KISTI at CLEF eHealth 2015 Task 2

Heung-Seon Oh, Yuchul Jung, and Kwang-Young Kim

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

Abstract. Laypeople (e.g., patients and their caregivers) usually use queries which describe a sign, symptom or condition to obtain relevant medical information on the Web. They can fail to find useful information for diagnosing or understanding their health conditions because the search results delivered by existing medical search engines do not fit the information needs of users. To deliver useful medical information, we attempted to combine multiple ranking methods, explicit semantic analysis (ESA), a cluster-based external expansion model (CBEEM), and concept-based document centrality (CBDC), using external medical resources to improve retrieval performance. As a first step, initial documents are searched using a baseline method. Based on the initial documents, ranking methods are selectively applied. Our experiments with combinations of ranking methods aim to find the best means of computing accurate similarity scores using different external medical resources. The best performance was obtained when the CBEEM and the CBDC were used together.

Keywords: medical information retrieval, external expansion model, concept-based retrieval

1 Introduction

The general public searches the Web to acquire medical information to diagnose their symptoms and find related health information. Unfortunately, searchers such as laypeople without medical knowledge can fail to find the necessary information in a search query because they are often not only unfamiliar with medical terminology but also uncertain about their exact questions. Tackling queries for laypeople has been a challenging issue with regard to medical information retrieval (IR) because existing Web search engines often fail to deliver satisfactory search results because the required information is not properly understood. To mitigate the difficulties of laypeople (e.g., patients and their relatives), Conference and Labs of the Evaluation Forum (CLEF) launched the eHealth Evaluation Lab [4]. Specifically, Task 2 of CLEF 2015 eHealth [10] explores circumlocutory queries consisting of the signs and symptoms of a medical condition.

As a participant in task 2, this paper introduces a re-ranking framework which attempts to combine selectively different ranking components, such as explicit semantic analysis (ESA), a cluster-based external expansion model (CBEEM), and concept-based document centrality (CBDC). The main goal of our framework is an accurate

estimation of the similarity score by combining different ranking methods using external medical resources.

Within our re-ranking framework, a query-likelihood method with Dirichlet smoothing as a baseline was utilized to obtain the initial document set. D_{init} is re-ranked with the help of ranking components using external medical resources, two biomedical collections (i.e., TREC CDS [11] and OHSUMED [5]) and ICD-10 [1]extracted from Wikipedia. In our experiments, we designed eight runs which combine more than one re-ranking components, except run 1, which represents the baseline. Among the eight runs, the best performance was observed in runs 6 and 8, when the CBEEM and the CBDC were combined. The best performances, in runs 6 and 8, were 0.3864 (P@10) and 0.3464 (NDCG@10).

The rest of this paper is organized as follows. Section 2 presents our ranking framework in detail. The experimental results are described in Section 3. Section 4 concludes with a short summary.

2 Method

2.1 Re-ranking framework

The key idea of our method is to devise a re-ranking framework which estimates an accurate similarity score between a query and a document using external medical resources. To do this, we build a pool of re-ranking components with external resources. Figure 1 shows an overview of our re-ranking framework. For a given query *Q*, a set of documents, $D_{init} = \{D_1, D_2, \dots, D_k\}$, is retrieved from collection *C* using a search engine. In this paper, a query-likelihood method with Dirichlet smoothing (QLD) [14] is utilized to obtain D_{init}. Then, we focus on re-ranking D_{init} using external resources to improve the performance. Specifically, two biomedical collections, TREC CDS and OHSUMED, and ICD-10 as extracted from Wikipedia were used as external resources. Based on D_{init}, re-ranking is performed through a series of ranking components in the pool.

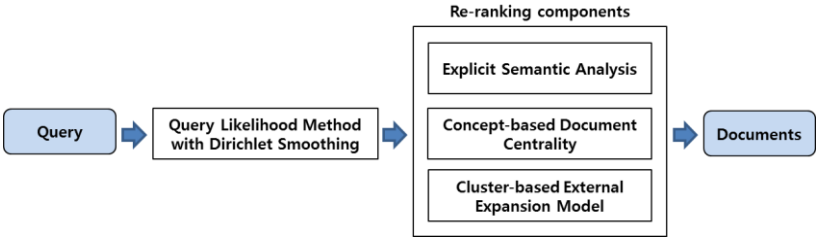

Fig. 1. Overview of the re-ranking framework

[1] http://apps.who.int/classifications/icd10/browse/2015/en

2.2 Basic Foundation

Before explaining the details of the three different re-ranking components, we introduce the basic foundation of the language modeling framework for IR to provide a deeper explanation. In language modeling for IR, the KL-divergence method (KLD) is a popular scoring function to compute similarity scores by estimating unigram language models for a query *Q* and a document *D* [6, 7, 9]:

$$score_{KLD}(Q, D) = \exp\left(-KL(\theta_Q||\theta_D)\right) = \exp\left(-\sum_w p(w|\theta_Q)\, log\, \frac{p(w|\theta_Q)}{p(w|\theta_D)}\right) \quad (1)$$

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

KLD has been attractive because effective pseudo-relevance feedback methods have been proposed to estimate more accurate query language models in an effort to improve performance. The research questions are how to estimate accurate query and document language models to improve the retrieval performance.

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

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

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

A document model is estimated using Dirichlet smoothing to avoid zero probabilities and to improve the retrieval performance through an accurate estimation [14]:

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

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

Query expansion aims to reveal information needs not expressed in *Q* by adding more useful words. Pseudo-relevance feedback (PRF) is a popular query expansion approach to update a query. Updating a query with PRF assumes that the top-ranked documents $F = \{D_1, D_2, \ldots, D_{|F|}\}$ in the initial search results relevant to a given query and the words in *F* are useful to modify a query for a better representation. A relevance model (RM) serves to estimate a multinomial distribution $p(w|q)$, which is the likelihood of a word *w* in query *Q*. The first version of the relevance model (RM1) is defined as follows:

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

RM1 is composed of three components: the document prior $p(\theta_D)$, the document weight $p(Q|\theta_D)$, and the term weight in a document $p(w|\theta_D)$. In general, $p(\theta_D)$ is assumed to have a uniform distribution without knowledge of 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 the retrieval performance, a new query model can be estimated by combing the relevance model and the original query model. RM3 [1] is a variant of a relevance model which is used here to estimate a new query model with RM1,

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

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

2.3 Re-ranking Components

Component 1 - Explicit Semantic Analysis: Concept-based IR using an explicit semantic analysis (ESA) [3] is a well-known approach used to deal with a vocabulary mismatch problem between a query and a document, where the words in the query and document are mapped to concepts. In medical IR, methods [2, 12] employ MetaMap to map words to concepts in the Unified Medical Language System (UMLS). Processing millions of documents in a collection using MetaMap involves a considerable amount of time complexity. To avoid this difficulty, concepts relevant to International Classification Diseases (ICD-10) were used as a concept resource because they are closely related to diseases. These concepts were collected from Wikipedia. Articles linked to the name of the section and the sub-section of ICD-10 were crawled. As a result, 3,784 articles with 93,756 unique words were obtained. The title of an article was used as a medical concept. Figure 2 shows an example of the medical concept *Bubonic plague* [2] in Wikipedia. Based on the concepts, a word-concept matrix filled with standard TF-IDF values was constructed. Then, a similarity score between a query and a document is computed after concept mapping, as shown in Figure 3. Cosine similarity was utilized as a scoring function.

[2] http://en.wikipedia.org/wiki/Bubonic_plague

Fig. 2. An example of the Wikipedia article of the medical concept *bubonic plague*

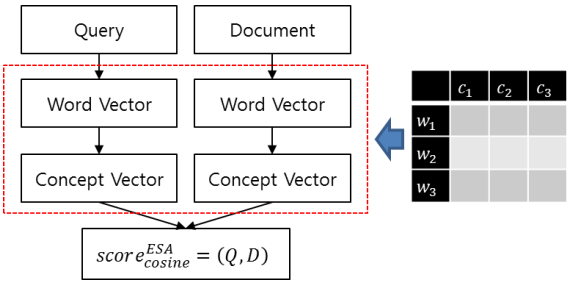

Fig. 3. Similarity computation using concept mapping

Component 2 - Cluster-based External Expansion Model: There are several medical collections, TREC CDS and OHSUMED, available to researchers, as medical collections have been developed for different purposes. For re-ranking purposes, these collections can be used as textual resources to build more robust external expansion models [13]. To this end, we revised an existing external expansion model (EEM) by combining it with a cluster-based document model [8]. The key idea of the EEM is to generate a feedback model by determining the proper contributions of multiple collections for a given query. Formally, the EEM is defined as follows:

$$p_{EEM}(w|Q) \propto \sum_{C \in E} p(Q|\theta_C) \cdot p(\theta_C) \sum_{D \in C} p(w|\theta_D) \cdot p(Q|\theta_D) \cdot p(D|\theta_C). \quad (6)$$

Specifically, the EEM consists of five components: the prior collection probability, document relevance, collection relevance, document importance, and word probability. Prior collection probability $p(\theta_C)$ is the prior importance of a collection among all the collections in use. Without the prior knowledge of collections, it can be ignored by setting a uniform probability $p(\theta_C) = \frac{1}{|E|}$. Document relevance $p(Q|\theta_D)$ is the relevance of a document *D* to a given query *Q*. Precisely, it is a query-likelihood score given to a document. Collection relevance $p(Q|\theta_C)$ is the relevance of a query *Q* with respect to a collection *C*. This component determines the query-dependent contribution of a collection when constructing the EEM. To avoid time-consuming iteration over a collection *C*, it can be estimated using the most highly relevant documents with the assumption that documents are equally important in a given collection *C*. Thus, it is the

average score of the feedback documents in D_{init}. Document importance $p(D|\theta_C)$ refers to the importance of a document *D* in a collection *C*. This is also ignored by setting to a uniform probability $p(D|\theta_C) = \frac{1}{|C|}$ without the prior knowledge of documents in a collection *C*. Word probability $p(w|\theta_D)$ is a probability of observing a word *w* in a document *D*. In [13], the MLE is utilized to estimate this component.

In the cluster-based document model [8], a document model is smoothed with cluster and collection models in which the clusters are generated with the K-means algorithm. Therefore, we can obtain more accurate document models because the probabilities of words which occur frequently in a cluster or a collection are decreased. Similarly, we can assume that each collection corresponds to a cluster explicitly partitioned over *E*. This assumption allows the use of the cluster-based document model without any additional computations with K-means clustering, as *K* is determined via $|E|$, and each collection is a cluster. All that is required is to utilize the statistics of a collection *C* for a cluster. Then, a document model is defined as follows:

$$\begin{aligned} p(w|\theta_D) &= (1-\lambda_E)\cdot\frac{c(w,D)+\mu\cdot p(w|C)}{|D|+\mu}+\lambda_E\cdot p(w|E) \\ &= (1-\lambda_E)\cdot\left[\frac{|D|}{|D|+\mu}p(w|D)+\frac{\mu}{|D|+\mu}p(w|C)\right]+\lambda_E\cdot p(w|E), \end{aligned} \tag{7}$$

where λ_E is a control parameter for all collections in *E*.

Our CBEEM is defined by revising $P(w|\theta_D)$ in Equation 6 and replacing it with that of Equation 7. Based on this revision, the CBEEM is expected to be a probability distribution over topical words because it is combined with individual RMs owing to the decrease in the probability of common words in the feedback documents. Then, a new query model is estimated with the CBEEM as follows:

$$p\left(w\middle|\theta_Q'\right) = (1-\beta)\cdot p\left(w\middle|\theta_Q\right)+\beta\cdot p_{CBEEM}(w|Q) \tag{8}$$

Component 3 - Concept-based Document Centrality: To utilize external resources, we designed a concept-based document centrality method (CBDC) as an additional re-ranking component. The key idea originated from centrality-based document scoring, which utilizes the associations among documents in the search results [6]. The centralities are computed through two steps - similarity matrix construction and a random-walk step. Among the initial documents, implicit links are generated because there are no explicit links among them. Then, the documents are re-ranked by combining the initial and centrality scores, as follows:

$$score(Q,D) = score_{QLD}(Q,D)\cdot score_{DC}(Q,D) \tag{9}$$

However, the CBDC differs from previous approaches [6] in two aspects. First, we attempted to capture the associations among a query and documents explicitly when computing document centralities, while the previous method only considered the associations among documents. Second, the CBDC captures the associations at the concept level while the previous method focused on the word level. The CBDC is estimated as follows. First, the document-concept weight matrix is constructed by concept mapping.

In this matrix, the query is augmented at the ends of the rows. Then, a document-document similarity matrix is computed using the document-concept weight matrix. Due to the need to augment the query, the CBDC considers the associations of documents with respect to a query. Next, a random walk was performed to compute centrality scores. We only utilized the centrality scores of documents.

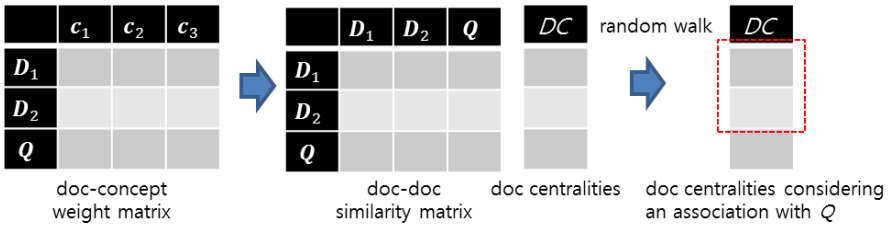

Fig. 4. Computation of concept-based document centralities

3 Experiments

3.1 Data

We utilized three medical external resources, TREC CDS, OHSUMED, and ICD-10, which were extracted from Wikipedia. Tables 1 show a summary of the TREC CDS and OHSUMED collections. TREC CDS consists of biomedical literature, specifically a subset of PubMed Central. A document is a full-text XML of a journal article. OHSUMED consists of biomedical literature which is a subset of the clinically oriented MEDLINE. Clearly, $E = \{C_{eHealth}, C_{CDS}, C_{OHSUMED}\}$ for the CBEEM.

Table 1. Data Statistics (The lengths are counted after stop-word removal.)

	CLEF eHealth	TREC CDS	OHSUMED
#Docs	1,102,289	732,451	348,566
Voc. Size	2,647,062	6,931,356	122,512
Avg. Doc. Len	540.0	1779.0	68.0

3.2 Evaluation Settings

Lucene[3] was exploited to index and search the initial documents D_{init}. Stop-words were removed using 419 stop-words[4] in INQUERY. In addition, numbers were normalized to NU<# of DIGITS>. A query-likelihood method with Dirichlet smoothing was chosen

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

[4] http://sourceforge.net/p/lemur/galago/ci/default/tree/core/src/main/resources/stopwords/inquery

as a scoring function. $|D_{init}|$ was set to 1000. Based on D_{init}, we performed eight runs by differentiating the combining components of our re-ranking framework. Table 2 shows the descriptions and Tables 3 and 4 summarize the performances of the submitted runs. The performances were measured by P@10, NDCG@10, rank-biased precision (RBP), and two different variants of RBP (i.e., uRBP, and uRBPgr). In contrast to the evaluation settings used in previous years, the readability of the retrieved medical content, along with the common topical assessments of relevance, is added as new evaluation measure [15].

3.3 Results

Table 2 describes our submitted runs for CLEF 2015 eHealth Task 2 and Table 3 summarizes our results obtained from the task's official standard evaluation set. Runs 7 and 8 are different from runs 5 and 6, as the experiments were performed with expanded queries produced from the CBEEM for ESA and CBDC, while runs 5 and 6 used original queries.

According to Table 3, ESA and CBDC using the concept relevant to ICD-10 are not helpful according to a comparison of runs 1, 2 and 3. It can be concluded that the reduction of the concept space without precise ICD-10 concepts resulted in low discrimination power. On the other hand, the CBEEM showed consistent improvements over QLD.

The best performance was obtained in runs 6 and 8, where the CBEEM and the CBDC were combined. This finding indicates that the use of external medical resources when also considering concept-level associations can have synergetic effects on the re-ranking of documents when they are in the proper right sequence. Moreover, the CBDC is not apparently affected by the query expansion results.

Table 2. Descriptions of our Submitted Runs

Run	Description
1	Query likelihood method with Dirichlet smoothing (QLD)
2	QLD + Explicit semantic analysis (ESA)
3	QLD + Concept-based document centrality (CBDC) using ESA
4	QLD + Cluster-based external expansion model (CBEEM)
5	QLD + CBEEM+ ESA
6	QLD + CBEEM+ CBDC
7	QLD + CBEEM + ESA with expanded query
8	QLD + CBEEM + CBDC with expanded query

Table 3. Performances of the Submitted Runs for Topical Relevance

Run	P@10	NDCG@10
1	0.3606	0.3352
2	0.3455	0.3223

3	0.3591	0.3395
4	0.3788	0.3424
5	0.3606	0.3362
6	**0.3864**	**0.3464**
7	0.3727	0.3459
8	**0.3864**	**0.3464**

In comparison with the readability-based measures (i.e., uRBP and uRBPgr), the best results in RBP were obtained from runs 6 and 8. However, the best performances of the two readability-based measures were observed from run 7.

Table 4. Performances of Submitted Runs for Readability-Biased Relevance

Run	RBP	uRBP	uRBPgr
1	0.3222	0.2593	0.2646
2	0.3038	0.2607	0.2614
3	0.3295	0.2596	0.2666
4	0.3306	0.2644	0.2709
5	0.3203	0.2702	0.2725
6	**0.3332**	0.2607	0.2695
7	0.3299	**0.2703**	**0.2739**
8	**0.3332**	0.2607	0.2695

The results show that the selection of re-ranking components is important because some of them can degrade previously achieved levels of moderated performance. In addition, we can expect additional performance improvements by combining two different re-ranking components if their application sequence is appropriate.

4 Conclusion

This working note describes our efforts to find high-performance combinations of different re-ranking components which utilize external medical resources. Among the different runs we attempted, runs 6 and 8 (where our proposed CBEEM and CBDC were used) showed the best performance in P@10, NDCG@10, and RBP. These results imply that the effective use of external medical resources for re-ranking can overcome the innate limitations of naïve queries by laypeople. As our future work, to enhance the proposed re-ranking components, we plan systematically to analyze symptom-wise evidence residing in promising external medical resources.

References

1. Abdul-Jaleel, N. et al.: UMass at TREC 2004: Novelty and HARD. Proceedings of Text REtrieval Conference (TREC). (2004).
2. Choi, S. et al.: Semantic concept-enriched dependence model for medical information retrieval. Journal of biomedical informatics. 47, 18–27 (2014).
3. Egozi, O. et al.: Concept-Based Information Retrieval Using Explicit Semantic Analysis. ACM Transactions on Information Systems. 29, 2, 1–34 (2011).
4. Goeuriot, L. et al.: Overview of the CLEF eHealth Evaluation Lab 2015. CLEF 2015 - 6th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS), Springer (2015).
5. Hersh, W. et al.: OHSUMED: an interactive retrieval evaluation and new large test collection for research. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '94. pp. 192–201 (1994).
6. Kurland, O., 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. pp. 306–313 ACM Press, New York, New York, USA (2006).
7. Lafferty, J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01. pp. 111–119 ACM Press, New York, New York, USA (2001).
8. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04. pp. 186–193 ACM Press, New York, New York, USA (2004).
9. Oh, H.-S., Myaeng, S.-H.: Utilizing global and path information with language modelling for hierarchical text classification. Journal of Information Science. 40, 2, 127–145 (2014).
10. Palotti, J. et al.: CLEF eHealth Evaluation Lab 2015, task 2: Retrieving Information about Medical Symptoms. CLEF 2015 Online Working Notes. CEUR-WS (2015).
11. Simpson, M.S. et al.: Overview of the TREC 2014 Clinical Decision Support Track. Proceedings of Text REtri eval Conference (TREC). (2014).
12. Wang, Y. et al.: A Study of Concept-based Weighting Regularization for Medical Records Search. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 603–612 (2014).
13. Weerkamp, W. et al.: Exploiting External Collections for Query Expansion. ACM Transactions on the Web. 6, 4, 1–29 (2012).

14. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems. 22, 2, 179–214 (2004).
15. Zuccon, G., Koopman, B.: Integrating Understandability in the Evaluation of Consumer Health Search Engines. Proceedings of the SIGIR Workshop on Medical Information Retrieval (MedIR). (2014).