Институт динамики систем и теории управления имени В.М. Матросова,
Сибирское отделение Российской академии наук (ИДСТУ СО РАН)

Семестровая работа

Дородных Никита Олегович

Кандидат технических наук, старший научный сотрудник
Лаборатории 4.2 и Молодёжной лаборатории по ИИ, обработке и анализу данных
ИДСТУ СО РАН

Иркутск, 2025

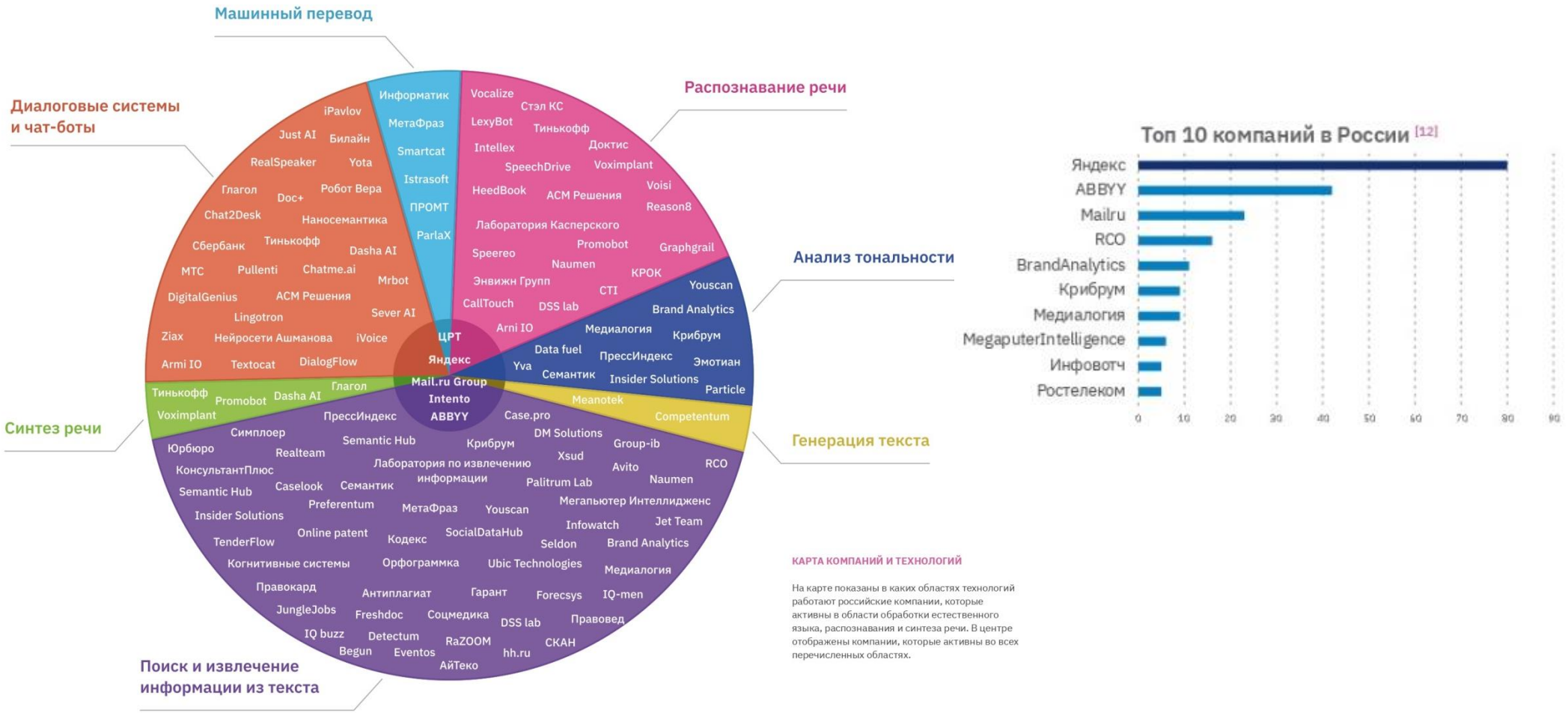
Искусственный интеллект

Искусственный Интеллект (ИИ) — наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ.

Области ИИ:

- Обработка естественного языка (*Natural Language Processing*) → **Извлечение информации**
- Инженерия знаний (*Knowledge Engineering*) → **Разработка баз знаний**
- Машинное обучение (*Machine Learning*) → **Нейронные сети**
- Интеллектуальный анализ данных (*Data Mining* & *Knowledge Discovery*)
- Мультиагентные системы (*Multi-Agent Systems*)
- Нечеткая логика (*Fuzzy Logic*)
- Робототехника (*Robotics*)
- Когнитивное моделирование (*Cognitive Modeling*) и др.

Обработка естественного языка (Natural Language Processing – NLP)



КАРТА КОМПАНИЙ И ТЕХНОЛОГИЙ

На карте показаны в каких областях технологий работают российские компании, которые активны в области обработки естественного языка, распознавания и синтеза речи. В центре отображены компании, которые активны во всех перечисленных областях.

Извлечение информации из текстов

Основное направление в *обработке естественного языка* **(NLP)** – это извлечение информации из текстов (**Information Extraction**).

В качестве извлекаемых из текстов данных обычно выступают:

- **Значимый объект**: *имя персоналии*, *название компании* и пр. для новостных сообщений, *термин предметной области* специального текста, *ссылка на литературу* для научно-технических документов и т. д.
- **Атрибуты объекта**, дополнительно характеризующие его, например, для компании – это *юридический адрес*, *телефон*, *имя руководителя* и т.п.
- **Отношение между объектами**: к примеру, отношение «*быть владельцем*» связывает компанию и персону-владельца, «*быть частью*» соединяет факультет и университет.
- **Событие/факт**, связывающее несколько объектов, например, событие «*прошла встреча*» включает *участников встречи*, а также *место* и *время* ее проведения.

Основные задачи извлечения информации из текстов

Согласно видам извлекаемой информации общая задача извлечения информации из текстов включает следующие основные подзадачи:

- **Распознавание и извлечение именованных сущностей (*named entities*):** А.П. Чехов, Нижний Тагил, ПКО «Картография» и т.п.;
- **Выделение атрибутов (*attributes*) объектов и семантических отношений (*relations*) между ними:** даты рождения персоны, отношения «работать в» и т.д.;
- **Извлечение фактов и событий (*events*)**, охватывающих несколько их параметров (*атрибутов*), например, событие «*кораблекрушение*» с атрибутами дата, время, место и т.п.

Задача **извлечения (*распознавания*) именованных сущностей** считается самой проработанной из области извлечения информации.

Распознавание именованных сущностей (Named Entity Recognition – NER)

Распознавание именованных сущностей (Named Entity Recognition – NER) – это задача, направленная на поиск и классификацию значимых фрагментов текста по заранее определенным категориям. Эти категории могут включать: *имена людей*, *названия организаций*, *места*, *даты*, *денежные суммы* и многое другое.

Цель NER – это выделить из неструктурированного текста ключевую информацию, идентифицировав такие сущности и определив их тип.

Популярными примерами реализации **NER** являются следующие программы (*библиотеки*):

- **Stanford CoreNLP** [*https://stanfordnlp.github.io/CoreNLP/*].
- **Apache OpenNLP** [*https://opennlp.apache.org/*] и **AllenNLP** [*https://allennlp.org/*].
- **spaCy** [*https://spacy.io/*].
- Российским решением в данной области можно назвать библиотеку с открытым исходным кодом – **DeepPavlov** [*https://deeppavlov.ai/*].

Связывание именованных сущностей (Named Entity Linking – NEL)

Логическим продолжением **NER** является задача – **связывания именованных сущностей (*Named Entity Linking, NEL*)**, представляющая собой поиск соответствий отдельных слов (*распознанных сущностей*) по тексту с конкретными сущностями (*экземплярами*) из базы знаний, графа знаний или онтологии (например, **DBpedia**, **Wikidata**, **YAGO**). Результатом данного процесса является текст с аннотациями (*референтными сущностями*), что позволяет:

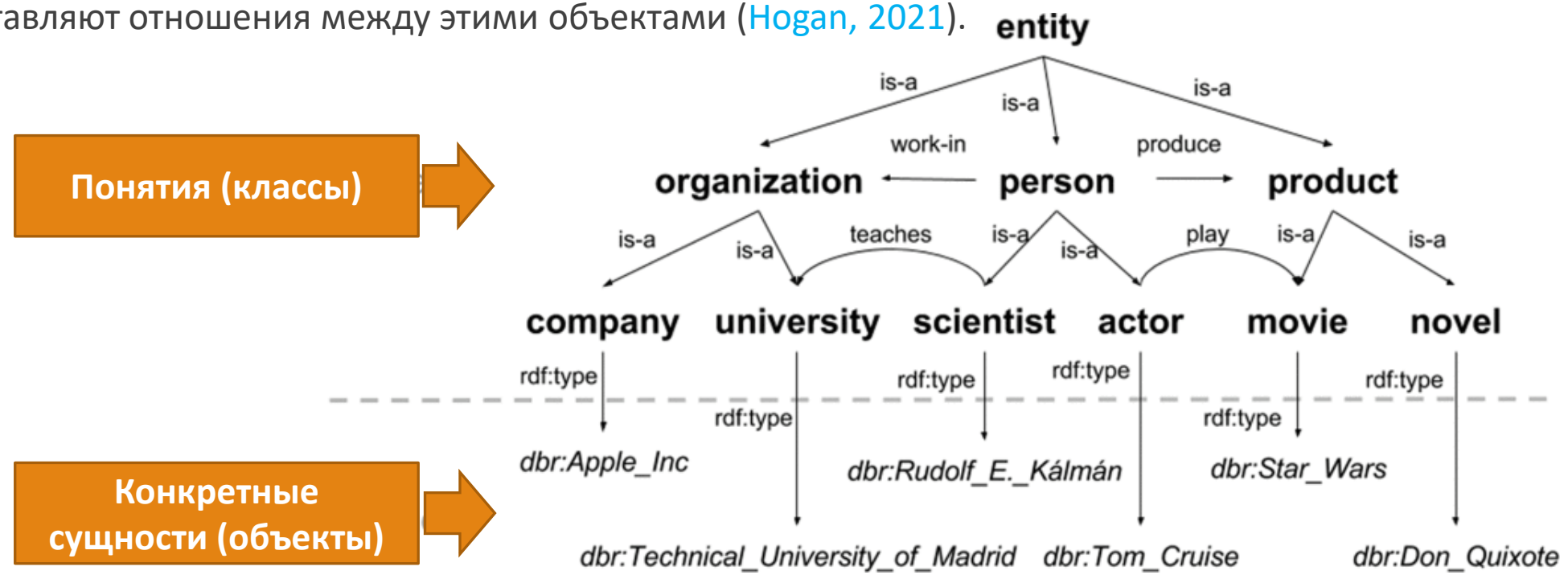
- Осуществлять более эффективный поиск (*Information Retrieval*).
- Применять эти размеченные данные в вопросно-ответных (*Question Answering*) и рекомендательных системах.
- Осуществлять автоматическое заполнение и обогащение существующих баз знаний новыми извлеченными фактами (*Knowledge Base Population*).
- Проводить интеллектуальный анализ содержания (*Content Analysis*) и т.д.

Широко распространенными открытыми средствами для решения задачи **NEL** являются: **AIDA**, **DBpedia Spotlight**, **Dexter**, **Babelfy**, **DoSeR**, **Yahoo FEL**.

Графы знаний

Одним из видов баз знаний является – **граф знаний**.

Граф знаний (knowledge graph) – это граф данных, предназначенный для накопления и передачи знаний о реальном мире, узлы которого представляют интересующие объекты, а ребра представляют отношения между этими объектами (Hogan, 2021).



(Hogan, 2021) *Hogan et al., Knowledge Graphs. ACM Computing Surveys, 54(4), 1–37 (2021).*

Таблицы – Источник больших данных

Часть дневника Мерера
(около 2600 г. до н. э.)

1. Примерно до **40%** всех таблиц расположенных в **Вебе**, обладают реляционной природой и содержат потенциально полезные факты (**Peeters, 2024**).
2. Русскоязычная версия Википедии 2021 года содержит ~ **1.4 млн.** таблиц (**Fedorov, 2023**).
3. Таблицы в формате **Google Sheets** используют около **2 млрд.** пользователей ежемесячно (**1**).
4. **Microsoft Excel** имеет от **750 млн.** до **1,2 млрд.** ежемесячных пользователей по всему миру (**1**).

(Peeters, 2024) *Peeters R., et al. The Web Data Commons Schema.org Table Corpora. Proc. the ACM Web Conference, 1079-1082 (2024).*
(Fedorov, 2023) *Fedorov P.E. et al. RWT: A Pub. Corpus of Web Tables for Rus. Lang. Based on Wiki.. Lob. Jour. of Math., 44, 111-122 (2023).*
(1) *https://askwonder.com/research/number-google-sheets-users-worldwide-eoskdoxav*

Задание для семестровой работы (1)

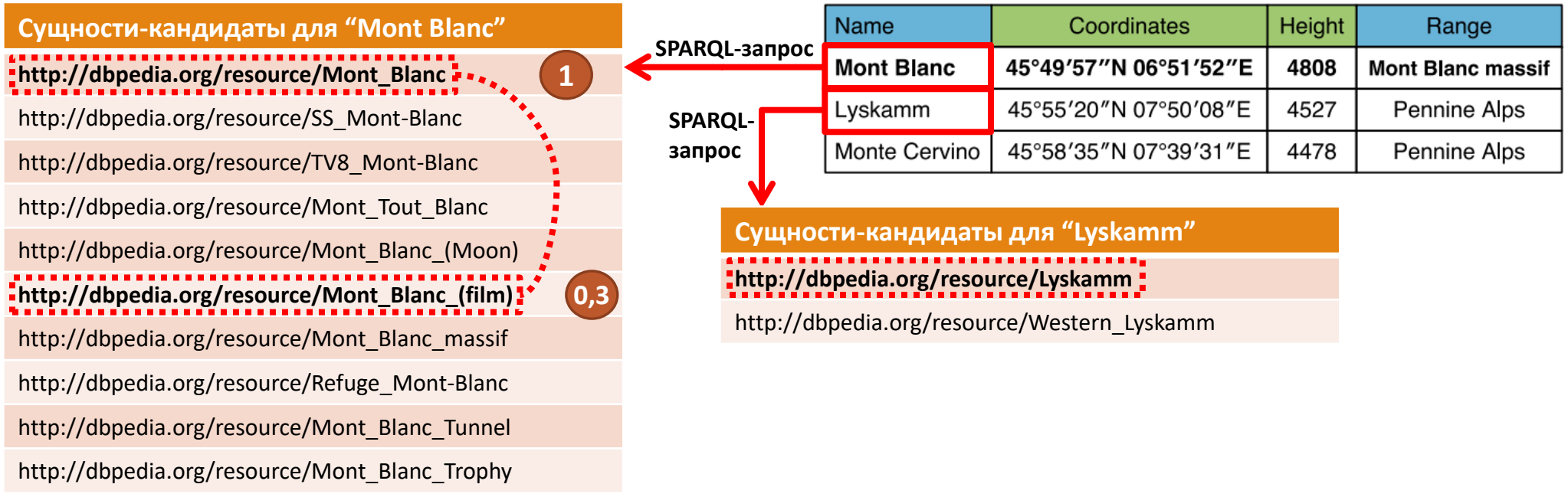
1. Решение задачи **NER** для ***табличных данных***, включая:

- программную реализацию выполнения процедуры **NER** на основе *стандартных, специализированных **NLP**-библиотек*;
- программную реализацию на основе *больших языковых моделей* с использованием *промптинга* (как **zero-shot**, так и **few-shot** настройки).

2. Решение задачи **NEL** для ***табличных данных*** на основе использования графов знаний общего назначения (**DBpedia** или **Wikidata**). Программная реализация может использовать как *стандартные библиотеки* для решения **NEL**-задачи (например, **DBpedia Spotlight** или **DBpedia Lookup**), так и *большие языковые модели*.

Связывание сущностей для ячеек таблиц

Поиск ***сущностей-кандидатов*** в графе знаний для каждого значения ячейки столбца и ***снятие неоднозначности*** (например, использовать *SPARQL-запрос* и *метрику расстояния Левенштейна* или использовать специализированное *средство поиска сущностей DBpedia Lookup*).



Задание для семестровой работы (2)

3. Создание небольшого **тестового набора табличных данных** путем ручной разметки (*аннотирования*) ячеек таблиц определенными **NER**-категориями (*метками*).

4. Разработка скриптов **получения экспериментальной оценки** производительности решения задачи **NER** (*для сравнения обоих вариантов*) с использованием созданного тестового набора табличных данных и стандартных метрик оценки качества таких как: ***точность*** (*precision*), ***полнота*** (*recall*) и ***F-мера*** (*F1*).

Основные требования к заданию (1)

1. В качестве исходных наборов табличных данных предлагается использовать:

- русскоязычный набор таблиц – **RF-200 (ru-facts-200)** [*https://github.com/YRL-AIDA/ru-facts-200*];
- англоязычные наборы таблиц такие как: **T2Dv2** [*https://webdatacommons.org/webtables/goldstandardV2.html*] и **Tough Tables (2T)** [*https://vcutrona.github.io/publication/2t/*].

2. Создание размеченного набора табличных данных для тестирования задачи **NER** должно осуществляться на основе выборки некоторых таблиц из приведенных выше наборов. Предполагается не менее **10** проверочных таблиц (*точное количество уточняется с преподавателем*).

Основные требования к заданию (2)

3. Программная реализация должна быть выполнена с использованием языка **Python** и библиотек **PyTorch**, **Transformers** и **LangChain** (*для работы с языковыми моделями*).

4. Разработанная программа должна обеспечивать **консольное взаимодействие** с пользователем и сохранение всех результатов обработки в виде отдельных **json-файлов** (*дополнительно к отображению результатов на консоль*).

5. Разработанная программа должна включать документацию (*подробное описание программы + руководство пользователя*).

Институт динамики систем и теории управления имени В.М. Матросова,
Сибирское отделение Российской академии наук (ИДСТУ СО РАН)

Семестровая работа

Спасибо за внимание!

Дородных Никита Олегович

Телефон: **+7 950 104-99-45**
Эл. почта: DorodnyxNikita@gmail.com