Evaluating Tokenization Strategies for Embedding based document Re-Ranking in Biomedical retrieval

Submitted in partial fulfillment of the requirements
for the award of the degree of

Master of Technology
in
Information Technology
With specialization in MLIS

Submitted by

Manu Pande
(Enrollment No. MML2023005)

Under the Supervision of
Dr. Muneendra Ojha

DEPARTMENT OF INFORMATION TECHNOLOGY
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
ALLAHABAD
Prayagraj-211015, India

SEPTEMBER 26, 2024

CANDIDATE DECLARATION

I, Manu Pande, MML2023005, certify that this thesis work titled **"Evaluating Tokenization Strategies for Embedding based document Re-Ranking in Biomedical Retrieval"**, submitted towards fulfillment of MASTER'S THESIS report of M.Tech. IT at Indian Institute of Information Technology Allahabad, is an authenticated record of our original work carried out under the guidance of Dr. Muneendra Ojha. Due acknowledgements have been made in the text to all other material used. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Manu Pande

Manu Pande
MML2023005
Machine Learning and Intelligent Systems

Acknowledgements

I reserve a special place in my heart for my beloved parents, whose unwavering love, unwavering support, and unwavering belief in my abilities have been the bedrock upon which my dreams have flourished. Their persistent support, sacrifices, and unshakable trust in my abilities have been the driving factors behind my quest for knowledge and academic pursuits.

I owe a tremendous debt of gratitude to my esteemed supervisor, **Dr. Muneendra Ojha** , whose guidance and advice have been the compass guiding me through the many twists and turns of this thesis. His stimulating conversations, insightful feedback, kind advice, and boundless forbearance have challenged me to push the boundaries of my capabilities and inspired me to strive for academic excellence. I am very thankful for the trust you put in me and the chances you gave me to grow both professionally and personally. I am grateful beyond words for the opportunity to have worked under your guidance, and I hope my thesis serves as a fitting tribute to your hard work, knowledge, and encouragement.

Finally, I want to thank the Faculty who helped me grow as a scholar.

Manu Pande

ABSTRACT

This study explores the effectiveness of embedding-based document re-ranking for biomedical information retrieval using different tokenization strategies. We implement large language models (LLMs) to generate semantic embeddings for both documents and queries and re-rank documents based on similarity scores. Our approach systematically compares general-purpose tokenizers with domain-specific ones to assess their impact on re-ranking performance. Additionally, we fine-tune an LLM model on a biomedical corpus to enhance domain-specific relevance. The evaluation includes standard information retrieval metrics, comparing the performance of our approach with baseline methods. By testing various tokenizers and assessing their effect on document ranking, this work aims to optimize re-ranking processes and provide insights into improving search relevance in biomedical applications.

Contents

List of Abbreviations

LLM	Large Language Model
TF-IDF	Term Frequency-Inverse Document Frequency
BM25	Best Matching 25

Chapter 1

Introduction

The effectiveness of document ranking in information retrieval systems, particularly in domains like biomedical research, is often limited by traditional techniques that rely heavily on keyword matching. While methods such as BM25 [1] or TF-IDF [2] are widely used, they are inherently constrained by their inability to capture the full semantic meaning of both user queries and document content. As a result, users often retrieve documents that may match keywords but lack contextual relevance to their actual information needs. This shortcoming is particularly evident in complex queries involving technical jargon, synonyms, or multi-conceptual relationships, which are common in domains like biomedicine.

There is a growing need for more advanced document re-ranking techniques that go beyond surface-level keyword matching. By integrating large language models (LLMs) capable of semantic understanding, we can re-rank the initial results retrieved from data stores in a way that prioritizes documents based on their deeper relevance to the user's query. LLMs can analyze the context of words within a document, allowing them to rank documents that may not contain direct keyword matches but are semantically aligned with the intent behind the query. This represents a significant advancement over traditional methods that focus solely on frequency or proximity of keywords.

Moreover, the choice of tokenization strategy—the way text is broken into smaller units for processing—plays a crucial role in the effectiveness of LLM-based re-ranking. Different tokenizers handle domain-specific vocabulary and complex terminologies with varying degrees of success. General-purpose tokenizers may not fully capture the nuances of specialized language in fields like biomedicine. In contrast, domain-specific tokenizers, are tailored to handle the intricate terminology of biomedical texts. By experimenting with multiple tokenization strategies, we aim to optimize document re-ranking for better performance in retrieving contextually relevant documents.

In this study, we explore embedding-based re-ranking methods using LLMs, combined with various tokenization strategies, to improve the relevance of search results in biomedical information retrieval. We evaluate the performance of general-purpose and domain-specific tokenizers to understand their impact on document ranking and determine how fine-tuning an LLM for domain-specific queries can further enhance retrieval accuracy. By doing so, we aim to address the limitations of traditional keyword-based ranking systems and provide more accurate, contextually relevant search results for users.

Chapter 2

Literature Review

2.1 Literature Table

Authors	Paper	Findings
Z. Qin, R. Jagerman, K. Hui, et al. [3]	Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting	This paper introduces "pairwise ranking prompting" for document ranking. It takes two documents and a query, and the LLM is asked to rank the two based on relevance to query. It uses off the shelf LLMs without domain-specific fine tuning. Context is given to the LLM about the task which is - ranking documents making use of few-shot learning.
B. Nourinanloo and M. Lamothe [4]	Re-Ranking Step by Step: Investigating Pre-Filtering for Re-Ranking with Large Language Models	This paper introduces an LLM-based pre-filtering step before re-ranking, where passages or documents are filtered out based on their relevance to a query. The pre-filtering helps reduce the noise that could misguide the re-ranking process. Then use of LLM is made for these best candidates to re-rank them. This led to development of re-rankers which were much smaller yet effective
S. Zhuang, B. Liu, B. Koopman, and G. Zuccon [5]	Open-source Large Language Models are Strong Zero-shot Query Likelihood Models for Document Ranking	This paper explores using open-source LLMs to estimate how likely a document is to be relevant to a given query. Higher likelihood means higher rank. The models perform effectively in zero-shot settings, where no task-specific training is provided. The paper also shows that additional instruction fine-tuning may hinder effectiveness
A. Drozdov, H. Zhuang, Z. Dai, et al. [6]	PaRaDe: Passage Ranking using Demonstrations with Large Language Models	This research addresses the limitations of zero-shot learning. Incorporates few-shot demonstrations into prompt. Demonstrations are pairs of passages and their relevance scores. The paper argues that presenting the LLM with difficult examples yield better results in ranking tasks. Difficult here means queries or passages that present more challenge to LLM to correctly rank

Table 2.1: Literature Table

Chapter 3

Methodology

Step 1: Data Collection

- **Biomedical Documents:** Collect a corpus of biomedical research papers
- **Queries:** Use a dataset which contains query-document pairs and relevance judgments. Or, create queries based on the domain.
- **Initial Document Retrieval:** Use an initial retrieval system to get a set of candidate documents for re-ranking.

Step 2: Preprocessing and Tokenization

- For each document and query, tokenize using each tokenizer and measure the difference in token length, word split, and vocabulary coverage.
- Analyze how tokenization handles biomedical terms like disease names, gene names, or abbreviations.

Step 3: Embedding Generation

- Use pre-trained LLMs to generate embeddings for the documents and queries.
- For each tokenizer, generate embeddings for both the documents and queries.
- Compare embeddings of documents and queries from different tokenizers using cosine similarity and computational efficiency

Step 4: Re-Ranking Process

- Start with retrieving an initial list of candidate documents.
- Re-rank the documents based on their similarity scores, using embeddings generated from each tokenizer.

Step 5: Fine-Tuning for Domain Specificity

- Fine-tune the LLM using a pairwise ranking loss function
- Where the model learns to rank more relevant documents higher

Step 6: Evaluation

- Compare the re-ranking results with the baseline traditional ranking methods.
- Evaluate how different tokenizers (general-purpose vs. domain-specific) affect the rankings.

References

[1] Wikipedia contributors, "Okapi bm25 — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Okapi_BM25&oldid=1194828429, 2024, [Online; accessed 22-September-2024].

[2] ——, "Tf–idf — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Tf%E2%80%93idf&oldid=1236851603, 2024, [Online; accessed 21-September-2024].

[3] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, L. Yan, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, and M. Bendersky, "Large language models are effective text rankers with pairwise ranking prompting," 2024. [Online]. Available: https://arxiv.org/abs/2306.17563

[4] B. Nouriinanloo and M. Lamothe, "Re-ranking step by step: Investigating pre-filtering for re-ranking with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.18740

[5] S. Zhuang, B. Liu, B. Koopman, and G. Zuccon, "Open-source large language models are strong zero-shot query likelihood models for document ranking," 2023. [Online]. Available: https://arxiv.org/abs/2310.13243

[6] A. Drozdov, H. Zhuang, Z. Dai, Z. Qin, R. Rahimi, X. Wang, D. Alon, M. Iyyer, A. McCallum, D. Metzler, and K. Hui, "Parade: Passage ranking using demonstrations with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2310.14408