JBNU at TREC 2022 Clinical Trials Track

Dalya Sin, Woo-Kyoung Lee, Seung-Hyeon Jo, Kyung-Soon Lee

Division of Computer Science and Engineering, CAIIT
Jeonbuk National University, Republic of Korea
{dalyasin, ukyoung147, jackaa, selfsolee}@jbnu.ac.kr

Abstract

This paper describes the participation of the JBNU team at the TREC 2022 Clinical Trials Track. Our approach is to focus on clinical terms detected from the ClinicalBERT and BioBERT. In order to expand clinical terms, the synonym terms are extracted by BERT embedding model. Our experimental results showed 0.4527, 0.3220 and 0.5543 by NDCG@10, P@10 and MRR, respectively.

1. SUBMITTED RUNS

In the TREC 2022 Clinical Trials track (Roberts et al., 2021), given a topic of a synthetic patient case, the system's goal is to retrieve clinical trials with a judgment of eligible, excludes, and not relevant. We submitted two runs, jbnu1 and jbnu2 to TREC 2022 Clinical Trials Track. Since a patient health record is provided as a topic, it is possible to use clinical terms to obtain suitable clinical trials data for exact patients. In our experiments, we used the ClinicalBERT and BioBERT model to detect clinical terms. For query expansion, we use terms from the BERT embedding for a topic to expending clinical terms up to two synonym words.

Clinical BERT(Alsentzer, Emily, et al., 2019) pre-trained over MIMIC patient notes about 2 million notes of clinical text and fine-tuned. Clinical BERT detects named entities such as Problem, Treatment and Test. In our experiments, we only use clinical terms which belong to the Problem entity. The Problem entity recognizes disease names and symptoms of a patient. This is good to retrieve similar patient's clinical trials.

BioBERT(Lee et al., 2019) pre-trained over a large-scale biomedical on corpora. BioBERT initialized over BERT's pre-trained 108M parameters, 4.5B of words trained on PubMed abstracts, 13.5B of words trained over PubMed Central full-text articles. In our experiments, we used the clinical terms from BioBERT such as Disease, Drug or Chemical, Gene or protein, and Species entities. We use this model to extract biomedical terms for a topic, since it has better results on biomedical text mining tasks.

BERT(Devlin et al.,2018) parameters are trained on English Wikipedia and BookCorpus. It has a great performance on non-contextual embedding tasks. We use this model to get a similar context of clinical terms to expand the scope of search in clinical terms.

We use Elasticsearch for the BM25 model for retrieval. All the fields in a document are used for retrieval such as brief title, brief summary, detailed description, inclusion criteria, and exclusion criteria fields. The parameters for BM25 model are set to 2.0 and 0.8 for k1 and b, respectively.

Our approaches are described as follows:

- **baseline:** All the terms are used for retrieval for each topic.
- **jbnu1**: The terms for a topic and clinical terms detected by ClinicalBERT and BioBERT are used for retrieval. For exact phrase matching, clinical terms are retrieved by proximity match.
- **jbnu2**: The synonym terms extracted by BERT embedding model are expanded to **jbnu1**.

2. Experimental Results

The experimental results are shown in Table 1.

Run Name	NDCG@10	P@10	MRR
Baseline: All the terms for a topic	0.2641	0.1820	0.3907
Baseline + BioBERT terms	0.3786	0.2660	0.4681
Baseline + ClinicalBERT terms	0.4350	0.3140	**0.5775**
jbnu1: Baseline + BioBERT + ClinicalBERT	**0.4530**	0.3200	0.5296
jbnu2: jbnu1 + BERT expanding terms	0.4527	**0.3220**	0.5543

Table 1. Our experimental results for TREC 2022 Clinical Trials Track.

Figure 1 shows our results compared with the median performance at NDCG@10, P@10.

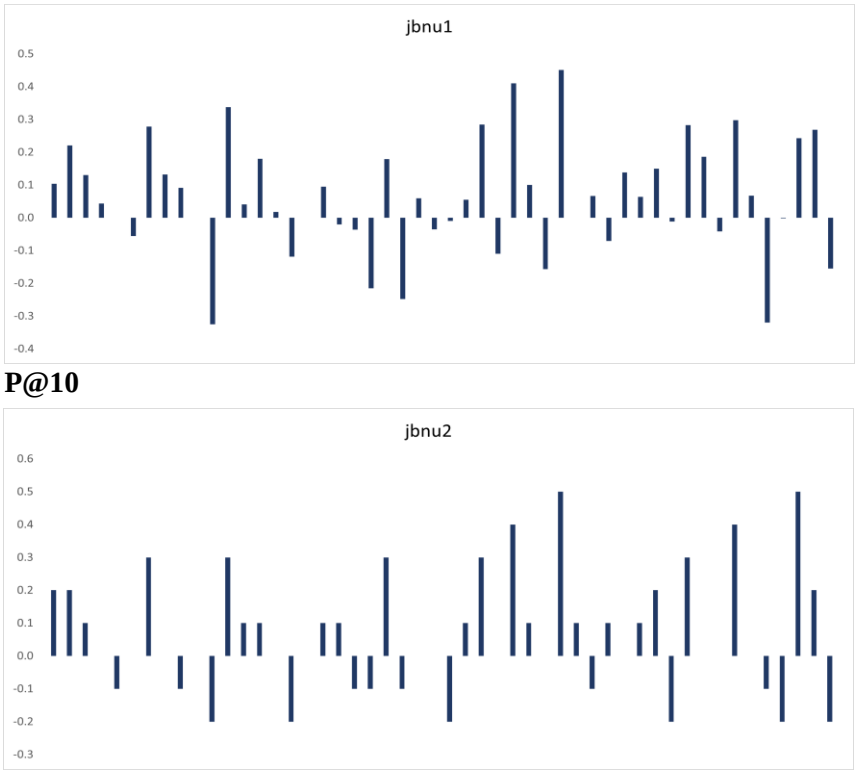

Figure 1. Our results compared with the median performance.

3. RESULT ANALYSIS

We have analyzed our results for a topic 44. It showed 1.0 for both P@5 and NDCG@5. The clinical terms and expanded terms for the topic are as follows:

Topic 44 A 48-year-old man comes to the office complaining of heartburn and acid reflux. He has taken over-the-counter antacids but sees no relief. Other medical history is unremarkable. The patient does not use tobacco, alcohol, or illicit drugs. Vital signs are within normal limits. BMI is 31 kg/m2. Physical examination is positive for mild tenderness in upper stomach. Chest x-ray shows an air-fluid opacity behind the heart. A barium swallow study reveals approximately 1/3 of the stomach herniating through the esophageal hiatus.
Clinical terms by ClinicalBERT: heartburn, acid reflux, mild tenderness in upper stomach, an air-fluid opacity behind the heart, the stomach
Biomedical terms by BioBERT: heartburn, antacids, tobacco, illicit drugs, BMI, stomach, x-ray, barium, stomach
Expanded synonym terms by BERT embeddings: 'indigestion', 'painkillers' are expanded for ‘heartburn’ 'sulfuric', 'hydrofluoric' are expanded for ‘acid’ 'cigarette', 'marijuana' are expanded for ‘tobacco’ 'sulfate', 'hydroxide’ are expanded for ‘barium’ …

Table 2. Example of detected clinical terms and synonym expansion.

4. CONCLUSION

In this paper, we describe our experiments using clinical terms detected by ClinicalBERT and BioBERT for a topic to retrieve eligible clinical trials. Our experimental results showed higher performance by combining clinical terms from ClinicalBERT and BioBERT and expanding synonym terms by BERT. For future work, the clinical terms detected from clinical trials can be used for considering contexts for a topic and documents.

ACKNOWLEDGMENT

This research was supported by MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW)(2022-0-01067) supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation).

REFERENCES

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available Clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Elasticsearch (2015). elasticsearch/elasticsearch. https://github.com/elasticsearch/elasticsearch.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S., & Hersh, W. R. (2022). Overview of the TREC 2022 Clinical Trials Track. In *Proceedings of the 31st Text REtrieval Conference*.