IBM @ TREC Clinical Trials Track 2021

Laura Biester*[1], Venkata Joopudi[2], and Bharath Dandala[2]

[1]Computer Science & Engineering, University of Michigan
[2]IBM Research
lbiester@umich.edu, {vnjoopud,bdand}@us.ibm.com

Abstract

Although clinical trials are crucial to the advancement of medical science, many clinical trials fail because they do not meet recruitment targets. This problem engenders a need for automated systems that can match patients to ongoing trials. The potential benefits of such systems are twofold: first, they would allow for the systematic study of new treatments through completed clinical trials and second, they could improve or even save the lives of patients for whom existing treatments are ineffective. We participate in the TREC Clinical Trials (CT) Track, for which the aim is to match synthetic 5-10 sentence patient descriptions with clinical trials from ClinicalTrials.gov, a clinical trials repository that includes all clinical trials in the United States. Our system uses BM25 and semantic textual similarity (STS) models to retrieve two thousand candidates from hundreds of thousands of clinical trials. We then proceed to rerank those trials using neural reranking models with BERT-based encoders and novel attention mechanisms. In addition to training our models on an existing related corpus [Koopman and Zuccon, 2016], we leverage data from MIMIC-III to generate a larger training corpus. In the end, we found that our BM25-based ranker utilizing a Lucene index outperformed our neural models, likely due to a lack of high-quality training data.

1 Introduction

In the TREC 2021 CT-Track, we are tasked with matching free-text patient descriptions consisting of 5-10 sentences with real clinical trials from ClinicalTrials.gov. The problem of clinical trial matching is important because many clinical trials fail when they do not meet their recruitment targets. An automated system that matches patients to appropriate clinical trials would help clinicians and patients alike.

Several works have approached the task of matching patients to clinical trials, but most differ from the TREC 2021 CT-Track in how the data are formatted. COMPOSE [Gao et al., 2020] and DeepEnroll [Zhang et al., 2020] use deep learning methods to match trials with patient records; their data notably differ from ours in that the patients are represented with structured electronic health records (EHRs) as opposed to free-text descriptions. The 2018 n2c2 shared task, "Cohort selection for clinical trials," used clinical narratives for 288 patients, but annotated a set of only 13 specific eligibility criteria [Stubbs et al., 2019]. In 2017, 2018, and 2019, the TREC Precision Medicine (PM) Track involved retrieving relevant clinical trials for a set of patients; however, the topics were limited to cancer patients who were represented with semi-structured data (disease, variant, and demographics) as opposed to unconstrained free-text descriptions [Roberts et al., 2017, 2018, 2019].

The setup that is most similar to the TREC 2021 CT-Track comes from Koopman and Zuccon [2016]; in this setup, free-text patient descriptions are matched with a snapshot of trials from ClinicalTrials.gov. However, in addition to the longer patient descriptions, the authors also obtain ad-hoc queries from medical professionals, which represent search terms that those professionals would use to find appropriate trials for each patient. Using a number of baseline models, the authors find that the results using ad-hoc queries exceed those using longer text descriptions for each model and metric evaluated.

*Work was performed as an intern at IBM Research.

Dataset	Patient Descriptions	Clinical Trials	Labeled Pairs
TREC	75	375K	0
SIGIR	60	204K	3870
AutoGT	18K	375K	700K

Table 1: Statistics of three datasets used by our system; the TREC dataset was used for our submitted results and final evaluation, while the SIGIR and AutoGT datasets were used for training our neural rerankers.

The prior work in this field reveals an opportunity to build better models for linking free-text patient descriptions to clinical trials without physician intervention.

2 Data

We evaluate our system on the official 2021 TREC CT-Track topics and clinical trials collection; we also utilize two external datasets. Statistics for the datasets are displayed in Table 1.

TREC The TREC 2021 CT-Track data consists of 75 topics and 375,580 clinical trial XML files from ClinicalTrials.gov. The topics are 5-10 sentence synthetic patient descriptions, developed by individuals with medical training. They are intended to mimic an admission note.

The trials are XML files corresponding to current and historical clinical trials in the United States and elsewhere. Most notably, the trial XML includes an eligibility criteria field. We parse the criteria to separately extract the inclusion and exclusion criteria, relying on the fact that the criteria typically contains a header such as "Exclusion Criteria:" which delimits the two sections. We also make use of additional information provided in the trial XML files for determining the relevance of each trial to the patients, including the conditions, interventions, MeSH terms, and keywords.

In the evaluation phase, each evaluated topic-trial pair is given one of three labels by experts: "eligible" (meets all inclusion criteria and doesn't meet any exclusion criteria), "excludes" (meets all inclusion criteria but also meets some exclusion criteria), and "not relevant." Because the TREC collection did not include any labeled data prior to submission, we rely on two external datasets for training our neural rerankers.

SIGIR The first external dataset we utilize is the SIGIR data collection from Koopman and Zuccon [2016]. The dataset includes 60 topics in three different formats: ad-hoc queries (short queries generated by medical assessors), summaries (average of 22 words), and descriptions (average of 78 words). We use the descriptions in our work as they most closely mimic the TREC topics. The topics were drawn from the 2014 TREC Clinical Decision Support Track [Simpson et al., 2014]. The trials are again extracted from ClinicalTrials.gov, but as they were collected on December 16, 2015, there are fewer trials included than there are in the TREC collection from April 27, 2021.

We note that the labeling scheme differs from that of the TREC CT-Track. In both cases, there are some trials labeled as not relevant and some labeled as fully relevant, e.g. "Highly likely to refer this patient for this clinical trial" or "eligible." However, Koopman and Zuccon [2016] includes an intermediate label indicating "Would consider referring this patient to this clinical trial upon further investigation," while TREC's intermediate label indicates that a patient is excluded, as they meet the inclusion criteria but also some exclusion criteria. In practice, the dataset is so small[1] that we combine the eligible and intermediate labels to create a single positive label.

AutoGT To supplement the small training set, we also create an auto-generated dataset. This dataset is designed by matching the primary diagnosis of MIMIC-III records (notes) with condition MeSH terms in the TREC 2021 clinical trial corpus. Specifically, the diagnosis codes in the structured data and priority codes associated with those diagnosis codes are used to determine the primary diagnosis. To replicate the format of TREC data, we randomly used the history of present illness (HPI) section for some records and a

[1] 4000 judged documents, most of which are judged to be irrelevant

combination of the HPI and past medical history (PMH) sections for others. Using this strategy, we form over 700,000 patient description - clinical trial pairs.

3 Methods

3.1 Query Generation

Our query generation module relies on the IBM Watson Annotator for Clinical Data (ACD) service, a medical domain NLP service featuring a variety of annotation tools that detect, normalize, and code medical and social findings from unstructured clinical data. Based on the observations we made through experimentation on SIGIR-2016 corpus, we developed several heuristics to assign weights to the extracted metadata; these weights along with the text spans serve as weighted ad-hoc queries. Specifically, we used the potential diagnosis, diagnosis, patient reported conditions, and therapeutic procedures from ACD annotations of patient descriptions as candidate concepts.

Next, we derived several features specific to each of these concepts such as entity type, inverse document frequency measured using MIMIC and PubMed, a rare disease boolean derived using a resource from Orphadata, and the section in which the concept is mentioned in the description (e.g., PMH, HPI, symptom probability). As a final step, we normalized and standardized these features and used the sum of the resulting feature values to determine the importance weight of each candidate concept present in the description.

3.2 Retrieval Modules

3.2.1 Lucene

The first model that we use is based on retrieval from a Lucene index using BM25 [Robertson et al., 1995]. We index the trials using the condition and intervention fields from the clinical trial XML files. Then, we retrieve the top n trials for each topic, matching the queries we generated for each topic with the trials in the index. We use query level boosting to boost terms according to the weights extracted from our query generation module.

3.2.2 STS Ranker

Employing the queries and weights from our query generation module, we use a transformer-based semantic textual similarity (STS) model to retrieve trials. First, low-dimensional representations are obtained for all the conditions and interventions present in the trials and the extracted metadata of each topic. Then, we measure weight normalized pairwise cosine similarity between a topic and the low-dimensional representations of all trails to rank them.

Specifically, we compute two versions of the STS score, $STS1$ and $STS2$. For each patient description, we use the set P of representations of query terms extracted by our query generation module, and corresponding weights W^P. Similarly, for each trial, we extract a set T of representations of interventions and conditions and corresponding weights W^T. Then, we use the following two equations to compute the STS scores:

$$STS1 = \sum_{i=1}^{|P|} W_i^P \cdot \left[\max_{j=1}^{|T|} cosine(P_i, T_j) \right]$$

$$STS2 = \frac{1}{|T|} \cdot \sum_{i=1}^{|T|} W_i^T \cdot \left[\max_{j=1}^{|P|} cosine(T_i, P_j) \right]$$

3.2.3 Combined Candidates

In order to combine candidates from the Lucene and STS modules, we compute the average rank for each topic-trial pair using the individual ranks R_{Lucene}, R_{STS1} and R_{STS2} as follows:

$$R_{combo} = 0.5 * min(R_{Lucene}, n) + 0.25 * min(R_{STS1}, n) + 0.25 * min(R_{STS2}, n)$$

where n is the number of total reranked trials we want as output. We return the top n trials according to the R_{combo} score.

3.3 Neural Rerankers

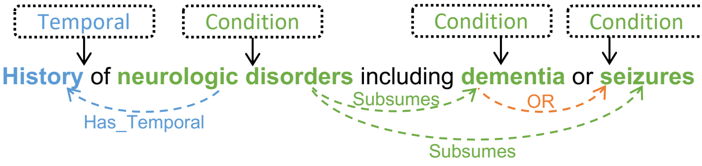

Figure 1: An example of the entities and relations extracted by our CHIA model on criteria from clinical trials.

To rerank results from our retrieval modules, we introduce a novel neural reranker. First, we extract entities from the text of the topic as well as the trial using a system trained on the CHIA dataset, a large annotated corpus of clinical trial eligibility criteria for concept and relation extraction [Kury et al., 2020]. To extract the entities, we used a joint entity and relation extraction model from Wang and Lu [2020]; the system is trained on trial inclusion and exclusion criteria. An example of this extraction is shown in Figure 1.

We encode the text using a BERT-based model; the text includes the entities extracted by CHIA from the topic, inclusion criteria, and exclusion criteria. Additionally, we encode keywords from the trial's interventions, MeSH terms, keywords, and conditions fields.[2] For our SIGIR model, we encode using ClinicalBERT [Alsentzer et al., 2019] and for our AutoGT model we encode using BlueBERT with PubMed abstracts and MIMIC-III [Peng et al., 2019]. We then extract the embeddings for each span, using a convolutional neural network with kernel sizes of 1, 2, 3, and 4 to combine representations for spans with multiple tokens.

Next, inspired by Zhang et al. [2020], we use attention mechanisms to compute alignment between spans in the topics and the criteria, interventions, MeSH terms, keywords, and conditions. These alignments are computed separately; in a final layer, we combine agreement scores across the various parts of the trial. A diagram of our reranking module is provided in Figure 2.

As our loss function, we use WARP loss [Weston et al., 2011], which was designed for learning to rank. We input one positive example and three negative examples to the warp loss function.

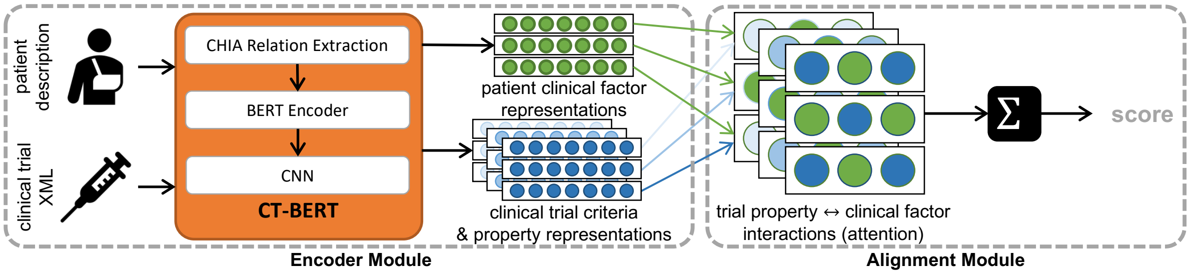
patient description
clinical trial XML
CHIA Relation Extraction
BERT Encoder
CNN
CT-BERT
patient clinical factor representations
clinical trial criteria & property representations
Encoder Module
trial property ↔ clinical factor interactions (attention)
Σ
score
Alignment Module

Figure 2: Our reranking architecture. First, patient descriptions and clinical trial descriptions are fed into our CT-BERT module, which extracts relevant entities using our CHIA relation extraction module and encodes them using a BERT-based encoder. The resulting representations are fed into the alignment module, which computes attention scores between the criteria and clinical factors from the patient descriptions.

						Included in Neural Reranker					
	Lucene	STS	Reranker	Training Data	BERT Model	Inclusion	Exclusion	Conditions	Interventions	MeSH	Keywords
IBMLucene	✓										
IBMSTS		✓									
IBMSIGIR	✓	✓	✓	SIGIR	ClinicalBERT	✓	✓	✓	✓	✓	✓
IBMAUTOGT	✓	✓	✓	AutoGT	BlueBERT	✓			✓	✓	✓
IBMSIGIRACT	✓	✓	✓	SIGIR	ClinicalBERT	✓	✓	✓	✓	✓	✓

Table 2: Summary of the components that make up our five submitted runs.

System	NDCG@10	PREC@10	Reciporical Rank
IBMLucene	**0.3174**	**0.1973**	**0.3913**
IBMSTS	0.2238	0.1480	0.2700
IBMSIGIR	0.1402	0.0893	0.1865
IBMAUTOGT	0.1318	0.0880	0.1350
IBMSIGIRACT	N/A	0.0573	0.1326

Table 3: The results of our runs (NDCG@10, PREC@10, Reciporical Rank) across all topics. The best performance across all metrics is from the IBMLucene system.

4 Results

We submitted five runs for evaluation (see a summary in Table 2). IBMLucene and IBMSTS use the retrieval methods described in Section 3.2.1 and Section 3.2.2 respectively. We retrieve the top 1000 results per topic using each method. In IBMSIGIR, we train the neural reranker described in Section 3.3 with the SIGIR data for training. We choose the best checkpoint based on evaluation on a subset of held out SIGIR data, then rerank the top 2000 results per topic after combining the Lucene and STS results with the method described in Section 3.2.3. In IBMAUTOGT, we use the auto-generated data as training data for the reranking system, then rank the same 2000 trials for each topic. IBMSIGIRACT uses the same model as IBMAUTOGT, but only includes currently active trials, to mimic a real-world scenario.

Somewhat surprisingly, we found that our simplest system (querying the Lucene index) yielded the best results across three metrics (NDCG@10, PREC@10, Reciprocal Rank). The full results for these three metrics are listed in Table 3. The results are consistent across these metrics, with IBMLucene > IBMSTS > IBMSIGIR > IBMAUTOGT > ABMSIGIRACT. Given the limited labeled training data, our results showed that we were unable to leverage deep learning systems for this task.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL `https://aclanthology.org/W19-1909`.

Junyi Gao, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Compose: Cross-modal pseudo-siamese network for patient trial matching. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 803–812, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403123. URL `https://doi.org/10.1145/3394486.3403123`.

Bevan Koopman and Guido Zuccon. 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*, SIGIR '16, page 669–672, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340694. doi: 10.1145/2911451.2914672. URL `https://doi.org/10.1145/2911451.2914672`.

[2] Exclusion criteria and conditions are not used with the AutoGT model based on results from initial experiments on a set of the SIGIR data.

Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific Data*, 7(1): 281, Aug 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00620-0. URL `https://doi.org/10.1038/s41597-020-00620-0`.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5006. URL `https://aclanthology.org/W19-5006`.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, and Shubham Pant. Overview of the TREC 2017 precision medicine track. In *TREC*, 2017. URL `https://trec.nist.gov/pubs/trec26/papers/Overview-PM.pdf`.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, and Alexander J Lazar. Overview of the TREC 2018 precision medicine track. In *TREC*, 2018. URL `https://trec.nist.gov/pubs/trec27/papers/Overview-PM.pdf`.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. Overview of the TREC 2019 precision medicine track. In *TREC*, 2019. URL `https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.PM.pdf`.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

Matthew S Simpson, Ellen M Voorhees, and William Hersh. Overview of the trec 2014 clinical decision support track. In *TREC*, 2014.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 09 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz163. URL `https://doi.org/10.1093/jamia/ocz163`.

Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.133. URL `https://aclanthology.org/2020.emnlp-main.133`.

Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

Xingyao Zhang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Deepenroll: Patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 1029–1037, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380181. URL `https://doi.org/10.1145/3366423.3380181`.