# A Biomedical Question Answering System in BioASQ 2017

**Mourad Sarrouti, Said Ouatik El Alaoui**
Laboratory of Computer Science and Modeling
Faculty of Sciences Dhar El Mahraz
Sidi Mohammed Ben Abdellah University
Fez, Morocco
mourad.sarrouti@usmba.ac.ma

## Abstract

Question answering, the identification of short accurate answers to users questions, is a longstanding challenge widely studied over the last decades in the open-domain. However, it still requires further efforts in the biomedical domain. In this paper, we describe our participation in phase B of task 5b in the 2017 BioASQ challenge using our biomedical question answering system. Our system, dealing with four types of questions (i.e., yes/no, factoid, list, and summary), is based on (1) a dictionary-based approach for generating the exact answers of yes/no questions, (2) UMLS metathesaurus and term frequency metric for extracting the exact answers of factoid and list questions, and (3) the BM25 model and UMLS concepts for retrieving the ideal answers (i.e., paragraph-sized summaries). Preliminary results show that our system achieves good and competitive results in both exact and ideal answers extraction tasks as compared with the participating systems.

## 1 Introduction

Finding accurate answers to biomedical questions written in natural language from the biomedical literature is the key to creating high-quality systematic reviews that support the practice of evidence-based medicine (Kropf et al., 2017; Wang et al., 2017; Sarrouti and Lachkar, 2017) and improve the quality of patient care (Sarrouti and Alaoui, 2017b). However, with the large and increasing volume of textual data in the biomedical domain makes it difficult to absorb all relevant information (Sarrouti and Alaoui, 2017a). Since time and quality are of the essence in finding answers to biomedical questions, developing and improving question answering systems are desirable. Question answering (QA) systems aim at directly producing and providing short precise answers to users questions by automatically analyzing thousands of articles using information extraction and natural language processing methods.

Although different types of QA systems have different architectures, most of them, especially in the biomedical domain, follow a framework in which (1) question classification and query formulation, (2) document retrieval, (3) passage retrieval, and (4) answer extraction components play a vital role (Athenikos and Han, 2010; Neves and Leser, 2015; Abacha and Zweigenbaum, 2015).

Question answering in the open-domain is a longstanding challenge widely studied over the last decades (Green et al., 1961; Katz et al., 2002). However, it still remains a real challenge in the biomedical domain. As has been extensively documented in the recent research literature (Athenikos and Han, 2010), open-domain QA is concerned with questions which were not restricted to any domain, while in restricted-domain QA such as the biomedical one, the domain of application provides a context for the QA process. Additionally, Athenikos and Han (2010) report the following characteristics for QA in the biomedical domain: (1) large-sized textual corpora, (2) highly complex domain-specific terminology, and (3) domain specific format and typology of questions.

Since the launch of the biomedical QA track at the BioASQ[1] challenge (Tsatsaronis et al., 2015), various approaches in biomedical QA have been presented. The BioASQ challenge, within 2017 edition, comprised three tasks: (1) task 5a on large-scale online biomedical semantic indexing, (2) task 5b on biomedical semantic QA, and (3)

[1] http://bioasq.org/

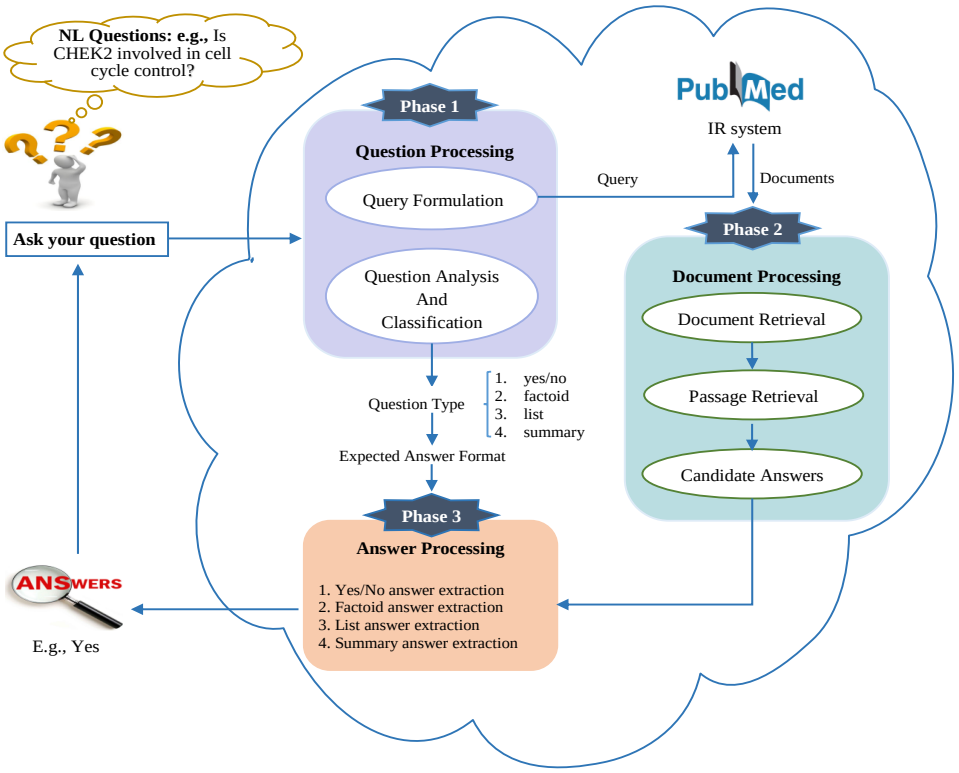


Figure 1: Overall architecture of the proposed biomedical question-answering system

task 5c on funding information extraction from biomedical literature. Task 5b consists of two phases: In phase A, BioASQ released questions in English from benchmark datasets. There were four types of questions: yes/no, factoid, list and summary questions (Balikas et al., 2013). Participants had to respond with relevant concepts, relevant documents, relevant snippets retrieved from the relevant documents, and relevant RDF triples. In phase B, the released questions contained the golden answers for the required elements (documents and snippets) of the first phase. The participants had to answer with exact answers (e.g., biomedical entity, number, list of biomedical entities, yes, no, etc.) as well as with ideal answers (i.e., paragraph-sized summaries) (Krithara et al., 2016). In this paper, we describe our participation in the phase B (i.e., exact and ideal answers) of task5b in the 2017 BioASQ challenge. In our biomedical QA system, we have used (1) a dictionary-based approach to generate the exact answers to yes/no questions, (2) the unified medical language system (UMLS) metathesaurus and term frequency metrics for extracting the exact answers of factoid and list questions, and (3) the BM25 model and UMLS concepts for retrieving the ideal answers. Figure 1 illustrates the generic architecture of our biomedical QA system.

The remainder of the paper is organized as follows. Section 2 introduces related work and discussion about the main biomedical QA approaches with a particular focus on BioASQ participants. Section 3 describes the answer extraction methods used in our biomedical QA system. Section 4 presents the preliminary results we obtained in the 2017 BioASQ challenge. Finally, the conclusion and future work are made in Section 5.

## 2 Related work

Since the launch of the BioASQ challenge (Tsatsaronis et al., 2015), QA in the biomedical domain has received much attention from the research community. The BioASQ challenge, which takes place regularly every year since 2013, is an EU-funded support action to set up a challenge on biomedical semantic indexing and QA. Yenala et al. (2015) have presented IIITH biomedical QA system in BioASQ 2015 based on PubMed search engine, leverage web search results, and domain words. The authors have relied on the PubMed search engine to retrieve relevant documents and then applied their own snippet extraction meth-

ods, which is based on number of common domain words of the top 10 sentences of the retrieved documents and the question. Zhang et al. (2015) have described USTB biomedical QA system in the 2015 BioASQ challenge. They have built a generic retrieval model based on the sequential dependence model, word embedding and ranking model for document retrieval. After splitting the top-ranked documents into sentences, the authors then have applied the same approach for snippets retrieval. Yang et al. (2016) have described the OAQA system in BioASQ 4b based on NLP annotators, machine learning algorithms for search result scoring, collective answer re-ranking, and yes/no answer prediction. Schulze et al. (2016) have presented HPI biomedical QA system based on NLP functionality from a in-memory database (IMDB). The authors have participated in phase A and B of BioASQ 4b. They have used the LexRank algorithm and biomedical entity names for generating ideal answers. Lee et al. (2016) have described KSAnswer biomedical QA system that returns relevant documents and snippets in BioASQ 4b. KSAnswer, which is participated in phase A of task 4b in the 2016 BioASQ challenge, retrieves candidate snippets using a cluster-based language model. Then, it reranks the retrieved top-$N$ snippets using five independent similarity models based on shallow semantic analysis.

## 3 Methods

In this section, we describe the answer extraction module of our biomedical QA system. Although our biomedical QA system is composed of many components (cf. Figure 1) which are included in three main phases, i.e., question processing, document processing, and answer processing, we have only used its answer extraction module since we have participated only in phase B (i.e., exact and ideal answers) of task 5b in BioASQ 2017.

During phase B, BioASQ organizers released the test set of biomedical questions along with their relevant documents, relevant snippets, and questions types, i.e., whether yes/no, factoid, list or summary. For each question, each participating system may return an ideal answer, i.e., a paragraph-sized summary of relevant information. In the case of yes/no, factoid, and list questions, the systems may also return exact answers; for summary questions, no exact answers will be returned. In the following sections (cf. Sections 3.1 and 3.2), we will provide a detailed description of the proposed methods used to extract exact and ideal answers for yes/no, factoid, list and summary questions.

### 3.1 Exact answers

As it has already been described by the BioASQ challenge, the participating systems in phase B of task 5b may return exact answers for yes/no, factoid, and list questions, while no exact answers will be returned for summary questions.

**Yes/No questions:** For each yes/no question, the exact answer of each participating system will have be either "yes" or "no". The decision for the answers "yes" or "no" in our system is obtained by a sentiment analysis-based approach. Indeed, we first have used the Stanford CoreNLP (Manning et al., 2014) for tokenization and part-of-speech tagging one by one the $N$ retrieved snippets $(s_1, s_2, ..., s_n)$ from benchmark datatsets. We then have assigned a sentiment score using the SentiWordNet (Baccianella et al., 2010) lexical resource to each word in the set of retrieved snippets. Finally, the decision for the answers "yes" or "no" is based on the number of positive and negative snippets.

**Factoid questions:**

For each factoid question, each participating system will have to return a list of up to 5 entity names (e.g., up to 5 names of drugs), numbers, or similar short expressions, ordered by decreasing confidence. To answer a factoid question in our biomedical QA system, we have first mapped both the given question and its $N$ relevant snippets retrieved from benchmark datasets to the UMLS metathesaurus in order to extract a set of biomedical entity names. To do so, the MetaMap[2] program was used (Aronson, 2001). We then re-ranked the obtained set of biomedical entity names based on term frequency metrics, i.e., the number of times an entity name appeared in the set of biomedical entity names. Indeed, the biomedical entity names appeared in the question are ignored. We finally kept the 5 top-ranked biomedical entity names as answers. A factoid question has one correct answer, but up to five candidate answers are allowed in BioASQ 2017.

**List questions:** For each list question, each participating system will have to return a single list of entity names, numbers, or similar short expres-

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

sions. The proposed method used to answer list questions in our system is similar to the one described for factoid questions. Indeed, the exact answer is the same of factoid questions, only the interpretation is different for list questions: All $N$ top-ranked biomedical entities are considered part of the same answer for the list question, not as candidates. In this work, we have used the five top-ranked ($N = 5$) entities as answers for list questions.

### 3.2 Ideal answers

To formulate and generate the ideal answers for a given yes/no, factoid, list or summary question, we have used the proposed retrieval model presented in (Sarrouti and Alaoui, 2017b). More specifically, after retrieving the $N$ relevant snippets from benchmark datasets to a given biomedical question, we have re-ranked them based on the BM25 model as retrieval model, stemmed words and UMLS concepts as features. First, we have preprocessed the retrieved set of snippets, including tokenization using the Stanford CoreNLP (Manning et al., 2014), removing stop words[3], and applying Porter' stemmer (Porter, 1980) to extract stemmed words. Additionally, we have used MetaMap to map both questions and snippets to UMLS metathesaurus concepts so as to extract UMLS concepts. Then, we have re-ranked the set of snippets using stemmed words and UMLS concepts as features for the BM25 model. Finally, the ideal answer is obtained by concatenating the two top-ranked snippets.

## 4 Experimental results and discussion

In this section, we present the preliminary results we obtained in BioASQ 2017. We first introduce the evaluation metrics, then give the experimental results, and finally discuss the results.

### 4.1 Evaluation metrics

The evaluation metrics used for the exact answers in phase B of task 5b are accuracy, strict accuracy and lenient accuracy, mean reciprocal rank (MRR), mean precision, mean recall, and mean F-measure. Accuracy, MRR and F-measure are the official measures used for evaluating the exact answers of yes/no, factoid and list questions, respectively. ROUGE-2 and ROUGE-SU4, on the other hand, are the main measures for an automatic evaluation of ideal answers. Details of these evaluations metrics appear in (Balikas et al., 2013).

### 4.2 Results and discussion

Table 1 highlights the preliminary results of our system in phase B (i.e., exact and ideal answers) of BioASQ task 5b. More details on the results can be found in the BioASQ web site[4].

Our system performed well in the challenge ranking as compared with the participating systems. In batch 1, it achieved the third and the fifth position within the 15 participating systems for extracting the exact answers of list and factoid questions respectively. More specifically, our system obtained the second and the third position when considering results by teams, instead of each individual run. In batch 2, considering results by teams, our system obtained the second and the fourth position for extracting the exact answers of list and factoid questions respectively. For yes/no questions, our system achieved the first and the second position respectively in batch 3 and batch 4, while it obtained the fifteenth position in batch 5.

On the other hand, for the ideal answers, our system in terms of ROUGE-2 achieved the fourth position as compared to the 15 and 21 participating systems in batch 1, batch 2 and batch 3 respectively. While in terms of ROUGE-SU4, our system obtained the third position in batch 1 and the fourth position in batch 2. In batch 4 and batch 5, our systems achieved respectively the second and third position within the 27 participating systems in terms of ROUGE-2 and ROUGE-SU4 when considering results by teams, instead of each individual run. This proves that the proposed method could effectively identify the ideal answers to a given biomedical question.

Overall, from the results and analysis on fives batches of testing data of BioASQ task 5b, we can draw a conclusion that our system is very competitive as compared with the participating systems in both exact and ideal answers tasks.

## 5 Conclusion and future work

In this paper, we presented the obtained results for the answer extraction module of our biomedical QA system that participated in task 5b of

[3] http://www.textfixer.com/resources/common-english-words.txt

[4] http://participants-area.bioasq.org/results/5b/phaseB/

| Datasets | Exact answers | | | | | Idial answers | |
|---|---|---|---|---|---|---|---|
| | Yes/No | Factoid | List | | | | |
| | **Accuracy** | **MRR** | **P** | **R** | **F** | ROUGE-2 | ROUGE-SU4 |
| Batch 1 | 0.7647 | 0.2033 (5/15) | 0.1909 | 0.2658 | 0.2129 (3/15) | 0.4943 (4/15) | 0.5108 (3/15) |
| Batch 2 | 0.7778 | 0.0887 (10/21) | 0.2400 | 0.3922 | 0.2920 (6/21) | 0.4579 (4/21) | 0.4583 (4/21) |
| Batch 3 | 0.8387 (1/21) | 0.2212 (9/21) | 0.2000 | 0.4151 | 0.2640 (6/21) | 0.5566 (4/21) | 0.5656 (4/21) |
| Batch 4 | 0.6207 (2/27) | 0.0970 (13/27) | 0.1077 | 0.2013 | 0.1369 (12/27) | 0.5895 (4/27) | 0.5832 (4/27) |
| Batch 5 | 0.4615 (15/25) | 0.2071 (9/25) | 0.2091 | 0.3087 | 0.2438 (11/25) | 0.5772 (7/25) | 0.5756 (7/25) |

Table 1: The primarily results of our system in phase B of BioASQ task 5b. P, R, and F indicate precision, recall, and F-measure, respectively. The values inside parameters indicate our current rank and the total number of submissions for the batch.

the 2017 BioASQ challenge. The proposed approach is based on (1) the SentiWordNet lexical resource to generate the exact answers for yes/questions, (2) UMLS metathesaurus and term frequency metrics for answering factoid and list questions, (3) our retrieval model based on UMLS concepts and the BM25 model for generating the ideal answers. The preliminary results show that our system achieved good performances and is very competitive as compared with the participating systems.

In future research, we intend to present the end-to-end evaluations of our biomedical QA system which includes question classification, document retrieval, passage retrieval, and answer extraction components.

## References


Asma Ben Abacha and Pierre Zweigenbaum. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information Processing & Management* 51(5):570–594. https://doi.org/10.1016/j.ipm.2015.04.006.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, page 17.

Sofia J Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer methods and programs in biomedicine* 99(1):1–24. https://doi.org/10.1016/j.cmpb.2009.10.003.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. volume 10, pages 2200–2204.

Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artiere, and Patrick Gallinari. 2013. Evaluation framework specifications. project deliverable d4.1, 05/2013.

Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball. In *western joint IRE-AIEE-ACM computer conference on - IRE-AIEE-ACM 61 (Western)*. Association for Computing Machinery (ACM). https://doi.org/10.1145/1460690.1460714.

Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimmy Lin, Gregory Marton, Alton Jerome McFarland, and Baris Temelkuran. 2002. Omnibase: Uniform access to heterogeneous data for question answering. In *Natural Language Processing and Information Systems*, Springer Nature, pages 230–234. https://doi.org/10.1007/3-540-36271-1_23.

Anastasia Krithara, Anastasios Nentidis, George Paliouras, and Ioannis Kakadiaris. 2016. Results of the 4th edition of BioASQ challenge. In *Proceedings of the Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics*. pages 1–7.

S. Kropf, P. Krcken, W. Mueller, and K. Denecke. 2017. Structuring legacy pathology reports by openEHR archetypes to enable semantic querying. *Methods of Information in Medicine* 56(2). https://doi.org/10.3414/me16-01-0073.

Hyeon-gu Lee, Minkyoung Kim, Harksoo Kim, Juae Kim, Sunjae Kwon, Jungyun Seo, Jungkyu Choi,

and Yi-reun Kim. 2016. KSAnswer: Question-answering system of kangwon national university and sogang university in the 2016 BioASQ challenge. *ACL 2016* page 45.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://doi.org/10.3115/v1/p14-5010.

Mariana Neves and Ulf Leser. 2015. Question answering for biology. *Methods* 74:36–46. https://doi.org/10.1016/j.ymeth.2014.10.023.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems* 14(3):130–137. https://doi.org/10.1108/eb046814.

Mourad Sarrouti and Said Ouatik El Alaoui. 2017a. A machine learning-based method for question type classification in biomedical question answering. *Methods of Information in Medicine* 56(3). https://doi.org/10.3414/me16-01-0116.

Mourad Sarrouti and Said Ouatik El Alaoui. 2017b. A passage retrieval method based on probabilistic information retrieval and UMLS concepts in biomedical question answering. *Journal of Biomedical Informatics* 68:96–103. https://doi.org/10.1016/j.jbi.2017.03.001.

Mourad Sarrouti and Abdelmonaime Lachkar. 2017. A new and efficient method based on syntactic dependency relations features for ad hoc clinical question classification. *International Journal of Bioinformatics Research and Applications* 13(2):161–177. https://doi.org/10.1504/ijbra.2017.10003490.

Frederik Schulze, Ricarda Schüler, Tim Draeger, Daniel Dummer, Alexander Ernst, Pedro Flemming, Cindy Perscheid, and Mariana Neves. 2016. HPI question answering system in bioasq 2016. In *Proceedings of the Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics*. pages 38–44.

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 Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* 16(1):1–28. https://doi.org/10.1186/s12859-015-0564-6.

Liqin Wang, Guilherme Del Fiol, Bruce E. Bray, and Peter J. Haug. 2017. Generating disease-pertinent treatment vocabularies from MEDLINE citations. *Journal of Biomedical Informatics* 65:46–57. https://doi.org/10.1016/j.jbi.2016.11.004.

Zi Yang, Yue Zhou, and Eric Nyberg. 2016. Learning to answer biomedical questions: OAQA at BioASQ 4b. *ACL 2016* page 23.

Harish Yenala, Avinash Kamineni, Manish Shrivastava, and Manoj Kumar Chinnakotla. 2015. IIITH at BioASQ challange 2015 task 3b: Bio-medical question answering system. In *CLEF 2015*.

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