Exploring post retrieval techniques for bio-medical domain retrieval

Yue Wang and Hui Fang

Department of Electrical and Computer Engineering
University of Delaware
140 Evans Hall, Newark, Delaware, 19716, USA
{wangyue,hfang}@udel.edu

Abstract. Retrieval is a common way to access the huge data collection in bio-medical domain. For instance, the physicians would need to retrieve relevant documents to treat cancer for the patients. With the huge number of documents in the collection, a reliable retrieval system is critical for a satisfying performance. This year's TREC Precision Medicine track continues the same procedure as last year by providing us the platform for testing different methods for bio-medical retrieval. For this year, we used similar methods as we used in TREC PM17, i.e., term based representation and concept based representation. In addition, we also tried to combine the different representation with results filtering and two-round retrieval. The results show that the modification on the methods, i.e., results filtering and two round retrieval, did not outperform the baseline methods.

1 Introduction

Bio-medical domain is a fast growing domain. It has been reported that there are 75 clinical trails and 11 systematic reviews being published per day [1]. It is clear that a effective way of accessing these information is in urgent need. Existing studies tried to solve this domain specific retrieval in two directions based on how the documents are represented, i.e., term based representation and concept based representation. As the name suggests, the term based representation treat each document as a "bag of terms", while the concept based representation consider the documents are composed by concepts identified by the NLP toolkits. Previous work show that the concept based representation would outperform the term based representation [2, 3]. However, they mostly focus on verbose queries, the performance of concept based representation on key term queries need to be further explored.

As a continues year of precision medicine track in TREC, PM18 follows the similar format as PM17, i.e., finding relevant documents based on the key terms queries submitted by physicians. We proposed several new methods together with some existing methods to test the performance with this year's query set. Specifically, we first tested the baseline form of term based and concept based representation separately. We then expand the original query using two well organized knowledge system for query expansion. In addition, we also tried result filtering and two-round retrieval methods as post-retrieval modification. The results show that query expansion using external resources would improve the performance, although the improvement is marginal. The post-retrieval modifications failed to further improve the performance.

2 Method

As a continuous task of last year's PM track, the goal of this year's track remains retrieving relevant information to treat cancers based on patients' clinical situations. The name of the cancer, together with affected genes, and demographic information are provided as queries. The "other" field used in last year, which contains additional information for the patient, is removed from this year's query. An example of this year's query is shown as in Figure 1.

```
<topic number="1">
  <disease>melanoma</disease>
  <gene>BRAF (V600E)</gene>
  <demographic>64-year-old male</demographic>
</topic>
```

Fig. 1. An example query of TREC PM track 2018.

2.1 Term base representation

One intuitive method is to finish the retrieval task in term based representation with the different query fields merged as one combined query. We applied this method as a baseline method in our experiment. In addition to this method, we also explored two online resources namely GeneCards[1] and Disease Ontology[2], for query expansion. Specifically, we extracted the expansion key terms as follows. GeneCards is a database about human genes. For each gene entry in the GeneCards database, the gene is further explained by several fields such like aliases, disorders, summary, etc. We first extracted the gene from the original query and submit them to the GeneCards. We then selected the alias and summary for each gene entry to form the pool for candidate expansion terms. The alias are directly used as the expansion term, while the top 10 terms from summary are selected as expansion term.

Disease Ontology has been developed as a standardized knowledge base for human disease. For the disease name mentioned in the query, we first curled the page of the target page on disease ontology. We then extract the names and synonyms of parent and children of the target disease as the candidate expansion term. We repeat the same procedure to acquire the names and synonyms of the grand-parent and grand-children of the target disease. These mentioned names are then used as the expansion terms for the original query.

2.2 Concept based representation

Different from term based representation, the documents and queries are treated as "bag of concepts". This is done by utilizing NLP tools to extract important medical concepts from the documents and queries. We followed the procedure described in previous work [4] to build the document collection. To be specific, we first retrieved top 5,000 documents for each query using term based representation. We then converted these documents as concept based representation using MataMap. Query expansion is also applied for concept based representation. We converted the expansion terms selected from Disease Ontology as concepts and append them as expansion terms for concept based representation.

2.3 Two-round retrieval

Since the gene field provides more detailed information for the disease, one intuitive way is to conduct a two-round retrieval. In the first round, as much as possible relevant documents are expected to be retrieved to form a candidate pool. Then the gene field is used as a filter to select the documents which are truly relevant to the query. Specifically, we first retrieved 5,000 documents using the disease name only, then these documents are filtered using the gene name mentioned in the query.

2.4 Result filtering

Demographic information is an important field to be considered for the clinical trail task. We included this information as a filtering step in the retrieval task. Firstly, top 5,000 documents are retrieved for each query. The demographic field of these documents are extracted and compared with the ones specified in the query. Only the ones satisfy the query would be kept.

[1] http://www.genecards.org/
[2] http://disease-ontology.org/

3 Experiment

3.1 Data set pre-processing and index building

The two data sets are collected from the track home page and pre-processed as follows.

Scientific abstracts The PubMed abstracts and the AACR/ASCO proceedings are merged as one data set. This data set is cleaned by only keep the content between the xml tags. The tags themselves are removed. We built the index with Indri. The stopword are not removed and no stemming is applied. This forms the term based index. We then retrieved top 5K results using the disease and gene for each query. The unique documents are kept and converted to concept based documents. No stopword removal and no stemming applied when we create the concept based index.

Clinical Trails We built a parser to extracts the following fields from the NCT data set: *brief title*, *acronym*, *official title*, *brief summary*, *detailed description*, *keyword*, *condition*, *intervention*, *condition browse*, *intervention browse*, *primary outcome*, *secondary outcome*, *other outcome*, *arm group*, *gender*, *min age*, *max age*, *inclusion*, and *exclusion*. Then we created the index with every field listed as a separated field using Indri. The whole collection is converted to concept based representation using MetaMap with the meta field information. We then created the index for concept based representation using indri as well.

3.2 Submitted runs

We submitted 5 runs for each track. We applied same basic retrieval function for all runs, with different expansion and weighting techniques introduced as follow:

Scientific Abstracts The details of the 5 submitted scientific abstracts runs are:

UDelInfoPMSA1: A term based run. We used the query as it is, with both disease field and gene field included as the query. This is served as a baseline run.

UDelInfoPMSA2: A term based run. In addition to the disease and gene field, we also included expansion terms selected from disease ontology and GeneCards.

UDelInfoPMSA3: A concept based run. We used query as it is, with both disease field and gene field included.

UDelInfoPMSA4: A concept based run. We expand the original query using the terms selected from disease ontology.

UDelInfoPMSA5: A term based run. We conducted the two round retrieval as described in section 2.3.

Clinical Trails The details of the 5 submitted clinical trails runs are:

UDelInfoPMCT1: A term based run. We used the query as it is, with both disease field and gene field included as the query. This is served as a baseline run.

UDelInfoPMCT2: A term based run. In addition to UDelInfoPMCT1, we also applied result filtering as described in section 2.4.

UDelInfoPMCT3: A term based run. In addition to the disease and gene field, we also included expansion terms selected from disease ontology and GeneCards.

UDelInfoPMCT4: A concept based run. We used query as it is, with both disease field and gene field included. We then filter the results using the demographic information as described in section 2.4

UDelInfoPMCT5: A term based run. We conducted the two round retrieval as described in section 2.3.

Table 1. Performance of submitted Scientific Abstracts runs.

	infNDCG	R-prec	P10
UDelInfoPMSA1	0.5016	**0.3289**	0.5720
UDelInfoPMSA2	**0.5081**	0.3253	**0.5800**
UDelInfoPMSA3	0.2479	0.1622	0.4020
UDelInfoPMSA4	0.2499	0.1583	0.4060
UDelInfoPMSA5	0.4954	0.3181	**0.5800**
TREC-Median	0.4290	0.2672	0.5460

Table 2. Performance of submitted Clinical Trails runs.

	infNDCG	R-prec	P10
UDelInfoPMCT1	**0.5057**	0.3952	0.5120
UDelInfoPMCT2	0.4686	0.3698	0.5120
UDelInfoPMCT3	0.4976	**0.3967**	0.5040
UDelInfoPMCT4	0.3354	0.2551	0.3640
UDelInfoPMCT5	0.4782	0.3810	**0.5240**
TREC-Median	0.4297	0.3267	0.4680

3.3 Experiment results

We first present the results of the performance of the submitted runs as shown in Table 1 and 2.

It is clear from Table 1 that one of the baseline system, UDelInfoPMSA2, performs the best for the scientific abstract track, although the improvement is marginal. It shows that including the expansion terms from the external resources could boost the performance, but the parameters need to be tuned to achieve the best performance. Comparing the two term based baseline runs with the two concept based baseline runs, we could easily tell that the term based methods are much better. This shows the concept based representation does not work for the scientific abstract task. The two rounds retrieval method introduced in 2.3 did not perform as expected neither. Although the precision at 10 achieved the same as the baseline methods, the infNDCG drops about 2%. This indicates that, instead of filtering, other re-ranking methods should be explored to further improve the performance.

The performance of the proposed methods on clinical trail runs are different. By including the expansion terms selected GeneCards and disease ontology, the performance of UDelInfoPMCT3 did not outperform the baseline method (UDelInfoPMCT1) in terms of infNDCG. Result filtering also lowered the performance comparing to the baseline method. Similar as shown in scientific abstract runs, the two round retrieval could achieve a higher performance in terms of precision, but not on infNDCG.

4 Conclusion

We evaluated the performance of term based representation and concept based representation together with query expansion and post-retrieval modifications for this year's PM track. Query expansion using the terms selected from GeneCards and Disease Ontology could improve the performance comparing with the baseline method, although the improvement is marginal. However the post-retrieval modification can not outperform the baseline as expected. Other variations should be tested for post retrieval modifications.

References

1. Bastian, H., Glasziou, P., Chalmers, I.: Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS medicine **7**(9) (2010) e1000326

2. Wang, Y., Liu, X., Fang, H.: A study of concept-based weighting regularization for medical records search. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014. (2014)
3. Wang, Y., Lu, K., Fang, H.: Learning2extract for medical domain retrieval. In: Information Retrieval Technology, Cham, Springer International Publishing (2017) 45–57
4. Wang, Y., Fang, H.: Exploring the query expansion methods for concept based representation. Technical report, DELAWARE UNIV NEWARK DEPT OF ELECTRICAL AND COMPUTER ENGINEERING (2014)