С. В. Савошинский, А. Д. Савченко, В. С. Воинов

АНАЛИЗ И ОПТИМИЗАЦИЯ ЭТАПОВ RETRIEVAL-AUGMENTED GENERATION СИСТЕМ

В работе анализируются этапы retrieval в системах Retrieval-Augmented Generation: chunking, embeddings, индексация и оценка качества поиска. В качестве baseline рассматривается fixed-size разбиение на 100-словные пассажи. Сравнение выполняется для dense, sparse и hybrid retrieval с использованием метрик Precision@K, Recall@K и MRR.

Введение

Retrieval-Augmented Generation (RAG) – архитектура, в которой генеративная модель использует внешнюю базу знаний через механизм поиска релевантных документов. Качество retrieval-этапа напрямую влияет на итоговое качество генерации, поскольку модель ограничена извлечённым контекстом.

I. Границы исследования

Работа ограничена retrieval-контуром: ingestion → chunking → embeddings → indexing → retrieval → reranking. Генеративная часть системы не рассматривается.

II. Теоретические основы retrieval

Dense Passage Retrieval реализует retrieval как bi-encoder: запрос и документ кодируются в общее embedding-пространство и сравниваются по скалярному произведению. Практически retrieval включает быстрый поиск кандидатов и последующее ранжирование.

III. Стратегии chunking

Базовым подходом является fixed-size chunking – разбиение текста на неперекрывающиеся блоки по 100 слов, применяемое в DPR и RAG. Альтернативой является semantic chunking, основанный на семантической близости предложений, а также recursive chunking, сохраняющий структуру документа через иерархию разделителей.

IV. Embedding-подходы

MTEB показывает отсутствие универсально лучшей embedding-модели, а BEIR подтверждает эффективность BM25 как baseline для sparse retrieval. В dense retrieval основным подходом являются трансформерные encoder-модели на основе BERT, учитывающие контекст текста.

V. Индексация и ANN-поиск

Для retrieval в RAG широко применяются FAISS и HNSW. При сравнении retrieval-индексов необходимо учитывать качество поиска, latency и потребление памяти.

VI. Метрики оценки retrieval

Для оценки retrieval используются метрики Precision@K, Recall@K и MRR:

Recall@K характеризует полноту извлечения релевантных документов, Precision@K – качество верхней части выдачи, а MRR – позицию первого релевантного результата.

VII. Методика эксперимента

Сравнение проводится для chunking-стратегий и embedding-моделей с использованием метрик Precision@K, Recall@K и MRR с учётом latency и времени индексации.

VIII. Заключение

Оптимизация retrieval в RAG должна основываться на воспроизводимом эксперименте. Fixed-size chunking рассматривается как baseline, а сравнение retrieval-подходов должно учитывать качество и вычислительные характеристики индексов.

Список литературы

1. Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // NeurIPS. 2020.
2. Karpukhin V. et al. Dense Passage Retrieval for Open-Domain Question Answering // EMNLP. 2020.
3. Hearst M. A. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages // Computational Linguistics. 1997.
4. Muennighoff N. et al. MTEB: Massive Text Embedding Benchmark // EACL. 2023.
5. Thakur N. et al. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. 2021.
6. FAISS Documentation. IndexHNSW [Electronic resource].

Савошинский Станислав Викторович, студент БГУИР, e-mail: stanislavsavoshinsky@gmail.com.
Савченко Алексей Дмитриевич, студент БГУИР, e-mail: asavchenko792@gmail.com.
Воинов Владислав Сергеевич, студент БГУИР, e-mail: v.voinov441@gmail.com.
Научный руководитель: Зотов Никита Владимирович, младший научный сотрудник НИЛ 3.3, ассистент кафедры интеллектуальных информационных технологий БГУИР, n.zotov@bsuir.by.