Hybrid Re-ranking for Biomedical Information Retrieval at the TREC 2021 Clinical Trials Track

Ming-Xuan Shi[1], Tsung-Hsuan Pan[1], Hen-Hsen Huang[2], and Hsin-Hsi Chen[1]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[1]{mxshi, tspan}@nlg.csie.ntu.edu.tw hhchen@ntu.edu.tw
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]hhhuang@iis.sinica.edu.tw

Abstract

This paper presents our methodology for the task of TREC 2021 Clinical Trials Track, which requires a system to retrieve and return the most relevant biomedical articles after giving queries. We propose a novel approach to biomedical information retrieval by leveraging the term-based and the embedding-based retrieval models with a re-ranking strategy. In our hybrid framework, all the documents will be indexed by using a term-based, efficient search engine. For the given query, a smaller set of candidate results are retrieved from the search engine. The ranking is determined not only by the term-based ranking score but also by the term relationships labeled by the Amazon Comprehend service[1] for refinement. Then, the candidate results are further scored by using the embedding-based model. We represent the document and the query with bioBERT and compute the cosine similarity between a pair of the document embedding and the query embedding as their relevance score. The final score is a linear combination of the term-based and the embedding-based scores. Experimental results show that our hybrid re-ranking method improves both Precision@k and R-precision scores on the 2016 Clinical Decision Support Track and 2021 Clinical Trials Track dataset.

1 Introduction

In handling challenging cases, clinicians often seek out relevant biomedical articles to assist in making a better decision. To better enable access to the scientific literature in the clinical scenario, it is necessary to explore information retrieval methods that connect clinical notes with the published literature. TREC Clinical Trials Track (CT) [1] provides an opportunity for the development of related information retrieval technology. For the purpose of this track, TREC Clinical Traits Track provided retrieved clinical trials from ClinicalTrials.gov. Clinical trial descriptions can be a long document, in which the core aspect of the trial description are the inclusion/exclusion criteria. Each of the topics, also a long free text description consisting of five to ten sentences, presents a patient's admission statement in an EHR (electronic health record).[2]

For the aim of this task, we develop a hybrid model that leverages the advantages of the term-based and the embedding-based ranking models. The term-based model, which determines the relevance between a document and the query by lexical matching, has advantages in retrieval efficiency thanks to inverted indexing. However, the lexical matching fails to retrieve the semantically relevant documents that do not contain any query terms explicitly. On the other hand, the embedding-based model, which determines the relevance between a document and the query by computing the similarity between their distributed representations, is capable of finding semantically relevant results. The downside of the

[1]https://aws.amazon.com/comprehend/
[2]http://www.trec-cds.org/2021.html

embedding-based model is the lack of efficient indexing mechanisms for querying a large collection of documents.

In this work, we propose an approach to biomedical information retrieval by integrating the term-based and the embedding-based retrieval models with a re-ranking strategy. In our hybrid framework, all the documents will be indexed by a search engine based on ElasticSearch,[3] which is a toolkit for building efficient term-based search engines. For the given query, a set of candidate results are retrieved from the search engine. The ranking is initially determined by the term-based ranking score, and further refined by considering the term relationships labeled by the Amazon Comprehend service[4]. From the smaller set of candidate results, we further apply the embedding-based model for re-ranking. We represent the document and the query with bioBERT [2] and compute the cosine similarity between a pair of the document embedding and the query embedding as their relevance score. The final score is a linear combination of the term-based and the embedding-based scores. We conduct the evaluation on the dataset used for the 2016 Clinical Decision Support Track. Experimental results show that our hybrid re-ranking method improves both Precision@k and R-precision scores on two test sizes. The performances of our model in the formal run will be presented when the evaluation of the 2021 Clinical Trials Track releases.

2 Dataset

We choose the dataset previously used for the 2016 Clinical Decision Support Track [5] as our dataset for building our hybrid information retrieval framework. The topics provided between 2017 and 2020 only have three specific attributes, but the topics provided in 2016 and 2021 are both free contexts. There are three tags in 2016's topics, which are note, description, and summary. We choose the contexts in the description tag as our queries because the text length, narrative style, and readability of the such contexts are similar to the topics in the 2021 task. Furthermore, there are 1.25 million documents in the 2016 CDS Task, providing sufficient data to fine-tune and evaluate our approach.

3 Methodology

Our idea is to leverage the efficiency of the term-based model and the semantic, fuzzy matching supported by the embedding-based model. The process of retrieval in our framework is two-staged. In the first stage, an efficient term-based model retrieves the candidate results from the full document collection. Query expansion is applied to include the potentially relevant documents that do not contain query terms. The candidate results are then further re-ranking by using an embedding-based model that is fine-tuned to represent the biomedical documents in contextual embeddings. We detail the term-based model and the embedding-based model in Section 3.1 and Section 3.2, respectively, and describe our final model in Section 3.3.

3.1 Term-based Model

The term-based model is constructed by using the ElasticSearch toolkit. We adopt the Okapi BM25[6] similarity module for relevant scoring. For every document, we build the indexes for the title, the abstract, and the PMID tags. In the beginning, we index the title and the abstract in different indexing attributes. Due to inconsistent document format (e.g. some contents are empty), we finally decide to concatenate the title and the abstract as a single document. Furthermore, we use the contents in brief_title, officl_title, and text block tag as our document content and nct_id as our document identifier for creating the index of the documents. The other settings remain identical to those in 2016. To improve the performance, we also consider the additional information automatically labeled by using Amazon Comprehend Medical service. Amazon Comprehend Medical service has the ability to label the biomedical terms in the text and give their confidence scores. We extract the terms labeled by Amazon Comprehend Medical service as our queries and use the confidence score of each term multiply the correspond retrieval score as our final scores.

[3] https://www.elastic.co/elasticsearch/
[4] https://aws.amazon.com/comprehend/
[5] http://www.trec-cds.org/2016.html
[6] https://www.elastic.co/guide/en/elasticsearch/reference/current/index-modules-similarity.html

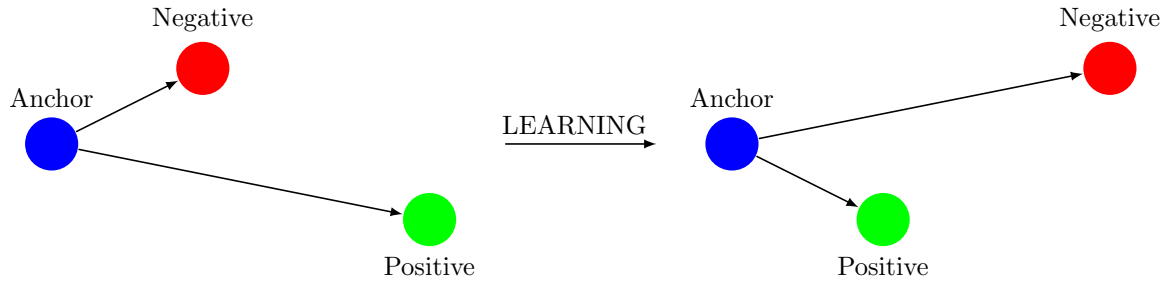

Figure 1: The **Triplet Loss** has the ability to make the distance between Negative Sample and Anchor Sample farther, and make the distance between Positive Sample and Anchor Sample closer.

A part of relevant documents may be ignored by the term-based model if they do not contain any query terms. To extend the search range, one important step in our framework is query expansion. We extend the query terms by consulting biomedical online dictionaries, such as WordNet,[7] RongYang Dictionary,[8] and mesh.[9] These biomedical online dictionaries provide the synonyms of biomedical terms. In order to include the words in the document into the scope of synonym expansion, we also use word embedding to find the synonyms. The last hidden layer of the pretrained bioBERT model is subject to be the representation of the word. When searching for synonyms, we will calculate the cosine similarity between words to assess similarity.

3.2 Embedding-based Model

We use document embedding to consider the relationship between clinical trials and topics. We consider the last hidden layer of the `[CLS]` token in the bioBERT model as the document embedding. The length of a document may exceed the maximum input length of the bioBERT model. Thus, we split the whole document into several paragraphs and represent them separately. We keep a small overlap of text between each adjacent paragraphs. For two paragraphs with similar meanings, the cosine similarity between the two paragraph embedding would be close to 1. When we judge whether document A is semantically similar to paragraph B, we will calculate the cosine similarity between each paragraph in document A and paragraph B, then sum and calculate the average. If the average value is close to 1, the semantics of document A and paragraph B are similar, otherwise the semantics are not similar. Thus, if we want to find the top-k related documents for a paragraph, then we can sort the calculated cosine similarity average and take the top k document as the search results.

We use bioBERT as our pretrained model [3]. There are 30 topics in 2016 Clinical Decision Support Track(CDS). For each topic, we use the relevant documents released by the 2016 CDS and the most irrelevant documents retrieved from our term-based model as the fine-tune dataset. We generate tight negative samples by extracting the documents that are ranked the highest by our term-based model but not listed in the ground-truth. As shown in Figure 1, the fine-tuning will strengthen the similarity between relevant document/query pairs and weaken the similarity between irrelevant document/query pairs even though they are lexically similar.

In our fine-tune phase, we used Triplet Loss as our loss function [4]. Formally speaking, we use the following formula as the loss function. The variable a, p, n appearing in the following formula respectively represent anchor (topic), relevant document, and irrelevant document. The `margin` variable represents the upper bound of distance, which means that when the difference between $\texttt{cosine similarity}(a, n)$ and $\texttt{cosine similarity}(p, n)$ is greater than or equal to the margin, the loss value is 0. The reason we need this is because we only need to focus on those samples that are not very different. When the differences between the two are identified, we will deal with other samples with small differences. As for the reason why we use $\texttt{cosine similarity}(a, n) - \texttt{cosine similarity}(a, p)$ as the loss function, we hope that the model can learn the cosine similarity between topic and relevant document should move as close as possible to 1 and the cosine similarity between topic and irrelevant

[7] https://wordnet.princeton.edu/
[8] http://libs2.vghtpe.gov.tw/digital_new/portal_d1.php?button_num=d1
[9] https://www.ncbi.nlm.nih.gov/mesh/

document should stay as far away as possible to -1.

$$L(a, p, n) = \max\{\texttt{cosine similarity}(a, n) - \texttt{cosine similarity}(a, p) + \texttt{margin}, 0\}$$

3.3 The Final Model

After we fine-tune our embedding-based model, we compute the average of the cosine similarity between each paragraph in the document and the topic. For this purpose, we normalize the paragraph embedding obtained from the bioBERT encoder and make them into unit vectors. So far, we can compute the average of the cosine similarity between each paragraph in a document and a topic easily, which equal to the cosine similarity between the topic embedding vector and the mean vector of paragraph embedding in the document. In order to give the score for each document, we use the averaged cosine similarity between each paragraph and topic as the score. If the embedding-based model is used alone, however, the performance is only slightly better than the term-based model. Thus, we combine these the two scores obtained by both the term-based and the embedding-based models by calculating their linear combination as follows.

$$s = (1 - \alpha) \cdot s_e + \alpha \cdot s_t \tag{1}$$

where s_e is the relevant score obtained by our embedding-based model, and s_t is the relevant score obtained by our term-based model. The value of α can be explored by grid search.

4 Experimental Results

This section shows the performance of our model evaluated on the 2016 Clinical Decision Support Track and 2021 TREC Clinical Trials Track dataset. In the 2016 CDS dataset, there are two version relevant file sets. The average number of their relevant document is about 1,000 and 3,600, respectively. In Table 1, Table 2 and Table 3, we show the performances of our hybrid model, compared with the term-based model as the baseline. In the 2016 dataset experiments, we set the value of α as 0.6. As for the 2021 evaluation results, the model with 0.5 alpha value achieved the best performance. Experimental results show that our hybrid model outperforms the term-based model, confirming that the information implied by the embeddings can be used to distinguish the semantic difference and refine the ranking.

Metric	Term-based Model	Hybrid Model
P@10	0.909	0.940
P@20	0.871	0.890
P@1000	0.285	0.302
R-prec	0.259	0.2731

Table 1: Experimental results on the 1,000 relevant document in 2016 dataset.

Metric	Term-based Model	Hybrid Model
P@10	0.990	0.996
P@20	0.983	0.993
P@1000	0.582	0.614
R-prec	0.339	0.350

Table 2: Experimental results on the 3,600 relevant document in 2016 dataset.

In this work, the negative samples for fine-tuning our embedding model are collected by using the term-based model. That is, they are the documents with a high s_t but labeled as irrelevant in the ground-truth. A potential direction to improve our hybrid model could be made by introducing another kind of negative samples, which can be obtained from the documents with a high initial embedding score but labeled as irrelevant in the ground-truth. Regarding this point, we hope this direction can be explored in the future.

Metric	Term-based Model	Hybrid Model
P@10	0.2507	0.2760
NDCG@5	0.4679	0.4899
NDCG@10	0.4341	0.4665
R-prec	0.1573	0.1624

Table 3: Final results on the 1,000 relevant document in 2021 dataset.

5 Conclusions

This paper shows our attempt to biomedical information retrieval at the TREC 2021 Clinical Traits Track. We propose a hybrid information retrieval framework that enhances the efficiency term-based model with the semantic awareness benefited by the embedding-based model. The embedding-based model has the ability to capture the relationship between document/query pairs in a different perspective, helping to distinguish the subtle semantic difference. Our hybrid approach has been shown effective on the 2016 datasets and can be easily applied to other information retrieval tasks.

References

[1] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. Overview of the trec 2016 clinical decision support track. In *TREC*, 2016.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019.

[3] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019.

[4] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.