ECNU at 2017 eHealth Task 2: Technologically Assisted Reviews in Empirical Medicine

Jiayi Chen[1], Su Chen[1], Yang Song[1], Hongyu Liu[1], Yueyao Wang[1], Qinmin Hu[1], Liang He[1], and Yan Yang[1,2]

Department of Computer Science & Technology, East China Normal University, Shanghai, 200062, China
Shanghai Engineering Research Center of Intelligent Service Robot,Shanghai,China
{jychen,schen,ysong,liuhy,yywang}@ica.stc.sh.cn,
{qmhu,lhe,yyang}@cs.ecnu.edu.cn

Abstract. The 2017 CLEF eHeath Task2 requires to rank the retrieval results given by medical database. The purpose is to reduce efforts that experts devote to finding indeed relevant documents. We utilize a customized Learning-to-Rank model to re-rank the retrieval result. Additionally, we adopt word2vec to represent queries and documents and compute the relevant score by cosine distance. We find that the combination of the two methods achieves a better performance.

Keywords: Learning to Rank; Word2vec; Health Information Retrieval

1 Introduction

The East China Normal University, participated in Task 2, Technologically Assisted Reviews in Empirical Medicine [1], of the CLEF 2017 eHealth Evaluation Lab [2]. This task aims at a ranking problem in Systematic Reviews. Systematic Reviews contains three stages: Boolean Search, Title and Abstract Screening, and Document Screening. This task requires us to rank the documents retrieved from the Boolean Search stage.

In the Boolean Search stage, experts build a boolean query including relevant information. Then they submit it to a medical database containing titles and abstracts of medical studies. The database returns a set of potential relevant studies. In the following two stages, experts decide which ones are indeed relevant by screening titles, abstracts and full documents.

There are two goals for this task. One is to rank the documents retrieved in the Boolean Search stage so that the relevant abstracts are retrieved as early as possible. The other one is to provide a subset of studies containing all or as many of relevant abstracts for the least effort[1].

[1] https://sites.google.com/site/clefehealth2017/task-2

2 Methods

In this task, we first customize a Learning-to-Rank (L2R) model[3]. Furthermore, we apply word2vec to represent queries and documents.

2.1 Learning-to-Rank Model

The Learning-to-Rank model has shown good performance [3, 5]. The architecture of L2R model is shown in Fig.1:

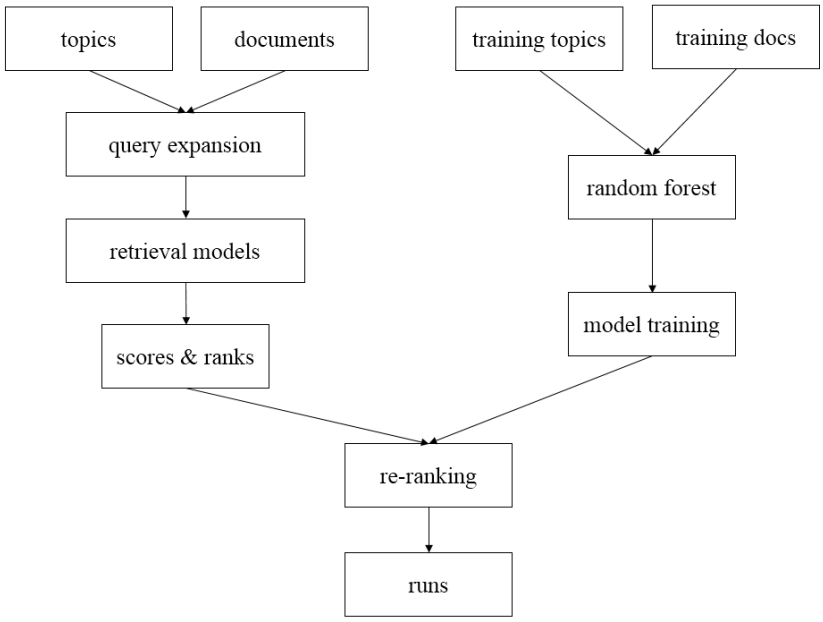

Fig. 1. Architecture of L2R model

There are three stages in the L2R model: Query Expansion, Feature Extraction, Model Training. In the L2R model, we combine each document and each query into a query-document pair. The L2R model gives a relevance score for each query-document pair.

Query Expansion: In the query expansion stage, we intend to improve retrieval precision by expanding queries. We apply the similar model proposed in the 2014 TREC Microblog track[4], 2015 TREC Clinical Decision Support track[5], and 2015 CLEF eHealth Task 2[3].

– Query is submitted to Google and the top-10 concurrent web titles and snippets(if exist) is crawled.
– The MeSH database is applied to extract medical terms from titles and snippets.

Feature Extraction: In this stage,we need to extract features of each query-document pair. When a document is retrieved under a query, it is attached with a weighting score and a rank. So we use the weighting score and the rank from the first retrieval round as features. To take advantages of different retrieval models, we adopt BM25[6], PL2[7] and BB2[8] models to obtain the scores and ranks of the query-document pair. Hence the dimension of the feature vector is six.

Model Training: The L2R model judges the relevance of a query-document pair by using the random forest classifier. We choose the topics and documents of the 2013 and 2014 tasks as the training data. The aforementioned feature vectors are applied to represent query-document pairs in this stage.

2.2 Word2vec Model

Assuming a document of n words is $D = \{d_1, d_2, ..., d_n\}$, we can represent each word d_i in D as a vector $\boldsymbol{d_i}$. Hence the vector of the whole document vector $\boldsymbol{D}$ can be calculated by the average of vectors $\boldsymbol{d_i}$:

$$\boldsymbol{D} = \frac{1}{n} \sum_{1 \leq i \leq n} \boldsymbol{d_i}. \tag{1}$$

Similarly, a query q could also be represented as a vector $\boldsymbol{q}$. We can compute the similarity between query q and document D. In this task, we use the cosine distance to compute the similarity between document D and query q:

$$sim(D, q) = cos(\boldsymbol{D}, \boldsymbol{q}) = \frac{\boldsymbol{D} \cdot \boldsymbol{q}}{\|\boldsymbol{D}\| \cdot \|\boldsymbol{q}\|}. \tag{2}$$

After similarities between the query and documents listed are computed, we can rank these documents in a descend order.

2.3 Combination

We use $S_L(D, q)$ to denote the score of document D with query q from L2R model, and $S_W(D, q)$ to that from Word Vector model. α is the weight of $S_L(D, q)$ and β is the weight of $S_W(D, q)$. The final score is computed as below:

$$S(D, q) = \alpha S_L(D, q) + \beta S_W(D, q), \tag{3}$$

$$\alpha + \beta = 1. \tag{4}$$

3 Experiments

3.1 Dataset

We are provided with development set and test set. In development set there are twenty topics while in test set there are thirty topics. Each topic file contains four parts:

- Topic-id
- The title of review written by experts
- The boolean query manually constructed by experts
- The set of PubMED Document Identifiers (PID's) returned by MEDLINE.

Since the query of a topic is a boolean query, we remove three words near the negation word "not" to avoid misleading the intension of the query .

3.2 Runs

We submit three runs whose descriptions are followed below:

run-1: result of the Word Vector model. We use the pre-trained word vectors from Stanford University trained by GloVe model[9]. The size of vocabulary is 2.2M and the dimension of each vector is 300. The vector of the word that does not occur in the pre-trained word vectors is **0**.

run-2: result of L2R model.We adopt terrier-4.0.0 to run BM25, BB2 and PL2 model. We select top-1000 PIDs for each topic.

run-3: result of the combination of L2R model and Word Vector model. The parameters are tuned on training set. Finally we choose $\alpha = 0.8$ and $\beta = 0.2$ in equation (4). However, a PID of a topic may not occur in the result of L2R model. In this case, $\alpha = 0$ and $\beta = 1$. Similar to run-2, we choose top-1000 PIDs for each topic.

The evaluation results of three runs are shown in Table.1. These results are provided by the organizer.

4 Conclusions and Future Work

In the 2017 CLEF eHealth Task 2, we ECNU_ICA team take advantages of the Learning-to-Rank model. We also adopt word2vec to represent queries and documents and compute their similarities by cosine distance. Although the combination of two methods performs well, the performance of our word2vec model needs to be improved. In the future, we will apply better methods which can avoid losing semantic information.

Table 1. Evaluations of 3 runs

Runs	ap	$lastrel$	wss_{100}	wss_{95}	$norm\_area$	$recall$
run-1	0.091	3633.167	0.099	0.121	0.627	1.0
run-2	0.16	699.167	0.067	0.159	0.644	0.708
run-3	0.166	725.1	0.077	0.175	0.651	0.716
Runs	$cost_{total}$	$cost_{uniform}$	$cost_{weighted}$	$loss_e$	$loss_r$	$loss_{er}$
run-1	3918.7	3918.7	3918.7	0.544	0.0	0.544
run-2	1000	6641.843	4003.28	0.292	0.153	0.444
run-3	1000	4016.12	6717.827	0.292	0.152	0.443

Runs	NCG@10	NCG@20	NCG@30	NCG@40	NCG@50
run-1	0.202	0.376	0.504	0.587	0.679
run-2	0.337	0.422	0.458	0.47	0.482
run-3	0.339	0.425	0.458	0.471	0.481
Runs	NCG@60	NCG@70	NCG@80	NCG@90	NCG@100
run-1	0.771	0.842	0.906	0.952	0.998
run-2	0.492	0.496	0.498	0.499	0.501
run-3	0.494	0.497	0.501	0.505	0.507

5 Acknowledgement

This research is funded by the National Nature Science Foundation of China (No. 61602179) and the Science and Technology Commission of Shanghai Municipality (No.15PJ1401700).
This work was supported by Xiaoi Research, by Shanghai Municipal Commission of Economy and Information Under Grant Project No.201602024.

References

1. E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. Overview of the CLEF technologically assisted reviews in empirical medicine. In Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017., CEUR Workshop Proceedings. CEUR-WS.org, 2017.
2. H. Suominen, L. Kelly, L. Goeuriot, E. Kanoulas, A. Neveol, G. Zuccon, and J. R. M. Palotti. Overview of the CLEF ehealth evaluation lab 2017. In Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11- 14, 2017, Proceedings, Lecture Notes in Computer Science. Springer, 2017.
3. Yang, S., He, Y., Hu, Q.M., He, L., Haacke, E.M.: ECNU at 2015 eHealth Task 2: User-centred Health Information Retrieval. Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2015).
4. Chen, Q., Hu, Q.M., Pei, Y.J., Yang, Y., He, L.:ECNU at TREC 2014: Microblog Track. (2014)
5. Yang, S., He, Y., Hu, Q.M., He, L.: ECNU at 2015 CDS Track: Two Re-ranking Methods in Medical Information Retrieval. Proceedings of the 2015 Text Retrieval Conference (2015).

6. Stephen E., Robertson, S.W., Susan J., Micheline H.B., Mike G.: Okapi at TREC-3. Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg, USA
7. Amati, Gianni, Cornelis Joost, and Van Rijsbergen.: Probabilistic models for information retrieval based on divergence from randomness. (2003).
8. Amati, G., Cornelis J., Van R.: Probabilistic models for information retrieval based on divergence from randomness. (2003).
9. Jeffrey P.,Richard S.,Christopher D.:2014. GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP) (2014).