KISTI at TREC 2014 Clinical Decision Support Track: Concept-based Document Re-ranking to Biomedical Information Retrieval

Heung-Seon Oh and Yuchul Jung

Korea Institute of Science and Technology Information
{ohs, jyc77}@kisti.re.kr

Abstract

With fast development of medical information systems and software, clinical decision support (CDS) systems continue to develop new methods to deal with diverse information coming from heterogeneous sources such as a large volume of electronic medical records (EMRs), patient genomic data, existing genomic pharmaceutical databases, curated disease-specific databases, peer-reviewed research, etc. As an avenue towards advanced clinical decision-making, TREC CDS track focuses on developing new techniques to find medical cases that are useful for patient care from biomedical literature. Meanwhile, given the volume of the existing literature, and the diversity in biomedical field, finding & delivering relevant medical cases for a particular clinical need is a non-trivial task. Moreover, understanding three kinds of different topics (i.e. diagnosis, test, and treatment) and retrieving appropriate biomedical research articles are quite challenging. To address these problems, we propose concept-based document re-ranking approaches to clinical documents. We basically use pseudo relevance feedback for query expansion to retrieve initial relevant documents. In addition, we considered two different concept-based re-ranking approaches which utilize popular external biomedical knowledge resources (i.e. Wikipedia and UMLS) for improving biomedical information retrieval. Our concept-based re-ranking approaches are to bridge the gaps between queries and biomedical research articles in semantic level.

1 Introduction

TREC Clinical Decision Support Track (CDS) aims to investigate techniques for linking medical cases to information that are relevant for patient care from published biomedical literature. The published biomedical literature which can be searched through PubMed is a trustable, comprehensive source for exploratory analysis and clinical decision-making support because it maintains a number of biomedical research articles including various information such as patient demographics, laboratory test results, radiology reports, clinical demonstration, medicine treatment, etc. The task of CDS is to find biomedical research articles published in PubMed Central (PMC) with a given query which requires expertise to make a decision for treating a patient. It provides a PMC collection with 733,138 XML articles and 30 test queries classified into one of three classes: diagnosis, test, and treatment.

In our participation to CDS, we propose concept-based document re-ranking approaches to retrieval biomedical documents. First, a set of documents are obtained from an initial search. Then, a first stage of re-ranking is performed using pseudo relevance feedback by expanding a query. At the second stage, we devised concept-based re-ranking approaches that utilize two different external biomedical knowledge resources (i.e. Wikipedia and UMLS) for more accurate biomedical information retrieval. Our concept-based re-ranking approaches are to show the potentials of using external knowledge resources in aspects of understanding the input queries and the retrieved biomedical research articles in semantic level based on the concepts of knowledge resources.

The rest of this paper is organized as follows. Section 2 explains our proposed approaches in details. Section 3 presents experimental results among different avenues towards effective CDS. In Section 4, we summarize our entire work and introduce future search direction.

2 Method

Our method is to re-rank documents obtained from an initial search with two stages. In the first stage, pseudo relevance feedback (PRF) is applied to obtain accurate ranking by expanding a query based on initial search results. In the second stage, concept-based ranking with two different medical resources are performed. Next subsection describes our method in detail.

2.1 Pseudo Relevance Feedback

For a given query Q, a set of documents, $D_{init} = \{D_1, D_2, \ldots, D_k\}$, are retrieved from a document collection COL using a search engine. Lucene[1] is employed with query-likelihood method using Dirichlet smoothing. Then, ranking is performed on D_{init}.

In this stage, KL-divergence method is used to compute a similarity score between a query and a document [10,13]:

$$\begin{aligned} score(Q,D) &= exp\left(-KL\left(\theta_Q||\theta_D\right)\right) \\ &= exp\left(-\sum_{w} p\left(w\middle|\theta_Q\right) log\frac{p(w|\theta_Q)}{p(w|\theta_D)}\right) \end{aligned} \quad (1)$$

where θ_Q and θ_D are query and document language models, respectively.

In general, a query model is estimated by maximum likelihood estimate (MLE) as below:

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

$$p(w|\theta_Q) = \frac{c(w,Q)}{|Q|} \tag{2}$$

where $c(w,Q)$ is a count of a word w in a query Q and $|Q|$ is the number of words in Q.

To avoid zero probabilities and improve retrieval performance, a document model is estimated using Dirichlet smoothing [15]:

$$p(w|\theta_D) = \frac{c(w,D) + \mu \cdot p(w|COL)}{\sum_t c(t,D) + \mu} \tag{3}$$

where $c(w,D)$ is a count of a word w in a document D, $p(w|COL)$ is a probability of a word w in a collection COL, and μ is the Dirichlet prior parameter.

PRF is a popular way of expanding a query. It is assumed that top-ranked documents $F = \{D_1, D_2, \dots, D_{|F|}\}$ in initial search results relevant to a given query and terms in F are useful to modify a query for a better representation. Relevance model (RM) is to estimate a multinomial distribution $p(w|Q)$ that is the likelihood of a term w given a query Q. The first version of relevance model (RM1) is defined as follows:

$$\begin{aligned} p_{RM1}(w|Q) &= \sum_{D \in F} p(w|\theta_D) \cdot p(\theta_D|Q) \\ &= \sum_{D \in F} p(w|\theta_D) \cdot \frac{p(Q|\theta_D) \cdot p(\theta_D)}{p(Q)} \\ &\propto \sum_{D \in F} p(w|\theta_D) \cdot p(\theta_D) \cdot p(Q|\theta_D) \end{aligned} \tag{4}$$

RM1 is composed with three components: document prior $p(\theta_D)$, document weight $p(Q|\theta_D)$, and term weight in a document $p(w|\theta_D)$. In general, $p(\theta_D)$ is assumed to be a uniform distribution without the knowledge of a document D. $p(Q|\theta_D) = \prod_{w \in Q} p(w|\theta_D)^{c(w,Q)}$indicates the query-likelihood score. $p(w|\theta_D)$ can be estimated using various smoothing methods such as Dirichlet-smoothing. Various strategies are applicable to estimate these components.

To improve retrieval performance, a new query model can be estimated by combing a relevance model and an original query model. RM3 [1] is a variant of a relevance model to estimate a new query models with RM1:

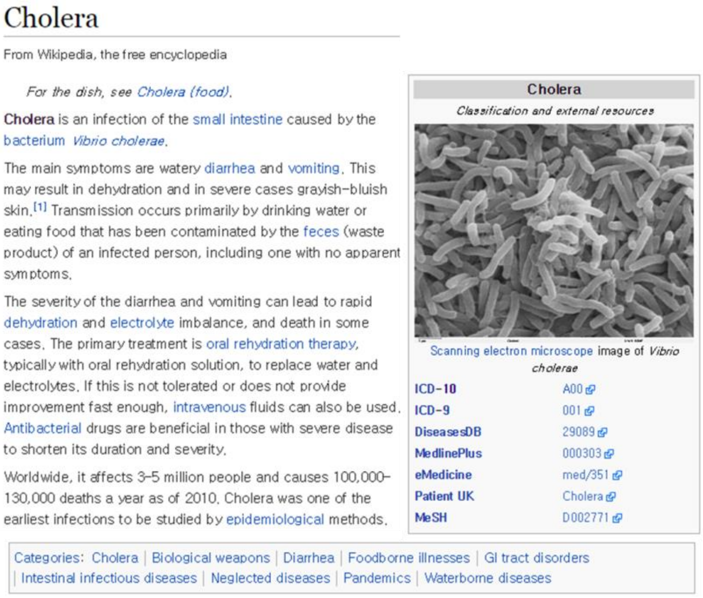
$$p(w|\theta_Q') = (1-\beta) \cdot p(w|\theta_Q) + \beta \cdot p_{RM1}(w|Q) \tag{5}$$

where β is a control parameter between the original query model and the feedback model.

2.2 Concept-based Ranking

The key idea of concept-based IR is to find documents by representing them with concepts rather than words. It is a popular solution to deal with synonymy and polysemy problems occur in IR tasks based on bag-of-words representation [6]. In general, Wikipedia is utilized as a resource of concepts because it has millions of concepts in the world while UMLS having medical-specific concepts which include SNOMED-CT and MeSH is dominantly used in biomedical IR tasks. We utilized two resources in different ways for concept-based IR.

Concept mapping with Wikipedia. Wikipedia is utilized as a concept resource. We assumed that a subset of concepts relevant to medical domain in Wikipedia are useful to CDS. To retain useful medical concepts, those belonging to International Classification Diseases (ICD)-10 [2] are selected for concept mapping. ICD-10 is a hierarchical classification scheme of diseases and other health problems defined by World Health Organization (WHO). Thus, coverage and granularity of concepts in ICD-10 are assumed to be suitable to CDS. Unfortunately, all concepts of ICD-10 don't exist in Wikipedia. From more than 14,400 ICD-10 concepts, 7,162 concepts are retained since they have an article in Wikipedia. Fig. 1 shows a Wikipedia article for ICD-10 concept Cholera.



Cholera

From Wikipedia, the free encyclopedia

For the dish, see Cholera (food).

Cholera is an infection of the small intestine caused by the bacterium *Vibrio cholerae*.

The main symptoms are watery diarrhea and vomiting. This may result in dehydration and in severe cases grayish-bluish skin.[1] Transmission occurs primarily by drinking water or eating food that has been contaminated by the feces (waste product) of an infected person, including one with no apparent symptoms.

The severity of the diarrhea and vomiting can lead to rapid dehydration and electrolyte imbalance, and death in some cases. The primary treatment is oral rehydration therapy, typically with oral rehydration solution, to replace water and electrolytes. If this is not tolerated or does not provide improvement fast enough, intravenous fluids can also be used. Antibacterial drugs are beneficial in those with severe disease to shorten its duration and severity.

Worldwide, it affects 3–5 million people and causes 100,000–130,000 deaths a year as of 2010. Cholera was one of the earliest infections to be studied by epidemiological methods.

Cholera

Classification and external resources

Scanning electron microscope image of *Vibrio cholerae*

ICD-10	A00
ICD-9	001
DiseasesDB	29089
MedlinePlus	000303
eMedicine	med/351
Patient UK	Cholera
MeSH	D002771

Categories: Cholera | Biological weapons | Diarrhea | Foodborne illnesses | GI tract disorders | Intestinal infectious diseases | Neglected diseases | Pandemics | Waterborne diseases

Fig. 1. An example Wikipedia article of ICD-10 concept *Cholera*[3]

[2] http://apps.who.int/classifications/icd10/browse/2010/en

[3] http://en.wikipedia.org/wiki/Cholera

Based on the selected concepts, ranking is performed by scoring documents with concept mapping method introduced in [7]. The method is adaption of concept mapping to document clustering with Wikipedia. In our case, we do ranking than clustering with documents. A document is represented by a word vector. Words are stemmed and lower-cased after stop-words are removed using a stop-words list[4]. The words in the word vector are mapped to ICD-10 concepts. In addition, a category vector can be derived from a concept vector by similar mapping because an article corresponding to a concept has a set of categories at the end of an article as shown in Fig. 1. This is a decomposition of a document-category matrix into three components, document-word, word-concept, and concept-category matrices, shown in Fig. 2. Entries are filled with standard TF-IDF values in document-word matrix while they are filled with modified versions of TF-IDF values for concepts and categories in others.

$$\begin{array}{c|ccc} & cat_1 & cat_2 & cat_3 \\ \hline D_1 & & & \\ D_2 & & & \\ D_3 & & & \end{array} \rightarrow \begin{array}{c|ccc} & w_1 & w_2 & w_3 \\ \hline D_1 & & & \\ D_2 & & & \\ D_3 & & & \end{array} \times \begin{array}{c|ccc} & c_1 & c_2 & c_3 \\ \hline w_1 & & & \\ w_2 & & & \\ w_3 & & & \end{array} \times \begin{array}{c|ccc} & cat_1 & cat_2 & cat_3 \\ \hline c_1 & & & \\ c_2 & & & \\ c_3 & & & \end{array}$$

Fig. 2. Decomposition of document-category matrix

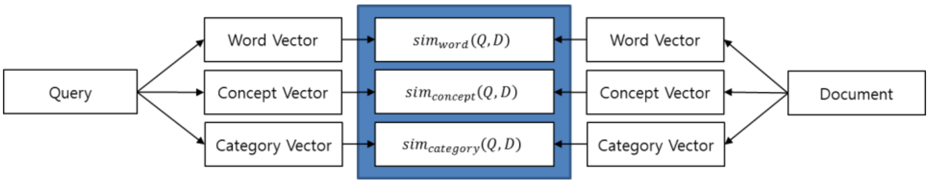


Fig. 3. Final score computation by combining three different scores

Therefore, as shown in Fig. 3, we can compute three scores based on different representations of a document and a query using cosine similarity function.

A final score is computed by a linear combination of three scores:

$$score(Q,D) = \alpha_1 \cdot sim_{word}(Q,D) + \alpha_2 \cdot sim_{concept}(Q,D) + \alpha_3 \cdot sim_{category}(Q,D) \qquad (6)$$

where $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$ and $\alpha_1, +\alpha_2 + \alpha_3 = 1$

In this paper, we set them uniformly as $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$.

Concept mapping with UMLS. Unified Medical Language System (UMLS) [4] is utilized as a domain-specific concept resource. It contains about 900,000 biomedical concepts integrating various resources such as the NCBI taxonomy, Gene Ontology,

[4] http://mallet.cs.umass.edu

Medical Subject Headings (MeSH), OMIM and the Digital Anatomist Symbolic Knowledge Base. In addition, it also contains the mapping between about 900,000 concepts and over 2 million. Due to the large volume of UMLS, it is often utilized for concept-based IR in biomedical domain [8,12]. We employ MetaMap [2] to identify UMLS concepts from texts. One characteristic of using MetaMap is that we can take into consideration the negation of concepts. Dealing with negations is an important issue in other research [3,9,11,14]. We handle the negations of concepts in documents with respect to those in a query while negated concepts are penalized in most of the previous work.

Consider a case with a query Q and two candidate documents D_1 and D_2. Let's suppose that a concept C is contained in Q and the two documents D_1 and D_2 having the same number of occurrences of C. If C in D_1 is mostly negated while it is mostly affirmed in D_2, it is natural to say that the document D_2 should be ranked higher than the document D_1 in search results. On the other hand, if C appears in a negated form in Q, D_1 should be ranked higher than D_2. Based on such idea, we propose a strategy for handling the negations of concepts as follows:

1. If a concept negated in a query appears in a document with affirmation, decrease the score of the document with respect to the query.
2. If a concept negated in a query appears in a document with negations, increase the score of the document with respect to the query
3. Take into account the number of times where a term is affirmed or negated in a document.

Negations are identified using NegEx [5] which is embedded in MetaMap.

In order to highlight the effect of the proposed strategy, we selected six UMLS semantic types that are often negated in the test queries and used only the UMLS concepts belonging to those types for document re-ranking. Table 1 shows the selected semantic types. We can see that the selected semantic types do not include those related to qualification such as *qualitative concepts*, *spatial concepts*, *body location or region*. Although MetaMap often proposes groups of concepts for a given phrase[5], the characteristics of the selected semantic types allow us to focus on individual concepts rather than groups of concepts.

Table 1. Selected UMLS Semantic Type

Semantic Type	Abbreviation	ID
Disease or Syndrome	dsyn	T047
Finding	fndg	T033
Sign or Symptom	sosy	T184
Pathologic Function	patf	T046

[5] For instance, for the phrase "mild dyspnea", MetaMap proposes a concept group that consists of two concepts; 1) concept 'mild' of 'Qualitative Concept' semantic type and 2) concept 'dyspnea' of 'Sign or Symptom' sematic type.

Injury or Poisoning	inpo	T037
Anatomical Abnormality	anab	T190

Given a document D, a concept vector $CV_D = \{v_1, v_2, \dots, v_n\}$ is constructed where $v_i = \sum_{p \in D} \sum_{j=1}^{k} (Neg_{i,j,p} Conf_{i,j,p}) / k$. Here, p represents a phrase that conveys a biomedical concept, and k represents the number of candidates (Meta Mappings) proposed for p. $Conf_{i,j,p}$ is the confidence score for the ith concept in jth candidate for p. We merge all the candidates into a single normalized version rather than selecting the most probable one among the candidates. We assumed that MetaMap would produce the same candidates when it is given the same phrase in similar contexts. $Neg_{i,j,p}$ is a term to handle the negations as proposed above. The value is -1 if the ith concept in jth candidate for p is identified as negated, and 1 if affirmed. Concept vector for Q is constructed in the same way. Then, cosine similarity between two concept vectors is computed. A final score is a combination of scores from PRF and the cosine similarity:

$$score(Q, D) = score_{PRF}(Q, D) + \alpha \cdot sim(Q, D) \quad (7)$$

where $\alpha > 0$ is a weight of the similarity from concept mapping

3 Results

Table 2 shows the descriptions of the submitted runs for evaluation. Run1 is obtained using language model with Dirichlet smoothing implemented Lucene. Then, re-ranking with PRF is performed on the initial search results in Run2. Run3, Run4, and Run5 are the results of concept-mapping with Wikipedia and UMLS. In all runs, 1,000 documents for each query are retrieved and re-ranked. For PRF with RM, the numbers of feedback documents and words are set to 10 and 100, respectively. Mixture weights for Dirichlet smoothing (μ) and RM (β), are set to 0.1 and 1,500, respectively.

Table 2. Descriptions of submitted runs for evaluation

ID	Description
Run1	Initial search
Run2	Initial search + pseudo relevance feedback
Run3	Initial search + pseudo relevance feedback + Wikipedia
Run4	Initial search + pseudo relevance feedback + MetaMap ($\alpha = 1$)
Run5	Initial search + pseudo relevance feedback + MetaMap ($\alpha = 2$)

Table 3 shows the performance summary for five runs. We can see that performances of Run2, Run4, and Run5 are improved against Run1 while it is degraded in Run3. We think that the degradation in Run3 comes from improper concept mapping to ICD-10 in Wikipedia. Our restriction of ICD-10 may result in insufficient coverage of concepts (about 7,000 concepts). From Run4 and Run5, concept-mapping to UMLS improves performance. However, they are not high as we expected. We think that these little improvements show the limitation of re-ranking on the initial search. According to our investigation of the initial search, the number of relevant documents is relatively low in the initial search results by comparing the all documents judged as relevant. Thus, it is necessary to perform re-ranking after initial search with query expansion to contain many relevant documents.

Table 3. Summary of evaluation results

	Run1	**Run2**	**Run3**	**Run4**	**Run5**
map	0.1054	0.1085	0.0933	0.1086	0.1086
R-prec	0.1665	0.1667	0.1443	0.1666	0.1659
P10	0.2933	0.2933	0.2200	0.2933	0.2933
infAP	0.0462	0.0491	0.0424	0.0492	0.0492
infNDCG	0.1911	0.193	0.1759	0.1946	0.1938

4 Conclusion

For TREC Clinical Decision Support track, we proposed two different concept-based re-ranking approaches which utilize Wikipedia and UMLS as a concept resource. We observed small performance improvements from the concept-based re-ranking by using UMLS (i.e., MetaMap). However, in order to achieve higher performances, a number of issues remained unresolved should be tackled further. As our future work, we plan to develop more effective way to utilizing biomedical knowledge resources and sophisticated negation handing strategy towards advanced concept-based ranking.

5 References

1. Abdul-Jaleel, N., Allan, J., Croft, W., Diaz, F., and Larkey, L. UMass at TREC 2004: Novelty and HARD. *Proceedings of Text REtrieval Conference (TREC)*, (2004).
2. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of AMIA Symposium*, (2001), 17–21.
3. Auerbuch, M., Karson, T.H., Ben-Ami, B., Maimon, O., and Rokach, L. Context-sensitive medical information retrieval. *Studies in health technology and informatics 107*, (2004), 282–6.
4. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research 32*, Database issue (2004), 267–270.

5. Chapman, W.W., Hillert, D., Velupillai, S., et al. Extending the NegEx lexicon for multiple languages. *Studies in health technology and informatics 192*, (2013), 677–81.
6. Egozi, O., Markovitch, S., and Gabrilovich, E. Concept-Based Information Retrieval Using Explicit Semantic Analysis. *ACM Transactions on Information Systems 29*, 2 (2011), 1–34.
7. Hu, X., Zhang, X., Lu, C., Park, E.K., and Zhou, X. Exploiting Wikipedia as external knowledge for document clustering. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, ACM Press (2009), 389–396.
8. Koopman, B., Bruza, P., Sitbon, L., and Lawley, M. AEHRC & QUT at TREC 2011 Medical Track: a concept-based information retrieval approach. *Proceedings of Text REtrieval Conference (TREC)*, (2011).
9. Koopman, B. and Zuccon, G. Understanding negation and family history to improve clinical information retrieval. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*, ACM Press (2014), 971–974.
10. Kurland, O. and Lee, L. PageRank without hyperlinks: Structural re-ranking using links induced by language models. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, (2006), 306–313.
11. Limsopatham, N., Macdonald, C., McCreadie, R., and Ounis, I. Exploiting term dependence while handling negation in medical search. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, ACM Press (2012), 1065–1066.
12. Martinez, D., Otegi, A., Soroa, A., and Agirre, E. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. *Journal of biomedical informatics*, (2014).
13. Oh, H.-S. and Myaeng, S.-H. Utilizing global and path information with language modelling for hierarchical text classification. *Journal of Information Science 40*, 2 (2014), 127–145.
14. Voorhees, E.M. and Hersh, W. Overview of the TREC 2012 Medical Records. *Proceedings of Text REtrieval Conference (TREC)*, (2012).
15. Zhai, C. and Lafferty, J. A study of smoothing methods for language models applied to Ad Hoc information retrieval. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, ACM Press (2001), 334–342.