Integrating Feedback-based Semantic Evidence to Enhance Retrieval Effectiveness for Clinical Decision Support

Chenhao Yang[1] and Ben He[2] Jungang Xu[3]

University of Chinese Academy of Sciences, Beijing, China,
yangchenhao14[1]@mails.ucas.ac.cn, {benhe[2], xujg[3]}@ucas.ac.cn

Abstract. The goal of Clinical Decision Support (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 evidence, 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 CDS tasks of TREC 2014 & 2015.

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, the patient records are considered as queries and the biomedical articles are retrieved in response to the queries. With the development of medical research, the volume of the published biomedical articles is growing rapidly, resulting in the difficulty in seeking out the most relevant and timely information for a particular clinical case.

Most of the existing CDS methods retrieve biomedical articles using the frequency-based statistical models [1, 3, 7, 10]. 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 [22], they ignore the semantic evidence of relevance. 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 texts 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 evidence into consideration. Benefiting from recent advances in natural language processing (NLP), words and documents can be represented with semantically distributed real-valued vectors, i.e. *embeddings*, which are generated by neural network models [4, 14, 18, 19]. 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 [18]. In this study, we utilize the Word2Vec technique proposed by Mikolov et al. [14, 18] to generate embeddings of words and biomedical articles, which is widely considered as an effective embedding method in NLP applications [9, 17, 26]. As a state-of-the-art topic model, latent Dirichlet allocation (LDA) [6] 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 [27]. Similar idea is presented in [28], where the semantic similarity between the embeddings of the patient record and biomedical article is utilized to improve the CDS system. Note that the patient records are used as queries in CDS as described above. 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 evidence to further enhance retrieval effectiveness. 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 runs in both 2014 & 2015 CDS tasks.

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 [23]:

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

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 [23]. 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. PRF provides a feedback-based automatic method for improving the retrieval performance by expanding the original user input query [15].

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 expansion method, increasing the weights of relevant concepts and expanding the original query with concepts extracted by MetaMap [20]. MetaMap is a highly configurable tool for recognizing the Unified Medical Language System (UMLS) concepts in text, which is usually utilized in the existing CDS methods. Song et al. proposed a customized learning-to-rank algorithm and a query term position based re-ranking model to improve the retrieval performance [25]. As biomedical articles are usually full-text scientific articles which are much longer than Web documents, Cummins et al. applied the recently proposed SPUD language model [11] to CDS for retrieving longer documents more fairly [10]. Abacha and Khelifi investigated several query reformulation methods utilizing Mesh and DBpedia. In addition, they applied rank fusion to combine different ranked document lists into a single list to improve the retrieval performance [1].

2.3 The Best Methods in the TREC 2014 & 2015 CDS Tasks

Choi and Choi proposed a three-step biomedical article retrieval method, which obtains the best run in the TREC 2014 CDS task [7]. Firstly, the method

utilizes external knowledge resource to apply query expansion, and uses the query likelihood (QL) language model [21] to rank articles. Secondly, a text classification based method is used for the topic-specific ranking. Note that the topics used in the TREC CDS task are classified into three categories, i.e. *diagnosis*, *test* and *treatment*. Finally, the method combines the relevance ranking score and the topic-specific ranking score with Borda-fuse method [7].

The CDS methods proposed by Balaneshin-kordan et al. [3] obtained both the best automatic and manual runs in the TREC 2015 CDS task. Their method extracts unigrams, bigrams and multi-word UMLS concepts from queries, the pseudo relevance feedback documents or external knowledge resources, and then uses the Markov Random Field (MRF) model [16] for document ranking. The relevance score of a document d given a query Q is computed as follows [3]:

$$\begin{aligned} score(d, Q) &= \sum_{c \in \mathbb{C}} \mathbb{1}_c score(c, d) \\ &= \sum_{c \in \mathbb{C}} \mathbb{1}_c \sum_{T \in \mathbb{T}} \lambda_T f_T(c, d) \end{aligned} \tag{3}$$

where $score(c, d)$ is the contribution of concept c to the relevance score of document d. $\mathbb{1}_c$ is an indicator function which determines whether the concept c is considered in the relevance weighting. $\mathbb{C}$ is the set of concepts. $\mathbb{T}$ is the set of all concept types, to which concept c belongs. Note that a concept can belong to multiple concept types at the same time. λ_T is the importance weight of concept type T, and $f_T(c, d)$ is a real-valued feature function.

The existing CDS methods retrieve biomedical articles based on the frequencies of concepts. As discussed in Section 1, the lack of semantic evidence of relevance may lead to limited retrieval performance. A recent work [28] integrates semantic similarity between the embeddings of the patient record and biomedical article to improve the CDS system, which is given by:

$$Sim(d, Q) = 0.5 \cdot \frac{\overrightarrow{d} \cdot \overrightarrow{Q}}{|| \overrightarrow{d} || \times || \overrightarrow{Q} |} + 0.5 \tag{4}$$

where $\overrightarrow{d}$ and $\overrightarrow{Q}$ are the embeddings of biomedical article d and patient record Q, respectively. $Sim(d, Q)$ is the semantic similarity which is integrated into BM25 model [23] by a linear interpolation. As the patient records are usually much shorter than the full-text biomedical articles, they do not necessarily contain sufficient amount of semantic evidence of relevance. Therefore, the approach in [28] leads to limited improvement on the CDS task. To deal with this problem, in the next section, we propose a feedback-based approach that considers the semantic similarity between a retrieved article and a set of feedback articles, which is a better indicator of relevance than patient record.

3 Feedback-based Semantic Evidence

The methods for generating the embeddings of biomedical articles are introduced in Section 3.1. The generated embeddings are utilized for enhancing the retrieval performance of CDS in Section 3.2.

3.1 Generating Embeddings of Biomedical Articles

The Word2Vec technique proposed by Mikolov et al. [14, 18] is a state-of-the-art neural embedding framework, which has been shown to be effective and efficient in many NLP tasks. In this study, Word2Vec is also utilized to generate embeddings of words and biomedical articles. A unique advantage of Word2Vec is that the semantic relationships can be preserved in vector operations, such as addition and subtraction [18]. Therefore, the embeddings of biomedical articles can be generated through vector operations of word embeddings such that they are applicable to the CDS task. Considering the fact that informative words are usually infrequent in biomedical articles, we utilize the Skip-gram architecture of Word2Vec, which shows better performance for infrequent words than the CBOW architecture of Word2Vec in generating embeddings [18]. Besides, the negative sampling algorithm is used to train embeddings [18].

The Skip-gram architecture is composed of three layers, i.e. an input layer, a projection layer and an output layer. The basic idea of Skip-gram is to predict the context of a given word w. Considering the conditional probability $p(c(w)|w)$ given a word w and the corresponding context $c(w)$, the goal of Skip-gram model is to maximize the likelihood function as follows [12]:

$$\arg\max_{\theta} \prod_{(w,c(w))\in D} p(c(w)|w;\theta) \tag{5}$$

where w and $c(w)$ denote a word and the corresponding context, respectively. $(w, c(w))$ is a training sample, and D is the set of all training samples. θ is the parameter set that needs to be optimized. In addition, the conditional probability $p(c(w)|w)$ is modeled as Softmax regression, which is given as follows:

$$p(c(w)|w;\theta) = \frac{e^{v_w \cdot v_{c(w)}}}{\sum_{c(w)'\in C} e^{v_w \cdot v_{c(w)'}}} \tag{6}$$

where v_w and $v_{c(w)}$ are the embeddings of word w and the corresponding context $c(w)$, respectively. Substituting Equation (6) back into Equation (5), the final objective function of Skip-gram is given by:

$$\begin{aligned} &\arg\max_{\theta} \prod_{(w,c(w))\in D} \log p(c(w)|w) \\ &= \sum_{(w,c(w))\in D} \left(\log e^{v_w \cdot v_{c(w)}} - \log \sum_{c'} e^{v_w \cdot v_{c(w)'}} \right) \end{aligned} \tag{7}$$

where the parameters in Equation (7) are trained by stochastic gradient ascent.

A major challenge of the application of the word embeddings to CDS is how to generate effective embeddings for biomedical articles. In this paper, we adopt two ways of generating embeddings for biomedical articles, namely *Term Summation* and *Paragraph Embeddings*, abbreviated as *Sum* and *Para* respectively. As the semantic relationships are preserved in the embedding operations, one way of generating embeddings of biomedical articles is to sum up the word embeddings of the top-k most informative words in a given article, i.e. *Term Summation*, which is given by:

$$\overrightarrow{d} = \sum_{w \in W_k^d} tf\text{-}idf(w) \cdot \overrightarrow{w} \tag{8}$$

where $\overrightarrow{w}$ and $\overrightarrow{d}$ are the embeddings of word w and biomedical article d, respectively. W_k^d is the set of the top-k terms with the highest *tf-idf* weights in d. $tf\text{-}idf(w)$ is used to measure the amount of information carried by word w, which is given by:

$$tf\text{-}idf(w) = tf \cdot \log_2 \frac{N - df_w + 0.5}{df_w + 0.5} \tag{9}$$

where tf is the term frequency of w in d. N is the total number of biomedical articles in the whole collection, and df_w is the document frequency of word w.

In addition to *Term Summation*, we adopt the *Paragraph Embeddings* technique [14] to generate embeddings of biomedical articles. *Paragraph Embeddings* is an improved version of Word2Vec, in which each document is marked with a special word called *Paragraph id*. The *Paragraph id* participates in the training of each word as part of each context, acting as a memory that remembers what is missing from the current context. The training procedure of *Paragraph Embeddings* is the same as *Word2Vec*. Finally, the embedding of the special word *Paragraph id* is used to represent the corresponding biomedical article. We denote embeddings of biomedical articles generated by *Term Summation* and *Paragraph Embeddings* as $\overrightarrow{d_{Sum}}$ and $\overrightarrow{d_{Para}}$, respectively.

3.2 Using Embeddings for CDS

In this section, we introduce our proposed feedback-based CDS method, which considers the semantic similarity between a biomedical article to be scored and a pseudo feedback set. As Mikolov et al. demonstrated that words can have multiple degrees of similarity [18], integrating semantic associations by directly measuring the similarity between the embeddings of patient records and biomedical articles may only lead to limited improvement in retrieval performance (as used in [28]). Instead, we estimate the semantic relevance of a biomedical article by measuring the semantic similarity between the article and a pseudo relevance feedback set. Once we obtain the preliminary retrieval results

returned by BM25, the semantic relevance score of biomedical articles can be utilized to improve the retrieval performance, which is given as follows:

$$score(d, Q) = \lambda \cdot BM25(d, Q) + (1 - \lambda) \cdot SEM(d, D^k_{PRF}(Q)) \tag{10}$$

where $BM25(d, Q)$ is the ranking score of document d given by a baseline retrieval model, e.g. the classical BM25 model with PRF. $D^k_{PRF}(Q)$ is the pseudo relevance feedback set of biomedical articles, which is composed of the top ranked k articles returned by the baseline model. It is usually assumed by the PRF technique that most of the documents in $D^k_{PRF}(Q)$ are relevant to query Q, thus $D^k_{PRF}(Q)$ can be considered as a better indicator of relevance than patient records. $SEM(d, D^k_{PRF}(Q))$ measures the semantic similarity between document d and the pseudo relevance feedback set $D^k_{PRF}(Q)$, which is given as follows:

$$SEM(d, D^k_{PRF}(Q)) = \sum_{d' \in D^k_{PRF}(Q)} w_{d'} \cdot Sim(d', d) \tag{11}$$

where d' is one of the biomedical articles in $D^k_{PRF}(Q)$. $w_{d'}$ is the importance weight of d', which is given as follows:

$$w_{d'} = BM25(d', Q) + \max_{d'' \in D^k_{PRF}(Q)} BM25(d'', Q) \tag{12}$$

$Sim(d', d)$ denotes the semantic similarity between d' and d, which is given by Equation (4). In Equation (12), the maximum relevance score is added to normalize the gap between the relevance scores of different articles. Note that both $BM25(d, Q)$ and $SEM(d, D^k_{PRF}(Q))$ in Equation (10) are normalized by Min-Max normalization, such that the two scoring features are on the same scale.

4 Experimental Settings

In this section, we introduce the datasets used in the experiments and the experimental design.

4.1 Datasets

All our experiments are conducted on the standard datasets used in the TREC CDS tasks of 2014 and 2015. The target document collection used in the two years is an open access subset[1] of PubMed Central[2] (PMC), containing 733,138 full-text biomedical articles. We extract the *title*, *abstract*, *keywords* and

[1] `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`
[2] `http://www.ncbi.nlm.nih.gov/pmc`

body fields from each article as the source of the index. We use the open source Terrier toolkit version 4.1 [17] to index the collection with the recommended settings of the toolkit. Standard English stopwords are removed and the collection is stemmed using Porter's English stemmer. Using Porter's stemmer, inflected or derived words are reduced to their word stem, base or root forms.

There are 30 topics in each year, and each topic is a medical record narrative that serves as an idealized representation of actual patient record. These topics are classified into three categories, i.e. *diagnosis*, *test* and *treatment*, with 10 topics in each category. According to [24], there is little difference observed in retrieval performance when the three topic types are taken into account. Thus the topic types are not considered in our study. There are two versions of the medical record narratives, i.e. *Summary* and *Description* fields. Table 1 presents an example of the *Summary* and *Description* fields. The *Description* field is much longer than the *Summary* field, and has more detailed information about a patient. However, the *Description* field may contain more irrelevant information than the *Summary* field. In the experiments, both the *Summary* and *Description* fields are used as queries.

Table 1. Example of *Summary* and *Description* fields.

Topic type - diagnosis
Summary: A 62-year-old immunosuppressed male with fever, cough and intranuclear inclusion bodies in bronchoalveolar lavage.
Description: A 62 yo male presents with four days of non-productive cough and one day of fever. He is on immunosuppressive medications, including prednisone. He is admitted to the hospital, and his work-up includes bronchoscopy with bronchoalveolar lavage (BAL). BAL fluid examination reveals owl's eye inclusion bodies in the nuclei of infection cells.

As described in Section 3.1, the Skip-gram model of Word2Vec[3] toolkit is utilized to generate embeddings of words and biomedical articles, which are trained using the negative sampling algorithm [18]. Note that the *title*, *abstract*, *keywords* and *body* fields of each biomedical article are extracted as the training set of Word2Vec, and the stopword removal and stemming are applied. As recommended in [18], the window size is set to 10 for Skip-gram model. As documents in the target collection are full-text long biomedical articles, the number of dimensions of the embeddings are set to 300, a value that is larger than the recommended 100 in [2].

4.2 Experimental Design

In our study, we evaluate our CDS method against two baselines. As described in Section 3.2, we use the BM25 model [23] with applying PRF as one of the

[3] The learned embeddings of words and biomedical articles can be downloaded from `http://gucasir.org/CDS.tgz`.

Table 2. Evaluation results on the *Summary* field on the TREC 2014 CDS task. The difference in percentage is measured against the baseline retrieval model *BM25*. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Model	infNDCG	infAP	MAP	R-Prec
BM25	0.2524	0.0805	0.1537	0.2004
$BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.2698 +6.89%*	0.0935 16.15%*	0.1628 +5.92%*	0.2067 +3.14%
$BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.2748** **+8.87%***	**0.0953** **+18.39%***	**0.1645** **+7.03%***	**0.2083** **+3.94%**

Table 3. Evaluation results on the *Description* field on the TREC 2014 CDS task. The difference in percentage is measured against the baseline retrieval model *BM25*. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Model	infNDCG	infAP	MAP	R-Prec
BM25	0.2460	0.0700	0.1440	0.2065
$BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.2751 +11.83%*	**0.0918** **+31.14%***	0.1623 +12.71%*	0.2196 +6.34%*
$BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.2830** **+15.04%***	0.0911 +30.14%*	**0.1661** **+15.35%***	**0.2206** **+6.83%***

baselines. In addition, we use the CDS method proposed in [28] as another baseline.

The parameters k_1 and k_3 of BM25 (See Equation (1)) are set to default values and b is set to the optimal value on training data by grid search algorithm [5]. As described in Section 3.1, we adopt two methods for generating embeddings of biomedical articles, which are denoted as $\overrightarrow{d_{Sum}}$ and $\overrightarrow{d_{Para}}$ respectively. For convenience, we denote our proposed CDS method applying *Term Summation* and *Paragraph Embeddings* as $BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$ and $BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$, respectively. Besides, the previously proposed CDS method [28] is denoted as $BM25 + Sim_{d_{Para}\text{-}Q}$, which only uses *Paragraph Em-*

Table 4. Evaluation results on the *Summary* field on the TREC 2015 CDS task. The difference in percentage is measured against the baseline retrieval model *BM25*. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Model	infNDCG	infAP	MAP	R-Prec
BM25	0.2695	0.0736	0.1650	0.2198
$BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.2980 +10.58%*	0.0831 12.91%*	0.1758 +6.55%*	0.2345 +6.69%*
$BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.2986** **+10.80%***	**0.0842** **+14.40%***	**0.1791** **+8.55%***	**0.2408** **+9.55%***

Table 5. Evaluation results on the *Description* field on the TREC 2015 CDS task. The difference in percentage is measured against the baseline retrieval model *BM25*. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Model	infNDCG	infAP	MAP	R-Prec
BM25	0.2724	0.0733	0.1641	0.2184
$BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.2877 +5.62%*	0.0837 14.19%*	0.1762 +7.37%*	0.2325 +6.46%*
$BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.3016 +10.72%***	**0.0873 +19.10%***	**0.1806 +10.05%***	**0.2370 +8.52%***

Table 6. Comparison between our approach and $BM25 + Sim_{d_{Para}\text{-}Q}$ [28] on the TREC 2014 CDS task. The results are obtained based on the *Summary* field.

Method	infNDCG	infAP	MAP	R-Prec
$BM25 + Sim_{d_{Para}\text{-}Q}$	0.2618	0.0763	0.1579	0.1518
$BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.2698 +3.06%	0.0935 +22.54%*	0.1628 +3.10%	0.2067 +36.17%*
$BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.2748 +4.97%***	**0.0953 +24.90%***	**0.1645 +4.18%**	**0.2083 +37.22%***

beddings for generating embeddings of biomedical articles. Note that our method has the following tunable parameters, i.e. hyper-parameter λ (see Equation (10)), top k terms to generate embeddings of biomedical articles when applying *Term Summation* (*# Terms*) and top k articles in $D^k_{PDF}(Q)$ (*# PRF Documents*). All the parameters are tuned on training data by grid search algorithm [5].

The evaluation results are obtained by a two-fold cross-validation, where the topics are split into two equal-size subsets by parity in odd or even topic numbers. In each fold, we use one subset of the topics for training, and the remaining subset for testing. There is no overlap between the training and testing topics. Then the overall retrieval performance is obtained by averaging over the two test subsets of topics. Apart from the official TREC measure inferred NDCG (infNDCG) [24], we also report on other popular evaluation metrics in the CDS task, including Mean Average Precision (MAP) [8], R-Precision (R-Prec) [8]

Table 7. Comparison between our approach and $BM25 + Sim_{d_{Para}\text{-}Q}$ [28] on the TREC 2015 CDS task. The results are obtained based on the *Summary* field.

Method	infNDCG	infAP	MAP	R-Prec
$BM25 + Sim_{d_{Para}\text{-}Q}$	0.2742	0.0657	0.1642	0.1491
$BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.2980 +8.68%*	0.0831 +26.48%*	0.1758 +7.06%*	0.2345 +57.28%*
$BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.2986 +8.90%***	**0.0842 +28.16%***	**0.1791 +9.07%***	**0.2408 +61.50%***

Table 8. Comparison to *SNUMedinfo*, the best automatic run in the TREC 2014 CDS task. Results of *SNUMedinfo* are taken from those reported in [7]. $BM25+SEM_{d\text{-}D^k_{PRF}}$ is the best result of our approach on this dataset, as in Table 3. No statistical test is conducted due to unavailability of the per-query result of *SNUMedinfo*.

Method	infNDCG	infAP
SNUMedinfo	0.2674	0.0659
$BM25 + SEM_{d\text{-}D^k_{PRF}}$	**0.2830**	**0.0911**

Table 9. Comparison to *WSU-IR*, the best automatic run in the TREC 2015 CDS task. Results of *WSU-IR* are taken from those reported in [3]. $BM25 + SEM_{d\text{-}D^k_{PRF}}$ is the best result of our approach on this dataset, as in Table 5. The difference between the two approaches is not statistically significant.

Method	infNDCG	infAP
WSU-IR	0.2939	0.0842
$BM25 + SEM_{d\text{-}D^k_{PRF}}$	**0.3016** **+2.62%**	**0.0873** **+3.68%**

and inferred Average Precision (infAP) [24]. All statistical tests are based on the t-test at the 0.05 significance level.

5 Evaluation Results

In this section, we present the evaluation results of our proposed CDS method. Tables 2 and 3 present the evaluation results of the TREC 2014 CDS task using the *Summary* and *Description* fields respectively, and Tables 4 and 5 present the evaluation results of the TREC 2015 CDS task A. Note that all the evaluation results are obtained by a two-fold cross-validation based on the parity of the topic numbers. As described in Section 4.2, $BM25 + SEM_{d_{Para}\text{-}D^k_{PRF}}$ and $BM25 + SEM_{d_{Sum}\text{-}D^k_{PRF}}$ denote two different applications of our proposed CDS method, in which the embeddings of biomedical articles are generated by *Paragraph Embeddings* and *Term Summation*, respectively. Tables 6 and 7 present the comparison between our CDS method and the $BM25 + Sim_{d_{Para}\text{-}Q}$ method proposed in [28]. BM25 is the baseline retrieval model used for verifying the effectiveness of our proposed feedback-based semantic relevance score. In addition, the comparisons between our approach and the best methods in the TREC 2014 & 2015 CDS tasks are presented in Tables 8 and 9, respectively. According to the results, we have the observations as follows.

First, our proposed feedback-based CDS method has statistically significant improvements over the baseline retrieval model BM25 in most cases, which indicates the effectiveness of integrating semantic evidence into the frequency-based statistical models. Besides, according to Tables 8 and 9, our CDS method outscores the best automatic methods in both TREC 2014 and 2015 CDS tasks. This observation is promising in that a simple linear interpolation of the classical

BM25 model and our proposed semantic relevance score could have scored the best run in those tasks.

Second, according to Tables 6 and 7, our CDS method outperforms the method $BM25 + Sim_{d_{Para}\text{-}Q}$ proposed in [28], which integrates semantic evidence by measuring the cosine similarity between the embeddings of the patient record and biomedical article. As described in Section 1, patient records are much shorter than full-text biomedical articles, such that patient record is a weak indicator of relevance, thus our feedback-based CDS method is expected to outperform the method $BM25 + Sim_{d_{Para}\text{-}Q}$.

Third, comparing the two different ways of generating the article embeddings, *Term Summation* has a better performance than *Paragraph Embeddings* in most cases. As the full-text biomedical articles are usually very long, which contain large amount of irrelevant information, the mechanism of *Paragraph Embeddings* that considering the entire verbose texts while training embeddings may results in the sparse distribution of the semantic information in the embeddings of articles, such that the embeddings of articles generated by *Paragraph Embeddings* is not suitable to represent semantic relevance for long texts. In contrast, *Term Summation* generates embeddings of biomedical articles by only considering the top-k most informative words in the articles, which effectively reduces irrelevant information in the embeddings of biomedical articles.

Finally, comparing between the evaluation results obtained by using the *Summary* and *Description* fields, although using the *Description* field as queries obtained worse baseline retrieval results, the final performance of using *Description* field by integrating semantic evidence is better than using the *Summary* field in most cases. One possible reason is that the *Description* field is much longer than the *Summary* field, such that the relevant biomedical articles are returned by content-based retrieval models with relatively low ranking. By integrating the semantic evidence of relevance, the lowly ranked relevant documents are promoted in the ranking list which leads to improved retrieval performance.

Table 10. The evaluation results on the TREC 2015 CDS task A - automatic. The difference in percentage is measured against the baseline retrieval model *wsuirdaa* [3]. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Method	infNDCG	infAP	MAP	R-Prec
wsuirdaa	0.2939	0.0842	0.1864	0.2306
$wsuirdaa + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.3130 +6.50%*	0.0896 +6.41%*	0.1905 +2.20%	0.2396 +3.90%
$wsuirdaa + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.3157** **+7.42%***	**0.0898** **+6.65%***	**0.1926** **+3.33%**	**0.2469** **+7.07%***

Table 11. The evaluation results on the TREC 2015 CDS task A - manual. The difference in percentage is measured against the baseline retrieval model *wsuirdma* [3]. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Method	infNDCG	infAP	MAP	R-Prec
wsuirdma	0.3109	0.0880	0.1968	0.2493
$wsuirdma + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.3265 +5.02%*	0.0940 +6.82%*	0.2015 +2.39%	0.2605 +4.49%
$wsuirdma + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.3335 +7.27%***	**0.0963 +9.43%***	**0.2054 +4.37%**	**0.2643 +6.02%***

Table 12. The evaluation results on the TREC 2015 CDS task A - automatic. The difference in percentage is measured against $wsuirdaa + SEM_{d_{LDA}\text{-}D^k_{PRF}}$. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Method	infNDCG	infAP	MAP	R-Prec
$wsuirdaa + SEM_{d_{LDA}\text{-}D^k_{PRF}}$	0.2963	0.0853	0.1864	0.2306
$wsuirdaa + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.3130 +5.64%*	0.0896 +5.04%*	0.1905 +2.20%	0.2396 +3.90%
$wsuirdaa + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.3157 +6.55%***	**0.0898 +5.28%***	**0.1926 +3.33%**	**0.2469 +7.07%***

6 Application of the Semantic Relevance Score to Other State-of-the-art Methods

In this section, we use the best TREC run in 2015, WSU-IR, as the baseline to examine if our proposed method can still improve over the strongest baseline as far as we are aware of. We do not conduct the same comparison to *SNUMedinfo*, the best TREC CDS run in 2014, due to unavailability of per-query results. In addition, the latent Dirichlet allocation (LDA) model [6] is applied to generate the distributed representations of biomedical articles for comparison with the neural embedding model Word2Vec in our study.

Table 13. The evaluation results on the TREC 2015 CDS task A - manual. The difference in percentage is measured against $wsuirdma + SEM_{d_{LDA}\text{-}D^k_{PRF}}$. A statistically significant difference is marked with a *. The best result of each evaluation metric is in **bold**.

Method	infNDCG	infAP	MAP	R-Prec
$wsuirdma + SEM_{d_{LDA}\text{-}D^k_{PRF}}$	0.3117	0.0887	0.1970	0.2494
$wsuirdma + SEM_{d_{Para}\text{-}D^k_{PRF}}$	0.3265 +4.75%*	0.0940 +5.98%*	0.2015 +2.28%	0.2605 +4.45%
$wsuirdma + SEM_{d_{Sum}\text{-}D^k_{PRF}}$	**0.3335 +6.99%***	**0.0963 +8.57%***	**0.2054 +4.26%**	**0.2643 +5.97%***

Tables 10 and 11 present the evaluation results based on the automatic and manual runs submitted by WSU-IR [3] in the TREC 2015 CDS Task A, respectively. *wsuirdaa* and *wsuirdma* in Tables 10 and 11 are the submitted automatic and manual runs respectively, and are used as our strong baselines. $wsuirdaa + SEM_{d_{Para}\text{-}D^k_{PRF}}$ and $wsuirdaa + SEM_{d_{Sum}\text{-}D^k_{PRF}}$ correspond to applying *Paragraph Embeddings* and *Term Summation* respectively, the same to $wsuirdma + SEM_{d_{Para}\text{-}D^k_{PRF}}$ and $wsuirdma + SEM_{d_{Sum}\text{-}D^k_{PRF}}$. According to the results, we can see that there are still statistically significant improvements over the strong baselines *wsuirdaa* and *wsuirdma* in most cases when applying both *Paragraph Embeddings* and *Term Summation*, indicating the effectiveness of our proposed semantic relevance score.

Tables 12 and 13 present the comparison between Word2Vec and LDA in generating the distributed representations of biomedical articles. The number of topics in LDA is set to 100 as used in [13]. $wsuirdaa + SEM_{d_{LDA}\text{-}D^k_{PRF}}$ and $wsuirdma + SEM_{d_{LDA}\text{-}D^k_{PRF}}$ in Tables 12 and 13 correspond to applying LDA model for generating the article representations, on top of the best TREC CDS runs in 2015. According to the results, there are statistically significant improvements over LDA in most cases when Word2Vec is utilized to generate the article embeddings. In fact, our experience indicates that the optimal value of hyperparameter λ (See Equation (10)) when applying LDA model is usually 1, such that the semantic relevance score does not work when LDA is used. Therefore, we may conclude that Word2Vec is more suitable than LDA for estimating the semantic similarity between biomedical articles.

7 Conclusions and Future work

In this paper, we have proposed a novel feedback-based CDS method, which integrates the semantic similarity between a biomedical article and the corresponding pseudo relevance feedback set into frequency-based models. Experimental results show that integrating semantic evidence of relevance can indeed significantly improve the retrieval performance over the existing CDS approaches, including the best TREC results. In addition, a simple linear combination of the classical BM25 model with our proposed semantic relevance score ($BM25 + SEM_{d\text{-}D^k_{PRF}}$) would have achieved the best automatic runs on the TREC 2014 and 2015 CDS tasks. Compared to *Paragraph Embeddings*, *Term Summation* is more suitable to generate the embeddings of biomedical articles, due to the ability of reducing irrelevant information in the embeddings of biomedical articles. The comparison between Word2Vec and LDA shows that Word2Vec is more suitable than LDA for estimating the semantic similarity between biomedical articles.

In future research, we plan to utilize the semantic evidence for query expansion to further improve the performance of a CDS system.

Acknowledgments: This work is supported by the National Natural Science Foundation of China (61472391). We would like to thank the authors of [3] for kindly sharing their TREC runs with us.

Bibliography

[1] A. Abacha and S. Khelifi. LIST at TREC 2015 Clinical Decision Support Track: Question Analysis and Unsupervised Result Fusion. In *TREC*, 2015.

[2] A. Dai, C. Olah, and Q. Le. Document Embedding with Paragraph Vectors. *CoRR*, abs/1507.07998, 2015.

[3] S. Balaneshinkordan, A. Kotov, and R. Xisto. WSU-IR at TREC 2015 Clinical Decision Support Track: Joint Weighting of Explicit and Latent Medical Query Concepts from Diverse Sources. In *TREC*, 2015.

[4] Y. Bengio, H. Schwenk, J. Senécal, F. Morin, and J. Gauvain. *Neural Probabilistic Language Models*, pages 137–186. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[5] J. Bergstra, R. Bardenet, B. Kgl, and Y. Bengio. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

[6] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] S. Choi and J. Choi. SNUMedinfo at TREC CDS Track 2014: Medical Case-based Retrieval Task. Technical report, DTIC Document, 2014.

[8] G. Chowdhury. TREC: Experiment and evaluation in information retrieval. *Online Information Review*, (5), 2007.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural Language Processing (Almost) from Scratch . *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[10] R. Cummins. Clinical Decision Support with the SPUD Language Model. In *TREC*, 2015.

[11] R. Cummins, J. Paik, and Y. Lv. A pólya Urn Document Language Model for Improved Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 33(4):21, 2015.

[12] Y. Goldberg and O. Levy. word2vec Explained: Deriving Mikolov et al.s Negative-Sampling Word-Embedding Method. *CoRR*, abs/1402.3722, 2014.

[13] T. Goodwin and S. Harabagiu. UTD at TREC 2014: Query Expansion for Clinical Decision Support. Technical report, DTIC Document, 2014.

[14] Q. Le and T. Mikolov. Distributed Representations of Sentences and Documents. *CoRR*, abs/1405.4053, 2014.

[15] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[16] D. Metzler and W. Croft. A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM, 2005.

[17] T. Mikolov, W. Yih, and G. Zweig. Linguistic Regularities in Continuous Space Word Representations. *In HLT-NAACL*, 2013.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013.

[19] A. Mnih and G. Hinton. A Scalable Hierarchical Distributed Language Model. In *Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December*, pages 1081–1088, 2008.

[20] J. Palotti and A. Hanbury. TUW @ TREC Clinical Decision Support Track 2015. In *TREC*, 2015.

[21] J. Ponte and W. Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[22] K. Roberts, M. Simpson, E. Voorhees, and W. Hersh. Overview of the TREC 2015 Clinical Decision Support Track. In *TREC*, 2015.

[23] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. *TREC*, pages 73–96, 1996.

[24] M. Simpson, E. Voorhees, and W. Hersh. Overview of the TREC 2014 Clinical Decision Support Track. Technical report, DTIC Document, 2014.

[25] Y. Song, Y. He, Q. Hu, and L. He. ECNU at 2015 CDS Track:Two Re-ranking Methods in Medical Information Retrieval. In *TREC*, 2015.

[26] J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[27] I. Vulić and M. Moens. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In *The International ACM SIGIR Conference*, pages 363–372, 2015.

[28] C. Yang and B. He. A Novel Semantics-based Approach to Medical Literature Search. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 1616–1623. IEEE, 2016.