UCAS at TREC-2017 Precision Medicine Track

Canjia Li, Ben He, Yingfei Sun, and Jungang Xu

School of Computer and Control Engineering
University of Chinese Academy of Sciences
{licanjia17}@mails.ucas.ac.cn, {benhe, yfsun, xujg}@ucas.ac.cn

Abstract. The participation of UCAS at TREC Precision Medicine 2017 aims to evaluate the effectiveness of integrating semantic evidence to enhance medical information retrieval system. Benefited from the success of distributed semantic representation of words and documents in the natural language process (NLP) domain, two methods on generating document vectors are proposed. Based on the hypothesis that pseudo relevant feedback for a given query would be a better representation of the query in the semantic vector space, we propose a framework that integrates the semantic features to the final ranking process. In addition, query expansion using Medical Subject Headings (MeSH) and pseudo relevance feedback (PRF) are used. Experimental results show that our method achieves significant improvement over the PRF baseline for clinical trials, while full text articles might be required for learning local document embeddings that are effective for retrieval from abstracts.

1 Introduction

TREC Precision Medical track 2017 (PM2017) focuses on connecting patients with existing articles from PubMed Central (PMC) and experimental treatments in clinical trials from ClinicalTrials.gov website. There are 30 topics concerning patients' condition: disease, genetic variants, demographic and potential other information. For each collection, participants are allowed to submit a maximum of five runs.

Many existing ranking methods in information retrieval (IR) are based on frequency-based statical models. Benefiting from advances in natural language processing (NLP) tasks, words and documents can be represented in high dimension vector space, i.e. embeddings [1]. By taking word context into account, word embeddings capture not only structural relationships but also meaningful semantic relationships between words [2]. Positive results have been seen in researches which emphasize on the importance of semantics in IR tasks [3–5].

In our experiments, we adopt two methods of generating semantic document embeddings. Based on the hypothesis that pseudo relevant feedback for a given query would be a better representation of the query in the semantic vector space, semantic score for a document is obtained by calculating its semantic relevance with pseudo relevant feedback set. These features are then integrated with the baseline model in the final ranking process. In addition, query expansion using Medical Subject Headings (MeSH) and pseudo relevance feedback is performed.

The rest of the paper is organized as follows. Section 2 gives a detailed introduction to the techniques used in our experiments. Section3 presents the experimental settings, results and analysis. Finally, Section 4 concludes our experiments.

2 Method

In this section, we summarize the key techniques used in our medical retrieval system.

2.1 Parameter-free Model: DPH

Derived from the DFR framework, DPH is a parameter-free retrieval model. DPH can obtain comparable results with language models (LM) while results from LM are obtained under optimal parameters [6]. Using DPH, the relevance score of a document d for a query Q is given by [7]:

$$score(d,Q) = \sum_{t \in Q} \frac{qtw(1-F)^2}{tf+1} \cdot (tf \cdot \log_2(tf \cdot \frac{avg_l}{l} \frac{N}{TF})) + 0.5 \cdot \log_2(2\pi \cdot tf \cdot (1-F)) \tag{1}$$

where F is given by tf/l, tf is term frequency within the document. l is the document length and avg_l is the average document length in the collection. N is the number of documents while TF is term frequency in the collection. All variables in the formula can be directly obtained from the collection statistics, thus no parameter tuning is required.

2.2 Probability Retrieval Model: BM25

The probability retrieval model BM25 has been widely used in the information retrieval domain due to its effectiveness [8]. Given a document d and a query Q, the ranking function is

$$score(d,Q) = \sum_{t \in Q} w_t \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \tag{2}$$

where t denotes one term in the query, and qtf is the term 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 . w_t is the weight of query term t, which is given by:

$$w_t = \log_2 \frac{N - df_t + 0.5}{df_t + 0.5} \tag{3}$$

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 contains t.

2.3 Pseudo Relevance Feedback

Under the assumption that the top K documents in the initial retrieval result are a good feedback of user's information need, the pseudo relevance feedback technique automatically extracts important terms in these documents based on various methods such as KL and Bo1, etc [9].

2.4 MeSH Terms Query Expansion

MeSH (Medical Subject Headings) has been widely used in the previous CDS track to improve medical information retrieval [10, 11]. Given a query, We adopt the MetaMap toolkit[1] from UMLS program [12] to extract relevant medical concepts . Since UMLS has more than 100 semantic types, only the following semantic types are considered to be relevant and added to the given query: disease or syndrome, sign or symptom, pathologic function, diagnostic procedure, anatomical abnormality, laboratory procedure, pharmacologic substance, neoplastic process, therapeutic or preventive procedure.

2.5 Document Embedding

Word2Vec is a state-of-the-art neural embedding framework, which aims to generate high quality word vectors in high-dimension space [1]. By calculating the cosine distance between two words, their similarity can be captured. Moreover, it has been stated that simple algebraic operations can also be performed on word vectors [2]. For example, vector(King) - vector(Man) + vector(Woman) results in a vector that is closest to the vector representation of the word Queen. As the semantic relationships are preserved in the embedding operations, one way of generating document embedding is to add up the most informative word embeddings, say top k, within the document, which is given by:

$$\overrightarrow{d} = \sum_{w \in W_k^d} tf\text{-}idf(w) \cdot \overrightarrow{w} \tag{4}$$

where $\overrightarrow{w}$ and $\overrightarrow{d}$ are the embeddings of word w and the corresponding document d, respectively. W_k^d is the set of the top-k terms with the highest *tf-idf* weights in d. We denote the above way of generating word embeddings as *Term Addition*.

We also adopt the unsupervised neural network model, modified from the Word2Vec framework, to learn distributed vector representations for documents [13]. The learned vectors are denoted as *Paragraph Vector*.

[1] https://metamap.nlm.nih.gov/

2.6 Feedback-based Ranking Approach

Both DPH and BM25 are frequency-based statistical models, in which semantic evidence of relevance is not taken into account. Intuitively, We can enhance retrieval system performance by integrating semantic evidence. In this work, we utilize the feedback-based approach in [14] to rerank the results given by the above classical retrieval models. Given a query, similar to PRF, we assume that the top k documents in the initial querying result provide an abundant semantic evidence about the query. Therefore, we estimate the semantic relevance of a document by measuring its semantic relevance with the top k documents. The enhanced ranking score is given as follows:

$$score(d, Q) = \lambda \cdot BM(d, Q) + (1 - \lambda) \cdot SEM(d, D^k_{PRF}(Q)) \tag{5}$$

where $BM(d, Q)$ is the ranking score of document d given by a baseline retrieval model, e.g. the classical BM25 or DPH ranking model with PRF. $D^k_{PRF}(Q)$ is the pseudo relevance feedback set of documents, which is composed of the top ranked k articles returned by the baseline model. $SEM(d, D^k_{PRF}(Q))$ measures the semantic similarity between document d and the pseudo relevance feedback set $D^k_{PRF}(Q)$, which is given as follows:

$$SEM(d, D^k_{PRF}(Q)) = \sum_{d' \in D^k_{PRF}(Q)} w_{d'} \cdot Sim(d', d) \tag{6}$$

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

$$w_{d'} = BM(d', Q) + \max_{d'' \in D^k_{PRF}(Q)} BM(d'', Q) \tag{7}$$

$Sim(d', d)$ denotes the semantic similarity between d' and d, which is measured by the cosine distance. In Equation (7), the maximum relevance score is added to normalize the gap between the relevance scores of different articles. Note that both $BM(d, Q)$ and $SEM(d, D^k_{PRF}(Q))$ in Equation (5) are normalized by Min-Max normalization, so that the two scoring features are on the same scale.

3 Experimental Setting and Results

3.1 Query Preprocessing and Index

The topics consist of the disease, genetic variants, demographic, and the potentially other information about patients. We assume that disease and genetic variants are better indicator of patients' conditions, especially their diseases. Therefore, the components of a topic are assigned with different weights to reflect their relevance to a patient's information need. The weights for disease, genetic variants, demographic and other information are $3.0, 2.0, 1.0, 1.0$, respectively. For example, topic 1 is reformed as the following:

$$\boxed{liposarcoma^{3.0}\, cdk4^{2.0}\, amplification^{2.0}\, male^{1.0}\, gerd^{1.0}}$$

There are two target collections for the Precision Medicine track: scientific abstracts and clinical trials. The scientific abstracts consist of a January 2017 snapshot of PubMed abstracts and abstracts from AACR and AASCO proceedings. For abstracts from PubMed, we also extract their *title* fields. The clinical trials are an April 2017 snapshot of ClinicalTrials.gov website. The *title, summary, detail, criteria* fields are extracted for indexing after stemming with Porters stemmer and the removal of stopword. All experiments are conducted with Terrier [15].

3.2 Results

We submitted five official runs for each document collection, which is summarized in Table 1 (SA denotes scientific abstracts, CT denotes clinical trials):

Table 1. Run submission summary

RunID	Collection	Baseline Model	Query Expansion	Document Vector
UCASBASE	SA	DPH	Bo1	None
	CT	DPH	Bo1	None
UCASSEM1	SA	DPH	Bo1	Term Addition(ALL)
	CT	DPH	Bo1	Term Addition(ALL)
UCASSEM2	SA	DPH	Bo1	Paragraph Vector
	CT	DPH	Bo1	Paragraph Vector
UCASSEM3	SA	BM25	KL	Term Addition(TOP-5)
	CT	BM25	KL	Term Addition(TOP-5)
UCASSEMUMLS	SA	BM25	KL	Paragraph Vector
	CT	BM25	KL	Term Addition(TOP-5)

Table 2. Evaluation results

runID	CT			SA		
	P5	P10	P15	infNDCG	P10	R-prec
UCASBASE	0.4143	0.3750	0.3429	**0.3271**	**0.4276**	**0.2227**
UCASSEM1	0.4286	0.3607	0.3262	0.3172	0.4000	0.2043
UCASSEM2	**0.4429**	**0.3786**	**0.3548**	0.3101	0.4172	0.2019
UCASSEM3	0.4286	0.3571	0.3286	0.3106	0.4103	0.2057
UCASSEMUMLS	0.3357	0.3000	0.2619	0.2825	0.3690	0.1874

The evaluation results of our runs for clinical trials and scientific abstracts are shown in Table 2. The best values are highlighted in boldface. From the results above, we found that integrating semantic evidence to the ranking process

enhances the system performance for clinical trials according to the $P@5$ metric. In particular, *UCASSEM2* outperforms the baseline method in all metrics with improvements of $6.8\%, 1.0\%, 3.4\%$ respectively, which demonstrates the effective of our approach.

For scientific abstracts, no improvement was seen against the baseline method. As the average document length of scientific abstracts is much shorter than clinical trials, we assume that the proposed embedding-based method may not apply to short documents. In our approach, the top k term embedding with highest tfidf within the document are summed up in the $TermAddition$ method. However, this method may not be able to recognize the most important terms because term frequencies of different terms tend to be similarly low in a short document. Besides, if all terms are summed up, the embedding centroid dose not represent the theme of this document because many meaningless words force the centroid to move randomly in the high-dimension space. In contrast, the embedding centroid of a long document retain its position in the space due to the frequent occurrences of informative words.

Query expansion using the domain knowledge with UMLS was explored in the run $UCASSEMUMLS$. However, system performance dropped rapidly for both collections, which suggests that domain knowledge should be used carefully in the medical IR research task.

In the future, we plan to investigate effective and meaningful embeddings for short documents, i.e., scientific abstracts, to better respond to patients' information needs.

4 Conclusions

In order to exploit semantic features in the medical document ranking system, document vector representation methods derived form Word2Vec technique are proposed. In the ultimate ranking process, semantic relevance features are combined with baseline model to enhance the system performance. Evaluation results show that the proposed method outperforms the traditional retrieval models for long documents, i.e. clinical trials. Still, better representations of short documents are required for effective retrieval from abstracts.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61472391).

References

1. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

2. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
3. Q. Ai, L. Yang, J. Guo, and W. B. Croft, "Improving language estimation with the paragraph vector model for ad-hoc retrieval," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 869–872, ACM, 2016.
4. X. Tu, J. X. Huang, J. Luo, and T. He, "Exploiting semantic coherence features for information retrieval," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 837–840, ACM, 2016.
5. I. Vulić and M.-F. Moens, "Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 363–372, ACM, 2015.
6. Z. Ye, B. He, X. Huang, and H. Lin, "Revisiting rocchios relevance feedback algorithm for probabilistic models," *Information Retrieval Technology*, pp. 151–161, 2010.
7. G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi, "Fub, iasi-cnr and university of tor vergata at trec 2008 blog track," tech. rep., FONDAZIONE UGO BORDONI ROME (ITALY), 2008.
8. S. Robertson, H. Zaragoza, *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
9. G. Amati, *Probability models for information retrieval based on divergence from randomness*. PhD thesis, University of Glasgow, UK, 2003.
10. S. Balaneshin-Kordan, 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," tech. rep., Wayne State University Detroit United States, 2015.
11. H. Gurulingappa, L. Toldo, C. Schepers, A. Bauer, and G. Megaro, "Semi-supervised information retrieval system for clinical decision support.," in *TREC*, 2016.
12. O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.
13. Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196, 2014.
14. C. Yang, B. He, and J. Xu, "Integrating feedback-based semantic evidence to enhance retrieval effectiveness for clinical decision support," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, pp. 153–168, Springer, 2017.
15. C. Macdonald, R. McCreadie, R. L. Santos, and I. Ounis, "From puppy to maturity: Experiences in developing terrier," *Open Source Information Retrieval*, vol. 60, 2012.