BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Document Retrieval with Fine-grained Relevance Cues

Bc. Antonín Jarolím*

Abstract

Fine-grained retrieval is critical for improving downstream NLP tasks such as question answering and retrieval-augmented generation, but existing methods either lack deep document understanding or incur prohibitive computational costs. This paper aims to develop an efficient, explainable retrieval system that identifies minimal evidence spans supporting document relevance to a query. We collect a small human-annotated dataset and generate large-scale token-level relevance annotations using the Gemma 2 (27B) model on MS-MARCO. We then extend ColBERT with lightweight token-level heads supervised by a weighted Binary Cross Entropy and KL loss to predict per-token relevance cues. We propose three architectural variants—Embedding De-normalization, Separate Linear Projections, and Non-linear Transformation—that balance retrieval efficiency with interpretability. Our ColBERT-based model (120M) matches the fine-grained relevance cue quality of the much larger Gemma 2 (27B) model on the human-annotated evaluation set, despite being over 270 times smaller. We show that our extensions enable accurate, token-level explainability with minimal computational overhead. Our work provides a scalable, efficient pathway to more trustworthy and interpretable retrieval systems, reducing latency, minimizing hallucinations, and making fine-grained explainability feasible for real-world NLP applications.

*xjarol06@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Research shows that highlighting or extracting the most relevant parts of documents not only accelerates human understanding of relevance but also improves retrieval efficiency [1, 2]. However, while large language models (LLMs) can generate summaries to aid comprehension, hallucinated or inaccurate summaries risk misleading users. A promising solution is to limit generation to the most relevant document spans, underlining the importance of fine-grained retrieval.

In this work, we aim to extend interpretability into information retrieval by developing models that not only retrieve relevant passages but also explain their relevance at a fine-grained level. Fine-grained explanations—such as token-level scores—offer two key benefits: they allow users to judge passage relevance more quickly, and they can sometimes fully satisfy the information need without reading the entire document. This is particularly crucial in cases where only a small snippet, like a phone number or a short fact, is sufficient. Therefore, producing explanations at the lowest possible granularity is essential for making retrieval both more interpretable and more efficient.

2. Proposed Approach

To train a model that jointly predicts the relevance of a passage to a given query and identifies the most relevant parts of the passage, two types of annotations are required. The first annotation serves as supervision for the retrieval task, indicating which documents are relevant to the query. The second provides fine-grained information to guide the extraction task, specifying which parts of the relevant document are most relevant.

Currently, no large-scale retrieval dataset with token-level annotations exists. To address this gap, we extend MS-MARCO [3] dataset by incorporating fine-grained relevance information. We noticed that LLM can do much better in fine-grained relevance extraction, therefore we evaluate multiple LLM on human annotated dataset to select best LLM for this task. Evaluation proved LLM Gemma 2 (27B) [4] to be the best performing model for this task, as it matches with human annotations the most Given the cost and inefficiency of manual extraction annotation, we utilize large language models (LLMs) to identify the most informative and contextually relevant spans within a

relevant document for a given query.

The proposed approach modifies ColBERT [5, 6] to support fine-grained extraction without sacrificing retrieval performance. This change explicitly biases the model towards identifying fine-grained relevance signals within passages.

3. Query-Document Interaction

In the original ColBERT, relevance scores are computed by matching query tokens to the most relevant document tokens. Our modification introduces an additional training objective: searching for the best matching query token for each passage token. The magnitude of this match is interpreted as an indication of how relevant each passage token is to the entire query. A sigmoid function and Binary Cross-Entropy loss are used to obtain the probability of passage token relevance, guiding the model's supervision.

1. **Embedding De-normalization:** almost preserves retrieval footprint efficiency while recovering richer representations for interpretability by scaling retrieval representations.
2. **Separate Linear Projections:** maximizes task-specific expressivity but at the cost of doubling the index size, saving both retrieval and interpretability representations.
3. **Non-linear Transformation:** preserves the retrieval footprint by applying a lightweight feed-forward network (FFN) to the retrieval representations of the top-k retrieved documents.

These variants offer trade-offs between retrieval efficiency, interpretability, and deployment flexibility.

4. Training And Evaluation

The de-normalization architecture is most affected by the trade-off between retrieval and extraction tasks. For example, in the de-normalized architecture, the model initially performs well on retrieval (F1 score 0.98) but its performance drops to 0.83 as it focuses more on fine-grained extraction, with the F1 score increasing to 0.67. This pattern is consistent across other de-normalized runs, where the retrieval performance decreases as the model shifts focus to extraction due to increasing parameter values.

In contrast, the Separate Linear Projections architecture retains retrieval performance during training. Although its extraction performance reaches an F1 score of 0.677, it remains one of the best models, offering a Pareto optimal solution. When initialized from BERT, this architecture achieved better extraction performance, likely due to fewer constraints, though it couldn't match the retrieval performance of other models after similar training steps.

The Separate Linear Transformation with FFN architecture (with a skip connection) starts from pre-trained ColBERT and retains retrieval performance while achieving the best extraction F1 scores. However, adding layer normalization to the FFN network (in another version) negatively impacted retrieval performance, showing a drop right after training began. Unfortunately, training with this architecture starting from BERT did not yield a well-performing model.

5. Conclusions

We present a modified retrieval approach that provides built-in relevance explainability through token-level cues. Our ColBERT-based retrieval model (120M) achieves fine-grained relevance cue quality comparable to that of Gemma 2 (27B), despite being over 270 times smaller. To support different deployment needs, we propose three architectural variants—Embedding De-normalization, Separate Linear Projections, and Non-linear Transformation—offering a trade-off between retrieval efficiency and interpretability.

References

[1] Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 229–239, New York, NY, USA, 2019. Association for Computing Machinery.

[2] Jurek Leonhardt, Koustav Rudra, and Avishek Anand. Extractive explanations for interpretable text ranking. *ACM Trans. Inf. Syst.*, 41(4), March 2023.

[3] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016.

[4] Gemma Team. Gemma: Open models based on gemini research and technology, 2024.

[5] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. *CoRR*, abs/2004.12832, 2020.

[6] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient re-

trieval via lightweight late interaction. *CoRR*, abs/2112.01488, 2021.