DUTIR at TREC 2019: Precision Medicine Track

Jingkun Feng, Zhihao Yang, Zhiqiang Liu, Ling Luo, Hongfei Lin, Jian Wang College of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, China

Abstract: This paper describes the system developed for the TREC 2019 Precision Medicine Track by the Team DUTIR from Dalian University of Technology. In the system, we applied a hybrid method to score the related documents for each topic. First, we used Elasticsearch, an open-source Lucene-based full-text search engine, to obtain the initial retrieval results. Then we trained several deep models using TREC 2017 PM data. Finally, we applied the pre-trained models to reorder the initial search results. The performance of our system is above the median for the scientific abstracts subtask and below median for the clinical trials subtask.

INTRODUCTION

The 2019 Text Retrieval Conference (TREC) Precision Medicine (PM) track continues the prior 2017[1] and 2018[2] Precision Medicine tracks, which aims to retrieve biomedical articles and clinical trials which can provide useful and precision medicine-related information for treating cancer patients. Given 40 cancer patient topics in this track, participants of the track will be challenged with retrieving (1) scientific abstracts, in the form of biomedical article abstracts, addressing relevant treatments for the given patient, and (2) clinical trials, addressing relevant clinical trials for which the patient is eligible. Same as the 2017 and 2018 PM tracks, each

topic includes information on the patient's disease (type of cancer), the relevant genetic variants (which genes), and some basic demographic information (age, sex). Table 1 shows a topic example. In addition, the biomedical articles are largely from MEDILINE/PubMed, while the clinical trials are a snapshot of the ClinicalTrials.gov. The specific descriptions of the PM track are available online¹.

Table 1. An example of cancer patient topics

<topic number="1">
<disease>Acute lymphoblastic leukemia</disease>
<gene>ABL1, PTPN11</gene>
<demographic>12-year-old male</demographic>
</topic>

In this report, we present our system for the two subtasks. In scientific abstracts task, we use a hybrid method to retrieve related articles from the collections. We first indexed all documents with Elasticsearch², and then we applied query templates and query expansion to get the initial search results for each topic. Finally, we used deep models which were trained by using the 2017 PM track data to reorder documents of the initial retrieval results. Different from the scientific abstracts task, we only used Elasticsearch with the same query templates and query expansion to obtain retrieval results for clinical trials track. In next section our method used in

_

¹ http://www.trec-cds.org/2019.html

² https://www.elastic.co/cn/products/elasticsearch

scientific abstracts task is introduced in detail.

METHODS

This section describes our IR system designed for the scientific abstracts retrieval subtask of the PM track. Figure 1 shows the architecture of our system. As shown in the figure 1, our system consists of two parts, initial relevant documents retrieval and retrieved documents re-ranking.

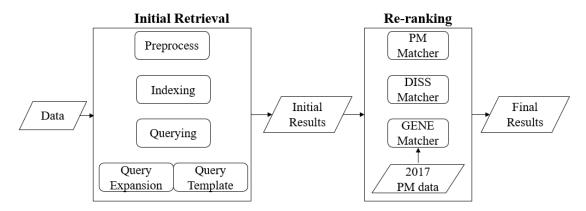


Figure 1. The overview of system for scientific abstracts task

For retrieving the initial relevant documents, we relied on Elasticsearch, an open source Lucene-based full-text search engine, to index and query the collection of scientific abstracts contemplated in the biomedical articles task. In particular, we preprocess the collection of documents before building the index. Then, we utilized query expansion strategy for expanding terms that will be exploited in Elasticsearch queries to refer to diseases and genes. Moreover, we wrote some query templates using structured query language provided by Elasticsearch and combined their search results in different weighting ways.

For reordering the retrieved documents from the initial results, we used deep learning models to reorder the initial retrieval results in order to further improve the retrieval performance. Specifically, we applied the text classification model to determine whether documents are therapeutically relevant. In addition, we used deep matching models to score documents from disease and gene dimensions.

In the rest of this section, the main steps of our scientific abstracts retrieval system are introduced in detail.

Preprocessing & Indexing

To make retrieval more efficient, we preprocess the original documents before indexing. Specifically, we removed fields in each document that were not relevant for retrieval. For each XML document, we took into account the following fields:

- 1) PMID: the ID of the article;
- 2) Title: the title of the article;
- 3) Abstract: the abstract of the article;
- 4) Chemical List: the chemicals that appears in the article;
- 5) Mesh Heading List: Mesh Concepts that article contains.

For the text content in the above fields, we performed a series of processing such as tokenization, stop word removal and stemming. Finally, we use Elasticsearch's default options for indexing.

Query Expansion & Query Templates

The purpose of query expansion is to support the formulation of queries with a higher recall by gathering alternative expressions useful to the terms mentioned in queries. We paid more attention to the fields of disease and gene of the topic to conduct query expansion.

For the disease field, we expanded disease with its synonyms and hyponyms via a disease dictionary. While for the gene field, we expanded gene with its synonyms via a gene dictionary. In particular, the two dictionaries mentioned above were built from the disease and gene vocabularies provided by the CTD base [3].

To improve the quality of the initial retrieval results, we wrote a variety of query templates with different focuses. For the same topic, we used all the query templates in turn to obtain the retrieval results, and finally summarized all the results of the same topic with different weights.

Re-Ranking

The objective of re-ranking is to improve the retrieval performance by adjusting the position of some documents in the initial retrieval result.

Due to the requirement that retrieved documents should focus on precision medicine, we need to consider whether a document is therapeutically relevant. Specifically, we integrated some classic text classification models such as TextCNN [4] and TextRCNN [5] to determine whether a retrieved document focuses on precision medicine. The training data is about 20000 biomedical articles from TREC 2017 PM Track which were annotated with

"PM" for "Precision Medicine" and "Not PM" for "Not Precision Medicine". For each retrieved document, we used pre-trained model named PM Matcher to predict whether it is a "PM" document. If a document is "PM", we boost its initial score.

To assess the relevance of the document from both disease and gene dimensions, we also utilized deep matching models such as DRMM [6] and PACRR [7] to score retrieved documents. To train these models, we used the same training corpus as above which had different labels. In particular, we used pre-trained model named DISS Matcher to score documents from disease dimension. Meanwhile, another pre-trained model named GENE Matcher was applied to score documents from gene dimension.

To obtain the final ranking results, we combined above three scores with different weights.

RESULTS

We submitted five runs for scientific abstracts subtask and two runs for clinical trials subtask. For both tasks, metrics used are inferred non discounted cumulative gain (infNDCG), precision at rank 10 (P@10) and R-precision (R-Prec).

For scientific abstracts subtask, we list the average performance of each of our runs and the official median result in table 2. As can be seen from table 2, our results are better than the official median result, which proves the effectiveness of our method.

Table 2. Results for Scientific Abstracts Subtask

	InfNDCG	P10	R-Prec
DutirRun1	0.4980	0.5800	0.3266
DutirRun2	0.5103	0.5975	0.3273
DutirRun3	0.5053	0.5825	0.3227
DutirRun4	0.5035	0.5825	0.3257
DutirRun5	0.5108	0.5750	0.3250
Median	0.4559	0.5450	0.2806

In general, the five runs for retrieving scientific abstracts are described in detail as follows:

- 1) DutirRun1: we took this run as a baseline. In this run, we only applied Elasticsearch with query templates and query expansion to obtain retrieval results.
- 2) DutirRun2: in order to explore whether the PM Matcher works, we improved documents' scores which were judged as "PM" by PM Matcher. The result of this run proves the effectiveness of PM Matcher.
- 3) DutirRun3: in this run, we only utilized DISS Matcher to re-ranking the initial results. Compared with baseline, the result of this run is slightly improved.
- 4) DutirRun4: GENE Matcher was applied in this run. There was also a

- slight improvement during this run.
- 5) DutirRun5: we combined PM Matcher, DISS Matcher and GENE Matcher to re-ranking the initial results in this run. As we can see from the results, the NDCG value increased slightly, but the other two indicators did decrease.

For clinical trials subtask, we also show our results and the median result in table 3.

Table 3. Results for Clinical Trials Subtask

	InfNDCG	P10	R-Prec
Dutir_Cli1	0.5022	0.4579	0.3453
Dutir_Cli2	0.5038	0.4553	0.3420
Median	0.5136	0.4657	0.3477

Our results for clinical trials subtask are not satisfactory, so we can interface that retrieving relevant documents only with Elasticsearch was not a good method.

CONCLUSIONS

This paper describes the IR system developed for the TREC 2019 Precision Medicine track. In our system, we adopt a hybrid method to retrieve biomedical articles. For scientific abstracts task, we used Elasticsearch with query templates and query expansion to obtain initial results, then we utilize deep models to re-rank the retrieved documents. Experimental

by three models we trained. For clinical trials task, we only used Elasticsearch to search relevant articles. Experimental results show that only using Elasticsearch is not a good method.

ACKNOWLEGEMENTS

This work was supported by the grant from the National Key Research and Development Program of China (No. 2016YFC0901902).

REFERENCES

- [1] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 precision medicine track. In TREC. Gaithersburg, MD.
- [2] Kirk Roberts. 2018. Overview of the TREC 2018 precision medicine track. In TREC. Gaithersburg, MD.
- [3] Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 2018 Sep 24.
- [4] Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820 (2015)

- [5] Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: AAAI. vol. 333, pp. 2267–2273 (2015) [6] Guo J, Fan Y, Ai Q, et al. A deep relevance matching model for ad-hoc retrieval[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016: 55-64.
- [7] Hui K, Yates A, Berberich K, et al. PACRR: A Position-Aware Neural IR Model for Relevance Matching[J]. 2017.