Building Retrieval Benchmarks using Retrieval Augmented Generation

Junyoung Son°, Taemin Lee[1,*]
Korea University, Human-Inspired AI Research
{s0ny,taeminlee}@korea.ac.kr

Abstract

Since text retrieval can significantly influence the overall performance of RAG applications, it is crucial to objectively evaluate the performance of retrieval models across various search tasks. However, compared to embedding benchmarks for English, the development and research of Korean embedding model benchmarks remain insufficient. Addressing this need, this paper proposes a methodology for constructing an automated retrieval model benchmark based on LLMs and RAG models. Additionally, using this methodology, we build benchmark data and evaluate various retrieval models, commercial embedding models, and open-source embedding models.

Keywords: Retrieval Benchmark, Retrieval, Retrieval Augmented Generation

1. Introduction

With the advancement of Large Language Models (LLMs), text retrieval technologies are being utilized in various fields, including Retrieval-Augmented Generation (RAG). One of the challenges when building such a RAG system is to ensure that the retrieval model soes not become the bottleneck of the entire RAG system. In other words, the performance of the overall RAG system can heavily depend on how effectively the retrieval model provides results that are relevant to the query's intent [1].

Various benchmark datasets have been constructed globally to evaluate the performance of retrieval models. Notably, BEIR released an evaluation dataset encompassing nine information retrieval tasks, such as fact verification, query retrieval, and entity retrieval, using 18 open datasets [2]. MTEB, focusing on the generalization performance of text embedding models, published an extensive benchmark dataset comprising 58 datasets for eight embedding tasks [3]. While these benchmark datasets have significantly contributed to the evaluation of numerous retrieval/embedding models, they are primarily composed of English-language data. Consequently, benchmark datasets for evaluating Korean retrieval models remain insufficient.

To response to this need, we propose a pipeline for constructing a benchmark for Korean retrieval models. In details, we generate queries from a collection of documents and try to answer the query using existing RAG model to measure the extent to which each document contributes to the response.

2. Method

2.1 Query Generation

We utilizes OpenAI gpt-4o [4] for the query generation. Given a document, 4 types of queries are generated: 1) signle-document queires, 2) multi-document queries, 3) queries containing typographical errors, and 4) ambiguous queries.

2.2 Query-Document Pair Selection

Considering that the entire document collection may contain potential positive candidate documents that can be used for answering than the document used for query generation, we perform k-nearest neighbor searches over the entire document collection for each query to construct (query, candidate document set) pairs. If the document used to generate the query is not included in the search results, we manually added.

Since not all documents within the candidate set may contribute to answering the query, we identify and remove such documents. In the filtering stage, we utilize Cohere Command-R model [5], an RAG applications. this model is

*Corresponding author

Models	Benchmark Result		
	Recall@5	**Precision@5**	**NDCG@5**
BM25	41.44	12.07	34.35
text-embedding-3-small	37.79	11.18	31.21
text-embedding-3-large	43.5	12.74	36.71
multilingual-e5-large	48.28	14.17	41.84
e5-mistral-7b-instruct	**48.66**	**14.23**	**41.99**

Table 1. Experimental results on our synthesized benchmark.

trained to generate answer with supporting documents following reference spans. By using this information, we filter out the documents that there is no contributes to answering the query. In addition, we calculate word co-occurrence between the query and the supporting documents to filter out potential noises.

3. Experiment

3.1 Experimental Setting

The documents used for synthesizing the benchmark were drawn from a financial domain set. Each document is composed of multiple paragraphs. We employ a document-level input to enable the generation of complex queries that reference multiple documents. We observed that a significant number of documents not referenced in the answers were removed after filtering using RAG (approximately 5 to 2).

To validate the synthesized benchmark, we employ the following representative retrieval models that support the Korean language: BM25 [6], OpenAI Embedding Models [4], multilingual-e5 [7, 8].

3.2 Experimental Result

The experimental results using our synthesized benchmark are presented in Table 1. Supervised Embedding Models demonstrate superior performance, with the largest model, e5-mistral-7b-instruct, achieving the highest average performance. On the other hand, OpenAI embedding models shows significantly lower performance compared to the word-based model such as BM25. Overall, In summary, the quantitative results indicate a moderate level of difficulty, neither exceptionally easy nor overly challenging.

4. Conclusion

In this study, we proposed a methodology for synthesizing a retrieval benchmark in response to the need for evaluating the performance of Korean retrieval models. The proposed methodology utilizes Large Language Models and Retrieval-Augmented Generation models to synthesize retrieval benchmark. LLMs were employed to generate various types of queries given a document, while RAG models were used to verify the documents that can answer the generated queries. We hope that this work will facilitate the development of benchmarks for evaluating Korean retrieval models across various domains and embedding tasks in the future.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425), and ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT)(IITP-2024-2020-0-01819)

Reference

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in*

Neural Information Processing Systems, Vol. 33, pp. 9459–9474, 2020.

[2] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models."

[3] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

[4] OpenAI, "Chatgpt 4o," 2024, accessed: 2407. [Online]. Available: https://openai.com/chatgpt

[5] Cohere, "Cohere command-r models," 2024, accessed: 2407. [Online]. Available: https://cohere.com/command

[6] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2009.

[7] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Improving text embeddings with large language models," *arXiv preprint arXiv:2401.00368*, 2023.

[8] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," *arXiv preprint arXiv:2402.05672*, 2024.