Pandemic Literature Search: Finding Information on COVID-19

Vincent Nguyen[1,2], **Maciek Rybinski**[2], **Sarvnaz Karimi**[2], **Zhenchang Xing**[1]

Australian National University[1] and CSIRO's Data61[2]

Vincent.Nguyen@anu.edu.au
https://ngu.vin

Background

Motivation

- Finding information in a pandemic scenario presents new challenges as information gradually becomes more available
- Contemporary deep learning LTR (Learning-To-Rank) methods rely on presence of large labeled corpora which is not available in a pandemic search scenario

Dataset

*TREC COVID search The TREC COVID search task was organized, soon after COVID-19 was declared a pandemic, by several institutions, such as NIST and Allen Institute for AI. New search topics were added every few weeks as the search needs of the population changed.

```
Topic 3
Query:      coronavirus immunity
Question:   will SARS-CoV2 infected
            people develop immunity?
            Is cross protection possible?
Narrative:  seeking studies of immunity
            developed due to infection with
            SARS-CoV2 or cross protection
            gained due to infection with
            other coronavirus types
```

Figure 1: A sample topic from the TREC COVID.

Documents

CORD-19 (The Covid-19 Open Research Dataset)[1] is a dataset of research articles on coronaviruses (COVID19, SARS and MERS). It is compiled from three sources: PubMed Central (PMC), the WHO articles, and bioRxiv and medRxiv.

Round	No. Documents	No. Judgments	No. Topics
1	51103	8691	30
2	59851	12037	35
3	128492	12993	40
4	157817	13312	45
5	191175	23373	50

Table 1: Statistics for each TREC-COVID round.

[1]Lu Wang L, et al. CORD-19: The Covid-19 Open Research Dataset. Preprint. ArXiv. 2020;arXiv:2004.10706v2. Published 2020 Apr 22.

Methodology

Main Problems

- In a pandemic "zero-day" scenario, there is no training data for deep learning re-ranking solutions to operate.
- **Solution:** A universal sentence embedding model trained over the unlabeled corpora to estimate document relevance.
- **Compromise:** The model hasn't been trained to rank, thus we need to supplement it with a feature-based ranking/scoring function.

Hybrid Search Model

We use a neural language model known as a sentence transformer, that can compute universal sentence embeddings. We supplement it with a log-normalized BM25 scoring function.

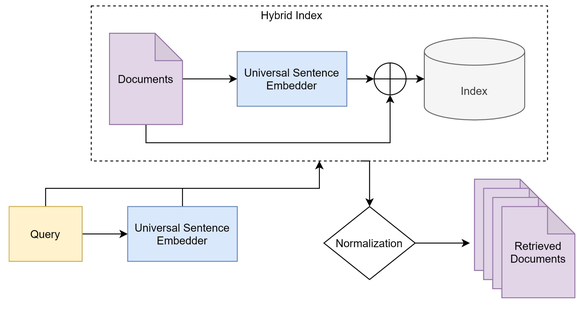
$$\psi(T_i, d) = \log_z(\sum_{t \in T_i} \sum_{f \in d} BM25(t, f)) + \sum_{t \in T_i} \sum_{f \in d} cos(v(t), v(f)), \quad (1)$$

where z is a hyper-parameter, $t \in T_i$ represents fields of the topic (i.e., query, narrative and question), $f \in d$ represents facets of the document (i.e., abstract, title, body), BM25 denotes the BM25 scoring function, $v(t)$ is the neural representation of the topic field, $v(f)$ denotes the neural representation of the document facet, and cos is cosine similarity.
The hyper-parameter z is solved for each topic with the formula:

$$z = \sqrt[R_{cos}]{max(BM25(t, f))} \quad (2)$$

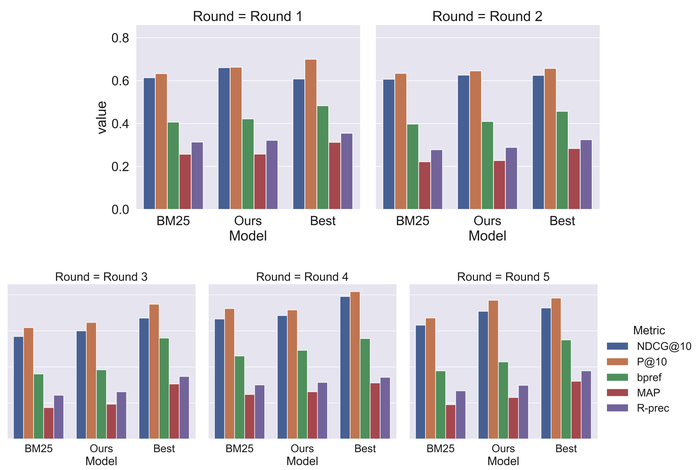
where R_{cos} is the upper range of the summed cosine function:

$$R_{cos} = max(\sum_{t \in T_i} \sum_{f \in d} cos(v(t), v(t))) \quad (3)$$



Hybrid Index flowchart

Results



Key findings

- The neural components finds otherwise undiscovered relevant documents as it can find documents with no word overlap with the search query.
- The neural component acts as a pseudo-reranking model.
  It can efficiently rerank the entire corpus as the main performance penalty is during indexing.
- Although not trained on any additional data, as the document corpus size increases, the performance of the model increases.

Future Work

- Training the model with explicit ranking signals
- Apply to the model to more general tasks such as the biomedical domain

Acknowledgments

This research is supported by the Australian Research Training Program and the CSIRO Postgraduate Scholarship and CSIRO's Future Science platform for Precision Health.