УДК 004.89 **DOI** 10.15622/ia.21.6.4

Р.Х. Зулкарнеев, Н.И. Юсупова, О.Н. Сметанина, М.М. Гаянова, А.М. Вульфин

МЕТОДЫ И МОДЕЛИ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ МЕДИЦИНСКИХ ДОКУМЕНТОВ

Зулкарнеев Р.Х., Юсупова Н.И., Сметанина О.Н., Гаянова М.М., Вульфин А.М. **Методы и модели извлечения знаний из медицинских документов.**

Аннотация. В работе выполнен анализ современного состояния проблемы извлечения знаний из клинических рекомендаций, представленных в виде слабоструктурированных корпусов текстовых документов на естественном языке с учетом их периодического обновления. Рассматриваемые методы интеллектуального анализа накопленных массивов медицинских данных позволяют автоматизировать ряд задач, направленных на повышение качества медицинской помощи за счет значимой поддержки принятия решений в процессе диагностики и лечения. Выполнен обзор известных публикаций, освещающий подходы к автоматизации построения нейросетевых языковых моделей, онтологий и графов знаний в задачах семантического моделирования проблемно-ориентированного корпуса текстов. Представлена структурно-функциональная организация системы извлечения знаний и автоматического построения онтологии и графа знаний проблемно-ориентированного корпуса для конкретной предметной области. Рассмотрены основные этапы извлечения знаний и динамического обновления графа знаний: извлечение именованных сущностей, семантическое аннотирование, извлечение терминов, ключевых слов, тематическое моделирование, идентификация тем и извлечение отношений. Формализованное представление текстов получено с помощью предобученной модели-трансформера BERT. Использовано автоматическое выделение триплетов «объект»-«действие»-«субъект» на основе частеречной разметки корпуса текстов для построения фрагментов графа знаний. Проведен эксперимент на корпусе медицинских текстов заданной тематики (162 документа обезличенных историй болезни пациентов педиатрического центра) без предварительной разметки с целью проверки предложенного решения по извлечению триплетов и конструирования на их основе графа знаний. Анализ экспериментальных результатов подтверждает необходимость более глубокой разметки корпуса текстовых документов для учета специфики медицинских текстовых документов. Показано, что модели общего назначения не позволяют приблизиться по качеству выделения именованных сущностей к специализированным моделям, однако, позволяют предварительно разметить корпус для дальнейшей верификации и уточнения разметки (оценка F1-меры для модели общего назначения – 20,4% по сравнению с вариантом использования словаря – 16,7%). Для неразмеченного корпуса текстов предложенное решение демонстрирует удовлетворительную работоспособность ввиду выделения атомарных фрагментов, включаемых в автоматически формируемую онтологию.

Ключевые слова: клинические тексты, извлечение информации, машинное обучение, интеллектуальный анализ медицинских данных, автоматическое построение онтологий, графы знаний.

1. Введение. Методы и системы интеллектуального анализа медицинских данных применяются для поддержки принятия решений в процессе диагностики заболеваний [1], контроля выполнения

лечебных протоколов и координации действий медицинского персонала, а также для предиктивного анализа и выявления на ранних этапах потенциально опасных состояний пациентов [2].

Медицинские данные можно условно разделить на две крупные категории [1]:

– структурированные данные, имеющие заранее определенный формат представления и хранения, хорошо поддающиеся формализации и последующей обработки с привлечением технологий интеллектуального анализа данных (результаты анализов и пр.);

– слабоструктурированные данные, представленные на естественном языке, со слабо выраженной или отсутствующей жесткой структурой (форматом) представления и хранения (анамнезы, протоколы осмотров, результаты обследований и так далее), для автоматизации анализа которых необходимо применение методов естественно-языковой обработки, формализации и извлечения структуры для последующего применения интеллектуального анализа и построения онтологии и графа знаний проблемной области.

Одним из актуальных направлений работы с данными, в том числе, с «большими данными» в медицинской практике является оперативный анализ (сбор, хранение, формализация, постоянное обновление, анализ, интерпретация) с целью создания регулярно пополняемых баз – клинических регистров. Высокая загруженность специалистов здравоохранения осложняет процесс принятия решений в сложных случаях, ввиду существенных временных затрат на поиск и анализ соответствующих источников. Методы интеллектуального анализа накопленных массивов медицинских данных позволяют автоматизировать подобные задачи, встречающиеся в клинической практике, повысив тем самым общий уровень качества медицинской помощи [1]. Внедрение интеллектуальных технологий направлено на повышение информационной осведомленности врача, помощь в быстром и обоснованном принятии клинического решения путем предоставления экспертных мнений и рекомендаций [3].

Ключевой проблемой при обработке и анализе медицинских данных является необходимость их формализации и извлечения знаний из периодически обновляемых клинических рекомендаций [4, 5]. Интеллектуальный анализ клинических текстов и извлечение знаний из накопленных массивов периодически меняющихся данных является одним из перспективных научных направлений на стыке компьютерной лингвистики, машинного обучения и медицины [6, 7], направленных на решение данной проблемы. На сегодняшний день

существует достаточно много технологий лингвистического анализа текстов [8], но, как показала практика, анализа текста на уровне только лингвистических правил недостаточно для корректного и полного извлечения фактов из корпуса медицинских документов [3]. Для эффективного извлечения фактов из текста база знаний должна содержать информацию, включающую медицинские онтологии, классификаторы, систематизированные знания в области анатомии, физиологии и патофизиологии человека. При сопровождении созданной базы знаний необходима постоянная актуализация информации с применением технологий анализа и сбора данных из первичных источников [9, 10, 11].

В работе [1] представлена комплексная система интеллектуальной обработки данных в многопрофильном педиатрическом центре, которая решает задачи автоматизации диагностики и выявления значимых признаков из накопленных слабоструктурированных данных. Из медицинских текстов извлекаются: названия заболеваний, симптомы, области тела, к которым относится заболевание, а также применяемые лекарственные препараты. Для извлечения знаний использованы медицинские тезаурусы, набор вручную составленных шаблонов, а также различные методы на основе машинного обучения.

Особенностью рассмотренного решения является применение методов глубокой иерархической разметки корпуса клинических текстов с широким привлечением экспертов предметной области. Как показывает анализ работ, проблемой является высокая трудоемкость подготовки исходных данных: создание и разметка соответствующих корпусов текстов, формирование баз правил, последующая верификация моделей машинного обучения. Перспективным является подход по извлечению знаний непосредственно из данных с помощью интеллектуальных алгоритмов, когда роль человека-эксперта сводится к верификации автоматически построенных онтологических моделей («обучение онтологий», ontology learning).

В статье рассмотрена задача анализа и разработки методов и механизмов извлечения знаний из периодически обновляемых клинических рекомендаций с целью извлечения знаний на основе технологий автоматизации построения онтологии проблемной области и формирования графа знаний.

Для решения имеющейся задачи в работе проведены следующие действия:

– во втором разделе проведен обзор известных публикаций по тематике автоматизации построения онтологий, графов знаний как

инструментов семантического моделирования проблемно-ориентированного корпуса текстов;

– в третьем разделе разработана структурно-функциональная организация системы извлечения знаний и автоматического построения онтологии и графа знаний проблемно-ориентированного корпуса для конкретной предметной области с целью последующего построения системы поддержки принятия решений при анализе клинических рекомендаций;

– в четвертом разделе представлены предварительные результаты эксперимента на корпусе медицинских текстов заданной тематики (пульмонология, история болезней пациентов) при извлечении знаний из неразмеченного корпуса медицинских текстов;

– в пятом разделе отражены анализ и обсуждение результатов исследования.

2. Методы извлечения знаний из слабоструктурированных данных на основе автоматизации построения онтологий и графов знаний

2.1. Подходы к автоматизации построения онтологий. Онтологии приобрели большую популярность и признание и считаются качественным источником семантики и интероперабельности во всех интеллектуальных системах обработки слабоструктурированных и неструктурированных данных.

Для представления (хранения) медицинских знаний разработаны специальные онтологии, которые условно разделены на две группы [4]:

1. онтологии формирования медицинских признаков из элементарных терминов;
2. онтологии описания патологических процессов и других медицинских явлений.

Онтологии являются фундаментом для большинства существующих медицинских экспертных систем.

В традиционном подходе к построению онтологии в качестве основного источника знаний выступает эксперт – специалист в предметной области. Данный подход имеет множество недостатков, связанных с серьезными трудозатратами и ограниченными возможностями экспертов предметной области на этапе сбора, подготовки и последующего анализа данных.

Бизнес-процессы современной цифровой экономики генерируют значительные объемы данных, что существенно снижает эффективность эксперта как непосредственного и единственного источника знаний. Экспоненциальный рост объемов доступных

слабоструктурированных или неструктурированных данных в глобальных и локальных базах существенно повысил актуальность проблемы автоматического получения онтологии на основе анализа проблемно-ориентированных корпусов текстов [12].

Становится перспективным подход по извлечению знаний непосредственно из существующих структурированных и неструктурированных источников данных с помощью методов и технологий интеллектуального анализа [13]. При реализации данного подхода для человека-эксперта отводится роль проектирования концептуальных верхнеуровневых абстракций, частичная разметка исходных данных и валидация полученных результатов (верификация автоматически построенных онтологических моделей).

В [14] предлагается несколько методологий, использующих методы из различных областей (машинное обучение, интеллектуальный анализ текста, представление знаний и рассуждения, поиск информации и обработка естественного языка), для обеспечения определенного уровня автоматизации процесса получения онтологий из неструктурированного текста. Описывается процесс изучения онтологий и дальнейшая классификация методов изучения онтологий на три класса (лингвистические, статистические и логические) и обсуждается множество алгоритмов в каждой категории.

В работе [15] предложено рассматривать «обучение онтологий» на основе слабоструктурированных данных как некоторую последовательность согласованных действий по извлечению знаний из данных, проектированию и построению отдельных фрагментов онтологий. Первым шагом является извлечение из текста основных терминов. Множество выделенных терминов на основе поиска синонимов трансформируется во множество концептов. Последующее структурирование концептов позволяет построить иерархию концептов. На заключительном этапе строится совокупность аксиом для проектируемой онтологии. Подобный подход позволяет строить онтологии без трудоемкого ручного проектирования, что стало возможным благодаря стремительному развитию технологий интеллектуальной обработки текстов на основе методов машинного обучения, что позволяет вывести качество извлекаемых иерархий концептов на принципиально новый уровень. Предваряющие исследования основывались на базовой версии онтологии, разработанной экспертами вручную, на основе которой выполнено извлечение знаний из слабоструктурированных текстовых данных с помощью методов машинного обучения [16].

В работе [17] описаны подходы к обучению онтологий на основе анализа метаданных и контекста слабоструктурированного содержания. Предложена модель совместного представления контента и его метаданных в системе управления контентом. Для извлечения терминов был использован ансамблевый метод. Описаны методы построения таксономических отношений на основе векторного представления слов и нетаксономических отношений на основе анализа универсальных зависимостей с помощью алгоритмов обработки естественного языка с применением машинного обучения.

В работе [18] предложена схема применения методов кластеризации в задаче формирования концептов на основе кластеров семантически замкнутых терминов. Для решения проблемы построения кластеров, специфичных для конкретной предметной области или при определении соответствующих концептуальных обозначений для каждого кластера, предложено использовать основные понятия из онтологии предметной области в качестве предварительных знаний и адаптировать кластеризацию терминов с помощью моделей LDA (Latent Dirichlet allocation – латентное размещение Дирихле), основанных на начальных знаниях, чтобы учесть эти основные понятия. На первом этапе выделенная тема связана с набором начальных терминов одной основной концепции, затем обучение модели руководствуется этими начальными понятиями, чтобы собрать в одной и той же теме термины, которые относятся к ее основной концепции.

Предлагаемый в [19] подход автоматизирует процесс создания баз знаний, основываясь на принципах адаптивности к специфике проблемной области экспертизы, аспектам рассматриваемой задачи и глобальным базам знаний. Приводится онтологически управляемая архитектура инструментальной среды, автоматизирующей создание продукционных экспертных систем. На основе заданных с помощью онтологий сценариев естественно-языкового диалога процесс извлечения знаний позволяет существенно снизить трудозатраты эксперта и инженера по знаниям на построение и верификацию базы знаний.

В статье [20] рассматриваются вопросы применения алгебраических методов представления и обработки знаний в медицинских интеллектуальных информационных системах. Для представления знаний предлагается использовать аппарат Е-структур для построения процедур обеспечения целостности баз знаний.

Статья [21] посвящена обобщению методов обработки текстов на естественном языке, в основе которых лежит формирование и

использование ассоциативно-онтологического представления данных. Предлагаемый метод расширяет методы лингвистической статистики и логико-статистические методы для извлечения знаний и построения ассоциативной онтологии заданной предметной области.

В работе [22] предложена методика обработки обращений пациентов на основе применения инфологической системы, позволяющей организовать выявление семантического содержания жалоб на состояние здоровья. В основу предлагаемой методики положен инфологический подход к обработке текстовых документов на основе итерационного процесса формирования тематических знаний посредством формирования тематических антологий – т.е. на основе предметно-ориентированных корпусов, их тезаурусов и глоссариев производится уточнение области и оценка сходства с ними новых текстовых документов.

Рассмотренные работы по автоматизации построения онтологий предлагают различные подходы и инструментарий для снижения нагрузки на экспертов предметной области и инженеров по знаниям, однако, как отмечается в актуальных исследованиях, применение моделей машинного обучения и интеллектуального анализа позволит на основе тонкой настройки существующих нейросетевых лингвистических моделей, построенных на обобщенных корпусах текстов, существенно повысить качество анализа исходных, «сырых» слабоструктурированных данных и снизить требования к предварительно построенным глоссариям и кодификаторам [23, 24, 25, 26].

2.2. Языковые модели в контексте инженерии знаний. Предобученные языковые модели обладают знаниями об отношениях, содержащихся в обучающей выборке [24]. В [27] отмечается, что языковые модели имеют множество преимуществ перед структурированными базами знаний, например, в том, что они не требуют проектирования структуры, свободно расширяемы новыми данными и не требуют предварительной разметки.

Предобученные языковые модели BERT [28] демонстрируют существенно более высокие результаты в решении задач обработки естественного языка. Для существующих версий предобученных языковых моделей семейства BERT особо актуальным является вопрос модификации и разработки методов непрерывной тонкой настройки и аугментации внешними данными для поддержания их актуальности в решаемых задачах с наименьшими временными затратами [24].

Модель BioBERT [29] предварительно обучена на корпусе медицинских текстов (аннотации статей PubMed и PMC) и широко

используется для решения задач извлечения именованных сущностей (Named Entity Recognition, NER), извлечения отношений между сущностями (Relationship Extraction, RE) и построения вопросно-ответных систем (Question Answering System, QA).

Клинические заметки содержат информацию о пациентах, которая выходит за рамки структурированных данных, таких как лабораторные показатели и лекарства. Тем не менее, клинические записи использовались недостаточно по сравнению со структурированными данными, поскольку они очень многомерны и разрежены.

Модель ClinicalBERT [30] – это вариант BERT, предобученный на корпусе клинических документов. Модель способна выделять отношения между медицинскими концепциями. Модель предварительно обучена на наборе данных Medical Information Mart for Intensive Care III из электронных медицинских карт 58 976 уникальных госпитализаций 38 597 пациентов в отделении интенсивной терапии в период с 2001 по 2012 год. Содержит 2 083 180 обезличенных заметок, связанных с госпитализациями. Модель Bio-Discharge-Summary [30] является дообученным вариантом BioBERT и предназначена для решения нескольких задач обработки слабоструктурированных проблемно-ориентированных текстов с минимальными архитектурными модификациями.

В работах [31, 32] представлен полноразмерный русскоязычный корпус отзывов пользователей Интернета со сложной маркировкой NER, а также оценка уровней точности, достигнутых в этом корпусе нейронными сетями глубокого обучения для извлечения фармакологически значимых сущностей из русских текстов: Medication (33005 высказываний), Adverse Drug Reaction (1778), Disease (17403), и Note (4490).

В работе [33] представлен российский корпус реакций на лекарства (The Russian Drug Reaction Corpus, RuDReC) – частично аннотированный корпус отзывов потребителей на русском языке о фармацевтических продуктах для выявления именованных объектов, связанных со здоровьем, и эффективности фармацевтических продуктов. Представлена базовая модель для задач распознавания именованных сущностей (NER) и классификации предложений с несколькими метками в этом корпусе. Макро-оценка меры F_1 в задаче NER составляет 74,85% и была достигнута с помощью модели RuDR-BERT.

Разработанный в [1] метод позволяет находить в тексте различные варианты использования медицинских терминов по

заданным кодификаторам, аналогично системе MetaMap [34]. В качестве кодификаторов использовались Unified Medical Language System (UMLS) Metathesaurus [35] (русскоязычный вариант представлен MeSH в [36]), а также подготовленный государственный реестр лекарственных средств [37].

В работе [38] рассмотрена задача обработки текстов и подготовки моделей векторизации для классификации научных текстов по научной специальности. Проведено сравнение разных способов подготовки текстов и выявлена наиболее эффективная их комбинация, приведены результаты векторизации корпуса текстов на основе метод TF-IDF, оценено влияние гиперпараметров на результаты классификации с помощью предложенной модели машинного обучения.

В работе [1] создан размеченный корпус клинических текстов на русском языке. В состав корпуса вошли более 120 деперсонализованных историй болезни пациентов педиатрического центра с аллергическими, ревматическими и нефрологическими заболеваниями, а также болезнями органов дыхания. При составлении инструкций по разметке учитывался опыт зарубежных семинаров, таких как CLEF eHealth [39], для которых создавались схожие ресурсы. Специалисты в области медицины разметили в корпусе более 18 000 сущностей, а также более 12 000 атрибутов и связей. Оценки метода извлечения лекарственных препаратов из клинических текстов: точность – 84,3%, полнота – 74,6%, F_1-мера – 79,2%.

В работе [26] рассматривается новый способ извлечения понятий из текстов предметной области на основе комбинации анализа формальных понятий и бутстрап-технологии информационного поиска. Предложен новый способ автоматического извлечения понятий из текстов медицинской тематики, основанный на заполнении пропусков в сильно разреженных матрицах совместной встречаемости терминов, удовлетворяющих лексико-синтаксическим шаблонам вида «Существительное + Существительное» или «Существительное + Существительное в родительном падеже».

Анализ работ показывает, что для англоязычного домена документов созданы и исследованы как проблемно-ориентированные корпусы текстов, так и передовые языковые модели, предназначенные для решения целого спектра задач анализа слабоструктурированных данных. Для русскоязычного домена предприняты успешные попытки создания корпусов документов и построения языковых моделей, однако их использование для построения онтологических моделей и графов знаний изучено недостаточно – требуются значительные

усилия для расширения и разметки корпусов специализированных текстов, построения глоссариев и применения методов выделения NER и связей между сущностями для качественного перехода к автоматизированным вариантам построения онтологий.

2.3. Граф знаний как инструмент семантического моделирования проблемно-ориентированного корпуса текстов. Основываясь на онтологии, граф знаний позволяет формализовать и интегрировать гетерогенные источники данных и знаний в общую базу, одновременно обеспечивая их непротиворечивость.

Граф знаний (Knowledge Graph, KG) [40] – это структурированное графическое представление семантических знаний и отношений, где узлы в графе представляют сущности, а ребра представляют отношения между ними. Построение графа знаний предполагает извлечение связей из неструктурированного текста с последующим эффективным хранением в граф-ориентированных базах данных. Современное использование графов знаний возникло и развивалось в рамках направления Semantic Web [41], во многих работах граф знаний определяется как множество «триплет» в виде (субъект, предикат, объект), образующих RDF-граф [42, 43], в котором вершинами являются субъекты и объекты, а рёбра отображают отношения между ними. Модель данных RDF (Resource Description Framework – «среда описания ресурса») [44] является утверждением о ресурсах (информационные и неинформационные сущности) в машиночитаемом формате и имеет вид «субъект предикат – объект» (триплет) (рисунок 1).

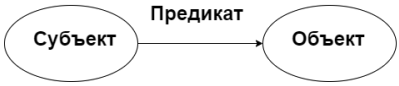


Рис. 1. Триплет RDF

Возможная структура триплета имеет следующий вид: СУБЪЕКТ (например, инсулин), ПРЕДИКАТ (например: может понижать), ОБЪЕКТ (например: уровень глюкозы в крови).

В [40] показано, что граф знаний должен быть источником достоверного знания, а не набором утверждений.

Особенностью графа знаний является не только способ представления знаний, но и способ получения новых знаний [45, 46]: «граф знаний собирает и интегрирует информацию в онтологию и применяет подсистему вывода для получения новых знаний». Существует множество исследований, предлагающих механизмы порождения нового знания: логические методы, основанные на

применении правил вывода [47] и статистические методы, основанные на векторных вложениях (embeddings) графов знаний [48], а также различные комбинации этих двух методов. В обзоре [24] показано, что крупномасштабные графы знаний эффективно используются для решения следующих задач: семантический поиск, поддержка принятия решений и генерирование ответов на вопросы [49]. Особо отмечено, что понятие «граф знаний» можно считать синонимом «базы знаний» [50] или «понятием, находящимся на уровень выше понятия базы знаний» [51].

Концепция открытых графов знаний была впервые реализована в 2007 г. в базе знаний DBpedia [52], построенной на основе интеллектуального анализа статей из онлайн-энциклопедии Wikipedia [53]. Непосредственно термин «граф знаний» введен компанией Google и связан с предложенным решением Google Knowledge Graph [54, 55].

Графы знаний, такие как Freebase и YAGO, широко используются в различных задачах обработки естественного языка (Natural Language Processing, NLP). Обучение представлению графов знаний направлено на отображение сущностей и отношений в непрерывное низкоразмерное векторное пространство. Обычные методы встраивания KG используют только триплеты KG и, таким образом, страдают от разреженности структуры. Проблема решается путем включения вспомогательных текстов сущностей, обычно описаний сущностей. Однако эти методы обычно фокусируются только на локальных последовательных последовательностях слов, но редко явно используют глобальную информацию о совпадении слов в корпусе.

Одной из самых больших проблем современной медицины является предоставление соответствующих, персонализированных и точных диагнозов и методов лечения на основе обработки данных, необходимых для персонализации медицины. В статье [56] рассмотрен подход к персонализации подбора лечения на основе графа знаний.

В работе [55] предлагается моделировать весь вспомогательный текстовый корпус и представить сквозную модель встраивания KG с улучшенным текстовым графом.

В работе [57] предложен метод извлечения отношений с использованием семантической регулярности в распределенном пространстве вложения векторов слов. Такой полууправляемый подход не зависит от синтаксиса языка и может быть использован для извлечения отношений из любого языка. Исследованы различные показатели сходства для маркировки извлеченных отношений оценкой достоверности семантической связи.

В работе [58] описывается система представления и интеллектуального анализа знаний, названная авторами семантическим графом знаний. В основе семантического графа знаний лежит использование инвертированного индекса наряду с дополнительным не инвертированным индексом для представления узлов (терминов) и ребер (документов в списках пересекающихся проводок для нескольких терминов/узлов). Предлагаемый семантический граф знаний способен динамически обнаруживать и оценивать взаимосвязи между заданным множеством сущностей посредством динамического создания множества вершин и ребер из компактного графа, автоматически сформированного в процессе анализа корпуса текстовых данных проблемной области.

В работе [59] предложено расширенное представление масштабируемого графа знаний на основе автоматического извлечения информации из корпуса новостных статей и анализ возможности использования графа знаний в качестве эффективного приложения для анализа и генерации представления знаний из извлеченного корпуса. Граф знаний состоит из базы знаний, построенной с использованием триплетов – выделенных отношений.

В работе [60] исследуется идея навигации по семантическим связям между извлеченными объектами как способ поиска в текстовом корпусе.

В работе [61] рассматриваются возможности применения концептуальных графов знаний для семантической разметки корпусов текстов. Построение метаданных на основе подобной разметки направлено на совершенствование алгоритмов решения определенных классов задач извлечения знаний и слабоструктурированных текстов. Предложен алгоритм автоматического построения концептуальных графов знаний, приводятся результаты экспериментов на текстах аннотаций научных статей.

В [62] предложена Knowledge Graph Language Model (KGLM) – нейросетевая языковая модель, аугментированная механизмами выбора и копирования информации из внешнего графа знаний, способная обращаться к внешнему источнику фактов, для генерирования фактически корректного текста. KGLM, в отличие от других новейших языковых моделей, требует размеченного набора обучающих данных.

В [63] предлагают методику предобучения Retrieval-Augmented Language Model (REALM), аугментирующую алгоритмы предобучения языковых моделей обученной системой поиска текстовых знаний. Как утверждают авторы, в отличие от моделей, которые содержат знания в своих параметрах, их подход эксплицитно выявляет роль знаний, так

как модели требуется решить, какие знания ей потребуются для рассуждений.

В [64] авторы отмечают, что, несмотря на значительный успех предобученных языковых моделей в эмпирических исследованиях, такие модели, будучи предобученными без учителя, не справляются с извлечением больших объемов знаний. Кроме того, авторы подчеркивают трудности, связанные с «внедрением» многообразных знаний в единую предобученную модель с помощью изменения исходных параметров таких моделей, в частности, риск катастрофической забывчивости.

В [65] приведены недостатки традиционных методов использования баз знаний для улучшения производительности рекуррентных нейронных сетей в задачах машинного чтения – низкая способность к обобщению признаков и необходимость конструирования признаков для достижения оптимальной производительности в отдельных задачах.

В [66] указано, что в совокупности связей между вершинами графов знаний содержатся дополнительные знания, в то же время, традиционные методы обучения на представлениях знаний (Knowledge Representation Learning, KRL) используют только триплеты, игнорируя контекстуализированную информацию. Предложена модель BERT-MK (BERT-based language model with Medical Knowledge) – предобученная языковая модель BERT, тонко настроенная с помощью обучения на крупномасштабном медицинском корпусе и аугментированную медицинскими знаниями, с помощью представлений, построенных на основе авторского подхода.

В [67] утверждается, что предобученные языковые модели наподобие BERT и RoBERTa, несмотря на высокие результаты в задачах обработки текста на естественном языке и способности к извлечению лингвистических знаний из неразмеченных текстовых корпусов, как правило, недостаточно способны к захвату фактов о мире. Авторы предлагают рассмотреть взаимную аугментацию языковых моделей с графами знаний, предлагая модель Knowledge Embedding and Pre- Trained LanguagE Representation (KEPLER).

В [68] авторы раскрывают проблему завершенности графов знаний и процесса их дополнения на основе оценки правдоподобности новых триплетов. Авторы предлагают рассматривать триплеты как текстовые последовательности и представляют Knowledge Graph Bidirectional Encoder Representations from Transformer (KG-BERT) – предобученную языковую модель BERT, дообученную для решения задачи оценки достоверности триплетов и их отношений.

Концептуальная модель графа знаний в [69] включает следующие сущности:

– «**POSOLOGY**» – описание схемы назначения лекарственного препарата и дополнительные атрибуты (dosage, duration, form, frequency, route of administration of the drug (route), name and identifier (id) of this node and special attribute (pos));

– «**PATIENT**» – пациент;

– «**DRUG**» – лекарственный препарат и атрибуты (the drug concentration (Strength), the node identifier, and the drug name and the special attribute (str));

– «**ADE**» – побочные эффекты и атрибуты (the name of the side effect and the node identifier);

– «**REASON**» – причина назначения и его атрибуты: the name of the reason and the node identifier,

что позволяет извлекать и формализовывать на основе техник выделения NER и отношений между сущностями с помощью предобученных моделей BERT знания о схемах лечения в виде графа знаний.

Таким образом, в науке и практике семантического анализа текстов накоплены определённые результаты, которые позволяют решать различные задачи. Задача извлечения знаний из медицинских текстов имеет свою специфику:

– для русскоязычного сегмента представлены лишь отдельные достаточно скромные по размеру корпусы размеченных медицинских текстов, что объясняется высокой трудоемкостью их сбора и отсутствием общепринятого протокола разметки (по сравнению с англоязычными решениями), учитывающего структуру и номенклатуру отечественной документации;

– отдельной проблемой является формирование открытых кодификаторов и глоссариев именованных сущностей для построения моделей NER и дальнейшей автоматизации конструирования графов знаний (для русскоязычного сегмента);

– недостаточное количество полилингвальных языковых предметно-ориентированных моделей семейства BERT, T5 и т.д.;

– необходимость интеграции и адаптации методов построения графовых моделей, языковых моделей и традиционных моделей баз знаний из других предметных областей.

3. Система извлечения знаний и автоматического построения онтологии и графа знаний проблемно-ориентированного корпуса клинических текстов. Знания являются динамической структурой, имеющей свой жизненный цикл, что

требует постоянной модификации и обновления данных в графах знаний на основе применения комплекса методов машинного обучения [70].

Основные этапы извлечения знаний и динамического обновления графа знаний включают:

1. распознавание/извлечение именованных сущностей (Named Entity Recognition/Extraction) – разграничение позиций упоминаний сущностей во входном тексте;
2. связывание/снятие омонимии сущностей, или семантическое аннотирование (Entity Linking/Disambiguation, Semantic Annotation) – ассоциирование упоминаний сущностей с подходящим и однозначным идентификатором в базе знаний;
3. извлечение терминов (Term Extraction) – извлечение основных фраз, которые обозначают концепты, релевантные к выбранной предметной области и описанные в корпусе, иногда включая иерархические отношения между концептами;
4. извлечение ключевых слов/фраз (Keyword/Keyphrase Extraction) – извлечение основных фраз, которые позволяют категоризировать тематику текста (в отличие от извлечения терминов, задача извлечения ключевых фраз заключается в описании именно текста, а не предметной области). Ключевые фразы также могут быть связаны с базой знаний;
5. тематическое моделирование/классификация (Topic Modeling, Classification) – кластеризация слов/фраз, которые часто встречаются совместно в сходном контексте. Эти кластеры затем ассоциируются с более абстрактными темами, с которыми связан текст;
6. маркирование/идентификация темы (Topic Labeling/Identification) – для кластеров слов, идентифицированных как абстрактные темы, извлечение одиночного термина или фразы, наилучшим образом характеризующей эти темы;
7. извлечение отношений (Relation Extraction) – извлечение потенциальных n-арных отношений из неструктурированных или полуструктурированных (таких как HTML-таблицы) источников.

На основе анализа основных этапов извлечения знаний и динамического обновления графа знаний предложена структурная схема системы автоматизированного построения онтологий и графов знаний (рисунок 2).

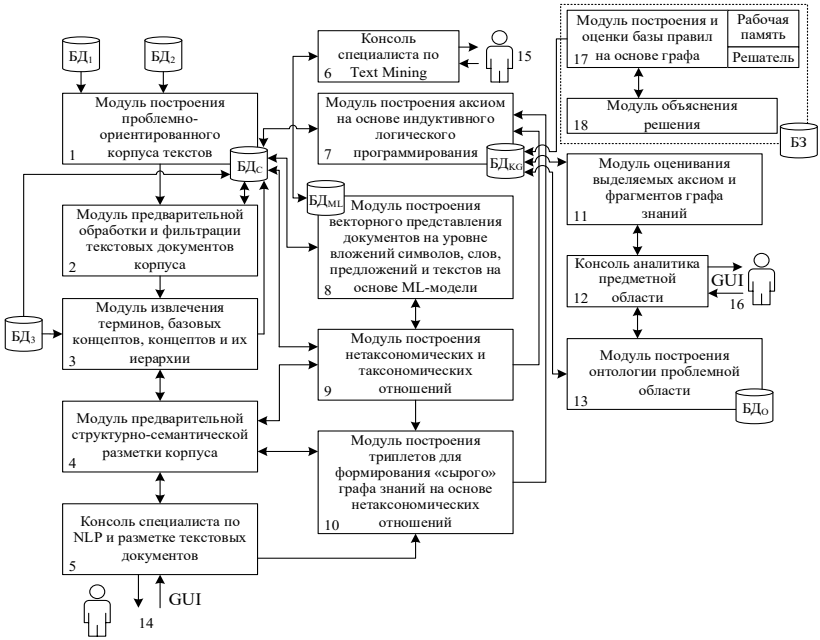


Рис. 2. Структурная схема системы автоматизированного построения онтологий и графов знаний (ML – модели машинного обучения, GUI – графический пользовательский интерфейс, NLP – обработка естественного языка, KG – граф знаний)

Модуль (1) построения проблемно-ориентированного корпуса текстов на основании данных их внешних источников ($БД_1$ — специализированные тексты – клинические описания и истории болезней пациентов, включая: эпикризы, результаты функциональной диагностики, осмотров и рекомендации врачей по лечению и $БД_2$ — описания клинических рекомендаций по лечению) позволяет собирать в документ-ориентированной $БД_C$ текстовые документы. Предварительная обработка и фильтрация собранных данных проводится в модуле (2) с помощью инструментов символьной фильтрации, фильтрации с помощью стоп-словарей и удаления нерелевантных фрагментов текстов, также применяются нейросетевые модели лемматизации (приведения в исходную форму) и стемминга, частеречной разметки. Модуль (3) позволяет извлекать из корпуса текстов с помощью инструментов NLP основные термины предметной области, сформировать кортеж основных концептов и их предварительную иерархию.

Модуль (4) позволяет выполнить предварительную структурно-семантическую разметку для выделения списка аннотаций, характеризующих заболевания, симптомы, лекарственные препараты, методы лечения, результаты применения методов лечения и т.д. Разметка выполняется с привлечением специалистов предметной области и специалистов по обработке естественно-языковых текстов (14) посредствам графической консоли (5). Специалист корректирует предварительно выделенные термины и иерархию концептов, собранные в граф-ориентированной $БД_3$.

Модуль (8) построения векторного представления документов на различных уровнях – от символьного до уровня отдельных текстов – с помощью нейросетевых моделей вложений позволяет получить формализованные текстовые описания, используемые для дальнейшей оценки семантического сходства и корректировки иерархии концептов. Построение векторного представления выполняется с помощью дообучаемых моделей машинного обучения, обновляемых и подготовленных в $БД_{ML}$ под контролем специалиста (15) по анализу и извлечению знаний из корпуса, корректирующего процесс посредствам консоли (6).

Модуль (10) позволяет для предварительно размеченного (автоматически и с привлечением эксперта) корпуса текстов формировать триплеты *«объект»-«действие»-«субъект»*, на основании анализа которых стоится «сырой» граф знаний. Модуль (9) выполняет построение нетаксономических и таксономических отношений для иерархии выделенных концептов и фрагментов текстов, что позволит перейти к более качественному построению графа знаний в модуле (7) формирования аксиом на основе индуктивно-логического подхода. Верификация множества сформулированных аксиом и фрагментов графа знаний (11) с привлечением специалиста предметно области (16) посредствам графического интерфейса консоли доступа (12) позволяет отфильтровать нерелевантные данные в граф-ориентированной $БД_{KG}$, предназначенной для хранения графа знаний.

Модуль (13) в процессе построения аксиом и формирования «сырого» графа знаний на основе триплетов позволяет представить иерархию концептов и отношений в виде автоматически построенной онтологии анализируемой проблемной области.

На рисунке 3 представлен верхний уровень функциональной модели автоматизированного построения онтологии и графа знаний, раскрывающий процесс сбора и обработки текстовых данных.

Декомпозиция процесса автоматизированного построения онтологий и графа знаний приведена на рисунке 3.

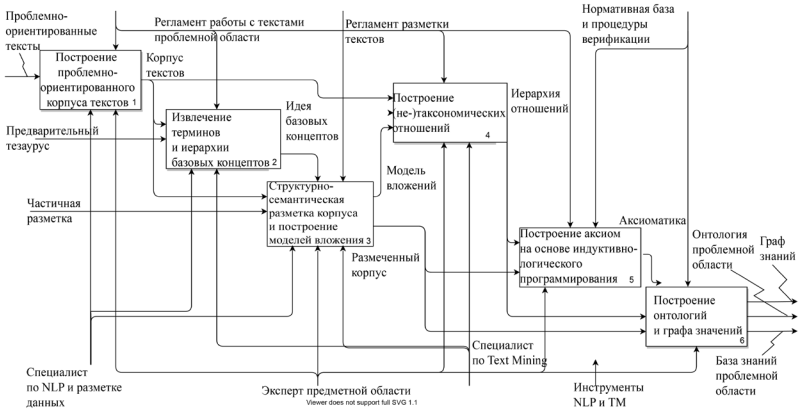


Рис. 3. Первый уровень декомпозиции функциональной модели процесса автоматизированного построения онтологии и графа знаний (NLP – обработка естественного языка, TM – извлечение знаний из текстов)

Первый уровень декомпозиции реализует основные этапы извлечения знаний и динамического обновления графа знаний с учетом задействованных ресурсов и инструментов, ограничений и условий использования, а также формируемого потока промежуточных данных. Структурно-функциональная организация позволяет перейти к проектированию архитектуры прототипа программной системы.

4. Эксперимент автоматизированного построения онтологии и графа знаний для корпуса проблемно-ориентированных текстов. Исходный корпус текстов проблемной области детально описан в [71, 1] и включает 162 документа обезличенных историй болезни пациентов педиатрического центра с болезнями органов дыхания, с аллергическими, нефрологическими и ревматическими болезнями. Пример документа – клинического описания – приведен ниже:

```
'ДИАГНОЗ: Бронхиальная астма, атопическая форма, легкое
интермиттирующее течение, неполная ремиссия.
Круглогодичный аллергический ринит, ремиссия.
Идиопатическая эпилепсия. Нарушение эмоционально-волевой
сферы. Парциальный дефицит когнитивных функций.
Ангиопатия сетчатки обоих глаз со спазмом артерий.
Хронический компенсированный тонзиллит. Остеохондропатия
позвоночника. <…>'
```

Комплекс алгоритмов обработки и структуризации текстовых данных, извлечения внутренней структуры на уровне частей документа (разделы, абзацы, предложения), частеречной разметки, семантической

разметки (извлечение именованных сущностей, терминов), нормализации и формализации на основе алгоритмов векторных вложений различного уровня с привлечением нейросетевых предобученных моделей позволяет выполнить для данного корпуса текстов основные этапы предобработки и формализации с применением известных инструментов [72, 73] описан в таблице 1.

Таблица 1. Структура конвейера NLP-конвейер

Этап	**Шаги**	**Действия**	**Инструменты**
Предобработка	Символьная фильтрация	Удаление нерелевантных символов, HTML-тегов	Набор регулярных выражений
	Токенизация	Разбивка текста на токены с помощью предобученной для русского языка нейросетевой модели	Razdel (фреймворк Natasha), Spacy, Stanza, nltk
	Фильтрация нерелевантных токенов	Удаление ссылок, нерелевантных сокращений	Регулярные выражения
Нормализация	Лемматизация	Приведение слов в исходную форму с помощью предобученной нейросетевой модели	Morph (фреймворк Natasha), pymorphy2, spacy
Постобработка	Частеречная фильтрация	Остаются только существительные, глаголы, прилагательные, наречия, местоимения	Morph (фреймворк Natasha)
	Извлечение именованных сущностей	Разметка тегами выделенных типов именованных сущностей	Natasha, spacy
	Фильтрация на основе стоп-словарей	Фильтрация нерелевантных лемм с помощью составного стоп-словаря, включающего наиболее часто встречающиеся слова корпуса текстов	NLTK-russian, english
	Формирование документа-строки	Объединение лемм в нормализованную строку-документ	

Например, расширенный стоп-словарь включает 343 токена: 'мм', 'мг', 'р', 'д', 'г', 'чсс', 'ст', 'год', 'рт', 'фв' и т.п.

В качестве разрешенных тегов частеречной разметки [74, 75] и последующего построения триплетов *«объект»-«действие»-«субъект»* использован кортеж: ['ADV', 'VERB', 'ADJ', 'NOUN', 'PROPN'] (таблица 2). Выбор тегов зависит от специфики текстового корпуса и может быть иным.

Таблица 2. Теги частеречной разметки

Тег-POS	**Описание**	**Пример**
ADJ	adjective, имя прилагательное	большой, старый, зеленый, непонятный
ADV	adverb, наречие	очень, завтра, вниз
AUX	auxiliary, вспомогательный глагол	есть, будет
NOUN	noun, имя существительное	девушка, кошка, земля
NUM	numeral, имя числительное	1, 20200, один, двадцать восемь, IV, MMXIV
PRON	pronoun, местоимение	я, ты, он, она, я, себя, кто-то
PROPN	proper noun, имя собственное	РФ, Людвиг Витгенштейн
VERB	verb, глагол	бежать, бежит, ест

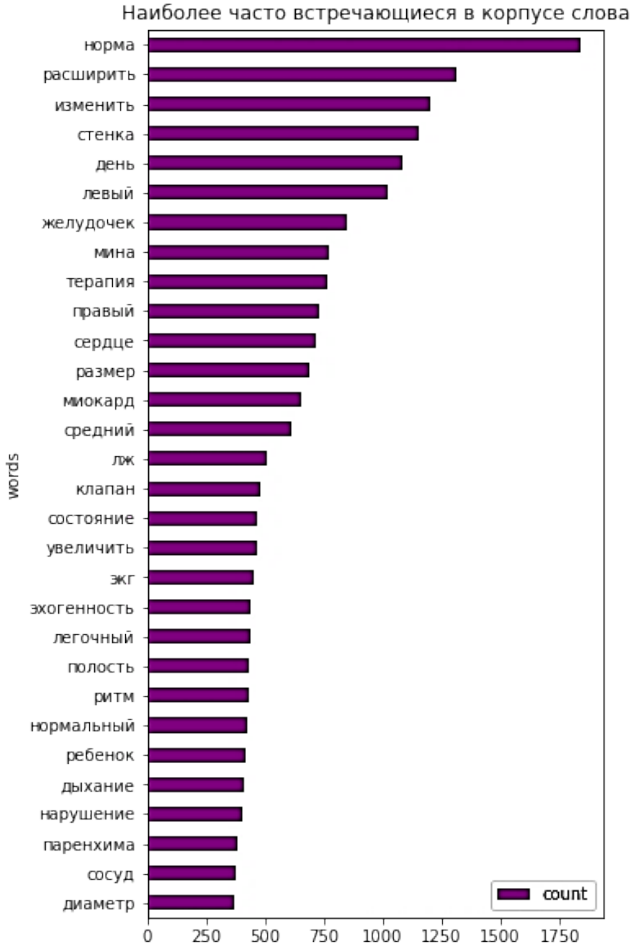
Пример фрагментов исходных текстов, их префильтрованный и нормализованный вид приведены в таблице 3.

В процессе анализа строится словарь корпуса текстов, визуализация которого в виде «облака слов» и диаграммы вхождения в корпус приведены на рисунке 4. «Облако слов» является удачным инструментом разведочного анализа, позволяющим визуально оценить частотность распределения ключевых слов и словосочетаний. Последующий количественный анализ диаграмм вхождения позволяет уточнить стоп-словарь и необходимость корректировки разметки.

Таблица 3. Иллюстрация фрагментов исходных, префильтрованных и нормализованных текстов корпуса

	Фрагменты исходных текстов	Префильтрованные тексты	Нормализованные тексты
0	НАХОДИЛСЯ НА ЛЕЧЕНИИ С 9.10.41 ПО 20.10...	находился на лечении с по г диагно...	[находиться, лечение, 20.10.41, диагноз, по...
1	ВОЗРАСТ: 9 лет (16.01.2021). \nНАХОДИЛСЯ НА ЛЕЧ...	возраст лет находился на лечении с ...	[возраст, год, находиться, лечение, 26.02.2031...
2	ДИАГНОЗ: Бронхиальная астма, атопическая, легк...	диагноз бронхиальная астма атопическая ...	[диагноз, бронхиальный, астма, атопический, ле...
3	ДИАГНОЗ: Бронхиальная астма, атопическая форма...	диагноз бронхиальная астма атопическая ...	[диагноз, бронхиальный, астма, атопический, фо...
4	ДИАГНОЗ: Бронхиальная астма, тяжелое персистир...	диагноз бронхиальная астма тяжелое пер...	[диагноз, бронхиальный, астма, тяжелый, персис...

а)



б)

Рис. 4. Визуализация частотного словаря корпуса текстов в виде «облака слов»: а) «облако слов» для корпуса текстов; б) диаграмма наиболее часто встречающихся слов (нормализованная форма)

Одной из обозначенных в ходе анализа проблем была названа трудоемкость выделения именованных сущностей и необходимость глубокой иерархической разметки корпуса текстов. Далее оценивается применение распространенных для русскоязычного домена моделей выделения именованных сущностей на основе нейронных сетей

(рекуррентных и глубоких архитектур), представленных в фреймворках Natasha и spacy, общего назначения (таблица 4).

Таблица 4. Количественный анализ выделенных именованных сущностей в корпусе текстов

Параметр	**Характеристика**			
Фреймворк	Natasha	Spacy	Базовый вариант I [1]	Модель II [1]
Количество выделенных именованных сущностей	773	1164	-	-
Общее пересечение	335		-	-
Оценка F1 меры выделения именованных сущностей	20,4 %	19,1 %	16,7 %	81,9 %
Пример выделяемых именованных сущностей (с меткой типа)	'Квинке': 'PER', 'ГКС': 'ORG', 'Институте иммунологии': 'ORG', 'ОВЛД': 'ORG', 'АБ': 'ORG', 'ЗОД': 'ORG', 'ФВД': 'ORG', 'НИИ педиатрии': 'ORG', 'И.И. Балаболкина': 'PER', 'Зев': 'PER', 'НЦЗД РАМН': 'ORG',	'АЛТ': 'ORG', 'Цитофлавин': 'PER', 'НИИ паразитологии': 'ORG', 'МБТ': 'ORG', 'Великий Новгород': 'LOC', 'Смекта': 'PER'	-	-

Из таблицы видно, что модели общего назначения, не дообученные на размеченном корпусе, не позволяют приблизиться по качеству выделения именованных сущностей к специализированным моделям, однако позволяют предварительно разметить корпус для дальнейшей верификации и уточнения разметки.

Формализованное представление текстов получено с помощью предобученной модели-трансформера BERT Large Model Multitask (cased) for Sentence Embeddings in Russian Language – предложенная специалистами RnD NLP SberDevices модель-трансформер многозадачного обучения для построения универсальной модели естественного языка на основе модели SBERT. T-SNE проекция формализованного корпуса текстов представлена ни рисунке 5.

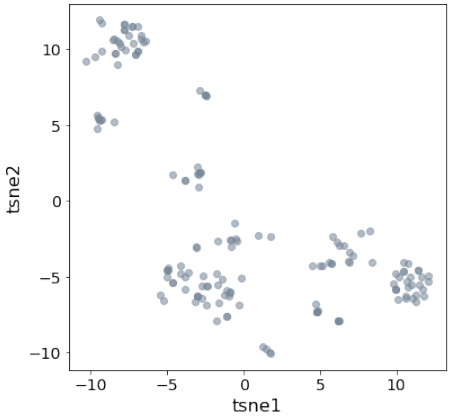


Рис. 5. T-SNE проекция формализованного с помощью предобученной модели-трансформера BERT корпуса текстов

Из рисунка видно, что тексты сгруппированы в устойчивые кластеры по степени их семантического сходства, что позволяет положительно оценить предыдущие этапы предобработки, нормализации и формализации текстовых описаний.

Автоматическое выделение триплетов «объект»-«действие»-«субъект» на основе частеречной разметки корпуса текстов позволяет выделить атомарные фрагменты «сырого» графа знаний. Основой триплетов являются зависимости [76, 77], представленные в таблице 5. Выбор зависимостей обусловлен тематикой и окраской текста.

Таблица 5. Универсальные синтаксические отношения для построения триплетов

Отношение	**Пояснение отношения**
nsubj	Именное подлежащее, которое является синтаксическим подлежащим.
nsubj:pass	Именная группа, которая является синтаксическим подлежащим пассивного предложения.
obj	Именная группа, обозначающая объект, на который воздействуют или который претерпевает изменение состояния или движения.
obl	Отношение используется для именных (существительное, местоимение, именное словосочетание), функционирующих как неосновной (косвенный) аргумент или дополнение.
nmod	Отношение используется для номинальных зависимостей другого существительного или именной фразы и функционально соответствует атрибуту или дополнению родительного падежа.
nummod	Числовой модификатор существительного — это любая числовая фраза, которая служит для изменения значения существительного с помощью количества.

Как правило, Subject и Object являются существительными, а Relation – глаголом.

Фрагмент исходной базы триплетов приведен в таблице 6.

Таблица 6. Фрагмент исходной базы триплетов

	Полное предложение	**subject**	**verb**	**object**	**Subj (нормальная форма)**	**Obj (нормальная форма)**
0	ЭПИКРИЗ Ребенок поступил в отделение впервые с...	ЭПИКРИЗ	поступил	отделение	эпикриз	отделение
1	НАХОДИЛСЯ НА ЛЕЧЕНИИ с 26 02 2031 по 4 02 2...	ДИАГНОЗ	НАХОДИЛСЯ	ЛЕЧЕНИИ	диагноз	лечении
2	ЭПИКРИЗ Ребенок поступил в отделение впервые с...	ЭПИКРИЗ	поступил	отделение	эпикриз	отделение
3	ЭПИКРИЗ Ребенок поступил в отделение повторно ...	ЭПИКРИЗ	поступил	отделение	эпикриз	отделение
4	ЭПИКРИЗ Мальчик поступил в клинику впервые с ...	Мальчик	поступил	клинику	мальчик	клинику

Фрагмент базы триплетов после фильтрации представлен в таблице 7.

Таблица 7. Фрагмент базы триплетов после фильтрации

	Subj (нормальная форма)	**Obj (нормальная форма)**	**verb**	**Полное предложение**
1	проба	фтизиатром	запрещена	Проба запрещена фтизиатром
2	орви	форме	болел	Дважды болел ОРВИ в легкой форме в марте 2054...
3	талия	ушка предсердие	сглажена	Талия сердца сглажена за счет ушка левого пред...

Продолжение Таблицы 7

	Subj (нормальная форма)	**Obj (нормальная форма)**	**verb**	**Полное предложение**
4	галотерапия	условиях галокамера	Рекомендовано	Консультация врача физиотерапевта Рекомендова...
5	бронх	уровня ветвь	прослежены	Бронхи прослежены до уровня субсегментарных ве...
6	бронх	уровня ветвь	прослежены	Бронхи прислежены до уровня субсегментарных ве...
7	бронх	уровня ветвь	прослеживаются	Бронхи прослеживаются до уровня субсег ментарн...
11	девочка	улучшением	выписана	Девочка выписана с улучшением
12	мальчик	терапию стационар	получал	Мальчик регулярно получал терапию в стационаре...
18	лимфоузел	сторон	структурны	Лимфоузлы структурны с двух сторон
19	мониторирование	стационарных	проведено	Исследование проведено на аппарате BPLab Мони...
21	жалоба	стабильным	оставалось	За период пребывания в отделении состояние реб...
25	преднизолон	снижением доз	добавлен	Терапия в отделении Дигоксин 0 000035 мг х 2 ...
32	голова	размере	увеличена	Голова резко увеличена в размере
38	вено-венозный	признаками эпитализация	кондуит	Кавапульмональный анастомоз проходим диаметро...
42	терапия	показаниям	назначена	Повторная госпитализация в НЦЗД РАМН при стаби...
54	эпикриз	отделение	поступила	ЭПИКРИЗ Девочка поступила в отделение впервые ...
63	терапия	объеме	продолжена	Терапия продолжена в прежнем объеме Следующа...
64	терапия	объеме	оставлена	Терапия оставлена в прежнем объеме После вып...

Отдельные выделенные типовые узлы «субъект», «объект», «действие» приведены ниже в порядке убывания частоты встречаемости:

знак	51
наблюдение	15
контроль	14
консультация	13
осмотр	11
госпитализация	10
инфекция	6
показатель	3
выздоровление	2
выполнение	2
отделение	2
условие	2

Выделенные триплеты (всего 162) сгруппированы в зависимости от выделенных узлов «субъект» и «объект». Вариация взаимной привязки триплетов приведена на рисунке 6.

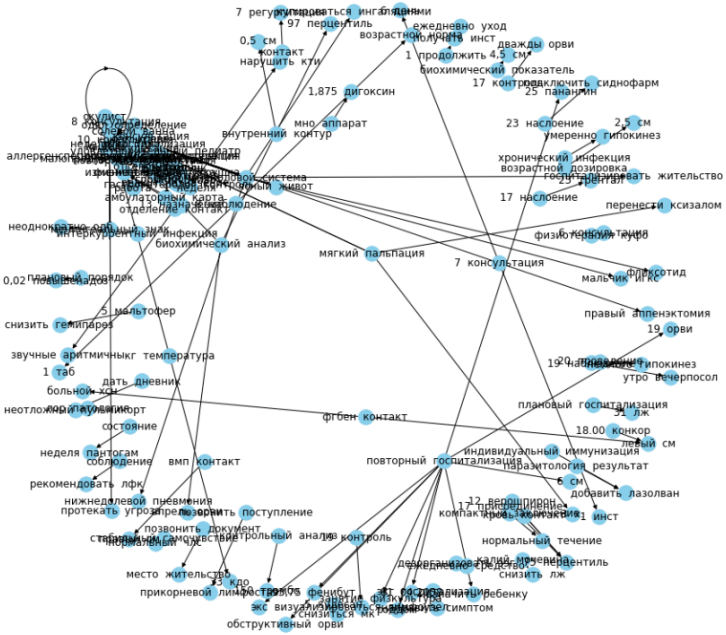


Рис. 6. Вариация взаимной привязки триплетов

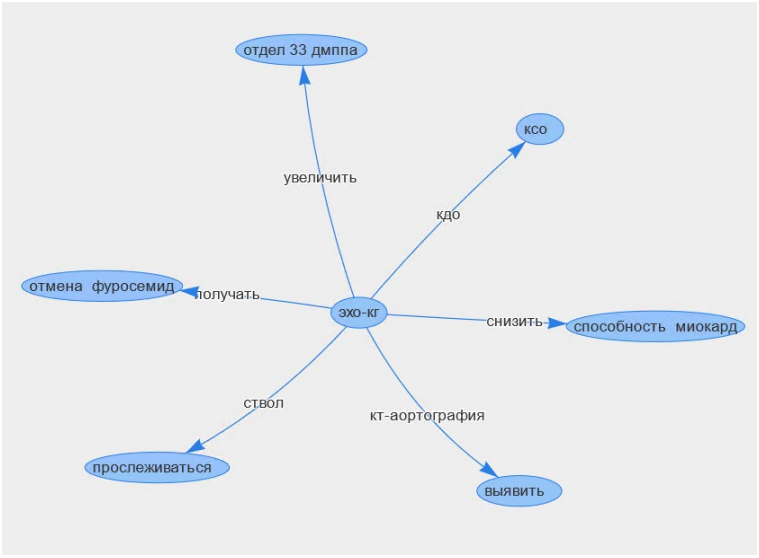
Для последующего анализа и построения, например, «вопрос-ответной» системы возможно построение отдельных подграфов на основе выделенного отношения (рисунок 7 – отношение «наблюдение»).

Рис. 7. Подграф с типом ребра «наблюдение»

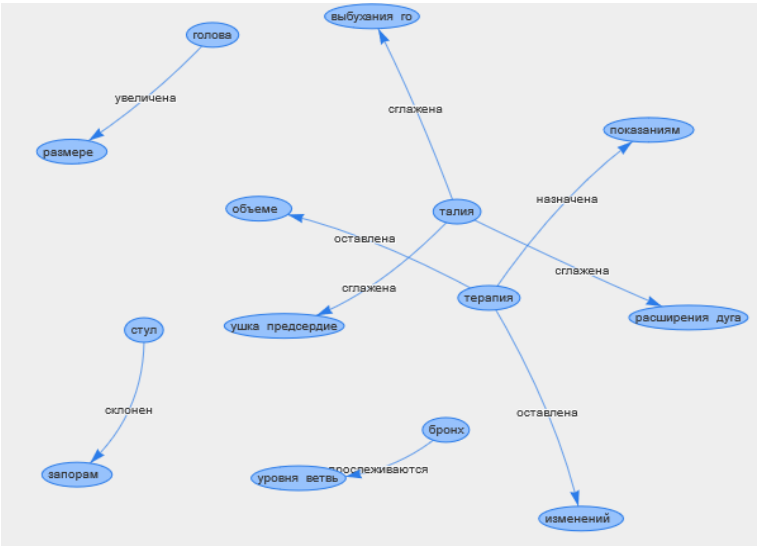
Далее для корпуса без предварительной разметки на основе извлеченного набора триплетов построен граф знаний (с фильтрацией по частоте встречаемости).

Фрагменты автоматически извлеченного графа знаний приведены на рисунке 8.

Дальнейшие анализ и верификация полученных триплетов при включении в конструируемую онтологию и граф знаний предлагается выполнить с помощью методов семантической оценки текстовых фрагментов, содержащихся в триплетах, с применением предобученной нейросетевой модели трансформера. Построение подобной модели требует расширенного корпуса текстов и является темой дальнейшего исследования.



а)



б)

Рис. 8. Фрагменты автоматически извлеченного графа знаний: а) Фрагмент 1 автоматически извлеченного графа знаний; б) Фрагмент 2 автоматически извлеченного графа знаний

5. Анализ и обсуждение. Проведен эксперимент на корпусе медицинских текстов заданной тематики (пульмонология, история болезней пациентов) без предварительной разметки с целью проверки предложенного решения по извлечению триплетов и конструирования на их основе графа знаний.

Показано, что модели общего назначения не позволяют приблизиться по качеству выделения именованных сущностей к специализированным моделям, однако, позволяют предварительно разметить корпус для дальнейшей верификации и уточнения разметки (оценка F1-меры для модели общего назначения – 20,4% по сравнению с вариантом использования словаря – 16,7%).

Применение языковых моделей трансформеров в сочетании с традиционным подходом по выделению триплетов, исходя из мировой практики, позволяет существенно расширить возможности по формализации знаний, построению графов знаний и решению задач построения систем поддержки принятия решений в клинической практике.

Результаты показывают необходимость более глубокой разметки корпуса текстовых документов для учета специфической лексики и обилия сокращений. На неразмеченном корпусе текстов предложенный инструмент показал свою удовлетворительную работоспособность ввиду выделения атомарных фрагментов, включаемых в автоматически формируемую онтологию.

6. Заключение. Результаты анализа проблемы извлечения знаний из периодически обновляемых клинических рекомендаций, а также анализ современного состояния подходов к автоматизации построения онтологий и графов знаний в задачах семантического моделирования проблемно-ориентированного корпуса текстов показали:

– применение моделей машинного обучения и интеллектуального анализа позволит на основе тонкой настройки существующих нейросетевых лингвистических моделей, построенных на обобщенных корпусах текстов, существенно повысить качество анализа исходных, «сырых» слабоструктурированных данных и снизить требования к предварительно построенным глоссариям и кодификаторам;

– для русскоязычного сегмента представлены лишь отдельные достаточно скромные по размеру корпусы размеченных медицинских текстов, что объясняется высокой трудоемкостью их сбора и отсутствием общепринятого протокола разметки (по

сравнению с англоязычными решениями), учитывающего структуру и номенклатуру отечественной документации;

– отдельной проблемой является формирование открытых кодификаторов и глоссариев именованных сущностей для построения моделей NER и дальнейшей автоматизации конструирования графов знаний (для русскоязычного сегмента);

– недостаточное количество полилингвальных языковых предметно-ориентированных моделей семейства BERT, T5 и т.д;

– необходимость интеграции и адаптации методов построения графовых моделей, языковых моделей и традиционных моделей баз знаний из других предметных областей.

Предлагаемая структурно-функциональная организация системы извлечения знаний и автоматического построения онтологии и графа знаний проблемно-ориентированного корпуса для конкретной предметной области основана на применении комплекса методов машинного обучения и позволяет перейти к проектированию архитектуры прототипа программной системы.

Экспериментальные исследования проведены на корпусе медицинских текстов заданной тематики (пульмонология, история болезней пациентов). Показано, что модели общего назначения позволяют предварительно разметить корпус для дальнейшей верификации и уточнения разметки, а также построения специализированных моделей (оценка F1-меры для модели общего назначения – 20,4% по сравнению с вариантом использования словаря – 16,7%). Применение языковых моделей трансформеров для выделения триплетов позволяет существенно расширить возможности по формализации знаний, построению графов знаний и решению задач построения систем поддержки принятия решений в клинической практике. Необходимым этапом является детализированная разметка корпуса текстовых документов для учета специфической лексики и обилия сокращений.

Литература

1. Баранов А.А. и др. Технологии комплексного интеллектуального анализа клинических данных // Вестник Российской академии медицинских наук. 2016. Т. 71. №. 2. С. 160-171.
2. Musen M.A., Middleton B., Greenes R.A. Clinical decision-support systems. In: Biomedical informatics. Springer. 2014. pp. 643–674. doi: 10.1007/978-1-4471-4474-8_22.
3. Rencis E. Natural language-based knowledge extraction in healthcare domain // Proceedings of the 2019 3rd International Conference on Information System and Data Mining. 2019. pp. 138-142.

4. Бледжянц Г.А., Саркисян М.А., Исакова Ю.А., Туманов Н.А., Попов А.Н., Бегмуродова Н.Ш. Ключевые технологии формирования искусственного интеллекта в медицине // Ремедиум. 2015. № 12. С. 10-15.
5. Рубрикатор клинических рекомендаций. URL: https://cr.minzdrav.gov.ru/clin_recomend (дата обращения: 01.10.2022).
6. Dligach D., Bethard S., Becker L., Miller T.A., Savova G.K. Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association (JAMIA). 2014. pp. 448–454. doi: 10.1136/amajnl-2013-001766.
7. Chikka V.R., Mariyasagayam N., Niwa Y., Karlapalem K. Information Extraction from Clinical Documents: Towards Disease/Disorder Template Filling. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer. 2015. pp. 389–401. doi: 10.1007/978-3-319-24027-5_41.
8. Shelmanov A.O, Smirnov I.V, Vishneva E.A. Information extraction from clinical texts in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue (2015). Issue 14 (21). 2015. pp. 560–572.
9. Кушнерова И.А., Акимов С.С. Перспективы применения искусственного интеллекта в медицине // Компьютерная интеграция производства и ИПИ-технологии: Сб. научн. тр. VIII Всероссийской научн. -практ. конф. (Оренбург, 16–17 ноября 2017 г.). Оренбург: ОГУ. 2017. С. 249–250.
10. Берестнева Е.В., Шаропин К.А., Жаркова О.С. Создание медицинских баз знаний с использованием деревьев решений // Успехи современной науки. 2016. Т. 2. № 10. С. 69–72.
11. Катасёв А.С., Ахатова Ч.Ф. Гибридная нейронечеткая модель интеллектуального анализа данных для формирования баз знаний мягких экспертных диагностических систем // Наука и образование: научное издание МГТУ им Н.Э. Баумана. 2012. № 12. С. 34–43.
12. Климов А.А., Куприяновский В.П., Гринько О.В., Покусаев О.Н. К вопросу обратного инжиниринга - путь от бумаги до цифровых онтологических правил для образовательных технологий // International Journal of Open Information Technologies. 2019. Т. 7. № 9. С. 82-91.
13. Муромцев Д., Волчек Д., Романов А. Индустриальные графы знаний - интеллектуальное ядро цифровой экономики // Control Engineering Россия. 2019. № 5(83). С. 32-39.
14. Asim M.N., Wasim M., Ghani Khan M.U., Mahmood W., Abbasi H.M. A survey of ontology learning techniques and applications // Database. 2018. vol. 2018. Bay101. https://doi.org/10.1093/database/bay101 (дата обращения: 26.06.2022).
15. Al-Aswadi F.N., Chan H.Y., Gan K.H. Automatic ontology construction from text: a review from shallow to deep learning trend // Artificial Intelligence Review. 2020. Т. 53. №. 6. pp. 3901-3928.
16. Ding Y., Foo S. Ontology research and development. Part 1-a review of ontology generation // Journal of information science. 2002. Т. 28. №. 2. pp. 123-136.
17. Волчек Д.Г., Романов А.А. Создание и обучение онтологий на основе анализа контекста и метаданных слабоструктурированного контента // Экономика: вчера, сегодня, завтра. 2020. Т. 10. № 1А. С. 303–312. doi: 10.34670/AR.2020.91.1.033.
18. Huang H. et al. Core-Concept-Seeded LDA for Ontology Learning // Procedia Computer Science. 2021. Т. 192. pp. 222-231.
19. Минин А.С., Чуприна С.И. Методы и средства построения онтологически управляемых систем приобретения знаний // Вестник пермского университета. Математика. Механика. Информатика. 2021. №. 4 (55). С. 25-34.

20. Максимов А.И., Молодов В.А., Рунов С.С. Об одном способе представления знаний в медицинских интеллектуальных системах // Современные инновации. 2021. № 1 (39). С. 48–50.
21. Кулешов С.В., Зайцева А.А., Марков В.С. Ассоциативно-онтологический подход к обработке текстов на естественном языке // Интеллектуальные технологии на транспорте. 2015. № 4 (4). С. 40–45.
22. Михайлов С.Н, Малашенко О.И., Зайцева А.А. Методика инфологического анализа семантического содержания обращений пациентов для организации электронной записи // Труды СПИИРАН. 2015. № 5 (42). С. 140–154.
23. Harnoune A. et al. BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis // Computer Methods and Programs in Biomedicine Update. 2021. vol. 1. no. 100042.
24. Понкин Д.И. Концепт предобученных языковых моделей в контексте инженерии знаний // International Journal of Open Information Technologies. 2020. № 9. С. 18–29. URL: http://injoit.org/index.php/j1 (дата обращения: 24.09.2022).
25. Землянский С.А., Аксёнов С.В., Лызин И.А., Берестнева О.Г. Тематическое моделирование в контексте медицинских текстов // Доклады ТУСУР. 2021. Т. 24. № 4. С. 58–64.
26. Нугуманова А.Б., Байбурин Е.М., Мансурова М.Е., Барахнин В.Б. Автоматическое извлечение решеток понятий из медицинских текстов на основе комбинации анализа формальных понятий и технологий бутстраппинга // Вестник НГУ. Серия: Информационные технологии. 2018. Т. 16. № 4. С. 140-152.
27. Petroni F., Rocktaschel T., Lewis P. Language Models as Knowledge Bases? // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'2019). ong Kong (China): Association for Computational Linguistics. 2019. pp. 2463–2473.
28. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. URL: https://arxiv.org/abs/1810.04805 (дата обращения: 24.09.2022).
29. Lee J., Yoon W., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining Bioinformatics // arXiv preprint arXiv: 1901.08746. URL: https://arxiv.org/abs/1901.08746 (дата обращения: 24.09.2022).
30. Alsentzer E., Murphy J.R., Boag W., Weng W.-H., Jin D., Naumann T., McDermott M. Publicly available clinical bert embeddings // arXiv preprint arXiv:1904.03323. URL: https://arxiv.org/pdf/1904.03323.pdf (дата обращения: 24.09.2022).
31. Sboev A. et al. An analysis of full-size Russian complexly NER labelled corpus of Internet user reviews on the drugs based on deep learning and language neural nets //arXiv preprint arXiv:2105.00059. URL: https://arxiv.org/pdf/2105.00059.pdf (дата обращения: 24.09.2022).
32. Russian Drug Review corpus by Sag team (RDRS). URL: https://sagteam.ru/med-corpus/stata/#ours-Pharm2021arxiv (дата обращения: 24.09.2022).
33. Tutubalina E. et al. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews // Bioinformatics. 2021. Т. 37. № 2. С. 243–249.
34. Aronson A.R, Lang F.M. An overview of MetaMap: historical perspective and recent advances // Journal of the American Medical Informatics Association. 2010. №17 (3). pp. 229–236. doi:10.1136/jamia.2009.002733.

35. Schuyler P.L, Hole W.T, Tuttle M.S, Sherertz D.D. The UMLS Metathesaurus: representing different views of biomedical concepts // Bulletin of the Medical Library Association. 1993. № 81 (2). pp. 217–222.
36. Unified Medical Language System (UMLS). URL: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/ (дата обращения: 04.10.2022).
37. Государственный реестр лекарственных средств. URL: http://grls.rosminzdrav.ru/Default.aspx (дата обращения: 24.09.2022).
38. Гусев П.Ю. Обработка текстов и подготовка моделей векторизации для программного комплекса классификации научных текстов // Моделирование, оптимизация и информационные технологии. 2021. Т. 9. № 1. С. 6–7.
39. Kelly L., Goeuriot L., Suominen H., Schreck T., Leroy G., Mowery D.L. et al. Overview of the SHARE/CLEF eHealth evaluation lab 2014 // Springer. 2014. pp. 172–191. doi:10.1007/978-3-319-11382-1_17.
40. McCusker J.P., Erickson J.S., Chastain K., Rashid S., Weerawarana R., Bax M., McGuinness D.L. What is a knowledge graph? URL: https://www.semantic-web-journal.net/ (дата обращения: 25.09.2022).
41. Апанович З.В. Эволюция понятия и жизненного цикла графов знаний // Системная информатика. 2020. №.16. С. 57–74.
42. Färber M., Bartscherer F., Menne C., Rettinger A. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago // Semantic Web. 2016. pp. 1–53.
43. Huang Z., Yang J., Harmelen F.V., Hu Q. Constructing disease-centric knowledge graphs: a case study for depression (short version) // Proceedings of the Conference on Artificial Intelligence in Medicine in Europe. Springer. 2017. pp. 48–52.
44. World Wide Web Consortium (W3C). URL: https://www.w3.org/ (дата обращения: 25.09.2022).
45. Ehrlinger L., Woß W. Towards a definition of knowledge graphs // SEMANTiCS (Posters, Demos, SuCCESS). 2016. no. 48.
46. Ernst P., Siu A., Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences // BMC bioinformatics. 2015. № 16 (157). https://doi.org/10.1186/s12859-015-0549-5.
47. Stepanova D., Gad-Elrab M.H., Ho T.V. Rule Induction and Reasoning over Knowledge Graphs // Reasoning Web International Summer School // Springer, Cham. 2018. pp. 142-172.
48. Nickel M., Murphy K., Tresp V., Gabrilovich E. A review of relational machine learning for knowledge graphs // Proceedings of the IEEE, 104(1). 2016. vol. 104 (1). pp. 11–33.
49. Yao L., Mao C., Luo Y. KG-BERT: BERT for Knowledge Graph Completion // arXiv preprint arXiv: 1810.04805. URL: https://arxiv.org/abs/1810.04805 (дата обращения: 24.09.2022).
50. Ji S., Pan S., Cambria E. et al. A Survey on Knowledge Graphs: Representation, Acquisition and Applications // arXiv preprint arXiv: 2002.00388. URL: https://arxiv.org/abs/2002.00388 (дата обращения: 24.09.2022).
51. Yoo S.-Y., Jeong O.-K. Automating the expansion of a knowledge graph // Expert Systems with Applications. 2020. vol. 141. no. 112965.
52. Глобальный и единый доступ к графам знаний. URL: https://www.dbpedia.org/ (дата обращения: 07.07.2022).
53. Википедия. Свободная энциклопедия. URL: www.en.wikipedia.org/wiki/Main_Page (дата обращения: 08.07.2022).
54. Adams T. Google and the future of search: Amit Singhal and the knowledge graph // The Guardian. 2013. Т. 19.
55. Ehrlinger L., Wöß W. Towards a definition of knowledge graphs // SEMANTiCS (Posters, Demos, SuCCESS). 2016. Т. 48. №. 1-4. p. 2.

56. Silva M.C., Faria D., Pesquita C. Matching Multiple Ontologies to Build a Knowledge Graph for Personalized Medicine // European Semantic Web Conference. – Springer, Cham. 2022. pp. 461-477.
57. Kumar K., Manocha S. Constructing knowledge graph from unstructured text // Self. 2015. T. 3. 4 p.
58. Grainger T. et al. The Semantic Knowledge Graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain // 2016 IEEE international conference on data science and advanced analytics (DSAA). IEEE. 2016. pp. 420-429.
59. Lakshika M., Caldera H.A. Knowledge Graphs Representation for Event-Related E-News Articles // Machine Learning and Knowledge Extraction. 2021. T. 3. №. 4. pp. 802-818.
60. Bernasconi E., Ceriani M., Mecella M. Exploring a Text Corpus via a Knowledge Graph // IRCDL. 2021. pp. 91-102.
61. Богатырев М.Ю., Тюхтин В.В. Построение концептуальных графов как элементов семантической разметки текстов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог – 2009».
62. Logan R., Liu N.F., Peters M.E. et al. Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Italy: Association for Computational Linguistics. 2019. pp. 5962–5971.
63. Guu K., Lee K., Tung Z. et al. REALM: Retrieval Augmented Language Model Pre-Training // arXiv preprint arXiv: 2002.08909. URL: https://arxiv.org/abs/2002.00388 (дата обращения: 24.09.2022).
64. Wang R., Tang D., Duan N. etc. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters // arXiv preprint arXiv:2002.01808. https://arxiv.org/abs/2002.01808 (дата обращения: 24.09.2022).
65. Yang B., Mitchell T. Leveraging Knowledge Bases in LSTMs for Improving Machine Reading // arXiv preprint arXiv:1902.09091. https://arxiv.org/abs/1902.09091 (дата обращения: 24.09.2022).
66. He B., Zhou D., Xiao J. et al. Integrating Graph Contextualized Knowledge into Pre-trained Language Models // arXiv preprint arXiv:1912.00147. https://arxiv.org/abs/1912.00147 (дата обращения: 24.09.2022).
67. Wang X., Gao T., Zhu Z. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation // arXiv preprint arXiv:1911.06136. https://arxiv.org/abs/1911.06136 (дата обращения: 24.09.2022).
68. Weng J., Gao Y., Qiu J. et al. Construction and Application of Teaching System Based on Crowdsourcing Knowledge Graph // Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference (CCKS 2019). China. Singapore: Springer. 2019. pp. 25 – 37.
69. Harnoune A. et al. BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis // Computer Methods and Programs in Biomedicine Update. 2021. vol. 1. no. 100042.
70. Martinez-Rodriguez J.L., Hogan A., Lopez-Arevalo I. Information extraction meets the semantic web: a survey // Semantic Web. 2020. T. 11. №. 2. pp. 255-335.
71. Баранов А.А. и др. Методы и средства комплексного интеллектуального анализа медицинских данных // Труды Института системного анализа Российской академии наук. 2015. Т. 65. №. 2. С. 81-93.
72. Васильев В.И. и др. Методика оценки актуальных угроз и уязвимостей на основе технологий когнитивного моделирования и Text Mining // Системы управления, связи и безопасности. 2021. №. 3. С. 110-134.

73. Васильев В.И., Вульфин А.М., Кучкарова Н.В. Автоматизация анализа уязвимостей программного обеспечения на основе технологии Text Mining // Вопросы кибербезопасности. 2020. №. 4 (38). С. 22-31.
74. Веб-сервис для хостинга IT-проектов и их совместной разработки. URL: https://github.com/Koziev/rupostagger (дата обращения: 26.09.2022).
75. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных // М.: Изд-во НИУ ВШЭ. 2017. с. 269.
76. De Marneffe M.C. et al. Universal Stanford dependencies: A cross-linguistic typology // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 4585-4592.
77. Простой граф знаний на текстовых данных. Хабр: Коллективный блог. URL: https://habr.com/ru/post/559110/. (дата обращения 08.07.2022).

Зулкарнеев Рустэм Халитович — д-р мед. наук, профессор, профессор, кафедра пропедевтики внутренних болезней с курсом физиотерапии, Башкирский государственный медицинский университет. Область научных интересов: исследования в области кардиореспираторной физиологии, пульмонологии, кардиологии, медицинской информатики. Число научных публикаций — 230. zrustem@ufanet.ru; улица Ленина, 3, 450000, Уфа, Россия; р.т.: +7(917)420-6925.

Юсупова Нафиса Исламовна — д-р техн. наук, профессор, профессор, кафедра вычислительной математики и кибернетики, Уфимский государственный авиационный технический университет. Область научных интересов: интеллектуальные методы обработки информации и управления с приложениями в социальных, экономических и технических системах. Число научных публикаций — 560. yussupova@ugatu.ac.ru; улица Карла Маркса, 12, 450000, Уфа, Россия; р.т.: +7(917)343-5953.

Сметанина Ольга Николаевна — д-р техн. наук, профессор, профессор, кафедра вычислительной математики и кибернетики, Уфимский государственный авиационный технический университет. Область научных интересов: интеллектуальные методы обработки информации и управления с приложениями в социальных, экономических и технических системах. Число научных публикаций — 250. smoljushka@mail.ru; улица Карла Маркса, 12, 450000, Уфа, Россия; р.т.: +7(917)755-2214.

Гаянова Майя Марсовна — канд. техн. наук, доцент, доцент, кафедра вычислительной математики и кибернетики, Уфимский государственный авиационный технический университет. Область научных интересов: интеллектуальные методы обработки информации и управления с приложениями в социальных, экономических и технических системах. Число научных публикаций — 100. maya.gayanova@gmail.com; улица Карла Маркса, 12, 450000, Уфа, Россия; р.т.: +7(917)409-7014.

Вульфин Алексей Михайлович — канд. техн. наук, доцент, доцент, кафедра вычислительной техники и защиты информации, Уфимский государственный авиационный технический университет. Область научных интересов: исследования в области интеллектуального анализа данных и моделирования сложных технических систем. Число научных публикаций — 160. vulfin.am@ugatu.su; улица Карла Маркса, 12, 450000, Уфа, Россия; р.т.: +7(917)400-2189.

Поддержка исследований. Работа выполнена при финансовой поддержке РНФ (проект № 22-19-00471).

DOI 10.15622/ia.21.6.4

R. ZULKARNEEV, N. YUSUPOVA, O. SMETANINA, M. GAYANOVA, A. VULFIN

METHOD AND MODELS OF EXTRACTION OF KNOWLEDGE FROM MEDICAL DOCUMENTS

Zulkarneev R., Yusupova N., Smetanina O., Gayanova M., Vulfin A. **Method and Models of Extraction of Knowledge from Medical Documents.**

Abstract. The paper analyzes the problem of extracting knowledge from clinical recommendations presented in the form of semi-structured corpora of text documents in natural language, taking into account their periodic updating. The considered methods of intellectual analysis of the accumulated arrays of medical data make it possible to automate a number of tasks aimed at improving the quality of medical care due to significant decision support in the treatment process. A brief review of well-known publications has been made, highlighting approaches to automating the construction of ontologies and knowledge graphs in the problems of semantic modeling of a problem-oriented text corpus. The structural and functional organization of the system of knowledge extraction and automatic construction of an ontology and a knowledge graph of a problem-oriented corpus for a specific subject area is presented. The main stages of knowledge extraction and dynamic updating of the knowledge graph are considered: named entity extraction, semantic annotation, term and keyword extraction, topic modeling, topic identification, and relationship extraction. The formalized representation of texts was obtained using a pre-trained BERT transformer model. The automatic selection of triplets "object" - "action" - "subject" based on part-of-speech markup of the text corpus was used to construct fragments of the knowledge graph. An experiment was carried out on a corpus of medical texts on a given topic (162 documents of depersonalized case histories of patients of a pediatric center) without preliminary markup in order to test the proposed solution for extracting triplets and constructing a knowledge graph based on them. An analysis of the experimental results confirms the need for a deeper markup of the corpus of text documents to take into account the specifics of medical text documents. For an unmarked corpus of texts, the proposed solution demonstrates satisfactory performance in view of the selection of atomic fragments included in the automatically generated ontology.

Keywords: clinical texts, information extraction, machine learning, medical data mining, automatic ontology building, knowledge graphs.

References

1. Baranov A.A. et al. [Technologies of complex intellectual analysis of clinical data]. Vestnik Rossijskoj akademii medicinskih nauk - Bulletin of the Russian Academy of Medical Sciences. 2016. vol. 71. №. 2. pp. 160-171. (In Russ.).
2. Musen M.A., Middleton B., Greenes R.A. Clinical decision-support systems. In: Biomedical informatics. Springer, 2014. pp. 643–674. doi: 10.1007/978-1-4471-4474-8_22.
3. Rencis E. Natural language-based knowledge extraction in healthcare domain. Proceedings of the 2019 3rd International Conference on Information System and Data Mining. 2019. pp. 138-142.
4. Bledzhjanc G.A., Sarkisjan M.A., Isakova Ju.A., Tumanov N.A., Popov A.N., Begmurodova N.Sh. Kljuchevye tehnologii formirovanija iskusstvennogo intellekta v medicine [Key technologies of artificial intelligence formation in medicine]. Remedium. 2015. № 12. pp.10–15. (In Russ.).
5. Rubrikator klinicheskih rekomendacij [Rubricator of clinical recommendations]. Available at: https://cr.minzdrav.gov.ru/clin_recomend (accessed 01.10.2022). (In Russ.).

6. Dligach D., Bethard S., Becker L., Miller T.A., Savova G.K. Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association (JAMIA). 2014. pp. 448–454. doi: 10.1136/amajnl-2013-001766.
7. Chikka V.R., Mariyasagayam N., Niwa Y., Karlapalem K. Information Extraction from Clinical Documents: Towards Disease/Disorder Template Filling. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer. 2015. pp. 389–401. doi: 10.1007/978-3-319-24027-5_41.
8. Shelmanov A.O, Smirnov I.V, Vishneva E.A. Information extraction from clinical texts in Russian. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue (2015). Issue 14 (21). 2015. pp. 560–572.
9. Kushnerova I.A., Akimov S.S. [Prospects for the use of artificial intelligence in medicine] Komp'juternaja integracija proizvodstva i IPI-tehnologii. Sb. nauchn. tr. VIII Vserossijskoj nauchn. -prakt. konf. [Computer integration of production and IPI technology]. Orenburg, 2017, pp. 249–250. (In Russ.).
10. Berestneva E.V., Sharopin K.A., Zharkova O.S. [Creating medical knowledge bases using decision trees]. Uspehi sovremennoj nauki – Successes of modern science. 2016. № 10. pp. 69–72. (In Russ.).
11. Katasjov A.S., Ahatova Ch.F. [Hybrid neuro fuzzy data mining model for the formation of knowledge bases of soft expert diagnostic systems]. Nauka i obrazovanie: nauchnoe izdanie MGTU im N.Je. Baumana – Science and Education: scientific publication of Bauman Moscow State Technical University. 2012. № 12. pp. 34–43. (In Russ.).
12. Klimov A.A., Kuprijanovskij V.P., Grin'ko O.V., Pokusaev O.N. [On the issue of reverse engineering - the path from paper to digital ontological rules for educational technologies] International Journal of Open Information Technologies. 2019. vol. 7. № 9. pp. 82-91. (In Russ.).
13. Muromcev D., Volchek D., Romanov A. [Industrial knowledge graphs - the intellectual core of the digital economy]. Control Engineering Rossija. - Control Engineering Russia. 2019. № 5(83). pp. 32-39. (In Russ.).
14. Asim M.N., Wasim M., Ghani Khan M.U., Mahmood W., Abbasi H.M. A survey of ontology learning techniques and applications. Database. vol. 2018. Bay101. Available at: https://doi.org/10.1093/database/bay101 (accessed 26.06.2022)
15. Al-Aswadi F.N., Chan H.Y., Gan K.H. Automatic ontology construction from text: a review from shallow to deep learning trend. Artificial Intelligence Review. 2020. vol. 53. №. 6. pp. 3901-3928.
16. Ding Y., Foo S. Ontology research and development. Part 1-a review of ontology generation. Journal of information science. 2002. vol. 28. №. 2. pp. 123-136.
17. Volchek D.G., Romanov A.A. [Creation and training of ontologies based on the analysis of the context and metadata of semi-structured content]. Jekonomika: vchera, segodnja, zavtra - Economics: yesterday, today, tomorrow. 2020. vol. 10. № 1A. pp. 303-312. doi: 10.34670/AR.2020.91.1.033 (In Russ.).
18. Huang H. et al. Core-Concept-Seeded LDA for Ontology Learning. Procedia Computer Science. 2021. vol. 192. pp. 222-231.
19. Minin A.S., Chuprina S.I. [Methods and tools for constructing ontologically controlled knowledge acquisition systems]. Vestnik permskogo universiteta. Matematika. Mehanika. Informatika. - Perm university bulletin. Maths. Mechanics. Informatics. 2021. №. 4 (55). pp. 25-34. (In Russ.).
20. Maksimov A.I., Molodov V.A., Runov S.S. [About one way of presenting knowledge in medical intelligent systems]. Sovremennye innovacii – Modern innovations. 2021. №1 (39). pp. 48–50. (In Russ.).

21. Kuleshov S.V., Zajceva A.A., Markov V.S. [Associative-ontological approach to natural language text processing]. Intellektual'nye tehnologii na transporte – Intelligent technologies in transport. 2015. № 4 (4). pp. 40–45. (In Russ.).
22. Mihajlov S.N, Malashenko O.I., Zajceva A.A. [Methodology of infological analysis of the semantic content of patients' appeals for the organization of electronic records]. Trudy SPIIRAN – Works of SPIIRAN. 2015. № 5 (42). pp. 140–154. (In Russ.).
23. Harnoune A. et al. BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update. 2021. vol. 1. no. 100042.
24. Ponkin D.I. [The concept of pre-trained language models in the context of knowledge engineering]. International Journal of Open Information Technologies. 2020. № 9. pp. 18–29. URL: http://injoit.org/index.php/j1 (accessed: 24.09.2022). (In Russ.).
25. Zemljanskij S.A., Aksjonov S.V., Lyzin I.A., Berestneva O.G. [Thematic modeling in the context of medical texts]. Doklady TUSUR – TUSUR reports. 2021. vol. 24. № 4. pp. 58–64. (In Russ.).
26. Nugumanova A.B., Bajburin E.M., Mansurova M.E., Barahnin V.B. [Automatic extraction of concept lattices from medical texts based on a combination of formal concept analysis and bootstrapping technologies]. Vestnik NGU. Serija: Informacionnye tehnologii – Bulletin of the NSU. Series: Information Technology. 2018. vol. 16. № 4. pp. 140-152. (In Russ.).
27. Petroni F., Rocktaschel T., Lewis P. Language Models as Knowledge Bases? Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'2019). Hong Kong (China): Association for Computational Linguistics. 2019. pp. 2463–2473.
28. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. URL: https://arxiv.org/abs/1810.04805 (дата обращения: 24.09.2022).
29. Lee J., Yoon W., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining Bioinformatics. arXiv preprint arXiv: 1901.08746. URL: https://arxiv.org/abs/1901.08746 (дата обращения: 24.09.2022).
30. Alsentzer E., Murphy J.R., Boag W., Weng W.-H., Jin D., Naumann T., McDermott M. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323. URL: https://arxiv.org/pdf/1904.03323.pdf (дата обращения: 24.09.2022).
31. Sboev A. et al. An analysis of full-size Russian complexly NER labelled corpus of Internet user reviews on the drugs based on deep learning and language neural nets. arXiv preprint arXiv:2105.00059. URL: https://arxiv.org/pdf/2105.00059.pdf (дата обращения: 24.09.2022).
32. Russian Drug Review corpus by Sag team (RDRS). URL: https://sagteam.ru/med-corpus/stata/#ours-Pharm2021arxiv (дата обращения: 24.09.2022).
33. Tutubalina E. et al. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. Bioinformatics. 2021. vol. 37. № 2. pp. 243–249.
34. Aronson A.R., Lang F.M. An overview of MetaMap: historical perspective and recent advances // Journal of the American Medical Informatics Association. 2010. № 17 (3). pp. 229–236. doi:10.1136/jamia.2009.002733.
35. Schuyler P.L, Hole W.T, Tuttle M.S, Sherertz D.D. The UMLS Metathesaurus: representing different views of biomedical concepts. Bulletin of the Medical Library Association. 1993. № 81 (2). pp. 217–222.

36. Unified Medical Language System (UMLS). Available at: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/ (accessed: 04.10.2022).
37. Gosudarstvennyj reestr lekarstvennyh sredstv [State Register of Medicines]. Available at: http://grls.rosminzdrav.ru/Default.aspx (accessed: 24.09.2022). (In Russ.).
38. Gusev P.Ju. [Text processing and preparation of vectorization models for the scientific text classification software package]. Modelirovanie, optimizacija i informacionnye tehnologii – Modeling, optimization and information technology. 2021. vol. 9. №1. pp. 6–7. (In Russ.).
39. Kelly L., Goeuriot L., Suominen H., Schreck T., Leroy G., Mowery D.L. et al. Overview of the SHARE/CLEF eHealth evaluation lab 2014. Springer. 2014. pp. 172–191. doi:10.1007/978-3-319-11382-1_17.
40. McCusker J.P., Erickson J.S., Chastain K., Rashid S., Weerawarana R., Bax M., McGuinness D.L. What is a knowledge graph? URL: https://www.semantic-web-journal.net/ (дата обращения: 25.09.2022).
41. Apanovich Z.V. [Evolution of the concept and life cycle of knowledge graphs]. Sistemnaja informatika – System Informatics. 2020. №.16. pp. 57–74. (In Russ.).
42. Färber M., Bartscherer F., Menne C., Rettinger A. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web. 2016. pp. 1–53.
43. Huang Z., Yang J., Harmelen F.V., Hu Q. Constructing disease-centric knowledge graphs: a case study for depression (short version). Proceedings of the Conference on Artificial Intelligence in Medicine in Europe. Springer. 2017. pp. 48–52.
44. World Wide Web Consortium (W3C). URL: https://www.w3.org/ (дата обращения: 25.09.2022).
45. Ehrlinger L., Woß W. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS). 2016. no. 48.
46. Ernst P., Siu A., Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC bioinformatics. 2015. №16 (157). https://doi.org/10.1186/s12859-015-0549-5.
47. Stepanova D., Gad-Elrab M.H., Ho T.V. Rule Induction and Reasoning over Knowledge Graphs. Reasoning Web International Summer School. Springer, Cham. 2018. pp. 142-172.
48. Nickel M., Murphy K., Tresp V., Gabrilovich E. A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1). 2016. vol. 104 (1). pp. 11–33.
49. Yao L., Mao C., Luo Y. KG-BERT: BERT for Knowledge Graph Completion. arXiv preprint arXiv: 1810.04805. URL: https://arxiv.org/abs/1810.04805 (дата обращения: 24.09.2022).
50. Ji S., Pan S., Cambria E. et al. A Survey on Knowledge Graphs: Representation, Acquisition and Applications.arXiv preprint arXiv: 2002.00388. URL: https://arxiv.org/abs/2002.00388 (дата обращения: 24.09.2022).
51. Yoo S.-Y., Jeong O.-K. Automating the expansion of a knowledge graph. Expert Systems with Applications. 2020. vol. 141. no. 112965.
52. Global and Unified Access to Knowledge Graphs. Available at: https://www.dbpedia.org/ (accessed 07.07.2022).
53. Vikipedija. Svobodnaja jenciklopedija. [Wikipedia. The Free Encyclopedia]. Available at: www.en.wikipedia.org/wiki/Main_Page (accessed 08.07.2022). (In Russ.).
54. Adams T. Google and the future of search: Amit Singhal and the knowledge graph. The Guardian. 2013. vol. 19.
55. Ehrlinger L., Wöß W. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS). 2016. vol. 48. №. 1-4. p. 2.

56. Silva M.C., Faria D., Pesquita C. Matching Multiple Ontologies to Build a Knowledge Graph for Personalized Medicine. European Semantic Web Conference. Springer, Cham. 2022. pp. 461-477.
57. Kumar K., Manocha S. Constructing knowledge graph from unstructured text. Self. 2015. vol. 3. p. 4.
58. Grainger T. et al. The Semantic Knowledge Graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain. 2016 IEEE international conference on data science and advanced analytics (DSAA). IEEE. 2016. pp. 420-429.
59. Lakshika M., Caldera H.A. Knowledge Graphs Representation for Event-Related E-News Articles. Machine Learning and Knowledge Extraction. 2021. vol. 3. №. 4. pp. 802-818.
60. Bernasconi E., Ceriani M., Mecella M. Exploring a Text Corpus via a Knowledge Graph. IRCDL. 2021. pp. 91-102.
61. Bogatyrev M.Ju., Tjuhtin V.V. [Construction of conceptual graphs as elements of semantic markup of texts. Computational Linguistics and Intelligent Technologies]. Po materialam. ezhegodnoj Mezhdunarodnoj konferencii «Dialog – 2009». [By materials. annual International Conference "Dialogue - 2009"]. (In Russ.).
62. Logan R., Liu N.F., Peters M.E. et al. Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Italy: Association for Computational Linguistics. 2019. pp. 5962–5971.
63. Guu K., Lee K., Tung Z. et al. REALM: Retrieval Augmented Language Model Pre-Training. arXiv preprint arXiv: 2002.08909. URL: https://arxiv.org/abs/2002.00388 (дата обращения: 24.09.2022).
64. Wang R., Tang D., Duan N. etc. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. arXiv preprint arXiv:2002.01808. https://arxiv.org/abs/2002.01808 (дата обращения: 24.09.2022).
65. Yang B., Mitchell T. Leveraging Knowledge Bases in LSTMs for Improving Machine Reading. arXiv preprint arXiv:1902.09091. https://arxiv.org/abs/1902.09091 (дата обращения: 24.09.2022).
66. He B., Zhou D., Xiao J. et al. Integrating Graph Contextualized Knowledge into Pre-trained Language Models. arXiv preprint arXiv:1912.00147. https://arxiv.org/abs/1912.00147 (дата обращения: 24.09.2022).
67. Wang X., Gao T., Zhu Z. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. arXiv preprint arXiv:1911.06136. https://arxiv.org/abs/1911.06136 (дата обращения: 24.09.2022).
68. Weng J., Gao Y., Qiu J. et al. Construction and Application of Teaching System Based on Crowdsourcing Knowledge Graph. Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference (CCKS 2019). China. Singapore: Springer. 2019. pp. 25 – 37.
69. Harnoune A. et al. BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update. 2021. vol. 1. no. 100042.
70. Martinez-Rodriguez J.L., Hogan A., Lopez-Arevalo I. Information extraction meets the semantic web: a survey. Semantic Web. 2020. vol. 11. №. 2. pp. 255-335.
71. Baranov A.A. et al. [Methods and means of complex intellectual analysis of medical data.] Trudy Instituta sistemnogo analiza Rossijskoj akademii nauk. 2015. vol. 65. №. 2. pp. 81-93. (In Russ.).
72. Vasil'ev V.I. et al. [Methodology for assessing current threats and vulnerabilities based on cognitive modeling and Text Mining technologies]. Sistemy upravlenija, svjazi i bezopasnosti. – Control, communication and security systems. 2021. №. 3. pp. 110-134. (In Russ.).

73. Vasil'ev V.I., Vul'fin A.M., Kuchkarova N.V. [Automation of software vulnerability analysis based on Text Mining technology]. Voprosy kiberbezopasnosti. - Cyber security issues. 2020. №. 4 (38). pp. 22-31. (In Russ.).
74. Veb-servis dlja hostinga IT-proektov i ih sovmestnoj razrabotki [A web service for hosting IT projects and their joint development]. Available at: https://github.com/Koziev/rupostagger (In Russ.).
75. Bol'shakova E.I., Voroncov K.V., Efremova N.Je., Klyshinskij Je.S., Lukashevich N.V., Sapin A.S. Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i analiz dannyh [Automatic natural language text processing and data analysis]. Moscow: HSE Publishing House. 2017. p. 269 (In Russ.).
76. De Marneffe M.C. et al. Universal Stanford dependencies: A cross-linguistic typology. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 4585-4592.
77. Prostoj graf znanij na tekstovyh dannyh. Habr: Kollektivnyj blog. [A simple knowledge graph on textual data. Habr: Collective blog]. Available at: https://habr.com/ru/post/559110/. (accessed 08.07.2022). (In Russ.).

Zulkarneev Rustem — Ph.D., Dr.Sci., Professor, Professor, Department of propaedeutics of internal diseases with a course of physiotherapy, Bashkir State Medical University. Research interests: research in the field of cardiorespiratory physiology, pulmonology, cardiology, medical informatics. The number of publications — 230. zrustem@ufanet.ru; 3, Lenina St., 450000, Ufa, Russia; office phone: +7(917)420-6925.

Yusupova Nafisa — Ph.D., Dr.Sci., Professor, Professor, Department of computational mathematics and cybernetics, Ufa State Aviation Technical University. Research interests: intelligent methods of information processing and management with applications in social, economic and technical systems. The number of publications — 560. yussupova@ugatu.ac.ru; 12, Karl Marx St., 450000, Ufa, Russia; office phone: +7(917)343-5953.

Smetanina Olga — Ph.D., Dr.Sci., Professor, Professor, Department of computational mathematics and cybernetics, Ufa State Aviation Technical University. Research interests: intelligent methods of information processing and management with applications in social, economic and technical systems. The number of publications — 250. smoljushka@mail.ru; 12, Karl Marx St., 450000, Ufa, Russia; office phone: +7(917)755-2214.

Gayanova Maya — Ph.D., Associate Professor, Associate professor, Department of computational mathematics and cybernetics, Ufa State Aviation Technical University. Research interests: intelligent methods of information processing and management with applications in social, economic and technical systems. The number of publications — 100. maya.gayanova@gmail.com; 12, Karl Marx St., 450000, Ufa, Russia; office phone: +7(917)409-7014.

Vulfin Alexey — Ph.D., Associate Professor, Associate professor, Department of computer science and information protection, Ufa State Aviation Technical University. Research interests: research in the field of data mining and modeling of complex technical systems. The number of publications — 160. vulfin.am@ugatu.su; 12, Karl Marx St., 450000, Ufa, Russia; office phone: +7(917)400-2189.

Acknowledgements. This work was supported by the Russian Science Foundation (project no. 22-19-00471).