Combining Term-based and Concept-based Representation for Clinical 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. Biomedical domain retrieval has been a trending topic that attracts many IR researchers. Different document representation methods, i.e., term based representation and concept based representation, have been proposed to solve this question. However, previous studies have focused the verbose queries. In this year's Precision Medicine track, we evaluated the performance of these two basic document representation methods on short queries. We also explored possible ways to combine these two methods. The results show that these two representations perform differently on the scientific abstract and clinical trail data sets. Simply merge the results list may not leads to optimal performance, while term based filtering on top of the concept based results could significantly improve the performance.

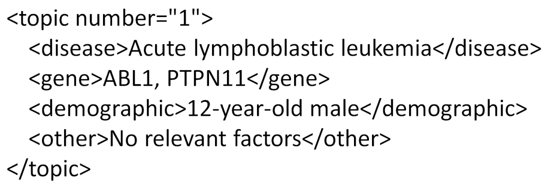
1 Introduction

Since the genetic and environmental situations for each patient could be different, there is no universal solution of a certain disease. Therefore, finding more relevant and accurate information for the physicians based on the patients' situation is a critical task. Existing studies solve the clinical related retrieval task in two directions based on how the documents and queries are represented, namely term based representation and concept based representation. Same as the retrieval task in the other domain, term based representation considers the documents and queries as "bag of terms", and then apply various types of techniques to improve the retrieval performance. The concept based representation, on the other hand, treats the data collections as "bag of concepts". The effectiveness of concept based representation has been studied for clinical domain [1, 2]. However, most of the works focus on improving the performance of the verbose queries using concept based representation, instead of the keyword queries.

This year's TREC provides us a platform to test the performance of keyword queries using different forms of representations. Different from the verbose queries used in previous years' Clinical Decision Support track, it is not clear that how the keyword query would perform with the term based representation or the concept based representation, and a combination of these two. Therefore, we proposed several experiments to evaluate the performances. Specifically, we first test different method of using term based and concept based representation separately. We then proposed different methods to combine the results. The results show that these two document representation methods perform differently on the two provided datasets. The term based representation works better on scientific abstract documents, while the concept based representation outperforms on clinical trail documents. In addition, simply merge the results lists does not lead to a better performance for the two representations.

2 Method

This year's PM track focuses on retrieving relevant information for physicians to treat cancer for patients. The queries are represented as several keywords, including the patients' disease, relevant genetic variants, basic demographic information, and other information. An example of this year's query is shown as in Figure 1.



```
<topic number="1">
   <disease>Acute lymphoblastic leukemia</disease>
   <gene>ABL1, PTPN11</gene>
   <demographic>12-year-old male</demographic>
   <other>No relevant factors</other>
</topic>
```

Fig. 1. An example query of TREC PM track 2017.

2.1 Term base representation

The straight forward method is to merge the information from different fields together and retrieve the document in term based representation using the merged keyword query. However, this would certainly weaken the effectiveness of the information conveyed in the fields. Therefore, we tried to apply different weights on the terms from different query fields.

In addition, we also applied the query expansion with the GeneCards[1]. The GeneCards is a searchable database of human genes. It contains numerous and concise genomic related information. For each gene entry in the GeneCards database, there are several fields, such like aliases, disorders, summary, proteins, expression, etc, further explains this gene in details. We believe it is a reliable and fruitful resource and utilize it to expand our query. Specifically, all genes mentioned in the queries are submitted to GeneCards to find their own page. The alias and summaries are extracted from the database. The alias are directly used as the expansion term, the indri query language #1() is used to make sure the expansion terms are shown as exact form in the document. The top 10 terms (based on the term frequency) from the summaries are also selected. The common stop words and some domain stop words are removed from the top 10 terms.

2.2 Concept based representation

The general idea of concept based representation is utilizing NLP tools extract the important medical concepts from the documents, and used those identified concepts to represents the original documents. Previous works have shown that it could outperform the term based representation when the queries are longer [3]. Therefore, we also want to explore the effectiveness of the concept based representation when the queries are shorter. To be specific, we followed the same method as the previous by first conduct a initial retrieval within term based collection and collected the top 5K documents for each query. We then converted these documents using MetaMap by only keep the identified concepts. Directly perform the retrieval task in the concept based representation index could suffer from the problems of inaccurate mapping. Therefore, we also applied the Unified method and Balanced methods as introduced in [1]: the Unified method replaces the variants of the CUIs that belong to the same aspect using the one with the highest IDF value, while the Balanced method assigns higher weight to the concepts from an important aspect.

2.3 Results combination

In addition to test the performance of the concept based representation on short queries, we also want to explore the possible methods to combine the term based and concept based results. One of the straight forward way is to directly combine the two ranked lists. We applied the normalized combination method as follow: let d_i be the current document, $S^A_{d_i}$ be the score of the document d_i in

[1] http://www.genecards.org/

results list A, S^A_{min} be the score of the last document in the results list A. Then the combined score for d_i from ranked list A and B would be represented as:

$$S_{d_i} = \frac{S^A_{d_i} - S^A_{min}}{S^A_{min}} + \frac{S^B_{d_i} - S^B_{min}}{S^B_{min}} \tag{1}$$

If d_i is not return in one of the ranked list, the corresponding component will be set to 0. The results will then be re-ranked based on the combined normalized score.

2.4 Result filtering

Each document in the clinical trail collection is associated with the age and gender information of the target patients. Therefore, it requires not only the document should match the query, but the demographic information should also satisfy for the clinical trail task. Therefore, we filtered the results based on the demographic information in the term based representation. If the target gender/age is not in the clinical trail, this result will be dropped.

3 Experiment

3.1 Data set pre-processing and index building

We crawled the data collections from the track home page. The documents are pre-processed in the following way.

Scientific abstracts Both data collections, i.e., PubMed abstracts and AACR/ASCO proceedings, are merged as one data set. We parsed this data set to remove everything in the xml tag field, only kept the content of each field. The data is converted to indri document format without changing anything in the content.

We first built a index that with the these two data collection in term based representation. No stopword removal and no stemming applied. We then used three types of queries, i.e., the disease+gene, disease+other, and disease+other+gene to retrieve the top 5K documents for each query. The unique ones are kept and converted to Concept based index. When creating the concept based index, no stopword being removed and no stemming applied.

Clinical Trails The NCT data is downloaded and extracted. We created a parser that extracts the following fields from the raw data: *brief title*, *acronym*, *official title*, *detailed description*, *keyword*, *condition*, *intervention*, *condition browse*, *intervention browse*, *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:

UDelInfoPMSA2: Term based retrieval. The queries are parsed to search using the disease field, gene field, and other field separately. The weight on disease is 0.5, gene is 0.3, and other is 0.2.

UDelInfoPMSA3: Term based retrieval. The terms from the disease field, gene field, and other field are merged. Query expansion is done as described in section 2.1. The weight on original terms is 0.8, and expansion term is 0.2.

UDelInfoPMSA5: Concept based retrieval. The queries are parsed using the same way as UDelInfoPMSA2.

UDelInfoPMSA6: Concept based retrieval. The CUIs are merged from different query field. Unified and Balance method are applied on the identified concepts as described in section 2.2.

UDelInfoPMSA7: Combined run. Combine the result lists of UDelInfoPMSA2 and UDelInfoPMSA5 together using the method introduced in section 2.3.

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

UDelInfoPMCT3: Term based retrieval. Search the disease, gene, and other information in detailed description, inclusion, keyword, condition browse, and intervention browse fields. The results are filtered based on gender and age information as described in section 2.4.

UDelInfoPMCT4: Term based retrieval. Search the disease, gene, and other information in detailed description, inclusion, keyword, condition browse, and intervention browse fields. Query expansion using the alias, and top 10 terms from summaries. The weight on original terms is 0.8, and on the expansion terms are 0.2. The results are filtered based on gender and age information as described in section 2.4.

UDelInfoPMCT6: Concept based retrieval. Search the disease, gene, and other information in detailed description, inclusion, keyword, condition browse, and intervention browse fields.

UDelInfoPMCT8: Filtered run. Search the disease, gene, and other information in concept based representation. and filter the results using demographic data in term based representation as described in section 2.4.

UDelInfoPMCT10: Combined run. Combine the result lists from method UDelInfoPMCT6 and method UDelInfoPMCT3 using the method as described in section 2.4.

3.3 Experiment results

We first present the results of the performance of the submitted runs as shown in Table 1 and 2. Since the judgement of query 12 was not released before the organizer released the overall results, we only

Table 1. Performance of submitted Scientific Abstracts runs (including topic 12).

	infNDCG	**R-prec**	**P10**
UDelInfoPMSA2	**0.3897**	**0.2503**	**0.5067**
UDelInfoPMSA3	0.3328	0.2245	0.4233
UDelInfoPMSA5	0.2631	0.1775	0.3567
UDelInfoPMSA6	0.1760	0.1226	0.2400
UDelInfoPMSA7	0.3530	0.2392	0.4367
TREC-Median	0.2702	0.1704	0.3655

It is clear from Table 1 that one of the baseline system, UDelInfoPMSA2, performs the best for the scientific abstract track. UDelInfoPMSA2 outperforms the UDelInfoPMSA3 indicating that applying the Genecards for query expansion did not perform as well as expected. This could be due to the #1() restriction we applied to the expansion terms. In addition, the term based methods are generally better than the concept based method. Comparison between the performance of UDelInfoPMSA2 and

Table 2. Performance of submitted Clinical Trails runs (including topic 12).

	P5	**P10**	**P15**
UDelInfoPMCT3	0.0966	0.0793	0.0690
UDelInfoPMCT4	0.1034	0.0793	0.0759
UDelInfoPMCT6	0.1310	0.1034	0.1034
UDelInfoPMCT8	**0.3241**	**0.2862**	**0.2506**
UDelInfoPMCT10	0.0690	0.0621	0.0644
TREC-Median	0.2684	0.2448	0.2212

UDelInfoPMSA5 shows that the concept based representation does not work for the scientific abstract, and merging the different query fields together (UDelInfoPMSA6) further hurt the performance. The combined run (UDelInfoPMSA7) is expected to perform the best. However, we believe it is affected by the low performance of the concept based representation.

For the clinical trail track, query expansion does not improve the performance significantly in term based representation. By comparing the the UDelInfoPMCT6 with UDelInfoPMCT3 we could claim that the concept based representation is better than the term based representation for this track. The same as shown in scientific abstract track, the effectiveness of the query expansion using Genecards is not significant. The combined run, i.e., UDelInfoPMCT8, significantly improves the performance of the concept based baseline method, which indicating that the results filtering in term based could successfully find the target medical trail document for the patient.

One clear observation we could make is that the concept based representation and term based representation perform very differently in these two tracks. To the best of our knowledge, we believe there are two reasons: On one hand, the effectiveness of concept based representation is highly associative with the correctness of the mapping. The documents in the scientific abstract, which are all medical domain research paper, are more complicated. Therefore the mapping results are not as good as the clinical trail documents, which are short and concise. One the other hand, concept based representation is more suitable for the clinical trail documents. This is where the authors may use different forms of names when referring the same disease or symptom. The concept based representation could convert those variants to the same form to bridge the vocabulary gap.

4 Conclusion

We tried both term based and concept based representation for this year's Precision Medicine track. The term based representation outperformed the concept based representation on the scientific abstract track, but the concept based representation achieved better performance on clinical trail track. We believe that it is due to the nature of the document collections which leads to the performance differences.

References

1. 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)
2. Wang, Y., Lu, K., Fang, H.: Learning2extract for medical domain retrieval. In: Information Retrieval Technology, Cham, Springer International Publishing (2017) 45–57
3. 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)