



МГУ имени М.В. Ломоносова

Научно-исследовательский вычислительный центр

Лаборатория анализа информационных ресурсов (404)

Лаборатория информационных систем математических наук (418)

Отчет за 2024 год

Тема:

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

Этап 5. (2024 г.) Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий

Кадровый состав (по теме)

➤ ЛАИР (лаб. 404):

- Добров Б.В., Лукашевич Н.В., Мячина А.В., Рожков И.С., Сидоров А.В., Тихомиров М.М., *Чернышев Д.И.*, Штернов С.В., Кремнева М.Д., Ионова Н.Р., *Ярошенко П.В.*
Лютикова Е.А.

- совм. Волков Е.В. [1/4]

➤ ЛАЛС (лаб. 415):

- Гращенко П.В., Герасимова А.А., Студеникина К.А.

Планы на 2024 год

Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий

- разработать методы ведения текстовых энциклопедий в предметной области ("корпоративных энциклопедий");
- исследовать методы формирования энциклопедической статьи по массиву новостных текстов;
- исследовать возможности современных нейросетевых языковых моделей для задачи ведения текстовых энциклопедий в предметной области.

Практическая значимость ожидаемых результатов заключается в создании новых инструментов для формирования графов знаний в форме связанных текстовых статей.

Формы завершения: Отчет, WoS+Scopus=6, другие=2.

Публикации 2024

	Тема	
	ЛАИР (404)	ЛАЛС (415)
Q1	1	
WoS, Scopus	14	1
РИНЦ	1	2
Всего	19	

	Q1	Wos, Scopus	РИНЦ	
2021	1	6	3	11
2022	2	10	1	13
2023	1	14	3	19
				53 ⁴

Поддержка исследований

- **Чернышёв Д.И.** получил грантовую поддержку Фонда «Интеллект» (аспирантам) по теме «Улучшение качества автоматических методов порождения сложноструктурированных повествовательных текстов на основе интеграции в нейросетевые модели абстрактного реферирования общих и специальных знаний» (2024)
- **Тихомиров М.М.** получил грантовую поддержку Фонда «Интеллект» (молодым ученым) по теме «Автоматические методы адаптации мультязычных ресурсов и моделей на конкретный домен (язык, предметную область)» (2024)
- **Рожков И.С.** получил стипендию имени Б.И.Березина, а также стипендию МГУ

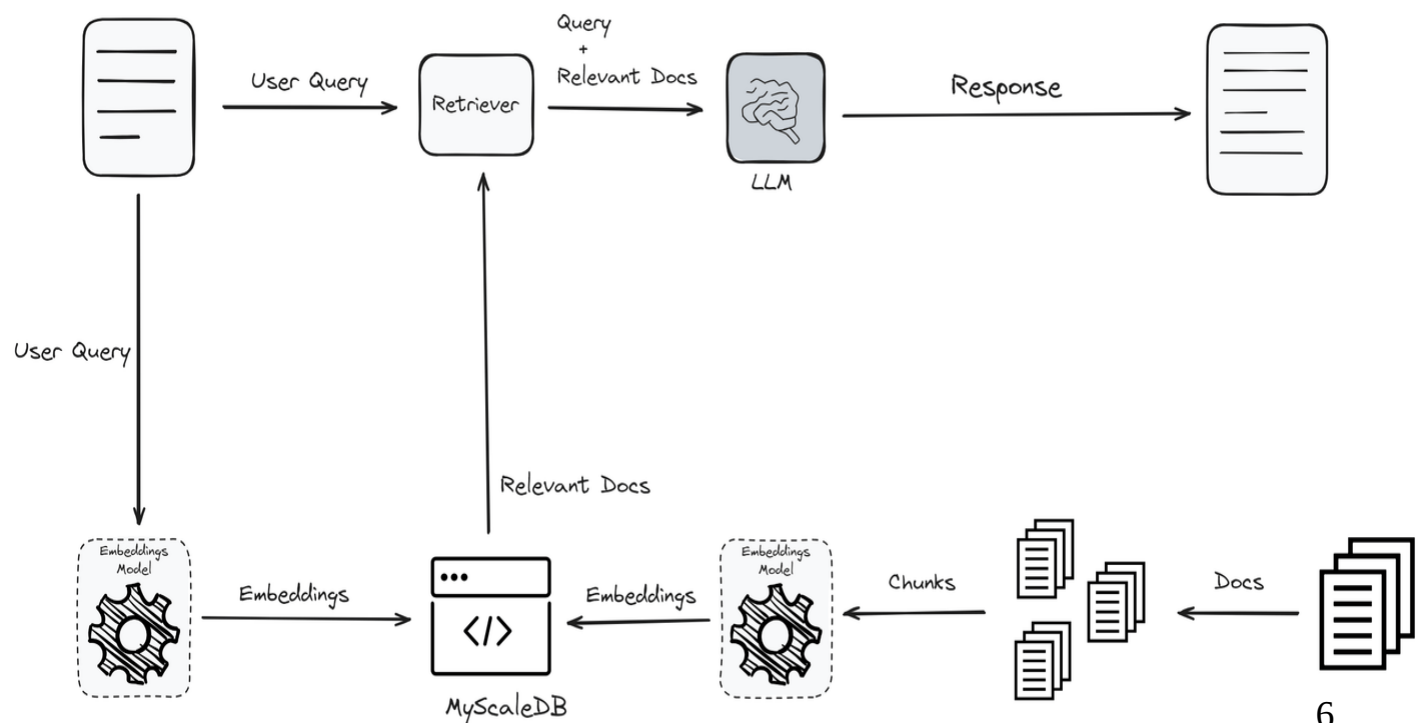
➤ Грант РФФ

РФФ 25-71-00206

25 декабря 2024 - 16 декабря 2026 **Семантический поиск на основе нейросетевых моделей информационного-поиска для русского языка: создание бенчмарка и разработка новых моделей**

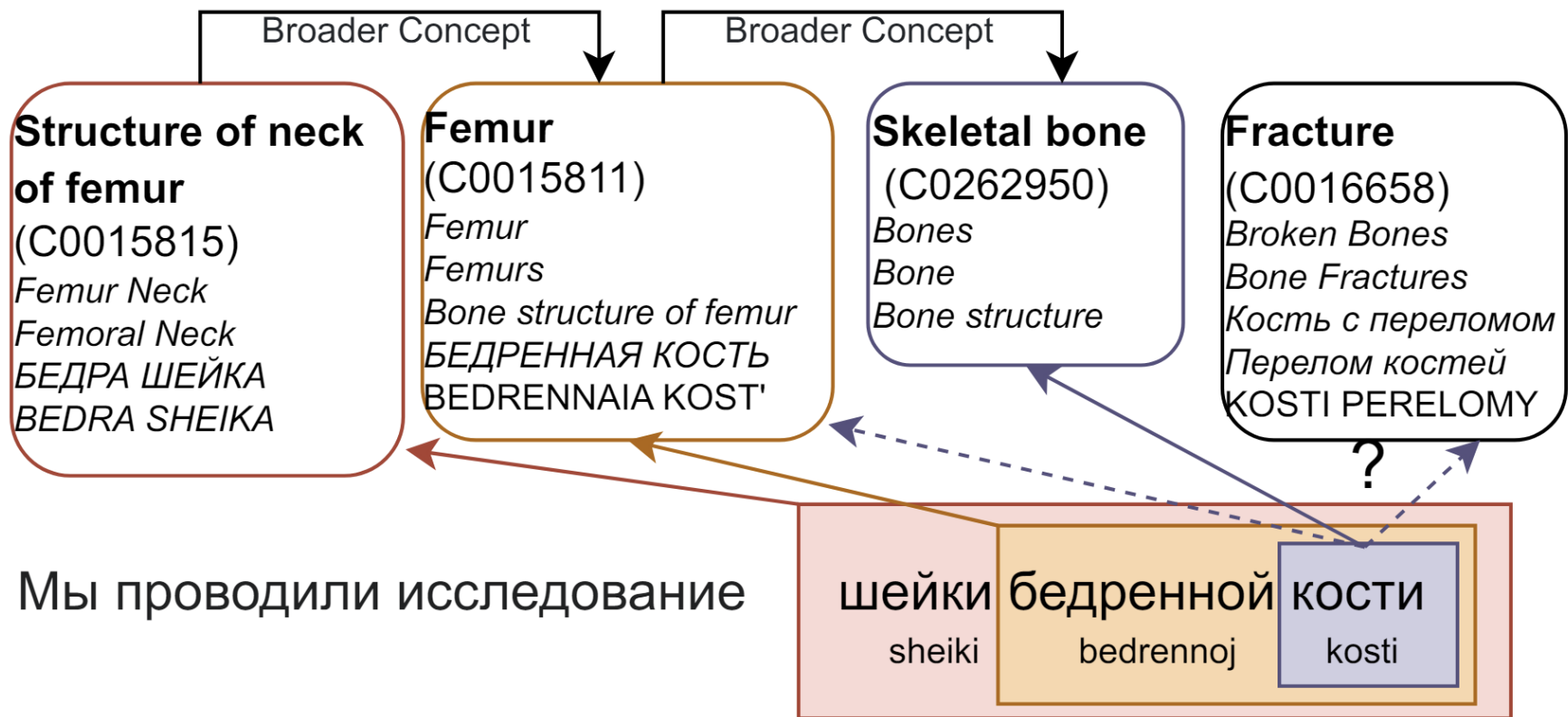
Руководитель: Добров Б.В. Участники : Лукашевич Н.В., Тихомиров М.М. Ковалев Г.П.

➤ + 1 НИР



Связывание русских медицинских текстов с базой медицинских понятий UMLS

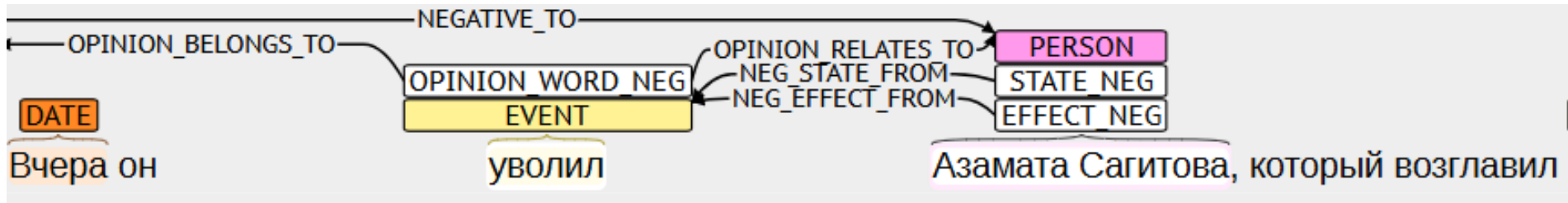
Проблема: в UMLS недостаточно русских переводов.
Попытка использовать вложенность понятий



Мы проводили исследование

We performed a study of the femoral neck

Извлечение структурированных мнений из текстов. Тестирование RuOpinion-2024



- Вход: предложение
- Результат: кортеж
- Сущности-объекты не заданы, но типы их известны
 - Person, Organization, Country...
- Мера качества:
 - F-мера
 - Позволяется неточное извлечение

```
{  
  "sent_id": 1,  
  "text": "Вчера он уволил Азамата Сагитова, который возглавил башкирскую администрацию год назад после вынужденно  
й отставки Радия Хабирова, сейчас занимающего пост замест  
ителя начальника управления президента РФ по внутренней п  
олитике.",  
  "opinions": [  
    {  
      "Source": [["NULL"], ["0:0"]],  
      "Target": [["Азамата Сагитова"], ["16:32"]],  
      "Polar_expression": [["уволил"], ["9:15"]],  
      "Polarity": "NEG"  
    }  
  ]  
}
```

Проект Ruadapt

Цель - перенос мультязычных LLM на русский язык

Основные шаги:

- Подготовка более подходящей для русского языка токенизации
- Адаптация базовой версии модели на новую токенизацию
- Отображение из базовой модели в инструктивную (LFR)
- Дополнительное дообучение на инструкциях

На текущий момент этап адаптации базы происходит на датасете в 60-80GB:

- Википедия
- Новости
- Хабр
- Книги (либрусек)

Почему недостаточно разработанных методов

- Адаптация происходит для **базовых моделей**, не инструктивных
- Соответственно **требуется воспроизводить** процедуру **instruct-tuning** с базы
 - LLaMa-3 обучалась на 10 миллионах инструкций (датасет не опубликован!)
 - Лучшая версия модели mistral openchat-3.5 обучалась на датасете, который также **закрыт**

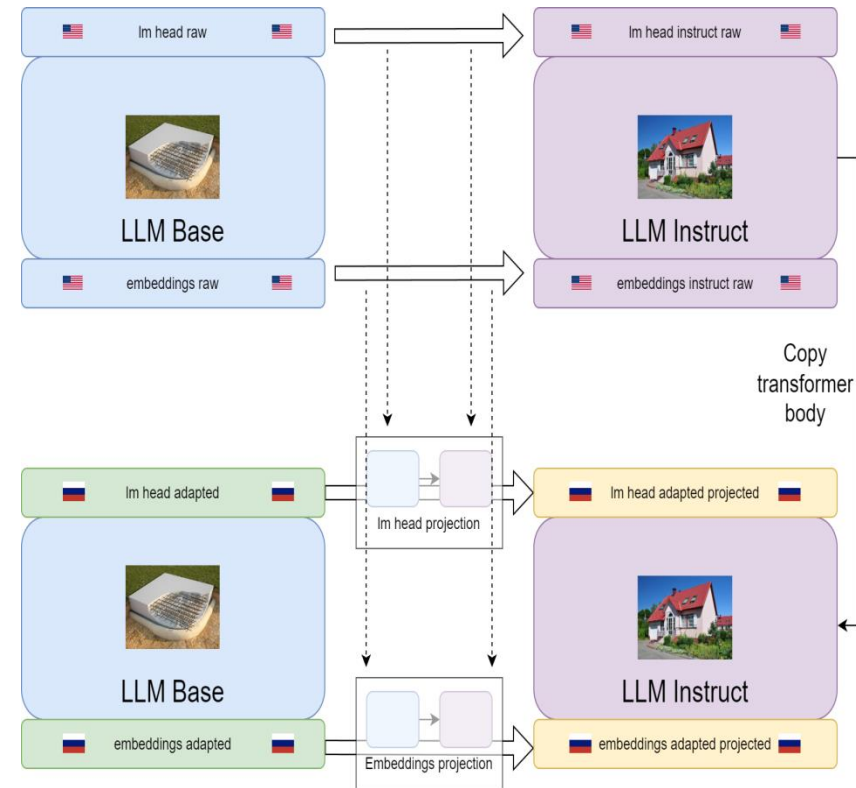
Ruadapt: Learned Embeddings Propagation (LEP)

Решение описанных недостатков!

- Как результат - позволяет **адаптировать инструктивные версии моделей**

Основная идея:

- **Адаптировать базовую модель**
- **Рассчитать проекцию из эмбедингов исходной базовой в инстракт версию**
- **Применить проекцию на адаптированную базу**



OpenChat-3.5

Фраза "моя хата с краю" не имеет прямого значения на русском языке. Однако, если рассматривать её в контексте украинского языка, то "хата" означает "изба" или "хутор", а "краю" означает "краю" или "границе". Таким образом, можно предположить, что эта фраза означает "моя изба на границе" или "моя изба вдали от города". Однако, без контекста, это лишь предположение, и значение может быть разным в зависимости от контекста и языка.

OpenChat-3.5 + LEP-Extended

Фраза "моя хата с краю" является частью **русской народной мудрости** и имеет несколько значений. В целом, она означает что-то в духе "я живу на окраине, вдали от центра событий или общества", но может иметь и другие толкования в зависимости от контекста.

OpenChat-3.5 + LEP-Extended + Calibration (saiga d7 + copy task)

Фраза "моя хата с краю" означает, что **человек не хочет вмешиваться** в дела других людей или не хочет принимать участие в конфликте, предпочитая держаться подальше от него. Фраза происходит из **русской народной мудрости**, которая гласит: "Моя хата с краю, а я не влезу".

RuadaptQwen2.5: промежуточный итог

- RuadaptQwen2.5-1.5B-instruct
- RuadaptQwen2.5-3B-instruct
- RuadaptQwen2.5-7B-instruct
- **RuadaptQwen2.5-32B-instruct**
 - ~5000 GPU часов на 1 “прогон” эксперимента
- RuadaptQwQ 32B
 - Только LEP версия

Ускорение по сравнению с Qwen2.5 версиями до **30-60%**
при генерации **русского текста** (в символах)

Russian SuperGLUE

Rank	Name	Team	Link	Score	LiDiRus	RCB	PARus	MuSeRC
1	HUMAN BENCHMARK	AGI NLP	i	0.811	0.626	0.68 / 0.702	0.982	0.806 / 0.42
2	ruadapt Solar 10.7 twostage	RCC MSU	i	0.805	0.591	0.597 / 0.594	0.916	0.946 / 0.837
3	Mistral 7B LoRA	Saiga team	i	0.763	0.46	0.529 / 0.573	0.824	0.927 / 0.787
4	FRED-T5 1.7B finetune	SberDevices	i	0.762	0.497	0.497 / 0.541	0.842	0.916 / 0.773
5	Golden Transformer v2.0	Avengers Ensemble	i	0.755	0.515	0.384 / 0.534	0.906	0.936 / 0.804
6	LLaMA-2 13B LoRA	Saiga team	i	0.718	0.398	0.489	0.784	0.919 /

RuadaptQwen2.5: результаты

Ru Arena General

model	▲ score ▲	95% CI	▲ avg_tokens
-chatgpt-4o-latest	94.74	+0.8 / -0.8	693.15
o1-mini	93.46	+0.9 / -0.8	791.18
yi-lightning	93.46	+1.0 / -1.1	636.68
claude-3-opus-20240229	91.31	+0.9 / -1.1	468.69
gpt-4-1106-preview	90.89	+1.2 / -1.2	541.66
o1-preview	90.8	+1.1 / -1.1	664.89
<u>RefalMachine-RuadaptQwen2.5-32B-instruct-v1</u>	90.47	+1.3 / -1.1	527.86
SberDevices-GigaChatMaxWithoutFilter	89.96	+1.1 / -1.2	523.95
DeepSeek, Inc.-DeepSeek-V2-Chat-0628	89.67	+1.0 / -1.0	514.79
gemini-1.5-pro-002	89.07	+1.2 / -1.1	639.51
gemini-1.5-pro-exp-0801	88.88	+1.1 / -1.0	547.91

1 2	Модель, команда ↓	Результат ↓
1	GPT4o MERA	0.642
2	RuadaptQwen-32B-instr... 32.0B RCC MSU	0.615
3	Qwen2.5-32B-Instruct 32.0B RCC MSU	0.603
4	Qwen2.5-72B-Instruct 72.7B MERA	0.601
5	Meta-Llama-3.1-405B-In... 405.0B MERA	0.59
6	GigaChat Max GIGACHAT	0.588
7	Mistral-Large-Instruct-2... 123.0B MERA	0.574
8	GPT4o-mini MERA	0.57

Публикации по теме

1. Chernyshev D.I. Improving Abstractive Summarization with Unsupervised Dynamic LoRA Mixtures //Lobachevskii Journal of Mathematics, издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), 2024, том 45, № 7, с. 2995-3006. - DOI: 10.1134/s1995080224603710 (Scopus Q2, WoS Q2)
2. Chernyshev Daniil, Dobrov Boris. Content Selection in Abstractive Summarization with Biased Encoder Mixtures //Lecture Notes in Computer Science, издательство Springer Nature (Switzerland), с. 64-77. 2024. - DOI: 10.1007/978-3-031-54534-4_5 (Scopus)
3. Chernyshev Daniil, Dobrov Boris. Investigating the Pre-Training Bias in Low-Resource Abstractive Summarization // IEEE ACCESS, том 12, 2024. С. 47219-47230. - DOI:10.1109/ACCESS.2024.3379139 (Scopus Q1, WoS Q1)
4. Davydova V., Loukachevitch N., Tutubalina E. Overview of BioNNE Task on Biomedical Nested Named Entity Recognition at BioASQ 2024 //CLEF Working Notes (Scopus)
5. Loukachevitch Natalia, Sakhovskiy Andrey, Tutubalina Elena. Biomedical Concept Normalization over Nested Entities with Partial UMLS Terminology in Russian //Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), место издания ELRA and ICCL Torino, Italia, 2024. С. 2383-2389. (Scopus)
6. Nentidis Anastasios, Krithara Anastasia, Paliouras Georgios, Krallinger Martin, Sanchez Luis Gasco, Lima Salvador, Farre Eulalia, Loukachevitch Natalia, Davydova Vera, Tutubalina Elena. BioASQ at CLEF2024: The Twelfth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge //Lecture Notes in Computer Science, издательство Springer Nature (Switzerland), с. 490-497. 2024. - DOI: 10.1007/978-3-031-56069-9_16 (Scopus)

Планы на 2025 год

Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий

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

Практическая значимость ожидаемых результатов заключается в создании новых инструментов для формирования графов знаний в форме связанных текстовых статей.

Формы завершения: Отчет, WoS+Scopus=6, другие=2.

Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий

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

Разработка методов автоматизированного формирования графов знаний в форме корпоративных энциклопедий (2)

- 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 некоторого 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;
- 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;
- 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.
- 19 статей (1 Q1, 13 Wos+Scopus, 4 РИНЦ)

Публикации по теме

1. Chernyshev D.I. Improving Abstractive Summarization with Unsupervised Dynamic LoRA Mixtures //Lobachevskii Journal of Mathematics, издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), 2024, том 45, № 7, с. 2995-3006. - DOI: 10.1134/s1995080224603710 (Scopus Q2, WoS Q2)
2. Chernyshev Daniil, Dobrov Boris. Content Selection in Abstractive Summarization with Biased Encoder Mixtures //Lecture Notes in Computer Science, издательство Springer Nature (Switzerland), с. 64-77. 2024. - DOI: 10.1007/978-3-031-54534-4_5 (Scopus)
3. Chernyshev Daniil, Dobrov Boris. Investigating the Pre-Training Bias in Low-Resource Abstractive Summarization // IEEE ACCESS, том 12, 2024. С. 47219-47230. - DOI:10.1109/ACCESS.2024.3379139 (Scopus Q1, WoS Q1)
4. Davydova V., Loukachevitch N., Tutubalina E. Overview of BioNNE Task on Biomedical Nested Named Entity Recognition at BioASQ 2024 //CLEF Working Notes (Scopus)
5. Loukachevitch Natalia, Sakhovskiy Andrey, Tutubalina Elena. Biomedical Concept Normalization over Nested Entities with Partial UMLS Terminology in Russian //Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), место издания ELRA and ICCL Torino, Italia, 2024. С. 2383-2389. (Scopus)
6. Nentidis Anastasios, Krithara Anastasia, Paliouras Georgios, Krallinger Martin, Sanchez Luis Gasco, Lima Salvador, Farre Eulalia, Loukachevitch Natalia, Davydova Vera, Tutubalina Elena. BioASQ at CLEF2024: The Twelfth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge //Lecture Notes in Computer Science, издательство Springer Nature (Switzerland), с. 490-497. 2024. - DOI: 10.1007/978-3-031-56069-9_20 (Scopus)

Публикации по теме (2)

7. Rogov A., Loukachevitch N. Automatic Evaluation of Interpretability Methods in Text Categorization //Journal of Mathematical Sciences. – 2024. – Т. 285. – №. 2. – С. 201-209. – DOI: 10.1007/s10958-024-07426-0 (Scopus Q3)
8. Rogov A.A., Loukachevitch N.V. Evaluating the Performance of Interpretability Methods in Text Categorization Task //Lobachevskii Journal of Mathematics, - издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), том 45, № 3, 2024. С. 1234-1245. DOI: 10.1134/S1995080224600699 (Scopus Q2, WoS Q2)
9. Rusnachenko Nicolay, Golubev Anton, Loukachevitch Natalia. Large Language Models in Targeted Sentiment Analysis //Lobachevskii Journal of Mathematics, издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), том 45, № 8, 2024. С. 3148-3158. - DOI: 10.1134/S1995080224603758 (Scopus Q2, WoS Q2)
10. Solomatina Yuliana, Loukachevitch Natalia. Prompt-Tuning for Targeted Sentiment Analysis in Russian //Communications in Computer and Information Science, издательство Springer Nature (Switzerland), с. 111-124. 2024. - DOI: 10.1007/978-3-031-67008-4_9 (Scopus Q4)
11. Sulzhenko R.V., Dobrov B.V. Approximation of the meaning for thematic subject headings by simple interpretable representations //Lobachevskii Journal of Mathematics, издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), том 45, № 3, 2024. С. 1261-1274. - DOI: 10.1134/S1995080224600778 (Scopus Q2, WoS Q2)
12. Tikhomirov M.M., Chernyshev D.I. Improving Large Language Model Russian Adaptation with Preliminary Vocabulary Optimization //Lobachevskii Journal of Mathematics, издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), 2024, том 45, № 7, с. 3211-3219. - DOI: 10.1134/s1995080224604120 (Scopus Q2, WoS Q2)

Публикации по теме (2)

13. Yandutov Alexey V., Loukachevitch Natalia V. Document-Level Relation Extraction in Russian // Communications in Computer and Information Science, издательство Springer Nature (Switzerland), с. 136-148. - DOI: 10.1007/978-3-031-67008-4_11 (Scopus Q4)
14. Бубнов А.С., Галлини Н.И., Гришин И.Ю., Кобозева И.М., Лукашевич Н.В., Панич М.Б., Раевский Е.Н., Садковский Ф.А., Тимиргалеева Р.Р. Онтологии как фундамент формализации научной информации и извлечения новых знаний // Доклады Российской академии наук. Математика, информатика, процессы управления, 2024, Том 520, с. 87-95 (РИНЦ, УБС4)
15. Герасимова А.А., Лютикова Е.А., Паско Л.И. Языковая компетенция сквозь призму грамматической вариативности. Часть 1. Теоретические и методологические соображения // Вестник Московского университета. Серия 9: Филология, издательство Изд-во Моск. ун-та (М.), 2024. № 4. - DOI: 10.55959/MSU0130-0075-9-2024-47-04-1 (РИНЦ)
16. Герасимова А.А., Лютикова Е.А., Паско Л.И. Языковая компетенция сквозь призму грамматической вариативности. Часть 2. Количественная оценка // Вестник Московского университета. Серия 9: Филология, издательство Изд-во Моск. ун-та (М.), № 5, 2024. С. 111-128. - DOI: 10.55959/MSU0130-0075-9-2024-47-05-9 (РИНЦ)
17. Гращенков П.В., Паско Л.И., Студеникина К.А. Ограничение сочинительного острова в лингвистической компетенции больших языковых моделей // Вестник Санкт-Петербургского университета. Язык и литература, издательство Федеральное государственное бюджетное образовательное учреждение высшего образования "Санкт-Петербургский государственный университет" (Санкт-Петербург), том 21, № 3, 2024. С. 668-689. – DOI: 10.21638/spbu09.2024.309 (Scopus)
18. Гращенков П. Вариативное согласование и структура сочиненной составляющей // Вопросы языкознания. – 2024. - №3. – С.7–30. – DOI: 10.31857/0373-658X.2024.3.7-30 (РИНЦ)
19. Лемаев В.И., Лукашевич Н.В. Автоматическая классификация эмоций в речи: методы и данные // Litera, издательство Nota bene (М.), 2024. – DOI: 10.25136/2409-8698.2024.4.70472