Pandemic Literature Search: Finding Information on COVID-19

Vincent Nguyen[1,2] **Maciej Rybinski**[1] **Sarvnaz Karimi**[1] **Zhenchang Xing**[2]
[1]CSIRO Data61, Sydney, Australia
[2]The Australian National University, Canberra, Australia
{firstname.lastname}@csiro.au
{zhenchang.xing}@anu.edu.au

Abstract

Finding information related to a pandemic of a novel disease raises new challenges for information seeking and retrieval, as the new information becomes available gradually. We investigate how to better rank information for pandemic information retrieval. We experiment with different ranking algorithms and propose a novel end-to-end method for neural retrieval, and demonstrate its effectiveness on the TREC COVID search.[1] This work could lead to a search system that aids scientists, clinicians, policymakers and others in finding reliable answers from the scientific literature.

1 Introduction

As COVID-19—an infectious disease caused by a coronavirus—led the world to a pandemic, a large number of scientific articles appeared in journals and other venues. In a span of five months, PubMed alone indexed over 60,000 articles matching coronavirus related search terms such as `SARS-CoV-2` or `COVID-19`. This volume of published material can be overwhelming. There is a need for effective search algorithms and question answering systems to find relevant information and answers. In response to this need, an international challenge—*TREC COVID Search* (Roberts et al., 2020; Voorhees et al., 2020)—was organised by several institutions, such as NIST and Allen Institute for AI, where research groups and tech companies developed systems that searched over scientific literature on coronavirus. Through an *iterative* setup organised in different rounds, participants are presented with several topics. The evaluations measure the effectiveness of these systems in finding the relevant articles containing answers to the questions in the topics.

We propose a method that improves the systems developed for the TREC-COVID challenge by adopting a novel hybrid neural end-to-end approach for ranking of search results. Our method combines a traditional inverted index and word-matching retrieval with a neural indexing component based on BERT architecture (Devlin et al., 2019). Our neural indexer leverages the Siamese network training framework (Reimers and Gurevych, 2019) fine-tuned on an auxiliary task (unrelated to literature retrieval) to produce universal sentence embeddings. This means that neural indexing can be performed offline for the entire document collection and does not need to be retrained on additional queries. This allows for incorporating the neural component for the entire retrieval process, contrasting with the typical multi-stage neural re-ranking approaches (Li et al., 2020; Zhang et al., 2020; Liu et al., 2017; Wang et al., 2011).

Our method is competitive with the top systems presented in TREC COVID [2]. It improves as corpus size increases despite not being trained on additional data which is a useful property in pandemic information retrieval.

2 Related Work

The use of neural networks in search has mostly been limited to reranking top results retrieved by a 'traditional' ranking mechanism, such as Okapi BM25 (Robertson et al., 1995). Only a portion of top results is rescored with a neural architecture (McDonald et al., 2018). Since the most successful neural reranking models depend on joint modelling of both documents and the query, rescoring the entire collection becomes costly. Moreover, the effectiveness gains achieved with neural reranking are debated (Yang et al., 2019) until recently (Lin, 2019).

Since late 2018, large neural models pre-trained on language modeling—specifically BERT (Devlin et al., 2019) which uses bi-directional transformer

[1] We release our code for the neural index in GitHub: `https://git.io/JkZ7I`

[2] `https://git.io/JkZ7m` Accessed: 10 Oct 2020

architecture—achieve state-of-the-art for several NLP tasks. The architecture is successfully applied to ad-hoc reranking (Nogueira and Cho, 2019; Akkalyoncu Yilmaz et al., 2019; Dai and Callan, 2019).

The existing applications of BERT in search share the limitation of being restricted to reranking because they rely on its next sentence prediction mechanism for a regression score. However, our approach builds on Reimers and Gurevych (2019), where a BERT architecture is trained to produce sentence embeddings. Leveraging these embeddings allows for a cost-efficient application of BERT to neural indexing.

Neural indexing is a less explored field. Whereas Zamani et al. (2018) leverages sparse neural representations for retrieval, Seo et al. (2019) uses sparse and dense representations for learning to rank. These methods rely on networks trained to produce representations directly for ranking documents. For our proposed method, we use universal embeddings[3] generated from transformer encoders trained on an auxiliary task of semantic similarity scoring or Natural Language Inference[4].

3 Dataset

Documents CORD-19 (The Covid-19 Open Research Dataset) (Wang et al., 2020) is a dataset of research articles on coronaviruses (COVID-19, SARS and MERS). It is compiled from three sources: PubMed Central (PMC), the WHO articles, and bioRxiv and medRxiv. Evaluations in subsequent stages (referred to as *rounds*) of TREC COVID Search task are performed on growing snapshots of CORD-19 dataset (Table 1). The collection grew to over 68,000 articles by mid-June 2020. The growth of CORD-19 continues with weekly updates (Roberts et al., 2020).

Topics As part of the TREC COVID search challenge, NIST provides a set of important COVID-related topics. Over five rounds, the topic set is augmented. Round 1 has 30 topics, with five new topics added per subsequent round. Each topic consists of three parts: query, question, and narrative (see Figure 1).

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.

```
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.

Relevance Judgements and Evaluation TREC organises manual judgements per each round of the shared task, using a pooling method over a sample of the submitted runs (Voorhees et al., 2020). Given a topic, a document is judged as: irrelevant (0), partially relevant (1), and relevant (2). As judgements are manually annotated by biomedical experts, only a subset of runs submitted to the track are judged.

The evaluation procedure in each of the subsequent rounds discards (topic, document) pairs included judged in previous rounds. We use this procedure (referred to as *residual scoring*) when comparing against the top-performing runs in the competitive.

In additional experiments, we use *cumulative scoring*, which means evaluating topics for round 2 using human judgments for rounds 1 and 2. Topics of round 3 are evaluated using judgments of rounds 1–3, and so on. Using cumulative scoring allows us to use a larger proportion of judged documents for the topic sets corresponding to subsequent rounds.

Metrics Four precision focused metrics are used to evaluate the rankings: NDCG (Järvelin and Kekäläinen, 2002) at rank 10 (NDCG@10), precision at rank 10 (P@10), mean average precision (MAP) and recall-precision (R-prec). BPref takes into account the noisy and incomplete judgements.

[3] The main property we are interested in for universal embeddings, is that pairs of embeddings can be compared directly via cosine similarity rather than indirectly comparing them through a task-specific network which requires additional training

[4] We do not directly use embeddings as a ranker as they are not trained for retrieval; instead, we use them in combination with a traditional inverted index.

4 Methods

Neural Index Retrieval (NIR) We built a hybrid neural index by appending neural representation vectors to document representations of a traditional inverted index. The neural representations are created using an average over individual representations of sentences (bag-of-sentences) from a BERT-based universal sentence encoder for the title, abstract and full-text facets. Sentence representations are created by averaging token-level representations produced by the encoder (average pooling strategy outlined in Reimers and Gurevych (2019)). We investigate a selection of models derived from applying the training of the Sentence Transformer (Reimers and Gurevych, 2019), a Siamese network built to enable cosine comparability between transformer sentence embeddings, and the biomedically-themed BERT-based pre-trained models, such as BioBERT (Lee et al., 2019). To obtain individual sentences, we use a neural sentence segmentation model, *ScispaCy* (Neumann et al., 2019).

For retrieval, we propose a hybrid approach. We score (topic, document) pairs by combining: (1) Okapi BM25 scores for all pairs of topic fields and document facets; and, (2) cosine similarities calculated for neural representations of all pairs of topic fields (calculated *ad hoc*) and document facets stored in the index[5]. The final score adds a log-normalised sum of BM25 scores to the sum of neural scores. Formally, the relevance score ψ for ith topic T_i and document $d \in D$ is

$$\begin{aligned}\psi(T_i, d) = \log_z(\sum^{t \in T_i} \sum^{f \in d} BM25(t, f)) \\ + \sum^{t \in T_i} \sum^{f \in d} cos(v(t), v(f)),\end{aligned} \tag{1}$$

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:

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

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

$$R_{cos} = max(\sum^{t \in T_i} \sum^{f \in d} cos(v(t), v(t))) \tag{3}$$

The z hyper-parameter normalizes the BM25 score such that its range will be the same as the range of the summed cosine similarity score, R_{cos}. This is to ensure both components, neural and BM25, have equal contribution to the final score.

We also filter by date. The documents created before December 31st 2019 (the first reported COVID-19 case) are removed.

Sentence Embedding Models We compare four different embedding models. We choose our models based on differences in pre-training corpora (PubMed vs. different COVID-specific corpora) and Siamese fine-tuning task (NLI, Natural Language Inference), and STS (Semantic Textual Similarity). We evaluate BioBERT-NLI and BioBERT-STS (pre-trained on PubMed corpus, before COVID), CovidBERT-NLI (pre-trained on a small subset of CORD corpus), and ClinicalCovidBERT-NLI (pre-trained on a larger subset of the CORD corpus)[6].

As a baseline, we use our method with a BioBERT model fine-tuned on an *ad hoc* retrieval task on MS Marco dataset (Nguyen et al., 2016). BioBERT-msmarco is not a universal sentence encoder, and its inclusion is to provide perspective on the significance of using the Siamese fine-tuning in our neural indexing approach. Additionally, we include BM25 as a baseline.

BM25 and top-run baselines For each evaluation round, we report an unmodified BM25 (no neural index) baseline together with a top automatic run (per official *leaderboard*) from the TREC evaluations. Note that the best run baseline does not refer to one specific system, but the best performing run for each round of the evaluation.

5 Experimental Results

We present: (1) a comparison of retrieval effectiveness of our method with different embedding models using cumulative scoring on rounds 1–3 (Table 2); (2) a comparison of our most effective system to the BM25 baseline and best runs from the official shared task evaluations using residual

[5] We emphasise that our model is not a re-ranking model but a ranker model as it scores the entire collection during retrieval, rather than re-ranking a retrieved list.

[6] A directory to the models: `https://git.io/JTfz2` Accessed: 10 Oct 2020

Model	Round 1					Round 2					Round 3				
	NDCG@10	P@10	bpref	MAP	R-prec	NDCG@10	P@10	bpref	MAP	R-prec	NDCG@10	P@10	bpref	MAP	R-prec
BM25 (baseline)	0.624	0.650	0.409	0.258	**0.316**	**0.666**	0.694	0.380	0.240	0.304	0.717	0.762	0.412	0.235	0.313
BioBERT-NLI	0.614	0.597	0.384	0.219	0.279	0.608	0.671	0.374	0.219	0.291	0.726†	0.772	0.410	0.237	0.311
Covid-NLI	0.582	0.597	0.409	0.249	0.309	0.522	0.597	0.347	0.193	0.274	0.736	0.780	0.413	0.239	0.316
ClinicalCovid-NLI	**0.641**†	**0.663**	**0.408**	**0.258**	0.315	0.650	**0.710**	**0.397**	**0.264**	**0.320**	**0.739**	**0.780**	**0.420**‡	**0.252**‡	**0.327**
BioBERT 1.1 STS	0.612	**0.633**	**0.408**	0.246	0.302	0.613	0.663	0.396	0.251	0.314	0.722	0.762	0.398	0.228	0.302
BioBERT msmarco	0.528	0.530	0.366	0.197	0.255	0.593	0.666	0.372	0.232	0.303	0.691	0.743	0.389	0.218	0.296

Table 2: Results for our runs for Round 1–3. Best run is as reported by organisers per that round. † denotes statistical significance at 95% and ‡ at 99% over the baseline.

Model	Round 1					Round 2					Round 3					Round 4					Round 5				
	NDCG@10	P@10	bpref	MAP	R-prec	NDCG@10	P@10	bpref	MAP	R-prec	NDCG@10	P@10	bpref	MAP	R-prec	NDCG@10	P@10	bpref	MAP	R-prec	NDCG@10	P@10	bpref	MAP	R-prec
BM25	0.614	0.633	0.407	0.257	0.314	0.607	0.634	0.398	0.222	0.278	0.569	0.618	0.361	0.174	0.243	0.666	0.724	0.461	0.247	0.300	0.632	0.672	0.378	0.190	0.267
ClinicalCovid-NLI	**0.661**	0.663	0.422	0.258	0.322	**0.626**	0.646	0.410	0.228	0.289	0.600	0.647	0.384†	0.193†	0.262	0.685	0.716	0.492‡	0.262	0.314	0.709‡	0.770‡	0.428	0.230‡	0.298‡
Best Automatic Run	0.608	**0.700**	**0.483**	**0.313**	**0.355**	0.625	**0.657**	**0.457**	**0.284**	**0.325**	**0.671**	**0.748**	**0.560**	**0.305**	**0.347**	**0.791**	**0.818**	**0.557**	**0.311**	**0.342**	**0.727**	**0.782**	**0.550**	**0.320**	**0.378**

Table 3: Our proposed method with comparison to a BM25 baseline and the top automatic run for that round. All evaluation is performed with residual document scoring

scoring on rounds 1–5 (Table 3); and, (3) an ablation test of our most effective system on round 5 topics using cumulative scoring (Table 4).

Choice of the embedding model Table 2 provides insights into the selection of the sentence embedder: (1) the importance of domain-adaptive pre-training for neural re-ranking, that is using a model pre-trained on a task-specific corpus. We believe it is especially important in our setup, as there is no other task-specific training involved at any stage. Unsurprisingly, using a larger domain-specific corpus in pre-training yields better results; (2) there is no apparent difference between NLI and STS fine-tuning. Notably, BioBERT-msmarco performs worse than other evaluated models and the baseline, showing the importance of adapting BERT to act as a universal sentence encoder at the fine-tuning stage.

Ablations Table 4 confirms that the combination of BM25 with the neural indexing yields best overall results as removing either component leads to a significant loss in performance. Removal of facets makes no significant differences. Removal of the date filter significantly degrades NDCG@10.

Comparisons with best runs Aside from rounds 3 and 4, our models remain competitive with the top run. Our model scored higher NDCG@10 for rounds 1 and 2 over the baseline automatic runs. Most of the top runs used neural re-rankers which have been specifically trained on related tasks such as med-marco (MacAvaney et al., 2020).

Where does the model succeed or fail? Our model consistently outperforms the BM25 baseline (Table 3).

Model	**P@10**	**NDCG@10**
NIR	0.852	0.796
no neural	0.744†	0.808†
no BM25	0.668‡	0.706‡
no title	0.848	0.784
no abstract	0.848	0.785
no fulltext	**0.856**	**0.799**
no date filter	0.834	0.775†

Table 4: Ablation studies for our proposed method where document facets, query facets and other aspects of the model are removed.

The model can retrieve documents undiscovered by the BM25 component or a pipeline model which uses word-overlap scoring in its initial retrieval. It computes scores over the entire collection as a hybrid inverted index which leads to an average increase of in 6% R-prec values (Table 3) over the BM25 baseline. The improvement in early recall is also a desirable feature if we were to pair our model with a task-specific neural re-ranker.

We expect that the top ranked documents are scored highly by both components, however, we found that our model placed an irrelevant document at the rank one for Topic 3. This document was scored highly by BM25 but much lower in the neural/cosine component. It saturated the scoring function as it repeated many of the keywords in the query, however, the semantic content of the text was irrelevant to the query itself as it discussed "coronavirus crossing continents" rather than "coronavirus cross protection".

On the other hand, for topic 1, "coronavirus origins", we found that the neural index overcame semantic mismatches of the BM25 scoring. In the dataset, most documents are related to coronavirus, the word “origin” contributes more to the final score and BM25 retrieved an irrelevant document at rank three which is a document that discusses origins of a different virus. However, when using the neural scorer, this document is placed at rank 42.

From Table 2 and 3, although our model is not trained on any additional data, it improves in ranking as the corpus size increases. This is a useful property in pandemic information retrieval as the model does not need to be continually retrained, and each document is embedded once.

6 Conclusions

We propose a novel neural ranking approach (NIR) for pandemic information retrieval. Experimenting with the TREC COVID search challenge, we show that our method is competitive compared to other automatic systems. We show that a neural scoring is beneficial in alleviating some of the shortcomings of the keyword-based retrieval. Empirically, our model shows improvements with time in a pandemic scenario without additional training data. A balanced scoring function combines the strengths of the inverted and neural indices. A neural index explicitly trained for ranking would be a suitable avenue for future research.

Acknowledgements

Vincent is supported by the Australian Research Training Program and the CSIRO Research Office Postgraduate Scholarship. This work is funded by the CSIRO Precision Health Future Science Platform.

References

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *EMNLP*, pages 3490–3496, Hong Kong, China.

Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *SIGIR*, pages 985–988, Paris, France.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186, Minneapolis, MN.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: passage representation aggregation for document reranking. *arXiv:2008.09093*.

Jimmy Lin. 2019. Neural hype, justified! A recantation. *ACM SIGIR Forum*, 53.

Shichen Liu, Fei Xiao, Wenwu Ou, and Luo Si. 2017. Cascade ranking for operational e-commerce search. In *SIGKDD*, page 1557–1565, New York, NY, USA.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. SLEDGE: A simple yet effective baseline for COVID-19 scientific knowledge search. *arXiv:2005.02365*.

Ryan McDonald, George Brokos, and Ion Androutsopoulos. 2018. Deep Relevance Ranking Using Enhanced Document-Query Interactions. In *EMNLP*, pages 1849–1860, Brussels, Belgium.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *BioNLP*, pages 319–327, Florence, Italy.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *NIPS Cognitive Computations Workshop*, volume 1773, Barcelona, Spain.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv:1901.04085*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*, pages 3982–3992, Hong Kong, China.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William Hersh. 2020. TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19. *J. Am. Med. Inform. Assoc.*

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *TREC*, Gaithersburg, MD, US.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *ACL*, pages 4430–4441, Florence, Italy.

Ellen Voorhees, Alam Tasmeer, Demner-Fushman Dina, Hersh William, and Kyle Lo. 2020. TREC-COVID: Constructing a pandemic information retrieval test collection. *ACM SIGIR Forum*, 54:1–12.

Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *SIGIR*, page 105–114, Beijing, China.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. In *ACL NLP-COVID Workshop*, Online.

Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *SIGIR*, pages 1129–1132, Paris, France.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *CIKM*, page 497–506, Torino, Italy.

Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, and Jimmy Lin. 2020. Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 open research dataset. *arXiv:2007.0784*.