WHUIRGroup at TREC 2016 Clinical Decision Support Task

Ruixue Wang[1,2], Heng Ding[1,2], Wei Lu[1,2]

1. Institute for Information Retrieval and Knowledge Mining,
2. School of Information Management, Wuhan University,
No. 299 Bayi Road, 430072 Wuhan, Hubei, China
{ruixue_wang , hengding , weilu }@whu.edu.cn

Abstract. The goal of Clinical Decision Support(CDS) is to help physicians find academic papers in PubMed, and link medical records to information relevant for patient care. Participants attending the track have access to a data collection which is a snapshot of 1.25 million articles from PubMed Central（PMC）and a set of 30 topics which are EHR admission notes curated by physicians in the real condition from the MIMIC-III data. Our CDS systems are based on Indri toolkit, using Continuous Space Word Vectors expanding the queries. The 2015 topics and results are used to train the re-rank model, using the LambdaMART rank algorithms. The evaluation of run submissions are used partially marked results which is a promising methodologies.

Keywords: learning-to-rank, language model, query expand

1 Introduction

According to the various requirements of end users, medical information retrieval can be categorized into three groups depending on the end users: searching from physicians, searching from members of the general public (patients and their relatives) and searching from radiologists (a subset of physicians for which search in images) [1]. CLEF organizes patient-centered information retrieval, which is concentrating more on general public's medical information seeking needs when searching content on the web [2], while TREC organizes Clinical Decision Support paying more attention on physician's information needs [3].

Clinical Decision Support(CDS) is run for the third time in 2016. It aims to link medical records to information relevant for patient care and help physicians finding necessary information in medical research literature. The topics are admission notes extracted from MIMIC-III, which is the real data generated by clinicians and different from the synthetic cases used in previous years, which is more representative of the physician's search in real situation.

The rest of this paper is organized as follows. Section 2 presents the data collection used in this track. Section 3 presents the method used to solve the task in CDS system. The experimental results are described in Section 4. Conclusions are summarized and discussed in Section 5.

2 Data

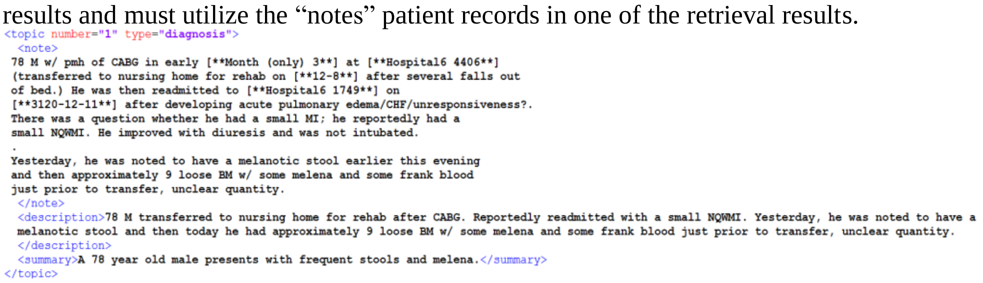
2.1 Document collection

Document collection for Clinical Decision Support Task is the Open Access Subset of PubMed Central (PMC1), which is an online digital database of freely available full-text biomedical literature. Unlike last years' collection, this collection is a new snapshot of the PMC which contains 1.25 million articles. Each article is represented as an NXML file and is identified by a unique number.

[1] http://www.ncbi.nlm.nih.gov/pmc/

2.2 Topics

The topics for the track are EHR admission notes curated by physicians from the MIMIC-III data. Each topic belongs to a generic clinical question types (diagnosis, test, treatment) and includes three versions of the patient records (notes, description, summary). There are to 30 topics in total, and an example is shown in Figure 1. Participants could submit five retrieval results and must utilize the “notes” patient records in one of the retrieval results.



```
<topic number="1" type="diagnosis">
  <note>
 78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital6 4406**]
 (transferred to nursing home for rehab on [**12-8**] after several falls out
 of bed.) He was then readmitted to [**Hospital6 1749**] on
 [**3120-12-11**] after developing acute pulmonary edema/CHF/unresponsiveness?.
 There was a question whether he had a small MI; he reportedly had a
 small NQWMI. He improved with diuresis and was not intubated.
 .
 Yesterday, he was noted to have a melanotic stool earlier this evening
 and then approximately 9 loose BM w/ some melena and some frank blood
 just prior to transfer, unclear quantity.
  </note>
  <description>78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI. Yesterday, he was noted to have a
  melanotic stool and then today he had approximately 9 loose BM w/ some melena and some frank blood just prior to transfer, unclear quantity.
  </description>
  <summary>A 78 year old male presents with frequent stools and melena.</summary>
</topic>
```

Figure 1 Example of topic

3 Method

3.1 Framework of our system

Figure 2 illustrates the framework of the CDS retrieval system. Because the verboseness of the topics, first step was to simplify the topics and remove the useless words. Then Continuous Space Word Vectors [4] method were applied to expand queries.

Indri is used to index our documents with root form achieved by Porter stemming and stop-words removed from documents in the process. Two indices are generated, named as index_nxml and index_txt. For index_nxml, only "abstract”, “title" and "body" field of PubMed documents are used. For index_txt, the nxml2txt[2] tool, which transfers nxml format into .txt, pre-process the PubMed documents before indexing.

The topics process with the same stemmer and stop-words. Our system use the Indri Language Model, TFIDF and BM25 to get base results. For run4 and run3, LambdaMART [5] are used to re-rank the results. For run2 and run5, after simplifying the “notes” or “summaries” to get the initial queries, we use the Continuous Space Word Vectors[3] expand the queries. Continuous Space Word Vectors is obtained by applying Word2Vec to a corpus of 10,876,004 English abstracts of biomedical articles from PubMed. The resulting vectors has 1,701,632 distinct words(types). Finally, using the expanded queries search the documents to get the retrieval results.

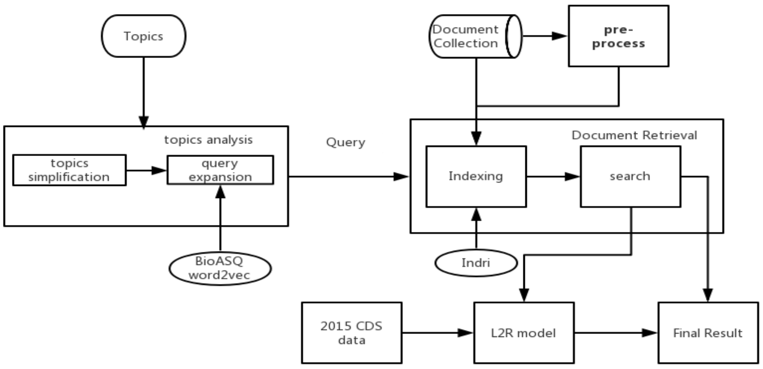


Fig. 2. Framework of our retrieval system

[2] https://github.com/spyysalo/nxml2txt

[3] http://bioasq.org/news/bioasq-releases-continuous-space-word-vectors-obtained-applying-word2vec-pubmed-abstracts

3.2 Learning-to-rank Algorithm

In CDS 2014 &2015, there are some teams using learning-to-rank algorithm to re-rank the results. The Michigan team [6] used Random Forest to re-rank their results, GRIUM [7] re-ranked their result based on SVM algorithm and HIT-WI [8] re-ranked the result based on cooccurrence network. In our CDS system, LambdaMART rank algorithms is applied to re-rank the results.

Feature Extraction. For each document-query pair, the rank and weighting score are extracted as features. These two features are got from BM25, TFIDF and LM model separately in both two indices, index_txt and index_nxml. So, 12 features are obtained in our learning-to-rank model. In this way, we can utilize the advantage of different retrieval models.

Model Training. The training data generated from the 2015 topics and run results. We extracted the features from the results to construct the training data collection.

Re-rank results. Firstly, all 12 features are obtained from retrieval results using BM25, TFIDF and LM in both two indices. Secondly, we construct the predicting data and utilize the model to grade the input results. Finally, rank the results by their new scores to get final run submission.

3.3 Summary of runs

As table 1 shows, we summarize the differences and similarities among our five runs to compare different methods to find out the best way for CDS.

Table 1 summary of 5 runs' method

No.	Field	Pre-process	Index	Model	Re-rank	Query Expand
WHUIR-Group1	notes	nxml2txt tools process the PubMed documents	index_txt	LM	×	×
WHUIR-Group4	notes	Same as 1	index _txt&index_nxml	LM BM25 TFIDF	q-d‘s score and rank position	×
WHUIR-Group5	notes	×	index_txt	LM	×	word2vec
WHUIR-Group2	summaries	×	index_nxml	LM	×	word2vec
WHUIR-Group3	summaries	Same as 1	index _txt&index_nxml	Same as 4	Same as 4	×

4 Experiments and results

In CDS 2016, participants may submit a maximum of five run submissions. At least, one of the five run submissions must utilize the "notes" information which is new in this year topic. Each run consists of a ranked list of up to 1000 documents' PMCID, using the standard trec_eval format.

The evaluation metrics used for the CDS are infAP, infNDCG, R-prec and P@10. The judgment sets were created from the highest ranked articles for each topic submitted by the participants and judged by medical librarians and physicians trained in medical informatics.

All the run submissions reported in this work were obtained using Indri [9] toolkit. Table 2 and Table 3 separately present our official result for using notes and summaries.

According to Table 2 and Table 3, the best and median results using "summaries" are better than "notes". "Summaries" consists just 1-2 sentence summary of the description. It is different from the "note" which has much verbose information. This may lead "summaries" results better than "notes". It is also affect our run submissions.

Comparing between WHUIRGroup2 and WHUIRGroup3 results, the learning-to-rank method have an increase when using "summaries". But the learning-to-rank method have a decrease when using "notes"

Table 2 WHUIRGroup-result using notes

No.	infAP	infNDCG	R-prec	P10
Best	0.0599	0.3302	0.1993	0.51
Median	0.0098	0.1227	0.0791	0.1833
WHUIRGroup1	***0.0104***	***0.119***	***0.0869***	***0.17***
WHUIRGroup4	0.0052	0.0778	0.0435	0.1633
WHUIRGroup5	0.005	0.0748	0.0309	0.14

Table3 WHUIRGroup-result using summaries

No.	infAP	infNDCG	R-prec	P10
Best	0.08685	0.43767	0.25535	0.63
Median	0.01959	0.18589	0.12197	0.2633
WHUIRGroup2	0.0068	0.0981	0.0489	0.1467
WHUIRGroup3	***0.0179***	***0.1654***	***0.0989***	***0.2233***

In table 4, we show the different types of topics using notes' best performance. From the table, we find that different types perform best in different evaluation metrics. UWM-UO also discussed the different among the types [10].

Table 4 Different Types of topics using notes' best performance

No.	infAP	infNDCG	R-prec	P10
all-best	0.0599	0.3302	0.1993	0.51
diagnosis-best	0.05372	0.28011	0.19569	0.47
test-best	0.03435	0.31848	0.1782	***0.54***
treatment-best	***0.05893***	***0.3283***	***0.19857***	0.5096

For the 30 topics, ***WHUIRGroup1*** generated results with the highest scores in evaluation metrics of infAP, infNDCG, R-prec and P@10 among using notes runs. ***WHUIRGroup3*** generated results with the highest scores among using summaries runs.

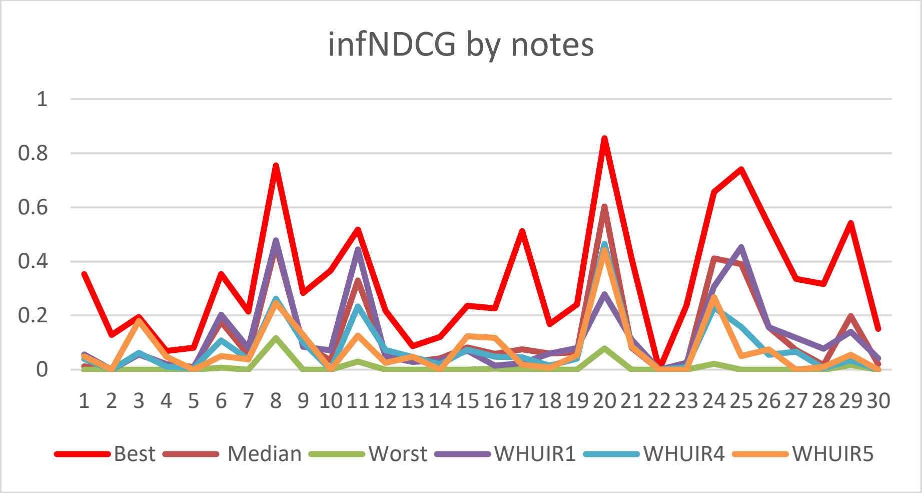


Figure 2 infNDCG scores by using notes as queries

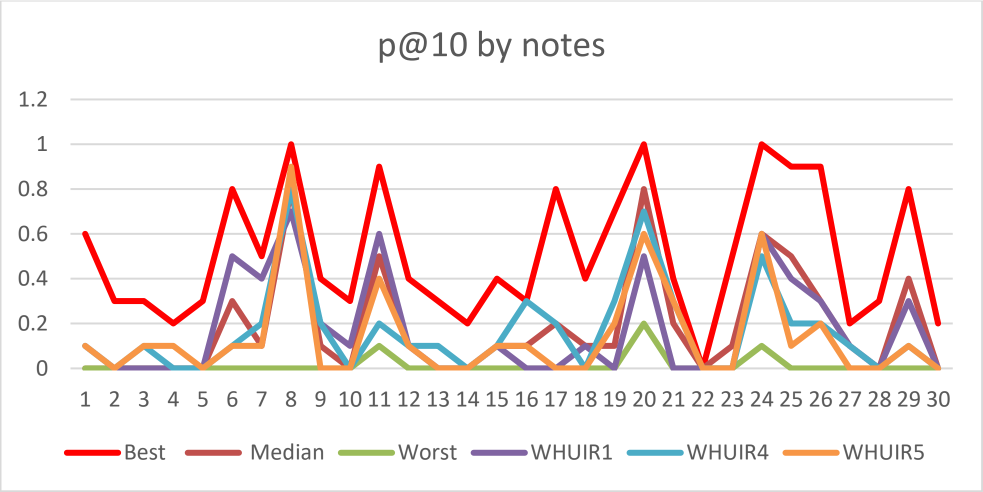


Figure 3 p@10 scores by using notes as queries

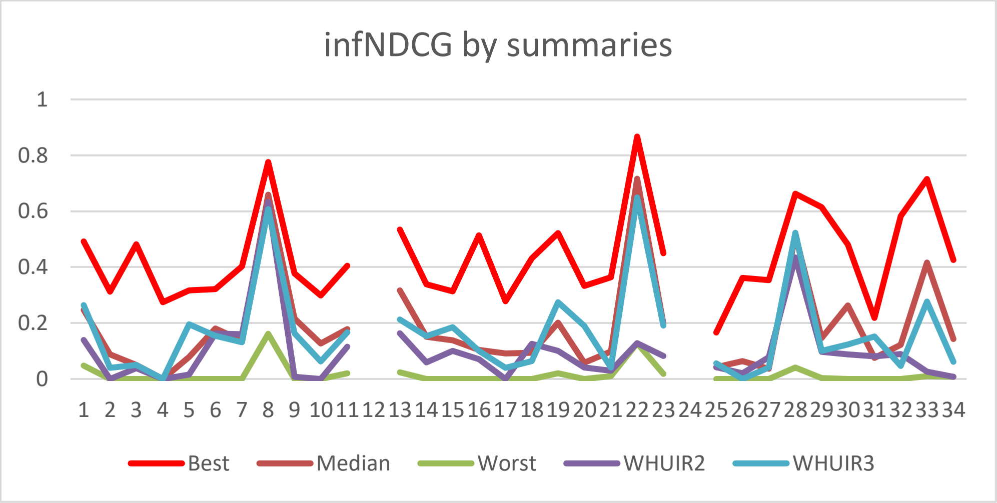


Figure 4 infNDCG scores by using summaries as queries

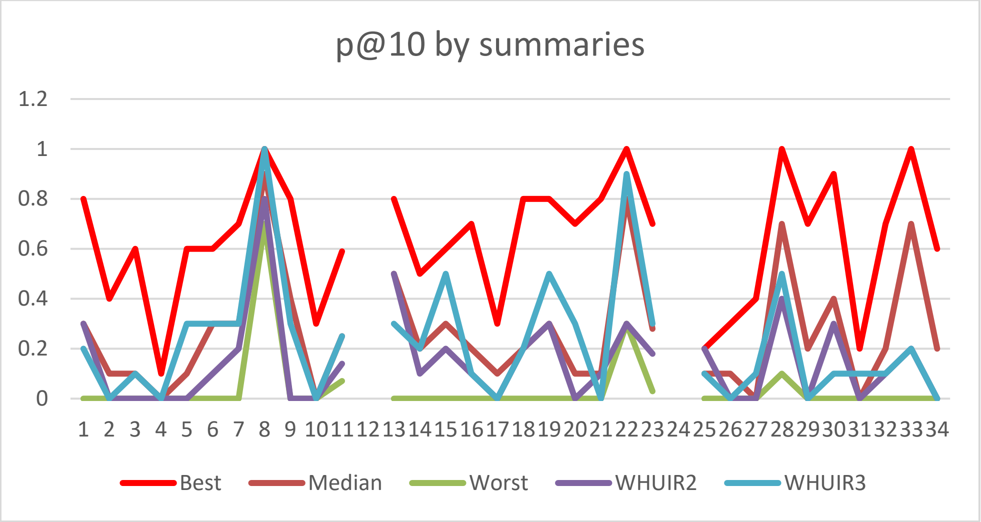


Figure 5 p@10 scores by using summaries as queries

5 Conclusions

This paper describes the participation of WHUIRGroup in the TREC 2016 Clinical Decision Support track. Our best runs outperformed the median 43% among the total 30 topics. Comparing the retrieval results between using "summaries" and "notes" versions of the patient records as queries, the "summaries" performs better. In the future, we will consider the different versions of patient records. Additionally, various types of topics have different performance, we should apply different retrieval system according the type of the topics. This information could be utilized to help to improve health information retrieval .

6 ACKNOWLEDGMENTS

This work is supported by the National Social Science Fund of China (Grant No. 12&ZD1221) and the National Natural Science Foundation of China (Grant No. 71173164).

References

1. Hanbury A. Medical information retrieval: an instance of domain-specific search. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012: 1191-1192.
2. Kelly L, Goeuriot L, Suominen H, et al. Overview of the clef ehealth evaluation lab 2016. International Conference of the Cross-Language Evaluation Forum for European Languages. Springer International Publishing, 2016: 255-266.
3. Roberts K, Simpson M S, Voorhees E, et al. Overview of the TREC 2015 Clinical Decision Support Track.
4. Tsatsaronis G, Balikas G, Malakasiotis P, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics, 2015, 16(1): 1.
5. Q. Wu, C.J.C. Burges, K. Svore and J. Gao. Adapting Boosting for Information Retrieval Measures. Journal of Information Retrieval, 2007.
6. Hu F, Wu D T Y, Mei Q, et al. Learning from Medical Summaries: The University of Michigan at TREC 2015 Clinical Decision Support Track.
7. Liu X J, Nie J Y. Investigation of Concept-based Proximity Matching-GRIUM@ Clinical Decision Support Track 2015 Task 1a.
8. Jiang J, Guan Y, Su J, et al. HIT-WI at TREC 2015 Clinical Decision Support Track.
9. T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language modelbased search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis, volume 2, pages 2–6. Citeseer, 2005.
10. Mu X, You S. TREC 2015 paper submission UWM-UO@ 2015 Clinical Decision Support Track: QE by Weighted Keywords using PRF. TREC. 2015.