AQA Vector to LLM Rerank

YunGui Zhuang

Algorithm Engineer,Tgnet,GuangZhou

Abstract

Academic data mining is rich in many entity-centric applications, such as paper retrieval, expert discovery and journal recommendation. However, the lack of data benchmarks related to academic knowledge graph mining has severely limited the development of the field. The dataset is derived from OAG-QA, which retrieves question posts from StackExchange and Zhihu websites, extracts the URLs of papers mentioned in the answers, and matches them with the papers in OAG. Participants are provided with a dataset of questions and need to find the papers that best match those questions. We propose a bge vector model first, which is fine-tuned and then rearranged by LLM to get the final result.

1 Introduction

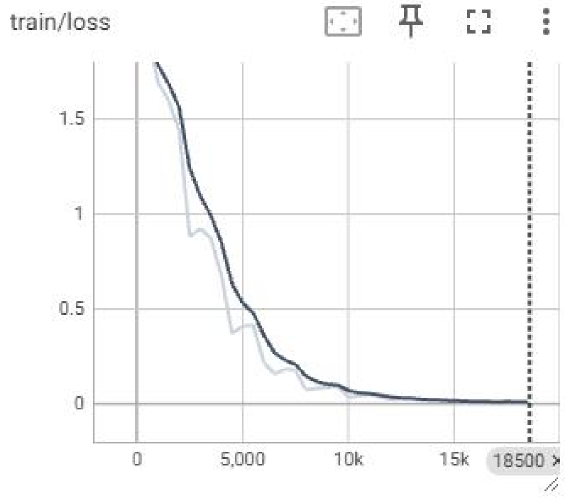
The ultimate goal of academic data mining is to deepen our insights into scientific developments, nature and trends. By unlocking the deep value of academic data, we can tap and unleash the potential power of science, technology and education. Specifically, in-depth analyses of academic data can not only provide strong support to governments in formulating science policies, but also help companies identify and nurture talent, while facilitating researchers to explore and acquire new knowledge more efficiently. This process not only promotes the innovation and dissemination of knowledge, but also accelerates the progress of the scientific community as a whole.

For this task, we are targeting data sourced from OAG-QA, which retrieves question posts from StackExchange and Zhihu websites, extracts the URLs of the papers mentioned in the answers, and matches them with the papers in OAG. Using question-paper pairs to train the retrieval model requires finding the papers that best match these questions.

2 Methodology

2.1 BGE Vector Model Fine-Tuning

We use the vector retrieval model of bge-large-en-v1.5 [1] as a base, and then build positive and negative samples for training based on the given training set, in which building positive and negative samples has a large impact on the coarse recall, we recall the top500 based on the original bge-large-en-v1.5 model,and then select from top20-500 between 10 data as a negative sample, in the preliminary round, this sample selection has the best impact on the results.



2.2 LLM-Rerank

NV-Embed [2] presents several new designs, including having the LLM attend to latent vectors for better pooled embedding output, and demonstrating a two-stage instruction tuning method to enhance the accuracy of both retrieval and non-retrieval tasks.

We used a model fine-tuned with BGE (Bilateral Graph Enhancement) to perform an initial coarse recall to filter out the top 100 candidate documents. Subsequently, a fine-grained re-ranking of these candidate documents was performed using NV-Embed to optimise the relevance and accuracy of the retrieval results. This strategy has demonstrated significant benefits in real-world applications, not only improving retrieval efficiency, but also significantly enhancing the quality and depth of results. With this two-stage approach combining advanced fine-tuning and intelligent rearrangement, we are able to capture users' needs more accurately and provide them with richer and more valuable academic resources.

3 Results

In the end, we used a two-stage implementation of coarse recall and LLM for re-ranking. bge fine-tuned score is 0.1317; using the fine-tuned bge for recalling top100 candidate papers, and then using the NV-Embed-v1 model for re-ranking these top100 data to calculate the score, and the final score after re-ranking is 0.17478. Thus, it also proves the feasibility of this two-stage approach of coarse recall followed by rearrangement.

4 Discussion

Our study employs an innovative two-stage approach that combines fine-tuning of BGE vector models and LLM re-ranking to improve the accuracy of matching academic questions with relevant papers. This strategy is a novel attempt in the field of academic data mining, which not only improves the retrieval efficiency, but also enhances the depth and quality of the results. However, this approach also poses some challenges. For example, the selection of positive and negative samples in the fine-tuning process has a significant impact on model performance and needs to be carefully designed to avoid bias. In addition, the computational cost of the LLM reshuffling phase is high, which may limit the application of the model in resource-constrained environments. We discuss how these challenges can be overcome by optimising the algorithm and utilising more efficient computational resources.

The experimental results show a significant improvement in retrieval scores after BGE model fine-tuning and NV-Embed model rearrangement. This result validates the effectiveness of our approach, but also triggers thoughts on further optimisation of the model performance. We provide an in-depth analysis of the possible reasons for the score improvement, including the model's deeper understanding of academic data and a more refined rearrangement strategy. In addition, we have explored the potential of this approach to be applied in different academic domains and how further research can extend the applicability of the model. Future work will include exploring more data sources, improving the model architecture, and developing more efficient algorithms to handle larger datasets.

5 Conclusion

My research not only provides a new solution to the field of academic data mining, but also demonstrates the potential of this approach in improving retrieval efficiency and quality of results through practical applications. I believe that this work can provide strong support for science policy making, talent identification and knowledge acquisition, thus accelerating the overall progress of the scientific community.

Despite the positive results of my research, I recognise that there is room for further optimisation and expansion. In the future, I plan to explore additional data sources and model architectures to further improve the system's generalisation capabilities and application scope.

Overall, my thesis demonstrates an academic data mining approach that combines advanced fine-tuning and intelligent rearrangement, which has significant advantages in improving retrieval efficiency and result quality, and opens up new paths for the construction and application of academic knowledge graphs.

References

[1] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. ArXiv, abs/2402.03216.

[2] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. ArXiv, abs/2402.03216.Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., & Ping, W. (2024). NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. ArXiv, abs/2405.17428.