

МГУ имени М.В. Ломоносова

4.04.Лаборатория анализа информационных ресурсов

№ госрегистрации
АААА-А20-120121690111-5

УТВЕРЖДАЮ
Директор/декан

«__» _____ Г.

УДК
004.8 Искусственный интеллект
519.767.6 Способы выделения семантической информации из текста

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

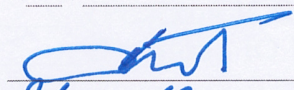
по теме:

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

Зам. директора/декана
по научной работе

«__» _____ Г.

Руководитель темы
Добров Б.В.

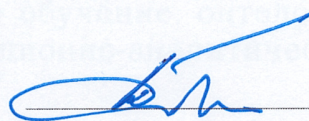

«21» 12 2023 Г.

Москва 2023

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель темы:

заведующий лабораторией, кандидат физико-математических наук



(Добров Б.В.)

Исполнители темы:

специалист

техник

ведущий специалист, кандидат филологических наук

заведующий лабораторией, кандидат филологических наук, доктор филологических наук, доцент по кафедре программист 2 категории

специалист

ведущий научный сотрудник, кандидат физико-математических наук, доктор технических наук

ведущий научный сотрудник, кандидат филологических наук, доктор филологических наук, доцент/с.н.с. по специальности

специалист

специалист

специалист, кандидат технических наук

специалист

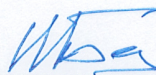
специалист

специалист

младший научный сотрудник, кандидат физико-математических наук

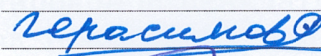
стажер

специалист

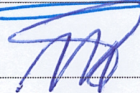


(Болдырев И.А.)

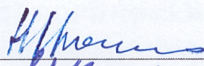
(Волков Е.В.)



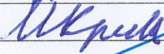
(Герасимова А.А.)



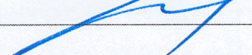
(Гращенко П.В.)



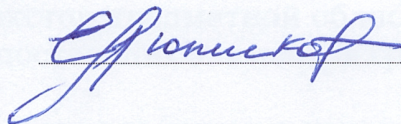
(Ионова Н.Р.)



(Кремнева М.Д.)



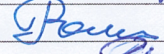
(Лукашевич Н.В.)



(Лютикова Е.А.)



(Мячина А.В.)



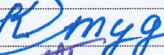
(Рожков И.С.)



(Рубцова Ю.В.)



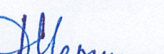
(Сидоров А.В.)



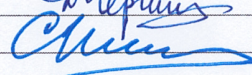
(Скворцов Н.А.)



(Студеникина К.А.)



(Тихомиров М.М.)



(Чернышев Д.И.)



(Штернов С.В.)

РЕФЕРАТ

Ключевые слова:

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

Ключевые слова по-английски:

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

Объектом исследования являются графы знаний – структуры, обобщающие онтологии, путем учета более широкой номенклатуры объектов – именованных сущностей различных типов, текстовых фрагментов, отражающих знания, содержащиеся в коллекции документов. Целью работы является разработка методов автоматизированного формирования и сопровождения графов знаний большого размера с использованием методов глубокого обучения на основе содержательной обработки больших массивов текстов, и на основе ранее созданных больших онтологических ресурсов. А также исследование методов использования больших графов знаний для поддержки решения информационно-аналитических задач в реальных социальноэкономических и научно-технических предметных областях.

В течение 2023 года при выполнении 4го этапа «Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области» включали:

- разработка методов поиска недостающих элементов графа знаний и наполнение графа знаний;
- разработка методов выявления в анализируемых текстах элементов графов знаний, проверки наличия новой информации и актуализации информации в графах знаний;
- исследование методов выявления возможных противоречий между новыми и хранимыми данными, разработка методов представления противоречивых данных в графе знаний.

Практическая значимость ожидаемых результатов заключается в создании новых инструментов для формирования графов знаний.

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

ВВЕДЕНИЕ

Целью работы является разработка методов автоматизированного формирования и сопровождения графов знаний большого размера с использованием методов глубокого обучения на основе содержательной обработки больших массивов текстов, и на основе ранее созданных больших онтологических ресурсов. А также исследование методов использования больших графов знаний для поддержки решения информационно-аналитических задач в реальных социально-экономических и научно-технических предметных областях.

В течение 2023 года при выполнении 4го этапа «Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области» включали:

- разработка методов поиска недостающих элементов графа знаний и на-
полнение графа знаний;
- разработка методов выявления в анализируемых текстах элементов гра-
фов знаний, проверки наличия новой информации и актуализации инфор-
мации в графах знаний;
- исследование методов выявления возможных противоречий между но-
выми и хранимыми данными, разработка методов представления противो-
речивых данных в графе знаний.

ОСНОВНАЯ ЧАСТЬ

Задачи НИР в 2023 году в ходе этапа «Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области» включали:

- разработать методы поиска недостающих элементов графа знаний и наполнение графа знаний;
- разработать методы выявления в анализируемых текстах элементов графов знаний, проверки наличия новой информации и актуализации информации в графах знаний;
- исследовать методы выявления возможных противоречий между новыми и хранимыми данными, разработка методов представления противоречивых данных в графе знаний.

По направлению разработки методов поиска недостающих элементов графа знаний и наполнение графа знаний исследовались методы установления отношений в графе знаний:

- исследованы новые методы установления таксономических отношений между существующими концептами лингвистических онтологий и терминоподобными сущностями [5, 9, 10];
- 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, 4];
- также исследовались задачи выявления отношений в сложных случаях вложенных именованных сущностей [7, 8, 11].

В работе [5] представлен обзор подходов к автоматическому извлечению таксономических отношений (IS_A отношений) из текстов, включая классические методы на основе лексико-семантических шаблонов, векторных представлений слов и их современное развитие. Также описаны новые подходы, основанные на языковых моделях типа BERT и использующие новые варианты работы с лексико-семантическими шаблонами, в которых модель должна заполнить нужную позицию в шаблоне. В статье также представлены различные способы тестирования качества извлечения таксономических отношений, приведены результаты этих тестирований. По достигаемым в настоящее время результатам можно сделать вывод, что, несмотря на прогресс в развитии методов извлечения таксономических отношений, качество извлечения является недостаточным для полностью автоматического построения или пополнения таксономий.

В статьях [9, 10] представлен подход к созданию интерпретируемых векторных представлений слов, в которых каждому компоненту вектора соответствует определенная интерпретируемая семантическая категория. Для получения таких категорий используется лексико-семантический ресурс в виде семантической сети RuWordNet, а также репрезентативный корпус русскоязычных текстов для формирования векторных представлений. Полученные интерпретированные векторные представления были протестированы на возможность отображения разных моделей в одно и то же векторном пространстве.

В работе [11] изучаются подходы к извлечению отношений между так называемыми вложенными именованными сущностями, то есть именованными сущностями, которые могут содержать другие именованные сущности. Большинство наборов данных, аннотированных вложенными именованными

ми сущностями, не имеют разметки отношений. Поэтому некоторые современные модели извлечения отношений не могут извлекать отношения между вложенными или перекрывающимися объектами из-за архитектуры или формата ввода. Используется корпус NEREL для русского языка, аннотированный вложенными именованными сущностями и отношениями между ними.

В статье [7] изучаются возможности модели машинного понимания (MRC) по переносу модели между предметными областями в задаче распознавания именованных объектов. В такой задаче модели требуется извлечь некоторые знания из текстов с подсказками, чтобы извлечь именованные сущности. Обычно для этого требуется обучение на большом наборе данных. Однако в сценариях с ограниченными ресурсами или в сценариях, специфичных для предметной области, создание набора обучающих данных может быть трудным и неэффективным. Именно здесь может быть полезен перенос модели из доступных наборов данных. В работе сравниваются различные запросы (prompts) к модели машинного чтения для передачи знаний между областями. Мы исследуем подход к переходу от общего набора данных NEREL к биомедицинскому набору NEREL-BIO, который имеет некоторый набор общих типов сущностей. В статье [8] исследуются различные методы генерации вопросов (подсказок) в вопросно-ответных подходах к распознаванию названных объектов. В таких методах распознавание понимается как ответ на вопросы об названных сущностях. Система должна возвращать допустимые именованные сущности требуемого типа, на которые есть ссылка в предложении. В частности, мы изучаем влияние подсказок на распознавание вложенных именованных объектов в условиях нескольких попыток. Мы тестируем различные подходы на основе подсказок на данных конкурса по извлечению вложенных именованных объектов RuNNE-2022.

По направлению разработки методов выявления в анализируемых текстах элементов графов знаний, проверки наличия новой информации и актуализации информации в графах знаний рассматривались задачи абстрактивного реферирования – выделение наиболее значимой информации в текстах.

Создание коллекции примеров для обучения систем абстрактивного аннотирования является затратным процессом из-за больших временных затрат и высоких требований к квалификации экспертов необходимых для написания качественной аннотаций. Для задачи абстрактивного аннотирования новостных кластеров в [2] предлагается новый метод создания коллекций для обучения нейросетевых методов аннотирования – ClusterVote, предназначенный моделировать особенности задачи путем учета информации в связанных документах. Метод может быть использован для формирования абстрактивных аннотаций различного уровня детализации, а также для получения экстрактивных аннотаций. Используя метод ClusterVote, была сформирована новая коллекция на английском и русском языках для обучения систем аннотирования новостных статей – Telegram News*CV. Экспериментальные результаты показывают, что при определенных параметрах сформированные ClusterVote коллекции имеют схожие экстрактивные характеристики с такими известными наборами данных как CNN/Daily Mail и при этом обладают более высокими показателями “фактической достоверности” – воспроизведения в аннотациях именованных сущностей исходных текстов, а также их связей.

В работе [1] исследовались свойства предварительно обученных моделей

абстрактного реферирования в условиях ограниченных ресурсов.

Недавние достижения в области абстрактного суммирования демонстрируют важность задач предварительного обучения, однако языковые модели общего назначения превосходят по эффективности специализированные подходы предварительного обучения. Проведено сравнение пяти современных предварительно обученных языковых моделей на пяти однодокументных наборах данных абстрактного суммирования из разных областей. Для проверки моделей мы предлагаем четыре новых теста на понимание задач, которые оценивают основные компоненты моделей суммирования. Наши эксперименты показывают, что предварительная подготовка псевдосуммирования смещает модели в сторону более экстрактивного поведения и препятствует их способности правильно фильтровать существенное содержание, что приводит к худшему обобщению.

По направлению исследования методов выявления возможных противоречий между новыми и хранимыми данными, разработка методов представления противоречивых данных в графе знаний были выполнены исследования по воспроизведению лингвистических характеристиках текстов при применении современных методов автоматической обработки.

В работе [6] рассматриваются условия местности на Негативном Согласии (НК) на русском языке. Изучалось удаленное лицензирование отрицательных ни-местоимений в инфинитивах, контролирующих субъект и объект. Существующие теории отрицательного согласия так или иначе опираются на движение элементов отрицательного согласия (NCI): либо как прямое условие лицензирования, либо как предпосылку для входа в локальную конфигурацию с лицензиаром. Следовательно, теории предсказывают, что перемещение местоимения ni в матричное положение (т. е. в локальный домен матричного лицензиара) повысит приемлемость ni -лицензирования. Используя экспериментальные данные, мы исследуем различные линейные позиции, доступные для ни-местоимений в матричном предложении и во встроенных инфинитивных предложениях. Показано, что для лицензирования на большие расстояния не требуется передвижение и что производные позиции ни-местоимений не зависят от ни-лицензирования. Хотя и отказ от лицензирования, и перемещение подчиняются местным ограничениям, они различаются по типу требуемого местоположения. Представлены доказательства недавних предположений о том, что движение и соглашение имеют разные свойства в отношении ограничений местоположения.

В статьях [12, 13] рассматривается проблема согласования сочинительных конструкций и чередования форм чисел. Цель, согласуясь с двумя соединенными существительными в единственном числе, копирует признак числа либо во множественном, либо в единственном числе. Рассматривается синтаксис русских именных словосочетаний с согласованными модификаторами, демонстрирующий числовое изменение контролера соглашения. Если два соединенных прилагательных в единственном числе имеют разное толкование, допустимы существительные как в единственном, так и во множественном числе. В отличие от предыдущих исследований, основанных на самоанализе и корпусе, здесь параметризуется изменение числа на основе результатов экспериментов по приемлемости в самостоятельном темпе (шкала Лайкерта 1-7). Сравниваются данные о человеческом восприятии с языковыми вероятностями, предсказанными нейронной моделью генерации текста `gpt-3`. Два тематических исследования были проведены для

анализа морфологических и синтаксических факторов, параметризующих вариацию. Первое исследование рассматривает влияние морфологии существительного на чередование чисел. Второе исследование изучает влияние атрибутивного соглашения премодификатора. Как оценки приемлемости в автономном режиме, так и время онлайн-чтения показывают, что наблюдаемые морфологические и синтаксические факторы следует учитывать при параметризации числовых вариаций в русских именных группах с согласованными модификаторами. Языковая модель ruGPT-3, обученная на обширной коллекции русских текстов, способна предсказать правильную вероятность для весьма приемлемых и крайне неприемлемых предложений, но не может назначить точные значения вероятности для случаев вариации.

СПИСОК ПУБЛИКАЦИЙ

1. Chernyshev D.I., Evaluating the Summarization Comprehension of Pre-Trained Language Models / Chernyshev D.I., Dobrov B.V. // Lobachevskii Journal of Mathematics. - Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation). - 2023. - том 44, № 8, с. 3028-3039. (WoS, Scopus)
2. Chernyshev D.I., The Method of Automatic Construction of Training Collections for the Task of Abstractive Summarization of News Articles / Chernyshev D.I., Dobrov B.V. // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. - Pleiades Publishing, Ltd (Road Town, United Kingdom), - 2023. - том 33, № 3, с. 255-267. (Scopus)
3. Loukachevitch N., NEREL: a Russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links / Loukachevitch Natalia, Artemova Ekaterina, Batura Tatiana, Braslavski Pavel, Ivanov Vladimir, Manandhar Suresh, Pugachev Alexander, Rozhkov Igor, Shelmanov Artem, Tutubalina Elena, Yandutov Alexey // Language Resources and Evaluation. - Springer Verlag (Germany). - 2023. (Scopus)
4. Loukachevitch N., NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities / Loukachevitch Natalia, Manandhar Suresh, Baral Elina, Rozhkov Igor, Braslavski Pavel, Ivanov Vladimir, Batura Tatiana, Tutubalina Elena // Bioinformatics, издательство Oxford University Press (United Kingdom). - 2023. - том 39, № 4. (Q1)
5. Loukachevitch N.V., Automatic Methods for Extracting Taxonomic Relationships from Texts / Loukachevitch N.V. // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. - Pleiades Publishing, Ltd (Road Town, United Kingdom), - 2023. - том 33, № 3, с. 398-406. (Scopus)
6. Lyutikova E., Negative concord and locality in Russian / Lyutikova E., Gerasimova A. // Canadian Journal of Linguistics. — 2023. — P. 1-43. (Scopus)
7. Rozhkov I.S., Machine Reading Comprehension Model in Domain-Transfer Task / Rozhkov I.S., Loukachevitch N.V. // Lobachevskii Journal of Mathematics. - Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation). - 2023. - том 44, № 8, с. 3159-3167. (WoS, Scopus)
8. Rozhkov I.S., Prompts in Few-Shot Named Entity Recognition / Rozhkov I.S., Loukachevitch N.V. // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, издательство Pleiades Publishing, Ltd (Road Town, United Kingdom). - 2023. - том 33, № 2, с. 122-131. (Scopus)
9. Tikhomirov M.M., Generation of Interpreted Vector Representations of Words Based on Supersenses / Tikhomirov M.M., Loukachevitch N.V. // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. - Pleiades Publishing, Ltd (Road Town, United Kingdom). - 2023. - том 33, № 3,

с. 517-524. (Scopus)

10. Tikhomirov M.M., Method for Generating Interpretable Embeddings Based on Superconcepts / Tikhomirov M.M., Loukachevitch N.V. // Lobachevskii Journal of Mathematics. - Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation). - 2023. - том 44, № 8, с. 3168-3176. (WoS, Scopus)

11. Yandutov A.V., Approaches to Relation Extraction for Nested Named Entities / Yandutov A.V., Loukachevitch N.V. // Lobachevskii Journal of Mathematics. - Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), - 2023. - том 44, № 1, с. 249-258. (WoS, Scopus)

12. Студеникина К. А., Parametrizing number variation in russian noun phrases with experimental studies and language modeling / Студеникина К. А. // Научный результат. Вопросы теоретической и прикладной лингвистики. — 2023. — Vol. 9, no. 1. — P. 192-205. (РИНЦ)

13. Студеникина К. А., Иерархия согласования при сочинении русских именных групп: корпусное и экспериментальное исследование / Студеникина К. А. // Русский язык в научном освещении. — 2023. — Т. 46, № 2. — С. 25-37. (РИНЦ, WoS)

ЗАКЛЮЧЕНИЕ

При выполнении НИР в 2023 году в ходе этапа «Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области» получены следующие результаты:

- исследованы новые методы установления таксономических отношений между существующими концептами лингвистических онтологий и терминоподобными сущностями;

- 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;

- исследованы задачи выявления отношений в сложных случаях вложенных именованных сущностей;

- рассматривались задачи абстрактивного аннотирования: для задачи абстрактивного аннотирования новостных кластеров предложен новый метод создания коллекций для обучения нейросетевых методов аннотирования, предназначенный моделировать особенности задачи путем учета информации в связанных документах;

- исследованы предварительно обученных моделей абстрактивного реферирования в условиях ограниченных ресурсов;

- 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)