

0.1. Мазин Д., Латушко А.П. Разработка методов предварительной разметки для извлечения синонимов и гиперонимов в русскоязычных текстах

Современные задачи обработки естественного языка часто решаются с помощью больших языковых моделей, требующих больших объемов данных [1], размеченных в соответствии с требованиями задачи [2]. Предварительная автоматизированная разметка может снизить трудоемкость и повысить эффективность ручного аннотирования большого объема данных, предоставляя основу для дальнейшей ручной корректировки.

В данной работе были разработаны и протестированы методы автоматической разметки для извлечения синонимов и гиперонимов из текстов аннотаций статей "Википедии" на русском языке. Первый разработанный метод основан на формальных правилах. В результате его использования была достигнута метрика F1 для синонимов — 0.51, для гиперонимов — 0.39 на выборке из 1 тыс. текстов.

Следующий этап исследования заключался в использовании больших языковых моделей (LLMs): Mistral из библиотеки GPT4All (F1 для синонимов — 0.4, для гиперонимов — 0.14), YandexGPT (F1 для синонимов — 0.37, для гиперонимов — 0.19) и GigaChat (F1 для синонимов — 0.68, для гиперонимов — 0.33). При объединении двух наиболее доступных методов с возможностью локального запуска (разметки формальных правил и модели Mistral) были получены следующие метрики: F1 для синонимов — 0.58, для гиперонимов — 0.38.

Таким образом, предварительная разметка, полученная с использованием разработанных методов, существенно снижает трудозатраты разметчика, сводя его задачу к корректировке результатов автоматизированной разметки. В зависимости от уровня формализации задачи могут применяться как формальные правила, так и языковые модели, результаты которых сопоставимы по качеству. Данное исследование демонстрирует перспективы автоматической разметки в NLP-задачах и намечает пути дальнейшего усовершенствования методов.

Научный руководитель — канд. техн. наук Бручес Е. П.

Список литературы

- [1] BROWN T., MANN B. Language models are few-shot learners // Proc. Intern. Conf. «Conference and Workshop on Neural Information Processing Systems». Vancouver: Curran Associates, Inc., 2020. Vol. 33. P. 1877–1901.
- [2] VILLALOBOS P., HO A. Will we run out of data? Limits of LLM scaling based on human-generated data // Machine Learning (cs.LG). arXiv. 2022. DOI:arXiv.2211.04325.