Matching a Patient from An Admission Note to Clinical Trials: Experiments with Query Generation and Neural-Ranking

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

ABSTRACT

Many clinical trials fail to attract enough eligible participants. The TREC 2022 Clinical Trials track set a task where patient data, in the form of clinical notes, can be used to match eligible patients to a relevant clinical trial. We explore a number of dense retrieval methods using Bidirectional Encoder Representations from Transformers (BERT). Our best method used BERT reranking using models based on monoBERT architecture. Our self-supervised monoBERT run achieved effectiveness competitive to that of a fully-tuned monoBERT run.

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:
Vincent Nguyen   Maciej Rybinski   Sarvnaz Karimi. 2021. Matching a Patient from An Admission Note to Clinical Trials: Experiments with Query Generation and Neural-Ranking. In *TREC'22: TREC, November, 2022.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Clinicial trials are vital to the development of new treatments and benefit both the patient and the wider medical community [2]. However, there are requirements to ensure that clinical trials enrol many patients to ensure the study can draw reliable conclusions [12], and to ensure that the treatments work for a wider population demographic. However, few patients are presented with opportunities to join clinical trials and fewer agree to participate [11]. This is a result of patients not being exposed to clinical trials that they are eligible for, as often their physician will only actively recruit the patients to trials where they are an investigator [10]. There is a necessity for the development of tools to increase participation in clinical trials as up to 80% of clinical trials fail to reach the required participation numbers.

The TREC Clinical Trials (CT) 2022, is the second edition of the yearly track. Previously held biomedical retrieval tracks, TREC Precision Medicine, introduced clinical trial retrieval tasks [4–6] in its 2017, 2018, and 2019 editions. The task of this year's track is to link a synthetic patient's electronic health record (EHR), in free text, to relevant clinical trials. TREC CT's goal is to study the use of automatic retrieval systems to expose patients to relevant clinical trials to increase participation.

In our submission to this year in the TREC CT track, we build upon our last year's submission [7]. Our experiments this year focus on neural ranking using resource-effective self-supervision and supervision signals from last year's judgement pool. Our experiments with a neural reranking pipeline centered around resource-effective learning, used a reranker trained on labeled data (from last year's edition of the track) compared with a self-supervised model trained using the target document corpus. We also experiment with efficient end-to-end neural ranking (where document representations can be pre-computed) with bi-encoders and with neural query expansion. Finally, we also probe the effect of a simple heuristic for matching the patient note with the demographic profile specified in the clinical trials, which we apply to one of the bi-encoder runs.

2 DATASET

The TREC 2022 CT dataset consists of 50 topics with 35,394 relevance judgments. The corpus for the task is a 2020 snapshot of ClinicalTrials.gov database[1], with over 375K registered clinical trial records. Each topic simulates a patient's admission note. An example topic is shown in Figure 1.

For each topic-document pair in the dataset, a relevance judgment assigns a score of 0 for *not relevant*, 1 for *excluded* and 2 for *eligible*.

For training our runs using supervised learning reranking models, we use the TREC Clinical Trials 2021 track collection as in-domain training data. We also have submitted runs with neural rerankers trained exclusively using the document collection (i.e., the clinical trials corpus). We have also included a baseline run with automated query expansion, based on the methodology successfully applied in TREC CT by Pradeep et al. [3] as their initial retrieval step.

TREC 2022, November, 2022, Online
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00

[1] http://clinicaltrials.gov/

Topic 8: A 7-month-old boy is brought to emergency by his parents due to irritability and inability to defecate for the past 3 days. The patient has had constipation and discomfort with bowel movements since birth. His symptoms worsened after eating semi-solid foods. Vital signs are normal. Abdominal examination shows distension and tenderness to palpation with presence of bowel sounds. Xray with barium shows a narrow rectum and rectosigmoid area. The rest of the colon proximal to this segment is dilated. Digital rectal exam revealed burst of feces out of the anus. The biopsy showed absence of submucosal ganglia in the last segment of the large intestine.

Topic 38: A 60-year-old man comes to the clinic complaining of hand tremor that started few months ago. It is most bothering when he wants to drink from a glass or pour from a bottle. He does not smoke, but drinks occasionally. He recently started consuming more alcohol as his tremor subsides somewhat when he drinks small amounts of alcohol. Family history is significant for similar problems in his mother. Vital signs are normal and the patient has no other medical conditions. Neurologic examination shows bilateral tremor in the upper extremities. The diagnosis of essential tremor is confirmed.

Figure 1: Two example topics from the TREC CT 2022 track.

3 METHODS

3.1 MonoBERT baseline (monobert500 run)

The MonoBERT baseline is a BM25 initial ranking followed by reranking top 500 documents per topic with a MonoBERT-style [1] neural reranker. For the initial BM25 retrieval we have used a setup based on our previous experiments [9].

The reranker was initialised from a SciBERT checkpoint (allenai/scibert_scivocab_uncased) and fine-tuned using binarised relevance scores of the TREC CT 2021 collection. Following our previous experience in clinical trials retrieval [9], we interpolate the normalised BM25 score with a softmax over BERT's output at a 1:9 ratio.

One point of difference to a typical MonoBERT set-up is the order of input sequences—in our experiments for TREC CT 2022 document representations are fed to the model (both in training and inference) as Sentence A, while the query (so, the medical note) is presented as Sentence B part of the input. Our preliminary results have shown that it does not hurt the model's effectiveness, and it gives us a more direct comparison to the self-supervised run. A brief_summary field of the clinical trials was used as input representing the document (Sentence A).

3.2 Self-supervised MonoBERT (zs_bet_500 run)

The self-supervised MonoBERT follows the setup already described for the MonoBERT baseline. The core difference is that the reranker model has been trained without the TREC CT 2021 as training data. That is, the reranking model for this run was trained for predicting relevance for brief_summary—inclusion_criteria pairs drawn from the ClinicalTrials.gov snapshot. Labels for this classification task were defined as 1 for the summary–inclusion pairs coming from the same document and 0 for the pairs coming from different documents. We drew the positive pairs from the entire corpus. The negative instances were sampled randomly in a 2:1 ratio.

3.3 Contrastive Learning with Bi-Encoders

The Bi-Encoder setup, CSIROmedANIR run, used was a SqueezeBERT model with an embedding space warm-started by a version of MS Marco filtered through MetaMap. To achieve this, we kept the subset of queries, around 100,000, which contained biomedical entities found by MetaMap and used this subset for representation learning for one epoch. The model was then tuned on TREC CT 2021 using a triplet loss function where the query's cosine distance is reduced to a relevant document and maximised from an irrelevant document. We used SqueezeBERT as preliminary results indicated this was a strong candidate over other larger language models. Similar to our previous year's submission, we use a log normalised sum with BM25 and cosine similarity between query and document as the final ranking score. The scoring is done end-to-end on the search engine node with no reranking step.

We submitted a second run of this method, ANIR_demo, where we reranked the final results using eligibility criteria based on age and sex. Clinical trials with an age/gender requirement matching age and gender parsed from the topics had their scores boosted so that the lowest scored matching trial within this top 10 results ended up with a higher score than that of the highest scoring non-matching trial (so, the order within the matching/non-matching subgroups was preserved).

3.4 Query expansion using DocT5Query trained on MS Marco (doct5query run)

We followed the methodology outlined by Pradeep et al. [3]. For this run we expanded each of the topics 40 times using a pretrained docT5query model[2]. The original topic and each of the expansions were posted as queries and scored using BM25. The individual results were then combined using a reciprocal rank fusion. Based on our hyperparameter tuning experiments on the TREC CT 2021 we fused the scores assigning a higher weight to the results of the original query in a 20:1 ratio.

[2]'castorini/doc2query-t5-base-msmarco'; we note it is a smaller version of the model, compared to the one used by Pradeep et al. [3]

		Metrics			
Method	**Run name**	**NDCG@5**	**NDCG@10**	**P@10**	**RR**
BM25 w. MonoBERT	monobert500	0.5090	**0.4912**	**0.3620**	0.5273
BM25 w. self-supervised MonoBERT	zs_bert_500	**0.5308**	0.4815	0.3280	**0.6117**
Contrastive Learning with Bi-Encoders	CSIROmedANIR	0.3394	0.3083	0.2020	0.4085
	ANIR_demo	0.3954	0.3407	0.2380	0.4709
Query expansion using DocT5Query	doct5query	0.3626	0.3374	0.2420	0.3912
BM25		0.4359	0.4022	0.2780	0.5150
TREC Median			0.3922	0.2580	0.4114

Table 1: A comparison of our submitted runs and the official TREC median. We add a post-TREC BM25 run ($b = 0.7$ and $k1 = 1.2$; the same parameters were used in the reranked runs) for comparison. Bold indicates the highest value in a column.

4 EVALUATION METRICS

For this track, three metrics are used for evaluation: Normalized Discounted Cumulative Gain at rank 10 (NDCG@10), precision at rank 10 (P@10) and reciprocal rank (RR). We also report NDCG at rank five for completeness.

5 EXPERIMENTS AND RESULTS

For our two monoBERT runs and DocT5Query run we use A2A API [8] to obtain the initial BM25-ranked lists.

A comparison of our submitted runs to the TREC median and a BM25 baseline is shown in Table 1. Our two MonoBERT-based rerankers lead to higher values for all three official metrics compared to the TREC median and BM25 baseline. The other three runs, however, were less effective and performed around the median level. We note that a plain untuned BM25 which is the basis for our reranking provides above-median results in all three metrics.

To gain insight on how our MonoBERT run compares on a single topic basis against median and best submissions, we visualise the NDCG@10 values per topic in Figure 2. For four (8%) topics (20, 21, 38 an 47), our run achieves perfect score (NDCG@10=1). On the flip side, there were eight topics (16%) which we did not retrieve any of the known relevant documents. For the remaining 38 topics (76%), our system's output for that topic mostly sits between median and best. In the example topics in Figure 1, we see the two extreme ones, with Topic 8 leading to our run not ranking any of the relevant documents in top-10. This topic had only 14 relevant documents, which our system retrieved 3 of those (21.4% recall). Topic 38, on the other hand, had 139 relevant documents, we retrieved 131 of those (94.9% recall) and achieved a score of 1 for RR and NDGC@10.

6 SUMMARY

We reported on our CSIROmed team's participation in the TREC 2022 Clinical Trials track. Our team submitted five runs with three different strategies. Our MonoBERT-based rerankers resulted in better ranking compared to other methods that we experimented with. The reranked runs also led to well-above TREC median results in all three official metrics of NDCG@10, Precision at rank 10 and Reciprocal Rank.

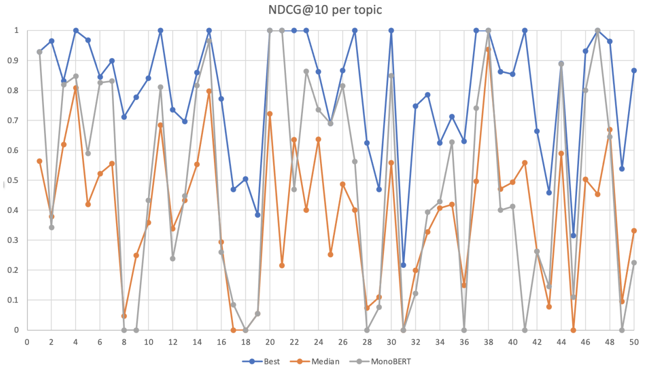

Figure 2: Per topic comparison of our best submitted run (MonoBERT) with the TREC median and best. Note that the best per topic of TREC is from any submitted run of all the participants and does not represent a single submission.

ACKNOWLEDGEMENTS

This work is supported by The Commonwealth Scientific and Industrial Research Organisation (CSIRO) Precision Health Future Science Platform (FSP). We would like to acknowledge the contributions of our colleagues Xiang Dai and Liam Watts.

REFERENCES

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

[2] Jill M Novitzke. 2008. The significance of clinical trials. *J Vasc Interv Neurol* 1, 1 (Jan. 2008), 31.

[3] Ronak Pradeep, Yilin Li, Yuetong Wang, and Jimmy Lin. 2022. Neural Query Synthesis and Domain-Specific Ranking Templates for Multi-Stage Clinical Trial Matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2325–2330. https://doi.org/10.1145/3477495.3531853

[4] 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.

[5] 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.

[6] 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.

[7] Maciej Rybinski, Vincent Nguyen, and Sarvnaz Karimi. 2021. CSIROmed Team Report of TREC 2021 Clinical Trials track: Experiments with BERT Reranking Methods. In *TREC*. https://trec.nist.gov/pubs/trec30/papers/CSIROmed-CT.pdf

[8] Maciej Rybinski, Liam Watts, and Sarvnaz Karimi. 2022. A2A-API: A Prototype for Biomedical Information Retrieval Research and Benchmarking. In *SIGIR*. 3318–3322. https://doi.org/10.1145/3477495.3531667

[9] 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.

[10] José A Sacristán, Alfonso Aguarón, Cristina Avendaño-Solá, Pilar Garrido, Juan Carrión, Alipio Gutiérrez, Robert Kroes, and Angeles Flores. 2016. Patient involvement in clinical research: why, when, and how. *Patient Prefer Adherence* 10 (April 2016), 631–640.

[11] Joseph M. Unger, Dawn L. Hershman, Kathy S. Albain, Carol M. Moinpour, Judith A. Petersen, Kenda Burg, and John J. Crowley. 2013. Patient Income Level and Cancer Clinical Trial Participation. *Journal of Clinical Oncology* 31, 5 (2013), 536–542. https://doi.org/10.1200/JCO.2012.45.4553 arXiv:https://doi.org/10.1200/JCO.2012.45.4553 PMID: 23295802.

[12] Joseph M Unger, Dawn L Hershman, Cathee Till, Lori M Minasian, Raymond U Osarogiagbon, Mark E Fleury, and Riha Vaidya. 2021. "When Offered to Participate": A Systematic Review and Meta-Analysis of Patient Agreement to Participate in Cancer Clinical Trials. *J Natl Cancer Inst* 113, 3 (March 2021), 244–257.