УДК 004.89

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Анохин И.И. (ИТМО)
Научный руководитель – кандидат технических наук, Балакшин П.В.
(ИТМО)

Введение. Распознавание именованных сущностей (Named Entity Recognition, NER) [1] является важной задачей обработки естественного языка (Natural Language Processing, NLP). В неё включается идентификация и категоризация сущностей в тексте: имена людей, названия организаций, географические места, даты, числа и другие специфические объекты. Полученные в результате обработке сущности используются как базовые блоки для дальнейших NLP задач, таких как поиск информации, извлечение информации, автоматическое реферирование, вопросно-ответные системы. Важную роль распознавание именованных сущностей может иметь и в задачах анализа резюме кандидатов.

Основная часть. На данный момент существует большое количество решений для извлечения именованных сущностей в англоязычных текстах. Однако для русских текстов было создано не так много хороших решений с открытым исходным кодом [2]. В последние годы популярным решением стал набор инструментов Natasha, однако он использует внутри готовые словари для категоризаций и наборы правил, основанные на регулярные выражениях, для извлечения сущностей. Использование в этих задачах больших лингвистических моделей имеет ряд преимуществ:

1) Большие языковые модели легко адаптируются к различным областям.

2) Большие языковые модели превосходят по производительности традиционные системы NER, основанных на правилах [3].

3) Большие языковые модели являются контекстно ориентированными, благодаря чему могут лучше воспринимать связи между разными предложениями.

4) Отсутствие необходимости писать правила с помощью регулярных выражений делает их более простыми в использовании.

Выводы. Проведен анализ возможности использования больших языковых моделей для анализа именованных сущностей в не структурированном тексте, получены метрики и результаты.

Список использованных источников:

1. Named Entity Recognition with LLMs — Extract Conversation Metadata [Электронный ресурс] – Режим доступа: https://medium.com/@grisanti.isidoro/named-entity-recognition-with-llms-extract-conversation-metadata-94d5536178f2. Дата обращения: 01.02.2024.

2. Yargy-парсер и библиотека Natasha. Извлечения структурированной информации из текстов на русском языке [Электронный ресурс]. – Режим доступа: https://habr.com/ru/articles/349864. Дата обращения: 01.02.2024.

3. Using ChatGPT to Pre-annotate Named Entities Recognition Labeling Tasks [Электронный ресурс]. – Режим доступа: https://kili-technology.com/data-labeling/machine-learning/using-chatgpt-to-pre-annotate-named-entities-recognition-labeling-tasks#what-are-the-benefits-of-using-a-large-language-model-for-named-entity-recognition. Дата обращения: 01.02.2024.