Indian Institute of Technology, Kanpur

CS657A - Information Retrieval

COVID Information Retrieval

April 29, 2022

Project Report

Group Number - 20

Submitted by:
Aviral Agarwal (180167)
Moksh Shukla (180433)
Nitik Jain (17807448)
Shubham Gupta (180749)
Archi Gupta (21111014)

Contents

Acknowledgements

We would like to thank all the people who contributed in some way or the other to the the project. First and foremost we offer my sincerest gratitude to Prof. Arnab Bhattacharya, who introduced us to these topics and provided his valuable guidance throughout this project. We would also like to extend our gratitude to the organizers of TREC-COVID Information Retrieval challenge, who provided us with the document set and the relevant queries etc to build our problem statement upon. Lastly, we would like to acknowledge the tremendous need of the CORD-19 dataset, and thank every personnel associated with the same.

1 Motivation

During the last major global pandemic, the 1918-19 influenza ("Spanish Flu"), the information landscape was very different than today: flu viruses had not yet been discovered; worldwide iteracy was considerably lower; information spread largely by word-of-mouth; and the digital content we depend on so greatly today for scientific advancement did not exist from PubMed and preprints to social media. Medically, COVID-19 itself is different: rapidly spreading through many asymptomatic individuals but also having high morbidity and mortality, especially for certain groups, such as the elderly, infirm, and those facing existing health disparities. However, another key difference in this pandemic is the quantity of information, including the use of preprints and rapid publication policies, which has resulted in a scientific corpus that grows by hundreds of COVID-19 articles per day [1].

The Coronavirus Disease 2019 (COVID-19) pandemic has resulted in an enormous demand for and supply of evidence-based information. On the demand side, there are numerous information needs regarding the basic biology, clinical treatment, and public health response to COVID- 19. On the supply side, there have been a vast number of scientific publications, including preprints. 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 [2].

Researchers, physicians, and policymakers working on the COVID-19 response are continually looking for reliable information about the virus and its impact. This presents a once-in-a-lifetime opportunity for the information retrieval (IR) and text processing groups to contribute to the response to this pandemic while also researching strategies for quickly establishing information systems in the event of similar future crises. We discovered the TREC-COVID Information Retrieval Challenge and decided to contribute towards part of the same challenge.

2 Introduction

A key component in identifying available evidence is by accessing the scientific literature using the best possible information retrieval (IR, or search) systems. As such, there was a need for rapid implementation of IR systems tuned for such an environment and a comparison of the efficacy of those systems. The COVID-19 global health crisis has in-turn worsened the problem. The rapid need of information globally led to an exponential surge in published scientific literature. Finding information which is relevant and reliable became need of then hour.

There are many important basic research questions surrounding the use of IR in such a pandemic situation such as finding important IR modalities, effective ways to tailor and improve domain specific search

engines and quantitatively evaluating search engine's performance. In an attempt to tackle the pandemic, extremely large COVID-19–related corpora are being created, sometimes with inaccurate information, which is no longer at scale of human analyses [3].

A common approach for large-scale comparative evaluation of IR systems is the challenge evaluation, with the largest and best-known approach coming from the Text Retrieval Conference (TREC) organized by the US National Institute of Standards and Technology (NIST)[4]. The TREC framework was applied to the COVID- 19 Open Research Dataset (CORD-19)[5], a dynamic resource of scientific papers on COVID-19 and related historical coronavirus research. We describe the dataset in detail in Section 4. 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. For the 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.

In this project, we first use document data from the CORD-19 dataset, and pre-process it using its meta-data. Since the manually classified data categorizing the documents as `relevant`, `partially relevant` and `not relevant` was not available, we decided to compare various IR models to each other. Our baseline model is BM25, where we use compare the relevancy of queries to the document by using the Abstract and Title of document. We then also retrieve the list of relevant documents from other quicker models and compare the common documents to that from each other. Lastly, we also compare our results with those manually labeled as mentioned in TREC-COVID IR Challenge.

3 Task Description

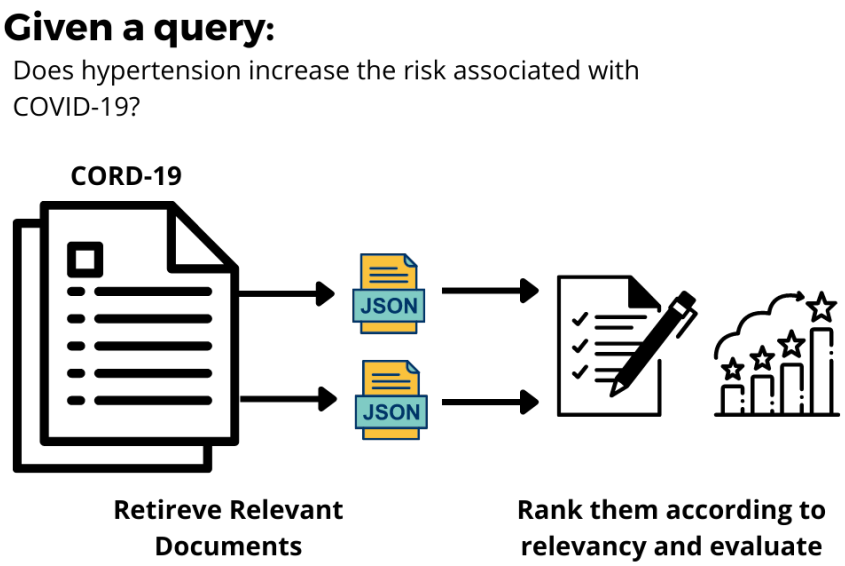

Figure 1: Given an input query, the system identifies relevant papers (yellow highlighted rows) based on supporting evidence from CORD-19

The systems task in TREC-COVID is a classic ad hoc search task. Systems are given a document set and a set of information needs called *topics*. They produce a ranked list of documents per topic where each

list is ordered by decreasing likelihood that the document matches the information need (this is called a *run*). Originally, human annotators will judge a fraction of the documents for relevance, and those relevance judgments will be used to score runs. However, in the context of searching for scientific knowledge in the deluge of COVID-19–related literature, we will implement and compare multiple information retrieval methodologies for effective identification of relevant sources to answer biomedical or any COVID related queries posed using natural language. We then also compare the results with the partial manually annotated TREC-COVID labels.

4 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.

4.1 File Structure:

1. `CORD-19` - folder containing the May 19, 2020 version of CORD-19 documents and metadata.
2. `topics-rnd3.csv` - file containing the *topic-id, query, question,* and *narrative* for each topic.
3. `docids-rnd3.txt` - the list of documents that are can be predicted as relevant; these exclude documents that have been scored in previous rounds for a topic.
4. `qrels.csv` - human annotated relevant and non-relevant categorization of documents for the 40 queries.

4.2 Topics:

The topics used in TREC-COVID have been written by its organizers with biomedical training, inspired by consumer questions submitted to the National Library of Medicine, discussions by medical influencers on social media, and suggestions solicited on Twitter via the #COVIDSearch tag in late March 2020. They are representative of the high-level concerns related to the pandemic. An initial set of 30 topics were created, with 5 new topics being added for each additional round. Hence, we performed our retrieval on a total of **40** topics, which can be found in the `topics-rnd3.csv` file.

The topic file is an xml file that contains all of the topics to be used in the round. The format of a topic is as follows (this is an example topic, not part of the official topic set):

```
<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>
```

Each topic is composed of three fields:

1. `query:` a short keyword query
2. `question:` a more precise natural language question
3. `narrative:` a longer description that further elaborates on the question, often providing specific types of information that would fall under the topic score
4. `average:` additionally, we also create another set of queries which are the average of the embeddings of the above three. This is widely used and accepted in NLP.

Query	Question	Narrative
Coronavirus response to weather changes	How does the coronavirus respond to changes in the weather?	Seeking range of information about virus viability in different weather/climate conditions as well as information related to transmission of the virus in different climate conditions
Coronavirus social distancing impact	Has social distancing had an impact on slowing the spread of COVID-19?	Seeking specific information on studies that have measured COVID-19's transmission in one or more social distancing (or non-social distancing) approaches
Coronavirus outside body	How long can the coronavirus live outside the body?	Seeking range of information on the virus's survival in different environments (surfaces, liquids, etc.) outside the human body while still being viable for transmission to another human
coronavirus asymptomatic	What is known about those infected with Covid-19 but are asymptomatic?	Studies of people who are known to be infected with Covid-19 but show no symptoms?
Coronavirus hydroxy-chloroquine	What evidence is there for the value of hydroxychloroquine in treating Covid-19?	Basic science or clinical studies assessing the benefit and harms of treating Covid-19 with hydroxychloroquine.

Table 1. Few illustrative examples of topics for TREC-COVID task.

5 Methodology

5.1 Pre-processing

In the data-set, for each document, we are provided with two types of JSON files: `pdf_json` and `pmc_json`, which differ only in their format. For simplicity, we used only the `pdf_json` form. We first cleared the data set from NaN values. Then we extracted the abstract of each document and stored it separately along with the document ID. Later, we used this data for various tasks detailed in his report.

For the BM25 model as well as the Bag of Embeddings model the pre-processing done was very similar to the one we did in our class assignments. The corpus consisted of .json parsed files for each research article as mentioned above and we leveraged that to read the abstract, title and the body text as well (only for BM25), created a single string for the whole article and moved ahead with the pre-processing and cleaning part. The pre-processing and the cleaning involved the following things:-

- Lower Casing

- Email and url removal
- Non ASCII removal
- Special Characters removal
- Stop words removal
- Stripping extra white space and stemming
- Tokenizing

For the queries also same pre-processing was done before running the query

5.2 Models

We implemented and compared the following IR models:

5.2.1 BM25 - The Baseline Model [6]

A family of ranking function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. The BM25 term weighting formulas have been used quite widely and quite successfully across a range of collections and search tasks.

The ranking function: 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. One of the most prominent instantiations of the function is as follows. Given a query Q, containing keywords $q_1, \ldots, q_n$, the BM25 score of a document D is:

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

where $f\left(q_i, D\right)$ is q_i 's term frequency in the document $D, |D|$ is the length of the document D in words, and avgdl is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters, usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75\text{IDF}\left(q_i\right)$ is the IDF (inverse document frequency) weight of the query term q_i. It is usually computed as:

$$\text{IDF}\left(q_i\right) = \ln\left(\frac{N - n\left(q_i\right) + 0.5}{n\left(q_i\right) + 0.5} + 1\right)$$

where N is the total number of documents in the collection, and $n\left(q_i\right)$ is the number of documents containing q_i.

5.2.2 Contriever [7]

Information retrieval based on neural networks have obtained state-of-the-art results on datasets and tasks where large training sets are available. However, they do not transfer well to new domains or applications with no training data, and are often outperformed by term-frequency methods such as BM25 which are not supervised. Contriever is a simple contrastive learning based self supervised IR model and is competitive with BM25 .

Contrastive Learning: 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. The resulting algorithm learns by discriminating between documents, using a contrastive loss ([8]). This loss compares either positive (from the same document) or negative (from different documents) pairs of document representations. More formally, given a positive pair of representations (q, k_+) and a set of negative pairs $(q, k_i)_{i=0..K}$, the contrastive InfoNCE loss is defined as:

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

where τ is a temperature parameter. This loss encourages the relevance score of similar examples to be high and that of dissimilar examples to be low. 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 k_i. In the following, we refer to the left-hand side representations in the score s as queries and the right-hand side representations as keys.

Building positive pairs from a single document: A crucial element of contrastive learning is how to build positive pairs from a single input. In computer vision, this step relies on applying two independent data augmentations to the same image, resulting in two "views" that form a positive pair. While contriever primarily consider similar independent text transformation, they also explore dependent transformations designed to reduce the correlation between views. Those explored are:

1. Inverse Cloze Task
2. Independent cropping
3. Additional data augmentation.

Building large sets of negative pairs: An important aspect of contrastive learning is to maintain a large set of negative pairs. Most standard frameworks differ from how the negatives are handled, and the two used here are

1. Negative pairs within a batch.
2. Negative pairs across a batch.

5.2.3 Bag-of-Embeddings [9]

Instead of word being represented as just a single point in the semantic space, the word is represented as a probability distribution that covers the entire semantic space. Specifically, it is mapped to the multivariate normal distribution with spherical covariance. Given a second embedded word, we can measure their similarity using the "overlap" between these two distributions to get a sense for how close the two words are.

5.2.4 BERT Embeddings [10]

Bidirectional Encoder Representation from Transformers (BERT) is a state of the art technique for natural language processing pre-training developed by Google. We will use a pre-trained BERT embedding to embed our corpus.

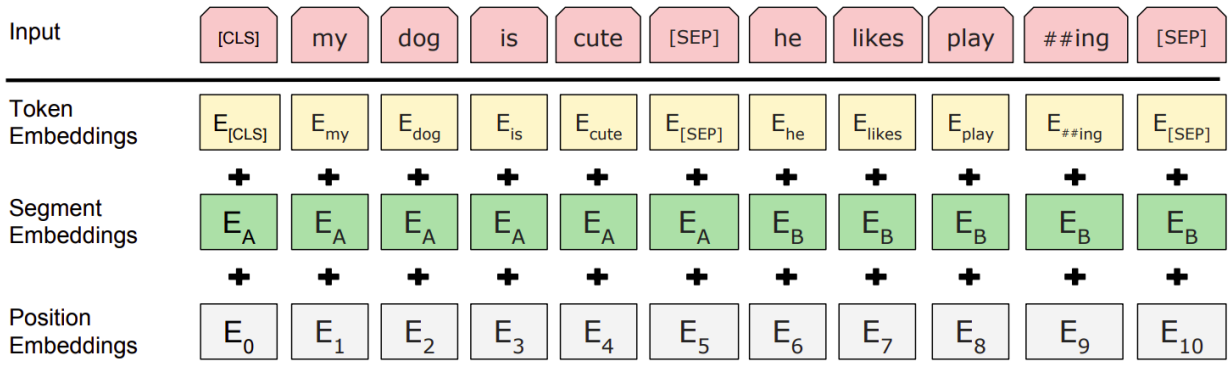

Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings

6 Results

What exactly the below results mean is that we extracted top 50 documents for each of the 40 queries using the different models namely BERT, Contriever and BM25. Once the relevant document id's were obtained for the average score (as mentioned above) we ran an intersection for $\binom{3}{2}$ 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 different queries.

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

Table 1: Document retrieval results for the three models: BERT, BM25, Contriever

7 Challenges and Future Work

Right now the major challenge we faced was of computational resources to process such huge data set, due to which we had to try different things such as taking only abstract and summary or summarised version of the abstract or running till a particular number of documents, but either way computational resource was a major constraint. The next challenge that we faced was to use Sentence-BERT on a whole paragraph for which we used Pegasus first and then the S-BERT.

Coming to future work we can look at more sophisticated algorithms along with hybrid models for better retrieval results. Right now we have used Vanilla models for all three. Fine tuning BERT specifically for the COVID data-set is another thing that we can do.

8 Conclusion

We compared various types of the state of the art information retrieval models. The main aim was to increase the overall data-set, but we could not do it precisely due to some resource constraints. We tried different conventional models, and in the end, we came up with a method to increase the data-set, which is to consider those documents which are retrieved as relevant by maximum models i.e the maximum overlap. The overlaps are the ones mentioned in the results above, and we believe that when three models get the same documents

as the relevant ones, they would be relevant to the whole data-set, and then we can add them to the data-set.

References

[1] Lucy Lu Wang et al. "Cord-19: The covid-19 open research dataset". In: *ArXiv* (2020).

[2] Kirk Roberts et al. "Searching for scientific evidence in a pandemic: An overview of TREC-COVID". In: *Journal of Biomedical Informatics* 121 (2021), p. 103865.

[3] Douglas Teodoro et al. "Information retrieval in an infodemic: the case of COVID-19 publications". In: *Journal of medical Internet research* 23.9 (2021), e30161.

[4] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. Citeseer, 2005.

[5] AIF Ai. "Covid-19 open research dataset challenge (cord-19)". In: *Allen Institute for Artificial Intelligence, https://www. kaggle. com/alleninstitute-for-ai/CORD-19-research-challenge* (2020).

[6] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[7] Gautier Izacard et al. *Towards Unsupervised Dense Information Retrieval with Contrastive Learning*. 2021. DOI: 10.48550/ARXIV.2112.09118. URL: https://arxiv.org/abs/2112.09118.

[8] Zhuofeng Wu et al. *CLEAR: Contrastive Learning for Sentence Representation*. 2020. DOI: 10.48550/ARXIV.2012.15466. URL: https://arxiv.org/abs/2012.15466.

[9] Peng Jin et al. "Bag-of-Embeddings for Text Classification". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. New York, New York, USA: AAAI Press, 2016, pp. 2824–2830. ISBN: 9781577357704.

[10] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.