Data Sci. Eng. (2017) 2:316–327
https://doi.org/10.1007/s41019-017-0052-2

A Feedback-Based Approach to Utilizing Embeddings for Clinical Decision Support

Chenhao Yang[1] · Ben He[1] · Canjia Li[1] · Jungang Xu[1]

Received: 24 August 2017 / Revised: 18 October 2017 / Accepted: 31 October 2017 / Published online: 10 November 2017

Abstract Clinical Decision Support (CDS) is widely seen as an information retrieval (IR) application in the medical domain. The goal of CDS is to help physicians find useful information from a collection of medical articles with respect to the given patient records, in order to take the best care of their patients. Most of the existing CDS methods do not sufficiently consider the semantic relation between texts, hence the potential in improving the performance in biomedical articles retrieval. This paper proposes a novel feedback-based approach which considers the semantic association between a retrieved biomedical article and a pseudo feedback set. Evaluation results show that our method outperforms the strong baselines and is able to improve over the best runs in the TREC CDS tasks.

Keywords Clinical Decision Support · Semantic association · Relevance feedback

1 Introduction

The goal of Clinical Decision Support (CDS) is to efficiently and effectively link relevant biomedical articles to meet physicians' needs for taking better care of their patients. In CDS applications, the patient records are considered as queries and the biomedical articles are retrieved in response to the queries. A major difference between CDS and traditional IR tasks is that the documents, mostly scientific articles, are very long and contain comprehensive information about a specific topic such as a treatment for a disease, or a patient case. As a result, the CDS queries, although longer than those in other IR tasks, may not cover the various aspects of the user information need, and simple document-query matching does not lead to optimal effectiveness in the CDS task.

Most of the existing CDS methods retrieve biomedical articles using the frequency-based statistical models [1, 2, 6, 9]. Those methods extract concepts from queries and biomedical articles, and further utilize concepts to apply query expansion or document ranking. Then, the relevance score of a given article is assigned based on the frequencies of query terms or concepts. Despite the fact that the frequency-based CDS methods have been shown to be effective and efficient in the CDS task [25], they ignore the semantic associations between texts. We argue that the retrieval effectiveness of the CDS systems can be further improved by integrating the semantic information. For instance, suppose two short medical-related texts as follows:

- The child has symptoms of strawberry red tongue and swollen red hands.
- This kid is suffering from Kawasaki disease.

Though the two short sentences have no terms in common, they convey the same meaning and are considered to be related to each other. However, the two sentences above are considered completely unrelated by the existing frequency-based CDS methods. In this paper, we aim to further enhance the retrieval performance of the CDS systems by taking the semantic association between texts into consideration. Benefiting from recent advances in natural language processing (NLP), words and documents can be represented with semantically distributed real-valued vectors, namely the *embeddings*, which are generated by

✉ Ben He
benhe@ucas.ac.cn

[1] University of the Chinese Academy of Sciences, Beijing, China

neural network models [3, 17, 21, 22]. The embeddings have been shown to be effective and efficient in many NLP tasks due to the ability in preserving semantic relationships in vector operations such as summation and subtraction [21]. In this study, we utilize the Word2Vec technique proposed by Mikolov et al. [17, 21] to generate embeddings of words and biomedical articles, which is widely considered as an effective embedding method in NLP applications [8, 20, 30]. As a state-of-the-art topic model, latent Dirichlet allocation (LDA) [5] is also used for comparison with Word2Vec in generating distributed representations of biomedical articles in this study.

There have been efforts in utilizing the embeddings to improve IR effectiveness. For example, Vulić and Moens estimate a semantic relevance score by the cosine similarity between the embeddings of the query–document pair to improve the performance of monolingual and cross-lingual retrieval [31]. Similar idea is presented in [32], where the semantic similarity between the embeddings of the patient record and biomedical article is utilized to improve the CDS system. We argue that query is a weak indicator of relevance in that query is usually much shorter than the relevant documents, such that the use of semantic associations of the query-document pairs may only lead to limited improvement in retrieval performance. To this end, this paper proposes a feedback-based CDS method which integrates semantic associations between texts to further enhance retrieval effectiveness. To the best of our knowledge, this paper is the first to estimate the relevance score for IR tasks based on document-to-document (D2D) embedding similarity. Experimental results show that our proposed CDS method can have significant improvements over strong baselines. In particular, a simple linear combination of the classical BM25 weighting function with the semantic relevance score generated by our method leads to effective retrieval results that are better than the best TREC CDS runs.

A conference version of this paper was published in [33]. Extensions to the conference version include:

- The experiments conducted on the recent TREC 2016 CDS task dataset. The results obtained on this new dataset are consistent with those on the TREC 2014 and 2015 CDS datasets.
- The proposed approach is further evaluated on five standard IR test collections. Results show that our approach is able to outperform strong baselines for IR tasks other than CDS.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related work. Section 3 describes the proposed feedback-based approach in details. For the evaluation of the proposed approach on the CDS datasets, the experimental settings and results are presented in Sects. 4 and 5, respectively. The proposed approach is further evaluated on other standard TREC IR collections in Sect. 6. Finally, Sect. 7 concludes this work and suggests possible future research directions.

2 Related Work

2.1 BM25 and PRF

As our CDS method is to integrate the semantic relevance score into the classical BM25 model with applying pseudo-relevance feedback (PRF), we introduce BM25 model and PRF in this section. The ranking function of BM25 given a query Q and a document d is as follows [26]:

$$score(d, Q) = \sum_{t \in Q} w_t \frac{(k_1 + 1)tf}{K + tf} \frac{(k_3 + 1)qtf}{k_3 + qtf} \quad (1)$$

where t is one of the query terms, and qtf is the frequency of t in query Q. tf is the term frequency of query term t in document d. K is given by $k_1((1 - b) + b \cdot \frac{l}{avg\_l})$, in which l and avg_l denote the length of document d and the average length of documents in the whole collection, respectively. k_1, k_3 and b are free parameters whose default setting is $k_1 = 1.2$, $k_3 = 1000$ and $b = 0.75$, respectively [26]. w_t is the weight of query term t, which is given by:

$$w_t = \log_2 \frac{N - df_t + 0.5}{df_t + 0.5} \quad (2)$$

where N is the number of documents in the collection, and df_t is the document frequency of query term t, which denotes the number of documents that t occurs.

Pseudo-relevance feedback (PRF) is a popular method for improving IR effectiveness by using the top-k documents as pseudo-relevance set[18]. One of the best-performing PRF methods on top of BM25 is an adoption of Rocchio's algorithm presented in [16], which is able to provide state-of-the-art retrieval effectiveness on standard TREC test collections [16]. BM25 with PRF is denoted as $BM25_{PRF}$ in this paper.

2.2 State-of-the-Art CDS Methods

Due to the specificity of medical healthcare field, most of the existing CDS methods retrieve biomedical articles based on concepts, including unigrams, bigrams and multi-word concepts. These concepts are extracted from different resources, such as queries, biomedical articles, external medical databases, etc. These content-based CDS methods usually utilize concepts to apply query expansion or document ranking based on the frequencies of the concepts. Palotti and Hanbury proposed a concept-based query