УДК 004.83

Метод информационного поиска в векторных пространствах графов знаний для задач разработки вопросно-ответных систем

Меньщиков М.А. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Муромцев Д.И. (ИТМО)

Введение. Разработка и обучение такого семейства больших языковых моделей как GPT, Qwen и LLaMA вызвало революцию в области технологий искусственного интеллекта и фундаментально изменило подходы по обработке естественного языка. Несмотря на свои возможности, связанные с пониманием языка и генерацией текста, LLM сталкиваются с трудностями при обработке вопросов, требующих знаний из специализированной области, информация по которым отсутствовала в их тренировочной выборке.

Подход Retrieval-Augmented Generation (RAG) [1] является естественным решением описанной проблемы, который направлен на повышение фактологической точности и специфичности ответов путем интеграции компоненты поиска в процесс генерации. Несмотря на свои явные преимущества у стандартных RAG-методов есть ряд недостатков: (1) пренебрежение взаимосвязями между фрагментами текста; (2) избыточность информации; (3) отсутствие механизмов обобщения информации.

Подход Graph Retrieval-Augmented Generation (GraphRAG) [2] выступает в качестве одного из ответвлений RAG-подхода, направленный на решение вышеописанных проблем. В отличие от традиционных RAG-техник, GraphRAG извлекает подмножество элементов, содержащих информацию о взаимосвязи друг с другом, из предварительно-построенной графовой структуры данных. Кроме того, хранение информации в виде графа предполагает приведение неструктурированных текстов на естественном языке к набору коротких, ёмких фактов, что позволяет сократить длину, подаваемого на вход LLM, контекста и смягчить проблему излишней информации. Также из-за более концентрированного способа хранения можно на этапе извлечения охватить больше знаний и тем самым повысить качество решения QFS-задачи (Query-Focused Summmarization).

В рамках данного исследования внимание будет сконцентрировано на способах комбинации/модификации существующих GraphRAG-методов с целью повышения качества работы вопросно-ответных систем.

Основная часть. В результате литературного поиска был получен список работ, описывающих алгоритмы построения путей (извлечения триплетов) на графах знаний для решения QA-задачи: ToG, RoG, PMKGR, KG-RAG, GRAG, GNN-RAG, ToGv2, DoG, GCR, PDA. На их основе можно выделить следующие идеи для повышения качества по сравнению со стандартными RAG-модификациями:

1. Построение графа знаний (в качества базы со специализированной информацией, по которой далее будет осуществляться поиск) по набору неструктурированных текстов на естественном языке с помощью LLM.
2. Сопоставление ключевых сущностей из user-вопроса с вершинами в графе знаний в качестве способа инициализации процесса поиска.
3. Поиск/извлечение релевантных триплетов/путей на основе оценки их семантической близости к вопросу с помощью LLM- или Embedding-моделей.
4. Итеративный поиск на графе с использованием текущего набора извлечённой

информации.

5. Динамическое планирование поиска на графе за счёт генерации промежуточных вопросов.
6. Использование нескольких LLM-агентов для генерации/корректировки промежуточных вопросов.

При этом упомянутые методы имеют ряд недостатков:

1. Используются наивные алгоритмы по сопоставлению каждой сущности из user-вопроса по одной вершине из графа: не рассматриваются случаи, когда сущности может соответствовать переменное число вершин.
2. Отсутствует агрегация/суммаризация извлечённой информации для сокращения длины, подаваемого на вход LLM, контекста: “Lost in the Middle”-дилемма [3].
3. Отсутствует стадия по корректировке и разбиению на независимые подвопросы исходного user-вопроса для ускорения и упрощения процесса поиска: не рассматриваются user-вопросы вида “Когда родился Пушкин и Тургенев?”, “Пушкин”.

Таким образом, если использовать вышеупомянутые идеи, а также устранить вышеупомянутые проблемы при разработке GraphRAG-пайплайна, то ожидается получить прирост по качеству работы в рамках QA-задачи.

Сравнение полученного решения будет выполнено как с существующими GraphRAG-методами, так и с рядом стандартных RAG-модификаций: kNN-LM, REALM, DPR, RAG, COLBert-QA, FiD, EMDR2, RETRO, Atlas, REPLUG. В рамках данного исследования выдвигается следующая гипотеза: полученный GraphRAG-метод будет лучше большинства указанных GraphRAG-методов и стандартных RAG-модификаций. Для оценки качества решения QA-задачи выбраны следующие датасеты: TriviaQA, HotpotQA, WebQSP. В качестве основных метрик для оценки близости сгенерированного ответа к “ground truth”-варианту используются следующие варианты: ExactMatch, LLM-as-a-Judge (бинарная оценка). В качестве LLM-моделей для построения графа и осуществления поиска выбраны следующие варианты: Llama3.1 8B, QWEN2.5 7B и DeepSeek-r1 8B.

Выводы. В результате проведённого исследования будет получена оценка качества предложенного GraphRAG-метода и сделан вывод о выполнимости выдвинутой гипотезы. Ожидается, что комбинированный метод по качеству будет лучше в сравнении с заданным списком baseline-решений. Разработанный алгоритм может быть интегрирован в специализированные вопросно-ответные системы для повышения точности и полноты генерируемого текста.

Список использованных источников:

1. Gao Y. et al. Retrieval-augmented generation for large language models: A survey //arXiv preprint arXiv:2312.10997. – 2023.

2. Peng B. et al. Graph retrieval-augmented generation: A survey //arXiv preprint arXiv:2408.08921. – 2024.

3. Liu N. F. et al. Lost in the middle: How language models use long contexts, 2023 //URL https://arxiv. org/abs/2307.03172.