Information Retrieval on CORD-19 with Question Answering and Summarization

Team 15

Naveen Shenoy - 191IT134

Sarthak Jain - 191IT145

Pratham Nayak - 191IT241

Introduction

- Due to rapid spread, COVID-19 resulted in a pandemic. Considerable research has been made to understand its effects and causes.
- Due to continuous research we have large document space from which valuable information can be retrieved.
- We require a system that can provide answer to queries of users from large document space.
- It is challenging to find important documents which contains answer to the queries of users.
- We aim to implement an model that retrieves relevant sentences based on user query and provides them with the summary and answer to the queries.

Problem Statement & Objectives

To develop and implement an information retrieval system for COVID-19 articles that supports question-answering and summarization.

Objectives:-

- To reduce the search space using a computationally inexpensive ranking algorithm.
- To shortlist top 10-20 candidate sentences for summarization.
- To generate a summary of the final candidates in a few lines.
- To find the exact answer for the given query in a few words.

Original Methodology

The existing methodology is divided into the following sub modules.

- **Graph Formation:** If Paper A and Paper B cite same paper then this indicates A and B are discussing a similar topic. For semantic similarity, BioSentVec embeddings is used to represent the abstracts and articles after which cosine similarities between them is considered. If this value is greater than a threshold or their is a common citation between the two papers, an edge is added between them.
- **Clustering:** Ego-splitting is used to clusters the documents based on the graph structure obtained in the previous step.
- **Mapping queries to documents**: Cosine similarity between the BioBERT embedding of the paper title and the query terms are used to identify important clusters and all the documents from those clusters are considered for further processing.

Original Methodology

- **Shortlisting Sentences**: Based on the cosine similarity between the BioBERT embeddings of the query and the documents, top-100 documents are shortlisted and all the sentences in these documents are taken for further processing.
- **Returning Best Matching Sentences**: The pairwise cosine similarity between the selected sentences is used to add edges between them. PageRank algorithm is used to rank these sentences based on the constructed graph and the top-7 sentences are returned as the best matching sentences.
- **Summarization**: BERT embeddings of the top matching sentences are clustered using K-means and the sentences closest to the centroid are returned as the summaries.

Proposed Methodology

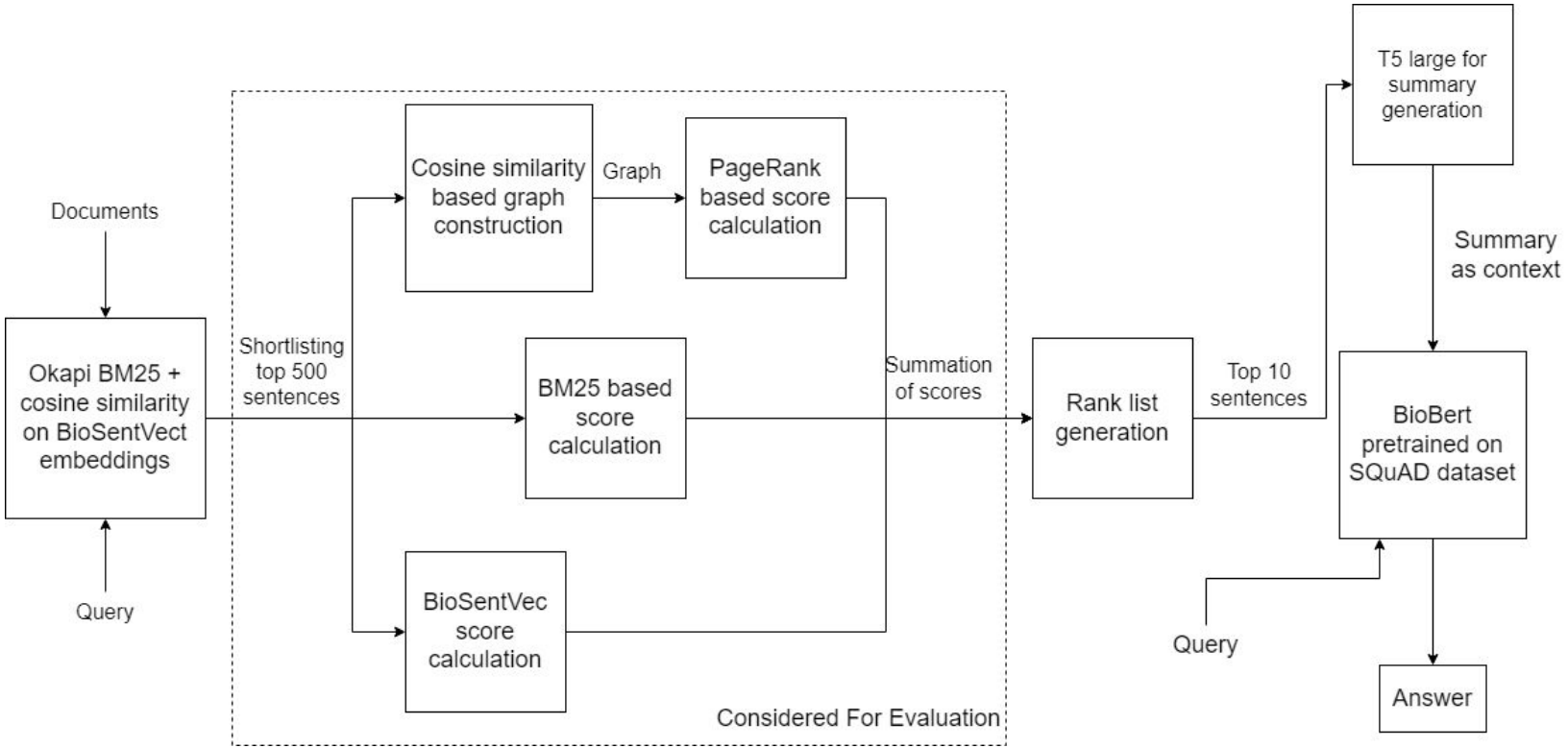
The project is divided into the following sub modules.

- **Search Space Reduction:-** We use Okapi BM25 model scores for each document as well as the cosine similarity using BioSentVec embeddings. Based on the final score for a sentence, which is the arithmetic sum of the BM25 score and the similarity of the BioSentVec embeddings of the sentence and the query, 500 sentences are shortlisted.
- **Further shortlisting using PageRank:-** A graph between the sentences is created where an edge between two sentences exists if the the cosine similarity between two sentences as well as the cosine similarity between the query and both the sentences is above a specific threshold. PageRank scores each of the sentences based on how influential they are. The arithmetic sum of the PageRank score, BM-25 score and BioSentVec scores are used to rank the sentences. The top-10 sentences are shortlisted from the generated rank list.

Proposed Methodology

- **Summarization:-** The T5-large (Text-to-Text transfer transformer) model is used for text summarization. T5-large is used to summarize the top-10 sentences from the previous stage in an abstractive manner to provide a short summary of 10 - 60 words.
- **Question Answering:-** For question answering, a BioBERT model pre-trained on the SQuAD dataset is used. The model takes the input query along with a limited sized context as the input. The generated summary from the previous stage is used as context to the BioBERT question-answering model to get the exact answer for the given query.

Proposed Methodology – Flow Diagram



Work Done

1. **Dataset:**
    a. **COVID-19 Open Research Dataset (CORD-19):** Contains over 4,00,000 COVID-19 related research articles. These are used in building the summarization and QA model.
    b. **CovidQA Dataset**: Question Answering dataset created using CORD-19. It contains (query, document, answer) tuples. 130 such tuples are present. The answers to the queries are from a set of 85 documents from CORD dataset.
2. **Preprocessing:**
    a. All the given documents (articles) are first parsed to extract the abstract, body and ref entries.
    b. The content is then split into sentences using a heuristic algorithm & converted to lower case.
    c. For Okapi BM25, the sentences are further preprocessed using basic normalization, lemmatization and stop words removal pipeline.
    d. For BioSentVec & BertQA, the sentences are used as it is to preserve the semantics of the text.

Work Done

Models/Techniques Used:-

- **Okapi BM25** and **PageRank**
- **BioSentVec:** It is a 700-dimensional sentence embeddings computed using PubMed and MIMIC-III dataset.
- **T5 (Text-to-Text transfer transformer)**: A transformer model that can perform text classification, question answering, machine translation, and abstractive summarization.
- **BioBERT (SQuAD)**: BioBERT is a BERT model pre-trained on biomedical corpus. BioBERT trained on the Standard Question Answering Dataset is used in the given approach.

Evaluation Design & Metrics

The **CovidQA** dataset is used for evaluation of the proposed model on the question answering task. Given a query and a document containing the answer to the query, the system ranks all the sentences. To be deemed correct, the chosen sentence must contain the answer in the form of a substring.

Evaluation Metrics:-

- **Precision@1**: For a single query, P@1 = 1 if the top ranked document in the generated rank list contains the answer else P@1 = 0
- **Recall@3**: For a single query, Recall@3 is the total number of relevant sentences generated in the top 3 of the ranklist divided by the total number of relevant sentences.
- **Mean Reciprocal Rank (MRR)**: MRR for a query is the inverse of rank of the first sentence containing the answer in the generated rank list.

Results

Table 1: Performance on Natural Language Query

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
BioBERT	0.097	0.142	0.170
BioBERT (SQuAD fine tuned)	0.161	**0.403**	**0.336**
BM25 + BioSentVec + PageRank	**0.177**	0.244	0.268

Table 1 shows the comparison of the proposed approach with the baseline models for NL query.

- The proposed model outperforms all baseline models for P@1 metric.
- For R@3 and MRR, the proposed model outperforms all baselines except the SQuAD fine-tuned BioBERT model.

Results

Table 2: Performance on Keyword Query

Model	P@1	R@3	MRR
Random	0.012	0.034	-
BM25	0.150	0.216	0.243
BERT	0.073	0.164	0.187
SciBERT	0.024	0.064	0.094
BioBERT	0.129	0.145	0.185
BioBERT (SQuAD fine tuned)	0.056	0.093	0.135
BM25 + BioSentVec + PageRank	**0.162**	**0.236**	**0.258**

Table 2 shows the comparison of the proposed approach with the baseline models for keyword query.

- The proposed model outperforms all baseline models for all metrics.
- Low performance of BERT based models can be because the query lacks natural language semantics around which they heavily rely on.

Results

Table 3: Performance on Natural Language Query

Model	P@1	R@3	MRR
BioSentVec + PageRank	0.138	0.216	0.247
BM25 + PageRank	0.154	0.215	0.248
BM25 + BioSentVec	0.169	**0.244**	0.264
BM25 + BioSentVec + PageRank	**0.177**	**0.244**	**0.268**

Table 4: Performance on Keyword Query

Model	P@1	R@3	MRR
BioSentVec + PageRank	0.138	0.208	0.248
BM25 + PageRank	0.146	0.215	0.244
BM25 + BioSentVec	0.154	**0.236**	0.254
BM25 + BioSentVec + PageRank	**0.162**	**0.236**	**0.258**

Table 3 & Table 4 show the results of **ablation study** performed for NL and keyword queries.

- Based on the observations, the final model performs optimally.
- Major performance boost is provided by BM-25, followed by BioSentVec and then PageRank.
- The performance drop is maximum without BM-25 and minimum without PageRank.

Results

Query	Summary	Answer
common symptoms of covid 19	the most common symptoms of covid-19 are fever, cough and tiredness. the most common symptoms of covid-19 are fever, dry cough, and malaise/fatigue. covid screening for that visitor depends on resource availability and symptoms.	fever, cough and tiredness
how to mitigate covid transmission?	social measures for mitigating covid transmission, including social distancing, and sheltering in place, have not been examined. risk of transmission during rugby matches is very low but efforts should be made to further mitigate disease transmission within the environment.	travel bans and sheltering in place
various assistance programmes for covid 19	social assistance programmes such as cash transfers or basic income schemes should also be considered. routine surveillance programmes are planned and conducted on a yearly basis. the absence of robust cost data for handwashing programmes makes financial planning and resource allocation difficult.	social assistance programmes such as cash transfers or basic income schemes

Conclusion

- Information retrieval from COVID-19 corpus of documents is a complex task.
- The proposed techniques provide a way to effectively retrieve information from COVID-19 articles catering to both keyword as well as context/semantic similarity techniques.
- Providing summary of the most relevant sentences across all documents thus allows to obtain a quick understanding of the topic from the corpus.

Individual Contribution

Naveen Shenoy:- Literature Survey, Proposed Model, Search space reduction, Evaluation on CovidQA, Question Answering

Pratham Nayak:- Literature Survey, Proposed Model, Search Space reduction, Page Rank, Evaluation on CovidQA, Question Answering, Summarization

Sarthak Jain:- Literature Survey, Proposed Model, Page Rank, Evaluation on CovidQA, Summarization

References

Base paper:-

- Das, D., Katyal, Y., Verma, J., Dubey, S., Singh, A., Agarwal, K., Bhaduri, S., and Ranjan, R. (2020). Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-bert embeddings. In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.

CovidQA Dataset:-

- Tang, R., Nogueira, R., Zhang, E., Gupta, N., Cam, P., Cho, K., and Lin, J. (2020). Rapidly bootstrapping a question answering dataset for covid-19

THANK YOU