

Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В.ЛОМОНОСОВА»
(МГУ)

Научно-исследовательский вычислительный центр

УДК
004.8, 519.767.6
Рег. № НИОКТР
AAAA-A20-120121690111-5
Рег. № ИКРБС

УТВЕРЖДАЮ
Директор
Научно-исследовательский
вычислительный центр,
д.ф.-м.н., проф., член-корр. РАН

_____ В.В. Воеводин

«_____» _____ Г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:

Методы структурирования трудноформализуемых предметных областей на
основе автоматизированного формирования больших графов знаний и
онтологий по разнородным потокам текстовых данных
(промежуточный)

Руководитель НИР:

заведующий лабораторией,
кандидат
физико-математических наук,



Добров Б.В.

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель НИР,
заведующий ла-
бораторией, кан-
дидат физико-
математических наук



Добров Б.В.
(все разделы)

Исполнители:

научный сотрудник,
кандидат филологи-
ческих наук



Герасимова А.А.
(все разделы)

заведующий лабо-
раторией, кандидат
филологических наук,
доктор филологиче-
ских наук, доцент по
кафедре



Гращенко П.В.
(все разделы)

программист 2 кате-
гории



Григорьев Д.А.
(все разделы)

программист 2 кате-
гории



Ионова Н.Р.
(все разделы)

стажёр-
исследователь



Ковалев Г.П.
(все разделы)

программист 1 кате-
гории



Кремнева М.Д.
(все разделы)

ведущий науч-
ный сотрудник,
кандидат физико-
математических наук,
доктор технических
наук



Лукашевич Н.В.
(все разделы)

ведущий научный сотрудник, канди- дат филологических наук, доктор фило- логических наук, доцент/с.н.с. по спе- циальности		Лютикова Е.А. (все разделы)
специалист		Мячина А.В. (все разделы)
стажёр- исследователь		Паско Л.И. (все разделы)
специалист		Рожков И.С. (все разделы)
специалист		Сидоров А.В. (все разделы)
специалист		Студеникина К.А. (все разделы)
старший науч- ный сотрудник, кандидат физико- математических наук		Тихомиров М.М. (все разделы)
стажер		Чернышев Д.И. (все разделы)
специалист		Штернов С.В. (все разделы)
научный сотрудник, кандидат филологи- ческих наук		Ярошенко П.В. (все разделы)

РЕФЕРАТ

Ключевые слова:

большие данные, обработка естественного языка, графы знаний, глубокое обучение, онтологии, представление знаний, текстовые данные, информационно-аналитические системы, лингвистические онтологии, пре-добученные языковые модели, большие языковые модели, нейронные сети

Ключевые слова по-английски:

bigdata, information-analytical system, text data, knowledge graph, linguistic ontology, knowledge representation, pretrained language models, natural language processing, ontology

Представлен научно-технический отчет по Этапу 6 2025 г. «Исследование методов применения больших нейросетевых языковых моделей для задач автоматизированного формирования графов знаний» НИР «Методы структурирования трудноформализуемых предметных областей на основе автоматизированного формирования больших графов знаний и онтологий по разнородным потокам текстовых данных». Проведены исследования по следующим направлениям:

(1) исследование методов улучшения качества результатов информационного поиска для решения задач RAG (подключения новых знаний) для больших языковых моделей;

(2) исследование методов автоматического структурирования материалов при формировании аналитических статей;

(3) исследование методов организации научно-технических знаний.

Основным направлением исследований является рассмотрение возможностей совместного применения феноменологических знаний (онтологий, графов знаний), явно описывающих свойства сущностей, и знаний о контекстах сущностей, неявно представленных в больших нейросетевых языковых моделях (далее – LLM, от англ. Large Language Models).

В рамках Этапа 6 наибольший интерес представляют исследования методов обработки, анализа и организации результатов анализа научно-технических неструктурированных данных. В том числе с точки зрения формирования графов знаний (как взаимосвязанной сети понятий онтологии, именованных и иного типа сущностей, а также фрагментов документов), прежде всего, в формате вики-ресурсов, включающих, по сути, аналитические обзорные рефераты, формируемые в результате поиска релевантных документов в больших базах данных.

По результатам Этапа 6 НИР опубликовано 24 научных работы. Из них: Q1 1, WoS 2, Scopus 14, RSCI 4, Ядро РИНЦ 6, РИНЦ 7.

ВВЕДЕНИЕ

В 2025 году работа по НИР «Методы структурирования трудноформализуемых предметных областей на основе автоматизированного формирования больших графов знаний и онтологий по разнородным потокам текстовых данных» шла в рамках Этапа 6 «Исследование методов применения больших нейросетевых языковых моделей для задач автоматизированного формирования графов знаний» по следующим направлениям:

(1) исследовать методы улучшения качества результатов информационного поиска для решения задач RAG (подключения новых знаний) для больших языковых моделей;

(2) исследовать методы автоматического структурирования материалов при формировании аналитических статей;

(3) исследовать методы организации научно-технических знаний.

Данная работа продолжает цикл исследований, выполненных в 2020-2024 гг. (см. Приложение Б):

- Этап1. 2020 г. «Разработка методов автоматического пополнения больших лингвистических онтологий таксономическими отношениями, методов извлечения редких типов именованных сущностей»;

- Этап2. 2021 г. «Разработка методов автоматизированного формирования больших лингвистических предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний»;

- Этап 3. 2022 г. «Разработка методов автоматизированного формирования больших графов знаний предметной области»;

- Этап 4. 2023 г. «Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области»;

- Этап 5. 2024 г. «Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий».

Основное направление исследований смещается в сторону рассмотрения возможностей совместного применения феноменологических знаний (онтологий, графов знаний), явно описывающих свойства сущностей, и знаний о контекстах сущностей, неявно представленных в больших нейросетевых языковых моделях (далее – LLM, от англ. Large Language Models).

В рамках Этапа 6 наибольший интерес продолжает представлять исследование методов обработки, анализа и организации результатов анализа научно-технических неструктурированных данных. В том числе с точки зрения формирования графов знаний (как взаимосвязанной сети понятий онтологии, именованных и иного типа сущностей, а также фрагментов документов), прежде всего, в формате вики-ресурсов, включающих, по сути, аналитические обзорные рефераты, формируемые в результате поиска релевантных документов в больших базах данных.

ОСНОВНАЯ ЧАСТЬ

1.1. План работ на 2025 год по теме

План работ по теме на 2025 год предполагал проведение исследований по трем направлениям:

- Исследование методов улучшения качества результатов информационного поиска для решения задач RAG (подключения новых знаний) для больших языковых моделей;
- Исследование методов автоматического структурирования материалов при формировании аналитических статей;
- Исследование методов организации научно-технических знаний.

1.2. Исследование методов улучшения качества результатов информационного поиска для решения задач RAG (подключения новых знаний) для больших языковых моделей

Задачи информационного поиска находились в центре интереса большого количества исследователей в 1990е и в начале 2000х годов. Однако, в связи с существенной монополизацией рынка информационного поиска компаниями типа Google или Яндекс, которые существенно использовали данные о пользовательских переходах, недоступные обычным исследователям, для улучшения моделей релевантности массового поиска, а также плохой масштабируемости моделей корпоративного поиска в специальных предметных областях, интерес научного сообщества к проблематике информационного поиска уменьшился.

Вместе с тем, в 2010-2020е годы произошли значимые перемены в представлениях о структуре человеческих знаний, отражаемых в слабоструктурированных данных:

- появление нейросетевых «плотных» векторных представлений («эмбедингов») контекстов слов и высказываний, а также развитие моделей поиска методами «малых миров», ввело в рассмотрение новый вид поиска – векторный (Dense Search), когда эмбединги запроса сопоставляются с эмбедингами фрагментов документов;
- появление больших предобученных нейросетевых языковых моделей (далее – LLM, от англ. Large Language Models), которые в состоянии уложить огромные массивы текстов в матрицы весов нейросетевых моделей, моделирующих глубокие контексты высказываний, позволяет ставить вопрос об автоматической оценке релевантности смысла запроса смыслу высказываний в документе, что нивелирует преимущество накопленных массивов пользовательских оценок больших корпораций.

Довольно быстро выяснилось, что стандартные «фундаментальные» (или «базовые») LLM имеют ограничения по «понимаю» в специальных предметных областях, в силу того, что тексты специальных предметных областей им предъявлялись в меньшем объеме. Также имеется проблема актуализации знаний, например, содержащихся в новостных текстах, которые с большим запаздыванием попадают в «базы неявно представленных знаний контекстов» LLM (и быстро устаревают).

Естественным было появление концепции RAG (Retrieval Augmented Generation, «поисковой дополненной генерации»), когда недостающие знания для LLM получаются в результате поиска, прежде всего векторного, с использованием информационно-поисковой системы. При этом качество поиска в организации технологической цепочки, использующей RAG является ключевым.

Базисом для исследования качества поиска являются наборы данных, на которых можно в автоматическом режиме исследовать качество удовлетворения различных информационных потребностей пользователей.

В рамках НИОКР был [3] создан RusBEIR, комплексный бенчмарк, разработанный для оценки моделей информационного поиска (ИП) на русском языке без предварительного обучения. Он включает 17 наборов данных из различных областей, интегрируя адаптированные, переведенные и вновь созданные наборы данных, что позволяет систематически сравнивать лексические и нейронные модели. Полученные результаты показывают важность предварительной обработки для лексических моделей в морфологически богатых языках и подтверждают, что классические методы типа BM25 является надежным базовым набором данных для поиска полных документов и во многих ситуациях успешно конкурируют с моделями векторного поиска. Нейронные модели, такие как mE5-large и BGE-M3, демонстрируют превосходную производительность на большинстве наборов данных, но сталкиваются с проблемами при поиске длинных документов из-за ограничений по размеру входных данных. RusBEIR предлагает унифицированную платформу с открытым исходным кодом, которая способствует исследованиям в области информационного поиска на русском языке.

В рамках темы продолжались работы по углублению понимания природы представления знаний в LLM. В работе [20] представлены результаты исследований по адаптации мультязычных LLM для обработки русского языка. Прямая адаптация таких моделей к новому языку, например русскому, сопряжена с риском катастрофического забывания исходных знаний и требует значительных вычислительных ресурсов. Представлена комплексная и вычислительно эффективная методология Ruadapt для языковой адаптации LLM с заменой токенизации. Полная адаптация одного варианта Qwen3-8B модели по этой методологии требует менее 2000 GPU-часов, а последующая адаптация других вариантов в 10 раз меньше, благодаря делимости результатов каждого шага процедуры. Оптимальная конфигурация процедуры адаптации позволяет добиться до 80% ускорения генерации с полным сохранением навыков работы с длинным контекстом и незначительными потерями эффективности анализа пользовательских инструкций. Приведено подробное эмпирическое исследование каждого шага адаптации с целью определения оптимальных гиперпараметров, а также ключевых этапов и их влияния на итоговое качество. Выработанные рекомендации применяются в текущем поколении моделей Ruadapt, включая RuadaptQwen3-32B-Hybrid. Модели, код и наборы данных опубликованы в открытом доступе, предлагая научному сообществу проверенную и экономически целесообразную стратегию создания высококачественных языковых моделей.

1.3. Исследование методов автоматического структурирования материалов при формировании аналитических статей

1.3.1. Формирование энциклопедических статей

В связи с прогрессом в области абстрактного реферирования с использованием возможности генеративных LLM, в том числе обзорного реферирования, стала актуальной задача формирования аналитических статей с качеством, сопоставимым с результатами работы профессиональных экспертов.

Одной из перспективных является задача автоматического формирования энциклопедических статей в стиле Википедии (в том числе по причине наличия большого свободно доступного материала для тестирования).

Создание текстов, требующих высокой фактической точности и строгого форматирования, таких как энциклопедические статьи, сопряжено с многочисленными трудностями: как и где собрать необходимую информацию, как структурировать её в связный и хорошо отформатированный текст, и как обеспечить отсутствие фактических ошибок в составленной статье. Предложено [4] решение этих проблем для русского языка в виде системы генерации энциклопедических статей, которая извлекает самые последние и актуальные знания из научных публикаций в онлайн-библиотеке eLIBRARY.RU и структурирует их как единый контекст для ввода в генеративную модель. Для оценки как влияния извлеченных знаний на содержание итоговых текстов, так и общего качества генерации, рассмотрено несколько стратегий подсказок, некоторые из которых не используют контекст, найденный в публикациях, и сравнили эти подходы с помощью автоматических метрик и оценки экспертов.

В работе [13] представлен RuWikiBench - открытый бенчмарк на основе “Рувики” для оценки способностей больших языковых моделей воспроизводить статьи в стиле Википедии, основанный на трех задачах: отбор релевантных источников, построение структуры статьи и генерация секций. Результаты тестирования популярных открытых LLM показывают, что даже в идеальных условиях лучшие модели не всегда следуют экспертной логике составления энциклопедических материалов: даже при совершенной работе системы подбора материалов модели не могут воспроизвести эталонное оглавление, а качество генерации секций практически не зависит от числа параметров.

1.3.2. Интерпретация возможностей LLM

Проблемы генерации аналитических статей во многом определяются недостаточным пониманием природы LLM и недостаточным развитием инструментов управления генерации. Это заставляет исследовать различные лингвистические возможности генерации с использованием LLM.

В работе [19] были исследованы сходства и различия лингвистической компетенции носителей языка и больших языковых моделей. Материалом для сравнения служит созданный участниками НИР Корпус Вариативного Согласования (КВас). Он содержит 6803 предложения с оценками по шкале от 1 до 7, полученными при проведении 20 синтаксических экспериментов по изучению вариативного согласования. Корпус фиксирует средние оценки русских предложений с различными условиями согласования, полученные от носителей языка, и позволяет выяснить, как LLM справляются с градуальной оценкой приемлемости.

Был проведен эксперимент, в рамках которого тестировались четыре большие языковые модели: преимущественно русскоязычные YandexGPT 5 Pro и GigaChat 2 Max, а также мультиязычные Llama 3.3 70B и Mistral Large. Для каждой модели было опробовано два режима тестирования: zero-shot, содержащий только инструкцию, и few-shot, где добавлены тренировочные предложения и их оценки.

Поскольку данные в КВас демонстрируют различный уровень приемлемости в зависимости от экспериментальных условий, подсчет только средней ошибки для

предсказанных моделями оценок мог бы оказаться недостаточно показательным. По этой причине была разработана метрика, позволяющая оценить, какая доля контрастов между экспериментальными условиями, релевантными для людей, выявляется с помощью LLM. Результаты показывают, что среднее значение ошибки меньше для предложений без вариативного согласования. Примеры с согласовательной вариативностью оказываются сложнее для LLM. Качество моделей проседает для одного и того же типа конструкций – сочинения. Модели значительно лучше определяют контрасты для конструкций с постпозитивными относительными предложениями, количественными конструкциями и управляющими квантификаторами. Наиболее точное совпадение при выделении значимых контрастов по сравнению с носителями языка демонстрирует Mistral. Наименьшее количество контрастов выделяет модель Llama. Русскоязычные модели занимают промежуточную позицию, при этом YandexGPT превосходит GigaChat. Добавление примеров в режиме few-shot улучшает среднее качество, но различие незначительно.

Таким образом, результаты показывают, что качество решения задачи градуальной оценки приемлемости сильно отличается для разных классов лингвистических феноменов.

Сравнение моделей демонстрирует, что для достижения лучшего качества наиболее важным оказывается количество параметров модели, которое, однако, может быть компенсировано объемом русскоязычных данных при обучении.

Работа [22] посвящена исследованию влияния таких ситуаций на применимость стандартных методов оценки качества синтаксического анализа к результатам работы синтаксического анализатора на основе больших языковых моделей.

При применении больших языковых моделей для задачи синтаксического анализа возникают принципиально новые ситуации: завершение синтаксического анализа с ошибкой и существенное изменение предложения, для которого выполняется синтаксический анализ (в том числе удаление или добавление слов). Экспериментальная оценка проведена для синтаксического анализатора U-DepPLLaMA на тестовой выборке датасета предложений с синтаксической разметкой Taiga. Установлено, что к результатам синтаксического анализа на основе больших языковых моделей нельзя применять метод оценки на основе вычисления среднего значения метрик UAS и LAS на тестовых предложениях. Кроме того, без существенной модификации нельзя использовать стандартный алгоритм выравнивания наборов токенов. Реализация исследования доступна по адресу https://github.com/Derinhelm/parser_stat/tree/llm_taiga.

С развитием предобученных языковых моделей, таких как BERT, их применение охватывает всё более ответственные сферы — от рекомендательных систем до медицинской диагностики. Однако рост сложности моделей требует не менее активного развития методов интерпретации, позволяющих понять, на основе каких признаков модель принимает решения. В данной работе [17] исследуются подходы к объяснению поведения BERT в задачах текстовой классификации. Основное внимание уделено сравнению двух парадигм: современных методов, основанных на маскированном языковом моделировании и промпт-обучении, и традиционных техник, таких как LIME и методы на основе векторной близости. Экспериментальный анализ проводится на датасетах Web of Science и 20Newsgroups. Для оценки качества интерпретаций

используется построение графиков активации, что позволяет визуализировать значимость входных токенов для конечного предсказания.

1.3.3. Исследование методов определения тональности

Оценка тональности является важной частью решения задачи формирования аналитических материалов. Потому что, с одной стороны, перспективность, достоинства и недостатки рассматриваемых в аналитических статьях процессов может быть оценена по динамике встречаемости выделяемых объектов (понятий, именованных сущностей), с другой стороны, текст может содержать явные или неявные оценки авторов первичных документов или оценки упоминаемых акторов, или указание позитивных и негативных характеристик рассматриваемых процессов.

Поэтому в рамках НИР уделяется значительное внимание исследованию вопросов оценки тональности, которая может быть выражена с использованием большого количества вариаций.

В статье [5] описана процедура организации и результаты научного соревнования RuOpinionNE-2024 в рамках методологии Dialogue Evaluation по извлечению структурированных мнений из российских новостных текстов. Задача конкурса состоит в извлечении кортежей мнений для заданного предложения; кортежи состоят из носителя тональности, его цели, выражения и тональности от носителя к цели. В общей сложности на задачу было подано более 100 заявок, что позволяет говорить о получении представления о современном уровне решения задачи соревнования. Участники экспериментировали в основном с большими языковыми моделями в форматах с нулевым количеством примеров, малым количеством примеров и тонкой настройкой. Лучший результат на тестовом наборе был получен при тонкой настройке большой языковой модели. В работе также сравнено 30 подсказок и 11 языковых моделей с открытым исходным кодом с 3-32 миллиардами параметров в режимах с одним и десятью примерами и выявили лучшие модели и подсказки.

В статье [21] на данных соревнования RuOpinionNE-2024 описывается исследование применения больших языковых моделей (LLM) в различных постановках задачи анализа тональности, как например: поиск мнений конкретного лица или мнений о конкретном лице; классификация (на негативную, нейтральную или позитивную) тональности мнения по отношению к цели; и выделение отношения между двумя сущностями в тексте. Исследования были проведены с использованием нейросетевых моделей типа BERT и Ruadapt-Qwen2.5. Была проведена серия экспериментов с использованием техники обучения моделей и подсказок (промптов).

1.3.4. Автоматическое распознавание эмоций

Близкой к задаче определения тональности является задача автоматического распознавания эмоций, что также можно использовать для оценки мнений авторов и упоминаемых акторов относительно достоинств и недостатков рассматриваемых в аналитических статьях процессов.

Неотъемлемой частью этой задачи является описание и систематизация эмоциональной лексики, которая может служить в качестве маркера, указывающего на то, что в тексте выражается та или иная эмоция. На сегодня уже существует достаточно

много англоязычных ресурсов эмоциональной лексики в открытом доступе, однако для других языков, в частности русского, размеченных данных значительно меньше.

Цель работы [23] - описать доступные для использования русскоязычные ресурсы эмоциональной лексики и представить созданный на их основе новый объединяющий эмоциональный лексикон - RusEmoLex (Russian Emotion Lexicon). Описаны следующие типы русскоязычных источников: лексикографические ресурсы (словари и тезаурусы); корпусные данные; данные, полученные в результате опросов (зафиксированные ассоциации слов с эмоциями); наборы данных, используемые в машинном обучении. Описана методика создания ресурса RusEmoLex на основе доступных русскоязычных источников. Методика включает следующие этапы: формирование исходных списков эмоциональной лексики на основе доступных источников; объединение списков; отбор лексики в RusEmoLex. Для включения в RusEmoLex лексическая единица должна удовлетворять следующим критериям: (1) количество источников, где зафиксирована лексическая единица (лексическая единица должна входить в два или более источников); (2) качество источников (как минимум один из источников, куда вошла лексическая единица, должен относиться к словарным или корпусным ресурсам и не являться переводным); (3) наличие метки класса (лексическая единица должна иметь установленный на основании большинства источников эмоциональный класс). Основным результатом исследования является эмоциональный лексикон RusEmoLex, включающий 1024 лексических единиц, размеченных по частеречной принадлежности; эмоциональному классу; категории источников, где встречается слово; количеству вхождений в источники. RusEmoLex находится в открытом доступе и может использоваться как для собственно лингвистических задач, так и в области обработки естественного языка.

Задача автоматического распознавания эмоций в разговорной речи в последнее время приобретает все больше популярности, в первую очередь благодаря появлению больших датасетов размеченных данных и развитию моделей машинного обучения. Наиболее распространенным подходом для решения данной задачи на текущий момент является применение речевых моделей с архитектурой трансформера, при этом классификация обычно производится на основе небольшого набора из 4–9 базовых эмоций. Однако даже проявление этих базовых эмоций у разных людей может в значительной степени отличаться в зависимости от языка говорящего и его культурного окружения.

В работе [15] исследуются возможности современных моделей машинного обучения, обученных на речевых эмоциональных данных разных языков, в частности то, какие лингвистические характеристики языков наибольшим образом влияют на их поведение. Для этого было проведено сравнение между собой качества автоматической классификации эмоций при применении речевых моделей, обученных на данных разных языков с дальнейшим их переносом на русский язык. В исследовании использовалась речевая модель HuBERT, отдельно обученная на данных польского, китайского и японского языков и затем протестированная на данных русского языка. В качестве обучающих данных для иностранных языков были соответственно взяты датасеты эмоциональной речи nEMO, ESD и JNVN. Тестирование обученных моделей производилось на данных русского датасета Dusha. Полученные результаты указывают на первоочередную зависимость речевой модели от просодических характеристик

данных, в то время как различия в синтаксисе языков не оказывают большого влияния на качество классификации.

В работе [16] было проведено исследование влияние различий в языках и обучающих данных на качество межязыкового переноса обученной модели на русский язык в задаче автоматического распознавания эмоций в речи. В качестве языковых источников, данные которых были использованы на этапе обучения, выступили английский, польский, китайский и японский языки, для которых соответственно использовались датасеты эмоциональной речи IEMOCAP, nEMO, ESD и JNVN. В качестве самой модели была использована речевая модель HuBERT на архитектуре трансформера. Все обученные на соответствующем датасете модели были протестированы на укороченной выборке из датасета русской эмоциональной речи Dusha. На основе полученных данных были описаны основные тенденции при выборе разных языков для обучения речевой модели и её дальнейшего переноса на русский язык, а также были проанализированы различия в датасетах, которые указывают на необходимость дальнейшей работы по сбору и разметки качественных данных эмоциональной речи.

1.4. Исследование методов организации научно-технических знаний

Решение проблемы организации научно-технических знаний осложняется несколькими факторами: (а) сложностью получения доступа к представительным коллекциям данных; (б) наличием разнообразных интерпретаций одних и тех же наблюдений у разных экспертов, причем мнения которых обычно выражены недостаточно формализовано; (в) как следствие, сложность оценки результатов вычислительного моделирования.

1.4.1. Общие вопросы организации научного знания

В рамках поддержки направления формирования представительных и достоверных коллекций научно-технических данных коллектив НИР участвует в деятельности Национальная платформа «Ковчег знаний МГУ» [14]. Трёхуровневая архитектура платформы включает: (1) создание обширного верифицированного русскоязычного корпуса объемом в миллиарды токенов; (2) разработку многоуровневой метаонтологии, предназначенной для описания миллионов сущностей и их семантических отношений; (3) внедрение слоя сервисов, включающего конвейер автоматической обработки текстов и поиска информации, а также образовательные лаборатории.

Одним из ключевых вопросов является разработка онтологий, как модели описания научно-технических знаний.

В работе [18] впервые исследуется потенциал больших языковых моделей для задачи автоматического пополнения таксономий на материале русского языка. Авторы адаптировали передовую методику TaxoLLaMA, которая ранее показала высокую эффективность для английского языка, используя для этого данные русскоязычного тезауруса RuWordNet. В рамках исследования было проведено сравнение производительности мультиязычных моделей и моделей, специально адаптированных для русского языка. Эксперименты подтвердили успешную применимость метода к русскоязычным данным и выявили значительное преимущество русскоязычных моделей. После дообучения модель YandexGPT-5-8B-Lite превзошла лучший предыдущий результат, достигнув значения MAP 55.42.

В работе [10] рассматривается проблема искажений, вносимой когипонимами (узлами таксономии с одинаковым общим гиперонимом-родителем) в обучающих наборах данных для задач извлечения гиперонимов. Хотя удаление тестовых элементов из обучающей выборки необходимо для предотвращения утечки данных, мы показали, что исключение когипонимов не менее важно. При тонкой настройке модели на наборе данных, состоящем из пар <гипоним-гипероним>, извлечённых из таксономического ресурса WordNet, простого исключения тестовых узлов недостаточно для адекватной оценки качества модели на тестовых данных. Когипонимы ведут себя как неявные подсказки для определения гиперонимов, искусственно завышая показатели качества модели и искажая её эффективность в реальных сценариях разработки таксономий. Мы обучили модель LLaMA-2 с использованием процедуры TaxoLLaMA, предложенной в Moskvoretskii et al. (2024), на обширном корпусе пар гипоним-гипероним, извлечённых из WordNet, с их определениями и без. Оценка на бенчмарке SemEval-2018 показала, что включение когипонимов в тренировочные данные искусственно завышает показатели качества.

1.4.2. Формирование данных для организации научных соревнований

В настоящее время методология сравнительного тестирования разных методов для решения задач обработки и анализа данных в согласованной постановке доказала свою эффективность. Обмен информацией о достоинствах и недостатках различных подходов и методов значительно ускоряет общее продвижение фронта исследований.

Организация научных соревнований требует обычно объединения экспертизы группы исследователей, и включает в себя: (а) согласование нюансов постановки задач, подготовки данных, метрик оценки и интерпретации результатов; (б) формирования наборов данных; (в) проведение тестирования; (г) интерпретация результатов.

Участники коллектива авторов НИР принимают участие в деятельности международного коллектива по формированию наборов данных (датасетов) BioASQ для поддержки решения задач семантического индексирования биомедицинских данных и автоматического формирования ответов на вопросы по таким данным.

В 2025 году [6, 7] BioASQ включал в себя новые версии двух уже существующих задач, b и Synergy, а также четыре новые задачи: а) задача MultiClinSum по многоязычному клиническому суммированию; б) задача BioNNE-L по связыванию вложенных именованных сущностей на русском и английском языках; с) задача ELCardioCC по клиническому кодированию в кардиологии; d) задача GutBrainIE по извлечению информации о взаимодействии кишечника и мозга.

В этом научном соревновании на данных BioASQ приняли участие 83 команды, представившие в общей сложности более 1000 различных работ по шести различным общим задачам конкурса. Как и в предыдущих выпусках, несколько участвующих систем показали конкурентоспособные результаты, что свидетельствует о постоянном развитии передовых технологий в этой области.

Также коллектив авторов НИР принял участие в организации формирования нового бенчмарка ruSciFact [11] для проверки фактов в научных утверждениях на русском языке.

ruSciFact структурирован как задача NLI (Non-Language Intensive Line), цель которой — проверить, подтверждается ли факт заданным аннотацией. Для генерации фак-

тов использовался трехэтапный конвейер на основе LLaMA-405B, проверяя полученные предложения с помощью экспертов-терминологов. Набор данных ruSciFact состоит из 1128 пар в формате <аннотация, утверждение>, которые выпущены в качестве открытого исходного кода вместе с кодом бенчмарка. Кроме того, открыт исходный код конвейера генерации фактов, что облегчает расширение набора данных на конкретные научные области.

1.4.3. Задачи анализа научно-технических текстов

С учетом сложности анализа научно-технических текстов в рамках НИР продолжаются работы по разработке методов улучшения качества решения задач обработки такого рода документов.

В работе [24] предложен многоэтапный подход к автоматическому порождению ключевых слов для русскоязычных научных статей. Метод основан на дообучении трансформеров с использованием псевдоразметки и контрастивного обучения, а также включает фильтрацию порождённых кандидатов. Реализованы две стратегии генерации псевдоразметки и архитектура с биэнкодером для отбора релевантных ключевых слов. Эксперименты на корпусе математики и компьютерных наук демонстрируют превосходство предложенного подхода над классическими и нейросетевыми методами по метрикам F1, ROUGE-1 и BERTScore.

В работе [9] описываются результаты участия в конкурсе RuTermEval, посвященном извлечению вложенных терминов. Для извлечения вложенных терминов мы применяем модель Binder, которая ранее успешно использовалась для распознавания вложенных именованных сущностей. Получены лучшие результаты распознавания терминов во всех трех направлениях конкурса RuTermEval. Кроме того, исследуется новая задача распознавания вложенных терминов на основе плоских обучающих данных, аннотированных терминами без вложенности. Можно заключить, что несколько предложенных в работе [9] подходов достаточно эффективны для извлечения вложенных терминов без их вложенной маркировки.

1.4.4. Базовые проблемы анализа текстов в специальных предметных областях

Одной из проблем как автоматического анализа текстов в общем домене, так, особенно, текстов в специальной предметной области, является решение задачи разрешения многозначности, что поддерживает интерес к улучшению качества решения данной задачи.

Различные статистические и нейронные подходы к разрешению неоднозначности значений слов (WSD) в русском языке были многократно рассмотрены, однако мало внимания уделялось использованию LLM для этой задачи. Способность генеративных LLM делать содержательные выводы делает их многообещающим инструментом для исследований. Цель исследования [1] — проверить способность генеративных моделей, особенно адаптированных для русского языка, различать значения неоднозначных русских слов. Используемая экспериментальная установка включает в себя побуждение модели выбрать правильное значение неоднозначного слова из предопределенного списка. Проведены эксперименты на ряде LLM на основе русскоязычных декодеров, от самых современных моделей до более мелких. Цель этого исследования двояка. Во-первых, установить базовый уровень производительности генеративных LLM в разрешении неоднозначности значений слов в русском языке. Во-вторых, про-

вести лингвистический анализ, изучив, как различные модели обрабатывают различные типы лексической неоднозначности, а именно омонимию и полисемию, а также их способность обрабатывать разреженные данные. В результате описан диапазон производительности различных моделей, при этом показатели F1-меры варьировались от 0,45 до более 0,9. Выявлены лингвистические факторы, влияющие на сложность разрешения неоднозначности, а именно семантическую связь и представление значений. Наибольший средний показатель F1, равный 0,75, достигается для омонимов с хорошо представленными значениями. Эти результаты указывают на ключевые области для дальнейшего улучшения обработки лексической неоднозначности в русском языке генеративными моделями с лингвистическим знанием.

В работе [2] представлен RuSemCor, открытый корпус для разрешения многозначности слов (WSD) для русского языка. Корпус был создан путем ручного сопоставления токенов из корпуса OpenCorpora со значениями слов в русской сети WordNet RuWordNet. Он состоит из 869 документов с 121 710 токенами, из которых 51 588 аннотированы в WordNet. Ресурс представлен с использованием онтологий NIF, OLiA, OntoLex и Global WordNet и интегрирован в облако лингвистических связанных открытых данных (Linguistic Linked Open Data). Авторы использовали RuSemCor в качестве диагностического эталона для оценки ряда методов разрешения многозначности слов. Проведенные эксперименты дали три основных результата: 1) Генеративные LLM-модели значительно превосходят традиционные методы, основанные на знаниях, такие как персонализированный PageRank. 2) Несмотря на свои преимущества, генеративные LLM-модели не превосходят модели на основе кодировщиков, специально обученные для разрешения многозначности слов. 3) Включение лексико-семантических отношений из RuWordNet дает неоднозначные результаты: оно повышает производительность моделей на основе кодировщиков и ведущих моделей линейного программирования, таких как GPT-4, DeepSeek и Mistral 24B, но, как правило, снижает точность для более мелких генеративных моделей, таких как GPT-3 и Mistral 7B. Ресурс распространяется под открытой лицензией CC BY-SA и доступен по адресу: <https://github.com/LLOD-Ru/rusemcor>.

1.4.5. Анализ мультимодальных данных

Как уже указывалось, одна из проблем улучшения результатов анализа научно-технических данных – сложности с формированием представительных коллекций.

Одним из перспективных подходов решения этой проблемы является привлечение мультимодальных данных, прежде всего аудиоданных подкастов, звуковых дорожек для видео.

Современные нейронные позволили значительно улучшить качество распознавания видео- и аудио- данных. Однако, сохраняется еще большое количество нерешенных вопросов, методы решения которых исследуются в рамках НИР.

Одной из проблем является распознавание незнакомых слов. Одним из популярных методов улучшения качества распознавания слов, незнакомых модели (не входящих в словарный запас), является метод расширения словарного запаса. Слова, не входящие в словарный запас, особенно распространены в предметных областях. Это термины и словосочетания, подобные терминам. Многие из них в русской речи были недавно заимствованы из английского языка и пока не имеют общепринятого написа-

ния в русском языке. Это затрудняет добавление таких слов в словарь модели для языковых моделей, снижая качество их звучания в предметных областях.

В статье [12] представлен метод повышения качества распознавания речи, содержащий такие термины, основанный на алгоритме автоматического построения так называемых “русских транскрипций” для произвольных английских слов. Русская транскрипция означает слово со схожим звучанием, написанное русскими буквами, которое могло бы заменить оригинальный термин при добавлении в словарь модели и, таким образом, было бы правильно распознано. Результаты проведенных нами экспериментов позволяют предположить, что описанный здесь алгоритм может быть использован для улучшения качества систем распознавания речи.

В работе [8] предлагается подход, основанный на данных, для прогнозирования частоты ошибок распознавания слов (WER) без необходимости использования эталонных транскрипций. Предложенный метод включает создание разнообразных аудиоданных путем применения различных типов шума, акустических искажений и импульсных характеристик помещения к чистым образцам речи на многих уровнях качества и разборчивости. В отличие от предыдущих работ, извлекается и анализируется полный набор характеристик качества речи, включая оценки отношения сигнал/шум (SNR), современные нейронные метрики качества звука (такие как NISQA) и оценки достоверности моделей автоматического распознавания речи (ASR) для обучения моделей прогнозирования WER. Проведены эксперименты на нескольких языках с использованием современных архитектур ASR (Whisper и FastConformer), чтобы продемонстрировать эффективность метода в прогнозировании WER в различных акустических условиях. Показано, что наш подход обобщается в многоязычной среде с унифицированной моделью. Проведен анализ важности признаков для определения ключевых метрик, необходимых для прогнозирования WER. Данная работа позволяет создавать практические приложения, такие как фильтрация аудиовходов на основе качества, что дает системам автоматического распознавания речи возможность оценивать ожидаемую производительность и определять надежность транскрипции без использования эталонных текстовых данных.

СПИСОК ЛИТЕРАТУРЫ

1. Gousyatskaya Polina, Loukachevitch Natalia, Word Sense Disambiguation in Russian: A Generative LLM Approach //Communications in Computer and Information Science, место издания Springer Nature Switzerland Cham, том 2671, с. 145-156. – 2025. – DOI 10.1007/978-3-032-04958-2_11 (Scopus)
2. Kirillovich Alik, Karpov Ilya, Ilvovsky Dmitry, Kulaev Maxim, Loukachevitch Natalia, RuSemCor: a Word Sense Disambiguation corpus for Russian //ACM International Conference on Information and Knowledge Management. – 2025. (Scopus)
3. Kovalev G., Tikhomirov M., Kozhevnikov E., Kornilov M., Loukachevitch N., Building Russian Benchmark for Evaluation of Information Retrieval Models //In Proceedings of the International Conference "Dialogue-2025". – 2025. (Scopus)
4. Lee G., Loukachevitch N., Khokhlov A., Generating Encyclopedic Articles Based on a Collection of Scientific Publications //Proceedings of the International Conference "Dialogue-2025". – 2025. (Scopus)

5. Loukachevitch N., Tkachenko N., Lapanitsyna A., Tikhomirov M., Rusnachenko N., RuOpinionNE-2024: Extraction of Opinion Tuples from Russian News Texts //Proceedings of the International Conference "Dialogue-2025". – 2025. (Scopus)
6. Nentidis A., Katsimpras G., Krithara A., Krallinger M., Ortega M.R., Loukachevitch N., BioASQ at CLEF2025: The Thirteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge //Proceedings of European Conference on Information Retrieval 2025, Springer (New York), c. 407-415. – 2025. (Scopus)
7. Nentidis Anastasios, Katsimpras Georgios, Krithara Anastasia, Krallinger Martin, Rodriguez-Ortega Miguel, Rodriguez-Lopez Eduard, Loukachevitch Natalia, Sakhovskiy Andrey, Tutubalina Elena, Dimitriadis Dimitris, Tsoumakas Grigorios, Giannakoulas George, Bekiaridou Alexandra, Samaras Athanasios, Nunzio Giorgio Maria Di, Ferro Nicola, Marchesin Stefano, Martinelli Marco, Silvello Gianmaria, Paliouras Georgios, Overview of BioASQ 2025: The Thirteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering //Lecture Notes in Computer Science, Springer Nature (Switzerland), c. 173-198. – 2025. – DOI: 10.1007/978-3-032-04354-2_12 (Scopus)
8. Polevoi Anton, Kragin Alexander, Loukachevitch Natalia, Ground Truth-Free WER Prediction for ASR via Audio Quality and Model Confidence Features //International Conference on Speech and Computer, Springer Nature (Switzerland), c. 10-1007. – 2025. – DOI: 10.1007/978-3-032-07959-6_3 (Scopus)
9. Rozhkov I., Loukachevitch N., Methods for Recognizing Nested Terms //Proceedings of the International Conference "Dialogue-2025". – 2025. (Scopus)
10. Sadkovskii F., Loukachevitch N., Grishin I. Fine-tuning large language models for hypernym discovery task: Sister terms do their part // Записки научных семинаров Санкт-Петербургского отделения математического института им. В.А. Стеклова РАН. — 2025. — Vol. 546. — P. 146
11. Vatolin A., Gerasimenko N., Loukachevitch N., Ianina A., Vorontsov K., ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian //In Proceedings of the International Conference "Dialogue-2025". – 2025. (Scopus)
12. Volkov E.V., Dobrov B.V., Automated Transcription of Unfamiliar Words to Improve Recognition of Terms in the Subject Domain //Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Pleiades Publishing, Ltd (Road Town, United Kingdom), том 35, № 3, с. 245-254. – 2025. – DOI: 10.1134/S1054661825700154 (Scopus, Ядро РИНЦ, RSCI)
13. Григорьев Д.А., Чернышев Д.И., RUWIKIBENCH: оценка больших языковых моделей посредством воспроизведения энциклопедических статей //Доклады Российской академии наук. Математика, информатика, процессы управления, Российская академия наук (Москва), том 527, № 5, с. 171-181. – 2025. – DOI: 10.7868/S2686954325070148 (Ядро РИНЦ, переводная версия Web of Science, Scopus)
14. Зубарева Е.В., Лукашевич Н.В., Пичугов А.А., Семенов А.Л., Национальная платформа «Ковчег знаний МГУ»: онтологический компас цифрового суверенитета и трансформации образования //Информатизация образования и методика электронного обучения : цифровые технологии в образовании: материалы IX Междунар. науч. конф. Красноярск, 23–26 сентября 2025 г.: в 4 ч. Ч. 1, место издания Красноярский государ-

ственный педагогический университет им. В.П. Астафьева г.Красноярск, том 1, с. 215-219. – 2025. (РИНЦ)

15. Лемаев В.И., Лукашевич Н.В., Выражение эмоций в разных языках: возможности межъязыкового переноса моделей распознавания //Вестник Московского университета. Серия 9: Филология, Изд-во Моск. ун-та (М.), № 4, с. 9-20. – 2025. (Ядро РИНЦ, RSCI)

16. Лемаев В.И., Лукашевич Н.В., Межъязыковой перенос моделей автоматического распознавания эмоций в речи на русский язык: данные и тенденции //Научно-техническая информация. Серия 2: Информационные процессы и системы, ВИНТИ (М.), № 5, с. 28-38. – 2025. (Ядро РИНЦ, RSCI)

17. Рогов А.А., Лукашевич Н.В., Сравнение подходов к интерпретации языковых моделей: анализ метода на основе маскированного языкового моделирования и традиционных методов //Труды XII национальной конференции по искусственному интеллекту с международным участием, КИИ-2025, с. 287-297. – 2025. (РИНЦ)

18. Садковский Ф.А., Лукашевич Н.В., Гришин И.Ю., Автоматическое пополнение таксономий на русском языке с помощью больших языковых моделей //Труды XII национальной конференции по искусственному интеллекту с международным участием, том 3, с. 152-163. – 2025. (РИНЦ)

19. Студеникина К.А., Лютикова Е.А., Герасимова А.А., Оценка лингвистической компетенции больших языковых моделей на материале корпуса согласовательной вариативности //Компьютерная лингвистика и вычислительные онтологии, издательство Университет ИТМО (Санкт-Петербург). – 2025. (РИНЦ)

20. Тихомиров М.М., Чернышев Д.И., Ruadapt: вычислительно эффективная языковая адаптация больших языковых моделей //Доклады Российской академии наук. Математика, информатика, процессы управления, Российская академия наук (Москва), том 527, с. 290-300. – 2025. – DOI: 10.7868/S2686954325070252 (Ядро РИНЦ, переводная версия Web of Science, Scopus)

21. Уразов С.О., Лукашевич Н.В., Задачи анализа тональности в новостных текстах //Труды XII национальной конференции по искусственному интеллекту с международным участием КИИ-2025, с. 319-329. – 2025. (РИНЦ)

22. Шамаева Е.Д., Лукашевич Н.В., Применимость методов оценки качества синтаксического анализа к результатам синтаксического анализатора на основе большой языковой модели //Труды XII национальной конференции по искусственному интеллекту с международным участием КИИ-2025, том 1, с. 1-330. – 2025. (РИНЦ)

23. Ярошенко Полина Владимировна, Лукашевич Наталья Валентиновна, RusEmoLex: эмоциональный лексикон для русского языка //Russian Journal of Linguistics, Российский университет дружбы народов (РУДН) (Москва), том 29, № 3, с. 659-687. – 2025. – DOI: 10.22363/2687-0088-44439 (Scopus Q1)

24. Яушев К.Ш., Лукашевич Н.В., Автоматическая генерация ключевых слов для русскоязычных научных статей с использованием псевдоразметки и контрастивного обучения //Труды XII национальной конференции по искусственному интеллекту с международным участием, том 1, с. 341-352. – 2025. (РИНЦ)

ЗАКЛЮЧЕНИЕ

В рамках Этапа 6 НИР 2025 г. «Исследование методов применения больших нейросетевых языковых моделей для задач автоматизированного формирования графов знаний» получены следующие основные результаты.

1) По направлению «Исследование методов улучшения качества результатов информационного поиска для решения задач RAG (подключения новых знаний) для больших языковых моделей»:

- Создан RusBEIR, комплексный бенчмарк, разработанный для оценки моделей информационного поиска на русском языке без предварительного обучения. Он включает 17 наборов данных из различных областей, интегрируя адаптированные, переведенные и вновь созданные наборы данных, что позволяет систематически сравнивать лексические и нейронные модели. Полученные результаты показывают важность предварительной обработки для лексических моделей в морфологически богатых языках и подтверждают, что классические методы типа BM25 является надежным базовым набором данных для поиска полных документов и во многих ситуациях успешно конкурируют с моделями векторного поиска. Нейронные модели, такие как mE5-large и BGE-M3, демонстрируют превосходную производительность на большинстве наборов данных, но сталкиваются с проблемами при поиске длинных документов из-за ограничений по размеру входных данных.

- Описаны методы адаптации мультязычных LLM для обработки русского языка. Представлена комплексная и вычислительно эффективная методология Ruadapt для языковой адаптации LLM с заменой токенизации. Приведено подробное эмпирическое исследование каждого шага адаптации с целью определения оптимальных гиперпараметров, а также ключевых этапов и их влияния на итоговое качество. Модели, код и наборы данных опубликованы в открытом доступе, предлагая научному сообществу проверенную и экономически целесообразную стратегию создания высококачественных языковых моделей.

2) По направлению «Исследование методов автоматического структурирования материалов при формировании аналитических статей»:

- Разработаны методы автоматического формирования энциклопедических статей методами абстрактного реферирования с использованием больших нейросетевых языковых моделей на материалах ресурсов Wikipedia и РуВики;

- Проведены исследования для лучшего понимания возможностей LLM в различных задачах базовых методов обработки текстов: синтаксических экспериментов по изучению вариативного согласования на материале созданного участниками НИР Корпуса Вариативного Согласования (КВас), оценки качества синтаксического анализа к результатам работы синтаксического анализатора на основе больших языковых моделей, исследованы подходы к объяснению поведения BERT в задачах текстовой классификации.

- Исследованы методы определения тональности и выражения эмоций, что является важным для извлечения экспертных оценок о перспективности, достоинствах и недостатках рассматриваемых процессов для формирования информационно-аналитических материалов. В частности, описаны доступные для использования русскоязычные ресурсы эмоциональной лексики и представить созданный на их основе

новый объединяющий эмоциональный лексикон - RusEmoLex (Russian Emotion Lexicon). Описана методика создания ресурса RusEmoLex на основе доступных русскоязычных источников.

3) По направлению «Исследование методов организации научно-технических знаний»:

- В рамках поддержки направления формирования представительных и достоверных коллекций научно-технических данных коллектив НИР участвует в деятельности Национальная платформа «Ковчег знаний МГУ». Трёхуровневая архитектура платформы включает: (1) создание обширного верифицированного русскоязычного корпуса объемом в миллиарды токенов; (2) разработку многоуровневой метаонтологии, предназначенной для описания миллионов сущностей и их семантических отношений; (3) внедрение слоя сервисов, включающего конвейер автоматической обработки текстов и поиска информации, а также образовательные лаборатории.

- Исследован потенциал больших языковых моделей для задачи автоматического пополнения таксономий на материале русского языка. Адаптирована методика TaxoLLaMA, которая ранее показала высокую эффективность для английского языка, используя для этого данные русскоязычного тезауруса RuWordNet. Эксперименты подтвердили успешную применимость метода к русскоязычным данным и выявили значительное преимущество русскоязычных моделей.

- Участники НИР активно участвуют в кооперации с другими научными коллективами по организации научных соревнований по решению задач обработки и анализа научно-технических текстов. Проведение такого рода научных конкурсов является эффективной формой поиска новых методов решения исследовательских задач и быстрого распространения лучших практик. Участники коллектива авторов НИР принимают участие в деятельности международного коллектива по формированию наборов данных (датасетов) BioASQ для поддержки решения задач семантического индексирования биомедицинских данных и автоматического формирования ответов на вопросы по таким данным. Также коллектив авторов НИР принял участие в организации формирования нового бенчмарк ruSciFact для проверки фактов в научных утверждениях на русском языке.

- С учетом сложности анализа научно-технических текстов в рамках НИР продолжаются работы по разработке методов улучшения качества решения задач обработки такого рода документов.

Предложен многоэтапный подход к автоматическому порождению ключевых слов для русскоязычных научных статей. Метод основан на дообучении трансформеров с использованием псевдоразметки и контрастивного обучения, а также включает фильтрацию порождённых кандидатов. Реализованы две стратегии генерации псевдоразметки и архитектура с биэнкодером для отбора релевантных ключевых слов. Эксперименты на корпусе математики и компьютерных наук демонстрируют превосходство предложенного подхода над классическими и нейросетевыми методами по метрикам F1, ROUGE-1 и BERTScore. Описан опыт участия в конкурсе RuTermEval, посвященном извлечению вложенных терминов. Для извлечения вложенных терминов применялась модель Binder. Получены лучшие результаты распознавания терминов во всех трех направлениях конкурса RuTermEval. Также предложены и исследованы новые методы разрешения многозначности лексики.

- Для решения проблемы сложности с формированием представительных коллекций научно-технических текстов, рассматривались методы распознавания речи, как источника информации.

Исследовалась задача распознавания незнакомых слов (не входящих в словарный запас), что характерно для специальных предметных областей, особенно научно-технических. Представлен метод повышения качества распознавания речи, содержащей такие термины, основанный на алгоритме автоматического построения так называемых “русских транскрипций” для произвольных английских слов. Также предложен подход для прогнозирования частоты ошибок распознавания слов без необходимости использования эталонных транскрипций. Предложенный метод включает создание разнообразных аудиоданных путем применения различных типов шума, акустических искажений и импульсных характеристик помещения к чистым образцам речи на многих уровнях качества и разборчивости. В отличие от предыдущих работ, извлекается и анализируется полный набор характеристик качества речи, включая оценки отношения сигнал/шум (SNR), современные нейронные метрики качества звука (такие как NISQA) и оценки достоверности моделей автоматического распознавания речи (ASR) для обучения моделей прогнозирования WER. Данная работа позволяет создавать практические приложения, такие как фильтрация аудиовходов на основе качества, что дает системам автоматического распознавания речи возможность оценивать ожидаемую производительность и определять надежность транскрипции без использования эталонных текстовых данных.

По результатам Этапа 6 НИР опубликовано 24 научных работ. Из них: Q1 1, WoS 2, Scopus 14, RSCI 4, Ядро РИНЦ 6, РИНЦ 7.

ПРИЛОЖЕНИЕ Б. РЕЗУЛЬТАТЫ, ПОЛУЧЕННЫЕ ПО ТЕМЕ ЗА 2020-2024 ГГ.

В настоящем разделе приведены результаты, полученные по теме за период 2020-2024 гг.

- Этап 2020 г. «Разработка методов автоматического пополнения больших лингвистических онтологий таксономическими отношениями, методов извлечения редких типов именованных сущностей»;

- Этап 2021 г. «Разработка методов автоматизированного формирования больших лингвистических предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний»;

- Этап 2022 года «Разработка методов автоматизированного формирования больших графов знаний предметной области»;

- Этап 2023 года «Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области»;

- Этап 2024 года «Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий».

Этап 2020 г. «Разработка методов автоматического пополнения больших лингвистических онтологий таксономическими отношениями, методов извлечения редких типов именованных сущностей»

В 2020 году были получены следующие результаты.

1) В сотрудничестве с коллегами из Сколтеха было организовано и проведено научное соревнование по методам автоматического пополнения таксономических отношений больших лингвистических онтологий. Основой соревнования являлось сформированное обучающий и тестовый наборы данных. Участники тестирования должны были дополнить существующую таксономию RuWordNet новыми словами: для каждого нового слова их системы должны предоставлять ранжированный список возможных гиперонимов, т.е. ближайших родовых слов. По сравнению с предыдущими заданиями для других языков, данное тестирование имеет более реалистичную постановку задания: новые слова предоставлены без толкований. Вместо этого был предоставлен текстовый корпус, в котором встречаются эти новые слова. Для проведения тестирования был создан новый набор данных на основе неопубликованных данных тезауруса RuWordNet. Задача тестирования состоит из двух подзадач: «существительные» и «глаголы». В задании приняли участие 16 исследовательских групп, показавших высокие результаты, более половины из них превзошли базовый подход, рассчитанный организаторами тестирования.

2) Проведено исследование моделей и методов пополнения больших лингвистических онтологий с использованием методов машинного обучения. Исследованы подходы для извлечения отношений гипоним-гипероним (класс-подкласс), которые являются основной большинства онтологий и графов знаний. Существенной является

задача автоматического пополнения онтологий на основе больших текстовых корпусов. В рамках тестирования RUSSE-2020 был реализован метод для пополнения существующей таксономии в тезаурусе RuWordNet. Метод включал использование следующих признаков для пополнения таксономии:

- Дистрибутивные векторные модели (word2vec, PMI+SVD), -- Специальные типы шаблонов,
- Использование структуры существующего тезауруса,
- Нейросетевая архитектура transformer в виде модели BERT для решения задачи классификации.

Результатом алгоритма является ранжированный список из 10 кандидатов гиперонимов. Оценка качества проводилась на основе метрик MAP и MRR. В результате описанный подход получил 4 место в соревновании по предсказанию гиперонимов среди существительных. Особенностью подхода является то, что среди первых 5 решений участников, только в данном решении не использовались сторонние словари и внешние векторные представления, обученные на других, более крупных, наборах данных. Это важно по той причине, что приближает к реальной ситуации, когда необходимо расширить существующий тезаурус на новый набор данных. Представленный подход является новым и уникальным для задачи предсказания гиперонимии для расширения тезауруса. То, что данный подход получил высокие результаты, не используя внешние словари и векторные представления по другим наборам данных, также является преимуществом данного подхода.

3) Разработана и опубликована обновленная версия лингвистической онтологии RuWordNet. В рамках сотрудничества с сообществом Global WordNet Association планируется публикация версии RuWordNet в формате Open Multilingual WordNet. Объем новой версии составляет более 135 тысяч слов и выражений.

4) Исследованы возможности больших предобученных моделей для решения актуальных задач при построении информационно-аналитических систем – извлечение именованных сущностей «редких» типов. Исследованы подходы к улучшению качества извлечения именованных сущностей в конкретной предметной области за счет автоматической доразметки текстовой коллекции и обучения специализированной версии юольшой языковой модели BERT для заданной предметной области. Для экспериментов был использован корпус новостных статей и комментариев в области компьютерной безопасности Sec_col. Для этого модель RuBERT была дообучена на текстовой коллекции новостей и комментариев в области компьютерной безопасности (RuCyBERT). Замена исходного RuBERT на дообученный RuCyBERT приводит к значительному росту качества извлечения именованных сущностей. Кроме того, были исследованы возможности пополнения обучающей коллекции за счет использования списка дескрипторов (слов, стоящих перед именем, например: вирус РЕТУА), соответствующих каждому типу именованных сущностей. Основная идея метода состоит в том, что неразмеченные предложения автоматически модифицируются, путем добавления именованных сущностей рядом или вместо дескриптора. Таким образом можно генерировать большое количество предложений с псевдо разметкой. Подобное можно сделать и уже с размеченными данными, добавляя в них новые сущности. В экспериментах было показано, что использование модели BERT, настроенной на коллекции текстов заданной предметной области и предварительно обученной на сочета-

нии общего набора данных и дополнительно порожденных данных, обеспечивает наилучшие результаты распознавания именованных сущностей. Была также изучена вычислительная производительность модели BERT в так называемом режиме смешанной точности. Был обучен новый вариант модели BERT для русского языка: RuNewsBERT. Обучение было выполнено следующим образом: (а) Инициализация весов от RuBERT, (б) Текстовая коллекция: 8 миллионов новостей, собранных с различных русскоязычных источников, (в) Обучение проводилось на системе DGX-2 на 16 видеокартах V100, (г) Обучение происходило только на задаче MLM, в каждом документе обрабатывались первые 512 токенов, (д) Для обучения потребовались 4 миллиона итераций, что заняло примерно один месяц.

5) Проведены исследования методов определения тональности с использованием нейросетевых методов с механизмом «внимания». Создана и опубликована новая версия словаря оценочной лексики RuSentiFrames. Тексты могут передавать несколько типов взаимосвязанной информации, касающейся мнений и отношений. Такая информация включает отношение автора к упомянутым сущностям, отношение сущностей друг к другу, положительное и отрицательное влияние на сущности в описанных ситуациях. В лексиконт RuSentiFrames для русского языка предикатные слова и выражения собраны и связаны с так называемыми оценочными фреймами, передающими несколько типов предполагаемой информации об установках и эффектах. Мы применили созданные фреймы для извлечения оценочных отношений между именованными сущностями из большой коллекции новостей. Исследованы возможности недавно появившейся архитектуры BERT по сравнению с традиционными подходами на основе нейронных сетей (CNN, LSTM, BiLSTM) на существующих размеченных наборах данных для анализа тональности на русском языке. Сравнивались два варианта архитектуры BERT, дообученной на русском языке: (а) обученный на новостях и Википедии и (б) обученный на комментариях, постах в социальных сетях (разговорный вариант). Было показано, что для всех рассмотренных задач тональности в этом исследовании разговорный вариант русского BERT работает лучше. Наилучшие результаты были достигнуты с помощью модели BERT-NLI, которая рассматривает задачи классификации тональности как задачу логического вывода на естественном языке. По одному из наборов данных эта модель практически достигает человеческого уровня. Рассмотрена задача извлечения оценочных отношений между именованными сущностями, упомянутыми в тексте. Предлагается подход на основе нейросетевых кодировщиков контекста, основанных на внимании. Для этой задачи были адаптированы кодировщики контекста двух типов: (а) функционально-ориентированные; (б) основанные на самовнимании. В исследовании использовался корпус русскоязычных аналитических текстов RuSentRel и автоматически построенный новостной датасет RuAttitude для обогащения обучающей выборки. Задача выделения отношения рассматривалась как двухклассовая (положительный, отрицательный) и трехклассовая (положительный, отрицательный, нейтральный) для всего документа. Эксперименты с корпусом RuSentRel показали, что трехклассовые модели классификации, которые используют корпус RuAttitude для обучения, приводят к увеличению на 10% и дополнительным 3% на F1, когда архитектуры моделей включают механизм внимания. Также были проанализированы распределения весов внимания в зависимости от типа контекста.

Этап 2021 г. «Разработка методов автоматизированного формирования больших лингвистических предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний»

Результаты этапа 2021 года.

1) По направлению выявления сложных текстовых образов именованных сущностей – вложенных, разрывных, неполных - сформирован NEREL - новый датасет на русском языке с размеченными именованными сущностями и отношениями между ними. Особенностью NEREL является разметка вложенных именованных сущностей и их отношений. Отношения между сущностями размечаются в рамках связного текста и не ограничиваются уровнем предложения.

2) По направлению разработки методов автоматического пополнения больших лингвистических онтологий (с небольшим количеством фиксированных отношений) предметной области - получен результат, что комбинации векторных представлений, обученных на общей предметной области, рассчитанные на больших текстовых коллекциях из сети Интернет, оказывают существенное влияние на качество пополнения таксономий, таких как WordNet, RuWordNet, Онтологии Естественных Наук и Технологий (ОЕНТ).

3) По направлению разработки методов глубокого машинного обучения для интегрирования большой номенклатуры типов именованных сущностей с понятиями онтологии - реализована система предсказания гиперонимов для неизвестных заранее именованных сущностей и веб-сервис для работы с ней.

4) Велись исследования методов наполнения «текстовых вершин» графа знаний, когда элемент графа знаний представляет собой фрагмент текста, содержащий неструктурированное знание по заданной теме. Исследовались методы абстрактного аннотирования извлечения значимых текстовых фрагментов с использованием современных нейросетевых подходов.

5) Были рассмотрены методы анализа текстовых материалов вида «резюме и вакансии», учебные курсы. Для онтологии ОЕНТ получен результат, что отношения «пререквизит» могут автоматически выводиться по иерархии существующих отношений лингвистических онтологий типа РуТез, возможно, с добавлением небольшого количества отношений вручную.

6) Проводились исследования по интеграции информационных методов в биологические исследования. Практическая значимость полученных результатов заключается в снижении трудоемкости для формирования больших онтологических ресурсов, создании новых инструментов для информационно-аналитических систем, в том числе для новых предметных областей.

Этап 2022 года «Разработка методов автоматизированного формирования больших графов знаний предметной области»

В течение 2022 года получены следующие результаты.

1) По направлению исследования методов автоматического извлечения неизвестных отношений показано, что применение выделяемых с использованием нейросетевых методов именованных сущностей и отношений с ними позволяет ввести метрики фактологической достоверности оценки качества экстрактивных и абстрактных аннотаций. Разработан новый метод построения псевдо-аннотаций на основе кластеров – ClusterVote. Метод апробирован для обучения русскоязычных предобученных генеративных моделей общего назначения: mBART, ruT5. С помощью метода собрана самая большая коллекция для аннотирования русскоязычных новостей – Telegram News*CV(RU).

2) По направлению разработки методов автоматизированного формирования больших онтологий предметной области с развитым набором отношений были проведены эксперименты по извлечению отношений (49 типов) на датасете NEREL. Особенностью датасета является то, что он размечен вложенными именованными сущностями, что позволяет увеличивать полноту извлечения отношений из текстов. Была проведена коррекция входного формата данных, после чего качество извлечения отношений внутри предложения с помощью пакета OpenNRE с использованием контекстуализированных эмбедингов RuBERT, достигло 80.5% F-меры. Для исследования извлечения таксономических отношений из текстовых коллекций в рамках проекта был создан датасет Diachronic wordnets. Был исследован подход на основе мета-эмбедингов с функцией потерь триплет-лосс, комбинирующий векторные представления слов (word2vec, glove, fasttext) и графовые представления, с помощью которого получены лучшие результаты извлечения гиперонимов для существительных во всех вариантах датасета.

3) По направлению разработки методов связывания различных текстовых вариантов извлеченных именованных сущностей на основе результатов обработки больших текстовых коллекций были проведены эксперименты по связыванию упоминаний именованных сущностей из набора данных NEREL с объектами графа знаний Викиданные. Показано, что наиболее эффективным из рассмотренных способов оценки неопределенности является score-based подход. Для ряда категорий рассматриваемого набора данных, более высокую эффективность показывают методы, основанные на ансамблях моделей. 4) По направлению разработки методов разрешения многозначности текстового выражения именованных сущностей в разных документах был изучен подход, учитывающий априорную многозначность именованных сущностей при связывании сущностей с Викиданными. В результате комбинирования score-based оценки с предложенным методом удалось увеличить точность предсказания правильной ссылки сущности в Викиданных. Практическая значимость результатов заключается в снижении трудоемкости для формирования больших графов знаний в части подключения именованных сущностей, а также текстовых объектов в виде аннотаций.

Этап 2023 года «Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области»

Результаты этапа 2023 года.

1) Исследованы новые методы установления таксономических отношений между существующими концептами лингвистических онтологий и терминоподобными сущностями;

2) Сформированы, опубликованы и исследованы, в том числе в результате участия в организации научных соревнований, новые наборы данных для выявления сложных отношений в специальных предметных областях;

3) Исследованы задачи выявления отношений в сложных случаях вложенных именованных сущностей;

4) Рассматривались задачи абстрактного аннотирования: для задачи абстрактного аннотирования новостных кластеров предложен новый метод создания коллекций для обучения нейросетевых методов аннотирования, предназначенный моделировать особенности задачи путем учета информации в связанных документах;

5) Исследованы предварительно обученных моделей абстрактного реферирования в условиях ограниченных ресурсов;

6) Выполнены исследования по воспроизведению лингвистических характеристик текстов при применении современных методов автоматической обработки.

По теме опубликовано 13 статей (1 Q1, 5 Wos+Scopus, 1 WoS, 5 Scopus, 1 РИНЦ)

Этап 2024 года «Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий»

Результаты этапа 2024 года.

1) Для исследования вопросов сочетания в ресурсах энциклопедической информации и онтологической структуры были проанализированы подходы к структуризации знаний в онлайн энциклопедиях: Большой Российской Энциклопедии (БРЭ), Википедии, «энциклопедии данных» Wikidata (Викиданные), проекте «Ковчег знаний» МГУ имени М.В. Ломоносова.

2) В рамках исследований по направлению методов ведения энциклопедий в заданной предметной области были исследованы возможности больших языковых моделей (далее также – LLM, от англ. Large Language Models) порождать заготовки энциклопедических статей (энциклопедических справок) по материалам научных публикаций.

3) Произведено исследование эффективности различных схем оптимизации словаря и настройки адаптеров для адаптации больших нейросетевых языковых моделей на русский язык.

4) Исследованы методы улучшения решения задачи абстрактного реферирования (англ. abstractive summarization) с применением больших языковых моделей, в том числе инструктивных. Предложена модель Pegasus-SP, которая объединяет псевдосуммирование с перестановкой предложений. Новая модель превосходит существующие аналоги в условиях ограниченных ресурсов и демонстрирует лучшую адаптивность;

5) Для пополнения онтологий и графов знаний исследована задача связывания извлеченных сущностей с единицами баз знаний. С этой целью исследовались методы нормализации медицинских понятий, то есть привязывания упоминания в тексте не-

которого варианта названия понятий к его нормативному наименованию. Особенностью исследования было то, что понятия могли быть вложенными друг в друга. Исследование было выполнено на новом вручную аннотированном наборе данных аннотаций PubMed;

6) Исследованы возможности языковой модели FlanT5 по тематической классификации текстов и объяснению принятых решений в виде наиболее значимых слов для классификации; - Исследован интерпретируемый метод классификации текстов - алгоритм построения формул, который конструирует представление текстовой темы в виде логической формулы. Представленный алгоритм показал хорошие результаты при сравнении с современными методами машинного обучения на реальных коллекциях с зашумленной экспертной разметкой.

7) Проанализированы модели грамматической компетенции, представленные направленными графами. Оценивались их способности к генерации, ограничения и возможности для учета грамматического варьирования и других ограничений больших языковых моделей.

По результатам этапа 2024 года опубликовано 19 научных работ, из них 14 индексируются Scopus, 6 Web Of Science.