The 8th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2024)

Rethinking Retriever Evaluation in Retrieval-Augmented Generation: A Document- and Word-Level Analysis

Chanhee Park[†], Heuiseok Lim[*‡†]
[‡]Department of Computer Science and Engineering, Korea University, [†]Human-Inspired AI Research
{pch7678, limhseok}@korea.ac.kr

Abstract

Evaluating retriever performance in retrieval-augmented generation (RAG) systems requires going beyond traditional document-level metrics to capture the nuanced interplay between retrieved information and generated output. In this paper, we propose a dual labeling evaluation approach that assesses retriever accuracy at both the document and word levels, providing a more granular understanding of performance. Using the KLUE MRC benchmark, which includes queries, contexts, and answers, we compute probabilities reflecting the retriever's ability to identify the correct document and pinpoint the crucial answer word within it. We then evaluate three Korean-language retrievers, demonstrating that this method reveals finer-grained performance differences overlooked by conventional methods. Our analysis highlights the strengths and weaknesses of each retriever, emphasizing the importance of not only retrieving relevant documents but also effectively locate essential information within them. This study provides valuable insights for developing and improving retrievers tailored for the specific demands of RAG tasks.

Keywords: Retriever Augmented Generation, Natural Language Processing, Evaluation, Information Retrieval

1. Introduction

Retrieval-augmented generation (RAG) has emerged as a powerful framework, enhancing generative models by incorporating external knowledge[1]. However, evaluating RAG systems poses significant challenges due to the complex interplay between retriever and generator components. Existing evaluation methods often fall short of accurately assessing RAG performance, primarily focusing on document-level retrieval accuracy without delving into the granular interaction between retrieved information and generated output. Specifically, these metrics do not account for the nuanced role of each retrieved chunk in supporting the generator's task, especially when the relevance of a document does not guarantee the presence of the exact answer needed by the generator[2].

This paper addresses this gap by proposing a novel evaluation framework that employs dual labeling on the retrieved results at both document and word levels. This approach provides a multifaceted view of retriever performance, allowing us to differentiate between a retriever's ability to locate relevant documents and its effectiveness in pinpointing specific answer words within those documents. By incorporating both levels of granularity, this method offers a deeper understanding of how well retrieval aligns with generation, providing insights previously overlooked by conventional methods. We demonstrate the effectiveness of this approach on the KLUE MRC dataset, offering a comprehensive analysis of retriever strengths and weaknesses that can inform future RAG system development.

2. Evaluation Method

Traditional retrieval evaluation metrics, such as F1 Score, MAP, MRR and NDCG[3, 4], evaluate retrievers based on document relevance but do not account for whether the retrieved documents contain the exact answer words needed for generation. Table 1 shows an example of a retrieved chunk whose document is correct but does not provide the specific information necessary for the generation process. Conversely, Table 2 shows a chunk from an irrelevant document that still contains the correct answer word. These two examples underscore the limitations of document-level evaluation.

To address these shortcomings, we propose a dual labeling approach that evaluates the retriever's performance at both

[QA data]
'document': **'José Carreras'**
'question': 'In which city is the hospital located where José Carreras was treated with the help of the foundation?'
'answers': 'Madrid'
[retrieve result]
'document': **'José Carreras'**
'chunk': 'The Hermosa Foundation was founded by Plácido Domingo. Domingo established the foundation to help treat José Carreras' illness. He wanted to help anonymously so as not to hurt the pride of his rival, Carreras. Carreras was moved by Domingo's friendship, and they became close friends afterward. This experience also led Carreras to establish his own leukemia foundation.'

Table 1. An example with a relevant document without the answer word.

the document and word levels. For each query, the retriever returns the most relevant chunk of text, on which we assign two labels. To be specific, **doc_label** is set to 1 if the chunk originates from the gold-standard answer document and 0 otherwise. **word_label** is set to 1 if the chunk contains the correct answer word and 0 otherwise.

Using these labels, we compute the following probabilities:

- $\mathbf{P}(\mathbf{Doc})$: Probability of the retrieved chunk belonging to the correct document.
- $\mathbf{P}(\mathbf{Word})$: Probability of the retrieved chunk containing the correct answer word.
- $\mathbf{P}(\mathbf{Doc} \cap \mathbf{Word})$: Probability of the chunk being from the correct document and containing the correct answer word.
- $\mathbf{P}(\mathbf{Doc} \mid \mathbf{Word})$: Probability of the chunk belonging to the correct document given it contains the correct answer word.
- $\mathbf{P}(\mathbf{Word} \mid \mathbf{Doc})$: Probability of the chunk containing the correct answer word given it comes from the correct document.

This dual labeling scheme provides a more comprehensive view of retriever performance. High $\mathbf{P}(\mathbf{Doc} \cap \mathbf{Word})$ indicates excellent overall performance, while high $\mathbf{P}(\mathbf{Doc} \mid \mathbf{Word})$ suggests strong document discrimination ability. A high $\mathbf{P}(\mathbf{Word} \mid \mathbf{Doc})$ signifies the retriever's effectiveness in pinpointing crucial information within relevant documents.

[QA data]
'document': 'Yeo Woon-hyung'
'question': 'Who was dispatched by the Shinhan Youth Party to the Paris Peace Conference?'
'answers': **'Kim Kyu-sik'**
[retrieve result]
'document': 'March 1st Movement'
'chunk': 'At the time, Yeo Woon-hyung and Shin Gyu-sik, who were studying in China, judged that the declaration and the subsequent Paris Peace Conference would be a decisive event for Korea's future, regardless of whether they achieved independence. They organized the Shinhan Youth Party on paper and dispatched **Kim Kyu-sik**, who was fluent in French, to the Paris Peace Conference, while sending Jang Deok-soo, proficient in Japanese, to Japan. This news was a significant piece of information for independence activists in and outside Korea.[3]'

Table 2. An example with an irrelevant document with the answer word.

3. Experiments

3.1 Dataset and Models

We utilize a subset of the KLUE MRC[5] validation set, specifically 1825 data points sourced from Wikipedia. Each query is paired with an answer span located within a Wikipedia page. To evaluate retriever performance in isolation, we avoid using the provided contexts and instead crawl the Wikipedia pages corresponding to the document titles. The crawled documents are then chunked using LangChain[6]'s RecursiveCharacterTextSplitter with a chunk size of 300 and an overlap of 20.

We evaluate three retrievers capable of processing Korean

Model	**P(Doc)**	**P(Word)**	**P(Doc ∩ Word)**	**P(Doc | Word)**	**P(Word | Doc)**
BM25	0.7025	0.5447	0.5342	0.9809	**0.7605**
BGE	0.7403	0.5364	0.5266	0.9817	0.7113
e5	**0.7737**	**0.5732**	**0.5633**	**0.9828**	0.7280

Table 3. Retriever Accuracy Results

input: BM25 with Kiwi tokenizer[7], BAAI/bge-m3[8], and Infloat/multilingual-e5-large[9]. These models are assessed using the proposed dual labeling approach, providing insights into their document-level and word-level retrieval performance.

3.2 Results

Table 3 presents the evaluation results for each retriever using our proposed metrics. As evident from the results, our method provides a deeper understanding of retriever behavior:

Overall Performance: While e5 achieves the highest scores across most metrics, indicating superior overall performance, BM25, despite a lower **P(Doc)**, exhibits the highest **P(Word | Doc)**. This suggests that BM25, once it retrieves the correct document, excels at identifying the crucial answer word within it.

Document Discrimination: The consistently high **P(Doc | Word)** across all models indicates their proficiency in differentiating between documents based on the presence of the answer word.

Targeting Relevant Information: The relatively lower **P(Word | Doc)** scores compared to **P(Doc | Word)** suggest room for improvement in pinpointing the exact answer word within a relevant document. This highlights the need for future research focusing on finer-grained retrieval within documents.

4. Conclusion

This paper presents a dual labeling evaluation framework that provides a more granular understanding of retriever performance in retrieval-augmented generation tasks. By distinguishing between document-level and word-level accuracy, this method reveals strengths and weaknesses in different retriever models that are often missed by traditional metrics. Our experiments show that even when a model excels in document retrieval, it may struggle to identify the exact information needed for generation, highlighting the importance of fine-grained evaluation. Future work will explore this approach in more complex multi-chunk retrieval scenarios and assess its applicability across diverse datasets. This work contributes to a more comprehensive evaluation of RAG systems, offering actionable insights for improving information retrieval tailored to generation tasks.

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). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

Reference

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Vol. 33, pp. 9459–9474, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[2] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2405.07437

[3] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models," *arXiv preprint arXiv:2104.08663*, 2021.

[4] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human-generated machine reading comprehension dataset," 2016.

[5] S. Park, "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.

[6] LangChain, "Langchain documentation," https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter, 2024, accessed: September 5, 2024.

[7] TeddyNote, "랭체인langchain 노트 - langchain 한국어 튜토리얼kr," https://wikidocs.net/251980, 2024, accessed: September 5, 2024.

[8] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," *arXiv preprint arXiv:2402.03216*, 2024.

[9] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," 2024. [Online]. Available: https://arxiv.org/abs/2402.05672