A Multi-strategy Query Processing Approach for Biomedical Question Answering: USTB_PRIR at BioASQ 2017 Task 5B

Zan-Xia Jin, Bo-Wen Zhang*, Fan Fang, Le-Le Zhang and **Xu-Cheng Yin***
Pattern Recognition and Information Retrieval lab (PRIR)
Department of Computer Scicene, University of Science and Technology Beijing
bowenzhang@xs.ustb.edu.cn, xuchengyin@ustb.edu.cn

Abstract

This paper describes the participation of USTB_PRIR team in the 2017 BioASQ 5B on question answering, including document retrieval, snippet retrieval and concept retrieval task. We introduce different multimodal query processing strategies to enrich query terms and assign different weights to them. Specifically, sequential dependence model *(SDM)*, pseudo relevance feedback *(PRF)*, fielded sequential dependence model *(FSDM)* and Divergence from Randomness model *(DFRM)* are respectively performed on different fields of PubMed articles, sentences extracted from relevant articles, the five terminologies or ontologies (MeSH, GO, Jochem, Uniprot and DO) to achieve better search performances. Preliminary results show that our systems outperform others in the document and snippet retrieval task in the first two batches.

1 Introduction

Due to the continuous growth of information produced in the biomedical domain, there is a particularly growing demand for biomedical QA from the general public, medical students, health care professionals and biomedical researchers (Zweigenbaum, 2003). They consult knowledge about the natures, the preventions or the treatments of diseases, or learn from research results of other researchers. To some extent, biomedical QA is one of the most significant applications of the existing real-world biomedical systems (Han and Athenikos, 2010).

Since 2013, BioASQ organizers has proposed a community-based shared task which aims to evaluate the current solutions of a variety of QA subtasks. Several benchmarks have been provided for researchers to evaluate their QA systems. BioASQ 2017 Task 5B challenge (Tsatsaronis et al., 2015a) is the fifth edition of the question answering task, of which the phase A requires the evaluated system to (i) semantically annotate the questions with concepts from a set of designated terminologies and ontologies (MeSH, GO, Jochem, Uniprot and DO); and (ii) retrieve relevant articles, text snippets, and RDF triples from designated article repositories and ontologies (PubMed/MEDLINE articles) with biomedical questions in natural language provided by biomedical professionals or researchers. The ground truth are manually annotated by these experts with some annotated tools. There are five batches of evaluation and in each batch participants are provided with 100 natural language questions and required to return at most 10 relevant documents, snippets, concepts to the questions within 24 hours.

Over the past decade, a variety of approaches have been proposed for biomedical question answering (Bauer and Berleant, 2012). Generally, a QA system typically consists of question processing, document processing, and answer processing phases, which are respectively in charge of 1) converting natural language questions into queries, 2) searching relevant documents, and 3) extracting, ranking candidate answers and formatting them into expected answer type. (Han and Athenikos, 2010; Holzinger et al., 2014). There are several studies concerning the improvements on query processing phase (Huang et al., 2006; Yu et al., 2005; Kobayashi and Shyu, 2006) and document processing phase (Cairns et al., 2011; Yu and Cao, 2008). However for answer processing phase, especially answer matching and ranking, only some simple approaches in previous BioASQ challenge have been proposed (Tsatsaronis et al., 2015a; Mao and Lu, 2015). According to the above researches, the most challenges of biomedical QA are three main issues, specifically 1) how to generate query terms appropriately from natural language questions, 2) how to match relevant documents or sentences when they use different expressions (maybe synonyms of keywords) and 3) how to measure and utilize the difference in importance of query terms.

In order to address these challenges, in this paper we propose a multi-strategy query processing approach which combines several mature query processing models according to the different characteristics of data sources, which is also actually the participation of our USTB_PRIR team in the BioASQ Task 5B phase A challenge[1]. Specifically, in order to extract proper keywords and generate queries, we perform stop-words removal, noun extraction with Pos-of-Tagger (POS) and stemming. For the missing issue caused by expressions, we utilize a thesaurus which is produced through computing the similarities between the vector representations of each pairs of words. Moreover for query keyword weighting, we take the word sequences, different fields of appearance, TF-IDF, etc into consideration for different BioASQ tasks. We evaluate our approach on the BioASQ 2016 and 2017 benchmarks for document, snippet, concept retrieval and experimental results demonstrate our method outperforms the baseline methods or other participants so far on document, snippet and concept retrieval tasks.

2 Related Work

The participants of previous BioASQ challenge have proposed several approaches for searching relevant documents, snippets and concepts for biomedical QA. One of the participants(Choi, 2015) proposed to utilize semantic concept enriched dependence model where the recognised UMLS concepts in the query are used as additional dependence features for ranking documents. Another team(Papanikolaou et al., 2014) developed a figure-inspired text retrieval method as a way of retrieving documents and text passages from biomedical publications. For matching relevant snippets, most participants works on similar methods of searching articles. An exception is the framework proposed by NCBI(Mao et al., 2014), which directly compute the cosine similarities between the questions and the sentences.

However, these methods focus on the matching function or the ranking process, which ignores the three challenges mentioned in the Introduction section. The natural language questions are too raw to be regarded as query keywords and the difference in importance of keywords should be considered. Some re-ranking or learning-to-rank based approaches works not well either for the same reason because they rely much on the initial ranking results.

3 Task 5B Phase A: Document Retrieval

3.1 The Framework Architecture

The framework of searching relevant documents is shown in Figure 1, which includes document pre-processing, query pre-processing, several ranking models based on query expansion and term weighting strategies.

3.2 Pre-Processing

3.2.1 Document Pre-processing

We download the entire database of MEDLINE updated in Feb 2017 through the FTP service of National Institutes of Health (NIH) which contains 26,759,010 citations. These documents are represented in JSON files which contains a variety of information, including journal information, contents of title, author, abstract and keywords, similar articles and comments. We analyze the resources and select the following fields to represent the documents: *ArticleTitle*, *AbstractText*, *Title*, *MedlineTA*, *NameOfSubstance*, *DescriptorName*, *QualifierName*, *Keyword* and *ISOAbbreviation*. These fields are extracted from the document resources and indexed with Galago, an open source search engine[2], which is developed as an improved JAVA version of Indri. We also perform stemming and stop-words removal work like other IR applications, however unfortunately, the performances seems worse during the training process. As a result, we decide not to utilize these strategies for document pre-processing.

3.2.2 Query Pre-Processing

As is mentioned above, one of the challenges is how to automatically generate the query terms from a natural language question. During query pre-processing, we carry out a series of work to extract the keywords of the user queries. There

[1] http://bioasq.org

[2] http://www.lemurproject.org/galago.php

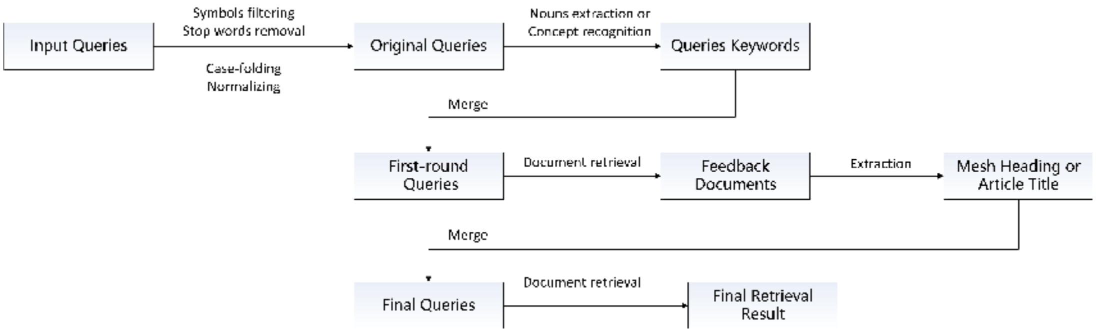

Figure 1: The whole framework architecture of query generation method based on multimodal document retrieval strategies

are several symbols which is unnecessary and unrelated to the requests so we filter out the symbols in the first step. Note that the symbols which may be a part of named entity cannot be removed. Afterwards, stop words like "what" or "are" are common in natural language questions and are not suitable to feed into search engine so they are removed according to a stop-words list. As usual, the query terms are case-folded and normalized. In addition, we used the Stanford-Postagger package to identify nouns from queries and the MetaMap to identify the biomedical concepts in query terms.

3.3 Ranking Models

3.3.1 Sequential Dependence Model

The traditional IR techniques in biomedical domain rely on a unigram Bag of Words (BoW) retrieval model. Each document in the collection of candidates, as well as each query, is represented by a set of words and the corresponding frequency based on the assumption that the appearance of each pair of words are independent. Different sequence of queries is regarded as the same. Consider an example of two documents that contain all query keywords. It is obvious that the document with the right sequence of terms appearing in query is more likely to meet the demand. Therefore, we introduce the Sequence Dependency Model (SDM) (Bonnefoy et al., 2012) to take the sequence information into account when computing the relevance between a document and a query.

SDM is a special case of the Markov Random Field (MRF) (Metzler and Croft, 2005). In order to capture the information of a sentence, this model extracts the phrases in different ways, and gives corresponding weights to different types of phrases to indicate their importance.

There are three features in the SDM to be considered: single-word features (a collection consists of single-word, Q_T), ordered bi-words phrase features (the two words in a phrase appearing in order, Q_O) and unordered window features (one or several words can be allowed appearing between the two words, Q_U). Generally, the potential function for unigrams (single-word feature) looks as follows:

$$f_T(q_i, D) = \log P(q_i|\theta_D) = \log \frac{tf_{q_i,D} + \mu\frac{cf_{q_i}}{|C|}}{|D| + \mu} \tag{1}$$

where q_i is a query term, D is a document, $tf_{q_i,D}$ is the frequency of q_i in D, $|D|$ is the document length, μ is a Dirichlet prior, that is usually set to the average document length in the collection, cf_{q_i} is the collection frequency of q_i and $|C|$ is the total number of terms in the collection. Similarly, for ordered and unordered bi-grams, the potential functions are respectively as follows:

$$\begin{aligned} f_O(q_i, q_{i+1}, D) &= \log P(\#1(q_i, q_{i+1})|\theta_D) \\ &= \log \frac{tf_{\#1(q_i,q_{i+1})} + \mu\frac{cf_{\#1(q_i,q_{i+1})}}{|C|}}{|D| + \mu} \end{aligned} \tag{2}$$

$$\begin{aligned} f_U(q_i, q_{i+1}, D) &= \log P(\#uwN(q_i, q_{i+1})|\theta_D) \\ &= \log \frac{tf_{\#uwN(q_i,q_{i+1})} + \mu\frac{cf_{\#1(q_i,q_{i+1})}}{|C|}}{|D| + \mu} \end{aligned} \tag{3}$$

where $\#1(q_i, q_{i+1})$ and $\#uwN(q_i, q_{i+1})$ are respectively the appearances of the exact phrase $q_i q_{i+1}$ and the term $q_i, q_{i+1}]$ within a window N terms. Hence, the scoring function of a document in SDM is the combination of the above three functions, shown as follows:

$$
\begin{aligned}
score_{\text{SDM}} &= score_{\text{SDM}}(Q_T, Q_O, Q_U, D) \\
&= \lambda_T \sum_{i=1}^{|Q|} f_T(q_i, D) \\
&+ \lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i, q_{i+1}, D) \\
&+ \lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i, q_{i+1}, D)
\end{aligned} \tag{4}
$$

Where Q is a sequence of keywords extracted from a user query, D is a candidate document, q_i is the i-th query keyword of Q. f_T,f_O,f_U are the maximum likelihood estimations of the corresponding feature terms in document D. λ_T,λ_O,λ_U are the features weights satisfy these conditions:

(1)$0 \leq \lambda_T, \lambda_O, \lambda_U \leq 1$ and $\lambda_T + \lambda_O + \lambda_U = 1$

(2)$\lambda_T \geq 0.6$

(3)$\lambda_O = 2\lambda_U$

Often, $\lambda_T = 0.85$,$\lambda_O = 0.1$, $\lambda_U = 0.05$.

3.3.2 Fielded Sequential Dependence Model

As is mentioned, the candidate documents are structured into several fields which contains different types of information. One of the limitations of standard SDM for structured document retrieval is that it considers term matches in different parts of a document as equally important (i.e. having the same contribution to the final relevance score of a document), thus disregarding the document structure.

To adapt the MRF framework to multi-fielded entity descriptions, we introduce (Zhiltsov et al., 2015)'s approach from their FSDM model to replace a single document language model $P(q_i|\theta_D)$ with a mixture of language models (MLM) for each document field. Consequently, the potential function for unigrams in case of FSDM is:

$$
\widetilde{f}_T(q_i, D) = \log \sum_j w_j P(q_i|\theta^j) \tag{5}
$$

where j represents the different fields, and the $P(q_i|\theta^j)$ is the language model in each individual field. Similarly, we can compute $\widetilde{f}_O(q_i, q_{i+1}, D)$ and $\widetilde{f}_O(q_i, q_{i+1}, D)$. Therefore, the scoring function of FSDM is as follows:

$$
\begin{aligned}
score_{\text{FSDM}} &= score_{\text{FSDM}}(Q_T, Q_O, Q_U, D) \\
&= \lambda_T \sum_{i=1}^{|Q|} \widetilde{f}_T(q_i, D) \\
&+ \lambda_O \sum_{i=1}^{|Q|-1} \widetilde{f}_O(q_i, q_{i+1}, D) \\
&+ \lambda_U \sum_{i=1}^{|Q|-1} \widetilde{f}_U(q_i, q_{i+1}, D)
\end{aligned} \tag{6}
$$

3.3.3 Pseudo Relevance Feedback

With the first-pass retrieval results, we assume that the initially retrieved top-K documents are relevant to questions, and their title and mesh fields contain relevant terms to the original query (Zhang et al., 2015a). Thus, for document retrieval, we use the Pseudo Relevance Feedback (PRF) to enrich query terms from the top-K document initially retrieved. The titles or mesh headings of the top-K documents are extracted and then added to the original query term set. However, the performance of PRF can be affected by the quality of the initial result, the number of pseudo-relevant documents (top K), the number of expansion terms, and the term re-weighting method applied. In our experiments, we use $K = 3$ and extract all the words in title or mesh headings as the expansion terms, which results in the best performance.

3.3.4 Multimodal Strategies Combination

Since there are several strategies to enrich query terms and optimize their weights, the final scoring function is expected to make full use of these strategies and combine these strategies effectively. We take the importance of nouns, sequence orders and crucial fields into consideration so our weight optimization of query terms is based on to noun extraction, sequential dependence model (SDM), Fielded sequential dependence model (FSDM), and Pseudo Relevance Feedback (PRF). According to massive experiments we find out that for some questions, the original queries, noun queries and enriched queries with PRF from relevant articles are all useful to some degree. Furthermore, we also find out that it is necessary to both search in the full text of the document, and to assign different weights to different fields at meanwhile. Hence, the final scoring function to search relevant documents is shown as follows:

Table 2: MAP performances compared with BioASQ Task 5B document retrieval participants.

System	Batch 1	Batch 2
sdm + NN + fsdm	0.1049	0.0850
sdm + NN + fsdm + PRF (mesh)	**0.1086**	0.0863
sdm + fsdm + PRF (mesh)	0.1032	0.0859
sdm + w2v	0.0928	**0.0874**
sdm	0.0952	0.0866
best of fdu	0.1072	0.0834
best of UNCC	0.1080	-
best of Olelo	0.0465	0.0318
best of KNU-SG	0.0413	0.0419
best of HPI	0.0307	0.0329
best of Others	0.0437	0.0265

$$score(Q,D) = \lambda_1 score_{\text{SDM}}(Q,D) + \lambda_2 score_{\text{FSDM}}(Q',D) + \lambda_3 score_{\text{SDM}}(Q'',D) \quad (7)$$

where Q is the original query term set after query pre-processing, Q' represents the noun query term set with noun extraction and the Q'' stands for the enriched query term set with PRF.

3.4 Experments

We evaluate our proposed method by using both the benchmark datasets from the previous BioASQ challenges and the current challenge. The optimization of all parameters, including the weighting paramters like w_j in FSDM function and hyper-parameters (e.g. λ_T, λ_O, λ_U, λ_1, λ_2, λ_3) are processed through tuning with the rules on training set (when evaluated on BioASQ Task 4B, the training set includes 800 questions from BioASQ 2B and 3B; for BioASQ Task 5B, the training set contains 500 more questions on BioASQ 4B). Table 1 provides the results of our experiments in BioASQ task 4B, and Table 2 provides the results of our experiments in BioASQ Task 5B. The $sdm + w2v$ approach refers to our previous approach in (Zhang et al., 2015b). Obviously, our proposed method shows greater performance compared with baseline, SDM and FSDM and outperform than other participants in current challenge.

4 Task 5B Phase A: Snippet Retrieval

4.1 The Framework Architecture

The framework of searching relevant snippets is shown in Figure 2, which includes pre-processing, some additional ranking models which is different from document retrieval.

4.2 Pre-Processing

The query pre-processing for snippets retrieval is the same to the strategies for document retrieval, which includes unnecessary symbol removal, stop-words removal, case-folding, noun extraction and concept extraction with Metamap.

For the snippet pre-processing, we choose the candidate snippets from the top-K documents of the best performed document retrieval approach on the basis of results of document retrieval. The sentences with the field ArticleTitle and the field abstract of these articles are separated through some specific rules, which can be regarded as "small documents". These sentences make up a pile of new files with unstructured text. They are then indexed by Galago for search in the next step.

4.3 Ranking Models

Different from document retrieval, the candidate snippets are represented in unstructured text, which makes some ranking models more difficult to utilize (e.g. FSDM). Moreover, since they are much shorter in length, they are more likely express similar meaning with different expressions (e.g. synonyms) which may emphasize the importance of the issue of recognizing these relevant results. Furthermore, the PRF method generally provides massive expansion query terms, which may affect the search performance of the short text so we give up applying PRF as query expansion method.

In addition, we introduce DFRM from (Clinchant and Gaussier, 2011) as an additional term weigting model to optimize the most appropriate weight for query terms.

4.3.1 Divergence from Randomness Model

The Divergence from Randomness models (DFRM) are based on this simple idea: "The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d". In other words the term-weight is inversely related to the probability of term-frequency within the document d obtained by a model M of randomness:

$$\text{weight}(t|d) \propto -\log \text{Prob}_M(t \in d|\text{Collection}) \quad (8)$$

where the subscript M stands for the type of model of randomness employed to compute the probability. In order to choose the appropriate model M of randomness, we can use different urn models.

Table 1: MAP performances of system components on BioASQ Task 4B document retrieval.

	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
baseline	0.2056	0.2593	0.228	0.2324	0.2516
sdm	0.2214	0.2577	0.2436	0.2517	0.2935
fsdm	0.2156	0.2621	0.2228	0.2469	0.2728
sdm + fsdm	0.2269	0.2768	0.2447	0.2608	0.2968
sdm + NN + fsdm	0.2307	0.2741	0.2454	0.2632	0.2926
sdm + fsdm + PRF(title)	0.2337	0.2778	0.2455	0.265	0.2931
sdm + fsdm + PRF(mesh)	0.2372	0.2863	**0.2564**	0.2762	0.3019
sdm + NN + fsdm + PRF(mesh)	0.2436	0.2859	0.2465	0.2773	**0.3083**
sdm + NN + fsdm + PRF(title)	0.2377	0.2767	0.2429	0.2681	0.2985
sdm + NN + fsdm + mesh + PRF(mesh)	**0.2440**	**0.2876**	0.2505	**0.2821**	0.3059

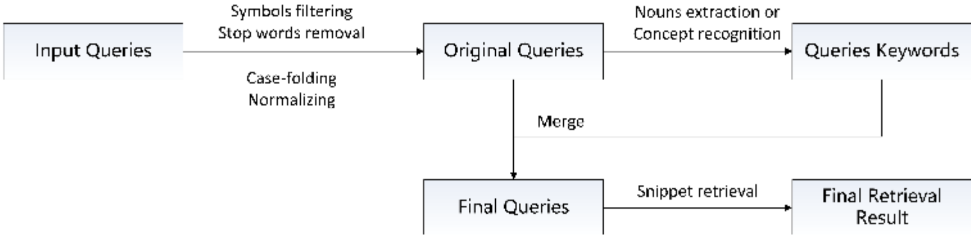

Figure 2: The whole framework architecture of query generation method based on multimodal snippet retrieval strategies

Table 3: Basic DFR Models.

D	Divergence approximation of the binomial
P	Approximation of the binomial
B_E	Bose-Einstein distribution
G	Geometric approximation of the Bose-Einstein
$I(N)$	Inverse Document Frequency model
$I(F)$	Inverse Term Frequency model
$I(n_e)$	Inverse Expected Document Frequency model

There are many ways to choose M, each of these provides a basic DFR model. The basic models are derived in Table 4.

If the model M is the binomial distribution, then the basic model is P and the value can be computed approximately as follows:

$$-\log \text{Prob}_P(t \in d|\text{Collection}) = -\log(\genfrac{}{}{0pt}{}{TF}{tf})p^{tf}q^{TF-tf} \tag{9}$$

where TF is the term-frequency of the term t in the collection, tf is the term-frequency of the term t in the document d, N is the number of documents in the collection, and p is $\frac{1}{N}$ and $q = 1 - p$.

4.3.2 Multimodal Strategies Combination

Similar to document retrieval, the final scoring function of snippet retrieval is expected to combine these strategies together effectively. Due to the reason of shorter text length the FSDM model cannot be used and the default IR language model performs not so satisfying for returning relevant snippets, we construct the merging scoring function to optimize the query term weights according to the Term Frequency—Inverse Document Frequency (TF-IDF), sequential dependence model (SDM) and Divergence from Randomness model (DFRM). As mentioned above, we control the length of queries to guarantee the performance, thus we no longer use PRF for snippets retrieval when merging the strategies. As the length of the queries decreases, the divergence of importance of each word becomes larger, so it is necessary to assign the weights of query terms according to the different importance. So we apply the DFRM method or the TF-IDF method along with SDM to achieve the results, which are respectively shown as follows:

$$\begin{aligned} score(Q,D) =&(1-\lambda_1-\lambda_2)score_{\text{SDM}}(Q,D) \\ &+\lambda_1 score_{\text{TF-IDF}}(Q,D) \\ &+\lambda_2 score_{\text{DFRM}}(Q,D) \end{aligned} \tag{10}$$

where the terms are weighted according to corresponding strategies through the following weighting function:

$$score_{\text{TF-IDF/DFRM}}(Q,D) = \sum_t score_{\text{TF-IDF/DFRM}}(t,D) \tag{11}$$

where t is the query term appearing in query Q. It is worth noting that when conducting the experiments we only consider $\lambda_1 = 0$ or $\lambda_2 = 0$ for tuning parameters.

Table 4: MAP performances of system components on BioASQ Task 4B snippet retrieval.

	Batch 1	Batch 2	Batch 3
baseline	0.1003	0.1361	0.1275
sdm	0.1047	0.1368	0.1338
sdm + PRF	0.1044	0.1370	0.1327
sdm + NN	0.1023	0.1402	0.1347
sdm + NN + PRF	0.1030	0.1357	0.1307
sdm + DFRM	**0.1193**	**0.1520**	**0.1469**
sdm + TF-IDF	0.1087	0.1424	0.1357

Table 5: MAP performances compared with BioASQ Task 5B snippet retrieval participants.

System	Batch 1	Batch 2
sdm + NN	0.0458	0.0811
sdm + NN + PRF(mesh)	0.0439	0.0716
sdm + DFRM	**0.0467**	**0.0898**
sdm + TF-IDF	0.0463	0.0874
sdm	0.0452	0.0736
best of fdu	-	0.0621
best of UNCC	-	-
best of Olelo	0.0260	0.0318
best of KNU-SG	0.0181	0.0362
best of HPI	0.0323	0.0335
best of Others	0.0249	0.0262

4.4 Experments

Similar to document retrieval, we evaluate the method on the first 3 batches from the previous BioASQ challenge and the current challenge. Similar to document retrieval, the optimization of all parameters are processed through tuning with the rules on training set. Table 4 provides the results of our experiments in BioASQ task 4B, and Table 5 provides the results of our experiments in BioASQ Task 5B. Obviously, our proposed merging strategy shows greater performance compared with various components and achieve better results than other particatants.

5 Task 5B Phase A: Concept Retrieval

Unlike the previous two tasks, the concept retrieval task is more like a named entity recognition task than an IR task. For each natural language question, participants are required to return relevant concepts from five ontologies or terminologies: MeSH, GO, Jochem, Uniprot and DO. In other words, the task aims at recognizing relevant biomedical concept within the question and matching them with the concepts in the data sources.

Since we have few experience in named entity recognition, we have to regard the task as an IR problem and design three query processing approaches to generate appropriate query keywords for the web search services provided by BioASQ officials and implement the requested JSON file according to the examples in the guidelines (Neves, 2014). The five URLs of web services are utilized to post search requests for concepts and obtain search results. The request consists of two basic elements: keywords, the query to feed into search engine. Typically, this is a simple set of phrases separated by spaces acting as queries which may contain alphanumeric and punctuation characters; *page* and *concepts-per-page*, to control the number of results since the search engine may return thousands of concepts for one query. Thus, a pagination mechanism is used. Specifically, *Page* is a number representing the page (batch of concepts) to be retrieved, and *concepts-per-page* is a number representing the number of concepts per page (Tsatsaronis et al., 2015b).

Table 6: MAP performances of system components on BioASQ Task 4B concept retrieval.

	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Ours	0.1094	**0.1124**	**0.1386**	0.1174	0.1031
fdu	-	-	0.1566	0.1319	0.1004
HPI	0.0860	-	0.0863	0.0721	0.0439
oaqa	-	-	0.1067	0.1332	0.0915
auth	**0.1433**	0.0814	0.1361	**0.1376**	**0.1066**

For concept retrieval, noun extraction, synonym query expansion and pseudo relevance feedback are respectively used. On the purpose of obtaining the synonyms of query keywords, we download the vector representations of vocabularies produced through google word2vec tool (a word embedding tool to train word vectors on corpora), provided by BioASQ officials. We compute the cosine similarity between each query keyword and the word in the word list to find out the most semantic related words. These words are regarded as synonyms of the query keywords. We select the top 10 concepts as the submitted results ordered by descending predicted relevance score to the corresponding queries.

Since the results for this subtask will only be available after the manual assessment phase, we only evaluate the proposed method on the BioASQ 4B with other participants or any runs submitted off the evaluation. Table 6 provides the results of our experiments in BioASQ Task 4B and the statistics indicate our approach shows fairly good performance on all batches.

6 Conclusion

In this paper, we describe how to utilize multi-modal query processing strategies for biomedical question answering applied to the participation of our USTB_PRIR team on phase A of BioASQ Task 5B. According to the official results, our system shows great robustness and effectiveness with competitive performance among the participating systems.

During the study of concept retrieval, we realize that named entity recognition of biomedical concepts may be helpful for the other tasks and so we may focus on utilizing this in the future.

References

Michael A Bauer and Daniel Berleant. 2012. Usability survey of biomedical question answering systems. *Human genomics*, 6(1):17.

Ludovic Bonnefoy, Romain Deveaud, Patrice Bellot, P Forner, J Karlgren, and C Womser-Hacker. 2012. Do social information help book search? In *INEX*, volume 12, pages 109–113.

Brian L Cairns, Rodney D Nielsen, James J Masanz, James H Martin, Martha S Palmer, Wayne H Ward, and Guergana K Savova. 2011. The mipacq clinical question answering system. In *AMIA Annual Symposium Proceedings*, volume 2011, page 171. American Medical Informatics Association.

Sungbin Choi. 2015. Snumedinfo at CLEF bioasq 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

Stphane Clinchant and Eric Gaussier. 2011. Bridging language modeling and divergence from randomness models: A log-logistic model for ir.

Hyoil Han and Sofia J Athenikos. 2010. Biomedical question answering: a survey. *Computer Methods & Programs in Biomedicine*, 99(1):1–24.

Andreas Holzinger, Matthias Dehmer, and Igor Jurisica. 2014. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, 15(Suppl 6):I1.

X. Huang, J. Lin, and D Demnerfushman. 2006. Evaluation of pico as a knowledge representation for clinical questions. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2006:359–63.

T Kobayashi and C. R. Shyu. 2006. Representing clinical questions by semantic type for better classification. *Proceedings of the AMIA 2006 Symposium*, 2006:987.

Yuqing Mao and Zhiyong Lu. 2015. NCBI at the 2015 bioasq challenge task: Baseline results from mesh now. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

Yuqing Mao, Chih-Hsuan Wei, and Zhiyong Lu. 2014. NCBI at the 2014 bioasq challenge task: Large-scale biomedical semantic indexing and question answering. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 1319–1327.

Donald Metzler and W. Bruce Croft. 2005. A markov random field model for term dependencies. In *SIGIR*, pages 472–479.

Mariana L Neves. 2014. Hpi in-memory-based database system in task 2b of bioasq. In *CLEF (Working Notes)*, pages 1337–1347.

Yannis Papanikolaou, Dimitrios Dimitriadis, Grigorios T-soumakas, Manos Laliotis, Nikos Markantonatos, and Ioannis P. Vlahavas. 2014. Ensemble approaches for large-scale multi-label classification and question answering in biomedicine. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 1348–1360.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel Ngonga, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015a. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015b. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

H. Yu and Y. G. Cao. 2008. Automatically extracting information needs from ad hoc clinical questions. *Amia Annual Symposium Proceedings*, 2008:96–100.

Hong Yu, Carl Sable, and Hai Ran Zhu. 2005. Classifying medical questions based on an evidence taxonomy. In *National Conference on Artificial Intelligence*.

Yanchun Zhang, S Peng, R You, Z Xie, B Wang, and Shanfeng Zhu. 2015a. The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering. In *CEUR Workshop Proceedings*, volume 1391. CEUR Workshop Proceedings.

Zhijuan Zhang, Tiantian Liu, Bo-Wen Zhang, Yan Li, Chun Hua Zhao, Shao-Hui Feng, Xu-Cheng Yin, and Fang Zhou. 2015b. A generic retrieval system for biomedical literatures: Ustb at bioasq2015 question answering task. In *CLEF (Working Notes)*.

Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. 2015. Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–262.

Pierre Zweigenbaum. 2003. Question answering in biomedicine. In *Proceedings Workshop on Natural Language Processing for Question Answering, EACL*, volume 2005, pages 1–4. Citeseer.