Exploring the Query Expansion Methods for Concept Based Representation

Yue Wang and Hui Fang

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

Abstract. The CDS track investigates methods that could help physicians find relevant medical cases for patients they are dealing with. Concept based representation has been shown to be effective in biomedical domain. However, the basic concept based retrieval method may not perform well because of the additional restriction on each clinical cases. Therefore, in this paper, we explored two external resources to perform query expansion for the basic concept based representation method, and discussed the performance of them.

1 Introduction

The Clinical Decision Support track (CDS track) is a new track in 2014 TREC. The goal of this track is to help the physicians by connecting the medical cases with the relevant information for patient care. The task for this year is to retrieve the relevant biomedical articles that could help answer the generic clinical questions about the medical records[1]. This is similar as the Medical Record track in TREC in 2011 and 2012, in the sense that both of these tracks focus on the biomedical domain. However, the differences are also obvious. On one hand, the queries, or topics, in CDS track are much longer, which contains more information about the patients. On the other hand, each topic is associated with a query type, which requires the document to not only mention the terms in the query, but could also be used to answer the questions with the query type. We consider this problem as a retrieval task, where the returned documents should be relevant to the original query, and also following the restriction that specified in the query type.

Concept based representation has been proposed and showed to be more effective than traditional term based representation in biomedical domain [1, 2]. The so-called concept based representation first extracts the concepts from both the documents and queries using existing NLP tools, and then perform retrieval task over those identified concepts. We followed the same direction in the CDS track. In addition, we explored the basic concept based representation method with query expansion. To be specific, we first followed the method proposed in [1] to convert the documents from term based representation to concept based representation. We then utilized the Cases Database and UMLS relations to expand the key concepts in the original query, and the expanded queries were submitted to the retrieval system as the final query. To compare with, we also applied the same query expansion techniques to the term based representation. The results show that using the UMLS relation could help to improve performance.

2 Query expansion in concept based representation

The queries in CDS track is the narrative of the case report for a patient, which may contain the patient's medical history, current symptoms, test results and others. An example of the narrative is shown as in Fig 1. There are two types of narrative provided by the organizers, the description and

[1] http://www.trec-cds.org

the summary. These two narratives are equivalent based on the organizers introduction, while the summary narrative contains less irrelevant information. Therefore, the summary narratives are used in our experiments. However, even with the summary queries, there are some terms that are not relevant to the diseases/symptoms of the patients, for instance, the word "shows" and "on" in the example query. In order to remove the non-relevant terms from the query, we applied MetaMap to identify the key concepts from the summary query. The key concepts are the ones with certain semantic types, such as "Sign or Symptom" and "Diagnostic Procedure".

There are three types of queries, i.e., diagnosis, test, and treatment. Each query is associated with one query type. For a particular query, the returned document would be considered as relevant only if it can help to answer the question related to the query type. For instance, a document mentioned the name of the disease that has the same symptoms shown in Fig 1 would be relevant, but the other document only contains what test should be performed based on these symptoms is not relevant, as the query is asking for the patient's diagnosis.

```
<topic number="3" type="diagnosis">
 - <description>
    A 58-year-old nonsmoker white female with mild exertional dyspnea and occasional cough is found to
    have a left lung mass on chest x-ray. She is otherwise asymptomatic. A neurologic examination is
    unremarkable, but a CT scan of the head shows a solitary mass in the right frontal lobe.
   </description>
 - <summary>
    58-year-old female non-smoker with left lung mass on x-ray. Head CT shows a solitary right frontal lobe
    mass.
   </summary>
</topic>
```

Fig. 1. An example narrative of TREC CDS track 2014.

2.1 Query expansion with Case databases

Cases database[2] is a freely accessible tool developed by BioMed Central, which allows users to explore thousands of medical case reports online. For the case reports in the system, key information, including Condition, Symptom, Medication, Intervention, Pathogen, and Subject Area, are extracted from the original documents. An example of these extracted information is shown as Fig 2. Since the reports are uploaded by the domain experts and the key concepts of each report has been identified, it makes the Cases Database a valuable and reliable resource to perform query expansion. In order to do this, we crawled the Cases Database to get the case reports that are relevant to those key concepts identified from the query narratives. We started with a query made with all key concepts identified in a narrative, and recursively removed one concept from the query. For each query we submitted to cases database, we only retrieved the top 200 reports as the candidate documents. With the returned case reports, we extracted the co-occurred key terms/concepts from each field in the returned cases. We kept doing this until the number of concepts in the query is less than 3. At last, all the extracted co-occurred terms are sorted based on the times it shows in the returned results. The top 20 terms from each field were selected and the count was then normalized as the weight.

The key information has been clustered into different categories in Cases Database. This is very useful when the query type should be considered. To be specific, for the diagnosis queries, the terms from "Pathogen" and "Condition" should have a higher weight than others, and if it is a test query, the terms in "Intervention" will be favored, while the terms in "Medication" and "Intervention" will gain a higher weight if it is a treatment query.

We applied the similar method for the concept based representation as well. For the returned co-occurred terms, we mapped them to its CUI using MetaMap. We only kept the MetaMapping but not the candidates, because the candidates will contain noisy CUIs which only covers a part of the original meaning if the there are more than one term in the term based representation. If there were more than one Mappings, we kept all of them and assigned the same weight to each of them.

[2] http://www.casesdatabase.com/. Currently closed.

Fig. 2. An example case report on Cases Database.

2.2 Query expansion with UMLS relationships

Concepts are connected with each other by the semantic relations contribute in the UMLS system. Therefore we can use the semantic connection among concepts to identify similar CUIs for query expansion. There are 476 types of relations in UMLS, but not all of them are useful in expanding the queries. One reason is the types have directions. For example, the semantic type "has ingredient" and "ingredient of" are equivalent if the order of the two concepts is not a important factor. Therefore, for a given query with a certain type, we can expand the query with the concepts that are semantically related to the concepts contained in the original query. We expanded the diagnosis queries with the following semantic types: *may_diagnose, diagnoses, is_finding_of_disease, may_be_finding_of_disease, manifestation_of, is_abnormal_cell_of_disease*. We expanded the treatment queries using the following semantic types: *may_be_treated_by, Treated_by, disease_has_accepted_treatment_with_regimen, may_be_prevented_by*. For the test query, because no clear relation links can be categorized as test, we used another method. We first identified all concepts that can be considered as test by looking at the semantic type of this concept. The concepts with the semantic type "Diagnostic Procedure" or "Laboratory Procedure" are included. This gives more than 30 thousands concepts. We further removed the concepts occur less than 100 times in the collection. This reduces the total number of candidate concepts to 968. For each concepts mentioned in the query, we ranked these test concepts based on the mutual information of the candidate concepts and the concepts mentioned in the query, and the top 5 concepts are chosen as the expansion concepts.

This method can be applied to term based representation as well. For the term based representation, we could mapped the concepts back to the preferred names and use that as the expansion terms.

3 Experiment

3.1 Collection crawling and pre-processing

The collection is crawled from the track home page by downloading the four individual bundles. After decompressed the files and merged them together, the original documents are parsed to extract the "title" and "body" fields. The "back" field is dropped from the original documents because it only contains the acknowledgement and reference of the original document, which is less helpful in answering the clinical questions. The field tags are removed from the original documents.

3.2 Index building

For the term based representation, we built the index based on the parsed data. Stemmer is not applied and the stopwords are not removed. The redundant files provided on the official website are removed from the collection.

For the concept based representation, because the time is limited for converting all the documents to concept based representation, we built a subset of collection using the top 5000 results from each query in term based representation. To be specific, we first applied the baseline method to both the summary narrative and description narrative, and then merged the top 5000 results from each of these two methods together. The redundant results are removed from the collection. We then convert these documents to concept based representation. When do the converting, the negated concepts from the original concept are identified and replaced with a different form.

3.3 Submitted runs and results

We submitted 5 runs this year with the methods we described in Section 2.1 and 2.2. Table 1 summarizes the performances of the official runs and two additional results: **Concept-BL**, which is a concept base retrieval method with no query expansion applied, and **TREC-Median**, which is the average of the median performance of all 91 automatic runs provided by the organizers.

Table 1. Performance of submitted runs.

	Representation	**Expansion**	**infAP**	**infNDCG**	**R-Prec**	**P@10**
UDInfoCDS1	Concept	Cases Database	0.0433	0.1706	0.1348	0.2900
UDInfoCDS2	Concept	Case Database + UMLS	0.0506	**0.1931**	**0.1579**	**0.3167**
UDInfoCDS3	Term	Cases Database	0.0335	0.1461	0.1156	0.2433
UDInfoCDS4	Concept	UMLS	**0.0511**	0.1890	0.1534	0.3067
UDInfoCDS5	Term	No	0.0384	0.1617	0.1339	0.2633
Concept-BL	Concept	No	0.0440	0.1682	0.1350	0.2867
TREC-Median	–	–	0.0316	0.1514	0.1257	0.2333

It is clear that UDInfoCDS2 outperforms than the other methods on most of the measures, which indicates that utilizing the information from two collection together to expand the query could improve the performance. A further analysis would reveal that, by taking the results of UDInfoCDS1, UDInfoCDS2, UDInfoCDS4 and Concept-BL into consideration, the improvement of UDInfoCDS2 is mainly from the concepts identified using UMLS relationship. In addition, the improvement when only using Cases Database is limited on concept base representation, and it actually hurt performance over the term based representation. Moreover, the performance of UDInfoCDS1 is better than UDInfoCDS3, which shows that using concept base representation could help to improve the performance. This is consistent with the findings in the previous work[1, 2]. Last but not the least, comparing the performance of the submitted runs with TREC-Median shows that the submitted methods are better than the median of all submitted runs.

4 Conclusion

In this year's track, we explored the concept based retrieval method with query expansion using external resource in biomedical domain. The experiment results show that using the concepts identified in UMLS system with certain relation as the expanded query concept could improve the performance. However, the concepts from Case Database do not perform well. This reveals the limitation of the how the external resources are utilized. Because this is the first year of this track, we are lack of training topics for the proposed methods. In the future, we plan to study how to better incorporate the external resources into the concept base retrieval.

References

1. A Miguel, P.C., Wang, Y., Fang, H.: Exploiting Domain Thesaurus for Medical Record Retrieval. In: Proceedings of the Twenty-First Text REtrieval Conference, TREC 12. (2012)
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)