KISTI at CLEF eHealth 2017 Patient-Centered Information Retrieval Task-1: Improving Medical Document Retrieval with Query Expansion

Heung-Seon Oh and Yuchul Jung

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

Abstract. In this report, we describe our retrieval framework for participating in CLEF eHealth 2017 Patient-Centered Information Retrieval Task-1: Ad-hoc Search. Our retrieval framework is a query expansion approach which adopts relevance and pseudo relevance feedback to improve retrieval performance.

Keywords: language model, feedback model, query expansion

1 Introduction

This report summarizes our approaches to CLEF eHealth 2017 [2] Patient-Centered Information Retrieval Task-1, a standard ad-hoc task [7] . As same with 2016, this task utilizes a large web corpus (ClueWeb12 B13) and topics developed by mining health web forums where users were seeking advice about specific symptoms, diagnosis, conditions or treatments.

The main goal of the task is to improve the relevance assessment pool and the collection reusability. To meet the evaluation requirements of this year, we explicitly exclude documents that have been already assessed in 2016 from our search results. Meanwhile, to enhance the relevance of the searched, we utilize the already assessed documents in our proposed approaches by following the suggested guideline.

Based on the above considerations, we've designed a medical information retrieval framework which is characterized with relevance feedback for initial search and query expansion for re-ranking.

2 Method

2.1 Retrieval framework

Our proposed framework basically performs selective query expansion techniques in the initial retrieval and re-ranks the retrieval results based on the more accurate query expansion methods developed. Figure 1 shows the overview of our retrieval framework. First, we employ relevance feedback (RF) based on the relevance judgements built in last year since it is encouraged to improve retrieval performance and relevance assessment pool. For a query Q, a feedback model, Q_F, is constructed and combined to produce a new query model, Q_{RF}. Second, an initial search is performed using Q_{RF} and produces a set of documents, $D_s = \{D_1, D_2, \dots, D_{|D_s|}\}$, from a collection C. For the retrieved documents, we perform re-ranking with new queries built via two different query expansion methods.

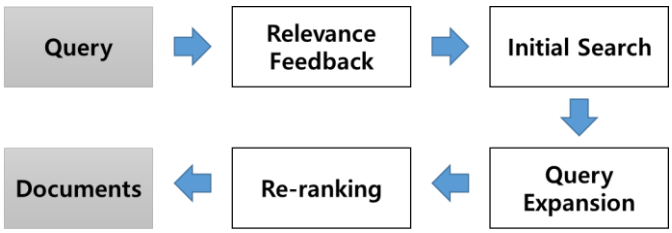

Fig. 1. Overview of retrieval framework

As summarized above, our framework starts with the relevance feedback to improve retrieval performance and relevance assessment pool. Let D_R is a set of documents relevant to a query Q. A relevance model, i.e. RM1 [4], is constructed with D_R scored by KL-divergence method (KLD) [3, 6, 9]. There exists two differences compared to standard RM1 since it is built using the relevance judgements. First, all documents in D_R are used to involve in a feedback model because they are explicitly relevant. Second, the relevance are employed as document priors. From the differences, it is expected that a query model includes all relevant information in D_R. Finally, a new query Q_{RF} is constructed via RM3 [1]. After that, the initial search is performed using KLD method on the entire collection C and obtain a set of retrieved documents D_I which are target for re-ranking.

Before performing the re-ranking, two different query expansion techniques are considered based on D_I. The first query expansion approach adopts random-walk based centrality scores [5] with a different transition matrix. This strategy is to estimate the query model by considering the associations of words in a query. The major difference is that an association between two words w and u where $w, u \in Q_{RF}$ is computed using two corresponding word vectors rather than co-occurrences. The word vectors are an accurate representation obtained through GloVe [8], an unsupervised learning algorithm for obtaining vector representations for words, so call word embedding. The GloVe is known to outperform word2vec models on similarity tasks

and named entity recognition tasks. The word vectors were computed on TREC CDS 2016 collection [8] which contains about 1.2M biomedical journal articles. We expect that the word vectors are more representative in medical domain than other domains. Then, centrality scores are computed using random-walk based on the transition matrix and regarded as a query model. Similar to RM3 above, a query model Q_C are generated by combining Q_{RF} and the centrality scores. Finally, documents in D_{I} are re-ranked according to Q_C with KLD method.

The second query expansion approach follows cluster-based external expansion model (CBEEM) [6] which is an advanced version of using external collections in pseudo relevance feedback (PRF). The key idea of CBEEM is to estimate an accurate feedback model using not only the original collection but also other benchmark collections. Again, TREC CDS 2016 collection was employed as an external collection. As a result, re-ranking is performed with a new query Q_{CBEEM} with D_{I}.

3 Experiments

3.1 Data

Two different collections are used for target and external collections, respectively. The target collection is ClueWeb12-Disk-B (ClueWeb12B) including about 52M web pages while the external collection is TREC CDS 2016 including about 1.2M biomedical journal articles. In both collections, text of pages were extracted by removing HMTL and XML tags using JSOUP[1] parser. Table 1 shows the summary of data statistics of ClueWeb12B and TREC CDS 2016, respectively. Words occur less than 5 and longer than 100 characters are replaced with <UNK>. Numbers are normalized to <NUx> where x is length of a number. Finally, all words are lowercased. This normalization reduces noisy words. Stop-words were removed using 419 stop-words[2] in INQUERY on query time but not on indexing time.

Table 1. Data Statistics

	ClueWeb12B	TREC CDS 2016
#Docs	52,051,844	1,255,260
Voc. Size	20,139,450	2,938,617
Tokens	44,291,018,290	5,663,660,754
Avg. Doc. Len	850.9	4,511.9

[1] https://jsoup.org/

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

3.2 Evaluation Settings

All mixtures for combining the query and feedback models are set to 0.5. Dirichlet prior is set to 2500. In relevance feedback (RF), the size of feedback words is set to 50 while the size of feedback documents corresponds to the number of relevant documents. In two query expansion approaches, they are fixed as 5 and 50, respectively. Word vectors are estimated using GloVe with ADAM optimizer where the vector size is 200.

3.3 Submitted Runs

We submitted three runs for this task. Run1, considered as our baseline, is the results of applying RF. Run2 and Run3 employed centrality scores and CBEEM, respectively. Table 2 summarized three runs.

Table 2. Descriptions of our Submitted Runs

Run	Description
1	Relevance feedback (RF)
2	RF + Random-walk based centrality scores
3	RF + Cluster-based external expansion model

4 References

1. Abdul-Jaleel, N. et al.: UMass at TREC 2004: Novelty and HARD. In: Proceedings of Text REtrieval Conference (TREC). (2004).
2. Goeuriot, L. et al.: CLEF 2017 eHealth Evaluation Lab Overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer (2017).
3. Kurland, O., Lee, L.: PageRank without hyperlinks: Structural re-ranking using links induced by language models. In: 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).
4. Meij, E. et al.: Conceptual language models for domain-specific retrieval. Inf. Process. Manag. 46, 4, 448–469 (2010).
5. Oh, H.-S. et al.: A Multiple-Stage Approach to Re-ranking Medical Documents. In: Proceedings of CLEF. pp. 166–177 (2015).

6. Oh, H.-S., Jung, Y.: Cluster-based query expansion using external collections in medical information retrieval. J. Biomed. Inform. 58, 70–79 (2015).
7. Palotti, J. et al.: CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017).
8. Roberts, K. et al.: Overview of the TREC 2016 Clinical Decision Support Track. In: In Proceedings of The Twenty-Fifth Text REtrieval Conference (TREC 2016). (2016).
9. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the tenth international conference on Information and knowledge management. pp. 403–410 ACM, New York, New York, USA (2001).