Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240539

Comparative Evaluation of Pre-Trained Language Models for Biomedical Information Retrieval

Franziska WEBER[a] and Dennis TODDENROTH[a,1]
[a] *Medical Informatics, University Erlangen-Nuremberg, Germany*

Abstract. Finding relevant information in the biomedical literature increasingly depends on efficient information retrieval (IR) algorithms. Cross-Encoders, SentenceBERT, and ColBERT are algorithms based on pre-trained language models that use nuanced but computable vector representations of search queries and documents for IR applications. Here we investigate how well these vectorization algorithms estimate relevance labels of biomedical documents for search queries using the OHSUMED dataset. For our evaluation, we compared computed scores to provided labels by using boxplots and Spearman's rank correlations. According to these metrics, we found that Sentence-BERT moderately outperformed the alternative vectorization algorithms and that additional fine-tuning based on a subset of OHSUMED labels yielded little additional benefit. Future research might aim to develop a larger dedicated dataset in order to optimize such methods more systematically, and to evaluate the corresponding functions in IR tools with end-users.

Keywords. Biomedical Information Retrieval, Sentence-BERT, Cross-Encoder, ColBERT, OHSUMED

1. Introduction

As the amount of biomedical literature continues to grow, finding relevant information increasingly depends on efficient information retrieval (IR) algorithms. PubMed currently uses a two-stage retrieval algorithm called *Best Match*, which combines term-based BM25-scoring with a machine-learning model based on features that characterize candidate documents and queries [1]. Even though *Best Match* may improve upon the previously used reverse chronological order, more recent methods from computational linguistics such as neural architectures promise to underpin even better relevance models [2]. Popular general search engines have already deployed such neural retrievers since 2019 [3].

Bidirectional Encoder Representations from Transformers (BERT) models constitute a family of language representation models that compute contextualized vector embeddings of the input tokens by passing them through multiple transformer encoder layers. BERT models are bidirectional, which means that the embedding of a token depends on its right and its left attention-weighted context in every layer [4]. Word-masking experiments indicate that this bidirectionality is a key factor for nuanced

[1] Corresponding Author: Dennis TODDENROTH; E-mail: dennis.toddenroth@fau.de.

semantic representations [4]. Furthermore, unlike classical term-based approaches, BERT models do not filter out determiners, conjunctions, or prepositions as stop words. Considering such commonly used words might be beneficial for grasping the subtleties of natural language in intuitive search queries.

By adding an additional output layer, BERT models can be fine-tuned for different tasks like document classification or named entity recognition [4]. There is also a range of BERT variants designed to compute vector representations for IR applications. Here we evaluate how well three of these IR-specific BERT variants reflect the relevance of biomedical documents for a given query.

2. Methods and Data

To evaluate BERT-derived relevance estimates for biomedical IR, we consider three task-specific vectorization algorithms, namely Cross-Encoders, Sentence-BERT, and ColBERT. Cross-Encoders run concatenations of queries and documents through the network simultaneously and derive scores with a simple regression. Due to the large number of embedded pairs, Cross-Encoders are computationally expensive and not suited for large-scale IR [5]. Sentence-BERT calculates separate fixed-size embeddings for queries and documents via pooling operations on the tokens. The resulting high-dimensional vectors can encode rich semantics, so that simple operations such as dot products may effectively represent the relevance of a query-document-pair. Pre-computing document embeddings makes Sentence-BERT potentially more efficient than Cross-Encoders for large-scale IR applications [5].

ColBERT estimates relevance scores as the sum of maximum similarities between query and document tokens, so document tokens can again be precomputed. Instead of having to capture complex query-document relationships in a single vector-based operation, ColBERT thus considers a finer-grained token level. While being beneficial for the quality of ColBERT models, this token-level computation increases the space footprints of the embeddings by an order of magnitude. Instead of the original ColBERT we use its successor model ColBERTv2, which aims to reduce the storage requirements for the embeddings while approximately preserving their quality [6].

We evaluated 21 Sentence-BERT models and 12 Cross-Encoders, which were trained on different general-purpose or biomedical data and provided on the platform Hugging Face[2]. For ColBERT, we only worked with the pre-trained ColBERTv2 checkpoint from the ColBERT GitHub repository[3] because there were no suitable models available on Hugging Face. To investigate how well these models judge the relevance of a document for a query, we computed the agreement of their produced scores with relevance labels from the TREC-9 version [7] of the OHSUMED dataset. This dataset was created in 1994 at the General Medicine Clinic at Oregon Health Sciences University [8] and contains 63 queries and 3875 query-document-pairs labeled as *'possibly relevant'* (`1`) or *'definitely relevant'* (`2`).

We calculated relevance scores with Python 3.11.4. For Sentence-BERT, we called `SentenceTransformer.model.encode` from the `sentence-`

[2] https://huggingface.co/

[3] https://github.com/stanford-futuredata/ColBERT

`transformers` package[4], version 2.2.2, to compute query and document embeddings. As relevance measures, we used the dot product and the cosine similarity of the embeddings. For Cross-Encoders, we applied the function `CrossEncoder.model.predict` from the same package. For ColBERT relevance scores, we used the function `colbert.modeling.colbert.colbert` from the ColBERT GitHub repository.

To visualize the relation between the computed relevance scores and the OHSUMED labels, we compared estimated relevance distributions with boxplots. The association between relevance scores and labels was quantified using Spearman's Rank Correlation Coefficient r_S. We plotted receiver operating characteristic (ROC) curves to display how well the models distinguished between the two relevance labels.

Lastly, we tried to improve the results of the best-performing model of each variant through fine-tuning. As proposed in the OHSUMED data set, we used the data from 1987 consisting of 670 labeled pairs for training. The remaining 3205 pairs from 1988 to 1991 formed the test set. We experimented with using only document titles or concatenations of titles and abstracts as inputs. Fine-tuning was implemented by calling the `fit` method of the `SentenceTransformer` and `CrossEncoder` classes of the `sentence_transformers` library using default training parameters.

3. Results

The performances of the different Sentence-BERT models and Cross-Encoders from Hugging Face varied greatly. Table 1 summarizes metrics achieved by the top-performing model of each of the three considered BERT-variants, while `https://github.com/franziskaweber/bert-biomedical-ir` provides a tabular overview of the results of all of the analyzed models.

The Sentence-BERT model demonstrating the highest performance was the multipurpose model `all-mpnet-base-v2`[5], which was trained on a diverse dataset of over one billion training pairs. For Cross-Encoders, the model `ms-marco-MiniLM-L-12-v2`[6] trained on the IR corpus MS MARCO performed best. The Sentence-BERT model `all-mpnet-base-v2`, the Cross-Encoder `ms-marco-MiniLM-L-12-v2` and the ColBERTv2 checkpoint all performed better with concatenated titles and abstracts than with just the titles. Usage of dot product or cosine similarity, however, made no difference for the results observed with the Sentence-BERT model `all-mpnet-base-v2`.

Table 1. Performances of the top-performing model of each of the considered BERT-variants.

Model Name	BERT variant	Documents	r_S	AUC
all-mpnet-base-v2	Sentence-BERT	Titles + Abstracts	0.28423	0.66
ms-marco-MiniLM-L-12-v2	Cross-Encoder	Titles + Abstracts	0.18925	0.61
ColBERTv2 checkpoint	ColBERT	Titles + Abstracts	0.18340	0.61

The boxplots in figure 1 visualize the difference between the relevance scores produced by these three models for the query-document pairs labeled as either *’possibly*

[4] https://www.sbert.net/docs/

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[6] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

relevant' (`1`) or *'definitely relevant'* (`2`). Furthermore, the corresponding ROC curves illustrate the ability of the models to distinguish between both label types, which achieved areas under the curve between 0.61 and 0.66.

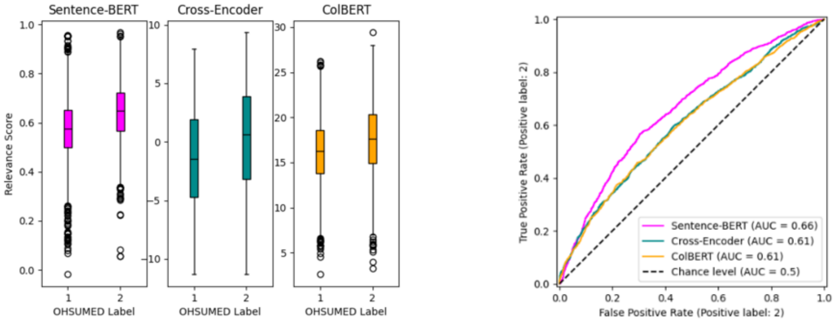

Figure 1. Performances of the top-performing models of each of the considered BERT variants. Visualization of the difference between relevance scores of query-document pairs labeled as 1 and as 2 (left), discrimination of relevance labels (right).

According to Spearman's Rank Correlation, the Sentence-BERT model `all-mpnet-base-v2` performed best with $r_S = 0.28$ compared to $r_S = 0.19$ for the Cross-Encoder `ms-marco-MiniLM-L-12-v2` and r_S=0.18 for the ColBERTv2 checkpoint. This is also reflected in the Sentence-BERT model's ROC-AUC of 0.66, which exceeds the AUCs of the other two models by 0.05. The capability of the evaluated models to distinguish between *'possibly relevant'* and *'definitely relevant'* documents for provided queries is thus moderate, but still substantially better than randomly choosing one of the labels.

The effects of our fine-tuning on the results of these best-performing models were minimal. For instance, in the case of using concatenated titles and abstracts as training and test data, r_S and AUC of the Cross-Encoder model increased by only 0.02 and 0.01 respectively. For the ColBERT model, the AUC also grew by 0.01 while r_S augmented by less than 0.01. The r_S of the Sentence-BERT model even decreased by 0.03 and its AUC by 0.02. Some other investigated models worked less well, and barely outperformed random choices. More detailed results can be found in the referenced repository.

4. Discussion

Although none of the investigated BERT models reproduced relevance labels particularly well, the observed scores still suggest a fundamental capability to assess document relevance. In our experiments, we found that one of the Sentence-BERT configurations outperformed the alternative models. The moderate overall performance could be due to insufficient pre-training for the biomedical domain. The varying results of the different Sentence-BERT and Cross-Encoder models from Hugging Face indicate that pre-training does in fact have a large influence on a model's performance.

We are aware that the OHSUMED dataset from 1994 that we used for our evaluations might no longer be completely up-to-date, and that it is relatively small. For some documents OHSUMED provides only the title but no abstract [7], no documents are explicitly labeled as *'irrelevant'*, and only about 11% of the judgments were duplicated to assess inter-observer reliability [8]. These factors, potentially combined with a sub-optimal choice of the training parameters such as the number of epochs, may also explain why our fine-tuning attempt did not have a more pronounced effect on performance.

5. Conclusions

We evaluated the ability of the three BERT-based vectorization algorithms to reproduce biomedical relevance labels, and found that Sentence-BERT outperformed the alternative approaches. Due to the current success of pre-trained language models, we suspect that some of these approaches will be applicable to biomedical IR tools. This could contribute to better representations of natural language queries and be helpful for exploring biomedical topics with precise retrieval functions. In the context of evolving methods for biomedical IR, a possible next step might aim to develop a dedicated dataset with a larger number of pertinent documents, natural language queries, and manually assigned relevance labels in order to optimize such models more systematically. Additionally, we propose that it would be interesting to evaluate the utility of such IR methods with real end-users in prototypical implementations.

References

[1] Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, Osipov M, Kholodov M, Ismagilov R, Mohan Sunil, Ostell J, Lu Z. Best Match: New relevance search for PubMed. PLoS biology. 2018 Aug;16(8):e2005343, doi: 10.1371/journal.pbio.2005343

[2] Luo M, Mitra A, Gokhale T, Baral C. Improving biomedical information retrieval with neural retrievers. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2022 Jun; Online: p. 11038-11046, doi: 10.1609/aaai.v36i10.21352

[3] Nayak P. Understanding searches better than ever before, 2019 Oct 25. https://blog.google/products/search/search-language-understanding-bert (accessed on 2023 Sept 4)

[4] 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; Minneapolis, USA: ACL; p. 4171-4186, doi: 10.18653/v1/N19-1423

[5] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In: InuiK, Jiang J, Ng V, Wan X, editors. Proceedings of EMNLP-IJCNLP; 2019 Nov; Hong Kong, China: ACL; p. 3982–3992, doi: 10.18653/v1/D19-1410

[6] Santhanam K, Khattab O, Saad-Falcon J, Potts C, Zaharia MA. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In: Marine C, de Marneffe MC, Ruiz M, Vladimir I, editors. Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies; 2022 Jul; Seattle, USA: ACL; p. 3715-3734, doi: 10.18653/v1/2022.naacl-main.272

[7] Robertson SE, Hull DA. The TREC-9 Filtering Track final report. In: Voorhees EM, Harman DK, editors. Proceedings of The Ninth Text Retrieval Conference; 2000 Nov 13-16; Gaithersburg, USA: NIST

[8] Hersh W, Buckley C, Leone TJ, Hickam D. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Croft BW, van Rijsbergen CJ, editors. SIGIR'94: Proceedings of ACM-SIGIR; 1994; London, UK: Springer; p. 192-201, doi: 10.1007/978-1-4471-2099-5_20