УДК 004.8

ТЕХНОЛОГИЧЕСКИЙ ПОДХОД К ИЗВЛЕЧЕНИЮ КЛЮЧЕВЫХ СВЕДЕНИЙ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ С ПРИМЕНЕНИЕМ МОДЕЛИ RAG

Жменько А.Ю. (ТюмГУ)

Научный руководитель – кандидат технических наук, Воробьева М.С. (ТюмГУ)

Введение. В современном мире объем текстовой информации стремительно растет, что делает процессы ее анализа и извлечения ключевых сведений (суммаризации) более актуальными. Такие технологии востребованы в юриспруденции, медицине, государственном управлении и других сферах, где важно быстро получать сжатую и информативную версию больших текстов. Для решения такой проблемы можно использовать большие языковые модели (LLM), способные генерировать краткое изложение текста [1]. Однако их основным ограничением является размер контекстного окна — слишком большой объём текста, содержащийся в документе, не может быть обработан моделью целиком, что приводит к потере важной информации, а также недостаточной релевантности результатов. Дополнением такого решения является подход Retrieval-Augmented Generation (RAG), который сочетает поиск семантически похожих на вопрос фрагментов текста с возможностями генеративной модели [2]. Это позволяет обрабатывать документы любого объема, выбирая ключевые части для передачи в LLM, а затем формировать осмысленное резюме, учитывая контекст документов. Зарубежные исследования демонстрируют эффективность RAG в повышении производительности обработки запросов и качества ответов на вопросы, требующих наукоёмких знаний [3], также указывается повышение точности и релевантности ответов на вопросы, специфичные для предметной области, с использованием документов [4].

Основная часть. Сервис суммаризации документов в качестве входных данных принимает текстовый документ, а также вопрос, на который необходимо ответить в контексте документа. На выходе формируется ответ, содержащий краткое резюме релевантных фрагментов документа с учетом запроса. Архитектура сервиса с использованием модели RAG состоит из нескольких частей:

1) Предварительная обработка текста: исходный документ разбивается на семантически целостные сегменты.
2) Векторизация сегментов и запроса: для каждого сегмента и запроса пользователя вычисляется эмбеддинг (векторное представление) с помощью модели Sentence-BERT (sbert_large_nlu_ru).
3) Поиск релевантных фрагментов: при помощи косинусного сходства определяются сегменты, семантически наиболее соответствующие запросу.
4) Составление промпта: из вопроса пользователя, выбранных из предыдущего шага сегментов текста и инструкции (указанием, в каком виде нужно дать ответ и какую информацию для этого использовать) формируется запрос для модели LLM.
5) Генерация ответа: собранный промпт отправляется в генеративную модель Llama 3.1 8B с квантизацией Q4_K_M, возвращается ответ на вопрос с использованием контекста в промпте.

Для оценки качества работы модели проводился анализ семантической схожести ответов операторов и модели на шаблонные вопросы, связанные с разделом кадастрового учёта недвижимости в документах многофункциональных центров (МФЦ). В качестве основы для тестирования использовались 10 сложно-структурированных документов, каждый из которых состоит из 46-91 сегментов с разделением на блоки, содержащие заголовок и текст. Для измерения степени соответствия эталонным ответам оператора была выбрана метрика BERTScore, так как измеряет степень схожести текстов на основе контекстуального

понимания. В результате тестирования получили BERTScore = 0,85 (диапазон от 0 до 1), что свидетельствует о высоком уровне совпадения смыслового содержания с ответами оператора.

Выводы. Сервис суммаризации текстовых документов с использованием поисковой расширенной генерации демонстрирует высокую эффективность в упрощении взаимодействия пользователей с официальной документацией МФЦ. Основные преимущества включают снижение времени на обработку запросов, получение более точной информации. Внедрение такого решения позволит снизить нагрузку на справочные центры компаний.

Для дальнейшего развития проекта предлагается доработка и тестирование сервиса для работы с различными категориями документов, использование разных методов разделения документа на сегменты.

Список использованных источников

1. Галеев, Д. Т. Экспериментальное исследование языковых моделей "трансформер" в задаче нахождения ответа на вопрос в русскоязычном тексте // Информатика и автоматизация. – 2022. – Т. 21, № 3. – С. 521-542. – DOI 10.15622/ia.21.3.3. – EDN EVCUJD.

2. Федоров, В. О. Большие языковые модели с поисковой расширенной генерацией: обзор и перспективы // Оригинальные исследования. – 2023. – Т. 13, № 12. – С. 43-47. – EDN GIXWMI.

3. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Электронный ресурс] // arXiv.org. – 2020. – URL: https://arxiv.org/abs/2005.11401 (дата обращения: 10.11.2025).

4. Sun H., Wang Y., Zhang S. Retrieval-Augmented Generation for Domain-Specific Question Answering: A Case Study on Pittsburgh and CMU [Электронный ресурс] // arXiv.org. – 2024. – URL: https://arxiv.org/abs/2411.13691 (дата обращения: 24.11.2024).

Автор _____ Жменько А.Ю.

Научный руководитель _____Воробьева М.С.