Solution Report for the KDD 2024 OAG-Challenge: Academic Question Answering Task

Jintang Li
lijt55@mail2.sysu.edu.cn
Sun Yat-sen University
Guangzhou, China

Xinzhou Jin
jinxzh5@mail2.sysu.edu.cn
Sun Yat-sen University
Guangzhou, China

Wangbin Sun
sunwb7@mail2.sysu.edu.cn
Sun Yat-sen University
Guangzhou, China

Abstract

This report details our solution for the OAG-Challenge, a competition aimed at advancing the state of academic knowledge graph mining technologies. We focused on the Academic Question Answering (AQA) task, which requires retrieving relevant papers that answer specialized questions. Our solution involves using pretrained large language models (LLMs) for generating text embeddings and employing similarity-based retrieval to identify the top 20 matching papers for each question. It is worth noting that our solution is built upon open-sourced LLMs for text embedding, making it training-free and resource-friendly for participants. Despite this, we achieved a top-9 rank in both the public and private leaderboards. Code is made public available at https://github.com/EdisonLeeeee/KDDcup24-AQA.

CCS Concepts

• **Information systems** → *Clustering*; **Information retrieval query processing**; • **Computing methodologies** → *Unsupervised learning*; *Learning latent representations*; **Natural language processing**; *Lexical semantics*.

Keywords

OAG-Challenge, Academic Question Answering, Large Language Models, Semantic Search

ACM Reference Format:

Jintang Li, Xinzhou Jin, and Wangbin Sun. 2018. Solution Report for the KDD 2024 OAG-Challenge: Academic Question Answering Task . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Academic data mining aims to deepen our understanding of scientific development, nature, and trends by uncovering significant scientific, technological, and educational values from academic data. This field leverages large-scale data sets from various academic sources to extract meaningful insights and patterns that can influence policy-making, support talent discovery, and enhance the efficiency of knowledge acquisition for researchers.

The OAG-Challenge [1], introduced at KDD Cup 2024, comprises three realistic and challenging academic tasks designed to promote the latest developments in academic knowledge graph mining [6]. These tasks are structured to push the boundaries of current technologies, encouraging participants to innovate and develop advanced methods for academic data analysis.

We have entered the OAG-Challenge and focus on the Academic Question Answering (AQA) task, which is particularly critical in today's fast-paced technological environment. The AQA task aims to provide high-quality, cutting-edge academic knowledge to researchers and the general public by developing a model to retrieve relevant papers in response to professional questions. Given the exponential growth in academic publications, it has become increasingly difficult for individuals to stay current with the latest research in their fields. An effective AQA system addresses this challenge by streamlining the process of finding pertinent research papers, thus saving time and enhancing the productivity of researchers.

To tackle this task, participants must build models capable of understanding and processing complex academic queries and mapping them to the most relevant academic papers. This involves sophisticated natural language processing techniques to accurately interpret the queries and a deep understanding of the vast corpus of academic literature to identify the most relevant responses. The dataset provided for this task includes question-paper pairs, derived from platforms such as StackExchange [2] and Zhihu [3], which link user queries to specific academic papers referenced in the answers.

The broader goal of the AQA task is not only to improve the efficiency of academic information retrieval but also to advance the overall capabilities of academic knowledge graphs. These graphs represent relationships between various academic entities such as papers, authors, and institutions, and are crucial for understanding the structure and dynamics of scientific research. By improving our ability to navigate and utilize these graphs, the AQA task contributes to a more connected and insightful academic ecosystem, fostering collaboration and innovation.

This report details our approach to the AQA task, outlining the solution used to develop a robust and accurate model for the OAG-Challenge. Through the use of pretrained text embedding models and advanced similarity search algorithms, our solution achieved top results with less memory and resource. The results and insights

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

[1] https://www.biendata.xyz/kdd2024/
[2] https://stackexchange.com/
[3] https://www.zhihu.com/

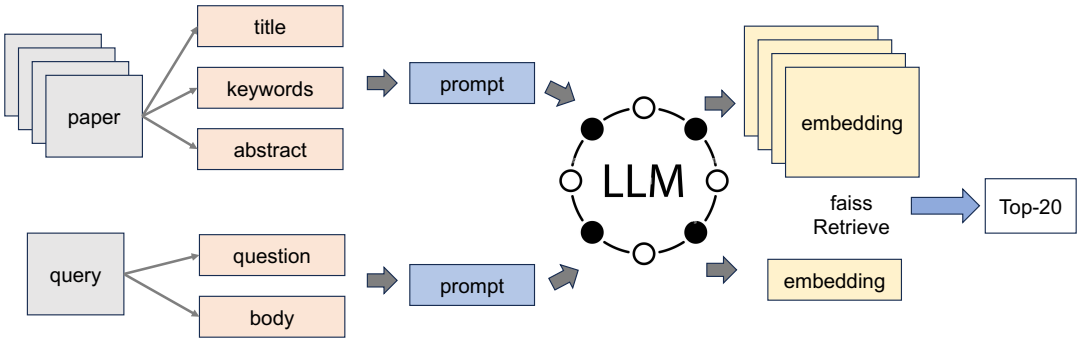

Figure 1: Overall framework.

gained from this challenge have the potential to significantly impact the field of academic data mining, paving the way for more intelligent and responsive research tools in the future.

2 Related Work: Semantic Search

Semantic search has revolutionized information retrieval and retrieval augmented generation (RAG) techniques [3]. By leveraging the power of language models and text embeddings, semantic search enables more accurate and efficient document retrieval. Semantic search goes beyond traditional keyword-based search methods by considering the meaning and intent behind a user's query. It leverages text embeddings, which are multi-dimensional numerical representations of "meaning" generated by language models. These embeddings capture the semantic relationships between words and phrases, allowing for more nuanced and context-aware document retrieval.

Semantic search involves representing both user queries and documents in an embedding space. By mapping the semantics of text onto this multi-dimensional space, it becomes possible to perform vector searches to find documents that align closely with the user's query intent. In recent years, semantic search, empowered by LLMs and text embeddings, has revolutionized the way we retrieve information. LLM-enabled semantic search offers several advantages over traditional search methods. Some of the key benefits include:

- **Enhanced Precision**: By understanding the meaning and intent behind queries, semantic search provides more precise and contextually relevant search results.
- **Improved Efficiency**: Leveraging the power of LLMs and text embeddings allows for faster retrieval of relevant documents, reducing search times.
- **Multilingual Support**: LLMs are capable of handling multiple languages, making semantic search effective across diverse linguistic contexts.
- **Versatility**: Semantic search can be applied to various domains, including web search, document retrieval, question answering systems, and chatbots, among others.

By incorporating the semantic meaning of text into the search process, we can achieve more accurate and efficient document retrieval. However, choosing the right embedding model and understanding the underlying technologies are essential for successful implementation. As LLMs continue to evolve, the future of semantic search holds even greater potential for advancing the field of information retrieval and natural language understanding.

3 Dataset Description and the AQA Task

Academic question answering. In an era of rapid technological advancement and information growth, it is crucial to provide high-quality, multi-domain academic knowledge. The AQA task challenges participants to train a retrieval model using question-paper pairs. Traditional keyword-based information retrieval cannot satisfy professional knowledge retrieval in the era of artificial intelligence. For instance, consider the question, "Can neural networks be used to prove conjectures?". How to retrieve answers and evidence from scholarly literature? Given an academic question q and a paper set $P^q = \{p_1^q, p_2^q, \ldots, p_N^q\}$, the goal of academic question answering is to select the most relevant papers from the candidate set P^q.

AQA dataset. The dataset, presented by OAG-challenge, includes questions from StackExchange and Zhihu, with answers that reference URLs of papers matched to the OAG dataset. The dataset comprises 17,948 question-paper pairs. In addition, Participants are provided with a question dataset and must find the papers most relevant to these questions. Questions cover 22 disciplines and 87 topics, forming a two-level hierarchical structure; that is, each topic belongs to a discipline. For each topic, 10,000 candidate papers, including the ground-truth papers in the answers, are included.

4 Proposed Solution

In this section, we detail our solution for the AQA task. The overall framework is presented in Figure 1. Specifically, we design prompts for LLMs to process the query and paper and then generate embeddings for them respectively. The top matches of queries and papers are indexed by the retrieval step using the efficient Faiss library.

Prompt Design. Given an input query with fields of *question* and *body*, we combine the text together with a query template as the output prompt. The query template is designed with a prefix as 'Given a question, retrieve passages that answer the question: {query}'. For the prompt of papers, we simply concatenate the texts from the title, abstract, and keywords from additional data [4].

[4]https://opendata.aminer.cn/dataset/DBLP-Citation-network-V15.zip

Embedding Generation Model. Embeddings play a crucial role in enabling semantic search. These numerical representations capture the semantic meaning of text by encoding it into vectors consisting of hundreds to thousands of numbers. The choice of an embedding model is crucial for the effectiveness of semantic search. We use Nv-Embed [4] as the pretrained LLM model for inference, which can be accessed from Hugging Face. Compared to existing LLMs, NV-Embed 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.

Retrieval. The final piece of the solution is the retrieval step. Given a query's semantic representation or embedding, we need to identify the most relevant papers based on their proximity in the embedding space. The proximity is typically measured with a Cosine similarity function. By calculating the similarity scores between query embeddings and paper embeddings, a retrieval system can rank and present the most suitable papers for each query. For ease of implementation, we use Faiss [2] as the backend for similarity scoring and semantic search of top matches. Faiss is widely used in academia and industry for tasks such as nearest neighbor search, clustering, and dimensionality reduction. Its ability to efficiently handle large-scale data and provide quick, accurate results makes it a valuable tool for many machine learning and data analysis applications.

Implementations and results. Our solution is implemented using PyTorch [1] and the Transformers [5] library, providing a robust framework for building and saving/loading LLMs. We leverage the pretrained Nv-Embed model, known for its advanced design and high performance in various natural language processing tasks. For our implementation, we set the batch size to 2 and the maximum sequence length to 4096 tokens, allowing us to manage memory consumption efficiently while effectively embedding individual queries or documents. The model is deployed on an NVIDIA RTX 4090 GPU with 24 GB of memory. Despite that our solution does not involve the fine-tuning step, we achieved a top-9 rank in both the public and private leaderboards.

5 Conclusion

This report presents our solution for the AQA task in the OAG-Challenge. By employing pretrained embedding models and using Faiss for similarity-based retrieval, we successfully enhanced the accuracy of matching questions to relevant papers. Our solution is training-free and resource-friendly, as it does not involve the training or fine-tuning of LLMs, making it accessible for follow-up participants. The success of our solution in the AQA task not only provides a framework for future research but also sets a benchmark for the development of more intelligent and responsive academic tools. Future work could further optimize the model and improve retrieval performance.

References

[1] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM. https://doi.org/10.1145/3620665.3640366

[2] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[3] Vikas Jindal, Seema Bawa, and Shalini Batra. 2014. A review of ranking approaches for semantic search on Web. *Inf. Process. Manag.* 50, 2 (2014), 416–425.

[4] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. arXiv:2405.17428 [cs.CL]

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[6] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, Tianyi Han, Yuwei An, Dan Zhang, Weng Lam Tam, Kun Cao, Yunhe Pang, Xinyu Guan, Huihui Yuan, Jian Song, Xiaoyan Li, Yuxiao Dong, and Jie Tang. 2024. OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009