Matching of Patient Questionnaires to Clinical Trials with Large Language Models

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

ABSTRACT

To assist with finding eligible participants for clinical trials, the TREC 2023 Clinical Trials track sets a task where patient data, in the form of patient questionnaires, can be used to match eligible patients to a relevant clinical trial. We explore several query expansion and reranking methods using large language models. Our best method uses query expansion with GPT 3.5-turbo and reranking with a fine-tuned version of the same model.

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 Sarvnaz Karimi. 2021. Matching of Patient Questionnaires to Clinical Trials with Large Language Models. In *TREC'23: TREC, November, 2023.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Development of new treatments heavily relies on clinical trials [2]. Clinical trials are required to enrol enough patients to ensure the study can be used to draw reliable conclusions [5] and that the treatments work for a wide range of demographics. However, in most cases, not many patients find opportunities to join clinical trials and from those who do, some may not agree to participate or even not be eligible to do so [3, 4]. It is therefore necessary to create tools that facilitate the matching of patients with potential trials to improve their success.

The TREC Clinical Trials (CT) 2023, is the third edition of the yearly track. The task is to link a synthetic patient's questionnaires, in a semi-structured format, 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. Our experiments this year focus on query expansion using large language models (LLMs) and LLM-based neural reranking in both zero-shot, and supervised settings.

TREC 2023, November, 2023, Online
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

```
<topic number="8" template="anxiety">
<field name="definitive diagnosis">no</field>
<field name="age">12yo</field>
<field name="proficient languages">English, Spanish</field>
<field name="SSASI">12</field>
<field name="HAM-A">25</field>
<field name="PHQ-9"/>
<field name="HAM-D">14</field>
<field name="GAD-7"/>
<field name="Beck Depression Inventory"/>
<field name="suicidal ideation">no</field>
<field name="dementia">no</field>
</topic>
```

Figure 1: An example topic from the TREC CT 2023 track.

In our experiments, we explore the effects of query expansion, where patient questionnaires are extended with *gpt-3.5-turbo* summaries of the diagnosis. We probe several reranking approaches: dense-retrieval-based method using *text-embedding-ada-002* dense vector representations, zero-shot pointwise reranking using *gpt-3.5-turbo*, and pointwise reranking using a fine-tuned *gpt-3.5-turbo* model. We compare our results against a BM25 baseline.

2 DATASET

The TREC CT 2023 dataset consists of 40 topics with corresponding relevance judgments. The corpus for the task is a 2023 snapshot of ClinicalTrials.gov database[1], with over 450K registered clinical trial records. Each topic simulates a patient's questionnaire with a condition-specific template. 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 our run which uses supervised learning to train reranking models, we use the TREC Clinical Trials 2021 track collection as training data (noting different structure of the topics).

3 METHODS

As per TREC guidelines all methods return 1000 documents per topic. All the reranking methods build on the query-expanded BM25 used as the first stage retrieval step (see

[1] http://clinicaltrials.gov/

detailed descriptions below). In all reranking experiments, we rerank the top 100 documents.

Indexing and BM25 baseline. The following fields of clinical trials are indexed: brief summary, brief title, identifier, detailed description, drug name, drug keywords, inclusion/exclusion criteria, gender, general keywords, intervention type, maximum age, minimum age, official title, and primary outcome. Intervention type, gender, and primary outcome refer to controlled vocabularies; age-related fields are numeric. All other fields except clinical trial ID are textual.

The BM25 retrieval is implemented using Solr [1] search engine's implementation with default hyper-parameter values ($b = 0.7$ and $k1 = 1.2$). We index clinical trials with all their fields, with inclusion/exclusion criteria split into two separate fields. Moreover, all textual fields are aggregated into a *text* field. For BM25, the documents are scored using this aggregated field. We use the complete topic text as a query. BM25 baseline is evaluated in the bm25_bsln run.

BM25 with query expansion. We employ a query expansion strategy which uses *gpt-3.5-turbo*. We prompt the model to generate text regarding the patient diagnosis (i.e., *What is the diagnosis for the following patient?*). The generated text is then appended to the original topic text for query formulation. The query expansion strategy is evaluated in the *qe* run.

Dense vector reranking. In dense vector reranking inclusion criteria and expanded queries are separately embedded using *text-embedding-ada-002* model and the reranking score is calculated as a cosine similarity between these vector representations. For the final ranking the cosine score of top 100 documents is interpolated with the normalised expanded-BM25 score (so, the first stage retrieval score). The interpolation is done in a 9:1 ratio. For the remaining 900 documents for each topic we keep the first stage retrieval score divided by 10 (i.e., we assume the cosine similarity of 0). The dense reranking is evaluated in the *qe_err* run.

Zero-shot prompt-based reranking. We obtain prompt-based ranking scores directly by prompting *gpt-3.5-turbo* model (with a prompt 'Consider the following patient: (...) Is the clinical trial below adequate for this patient (...) Reply only with a confidence score between 0 and 1.'). The topics are represented in the prompt with the original topic text, while the trials are represented with a concatenation of brief summaries and inclusion/exclusion criteria. Top 100 documents are scored with the output of the model incremented by 1, remaining documents are scored with the normalised score from the initial retrieval step.

Trained prompt-based reranking. The reranker and prompt are identical to the zero-shot setting. The key difference is that the model used on inference is previously fine-tuned on 10000 labelled topic-document pairs randomly sampled from TREC CT 2021 relevance judgements. Of note, 2021 topics were in a free-text format simulating an extract from a patient's electronic health record, so the topics presented to the model at inference are quite different to those used in training.

Another point of difference with the zero-shot prompt-based reranking is that in the fine-tuning process, the model learns to output scores matching to the task's graded relevance scale (0, 0.5, and 1; these match to human judgements of 0, 1, and 2, respectively). To counteract multi-document ties in the reranked results we add the inverse rank of the document in the first stage retrieval to the reranker score for the top 100 documents. The remaining 900 documents are scored with the inverse rank alone.

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).

5 RESULTS

The results of our experiments are reported in Table 1. We obtain the best results for the trained prompt-based reranker, which is also our most resource-intensive approach (101 LLM requests per topic, requires fine-tuning and maintenance of a dedicated model). In contrast, our query expansion approach is the second-lightest of all evaluated methods (behind only stock-standard BM25), yet still offers a substantial improvement in the primary evaluation metric (NDCG@10) over the BM25 baseline. We believe that our results highlight the promise of query expansion using LLMs.

6 SUMMARY

We reported on our CSIROmed team's participation in the TREC 2023 Clinical Trials track. Our team submitted five runs, with the best zero-shot results attained with our query expansion strategy without reranking. The best overall results were obtained by the run combining query expansion and rank fusion with a trained *gpt-3.5-turbo* reranker. The team ranked highest amongst all the participants in this round based on the main metric of NDCG.

ACKNOWLEDGEMENTS

This work is supported by The Commonwealth Scientific and Industrial Research Organisation (CSIRO) Precision Health Future Science Platform (FSP).

REFERENCES

[1] Apache. 2016. http://lucene.apache.org/solr/. [Version: 6.0.1].
[2] Jill M Novitzke. 2008. The significance of clinical trials. *J Vasc Interv Neurol* 1, 1 (Jan. 2008), 31.
[3] 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.
[4] 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.
[5] 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

		Metrics		
Method	**Run name**	**NDCG@10**	**P@10**	**RR**
BM25 baseline	bm25_bsln	0.619	0.330	0.563
Query expansion (QE)	qe	0.699	0.354	0.535
QE + dense ranking	qe_err	0.652	0.368	0.483
QE + prompt ranking	qe_prr_zs	0.593	0.338	0.452
QE + tuned prompt ranking	qe_prr_ft	0.738	0.527	0.667
TREC median	-	0.648	0.397	0.538

Table 1: A comparison of our submitted runs and the official TREC median.

Agreement to Participate in Cancer Clinical Trials. *J Natl Cancer Inst* 113, 3 (March 2021), 244–257.