A Multiple-stage Approach to Re-ranking Clinical Documents

Heung-Seon Oh and Yuchul Jung

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

Abstract. This paper presents our approach to medical information retrieval and experimental results of participating in eHealth Task 3-A of CLEF 2014. The task is to retrieve relevant documents from a medical collection given a query generated from a discharge summary. The key idea of our method is to compute accurate similarity scores via multiple stages of re-ranking documents from initial documents retrieved by a search engine.

Keywords: medical information retrieval, language models, abbreviations, query expansion

1 Introduction

Health-related content is one of the most searched-for topics on the internet. This became an important domain for research in information retrieval (IR). Recently, medical IR is actively researched to tackle diverse medical information sources including the general web, journal articles, social media, and hospital records. However, medical IR is still challenging because it should consider various information needs from a wide range of users including patients and their care givers, researchers, clinicians, practitioners, etc. Moreover, it is highly co-related with those users' background medical knowledge and language skills.

eHealth Task 3-A of Conference and Labs of the Evaluation Forum (CLEF) 2014 [1, 2] aims at improving the effectiveness of medical IR systems to support laypeople (e.g., patients and their relatives) who have different information needs. Most of previous researches focus on utilizing external medical resources such as MetaMap [3], NegEx [4], and international classification of diseases(ICD)-9 and natural language processing (NLP) [5] to understand the meanings of medical words at semantic level.

This paper presents a multiple-stage re-ranking method which focuses on utilizing various retrieval techniques rather than exploiting utilizing external resources and NLP techniques.

In particular, our proposed method passes through multiple re-ranking stages to elevate the ranked position of most relevant documents. Basically, we first perform query expansion with abbreviations, and pseudo relevance model in the end. In the

middle of the re-ranking, query expansion with discharge summary, clustering-based document scoring, and centrality-based document scoring can be combined selectively or sequentially.

The rest of this paper is organized as follows. Section 2 delivers related researches of medical information retrieval. Section 3 presents our re-ranking method in detail. The experimental results are described in Section 4. Section 5 concludes with short summary.

2 Related Work

Recently, many IR researches have been performed with different types of medical collections. TREC held medical track in 2011 and 2012. A research [6] presents two-stages method. They extract useful attributes such as age and gender from a collection by NLP techniques and hand-crated regular expressions. In search time, a query is expanded using unified medical language system (UMLS). In [4], several ranking functions are proposed to combine several evidence of different levels including various external medical resources. The results show that the proposed methods achieved the best performance. A research [5] presents a negotiation detection method using syntactic information and shows the effective way of handling negations.

CLEF held eHealth Lab in 2013. A research [7] presents a two-step ranking system utilizing three different external resources: external medical collections, medical concept mapper, and discharge summaries. It first retrieves documents in text-space and re-rank them in concept-space.

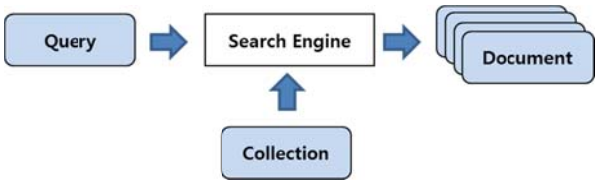
MedSearch system [8] addresses three distinctions compared to traditional systems. First, it provides query reformulation which makes a long descriptive query to a moderate-length query. Second, it supports the diversification of web search results. Third, it provides medical phrases semantically related to a query from Medical Subject Headings (MeSH) ontology.

3 Methods

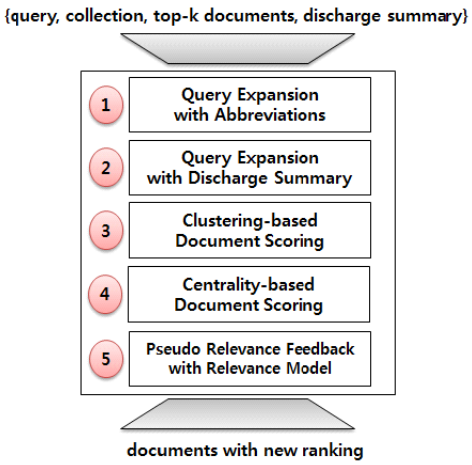
The key idea of our method is to re-rank top-k documents via multiple stages for computing more accurate similarity scores with respect to a query.

Figure 1 shows the overview of our multiple stages re-ranking method. For a given query Q, a set of documents, $D_{init} = \{D_1, D_2, \dots, D_k\}$, are retrieved from a collection C using a search engine. In our implementation, initial documents are retrieved by Lucene[1] using a query-likelihood method with Dirichlet smoothing [9]. Based on the initial documents, re-ranking is performed via multiple stages. The rest of this paper explains the details of the re-ranking method.

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



(a) Searching Initial documents



(b) Re-ranking initial documents

Fig. 1. Overview of our document re-ranking method: documents are re-ranked via multiple stages (b) based on the initially retrieved documents (a)

Throughout re-ranking, KL-divergence method is utilized to compute a similarity score between a query and a document [10, 11]:

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

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

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

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

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 [9]:

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

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

The first stage aims at expanding a query with abbreviations. In numerous numbers of medical documents, abbreviations are widely used to represent important meanings. Unfortunately, the clear interpretation of abbreviations is quite difficult due to the existences of several different meanings for a same abbreviated expression. Similarly, medical queries generated by users may also contain abbreviations. If we submit a query including abbreviations, it may not match relevant documents due to term mismatch problem or may match documents with abbreviations implying different meanings. To deal with this problem, query expansion considering abbreviations is considered. To do that, we extract pairs of abbreviation and corresponding full representation with an occurrence count using simple rule-based extraction method [12] from the entire collection. Then, a query model is estimated by incorporating words from the full representations of an associated abbreviation:

$$p(w|\theta'_Q) = (1 - \lambda_{abbr}) \cdot p(w|\theta_Q) + \lambda_{abbr} \cdot \sum_{t \in full(w)} p_{abbr}(t|w) \quad (4)$$

where $p(w|\theta_q)$ is MLE, λ_{abbr} is a control parameter, $full(w)$ is a set of words consisting of a full representation for an abbreviation w, and $p_{abbr}(t|w)$ is estimated by $\frac{count(t)}{count(t,w)}$.

The second stage is to reflect information from a discharge summary. A query used in CLEF eHealth Task-3 is generated by a human expert after reading a discharge summary corresponds to the query. Therefore, it has hidden but useful information not captured by a query. The use of a discharge summary can improve retrieval performance by utilizing such hidden information. To do that, a query model is expanded by combining a random-walk based discharge summary model. First, we should compute word-to-word transition matrix to measure the associations among words in a discharge summary. A simple solution is to use a co-occurrence count between two words among all sentences [13]. However, words are strongly associated when they appear closely in a sentence. In addition, associations between topical words are important than those between common words. To resolve this situation, we utilize hyperspace analogue to language (HAL) [14] function with inverse document frequency (IDF):

$$HAL(w,u) = \sum_{n=1}^{N} wt(n) \cdot co(w,u,n) \cdot IDF(w) \cdot IDF(u) \quad (5)$$

where n is a distance between words w and u, N is a window size, $\mathrm{wt(n)} = N - n + 1$, $co(w, u, n)$ is a co-occurrence count of w and u within k-distance, and $IDF(w) = \log\left(\frac{|C|}{doc\_freq(w)}\right)$

Then, a transition probability is computed:

$$p_{HAL}(w|u) = \frac{HAL(w,u)}{\sum_{t \in DS} HAL(t,u)} \tag{6}$$

where DS is a discharge summary document

Based on the translation matrix $p_{HAL}(w|u)$, word centralities are computed using random-walk:

$$cent(w) = \frac{\lambda_{DP}}{|V_{DS}|} + (1 - \lambda_{DP}) \cdot \sum_{u \in DS} \frac{cent(u)}{p_{HAL}(w|u)} \tag{7}$$

where $|V_{DS}|$ is the number of unique words in a discharge summary DS and λ_{DP} is a damping factor .

We approximate the resulting $cent(w)$ as a discharge summary model $p(w|\theta_{DS})$ and update the query model with it:

$$p\left(w\middle|\theta_Q''\right) = (1 - \lambda_{DS}) \cdot p\left(w\middle|\theta_Q'\right) + \lambda_{DS} \cdot p(w|\theta_{DS}) \tag{8}$$

where λ_{DS} is a control parameter and $p(w|\theta_{DS})$ is a discharge summary model.

The third stage is to incorporate cluster information of documents. Namely, a score for a document is computed by incorporating the membership of the document to a cluster we constructed. Bottom-up hierarchical agglomerative clustering [15] is applied to partition the top-k documents, D_{init}, into a set of disjoint clusters. At first, k-clusters for every document in D_{init} are constructed. Then, two clusters which have the highest similarity are selected and merged to a single cluster if the similarity is above a threshold. This procedure stops when there are no clusters with the threshold above. Similarity scores are computed using KL divergence method between a query model and Dirichlet-smoothed cluster model.

A new score is computed by combining the initial search score and the cluster score:

$$score_{sc}(Q, D) = score_{search}(Q, D) \cdot score_{cluster}(Q, CL_D) \tag{9}$$

where CL_D is a cluster of a document D. $score_{sc}(Q, D)$ is used after normalization over all document scores.

The fourth stage is to utilize the associations among documents by generating implicit links [10]. This stage consists of two steps: similarity matrix construction and random-walk. For each document $d \in D_{init}$, α documents in D_{init} are selected according to high generation probabilities:

$$score(D_1, \mathrm{D}_2) = \exp\left(-KL\left(\theta_{D_1} || \theta_{\mathrm{D}_2}\right)\right) \tag{10}$$

Based on those results, a similarity matrix with the initial documents and corresponding α documents is constructed. Then, random-walk is executed on this matrix to produces centrality scores for the initial documents. This score is multiplied with the previous score:

$$score_{scs}(Q, D) = score_{sc}(Q, D) \cdot score_{centrality}(Q, D) \tag{11}$$

The fifth stage is pseudo relevance feedback. A popular way of query expansion is to update a query based on pseudo-relevance feedback (PRF). Updating a query with PRF assumes that top-ranked documents $F = \{D_1, D_2, \ldots, D_{|F|}\}$ in an 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:

$$\begin{aligned} p_{RM1}(w|Q) &= \sum_{D \in F} p(w|\theta_D) p(\theta_D|Q) \\ &= \sum_{D \in F} p(w|\theta_D) \frac{p(Q|\theta_D) p(\theta_D)}{p(Q)} \\ &\propto \sum_{D \in F} p(w|\theta_D) p(\theta_D) p(Q|\theta_D) \end{aligned} \tag{12}$$

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 the retrieval performance, a new query model can be estimate by combing a relevance model and an original query model. RM3 [16] is a variant of a relevance model to estimate a new query models with RM1:

$$p\left(w\middle|\theta_Q''\right) = (1 - \beta) \cdot p\left(w\middle|\theta_Q''\right) + \beta \cdot p_{RM1}(w|Q) \tag{13}$$

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

Based on this query model, final scores for documents are computed.

4 Experiments

As mentioned, initial documents are retrieved by Lucene using a query-likelihood method with Dirichlet smoothing. We limited the size of the initial documents to 100. Based on the initial documents, we submitted 7 runs by differentiating the components of our re-ranking method.

Table 1 shows the parameter and corresponding values for each component in the experiments. Table 2 describes involving components at each run and evaluation results from corresponding runs. Basically, component 5 which indicate the use of PRF is applied to all runs thus regarded as baseline of our experiments. Except Run01, all runs utilize component 1. The distinction between Run02-04 and Run05-07 is that the former uses discharge summary while the latter doesn't. Precision and normalized discounted cumulative gain (NDCG) are used to measure the performance of top-10 ranked documents from 100 initial documents. They are denoted as P@10 and NDCG@10, respectively.

Our baseline achieved 0.7300 and 0.7235 in P@10 and NDCG@10, respectively. It shows that PRF is an effective solution to find correct medical documents. For precision, the best performance, 0.7400 is obtained from Run02 which utilizes abbreviations and discharge summary. For NDCG, the best performance, 0.7333, is obtained from Run04 which uses all components in the re-ranking method. This shows that sequentially combining the proposed components is contributed to achieve the best performance in NDCG measure. However, clustering and centrality-based document scoring were not effective in enhancing precision measure.

Table 1. Parameters setup used in re-ranking method

Component	Description	Parameters
1	Query expansion with abbreviations	abbreviation_mixture =0.15
2	Query expansion with discharge summary	hal_window_size=3 random_walk_damping_factor =0.85
3	Clustering –based document scoring	clustering_similarity threshold = 0.9
4	Centrality-based document scoring	random_walk_damping_factor =0.85 alpha_doc_size=10
5	Pseudo relevance feedback with relevance model	feedback_doc_size = 10 feedback_word_size=100 feedback_mixture=0.1 dirichlet_mixture = 1500

Table 2. Evaluation results

Run Id	Components					Evaluation Measures	
	1	2	3	4	5	P@10	NDCG@10
KISTI_EN_RUN01					O	0.7300	0.7235
KISTI_EN_RUN02	O	O			O	**0.7400**	0.7301

KISTI_EN_RUN03	O	O	O		O	0.7160	0.7171
KISTI_EN_RUN04	O	O	O	O	O	0.7380	**0.7333**
KISTI_EN_RUN05	O				O	0.7280	0.7211
KISTI_EN_RUN06	O		O		O	0.7240	0.7187
KISTI_EN_RUN07	O		O	O	O	0.7260	0.7233

Due to quite high baseline (i.e., Run01) obtained by PRF with relevance model and lack of in-depth study on the provided healthcare dataset, our experiments fail in showing drastic improvements in evaluation measures. Meanwhile, the moderate performances observed in our multi-stage approach to re-ranking documents (i.e., Run04) may arise from synergistic effects between involved components. The detailed analysis on the involved components in terms of causal and sequential effects is remained as our future work.

5 Conclusion

This paper shows a multiple stage approach to re-ranking medical documents. Our method focuses on utilizing various retrieval techniques rather than utilizing medical dependent external resources and natural language processing to understand medical meanings. We found that using abbreviations and discharge summary play an important role to find correct medical documents. Our future works include further development of two components and in-depth error analysis based on standard assessment dataset.

References

1. Kelly, L., Goeuriot, L., Suominen, H., Schrek, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. Proceedings of CLEF 2014. Springer (2014).

2. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Mueller, H.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. Proceedings of CLEF 2014 (2014).

3. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA. 17, 229−36 (2010).

4. Zhu, D., Carterette, B.: Exploring evidence aggregation methods and external expansion sources for medical record search. Proceedings of Text REtrieval Conference (TREC). 1–9 (2012).

5. Dıaz, A., Ballesteros, M., Carrillo-de-Albornoz, J., Plaza, L.: UCM at TREC-2012: Does negation influence the retrieval of medical reports? Proceedings of Text REtrieval Conference (TREC) (2012).

6. King, B., Wang, L., Provalov, I., Learning, C., Zhou, J.: Cengage Learning at TREC 2011 Medical Track. Proceedings of Text REtrieval Conference (TREC) (2011).

7. Zhu, D., Stephen, W., James, M., Carterette, B., Liu, H.: Using Discharge Summaries to Improve Information Retrieval in Clinical Domain. ShARe/CLEF eHealth Evaluation (2013).

8. Luo, G., Tang, C., Yang, H., Wei, X.: MedSearch. Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM ’08. p. 143. ACM Press, New York, New York, USA (2008).

9. Zhai, C., 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. pp. 334–342. ACM Press, New York, New York, USA (2001).

10. Kurland, O., Lee, L.: PageRank without hyperlinks. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’05. p. 306. ACM Press, New York, New York, USA (2005).

11. Oh, H.-S., Myaeng, S.-H.: Utilizing global and path information with language modelling for hierarchical text classification. Journal of Information Science. 40, 127–145 (2013).

12. Schwartz, A.S., Marti A. Hearst: A simple algorithm for identifying abbreviation definitions in biomedical text. Proceedings of Pacific Symposium on Biocomputing. pp. 451–462 (2003).

13. Weeds, J., Weir, D.: Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. Computational Linguistics. 31, 439–475 (2005).

14. Song, D., Bruza, P.: Discovering information flow suing high dimensional conceptual space. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’01. pp. 327–333. ACM Press, New York, New York, USA (2001).

15. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval.

16. Abdul-Jaleel, N., Allan, J., Croft, W., Diaz, F., Larkey, L.: UMass at TREC 2004: Novelty and HARD. (2004).