УДК 004.89

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕТОДА RETRIEVAL AUGMENTED GENERATION (RAG) ДЛЯ ПЕРСОНИФИКАЦИИ ДИАЛОГОВОГО АГЕНТА.

Жердева А.Х. (Университет ИТМО)
Научный руководитель – кандидат технических наук, доцент Махныткина О.В.
(Университет ИТМО)

Введение. Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №623088 «Разработка русскоязычного персонифицированного эмоционального диалогового агента». При разработке диалоговых агентов использование глубокого обучения позволяет эффективно использовать большие объемы данных для выделения значимых характеристик и стратегий генерации ответов, минимизируя необходимость в ручной обработке. Для придания диалоговому агенту определенных личностных черт необходимо, чтобы его диалоги были похожи на человеческие. Стоит отметить, что создание агентов с определенной личностной характеристикой является важным аспектом в области развития искусственного интеллекта, так как это может значительно повысить удовлетворенность пользователей и улучшить качество взаимодействия с диалоговыми агентами. [1]

Основная часть. В данной работе рассмотрен метод Retrieval Augmented Generation (RAG) [2] для персонификации диалогового агента. Метод RAG представляет собой комбинацию генеративной и поисковой модели, что позволяет улучшить качество генерации ответов. Для проведения экспериментов была выбрана Large Language Model (LLM) saiga2_7b_lora, а в качестве данных – датасет на русском языке Toloka RuPersonaChat, который содержит диалоговые данные с описанием персоны. Поисковый компонент преобразует входной текст в вектор с использованием кодировщика запроса и сохраняет закодированные документы в виде поискового индекса, далее осуществляется поиск векторов документов в индексе, связанных с входным вектором. После этого поисковый компонент возвращает текстовое представление найденных документов [3]. Генератор в виде LLM принимает текст пользователя и соответствующие ему документы, формулирует запрос и предлагает языковой модели сформулировать ответ на вопрос пользователя, учитывая информацию из документов. Документы представляют собой список фактов о персоне и истории диалогов с использованием этих фактов. Реализация RAG для персонификации позволяет создавать более релевантные ответы на запросы пользователей. Были сделаны тест-кейсы для оценки качества модели, где были представлены вопросы и ответы модели с разными промтами: с использованием и без использования метода RAG. Оценка качества проводилась вручную по методике SSA.

Выводы. В данной работе был проведен сравнительный анализ LLM saiga2_7b_lora, ее модификации с применением метода RAG и промт-тюнинга для персонификации диалогового агента. Эксперименты показали, использование RAG привело к улучшению качества генерации персонифицированных ответов диалогового агента.

Список использованных источников:

1. Zhang S., Dinan E., Urbanek J., Szlam A., Kiela D., Weston J. Personalizing Dialogue Agents: I have a dog, do you have pets too? // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. – 2018. - vol 1. – p. 2204-2213
2. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. // Advances in Neural Information Processing Systems. – 2020. – vol 33. – p. 9459–9474
3. Abil H., Marjan G., and Luke Z. Simple and effective retrieve-edit-rerank text generation. // 58th Annual Meeting of the Association for Computational Linguistics. – 2020. - p. 2532–2538