The Solution for The AQA-KDD-2024 OAG-Challenge

Zhe Zhang
Nanjing University of Science and Technology
Nanjing Shi, China
zzqianlan@163.com

WeiLi Guo
Nanjing University of Science and Technology
Nanjing Shi, China
wlguo@njust.edu.cn

YiZhe Zhang
Nanjing University of Science and Technology
Nanjing Shi, China
zhangyizhe@njust.edu.cn

Abstract

This report presents our solution for the OAG-Challenge's Academic Question Answering (AQA) task, which aims to retrieve relevant papers that answer specialized questions. We utilized vector models to embed papers, simplified questions using large language models (LLMs), and employed FAISS for vector-based retrieval. Finally, we applied a re-ranking model for detailed sorting of the results. Our approach leverages pretrained LLMs for efficient and effective text embedding and retrieval. The implementation details and code are publicly available at https://github.com/qianlanzz/AQA-KDD-2024.git.

ACM Reference Format:
Zhe Zhang, WeiLi Guo, and YiZhe Zhang. 2024. The Solution for The AQA-KDD-2024 OAG-Challenge. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The rapid growth of academic publications has made it increasingly difficult for researchers to stay updated with the latest advancements. The OAG-Challenge at KDD Cup 2024 aims to advance academic knowledge graph mining, with the Academic Question Answering (AQA) task focusing on retrieving relevant papers to answer specialized questions [5].

To address this challenge, we developed a model that uses vector embeddings for papers, simplified questions with large language models (LLMs), and FAISS [4] for efficient vector-based retrieval. A re-ranking model then refines the results. Our solution leverages pretrained LLMs for text embedding, ensuring efficiency and resource-friendliness.

This report details our approach, highlighting the methods and technologies used to create an effective AQA system, ultimately contributing to more intelligent academic information retrieval. Our main contributions are as follows:

- Integration of LLMs for Question Simplification: We utilized large language models to effectively simplify complex academic queries, enhancing the accuracy of subsequent paper retrieval by aligning the question format with the embedding model's capabilities.
- Efficient Vector-Based Retrieval and Re-ranking: Our approach employs FAISS for efficient vector-based retrieval, followed by a re-ranking model to refine the results. This combination ensures high precision in identifying the most relevant papers, optimizing both retrieval speed and accuracy.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Work

The field of academic question answering (AQA) and paper retrieval has seen significant advancements with the advent of large-scale language models and advanced retrieval techniques. Several key approaches have been explored in recent literature. Large Language Models (LLMs) for Text Embedding: Recent advancements in natural language processing have shown the effectiveness of pretrained large language models, such as BERT, GPT, and their variants, for generating high-quality text embeddings [3]. These embeddings capture the semantic meaning of text and have been successfully applied in various retrieval tasks. For instance, models like SciBERT [1], specifically trained on scientific literature, have shown improved performance in academic contexts.

Similarity-Based Retrieval Methods: Traditional retrieval methods often rely on term frequency-inverse document frequency (TF-IDF) or BM25 for text matching [6]. However, these methods may fall short in capturing the semantic relationships between queries and documents. Recent approaches leverage vector-based retrieval systems, such as FAISS, which use dense embeddings to perform efficient and accurate similarity searches. This shift towards vector-based retrieval has significantly improved the performance of information retrieval systems.

Re-ranking Strategies: To further refine retrieval results, re-ranking techniques have been developed. These techniques typically involve training a secondary model to re-evaluate and order the initial set of retrieved documents based on their relevance to the query. Methods such as Learning to Rank (LTR) and neural re-ranking models have demonstrated enhanced performance by considering the contextual relevance and deeper semantic connections between queries and documents.

Our approach builds upon these advancements by integrating LLMs for question simplification and utilizing FAISS for efficient vector-based retrieval. The combination of these techniques, along with a robust re-ranking model, ensures high precision in identifying the most relevant academic papers for specialized queries.

3 Method

To effectively address the Academic Question Answering (AQA) task, we developed a comprehensive multi-step method that incorporates advanced natural language processing techniques and

efficient retrieval strategies. Our approach consists of the following key components:

3.1 BGE Model Tuning and Fusion

We began by selecting the BGE model [2], which is well-regarded for its robust performance in generating high-quality text embeddings. Recognizing the potential for further enhancement, we employed a process known as hard example mining, wherein we identified instances that posed significant challenges to the initial model. By fine-tuning the BGE model using these difficult examples, we enabled it to learn from complex and nuanced data points, thereby improving its overall robustness and accuracy. Following the fine-tuning phase, we proceeded to fuse the enhanced BGE model with the original version, effectively combining their respective strengths. This fusion process was designed to maintain the fine-tuned model's improved accuracy while preserving the original model's broad generalization capabilities, resulting in a more balanced and effective solution.

3.2 Paper Embedding with BGE Model

To facilitate efficient and accurate paper retrieval, we generated embeddings for all papers in the corpus using the newly enhanced BGE model. These embeddings provided a rich semantic representation of the academic texts, capturing intricate details and relationships within the content. We then stored these embeddings in a FAISS (Facebook AI Similarity Search) vector database, which is specifically optimized for high-speed similarity searches. The use of FAISS enabled us to handle large-scale datasets effectively, ensuring rapid retrieval of relevant papers based on their vector representations. This step was crucial in creating a scalable and efficient retrieval system that could handle the extensive corpus of academic literature involved in the AQA task.

3.3 Query Simplification and Embedding

Given the inherent complexity of the questions in the AQA task, which often included numerous hyperlinks and irrelevant content, it was imperative to preprocess these queries to enhance retrieval accuracy. We employed a large language model to simplify and summarize the questions, effectively distilling them to their core informational needs while removing extraneous elements. This preprocessing step ensured that the queries were more focused and aligned with the relevant academic content. The instructions provided to llama3 are as follows: Refine and summarize the given question without leaving out details and proper nouns.Please use a paragraph and don't generate irrelevant content.

The simplified questions were then embedded using the BGE model, creating a consistent and comparable vector representation for both queries and papers. This approach was critical in ensuring that the retrieval system could accurately match the simplified queries with the most pertinent academic papers.

3.4 Retrieval and Re-ranking

Utilizing the FAISS vector database, we conducted an initial retrieval process to identify the top 100 candidate papers based on the query embeddings. This step leveraged the efficiency of FAISS to quickly sift through the vast corpus and generate a broad set of potentially relevant results. To refine these initial results, we applied the BGE-Reranker-v2-minicpm-layerwise model, a sophisticated re-ranking tool designed to evaluate the top 100 papers with greater precision. The re-ranking model considered deeper semantic connections and contextual relevance, reordering the initial set of papers to prioritize the most relevant and high-quality ones. This final re-ranking step ensured that our system provided highly accurate and contextually appropriate responses to the academic queries, significantly enhancing the overall effectiveness of the AQA task.

By integrating these components, our approach effectively combines the strengths of BGE models for both embedding generation and re-ranking, along with the efficiency of FAISS for vector-based retrieval. The use of query simplification and model fine-tuning with hard examples further enhances the accuracy and relevance of the retrieved academic papers, making our solution both robust and resource-efficient.

4 Experiments

To evaluate the effectiveness of our approach, we conducted a series of experiments focusing on the impact of re-ranking models on retrieval performance. We used three configurations: retrieval without re-ranking, retrieval with the bge_reranker model, and retrieval with the bge-reranker-v2-minicpm-layerwise model. The performance of each configuration was measured using a specific evaluation metric, with the results summarized below.

4.1 Without Reranker

In the baseline configuration, we conducted the retrieval process without applying any re-ranking model. This setup served as a control to understand the inherent performance of our initial retrieval system. The score obtained in this scenario was 0.134. This relatively low score highlighted the need for further refinement to improve the relevance and accuracy of the retrieved papers.

4.2 With BGE Reranker

The second configuration involved the use of the bge_reranker model for re-ranking the initially retrieved results. The bge_reranker is designed to enhance the relevance of the top retrieved papers by re-evaluating their semantic connections to the query. Applying this model improved the performance, resulting in a score of 0.146. This improvement demonstrated the positive impact of re-ranking on retrieval accuracy, though it indicated that further enhancements were possible.

4.3 With LLM-based Reranker

The third configuration utilized the bge-reranker-v2-minicpm-layerwise model, which combines advancements in large language model-based re-ranking techniques. Our rerankers are initialized from google/gemma-2b for the LLM-based reranker and openbmb/MiniCPM-2B-dpo-bf16 for the LLM-based layerwise reranker. These models were trained on a mixture of multilingual datasets, allowing them to handle diverse linguistic contexts effectively. This configuration achieved the best performance, with a score of 0.166. The significant improvement in this scenario highlighted the superior capability

Table 1: Experimental results of different methods

Method	Score on test set
No_Reranker	0.138
BGE_Reranker	0.146
LLM_Reranker	0.166

of the bge-reranker-v2-minicpm-layerwise model in enhancing retrieval relevance and accuracy.

Results. Table 1 shows the performance of our method on the official final test set.The experimental results clearly demonstrate the importance of re-ranking in academic paper retrieval tasks. The progressive improvements from the baseline to the advanced re-ranking model underscore the value of leveraging sophisticated re-ranking techniques to achieve higher precision in retrieving relevant academic papers.

5 Conclusion

In this study, we addressed the Academic Question Answering (AQA) task from the OAG-Challenge, focusing on retrieving relevant academic papers in response to professional queries. Our approach involved generating text embeddings using the bge model, fine-tuning it with hard samples, and integrating these embeddings into a faiss vector store for efficient retrieval. We simplified complex queries using a large language model and employed advanced re-ranking techniques to enhance retrieval accuracy. Our experiments demonstrated significant improvements in performance with the use of re-ranking models, achieving the highest score of 0.166. These findings underscore the effectiveness of our methods in improving academic paper retrieval, providing a valuable framework for future research and development in this domain.

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: a pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

[2] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

[5] Saeed-Ul Hassan, Anam Akram, and Peter Haddawy. 2017. Identifying important citations using contextual information from full text. (2017).

[6] Ammar Ismael Kadhim. 2019. Term weighting for feature extraction on twitter: a comparison between bm25 and tf-idf. In *2019 international conference on advanced science and engineering (ICOASE)*. IEEE, 124–128.