УДК 004.8

ГИБРИДНЫЙ ВОПРОСНО-ОТВЕТНЫЙ ПОИСК С ИСПОЛЬЗОВАНИЕМ НЕРАЗМЕЧЕННЫХ И СТРУКТУРИРОВАННЫХ ДАННЫХ

Ефимов П.В. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., доцент Муромцев Д.И.
(Университет ИТМО, г. Санкт-Петербург)

Введение. В условиях непрерывного увеличения объемов информации растет потребность в оперативном удовлетворении информационных потребностей пользователей. На смену традиционным поисковым системам, возвращающим список ранжированных документов, приходят вопросно-ответные системы, от которых пользователь ожидает получить краткий ответ на свои запросы. Подходы к современному вопросно-ответному поиску разделяются на две группы: поиск в текстовой коллекции (Open-Domain Question Answering, ODQA) и поиск в базе знаний (Knowledge Base Question Answering, KBQA). При этом на заре развития вопросно-ответных систем предлагались методы, объединяющие данные группы методов [1, 2]. В настоящий настоящий момент наблюдаются попытки развития методов объяснимого искусственного интеллекта (Explainable AI), а также потребность в построении многоязычных систем и заполнении пропусков в структурированных данных (в том числе на разных языках). В связи с этим в данной работе предлагается рассмотреть возможность построения гибридной вопросно-ответной системы с использованием современных методов глубокого обучения для обработки естественного языка.

Основная часть. Существуют различные подходы к комбинированию вопросно-ответного поиска в текстовой коллекции и графе знаний. В [1] предлагалось генерировать кандидатов на ответ из текстовой коллекции, а затем ранжировать их при помощи базы знаний. Тем временем в [2] кандидаты из базы знаний ранжировались на основе данных из поисковой выдачи. Появление новых наборов данных для оценки вопросно-ответных систем, которые содержат как текстовую разметку, так и ссылки на таблицы или базы знаний, говорит о росте интереса к гибридному вопросно-ответному поиску [3—6].

В настоящий момент в задаче вопросно-ответного поиска в текстовой коллекции были получены наилучшие результаты при помощи предобученных моделей на базе архитектуры «Трансформер». Такой поиск состоит из двух этапов: поиск документов или фрагментов текста (retriever) и выделение ответа в найденных фрагментах текста (reader). На первом этапе может использоваться поиск с использованием разреженных векторов (BM25) или плотных векторных представлений (Dense Passage Retrieval, DPR). На этапе выделения ответа возможно использование одной из двух моделей: BERT на базе кодировщика из архитектуры «Трансформер» и seq2seq модель T5. Первая модель предсказывает границы ответа в тексте, а вторая — генерирует ответы, используя информацию из разных фрагментов текста. Обе модели для выделения ответа в тексте генерируют различных кандидатов на ответ, неявно определяя тип вопроса и класс потенциальной сущности в качестве ответа. Существующие базы знаний могут помочь переранжировать кандидатов, используя сущности и отношения выделенные в тексте вопроса.

Также возможны другие применения базы знаний в гибридной вопросно-ответной системе. Например, используя методы оценки неопределенности вопросно-ответная система может выбирать, какую подсистему использовать для ответа на вопрос: на основании текстовой коллекции или базы знаний. Ещё одно применение базы знаний — получение многоязычных ответов при помощи наименований сущностей на разных языках.

Выводы. В данном докладе было описано, как современный вопросно-ответный поиск в большой текстовой коллекции может быть улучшен и расширен при помощи структурированных данных из базы знаний. Объединив подходы для разных источников информации, ожидается получение новых результатов на современных наборах данных [4—6].

Список использованных источников:

1. Ferrucci D. et al. Building Watson: An overview of the DeepQA project //AI magazine. – 2010. – Т. 31. – №. 3. – С. 59-79.
2. Savenkov D., Agichtein E. When a knowledge base is not enough: Question answering over knowledge bases with external text data //Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. – 2016. – С. 235-244.
3. Chen W. et al. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data //Findings of the Association for Computational Linguistics: EMNLP 2020. – 2020. – С. 1026-1036.
4. Rybin I. et al. RuBQ 2.0: an innovated Russian question answering dataset //The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18. – Springer International Publishing, 2021. – С. 532-547.
5. Longpre S., Lu Y., Daiber J. MKQA: A linguistically diverse benchmark for multilingual open domain question answering //Transactions of the Association for Computational Linguistics. – 2021. – Т. 9. – С. 1389-1406.
6. Sen P., Aji A. F., Saffari A. Mintaka: A Complex, Natural, and Multilingual Dataset for End-to-End Question Answering //Proceedings of the 29th International Conference on Computational Linguistics. – 2022. – С. 1604-1619.