A Generic Document Retrieval Framework Based on UMLS Similarity for Biomedical Question Answering System

Mourad Sarrouti and Said Ouatik El Alaoui

Abstract Biomedical document retrieval systems play a vital role in biomedical question answering systems. The performance of the latter depends directly on the performance of its biomedical document retrieval section. Indeed, the main goal of biomedical document retrieval is to find a set of citations that have high probability to contain the answers. In this paper, we propose a biomedical document retrieval framework to retrieve the relevant documents for the biomedical questions (queries) from the users. In our framework, we first use GoPubMed search engine to find the top-K results. Then, we re-rank the top-K results by computing the semantic similarity between questions and the title of each document using UMLS similarity. Our proposed framework is evaluated on the BioASQ 2014 task datasets. The experimental results show that our proposed framework has the best performance (MAP@100) compared to the existing state-of-the-art related document retrieval systems.

Keywords Information retrieval · Biomedical question answering system · Gopubmed · Unified modeling language system · Semantic similarity

1 Introduction

By the rapidly increasing of knowledge in the biomedical domain, it becomes very difficult even for experts to absorb all the relevant information in their field of interest. Information Retrieval (IR) systems present a list of document that might have the associated information, but the majority of them leave it to the user to find and extract the required information [8]. For example, the biomedical question "Is the PTPN22 gene a biomarker for Rheumatoid Arthritis?" from BioAsk training datasets, should get back the response "Yes", but instead the user is presented with a large number of

M. Sarrouti (✉) · S.O. El Alaoui
Laboratory of Computer Science and Modeling, FSDM, Sidi Mohammed Ben Abdellah University, Fes, Morocco
e-mail: mourad.sarrouti@usmba.ac.ma

S.O. El Alaoui
e-mail: said.ouatikelalaoui@usmba.ac.ma

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies 2016*,
Smart Innovation, Systems and Technologies 57,
DOI 10.1007/978-3-319-39627-9_18

documents that are potentially relevant to explore in the quest of an accurate answer. Unlike IR systems, Question Answering (QA) systems aim to provide inquirers with direct and precise answers to their questions, by employing Information Extraction (IE) and Natural Language Processing (NLP) methods [3]. In other words, QA systems allow to quickly get precise answers to user's questions with the least amount of reading required.

Typically an automated QA system consists of three main elements, which independently can be studied and developed, [3, 9, 11, 14]: Questions Processing, Documents Processing and Answers Processing. Figure 1 illustrates the generic architecture of a biomedical QA system. For a given biomedical question written in natural language, the Question Processing phase aims to analyze the question and create IR query, identifying the type of question as well [16] . Indeed, the first task is called Query Reformation and the second is called Question Classification [11, 16]. The document Processing phase allows to process the returned documents by an IR system and provides candidate passages which could probably contain the answer. Finally, the type of question identified in the question processing and the candidates answers generated in the document processing are used in the Answer Processing phase in order to extract the final answer.

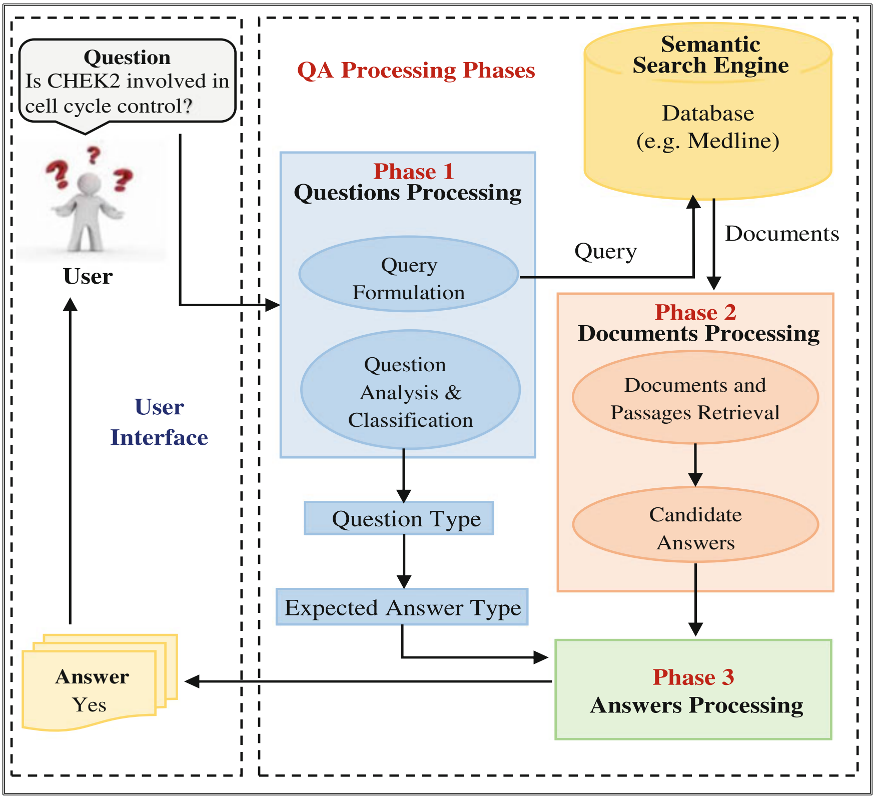

Fig. 1 Generic architecture of questions answering system

Previously, we have addressed the problem of biomedical question classification [16] and in this paper we are interested on biomedical document retrieval, which is an important component of biomedical QA systems. As we mentioned earlier, the task of biomedical document retrieval is to find a list of relevant documents that are likely to contain the answer. In other words, if a list of relevant documents is determined correctly, it can be useful for finding the location of the answer. Therefore, the task of document retrieval has a significant impact on the overall performance of the biomedical QA system.

In light of this, GoPubMed [7] is considered one of the information retrieval tools most widely used in biomedical QA systems to access the MEDLINE database. MEDLINE is a major biomedical literature database repository, which is supported by the U.S. National Library of Medicine (NLM). In other words, the goal of GoPubMed, like all other search engines, is to retrieve documents considered relevant to a user query. Researches have done great effort to optimize retrieval result rankings, hoping to place the most relevant ones at the top of the ranking list. Nevertheless, no ranking solution is perfect, due to the inherent complexity of ranking search results. For instance, in [13], GoPubMed has been used as biomedical document retrieval. The query reformulation component includes sentence splitting, tokenization, part-of-speech tagging and chunking using the Stanford CoreNLP. They have also kept the top 100 documents returned by GoPubMed.

In this paper, we propose a novel biomedical document retrieval framework. Comparing with the previous works, our proposed method has the following contributions:

1. The novel framework is based on GoPubMed semantic search engine and our query reformulation method.
2. In our knowledge, this is for the first time that UMLS similarity has been used for re-ranking the top-K citations and keep the top-k ones which have high probability to contain the answers.

The remainder of the paper is organized as follows. Section 2 introduces related work and discussion about the main biomedical QA approaches. Section 3 describes the overall architecture of the proposed framework. Section 4 presents our experiments on a benchmark dataset and the results of our biomedical document retrieval framework. Finally, conclusion and future work are made in Sect. 5.

2 Related Work

Although research on QA systems has boomed in recent years, document retrieval has been a large part in the research community of text mining after the introduction of QA Track in the Text REtrieval Conference (TREC[1]) in 1999 as well as the presentation of biomedical QA in the BioASK[2] [18].

[1] http://trec.nist.gov/.

[2] http://bioasq.org/.

However, QA system has been a well studied research area [15]. Biomedical QA system has its own challenges such as the presence of complex technical terms, compound words, domain specific semantic ontologies, domain-specific format and typology of questions [3].

MedQA [10] is a biomedical QA system which generates paragraph-level answers from the MEDLINE collection. The system consists of information retrieval, extraction, and summarization techniques to automatically generate paragraph-level answers for definitional questions. For query formulation and document retrieval, the system use a shallow syntactic parser and a standard IR engine.

Abacha and Zweigenbaum [1] have described their implemented medical QA system called MEANS. The system consists of three main steps: corpora annotation, question analysis, and answer search. The authors have exploited natural language processing techniques as well as biomedical and semantic resources (e.g. UMLS) to build RDF annotations of the source documents and SPARQL queries representing the users questions.

In [6], SNUMedinfo team has leased 2014 MEDLINE/PubMed Journal Citations and used Indri search engine [17]. In fact, they have experimented with semantic concept-enriched dependence model and sequential dependence model. They have also shown that the semantic concept enriched dependence model showed significant improvement over baseline.

In [19], Weissenborn et al. have presented a biomedical QA system which is composed of three subsystems: question analysis, document retrieval, and answer extraction. They have used GoPubMed search engine in order to find the relevant documents to the question. in addition, the authors have completely depended on GoPubMed ranking of documents.

Continuing, Neves [13] has presented a biomedical document retrieval system. The latter includes sentence splitting, tokenization, part-of-speech tagging and chunking using the Stanford CoreNLP package for query reformulation. The approach is based on GoPubMed search engine. Neves has also completely depended on GoPubMed ranking of documents.

To our knowledge, all the above methods have not taken into account the semantic similarity between question and each title of document returned by a search engine in order to re-rank these documents again. In this paper, in order to improve the performance of document retrieval of biomedical QA systems, we propose a generic biomedical document framework based on UMLS similarity.

3 Proposed Method

In this section, we describe in detail our proposed document retrieval for biomedical QA system. Indeed, the main goal of this method is to find the high informative documents for a given biomedical question from PubMed articles.

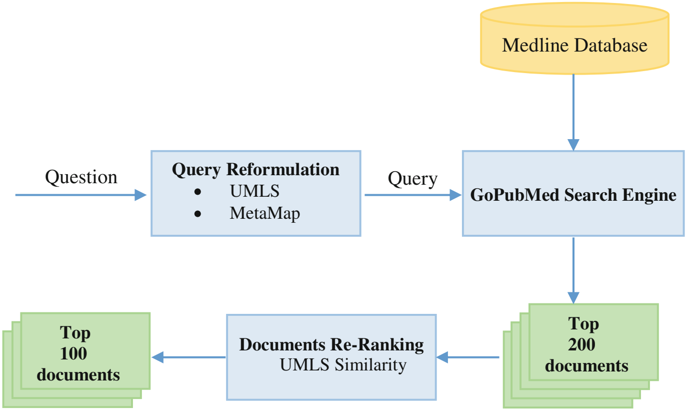

Fig. 2 Flowchart of our biomedical documents retrieval framework

Table 1 Example of mapping question to UMLS concepts

Question	UMLS concepts	CUI
Which are the	Cardiac	C0018787
Cardiac manifestation of	Manifestation of	C1280464
Marfan syndrome	Marfan syndrome	C0024796

CUI indicates Concept Unique Identifier

In order to achieve this goal, we construct a biomedical document retrieval framework to solve the semantic search by reformulating the query and using UMLS Similarity [12] to rank the returned biomedical documents. Moreover, we have proposed two algorithms. The first one allows to reformulate the query and search the top 200 documents using GoPubmed[3] web service [7]. The second one aims to re-rank the top 200 documents and keep only a set of 100 PubMed documents which have high probability to contain the answer. The flowchart of our biomedical document retrieval framework is presented in Fig. 2 and below are the various steps described in details.

1. **Query Reformulation**: in this step, we process the biomedical question, written in natural language, to make it efficient and optimized for searching. Indeed, We have used MetaMap [2] for mapping terms in questions to Unified Medical Language System (UMLS) in order to extract the Biomedical Entity Names (BENs) and connect them with the "AND" operator. The UMLS [5] is a repository developed by the US National Library of Medicine, integrating over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies as well as 12 million relations among these. Table 1 illustrates an example of mapping question to UMLS using Metamap. Besides, based on the above definition, the query reformulation method is defined in the proposed Algorithm 1.

[3] http://gopubmed.org/web/gopubmedbeta/bioasq/pubmedmedline.

Algorithm 1 Query Reformulation and List of Top Documents

```
Input ← Question
Output ← List_of_Top_Documents(200)
function SEARCHDOCUMENTS(Question)
    Query ← QUERYREFORMULATION(Question)
    List_of_Top_Documents[200] ← GoPubMed_Web_Service(Query)
    return List_of_Top_Documents
end function
function QUERYREFORMULATION(Question)
    Biomedical_Entity_Names[N] ← Mapping_Qestion_to_UMLS(Question)
    do
        Query ← Query + "AND" + Biomedical_Entity_Names[i]
        $i \leftarrow i + 1$
    while i < N
    return Query
end function
```

2. **Pubmed Document Retrieval Using GoPubMed**: As it was shown in Algorithm 1, the query generated in the query reformulation phase will be fired to GoPubMed semantic search engine [7] in order to find the top 200 documents.
3. **Biomedical Document Re-Ranking**: the document re-ranking is the main and important step in the proposed framework. Indeed, we do not completely depend on GoPubMed ranking of documents. So we re-rank the obtained 200 documents again by computing the similarity between a given question and the title of each document. We have used UMLS similarity package[4] [12] to obtain similarity between biomedical concepts of a question and the concepts of document title. In fact, we have used path length as similarity measure where the similarity score is inversely proportional to the number of nodes along the shortest path between the concepts. Moreover, our proposed biomedical document Re-Ranking is presented in Algorithm 2.

4 Experimental Results and Discussion

In this section, we conduct our experiments on benchmark dataset to show the performance of our biomedical document retrieval framework. We first describe the dataset, then we present the experimental results, and finally discuss the results.

4.1 Datasets

To demonstrate the efficiency of the proposed framework, we perform experiments on benchmark dataset provided by biomedical experts. Actually, The experimental dataset comes from the official dataset of biomedical semantic QA Taskb phase

[4] http://maraca.d.umn.edu/cgi-bin/umls_similarity.cgi.

Algorithm 2 Biomedical Document Re-Ranking

```
Input ← Question, List_of_Top_Documents(200)
Output ← Top_Documents(100)
Q ← Question
relDocs ← List_of_Top_Documents(200)
function RankDocuments(Q, relDocs)
    scores ← {}
    do
        T[i] ← relDocs[i].title
        scores[i] ← ComputeSimilarity(Q, T[i])
        i ← i + 1
    while i < N
    scores, TopDocument[100] ← SortScores(scores, relDocs)
    return TopDocument[100]
end function
function ComputeSimilarity(Q, docTitle)
    question_concepts_CUI[N] ← Mapping_Question_to_UMLS(Question)
    docTitle_concepts_CUI[M] ← Mapping_Title_to_UMLS(docTitle)
    similarity ← 0
    sumSimilarity ← 0
    do
        QCUI ← Question_Concepts_CUI[i]
        do
            TCUI ← docTitle_Concepts_CUI[j]
            similarity ← UMLS :: Similarity(QCUI, TCUI)
            if similarity ≠ −1 then
                sumSimilarity ← sumSimilarity + similarity
            end if
            j ← j + 1
        while j < M
        i ← i + 1
    while i < N
    return sumSimilarity
end function
```

A [18]. In the dataset, there are five batches of questions in the testing set where includes 100 questions in each batch.

4.2 Evaluation Metrics

The typical evaluation measures used in IR are: mean precision, mean recall, mean F-measure and mean average precision (MAP) [18]. In fact, MAP is our main evaluation measure. For the test in 2014, the first 100 documents from the resulting list are permitted to be submitted.

Table 2 TOP 10 MAP@100 results of document retrieval systems [4] on batch 1 of BIOASQ 2014 and the results of our proposed framework

System	Mean precision	Mean recall	Mean F-measure	MAP
SNUMedinfo1	0.0457	0.5958	0.0826	0.2612
SNUMedinfo3	0.0457	0.5947	0.0826	0.2587
SNUMedinfo2	0.0451	0.5862	0.0815	0.2547
SNUMedinfo4	0.0457	0.5941	0.0826	0.2493
SNUMedinfo5	0.0459	0.5947	0.0829	0.2410
Top 100 baseline	0.2274	0.4342	0.2280	0.1911
Top 50 baseline	0.2290	0.3998	0.2296	0.1888
Main system	0.0413	0.2625	0.0678	0.1168
Biomedical text ming	0.2279	0.2068	0.1665	0.1101
Wishart-S2	0.1040	0.1210	0.0793	0.0591
Our system	**0.2331**	**0.3644**	**0.2253**	**0.2758**

4.3 *Results and Discussion*

To conduct the experiments, we have used the batch 1 of testing datasets (Benchmark dataset) of BioASQ 2014 task [18]. We first have applied MetaMap [2] to extract the Biomedical Entity Names of a given biomedical question and connect them with the "AND" operator in order to construct the query. Then, the latter will be fired to GoPubmed[5] semantic search engine [7] in order to find the top 200 documents (see Algorithm 1). After that, as we have not depended on GoPubMed ranking of documents, the proposed Algorithm 2 has been used in order to re-rank the 200 documents and keep only the top 100 documents. Table 2 presents the comparison between our results and the top 10 results on batch 1 of testing datasets in BioASQ 2014.

Overall, from Table 2, it can be seen clearly that the results of our proposed framework have an absolute competitiveness with the top 10 results in term of MAP. Indeed, the performance of our system was 0.2758 of MAP. Moreover, Our proposed framework significantly outperforms the baseline system (Top 100 Baseline) by a wide margin in term of mean average precision (0.0847 MAP).

In addition, Table 3 shows the results obtained by GoPubMed document ranking and our proposed algorithm (see Algorithm 2) for biomedical document re-ranking. We can see that when using GoPubMed document ranking, the mean average precision was 0.1439. While Algorithm 2 increased the performance to 0.2758 of MAP and the improvement is statistically significant. Hence, the proposed algorithm for document re-ranking plays a vital role on the overall performance of our framework.

Studies have shown that the biomedical document retrieval systems can improve the performance of biomedical QA systems, because the answers extraction is

[5] http://gopubmed.org/web/gopubmedbeta/bioasq/pubmedmedline.

Table 3 Results obtained by GoPubMed document ranking and our proposed algorithm 2 on batch 1 of BIOASQ 2014

System	Mean precision	Mean recall	Mean F-measure	MAP
Our query reformulation and GuPubMed document ranking	0.2253	0.3111	0.1913	0.1439
Proposed framework	**0.2331**	**0.3644**	**0.2253**	**0.2758**

based on the documents returned by document retrieval systems which have high probability to contain answers. Therefore, our proposed biomedical document retrieval framework can be used in order to find relevant documents to the biomedical question with high mean average precision. Moreover, the importance of our results using the proposed framework thus lies both in their generality and their relative ease of application to biomedical QA systems.

5 Conclusion and Future Work

In this paper, we have tackled an original biomedical document retrieval framework. First, we have used Metamap to extract biomedical named entities and connect them in order to generate queries. Then, the top 200 relevant documents are retrieved by GoPubMed search engine. Next, we have kept only the top 100 documents after re-ranking the top 200 documents by computing the semantic similarity between question and documents title. Finally, the experiments on the BioASQ 2014/2015 document retrieval task have demonstrated that our proposed framework is proved to be effective and competitive for biomedical documents retrieval compared to several state-of-the-art systems.

In our future work, we will focus on integrating our biomedical document retrieval framework in a biomedical QA system.

Acknowledgments The authors would like to thank BioASK challenges [18] for providing us with benchmark datasets.

References

1. Abacha, A.B., Zweigenbaum, P.: Means: a medical question-answering system combining nlp techniques and semantic web technologies. Inf. Process. Manag. **51**(5), 570–594 (2015)
2. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association (2001)

3. Athenikos, S.J., Han, H.: Biomedical question answering: a survey. Comput. Methods Programs Biomed. **99**(1), 1–24 (2010)
4. Balikas, G., Partalas, I., Ngomo, A.C.N., Krithara, A., Gaussier, E., Paliouras, G.: Results of the bioasq track of the question answering lab at clef 2014. Results of the BioASQ Track of the Question Answering Lab at CLEF 2014, 1181–1193 (2014)
5. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucl. Acids Res. **32**(suppl 1), D267–D270 (2004)
6. Choi, S., Choi, J.: Classification and retrieval of biomedical literatures: Snumedinfo at clef qa track bioasq 2014. In: Proceedings of Question Answering Lab at CLEF (2014)
7. Doms, A., Schroeder, M.: Gopubmed: exploring pubmed with the gene ontology. Nucl. Acids Res. **33**(suppl 2), W783–W786 (2005)
8. Dwivedi, S.K., Singh, V.: Research and reviews in question answering system. Procedia Technol. **10**, 417–424 (2013)
9. Gupta, P., Gupta, V.: A survey of text question answering techniques. Int. J. Comput. Appl. **53**(4), 1–8 (2012)
10. Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., Yu, H.: Beyond information retrieval medical question answering. In: AMIA Annual Symposium Proceedings, vol. 2006, p. 469. American Medical Informatics Association (2006)
11. Loni, B.: A Survey of State-of-the-Art Methods on Question Classification, pp. 01–40. Delft University of Technology, Delft (2011)
12. McInnes, B.T., Pedersen, T., Pakhomov, S.V.: Umls-interface and umls-similarity: open source software for measuring paths and semantic similarity. In: AMIA Annual Symposium Proceedings, vol. 2009, p. 431. American Medical Informatics Association (2009)
13. Neves, M.: Hpi in-memory-based database system in task 2b of bioasq. In: Proceedings of Question Answering Lab at CLEF (2014)
14. Neves, M., Leser, U.: Question answering for biology. Methods **74**, 36–46 (2015)
15. Ryu, P.M., Jang, M.G., Kim, H.K.: Open domain question answering using wikipedia-based knowledge model. Inf. Process. Manag. **50**(5), 683–692 (2014)
16. Sarrouti, M., Lachkar, A., Ouatik, S.E.: Biomedical question types classification using syntactic and rule based approach. In: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 265–272 (2015)
17. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis, vol. 2, pp. 2–6. Citeseer (2005)
18. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., et al.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. **16**(1), 138 (2015)
19. Weissenborn, D., Tsatsaronis, G., Schroeder, M.: Answering factoid questions in the biomedical domain. BioASQ@ CLEF 1094 (2013)