SCUT-CCNL at TREC 2019 Precision Medicine Track

Xiaofeng Liu[1], Lu Li[1], Zuoxi Yang[1], Shoubin Dong[1]

[1]Communication and Computer Network Key Laboratory of Guangdong,
School of Computer Science and Engineering,
South China University of Technology, Guangzhou, China

Xiaofeng Liu: cscellphone@mail.scut.edu.cn
Lu Li: antoine.pzc@gmail.com
Zuoxi Yang: yzxstruggle@163.com
Shoubin Dong: sbdong@scut.edu.cn

Abstract

This paper describes a retrieval system developed for the TREC 2019 Precision Medicine Track. For two tasks of Scientific Abstracts and Clinical Trials, we applied the same structure, including the retrieval model, the query expansion and the re-ranking model, to generate the final retrieval results. The experiment results show that the re-ranking model based on SciBERT is of great benefit for retrieval tasks.

Keywords: Precision Medicine, Re-ranking model

1 Introduction

There are two tasks, the Scientific Abstracts and the Clinical Trials, released by the TREC 2019 Precision Medicine Track. For the task of Scientific Abstracts, it focuses on extracting the abstracts of biomedical articles from PubMed Central and providing patients with related treatments, prevention and prognosis. For the task of Clinical Trials, it addresses experiment treatments from ClinicalTrials.gov website for patients if the existing treatments are invalid. In addition, unlike the TREC 2017 and 2018 Precision Medicine Track [1-2], this year adds a sub-task to the scientific abstracts task to find out the particular treatment recommendation for biomedical articles. This sub-task is optional and we have not done it. The 2019 track provides 50 topics related to the patient's disease, genetic variants and demographics.

In this paper, we first indexed the document set and process the query, and then used the BM25 model to extract the relevant document based on the given patient information to form a candidate document set. Finally, we applied SciBERT to re-rank the candidate document set to get the final results.

The rest of the paper is organized as follows. Section 2 gives a detailed description of our approach. Section 3 is the result of our approach on two tasks. In Section 4, we make a conclusion about our work.

2 Methodology

2.1 Collection Indexing

The collections of scientific abstracts and clinical trials are respectively indexed by Apache Luence [3]. For scientific abstract, the indexed fields are "PMID", "ArticleTitle", "AbstractText", "MeshHeadingList", "ChemicalList", while for clinical trials, the indexed fields are "NCT ID", "Title", "Brief Title", "Brief Summary", "Detailed Description", "Study Type", "Intervention Type", "Inclusion Criteria", "Exclusion Criteria", "Healthy Volunteers", "Keywords" and "MeSH Terms". Before indexed scientific abstracts and clinical trials, we applied standard processing including tokenization, stemming and stop word removal to these indexed fields.

2.2 Query Expansion

Referring to [3], we use the same tools to expand disease fields and gene fields. The diseases are enriched into preferred term and synonyms by Lexigram[1] a proprietary API based on the Systematic Nomenclature of Medicine-Clinical Terms (SNOMED CT), the Medical Subject Headings (MeSH) and the International Classification of Diseases (ICD). And the genes are expanded with its synonyms and hyponym by the National Center for Biotechnology Information (NCBI) gene list[2]. The query expansions are used by re-ranking model. Table 1 and Table 2 shows an example of the expansion of disease and gene variant.

Table 1: An example of disease expansion

Disease	dilated cardiomyopathy
Preferred term	primary dilated cardiomyopathy
Synonyms	congestive cardiomyopathy, cocm, ccm, dilated cardiomyopathy, congestive dilated cardiomyopathy, dcm

Table 2: An example of gene expansion

Gene	LMNA
Synonyms	CDCD1, CDDC, CMD1A, CMT2B1, EMD2, FPL, FPLD, FPLD2, HGPS, IDC, LDP1, LFP, LGMD1B, LMN1, LMNC, LMNL1, MADA, PRO1
Hyponyms	lamin, 70 kDa lamin, epididymis secretory sperm binding protein, lamin A/C-like 1, mandibuloacral dysplasia type A, prelamin-A/C, renal carcinoma antigen NY-REN-32

2.3 Document Retrieval

In this section, we will introduce our retrieval model applied for the tasks.

2.3.1 Query Generation

We employed the topics of disease and gene variant to generate a disjunctive Boolean

[1] https://www.lexigram.io

[2] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE INFO/Mammalia/Homo_sapiens.gene_info.gz

query and a conjunctive Boolean query, respectively noted as $Q_d \cup Q_g$ and $Q_d \cap Q_g$, where disease is denoted as Q_d and gene variant is denoted as Q_g.

2.3.2 Retrieval Model

To retrieve relevant clinical trials and scientific abstracts, we used BM25 [4], a language model with Jelinek-Mercer smoothing and one with Dirichlet smoothing, with two Boolean queries. In each query, we reserved the top 20,000 results for each topic, noted as R_0 and R_1.

Finally, we combined R_0 and R_1 with corresponding weights, which are set by grid searching on TERC 2017 and 2018 Precision Medicine Track according to the highest recall. We reserved the top 2000 results for each topic as candidate documents.

2.4 Re-ranking Method

In this section, we will describe our re-ranking method in detail. In TREC 2017 and 2018 Precision Medicine Track, a list of related documents is provided for each topic. Some of the related documents are marked as 2, which is definitely relevant to "Precision Medicine" and some of them are marked as 1, which is partially relevant to "Precision Medicine". The rest of them are marked as 0, which is not relevant to "Precision Medicine".

With the annotation information, we considered document ranking as a two-category problem-given a document, by judging whether it is related to PM. Documents, definitely relevant and partially relevant to given topics, are considered as positive examples. The irrelevant documents are considered as negative examples. The 2017 Track has qrels on 30 topics, and the 2018 Track has qrels on 50 topics. The data set used for model training consists of a list of related documents about 80 topics. According to the ratio of 9 to 1, we randomly selected 72 topic related documents as the training set and 8 topic related documents as the verification set. The candidate document set is a test set described in Section 2.3.2.

For this classification problem, we selected SciBERT [5] as the re-ranking model. SciBERT is a self-attention model pre-trained on a large-scale biomedical literature corpus and is one of the best models for classification problems. SciBERT is rarely used to document retrieval, however its own self-contained structure (Transformer) [6] can learn the complex semantic information from long articles, which can improve the performance of the document retrieval. In addition, by pre-training on biomedical data sets, the model can learn plenty of additional biomedical knowledge.

The SciBERT applied to the re-ranking step needs to process two parts of input, one of which is the input of the topic and the other is the input of the candidate retrieval documents. Therefore, for the input of the model, we need to follow the steps described as below.

a) For the processing of the topic, we only use the disease field and the gene variant field of topics for the document re-ranking, and then we expanding these fields. We arrange them in order of "disease, the preferred term of disease, the synonyms of disease, gene variant, the synonyms of gene variant, the hyponym of gene variant" and form the input A.

b) Due to the difference of contents of the documents in the two tasks, they are processed separately. For Scientific abstracts, we select "AbstractTitle" and "AbstractText" as input B in a document. For Clinical Trials, we select "Brief Title" and "Brief Summary" as input C in a document.

Figure 1: The re-ranking model of input

[CLS] “A” [SEP] “B” or “C” [SEP]

c) Finally, according to qrels of TREC 2017 and 2018 Precision Medicine, a query statement, input A, corresponds to a document, input B or input C. We combined input A and input B or input A and input C as shown in Figure 1. Finally, the number of samples generated by each task is shown in Table 3.

Table 3: The number of samples generated by two tasks

	Train	Dev	Test
Scientific Abstracts	37,361	4,025	80,000
Clinical Trials	24,810	2,397	80,000

After the re-ranking model is trained on the training set, the candidate documents are classified and got a score, and then the top 1000 documents for each topic are reserved as final result.

3 Results

3.1 Run description

For Scientific Abstracts, our team submitted five runs. For the first four runs, we used the retrieval model and the re-ranking model, and the last run we only used the retrieval model. The runs of the first four times are also slightly different. Some of them only extended the synonym or has no extension when expanding the topic. There are also some results in which 4,000 candidate documents are built for each topic. The specific differences are shown in Table 4:

Table 4: The differences of five runs for Scientific Abstracts

	Re-ranking model	Topic Expansion	The number of re-ranking document
ccnl_sa1	Yes	All expansion	80,000
ccnl_sa2	Yes	All expansion	160,000
ccnl_sa3	Yes	Only Synonyms	80,000
ccnl_sa4	Yes	No expansion	80,000
ccnl_sa5	No	No expansion	80,000

For the Clinical Trials, our team submitted two runs. Both runs used a retrieval model, a re-ranking model, and the same topic extension. The difference is in the number of final re-ranked documents, shown in Table 5.

Table 5: The differences of two runs for Clinical Trials

	The number of re-ranking document
ccnl_trials1	80,000
ccnl_trials2	160,000

3.2 Evaluation Results

The evaluation of Scientific Abstracts and Clinical Trials are given in Table 6 and Table 7, respectively. The best results are all marked in bold. In the Scientific Abstracts task, the run "ccnl_sa1" is optimal for infNDCG. The run "ccnl_sa2" is optimal for P@10 and R-prec. In the Clinical Trials task, the run ccnl_trials1 is optimal for infNDCG and P@10, and the run "ccnl_trials2" is optimal for R-prec.

Table 6: The evaluation of five runs for Scientific Abstracts

	infNDCG	P@10	R-prec
ccnl_sa1	**0.5309**	0.6450	0.3032
ccnl_sa2	0.5222	**0.6500**	**0.3066**
ccnl_sa3	0.4569	0.5475	0.2858
ccnl_sa4	0.4571	0.5775	0.2886
ccnl_sa5	0.4463	0.5025	0.2770

Table 7: The evaluation of two runs for Clinical Trials

	infNDCG	P@10	R-prec
ccnl_trials1	**0.4862**	**0.5947**	0.3414
ccnl_trials2	0.4845	0.5789	**0.3440**

Conclusions

In this paper, we described the method used on the TREC 2019 Precision Medicine. We performed standard processing including tokenization, stemming and stop word removal and indexing on the document set of the two tasks. Then we used BM25 to retrieve the document sets to generate candidate document sets, and finally we used SciBERT to re-rank the candidate document sets. The experimental results show that SciBERT can be used to improve the optimal ranking of the search results. In addition, the expansion of diseases and genetic variants has an important role in ranking.

References

1. Roberts, K., Demner-Fushman, D., et al. Overview of the TREC 2017 Precision Medicine Track. In Proceedings of the Twenty-Sixth Text Retrieval Conference.
2. Roberts, K., Demner-Fushman, D., et al. Overview of the TREC 2018 Precision Medicine Track. In Proceedings of the Twenty-Seventh Text Retrieval Conference.
3. M. Oleynik, E. Faessler, A. et al. HPI-DHC at TREC 2018: Precision Medicine Track. In Proc. of the Twenty-Seventh Text REtrieval Conference, TREC 2018.

4. Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. Nist Special Publication Sp, 109:109, 1995.
5. Beltagy, I, Cohan, A, Lo, K. SciBert: A Pretrained Language Model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
6. Tang GB, Mathias M, Annette R, et al. Why self-attention? a targeted evaluation of neural machine translation architectures. In Proc. of EMNLP, 2018.