CSIROmed Team Report of TREC 2021 Clinical Trials track: Experiments with BERT Reranking Methods

Maciej Rybinski Vincent Nguyen Sarvnaz Karimi
CSIRO Data61
Sydney, Australia
firstname.lastname@csiro.au

ABSTRACT

A large body of clinical trials fail to attract enough eligible participants. TREC 2021 Clinical Trials track set a task to use patient data in the form of clinical notes as a way to identify patients eligible for clinical trials. We explore a number of reranking methods as well as query rewighting using Bidirectional Encoder Representations from Transformers (BERT). Our best method used BERT reranking trained on scientific literature.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; *Language models*; *Decision support systems*; • **Applied computing** → *Health informatics*.

KEYWORDS

Clinical trials search; Medical information retrieval; Learning-to-rank; Evidence-based medicine

ACM Reference Format:
Maciej Rybinski Vincent Nguyen Sarvnaz Karimi. 2021. CSIROmed Team Report of TREC 2021 Clinical Trials track: Experiments with BERT Reranking Methods. In *TREC'21: TREC, November, 2021.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

TREC Clinical Trials (CT) is the first edition of the track run in TREC 2021. Previously held biomedical retrieval track, TREC Precision Medicine, introduced clinical trial retrieval tasks [8, 10, 11] in its 2017, 2018, and 2019 editions. New in the 2021 edition, the task focuses on retrieval of clinical trials given EHR extracts (ER admission statements; 5-10 sentences) for synthetic patients. Specifically, the task is one of matching trials to a given patient, where patient is represented exclusively with a free text extract of their EHR.

2 DATASET

The TREC 2021 CT dataset consists of 75 topics with 35,832 manual judgments. The corpus for the task is a 2020 snapshot of ClinicalTrials.gov database[1], with over 375K registered clinical trial records. Each topic contains a narrative simulating a patient's admission note. An example topic is shown in Figure 1.

```
A 2-year-old boy is brought to the emergency department by his parents for 5 days of high fever and irritability. The physical exam reveals conjunctivitis, strawberry tongue, inflammation of the hands and feet, desquamation of the skin of the fingers and toes, and cervical lymphadenopathy with the smallest node at 1.5 cm. The abdominal exam demonstrates tenderness and enlarged liver. Laboratory tests report elevated alanine aminotransferase, white blood cell count of 17,580/mm, albumin 2.1 g/dL, C-reactive protein 4.5 mg, erythrocyte sedimentation rate 60 mm/h, mild normochromic, normocytic anemia, and leukocytes in urine of 20/mL with no bacteria identified. The echocardiogram shows moderate dilation of the coronary arteries with possible coronary artery aneurysm.
```

Figure 1: An example topic.

Relevance judgments assign a score of 0 for *not relevant*, 1 for *excluded* and 2 for *eligible* to a topic-document pair.

For training our runs using supervised learning re-ranking models, we use past TREC Clinical Decision Support (CDS) 2014-16 collections for literature retrieval (2017-2019) with similarly defined EHR extracts [9, 12, 15] (hereafter referred to collectively as *training collections*). The key difference between these training collections and the 2021 dataset is that CDS track focused on retrieval of relevant scientific literature abstracts (with PubMed snapshots), rather than searching clinical trials. We also used the dataset introduced by Koopman and Zuccon [3] as an auxiliary source of training and evaluation data. This dataset uses topics from TREC CDS 2014-2015, but search task and a shallower pool of relevance judgments are defined over a historic snapshot of the ClinicalTrials.gov database.

3 METHODS

Our submission broadly addresses the applicability of established neural retrieval methods to TREC CT task. Our submission for the official evaluations included the following runs:

- CSIROmed_DCT—DeepCT-query [2] method applied to the topic texts (i.e., the clinical notes) for query terms (neural) re-weighting; DeepCT-query model was trained on TREC CDS collections. We used BM25 for the retrieval with reweighted queries.

[1] http://clinicaltrials.gov/

TREC 2021, November, 2021, Online
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00

- CSIROmedNIR—our neural indexing method where scores are derived from an interpolation of BM25 (sparse) scores and cosine similarities between universal sentence embedding vectors obtained with Sentence-BERT [7]; this run follows our approach originally applied in 2020 COVID-TREC evaluation [5].
- CSIROmed_inc—a MonoBERT [6] style pipeline with Divergence From Randomness (DFR) used in initial retrieval step and a pointwise BlueBERT reranker used to re-score top 100 documents; inclusion criteria CT field is used to represent documents in the reranker input.
- CSIROmed_abs—a MonoBERT style pipeline with DFR initial retrieval step and a pointwise BlueBERT reranker used to re-score top 100 documents; 'brief summary' field is used to represent documents in the reranker input.
- CSIROmed_brd—a Borda rank fusion of the four other runs and an additional baseline run (DFR retrieval without reranking).

We also report results for the DFR baseline for reference. Within our official runs we cover a selection of neural approaches relevant to this year's TREC CT track.

We use DeepCT-query in this setting for two reasons. Firstly, it is an attempt to reweight/prune long clinical narratives, which were already shown to result in poor effectiveness when used as queries in the context of historic TREC CDS tasks. Secondly, we evaluate this method in clinical settings and with limited training data, which provides additional insight on the adaptability of DeepCT-query to specialised domains.

Similarly, with the CSIROmedNIR run we attempt to improve the initial retrieval step with a BERT-based encoder. The method has already proved successful in a scientific retrieval task with natural language questions, so our experiment tests its applicability to clinical narratives.

The MonoBERT-style runs re-use a pipeline successfully employed in TREC PM 2020 [13], where it was used successfully in a zero/few shot setting. In TREC CT experiment we expect these runs to provide a strong neural baseline that can be potentially used in combination with improvements in the initial retrieval.

Borda rank fusion run was intended as a simple way to combine the results of the very different methods used in the remaining runs.

Indexing. We index clinical trials with the following fields: brief summary, brief title, clinical trial ID, detailed description, drug name, drug keywords, exclusion, gender, general keywords, inclusion, intervention type, maximum age, minimum age, official title, and primary outcome. Age-related fields are numeric, all other fields except clinical trial ID are copied into an aggregate *text* field, which is also indexed.

CSRIOmed_DCT. In DeepCT-query a BERT model takes a query as input and returns a vector of weights (one weight for each of the query terms). These weights are used as term weights in a subsequent sparse retrieval step with BM25, with documents indexed as described in the previous paragraph. Our DeepCT-query model was trained on TREC CDS data. In the original implementation a BERT model is trained as a regressor for query term recall signal, defined for term t as $r_t = \frac{|d_{pos}(t)|}{|d_{pos}|}$, where d_{pos} denotes a set of documents relevant to a given input query, and $d_{pos}(t)$ denotes a subset of these documents that contain term t. The only important difference between our experiment and the methodology outlined in the original paper is that in our experiments we incorporate signal from non-relevant documents d_{neg} into the regression training target for term t as follows: $r_t = max(0, \frac{|d_{pos}(t)|}{|d_{pos}|} - \frac{|d_{neg}(t)|}{|d_{neg}|})$. The sparse retrieval step is carried out using BM25 scoring with the default parameters ($k1 = 1.2$, $b = 0.7$) over the aggregated *text* field.

CSIROmedNIR. We propose an end-to-end retrieval framework using a hybrid scoring function from [5] consisting of: (1) a cosine similarity over textual fields using a universal sentence encoder; and, (2) BM25 scoring. Due to the lack of training data, we do all our tuning and pruning on a test collection for matching patients to clinical trials [3]. We tune BM25 maximising recall precision giving two hyperparameters $b = 0.9$ and $k1 = 1.4$. Due to the availability of many fields, we prune fields based on recall precision using a leave-one-out approach to estimate significance. We do this pruning independently for BM25 and cosine similarity. Our final scores for each document and query is generated based on the hyperparameters and fields found during pruning/training. Log-normalization is used to combine cosine similarity scores and BM25 scores.

CSIROmed_abs. We propose a two-step search framework consisting of: (1) an initial retrieval using DFR [1]; and, (2) neural re-ranking with BioBERT [4] for the top 100 documents of the initial ranking. We adapt the re-ranking setup proposed by Rybinski et al. [14] for clinical trials retrieval in TREC PM task. Specifically, as a neural relevance scorer we use an output of a fine-tuned BlueBERT with binary linear layer connected to the encoder's pooled layer with dropout. BlueBERT is a domain-adapted BERT variant with pre-training on a snapshot of PubMed and MIMIC-III clinical notes. The re-ranking model is fine-tuned using cross-entropy loss and a pointwise re-ranking approach, so we essentially train a binary classifier on binarized human judgments from 2014—2016 TREC CDS datasets (an abstracts retrieval task). For search we use a softmax over the classifier output as BERT-based relevance score. In training (with the CDS topics and documents) we represent the queries (Sentence A inputs) with the *description* topic field (the longest narrative), and the documents (Sentence B inputs) with a title and abstract. In inference A input is the entire topic and B input is a concatenation of brief titles and brief summaries of the respective clinical trials. Inputs are capped to the maximum token length for BERT variants (512 token). For a final score DFR scores

Run	Description
abs	BERT reranking trained on scientific literature
brd	Majority voting
DCT	Query re-weighting with BERT
inc	BERT reranking trained on literature applied to inclusion criteria
NIR	Neural indexing
DFR	DFR baseline run (post-TREC)

Table 1: CSIROmed automatic runs with their short definitions.

	Metrics		
Run	NDCG@10	RR	P@10
NIR	0.2281	0.2925	0.1240
DCT	0.3658	0.4022	0.2133
brd	0.4777	0.5034	0.2800
abs	0.5285	**0.5114**	**0.3240**
inc	**0.5320**	0.5007	0.3173
DFR	0.4196	0.4023	0.2240
Automatic	0.3040	0.2942	0.1613
Manual	0.6212	0.7213	0.4573

Table 2: Comparison of submitted CSIROmed automatic runs and a post-TREC DFR baseline against TREC median (mean of medians for individual topics) on both automatic and manual runs as shown in the last two rows. Bold numbers indicate the highest values in our runs.

are interpolated with the BERT-based scoring (in a 1:9 ratio for normalised scores).

CSIROmed_inc. The pipeline for CSIROmed_inc is identical to that of CSIROmed_abs in its general outline. The main difference here is that the model is trained on the auxiliary training dataset and documents (here, clinical trials for both training and inference) are represented with a concatenation of their brief titles and inclusion criteria.

CSIROmed_brd. For the rank fusion we use a Dowdall system, i.e. each document-topic pair is scored as a sum of inverse ranks across initial rankings. As initial rankings we used the remaining 4 runs, a DFR baseline and another run similar to CSIROmed_inc, but trained on exclusion criteria. As in all runs, we return 1000 documents per topic.

4 RESULTS

A comparison of our runs against TREC median across automatic and manual runs of all the teams is shown in Table 2. While all our submitted runs are automatic, the manual median values are listed as a potential upper-bound of what can be achieved with human-in-the-loop. Following the official evaluation in this track, we report NDCG@10 calculated over graded judgments as the main evaluation metric. RR and P@10 are calculated over binarised judgments (*eligible* only).

From our five submitted runs, all but the NIR run achieved scores above the TREC median for automatic runs. Strongest runs, however, were CSIROmed_abs and CSIROmed_inc.

REFERENCES

[1] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *TOIS* 20, 4 (2002), 357–389.

[2] Zhuyun Dai and Jamie Callan. 2019. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687* (2019).

[3] Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 669–672.

[4] 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 (09 2019), 1234–1240.

[5] Vincent Nguyen, Maciej Rybinski, Sarvnaz Karimi, and Zhenchang Xing. 2020. Pandemic Literature Search: Finding Information on COVID-19. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*. 92–97.

[6] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv:1901.04085* (2019). arXiv:1901.04085 [cs.IR]

[7] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*. Hong Kong, China, 3982–3992. https://www.aclweb.org/anthology/D19-1410.pdf

[8] Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, William R. Hersh, Steven Bedrick, Alexander Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *TREC*. Gaithersburg, MD.

[9] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. 2016. Overview of the TREC 2016 Clinical Decision Support Track. In *TREC*. Gaithersburg, MD.

[10] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, and Alexander J. Lazar. 2018. Overview of the TREC 2018 Precision Medicine Track. In *TREC*. Gaithersburg, MD.

[11] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, Shubham Pant, and Funda Meric-Bernstam. 2019. Overview of the TREC 2019 Precision Medicine Track. In *TREC*. Gaithersburg, MD.

[12] Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, and William R. Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track. In *Text REtrieval Conference*. Gaithersburg, MD.

[13] Maciej Rybinski and Sarvnaz Karimi. 2021. Will Sorafenib help? Treatment-aware Reranking in Precision Medicine Search. In *Proceedings of the 30th International ACM Conference on Information and Knowledge Management (CIKM)*.

[14] Maciej Rybinski, Jerry Xu, and Sarvnaz Karimi. 2020. Clinical trial search: Using biomedical language understanding models for re-ranking. *Journal of Biomedical Informatics* 109 (2020), 103530.

[15] M. Simpson, E. Voorhees, and W. Hersh. 2014. Overview of the TREC 2014 Clinical Decision Support Track. In *TREC*. Gaithersburg, MD.