TREC 2021 Clinical Trials
Submission for Universidad del País Vasco

Jordan Koontz
Ixa (www.ixa.eus)
(UPV/EHU)
jkoontz001@ikasle.ehu.eus

Maite Oronoz
HiTZ (www.hitz.eus)
(UPV/EHU)
maite.oronoz@ehu.eus

Alicia Pérez
HiTZ (www.hitz.eus)
(UPV/EHU)
alicia.perez@ehu.eus

Abstract

This paper describes the University of the Basque Country's submission to the TREC 2021 Clinical Trials Track. We begin by summarizing the documents by extracting medical entities. Next, we utilize multilingual and scientific domain sentence embeddings to represent the summarized clinical trials descriptions and the patient topic documents. Lastly, we rank the clinical trial relevance by calculating the cosine similarities between texts.

1 Introduction

The goal of the TREC 2021 Clinical Trials track was to retrieve relevant clinical trials for a given topic.

Our objective was to produce a simple automatic system that was capable of accurately retrieving relevant clinical trials with minimal feature engineering. Moreover, we set out to:

- Summarize the clinical trials and the topic documents.
- Produce semantically meaningful sentence embeddings.

2 Methods

In order to distill the documents to dense yet meaningful representations, we first trained and applied a named entity recognition (NER) system as described in section 2.1. Next, with these summarized texts, we produced sentence embeddings. Details are given in section 2.2. Based on the aforementioned representations, we set, as relevance metrics to retrieve relevant documents, alternative similarity approaches presented in section 2.3.

2.1 Document Summarization

For our summarization strategy, we decided to extract named entities from both the clinical trial descriptions and the patient topic documents. Specifically, we extracted the following medical entity types: Problem, Treatment, and Test. For training data, we used the i2b2/VA 2010 corpus (Uzuner et al., 2011), which we converted to CoNLL format (Sang and Meulder, 2003). Next, we fine-tuned the Bio+ClinicalBERT model (Alsentzer et al., 2019), which was initialized from BioBERT (Lee et al., 2020) and trained on all MIMIC notes (Johnson et al., 2016). Lastly, we applied our fined-tune model to the clinical trial descriptions and the patient topic documents to obtain a distilled corpus.

2.2 Sentence Embeddings

To represent the summarized clinical descriptions and patient topic documents, we utilized pre-trained sentence embeddings. Moreover, we wished to evaluate the effectiveness of multilingual and scientific domain models. For the multilingual model, we selected the Language-agnostic BERT Sentence Embedding (LaBSE) which supports 109 languages (Feng et al., 2020). For the scientific domain embeddings, we selected the SPECTER (allenai-specter) model which can be used to produce document-level embeddings for scientific documents without the need for task-specific fine-tuning (Cohan et al., 2020).

2.3 Selection criterion: similarity

The difference between each of our runs rests on the similarity employed to rank the retrieved documents. Starting from cosine similarity we figured out different re-rankings.

For our first two runs (denoted as 'LaBSE' and 'specter'), we used, respectively, LaBSE and the allenai-specter embeddings. Next, we strictly compare text similarity between the clinical trials and the patient topic documents with cosine similarity and select the top 1000 most similar documents.

For the next two runs (denoted as 'LaBSE rerank' and 'specter rerank'), we produced a binary bag-of-words (BoW) representation of the texts and then re-ranked the scores from runs 1 and 2. The binary BoW representation involves as many features as the size of the vocabulary (i.e. the set of unique words present in the patient topic document). The text vectorization is merely a binary vector with each component representing the presence or absence of the corresponding word in the vocabulary (defined as 1 if the word is present in the vocabulary else 0). Next, cosine similarities are calculated as in runs 1 and 2. Finally, the weighted average of the cosine similarities between the BoW representations and those of the sentence embeddings from runs 1 and 2 (0.2 and 0.8 respectively) is calculated to produce a reranked score. The top 1000 documents with the highest scores are again selected.

Lastly (the run denoted 'specter rerank2'), another BoW representation is created, by contrast, this is not binary, instead, we used the Term Frequency (aka TF). The cosine similarity between texts is calculated and the weighted average calculated between the TF scores and the allenai-specter sentence embedding scores from run 2 (i.e. specter); The top 1000 scores are again selected.

3 Experimental results

Table 1 shows the NDCG and precision at 10 results for the five runs.

Run Name	NDCG@10	PREC@10
LaBSE	0.2551	0.1347
specter	0.2555	0.1480
LaBSE rerank	0.2900	0.1413
specter rerank	0.3614	0.2093
specter rerank2	0.2694	0.1547

Table 1: Experimental results for each run

It is clear that representing the summarized texts with sentence embeddings alone has its limitations. LaBSE and specter achieve nearly identical NDCG@10 scores, while specter has a slightly improved PREC@10 score.

The inclusion of the binary BoW representation yields notable improvements for the runs 2 and 3 (LaBSE rerank and specter rerank). Specter rerank achieves an improvement of 0.1059 and 0.0613 in NDCG@10 and PREC@10 respectively.

However, the TF representation for the final run, specter rerank2, results in modest improvements from specter: 0.0139 and 0.0067 for NDCG@10 and PREC@10 respectively.

4 Concluding remarks and future work

While a classical approach for this type of tasks might have included Information Retrieval as the core engine, instead, we turned to a simple approach, based on document similarity. Distilling the documents to only medical problems, treatments, and tests, undoubtedly results in the exclusion of relevant information about the patient, such as age, gender, and language. For future work, a NER system trained to extract a broader set of entities should be evaluated. We observe that the use of the binary BoW representation helps to ameliorate the exclusion of relevant terms that are discarded by the NER system.

5 Acknowledgemetns

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research. This work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED), European Commission (FEDER) and the Basque Government (IXA IT-1343-19).

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.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. Scientific data.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. CoRR, cs.CL/0306050.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text". Journal of the American Medical Informatics Association : JAMIA, 18(5):552–6.