Enhancing Clinical Decision Support Systems with Public Knowledge Bases

Danchen Zhang, Daqing He[1]
Department of Informatics and Networked Systems
School of Computing and Information
University of Pittsburgh

Abstract. With vast amount of biomedical literature available online, doctors have the benefits of consulting the literature before making clinical decisions, but they are facing the daunting task of finding needles in haystacks. In this situation, it would help doctors if an effective clinical decision support system could generate accurate queries and return a manageable size of highly useful articles. Existing studies showed the usefulness of patients' diagnosis information in such scenario, but diagnosis is often missing in most cases. Furthermore, existing diagnosis prediction systems mainly focus on predicting a small range of diseases with well-formatted features, and it is still a great challenge to perform large-scale automatic diagnosis predictions based on noisy patient medical records. In this paper, we propose automatic diagnosis prediction methods for enhancing the retrieval in a clinical decision support system, where the prediction is based on evidences automatically collected from publicly accessible online knowledge bases such as Wikipedia and Semantic MEDLINE Database (SemMedDB). The assumption is that relevant diseases and their corresponding symptoms co-occur more frequently in these knowledge bases. Our methods performance was evaluated using test collections from the Clinical Decision Support (CDS) track in TREC 2014, 2015 and 2016. The results show that our best method can automatically predict diagnosis with about 65.56% usefulness, and such predictions can significantly improve the biomedical literatures retrieval. Our methods can generate comparable retrieval results to the state-of-art methods, which utilize much more complicated methods and some manually crafted medical knowledge. One possible future work is to apply these methods in collaboration with real doctors.

Notes: a portion of this work was published in iConference 2017 as a poster, which won the best poster award. This paper greatly expands the research scope over that poster.

Keywords: Medical text retrieval; diagnose prediction; query expansion

1 Introduction

To accurately make clinical decisions, doctors may sometimes have to consult external information for reference. The published biomedical articles, which are expert written

[1] Corresponding authors: dah44@pitt.edu

materials that cover nearly all topics in the medical area, are the most commonly source of reference [2, 3, 25]. Many biomedical literature search engines, such as PubMed[2], have been developed to facilitate this data access task. However, most platforms only support keywords retrieval, thus impose a high requirement for the doctors to accurately construct their queries.

However, it is often hard for doctors to generate accurate queries because their information needs are often complicated and exploratory. For example, Ely et al. [1] identified that doctors' clinical questions could touch the following three aspects at the same time:

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

Because doctors often only have limited information about the patient's current condition, such as symptoms or disease history, they usually find it hard to generate an accurate query. Therefore, better search support technologies are critical needed in such scenarios.

It is under this goal of providing doctors with more effective biomedical text retrieval technologies, Text Retrieval Conference (TREC) hosted Clinical Decision Support (CDS) track[3] between 2014 and 2016. The participants of the track have worked on retrieving relevant biomedical articles to answer the above-mentioned three questions with 90 sample electronic health records (EHRs) [2, 3].

The outcomes of TREC CDS provide two important insights. The first one is that terms extracted from EHRs alone are too ambiguous to reflect the true information needs. For example, Balaneshin-kordan et al. [8] found through their 2014 CDS participation that non-relevant documents talking about wrong diseases were returned when the queries contained evidences only from the EHRs. This is because the provided EHRs data only contains some incomplete disorder related information such as disease history, symptoms, and testing results (see Figure 1). However, the same symptoms and disease history information might partially be shared by different diseases. An example is that patients of either hypothyroidism or hyperthyroidism might both have symptoms of dyspnea, hair loss and fatigue. Thus, extra information is needed to enrich the queries.

The second insight is that correctly identified diagnosis, which clearly state the possible disease, would significantly improve the retrieval performance. For example, 2015 TREC CDS had task A and task B, both of which have EHRs for 30 patients, but 20 out of 30 EHRs in task B were also provided with diagnosis information. The evaluation results showed that, compared with that of task A, both the median and mean performance of task B increased by 8%, reflecting the benefits of having the diagnosis in helping biomedical literature retrieval [3].

---

[2] http://www.ncbi.nlm.nih.gov/pubmed

[3] http://trec-cds.appspot.com/

Figure 1: An example topic from CDS task B of TREC 2015

```
<topic number="13" type="test">
  <description>
    A 5-year-old boy presents to the emergency department with complaints of progressively worsening
    dysphagia, drooling, fever and vocal changes. He is toxic-appearing, and leans forward while
    sitting on his mother's lap. He is drooling and speaks with a muffled "hot potato" voice. The
    parents deny the possibility of foreign body ingestion or trauma, and they report that they are
    delaying some of his vaccines.
  </description>
  <summary>
    A 5-year-old boy presents with difficulty in breathing, stridor, drooling, fever, dysphagia and
    voice change.
  </summary>
</topic>
```

These above-mentioned findings motivated us to conduct diagnosis predication before generating queries for helping doctors in their biomedical text retrieval. Particularly, we are interested in obtaining medical diagnosis information from public available knowledge bases because large quantity of EHRs are usually not available openly.

In this paper, we will to explore two types of large scale open knowledge bases. Both cover information about wide range of diseases and related information, so they are suitable to be mined for possible diagnosis related information. The first one is Wikipedia, which represents the type of open corpus of free-text. Wikipedia presents information at word level, with rich context information about each disease and its related symptoms, tests, and treatments. The second one is Semantic MEDLINE Database (SemMedDB), which consists of medical concepts extracted from PubMed[4] literatures. Different to Wikipedia, SemMedDB organizes its information around medical concepts and their relationships. This conceptual oriented expression can be useful because doctors, medical literature and Wikipedia articles might express the same concepts but with different words or phrases. However, SemMedDB has its own problems. As Kilicoglu et al. [22] reported that NLM's SemRep, the tool used to build SemMedDB, only achieves a 75% extraction accuracy. This means that SemMedDB itself contains many noisy extraction results too. Consequently, our method uses the combination of Wikipedia and SemMedDB to extract diagnosis related information for support doctors' biomedical literature retrieval.

Once the diagnosis is predicted, we view the rest of the method as a query expansion problem. The expanded query contains the original parts that is generated from the disorder related medical concepts recognized in the EHRs, and the expanded part is the predicted diagnosis. Therefore, our model consists of five modules: medical concept extraction with MetaMap, Wikipedia based diagnosis predictor, SemMedDB based diagnosis predictor, prediction fusion, and query expansion with diagnosis.

There are quite a few works studying the large-scale diagnosis prediction, so we will compare our work with them in this paper. Since we have no correct diagnosis, we will validate our method on CDS retrieval performance extrinsically.

The remainder of this paper is presented as follows. Section 2 shows the related works. Section 3 describes our proposed methods. Sections 4 and 5 give the experiments and discussions. Finally, Section 6 presents the conclusion and future work.

---

[4] PubMed is a medical literature corpus, comprising more than 27 million citations. Available at https://www.ncbi.nlm.nih.gov/pubmed/.

2 Related Works

The related work can be divided into two parts. The first one examines the existing studies on achieving automatic diagnosis prediction, and the second part talks about query expansion in medical retrieval.

2.1 Automatic Diagnosis Prediction

Automatic diagnosis prediction is a very popular research topic that attracts many researchers. Esfandiari et al. [11] gave a comprehensive review of the studies in this area. According to their summarization, most studies regarded diagnosis prediction as a classification task. For example, Yeh et al. [12] utilized the patients' history diseases, blood test results and physical exam results as the features, and trained classifiers to predict the probability of getting a cerebrovascular disease. Besides, other studies tried to predict the diagnosis with regression methods, clustering methods, association rules or hybrid systems [11]. For example, Brines tried to predict the risk of Alzheimer disease through a regression model [17]. However, these works usually targeted on sensitively and accurately predicting a small range of diseases. For instance, Yeh et al. [12] only concentrated on the cerebrovascular disease prediction, and obtained an accuracy 98.01% and sensitivity 94.68%.

There have been a few works exploring the large-scale auto-diagnosis. Isola et al. [18] proposed to a neural network and a KNN based system to predict diseases. But their work focused on the implementation of such a system, and did not provide performance data on the prediction. Gomathi et al. [19] constructed a well-structured database to compute the probability of a disease based on a patient's symptoms, but also did not provide prediction performance. Liu et al. [20] proposed a quite interesting android platform, allowing users interactively to communicate with the auto-diagnosis application, and used user's explicit feedback in diagnosis probability computation. Nie et al. [26] proposed a deep learning based framework to predict large-scale diseases for users on Question-Answer (QA) platforms. They trained the network on collected QA data, and carefully refined the network on different disease prediction with precisely collected disease related QA data. They validated their system on predicting 20 diseases, and found that their method outperformed KNN, SVM, Decision Tree and Naïve Bayes. Although their system worked on large-scale disease prediction, their system highly relied on good quality training data. Koopman et al. [21] found that, when the training data is imbalanced, their system could encounter the classification/prediction failure for the rare but important diseases which might have little data for training.

In comparison to the related work on this part, our methods work on predicting wide range of diseases. This is because doctors' medical literature search can be on all sorts of diseases so that our automatic diagnosis prediction system cannot afford to work only on a small number of diseases. Furthermore, we concentrate on proposing a diagnosis prediction system with no need on large quantity of training data. Finally, different to past studies which only accepted well-formatted features, our system can deal with noisy medical free text. Overall, our goal is to make the prediction system quite robust and need little effort when changing scenarios.

2.2 Query Expansion in Biomedical Text Retrieval

In the generic information retrieval area, query expansion is a common module to enhance the original search, but it may lead to query drift if not design carefully [4, 5]. In CDS task, nearly all previous studies utilized query expansion to enhance the original query [6-10].

Most of them extracted expanded terms from Pseudo Relevance Feedback (PRF) documents within the collection [6, 8]. For example, Choi et at. [6] presented the best result in CDS 2014. They utilized the most frequent Medical Subject Headings (MeSH) terms that label the PRF documents to expand the original query, and then used the classifiers to re-rank the returned document list. Both steps showed significant improvement. Balaneshin-kordan et al. [8] expanded their query with terms selected both from PRF files and Google search results.

There were also other works using important medical concepts extracted from external resources to expand the query [7, 8, 10]. Oh et al. [7] proposed to enhance external expansion model (EEM) with cluster-based document model (CBEEM) to expand the query more accurately. Their PRF model consisted of top-ranked documents both in the target dataset and in the external collection, Wikipedia. Their results outperformed the best runs in CDS task of TREC 2014. Song et al. [10] extracted the most frequent MeSH terms appearing in the Google search result returned with the original query.

Past works explored different retrieval models in biomedical text retrieval task. Both Balaneshin-kordan et al. [8] and Xie et al. [9] used the Markov Random Field (MRF) model and got very high retrieval performance. Song et al. [10] proposed to retrieve relevant biomedical articles through combining three retrieval models, including BM25, PL2 and BB2, and their result performed the best in 2015 CDS task B.

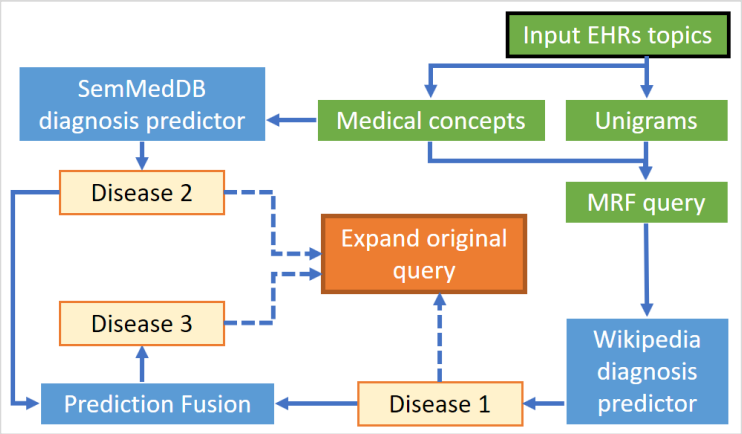
In the query expansion procedure of all these previous literatures, it showed that the quality of the expanded terms is important. Since the diagnosis can better reflect users' true information needs, we believe that adding the diagnosis predicted with Wikipedia and SemMedDB to the original query can help to clarify the original query and does not introduce too much noise.

3 Our Methods

As stated, we propose in this paper novel methods to enhance the clinical decision system by automatically predicting patient disease with the public online knowledge base. To be able to assist doctors in their clinical decisions on a wide range of diseases, the diagnosis prediction methods we build here can make predictions for large number and wide scale of diseases rather than concentrating on a small number of narrowly defined diseases. Therefore, our diagnosis prediction algorithms need to draw evidence and help from large-scale public accessible data and knowledge bases, and our methods concentrate on two knowledge bases, Wikipedia, the free text collection that provides rich contextual information but at the word level, and SemMedDB, the conceptual level knowledge bases consisting of the medical concepts and their relationships extracted from PubMed. Figure 2 shows our methods, the integration of some or all of the following functional modules:

- *Medical concepts extractor*, which automatically identifies the medical concepts in the given EHRs for a search task.
- *Wikipedia-based diagnosis predictor*, which utilizes Wikipedia knowledge to predict the most probable disease diagnosis based on the EHRs of a search task.
- *SemMedDB-based diagnosis predictor*, which utilizes SemMedDB knowledge to predict the most probable disease diagnosis based on the EHRs of a search task.
- *Fusion-based diagnosis predictor*, which combine the ranking list of above two predictors to get most probable disease diagnosis.
- *Query expansion with diagnosis*, which expands the original query with the predicted diagnosis.

Figure 2: Methods of enhancing CDS system by predicting the patient's diagnosis



3.1 Medical Concept Extractor (MCE)

Following previous literatures [6-10], the medical concept extractor (MCE) module relies on Unified Medical Language System (UMLS) vocabulary to identify medical concepts in EHRs. UMLS is a knowledge base published by U. S. National Library of Medicine (NLM), which contains a full list of medical concepts [13]. Each UMLS concept has a Semantic Type (ST) attribute. For example, concept "Dyspnea" has its ST as "sosy", and the ST of "woman" is "popg". In this work, medical concepts with following STs are kept: *acab, anab, bact, bdsu, blor, bpoc, bpoc, celc, cgab, comd, diap, dsyn, emod, euka, fndg, fngs, food, ftcn, hlca, inpo, lbpr, lbtr, menp, mobd, neop, ortf, patf, phsu, sosy, tisu, tmco, topp, virs.* These STs are selected because such medical concepts provide important patient disease related information, and frequently appears in EHRs.

The implementation of MCE relies on MetaMap to extract the concepts from EHRs. MetaMap is a popular medical text mining tool developed by NLM [14]. MetaMap can detect the negation expression. For example, "trauma" will be ignored in "she has no history of trauma". Our extractor module ignores negation expressions, and assigns a

unique medical concept identifier - Concept Unique Identifier (CUI) in the UMLS Metathesaurus - to each extracted medical concept.

3.2 Wikipedia Based Diagnosis Predictor (Wiki-DP)

Wikipedia-based diagnosis Predictor (Wiki-DP) module relies on information extracted from Wikipedia for diagnosis prediction. As an open and rich knowledge base, Wikipedia covers wide range of diseases and their related information, which in most cases sufficient to act as an external resource for biomedical retrieval. Typically, a wiki page titled with a disease name would contain information about the causes, symptoms, pathophysiology and diagnosis of the disease. Through such information, we can calculate the co-occurrence between a disease and certain symptoms, which can then be used to create models for ranking possible diseases based on the extracted symptoms.

3.2.1 Initial Query Composition

After the medical concepts are extracted, we can construct a query, which will be used in disease predictor. The query is constructed with Markov Random Field (MRF) model [23]. Works in [8, 9] also use MRF model in medical retrieval. In MRF, query consists of cliques, for example, EHR query consists of extracted medical concepts. If a clique contains several terms, term dependence information can help retrieve the relevant documents. Such term dependence in one clique can be described as terms appearing in the document in an ordered/unordered sequence within a window size. Given a query Q, document D can be ranked as:

$$P(D|Q) \overset{rank}{\Longleftrightarrow} \sum_{c\in T} \lambda_T f_T(D|c) + \sum_{c\in O} \lambda_O f_T(D|c) + \sum_{c\in U} \lambda_U f_T(D|c) \quad (1)$$

Where, T is the unigram clique set in the query which has no term dependency, O is the cliques of ordered terms having sequential dependency, and U is cliques of unordered terms having sequential dependency; $\lambda_T, \lambda_O, \lambda_U$ is weighting parameters for unigram, ordered cliques, and unordered cliques, respectively, and they add up to 1. $f(D|c)$ is the probability of the document appearing given the clique. For example, for an unordered clique "chest pain", $f_T(D|c)$ can be described in Indri language [16] as "#uw(chest pain)".

In this work, according to the experiment results of training data, we only consider the unordered term dependency and independent unigrams. Each extracted medical concepts is a clique. In Indri query language, such query can be written as:

#weight (λ_T *#combine(unigrams)* λ_U *#combine(medical concepts))* (2)

Where $\lambda_T + \lambda_U = 1$. For example, after processing the summary text in Figure 1, we will have a set of terms: *breathing difficulty, dysphagia, fever, drooling, stridor,* and *voice change*. Thus, the query is:

#weight(λ_T *#combine(breathing difficulty dysphagia fever drooling stridor voice change)* λ_U *#combine(#uw(breathing difficulty) dysphagia fever drooling stridor #uw(voice change)))*

3.2.2 Ranking the candidate diseases

We assume that the disease (predicted diagnosis) co-occurs frequently with its symptoms in the Wikipedia articles. Therefore, using the above obtained query, we can retrieve the most relevant Wikipedia articles, and expect that the most relevant wiki pages contain the diagnosis. We downloaded Wikipedia and index it with Indri (See more details in next section). Although there could be many Wikipedia articles talking about entities in other domains, such as foods, traveling, and policy, these noisy entities do not bother our predictions. This is probably because few terms are shared between the query and the unrelated entities. Indri evaluates document relevance by language modeling with Dirichlet smoothing, in which we have a smoothing parameter μ defining the degree to overcome data sparseness and 'zero-probability' problem [16].

After wiki pages are returned, MetaMap is used again to identify if the article title is a disease name. If yes, this title is selected as diagnosis. If not, current article is ignored and next article is considered. In total top 10 wiki articles participate in provide possible diagnosis.

3.3 SemMedDB-Based Diagnosis Predictor (SMDB-DP)

Wiki-DP draws disease information based on a search to the free text content of Wikipedia, but SemMedDB-Based Diagnosis Predictor (SMDB-DP) utilizes SemMedDB, which is a repository of concepts and their relationships extracted from PubMed using a tool called SemRep[5]. SemMedDB is a network of medical concepts, with concepts represented as nodes and the co-occurrence of concepts in the medical literature represented as edges. Like Wiki-DP, SMDB-DP also assumes that the true diagnosis should be the disease co-occur frequently with the extracted symptoms (or signs or history diseases). Therefore, the disease concepts in SemMedDB can be ranked based on their co-occurrence with the extracted symptoms.

However, for SMDB-DP to achieve good performance in diagnosis prediction, two important issues should be resolved. The first one is about partial matching of extracted symptoms and those mentioned in SemMedDB. Because all extracted medical concepts might not co-occur with a disease in one medical article, it is common that only parts of the extracted symptoms are mentioned together with a disease in one document. Therefore, partial matching needs to be handled.

The second issue is that, although it would make the model much simpler, the symptoms associated with a disease cannot be regarded independent to each other. The risk of viewing those symptoms to be independent is that it can cause sever topic drift (disease drift) problem. For example, popular disease can appear in thousands of papers, such as *fever* appears in 126,396 articles, while rare diseases are very infrequently, for example, *voice change* only appear in 51 articles. If fever and voice change is independent of each other, the predicted candidate disease will be dominated by fever related popular diseases.

Hence, suppose *n* medical concepts are extracted from the medical record, and we assume that they are dependent of each other, SMDB-DP would consider a disease to

[5] SemMedDB is accessible in https://skr3.nlm.nih.gov/SemMedDB/dbinfo.html

be true only when at least $\lceil\sqrt{n}\rceil$ concepts co-occur. And the probability of a disease being the true diagnosis is calculated as:

$$\text{P(disease | extracted concepts)} = \frac{Co{-}occur\ doc\ count\ contains\ disease}{Co{-}occur\ doc\ count} \quad (3)$$

In this way, we can get a disease ranking list for each query.

3.3 Prediction Fusion

We assume that a disease is highly probable to be correct if it is predicted as true diagnosis by both SemMedDB and Wikipedia. From the experiments results in next section, we find that Wikipedia has a better and more robust prediction across three datasets, and hence we use the following rules to combine the prediction outputs from Wiki-DP and SMDB-DP:

- Only top 10 diseases are considered in both ranking lists.
- If the two lists share the same diseases, the shared diseases are kept and ranked with Wikipedia ranking score.
- If the two list do not share diseases, select the top disease in Wikipedia ranking list.

3.4 Query Expansion with Diagnosis

After obtaining the predicted diagnosis, we can use it to expand the original query. In the format of Indri query language, this combination is shown as follows:

#weight ((1-α) #combine (original query) α #combine (predicted diagnosis)) (4)

where weighting parameters α ranges from 0 to 1. Original query is the query used in Wiki-DP module to retrieve Wikipedia articles, which contains patient symptom information but without diagnosis.

4 Experiments

We conducted a set of evaluations to validate the effectiveness of our proposed method.

4.1 Dataset and Metrics

Our study included five datasets (see Table 1). Two datasets were used for building diagnosis prediction algorithms. The English Wikipedia collection (enwiki)[6] was used for Wiki-DP. It was downloaded on March 5th, 2016, and contains 5.79 million articles. Only the title and the content of each article were kept. Tags, references, external links and see also parts were all removed. The Wikipedia collection was firstly performed stop word removal and stemming using Porter stemmer, then it was indexed by Indri,

[6] https://dumps.wikimedia.org/enwiki/20160701/

The SemMedDB database contains medical concepts extracted from 26.7 million PubMed citations. We downloaded the PREDICTION table, which was published on Dec 31, 2016. This table contains the CUI pairs appearing in each PubMed article.

The three data collections used for evaluating our method were from the CDS track of TREC 2014, 2015 and 2016. The data collections for the CDS track in TREC 2014 and 2015 contain the same set of 744,138 articles from PubMed Central. The data collection for CDS in TREC 2016 has 1.25 million articles. Each article in these three collections contains only title, abstract and article content. The CDS track in each year provided 30 patients' EHRs as the search topics, each of which, as shown in Figure 1, has three elements: description, summary, and either diagnosis in 2014/2015 or notes in 2016. To compare with past studies, we extracted the query from the summary area, which is the most popular approach. Description and notes can be directly processed by our system with minor modification. Among the 30 EHRs, the first ten EHRs require the CDS system to provide articles related to the patient's diagnosis (Q1), the second ten EHRs require articles on what test should the patient receive (Q2), and last ten EHRs need system give the treatment plan articles (Q3).

Table 1: Summary of experiment topics and collections

Dataset	Usage	Collection size	Indexing	Topics
2014 CDS Track	Training	733,138 articles	Indri	30
2015 CDS Track	Testing	733,138 articles	Indri	30
2016 CDS Track	Testing	1.25 million articles	Indri	30
Wikipedia	Knowledge base	5.79 million articles	Indri	-
SemMedDB	Knowledge base	Data extracted from 26.7 million articles	-	-

We used TREC 2014 data for training the parameters in our method, and used TREC 2015 and 2016 data for testing. The statistical tests were performed using Wilcoxon Signed Ranks Test.

Following the TREC CDS track's setting, the evaluation metrics we used include (1) infNDCG, inferred Normalized Discounted Cumulative Gain [15], (2) infAP, inferred averaged precision [15], (3) P@10, precision considering only the top 10 ranked documents, and (4) MAP, mean averaged precision of all topics in the task.

4.2 Baselines

To make comparison, we employed four basic baselines. Two low baselines, called *Baseline_unigram* and *Baseline_MRF*, only rely on the information available in the provided EHRs for generating the queries. The former one only considers the unigrams in EHRs, while the latter one, as above introduced, consider both independent unigrams and the term dependency in medical concept clique. Between these two baselines, as shown in Table 2, *Baseline_MRF* significantly outperforms *Baseline_unigram* on 2014 and 2015 topics (p-value<0.05), but it performs significantly inferior for 2016 topics (p-value<0.05). Maybe MRF failure on 2016 topics is due to the query style changes a lot. In 2016 EHRs, patient's vast disease history information appears, while in past two

years, query is mostly composed by symptoms. History disease might lead the important symptom information less weighted. Also, the target collection size is nearly doubled. The TREC CDS track overview [3, 25] shows the mean infNDCG performance of all participant systems drops from 20.99% in 2015 to 18.59% in 2016. Overall, unigram queries maybe are relatively more stable than queries generated from MRF. Trained on 2014 topic set, parameters on best infNDCG is $\lambda_T = 0.4$, and $\lambda_U = 0.6$, with smoothing parameter being 2000.

The rest two baselines employed pseudo relevance feedback to act as high baselines. Relevance Feedback Model 3 (RM3) is a popular, stable and effective pseudo relevance feedback model [27]. These two higher baselines, *Baseline_unigram_RM3* and *Baseline_MRF_RM3*, provide direction comparison to our proposed query expansion methods. Trained on 2014 topic set, best parameters are extracting most informative 3 words from top 5 documents with smoothing parameter being 500. Again, as shown in Table 2, *Baseline_unigram_RM3* and *Baseline_MRF_RM3* both significantly improve over *Baseline_unigram* and *Baseline _MRF* (p-value<0.05), respectively on TREC CDS 2014 and 2015 topics, which indicates that they are indeed higher baselines. However, their performance improves over the two non-PRF baselines on TREC CDS 2016 data, but the improvement is not significant.

4.3 How well can predicted diseases improve CDS system performance?

In this study, for each topic, we have three predicted diseases: one from Wiki-DP, one from SMDB-DP, and the last one from the fusion of two disease ranking list. From Table 2, we find that all three kinds of diagnosis prediction significantly improve the retrieval performance in 2014, 2015, and 2016 topics (p-value<0.05). Further, Wiki-DP and fused predicted diseases even outperforming the RM3 model in 2014 and 2015.

In the results of 2016 CDS, performance is much lower than 2014 and 2015. As stated above, it might be caused by that the topics in 2016 is quite different from past years with history disease information introduced, and the target collection size is nearly doubled.

4.3.1 The effectiveness of having Wiki-DP

In examining the effectiveness of having our diagnosis prediction algorithm using Wikipedia, we compare its performance with the baselines and some state of the art systems. As shown in Table 2, our two wiki runs (*Unigram_wiki* and *MRF_wiki*) outperform their corresponding low baselines (*Baseline_unigram* and *Baseline_MRF*) as well as their corresponding high baselines (*Baseline_unigram_RM3* and *Baseline_MRF_RM3*) in all three TREC datasets (2014, 2015 and 2016). Statistical tests show that all improvements are significant.

The last three rows in Table 2 show the best performed system among all TREC participants in each year. The performance values presented in Table 2 are based on their published work notes. We can see that, for 2014 and 2015, our two wiki runs achieve much higher performance than these best systems. Even for TREC 2016, our methods are very close to the state-of-art runs. This validate the effectiveness of having Wiki-DP.

Table 2: Performance comparison on CDS task with the state-of-art runs. * means significantly outperforms Baseline_unigram or Baseline_MRF; ** means significantly outperforms Baseline_unigram_RM3 or Baseline_MRF_RM3.

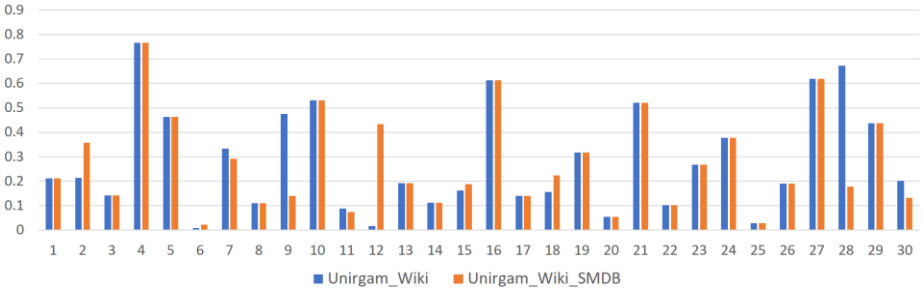
2014 (Training data)	infNDCG	infAP	MAP	P@10
Baseline_unigram	22.02%	5.17%	8.53%	32.00%
Baseline_unigram_RM3	26.45%*	8.46%*	11.96%*	34.33%*
Unigram_wiki	28.44%**	9.41%**	14.07%**	36.00%**
Unigram_SMDB	23.44%*	6.33%*	10.56%*	36.00%*
Unigram_wiki_SMDB	28.31%**	8.15%*	13.15%**	35.00%*
Baseline_MRF	24.11%	5.96%	9.77%	37.00%
Baseline_MRF_RM3	28.27%*	9.18%*	13.27%*	39.00%*
MRF_Wiki	29.31%**	9.39%*	12.87%*	33.00%
MRF_SMDB	25.43%*	6.89%*	11.26%*	39.00%*
MRF_Wiki_SMDB	28.59%**	8.20%*	12.60%*	34.00%
SNUMedinfo [6]	26.74%	-	6.59%	36.33%
2015 (Testing data)	infNDCG	infAP	MAP	P@10
Baseline_unigram	18.24%	3.75%	8.47%	33.33%
Baseline_unigram_RM3	22.88%*	5.91%*	12.23%*	36.67%*
Unigram_wiki	26.67%**	6.28%*	14.30%**	37.67%**
Unigram_SMDB	21.73%*	5.01%*	11.05%*	38.33%**
Unigram_wiki_SMDB	29.32%**	7.49%*	15.68%**	39.33%**
Baseline_MRF	19.90%	3.88%	8.69%	32.00%
Baseline_MRF_RM3	25.34%*	6.38%*	13.43%*	40.00%*
MRF_Wiki	29.67%**	6.98%*	14.77%**	37.67%*
MRF_SMDB	22.11%*	4.87%*	10.79%*	38.33%*
MRF_Wiki_SMDB	29.36%**	7.24%*	15.22%**	36.67%*
WSU-IR [8]	29.39%	8.42%	18.64%	46.67%
2016 (Testing data)	infNDCG	infAP	MAP	P@10
Baseline_unigram	18.40%	1.91%	4.82%	25.33%
Baseline_unigram_RM3	18.45%	2.47%	4.94%	25.33%
Unigram_wiki	22.92%**	2.88%*	6.11%**	31.67%**
Unigram_SMDB	18.43%*	1.98%	4.82%	26.00%
Unigram_wiki_SMDB	22.09%**	2.86%*	5.74%**	31.00%**
Baseline_MRF	17.35%	1.86%	4.50%	21.00%
Baseline_MRF_RM3	17.75%	2.14%	4.25%	24.33%
MRF_Wiki	22%**	2.86%**	5.42%**	31.67%**
MRF_SMDB	17.64%	1.82%	4.51%	23.33%
MRF_Wiki_SMDB	21.9%**	2.92%**	5.54%**	32.00%**
MERCKKGAA [8]	24.93%	3.15%	-	35.00%

4.3.2 The effectiveness of having SMDB-DP

The improvement obtained from using SMDB-DP alone is not as good as wiki-DP, but *Unigram_SMDB* and *MRF_SMDB* still significantly outperforms *Baseline_unigram* and *Baseline_MRF* in 2014, and 2015 (p-value<0.001). However, Table 2 shows RM3

is much better than SMDB-DP (p-value<0.001). This indicates the concept level diagnosis prediction is harder than word level prediction. We think there might be several reasons. First, the medical concepts are not precisely extracted (above mentioned 75% accuracy), while word level retrieval (wiki-DP) do not have this problem. Second, through MetaMap, one disease name might be identified with several concepts, for example, "Hypotension" has two CUIs, *C0020649* and *C3163620*. In this work, we simply use the MetaMap's top recommendation, but maybe other concepts also work or even better. In addition, since SMDB-DP has a very bad performance in 2016, it can be inferred the vast history diseases severely affect the SMDB-DP.

Figure 3: Individual topic performance comparison among Unigram_Wiki and Unigram_Wiki_SMDB on CDS 2014 (infNDCG)



4.3.3 The Combined Effectiveness of Wiki-DP and SMDB-DP

We conducted two runs on top of unigram query and MRF query with the expansion of fused results from the diagnoses from both Wiki-DP and SMDB-DP. As introduced in Section 3.4, such fusion is heavily relying on the Wiki-DP's performance. As shown in Table 2, runs *Unigram_wiki_SMDB* and *MRF_Wiki_SMDB* have quite similar performance with *Unigram_Wiki* and *MRF_Wiki*. Although in 2015 CDS, *Unigram_Wiki_SMDB* significantly outperforms *Unigram_Wiki* on infNDCG (29.32% vs 26.67%), it is probably because the parameters were trained on 2014 CDS. If it were trained on 2015 CDS data, *Unigram_Wiki* can get infNDCG at 29.43%, basically it is the same with *Unigram_Wiki_SMDB*.

Among the 90 topics test in our experiments, the results from the fused diagnosis differ from the Wiki-DP predictions by about 30% (10 different predictions in 2014, 7 in 2015, and 12 in 2016). Figure 3 shows the infNDCG performance of the 30 topics on 2014 CDS data. The fused predictions of three topics further improve the retrieval performance by a large degree, whereas the fused predictions on two topics generated inferior results against the Wiki only approach. This indicates that SMDB-DP can work as a supplementary module for the Wiki-DP, but the fused data is not promised to be always correct.

4.4 How accurate is the predicted diagnosis?

Although the experiments presented in Section 4.3 show the significant improvement contributed by the predicted diagnosis in helping biomedical retrieval. Intrinsically, it

is hard to identify whether or not the predicted diagnosis is correct. Firstly, this is because we do not have ground-truth diagnosis for 70 out of 90 topics. Secondly, even among the 20 topics that 2015 CDS provides the correct diagnosis, it is still hard to judge the correctness of our predicted disease. For example, the TREC provided diagnosis for topic 15 is "Paroxysmal Atrial fibrillation", and our predicated disease is "Atrial fibrillation". This partially matched prediction can improve infNDCG by 0.16, but it is not the same as true diagnosis. Therefore, we define in this paper the usefulness of the predicted disease rather than the correctness, and state that only when a prediction can improve the topic retrieval performance by at least 1.00% on infNDCG, will we state that the prediction is useful.

Table 3: Disease prediction usefulness on CDS topics from 2014 to 2016

Diagnosis prediction methods	2014	2015	2016	Mean
Wiki-DP	70.00%	66.67%	60.00%	65.56%
SMDB-DP	53.33%	43.33%	33.33%	43.33%
Prediction Fusion	63.33%	66.67%	56.67%	62.22%

As shown in Table 3, Wiki-DP generated the highest portion of useful predictions, with mean portion of usefulness prediction to be 65.56%, and give relatively robust performance across the three years' topic sets. Figure 4 further shows the predicated diseases by Wiki-DP. SMDB-DP generated the lowest portion of useful predictions (43.33% mean value), indicating the difficulty of concept level diagnosis prediction.

5 Discussion

Through the experiment, we have demonstrated that our method of utilizing online open knowledge bases for diagnosis prediction to improve the medical literature retrieval can significantly improve the performance, and reach to the comparable level of the start of the art methods. In this section, we want to review the methods in more detail in terms of the places where it fails and the comparison with some existing approaches.

5.1 Further Analysis of Diagnosis Prediction

As shown in the results, most predicted diagnoses made by our methods are correct. Even when the retrieval from one knowledge base fails, the results from the other one often can help to recover to the correct prediction. For example, with the initial query extracted from Figure 1, "Epiglottitis" is ranked 8th in the results obtained through SemMedDB knowledge base, but it is the first in the list from Wikipedia. This helps to make it the correct prediction. Another disease appearing in both results is "Retropharyngeal abscess", which ranks at the 7th in the SemMedDB results, but is at 14th rank in the Wikipedia list. So, drawing evidence from two sources does make our diagnosis prediction methods more robust. Our experiments results show that, in general, the results from Wikipedia are usually more reliable, but a confirmation from the SemMedDB results makes the predictions even more accurate.

Figure 4: Wiki-DP prediction results for 90 topics in CDS task from 2014 to 2016

	2014 CDS Task	2015 CDS Task	2016 CDS Task
1	Coronary artery disease	circulatory shock	Lower gastrointestinal bleeding
2	Middle East respiratory syndrome	Eosinophilic pneumonia	Osteoarthritis
3	Solitary pulmonary nodule	Pulmonary embolism	Sepsis
4	Kawasaki disease	Takotsubo cardiomyopathy	Chronic obstructive pulmonary disease
5	Pulmonary embolism	Rheumatic fever	Pneumonia
6	Traumatic injury wound	Hyperthyroidism	Cholecystitis
7	Bipolar II disorder	Major depressive episode	Acute pancreatitis
8	Multiple sclerosis	Obstructive sleep apnea	Hyperemesis gravidarum
9	Anatomical Abnormality	Trichinosis	Acute respiratory distress syndrome
10	Pseudoaneurysm	Vaginal bleeding	Cirrhosis
11	Compartment syndrome	Hypothyroidism	Angina pectoris
12	Anorexia nervosa	Meningitis	Head injury
13	Pulmonary embolism	Epiglottitis	Pressure ulcer
14	Traumatic brain injury	Megaloblastic anemia	Dyspnea
15	Ovarian cancer	Atrial fibrillation	Lung cancer
16	Rabies	Asthma	Apraxia of speech
17	Lemierre syndrome	Cervical cancer	Lutembacher syndrome
18	Azotaemia	Congestive heart failure	Pancreatitis
19	Esophageal dysphagia	Chronic obstructive pulmonary disease	Heart failure
20	sternum fracture	General Paralysis	Gallstone
21	Systemic lupus erythematosus	Traveller diarrhea	Pneumocystis pneumonia
22	Appendicitis	Bronchopulmonary Aspergillosis	Cardiology
23	Chronic obstructive pulmonary disease	Dengue fever	Colorectal cancer
24	Blunt splenic trauma	Pneumonia	Bowel obstruction
25	Traumatic brain injury	Langerhans-cell histiocytosis	atrial fibrillation
26	Malaria	Ectopic pregnancy	atrial fibrillation
27	Familial adenomatous polyposis	Iron-deficiency anemia	Headache
28	Prolactinoma	Lyme disease	Gastrointestinal bleeding
29	Osteoporosis	Kawasaki disease	Idiopathic pulmonary fibrosis
30	Peripheral artery disease	Rotator cuff	atrial fibrillation

However, our diagnosis prediction method does make errors. There are two types of causes to the errors in Wiki-DP. The first one is related to the insufficient information in Wikipedia data. Typical scenarios include that the correct diseases do not have sufficient content in their Wiki pages, or the terms in the query to search in Wikipedia fail to distinguish the correct diseases from irrelevant diseases. For example, Topic 25 in 2015 CDS should have a correct associated disease “Osteomyelitis”, but the Wiki page of “Osteomyelitis” contains no symptom information at all. In the meantime, the Wiki page of “Langerhans-cell histiocytosis” contains some symptoms mentioned in the given EHRs. This causes the prediction method to wrongly select “Langerhans-cell histiocytosis” rather than “Osteomyelitis”.

The second type of cause is related to the limitation of current retrieval mechanism, particularly the handling of negation. For example, the EHRs associated with Topic 12 in 2014 CDS show that the patient has a symptom of “weight gain”, and the Wiki page of “Anorexia nervosa” talks about patients wanting to “prevent weight gain”, “fear weight gain” or “avoid weight gain”. Because of lacking processing negation in collection text, our method wrongly ranks “Anorexia nervosa” as the most plausible disease for the patient.

To combat these problems, we need to enlarge our external knowledge bases to include more publicly available online resources. For example, there are published medical concept relationship dataset, such as MayoClinic[7] or WebMD[8]. This will be a future work to explore.

In terms of SMDB-DP, there are three causes to the errors. First, the medical concepts in SemMedDB corpus are not precisely extracted (75% accuracy [22]), while Wiki-DP does not have this procedure and information loss. Second, through MetaMap, one disease name might be identified with several concepts. For example, MetaMap will map "Hypotension" into two medical concepts, as shown in Figure 5. In this work, we simply use the top identification from MetaMap, but other concepts should also be considered. Third, there are many medical concepts extracted from query not existing in SemMedDB. For example, "weight loss" is a symptom appearing in 4 topics, but SemMedDB do not have such a symptom in extracted concepts, making the information need not fully provided to the SMDB-DP module. In conclusion, concept level diagnosis prediction is harder and more complicated than the word level.

Figure 5: MetaMap maps "Hypotension" to the two medical concepts

```
>>>>> Mappings
Meta Mapping (1000):
  1000   C0020649:HYPOTENSION (Hypotension) [Finding]
Meta Mapping (1000):
  1000   C3163620:Hypotension (Hypotension Adverse Event) [Finding]
<<<<< Mappings
```

5.2 Further Comparison to the State-of-the-Art Biomedical Retrieval Systems

Although our methods only reach to the comparable performance with the state-of-art algorithms, our retrieval model is much simpler. For example, the best performed system in the 2014 CDS task was developed by Choi et al. [6]. They used pseudo relevance feedback (PRF) to obtain most frequently mentioned MESH terms from the top ranked articles, and expand their queries with such terms. This increased their algorithm's performance from 19.21% to 22.24% on infNDCG. Then their method used classification method to re-rank the results to achieve their best results 26.74%. In contrast, our method only performs query expansion with predicted diagnosis, which increases infNDCG from 21.88% to 28.44%. We believe that our approach is simpler in retrieval model. At the same time, we regard the CDS task as a concept based information retrieval task, in which our query can recognized the concept associated with a disease but is not limited by the text expression in the document.

The best run in 2015 CDS task was submitted by Balaneshin-kordan et al. [8]. They explored a lot on medical concept detection and selection, and expanded their queries with the most important unigrams in PRF documents. They attributed their success to the MRF model and Parameterized Query Expansion (PQE). However, we cannot tell them apart since they did not publish the intermediate performance.

---

[7] http://www.mayoclinic.org/

[8] http://www.webmd.com/

The best run in 2016 CDS task was submitted by Gurulingappa et al. [24]. They first expand the original query with extracted UMLS medical concepts from EHRs, and with the most important words in PRF. Then they measured the document similarity based on word embeddings, and combine these features with a learning-to-rank model, which improved the infNDCG from 22.61% to 24.93%. This method indicates that the word embedding can help search the relevant documents that use different words but keep the same relevant information.

In summary, our method takes a different route to the existing state of the art methods. It is possible that our method can be combined with these state of the art approaches or even more advanced retrieval methods. Under such situation, the retrieval improvement can be even larger.

6 Conclusion

In this paper, we target to enhance the current CDS systems with public knowledge bases, Wikipedia and SemMedDB. To be specific, a word level and a concept level diagnosis prediction methods are proposed to automatically find the disease of the patient, which are used to perform query expansion in medical text retrieval. This idea has been proven to be effective by the significantly improved retrieval performance using our methods through the validation on TREC CDS track data of 2014, 2015 and 2016. Our disease prediction usefulness can reach 65.56%. In the future, we will incorporate the word embedding techniques to enhance the diagnosis prediction methods.

Reference

1. Ely, John W., et al. "A taxonomy of generic clinical questions: classification study." Bmj 321.7258 (2000): 429-432.
2. Simpson, Matthew S., Ellen M. Voorhees, and William Hersh. Overview of the trec 2014 clinical decision support track. LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD, 2014.
3. Roberts, Kirk, et al. "Overview of the TREC 2015 Clinical Decision Support Track."
4. Zighelnic, Liron, and Oren Kurland. "Query-drift prevention for robust query expansion." Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008.
5. Carpineto, Claudio, and Giovanni Romano. "A survey of automatic query expansion in information retrieval." ACM Computing Surveys (CSUR) 44.1 (2012): 1.
6. Choi, Sungbin, and Jinwook Choi. SNUMedinfo at TREC CDS track 2014: Medical case-based retrieval task. SEOUL NATIONAL UNIV (REPUBLIC OF KOREA), 2014.
7. Oh, Heung-Seon, and Yuchul Jung. "Cluster-based query expansion using external collections in medical information retrieval." Journal of biomedical informatics 58 (2015): 70-79.
8. Balaneshin-kordan, Saeid, Alexander Kotov, and Railan Xisto. "WSU-IR at TREC 2015 Clinical Decision Support Track: Joint Weighting of Explicit and Latent Medical Query Concepts from Diverse Sources." Proceedings of the 2015 Text Retrieval Conference. 2015.
9. Xie, Zhongda, Yunqing Xia, and Qiang Zhou. "Incorporating Semantic Knowledge with MRF Term Dependency Model in Medical Document Retrieval." National CCF Conference on Natural Language Processing and Chinese Computing. Springer International Publishing, 2015.

10. Song, Yang, et al. "ECNU at 2015 CDS Track: Two Re-ranking Methods in Medical Information Retrieval." Proceedings of the 2015 Text Retrieval Conference. 2015.
11. Esfandiari, Nura, et al. "Knowledge discovery in medicine: Current issue and future trend." Expert Systems with Applications 41.9 (2014): 4434-4463.
12. Yeh, Duen-Yian, Ching-Hsue Cheng, and Yen-Wen Chen. "A predictive model for cerebrovascular disease using data mining." Expert Systems with Applications 38.7 (2011): 8970-8977.
13. Bodenreider, Olivier. "The unified medical language system (UMLS): integrating biomedical terminology." Nucleic acids research 32.suppl 1 (2004): D267-D270.
14. Aronson, Alan R. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001.
15. Voorhees, Ellen M. "The effect of sampling strategy on inferred measures."Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014.
16. Strohman, Trevor, et al. "Indri: A language model-based search engine for complex queries." Proceedings of the International Conference on Intelligent Analysis. Vol. 2. No. 6. 2005.
17. Briones, Natalia, and Valentin Dinu. "Data mining of high density genomic variant data for prediction of Alzheimer's disease risk." BMC medical genetics 13.1 (2012): 7.
18. Isola, Rahul, Rebeck Carvalho, and Amiya Kumar Tripathy. "Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and $ k $-NN." IEEE transactions on information technology in biomedicine 16.6 (2012): 1287-1295.
19. Gomathi., P., Nithya, NS. "Medical Disease Diagnosis Using Structuring Text." International Journal of Computer Science & Engineering Technology 1.5: 591-594.
20. Liu, Chaochun, et al. "Augmented LSTM Framework to Construct Medical Self-diagnosis Android." Data Mining (ICDM), 2016 IEEE 16th International Conference on. IEEE, 2016.
21. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 classification of cancers from free-text death certificates. International journal of medical informatics. 2015 Nov 30;84(11):956-65.
22. Kilicoglu, Halil, et al. "SemMedDB: a PubMed-scale repository of biomedical semantic predications." Bioinformatics 28.23 (2012): 3158-3160.
23. Metzler, Donald, and W. Bruce Croft. "A Markov random field model for term dependencies." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.
24. Gurulingappa, Harsha, et al. "Semi-Supervised Information Retrieval System for Clinical Decision Support." TREC. 2016.
25. Roberts, Kirk, et al. "Overview of the TREC 2016 Clinical Decision Support Track." TREC. 2016.
26. Nie, Liqiang, et al. "Disease inference from health-related questions via sparse deep learning." IEEE Transactions on Knowledge and Data Engineering 27.8 (2015): 2107-2119.
27. Lv, Yuanhua, and ChengXiang Zhai. "A comparative study of methods for estimating query language models with pseudo feedback." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.