Enhancing Information Retrieval and Summarization Accuracy in Biomedical Literature Systems Using PageRank

Shyam Saktawat, Ayush Kumar Sahu, Abhishek Choudhary and Priyesh Tandel

Abstract

During the COVID-19 pandemic, significant efforts were made to gather research papers on SARS-CoV-2 and other coronaviruses to aid doctors and researchers. The COVID-19 Open Research Dataset (CORD-19) contains over 400,000 articles. To facilitate easier access to relevant information for healthcare workers and researchers, effective systems for searching and extracting data are essential. We developed a system leveraging transformer-based models for question answering and summarization. By combining keyword-based methods with neural network models, we narrowed the search and identified the most relevant information. For summarization, various scoring methods were utilized to select the best sentences. These summaries, along with user queries, provided accurate answers. Our system was evaluated using the CovidQA dataset, demonstrating improved performance over existing methods for both natural language and keyword-based searches.

Keywords

Information retrieval, PageRank, Question-answering, Abstractive Summarization, Transformer models

1. Introduction

Since its emergence in late 2019, the Novel Coronavirus Disease (COVID-19) has escalated into a worldwide pandemic swiftly. Despite stringent protocols adopted by most countries to contain its spread, the mutating nature of the virus posed serious challenges. Researchers have extensively studied the Coronaviridae family, resulting in a vast volume of scientific literature. During the pandemic, research efforts surged to understand the causes, impact, genetics, and mitigation strategies. This led to initiatives like the COVID-19 Open Research Dataset (CORD-19), comprising over 1,000,000 documents with more than 400,000 full-text research articles on SARS-CoV-2, COVID-19, and related topics.

2. Problem Statement

Developing an effective information retrieval (IR), question-answering (QA), and abstractive summarization system for large-scale biomedical datasets like CORD-19 presents significant challenges. Managing and extracting relevant information from over 400,000 research articles requires highly efficient and accurate systems to meet the needs of researchers and healthcare professionals.

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202311048@daiict.ac.in (S. Saktawat); 202311066@daiict.ac.in (A. K. Sahu); 202311067@daiict.ac.in (A. Choudhary); 202101222@daiict.ac.in (P. Tandel)

3. Dataset

- **CORD-19**: Over 400,000 full-text research articles related to SARS-CoV-2, COVID-19, and other coronaviruses. Initially containing around 1,000,000 abstracts, filtered down to 600,000 entries based on keywords and publication year, and further refined to 50,000 abstracts from top journals, totaling **450,000 sentences**.
- **CovidQA**: A dataset for evaluating QA systems, comprising approximately 600 question-answer pairs across general, biomedical, and expert-level queries.

4. Methodology

We developed an integrated framework utilizing multiple models and algorithms to enhance IR and summarization in biomedical corpora:

4.1. Search Space Reduction

Efficiently reducing the search space in the vast CORD-19 dataset was essential. We employed:

- **Okapi BM25**: A probabilistic IR model ranking sentences based on query term frequency and document frequency.
- **BioSentVec**: Generates sentence embeddings for semantic similarity comparisons [1].
- **PageRank**: Ranks sentences based on their interconnectedness within a similarity graph.

Combining BM25 scores with BioSentVec cosine similarities narrowed down candidate sentences for further processing.

4.2. Ranking Optimization

Shortlisted sentences were refined using the PageRank algorithm, prioritizing influential sentences within a similarity graph constructed by connecting semantically similar sentences.

4.3. Summarization and Question Answering

Top-ranked sentences were input into the T5-large model for abstractive summarization [2], generating concise summaries. These summaries, along with user queries, were then processed by a BioBERT model fine-tuned on the SQuAD dataset to provide accurate answers to complex medical questions [3].

5. Evaluation Metrics

We utilized the following metrics to assess system performance:

5.1. Precision@1 (P@1)

Indicates whether the top-ranked sentence contains the answer:

$$P@1 = \frac{\text{Number of correct answers in top 1}}{\text{Total number of queries}}$$

202311048, 202311066, 202311067, 202101222@daiict.ac.in

5.2. Recall@3 (R@3)

Measures the percentage of correct answers within the top three results:

$$R@3 = \frac{\text{Number of correct answers in top 3}}{\text{Total number of relevant answers}}$$

5.3. Mean Reciprocal Rank (MRR)

Represents the average rank position of the first correct answer across all queries:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where Q is the set of queries and $rank_i$ is the rank of the first relevant document for query i.

6. Results

Our system was evaluated against baseline models using the CovidQA dataset. The results are summarized in Table 1.

Table 1
Performance Comparison on CovidQA Dataset

Model	P@1	R@3	MRR
Random	0.012	0.034	-
BM25	0.150	0.216	0.243
BERT	0.081	0.117	0.159
SciBERT	0.040	0.056	0.099
BioB	0.097	0.142	0.170
BioBERT (SQuAD fine-tuned)	0.161	0.403	0.258
BM25 + BioSentVec + PageRank (Proposed)	**0.169**	**0.267**	**0.344**

7. Key Challenges and Learnings

1. **Handling Large Datasets:** The extensive dataset posed computational and memory challenges. **Solution:** Implemented data loading in smaller batches and used efficient data structures.
2. **Indexing Problems:** Initial indexing led to repetitive sentences. **Solution:** Fine-tuned indexing parameters and incorporated deduplication techniques.
3. **Building a Complex Pipeline:** Integrating multiple components introduced challenges. **Solution:** Adopted a modular approach, developing and testing each component independently.
4. **Missing Details in Research Papers:** Some methodologies lacked comprehensive explanations. **Solution:** Utilized alternative resources, open-source implementations, and consulted experts.

202311048, 202311066, 202311067, 202101222@daiict.ac.in

References

[1] Q. Chen, Y. Peng, Z. Lu, Biosentvec: creating sentence embeddings for biomedical texts, in: 2019 IEEE International Conference on Healthcare Informatics, IEEE, 2019, pp. 1–5.

[2] C. Raffel, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[3] J. Lee, et al., Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

202311048, 202311066, 202311067, 202101222@daiict.ac.in