A generic retrieval system for biomedical literatures: USTB at BioASQ2015 Question Answering Task

Zhi-Juan Zhang, Tian-Tian Liu, Bo-Wen Zhang*, Yan Li,
Chun-Hua Zhao, Shao-Hui Feng, Xu-Cheng Yin*, and Fang Zhou

Department of Computer Science and Technology,
University of Science and Technology Beijing (USTB), Beijing 100083, China
xuchengyin@ustb.edu.cn
zbw292@126.com

Abstract. In this paper we describe our participation in the 2015 BioASQ challenge task on question answering (Phase A). Participants need to respond to the natural language questions in the format of documents, snippets, concepts and RDF triplets. In document retrieval, we build a generic retrieval model based on the sequential dependence model, Word Embedding and Ranking Model. In addition, from the view of the special significance of titles(Title Significance Validation), we re-rank the *top-K* results by counting the meaningful nouns in the titles. The *top-K* documents are split into sentences and indexed for snippets retrieval. The similar models of document retrieval are applied for this part. To extract the biomedical concepts and corresponding RDF triplets, we use concept recognition tools *MetaMap* and *Banner* [1]. Statistics indicate that our systems outperform other results.

Keywords: generic retrieval, sequential dependence model, Word Embedding, Ranking, *MetaMap*, *Banner*

1 Introduction

The challenge of BioASQ consists of two tasks [1]: a large-scale semantic indexing task (Task 3a) and a question answering task (Task 3b). We only focus on phase A of Task 3b which includes four parts: retrieving the gold relevant articles and the most relevant snippets from the benchmark datasets, retrieving relevant concepts from designated terminologies and ontologies and RDF triples from designated ontologies. For this task, participants are provided with about 100 questions in each batch and required to return at most 10 answers for each part.In all of the following experiments,we utilize the training datasets 3b which includes 810 queries.

[1] http://ikmbio.csie.ncku.edu.tw/GN/

2 Methodology

In our system, we deploy *Galago* [2], an open source search engine developed as an improved JAVA version of Indri, over large clusters for indexing and retrieval. We lease 2015 *MEDLINE/PubMed Journal Citations* from the U.S. National Library of Medicine, composed of about 22 million *MEDLINE citations*.

2.1 Data Pre-Processing

For documents retrieval, the fields of title and abstract are extracted from document resources and indexed with *Galago*. On the basis of experimental results of document retrieval, the *top-K* documents are chosen from the candidates as the source of retrieval for snippets retrieval part. Titles and abstracts of the articles are separated into several sentences according to some specific rules. These sentences make up a pile of new files with the field name Text for indexing. For concepts retrieval part, participants are required to return relevant concepts in five ontologies or terminologies: *MeSH*, *GO*, *SwissProt*, *Jochem* and *DO*. We download all the resources and index the fields of term and ID.

2.2 Query Pre-Processing

Except the experiment of triples retrieval, original queries are processed with the same approaches. The stop words in queries are removed and the queries are case-folded, stemmed and tagged with *Porter Stemmer* and *Part-Of-Speech*. Finally we filter out the special symbols.

MetaMap [2, 3], which is a semantic tool in medical text processing, maps concepts in the *UMLS Metathesaurus*. Biomedical terminologies and ontologies are identified from queries by *MetaMap* and composed new queries to retrieval concepts. Linked life Data is aggregation of more than 25 popular biomedical data sources. Users are able to access 10 billion RDF statements through a single *SPARQL* endpoint.

In the following sections, the procedures of retrieval models sequential dependence model(SDM), Word Embedding($Word2Vec$), Ranking Model(RM) and Title Significance Validation(TSV) are introduced in detail.

2.3 Searching

2.3.1 Sequential Dependence Model

Our baseline of documents retrieval is the unigram language model referred as query likelihood model (QL). In this model, the likelihood of query term q_i occurring assumed is that not affected by the occurrence of any other query terms. But for a natural language query, the terms depend on each other. So our retrieval models should take the sequence of terms into account.

[2] `http://www.galagosearch.org/`

Metzler and *Crofts Markov Random Field* (*MRF*) model [4, 5], also called undirected graphical models, is commonly used in the statistical machine learning domain to succinctly model joint distributions. The sequential dependence model (*SDM*) is a special case of the *MRF*. It assumes the occurrences of adjacent query terms are related.

Three types of features are considered in *SDM*: single term features (standard unigram language model features, f_T), exact ordered phrase features (words appearing in sequence, f_O) and unordered window features (require words to be close together, but not necessarily in an exact sequence order, f_U).

For the query Q after pre-processing, Q=q_1, q_2,...,q_i,.... Document D is ranked according to the following equation (1):

$$\begin{aligned} score_{SDM}(Q,D) = \lambda_T \sum_{q \subset Q} f_T(q,D) \\ +\lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i, q_i+1, D) \\ +\lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i, q_i+1, D) \end{aligned} \tag{1}$$

(λ_T, λ_O, λ_U) are weight parameters for single terms, ordered terms and unordered terms.

2.3.2 Word2Vec

One of the most critical language issues for retrieval performance is the term mismatching problem. The 810 queries of training datasets 3b after pre-processing contains 4609 terms. There are about 5.7 terms on average for each query. The queries are short and the natural language is inherently ambiguous. The queries may not use the same terms as the retrieval sources. Query expansion is usually utilized to select the golden relevant terms to the original queries. However, the main challenge of the query expansion is to find the expansion terms, especially in specific areas, such as biomedicine.

The result vectors of words offered by BioASQ officials can be used to estimate the relatedness of two words. With the similarities of each two words, the query expansion can be easily applied. The resulting vectors of 1,701,632 distinct words (types) is trained by the *Word2Vec* [3] tool which processes a large corpus and maps the words in the corpus to vectors of a continuous space. We use these word vectors based on *SDM*. The feature f_T is replaced with f_W, which represents the expansion terms feature.

For a query Q=q_1, q_2,...,q_i,..., we calculate the distance between the term q_i and all the distinct terms from the dictionary by cosine similarity. Then all the terms are sorted by the distances with q_i. The nearest k terms are

[3] `https://code.google.com/p/word2vec/`

chosen to enrich the original query. The original terms q_i with the additional terms q_{i1},q_{i2},. . . ,q_{ik}, used as expansion terms with corresponding weights w_i(i=1,2,. . .).A new query can be reformulated as Q_{new}=(t_1,t_2,. . . ,t_i,. . .),where $t_i \epsilon$ T_i =q_i,q_{i1},q_{i2},. . . ,q_{ik}.

Documents are ranked by the enriched *SDM* query T according to the following scoring equation (2):

$$\begin{aligned} score_{Word2Vec}(Q, D) = \lambda_W \sum_{t \subset T} f_W(T, D) \\ +\lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i, q_i + 1, D) \\ +\lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i, q_i + 1, D) \end{aligned} \tag{2}$$

(λ_W, λ_O, λ_U) are the weights for expansion terms, ordered phrases and unordered phrases in *SDM*.

2.3.3 Re-ranking of Word2Vec Results

In order to improve the retrieval performance, we propose a Reranking Model (RM)[6, 7] based on the Word2Vec results. For each query, a subset D of results composed by the *top-K* documents is represented by a vector according to *TF-IDF*. Then the similarity of each two documents is calculated by the cosine similarity of corresponding vectors. The similarity of the K documents make up the K*K dimension matrix M.M[i][j] represent the similarity of the D_i and D_j. Via these similarities, we update the score of the documents for each query by Equation (3).$score_i$ is the initial score for the D_i,

The updated score of the document D_i for the query Q is calculated by the following equation:

$$score_{RM}(Q, D_i) = \lambda score_i + \frac{1}{k-1}(1-\lambda) \sum_{j=0, j \neq i}^{k} (score_j * M[i][j]) \tag{3}$$

2.3.4 Title Significance Validation

With a specific request and several relevant literatures, people usually directly judge the titles rather than carefully reading the full text of the abstract. In order to investigate the special significance of titles, we design an interesting experiment to validate it. We pick *top-K* documents retrieved by the *Word2Vec* model and look up the corresponding titles. Then we compare these titles with the processed query. Different from other type of words, nouns are a meaningful linguistic unit and have virtual influence in natural language.

Hence, we filter out all types of words from the queries other than the nouns labeled by the *Stanford-POS tagger* when processing the queries. The frequency with which the nouns occur in the titles are counted as *title-hit*. We combine the *title-hit* and initial score by linear combination. We respectively compare (stemmed query, stemmed titles) and (non-stemmed query, non-stemmed titles) to see if *title-hit* can influence the performance.

3 Experiments on Generic Retrieval Models

We train and validate our methods on the training datasets 3b which contains 810 queries based on the 22 million *MEDLINE* documents. We utilize *trec_eval* [8] to evaluate the top 100 ranked search lists. Mean average precision(MAP)[9, 10] serves as our evaluation metric.In the previous years,we are required to return at most 100 relevant results.But the participating systems are required to return at most 10 relevant results in 2015.So we select the best parameters through the training datasets 3b,then the parameters are utilized to the testsets 3b.The results with testsets 3b are offered by BioASQ officials. The scalar μ which is a hyper-parameter controls the amount of collection smoothing applied. We set the value in the range between 500 and 5000.The following tables are only parts of our documents experiments for setting up the parameters on training datasets 3b.

In *SDM*, there are three weighting parameters (λ_T, λ_O, λ_U) to be trained. We set each of the parameters values from 0.00 to 1.00 in steps of 0.01. Based on this, w_i in the *Word2Vec* model, which is the weights for expansion terms, needs to be set. In addition, another issue for query expansion is to confirm how many expansion terms are suitable for retrieval. As a comparison, on all training data, the performance with the expansion terms from 1 to 10 are measured for an optimum parameter. The results are shown in Table 1.

Table 1. Comparison of *QL*, *SDM*, *Word2Vec* measured with *MAP@100*

Method	QL	SDM	Word2Vec
Training Set2b	0.2235	0.2381	**0.2438**
Batch1	0.2726	0.2947	**0.3014**
Batch2	0.2608	0.2771	**0.294**
Batch3	0.2588	0.2723	**0.2767**
Batch4	0.2476	0.2606	**0.2739**
Batch5	0.2389	0.2681	**0.2781**

Overall, it works well when those parameters are optimized in the datasets 3b.Obviously, *Word2Vec* shows greater performance compared with *QL* and *SDM*. Especially,the average result with *Word2Vec* is higher than the other two.So Reranking Model and Title Significance Validation are evaluated based on this model.

Afterwards, the *top-K* documents determined by the initial ranking are re-ranked by *RM*. The value of K is trained by groups of experiments.The initial scores and similarities are also taken into account. The value λ is changed from 0.000 to 1.000 in steps of 0.001.After many experiments, we get the stable parameter values.The parts of comparison results are shown in Table 2.

Table 2. Results of 810 queries for Reranking Model with MAP@100

λ	0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00
RM	0.2879	0.2884	0.2890	0.2894	0.2897	0.2899	0.2899	**0.2901**	0.2891	0.2886	0.2878

The *RM* performs well compared with *Word2Vec* which the value of λ is 1.

Results for the *TSV* model which contains non-stemmed queries and stemmed queries are presented in Table 3.

Table 3. MAP@100 results for TSV

Method	Word2Vec	non_Stem	Stem
Training Datasets 3b	0.2878	0.2932	**0.2988**

Experimental results show that the effectiveness is improved when applying title significance validation appropriately.

We choose the parameters in *RM* model with best result on the five Batch respectively and then compare to the official results of top 3 winning participants in BioASQ 2014 [4].Table 4 shows the results of our system and top 3 participants.

Table 4. Comparison between our system and Top 3 participants in BioASQ 2014(Phase A) measured with *MAP@100*

Method	Top 1	Top 2	Top 3	Our system
Batch1	0.2794	0.1108	0.1040	**0.3067**
Batch2	0.3016	0.2508	0.0797	**0.3059**
Batch3	**0.2918**	0.2773	0.1022	0.2793
Batch4	0.2713	0.0898	0.0881	**0.2850**
Batch5	0.2661	0.0889	0.0883	**0.2806**
Mean Value	0.2820	0.1635	0.0925	**0.2915**

[4] `http://participants-area.bioasq.org/oracle/results/taskB/phaseA/`

Form the Table 4,we find our results are better than the Top 1 except the Batch3 because of the random data.So our generic retrieval system is more effective in biomedical retrieval.

4 Conclusion

Due to the limited time, we only participate in the phase A of task 3b. But our approaches performs competitive especially during the documents and snippets retrieval. We adopt various retrieval models and adjust almost all possible parameters to improve the final performance. Although our trained system performs stable on the training set 2015 (810 queries), the *MAP* value on batch 3(testsets 3b) is unusual. Giving a deep analysis of the query set of batch 3, we think the cause may be the count of terms and biomedical nouns in each query.

In the future, we will focus on the strategies of query expansion on biomedical text, probabilities of improving the document retrieval accuracy through the feedback results of snippets retrieval. Besides, our research will add natural language processing (*NLP*) into our system to improve the performance.

References

1. Georgios Balikas, Ioannis Partalas, Axel-Cyrille Ngonga Ngomo, Anastasia Krithara, Georgios Paliouras:Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. CLEF (Working Notes) 2014: 1181-1193
2. Aronson A R. MetaMap: Mapping Text to the UMLS Metathesaurus[J]. Bethesda, 2006.
3. Aronson A R, Lang F M. An overview of MetaMap: historical perspective and recent advances[J]. J Am Med Inform Assoc, 2010, 17(3):: 229C236.
4. Metzler D, Croft W B. A Markov random field model for term dependencies[C]// In Proceedings of SIGIR 2005. 2005:472-479.
5. Sungbin Choi, Jinwook Choi: Classification and Retrieval of Biomedical Literatures: SNUMedinfo at CLEF QA track BioASQ 2014. CLEF (Working Notes) 2014: 1283-1295
6. Bo-Wen Zhang, Xu-Cheng Yin, Xiao-Ping Cui, Bin Geng, Jiao Qu, Fang Zhou, Li Song and Hong-Wei Hao. Social Book Search Reranking with Generalized ContentBased Filtering. Submitted to CIKM14.
7. Bo-Wen Zhang, Xu-Cheng Yin, Xiao-Ping Cui, Jiao Qu, Bin Geng, Fang Zhou, Hong-Wei Hao:USTB at INEX2014: Social Book Search Track. CLEF (Working Notes) 2014: 536-54
8. Buckley C. trec eval IR evaluation package; 1999.
9. Manning CD, Raghavan P, Schutze H. Introduction to information retrieval, vol.1. Cambridge University Press Cambridge; 2008.
10. Buckley C, Voorhees EM. Evaluating evaluation measure stability. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. Athens, Greece: ACM;2000. p. 33C40.