Graph-Based Multi-Modality Learning for Clinical Decision Support

ABSTRACT

The task of clinical decision support (CDS) involves retrieval and ranking of medical journal articles for medical records of diagnosis, test or treatment. Previous studies on this task are based on bag-of-words representations of document texts and general retrieval models. In this paper, we propose to use the paragraph vector technique to learn the latent semantic representation of texts and treat the latent semantic representations and the original bag-of-words representations as two different modalities. We then propose to use the graph-based multi-modality learning algorithm for document re-ranking. Experimental results on two TREC-CDS benchmark datasets demonstrate the excellent performance of our proposed approach.

CCS Concepts

• Information systems → Information systems applications → Enterprise information systems • Information systems → Information retrieval → Specialized information retrieval

Keywords

Clinical Decision Support; Paragraph Vector; Graph-based Multi-Modality Learning; TREC.

1. INTRODUCTION

The task of Clinical Decision Support (CDS)[1] is designed to assess the ability of search engines to retrieve biomedical journal articles relevant for answering generic clinical questions about medical records[10][11]. Each topic in this track consists of a sentence-long summary and a paragraph-long description of a patient case, along with one of three types of clinical information: diagnosis (i.e., "*What is this patient's diagnosis?*"), test ("*What diagnostic test is appropriate for this patient?*") or treatment ("*What treatment is appropriate for this patient?*"), which is likely to support physicians and other health professionals with clinical decision-making tasks by medical records such as a list of symptom, a treatment plan and a particular test. The CDS task aims to retrieve articles to particular case based on the summary, description, or both. The task has become one important evaluation track on TREC 2014 and TREC 2015.

CDS is a newly established task and it is considered a special kind of document retrieval task. Most previous studies adopted existing popular retrieval models for addressing this task [10][11]. In many retrieval models, texts are required to be represented as a feature vector and one of the most common representation is bag-of-words (BOW) [7]. Despite its popularity, BOW model casts off the order of words and thus cannot reflect the genuine semantics of word sequences. Moreover, BOW model cannot deal with synonymy or polysemy. There is a big semantic gap between topic and document based on the BOW model, and there is also a big document-to-document semantic gap.

In recent years, a variety of neural network methods [12][13] have been proposed to map words, paragraphs or documents to dense vectors which are considered to embody their latent semantics. In this study, we investigate leveraging the Paragraph Vector model [13] to map texts into latent semantic representations. After that, we have two different representations (i.e., BOW representations and latent semantic representations) for each topic and each document. We then propose to use the graph-based multi-modality learning algorithm for document re-ranking by considering the two different representations as two modalities.

Evaluation results on two benchmark datasets (TREC CDS 2014 and TREC CDS 2015) show that our proposed method can significantly improve the retrieval performance, and the overall performance is very competitive as compared with the TREC participating systems.

The contributions of this paper are summarized as follows:

1) We propose a new method consisting of three steps for addressing the CDS task.

2) We propose to leverage Paragraph Vector to learn the latent semantics of biomedical journal articles.

3) We propose to use the graph-based multi-modality learning algorithm for document re-ranking in the CDS task.

4) Evaluation results on two benchmark datasets are very competitive.

2. OUR PROPOSED METHOD

2.1 Overview

Our proposed method for the CDS task consists of three steps: initial document retrieval, latent semantic learning and multi-modality re-ranking. The first step aims to retrieve a small set of candidate documents from the large corpus by using typical retrieval models. The second step aims to learn the latent semantic representations of the documents and paragraphs. The last step aims to re-rank the candidate documents by using the graph-based multi-modality learning algorithm. We can see that the first step helps to largely reduce the document set to be re-ranked, and thus it can improve the efficiency of the re-ranking process. Moreover, the candidate document set retrieved in the first step contain less noisy documents and thus it is beneficial to the graph-based re-ranking results. The second step is used to obtain another different representation of each document or paragraph, besides the conventional BOW representation. After that, we have two representations for each document or paragraph, and these two representations can be considered two modalities, which are then used in the third step. The details of the three steps will be described in next sections, respectively.

2.2 Initial Document Retrieval

For initial document retrieval, we use the LemurTFIDF model [2] as retrieval model. In this model, a word's TF score is computed

[1] http://www.trec-cds.org/

with the Okapi TF formula [15] and then a document or topic description is represented by a TFIDF-based term vector. Note that we use the concatenation of the summary and description of a topic to represent the topic and simply use the term "topic description" to denote the concatenated text in this paper. As a preprocessing step for our approach, each word was stemmed using Porter's stemmer. The rank scores of documents are computed as the cosine similarity between the TFIDF vectors of documents and the topic description for each topic.

What's more, we used MeSH[2] [9] to enrich topic description. MeSH is the National Library of Medicine's (NLM) controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. We simply increase the occurrence number of a word in the topic description by one if the word is a MeSH word, and then compute the TF score of this word using the Okapi TF formula for topic enrichment.

In order to obtain a better candidate document set, we further apply the pseudo relevance feedback technique. First, we apply the LemurTFIDF model implemented in Terrier [8], an open source retrieval system, to get prior ranking results. Then, we select Top-*K* documents in the prior ranking results, choose 20 words from each Top-*K* document with highest LemurTFIDF scores, and then add these items to the topic description representation as feedback model. Finally, we rank all documents according to their cosine similarity with the new topic representation. We simply set *K* to 5.

After initial document retrieval, we keep and use only Top-2000 documents of each topic for multi-modality re-ranking.

2.3 Latent Semantic Learning with Paragraph Vector

Paragraph vector [13], which was inspired by word vector [12] and has been proved to be one of the state-of-the-art methods for document modeling, is an unsupervised framework to learn continuous distributed vector representations for pieces of variable-length texts. In the paragraph vector framework, both documents and words are mapped to unique vectors. Each document is treated as a unique token which is the context of all the words in the document.

More formally, given a sequence of training word $w_1, w_2, \dots, w_T$ and the paragraph p, the aim is to maximize the average log probability

$$\frac{1}{T}\sum_{t=k}^{T-k} log\, p(w_t|w_{t-k}, \dots, w_{t+k}, p)$$

In paragraph vector model, the objective can be typically achieved by the softmax function, a generalization of the logistic function. It has the form as follows:

$$p(w_t|w_{t-k}, \dots, w_{t+k}, p) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

Each of y_i is un-normalized log-probability for each output word i, computed as

$$y = b + Uh(w_{t-k}, \dots, w_{t+k}, p; W, D)$$

where U, b are the softmax parameters. h is constructed by a concatenation or average of word vectors and paragraph vector looked up from matrices *W* and *D*.

The paragraph vector model considers the order of word and represents paragraph as a vector in lower-dimension space compared with the BOW vector. Intuitively, the paragraph vector contains latent semantic information and we believe considering this latent vector can improve document ranking results.

In our study, the open source toolkit *word2vec*[3] is used to generate paragraph vectors. This toolkit was firstly developed to implement the word vector model and then modified to obtain paragraph vector. The vector dimension is simply set to 100 and the window size of words is set to 10. The negative sampling method is used and the number of negative samples is set to 5 and the sampling threshold is set to 1e-3. The iteration time is set to 20. All the above parameter values were suggested by Quoc Le[4].

In our study, we segment each document into paragraphs by paragraph breaks (i.e. <*p*> tags) and map each paragraph to a unique vector. Document vectors are generated by simple addition of vectors of paragraphs with normalization. Note that we will use "paragraph vector" to denote the document vector of a document in subsequent sections.

2.4 Graph-Based Multi-Modality Re-Ranking

2.4.1 Basic Manifold-Ranking Algorithm

The manifold-ranking method [3][4] is a universal ranking algorithm and it is initially used to rank data points by their graph-based relationships with the prior assumption that nearby points and points on the same structure (typically referred to as a cluster or a manifold) are likely to have the same ranking scores. The manifold-ranking method has been used for information retrieval, such as summarization, image retrieval and this method has been used for topic-focused multi-document summarization to rank sentences [18][19]. In our study, we use the manifold-ranking method to re-rank the initially retrieved documents. In other words, the data points refer to the topic description and all documents in the candidate set. The manifold-ranking process of the document re-ranking task can be formalized as follows:

Given a set of vectors $\chi = \{x_0, x_1, x_2, \dots, x_n\} \subset R^m$ and these vectors can be either BOW vectors or paragraph vectors. The BOW vectors are generated in the same way of the initial document retrieval process. The only change is that BOW vectors of topics are enriched on the basis of Top-*K* documents of initial retrieval results instead of the ranking results by Terrier. The paragraph vectors are generated by the method we introduce in Section 2.3. The first point x_0 represents the topic description and the rest *n* points represent all the initially retrieved documents (data point to be ranked).

The ranking scores are expressed as a vector $f = [f_0, f_1, \dots, f_n]^T$ and the initial values are expressed as a vector $y = [y_0, y_1, \dots, y_n]^T$ with $y_0 = 1, y_i = 0\ (i \neq 0)$. The similarity relationship is constructed as a matrix $W = (W_{ij})_{(n+1)\times(n+1)}$, in which each W_{ij} corresponds to the cosine similarity between x_i and x_j and $W_{ii} = 0$. $S = D^{-1/2}WD^{-1/2}$ denotes the normalization of matrix W and $D = (D_{ij})_{(n+1)\times(n+1)}$ where $D_{ii} = \sum_{j=0}^{n} W_{ij}$ is a diagonal matrix.

[2] https://www.nlm.nih.gov/mesh/MBrowser.html

[3]https://groups.google.com/d/msg/word2vec/toolkit- Q49FIrNOQRo/J6KG8mUj45sJ

[4] https://groups.google.com/d/msg/word2vec-toolkit/Q49FIrNOQRo/CJLWzmr0LaUJ

According to Zhou et al.'s work, f can be optimized by iterating process $f^{t+1} = \alpha S f^t + (1-\alpha)y$ where $0 < \alpha < 1$. After convergence, $f^* = lim_{t\to\infty} f^t$, and f_i^* stands for the final ranking score of document i.

2.4.2 Multi-Modality Learning Algorithm

The basic manifold-ranking algorithm makes use of the similarity relationships based on uniform vectors. In our study, we expect to take into account not only statistical features but also semantic features during our re-ranking process. Specifically, we wish to use both traditional BOW vectors and paragraph vectors to improve ranking results. The two modalities are very different and they have unique characteristics. Therefore, it would be more appropriate to consider both the two modalities and the multi-modality manifold-ranking algorithm [6] is thus applied for document re-ranking.

In our multi-modality re-ranking process, we construct two matrices W^a and W^b to represent the similarity relationships of documents based on the BOW vectors and the paragraph vectors, respectively. The definitions of W^a and W^b are the same with W in the basic manifold-ranking algorithm. In particular, W^a is built from BOW vectors and W^b is built from paragraph vectors. S^a and S^b are the normalization of W^a and W^b in the same way. By the work of [6][18], there are three iterating schemes for fusing two modalities: linear fusion scheme, sequential scheme and score combination scheme.

Linear fusion scheme:

$$f^{t+1} = \mu S^a f^t + \eta S^b f^t + (1-\mu-\eta)y$$

where $0 < \mu, \eta$ and $\mu + \eta < 1$

$$f^* = lim_{t\to\infty} f^t$$

Sequential fusion scheme:

$$f^{t+1} = \mu S^a f^t + \eta S^b f^t - \mu\eta S^b S^a f^t + (1-\mu)(1-\eta)y$$

where $0 < \mu, \eta < 1$

$$f^* = lim_{t\to\infty} f^t$$

Score combination scheme:

$$f_a{}^{t+1} = \mu S^a f_a{}^t + (1-\mu)y$$

$$f_b{}^{t+1} = \eta S^b f_b{}^t + (1-\eta)y$$

$$f_a{}^* = lim_{t\to\infty} f_a{}^t$$

$$f_b{}^* = lim_{t\to\infty} f_b{}^t$$

$$f^* = \lambda f_a{}^* + (1-\lambda) f_b{}^* \text{ where } 0 < \lambda < 1$$

By using the three schemes above, we obtain the re-ranking results by considering both statistical and semantic characteristics.

3. EVALUATION

3.1 Data Set and Evaluation Metrics

In 2014 and 2015, the TREC-CDS track was designed to assess the ability of search engines to retrieve biomedical journal articles relevant for answering generic clinical questions about medical records. There are 30 topics in the track of each year and each topic within this track consists of a sentence-long summary and a paragraph-long description of a patient case, along with one of three types of clinical information: diagnosis, test or treatment. Table 1 gives three example topics in TREC 2014. The top 100 documents in the retrieval results are used for evaluation by evaluation toolkits. We used the two TREC-CDS datasets for evaluation in this study.

In TREC-CDS 2015, there are two tasks (task A and task B) with the topics. The only difference is that the topic descriptions of task B contain the description of diagnosis. In order to keep the consistency of two datasets, we only experimented with task A in TREC-CDS 2015.

We used *trec_eval*[5] and *sample_eval*[6] toolkits for evaluation, which was officially adopted by TREC-CDS. We used infNDCG [5] as main metric and also provided results of infAP, Precision@10 and R-Prec measures.

Table 1. Example topic summaries from TREC 2014

Topic	Type	Summary
1	Diagnosis	58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.
11	Test	40-year-old woman with severe right arm pain and hypotension. She has no history of trauma and right arm exam reveals no significant findings.
21	Treatment	21-year-old female with progressive arthralgias, fatigue, and butterfly-shaped facial rash. Labs are significant for positive ANA and anti-double-stranded DNA, as well as proteinuria and RBC casts.

3.2 Evaluation Results

In the experiments, the proposed multi-modality re-ranking methods with the three fusion schemes introduced in Section 2.4 are denoted as "DoubleM(SEQ)", "DoubleM(LIN)" and "DoubleM(COM)". The basic manifold-ranking methods based on BOW vectors and paragraph vectors are denoted as "SingleM(BOW)" and "SingleM(P2V)", and the baseline corresponds to the initial document retrieval results. In the experiments, the regularization parameter μ for the BOW modality is fixed at 0.65 and parameter η for the paragraph vector modality is fixed at 0.15. Therefore, we set $\alpha = 0.65$ for "SingleM(BOW)", $\alpha = 0.15$ for "SingleM(P2V)", $\mu = 0.65$ and $\eta = 0.15$ for all multi-modality re-ranking methods. The combination weight parameter λ is simply set to 0.5 for the score combination scheme.

To compare with state-of-the-art methods, we also list the two best results of automatic participants [10][11]. In 2014, they are SNUMedinfo6 [17] and NovaSearch [9]. In 2015, they are LIST_LUX [1] and CAMspud1 [14]. Table 2 shows the comparison results (infNDCG) of different re-ranking schemes and Table 3 shows the comparison results between our results and the best results of automatic participants.

Seen from the tables, the basic manifold-ranking algorithm performs as well as other top-performing retrieval methods and the proposed multi-modality algorithms can significantly outperform the baseline and "SingleM" methods. The proposed method with the sequential fusion scheme (i.e. "DoubleM(SEQ)") performs the best on the TREC-CDS 2014 and 2015 datasets and it performs better than all automatic participants in TREC-CDS over the main metric.

[5] http://trec.nist.gov/trec_eval/index.html

[6] http://trec.nist.gov/data/clinical/sample_eval.pl

Table 2. Comparison results of different re-ranking schemes

	TREC-CDS 2014	TREC-CDS 2015
Baseline	0.2357	0.2350
SingleM(BOW)	0.2786	0.2703
SingleM(P2V)	0.2675	0.2450
DoubleM(LIN)	0.2790	0.2689
DoubleM(COM)	0.2950	0.2903
DoubleM(SEQ)	**0.2967**	**0.2957**

Table 3. Comparison results between our results and best automatic participants' results in 2014 ("n/a" means the performance value has not been reported in literatures)

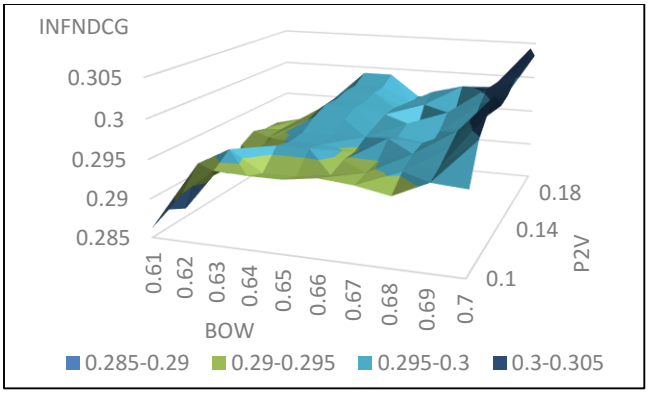
	infNDCG	infAP	Prec10	R-Prec
SNUMedinfo6	0.2674	0.0659	0.3633	n/a
NovaSearch	0.2631	0.0757	**0.3900**	0.2165
SingleM(BOW)	0.2786	0.0870	0.3533	0.2140
DoubleM(SEQ)	**0.2967**	**0.0934**	0.3633	**0.2275**

Table 4. Comparison results between our results and best automatic participants' results in 2015

	infNDCG	infAP	Prec10	R-Prec
Run2DBpComb	0.2894	**0.0787**	0.4533	**0.2299**
CAMspud1	0.2823	0.0758	n/a	n/a
SingleM(BOW)	0.2703	0.0636	0.4667	0.2062
DoubleM(SEQ)	**0.2957**	0.0713	**0.480**0	0.2162

In order to investigate the influence of parameters in multi-modality re-ranking, the parameter μ is varied from 0.61 to 0.7 and the parameter η is varied from 0.11 to 0.2 in "DoubleM(SEQ)" on the two datasets. Figures 1 and 2 show the surface charts on the two datasets and the BOW axis stands for μ and the P2V axis stands for η. We can see from the figures that both two modalities contribute to the improvement of the results, and our proposed method can achieve very good performance over a relatively wide range of parameter values, which demonstrates the robustness of our proposed method.

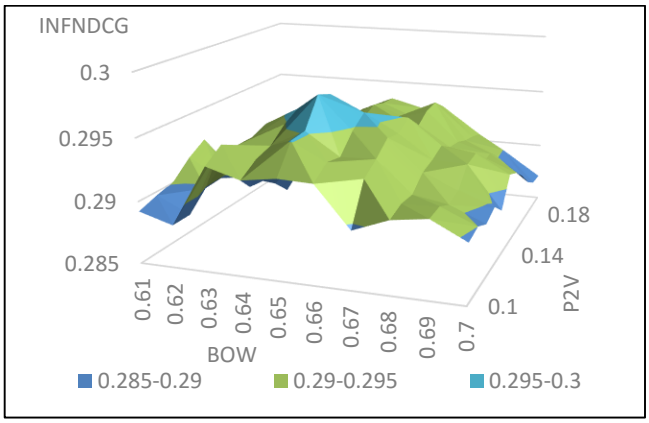
Figure 1. InfNDCG vs. μ and η for "DoubleM(SEQ)" in 2014



4. CONCLUSIONS

In this paper, we propose to use the paragraph vector technique and the multi-modality ranking algorithm for addressing the task of clinical decision support. The results on two benchmark datasets are very impressive.

Figure 2. InfNDCG vs. μ and η for "DoubleM(SEQ)" in 2015



5. REFERENCES

[1] Asma Ben Abacha and Saoussen Khelifi. LIST at TREC 2015 Clinical Decision Support Track: Question Analysis and Unsupervised Result Fusion. In Notebooks of TREC2015.

[2] Chengxiang Zhai. Notes on the Lemur TFIDF model. http://www.cs.cmu.edu/~lemur/1.0/tfidf.ps

[3] D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. SchÖlkopf. Ranking on data manifolds. In Proceedings of NIPS-03.

[4] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. SchÖlkopf. Learning with local and global consistency. In Proceedings of NIPS-03.

[5] Emine Yilmaz, Evangelos Kanoulas, Javed A. Aslam. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In SIGIR'08, July 20–24, 2008, Singapore.

[6] Hanghang Tong, Jingrui He, Mingjing Li, Changshui Zhang, Wei-Ying Ma. Graph Based Multi-Modality Learning. In MM'05, November 6–11, 2005, Singapore.

[7] Harris, Zellig. Distributional structure. Word, 1954.

[8] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, et al. Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings of OSIR 2006.

[9] J. Gobeill, A. Gaudinat, E. Pasche, P. Ruch. Full-texts representation with Medical Subject Headings, and co-citations network reranking strategies forTREC 2014 Clinical Decision Support Track. In Proceedings of TREC 2014.

[10] Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, William R. Hersh. Overview of the TREC 2015 Clinical Decision Support Track. In Notebooks of TREC 2015.

[11] Matthew S. Simpson, Ellen M. Voorhees, William Hersh. Overview of the TREC 2014 Clinical Decision Support Track. In Proceedings of TREC 2014.

[12] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Distributed representations of phrases and their compositionality. In Advances on Neural Information Processing Systems, 2013.

[13] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014.

[14] Ronan Cummins. Clinical Decision Support with the SPUD Language Model. In Notebooks of TREC 2015.

[15] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In TREC-8.

[16] S. E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford . Okapi at TREC-3. In TREC-3.

[17] Sungbin Choi, Jinwook Choi. SNUMedinfo at TREC CDS track 2014: Medical case-based retrieval task. In Proceedings of TREC 2014.

[18] Xiaojun Wan and Jianguo Xiao. Graph-Based Multi-Modality Learning for Topic-Focused Multi-Document Summarization. In IJCAI 2009.

[19] Xiaojun Wan, Jianwu Yang and Jianguo Xiao. Mani- fold-ranking based topic-focused multi-document summarization. In Proceedings of IJCAI-07.