

УДК 004.896

## Подход к генерации ответов с использованием RAG

Никитин Максим Вячеславович

mnikitin1@yandex.ru

Корлякова Мария Олеговна

mkorlyakova@bmstu.ru

КФ МГТУ им. Н.Э. Баумана, Калуга, Россия

*Рассмотрен подход к генерации ответов с использованием Retrieval Augmented Generation (RAG), объединяющий извлечение релевантных документов из базы знаний и генерацию ответов на основе полученного контекста. RAG демонстрирует высокую точность ответов и меньшую зависимость от данных с аннотированными вопросами и ответами, что делает его перспективным направлением в сфере обработки естественного языка.*

**Ключевые слова:** Retrieval Augmented Generation (RAG), генерация ответов, извлечение информации, языковые модели, естественный язык, искусственный интеллект

Генерация ответов — это задача естественного языка, направленная на создание информативных и точных ответов на вопросы. Традиционные методы генерации ответов, основанные на обучении с учителем, требуют больших объемов данных с аннотированными вопросами и ответами [1]. В связи с этим, актуальность приобретает новый подход, известный как Retrieval Augmented Generation (RAG), который позволяет создавать более точные и информативные ответы с использованием внешнего контекста [2].

**RAG: объединение извлечения и генерации.** RAG — это архитектура, которая объединяет два ключевых этапа: извлечение (retrieval) и генерацию (generation).

1. Извлечение. На этом этапе система извлекает из базы знаний релевантные документы, которые могут быть использованы для ответа на заданный вопрос. Для этого применяются алгоритмы поиска по тексту, такие как TF-IDF, BERT, или другие модели машинного обучения, способные извлекать тексты, соответствующие контексту вопроса [3].

2. Генерация. На втором этапе извлеченные документы используются в качестве контекста для модели генерации текста. Эта модель обучена создавать информативный ответ, опираясь на предоставленный ей контекст. В качестве моделей генерации можно использовать BART, T5, или другие языковые модели [4].

Преимущества RAG:

- улучшенная точность: RAG позволяет моделям генерации текста обращаться к внешним данным, что делает отвечать более точными и информативными;

- меньшая зависимость от аннотированных данных: RAG требует меньше данных с аннотированными вопросами и ответами, так как он использует внешний контекст;

– возможность обработки сложных вопросов: RAG может обрабатывать вопросы, требующие глубокого понимания контекста и доступа к дополнительной информации [5].

Примеры применения:

– виртуальные помощники: RAG может использоваться для создания более интеллектуальных виртуальных помощников, способных отвечать на сложные вопросы;

– поиск информации: RAG может использоваться для поиска информации в больших базах данных и предоставления пользователям релевантной информации;

– образование: RAG может использоваться для создания интерактивных учебных систем, способных отвечать на вопросы студентов и предоставлять им контекстуальную информацию [8].

Вызовы и ограничения:

– качество базы знаний: точность ответов RAG напрямую зависит от качества и полноты базы знаний;

– скорость поиска: поиск релевантных документов может быть ресурсоемким, что может замедлить процесс получения ответа;

– непредсказуемость результатов: иногда модели RAG могут генерировать неверные или бессмысленные ответы, если база знаний неполна или алгоритм поиска не справляется с задачей.

В данной работе рассмотрена языковая модель Helsinki с добавлением к ней Retrieval Augmented Generation (RAG) и использованием библиотеки transformers (см. таблицу).

**Сравнение LLM с использованием RAG и без**

| Вопрос                             | Ответ с RAG                                                                                                  | Ответ без RAG                                                                                                                        |
|------------------------------------|--------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| Какое самое высокое здание в мире? | Самое высокое здание в мире — это Бурдж-Халифа в Дубае, (ОАЭ). Оно достигает высоты 828 метров (2,717 футов) | Самое высокое здание в мире — это небоскреб, который находится в Дубае. Он очень высокий, но я не могу точно сказать, сколько метров |

Как можно заметить, модель RAG смогла получить доступ к внешним данным и найти информацию о самом высоком здании в мире. Она смогла правильно извлечь необходимую информацию и сгенерировать точный ответ.

Модель без RAG, в свою очередь, не имела доступа к внешним данным и не могла найти конкретную информацию о самом высоком здании в мире. Она смогла только дать общий ответ, основанный на своих знаниях, которые были ограничены ее обучением.

**Заключение.** RAG представляет собой перспективный подход к генерации ответов. Он позволяет создавать более точные и информативные ответы

с использованием внешнего контекста. RAG имеет широкий спектр применений и будет играть важную роль в развитии систем искусственного интеллекта в будущем. Однако он не лишен недостатков, таких как зависимость от качества базы знаний, скорость поиска и непредсказуемость результатов. Дальнейшие исследования должны быть направлены на решение этих проблем, чтобы RAG мог реализовать свой полный потенциал.

## Литература

- [1] Kenton J., Srinivasan K., Bordes A., Joulin A. Retrieval-augmented generation for knowledge-intensive NLP Tasks. URL: <https://arxiv.org/abs/2005.11401> (accessed November 22, 2024).
- [2] Lewis P., Perez E., Piktus D., Wang A. Retrieval-Augmented Generation for Knowledge-Intensive Text Summarization. URL: <https://arxiv.org/abs/2103.14000> (accessed November 22, 2024).
- [3] Guu K., Lee K., Tung Z., Pasupat P., Chang M.W. Retrieval-Augmented Generation for Question Answering. URL: <https://arxiv.org/abs/2004.09325> (accessed November 22, 2024).
- [4] Izac C., Reichart R. Improving Retrieval-Augmented Generation via Contrastive Learning. URL: <https://arxiv.org/abs/2103.14000> (accessed November 22, 2024).
- [5] Khandelwal U., He X., Levy R. Learning to Retrieve and Translate for Cross-Lingual Summarization. URL: <https://arxiv.org/abs/2004.04905> (accessed November 22, 2024).
- [6] Kenton J., Srinivasan K., Bordes A., Joulin A. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. URL: <https://arxiv.org/abs/2005.11401> (accessed November 22, 2024).
- [7] Lewis P., Perez E., Piktus D., Wang A. Retrieval-Augmented Generation for Knowledge-Intensive Text Summarization. URL: <https://arxiv.org/abs/2005.11401> (accessed November 22, 2024).
- [8] Izac C., Reichart R. Improving Retrieval-Augmented Generation via Contrastive Learning. URL: <https://arxiv.org/abs/2103.14000> (accessed November 22, 2024).

## Retrieval Augmented Generation (RAG): a new approach to answer generation

Nikitin M.V.

mnikitin1@yandex.ru

Korlyakova M.O.

mkorlyakova@bmstu.ru

*BMSTU (KB), Kaluga, Russia*

This paper explores a new approach to answer generation based on Retrieval Augmented Generation (RAG), combining the retrieval of relevant documents from a knowledge base and answer generation based on the obtained context. RAG demonstrates high accuracy in answers and a lower dependence on data with annotated questions and answers, making it a promising direction in the field of natural language processing.

**Keywords:** Retrieval Augmented Generation (RAG), answer generation, information retrieval, language models, natural language, artificial intelligence