TREC 2021* Clinical Trials Retrieval, Duisburg-Essen University submission

Sameh Frihat and Norbert Fuhr

University of Duisburg-Essen, Duisburg, Germany
{firstname.lastname}@uni-due.de

Abstract. Clinical trials are human research studies that aim to evaluate a medical, surgical, or behavioral intervention that is critical to the advancement of medical science. The majority of clinical trials fail because recruitment goals are not met. This issue necessitates the incorporation of automated systems capable of matching patients to ongoing clinical trials. This paper summarizes our participation in the TREC 2021 clinical trials track, which provided all participants with a 5–10 sentence patient description and a clinical trials database from ClinicalTrials.gov for matching. Our submission consists of a variety of retrieval techniques, including BM25, entity recognition, BERT, and others. The results show that a simple BM25 ranking algorithm could outperform neural network-based models, mainly due to the absence of quality training data.

Keywords: Clinical Trials · Patients Recruitment · Information Retrieval

1 Introduction

TREC Biomedical Tracks in 2021 presents the clinical trials challenge for the first time[1]. This track aims to match clinical trial documents with patient case descriptions. On one hand, clinical trial documents contain different fields of information, mainly description, eligibility criteria, and others. Eligibility criteria, which are the most important factors in deciding whether or not to enrol a patient in a clinical trial, have four key components. It begins with demographic information, namely age and gender, and ends with inclusion and exclusion criteria in the form of free text. Inclusion-exclusion criteria contain various types of information such as lab values, genes, diseases, and others. On the other hand, the patient description simulates the first admission statement in an electronic health record (EHR) as 5–10 sentences, which contains the age, gender, and other information.

Most clinical trials, which are important for medical knowledge development, fail to recruit the minimum number of patients required to power the

* This project was supported in part by the National Science Foundation and in part by the National Security Agency.

[1] http://www.trec-cds.org/

study, causing the trial to be delayed or terminated. The idea of using EHRs to match patients with trials is introduced, burying the issue of recruiting. We aimed to investigate the performance of five simple information retrieval models for matching patients with trials in our submissions. Filtering and probabilistic techniques, as well as entity extraction, membership functions, and unsupervised BERT models, are used to build the models.

In Section 3, we present the IR architectures with a deeper look into the results. Results show that a simple probabilistic model based on a combination of BM25 and entity extraction can outperform more complex models.

2 Data

The 2021 clinical trials track provided 75 topics and 375,580 clinical trial documents for evaluating the models. Each topic contains 5–10 sentences describing a synthetically created patient, inspired by actual patients and modified. This is because extracting real patient data from EHRs was deemed too risky due to ethical and legal constraints.

The clinical trial document comes in XML format provided by the U.S. National Library of Medicine ClinicalTrials.gov. Each document contains the trial description and the eligibility criteria. The eligibility criteria, which are provided in the form of free text containing the inclusion and exclusion criteria, along with eligible genders and ages.

3 Methods

The clinical trials track is now operational for the first time. As a result, we chose to modify the baseline model (BM25) [5] rather than apply new techniques that necessitate training data in advance, which is currently unavailable. We attempted five different automatic runs, each combining filtering techniques (based on demographic data provided) with probabilistic retrieval techniques such as BM25 and BERT Embeddings [3].

3.1 Preprocessing

We automatically extracted the gender and age for each topic prior to running the retrieval model. Based on this demographic information, we filtered out clinical trials that do not accept the mentioned patient at the topic. Then, the resulting trial subcollection is fed into the retrieval models for ranking clinical trials based on their relevance.

3.2 Runs[2]

We performed five runs (retrieval models), described as follows:

[2] We used the same run names as provided in the TREC submission results.

- **First_run** This run is titled with *BM25 on eligibility criteria*, which could be described as the following:
  - Retrieve the top 1000 trials using BM25 on the topic and eligibility inclusion criteria from the filtered subcollection for calculating the positive scores.
  - Assign the BM25 score for each of the 1000 trials on the topic and eligibility exclusion criteria as a negative score.
  - Re-rank the documents using subtracting the "negative score" from the "positive score", which gives inclusion and exclusion criteria the same weight.

- **Second_run** This run is titled with *BM25 and membership function*, which could be described as the following:
  - Retrieve the top 1000 trials using BM25 on the topic and trial's description from the filtered subcollection.
  - Assign to each trial the BM25 score on the eligibility inclusion criteria as a positive score.
  - Assign to each trial the BM25 score on the eligibility exclusion criteria as a negative score.
  - Re-rank the documents using subtracting the "negative score" from the "positive score", which gives inclusion and exclusion criteria the same weight.

- **Third_run** This run is titled with *BM25 with Named-entity recognition on topics (NER)* [4], which could be described as the following:
  - Extract topic entities using SciBert [2].
  - Calculate the BM25 score between extracted *entities* and inclusion criteria.
  - Calculate the BM25 score between the topic and *trial description*.
  - Re-rank the documents using this formula: $2 * entities\ score + trial\ description\ score$.

- **Fourth_run** This run is titled *Bio Clinical BERT*, which is based on publicly available clinical BERT embeddings [1] and trained on medical articles and clinical notes. The model could be described as the following:
  - Use BM25 as an initial retrieval for clinical trials, using the topic text with the trial description and inclusion criteria from the filtered subcollection.
  - Calculate the clinical BERT embeddings for the retrieved trials as well as the topic.
  - Re-rank the trials based on the cosine similarity between the trial embeddings and topic embeddings.

- **Fifth_run** This run is titled with *BM25 on trial description with NER as membership function*, which could be described as the following:
  - Retrieve the top 1000 trials using BM25 on the topic and trial's description.

* Assign to each trial the BM25 score on the extracted entities from the eligibility inclusion criteria text as a *positive score*.
* Assign to each trial the BM25 score on the extracted entities from the eligibility exclusion criteria text as a *negative score*t.
* Re-rank the documents using this formula: $0.5 * BM25\ score + 0.2 * positive\ score - 0.3 * negative\ score$.

4 Results

Our main results can be summarized in table 1 in terms of NDCG@5 and NDCG@10. Third_run outperforms the other runs as well as the average of the TREC medians for automatic runs, according to the results. Furthermore, Figure 1 compares our best submission to the median and best NDCG scores for 101 automatic TREC runs for each topic, providing a more in-depth look at our best submission. This shows that it performed quite well in most topics, but it failed in a few topics like 34, 69, and 71. However, when compared to the average of the TREC medians for manual runs, all runs performed poorly. Moreover, when our best run is compared with the best and median of 12 manual runs per topic (Figure 2), in some topics like 17, 48, 56, and 67, it outperformed the median and is very comparable to the best-achieved score.

Table 1. Our runs' results in comparison to the TREC average of the median

Submission	NDCG@5	NDCG@10
First_run	0.1959	0.1704
Second_run	0.2103	0.1951
Third_run	**0.4529**	**0.4214**
Fourth_run	0.1991	0.2023
Fifth_run	0.3262	0.2973
AVG_Median (manual)	-	0.6212
AVG_Median (auto)	-	0.3040

5 Discussion

The best two of our five submissions use the NER technique, which was successful in retrieving most of the topics with relatively good scores. This could imply that there is a link between the use of NER and the results, which worth further investigation. Moreover, it would be very interesting to see how much weight tuning could improve the result, i.e., cross-validation with the data that will be available soon.

Fig. 1. NDCG@10 of the best submitted run compared to the official median NDCG@10 of 101 automatic runs and best achieved score per topic.

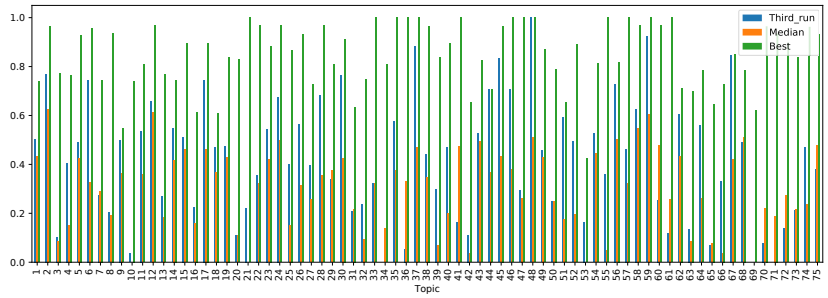
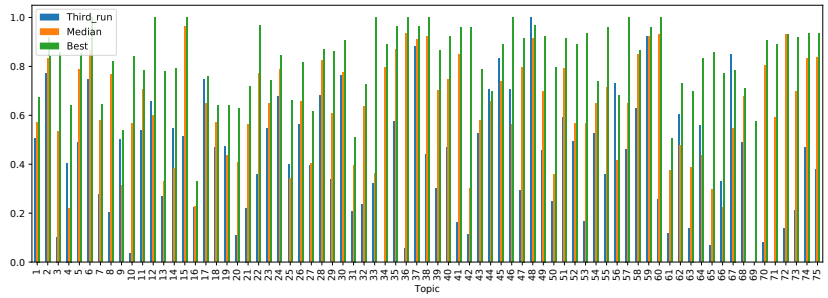


Fig. 2. NDCG@10 of the best submitted run compared to the official median NDCG@10 of 12 manual runs and best achieved score per topic.



The result also shows that the idea of using negative weights for the trial exclusion criteria as in First_run and Second_run is limited and should be implemented differently. Our best IR system failed in topics 10, 34, 36, 69, and 71 with very low recalls. However, the median of all TREC runs (manual and automatic) also failed on some of these topics. Therefore, we would like to investigate the causes by observing these topics.

6 Conclusion

This paper describes the procedures and outcomes of our participation in the TREC 2021 Clinical Trials track. The results show that one of our IR models works well. This run outperformed the official TREC medians (automatic runs) and contributed the best results in a variety of topics. The results, however, show that our negative score for exclusion criteria methods are not as effective

as we had hoped. Some topics should be thoroughly researched in future studies before implementing other IR methods.

7 Acknowledgemetns

This work was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalization of medicine at the point of care (WisPerMed), University of Duisburg-Essen, Germany.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Mohit, B.: Named entity recognition. In: Natural language processing of semitic languages, pp. 221–245. Springer (2014)
5. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc (2009)