Can Word Embedding Help Term Mismatch Problem? – a Result Analysis on Clinical Retrieval Tasks

Danchen Zhang, Daqing He
University of Pittsburgh
{daz45, dah44}@pitt.edu

Abstract. Clinical Decision Support (CDS) systems assist doctors to make clinical decisions by searching for medical literature based on patients' medical records. Past studies showed that correctly predicting patient's diagnosis can significantly increase the performance of such clinical retrieval systems. However, our studies showed that there are still a large portion of relevant documents ranked very low due to term mismatch problem. Different to other retrieval tasks, queries issued to this clinical retrieval system have already been expanded with the most informative terms for disease prediction. It is therefore a great challenge for traditional Pseudo Relevance Feedback (PRF) methods to incorporate new informative terms from top K pseudo relevant documents. Consequently, we explore in this paper word embedding for obtaining further improvements because the word vectors were all trained on much larger collections and they can identify words that are used in similar contexts. Our study utilized test collections from the CDS track in TREC 2015, trained on 2014 data. Experiment results show that word embedding can significantly improve retrieval performance, and term mismatch problem can be largely resolved, particularly for the low ranked relevant documents. However, for highly ranked documents with less term mismatching problem, word emending's improvement can also be replaced by a traditional language model.

Keywords: Clinical decision support, word embedding, term mismatch.

1 Introduction

During their clinical decision making process, doctors often consult external literature for reference. The published biomedical literature, which contains expert written materials on nearly all topics in the medical area, are the most common source of reference [1]. TREC has hold Clinical Decision Support (CDS) track since 2014 to support medical text retrieval, based on which we proposed a diagnosis prediction enhanced retrieval model (MRF-Wiki) [3, 4], which outperforms the state-of-art models.

However, our MRF-Wiki model still has room to improve, because many relevant articles are either lowly ranked or not returned at all. Different to other retrieval tasks, queries in disease prediction-based retrieval systems, such as ours, have already included the most informative terms with the help of predicted diseases. It is hard for traditional Pseudo Relevance Feedback (PRF) model to find new and more informative words to expand the original query, not to mention its topic drift problem [7].

In this paper, we will firstly examine the failure cases in the current model and particularly explore the effects from term mismatch, which is a common problem in retrieval tasks [5, 6]. In medical domain, the term mismatch problem can appear like this. Disease in the query might not appear in the relevant document, and the virus causing the disease could be an important evidence for making the document relevant. But they cannot be matched with traditional language models.

In this situation, the word embedding model, trained on the large collections, can identify words that are used in similar contexts with respect to a given word [6]. It is expected that word embedding could introduce a full list of reasonably weighted new words closely relevant to the query terms, which might help in resolving term mismatch problem in medical domain.

Consequently, we try to answer the following research questions in this study: 1) To what extent does term mismatch problem affect the diagnosis-based clinical text retrieval models? 2) In what situation can word embedding model solve term mismatch problem? And 3) What are the problems still limiting the system enhanced by word embedding?

2 Related works

Pseudo Relevance Feedback extracts new informative terms from top ranked documents, and it is one of the traditional ways to resolve word mismatch problem in information retrieval. In CDS task, majority previous studies utilized PRS to enhance the original queries. For example, Limsopatham et al. [2] explored collecting terms from different knowledge sources to expand the query. Choi et al. [8], the best run in CDS 2014, utilized the most frequent Medical Subject Heading label terms inside the PRF documents to expand the original query. Balaneshin-kordan et al. [9] expanded the queries with terms selected from both PRF documents and Google search results. All these methods of using PRF significantly improved the retrieval performance. However, different to these systems, queries in our disease prediction based clinical retrieval system have already included the most informative terms (i.e., predicted disease), which makes it challenging for PRF to find new and more informative terms.

Word embedding models, which can leverage the underlying word semantic similarities, have been widely used in information retrieval [5, 6, 10, 11]. Zhou et al. [5] demonstrated that word embedding can significantly improve the performance of a question-answer system by alleviating term mismatch problems. Ganguly et al. [6] proposed a word embedding-based word transformation model to address the term mismatch problem. In this study, we want to explore, after the most informative words have already existed in the query, whether word embedding can further improve the retrieval performance by solving the remaining term mismatch problems.

3 Methods

To answer the above-mentioned research questions, we conducted retrieval experiments using TREC CDS 2014 and 2015 collections. The three retrieval models employed in the study are presented below.

3.1 Baseline: Markov Random Field with Wikipedia Based Diagnosis Prediction (MRF-Wiki) model

MRF-Wiki was a model we proposed in [3, 4]. The patient's disease related information is extracted from the topic, and a Wikipedia based disease predictor model is called to predict the patient's disease diagnosis, which is used to expand the query generated by Markov Random Field (MRF). The query can be written in the form of Indri query language as in Formula (1). This model is the baseline in this study.

#weight ((1-α) #combine (MRF query) α #combine (predicted diagnosis)) (1)

3.2 PRF Enhanced Document Ranking (PRF-DR) model

The PRF model used in this study is Relevance Model 3 (RM3), a classic PRF model presented in [12]. RM3 selects the most informative terms form topK documents, and each term is weighted by their importance. It can be written in Indri query language:

#weight ((1- β) #combine (MRF-WIKI query) β #combine (weighted terms)) (2)

3.3 Word Embedding Enhanced Document Ranking (WE-DR) model

Trained on large collections, word embedding models can learn high-quality dense word vector representations from the contextual information of the word. These dense word vectors keep the semantic relationships, which provide us the basis for resolving the term mismatch problem.

In our WE-DR model, only the title and keywords parts are used to represent the whole document. This is because these two parts contain the most informative terms and disclose the document's main topics. Through analyzing MRF-WIKI queries, we found that patient's diagnosis information is more important than symptom information because nearly all relevant documents contain the diagnosis but usually do not mention the symptoms. Thus, predicted diagnosis terms were used as the surrogate of the query in the enhancement with word embedding. For a text (i.e., query or document) with n terms, vectors are calculated as the averaged accumulated word vectors:

$$text\_vec = \frac{1}{n}\sum_{term \in text} word\_vec \quad (3)$$

Cosine similarity is commonly chosen to evaluate the association between two vectors [10, 11]. We used it here for calculating the similarity between query and document. The relevance score of the document is combined by the score in MRF-Wiki model and the cosine similarity, which are combined with a parameter γ:

$$Score(d) = (1-\gamma)\, MRF\text{-}WIKI(d) + \gamma\, Cos(doc\_vec, query\_vec) \quad (4)$$

We used a collection of pre-trained word vectors of 200 dimensions that were generated using skip-gram model with a window size of 5 on PubMed and PMC texts [13]. The collection contains vectors for 2,515,686 words.

4 Experiments and Discussions

4.1 Dataset and Metrics

In this study, the target collection is a corpus of 744,138 PubMed articles, published in TREC 2014 CDS track. It was preprocessed with stop word removal and Porter stemming, and was indexed with Indri. We used the 30 topics from TREC 2014 CDS track to train the models, and used 30 topics from TREC 2015 CDS track for testing. In this study, α *is set 0.5;* β *is set 0.5, and* γ *is set 0.25. For RM3 model,* 3 most important terms from top 5 retrieved documents are expanded to MRF-WIKI.

Following TREC CDS track, the evaluation metrics we used include infNDCG (inferred Normalized Discounted Cumulative Gain), P@10, and MAP.

4.2 Results: Impacts of Word Embedding

WSU-IR [14] was the best performed model in CDS track 2015, and the three methods mentioned in Section 3 (i.e., MRF-WIKI, PRF-DR and WE-DR) all achieved better infNDCG than it. But only WE-DR has higher P@10 and MAP that WSU-IR (see Table 1). As only a final performance is provided for WSU-IR [14], we cannot conduct significant test for further comparation. However, using Wilcoxon Signed Ranks Test to examine among our three models, we find that WE-DR significantly outperforms MRF-WIKI on infNDCG, and significantly outperforms both MRF-WIKI and PRF-DR on P@10 and MAP (p-value<0.05). PRF-DR shows no significant difference from the baseline MRF-WIKI in all three metrics.

Table 1. Performance comparison on CDS 2015 task. * indicates significantly outperform MRF-WIKI; ** means significantly outperform MRF-WIKI and PRF-DR;

	infNDCG	P@10	MAP
WSU-IR [4]	29.39%	46.67%	18.64%
MRF-WIKI	31.04%	42.33%	18.61%
PRF-DR	32.71%	44.67%	17.74%
WE-DR	32.26%*	49.67%**	19.52%**

4.3 Results: Analysis of Term Mismatch Problem

We performed manual analysis on the results of six topics to further explore the effect of word embedding model, and the six topics are randomly selected from the topics with correct diagnosis prediction. During the analysis, we selected three groups of

documents. Firstly, we selected top ranked relevant documents using those relevant documents appearing within the top 10 ranks. Secondly, we then selected low ranked relevant documents by identifying the last 5 relevant documents from ranks 500 to 1000. Finally, for each topic, we randomly selected 5 relevant but not returned (i.e., false negative) documents. There are totally 98 documents selected from these topics, as shown in Table 2.

Table 2. Selected 98 documents in result analysis

	Doc Count	Term mismatch affected Docs
Top relevant documents	28	2
Low relevant documents	35	14
False negative documents	35	10
Total documents	98	26

To what extent does term mismatch problem affect the diagnosis based clinical text retrieval system (baseline)?

In the selected 98 documents, if a document does not contain the query terms, it is labeled as a document affected by term mismatch. From Table 2 we can see that the number of highly ranked documents affected by this problem is small, but the lowly ranked relevant documents and false negative ones are much commonly affected by term mismatch frequently appears. Totally we collected 26 term mismatch cases.

In what situation can word embedding model resolve term mismatch problem?

Highly ranked documents. P@10 is significantly improved in WE-DR, implying word embedding helps highly ranked documents. In our selected document set, half of top relevant documents ranking are boosted, while most of the other half ranking stay at the same. After exploration, we find that those boosted documents usually have short title/keywords but the query terms appear several times. In contrast, those few declined documents usually have long title/keywords and the query terms only appear one time. This implies that embedding model boosts the documents with more query terms, so it works like a traditional language model in highly ranked documents, where term mismatch seldom appears.

Lowly ranked relevant documents. Table 2 shows that, among the 14 lowly ranked relevant documents suffering term mismatch problem, 13 of them have their ranks boosted by the word embedding model, The only document with declined ranking has a rhetoric title “How 40 kilograms of fluid retention can be overlooked: two case reports”, even though its main topic is about the diagnosis of “heart failure”. This relevant document is even hard to identify using manual methods. In addition, P@1000 of WE-DR is 8.84%, which significantly outperforms MRF-Wiki’s 8.68% (p-value<0.05). This indicates that the embedding model can improve the lowly ranked document ranking by resolving the word mismatch problem.

What are the problems still limiting the system enhanced with word embedding?

We further analyzed the non-relevant documents ranked within the top 10, and identified three main reasons. First, **documents discussing irrelevant patient situations**. For example, topic 11 is related to a 56-year-old lady, but some non-

relevant articles talk about patients as pregnant female, male, adolescent, or even animals. Second, **document concerns another disease**. For example, topic 3 seeks documents of "Pulmonary embolism", and some top ranked documents talk about "Pulmonary hypertension", which is a different disease. The last one is **different aspects of disease.** For example, topic 17 seeks for information about what test cervical cancer patient should receive, but some non-relevant articles talk about treatment plan, or study people's attitude towards cervix cancer, which make them non-relevant.

5 Conclusion

In this paper, we presented and examined a diagnosis prediction-based clinical decision support system with a word embedding model. The embedding model aims to resolve the term mismatch problem. Our results show that for highly ranked documents, word emending's improvement can also be replaced by a traditional language model, however, for the lowly ranked documents, improvement comes from overcoming the term mismatch problem. Overall, our system outperforms the state-of-the-art performance. In next step, we will explore ideas on how to filter out top non-relevant documents.

References

1. Roberts, Kirk, et al. Overview of the TREC 2016 Clinical Decision Support Track. TREC. (2016).
2. Limsopatham, Nut, et al. Modelling the usefulness of document collections for query expansion in patient search. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, (2015).
3. Zhang, Danchen, and Daqing He. Enhancing Clinical Decision Support Systems with Pub-lic Knowledge Bases. Data and Information Management (2017).
4. Zhang, Danchen, et al. Wikipedia-Based Automatic Diagnosis Prediction in Clinical Deci-sion Support Systems. iConference 2017 Proceedings (2017).
5. Zhou, Guangyou, et al. Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering. ACL (1). (2015).
6. Ganguly, Debasis, et al. Word embedding based generalized language model for information retrieval. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, (2015).
7. Carpineto, Claudio, and Giovanni Romano. A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR) 44.1 (2012).
8. Choi, Sungbin, and Jinwook Choi. SNUMedinfo at TREC CDS track 2014: Medical case-based retrieval task. SEOUL NATIONAL UNIV (REPUBLIC OF KOREA), (2014).
9. Balaneshin-kordan, Saeid, Alexander Kotov, and Railan Xisto. WSU-IR at TREC 2015 Clinical Decision Support Track: Joint Weighting of Explicit and Latent Medical Query Concepts from Diverse Sources. Proceedings of the 2015 Text Retrieval Conference. (2015).
10. Gurulingappa, Harsha, et al. Semi-Supervised Information Retrieval System for Clinical Decision Support. TREC. (2016).

11.Mitra, Bhaskar, et al. A dual embedding space model for document ranking. arXiv preprint arXiv:1602.01137 (2016).
12.Lv, Yuanhua, and ChengXiang Zhai. A comparative study of methods for estimating query language models with pseudo feedback. Proceedings of the 18th ACM conference on In-formation and knowledge management. ACM, (2009).
13.Moen, S. P. F. G. H., and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing. (2013).
14.Balaneshin-Kordan, Saeid, Alexander Kotov, and Railan Xisto. WSU-IR at TREC 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. Wayne State University Detroit United States, (2015).