CS657: Information Retrieval

Course Project Presentation

COVID INFORMATION RETRIEVAL

Group Number 20

Members:

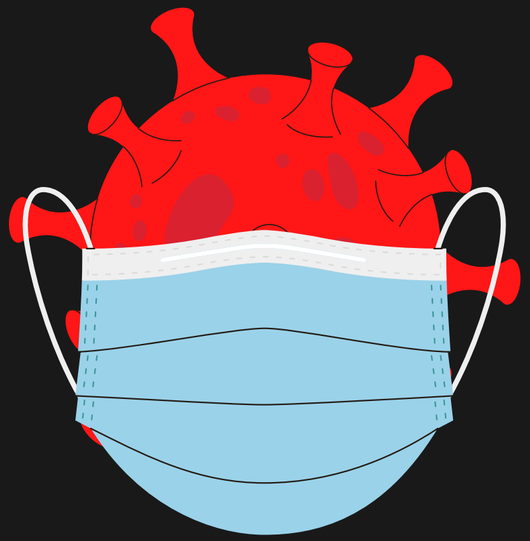
Nitik Jain (170448)

Moksh Shukla (180433)

Aviral Agarwal (180167)

Shubham Gupta (180749)

Archi Gupta (21111014)



CS657 PROJECT PRESENTATION | SPRING 2022

MOTIVATION

- The Coronavirus Disease 2019 (COVID-19) pandemic has resulted in an enormous need and demand for information needs regarding the basic biology, clinical treatment, and public health response to COVID-19.
- Despite the large supply of available scientific evidence, beyond the medical aspects of the pandemic, COVID- 19 has resulted in an “infodemic” as well with large amounts of confusion, disagreement, and distrust about available information.
- This has presented a once-in-a-lifetime opportunity for the information retrieval (IR) and text processing groups to contribute to the response to this pandemic.
- The TREC-COVID Information Retrieval Challenge is one of the challenge which was created to suffice the needs for the same and decided to contribute towards part of the same challenge in extending the modalities of it.

INTRODUCTION

- A common approach for large-scale comparative evaluation of IR systems is the challenge evaluation, with the largest and best-known approach for COVID-19 corpora came from the Text Retrieval Conference (TREC)

- The TREC framework was applied to the COVID- 19 Open Research Dataset (CORD-19), a dynamic resource of scientific papers on COVID-19 and related historical coronavirus research.

- The primary goal of the TREC-COVID Challenge was to build a test collection for evaluating search engines dealing with the complex in formation landscape in events such as a pandemic.

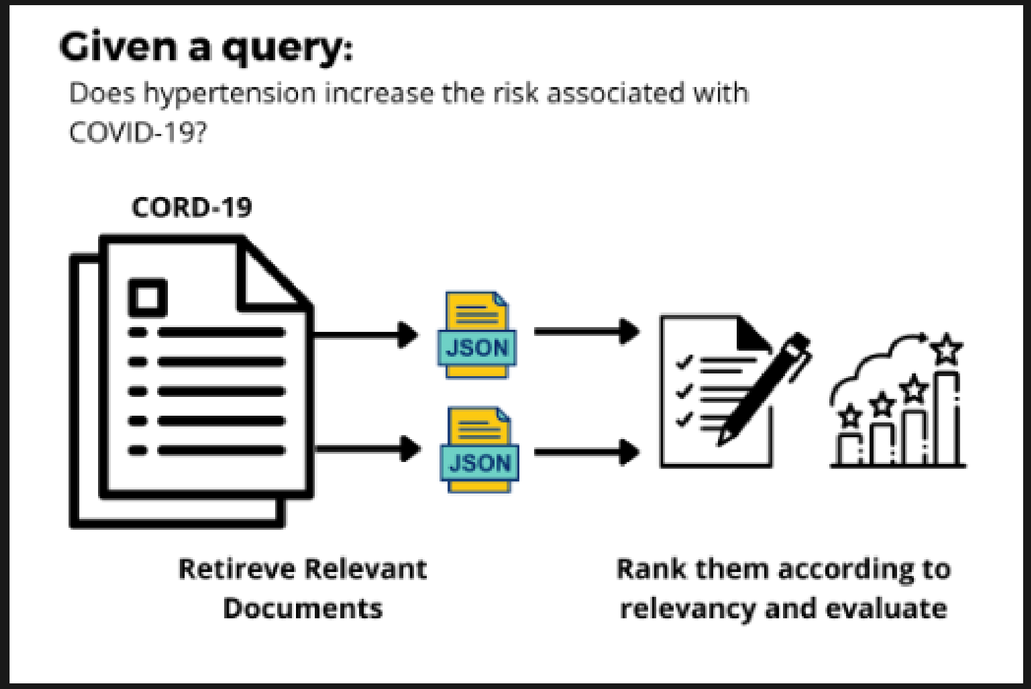
- In this project, we have focused on evaluating various Information Retrieval frameworks such as BM25, Contriever, Bag of Embeddings etc. to rank documents on the basis of relevancy to input queries.

TASK DESCRIPTION

- Systems given a document set and a set of information needs called topics will produce a ranked list of documents per topic where each list is ordered by decreasing likelihood that the document matches the information need (called run).

- In the context of searching for scientific knowledge in the absence of annotated COVID-19 literature, we implement and compare multiple IR methodologies for effective identification of relevant sources to answer COVID related queries posed using natural language.

- We then also compare the results with the partial manually annotated TREC-COVID labels.



DOCUMENT SET DESCRIPTION

- TREC-COVID uses the document set provided by CORD-19. CORD-19 consists of new publications and preprints on the subject of COVID-19, as well as relevant historical research on coronaviruses, including SARS and MERS.

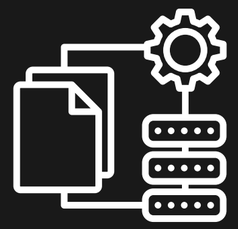
- The April 10, 2020 release of CORD-19, which is used for the first round of TREC-COVID, includes 51K papers, with full text available for 39K. We use complete 51000 documents for training and evaluating our models.

- **Topics - topic file is an xml file that contains all of the topics to be used in the round.**

```
<topic number="1000">
<query>covid effects, muggles vs. wizards</query>
<question>Are wizards and muggles affected differently by COVID-19?</question>
<narrative>
Seeking comparison of specific outcomes regarding infections in
wizards vs. muggles population groups.
</narrative>
</topic>
```

METHODOLOGY

DATA PRE-PROCESSING



RETRIEVAL

Implemented IR models to retrieve documents via relevancy

VALIDATION

PRE-PROCESSING

- The dataset contains 2 types of JSON files: *pdf_json* and *pmc_json*, with a difference in only format. For simplicity, we used only the *pdf_json* form.

- We first cleared the datset from *NaN* values and, extracted the abstract of each document and stored it separately along with the document ID.

- The task corpus consisted of .json parsed files for each research article and we leveraged that to read the abstract, title and the body text as well created a single string for the whole article and moved ahead with the pre-processing and cleaning part.

:

- The further pre-processing and the cleaning involved the following:
  - Lower Casing
  - Email and url removal
  - Non ASCII removal
  - Special Characters removal
  - Stop words removal
  - Stripping extra white space and stemming
  - Tokenization

RETRIEVAL MODELS - BM25

- BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document. It is a family of scoring functions with slightly different components and parameters.

- Given a query Q, containing keywords q1, . . . , qn , the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{DF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{svgdl}}\right)}$$

RETRIEVAL MODELS - CONTRIEVER

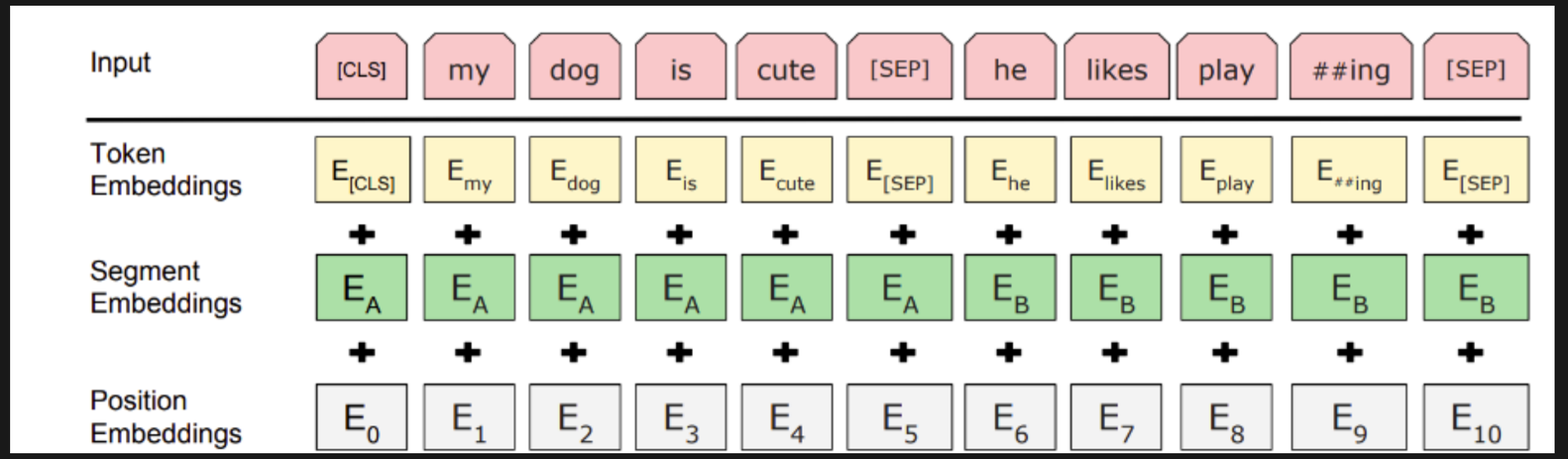
- Contrastive learning is an approach that relies on the fact that every document is, in some way, unique.
- This signal is the only information available in the absence of manual supervision, hence this method is unspervised.
- The resulting algorithm learns by discriminating between documents, using a contrastive loss
- This loss compares either positive (from the same document) or negative (from different documents) pairs of document representations:

$$\mathcal{L}(q, k_+) = \frac{\exp(s(q, k_+)/\tau)}{\sum_{i=0}^{K} \exp(s(q, k_i)/\tau)}$$

Another interpretation of this loss function is the following: given the query representation q, the goal is to recover, or retrieve, the representation k+ corresponding to the positive document, among all the negatives ki .

RETRIEVAL MODELS - BERT-Embeddings

- BERT (Bidirectional Encoder Representations from Transformers) is based on Transformers in which every output element is connected to every input element, and the weightings between them are dynamically calculated based self attention mechanism.
- BERT designed to read in bidirectionally unlike historic language models capability.
- Using this bidirectional capability, BERT is pre-trained on two different NLP tasks: Masked Language Modeling and Next Sentence Prediction.
- We have used a pre-trained BERT embedding to embed our corpus.



EXPERIMENTAL RESULTS

- We extracted top 50 documents for each of the 40 queries using the different models BERT, Contriever and BM25.
- Once the relevant document id's were obtained for the average score (as mentioned above) we ran an intersection for (3C2) pair i.e 3 pairs - BERT & BM25, BERT & Contriever, BM25 & Contriever for each query and took the number of same documents retrieved for each query for each pair of models. At the end we reported the mean and standard deviation for each pair across all queries so that the number of common documents can be normalised across the board.

Model Comparison	BERT-Contriever	BERT-BM25	BM25-Contriever
Mean	36.725	18.1	11.775
Standard Deviation	17.927	13.647	11.014

CHALLENGES AND FUTURE WORK

- Major challenged faced was of computational resources for processing the huge dataset.
- We had to try several approaches, such as taking only the abstract and summary or summarised version of the abstract, or running till a certain amount of documents, but either way, resources were a major constraint.
- The other challenge that was to use Sentence-BERT on a whole paragraph for which we used Pegasus first and then the S-BERT.
- In future we can look to work with sophisticated algorithms along with hybrid models for better retrieval results.
- Fine tuning BERT specifically for the COVID data-set is another thing that we can do.

Thank you!

CS657 PROJECT PRESENTATION | SPRING 2022