УДК 004

Применение RAG-систем для анализа данных диалога

Блохина А.Ю., Федоров Д.А. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Федоров Д.А.
(ИТМО)

Введение. Видеозапись совещаний, интервью и переговоров становится распространенной практикой, в результате чего объем такого контента постоянно увеличивается. Извлечение информации из этих массивов данных традиционными методами – ручным протоколированием или полным просмотром – требует значительных временных затрат. Технология Retrieval-Augmented Generation (RAG), комбинирующая генеративные способности больших языковых моделей и механизм семантического поиска по внешним источникам данных [1, 2], предлагает подход к созданию вопросно-ответных систем, способных работать с подобным контентом. Однако стандартные RAG-решения ориентированы преимущественно на монологический текст и не учитывают специфику диалоговых данных.

Основная часть. Диалоговые данные, к которым относятся записи интервью, совещаний и переговоров, обладают рядом особенностей. Информация в них передается через реплики участников, которые могут перебивать друг друга, возвращаться к ранее обсуждавшимся темам, использовать отсылки к сказанному ранее, что делает неприменимыми стандартные подходы к сегментации и индексированию, разработанные для монологических текстов. Для обработки таких данных предлагается RAG-архитектура, в которой большие языковые модели решают задачи сегментации диалога на смысловые блоки и генерации ответов по запросам пользователя. В состав архитектуры включены модуль транскрибации и дизаризации аудиопотока, сегментации диалога на смысловые блоки с использованием больших языковых моделей, модуль векторного индексирования и семантического поиска, а также генерации ответов. Ключевым элементом является модуль сегментации, поскольку качество разбиения диалога напрямую влияет на эффективность последующего поиска [3]. Для определения границ смысловых блоков используется большая языковая модель, что позволяет учитывать не только формальные признаки, но и семантические связи между репликами. Полученные результаты демонстрируют, что учет диалогической структуры при сегментации позволяет повысить полноту извлечения релевантной информации по сравнению с традиционными подходами.

Выводы. В работе представлена архитектура RAG-системы для анализа данных диалога. Предложенное решение обеспечивает повышение полноты и точности поиска релевантной информации по сравнению с традиционными подходами. Полученные результаты могут быть применены при разработке интеллектуальных вопросно-ответных сервисов на базе больших языковых моделей для работы с архивами видеозаписей совещаний и интервью.

Список использованных источников:

1. Gao Y. et al. Retrieval-augmented generation for large language models: A survey //arXiv preprint arXiv:2312.10997. – 2023.
2. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks//Advances in neural information processing systems. – 2020. – Т. 33. – С. 9459-9474.
3. Setty S. et al. Improving retrieval for rag based question answering models on financial documents //arXiv preprint arXiv:2404.07221. – 2024