Elsevier Data Science Health Sciences at TREC 2022 Clinical Trials: Exploring Transformer Embeddings for Clinical Trial Retrieval

Drahomira Herrmannova, Sharvari Jadhav, Harsh Sindhwa,
Hina Nazir, Elia Lima-Walton
Data Science Health Sciences, Elsevier, USA
{d.herrmannova, s.jadhav.1, h.sindhwa}@elsevier.com
{h.nazir, e.lima}@elsevier.com

Abstract

In this paper, we describe the submissions of Elsevier Data Science Health Sciences to the TREC 2022 Clinical Trials Track. Our submissions explored the applicability of transformer embeddings to the task and demonstrated a straightforward retriever using the MiniLM model can achieve competitive performance. Additionally, we share observations from a manual evaluation we performed to better understand the performance of our embedding-based retrievers.

1 Introduction

Clinical trials are the cornerstone of evidence-based medicine, ensuring the availability of safe and effective treatments by studying their effects on human subjects. Matching enough eligible patients to clinical trials is essential for achieving statistically significant results; however, the recruitment of patients represents a bottleneck in clinical research. This is due to many factors including the sheer number of recruiting trials and the often complicated eligibility criteria. The TREC 2022 Clinical Trials (CT) Track[1] provides a venue for developing systems for automated patient-to-trial matching. In this paper, we describe the submissions of Elsevier Data Science Health Sciences to the TREC 2022 CT Track.

The TREC 2022 CT Track provides participants with descriptions of 50 patients (also referred to as topics) and a historic snapshot of the ClinicalTrials.gov database. Participating teams are asked to retrieve a set number of clinical trials relevant to each topic, the relevance of retrieved trials is then evaluated by TREC. The topics are plain text descriptions of 5-10 sentences created to resemble EHR admission statements. The ClinicalTrials.gov snapshot is a set of 375,580 clinical trials in XML format published between November 1999 and April 2021. The clinical trial descriptions can be quite long and complex, often including inclusion/exclusion criteria which define trial eligibility. This can make matching patients to clinical trials a challenging task as eligible patients should meet inclusion criteria but no exclusion criteria, based on the plain text topic descriptions provided.

In addition to the above data, we used the dataset introduced in (Koopman and Zuccon, 2016) as an auxiliary source of evaluation data. This dataset uses topics from the TREC 2014-2015 Clinical Decision Support (CDS) Track, and provides relevance judgments over a historic snapshot of the ClinicalTrials.gov database.

We submitted four submissions for this challenge that involved experimenting with different retrievers and rerankers. We experimented with both lexical and embedding based retrieval. More specifically, we used BM25 as well as two different Transformer models for retrieving relevant trials: MiniLM[2] (Wang et al., 2020), a distilled Transformer model that uses self-attention distillation, and a DistilBERT model[3] trained using Generative Pseudo Labeling (GPL) for domain adaptation (Wang et al., 2022). We then experimented with two re-ranking methods for re-ordering the retrieved trials: a naïve reranker which averages scores from different retrievers, and a MiniLM-based reranker. To better understand the limitations of our approach, we performed a manual evaluation of 600 topic-trial recommendations, demonstrating the importance of accounting for the trial exclusion criteria in retrieval.

2 Methods

Our submissions broadly address the applicability of established retrieval methods to the task. Our

[1] http://www.trec-cds.org/2022.html

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://huggingface.co/GPL/bioasq-msmarco-distilbert-gpl

system uses a retriever and a reranker. In addition, we apply basic preprocessing and data filtering. As part of preprocessing, we split the trial criteria field into two separate fields for inclusion and exclusion criteria. During retrieval we used two basic inclusion/exclusion filters – age and gender. Age and gender play an important role in identifying whether a patient is eligible for a trial. Applying an age and a gender filter has also been shown to improve the performance of systems submitted to the previous TREC 2021 CT Track (Vu and Wu, 2021).

2.1 Retrievers

2.1.1 BM25

Our baseline retriever (submission *bm25_bi_filtered*) uses BM25. In our final submission, stopwords and punctuation were removed, all terms were lowercased, and scoring was computed using uni- and bi-gram tokens. Additionally, trials were filtered using patient age and gender information.

In addition to the aforementioned BM25 model, we experimented with query expansion using our in-house medical taxonomy. This taxonomy is composed of medical concepts, which are classified into different semantic types (e.g., drug brand names) and semantic groups (e.g., diseases). Query expansion was done by mapping terms in the topics to concepts in the taxonomy and appending concept synonyms to the topic text. Query expansion was evaluated using the dataset introduced in (Koopman and Zuccon, 2016). However, as it did not demonstrate improved performance over our BM25 model without query expansion, we have not used query expansion in the final submission.

2.1.2 Embedding-Based Retrieval

As our embedding-based retrievers, we used two different transformer models in a bi-encoder fashion. Specifically, we used MiniLM (Wang et al., 2020) (submission *senttr*), and a DistilBERT model trained using GPL (Wang et al., 2022) (submission *st_distilbert*). While MiniLM is trained on a broad range of different domains (a complete list of training datasets is provided on the MiniLM model card[2]), the DistilBERT model was trained using the MS MARCO dataset (Nguyen et al., 2016) and adapted to BioASQ (Tsatsaronis et al., 2015) (biomedical QA dataset) using GPL. We selected this particular GPL model due to the similarity between BioASQ and clinical trials domains. The topic and the trial inclusion criteria were encoded using either MiniLM or using DistilBERT and cosine similarity between the two vectors was used for retrieval.

2.2 Rerankers

2.2.1 Naïve Reranker

Our naïve reranker (submission *bm25_st_bienc*) computes scores by averaging of BM25 scores and cosine similarities between embedding vectors obtained using MiniLM.

2.2.2 MiniLM Reranker

Finally, we used a MiniLM reranker trained on MS MARCO to re-score the top 100 documents. The inclusion criteria field was used to represent documents in the reranker input. Due to time restrictions, we were unable to finalize this submission; however, we hope to perform an evaluation of this model in our future work.

3 Results

3.1 TREC 2022 Results

Table 1 shows results of our submissions along with the best performing run in the TREC 2022 CT Track (*frocchio_monot5_e* by team *h2oloo*) and the TREC median score. Our best submission *senttr* uses MiniLM-L6-v2[2] and placed fifth by NDCG@10 out of all teams' best submissions. Two of our submissions (*senttr* and *bm25_st_bienc*) performed better than the TREC median score. The MiniLM model performs significantly better than all other models including DistilBERT. Interestingly, the GPL DistilBERT achieved the worst scores of our four submissions and performed significantly worse than MiniLM. This may be due to the domain shift between the BioASQ (biomedical QA) and clinical trials domains, despite similarities between the domains. Re-ranking by combining scores from BM25 and MiniLM (*bm25_st_bienc*) also did not improve performance over using MiniLM only.

3.2 Manual Evaluation

To understand whether our setup using transformer encoders like MiniLM and DistilBERT is capable of retrieving relevant trials, we performed a manual evaluation using the (Koopman and Zuccon, 2016) data. For this evaluation, we used the topics and the snapshot of ClinicalTrials.gov used in (Koopman and Zuccon, 2016) and generated top 10

Table 1: Evaluation results of our submissions along with the best performing run in the TREC 2022 CT Track (frocchio_monot5_e by team h2oloo) and the TREC median score, ranked by NDCG@10.

Submission	NDCG@5	NDCG@10	Precision@5	Precision@10	Reciprocal Rank
frocchio_monot5_e	-	0.6125	-	0.5080	-
senttr	0.4973	0.4758	0.3680	0.3540	0.5341
bm25_st_bienc	0.4774	0.4391	0.3400	0.3140	0.5331
TREC median	-	0.3922	-	0.2580	0.4114
bm25_bi_filtered	0.3566	0.3275	0.2480	0.2240	0.4697
st_distilbert	0.3441	0.3194	0.2520	0.2400	0.3688

predictions for each topic in the data by calculating consine similarity between the MiniLM[2] encoded topic description and the trial criteria, generating 600 topic-trial pairs. These predictions were provided to six subject matter experts (SMEs) – medical doctors and students, who were asked to assign each patient trial pair a score of 0 (would not refer patient), 1 (would consider referring patient), or 2 (highly likely to refer patient) to match the scoring used by the (Koopman and Zuccon, 2016) dataset. The SMEs were provided with the following information: topic description, trial title, trial criteria, trial condition, and trial keywords. Table 2 provides summary results of the evaluation, while Figure 1 provides results per topic.

Table 2: Summary results of our manual evaluation.

NDCG@10	Precision@10	MRR
0.5468	0.2800	0.4906

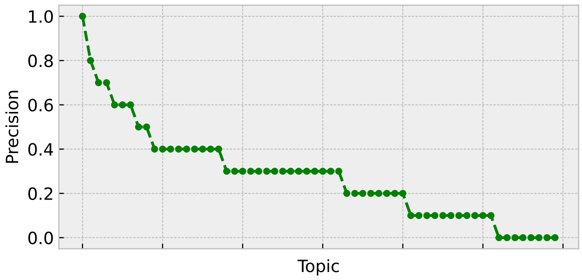

Figure 1: Precision per topic obtained from our manual evaluation.

We observed two common reasons for the disparity in the number of relevant retrieved trials between different topics: 1) the number of trials available for a specific condition, and 2) trial exclusion criteria. The topics with a high number of retrieved trials that are relevant tended to mention more common conditions such as diabetes, while topics with little to no relevant trials tended to mention less common conditions such as amenorrhea. Searching ClinicalTrials.gov for these two conditions reveals thousands of trials related to diabetes, but less than 100 related to amenorrhea. The SMEs also tended to mark trials as not relevant due to the trial exclusion criteria, such as patient history. Based on these findings, in our future work we intend to incorporate more granular matching based on the exclusion and inclusion criteria.

4 Conclusion

In this paper, we describe the submissions of Elsevier Data Science Health Sciences to the TREC 2022 CT Track. Our submissions explored the applicability of transformer embeddings and demonstrated a straightforward retriever using the MiniLM model can achieve competitive performance. As our future work we plan to incorporate trial exclusion criteria in the retrieval.

References

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*, pages 669–672.

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 *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 colocated with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, volume 1773.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Hoang Vu and Danny T.Y. Wu. 2021. CincyMedIR at TREC 2021 Clinical Trial track. In *NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021)*.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.