The Application of Traditional IE as a Non-traditional Method in an IR Task: TDMINER at 2021 TREC Clinical Trials

Chengyi Zheng

Kaiser Permanente Southern California, Pasadena, CA[1]

tdminer@gmail.com

Abstract

The 2021 TREC Clinical Trials (CT) task focused on finding appropriate trials based on the health profiles of individual patients. This notebook details our participation in the 2021 TREC CT. In this paper, we presented the findings of our first participation in the TREC task. The TREC Clinical Trials (CT) goal for 2021 was to discover appropriate trials based on the health characteristics of individual patients. This notebook details our participation in the TREC CT in 2021 (team TDMINER). We presented the findings of our initial participation in the TREC task in this publication. Traditional information retrieval approaches, such as Elasticsearch with BM25 or DFR with query expansion, and machine learning-based rerankers, such as BERT, were used in previous efforts. Unlike these methods, we concentrated on developing an IE-based baseline that could be utilized as a starting point for future research. As part of our two-stage IR process, we implemented a basic weight-scaled reranking method. We submitted our results in the manual run category since we manually reranked the IE-identified concepts for each topic.

There were 26 teams in total, with 101 automatic runs and 12 manual runs submitted. In terms of NDCG@10, PREC@10, and mean reciprocal rank (MRR), we achieved final ranking scores of 0.715, 0.576, and 0.834, respectively. In manual runs, the averaged median scores for these assessment criteria were 0.621, 0.457, and 0.721; in automatic runs, 0.304, 0.161, and 0.294. Our system ranked first on the NDCG@10 and MRR, second on the PREC@10 among all the submissions.

Keywords: Clinical Trial; Information Extraction; Information Retrieval; TREC; Search; Eligibility criteria

[1] Kaiser Permanente was the author’s employer at the time of performing this task. This work was neither sponsored nor endorsed by Kaiser Permanente. The author is solely responsible for the work and content of this paper.

Introduction

Clinical trials are the backbone of modern medicine's search for effective interventions and treatments. Clinical trials, on the other hand, are expensive and time-consuming. Many trials failed not because the medicines were ineffective, but because they were unable to enroll enough patients. The difficulty in locating matching trials for patients or finding matching patients for trial sponsors is often the cause of recruitment failure, not a lack of individuals who fit the selection criteria. The TREC Clinical Trials Task for 2021 is modeled after clinicians or patients looking for appropriate trials based on individual patients' health profiles.

For the first time, our team took part in the TREC task. We focus on establishing a workable basic system within the task timetable. We highlighted some of the lessons we gained while summarizing our approach and results.

Method

Matching clinical trial participants was the subject of many publications.[1-7] The past four years, TREC biomedical tracks have focused on finding relevant literature and clinical trials in the oncology domain. For the previous four years, TREC biomedical tracks have concentrated on locating relevant literature and clinical trials in the oncology sector, with Precision Medicine as a theme.[7-10] In the 2020 Precision Medicine track, traditional information retrieval (IR) methods such as Elasticsearch and BM25 for baseline search were dominant, followed by machine learning-based rerankers, the majority of which employed DNN.[7]

We decided to try out a new way that wasn't based on the classic IR method or the now-popular BERT-based DNN model. We wanted to see if an information extraction (IE)-based method[11-14] could handle this kind of task. We believe that a good IE system could lay the groundwork for IR tasks.

In the pre-preprocessing step, we extracted the inclusion and exclusion criteria from the clinical trials XML files. Multiple inclusion and exclusion criteria may be used in each clinical trial. We determined the number of inclusion and exclusion criteria for each experiment. The numbers of inclusion and exclusion criteria were used to weight and rerank the retrieved trials.

We indexed the topic text and the clinical trial documents using a set of ontologies from the UMLS. We used search queries to extract concepts and their associated features such as temporality, person, and numeric value from both the topic text and the trial document. We narrowed our search for clinical trial materials to a few fields, such as inclusion and exclusion criteria, title, condition, and keywords. We checked the returning search results for the topic text, looking for any essential text that didn't have any matching concepts, such as disease abbreviations that weren't in the ontologies. We constructed a customized ontology for these unmatched texts to improve search sensitivity.

For each topic, we searched all of the trials and assigned a score to those that had matching concepts. Because there were no labeled data to train a reranking model, we created a scoring system to rank returned trials. We initially calculated a scaling score for each matched concept in each trial based on the number of concepts in the topic and the section of the trial document:

$$scale = log_2(\text{number of concepts in the trial section } + \ 1)$$

Then we assigned section-based factor scores based on the section of the matching concepts:

factor = 30, 20, 20, and 15 for the section in title, condition, keyword, and inclusion criteria, respectively.

The concept-based score was calculated based on a reversed sigmoid function:

$$r\_sigmoid\_score = 1/(\frac{1}{1+\ \mathrm{e}^{-\mathrm{x}}}),$$

where

$$x = \frac{\text{concept ranking in the topic}}{\text{number of concepts in the topic}}$$

We implemented a weighting scheme by ranking the IE identified concepts from 1 to n. The concept ranking in the topic was manually assigned based on the author’s judgment. For example, the diseases were ranked higher than the treatments. Within diseases, they were ranked by their severity, temporality, and likelihood of being part of the main diagnoses. Clinicians and patients could easily apply this simple weighting method to any future topic.

The final score for one matching concept is:

$$concept\_score = factor/\text{scale} * r\_sigmoid\_score$$

The overall scores for matched concepts between one topic and one clinical trial were the sum of individual matched scores:

$$Matched_{score} = \sum_{k=0}^{n} concept_{score_k} * polarity;$$

$$polarity = 1\ for\ positive\ matched\ pair;\ -1\ for\ negative\ matched\ pair$$

Because the call for participants of this TREC track stated the evaluation measure is the normalized discounted cumulative gain (NDCG), we assumed that all the 1000-allowable submitted trials would affect the NDCG score. Therefore, it is advantageous to submit as many trials as possible for each topic. For topics with less than 1000 matched trials, we used synonym expansion approaches such as relationship-based synonym discovery and word vector-based similarity algorithms. We also performed concept expansion to increase the number of matched trials. For example, we added the upper-level concepts "thyroid neoplasms" and "thyroid carcinoma" to topic 11's initial disease concept "metastatic papillary thyroid cancer."

Most of the topics had over 1000 hits in the initial run. We applied methods to exclude or downrank trials. For example, we made two assumptions: if the qualifying disease condition was not indicated in the topic text, we presumed the patient did not have it; and the trial title had the must-met inclusion criteria for a trial. As a result, trials whose titles contained disease concepts that did not occur in the topic were excluded.

Result

The results we received from the organizer did not have the scores of other participating teams. The results did include the best, median, and worst scores for each topic separated by the automatic and manual runs. Overall, there were submissions from 101 automatic runs and 12 manual runs. Our PREC@10 and MRR scores were 0.576 and 0.834.

Because the NDCG score was the only mentioned measurement on the track website. We focused our analyses on the NDCG score. From the summarized results, we calculated the overall averaged scores based on the best, median, and worst scores for each topic for the auto and manual runs (Table 1). We further compared the scores from our best run submission to the best and median scores of the automatic and manual runs.

Our average NDCG@10 score is 0.715, which is 0.109 and 0.134 lower than the averaged best scores of the manual (0.823) and automatic runs (0.849). Because of the larger number of submitted runs in the automatic category, the averaged best score for automatic runs could deviate more from the best score of individual runs. We expected the difference between the averaged best scores of the automatic and manual runs would be smaller than our calculated number (0.026). Compared to the averaged median scores, our score is 0.094 and 0.411 better than the scores of manual (0.621) and automatic (0.304) runs. Because our team submitted four manual runs out of the total 12 manual runs. Our scores could significantly associate with the best or median scores of the manual runs.

Our scores were among the best in 36 topics and poorer than the median in seven topics when compared to the automatic or manual submissions. In the disease areas of autoimmune, gastrointestinal, neurology, cancer, and psychiatry, our system performed better; in cardiology, endocrinology, and healthy patients, it did worse.

		NDCG@10	Delta values of our score minus the best and median scores of the manual (m) and automatic (a) runs			
Topic	**Disease area or specialty**	**tdminer4**	**Best (m)**	**Best (a)**	**Median (m)**	**Median (a)**
29	Autoimmune	**0.861**	0.000	0.123	0.253	0.426
30	Autoimmune	0.776	-0.130	-0.188	0.000	0.150
37	Autoimmune	**0.963**	0.000	0.189	0.053	0.878
38	Autoimmune	**0.921**	-0.079	0.155	0.000	0.768
65	Autoimmune	0.794	-0.064	-0.135	0.496	0.367
66	Autoimmune	**0.770**	0.000	-0.187	0.546	0.442
72	Autoimmune	**0.932**	0.000	0.187	0.000	0.641
2	Cardiology	0.834	-0.086	-0.103	0.000	0.642
5	Cardiology	**0.789**	-0.068	0.241	0.000	0.424
16	Cardiology	0.229	-0.099	-0.512	0.000	0.229
19	Cardiology	0.436	-0.206	-0.373	0.000	0.077
47	Cardiology	0.796	-0.120	-0.173	0.000	0.181
62	Cardiology	0.475	-0.255	-0.292	0.000	0.291
20	Endocrinology	0.470	-0.157	-0.275	0.064	0.051
40	Endocrinology	0.747	-0.177	-0.147	0.000	0.286
53	Endocrinology	**0.931**	-0.006	0.317	0.364	0.772
56	Endocrinology	0.351	-0.332	-0.545	-0.067	-0.112
74	Endocrinology	**0.833**	-0.103	0.223	0.000	0.468
6	Gastroenterology	**1.000**	0.000	0.161	0.131	0.570
7	Gastroenterology	**0.646**	0.000	-0.183	0.069	0.536
21	Gastroenterology	**0.718**	0.000	-0.283	0.155	0.718
22	Gastroenterology	**0.967**	0.000	0.000	0.195	0.646

26	Gastroenterology	0.656	-0.160	-0.228	0.000	0.235
27	Gastroenterology	0.391	-0.226	-0.576	-0.012	-0.107
51	Gastroenterology	**0.915**	0.000	0.048	0.123	0.764
57	Gastroenterology	0.649	-0.352	-0.282	0.000	0.336
58	Gastroenterology	**0.865**	0.000	0.138	0.015	0.608
63	Gastroenterology	0.388	-0.311	-0.580	0.000	0.035
64	Gastroenterology	**0.832**	0.000	0.023	0.395	0.458
12	Genetic	0.599	-0.401	-0.314	0.000	0.173
45	Genetic	**0.889**	0.000	0.254	0.149	0.670
49	Genetic	0.696	-0.229	-0.052	0.000	0.601
55	Genetic	**0.961**	0.000	-0.039	0.247	0.639
33	Healthy	0.362	-0.638	-0.446	0.000	0.223
36	Healthy	**1.000**	0.000	0.000	0.065	0.624
69	Healthy	0.000	-0.575	-1.000	0.000	-0.330
71	Healthy	0.590	-0.301	-0.411	0.000	0.119
10	Hematology	**0.841**	0.000	-0.125	0.272	0.493
43	Infectious	**0.786**	0.000	-0.052	0.207	0.716
46	Infectious	0.497	-0.503	-0.397	-0.065	0.297
48	Infectious	0.931	-0.036	-0.069	0.015	0.456
52	Infectious	0.569	-0.323	-0.084	0.000	0.532
39	Neonatology	0.704	-0.163	-0.122	0.000	0.211
73	Neonatology	0.697	-0.221	-0.010	0.000	0.328
3	Neurology	**0.861**	0.000	-0.105	0.325	0.429
9	Neurology	0.311	-0.228	-0.689	0.000	-0.067
35	Neurology	**0.963**	0.000	-0.037	0.093	0.702
41	Neurology	0.882	-0.076	-0.119	0.032	0.371
75	Neurology	**0.932**	-0.002	0.062	0.097	0.501
42	OBG	**0.957**	-0.003	0.168	0.655	0.709
67	OBG	0.545	-0.240	-0.107	0.000	0.370
1	Oncology	0.619	-0.056	-0.272	0.046	0.424
4	Oncology	**0.642**	0.000	0.219	0.422	0.642
8	Oncology	0.775	-0.045	-0.037	0.009	0.332
11	Oncology	0.706	-0.079	-0.294	0.000	0.659
15	Oncology	**0.967**	-0.033	0.148	0.004	0.465
17	Oncology	0.674	-0.084	-0.326	0.025	0.350
25	Oncology	**0.661**	0.000	-0.307	0.318	0.112

31	Oncology	**0.512**	0.000	-0.489	0.117	-0.095
61	Oncology	**0.507**	0.000	-0.460	0.132	0.029
44	Orthopedics	0.656	-0.040	-0.344	0.000	0.398
50	Pediatrics	0.359	-0.438	-0.354	0.000	-0.072
34	Psychiatry	**0.890**	0.000	0.191	0.096	0.805
70	Psychiatry	**0.906**	0.000	0.121	0.103	0.646
54	Rare	**0.739**	0.000	0.091	0.091	0.662
14	Respiratory	0.383	-0.407	-0.345	0.000	0.346
23	Respiratory	0.727	-0.017	-0.122	0.080	0.305
59	Respiratory	**0.957**	0.000	0.172	0.033	0.448
13	Urology	**0.779**	0.000	0.159	0.448	0.779
18	Urology	0.569	-0.072	-0.396	0.000	0.349
24	Urology	0.829	-0.018	-0.108	0.042	0.640
28	Urology	0.863	-0.008	-0.036	0.039	0.589
32	Urology	**0.725**	0.000	-0.114	0.087	0.507
60	Urology	**1.000**	0.000	0.037	0.069	0.763
68	Urology	0.679	-0.030	-0.252	0.000	0.199
all	Average of all topics	0.715	-0.109	-0.134	0.094	0.411

Table 1. Comparison of NDCG@10 scores for our best runs with the best and median scores for auto and manual runs.

Note: The disease area or specialty labeling for each topic was created by the author for analyses purpose. It was based on the interpretation of the possible main diagnoses. Scores highlighted in orange color indicate that our score was better or equal to either the best scores of auto or manual runs. Scores highlighted in red color indicate our score was worse than the median scores of either auto or manual runs.

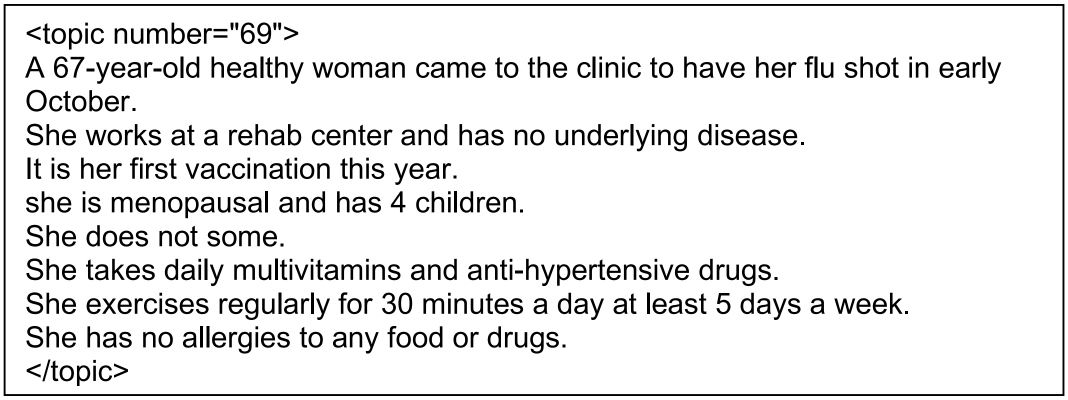
Discussion

We created an IE-focused IR system that is flexible and requires minimal human intervention. Due to the time constraints, we spent the majority of our work developing the baseline IE system. We implemented a simple reranking algorithm to select the candidate trials. Other down streaming modules, such as the machine learning-based reranker,[15] might be built on top of this IE baseline system. There were notable methods differences compared to other top-performing teams. For example, the University of Waterloo team used the deep learning-based models for both their first-stage retrieval and second-stage reranker. They trained a T5-3B model on the MS MARCO data and created 40 query runs for each topic. Their rerank model was based on the MS MARCO V2 model. They observed the biggest performance gain by retraining the base model on the clinical trial collection curated by Koopman and Zuccon.[16] The Alibaba team used an embedding-based (ClinicalBert) retriever and active learning retrained reranker which was re-trained on 1.7k manual annotated instances.

Besides the resources mentioned above,[16] many teams also used the description field of the clinical trial document. For example, the Waterloo team gained 0.033 on the NDCG@10 by adding the description field. Due to time constrain, we did not utilize these resources.

We were unaware that the ultimate measurement was NDCG@10, therefore we focused our efforts on improving the selection and ranking of the top 1000 results. Some of the techniques, particularly those that matched both the inclusion and exclusion criteria, may have been excessively harsh in removing some prospective hits.

We were only able to undertake a limited analysis of our results because we received the results from the organizer a week before the notebook submission date. We observed great performance variation among topics. Nevertheless, topic 69 was the only topic with zero median NDCG@10 and PREC@10 scores in both the automatic and manual runs, therefore we looked at its data.



```
<topic number="69">
A 67-year-old healthy woman came to the clinic to have her flu shot in early
October.
She works at a rehab center and has no underlying disease.
It is her first vaccination this year.
she is menopausal and has 4 children.
She does not some.
She takes daily multivitamins and anti-hypertensive drugs.
She exercises regularly for 30 minutes a day at least 5 days a week.
She has no allergies to any food or drugs.
</topic>
```

Figure 1: Topic 69 as provided by the track organizer

The subject of topic 69 was a healthy 67-year-old female with no underlying condition, as illustrated in Figure 1. Anti-hypertensive medicines may imply that she had hypertension. She was menopausal and exercised five times a week for at least 150 minutes. She took multivitamins and was recently vaccinated.

The eligibility requirements for our top three chosen trials for this topic are listed below (Figure 2). These three trials were deemed "not relevant" in the judged results. However, these trials, according to our interpretation, fit the characteristics of topic 69. We expected other participants could have similar questions because the majority of the runs had zero NDCG@10 and PREC@10 scores. If the organizers could provide more information regarding the judging criteria, that would be great. Topic-level judgment summary information, for example, could aid us in better understanding the reasoning behind their decisions. We understand the difficulties of judging nearly 36,000 documents. In the future, we hope the organizers could consider performing limited double-adjudication for each topic to reduce the potential error and bias. Finally, we appreciated the efforts of the organizers and adjudicators.

NCT03831373

Ages Eligible for Study:	60 Years and older (Adult, Older Adult)
Sexes Eligible for Study:	All
Accepts Healthy Volunteers:	Yes

Criteria

Inclusion Criteria:

- No medical contraindications to be involved in the exercise programs.
- To be able to follow the exercise classes
- Signed informed consent

Exclusion Criteria: -

NCT00894205

Ages Eligible for Study:	60 Years and older (Adult, Older Adult)
Sexes Eligible for Study:	All
Accepts Healthy Volunteers:	Yes

Criteria

Inclusion Criteria:

- No medical contraindications to be involved in the exercise programs.
- To be able to follow the exercise classes
- Signed informed consent

Exclusion Criteria: -

NCT02250950

Ages Eligible for Study:	18 Years and older (Adult, Older Adult)
Sexes Eligible for Study:	All
Accepts Healthy Volunteers:	Yes

Criteria

Inclusion Criteria:

- participants were were willing to attend an exercise class once a week for 12 weeks, were willing to complete questionnaires at baseline and 12 weeks, allowed the intervention staff to monitor their attendance at the YMCA for 6 months post intervention, and allowed the exercise instructor to create an audio recording of all of the intervention sessions.

Exclusion Criteria:

- not have any illnesses that would prevent them from exercising once a week, not be pregnant or planning to get pregnant within the next 3 months

Figure 2: Top 3 selected trials in run tdminer4

Conclusion

In this paper, we presented the findings of our first participation in the TREC task. We achieved the highest NDCG@10 and MRR scores, and 2nd ranked PREC@10. Unlike approaches used in past Precision Medicine challenges, we did not use traditional IR methods. Our methodology focused on creating an IE-based baseline that could be used as a starting point for future research. We spent less time designing the reranking algorithm as part of our two-stage IR pipeline. We also did not test many of the ideas we had during the competition due to time restrictions. Given the availability of relevant judgment data, we could investigate these ideas, as well as the effectiveness of our reranking method and alternative reranking approaches. Our findings showed that a competent IE-based method could achieve performance comparable to the best deep learning-based approaches. Our approach is resource-friendly and can be executed on a laptop, unlike those resource-hungry and computing-intensive deep learning-based systems. Given their drastically diverse methodologies, their results could be very complementary. In the future, we are interested in combining these deep learning-based approaches with our IE-based method.

References

1. Tu SW, Peleg M, Carini S, et al. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform.* 2011;44(2):239-250.
2. Kang T, Zhang S, Tang Y, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc.* 2017;24(6):1062-1071.
3. Zhang H, He Z, He X, et al. Computable Eligibility Criteria through Ontology-driven Data Access: A Case Study of Hepatitis C Virus Trials. *AMIA Annu Symp Proc.* 2018;2018:1601-1610.
4. Xu C, Forkel W, Borgwardt S, Baader F, Zhou B. Automatic Translation of Clinical Trial Eligibility Criteria into Formal Queries. Paper presented at: JOWO2019.
5. Inan OT, Tenaerts P, Prindiville SA, et al. Digitizing clinical trials. *npj Digital Medicine.* 2020;3(1):101.
6. Rosa C, Marsch LA, Winstanley EL, Brunner M, Campbell ANC. Using digital technologies in clinical trials: Current and future applications. *Contemp Clin Trials.* 2021;100:106219.
7. Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020. 2020.
8. Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019. 2019.
9. Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018. 2018.
10. Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017. 2017.
11. Zheng C, Huang BZ, Agazaryan AA, Creekmur B, Osuj T, Gould MK. Natural Language Processing to Identify Pulmonary Nodules and Extract Nodule Characteristics From Radiology Reports. *Chest.* 2021.
12. Zheng C, Sun BC, Wu YL, et al. Automated abstraction of myocardial perfusion imaging reports using natural language processing. *J Nucl Cardiol.* 2020.
13. Zheng C, Yu W, Xie F, et al. The use of natural language processing to identify Tdap-related local reactions at five health care systems in the Vaccine Safety Datalink. *Int J Med Inform.* 2019;127:27-34.

14. Zheng C, Luo Y, Mercado C, et al. Using natural language processing for identification of herpes zoster ophthalmicus cases to support population-based study. *Clin Exp Ophthalmol.* 2018.

15. Zheng C, Rashid N, Wu YL, et al. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. *Arthritis Care Res (Hoboken).* 2014;66(11):1740-1748.

16. Koopman B, Zuccon GJPottIAScoR, Retrieval DiI. A Test Collection for Matching Patients to Clinical Trials. 2016.