УДК 004.82

Анализ RAG-методов для разработки диалоговой системы по поиску релевантных научных публикаций

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

Введение. На сегодняшний день специалисту в области машинного обучения необходимо за короткие сроки предложить и разработать модель на основе нейросетевых технологий, которая бы оптимально-возможным образом решала поставленную задачу заказчика. Хорошей практикой является брать за основу существующие решения и модифицировать их под новые поставленные требования. Необходимо выполнить поиск и анализ научных публикаций, в результате которого будет получено представление о текущем состоянии технологий по заданной тематике. При этом сам поиск информации в ранее не разобранной области знаний может занять продолжительное время (от 3 до 9 недель), которое уходит на чтение и ручную сортировку работ по их релевантности.

За последние 5-8 лет появилось множество специализированных поисковых систем, которые используют технологии машинного и глубокого обучения для облегчения процессов анализа научных работ: они реализуют функционал QA-систем, суммаризируют информацию из N релевантных статей в виде нескольких абзацев текста, извлекают определённую информацию из публикаций. В качестве примеров таких систем можно выделить следующие: SciSpace [1], Scite [2], Paperdigest [3]. При этом у существующих решений есть проблема, связанная с слабо-контролируемым процессом поиска: пользователь может только ввести текстовый запрос и выполнить фильтрацию по статическим критериям.

В связи с этим планируется разработать диалоговую систему для итеративного поиска и суммаризации научных исследований по конкретной задаче на основе открытых архивов публикаций с применением методов глубокого обучения. Под итеративным поиском понимается последовательное уточнение у пользователя, при помощи обмена текстовыми сообщениями (диалога), особенностей решаемой задачи для получения релевантного конечного результата. Архитектура системы была разбита на 7 независимых частей:

1. Получение релевантных документов из базы знаний по текстовому запросу на естественном языке.
2. Агрегация отобранных документов.
3. Генерация ключевой особенности каждого кластера в виде текста по заданному критерию.
4. Формирование ответа системы в виде сообщения с выбором варианта ответа для продолжения диалога на основании полученных именованных кластеров.
5. Принятие решения системой о прекращении дальнейшей фильтрации множества документов.
6. Формирование конечной поисковой выдачи.
7. Генерация обзора (суммаризации) материалов на основе первых M отобранных релевантных статей.

В данной работе внимание будет сконцентрировано на задаче 1.

RAG-методы для разработки диалоговых систем. В некоторых работах для получения задача-ориентированных диалоговых систем (task-oriented dialogue system) с фиксированным набором знаний (knowledge-grounded) авторы предлагают дообучить

одну языковую модель, которая бы содержала в себе всю необходимую информацию для генерации в end2end-формате релевантного ответа. Модель ChatGPT в явном виде показала проблему, которая возникает при генерации фактоидной информации на основе нейронных сетей: галлюцинации. Это связано с тем, что знания в сети не хранятся в явном виде, а как значения весов её параметров, что затрудняет процесс интерпретации её результатов. При этом языковые модели обладают хорошей обобщающей способностью, за счёт которой им удаётся понимать синтаксические и семантические зависимости естественного языка и генерировать связный текст.

В данной работе будут рассматриваться способы комбинации языковых моделей и более строгих структур хранения информации (баз знаний) для поиска релевантной информации в множестве научных публикаций по заданному вопросу и генерации ответа. Планируется выполнить анализ следующих RAG-методов: REALM [4], RETRO [5], REPLUG [6], SPLADE [7], FLARE [8]. По результатам анализа лучшее решение будет протестировано на области научных публикаций.

Выводы. В результате проведённого исследования будет сделан вывод о возможности применения существующих RAG-методов для получения релевантных документов из заданной базы знаний для области научных публикаций.

Список использованных источников:

1. SciSpace [Электронный ресурс]. – 2024. – URL: https://scispace.com/ (дата обращения 07.02.2024).
2. Scite [Электронный ресурс]. – 2024. – URL: https://scite.ai/ (дата обращения 07.02.2024).
3. PaperDigest [Электронный ресурс]. – 2024. – URL: https://www.paperdigest.org/ (дата обращения 07..02.2024).
4. Guu K. et al. Retrieval augmented language model pre-training //International conference on machine learning. – PMLR, 2020. – С. 3929-3938.
5. Borgeaud S. et al. Improving language models by retrieving from trillions of tokens //International conference on machine learning. – PMLR, 2022. – С. 2206-2240.
6. Shi W. et al. Replug: Retrieval-augmented black-box language models //arXiv preprint arXiv:2301.12652. – 2023.
7. Formal T., Piwowarski B., Clinchant S. SPLADE: Sparse lexical and expansion model for first stage ranking //Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2021. – С. 2288-2292.
8. Jiang Z. et al. Active retrieval augmented generation //arXiv preprint arXiv:2305.06983. – 2023.