Alibaba DAMO Academy at TREC Clinical Trials 2021: Exploring Embedding-based First-stage Retrieval with TrialMatcher

Qiao Jin[1,2], **Chuanqi Tan**[2], **Zhengyun Zhao**[1], **Zheng Yuan**[1,2], **Songfang Huang**[2]
[1]Tsinghua University, [2]Alibaba Inc.
{jqa14,zhao-zy15,yuanz17}@mails.tsinghua.edu.cn
{qiao.jqa,chuanqi.tcq,songfang.hsf}@alibaba-inc.com

Abstract

This paper describes the submissions of Ailbaba DAMO Academy to the TREC 2021 Clinical Trials Track, where the task is to match eligible clinical trials for given patient notes. Our systems follow the standard retrieval-reranking procedure. We propose a novel embedding-based retrieval model, TrialMatcher, as the retriever. TrialMatcher contains a patient note encoder and a clinical trial encoder pre-trained by 370k clinical trial documents. It retrieves relevant clinical trials based on embedding space distances. We then use different re-rankers to reorder the candidates returned by TrialMatcher. In automatic runs, the re-rankers are trained by a relevant dataset or a synthetic patient-trial relevance dataset. In manual runs, the re-rankers are trained by annotations derived from a human-in-the-loop active learning strategy. Our automatic runs rank the second in all participants on all four metrics. Our manual runs rank the first on one metric, and the second on three other metrics.

1 Introduction

Clinical trials are studies that prospectively study the effects of different treatments on human subjects[1], which is conceptually similar to the A/B testing in software engineering. In most countries, new therapies should be evaluated by clinical trials before approved. It is vitally important to match eligible patients to clinical trials, because: 1) patients have the potentials to greatly benefit from novel treatments, especially when there is no effective therapy available. For example, late-stage cancer patients can be unresponsive to most first- and second-line therapies (Enzinger et al., 2014), and there is even no approved drugs for many rare diseases (Griggs et al., 2009); 2) clinical trials need to recruit enough patients to continue and achieve statistically significant results.

To facilitate automatic patient-to-trial matching, TREC 2021 Clinical Trials (CT) Track[2] releases 75 patient notes (known as topics) and evaluates the relevance of clinical trials returned by participating systems. TREC CT is a challenging task: In terms of topics, they are lengthy, usually containing 5-10 complete sentences. Some are noisy, e.g. with mentions of the husband's disease, or contain ambiguous abbreviations. In terms of the clinical trial documents, they are even longer with many sections such as titles, summaries, inclusion and exclusion criteria. Matching patients with the trials requires entailment-like reference (Zhang et al., 2020), since ideally eligible patients should meet all inclusion criteria and no exclusion criteria.

Traditional first-stage retrievers for this task typically first extract keywords (e.g. using NER models) since the topics are long and noisy, then query the inverted index of clinical trials using BM25 with or without query expansion. However, such methods might not work, since: 1. Keyword extraction tools are imperfect, especially for abbreviations; 2. Diagnoses are not always directly available, so reasoning from symptoms is required; 3. Some records contain many noisy terms that will also be extracted, such as the diseases of relatives and certain past medical histories.

In this paper, we describe the submissions of Alibaba DAMO academy to the TREC CT 2021 track. We first apply TrialMatcher, a novel method that performs Embedding-Based Retrieval (EBR), to find eligible clinical trial candidates for given patients. We then use different re-rankers based on pre-trained language models to re-order the candidates. Our automatic submissions rank the second among all participating teams in 4/4 metrics. Our manual submissions rank the first in 1/4 metric, and the second in 3/4 metrics. The results show that EBR for clinical matching is a promising future direction to explore.

[1] https://grants.nih.gov/policy/clinical-trials/definition.htm

[2] http://www.trec-cds.org/2021.html

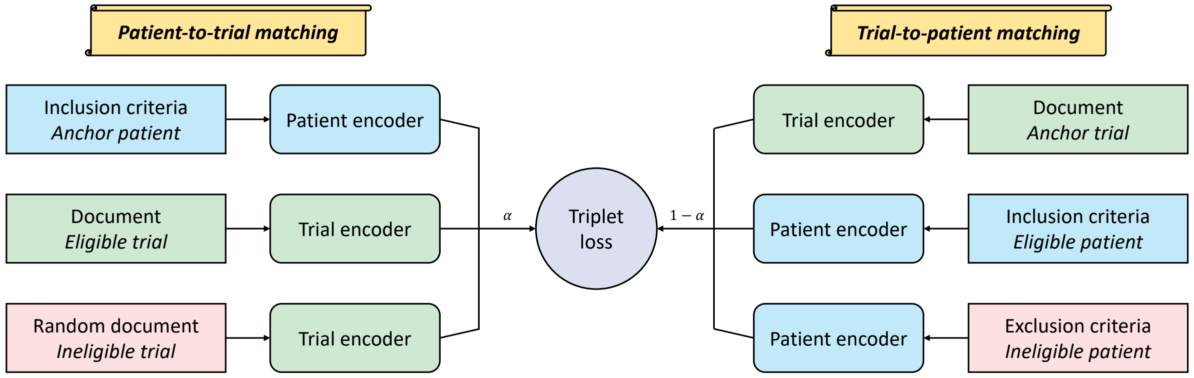

Figure 1: The TrailMatcher pre-training architecture.

2 Methods

Following the standard practice, our systems contain a retriever and a re-ranker. The retriever is denoted as TrialMatcher, performing EBR. The re-ranker is based on the human-in-the-loop active learning method that won the TREC 2020 Precision Medicine Track (Jin et al., 2020).

2.1 TrialMatcher

TrialMatcher performs EBR with pre-trained patient note and clinical trial encoders. The pre-training architecture is shown in Figure 1 and the inference architecture is shown in Figure 2.

2.1.1 Pre-training

Let $\text{Enc}^{\text{patient}}$ and $\text{Enc}^{\text{trial}}$ denote the encoder for free-text patient notes and clinical trial descriptions (consisting of title, condition texts and brief summaries), respectively. They are essentially transformer encoders (Vaswani et al., 2017) initialized by ClinicalBERT (Alsentzer et al., 2019), a clinical version of BERT (Devlin et al., 2019). We have:

$$\text{Enc}^{\text{patient}}(\text{patient note}) \in \mathbb{R}^d$$

$$\text{Enc}^{\text{trial}}(\text{clinical trial description}) \in \mathbb{R}^d$$

where d is the embedding dimension.

We use all clinical trial entries (over 370k) from `clinicaltrials.gov` to pre-train the encoders. Specifically, for each clinical trial entry indexed by i, let's denote its description as D_i, inclusion criteria as I_i, exclusion criteria as E_i. Let D_r be the description of another randomly sampled clinical trial description.

In pre-training, we use inclusion and exclusion criteria as proxies for eligible and ineligible patient notes, respectively. Based on it, we design two contrastive learning tasks: *patient-to-trial* matching and *trial-to patient* matching. We optimize the Triplet loss, where the goal is to minimize the distance between the anchor instance embedding and the positive instance embedding and maximize the distance between the anchor instance embedding and the negative instance embedding:

$$\mathcal{L}_{\text{triplet}}(\mathbf{anc}, \mathbf{pos}, \mathbf{neg}) = \max(0, \text{dist}(\mathbf{anc}, \mathbf{pos}) - \text{dist}(\mathbf{anc}, \mathbf{neg}) + m) \tag{1}$$

where $\mathbf{anc}, \mathbf{pos}, \mathbf{neg}$ denote anchor, positive instance and negative instance embeddings, respectively. dist denotes a specific distance metric, e.g. Euclidean distance. m is the margin hyper-parameter in the triplet loss. The losses for the patient-to-trial matching and the trial-to-patient matching tasks are:

$$\mathcal{L}_{\text{patient-to-trial}} = \mathcal{L}_{\text{triplet}}(\text{Enc}^{\text{patient}}(I), \text{Enc}^{\text{trial}}(D), \text{Enc}^{\text{trial}}(D_r)) \tag{2}$$

$$\mathcal{L}_{\text{trial-to-patient}} = \mathcal{L}_{\text{triplet}}(\text{Enc}^{\text{trial}}(D), \text{Enc}^{\text{patient}}(I), \text{Enc}^{\text{patient}}(E)) \tag{3}$$

The final pre-training loss is a weighted sum of them:

$$\mathcal{L}_{\text{pre-training}} = \alpha\mathcal{L}_{\text{patient-to-trial}} + (1-\alpha)\mathcal{L}_{\text{trial-to-patient}} \tag{4}$$

where α is the multi-task hyper-parameter to tune.

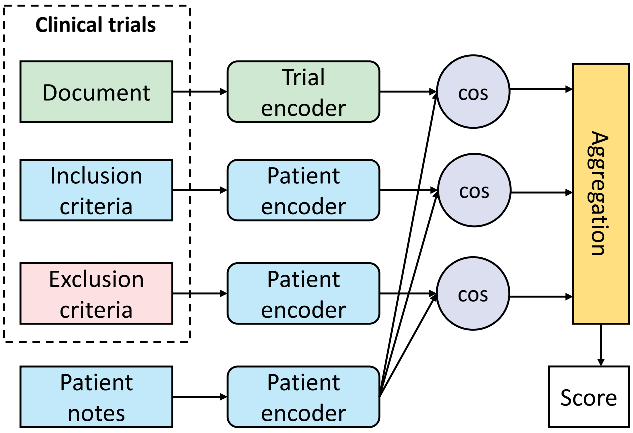

Figure 2: The TrialMatcher inference architecture.

2.1.2 Inference

During inference, let's denote the given patient note as P.

TrialMatcher computes a relevance score s for each candidate clinical trial. For clinical trial j,

$$s_j = \text{Aggregate}(s_{P,D_j}, s_{P,I_j}, -s_{P,E_j}) \quad (5)$$

where Aggregate denotes an aggregation method (e.g. averaging), and:

$$s_{P,D_j} = \cos(\text{Enc}^{\text{patient}}(P), \text{Enc}^{\text{trial}}(D_j)) \quad (6)$$

$$s_{P,I_j} = \cos(\text{Enc}^{\text{patient}}(P), \text{Enc}^{\text{patient}}(I_j)) \quad (7)$$

$$s_{P,E_j} = \cos(\text{Enc}^{\text{patient}}(P), \text{Enc}^{\text{patient}}(E_j)) \quad (8)$$

where cos denotes the cosine similarity.

The clinical trial candidates are ranked by s and the highest ranked ones will be retrieved for re-ranking.

2.2 Re-ranking

We have submitted five runs, namely damoebr, damoebrsigir, damoebrtog, damohitl and damohitls. They all use the same TrialMatcher retriever, and only differ in the re-ranking step. Damoebr does not re-rank the TrialMatcher results, so it provides a baseline performance of TrialMatcher. Damoebrsigir and damoebrtog use automatic re-rankers, which will be introduced in Section 2.2.1. Damohitl and damohitls use manual re-rankers, which will be described in Section 2.2.2.

2.2.1 Automatic re-rankers

For damoebrsigir, we fine-tune a clinicalBERT with a similar patient-to-trial matching dataset (Koopman and Zuccon, 2016) that contains 60 patients and 4000 relevance annotations (which we refer to as the SIGIR dataset in this paper), and use the fine-tuned clinicalBERT as the re-ranker. Instead of encoding the patient notes and clinical trials separately, the concatenation of patient notes and clinical trials is sent to clinicalBERT during fine-tuning and re-ranking.

For damoebrtog, we fine-tune another clinicalBERT to predict whether a set of criteria is the inclusion or exclusion criteria of a given clinical trial description. The motivation is similar to that of the TrialMatcher pre-training, where we consider a set of criteria as a patient summary, so the ability to predict whether the set is inclusion or exclusion criteria can help the prediction of whether a patient is eligible. We use the whole `clinicaltrials.gov` to construct the fine-tuning dataset. During fine-tuning, the concatenation of I or E and D are sent to clinicalBERT to predict inclusion as 1 or exclusion as 0. During inference, the concatenation of P and D is fed to the re-ranker to predict the eligibility.

2.2.2 Manual re-rankers

The damohitl and damohitls re-rankers are based on the human-in-the-loop active learning method that won the TREC 2020 Precision Medicine Track (Jin et al., 2020). In this year, we have annotated about 1.7k instances in active learning.

Submissions	NDCG@10	Precision@10	Mean Reciprocal Rank	R-Precision
Automatic runs				
f_t_mt5_2	**71.18**	**59.33**	**81.62**	26.28
f_t_mt5	67.92	54.93	71.61	**26.39**
damoebrtog (ours)	59.53	40.93	60.83	21.91
CSIROmed_inc	53.20	31.73	NA	NA
CSIROmed_abs	52.85	32.40	NA	NA
RM3Filtered	51.49	33.60	49.36	20.78
ielabr2	49.92	32.40	51.19	19.91
damoebrsigir (ours)	48.21	40.40	58.41	16.81
IKR3_BSL	47.85	NA	NA	NA
damoebr (ours)	34.09	25.87	42.93	18.65
Manual runs				
tdminerrun3	**71.50**	**57.60**	NA	NA
tdminerrun4	71.50	57.07	82.57	24.40
tdminerrun2	71.08	56.80	**83.15**	24.55
tdminerrun1	70.78	NA	82.53	24.52
damohitl (ours)	70.38	56.93	75.75	**27.82**
damohitls (ours)	70.28	56.93	75.30	27.73

Table 1: Evaluation results of different submissions, ranked by NDCG@10. **Bolded** numbers denote best performance, and underlined numbers denote the second best (team-wise) results. All numbers are percentages.

3 Results

The results of different submissions are shown in Table 1.

Automatic submissions: Damoebrtog ranks the second by 4/4 metrics among all participating teams. However, the best submissions (f_t_mt5 and f_t_mt5_2) largely outperform the damoebrtog, indicating that there is still much room for improvements. Within our submissions, damoebrtog is much better than damoebrsigir, which in turn is much better than damoebr. These results show that: 1. clinicalBERT that is fine-tuned on the SIGIR dataset improves the original EBR results; and 2. self-supervised fine-tuning with large-scale instances from `clinicaltrials.gov` is better than fine-tuning with the clean but small SIGIR dataset.

Manual submissions: The best manual submissions are better than the best automatic submissions by NDCG@10, R-Prec and MRR, but the best automatic run (f_t_mt5_2) surprisingly outperforms the best manual run (tdminerrun4) by P@10 (0.5933 v.s. 0.5760). Damohitl achieves the highest RPrec of 0.2782, outperforming the second team (0.2455) by a large margin. Damohitl and damohitl rank the second by the other 3/4 metrics, and the results are comparable to the best team (0.7038 v.s. 71.50 by NDCG@10, 0.5693 v.s. 0.5760 by P@10). As expected, our manual submissions damohitl and damohitls are much better than our automatic submissions, showing the importance of training re-rankers with task-specific annotations.

4 Conclusion

In this paper, we describe the submissions of Alibaba DAMO Academy to the TREC Clinical Trials 2021 track. We propose a novel embedding-based retrieval model, TrialMatcher, and use it with several re-rankers fine-tuned by different datasets. The submissions rank the first by 1/8 metric, and the second by all other 7/8 metrics. Overall, our results show that: 1. embedding-based retrieval is useful but not sufficient; 2. re-rankers trained by in-domain annotations can largely improve the performance; 3. a large set of noisy instances might be better than a small set of clean instances for training the re-ranker. Potential future works include thoroughly characterizing the embedding-based retrieval model, and exploring other pre-training schemes for the patient-to-trial matching task.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea C Enzinger, Baohui Zhang, Jane C Weeks, and Holly G Prigerson. 2014. Clinical trial participation as part of end-of-life cancer care: associations with medical care and quality of life near death. *Journal of pain and symptom management*, 47(6):1078–1090.

Robert C Griggs, Mark Batshaw, Mary Dunkle, Rashmi Gopal-Srivastava, Edward Kaye, Jeffrey Krischer, Tan Nguyen, Kathleen Paulus, Peter A Merkel, et al. 2009. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular genetics and metabolism*, 96(1):20–26.

Qiao Jin, Chuanqi Tan, Mosha Chen, Ming Yan, Songfang Huang, Ningyu Zhang, and Xiaozhong Liu. 2020. Aliababa DAMO academy at TREC precision medicine 2020: State-of-the-art evidence retriever for precision medicine with expert-in-the-loop active learning. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xingyao Zhang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2020. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, pages 1029–1037.