Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210848

An Evaluation of Pretrained BERT Models for Comparing Semantic Similarity Across Unstructured Clinical Trial Texts

Jessica PATRICOSKI[a], Kory KREIMEYER[b], Archana BALAN[b], Kent HARDART[b], Jessica TAO[b], the Johns Hopkins Molecular Tumor Board Investigators[b], Valsamo ANAGNOSTOU[b] and Taxiarchis BOTSIS[a,b,1]
[a] *Biomedical Informatics and Data Science Section, Johns Hopkins University School of Medicine, Baltimore, MD*
[b] *Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD*

Abstract. Processing unstructured clinical texts is often necessary to support certain tasks in biomedicine, such as matching patients to clinical trials. Among other methods, domain-specific language models have been built to utilize free-text information. This study evaluated the performance of Bidirectional Encoder Representations from Transformers (BERT) models in assessing the similarity between clinical trial texts. We compared an unstructured aggregated summary of clinical trials reviewed at the Johns Hopkins Molecular Tumor Board with the ClinicalTrials.gov records, focusing on the titles and eligibility criteria. Seven pre-trained BERT-Based models were used in our analysis. Of the six biomedical-domain-specific models, only SciBERT outperformed the original BERT model by accurately assigning higher similarity scores to matched than mismatched trials. This finding is promising and shows that BERT and, likely, other language models may support patient-trial matching.

Keywords. Clinical trial, word embeddings, bidirectional coder representations

1. Introduction

Clinical texts often contain unstructured information that requires applying advanced text processing methods to support specific tasks like patient-trial matching. Various language models have the potential to process biomedical and clinical texts to aid in these challenges. The Bidirectional Encoder Representations from Transformers (BERT) family has shown promise in solving multiple problems, including semantic similarity. BERT models have been trained on multiple corpora, including PubMed abstracts, PMC full-text articles, clinical notes, and synthetic vocabularies [1-6].

In this study, we report the performance of pre-trained BERT-based language models by assessing the level of similarity between clinical trial raw text (official titles and eligibility criteria) drawn from two sources: an Institutional unstructured aggregated summary of clinical trials and ClinicalTrials.gov.

[1] Taxiarchis Botsis, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, 1550 Orleans St, CRB2-Rm 153, Baltimore, MD 21287, USA; E-mail: tbotsis1@jhmi.edu.

2. Methods

We created an unstructured aggregated summary of clinical trials reviewed at the Johns Hopkins (JH) Molecular Tumor Board at the Sidney Kimmel Cancer Center. Our corpus included all clinical trials registered as of June 1st, 2021. We chose to focus on trials with no provided National Clinical Trial identifier (NCT ID), incorrectly formatted identifiers, or unrecognized identifiers, in order to capture the challenges of missing data. Figure 1 outlines a breakdown of the dataset and the final subset used to create a data frame of the official study titles and eligibility criteria (EC) for the 689 trials with non-null EC fields.

To create pairs of clinical trial data between the clinical trials in our registry and those registered with ClinicalTrials.gov, we downloaded a corpus of official study titles (and associated NCT IDs) for all clinical trials registered with ClinicalTrials.gov from the Clinical Trials Transformation Initiative's Aggregate Analysis database (AACT) [7]. After basic preprocessing, embeddings (i.e., real-valued vector representations for each word based on context) for each querying title (from our clinical trials) and the 387,486 corpus titles were created using Sentence-Transformers [6, 8].

For each querying title, the corpus title with the highest cosine similarity score was assigned as a potential match. After manually reviewing the potential matches, each pair was labeled as either belonging to the same trial (hereafter, a "match") or different trials (hereafter, a "mismatch"). Of the pairs created, 603 pairs were matches and 86 were mismatches. Finally, the eligibility criteria for each paired ClinicalTrails.gov trial were retrieved through ClinicalTrials.gov's Application Programming Interface URLs using the linked NCT IDs from the AACT database.

The first pre-trained model used to calculate the similarity between the created pairs was an uncased version of BERT-Base model [1]. We also used six more models from the BERT family with the same architecture as the Base model but domain-specific pre-training and fine-tuning using different corpora. Each of the BioBERT [2], BlueBERT [3], Clinical BioBERT [4], SciBERT [4], PubMedBERT [5], and CODER [6] models created embeddings for and checked the cosine similarity between pair titles and between pair EC. The difference in mean similarity scores for matched and mismatched pairs was compared across all models to assess performance. It is important to note that because treatment information in our aggregated summary was not separated from the EC, the EC field contained noise that might have affected similarity calculations.

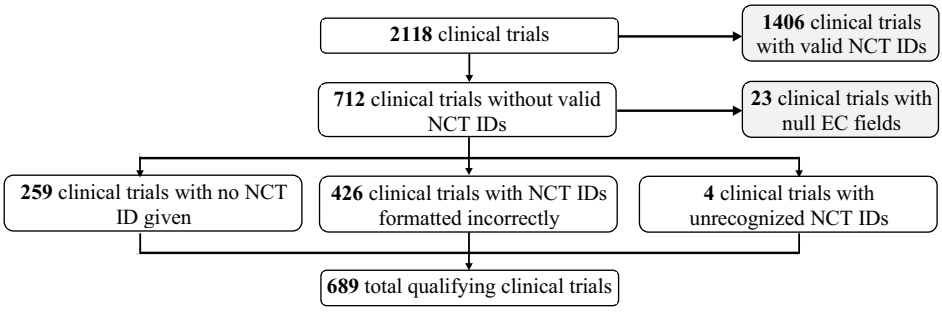

Figure 1. A visual breakdown of the clinical trial summaries. Grey-shaded boxes indicate excluded trials.

3. Results

Visual representations of the similarity scores calculated by the BERT-Base and other pre-trained models for titles and eligibility criteria are shown in Figure 2. SciBERT, while ranking lowest in mean similarity of matching pairs, demonstrated the largest overall difference in mean similarity between matched and mismatched pairs (0.141 for titles and 0.064 for EC) and assigned lower values for mismatched pairs than all other models. Thus, we consider that SciBERT had the highest performance for the given task. Outside of SciBERT, no other model outperformed BERT, which had an overall difference in mean similarity between matched and mismatched pairs of 0.102 for titles and 0.054 for EC. Table 1 includes two examples of matched and mismatched trial titles and the similarity scores for the seven BERT models. As expected, the mean similarity scores for trial EC were generally lower than those for title scores, likely due to the noise present in the EC field of our aggregated clinical trial summaries.

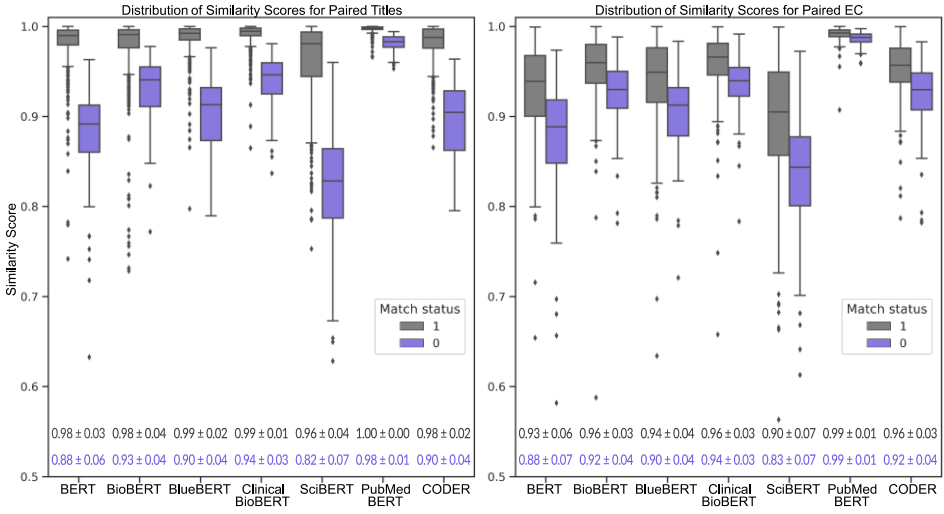

Figure 2. Boxplots showing the spread of similarity scores for paired clinical trial titles (left) and eligibility criteria (right), grouped by match status. A match status of 1 indicates the content assessed belonged to the same clinical trial, while a match status of 0 indicates the content belonged to separate trials. The mean similarity scores and standard deviations, also grouped by match status, are listed along the x-axis.

Table 1. Examples of compared trial titles and their similarity scores for matched and mismatched pairs.

Example Match (Title Similarity)		
"A Ph. II Study of the Efficacy and Safety of SU011248 in Patients with Metastatic Breast Cancer"	BERT: 0.986 BioBERT: 0.973 BlueBERT: 0.973 Clin. BioBERT: 0.974	SciBERT: 0.977 PubMedBERT: 0.997 CODER: 0.984
"A Phase 2 Study Of The Efficacy And Safety of SU011248 In Patients With Metastatic Breast Cancer"		
Example Mismatch (Title Similarity)		
"Donor Lymphocyte Infusions (DLI) plus Rapamycin to Decrease Toxicity Associated with DLI"	BERT: 0.858 BioBERT: 0.901 BlueBERT: 0.875 Clin. BioBERT: 0.916	SciBERT: 0.649 PubMedBERT: 0.957 CODER: 0.862
"Rapamycin in Relapsed Acute Lymphoblastic Leukemia"		

4. Discussion

Overall, SciBERT performed best in terms of distinguishing between sets of trial texts belonging to the same clinical trial and sets of trial text belonging to different clinical trials. The largest difference between SciBERT and most of the other models is its use of SciVocab, which overlaps with BaseVocab used in the BERT-Base by only 42% [9]. SciBERT was trained from scratch and did not use BERT's weights as initialization [4, 9]. After SciBERT, BERT had the best performance. Although trained with biomedical texts, the remaining models did not show promise in efficiently supporting our matching task.

A significant limitation of this study was that due to the inconsistent format of our corpus, treatment information could not be separated from the EC, resulting in an intangible level of interference in similarity assessment. However, the handling of noisy information is a standard challenge in free-text processing and comparison that was successfully handled by some of the selected models. The second major limitation is that the BERT architecture has a maximum token length of 512, and as a result, most EC texts were not compared in full. We acknowledge that this limitation may have introduced a bias to the similarity calculations but did apply to all models that used the same token length.

Future research is needed to address the above limitations and investigate other biomedical language model architectures, the impact of pre-training on clinical trials data, and the feasibility of integrating semantic similarity techniques for comparing clinical trial identities and best utilizing them into the patient matching process.

References

[1] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of NAACL-HLT; 2019 Jun 2-7; Minneapolis, MN. Stroudsburg (PA): ACL; c2019;4171-4186.

[2] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.

[3] Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In: Demner-Fushman D, Cohen KB, Ananiadou S, Tsujii J, editors. Proceedings of the BioNLP workshop; 2019 Aug 1; Florence, IT. Stroudsburg (PA): ACL; c2019. p. 58-65.

[4] Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. Journal of Biomedical Informatics. 2019;100:100057.

[5] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. arXiv: 2007.15779v5 [Preprint]. 2020 [cited 2021 Aug 20]: [24 p.]. Available from: https://arxiv.org/abs/2007.15779.

[6] Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge infused cross-lingual medical term embedding for term normalization. arXiv: 2011.02947 [Preprint]. 2017 [cited 2021 Aug 20]: [11 p.]. Available from: https://arxiv.org/pdf/2011.02947.pdf.

[7] Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, Pietrobon R. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. PLoS One. 2012;7(3):e33677.

[8] Reimers J, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2019 Nov 3-7; Hong Kong, CN. Stroudsburg (PA): ACL; c2019;3892-3992.

[9] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2019 Nov 3-7; Hong Kong, CN. Stroudsburg (PA): ACL; c2019;3615-33620.