях замкнутости это определяет подкатегорию категории множеств. Это соображение навело нас на мысль перевести всю теорию образцов, сопоставления и конкретизации на язык теории категорий. Роль образцов, в этом случае будут играть морфизмы категории. В этом случае пришлось отказаться от представления о том, что ситуация – элемент множества S, ибо объекты категории не обязаны быть множествами. Вместо этого ситуация понимается как частный случай образца. Это соответствует смыслу этих понятий: категория – это полностью конкретизированный образец, в котором не осталось опущенных деталей. Какие именно образцы считаются ситуациями – определяется спецификой конкретной задачи.

Все это приводит к следующему определению.

Системой образцов называется категория[1] $\mathcal{C}$, в которой для любой пары объектов X и Y и множества морфизмов $\mathcal{C}(X,Y)$ определено множество $\mathcal{C}_S(X,Y) \subset \mathcal{C}(X,Y)$. Точное определение, приведенное в наших более ранних работах, содержит дополнительные требования к множествам $\mathcal{C}_S(X,Y)$, которые мы здесь не приводим, чтобы не усложнять изложение. Образцом с областью конкретизаторов X и областью значений Y считаем любой морфизм $\varphi \in \mathcal{C}(X,Y)$, ситуацией – морфизм $\alpha \in \mathcal{C}_S(X,Y)$. (Такой образец будем в дальнейшем называть Y-образцом, а ситуацию – Y-ситуацией.) Ситуация, таким образом, представляет собой частный случай образца. Ситуация $\alpha \in \mathcal{C}_S(A,S)$ сопоставима с образцом $\varphi \in \mathcal{C}(X,S)$, тогда и только тогда, когда существует морфизм $\beta \in \mathcal{C}(A,X)$ такой, что $\alpha = \beta\varphi$. Пусть S, T – объекты категории $\mathcal{C}$. Продукцией из S в T называется пара образцов (φ,ψ), $\varphi \in \mathcal{C}(X,S)$, $\psi \in \mathcal{C}(X,T)$, X – объект категории $\mathcal{C}$. Будем говорить, что продукция применима к ситуации α, если эта ситуация сопоставима с образцом φ, т.е. если $\alpha = \beta\varphi$ для некоторого морфизма $\beta \in \mathcal{C}(A,X)$. В этом случае результатом применения продукции к ситуации α является ситуация $\beta\psi$.

2. Порядок на множестве образцов

Образец можно рассматривать как способ единообразно описать множество ситуаций. Именно, образец может служить для описания тех ситуаций, которые сопоставимы с этим образцом. Определение сопоставимости ситуации с образцом легко обобщается на определение сопоставимости образца с образцом: говорят, образец $\psi \in \mathcal{C}(Y,S)$ сопоставим с образцом $\varphi \in \mathcal{C}(X,S)$, если когда существует морфизм $\xi \in \mathcal{C}(,X)$ такой, что $\psi = \xi\varphi$. В этом случае будем также говорить, что образец ψ является частным случаем образца φ. Логичность такой терминологии следует из легко доказуемого факта: если образец ψ сопоставим с образцом φ, то всякая ситуация, сопоставимая с образцом ψ, сопоставима также и с образцом φ, т.е. образец φ описывает более широкое множество ситуаций.

Пользуясь понятием сопоставимости, можно для объекта S категории $\mathcal{C}$ ввести частичный порядок [Биркгоф, 1984] на множестве S-образцов, полагая, что для S-образцов ψ и φ условие $\psi \leq \varphi$ означает, что образец ψ является частным случаем образца φ.[2] Как будет понятно из дальнейшего,

[1] Будем считать, что все рассматриваемые категории являются малыми, чтобы избежать сложостей с теоретико-множественной аксиоматикой.

[2] Строго говоря, такое определение дает не порядок, а предпорядок. Для перехода к порядку надо провести некоторую факторизацию, как это всегда делается. Здесь мы не будем углубляться в эти тонкости, ибо дальнейший материал не содержит стррогих математических доказательств и является достаточно понятным и без таких уточнений.

нас будет интересовать случай, кода получившееся упорядоченное множество является решеткой [Биркгоф, 1984]. Это бывает, естественно, не всегда, но во многих практически важных случаях это так.

Пусть φ, ψ – S-образцы, $\varphi \in C(X,S)$, $\psi \in C(Y,S)$. Будем называть образец $\sigma: Z \longrightarrow S$ обобщением образцов φ и ψ, если существуют морфизмы λ и μ такие, что $\lambda\sigma = \varphi$, $\mu\sigma = \psi$. Будем записывать такую ситуацию как обобщение $(Z, \sigma, \lambda, \mu)$. Очевидно, что в этом случае $\varphi \leq \sigma$, $\psi \leq \sigma$. Будем называть обобщение $(Z, \sigma, \lambda, \mu)$ наименьшим обобщением образцов φ и ψ, если для любого другого обобщения $(Z', \sigma', \lambda', \mu')$ существует морфизм $\xi: Z \longrightarrow Z'$ такой, что $\sigma'\xi = \sigma$. Наименьшее обобщение образцов φ и ψ соответствует верхней грани $\varphi \vee \psi$ в теории решеток.

3. Обучение на основе обобщения

Изложив некоторые детали необходимого нам математического аппарата мы переходим к изложению методов обучения, основанных на обобщении.

В общих чертах обучающаяся система работает следующим образом. Система должна научиться решать задачи определенного типа. Для этого используется процедура обучения с учителем. Учитель предлагает системе задачи, решения которых ему известны. Система пытается их решить. Если система не может решить задачу – учитель предоставляет ей решение, которое она использует для обучения, изменяя свою базу знаний. Говоря на языке продукционных систем решение, предоставляемое учителем – это просто описание целевой ситуации, которую система получила бы, если бы обладала достаточныеми знаниями. Если система решает задачу, предложенное ей решение проверяется учителем. В случае, если решение – правильное, никаких действий не предпринимается. В случае, если решение неверное, учитель информирует систему об этом, и система использует эту информацию для обучения. Как именно использовать такую «отрицательную» информацию – вопрос весьма непростой. Частично эта тема раскрывается в разделе 5, но на настоящий момент она проработана значительно слабее, чем использование «положительной» информации.

Встретив задачу, которую она не может решить, и получив от учителя решение, система первым делом создает и записывает в базу знаний правило, левой частью которого является условие задачи, правой – полученное от учителя решение. В дальнейшем она может использовать это правило только в одном случае: если на вход поступит точно такая же задача. Тогда она выдаст такое же решение. Ни для чего другого это правило не годится. Однако система сравнивает каждое новое правило с уже существующими. Если вновь созданное правило оказывается близким к одному из созданных ранее и эти два правила имеют много общего, система строит обобщение двух правил, помещает это обобщение в базу знаний, а исходные правила из базы удаляет.

Используем описанный выше формализм чтобы конкретизировать процесс обучения.

Проще всего изложить алгоритм обучения для задачи распознавания. Эта задача состоит в следующем. На вход системы распознавания поступают объекты определенного вида. Система должна научиться распознавать объекты, относящиеся к определенному классу, т.е. задача распознавания может рассматриваться также как задача классификации.

В нашей терминологии поступающие на вход объекты будут рассматриваться как ситуации и система должна определить, относится ли поступившая на вход ситуация к некоторому классу. Для этого формируется образец, описывающий ситуации этого класса. Система производит сопоставление поступившей ситуации с этим образцом. Если сопоставление прошло успешно, считается, что ситуация относится к нужному классу, если нет – не относится. В сложных случаях одного образца может не хватить, система формирует базу образцов и решение о том, что поступившая ситуация относится к классу, принимается в том случае, если ситуация сопоставима с одним из этих образцов из этой базы.

База образцов, используемая для распознавания, строится следующим образом. Если на вход поступает ситуация α, относящаяся к исследуемому классу. Эта ситуация добавляется к базе как образец (выше говорилось, что ситуация – частный случай образца). Затем эта ситуация сравнивается со всеми образцами, содержащимися в базе. Если для ситуации α и некоторого образца φ возможно построить нетривиальное наименьшее обобщение $\alpha \vee \varphi$, то из базы удаляются и ситуация α, и образец φ, вместо них в базу добавляется образец $\alpha \vee \varphi$.

Тут возникает очень много вопросов.

Что такое нетривиальное наименьшее обобщение? Если упорядоченное множество образцов является решеткой, то в нем любая пара элементов имеет точную верхнюю грань. Следовательно, любая пара образцов имеет наименьшее обобщение. Какое из них считать считать нетривиальным? Этот вопрос может быть решен только в результате экспериментов. (Отметим, что в современной теории обучения многие решения принимаются в результате экспериментов. Иногда эти решения кажутся странными, но эксперименты показывают, что они работают.) В некоторых наших моделях обобщение считалось нетривиальным, если оно не совпадало с максимальным элементом решетки. В других случаях надо было вводить некоторые ограничения, требующие, чтобы обобщение было бы не слишком общим в каком-то смысле.

Более серьезный вопрос: откуда следует, что если ситуация α относится к некоторому классу и всякая ситуация, сопоставимая с образцом φ, относится к этому классу, то и всякая ситуация, сопоставимая с образцом $\alpha \vee \varphi$, тоже относится к этому классу? Ни откуда. Это утверждение нельзя

доказать, в общем случае оно неверно. Это – всего лишь разумная гипотеза, состоящая в том, что если ситуация α относится к некоторому классу, ситуации, описываемые образцом φ, относится к этому классу, то условием принадлежности к классу может быть то общее, что содержится в описании ситуации α и образца φ, а именно это общее содержит образец $\alpha \vee \varphi$. Но всякая гипотеза может оказаться ложной. В этом случае база должна быть подправлена, об этом говорится в следующем разделе.

Система распознавания проще, чем продукционная система. В такой системе правило состоит из одной лишь левой части. При сопоставлении решается вопрос применимо – не применимо правило, а вопрос о том, что делать, если применимо, не рассматривается. Вопрос о том, как создавать правила в процессе обучения, в первую очередь – как определить операцию обобщения двух правил.

Пусть в результате обработки информации, поступившей от учителя, было создано правило (φ, ψ), где φ, ψ – образцы. Допустим, в базе знаний уже существует правило (λ, μ). Строим образец ξ – наименьшее обобщение образцов φ и λ, образец η – наименьшее обобщение образцов ψ и μ, и заменяем пару правил (φ, ψ) и (λ, μ) на правило (ξ, η).

Авторы отдают себе отчет, что столь краткое изложение не может дать сколько-нибудь полное изложение алгоритмов обучения. Но описать алгоритм детальнее пока не представляется возможным. Во-первых, мы находимся на стадии разработки основных идей обучения на основе обобщения. Очень многие вопросы пока не решены, многие детали не проработаны. Во-вторых, даже то, что проработано – слишком велико по объему для небольшого доклада.

4. Неоправданно широкие обобщения

Обучение на основе обобщения основано на следующей идее. Пусть p и q – описания двух множеств ситуаций (например, образцы). Пусть r – обобщение этих описаний, т.е. описание, включающее в себя то общее, что есть и в p, и в q. Допустим, мы знаем, что как ситуации, подходящие под описание p, так и ситуации, подходящие под описание q, удовлетворяют некоторому условию. Логично предположить, что выполнение этого условия является следствием той общей части, которая содержится в обоих описаниях. Из этого, вроде бы, следует, что ситуации, подходящее под описание r, тоже удовлетворяют условию. Несмотря на некоторую логичность, высказанное утверждение отнюдь не всегда является верным. Да, под описание r подходят все ситуации, подходящие под описание p, и все ситуации, подходящие под описание q, но под описание r могут походить и многие другие ситуации, не подходящие ни под p, ни под q, и эти ситуации могут условию не удовлетворять. В этом случае мы говорим, что r – неоправданно широкое обобщение p и q.

Совершая такие обобщения, обучающаяся система может прийти к неверным выводам, и нам не известен способ, позволяющий этого избежать. То же самое происходит при обучении человека: увидев, что нечто происходит часто, он может сделать вывод, что это происходит всегда, и пребывать в этом заблуждении то тех пор, пока не попадет в ситуацию, когда этого не произошло, или пока кто-то более образованный его не поправит.

Таким образом, в процессе обучения система формирует ни безусловные знания, а гипотезы. Эти гипотезы надо еще проверять каким-либо способом. Отметим, что умение формулировать разумные гипотезы – важная способность интеллекта, в том числе и искусственного интеллекта.

Борьба с неоправданно широкими обобщениями – возможно, более сложная задача, чем само построение обобщений. Для того, чтобы уменьшить возможность появления таких обобщений и тем самым приблизить формируемые при обучении гипотезы к реальным знаниям используются два приема: первый – возможное уменьшение степени обобщения, второй – доработка базы знаний в процессе обучения. Об уменьшении степени обобщения мы напишем ниже, доработке базы посвящен следующий раздел.

Первым моментом, связанным с минимизацией степени обобщения, является использование наименьшего обобщения в терминах теории категорий или верхней грани – в терминах теории решеток. Но этого не достаточно. Полная решетка всегда имеет наибольший элемент. Это означает, что даже если два образца не имеют ничего общего, по ним все равно можно построить верхнюю грань – она будет совпадать с наибольшим элементом решетки и с ней будет сопоставима любая ситуация.

Чтобы этого избежать, надо несколько уточнить процедуру, содержащуюся в последнем абзаце предыдущего раздела. При описанном там построении наименьшего обобщения образцов φ и λ и наименьшего обобщения образцов ψ и μ, проверить, не является ли это обобщение слишком общим, и, если является – отказаться от обобщения. Что значит – слишком общим? Можно просто считать слишком общим обобщение, совпадающее с наибольшим элементом решетки. Но возможно, следует выработать более тонкий критерий. На сегодняшний момент готовых решений у авторов доклада нет.

5. Доработка базы знаний в процессе обучения

Если предыдущие разделы содержат научные результаты, то этот посвящен нерешенным пока вопросов, и их больше, чем вопросов решенных.

Одна из главных проблем – проблема неоправданно широких обобщений. Один из способов борьбы с ними – удаление из базы знаний правил, возникших в результате таких обобщений. Если предложенное системой решение будет отвергнуто учителем – правило, на основании которого было получено это решение, должно быть удалено.

Тут возникает целый ряд проблем. Система приходит к заключению, используя цепочку правил. Как понять, какое из правил цепочки, приведшей к неверному заключению, является ошибочным? Есть целый ряд соображений, посвященных этому вопросу, но ни одно из них не может считаться решением проблемы.

Вторая проблема состоит в том, что при создании по двум правилам их обобщения это обобщение замещает исходные правила, которые удаляются из базы. Это означает, что если обобщение будет впоследствии удалено из базы – там не останется ничего, ни исходных правил, которые были обобщены, ни результата. Как этого избежать? Один из способов – хранить вместе с каждым обобщением ссылку на те правила, из которых оно получено. Это позволит при удалении обобщения восстановить те правила, из которых оно было получено. Другой способ – ранжировать все правила по полезности и использовать сперва более полезные, и только в случае, когда они не подошли – менее полезные. Самый простой способ – расположить все правила базы знаний в виде последовательности и при поиске применимого правила перебирать правила в порядке, в котором они расположены в этой последовательности. Обобщение должно размещаться в этой последовательности раньше, чем правила, из которых оно было получено, так что после включения в базу знаний обобщения эти правила не буду использоваться, однако после удаления обобщения они снова начнут работать. Возможно, следует строить базу знаний не как последовательность правил, а как дерево, в котором правила расположены в вершинах и упорядочены по степени общности. Это значительно ускорит и выбор правила, применимого в некоторой ситуации.

Обучающаяся система должна постоянно анализировать базу знаний и, возможно, вносить в нее изменения. Рассмотрим, к примеру, правило, полученное обобщением двух других. Исходные правила при этом удаляется из базы или передвигаются в более далекие позиции в последовательности или в дереве правил. Может оказаться, что созданное обобщение является более обобщим, чем некоторые другие правила, не только те два, из которых обобщение было получено. В этом случае и эти правила следует переместить подальше или удалить, сохранив возможность восстановления.

Мы перечислили далеко не все нерешенные проблемы. Их решение – предмет дальнейшего исследования.

Заключение

Обучение, основанное на обобщении – перспективный подход к обучению, который может быть использован при создании интеллектуальных компьютерных систем. Он может быть применен в традиционных для Искусственного интеллекта системах, таких, например, как экспертные

системы. Это позволяет объединить его с разработанными ранее методами построения интеллектуальных систем, включив в работу этих систем и обучение, и прямую передачу знаний от эксперта. Знания, полученные в результате обучения, представляются в форме, понятной человеку. Обучение, основанное на обобщении, является достаточно естественным, оно повторяет приемы обучения, свойственные человеку.

В настоящий момент у авторов нет сколько-нибудь завершенной системы, обучающейся на основе обобщения. Предложенные методы требуют апробации, которая и покажет, насколько они эффективны. Отсутствие законченной системы не позволяет провести сравнение такого алгоритма обучения с другими известными, прежде всего – с обучающимися нейронными сетями. Авторы могут высказывать лишь предположения, которые нуждаются в экспериментальной проверке.

Одно из таких важных на наш взгляд предположений мы выскажем прямо здесь. Знания, которые создаются при обучении на основе обобщения, много понятнее для человека, чем знания, создающиеся в результате обучения нейронной сети. Результатом обучения нейронной сети является набор весов. Понять его смысл про хото сколько-нибудь значительном размере сети для человека невозможно. Результатом обучения продукционной системы является набор правил, хорошо понятных для человека. Хотя в реальных задачах, когда количество правил исчисляется тысячами, разобратьс в них тоже непросто, это все-таки много проще, чем оценивать веса, связанные с ребрами сети. Получив некоторый результат, продукционная система может объяснить свое решение, приведя те правила, на основании которых она пришла к этому результату. Ничего подобного нейронная сеть сделать не может.

Есть надежда, что такое обучение на основе обобщения будет идти быстрее, чем обучение нейронных сетей, и не потребует столь большого обучающего материала.

Отметим один существенный момент. Если нейронная сеть оперирует вещественными числами и обучении состоит в подборе этих чисел, в основе продукционной системы лежит бинарная логика да/нет. Достоинство это, или недостаток – покажут дальнейшие исследования.

Список литературы

[Стефанюк, 1999] Стефанюк В.Л., Жожикашвили А.В. Теоретико-категорные образцы для задач искусственного интеллекта, новые результаты // Известия РАН. Теория и системы управления. – 1999. – № 5. – С. 5-16.

[Маклейн, 2004] Маклейн С. Категории для работающего математика. – М.: Физматлит, 2004.

[Zhozhikashvili, 2023] Zhozhikashvili A.V., Stefanuk V.L. Category Technology to Design Intelligent Systems for Complicated Decisions, // Lecture Notes in Networks and Systems book series (LNNS,volume 566), Springer Nature.

[**Биркгоф, 1984**] Биркгоф Г. Теория решеток. – М.: Наука, 1984.

УДК 004.81:004.588
doi: 10.15622/rcai.2025.008

МОДЕЛЬ УЧЕБНОЙ СИТУАЦИИ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО ПЛАНИРОВЩИКА ОБУЧАЮЩЕЙ СИСТЕМЫ

В.А. Углев (*uglev-v@yandex.ru*) [A,B]

[A] Отделение интеллектуальных систем в гуманитарной сфере Российского государственного гуманитарного университета, Москва

[B] Кафедра прикладной физики и космических технологий Сибирского федерального университета, Железногорск

В работе описывается модель учебной ситуации, обрабатывающаяся планировщиком интеллектуальной автоматизированной обучающей системы (ИАОС). Обсуждаются вопросы её представления в процессе работы обучающей системы и последующей обработки. На примере учебного процесса в экспериментальной ИАОС, показан подход к обработке ученой ситуации, используя встроенные экспертные системы и средства картирования. Показан механизм того, как взаимодействие различных моделей по-разному интерпретирует учебную ситуацию и влияет на дальнейшее принятие решений.

Ключевые слова: инженерия знаний, электронное обучение, интеллектуальная автоматизированная обучающая система, учебная ситуация, когнитивная карта диагностики знаний.

Введение

Электронное обучение, как инструмент, выполняет не только функцию выдачи (трансляции) дидактического материала и проверки знаний, но и управления учебным процессом [Беспалько, 1970]. Это заставляет разработчиков современных интеллектуальных автоматизированных обучающих систем (ИАОС, Intelligent Tutoring Systems) включать в свой состав не только подсистему интеллектуального решателя (планировщика) с базой знаний, но и модели различных сущностей. От качества составления базы знаний инженером по знаниям и адекватности построения этих самых моделей и зависит во многом результативность применения ИАОС [Рыбина, 2023]. А так как обучение в электронной среде носит индиви-

дуализированный характер, то и модели должны гибко адаптироваться к особенностям обучаемого и обстоятельствам учебной ситуации. О составлении и использовании интеллектуальным решателем ИАОС одной из таких моделей и пойдет речь в данной статье.

В состав ИАОС традиционно входят модели обучающегося (модель ученика), учебного материала (модель методиста) [Karpenko, 2011]. Примером может быть машина Скиннера [Skinner, 1986]. В развитых системах отдельно выделяются модели учебного воздействия (модель учителя) и модель балансировки целей (модель тьютора). Таким образом, модели ученика и учителя имеют возможность оценивать сложившуюся учебную ситуацию с различных точек зрения, а модель тьютора – искать компромиссное решение [Uglev, 2024b]. Но для того, чтобы все эти модели сущностей корректно работали, требуется проработанная модель учебной ситуации.

Учебную ситуацию в ИАОС анализируют либо через уже имеющиеся модели (не выделяют как отдельную сущность), либо формализуют. Например, в моделях типа Cognitive Tutors [Anderson et al., 1995], [Koedinger et al., 2007], [Lieder, 2019] реализованы развитые методы анализа учебной ситуации, хотя они жестко привязаны к предметному материалу. Ту же проблему имеют ИАОС на базе онтологий предметных областей (см, например, [Сычёв и др., 2025]). Использование схематики (ментальных моделей и когнитивных стратегий) в модели 4C/ID [Van Merriënboer et al, 2002], [Frerejean et al., 2019] также предполагает детальную проработку вспомогательной информации (Supportive) и когнитивных правил (JIT) относительно изучаемого дидактического материала. Если исходить из того, что интеллектуальный планировщик ИАОС по возможности должен оперировать с любым учебным курсом, то формализация и методическая проработка модели учебной ситуации без привязки к изучаемой предметной области представляется актуальной.

1. Модельное обеспечение ИАОС и учебная ситуация

Основой для анализа учебной ситуации является модель дидактического материала, относительно содержания которого обучающийся-человек производит какие-либо действия. При этом в электронном курсе выделяется как структурная часть (иерархия дидактических единиц), так и семантика взаимозависимостей её компонентов, а также целевые параметры. Даже с учётом процесса индивидуализации, в модели курса формируется траектория изучения материала, устанавливаются нормы контроля и прочие параметры (денотат). По сути, это уже объект для полноценной обработки в контексте прикладной семиотики [Поспелов и др., 1999].

Из протокола событий в обучающей среде формируется цифровой образовательный след, а прочие данные об обучающемся формируются в результате диалогового взаимодействия (заполнения анкет, ответов на предметные или методические вопросы по ходу процесса обучения).

Оценка обстановки, позволяющей принять решения, была приведена в модели П.К. Анохина (процесс афферентного синтеза [Анохин, 1975]). Её адаптированный для ИАОС вариант (текст в скобках) приведен на рис. 1. Модель учебной ситуации, в отличии от моделей сущностей, актуализируется в рабочей памяти АИОС после конкретного события, и она обязана включает в себя как параметры, связанные с электронным курсом, так и параметры, связанные с моделью обучающегося. Тактом образом, там должна отразиться как семиотическая структура вовлечённых в анализ сущностей, так и временная декомпозиция, позволяющая реализовать афферентный синтез со стороны интеллектуального планировщика.

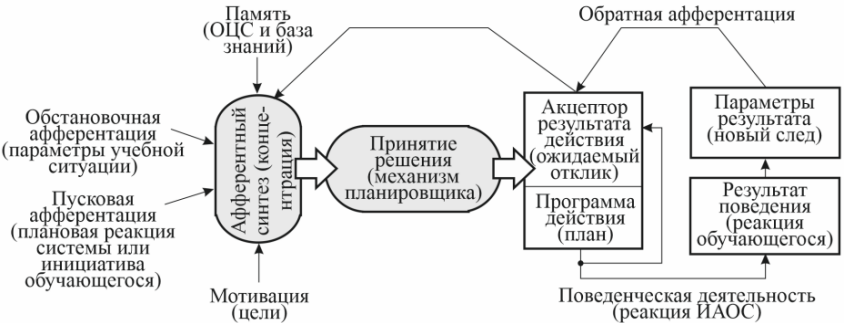


Рис. 1. Модель афферентного синтеза с адаптацией для ИАОС

Сведем всё это воедино, составив табл. 1. Из неё видно, что одни и те же компоненты источников данных имеют зависимости многие-ко-многим как в отношении модели поведенческого акта Анохина, так и элементов квадрата Поспелова. Таким образом появляется возможность интерпретировать учебный процесс в ИАОС сразу с точки зрения двух моделей.

Таблица 1

Объекты, сущности и данные, которые включены в ИАОС	Компоненты афферентного синтеза	Компонент квадрата Поспелова
1. Трек перемещений по дидактическому материалу	Обстановочная афферентация	Денотат
2. Ответы на контрольно-измерительные материалы	Обстановочная афферентация	Денотат
3. Характеристики процесса решения заданий	Обстановочная афферентация	Денотат
4. Данные анкетирования (предпочтения, целеполагание)	Мотивационное возбуждение	Прагматика и денотат

Окончание табл. 1

5. Траектория и характер диалога	Обстановочная афферентация	Денотат
6. Событие в ИАОС	Пусковая афферентация	Денотат
7. Иерархия дидактического материала	Память	Синтактика
8. Связи внутри учебного материала, нормативами, целями обучения, целями обучающегося	Память и Акцептор результата действий	Семантика
9. База знаний ИАОС	Память, Механизмы принятия решений	Семантика
10. Алгоритмы педагогического воздействия	Программа действий и Память	Прагматика

Как следует из табл. 1, данные первых шести строк соответствуют хранимому в протоколах ИАОС цифровому образовательному следу, строки 7 и 8 задаются моделью электронного курса, а последние две строки отражают данные из моделей сущностей (обучающегося, учителя и тьютора) и общей логики работы системы.

Пусть в момент времени t наш обучающийся с индексом k инициировал событие ω (характеризуется масштабом, уровнем, аспектом и динамикой [Углев и др., 2022]). Тогда параметрическую *модель учебной ситуации* обозначим через $V(t,k,\omega)$. В её состав войдут следующие компоненты, характеризующие текущую обстановку:

- базовое состояние модели методиста – Ω (совокупность нормативных параметров организации обучения);
- текущее состояние индивидуализированной модели курса – Ω^* (совокупность параметров дидактического материала с учётом потребности учащегося, полученные в результате работы интеллектуального планировщика);
- текущее состояние модели обучающегося – U (совокупность параметров из ОЦС и результаты проверки гипотез об учащемся со стороны ИАОС);
- данные из образовательного цифрового следа, относящиеся к текущему состоянию процесса обучения – P (выборки их протоколов событий в ИАОС);
- данные из образовательного цифрового следа, относящиеся к истории процесса обучения (динамика, $P`$);
- база знаний ИАОС, включающая логику принятия решений моделей учителя и тьютора (Kb), в виде коллекции правил, прецедентов и пр.

Тогда общая модель учебной ситуации будет представлена в виде кортежа (1).

$$V(t,k,\omega)=< U_k, P, P` \mid \Omega, \Omega^*, Kb>. \quad (1)$$

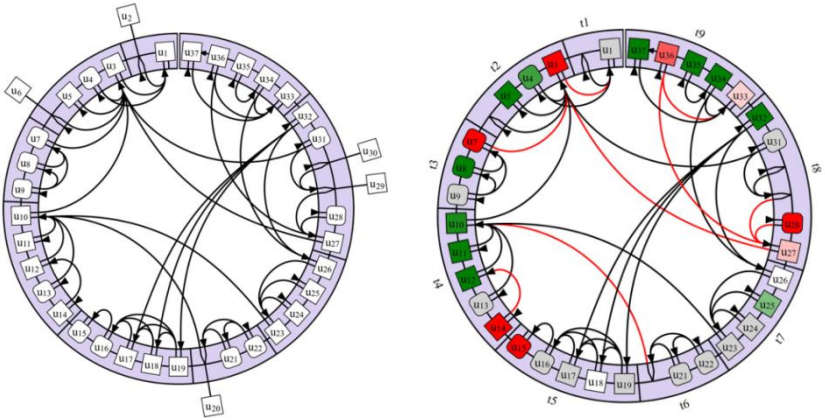
Так как *V* группирует в рабочей памяти ИАОС данные для конкретной учебной ситуации, предполагающей выполнение акта принятия решений, то мы будем совокупность этих данных называть *параметрической картой* учебной ситуации. Предложенный подход позволяет не только рассмотреть далее оригинальные методы обработки *V* для управления учебным процессом, но и для автоматической выработки пояснений принятых решений.

2. Картирование учебной ситуации

Параметрическая карта, как сгруппированная выборка данных для принятия решения, обуславливает только структурный состав анализируемых параметров. Собственно, механизм принятия решений при наступлении ω-го события реализуется интеллектуальным планировщиком ИАОС. Если мы обозначим через *M* модель реакции системы (педагогического воздействия), то переход от *V* к *y* (решению) в модели может быть реализовано за произвольное число шагов. В работе [Углев и др., 2022] мы предложили концепцию сквозного подхода к анализу учебной ситуации. Она предполагает, что синтактический, семантический и прагматический слои объединяются в виде единой графической структуры – *когнитивной карты диагностики знаний* (ККДЗ) [Uglev, 2024a].

Если кратко описать сущность ККДЗ, то это таким образом сгруппированные данные синтактики учебного материала, которые, дополненные семантическими связями, формируют подложку (контур) для наложения данных об оперативной обстановке (денотат), различных аспектах рассмотрения, а также позволяет выделять выявленные автоматически интеллектуальной системой акценты. Визуально такую карту можно представить в виде модели маленького мира [Milgram et al., 1967].Например, для картирования учебного курса на карте выделяются отдельные дидактические единицы (u_i), сгруппированные по темам и расположенные так, как предписывает образовательная траектория из Ω^*. Поэтому если подложка карты может содержать только данные из *V* без компонентов образовательного цифрового следа *P* (индивидуализированная карта, рис. 2,а), то для принятия решения в текущей ситуации на карту наносятся оперативные данные, отражающие текущую диспозицию в одном из аспектов рассмотрения (частная карта, знаниевый аспект, рис. 2,б). Цветом на рисунке выделены оценки результатов деятельности обучающегося относительно каждой подлежащей проверке u_i: красный цвет соответствует низкой оценке показателя в рассматриваемом аспекте; зелёный – высокой;

переход от красной к белой (нейтральной) и от бело к зелёной – промежуточным вариантам оценки; серый обозначает не оцениваемые компоненты. Анализ карты происходит автоматически, но и на визуальном отображении видно, например, что проблемы в освоении дидактической единицы u_{28} «Структуры данных при формализации баз знаний и алгоритмы их обхода» обуславливается не только результатами изучения u_{27} «Интеллектуальный решатель и цикл его работы», но и, с высокой долей вероятности, изучения элемента карты u_3 «Схема цикла разработки интеллектуальных информационных систем». Такая гипотеза проверяется сразу в нескольких аспектах анализа и на этом основании в дальнейшем ИАОС синтезирует рекомендации. Обновление содержимого параметрической карты происходит по каждому событию в ИАОС, что позволяет сформировать актуальное визуальное отображение *V*.



а б

Рис. 2. Индивидуализированная и частная ККДЗ на примере учебного курса «Системы искусственного интеллекта» магистра специальности «Информатика и вычислительная техника» [Uglev, 2024a]

Специфика такого представления данных об учебной ситуации позволяет интеллектуальному решателю независимо проанализировать данные относительно базы знаний любой из сущностей (модели ученика, учителя и тьютора), найдя компромиссное решение и выработать аргументы для объясняющего (методического) диалога с обучающимся. Подробнее об обосновании подхода к выработке решений см. в [Углев и др., 2022]. А вот механизм реализации работы с учебной ситуацией следует раскрыть подробнее.

3. Реализация механизма оценки учебной ситуации

Рассмотрим процесс обработки данных из *V* на примере экспериментальной ИАОС AESU. В её основе лежит подход, который предполагает использование экспертных систем на базе технологии доски объявлений (black board) [Jackson, 1999].

Типовым компонентом, реализующим принятие решение в ИАОС AESU, является модель решений *M*, которая ассоциируется с определённым событием в обучающей системе и предполагает выгрузку из параметрической карты необходимой выборки данных. Такие данные представляются в виде ККДЗ, но, без необходимости, не выводятся на экран пользователя. Сборка различных типов ККДЗ реализуется подключаемым компонентом продукционных экспертных систем для каждой сущности по-отдельности: сначала карта для модели ученика, затем учителя, а потом ищется компромиссная конфигурация с помощью логики модели тьютора.

Модель решений поддерживает процесс фазификации входных количественных данных [Zadeh et al., 2018]. Для этого в графе модели настраиваются соответствующие характеристические функции, а логика соответствующего фрагмента базы знаний работает уже с качественными значениями (результатом фазификации), дополненными коэффициентом уверенности [Uglev, 2024c]. Пример графа решений, позволяющий синтезировать индивидуализированный состав электронного курса и сборки для него ККДЗ (см. рис. 2,а), приведен на рис. 3 (правая часть окна), а характеристические функции для фазификации одного из входов – в левой части окна программы.

Рис. 3. Интерфейс конструктора модели решений в окне программы FLM_Builder с загруженной моделью по индивидуализации состава учебного курса [Uglev, 2024c]

Приведем пример: пусть имеется запрос о помощи со стороны учащегося при выполнении задания (ω_{17}) из дидактической единицы u_{28} с рис. 2,б. В соответствии с этим событием, ассоциированным в ИАОС с

моделью M_{17}, априори имеется соответствующая часть базы знаний (правила Kb_{17}), а также модель электронного курса (дидактические единицы с параметрами важности в теме и курсе, сложности изучения, семантических связей и пр. метаданные из Ω), имеющих отношение к u_{28} и связанным с ней элементам. Уточнение текущей конфигурации курса Ω^*, соответствующего индивидуальной образовательной траектории, ранее выработанной интеллектуальным планировщиком и согласованной с обучающимся, описывается в виде подмножества параметров из Ω (см, например, отсутствие элементов u_{29} и u_{23} на картах с рис. 2).

С целью концентрации ситуационных данных подгружается статистика работы обучающегося с u_{28} в виде P и $P`$ (частотно-временные параметры обращений, результативность контроля и характер ошибок). Учёт личностных потребностей учащегося (предпочтения в знаниевом, компетентностном и целевом аспектах по отношению к рассматриваемой дидактической единице, а также текущая модель поведения) формируют выборку параметров из U_k. Таким образом, для оценки учебной ситуации V был сформирован набор параметров, которые затем передаются в M_{17} (на примере с рис. 3 – это узлы графа решений, окрашенные зелёным цветом). Очевидно, что альтернативы реакции ИАОС для M_{17} выявлены инженером по знаниям заранее. Тогда получается, что, при известном наборе анализируемых факторов, задача выработки решения представляется в виде чёрного ящика, который и следует раскрыть в виде дерева выработки решений (см. на рис. 3 пример организации связей через промежуточные гипотезы, окрашенные розовым цветом). Так как в рассматриваемом примере помощь требуется для компонента u_{28}, то модель M_{17} будет запущена сначала для этой дидактической единице, а затем для всех тех, которые сазаны прямо или косвенно с ней по ККДЗ. В результате оцениваются, ранжируются и предъявляются те меры и элементы курса для повторного обращения, которые были интеллектуальным планировщиком выбраны как наиболее значимые [Углев и др., 2022].

Механизм анализа учебной ситуации с помощью M-моделей позволяет реализовать фрагментированную базу знаний. Это даёт возможность быстро корректировать логику ИАОС со стороны инженера по знаниям, минимизируя работу программиста. При этом первичная оценка эффективности каждой отдельно взятой модели производится отдельно от остальных: вырабатываются критерии и по ним оцениваются результаты специально организуемых педагогических экспериментов. Для примера с рис. 3 из [Uglev, 2024c] оценивалась степень согласия предложенной планировщиком индивидуализированной модели учебного курса с мнением учащегося и оценивался уровень удовлетворённости человека разъяснениями со стороны ИАОС.

4. Результаты

Использование интеллектуальным решателем различных *M*-моделей обработки событий и их оперативное обновление стало возможным благодаря введению в экспериментальную ИАОС предложенной выше модели учебной ситуации. Её специфика заключается в том, что комплексный сбор данных о процессе обучения, объединяющий акценты прикладной семиотики и теории функциональных систем (афферентный синтез), позволил сконцентрировать необходимые для принятия решений данные в единой структуре (параметрической карте). Это привело к тому, что за последние два года число независимых разработчиков фрагментов баз знаний (инженеров по знаниям) увеличилось в трое, а у программиста, занимающегося поддержкой системы, снизилась нагрузка в вопросах настройки логики работы ИАОС AESU (до 40%). В целом это ускорило испытания новых методик, опираясь на возможность оперировать распределённой базой знаний. Ограничениями данного подхода является зависимость от качественной формализации метаданных об учебном материале (модель методиста) и объёма ОЦС учащегося, подлежащего анализу при обработке учебной ситуации.

Особенно ценным для проекта результатом стал механизм синтеза и визуализации различных типов ККДЗ. Синтез параметрической карты по сути стал элементом процесса метрической концентрации из [Углев и др., 2022]. Это позволило автоматически получать более убедительные аргументы при пояснении решений АИОС обучающемуся в ходе методического диалога. Данный результат также показывает, что в этой области следует продолжать исследования, являющиеся элементом более широкого направления XAI [Arrieta et al., 2020].

Заключение

Результативность работы инженера по знаниям в реальных проектах во многом определяется возможностями интеллектуального решателя извлекать исходные данные из источников предметных знаний [Гаврилова 2001]. Гибкие модели предобработки и концентрации знаний ускоряют не только формирование моделей принятия решений, но и последующую актуализацию баз знаний. Предложенная нами модель учебной ситуации для интеллектуальных автоматизированных обучающих систем продемонстрировала возможности гибкого использования при решении широкого класса задач принятия решений о педагогическом воздействии и его пояснении.

Полученный результат, базирующихся в том числе и на модели учебной ситуации, был получен в результате целого цикла работ по исследованию механизмов картирования. В частности, были детально проработа-

ны новые типы когнитивных карт диагностики знаний [Uglev, 2024a]. Как следствие, это привело к углублению знаний о применении средств когнитивной визуализации в ИАОС, что соответствует общемировым тенденциям в исследованиях по данному направлению [Ilves, 2018]. Значительную роль в этих исследованиях сыграл подход к модульной организации логики работы планировщика, включая анализ учебной ситуации.

Список литературы

[Anderson et al., 1995] Anderson J.R. et al. Cognitive tutors: Lessons learned //The journal of the learning sciences. – 1995. – Vol. 4(2). –P. 167-207.

[Arrieta et al., 2020] Arrieta A., Díaz-Rodríguez N., Del Ser et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information fusion. – 2020. – Vol. 58.

[Frerejean et al., 2019] Frerejean J. et al. Designing instruction for complex learning: 4C/ID in higher education // European Journal of Education. – 2019. – Vol. 54(4). – P. 513-524. –DOI: 10.1111/ejed.12363.

[Ilves et al., 2018] lves K., Leinonen J., Hellas A. Supporting self-regulated learning with visualizations in online learning environments // Proceedings of the 49th ACM Technical Symposium on Computer Science Education. – 2018. – P. 257-262. – DOI: 10.1145/3159450.3159509.

[Jackson, 1999] Jackson P. Introduction to Expert Systems, Addison-Wesley Pub. Co., Reading, 1999.

[Karpenko, 2011] Карпенко А.П., Добряков А.А. Модельное обеспечение автоматизированных обучающих систем. Обзор // Машиностроение и компьютерные технологии. – № 7. – С. 1-63.

[Koedinger et al., 2007] Koedinger K.R., Aleven V. Exploring the assistance dilemma in experiments with cognitive tutors // Educational psychology review. – 2007. – Vol. 19(3). – P. 239-264. – DOI: 10.1007/s10648-007-9049-0.

[Lieder, 2019] Lieder F. et al. A cognitive tutor for helping people overcome present bias. – 2019.

[Milgram et al., 1967] Milgram S. et al. The small world problem // Psychology today. – 1967. – Vol. 2(1). – P. 60-67.

[Skinner, 1986] Skinner B.F. Programmed instruction revisited // Phi Delta Kappan. – 1986. – Vol. 68(2).

[Uglev, 2024a] Uglev V.A. Cognitive Maps of Knowledge Diagnosis (CMKD): the essence of the method, classification, characteristics and synthesis principles // Novel & Intelligent Digital Systems: Proceedings of the 4th International Conference. NiDS 2024. LNNS. Vol 1170. – Springer, Cham. – DOI: 10.1007/978-3-031-73344-4_51.

[Uglev et al., 2024b] Uglev V., Smirnov G. A Cross-Cutting Approach to Analysis of the Learning Situation in ITS Using a Mapping Mechanism // Journal of Integrated Design and Process Science. – Vol. 27(3-4). – DOI: 10.1177/10920617241289777.

[Uglev, 2024c] Uglev V.A. Implementation of Decision-Making Mechanism in the Intelligent Tutoring System Based on the Expert Systems Module // Pattern Recognition and Image Analysis. – Vol. 34, No. 3. – DOI; 10.1134/S1054661824700615.

[Van Merriënboer et al., 2002] Van Merriënboer J.J.G., Clark R.E., De Croock M.B.M. Blueprints for complex learning: The 4C/ID-model // Educational technology research and development. – 2002. – Vol. 50(2). – P. 39-61.

[Zadeh et al., 2018] Zadeh L.A., Aliev R.A. Fuzzy logic theory and applications: part I and part II, World Scientific Publishing, 2018.

[Анохин, 1975] Анохин П.К. Очерки по физиологии функциональных систем. – Рипол Классик, 1975.

[Беспалько, 1970] Беспалько В.П. Программированное обучение. Дидактические основы. – М., 1970.

[Гаврилова и др., 2001] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001.

[Поспелов и др., 1999] Поспелов Д.А., Осипов Г.С. Прикладная семиотика // Новости искусственного интеллекта. – 1999. – № 1.

[Рыбина, 2023] Рыбина Г.В. Интеллектуальные обучающие системы на основе интегрированных экспертных систем: учеб. пособие. – М.: Директ-Медиа, 2023.

[Сычёв и др., 2025] Сычёв О.А., Пенской Н.А., Терехов Г.В. Метод разработки интеллектуальных тренажеров на основе онтологии предметной области // Онтология проектирования. – 2025. – Т. 15, № 1(55). – С. 67-81.

[Углев и др., 2022] Углев В.А., Гаврилова Т.А. Подход к реализации сквозной визуальной поддержки процессов принятия решений для интеллектуальных автоматизированных обучающих систем // XX национальная конференция по искусственному интеллекту с международным участием (КИИ-2022). В 2 т. Т. 2. – М.: Изд-во МЭИ, 2022. – С. 413-426.

УДК 004.81
doi: 10.15622/rcai.2025.009

ЭМБЕДДИНГ ИНТЕНСИЙ РЕЧЕВЫХ АКТОВ В СЕМАНТИЧЕСКОЕ ПРОСТРАНСТВО

Д.Л. Хабаров (*hdl001@campus.mephi.ru*)
А.В. Самсонович (*avsamsonovich@mephi.ru*)

Национальный исследовательский ядерный университет «МИФИ», Москва

В работе представлена семантическая карта интенсий, понимаемых как коннотации речевых актов. Результат включает набор 125 пар интенсий, вложенных в семантическое пространство, и граф отношений между ними, а также нейронную сеть, обученную распознавать заданные интенции в высказываниях. Этот инструмент может использоваться для создания формальных представлений социально-реляционных аспектов речевых актов в диалоге. Метод построения карты основан на использовании OpenAI ChatGPT, тонкой настройке большой языковой модели (БЯМ), линейной алгебре и теории графов. Построенная модель выходит за рамки популярных подходов к анализу тональности или эмоциональной окраски текстов на естественном языке. Будучи общей моделью, она может использоваться для создания специализированных моделей для ограниченных парадигм. Это позволяет эффективно интегрировать БЯМ с символьными моделями, такими как когнитивная архитектура eBICA, для создания социально-эмоциональных диалоговых агентов.

Ключевые слова: семантическое картирование, LLM, BICA, DistilBERT, нейро-символьная интеграция, социальные интеллектуальные агенты.

Введение

Появление генеративного искусственного интеллекта (ИИ) открыло новые возможности взаимодействия человека и компьютера, выявив и существенные ограничения [Borazjanizadeh et al., 2024]. Рассмотрим два основных подхода в ИИ: «статистический ИИ», включая большие языковые модели (БЯМ), развивающий когнитивные функции через статистическое обучение данным без учета знаний предметной области, и «когни-

тивный ИИ», включая когнитивные архитектуры (КА), где семантика и функциональность закладываются разработчиками «вручную». КА уникальны благодаря их способности воспроизводить высшие функции разума (человекоподобную мотивацию, социально-эмоциональный, метакогнитивный, телеологический интеллект и т.п.), в чем статистический ИИ слаб. Однако КА, будучи созданными вручную, ограничены узкими парадигмами, тогда как БЯМ универсальны. Напрашивается мысль, что два подхода могли бы дополнить друг друга. Раскрытие данного потенциала возможно путем интеграции подходов, требующей двустороннего преобразования между естественным языком и формальными представлениями символьных моделей. Это преобразование может выполняться специально обученными БЯМ [Liu et al., 2024].

Ключевая область применения – социальные когнитивные агенты с эмоциональным интеллектом [Marsella et al., 2010], [Jung et al., 2011]. Такие агенты должны не только владеть языком, но и понимать психологическое состояние собеседника, включая его социально-реляционные установки, для генерации адекватной многомодальной реакции. Достижение уровня человека требует выхода за рамки базовых эмоций и популярных моделей [Russell, 1980], [Mehrabian, 1995], [Ekman, 1992], [Plutchik, 1982], [Ortony et al., 1988].

1. Основные понятия

В данной работе предлагается нейросетевой подход к формальному описанию множества интенциональностей, значимых для социальных отношений в диалоге [Jacob, 2023], [Zhuravlev et al., 2016]. Далее в фокусе внимания будет понятие «интенсия», определенная как социально-реляционная коннотация речевого акта (в отличие от понятия «интенция», означающего «намерение» [Scheer, 2004]). Интенсия же, понимаемая как коннотация, обозначает смысл речевого акта как "знака" [Wikipedia Contributors, 2024]. Это понятие близко к понятию перлокутивной интенции (ожидаемый эффект на слушателя) [Frijda, 1993б Searle, 1979].

Цель данной работы – построение семантической карты, характеризующей вербальные коммуникации. Термин "семантическая карта" здесь означает представление множества интенсий векторами в семантическом пространстве вместе с графом семантических отношений между ними [Huth et al., 2016], [Huth et al., 2012], [Samsonovich, 2018a]. Под семантическим пространством понимается линейное пространство, элементы которого несут смысловую нагрузку, а его геометрические свойства отражают семантические связи между элементами [Cowen, Keltner 2021], [Samsonovich, 2018b], [Samsonovich et al., 2010]. Близкое по смыслу понятие – концептуальное пространства [Gardenfors, 2004].

Отношения между интенсиями. После формирования списка интенсий определяются семантические и прагматические отношения между ними. Мы ограничимся четырьмя видами отношений: антонимия, комплементарность, антикомплементарность, активность/реактивность (контрапарты). Примеры будут указаны в следующем разделе.

2. Построение исходного списка интенсий

Здесь мы описываем, как в этой работе был построен набор интенсий.

2.1. Источники и требования для списка интенсий

Изучение конкретных случаев различных социальных ситуаций было основным источником для составления набора практически значимых интенсий. Другие источники включают данные анализа тональности и настроений (sentiment and tone analysis), сложные модели эмоций и связанные с ними наборы данных [HUMAINE, 2006].

Основные требования к списку интенсий следующие: интенсия не должна быть одним словом, а представлять собой краткое описание действия. Интенсия должна быть социально направленной – включать получателя действия. Интенсия должна выражаться в предикативной форме: *P*(*a*,*b*,*c*,*d*), где *a* – агент, *b* – получатель, *c* – действие (например, *express, ask, offer*), а *d* – дополнительная детализация действия (например, *express sympathy, ask for help*). Допускаются уточнения для *d*, например: *Express interest in person's thoughts*. Интенсия не должна состоять из логической комбинации нескольких предикатов. Не должны включаться предварительные условия, контекст ситуации, конкретные формулировки или ожидаемые реакции получателя.

2.2. Процедура составления

Экспертами, которые обладали достаточными знаниями в предметной области, был создан список из 43 интенсий на основе работ по речевым актам [Searle, 1979] и интенциальностям в дискурсе [Zhuravlev et al., 2016], анализа корпусов диалогов (таких как Switchboard), исследования различных парадигм социального взаимодействия. Также моделировались диалоги в установленных парадигмах, чтобы учесть различные социальные ситуации, которые были источником новых интенсий.

Параллельно, с использованием специального промпта, модель GPT-4o сгенерировала приблизительно 1000 интенсий. После ручной фильтрации по тем же критериям, что и для эталона, был получен очищенный БЯМ-список из 98 интенсий. Детальное сравнение двух списков показало высокую репрезентативность и точность БЯМ-результатов: 93% эталонных интенсий (40 из 43) имеют прямой или близкий семантический аналог в БЯМ-списке. На основе текущего списка БЯМ модель генерировала новые примеры, процесс продолжался несколько десятков итераций, пока

все новые примеры еще не были в существующем списке. Таким образом было построено максимально возможное покрытие. Фрагмент результирующего набора данных представлен в табл. 2.

Таблица 2

Интенсия	Антоним
Encourage positive emotions in the interlocutor	Evoke negative emotions or discourage enthusiasm
Express a desire to communicate	Show disinterest in communication
Express doubt for something	Express certainty or unquestioning belief
End the conversation in a polite way	End the conversation abruptly or rudely

3. Установление отношений и формирование графа интенсий

Следующий этап включал выявление комплементарных и антикомплементарных связей между интенсиями. Примеры комплементарных отношений представлены в табл. 3. Для каждой интенсии и ее антонима эксперт назначал комплементарную интенсию – социально правдоподобный ответ реципиента. Если подходящая формулировка уже существовала в наборе, то она же использовалась для комплементарной; иначе создавалась новая с соблюдением требований. Интенсии в колонке "Комплементарная" рассматриваются как реактивные (пассивные), поскольку являются реакциями на исходное действие. Некоторые из них могут проявляться и как активные в зависимости от контекста – быть инициированы агентом по собственному желанию.

Таблица 3

Интенсия	Комплементарная	Антикомплементарная
Acknowledge a person’s work	Express appreciation to a person	Show ingratitude or indifference
Adopt a formal tone in a conversation	Adopt a formal tone in a conversation	Adopt an informal tone in a conversation
Ask for clarification from a person	Clearly explain the details	Refuse to explain anything

Комплементарные связи стали основным механизмом расширения набора: исходные 98 интенсий увеличились до 125. Определение интенсий, которые являются реакциями на исходные интенсии, замыкает различные социальные ситуации, представленные этими же самыми интенсиями. Таким образом полученный список отражает набор ситуаций в интересующих нас парадигмах.

Каждая антонимическая пара активных интенсий определяет ось координат в семантическом пространстве для оценки речевых актов.

Итоговый граф (рис. 1) содержит три компоненты связности: основную (436 вершин), кластер формальный/неформальный тон (9 узлов) и кластер вежливый/невежливый стиль (4 узла). Распределение связей: 317 антонимических, 230 комплементарных, 243 антикомплементарных, 132 активно-пассивных.

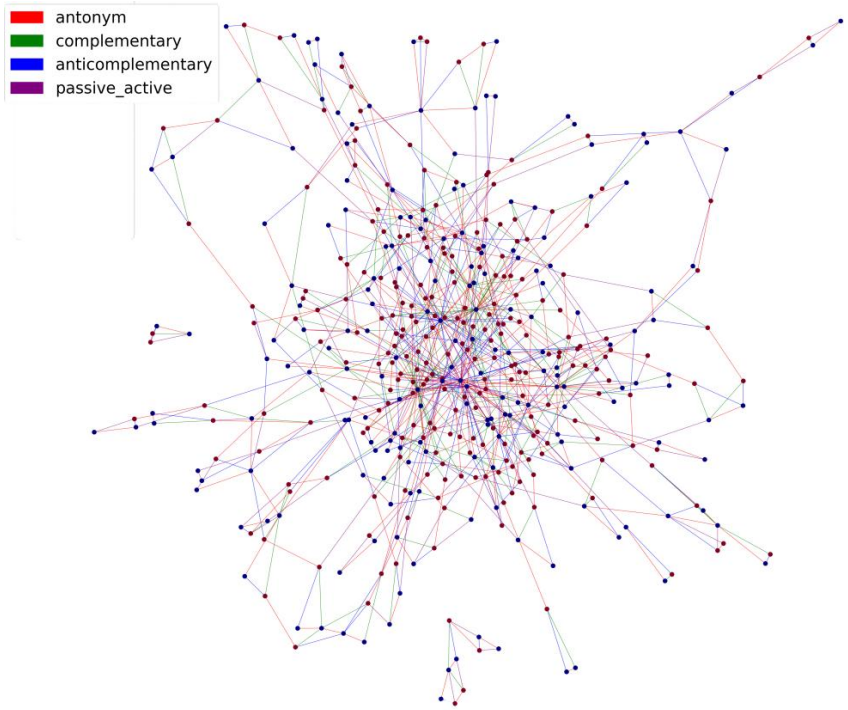


Рис. 1. Визуализация графика интенсий, легенда показывает цвета ребер

Степени вершин: min=2, max=47, средняя=4.11. Основная компонента: диаметр=13, радиус=7, центральных вершин=15. Точки сочленения: 21, мосты: 3. Минимальное внешне устойчивое множество: 99 вершин, минимальное вершинное покрытие: 241. Цикломатическое число: 476, хроматическое число: 4.

Две меньшие компоненты представляют собой обособленные подпространства, связанные со стилистическими особенностями (формальность и вежливость), что указывает на их относительную независимость от ядра. Граф обладает умеренной плотностью (средняя степень вершины 4.11), а также 21 точкой сочленения и 15 центральными вершинами. Это указыва-

ет на существование блоков в графе, которые могут функционировать как самостоятельные семантические единицы, несмотря на включенность в общую структуру.

4. Сокращение размерности семантического пространства

4.1. Маркировка полученного набора данных с использованием глубокой нейронной сети

Для матрицы оценок $Z \in R^{m \times n}$, где Z_i^j представляет оценку элемента i по шкале j, цель состоит в том, чтобы вложить элементы и шкалы в общее линейное пространство $A \simeq \mathbb{R}^d$. В этом пространстве элементы представлены векторами $X_i \in \mathbb{R}^d$, а шкалы – векторами $Y_i \in \mathbb{R}^d$, так что:

$$Z = XV^T. \quad (3)$$

Предполагая, что данные X декоррелированы и стандартизированы с помощью аффинного преобразования:

$$\frac{1}{m} X^T X - \bar{X}^T X \approx I, \ \bar{X} = \frac{1}{m} \sum_{i=1}^{m} X_i. \quad (4)$$

Эмпирическая матрица Грама для шкал аппроксимируется ковариацией Z:

$$\hat{G} = \frac{1}{m} Z^T Z - \bar{Z}^T Z \approx VV^T. \quad (5)$$

Для восстановления векторов шкал V выполняется спектральное разложение:

$$\hat{G} = U\lambda U^T, \hat{V} = \sqrt{\lambda} U^T. \quad (6)$$

Решение $\hat{V}$ определено с точностью до ортогонального преобразования Q, $\hat{V} \rightarrow Q\hat{V}$, а его размерность d определяется рангом $\hat{G}$. Отрицательные собственные значения в λ (обусловленные шумом) отбрасываются в качестве регуляризации.

Архитектура нейронной сети. Для оценки того, выражает ли речевой акт конкретную интенсию или ее антоним вдоль заданной семантической шкалы, мы разработали бинарную классификационную модель на основе **DistilBERT** (рис. 2) – облегченной и вычислительно эффективной версии BERT, работающей на 60% быстрее при потере точности не более 3% [SanhSanh et al, 2019]. Большое число интенсий и классификаторов было основной причиной использовать более вычислительно эффективную архитутектуру.

Входом модели является текстовая фраза, представляющая речевой акт. Токены преобразуются в эмбеддинги (768D), маска внимания отделяет реальные токены от паддинга. DistilBERT обрабатывает последовательность, выдавая контекстуализированные эмбеддинги токенов. Для классификации общего смысла применяется **mean pooling** для всей последовательности эмбеддингов. В результате получается единый вектор, представляющий семантику фразы.

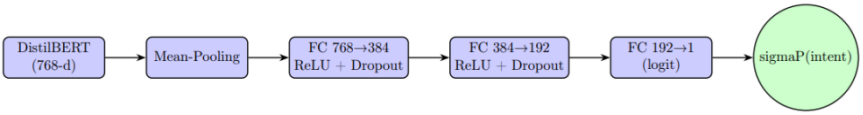


Рис. 2. Архитектура нейросети

Объединенный вектор предложения затем пропускается через трехслойную нейронную сеть прямого распространения.

Архитектура модели включает следующие слои (рис. 2):

- Первый слой преобразует 768-мерный вход в 384 измерения с последующей функцией активации ReLU и слоем dropout (rate=0.3) для предотвращения переобучения.
- Второй слой уменьшает вектор с 384 до 192 измерений, также с ReLU и dropout (0.3).
- Финальный слой – полносвязное линейное преобразование из 192 в единое скалярное значение (логит) без функции активации.

Для стабильности и эффективности обучения все линейные слои инициализируются с помощью инициализации униформы Ксавье. Логистическая функция – сигмоида с порогом 0.5 – определяет бинарную классификацию (антонимический класс при < 0.5).

Процесс обучения. Основная сложность заключается в отсутствии размеченных данных, или данных, которые потенциально содержали бы интенсии из нужного списка. Для создания датасета был сгенерирован набор примеров для каждой интенсии с использованием OpenAI API и следующего промпта:

«Сгенерируй 20 разнообразных реплик, выражающих интенцию [НАЗВАНИЕ ИНТЕНСИИ] в диалоге. Реплики должны быть естественными и содержать явные признаки этой интенсии.»

Множество примеров было расширено через парафразирование: GPT-4o сгенерировал по 4 парафраза для каждой базовой фразы. Итоговый набор содержал по 80 примеров на интенсию. Каждый речевой акт проверялся экспертом на содержание указанных интенсий. Было произведено сравнение примеров, генерируемых GPT-4o и DeepSeek. Среднее попарное косинусное сходство примеров одной интенсии для рассмотренных интенсий = **0.66**, а косинусное сходство усредненных эмбеддингов примеров = **0.97**, что показывает очень сильную семантическую близость примеров, созданных разными БЯМ.

Кроме того, был проведен эксперимент, в котором сгенерированные примеры речевых актов для набора интенсий размечала группа экспертов, также обладающих знаниями в этой области, на предмет содержания интенсии в представленном речевом акте: 95% доверительный интервал для коэффициента корреляции Пирсона - [0.84, 0.87], что показывает высокий уровень согласованности оценок экспертов и БЯМ.

Оценка результатов. Каждая из 125 семантических шкал оценена отдельным бинарным классификатором, обученным различать целевую интенсию и ее антоним. Средняя точность классификации по всем шкалам на тестовой выборке составила **94.25%** (SD = 5.05%), демонстрируя высокую точность классификации.

Для каждого сгенерированного речевого акта были применены все бинарные классификаторы, а выбор интенсии определялся как первый по степени уверенности. Параллельно с этим тестировалась модель, обученная как многоклассовый классификатор, хотя такой подход не отвечает поставленной цели. Многоклассовый классификатор нормирует вероятности и выбираем одну лучше представленную интенсию, а для текущей задачи необходимо для каждой интенсии оценить степень ее присутствия в интервале [0,1], и только потом уже можно определить лидирующую интенсию. Кроме того, такая система не является гибкой в случае удаления или добавления новых итенсий в списке.

Для части данных, размеченных экспертами, были применены все 125 классификаторов и точность по вхождению топ-1 составила **87,12**%, а для вхождения в топ-3 **91,56%**, что показывает хороший уровень определения лидирующих интенсий, а также помогает оценить степень присутствия любой другой.

Разметка данных. После обучения классификаторов создан набор из 1750 высказываний, оцененных по всем 125 семантическим шкалам. Каждое высказывание пропущено через каждый бинарный классификатор, получая "сырые" оценки на интервале [0, 1]. Оценки линейно преобразованы в интервал [-1, 1] для последующих преобразований (1)–(4). Результирующая матрица Z формирует начальное семантическое представление высказываний в пространстве интенсий.

4.2. Сокращение размерности семантического пространства

После преобразования матрицы Грама была получена матрица размерностью 125×92, которая является вложением исходного набора шкал V. Результат – семантическое пространство как вложение интенсий, с помощью которого можно определить углы между шкалами для выявления семантической близости или ортогональности. Примеры минимальных углов между шкалами:

- "Encourage hope in interlocutor" – "Encourage positive emotions in interlocutor" (32.71°).
- "Express care for person" – "Express empathy to person" (33.33°).
- "Inspire person" – "Motivate person" (33.41°).

Найденные ортогональные координаты (которые являются линейными комбинациями исходных шкал) затруднительно интерпретировать семантически. Поэтому для построения семантических подпространств предла-

гается выбирать подмножество шкал, образующих косоугольный полный базис (например, можно объединить семантически близкие шкалы, а также те шкалы, которые ближе всего к главным компонентам).

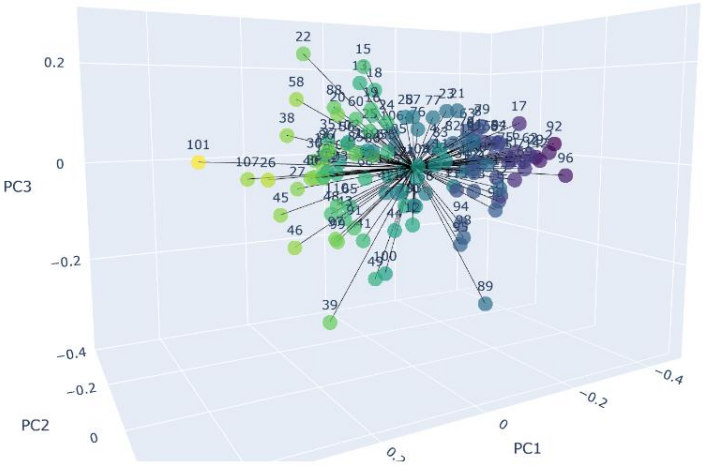


Рис. 3. Проекция шкал на первые три компоненты во вложении, линии идут от нулевых координат

Полученный эмбеддинг (рис. 3) показывает, насколько близки шкалы друг к другу. Минимальный угол (32.71°) не позволяет напрямую объединять отдельные пары. Тем не менее с их помощью все еще можно выделить группы семантически близких шкал (например: **Inspire a person in something**, **Motivate a person for something** и **Help a person restore their confidence**). Эти группы могут быть объединены путем редактирования списка интенсий. Альтернативно, можно работать с полным набором 125 шкал.

Полученная модель позволяет строить специализированные семантические подпространства для конкретных социальных ситуаций (при помощи выбора подмножеств координат), а также строить функции перехода между подпространствами, используя полученное вложение. Это открывает возможности для моделирования динамики диалога с правилами перехода между его этапами.

Заключение

Результатом данной работы является методика построения семантической карты коммуникативных интенсий, а также сама построенная семантическая карта, включающая список интенсий, их эмбеддинг в семантическое пространство и граф их отношений. Охарактеризованы топология графа и размерность пространства. Составлен набор из 125 антонимиче-

ских пар активных интенсий, интерпретированных как семантические шкалы. Граф их отношений содержит 449 узлов и 922 ребра. Разделение графа на три компоненты связности указывает, что некоторые подмножества интенсий могут формировать обособленные семантические подпространства.

Для оценки естественных высказываний относительно этих шкал обучена серия бинарных классификаторов на архитектуре DistilBERT, а также реализован многоклассовый классификатор на данной основе. Модели продемонстрировали высокую точность со средним значением 88.12%, полученным на данных реальных диалогов, подтвердив надежное распознавание нюансов социальных коннотаций в естественном языке.

Кроме того, методами линейной алгебры получено 92-мерное вложение шкал интенсий. Это вложение позволяет сократить размерность семантической карты путем отбрасывания менее информативных координат, сохраняя при этом все 125 исходных шкал в новом пространстве меньшей размерности как переопределенный базис. Данные модели могут использоваться для определения правил перехода между семантическими подпространствами, что полезно при построении моделей социальной динамики на базе КА типа eBICA [Samsonovich, 2020].

Перспективные направления включают: расширение набора интенсий через новые типы семантических отношений; сжатие набора путем удаления семантически избыточных концептов с использованием вложений; улучшение классификаторов путем добавления данных, включая примеры речевых актов без целевых интенсий; построение специализированных семантических подпространств для конкретных социальных сценариев (фаза знакомства, поиск общих интересов).

Список литературы

[Borazjanizadeh et al., 2024] Borazjanizadeh N., Piantadosi S.T. Reliable reasoning beyond natural language. arXiv:2407.11373v2 (19 July 2024). – doi: 10.48550/arXiv.2407.11373.

[Liu et al., 2024] Liu S., Xu J., Tjangnaka W., Semnani S.J., Yu C.J., Lam M.S. SUQL: Conversational search over structured and unstructured data with large language models. arXiv:2311.09818v2 (2024). – doi: 10.48550/arXiv.2311.09818.

[Marsella et al., 2010] Marsella S., Gratch J., Petta P. Computational models of emotion / In: Scherer, K.R., Bänziger, T., Roesch, E. (Eds.) A Blueprint for Affective Computing: A Sourcebook and Manual. – Oxford: Oxford University Press, 2010.

[Jung et al., 2011] Jung Y., Kuijper A., Fellner D., Kipp M., Miksatko J., Gratch J., Thalmann D. Believable virtual characters in human-computer dialogs: State of the art report // In: Proc. 32nd Annual Conference of the European Association for Computer Graphics. – ACM Press, 2011. – P. 1-26.

[Russell, 1980] Russell J.A. A circumplex model of affect // Journal of Personality and Social Psychology. – 1980. – Vol. 39(6). – P. 1161-1178.

[Mehrabian, 1995] Mehrabian A. Framework for a comprehensive description and measurement of emotional states // Genetic, Social, and General Psychology Monographs. – 1995. – Vol. 121(3). – P. 339-361.

[Ekman, 1992] Ekman P. An argument for basic emotions // Cognition and Emotion. – 1992. – Vol. 6(3). – P. 169-200. – doi: 10.1080/02699939208411068.

[Plutchik, 1982] Plutchik R. A psychoevolutionary theory of emotions // Social Science Information. – 1982. – Vol. 21. – P. 529-553.

[Ortony et al., 1988] Ortony A., Clore G.L., Collins A.M. The Cognitive Structure of Emotions. – Cambridge: Cambridge University Press, 1988.

[Jacob, 2023] Jacob P. Intentionality. The Stanford Encyclopedia of Philosophy (Spring 2023 Edition) / E.N. Zalta, U. Nodelman (Eds.). – URL: https://plato.stanford.edu/archives/spr2023/entries/intentionality.

[Zhuravlev et al., 2016] Zhuravlev A.L., Pavlova N.D., Zachesova I.A. On discourse, discursive influence and information-psychological security: Instead of a preface / In: A.L. Zhuravlev, N.D. Pavlova, I.A. Zachesova (Eds.). Psychology of Discourse: Problems of Determination, Influence, Security [Psikhologiya Diskursa: Problemy Determinatsii, Vozdeystviya, Bezopasnosti]. – Moscow: Institute of Psychology RAS, 2016. – P. 5-10 (in Russian).

[Huth et al., 2016] Huth A.G., De Heer W.A., Griffiths T.L., Theunissen F.E., Gallant J.L. Natural speech reveals the semantic maps that tile human cerebral cortex // Nature. – 2016. – Vol. 532(7600). – P. 453-458. – doi: 10.1038/nature17637.

[Huth et al., 2012] Huth A.G., Nishimoto S., Vu A.T., Gallant J.L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain // Neuron. – 2012. – Vol. 76. – P. 1210-1224. – doi: 10.1016/j.neuron.2012.10.014.

[Samsonovich, 2018a] Samsonovich A.V. On semantic map as a key component in socially-emotional BICA // Biologically Inspired Cognitive Architectures. – 2018. – Vol. 23. – P. 1-6. – doi: 10.1016/j.bica.2017.12.002.

[Cowen, Keltner 2021] Cowen A.S., Keltner D. Semantic space theory: A computational approach to emotion // Trends in Cognitive Sciences. – 2021. – Vol. 25(2). – P. 124-136. – doi: 10.1016/j.tics.2020.11.004.

[Samsonovich, 2018b] Samsonovich A.V. On semantic map as a key component in socially-emotional BICA // Biologically Inspired Cognitive Architectures. – 2018. – Vol. 23. – P. 1-6. – doi: 10.1016/j.bica.2017.12.002.

[Samsonovich et al., 2010] Samsonovich A.V., Goldin R.F., Ascoli G.A. Toward a semantic general theory of everything // Complexity. – 2010. – Vol. 15(4). – P. 12-18. – doi: 10.1002/cplx.20293.

[Gardenfors, 2004] Gardenfors P. Conceptual Spaces. Cambridge. – MA: MIT Press, 2004.

[Scheer, 2004] Scheer Richard. The ‘mental state’theory of intentions // Philosophy. – 2004. – 79.1. – P. 121-131.

[Wikipedia Contributors, 2024] Wikipedia Contributors. Intension. Wikipedia, The Free Encyclopedia. – 2024. – URL: https://en.wikipedia.org/w/index.php?title=Intension&oldid=1242642153.

[Vanderveken, Searle, 1985] Vanderveken D., Searle J.R. Foundations of Illocutionary Logic. – Cambridge: Cambridge University Press, 1985.

[Samsonovich, 2013] Samsonovich, A.V. Emotional biologically inspired cognitive architecture // Biologically Inspired Cognitive Architectures. – 2013. – Vol. 6. – P. 109-125. – doi: 10.1016/j.bica.2013.07.009.

[Samsonovich, 2020] Samsonovich A.V. Socially emotional brain-inspired cognitive architecture framework for artificial intelligence // Cognitive Systems Research. – 2020. – Vol. 60. – P. 57-76. – doi: 10.1016/j.cogsys.2019.12.002.

[HUMAINE, 2006] HUMAINE Emotion Annotation and Representation Language (EARL). – 2006. – URL: http://emotion-research.net/projects/humaine/earl.

[Searle, 1979] Searle John R. Expression and meaning: Studies in the theory of speech acts. – Cambridge University Press, 1979.

[SanhSanh et al, 2019] Debut L., Chaumond J., & Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter // arXiv preprint arXiv:1910.01108. – 2019.

Секция 2 | ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

УДК 620.22–4:004.65:004.82
doi: 10.15622/rcai.2025.010

МАЛЫЕ ДАННЫЕ – ЭТО ВСЕ ЧТО У ВАС ЕСТЬ

А.В. Аментес (*Artem.amentes@yandex.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В работе описывается подход к решению задач связанных с расширяющимися малыми наборами данных (МНД). Приведены описания малых наборов данных, их размеры и ограничения, с которыми сталкиваются исследователи. Также приведены методы интеллектуального анализа данных (ИАД), глубокого обучение и некоторые другие методы работы с МНД. ДСМ-метод рассмотрен как эффективная методология построения интеллектуального анализа данных и построения прогнозных моделей даже при очень малом размере data set'ов.

Ключевые слова: малые наборы данных, искусственный интеллект (ИИ), нейронные сети, большие данные, ДСМ-метод.

Введение

Модели искусственных нейронных сетей часто зависимы от больших объемов данных, необходимых для обучения. Актуальность представляемой работы – в том, что она посвящена отдельно стоящему подклассу задач, связанных с МНД. Существуют предметные области, данные в которых накапливаются крайне медленно, или вообще находятся в очень ма-

лом количестве, а собрать больше данных здесь не представляется возможным [Faraway, 2017]. Избежать обработки МНД не получится. Вытянутые вправо таблицы данных ОБЪЕКТЫ х ПРИЗНАКИ всегда есть в отраслях с глубокой экспертизой. Чем выше экспертность специалиста тем структурно сложнее данные, являющиеся не только многопараметричными, но и, нередко, разнородными. При этом в реальных задачах модели ИИ часто сталкиваются с примерами, которые не похожи на обучающие, что позволяет говорить о проблеме нехватки репрезентативных данных и необходимости надёжного выявления таких случаев для повышения устойчивости и надёжности ИИ-систем и решений. Ошибки на данных, не похожих на обучающие, являются серьезным вызовом при обработке МНД. За годы исследований методов работы с МНД, накоплено значительное количество подходов. При этом, многие идеи, разработанные еще в прошлом веке, только сегодня стали активно применяться для работы с накапливаемыми данными. Вместе с тем, в 2020-е годы все чаще появляются работы, связанные исследованием причинно-следственных связей в постоянно накапливаемых эмпирических данных. Все дело в том, что «объяснимость» как одно из свойств надежного ИИ должна отвечать пользователю на вопрос «*Почему*?», что связано с причинами формирования моделью ИИ конкретного прогноза или ответа. Новизна нашего исследования заключается в демонстрации методологии, позволяющей не только давать уверенные ответы относительно наличия или отсутствия целевого свойства у изучаемого объекта посредством ИЭ модели ИАД, но и отвечать на вопрос «*Почему получен именно такой результат*?». А это решает проблему объяснимости работы такого класса программного обеспечения.

Практическая значимость предлагаемых методов, основанных на анализе причинно-следственных связей, заключается в том, что такие виды деятельности как медицина, военное дело, кибербезопасность, инженерное дело и др. предъявляют к интеллектуальным системам особые требования:

1) надежность – интерпретация работы модулей и функций;
2) устойчивость – сохранение интерполяционных связей каузального характера при пополнении базы прецедентов;
3) объяснимость – демонстрация причинных связей в найденном решении.

Эти три понятия создают основы для доверия к результату работы программного обеспечения. Лица принимающие решения должны обладать исчерпывающими обоснованиями. Современные модели искусственного интеллекта, в особенности обученные на ограниченном числе прецедентов, часто не обеспечивают качественной поддержки в этих вопросах.

1. Малые наборы данных

Современная гонка за большими данными сформировалась во многом благодаря особому классу ИЭ задач, когда на больших данных появляется возможность прогнозировать явления или свойства с достаточно большой статистической уверенностью. Сегодня приоритетом в ИИ принято считать интерполяционно-экстраполяционные модели (ИЭМ), которые в наибольшей мере представлены нейросетевыми архитектурами. Считается что, набор методов, связанных с нейросетевым обучением, позволяет сугубо вычислительными средствами имитировать человеческие познавательные когнитивные функции, а именно индуктивное обучение на примерах и экстраполяцию интерполированных зависимостей на новые данные. Современные ИЭМ часто зависимы от больших объемов данных, необходимых для их обучения (см., например, искусственные нейронные сети – ИНС). Актуальность представляемой работы связана с необходимостью обрабатывать расширяющиеся data set'ы малого объема и заключается в том, что требуется работать с отдельно стоящим подклассом задач, где предметом анализа оказываются МНД. Существуют предметные области, данные в которых накапливаются крайне медленно, или вообще находятся в очень малом количестве, а собрать больше данных не представляется возможным. Грубо говоря, существование МНД обусловлено отсутствием возможности собрать хоть сколько бы то ни было значимую выборку [Ballantyne, 2023].

МНД – это самостоятельно направление в науке об искусственном интеллекте. Природа данных такова, что существуют ограничения по сбору обучающих данных для некоторых событий, которые могут случаться раз в столетие, и собрать даже 100 таких примеров не представляется возможным. Также, существует ресурсный дефицит у исследователей, который не дает возможность ученым разметить 200 000 фото деревьев, когда ресурсов достаточно только для 200 образцов. Специфические задачи связанные, например, с редкими (мертвыми) языками сильно зависят от найденных артефактов. Существуют и другие регуляторные, политические, военные и этические барьеры (согласие пациентов, конфиденциальность, гостайна). Кроме того, данные должны быть надлежащего качества, что в значительной степени усложняет их сбор. МНД также могут расширяться во времени, что порождает новые вызовы при работе не только с МНД, но и с поступлением новых данных, которые могут не принадлежать ранее известному распределению данных в обучающей выборке. Вообще говоря, было бы полезно детально проанализировать и подходы к определению МНД [Abualigah, 2025]. К сожалению, на данный момент единой методологии отнесения данных к МНД не существует. Каждое направление деятельности характеризуется своим потоком данных и их размерностью.

Авторы чаще рассматривают МНД с точки зрения проблемы, которую они могут вызвать при использовании различных методов машинного обучения. Так, в статье [Wang, 2023] предлагается считать, что малая выборка это – такой набор данных, в котором мало аннотированных данных, или же аннотация является сложной или дорогостоящей. Описание малой выборки носит скорее качественный характер, нежели количественный. В статье [Hollmann et al., 2025], к МНД относят табличные данные, содержащие менее 10 000 строк. В работе [Сафонова, 2023] предлагается конкретное определение МНД в контексте глубокого обучения: МНД – это наборы данных, которые содержат менее 1,000 аннотированных примеров или те, которые плохо покрывают распределение признаков. Часто это приводит к недостаточности информации для эффективного извлечения значимых признаков методами глубокого обучения. Существуют также случаи, называемые авторами «extra-small» (сверхмалые данные), когда набор данных включает от 1 до 10 аннотированных примеров (например, при изучении редких природных катастроф [13]. В работе по изучению Байесовской модели для локализации модификаций белков (PTMProphet), модели глубокого обучения для контроля инсулина, иммунные анализы на МНД, сегментация сердечных МРТ с использованием 3D моделей на маленьком датасете (150 пациентов) столкнулись с ограничениями применения ИЭМ в медицине для создания клинических и программных решений. Основные проблемы здесь – низкая статистическая значимость, ограниченность признаков, высокая вероятность переобучения, ограниченная обобщаемость, и наличие смещений [Mikołajewski et al, 2023]. Медицинские приложения являются ярким примером использования МНД. Медицинские МНД часто состоят из нескольких десятков или сотен прецедентов, что препятствует обучению больших моделей без переобучения. Так в одной из работ, посвященных построению модели для диагностики опухолей мозга [Piffer, 2024], описывают МНД как ограниченное количество аннотированных образцов, обусловленное сложностью, дороговизной и трудностью сбора данных, часто ограниченное несколькими десятками или сотнями пациентов. Малый размер определяется в относительных терминах по отношению к сложности модели и числу параметров, подлежащих обучению (когда данных недостаточно для надежного обучения глубоких нейросетей). Авторами проведен систематический обзор 77 исследований с МНД (в среднем ~16 600 образцов, минимально 16). Представленные исследования относительно достаточности наборов данных для обучения искусственных нейронных сетей (ANN) показали, что правило «в 10 раз больше параметров» [Pasini, 2015] недостаточно консервативно для дискретного выбора. Обширные эксперименты Монте-Карло на синтетических данных с разной сложностью и уровнем шума доказали, что для стабильного обучения рекомендуют правило «в 50 раз

больше параметров», особенно при оценке качества модели по логарифму правдоподобия [Alwosheel et al, 2018]. В свою очередь в контексте работы с пищевыми продуктами, МНД возникают по причине «усталости вкуса», когда эксперты не могу обеспечить значительное количество экспериментов. Кроме того, уточняется, что такие данные, которые потребители собирают, пробуя разные продукты, обычно представлены таблицами, вытянутыми вправо. Количество признаков значительно, а количество наблюдений не велико [Corney, 2002]. Гетерогенность данных приводит к проблеме, когда внутри набора данных, как это было в пищевых исследованиях, находятся разнородные признаки. Предложенный метод работы с разбивкой данных на кластеры в соответствии с их природой косвенно можно считать с переходом на работу с МНД [Noroozi, 2023]. В обзорной статье [Nivedhaa, 2024] дефицит данных (small data) понимается как недостаточное количество**,** низкое качество или предвзятость (bias) тренировочного набора, что приводит к плохой обобщающей способности моделей, а отсутствие репрезентативности и дисбаланс классов усиливают риски переобучения и некорнректных прогнозов. МНД также можно понимать как недостаточно полные, нерепрезентативные или с сильными дисбалансами, что ведет к системным ошибкам и смещениям в моделях [de Miguel Beriain, 2022]. Это ограниченный объём данных с очень низкой долей неудачных примеров [Pajić, 2023], где количество доступных меток и образцов недостаточно для применения традиционных ML-моделей большого объёма. В данном случае речь о нескольких сотнях размеченных примеров при 1.4% неудач. Ряд авторов подчеркивают важность «data centric AI» – подхода, где фокус смещается с улучшения алгоритмов на улучшение качества данных. [Nisheva-Pavlova, 2022]. Данные должны быть связаны с людьми через своевременные, значимые инсайты, часто визуализированные и структурированные, чтобы быть понятными и полезными для повседневных задач. В ряде научных областей МНД являются нормой [Mendes, 2020].

Итак, мы рассмотрели некоторые подходы к определению понятия «малые данные» с количественной и качественной точек зрения. Объем таких data set’ов недостаточен для обучения искусственных нейронных сетей, а также порождает проблему переобучения в классических моделях ML. При этом авторы сходятся во мнении, что МНД неизбежны, так как всегда будут существовать редкие явления, узкоспециализированные направления деятельности, требующие глубокой экспертизы в сборе и разметке данных. Представленные методы интеллектуального анализа данных и получение надежных, устойчивых и объяснимых результатов работы над данными не обеспечивают качества, достаточного для принятия надежных решений.

2. Методы обучения интерполяционно-экстраполяционных моделей на малых выборках данных

Так как избежать работы с МНД не удается, то исследователи ИИ создают методологии и методы для эффективного решения задач ИАД и прогнозирования искомых свойств на данных. Любое использование программного обеспечения, которое обладает продвинутыми интеллектуальными особенностями не обходится без набора обучающих данных (примеров).

За годы исследований методов работы МНД, накоплено значительное количество подходов [Wang, 2023]: Data augmentation (1990-е); Transfer Learning (1990-е); Self-Supervised Learning (2010-е); Semi-Supervised Learning (1970-е); Few-Shot / Zero-Shot Learning (2000-е); Active Learning (1990-е); Weakly Supervised Learning (2000-е); Multi-Task Learning (1997-е); Ensemble Learning (1990-е); Process-Aware / Physics-Informed Learning (2010-е); Spatial Cross-Validation (2010-е); Causal Discovery (2020-е); Data-Centric AI (2020-е).

Не смотря на разнообразие методов улучшения качества работы моделей на МНД существует ряд проблем, которые на данный момент не решены надежным способом. Так проблемой является выявление, аудит и смягчение различных видов предвзятости (bias) и искажений в наборах данных, используемых для обучения алгоритмов принятия решений. Малое количество примеров для каждого целевого свойства влияют на появление некорректности, дискриминации и потери доверия к ИИ-системам. Примечательно, что помимо классических методов преодоления проблемы МНД, некоторые авторы предлагают интерактивный подход с «человеком в петле» (human-in-the-loop) и формирование разнообразных, мультидисциплинарных команд. Привлечение экспертов для повышения надежности работы моделей на МНД набирает популярность в ряде направлений применения ИЭМ [de Miguel Beriain, 2022]. Классические нейросетевые модели подвержены переобучению и нестабильности из-за малого числа примеров относительно сложности модели. Подчёркивается, что МНД имеют низкую вариативность, невысокую скорость обновления и ограниченный объём, но содержат высоко структурированные и информативные сведения, критичные для понимания анализируемых явлений. Улучшение процессов международного обмена данными позволит исследователям накапливать достаточного размера выборки для использования современных методов анализа [Mendes, 2020]. Глубокие нейросети требуют десятков тысяч размеченных образцов, которые часто недоступны [Piffer, 2024]. Важная проблема, по мнению авторов, заключается в несбалансированности классов. МНД часто характеризуются значительной несбалансированностью классов, что усложняет обучение модели. Несба-

лансированность классов приводит к тому, что модель чаще предсказывает преобладающий класс, что ведёт к ухудшению качества предсказаний редких классов [Safonova, 2023]. Существуют подходы, которые характеризуются меньшей зависимостью от объёма данных, что важно в прикладных областях с ограниченными данными. В описаниях этих подходов, авторы [Jiao, 2024] приводят в своем обзоре преимущества методов в различных задачах: визуальное представление, обработка текста, мультимодальные данные, медицинские приложения. Некоторые авторы [Kimpimäki, 2023] используют внедрение и адаптацию методов вычислительной абдукции (комбинация вычислительных методов и абдуктивного рассуждения) в области менеджмента и организационных исследований, в частности в поисках устойчивой стратегии управления.

Многие идеи, разработанные еще в прошлом веке, только сегодня стали активно применяться для работы с накопленными данными, тем не менее, в 2020-е годы все чаще появляются работы, связанные с исследованием причинно-следственных связей в данных. Дело в том, что «объяснимость» как одно из свойств надежного ИИ должна отвечать пользователю на вопрос «*Почему*?», что связано с поиском причин получения системой ИИ конкретного прогноза или ответа. Большинство методов здесь посвящены корреляционным моделям, опирающимся на различные манипуляции со статистическим подходами. Авторы пытаются увеличить надежность, сделать модели устойчивыми к аномалиям, а также увеличить набор данных для построения уверенных корреляций. Вместе с тем, существуют и другие подходы, которые опираются на логику рассуждений и позволяют преодолевать барьеры МНД не только в статике, но и в динамике их изменений. К сожалению, как показало углубленное изучение (в том числе – экспериментальный анализ на представленных ниже данных НМИЦ НХ им. Н.Н. Бурденко), рассмотренные в Разделах 1-2 известные подходы к обработке МНД не позволяют получать неоспариваемые прикладные результаты.

3. Метод интеллектуального анализа данных с учетом причинно-следственных связей

ДСМ-подход и базирующаяся на нем методология ИАД (ДСМ-ИАД) используют эвристику причинного сходства для идентификации факторов влияния, «вынуждающих» наличие целевых свойств у тех прецедентов в анализируемом data set’е, которые таковыми свойствами обладают.

Логико-математическими средствами ДСМ-рассуждения обеспечивается восстановление скрытых, т.е. представленных в неявном виде в анализируемых исходных эмпирических данных, причинно-следственных зависимостей вида *набор значений параметров => целевые свойства*.

ДСМ-метод может рассматриваться как методология организации интеллектуального анализа данных, ориентированного на выделение в анализируемых данных эмпирических зависимостей каузального типа, и позволяет формировать результативные решения для преодоления всех пяти типов проблем («барьеров») – классов ограничений математических ИЭ техник компьютерного анализа данных, о которых шла речь в Разделе I. При этом ДСМ-подход обеспечивает использующему его исследователю гибкость в выборе инструментария формализованного описания и анализа данных [Финн, 2010].

В части предлагаемых в ДСМ-подходе инструментальных средств представления данных и знаний это, в первую очередь это – возможности единообразной «логикой» (и алгоритмикой рассуждения) обрабатывать разнотипные данные, в том числе: булевские данные; значения в шкалах наименований; значения в порядковых шкалах; значения в метрических шкалах; числовые значения параметров; текстовые описания; семантически связанные между собою значения в описании каждого конкретного анализируемого прецедента.

Не менее существенна также и гибкость при выборе конкретных ДСМ-инструментов интеллектуального анализа данных [Финн, 2021].

Завершая рассмотрение возможностей и ограничений использования в компьютерном анализе данных наиболее распространенных математических моделей, выделим ДСМ-подход, еще раз фокусируясь на его характеристиках и преимуществах для эксперта в области ИАД. Суммируя рассмотренные выше доводы и аргументы, отметим, что ДСМ-подход предоставляет возможности для:

- интеграции в процесс экспертизы всех тех данных, которые эксперт считает релевантными наличию или, наоборот, отсутствию целевых эффектов, а также проактивной идентификации таких эффектов;
- обеспечения неформальной интерпретации и объяснения результатов ДСМ-ИАД-экспертизы с помощью эмпирических зависимостей причинно-следственного типа, выделяемых из анализируемых данных;
- оперирования актуальными, постоянно пополняемыми новыми элементами, обучающими коллекциями прецедентов, в том числе – ограниченными по своим текущим размерам [Забежайло, 2023].

В медицине онкологических заболеваний головного мозга, количество прецедентов, как правило, незначительное и позволяет накапливать новые данные крайне медленно. Такая ситуация требует использования адекватных средств анализа данных. Авторы работы [Аментес, 2024] представляют результаты ИАД накопленного в НМИЦ НХ им. Н.Н.Бурденко (г. Москва) примерно за 15 лет клинической практики data set’а размером в 250 прецедентов, каждый из которых характеризуется более чем 225 признаками. Такой набор данных (как таблица ОБЪЕКТЫ х ПРИЗНАКИ)

сильно «вытянут вправо». Кроме того, набор данных пополняется новыми примерами крайне редко (всего 10-15 случаев в год). Малая по числу прецедентов выборка не позволяет использовать классические статистические методы анализа данных. Представленная в докладе модель ИАД на базе ДСМ-метода, предлагает результативное решение этой проблемы. Интерполяционно-экстраполяционная модель на базе эвристики причинного сходства позволяет интерполировать обучающий data_set каузальными эмпирическими зависимостями и «диагностировать» новые объекты проверкой экстраполируемости на каждый из них тех причинно-следственных зависимостей (биомаркеров исследуемого эффекта), которые получены при интерполяции анализируемого data_set'а.

Методология ДСМ-метода реализована через стандартные методы программного языка python. Так, для построения сходства используются базовые модули фильтров данных, а для получения замыкания Галуа – алгоритмы перебора и сортировок столбцов и строк датафреймов. Универсальность метотодологии ДСМ позволяет решать задачи интеллектуального анализа данных существующими библиотеками, вместе с тем получая достоверные (проверяемые) и надежные (устойчивые) результаты сравнения внутри группы прецедентов методом рассуждений. Алгоритмика проверки «запрета на контрпримеры (ЗКП)» реализуется через порождения подмножеств на данных, обладающих противоположным свойством. Использование комбинации датафреймов, хранящихся в переменных среды обработки данных позволяет быстро проводить необходимые манипуляции с табличными данными. Механизмы сохранения найденных неподвижных точек, формируют локальную базу знаний (атлас зависимостей – маркеров целевого эффекта[1]). Такими инструментальными средствами выполняется весь пайплайн работы методологии рассуждения ДСМ-метода: поиск сходств, построение неподвижных точек, проверка на контрпримеры, построение интерпретируемого, объяснимого в терминах постановки задачи анализа, доступного для понимания эксперту, работающему с когнитивным интерфейсом программы.

Совместно с экспертами НМИЦ НХ Н.Н. Бурденко разработано программное обеспечение, реализующее методологию ДСМ-метода работы с малыми выборками [Аментес, 2024]. В ходе проведенных испытаний удалось получить прикладные результаты, которые не удавалось сформировать другими (см. Разделы 1-2 выше) средствами:

а) сформированы причинно-следственные зависимости – маркеры исследуемых эффектов (логические условия, выполненные на примерах и невыполненные на контрпримерах отдельно для двух эффектов – позитивного и негативного исходов операций);

[1] В обсуждаемых экспериментах – позитивного или, наоборот, негативного исхода нейрохирургической операции.

b) неоспариваемым образом (т.е. с выполнением условия ЗКП) найденными в процессе ДСМ-анализа факторами причинности (маркерами позитивного и негативного исходов операций) разделены примеры и контрпримеры из анализируемого data set'a.

Результаты, выдаваемыĕ работоӣ используемой компьютерно-ориентированной модели ИАД, позволяет не только отнести исследуемыӣ объект к одному из целевых классов, но и выдать заключение о причинах наличия целевого свойства, объясняя эту причинность в содержательном контексте терминов и понятий языка экспертов-медиков, предоставивших исходные данные для проводимого анализа.

Заключение

Существуют предметные области, где накопление данных происходит медленно, а количество параметров, описывающих свойства даже одного явления-прецедента достаточно велико. Так, в медицине онкологических заболеваний головного мозга, количество прецедентов, как правило, незначительное и позволяет накапливать новые данные крайне медленно. Малая по числу прецедентов выборка не позволяет использовать классические статистические методы, а также глубокие нейронные сети. Даже деревья решений не дают здесь устойчивого, интерпретируемого результата требуемой точности.

Представленная в настоящей статье компьютерно-ориентированная модель интеллектуального анализа данных на базе ДСМ-метода предлагает результативное решение этой проблемы. Ее возможности в реальных приложениях продемонстрированы на примере интеллектуального компьютерного анализа реальных данных НМИЦ НХ им. Н.Н. Бурденко (г. Москва). Интерполяционно-экстраполяционная модель на базе эвристики причинного сходства позволяет интерполировать обучающий data_set каузальными эмпирическими зависимостями и «диагностировать» новые объекты проверкой экстраполируемости на каждый из них тех причинно-следственных зависимостей, которые получены при интерполяции анализируемого data_set'a. Результат, выдаваемый работой такой модели интеллектуального анализа данных, позволяет не только отнести исследуемый объект к одному из целевых классов, но и сформировать заключение о наличии целевого свойства, объясняя причины его наличия у конкретного объекта в содержательном контексте терминов и понятий языка экспертов (медиков), предоставивших реальные исходные данные для выполняемого анализа.

Список литературы

[Аментес, 2024] Аментес А.В., Забежайло М.И. Об опыте разработки атласа биомаркеров исхода нейрохирургических операций // НТИ. Сер. 2. Инф. процессы и системы/ ВИНИТИ РАН. – 2024. – № 8. – ISSN 0548-0027.

[Забежайло, 2023] Забежайло М.И., Аментес А.В. О некоторых особенностях интеллектуального анализа коллекций эмпирических данных, пополняемых новыми сведениями, но ограниченных по своим размерам // Научно-техническая информация. Сер. 2. – 2023. – № 6. – С. 19-24.

[Финн, 2010] Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта // Искусственный интеллект и принятие решений. Ч. I. – 2010. – № 3. – С. 3-21; Ч. II // Там же. – № 4. – С. 14-40.

[Финн, 2021] Финн В.К. Искусственный интеллект (методология, применения, философия). – 2- е изд., испр. – М.: ЛЕНАНД, 2021. – 468 с.

[Abualigah, 2025] Abualigah L. Enhancing Real-Time Data Analysis through Advanced Machine Learning and Data Analytics Algorithms // International Journal of Online and Biomedical Engineering (iJOE). – 2025. – Vol. 21, No. 1. – P. 4-25.

[Alwosheel, 2018] Alwosheel A., van Cranenburgh S., Chorus C.G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis // Journal of Choice Modelling. – 2018. – Vol. 28. – P. 167-182.

[Ballantyne, 2023] Ballantyne P., Berragan C. Overture poi data for the united kingdom: a comprehensive, queryable open data product // A PREPRINT. – Liverpool: Geographic Data Science Lab, University of Liverpool, 2023.

[Conant, 2015] Conant D.D. Developing a Scalable Model to Analyze Expanding Data Sets // INFORMS Transactions on Education. – 2015. – Vol. 15, No. 3. – P. 215-223.

[Cordelli, 2023] Cordelli E., Guarrasi V., Iannello G., Ruffini F., Sicilia R., Soda P., Tronchin L. Making AI trustworthy in multimodal and healthcare scenarios // CEUR Workshop Proceedings. – 2023. – Vol. 3265. – P. 1-18.

[Corney, 2002] Corney D.P. A. Intelligent analysis of small data sets for food design: PhD: Computer Science / David Peter Alfred Corney; University College London. – London, 2002. – 168 p. – URL: ProQuest Dissertations.

[Faraway, 2017] Faraway J., Augustin N. When small data beats big data // Department of Mathematical Sciences, University of Bath. – Bath, 2017.

[Frontoni, 2022] Frontoni E., Paolanti M., Lauriault T.P., Stiber M., Duranti L., Abdul-Mageed M. Trusted Data Forever: Is AI the Answer? // Proceedings of the 25th EDBT. – 2022. – P. 1-12.

[Hollmann, 2025] Hollmann N., Müller S., Purucker L., Krishnakumar A., Körfer M., Hoo S.B., Schirrmeister R.T., Hutter F. Accurate predictions on small data with a tabular foundation model // Nature. – 2025. – Vol. 637, No. 7979. – P. 319-326.

[Jiao, 2024] Jiao L., Wang Y., Liu X., Li L., Liu F., Ma W., Guo Y., Chen P., Yang S., Hou B. Causal Inference Meets Deep Learning: A Comprehensive Survey // Research. – 2024. – Vol. 7. – Article 0467.

[Kimpimäki, 2023] Kimpimäki J.-P. From observation to insight: Computational abduction and its applications in sustainable strategy research: dissertation. – Lappeenranta-Lahti University of Technology LUT, 2023. – 116 p. – ISBN 978-952-412-004-3.

[de Miguel Beriain, 2022] de Miguel Beriain I., Nicolás Jiménez P., Rementería M.J., Cirillo D., Cortés A., Saby D., Lazcoz Moratinos G. Auditing the quality of datasets used in algorithmic decision-making systems // Panel for the STOA, European Parliamentary Research Service. – Brussels, 2022. – 41 p.

[Mendes, 2020] Mendes P.S.F., Siradze S., Pirro L., Thybaut J.W. Open data in catalysis: from today's big picture to the future of small data // ChemCatChem. – 2020.

[Mikołajewski, 2023] Mikołajewski D., Mikołajewska E. Artificial intelligence-based analysis of small data sets in medicine // Studia i Materiały Informatyki Stosowanej. – 2023. – Vol. 15, No. 2. – P. 18-23.

[Nisheva-Pavlova, 2022] Nisheva-Pavlova M., Dobreva B. Small Data and Data Centric AI: Case Study from the Master's Program in Artificial Intelligence at Sofia University // Proceedings of the Fifteenth International Conference on Information Systems and Grid Technologies (ISGT'2022). Sofia, Bulgaria, May 27–28, 2022. CEUR Workshop Proceedings. Vol. 3154. – P. 171-179.

[Nivedhaa, 2024] Nivedhaa N. A comprehensive review of AI's dependence on data // International Journal of Artificial Intelligence and Data Science (IJADS). – 2024. – Vol. 1, No. 1. – P. 1-11.

[Noroozi, 2023] Noroozi G. Data Heterogeneity and Its Implications for Fairness: MSc: Computer Science. Western University. – Ontario, 2023.

[Pajić, 2023] Pajić N., Djapan M., Buluschek E., Fahrenbruch W., Đorđević A., Stefanović M. Machine Learning Prediction Model for Small Data Sets Instead of Destructive Tests for a Case of Resistance Brazing Process Verification // IJIEM. – 2023. – Vol. 30, No. 3. – P. 797-814.

[Piffer, 2024] Piffer S., Ubaldi L., Tangaro S., Retiko A., Talamonti H. Tackling the small data problem in medical image classification with artificial intelligence: a systematic review // Progress in Biomedical Engineering. – 2024. – Vol. 6, No. 032001.

[Pasini, 2015] Pasini A. Artificial neural networks for small dataset analysis // Journal of Thoracic Disease. – 2015. – Vol. 7, No. 5. – P. 953-960.

[Safonova, 2023] Safonova A., Ghazaryan G., Stiller S., Main-Knorn M., Nendel C., Ryo M. Ten deep learning techniques to address small data problems with remote sensing // International Journal of Applied Earth Observation and Geoinformation. – 2023. – Vol. 125. – Art. 103569.

[Wang, 2023] Wang H., Duentsch I., Guo G., Khan S.A. Special issue on small data analytics // International Journal of Machine Learning and Cybernetics. – 2023. – Vol. 14, No. 1. – P. 1-2.

УДК 004.85
doi: 10.15622/rcai.2025.011

РЕШЁТОЧНОЕ ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

Д.В. Виноградов (*vinogradov.d.w@gmail.com*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

Работа посвящена изложению теоретико-решеточного подхода к порождению правил, образующих (суб-)оптимальную стратегию в парадигме обучения с подкреплением. Мы аргументируем за возврат к использованию метода Монте-Карло с поиском по дереву. Вероятностно-комбинаторный формальный метод, основанный на теории решеток, устраняет основной недостаток метода Монте-Карло – отсутствие обобщающей способности. Будут обсуждаться проблемы широко применяемых сейчас нейросетевых подходов и указаны достоинства метода Монте-Карло. Наконец, с использованием теории категорий будет представлена строгая формализация предлагаемого подхода.

Ключевые слова: обучение с подкреплением, метод Монте-Карло, решётки, теория категорий.

Введение

Обучение с подкреплением (см., например, [Саттон и др., 2020]) является одной из парадигм машинного обучения, нацеленной на порождение (суб-)оптимальной стратегии поведения агента (агентов в многоагентной постановке) в вероятностно изменяемой среде на основании наград (и наказаний). Стандартной математической моделью для обучения с подкреплением является Марковский процесс принятия решений – Markov Decision Process. В зависимости от выбора действий агента и случайного ответного поведения среды возникает дерево возможных траекторий.

Основными проблемами для возможности эффективного нахождения (суб-)оптимальной стратегии являются

1. Накопительный характер оценок качества действий в промежуточных состояниях, когда окончательная оценка возникает только по окончании всей траектории.
2. Экспоненциально большое число состояний в дереве траекторий.

С 2015 года в обучении с подкреплением доминирующим является нейросетевой подход, когда нейросеть того или иного вида аппроксимирует или оценки действий (с последующим применением жадного алгоритма выбора действия в наблюдаемом состоянии), или даже напрямую функцию выбора действий.

Ранее для оценок качества действий обучающегося агента использовалась классическая техника Монте-Карло, основанная на поиске в дереве (МКПД) – Monte Carlo Tree Search [Świechowski et al., 2023]. В ней для оценки значения действия в текущем состоянии мы вычисляем награды агента посредством усреднения вознаграждений для некоторого (статистически значимого) количества частичных траекторий, запущенных из текущего состояния с помощью выбранного действия.

Смена парадигмы с МКПД на глубокие нейросети произошла в 2014 году под влиянием впечатляющих успехов систем AlphaGo и AlphaStar фирмы DeepMind.

Теоретическим обоснованием этого перехода является отсутствие у МКПД обобщающей способности, когда для оценки ранее ненаблюдавшегося действия необходимо заново запускать поиск по дереву. В то же время нейросети, как универсальные аппроксиматоры, всегда обеспечивают оценку (возможно, очень плохую) для любого действия!

Однако МКПД имеет и свои неоспоримые достоинства:

1. Он обеспечивает контролируемый уровень дисперсии оценки в каждой оцениваемой точке, а не только оптимизацию единственной глобальной функции потерь, используемой для обучения нейросети.
2. Он работает напрямую с состояниями и действиями агента, а не с их векторными представлениями в скрытых слоях нейросети.

Для добавления обобщающей способности к МКПД предлагается использовать вероятностно-комбинаторный формальный метод (ВКФ-метод) (см., например, [Виноградов, 2022]).

Нейросетевой подход, доминирующий в настоящее время, не лишён существенных недостатков, среди которых следует упомянуть:

1. Невозможность предварительно учитывать правила, сформулированные экспертами.
2. Трудность передачи знаний от одной модели другой.
3. Плохая объясняемость принимаемых решений.
4. Наличие “галлюцинаций” нейросетей.

У любого варианта обучения с подкреплением имеются известные трудности:

1. Нестационарность распределения ответов среды (она может изменять вероятности переходов с течением времени).
2. Возможная экспоненциальная малость числа траекторий с большими наградами.

3. Нахождение баланса между исследованием качества выбранного действия и поиском более успешного альтернативного действия в текущем состоянии.

Наиболее ярко нестационарность проявляется при обучении стратегии игры с полной информацией, когда вероятности перехода в следующее состояние игры определяются как выбором действия учащегося, так и его оппонента. Чтобы справиться с первой трудностью в обучении с подкреплением обычно предполагают, что время изменения распределения ответов оппонента значительно превосходит время перемешивания цепей Маркова, соответствующих применяемым алгоритмам. Мы тоже будем делать это допущение.

Для учёта второй проблемы переходят на точку зрения вероятно-приближенно корректного (ВПК-) обучения – PAC-learning – Л. Вэльянта [Valiant, 1984].

Наконец, третью трудность мы исключим из обсуждения (хотя имеются некоторые теоретические результаты), ограничившись случаем автономного обучения, когда этапы обучения и применения обученной стратегии разнесены во времени.

1. Элементы теории обучения с подкреплением

В обучении с подкреплением (ОсП) обучающийся, называемый *агентом*, взаимодействует с окружением. *Окружение* может находиться в одном из своих состояний $s \in S$, наблюдаемых игроком. В каждом из состояний агент может совершить одно из нескольких *действий* $a \in A$, зависящих от текущего состояния.

Например, для игры состояниями окружения являются только вершины дерева игры, соответствующие состояниям, в которых решение принимает обучающийся.

Так как поведение среды стохастично (например, для игры оппонент может использовать рэндомизованную стратегию), то переход к новому наблюдаемому состоянию имеет условное вероятностное распределение $P_a(s, s') = \mathbf{P}[S_{t+1} = s' | S_t = s, A_t = a]$ оказаться в состоянии s' при выборе (допустимого) действия a в состоянии s. Однако это распределение нам неизвестно. Более того, указанная выше трудность (1) говорит о том, что это распределение может быть нестационарным. В дальнейшем мы будем явно предполагать его стационарность.

В ОсП агент получает *непосредственное вознаграждение* $r_a(s, s') \in \mathbf{R}^1$ за попадание в состояние s' при выборе действия a в состоянии s. Конечно, вознаграждение может быть отрицательным: правильно назвать его в этом случае *наказанием*. В играх вознаграждение обычно получается только в финальных состояниях. Это условие упрощает как формулы, так и вычисления.

Целью агента в ОсП является выбор оптимальной политики $\pi: S \to \Delta^A$ – вероятностной стратегии выбора действий для каждого состояния, в котором он может находиться. Оптимальность вычисляется относительно (вообще говоря, дисконтированной со скоростью дисконта $\gamma > 0$) накопленной награды

$$V^\pi(s) = \mathbf{E}[\Sigma\, \gamma^t \cdot \Pi\, r_a(s', s'') \cdot \pi(a|\, s') \cdot \mathbf{P}[S_{j+1} = s''|\, S_j = s', A_j = a] \mid S_0 = s].$$

Дисконтирование вводилось для доказательства существования (даже детерминированной) оптимальной стратегии посредством неподвижной точки сжимающего оператора. Переходом $\gamma \to 1$ существование детерминированной оптимальной стратегии доказывается и в этом предельном случае. В дальнейшем, мы ограничимся случаем без дисконта ($\gamma = 1$). Суммирование идет по всем траекториям $s_0, \ldots, s_n, s_{n+1}, \ldots$ до остановки и всем последовательностям действий $a_0, \ldots, a_n, a_{n+1}$, умножение – по j от 0 до t - 1.

Для выбора оптимального действия в обучении с подкреплением рассматривают функцию Q-values

$$Q^\pi(s_0, a) = \mathbf{E}[\Sigma\Pi\, r_a(s', s'') \cdot \pi(a'|\, s') \cdot \mathbf{P}[S_{j+1} = s''|\, S_j = s', A_j = a'] \cdot \mathbf{P}_a(s_0, s_1)|\, S_0 = s_0, A_0 = a].$$

Она соответствует выбору действия a в начальном состоянии s_0, а затем применению политики π. Здесь умножение идет по j от 1 до t - 1.

Неизвестность распределения вероятностей переходов заставляет заменять точное значение функции $Q^\pi(s_0, a)$ на ее приближение. Первоначально применялся метод Монте-Карло, после 2015 года – аппроксимация нейросетью.

2. Категория решёток и сходство

Мы лишь напомним ключевые понятия про категорию полных полурешеток. Для более подробного знакомства читатель отсылается к статье [Виноградов, 2021].

Категория **Set** множеств и отображений между ними допускает эндофунктор $2\hat{}: \mathbf{Set} \to \mathbf{Set}$, который отображает множество X во множество-степень $2\hat{}X = \{A: A \subseteq X\}$, а отображение $f: X \to Y$ в отображение $2\hat{}f: 2\hat{}X \to 2\hat{}Y$, где $2\hat{}f(A) = \{f(x): x \in A\} \subseteq Y$, для любого $A \subseteq X$.

Имеется естественное преобразование $\cup: 2\hat{}\cdot 2\hat{} \to 2\hat{}$, которое отображает каждое семейство $S \subseteq 2\hat{}X$ подмножеств множества X в их объединение $\cup S = \cup\{A: A \in S\} \in 2\hat{}X$.

Существует естественное преобразование $\{\}: I_{\mathbf{Set}} \to 2\hat{}$ тождественного функтора $I_{\mathbf{Set}}$ в функтор $2\hat{}$, которое отображают каждый $x \in X$ в одноэлементное подмножество $\{x\} \in 2\hat{}X$.

Тройка $\langle 2\hat{}, \{\}, \cup\rangle$ задаёт монаду в категории **Set**, для которой можно определить категорию полных (полу-)решёток **Lat** как множество пар $\langle L, \wedge\rangle$, где объект (множество) L называется *носителем решётки*, а морфизм $\wedge: 2\hat{}L \to L$ называется *сходством*, причём должны выполняться тождества $\wedge\cdot\cup = \wedge\cdot 2\hat{}\wedge$ и $\wedge\cdot\{\} = \mathrm{id}$.

Лемма 1 из статьи [Виноградов, 2021] утверждает, что элементами этой категории являются полные полурешётки и только они.

Эта категория алгебр допускает свободные алгебры. Они соответствуют образу сопряженного функтора G: **Set** → **Lat** к забывающему функтору F: **Lat** → **Set**, который сопоставляет полной полурешётке $\langle L,\wedge\rangle$ ее носитель L.

Теорема 1 из работы [Виноградов, 2021] доставляет явную конструкцию свободной алгебры над множеством X: ее носителем является множество-степень 2^X, а сходством является объединение $\cup$: $2^X \cdot 2^X \to 2^X$.

Свободность образа G: **Set** → **Lat** (сопряженность с функтором F: **Lat** → **Set**) означает наличие биекции между отображениями множеств $S \to L = F\langle L,\wedge\rangle$ и отображениями полурешеток $G(S) = \langle 2^S,\cup\rangle \to \langle L,\wedge\rangle$.

Этот факт имеет два важных следствия! Во-первых, при работе с решётками важно учитывать множества родителей сходства, а не только структурное описание. На этой идее базируется Анализ формальных понятий (АФП) [Ganter et al., 1999] – современный раздел теории решёток. Во-вторых, если имеется некоторое описание объектов (в нашем случае состояний среды обучения с подкреплением), на котором задана операция сходства, то она может быть распространена на всё множество-степень 2^S.

3. Общая схема решёточного обучения с подкреплением

Из фундаментальной теоремы АФП следует, что любую полную решётку можно породить как сходства из списка битовых строк – формального контекста.

Чтобы применить ВКФ-метод обучения, мы будем предполагать, что состояния игры допускают кодирование бинарными признаками f_j (j =1, …, n) так, что сходство между состояниями представляется побитовым умножением. Так как побитовое умножение является базовой операцией современных процессоров, это приводит к значительному ускорению вычислений, которое осуществляется компилятором.

Группируя оценки действий в разных состояниях, но с одинаковым действием, получаем набор обучающих выборок, как представлено в табл. 1 ниже.

Таблица 1

	a_1	a_2	…	a_k	Q	f_1	…	f_n
s_1	1	0	…	0	$Q(s_1, a_1)$	$\delta_{1,1}$	…	$\delta_{1,n}$
…	…		…	…	…	…	…	…
s_m	1	0	…	0	$Q(s_m, a_1)$	$\delta_{m,1}$	…	$\delta_{m,n}$
s_1	0	1	…	0	$Q(s_1, a_2)$	$\delta_{1,1}$	…	$\delta_{1,n}$
…	…		…	…	…	…	…	…
s_m	0		…	1	$Q(s_m, a_k)$	$\delta_{m,1}$	…	$\delta_{m,n}$

Для фиксированного действия (например, a_1) обучающими примерами будут те строчки, где $Q(s_j, a_1) \geq 0$, а контр-примерами – дополнительные (где $Q(s_j, a_1) < 0$). На битовых строчках $\delta_{i,1} \ldots \delta_{i,n}$, и $\delta_{j,1} \ldots \delta_{j,n}$, соответствующих обучающим примерам, есть бинарная операция сходства – побитовое умножение.

При адекватном кодировании состояний битовыми строками $\delta_{i,1} \ldots \delta_{i,n}$ можно надеяться, что общие признаки будут определять обобщение конкретных ситуаций (обучающих примеров), в которых применение выбранного действия приводит к большой награде. Контр-примеры необходимы для уменьшения числа сходств – устранения переобучения. Поэтому присоединение ВКФ-метода, вероятностно порождающего такие сходства, может превратить МКПД в хорошую альтернативу нейросетевым методам.

Заключение

В настоящей работе представлено формальное описание теоретико-решеточного подхода к обучению с подкреплением.

Возможность применения вероятностно-комбинаторного формального метода к задаче обобщения результатов оценивания методом Монте-Карло с поиском по дереву состояний позволяет надеяться на возрождение использования классического метода МКПД для ОсП.

Для апробации предложенного подхода автор запрограммировал на языке C++23 с использованием библиотеки oneTBB 2022.1 прототип системы ОсП для игр двух лиц с полной информацией Noughts&Crosses и нескольких вариантов игры Nim. На первом этапе существенную помощь ему оказала его аспирантка в ФИЦ ИУ РАН Л.А. Якимова. В дальнейшем планируется сравнение с известными нейросетевыми алгоритмами PPO [Schulman et al., 2017] от OpenAI и Deep Graph Network RL [Munikoti et al., 2022] при сотрудничестве с магистрантом МФТИ А.С. Мисником.

Благодарности. Идея написать настоящую работу возникла у автора в результате обсуждения со студентом МФТИ А.С. Мисником результатов совместной работы автора со своей бывшей аспиранткой ФИЦ ИУ РАН Л.А. Якимовой [Виноградов и др., 2024]. Им обоим автор выражает свою благодарность за сотрудничество и интерес к его работе. Автор считает своим приятным долгом поблагодарить своих коллег по ВЦ им. А.А. Дородницына ФИЦ ИУ РАН за поддержку и конструктивные дискуссии.

Список литературы

[Виноградов, 2021] Виноградов Д.В. Проекции полурешеток: язык теории категорий // Научно-техническая информация. Серия 2. – 2021. – № 6. – С. 27-31.

[Виноградов, 2022] Виноградов Д.В. Алгебраическое машинное обучение: упор на эффективность // Автоматика и телемеханика. – 2022. – № 6. – С. 5-23.

[Виноградов и др., 2024] Виноградов Д.В., Якимова Л.А. Вероятностный подход к «доброму старомодному» искусственному интеллекту // Научно-техническая информация. Серия 2. – 2024. – № 3. – С. 21-26.

[Саттон и др., 2020] Саттон Р.С., Барто Э.Дж. Обучение с подкреплением. – М.: ДМК Пресс, 2020. – 552 с.

[Ganter et al., 1999] Ganter B., Wille R. Formal Concept Analysis. – Berlin: Springer, 1999.

[**Munikoti et al., 2022**] Munikoti S., Agarwal D., Das L., Halappanavar M., Natarajan B. Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications // arXiv preprint. – 2022. – doi: 10.48550/arXiv.2206.07922.

[Schulman et al., 2017] Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms // arXiv preprint. – 2017. – doi: 10.48550/arXiv.1707.06347.

[Świechowski et al., 2023] Świechowski M., Godlewski K., Sawicki B., Mańdziuk J. Monte Carlo Tree Search: a review of recent modifications and applications // Artificial Intelligence Review. – 2023. – Vol. 56, – P. 2497-2562.

[Valiant, 1984] Valiant L.G. An theory of the learnable // Communications of the ACM, – 1984. – Vol. 27, No. 11. – P. 1134-1142.

УДК 004.89
doi: 10.15622/rcai.2025.012

ОБ ЭВРИСТИЧЕСКОМ ПОТЕНЦИАЛЕ НЕКОТОРЫХ ПОЗНАВАТЕЛЬНЫХ ПРОЦЕДУР

М.А. Михеенкова (*m.mikheyenkova@yandex.ru*)
С.М. Гусакова (*svem45@yandex.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В работе охарактеризованы проблемы формализации исследовательских эвристик для решения задач в открытых эмпирических областях, лишённых систематического формального аппарата. Рассматриваются особенности применения некоторых познавательных процедур ДСМ-метода автоматизированной поддержки исследований в соответствии с семантической и прагматической спецификой конкретной исследовательской ситуации. Приводятся примеры адекватного использования эвристического потенциала формальных средств для получения интерпретируемых результатов в нескольких предметных областях.

Ключевые слова: точная эпистемология, формализованная эвристика, познавательные процедуры, ДСМ-метод, операция сходства, метод различия, ситуационный подход.

Введение

Эффективность методов ИИ и создаваемых на их основе интеллектуальных систем связаны в широком смысле с исследованиями в двух направлениях: эпистемологическом и эвристическом [McCarthy, 2014]. Первое решает задачи создания языка представления эмпирических фактов, формирования системы знаний и организации баз фактов и баз знаний. В свою очередь внимание к теоретическим (эпистемологическим) основаниям ИИ служит залогом успешного развития второго (эвристического) направления: создания формальных подходов к конструированию эвристик (систем правил с использованием правдоподобных рассуждений [Пойа, 1975]) для порождения нового знания и его принятия. Иными словами, можно утверждать, что идейным основанием ИИ как научной и прикладной области является точная эпистемология – метатеоретическая

дисциплина, исследующая взаимодействие познающего субъекта и соответствующего объекта познания конструктивными средствами – формализованными эвристиками и логиками рассуждений [Финн, 2023, с. 27-98]. Основным конечным продуктом ИИ являются интеллектуальные системы (ИС), реализующие формальные эвристики (сформированные на основе разработанных эпистемологических принципов) и тем самым обеспечивающие поддержку эмпирических исследований. Более того: в плохо формализованных областях, лишённых систематического формального аппарата, ИС в значительной степени создают условия для организации собственно познавательного процесса и объективизации его результатов – возможно, лишь, частичной[1]. Учитывая характерные особенности таких областей, здесь не приходится говорить, однако, о возможности полного воплощения семантических особенностей структуры данных и знаний в синтаксических конструкциях (см., например, [Гусакова и др., 2023]). Соответственно, в реальной исследовательской ситуации приходится привлекать также и прагматические соображения целесообразности применения тех или иных процедурных решений при реализации формальных эвристик.

1. Эвристика синтеза познавательных процедур

Научные принципы точной эпистемологии конструктивно и достаточно полно реализованы в ДСМ-методе автоматизированной поддержки исследований[2], представляющем собой полноценную теоретическую и практическую методологию создания средств формализации, имитации и усиления познавательной деятельности и соответствующей реализации в интеллектуальных системах (ИС-ДСМ). Методологическим основанием ДСМ-метода является конструктивная эволюционная эпистемология [Финн, 2024b], расширяющая и уточняющая схему роста знания К. Поппера [Поппер, 2019, с. 12-48].

Формализованная эвристика порождения нового знания в ДСМ-методе представляет собой эвристику синтеза неэлементарных познавательных процедур: эмпирической индукции, структурной аналогии и абдуктивного объяснительного принятия гипотез. Взаимодействие таких процедур соответствует традиционному познавательному процессу, представленному исследовательской эвристикой «анализ данных – порождение гипотез – предсказание – объяснение результатов – принятие результатов». Аналитическая часть, формирующая основание для порождения гипотез о кау-

[1] Частичный характер объективизации связан с неизбежной субъективностью эмпирических данных в плохо формализованных областях.

[2] Изложение формальных средств ДСМ-метода и примеров решения задач в различных предметных областях средствами ИС-ДСМ представлены в многочисленных публикациях (см., например, [Автоматическое порождение…, 2020], [Финн, 2023] и др.).

зальных зависимостях в данных, обеспечивается формальным представлением уточнений и расширений индуктивных канонов Д.С. Милля [Милль, 2022].

Согласно онтологическим допущениям ДСМ-метода, в исходной базе фактов (БФ) должны быть представлены примеры как наличия имеющихся эффектов у рассматриваемых объектов ((+)-факты), так и их отсутствия ((–)-факты)[3], а в описание данных должны быть включены значимые факторы влияний ((+)- и (–)-причины). В логическом языке метода БФ = $БФ^{+} \cup БФ^{-} \cup БФ^{\tau}$, где $БФ^{\tau} = \{\langle X, Y\rangle \mid J_{\langle \tau, 0\rangle}(X \Rightarrow_1 Y)\}$, $БФ^{\sigma} = \{\langle X, Y\rangle \mid J_{\langle \nu, 0\rangle}(X \Rightarrow_1 Y)\}$ ($\nu = 1$ для $\sigma = +$, $\nu = -1$ для $\sigma = -$), $\nu \in \{1, -1, 0, \tau\}$ – типы внутренних истинностных значений для представления фактов и гипотез о фактах – фактические истина, ложь, противоречивость и неопределенность, соответственно. Для представления фактов с оценками (высказываний о фактах) используются внешние истинностные значения *t, f* – логические истина и ложь. *J*-операторы определяются следующим образом: $J_\nu\varphi = t$, если оценка $\nu[\varphi] = \nu$, $J_\nu\varphi = f$, если оценка $\nu[\varphi] \neq \nu$. К примеру, $J_{\langle 1, 0\rangle}(X \Rightarrow_1 Y)$ интерпретируется как «высказывание “объект *X* является примером эффекта *Y*” в начальном состоянии БФ имеет истинностную оценку 1». При этом как *X*, так и *Y* могут/должны иметь структурированное описание.

Наименьшими базисными предикатами для правил индуктивного вывода (для порождения возможных гипотез о причинах) являются предикаты $M^{\sigma}_{a,n}(V, W)$ ($\sigma = +, –$; *n* – число применений правил правдоподобного вывода к БФ, «*a*» – от «agreement» – «сходство» у Д.С. Милля), где V (алгебраически определённое сходство объектов *X*) – потенциальная причина *W*. Добавление дополнительных условий – запрета на контрпримеры $(b)^{\sigma}$, различия $(d_0)^{\sigma}$ и т.д. – порождает множество предикатов $M^{+}_{x,n}(V, W), M^{-}_{y,n}(V, W)$, $I^{\sigma} = \{a^{\sigma}, (ab)^{\sigma}, (ad_0)^{\sigma}, (ad_0b)^{\sigma}\}$, $x \in I^{+}, y \in I^{-}$; $M^{\sigma}_{ab,n}(V, W) \rightleftharpoons M^{\sigma}_{a,n}(V, W)$ & $(b)^{\sigma}$ и т.д., частично упорядоченное по отношению логической выводимости ([Финн, 2023, с. 381-436]), что обеспечивает возможность вариативного выбора адекватных эвристик в зависимости от прагматической ориентации исследования.

Индуктивный анализ средствами ДСМ-метода может осуществляться в рамках двух разнонаправленных стратегий: прямой («от причины – к следствию») и обратной («от следствия – к причине»). Предикаты прямого $\widetilde{M}^{\sigma}_{a,n}(V, W, k)$ и обратного $\widetilde{\widetilde{M}}^{\sigma}_{a,n}(V, W, k)$ методов сходства ($\sigma = +, –$), использующиеся при формулировке индуктивных правил правдоподобно-

[3] Ср. с известной идеей контр-фактического вывода в причинном анализе [Pearl et al., 2018], фактически представляющей собой интерпретацию хорошо известного в теории познания принципа фальсифицируемости – критерия демаркации К.Р. Поппера [Поппер, 2005], отделяющего научное знание от ненаучного.

го вывода, имеют одинаковую структуру. Они распознают локальное сходство на множестве примеров $J_{(\nu,\, n)}(C_i \Rightarrow_1 Q_i)$ (i = 1, ..., k, $k \geq 2$, ν = 1 для σ = +, ν = –1 для σ = –), которое является основанием для правдоподобного вывода о причинах рассматриваемого явления. Предикат прямого сходства $\widetilde{M}^{+}_{a,n}(V,\ W,\ k)$ описывает для (+)-примеров эмпирическую зависимость $\forall X \forall Y((J_{(1,\, n)}(X \Rightarrow_1 Y) \& \ \forall U(J_{(1,\, n)}(X \Rightarrow_1 U) \rightarrow U \subseteq Y) \& V \subset X) \rightarrow (W \subseteq Y \& W \neq \varnothing))$, содержательно интерпретируемую как «сходство V (подобъект) объектов X в (+)-примерах есть причина наличия свойств W, общих для объектов X» (для (–)-примеров – аналогично). Подформула $(\bigvee_{i=1}^{k}(X = X_i))$ предиката описывает так называемое «условие исчерпываемости»: требование рассмотрения всех элементов БФ, содержащих V.

Эмпирическая зависимость в предикате $\widetilde{\widetilde{M}}^{+}_{a,n}(V,\ W,\ k)$ обратного сходства представлена подформулой $\forall X \forall Y((J_{(1,\, n)}(X \Rightarrow_1 Y) \& \forall U(J_{(1,\, n)}(X \Rightarrow_1 U) \rightarrow U \subseteq Y) \& W \subseteq Y) \rightarrow (V \subset X \& V \neq \varnothing))$, которая интерпретируется как «сходство свойств W объектов X в (+)-примерах есть следствие сходства V самих объектов X». Условие исчерпываемости $(\bigvee_{i=1}^{k}(Y = Y_i))$ требует рассмотрения всех элементов БФ, содержащих W. В правила индуктивного вывода для прямого метода входят непараметрические предикаты $M^{\sigma}_{a,n}(V,W) \rightleftharpoons \exists k\ \widetilde{M}^{\sigma}_{a,n}(V,W,k)$, для обратного – $\widetilde{M}^{\sigma}_{a,n}(V,W) \rightleftharpoons \exists k\ \widetilde{\widetilde{M}}^{+}_{a,n}(V,W,k)$.

Несмотря на структурное единство предикатов сходства прямого и обратного метода, семантика каузального отношения приводит к несимметричности интерпретаций порождаемых эмпирических зависимостей. Кроме того, различные условия исчерпываемости этих методов обуславливают разницу их прагматического значения. Каузальное отношение – следствие выявленной эмпирической зависимости – в прямом методе является функциональным (см. [Забежайло, 2013]): если выполняется предикат $M^{+}_{a,n}(V,\ W)$, то имеет место $\forall U M^{+}_{a,n}(V,\ U) \rightarrow (U = W)$. Одновременно может выполняться также предикат $M^{+}_{a,n}(V',\ W)$, где $V \cap V' = \varnothing$. Это означает множественность причин для W в представленной БФ, т.е. каждая причина является достаточной, но не необходимой. Иными словами, возможны различные механизмы проявления свойств (эффектов) W. Эти особенности прямого метода могут быть использованы, к примеру, при решении задачи построения типологии социума относительно ядра W [Михеенкова, 2024]. Типологизация заключается в том, что личностные свойства, формирующие класс групповых признаков (сходство в ДСМ-методе – общие характеристики V индивидуумов), рассматриваются как идеально-типические свойства. Множественный характер типических оснований V в этом случае оказывается источником построения содержательных социологических теорий.

Особенностью применения обратного метода является единственность выявляемой эмпирической зависимости: каждое W в $\breve{M}^{+}_{a,n}(V,W)$ есть следствие единственного V, т.е. если выполняется предикат $\breve{M}^{+}_{a,n}(V,W)$, то имеет место $\forall Z\breve{M}^{+}_{a,n}(Z,W)\rightarrow(Z=V)$, иными словами, выполняется условие, совпадающее с условием единственности причины в прямом ДСМ-методе [Гусакова и др., 2016], которое имеет вид $\forall ZM^{+}_{a,n}(Z,W)\rightarrow(Z=V)$ (следствие W имеет единственную причину V). Это условие, обозначаемое $(s)^{+}$, – одно из так называемых «усилений» предиката сходства, позволяющее сформулировать предикат единственного сходства $M^{+}_{as,n}(V,\ W) \rightleftharpoons M^{+}_{a,n}(V,\ W)\&(s)^{+}$, $M^{-}_{as,n}(V,\ W)$ аналогично. Отметим, однако, что выполнимость предиката $\breve{M}^{+}_{a,n}(V,W)$ не обязательно означает выполнимость и предиката $M^{+}_{a,n}(V,\ W)$ для рассматриваемой БФ и, соответственно, выполнимость $M^{+}_{as,n}(V,\ W)$.

Обратный метод допускает одновременное выполнение предикатов $\breve{M}^{+}_{a,n}(V,W)$ и $\breve{M}^{+}_{a,n}(V,W')$ таких, что $W\cap W'=\varnothing$: W' также является следствием именно V, т.е. выявленные зависимости не являются функциональными и не могут служить основанием для построения типологии. Иными словами, прямой и обратный метод имеют различное значение с точки зрения представления эвристики порождения нового знания для конкретной исследовательской проблемы.

2. Выбор эвристических подходов

Увеличение степени правдоподобия индуктивных гипотез о причинах явлений (эффектов, свойств) и доопределяемых с их помощью гипотез о свойствах объектов повышает доверие к результатам работы ИС. В ДСМ-методе выбор познавательных процедур обусловлен наличием различных типов каузальных вынуждений и опирается как на специфику предлагаемых формализмов, так и на метатеоретические исследования предметных областей.

Теоретическим основанием усиления гипотез (повышения степени их правдоподобия) является формирование стратегий на основании дистрибутивной решётки индуктивных процедур ([Финн, 2023, с. 381-436]). Одним из эффективных способов усиления предикатов сходства является условие запрета на контрпримеры, в соответствии с которым гипотеза $J_{(\nu,n)}(V \Rightarrow_2 W)$ фальсифицируется, если в базе фактов существует хотя бы один пример $J_{(\mu,n)}(X\Rightarrow_1 Y)$ ($\mu\neq\nu$, $\mu, \nu\in\{1, -1, 0\}$) такой, что $V\subset X$ и $W\subseteq Y$ (для обратного метода аналогично). Добавление к предикату сходства $M^{\sigma}_{a,n}(V,\ W)$ так называемого условия различия d^{σ}_{0} (σ = +, –) формализует уточнение и расширение метода различия Д.С. Милля [Милль, 2022], $M^{\sigma}_{ad_0,n}(V,\ W) \rightleftharpoons M^{\sigma}_{a,n}(V,\ W)\&d^{\sigma}_{0}$.

$d_0^+ \rightleftharpoons X\forall Y\forall Z\forall U((J_{(1,n)}(X\Rightarrow_1 Y)\&(W\subseteq Y)\&(V\subset X))$ & $((X\backslash V)\subset Z)$ & $((X\backslash V)\neq\varnothing$ & $\neg\,(V\subset Z\,))) \rightarrow(\neg\, J_{(1,n)}(X\Rightarrow_1 U)\vee\neg\,(W\subseteq U))$ (это один из вариантов, другой можно найти в [Финн, 2024а]), d_0^- аналогично.

Содержательно это означает, что подобъект V, удовлетворяющий простому прямому предикату сходства, признается причиной W только тогда, когда в базе фактов не существует ни одного положительного примера $J_{(1,n)}(X_0\Rightarrow_1 Y_0)$ такого, что $(X\backslash V)\subset X_0$ и $W\subset Y_0$ (X – любой из объектов, содержащих V). Причина, удовлетворяющая предикату $M^+_{ad_0,n}$, имеет более высокую степень правдоподобия, чем полученная простым методом сходства, так как для ее нахождения учитывается и сходство объектов, и различие в проявлении свойств объектов, отличающихся только наличием или отсутствием V. Поскольку может существовать и другая причина проявления свойств W, отсутствие в объекте подобъекта V не позволяет заключить, что у этого объекта отсутствуют свойства W. Следовательно, в общем случае, наличие причины, полученной методом различия остается достаточным, но не необходимым условием выполнения свойств W.

Если причина найдена с помощью обратного метода простого сходства, усиливать ее условием различия не имеет смысла, так как эта причина получена как результат сходства исчерпывающего по W множества примеров, и для любого примера $J_{(1,n)}(Z\Rightarrow_1 U)$, не входящего в это множество, будет выполнено условие $\neg(W\subseteq U)$. Т.е. причина, полученная обратным методом, всегда удовлетворяет предикату различия.

Условие единственности причины при определенных условиях тоже может быть использовано как способ фальсификации гипотез. Это имеет смысл только в задачах, где существует единственный механизм, вынуждающий проявиться свойству W. В общем случае причина для W, полученная прямым методом сходства, может быть не единственной, так как исчерпываемость по объектам не означает исчерпываемости по свойствам, и примеры, содержащие W и не содержащие V, могут быть сходны по V' (см. выше) Если по содержательным соображениям к предикату простого прямого сходства добавлено условие единственности причины, а найдено две причины, они должны быть фальсифицированы. Но одна из них может быть реальной причиной. Здесь может помочь метод различия. Для каждой причины должно быть проверено условие различия. Очевидно, что степень доверия к той причине, которая удовлетворяет условию различия, выше, чем к не удовлетворяющей этому условию. В этом случае последняя должна быть фальсифицирована, а оставшаяся признана единственной причиной для свойства W.

В том случае, когда для W существует единственная причина V, а из V следует только W, наличие подобъекта V в любом объекте X является необходимым и достаточным условием проявления свойства W.

Поскольку в этом случае результат прямого и обратного методов совпадает, условие различия выполняется автоматически. Степень доверия к такой гипотезе существенно повышается.

Описанные выше сценарии одновременного выполнения предикатов $M^{+}_{a,n}(V,\ W)$ и $M^{+}_{a,n}(V',\ W)$, где $V \cap V' \neq \varnothing$, а также $\widetilde{M}^{+}_{a,n}(V,W)$ и $\widetilde{M}^{+}_{a,n}(V,W')$ таких, что $W \cap W' \neq \varnothing$, заставляют задуматься о том, что проявление эффектов может детерминироваться не только внутренней структурой самих объектов (респондентов), но и внешними влияниями – контекстом проявления эффекта. Для анализа такого типа каузальности было предложено расширение формального языка ДСМ-метода: введение переменных для ситуационных (контекстных, внешних по отношению к объекту) параметров S и представление исходной БФ в виде множеств $\text{БФ}^{\tau} = \{\langle X,\ Y,\ S\rangle \mid J_{(\tau,\ 0)}P(X,Y,S)\}$, $\text{БФ}^{\sigma} = \{\langle X,\ Y,\ S\rangle \mid J_{\langle \nu,\ 0\rangle}P(X,Y,S)\}$ («объект X проявляет/не проявляет эффекты Y в ситуации S»; $\nu = 1$ для $\sigma = +$, $\nu = -1$ для $\sigma = -$). Соответственно, для индуктивного порождения гипотез о причинах вида $R_i(\langle V,S\rangle, W)$ – «пара ⟨подмножество характеристик объекта V и характеристики ситуации S⟩ есть причина проявления эффектов W» – формулируются предикаты сходства ${}_i M^{\sigma}_{a,\ n}$ $(V,\ W,\ S)$ (прямой и несколько иначе сформулированные два предиката обратного метода) [Автоматическое порождение…, с. 428-445]. Здесь индекс i характеризует природу причинности в рассматриваемой предметной области: зависимость/независимость свойств W объекта от его характеристик (структурированного описания) V и внешних условий S (ситуации).

Структура предиката ситуационного сходства ${}_i \widetilde{M}^{\sigma}_{a,\ n}(V,\ W,\ S,\ k)$ аналогична структуре предиката простого сходства $\widetilde{M}^{\sigma}_{a,n}(V,\ W,\ k)$: предикат обнаруживает локальное сходство объектов и ситуаций ($V = X_1 \cap \ldots \cap X_k$) $\&(S_0 = S_1 \cap \ldots \cap S_k)$ $\&V \neq \varnothing \& S_0 \neq \varnothing$ на множестве $k \geq 2$ примеров $(\&_{i=1}^{k}\ J_{(1,n)}P(X_i,Y_i,S_i))$ из БФ (здесь – для ${}_2\widetilde{M}^{+}_{a,\ n}(V,\ W,\ S,\ k)$) и порождает эмпирическую зависимость $\forall X \forall Y \forall S$ $((J_{(1,n)}P(X,Y,S) \& \forall U(J_{(1,n)}P(X,U,S) \rightarrow U \subseteq Y) \&$ $(V \subset X) \& (S_0 \subseteq S)) \rightarrow$ $(W \subseteq Y$ $\& W \neq \varnothing)))$ с условием исчерпываемости $(\vee_{i=1}^{k}(X = X_i))$.

Ситуационный подход оказался эффективным инструментом анализа социологических данных (примеры см. [Михеенкова, 2023]). Результаты первого этапа исследований с применением предиката ситуационного сходства для анализа трудовых отношений на двух российских предприятиях (А и Б) выявили интерес исследователей к формированию типологических единиц не только по отношению к той или иной стратегии отстаивания трудовых прав и сохранения лояльности предприятию, но и выборе таких стратегий в зависимости от различия положений, характеризующих условия труда и зарплатные ожидания (т.е. ситуацию) на этих

предприятиях. Соответственно, предикат ${}_2\widetilde{M}^{+}_{a,\,n}(V,\ W,\ S,\ k)$ был усилен условием $d^{+}_{0S} \rightleftharpoons \forall X \forall Y \forall S \forall S' \forall U((J_{(1,n)}P(X,Y,S) \& (W \subseteq Y) \& (V \subset X)\ \ \& (S_0 \subseteq S)\ \ \& ((S \backslash S_0) \subset S') \ \& \ ((S \backslash S_0) \neq \varnothing \ \& \ \neg\ (S_0 \subset S'\ ))) \rightarrow (\neg J_{(1,n)}P(X,\ U,\ S') \vee \neg\ (W \subseteq U)))$.

Оказалось, к примеру, что только на предприятии Б среди работников, не готовых отстаивать свои права, выделилась группа, полагающая, что их трудовые права не нарушаются. Среди тех, кто готов добиваться справедливости, только на предприятии Б оказались «нацеленные на заработок» и «нацеленные на карьеру». Выделенные типы – это номинальные группы людей, объединенные общими характеристиками. Сочетание характеристик, составляющих типы, позволяет понять проблемы этих людей и принимать соответствующие управленческие решения по отношению к каждой из выделенных групп применительно к конкретной ситуации.

Аналогичные эффекты влияния внеличностных факторов (внешних условий, контекста) на социальное поведение с учётом их различия были выявлены в исследованиях гражданского участия в малых и средних городах России, а также выбора индивидуальных стратегий гражданской активности (см. [Михеенкова, 2023]).

Вариант ситуационного расширения используется также в криминалистике при решении задачи выявления влияния психологических характеристик на особенности подписи [Гусакова, 2023]. Для описания психологических характеристик привлекаются несколько психологических опросников, так как в них отражены разные стороны психологии личности. Один опросник выбирается как основной, остальные как дополнительные параметры. В качестве эффектов рассматриваются признаки подписи респондента. Работа ДСМ-системы аналогична работе с ситуационным расширением в социологии. Этот метод позволяет выявить, какие стороны личности влияют на особенности подписи. В случае использования двух опросников – ОСТ (структура темперамента) и ОЧХ (черты характера) как дополнительного, была получена гипотеза: «высокие значения предметных шкал опросника структуры темперамента и низкие значения шкал «Экзальтированность» и «Тревожность» опросника черт характера, дают значение частных признаков подписи *Протяженность движений по вертикали* и *горизонтали – увеличена.* Причем это значение *устойчиво.* При таких же значениях предметных шкал, но высоком уровне экзальтированности и тревожности значения указанных признаков подписи вариативны.

В другом случае гипотеза с исчерпываемостью по чертам характера, не прошла проверку методом различия, что свидетельствует об отсутствии влияния выбранных черт характера на признаки подписи.

Таким образом использование дополнительного параметра позволило выявить черты характера, влияющие и не влияющие на почерк. Следует заметить, что при объединении шкал двух опросников в одно описание в

первом случае прямым методом гипотеза не была бы получена, а обратный метод дал бы два следствия одной причины. Влияние черт характера не было бы выявлено.

Отметим, что необходимость учёта контекста явлений, порой играющего решающую роль в их проявлении, характерна для множества областей: анализа естественного языка, компьютерного зрения, транспорта, здравоохранения, адаптивного поведения роботов, военного дела и т.д.

Приведённые примеры, как и весь многолетний опыт применения ДСМ-метода в различных областях, свидетельствуют о необходимости тщательной проработки исходной содержательной модели исследуемой проблемы для выделения значимых факторов и разработки адекватного формального языка представления данных. Окончательно можно говорить об оптимальном использовании всех возможностей применяемых процедур лишь после эмпирической верификации и экспертной оценки полученных результатов.

Заключение

Эвристический потенциал методов ИИ и их реализаций в интеллектуальных системах полноценно может быть актуализирован лишь при переходе от оппозиции «человек – машина» к партнёрским человеко-машинным системам [AI Index…, 2018]. Это является непосредственным следствием необходимости воспроизведения элементов ключевого для точной эпистемологии понятия теоретического (идеального) естественного интеллекта (ТЕИ) [Финн, 2023, с. 27-98] и, соответственно, перехода от логики доказательств к логике рассуждений [van Benthem, 2008]. В ДСМ-методе автоматизированной поддержки исследований логика рассуждений представляет собой логику синтеза познавательных процедур – амплиативных выводов на достаточном основании. Фундаментальная для нынешнего развития ИИ проблема доверия к результатам этих выводов (в отсутствии вывода как доказательства) решается здесь как эвристика приближения к достоверности гипотез [Финн, 2024b]. Тем самым обеспечивается решение проблемы создания эпистемологических оснований и эвристических средств организации и решения исследовательских задач – прежде всего, в лишённых развитого формального аппарата эмпирических областях. Являясь технологическим средством точной эпистемологии, партнёрские интеллектуальные ДСМ-системы (ИС-ДСМ) оказываются средством конструктивной аппроксимации человеческой познавательной деятельности в открытом мире. Однако удовлетворительная достижимость этой аппроксимации возможна лишь в условиях междисциплинарного взаимодействия исследователей в области ИИ и специалистов предметной области.

Список литературы

[Автоматическое порождение…, 2020] Автоматическое порождение гипотез в интеллектуальных системах / под общ. ред. В.К. Финна. – М.: Книжный дом «ЛИБРОКОМ», 2020 (изд. 2-е стереотип.). – 528 с.

[Гусакова, 2023] Gusakova S.M. Expansion of the functionality of the intellectual psychological and forensic system to solve new problems // Pattern Recognit. Image Anal. – 2023. – 33. – P. 345-349. – https://doi.org/10.1134/s1054661823030185.

[Гусакова и др., 2023] Гусакова С.М., Михеенкова М.А. О формировании эмпирических теорий в плохо формализованных областях // Двадцать первая Национальная конференция по искусственному интеллекту с международным участием КИИ-2023 (Смоленск, 16–20 октября 2023 г.): Труды конференции. В 2-х т. Т. 1. – Смоленск: Принт-Экспресс, 2023. – 410 с. – С. 169-179.

[Забежайло, 2013] Забежайло М.И. О функциональности отношения причинности, используемого в ДСМ-рассуждениях // НТИ, сер. 2. – 2013. – № 7. – С. 33-38.

[Милль, 2022] Милль Д.С. Система логики силлогистической и индуктивной (изд. стереотип.). – М.: URSS, 2022. – 832 с.

[Михеенкова, 2023] Михеенкова М.А. Формализация исследовательских эвристик для задач когнитивной социологии // НТИ, сер. 2.. – 2023. – № 8. – С. 1-10.

[Михеенкова, 2024] Михеенкова М.А. Отношение причинности как основа построения социальной типологии // Интегрированные модели и мягкие вычисления в искусственном интеллекте: Сборник научных трудов XII Международной научно-практической конференции (ИММВ-2024, Коломна, 14-17 мая 2024 г.). В 2-х т. Т. 2. – Смоленск: Универсум, 2024. – 315 с. – С. 115-125.

[Пойа, 1975] Пойа Д. Математика и правдоподобные рассуждения (под ред. С.А. Яновской). – М.: Наука, Главная редакция физико-математической литературы, 1975 (изд. 2-е). – 464 с.

[Поппер, 2005] Поппер К.Р. Логика научного исследования: пер. с англ. / под общей ред. В.Н. Садовского. – М.: Республика, 2005. – 447 с.

[Поппер, 2019] Поппер К.Р. Вся жизнь – решение проблем. О познании, истории, политике. Ч. I: Вопросы познания природы: пер. с нем. – М.: УРСС: ЛЕНАНД, 2019. – 200 с.

[Финн, 2023] Интеллект, информационное общество, гуманитарное знание и образование. – М.: ЛЕНАНД, 2023 (Изд. стереотип.). – 464 с.

[Финн, 2024a] Финн В.К. Об эмпирических закономерностях ранга *r* в ДСМ-методе автоматизированной поддержки исследований // НТИ, сер. 2. – 2024. – № 1. – С. 11-33.

[Финн, 2024b] Финн В. К. ДСМ-метод автоматизированной поддержки исследований и аппарат понятий для искусственного интеллекта // Искусственные общества. – 2024. – Т. 19. – Вып. 2. – DOI: 10.18254/S207751800030907-6. – URL: https://artsoc.jes.su/s207751800030907-6-1/.

[AI Index…, 2018] AI Index 2018 Report. – https://hai.stanford.edu/sites/default/files/2020-10/AI_Index_2018_Annual_Report.pdf (дата обращения: 25.05.2025).

[van Benthem, 2008] Benthem Van J. Logic and reasoning: do the facts matter? // Studia Logica. – 2008. – Vol. 88. – P. 67-84.

[McCarthy, 2014] McCarthy J. Epistemological problems of artificial intelligence // Readings in Artificial Intelligence (Ed. By Webber, B.L., Nilsson, N.J.). – Ca.: Morgan Kaufmann, 2014. – 547 p. – P. 459-465. – https://doi.org/10.1016/B978-0-934613-03-3.50035-0.

[Pearl et al., 2018] Pearl J., Mackenzie D. The Book of Why: The New Science of Cause and Effect. – New York: Basic Books, 2018.

УДК 004.852
doi: 10.15622/rcai.2025.013

АНАЛИЗ МЕТОДОВ ОЦЕНКИ ВАЖНОСТИ ПРЕДИКТОРОВ НЕБЛАГОПРИЯТНЫХ СОБЫТИЙ В КАРДИОХИРУРГИИ[1]

Б.В. Потапенко (*bvpotapenko@gmail.com*) [A]
К.И. Шахгельдян (*carinashakh@gmail.com*) [A,B]
Б.И. Гельцер (*boris.geltser@vvsu.ru*) [A,B]

[A] Владивостокский государственный университет, Владивосток
[B] Дальневосточный федеральный университет, Владивосток

В работе исследованы методы оценки важности предикторов моделей машинного обучения. Рассмотрены как подходы зависящие от архитектуры модели, так и независящие от неё. Модели обучались предсказывать вероятность наступления летальности в послеоперационный период для пациентов с инфарктом миокарда с подъемом сегмента ST, которым выполнено чрескожное коронарное вмешательство. Результаты демонстрируют заметное расхождение в ранжировании признаков по их важности в зависимости от применяемых методов. Особенно это заметно для признаков, которые влияют на предсказание нелинейно, и связаны с другими признаками. В работе поднимаются вопросы проблемы интерпретации важности в клинической медицине. Результаты указывают в пользу применения комбинированных методов оценки важности для повышения доверия к системам поддержки принятия врачебных решений.

Ключевые слова: объяснимый искусственный интеллект, важность предикторов, прогностические модели в клинической медицине, интеллектуальный анализ данных, машинное обучение.

Введение

Искусственный интеллект (ИИ) нашёл широкое применение в системах поддержки принятия врачебных решений (СППВР) в последнее десятилетие [Chen et al., 2023]. В здравоохранении применяются классические алгоритмы машинного обучения (МО): линейной и логистической регрес-

[1] Работа выполнена при финансовой поддержке проекта FZNS-2023-0010 Госзадания Дальневосточного федерального университета (ДВФУ).

сии, SVM, системы, основанные на правилах, деревья решений случайный лес [Papadopoulos et al., 2022], стохастический градиентный бустинг, методы глубокого обучения нейросетей [Shamshirband et al., 2021].

Для системы здравоохранения и клинической медицины особенно важны объяснения генерируемых моделями МО заключений [Pierce et al., 2022]. Сложность объяснения результатов работы моделей ограничивает применимость моделей машинного обучения в здравоохранении [Wubineh et al., 2024], [Khan et al., 2024], [Albahri et al., 2023].

Одной из концепций объяснения заключений, генерируемых моделями МО, является оценка важности признаков [Saarela et al., 2021], а именно какие признаки являются для модели наиболее важными в целом (глобальная важность), и что побудило модель сделать конкретное предсказание в частном случае (локальная важность). Важность, представленная численно, может быть отображена на диаграмме, сравнена с важностью другого признака; может быть измерено её изменение при обновлении параметров модели, или в процессе обучения, делая работу эксперта более эффективной [Wang et al., 2021].

Самыми популярными подходами к оценке важности являются оценки весовых коэффициентов логистической регрессии, методы, встроенные в ансамбли деревьев решений, семейство методов основанных на методе аддитивного объяснения Шепли (SHAP) [Lundberg et al., 2017], [Lundberg et al., 2020], оценка важности методом перестановок значений [Breiman et al., 2001].

Целью данного исследования является сравнительный анализ методов оценки важности предикторов, популярных в индустрии, их сильные и слабые стороны на примере задачи бинарной классификации на данных клинической медицины.

1. Материалы и методы

1.1. Датасет

Для анализа методов оценки важности предикторов мы использовали данные о больных инфарктом миокарда с подъемом сегмента ST (ИМпST), которым была выполнена операция чрескожного коронарного вмешательства (ЧКВ) в Краевой клинической больнице №1 г. Владивостока в период с 2016 до 2022 гг. Перед обработкой все данные были обезличены. В предыдущих исследованиях авторами были определены и валидированы предикторы внутригоспитальной летальности больных ИМпST после ЧКВ: возраст (Age), класс острой сердечно-сосудистой недостаточности (ОССН) по по T.Killip выше 2 (Killip_gt_2, бинарный признак), частота сердечных сокращений (HBR), систолическое артериальное давление (Systolic AP), уровень креатинина в крови (Creatinine), фракция выброса левого желудочка (EFLV), количество лейкоцитов в крови

(WBC), относительное значение количества нейтрофилов (Neutrophils), относительное значение количества эозинофилов (Eosinophils), тромбокрит (Thrombocrit) [Shakhgeldyan et al., 2024].

В качестве зависимой переменной рассматривается внутригоспитальная летальность (ВГЛ) – летальность в больнице или в 30-дневный период после проведения операции. В итоговую выборку вошли: 4668 записей о пациентах с ИМпST после ЧКВ, из которых 4355 (93.3%) относились к группе выживших, а 313 (6,7%) – к группе ВГЛ.

1.2. Методы

Мы применили оценки статистической значимости: t-тест; U-тест; Хи2-тест. Модель-зависимые методы МО: однофакторная (LR) и многофакторная логистические регрессии (MLR); случайный лес (RF), CatBoost (CB), стохастический градиентный бустинг (XGB) и искусственная нейронная сеть – многослойный персептрон (NN). Для оценки важности предикторов в первых двух методах использовали весовые коэффициенты моделей, в оставшихся – характеристику importance, которая вычисляется в процессе обучения моделей. Модель-независимые методы: важность на основе метода перестановок и аддитивного объяснения Шепли (SHAP) [Lundberg et al., 2020].

1.3. Дизайн исследования

На первом этапе исследования статистическими методами оценивалось влияние признаков на зависимую переменную. Для непрерывных признаков были использовались: тест Стьюдента, тест Манна-Уитни. Для категориальных признаков – тест хи-квадрат. Для сравнения важности предикторов использовали p-value, минимальное значение которого соответствует максимальной важности предиктора. Кроме того, вычислены весовые коэффициенты LR на основе нормированных методом MinMax предикторов.

На втором этапе были применены методы МО. Данные были разделены на обучающие и валидационные с учётом стратификации. Для обучения методами MLR и NN данные были нормированы методом MinMax. В процессе обучения моделей выполнялся подбор гипер-параметров с помощью стратифицированной k-fold кросс-валидации. Выбор лучшей модели осуществлялся с на основе максимизации метрики площади под ROC-кривой (AUC). Для обучения моделей использовали MLR, RF, CB, XGB, NN. Сравнили нормированные коэффициенты регрессии и важность признаков по собственной оценке моделей на основе деревьев решений.

На третьем этапе оценили важность признаков для этих моделей методами SHAP и методом перестановки. Метрикой для оценки важности методом перестановок была выбрана AUC на валидационном наборе данных.

2. Результаты

2.1. Статистические метрики важности

Первый подход к оценке важности основывается на сравнении статистических параметров: t-статистика, r-значение и p-value, полученные методами межгрупповых сравнений – тестом Стьюдента, Манна-Уитни. Нормализованные значения обоих тестов и оценки p-value представлены в табл. 1. Результаты t-теста показали, что наиболее значимыми признаками являются: Creatinine, WBC, EF LV и Neutrophils . Наименее значимый – Thrombocrit. С точки зрения теста Манна-Уитни наиболее значимые предикторы – Neutrophils, Eosinophils, Systolic AP и Age, а наименее – Thrombocrit. Признаки, в важности которого уверены оба теста – это Neutrophils. Оба теста верифицировали Thrombocrit, как наименее важный предиктор. Самым противоречивым оказался признак Eosinophils.

Таблица 1

Статистики межгрупповых сравнений тестами Стьюдента и Манна-Уитни				
Предикторы	t-test rank	t-test p-value	U-test rank	U-test p-value
Age	0.57	2.676e-37	0.72	6.271E-35
HBR	0.77	7.162e-55	0.68	5.567E-32
Systolic AP	0.84	1.516e-61	0.75	2.342E-37
Creatinine	1.00	9.353e-79	0.72	1.532E-34
EF LV	0.85	4.705e-63	0.74	5.185E-35
WBC	0.89	1.759e-66	0.70	4.184E-33
Neutrophils	0.84	4.162e-61	1.00	2.622E-55
Eosinophils	0.30	6.130e-19	0.88	1.082E-44
Thrombocrit	0.00	5.603e-06	0.00	5.707E-03

Для категориального признака (Killip_gt_2) вычислили значения χ^2, (p-value =5.699e-80). Согласно p-value Killip_gt_2 является наиболее значимым среди всех предикторов ВГЛ.

2.2. Коэффициенты однофакторной логистической регрессии

Важность предикторов может ассоциироваться с модулем весового коэффициента LR. На первое место по значимости выходят Systolic AP (7.02), Neutrophils (6.97), на третьем месте по важности Creatinine (6.76). Далее WBC (5.94), HBR (b) (5.81), EF LV (6.02), Eosinophils (5.35), Age (4.26), Thrombocrit (2.11), Killip_gt_2 (2.01). Таким образом, наиболее важными с точки зрения LR являются Systolic AP, Neutrophils и Creatinine, наименее Thrombocrit и Killip_gt_2.

2.3. Многофакторные модели машинного обучения

Для оценки важности мы рассматриваем не только изолированное влияние предикторов на конечную точку, но и то, как это влияние проявляется при работе в многофакторных моделях. Для MLR – мы рассматриваем весовые коэффициенты, для ансамблевых методов – внутренний механизм расчета важности (importance). Для сравнения все оценки были нормированы на максимальнее значения (табл. 2).

Среди коэффициентов MLR с заметным отрывом лидирует Creatinine (относительная важность =1), за ним следуют EF LV (относительная важность = 0.379) и Neutrophils (относительная важность = 0.374). Наименьший коэффициент был получены при Killip_gt_2 (относительная важность = 0). По мнению трех ансамблевых моделей на основе деревьев решений Neutrophils – самый важный признак, за ним следует Eosinophils. Признак Age занимает третье место в оценке CatBoost, при этом XGBoost относит его к наименее важным, при этом на третье место последний ставит Killip_gt_2. Показатель Creatinine занимает третье место согласно RF. Наименее важными признаками являются согласно ансамблевым методам – Systolic AP и Thrombocrit.

Таблица 2

Оценка важности предикторов многофакторных моделей				
Предикторы	MLR	RF	CB	XGB
Age	0.278	0.197	0.634	0.045
Killip_gt_2	0.000	0.111	0.136	0.287
HBR	0.304	0.140	0.248	0.003
Systolic AP	0.183	0.050	0.000	0.000
Creatinine	1.000	0.518	0.441	0.088
EF LV	0.379	0.470	0.490	0.267
WBC	0.239	0.143	0.065	0.019
Neutrophils	0.374	1.000	1.000	1.000
Eosinophils	0.178	0.782	0.873	0.377
Thrombocrit	0.227	0.000	0.009	0.008

2.4. Глобальные оценки SHAP

Метод SHAP позволяет оценить важность признаков в многофакторной модели, независимо от модели и после ее обучения. Метод оценивает степень влияния признака по величине shap-value, которая при положительном значении описывает влияние на риск развития неблагоприятного события. По оценке SHAP для положительного класса в MLR самыми важными являются признаки Neutrophils, EF LV, HBR (табл. 3). Данный результат расходится с оценкой на основании весовых коэффициентов регрессии, где лидером был Creatinine. В то же время 2 других признака – Neutrophils, EF LV повторяются. Согласно SHAP в MLR Creatinine занимает 5 место в ранге важности.

Таблица 3

Глобальная оценка SHAP для положительного класса					
Предикторы	MLR	RF	CB	XGB	NN
Age	0.52	0.04	0.35	0.32	4.04
Killip_gt_2	0.32	0.04	0.35	0.27	3.98
HBR	0.61	0.04	0.39	0.32	4.5
Systolic AP	0.24	0.02	0.14	0.13	1.56
Creatinine	0.38	0.07	0.46	0.42	1.48
EF LV	0.64	0.06	0.48	0.39	2.57
WBC	0.29	0.03	0.18	0.16	2.15
Neutrophils	0.72	0.11	0.54	0.75	9.26
Eosinophils	0.13	0.09	0.41	0.41	10.83
Thrombocrit	0.24	0.01	0.13	0.14	1.82

Для моделей RF, CatBoost и XGBoost важность по SHAP предиктора Neutrophils подтверждается, но заметно отличается от MLR переходом признака Eosinophils с последнего места на второе, четвёртое и третье соответственно. Среди значений SHAP для признаков NN выделяется вышедший на первое место предиктор Eosinophils, Neutrophils на вторую позицию и заметное изменение рангов для остальных признаков.

2.5. Важность на основе метода перестановок

Метод перестановок подразумевает искажение по очереди каждого признака и оценку снижения при этом точности прогностической модели. В качестве базовой метрики качества моделей используется площадь под ROC-кривой (AUC). Важными признаками для всех моделей по этому методу оценки являются: Neutrophils, Eosinophils, HBR, Age, Creatinine (табл. 4). Остальные признаки не попали в тройку наиболее важных ни для одной из моделей.

Таблица 4

Важность признаков по методу перестановок					
Предикторы	LR	RF	CB	XGB	NN
Age	**0.030**	0.014	0.020	**0.023**	**0.035**
Killip_gt_2	0.001	0.004	0.001	0.001	0.002
HBR	**0.022**	0.016	**0.020**	0.017	**0.035**
Systolic AP	0.005	0.003	0.004	0.002	0.000
Creatinine	0.016	**0.022**	**0.020**	0.014	0.019
EF LV	0.011	0.009	0.011	0.014	0.021
WBC	0.010	0.006	0.003	0.002	0.022
Neutrophils	**0.026**	**0.037**	**0.024**	**0.057**	**0.100**
Eosinophils	0.004	**0.023**	0.015	**0.023**	**0.079**
Thrombocrit	0.001	0.000	-0.001	0.000	-0.003

Метод перестановок подтвердил значимость Neutrophils, поставив его на первое место для всех моделей кроме MLR, где он поставил его на второе место. Показатель Eosinophils занимал вторую позицию за исключением MLR и CatBoost. Все модели демонстрировали низкий уровень значимости для показателей Killip_gt_2 и Thrombocrit.

Можно заметить различия между глобальной важностью по SHAP и важностью на основе перестановок. Так для MLR Age и HBR оказались важнее, чем EF LV (второй по важности по оценке SHAP); у модели CatBoost признаки EF LV и Eosinophils в оценке перестановками уступили HBR; у модели XGBoost признак Age по оценке перестановками оказался важнее, чем EF LV и Creatinine; а для NN при оценке перестановками признак Neutrophils оказался важнее, чем Eosinophils.

Обсуждение

В данной работе мы рассмотрели несколько подходов к оценке важности предикторов на примере задачи прогнозирования ВГЛ у пациентов с ИМпST после ЧКВ. Обобщая важность предикторов, полученных разными методами, можно представить их ранг (табл. 5).

Таблица 5

Обобщенный ранг важности предикторов						
Предикторы	Статистика	LR	MLR	Ансамбли	SHAP	Перестановки
Age	**8**	**8**	**5**	5	6	**4**
Killip_gt_2	1	10	10	6	7	7
HBR	7	5	4	7	5	3
Systolic AP	5	1	8	10	10	8
Creatinine	2	3	1	3	2	5
EF LV	**4**	**6**	**2**	**4**	**4**	6
WBC	3	4	6	8	8	**9**
Neutrophils	**6**	**2**	**3**	**1**	**1**	**1**
Eosinophils	9	7	9	**2**	**3**	2
Thrombocrit	10	9	7	9	9	10

Анализ рейтинга важности предикторов в их влиянии на конечную точку позволяет сделать несколько замечаний. Разные подходы к оценке важности обеспечивают противоречивые результаты, вплоть до прямо противоположных. Так, например, предиктор Systolic AP имеет максимальную важность при сравнении весовых коэффициентов LR, и наименьшую важность при многофакторных ансамблевых моделях и при использовании SHAP. Класс ОССН по T.Killip имеет максимальную важно при статистической оценке и минимальную – при работе в MLR. Оценка важности зависит от методов МО, включая использования методов SHAP или перестановки.

Наиболее существенные различия мы можем наблюдать в сравнении методов MLR и ансамблевых методов на основе деревьев решений. Это согласуется с выводами коллег [Saarela et al., 2021], [Khan et al., 2024]. Ансамблевые методы МО обеспечивают устойчивую оценку для наиболее важных предикторов, которая подтверждается при наложении методов SHAP и перестановки. Так, наиболее важными в этом случае были Neutrophils, Eosinophils и Creatinine. Наименее важные предикторы подтверждают свой рейтинг в большинстве методов оценки. Например, Thrombocrit имеют низкую важность согласно всем рассматриваемым подходам.

Наша работа подтверждает высокую значимость показателей EF LV и Creatinine, которая была оценена в работе по эпидемиологии сердечно сосудистых заболеваний [Ziaeian et al., 2016]. При этом мы показали, что признак Neutrophils является одним из наиболее важных предикторов ВГЛ у пациентов с ИМпST после ЧКВ. Существенные различия в оценке методов, использующих MLR и ансамблевых методов на основе деревьев решений можно объяснить учетом линейных и нелинейных отношений между предикторами и конечной точкой. При наличии линейных отношений рейтинговая оценка важности подтверждается разными подходами. Примером такой взаимосвязи служит Age, увеличение которого ведет к росту вероятности ВГЛ у пациентов с ИМпST после ЧКВ. Противоположным примером является признак Eosinophils, который демонстрирует низкий уровень значимости в MLR и статистике, но высокий – для ансамблевых методов. Аналогичную нелинейную зависимость можно предположить у Creatinine, который обладал наибольшим коэффициентом в MLR, но при этом в подходе на основе перестановки был лишь на 5-6 позиции.

Наше исследование демонстрирует, что рассмотренные оценки важности тем более стабильны, чем более сильна и более линейна связь зависимой переменной с конкретным признаком. И менее эффективны там, где предиктор нелинейно влияет на зависимую переменную, что типично для данных клинической медицины.

Заключение

В данной работе мы показали, что разные подходы к оценке важности предикторов прогностических моделей обеспечивают противоречивые результаты, вплоть до прямо противоположных. Оценка важности зависит от методов МО, выбранной архитектуры моделей. Ансамблевые методы МО обеспечивают устойчивую оценку для наиболее важных предикторов, которая подтверждается при наложении методов SHAP и перестановки. Методы оценки важности предикторов имеют демонстрируют разные результаты, когда связь между предикторами и предсказанием нелинейна. Таким образом, проблема выбора метода оценки важности актуальна, как и задача разработки новых более универсальных алгоритмов.

Список литературы

[Albahri et al., 2023] Albahri A.S. [et al.]. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion // Information Fusion. – 2023. – Vol. 96. – P. 156-191.

[Breiman, 2001] Breiman L. Random Forests // Machine Learning. – 2001. – Vol. 45(1). – P. 5-32. – doi: 10.1023/A:1010933404324.

[Chen et al., 2023] Chen Z. [et al.]. Harnessing the power of clinical decision support systems: challenges and opportunities // Open Heart. – 2023. – doi: 10.1136/openhrt-2023-002432.

[Khan et al., 2024] Khan N. [et al.]. Guaranteeing Correctness in Black-Box Machine Learning: A Fusion of Explainable AI and Formal Methods for Healthcare Decision-Making // IEEE Access. – 2023. – doi: 10.1109/ACCESS.2024.3420415.

[Lundberg et al., 2017] Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems. – 2017. – P. 4768-4777. – doi: 10.5555/3295222.3295230.

[Lundberg et al., 2020] Lundberg S.M. [et al.]. From local explanations to global understanding with explainable AI for trees // Nature Machine Intelligence. – 2020. – Vol. 2. – P. 56-67. – doi:10.1038/s42256-019-0138-9.

[Papadopoulos et al., 2022] Papadopoulos M [et al.]. A systematic review of technologies and standards used in the development of rule-based clinical decision support systems // Health Technology. – 2022. – Vol. 12. – P. 713-727. – doi: 10.1007/s12553-022-00672-9.

[Pierce et al., 2022] Pierce R.L. [et al.]. Explainability in medicine in an era of AI-based clinical decision support systems // Frontiers in Genetics. – 2022. – Vol. 13. – 903600. – doi:10.3389/FGENE.2022.903600/BIBTEX.

[Saarela et al., 2021] Saarela M., Jauhiainen S. Comparison of feature importance measures as explanations for classification models // SN Applied Sciences. – 2021. – Vol. 3(2). – P. 1-12. – doi:10.1007/s42452-021-04148-9/TABLES/4.

[Shakhgeldyan et al., 2024] Shakhgeldyan K.J., Kuksin N.S., Domzhalov I.G., Geltser B.I. Methods of prognostic analysis for the prediction of in-hospital mortality in patients with acute st-elevation myocardial infarction after percutaneous coronary interventions // Pattern Recognition and Image Analysis. – 2024. – T. 34. – Vol. 3. – P. 786-796. – doi: 10.1134/S1054661824700676.

[Shamshirband et al., 2021] Shamshirband S. [et al.]. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues // Journal of Biomedical Informatics. – 2021. – Vol. 113. – 103627. – doi: 10.1016/J.JBI.2020.103627.

[Wang et al., 2021] Wang X., Yin M. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making // International Conference on Intelligent User Interfaces, Proceedings IUI. – 2021. – P. 318-328. – doi: 10.1145/3397481.3450650/SUPL_FILE/P318-WANG.PDF.

[Wubineh et al., 2024] Wubineh B.Z., Deriba F.G., Woldeyohannis M.M. Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: A systematic literature review // Urologic Oncology: Seminars and Original Investigations. – 2024. – Vol. 42(3). – P. 48-56. – doi: 10.1016/J.UROLONC.2023.11.019.

[Ziaeian et al., 2016] Ziaeian B., Fonarow G. Epidemiology and aetiology of heart failure // Nature Reviews Cardiology. – 2016. – Vol. 13(6). – P. 368-378. – doi: 10.1038/nrcardio.2016.25.

УДК 004.94
doi: 10.15622/rcai.2025.014

ИССЛЕДОВАНИЕ И РЕАЛИЗАЦИЯ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ОБВОДНЁННОСТИ ПРОДУКЦИИ НЕФТЯНЫХ СКВАЖИН[1]

И.Б. Фоминых (*igborfomin@mail.ru*)
И.С. Михайлов (*fr82@mail.ru*)
К.О. Сидоров (*kirill.sidoroff2014@yandex.ru*)
Мьо Хлайн Вин (*myohlaingwin69287@gmail.com*)

Национальный исследовательский университет «МЭИ», Москва

В работе предлагается решение актуальной задачи прогнозирования обводнённости продукции нефтяных скважин с использованием методов интеллектуального анализа данных. Рассмотрены методы интеллектуального анализа данных: LSTM, BiLSTM, Prophet, ARIMA, XGBoost, NNAR и TBATS. Выявлены их сильные и слабые стороны. Выполнена реализация и тестирование данных методов с использованием реальных данных, полученных на нефтяных месторождениях. Показано, что для краткосрочного прогнозирования лучше использовать модели LSTM, BiLSTM и NNAR, для долгосрочного прогнозирования изменения обводнённости продукции скважин лучше использовать LSTM модель.

Ключевые слова: интеллектуальный анализ данных, искусственная нейронная сеть LSTM, прогнозирование, обводнённость продукции нефтяных скважин.

Введение

Уровень применения методов интеллектуального анализа данных в современных информационных системах неуклонно растёт. Данные методы используются для решения различных задач, связанных с классификацией, регрессией, прогнозированием, ассоциацией и другими. Наиболее сильно данная тенденция выражена в информационных системах, которые

[1] Работа выполнена при финансовой поддержке РНФ (проект № 24-11-00285), https://rscf.ru/project/24-11-00285/.

отвечают за анализ и управление различными технологическими процессами, поскольку объём обрабатываемой ими информации весьма велик и обычные математические модели требуют слишком детальной настройки на каждый отдельный случай [Барабанов и др., 2019]. В частности, к таким задачам относится задача анализа изменения обводнённости продукции скважин в нефтяной промышленности. Данный параметр наблюдается в режиме реального времени с помощью многофазного расходомера, устанавливаемого на устье нефтяных скважин. Необходимо отметить, что при увеличении обводнённости продукции скважины необходимо принимать управляющие воздействия, направленные на изменение режима добычи или технологии добычи нефти на ней, поскольку в противном случае скважина может стать нерентабельной. Также, в случае роста обводнённости может измениться режим течения продукции скважины, что приведёт к нештатной работе добывающего оборудования и возможной остановке добычи. Поэтому очень важно отслеживать данный параметр и качественно предсказывать его значение.

Однако этот процесс осложняется множеством факторов, включая свойства пласта, параметры оборудования и режим эксплуатации, что делает традиционные методы анализа недостаточно точными [Bian et al, 2022].

В связи с большим количеством обрабатываемых данных возникает необходимость в применении методов интеллектуального анализа, способных учитывать нелинейные зависимости и выявлять скрытые закономерности [Dong, 2022].

Данное исследование посвящено решению задачи прогнозирования изменения обводнённости продукции нефтяных скважин на основе методов машинного обучения. При этом необходимо выполнить анализ существующих подходов для решения задачи прогнозирования временных рядов, выявить их преимущества и ограничения, реализовать алгоритмы машинного обучения для решения поставленной задачи, выполнить тестирование и сравнение моделей на реальных данных измерительных систем, установленных на нефтяных скважинах [Sarker, 2021].

1. Измерение обводнённости продукции скважин

В качестве исходных данных для решения задачи прогнозирования обводнённости продукции скважин рассматриваются показания многофазных расходомеров, установленных на нефтяных скважинах. Они регистрируют в режиме реального времени следующий набор параметров: скорость нефте-водо-газового потока, газосодержание потока, скорость прохождения ультразвука в контролируемом объёме, температуру, давление. Расчётными параметрами являются: расход жидкости, расход газа, обводнённость продукции. Согласно разработанным математическим моделям

показано, что указанный набор параметров образует многомерное пространство, определяющее расходные параметры потока, в том числе обводнённость потока.

Массив обучающих данных представляет собой набор результатов измерений расходомеров на скважинах с различными химическими параметрами продукции, с различным оборудованием добычи и различными режимами течения продукции.

Для повышения достоверности анализа проведена предобработка данных, включающая удаление следующих строк:

1) постоянных в течение длительного времени данных, поскольку они соответствуют остановке нефте-водо-газового потока на измерительном участке;
2) с нулевой скоростью ультразвука (ошибки измерений);
3) с одинаковой скоростью в разных точках измерения (некорректные данные);
4) с 100% содержанием газа, поскольку это означает, что измерительный участок заполнен попутным нефтяным газом;
5) с неизменной скоростью ультразвука при изменяющихся других параметрах (возможное зависание сенсора).

Очищенные данные сохранены в CSV-файл для дальнейшего анализа.

2. Методы интеллектуального анализа данных для решения задачи прогнозирования

В данном исследовании рассматриваются следующие наиболее популярные методы решения задачи прогнозирования.

2.1. LSTM

Long short-term memory (LSTM) – реккурентная нейронная сеть (RNN), которая используется для решения задач обработки естественного языка, распознавания речи, музыкальной композиции и других задач, где есть последовательность входных данных. LSTM способна обрабатывать длинные последовательности данных с запоминанием предыдущих состояний. Данные сети были разработаны для решения проблемы исчезающего градиента, с которой можно столкнуться при обучении традиционных RNN [Hochreiter et al., 1997].

В отличие от других алгоритмов машинного обучения, рекуррентные нейронные сети с долгой кратковременной памятью способны автоматически выявлять признаки из временных последовательностей, обрабатывать многомерные данные, а также выводить последовательности переменной длины, благодаря чему их можно использовать для прогнозирования временных рядов.

LSTM-сети демонстрируют высокую эффективность при работе с временными рядами благодаря своей способности анализировать долгосрочные временные зависимости. Их архитектура специально разработана для решения задач классификации, обработки и прогнозирования последовательных данных, где критически важные события могут быть разделены произвольными временными интервалами. LSTM-архитектура была создана в качестве решения ключевой проблемы традиционных рекуррентных сетей – эффекта исчезающих градиентов. В отличие от стандартных RNN и скрытых марковских моделей, LSTM демонстрируют устойчивую работу с длинными временными зависимостями, что обеспечивает их превосходство в различных задачах обработки последовательностей.

2.2. BiLSTM

Бидирекциональная (двунаправленная) сеть с долгой кратковременной памятью (BiLSTM) представляет собой архитектуру рекуррентных нейронных сетей, которая расширяет традиционную LSTM за счёт одновременной обработки входных данных в двух направлениях: прямом и обратном. В отличие от классической LSTM, выполняющей анализ последовательности только в хронологическом порядке (с начала к концу), BiLSTM использует две параллельные LSTM-сети – одна обрабатывает данные в прямом направлении, другая – в обратном. Результаты обеих сетей объединяются, что позволяет модели учитывать контекст как из предыдущих, так и из последующих элементов последовательности.

Данная архитектура особенно эффективна в задачах, где значение текущего элемента зависит от информации, расположенной как в прошлом, так и в будущем, например, при обработке естественного языка, распознавании речи, анализе временных рядов и других последовательных данных. Благодаря двунаправленному подходу BiLSTM способна лучше выявлять долгосрочные зависимости и улучшать качество предсказаний по сравнению с однонаправленными моделями [Schuster et al., 1997].

Принцип работы BiLSTM заключается в следующем: входная последовательность сначала подаётся на прямую LSTM-сеть, которая формирует скрытые состояния, учитывающие контекст слева направо. Затем та же последовательность обрабатывается в обратном порядке другой LSTM-сетью, формирующей скрытые состояния с учётом контекста справа налево. Итоговые представления, полученные из обеих направлений, объединяются (например, посредством конкатенации), что обеспечивает более полное описание каждого элемента с учётом двунаправленного контекста.

2.3. Prophet

Prophet – это метод прогнозирования временных рядов, основанный на аддитивной модели, которая учитывает нелинейные тренды, отражающие годовые, недельные и суточные сезонные колебания, а также влияние

праздничных дней. Данный инструмент особенно эффективен при анализе временных рядов с выраженными сезонными паттернами и наличием нескольких циклов исторических данных. Prophet устойчив к пропущенным значениям и резким изменениям тренда, а также обычно хорошо справляется с аномалиями в данных [Taylor, Letham, 2017].

Prophet имеет несколько нижеприведенных преимуществ:

- Prophet, благодаря точности и скорости, используется во многих приложениях в Facebook для создания надежных прогнозов для планирования и постановки целей.
- Prophet создает настраиваемые прогнозы, представляет множество возможностей для пользователей создавать и корректировать прогнозы. Возможно использовать интерпретируемые человеком параметры для улучшения прогноза, добавляя знания о предметной области.
- Доступен в R или Python.
- Хорошо обрабатывает сезонные колебания. Prophet учитывает сезонность с несколькими периодами.
- Устойчив к выбросам (обрабатывает выбросы, удаляя их).

2.4. ARIMA

ARIMA – это аббревиатура, которая расшифровывается как Auto Regressive Integrated Moving Average (Авторегрессивное интегрированное скользящее среднее). Это алгоритм прогнозирования, основанный на идее, что информация в прошлых значениях временного ряда может использоваться сама по себе для прогнозирования будущих значений [Box et al., 1994].

Любой несезонный временной ряд, который демонстрирует закономерности и не является случайным белым шумом, может быть смоделирован с помощью моделей ARIMA.

Аббревиатура ARIMA является описательной и отражает ключевые аспекты самой модели. Вкратце, они таковы:

AR: Авторегрессия. Модель, которая использует зависимую связь между наблюдением и некоторым количеством наблюдений с лагом.

I: Интегрированная. Использование дифференциации сырых наблюдений (например, вычитание наблюдения из наблюдения на предыдущем временном шаге) для того, чтобы сделать временной ряд стационарным.

MA: Скользящее среднее. Модель, которая использует зависимость между наблюдением и остаточной ошибкой из модели скользящего среднего, применяемой к запаздывающим наблюдениям. Каждый из этих компонентов явно указан в модели как параметр. Используется стандартная нотация ARIMA(p,d,q), где параметры заменяются целыми значениями для быстрого указания конкретной используемой модели ARIMA.

Параметры модели ARIMA определяются следующим образом: d - количество разностей, необходимых для того, чтобы сделать временной ряд стационарным, p – количество наблюдений запаздывания, включенных в модель, т.е. порядок члена AR, q – размер окна скользящего среднего, т. е. порядок члена MA.

Таким образом, цель обучения модели состоит в том, чтобы определить значения p, d и q.

2.5. XGBoost

XGBoost (eXtremeGradientBoosting) – это высокоэффективная реализация алгоритма градиентного бустинга на основе деревьев решений.

Бустинг – это метод, при котором модели (чаще всего деревья решений) обучаются последовательно, а не независимо, как в случайных лесах. Каждое последующее дерево концентрируется на исправлении ошибок, сделанных предыдущими моделями. Итоговый прогноз получается, как взвешенная сумма предсказаний всех деревьев [Chen et al., 2016].

К основным параметрам модели относятся: количество деревьев, максимальная глубина дерева, скорость обучения, минимальное уменьшение функции потерь для разбиения узла, доля данных для обучения каждого дерева, доля признаков для построения дерева, коэффициенты L1 и L2 регуляризации, функция потерь.

Основные особенности XGBoost, отличающие его от других алгоритмов градиентного бустинга:

- Адаптивный штраф при построении деревьев.
- Пропорциональное уменьшение узлов листьев.
- Метод Ньютона в оптимизации.
- Дополнительный параметр рандомизации.
- Автоматический отбор признаков.

2.6. NNAR

NNAR (NNETAR**) (**Neural Network AutoRegression) – это гибридная модель, сочетающая:

- Авторегрессию (AR) – использование лагов временного ряда как признаков. В классической модели AR прогноз строится на основе предыдущих значений ряда. Например, значение в момент t зависит от значений в моментах $t-1, t-2, \ldots, t-p$.
- Нейронную сеть с одним скрытым слоем – для учета нелинейных зависимостей. Вместо простой линейной комбинации прошлых наблюдений, NNETAR использует многослойную перцептронную нейросеть, которая способна моделировать сложные нелинейные зависимости [Gregor et al., 2014].

2.7. TBATS

TBATS – это аббревиатура для обозначения тригонометрической сезонности (Trigonometricseasonality) преобразования Бокса-Кокса (Box-Coxtransformation) ошибки ARMA (ARMA errors) тренда (Trend) сезонных компонент (Seasonalcomponents). TBATS был разработан для прогнозирования временных рядов с несколькими сезонными периодами. Например, ежедневные данные могут иметь как недельный, так и годовой шаблон. Или часовые данные могут иметь три сезонных периода: ежедневный шаблон, недельный шаблон и годовой шаблон [De Livera et al., 2011].

В модели TBATS к исходному временному ряду сначала применяется преобразование Бокса-Кокса, после чего ряд описывается как линейное сочетание экспоненциально сглаженного тренда, сезонных компонентов и компонента ARMA. Для моделирования сезонности используются тригонометрические функции, основанные на разложении в ряды Фурье. Выбор и настройка гиперпараметров модели, таких как включение или исключение отдельных компонентов, осуществляется с помощью критерия информационного критерия Акаике (AIC).

3. Реализация рассмотренных методов и сравнение результатов тестирования

Рассмотренные методы были реализованы на языке Python. Также было выполнено их тестирование на подготовленных наборах данных и произведено их сравнение по основным метрикам для поиска оптимальной модели. Прогнозирование осуществлялось на интервалах различной длительности.

Результаты сравнения представлены в табл. 1.

Таблица 1

Сравнение оценок результатов прогнозирования с помощью различных моделей

Модель	Интервал	MAE	MSE	RMSE	R^2	MAPE (%)	WAPE (%)
LSTM	30 сек	1.2513	4.2588	2.0637	0.8829	1.51	1.38
	1 мин	1.8107	6.7384	2.5958	0.9584	2.58	2.29
	2 мин	1.8252	5.6176	2.3702	0.9456	2.53	2.41
	5 мин	1.9252	5.4583	2.3363	0.9614	2.56	2.49
	1 час	2.7792	13.2882	3.6453	0.9661	4.98	4.07
BiLSTM	30 сек	2.1495	8.6997	2.9495	0.7608	2.53	2.37
	1 мин	2.4367	9.5888	3.0966	0.9407	3.35	3.09
	2 мин	2.4726	8.9197	2.9866	0.9136	3.35	3.26
	5 мин	2.2178	7.0866	2.6621	0.9499	2.97	2.87
	1 час	2.7607	14.1673	3.7640	0.9639	5.09	4.04

Окончание табл. 1

Модель	Интервал	MAE	MSE	RMSE	R^2	MAPE (%)	WAPE (%)
Prophet	30 сек	28.4125	843.6725	29.0460	-22.2004	31.06	31.38
	1 мин	16.7951	444.0584	21.0727	-1.7440	19.20	21.27
	2 мин	13.7081	291.3234	17.0682	-1.8212	16.69	18.07
	5 мин	15.2648	373.0373	19.3142	-1.6370	17.92	19.72
	1 час	17.5568	428.0774	20.6900	-0.0919	32.25	25.68
ARIMA	30 сек	7.5700	99.8300	9.9900	0.7977	17.42	11.61
	1 мин	11.9700	256.3900	16.0100	0.4804	28.00	18.34
	2 мин	15.5000	388.7300	19.7200	0.2122	38.89	23.74
	5 мин	18.1800	498.1300	22.3200	-0.0096	44.78	27.85
	1 час	21.2600	632.9900	25.1600	-0.2829	42.63	32.57
XGBoost	30 сек	2.4243	7.9973	2.8280	0.7801	2.74	2.68
	1 мин	2.7073	13.1424	3.6252	0.9188	3.66	3.43
	2 мин	3.1110	14.4441	3.8005	0.8601	4.26	4.10
	5 мин	3.0611	14.4515	3.8015	0.8978	4.13	3.95
	1 час	3.6273	21.8579	4.6752	0.9442	6.55	5.31
NNAR	30 сек	2.6841	7.6938	2.7738	0.7884	3.02	2.96
	1 мин	2.6474	8.6734	2.9451	0.9464	3.50	3.35
	2 мин	2.1953	7.6499	2.7659	0.9259	2.93	2.89
	5 мин	2.3024	8.0152	2.8311	0.9433	3.07	2.97
	1 час	3.3613	20.0977	4.4830	0.9487	6.07	4.92
TBATS	30 сек	8.2331	87.3586	9.3466	-1.4023	9.47	9.09
	1 мин	11.9814	174.5704	13.2125	-0.0787	16.39	15.18
	2 мин	10.8501	162.5795	12.7507	-0.5744	14.54	14.30
	5 мин	11.8181	193.0253	13.8934	-0.3645	15.72	15.27
	1 час	17.6330	526.5015	22.9456	-0.3429	39.87	25.80

На основе полученных результатов можно сделать следующие выводы.

3.1. LSTM

Обладает хорошей производительностью (R^2 около 0.88–0.96), что указывает на сильную способность модели объяснять данные.

MAE и RMSE увеличиваются с увеличением интервала времени, что нормально для временных рядов.

По MAPE и WAPE модель демонстрирует хорошие результаты с низкими значениями данных оценок.

Высокая частота данных позволяет LSTM эффективно улавливать краткосрочные зависимости.

LSTM запоминает предыдущие значения, что важно для плавно меняющихся метрик.

Dropout и Dense слои помогают обобщить данные и сгладить влияние шума.

3.2. BiLSTM

R^2 выше для интервалов времени 1 минута и больше (0.94 и выше), что говорит о хорошей объясняющей способности.

Однако MAE и RMSE в некоторых случаях хуже, чем у LSTM, особенно на 30 секунд и 1 минуту.

MAPE и WAPE немного выше по сравнению с LSTM.

3.3. Prophet

Очень слабая производительность по R^2 (от -22 до -0.09), что указывает на плохую способность модели объяснять данные. Модель явно не подходит для данной задачи.

MAE, MSE, RMSE, MAPE и WAPE также имеют очень высокие значения, что делает эту модель непригодной для прогнозирования в данном контексте. Prophet плохо работает с высокочастотными данными, он ориентирован на дневной и недельный масштаб изменения наблюдаемых данных. Модель не учитывает краткосрочные зависимости, важные при интервалах 30 сек – 5 мин. Отсутствие автокорреляции в ядре модели делает её уязвимой к шуму и быстрым случайным отклонениям какой-либо величины.

3.4. ARIMA

R^2 ухудшается с увеличением интервала времени (от 0.80 до отрицательных значений), что указывает на плохую предсказательную способность модели на длительные интервалы.

MAE и RMSE также показывают большие ошибки на больших интервалах времени.

MAPE и WAPE также высоки, что делает модель менее эффективной по сравнению с нейронными сетями.

Приведение ряда к стационарному виду улучшает качество, но ARIMA всё ещё линейная модель, и не справляется с нелинейными и асимметричными паттернами в метрике.

При дифференцировании модель теряет часть долгосрочной динамики, что делает её прогноз менее устойчивым на горизонтах >2 минут.

Также, ARIMA не учитывает взаимодействие с другими переменными, и не имеет механизма адаптации к смене режима генерации данных (например, резкие падения).

3.5. XGBoost

XGBoost характеризуется хорошим значением R^2 (0.78–0.94), с улучшением на более длинных интервалах времени.

MAE и RMSE показывают стабильные результаты по сравнению с другими моделями.

Значения MAPE и WAPE невысокие, что делает эту модель достаточно эффективной для прогнозирования.

XGBoost использует бустинг деревьев решений, которые обнаруживают сложные нелинейные зависимости между лагами.

Хорошо обрабатывает высокочастотные шумные данные, если заранее заданы временные признаки (лаги, rollingwindow и др.).

Если нет предположений о стационарности, то модель легко адаптируется к изменяющимся паттернам.

Качество зависит от исходных признаков, поэтому при хорошем представлении признаков XGBoost близок к LSTM по точности.

3.6. NNAR

NNAR характеризуется хорошими значениями R^2 (0.95 и выше) на интервалах более одной минуты.

MAE и RMSE относительно стабильны и показывают хорошие результаты.

MAPE и WAPE также находятся на приемлемом уровне, с улучшением на более длительных интервалах времени.

Использует отложенные лаги как входы, что приближает данную модель к MLP-архитектуре, ориентированной на временные ряды.

Простая архитектура снижает риск переобучения и хорошо усваивает шумную, но структурированную динамику.

В отличие от ARIMA, не требует стационарности и может обучиться нелинейным трансформациям.

3.7. TBATS

TBATS показывает очень слабую производительность (R^2 находится в интервале от отрицательных значений до 0.70), что делает модель неэффективной для данной задачи.

MAE, MSE, RMSE, MAPE и WAPE также показывают очень высокие значения, что подтверждает её плохие результаты.

TBATS предназначен для выраженных сезонных рядов с длинными периодами (например, годовые/недельные циклы). В рассматриваемых данных не наблюдается сезонность.

Использует сложные компоненты (Box-Cox, ARMA, тройную экспоненту), которые приводят к переобучению при использовании на коротком и шумном временном ряде.

Высокая вариативность и отсутствие периодичности делают модель TBATS неэффективной для данной задачи.

Заключение

В работе отмечена актуальность разработки методов и программных средств, предназначенных для решения задачи изменения обводнённости продукции нефтяных скважин. Для решения указанной задачи были рассмотрены следующие методы интеллектуального анализа данных: LSTM, BiLSTM, Prophet, ARIMA, XGBoost, NNAR и TBATS. Выявлены сильные и слабые стороны данных методов. Выполнена реализация указанных методов. Выполнено тестирование реализованных методов с использованием реальных данных, полученных на нефтяных месторождениях. Показано, что при решении поставленной задачи для краткосрочного прогнозирования (от 1 до

5 минут) лучше использовать модели LSTM, BiLSTM и NNAR. Для долгосрочного прогнозирования изменения обводнённости продукции скважин лучше использовать LSTM модель. Также отмечено, что модели Prophet, XGBoost, ARIMA и TBATS не подходят для решения поставленной задачи из-за специфики изменения обводненности потока продукции скважин.

Таким образом, полученные результаты подтверждают целесообразность применения рекуррентных нейросетевых моделей для решения задач прогнозирования в нефтедобывающей отрасли. Дальнейшие исследования могут быть направлены на усовершенствование архитектур нейросетей с учётом специфики процессов обводнённости, а также на интеграцию гибридных подходов, способных повысить точность и надёжность прогнозов. Практическая значимость работы заключается в возможности использования разработанных моделей для оптимизации технологических процессов и повышения эффективности эксплуатации нефтяных скважин.

Список литературы

[Барабанов и др., 2019] Барабанов А.О., Гужов С.В. Прогнозирование тепловой нагрузки на отопление с использованием ИНС // С.О.К. – 2019. – № 11. – С. 28-30.

[Bian et al, 2022] Haihong Bian, Qian Wang, Guozheng Xu, Xiu Zhao. Load forecasting of hybrid deep learning model considering accumulated temperature effect // Energy Reports. – 2022. – Vol. 8. – P. 205-215. – DOI: 10.1016/j.egyr.2021.11.082.

[Dong, 2022] Dong X., Deng S. & Wang D. A short-term power load forecasting method based on k-means and SVM // J Ambient Intell Human Comput. – 2022. – Vol. 13. – P. 5253-5267. – DOI: 10.1007/s12652-021-03444-x.

[Sarker, 2021] Sarker I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. // SN COMPUT. SCI. – 2021. – Vol. 2, 420. – https://doi.org/10.1007/s42979-021-00815-1.

[Hochreiter et al., 1997] Hochreiter S., & Schmidhuber J. Long short-term memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735-1780. – doi: 10.1162/neco.1997.9.8.1735.

[Schuster, Paliwal, 1997] Schuster M., Paliwal K.K. Bidirectional recurrent neural networks // IEEE Transactions on Signal Processing. – 1997. – Vol. 45, No. 11. – P. 2673-2681.

[Taylor et al., 2017] Taylor S.J., Letham B. Forecasting at scale // PeerJ Preprints. – 2017. – Vol. 5. – e3190v2. – DOI: 10.7287/peerj.preprints.3190v2.

[Box et al., 1994] Box G.E.P., Jenkins G.M., Reinsel G.C. Time Series Analysis: Forecasting and Control, 3rd edition // Prentice Hall. – 1994.

[Chen et al., 2016] Chen T., Guestrin C. XGBoost // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – DOI: 10.1145/2939672.2939785.

[Gregor et al., 2014] Gregor K., Danihelka I., Mnih A., Blundell C., Wierstra D. Deep AutoRegressive Networks // Proceedings of the 31st International Conference on Machine Learning (ICML), JMLR: W&CP. – 2014. – Vol. 32.

[De Livera et al., 2011] De Livera A.M., Hyndman R.J., Snyder R.D. Forecasting time series with complex seasonal patterns using exponential smoothing // Journal of the American Statistical Association. – 2011. – Vol. 106, No. 496. – P. 1513-1527.

УДК 004.89
doi: 10.15622/rcai.2025.015

ОБ УСЛОВИИ НЕЗАВИСИМОСТИ ПРИЧИНЫ ОТ КОНТЕКСТА В ДСМ-МЕТОДЕ АВТОМАТИЗИРОВАННОЙ ПОДДЕРЖКИ ИССЛЕДОВАНИЙ

О.П. Шестерникова (*oshesternikova@frccsc.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В статье описывается условие независимости причины от контекста в ДСМ-методе автоматизированной поддержки исследований, предложенное для анализа ситуации множественности причины (существования нескольких достаточных компонентных причин). Приводится определение контекста причины, представлен алгоритм для проверки условия независимости от контекста, а также пример выполнения условия для медицинской задачи установления целесообразности направления на компьютерную томографию у больных хроническим панкреатитом.

Ключевые слова: ДСМ-метод АПИ, причинность, достаточная компонентная причина, множественность причин, хронический панкреатит.

Введение

ДСМ-метод автоматизированной поддержки исследований (ДСМ-метод АПИ) направлен на обнаружение причинно-следственных отношений в данных. Развиваемая в ДСМ-методе АПИ общая теория причинности имеет много общего с концепцией так называемой *детерминированной причинности* в медицине. Обычно детерминированную причинность связывают с моделью достаточной компонентной причины (*Sufficient-Component Cause model, SCC*), в которой факторы риска некоторого заболевания являются «компонентами, составляющими достаточные причины, но сами по себе они недостаточны» [Котеров и др., 2022]. Достаточной для наступления исхода (возникновения заболевания) является их комбинация. Графическое представление такой модели реализуется в виде диаграммы, называемой «пирог причинности» (*Causal Pie*), где сектора соот-

ветствуют необходимым компонентам причины (факторам риска), а весь круг – непосредственно достаточной компонентной причине [Котеров и др., 2022]. Таким образом, в модели достаточной компонентной причины причина – это многофакторное явление [VanderWeele, 2017]. При этом важно отметить, что для исследуемого явления может существовать больше одной достаточной причины.

В ДСМ-методе АПИ такая ситуация называется «множественностью причин» (используется термин философа Д.С. Милля): «неверно, будто каждое единичное следствие должно быть связано с одной только причиной, с одним рядом условий, будто всякое явление может быть произведено лишь одним путём. Часто существует несколько независимых друг от друга способов, при помощи которых можно вызвать одно и то же явление» [Милль, 2011].

Таким образом, развитие средств анализа множественности причин в ДСМ-методе АПИ представляется практически значимой для использования в медицинских исследованиях, в частности, при детерминированном подходе. Одним из таких средств является рассмотрение контекста причины, в который входят другие причины изучаемого эффекта.

1. Определение условия независимости причины от контекста

Под контекстом причины V понимаются другие причины $\{V_1, .., V_m\}$, также являющиеся причинами наличия (или отсутствия) изучаемого эффекта [Финн, 2024]. Контекст для причины V эмпирической предзакономерности (ЭПЗК) в ДСМ-методе АПИ формально определяется как $\bar{V} \rightleftharpoons \{V \mid J_{<\upsilon,2n+1>}\ H_2(V, Y, p, h)\}$, где υ=1, если σ=+ и υ=−1, если σ=−, а предикат $H_2(V,Y,p,h)$ означает, что V является причиной Y для некоторого расширения номер p базы фактов (БФ) в некоторой истории возможных миров с номером h [Финн, 2023].

Условие независимости причины V от своего контекста $\bar{V} = \{V_1, .., V_m\}$ определяется следующим образом:

$$\exists X((V \subset X) \wedge \neg\big((V_1 \subset X) \vee ... \vee (V_m \subset X)\big). \qquad (1.1)$$

Это условие говорит о том, что в исходной БФ должен существовать такой пример X, в котором рассматриваемая причина V содержится но не содержится ни одна другая причина из контекста. То есть если «изолировать» эту причину от других, то эффект сохранится.

Заметим, что, если контекст содержит некоторую причину V_0, которая содержится в исследуемой причине V ($V_0 \subset V$), то условие независимости причины не выполняется, так как $\forall X\ ((V \subset X) \rightarrow (V_0 \subset X))$. Причины, которые не содержатся в других причинах, называются *минимальными* [Гусакова и др., 2016]. Таким образом, для выполнения условия независимости минимальность является необходимым, но не достаточным требованием.

Минимальность является осмысленным требованием с практической медицинской точки зрения: необходимо находить такие достаточные компонентные причины, что «устранение любого из входящего в них условия может быть достаточно, чтобы сделать эту достаточную причину недействующей» [VanderWeele, 2017]. Кроме того, целенаправленный поиск только минимальных причин позволяет использовать некоторые оптимизации для комбинаторного перебора при порождении гипотез о причине в ДСМ-методе АПИ (например, [Забежайло, 2014]).

2. Условия независимости причины от контекста и типы контекста

Определяются 7 возможных типов контекста причины [Шестерникова и др., 2025] на основе трёх возможных условий:

1) причина V_1 из контекста несравнима с рассматриваемой причиной V ($V_1 \parallel V$);
2) причина V_1 из контекста содержится в рассматриваемой причине V ($V_1 \subset V$);
3) причина V_1 из контекста содержит рассматриваемую причину ($V \subset V_1$).

Возможные комбинации приведенных условий соответствуют типам контекста (обозначаются буквами A-G):

A) все причины из контекста несравнимы с рассматриваемой причиной V (для всех причин выполняется только условие 1): $\forall V_1 (V_1 \parallel V)$;

B) все причины из контекста включаются в рассматриваемую причину V (для всех причин выполняется только условие 2): $\forall V_1 (V_1 \subset V)$;

C) все причины из контекста содержатся в рассматриваемой причине V (для всех причин выполняется только условие 3): $\forall V_1 (V \subset V_1)$

D) все причины из контекста сравнимыми с рассматриваемой, но не удовлетворяют случаям B и C (для всех причин выполняется либо условие 2, либо условие 3): $\forall V_1 ((V_1 \subset V) \vee (V \subset V_1)) \,\&\, \exists V_2\ (V_2 \subset V) \,\&\, \exists V_3 (V \subset V_3)$

E) все причины из контекста либо несравнимы с рассматриваемой, либо включаются в неё (для всех причин выполняется либо условие 1, либо условие 2): $\forall V_1 ((V_1 \parallel V) \vee (V \subset V_1)) \,\&\, \exists V_2 (V_2 \parallel V) \,\&\, \exists V_3\ (V \subset V_3)$

F) все причины из контекста либо несравнимы с рассматриваемой, либо её включают (для всех причин выполняется либо условие 1, либо условие 3): $\forall V_1 ((V_1 \parallel V) \vee (V_1 \subset V)) \,\&\, \exists V_2 (V_2 \parallel V) \,\&\, \exists V_3 (V_3 \subset V)$

G) причины из контекста могут быть любыми (для причин выполняется любое из условий 1-3): $\forall V_1 ((V_1 \parallel V) \vee (V_1 \subset V) \vee (V \subset V_1)) \,\&\, \exists V_2 (V_2 \parallel V) \,\&\, \exists V_3 (V_3 \subset V) \,\&\, \exists V_4 (V \subset V_4)$

Очевидно, что условию минимальности причины V удовлетворяют три типа её контекста: A, C и F. Однако, для выполнения условия независимости от контекста, все минимальные причины должны быть дополнительно проверены алгоритмом, приведенным далее.

3. Алгоритм проверки условия независимости причины от контекста

Ниже представлен алгоритм для проверки выполнения условия независимости причины от контекста.

Вход: Причина *V,* которую нужно проверить на выполнение условия

1. Формируем множество предсказаний, которые делает причина *V* Pred(*V*) = {*Z*| *V* ⊂ Z}.
2. Определим множество X1 примеров из исходной базы фактов (БФ), которые объясняет причина *V,* X1 = {*X* | *X* ∈ БФ & *V* ⊂ *X*}.
3. Для каждого предсказания *Z* из Pred(*V*) проверка контекста:
   a. Формируем контекст Ctxt(*V*, *Z*): множество гипотез о причине, которые также делают предсказание *Z.*
   b. Определим множество X2 примеров из БФ, которые объясняют гипотезы о причине из контекста X2 = {*X* | ∀ V_0 ($V_0 \in Ctxt(V, Z)$ & $V_0 \subset X$)}.
   c. Если не существует примера, который объясняет причина *V* и не объясняет контекст (|X1 − X2| = 0), то условие независимости причины для *V* не выполняется.
4. Если проверили все предсказания *Z*, то условие независимости причины для *V* выполняется.

При обнаружении эмпирических закономерностей (ЭЗК) – регулярностей в расширяющихся (динамических) массивах данных – посредством ДСМ-метода АПИ условие независимости проверяется для гипотезы о причине (об отсутствии причины) *V* во всех расширениях массива (и во всех построенных для обнаружения ЭЗК историях возможных миров) [Финн, 2023]. Кроме того, обнаружение ЭЗК подразумевает рассмотрение не всех порождаемых гипотез о причине (об отсутствии причины), а только тех, которые делают верные предсказания и не делают неверных [Финн, 2023].

4. Пример эмпирической закономерности с условием независимости причины от контекста

Ниже представлен пример выполнения условия независимости причины от контекста для гипотезы о причине в задаче определения целесообразности направления на компьютерную томографию (КТ) у больных хроническим панкреатитом [Интеллектуальная система, 2022].

Положительные гипотезы о причине (есть исследуемый эффект, нужно направление на КТ):

(1) Табакокурение И индекс массы тела выше нормы И длительность клинических проявлений меньше года являются основанием направления на КТ.

(2) Отсутствие алкогольной зависимости И длительность не больше пяти лет И сильная боль И сахарный диабет являются основанием направления на КТ.

(3) Индекс массы тела выше нормы И длительность клинических проявлений меньше года И сахарный диабет являются основанием направления на КТ.

верно предсказывают пациента №278

(278) Возраст 50-59 лет И отсутствие алкогольной зависимости И табакокурение И индекс массы тела выше нормы И длительность клинических проявлений меньше года И сильная боль И сахарный диабет.

При этом для каждой из гипотез (1)-(3) существует пациент из исходной базы фактов (БФ), который предсказывается рассматриваемой гипотезой, но не предсказывается остальными двумя, составляющими контекст.

Для (1) (контекст $\bar{V} = \{(2), (3)\}$) существует пациент №115

(115) Возраст 60-69 лет И отсутствие алкогольной зависимости И табакокурение И индекс массы тела выше нормы И длительность клинических проявлений меньше года И сильная боль И отсутствие сахарного диабета.

Для (2) (контекст $\bar{V} = \{(1), (3)\}$) существует пациент №15

(15) Возраст 70-79 лет И отсутствие алкогольной зависимости И отсутствие табакокурение И индекс массы тела норма И длительность клинических проявлений не больше 2-х лет И сильная боль И сахарный диабет.

Для (3) (контекст $\bar{V} = \{(1), (2)\}$) существует пациент №152

(152) Возраст 40-49 лет И алкогольная зависимость И отсутствие табакокурения И индекс массы тела выше нормы И длительность клинических проявлений меньше года И несильная боль И сахарный диабет.

Все приведенные гипотезы имеют в исследовании тип контекста F.

Случаев существования минимальной причины, для которой не выполнялось бы условие независимости, в рассматриваемой задаче не было обнаружено.

Заключение

Представленное в статье условие независимости причины от контекста в ДСМ-методе АПИ выделяет «изолированные» причины в ситуации множественности причин (существует несколько причин изучаемого эффекта). Обнаружение таких причин может быть полезно в медицинских исследованиях, направленных на анализ детерминированной причинности.

Список литературы

[Гусакова и др., 2016] Гусакова С.М., Михеенкова М.А. Интеллектуальный анализ данных как инструмент формирования структуры социума // Научно-техническая информация. Серия 2. – 2016. – № 8. – С. 9-18.

[Забежайло, 2014] Забежайло М.И. Приближенный ДСМ-метод на примерах // Научно-техническая информация. Серия. 2. – 2014. – № 10. – С. 1-12.

[Интеллектуальная система, 2022] Шестерникова О.П., Финн В.К., Лесько К.А., Винокурова Л.В. Интеллектуальная система прогнозирования необходимости применения компьютерной томографии // Искусственный интеллект и принятие решений. – 2022. – № 2. – С. 3-16. – doi: 10.14357/20718594220201.

[Котеров и др., 2022] Котеров А.Н., Ушенкова Л.Н. Критерии причинности в медико-биологических дисциплинах: история, сущность и радиационный аспект. Сообщение 4, часть 2: Иерархия критериев, их критика и иные методы установления причинности // Радиационная биология. Радиоэкология. – 2022. – Т. 62, № 4. – С. 339-398. – doi: 10.31857/S0869803122040051.

[Милль, 2011] Милль Д.С. Система логики силлогистической и индуктивной: Изложение принципов доказательства в связи с методами научного исследования: пер. с англ. / Предисл. и прил. В.К. Финна. – М.: ЛЕНАНД, 2011. – 832 с.

[Финн, 2023] Финн В.К. Об эмпирических закономерностях в ДСМ-методе автоматизированной поддержки исследований // Научно-техническая информация. Серия 2. – 2023. – № 12. – С. 14-33. – doi: 10.36535/0548-0027-2023-12-2.

[Финн, 2024] Финн В.К. Об эмпирических закономерностях ранга r в ДСМ-методе автоматизированной поддержки исследований // Научно-техническая информация. Сер. 2. – 2024. – № 1. – С. 11-33. – doi: 10.36535/0548-0027-2024-01-2.

[Шестерникова и др., 2025] Шестерникова О.П., Финн В.К. О когнитивном интерфейсе интеллектуальных систем, реализующих ДСМ-метод автоматизированной поддержки исследований // Научно-техническая информация. Серия 2. – 2025. – № 2. – С. 10-17. – doi: 10.36535/0548-0027-2025-02-2.

[VanderWeele, 2017] VanderWeele T.J. Invited Commentary: The Continuing Need for the Sufficient Cause Model Today // American journal of epidemiology. – 2017. – Vol. 185(11). – P. 1041-1043. – doi: 10.1093/aje/kwx083.

Секция 3 | МОДЕЛИРОВАНИЕ РАССУЖДЕНИЙ

УДК 004.83
doi: 10.15622/rcai.2025.016

О ВОЗМОЖНОСТИ ПОРОЖДЕНИЯ ОБОСНОВАННЫХ КАУЗАЛЬНЫХ ГИПОТЕЗ В ДСМ-МЕТОДЕ ДЛЯ ЗАДАЧИ НАСЛЕДСТВЕННОЙ БОЛЕЗНИ

Г.С. Вельмакин (*grigoryyii@gmail.com*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В настоящей работе в ДСМ-методе, на примере обратного предиката сходства, будет построен предикат сходства с композицией отношений. Будет показано, что построенный предикат выразим в Граф-АФП – расширения анализа формальных понятий путём добавления отношений между объектами. В конце, посредством Граф-АФП, будет показана связь между предикатом сходства с композицией отношений и логикой описания. Всё вышеописанное позволит адекватно описать порождение каузальных гипотез в задаче изучения причин наследственной болезни в ДСМ-методе и Граф-АФП.

Ключевые слова: ДСМ-метод, предикат сходства с композицией отношений, Граф-АФП, логика описания.

Введение

Порождение обоснованных гипотез о причине той или иной наследственной болезни является актуальной задачей. Для её адекватного решения тем или иным методом необходимо, чтобы средствами используемого

метода можно было выразить композицию таких отношений как «родитель-ребёнок». ДСМ-метод [Финн, 2009], как метод, в котором происходит порождение (посредством индукции (п.п.в. I-го рода) и аналогии (п.п.в. II-го рода)) и обоснование (посредством абдукции) каузальных гипотез, мог бы быть использован для решения указанной задачи, однако в настоящей момент в нём отсутствуют средства, позволяющие схватит идею родословной ветви, как следствие, идею устойчивости паттерна «подобъект C′ есть причина множества свойств D» при наследовании, из-за чего, в частности, используя ДСМ-метод без отношений, в данной задаче, на основе такой БД, в которой есть два независимых родителя с исследуемой болезнью, но ни у их родителей, ни у их детей изучаемой болезни нет, как нет и у других людей из БД, посредством п.п.в. I-го рода мы сделаем грубый вывод, что болезнь наследственная. Настоящая работа призвана заполнить указанный пробел.

В первой части работы будет описана модельная задача наследования болезни. Для неё будет построен предикат сходства с композицией отношений $\widetilde{M}_n^{rel,comp,+}(V',\mathrm{W})$; его «объектами» являются множества упорядоченных троек $\langle \mathrm{C}_1,\mathrm{C}_2,\mathrm{C}_3\rangle$, а операцией сходства $\odot$ является поэлементное сходство объектов, т. е. (для простоты восприятия мы оставили операцию сходства в виде $\cap$):

$$\langle C_1, C_2, C_3\rangle \odot \langle C_4, C_5, C_6\rangle := \langle C_1 \cap C_4, C_2 \cap C_5, C_3 \cap C_6\rangle.$$

Во второй части работы будет показано, что построенный предикат выразим в Граф-АФП, тем самым мы покажем что Граф-АФП также допускает решение этой задачи.

В третьей части работы будет показано, как с использованием оператора int из Граф-АФП мы можем переходить от объекта g к его внутренней структуре C, и тем самым мы установим связь между ЛО и ДСМ-методом, что позволит естественным образом задавать более тонкие правила для предиката сходства с композицией отношений.

1. О построении предиката сходства с композицией отношений в ДСМ-методе

1.1. Предварительные обозначения, понятия и определения

Опишем **алфавит**.

- Константы: (1) Константные символы для объектов (без штрихов) и подобъектов (со штрихами): C, C′, C_1, C'_1, C_2, C'_2, ...; (2) Константные символы для признаков: D, D_1, D_2, ...; (3) Константные символы для бинарных отношений: R, S, R_1, S_1, ...;

- Переменные: (1) Переменные для объектов и подобъектов: V, V′, X, X′, Z, Z′ (быть может, с нижними индексами); (2) Переменные для признаков: U, W, Y (быть может, с нижними индексами); (3) Переменные для бинарных отношений: r, s (быть может, с нижними индексами);
- Логические **операторы Россера-Тюркетта**: J_t, J_f, $J_{\langle \nu,n \rangle}$ и $J_{(\tau,n)}$, где $\nu \in \{+1,-1,0\}$, $n \in \mathbb{N}$;
- Логические связки (внешние): $\neg$, $\wedge$, $\vee$, $\longrightarrow$;
- Вспомогательные символы: (,), **,** **..**

Обозначение.

$$J_{(\nu,n)}\Phi := \bigvee_{i=0}^{n} J_{\langle \nu,i \rangle}\Phi,$$

где Φ – формула.

В выражении $C \Rightarrow_1 D$, которое является внутренней формулой (т.е. принимает истинностные значения $\langle \nu,n \rangle$ и (τ,n), где $\nu \in \{+1,-1,0\}$, $n \in \mathbb{N}$), говорится, что объект C обладает множеством свойств D. В выражении $C' \Rightarrow_2 D$, которое является внутренней формулой, говорится, что подобъект C' есть причина (каузально вынуждает) множества свойств D.

В выражении $C_1 R C_2$, которое является внешней формулой (т.е. принимает классические истинностные значения t и f), говорится, что объекты C_1 и C_2 находятся в отношении R, т.е. $\langle C_1,C_2 \rangle \in R$.

Определим три универсума:

- $U^{(1)}$ – множество элементов, из которых образуются объекты и подобъекты;
- $U^{(2)}$ – множество заданных свойств, присущих объектам;
- $U^{(rel)}$ – множество бинарных отношений между объектами. Всех их мы считаем антирефлексивными и асимметричными.

1.2. Предикат сходства с композицией отношений

1.2.1. Модельная задача. Опишем модельную задачу, в которой будет учитываться композиция между отношениями, а именно будут изучаться множество людей, для каждого из которых хотя бы в одной родословной ветви сохраняется изучаемая болезнь. Для данной задачи ниже будет построен соответствующий предикат сходства.

Пусть у нас есть множество G, состоящее из непересекающихся множеств:

$$G = G_1 \cup G_2 \cup G_3,$$

где G_1 множество мужчин и женщин, G_2 множество детей людей из G_1, G_3 множество детей людей из G_2. Также даны два отношения R_1 и R_2, а именно, выражение $C_1 R C_2$ означает, что человек $C_1 \in G_1$ есть родитель

человека $C_2 \in G_2$, а выражение C_2RC_3 означает, что человек $C_2 \in G_2$ есть родитель человека $C_2 \in G_2$. Сами люди C могут быть выражены, например, в виде ДНК, т.е. посредством трёхмерных помеченных графов.

Пусть мы построили положительный предикат сходства с композицией отношений $\widetilde{M}_n^{rel,comp,+}$. Тогда выполнение предиката $\widetilde{M}_n^{rel,comp,+}(C',\mathrm{D})$ означает, что нашлись $\geqslant 2$ человека из G_3, имеющих общую часть ДНК C' и общие симптомы D, для каждого из которых найдётся хотя бы один родитель из G_2, причём количество всех найденных родителей $\geqslant 2$, и каждый из них также имеет общую часть ДНК C' и общие симптомы D, и у каждого найденного родителя из G_2 также найдётся хотя бы один родитель из множества G_1, причём количество всех найденных родителей $\geqslant 2$, и каждый из них также имеет общую часть ДНК C' и общие симптомы D.

1.2.2. Положительный предикат сходства с отношениями. Положительный предикат сходства с композицией отношений $\widetilde{\boldsymbol{M}}_{\boldsymbol{n}}^{\boldsymbol{rel,comp,+}}$ будет строиться на основе положительного предиката сходства с отношениями $\widetilde{\boldsymbol{M}}_{\boldsymbol{n}}^{\boldsymbol{rel,+}}$, который в свою очередь будет строиться на основе предиката сходства без отношений $\widetilde{\boldsymbol{M}}_{\boldsymbol{n}}^{\boldsymbol{+}}$. Для наглядности мы выберем конкретный предикат сходства без отношений, а именно обратный положительный предикат сходства $\widetilde{\boldsymbol{M}}_{\boldsymbol{in,n}}^{\boldsymbol{+}}$, отрицательная версия которого была представлена в [Финн, 2009] (перепечатана из сборника 1999 года). Такой выбор обусловлен тем, что ниже мы будем обобщать результат выразимости предиката сходства в Граф-АФП из работы [Кузнецов, 2006], в которой был использован именно он. Конечно, идеи, которые будут приведены ниже, будут справедливы и для остальных предикатов сходства, а также для их усилений, например, для запрета на контрпримеры (ниже мы также дадим это усиление). Так, если мы хотим подчеркнуть, что идея применима для всех предикатов сходства, то мы будем писать просто $\widetilde{\boldsymbol{M}}_{\boldsymbol{n}}^{\boldsymbol{+}}$.

Предварительно сделаем пояснения относительно индексов у константных символов, а именно, константа «$C^{1,j}$» мыслится стоящей в области определения (индекс «1») отношения R_j (индекс «j»), а константа «$C^{2,j}$» – в области значения (индекс «2») отношения R_j (индекс «j»). Аналогичные соображения верны для констант D и переменных X и Y.

Опишем части предиката сходства с отношениями $\widetilde{M}_n^{rel,+}$.

Часть

$$\Phi_R^{1,j}(X_1^{1,j},\dots,X_{k_{1,j}}^{1,j},R_j,X_1^{2,j},\dots,X_{k_{2,j}}^{2,j}) :=$$

$$\forall X^1 \exists X^2 \Big(\bigvee_{l=1}^{k_{1,j}}(X^1 = X_l^{1,j}) \to \bigvee_{l=1}^{k_{2,j}}(X^2 = X_l^{2,j}) \wedge X^1 R_j X^2\Big),$$

а также симметричная ей $\Phi_R^{2,j}$ заменяет условие рассмотрения кортежей объектов $\langle C_1,C_2\rangle$, которые являются «объектами» для предиката сходства $\widetilde{M}_n^{rel,+}$, а операцией сходства $\odot$ является поэлементное сходство объектов, т.е.:

$$\langle C_1, C_2\rangle \odot \langle C_3, C_4\rangle := \langle C_1 \cap C_3, C_2 \cap C_4\rangle.$$

Вместо этого мы по-отдельности для множеств первых и вторых компонент строим предикат сходства, после чего соединяем их вместе, в частости, этим условием.

Теперь опишем, как нужно изменить $(ЭЗ)^+$ из предиката $\widetilde{M}_n^+$, если объекты $X_1^{1,j}, \dots, X_{k_{1,j}}^{1,j}$ берутся из области определения отношения R_j. Т.к. теперь на объекты накладывается дополнительное условие в виде $\Phi_R^{1,j}$, имеющий вид $\Phi_R^{1,j} := \forall X^1 \exists X^2 \Phi'^{1,j}_R$, то, представив $(ЭЗ)^+$ в виде $\Psi^+ := \forall X \Psi'^+$, мы видим, что его необходимо заменить на:

$$\Psi^{1,+} := \forall X \exists X^2 \left(\bigvee_{l=1}^{k_{2,j}} (X^2 = X_l^{2,j}) \wedge X R_j X^2 \to \Psi'^+\right),$$

получив $(ЭЗ)^{1,+}$ (конечно, т.к. само Ψ'^+ имеет вид импликации, то, используя эквиваленции классической логики, мы можем поместить описанную добавочную формулу в посылку импликации Ψ'^+). Симметрично добавим

$$\exists X^1 \left(\bigvee_{l=1}^{k_{1,j}} (X^1 = X_n^{1,j}) \wedge X^1 R_j X\right)$$

в $(ЭЗ)^+$, получив $(ЭЗ)^{2,+}$ $\Psi^{2,+}$.

Примечание. Представленная здесь идея верна и для усилений предикатов сходства. Например, для запрета на контрпримеры, который имеет вид $(b)_n^+ := \forall X (b)'^+_n$, мы получаем (она добавляется к $\widetilde{N}_n^{1,+}$ (см. ниже) через $\wedge$):

$$(b)_n^{1,+} := \forall X \exists X^2 \left(\bigvee_{l=1}^{k_{2,j}} (X^2 = X_l^{2,j}) \wedge X R_j X^2 \to (b)'^+_n\right).$$

Теперь определим предикат $\widetilde{N}_n^{1,+}$ как предикат $\widetilde{M}_n^+$ без кванторной приставки (т.к. в дальнейшем мы будем иметь дело с обратным предикатом сходства, то кванторная приставка будет иметь вид «$\exists X_1 \dots \exists X_k \exists Y_1 \dots \exists Y_k$»), с дополнительным условием $\Phi_R^{1,j}$ и $\Psi^{1,+}$ вместо Ψ^+. Симметрично определим $\widetilde{N}_n^{2,+}$.

Т.к. ранее мы договорились рассматривать «объекты» $\langle C_1,C_2\rangle$ покомпонентно, то теперь, помимо объединения посредством $\Phi_R^{1,j}$ и $\Phi_R^{2,j}$, их необходимо соединить кванторно (в $\widetilde{N}_n^{1,+}$ (симметрично для $\widetilde{N}_n^{2,+}$), помимо переменных до R_j, которые соответствуют переменным для $\widetilde{M}_n^+$, за счёт $\Phi_R^{1,j}$ появились переменные после R_j).

$$\widetilde{M}'^{\,rel,+}_{in,n}(V',W,R_j) :=$$
$$\exists k_{1,j}\exists X_1^{1,j}\dots\exists X_{k_{1,j}}^{1,j}\exists Y_1^{1,j}\dots\exists Y_{k_{1,j}}^{1,j}$$
$$\exists k_{2,j}\exists X_1^{2,j}\dots\exists X_{k_{2,j}}^{2,j}\exists Y_1^{2,j}\dots\exists Y_{k_{2,j}}^{2,j}$$
$$(\widetilde{N}_{in,n}^{1,+}(V',W,k_{1,j},X_1^{1,j},\dots,X_{k_{1,j}}^{1,j},Y_1^{1,j},\dots,Y_{k_{1,j}}^{1,j},R_j,X_1^{2,j},\dots,X_{k_{2,j}}^{2,j})\wedge$$
$$\wedge\widetilde{N}_{in,n}^{2,+}(V',W,k_{2,j},X_1^{2,j},\dots,X_{k_{2,j}}^{2,j},Y_1^{2,j},\dots,Y_{k_{2,j}}^{2,j},R_j,X_1^{1,j},\dots,X_{k_{1,j}}^{1,j})).$$

Теперь определим $\widetilde{M}_n^{rel,+}$ как

$$\widetilde{M}_n^{rel,+}(V',W) := \widetilde{M}'^{\,rel,+}_{\,n}(V',W,R_1)\wedge\widetilde{M}'^{\,rel,+}_{\,n}(V',W,R_2).$$

На основе положительного предиката $\widetilde{M}_n^{rel,+}$ может быть построен отрицательный предикат путём замены +1 на -1 в $J_{\langle+1,n\rangle}$ (или могут браться другие комбинации предикатов сходства с их усилениями, например, как предложено в работе [Финн, 2009], взять простой положительный и обратный отрицательный), что позволит породить п.п.в. I-го рода с отношениями (для наглядности выпишем только один):

$$(I)_n^{rel,+}\quad\frac{J_{(\tau,n)}(V'\Rightarrow_2 W)\quad\widetilde{M}_n^{rel,+}(V',W)\wedge\neg\widetilde{M}_n^{rel,-}(V',W)}{J_{\langle+1,n+1\rangle}(V'\Rightarrow_2 W)}.$$

Аналогичные соображения для п.п.в. I-го рода верны и для предиката сходства с композицией отношений.

1.2.3. Положительный предикат сходства с композицией отношений. Теперь перейдём к описанию частей $\widetilde{M}_n^{rel,comp,+}$.

Вместо части

$$\exists k_{2,1}\exists X_1^{2,1}\dots\exists X_{k_{2,1}}^{2,1}\exists Y_1^{2,1}\dots\exists Y_{k_{2,1}}^{2,1}$$
$$\exists k_{1,2}\exists X_1^{1,2}\dots\exists X_{k_{1,2}}^{1,2}\exists Y_1^{1,2}\dots\exists Y_{k_{1,2}}^{1,2}$$
$$(k_{2,1}=k_{1,2}\wedge\bigwedge_{l=1}^{k_{2,j}}(X^2=X_l^{2,j})\wedge\dots),$$

в которой говорится, что искомые объекты $C_1^{2,1},\dots,C_{k_{2,1}}^{2,1}$ из области значения отношения R_1 и объекты $C_1^{1,2},\dots,C_{k_{1,2}}^{1,2}$ из области определения отношения R_2, должны совпадать, мы выберем верхнюю кванторную приставку (можно было и нижнюю).

В части

$$\Phi_R^{comp}(X_1^{2,2},\dots,X_{k_{2,2}}^{2,2},R_2,X_1^{2,1},\dots,X_{k_{2,1}}^{2,1},R_1,X_1^{1,1},\dots,X_{k_{1,1}}^{1,1}) :=$$

$$\forall X^3\exists X^2\exists X^1(\bigvee_{l=1}^{k_{2,2}}(X^3=X_l^{2,2})\to\bigvee_{l=1}^{k_{2,1}}(X^2=X_l^{2,1})\wedge X^2R_2X^3\wedge$$

$$\wedge\bigvee_{l=1}^{k_{1,1}}(X^1=X_l^{1,1})\wedge X^1R_1X^2),$$

схватывается идея родословной ветви. Заметим, что в данном условии уже содержится условие $\Phi_R^{2,2}$.

Распишем, как нужно изменить $(ЭЗ)^+$ для каждой из области G:

(1) Для области G_1 необходимо добавить

$$\exists X^2(\bigvee_{l=1}^{k_{2,1}}\left(X^2=X_l^{2,1}\right)\wedge XR_2X^2),$$

чтобы получить нужный вид ограничения $\Phi_R^{1,1}$.

(2) Для области G_2 необходимо добавить как

$$\exists X^1(\bigvee_{l=1}^{k_{1,1}}\left(X^1=X_l^{1,1}\right)\wedge X^1R_2X),$$

чтобы получить нужный вид ограничения $\Phi_R^{2,1}$, так и

$$\exists X^3(\bigvee_{l=1}^{k_{2,2}}\left(X^2=X_l^{2,2}\right)\wedge XR_2X^3),$$

чтобы получить нужный вид ограничения $\Phi_R^{1,2}$.

(3) Для области G_3 необходимо добавить

$$\exists X^2\exists X^1(\bigvee_{l=1}^{k_{2,1}}(X^2=X_l^{2,1})\wedge X^2R_2X\wedge\bigvee_{l=1}^{k_{1,1}}(X^1=X_l^{1,1})\wedge X^1R_1X^2),$$

чтобы получить нужный вид ограничения Φ_R^{comp}. Заметим, что при добавлении мы также получаем вид ограничения $\Phi_R^{2,2}$.

Формулы $\widetilde{N}_n^{1,+}$, $\widetilde{N}_n^{2,+}$ и $\widetilde{N}_n^{3,+}$ определяются как выше; отметим только, что в $\widetilde{N}_n^{2,+}$ нужно добавить как $\Phi_R^{2,1}$ так и $\Phi_R^{1,2}$, а в $\widetilde{N}_n^{3,+}$ достаточно добавить Φ_R^{comp} (т.к. $\Phi_R^{2,2}$ автоматически будет выполнено).

В части

$$\widetilde{M}'^{\,rel,comp,+}_{in,n}(V',W,R_1,R_2) :=$$
$$\exists \mathrm{k}_{1,1} \exists \mathrm{X}_1^{1,1} \dots \exists \mathrm{X}_{\mathrm{k}_{1,1}}^{1,1} \exists \mathrm{Y}_1^{1,1} \dots \exists \mathrm{Y}_{\mathrm{k}_{1,1}}^{1,1}$$
$$\exists \mathrm{k}_{2,1} \exists \mathrm{X}_1^{2,1} \dots \exists \mathrm{X}_{\mathrm{k}_{2,1}}^{2,1} \exists \mathrm{Y}_1^{2,1} \dots \exists \mathrm{Y}_{\mathrm{k}_{2,1}}^{2,1}$$
$$\exists \mathrm{k}_{2,2} \exists \mathrm{X}_1^{2,2} \dots \exists \mathrm{X}_{\mathrm{k}_{2,2}}^{2,2} \exists \mathrm{Y}_1^{2,2} \dots \exists \mathrm{Y}_{\mathrm{k}_{2,2}}^{2,2}$$
$$\widetilde{\mathrm{N}}^{1,+}_{\mathrm{in,n}}(\mathrm{V}', \mathrm{W}, \mathrm{k}_{1,1}, \mathrm{X}_1^{1,1}, \dots, \mathrm{X}_{\mathrm{k}_{1,1}}^{1,1}, \mathrm{Y}_1^{1,1}, \dots, \mathrm{Y}_{\mathrm{k}_{1,1}}^{1,1}, \mathrm{R}_1, \mathrm{X}_1^{2,1}, \dots, \mathrm{X}_{\mathrm{k}_{2,1}}^{2,1}) \wedge$$
$$\wedge \widetilde{N}^{2,+}_{in,n}(V', W, k_{2,1}, X_1^{2,1}, \dots, X_{k_{2,1}}^{2,1}, Y_1^{2,1}, \dots, Y_{k_{2,1}}^{2,1},$$
$$R_1, X_1^{1,1}, \dots, X_{k_{1,1}}^{1,1}, R_2, X_1^{2,2}, \dots, X_{k_{2,2}}^{2,2}) \wedge$$
$$\wedge \widetilde{N}^{3,+}_{in,n}(V', W, k_{2,2}, X_1^{2,2}, \dots, X_{k_{2,2}}^{2,2}, Y_1^{2,2}, \dots, Y_{k_{2,2}}^{2,2},$$
$$R_2, X_1^{2,1}, \dots, X_{k_{2,1}}^{2,1}, R_1, X_1^{1,1}, \dots, X_{k_{1,1}}^{1,1}))$$

говорится об устойчивости паттерна C′⇒$_2$D при наследовании.

Теперь определим $\widetilde{M}^{rel,comp,+}_{in,n}(V',W)$ как

$$\widetilde{M}^{rel,comp,+}_{in,n}(V',W) := \widetilde{M}'^{\,rel,comp,+}_{in,n}(V',W,R_1,R_2).$$

2. О выразимости предиката сходства с композицией отношений в Граф-АФП

2.1. Предварительные обозначения, понятия и определения

Анализ формальных понятий (АФП) – ветвь алгебраической теории решёток, метод анализа данных и модель машинного обучения, основанного на отношениях поглощения (порядка общности) [Kuznetsov, 2019], [Ganter et al., 1999]. Одним из расширений АФП является Граф-АФП [Ferré et al., 2020].

Определение 2.1. **Графовым контекстом** в Граф-АФП называют тройку K=(G,M,I), где G – множество объектов, M={M^1,M^2,...,M^n} – множество признаков, разбитых на подмножества M^i, I⊆G*×M – отношение инцидентности между кортежами объектов и признаками, где i местный кортеж может иметь признак только из множества M^i. $\mathcal{V}$ есть множество переменных, $G \subseteq \mathcal{V}$.

Определение 2.2. **Спроектированный графовый узор** (СГУ) есть пара $Q = (\bar{x}, P)$, где $P \subseteq \mathcal{V}^* \times M$ есть **графовый узор** (ГУ), а $\bar{x} \in \mathcal{V}^*$ есть кортеж переменных, который мы будем называть **проецирующим кортежом**.$|Q| := |\bar{x}|$ есть **местность** СГУ.

Мы будем **обозначать** множество СГУ через $\mathcal{Q}$, а через $\mathcal{Q}_k$ подмножество k-СГУ, т.е. СГУ, имеющих местность k.

Определение 2.3. k местным **отношением объектов** R, т.е. |R|=k, будем называть подмножество k местных кортежей, т.е. $R \subset G^k$.

Мы будем **обозначать** через $\mathcal{R}$ множество всевозможных отношений, а через $\mathcal{R}_k$ подмножество k местных отношений.

Для произвольных $R \in \mathcal{R}_k$ и $Q \in \mathcal{Q}_k$ определим операторы и int и ext:

$$\text{int}(R) := \cap_q \{Q_K(\bar{g}) \mid \bar{g} \in R\},$$

$$\text{ext}(Q) := \{\bar{g} \in G^k \mid Q \subseteq_q Q_K(\bar{g})\},$$

где СГУ $Q_K(\bar{g}) := (\bar{g}, I)$ есть **описание кортежа объектов** $\bar{g} \in G^*$.

В дальнейшем мы ограничимся случаем графового контекста с $M=\{M^1,M^2\}$. Договоримся вместо операторов ext и int писать просто $\cdot^I$, если $K=(G,\{M^1,M^2\},I)$ – графовый контекст. Также введём два оператора:

- Оператор $\partial_{\bar{x}}$, действие которого на СГС определим как:

$$\partial_{\bar{x}}(\bar{z}, P) := \{w \mid (\bar{x}, w) \in P\}.$$

- Оператор $\int d\bar{x}$, действие которого на $A=\{w_1,...,w_n\} \subseteq M^k$ определим как:

$$\int A d\bar{x} := (\bar{x}, \{(\bar{x}, w) \mid w \in A\}),$$

где $|\bar{x}| = k$.

Утверждение 2.1. $\partial_{\bar{x}}(\bar{z}, P_1) \cap \partial_{\bar{x}}(\bar{z}, P_1) = \partial_{\bar{x}}((\bar{z}, P_1) \cap_q (\bar{z}, P_2))$.

Утверждение 2.2. $\int A d\bar{x} \subseteq_q (\bar{x}, P) \Leftrightarrow A \subseteq \partial_{\bar{x}}(\bar{x}, P)$.

Договоримся, если не оговорено противное, считать, что если мы пишем $\partial_{\bar{x}} R^I$, где $R \in \mathcal{R}_k$ и $|\bar{x}| = |R|$, то $R^I = (\bar{x}, P)$.

Предварительно переведём множество бинарных отношений $U^{(rel)}$ в множество $U^{*(rel)}$, элементами которого будут такие объекты R*, что $C^1 R C^2 \Leftrightarrow (C^1, C^2) \Longrightarrow_{rel} R^*$, где $\Longrightarrow_{rel} \subseteq \mathcal{P}(U^{(1)}) \times \mathcal{P}(U^{(1)}) \times \mathcal{P}(U^{*(rel)})$.

Примем предположение о том, что примеры являются либо положительными, либо отрицательными относительно каждого функционального (целевого) признака, т.е. элемента из $U^{(2)}$. Эту ситуацию можно представить следующим образом:

- Графовый контекст структуры и отношений $K_M(G,\{M^1,M^2\},I)$: (1) M^1 есть множество структурных признаков, $M^1=U^{(1)}$, т.е. для объекта $g \in G$ множество $\partial_{(x)}\{(g)\}^I$ есть множество элементов, из которых он состоит, т.е. $\partial_{(x)}\{(g)\}^I = C$ для некоторого $C \in \mathcal{P}(U^{(1)})$; (2) M^2 есть множество отношений, $M^2=U^{*(rel)}$, т.е. для упорядоченной пары $(g^1,g^2) \in G \times G$ множество $\partial_{(x^1,x^2)}\{(g^1, g^2)\}^I$ есть множество отношений, в которых состоит пара (g^1,g^2), т.е. $\partial_{(x^1,x^2)}\{(g^1, g^2)\}^I = \{R_1^*, \dots, R_k^*\}$ для некоторых $R_1^*, \dots, R_k^* \in U^{*(rel)}$, где $C^1 = \partial_{(x^1)}\{(g^1)\}^I$, $C^2 = \partial_{(x^2)}\{(g^2)\}^I$ и $C^1 R_1 C^2$, ..., $C^1 R_k C^2$.

- Формальный контекст свойств $K_P(G,P,J)$, где P есть множество свойств, $P=U^{(2)}$ и $J=\{+,-\}$, т.е. для объекта $g \in G$ объектное содержание $\{g\}^+$

есть множество свойств, которыми он обладает, а объектное содержание $\{g\}^-$ есть множество свойств, которыми он не обладает, т.е., если $C = \partial_{(x)}\{(g)\}^I$ и $J_{(+1,n)}(C \Rightarrow_1 D)$, то $\{g\}^+$=D, а если $J_{(-1,n)}(C \Rightarrow_1 D)$, то $\{g\}^-$=D для некоторого $D \in \mathcal{P}(U^{(2)})$.

Пусть A – множество, а P, Q – бинарные отношения. Тогда определим:

- $A \circ Q := \{\langle x,y\rangle \mid x \in A \wedge xQy\}$;
- $P \circ A := \{\langle x,y\rangle \mid y \in A \wedge xPy\}$;
- $P \circ A \circ Q := \{\langle x,y\rangle \mid \exists z(z \in A \wedge xPz \wedge zQy)\}$.

2.2. Теорема о выразимости обратного положительного предиката сходства с композицией отношений в Граф-АФП

Сформулируем теорему, которая будет являться обобщением теоремы из [Кузнецов, 2006]. Договоримся, что мы не различаем объект g и одноместный кортеж (g) когда из контекста ясно, что должно быть.

Теорема 2.1 ((+)-обратный метод с композицией отношений через соответствия Галуа). Для произвольных $V' \subseteq M^1$, $W \subset P$, $V' \neq \emptyset$, $W \neq \emptyset$, следующие два утверждения эквивалентны:

$$\widetilde{M}_{in,n}^{rel,comp,+}(V', W).$$

Имеет место $|E_i| \geqslant 2$ и $\partial_{(x^i)} E_i^I = V'$ и $E_i^+ = W$, где $i \in \{1,2,3\}$, а также $E_1 \subseteq (\int V' d(x))^I \cap A$ и $E_2 \subseteq (\int V' d(x))^I \cap B \cap C$ и $E_3 \subseteq (\int V' d(x))^I \cap D$, где $A=\mathrm{dom}(R_1 \circ E_2)$, $B=\mathrm{rng}(E_1 \circ R_1)$, $C=\mathrm{dom}(R_2 \circ E_3)$ и $D=\mathrm{rng}(E_1 \circ (R_1 \circ E_2 \circ R_2))$, а E_i находятся из решения следующей теоретико-множественной системы уравнений:

1. $E_1 = W^+ \cap A$.
2. $E = W^+ \cap B \cap C$.
3. $E_1 = W^+ \cap D$.

Если мы добавим запрет на контрпримеры, то во вторую часть теоремы выше к выполняемым требованиям мы должны добавить $A \cap (\int V' d(x))^I \subseteq W^+$ и $B \cap C \cap (\int V' d(x))^I \subseteq W^+$ и $D \cap (\int V' d(x))^I \subseteq W^+$.

3. Об усилении предиката сходства с композицией отношений средствами ЛО

Логика описания (ЛО) (описательная логика, дескрипционная логика) является одним из языков представления знаний, позволяющая создавать онтологии [Rudolph, 2011], [Baader et al., 2017].

Сделаем важное **наблюдение**. В ДСМ-методе мы имеем дело с внутренней структурой C объекта g, однако средств ДСМ-метода недостаточно чтобы перейти к самому объекту g; в ЛО мы имеем дело с объектом g, однако средств ЛО недостаточно чтобы перейти к его внутренней структуре C. АФП, через оператор int, позволяет переходить от g к C, и наобо-

рот (в силу уникальности внутренних структур). Если внутреняя структура C имеет вид, отличный от множества, необходимо воспользоваться оператором int из [Ferré, 2023].

Запишем $\Phi_R^{1,j}$ с использованием Граф-АФП и ЛО (т.к. интерпретация $\mathcal{I}$ фиксирована, аксиомы ЛО принимают классические истинностные значения t и f). Предварительно необходимо произвести перевод отношений R из множеств упорядоченных пар вида $\langle C_1, C_2 \rangle$ в множество упорядоченных пар вида $\langle g_1, g_2 \rangle$.

$$\Phi_R^{1,j}(X_1^{1,j}, \dots, X_{k_{1,j}}^{1,j}, R_j, X_1^{2,j}, \dots, X_{k_{2,j}}^{2,j}) := \exists g_1^{1,j} \dots \exists g_{k_{1,j}}^{1,j} \exists g_1^{2,j} \dots \exists g_{k_{2,j}}^{2,j}$$

$$(g_1^{1,j} \in \mathrm{G} \wedge \partial_{(g_1^{1,j})}\{(g_1^{1,j})\}^I = X_1^{1,j} \wedge \dots \wedge g_{k_{1,j}}^{1,j} \in \mathrm{G} \wedge \partial_{(g_{k_{1,j}}^{1,j})}\{(g_{k_{1,j}}^{1,j})\}^I = X_{k_{1,j}}^{1,j} \wedge$$

$$\wedge g_1^{2,j} \in \mathrm{G} \wedge \partial_{(g_1^{2,j})}\{(g_1^{2,j})\}^I = X_1^{2,j} \wedge \dots \wedge g_{k_{2,j}}^{2,j} \in \mathrm{G} \wedge \partial_{(g_{k_{2,j}}^{2,j})}\{(g_{k_{2,j}}^{2,j})\}^I = X_{k_{2,j}}^{2,j} \wedge$$

$$\wedge J_t(\{g_1^{1,j}, \dots, g_{k_{1,j}}^{1,j}\} \sqsubseteq \exists R_j.\{g_1^{2,j}, \dots, g_{k_{2,j}}^{2,j}\})).$$

Мы можем усилить аксиому ЛО, заменив «$\exists R_j$» на, например, «$\geqslant 2R_j$», что будет означать, что мы смотрим на тех людей из G, у которых как минимум два ребёнка унаследовали изучаемую болезнь.

Заключение

Результаты, полученные в настоящей работе имеют как локальный характер, а именно показано, что средствами ДСМ-метода и Граф-АФП можно порождать обоснованные гипотезы о причине в задаче наследственной болезни, так и глобальный. Последнее выражается, во-первых, в том, что мы построили предикат сходства с отношениями, как следствие, теперь средствами ДСМ-метода можно решать задачи в которых без учёта отношений мы бы получали неправильный результат. Во-вторых, мы обогатили Граф-АФП. В-третьих, мы решили задачу объединения ДСМ-метода и ЛО.

В дальнейшем автор планирует создать необходимую программу, в которой было бы реализовано всё то, что было описано в настоящей работе, и провести следующее сравнение: насколько лучше будет проявлять себя ДСМ-метод, если в модельной задаче мы будем учитывать наследственность, а именно (рис. 1), сравнить случай 1 (ДСМ-метод без отношений), случай 2 и случай 3 (ДСМ-метод с одним отношением), случай 4 (ДСМ-метод с двумя отношениями) и случай 5 (ДСМ-метод с композицией двух отношений).

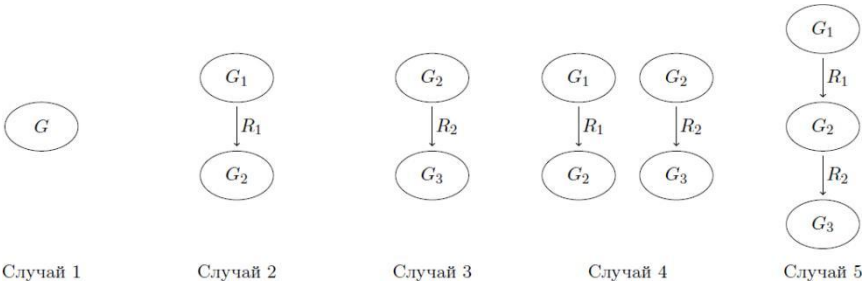


Рис. 1

Благодарности. Автор выражает благодарность д.т.н., профессору, В.К. Финну за указание, что построенные предикаты являются предикатами сходства; их «объектами» являются кортежи объектов, а их сходством является поэлементное сходство объектов.

Список литературы

[Кузнецов, 2006] Кузнецов С.О. ДСМ-метод на языке соответствий Галуа // Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2006. – №. 12. – С. 1-7.

[Финн, 2009] Финн В.К. Синтез познавательных процедур и проблема индукции // Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2009. – № 6. – С. 1-37.

[Baader et al., 2017] Baader F. et al. Introduction to description logic. – Cambridge University Press, 2017.

[Ferré et al., 2020] Ferré S., Cellier P. Graph-FCA: An extension of formal concept analysis to knowledge graphs // Discrete applied mathematics. – 2020. – Vol. 273. – P. 81-102.

[Ferré, 2023] Ferré S. Graph-FCA Meets Pattern Structures //International Conference on Formal Concept Analysis. – Cham: Springer Nature Switzerland, 2023. – P. 33-48.

[Ganter et al., 1999] Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. – 1999.

[Kuznetsov, 2019] Kuznetsov S.O. Ordered Sets for Data Analysis // arXiv preprint arXiv:1908.11341. – 2019.

[Rudolph, 2011] Rudolph S. Foundations of description logics // Reasoning Web International Summer School. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. – P. 76-136.

УДК 007.5:519.816:681.3.016
doi: 10.15622/rcai.2025.017

ПРИМЕНЕНИЕ НЕЙРОСЕТЕВОГО ПОДХОДА И ПАРАЛЛЕЛЬНОЙ ОБРАБОТКИ ИНФОРМАЦИИ ДЛЯ ОПРЕДЕЛЕНИЯ ВЫПОЛНИМОСТИ ФОРМУЛ ЛОГИКИ ВЕТВЯЩЕГОСЯ ВРЕМЕНИ НА СТРУКТУРАХ КРИПКЕ[1]

А.П. Еремеев (*eremeev@appmat.ru*)
Н.Ю. Филинов (*filinov.n@yandex.ru*)

Национальный исследовательский университет «МЭИ», Москва

В работе представлен подход к задаче проверки выполнимости формул темпоральной вычислительной древовидной логики CTL на структурах Крипке с использованием графовых нейронных сетей. Выполнимость формулы формализуется как задача бинарной классификации по паре (формула CTL, структура Крипке). Предложена архитектура модели, объединяющая графовую и формульную части с последующей классификацией. Осуществлена генерация синтетического обучающего набора данных с автоматической разметкой с использованием классического алгоритма model checking. Представлены результаты экспериментов, подтверждающие высокую точность модели и её преимущество по скорости работы по сравнению с традиционными методами. Работа выполняется в рамках разработки инструментальных средств для интеллектуальных систем поддержки принятия решений реального времени.

Ключевые слова: искусственный интеллект, темпоральная логика, model checking, структура Крипке, графовая нейросеть, реальное время.

Введение

Верификация свойств систем с помощью формальных логик остаётся ключевым направлением в теории программирования и автоматической проверке моделей. Одним из наиболее используемых формализмов в этой

[1] Работа выполнена при финансовой поддержке РНФ (проект № 24-11-00285), https://rscf.ru/project/24-11-00285/.

области является вычислительная древовидная логика (Computational Tree Logic, CTL), формулы которой интерпретируются на структурах Крипке [Clarke et al., 1981], [Kupferman et al., 2000]. Задача проверки выполнимости (satisfiability) формул CTL заключается в определении, существует ли структура Крипке, на которой формула выполняется. Эта задача, несмотря на свою выразительность, остаётся алгоритмически сложной.

С развитием методов машинного обучения, в частности, нейросетевых архитектур, возникает естественный вопрос: можно ли использовать нейросети для решения задач логического вывода, таких как выполнимость формул? Особенно интересным представляется применение графовых нейросетей (Graph Neural Network, GNN), так как структуры Крипке по сути являются помеченными графами.

В данной работе рассматривается подход, при котором задача определения выполнимости CTL-формул моделируется как задача бинарной классификации на графах. Предлагается способ кодирования формул и моделей, рассматриваются возможные архитектуры нейросетей, и приводятся результаты экспериментов, демонстрирующие перспективность подхода.

Данные исследования продолжают исследования и разработки по темпоральной логике ветвящегося времени, описанные в работах [Еремеев и др., 2011; 2017; 2023].

1. Логика CTL

Логика CTL является *темпоральной логикой ветвящегося времени.* В этой логике допустимы только формулы, в которых каждый темпоральный оператор *X, U, F* и *G* (характеризующий некоторое вычисление) предваряется квантором пути – *A* или *E*, что превращает любую темпоральную формулу пути в формулу, характеризующую состояние. На рис. 1 представлена схема вложенности логик, где линейные темпоральные логики LTL [Pnueli, 1977] и CTL [Ben-Ari et al., 1983], [Clarke et al., 1980] являются подмножеством более широкого класса логики CTL*.

Грамматика CTL:

$$\phi ::= \bot \mid \top \mid p \mid (\neg\phi) \mid (\phi \wedge \phi) \mid (\phi \vee \phi) \mid (\phi \Rightarrow \phi) \mid (\phi \Leftrightarrow \phi)$$
$$\mid \quad \mathbf{AX}\,\phi \mid \mathbf{EX}\,\phi \mid \mathbf{AF}\,\phi \mid \mathbf{EF}\,\phi \mid \mathbf{AG}\,\phi \mid \mathbf{EG}\,\phi \mid \mathbf{A}\,[\phi\,\mathbf{U}\,\phi] \mid \mathbf{E}\,[\phi\,\mathbf{U}\,\phi].$$

Синтаксическое значение кванторов пути в CTL:

- *A* означает "по всем путям" (неизбежно);
- *E* означает "по крайней мере (существует) один путь" (возможно).

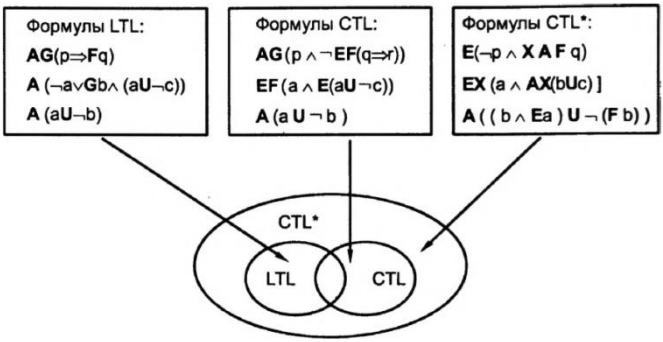


Рис. 1. Общая схема вложенности логик

Кванторы, зависящие от пути:

- φ X – Next: φ должно сохраняться в следующем состоянии (этот оператор иногда указывается N вместо X);
- φ G – Globally: φ должно сохраняться на всем последующем пути;
- φ F – Finally: φ в конечном итоге должно выполняться (где-то на последующем пути);
- ψ U φ – Until: φ должно сохраняться *по крайней мере* до тех пор, пока не будет сохранено некоторое положение ψ. Это подразумевает, что ψ будет проверяться в будущем;
- ψ W φ – Weak until - φ должно выполняться до тех пор, пока не будет выполнено ψ. Разница с U заключается в том, что нет гарантии, что ψ когда-либо будет проверено. Оператор W иногда называют "unless (пока не)".

2. Структура Крипке

Структура Крипке M – это пятерка $M = (S, S_0, R, AP, L)$, где:

- S – конечное непустое множество состояний;
- $S_0 \subseteq S$ – непустое множество начальных состояний;
- $R \subseteq SxS$ – тотальное отношение на S, т.е. множество переходов, удовлетворяющее требованию: $(\forall s \in S)(\exists s' \in S)(s, s') \in R$ (из любого состояния есть переход);
- AP – конечное множество атомарных предикатов;
- $L: S \rightarrow 2^{AP}$ – функция пометок (каждому состоянию отображение L сопоставляет множество истинных в нем атомарных предикатов).

В работе используется алгоритм model checking для CTL со сложностью $O\big(|\Phi| * (|S| + |R|)\big)$ [Еремеев и др., 2023].

3. Постановка задачи машинного обучения

Рассматривается задача определения выполнимости формулы CTL на заданной структуре Крипке. Формально задача может быть сформулирована следующим образом.

Дано:

- формула CTL: φ;
- структура Крипке: $M = (S, S_0, R, AP, L)$

Требуется:

определить, существует ли состояние $s \in S$, такое что $M, s \vDash \varphi$, т.е. φ выполнима в M.

В классическом варианте задача решается с помощью алгоритмов model checking, которые проверяют истинность формулы φ в каждом состоянии структуры. Однако в данном исследовании предлагается аппроксимировать эту проверку с помощью машинного обучения, предсказывая выполнимость φ на M как задачу бинарной классификации.

4. Описание предлагаемого подхода

Подготовительный этап. Для применения методов машинного обучения необходимо перевести логическую и графовую информацию в числовое или структурированное представление.

Формула представляется в виде абстрактного синтаксического дерева (Abstract Syntax Tree, AST), которое затем кодируется с использованием одного из следующих подходов:

- Bag-of-Operators: частотный вектор по используемым операторам (*EX, AG,* ∧, и т.д.).
- Position encoding: информация о глубине вложенности и структуре формулы.
- Tree embeddings: использование рекурсивных нейросетей или трансформеров для кодирования структуры формулы.

Поскольку структура Крипке – это ориентированный помеченный граф, она кодируется как: граф с вершинами (состояниями), дугами (переходами) и метками (атомарные предикаты); adjacency matrix + node labels: представление через матрицу смежности и признаковые векторы вершин. В случае применения GNN используется n-мерное признаковое пространство на узлах и итеративная агрегация признаков по соседям.

Формулировка в виде задачи классификации. Преобразованная пара (φ, M) подаётся на вход нейросетевой модели (например, GNN + MLP), задача которой – предсказать метку $y \in \{0, 1\}$, где: $y = 1$, если существует состояние s в M, такое что $M, s \vDash \varphi$ (т.е. φ выполнима); $y = 0$, в противном случае. Таким образом, подход работает как приближённый предсказатель

выполнимости, позволяя быстро фильтровать нерелевантные пары или подсказывать возможные области интереса для дальнейшей формальной проверки.

Основная цель нейросетевого подхода – снижение вычислительной нагрузки на классические алгоритмы model checking, особенно в случае больших систем. Модель может использоваться как предварительный фильтр (pre-check), приближённый классификатор или компонент гибридной системы верификации.

Целью данного этапа является подготовка качественной обучающей выборки, состоящей из пар (формула CTL, структура Крипке) и метки, определяющей, выполнима ли формула на данной структуре. Поскольку таких данных не существует в открытом доступе, необходима автоматическая генерация и разметка.

Генерация структур Крипке и CTL формул. Для получения разнообразных графов, имитирующих модели программ, используется генерация случайных структур:

- состояния (S): фиксируются или варьируются количество вершин (например, от 5 до 100);
- переходы (R): сгенерированы случайно с гарантией достижимости и связности (например, вероятность ребра p выбирается из диапазона [0.1, 0.4]);
- метки (L): каждый узел помечается подмножеством из фиксированного множества атомарных предикатов $AP = \{p, q, r, ...\}$.

Дополнительно: можно добавлять вариации – циклы, ветвления, сильно связные компоненты (SCC) – чтобы разнообразить типы систем.

Реализация: с использованием библиотеки networkx, можно параллельно генерировать тысячи таких графов.

Формулы CTL генерируются синтаксически с использованием шаблонов:

- используются глубины формулы как параметр сложности: глубина 2–5;
- операторы: EX, AX, EF, AF, EG, AG, логические связки ($\neg, \wedge, \vee, \rightarrow$);
- формулы создаются рекурсивно:
  - базовые случаи: атомарные p, q, $\neg p$, и т.д.;
  - рекурсивные: EX φ, $\varphi \wedge \psi$, AG($\varphi \rightarrow$ EF ψ), и т.д.

Формулы можно генерировать с контролем по:

- количеству переменных;
- глубине вложенности;
- частоте использования определённых операторов

Чтобы избежать переобучения, генерация должна обеспечивать структурное разнообразие. При генерации формул используются классовые представления формул из PyModelChecking [10].

Разметка данных. Для каждой пары (структура, формула) вычисляется метка. Используется model checker, предложенный в [Еремеев и др., 2023].

Если хотя бы в одном состоянии s верно, что $M, s \vDash \varphi$, то метка есть 1 (выполнима), иначе 0. Для расширенного исследования можно менять метрику выполнимости. Например, можно считать, что формула выполнима на структуре, если формула истинна во всех состояниях S структуры Крипке.

Параллельность. Поскольку каждая пара независима, проверку можно распараллелить по ядрам или в кластере. В языке python данная генерация реализована через Multiprocessing. Распараллеливание этого процесса особенно актуально на данных с большим количеством состояний, переходов в структуре Крипке и подформул формулы CTL.

При генерации формул возникает проблема в ситуации, когда классы сильно разбалансированы. Эту проблему можно решить отбрасыванием некоторой части сгенерированных данных, где превалирует та или иная метка (0 или 1). Для оценки качества перекоса данных при обучении так же используется метрика ROC-AUC [Hanley et al., 1982].

5. Архитектура модели GNN

Выбрана параллельная архитектура. Модель объединяет в себе два потока данных – графовую составляющую и формульную составляющую.

Графовая составляющая. Структура Крипке представляется графом, узлы которого имеют бинарные признаки — наличие или отсутствие атомарных предикатов (p, q, r, s). Эта часть обрабатывается двумя слоями GCNConv и агрегируется через global mean pool, формируя глобальное представление графа.

Global mean pool – это агрегирующая операция, используемая в сетях (GNN), в частности, в библиотеке PyTorch Geometric [PyG Documentation, 2024], для преобразования признаков узлов графа в один вектор фиксированной длины для каждого графа в батче. Эта операция является одним из простых эффективных способов в своем роде.

Формульная составляющая. Формула токенизируется и передается через слой Embedding, затем поступает в длительную кратковременную память (Long Short-Term Memory, LSTM). Последнее скрытое состояние LSTM служит эмбеддингом формулы. Затем два эмбеддинга – графа и формулы – объединяются конкатенацией и передаются в классификатор на полносвязных слоях для предсказания результата.

Общий план forward-прохода следующий:

- x = ReLU(GCN1(x)) → ReLU(GCN2(x)) → global_mean_pool → g_embed;
- formula → Embedding → Packed LSTM → f_embed;
- combined = concat(g_embed, f_embed) → classifier → output.

Предложенный подход имеет следующие преимущества. Модель обучается в end-to-end режиме. Это значит, что модель учится одновременно распознавать структуры графов и понимать логическую формулу с учётом

взаимосвязи между ними. Это ключевое преимущество end-to-end подхода. Разделение графа и формулы позволяет гибко подбирать архитектуры для каждого потока. LSTM хорошо справляется с переменной длиной формул, особенно в сочетании с pack_padded_sequence. Функция pack_padded_sequence нужна для игнорирования паддинга, чтобы обрабатывались только непустые токены.

Архитектура GNN с визуализацией процесса представлена на рис. 2.

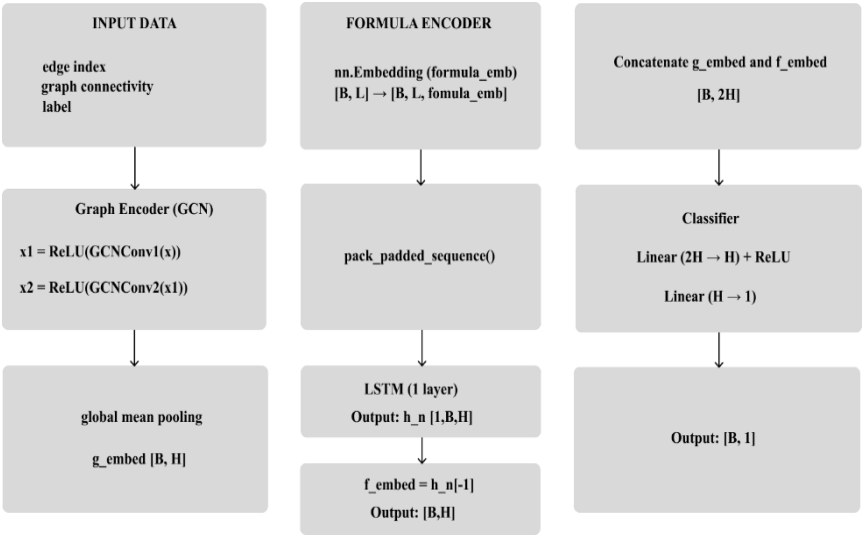


Рис. 2. Архитектура GNN

6. Обучение

Для обучения сети и запуска тестовых примеров использовалась следующая конфигурация ПК.

- Processor 12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz;
- Installed RAM 32,0 GB (31,7 GB usable);
- System type 64-bit operating system, x64-based processor.

Обучение проводится с использованием функции потерь BCEWithLogitsLoss, дополнительно используется параметр pos_weight для компенсации дисбаланса классов. Оценка проводится по метрикам Accuracy (сравнение предсказанных и истинных значений один к одному) и ROC-AUC (площадь под кривой ошибок). ROC-AUC нужна, чтобы оценить качество бинарного классификатора в случае, если присутствует большой перекос классов в исходных данных.

Процесс обучения проведен на разных датасетах. Результаты приведены в табл. 1. Заметим, что с ростом глубины генерации формулы увеличивается сложность задачи и время обучения. Это связано с тем, что формула глубины, например, 40, может иметь около 50000 токенов, что является объемным для вычислений. В целом результаты обучения хорошие, есть возможность разнообразить датасеты и увеличить количество эпох для увеличения точности. Наглядно результаты представлены на рис. 3.

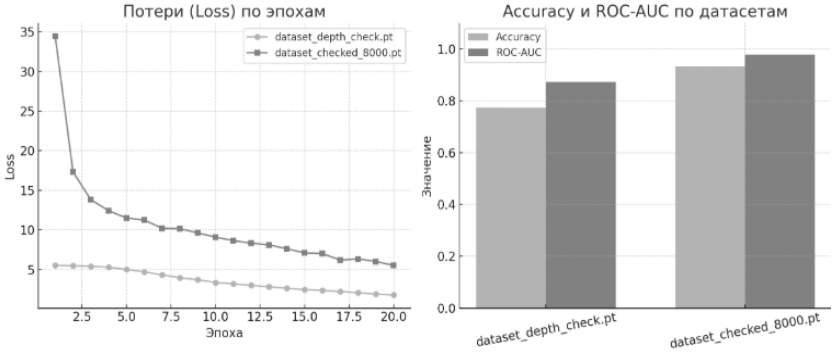


Рис. 3. Результаты обучения GNN

7. Анализ результатов

Будем сравнивать результаты работы алгоритма из [PyModelChecking's, 2024] с результатами GNN на одних и тех же наборах данных.

Предложенная графовая нейронная сеть GNN уверенно обходит классический подход по скорости. Даже при длинных формулах её время редко превышает 1–1.2 секунды. Однако оно также растёт с увеличением размера входных данных. Результаты сравнения даны в табл. 2.

Подход с использованием GNN в целом оказался эффективным – даже простые модели могут достигать высоких результатов и обходить по скорости классические алгоритмы модельной проверки, особенно, если использовать распараллеливание по ветвям. Отметим, что нейросетевой метод дает приближенные результаты и не гарантирует строгой корректности. Для минимизации рисков предлагается встраивать в систему многократный запуск метода и использовать, например, среднее среди всех полученных результатов. В случаях, когда метод работает кратно быстрее классических алгоритмов, такой подход может быть целесообразен.

Таблица 1

Результаты обучение GNN на разных датасетах

Параметр	Датасет 1	Датасет 2
Количество данных	1000	5160
Глубина формулы при генерации	40	20
Баланс классов	318 / 318	2580 / 2580
Эпох обучения	20	20
Потери (Loss) на 1-й эпохе	5.5344	34.4617
Потери (Loss) на 20-й эпохе	1.7991	5.5247
Accuracy на тесте	0.7734	0.9471
ROC-AUC на тесте	0.8728	0.9787
Время обучения (сек)	257.39	45.70
Примечание	Глубокие формулы, меньше данных	Много данных, формулы проще

Таблица 2

Сравнение времени работы: Model Checking vs GNN

Кол-во узлов	Глубина формулы	Длина формулы (токенов)	Model Checking (с)	GNN (с)
60	40	~5000	0.110	0.030
80	40	~5000	0.263	0.042
120	40	~5000	1.170	0.077
40	80	19512	5.078	1.036
80	80	—	12.481	1.229

Чем длиннее формула, тем эффективнее оказывается применение GNN по сравнению с обычным алгоритмом. Результаты приведены в табл. 3. При этом точность остаётся довольно высокой (около 90-92%), что делает GNN особенно удобными для задач, где важна скорость и обрабатывается большой поток формул. Заметим, что с увеличением глубины формулы значительно растёт нагрузка на память и вычисления, так как мы храним каждый токен формулы и при случайной генерации можем получить массивы порядка 10^5 элементов. LSTM обрабатывает последовательность за O(N) шагов (все токены параллельно через матричные операции). Проведённое сравнение демонстрирует практическое преимущество ML-подходов в задачах приближённой проверки CTL-формул. Хотя они не дают строгой гарантии корректности, но высокая точность и большой выигрыш по времени делают их подходящими применения в интеллектуальных системах реального времени (ИС РВ) (типа ИС РВ для поддержки принятия решений, ИСППР РВ), особенно когда требуется масштабируемость и оперативность.

Таблица 3

Влияние сложности формулы на время (фиксированное число узлов 40)

Макс. длина формулы	Длина токенов	Model Checking (с)	GNN (с)
120	8836	1.378	0.422
120	27040	3.970	1.271
1200	9826	1.514	0.459
1200	19830	3.007	0.890
2000	13629	2.082	0.617
2000	25416	3.781	1.226
5000	6036	0.926	0.270
5000	19804	3.005	0.886
5000	19292	2.927	0.866

Заключение

Предложен новый подход к задаче проверки выполнимости формул темпоральной логики CTL на структурах Крипке с использованием методов машинного обучения. Рассмотрена формализация задачи как бинарной классификации, где входом модели является пара (формула CTL, структура Крипке), а выходом – предсказание её выполнимости. Разработаны методы представления входных данных: формулы кодируются с помощью LSTM на основе токенизации; структуры Крипке – с помощью графовых признаков и обрабатываются графовой нейронной сетью (GNN). Предложена end-to-end архитектура, объединяющая оба потока данных. Предложена методика генерации синтетических датасетов – случайная генерация структур Крипке и формул CTL с последующей разметкой через классический алгоритм model checking. Это позволило получить объёмную и разнообразную обучающую выборку. Для ускорения реализована система параллельной генерации и верификации.

На экспериментальных данных продемонстрировано, что предложенная модель способна эффективно предсказывать выполнимость формул, достигая высоких значений метрик качества (Accuracy и ROC-AUC), и показывает значительное преимущество по времени работы по сравнению с классическим алгоритмом model checking, особенно при увеличении длины формулы или размера графа. Показана перспективность применения нейросетевых архитектур в задачах приближённой проверки логических свойств. Такие модели могут использоваться в качестве предварительных фильтров, ускоряющих классическую верификацию. Данные исследования и разработки выполняются в плане реализации базовых инструментальных (математических и программных) средств для конструирования ИС/ИСППР РВ для диагностики и мониторинга сложных технических (технологических) и организационных объектов и процессов.

Список литературы

[Еремеев и др., 2011] Еремеев А.П., Куриленко И.Е. Темпоральные модели на основе логики ветвящегося времени в интеллектуальных системах // Искусственный интеллект и принятие решений. – 2011. – № 1. – С. 14-26.

[Еремеев и др., 2017] Еремеев А.П., Куриленко И.Е. Реализация вывода в темпоральных моделях ветвящегося времени // Известия РАН. Теория и системы управления. – 2017. – № 1. – С. 107-127. – ISSN 0002-3388.

[Еремеев и др., 2023] Еремеев А.П., Филинов Н.Ю. Реализация алгоритма темпоральной ветвящейся логики в рамках инструментальных средств построения интеллектуальных систем поддержки принятия решений реального времени // Труды Межд. научно-техн. конг. «Интеллектуальные системы и информационные технологии – 2023 (IS&IT'23). Научн. изд. в 2-х т. Т. 1. – Таганрог: Изд-во Ступина С.А., 2023. – С. 320-330. – ISBN 978-5-6050434-1-6 (Т. 1).

[Ben-Ari et al., 1983] Ben-Ari M., Manna Z., Pnueli A. The temporal logic of branching time // Acta Informatica. – 1983. – 20 pp. – P. 207-220.

[Clarke et al., 1980] Clarke E.M., Emerson E.A. Characterizing correcteness properties of parallel programs using fixpoints // In Automata, Languages, and Programming. LNCS 85. – Springer, 1980. – P. 169-181.

[Clarke et al., 1981] Clarke E.M., Emerson E.A. Design and synthesis of synchronization skeletons using branching time temporal logic. In Logic of Programs: Workshop. LNCS 131. – Springer, 1981.

[Hanley et al., 1982] Hanley J.A., & McNeil B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve // Radiology. – 1982. –143(1). – P. 29-36.

[Kupferman et al., 2000] Kupferman O., Vardi M.Y. and Wolper P. An Automata-Theoretic Approach to Branching-Time Model Checking // In Journal of the ACM. – March 2000. – Vol. 47, No. 2. – P. 312-360.

[Pnueli, 1977] Pnueli. A. The Temporal Logic of Programs // In Proceedings of the 18th IEEE Symposium on Foundations of Computer Science, Providence, 31 October-2 Novemver 1977. – P. 46-67.

[PyG Documentation, 2024] PyG Documentation. – https://pytorch-geometric.readthedocs.io/en/latest/.

[PyModelChecking's, 2024] PyModelChecking's,documentation. – https://pymodelchecking.readthedocs.io/en/latest/.

УДК 004.89
doi: 10.15622/rcai.2025.018

К ПРОБЛЕМЕ ИНТЕГРАЦИИ СТАТИСТИЧЕСКИХ И ДЕТЕРМИНИСТСКИХ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ЭМПИРИЧЕСКИХ ДАННЫХ

М.И. Забежайло (*m.zabezhailo@yandex.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

Представлены некоторые возможности интеграции статистических и детерминистских методов, используемых в интеллектуальном анализе эмпирических данных. Подход ориентирован на поиск неявным образом заданных причинно-следственных зависимостей, связывающих «причину» и «контекст ее актуальности» с «вызываемыми» ими «следствиями». Обсуждается пример использования этого подхода в области высокотехнологичной медицинской диагностики.

Ключевые слова: интеллектуальный анализ данных, тернарное отношение причинности, медицинская диагностика

Введение

Наряду с обширным перечнем решений в области стремящихся к автономности своего функционирования прикладных систем искусственного интеллекта (ИИ) – различных аппаратно-программных комплексов, располагающихся на борту того или иного подвижного – передвигающегося по земле или по воде, летающего и т.п. – устройства, а также завоевавших сегодня фантастическую популярность Больших Языковых Моделей (БЯМ), все более заметный интерес как разработчиков так и влиятельных пользователей наблюдается в настоящее время к ИИ-системам и решениям, ориентированным на оперативный анализ больших объемов данных и поддержку управленческих решений. Актуальные примеры таких потребностей несложно найти, в частности, в области управления большими информационно-телекоммуникационными структурами (см., в том числе – [Verma et.al., 2015], [Tirmazi et.al., 2020], [Смирнов и др., 2024] и др.) обеспечения обороны и безопасности ([Entous, 2025a-b] и др.]), противодействия мошенничествам в финансовой сфере [Грушо и др., 2021], а также ряде других значимых областей приложений компьютерных методов и технологий.

Проблемно-ориентированный анализ Big Data и поддержка принятия управленческих решений в режиме жестких ограничений по времени сегодня – это область востребованности достаточно «тонко» организованных математических моделей, методов и прикладных систем ИИ. Уже при поверхностном взгляде на рассматриваемую проблематику исследований и разработок, возникает несколько вопросов фундаментального характера, в частности:

- Как именно «бороться» (см. ограничения по времени анализа данных и поддержки принятия решений) с эффектом *Big*, а также связанным с ним эффектом *Open*? (например, как интегрировать «быстрые» заключения статистического анализа данных «в среднем» – т.е. в определенном смысле – приближенные, дающие результат с точностью до подкласса статистически неразличимых, воспринимаемых средствами статистического анализа как «однородные» – объектов) с точными, однако, как правило, требующими при их поиске большого перебора возможных вариантов) детерминистскими средствами?
- Как оценивать достаточность оснований для принятия предлагаемых системой ИИ выводов и заключений? (Своего решения здесь требует основополагающая для ИИ проблема доверия к результатам, формируемым системой искусственного интеллекта).
- Как обеспечить принимающих ответственные решения лиц (ЛПР) неформальной интерпретацией и объяснением формируемых системой ИИ рекомендаций. Необходимость такого «сервиса» обусловлена ответственностью за последствия решений, которую ЛПР предстоит принять на себя?

Один из вариантов использования возможностей интеллектуального анализа данных (ИАД) для ответа на эти вопросы и будет предметом представленного ниже обсуждения.

1. Компьютерный анализ эмпирических данных: от логики доказательства к логикам рассуждений

Проблематика компьютерного анализа данных эмпирической природы – накапливаемых результатов лабораторных экспериментальных исследований того или иного физического или технического эффекта, клинических данных о возникновении и развитии у пациентов заболеваний той или иной определенной нозологии, данных мониторинга возникновений и распространения сбоев в крупных ИТ-структурах и т.п. – имеет ряд специфических особенностей, недостаточное внимание к которым может привести к формированию неадекватных заключений и выводов.

Так в целом ряде случаев задачи обсуждаемого типа требуют аккуратно учитывать открытый характер анализируемой предметной области – присущего Big Data так называемого эффекта *Open*, обусловленного возможностями появления в некоторый момент времени таких новых данных, кото-

рые ранее еще не приходилось анализировать и о существовании которых на текущий момент еще не было накоплено какой-либо значимой информации. С этим явлением, с легкой руки Н.Талеба получившим название эффекта *Черного лебедя* [Талеб, 2015], нам пришлось столкнуться, в т.ч., в период пандемии COVID-19, когда практически каждая очередная мутации вируса требовала изменений или же вообще новых подходов в борьбе с нею – новых препаратов, новых медицинских протоколов лечения и т.п.

Необходимость оперировать в условиях эффекта *Open* указала на неадекватность применения в таких условиях некоторых традиционных методов анализа данных, причем – еще «до компьютера», т.е. на уровне математической модели и алгоритмов анализа данных. Так, например, применение традиционных методов статистического анализа данных основано на идее формирования так называемой *генеральной совокупности*, затем выделения из нее *репрезентативных выборок* и последующего «обучения» на таких выборках для того, чтобы в рамках математических моделей интерполяционно-экстраполяционного типа (интерполяция «обучающей» выборки эмпирическими зависимостями заданного типа, например, регрессиями того или иного вида, а затем «диагностика» вновь анализируемого объекта проверкой экстраполируемости на него какой-либо из уже найденных интерполяционных зависимостей) обосновать «аналогию» анализируемого нового прецедента с теми, которые собраны в репрезентативной выборке из *генеральной совокупности*. Однако, в открытых предметных областях (см. эффект *Open*) само понятие генеральной совокупности как коллекции данных, в которой отражены все варианты проявления изучаемого эффекта, оказывается, вообще говоря, не универсальным. Потенциально возможное появление соответствующего *Черного лебедя* (см. [Талеб, 2015] и др.) ставит под сомнение *генеральный* характер той или иной конкретной *совокупности* данных. И дело здесь вовсе не использовании тех или иных компьютерных методов анализа данных, фундаментальная проблема – в *репрезентативности* используемых в каждом конкретном случае «обучающих» выборок из претендующей на универсальность конкретной совокупности эмпирических данных (ЭД).

Еще одной характерной особенностью компьютерного анализа ЭД является фокусировка внимания не на утверждениях универсального характера, справедливых для всей изучаемой предметной области или же ее «аналитически» выделяемого «целевого» фрагмента, а на специальном классе контекстно-определяемых утверждений – заключениях и выводах, принимаемых (т.е. оцениваемых как приемлемые, заслуживающие внимания) с точностью до релевантности заданному «контексту». Например, это могут быть заключения, которые релевантны набору эмпирических фактов (ЭФ), ранее уже «установленных» экспериментальным путем, – в частности, данным конкретных медицинских «тестов», результатам конкретных лабораторных «измерений» и т.п.

Таким образом, вполне естественным выглядит разделение процедурных «инструментов» – логико-математических моделей, методов и алгоритмов, ориентированных на формирование «универсальных» доказуемых утверждений (будем называть такие «инструментальные» средства анализа данных логиками доказательства), и тех, задача которых – сформировать «рациональные» следствия (причем – не только сугубо дедуктивного характера, но и те, которые используют формализованные варианты и других познавательных процедур – формирования индуктивных обобщений, рассуждений по аналогии, построения абдуктивных объяснений и др. [Финн, 2021, 2024]). Как следствие, актуальной оказывается задача порождения из накапливаемых ЭД не только «универсально» не оспариваемых утверждений (доказуемо корректных, как уже отмечалось выше, для всей анализируемой предметной области), но и тех – в некотором смысле «частных» (не универсальных) заключений, которые (как это удается продемонстрировать) оказываются неоспариваемыми относительно заранее заданного «контекста» – например, набора определенных теоретических утверждений («априорных» гипотез) и набора некоторых ЭФ, зафиксированных объективными средствами в ходе экспериментов, которые выполнены с соблюдением определенных «правил». Примеры востребованности подобного рода заключений и выводов нетрудно найти в задачах диагностического типа [Забежайло и др., 2021] – в медицинской или технической диагностике, идентификации и противодействию мошенничествам в финансовой сфере, в задачах обеспечения кибербезопасности и др.

Задействованные при этом формализованные средства формирования «рациональных» заключений и выводов из накапливаемых ЭД представляется естественным рассматривать как логики рассуждений – логико-математический «инструментарий» систематического порождения *следствий* из анализируемого *контекста* в его текущем состоянии. Таким образом, логика рассуждений возникает как объединение двух этапов (и двух типов «инструментальных» средств анализа данных):

- *получения следствий* (заключений, причем – не обязательно строго дедуктивного характера) из имеющегося на текущий момент (и потенциально – расширяющегося – см. эффект *Open*) контекста, а также
- *оценки достаточности оснований для принятия* порожденных на первом этапе следствий (*оценки доверия* к таким следствиям).

При этом наряду с разработкой «инструментов» собственно порождения следствий из контекстов – математических моделей, методов и алгоритмов, позволяющих использовать формализованные средства рассуждения для формирования неоспариваемых на заданном контексте выводов и заключений, критически значимой оказывается разработка средств оценки доверия к полученным в процессе выполненных рассуждений результатам («инст-

рументов» оценки достаточности оснований для принятия выводов и заключений, полученных формализованными средствами). Таким образом, важно иметь соответствующие процедурно-«инструментальные» возможности не только рассчитать – целенаправленно «вывести» из анализируемого набора имеющихся ЭД с помощью процедур контролируемого «наследования» корректности – соответствующие «следствия», но и оценить возможности доверять им (оценить их неоспариваемость на уже накопленном контексте – имеющихся на данный момент ЭД, а также, разумеется, если это окажется возможным, – идентифицировать[1] неоспариваемость этих заключений на всей анализируемой предметной области, т.е. обосновать «универсально»-доказуемый характер полученных «следствий»).

Логико-математический «инструментарий» подобного типа – специальные средства логики рассуждений уже достаточно давно привлекают внимание специалистов. Так, например, еще в конце 70-х годов XX века чешские математики П. Гаек и Т. Гавранек предложили [Hajek et.al., 1978] оригинальный подход, который позволял автоматизированными компьютерными средствами выдвигать гипотезы на основе накапливаемых ЭД, а затем оценивать статистическими средствами достаточность оснований для их принятия. В работе [ван Бентем, 2011] Йохан ван Бентем сформулировал необходимость создания логики рассуждений, дополняющей логику доказательства средствами поддержки взаимоотношений между логическими и эмпирическими фактами. Обширное отражение в научной литературе получили исследования методов и средств порождения зависимостей на базе накапливаемых ЭД (см., например, [Agrawal et. al., 1996], [Hajek, 2001]).

Однако, обсуждаемая проблема разработки инструментов формирования и последующей оценки достаточности оснований для принятия результатов интеллектуального анализа данных (ИАД) в постоянно расширяемых новыми сведениями контекстах – коллекциях ЭФ – по-прежнему далека от общепризнанного решения. Дополнительные «краски» к сложившейся на текущий момент «картине мира» добавили представления о границах доказуемого и недоказуемого в условиях эффекта *Open* (см. в частности, естественные ограничения дедуктивной доказуемости или уже упоминавшиеся выше проблемы с формированием *генеральной совокупности* при применении методов статистического анализа данных.). В свою очередь, представления о доказуемости как о возможности приведения к неоспариваемости на уже накопленных ЭД потребовало уточнения требуемых границ такого контекста. Возможности пополнения текущего набора ЭД релевантной цели анализа новой информацией привели к необходимости задуматься над задачей проверки сохранения неоспариваемости ранее уже полученных выводов и заключений при направленном

[1] Или же, наоборот, – опровергнуть.

пополнении анализируемых данных новыми сведениями. Естественным следствием этому стал интерес к неформальной интерпретации и объяснению результатов ИАД. При этом, по-видимому, наиболее продуктивным оказался подход к объяснению, трактуемому как ответ на вопрос *ПОЧЕМУ*? ([Pearl, 1995] [Zabezhailo, 2021] и др.). Именно этим, в свою очередь, можно объяснить сегодняшний интерес сообщества исследователей ИИ к задачам разработки проблемно-ориентированных математических моделей, методов и алгоритмов восстановления причинно-следственных зависимостей, которые изначально скрыты в анализируемых ЭД ([Pearl, 1999, 2000] и др.).

2. Компьютерный анализ постоянно пополняемых эмпирических данных

Потребность оперировать data_set'ами ограниченного размера при решении задач компьютерного анализа данных не нова и достаточно хорошо знакома специалистам. Еще без малого 60 лет назад Ю.И.Журавлев предложил элегантную математическую конструкцию – так называемые корректные алгебры над множеством некорректных (эвристических) алгоритмов [Журавлев, 1977], которая позволила получить эффективные решения в целом ряде важных для страны прикладных задач[2]. Однако, в тех задачах, как и в целом в классическом машинном обучении, речь шла о выборе наиболее «точного» (минимизирующего ошибки классификации) на заданном анализируемом data_set'е алгоритма в заданном семействе алгоритмов. В настоящее время ситуация принципиально иная: необходимо оперировать малыми[3] по числу элементов data_set'ами, которые последовательно пополняются новыми данными той же «природы». В формализованном виде это требует разработки подходов, математических моделей и методов анализа данных, представляемых расширяющимися (по крайней мере – в части числа их строк – перечня объектов-прецедентов, входящих на текущий момент в анализируемый data_set) «плоскими[4]» матрицами ОБЪЕКТЫ х ПРИЗНАКИ. Собственно, это и стало дифференцирующим фактором, позволяющим отделить классическое ML от ИАД, ориентированного, помимо прочего, и на учет динамики изменений в анализируемых данных. В свою очередь, работа с последовательностями динамически изменяемых (так называемых расширяющихся [Finn, 2019, 2023]) баз фактов средствами интерполяционно-экстраполяционных математических методов требует формирования проблемно-ориентированных «инструментов» контроля надежности ин-

[2] В т.ч., поиска промышленных месторождений золота [Журавлев, 2024].

[3] Одним из формальных уточнений этого понятия может быть статистическая незначимость размеров таких наборов данных.

[4] Где число сток, как правило, существенно меньше числа столбцов.

дуктивных (экстраполяционных) обобщений, формируемых на таких данных средствами ИАД. Результативным математическим «инструментарием», позволяющим выделять в множестве результатов ИАД подмножества таких, надежность которых не вызывает сомнений, стал поиск в анализируемых данных эмпирических зависимостей (ЭЗ) причинно-следственного типа – каузальных маркеров изучаемых эффектов.

Наиболее распространенным вариантом математических средств выявления в анализируемых данных эмпирических зависимостей – маркеров изучаемых эффектов – на сегодня являются методы статистического анализа данных (САД). Примером популярной математической техники САД может служить часто используемый в медицинских приложениях подход Каплана-Майера, в рамках которого результаты поиска эмпирических закономерностей в иерархии случайных величин (которыми в процессе выполняемого САД представлены клинические признаки) кладутся в основу прогнозирования исхода наблюдаемого заболевания ([Ардашев и др., 2024] и др.). Выявление взаимосвязей между группами клинических признаков (идентификация маркеров целевого эффекта), а также оценка меры их причинно-следственной взаимосвязанности позволяют целенаправленно формировать тактику персонализированных медицинских воздействий.

Однако, одним из уязвимых мест этой процедурной конструкции оказывается проблема «предсказательной силы» маркеров выбранного типа. Заключения, формируемые на базе частот встречаемости отдельных признаков в анализируемых data_set'ах, надежность (устойчивость) собираемых далее комбинаций таких признаков, оценки доверия и общая «прогностическая сила» финального заключения, к сожалению, далеко не всегда могут вести к выводам и рекомендациям неоспариваемого характера. Важным компонентом в множестве причин возникновения ситуаций этого типа оказывается то, что, хотя выделяемый в процессе САД фактор или комбинация факторов влияют на эффект «достаточно часто», тем не менее, есть прецеденты, в которых этот фактор\комбинация факторов наличествуют, а целевой эффект при этом отсутствует. Т.е. не смотря на то, что данная комбинация факторов «во *многих* случаях» позволяет указать на возможное присутствие изучаемого целевого эффекта у рассматриваемых объектов-прецедентов, она – эта комбинация факторов – в общем случае не годится для формирования на ее основе неоспариваемого заключения (например, неоспариваемого диагноза у конкретного пациента).

Что же – какие именно математические модели и методы САД – предложить в таких ситуациях для устранения некорректности в заключениях выполняемого ИАД (например, формирования неоспариваемого медицинского диагноза и т.п.). Как показали соответствующие исследования, одним из вариантов результативных действий может обеспечить ориентация на рабочую гипотезу вида:

помимо самого «базового» фактора влияния (или определенной *комбинации* таких *факторов*) для *наличия целевого эффекта* критически значим и определенный «*контекст*[5]», во взаимодействии с которым «*базовый*» *фактор* обеспечивает *появление целевого эффекта*.

Как следствие, возникает задача: какими средствами идентифицировать такого рода контекст по имеющимся data_set'у и целевому эффекту?

3. Предлагаемый вариант решения

В общем виде предлагаемый вариант решения, опирающийся на сформулированную выше рабочую гипотезу, характеризует следующая процедурная конструкция: будем использовать ДСМ-ИАД ([Финн, 2021, 2024], [Grusho et.al., 2024], [Грушо и др., 2021] и др.]) на базе тернарного отношения причинности вида:

< [*статистически значимая*[6] *комбинация факторов причинного влияния*]; [*контекст*] => => [*целевой эффект*] >.

Ключевые отличительные особенности предлагаемого подхода:

- неопровергаемость заключения ДСМ-ИАД на имеющихся на текущий момент эмпирических данных (в т.ч. – с учетом контрафактуальных ([Höfler, 2005], [Pearl, 1999, 2000] и др.) заключений, и выполнимости условия Запрета на КонтрПримеры ([Финн, 2021, 2024] и др.);
- неформальная интерпретируемость формируемых эмпирических зависимостей – маркеров целевого эффекта (и, как следствие, минимизация эффектов переобучения в процессе компьютерного анализа данных используемыми статистическими и детерминистскими средствами);
- объяснение результативности формируемых ЭЗ-маркеров целевого эффекта за счет использования (выявления) каузальных зависимостей, изначально скрытых в анализируемых эмпирических данных;
- интеграция статистических и детерминистских средств компьютерного анализа данных (идея Тьюки [Tukey, 1977]): сперва «быстрый» поиск базовых факторов влияния статистическим средствами, а затем – последующая идентификация «контекста» их «продуктивности» детерминистскими средствами (как оказалось, не требующая исчерпывающего перебора формируемых в процессе ДСМ-ИАД всех вариантов гипотез о причинах).

[5] Некоторые определенные дополнительные факторы, принимающие некоторые специальные – взаимосвязанные со значениями «базового» фактора – значения.

[6] Идентифицированная средствами той или иной версии стат.анализа данных.

4. Как это работает. Пример

В процессе изучения эффекта так называемой псевдопрогрессии (*ПсП*) опухолей головного мозга человека специалистами Национального медицинского исследовательского центра нейрохирургии имени академика Н.Н. Бурденко (НМИЦ НХ им. Н.Н. Бурденко) на ЭД 427 пациентов, накопленных почти за 20 лет клинических исследований, были идентифицированы 59 неоспариваемых прецедентов наличия эффекта *ПсП* и 368 случаями его отсутствия. При этом описание каждого из анализируемых прецедентов содержит значения более 150 отдельных параметров (клинических признаков[7]). У 48 из 59 пациентов с *ПсП* при этом зафиксировано наличие кисты в опухоли (45 имели чисто кистозную *ПсП*, а 3 - так называемую смешанную[8]). Таким образом, фактор наличие кисты в опухоли характеризовался высоким значением частоты встречаемости на прецедентах *ПсП* (что дает основания использовать его как указатель на соответствующую группу «риска»). Однако, у 78 из 368 пациентов без *ПсП* также наблюдалась киста в опухоли. Таким образом, только один этот клинический признак не представляется возможным использовать для надежного прогноза наличия\отсутствия эффекта *ПсП* в целом по рассматриваемому множеству пациентов.

В работе [Трунин, 2021] был предложен механизм (основанный на модели САД в духе метода Каплана-Майера ([Ардашев и др., 2024] и др.), позволяющий проводить дифференциальную диагностику эффектов *ПсП* и рецидива опухолей (*РО*). В его основе – полученный Ю.Ю. Труниным САД-критерий, интегрирующий три клинических параметра – *возраст пациента* (*младше 11 лет*, *11 лет и старше*), *локализация опухоли* (*супратенториальная*, *инфратенториальная*) и *наличие кисты* в опухоли (*да*, *нет*). Риск возникновения эффекта *ПсП*, как показал Ю.Ю. Трунин, становится высоким (54%) уже через 6 месяцев после лучевой терапии (ЛТ), через 12 мес. после ЛТ он возрастает до 95.2%, и подходит к 100% через 48 мес. после ЛТ. В свою очередь, на имеющихся ЭД рецидивы опухоли возникают через 2 и более года (медиана 29 мес.) после ЛТ. Все это позволяет сформировать эффективные критерии отличия эффектов *ПСП* и *РО*.

Однако, открытым оставался ряд вопросов, связанных с *персонифицированной* диагностикой эффектов *ПсП* и *РО* (т.е. с возможностью формировать для конкретного пациента неоспариваемое[9] заключение, учитывающее не только общие характеристики соответствующей целевой группы, но и его персональные особенности).

---

[7] Мультимодальной природы количественных, качественных и др.

[8] При которой имеются признаки как кистозного, так и с*о*лидного характера.

[9] Например, на имеющемся на текущий момент массиве эмпирических данных.

Для решения это задачи персонификации диагноза была использована представленная (см. Разделы II-III) математическая техника ИАД на основе тернарного отношения причинности. Было показано (см. [Забежайло и др., 2024]), что совместное использование определенных значений дополнительных факторов – *гистология опухоли* и *локализация опухоли*, объединенных в виде сопутствующего контекста вместе с фактором *киста в опухоли*, позволяет получить неоспариваемый на имеющихся ЭД прогноз возникновения эффекта *ПсП*. При этом сформированный контекст «выдержал» проверки контрафактуальности по всем остальным имеющимся эмпирическим данным (находящимся за пределами множеств значений параметров, выделенных при формировании этого контекста).

Таким образом, дополнение статистически значимого фактора *киста в опухоли* идентифицируемым детерминистскими средствами контекстом позволило формировать неоспариваемые на имеющемся эмпирическом материале персонифицированные заключения относительно возникновения эффекта *ПсП*, в том числе – у отдельно изучаемого пациента.

Заключение

Актуальной для эффективного применения методов ИАД в практически значимых приложениях по-прежнему остается на сегодняшний день проблема «управления» размерностью множества параметров, характеризующих анализируемый data_set. Активно разрабатываются различные методы сокращения размерности, позволяющие «склеивать» «малосущественные» переменные без существенной потери в точности результата, формируемого на «усеченных» таким способом данных. Не теряет своей популярности классическая идея выделения *главных компонент* в анализируемых многомерных данных, а метод PCA [Pearson, 1901] – был и остается прекрасным примером такого математического инструментария для работы с данными метрического характера.

Представленная в данной работе версия ИАД на базе контекстного отношения сходства также может рассматриваться как пример использования аналогичного подхода (в основе которого - «склейки» некоторых параметров в описаниях объектов-прецедентов) при работе с неметрическими данными в решении задач диагностического типа.

Таким образом, отдельного внимания, по-видимому, заслуживают возможности, которые открывает интеграция этих двух подходов в задачах анализа больших наборов мультимодальных данных. Практически значимый пример – задачи поддержания так называемой ситуационной осведомленности ЛПР в ряде актуальных приложений ([Entous, 2025a-b] и др.]).

Список литературы

[Ардашев и др., 2024] Ардашев В.Н., Калёнова И.Е., Ляпкова Н.Б., Потехин Н.П., Фурсов А.Н. (Под ред. Проф. Бояринцева В.В.) Доказательная медицина: обзор современных математических методов анализа. – М.: УД Президента РФ, 2013. – 224 с. – https://volynka.ru/Articles/Text/124?ysclid=m8lk72172l581580264.

[ван Бентем, 2011] Бентем ван, Й. Логика и рассуждение: много ли значат факты? // Вопросы философии. – 2011. – №12. – С. 63-76.

[Грушо и др., 2021] Грушо А.А., Забежайло М.И., Смирнов Д.В., Тимонина Е.Е. Интеллект. анализ пополняемых коллекций Big Data в режиме процессно-реального времени // Информатика и ее применения. – 2021. – Т. 15, № 2. – С. 36-43.

[Журавлев, 1977] Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. I-III // Кибернетика. – 1977. – № 4. – С. 5-17, 1977. № 6. С. 21-27, 1978. № 2. – С. 35-43.

[Журавлев, 2024] Журавлев Ю.И. Распознавая образ / Колл. авт.; ред. А.Л. Семенов, Ю.В. Чехович и Е.О. Самойлова. – М.: Буки Веди, 2024. – 512 с.

[Забежайло и др., 2021] Забежайло М.И., Грушо А.А., Грушо Н.А., Тимонина Е.Е. Поддержка решения задач диагностического типа // Системы и средства информатики. – 2021. – Т. 31, № 1. – С. 69-81.

[Забежайло и др., 2024] Забежайло М.И., Михеенкова М.А., Трунин Ю.Ю. О небинарной версии отношения причинности в интеллектуальном анализе онкологических данных // НТИ, Сер. 2 «Информ. проц. и системы». – 2024. – №. 6. – С. 13-20.

[Смирнов и др., 2024] Смирнов Д.В., Грушо А.А., Забежайло М.И. К задаче идентификации сбоев в информационно-технологической инфраструктуре путем мониторинга и анализа косвенных данных // Системы и средства информатики. – 2024. – Т. 34. – Вып. 3. – С. 14-22.

[Талеб, 2015] Талеб Н.Н. Черный лебедь. Под знаком непредсказуемости. – М: КоЛибри, 2015. - 36 С.

[Трунин, 2021] Трунин Ю.Ю. Стереотаксическое облучение в комплексном лечении пациентов с пилоидными астроцитомами: дис. … д-ра мед. наук: 14.01.18 – нейрохирургия и 14.01.13 – лучевая диагностика, лучевая терапия. – М.: НМИЦ НХ им. Н. Н. Бурденко, 2021. – 294 с.

[Финн, 2021] Финн В.К. Искусственный интеллект: методология, применение, философия. – М.: ЛЕНАНД, 2021. – 468 с.

[Финн, 2024] Финн В.К. ДСМ-метод автоматизированной поддержки исследований и аппарат понятий для искусственного интеллекта // Искусственные общества. – 2024. – Т. 19. – Вып. 2. – https://artsoc.jes.su/s207751800030907-6-1/.

[Agrawal et. al., 1996] Agrawal R., Manilla H., Sukent R., Toivonen A., Verkamo A. Fast discovery of Association rules // In: Advance in Knowledge Discovery and Data Mining. – P. 307-328. AAAI, Menlo Park, 1996.

[Entous, 2025a] Entous A. The Partnership: The Secret History of the War in Ukraine // The New York Times. – March 29, 2025. – https://www.nytimes.com/interactive/2025/03/29/world/europe/us-ukraine-military-war-wiesbaden.html.

[Entous, 2025b] Entous A. Key Takeaways From America's Secret Military Partnership With Ukraine // The New York Times. – March 30, 2025. – https://www.nytimes.com /2025/03/30/world/europe/us-ukraine-military-war-takeaways.html?unlocked_article_ code=1.8E4.nuPY.ksQThlmbAR9A&smid=nytcore-ios-share&referringSource= articleShare.

[Finn, 2019] Finn V.K. On the Heuristics of JSM Research (Additions to Articles) // Autom. docum. and mathematical linguistics. – 2019. – Vol. 53, No. 5. – P. 250-282.

[Finn, 2023] Finn V.K. On Empirical Regularities in the JSM Method of Automated Research Support // Automatic Documentation and Mathematical Linguistics. – 2023. – Vol. 57. – No. 6. – P. 362-381.

[Grusho et.al., 2024] Grusho A., Grusho N., Zabezhailo M., Timonina E. (2024). On Some Possibil. of Using AI Methods in the Search for Cause-And-Effect Relat. in Accumul. Empirical Data. // In: Kovalev S., Kotenko I., Sukhanov A., Li Y., Li Y. (eds) // Proc. of the 8th Int.. Scient. Conf. "Intell. Inform. Techn. for Industry" (IITI'24), Vol. 2. Lect. Not. in Netw. and Syst. V. 1210. – Springer, Cham. – P. 280-290.

[Hajek et.al., 1978] Hájek P., Havránek T. Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory). – Springer, Heidelberg, 1978. – https://link.springer.com/book/10.1007/978-3-642-66943-9. – P. 396.

[Hajek, 2001] Hájek P. The GUHA method and mining association rules // In: Proc. CIMA 2001, Bangor, Wales, 2001. – P. 533539..

[**Höfler, 2005**] Höfler M. Causal inference based on counterfactuals // BMC Med. Res.Methodol. – 2005. – 5, 28. – https://bmcmedresmethodol.biomedcentral.com/ articles/10.1186/1471-2288-5-28.

[**Pearl, 1995**] Pearl J. Causal diagrams for emp. Research // Biometrika. – 1995. – 82. – P. 669-710.

[**Pearl, 1999**] Pearl J. Probabilities of causation: three counterfactual interpretations and their identifications // Synthese. – 1999. – 121. – P. 93-149.

[**Pearl, 2000**] Pearl J. Causality: models, reasoning, and inference. – N.-Y.: Cambr. Univ. Press, 2000. – 384 p.

[Pearson, 1901] Pearson K. On lines and planes of closest fit to systems of points in space // Philos. Mag. – 1901. – 2(11). – P. 559-572.

[Tirmazi et.al., 2020] Tirmazi M., Barker A., Deng N., Haque Md Em., Qin Z.G., Hand S., Harchol-Balter M., Wilkes J. Borg: the Next Generation // EuroSys '20: Proc. of the 15th Europ. Conf. on Computer Systems. – 2020. – Article No.: 30. – P. 1-14.

[Tukey, 1977] Tukey J.W. Exploratory data analysis. – Reading, Mass.: Addison-Wesley Pub. Co., 1977. – 712 p.

[Verma et.al., 2015] Verma A., Pedrosa L., Korupolu M., Oppenheimer D., Tune E., Wilkes J. Large-scale cluster management at Google with Borg // EuroSys '15: Proc. of the10th Eur. Conf. on Computer Systems. – 2015. – Article No. 18. – P. 1-17.

[**Zabezhailo, 2021**] Zabezhailo M.I. Models of Explanation in Intelligent Data Analysis // IMSC-2021: Integr. Models and Soft Comp. in AI Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the IMSC-2021", Kolomna, Russia, May 17-20, 2021 (CEUR Workshop Proc. – Vol. 2965. – P. 59-63.

УДК 004.83
doi: 10.15622/rcai.2025.019

ДОГАДКА, ЭМПИРИЧЕСКАЯ ПРОВЕРКА И ОБЪЯСНЕНИЕ ПРИ ВТОМАТИЗИРОВАННОМ РЕШЕНИИ ЗАДАЧ

С.С. Курбатов (*curbatow.serg@yandex.ru*)

Научно-исследовательский центр электронной вычислительной техники, Москва

Анализируется различие в подходах к автоматизированному решению задач с использованием LLM и базирующемся на моделировании рассуждений. Анализ проводится на примере решения олимпиадной геометрической задачи. Рассматриваются такие аспекты как переход от текста задачи к компьютерному представлению, взаимодействие с нейросетью в процессе решения, качество объяснения доказательства, визуализация и её роль для эмпирических догадок.

Ключевые слова: LLM, Моделирование рассуждений, Олимпиадная задача.

Введение

Интерес к компьютерному моделированию рассуждений, заложенный ещё в пионерских работах Д.А. Поспелова, в последнее время возрос в связи с трудностями объяснения выводов и рекомендаций в нейронных сетях.

Использование больших лингвистических моделей (LLM) для решения сложных проблем инициировало комплексный подход, сочетающий поиск в нейронных сетях с логической (символьной) обработкой. В [Chen et. al., 2025] обсуждаются современные эффективные модели рассуждений (DeepSeek-R1-Zero и DeepSeek-R1), использующие обучение с подкреплением (RL). Отмечаются проблемы моделей и их преодоление путём многоступенчатого обучения и специального старта перед RL. В [Ma et. al., 2025] подчеркивается, что современные масштабные модели рассуждений (DeepSeek-AI, OpenAI o1 и ряд других) при подходе к сложным задачам в процессе поиска генерируют длинные цепочки рассуждений («размышлений»), комбинируя их с возвратом назад и «самооценкой». Тем не менее обход промежуточной логической обработки с

помощью простой подсказки («NoThinking», по терминологии авторов результата) показал эффективность в ряде сложных рассуждений, включая решение математических задач, формальное доказательство теорем и кодирование.

OpenAI o1 обучена для выполнения сложных рассуждений, прежде чем ответить пользователю система может создать длинную внутреннюю цепочку ассоциативных связей [Chen et. al., 2024]. По утверждению авторов модель ИИ, использующая такую цепочку, может рассуждать над сложными задачами и решать более трудные проблемы, чем предыдущие модели в области науки, кодирования и математики. Фреймворк ReSearch рассматривает поисковые операции в LLM как неотъемлемые компоненты цепочки рассуждений. Вопрос о специальном контроле над LLM, чтобы улучшить моделирование рассуждений и сделать их более антропоморфными, исследуется в [Højer et. al., 2025]. Такой контроль позволяет улучшить производительность в конкретных задачах, избегая трудоёмкого дообучения LLM. Обращение ученых к моделям ИИ для формирования новых гипотез, а частности, в таких областях как химия, биология, медицина обсуждается в [Bajorath, 2025]. Автор предупреждает о возможных заблуждениях при интерпретации результатов предсказательных алгоритмов. Подчеркивается необходимость объяснения предсказаний на уровне причинно-следственных связей.

Философские аспекты моделей для научных рассуждений и их роль в образовательных контекстах с эпистемологических позиций обсуждаются в [Rost, et. al., 2022]. Отмечается, что с этих позиций к числу необходимых характеристик знания должно относиться не только соответствие информации реальному положению дел, но и её обоснованность. Такая трактовка способствует более последовательному теоретическому пониманию моделирования и интерпретации результатов эмпирических исследований.

Вопросы естественно-научного образования в аспекте понимания того, как ученые мыслят и рассуждают обсуждаются в [Krell et. al., 2022]. Авторы рассматривают феномен научного мышления как компетенции, а не только как обладание способностями и знаниями. Подчёркивается, что такая интерпретация важна не только для ученого, но и для любого человека при принятии взвешенных решений в обыденной жизни. Абдуктивный подход к моделированию рассуждений исследуется в [Upmeier zu Belzen et. al., 2021]. Подход предполагает объяснение с помощью причины (отличается от индуктивных и дедуктивных рассуждений). Указывается роль подхода для компетентности в области моделирования рассуждений.

Подчёркивая ограничения на математические рассуждения моделей в [Mirzadeh et. al.] отмечается, что существующие LLM не выполняют полноценные логические рассуждения – они просто копируют шаги рассуждения,

извлекаемые из данных после обучения. Различные перефразировки одного и того же запроса (даже изменение только числовых характеристик) могут существенно повлиять на качество ответа. В относительно узкой предметной области взаимодействие LLM и символьного логического вывода («движки») демонстрируется в [Trinh et. al., 2024], [Zhang et. al., 2024]. Тем не менее эти работы будут нам интересны именно в аспекте такого взаимодействия и моделирования рассуждений.

Основная цель предлагаемого исследования – сопоставить возможности моделирования рассуждений в нейросетевой модели (LLM) и в системе, использующей традиционные методы ИИ. В нашей работе подчеркивается, что потенциал традиционных методов недооценён в связи с превалирующей тенденцией использования нейросетей. Не менее важна и прикладная сторона – ознакомление учащихся с элементами ИИ, непосредственно относящихся к учебному процессу, а также возможность прямого участия в компьютерной доработке. Выбранная предметная область – автоматическое решение сложных геометрических задач. Именно в геометрии наглядность и интуитивная очевидность сочетается с логической глубиной. В общей постановке проблема моделирования рассуждений весьма сложна. Ограничения предметной области позволяют сосредоточиться на рассуждения в «чистом» виде, не выхолащивая тем не менее суть дела. Далее обсуждаются детали сопоставления.

1. Методология

Концепции моделирования рассуждений пока не развиты в направлении упрощения алгоритмизации и масштабных прикладных возможностей. Во введении выделены работы, ориентированные на моделирование рассуждений в бурно развивающейся области нейронных сетей и больших лингвистических моделей (LLM). После пионерских работ Д.А. Поспелова велась многолетняя атака на эту проблему. В частности, в одной из работ в соавторстве с Д.А. Поспеловым анализировалась проблема представления знаний о времени и пространстве, предлагалась каузальная логика, формализующая причинно-следственные отношения.

В работах Э.В. Попова исследовалась модель участника общения с учетом его знаний о языке и предметной области, Р. Шенк выдвинул и экспериментально обосновал идею концептуальной обработки информации, использующей механизм умозаключений. Создавались системы, использующие логический вывод, экспертные системы, базы знаний (онтологии), условно формальные системы (когнитивные, индуктивные, по аналогии, правдоподобные умозаключения и т.п.), в той или иной степени затрагивающие проблему моделирования рассуждений. Однако, несмотря на отдельные и довольно серьёзные успехи, необходимость полноценного решения проблемы компьютерного моделирования рассуждений сохраняет свою актуальность.

Моделирование рассуждений затрагивает и проблему объяснимого ИИ (XAI), поскольку сформулированные в [Phillips et al., 2020] общие принципы (Объяснение, Понимание, Точность объяснения, Границы знания) безусловно значимы для рассуждений. Для конкретных областей XAI даже разработаны фреймворки на базе Python [Hu et al., 2021], однако они ориентированы скорее на интеграцию алгоритмов и обобщение API для данных. Разумеется, эти технологические аспекты важны, особенно при модификации системы, однако они лежат несколько в стороне от собственно моделирования рассуждений.

В медицинских приложениях используются вероятностные модели, основанные на байесовских сетях [Pradeepta, 2022], отражающих причинно-следственные связи и неопределённости. Более антропоморфные подходы обычно используют интерпретируемые визуализации (графы, таблицы, изображения) или тексты, призванные «объяснить», как система получила свои результаты. Используемый нами подход базируется на интерактивной визуализации, позволяющий кликом мыши по объекту чертежа выдать обоснование его появления.

В развиваемом автором подходе методологической основой является онтология, концентрирующая в когнитивных схемах лингвистические, логические (теоремы, их связи с ЕЯ и выводом) и визуальные знания, а также возможности их взаимодействия для эффективного вывода, ориентированного на человека (объяснительные возможности). Подход реализуется в компьютерной системе, базирующейся на идеях известного ученого и педагога Пойа [Polya, 1981]. Возникающие при этом проблемы требуют комплексного привлечения различных методов, некоторые аспекты рассмотрены в [Kurbatov et. al., 2024], [Kurbatov, 2023]. Пойа подчеркивал, что важно не только собственно решение, но и то, какими рассуждениями обоснованы шаги решения (выбор теорем, дополнительные построения и т.д.), какие догадки привлекаются, какие правдоподобные умозаключения используются. Все эти соображения сохраняют свою значимость и для XAI, но требуют детальной и весьма сложной проработки в конкретных приложениях.

Рассуждения на уровне «модель мира» практически не рассматриваются в рамках данной работы. Но в определённом смысле чертёж является «моделью мира» для семантического представления, определяя как «физически» можно менять «конструкцию» чертежа, сохраняя заложенные в семантике условия. Инструментальные средства реализации, используемые в проводимом исследовании: VBA EXCEL, JavaScript, библиотека JSXGRAPH. Средства программирования выбраны с учётом ориентации прикладной ориентации – создание образовательного ресурса для средней школы.

2. Эксперимент

С учетом выбранной предметной области сопоставление проведём на основе различного решения олимпиадной геометрической задачи [Trinh et. al., 2024]. Задача в течении ряда лет не поддавалась автоматическому решению, но недавно была решена системой AlphaGeometry с использованием LLM. Предварительное обучение языковой модели проводилось на 100 миллионах синтетически сгенерированных доказательств. Модель взаимодействует с решателем для генерации дополнительных построений и дедуктивных шагов.

Текст задачи: «Let ABC be an acute triangle. Let (O) be its circumcircle, H its orthocenter, and F the foot of the altitude from A. Let M be the midpoint of BC. Let Q be the point on (O) such that QH ⊥ QA and let K be the point on (O) such that KH ⊥ KQ. Prove that the circumcircles (O1) and (O2) of triangles FKM and KQH are tangent to each other».

Фрагмент решения приведён на рис. 1.

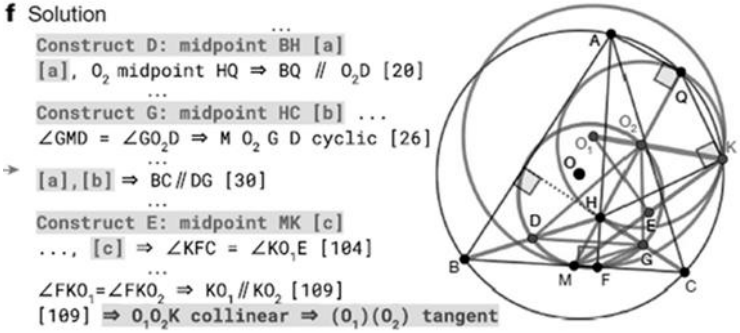


Рис. 1. Фрагмент решения (AlphaGeometry)

AlphaGeometry использует специальные средства («разности зависимостей») при генерации синтетических доказательств и выполняет около 10 миллионов синтетических шагов доказательства. При этом строятся вспомогательные точки, которые оцениваются как фактор очень сильного ветвления при чисто символьном выводе. Решение содержит 109 шагов, причём шаг 26 – спорный (рис. 1). Подробнее некоторые детали и сравнительные характеристики подходов рассматриваются в разделе 3.

Развиваемый нами подход предполагает ввод текста задачи на естественном языке (ЕЯ), лингвистическую обработку для получения семантического представления, автоматическое решение и его интерактивная визуализация. Подход базируется на компьютерном воплощении концепции Пойа, который подчеркивал, что в работе ученого догадка почти всегда

предшествует доказательству. Далее акцентируется идейная сторона нашего решения – догадки и их обоснования, а не теоремы – это техническая сторона. В силу некоторые моменты решения даются в упрощенном варианте.

В результате лингвистической обработки текста задачи (в русском переводе) формируется семантическое представление и визуализируется интерактивный чертёж, аналогичный приведённому на рис. 1. На первом шаге в нашем решении выполняется эмпирическая проверка корректности. Точки K, O1 и O2 находятся на одной прямой (с точностью возможностей графики), но главное, что при изменении параметров треугольника это расположение сохраняется. AlphaGeometry доказывает касание окружностей коллинеарностью KO1O2, наша система рассматривает равенство касательных (следует, если допустить доказываемое). Догадка и её эмпирическое подтверждение отображены на рис. 2.

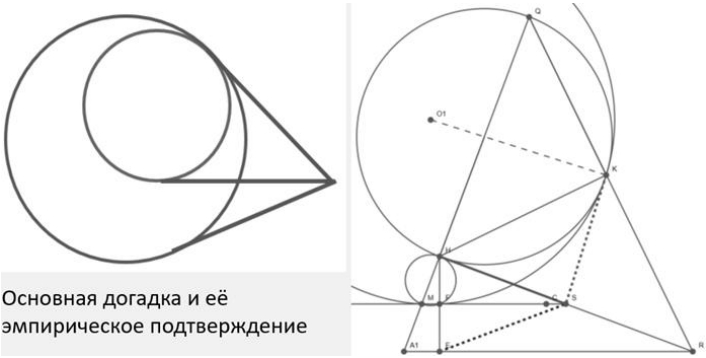


Рис. 2. Догадка и эмпирика

Элементы семантического описания и теоремы базы знаний снабжены ЕЯ-описаниями, что позволяет организовать их взаимодействие с помощью ключевых слов, перифраз или ассоциаций. Например, «три равных касательных из точки» имеет ассоциацию «окружность с центром в точке и радиусом, равным касательной». ЕЯ-описание семантической структуры задачи содержит фрагменты «окружности и равные касательные», «секущая окружности по точкам M и F» и т.п. По этим фрагментам из базы знаний извлекаются теоремы, которые могут оказаться полезными. Пример такой теоремы – на рис. 3.

Рис. 3. Граф теоремы и формализация ассоциации «часть подсказывает целое»

Важный шаг – нахождение центра эмпирически подтвержденной окружности отражён на рис. 4.

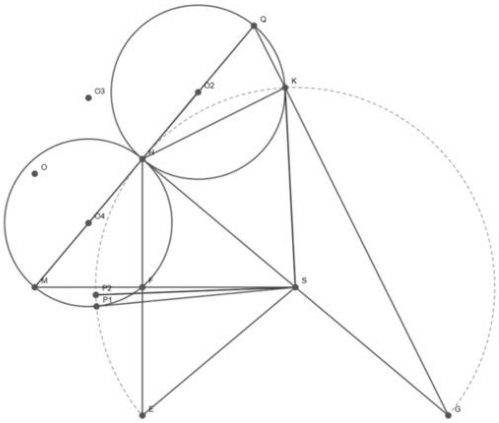


Рис. 4. Ключевая окружность с равными касательными

Текст теоремы: «общая секущая двух окружностей включает геометрическое место точек с равными касательными к этим окружностям». Разумеется, таких теорем может быть несколько. Важно, что если теорема не может быть непосредственно применена к текущей структуре задачи, то она может дать рекомендацию о включении в семантическую структуры недостающего фрагмента, с обоснованием ассоциации. Поясним конкретнее использование ассоциации. Пусть для структуры на рис. 3. в текущем семантическом представлении найдены элементы, которые означивают подграф, выделенный из общего графа теоремы. Тогда оставшаяся часть структуры является кандидатом на дополнительное построение (окружность и касательная).

После ряда неудачных попыток решатель, используя общее утверждение «центр окружности лежит на пересечении перпендикуляра из середины хорды и диаметра», строит центр S. Далее использование ассоциации (рис. 3.) позволяет построить окружность MFH, имеющую общую хорду с окружностью MFK. Это позволяет применить перспективную теорему и построить точки P1 и P2 (равные касательные на рис. 4). И, наконец, применяя свойства равных касательных, решатель доказывает их совпадение в точке K. Это завершает решение задачи (рис. 5).

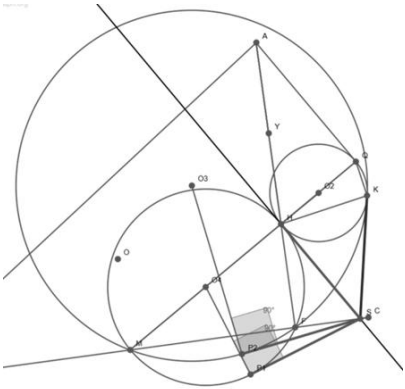


Рис. 5. Решение получено

Процесс решения выполнялся в диалоговом режиме. Однако практически все обращения к системе не являются подсказками решателю, шаги формируются с использованием эвристических, эмпирических и логических обоснований.

3. Обсуждение

Сравнение подходов проводится на весьма общем уровне, поскольку из публикаций неясны вопросы семантического представления, возможности интерактивной визуализации и эмпирической верификации, методы обоснования шагов решения, перспективы для образовательных целей. Можно согласиться с разработчиками AlphaGeometry в том, что современные подходы к геометрии (в области автоматизации решения задач) опираются в основном на символьные методы и разработанные человеком жестко закодированные эвристики поиска. Тем не менее развиваемый нами подход ориентирован на существенное уменьшение жесткости эвристик за счёт ассоциаций, эмпирических характеристик (чертёж, а в пределе рецепторы), онтология, снабженная ЕЯ-описаниями объектов. Последний пункт важен, поскольку для пользователя соответствующие применениям эвристик рассуждения должны быть предъявлены *вербально*.

Заключение

Комплексное использование нейронных сетей и традиционных методов ИИ (базы знаний, когнитивные подходы, прикладные онтологии и т.п.) обладает значительным потенциалом. Автор согласен с доктором Барнеттом [Matthew Barnett, 2025], в том, что в ближайшие три года будут разработаны системы ИИ, способные превзойти лучших математиков-людей в доказательстве произвольных математических теорем. Согласен автор и в том, что экономически ценные возможности ИИ будут отставать, а надежные агенты компьютерного управления крупными объектами появятся значительно позже, чем высококачественные модели математических рассуждений.

Прикладная значимость развиваемого исследования для образовательного процесса состоит в том, что постановка цели, средства её достижения (теоремы, доп. построения), учёт наводящих и индуктивных соображений (догадок), четкое разграничение догадки и доказательства, структуризация доказательства, вычленение ключевых моментов для устранения когнитивного диссонанса важны для любого специалиста. Создание образовательного ресурса, базирующегося на результатах данной работы, позволит продвинуть элементы ИИ непосредственно в среднее школьное образование.

Список литературы

[Chen et. al., 2025] Chen M., Tworek J., Jun H., Yuan Q., and others. DeepSeek-R1. Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. – https://arxiv.org/abs/2501.12948, Submitted on 22 Jan 2025.

[Ma et. al., 2025] Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, Matei Zaharia. Reasoning Models Can Be Effective Without Thinking. – https://doi.org/10.48550/arXiv.2504.09858, Submitted on 14 Apr 2025.

[Chen et. al., 2024] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu and others: Learning to reason with LLMs, September 12, 2024, Release. – https://openai.com/index/learning-to-reason-with-llms/.

[Huang et. al., 2006] Huang W., Hong S.H., Eades P. Predicting Graph Reading Performance: A Cognitive Approach // In: Proc. Asia Pacific Symposium on Information Visualization (APVIS2006), Tokyo, Japan, 2006. – P. 207-216. – doi: 10.1145/1151903.1151933 (*статья в сборнике трудов конференции на англ. языке*)

[Højer et. al., 2025] Bertram Højer, Oliver Jarvis, Stefan Heinrich. Improving Reasoning Performance in Large Language Models via Representation Engineering. – https://doi.org/10.48550/arXiv.2504.19483, Submitted on 28 Apr 2025.

[Bajorath, 2025] Jürgen Bajorath. From scientific theory to duality of predictive artificial intelligence models, April 2025 Cell Reports Physical Science 6(4):102516, [Iman Mirzadeh, et. al.] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, et. al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, Submitted on 7 Oct 2024, preprint. – https://arxiv.org/abs/2410.05229.

[Rost et. al., 2022] Marvin Rost, Tarja Knuuttila: Models as Epistemic Artifacts for Scientific Reasoning in Science Education Research, Educ. Sci. – 2022. – 12(4), 276; [Iman Mirzadeh, et. al.] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, et. al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, Submitted on 7 Oct 2024, preprint. – https://arxiv.org/abs/2410.05229.

[Krell et. al., 2022] Moritz Krell, Andreas Vorholzer, Andreas Nehring. Scientific Reasoning in Science Education: From Global Measures to Fine-Grained Descriptions of Students' Competencies // Education Sciences. – January 2022 – 12(97). – P. 1-8. – DOI: 10.3390/educsci12020097.

[Upmeier zu Belzen, et. al., 2021] Annette Upmeier zu Belzen, Paul Engelschalt, Dirk Krüger: Modeling as Scientific Reasoning – The Role of Abductive Reasoning for Modeling Competence // Education Sciences. – September 2021. – 11(9):495. – DOI: 10.3390/educsci11090495.

[Mirzadeh et. al.] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, et. al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, Submitted on 7 Oct 2024, preprint, [Iman Mirzadeh, et. al.] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, et. al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, Submitted on 7 Oct 2024, preprint. – https://arxiv.org/abs/2410.05229.

[Trinh et. al., 2024] Trinh T.H., Wu Y., Le Q.V., et al. Solving Olympiad geometry without human demonstrations // Nature. – 2024. – 625. – P. 476-482 – https://doi.org/10.1038/s41586-023-06747-5.

[Zhang et. al., 2024] Chi Zhang, Jiajun Song, Siyu Li, et. al. Proposing and solving olympiad geometry with guided tree search. – https://arxiv.org/abs/2412.10673, Submitted on 14 Dec 2024.

[Polya, 1981] Polya G. Mathematical Discovery: On Understanding, Learning and Teaching Problem, Solving (Combined Edition). – Willey, New York, 1981. – 432 p.

[Kurbatov et. al., 2024] Sergey S. Kurbatov, Mikhail A. Gilmendinov. Cognitive Dissonance in Solving Planimetric Problems // In: Proceedings of 8th Computational Methods in Systems and Soft-ware. – 2024. – Vol. 2. – P. 163-172. – https://doi.org/10.1007/978-3-031-77411-9_15.

[Kurbatov, 2023] Kurbatov S. Paraphrasing in the system of automatic solution of planimetric problems: data analytics in system engineering // In: Proceedings of 7th Computational Methods in Systems and Software. – 2023. – Vol. 3. – P. 217-225. – https://doi.org/10.1007/978-3-031-53552-9_20.

[Phillips et al., 2020] Jonathon Phillips P., et al. Four Principles of Explainable Artificial Intelligence, NIST. – 2020. – https://doi.org/10.6028/NIST.IR.8312-draft.

[Hu et al., 2021] Brian Hu, et al. XAITK: the explainable AI toolkit // Applied AI Letters. – October 2021. – Vol. 2(4). – https://doi.org/10.1002/ail2.40.

[Pradeepta, 2022] Pradeepta M. Practical Explainable AI Using Python: Artificial Intelligence Model Explanations Using Python-based Libraries. – Apress Media LLC, 2022. – 356 p. – ISBN13: 978-1-4842-7157-5.

[Barnett, 2025] Matthew Barnett: The promise of reasoning models, Gradient Updates – Epoch AI, Feb 28, 2025. – https://epoch.ai/gradient-updates/the-promise-of-reasoning-models.

УДК 004.89
doi: 10.15622/rcai.2025.020

ЕЩЕ ТРИ ВОПРОСА[1] (НА ПОНИМАНИЕ), АДРЕСОВАННЫЕ «ТОВАРИЩАМ ПО ПАРТИИ»

В.К. Финн (*v.k.finn@yandex.ru*)
М.А. Михеенкова (*m.mikheyenkova@yandex.ru*)
М.И. Забежайло (*m.zabezhailo@yandex.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

Обсуждаются представления об интеллектуальности систем искусственного интеллекта (ИИ), методах оценки качества формируемых ими результатов, а также некоторые проблемы организации экспертизы проектов и решений в области ИИ.

Ключевые слова: искусственный интеллект, исследования и разработки, экспертиза, оценки качества формируемых системами ИИ решений, подготовка кадров.

Сегодня словосочетание Искусственный Интеллект (ИИ) прочно вошло и закрепилось в самых различных составляющих нашей обыденной жизни. И дело не только в том, что на тему ИИ, как принято говорить, не высказывает тех или иных суждений, по-видимому, лишь только абсолютно ленивый. Бесспорный факт, что за последние примерно четверть века область внимания, а также все более смелых ожиданий, связанных с применениями активно развивающихся ИИ-технологий и решений, фантастически расширилась. Если 80 лет назад эта проблематика увлекала преимущественно лишь пионеров-исследователей да и, пожалуй, военных (для который во все времена характерен интерес к перспективным технологиям и решениям), то сегодня уже могут потребоваться существенные усилия, чтобы указать такую область окружающей нас жизни, в которой на обсуждались бы перспективы применения тех или иных ИИ-решений.

Сопутствующими эффектами в такой ситуации стали hype[2] вокруг проблематики ИИ вместе с целым набором мифов и легенд, циркулирую-

[1] См. также [Забежайло, 2023].

[2] Бесконтрольный ажиотаж (англ.).

щих в обществе (в том числе – даже в профессиональной среде исследователей). Перспективы дальнейшего развития в этом направлении текущей ситуации не могут не вызывать у экспертов вполне обоснованных опасений. Один из очевидных факторов риска здесь – это рост «пузыря» завышенных ожиданий общества в части практических результатов применения ИИ-систем и решений, дополняемый перспективами последующего глубокого разочарования «широких масс интересантов», уставших дожидаться обещанных им фантастических прорывов и достижений.

Настораживающие профессионалов элементы реальности окружающей нас действительности – это заполнившие инфосферу всевозможные обсуждения фантастических преимуществ интеллектуализации всех сфер нашей жизни: использования интеллектуальных пылесосов или холодильников, интеллектуальных[3] домов и интеллектуальных производственных технологий, необходимости прямо завтра перейти на использование интеллектуальных систем управления, возникающие повсеместно кафедры и отделы интеллектуальных систем, а также многое другое. (Справедливости ради следует отметить, что в ответственных ситуациях – в частности, при использовании ИИ-технологий и решений в критически значимых приложениях – обычно используется более корректная терминология: например, упоминаются *системы с элементами ИИ* и т.п.).

К сожалению, используемая на текущий момент (причем даже в профессиональном сообществе ИИ-специалистов) терминология далеко не всегда предлагает однозначное (и операционально[4] корректное) толкование задействованных понятийных конструкций. Следствием этого оказываются не только возникновение и поразительная живучесть целого ряда мифов о некоторых аспектах проблематики ИИ[5], но и ряд проблем, связанных с завышенными и, вообще говоря, мало оправданными ожиданиями от применения ИИ-систем в практически значимых приложениях.

Не менее запутанной оказалась ситуация с оценкой достаточности оснований для принятия результатов, формируемых системами ИИ. Ограниченная применимость традиционных средств обоснования результатов работы ИИ-систем – использования математического инструментария

[3] Справедливости ради следует заметить, что в актуально используемой лексике некоторых национальных языков части подобных смысловых неоднозначностей (а иногда – и очевидных коллизий) удается избежать. Так, например, английский термин *smart* позволяет «смягчить» толкование «интеллектуальности» поведения некоторых технических устройств (называемых *smart devices*) в соотнесении с представлениями об интеллектуальности поведения человека в тех или иных требующих «использования интеллекта» ситуациях.

[4] Позволяющее аргументированно отделить удовлетворяющие ему объекты от неудовлетворяющих

[5] См., в частности, [Финн и др., 2023] и др.

дедуктивных доказательств или же средств стандартного статистического анализа данных при решении прикладных задач в открытых, постоянно пополняемых новой информацией предметных областях – стала драйвером развития новых подходов, методов и математических моделей. Актуализировалась проблематика обеспечения доверия и доверенности систем ИИ. Использование инструментов доказательства, понимаемого как приведение формируемого ИИ-системой заключения к неоспаримости относительно заданного контекста (например, конкретного множества накапливаемых фактов эмпирического исследования и др.) расширяется за счет его интеграции со средствами формирования содержательной интерпретации получаемых результатов, а также их неформального объяснения и аргументации. Стандартом de facto стало использование так называемых контрафактуальных схем обоснования, учитывающих как сходства, так и различия в описаниях анализируемых эффектов. Бесспорный триумф использования технологий искусственных нейронных сетей (ИНС) и так называемых больших языковых моделей (БЯМ) при решении важных прикладных задач вывел в разряд критически значимых потребность в надежном и неоспариваемом обосновании формируемых ими в каждом конкретном случае результатов и заключений (потребность все-таки иметь возможности формировать приемлемый ответ на классический вопрос «*Что есть истина*?»).

Очевидным (причем теперь уже не только для разработчиков, но и для более широкого круга Лиц, Принимающих Решения (ЛПР), индустриальных заказчиков, ответственных руководителей органов государственного управления, представителей структур-регуляторов и др.) барьером на пути разработки и развития ИИ-систем для различных индустриальных приложений оказалась проблема подготовки квалифицированного заказчика – как собственно ЛПР, способного и готового принять на себя ответственность за принимаемые им решения и их последствия, так и обеспечивающих его компетентной поддержкой экспертов (в том числе – способностью обеспечить со стороны заказчика корректную постановку решаемой задачи для исполнителя, а также проконтролировать корректность представляемых исполнителем результатов решения). *Как*, *чему* именно, *где* и *какими силами* готовить таких специалистов – вот еще один комплекс вопросов, без эффективного решения которых отвечать на технологические вызовы развития ИИ-систем и решений буде крайне затруднительно.

Таким образом, представляется актуальным и полезным уточнить[6] понимание экспертами профессионального ИИ-сообщества (см. ранее, [Финн и др., 2023], [Забежайло, 2023] и др.) еще по трем позициям:

(i) Как понимать термин *интеллектуальность* систем ИИ (СИИ)?

[6] Например, через публикации, обсуждения на ИИ-конференциях и семинарах, …

(ii) Какими средствами следует воспользоваться для *оценки приемлемости* (надежности, обоснованности, …) заключений и рекомендаций, формируемых конкретной СИИ?

(iii) Что полезно предпринять для повышения уровня ИИ-подготовки ЛПР и экспертов, определяющих направления развития и приоритеты финансирования ИИ-исследований и разработок в нашей стране?

Вариант ответа на каждый из этих трех вопросов и будет предложен ниже.

1 Так в чем же, собственно, интеллект? (О критериях интеллектуальности систем ИИ)

К классу систем ИИ (СИИ) представляется целесообразным отнести все такие компьютерные системы, которые позволяют решать те или иные задачи, относимые к ИИ как области исследований и разработок. При этом достаточно часто вместе с термином СИИ также используется термин *интеллектуальные системы* (ИС). Некоторые авторы рассматривают эти два термина – СИИ и ИС – как синонимы, что на наш взгляд представляет собою глубокое заблуждение. Если рассмотреть понимание этих двух терминов в самом общем виде (так сказать с высоты птичьего полета), то в случае СИИ мы имеем дело с эмуляцией (своего рода «протезированием») в том числе и сугубо вычислительными средствами отдельных функций интеллекта человека (естественного интеллекта – ЕИ), используемых им при решении «трудных» (как принято считать. требующих «интеллекта») задач. В таком случае, в чем же *интеллектуальность* СИИ, позволяющих аргументированно отнести их к классу ИС?

Представляется вполне естественным считать, что интеллектуальность СИИ определяется возможностями *воспроизведения* (*имитации*) в них (*компьютерными средствами* – !) тех или иных *функций* (*способностей*) ЕИ. При таком подходе представления об ИИ – в т.ч., определение ИИ как исследования и разработки, которые направлены на имитацию и усиление функций ЕИ компьютерными средствами, – открывает возможности для целенаправленного и контролируемого[7] построения систем ИИ. Однако, при этом потребуется уточнить наши представления об ЕИ (в том числе – перечислить или же привести примеры способностей ЕИ, характеризующих собственно понятие *интеллект*). Не менее значимо и разделение таких способностей на *допускающие* и *не допускающие* (по крайней мере на текущем уровне развития) моделирование компьютерными средствами.

[7] Позволяющего обходить различные мифы и угрозы типа, например, «восстания интеллектуальных пылесосов» или же всеобщей безработицы, которую сулит широкое внедрение ИИ-систем и решений.

Двигаясь в обозначенном направлении – уточняя наши представления о характеристиках ЕИ, воспользуемся хорошо известным подходом[8] феноменологического характера. Еще в 20-е годы прошлого века психологи, изучая структуру ЕИ, предложили определение «*интеллект – это то, что измеряется тестами интеллекта*» (см., например, работу Э.Боринга [Boring, 1923], увидевшую свет в 1923 году, и др.). Процедурный каркас феноменологического подхода Э. Боринга – это психологические тесты и обработка их результатов математическими методами статистики (приоритет – метод главных компонент [Pearson, 1901], использованный для выявления факторов наибольшего влияния) позволил сформировать «экосистему» из 10 так называемых «широких» способностей (таких, например, как «*работа» с памятью* – запоминание и «извлечение» информации из памяти, *способности к восприятию* – визуальная и слуховая обработка информации, *чтение* и *письмо*, *способности* оперировать *количественными величинами*, *работа* со *знаниями*, а также способность выстраивать гибкие *последовательности логичных «шагов» мысли* и др.), дополненных сопутствующими им «узкими» способностями, каждая из которых детализирует соответствующие «широкие» способности. Современная теория структуры человеческого интеллекта, предложенная Кэттеллом, Хорном и Кэрроллом [Schneider, et al 2018], стала естественным развитием и обобщением исторически сложившегося (см. [Boring, 1923] и др.) подхода. При этом следует особо отметить, что центральная роль в теории Боринга-Кэттелла-Хорна-Кэрролла отводится умению человека *рассуждать* и *формировать понятия*, решая *новые задачи* в *незнакомых* ситуациях.

Возвращаясь к данному выше определению ИИ как имитации и усиления компьютерными средствами системообразующих способностей ЕИ, обратим внимание на разделение формализуемых и не формализуемых на текущий момент свойств и способностей ЕИ. В работах [Финн, 2023] и др. представлен перечень из 13 способностей ЕИ, эффективно переносимых сегодня на компьютер. Вот краткий вариант их перечня:

(1) обнаружение существенного в данных;
(2) порождение последовательности «цель – план – действие»;
(3) подбор посылок, релевантных цели рассуждения;
(4) способность к рассуждению: вывод следствия из посылок;
(5) синтез и взаимодействие познавательных процедур (например, индукции, аналогии и абдукции с последующим применением дедукции);

[8] См. предложенное Д.С.Миллем [Милль, 2007] определение понятия *theoretical economy* как область исследований, которыми занимаются *theoretical economists*, а также более поздние «реинкарнации» этого методологического приема (физика как то, чем занимаются физики, и т.д.).

(6) рефлексия – оценка знаний и действий (как рациональная и аргументированная реакция на состояние знаний и результаты действий)
(7) способность к объяснению – ответ на вопрос «*Почему*?»;
(8) аргументация при принятии решений;
(9) познавательное любопытство и способность к распознаванию;
(10) способность к обучению и использование памяти;
(11) способность к интеграции знаний для образования концепций и теорий;
(12) способность к уточнению неясных идей – преобразованию их в понятия;
(13) способность к изменению системы знаний при получении новых знаний и изменений познавательных ситуаций.

Параллельно следует принять во внимание, что (по крайней мере, на текущий момент) за рамками области компьютерной формализации способностей ЕИ оказываются не только ряд таких критически значимых характеристик интеллекта как, например, интуиция и воображение, но и совокупность функций мозга, характеризующих высшую нервную деятельность человека.

Следуя далее уже намеченным путем и принимая данное выше определение ИИ (как области исследований и разработок, ориентированных на имитацию и усиление познавательных функций человека компьютерными средствами), можно предложить операциональный *критерий интеллектуальности* систем ИИ:

> *отнесение СИИ к классу ИС обеспечивает реализация в оцениваемой СИИ рассматриваемых – см. выше и* [Финн, 2023] *– способностей ЕИ (в том числе – приближенная имитация тех или иных из перечисленных ЕИ-способностей компьютерными средствами*).

Упрощенная (и как следствие – «радикализованная») версия Критерия: *СИИ можно считать интеллектуальной, если она способна реализовать тот или иной вариант рассуждений* (т.е. ИС – это «*рассуждающая*» СИИ).

Разумеется, авторы допускают существование и других – отличных от предложенного ими выше – определений ИИ и, как следствие, других вариантов Критерия интеллектуальности СИИ. При этом предполагается, что каждая такая версия Критерия (прежде чем стать «инструментом» поддержки принятия соответствующих оценочных решений) должна пройти публичное критическое обсуждение в профессиональном сообществе специалистов, занятых практическими исследованиями и разработками в области ИИ.

2. «Чем же сердце, наконец, успокоится?» (Об оценке достаточности оснований для принятия результатов, формируемых системой ИИ)

Необходимость применять СИИ в анализе открытых – постоянно пополняемых новыми данными (т.е. так называемый эффект Open, характерный для Big Data) – предметных областей повлекла существенное расширение соответствующего «инструментария» – подходов, математических моделей, методов и алгоритмов оценки достаточности оснований для принятия результатов, формируемых СИИ в той или иной конкретной ситуации. Осознано отсутствие возможностей использовать в условиях эффекта Open целый ряд традиционных средства компьютерного анализа данных. Так, для применения доказательств дедуктивного типа открытый характер анализируемых данных порождает в общем случае непреодолимые препятствия на пути формирования аксиоматического описания предметной области, а уж о доказательстве утверждений типа теорем о полноте в таких ситуациях вообще нет оснований говорить. Другой пример – использование «инструментов» статистического анализа данных натыкается в условиях эффекта Open на отсутствие возможностей формировать генеральные совокупности, репрезентативные выборки из которых могли бы стать надежным основанием для обучения СИИ.

Как следствие, внимание исследователей и разработчиков развернулось в сторону подходов, ориентированных на содержательную интерпретацию, неформальное объяснение и аргументацию результатов, формируемых системами ИИ. Отсечение не интерпретируемых результатов стало эффективным инструментом борьбы с артефактами переобучения. Объяснение (как ответ на вопрос «Почему»?) позволяет использовать изначально скрытые в анализируемых данных эмпирические зависимости причинно-следственного типа как базу для обоснования приемлемости формируемых системой ИИ выводов и заключений. Все большую популярность завоевывают решений на основе так называемого контрафактуального подхода [Pearl, 1999], [Pearl, 2005], [Höfler, 2005 и др.], позволяющие проверять порождаемые рекомендации СИИ на предмет их фальсифицируемости. Активно развивается проблематика – подходы, модели и алгоритмы – оценки доверия (в значении *believe*) и доверенности (в значении *trust*) результатов, формируемых той или иной системой ИИ. В основе этих решений – те или иных механизмы сравнения сходства и различия в описаниях анализируемых (проверяемых на фальсифицируемость) свойств и эффектов.

К сожалению, по-прежнему открытым остается в общем случае вопрос о достаточности оснований для принятия результатов, порождаемых с использованием ИНС или Больших Языковых Моделей (БЯМ). В случае

БЯМ отсутствие надежных средств формирования ответа на классический вопрос «Что есть истина?», адресуемый к предлагаемым ими результатам, породил в последние несколько лет целую индустрию эвристических сервисов, которые призваны помочь пользователям в использовании БЯМ (см., например, prompt engineering, retrieval augmented generation и др.).

Все более широко востребованными при оценке достаточности оснований для принятия результатов, генерируемых системами ИИ, становятся аргументационные подходы. Среди направлений, тесно связанных с успехами в создании объяснимых (приемлемых) решений (а, следовательно, заслуживающих доверия систем), заметное место заняли формальные теории аргументации [Čyras et al., 2021]. Аргументация как таковая служит двум целям: *обоснование* и *убеждение*. В первом случае предъявление посылок вывода фиксирует состоятельность позиции. В этом смысле дедуктивное доказательство, *гарантирующее* истинность заключения, может также рассматриваться как частный случай аргументационной схемы. Однако, как уже говорилось, в открытом мире истинность заключения может ставиться под сомнение. Здесь можно рассчитывать лишь на ту или иную степень его вероятности или правдоподобия и, соответственно, приемлемости. Значительный вклад в развитие и применение аргументации в ИИ и информатике вносят так называемые «системы абстрактной аргументации» [Dung, 1995]. Ключевым понятием в этой структуре является атака аргументов – абстрактное формальное отношение, представляющее возможность опровержения одного аргумента другим. В дальнейших разработках системы абстрактной аргументации были существенно расширены: системы биполярной аргументации используют отношения атаки и поддержки, количественной – с учётом «веса» аргументов [Baroni et al., 2019], обобщённой аргументации [Gabbay, 2016] – любое количество отношений. Системы нечёткой аргументации позволяют представить относительную силу взаимосвязи между аргументами и степень их принятия [Janssen et al., 2008]. Генеративные модели с использованием байесовской аргументации (в рамках байесовского подхода к рассуждениям) оказываются действенными для достижения устойчивости к состязательным атакам [Cerutti, 2022]. Тесная связь логики и аргументации, известная ещё со времён Аристотеля, поддерживается последними исследованиями в области представления знаний и рассуждения в ИИ (см., например, [Вагин, 2019], [Besnard et al., 2020], [Dastani et al., 2020]). В [Финн, 2021а] предложены оригинальные отечественные варианты четырёхзначных логик с аргументационной семантикой и неассоциативными связками, в [Финн, 2021б] развита схема многоуровневой аргументации.

Аргументация как когнитивный феномен является предметом изучения различных дисциплин – философии, лингвистики, юриспруденции, политологии, когнитивных наук и т.д. Это расширяет горизонты фор-

мальных подходов к принятию решений с помощью аргументации, обоснованию на основе аргументов с учётом базовых (внешних) знаний, формированию иерархических систем аргументации (например, на основе доверия источникам или неких количественных оценок), аргументационного диалога «*pro*» и «*contra*» и т.д. На базе аргументации создаются рекомендательные системы, классификаторы, планировщики. В последнее время наблюдается также усиление таких систем моделями ML для формирования аргументации в рамках обучения на положительных и отрицательных примерах или разнонаправленных вознаграждений/штрафов в обучении с подкреплением.

Таким образом, выбор средств, которыми следует воспользоваться для *оценки приемлемости* (надежности, обоснованности, …) заключений и рекомендаций, формируемых конкретной СИИ, – еще один актуальный вопрос для обсуждения в профессиональном ИИ-сообществе.

3. «А судьи – кто?» (Об экспертизе и кадрах)

Опыт последних примерно 5 лет в части организации и проведения в нашей стране крупных конкурсов по проблематике ИИ высветил критически значимую проблему. Речь о несоответствии уровня квалификации (в обсуждаемой здесь области) участвовавших в проведении этих конкурсов ответственных руководителей и привлекаемых ими экспертов целям и задачам развития данного направления науки и технологий в РФ. Собственно, и ранее, например, в период существования РФФИ, нередко решения о финансировании или же, наоборот, не финансировании тех или иных групп исследователей принимались не на основе беспристрастной оценки качества и перспектив предлагаемых ими проектов, а по принципу «близости» к соответствующим ЛПР или же с учетом «конкуренции» определенных исследовательских сообществ (а также их лидеров). При этом, как это было, например, в случае конкурсов РФФИ второй половины 2010-х годов, «квантование» выделяемых в таким образом организованных процедурах в виде грантовой поддержки сумм в диапазоне 3-5 миллионов рублей не оказывало существенного влияния на состояние ИИ-исследований и разработок в целом по стране. Однако, при выходе на проекты с общим финансированием порядка 1 млрд. рублей (см., в частности, Конкурс на организацию так называемой первой волны Центров ИИ, запущенный постановлением Правительства РФ летом 2021 года) влияние уровня экспертизы на принимаемые решения приобрело критически значимый характер. Если, например, взглянуть на имеющую место ситуацию, рассуждая по аналогии, то любому непредвзятому наблюдателю было бы странно наблюдать за экспертизой и принятием решения в ситуации, когда ведущим в экспертизе плана проведения нейрохирургической операции на головном мозге человека был бы избран, например, хирург-онколог, который является бесспорным специалистом в области

хирургического лечения рака простаты. Мотивация ЛПР в крупных конкурсах обсуждаемого типа, сводимая к фигуре рассуждений вида «*в ведущих промышленно-развитых странах мира ИИ-методы и технологии класса NNN активно развиваются, следовательно, и мы в нашей стране должны будем пойти именно этим путем*[9]», разумеется, не лишена здравого смысла. Однако, она не учитывает ряд критически значимых для нашей страны факторов – наличия своих научных и инженерных школ с длительной успешной историей развития и собственными достижениями, асимметрии технической вооруженности отечественных исследователей в сравнении с основными зарубежными конкурентами[10] и др.

В чем причины такой ситуации? Одним из приоритетных факторов, на наш взгляд, является проблема недостаточного уровня подготовки, а более широко – формирования квалифицированного заказчика и консультирующих его экспертов. Где взять квалифицированных экспертов? Как, кому, по каким направлениям и где готовить кадры для ИИ как области исследований и разработки? Таким образом, хорошо известный вопрос «А судьи – кто?» легко трансформируется в рамках национальных усилий по развитию ИИ-решений и систем в достаточно жесткий вопрос об экспертизе и кадрах.

Опыт некоторых крупных российских ИТ-проектов последних пяти лет (см. например, формирование Центров НТИ, Лидирующих Исследовательских Центров по заданным направлениям и тематикам, 6 центров ИИ так называемых первой «волны», а также последовавшие за эти конкурсы «второй» и «третьей» волн и др.) продемонстрировал, что:

- реальной альтернативой формально открытым конкурсам в условиях небесспорной экспертизы (а также ориентации ЛПР такого конкурса в первую очередь на «статус» организаций победителей, а уж потом на содержание предлагаемых работающими в них исследовательскими группами проектных предложений) может быть, например, прямое включение в государственное задание «достойным» организациям соответствующих работ (обеспеченных соответствующим финансированием). При таком – альтернативном конкурсному – подходе можно избежать непродуктивного расхода сил и времени соискателей на подготовку объемных конкурсных заявок[11], а также дать возможности получателям соответствующего государственного заказ привлекать (разумеется, под свою ответственность) полезных им высоко квалифицированных соисполнителей по конкретным направлениям работ;

[9] Грубо говоря, схема принятия решения выглядит примерно следующим образом: «*Слышали про ИНС. Это – наше 'все'. Будем развивать это направление*»

[10] См., например, уровень обеспеченности вычислительными ресурсами отечественных разработчиков ИНС-решений в сравнении с конкурентами в США и КНР.

[11] Часто используемых лишь при аргументации отказа соискателю

- критически значимым для последующего выполнения проектных работ и последующих процедур сдачи-приемки получаемых результатов оказывается организационно-контентный процесс, позволяющий ответственно сформировать задачи для потенциальных исполнителей. Очевидные критичные аспекты здесь – это, в частности:
  - кто именно (какие требования к квалификации, в каких именно конкретных направлениях, …) и
  - как именно сформирует действительно необходимые государству цели и задачи (деньги ведь – бюджетные, и отчетность за их использование предусматривает, в том числе, и достаточно жесткие меры государственного контроля и ответственности);
  - кто именно (квалификация, в каких направлениях науки и технологий) и по каким критериям будет проводить экспертизу конкурсных заявок?
- вопрос о том, как (по каким критериям) принимать решение о выборе победителей (дать «своим»; выбрать необходимые для индустрии решения, отобрать те работы, которые способны обеспечить возврат инвестиций, …) на национальной уровне проектов, оказывается критически значимым;
- «тонкой» настройки и аккуратной организации требует мониторинг исполнения проектов: важно располагать «инструментами» управления, позволяющими не потерять вложенные инвестиции, и при этом иметь эффективные возможности обходить возникающие по ходу проекта трудности и проблемы;
- в каждом конкретном проекте следует принимать особое – учитывающее специфику данного проекта – решение о том, как организовать процедуры сдачи-приемки (в т.ч. – научную, техническую и финансовую экспертизу), чтобы обеспечить успешное завершение проекта.

Итак, отдельного обсуждения в профессиональном сообществе ИИ-исследователей и разработчиков заслуживают, на наш взгляд, три проблемы:

(1) подготовки квалифицированного заказчика (включая формирование среды, которая способна обеспечить заказчика-ЛПР профессиональной экспертизой в требующемся ему круге вопросов или предметной области);

(2) формирования профессиональной экспертизы (чтобы иметь неоспариваемый ответ на вопрос «*А судьи – кто*?»);

(3) систематической подготовки квалифицированных исполнителей (чтобы иметь неоспариваемое представление о том, кто именно и чему именно учит, ориентируясь, как минимум на 5 профилей подготовки специалистов по ИИ: «теоретики», разработчики программных систем ИИ, квалифицированные пользователи ИИ-систем, специали-

сты по управлению проектами в области разработки и внедрения ИИ-систем и решений, а также – специальные курсы повышения квалификации в области ИИ для специалистов различного профиля – от государственных чиновников до преподавателей учебных заведений, обучающих проблематике ИИ).

Заключение

Сегодня ИИ как область исследований и разработок по-прежнему находится в стадии становления, характеризуемой еще только формируемой собственной проблемно-ориентированной терминологией, а также специфической системой понятий и методов исследования. Именно по этой причине представляется важным определиться в нашей профессиональной среде ИИ-специалистов с актуальными именно для нас приоритетами понимания перспектив и выбора путей дальнейшего развития. Было бы полезно сформировать общее видение того, на чем именно сфокусироваться, чтобы не отстать «навсегда», наслаждаясь своими былыми достижениями и, не жалея сил и времени, поддерживая готовность «к лихой кавалерийской атаке на танки противника[12]».

Список литературы

[Вагин, 2019] Вагин В.Н. Элементы теории аргументации и её роль в интеллектуальном анализе данных // В кн.: Вагин В.Н. Знания и убеждения в интеллектуальном анализе данных. – М.: Физматлит, 2019. – 536 с.

[Забежайло, 2023] Забежайло М.И. Три вопроса (на понимание), адресованные «товарищам по партии» // КИИ-2022 (21–23 декабря 2022 г.). Труды конф. / под ред. В.В. Борисова, Б.А. Кобринского. – М.: РАИИ, 2022. – Т. 1. – С. 280-289.

[Милль, 2007] Милль Д.С. Об определении предмета политической экономии и о методе исследования, свойственном ей // В книге: Основы полит. экономии с нек. приложениями к социальной философии. – М.: Эксмо, 2007. – С. 985-1023.

[**ПК**] Польские кавалерийские бригады в 1939 году. – https://chestnut-ah.livejournal.com/842458.html.

[Финн, 2021а] Финн В.К. Стандартные и нестандартные логики аргументации // В.К. Финн. Искусственный интеллект: методология, применение, философия. – М.: ЛЕНАНД, 2021. – 468 с. – С. 337-363.

[12] Поучительным примером здесь может быть, в частности, ситуация с различными видами кавалерии в качестве основной ударной силы польской армии второй половины 30-х годов. По данным [ПК] в сентябре 1939 года относительная доля кавалерийских частей в польской армии была в 5 раз выше, чем в германской: примерно 10,5% к 2,1%. Такая разница была следствием как особенностей актуальной на тот момент военной доктрины Польши, так и старых национальных кавалерийских традиций.

[Финн, 2021б] Финн В.К. Двенадцать тезисов об аргументационных системах // Финн В.К. Интеллект, информационное общество, гуманитарное знание и образование. – М.: ЛЕНАНД, 2021. – 464 с. – С. 191-221.

[Финн, 2023] Финн В.К. Искусственный интеллект. Методология, применения, философия. – 2-е изд. испр. и доп. – М.: ЛЕНАНД, 2023. – 464 с.

[Финн и др., 2023] Забежайло М.И., Михеенкова М.А., Финн В.К. О некоторых актуальных мифах современного искусственного интеллекта // Труды XXI Нац. по ИИ с междунар., Смоленск 16-20 октября 2023 г. – Т. 1. – С.190-200.

[Baroni et al., 2019] Baroni P., Rago A., and Toni F. From fine grained properties to broad principles for gradual argumentation: A principled spectrum // Int. J. Approx. Reason. – 2019. – 105. – P. 252-286.

[Besnard et al., 2020] Besnard P., Cayrol C., Lagasquie-Schiex M.-C. Logical theories and abstract argumentation: A survey of existing works // Argument & Computation. – 2020. – Vol. 11(1-2). – P. 41-102.

[Boring, 1923] Edwin G. Boring. Intelligence as the Tests Test It // New Republic. – 1923. – 36. – P. 35-37.

[Cerutti, 2022] Cerutti F. Supporting Trustworthy Artificial Intelligence via Bayesian Argumentation // Lecture Notes in Computer Science. – 2022. – Vol. 13196. – P. 377-388. – https://doi.org/10.1007/978-3-031-08421-8_26.

[Čyras et al., 2021] Čyras K., Rago A., Albini E., Baroni P., Toni F. Argumentative XAI: A Survey // Proc. of the Thirt. Int. Joint Conf. on AI (IJCAI-21). – P. 4392-4399.

[Dastani et al., 2020] Dastani M., Dong H., van der Torre L. (eds). Logic and Argumentation // Lecture Notes in Artificial Intelligence. – 2020. – Vol. 12061.

[Dung, 1995] Dung P.M. On the Acceptability of Arguments and its Fundamental Role in Non-Monotonic Reasoning, Logic Programming, and *n*-person Games // Artificial Intelligence. – 1995. – Vol. 77. – P. 321-357.

[Gabbay, 2016] Gabbay D.M. Logical foundations for bipolar and tripolar argumentation networks: preliminary results // J. Log. Comput. – 2016. – 26(1). – P. 247-292.

[**Höfler, 2005**] Höfler M. Causal inference based on counterfactuals // BMC Med. Res. Methodol. – 2005. – 5, 28. – https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-5-28.

[Janssen et al., 2008] Janssen J., De Cock M., Vermeir D. (2008) Fuzzy Argumentation Frameworks // Proceedings of the 12th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. – P. 513-520.

[**Pearl, 1999**] Pearl, J. Causality: models, reasoning, and inference. – N.-Y.: Cambridge Univ. Press, 2000. – 384 p.

[**Pearl, 2005**] Pearl J. Probabilities of causation: three counterfactual interpretations and their identifications // Synthese. – 1999. – 121. – P. 93-149.

[Pearson, 1901] Pearson K. On lines and planes of closest fit to systems of points in space // Philos. Mag. – 1901. – 2(11). – P. 559-572.

[Schneider et al 2018] Schneider W.J. and McGrew K.S. The Cattell-Horn-Carroll Theory of Cognitive Abilities // in Contempor. Intellect. Assessm.: Theories, Tests, and Issues. – 4th ed., ed. D.P. Flanagan & E.M. McDonough (N-Y: Guilford, 2018). – P. 73-163.

Секция 4 ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ, БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ

УДК 004.912
doi: 10.15622/rcai.2025.021

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ АРГУМЕНТОВ НА ОСНОВЕ СИСТЕМАТИЗАЦИИ МОДЕЛЕЙ РАССУЖДЕНИЯ Д. УОЛТОНА[1]

И.Р. Ахмадеева (*i.r.akhmadeeva@iis.nsk.su*)
Ю.А. Загорулько (*zagor@iis.nsk.su*)
И.С. Кононенко (*irina_k@cn.ru*)
А.С. Серый (*alexey.seryj@iis.nsk.su*)
Е.А. Сидорова (*lsidorova@iis.nsk.su*)

Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск

Статья посвящена применению трансформерных моделей-энкодеров при разработке методов автоматической классификации аргументов. Представлена категоризация схем Д. Уолтона, включающая четыре классификатора, соответствующих различным уровням или аспектам аргументативной структуры. Были исследованы два подхода к решению задачи мультиклассовой классификации: (1) разработка классификатора для предсказания схемы аргумента и (2) разработка классификатора, предсказывающего схему и

[1] Работа выполнена при финансовой поддержке РНФ (проект № 23-11-00261, https://rscf.ru/ project/23-11-00261/).

категорию аргумента. Для получения векторных представлений использовалась модель ru-en-RoSBERTa. Эксперименты проводились на трех корпусах аннотированных аргументов: русскоязычном корпусе ArgNetSC и двух англоязычных корпусах Araucaria и NLAS (корпус автоматически сгенерированных аргументов). Наилучшие результаты на русскоязычном корпусе составили 41,3% F1-меры. Результаты по отдельным категориям (классам) аргументов – от 60% до 89%.

Ключевые слова: анализ аргументации, классификация аргументов, мультиклассовая классификация, категоризация схем Уолтона, трансформерная модель.

Введение

Теоретические и практические исследования аргументации опираются на данные о фактическом использовании аргументации в коммуникативной практике. В последнее время набирает обороты работа по получению таких данных, и прежде всего, – по созданию аннотированных корпусов реального аргументативного дискурса, что обусловлено во многом требованиями методов машинного обучения для автоматизированной обработки текста. Анализ данных, в свою очередь, базируется на типовых моделях аргументации – представленных преимущественно в виде схем и таксономий схем, специализированных для различных типов дискурса.

Таксономия схем аргументации Дугласа Уолтона [Walton et al., 2008, 2016] представляет собой эмпирически ориентированную классификацию схем, основанную на изучении очевидных конвенций аргументативной практики. На другом конце спектра находится Периодическая таблица аргументов Дж. Вагеманса [Wagemans, 2016]: она базируется на множественных априорных критериях, разработанных для исчерпывающего описания всех возможных комбинаций различных характеристик аргумента. Существуют и другие альтернативы, имеющие свои собственные преимущества и недостатки: [Feng et al., 2011], [Lawrence et al., 2016], [Musi et al., 2016], [Liga et al., 2020]. Авторы [Visser et al., 2021] аннотировали предвыборные дебаты с использованием схем Уолтона и таблицы Вагеманса. Исследование [Bezou-Vrakatseli et al., 2021] демонстрирует потенциальную совместимость двух подходов к классификации схем и выдвигает идею объединения их сильных сторон. Однако идея гибридной классификации до сих пор не получила широкого практического применения, более популярными остаются эмпирические модели, такие как модель Д. Уолтона.

Целью данной работы является разработка методов автоматической классификации аргументов, размеченных в соответствии с моделью Д.Уолтона в русскоязычных текстах научной коммуникации.

Для достижения поставленной цели в рамках данной работы были сформулированы следующие вопросы исследования.

RQ1. Какое качество мультиклассовой классификации аргументов в процессе анализа текстов научной коммуникации на русском языке можно получить с использованием нейросетевого подхода? Количество классов определяется количеством схем аргументов Уолтона, используемых при аннотировании наборов данных.

RQ2. Эффективно ли применение дополнительной систематизации при автоматической классификации аргументов? Под систематизацией в данном случае будет пониматься категоризация, объединяющая модели аргументов в группы по значимым признакам.

1. Обзор методов классификации аргументов

Задача классификации аргументов замыкает цепочку подзадач интеллектуального анализа аргументации (Argument Mining, AM) и может рассматриваться как задача мультиклассовой классификации аргументативных отношений (МКА) для заданных наборов посылок и тезисов. Трансформерные архитектуры языковых моделей с последующим дообучением (Supervised fine-tuning, SFT) на сегодняшний день наиболее эффективны для задач анализа текста. Однако этот подход требует репрезентативных наборов обучающих данных, что в случае МКА означает наличие представительного набора аннотированных аргументов для каждого класса из заданного классификатора.

Появление ресурсов, аннотированных схемами аргументации (таких как корпус Araucaria [Reed et al., 2008]) позволило сделать первые шаги по автоматическому выявлению схем аргументации с применением моделей классификации [Walton, 2011]. В дальнейшем исследования в значительной степени опирались на метод SVM и нейронные сети.

В [Feng et al., 2011] в рамках решения задачи реконструкции энтимем в качестве первого этапа проводится классификация аргументов по пяти наиболее частотным схемам Уолтона. При этом зафиксирована точность 0,63–0,91 в классификации «один против всех» и 0,80–0,94 в попарной классификации. При обучении использовался корпус Araucaria, в котором выбирались аргументативные сегменты, реализующие пять наиболее распространенных схем аргументации, и классификатор обучался как по признакам, специфичным для каждой отдельной схемы, так и по ряду общих лингвистических признаков.

В подходе [Lawrence, et al., 2015] идентифицируются отдельные компоненты схем, которые затем группируются в экземпляры схем. Здесь рассматриваются только две схемы (*From Expert Opinion* и *From Positive Consequences*), а классификаторы обучаются идентифицировать их от-

дельные компоненты, посылки и выводы. Учет признаков отдельных типов этих компонент дал F1-меру от 0,75 до 0,93 при идентификации хотя бы одной составляющей схемы.

Исследование [Stab et al., 2014], проведенное на материале 90 студенческих эссе, выделяет компоненты аргумента в многоклассовой классификации с помощью алгоритма SVM (классы: *Major Claim, Claim, Premise, None*) и признаков различных типов (структурные, лексические, синтаксические, индикаторные, контекстные). Получены F1 = 0,73 для выделения компонентов аргумента и 0,72 для обнаружения отношений в аргументе.

Исследование [Pimenov et al., 2024], проведенное на базе корпуса ArgNetSC, применяет классические методы машинного обучения – метод опорных векторов (SVM) и многослойный перцептрон (MLP) – для анализа 50 научных статей. Решается задача бинарной классификации для трех самых частотных типов аргументов: *From Part to Whole* (F1 = 0,64), *From Verbal Classification* (F1 = 0,68), *From Correlation to Cause* (F1 = 0,64).

Такие методы, как рекуррентные нейронные сети, сверточные нейронные сети и блоки долговременной и кратковременной памяти (LSTM) сыграли решающую роль во включении контекстных данных в процессы машинного обучения [Srivastava et al., 2022]. [Galassi et al., 2018] применяют LSTM для лучшего прогнозирования взаимосвязей компонентов в сложных структурах аргумента. Преодоление известных ограничений SVM и нейронных сетей, таких как необходимость разработки широкого набора признаков и трудности в захвате дальних зависимостей в тексте, связывают с появлением моделей на основе Transformer и особенно BERT: [Vaswani et al., 2017] и [Devlin et al., 2019].

2. Систематизация схем аргументов на основе модели Д. Уолтона

Д. Уолтон с коллегами неоднократно пересматривали первоначально предложенную таксономию схем. В [Walton et al., 2008] представлена система, состоящая из трех основных категорий. В последней версии системы [Walton et al., 2016] классическое различие между зависимыми от источника и независимыми от источника аргументами дает критерий для первой дихотомии. Зависимые от источника аргументы далее делятся на «эпистемические аргументы» и «практические аргументы», а первые подразделяются на те, которые заключаются в применении правил к случаям, и те, которые извлекают правила или сущности («аргументы открытия»). Авторы допускают, что одним из путей развития является ввод дополнительных классификаторов, которые позволят произвести “скрещивание” новой и существующей классификаций.

Отмечая сложности аннотирования на основе таксономии схем Уолтона, исследователи [Сидорова и др., 2024] обосновали необходимость дальнейшей систематизации схем аргументации и предложили их многоаспектную классификацию [Kononenko et al., 2023] (табл. 1). Выбор данного классификатора для проведения экспериментов связан с тем, что, во-первых, он основан на компендиуме Д.Уолтона и рассматривает достаточно большое количество схем аргументов, во-вторых, имеющиеся русскоязычные корпусы имеют согласованную с этим классификатором разметку, и, в-третьих, группы схем, выделяемых в многоаспектном классификаторе, зависят только от формы аргумента и не требуют никаких дополнительных данных.

Таблица 1

Типология схем аргументации

Основное отношение	Тип заключения	
	Практический	Теоретический
Гипер-Гипонимия	К действию	Устройство реальности
Элемент-множество	К цели	Установление реальности
Казуальность	К обязательству	
Коммуникация	Зависимость от источника аргумента	
Кондициональность	Внешний	Внутренний
Корреляция	От знания	От классификации
Меронимия	От цели	От объяснения
Аналогия	От ценности	От определения
Авторитетность	От человека	От факта
Способ	Направление атаки	
Противоречие	Нет атаки	На источник
	На тезис	На аргумент

Представленная категоризация позволяет настроить четыре разных классификатора и вычислить конечный тип аргумента на основании пересечения данных признаков.

3. Наборы данных

Для экспериментального исследования использовались три источника данных, основные характеристики которых представлены в табл. 2.

Таблица 2

Характеристики датасетов

Набор данных	Количество аргументов	Количество схем
Araucaria	730 (1746)*	17
NLAS	1893	20
ArgNetSC	9178	42

Выбор источников обусловлен общей для всех наборов данных разметкой текста с помощью схем аргументации Д. Уолтона. Из всех найденных наборов данных, с разметкой схемами Уолтона, в наше исследование не вошел только датасет EthiX [Bezou-Vrakatseli et al., 2024], поскольку утверждения, вошедшие в его состав, не составляют связного текста, а являются компиляцией аргументов, в то время как нашей задачей является классификация аргументов в тексте. Помимо этого, поддерживаемые или опровергаемые тезисы представлены не в явном виде, а в форме вопросов (*Would the world be a better place without humans?*) и в разметке не содержится указаний на то, какую из двух точек зрения доказывает каждый аргумент и, как следствие, невозможно осуществить промежуточную классификацию аргумента по признаку направление атаки.

Корпус Araucaria содержит англоязычные тексты из материалов газет и судебных дел; всего размечено 1746 аргументов, однако схема указана только для 730 из них. Ресурс характеризует высокая степень несбалансированности: для некоторых схем в корпусе имеется менее пяти примеров (*Inconsistent Commitment, Falsification of Hypothesis, Exceptional Case, Fear Appeal, Popular Practice*).

Англоязычная часть корпуса NLAS [Ruiz-Dolz et al., 2024] представляет 50 тематик, в которых автоматически сгенерированы 1893 аргумента с использованием 20 наиболее распространенных схем Уолтона; для NLAS характерна сбалансированность: имеется 75-100 примеров на каждую схему.

Корпус ArgNetSC [Ilina et al., 2021]) – русскоязычный корпус текстов, относящихся к области научной коммуникации. Тексты снабжены разметкой а) аргументов в соответствии с моделью Д.Уолтона – всего около 9 тыс. размеченных аргументов; б) классов аргументов. Корпус включает 286 текстов различных научных и научно-популярных жанров. Тексты имеют в среднем объем 3,5 тыс. токенов, средний объем комментария к ним – 2-5 предложений. На рис. 1 приведена статистика встречаемости аргументов каждого класса в корпусах.

Из рисунка видно, что классы остаются несбалансированными. При экспериментальных исследованиях классы с малым количеством встречаемости были исключены из рассмотрения, а аргументы, относящиеся к этим классам отнесены к другим классам в соответствии с их семантикой. Так, классы «*К цели*» и «*К действию*» не рассматривались, вместо этого использовался их родительский класс: «*Практический аргумент*», а схемы класса «*К обязательству*» перенесены в «*Установление реальности*».

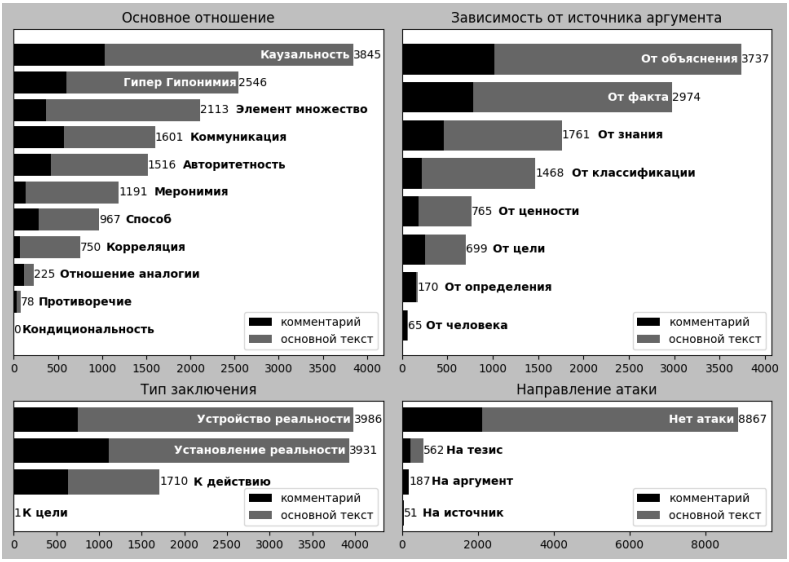


Рис. 1. Классы аргументов в корпусе ArgNetSC

4. Экспериментальное исследование

Датасеты с описанными выше свойствами позволяют провести следующие экспериментальные исследования:

а. Обучение классификации англоязычных текстов на сгенерированных аргументах (корпусе NLAS) и тестирование на “естественных” текстах корпуса Araucaria (на материале газет и судебных дел).

б. Обучение модели классификации аргументов и сравнение качества классификации с учетом и без учета дополнительной систематизации.

4.1. Архитектура эксперимента

В данной работе были исследованы два подхода к решению задачи мультиклассовой классификации аргументов: (1) классификация с использованием дообученной трансформерной модели и (2) классификация на основе систематизации схем Уолтона.

Для дообучения была использована модель ru-en-RoSBERTa [Snegirev et al., 2025], которая основана на модели ruRoBERTa [Zmitrovich et al., 2024] и является универсальной моделью для построения текстовых эмбеддингов для русского языка. Выбор данной модели также обусловлен тем, что по результатам предварительных экспериментов она продемонстрировала лучшее качество в задаче классификации схем аргументации по сравнению с другими русскоязычными и мультиязычными моделями, такими как BERTA, ruRoBERTa, mBERT.

В рамках данного подхода были предложены два способа интеграции информации о категориях схем аргументации в процесс классификации: многозадачное обучение и регуляризация эмбеддингов на основе категорий. Для их проверки были созданы классификаторы, реализующие один из методов или их комбинацию, которые сравнивались с базовой моделью на основе трансформерной архитектуры RoBERTa:

1. RoBERTa-sch – базовый многоклассовый классификатор по схемам аргументации.
2. RoBERTa-sch+cls-MT – модель, одновременно предсказывающая схему аргумента и его категории в рамках каждого из четырех классификаторов, обученная в многозадачном режиме.
3. RoBERTa-sch+cls-Reg – модификация модели (1), в которой векторные представления аргументов, принадлежащих одному классу, должны быть близки в векторном пространстве, что достигается с помощью контрастивной функции потерь.
4. RoBERTa-sch+cls-MT-Reg – модель, использующая обе предложенные стратегии.

Задача классификации аргументов в соответствии с систематизацией схем Уолтона проще детализированной классификации по конкретным схемам, поскольку количество категорий схем меньше, что обеспечивает более равномерное представление каждой категории в обучающей выборке. Однако задача усложняется тем, что один аргумент может одновременно принадлежать нескольким категориям, что требует решения проблемы классификации по нескольким меткам (multi-label). Для решения данной задачи была разработана архитектура из четырех классификаторов, обученных в режиме многозадачного обучения с использованием функции потерь на основе бинарной кросс-энтропии (Binary Cross Entropy), которая позволяет независимо оценивать принадлежность текста к каждой категории.

Для компенсации дисбаланса как в распределении схем, так и в распределении их категорий применялись специализированные функции потерь: для классификации схем – фокусная функция потерь (Focal Loss), уменьшающая влияние часто встречающихся и легко распознаваемых примеров, для классификации категорий – бинарная кросс-энтропия с весовой корректировкой, где каждой категории назначался вес, обратно пропорциональный частоте ее встречаемости.

4.2. Результаты экспериментов

Для оценки качества модели на корпусе ArgNetSC использовалась перекрестная проверка (cross-validation) с разбиением на 5 частей. Внутри каждой обучающей выборки дополнительно выделялась валидационная выборка (20%) для подбора гиперпараметров и применения ранней остановки. Результаты усреднялись по всем 5 частям.

Небольшой объем корпуса Araucaria не позволил использовать его для полноценного обучения моделей. Единицы примеров для некоторых аргументативных схем делает невозможным их корректную классификацию и обобщение моделью. В связи с этим была применена стратегия переноса обучения (transfer learning): в соответствующем эксперименте обучение осуществлялось на автоматически сгенерированном корпусе NLAS, а тестирование – на корпусе Araucaria. При этом на валидационной части NLAS получено значение F1 = 99%.

В табл. 3 представлены значения взвешенных метрик точности (P), полноты (R) и F1, агрегированные с учетом распределения классов.

Таблица 3

Результаты экспериментов по классификации аргументов

Model	**Araucaria**			**ArgNetSC**		
	P	**R**	**F1**	**P**	**R**	**F1**
RoBERTa-sch	38,3	18,22	21,63	42,21	**43,68**	**41,3**
RoBERTa-sch+cls MT	**42,37**	25,62	27,38	43,36	42,74	41,23
RoBERTa-sch+cls Reg	37,59	30,1	28,54	43,13	42,01	40,97
RoBERTa-sch+cls MT-Reg	38,73	**31,1**	**28,7**	**43,41**	41,7	40,63

Из таблицы видно, что наилучшие результаты для английского корпуса дает модель RoBERTa-sch+cls-MT-Reg, для которой получены лучшие полнота и F1. Для англоязычного корпуса точность выше, чем полнота, что, возможно, связано с тем, что в обучающей выборке отсутствовали примеры некоторых классов. В целом, результаты на русскоязычном корпусе лучше, что, по-видимому, объясняется тем, что модель для английского языка обучалась на синтезированных данных.

В табл. 4 представлены значения F1-меры для каждой категории многоаспектной классификации.

Таблица 4

Результаты классификации аргументов по дополнительным категориям

Корпус	Основное отношение	Тип заключения	Зависимость от источника аргумента	Направление атаки
Araucaria	17,8	49,6	25,93	86,76
ArgNetSC	**61,59**	**64,02**	**60,25**	**89,73**

Из таблицы видно, что результаты для русскоязычного корпуса по всем четырем классификаторам значительно лучше, чем для англоязычного.

4.3. Обсуждение результатов

В целом (RQ1) качество классификации не очень высокое. Наблюдается контраст по качеству между синтезированными текстами (99%) и естественными текстами, демонстрирующими значительно более низкие результаты. Одна из возможных причин – дисбаланс классов в русскоязычном корпусе, вторая причина связана с множественностью результата классификации (multi-label) для принятой систематизации.

Анализ и сравнение результатов работы моделей на разноязычных наборах данных показывают, что на текущий момент предлагаемый подход лучше работает для русского языка, что, по-видимому, связано с разным качеством датасетов.

В отношении влияния дополнительной систематизации схем аргументации (RQ2) на результаты классификации можно отметить, что модель демонстрирует положительные результаты для английского языка и в тоже время отсутствие значимых результатов для русского языка. В целом можно сделать вывод о полезности применения систематизации в условии недостаточности обучающих данных.

К причинам ошибок, не связанных с процедурой экспериментов, относятся факторы, которые в целом затрудняют решение задачи АМ:

- Естественной речи свойственно непрямое выражение мыслей, в особенности, в публицистических жанрах. Утверждения, реализующие посылки и заключение аргумента, иногда выражаются косвенным образом, и аннотаторы могут корректно восстановить смысл только в контексте целого графа.
- Ещё одно свойство естественной речи – большое количество энтимем (имплицитность, отсутствие посылки или заключения в явном виде). Значимость этого фактора подтверждается разницей результатов, полученных на синтетически сгенерированном корпусе NLAS и естественном ArgNetSC.
- Из-за диалогового характера взаимодействия в жанре комментариев к аналитическим статьям (часть корпуса ArgNetSC) элементы одного аргумента могут принадлежать разным участникам диалога и быть лексически и синтаксически неполными (эллипсис в комментариях) и неоднородными (неформальность языка комментариев).
- В текстах научной или другой сложной тематики в паре «посылка–заключение», помимо отношений, относящихся к аргументу, могут быть эксплицированы и другие семантические отношения.

Заключение

В работе представлено экспериментальное исследование процедуры автоматической классификации аргументов в соответствии с набором схем Д. Уолтона на основе нейросетевого подхода. Рассмотрены две до-

полнительные стратегии обучения моделей на основе систематизации схем аргументов: многозадачное обучение и регуляризация эмбеддингов на базе категорий.

Наилучшие результаты на русскоязычном корпусе достигли 41,3% F1-меры для мультиклассовой классификации по 42 классам. Использование многоаспектной классификации позволило улучшить качество классификации аргументов на англоязычных текстах более, чем на 7%, однако практически не повлияло на качество анализа русскоязычных текстов. В целом качество классификации по отдельным категориям достаточно хорошее – от 60% до 89%.

Дальнейшее развитие подхода может быть связано с развитием систематизации, балансировкой наборов данных и интеграцией с большими языковыми моделями.

Список литературы

[Сидорова и др., 2024] Сидорова Е.А., Кононенко И.С.. Онтологический анализ приемов аргументации в научном дискурсе // Информационные и математические технологии в науке и управлении. – 2024. – № 3(35). – С. 20-32. – DOI: 10.25729/ESI.2024.35.3.002.

[Bezou-Vrakatseli et al., 2021] Bezou-Vrakatseli E., Cocarascu O., & Modgil S. Towards an Argument Scheme Classification for Ethical Reasoning // In: CEUR Workshop Proceedings. – 2021. – 3205. – P. .13-17.

[Bezou-Vrakatseli et al., 2024] Bezou-Vrakatseli E, Cocarascu O, Modgil S. EthiX: A Dataset for Argument Scheme Classification in Ethical Debates // In 27th European Conference on Artificial Intelligence (ECAI). – 2024. – P. 3628-3635. – doi: 10.3233/FAIA240919.

[Devlin et al., 2019] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2019. – Vol. 1. – P. 4171-4186.

[Feng et al., 2011] Feng V.W., Hirst G. Classifying arguments by scheme // In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. – 2011. – P. 987-996.

[Galassi et al., 2018] Galassi A., Lippi. M., Torroni. P. Argumentative link prediction using residual networks and multi-objective learning // In: Proceedings of the 5th Workshop on ArgumentMining. – 2018. – P. 1-10.

[Ilina et al., 2021] Ilina Daria, Kononenko Irina, Sidorova Elena. On Developing a Web Resource to Study Argumentation in Popular Science Discourse // In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog-2021”. – 2021. – P. 318-327.

[Kononenko et al., 2023] Kononenko I.S., Sery A.S., Shestakov V.K., Sidorova E.A., Zagorulko Y.A. An Approach to Classifying Walton's Argumentation Schemes // In: Proc. 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, 2023. – P. 1540-1545. – doi: 10.1109/APEIE59731.2023.10347573.

[Lawrence, et al., 2015] Lawrence J., Reed C. Combining argument mining techniques // In Proceedings of the 2nd Workshop on Argumentation Mining. – 2015. – P. 127-136.

[Lawrence et al., 2016] Lawrence J., Reed C. Argument mining using argumentation scheme structures // In: Proceedings of Computational Models of Argument (COMMA). – 2016. – P. 379-390.

[Liga et al., 2020] Liga D., Palmirani M. Argumentation schemes as templates? Combining bottom-up and top-down knowledge representation // In: Proc. 20th CMNA workshop. – 2020. – P. 51-56.

[Musi et al., 2016] Musi E., Ghosh D., Muresan S. Towards feasible guidelines for the annotation of argument schemes // In: Proc. Third Workshop on Argument Mining, ArgMining@ACL, 2016. – P. 82-93.

[Pimenov et al., 2024] Pimenov I.S., Salomatina N.V. An Automatic Method for Standartizing Argumentative Annotations across Annotators Genres // 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM). – 2024. – P. 2260-2265. – DOI: 10.1109/EDM61683.2024.10615176.

[Reed et al., 2008] Reed C., Mochales Palau R., Rowe G., Moens M.F. Language resources for studying argument // In: Proc. 6th conference on language resources and evaluation (LREC 2008). – 2008. – P. 91-100.

[Ruiz et al., 2024] Ruiz-Dolz R., Taverner J., Lawrence J., Reed C. NLAS-multi: A multilingual corpus of automatically generated Natural Language Argumentation Schemes, Data in Brief. – 2024. – Vol. 57. – https://doi.org/10.1016/j.dib.2024.111087.

[Snegirev et al., 2025] Snegirev A., Tikhonova M., Maksimova A., Fenogenova A., Abramov A. The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) / ed. Chiruzzo L., Ritter A., Wang L. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. – P. 236-254.

[Srivastava et al., 2022] Srivastava P., Bhatnagar P., Goel A. ArgumentMining using BERT and Self-Attention based Embeddings // 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). – 2022. – https://doi.org/10.48550/arXiv.2302.13906.

[Stab et al., 2014] Stab C. and Gurevych I. Identifying argumentative discourse structures in persuasive essays // in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2014. – P. 46-56.

[Vaswani et al., 2017] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. Advances in neural information processing systems. – 2017. – 30.

[Visser et al., 2021] Visser J., Lawrence J., Reed C., Wagemans J., Walton D. Annotating argument schemes // Argumentation. – 2021. – 35. – P. 101-139. – doi: 10.1007/s10503-020-09519-x.

[Walton, 2011] Walton D. Argument mining by applying argumentation schemes // Studies in Logic. – 2011. – 4(1). – P. 38-64.

[Wagemans, 2016] Wagemans J.H.M. Constructing a Periodic Table of Arguments / Ed. Bondy P., Benacquista L. Argumentation, Objectivity, and Bias, Proc. of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA), Windsor, 2016. – P. 1-12.

[Walton et al., 2008] Walton D., Reed C., Macagno F. Argumentation schemes. – Cambridge, Cambridge University Press, 2008. – 456 p.

[Walton et al., 2016] Walton D., and Macagno F. A Classification System for Argumentation Schemes // Argument & Computation. – 2016. – P. 1-29.

[Zmitrovich et al.] Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov et al. A Family of Pretrained Transformer Language Models for Russian // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – 2024. – P. 507-524.

УДК 004.89
doi: 10.15622/rcai.2025.022

ИЗВЛЕЧЕНИЕ ИЗ ТЕКСТА ФРУСТРАЦИОННЫХ РЕАКЦИЙ С ПОМОЩЬЮ НЕЙРОСЕТЕВЫХ ПОДХОДОВ[1]

Д.А. Киреев (*kireev@isa.ru*)
Ю.М. Кузнецова (*kuzjum@yandex.ru*)
Н.В. Чудова (*nchudova@gmail.com*)
А.А. Чуганская (*anfi.chuganskaya@yandex.ru*)
И.В. Смирнов (*ivs@isa.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В статье представлены результаты применения нейросетевых моделей в задаче автоматического определения типов фрустрационных реакций в текстах сетевых дискуссий. Разработан корпус русскоязычных текстов с разметкой различных типов реакций на фрустрацию на основе типологии С. Розенцвейга. Рассмотрены два подхода к классификации текстов по типам реагирования: первый – последовательное определение наличия фрустрации, ее направления и типа реагирования, второй – одновременное определение всех типов реагирования. Эксперименты с нейросетевыми моделями на основе архитектуры «трансформер» и современными Большими Языковыми Моделями показали преимущество и эффективность второго подхода. Результаты демонстрируют, что используемые модели способны эффективно моделировать работу психодиагноста с речевыми проявлениями фрустрации.

Ключевые слова: сетевые дискуссии, реакция на фрустрацию, нейросетевые модели, трансформеры, большие языковые модели.

Введение

Последние годы отмечены повышенным интересом исследователей и практиков к возможностям использования методов машинного обучения в области поддержки психодиагностики. Значительная часть таких исследо-

[1] Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации, (проект № 075-15-2024-544).

ваний связана с определением предпочтений (политических, покупательских и др.) пользователей социальных сетей и основывается на анализе профилей пользователей и их поведения в сети. Такого рода работы опираются в теоретическом плане на представления о самопрезентации, группировании, инструментальных ценностях и других концептах социальной психологии и психологии личности. Но не менее значимой для социальной практики и интересной с точки зрения развития методов ИИ является область работ, связанная с анализом текстов сетевых дискуссий [Дымова и др., 2024], [Мамаев, 2024], [Отраднов и др., 2024], [Фокина и др., 2023], [Cau et al., 2025], [Mesquiti et al., 2025].

Фрустрация означает «расстройство (планов), уничтожение (замыслов), т.е. указывает на какую-то в известном смысле слова травмирующую ситуацию, при которой терпится неудача. Фрустрация должна рассматриваться в контексте более широкой проблемы выносливости по отношению к жизненным трудностям и реакций на эти трудности» [Левитов, 1967]. После десятилетий исследований и разноречивых трактовок этого понятия в современной психологии сформировалось представление о фрустрации как о психическом состоянии, вызванном неуспехом в удовлетворении потребности, желания [Мещеряков и др., 2003]. С. Розенцвейг предложил оценивать реакцию на препятствие по двум основаниям – направленность (на других, на себя, отрицание проблемы) и объект фиксации (на препятствии, на защите «я», на достижении цели); таким образом, любой ответ в ситуации фрустрации может быть отнесён к одному из девяти типов фрустрационного реагирования [Тарабрина, 1984].

Реакции на фрустрацию, согласно С. Розенцвейгу, определяются следующим образом:

- экстрапунитивные реакции (E) – тревожно-обвинительные реакции, представленные реакцией с фиксацией на препятствии («Какой ужас! Какое безобразие!»), реакцией с фиксацией на виновнике («Ему плевать на людей!») и реакцией с фиксацией на получении желаемого («Так сделайте это!»);
- интропунитивные реакции (I) – самообвинительные реакции, представленные реакцией с фиксацией на препятствии («Это даже хорошо, что у меня с первого раза не получилось»), реакцией с фиксацией на виновнике («Извините, не заметила») и реакцией с фиксацией на получении желаемого («Попробую поискать правильного врача»);
- импунитивные реакции (M) – реакции с отрицанием проблемы, представленные реакцией с фиксацией на препятствии («Ничего страшного»), реакцией с фиксацией на виновнике («С каждым может случиться») и реакцией с фиксацией на получении желаемого («Как-то договоримся, а может и вообще рассосётся»).

Проблема автоматического определения фрустрации у автора текста не является новой и уже была исследована ранее. В работе [Suri et al., 2020] представлен подход к определению фрустрации в пользовательских отзывах с использованием методов машинного обучения. Авторы исследуют фрустрацию как эмоцию, возникающую из разочарования или неудовлетворенности продуктом или услугой. Исследование фокусируется на анализе негативных отзывов в онлайн коммерции, где пользователи выражают глубокие чувства разочарования приобретенными товарами. Работа демонстрирует применимость алгоритмов машинного обучения для автоматического выявления фрустрации в текстовых данных отзывов покупателей. Исследование [Chhaya et al., 2018] предлагает подход к выявлению фрустрации из переписок по электронной почте с помощью количественной оценки чувств и тональности. Авторы идентифицируют лингвистические особенности, влияющие на человеческое восприятие фрустрации, и моделируют ее как задачу обучения с учителем. Работа представляет детальное сравнение между традиционными регрессионными и основанными на распределении слов моделями. В работе [Ļeonova et al., 2022] представлен сравнительный анализ предсказания интенсивности фрустрации в постах социальных сетей на разных языках с использованием нейросетевых моделей, комбинирующих лексические и нелексические способы выражения. Авторы тестировали различные конфигурации моделей на текстах диалогов поддержки клиентов на латышском и английском языках. Исследование демонстрирует, что модели с конфигурациями, использующими все доступные признаки на основе нелексических средств выражения, дают наилучшую точность.

Предыдущие исследования преимущественно опираются на лингвистические признаки и ориентированы на обработку англоязычных текстов. В настоящей работе мы решаем задачу определения типов фрустрационного реагирования в текстах на русском языке с применением различных предобученных нейросетевых моделей. Мы полагаем, что предобучение на больших корпусах русскоязычных текстов позволяет сформировать у нейросетевой модели общую речевую компетентность на уровне языковой системности, а дообучение на размеченном экспертами-психологами корпусе текстов сетевых дискуссий – профессиональную речевую компетентность на уровне речевой системности [Девяткин и др., 2023]. В наших прошлых исследованиях в качестве модельного примера работы психодиагноста с текстом была взята модифицированная нами полупроективная методика изучения фрустрационного реагирования (тест Розенцвейга) [Девяткин и др., 2021], [Devyatkin et al., 2021]. Для настоящего исследования был собран новый корпус русскоязычных реплик сетевых дискуссий, размеченных в соответствии с типологией С. Розенцвейга. Корпус использовался для автоматического определения типов фрустрационных реакций с использованием нейросетевых подходов и Больших Языковых Моделей, которые, насколько мы знаем, ещё не применялись для решения такой задачи.

1. Данные

Исходный корпус комментариев, содержащий около 8300 реплик, был собран из дискуссий в социальной сети ВКонтакте. Для создания обучающего корпуса было отобрано 6 тыс. реплик, которые были размечены тремя экспертами-психологами. Сначала был создан фоновый корпус (3202 реплики), содержащий те высказывания, в которых эксперт не видит явных признаков фрустрационного реагирования. Такие высказывания далее обозначаются классом «р». Далее, в обучающий корпус (2527 реплик) добавлялись только те высказывания, которые считаются несомненными проявлениями одного из девяти типов фрустрационного реагирования по мнению всех трёх экспертов. Наконец, был создан проверочный корпус (2275 реплики), содержащий оставшиеся реплики из исходного корпуса. В этом корпусе встречаются все типы высказываний – фоновые реплики, однозначные реплики-реакции (как в обучающем корпусе), неоднозначные реплики.

Получившиеся выборки оказались несбалансированными. Это вызвано тем, что экстрапунитивные реакции (испуг/возмущение, обвинение, требование) широко представлены в сетевом общении. Реплики с импунитивными реакциями (попытка успокоить, попытка примирить, выражение надежды на легкое разрешение проблемы) встречаются несколько реже и как правило в дискуссии противопоставляются репликам с экстрапунитивными реакциями. Интропунитивные реакции в целом нехарактерны для сетевого дискурса (как и вообще для публичной коммуникации), однако выражение готовности самостоятельно разрешить проблему встречаются всё же не так редко, как извинения или, тем более, размышления о пользе собственных промахов и неудач. Попытки увеличить количество примеров интропунитивных реакций (i, I, I’) путем генерации похожих реплик с помощью Больших Языковых Моделей не привели к успеху, т.к. сгенерированные реплики оказались неестественными и были отвергнуты экспертами-психологами.

Перед обучением имена собственные в репликах были удалены с помощью модуля NER (Named Entity Recognition) из библиотеки spacy [Honnibal et al., 2020]. Реплики из фонового и обучающего корпусов использовались при обучении моделей, а реплики из проверочного корпуса использовались при оценке результатов работы моделей. Реплики из фонового и обучающего корпусов были разделены на тренировочную и валидационную выборки в отношении 9 к 1 соответственно, со стратификацией по классу, то есть сохраняя соотношение классов, что было важно из-за их несбалансированности. В табл. 1 представлена информация о количестве классов в каждой из выборок.

Таблица 1

класс\выборка	тренировочный	тестовый	валидационный
p	2753	832	306
e	319	161	35
E	901	358	100
E'	671	317	74
i	289	27	32
I	115	13	13
I'	14	5	2
m	110	40	12
M	150	39	17
M'	248	68	28
всего	**5570**	**1860**	**619**

Получившиеся выборки опубликованы в открытом доступе на платформе HuggingFace[1].

2. Методы

Задача выявления типов реакции на фрустрацию решается как задача классификации текста. Хотя тест Розенцвейга предполагает параллельное выделение направленности и типа, предварительные эксперименты и экспертиза психологов-разметчиков показали, что определение типа реакции сразу по всем направленностям менее эффективно, чем определение типа для каждой направленности. Поэтому предлагается решать задачу в несколько шагов:

1. Определение наличия\отсутствия фрустрации
2. Определение направленности реакции на фрустрацию, если на Шаге 1 выявлена фрустрация
3. Определение типа реакции на фрустрацию в соответствии с объектами фиксации (мультиклассовая классификация для ранее предсказанного направления):
   a. Определение типа экстрапунитивных (E) реакций
   b. Определение типа интрапунитивных (I) реакций
   c. Определение типа импунитивных (M) реакций

Далее данный подход будет называться «Пошагово». Он включает 3 шага с 5-ю разными классификаторами. Такой подход имитирует работу эксперта при разметке.

Задачу извлечения типов реакции на фрустрацию можно решать и как одну задачу классификации. Такой подход подразумевает создание одного классификатора, который обучается на извлечение всех 10 классов (1 класс наличие\отсутствие фрустрации и 9 классов для типов реакции). Он имитирует процесс профессионального обучения специалиста, когда

[1] https://huggingface.co/datasets/isa-ras/frustration_dataset.

представление о фрустрации и девяти вариантах реагирования на неё даётся будущему специалисту сразу, в комплексе. Далее данный подход будет называться «Всё сразу».

Отметим, что в типологии Дж. Брунера указанные выше подходы можно описать как сканирующую (для подхода «Пошагово») и фокусирующую (для подхода «Всё сразу») стратегии приёма информации при образовании понятий [Брунер, 1977].

Данные подходы использовались при обучении нейросетевых моделей с архитектурой трансформер, предобученные на русском языке. Лучший подход так же используется при работе с Большими Языковыми Моделями.

3. Эксперименты

Для оценки качества решения задачи использовалась взвешенный показатель F1 (где весом является количество истинных примеров каждого класса), т. к. он позволяет учитывать несбалансированность выборок.

3.1. Сравнение подходов «Пошагово» и «Всё сразу» с использованием трансформеров.

Для сравнения подходов «Пошагово» и «Всё сразу» были использованы нейросетевые модели с архитектурой трансформера предобученные на русском языке: ruBert-base, ruBert-large, ruElectra-small ruElectra-medium, ruElectra-large, ruRoberta-large, представленные в работе [Zmitrovich et al., 2023]. Подход «Пошагово» состоит из 5 задач, каждая из которых рассматривались отдельно. Для всех моделей были подобраны гиперпараметры на 100 итерациях, где модели обучались на тренировочной и оценивались на валидационной выборках. В табл. 2 представлены результаты работы моделей с подобранными гиперпараметрами, лучшие показатели для каждой задачи выделены жирным.

Таблица 2

Задача \Модель	ruBert-base	ruBert-large	ruElectra-small	ruElectra-medium	ruElectra-large	ruRoberta-large
Наличие фрустрации	80.27	**81.07**	78.18	78.46	79.02	80.45
Направление фрустрации	86.28	88.28	77.19	81.62	86.81	**89.78**
Тип экстрапунитивных (E) реакций	81.35	82.27	68.36	76.99	78.51	**82.66**
Тип интрапунитивных (I) реакций	92.51	**96.78**	90	92.33	92.51	93.36
Тип импунитивных (M) реакций	73.44	**84.15**	65.37	72.29	82.04	80.12
Подход «Всё сразу»	67.17	66.85	53.01	61.84	66.59	**71.30**

Лучшие модели в обоих подходах оценивались на тестовой выборке. В табл. 3 представлены оценки качества работы обоих подходов.

Таблица 3

Класс \ Подход	«Пошагово»	«Всё сразу»
p	79.75	80.89
e	56.43	60.63
E	67.51	71.66
E'	57.89	62.78
i	40.00	51.61
I	18.18	42.11
I'	0.00	0.00
m	25.35	31.58
M	24.24	17.50
M'	36.50	34.43
Взвешенное среднее	**66.51**	**69.27**

3.2. Эксперименты с Большими Языковыми Моделями

Большие языковые модели применялись в подходе «Всё сразу», так как в разделе 3.1 он показал результаты лучше. Эксперименты были построены с помощью подхода «обучение-в-контексте» [Brown et al., 2020]: на вход модели подавался фрагмент диалога, где первое сообщение было системной инструкцией с описанием классов фрустрации и информацией о том, что нужно отвечать только классом фрустрации, а если его нет – классом p. Далее следовал диалог между пользователем и моделью, где пользователь отправляет реплику и модель отвечает правильным классом. Реплики и ответы были взяты из тренировочной выборки. Потом пользователь отправлял реплику из тестовой выборки и ожидал ответ модели. Такой диалог позволял дать модели описание необходимой информации, примеры и ожидаемых результатов. Всего рассматривалось 6 моделей: GPT-4.1-nano, GPT-4.1-mini и GPT-4.1 от OpenAI, Gemini-2.0-flash от Google и DeepSeek-v3, и LLaMa-3.3-70b-instruct из открытого доступа. Стоит заметить, что некоторые модели могли не обработать запрос из-за нарушения их политики использования и такие случаи рассматривались как неправильная разметка. В табл. 4 представлены результаты экспериментов.

Таблица 4

Модель \ Метрика	GPT-4.1-nano	GPT-4.1-mini	GPT-4.1	DeepSeek-v3	Gemini-2.0-flash	LLaMa-3.3-70b-instruct	Claude-3.7
p	76.72	75.71	78.77	70.50	69.84	69.36	65.38
e	18.89	52.12	60.92	46.28	50.50	40.96	55.81
E	20.05	70.67	73.92	65.28	71.93	67.10	67.84
E'	32.43	59.58	64.05	57.24	57.33	53.40	54.38
i	7.50	15.62	36.11	21.69	22.56	23.88	34.57
I	-	12.50	18.18	14.29	21.05	14.81	27.27
I'	0.00	5.97	9.09	11.32	9.84	2.72	6.32
m	25.49	30.93	36.73	30.46	34.41	39.08	29.14
M	4.08	19.23	23.40	15.38	29.70	16.90	32.00
M'	9.76	19.05	34.71	29.20	47.24	34.02	37.37
Взвешенное среднее	**46.27**	**64.23**	**68.88**	**60.35**	**62.82**	**59.48**	**59.77**

3.3. Обсуждение результатов

Как можно заметить из табл. 3 и 4, оценка качества извлечения класса I или I’ показывает 0 или -, что вызвано маленьким количеством примеров данных классов (всего 13 и 5 примеров в тестовой выборке соответственно).

Из результатов экспериментов в разделе 3.1 видно, что подход «Всё сразу» показывает результаты лучше, чем подход «Пошагово» на 3 процента. Это может быть объяснено тем, что подход «Всё сразу» обучается разнице между всеми классами, что может быть проще чем классификация узкого типа фрустраций. К тому же ошибки в подходе «Пошагово» делают результаты последующих шагов также ошибочными, то есть если выделен неправильный тип, то направление точно будет неправильным, и каждый шаг подхода уменьшает набор данных, доступных для обучения моделей, из-за чего качество работы моделей ухудшается. Однако стоит заметить, что подход «Пошагово» позволяет интерпретировать вероятностные выходы каждой модели как оценку уверенности на каждом шаге, что можно использовать, например, для раннего выхода: если уверенность определения типа реакции меньше порога, то вместо продолжения работы подхода, можно сразу вернуть класс «фрустрации», что будет означать наличие фрустрации, но невозможность определения типа и направления реакции.

Анализ матрицы ошибок этих подходов показал, что модели чаще всего делают ошибки при классификации высказываний без явных признаков фрустрационного реагирования (р). Это объясняется сложностью и неоднозначностью самого психологического понятия фрустрационного реагирования.

Из результатов экспериментов с Большими языковыми моделями в разделе 3.2 видно, что лучший результат показала модель GPT-4.1, который всего на 1 процент хуже работы метода «Всё сразу» с трансформерами, хотя она и не была дообучена и работала на значительно меньшем количестве данных: трансформерная модель была обучена на 5570 примерах фрустрации, в то время как для работы Большой языковой модели были представлены только 50 примеров реплик. Более того, можно увидеть разницу в работе моделей семейства GPT 4.1: nano модель показала результаты хуже mini, которая показала результаты хуже, чем полная модель, что соответствует разнице в размере этих моделей: чем больше модель, тем она лучше работает. Однако, рассматриваемые модели не дообучались на исследуемой задаче и могли не обрабатывать некоторые реплики из-за своей политики использования, поэтому их потенциал был раскрыт не полностью.

Заключение

Полученные результаты показали принципиальную возможность моделировать с помощью нейросетевого подхода работу психодиагноста с таким неоднозначным по своей природе материалом как речевые проявления фрустрационного реагирования. Достигнутое качество автоматического распознавания в тексте сетевых дискуссий различных типов фрустрационного реагирования соответствует уровню точности самих используемых психологических понятий. Эта особенность психодиагностического материала, с которым имеют дело специалисты по машинному обучению, проявляется, в частности, в том, что обучающий корпус, в котором были собраны только однозначно трактуемые всеми тремя экспертами реплики, составил около 90% от всех реплик, содержащих реакцию на фрустрацию.

Наше исследование подтвердило фундаментальную закономерность, описанную в когнитивной психологии: при формировании обобщений, более выигрышной оказывается стратегия выдвижения целостной гипотезы, чем стратегия последовательной проверки парциальных гипотез. В нашем случае это имеет понятное объяснение – ошибки моделей накапливаются, делая последующие результаты неверными. Более того, каждый шаг уменьшает набор данных, доступный для обучения моделей, из-за чего качество работы моделей ухудшается.

Благодарности. Работа выполнялась с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

Список литературы

[Брунер, 1977] Брунер Д. Психология познания. За пределами непосредственной информации: пер. с англ. – М.: Прогресс, 1977. – 413 с.

[Девяткин и др., 2021] Девяткин Д.А., Ениколопов С.Н., Салимовский В.А., Чудова Н.В. Речевые реакции на фрустрацию: автоматическая категоризация // Психологические исследования. – 2021. – Т. 14, № 78. – С. 1. – doi: 10.54359/ps.v14i78.160.

[Девяткин и др., 2023] Девяткин Д.А., Салимовский В.А., Чудова Н.В. Об эвристическом потенциале категории «стилистико-речевая системность» // Коммуникативная стилистика текста: итоги и перспективы: материалы Всероссийского научного семинара (Томск, 20 января 2023 г.) / под общ. ред. С.М. Карпенко; Томский государственный педагогический университет. – Томск: Изд-во ТГПУ, 2023. – С. 20-27. – ISBN 978-5-89428-987-8.

[Дымова и др., 2024] Дымова П.И., Домбровская А.Ю. Измерение социального самочувствия горожан по цифровым маркерам: апробация методики // Социальные новации и социальные науки. – 2024. – № 3(16). – С. 94-107. – doi: 10.31249/snsn/2024.03.07.

[Левитов, 1967] Левитов Н.Д. Фрустрация как один из видов психических состояний // Вопросы психологии. – 1967. – Т. 6. – С. 118-129.

[Мамаев, 2024] Мамаев И.Д. Кластерный анализ лингвистических профилей скрытых сообществ // Филологические науки. Вопросы теории и практики. – 2024. – Т. 17, № 5. – С. 1739-1747. – doi: 10.30853/phil20240250.

[Мещеряков и др., 2003] Мещеряков Б.Г., Зинченко В.П. Большой психологический словарь. – М.: Прайм-Еврознак, 2003. – 525 с.

[Отраднов и др, 2024] Отраднов К.К., Калинин В.Н., Лесько С.А., Платонова И.В. Организация сбора и обработки данных социодинамических процессов с возможной самоорганизацией и наличием памяти и анализ наблюдаемых характеристик их временных рядов // International Journal of Open Information Technologies. – 2024. – Т. 12, № 4. – С. 4-14.

[Тарабрина, 1984] Тарабрина П.В. Экспериментально-психологическая методика изучения фрустрационных реакций: Методические рекомендации. – 1984. – № 5. – С. 34-37.

[Фокина и др, 2023] Фокина А.И., Чеповский А.А., Чеповский А.М. Использование платформы ТХМ корпусного анализа для анализа текстов сообществ социальных сетей // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2023. – Т. 21, №. 2. – С. 29-38. – doi: 10.25205/1818-7900-2023-21-2-29-38.

[Brown et al., 2020] Brown T. et al. Language models are few-shot learners // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 1877-1901. – doi: 10.48550/arXiv.2005.14165.

[Cau et al., 2025] Cau E., Pansanella V., & Pedreschi D., & Rossetti G. Language-Driven Opinion Dynamics in Agent-Based Simulations with LLMs. – 2025. – doi: 10.48550/arXiv.2502.19098.

[Chhaya et al., 2018] Chhaya N., et al. Frustrated, polite, or formal: Quantifying feelings and tone in email // Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. – 2018. – P. 76-86.

[Devyatkin et al., 2021] Devyatkin D., Chudova N., Chuganskaya A., Sharypina D. Methods for Recognition of Frustration-Derived Reactions on Social Media / In: Kovalev S.M., Kuznetsov S.O., Panov A.I. (eds) // Artificial Intelligence. RCAI 2021. Lecture Notes in Computer Science. – Vol 12948. – Springer, Cham. – P. 17-30. – doi: 10.1007/978-3-030-86855-0_2.

[Honnibal et al., 2020] Honnibal M., Montani I., Van Landeghem S., Boyd A. spaCy: Industrial-strength natural language processing in python. – 2020. – doi:10.5281/zenodo.1212303.

[Ļeonova et al., 2022] Ļeonova V., Zuters J. Frustration Level Analysis in Customer Support Tweets for Different Languages. – 2022.

[Mesquiti et al., 2025] Mesquiti S., et al. Analysis of social media language reveals the psychological interaction of three successive upheavals // Scientific Reports. – 2025. – Vol 15(1). – P. 5740. – doi: 10.1038/s41598-025-89165-z.

[Suri et al., 2020] Suri S., Sharma K., Papneja S. Frustration detection on reviews using machine learning // 2020 International Conference for Emerging Technology (INCET). – IEEE, 2020. – P. 1-5.

[Zmitrovich et al., 2023] Zmitrovich D., et al. A family of pretrained transformer language models for Russian // arXiv preprint arXiv:2309.10931. – 2023. – doi: 10.48550/arXiv.2309.10931.

УДК 004.853
doi: 10.15622/rcai.2025.023

СРАВНЕНИЕ РУССКОЯЗЫЧНЫХ ТЕКСТОВ С ПОМОЩЬЮ ГРАФА СЛОВОСОЧЕТАНИЙ ДЛЯ ВЫЯВЛЕНИЯ ОТЛИЧИТЕЛЬНЫХ СМЫСЛОВЫХ ФРАГМЕНТОВ[1]

Н.В. Мелещенко (*meleshenko.nikolay@mail.ru*)
О.И. Федяев (*olegfedyayev@mail.ru*)

Донецкий национальный технический университет, Донецк

В работе рассматривается проблема обновления университетских учебных программ с учётом требований (рекомендаций) предприятий. Предложен подход, основанный на представлении текста в виде графа словосочетаний, который позволяет визуализировать связи между терминами и улучшить процесс анализа. Отличием от предыдущих работ является использование дерева составляющих вместо дерева зависимостей для формализации извлечения словосочетаний. Решены проблемы нормализации извлеченных терминов и обработки конъюнкций в тексте, что способствует более точному определению словосочетаний. Проведенные экспериментальные исследования подтверждают эффективность предложенного подхода.

Ключевые слова: естественный язык, сравнение текстов, требования предприятий, учебные программы дисциплин, граф словосочетаний, нормализация терминов, смысловые фрагменты.

Введение

Представленная работа затрагивает проблему инновации университетских учебных программ дисциплин путём учёта новых требований со стороны предприятий. Регулярное обновление учебных программ дисциплин обеспечивает высокий уровень профессиональной подготовки выпускников, востребованных на рынке труда. В связи с этим возникает необходимость в регулярной и профессионально-ориентированной кооперации кафедры и профильных предприятий для решения данной проблемы. Взаимодействие выпускающей кафедры с предприятиями осуществляется на

[1] Данная работа выполняется по плану Молодёжной научной лаборатории «Искусственный интеллект» ДонНТУ. Научная работа № FRRF-2024-0010.

уровне смыслового анализа текстовых документов (рекомендации предприятий, рабочие программы дисциплин) и извлечения новых компетенций (технологий, методов, инструментов и др.) для корректировки соответствующих рабочих программ дисциплин.

Появление новых методов обработки текстовой информации и соответствующих инструментов сделало возможным автоматизировать решение задачи по интеллектуальной поддержке формирования и обновления образовательных программ, в том числе и рабочих программ дисциплин (РПД). В одной из первых отечественных работ этого направления [Космачёва и др., 2016] была предложена интерактивная система формирования РПД, которая использовала простые трансформационные методы обработки информации и ограничивалась проверкой РПД на соответствие критериям качества ФГОС. Формальные методы описания текстового документа при помощи графа были рассмотрены в работе [Sheetal et. al., 2014]. В работе [Ботов, 2019] использовались современные нейросетевые модели языка word2vec, но они применялись только для оценки семантической близости анализируемых документов. Извлечению с помощью нейросетевой модели BERT коротких фрагментов знаний и навыков из текстов требований онлайн-вакансий посвящена интересная работа [Николаев, 2023]. Однако в ней не решён вопрос насколько эти знания будут новыми по отношению к РПД. В статье [Wu et. al., 2024] авторы предлагают семантику текста представлять с помощью графа знаний из фраз, полученных при помощи вероятностной контекстно-свободной грамматики и алгоритма CKY. Однако данный подход был протестирован авторами только в задачах классификации и кластеризации текстов.

В предшествующей работе авторов [Федяев и др., 2025] рассмотрен один из подходов к решению задачи инновации РПД. Процесс решения заключался в сравнении семантик текстов требований предприятия и рабочих программ дисциплин с целью получения разницы в знаниях, представленных в этих документах. Идея решения основывается на представлении текста в виде графа – семантической сети, отражающей синтаксические связи между словами в тексте. Разница в знаниях рассматривается как разность графов двух документов, при этом учитываются не только отдельные слова, но и словосочетания. В используемом подходе было выявлено несколько не решённых проблемных вопросов:

- процесс извлечения новых терминов из текста не был формализован, а использование дерева зависимостей накладывает неопределённость из-за своей непостоянной структуры вследствие ошибок частеречной и морфологической разметок [Демидов, 2023];
- выделение и преобразование смысловых фрагментов осуществлялось только в тексте требований предприятий, но лучшим решением будет представление текста требований и РПД в одном формате для сравнения их при помощи вычисления разности графов;

- отсутствие нормализации извлеченных терминов приводит к ухудшению смысловой оценки текста.

Поэтому целью данной работы является представление текстов рабочих программ дисциплин и требований предприятий в новой форме – в виде графов словосочетаний, позволяющих повысить формализацию выявления и сравнения представленных в них знаний.

1. Особенности представления текста в виде дерева составляющих

Текст может быть представлен при помощи синтаксического анализа в двух видах: дерева зависимостей и дерева составляющих [Кравченко и др., 2024]. В работе [Федяев и др., 2025] текст изначально представлялся в виде дерева зависимостей, но в данной работе мы предлагаем использовать дерево составляющих, которое можно получить при помощи набора жадных (greedy) регулярных выражений (см. листинг 1), являющихся подобием контекстно-свободной грамматики.

Дерево составляющих основано на формализме контекстно-свободных грамматик и может быть построено автоматически, используя программные средства (язык программирования Python, библиотеки SpaCy и NLTK) синтаксического анализа на основе регулярных выражений. В таком дереве предложение делится на составляющие, т.е. фразы, которые относятся к определенной категории в грамматике.

Листинг 1

```
AP: {<A>+}
ACC: {<CC|COMMA><AP>}
APCC: {<AP><ACC>+}
NPG: {<NG><NG|FN>*}
NP: {<N><NPG|FN>*}
ANP: {<AP|APCC><NP>}
ANPG: {<AP|APCC><NPG>}
TERM: {<ANP|NP><ANPG>*}
TERMCC: {<CC|COMMA><TERM>}
TERMG: {<ANPG|NPG><ANPG>*}
TERMGCC: {<CC|COMMA><TERMG>}
COMTERM: {<TERM><TERMCC>+}
COMTERMEG: {<COMTERM><TERMGCC>+}
COMTERMG: {<TERM><TERMGCC>+}
FNP: {<FN>+}
AFNP: {<AP|APCC><FNP>}
FTERM: {<AFNP|FNP>}
FTERMCC: {<CC|COMMA><FTERM>}
```

По своей сути грамматика позволяет строить правильные предложения и извлекать их синтаксическую структуру [Кравченко и др., 2024], [Полетаев и др., 2023].

Подводя итог, можно отметить, что дерево составляющих, в отличие от дерева зависимостей, содержит синтаксическое представление предложения в соответствии с заданной контекстно-свободной грамматикой, что позволяет формализовать извлечение словосочетаний из текста. Такое представление имеет чёткую иерархию и делит предложения на отдельные фразовые составляющие [Wu et. al., 2024].

Приведенные выше регулярные выражения содержат правила, включающие разные синтаксические категории (табл. 1).

Таблица 1

Категория	Описание	Пример
CC	Союз	и, или
COMMA	Запятая	,
A	Прилагательное	синий
FN	Иностранное существительное	data, text
N	Существительное	метод
NG	Существительное в родительном падеже	метода
AP	Последовательность прилагательных	большой синий
ACC	Конъюнкция прилагательного	и большой синий
APCC	Конъюнкция набора прилагательных	черный средний и большой синий
NPG	Именная группа в родительном падеже	анализа данных
NP	Именная группа	метод анализа данных
ANP	Группа прилагательного	эффективный метод анализа данных
ANPG	Группа прилагательного в родительном падеже	больших данных
TERM	Термин	основные стратегии обучения нейронных сетей
TERMCC	Конъюнкция термина	и бинарное дерево
TERMG	Термин в родительном падеже	структур нейронных сетей
TERMGCC	Конъюнкция термина в родительном падеже	и структур данных
COMTERM	Конъюнкция терминов	обработка и анализ больших данных
COMTERMEG	Конъюнкция терминов и конъюнкция терминов в родительном падеже	разработка и тестирование прикладного программного обеспечения и системных решений
COMTERMG	Термин и конъюнкция терминов в родительном падеже	разработка моделей, алгоритмов и программ
FNP	Иностранная именная группа	text mining

Дерево составляющих используется нами для выделения основных смысловых конструкций – терминов. Под словом «термин» мы условились понимать такое максимально длинное словосочетание, удовлетворяющее регулярному выражению, которое начинается с существительного в именительном падеже и заканчивается, по возможности, существительным в родительном падеже, включая прилагательные, описывающие каждое существительное в термине. Например, дерево составляющих для предложения «Основные стратегии обучения нейронных сетей» представлено на рис. 1.

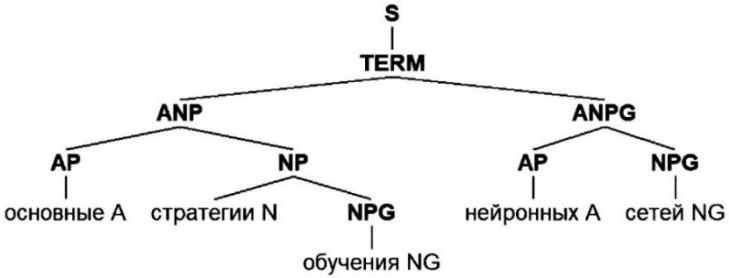


Рис. 1. Дерево составляющих предложения

Как видно из рисунка данное предложение содержит одну категорию TERM (термин). Термины рассматриваются нами как основные лексические единицы текста, поэтому текст можно представить как совокупность терминов [Wu et. al., 2024]. Эту совокупность можно в дальнейшем анализировать с целью выявления ключевых фраз и слов, а также осуществлять их поиск, если представить текст как граф терминов и их составляющих, что мы и используем в этой работе.

2. Преобразование дерева составляющих в граф словосочетаний

Главным недостатком, как дерева составляющих, так и дерева зависимостей является то, что они строятся для одного предложения, а не для всего текста [Wu et. al., 2024]. Поэтому необходимо определить структуру данных и алгоритм преобразования множества деревьев составляющих в общую структуру в качестве представления всего текста [Григорьева и др., 2023].

Представим текст в виде орграфа *G(V, E)*, где:

V – множество вершин, представляющих собой слова и образуемые из них словосочетания – термины, которые встречаются в исходном тексте;

E – множество ребер, которые указывают на формирование фразы $\mathrm{E} = \{\langle a, pos, b\rangle | a, b \in \mathrm{V}\}$, где pos – позиция части словосочетания a в словосочетании b. Графическое представление данного графа представлено на рис. 2.

Таким образом, данную структуру данных можно назвать графом словосочетаний, который представляет весь текст в виде набора терминов и словосочетаний, которые его формируют.

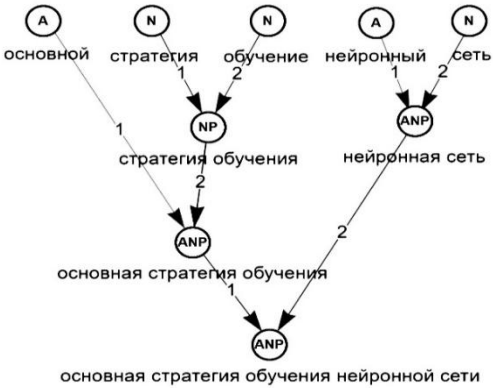


Рис. 2. Графическое представление графа словосочетаний

На рисунке видны существительные (стратегия, обучение и сеть), прилагательные (основной, нейронный), именная группа (стратегия обучения) и группы прилагательных (нейронная сеть, основная стратегия обучения, основная стратегия обучения нейронной сети). Данное представление текста имеет следующие преимущества:

1) возможность наглядной визуализации и простота интерпретации графа человеком;

2) данная структура данных позволяет применять различные алгоритмы на графах для анализа и обработки данных текста;

3) представление текста в виде графа даёт возможность при помощи операций над графами (разность, пересечение, объединение) определять различия в словосочетаниях, общие используемые термины и составить общий словарь словосочетаний для двух текстов [Sheetal et. al., 2014].

При построении графа для каждой вершины учитывается количество включений данного слова или словосочетания в другие словосочетания в тексте [Кравченко и др., 2024]. После данной операции мы можем для каждого ребра рассчитать частоту расширений словосочетаний, т.е., например, как часто слово «сеть» расширяется до «нейронная сеть» [Григорьева и др., 2023]. Таким образом, появляется возможность дальнейшего преобразования полученного графа путём выделения наиболее частых выражений в отдельные вершины. Программная модель графа реализована при помощи библиотеки NetworkX и визуализирована с помощью библиотеки PyVis.

3. Обработка конъюнкций

Конъюнкция – это операция образования сложных высказываний из более простых и по смыслу эквивалентная соединительному союзу «и» в естественном языке [Студеникина, 2018]. Конъюнкты нами считаются альтернативными составляющими для исходных словосочетаний, поэтому создаются различные вершины для одного и того же исходного выражения, но с различными составляющими так, как если бы это были отдельные выражения в тексте. На рис. 3 продемонстрировано дерево составляющих для предложения «Методы качественной оценки и способы обеспечения безопасности программ и данных».

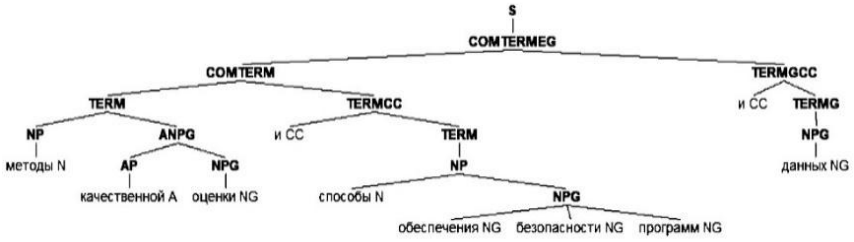


Рис. 3. Дерево составляющих для предложения с конъюнкциями

На рисунке видно, что в предложении обнаружен один комбинированный термин, состоящий из набора терминов в именительном падеже («методы качественной оценки», «способы обеспечения безопасности программ») и термина в родительном падеже («данных»), между терминами стоят союзы «и». Используя наш подход, из данного предложения можно выделить 4 термина (см. рис. 4): «метод качественной оценки безопасности программы», «способ обеспечения безопасности программы», «метод качественной оценки безопасности данных» и «способ обеспечения безопасности данных».

Для обработки конъюнкций использовалась такие операции сложения словосочетаний [Студеникина, 2018]:

$$N_1^1 NG_2^1 NG_3^1 + NG_1^2 = N_1^1 NG_2^1 NG_1^2$$

$$N_1^1 NG_2^1 + NG_1^2 = N_1^1 NG_1^2$$

$$N_1^1 NG_2^1 + NG_1^2 NG_2^2 = N_1^1 NG_1^2 NG_2^2$$

$$N_1^1 NG_2^1 + N_1^2 NG_2^2 NG_3^2 = N_1^1 NG_2^1 NG_3^2$$

Во всех иных случаях словосочетания считаются не принадлежащими конъюнкции и данные операции для формирования отдельных словосочетаний не применяются.

Данный подход позволяет выделять обособленные, несвязанные между собой термины, которые в тексте могли находиться в конъюнкции и, вследствие этого, могли быть захвачены как один связанный термин, что затрудняет смысловую оценку текста из-за наличия сочинительных союзов или запятых в термине.

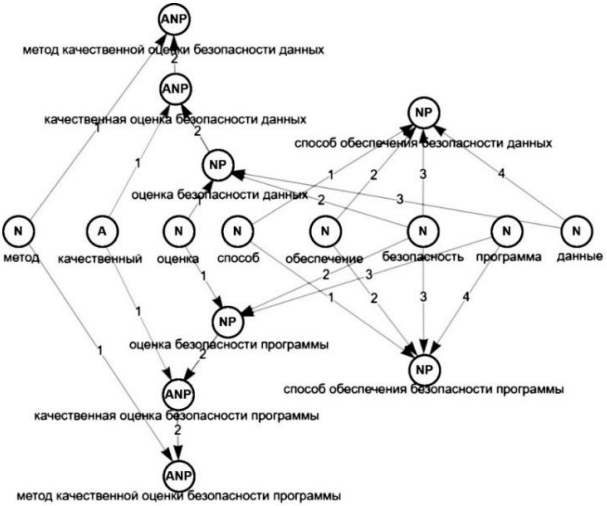


Рис. 4. Граф словосочетаний для предложения с конъюнкциями

4. Нормализация выделенных словосочетаний

Важным аспектом для понимания человеком результатов выделения словосочетаний является их нормализация. В работе [Федяев и др., 2025] при формировании графа смысловых фрагментов используются только исходные выражения из текста требований, что может привести к ошибочной интерпретации терминов из-за омонимии [Демидов, 2023], поэтому необходимо нормализовать данные выражения. Для этого в работе использовался морфологический анализатор Pymorphy2, написанный на языке Python. Схема нормализации представлена на рис. 5.

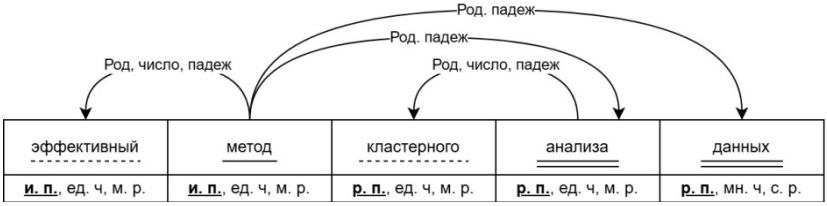


Рис. 5. Схема нормализации именной группы

На рисунке одной чертой выделено главное слово именной группы – существительное, оно всегда в именительном падеже, т.к. принимает форму леммы (нормальной формы слова), пунктиром выделены прилагательные, а двойной чертой – остальные существительные.

При нормализации в нашем случае используется две связи слов в словосочетаниях:

1) согласование, при котором зависимое слово согласуется в роде, числе и падеже с главным (например, «эффективный метод»);

2) управление, при котором зависимое слово ставится в том падеже, которого требует главное (например, «метод анализа»).

Главное слово управляет остальными существительными группы, поэтому они употребляются в родительном падеже. Все прилагательные согласуются со следующим существительным по тексту в роде, числе и падеже.

5. Определение новых словосочетаний

Для апробации рассмотренной идеи проведём эксперимент с текстовыми данными, рассмотренными в предыдущей статье авторов [Федяев и др., 2025]. Определим новые словосочетания для одного предложения из текста требований к специалисту по интеллектуальному анализу данных: «Знать методы дискриминантного и кластерного анализа данных». Результаты извлечения новых словосочетаний представлены на рис. 6. Серым выделены известные словосочетания по тексту рабочей программы, а черным – новые словосочетания. Можно сделать два вывода, во-первых, теперь для выявления новых словосочетаний достаточно вычислить разницу графов требований и рабочей программы дисциплины без промежуточных преобразований графа требований, т.к. оба графа имеют одинаковую структуру. Во-вторых, структура графа даёт чёткое понимание, что в тексте рабочей программы нет точной формулировки «методы дискриминантного и кластерного анализа данных», однако есть «кластерный анализ данных».

Это позволяет сказать, что несмотря на то, что все слова, из которых состоит данный термин, присутствуют в рабочей программе, как таковые понятия «метод кластерного анализа данных» и «метод дискриминантного анализа данных» не фигурируют в тексте рабочей программы. В отличие от предыдущей работы, где весь этот термин считается известным из текста рабочей программы (табл. 2).

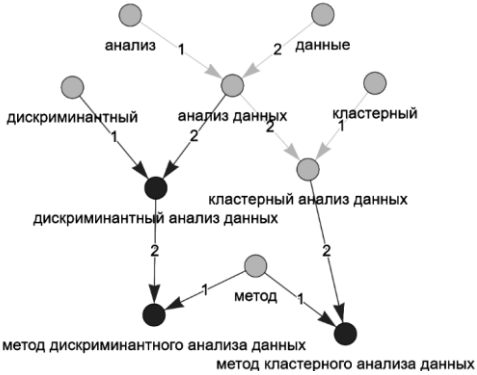


Рис. 6. Определение новых словосочетаний

Таблица 2

Пример анализируемого предложения в тексте требований: «Знать методы дискриминантного и кластерного анализа данных»	**Выделенные словосочетания в предложении**	
	Подход на основе дерева составляющих	**Подход на основе дерева зависимостей**
Новые словосочетания для РПД	Метод кластерного анализа данных, метод дискриминантного анализа данных, дискриминантный анализ данных	Не обнаружены
Известные словосочетания для РПД	Кластерный анализ данных, анализ данных	Методы дискриминантного и кластерного анализа данных

Таким образом, данный подход позволяет точнее идентифицировать части словосочетаний, которые присутствуют в обоих документах, и выделять как общие термины, так и те, которыми документы отличаются по смысловому содержанию. Кроме того, он выявляет в 2–3 раза больше словосочетаний, что подтверждает его более высокую детализацию разбиения на смысловые фрагменты.

Заключение

В работе предложен улучшенный способ сравнения текстовых документов, проиллюстрированный на примере извлечения новых смысловых фрагментов из текста требований предприятий по отношению к знаниям и навыкам, представленным в текстах рабочих программах дисциплин выпускающей кафедры. В качестве решения предлагается однородное представление сопоставляемых текстов в виде графов словосочетаний, которые позволяют при помощи разности графов получить отличительные смысловые фрагменты.

Извлечение словосочетаний формализовано при помощи набора правил с использованием регулярных выражений. Для представления всего текста в виде графа словосочетаний разработана структура самого графа и алгоритм преобразования набора деревьев составляющих в единый граф, представляющий весь текст.

Также было проведено экспериментальное сравнение предложенного подхода с ранее опубликованными результатами авторов данной работы. Эксперимент показал, что предложенный подход позволяет более правильно оценить смысловую новизну термина при сравнении документов требований предприятий и рабочей программы дисциплины. В отличие от подхода, используемого авторами в предыдущей работе [Федяев и др., 2025], данный подход благодаря разделению термина на словосочетания в чёткой синтаксической иерархии позволяет выделить конкретный фрагмент термина, который является новым по смыслу словосочетанием для текста рабочей программы дисциплины (см. пункт 5). К тому же, обработка конъюнкций позволила разделить термин с включением союза «и» на два отдельных обособленных термина, что было невозможно в прежнем подходе.

Таким образом, новизна и практическая значимость предложенного подхода заключается в применении дерева составляющих вместо дерева зависимостей для извлечения словосочетаний, обработке конъюнкций в тексте и нормализации извлечённых словосочетаний, что в целом позволяет более правильно определять новые смысловые фрагменты для текста рабочей программы дисциплины и повышает их интерпретируемость человеком (лектором).

Список литературы

[Ботов, 2019] Ботов Д.С. Интеллектуальная поддержка формирования образовательных программ на основе нейросетевых моделей языка с учетом требований рынка труда // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2019. – Т. 19, № 1. – С. 5-19. – doi: 10.14529/ctcr190101.

[Григорьева и др., 2023] Григорьева Е.Г., Клячин В.А., Помельников Ю.В., Попов В.В. Алгоритм выделения ключевых слов на основе графовой модели лингвистического корпуса // Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2017. – Т. 16, № 2. – С. 58-67. – doi: 10.15688/jvolsu2.2017.2.6.

[Демидов, 2023] Д.В. Демидов. Представление синтаксических структур с сочинительными конструкциями и омонимией // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2023. – Т. 21, № 4. – С. 17-45. – doi: 10.25205/1818-7900-2023-21-4-17-45.

[Космачёва и др., 2016] Космачёва И.М., Квятковская И.Ю., Сибикина И.В. Автоматизированная система формирования рабочих программ учебных дисциплин // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2016. – № 1. – С. 90-97.

[Кравченко и др., 2024] Кравченко Д.Ю. [и др.]. Алгоритм оптимизации извлечения ключевых слов на основе применения лингвистического парсера // Информатика и автоматизация. – 2024. – Т. 23, № 2. – С. 467-494. – doi: 10.15622/ia.23.2.6.

[Николаев, 2023] Метод извлечения знаний и навыков/компетенций из текстов требований вакансий // Онтология проектирования. – 2023. – Т. 13, № 2(48). – С. 282-293. – doi: 10.18287/2223-9537-2023-13-2-282-293.

[Полетаев и др., 2023] Полетаев А.Ю., Парамонов И.В., Бойчук Е.И. Алгоритм построения дерева синтаксических единиц русскоязычного предложения по дереву синтаксических связей // Информатика и автоматизация. – 2023. – Т. 22, № 6. – С. 1323-1353. – doi: 10.15622/ia.22.6.3.

[Студеникина, 2018] Студеникина К.А. Синтаксис сочинения русских именных групп: эллипсис или малые конъюнкты? // Типология морфосинтаксических параметров. – 2018. – Т. 1, № 2. – С. 115-133.

[Федяев и др., 2025] Федяев О.И., Мелещенко Н.В. Ролевые модели агентов системы моделирования процесса обновления учебных дисциплин с учётом требований предприятий // Проблемы искусственного интеллекта. – 2025. – Т. 36, № 1. – С. 12-25. – doi: 10.24412/2413-7383-12-25.

[Sheetal et. al., 2014] Sheetal S., Kulkarni P. Graph based Representation and Analysis of Text Document: A Survey of Techniques // International Journal of Computer Applications. – 2024. – Vol. 96. – P. 1-8. – doi: 10.5120/16899-6972.

[Wu et. al., 2024] Wu Y., Pan X., Li J., Dou S., Dong J., Wei D. Knowledge Graph-Based Hierarchical Text Semantic Representation // International Journal of Intelligent Systems. – 2024. – P. 1-14. – doi: 10.1155/2024/5583270.

УДК 378.4
doi: 10.15622/rcai.2025.024

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ НА КОЛЛЕКЦИЯХ НАУЧНЫХ ПУБЛИКАЦИЙ

Н.А. Назаров (*straider105@gmail.com*)
М.Р. Шарифуллин (*sharifullin2107@mail.ru*)
В.О. Толчеев (*tolcheevvo@mail.ru*)

Национальный исследовательский университет «МЭИ», Москва

Проанализированы прикладные задачи интеллектуального анализа текстовых данных, для решения которых важно проводить извлечение ключевых слов (КС). Отмечается, что наиболее часто и эффективно КС используются для обработки и анализа англоязычных научных (полнотекстовых и библиографических) документов. Проведено сравнение качества выявления КС (F1@K-мера) на общеизвестных и свободно распространяемых коллекциях текстовых научных данных (Inspec, SemEval-2010, Krapivin), а также датасете из новостных сообщений (DUC2001). В ходе экспериментальных исследований рассмотрены различные технологии извлечения КС: статистический алгоритм YAKE, графовые алгоритмы TopicRank и MultipartiteRank, нейросетевые модели KeyBERT и PromptRank. Проанализирована зависимость параметров от длины документов, оценены показатели качества.

Ключевые слова: извлечение ключевых слов, интеллектуальный анализ текстов, научные публикации, графовые алгоритмы, нейросетевые модели, статистический анализ.

Введение

В интеллектуальном анализе текстовых данных (Text Mining) большое внимание уделяется правильному выбору информативных терминов, которые способны наиболее полно описать смысл документов. Для этого обычно используются отдельные слова или словосочетания. Такие ключевые слова (КС) широко применяются в информационном поиске, суммаризации текстов (аннотирование-реферирование); анализе тональности, при выявлении структуры и визуализации документальных коллекций, ведении диалога в вопросно-ответных системах, мониторинге электрон-

ных сообщений, обнаружении тематических сообществ в сети Интернет, поисковой оптимизации и продвижении сайтов [Song et. al., 2023], [Ajallouda et. al., 2022]. В последнее время КС активно применяются в промпт-инжиниринге (Prompt Engineering) для составления запросов к большим языковым моделям (Large Language Model, LLM), а также при автоматизированной разметке (кластеризации) текстовых данных [Sahoo et. al., 2025], [Liu et. al., 2018], [Ванюшкин и др., 2018].

Дадим определение КС – неслучайно встречающиеся в документах важные понятия и/или их комбинации, отражающие содержание документа и формирующие его смысловое ядро [Шереметьева, 2015]. Объединение отдельных информативных терминов в общие конструкции чаще всего способно обеспечить «приращение смысла» и упростить интерпретацию анализируемых текстов.

Эффективность использования КС для решения задач Text Mining существенно зависит от стиля документа (художественный, публицистический, официально-деловой, научный, разговорный и т.п.) и его размера. Большие художественные и публицистические тексты допускают множественные трактовки, отражая субъективное восприятие и понимание прочитанного. В этом случае КС будут индивидуальными для каждого из читателей, заметно различаясь между собой. При обработке более формальных и частично структурированных материалов (научные статьи, юридические документы, медицинские карты и рецепты, новостные сообщения) извлекаемый набор КС чаще всего одинаково интерпретируется специалистами и представляет «общий код», позволяющий обмениваться важной информацией. По мнению ряда специалистов, научная публикация после прочтения «сворачивается» в ограниченное количество КС (8-10 слов и словосочетаний), которые хранятся в памяти человека и ассоциируются с конкретным текстом [Шереметьева, 2015], [Москвитина, 2009], [Москвитина, 2018].

Отнесение отдельных терминов и их сочетаний к КС зависит от ряда факторов: частоты и места совместной встречаемости, контекста появления, принадлежности к определенным частям речи (обычно к существительным и прилагательным), специфики предметной области. Чем более концентрировано излагаются сведения в документе, тем информативнее получаются выделяемые из него КС. Так, в полнотекстовых научных работах наибольшая «концентрация смыслов» содержится в названии, аннотации, ключевых словах, введении и заключении статьи, в библиографических описаниях – в названии, аннотации и ключевых словах (иногда КС отсутствуют, так как не указаны авторами и не присвоены при рубрикации в электронных библиотеках).

В данной работе рассматриваются англоязычные полнотекстовые и библиографические научные документы. Проводится комплексная экспериментальная оценка точностных и временных характеристик десяти из-

вестных методов выявления КС, определяются те подходы, которые обладают самыми высокими показателями качества обнаружения КС и наилучшим образом подходят для решения прикладных задач обработки научных текстов на основе КС (прежде всего кластеризация, визуализация и классификация).

1. Методы автоматического извлечения КС

К настоящему времени, несмотря на интенсивные исследования, не разработано универсального SOTA (State of The Art) подхода для выделения КС [Song et. al., 2023], [Митрофанова и др., 2022], [Musunuru et al., 2024]. В данной работе рассматриваются известные статистические, графовые и нейросетевые (эмбеддинговые) методы извлечения ключевых слов (МИКС). Их объединяет общий подход к выявлению КС, которое проводится в автоматическом режиме (т.е. используется обучение без учителя). В данной работе реализуется комплексное исследование десяти МИКС, часто используемых на практике [Ajallouda et. al., 2022], [Campos et. al., 2018], [Bougouin et. al., 2013], [Boudin, 2018], [Grootendorst, 2020], [Kong et. al., 2023]:

1) статистические алгоритмы RAKE, YAKE;
2) графовые методы TextRank, SingleRank, EmbedRank, MDERank, TopicRank, MultipartiteRank;
3) нейросетевые модели KeyBERT (на основе модели BERT) и PromptRank (на основе генерации КС с помощью большой языковой модели T5).

Путем экспериментальных исследований с использованием коллекций документов Inspec, SemEval-2010, Krapivin, DUC2001 выделены пять наиболее высокоточных МИКС (YAKE, TopicRank, MultipartiteRank, KeyBERT, PromptRank), которые далее подробно рассматриваются в данной работе.

1.1. Статистический метод YAKE

YAKE [Campos et. al., 2018] является многоязычным подходом (применим для различных языков, включая русский) и основан на извлечении статистических характеристик текста. Выделяемые КС могут быть названиями или аббревиатурами (W_{Case}). При выборе КС учитывается их местоположение в тексте (W_{Pos}), частота слова (W_{Freq}), количество предложений с кандидатом в КС ($W_{DifSentence}$), сходство со стоп-словами (W_{Rel}). Итоговый вес КС вычисляется по формуле:

$$S(w) = \frac{W_{Rel} * W_{Pos}}{W_{Case} + \frac{W_{Freq}}{W_{Rel}} + \frac{W_{DifSentence}}{W_{Rel}}}, \quad (1.1)$$

1.2. Графовый метод TopicRank

Этот метод [Bougouin et. al., 2013] является модификацией TextRank, однако вершинами (полного неориентированного) графа являются не слова (как в TextRank), а темы (кластер из похожих однословных и многословных выражений). Ключевые фразы-кандидаты выбираются из последовательности соседних существительных с одним или несколькими предшествующими прилагательными. Затем они группируются по темам с помощью иерархической агломеративной кластеризации [Boudin, 2018], чтобы выбрать КС, наилучшим образом «представляющие» получившиеся кластеры. Для каждого документа составляется ранжированный (по важности) список тем. Алгоритм TopicRank предусматривает выполнение следующих шагов: предобработка, выявление КС-кандидатов, составление тем (кластеризация КС-кандидатов), ранжирование тем, формирование наиболее информативных и релевантных КС. Полный граф позволяет учесть взаимозависимость тем, вес ребра $w_{i,j}$ между двумя темами (вершинами) t_i и t_j вычисляется на основе близости их ключевых фраз:

$$w_{i,j} = \sum_{c_i \in t_i} \sum_{c_j \in t_j} dist(c_i, c_j), \quad (1.2)$$

Расстояние между позициями ключевых фраз $c_{i,}$ и c_j рассчитывается по формуле:

$$dist(c_i, c_j) = \sum_{p_i \in pos(c_i)} \sum_{p_j \in pos(c_j)} \frac{1}{|p_i - p_j|}, \quad (1.3)$$

Здесь $pos(c_{i,})$, $pos(c_j)$ – множество позиций вхождений ключевых фраз $c_{i,}$и c_j в документе, $|p_i - p_j|$ – расстояние между позициями p_i (позиция КС c_i) и p_j (позиция КС c_j). Таким образом, чем больше значение $dist(c_i, c_j)$, тем сильнее связь между КС $c_{i,}$и c_j. Рейтинг тем в методе определяется по алгоритмам ранжирования PageRank или TextRank. Формула расчета PageRank имеет вид:

$$S(t_i) = (1 - \lambda) + \lambda \sum_{j \in In(t_i)} \frac{1}{|Out(t_j)|} S(t_j), \quad (1.4)$$

Здесь $S(t_i)$ – важность i-ой вершины, $In(t_i)$ – множество вершин, входящих в i-ую вершину ребра, $Out(t_i)$ –множество вершин, связанных с i-ой вершиной исходящими из неё ребрами, λ – коэффициент затухания, который задается пользователем (при значениях 0,1-0,3 выбираются средне связанные КС, при 0,8-1 – сильно связанные КС). В алгоритме ранжирования TextRank используется «улучшенная» формула PageRank:

$$S(t_i) = (1 - \lambda) + \lambda \sum_{t_j \in V_i} \frac{w_{j,i} S(t_j)}{\sum_{t_k \epsilon V_j} w_{j,k}}, \quad (1.5)$$

где $t_j \in V_i$ – число вершин V связанных с темой t_i, отношение под знаком суммы показывает, насколько сильно тема t_j поддерживает t_i.

После ранжирования тем выбираются ключевые фразы из самых важных тем. Для этого используется одна из стратегий: выбор фразы, которая первой встречается в документе; выбор наиболее частотной фразы в теме; выбор фразы – центроида, наиболее схожей с другими КС в кластере.

1.3. Графовый метод MultipartiteRank

Метод MultipartiteRank [Boudin, 2018] является модификацией TopicRank. Однако в нем используется многосторонний направленный граф, где вершины в отличии от TopicRank представляют КС-кандидаты, которые связаны только в том случае, если они относятся к разным темам. В алгоритме предусматривается корректировка весов кандидатов при вычислении их важности, чем выше значение, тем фразы из начала документа получают больший вес $w_{i,j}$:

$$w_{i,j} = w_{i,j} + \alpha e^{(\frac{1}{p_i})} \sum_{c_k \in T(c_j) \setminus \{c_j\}} w_{ki}, \quad (1.6)$$

1.4. Нейросетевой метод KeyBERT

Для выделения КС KeyBERT [Grootendorst, 2020] использует предварительно обученную модель BERT, что позволяет учитывать контекст появления КС в документе. В KeyBERT не вводятся ограничения на допустимые части речи и в качестве КС используются не только существительные и прилагательные, но и глаголы. Выбор репрезентативных КС осуществляется на основе расчета косинусного сходства между эмбеддингом (вложением) КС и эмбеддингом всего текста (отбираются КС, имеющие наибольшие значения косинусной меры).

1.5. Нейросетевой метод Promptrank (T5)

Этот метод [Kong A. et. al., 2023] извлечения ключевых фраз основан на использовании большой языковой модели T5 (Text-To-TextTransfer Transformer), имеющей архитектуру кодера-декодера, и применяет шаблоны (prompts) для выбора наилучших КС. Выбор фраз-кандидатов осуществляется с помощью выделения частей речи (PoS, Part-of-Speech). Для ранжирования КС-кандидатов документ вводится в кодировщик и вычисляется вероятность того, что будет сгенерирована такая же ключевая фраза. Чем выше вероятность, тем более точно КС соответствует документу и получает соответствующий ранг.

2. Описание используемых датасетов

Выявление КС – слабоформализованный процесс, который зависит от ряда факторов, в частности, от размера документов и количества КС, которые наилучшим образом описывают тексты.

Для настройки параметров и сопоставления различных методов в данной работе применяется F1@K-мера (модификация метрики F1-score), которая рассчитывается для топ-K ключевых фраз. F1@K-мера выбрана исходя из того, что она интуитивно понятна и используется в большинстве про-

фильных публикаций для сопоставления различных МИКС. При определении F1@K-меры необходимо вычислить значения точности и полноты для извлеченных КС ($Precision$@K и $Recall$@K), аналогично тому, как это делается в рекомендательных системах [Bjadon A. et. al., 2024]:

$$Precision@K = \frac{\text{Суммарное количество совпадений}}{\text{Общее количество предсказанных ключевых слов}(K)}; \quad (2.1)$$

$$Recall@K = \frac{\text{Суммарное количество совпадений}}{\text{Общее количество истинных ключевых слов}}; \quad (2.2)$$

$$F1@k = 2 * \frac{Precision@K * Recall@K}{Precision@K + Recall@K}. \quad (2.3)$$

Для корректного сравнения МИКС необходимо провести исследования на известных и общедоступных коллекциях научных документов, имеющих «золотой стандарт» – КС, сформированные экспертами (асессорами). В данной работе экспериментальные исследования проводятся на датасетах, которые содержат англоязычные научные публикации разного размера: Inspec, SemEval-2010 и Krapivin.

Набор данных Inspec содержит 2000 библиографических описаний публикаций в области информационных технологий и компьютерных наук с временным охватом 1998–2002 гг. Его основное предназначение – поддержка исследований в области обработки естественного языка, в частности, разработка методов выявления ключевых слов. Средний размер документа составляет 137 слов.

Набор данных SemEval-2010, созданный для проведения конкурса Semantic Evaluation по оценке качества МИКС, состоит из 244 научных статей из цифровой библиотеки ACM, временной охват 2000–2009 гг. Средний размер публикации составляет 230 слов.

Набор данных Krapivin включает 2000 полнотекстовых научных статей по тематикам Computer Science за 2009 год, которые были получены из цифровой библиотеки ACM. Чаще всего датасет используется для исследований в области извлечения ключевых слов и кластеризации текстов. Средний размер документов составляет 8940 слов.

3. Экспериментальное исследование методов автоматического извлечения КС на общедоступных текстовых коллекциях

3.1. Настройка гиперпараметров МИКС (с учетом различий в размерах документов)

Обычно в научной статье автор указывает не менее 5 КС. Для больших публикаций требуется предоставить расширенный список от 10 до 15 КС. Поэтому при сопоставлении МИКС чаще всего анализируются показатели качества для 5, 10 и 15 КС.

Прежде всего настроим гиперпараметры методов и выберем значения, которые обеспечивают наибольшие значения метрики качества F1@K. Экспериментальные исследования показали, что наилучшие гиперпраметры методов несущественно зависят от длины документа и практически идентичны для трех выборок (Inspec, SemEval-2010, Krapivin). В методах Yake и TopicRank все параметры оказались одинаковыми. В PromptRank(T5) различия имеются в значениях весового коэффициента учета позиции слов в тексте ('position_factor' равен 1.2e8 у Inspec и 1.2e9 у Krapivin и SemEval-2010). В методе MultipartiteRank различаются способы объединения кандидатов 'complete' и 'average', а также порог схожести для кластеризации кандидатов (threshold). В методе KeyBert имеются наиболее заметные различия: по числу выделяемых КС (Inspec, SemEval-2010 – 8 КС, Krapivin – 15 КС) и отбору КС (use_mmr). Большинство гиперпараметров методов не существенно зависят от размера документов и обладают высокой универсальностью, что позволяет их применять для различных датасетов.

Настройка гиперпараметров позволила в ряде случаев улучшить качество выявления КС по сравнению с результатами, которые приводятся в широкоизвестных профильных публикациях [Kong et. al., 2023]. Метод Yake показывает более высокие результаты, чем опубликованные ранее, на датасетах Krapivin и SemEval, TopicRank на датасетах Inspec и SemEval2010, MultipartiteRank на датасетах Inspec и SemEval2010, PromptRank на Inspec и Krapivin. KeyBert только на датасете SemEval2010 при показателях F1@10, F1@15.

В данной работе используются следующие настройки гиперпараметров:

1) Yake: n = 3, dedupLim =0,9, dedupFunc = seqm, windowsSize = 2, top = 15.
2) TopicRank: threshold = 0,2, method = average, heuristic = first, n = 15, edundancy_removal = False, stemming = False.
3) MultipartiteRank: threshold = 0,2, method = average, alpha = 1,1, n = 15, edundncy_removal = True, stemming = False.
4) KeyBert: keyphrase_ngram_range = (1,3), use_mmr = False, diversity = False, use_maxsum = False, nr_candidates = False.
5) PromptRank(T5): max_len = 512, temp_en = Book, temp_de = This book mainly this about, mode = base, enable_pos = True, position_factor = 1.2e9, length_factor = 0,6.

3.2. Сравнительный анализ показателей качества МИКС (в зависимости от размера документов)

Далее приводятся результаты исследований МИКС, полученные на разных наборах данных при указанных значениях гиперпараметров. Для оценки качества рассчитываются F1@5, F1@10, F1@15. Результаты представлены в табл. 1, а их визуализация на рис. 1.

Таблица 1

Inspec	Yake	TopicRank	MultipartiteRank	PromptRank	KeyBert
F1@5	6.16	16.69	22.69	32.02	5.94
F1@10	8.60	23.92	27.72	38.26	7.86
F1@15	9.65	27.79	29.64	38.93	9.03
Krapivin					
F1@5	10.50	5.44	7.72	16.13	4.32
F1@10	11.25	7.03	8.30	16.89	5.21
F1@15	10.65	7.46	7.90	17.45	6.13
SemEval					
F1@5	16.39	12.15	13.56	17.24	8.54
F1@10	19.40	15.12	16.43	20.66	13.05
F1@15	19.40	15.12	16.43	21.28	14.52

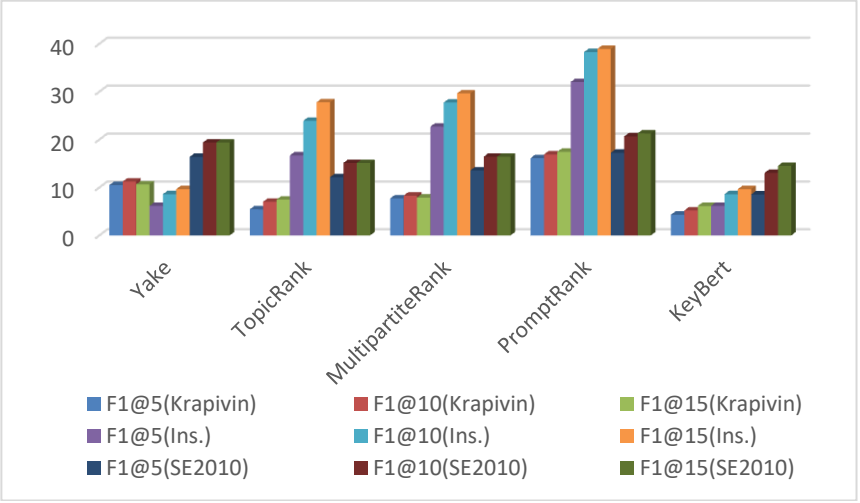


Рис. 1. Значения F1@K-меры на датасетах Inspec, SemEval-2010, Krapivin для различных методов

Анализ результатов исследований позволяет сделать выводы:

1) Показатели качества существенно зависят от размера документов. Все методы имеют низкие значения F1@5, F1@10, F1@15 на полнотекстовом датасете Krapivin.
2) Наилучшее качество продемонстрировал метод PromptRank(T5), который превзошёл остальные МИКС на рассмотренных коллекциях научных документов.
3) «Аутсайдером» практически на всех выборках оказался KeyBert.

3.3. Оценка точностных характеристик МИКС на коллекции новостных сообщений

В предыдущем эксперименте самые высокие показатели качества получены для частично структурированных коротких (библиографических) документов. Проанализируем, можно ли распространить сделанные выводы на другие небольшие тексты, в частности новостные сообщения. Для этого проведем экспериментальное изучение МИКС на коллекции DUC2001. Это набор данных, содержащий 308 новостных сообщений за 2001 год. Средний размер текстов – 845 слов.

Далее в табл. 2 приводятся значения F1@ 5,10, 15, полученные при использовании исследуемых методов для DUC2001.

Таблица 2

DUC2001	Yake	TopicRank	MultipartiteRank	PromptRank	KeyBert
F1@5	12.05	14.09	23.75	27.39	3.01
F1@10	14.44	19.31	25.83	31.59	3.10
F1@15	15.29	22.59	25.38	31.01	3.14

Исходя из результатов, представленных в табл. 1 и 2, можно сделать следующие выводы. На наборе новостных сообщений DUC2001 все методы достигают существенно более высоких показателей качества, чем в случае обработки коллекции полнотекстовых научных документов Krapivin. Однако PromptRank, который является «лидером» на всех датасетах, показывает меньшие значения F1@5, F1@10, F1@15 на DUC2001 по сравнению с Inspec, аналогичное поведение демонстрируют MultipartiteRank и TopicRank. Вместе с тем YAKE, в отличие от остальных МИКС, значительно улучшает показатели качества на DUC2001. Отметим также, что KeyBERT не способен обеспечить качественное выделение КС на DUC2001, показывая крайне низкие значения F1@5, F1@10, F1@15. В публикации [Rao S. et. al., 2022] причинами плохих результатов KeyBERT называют некорректное обрезание ключевых фраз и выделение большого числа КС, относящихся к глагольной группе. На рис. 2 (для данных из табл. 1 и 2) приведены показатели качества методов на датасетах Inspec (библиографические научные документы) и DUC2001 (короткие новостные сообщения).

Наряду с анализом точностных характеристик МИКС существенный интерес представляет их ресурсозатратность, прежде всего временная сложность. Теоретические оценки (с использованием О-нотации) указаны разработчиками в цитируемых нами публикациях, в данной работе приводятся значения процессного времени, что позволяет сравнить производительность методов на датасетах разного размера, состоящих из библиографических, полнотекстовых, новостных документов. Экспериментальные замеры (в секундах) на аппаратной платформе с моделью процессора E5-2640v3 и GPU - RTX 3060ti указаны в табл. 3.

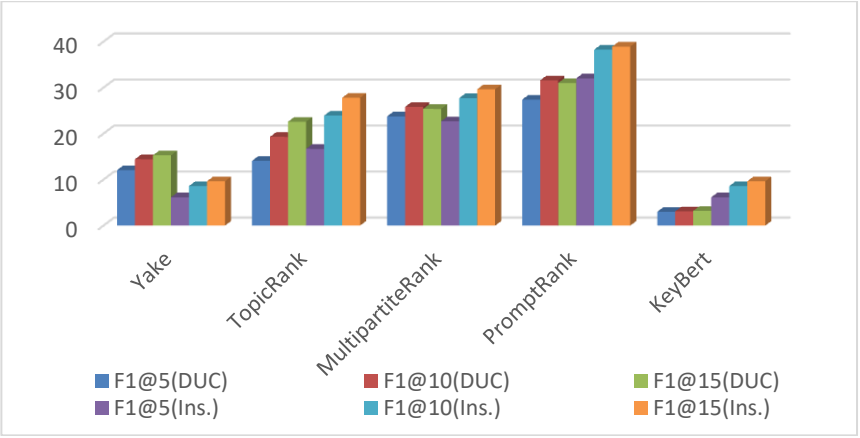


Рис. 2. Значения F1@K-меры на датасетах Inspec и DUC2001

Таблица 3

Датасет / Метод	Inspec	SemEval-2010	Krapivin	DUC2001
YAKE	45	6	63	8
TopicRank	273	30	344	69
MultipartiteRank	338	47	409	81
KeyBERT	895	99	782	163
PromptRank	2051	264	2107	342

Из табл. 3 следует, что наиболее быстрым МИКС является статистический метод YAKE, затем идут графовые методы (TopicRank и MultipartiteRank), наиболее медленными и ресурсозатратными являются нейросетевые модели (KeyBERT и PromptRank). Таким образом, при выявлении КС справедлива общеизвестная закономерность – чем точнее метод, тем больше его алгоритмическая сложность и дольше ожидание результатов.

Заключение

Как и в большинстве нетривиальных слабоформализуемых задач, при выявлении КС из текстов не удается найти единственное универсальное решение. В связи с этим для извлечения КС используются различные подходы, каждый из которых имеет как сильные стороны, так и определенные недостатки. В них по-разному выделяются наиболее существенные КС и даются не полностью совпадающие «приближения» к пониманию смысла научной публикации (или новостного сообщения).

В работе проведен сравнительный анализ наиболее известных и эффективных МИКС, оценена их точность и быстродействие на научных (полнотекстовых и библиографических) коллекциях документов и новостном датасете. Выявлены методы-«лидеры» (PromptRank) и методы-«аутсайдеры» (прежде всего KeyBERT), исследованы характеристики МИКС в зависимости от способа (правила) выявления КС, размера текста и вида документа (научный или новостной). Сформулированные при проведении исследований выводы и рекомендации позволяют сделать шаг в сторону большей формализации изучаемой проблемы, достичь лучшего понимания области применения каждого из методов, оценить степень их универсальности, а также разработать процедуры кластеризации, визуализации и классификации на основе использования извлеченных КС.

Список литературы

[Ванюшкин и др., 2018] Ванюшкин А.С. Гращенко Л.А. Опыт автоматизированной разметки текстов ключевыми словами // Материалы IV Всероссийской научно-практической конференции «Современные проблемы физико-математических наук». – 2018. – С. 320-325.

[Митрофанова и др., 2022] Митрофанова О.А., Гаврилюк Д.А. Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // TerraLinguistica. – 2022. – Т. 13, № 4. – С. 22-40.

[Москвитина, 2009] Москвитина Т.Н. Ключевые слова и их функции в научном тексте // Вестник Южно-Уральского государственного гуманитарно-педагогического университета. – 2009. – № 11. – С. 270-283.

[Москвитина, 2018] Москвитина Т.Н. Методы выделения ключевых слов при реферировании научного текста // Вестник Томского государственного педагогического университета. – 2018. – № 8. – С. 45-50.

[Шереметьева, 2015] Шереметьева С.О. Методы и модели автоматического извлечения ключевых слов // Вестник Южно-Уральского государственного университета. – 2015. – Т. 12, № 1. – С. 76-81.

[Ajallouda et. al., 2022] Ajallouda L., Fagroud F.Z., Zellou A., Benlahmar E.A Systematic Literature Review of Keyphrases Extraction Approaches // International Journal of Interactive Mobile Technologies (iJIM). – August 2022. – 16(16). – P. 31-58.

[BJadon et. al., 2024] BJadon A., Patil A. A Comprehensive Survey of Evaluation Techniques for Recommendation Systems // arXiv:2312.16015v2, 12 Jan 2024.

[Bougouin et. al., 2013] Bougouin A., Boudin F., Daille B. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction // International Joint Conference on Natural Language Processing (IJCNLP). – 2013. – P. 543-551.

[Boudin, 2018] Boudin F. Unsupervised Keyphrase Extraction with Multipartite Graphs // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2018. – P. 667-672.

[Campos et. al., 2018] Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Collection-independent automatic keyword extractor // Proceedings of 40th European Conference on Informational Retrieval (ECIR). – 2018. – P. 806-810.

[Grootendorst, 2020] Grootendorst M. KeyBERT: Minimal Keyword Extraction with BERT. – 2020. – URL: http://doi. org/10.5281/zendo.4461265 (дата обращения: 15.02.2025).

[Kong et. al., 2023] Kong A., Zhao S., Chen H., Li Q., Qin Y., Sun R., Bai X. PromptRank: Unsupervised Keyphrase Extraction Using Prompt // In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. – 2023. – P. 9788-9801.

[Liu et. al., 2018] Liu Q., Kawahara D., Li S. Scientific Keyphrase extraction: extracting candidates with semi-supervised data augmentation // Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. – 2018. – P. 183-194.

[Musunuru et al., 2024] Surveying Keyword Extractor: Classification, Applications, and Empirical Analysis // 2024 Parul International Conference on Engineering and Technology (PICET). – IEEE, 2024. – C. 1-8.

[Papagiannopoulou et. al., 2020] Papagiannopoulou E., Tsoumakas G. A review of keyphrase extraction // Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery. – 2020. – 10(2).

[Rao et. al., 2022] Rao S, Nasirian Sara, Ghoshal Parijat. Keyword Extraction in Scientific Documents // arXiv:2207.0188v2 7 Juk 2022.

[Song et. al., 2023] Song M., Feng Y., Jing L. A Survey on Recent Advances in Keyphrase Extraction from Pre-trained Language Models // In Findings of the Association for Computational Linguistics: EACL. – 2023. – P. 2153-2164.

[Sahoo et. al., 2025] Sahoo P., Singh A.K., Saha S., Jain V., Mondal S., Chadha A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications // arXiv:2402.07927v2 [cs.AI] 16 Mar 2025.

УДК 004.852+519.765
doi: 10.15622/rcai.2025.025

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ АРХИТЕКТУРЫ TRANSFORMER В ЗАДАЧЕ УПРОЩЕНИЯ ТЕКСТА

Н.А. Прокопьев (*nikolai.prokopyev@gmail.com*)
О.А. Невзорова (*onevzoro@gmail.com*)
Ф.М. Гафаров (*fgafarov@yandex.ru*)
А.А. Гафиатуллин (*arslan2911@mail.ru*)
А.Р. Зиастинов (*ziastinovalmaz@gmail.com*)

Казанский федеральный университет, Казань

В статье рассматривается задача упрощения текстов на русском языке с использованием моделей архитектуры Transformer. Выбор наилучшей модели производится с помощью стандартных оценочных метрик BLEU, ROUGE, SARI и уточненной метрики лексической сложности на основе оценки читаемости. Исследовались предобученные модели T5 и BART, обучение производилось на основе набора данных в области информатики. В результате было выявлено, что модель BART лучше справляется с задачей упрощения текста, а также генерирует тексты, более соответствующие по форме эталонным.

Ключевые слова: лексическая сложность текста, метрика, архитектура трансформер, набор данных.

Введение

Упрощение текста – это активно развивающаяся область обработки естественного языка (NLP), направленная на адаптацию текстов для различных групп читателей, в том числе людей с ограниченной грамотностью, носителей языка с низким уровнем владения, лиц с когнитивными нарушениями и широкую аудиторию, нуждающуюся в доступном изложении сложных материалов. Другой целевой аудиторией являются учащиеся средних школ и студенты высших учебных заведений. Усвоение учебного материала – это важнейшая составляющая процесса обучения, для успешного усвоения разрабатываются различные методики. Персонализированный подход к обучению включает этапы, связанные с оценкой уровня подготовки обучаемого и предъявления ему учебного материала в соответствии с уровнем подготовки.

Таким образом, актуальной является задача, связанная с подготовкой учебных материалов, отличающихся оценками лексической, синтаксической и терминологической сложности текста при сохранении его ключевого смысла. В статье рассматриваются основные технологии, применяемые для решения поставленной задачи снижения сложности учебных текстов, основные метрики оценивания сложности текстов, а также разработанный исследовательский прототип системы снижения сложности учебных текстов в области информатики.

Статья структурирована следующим образом: в разделе 1 приведен обзор основных методов, применяемых для решения задачи упрощения текста, в разделе 2 изложены основные метрики оценивания сложности текста, далее в разделе 3 описаны решения по разработке исследовательского прототипа системы снижения сложности текста и проведен сравнительный анализ применяемых вычислительных моделей, в заключении представлены основные выводы и направления будущих разработок.

1. Методы автоматического упрощения текста

В настоящее время разрабатываются методы автоматического упрощения текстов, включая модели на основе глубокого обучения [Sheang et al., 2021], [Das et al., 2025], гибридные архитектуры и специализированные подходы в обработке научных текстов [Anjum 2023].

Современные методы упрощения текстов основываются на предобученных моделях на основе архитектуры Transformer. Модель T5 (Text-to-Text Transfer Transformer) демонстрирует высокую гибкость благодаря унифицированному подходу "текст-в-текст", что позволяет настраивать степень упрощения [Sheang et al., 2021]. Модель BART (Bidirectional and Auto-Regressive Transformer) превосходит T5 по ряду метрик, особенно в задачах сохранения связности текста [Das et al., 2025].

Для улучшения качества упрощения текста разработаны также гибридные методы. Например, метод SIMPLEX [Truică et al., 2023], который сочетает Word2Vec и трансформеры (BERT, RoBERTa, GPT-2). Другой пример – MultiLS [North et al., 2024], первая многозадачная архитектура, поддерживающая прогнозирование сложности, генерацию и ранжирование замен. Управление упрощением через контрольные токены, позволяющие регулировать длину и сложность текста исследуется в [Dmitrieva, 2023]. Используя гибридные модели трансформеров, разрабатываются методы, которые учитывают потребности конкретных аудиторий, например, студентов с дислексией [Sukiman et al., 2023].

Для русского языка созданы специализированные наборы данных, такие как RuSimpleSentEval [Sakhovskiy et al., 2021] и корпус аннотированных предложений [Ivanov et al., 2023], что способствует развитию методов упрощения для славянских языков. Эксперименты с управляемым упрощением [Dmitrieva, 2023] подтвердили эффективность моделей mBART и T5 для решения задачи упрощения русских текстов.

2. Метрики оценивания сложности текста

2.1. Задача оценки результата упрощения текста

Одна из главных проблем в задаче упрощения текста – создание автоматической метрики оценки, которая могла бы заменить трудоемкую экспертную оценку. Хорошая метрика должна учитывать три ключевых свойства упрощенного текста: грамматичность, сохранение смысла и простоту (насколько легко воспринимается упрощенный вариант).

Исчерпывающий обзор по автоматическим метрикам оценки в задаче упрощения текста содержится в [Alva-Manchego et al., 2020]. Можно выделить несколько классов автоматических метрик оценки задачи упрощения текста, которые используются в настоящем проекте.

2.2. Оценочные метрики машинного перевода

2.2.1. Метрика BLEU (BiLingual Evaluation Understudy) оценивает качество перевода на основе точности совпадения n-грамм между текстом, полученным путем машинного перевода (кандидатом) и эталонным текстом [Papineni et al., 2002]. Задача упрощения текста в этом случае рассматривается как задача перевода сложного текста в более простой. Метрика принимает значения от 0 до 1, в некоторых источниках нормализуется в диапазон от 0 до 100, где больший показатель означает большее качество. Метрика BLEU ориентирована на точность (чем больше совпадающих n-грамм, тем выше оценка) и не учитывает порядок слов.

2.2.2. Метрика ROUGE (Recall-Oriented Understudy for Gisting Evaluation) представляет собой семейство метрик, предназначенных для оценки суммаризации текста [Lin, 2004]. Метрики вычисляют степень совпадения n-грамм, последовательностей и пар слов между алгоритмически сгенерированными аннотациями и эталонными аннотациями, составленными экспертами. Принимают значения от 0 до 1, в некоторых источниках нормализуются в диапазон от 0 до 100, где больший показатель означает большее качество.

В ROUGE-1 сравниваются единицы (слова) между сгенерированным и эталонным текстами. В ROUGE-2 сравниваются последовательности из двух слов, взятых из сгенерированного и эталонного текста. ROUGE-L ищет самую длинную последовательность, которая является общей для двух текстов, а затем измеряет её длину.

2.3. Оценочные метрики упрощения текста

Метрика SARI (System output Against References and Input sentence) сравнивает текст, полученный на выходе модели упрощения текстов (кандидат) с несколькими эталонными текстами и исходным текстом [Xu et al., 2016]. Идея SARI заключается в том, чтобы вознаграждать модели за добавление n-грамм, которые встречаются в любом из эталонных текстов, но не в исходных текстах; вознаграждать сохранение n-грамм как

в кандидатах, так и в эталонных текстах; вознаграждать сохранение важных n-грамм. Таким образом, метрика вычисляет среднее арифметическое оценок точности и полноты операций добавления, сохранения и удаления n-грамм. В настоящее время эта метрика стала стандартом для оценки и сравнения моделей упрощения текста. Она принимает значения от 0 до 100, где больший показатель означает большее качество.

2.4. Метрики, основанные на оценке читаемости текста

2.4.1. Метрика Флеша (Flesch Reading Ease) оценивает читаемость текста на основе двух ключевых параметров: средняя длина предложения (в словах) и средняя длина слова (в слогах). Диапазон: от 0 (очень сложно) до 100 (очень легко) [Flesch, 1948].

2.4.2. Метрика Флеша-Кинкайда (Flesch-Kincaid Grade Level, FKGL) Оценки Флеша-Кинкайда представляет собой пересчет метрики *FRE* с коэффициентами, полученными на основе процедур множественной регрессии в тестах по чтению [Kincaid et al., 1975].

2.4.4. Иные метрики. Помимо рассмотренных выше метрик, можно отметить более сложные метрики, например, метрику упрощения текста *SAMSA* [Sulem et al., 2018], которая использует семантический анализ исходного текста, а также метрику *Coh-Metrix* [McNamara et al., 2014], разработанную для анализа текстов по множественным характеристикам и уровням языка и дискурса.

3. Разработка исследовательского прототипа для системы снижения сложности текста

3.1. Общая схема последовательности этапов разработки

Для проведения сравнительного анализа моделей упрощения текста была проведена разработка исследовательского прототипа (общая схема разработки представлена на рис. 1), в рамках которой выполнено:

1. Подготовка общего набора данных в виде набора пар текстов, в которых один текст является сложным, а другой – упрощенной версией этого текста. В набор входят как данные на основе авторских текстов, так и полностью сгенерированные данные.
2. Разделение общего набора данных на набор эталонных данных и на набор данных для обучения моделей. Набор эталонных данных должен статистически соответствовать набору обучающих данных и состоять из проверенных экспертами вручную пар текстов.
3. Обучение выбранных моделей на основе обучающего набора. Этап включает предварительную подготовку конфигурации обучения, само обучение и проверку результатов обучения с использованием стандартной метрики потерь (loss). Отбираются наиболее удачные конфигурации и результаты обучения.
4. Оценка моделей на основе рассмотренных оценочных метрик, а также уточненной метрики лексической сложности.

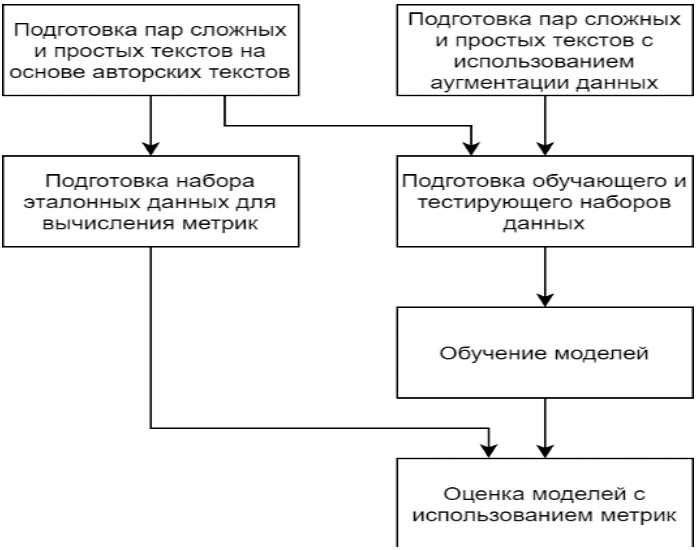


Рис. 1. Общая схема последовательности этапов разработки прототипа

3.2. Уточненная метрика лексической сложности

Для оценки сложности текстов была разработана формула для метрики лексической сложности (Lexical Complexity, LC) путем уточнения представленных ранее формул на основе метрики Флеша и иных оценок читаемости текста:

$$RI = \frac{FKLG/_{18} + FRE/_{100} + CLI/_{14} + SMOG/_{18} + ARI/_{14} + LIX/_{100}}{6}$$

$$LL = \frac{TTR + RTTR + CTTR}{3}$$

$$LC = \frac{RI + LL}{2} * 10.$$

В приведенных формулах RI – индекс читаемости, вычисляемый на основе метрик читаемости: FKGL (метрика Флеша-Кинкайда), FRE (метрика Флеша), CLI (метрика Колман-Лиау), SMOG (индекс SMOG), ARI (автоматический индекс удобочитаемости), LIX (индекс LIX).

LL – лексический уровень, вычисляемый на основе метрик лексического разнообразия: TTR (коэффициент лексического разнообразия, Token Type Ratio), RTTR (корневой коэффициент лексического разнообразия), CTTR (скорректированный коэффициент лексического разнообразия).

Параметры формул подробно описаны в статье [Поляков, 2024]. Все перечисленные параметры рассчитывались с использованием библиотеки Russian Texts Statistics[1], в которой все числовые коэффициенты адаптированы для русского языка на основе набора данных проекта Plain Russian[2]. Формулы для вычисления параметров также представлены на сайте[3].

Итоговая метрика лексической сложности не имеет установленной области допустимых значений, однако позволяет сравнивать сложность текстов, чем больше значение метрики, тем более сложным полагается текст.

3.3. Подготовка наборов данных

В качестве источника авторских текстов для общего набора данных использовалась Большая российская энциклопедия[4] в области информатики. Экспертная оценка извлеченных текстов, а также оценка на основе метрики LC показала, что эти тексты являются простыми. Для получения усложненных версий этих текстов была использована большая языковая модель Deepseek R1-Distill-Llama-70B-Q8[5] с промптом на усложнение текста (рис. 2).

Сгенерируй формальный, научный текст, содержащий сложные предложения и терминологию, который будет примерно равен по длине следующему простому тексту:
{target}
Ключевые требования:
– Текст должен быть формальным и научным по стилю,
– Текст должен содержать сложные предложения и терминологию,
– Длина сложного текста должна быть примерно равна длине простого текста.

Рис. 2. Промт на усложнение текста

Для аугментации данных использовалась та же модель: сначала с использованием отдельного промпта были сгенерированы простые тексты в энциклопедическом стиле в области информатики, затем были получены усложненные версии этих текстов. Полученный при помощи большой языковой модели набор текстов был валидирован экспертами.

Построение вручную эталонного набора пар текстов (простой-сложный) выполнялось на основе требований:

[1] https://github.com/SergeyShk/ruTS.

[2] https://github.com/infoculture/plainrussian.

[3] https://sergeyshk.github.io/ruTS/stats/readability_stats_funcs/.

[4] https://bigenc.ru/.

[5] https://ollama.com/library/deepseek-r1:70b-llama-distill-q8_0.

- Средняя LC пар текстов эталонного набора должна соответствовать средней LC пар текстов набора данных для обучения,
- Средняя длина простых текстов эталонного набора должна соответствовать средней длине простых текстов набора данных для обучения.

Тексты, не попавшие в эталонный набор, были оставлены в наборе данных для обучения. В табл. 1 представлены характеристики полученных наборов данных: количество пар, средняя длина текстов в словах, средняя лексическая сложность.

Таблица 1

Характеристика	Значение
Общее количество пар текстов	568
Предметная область	информатика
Набор данных для обучения	
Количество пар текстов в наборе	468
Количество пар на основе авторских текстов	241
Количество пар на основе аугментации данных	227
Средняя длина простых текстов	155 слов
Средняя длина простых авторских текстов	67 слов
Средняя длина простых аугментированных текстов	288 слов
Средняя LC сложных текстов	105,9
Средняя LC простых текстов	86,4
Набор эталонных данных	
Количество пар текстов в наборе	100
Средняя длина простых текстов	145 слов
Средняя LC сложных текстов	108,0
Средняя LC простых текстов	86,3

3.4. Обучение моделей

Для эксперимента по сравнительному анализу моделей архитектуры Transformer в задаче упрощения текста были выбраны модели T5 и BART, предобученные версии для русского языка: FRED-T5-large[6] и ru-bart-large[7]. Данные модели ранее были исследованы в решении этой задачи и показали хорошо применимые результаты [Das et al., 2025].

Модели были обучены на подготовленном наборе данных с различными конфигурациями обучения. В результате была выявлена наиболее удачная конфигурация, дающая наилучший результат обучения по стандартной метрике потерь (loss). В табл. 2 представлены характеристики этой конфигурации и полученные на ней результаты обучения для каждой модели.

[6] https://huggingface.co/ai-forever/FRED-T5-large.

[7] https://huggingface.co/sn4kebyt3/ru-bart-large.

Таблица 2

Характеристика	T5	BART
Конфигурация обучения		
Количество эпох	10	8
Разбиение данных	90/10%	90/10%
Размер батча	2	8
Скорость обучения	1e-4	3e-5
Стратегия оценки	по эпохам	каждые 50 шагов
Метрика обучения	eval_loss	eval_loss
Результаты обучения		
train_loss в начале	4,01	4,02
train_loss в конце	0,65	0,44
eval_loss в начале	0,76	0,94
eval_loss в конце	0,66	0,77

Результаты представлены в виде показателей метрики обучения в начале и в конце, после прохождения всех эпох. При успешном обучении происходит уменьшение метрики, чем ближе она к 0, тем успешнее результат.

3.5. Оценка моделей

Для обученных моделей была произведена оценка с использованием набора эталонных данных и оценочных метрик BLEU, ROUGE, SARI. Несмотря на то, что среди них непосредственно метрикой для задачи упрощения текста является SARI, метрика BLEU для машинного перевода и семейство метрик ROUGE для суммаризации текста тоже дают полезную информацию для сравнительного анализа, так как они иным образом проверяют соответствие сгенерированного моделью текста эталонному. Более подробно смысл сравнения моделей по метрикам BLEU и ROUGE раскрыт далее в разделе 4.

Кроме оценочных метрик была вычислена средняя разница лексической сложности LCdiff для сгенерированных и эталонных текстов:

$$LCdiff = |LC_c - LC_s| - |LC_c - LC_m|,$$

где LCc – это LC сложного эталонного текста, LCs – это LC простого эталонного текста, LCm – это LC сгенерированного простого текста.

Таким образом, положительное значение LCdiff означает, что модель сгенерировала текст более сложный по метрике LC, чем эталонный, а отрицательное значение – что модель сгенерировала более простой текст.

В табл. 3 представлены результаты оценки моделей T5 и BART.

Таблица 3

Метрика	T5	BART
BLEU	34,1	41,7
ROUGE-1	53,9	53,9
ROUGE-2	41,8	42,4
ROUGE-L	52,5	52,4
SARI	47,1	56,2
LCdiff	+5,45	+1,55

4. Сравнительный анализ моделей

Сравнение моделей по оценочным метрикам показывает, что модель BART показывает лучшие результаты по метрикам BLEU и SARI и сопоставимые результаты по метрикам семейства ROUGE (см. табл. 3).

Лучший результат по метрике BLEU означает, что модель BART генерирует упрощенные тексты, которые более точно соответствуют эталонным простым текстам в аспекте полного соответствия n-грамм. Таким образом, BART генерирует тексты, которые более соответствуют эталонным по форме, чем T5. Этот результат соответствует результату, полученному в [Das et al., 2025] с оценками BLEU/T5 37,8 и BLEU/BART 41,5.

Сопоставимые результаты по метрикам семейства ROUGE означают, что модели генерируют тексты, содержащие одинаковые униграммы (отдельные слова), биграммы (пары слов) и последовательности слов наибольшей длины. Следовательно, обе модели в равной степени сохраняют лексику и формулировки текстов, одинаково качественно суммаризуют смысл исходных сложных текстов. Этот результат не соответствует исследованию [Das et al., 2025], в котором было выявлено, что модель T5 имеет меньшие оценки ROUGE (например, ROUGE-1/T5 48,2 и ROUGE-1/BART 52,7).

Лучший результат по метрике SARI показывает, что модель BART лучше справляется с задачей упрощения текста, чем T5, что также не соответствует результату [Das et al., 2025], где показано, что разница по этой метрике между моделями несущественна (SARI/T5 39,4 и SARI/BART 38,7).

Сравнение по средней разнице лексической сложности LCdiff соответствует сравнению по метрике SARI – обе модели генерируют более сложные тексты по сравнению с эталонными, однако BART генерирует более близкие по лексической сложности тексты.

В табл. 4 представлены показательные примеры упрощения текстов эталонного набора моделями T5 и BART.

Таблица 4

Эталонный текст	Выход T5	Выход BART
Наиболее важным примером сложности алгоритма является время его работы, измеряемое числом элементарных шагов …	Алгоритмы, для которых максимальная временная сложность, наблюдаемая в экстремальных условиях…	Одним из ключевых показателей сложности алгоритма является время его работы…
Наиболее производительные универсальные микропроцессоры разрабатывают и производят компании Intel, AMD и IBM. В России пять основных разработчиков микропроцессоров…	В России основными производителями микропроцессоров являются МЦСТ, Baikal Electronics, Научно-технический центр «Модуль» …	Компании Intel, AMD и IBM разрабатывают и производят самые мощные универсальные микропроцессоры. В России наиболее крупными производителями универсальных микропроцессоров являются…

Общие выводы сравнительного анализа следующие:

- На основе экспертной оценки сгенерированных моделями упрощенных текстов на основе набора эталонных данных и метрик ROUGE, SARI, LCdiff можно утверждать, что обе модели приемлемо справляются с задачей упрощения текстов на русском языке, сохраняя при этом смысл, лексику и формулировки, однако BART справляется несколько лучше.
- Результаты проведенного эксперимента отражены в табл. 3 и частично соответствуют результатам, полученным в исследовании [Das et al., 2025], в частности, в ключевом выводе относительно модели BART по оценке BLEU. Разница в выводах по метрикам ROUGE и SARI может быть объяснена различием в источниках данных (использовались тексты на английском языке из ресурсов Simple Wikipedia и Newsela), в объеме наборов данных для обучения, в конфигурации обучения.
- Разница в метрике LCdiff корректно предсказывает разницу в метрике SARI, что является обоснованием корректности используемой уточненной формулы лексической сложности LC.

Заключение

В результате исследования было проведено дообучение моделей T5 и BART архитектуры Transformer с последующим сравнительным анализом на основе оценочных метрик BLEU, ROUGE, SARI и уточненной метрики лексической сложности. Анализ показал, что обе модели приемлемо справляются с задачей упрощения текста, при этом BART показывает лучшие результаты, а также вносит меньше изменений в форму текста. Кроме того, анализ показал соответствие результатов существующим аналогичным исследованиям и обоснованность используемой уточненной

формулы лексической сложности. Свойства метрик: грамматичность, сохранение смысла, простота требуют дальнейшего оценивания в последующих исследованиях.

Дальнейшее направление исследований связано с разработкой и применением метрики терминологической сложности, а также с расширением другими источниками текстов, предметными областями и тематиками.

Список литературы

[Alva-Manchego et al., 2020] Alva-Manchego F., Scarton C., Specia L. Data-Driven Sentence Simplification: Survey and Benchmark. Computational Linguistics. – 2020. – Vol. 46(1). – P. 135-187.

[Anjum et al., 2023] Anjum A., Lieberum N. Automatic Simplification of Scientific Texts using Pre-trained Language Models: A Comparative Study // In: Proc. CLEF Symposium. – 2023. –URL: https://ceur-ws.org/Vol-3497/paper-242.pdf (дата обращения: 29.05.2025).

[Das et al., 2025] Das S., Basak D., Bhattacharjee A. Text Simplification Using T5 Model and BART Model // In: Proc. 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech). Kolkata, India, 2025. – P. 1-6. – doi: 10.1109/IEMENTech65115.2025.10959517.

[Dmitrieva, 2023] Dmitrieva A. Automatic text simplification of Russian texts using control tokens // In: Proc. 9th Workshop on Slavic Natural Language Processing (SlavicNLP 2023). Dubrovnik, Croatia, 2023. – P. 70-77. – doi: 10.18653/v1/2023.bsnlp-1.9.

[Flesch, 1948] Flesch R. A new readability yardstick // Journal of Applied Psychology. – 1948. – Vol. 32(3). – P. 221-233.

[Ivanov et al., 2023] Ivanov V., Gamal E.M. A new dataset for sentence-level complexity in Russian // In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2023”. June 14–16, 2023.

[Kincaid et al., 1975] Kincaid J.P., Fishburne R.P., Rogers R.L., Chissom B.S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Institute for Simulation and Training. – 1975. – URL: https://stars.library.ucf.edu/istlibrary/56 (дата обращения: 30.05.2025).

[Lin, 2004] Lin C. ROUGE: a package for automatic evaluation of summaries // In: Proc. ACL 2004 Workshop on Text Summarization Branches Out. Barcelona, Spain, 2004. – P. 74-81.

[McNamara et al., 2014] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, USA.

[North et al., 2024] North K., Ranasinghe T., Shardlow M., Zampieri M. MultiLS: An End-to-End Lexical Simplification Framework // In: Proc. 3rd Workshop on Text Simplification, Accessibility and Readability (TSAR 2024). Miami, Florida, USA, 2024. – P. 1-11. – doi: 0.18653/v1/2024.tsar-1.1.

[Papineni et al., 2002] Papineni K., Roukos S., Ward T., Zhu W. Bleu: A method for automatic evaluation of machine translation // In: Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). Philadelphia, Pennsylvania, 2002. – P. 311-318. – doi: 10.3115/1073083.1073135.

[Sakhovskiy et all 2021] Sakhovskiy A., Izhevskaia A., Pestova A. et al. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian // In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. – 2021. – P. 607-617.

[Sheang et al., 2021] Sheang K.C., Saggion H. Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer // In: Proc. 14th International Conference on Natural Language Generation. Aberdeen, Scotland, UK, 2021. – P. 341-352. – doi: 10.18653/v1/2021.inlg-1.38.

[Sukiman et al., 2023] Sukiman S.A., Husin N.A., Hamdan H., Murad M.A.A. A Hybrid Personalized Text Simplification Framework Leveraging the Deep Learning-based Transformer Model for Dyslexic Students // Journal of Advanced Research in Applied Sciences and Engineering Technology. – 2023. – Vol. 34(1). – P. 299-313. – doi: 10.37934/araset.34.1.299313.

[Sulem et al., 2018] Sulem E., Abend O., Rappoport A. Semantic Structural Evaluation for Text Simplification // In: Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018). New Orleans, Louisiana, 2018. – P. 685-696. – doi: 10.18653/v1/N18-1063.

[Sun et al., 2023] Sun R., Xu W., Wan X. Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification. In: Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada, 2023. – P. 9345-9355. – doi: 0.18653/v1/2023.findings-acl.595.

[Truică et al., 2023] Truică C.O., Stan A.I., Apostol E.S. SimpLex: a lexical text simplification architecture // Neural Computing and Applications. – 2023. – Vol. 35. – P. 6265-6280. – doi: 10.1007/s00521-022-07905-y.

[Xu et al., 2016] Xu W., Napoles C., Pavlick E. et al. Optimizing statistical machine translation for text simplification // Transactions of the Association for Computational Linguistics. – 2016. – Vol. 4. – P. 401-415.

[Поляков, 2024] Поляков А.М., Зоидзе Э.А. Количественные критерии оценки сложности текста для методических целей // Наука в мегаполисе. – 2024. – № 3(59).

УДК 004.912
doi: 10.15622/rcai.2025.026

СРАВНЕНИЕ ПОДХОДОВ К ИНТЕРПРЕТАЦИИ ЯЗЫКОВЫХ МОДЕЛЕЙ: АНАЛИЗ МЕТОДА НА ОСНОВЕ МАСКИРОВАННОГО ЯЗЫКОВОГО МОДЕЛИРОВАНИЯ И ТРАДИЦИОННЫХ МЕТОДОВ

А.А. Рогов (*rogov.alisher@gmail.com*) [A]
Н.В. Лукашевич (*louk_nat@mail.ru*) [A,B]

[A] Московский государственный технический университет им. Н.Э. Баумана, Москва
[B] Московский государственный университет им. М.В. Ломоносова, Москва

С развитием предобученных языковых моделей, таких как BERT, их применение охватывает всё более ответственные сферы – от рекомендательных систем до медицинской диагностики. Однако рост сложности моделей требует не менее активного развития методов интерпретации, позволяющих понять, на основе каких признаков модель принимает решения. В данной работе исследуются подходы к объяснению поведения BERT в задачах текстовой классификации. Основное внимание уделено сравнению двух парадигм: современных методов, основанных на маскированном языковом моделировании и промпт-обучении, и традиционных техник, таких как LIME и методы на основе векторной близости. Экспериментальный анализ проводится на датасетах Web of Science и 20Newsgroups. Для оценки качества интерпретаций используется построение графиков активации, что позволяет визуализировать значимость входных токенов для конечного предсказания.

Ключевые слова: интерпретация нейросетевых моделей, метод LIME, маскированное языковое моделирование, вербализатор, классификация текста.

Введение

Современные языковые модели, такие как BERT [Devlin et. al., 2019], представляют собой сложные системы, которые часто воспринимаются как «чёрные ящики». Это создаёт барьеры для их применения в критически важных областях, где требуется уверенность в том, как модель принимает решения.

Одним из ключевых факторов, способных повысить доверие к моделям машинного обучения, является интерпретируемость. Чем лучше пользователь понимает логику работы системы, тем выше вероятность её принятия и эффективного использования. Особенно это важно в таких сферах, как медицина, финансы или юриспруденция, где ошибка может иметь серьёзные последствия.

В рамках этой работы мы продолжили наше предыдущее исследование [Rogov et. al., 2024] по интерпретации моделей в задаче классификации текста, делая акцент на человеко-ориентированном подходе. Подразумевается, что качественное объяснение должно быть семантически связано с предсказанной категорией – пользователь должен видеть явную связь между выделенными моделью признаками и смыслом класса.

Для исследования были выбраны три метода: PromptExplainer [Feng et. al., 2024], LIME [Ribeiro et. al., 2016] и метод на основе векторной близости. Все они предоставляют ранжированный список слов с весами, отражающими степень влияния каждого слова на финальное решение модели. Для сравнительного анализа использовался метод активационных графиков: в модель последовательно подавались фрагменты текста, содержащие по 10% самых значимых слов, и оценивалась вероятность сохранения исходного предсказания. Эксперименты проводились на датасетах Web of Science и 20Newsgroups.

1. Методы интерпретации

1.1. Метод на основе векторного представления слов

Для построения интерпретаций текста на основе семантической близости мы использовали подход, основанный на сравнении векторных представлений слов из текста с вектором, соответствующим целевой категории. Пусть $D = \{d_1, d_2, \dots, d_n\}$ – множество слов текста, подлежащего анализу, а q_i – вектор, представляющий метку категории i. В качестве меры схожести между словами и классом применялось косинусное расстояние между их векторными представлениями:

$$\boldsymbol{similarity}(\boldsymbol{d_j}, \boldsymbol{q_i}) = \mathbf{cos}(\overrightarrow{\boldsymbol{d_j}}, \overrightarrow{\boldsymbol{q_i}}),$$

где $\overrightarrow{\boldsymbol{d_j}}$ – вектор слова d_j из текста, а $\overrightarrow{\boldsymbol{qi}}$ – вектор метки класса q_i.

В качестве источников векторных представлений были выбраны две хорошо зарекомендовавшие себя модели: GloVe[1] [13] и fastText[2] [5]. Эти модели имеют широкое применение в задачах NLP благодаря своей спо-

[1] http://nlp.stanford.edu/data/glove.840B.300d.zip.

[2] https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.en.zip.

собности отражать лексическую и семантическую информацию. GloVe строит вложения на основе статистики совместной встречаемости слов, тогда как fastText учитывает внутреннюю структуру слов, что позволяет ему эффективно обрабатывать омонимы и редкие слова.

Несмотря на существование более современных моделей, таких как MiniLM-L12-H384[3], которые демонстрируют улучшенные характеристики по качеству и размеру, мы решили сосредоточиться на проверенных решениях. Это позволило нам получить простой базовый метод, необходимую для корректного сравнения с другими методами.

1.2. LIME

LIME (Local Interpretable Model-agnostic Explanations) – это метод локальной интерпретации, не зависящий от внутренней структуры модели. Его основная идея заключается в аппроксимации поведения сложной модели в окрестности анализируемого примера с помощью более простой и понятной модели, например, линейной регрессии.

В задачах классификации текста входной пример представляется бинарным вектором, где каждая компонента соответствует наличию или отсутствию определённого слова. При этом модель может использовать сложные признаки, такие как эмбеддинги, недоступные для прямой интерпретации.

Формально пусть $x \in R^d$ – объясняемый пример, $f : R^d \to R$ – функция, реализуемая моделью, а $g \in G$ – интерпретируемая модель, где G – класс простых моделей. Обычно используется линейная модель вида:

$$g(x') = \phi_0 + \sum_{i=1}^{d'} \phi_i x'_i,$$

где $x' \in \{0,1\}^{d'}$ – упрощённое представление примера, а ϕ_i – веса, характеризующие вклад признаков.

Для построения интерпретации минимизируется функционал:

$$\xi(x) = arg\ \min_{g \in G} L(f, g, \pi_x) + \Omega(g)\ ,$$

где L – мера ошибки аппроксимации, $\pi_x(z)$ – функция близости между образцами, а $\Omega(g)$ – штраф за сложность модели.

Таким образом, LIME позволяет выделить ключевые признаки, повлиявшие на конкретное предсказание, что особенно важно при работе с текстовыми данными.

1.3. PromptExplainer

PromptExplainer – метод интерпретации языковых моделей, основанный на парадигме промпт-обучения (prompt-based learning). Его ключевая идея – использовать внутреннюю задачу маскированного языкового моде-

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

лирования (masked language modeling, MLM) для оценки значимости токенов, вместо внешних аппроксимаций (таких как градиентные или внимательные методы). Такой подход позволяет получать объяснения, согласованные с реальными механизмами принятия решений модели. Метод состоит из двух основных этапов.

Первым этапом является проекция представлений токенов в пространство словаря. Представления всех токенов (включая незамаскированные) подаются в MLM-голову модели, которая проецирует их в пространство размером, равным количеству слов в словаре. Формально:

$$H_v = M_h(H),$$

где $H \in R^{n \times d}$ – матрица представлений токенов, M_h – MLM-голова, а $H_V \in R^{n \times V}$ – проекция в пространство словаря размера V. Каждый токен теперь представлен вектором, элементы которого указывают на вероятность его связи с каждым словом словаря.

Вторым этапом является извлечение дискриминативных признаков с помощью вербализатора. После получения пространства объясняющих признаков используется вербализатор – отображение между классами и набором слов-меток. Вербализатор помогает выделить те слова из пространства H_V, которые наиболее коррелируют с конкретным классом. Это позволяет сформировать окончательное объяснение в виде списка слов, релевантных данному предсказанию:

$$H_D = V(H_V),$$

где V – функция вербализации, а $H_D \in R^{n \times p}$ – матрица дискриминативных признаков для p классов. Затем для каждого класса вычисляется softmax-нормализованная вероятность которая служит мерой влияния токена на предсказание:

$$E_i = Softmax(H_D)[:, c_i].$$

В исследовании использовались два типа вербализаторов: KPT [Hu et al., 2022] и LogReg. Оба подхода формируют набор слов, связанных с каждым классом задачи, что позволяет строить интерпретации предсказаний модели.

KPT представляет собой расширенный вербализатор, использующий внешние источники знаний. Такой подход не ограничивается одним фиксированным словом на класс, а создаёт богатое семантическое представление, охватывающее разные уровни абстракции и связи между понятиями. Для построения вербализатора мы опирались на следующие источники[4,5,6]. Эти ресурсы, хотя и работают независимо, внутри используют та-

[4] relatedwords.org.
[5] describingwords.io.
[6] reversedictionary.org.

кие базы знаний, как WordNet и ConceptNet, а также учитывают информацию из предобученных векторных представлений слов. Автоматически собранные слова могут содержать шум, поэтому после этапа построения проводилось несколько шагов фильтрации:

- Relevance Refinement (RR) – отбор слов по релевантности целевому классу (удаляются слова с весом ниже порога).
- Frequency Refinement (FR) – удаление слов, редко встречающихся в корпусе. Это помогло отсеять термины, которые, несмотря на семантическую связь, практически не встречались в реальных примерах.
- Contextualized Calibration (CC) – корректировка с учётом частотности в MLM, что снижает риск систематических ошибок.
- Learnable Refinement (LR) – обучение весовых коэффициентов слов для их ранжирования по влиянию на итоговое предсказание.

LogReg – вербализатор на основе логистической регрессии, который строился без привлечения внешних источников знаний. На первом этапе тексты датасета векторизовались с помощью TF-IDF. Затем обучалась модель логистической регрессии, которая для каждого класса сохраняла веса признаков, отражающие вклад слова в вероятность отнесения текста к этому классу. Слова с наибольшими по модулю весами формировали список ключевых терминов класса. Такой подход позволяет учитывать специфику корпуса и адаптировать интерпретации под конкретную задачу.

2. Методы оценки объяснений: активация

Для оценки качества построенных интерпретаций мы использовали метод возмущения входного текста [Ali et. al., 2022] – один из подходов для анализа объяснимости моделей. Основная идея заключается в том, чтобы проверить, насколько точно интерпретация отражает реальные признаки, которыми модель руководствуется при принятии решения.

Процедура оценки:

1. Исходный текст заменяется на «пустую» версию, где все токены заменены на <unk>.
2. Слова исходного текста ранжируются по значимости на основе весов, выданных методом интерпретации.
3. В текст постепенно возвращаются наиболее важные слова – порциями по 10% от общего числа слов.
4. После каждого шага измеряется вероятность того, что модель выдаст исходное предсказание для целевого класса. Этот показатель называется активационной вероятностью.
5. Для итоговой оценки используется метрика активации – усреднённое значение этой вероятности по всем шагам добавления слов, что отражает скорость и качество восстановления правильного предсказания на основе ключевых признаков.

Чем быстрее растёт активационная вероятность при добавлении небольшого числа слов, тем точнее считается объяснение, что указывает на верное выделение моделью ключевых признаков.

Подход, ранее использованный в работах [Ali et al., 2022], показал чувствительность к качеству интерпретаций, устойчивость к шуму и универсальность для различных методов. В наших экспериментах все интерпретации строились на одной модели с использованием официальных реализаций, что исключало влияние архитектурных различий и позволяло сравнивать именно способность методов выделять значимые признаки.

3. Наборы данных

Для проведения экспериментов были выбраны два широко известных датасета: 20Newsgroups[7] и Web of Science (WOS) [Kowsari et. al., 2017]. Оба датасета предназначены для задачи многоклассовой классификации текста и имеют различия как в структуре данных, так и в сложности задачи.

Датасет WOS представляет собой коллекцию аннотаций научных публикаций, взятых из базы данных Web of Science. Он содержит три версии корпусов разного размера: 5736, 11967 и 46985 документов, соответствующих 11, 34 и 134 темам соответственно. В рамках исследования мы работали с версией, которая включает 11967 текстовых примеров, относящихся к 34 темам. Выборка была разделена в соотношении 70% / 30% на обучающую и валидационную части.

Особую ценность этого датасета представляет наличие двух уровней классификации, что позволяет анализировать методы интерпретации как на уровне широких научных областей, так и при работе с узкими специализациями:

- Первый уровень (WOS_L1) классификации включает семь обширных научных направлений, таких как «Информатика», «Электротехника», «Психология», «Машиностроение», «Строительная инженерия», «Медицинские науки» и «Биохимия».
- Второй уровень (WOS) классификации содержит 34 более специализированные категории, среди которых можно выделить такие темы, как «Обработка изображений», «Машинное обучение», «Социальное восприятие», «Гидравлика», «Генетика» и другие.

Для работы с составными названиями категорий, такими как «Machine learning» или «Water Pollution», мы использовали усреднение векторов отдельных слов. Это обеспечило корректную работу методов, основанных на векторных представлениях (например, GloVe и fastText), и позволило точно отразить семантику сложных меток.

[7] http://people.csail.mit.edu/jrennie/20Newsgroups/.

Датасет 20Newsgroups состоит из 18846 сообщений из новостных групп, охватывающих 20 различных тем. Данные отличаются большей свободой стиля и разнообразием тем, что усложняет классификацию по сравнению с WOS. Использовалось стандартное разбиение: 14846 примеров для обучения и 4000 для валидации.

Перед обучением из сообщений удалялись метаданные (заголовки, служебная информация), чтобы модель ориентировалась на содержимое текста. Длина документа была ограничена 1000 слов для ускорения обучения и оценки.

Оба датасета были выбраны по нескольким причинам:

- Они покрывают разные домены: научные аннотации (WOS) и пользовательские сообщения (20Newsgroups).
- Содержат разное количество классов и примеров, что позволяет оценить эффективность методов на задачах разной сложности.
- Хорошо исследованы и часто используются в литературе, что обеспечивает возможность сравнения с результатами других работ.

4. Эксперименты

В данной работе был реализован метод интерпретации на основе семантической близости слов из текста и названия класса, для чего использовались предобученные модели векторных представлений слов GloVe и fastText. Каждое слово сравнивалось с вектором класса по косинусной мере близости, после чего слова ранжировались по убыванию этой меры, формируя список наиболее релевантных слов для данной категории.

Для сравнения мы также применили традиционные методы интерпретации. Модель BERT (bert-base-uncased) [2] была дообучена на наших датасетах в стандартной задаче классификации. На основе этой дообученной модели с помощью библиотеки LIME строились объяснения для каждого текста. В LIME передавалась фактическая метка класса, чтобы обеспечить корректное сравнение и исключить влияние ошибок модели на результаты интерпретации. Параметры LIME были установлены следующим образом: максимальное количество признаков в объяснении – 256, размер окрестности для локальной модели – 300.

Для реализации PromptExplainer мы использовали фреймворк OpenPrompt [Ding et. al., 2022], который предоставляет удобную инфраструктуру для промпт-обучения. Здесь использовалась модель BERT (bert-base-uncased) и обучение проходило в условиях few-shot – по 5 примеров на класс. Формирование промпта осуществлялось по шаблону «[Category: <MASK>] Текст». После обучения вычислялась значимость слов относительно фактической метки класса, что позволяло исключить влияние ошибок классификации на качество интерпретаций. В качестве вербализаторов использовались два типа: основанный на внешних знаниях KPT и

построенный на весах модели логистической регрессии LR. Объяснения формировались с помощью softmax-весов слов из вербализатора, что позволяло ранжировать слова по степени их связи с целевой категорией.

В табл. 1 представлены результаты классификации с помощью промпт-обучения с вербализаторами KPT, LogReg (Prompt_KPT, Prompt_LogReg) и с помощью дообученной (Fine-tuned) модели BERT, измеренные метрикой accuracy. Табл. 2 показывает результаты оценки качества объяснений с помощью метрики активации. В таблице сравниваются методы PromptExplainer с двумя вербализаторами (LogReg и KPT), классический LIME, семантические методы на базе GloVe и fastText.

Таблица 1

	WOS	WOS_L1	20Newsgroups
Fine-tuned	86.3	93.1	71.3
Prompt_KPT	43.6	65.3	52.7
Prompt_LogReg	52.6	70.1	56.4

Таблица 2

	WOS	WOS_L1	20Newsgroups
LogReg	31.3	60.2	39.4
KPT	30.2	58.0	38.3
LIME	33.9	63.4	43.5
GloVe	28.9	52.7	35.6
fastText	28.0	56.4	36.6

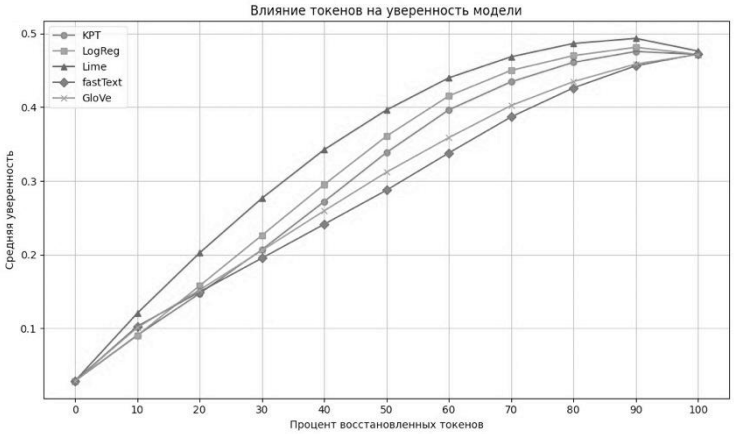


Рис. 1. График активации для датасета WOS

На рис. 1, 2 и 3 представлены графики активации для датасетов WOS, WOS_L1 и 20Newsgroups соответственно. Согласно приведённым данным, метод LIME демонстрирует наилучшие результаты среди всех рассмотренных подходов.

Хотя метод LIME показал лучшие результаты по метрике активации во всех рассмотренных экспериментах, это не означает, что PromptExplainer уступает концептуально. Различия в подходах могут объяснять наблюдаемые результаты.

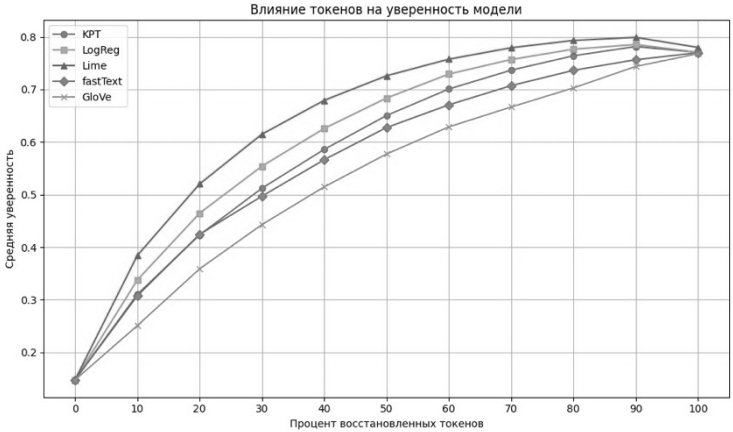


Рис. 2. График активации для датасета WOS_L1

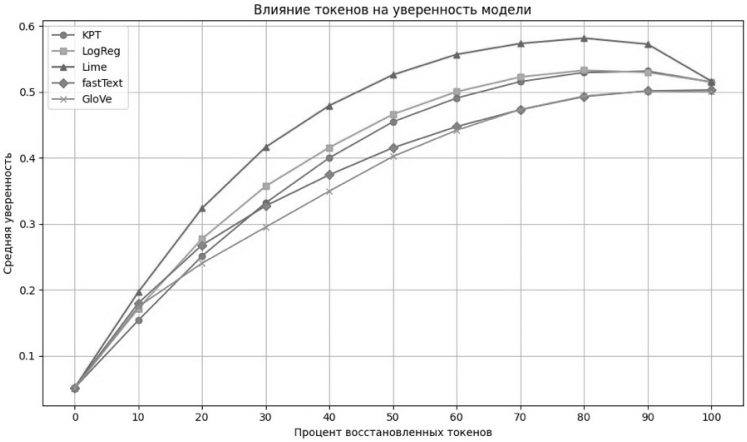


Рис. 3. График активации для датасета 20Newsgroups

LIME является локальным методом, который строит интерпретацию, аппроксимируя поведение модели в окрестности конкретного примера, что позволяет ему точнее выявлять релевантные признаки для данного конкретного текста. В то же время PromptExplainer опирается на глобально сформированные вербализаторы и промпты, которые могут не всегда достаточно гибко учитывать индивидуальные особенности каждого примера.

Кроме того, качество PromptExplainer сильно зависит от выбора и качества вербализатора. В данной работе использовались два типа вербализаторов: один на основе внешних знаний и один на основе весов логистической регрессии. Несмотря на то, что оба подхода обеспечивают адекватные результаты, возможно, они не отражают всех нюансов семантической релевантности для конкретных задач и текстов.

Также стоит отметить, что LIME напрямую использует предобученную BERT-модель, дообученную на конкретном датасете, что даёт ему преимущество в более точной локальной интерпретации. PromptExplainer в экспериментах работал в few-shot режиме, что ограничивает объем доступной информации для обучения промптов и, как следствие, может снижать качество интерпретаций.

Таким образом, разница в результатах скорее отражает особенности реализации и настройки методов, а не фундаментальные ограничения PromptExplainer. Перспективным направлением дальнейших исследований может стать улучшение вербализаторов. Это позволит более полно раскрыть потенциал PromptExplainer и сделать его более конкурентоспособным с существующими методами.

Заключение

В ходе исследования были рассмотрены и сравнены различные методы интерпретации предобученных языковых моделей, с особым акцентом на подход PromptExplainer, который базируется на внутренней задаче MLM и использовании вербализатора для построения объяснений. Также проведено сравнение с традиционными методами, такими как LIME и метод на основе косинусной близости слов.

Анализ экспериментов показал, что PromptExplainer способен формировать осмысленные и достаточно точные объяснения, однако его эффективность существенно зависит от качества и выбора вербализатора. В то же время метод LIME продемонстрировал более стабильные и высокие результаты во всех оценочных метриках, особенно при анализе активационных графиков, что связано с его локальным подходом к интерпретации и использованию дообученной модели BERT.

Полученные результаты указывают на существующие ограничения текущей реализации PromptExplainer и необходимость дальнейших исследований, направленных на улучшение вербализаторов и адаптацию метода к особенностям конкретных задач и данных.

Таким образом, данное исследование подчёркивает важность учёта как архитектурных особенностей предобученных моделей, так и качества вербализаторов при построении интерпретаций, а также открывает новые направления для развития методов объяснимости в области NLP.

Список литературы

[Ali et. al., 2022] Ali A., Schnake T., Eberle O., Montavon G., Müller K.R., & Wolf L. XAI for transformers: Better explanations through conservative propagation // In International conference on machine learning. – 2022, June. – P. 435-451. PMLR. (*статья в сборнике трудов конференции на англ. языке*).

[Devlin et. al., 2019] Devlin J., Chang M.W., Lee K., & Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Vol. 1 (long and short papers). – 2019, June. – P. 4171-4186). (*статья в сборнике трудов конференции на англ. языке*).

[Ding et. al., 2022] Ding N., Hu S., Zhao W., Chen Y., Liu Z., Zheng H., & Sun M. OpenPrompt: An Open-source Framework for Prompt-learning // In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. – 2022, May. – P. 105-113. (*статья в сборнике трудов конференции на англ. языке*)

[Feng et. al., 2024] Feng Z., Zhou H., Zhu Z., & Mao K. PromptExplainer: Explaining Language Models through Prompt-based Learning // In Findings of the Association for Computational Linguistics: EACL 2024. – 2024, March. – P. 882-895. (*книга на англ. языке*)

[Hu et. al., 2022] Hu S., Ding N., Wang H., Liu Z., Wang J., Li J., ... & Sun M. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification // In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2022, May. – P. 2225-2240. (*статья в журнале на англ. языке*)

[Kowsari et. al., 2017] Kowsari K., Brown D.E., Heidarysafa M., Meimandi K.J., Gerber M.S., & Barnes L E. Hdltex: Hierarchical deep learning for text classification. In 2017 16th IEEE international conference on machine learning and applications (ICMLA). – 2017, December. – P. 364-371. IEEE. (*статья в сборнике трудов конференции на англ. языке*)

[Ribeiro et. al., 2016] Ribeiro M.T., Singh S., & Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier // In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. – 2016, August. – P. 1135-1144. (*статья в сборнике трудов конференции на англ. языке*)

[Rogov et. al., 2024] Rogov A.A., & Loukachevitch N.V. Evaluating the Performance of Interpretability Methods in Text Categorization Task // Lobachevskii Journal of Mathematics. – 2024. – 45(3). – P. 1234-1245. (*статья в сборнике трудов конференции на англ. языке*).

УДК 004.912
doi: 10.15622/rcai.2025.027

КЛАССИФИКАЦИЯ АРГУМЕНТОВ С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ[1]

Е.А. Сидорова (*lsidorova@iis.nsk.su*)
А.С. Серый (*alexey.seryj@iis.nsk.su*)
И.Р. Ахмадеева (*i.r.akhmadeeva@iis.nsk.su*)
Д.В. Ильина (*dviljina@gmail.com*)

Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск

Статья посвящена разработке методов автоматической классификации аргументов в русскоязычных текстах с применением генеративных языковых моделей и промпт-инжиниринга. Исследование проводилось на трех датасетах с разметкой аргументации схемами Д. Уолтона: корпусе автоматически сгенерированных англоязычных аргументов NLAS, англоязычных корпусе Araucaria и русскоязычном корпусе ArgNetSC. Для классификации аргументов применялись три стратегии: 1) классификация по схемам Уолтона с использованием формальных определений, 2) классификация на основе систематизации схем аргументов, и 3) последовательный вывод с помощью диалога. Исследование с помощью моделей семейства Mistral показало, что наиболее эффективной является диалоговая модель общения с LLM и стратегия автоматического подбора семантически близких примеров для техники Few-Shot. Лучшие оценки составили 0.63 и 0.31 F_1-меры для англоязычных корпусов и 0.15 для русскоязычного корпуса. Для исследования качества предварительной фильтрации классов использовалась модель DeepSeek-R1-Distill-Llama-70B и лучшая стратегия, полученная на первом этапе. F_1-мера по классам схем составила 0.615, а истинная схема аргумента попала в отфильтрованный список схем в 78.5% случаев.

Ключевые слова: анализ аргументации, классификация аргументов, схемы аргументов Уолтона, систематизация схем аргументов, промпт-инжиниринг.

[1] Исследование выполнено при финансовой поддержке Российского научного фонда № 23-11-00261, https://rscf.ru/project/23-11-00261/.

Введение

Задача интеллектуального анализа аргументации (Argument Mining, AM) заключается в извлечении аргументов из текста и связывании их в единую структуру, позволяющую проводить дальнейшие исследования, опираясь на цельное представление аргументации заданного текста. Наличие аргументативной структуры позволяет сделать некоторые выводы о характере текста: связности, последовательности рассуждений, наличии или отсутствии обоснования и т.п. Добавление информации о типах используемых аргументативных приемов расширяет область применения AM на более значимые приложения, такие как анализ юридических споров [Walker et al., 2018], [Habernal et al., 2024] и политических дебатов [Lippi et al., 2016], выявление ошибок в рассуждениях и недостоверной информации [Yan et al., 2021], рецензирование проектов [Baimuratov et al., 2025], анализ мнений пользователей и др. Она позволяет оценить, насколько правомерны или убедительны используемые аргументы, поддерживают ли они тезис или опровергают, опираются ли они на факты или на мнения людей.

Существуют различные подходы к моделированию и систематизации аргументов [Lawrence et al., 2016]. Базовой моделью аргумента можно считать модель С. Тулмина [Toulmin, 2003], согласно которой аргумент описывается шестикомпонентой структурой: тезис, посылки, обоснование вывода, поддержка утверждений, границы применимости и степень уверенности. Многие последующие модели аргументации в той или иной степени опираются на эту модель [Freeman, 2011], но на практике имеют более простую структуру. Так, широко известные схемы рассуждения Дугласа Уолтона [Walton, 2009] оперируют только тремя компонентами: тезис, посылки и исключения. Создание онтологии аргументации на основе модели Уолтона [Walton et al., 2008], которая включает около 60 классов (схем аргументации), послужило стимулом многих практических исследований, связанных с разработкой аннотированных ресурсов и ручной разработкой наборов данных (датасетов) для автоматического анализа структурированной аргументации.

Многие исследования [Cabrio et al., 2018], [Zhang et al., 2022], [Shaefer et al., 2021] демонстрируют, что предобученные трансформерные архитектуры достигают высокого качества в различных задачах AM. Однако их эффективность существенно зависит от наличия репрезентативных датасетов. В области AM наблюдается дефицит аннотированных корпусов, а среди существующих лишь немногие содержат детальную классификацию аргументов. В этом контексте появление больших языковых моделей (Large Language Models, LLM) открывает новые перспективы. Благодаря их способности к обобщению языковых паттернов, полученной

в ходе претренинга на разнородных текстах, они потенциально могут решать задачи АМ в условиях few-shot или zero-shot обучения, что является особенно актуальным при недостатке размеченных данных.

Целью данной работы является разработка методов автоматической классификации аргументов в русскоязычных текстах с применением LLM и промпт-инжиниринга. Для достижения поставленных целей в рамках данной работы были сформулированы следующие вопросы исследования.

RQ1. Насколько эффективно можно применить LLM и методы промпт-инжиниринга для автоматической многоклассовой классификации аргументов?

RQ2. Эффективно ли применение систематизации схем аргументов для предварительного сужения множества классов при многоклассовой классификации аргументов?

Экспериментальное сравнительное исследование проводится на корпусах текстов, снабженных аргументативной разметкой в соответствии с моделью Д. Уолтона.

1. Обзор связанных работ

Классификация аргументов – отдельная подзадача АМ, направленная на типизацию уже выявленных аргументативно связанных утверждений согласно заданным категориям. В самом простом случае рассматриваются две категории: поддерживающий и атакующий тип. Классификация аргументов может также включать определение схемы аргумента или его качества (например, сильный аргумент против слабого) или другие признаки, определяемые структурными компонентами аргумента [Wagemans, 2016], [Kononenko et al., 2023].

С развитием средств глубокого обучения нейросетевые подходы стали широко применяться при решении задач АМ, таких как распознавание аргументации в тексте (argument detection) или классификация точки зрения (stance detection) [Ruggeri et al., 2021], Ying et al. 2018]. Более современные исследования все чаще направлены на оценку возможности применения генеративных LLM [Chen et al., 2024]. При этом работ, посвященных проблеме классификации аргументов, значительно меньше, что связано, в том числе, с небольшим количеством размеченных данных. В работе [Pietron et al., 2024] представлена ансамблевая модель, основанная на архитектуре BERT, и ChatGPT-4 в качестве модели тонкой настройки. Модель осуществляла классификацию пар утверждений на трех англоязычных наборах данных, при этом количество классов было невелико – от 2 до 4. Авторы [Bezou-Vrakatseli et al., 2025] проводят сравнительное исследование семи LLM разного размера на задаче Zero-Shot и Few-Shot классификации по семи схемам. При этом утверждается, что данная работа является первым исследованием подобного рода.

Сложность аннотирования богатых аргументами данных, ограничения, связанные с размерами корпусов, разнообразием языков и предметных областей, послужили мотивацией к разработке методики автоматической генерации мультиязычного корпуса текстов, размеченных с помощью схем аргументации. Корпус NLAS-multi [Ruiz-Dolz et al., 2024] – мультиязычный датасет, аннотированный 20 различными схемами аргументации Д. Уолтона. Корпус создавался с помощью генеративных моделей GPT-3.5-turbo и GPT-4 и ручной фильтрации результатов; соответствие примеров заданной схеме составило более 90%. Хотя этот корпус и является одним из самых больших и богатых корпусов с точки зрения схем аргументации, он состоит из скомпилированных схем, содержащих полные структуры рассуждений, которые редко используются в текстах, написанных людьми.

Качество работы LLM зависит от качества промптов и размера модели. Если выбор модели зависит от финансовых возможностей и ресурсов разработчиков, то выбор стратегии генерации промптов зависит только от них самих.

В работе [Pietron et al., 2024] используется простой запрос вида:

Это тезис "...". А это утверждение "...". Является ли утверждение аргументом за или против? Напиши ответ одним предложением.

Анализ ошибок показывает, что модель испытывает трудности при распознавании множественных отрицаний, сарказма, вопросительных утверждений и т.п. Добавление в запрос указания на необходимость объяснить ответ показывает, что модель испытывает трудности и при распознавании логических зависимостей и может противоречить сама себе.

Наша гипотеза заключается в том, что необходимо давать модели больше информации о типах логических связей, разбивать задачу на более простые шаги, приводить примеры (стратегия Few-Shot), использовать диалог для контроля вывода на основе дерева рассуждений.

2. Рассуждения на основе систематизации схем аргументов

Отмечая сложности аннотирования на основе существующей таксономии схем Уолтона, авторы [Сидорова и др., 2024] обосновали необходимость дальнейшей систематизации схем аргументации и предложили их многоаспектную классификацию по четырем основаниям[2]:

1. **Онтологическое отношение.** Включает 11 основных отношений, которыми могут быть связаны пары утверждений.

[2] Иерархию и описание схем аргументации можно найти на портале ArgNetBank Studio: https://uniserv.iis.nsk.su/arg/schemes.

2. **Направление атаки:** выделяются аргументы, которые могут атаковать тезис, его внешний источник или аргумент в целом, исходя из недостатков оппонента в диалоге. В отдельный класс («Нет атаки») включены аргументы, которые могут быть только поддерживающими.

3. **Тип заключения:** теоретический (вывод носит истинностный характер, сообщает о положении дел или истинности пропозиции) и практический (вывод сообщает о необходимости, долженствовании, возможности или (не)допустимости совершения того или иного действия).

4. **Источник аргумента:** внутренний (вывод сделан логически) или внешний (вывод сделан в силу качеств субъекта).

Таким образом, формально классификатор можно представить как совокупность из четырех ориентированных ациклических графов: $CL = \{CL_1, CL_2, CL_3, CL_4\}$, где $CL_i = < c_0^i, C^i, S, E^i >$ включает основание c_0^i, множество классов C^i (промежуточных вершин графа, являющихся дополнительными признаками схем, $c_0^i \in C^i$), множество схем аргументации S (конечные вершины графа, множество S общее для всех CL_i) и иерархические связи в графе E^i. Данная структура не является деревом, поскольку схема может соотноситься с несколькими классами (обладать несколькими признаками), но каждый граф обладает свойством: $CL_i \backslash S$ является деревом, что можно использовать для последовательного вывода схемы сверху-вниз.

Данная категоризация должна позволить генеративной модели последовательно уточнять тип аргумента на основании пересечения его признаков. Взаимодействие с моделью заключается в последовательной классификации аргумента по каждому из оснований классификатора и затем определении его схемы.

Вопросы к модели формируются по следующим принципам. Базовая инструкция ограничивает тематические рамки и сообщает общую задачу модели в диалоге (*Ты помогаешь исследовать приемы аргументации в текстах на русском языке*). Далее следуют формулировка задачи, примеры, список классов и входной аргумент, представляемый в виде пары **«посылка–заключение»**. В общем случае аргумент может принадлежать более чем одному классу, поэтому модели ставится задача классификации по многим меткам (Multi-Label). Если семантика классов является взаимоисключающей, это указывается в их описании. Дальнейший диалог продолжается без упоминания входного аргумента и примеров. На каждом шаге модели даются описания классов, являющихся потомками тех классов, которые она назвала на предыдущем шаге, и предлагается уточнить классификацию аргумента.

Для применения генеративных подходов большое значение имеет точность в выборе формулировок для определения каждого классификационного признака. Были применены следующие подходы: (1) выделение ин-

тегрального признака класса; (2) обобщение семантики экземпляров классов; (3) выделение отдельных признаков экземпляров в качестве условия вхождения в класс.

Наиболее корректным представляется использование первого подхода, так как он позволяет выделить семантические признаки аргументативной связи, не представленные в описаниях схем явно. Однако из-за высокой степени обобщенности семантики некоторых классов этот подход не всегда применим. Так, в описании класса «внутренний источник аргумента» реализован второй подход – перечислены возможные условия, при которых пара включается в этот класс: «тезис во фрагменте “заключение” обоснован во фрагменте “посылка” фактами, наблюдениями, логическими рассуждениями, указанием на причинно-следственные, условно-следственные, меронимические, классификационные связи или на последствия действия, о котором идет речь во фрагменте “заключение”».

3. Экспериментальное исследование

Для экспериментального исследования использованы три корпуса с разметкой схем аргументации.

1. **Araucaria** [Reed et al., 2008]: 730 аргументов на английском языке, аннотированные 17 схемами Уолтона. Коллекция сильно несбалансированна, для некоторых схем менее пяти примеров («От непоследовательности убеждений», «Опровержение гипотезы», «От исключения», «От страха», «От общей практики»).

2. **NLAS** (англоязычная часть корпуса NLAS-multi) – автоматически синтезированный корпус, содержащий 1893 аргумента на английском языке, аннотированных 20 схемами Уолтона. Корпус NLAS сбалансирован: 75–100 примеров на каждую схему.

3. **ArgNetSC** [Тимофеева и др., 2024] – русскоязычный корпус текстов, относящихся к научной коммуникации, размещенный на платформе ArgNetBank Studio (https://uniserv.iis.nsk.su/arg). Содержит более 9 тыс. русскоязычных аргументов, аннотированных 42 схемами Уолтона. Как и Araucaria, коллекция является несбалансированной: 90% аргументов покрывается 17 наиболее частотными схемами.

Основным корпусом, на котором проводились все эксперименты, был ArgNetSC; англоязычные корпуса использовались для получения контрастных метрик для английского языка и синтетических текстов.

3.2. Применяемые подходы

В данной работе исследованы три подхода к решению задачи мультиклассовой классификации аргументов на основе промпт-инжиниринга: (1) классификация по схемам (классам) с использованием определений, заданных в онтологии аргументации, (2) множественная классификация на основе дерева признаков и (3) классификация с помощью диалога с моделью на основе дерева признаков.

1. **Base-classification:** эксперимент, в котором модель сразу решала задачу многоклассовой классификации – определение аргументативной схемы, которой связана заданная пара «посылка–заключение».

2. **Tree-classification:** эксперимент, в котором вначале классификация осуществлялась отдельно по каждому основанию классификатора, а затем по множеству схем, соответствующих найденным классам. Каждый аргумент a классифицировался по каждому из классификаторов C^i. На данном этапе модель могла выбирать более одного класса. По результатам классификации $C_a^i \subset C^i$вычисляется ограниченное множество схем $S_a = \bigcap_i \{s \in S^i \big| \exists c \in C_a^i$: в CL_i существует путь из c в $s\}$. На последнем этапе модель классифицирует аргумент по множеству S_a.

3. **Dialog-classification:** эксперимент в технике Prompt Chaining и в диалоговом режиме, использующий всю иерархию классов схем. Для каждого основания c_0^i с моделью проводился диалог. На первом шаге требовалось классифицировать a по множеству $\{c \in C^i \big| (c_0^i, c) \in E^i\}$. На каждом шаге j модель уточняла классификацию по множеству $\{c \in C^i \big| \exists c \in C_a^{j-1}: (c', c) \in E^i\}$, где C_a^{j-1} – результат предыдущего шага. Диалог завершался по достижении уровня схем. Далее аналогично предыдущему подходу вычислялось множество S_a и проводился финальный этап классификации.

Во всех подходах задача классификации аргументов формулировалась в виде инструкции на естественном языке с применением стратегий Zero-Shot (без добавления примеров) и Few-Shot (с небольшим количеством примеров).

Также задачи разделялись по наличию или отсутствию контекста аргументов в инструкциях (**cont**) и методу подбора примеров в стратегии Few-Shot: случайные примеры (**rand**) или примеры, набираемые из аргументов, близких по смыслу к данному (**sim**). Для сравнения смысловой близости все аргументы были векторизованы моделью LaBSE (https://huggingface.co/sentence-transformers/LaBSE).

Результаты усреднялись по пяти случайным разбиениям исходных данных на обучающую и тестовую выборки. При этом каждый аргумент был представлен как минимум в одной тестовой выборке.

3.3. Результаты экспериментов

В качестве языковых моделей применялись модели семейства Mistral: Mistral 7b, Mistral NEMO и Mixtral 8x7B. Последняя на текущий момент крупнее большинства моделей, которые можно развернуть в среде с ограниченными вычислительными ресурсами, обладая при этом сравнимой скоростью вывода. Метрики меньших моделей не превысили значений 0.01, поэтому далее показаны только результаты модели Mistral 7x8B.

В табл. 1 представлены результаты классификации аргументов без предварительной фильтрации по классам (приводятся взвешенные F_1-меры классификации аргументов по схемам).

Таблица 1

Результаты классификации аргументов по схемам

Корпус	Стратегия				
	ZS	ZS+cont	FS-rand	FS-sim	FS-sim+cont
ArgNetSC	0.04	0.04	0.07	**0.14**	0.12
Araucaria	0.2	0.2	0.22	0.25	0.25
NLAS	0.43	–	0.29	0.63	–

Можно увидеть, что a) случайно выбранные примеры могут привести к увеличению числа ошибок модели даже по сравнению с Zero-Shot, b) правильный выбор примеров дает прирост качества, c) наличие или отсутствие контекста мало влияет на результат. В корпусе NLAS контекст аргументов недоступен, поскольку он содержит только синтетические аргументы. Высокие результаты на корпусе NLAS говорят о том, что искусственно сгенерированные согласно схемам аргументы хорошо распознаются моделью: на большинстве схем модель показала F_1 0.89–0.99, а общее значение в 0.63 ниже ожидаемого только за счет схем «От знака», «От примера» и «От свидетельских показаний», качество распознавания которых оказалось существенно ниже остальных.

Табл. 2 содержит результаты классификации с учетом классов схем, представленных в п. 2, и предварительной фильтрацией схем (приводятся усредненные взвешенные F_1-меры и точность фильтрации – доля примеров, когда истинная схема аргумента попала в сокращенный набор схем, полученный путем классификации аргумента по всем основаниям классификатора).

Таблица 2

Результаты классификации аргументов с предварительной фильтрацией по классам схем

Корпус	Стратегия		F_1 по классам	Точность фильтрации	F_1 по схемам
ArgNetSC	Tree	ZS	0.27	0.06	0.03
		ZS+cont	0.33	0.06	0.04
	Dialog	ZS	0.22	0.1	0.04
		FS-sim	0.55	**0.3**	**0.15**
		FS-sim+cont	**0.56**	**0.3**	0.14
NLAS	Dialog	FS-sim	0.51	0.37	0.37
Araucaria	Dialog	FS-sim	0.64	0.45	0.31

Как и в первом случае, наличие или отсутствие контекста мало влияет на результат, но правильный выбор примеров дает прирост качества. Для корпуса Araucaria приведены результаты подхода Dialog как показавшего лучший результат на первом этапе. Определение класса схемы на корпусе NLAS продемонстрировало качество, сопоставимое с результатами, полученными на других корпусах, что привело к снижению качества идентификации схем аргументов.

Качество классификации аргументов, составленных людьми, оказывается значительно ниже. Однако, несмотря на в целом низкое качество определения схемы аргумента, удалось добиться роста среднего качества определения его класса. Таким образом, можно говорить, что генеративная модель без предварительного дообучения показывает низкое качество классификации схемы аргументов, однако такие модели можно применять как помощники при разметке или предварительной фильтрации схем при автоматической классификации аргументов.

Для проверки последней гипотезы был проведен эксперимент по применению самой результативной стратегии (Dialog, FS-sim) в сочетании с моделью большего размера – DeepSeek-R1-Distill-Llama-70B, производной от DeepSeek-R1, способной размышлять последовательно и возвращать ход своих рассуждений; отвечая на вопрос и объясняя причины выбора, она может выступать в качестве ассистента. Модель большего размера, обученная подражать сверхбольшой модели DeepSeek-R1, дала значительный прирост качества: F_1-мера по классам аргументов составила 0.615, а истинная схема аргумента попала в отфильтрованный список схем в 78,5% случаев. Данный факт позволяет говорить, что генеративные модели могут быть применены в качестве вспомогательного средства при создании аннотированных корпусов аргументов.

Все вычисления выполнялись на следующей конфигурации: AMD Ryzen 9 7950X, 192 GB RAM, NVidia GeForce RTX 4090 24 GB VRAM.

3.4. Анализ результатов

Анализ результатов классификации по отдельным схемам показал, что модель наиболее часто ошибочно предсказывает схемы «От вербальной классификации» (398 из 1845), «Существующая практика» (273), «Вербальный скользкий склон» (219) и «Негативные последствия» (129).

Обнаружены устойчивые ошибки, когда модель некорректно интерпретирует примеры одной и той же схемы как примеры других одних и тех же аргументов. Так, 46.7% от всех ошибочных предсказаний схемы «От корреляции к причине» составляют предсказания «От вербальной классификации». Популярность этой схемы может быть связана с тем, что ее бóльшая посылка сформулирована не с указанием семантического отношения, а в общем виде, как modus ponens: «Для всех M, если M имеет свойство F, то M можно отнести к имеющим свойство G». Также частотно предсказание «Существующей практики» на месте «От позитивных последствий», «Практического вывода», «От метода» и «От цели к методу».

Это можно объяснить тем, что в каждой из них в одной из посылок эксплицирована причинно-следственная связь между действием и желаемым положением дел.

Значительно снижает общий результат низкий процент распознавания одной из самых частотных схем «От примера». Предполагаемая причина в том, что при отсутствии четких маркеров экземплификации (*пример, например, в частности*) можно увидеть другие отношения, например каузальности (схема «От причины к следствию») в примере:

> *Иногда мы все же заканчиваем работу раньше срока ... В итоге есть куча поводов сорвать срок задачи несмотря на заметный запас подстраховки.*

Отсутствие значительного улучшения при изменении стратегий может быть связано как со слишком общей семантикой классов, так и с несбалансированным подходом к формулированию промптов. Например, описание класса «Гипер-, гипонимия» включает 3 ключевых слова: «пропозиция фрагмента "посылка" является частным случаем, примером пропозиции фрагмента "заключение", и в паре есть пропозиции, связанные семантическим отношением род-вид», а «Каузальность» описана с помощью одного: «пара содержит пропозиции, связанные семантическим причинно-следственным отношением». В примере:

> *Экология давно вышла за рамки традиционного толкования, вобрав в себя компоненты других наук. Эти изменения отразились на терминологии отдельных отраслей экологии, которая обогатилась терминами из смежных областей научного знания и понятиями из общей лексики –*

есть указание на оба эти типа отношений, что в целом характерно для текстов со сложным содержанием. Истинным классом здесь является «Каузальность», но модель при любых параметрах возвращает класс «Гипер-, гипонимия» – вероятно потому, что в его описании больше ключевых слов, с которыми соотносится пример.

Кроме того, результат значительно снижает большое количество случаев, когда истинной схемы не оказывается на пересечении классов, что является частотной ситуацией при решении задачи классификации по многим меткам.

Заключение

В работе представлен подход к автоматической классификации аргументов в соответствии с набором схем Д. Уолтона на основе LLM. Сложность задачи связана как с большим количеством классов, так и с возможностью отнесения схем к нескольким надклассам, что часто приводит к некорректной предварительной фильтрации схем. Рассмотрены три стратегии разработки и применения инструкций: (а) непосредственная классификация аргументов по 48 схемам; (б) использование дополнительной

систематизации схем для вычисления классификационных признаков и предварительной фильтрации множества классов; (в) использование систематизации схем в качестве иерархии для диалогового режима.

Качество классификации аргументов достигает 0.63 F_1-меры на корпусе автоматически сгенерированных англоязычных аргументов NLAS, 0.31 F_1-меры на англоязычном корпусе Araucaria и 0.15 F_1-меры на русскоязычном корпусе ArgNetSC. Наиболее эффективной стратегией оказался автоматический подбор семантически близких примеров для техники Few-Shot и диалоговая модель общения с LLM. Использование дополнительной систематизации схем аргументов стабильно улучшало показатели на 1.5% на русскоязычных данных и на 6% на англоязычных. Дополнительный эксперимент с моделью большего размера показал, что размер модели существенно влияет на результаты и применение методов дистилляции моделей является перспективным направлением исследований. Можно сделать вывод, что пока генеративные LLM могут использоваться только как вспомогательный инструмент при анализе аргументации.

Возможные методы улучшения результатов включают стандартизацию описаний классов и общую оптимизацию инструкций; реализацию нечеткой фильтрации схем; интеграцию методов на основе SFT и компактных языковых моделей.

Список литературы

[Сидорова и др., 2024] Сидорова Е.А., Кононенко И.С. Онтологический анализ приемов аргументации в научном дискурсе // Информационные и математические технологии в науке и управлении. – 2024. – № 3(35). – С. 20-32.

[Тимофеева и др., 2024] Тимофеева М.К., Ильина Д.В., Кононенко И.С. Аргументативная разметка корпуса текстов научной интернет-коммуникации: жанровый анализ и исследование типовых моделей рассуждения с помощью платформы ArgNetBank Studio // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2024. – № 22(1). – С. 27-49.

[Baimuratov et al., 2025] Baimuratov I., Karpovich A., Lisanyuk E., Prokudin D. Argument Identification for Neuro-Symbolic Dispute Resolution in Scientific Peer Review // In: Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries. Association for Computing Machinery, New York, USA, 2025, Article 6. – P. 1-9.

[Cabrio et al., 2018] Cabrio E., Villata S. Five years of argument mining: a data-driven analysis // In: IJCAI. – 2018. – Vol. 18. – P. 5427-5433. – doi: 10.24963/ijcai.2018/766.

[Chen et al., 2024] Chen G., Cheng L., Luu A.T., Bing L. Exploring the Potential of Large Language Models in Computational Argumentation // In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. – 2024. – Vol. 1. – P 2309-2330, Bangkok, Thailand. ACL. – doi: 10.18653/v1/2024.acl-long.126.

[Freeman, 2011] Freeman J.B. Argument Structure: Representation and Theory. – Springer Science & Business Media, 2011. – Vol. 18.

[Habernal et al., 2024] Habernal I., Faber D., Recchia N. et al. Mining Legal Arguments in Court Decisions // Artificial Intelligence and Law. – 2024. – No. 32. – P. 1-38.

[Kononenko et al., 2023] Kononenko I.S., Sery A.S., Shestakov V.K., Sidorova E.A., Zagorulko Y. A. An Approach to Classifying Walton's Argumentation Schemes //

In: Proc. 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE). – 2023. – P. 1540-1545.

[Lawrence et al., 2016] Lawrence J., Reed C. Argument Mining Using Argumentation Scheme Structures. In: Proceedings of Computational Models of Argument (COMMA), Potsdam, Germany, 2016. – P. 379-390.

[Lippi et al., 2016] Lippi M., Torroni P. Argument Mining from Speech: Detecting Claims in Political Debates // In: Proc. Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, 2016. – P. 2979-2985.

[Pietron et al., 2024] Pietron M., Olszowski R., Gomułka J. Efficient argument classification with compact language models and ChatGPT-4 refinements, 2024 // arXiv preprint, https://arxiv.org/pdf/2403.15473v1.

[Reed et al., 2008] Reed C., Mochales Palau R., Rowe G., Moens M.F. Language resources for studying argument // In: Proc. 6th conference on language resources and evaluation (LREC 2008), Marrakech, Morocco, 2008. – P. 91-100.

[Ruggeri et al., 2021] Ruggeri F., Lippi M., Torroni P. TreeConstrained Graph Neural Networks for Argument Mining. 2021 // ArXiv preprint. doi: 10.48550/arXiv.2110.00124.

[Ruiz-Dolz et al., 2024] Ruiz-Dolz R., Taverner J., Lawrence J., Reed C. NLAS-multi: A multilingual corpus of automatically // Data in Brief. – 2024. – Vol. 57. – doi: 10.1016/j.dib.2024.111087.

[Shaefer et al., 2021] Schaefer R., Stede M. Argument mining on twitter: A survey // IT – Information Technology. – 2021. – 63(1). – P. 45-58. – doi: 10.1515/itit-2020-0053.

[Toulmin, 2003] Toulmin S. The Uses of Argument. – Cambridge University Press, Cambridge, 2003. – 262 p.

[Wagemans, 2016] Wagemans J.H.M. Constructing a Periodic Table of Arguments // In: Argumentation, Objectivity, and Bias. Proc. of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA), 2016. – P. 1-12.

[Walker et al., 2018] Walker V., Foerster D., Ponce J., Rosen M. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment // In: Proc. of the 5th Workshop on Argument Mining, Brussels, Belgium, 2018. – P. 68-78.

[Walton et al., 2008] Walton D., Reed C., Macagno F. Argumentation schemes. – Cambridge University Press, Cambridge, 2008. – 456 p.

[Walton, 2009] Walton D. Argumentation theory: A Very Short Introduction // In: Argumentation in Artificial Intelligence. – Springer, Boston, 2009. – P. 1-22.

[Yan et al., 2021] Yan M., Lin Y.R., Litman D. Argumentatively Phony? Detecting Misinformation via Argument Mining. In**:** Proc. 1st KDD Workshop on AI-enabled Cybersecurity Analytics, 2021.

[Ying et al. 2018] Ying R., You J., Morris C. el al. Hierarchical graph representation learning with differentiable pooling // ArXiv preprint. – doi: 10.48550/arXiv.1806.08804.

[Zhang et al., 2022] Zhang G., Nulty P., Lillis D.: A decade of legal argumentation mining: Datasets and approaches // In: Proc. International Conference on Applications of Natural Language to Information Systems. Springer, Cham, 2022. – P. 240-252.

УДК 004.8
doi: 10.15622/rcai.2025.028

ПРОБЛЕМА РАЦИОНАЛИЗАЦИИ И ЧРЕЗМЕРНОГО ПОЛАГАНИЯ НА ИНСТРУМЕНТЫ XAI: АНАЛИЗ ОБЪЯСНЕНИЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

А.В. Суворова (*asuvorova@hse.ru*)

Национальный исследовательский университет
"Высшая школа экономики", Санкт-Петербург

В работе исследуется проблема чрезмерного полагания (overreliance) пользователей на результаты интерпретации моделей машинного обучения, а также способов ее решения с помощью пояснений, генерируемых большими языковыми моделями (LLM). Результаты эксперимента показали, что большинство моделей, так же как и пользователи-люди в исходном эксперименте, игнорировали аномалии или предлагали правдоподобные, но ложные объяснения, рационализируя выводы. Это указывает на риски некритичного использования LLM для интерпретации моделей машинного обучения без дополнительных механизмов валидации.

Ключевые слова: объяснимый ИИ, машинное обучение, оценивание моделей пользователями.

Введение

Все более широкое применение интеллектуальных технологий, разработка и внедрение систем машинного обучения в различные сферы жизни, включая социальную сферу, приводят к необходимости объяснения решений, принимаемых с помощью таких систем. Вопросы интерпретируемости результатов моделей, построенных с помощью алгоритмов машинного обучения, выявления факторов, оказывающих влияние на решение, все чаще возникают со стороны как общества, так и исследователей, включая вопросы по поводу их корректности, отсутствия дискриминации и т.д. Как следствие, область исследований, связанная с различными аспектами интерпретируемости, объяснимости моделей (explainable AI, interpretable machine learning), очень быстро развивается, в частности, предлагаются различные алгоритмы для объяснения уже построенных моделей.

При этом в ряде статей [Bansal et al., 2021], [Ehsan et al., 2024] поднимается проблема чрезмерного полагания (overreliance) на результаты, предоставляемые различными моделями. Чаще всего авторы отмечают, что даже добавление пояснений в информационные системы не позволяет убрать этот эффект [Bansal et al., 2021], [Buçinca et al., 2021]. Однако в [Vasconcelos et al., 2023] авторы исследуют проблему подробнее и показывают, что эффективность пояснений зависит от многих факторов, включая сложность задачи (слепая уверенность в технологии усиливается для более сложных задач), формат представления пояснений (обычное текстовое пояснение не оказывает влияния на чрезмерное полагание, а сокращенное с выделением важных элементов – снижает его). Одновременно с этим, во многих работах предлагается использовать LLM для интерпретации моделей машинного обучения: как часть объясняющего инструмента (например, для преобразования естественно-языкового запроса к модели [Slack et al., 2023]) или для генерации пояснений к результатам объясняющих моделей, например, SHAP-графикам [Hsu et al., 2024], [Singh et al., 2024].

Важно отметить, что в указанных работах не обсуждается вопрос, насколько такие дополнительные пояснения от LLM свободны от ограничений, выявленных в предыдущих исследованиях, посвященных практикам использования объясняющих инструментов пользователями. Например, в статье [Kaur et al., 2020] авторы показали, что аналитики склонны слишком полагаться на результаты применения подобных инструментов интерпретации, даже не понимая принципы их работы, особенно, если результаты представлены в «научном» формате и подкреплены ссылками на публикации.

В докладе представлены результаты эксперимента, воспроизводящего задачу из [Kaur et al., 2020], но при условии, что «участником» эксперимента является LLM. Другими словами, исследуется, будут ли дополнительные пояснения от LLM, предлагаемые разными авторами для упрощения работы с результатами инструментов XAI, воспроизводить некорректные выводы пользователей или преодолевать их.

1. Особенности пользовательской оценки объяснений моделей машинного обучения

Методы объяснимого искусственного интеллекта (XAI) позволяют получить важную информацию о поведении модели, но из-за множества доступных инструментов интерпретации результатов конкретное решение может не соответствовать оптимальным требованиям целевых пользователей. Исследования показывают, что цели объяснимости и даже значение слова «быть объяснимым» различаются в зависимости от предметной области, опыта и роли конечного пользователя в рабочем процессе [Arrieta et al., 2020].

1.1. Типы пользователей в решениях, направленных на интерпретацию моделей

Авторы [Hong et al., 2020] выделяют три ключевые группы участников, вовлечённых в создание и применение моделей:

- создатели моделей (model builders) – специалисты по анализу данных и машинному обучению, которые разрабатывают модели;
- тестировщики и критики моделей (model breakers) – эксперты в предметной области, менеджеры продуктов, юристы, которые оценивают и проверяют модели;
- пользователи моделей (model consumers) – конечные пользователи, которые взаимодействуют с результатами работы моделей.

Большинство методов и инструментов объяснения моделей машинного обучения ориентированы либо на создателей (для отладки и улучшения моделей), либо на конечных пользователей (для прозрачности и интерпретируемости). В исследованиях [Hong et al., 2020], [Langer et al., 2021]; [Tomsett et al., 2018] показано, что конечным пользователям, не экспертам в предметной области или ML-специалистам, часто не нужны объяснения моделей, основными «потребителями» объяснений в моделях являются внутренние пользователи, включенные в процесс обсуждения, разработки, внедрения модели. Более того, у ML-специалистов тоже мало стимулов использовать модели объяснения [Langer et al., 2021], [Zakharova et al., 2021], т.к. их внедрение требует дополнительного времени и ресурсов, а оценивание адекватности модели чаще всего проводится самим разработчиком на основе формальных метрик качества, а внешнее оценивание базируется на доверии к самому разработчику и метриках ее эффективности (технических или бизнес-метриках), а не на результатах исследования модели. Таким образом, для более объективного оценивания нужен инструмент, который использовал бы существующие алгоритмы объяснения (и был соответственно разработан ML-специалистом), но при этом упрощал исследование модели для не-ML-специалиста. При этом потребности тестировщиков часто остаются без должного внимания. Основная сложность в разработке систем для этой группы заключается в том, что они, как правило, не обладают глубокими техническими знаниями о методах машинного обучения и не могут в полной мере использовать возможности интерпретируемого машинного обучения (IML).

1.2. Оценивание моделей интерпретации

Для выводов об эффективности того или иного решения, направленного на интерпретацию результатов моделей машинного обучения, независимо от того, какой тип пользователей является для этого решения целевым, необходимо иметь возможность оценить качество полученных объяснений.

В широко цитируемой классификации способов оценивания объяснений [Doshi-Velez et al., 2017] выделены три категории подходов к оценке:

- оценка итоговых инструментов пользователями. Этот вид оценки основан на экспериментах, в которых конечные пользователи решают реальные задачи, используя предложенные инструменты для принятия решений.
- упрощенная оценка на людях. Этот тип оценки также требует экспериментов с людьми, но, поскольку конечные пользователи (специалисты в предметной области, например, врачи) обычно труднодоступны, а их время дорого, то используется упрощенное приближение к реальной ситуации с непрофессионалами.
- оценка функциональности. Этот тип оценки не требует экспериментов с вовлечением людей и использует метрики, основанные на формальном определении интерпретируемости, которые можно оценить математически или с помощью моделирования [Agarwal et al., 2022].

Учитывая, что единого общепринятого определения объяснимости не существует, довольно часто она описывается довольно размытыми формулировками как «способность объяснить или представить в понятных терминах человеку» [Doshi-Velez et al., 2017]. При этом так как цель применения методов интерпретации результатов моделей в большинстве случаев связана с необходимостью специалисту принять решение, можно ли использовать анализируемую модель машинного обучения, нет ли в ней смещений и искажений, то оценивание именно со стороны конечных пользователей является существенным этапом проектирования системы. Как следствие, значительная часть исследований по оцениванию объяснений ориентирована на пользовательский опыт. В этих исследованиях описываются конкретные эксперименты, измеряющие удовлетворенность [Hoffman et al., 2018], доверие [Drozdal et al., 2020] или способность принимать решения на основе пояснений результатов модели [Kaur et al., 2020].

2. Искажения пользовательской оценки объяснений моделей

Несмотря на широкое распространение именно пользовательских оценок объяснений, корректность применения инструментов интерпретации конечными пользователями и способы предотвращения ошибочного использования обсуждаются не часто [Kaur et al., 2024]. Одна из таких работ – статья [Kaur et al., 2020], в которой изучались практики использования методов интерпретируемого машинного обучения аналитиками данных для изучения ML-моделей, в частности, для поиска ошибок в модели и данных. Именно эта работа взята за основу для исследования, будут ли дополнительные пояснения от LLM воспроизводить некорректные выводы пользователей или преодолевать их.

2.1. Описание исходного исследования

В рамках этого доклада рассмотрена только одна из выявленных в [Kaur et al., 2020] проблем – чрезмерное полагание на результатаы объяснения моделей и рационализация необычных паттернов, поэтому в последующем описании приведены только те элементы процедуры исследования, которые относятся к выявлению указанной проблемы.

Для определения того, смогут ли пользователи найти несоответствия в предложенной им модели, в набор данных Adult Income dataset были внесены различные искажения, включая замену для 10% наблюдений с высоким доходом реальных значений возраста на среднее значение по выборке – 38 лет. Затем на модифицированных данных была построена модель классификации с использованием алгоритма LightGBM для предсказания высокого или низкого дохода. К итоговой модели были применены инструменты интерпретации ML-моделей, включая SHAP, в частности, был построен график зависимости SHAP-значений от возраста, показывающий для каждого наблюдения (точка на графике) влияние возраста на предсказание дохода – чем дальше от 0, тем более влиятелен возраст для предсказание. Итоговый график показан на рис. 1. Стоит отметить, что цвет точки соответствует значению семейного статуса для наблюдения, но это не существенно для рассматриваемой задачи.

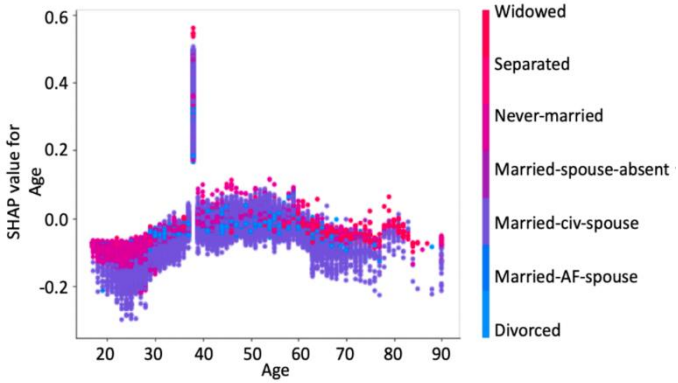


Рис. 1. SHAP-значения модели предсказания дохода относительно возраста

Участникам предоставлялось описание контекста – какие данные использованы, какая модель была построена, какое у нее качество, а также описание логики метода SHAP и полученные графики. Затем задавались вопросы на понимание модели, в том числе вопрос про возраст – «Как возраст влияет на доход (How does the feature Age affect output Income)?». Как отмечалось выше, часть участников в ответ на этот вопрос рациона-

лизировала выброс на значении 38 лет, объясняя его какими-то закономерностями в реальной жизни, а не ошибкой модели («Возраст 38 лет, похоже, имеет наибольшее положительное влияние на доход, исходя из графика. Не уверен, почему, но объяснение ясно показывает это... имеет смысл» [Kaur et al., 2020]).

2.2. Генерация пояснений к графикам с помощью LLM

Многие авторы [Omar et al., 2025] отмечают случаи некорректных выводов конечных пользователей по предоставляемым им SHAP-графикам, часто связывая это со сложностью подобных графиков. Как одно из возможных решений предлагается [Hsu et al., 2024] использовать инструменты на основе LLM для генерации дополнительных пояснений, что потенциально может предотвратить некорректные выводы. Для проверки этой идеи были сгенерированы пояснения к графикам из исходного эксперимента с конечными пользователями (см. раздел 2.1).

Для проведения эксперимента были использованы четыре модели – GigaChat, GPT 3.5, Sonar и DeepSeek. Запрос в каждую из них включал SHAP-график (рис. 1), контекст исходного эксперимента (описание данных и построенной модели) на английском языке, т.к. формулировки были напрямую скопированы из протокола исследования [Kaur et al., 2020], и целевой вопрос «Как возраст влияет на доход (How does the feature Age affect output Income)?». После ответа был задан уточняющий вопрос о возможном объяснении пика в возрасте 38 лет на графике. Запрос к каждой модели был повторен 10 раз. Также была проведена вторая итерация эксперимента, практически полностью повторяющая первую, за одним исключением – контекст задачи был расширен, передан в виде файла из исходного исследования, включающего не только описание данных и модели, но и объяснение принципов работы SHAP.

Результаты трех из четырех рассмотренных моделей (GigaChat, GPT 3.5, DeepSeek) оказались очень схожими – на первую часть задания был выдан очень обобщенный ответ примерно следующего содержания: «возраст оказывает нелинейное влияние на прогнозируемый доход. молодой возраст (20-40 лет): положительное влияние на доход; средний возраст (40-60 лет): нейтральное или слабо отрицательное влияние; пожилой возраст (60+ лет): отрицательное влияние на доход». Этот ответ часто дополнялся формальными описаниями метода: что отображается по оси x, чему соответствует ось y.

Ни одна из этих трех моделей не акцентировала внимание на пике в возрасте 38 лет при первоначальном запросе, в отличие от Sonar, где в дополнение к выводу примерно той же структуры про нелинейный эффект сразу выделялся выброс на значении 38 лет и приводилось возможное объяснение этому: «В этом конкретном возрасте (38 лет) признак "Возраст" оказывает сильное положительное влияние на прогнозирование

высокого дохода для многих людей. Это может быть артефакт данных или когортный эффект (например, случайная концентрация высокооплачиваемых людей этого возраста в выборке)».

Более того, при запросе уточнений в модели Sonar возможные искажения оставались основным объяснением (табл. 1).

Таблица 1

Возможная причина	**Описание**	**Вероятность**
Артефакт данных/ошибка ввода	Множество записей для возраста 38 лет из-за округления или ошибок ввода данных	высокая
Переобучение модели	Модель уловила специфическую закономерность, характерную только для 38-летних	средняя
Взаимодействие признаков	Сильное влияние других факторов, не отображённых на графике	средняя
Когортный эффект	Реальная экономическая или демографическая причина высоких доходов в возрасте 38 лет	низкая / средняя

Для остальных трех моделей дополнительный запрос о возможном объяснении пика в 38 лет приводил к рационализации видимой закономерности, повторяя поведение участников эксперимента [Kaur et al., 2020]. Среди возможных причин были указаны многие персональные, экономические и социальные факторы: стабильное карьерное положение, возможность иметь двойной семейный доход, накопленный профессиональный опыт, завершение всех уровней образования, спрос на рынке труда, особенности исторических данных. В некоторых результатах упоминалось возможное взаимодействие с другими переменными, но ни разу не появилась ошибка ввода.

Расширенное пояснение принципов SHAP в контексте на второй итерации не привело к улучшению выводов, обобщенные результаты были похожи.

Заключение

Проведенное исследование показало, что большие языковые модели (LLM) повторяют поведение пользователей, демонстрируя склонность к чрезмерному полеганию на результаты интерпретации моделей машинно-

го обучения. Как и в оригинальном эксперименте, в котором люди рационализировали аномалии в данных, три из четырех протестированных моделей либо игнорировали явный выброс на SHAP-графике, либо предлагали ему правдоподобные, но ошибочные объяснения.

Таким образом, несмотря на потенциал LLM в упрощении взаимодействия с ML-моделями, их прямое использование для объяснения решений требует осторожности. Чтобы избежать усиления эффекта слепого полагания на представленный инструментом результат, можно направить исследования на создание гибридных систем, сочетающих генерацию естественно-языковых пояснений с алгоритмами валидации и критического анализа.

Список литературы

[Agarwal et al., 2022] Agarwal C. et al. Openxai: Towards a transparent evaluation of model explanations // In: Advances in neural information processing systems. – 2022. – Vol. 35. – P. 15784-15799.

[Arrieta et al., 2020] Arrieta A.B. et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. – 2020. – Vol. 58. – P. 82-115.

[Bansal et al., 2021] Bansal G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance // In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. – 2021. – P. 1-16.

[Buçinca et al., 2021] Buçinca Z., Malaya M. B., Gajos K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI // in AI-assisted decision-making. In: Proceedings of the ACM on Human-computer Interaction. – 2021. – Vol. 5 (CSCW1). – P. 1-21.

[Drozdal et al., 2020] Drozdal J., Weisz J., Wang D., et al. Trust in automl: Exploring information needs for establishing trust in automated machine learning systems // In: Proceedings of the 25th International Conference on Intelligent User Interfaces. – 2020. – P. 297-307.

[Ehsan et al., 2024] Ehsan U. et al. Human-centered explainable AI (HCXAI): Reloading explainability in the era of large language models (LLMs) // In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. – 2024. – P. 1-6.

[Hoffman et al., 2018] Hoffman R.R., Mueller S.T., Klein G., Litman J. Metrics for explainable AI: Challenges and prospects // In: arXiv preprint. – 2018. – URL: arXiv:1812.04608.

[Hong et al., 2020] Hong S.R., Hullman J., Bertini E. Human factors in model interpretability: Industry practices, challenges, and needs // In: Proceedings of the ACM on Human-Computer Interaction. – 2020. – Vol. 4 (CSCW1). – P. 1-26.

[Hsu et al., 2024] Hsu C.C., Wu I.Z., Liu S.M. Decoding AI Complexity: SHAP Textual Explanations via LLM for Improved Model Transparency // In: 2024 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan). – IEEE, 2024. – P. 197-198.

[Kaur et al., 2020] Kaur H. et al. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning // In: Proceedings of the 2020 CHI conference on human factors in computing systems, Honolulu, USA, 2020. – P. 1-14. – doi: 10.1145/3313831.3376219.

[Kaur et al., 2024] Kaur H. et al. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning // In: Proceedings of the ACM on Human-Computer Interaction. – 2024. – Vol. 8 (CSCW1). – P. 1-34.

[Langer et al., 2021] Langer M. et al. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence. – 2021. – Vol. 296. – 103473.

[Omar et al., 2025] Omar Z. A. et al. Beyond Accuracy, SHAP, and Anchors--On the difficulty of designing effective end-user explanations // arXiv preprint arXiv:2503.15512. – 2025.

[Singh et al., 2024] Singh C. et al. Rethinking interpretability in the era of large language models // In: arXiv preprint arXiv:2402.01761. – 2024.

[Slack et al., 2023] Slack D. et al. Explaining machine learning models with interactive natural language conversations using TalkToModel // In: Nature Machine Intelligence. – 2023. – Vol. 5(8). – P. 873-883.

[Tomsett et al., 2018] Tomsett R., Braines D., Harborne D., Preece A., Chakraborty S. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems // In: arXiv preprint arXiv:1806.07552. – 2018.

[Vasconcelos et al., 2023] Vasconcelos H. et al. Explanations can reduce overreliance on ai systems during decision-making // In: Proceedings of the ACM on Human-Computer Interaction. – 2023. – Vol. 7 (CSCW1). – P. 1-38.

[Zakharova et al., 2021] Zakharova V., Suvorova A. Social Aspects of Machine Learning Model Evaluation: Model Interpretation and Justification from ML-practitioners' Perspective // In: 24th International Conference "Internet and Modern Society". –2021. – P. 230-234.

УДК 004.89
doi: 10.15622/rcai.2025.029

ЗАДАЧИ АНАЛИЗА ТОНАЛЬНОСТИ В НОВОСТНЫХ ТЕКСТАХ

С.О. Уразов (*urazov.msu@gmail.com*)
Н.В. Лукашевич (*louk_nat@mail.ru*)

Московский государственный университет
им. М.В. Ломоносова, Москва

В этой статье описывается исследование применения больших языковых моделей (LLM) в различных постановках задачи анализа тональности, как например: поиск мнений конкретного лица или мнений о конкретном лице; классификация (на негативную, нейтральную или позитивную) тональности мнения по отношению к цели; и выделение отношения между двумя сущностями в тексте. Исследования были проведены на датасете русских новостных текстов соревнования RuOpinionNE-2024, с использованием нейросетевых моделей типа BERT и Ruadapt-Qwen2.5. Была проведена серия экспериментов с использованием техники обучения моделей и подсказок (промптов).

Ключевые слова: Анализ тональности, Большие языковые модели, Промпт инжиниринг, Тонкая настройка с помощью техники LoRA.

Введение

Анализ тональности текста является активно развивающейся областью автоматической обработки естественного языка. Целью такого анализа обычно является выявление мнений относительно сущностей и определение общей тональности текста. Например, такой анализ лежит в основе политических исследований в рамках определения настроений в обществе и для отслеживания отношения к конкретным личностям или компаниям [Smetanin, 2020; Семина, 2020].

Анализ тональности в новостных текстах может использоваться для решения разных конкретных задач. Иногда необходимо собрать мнения по разным вопросам какого-то лица (например, недавно избранного на высокую должность). Или, наоборот, важны разные мнения по отношению к этому лицу. В третьей постановке могут быть важны отношения между людьми, организациями и т.д.

В данной статье рассматриваются различные постановки задач анализа тональности в новостных текстах. Для решения выделенных задач используются как энкодерные модели, такие как BERT [BERT, 2019], так и генеративные (декодерные) модели, например Ruadapt-Qwen2.5[1] [Qwen2.5, 2025], которые адаптированы под русский язык. Рассмотрены различные способы составления запросов для моделей (промпт-инжиниринг) и влияние подбора различных промптов во время обучения моделей.

Для проведения исследований использовался датасет новостных текстов на русском языке соревнования RuOpinionNE-2024 [RuOpinionNE-2024, 2025], в котором выделены кортежи мнений, включающие источник мнения, объект мнения, тональность мнения и оценочное выражение.

Вклад этой статьи следующий:

- Представлены различные постановки задач анализа тональности новостных текстов;
- Проведены эксперименты с использованием различных подходов к работе с моделями и к технике построения подсказок;
- Проанализированы зависимости, тенденции и различия в результатах использования различных подходов.

Дальнейшая структура статьи следующая. В главе 1 рассматриваются современные подходы, проводится краткий обзор работ по теме. В главе 2 детально описываются постановки задач, которые исследуются в этой статье. В главе 3 дается обзор датасета RuOpinionNE-24. В главе 4 описывается преобразование датасета под особенности каждой формулировки задачи. В главе 5 описываются и обсуждаются полученные результаты. В заключении подводится итог исследования.

1. Обзор работ по теме

Первые методы в области изучения анализа тональности были основаны на составлении правил, словарей, специальном представлении текста (мешок слов, опорные вектора и др.), а также на основе рекуррентных и сверточных нейронных сетей [Effective lstms, 2015]. Серьезный скачок в качестве автоматического анализа произошел при использовании методов, использующих трансформеры, применение которых во многих актуальных задачах дает высокие результаты [BERT, 2019; Golubev et al., 2020; Structured sentiment analysis, 2021; Semeval, 2022].

В настоящее время активно исследуется применение больших генеративных моделей, которые имеют возможности к более глубокому пониманию текста [GPT-4o, 2024] и способности к составлению сложного,

[1] https://huggingface.co/collections/RefalMachine/ruadapt-qwen-25-67124a497e75205228348919.

структурированного ответа. Были исследованы возможности больших языковых моделей в формате zero-shot и few-shot [Sentiment analysis, 2023] в различных постановках задачи анализа тональности. Это исследование показало, что применение серии моделей GPT-3.5 сравнимо по эффективности с применением тонко настроенной на соответствующую задачу Flan-UL2, несмотря на большую разницу в размерах моделей.

В соревновании RuOpinionNE-24 [RuOpinionNE-24, 2025] лучшие результаты в задаче выделения кортежей мнения среди русских новостных текстов были получены с использованием модели LlaMa-3.3-70B, тонко настроенной с помощью техники QLoRA [LoRA, 2021]. Кортежи мнений включали в себя источник мнения, цель мнения, тональность и выражение, которым описывается мнение источника к цели, и модель был настроена на то, чтобы выделять такие кортежи целиком [Vatolin et al., 2025]. Среди других участников соревнования высокие результатов были получены тоже с применением больших генеративных моделей при помощи тонкой настройки и few-shot формата [Rossyaykin, 2025].

2. Описание постановок задач

2.1. Задача извлечения мнений некоторого лица (источника)

В этой задаче необходимо находить фрагменты текстов (предложения), включающие явно или неявно выраженные мнения, которые были высказаны конкретным лицом, что может помочь составить картину взглядов этого лица, отследить их изменения. Данная задача представляет собой задачу бинарной классификации: есть в предложении мнение или нет.

2.2. Задача извлечения мнений по отношению к некоторой сущности (лицу, организации, событию).

В этой задаче необходимо находить фрагменты текстов (предложения), включающие явно или неявно выраженные мнения разных источников по отношению к заданному лицу. Данная задача также представляет собой задачу бинарной классификации.

Обе задачи возникают, например, при назначении на высокие должности новых лиц, когда требуется представить портрет их убеждений и взаимоотношений с другими людьми.

2.3. Классификация тональности отношения к объекту мнения

Данная задача заключается в определении того, какая именно тональность (нейтральная, позитивная или негативная) обращена к конкретной сущности в тексте, являющейся целью мнения.

Эту задачу также можно решить двумя способами:

- Подойти к задаче напрямую: провести классификацию на три класса;
- Используя результаты первой и второй задач, провести бинарную классификацию тех текстов, в которых есть полярное мнение.

2.4. Классификация отношений между сущностями

Четвертая задача – определить тональность взаимоотношения между сущностями, выделяя в них источник мнения, цель мнения и полярность мнения (негативное, нейтральное или позитивное).

3. Датасет RuOpinionNE-2024 и специализированные датасеты для задач

Датасет RuOpinionNE-2024 представляет собой набор размеченных отрывков новостных текстов, в которых выделены кортежи мнений (источник мнения, цель мнения, тональность мнения, выражение мнения). Источник мнения может отсутствовать ("NULL") или быть автором текста ("AUTHOR").

Мнения могут выражены как явно, так и имплицитно. Например, по предложению «Apple и Samsung нарушали патенты друг друга» должны быть извлечены кортежи:

(Apple, Samsung, NEG, “нарушали патенты друг друга”) и

(Samsung, Apple, NEG,” нарушали патенты друг друга”).

Датасет представлен в виде совокупности предложений, к каждому из которых приписано множество высказанных в них мнений. Предложение может не содержать никаких мнений, и тогда множество будет пустое. Источником и объектом мнения в RuOpinion-2024 могут быть сущности одного из следующих типов: PERSON, ORGANIZATION, PROFESSION, NATIONALITY, COUNTRY, REGION, CITY, IDEOLOGY.

Исходный датасет используется для формирования датасетов для всех конкретных задач. Например, в датасете для **первой задачи** сущностям заданных типов ставятся в соответствие значения 1 (источник мнения) или 0 (не источник мнения). В датасете **второй задачи** сущностям заданных типов сопоставляются значения 1 (объект мнения) или 0 (не объект мнения). **Для третьей задачи** каждой сущность ставится в соответствие одна из трех меток: POS, NEG, NEU.

Датасет четвертой задачи состоит из триплетов: две сущности и тональность отношения между ними. Соответственно, если пара источник-цель присутствовали в кортеже оригинального датасета, то ей ставилась в соответствие метка тональности и оценочное выражение. В противном случае, если такая пара не связана мнением, то ей сопоставлялся кортеж с нейтральной тональностью и пустым текстом (например, {NEU, [ ]}) [Willard et al., 2023].

4. Модели

В качестве методов для решения представленных задач использовались модели на основе архитектуры трансформер. В задачах 1-4 использовалась модель на основе многоязычного энкодера трансформера XLM-

RoBERTa-large. Данная модель может принимать на вход два фрагмента текста, разделенные служебным токеном [SEP]. Это позволяет использовать в модели специальный запрос (промпт), который формулирует задачу. Целевое предложение задается как первый фрагмент текста, а промпт как второй фрагмент текста после токена [SEP]. Для тонкой настройки все слои размораживались. Использовалась косинусная кривая обучения с разогревом в 3 эпохи. Процесс тонкой настройки для задач 1-3 проводился на 22 эпохах, для задачи 4 – на 8 эпохах. Размер батча – 16.

В задаче 4 также проводились эксперименты с генеративной моделью Ruadapt-Qwen2.5, которая представляет собой русифицированную версию модели Qwen2.5.

Для выделения именованных сущностей использовалась модель BINDER [Optimizing Bi-Encoder, 2022], обученная извлечению именованных сущностей на датасете NEREL[2].

5. Результаты экспериментов и обсуждение

5.1. Первая задача определения источника мнения

Для задачи определения источника мнения использовались "запросы" при идентификации источников в тексте.

Среди промптов были использованы следующие:

1. (имя сущности) – источник тональности;
2. (имя сущности) – выражает мнение;
3. (имя сущности) позитивно или негативно относится к сущности в тексте;
4. (имя сущности) – источник позитивного или негативного мнения;
5. *@(имя сущности)@* – источник позитивного или негативного мнения к объекту текста.

Результаты усреднялись по трем экспериментам. В каждом эксперименте проводилась тонкая настройка с нуля и тестирование модели. Пиковое значение скорости обучения: 1.5e-6. Результаты по разным промптам представлены в табл. 1.

Таблица 1

Промпт	Precision	Recall	F1 score
1	0.73 +-0.02	0.86 +-0.01	0.78 +-0.01
2	**0.77 +-0.01**	**0.84 +-0.01**	**0.80 +-0.01**
3	0.74 +-0.01	0.84 +-0.01	0.77 +-0.01
4	0.75 +-0.01	0.84 +-0.01	0.78 +-0.01
5	0.76 +-0.01	0.85 +-0.01	0.80-0.01

[2] https://github.com/fulstock/binder.

5.2. Вторая задача определения объекта мнения

Для извлечения объекта мнения использовались следующие промпты:

1. (имя сущности) – это цель тональности;
2. (имя сущности) – это цель мнения;
3. В тексте выражается позитивное или негативное мнение к сущности (имя сущности).

Пиковое значение скорости обучения: 2e-6. Результаты были получены следующие (см. табл. 2):

Таблица 2

Промпт	Precision	Recall	F1 score
1	**0.77 +-0.02**	**0.82 +-0.02**	**0.80 +-0.01**
2	0.76 +-0.02	0.80 +-0.02	0.78 +-0.01
3	0.77 +-0.01	0.81 +-0.01	0.79 +-0.01

Таким образом, видно, что в обеих подзадачах бинарной классификации определения источника мнения или объекта мнения качество достигает 80% по F-мере. Наблюдалось, что применении модели в запросе на конкретную сущность, каждая модель склонна отвечать правильно, но иногда путались роли источника и цели.

5.3. Классификация тональности мнения по отношению к сущности

Результаты работы классификации на три класса и на два класса (результаты усреднялись по трем экспериментам, пиковое значение скорости обучения: 3e-6) представлены в табл. 3:

Таблица 3

Классификация	Precision	Recall	F1 score
3 класса	0.65 +-0.01	0.70 +-0.01	0.68 +-0.01
2 задача + 2 класса	0.65 +-0.01	0.70 +-0.01	0.67 +-0.01

В работе над бинарной классификацией, для оценки работы были смешаны результаты поиска целей мнения из предыдущей задачи и совмещены с работой классификатора, чтобы провести оценку общей классификации на три класса.

В экспериментах наблюдалось, что в рамках текущего датасета работа модели слабо зависела от предлагаемого ей промпта. При этом лучшим оказался промпт вида: «Тональность (имя сущности) равна».

Результаты трехклассового и бинарного подходов сравнительно схожи и стабильны, что показывает равнозначность рассмотрения разбиения полной задачи на подзадачи. Это важный результат для перехода к рассмотрению третьей задачи, в которой сильная опора идет на результаты первой задачи.

5.4. Задача извлечения тональности отношений между сущностями

Для установления зависимости между источником мнения и целью мнения был использован пакет извлечения отношений OpenNRE [OpenNRE, 2019]. В нашем случае в качестве отношения задается тональность мнения источника к цели.

Аналогично предыдущим задачам, были добавлены следующие подсказки к основному тексту примера, чтобы помочь модели фокусироваться:

1. Какое отношение высказывает источник (имя источника) по отношению к цели (имя цели).
2. Какую тональность источник (имя источника) выражает к (имя цели);
3. Какую тональность источник (имя источника) явно или неявно выражает относительно цели (имя цели).
4. Какую тональность источник *@(имя источника)@* выражает к >>>(имя цели)<<<.
5. Отношение источника (имя источника) к цели (имя цели).

Пиковое значение скорости обучения: 9.5e-6. Результаты по разным запросам представлены в табл. 4:

Таблица 4

Модель	Precision	Recall	F1 score
1	0.83 +-0.01	0.62 +-0.01	0.71 +-0.01
2	0.86 +-0.01	0.63 +-0.01	0.73 +-0.01
3	**0.87 +-0.01**	**0.67 +-0.01**	**0.76 +-0.01**
4	0.87 +-0.01	0.67 +-0.01	0.76 +-0.01
5	0.84 +-0.01	0.62 +-0.01	0.71 +-0.01

Оценка работы проводилась по тем примерам, в которых выделен кортеж, соответствующий кортежам оригинального датасета.

Выделение сущностей в самой подсказке (промпт 4) не повлияло на результаты.

Для проведения экспериментов с декодер-моделями была использована модель Ruadapt/Qwen2.5-7B-ext-u48-instruct. Для тонкой настройки этой модели был использован подход LoRA с квантизацией 8bit. Lora rank: 8. Lora alpha: 32. Lora dropout: 0.1. Шаги накопления градиентов: 4. Пиковое значение скорости обучения: 6e-5.

Во многих примерах модель не была уверена в своем ответе, поэтому могла отвечать по-разному для одного и того же примера. Поэтому был использован прием, описанный в работе участников соревнования RuOpinionNE [Vatolin, 2025] – Most Common N, в котором среди 10 выданных тональностей бралась самая часто выделяемая тональность из них.

Сначала были проведены эксперименты со следующими системными промптами, используя пользовательский промпт 3 из предыдущей задачи:

1. Ты – полезный и умный инструмент для анализа тональности текста.

2. Ты – полезный и умный инструмент для анализа тональности текста. Твоя роль состоит в выявлении отношения в виде структуры {Polarity, Expression}, где Polarity это NEG, POS или NEU, а Expression это часть текста запроса.
3. Ты Qwen, созданный Alibaba Cloud. Ты – полезный помощник.
4. Ты – профессиональный журналист. Твоя роль состоит в выявлении отношения в виде структуры {Polarity, Expression}, где Polarity это NEG, POS или NEU, а Expression это часть текста запроса.
5. Ты – профессиональный психолог. Твоя роль состоит в выявлении отношения в виде структуры {Polarity, Expression}, где Polarity это NEG, POS или NEU, а Expression это часть текста запроса.

Были получены следующие результаты (табл. 5):

Таблица 5

Промпт	Precision	Recall	F1 score
1	**0.98 +-0.01**	**0.58 +-0.03**	**0.72 +-0.02**
2	0.98 +-0.01	0.56 +-0.03	0.70 +-0.02
3	0.97 +-0.01	0.56 +-0.02	0.71 +-0.03
4	0.98 +-0.01	0.58 +-0.03	0.72 +-0.02
5	0.97 +-0.01	0.58 +-0.04	0.72 +-0.02

Здесь наблюдается точность, практически равная 1 (precision ≈ 0.982), но при этом полнота оказывается значительно ниже (recall ≈ 0.574). Это говорит о том, что модель не улавливает многие зависимости между сущностями, однако если улавливает, то улавливает правильно.

Оказался интересным тот факт, что лучше всех показал себя достаточно обобщенный системный промпт 1. При этом он показывает результаты немного лучше, чем часто используемый промпт 3.

Эксперименты показали, что при уточнении и расширении промпта 1 (см. промпт 2) наблюдается следующее. Функция потерь на первых шагах тонкой настройки промпта 1 более чем в два раза выше, однако после нескольких шагов ситуация менялась, и функция потерь промпта 1 в среднем становилась меньше, чем у промпта 2 (см. рис. 1).

С выбранным лучшим системным промптом были проведены эксперименты со следующими пользовательскими промптами:

1. Какую тональность выражает (имя источника) явно или неявно выражает относительно объекта (имя цели)?;
2. Какое отношение источник (имя источника) явно или неявно выражает относительно объекта (имя цели)? Ответь в виде структуры {(Отношение), (Выражение)};
3. Какое отношение источник (имя источника) явно или неявно выражает относительно объекта (имя цели)? Ответь в виде структуры {(Отношение), (Выражение)}, где (Отношение) это NEG, POS или NEU, а (Выражение) это копия части текста запроса;

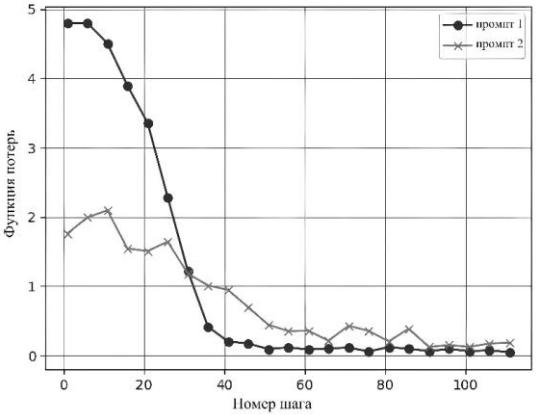


Рис. 1. Графики функций зависимости потерь от номера шага промпта 1 и 2

4. Какую тональность выражает (имя источника) явно или неявно выражает относительно объекта (имя цели)? Ответь в виде структуры {(Тональность), (Выражение)}, где (Тональность) это NEG, POS или NEU, а (Выражение) это копия части текста запроса;
5. С точки зрения журналистики, какую тональность выражает (имя источника) явно или неявно выражает относительно объекта (имя цели)? Ответь в виде структуры {(Тональность), (Выражение)};
6. С точки зрения психологии, какую тональность выражает (имя источника) явно или неявно выражает относительно объекта (имя цели)? Ответь в виде структуры {(Тональность), (Выражение)}.

На пользовательских промптах были получены следующие результаты (табл. 6):

Таблица 6

Промпт	Precision	Recall	F1 score
1	0.98 +-0.01	0.58 +-0.03	0.72 +-0.02
2	0.97 +-0.01	0.56 +-0.01	0.70 +-0.01
3	0.98 +-0.01	0.55 +-0.01	0.71 +-0.01
4	0.98 +-0.01	0.58 +-0.03	0.73 +-0.02
5	**0.97 +-0.01**	**0.59 +-0.04**	**0.73 +-0.04**
7	0.97 +-0.01	0.58 +-0.04	0.73 +-0.04

Таким образом, исходя из полученных результатов, лучшие результаты при тонкой настройке достигаются при использовании краткого лаконичного системного и более подробного пользовательского промптов.

При использовании большей модели той же архитектуры (Ruadapt/Qwen2.5-14B-instruct) результаты становятся значительно выше: Recall возрастает до ≈ 0.73, а общий F1 score – до ≈ 0.84, что уже превосходит возможности классификации моделей, основанных на моделях-энкодерах.

Заключение

В этой статье были исследованы способы разделения задачи анализа отношения между сущностями на подзадачи в рамках датасета русских новостных текстов соревнования RuOpinionNE-24. Каждая подзадача была проанализирована относительно использования разных моделей и соответствующих промптов.

Было изучено взаимодействие моделей в рамках рассмотренных задач. Рассмотрены различные подходы к тонкой настройке моделей на конкретных задачах.

По результатам исследований каждой подзадачи была получена система составления оценки отношения между сущностями в тексте, которая позволяет достаточно эффективно (F1 score ≈ 0.84 с использованием модели Ruadapt/Qwen2.5-14B-instruct) извлекать тональность отношения между сущностями в предложенном тексте.

Благодарности. Исследование выполнено в рамках государственного задания МГУ имени М. В. Ломоносова.

Список литературы

[Семина, 2020] Семина Т.А. Анализ тональности текста: современные подходы и существующие проблемы // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журнал. – 2020. – Вып. 4. – С. 47-64.

[Structured sentiment analysis, 2021] Barnes J. [et al]. Structured sentiment analysis as dependency graph parsing // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. – Vol. 1. –P. 3387-3402.

[Semeval, 2022] Barnes J. [et al]. Semeval 2022 task 10: Structured sentiment analysis / // Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). – 2022. – P. 1280-1295.

[BERT, 2019] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. – 2019. – URL: https://arxiv.org/abs/1810.04805 (дата обращения: 14.06.2025).

[Golubev et al., 2020] Golubev A., Loukachevitch N. Improving results on Russian sentiment datasets // Conference on artificial intelligence and natural language. – 2020. – P. 109-121. – URL: https://arxiv.org/abs/2007.14310 (дата обращения: 14.06.2025).

[OpenNRE, 2019] Han X. [et al]. OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. – 2019. – URL: https://openreview.net/pdf?id=9EAQVEINuum (дата обращения: 14.06.2025).

[LoRA, 2021] Hu E. J. [et al]. LoRA: Low-Rank Adaptation of Large Language Models. – 2021. – URL: https://arxiv.org/abs/2106.09685 (дата обращения: 14.06.2025).

[GPT-4o, 2024] Hurst A. [et al.]. GPT-4o System Card. – 2024. – URL: https://arxiv.org/pdf/2410.21276 (дата обращения: 14.06.2025).

[RuOpinionNE-2024, 2024] Loukachevitch N.V. [et al]. RuOpinionNE-2024: Extraction of Opinion Tuples from Russian News Texts. – 2025. – URL: https://arxiv.org/html/2504.06947v1 (дата обращения: 14.06.2025).

[Qwen, 2025] Qwen [et al]. Qwen2.5 Technical Report. – 2025. – URL: https://arxiv.org/abs/2412.15115 (дата обращения: 14.06.2025).

[Rossyaykin, 2025] Rossyaykin P. Structured sentiment analysis using few-shot prompting of an ensemble of LLMs. – 2025.

[Smetanin, 2020] Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives.IEEE Access. – 2020. – Vol. 8. – P. 110693-110719. – URL: https://api.semanticscholar.org/CorpusID:220078379.

[Effective lstms, 2016] Tang D., Qin B., Feng X., Liu T. Effective lstms for target-dependent senti ment classification. // International Conference on Computational Linguistics. – 2016. – URL: https://arxiv.org/abs/1512.01100 (дата обращения: 14.06.2025).

[Vatolin, 2025] Vatolin A. Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024. – 2025.

[Willard et al., 2023] Willard B. T., Louf R. Efficient guided generation for large language models. – 2023. – URL: https://arxiv.org/abs/2307.09702 (дата обращения: 14.06.2025).

[Optimizing Bi-Encoder, 2022] Zhang S., Cheng H., Gao J., Poon H. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning. – 2022. – URL: https://arxiv.org/pdf/2307.09702 (дата обращения: 14.06.2025).

[Sentiment analysis, 2023] Zhang W., Deng Y., Liu B., Pan S., Bing L. Sentiment analysis in the era of large language models: A reality check. Findings of the Association for Computational Linguistics: NAACL 2024. – P. 3881-3906.

УДК 004.89
doi: 10.15622/rcai.2025.030

ПРИМЕНИМОСТЬ МЕТОДОВ ОЦЕНКИ КАЧЕСТВА СИНТАКСИЧЕСКОГО АНАЛИЗА К РЕЗУЛЬТАТАМ СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА НА ОСНОВЕ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

Е.Д. Шамаева (*derinhelm@yandex.ru*)
Н.В. Лукашевич (*louk_nat@mail.ru*)

Московский государственный университет
им. М.В. Ломоносова, Москва

При применении больших языковых моделей для задачи синтаксического анализа возникают принципиально новые ситуации: завершение синтаксического анализа с ошибкой и существенное изменение предложения, для которого выполняется синтаксический анализ (в том числе удаление или добавление слов). Данная статья посвящена исследованию влияния таких ситуаций на применимость стандартных методов оценки качества синтаксического анализа к результатам работы синтаксического анализатора на основе больших языковых моделей. Экспериментальная оценка проведена для синтаксического анализатора U-DepPLLaMA на тестовой выборке датасета предложений с синтаксической разметкой Taiga. Установлено, что к результатам синтаксического анализа на основе больших языковых моделей нельзя применять метод оценки на основе вычисления среднего значения метрик UAS и LAS на тестовых предложениях. Кроме того, без существенной модификации нельзя использовать стандартный алгоритм выравнивания наборов токенов. Реализация исследования доступна по адресу https://github.com/Derinhelm/parser_stat/tree/llm_taiga.

Ключевые слова: синтаксический анализ, деревья зависимостей, большие языковые модели, токенизация.

Введение

Одной из классических задач обработки текстов является автоматический синтаксический анализ, в результате которого определяется синтаксическая структура предложения. В настоящее время синтаксический анализатор применяется как вспомогательный инструмент при разработке интерпретируемых методов решения таких задач, как распознава-

ние именованных сущностей [Li et al., 2024], [Vasiliev et al., 2023], [Alonso et al., 2021], [Nikolaev, 2023], оценка сложности текста [Morozov et al., 2024], перефразирование [Liu et al., 2023], выявление плагиата [Taufiq, 2023].

На качество синтаксического анализа влияет качество предшествующего этапа обработки текста, токенизации. На этапе токенизации происходит выделение токенов предложения (слов, знаков препинания и т.п.). Разбиение на токены существенно влияет не только на синтаксический анализ, но и на оценку качества синтаксического анализа.

Многие синтаксические анализаторы используют нейронные сети [Chen et al., 2014], [Andor et al., 2016], [Dozat et al., 2017]. Однако в последние годы для синтаксического анализа начали использовать и большие языковые модели (как без дообучения [Ezquerro et al., 2025], так и с дообучением [Hromei et al., 2024]). Данные анализаторы существенным образом отличаются от классических нейросетевых синтаксических анализаторов. Синтаксические анализаторы, не использующие большие языковые модели, устанавливают синтаксические отношения между токенами, выделенными из предложения на этапе токенизации. Синтаксические анализаторы на основе больших языковых моделей (Large Language Model, LLM) генерируют синтаксическую структуру предложения в текстовом виде. При этом, сгенерированное дерево зависимостей может не соответствовать исходному предложению.

В данной статье исследуется применимость стандартных методов оценки качества синтаксического анализа к результатам синтаксического анализатора на основе LLM. В качестве синтаксического анализатора на основе LLM выбран синтаксический анализатор U-DepPLLaMA, исследование проведено на тестовой выборке датасета предложений с синтаксической разметкой Taiga.

1. Методология исследования

1.1. Синтаксический анализ

Одним из широко используемых способов представления синтаксической структуры предложения является дерево зависимостей (пример дерева зависимостей приведен на рис. 1). Дерево зависимостей – это ориентированный граф вида дерево, в котором вершины соответствуют токенам предложения, а ребра – синтаксическим связям между токенами. Для каждого ребра предложения указан тип связи между соответствующими токенами. Каждому токену соответствует ровно[1] один главный токен (токен, от которого к данному токену проведено ребро). Главным токеном для корня дерева является вспомогательный токен ROOT.

[1] Следовательно, общее количество ребер в дереве зависимостей равно количеству токенов в предложении (без учета вспомогательного токена).

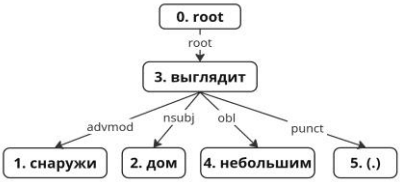


Рис. 1. Пример дерева зависимостей для предложения
«Снаружи дом выглядит небольшим»

Датасеты деревьев зависимостей используются как для обучения нейросетевых синтаксических анализаторов, так и для оценки качества их работы. Для русского языка существует несколько датасетов предложений с синтаксической разметкой. В частности, в проект Universal Dependencies[2] входят русскоязычные датасеты SynTagRus[3], GSD[4], PUD[5], Taiga[6] и Poetry[7]. Датасет Taiga основан на текстах электронной коммуникации, остальные датасеты – на литературных текстах: новостных, поэтических, публицистических и художественных. Для датасетов из проекта Universal Dependencies синтаксическая разметка производилась вручную. Кроме того, существуют датасеты, для которых синтаксическая разметка предложений выполнялась автоматически: датасеты PaRuS [Власова, 2019] и Nerus[8], банк синтаксических деревьев RSTB[9,10].

Предложения из датасетов проекта Universal Dependencies и других наборов предложений с синтаксической разметкой использовались и для соревнований синтаксических анализаторов, в частности, для наиболее известных соревнований CoNLL 2017 Shared Task [Zeman et al., 2017], CoNLL 2018 Shared Task [Zeman et al., 2018], GramEval 2020 Shared Task [Ляшевская, 2020]. В этих соревнованиях принимали участие только классические нейросетевые синтаксические анализаторы.

На соревнованиях для оценки качества синтаксических анализаторов используются такие метрики, как UAS (Unlabeled Attachment Score) и LAS (Labeled Attachment Score), MLAS (Morphology-aware Labeled Attachment

[2] https://universaldependencies.org/.

[3] https://universaldependencies.org/treebanks/ru_syntagrus/index.html.

[4] https://universaldependencies.org/treebanks/ru_gsd/index.html.

[5] https://universaldependencies.org/treebanks/ru_pud/index.html.

[6] https://universaldependencies.org/treebanks/ru_taiga/index.html.

[7] https://universaldependencies.org/treebanks/ru_poetry/index.html.

[8] https://github.com/natasha/nerus.

[9] http://otipl.philol.msu.ru/~soiza/testsynt/files/info.htm.

[10] Большая часть предложений в RSTB была размечена автоматически, 800 предложений – вручную.

Score) и BLEX (Bi-LEXical dependency score) [Zeman et al., 2017]. Семантически метрика UAS соответствует доле токенов предложения, для которых верно определен главный токен, метрика LAS – доле токенов, для которых верно определен главный токен и тип связи. Метрики MLAS и BLEX дополнительно учитывают морфологические характеристики токенов и корректность определения леммы соответственно.

1.2. Выравнивание наборов токенов

Вычисление метрики производится отдельно на каждом предложении: для каждого токена происходит сравнение результата синтаксического анализа с эталоном. Однако набор токенов, из которых состоит эталонное дерево зависимостей, может отличаться от набора токенов дерева зависимостей, построенного синтаксическим анализатором. Например, в датасете Taiga предложение «*Режиссёр-педагог: А. Вученович.*» разделено на 7 токенов: «*Режиссёр*», «-», «*педагог*», «*:*», «*А.*», «*Вученович*», «.», а результат работы синтаксического анализатора на основе LLM (U-DepPLLaMA) состоит из 4 токенов: «*Режиссёр-педагог*», «*:*», «*А.*», «*Вученович.*». Для решения проблемы несоответствия наборов токенов введен предварительный этап оценки качества синтаксического анализатора: выравнивание эталонного дерева зависимостей и дерева зависимостей, построенного синтаксическим анализатором.

Стандартным алгоритмом выравнивания является алгоритм, используемый для соревнования синтаксических анализаторов CoNLL-2017 [Zeman et al., 2017]. При использовании этого алгоритма два токена считаются одинаковыми, если совпадают индексы их начала и конца в исходном предложении. Сопоставление происходит за один проход слева направо по двум наборам токенов. Листинг алгоритма сопоставления наборов токенов приведен в Приложении А.

В классическом алгоритме выравнивания предполагается, что и для эталонного дерева зависимостей, и для построенного дерева зависимостей верно следующее: при объединении (с учетом порядка) токенов дерева зависимостей получается исходное предложение (с точностью до пробелов).

1.3. Применение больших языковых моделей для синтаксического анализа

Синтаксические анализаторы, основанные на большой языковой модели, сводят задачу синтаксического анализа к задаче генерации последовательности. Для этого используется текстовая запись дерева зависимостей. Например, при применении текстовой формы записи дерева зависимостей, разработанной в [Hromei, 2024], для дерева зависимостей предложения «Дети играют в прятки» будет создана текстовая запись «[root[nsubj[Дети]][играют][obl[case[в]][прятки]]]». Соответствующее дерево зависимостей приведено на рис. 2.

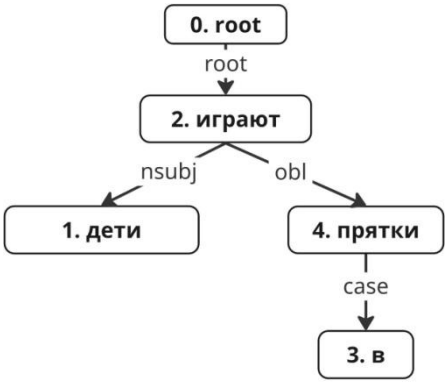


Рис. 2. Дерево зависимостей для предложения «Дети играют в прятки»

При использовании синтаксического анализатора на основе LLM в LLM передается промпт, содержащий исходное предложение. Пример такого промпта приведен в Листинге 1. В результате работы LLM возвращается исходный промпт, дополненный текстовой записью дерева зависимостей. Пример результата работы LLM приведен в Листинге 2. Далее по разработанному в [Hromei, 2024] алгоритму дерево зависимостей из текстовой записи преобразуется в стандартный формат.

Листинг 1

```
Input:
Дети играют в прятки
Answer:
```

Листинг 2

```
Input:
Дети играют в прятки
Answer:
[root[nsubj[Дети]][играют][obl[case[в]][прятки]]]
```

В классических нейросетевых синтаксических анализаторах набор токенов (и соответственно вершин дерева зависимостей) определяется до начала синтаксического анализа. Этот набор токенов при объединении совпадает с исходным предложением (с точностью до пробельных символов). В синтаксическом анализаторе на основе LLM текстовое представление дерева зависимостей генерируется LLM, поэтому для некоторых предложений сгенерированный набор токенов при объединении может не совпадать с исходным предложением. Предложение, построенное по сгенерированному дереву зависимостей, может отличаться от исходного предложения. Поэтому метод сопоставления токенов на основе индексов

начала и конца токена в предложении неприменим из-за потенциального несовпадения предложений. Следовательно, для этих случаев неприменим и классический алгоритм выравнивания.

1.4. Описание эксперимента

Целью данной работы является анализ предложений, для которых неприменим классический алгоритм выравнивания. Поэтому для данного эксперимента выбрана тестовая выборка датасета Taiga, состоящего из текстов электронной коммуникации. Для этих текстов достаточно часто встречается несовпадение токенизации (за счет существенного количества опечаток, хештегов и смайликов). Тестовая выборка датасета Taiga состоит из 881 тестового предложения.

Для данной работы выбран[11] синтаксический анализатор с дообучением U-DepPLLaMA[12] [Hromei et al., 2024]. Данный анализатор построен на основе большой языковой модели LLaMA и дообучен на предложениях с синтаксической разметкой. Дообучение происходило с помощью метода LoRA с квантованием (Q-LoRA), использовались преимущественно англоязычные предложения, но присутствовали и русскоязычные. В данной статье для проведения экспериментов выбрана версия анализатора с наименьшим количество параметров (7 миллиардов).

Для оценки качества синтаксического анализа использованы метрики UAS и LAS. Метрики MLAS и BLEX в данном исследовании неприменимы, поскольку анализатор U-DepPLLaMa не предоставляет информации о морфологических характеристиках токенов и их леммах.

2. Сравнение с существующими работами

На данный момент оценке качества синтаксических анализаторов на основе LLM посвящено небольшое количество статей. В этих статьях, в отличии от данной, не проводился детальный анализ результатов работы синтаксического анализатора. В [Hromei et al., 2024] для оценки качества синтаксических анализаторов используются только предложения, на которых синтаксический анализ завершился успешно и нет существенного несовпадения эталонного набора токенов и набора токенов, полученного в результате синтаксического анализа. В [Ezquerro et al., 2025] при оценке качества синтаксического анализа учитываются только предложения, для которых после преобразований удалось модифицировать построенное дерево зависимостей так, чтобы оно по количеству токенов совпадало с эталонным.

[11] Для этого анализатора в открытом доступе предоставлены и дообученные модели с разным количеством параметров, и программные реализации преобразования между графовым и текстовым форматами дерева зависимостей.

[12] https://github.com/crux82/u-deppllama; https://huggingface.co/sag-uniroma2/u-depp-llama-2-7b.

3. Результаты

При тестировании синтаксического анализатора U-DepPLLaMA на тестовой выборке датасета Taiga выявлены как предложения, для которых синтаксический анализ завершился неуспешно, так и предложения, для которых набор токенов существенным образом не совпадает с эталонным. Синтаксический анализ завершился неуспешно для 56 предложений (6.36%): LLM сгенерировала последовательность, которую невозможно корректно преобразовать в дерево зависимостей. Существенное несовпадение токенов выявлено для 87 предложений (10%). Эти предложения можно разделить на следующие группы:

1) с потерей токенов (слов, скобок, кавычек);
2) с добавлением лишних токенов;
3) с перестановкой токенов;
4) с заменой токенов.

Количество таких предложений и соответствующие примеры приведены в табл. 1 и 2. В данных таблицах используются следующие обозначения: Ист. – источник набора токенов, ЭТ – эталонный набор токенов, СА – синтаксический анализатор.

Таблица 1

Количество предложений с существенным несовпадением токенизации

Тип ошибки	Количество предложений	Доля от общего количества предложений
Потеря токенов (кроме скобок и кавычек)	10	1.14%
Лишний токен	4	0.45%
Перестановка токенов	28	3.18%
Потеря скобок	11	1.25%
Потеря кавычек	13	1.48%
Замена токена	15	1.70%

Таблица 2

Примеры существенного несовпадения токенизации

Тип ошибки	Ист.	Пример
Потеря токенов (кроме скобок и кавычек)	ЭТ	«За», «несколько», «лет», «,», «я», «видел», «множество», «постов», «с», «этими», «двумя», «песнями», **«,», «теперь», «пришла», «и», «моя», «очередь», «их», «поставить», «....», «)))»**
	СА	«За», «несколько», «лет», «,», «я», «видел», «множество», «постов», «с», «этими», «двумя», «песнями»
	ЭТ	«Кофе», «!», «Кофе», «!», «Кофе», «!», «Кофе», «!», «Кофе», «!», **«Кофе», «!»**
	СА	«Кофе!», «Кофе!», «Кофе!», «Кофе!», «Кофе!»

Окончание табл. 2

Тип ошибки	Ист.	Пример
Лишний токен	ЭТ	«Мы», «партию», «славим», «единороссов», «-», «Партию», «власти», «богатеньких», «боссов», «!»
	СА	«Мы», «партию», «славим», «единороссов», «-», «Партию», «власти», «богатеньких», «боссов!», **«богатеньких»**
Перестановка токенов	ЭТ	**«Их», «можно»**, «не», «сушить», «в», «духовке», «.»
	СА	**«можно», «Их»**, «не», «сушить», «в», «духовке.»
Потеря скобок	ЭТ	«-», «Покажи», «,», «как», «Ежик», «кушает», «яблоко», «(», «надуваем», «по», «очереди», «щечки», **«)»**, «;».
	СА	«-», «Покажи», «,», «как», «Ежик», «кушает», «яблоко», «(надуваем», «по», «очереди», «щечки», «;».
Потеря кавычек	ЭТ	«А», «кто», «там», «был», «"», «правее», **«"»**, «,», «время», «покажет», «.».
	СА	«А», «кто», «там», «был», «"», «правее», «,», «время», «покажет.».
Замена токенов	ЭТ	**«Также»**, «присутствует», «молодой», «Сергей», «Соседов», «.», «#сноб_news».
	СА	**«Такое»**, «присутствует», «молодой», «Сергей», «Соседов.», «#сноб_news».
	ЭТ	«3», «)», «Страница», «подписана», «НАСТОЯЩИМИ», «именем», «и», **«фамилией»**, «,», «а», «не», «вымышленными», «никнеймами», «.».
	СА	«3)», «Страница», «подписана», «НАСТОЯЩИМИ», «именем», «и», **«фамилии»**, «,», «а», «не», «вымышленными», «никнеймами.».
	ЭТ	«Приведенные», «нами», «артикуляционные», «упражнения», «используются», **«логопедами»**, «для», «стимуляции», «речевой», «активности», «детей», «.».
	СА	«Приведенные», «нами», «артикуляционные», «упражнения», «используются», **«лагопедами»**, «для», «стимуляции», «речевой», «активности», «детей.».
	ЭТ	«А», **«ватный»**, «диск», «не», «одолжите», «?».
	СА	«А», **«ваттный»**, «диск», «не», «одолжите?».

Дополнительно выявлено, что достаточно частой ошибкой является присоединение последнего знака пунктуации к предпоследнему токену. Например, в предложении «Теперь пришло время для объединения.» в датасете содержится 6 токенов («Теперь», «пришло», «время», «для», «объединения», «.»), а в результате работы синтаксического анализатора – 5 токенов, поскольку токены «объединения», «.» соединены в один («Теперь»,

«пришло», «время», «для», **«объединения.»**). На данных предложениях с объединением токенов значения метрик UAS и LAS не могут быть равно 1.0 вне зависимости от качества синтаксического анализа. Количество таких тестовых предложений – 610 (69.24% от общего количества).

Объединение с предшествующим токеном достаточно часто происходит и для предпоследнего токена, если предпоследний токен является знаком пунктуации, а последний токен – смайликом или хештегом. Например, предложение «И это только начало!;)» в датасете рассматривается как 6 токенов («И», «это», «только», **«начало», «!»**, «;)»), а результат синтаксического анализа содержит 5 токенов («И», «это», «только», **«начало!»**, «;)»). Количество таких предложений – 17 (1.93%).

Заключение

В данной статье исследуется применимость стандартных методов оценки качества синтаксических анализаторов к синтаксическим анализаторам на основе больших языковых моделей. Для исследования использованы синтаксический анализатор на основе большой языковой модели U-DepPLLaMA и тестовая выборка русскоязычного датасета предложений с синтаксической разметкой Taiga.

Установлено, что к результатам таких анализаторов неприменим стандартный алгоритм выравнивания наборов токенов. Причина этого заключается в существенном различии эталонного набора токенов и набора токенов, созданного синтаксическим анализатором: потери токенов, добавления лишних токенов, перестановки токенов и замены токенов на несуществующие в исходном тексте предложения.

Кроме того, неприменима оценка качества синтаксического анализа через среднее значение метрик UAS и LAS на множестве всех тестовых предложений. Вместо этого необходимо отдельно рассматривать предложения, для которых синтаксический анализ завершился неуспешно, и предложения, с существенным несовпадением токенизации (из-за которого невозможно применение алгоритма выравнивания). На остальных предложениях метрики UAS и LAS могут быть корректно вычислены.

Дополнительно установлено, что при применении синтаксического анализатора на основе большой языковой модели возможны галлюцинации большой языковой модели. Поэтому построенное дерево зависимостей может содержать токены, отсутствующие в исходном предложении.

В дальнейшем планируется проведение исследований по адаптации метрик UAS и LAS для результатов синтаксического анализа с существенным несовпадением токенизации. Кроме того, планируется более детальное исследование ошибок анализаторов (в частности, с применением методики, разработанной при создании банка синтаксических деревьев RSTB). Помимо этого, планируется провести сравнение работы больших

языковых моделей, дообученных для синтаксического анализа, и моделей без дообучения. Перспективным направлением исследований также является проверка применимости больших языковых моделей не только для грамматик зависимостей, но и для грамматик составляющих и гибридных подходов.

Список литературы

[Власова, 2019] Власова Н.А. [и др.]. PaRuS – синтаксически аннотированный корпус русского языка // Программные системы: теория и приложения. – 2019. – Т. 10, № 4(43). – С. 181-199. – doi: 10.25209/2079-3316-2019-10-4-181-199.

[Ляшевская, 2020] Ляшевская О.Н., Шаврина Т.О., Трофимов И.В., Власова Н.А. Grameval 2020: дорожка по автоматическому морфологическому и синтаксическому анализу русских текстов // Annual International Conference Dialogue. (Москва, 17 июня – 20 июня 2020 г.). Труды конференции. – С. 553-569. – doi: 10.28995/2075-7182-2020-19-553-569.

[Alonso et al., 2021] Alonso M.A., Gómez-Rodríguez C., Vilares J. On the use of parsing for named entity recognition // Applied sciences. – 2021. – Vol. 11(3). – P. 1090. – doi: 10.3390/app11031090.

[Andor et al., 2016] Andor D., Alberti C., Weiss D. et.al. Globally normalized transition-based neural networks // In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany, 2016. – P. 2442-2452. – doi: 10.18653/v1/P16-123.

[Chen et al., 2014] Chen D., Manning C.D. A fast and accurate dependency parser using neural networks // In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014. – P. 740-750. – doi: 10.3115/v1/D14-1082.

[Dozat et al., 2017] Dozat T., Manning C.D. Deep Biaffine Attention for Neural Dependency Parsing // International Conference on Learning Representations (ICLR). – 2017.

[Ezquerro et al., 2025] Ezquerro A., Gómez-Rodríguez C., Vilares D. Better Benchmarking LLMs for Zero-Shot Dependency Parsing // In: Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), Tallinn, Estonia, 2025. – P. 121-135.

[Hromei et al., 2024] Hromei C.D., Croce D., Basili R. U-DepPLLaMA: Universal Dependency Parsing via Auto-regressive Large Language Models. IJCoL // Italian Journal of Computational Linguistics. – 2024. – Vol. 10(1). – doi: 10.4000/125nm.

[Li et al., 2024] Li L., Chen Z., Liao S. et al. Event Extraction in Complex Sentences Based on Dependency Parsing and Longformer // In: Proceedings of 2024 International Conference on Machine Learning and Intelligent Computing, Wuhan, China, 2024. – P. 1-7.

[Liu et al., 2023] Liu T., Sun Y., Wu J. et al. Unsupervised Paraphrasing under Syntax Knowledge // In: Proceedings of the AAAI Conference on Artificial Intelligence. – 2023. – P. 13273-13281. – doi: 10.1609/aaai.v37i11.26558.

[Morozov et al., 2024] Morozov D., Lagutina K., Drozhashchikh G. et al. Exploring the Feature Space for Cross-Domain Assessing the Complexity of Russian-Language Texts // In: 2024 Ivannikov Ispras Open Conference (ISPRAS), Moscow, Russian Federation, 2024. – P. 1-8. – doi: 10.1109/ISPRAS64596.2024.10899137.

[Nikolaev, 2023] Nikolaev I.E. Knowledge and skills extraction from the job requirements texts // Ontology of Designing. – 2023. – Vol. 13(2). – P. 282-293. – doi: 10.18287/2223-9537-2023-13-2-282-293.

[Taufiq, 2023] Taufiq U., Pulungan R., Suyanto Y. Named entity recognition and dependency parsing for better concept extraction in summary obfuscation detection // Expert Systems with Applications. – 2023. – Vol. 217. – P. 119579. – doi: 10.1016/j.eswa.2023.119579.

[Vasiliev et al., 2023] Vasiliev S., Korobkin D., Fomenkov S. Extracting the Component Composition Data of Inventions from Russian Patents using Dependency Tree Analysis // In: 2023 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), Sochi, Russian Federation, 2023. – P. 1030-1034. – doi: 10.1109/ICIEAM57311.2023.10139170.

[Zeman et al., 2017] Zeman D., Hajic J., Popel M. et al. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies // In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, 2017. – P. 1-19. – doi: 10.18653/v1/K17-3001.

[Zeman et. al., 2018] Zeman D., Hajic J., Popel M. et al. CoNLL 2018 shared task: Multilingual Parsing from Raw Text to Universal Dependencies // In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, 2018. – P. 1-21. – doi: https://doi.org/10.18653/v1/K18-2001.

Приложение А. Листинг программы стандартного сопоставления токенов

В Листинге 3 приведен псевдокод программы стандартного сопоставления токенов, разработанный в [Zeman et al., 2017]. Данный псевдокод модифицирован для иллюстративных целей[13]. В переменной gold_words хранится эталонный список токенов, в переменной system_words – список токенов, построенный синтаксическим анализатором. Для каждого токена заданы атрибуты start и end – индексы начала и конца токенов в строке-предложении. В переменной alignment хранятся токены, которые есть и в эталонном наборе токенов, и в наборе токенов, построенном синтаксическим анализатором.

Листинг 3

```
gi, si = 0, 0
while gi < len(gold_words) and si < len(system_words):
    if (gold_words[gi].start, gold_words[gi].end) == (sys-
tem_words[si].start, system_words[si].end):
        alignment.append_aligned_words(gold_words[gi], sys-
tem_words[si])
        gi += 1
        si += 1
    elif gold_words[gi].start <= system_words[si].start:
        gi += 1
    else:
        si += 1
```

[13] Оригинальная реализация программы находится по адресу https://universaldependencies.org/conll17/evaluation.html.

УДК 004.912
doi: 10.15622/rcai.2025.031

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ КЛЮЧЕВЫХ СЛОВ ДЛЯ РУССКОЯЗЫЧНЫХ НАУЧНЫХ СТАТЕЙ С ИСПОЛЬЗОВАНИЕМ ПСЕВДОРАЗМЕТКИ И КОНТРАСТИВНОГО ОБУЧЕНИЯ

К.Ш. Яушев (*kyaush@mail.ru*) [A]
Н.В. Лукашевич (*louk_nat@mail.ru*) [B]

[A] Московский государственный технический университет им. Н.Э. Баумана, Москва
[B] Московский государственный университет им. М.В. Ломоносова, Москва

В работе предложен многоэтапный подход к автоматическому порождению ключевых слов для русскоязычных научных статей. Метод основан на дообучении трансформеров с использованием псевдоразметки и контрастивного обучения, а также включает фильтрацию порождённых кандидатов. Реализованы две стратегии генерации псевдоразметки и архитектура с биэнкодером для отбора релевантных ключевых слов. Эксперименты на корпусе математики и компьютерных наук демонстрируют превосходство предложенного подхода над классическими и нейросетевыми методами по метрикам F1, ROUGE-1 и BERTScore.

Ключевые слова: генерация ключевых слов, автоматическая разметка, контрастивное обучение, фильтрация ключевых слов.

Введение

В последние годы объём научных публикаций растёт стремительно [Cicero, 2025], что усложняет систематизацию и поиск информации. В условиях информационной перегрузки ключевые слова играют важную роль для навигации в библиографических системах (Scopus, Web of Science, eLIBRARY) и влияют на индексацию и цитируемость научных работ [Гендина и др., 2018].

Размеченные авторами ключевые слова могут быть слишком специфичными для узкой тематики статьи или, напротив, чрезмерно общими, что затрудняет их использование в поисковых системах и при тематиче-

ской классификации. Это делает актуальной задачу автоматической генерации ключевых слов, позволяющую формировать более сбалансированный и релевантный набор терминов для поиска и анализа.

Автоматическая генерация ключевых слов для русскоязычных текстов остаётся сложной из-за богатой морфологии, синтаксического разнообразия и отсутствия явных ключевых концептов в тексте [Glazkova et. al., 2024]. Для успешного решения задачи модели должны уметь выявлять скрытые семантические связи – неявные смысловые отношения между терминами, тематическими группами и контекстно связанными выражениями. Такие связи определяются на основе распределённых представлений и совместной встречаемости в корпусе. Современные большие языковые модели (LLM) на основе трансформеров доказали эффективность в генерации текстов и извлечении смысловых единиц [Vaswani et. al., 2017], однако их применение для русского языка ограничено дефицитом размеченных данных [Glazkova et. al, 2025].

В этой работе предложен многоэтапный метод, сочетающий псевдоразметку и контрастивное обучение, направленный на улучшение генерации как явно представленных, так и отсутствующих ключевых слов при ограниченных размеченных ресурсах.

1. Близкие работы

Методы выделения ключевых слов делятся на экстрактивные и абстрактивные. Экстрактивные (статистические: TF-IDF [Salton et. al., 1975], YAKE! [Campos et. al., 2020]; графовые: TextRank [Mihalcea et. al., 2004], TopicRank [Bougouin et. al., 2013]) выбирают значимые слова из текста, отличаются простотой и интерпретируемостью, но не могут генерировать отсутствующие в документе слова [Glazkova et. al., 2024].

Нейросетевые трансформеры (BERT [Devlin et. al., 2019], T5 [Raffel et. al., 2020], mBART [Tang et. al., 2021]) обучены на больших корпусах и способны выявлять глубокие семантические связи, генерируя новые релевантные ключевые слова. Среди русскоязычных моделей выделяются ruT5 [Zmitrovich et. al., 2024], Vikhr [Nikolich et. al., 2024] и Saiga [Gusev, 2023]. Главная сложность — генерация точных по смыслу ключевых слов при ограниченном объёме данных.

Ранее в русскоязычной тематике изучались модели mT5 и mBART, подтвердившие потенциал генеративных подходов [Glazkova et. al., 2024]. Инструктивные LLM-модели показали высокую эффективность в few-shot режиме, но требуют значительных ресурсов [Glazkova et. al., 2025].

Также развивается направление, в котором генеративные подходы дополняются механизмами фильтрации и ранжирования: например, с использованием биэнкодеров для отбора релевантных кандидатов по семантической близости [Choi et. al., 2023] или применения псевдоразметки для расширения обучающей выборки за счёт неразмеченных текстов [Kang et. al., 2024].

Настоящее исследование развивает эти направления, предлагая методы с псевдоразметкой и контрастивным обучением для повышения качества генерации при дефиците размеченных данных.

2. Методы автоматической разметки и архитектура обучения

Для преодоления дефицита размеченных данных предлагается двухэтапный подход, основанный на идеях [Kang et. al., 2024] и адаптированный под особенности русскоязычных научных текстов. На первом этапе проводится предобучение генеративной модели на корпусе, обогащённом псевдоразметкой, на втором этапе тонкая настройка на размеченных данных, чтобы развить способность модели не только извлекать, но и порождать ключевые слова.

Псевдоразметка формируется по двум стратегиям. Первая основана на *маскировании* слабо релевантных, но явно присутствующих ключевых слов. Для каждой аннотации библиотекой ruTermExtract[1] извлекается упорядоченный по релевантности список ключевых слов (именных форм) в корректных словоформах, учитывая морфологические особенности русского языка. Релевантность определяется косинусной близостью векторных представлений ключевых слов и текста, полученных с помощью модели E5[2]. Первые 5 наиболее релевантных ключевых слов считаются эталонными и сохраняются в тексте, а позиции с 6-й по 10-ю маскируются специальным токеном. Такой отбор исключает заведомо нерелевантные выражения и фокусирует модель на восстановлении терминов, связанных по смыслу, но не очевидных, что способствует генерации отсутствующих в тексте ключевых слов.

Вторая стратегия обучает генерацию релевантных, но явно непредставленных ключевых слов. Для этого используется «глобальная коллекция» – топ-5 ключевых слов, извлечённых ruTermExtract из остальных документов корпуса. Поиск семантически близких кандидатов выполняется методом HNSW [Malkov et. al., 2018], выбираются 5 наиболее близких по косинусному сходству и отсутствующих в исходном тексте ключевых слов. Они добавляются как целевые метки для обучения абстрактивной генерации.

Рассмотрим статью И.А. Чижова и Н.П. Заеца «Моделирование процесса теплопроводности многослойной конструкции для выполнения тепловизионного контроля» [Чижов и др., 2015]. Разметка выполняется по заголовку и аннотации. Ниже приведён текст с псевдоразметкой (цитата по [Чижов и др., 2015, с. 1]):

[1] https://github.com/igor-shevchenko/rutermextract.
[2] https://huggingface.co/d0rj/e5-large-en-ru.

[Моделирование процесса] [теплопроводности многослойной конструкции] для выполнения [тепловизионного контроля]. В статье представлен <этап моделирования> протекания [тепловых потоков] в многослойной конструкции, применяемые <численные решения> для <математического моделирования>. Перечислены учитываемые в модели <параметры> и проведён <анализ факторов> [теплового неразрушающего контроля].

В квадратных скобках представлены явно присутствующие ключевые слова; в угловых – маскируемые, предназначенные для генерации.

Ключевые слова, найденные глобально в коллекции, отсутствующие в тексте аннотации: *моделирование процесса теплопередачи*, *тепловизионный анализ*, *моделирование процессов теплопереноса*, *исследование проблем теплопередачи*, *численное моделирование процесса теплопередачи*.

Для повышения устойчивости к шуму и улучшения различения релевантных и нерелевантных ключевых слов применяется архитектура с элементами контрастивного обучения (рис. 1), предложенная в [Choi et. al., 2023]. Она включает экстрактор-генератор и ранкер.

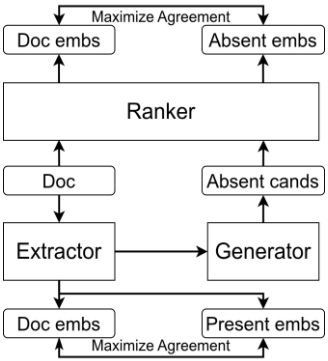


Рис. 1. Архитектура генерации и фильтрации ключевых слов

Экстрактор-генератор (рис. 2) – энкодер-декодерная модель, одновременно обучаемая на задаче извлечения и генерации с помощью комбинированной функции потерь с контрастивной частью (NT-Xent) и максимальным правдоподобием (MLE) [Chen et. al., 2020]:

$$\mathcal{L} = \gamma \cdot \mathcal{L}_{NT-Xent} + (1 - \gamma) \cdot \mathcal{L}_{MLE}. \quad (2.1)$$

Это позволяет объединить точность экстракции с обобщающей способностью генерации. Для формирования негативных примеров в контрастивном обучении применяется *hard negative mining*: в качестве «труд-

ных» негативов отбирается не более 5-ти кандидатных ключевых слов, извлечённые с помощью ruTermExtract, которые отсутствуют в эталонной разметке (псевдоразметке) датасета, но имеют высокое семантическое сходство с релевантными ключевыми словами. Такой подход помогает модели точнее проводить границу между действительно релевантными и нерелевантными ключевыми словами, минимизируя влияние тривиальных различий.

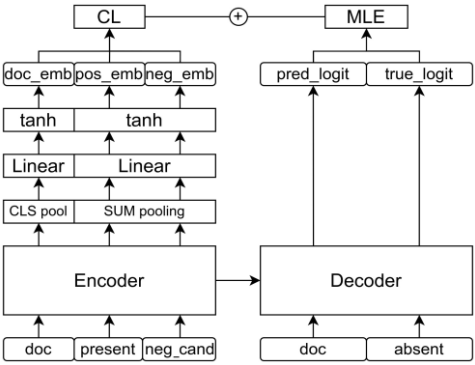


Рис. 2. Архитектура экстрактора-генератора с контрастивным модулем

Ранкер – биэнкодерная модель, контрастивно обученная различать релевантные и нерелевантные ключевые слова. После генерации кандидатных ключевых слов (например, с применением beam search) ранкер отбирает наиболее подходящие, фильтруя шум и снижая вероятность появления «галлюцинаций».

Общая стратегия обучения включает два этапа. Сначала модели обучаются на псевдоразмеченных данных, затем проводится тонкая настройка на вручную размеченном корпусе научных аннотаций, что позволяет повысить точность и адаптировать модели к специфике предметной области.

3. Эксперименты

3.1. Данные

В качестве основного размеченного корпуса для обучения и тестирования использовался датасет Math&CS[3], состоящий из 8348 аннотаций русскоязычных научных статей из области математики и компьютерных наук. Разделение на обучающую (5844) и тестовую (2504) выборки выполнено авторами датасета. В среднем каждая аннотация содержит от 3-х до 5-ти

[3] https://huggingface.co/datasets/aglazkova/keyphrase_extraction_russian.

(4.34±1.5) эталонных ключевых слов. Важной особенностью корпуса является то, что 53.66% эталонных ключевых слов отсутствуют явно в тексте аннотаций, что подчеркивает необходимость применения генеративных подходов.

Для псевдоразметки использовался обширный корпус ruSciBench[4], содержащий более 190 тысяч аннотаций научных статей из различных областей.

3.2. Используемые модели и их параметры

Для всестороннего анализа были выбраны модели из трех категорий:

1. Классические алгоритмы: YAKE![5] и ruTermExtract[6].
2. Нейросетевые модели: mT5-base 580M[7], mBART-large 610M[8], e5-large-en-ru 366M[9] (ранкер) а также несколько конфигураций модели ru-mbart-summ 380M[10], дообученной на задаче суммаризации:
   - &mask: обучение с маскированием;
   - &generator: модель с контрастивной функцией потерь;
   - &generator&ranker: полная модель с генератором и ранкером.
3. Инструктивные нейросетевые модели: Mistral-7B-Instruct[11], Vikhr-7B-Instruct[12] и Saiga-Mistral-7B-Lora[13] в режимах zero-shot и few-shot.

В качестве базовой была выбрана модель ru-mbart-summ – дообученная версия mbart_ru_sum_gazeta[14], изначально обученной на новостном корпусе и превосходящей T5 и GPT-3 по метрикам суммаризации. Дополнительное обучение на расширенном русскоязычном наборе повысило её способность обобщать тексты разных доменов. Архитектура «энкодер–декодер» и специализация на сжатии содержания делают её подходящей для генерации ключевых слов, поскольку эта задача требует выделения и формулирования основной идеи текста.

Настройка параметров следовала рекомендациям из оригинальных работ [Glazkova et. al., 2025]. Для энкодер–декодерных моделей (mT5, mBART, ru-mbart-summ и их модификаций) использовались: 10 эпох, максимальная длина входа 256 токенов, learning rate $4 \cdot 10^{-5}$, оптимизаторы Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 4 \cdot 10^{-8}$) и AdamW (с weight_decay =

[4] https://huggingface.co/datasets/mlsa-iai-msu-lab/ru_sci_bench.
[5] https://github.com/boudinfl/pke.
[6] https://github.com/igor-shevchenko/rutermextract.
[7] https://huggingface.co/google/mt5-base.
[8] https://huggingface.co/facebook/mbart-large-50.
[9] https://huggingface.co/d0rj/e5-large-en-ru.
[10] https://huggingface.co/d0rj/ru-mbart-large-summ.
[11] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2.
[12] https://huggingface.co/Vikhrmodels/Vikhr-7B-instruct_0.4.
[13] https://huggingface.co/IlyaGusev/saiga_mistral_7b_lora.
[14] https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta.

0.01 для моделей с дополнительной головой). Параметры генерации: repetition_penalty = 1.4, num_beams = 100, no_repeat_ngram_size = 2, num_return_sequences = 16 для моделей с ранжированием. Для инструктивных моделей (Mistral, Vikhr, Saiga) применялись zero-/few-shot режимы с промптами на русском по шаблону [Glazkova et. al., 2025], ограничение генерации 100 токенами и температура 0.5. Контрастивное обучение генераторов использовало $\tau = 0.1$ для потерь NT-Xent и $\gamma = 0.3$ при комбинировании с MLE (формула 2.1). При тестировании экстрактор выбирал 5 лучших кандидатов от ruTermExtract, генератор ограничивался 5 ключевыми словами, прочие генеративные модели – 10.

3.3. Метрики оценки

Качество генерации оценивалось с помощью трех метрик, рассчитанных для топ-10 сгенерированных ключевых слов:

1. F1-score@10: Гармоническое среднее точности и полноты для точных совпадений лемматизированных ключевых слов.
2. ROUGE-1@10: F1-мера перекрытия униграмм между сгенерированным и эталонным списками. Перед вычислением метрики ключевые слова лемматизировались и объединялись в одну строку через пробел.
3. BERTScore@10: Семантическая F1-мера, основанная на косинусном сходстве BERT[15]-эмбеддингов токенов. Перед вычислением метрики ключевые слова объединялись в одну строку через запятую.

4. Результаты и обсуждение

4.1. Сравнительный анализ моделей

Сводные результаты экспериментов приведены в табл. 1. В скобках указаны результаты из предыдущего исследования [Glazkova et al., 2025] для сравнения. При этом прямое сопоставление ограничено: в [Glazkova et al., 2025] для экстрактивных методов выбиралось лучшее значение среди топ-5, топ-10 и топ-15, тогда как в данной работе все модели оценивались при фиксированном топ-10. Кроме того, в [Glazkova et al., 2025] метрика ROUGE-1 вычислялась без лемматизации, а ключевые слова объединялись через запятую, тогда как в настоящем исследовании применялась лемматизация, а объединение выполнялось через пробел. Генеративные модели в предыдущей работе формировали не более 10 ключевых слов. Эти различия в методологии частично объясняют, почему при одинаковых моделях наши значения метрик, в том числе BERTScore, могут быть ниже, чем в [Glazkova et al., 2025].

[15] https://huggingface.co/google-bert/bert-base-multilingual-cased.

Таблица 1

Модель	F1@10	ROUGE1@10	BERTScore@10
ruTermExtract	10.19 (11.02)	29.97 (15.12)	71.26 (75.95)
YAKE!	04.04 (06.06)	26.27 (06.47)	70.56 (69.13)
mT5-base	4.42 (13.41)	17.99 (15.14)	67.86 (76.07)
mBart-large	16.43 (16.84)	33.65 (19.26)	72.34 (78.66)
ru-mbart-summ	16.46	33.61	74.32
&mask	17.80	**34.91**	75.30
&gen	02.04	06.41	77.71
&gen&rank	15.43	33.38	77.95
Mistral&few-shot	13.37 (15.08)	14.45 (16.30)	76.11 (74.85)
Vikhr&few-shot	14.57 (15.18)	17.67 (19.62)	77.14 (77.48)
Saiga&few-shot	15.04 (**20.16**)	17.98 (22.37)	78.11 (**79.50**)

В данном исследовании модификация ru-mbart-summ&mask показала лучшие значения F1 и ROUGE-1, что подтверждает эффективность двухэтапного обучения с маскированием. Подход обеспечивает генерацию точных и релевантных ключевых слов даже при ограниченном объёме размеченных данных.

Модель &gen&rank, несмотря на умеренные значения F1 и ROUGE-1, достигла наивысшего BERTScore среди моделей с учителем в данной работе. Однако по BERTScore абсолютным лидером остаётся Saiga в режиме few-shot из [Glazkova et al., 2025], что может быть связано с различиями в настройках генерации (например, температурой), вариативностью выходов инструктивных LLM и более агрессивной стратегией перефразирования, которая повышает семантическое сходство.

Следует отметить ограниченность сравнений: модель с фильтрацией (&ranker) не была применена к конфигурации с маскированием, которая лидирует по F1/ROUGE-1. Комбинирование этих подходов, вероятно, позволило бы достичь баланса между точностью и семантической насыщенностью, что представляет интерес для дальнейшей работы.

4.2. Примеры успешных и неудачных предсказаний

Для анализа был взят пример из статьи «Процедурное программирование, операторы ветвления и циклов в системе компьютерной алгебры» [Дубовицкий, 2012].

Эталонные ключевые слова: *процедурное программирование, оператор ветвления, операторы циклов, система компьютерной алгебры, система компьютерной алгебры mathpar*. Сводные результаты экспериментов представлены в табл. 2.

Таблица 2

Модель	Сгенерированные ключевые слова
YAKE!	процедурный программирование, оператор ветвление, компьютерный алгебра mathpar, создание оператор ветвление
ruTermExtract	процедурный программирование, компьютерный алгебра, система, алгоритм, создание оператор ветвление
mT5-base	информационная система, веб-технология, информатика, метод ветвления, цикл
mBart-large	алгебра, ветвление, цикл, программа, процедурный программирование
ru-mbart-summ	процедурное программирование, система компьютерной алгебры, структура системы, операторы ветвления и циклов
ru-mbart-summ&mask	процедурное программирование, система компьютерной алгебры, структура системы, операторы ветвления и циклов
ru-mbart-summ&gen	компьютерная алгебра, процедурное программирование, процедуры, алгоритмы, цикл, алгебраические операторы, архитектура системы
ru-mbart-summ&gen&rank	компьютерная алгебра, процедурное программирование, процедуры, алгоритмы, цикл, алгебраические системы, архитектура системы, ветвление, алгебраические методы, оператор ветвления
Mistral&few-shot	процедурное программирование, операторы ветвления и циклов, алгоритмы, компьютерная алгебра, mathpar

Классические методы, такие как YAKE! и ruTermExtract, показывают ограниченную способность к морфологически точной генерации и склонны к порождению лексически искажённых форм (оператор ветвление, компьютерный алгебра). Это отражается в низких значениях F1 и ROUGE-1, несмотря на поверхностную релевантность отдельных терминов.

Модель ru-mbart-summ&mask демонстрирует наилучшее соответствие эталону как в лексическом, так и в семантическом плане. Её предсказания почти полностью совпадают с референсными ключевыми словами, что объясняет высокие значения F1 и ROUGE-1 и делает её лидером по этим метрикам.

Модель ru-mbart-summ&gen&rank обеспечивает высокий уровень семантического разнообразия: она охватывает больше аспектов, включая термины второго порядка (алгоритмы, архитектура системы). Однако избыточность и наличие пересечений между фразами снижают точность на уровне F1/ROUGE, несмотря на один из самых высоких BERTScore, отражающий её семантическое богатство.

Инструктивные модели, такие как Mistral&few-shot, показывают высокое качество генерации: предсказания совпадают с эталоном и грамматически, и семантически. Это подтверждается их сильными значениями BERTScore. Однако, в отличие от ru-mbart-summ&mask, они уступают по F1 и ROUGE, что может быть связано с большей вариативностью выходов и нестабильностью при генерации.

Заключение

В работе представлен и экспериментально оценён комплексный подход к автоматическому порождению ключевых слов для аннотаций русскоязычных научных статей. Основной вклад заключается в разработке и анализе архитектур, эффективно работающих при ограниченном количестве размеченных данных.

Двухэтапная стратегия обучения – предобучение на псевдоразмеченных данных и донастройка на небольшом объёме эталонных аннотаций – существенно повышает качество генерации. Модель ru-mbart-summ&mask достигла наивысших значений F1 и ROUGE-1, демонстрируя высокую лексическую точность и стабильность, что делает её предпочтительной для задач с приоритетом точного совпадения с референсными ключевыми словами.

Контрастивная архитектура с генерацией кандидатов и последующей фильтрацией ранкером показала наивысший BERTScore, что подтверждает эффективность подхода, где генерация обеспечивает разнообразие, а фильтрация – семантическую релевантность.

Примечательно, что компактная ru-mbart-summ с минимальными изменениями достигает BERTScore, сопоставимого с крупными инструктивными моделями (например, Mistral 7B в few-shot режиме), обеспечивая высокое качество при умеренных вычислительных затратах.

Среди ограничений – отсутствие этапа фильтрации в конфигурации с маскированием и зависимость качества от точности псевдоразметки. Дальнейшие исследования следует направить на интеграцию фильтрации во все генеративные конфигурации, улучшение качества псевдоразметки для снижения уровня шума, проведение кросс-доменных экспериментов (например, в медицине, химии), углублённую настройку инструктивных моделей для повышения F1 и ROUGE, а также тесты на статистическую значимость (t-критерий Стьюдента) и привлечение экспертной оценки моделей с близкими метриками.

В целом, разработанные методы демонстрируют значительный шаг вперед в решении задачи автоматической генерации ключевых слов для русского языка и могут быть использованы для улучшения систем индексации и поиска в научных электронных библиотеках.

Список литературы

[Гендина и др., 2018] Гендина Н.И., Колкова Н.И. Методика формализованного аннотирования интернет-ресурсов // Научные и технические библиотеки. – 2018. – № 8. – С. 48-65. – doi: 10.33186/1027-3689-2018-8-48-65.

[Гусев, 2023] Гусев И. Проект RuLM. [Электронный ресурс] // GitHub. 2023. – URL: https://github.com/IlyaGusev/rulm (дата обращения: 01.03.2025).

[Дубовицкий, 2012] Дубовицкий Е.В. Процедурное программирование, операторы ветвления и циклов в системе компьютерной алгебры // Вестник российских университетов. Математика. – 2012. – Т. 17, № 2. – С. 598-602.

[Чижов и др., 2015] Чижов И.А., Заец Н.П. Моделирование процесса теплопроводности многослойной конструкции для выполнения тепловизионного контроля [Электронный ресурс] // Приложение математики в экономических и технических исследованиях. – 2015. – № 5. – С. 139-145. – URL: https://e.lanbook.com/journal/ issue/295367 (дата обращения: 01.03.2025).

[Bougouin et. al., 2013] Bougouin A., Boudin F., Daille B. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction // In: Proc. 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan, 2013. – P. 543-551.

[Campos et. al., 2020] Campos R., Mangaravite V., Pasquali A., Jorge A.M., Nunes C., Jatowt A. YAKE! Keyword Extraction from Single Documents using Multiple Local Features // Information Sciences. – 2020. – Vol. 509. – P. 257-289. – doi: 10.1016/j.ins.2019.09.013.

[Chen et. al., 2020] Chen T., Kornblith S., Norouzi M., Hinton G. A simple framework for contrastive learning of visual representations // In: Proc. 37th International Conference on Machine Learning (ICML 2020), Vienna, Austria, 2020. – P. 1597-1607. – doi: 10.1145/3701716.3715239.

[Choi et. al., 2023] Choi M., Gwak C., Kim S., Kim S., Choo J. SimCKP: Simple Contrastive Learning of Keyphrase Representations // In: Proc. Findings of the Association for Computational Linguistics (EMNLP 2023), Singapore, Singapore, 2023. – P. 3003-3015. – doi: 10.18653/v1/2023.findings-emnlp.199.

[Cicero, 2025] Cicero T. Forecasting the Scientific Production Volumes of G7 and BRICS Countries in a Comparative Analysis // Publications. – 2025. – Vol. 13(1). – Article 6. – doi: 10.3390/publications13010006.

[Devlin et. al., 2019] Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // In: Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, USA, 2019. – P. 4171-4186. – doi: 10.18653/v1/N19-1423.

[Glazkova et. al., 2024] Glazkova A., Morozov D. Exploring Fine-tuned Generative Models for Keyphrase Selection: A Case Study for Russian // In: Proc. 26th International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID 2024), Nizhny Novgorod, Russia, 2024. – URL: https://damdid2024.frccsc.ru/files/papers/DAMDID_2024_paper_11.pdf.

[Glazkova et. al., 2025] Glazkova A., Morozov D., Garipov T. Key Algorithms for Keyphrase Generation: Instruction-Based LLMs for Russian Scientific Keyphrases // In: Proc. Analysis of Images, Social Networks and Texts (AIST 2024), Bishkek, Kyrgyzstan, 2025. – P. 107-119. – doi: 10.1007/978-3-031-88036-0_5.

[Kang et. al., 2024] Kang B., Shin Y. Improving Low-Resource Keyphrase Generation through Unsupervised Title Phrase Generation // In: Proc. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 2024. – P. 8853-8865. – URL: https://aclanthology.org/2024.lrec-main.775.

[Malkov et. al., 2018] Malkov Y.A., Yashunin D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2020. – Vol. 42(4). – P. 824-836. – doi: 10.1109/TPAMI.2018.2889473.

[Mihalcea et. al., 2004] Mihalcea R., Tarau P. TextRank: Bringing Order into Text // In: Proc. Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, 2004. – P. 404-411. – URL: https://aclanthology.org/W04-3252.

[Nikolich et. al., 2024] Nikolich A., Korolev K., Bratchikov S., Kiselev I., Shelmanov A. Vikhr: Constructing a State-of-the-art Bilingual Open-Source Instruction-Following Large Language Model for Russian // In: Proc. Fourth Workshop on Multilingual Representation Learning (MRL 2024), Miami, Florida, USA, 2024. – P. 189-199. – doi: 10.18653/v1/2024.mrl-1.15.

[Raffel et. al., 2020] Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research. – 2020. – Vol. 21(140). – P. 1-67.

[Salton et. al., 1975] Salton G., Wong A., Yang C.S. A Vector Space Model for Automatic Indexing // Communications of the ACM. – 1975. – Vol. 18(11). – P. 613620. – doi: 10.1145/361219.361220.

[Tang et. al., 2021] Tang Y., Tran C., Li X., Chen P., Goyal N., Chaudhary V., Gu J., Fan A. Multilingual Translation from Denoising Pre-Training // In: Proc. Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021), Online, 2021. – P. 3450-3466. – doi: 10.18653/v1/2021.findings-acl.304.

[Vaswani et. al., 2017] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention Is All You Need // In: Proc. 31st Conference on Neural Information Processing Systems (NIPS'17), Long Beach, California, USA, 2017. – P. 6000-6010. – doi: 10.5555/3295222.3295349.

Abstracts of Plenary Reports

PRIORITY AREAS OF RESEARCH AND KEY TRENDS IN THE DEVELOPMENT OF AI TECHNOLOGIES

Yu.V. Visilter (*viz@gosniias.ru*)

GosNIIAS, Moscow

The trends in the development of AI methods and technologies at the current stage (2020-2025) are presented, grouped into several priority areas of AI research. The main trends and results in the following key areas and sub-areas are briefly described: LLM and other models for symbolic data, diffusion and other models for non-symbolic data, multimodal models, knowledge transfer methods with model adaptation, augmentation of LLM without model adaptation, reinforcement learning, agent-based and multi-agent systems, elements of general AI (AGI).

Keywords: Artificial Intelligence, Machine Learning, Computer Vision, NLP, LLM, Generative AI, RL, General AI.

SITUATIONAL MANAGEMENT: MODIFICATION OF A DECISION UNDER UNCERTAINTY

B.A. Kobrinskii (*kba_05@mail.ru*)

Federal Research Center "Informatics and Control"
of the Russian Academy of Sciences, Moscow

The paper considers a modified version of situational management, which assumes, instead of a decision based on random selection, the transfer of control to a hybrid intelligent system (HIS). The structure of an integrated system consisting of situational management technology and a HIS, including a rule base, a precedent library and a neural network, is presented. The possibility of obtaining context-dependent images by the user in the decision-making process is considered.

Keywords: situational management, hybrid intelligent system, integration of approaches, uncertainty situations, critical infrastructures.

AUTOMATIC MACHINE LEARNING FOR LARGE FUNDAMENTAL MODELS

A.V. Bukhanovsky (*avbukhanovskii@itmo.ru*)

NRU ITMO, Saint Petersburg

The intensive implementation of modern artificial intelligence technologies based on large fundamental models in various sectors of the economy and social sphere requires new mechanisms for customization and automation of the development of such solutions

for a specific task. If for artificial intelligence models using classical methods of working with data, this problem is successfully solved through the use of automatic machine learning (AutoML) technologies, then for large artificial intelligence models this issue remains open. The report proposes an approach to automating the design and training of multi-agent artificial intelligence systems based on large fundamental models that can take into account both the individual characteristics of agents and the topology of their interaction. To do this, we introduce a class of intelligent meta-agents capable of dynamically combining various application agents based on the principles of composite artificial intelligence. To implement this approach, it is proposed to use an "agent laboratory" – a software and hardware environment that provides agents with the ability to access data and calculations, as well as the effective use of various large language models. Based on the "agent laboratory", it is permissible, among other things, to simulate various cognitive effects inherent in systems of strong artificial intelligence. The approaches and technologies discussed will be illustrated by the challenges of creating large fundamental models for the oil and gas industry, urban planning, and supporting scientific activities.

Keywords: large fundamental models, automatic machine learning, multi-agent system.

DOMAIN ADAPTATION AND GENERALIZATION OF DEEP LEARNING MODELS

Abhinav Kumar (*abhinavkumar@ee.iith.ac.in*)

Indian Institute of Technology, Hyderabad, India

Deep learning models have been successful in areas such as robotics, medical imaging, and autonomous driving. However, their effectiveness often decreases when the domain is shifted, when the statistical properties of the deployment data differ from the training data. This shift can occur as a covariance, a priori, conditional, or mixed shift, and can significantly limit the reliability of the model in real pipelines. To address these issues, research has focused on unsupervised domain adaptation (UDA) when unmarked target data is available during training, and domain generalization (DG) when target data is unavailable. The report examines both unimodal (visual-only) and multimodal (visual-language) tasks, including single-source, multi-source, and multi-target UDA, as well as DG. Practical applications such as adapting RGB backlight to thermal imaging images for gesture recognition are also being considered. The report examines the limitations of existing adversarial and self-learning approaches and suggests targeted solutions for the development of UDA and DG methods. Deep learning models have been successful in areas such as robotics, medical imaging, and autonomous driving. However, their effectiveness often decreases when the domain is shifted, when the statistical properties of the deployment data differ from the training data. This shift can occur as a covariance, a priori, conditional, or mixed shift, and can significantly limit the reliability of the model in real pipelines. To address these issues, research has focused on uncontrolled domain adaptation when unlabeled target data is available during training, and domain generalization when target data is unavailable. The report examines both unimodal (visual only) and multimodal (visual language) tasks, including the adaptation of a subject area with

one source, multiple sources and multiple goals, as well as the generalization of the subject area. Practical applications such as adapting RGB backlight to thermal imaging images for gesture recognition are also being considered. The report examines the limitations of existing adversarial and self-learning approaches and suggests targeted solutions for developing domain adaptation and domain generalization methods.

Keywords: deep learning models, domain adaptations, domain generalization.

REASONING MODELS: PROS AND CONS

S.I. Nikolenko (*s.nikolenko@spbu.ru*)

Saint Petersburg State University, Saint Petersburg

2025 was the year of reasoning models for artificial intelligence. In October 2024, the o1-preview was released, in January 2025, DeepSeek released the R1 model, and almost immediately all the leading models became reasoning. In this report, we will talk about what reasoning models are, where they come from, and how they work. We will look at the latest results on how much reasoning really helps and how useful it is for other purposes, in particular for ensuring the security of artificial intelligence.

Keywords: reasoning models, security of artificial intelligence systems.

Abstracts of Sectional Reports

Section 1

Knowledge Engineering

CARTOGRAPHY AND SEMANTICS OF SCIENTIFIC KNOWLEDGE: A PILOT PROJECT

T.A. Gavrilova (*gavrilova@gsom.spbu.ru*) [A]
V. Shvankin (*v.shvankin@salesai.ru*) [B]
M.V. Kubelskiy (*m.kubelskiy@gsom.spbu.ru*) [A, B]
N.V. Ivanikova (*n.ivanikova@spbu.ru*) [C]
V. Luckov (*st098065@gsom.spbu.ru*) [A]

[A] Graduate School of Management, Saint-Petersburg State University
[B] Geropharm LLC, Saint-Petersburg
[C] Research Support Department, Saint-Petersburg State University

The paper presents a method for visualizing bibliometric data of researchers for one of the university subdivisions, taking into account the publications' semantics. By means of an in-depth analysis of the metadata of publications, including semantic analysis of titles and abstracts, a prototype of interactive maps of scientific interests and connections between authors was created. This approach helps identify key areas of research, interdisciplinary interactions and the structure of the research team, which contributes to more effective management of scientific activities and the development of cooperation within the university.

Keywords: knowledge maps, bibliometrics, semantic analysis.

ONTOLOGY-BASED ACCESS TO SYSTEMATIZED KNOWLEDGE AND RESOURCES ON MACHINE LEARNING

Yu.A. Zagorulko (*zagor@iis.nsk.su*) [A,B]
G.B. Zagorulko (*zagor@iis.nsk.su*) [A,B]
E.A. Sidorova (*lsidorova@iis.nsk.su*) [A,B]
I.O. Plotnikova (*i.plotnikova1@g.nsu.ru*) [B]

[A] A.P. Ershov Institute of Informatics Systems of SB RAS, Novosibirsk
[B] Novosibirsk State University, Novosibirsk

Despite the fact that the field of machine learning (ML) is actively developing, it is still poorly formalized, and the tools and resources developed within its framework are not sufficiently systematized. This not only prolongs the period of entry into the field of ML, but also makes it difficult for users to effectively select the tools and resources

necessary to solve their problems. This state of affairs in the field of ML necessitates the development of an information-analytical Internet resource that would provide systematization of knowledge and information resources on ML and content-based access to the tools, models, methods and data sets accumulated in this field. The paper describes an approach to the construction of such a resource based on the machine learning ontology developed by the authors.

Keywords: machine learning, ontology, information-analytical Internet resource, ontology design patterns.

METHODS OF CONSTRUCTING AN ECOSYSTEM OF KNOWLEDGE ON THE EXAMPLE OF ENERGY

L.V. Massel (*massel@isem.irk.ru*)
A.G. Massel (*amassel@isem.irk.ru*)
V.R. Kuzmin (*kuzmin_vr@isem.irk.ru*)

Melentiev Energy Systems Institute SB RAS, Irkutsk, Russia

The article considers an approach to building a knowledge ecosystem (on the example of energy) as an innovative approach to knowledge management. The architecture of a digital platform is proposed as the basis of a knowledge ecosystem. The authors' experience gained in building an IT infrastructure for systems research in the energy sector is used. The methods of building a knowledge ecosystem are determined by the architecture of the digital platform, partially tested in previous works by the authors. It is proposed to use the available author's scientific prototypes of tools that can become the basis of the services being developed.

Keywords: knowledge management, knowledge ecosystem, digital platform, energy.

CLASS-BASED TYPING OF NODES IN META-ASSOCIATIVE GRAPHS

A.E. Misnik (*anton@misnik.by*)

Inter-state educational institution of higher education
“Belarusian-Russian University” Belarus, Mogilev

The paper provides a theoretical foundation for class-based typing of nodes in meta-associative graphs—a hybrid knowledge-representation model that merges graph structures with object-oriented concepts. It details the formal composition of a node, the hierarchical and associative links among nodes, the mechanisms of inheritance and polymorphism, and the method for embedding a knowledge schema directly inside the model. The study demonstrates how explicit typing streamlines data-processing automation, optimizes query execution, and enhances the adaptability of information systems. Particular emphasis is placed on unified storage and access to knowledge, which enables

scalable ontology expansion without data migration. The findings broaden the theory of metagraphs and lay the groundwork for the development of semantically rich, high-performance knowledge-management systems.

Keywords: meta-associative graphs, knowledge engineering, ontologies, complex systems.

CATEGORY-THEORETIC APPROACH TO LEARNING BASED ON GENERALIZATION

V.L. Stefanuk (*ivan@ivanov.ru*)
A.V. Zhozhikashvili (*petr@petrov.ru*)

Institute for information transmission problems RAS, Moscow

The paper describes a new approach to machine learning – learning based on generalization. The paper presents a mathematical apparatus based on the language of category theory that can be used as a mathematical formalism for such learning.

Keywords: learning, generalization, pattern, category theory.

A LEARNING SITUATION MODEL FOR AN INTELLIGENT TUTORING SYSTEMS PLANNER

V.A. Uglev (*uglev-v@yandex.ru*) [A,B]

[A] Department of Intelligent Systems in the Humanities
of the Russian State University for the Humanities, Moscow
[B] Department of Applied Physics and Space Technologies,
Siberian Federal University, Zheleznogorsk

The paper describes a model of the learning situation, which is processed by the planner of an intelligent tutoring systems (ITS). The issues of its representation in the memory of the learning system and processing are discussed. Using the example of the educational process in the experimental ITS, an approach to processing the learning situation using the mechanism of expert systems and mapping tools is shown. The mechanism of how the interaction of different models interprets the learning situation in different ways and influences further decision-making is shown.

Keywords: Knowledge Engineering, e-learning, Intelligent Tutoring Systems, learning situation, Cognitive Maps of Knowledge Diagnosis.

SEMANTIC SPACE EMBEDDING OF SPEECH ACT INTENSIONS

D.L. Khabarov (*hdl001@campus.mephi.ru*)
A.V. Samsonovich (*avsamsonovich@mephi.ru*)

National Research Nuclear University MEPhI, Moscow

This work presents a semantic map of intensions, understood here as relational connotations of speech acts. The result is a tool that consists of a dataset of intensions, its embedding in a semantic space, and a graph of relations among the intensions, plus a neural network trained to recognize given intensions in utterances. The tool can be used for creating formal representations of social relational aspects of speech acts in a dialogue. The method of constructing the map is based on using OpenAI ChatGPT, fine-tuning a large language model (LLM), linear algebra, and graph theory. The constructed model of semantic space of intensions extends beyond the popular settings for sentiment or tonality analysis of texts in natural language. As a general model applicable to virtually any paradigm of social interaction, it can be used for constructing specialized models of limited paradigms. Therefore, the developed tool can enable efficient integration of LLMs with cognitive architectures, such as eBICA, for building socially emotional conversational agents.

Keywords: Semantic Mapping, LLM, BICA, DistilBERT, Neuro-Symbolic Integration, Social Intelligent Agents.

Section 2

Data Mining

SMALL DATA IS ALL YOU HAVE

A.V. Amentes (*Artem.amentes@yandex.ru*)

FRC "Computer Science and Control" RAS, Moscow

The paper describes an approach to solving problems related to expanding small datasets. The descriptions of small data sets, their sizes and limitations faced by researchers are given. The methods of data mining, deep learning and some other methods are also given. The JSM method is presented as an effective methodology for building data mining and predictive models, even with a very small data set size.

Keywords: small datasets, artificial intelligence, neural networks, big data, JSM method.

LATTICES REINFORCEMENT LEARNING

D.V. Vinogradov (*vinogradov.d.w@gmail.com*)

FRC "Computer Science and Control" RAS, Moscow

The paper presents a lattice-theoretic approach to generating rules that form a (sub-)optimal strategy for Reinforcement Learning. We argue for a return to the use of the method of Monte Carlo Tree Search. The probabilistic-combinatorial formal method based on lattice theory eliminates the main drawback of the Monte Carlo method – the lack of generalization ability. The problems of currently widely used neural network approaches will be discussed and the advantages of the Monte Carlo method will be indicated. Finally, a rigorous formalization of the proposed approach will be presented using Category Theory.

Keywords: Reinforcement Learning, Monte Carlo, lattices, Category Theory.

ON THE HEURISTIC POTENTIAL OF SOME COGNITIVE PROCEDURES

M.A. Mikheyenkova (*m.mikheyenkova@yandex.ru*)
S.M. Gusakova (*svem45@yandex.ru*)

FRC "Computer Science and Control" RAS

The paper discusses the challenges of formalizing research heuristics used to solve problems in open-ended empirical areas that lack a systematic formal framework. It considers the features of applying certain cognitive procedures of the JSM Method of

automated research support in accordance with the semantic and pragmatic specifics of a particular research context. Examples are given of adequate use of the heuristic potential of formal tools for obtaining interpretable results in several subject areas.

Keywords: exact epistemology, formalized heuristics, cognitive procedures, JSM Method, similarity operation, method of difference, situation approach.

ANALYSIS OF METHODS FOR ASSESSING THE IMPORTANCE OF PREDICTORS OF ADVERSE EVENTS IN CARDIAC SURGERY

B.V. Potapenko (*bvpotapenko@gmail.com*)[A]
K.J. Shakhgeldyan (*carinashakh@gmail.com*)[A,B]
B.I. Geltser (*boris.geltser@vvsu.ru*) [A,B]

[A] Vladivostok State University, Vladivostok, Russia
[B] Far Eastern Federal University, Vladivostok, Russia

This study investigates methods for assessing the importance of predictors in machine learning models. Both model-dependent and model-independent approaches were considered. The models were trained to predict the probability of mortality in the postoperative period for patients with ST-segment elevation myocardial infarction who underwent percutaneous coronary intervention. The results demonstrate a noticeable divergence in feature importance rankings depending on the applied methods. This is particularly evident for features that influence predictions non-linearly and are associated with other features. The study raises questions regarding the interpretation of feature importance in clinical medicine. The results indicate the advantage of using combined methods of importance assessment to increase trust in clinical decision support systems.

Keywords: explainable artificial intelligence, feature importance, clinical prognostic models, machine learning, clinical data mining.

DATA MINING METHODS DEVELOPMENT FOR OIL WELLS PRODUCTION WATER CUT FORECASTING

I.B. Fomynikh (*igborfomin@mail.ru*)
I.S. Mihailov (*fr82@mail.ru*)
K.O. Sidorov (*kirill.sidoroff2014@yandex.ru*)
Myo Hlaing Win (*myohlaingwin69287@gmail.com*)

National Research University “Moscow Power Engineering Institute”

The paper proposes a solution to the current problem of forecasting the water cut in oil well production using data mining methods. The following data mining methods are considered: LSTM, BiLSTM, Prophet, ARIMA, XGBoost, NNAR, and TBATS. Their strengths and weaknesses are identified. These methods are implemented and tested

using real data obtained from oil fields. It is shown that for short-term forecasting, it is preferable to use LSTM, BiLSTM, and NNAR models, while for long-term forecasting of changes in the water cut of well production, the LSTM model is more suitable.

Keywords: data mining, artificial neural network LSTM, forecasting, oil well production water cut.

ON THE CONTEXT INDEPENDENCE CONDITION IN THE JSM METHOD OF AUTOMATED RESEARCH SUPPORT

O.P. Shesternikova (*oshesternikova@frccsc.ru*)

Federal Research Center "Computer Science and Control"
of the Russian Academy of Sciences

The article describes the condition of independence of the cause from the context in the JSM-method of automated research support, proposed for the analysis of the situation of multiple causes (the existence of several sufficient component causes). An algorithm for checking this condition and an example for the medical research of establishing the appropriateness of referral for computed tomography in patients with chronic pancreatitis are given.

Keywords: JSM-method ARS, causality, sufficient component cause, multiple causes, chronic pancreatitis.

Section 3

Modeling Reasoning

ON THE POSSIBILITY OF GENERATING WELL-FOUNDED CAUSAL HYPOTHESES IN JSM-METHOD FOR THE PROBLEM OF HEREDITARY DISEASE

G.S. Velmakin (*grigoryyii@gmail.com*)

Federal Research Center "Computer Science and Control"
of the Russian Academy of Sciences, Moscow

In the paper, in JSM-method, using the inverse predicate of similarity as an exmplae, the predicate of similarity with composition of relations will be constructed. It will be shown that the constructed predicate can be expressed in Graph-FCA extensions of formal concept analysis by adding relationships between objects. In the end, using Graph-FCA, the relationship between the predicate of similarity with the composition of relations and description logic will be shown. All of the above will allow us to adequately describe the generation of causal hypotheses in the task of studying the causes of hereditary disease in thee JSM-method and Graph-FCA.

Keywords: JSM-method, predicate of similarity with composition of relations, Graph-FCA, description logic.

APPLICATION OF NEURAL NETWORK APPROACH AND PARALLEL INFORMATION PROCESSING TO DETERMINE THE FEASIBILITY OF BRANCHING-TIME LOGIC FORMULAS ON KRIPKE STRUCTURES

A.P. Eremeev (*eremeev@appmat.ru*)
N.Y. Filinov (*filinov.n@yandex.ru*)

NRU "MPEI", Moscow

In this paper, we present an approach to the satisfiability problem of temporal computational tree logic CTL formulas on Kripke structures using graph neural networks (GNN). The satisfiability task is framed as a binary classification problem over pairs (CTL formula, Kripke structure). We propose a model architecture combining graph and formula embeddings followed by a classifier. A synthetic dataset is generated and labeled using a classical model checking algorithm. Experimental results demonstrate high classification accuracy and significant runtime advantages over traditional model checking methods. This work is part of the development of tools for intelligent real-time decision support systems.

Keywords: artificial intelligence, temporal logic, model checking, Kripke structure, graph neural network, real time.

ON THE PROBLEM OF INTEGRATION OF STATISTICAL AND DETERMINISTIC METHODS OF EMPIRICAL DATA INTELLIGENT ANALYSIS

M.I. Zabezhailo (*m.zabezhailo@yandex.ru*)

FRC “Computer Science and Control” RAS, Moscow

Some possibilities to integrate statistical and deterministic methods used in the intelligent analysis of empirical data are presented. The approach is focused on identifying implicitly defined causal relationships linking the "cause" and the "context of its relevance" with the "effects" they "cause". An example of this approach application in the field of high-tech medical diagnostics is discussed.

Keywords: intelligent data analysis, ternary causality relation, medical diagnostics.

GUESSING, EMPIRICAL VERIFICATION AND EXPLANATION IN AUTOMATED PROBLEM SOLVING

S.S. Kurbatov (*curbatow.serg@yandex.ru*)

Research Centre of Electronic Computing NICEVT, Russia, Moscow

The difference in approaches to automated problem solving using LLM and based on reasoning modeling is analyzed. The analysis is carried out using the example of solving an Olympiad geometric problem. Such aspects as the transition from the problem text to a computer representation, interaction with a neural network in the solution process, the quality of the explanation of the proof, visualization and its role for empirical guesses are considered.

Keywords: LLM, modeling reasoning, olympiad problem.

THREE MORE QUESTIONS (FOR UNDERSTANDING), ADDRESSED TO "PARTY COMRADES"

V.K.Finn (*v.k.finn@yandex.ru*)
M.A.Mikheyenkova (*m.mikheyenkova@yandex.ru*)
M.I. Zabezhailo (*m.zabezhailo@yandex.ru*)

FRC “Computer Science and Control” RAS, Moscow

Some ideas about the intelligence of artificial intelligence (AI) systems, methods for assessing the quality of the results they generate, as well as some problems of organizing the expertise of projects and solutions in the field of AI are discussed.

Keywords: artificial intelligence, research and development, expertise, quality assessment of solutions generated by AI systems, personnel education and training.

Section 4

Text Mining, Large Language Models

AUTOMATIC ARGUMENT CLASSIFICATION BASED ON THE SYSTEMATIZATION OF D. WALTON'S ARGUMENTATION SCHEMES

I.R. Akhmadeeva (*i.r.akhmadeeva@iis.nsk.su*)
Yu.A. Zagorulko (*zagor@iis.nsk.su*)
I.S. Kononenko (*irina_k@cn.ru*)
A.S. Sery (*alexey.seryj@iis.nsk.su*)
E.A. Sidorova (*lsidorova@iis.nsk.su*)

A.P. Ershov Institute of Informatics Systems, SB RAS, Novosibirsk, Russia

This paper explores the application of transformer-based encoder models for the development of automatic argument classification methods. It introduces a categorization of D. Walton's argumentation schemes, comprising four classifiers corresponding to different levels or aspects of argumentative structure. Two approaches to solving the multiclass classification task were investigated: (1) a scheme prediction model and (2) a joint model predicting both the argumentation scheme and its category. The ru-en-RoBERTa model was used to obtain vector representations. Experiments were conducted on three annotated corpora: the Russian ArgNetSC corpus and two English-language corpora – Araucaria and NLAS (a corpus of automatically generated arguments). The best results on the Russian-language corpus reached an F1 score of 41,3%. The results for categories of arguments ranged from 55% to 89%.

Keywords: argumentation analysis, argument classification, multiclass classification, categorization of Walton's schemes, transformer model.

EXTRACTION OF FRUSTRATION REACTIONS FROM TEXT USING NEURAL NETWORK METHODS

D.A. Kireev (*kireev@isa.ru*)
Yu.M. Kuznetsova (*kuzjum@yandex.ru*)
N.V. Chudova (*nchudova@gmail.com*)
A.A. Chuganskaya (*anfi.chuganskaya@yandex.ru*)
I.V. Smirnov (*ivs@isa.ru*)

FRC «Computer Science and Control» RAS, Moscow

The article presents research results on the use of neural network models for automatic identification of frustration reactions in online discussions. Based on S. Rosenzweig's typology, a corpus of Russian-language texts containing reactions of various frustration types was developed and annotated by expert psychologists. Two

classification approaches were compared: a step-by-step approach (sequential determination of frustration presence, its direction, and type) and a single-stage approach (simultaneous determination of all classes). Experiments with transformer models (ruBert, ruElectra, ruRoberta) and modern language models (GPT, DeepSeek, Gemini, LLaMa, Claude) demonstrated the advantage and effectiveness of the single-stage approach. The results show that neural network models are capable of effectively modeling the work of a psycho-diagnostician with verbal display of frustration.

Keywords: web discussions, frustration reaction, neural networks, transformers, large language models.

RUSSIAN TEXTS COMPARISON USING COLLOCATION GRAPH TO IDENTIFY DISTINCTIVE SEMANTIC FRAGMENTS

N.V. Meleshchenko (*meleshenko.nikolay@mail.ru*)
O.I. Fedyaev (*olegfedyayev@mail.ru*)

Donetsk National Technical University, Donetsk

The paper considers the problem of updating university curricula, taking into account the requirements (recommendations) of enterprises. An approach based on the representation of the text in the form of a graph of phrases is proposed, which makes it possible to visualize the connections between terms and improve the analysis process. The difference from previous works is the use of a component tree instead of a dependency tree to formalize the extraction of phrases. The problems of normalization of extracted terms and processing conjunctions in the text have been solved, which contributes to a more accurate definition of phrases. The conducted experimental studies confirm the effectiveness of the proposed approach.

Keywords: natural language, text comparison, enterprise requirements, discipline curricula, collocation graph, term normalization, semantic fragments.

COMPARATIVE ANALYSIS OF KEYWORD EXTRACTION METHODS ON SCIENTIFIC PUBLICATION COLLECTIONS

N.A. Nazarov (*straider105@gmail.com*)
M.R. Sharifullin (*sharifullin2107@mail.ru*)
V.O. Tolcheev (*tolcheevvo@mail.ru*)

National Research University «MPEI», Moscow

This paper analyzes applied tasks of intelligent text data analysis, for which the extraction of keywords (KWs) plays a crucial role. It is noted that KWs are most frequently and effectively used for processing and analyzing English-language scientific documents (both full-text and bibliographic). A comparison of keyword extraction performance (F1@K metric) was conducted on widely known and publicly available collections of scientific textual data (Inspec, SemEval-2010, Krapivin), as well as on a dataset of news articles (DUC2001). The experimental studies considered various keyword

extraction technologies, including the statistical algorithm YAKE, graph-based algorithms TopicRank and MultipartiteRank, and neural network models KeyBERT and PromptRank. The parameters of these methods were tuned, the dependency of these parameters on document length was analyzed, and quality metrics were evaluated.

Keywords: rules, keyword extraction, text mining, scientific publications, method comparison, graph-based algorithms, neural network models, statistical analysis, bibliographic data, data visualization.

COMPARATIVE ANALYSIS OF TRANSFORMER MODELS FOR TEXT SIMPLIFICATION

N.A. Prokopyev (*nikolai.prokopyev@gmail.com*)
O.A. Nevzorova (*onevzoro@gmail.com*)
F.M. Gafarov (*fgafarov@yandex.ru*)
A.A. Gafiatullin (*arslan2911@mail.ru*)
A.R. Ziastinov (*ziastinovalmaz@gmail.com*)

Kazan Federal University, Kazan

This paper explores the text simplification problem in Russian using Transformer architecture models. Best model is selected using standard evaluation metrics BLEU, ROUGE, SARI and a refined lexical complexity metric based on readability evaluation. Pre-trained models T5 and BART were studied, training was performed on a dataset in computer science domain. It was found that BART model is superior in text simplification and generating texts are more consistent on structure with references.

Keywords: lexical complexity, metric, Transformer architecture, dataset.

COMPARISON OF APPROACHES TO INTERPRETING LANGUAGE MODELS: ANALYSIS OF A MASKED LANGUAGE MODELING-BASED METHOD AND TRADITIONAL TECHNIQUES

A.A. Rogov (*rogov.alisher@gmail.com*) [A]
N.V. Loukachevitch (*louk_nat@mail.ru*) [A,B]

[A] Bauman Moscow State Technical University, Moscow
[B] Lomonosov Moscow State University, Moscow

As pretrained language models such as BERT continue to grow in complexity and application – from recommendation systems to medical diagnostics – the demand for effective interpretation methods becomes increasingly important. Understanding how these models make decisions is crucial for building trust and ensuring their reliable use. In this work, we investigate approaches to explaining the behavior of BERT in text classification tasks. The focus is on comparing two paradigms: modern techniques based on masked language modeling and prompt-based learning, and traditional methods such as

LIME and vector similarity-based approaches. The experimental analysis is conducted on the Web of Science and 20Newsgroups datasets. Interpretation quality is evaluated using activation plots, which visualize the significance of input tokens in the final prediction.

Keywords: neural network model interpretation, LIME method, masked language modeling, verbalizer, text classification.

ARGUMENTS CLASSIFICATION USING LARGE LANGUAGE MODELS

E.A. Sidorova (*lsidorova@iis.nsk.su*)
A.S. Sery (*alexey.seryj@iis.nsk.su*)
I.R. Akhmadeeva (*i.r.akhmadeeva@iis.nsk.su*)
D.V. Ilina (*dviljina@gmail.com*)

A.P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia

The article is devoted to the development of methods for automatic classification of arguments in Russian-language texts using LLM and prompt engineering. The study was conducted on three datasets with argumentation markup using D. Walton's schemes: the NLAS corpus of automatically generated English-language arguments, the Araucaria corpus of annotated English-language arguments, and the Russian-language corpus of annotated arguments ArgNetSC. Three strategies were applied for argument classification: (1) classification using Walton's schemes with formal definitions, (2) classification based on systematization of argumentation schemes, and (3) sequential inference through dialogue. Research using Mistral family models showed that the most effective approach is the dialogue-based communication model with LLM and the strategy of automatic selection of semantically similar examples for Few-Shot technique. The best scores were 0.63 and 0.31 F_1-measure for English-language corpora and 0.15 for the Russian-language corpus. To investigate the quality of preliminary class filtering, the DeepSeek-R1-Distill-Llama-70B model and the best strategy obtained in the first stage were used. The obtained results allow us to conclude that a generative model can be applied as an assistant for annotation or automatic argument classification for preliminary filtering. The F1-measure for scheme classes was 0.615, and the true argumentation scheme was included in the filtered list of schemes in 78.5% of cases.

Keywords: argument analysis, argument classification, Walton’s argumentation schemes, systematization of argumentation schemes, prompt engineering.

RATIONALIZATION AND OVERRELIANCE ON XAI TOOLS: AN ANALYSIS STUDY OF LARGE LANGUAGE MODEL EXPLANATION

A.V. Suvorova (*asuvorova@hse.ru*)

HSE University, St. Petersburg

This study investigates the issue of users' overreliance on machine learning model interpretations and examines how explanations generated by large language models (LLMs) affect this tendency. Our experimental findings reveal that most LLMs either overlooked anomalies in the data or produced plausible yet misleading explanations, rationalizing model outputs reproducing user behaviour from initial experiment. These results underscore the potential dangers of employing LLMs for model interpretation purposes without implementing proper validation procedures.

Keywords: explainable AI, machine learning, user evaluation.

SENTIMENT ANALYSIS TASKS FOR NEWS TEXTS

S.O. Urazov (*urazov.msu@gmail.com*)
N.V. Loukachevitch (*louk_nat@mail.ru*)

Lomonosov Moscow State University, Moscow

This paper is dedicated to our research of the application of large language models (LLM) in various formulations of sentiment analysis task, such as finding opinions of and about a specific person; classifying (as negative, neutral or positive) the sentiment of an opinion towards a target; and extracting the relationship between two entities in a text. The study was conducted on the RuOpinionNE-2024 Russian news texts dataset using neural network models such as BERT and Ruadapt-Qwen2.5. A series of experiments were conducted using model learning techniques and prompts.

Key words: sentiment analysis, large language models, prompt engineering, LoRA fine tuning.

THE APPLICABILITY OF METHODS FOR EVALUATION OF SYNTACTIC PARSING TO THE RESULTS OF A LLM BASED PARSER

E.D. Shamaeva (*derinhelm@yandex.ru*)
N.V. Loukachevitch (*louk_nat@mail.ru*)

Lomonosov Moscow State University, Moscow

The use of large language models for the syntactic parsing is the reason for significantly new cases. For some sentences, the parsing fails with an error. For others, the original sentence is changed. The applicability of parser evaluation methods to these

sentences is explored in this article. The experiment was performed on parser U-DepPLLaMA and a test sample of Taiga treebank. It was found that calculating the average value of UAS and LAS metrics on test sentences is inapplicable for these sentences. Also the standart alignment algorithm is not applicable. The study program is available on the website https://github.com/Derinhelm/parser_stat/tree/llm_taiga.

Keywords: syntax parsing, dependency tree, large language models, tokenization.

AUTOMATIC KEYPHRASE GENERATION FOR RUSSIAN SCIENTIFIC TEXTS USING PSEUDO-LABELING AND CONTRASTIVE LEARNING

K.Sh. Yaushev (*kyaush@mail.ru*) [A]
N.V. Loukachevitch (*louk_nat@mail.ru*) [B]

[A] Bauman Moscow State Technical University, Moscow
[B] Lomonosov Moscow State University, Moscow

This paper presents a multi-stage approach to automatic keyphrase generation for Russian-language scientific articles. The method is based on fine-tuning transformer models using pseudo-labeling and contrastive learning, and incorporates filtering of generated candidates. Two pseudo-labeling strategies and a bi-encoder architecture for relevant keyphrase selection are proposed. Experiments on a corpus of mathematics and computer science articles demonstrate the superiority of the proposed approach over classical and neural methods in terms of F1, ROUGE-1, and BERTScore metrics.

Keywords: keyphrase generation, automatic keyphrase annotation, contrastive learning, keyphrase filtering.

АВТОРСКИЙ УКАЗАТЕЛЬ

СОДЕРЖАНИЕ

ПЛЕНАРНЫЕ ДОКЛАДЫ

Секция 1. ИНЖЕНЕРИЯ ЗНАНИЙ

Секция 2. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Секция 3. МОДЕЛИРОВАНИЕ РАССУЖДЕНИЙ

Секция 4. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ, БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ

Научное издание

Двадцать вторая Национальная конференция
по искусственному интеллекту с международным участием
КИИ-2025
Труды конференции в 3-х томах
Том 1

Подписано в печать 23.09.2025 г.
Формат 60x84 1/16. Тираж 300 экз. Усл. печ. л. 21,9.

Издательство СПб ФИЦ РАН

ISBN 978-5-6052274-4-1