



**МГУ имени М.В. Ломоносова**

**Научно-исследовательский вычислительный центр**

**Лаборатория анализа информационных ресурсов**

## **Отчет за 2022 год**

- **Тема:**  
Методы структурирования трудноформализуемых предметных областей на основе автоматизированного формирования больших графов знаний и онтологий по разнородным потокам текстовых данных
- **Этап 3. (2022 г.)** Разработка методов автоматизированного формирования больших графов знаний предметной области

# Кадровый состав (по теме)

## ➤ ЛАИР (лаб. 404):

- Добров Б.В., Лукашевич Н.В., Мячина А.В., Сидоров А.В., Тихомиров М.М., Чернышев Д.И., Штернов С.В.
- + Рубцова Ю.В. (РНФ НИВЦ)
- + Кремнева М.Д. (420-ЛКБ)
- + Болдырев И.А. [1/10], Волков Е.В. [1/4], Архипенко К.В. [1/10], Майоров В.Д. [1/10], Скворцов Н.А. [1/2]

## ➤ РНФ (Лукашевич Н.В.) [1/10] :

- Артемова Е.Л., Батура Т.В., Браславский П.И., Иванов В.В., Тутубалина Е.В., Шелманов А.О.

## ➤ РНФ (Швядас В.К., лаб. 407) [1/10 – 1/4] :

- Подшивалов Д.Д.

## ➤ ЛАЛС (лаб. 415):

- Гращенков П.В., Герасимова А.А., Лютикова Е.А.

# Планы на 2022 год

## Разработка методов автоматизированного формирования больших графов знаний предметной области

- исследовать **методы автоматического извлечения неизвестных отношений** (Open Information Extraction) с использованием методов машинного обучения и ранее созданных онтологических ресурсов
- разработать методы автоматизированного формирования больших онтологий предметной области с **развитым набором отношений**
- разработать **методы разрешения многозначности текстового выражения именованных сущностей** в разных документах
- разработать **методы связывания различных текстовых вариантов извлеченных именованных сущностей** на основе результатов обработки больших текстовых коллекций

# НИР (1)

- 28 мая 2020 г. - 30 декабря 2022 г. *РНФ 20-11-20166*  
**Автоматические методы построения и пополнения баз знаний на основе кросс-языковых технологий**  
Руководитель: Лукашевич Н.В. Участники НИР: ..., Тихомиров М.М.
- 1 октября 2019 г. - 1 октября 2022 г. *РФФИ 19-37-90119\19*  
**Автоматические методы улучшения качества векторных представлений лексических единиц за счет распространения контекстов по лексико-семантическим отношениям**  
Руководитель: Лукашевич Н.В. Отв. исполнитель: Тихомиров М.М.
- 19 февраля 2020 г. - 30 декабря 2022 г. *РФФИ 20-07-01059\20*  
**Автоматический анализ тональности текстов с множественными оценками на основе оценочных фреймов**  
Руководитель: Лукашевич Н.В. Участники НИР: ..., Тихомиров М.М., Штернов С.В.
- Чернышёв Д.И. получил грантовую поддержку Фонда «Интеллект» о  
**Улучшение качества автоматических методов порождения сложноструктурированных повествовательных текстов на основе интеграции в нейросетевые модели абстрактного реферирования общих и специальных знаний**

## НИР (2)

- 17 марта 2021 г. - 31 декабря 2024 г.

**РНФ 21-71-30003**

***Разработка и реализация принципов суперкомпьютерного кодизаина для развития сквозных цифровых технологий и создания высокопроизводительных вычислительных платформ для предсказательного моделирования и применения искусственного интеллекта: от микромира, до живых организмов и Земной системы***

Руководитель: Швядас В.К. Участники НИР: ... Волков Е.В. ...

Добров Б.В., ... Лукашевич Н.В., ... Тихомиров М.М. ... Чернышев Д.И.

- + 3 НИР

# Порождение графов знаний из текстов

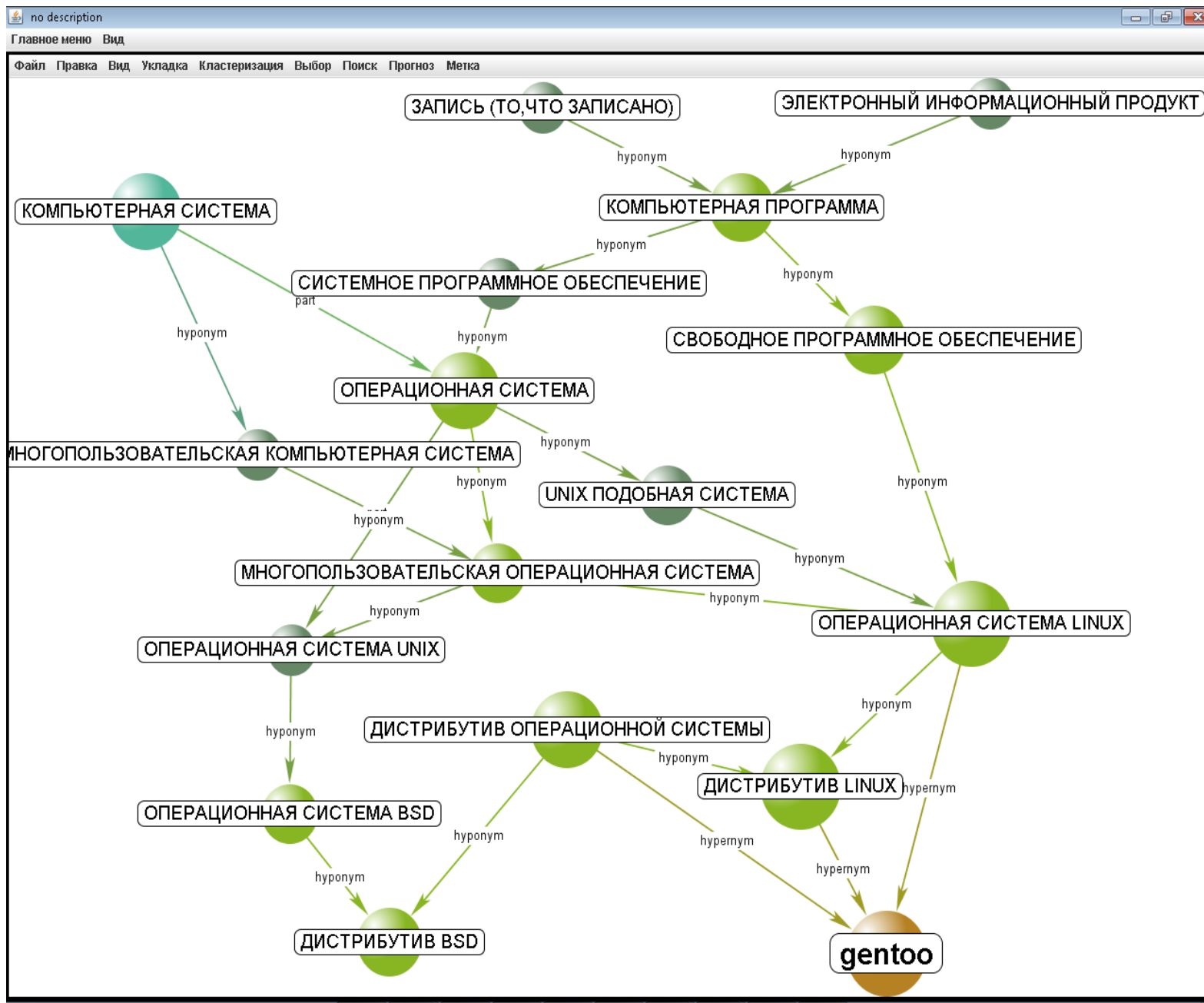
- Графы знаний стали важным инструментом в приложениях автоматической обработки текстов
  - Необходимость пополнения графов знаний по текстам
- Задачи, которые нужно решить при пополнении графа знаний
  - Извлечение сущностей (понятий и имен)
  - Извлечение отношений
  - Связывание текстовых сущностей с графом знаний (entity linking)
- Созданы новые размеченные наборы данных NEREL и NEREL-BIO с разметкой вложенных именованных сущностей
  - Вложенные именованные сущности позволяют увеличить полноту разметки, но требуют специальных моделей для извлечения

**М.М.Тихомиров (дисс. канд.физ.-мат.наук)**

## **Методы автоматизированного пополнения графов знаний на основе векторных представлений**

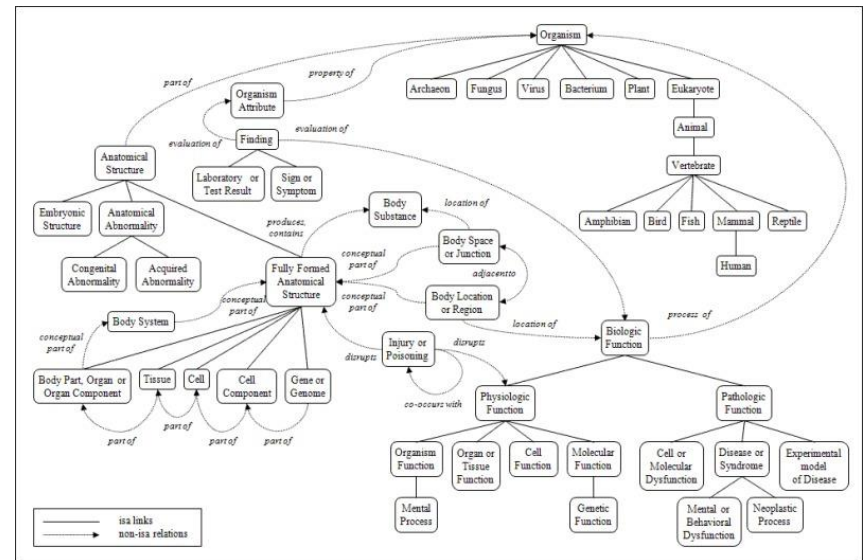
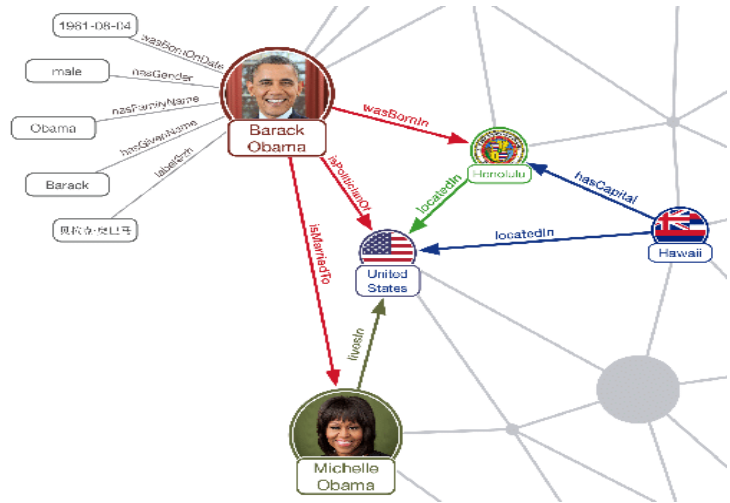
- Научный руководитель: Лукашевич Наталья Валентиновна, д.т.н., МГУ имени М.В. Ломоносова
- Защищена в совете МГУ.012.2(01.19) МГУ имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики
- Разработка методов пополнения графов знаний новыми понятиями и именованными сущностями
- Исследованы различные способы комбинирования векторных представлений слов и показано, что комбинация представлений с помощью автокодировщиков с учетом дополнительной информации о задаче приводит к улучшению качества векторных представлений, что в свою очередь приводит к улучшению качества решения целевой задачи
- Разработаны и реализованы подходы к пополнению таксономии графа знаний новыми понятиями и к извлечению именованных сущностей в предметной области информационной безопасности

# Предсказание гиперонимов и меронимов



# Графы знаний

- Граф знаний - это семантическая сеть с большим числом узлов (понятий) и отношений между ними
  - Содержит большой объем знаний о мире
  - Используется в автоматической обработке текстов, инф. поиске
- Примеры
  - Wikidata, Dbpedia (на основе Википедии) – много конкретных сущностей
  - Медицинская база знаний: UMLS

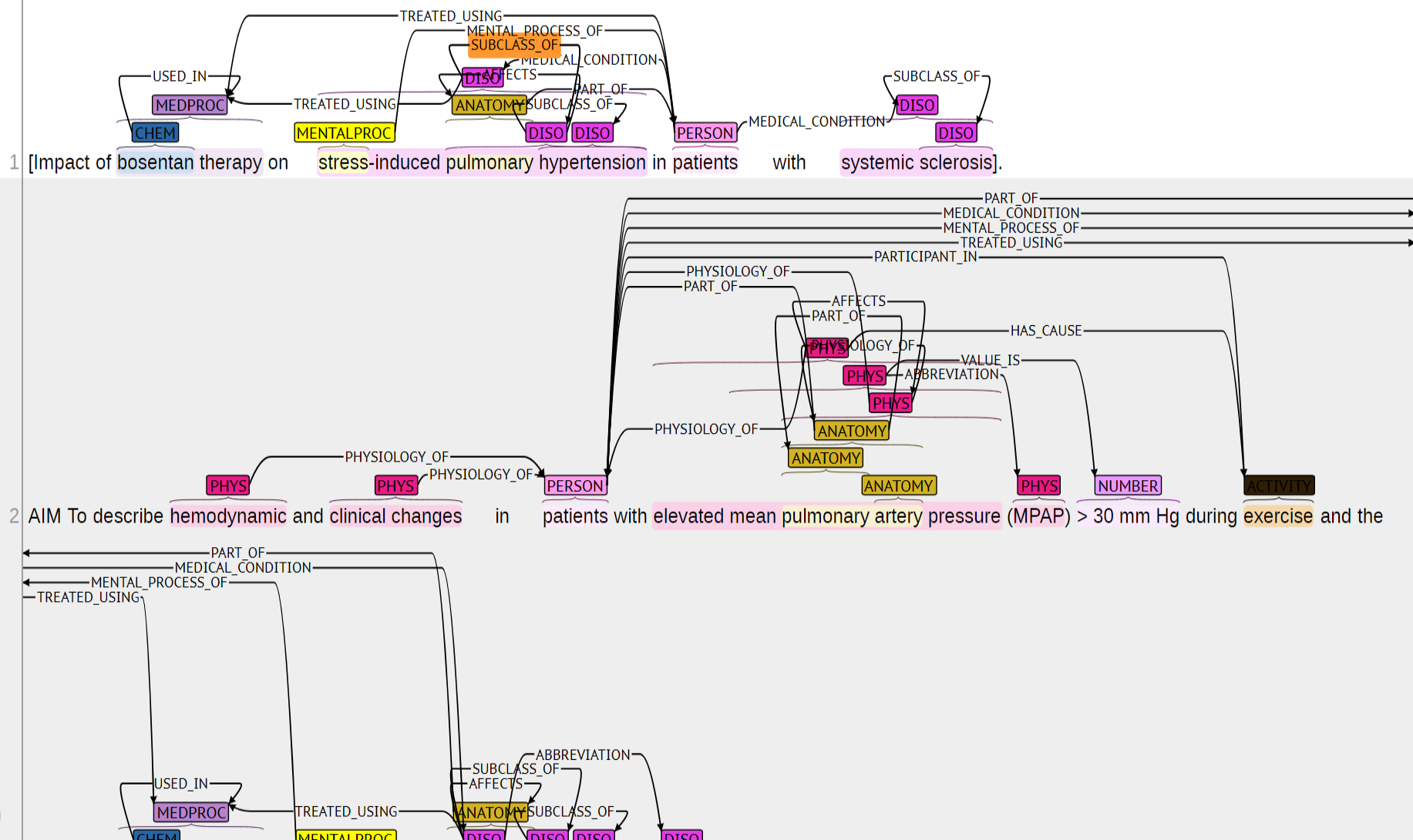


# Новые размеченные набор данных для пополнения графов знаний с тремя уровнями разметки: NEREL and NEREL-BIO

|                            | NEREL                                 | NEREL-BIO                  |
|----------------------------|---------------------------------------|----------------------------|
| Источник                   | Новостные тексты, в основном Wikinews | Pubmed abstracts           |
| Количество текстов         | 1983                                  | 1145                       |
| Количество типов сущностей | 29/100К упоминаний                    | 36/100К упоминаний         |
| Типы отношений             | 49/73 К отношений                     | 35/85К отношений           |
| Связывание с базой знаний  | Wikidata, 17 типов сущностей          | UMLS, 3 types<br>43K links |

(35 типов сущностей, 39 типов отношений)

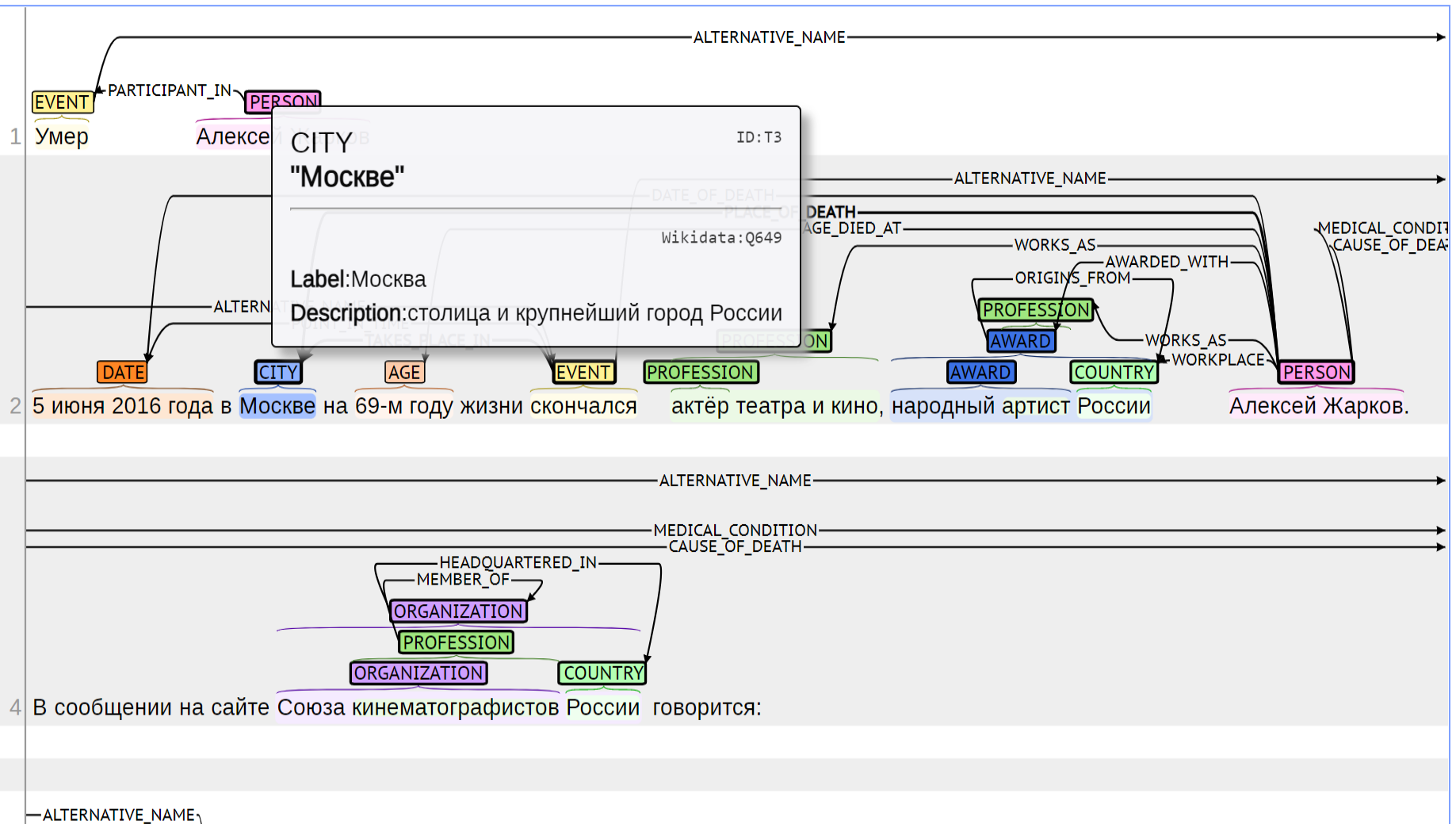
brat



# NEREL

(29 типов сущностей, 49 типов отношений)

 /brat\_user\_data/2021-11\_NEREL\_EL\_data/50texts\_tacred/102661\_text 

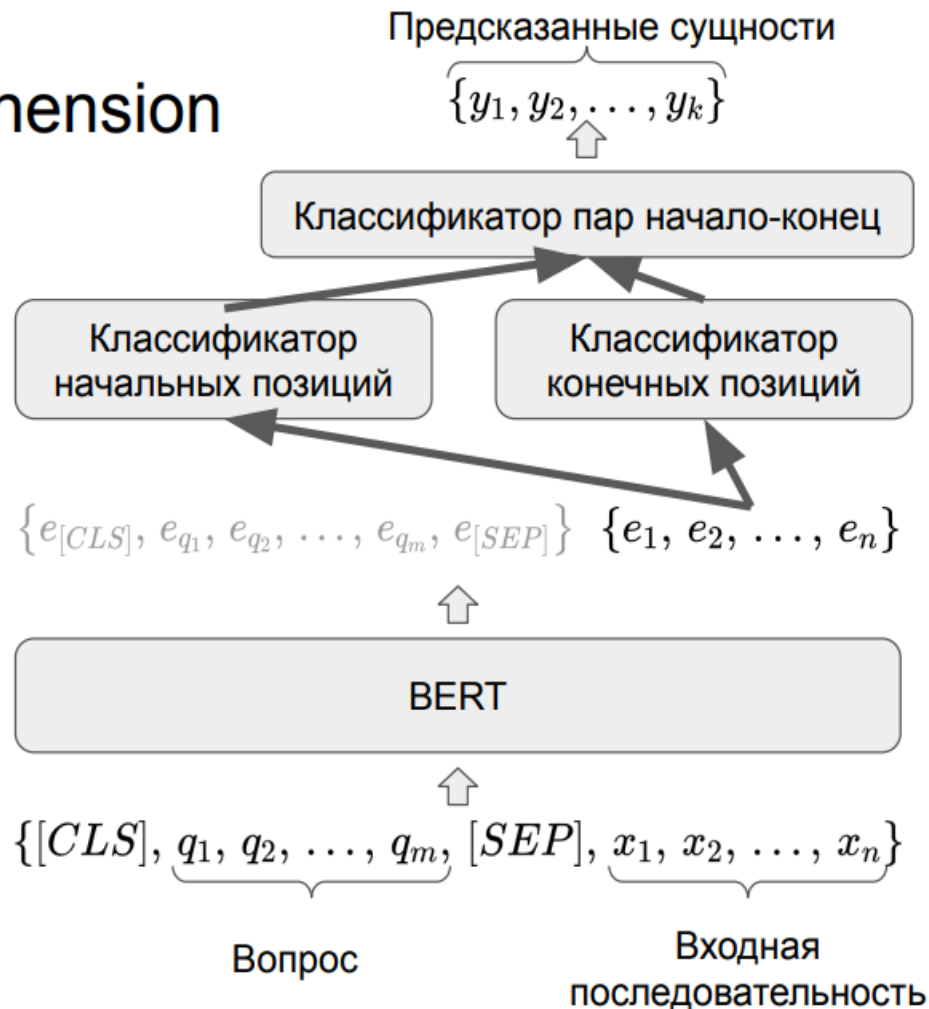


# Извлечение вложенных именованных сущностей

## Machine Reading Comprehension

Модель MRC даёт ответ A на вопрос Q, выбирая его из контекста C. Таким образом, ответом является подпоследовательность слов исходного предложения.

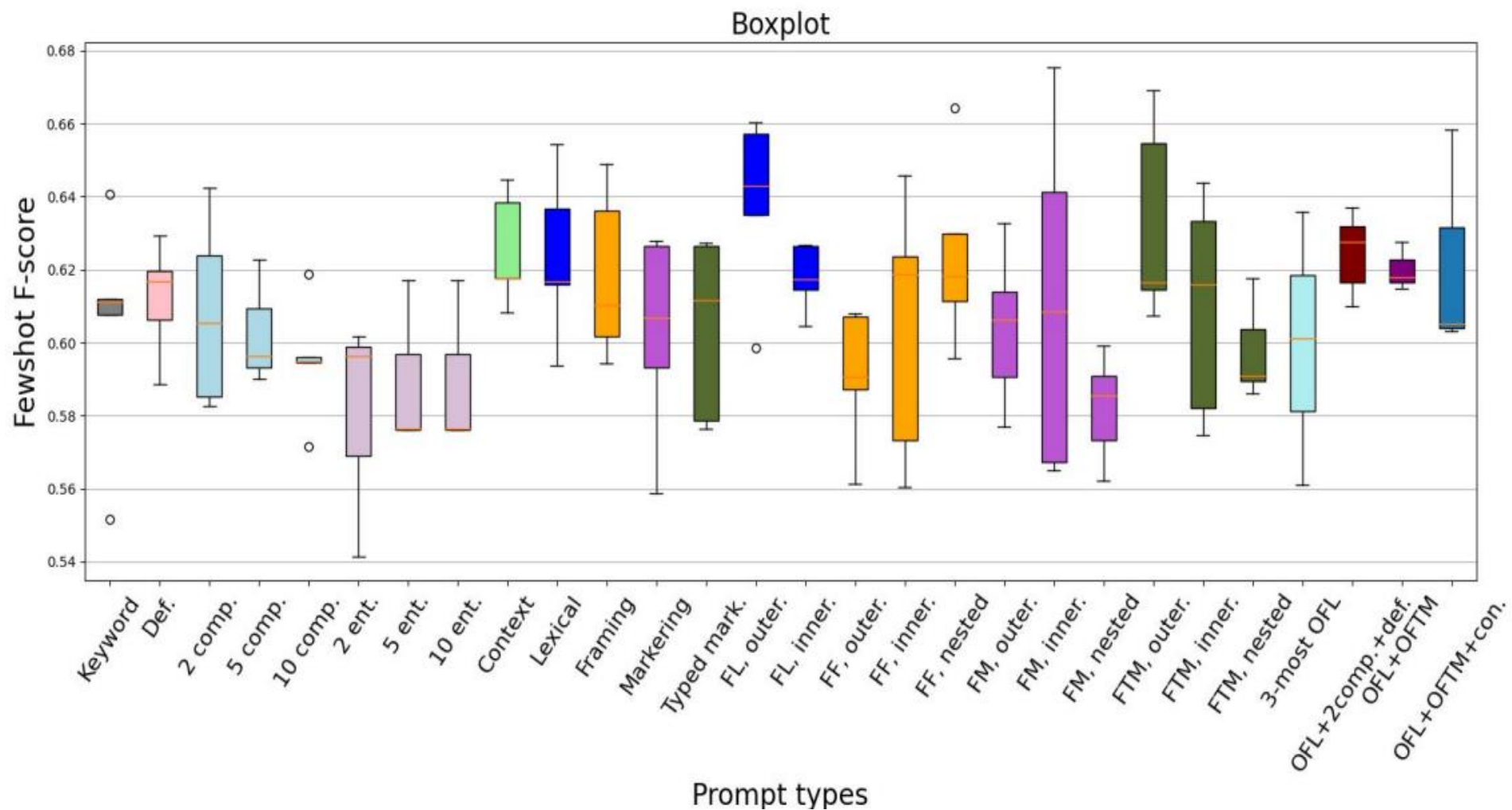
Задача извлечения именованных сущностей преобразуется в задачу ответа на вопросы (подсказки).



# Исследовались различные способы задания вопросов к модели

- Толкования, Ключевые слова, Лексические, Контекстные
  - Один из лучших вариантов полные лексические вопросы (подсказки)
    - В 1964 году Таунсу вместе с Николаем Басовым и Александром Прохоровым была присуждена премия по физике.
- =>
- В 1964 году PERSON вместе с PERSON и PERSON была присуждена Нобелевская премия по физике.

# Результаты тестирования вопросов на задаче обучения на малом числе примеров (few-shot)



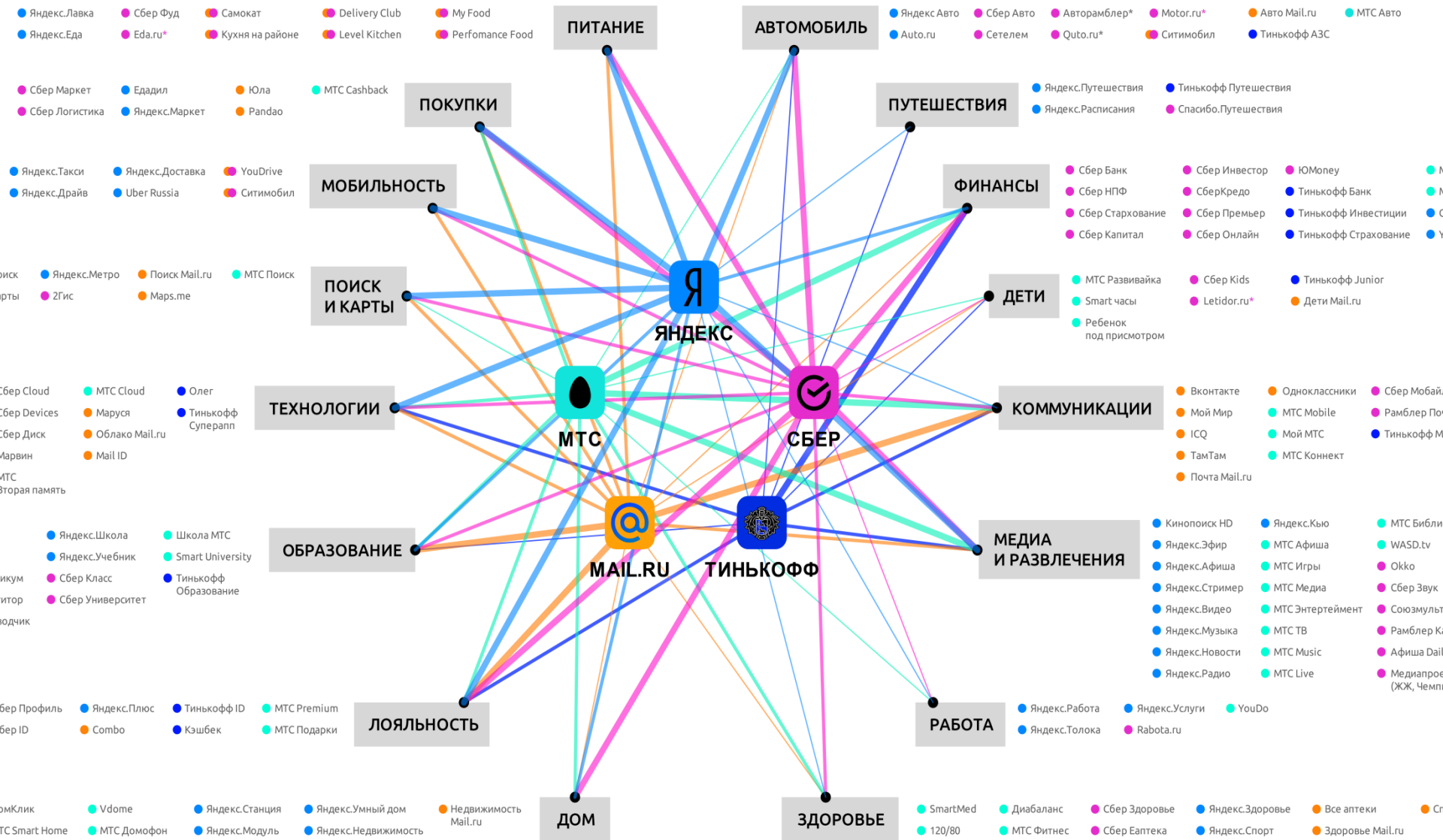
# Модели для извлечения отношений (RE)

| Representation                                                                                                                                                                                           | Nested relations | Model example |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|---------------|
| <b>Position features:</b><br>[People] <sub>1</sub> <u>moved</u> <sub>2</sub> into <sub>3</sub> [downtown] <sub>4</sub><br>For example, relative distances of subject and object from <u>moved</u> [-1,3] | +                | CNN for RE    |
| <b>Entity masking: replacing entities by some masks:</b><br>[CLS][SUBJ-PER] was born in [OBJ-LOC], Michigan..                                                                                            | -                | SpanBert      |
| <b>Entity markers: placing tags on the borders of entities</b><br><E1><E2> Moscow ><E2> State University<br><E1> was found in 1755                                                                       | +                | OpenNRE BERT  |

# Публикации 2022

|        | Тема       |            | +ЛАИР<br>(404) | ЛАИР (404)<br>всего |
|--------|------------|------------|----------------|---------------------|
|        | ЛАИР (404) | ЛАЛС (415) |                |                     |
| Q1     | 1          |            |                | 1                   |
| Scopus | 6          |            | 3              | 9                   |
| РИНЦ   | 1          | 2          | 1              | 2                   |
| Всего  | 11         |            |                | 12                  |

# Цифровые экосистемы



# Яндекс.Метрика

## платформы личных кабинетов lk.msu.ru

- Платформа представляет собой веб-приложение, объединяющее единой инфраструктурой следующие модули:
  - Личный кабинет студента МГУ
  - Личный кабинет преподавателя МГУ
  - Межфакультетские учебные курсы
  - ...

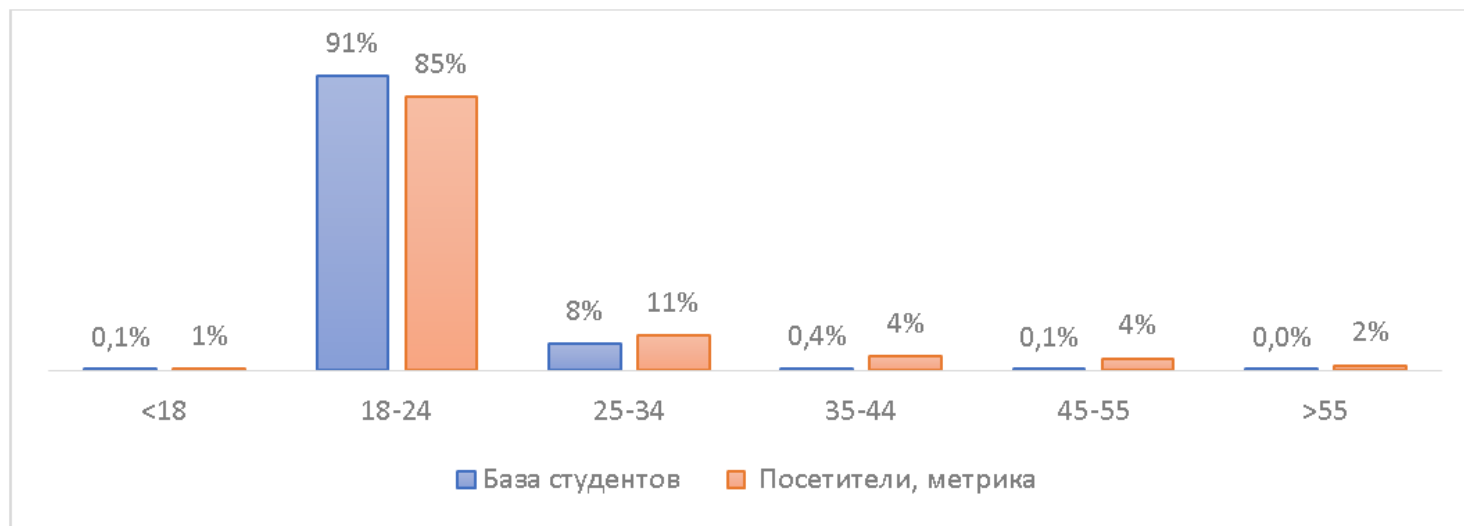
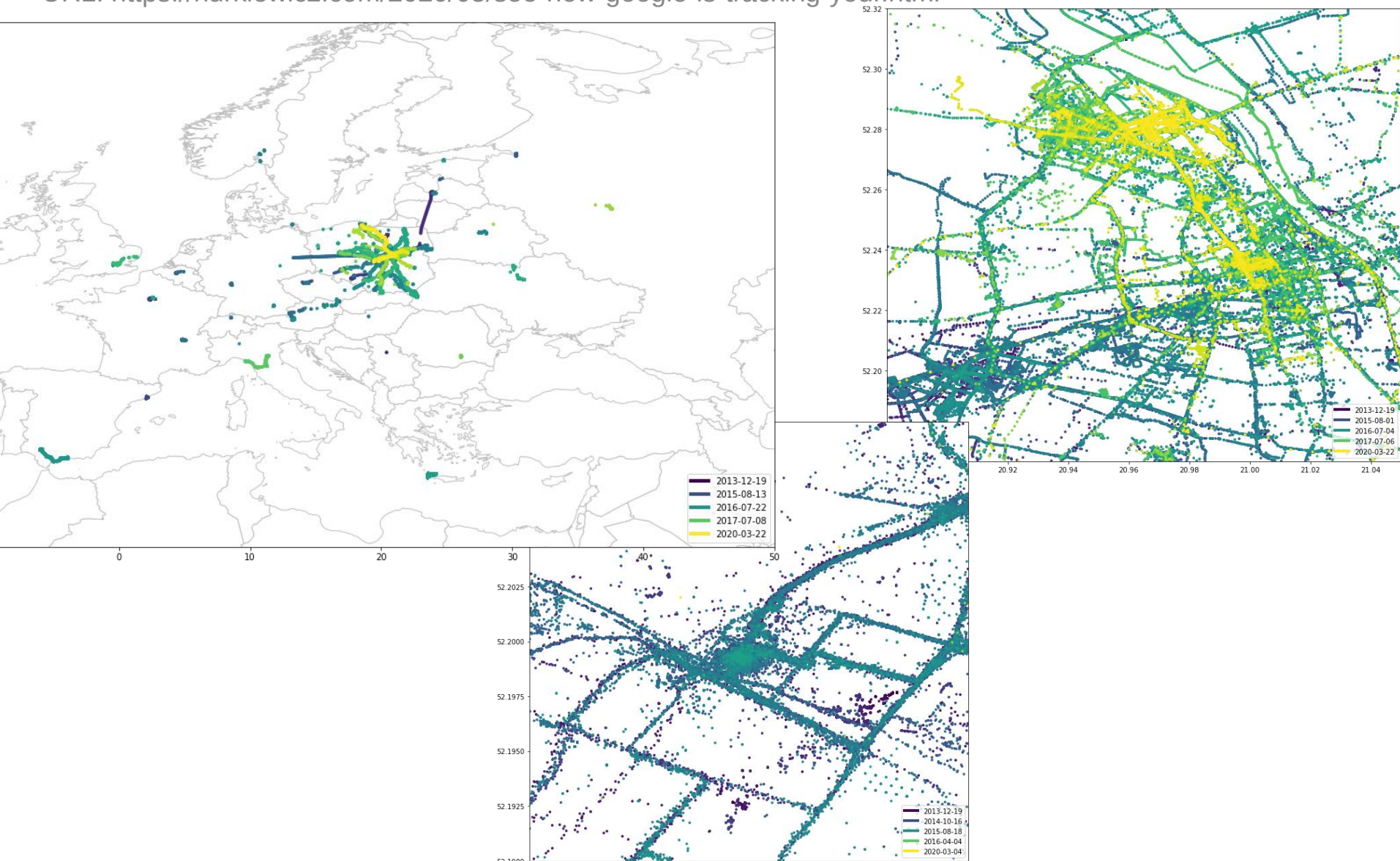


Рис. 6. Сравнение данных Метрики (кабинет студента) по сайту ЛК<sup>8</sup> и данных базы студентов<sup>9</sup>.

# Агрегирование данных пользователя

**Nurkiewicz T.** See how Google is tracking your location. With Python, Jupyter, Pandas, Geopandas and Matplotlib / Tomasz Nurkiewicz Текст : электронный // nurkiewicz.com. – [сайт]. – 2020. – 29 мар. - URL: <https://nurkiewicz.com/2020/03/see-how-google-is-tracking-your.html>



# Планы на 2023 год

**Разработка методов автоматизированного сопровождения больших графов знаний и больших онтологий по потоку разнородных текстов предметной области**

- 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

# Публикации по теме

- **(Scopus)** Daniil Chernyshev, Boris Dobrov, ClusterVote: Automatic Summarization Dataset Construction with Document Clusters //International Conference on Speech and Computer, Lecture Notes in Computer Science, Springer International Publishing AG (Cham, Switzerland), том 13721, с. 99-113. – DOI: [http://dx.doi.org/10.1007/978-3-031-20980-2\\_10](http://dx.doi.org/10.1007/978-3-031-20980-2_10)
- **(Scopus)** Daniil Chernyshev, Boris Dobrov Improving Neural Abstractive Summarization with Reliable Sentence Sampling /D. Chernyshev, B. Dobrov //International Conference on Data Analytics and Management in Data Intensive Domains, Communications in Computer and Information Science, Springer Cham, том 1620, с. 246-261 DOI [http://dx.doi.org/10.1007/978-3-031-12285-9\\_16](http://dx.doi.org/10.1007/978-3-031-12285-9_16) (Scopus)
- **(Scopus)** I. Dimov, B. Dobrov Methods for Automatic Argumentation Structure Prediction //International Conference on Data Analytics and Management in Data Intensive Domains, Communications in Computer and Information Science, Springer Cham, том 1620, с. 223-233. – DOI: [http://dx.doi.org/10.1007/978-3-031-12285-9\\_14](http://dx.doi.org/10.1007/978-3-031-12285-9_14) (Scopus)
- A. Kirillovich, N. Loukachevitch, M. Kulaev, A. Bolshina, D. Ilvovsky Sense-Annotated Corpus for Russian //Proceedings of CLIB-2022 conference, с. 130-136.
- **(Scopus)** /Evgeny Kotelnikov, Natalia Loukachevitch, Irina Nikishina, Alexander Panchenko RuArg-2022: Argument Mining Evaluation //Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022” Moscow, June 15–18, 2022. - Москва: РГГУ, том 21, с. 333-348 DOI: <http://dx.doi.org/10.28995/2075-7182-2022-21-333-348> (Scopus)
- **(Scopus)** Natalia Loukachevitch, Pavel Braslavski, Vladimir Ivanov, et. al. Entity Linking over Nested Named Entities for Russian //Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association Marseille, France, с. 4458-4466. (Scopus)
- **(Q1)** Irina Nikishina, Mikhail Tikhomirov, Varvara Logacheva, et. al. Taxonomy enrichment with text and graph vector representations //SEMANTIC WEB, том 13, № 3, с. 441-475 DOI <http://dx.doi.org/10.3233/sw-212955>. (Q1)
- **(Scopus)** Igor Rozhkov, Natalia Loukachevitch Machine Reading Comprehension Model in RuNNE Competition //Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022” Moscow, June 15–18, 2022/ - Москва: РГГУ, том 21, с. 488-496. DOI: <http://dx.doi.org/10.28995/2075-7182-2022-21-488-496> (Scopus)

## Публикации по теме (2)

- **(415, РИНЦ)** П.В. Гращенков Системы автоматической транскрипции русского текста: общая организация и решенные и нерешенные проблемы //Вестник Московского университета. Серия 9: Филология, издательство Изд-во Моск. ун-та (М.), № 3, с. 9-21. –URL: <https://elibrary.ru/item.asp?id=48866769>
- **(РИНЦ)** Н.В. Лукашевич Автоматический анализ тональности текстов. Проблемы и методы //Интеллектуальные системы. Теория и приложения (ранее: Интеллектуальные системы по 2014, № 2, ISSN 2075-9460), М., том 26, № 1.
- **(415, РИНЦ)** Е.А. Лютикова, А.А. Герасимова, Д.Д. Белова, и др Три аспекта непротиворечивости языковых данных: оценка и интерпретация //Интеллектуальные системы. Теория и приложения. - Том 26, Номер 1, 2022, С.196-202. - URL: <https://elibrary.ru/item.asp?id=49337593>

## +Публикации ЛАИР

- **(РИНЦ)** Калабихина И.Е., Лукашевич Н.В., Банин Е.П., Алибаева Камила. Автоматический анализ репродуктивных ценностей пользователей сети ВКонтакте // Интеллектуальные системы. Теория и приложения, 2022, том 26, № 1, с. 90-96
- **(Scopus)** Alibaeva Kamila, Loukachevitch Natalia. Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022”, с. 8-17 2022
- **(Scopus)** Loukachevitch N., Nevzorova O., Zlochevskaya D. Linking FrameNet to Russian Resources //Communications in Computer and Information Science, том 1625, с. 373-390. 2022
- **(Scopus)** Artemova E., Zmeev M., Loukachevitch N., Rozhkov I., Batura T., Ivanov V., Tutubalina E. RuNNE-2022 Shared Task: Recognizing Nested Named Entities //Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022” Moscow, June 15–18, 2022, место издания Москва: РГГУ, том 21, с. 33-44. 2022