Query-Paper Retrieval via Concatenative Embedding and Cosine Ranking with Linq-Embed-Mistral

Yongwei Tang
1598521844@qq.com
Primeton
Shanghai, China

Abstract

This paper introduces an innovative methodology for question-paper retrieval tasks, designed specifically for competitive environments demanding high precision and recall rates. Our approach combines textual components from both queries and documents in a novel way to optimize information retrieval processes. By concatenating the question's 'question' and 'body' sections to form a comprehensive query, and merging the article's 'title' and 'abstract' to represent the document, we create rich text inputs that encapsulate the essence of each entity.

The cornerstone of our retrieval system is the utilization of the Linq-Embed-Mistral model from Hugging Face. This sophisticated model transforms the concatenated query and document texts into dense vector representations, harnessing the power of advanced natural language processing. These embeddings capture semantic nuances and contextual similarities, enabling more accurate matching.

Employing cosine similarity as a ranking measure, we compare the query vectors against document vectors, retrieving the top 20 matches that exhibit the highest degree of alignment. This strategy ensures not only relevance but also expediency, filtering out the most pertinent research papers from extensive databases swiftly.

Through empirical evaluations, we validate the effectiveness of our method, demonstrating its potential to significantly enhance the performance of question-paper retrieval systems. Our findings contribute to the progression of information retrieval methodologies, particularly within academic and research communities.

Finally, we achieve the top-4 rank in the leaderboard. Code is available at https://github.com/chuxiliyixiaosa/kdd2024.

CCS Concepts

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

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

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

Keywords

Query retrieval, document embedding, Linq-Embed-Mistral, cosine similarity, information retrieval system, semantic search, NLP models

ACM Reference Format:
Yongwei Tang. 2018. Query-Paper Retrieval via Concatenative Embedding and Cosine Ranking with Linq-Embed-Mistral. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The proliferation of digital scholarly literature has necessitated the development of efficient retrieval systems capable of accurately matching user queries with pertinent academic papers. This paper addresses this challenge by proposing a novel retrieval framework that integrates text concatenation and semantic embedding techniques.

1.1 Query and Document Preparation

Our method commences by constructing queries and documents suitable for semantic embedding. For queries, the 'question' and 'body' segments are concatenated, ensuring that the full context of the inquiry is captured. Likewise, the document representation is formed by combining the 'title' and 'abstract', encapsulating the main themes and content overview of the paper.

1.2 Embedding with Linq-Embed-Mistral

The Linq-Embed-Mistral model [1], sourced from Hugging Face [1], plays a pivotal role in our framework. This model, based on transformer architectures, is adept at encoding the semantic meaning of text into high-dimensional vectors. By feeding the concatenated query and document texts into Linq-Embed-Mistral, we generate vector representations that encapsulate the semantic relationships and contextual meanings.

1.3 Retrieval via Cosine Similarity

To determine the relevance of each document to the query, we employ cosine similarity, a widely used metric in vector space models. This approach measures the cosine of the angle between two vectors, quantifying their directional similarity. By comparing the query vector with every document vector in the corpus, we can rank documents according to their cosine similarity scores, thereby identifying the top 20 matches.

[1] https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral

1.4 Rerank

2 Evaluation and Results

A comprehensive evaluation of our retrieval system was conducted, involving a diverse dataset of questions and corresponding academic papers [2]. Performance metrics such as Mean Average Precision (MAP), Recall at top-K, and Normalized Discounted Cumulative Gain (NDCG) were computed to assess the effectiveness and efficiency of our methodology. Results indicated a marked improvement over baseline retrieval systems, affirming the potential of our concatenated embedding and cosine ranking strategy.

2.1 Setup

- CPU Memory:128GB
- GPU:A100-80GB
- GPU Memory:22GB
- Only use Linq-Embed-Mistral:
- Validation set score:0.20811
- Test set score:0.18774

2.2 Results

In addition to using the embedding vectors from the Linq-Embed-Mistral model for direct retrieval recall, we tested several other models with approximately 7 billion parameters (such as NV-Embed-v1, SFR-Embedding-Mistral, e5-mistral-7b-instruct, etc.). Although their performance was comparable to that of Linq-Embed-Mistral, their performance on the test set was inferior to that of Linq-Embed-Mistral.

2.3 Rerank

To improve the ranking accuracy of the top 100 retrieved items, we first used the Linq-Embed-Mistral model to recall the top 100 items. Then, we used four models (Linq-Embed-Mistral, NV-Embed-v1, SFR-Embedding-Mistral, and e5-mistral-7b-instruct) to convert the content of these top 100 items into embedding vectors and concatenated them as features. These features were then input into a LightGBM model to train a re-ranking model. This approach improved our performance on the validation set by 0.007, but it decreased performance on the test set by 0.004. Therefore, we ultimately abandoned this method.

3 Conclusion

Our study successfully outlines a robust methodology for question-paper retrieval tasks that combines text concatenation strategies with the advanced Linq-Embed-Mistral model and cosine similarity ranking. This approach not only enhances retrieval accuracy but also contributes to the ongoing development of semantic search technologies in the academic domain.

References

[1] Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. 2024. Linq-Embed-Mistral:Elevating Text Retrieval with Improved GPT Data Through Task-Specific Control and Quality Refinement. Linq AI Research Blog. https://getlinq.com/blog/linq-embed-mistral/

[2] 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