Query Expansion with Automatically Predicted Diagnosis: iRiS at TREC CDS track 2016

Danchen Zhang, Daqing He, Sanqiang Zhao, Lei Li
School of Information Sciences, University of Pittsburgh
{daz45, dah44, saz31, lli}@pitt.edu

Abstract. This paper describes the participation of the iRiS team from University of Pittsburgh in the TREC Clinical Decision Support (CDS) track in 2016. According to the track requirements, 1,000 most relevant biomedical articles from the PubMed Collection were retrieved based on information needs of 30 patients with their electronic health records (EHR) notes. Our approach focuses on using MetaMap to extract medical concepts, and using Wikipedia knowledge base to predict the patient diagnosis. Consequently, the original query is expanded with the predicted diagnosis before sent to search PubMed articles. Parameters were tuned based on CDS 2014 and 2015, and Indri is used to construct the index of the collection. Our automatic runs on description ranks 2nd and our manual runs on notes ranks 3rd in all submitted runs.

Keywords: Medical text retrieval; diagnose prediction; query expansion

1 Introduction

The TREC 2016 CDS track focuses on the biomedical literature retrieval to support the physicians in making the clinical decisions. The participants need to provide relevant biomedical articles in terms of three most generic clinical questions (Simpson, et al., 2014):

- Q1: What is the patient's diagnosis?
- Q2: What tests should the patient receive?
- Q3: How should the patient be treated?

In past two years, the inputs to the participant retrieval system were the hospital summary and descriptions about the patients' visit. In this year TREC also provides the admission notes. This newly added information imposes challenges to the retrieval task because it contains a lot of medical abbreviations, which are hard to read even for people who have a little medical knowledge. However, the new topics do contain much more patient history information, while in past two years, the topics basically only have the patients' most urgent disease or symptoms.

In addition, the document collection in this year is much larger. This year's collection has 1.25 million articles from the Open Access Subset[1] of PubMed Central[2] (PMC), while the target document collection in CDS 2014 and 2015 contains only 733,138 articles.

An accurate query is important for effectively searching the relevant biomedical literatures from the PMC collection. In previous works, researchers constructed queries with the medical concepts recognized from the EHR notes, and enhanced the query with pseudo relevance feedback or online information (Roberts, et al., 2015). In this work, we propose to firstly expand the query with the automatically predicted diagnosis. This is under the assumption

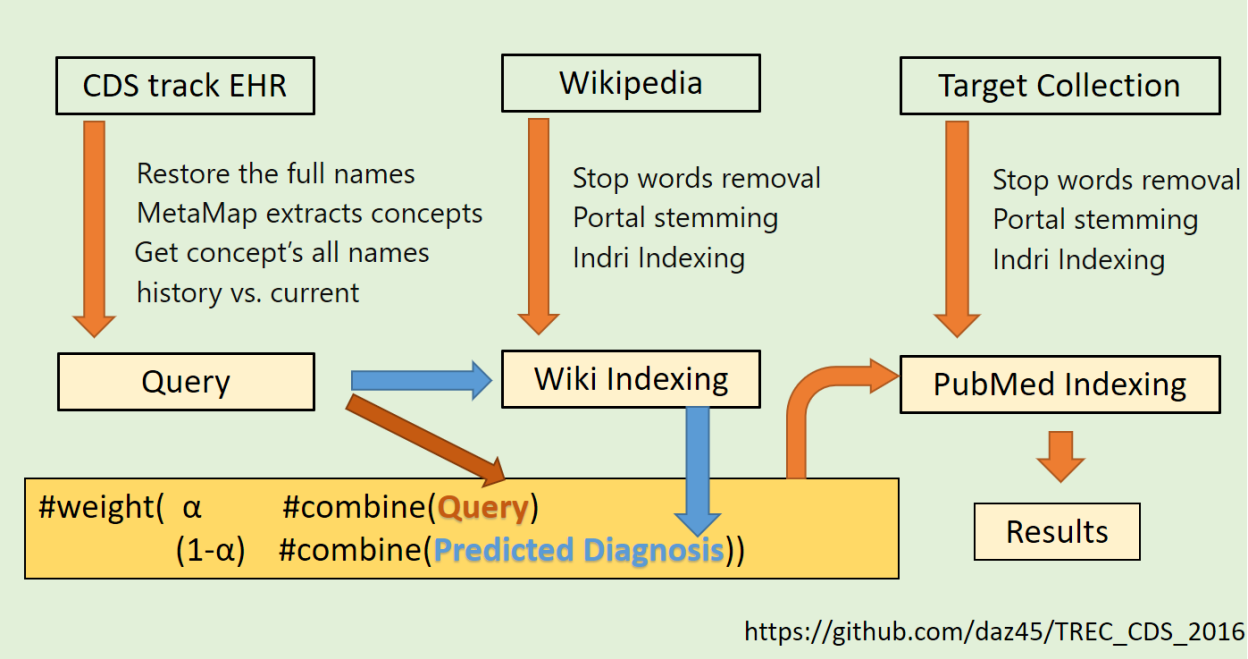
[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[2] https://www.ncbi.nlm.nih.gov/pmc/

that patient diagnosis information can better disclose the physician's true information need in making clinical decisions.

2 Methodology

We first preprocessing the target collection and all topics. Then the query is constructed with MetaMap. Next, we automatically generate the patient diagnosis with knowledge from Wikipedia. Finally, the diagnosis is used to expand the original query. The whole system is illustrated in Figure 1.

Figure 1. iRiS CDS track system framework



2.1 Data preprocessing

HER notes contain lots of abbreviations, which hinders the retrieval. Therefore, we replace the abbreviations with their full names using the UMLS vocabulary list. However, this approach could not resolve certain abbreviations that appear only in these EHR notes, such as 'c/b' for 'complicated by'. We, therefore, collected their full names via Google search results. To help other researchers, we have posted the whole abbreviations list in GitHub[3].

Target document collection was indexed by Indri (Strohman, et al., 2005), and all articles were preprocessed with stop word removal and Portal stemming. In diagnosis prediction procedure, we utilized Wikipedia as the prediction evidence. The English Wikipedia dump (enwiki)[4] was downloaded in March 5th, 2016, and all wiki pages were preprocessed and indexed in the same way with the target document collection.

2.2 Query construction

The EHR notes sometimes describe patient denies some symptoms. Therefore, negation terms are removed with the NegEx algorithm (Chapman, et al., 2001).

Then, following the past works, MetaMap[5], published by NLH, is used to extract the medical concepts from EHR notes. Summary and description can be directly processed by MetaMap, and we submitted three automatic runs.

However, the admission notes cannot be directly processed by the MetaMap. It contains many medical concepts, which might not be related to the patient symptoms. If the sentence is hard to understand which is usually caused by the bad format, or the sentence simply describe

[3] https://github.com/daz45/TREC_CDS_2016
[4] https://dumps.wikimedia.org/enwiki/20160701/
[5] https://metamap.nlm.nih.gov/

the patient body checking result, we delete them manually. Afterwards, the cleaned text is processed in the same way as the description and summary. For example, the note in topic 5, sentences are deleted after “In the ED, initial vs were: 80”.

For each recognized medical concept, we add all its names from UMLS knowledge base into the query. In addition, we believe current symptoms and signs is more important than patient disease history. Thus, we combine all history disease together with ‘#combine’ to give them lower weight. Further, if a medical concept has several names, we use ‘#combine’ to combine them. If a name has several terms, we use ‘#uw’ to combine them. For example, the final Indri query extracted from Topic 25 summary is shown as following:

```
#combine(
        #combine(AF afib #uw(atrial fibrillation))
        #combine(COPD #uw(Chronic obstructive pulmonary disease)
                        #uw(Chronic obstructive airway disease))
        #combine(hypertension HTN HBP #uw(high blood pressure))
        #combine(hyperlipidemia lipidemia)
        #combine(atrioseptoplasty #uw(repair atrial septum defect))
)
#combine(dyspnea SOB #uw(shortness of breath))
#combine(AF afib #uw(atrial fibrillation))
```

2.3 Wikipedia based automatic diagnosis prediction

The TREC CDS track 2015 overview (Roberts, et al., 2015) shows that: if the patient diagnosis information is utilized in the retrieval process, the mean infNDCG of submitted runs rapidly increases from 20.99% to 28.70%, and the median infNDCG increases from 22.88% to 32.12%. This observation proves the high utility of the patient diagnosis information, and motivates us to propose a method that automatically predicts the patient diagnosis.

Wikipedia has rich information on worldwide diseases. Usually, a wiki page named by a disease contains the symptoms, causes, pathophysiology, and diagnosis. We assume that the diagnosis of the patient is the disease whose wiki page is the most relevant to the query generated from patient EHR notes. We use the query extracted from EHR notes to search the most relevant disease wiki page. The page name is regarded as the predicted diagnosis.

For each Wikipedia page, only the title and content is kept, with tags, Reference, External Link and See Also sections removed from wiki article pages. In this retrieval task, Wikipedia data is indexed by Indri, and searched by the language model with Dirichlet smoothing.

2.4 Query Expansion with Diagnosis

Finally, the predicted diagnosis expands the original query. In Indri query language, such expansion is conducted as follows:

```
#weight (α #combine(original query)  (1- α) #combine(predicted diagnosis))
```

where α is the weighting parameter, ranging from 0 to 1. Similarly, the target document collection is searched by the language model with Dirichlet smoothing.

3 Experiments and discussion

3.1 Runs and results

We tune the parameters on CDS track 2014 and 2015 to get best infNDCG. We submitted five runs, as shown in Table 1. Run 1 to 3 are automatic runs, while Run 4 and 5 are manual runs. All 5 runs use the same method, only differs on query expansion parameter α and topic types. Dirichlet smoothing parameter mu is set as 4000.

Table 1. Results on CDS track 2016

	Run Name	Topic Type	Parameter	infNDCG	infAP	iP10
CDS 2016	Run1	Summary	α =0.8	20.18%	1.96%	28.67%
	Run2	Description	α =0.8	15.10%	1.39%	24.67%
	Run3	Description	α =0.7	15.88%	1.62%	25.67%
	Run4	Notes	α =0.8	16.71%	1.85%	24.33%
	Run5	Notes	α =0.7	18.17%	2.05%	27.00%

3.2 Is the predicted diagnosis correct and useful?

In Figure 2, we compare the basic language model (Baseline) with our proposed model (Run1) on all 30 topics on the summary. The best and median result of all participants are also shown in Figure 2. Totally, 21 out of 30 topics get improved by the predicted diagnosis information, 1 topic remains the same (infNDCG=0 in topic 27), and 8 decrease the retrieval performance. Through Wilcoxon signed-ranks test, Run1 significantly outperforms the baseline with p-value<0.05. The predicted diagnosis is shown in Table 2. For the topic 8 and 20, our proposed method obtained the best performance.

In Table 2, the disease names in bold character are those improve the retrieval performance, while only infNDCG of topic 27 remains same. Since we have no correct diagnosis information, we can only evaluate the predicted diagnosis correctness based on the retrieval performance. About 70% of predicted diagnosis is correct, i.e., useful in query expansion.

Figure 2. Baseline and Run1 on infNDCG

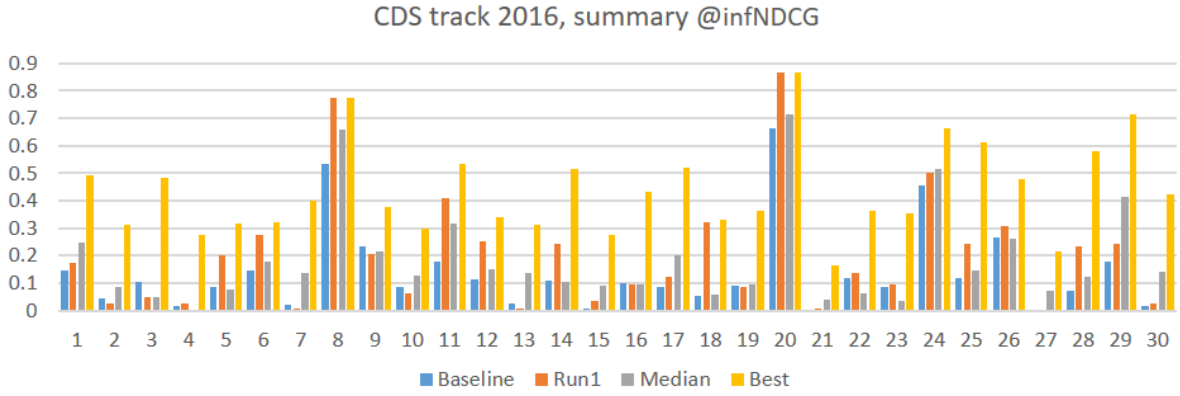


Table 2. Predicted diagnosis in Run1

1	**Fecal occult blood**	11	**Angina pectoris**	21	**Sepsis**
2	Heterotopic ossification	12	**Head injury**	22	**Cardiac arrest**
3	Anorexia nervosa	13	Iron-deficiency anemia	23	**Gastrointestinal bleeding**
4	**Pulmonary contusion**	14	**Pneumonia**	24	**Bowel obstruction ***
5	**Pneumonia**	15	**Lung cancer**	25	**Atrial fibrillation***
6	**Cholecystitis**	16	Apraxia of speech	26	**Atrial fibrillation ***
7	Cirrhosis	17	**Heart failure ***	27	Kernohan notch
8	**Diabetic ketoacidosis**	18	**Pancreatitis**	28	**Gastrointestinal bleeding**
9	Infant respiratory distress syndrome *	19	Hepatic encephalopathy	29	**Idiopathic pulmonary fibrosis ***
10	Syndrome of inappropriate antidiuretic hormone secretion	20	**Gallstone**	30	**Sinus bradycardia**

Also, from Table 2, we find that the predicted diagnosis help most in last 10 questions. 9 of 10 improve the retrieval performance, while only one topic remains same performance. It seems our prediction is more accurate in the “How should the patient be treated” related topics.

Further, we find that 6 generated diagnosis appear in the original query text, which are labeled with “*” in table 2, while the other 24 diagnosis doesn’t appear in the original query, and lead new concept in the expansion procedure. For these 6 topics, generated diagnosis give extra weight to the diagnosis tokens in the new query. It implies that the extracted medical concepts contribute differently. Some medical concepts are more important than others, and should have higher weight.

Our results for Run2 and Run3 is shown in Figure 3. We obtained best performance on topic 1, 5, 6, 9, 11, 14, 16 and 20. Result of Run 4 and Run 5 is shown in Figure 4. We obtained best performance on topic 5, 7, 14, and 20.

Figure 3. Run2 and Run3 on infNDCG

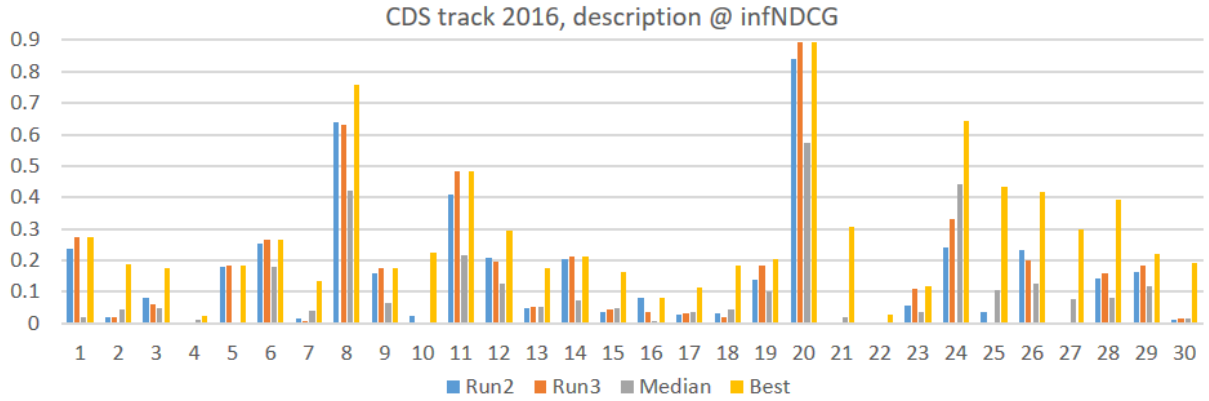
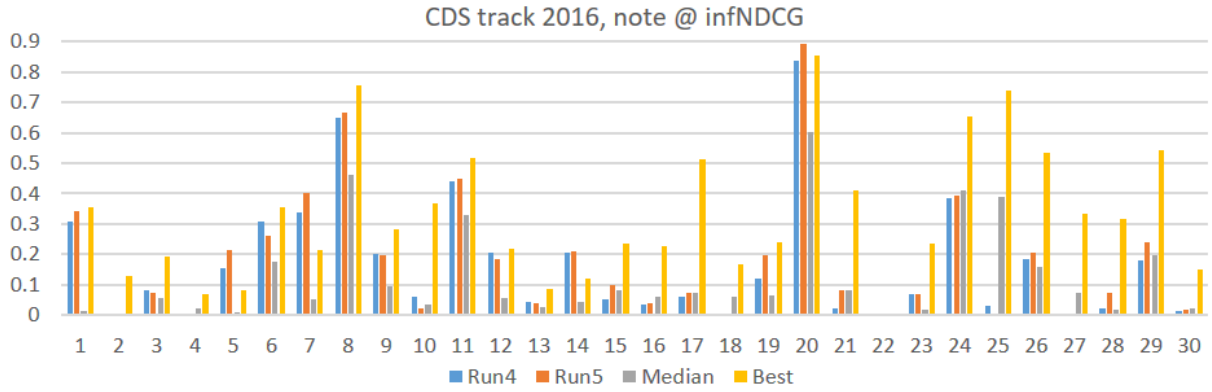


Figure 4. Run4 and Run5 on infNDCG



4 Discussion and Conclusion

A novel mechanism, Wikipedia based automatic diagnosis prediction, is proposed to enhance the clinical decision support system. Given patients’ disorder related information, we search through the Wikipedia collection to get the disease of highest probability, and use it to expand the original query. This idea has been proven to be effective.

However, there are still limitations in current works. For some topics, Wikipedia does not have the candidate disease wiki page, then the correct diagnosis cannot be obtained. Even provided with predicted diagnosis, it is still a bag-of-word retrieval system. In the next step, algorithms in deep learning area can be used to dig further on semantic relevance.

5 References

Roberts, Kirk, Simpson, Matthew S., Voorhees Ellen, Hesh R. William. (2015) "Overview of the TREC 2015 Clinical Decision Support Track."

Simpson, Matthew S., Ellen M. Voorhees, and William Hersh. (2014) Overview of the trec 2014 clinical decision support track. LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD, 2014.

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005, May). Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis (Vol. 2, No. 6, pp. 2-6).

Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001 Oct 31;34(5):301-10.