DOI: 10.15514/ISPRAS-2021-33(4)-8

Синтаксический анализ текстов предметной области при помощи онтологии

[1] *Б.И. Гельцер, ORCID: 0000-0002-9250-557X <boris.geltser@vvsu.ru>*
[2] *Т.А. Горбач, ORCID: 0000-0003-4380-6517 <tagorbachdv@gmail.com>*
[2] *В.В. Грибова, ORCID: 0000-0001-9393-351X <gribova@iacp.dvo.ru>*
[3] *О.В. Карпик, ORCID: 0000-0002-0477-1502 <parlak@mail.ru>*
[4] *Э.С. Клышинский, ORCID: 0000-0002-4020-488X <eklyshinsky@hse.ru>*
[4] *Н.А. Кочеткова, ORCID: 0000-0002-5346-0081 <nkochetkova@hse.ru>*
[2] *Д.Б. Окунь, ORCID: 0000-0002-6300-846X <okdm@iacp.dvo.ru>*
[2] *М.В. Петряева, ORCID: 0000-0002-1693-4508 <margaret@iacp.dvo.ru>*
[5] *К.И. Шахгельдян, ORCID: 0000-0002-4539-685X <carina.shahgeldyan@vvsu.ru>*

[1] *Дальневосточный Федеральный университет*
690922, Владивосток, о. Русский, п. Аякс, 10
[2] *Институт автоматики и процессов управления ДВО РАН*
690041, Владивосток, ул. Радио, д. 5
[3] *Институт прикладной математики им. М.В. Келдыша РАН*
125047, Москва, Миусская пл., д. 4
[4] *Национальный исследовательский университет «Высшая школа экономики»*
105066, Ст. Басманная ул., д.21/4, стр. 1
[5] *Владивостокский государственный университет экономики и сервиса*
690014, Владивосток, ул. Гоголя, д.41

Аннотация. В работе проводится сравнение трех методов синтаксического анализа текстов жалоб пациентов, извлеченных из электронных медицинских карт. В качестве контрольного теста используются существующие библиотеки синтаксического анализа текста. В качестве альтернативы предлагается использование онтологии для исправления ошибок, допущенных синтаксическим анализатором, либо полное формирование синтаксических зависимостей по данным, хранимым в онтологии. В статье показано что ограниченный набор правил, описывающих управление падежами зависимых слов, может показывать точность, сопоставимую с точностью современных синтаксических анализаторов, основанных на нейронных сетях.

Ключевые слова: синтаксический анализ; поверхностно-синтаксический анализ; онтологии; медицинские тексты

Для цитирования: Гельцер Б.И., Горбач Т.А., Грибова В.В., Карпик О.В., Клышинский Э.С., Кочеткова Н.А., Окунь Д.Б., Петряева М.В., Шахгельдян К.И. Синтаксический анализ текстов предметной области при помощи онтологии. Труды ИСП РАН, том 33, вып. 4, 2021 г., стр. 99-116. DOI: 10.15514/ISPRAS–2021–33(4)–8

Благодарности. Данная работа поддержана грантом РФФИ 18-29-03131.

Ontology-based syntactic analysis of domain-specific texts

[1] *B.I. Geltser, ORCID: 0000-0002-9250-557X <boris.geltser@vvsu.ru>*
[2] *T.A. Gorbach, ORCID: 0000-0003-4380-6517 <tagorbachdv@gmail.com>*
[2] *V.V. Gribova, ORCID: 0000-0001-9393-351X <gribova@iacp.dvo.ru>*
[3] *O.V. Karpik, ORCID: 0000-0002-0477-1502 <parlak@mail.ru>*
[4] *E.S. Klyshinskiy, ORCID: 0000-0002-4020-488X <eklyshinsky@hse.ru>*
[4] *N.A. Kochetkova, ORCID: 0000-0002-5346-0081 <nkochetkova@hse.ru>*
[2] *D.B. Okun, ORCID: 0000-0002-6300-846X <okdm@iacp.dvo.ru>*
[2] *M.V. Petryaeva, ORCID: 0000-0002-1693-4508 <margaret@iacp.dvo.ru>*
[5] *K.I. Shakhgeldyan, ORCID: 0000-0002-4539-685X <carina.shahgeldyan@vvsu.ru>*

[1] *Far Eastern Federal University*
10 Ajax Bay, Russky Island, Vladivostok, Russia, 690922
[2] *Institute of Automation and Control Processes, Far Eastern Branch of RAS*
5 Radio st., Vladivostok, Russia, 690041,
[3] *Keldysh Institute of Applied Mathematics*
4 Miusskaya square, Moscow, Russia, 125047,
[4] *HSE University*
21/4 building 1, Staraya Basmannaya st., Moscow, Russia, 105066,
[5] *Vladivostok State University of Economics and Service*
41 Gogolya st., Vladivostok, Russia, 690014

Abstract. The paper compares three methods for parsing of patients' chief complaints extracted from electronic medical cards. We propose two methods which are based on usage of an ontology: either as a method for correction of mistake made by a parser, or for constructing syntactical dependencies according to this ontology and a limited set of rules of syntactical governance. As a control test, we use existing natural text parsing libraries. The paper demonstrates that such a simple approach could achieve a high accuracy, which is comparable to modern parsers.

Keywords: parsing; shallow parsing; ontology; medical texts

For citation: Geltser B.I., Gorbach T.A., Gribova V.V., Karpik O.V., Klyshinskiy E.S., Kochetkova N.A., Okun D.B., Petryaeva M.V., Shakhgeldyan K.I. Ontology-based syntactic analysis of domain-specific texts. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 4, 2021, pp. 99-116 (in Russian). DOI: 10.15514/ISPRAS-2021-33(4)-8

Acknowledgments. This work was supported by the grant of RFBR no. 18-29-03131.

1. Введение

Медицинские информационные системы позволяют вывести общение доктора и пациента на новый уровень. Ситуация, когда информация об одном пациенте стекается из разных медицинских центров в одну точку, являющуюся сосредоточием самой полной информации об этом пациенте, удобна всем участникам. Становятся возможны консультации с врачами в других городах и странах без потери данных об истории болезни пациента. Одной из тем, активно развивающихся в рамках данного направления, является автоматическая обработка медицинских текстов, помогающая извлекать из медицинских карт пациентов информацию о течении заболевания. Полученные данные могут использоваться для решения различных задач: помощь в диагностике пациента по проявляемым симптомам и результатам обследования; поиск взаимосвязей между симптомами, диагнозом и прописанными лекарствами [1]; автоматизация оценки эффективности применения лекарств по результатам повторного обследования и т. д.

Одним из методов обработки медицинских текстов является извлечение фактов, суть которого состоит в выделении объектов и событий, а также взаимосвязей между ними.

Извлечение фактов, в свою очередь, базируется на извлечении сущностей (в том числе, именованных), которое помогает найти термины, названия лекарств, имена пациентов и врачей, единиц измерения и т.д. [2] Для нахождения связей между извлеченными сущностями используется синтаксический анализ, именно он помогает понять логику взаимодействия между извлеченными сущностями и суть производимых ими действий. Если отвлечься на другую предметную область, то недостаточно понять, что была произведена сделка между компаниями A, B и C, в которую были также вовлечены акции; необходимо понять, какая из компаний приобрела акции другой компании у третьей. Аналогичные проблемы ставятся и при анализе медицинских текстов. Например, требуется понять какая доза какого из лекарств была прописана при каком виде боли в какой именно части тела.

Обычно для синтаксического анализа применяют соответствующие программные библиотеки, точность работы которых достигает на реальных текстах порядка 95%. Такая точность достигается за счет использования нейронных сетей. Чаще используются ячейки LSTM [3], некоторые системы строятся по архитектуре трансформеров [4] с использованием уровня внимания [5].

Точность анализа медицинских текстов гораздо ниже. Это связано с тем, что медицинские тексты написаны весьма специфичным языком, с большим трудом поддающимся анализу: предложения могут строиться без применения глаголов; используемая терминология сложна, термины представляют собой длинные последовательности слов с большой глубиной подчинения; тексты содержат длинные перечисления подобных сложных терминов. В итоге, точность работы синтаксических анализаторов падает до 80-85%. Следствием низкой точности являются сложности в практическом применении методов анализа медицинских текстов. Извлекаемая из них информация служит входом для прочих этапов (диагностики, расчета статистики и т. д.). Получив на вход неточную информацию, эти этапы сами будут выдавать некорректные или неполные результаты, что вступает в противоречие высокой точностью анализа, который так важен в области медицины. Таким образом, требуется разработка новых методов синтаксического анализа медицинских текстов, которые позволили бы повысить его точность.

В данной статье мы хотели бы вернуться к старой идее из области автоматической обработки текстов, заключающейся в применении глубинной семантики в синтаксическом анализе. Текст содержит в себе некоторую семантику и логику повествования, которые выражаются в связях между словами. Аналогичные связи обычно закладываются в тезаурусы или онтологии. Следовательно, для синтаксического анализа текста может использоваться информация о семантических отношениях между терминами предметной области, описанных в онтологии этой предметной области. В чистом виде такой подход плохо применим на практике, так как для синтаксического анализа текста на любом языке необходимо иметь хотя бы некоторые представления о синтаксисе данного языка: преимущественном направлении связей между словами для разных конструкций и видов подчинения слов, согласовании слов в определенных конструкциях, глагольном и именном управлении слов и т. д. Особенности языков с развитой системой словоизменения, например, русского, польского или финского, требуют проведения морфологического анализа и снятия грамматической неоднозначности слов. Игнорируя эти особенности языка, мы не сможем получить качественную систему синтаксического анализа.

В данной работе мы предлагаем смешанную методику синтаксического анализа текста, опирающуюся на использование богатой онтологии предметной области и поверхностного синтаксического анализа (упрощенной и неполной его версии). Здесь мы используем тот факт, что два термина, связанные в тексте, должны быть связаны и в онтологии, а сам факт их связи будет выражен при помощи предлогов и грамматического согласования или управления. Мы утверждаем, что для проведения подобного анализа требуется короткий список правил и база информации об именном управлении зависимыми словами, однако сам метод подходит лишь для анализа коротких текстов узкой предметной области. Материалом для экспериментальной

части исследования послужили тексты жалоб пациентов. Тексты были записаны врачами и являлись частью истории болезни пациента.

2. Существующие решения

Прежде чем описать существующие решения в области анализа медицинских текстов, более внимательно рассмотрим сам этап синтаксического анализа. Целью данного этапа является построение дерева зависимостей, показывающего связи между словами в предложении. Так для фразы «Он видел их семью своими глазами» мы получим следующие связи: некто «он» производил действие «видеть», направленное на «семью», «семья» была «их», действие производилось при помощи инструмента «глаза», имеющего свойство «свои». Заметим, что фраза является неоднозначной, и альтернативный результат ее разбора будет следующим: действие направлено на «них» (видел их), действие производилось при помощи инструмента «глаза», количество глаз равно семи («семью глазами»), остальные связи совпадают. В обоих случаях семантическая информация об окружающем мире используется корректно («он» может производить действия, «семья» может являться объектом материального мира и ее можно увидеть, «семья» может относиться к кому-то, «глаза» могут кому-то принадлежать и быть в определенном количестве), равно как и правила языка (субъект действия предшествует глаголу, объект – следует за ним, свойства, выраженные прилагательным или местоимением, идут перед главным словом и согласовываются по грамматическим признакам). Если из полученных связей построить дерево, то мы получим структуру, называемую деревом зависимостей.

Как было сказано выше, современные системы синтаксического анализа строятся с использованием нейронных сетей. Заметим, что, применяя примерно одинаковый инструментарий, системы синтаксического анализа работают с разной эффективностью и скоростью. Альтернативой нейронным сетям являются поверхностный синтаксический анализ и сегментация [6, 9]. Они применяются, когда нет необходимости проводить полный синтаксический анализ всего предложения, а достаточно обнаружить границы фрагмента, например, при извлечении терминов, именованных сущностей или фактов. В такой ситуации может быть использована сегментация, которая даже не строит дерева зависимостей. Сегментация отличается более высокой скоростью работы, связанной с уменьшением числа правил или использованием более простых методов, например, скрытых Марковских моделей [7], условных случайных полей [8], контекстно-свободных грамматик или конечных автоматов [9]. В отличие от сегментации, поверхностный синтаксический анализ восстанавливает структуру зависимостей в фразе или предложении, используя сходные инструменты. Однако в его задачи не входит восстановление дерева зависимостей всего предложения или текста.

Применение поверхностно-синтаксического анализа позволяет сосредоточиться на определенных аспектах языка, не решая задачу анализа в общем случае. Это помогает упростить анализ конкретных языковых явлений, но требует написания или автоматической генерации набора правил. Такая работа занимает довольно продолжительное время и требует значительного ручного труда.

Как отмечалось выше, наша идея состоит в проведении синтаксического анализа с использованием знаний из онтологий. В целом, построение онтологий описано, например, в таких фундаментальных трудах, как [10, 11]. Обращаясь к онтологиям медицинской области, следует заметить, что успехи здесь достигли впечатляющих размеров: в связи с важностью и актуальностью предметной области, медицинские онтологии являются самыми проработанными среди всех. Самой большой из медицинских онтологий является Unified Medical Language System (UMLS) [12]. Она содержит в себе такие подсистемы, как Metathesaurus (иерархию понятий, собранных из различных словарей), Semantic Network (отношения между понятиями и категориями) и SPECIALIST Lexicon and Lexical Tools (большой словарь биомедицинских терминов и слов общего английского языка, используемый специально разработанными

инструментами для анализа текстов). Metathesaurus, называемый также MeSH (Medical Subject Headings), использовался, например, в системе MetaMap, предназначенной для извлечения данных из медицинских текстов [13]. Алгоритм MetaMap состоит из двух этапов: обработка медицинских текстов с извлечением фактов и уточнение извлеченных понятий. Первый этап содержит в себе такие стандартные шаги, как токенизация, синтаксический анализ, поиск сокращений и аббревиатур, поиск неоднословных терминов и проч. Результатом работы является размеченный медицинских текст, содержащий в себе ссылки на Metathesaurus. Сходный подход, основанный, однако, на поверхностном синтаксическом анализе, использован в работе [14]. Словарь Metathesaurus был переведен на 15 языков, считая русский [15]. Так система Exactus [16], использующая переведенную версию UMLS, проводит логический вывод относительно течения хронических заболеваний. Её алгоритм машинного обучения позволяет поднять точность извлечения фактов до 82% для определения тяжести заболевания и до 99% для течения заболевания.

Структура описанных выше онтологий предполагает объединение понятий в тематические группы или построение иерархии подобных групп, однако в них находит слабое отражение информация о связях между словами текста. В результате, такие онтологии хорошо подходят для выделения фактов или поиска сущностей, но плохо – для поиска связей между понятиями. В следующем разделе мы опишем структуру используемой нами онтологии, структура которой подходит для решения поставленных задач.

3. Используемые данные и инструменты

В нашем исследовании мы использовали «Базу медицинской терминологии и наблюдений» [17], которая обладает несколькими полезными для нас свойствами. Помимо описания отдельных понятий, данная онтология включает в себя три основных типа сущностей: признаки, характеристики и значения. Признаки объединяют характеристики в смысловые группы. Характеристики показывают текущий функциональный статус пациента и связаны с множеством принимаемых значений. Значения описывают течение заболевания и могут быть качественными, числовыми или интервальными. База медицинской терминологии и наблюдений имеет форму дерева, в котором значения являются листьями, подчиняющимися названию характеристики. Название характеристики подчиняется названию признака, все названия признаков подчинены общей вершине, описывающей жалобы пациентов. Например, *Боль в ноге → Локализация → правая нога*. Каждому термину может быть сопоставлен набор его синонимов или заменителей.

Онтология также содержит в себе другие разделы, например, раздел симптомов, в котором хранится информация о видах исследований. В данной работе мы использовали только секцию жалоб пациентов, описывающую субъективное мнение и ощущения пациентов. Этот раздел характеризует самочувствие пациентов и состояние систем организма: нервной, дыхательной, опорно-двигательной и др. Раздел жалоб содержит в себе подраздел «Общие жалобы», описывающие симптомы заболеваний: слабость, головокружение, тошнота, потливость и др., подраздел «Боли», включающий в себя головную боль, боль в спине, боль в шее, боль в горле и др., и другие подразделы. Для признаков определены такие характеристики как, например, «локализация», «причина», «частота», «время возникновения» и др.

Заметим, что среди прочих в онтологиях используется два вида связи: гипоним/гипероним и мероним/холоним. Гипоним выражает более частную сущность, чем данная. Гипероним, наоборот, показывает более общую сущность (то есть обратен гипониму). На некотором уровне абстракции можно сказать, что базовый класс является гиперонимом по отношению к наследуемому от него гипониму. Меронимы и холонимы отражают отношения часть-целое. Мероним является составной частью холонима (целое по отношению к мерониму). Так автомобиль является холонимом по отношению к меронимам двигатель, капот, руль и др.

Помимо этих двух видов отношений, вводится целый ряд других: функциональные, пространственные, временны́е, атрибутивные и др. [10] В используемой нами онтологии «База медицинской терминологии и наблюдений» используется несколько видов отношений. Так, например, связь между названием характеристики и его значением является атрибутивной (*Локализация → Правый глаз*), между признаком и характеристикой – функциональной (*Боль в глазу → Локализация*), между группой признаков и признаком – гиперонимической (*Боль → Боль в глазу*). Как это отмечалось выше, наша онтология представляет собой дерево (а не граф, как это бывает во многих онтологиях), содержащее в себе ссылки на другие ветви графа, используемые в разных местах. Так если локализация нескольких видов боли может совпадать, для них будет построено единственное поддерево, ссылки на которое будут размещены в соответствующих местах. Подобный подход помогает избегать дублирования информации.

Важным элементом наших экспериментов являлось использование программных библиотек синтаксического анализа, для которых предусмотрена возможность работы с русским языком: UDPipe версии 2.5 (выпущен в декабре 2019) [3] и spaCy 3.0 (выпущен в марте 2021) [18]. Оба синтаксических анализатора используют нейронные сети на основе LSTM. Языковая модель UDPipe была обучена только на русскоязычной части корпуса Universal Dependencies [19]. SpaCy обучалась на нескольких корпусах и обладает целым рядом моделей, лучше приспособленными для анализа текстов, написанных в разных стилях. Для своих экспериментов мы использовали модель «ru_core_news_sm», натренированную на новостной ленте и показывающую по отзывам создателей более точные результаты.

4. Алгоритмы синтаксического анализа медицинских текстов

В данной работе мы изложим алгоритмы работы трех методов разбора жалоб пациентов: синтаксический анализ при помощи стандарт ной библиотеки с коррекцией результатов при помощи данных из онтологии, поверхностно- синтаксический анализ на основании связей из онтологии и, в качестве контрольного метода, обычный синтаксический анализ без применения онтологии. В следующем разделе будет дана количественная оценка работы указанных методов.

4.1 Синтаксический анализ с коррекцией результатов при помощи онтологии

Как отмечалось выше, тексты жалоб пациентов пишутся весьма специфическим языком, анализ которого приводит к большому числу ошибок, связанных с соединением слов в дерево зависимостей. Причиной таких ошибок является некорректное взвешивание синтаксических связей анализатором, или следование наиболее вероятному решению, которое в данном конкретном тексте будет неверным. Неправильные синтаксические связи влекут ошибки в семантике, например, подчинение значения значению или характеристике из другой ветви онтологии. Итогом может быть дерево зависимостей, являющееся корректным с точки зрения синтаксиса (так как возможны несколько вариантов разбора данного текста), но некорректным с точки зрения семантики, так как подобные связи с ее точки зрения запрещены или противоречат связям онтологии.

Основная идея данного метода состоит в следующем. Требуется получить результаты синтаксического анализа текста, а потом откорректировать их в соответствии с иерархией терминов, хранимой в онтологии. Суть алгоритма состоит в последовательном перемещении вершин по дереву. Если есть связь между двумя терминами из разных ветвей онтологии, подчиненный термин должен быть перемещен выше по дереву с тем, чтобы найти там начало корректной ветви. Если мы видим некорректное подчинение терминов из одной ветви, следует исправить ситуацию хотя бы частично.

Предлагаемый алгоритм включает в себя следующие этапы: токенизация, морфологический анализ, извлечение терминов, синтаксический анализ, коррекция полученного дерева зависимостей. На этапе токенизации предложение разбивается на отдельные слова. Этап морфологического анализа приписывает этим словам грамматические характеристики (род, число и др.). Этап извлечения терминов анализирует последовательность токенов и выделяет из них непересекающиеся последовательности слов, являющиеся терминами. Далее все неоднословные термины будут анализироваться как единые элементы. В нашем случае нет необходимости решать проблему поиска новых терминов или снятия их многозначности, так как термином может являться только набор слов, присутствующий в онтологии. Все остальные слова и словосочетания рассматриваются как «заполнители», не имеющие отношения к предметной области. То есть здесь мы будем исходить из предположения о полноте онтологии предметной области, даже если это предположение неверно. Основными проблемами здесь являются возможное пересечение терминов, пропуск в них слов или написание слов с ошибкой. Под пересечением терминов мы понимаем ситуацию, когда начало и конец последовательности могут быть отнесены к разным терминам, а слова в середине относятся к ним обоим. В таком случае надо принять решение какой из терминов будет выделен из текста. Последними этапами алгоритма являются синтаксический анализ и коррекция дерева зависимостей по алгоритму, описанному ниже.

Здесь мы будем полагать, что два понятия из дерева зависимостей могут иметь общую дугу только случае, когда в онтологии есть прямой восходящий или нисходящий путь от одной вершины к другой. В противном случае дерево зависимостей должно быть исправлено при помощи следующих правил.

- Если родительская вершина в дереве зависимости имеет более низкий уровень в онтологии, чем дочерняя, следует поменять местами эти две вершины.
- Если в онтологии отсутствует прямой путь между двумя соединенными вершинами, следует подчинить дочернюю вершину следующему родителю, находящемуся на один уровень выше.
- Пусть две вершины дерева зависимостей подчинены одной родительской вершине, при этом одна из вершин является потомком другой вершины. В таком случае, первую вершину надо подчинить второй.

Эти правила применяются к дереву зависимо стей до тех пор, пока процесс не сойдется. После перестроения, дерево зависимостей должно соответствовать общей иерархической структуре онтологии, то есть родители и потомки должны относиться к одним и тем же ветвям онтологии и иметь корректный порядок подчинения. Часть вершин, которые отсутствуют в онтологии, должны оказаться на самом верхнем уровне дерева. Для остальных вершин факт связи между двумя вершинами будет означать, что в онтологии существует прямой путь вниз от родительской вершины к дочерней.

Рассмотрим пример работы алгоритма. Пусть дано следующее предложение: «*жалобы на выраженную боль в поясничном отделе позвоночника, ограничение движений в пояснично-крестцовом отделе, нарушение ходьбы, нарушение функции тазовых органов»*. После нахождения и объединения терминов мы получим следующую последовательность: «[***жалобы***] *на* [***выраженную***] [***боль***] *в* [***поясничном отделе позвоночника***], [***ограничение движений***] *в* [***пояснично-крестцовом отделе***], [***нарушение ходьбы***], [***нарушение функции тазовых органов***]». После синтаксического анализа будет получено следующее дерево зависимостей.


```
root
└─ 1 жалобы
   ├─ 4 боль
   │  ├─ 2 на
   │  └─ 3 выраженную
   ├─ 6 поясничном отделе позвоночника
   │  └─ 5 в
└─ 8 ограничение движений
   ├─ 10 пояснично-крестцовом отделе
   │  └─ 9 в
   ├─ 12 нарушение ходьбы
   │  └─ 14 нарушение функции тазовых
   │     органов
```

Вершины 1 и 8 находятся на одном уровне в дереве, но 8 является признаком в разделе 1. Следовательно, 8 необходимо подчинить 1. Вершина 6 является значением вершины 4, то есть первая должна быть подчинена второй. Вершина 12 является «братом» вершины 8, то есть должна быть поднята на один уровень вверх. После применения всех преобразований, дерево выглядит следующим образом.

```
root
└─ 1 жалобы
   ├─ 4 боль
   │  ├─ 2 на
   │  ├─ 3 выраженную
   │  └─ 6 поясничном отделе позвоночника
   │     └─ 5 в
   ├─ 8 ограничение движений
   │  ├─ 10 пояснично-крестцовом отделе
   │  │  └─ 9 в
   │  └─ 12 нарушение ходьбы
   │     └─ 14 нарушение функции тазовых
   │        органов
```

4.2 Синтаксический анализ, основанный на применении онтологии

Второй рассматриваемый алгоритм предполагает, что онтология описывает все разрешенные связи между понятиями, за исключением некоторых связок, предписываемых синтаксисом. Итоговое дерево разбора должно полностью соответствовать связям в онтологии и может изначально строиться по ним. Для определения порядка следования слов и связей между ними используются правила поверхностного синтаксиса. Данный алгоритм использует некоторые предположения о структуре русского предложения, описывающего жалобы пациента. Сформулируем их в терминах нашей онтологии.

- Признак вводится до перечисления своих характеристик.
- Характеристика вводится до упоминания значений. Например, «*Жалобы на* ***боль в спине***[признак] *с* ***локализацией*** [характеристика] *в области* ***поясничного отдела позвоночника***[значение]».
- Так как в тексте жалобы могут опускаться названия характеристик и признаков, по умолчанию значение может быть синтаксически подчинено как характеристике, так и признаку, а характеристика может быть подчинена признаку. Например, «*Жалобы на* ***боль в спине*** *с* ***локализацией*** *в области* ***поясничного отдела***

позвоночника» имеет то же значение, что и *Жалобы на* ***боль в спине*** *в области* ***поясничного отдела позвоночника***.

- В последовательности неоднословных терминов, обладающих одним и тем же первым или последним словом (или словосочетанием), повторяющееся слово или словосочетание может быть записано только один раз (*спинная и поясничная область позвоночника* вместо *спинная область позвоночника и поясничная область позвоночника*)[1]. Назовем такую форму записи упакованной формой использования терминов. Объединение терминов в упакованную форму может производиться при помощи союза или без него (например, *грудная аорта и артерия, спинная, поясничная область позвоночника*).

Для того чтобы создавать только корректные синтаксические связи, нам потребуется список терминов, предлогов и падежей, разрешенных для создания определенных связей с этими терминами. Термины должны соответствовать как главному, так и подчиненному словам, и могут задаваться не только начальными формами, но и разделами онтологии (здесь можно использовать регулярные выражения для проверки пути от корня онтологии к термину на соответствие шаблону). Подобные ограничения могут быть выражены в виде кортежа, задающего главное и подчиненное слово, шаблон их пути в онтологии от корня, возможный предлог и падеж, при помощи которых проводится подчинение зависимого слова главному. В качестве примера приведем правило <«», «.+/Боли», «», «», «.+/Боли/.+/Локализация/.+», «», «в», «Case=Loc»>, где любому главному слову, находящемуся в разделе «Боли», может подчиняться любое слово из подраздела «Локализация» раздела «Боли» в предложном падеже, соединенное через предлог «в» (боли в спинном отделе позвоночника). Всего было написано 44 таких правила, причем 41 из них описывало предлоги и падежи, задающие подчинение конкретных терминов.

С учетом описанных правил и предположений о структуре текста, алгоритм анализа жалоб пациентов на основе онтологии выглядит следующим образом.

1. Провести токенизацию и морфологическую разметку текста жалобы.
2. В размеченном тексте найти все термины. Каждый неоднословный термин должен быть свернут к одному токену с приписанными к нему грамматическими параметрами, взятыми от главного слова. Последовательность терминов в упакованной форме должна быть развернута в соответствующее перечисление полных терминов.
3. Соединить оставшиеся слова при помощи множества правил.
   a. Создать пустой список правил кандидатов.
   b. Пройти по всем словам текста.
   c. Начиная с хвоста списка правил кандидатов, найти первое, помеченное как «активное», и чье подчиняемое слово может быть применено к текущему слову. Если такое правило было найдено, нужно: 1) удалить все правила, находящиеся после него; 2) создать соединение между главным и зависимым словами; 3) подчинить предлог зависимому слову, если он есть в правиле.
   d. Поместить в список правил-кандидатов все правила, чье главное слово может быть успешно применено к текущему слову. Если правило требует предлога, пометить его как «неактивное», в противном случае – как «активное».
   e. Если текущее слово является предлогом, пометить все неактивные правила, которые ожидают этого предлога, как «активные»; все правила, не обладающие предлогом, пометить как «неактивные».

[1] Подобное явление описано, например, в [20, с. 240]

4. Перебирать все прилагательные и причастия в размеченном тексте, которые не были включены в неоднословные термины и не были соединены с другими словами на предыдущих шагах. Эти слова следует соединить с соседними существительными, согласующимися с данным прилагательным или причастием. Предпочтение отдается существительному, находящемуся справа.
5. По полученному списку связей слов создать дерево зависимостей предложения.

Удаляя правила в пункте 3c алгоритма, мы пытаемся поддерживать проективность дерева, то есть свойство, при наличии которого слова двух синтаксических групп находятся компактно и не перемешиваются. При соединении прилагательных и существительных, мы придерживаемся соответствующей статистики для русского языка [21, 22]. Мы «переиспользуем» предлоги, так как мы обнаружили, что зачастую при перечислении терминов предлог пишется только при первом упоминании, а дальше опускается. Аналогичная ситуация наблюдается и в группах терминов в упакованной форме. Заметим, что при анализе мы игнорируем все знаки препинания, так как в них делается больше всего ошибок. Также большое количество ошибок делается и в написании терминов, в связи с чем мы приме няем при их поиске алгоритмы нечеткого сравнения.

Рассмотрим пример разбора текста жалобы при помощи предложенного алгоритма.

Жалобы на выраженную боль в грудном и поясничном отделе позвоночника, усиление боли при физической нагрузке, нарушение ходьбы.

На первом этапе мы токенизируем текст и обрабатываем термины, как однословные, так и и неоднословные. Ниже термины даны в квадратных скобках, а остальные слова – в фигурных. Результат разбора будет следующим.

[Жалобы] {на} [выраженную] [боль] {в} [грудном отделе позвоночника] и [поясничном отделе позвоночника], [усиление боли] при [физической нагрузке], [нарушение ходьбы].

После анализа первых двух слов (*Жалобы на*) в списке будет одно правило, помеченное как активное, которое будет описывать соединения признака «*жалобы*» с его характеристиками. Слово «*выраженную*» пропускается без применения правил. По имеющемуся в списке правилу, термин «*боль*» может быть соединен с «*жалобы на*». Также на этом слове будет добавлено несколько правил вида «боль + в + локализация в предложном падеже». Эти правила будут активированы при анализе следующего слова – предлога «*в*». Активированные правила будут применены к терминам «*грудном отделе позвоночника*» и «*поясничном отделе позвоночника*», по этим правилам будут созданы соответствующие соединения между терминами. Термин «*усиление боли*» создаст соединение между признаком «*боль*» и характеристикой «*усиление боли».* На этом же слове будет удалено правило «боль + в + локализация» и добавлено правило «усиление боли + при + значение в предложном падеже». Последнее правило будет активировано предлогом «*при*», а на термине «*физической нагрузке*» будет создана связь между терминами. Наконец, термин «*нарушение ходьбы*» будет присоединен к термину «*жалобы*», а все правила, относящиеся к усилению боли, будут удалены из списка.

В результате работы алгоритма будут созданы следующие связи между терминами:

жалобы → боль → на,

боль → грудной отдел позвоночника → в,

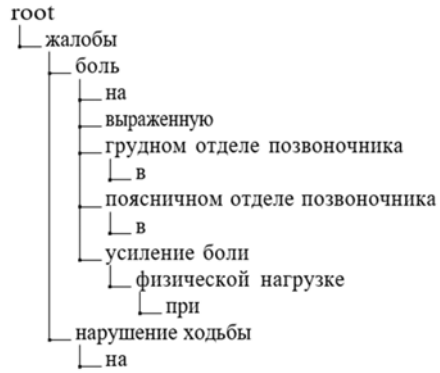
боль → поясничный отдел позвоночника → в,

боль → усиление боли,

усиление боли → физическая нагрузка → при,

жалобы → нарушение ходьбы → на

На следующем шаге мы соединим слова «*боль*» и «*выраженный*». После этого из полученных связей можно формировать итоговое дерево зависимостей.



Заметим, что правила глагольного и именного управления для терминов, которые мы писали вручную, могут быть извлечены из текстов жалоб в автоматизированном режиме. Для этого необходимо посчитать статистику совместной встречаемости терминов с учетом их связей в иерархии онтологии и выбрать наиболее вероятные варианты разбора. Однако детальная разработка такого метода сложна и является темой для отдельного проекта. Последовательности слов, которые не были найдены в онтологии, должны быть в нее добавлены вручную, так как для этого требуется определить их место в предметной области.

4.3 Разбор текста с использованием синтаксического анализатора

В качестве контроля мы использовали синтаксические анализаторы общего назначения. Заметим, что между деревьями зависимостей, которые строят наши алгоритмы, и деревьями зависимостей, построенными обычными анализаторами, имеется важное различие: последние строят деревья, вершины которых представляют собой отдельные слова, тогда как мы объединяем несколько слов неоднословного термина в одну вершину. Как следствие, прямое сравнение деревьев становится невозможным, и возникает необходимость ввести новый этап постобработки.

После разбора с использованием анализатора общего назначения мы находим термины в исходном тексте и вершины, соответствующие корням терминов. После этого мы удаляем все вершины, соответствующие словам каждого термина, оставляя лишь одну, находящуюся на самом верхнем уровне. Токен этой вершины заменяется на текст термина, а грамматические параметры заменяются на параметры слова, соответствующего корню дерева разбора термина. В случае, если в дереве зависимостей предложения находится не одно слово на верхнем уровне, а несколько, мы оставляем первое найденное. Такая ситуация означает, что синтаксический анализатор ошибся и разбил термин на несколько синтаксически не связанных фрагментов или некорректно определил связи термина с его потомками. Например, предложение *Жалобы на боль в поясничном отделе позвоночника с левой стороны* может разобраться *как жалоба → боль, боль → поясничный отдел* и *жалоба → позвоночник* (то есть позвоночник жалуется на боль в поясничном отделе). В этом примере термин был разорван на две части. Другим вариантом анализа является *боль → поясничный отдел позвоночника* и *боль → левая сторона* вместо *поясничный отдел позвоночника → левая сторона*. Здесь термины были выделены корректно, но связи построены неправильно. В первой ситуации наш алгоритм может создать некорректную связь (как мы это видели во втором примере), в результате чего будет сгенерирована ошибка (что является правильным поведением). Однако вероятна ситуация, когда наш алгоритм устранит ошибку и тем самым повысит точность работы синтаксического анализатора. Таким образом, полученная оценка для данного алгоритма будет являться скорее оценкой сверху, хотя и довольно точной.

5. Оценка точности методов

Итак, целью данной работы является сравнение трех описанных выше подходов к анализу текстов жалоб пациентов: традиционного синтаксического анализатора без каких-либо подсказок; традиционного синтаксического анализатора, результаты работы которого исправляются при помощи данных из онтологии; поверхностно-синтаксический анализ текста на основе определения связей между терминами из онтологии. Для экспериментов мы использовали два синтаксических анализатора: UDPipe 2.5 (выпущен в декабре 2019) и spaCy 3.0 (март 2021).

Мы вручную разметили «золотой стандарт», содержащий сто текстов жалоб пациентов, написанных на русском языке. Каждая жалоба была представлена в виде дерева зависимостей, всего жалобы содержали 1313 связей, соединяющих термины. Сравнивая результаты работы анализаторов с «золотым стандартом», мы не принимали во внимания слова, не являющиеся терминами, и не включили их в разметку «золотого стандарта». Несмотря на то, что предлоги в нашей нотации несут важную информацию о соединении терминов, они также были исключены из рассмотрения с тем, чтобы не вносить шум в статистику соединения терминов.

Так как нашей целью была оценка точности соединения терминов между собой, а не оценка правильности определения вида такого соединения, мы использовали метрику UAS (Unlabeled Attachment Score, доля правильных ответов при построении связей), а не LAS (Labeled Attachment Score, доля правильных ответов как для связей, так и их меток) [23]. Результаты наших экспериментов приведены в табл. 1.

Табл. 1. Результаты экспериментов
Table 1. Experimental results

Метод	**Правильных ответов**	**UAS**
UDPipe	975	0.743
Spacy	1080	0.823
UDPipe + онтология	930	0.708
Spacy + онтология	1021	0.778
Поверхностный синтаксический внализ + онтология	1084	0.826

Как видно из Табл. 1, применение онтологии ухудшает результаты синтаксического анализа, что связано с неполнотой применяемой нами онтологии. Наш анализ показал, что применение онтологии позволяет исправить некоторые ошибки анализатора, но привносит две большие проблемы. Первая проблема заключается в том, что если две вершины найдены в онтологии, но между ними нет прямой связи, то наш метод переносит подчиненную вершину выше к корню, где она и остается. Получается, что метод разрывает вершины, корректно соединенные синтаксическим анализатором, внося тем самым ошибку. Заметим, что такой же результат мы получим там, где синтаксический анализатор сам допустил ошибку. Второй проблемой является тот факт, что метод не включает в итоговое дерево те вершины, которые не были найдены в онтологии, тогда как синтаксический анализатор старается так или иначе соединить такую вершину с другими (причем может сделать это вполне успешно). Получается, что неполнота онтологии, в которой отсутствуют как понятия, так и связи между ними, приводит к увеличению числа ошибок.

Поверхностный синтаксический анализ с применением информации из онтологии менее чувствителен к ее неполноте. Вместо того чтобы искать прямые соединения между вершинами в онтологии, метод проверяет семантические метки, которыми являются фрагменты путей от корня онтологии. Например, если термин имеется в онтологии и записан в ней как «Локализация», то он будет трактоваться как таковая вне зависимости от наличия связи с нужным термином. Получается, что метод подразумевает неполноту онтологии: локализация для одного признака может являться локализацией для другого, даже если прямая связь в онтологии не обозначена по тем или иным причинам. Более того, отсутствие подобной связи может рассматриваться как сигнал о необходимости пополнения онтологии.

Отметим разницу между результатами работы UDPipe и spaCy – 0.08 UAS. Эта разница является видимым прогрессом в области синтаксических анализаторов за последние полтора года. Мы сравнили результаты работа spaCy 2 и UDPipe 2.5 и обнаружили, что точность их работы являлась сравнимой.

Также отметим разницу между точностью, которую показывает spaCy, и точностью нашего метода. Среди прочего такая маленькая разница связана с тем, что нам не удалось реализовать обработку последовательностей терминов в упакованной форме при анализе в spaCy из-за сложностей редактирования и ручного построения генерируемого им дерева зависимостей. Мы обнаружили, что в «золотом стандарте» имеется 20 последовательностей с союзом «и» (спинной и поясничный отдел позвоночника) и несколько, объединенных через запятую (спинной, поясничный отдел позвоночника). То есть результаты работы spaCy должны быть выше.

Одной из самых важных проблем метода поверхностного синтаксического анализа с использованием онтологии является отсутствие правил, описывающих глагольное управление (которому также подчиняются и причастия). Заметим, однако, что онтология содержит в себе по большей части информацию о понятиях, то есть существительных, а не о процессах, выражаемых глаголами. Как следствие, на текущем этапе мы можем игнорировать глаголы, так как в жалобах пациентов они встречаются относительно редко и не играют той важной роли, которая им присуща в других текстах. Ещё одним решением может быть преобразование глаголов в существительные по словарю.

Мы можем утверждать, что наши эксперименты подтвердили старую идею о том, что анализ текстов может проводиться с использованием семантической информации (в нашем случае – онтологии) и небольшого числа простых правил поверхностного синтаксического анализа (хотя текущие правила скорее хранят информацию об именном управлении). Подобные правила можно извлекать из текстов в автоматизированном режиме, но такая работа служит темой для отдельного исследования. Для автоматического получения кандидатов в добавляемые термины можно использовать подходы, подобные предложенному в [24], извлекающие их из формализованных описаний предметной области.

6. Используемые свойства онтологии и требования к ней

В данном проекте мы использовали несколько свойств связей онтологии, относящихся к синтаксису русского языка. Часть из этих свойств была отражена при построении онтологии, использованной в данном проекте – Базы медицинской терминологии и наблюдений [17]. Опишем разницу между разными подходами в разработке онтологий.

Онтология WordNet в исходном виде [25], использует ограниченное количество семантических помет для связей между словами: гипоним/гипероним, мероним/холоним, специальные наборы связей для глаголов и прилагательных. Сама структура онтологии не запрещает использование других видов связей, однако на практике они встречаются довольно редко. Этот недостаток зачастую исправляется при переводе онтологии или конвертации других онтологий и тезаурусов в формат WordNet. Так, например, онтология

ruWordNet, созданная на основе тезауруса РуТез, [26] содержит в себе такие связи как ассоциации, синонимия и др. В отличие от WordNet и ruWordNet, наша онтология содержит в себе функциональные и атрибутивные связи (см. разд. 3).

Теперь кратко рассмотрим, как эти семантические связи выражаются в русском языке. Связь гипоним/гипероним вводится в тексте при помощи специальной конструкции, например, «А – это В», «А, вид А – В» и т.д. (подробнее описано в [27, 28]), в противном случае, он будет содержать в себе тавтологию. То же касается синонимических и антонимических связей. Отношение меронимии показывает принадлежность одного объекта или понятия другому. В русском языке такая связь выражается при помощи генетивной конструкции (холоним подчиняется мерониму и находится в родительном падеже) – «капот автомобиля». Та же генетивная конструкция может использоваться и для выражения функциональной связи – «локализация боли». Также функциональная связь может быть выражена при помощи связи через предлог – «боль с локализацией» или с использованием связывающего глагола в разных формах – «боль, локализующаяся в». Для атрибутивных связей также может использоваться прямое дополнение или предложная связь – «локализация в правом глазу», «иррадиация в правую ногу». В отличие от других видов, атрибутивная связь позволяет связывать между собой существительное, выражающее название атрибута, и прилагательное, выражающее его значение – «сильная боль в глазу».

Заметим, что функциональных и атрибутивных связей достаточно, чтобы описать большую часть связей между словами в медицинских текстах. При этом отсутствие таких связей не позволяет получить приемлемый уровень полноты при их анализе. Таким образом, мы можем утверждать, что онтология жалоб пациентов обязана включать в себя атрибутивные и функциональные связи для обеспечения успешной работы синтаксического анализатора. Именно это свойство наблюдается у «Базы медицинской терминологии и наблюдений» [17], и именно благодаря ему нам удалось достигнуть достаточно высокого уровня точности результатов (хотя, говоря о точности, следует говорить скорее об ошибках, полученных вследствие неполноты онтологии и системы правил поверхностно-синтаксического анализа).

Также заметим, что синонимические и гипонимические связи оказываются очень полезны, когда один термин оказывается заменен другим. Для нашей онтологии такая замена означает пропуск уровня в связях одного термина с другим. Именно поэтому мы предъявляли требование принципиального наличия прямого пути между терминами по иерархии, а не наличия непосредственной связи. Требования наличия прямого пути связано с тем, что переход между соседями означает переход к понятию хотя и близкому, но имеющему отличия. Так, например, «Боль в глазу» имеет набор характеристик и их значений, отличающийся от характеристик понятия «Боль в спине», хотя оба этих термина являются соседями и прямыми потомками термина «Боль».

7. Заключение

В данной статье мы показали, что поверхностный синтаксический анализ русских текстов с определением связей при помощи онтологии узкой предметной области может показывать точность, сравнимую с точностью современных синтаксических анализаторов на основе нейронных сетей. Предложенный нами подход находится на одном уровне с одним из лучших современных синтаксических анализаторов – spaCy 3.0 в области анализа жалоб пациентов, извлеченных из историй болезни (разница составила лишь 0.003 UAS). Заметим, что предыдущее поколение синтаксических анализаторов (UDPipe 2 и spaCy 2) показывает гораздо более низкие результаты.

Подобная разница объясняется тем, что медицинские тексты обладают рядом особенностей, критически снижающих точность при их анализе: отсутствие глаголов; длинные последовательности именных групп, выражающих связи между терминами; сами термины.

Из-за таких особенностей, точность синтаксического анализа снижается с примерно 0.95 до примерно 0.85.

Недостатком предложенного метода является наличие онтологии предметной области, построенной по определенным принципам и содержащей в себе иерархию терминов. Создание подобной онтологии является серьезным трудом и занимает годы. Ручное создание правил глагольного и именного управления также занимает много времени. Подобная ситуация оправдывает себя в случаях, когда исходных данных недостаточно для применения методов машинного обучения. В то же время, развитие нейронных сетей, являющихся основой современных синтаксических анализаторов, ещё не исчерпало своего ресурса, в связи с чем мы можем ожидать нового прогресса, в том числе и в синтаксическом анализе. Применение предложенного нами метода для широкого спектра текстов требует появления новых методов автоматического выделения словарей глагольного и именного управления, а также автоматизированного пополнения онтологий, так как полнота и точность существующих методов всё ещё не удовлетворяет практических потребностей. В связи с этим можно утверждать, что предложенный нами метод подходит для текстов узкой предметной области с простым синтаксисом, существенно отличающимся от общепринятой нормы, и хорошо проработанной онтологией, отвечающей описанным в статье требованиям. В подобной ситуации он позволит получить выигрыш в скорости разработки системы и точности ее работы.

Список литературы / References

[1] Nugmanov R., Alimova I., Tutubalina E. Adverse drug reactions identification in social media posts and electronic health records with neural networks. European Journal of Clinical Investigation, vol.49, 2019, pp. 116-117.

[2] Chapman W.W., Gundlapalli A.V. et al. Natural Language Processing for Biosurveillance. In Infectious Disease Informatics and Biosurveillance: Research, Systems and Case Studies. Springer, 2011. pp. 279-310.

[3] Straka M., Straková J., Hajic J. UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging. In Proc. of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, 2019. pp. 95-103.

[4] Astudillo R.F., Ballesteros M. et al. Transition-based Parsing with Stack-Transformers. arXiv:2010.10669, 2020.

[5] Wang Y., Lee H.-Y., Chen Y.-N. Tree Transformer: Integrating Tree Structures into Self-Attention. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. pp. 1061-1070.

[6] Abney S. P. Parsing By Chunks. Studies in Linguistics and Philosophy, vol. 44, 1991. pp. 19-33.

[7] Molina A., Pla F. Shallow Parsing using Specialized HMMs. Journal of Machine Learning Research, vol. 2. 2002. pp. 595-613.

[8] Sha F., Pereira F.C. Shallow Parsing with Conditional Random Fields. In Proc. of 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003. pp. 134-141.

[9] Кобзарева Т.Ю., Лахути Д.Г., Ножов И.М. Модель сегментации русского предложения. Труды Международного семинара Диалог 2001, 2001 г., стр. 185–194 / Kobzareva T.Yu., Lakhuti D.G., Nozhov I.M. Segmentation model of the Russian sentence. In Proc. of the International Seminar Dialogue 2001, 2001, pp. 185-194 (in Russoan).

[10] Sowa J.F. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing, 2000, 594 p.

[11] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Изд-во Московского университета, 2011 г., 512 стр. / Lukashevich N.V. Thesauri in information retrieval problems. Publishing house of Moscow State University, 2011, 512 p. (in Russian).

[12] Current Bibliographies in Medicine. URL: https://www.nlm.nih.gov/archive/20040831/pubs/cbm/umlscbm.html.

[13] Aronson A.R., Lang F.-M. An overview of MetaMap: historical perspective and recent Advances. Journal of the American Medical Informatics Association, 2010, vol. 17, issue 3, pp. 229-236.

[14] Valdez J. An Ontology-Enabled Natural Language Processing Pipeline for Provenance Metadata Extraction from Biomedical Text (Short Paper). Lecture Notes in Computer Science, vol. 10033, 2016, pp. 699-708.

[15] MSHRUS (MeSH Russian) – Statistics. URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/stats.html

[16] Shelmanov A. O., Smirnov I. V., Vishneva E.A. Information Extraction from Clinical Texts in Russian. In Proc. of the International Conference on Computational Linguistics and Intellectual Technologies (Dialog-2015), 2015, pp. 560–572.

[17] Грибова В.В., Москаленко Ф.М. и др. Концепция гетерогенного хранилища биомедицинской информации. Информационные технологии, том 27, no. 2, 2019 г., стр. 97-106 / Gribova V.V., Moskalenko Ph.M. et al. A Concept for a Heterogeneous Biomedical Information Warehouse. Information technologies, vol. 25, no. 2, 2019, pp. 97-106 (in Russian).

[18] spaCy: What's New in v3.0. URL: https://spacy.io/usage/v3.

[19] Nivre J., de Marneffe M.-C. et al. Universal Dependencies v1: A Multilingual Treebank Collection. In Proc. of the 10th International Conference on Language Resources and Evaluation (LREC 2016), 2016, pp. 1659-1666.

[20] Апресян Ю.Д. Избранные труды, т. I. Лексическая семантика: 2 изд. М, Школа, 1995, 472 стр. / Apresyan Yu.D. Selected works, vol. I. Lexical semantics: 2nd ed. M, Shkola, 1995, 472 p. (in Russian).

[21] Клышинский Э.С. Степень свободы русского синтаксиса несколько преувеличена. Сборник трудов 20-го научно-практического семинара «Новые информационные технологии в автоматизированных системах», 2017 г., стр. 112-116 / Klyshinskiy E.S. The degree of freedom of Russian syntax is somewhat exaggerated. In Proc. of the 20th Scientific-Practical Seminar on New Information Technologies in Automated Systems, 2017, pp. 112-116 (in Russian).

[22] Клышинский Э.С., Логачева В.К. и др. Количественная оценка грамматической неоднозначности некоторых европейских языков. Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация, 2020, том 18, вып. 1, стр. 5-21 / Klyshinskiy E.S. Logacheva V.K. et al. Quantitative Estimation of Grammatical Ambiguity: Case of European Languages. NSU Vestnik. Series: Linguistics and Intercultural Communication, vol. 18. issue 1, 2020, pp. 5-21 (in Russian).

[23] Nivre J., Fang C.-T. Universal Dependency Evaluation. In Proc. of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), 2017, p. 86-95.

[24] Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке. Труды международной конференции Диалог-2010, 2010 г., стр. 136-143 / Zakharov V.P., Khokhlova M.V. Study of effectiveness of statistical measures for collocation extraction on Russian texts. In Proc. of the International Conference Dialogue 2010, 2010, стр. 136-143 (in Russian).

[25] Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, 1998, 449 p.

[26] Лукашевич Н.В., Лашевич Г. и др. Порождение тезауруса типа WordNet для русского языка. Труды Пятнадцатой национальной конференция по искусственному интеллекту с международным участием (КИИ-2016), 2016 г., стр. 89-97 / Loukachevitch N.V., Lashevich G. et al. Generating russian wordmet. In Proc, of the Fifteenth National Conference on Artificial Intelligence with International Participation (CAI 2016), 2016, pp. 89-97 (in Russian).

[27] Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов. Труды Десятой национальной конференция по искусственному интеллекту с международным участием (КИИ-2006), 2006 г., стр. 506-524 / Bolshakova E.I., Vasilieva N.E., Morozov S.S. Lexicosyntactic patterns for automatic text processing. In Proc, of the Tenth National Conference on Artificial Intelligence with International Participation (CAI 2006), 2006, pp. 506-524 (in Russian).

[28] Большакова Е.И., Баева Н.В. и др. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов. Труды международной конференции Диалог-2007, 2007 г., стр. 70-75 / Bolshakova E.I., Baeva N.V. et al. Lexicosyntactic patterns for automatic text processing. In Proc. of the International Conference Dialogue 2007, 2007, pp. 70-75 (in Russian).

Информация об авторах / Information about authors

Борис Израйльевич ГЕЛЬЦЕР – доктор медицинских наук, профессор, член-корреспондент РАН, директор Департамента клинической медицины Школы биомедицины ДВФУ. Научные интересы: доказательная медицина, клиническая медицина, методы машинного обучения в медицине, медицинские информационные системы.

Boris Israelevich GELTSER – Doctor of Medicine, professor corresponding member of RAS, head of Department of Clinical Medicine of School of Biomedicine FEFU. Research interests: evidence based medicine, clinical medicine, machine learning in medicine, medical information systems.

Татьяна Александровна ГОРБАЧ – кандидат медицинских наук, врач-невролог Медицинского центра ДВФУ. Научные интересы: неврология, когнитивные расстройства, системы представления знаний.

Tatiana Aleksandrovna GORBACH – PhD in Medicine, researcher at IACP FEB RAS. Research interests: neurology, cognitive disorders, knowledge representation.

Валерия Викторовна ГРИБОВА – доктор технических наук, заместитель директора по научной работе ИАПУ ДВО РАН. Научные интересы: Искусственный интеллект, принятие решений, экспертные системы, программные системы.

Valeriya Victorovna GRIBOVA – doctor of technical science, Research Deputy Director of IACP FEB RAS. Research interests: artificial intelligence, decision making, expert systems, software systems.

Олеся Владимировна КАРПИК – младший научный сотрудник ИПМ им. М.В. Келдыша. Научные интересы: лексикография, синтаксис, фонетика.

Olesya Vladimirovna KARPIK – junior researcher at Keldysh IAM RAS. Research interests: lexicography, syntax, phonetics.

Эдуард Станиславович КЛЫШИНСКИЙ – кандидат технических наук, доцент, доцент школы лингвистики НИУ ВШЭ. Научные интересы: искусственный интеллект, формальный синтаксис, автоматическая обработка текстов.

Eduard Stanislavovich KLYSHINSKIY – PhD in Computer Science, associated professor at School of Linguistics at NRU HSE. Research interests: artificial intelligence, formal syntax, natural language processing.

Наталия Александровна КОЧЕТКОВА – аспирант НИУ ВШЭ. Научные интересы: автоматическая обработка текстов, извлечение именованных сущностей, стилеметрия.

Natalia Aleksandrovna KOCHETKOVA – PhD student at NRU HSE. Research interests: natural language processing, named entities recognition, stylometrics.

Дмитрий Борисович ОКУНЬ – кандидат медицинских наук, научный сотрудник ИАПУ ДВО РАН. Научные интересы: онтологии, прикладные интеллектуальные системы, экспертные системы.

Dmitry Borisovich OKUN – PhD in Medicine, researcher at IACP FEB RAS. Research interests: ontology, application intelligent systems, expert systems.

Маргарита Вячеславовна ПЕТРЯЕВА – кандидат медицинских наук, научный сотрудник ИАПУ ДВО РАН. Научные интересы: биомедицина, прикладные интеллектуальные системы, экспертные системы.

Margaret Vyacheslavovna PETRYAEVA – PhD in Medicine, researcher at IACP FEB RAS. Research interests: biomedical system, applied intelligent systems, expert systems.

Карина Иосифовна ШАХГЕЛЬДЯН – доктор технических наук, профессор, директор института информационных технологий ВГУЭС. Научные интересы: системы представления знаний, машинное обучение, программные системы.

Carina Iosifovna SHAKHGELDYAN – doctor of technical science, professor, director of Institute of Information Technologies at VVSU. Research interests: knowledge representation, machine learning, software systems.