Enhancing Medical Information Retrieval with a Hybrid RAG System

Saxer Michael, Stuhlmann Linus
ZHAW School of Engineering
Bachelor's Program in Data Science

March 13, 2024

1 Introduction

Due to the exponential growth of medical literature, we investigate a new strategy for medical information retrieval. Our system combines traditional retrieval methods with state-of-the-art machine learning models to refine the accuracy and efficiency of information retrieval for healthcare professionals.

Our project centers on a Retrieval Augmented Generation system [1], which benefits from the integration of a conventional sparse vector space model like BM25 and a pretrained Neural Retriever. This system is enhanced by leveraging the text generation capability of LLaMA 2 [2]. A recent study by Luo et al. [3], have demonstrated significant enhancements in biomedical information retrieval through the combination of a neural and a traditional document retriever. Our integration seeks to provide an innovative solution that not only fetches medical information efficiently but also contextualizes it, delivering insights that are directly applicable to healthcare professionals' queries.

The efficacy of the proposed system will be measured using the BioASQ dataset [4], ensuring that our approach makes a contribution to the medical field by augmenting the precision and velocity of information access, thereby strengthening patient care and informed clinical decision-making.

2 Project Objective

The aim is to optimize a RAG-based QA system by investigating whether combining a traditional sparse vector space model with a trainable Neural retriever enhances the precision of the generated responses, ultimately attempting to outperform systems with traditional methods. We will do this by developing and evaluating a system that integrates a non-trainable ranking algorithm, such as BM25 with a trainable retriever like Dense Passage Retrieval (DPR), alongside LLaMA 2, to measure and compare the quality and relevance of retrieved medical information.

3 Methodology and Data

On the bases of the previously mentioned paper by Luo et al; [3] we attempt to reproduce the reported performance of their hybrid model combining DPR with BM25. This retriever will later be complemented by the generative capabilities of LLaMA 2 for enhanced question-answering performance. The system architecture is designed to address the specific challenges of large-scale medical data processing.

We will optimize a pretrained Neural Information Retrieval model, by fine-tuning on a corpus based on PubMed [5] articles. This process utilizes domain-specific data from medical literature Dataset BioASQ [4] for semantic learning. To address the extensive medical document corpus, we adopt data partitioning and parallel processing for improved scalability and efficiency. The system's performance is evaluated using the BioASQ dataset, leveraging ROUGE-L [6], BLEU [7], and word embedding cosine similarity metrics for benchmarking.

4 Timeline

This timeline gives a brief overview on the major milestones of the project.

Milestone	Weeks 1-3	Weeks 4-6	Weeks 7-9	Weeks 10-11
Literature research	X			
Create Prototype	X			
Run benchmark comparison		X		
Mid term presentation		X		
Optimizing RAG ensemble		X		
Debugging Application			X	
Finishing Code			X	
Final Project Presentations				X
Final Report Submission				X

Table 1: Milestone Tracking Table

References

[1] Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *NeurIPS*. Facebook AI Research; University College London; New York University. 2020.

[2] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: `2307.09288 [cs.CL]`.

[3] Man Luo et al. “Improving Biomedical Information Retrieval with Neural Retrievers”. In: *arXiv preprint arXiv:2201.07745* (2022).

[4] Anastasia Krithara et al. “BioASQ-QA: A manually curated corpus for Biomedical Question Answering”. In: *Sci Data* 10 (2023).

[5] *PubMed*. `https://pubmed.ncbi.nlm.nih.gov/`. National Center for Biotechnology Information, U.S. National Library of Medicine.

[6] Jun-Ping Ng and Viktoria Abrecht. “Better Summarization Evaluation with Word Embeddings for ROUGE”. In: *arXiv preprint arXiv:1508.06034* (2015). URL: `https://arxiv.org/pdf/1508.06034.pdf`.

[7] Elastic Search Labs. *Evaluating RAG Metrics*. 2023. URL: `https://www.elastic.co/search-labs/blog/articles/evaluating-rag-metrics` (visited on 12/01/2023).