Extracting Useful Information from Clinical Notes

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. A new type of query, i.e., note query, which contains plentiful information of the patients, is given in this year's CDS track. Previous results suggest that the additional information in the query may not lead to a better retrieval performance. Therefore, we proposed a method to extract important information from the clinical notes for retrieval. In addition, we also explored the expansion algorithms for abbreviation expansion.

1 Introduction

As the third year of Clinical Decision Support track (CDS track), similar to the ones in 2014 and 2015, this year's CDS track still focus on answering the clinical questions by retrieving the relevant biomedical articles for the users. However, different from the synthetic narratives of the case report for a patient in last two years, the actual electronic health record, named the *"note"* query, is given as a type of query in this year. The note query is extracted from the history of patient illness (HPI) section of the medical record, which contains the symptoms, medical history, treatments has been conducted on the patients, and other information. The note query is obviously different from the description query and the summary query given in the last two years since the redundant information it contains. By comparing the performance of using summary query and description query in last two year's CDS track, it is clear that the additional information provided in description query does not necessarily help to increase the performance. Therefore, an effective algorithm to extract the useful information from the verbose query is the key to successfully handle the new query type in this year's track.

Some existing works in CDS track solve the problem by converting both the query and document from term base representation to concept based representation [1] [2] [3], and then conduct the retrieval task in the concept domain. We argue that this may not fit for this year's problem set up since the preliminary goal is to automatically identify the useful information from the note query. Therefore, we propose to solve this problem in term based representation. To be specific, we use cTAKES to identified the noun phrases from the note query and use them as the query terms. In addition, we utilized the MediLexicon and MetaMap to expand the abbreviations identified from the original query. The expanded abbreviations and the extracted noun phrases are combined as the final query. The results show that using the UMLS expansion could improve the performance of the extracted noun phrases.

2 Useful information extraction and abbreviation expansion

The new query format in CDS 2016 is the admission note of a patient. The note query contains more information about the patients, including the chief complaint, medical history, lab work results, etc. An example of the note query is shown as Fig 1.

Comparing with the description query and summary query, the note query contains more abbreviations and other information about the patient. Two research questions need to be solved before the note query could be used to retrieve documents. The first one is what would be the useful information from the note query, and how to parse the note query in order to extract those useful information.

```
<topic number="1" type="diagnosis">
▼<note>
   78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital6 4406**] (transferred to nursing home for rehab
   on [**12-8**] after several falls out of bed.) He was then readmitted to [**Hospital6 1749**] on [**3120-12-11**]
   after developing acute pulmonary edema/CHF/unresponsiveness?. There was a question whether he had a small MI; he
   reportedly had a small NQWMI. He improved with diuresis and was not intubated. . Yesterday, he was noted to have a
   melanotic stool earlier this evening and then approximately 9 loose BM w/ some melena and some frank blood just
   prior to transfer, unclear quantity.
 </note>
▼<description>
   78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI. Yesterday, he was
   noted to have a melanotic stool and then today he had approximately 9 loose BM w/ some melena and some frank blood
   just prior to transfer, unclear quantity.
 </description>
▼<summary>
   A 78 year old male presents with frequent stools and melena.
 </summary>
</topic>
```

Fig. 1. An example query of TREC CDS track 2016.

The second one is how to identify the abbreviations from the regular terms, and how to expand the abbreviations in order to enrich the query term. We proposed two new methods to conquer these problems.

2.1 Noun phrase identification

Previous work has shown that, based on the semantic types of the terms, a controlled subset of the medical terms could be extracted as the query terms in order to improve the performance[4]. With a close look at the semantic types, such as "Sign or Symptom", "Disease or Syndrome", we could easily identify that the selected query terms are all noun phrases. Inspired by this observation, we assume that the useful information in the note query is captured by the noun phrases. Therefore, we could extract the key message from the note query by identifying the noun phrases.

Apache cTAKES[1] is a powerful toolkit that designed to extract information from electronic medical records. cTAKES is built using the UIMA framework and OpenNLP. One useful component of cTAKES is the noun phrase detection. Since the cTAKES is specially trained with the medical records, the identified noun phrase has both a high recall and a high precision. Therefore, we utilized the cTAKES to perform the noun phrase detection. To be specific, we first removed the de-identify information from the note query. We then feed the cTAKES with the cleaned note query. The cTAKES will output all the identified noun phrases from the query. One thing need to notice is that the identified noun phrases may contain each other. For instance, "nursing home for rehab" is a noun phrases and it also contains the "rehab" in it. In order to increase the coverage of the terms, the root of the noun phrases are kept as the query. We then use MetaMap to get the semantic type for each identified noun phrases. Only the noun phrases whose semantic type is related to biomedical domain would be kept as the query term.

2.2 Abbreviation expansion

Another observation we could get by looking at the example in Fig 1 is that there more abbreviations in the note query than in the description query. The abbreviation could be useful if we could find the the full terms it stands for. However, the challenge of abbreviation expansion is also clear. First, abbreviation identification is a difficult task by itself. It is straightforward to identify the ones with all capital letters, but the ones such as "pmh" (stands for "past medical history") is hard to be found out. We proposed a rule based abbreviation identification method to solve this problem. For each term, except the first word of each sentence, we check if it contains any capital letters. If there is at least

[1] http://ctakes.apache.org/

one letter is a capital letter, this word would be considered as abbreviation candidates. If the word does not contain any capital letters, we would check if it is a regular English term by looking it up in Oxford online dictionary and Merriam Webster online dictionary, respectively. If there is not valid entry for this term, it would also be considered as the abbreviation candidate.

With the identified abbreviation candidates, expanding the abbreviation to its correct form is also a challenge task. For instance, in medical domain, "BM" could mean "Bowel Movement" or "Bachelor of Medicine". Therefore, it is necessary to consider the syntax of the abbreviation and the commonness of expansion. In order to capture these, we included two resources, i.e, UMLS MetaMap[2] and MediLexion[3], for the expansion task. To be specific, we take the sentence that contains the abbreviation candidate as the input for MetaMap. If MetaMap could find the concept which is associated with the abbreviation, it would return the full form of it. We then could extract the full form from the MetaMap result as the expansion term. We denote this method as **MetaMap**. The MediLexion is a online medical terms search engine and abbreviation identification is one of its components. MediLexion would give all possible expansion form for a given abbreviation. The users, mostly the physicians, could pick the ones they think they are looking for by clicking a check mark on that expansion. Therefore, the most common expansion would get the highest pick rates. We fetched the top 3 expansions with highest pick rate for identified abbreviation candidate. We denote this method as **MediLexion**.

2.3 Combine the noun phrase detection with abbreviation expansion

Since the noun phrase detection and abbreviation expansion solve the problem from different angles, it is nature to consider combine the two lists of query terms together. The details of our system structure is shows in Figure 2. We used the extracted noun phrases as the basic query terms, while the abbreviation expansions are used as the expansion term. We trained our model on the description query and summary query on CDS 2014 and 2015 date set. The parameters, such as the weight of expansion, are tuned to get the best performance in terms of infNDCG. We then applied the trained model on 2016 data collection.

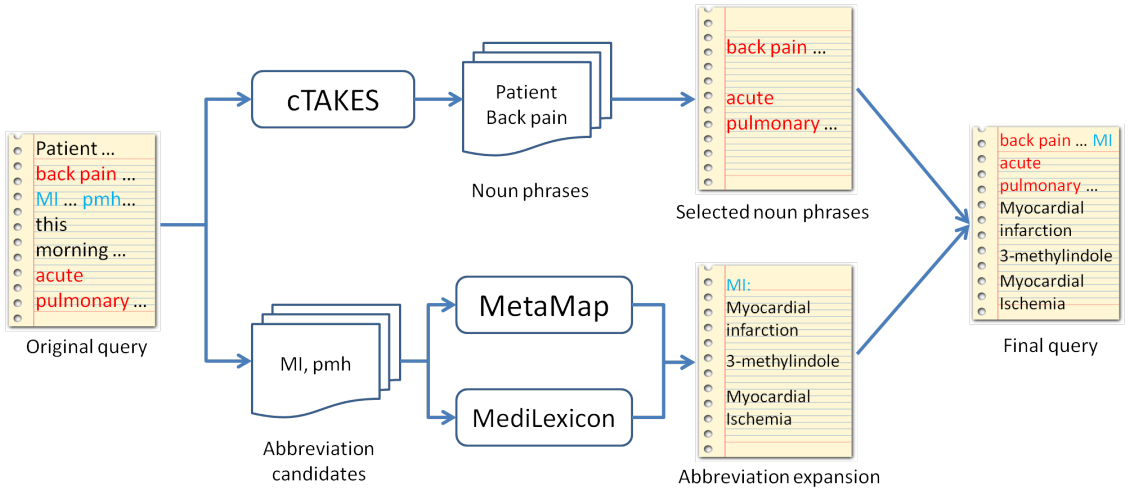

Fig. 2. The structure of our system.

[2] https://metamap.nlm.nih.gov/
[3] http://www.medilexicon.com/

3 Experiment

3.1 Collection pre-processing and index building

Different from the collection used in CDS 2014 and 2015, a new data collection is provided by the organizer. The new collection is a snapshot of the Open Access Subset of PubMed Central. Four compressed files are given, and In total, 1,255,260 documents have been crawled.

After the files has be decompressed and merged together, the original documents are parsed to build the index. To be specific, the field tags are removed from the original file, while all the other components, such as "title", "body" and "back", are kept to build the index. When building the index, we removed a set of stopword based on the default stopword list in Indri. In addition, Porter stemmer is applied when building the index.

3.2 Submitted runs and results

We submitted 5 runs this year: 3 runs using the note query, 1 run using the description query, and 1 run using the summary query. The details of each run can be found in Table 1. The **Noun Phrase**, **UMLS** and **MediLexicon** indicates the methods introduced in section 2.1 and 2.2. We set the weight on expansion of UDelInfoCDS2 to 0.1, while the weight on expansion of UDelInfoCDS3 to 0.15, based on the training results from CDS 2014 and 2015 data.

Table 1. Details of submitted runs.

	Query type	Noun Phrase	UMLS	MediLexicon
UDelInfoCDS1	Note	Yes	No	No
UDelInfoCDS2	Note	Yes	Yes	No
UDelInfoCDS3	Note	Yes	Yes	Yes
UDelInfoCDS4	Description	Yes	No	No
UDelInfoCDS5	Summary	No	No	No

The performance of each run is introduced in Table 2. To compare with, we also included the average of the median performance of all submitted run of each query type, they are noted as **TREC-Median**.

Table 2. Performance of submitted runs.

	Query type	infAP	infNDCG	R-Prec	P@10
UDelInfoCDS1	Note	0.0083	0.1068	0.0664	0.1833
UDelInfoCDS2	Note	0.0084	0.1095	0.0666	0.1667
UDelInfoCDS3	Note	0.0083	0.1042	0.0667	0.1700
TREC-Median	Note	0.0099	0.1227	0.0791	0.1833
UDelInfoCDS4	Description	0.0084	0.1153	0.0739	0.1700
TREC-Median	Description	0.0064	0.1043	0.0648	0.1533
UDelInfoCDS5	Summary	0.0311	0.2362	0.1527	0.3367
TREC-Median	Summary	0.0195	0.1859	0.1220	0.2633

3.3 Result analysis

The UDelInfoCDS4 and UDelInfoCDS5 are better than the TREC median on all measures, while the note runs perform worse than the average of the other groups. By comparing the performance of

UDelInfoCDS1-3 in Table 2, it is clear that the improvement of using UMLS and MediLexion as the abbreviation expansion is marginal. By including the MediLexion expansion, the performance even get worse, which suggests that this resources could be better utilized in other ways.

In order to better understand the reason that our proposed methods failed to parse the note correctly, we did the per query analysis on the note runs by comparing the extracted note query and the given summary query. The results show that our method may not perform well because it fail to identify the more important section from the note query. Not only the symptoms, but also other information, such as the "review of system" (a section contains the basic information of the patient), is also given in the note query. However, extracting the noun phrases from the less important sections lower the weight of the weight of the query terms, it may also introduce noisy term as well.

4 Conclusion

In this year's track, we explored the information extraction method on clinical notes, and also the abbreviation expansion techniques. The experiment results show that the proposed methods does not perform well on the new data collection. This reveals the limitation of extraction procedure. In the future, we plan to study how could we use the machine learning algorithm to help predict whether a given term is important or not.

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. 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)
3. Wang, Y., Fang, H.: Exploring the Query Expansion Methods for Concept Based Representation. In: Proceedings of the Twenty-Third Text REtrieval Conference, TREC 14. (2014)
4. Bedrick, S., Edinger, T., Cohen, A., Hersh, W.: Identifying Patients for Clinical Studies from Electronic Health Records: TREC 2012 Medical Records Track at OHSU. In: Proceedings of the Twenty-First Text REtrieval Conference, TREC 12. (2012)