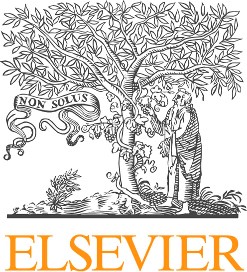
[Artificial Intelligence In Medicine 102 (2020) 101767](https://doi.org/10.1016/j.artmed.2019.101767)



Contents lists available at

ScienceDirec

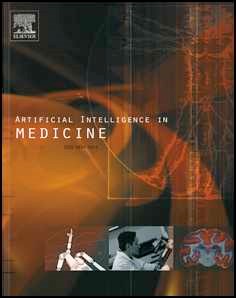
[t](http://www.sciencedirect.com/science/journal/09333657)

Artificial Intelligence In Medicine

journal homepage:

www.elsevier.com/locate/artme

[d](https://www.elsevier.com/locate/artmed)



SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions



[T](http://crossmark.crossref.org/dialog/?doi=10.1016/j.artmed.2019.101767&domain=pdf)

Mourad Sarroutia,\*, Said Ouatik El Alaouib,c

a *Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, U.S. National Institutes of Health, Bethesda, MD* b *National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco* c *Laboratory of Informatics and Modeling, FSDM, Sidi Mohammed Ben Abdellah University, Fez, Morocco*

A R T I C L E I N F O A B S T R A C T

|  |  |
| --- | --- |
| *Keywords:*  Biomedical question answering  Information retrieval  Passage retrieval  Natural language processing  Machine learning  Biomedical informatics  BioASQ | *Background and objective:* Question answering (QA), the identification of short accurate answers to users questions written in natural language expressions, is a longstanding issue widely studied over the last decades in the open-domain. However, it still remains a real challenge in the biomedical domain as the most of the existing systems support a limited amount of question and answer types as well as still require further efforts in order to improve their performance in terms of precision for the supported questions. Here, we present a semantic biomedical QA system named SemBioNLQA which has the ability to handle the kinds of yes/no, factoid, list, and summary natural language questions.  *Methods:* This paper describes the system architecture and an evaluation of the developed end-to-end biomedical QA system named SemBioNLQA, which consists of question classification, document retrieval, passage retrieval and answer extraction modules. It takes natural language questions as input, and outputs both short precise answers and summaries as results. The SemBioNLQA system, dealing with four types of questions, is based on (1) handcrafted lexico-syntactic patterns and a machine learning algorithm for question classification, (2) PubMed search engine and UMLS similarity for document retrieval, (3) the BM25 model, stemmed words and UMLS concepts for passage retrieval, and (4) UMLS metathesaurus, BioPortal synonyms, sentiment analysis and term frequency metric for answer extraction.  *Results and conclusion:* Compared with the current state-of-the-art biomedical QA systems, SemBioNLQA, a fully automated system, has the potential to deal with a large amount of question and answer types. SemBioNLQA retrieves quickly users’ information needs by returning exact answers (e.g., “yes”, “no”, a biomedical entity name, etc.) and ideal answers (i.e., paragraph-sized summaries of relevant information) for yes/no, factoid and list questions, whereas it provides only the ideal answers for summary questions. Moreover, experimental evaluations performed on biomedical questions and answers provided by the BioASQ challenge especially in 2015, 2016 and 2017 (as part of our participation), show that SemBioNLQA achieves good performances compared with the most current state-of-the-art systems and allows a practical and competitive alternative to help information seekers find exact and ideal answers to their biomedical questions. The SemBioNLQA source code is publicly available at [https://github.com/sarrouti/sembionlqa.](https://github.com/sarrouti/sembionlqa) |

# Introduction

Retrieving high-quality short answers to a given natural language question from the ever-increasing volume of biomedical literature is the key to creating high-quality systematic reviews that support the practice of evidence-based medicine [1–3] and improve the quality of patient care [4,5]. However, the explosive rise in the amount of the scientific literature in the biomedical domain has made it difficult to absorb all relevant information even for experts in their field of interest.

|  |  |
| --- | --- |
| |  | | --- | | ⁎ Corresponding author. Tel: +1 (301) 518-1033.  *E-mail address:* sarrouti.mourad@gmail.com (M. Sarrouti).  <https://doi.org/10.1016/j.artmed.2019.101767> |   Received 4 May 2018; Received in revised form 19 November 2019; Accepted 19 November 2019  0933-3657/ Published by Elsevier B.V. |

Therefore, to provide information seekers with relevant answers, more sophisticated technologies are needed [6,7] and automatic biomedical text mining methods are increasingly investigated [8,2,9,6,10,11]. Recent reports, such as the one by Hristovski et al. [12], have highlighted that the most widely used are specialized information retrieval (IR) systems such as PubMed. However, in the present settings, the set of the retrieved documents, potentially relevant, represents an answer size that is still too large to identify the precise information readily. Moreover, in classical IR systems, the users have often to deal with the burden of studying and filtering the returned citations of their queries so as to find the precise information they were looking for. In this context, an evaluation study presented by Ely and colleagues [13] showed that physicians spent on average less than two minutes looking for information to answer clinical questions, although many of their questions remained unanswered. In another study by Hersh et al. [14], at least 30 minutes are needed on average for medical and nurse practitioner students to answer clinical questions using MEDLINE.

Unlike IR, question answering (QA) systems have the potential to overcome the aforementioned shortcomings. They aims at directly producing and providing precise answers rather than entire documents to users questions, by automatically analyzing thousands of articles, ideally in less than a few seconds. Such systems, which can help users locate useful information quickly, need linguistically and semantically processing of both users questions and data sources in order to extract the relevant information. For instance, for the biomedical question “What is the name for anorexia in gymnasts?”, a QA system would provide as an answer particular name, i.e., Anorexia Athletica, for anorexia in gymnasts. A typical QA system can be viewed as a pipeline composed of three main processing phases [15–18] namely question processing, document processing, and answer processing each of which has to deal with specific challenges and issues.

While QA in the open-domain is a longstanding challenge widely studied over the last decades [19–22], it still requires further efforts in the biomedical domain so as to enable a user to have a better chance to find an accurate answer to his human natural language question. Open domain systems answer questions from diverse fields and domain-specific systems -focus on a specialized area- use specific linguistic resources that enable more precise answers to be given. Athenikos and Han [16] have reported the following characteristics for biomedicaldomain QA: (1) large-sized textual corpora, (2) highly complex domain specific terminology, and (3) domain-specific format and typology of questions. On the other hand, Bauer et al. [23] have shown that the amount of the supported question and answer types are the most important challenges and features for QA systems comparison.

Although the importance of answering questions in the biomedical domain, until now there are only few integral systems such as the ones described in [15,24–26,18,27] that can retrieve answers to biomedical questions written in natural language. While these systems have proven to be quite successful at answering biomedical questions, they provide a limited amount of question and answer types, for instance, some of them [15,24,26] handle only with definition questions or returns solely short summaries as answers for all types of questions, and the most of the other ones do not deal with yes/no questions which are one of the most complicated question types to answer as they are seeking for a clear “yes” or “no” answer. Furthermore, these systems still requires further efforts in order to improve their performance in terms of precision to currently supported question and answer types.

On the other hand, since the launch of the BioASQ[[1]](#footnote-1) challenge, theories and methods in biomedical QA continue to evolve to better meet the needs of users questions [28–32], thanks to the many editions of the BioASQ challenge. However, these systems suffer restrictions. For instance, in such systems the users have to give or select manually the question type to each given question. Furthermore, for a given biomedical question, these systems use the documents and passages released by the BioASQ challenge to produce the answer. A biomedical QA system designed to aid users in automatically finding the relevant answers in the biomedical domain, therefore, should be able to determine automatically the relevant documents, relevant passages and the types of questions and hence to answer the different types of questions.

In this paper, we present a fully automated system SemBioNLQA Semantic Biomedical Natural Language Question Answering - which has the ability to handle the kinds of yes/no questions, factoid

questions, list questions and summary questions that are commonly asked in the biomedical domain.

* Yes/no questions: questions that require either “yes” or “no” answer.
* Factoid questions: questions that require a particular entity name (e.g., of a disease, drug, or gene), a number, or a similar short expression as an answer.
* List questions: questions that expect a list of entity names (e.g., a list of gene names, a list of drug names), numbers, or similar short expressions as an answer.
* Summary questions: questions that expect short summaries as answer.

The aim of this study is to describe the process underlying the design and development of SemBioNLQA. SemBioNLQA is derived from our previously established methods in (1) question classification [33], (2) document retrieval [34], (3) passage retrieval [35], and (4) answer extraction system that was one of the winners2 in the 2017 edition of the BioASQ challenge [36]. We developed the SemBioNLQA system based on the integration of these methods and techniques.

SemBioNLQA first takes as its input a natural language biomedical question and includes preprocessing of the question, identification of the question type and the expected answer format based on handcrafted lexico-syntactic patterns and support vector machine (SVM), as well as building a query using UMLS entities. Then, based on PubMed and UMLS similarity, it retrieves documents satisfying the query from the MEDLINE database. After that, it extracts top-ranked passages from topranked documents based on the BM25 model, stemmed words and UMLS concepts. Finally, it generates and returns both “exact” (depending on the expected answer for each question type) and paragraphsized “ideal” answers from these passages based on the UMLS metathesaurus, BioPortal synonyms, SENTIWORDNET and term frequency metric. Compared with the previous works, our approach has the following contributions:

1. To the best of our knowledge, no one of the previous systems are able to automatically provide the specific exact answers for each question type.
2. SemBioNLQA, an end-to-end biomedical QA system, is aimed to be able to accept a variety of natural language questions and to generate appropriate natural language answers by providing both exact and ideal answers.
3. SemBioNLQA provides exact answers “yes” or “no” for yes/no questions, biomedical named entities for factoid questions, and a list of biomedical named entities for list questions. In addition to exact answers for yes/no, factoid and list questions, SemBioNLQA also returns ideal answers, while it retrieves only ideal answers for summary questions.
4. Finally, extensive experiments on benchmark datasets provided by the BioASQ challenges, show that SemBioNLQA is more effective and competitive as compared with the current state-of-the-art systems.

The remainder of the paper is organized as follows. First, related work concerning the main QA systems with a particular focus on the biomedical domain is reviewed in Section 2. Then, details on the overall architecture of the SemBioNLQA system are presented in Section 3. Several comprehensive experiments are performed to evaluate the effectiveness of the developed system in Section 4 where experimental settings, evaluation metrics, benchmark datasets, results, and discussion are presented. Conclusion and future work are finally presented in Section 5.

# Related work

Question answering has received much attention from the research community in recent years. Open-domain QA research area benefited from the Text Retrieval Conference (TREC) [37], which takes place regularly every year since 1999. Otherwise, biomedical QA has been a challenge from the past few years. There has been no significant progress toward this, because of the complexity of natural language in the biomedical domain. In this section, we will discuss the main integral QA systems with a particular focus on the biomedical domain.

In this context, Lee et al. [15] have developed a medical QA system named MedQA which is composed of five components including (1) question classification, (2) query generation, (3) document retrieval, (4) answer extraction, and (5) text summarization. In MedQA, at first, the question classification component automatically classifies medical questions into categories of the taxonomy created by Ely et al. [13] based on supervised machine-learning approaches. Next, the document retrieval component uses the query terms to retrieve relevant documents from either the Web documents using Google or the locally-indexed MEDLINE corpora using Lucene. Then, the answer extraction component identifies from the retrieved documents relevant sentences that answer the questions based on lexico-syntactic patterns. Finally, the text summarization component removes the redundant sentences and condenses the sentences into a coherent summary which is considered as answer. Although the MedQA system returns short summaries that could potentially answer medical questions, current MedQA's capacity is limited: it only provides answers to definitional questions.

Cruchet et al. [24] have built a biomedical QA system called HONQA which extracts sentences from Health On the Net Foundation (HON) certified websites and provides them as answers for biomedical questions. HONQA is based on a learning approach for identifying the question type and semantic resources such as UMLS to guide the system, particularly in the choice of answers, but no details are presented in the publication. In its current form, it is not able to provide exact answers to other question types, for instance, yes/no and factoid questions [23].

Gobeill et al. [25] have developed a biomedical QA system, EAGLi, which aims at extracting answers to biomedical questions from MEDLINE documents. Given a natural language question, EAGLi first analyzes the question in order to find the question type and to build the query based on a set of patterns. Then, it retrieves a set of relevant documents from MEDLINE using either PubMed or EasyIR. Finally, the system extracts and computes a score for each of concepts expressed in the most relevant documents, and finally outputs a ranked list of candidate answers. The current EAGLi's capacity is limited to Wh-type questions since it only covers the definitional and factoid questions.

Cao et al. [26] have presented a clinical QA system named AskHERMES that returns short summaries as answers of ad-hoc clinical questions. AskHERMES was developed through the main following steps: (1) question analysis, (2) document retrieval, (3) passage retrieval, and (4) summarization and answer presentation. In the question analysis step, the authors have first classified clinical questions into general topics (e.g., device, diagnosis, etc.) based on SVM classifier and then identified keywords that capture the most important content of the question using conditional random fields. In the document retrieval step, the BM25 model was used to retrieve relevant documents. After that, they have extracted candidate passages based on dynamically generates passage boundaries and scored them based on both wordlevel and word sequence-level similarity in the passage retrieval step. Finally, the answer was generated based on structural clustering using content-bearing terms. The AskHERMES system returns passages (short texts) that could potentially answer all types of clinical questions. However, it returns a large number of results, which tends to defeat the intent of a QA system in reducing the amount of information that must be read. Moreover, the system supports only a single answer type in form of multiple sentence passages for all questions types [23].

Abacha and Zweigenbaum [18] have presented a semantic medical QA system called MEANS. MEANS is composed of three main phases: (1) corpora annotation, (2) question analysis and classification, and (3) answer search. The authors have applied NLP methods, named entity recognition and relation extraction so as to build RDF annotations of the source documents and SPARQL queries representing the users questions. They further have defined the MESA ontology to represent the concepts and relations between them in order to construct SPARQL translations of natural language questions. To extract answers, the SPARQL queries were executed in order to interrogate RDF triples constructed in the corpus-annotation step. Even so, the authors have dealt with four questions types, they have focused on factoid and yes/ no questions since more specific processes are still required to deal with complex questions (e.g. why, when).

Hristovski et al. [12] have introduced a biomedical QA system, SemBT, based on semantic relations extracted from the biomedical literature. SemBT consists of three main processing steps: (1) preprocessing, (2) question processing, and (3) answer processing. During the preprocessing step, the authors first have extracted semantic relations using the SemRep natural language processing system from sentences retrieved from MEDLINE citations, and then stored them in a database. In the question processing step, the authors have constructed a query for searching in the database of the extracted semantic relations. Finally, in the answer processing phase, they have presented the resulting semantic relations as answers in a top-down fashion, first semantic relations with aggregated occurrence frequency, then particular sentences from which the semantic relations are extracted. The SemBT system returns answers in the form of semantic relations and particular sentences from which the semantic relations are extracted. However, in its current implementation, the questions must be in the form SubjectRelation-Object, and hence, it does not allow asking questions in natural language format, for example, the natural language question “What drugs can be used to treat diabetes?” can be asked in SemBT as “phsu treats diabetes” where “phsu” stands for “pharmacological substance” and “treats” is the name of the semantic relation (this example comes from [12]).

Kraus et al. [38] have developed the Olelo system for intuitive exploration of biomedical literature. The Olelo system consists of three main modules: (1) question processing, (2) document/passage retrieval, and (3) answer processing. In Olelo, the question processing module is based on a system described in [32]. In the second module, Olelo first uses the tokens and the matched terms to formulate a query to the database so as to retrieve abstracts of the retrieved documents and then ranks the obtained abstracts according to the occurrence and importance of the searched tokens. Finally, an answer is returned to the user depending on the type of the question by the third module. Although Olelo has proven to be quite successful at answering biomedical questions, currently, Olelo supports only three question types including factoid, list and summary questions. Indeed, it does not support yes/no questions which are one of the most complicated questions to answer as they are seeking for a clear “yes” or “no” answer. Moreover, according to the evaluation results of Olelo [27] in the 2017 BioASQ competition in which we also have participated [36], the system still requires further efforts in order to improve its performance in terms of precision to currently supported question types.

Since the launch of the biomedical QA track at the BioASQ challenge, various approaches have addressed QA in the biomedical domain. Based on learning-based approaches, Yang et al. [28] have described the development of a biomedical QA system that returns only the exact answers for factoid and list questions. They have designed two supervised algorithms that are shared by both question types: (1) answer type prediction and (2) candidate answer scoring. They have used several features (e.g., lemma, the semantic type of each concept, etc.) and logistic regression as a multi-class classifier to identify the semantic answer type of a given question. On the other hand, the authors have considered concepts as candidate answers and scored them based on multiplication method so as to combine the type coercion score and the occurrence count as well as logistic regression to learn the scoring function. Zhang et al. [29] have described a biomedical QA system that retrieves solely the exact answers for factoid and list questions. The system first used PubTator [39] for generating the candidate answers and then ranked them using term frequency metrics. Choi [30], on the other hand, has developed a biomedical QA system based on keyword terms that retrieves only the ideal answers for biomedical questions. In contrast, Neves [31] has proposed a biomedical QA system based on inmemory database (IMDB) and its built-in text analysis features to formulate both the exact and the ideal answers for biomedical questions. It consists of three main processing phases: (1) question processing, (2) document and passage processing, and (3) answer processing. They first have constructed the query to a given question in the question processing phase. They then have retrieved up to 200 top ranked documents, and fetching the title and abstract for each PMID using PubMed search engine and the BioASQ web services, respectively. They next have retrieved candidate passages based on the built-in information retrieval features available in the IMDB, which is based in approximated string similarity to match terms from the query to the words in the documents and ranked them using TFIDF metrics. They have dealt with four types of questions, i.e., yes/no questions, factoid questions, list questions, and summary questions. Indeed, the decision on either the answers “yes” or “no” was based on the sentiment analysis predictions provided by the IMDB. They have extracted factoid and list answers based also on built-in predictions provided by IMDB, more specifically, on the annotations of noun phrases and topics. For summary questions, they have built summaries for the ideal answers based on the phrases which contain sentiments. Schulze et al. [32] have presented an approach to biomedical QA which dealt solely with the ideal answers. The latter is based on the LexRank algorithm [40], but that solely used the named entities for the similarity function.

While the development of QA systems represents progress since the launch of the BioASQ challenge, the systems nevertheless suffer restrictions. Many were limited to very specific contexts and cover a limited type of questions and answers. Moreover, most of these systems are not end-to-end QA systems.

In this work, our goal is to go beyond the previous biomedical QA systems and develop a QA system with the ability to automatically handle with a large amount of question types including yes/no questions, factoid questions, list questions and summary questions that are commonly asked in the biomedical domain [41]. Furthermore, compared to the end-to-end QA systems such as Olelo, HONQA, AskHERMES, EAGLi, MEANS and MedQA, not only our system is able to accept a variety of questions types, but also it is capable to provide the specific answers to each question type. Table 1 summarizes the dimensions and comparisons of the aforementioned end-to-end QA systems and SemBioNLQA.

# Methods

In this paper, we propose a biomedical QA system based on a

**Table 1**

semantic approach using PubMed documents. The system has the ability to deal with four types of biomedical questions: yes/no questions, factoid questions, list questions, and summary questions [41]. Fig. 1 shows the architecture of SemBioNLQA and its main components which are: (1) question classification and query reformulation, (2) document retrieval, (3) passage retrieval, and (4) answer extraction.

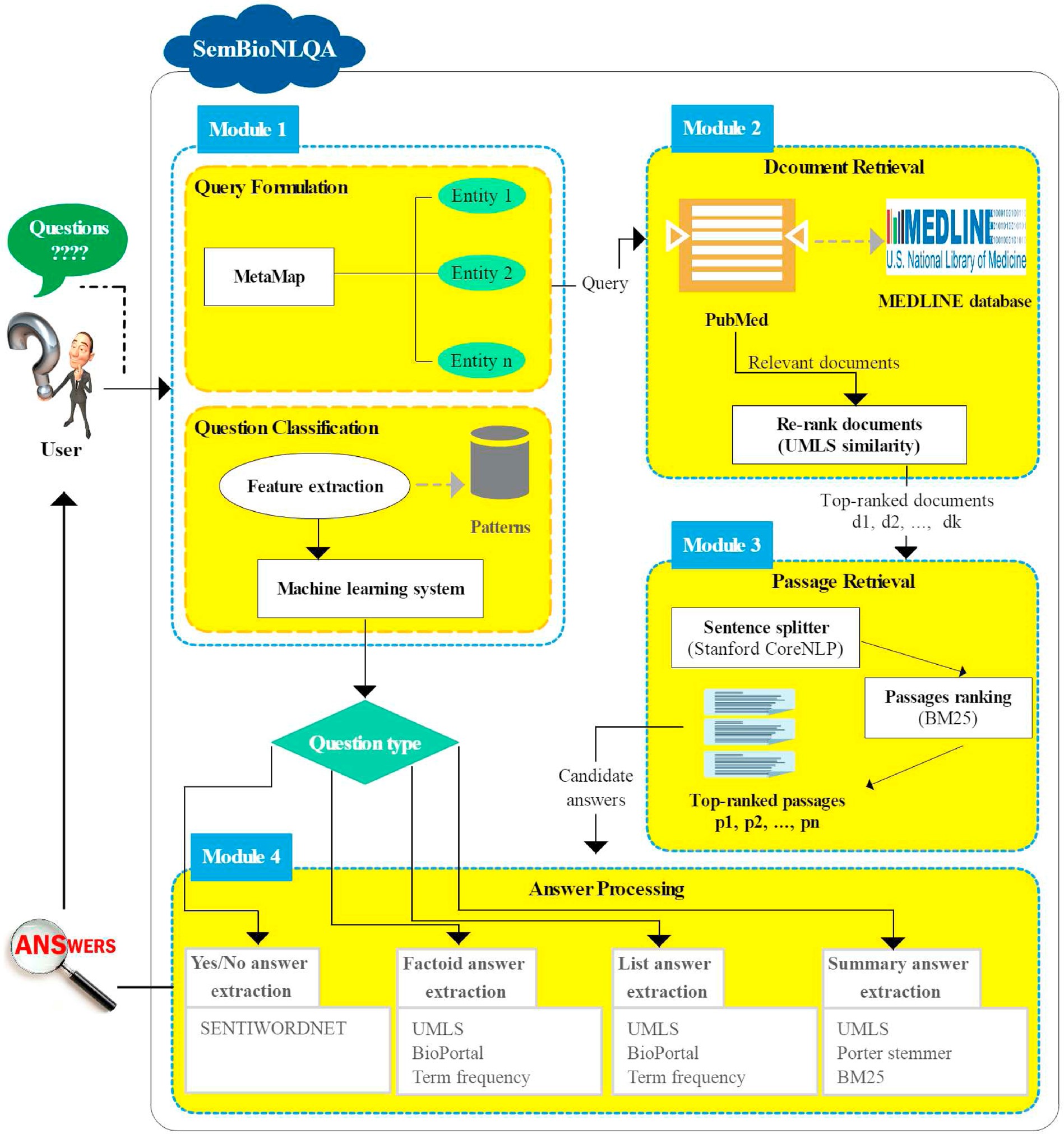
The first component receives the input entered by the user, i.e., a natural language question (e.g., what, why, where, etc.), and includes preprocessing of the question, identification of the question type and the expected answer format to be required, as well as building a query which is an input to document retrieval, the second component. A document retrieval system is used to retrieve documents satisfying the query. After that, top-ranked passages are extracted from top-ranked documents by the passage retrieval, the third component. The output of this component is a set of top-ranked passages which is used as a set of candidate answers and as input to the last component, answer extraction. In this step, the candidate answers are matched against the expected answer type generated by the first component and ranked by how well they satisfy the user question using an appropriate answer extraction algorithm. Finally, the top-ranked candidate answers and the raw texts from which the answers were extracted are shown to the user. In the following sections (cf. Sections 3.1, 3.2, 3.3, and 3.4), we will provide a detailed description of each component.

## Question classification and query reformulation

For a given biomedical question, the question classification task in SemBioNLQA aims at identifying the question type and therefore determine the expected answer format, to see whether the answer should be a biomedical entity name, a short summary, “yes ” or “no”, etc. The expected answer format allows using of type-specific answer retrieval algorithms and rejecting possible answers of the wrong type [42–46]. Accordingly, we have classified biomedical questions into one of the following four categories defined by the BioASQ challenge [41]: (1) yes/no, (2) factoid, (3) list, and (4) summary. The proposed question type classification method in SemBioNLQA, presented and evaluated in [33], uses our predefined set of handcrafted lexico-syntactic patterns (e.g., “[What] + [VBZ] + [\*] + [definition]+ [\*] +?” where “+”, VBZ and “\*” indicate the concatenation, any terms and verb 3rd person singular present, respectively) to extract learning features for machineslearning algorithms. We have used the 810 BioASQ training questions for training the classifier. The classification process consists of two steps. For a given question, it is first represented by a vector of terms. These terms are extracted using our set of handcrafted lexico-syntactic patterns. Then, a machine learning algorithm classifies the given question into one of the four aforementioned categories. Since a biomedical question can be assigned to one of four classes, a multi-class classification has been used. Moreover, we explored several learning algorithms such as naive bayes, decision tree and support vector machines (SVM), showing 4-class SVM with a linear kernel achieved the best result. The results have shown that linear kernel outperforms the other kernels such as RBF kernel, tree kernel, and composite kernel. The

|  |  |  |  |
| --- | --- | --- | --- |
| QA systems | Question format | Target question types | Answer types |
| MedQA [15] | Natural language | Definition | Summaries |
| HONQA [24] | Natural language | Definition, factoid | Sentence |
| EAGLi [25] | Natural language | Definition, factoid | Multi-phrase passages and a list of single Sentities |
| AskHERMES [26] | Natural language | - | Multiple sentence passages |
| MEANS [18] | Natural language | Definition, yes/no, factoid | Sentence, named entity, “yes” or “no” |
| SemBT [12] | Subject-Relation-Object | - | Semantic relations, sentences |
| Olelo [38] | Natural language | factoid, list, summary | MeSH term, list of Mesh terms, short summaries |
| SemBioNLQA | Natural language | Yes/no, factoid, list, summary | “Yes” or “no”, UMLS entity, list of UMLS entity, short summaries |

Question answering systems comparison matrix of features between the aforementioned systems and SemBioNLQA. The “-” indicates that the system did not include target question types classification



**Fig. 1.** Overall architecture of the proposed biomedical question answering system SemBioNLQA.

best value of C parameter is 1.01 which was fixed after 5-fold crossvalidation. Our evaluation on five batches of BioASQ annotated questions has shown that using the multi-class SVM classifier and our set of handcrafted lexico-syntactic patterns lead to the highest accuracy of 89.40%.

On the other hand, to build the query for document retrieval, we first have used MetaMap [47] for mapping terms of the question to unified medical language system (UMLS) metathesaurus in order to extract biomedical entity names. We then concatenated the extracted entities with the “+” operator. The UMLS, which consists of three knowledge resources: (1) the metathesaurus, (2) the semantic network, and (3) the SPECIALIST lexicon, is a repository of biomedical concepts developed by the U.S. National Library of Medicine (NLM). It integrates 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 concepts [48].

## Document retrieval

Document retrieval, which consists in retrieving the set of relevant documents that are likely to contain the answer to a given query constructed by the previous module [49], is one of the significant and important components of the proposed QA system. The correct answers can be only found when the set of retrieved documents is determined correctly. In this work, the approach for retrieving relevant PubMed documents is approximately similar to the one described in our recent document retrieval system, presented and evaluated in [34]. In this task, we first have sent the constructed query to PubMed search engine by calling E-utilities[[2]](#footnote-2) Web service from PubMed to retrieve relevant documents from the MEDLINE database. This database, maintained by the U.S. NLM, is considered as the authoritative source of medical evidence for medical professionals, biomedical researchers, and many other users [50]. We then have re-ranked the returned documents using the proposed algorithm [34] which is based on the UMLS similarity4 [51]. The idea is to compute the sum of the semantic similarity scores between biomedical concepts of a given question and each title of the returned documents. Finally, after re-ranking the returned documents, the *D* top-ranked ones are kept and used as the input data of the following component, passage retrieval. To compute the semantic similarity between the titles of the returned documents and the question, we first have mapped both the titles and the question to UMLS Metathesaurus concepts in order to extract UMLS concepts they refer to. We then have computed the semantic similarity between pair concepts using the UMLS similarity. Path length has been used as similarity measure where the similarity score is inversely proportional to the number of nodes along the shortest path between the concepts in Mesh5 ontology. Fig. 2 shows an example of simantic similarity scores between the biomedical question “Is Tuberous Sclerosis a genetic disease?” and the title “Tuberous sclerosis complex diagnosed from oral lesions.” of PubMed document (PMID=24310804).

## Passage retrieval

Once the *D* top-ranked documents (*d*1, *d*2, .. ., *dd*) are retrieved for a given question, the passage retrieval module retrieves and fixes the *P* top-ranked passages (pieces of texts) from the retrieved documents. These top-ranked passages are served as answer candidates and the SemBioNLQA system retrieves the answer from them. Several studies, such as the one reported in [52], have highlighted that a correct answer to a given biomedical question can be found with high probability when it already exists in one of the retrieved passages. Accordingly, we have developed a novel approach to extract relevant passages from PubMed documents to a given biomedical question. In the proposed passage retrieval approach, presented and evaluated in [35], we first have forwarded abstracts of the *D* top-ranked documents to Stanford CoreNLP [53] sentence splitter so as to segment them into sentences. Specifically, we define a passage in SemBioNLQA as one sentence. We then have preprocessed the obtained set of sentences (i.e., set of candidate passages) including tokenization, removing stop words,6 and applying Porter’ stemmer [54] to extract stemmed words. Additionally, we have used the MetaMap program for mapping both biomedical questions and candidate passages to UMLS concepts in order to extract biomedical concepts. Moreover, the MetaMap word sense disambiguation system [55] has been fixed to resolve ambiguities in the texts by identifying the meaning of ambiguous terms. Using stemmed words and UMLS concepts as features, we finally ranked the set of candidate passages and kept the *P* top-ranked ones using BM25 as retrieval model. The output of this component is a set of candidate answers, i.e., the *P* top-ranked passages (*p*1, *p*2, .. ., *pp*), which is an input to answer extraction, the last step. Details of this work appear in [35].

## Answer extraction

|  |  |
| --- | --- |
| 1. <http://maraca.d.umn.edu/cgi-bin/umls_similarity.cgi> 2. <https://www.nlm.nih.gov/mesh/> 3. <http://www.textfixer.com/resources/common-english-words.txt> | 1. <http://participants-area.bioasq.org/results/5b/phaseB/> 2. <http://www.bioasq.org/participate/fifth-challenge-winners> 3. <http://sentiwordnet.isti.cnr.it/> |

Answer extraction is the most challenging task of SemBioNLQA as this is when the precise answer has to be extracted from the candidates answers retrieved and selected by the passage retrieval component. The output from this module is a short accurate answer to the user's question. Here, the appropriate answer extraction method, which extracts the final answer from the candidate answers, is selected according to the question type that was automatically detected by the question classification module. Specifically, after retrieving the *P* candidate answers and identifying the question category to a given biomedical

question, the SemBioNLQA system applies the appropriate answer extraction method to extract final answers. As SemBioNLQA deals with four types of questions, i.e., yes/no questions, factoid questions, list questions, and summary questions, therefore, we developed novel answer presentation methods for each type of question. These methods are derived from our previously established answer extraction system [36] that performed well in the 2017 BioASQ7 challenge ranking [56,57] within the top tier submissions out of all teams, and was one of the fifth challenge winners8 . The proposed system provides exact answers (e.g., “yes”, “no”, a biomedical entity, etc.) and ideal answers (summaries) for yes/no, factoid and list questions, while it retrieves only the ideal answers for summary questions. In the following sections (cf. Sections 3.4.1, 3.4.2, 3.4.3, and 3.4.4), we will provide a detailed description of each answer extraction method.

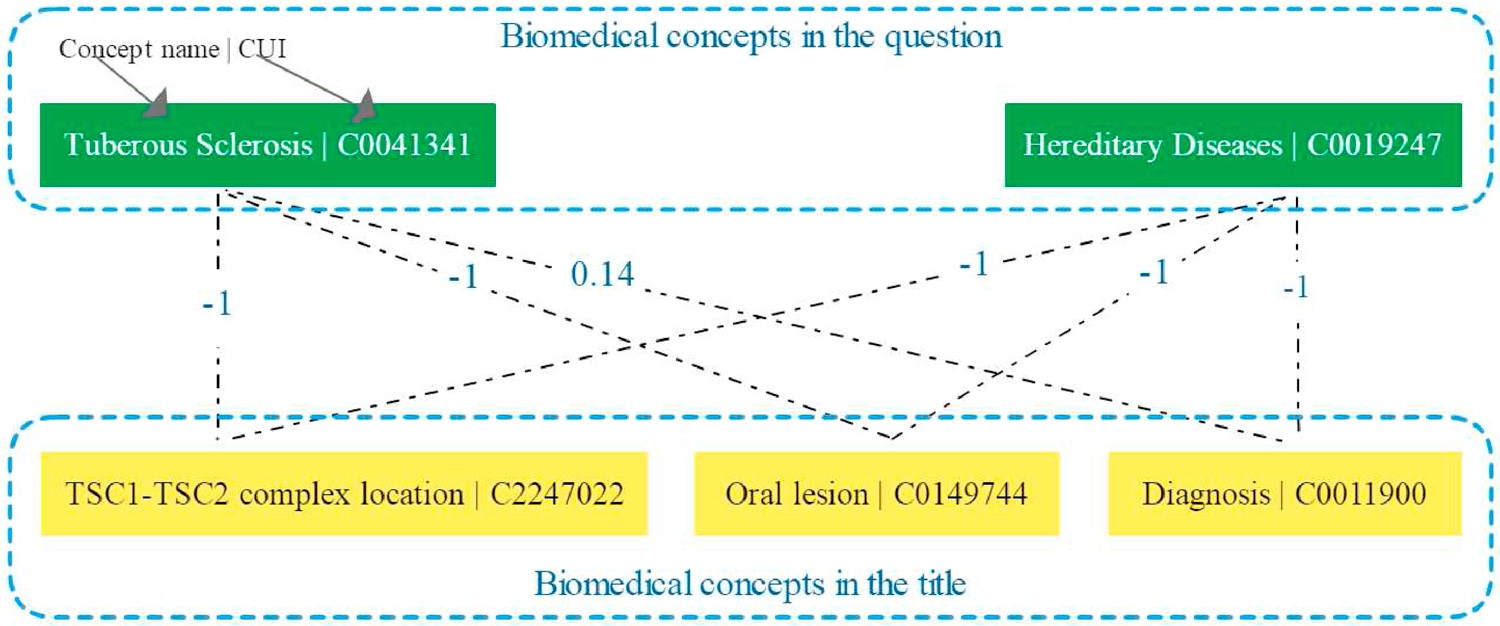
### Yes/no questions

After retrieving the *P* candidate answers to a given yes/no question, the decision on either the answer “yes” or “no” is obtained using our proposed yes/no answer extraction method, presented and evaluated in [58]. The latter is based on SENTIWORDNET9 : a lexical resource for sentiment analysis and opinion mining [59]. In this task, at first, the Stanford CoreNLP tools [53] have been used for tokenization and partof-speech tagging one by one the *P* retrieved candidate answers. Then, each word of the candidate answers is assigned its SENTIWORDNET score. Finally, the decision to output “yes” or “no” depends on the number of positive or negative candidate answers: “yes” for a positive final sentiment candidate answers score and “no” for a negative one. Algorithm 1 further illustrates how the proposed yes/no answer extraction method works.

**Algorithm 1** Biomedical yes/no answer generator [58]

|  |  |  |
| --- | --- | --- |
| 1: | **Input**: *yes/no question Q and set of candidate answers P* |  |
| 2: | **Output**: *answer: “yes” or “no”* |  |
| 3: | *postive* ⟵0;*negative* ⟵0 | ▷ number of positive candidate answers and negative ones |
| 4: | *i* ⟵1 |  |
| 5: | **function** PreProcessing(*p*:*answer condidate*) |  |
| 6: | *TOKEN*[1...*m*] ⟵ *TOKENIZATIONANDPOSTAGGING*(p) |  |
| 7: | **return** *TOKEN* |  |
| 8: | **end function** |  |
| 9: | **do** |  |
| 10: | *W*[1. .. *m*] ⟵ *PREPROCESSINGP*[*i*] | ▷ get a set of words and their POS tags of a candidate answer |
| 11: | *score* ⟵0.0;*j* ⟵1 |  |
| 12: | **do** |  |
| 13: | *score* ⟵ *score* + *SENTIWORDNET*W[j] |  |
| 14: | *j* ⟵ *j* + 1 |  |
| 15: | **while** *j* ≤ *m* | ▷ m is the size of the set of words W |
| 16: | **if** *score* ≥ 0 **then** |  |
| 17: | *positive* ⟵ *positive* + 1 |  |
| 18: | **else** |  |
| 19: | *negative* ⟵ *negative* + 1 |  |
| 20: | **end if** |  |
| 21: | *i* ⟵ *i* + 1 |  |
| 22: | **while** *i* ≤ *np* | ▷ np is number of candidate answers |
| 23: | **if** *positive* ≥ *negative* **then** | ▷ the decision for the answers “yes” or “no” is based on the number positive and negative candidate answers |
| 24: | *output* ⟵“*yes*″ |  |

|  |
| --- |
| **Fig. 2.** An example of semantic similarity scores between biomedical concepts extracted using the MetaMap tool. CUI indicates UMLS concept unique identifier. -1 |

means that there is not a semantic relationship between pair UMLS concepts.

25: **else**

26: *output* ⟵“*no*″

27: **end if**

### Factoid questions

With a view to answering factoid questions in SemBioNLQA, we have proposed a factoid answer extraction method based on UMLS metathesaurus, BioPortal synonyms and term frequency metrics. To do so, the *P* candidate answers are first mapped to the UMLS metathesaurus (2016AA knowledge source) using the MetaMap[[3]](#footnote-3) program [47] so as to extract the set of biomedical entity names *Es*. Next, the obtained set of biomedical entity names are ranked based on term frequency metrics *TF*(*ei*, *Es*), the number of times entity name *ei* appeared in the set of biomedical entity names *Es*. We have explored several term weighting methods such as TFIDF and BM25, showing term frequency achieved the best result for this task. We speculate that the answers are located in the first and second candidate answers. Then, synonyms for each of the *T* top-ranked entity names are retrieved using Web services from BioPortal[[4]](#footnote-4) . Finally, the *T* top-ranked biomedical entity names and their *T* top synonyms are displayed as answers, excluding entities also mentioned in the question. The idea behind excluding entities mentioned in the question is that after analysing the training set of questions and answers released by the BioASQ organizers, we found that the most entities that appear in questions are not part of the answers. For example, the answer of the factoid question “What is the name of Bruton's tyrosine kinase inhibitor that can be used for treatment of chronic lymphocytic leukemia?” which contains several entities (e.g., “Chronic Lymphocytic Leukemia”), is “Ibrutinib” which is not part of the question entities. As described by the BioASQ challenge, a factoid question has one correct answer, but up to five candidate answers and their synonyms are allowed. Fig. 3 shows an example of the whole process, i.e, mapping to UMLS metathesaurus, and synonyms extraction for a candidate answer of the factoid question “Which type of lung cancer is afatinib used for?”.

### List questions

The proposed method used to answer list questions in SemBioNLQA is similar to the one described for factoid questions. As it is shown by the BioASQ challenge, the main difference between factoid and list questions is that the former require a single list of answers while the latter expect a list of lists of entity names, numbers, or similar short expressions. Each entity may be accompanied by a list of synonyms. Therefore, a list of lists should be provided for each question by the proposed biomedical QA system. In other words, the exact answer is the same of factoid questions, but the interpretation is different for list questions: All *T* top-ranked entities are considered part of the same answer for the list question, not as candidates.

### Summary questions

As summary questions do not have precise answers, therefore, they are simply answered in SemBioNLQA by formulating short summaries (i.e., ideal answers) of relevant information. In this work, the ideal answers of questions are formed by concatenating the two top-ranked passages which were retrieved by the proposed passage retrieval approach based on Stanford CoreNLP sentence length as passage length, stemmed words and UMLS concepts as features for the BM25 model. We first have forwarded abstracts of the the *N* relevant documents of a given biomedical question to Stanford CoreNLP sentence splitter so as to segment them into sentences. We then have preprocessed the obtained set of candidate answer passages including tokenization, removing stop words, and applying Porter’ stemmer to extract stemmed words. Additionally, we have used the MetaMap program for mapping both biomedical questions and candidate passages to UMLS concepts in order to extract biomedical concepts. Moreover, the MetaMap word sense disambiguation system has been used to resolve ambiguities in the texts by identifying the meaning of ambiguous terms. Using stemmed words and UMLS concepts as features, we finally ranked the set of candidate answers passages using BM25 as retrieval model, and concatenated the two top-ranked candidate answer passages.

In particular, in addition to the exact answers returned for previous questions (i.e., yes/no, factoid and list), the system also provides ideal answers. Therefore, SemBioNLQA provides both exact answers and the ideal answers for yes/no, factoid and list questions, whereas it provides only ideal answers for summary questions.

# Experimental results and discussion

In this section, we first report a systematic evaluation of the SemBioNLQA system on benchmark datasets provided by the BioASQ challenge in order to compare with the Olelo system [38], which is fortunately evaluated at the 2017 BioASQ challenge [27] in which we also participated [36]. We then present a manual evaluation using BioASQ training questions and answers to compare with the AskHERMES [26], EAGLi [25] systems since as far as we know, they have not been recently evaluated on comprehensive biomedical QA benchmarks, such as the ones provided by BioASQ. We finally provide a discussion on the comparison of SemBioNLQA to these systems.

|  |
| --- |
| **Fig. 3.** An example of mapping to UMLS Metathesaurus and synonyms extraction. |

The BioASQ competition, within 2017 edition [56], comprised three tasks: (1) Task 5a on large-scale online biomedical semantic indexing, (2) Task 5b on biomedical semantic QA, and (3) Task c on funding information extraction from biomedical literature. The goal of Task 5b is to assess the performance of QA systems in different stages of the QA process. It is sub-divided into two phases: phase A and phase B. BioASQ created and released benchmark datasets of biomedical questions in English, along with gold standard answers. There were four types of questions: yes/no, factoid, list, and summary questions. In phase A participants had to respond with biomedical concepts, relevant documents, relevant passages, and RDF triples. In phase B participants were asked to answer with exact answers and ideal answers (paragraph-sized summaries). Exact answers are only required in the case of yes/no, factoid, list, while ideal answers are expected to be returned for questions. The released questions were accompanied by their types and the correct answers for the required elements (documents and passages) of the first phase.

## Evaluation metrics

The performance of the proposed biomedical QA system is evaluated using the evaluation measures described by the BioASQ challenge [41]. In the case of yes/no questions, the exact answers of SemBioNLQA had to be either “yes” or “no”. Therefore, accuracy is the main evaluation metric used to evaluate responses of yes/no questions. Let *n* be, the number of yes/no questions, and *k* the number of correctly answered yes/no questions, accuracy is computed as follows:

*k*

Accuracy *=*

*n* (1)

For factoid questions, mean reciprocal rank (MRR) is the main score used to evaluate factoid QA systems. Assuming that there are *n* factoid questions, MRR is defined as follows, where *ri* is the topmost position that contains the golden entity name (or one of its synonyms) in the returned list of possible responses.

1 *n* 1

MRR *=* \*

*n i=*1 *ri* (2)

On the other hand, to evaluate the exact answers of list questions, the mean average precision, mean average recall, and mean average Fmeasure metrics, which are computed by averaging precision (P), recall (R), and F1-measure (F1) over the list questions are used. These measures are computed as follows, where TP is the number of possible answers that are included both in the returned and the golden list; FP is the number of possible answers that are mentioned in the returned, but not in the golden list; and FN is the number of possible answers that are included in the golden, but not in the returned list. Indeed, mean average F-measure is the official score used by the BioASQ challenge for list questions.

 TP TP

*P= R= F*1 *=* 2\* TP *+* FP TP *+* FN(3)

Finally, the ideal answers of questions (all questions types: yes/no, factoid, list and summary), were automatically evaluated using Rouge measures. More details of these evaluations metrics appear in [60].

## Evaluation datasets

A publicly available evaluation infrastructure for biomedical QA developed by the BioASQ challenge [41], which includes benchmark datasets, is used to evaluate the performance of SemBioNLQA. The benchmark datasets, containing biomedical questions in English and golden standard answers, have been constructed and manually annotated by the BioASQ team of biomedical experts. There were four types of questions: yes/no, factoid, list, and summary questions. In phase B, Task b of the BioASQ challenge, within each edition, the organizers released a training set of questions-answers pairs and five batches of test questions. Each batch contains approximately 100 biomedical questions along with corresponding types, golden documents, and golden passages. In this study, the BioASQ 2015, 2016 and 2017 datasets are used to evaluate the effectiveness of the SemBioNLQA system. The following are some example of biomedical questions from the BioASQ training dataset:

* What medication were compared in the ROCKET AF Trial? (identifier 56bb616dac7ad10019000008)
* Which enzyme does MLN4924 inhibit? (identifier

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2**  The overall results of the proposed answer extraction methods in SemBioNLQA and comparison with the state-of-the-art systems on five batches of testing datasets provided by BioASQ 3b 2015. The “-” replace the scores of systems that did not evaluate on this batch or did not deal with this task, while “nr” indicated that the results are not reported for this evaluation measure. P, R, and F1 indicate precision, recall, and F1-measure respectively.  Datasets  System name  Exact answers  Idial answers  Yes/No  Factoid  List   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  | Accuracy | MRR | P | R | F1 | Rouge-2 | Rouge-SU4 | | Batch 1 | SemBioNLQA | 0.6970 | 0.1692 | 0.1545 | 0.2409 | 0.1830 | 0.2716 | 0.2860 | |  | Zhang et al. [29] | - | 0.1423 | nr | nr | 0.0756 | - | - | |  | Neves [31] | 0.6667 | - | 0.0292 | 0.0603 | 0.0364 | 0.1884 | 0.2008 | |  | Choi [30] | - | - | - | - | - | - | 0.3071 | | Batch 2 | SemBioNLQA | 0.6250 | 0.1776 | 0.1929 | 0.2714 | 0.2127 | 0.3123 | 0.3364 | |  | Zhang et al. [29] | - | 0.0859 | nr | nr | 0.1160 | - | - | |  | Neves [31] | 0.5625 | - | 0.0714 | 0.0161 | 0.0262 | 0.2026 | 0.2227 | |  | Choi [30] | - | - | - | - | - | - | 0.3710 | | Batch 3 | SemBioNLQA | 0.8621 | 0.1840 | 0.2353 | 0.2927 | 0.2524 | 0.3879 | 0.4078 | |  | Zhang et al. [29] | - | 0.0846 | nr | nr | 0.1319 | - | - | |  | Yang et al. [28] | - | 0.1615 | 0.0539 | 0.6933 | 0.0969 | - | - | |  | Neves [31] | 0.6207 | - | - | - | - | 0.1934 | 0.2189 | |  | Choi [30] | - | - | - | - | - | - | 0.3941 | | Batch 4 | SemBioNLQA | 0.7600 | 0.2960 | 0.2783 | 0.2713 | 0.2588 | 0.3917 | 0.4108 | |  | Zhang et al. [29] | - | 0.2299 | nr | nr | 0.2192 | - | - | |  | Yang et al. [28] | - | 0.5155 | 0.3836 | 0.3480 | 0.3168 | - | - | |  | Neves [31] | 0.5600 | 0.0345 | 0.1522 | 0.0473 | 0.0689 | 0.2504 | 0.2724 | |  | Choi [30] | - | - | - | - | - | - | 0.3906 | | Batch 5 | SemBioNLQA | 0.6071 | 0.1568 | 0.0583 | 0.0736 | 0.0625 | 0.3440 | 0.3533 | |  | Zhang et al. [29] | - | 0.2500 | nr | nr | 0.1340 | - | - | |  | Yang et al. [28] | - | 0.2727 | 0.1704 | 0.2573 | 0.1875 | - | - | |  | Neves [31] | 0.3571 | 0.0909 | 0.0625 | 0.0292 | 0.0397 | 0.1694 | 0.1790 | |  | Choi [30] | - | - | - | - | - | - | 0.3665 | |

56ed03862ac5ed1459000004)

## Results

In order to assess the effectiveness of the SemBioNLQA system and compare with the current integral biomedical QA systems, we present two different evaluations: (1) a systematic/automatic evaluation on benchmark datasets provided by the BioASQ challenges, and (2) a manual evaluation in terms of quality of the answers using BioASQ training questions and answers.

### Systematic evaluation

In this experiment, we present a systematic evaluation on biomedical questions provided by the BioASQ challenge so as to compare with BioASQ participant systems. As we previously noted, the BioASQ challenges in phase B (i.e., exact an ideal answers) of Task b provide the test set of biomedical questions along with their golden documents, golden snippets, and questions types [61,62,56] and participant systems [29,31,30,28,32] were asked to answer with exact answers and ideal answers using the golden documents, golden snippets, and golden questions types. 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. Therefore, to compare with such systems, we have also used gold-standard passages and questions types provided by BioASQ, instead of the ones retrieved by the proposed system.

Table 2 and Table 3 show the experimental results of the proposed answer extraction methods in SemBioNLQA and comparison with stateof-the-art studies presented in [29,31,30,28,32] on five batches of testing datasets provided by the BioASQ challenge in 2015 and 2016 respectively.

In addition, Table 4 and Table 5 present the results of our participation in Phase B, Task 5b of the 2017 BioASQ challenge using only the proposed answer extraction methods in SemBioNLQA. The values inside parameters indicate our current rank, the total number of submissions, and the total number of participated teams for the task. Note that in contrast to the 2015 and 2016 BioASQ challenges, biomedical QA systems are no longer allowed to provide an own list of synonyms in the 2017 challenge. The full results of phase B of task 5b are available online[[5]](#footnote-5) . Our system name for submission was “sarrouti”.

In particular, we also report the overall evaluation results of end-toend SemBioNLQA system on BioASQ datasets provided by the challenge in 2015 and 2016 editions in order to demonstrate the effectiveness of the SemBioNLQA, the impact of its different components on the overall performance, and also to make new comparisons easier. In this evaluation, we have used the relevant documents, relevant passages and question types returned by the proposed system. A direct comparison with the state-of-the-art methods presented in [31,29,32,28,30] and evaluated on either the 2015 or 2016 BioASQ challenges is not simple since the authors used the test set of biomedical questions along with their golden documents, golden snippets, and questions types released by the BioASQ challenges. As it has already been stated before, to a given biomedical question, SemBioNLQA first retrieves the *D* topranked documents, then finds the *P* top-ranked passages and finally applies the appropriate answer extraction according to the question type detected by the question classification module. In particular, we have decided to go with the *D* = 10 top-ranked documents and *P* = 10 top-ranked passages since only the 10 first ones from the resulting list are permitted for the test in the 2015 and 2016 BioASQ challenges. Table 6 highlights the obtained results on BioASQ 3b 2015 and BioASQ 4b 2016 datasets using the BioASQ Oracle system,13 an on-line system that allows uploading JSON result files and obtaining evaluations at any time. All answers returned by the SemBioNLQA system on either BioASQ 3b 2015 and BioASQ 4b 2016 datasets are available for download[[6]](#footnote-6) . On the other hand, Fig. 4 , Fig. 5 and Fig. 6 show the SemBioNLQA output for four biomedical questions which come from BioASQ training questions.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3**  The overall results of the proposed answer extraction methods in SemBioNLQA and comparison with the current state-of-the-art methods on five batches of testing datasets provided by BioASQ 4b 2016. The “-” replace the scores of systems that did not deal with this task, while “nr” indicated that the results are not reported for this evaluation measure. P, R, and F1 indicate precision, recall, and F1-measure, respectively.  Datasets  System name  Exact answers  Idial answers  Yes/No  Factoid  List   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  | Accuracy | MRR | P | R | F1 | Rouge-2 | Rouge-SU4 | | Batch 1 | SemBioNLQA | 0.8214 | 0.0726 | 0.2182 | 0.3939 | 0.2756 | 0.4772 | 0.4918 | |  | Schulze et al. [32] | - | - | - | - | - | nr | 0.2231 | | Batch 2 | SemBioNLQA | 0.8750 | 0.1452 | 0.2381 | 0.2505 | 0.2349 | 0.5021 | 0.5115 | |  | Schulze et al. [32] | - | - | - | - | - | nr | 0.2240 | | Batch 3 | SemBioNLQA | 0.8400 | 0.1218 | 0.2381 | 0.3627 | 0.2812 | 0.4978 | 0.5061 | |  | Schulze et al. [32] | - | - | - | - | - | nr | 0.2559 | | Batch 4 | SemBioNLQA | 0.8095 | 0.1129 | 0.1467 | 0.2231 | 0.1702 | 0.5192 | 0.5231 | |  | Schulze et al. [32] | - | - | - | - | - | nr | 0.2280 | | Batch 5 | SemBioNLQA | 0.8148 | 0.1136 | 0.1900 | 0.2353 | 0.1963 | 0.4979 | 0.5027 | |  | Schulze et al. [32] | - | - | - | - | - | nr | 0.3233 |   **Table 4**  Obtained results of our participation in “Exact Answers”, Phase B, Task 5b of the 2017 BioASQ challenge using the proposed answer extraction methods. The first value inside parameters indicates our current rank and the total number of submissions for the task, while the second indicates our current rank and the total number of participated teams.  Datasets  Yes/No  Factoid  List   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Accuracy | MRR | Precision | Recall | F-measure | | Batch 1 | 0.7647 | 0.1833 (6/15) (4/9) | 0.1909 | 0.2539 | 0.2037 (4/15) (3/9) | | Batch 2 | 0.7778 | 0.0887 (11/21) (5/9) | 0.2400 | 0.3922 | 0.2920 (7/21) (3/9) | | Batch 3 | 0.8387 (1/21) (1/10) | 0.2212 (9/21) (4/10) | 0.2000 | 0.4117 | 0.2625 (6/21) (3/10) | | Batch 4 | 0.6207 (2/27) (2/11) | 0.0970 (13/27) (5/11) | 0.1077 | 0.2013 | 0.1369 (12/27) (5/11) | | Batch 5 | 0.4615 | 0.2071 (9/25) (3/11) | 0.2182 | 0.3178 | 0.2529 (11/25) (6/11) | |

### Manual evaluation

This evaluation aims at comparing SemBioNLQA with end-to-end biomedical QA systems including AskHermes, EAGLi, Olelo. To do so, we have used the same strategy used in [27]. Indeed, we randomly selected 30 questions from the BioASQ training dataset and posed these to the four systems - AskHermes, EAGLi, Olelo and SemBioNLQA. This evaluation was carried out manually, and therefore, we needed to limit the number of questions and types. We decided to limit it to factoid, list and yes/no questions given that these types of answers are easier to check manually than summaries. This sequence of 30 questions, which are listed in Appendix A, contains 10 factoid questions, 11 list questions and 9 yes/no questions. In our evaluation, an answer is considered as correct if the first returned biomedical entity (for factual questions), at least one of the first five returned biomedical entities (for list questions) or the Boolean value, i.e., “yes” or “no”, (for yes/no questions) is correct. Indeed, we manually checked the results returned by each system to look for the correct standard answers as provided by the BioASQ challenge. Table 7 presents and compares the results of the aforementioned systems and SemBioNLQA. All answers returned by the systems are available for download[[7]](#footnote-7) . Please note that Medline database updates may result in different retrieved documents, snippets or answers.

## Discussion

While open-domain QA has been widely studied, few integral systems such as the ones described in [25,26,38] are currently able to automatically answers questions from the ever-increasing volume of peer-reviewed scientific articles in the biomedical domain. In this work, we addressed shortcomings of these systems, such as limited usability and performance in terms of the precision for the currently supported question and answer types. As results, unlike these systems, our developed biomedical QA system SemBioNLQA has the ability to handle with a large amount of questions and answers types such as yes/no, factoid, list and summary questions that may cover all types of questions.

The results of the systematic evaluation have shown that SemBioNLQA is more competitive as compared with current state-ofthe-art systems. As shown in Table 2, compared with the two approaches proposed in [31,29], one based on the in-memory database and another using PubTator, our approach gets better results. The presented system significantly outperformed the aforementioned systems in extracting both the exact answers and the ideal answers for yes/ no, factoid, list and summary biomedical questions. Moreover, the increased performance was statistically significant (the P-value is 7.6e-05, the result is significant at p < 0.01). Additionally, the proposed system is competitive compared with the approach presented in [30], which dealt only with the ideal answers (0.2860 against 0.3071, 0.3364 against 0.3710, 0.4078 against 0.3941, 0.4108 against 0.3906, and 0.3533 against 0.3665 of Rouge-SU4 in batch 1, batch 2, batch 3, batch 4, and batch 5, respectively). On the other hand, as shown in Table 3, SemBioNLQA still achieves good performance compared with the 2016 winning system developed by Schulze et al. [32] which dealt solely with the ideal answers of questions. The latter is based on the LexRank algorithm [40], but that solely used the named entities for the similarity function. The important thing to note here is that our system significantly outperforms the Schulze et al.'s [32] system in all batches of testing datasets. The largest difference in terms of Rouge-SU4 between SemBioNLQA and the aforementioned system was 0.2951 (0.5231 0.2280 in batch 4) which clearly indicates that our method is not only effective but robust in extracting the ideal answers. As part of our participation [36] in phase B (exact and ideal answers), Task 5B of the

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 5**  Obtained results of our participation in “Ideal Answers”, Phase B, Task 5b of the 2017 BioASQ challenge using the proposed answer extraction methods. The first value inside parameters indicates our current rank and the total number of participated teams for the task.  Datasets  Automatic scores  Manual scores   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | Rouge-2 | Rouge-SU4 | Readability | Recall | Precision | Repetition | | Batch 1 | 0.5087 (2/9) | 0.5247 (2/9) | 3.65 (2/9) | 4.42 (2/9) | 3.90 (2/9) | 3.89 (2/9) | | Batch 2 | 0.4823 (2/9) | 0.4828 (2/9) | 3.68 (2/9) | 4.59 (2/9) | 4.01 (2/9) | 3.91 (2/9) | | Batch 3 | 0.5658 (2/10) | 0.5729 (2/10) | 3.91 (2/10) | 4.64 (2/10) | 4.07 (2/10) | 4.00 (2/10) | | Batch 4 | 0.5667 (3/11) | 0.5640 (3/11) | 3.86 (3/11) | 4.51 (3/11) | 4.02 (2/11) | 3.95 (3/11) | | Batch 5 | 0.5616 (3/11) | 0.5595 (3/11) | 3.82 (3/11) | 4.53 (2/11) | 3.91 (3/11) | 3.90 (3/11) |   **Table 6**  The overall evaluation results of the SemBioNLQA system on five batches of biomedical questions provided by BioASQ 3b 2015 and BioASQ 4b 2016.  Datasets  Batches  Exact answers  Idial answers  Yes/No  Factoid  List   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  | Accuracy | MRR | Precision | Recall | F1-measure | Rouge-2 | Rouge-SU4 | | BioASQ 3b 2015 | Batch 1 | 0.7273 | 0.0128 | 0.0364 | 0.0682 | 0.0462 | 0.1039 | 0.1397 | |  | Batch 2 | 0.6875 | 0.0339 | 0.0714 | 0.1232 | 0.0891 | 0.1044 | 0.1413 | |  | Batch 3 | 0.8621 | 0.0641 | 0.0353 | 0.0368 | 0.0352 | 0.1269 | 0.1557 | |  | Batch 4 | 0.6800 | 0.0586 | 0.0696 | 0.0880 | 0.0751 | 0.1467 | 0.1702 | |  | Batch 5 | 0.6786 | 0.0795 | 0.0083 | 0.0208 | 0.0119 | 0.0915 | 0.1205 | | BioASQ 4b 2016 | Batch 1 | 0.8214 | 0.0534 | 0.1091 | 0.1545 | 0.1268 | 0.1552 | 0.1855 | |  | Batch 2 | 0.7188 | 0.0495 | 0.0476 | 0.0477 | 0.0462 | 0.1378 | 0.1720 | |  | Batch 3 | 0.8800 | 0.1186 | 0.0857 | 0.1667 | 0.1122 | 0.1430 | 0.1730 | |  | Batch 4 | 0.8095 | 0.0253 | 0.0400 | 0.0667 | 0.0500 | 0.1097 | 0.1289 | |  | Batch 5 | 0.8519 | 0.0687 | 0.0300 | 0.0517 | 0.0368 | 0.1609 | 0.1836 | |

2017 BioASQ challenge, the proposed answer extraction methods in SemBioNLQA performed well in the challenge ranking within the top tier teams as shown in Table 4 and Table 6 . A total of 16, 22, 21, 27, and 25 runs were submitted for batch 1, batch 2, batch 3, batch 4, and batch 5 respectively. Note that many runs were submitted by the same teams. The proposed system was one of the challenge winners[[8]](#footnote-8) .

From another side, as shown in Table 7 which presents and compares the results of the manual evaluation of SemBioNLQA, EAGLi, AskHERMES and also Olelo in terms of number of recognized questions and correct answers, SemBioNLQA gets better results and succeeded at answering the majority of randomly selected BioASQ questions. We have manually analyzed the answers provided for the biomedical questions by each system. In contrast to SemBioNLQA which has proven to be quite successful at extracting the exact answers depending on the expected answer for each question type, Olelo returned a summary as answer for the most questions and AskHERMES returned a multiple sentence passage as answer for all questions. Indeed, SemBioNLQA was able to detect the questions type were of the factoid, list or yes/no types, and thus generated exact answers depending on the expected answer for each question type. This indicates that its integration of our question classification method offers SemBioNLQA the ability to understand and correctly recognize the information needs of users. In contrast, even though Olelo was developed to handle with factoid, list and summary questions, it was not able to detect the types for given questions, and thus generated summaries for all questions, and therefore, the users have to read these summaries so as to find the precise answers. In particular, it only returns exact answers when both the headword and semantic types are detected, in addition to the candidate answers being of this same semantic type. On the other hand, as shown in Table 7, both SemBioNLQA and Olelo have succeeded in returning answers for all questions, while AskHERMES and EAGLi could not provide answers for the majority of the questions, instead, only the following messages “Nothing found! Please refine your question” in the former and “EAGLi did not understand your question. Try a popular example, or go to the manual mode.” in the latter.

Although the SemBioNLQA system could effectively answer a variety of biomedical questions, we found that there are still some mistakes that the proposed system cannot fix. For example, we found that the current form of the proposed system was not able to provide answers to some questions especially for these which expect a number as answer instead of biomedical entities. For instance, the answers for the biomedical question “What is the prevalence of short QT syndrome?” (identifier 52fb78572059c6d71c000067) and “What is the number of protein coding genes in the human genome?” (identifier 535d3c069a4572de6f000006) collected from BioASQ training questions, are “0.01% -0.1%”, “Between 20,000 and 25,000”, respectively. Such questions seem to be quite complicate and need more specific information extraction methods. On the other hand, it is clear from Table 6 that the different components of the SemBioNLQA system have a significant impact on the answer extraction task and therefore on the overall performance of SemBioNLQA since if the set of retrieved documents, passages and the type of a given question are not identified correctly, further processing steps to extract the answers will inevitably fail too. For instance, for the question “What is the association of spermidine with *α*-synuclein neurotoxicity?” (identifier 56c073fcef6e394741000020) from the batch 1 of test set of the 2016 BioASQ challenge, the returned type of question is “summary” whereas in the corpus the type of question is “factoid”. Therefore, the SemBioNLQA system will inevitably fail to extract and output the correct answer since extracting the answer to a factoid question, which is asking for a biomedical entity, is not the same as extracting the answer to a summary question which is looking only for an ideal answer.

|  |
| --- |
| **Fig. 4.** The SemBioNLQA output for the biomedical question “What symptoms characterize the Muenke syndrome?” (identifier 52bf1d3c03868f1b0600000d). |

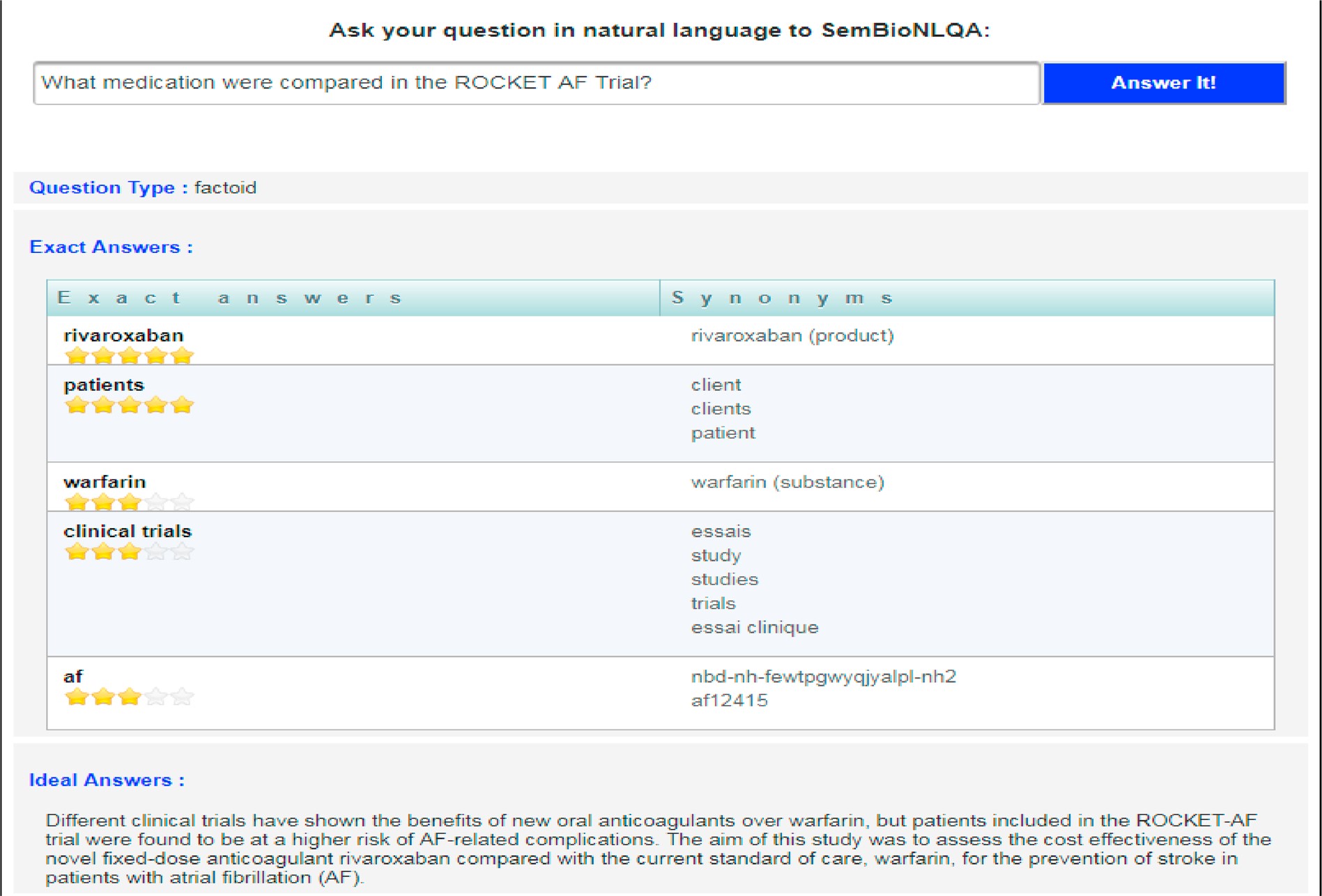
Overall, SemBioNLQA holds a number of advantages over the stateof-the-art systems. First, the integration of our question types classification method it offers have a clear advantage over Olelo in that it returns exact answers which depend on the expected answer of each question type. Second, SemBioNLQA provides an unbeatable advantage over AskHERMES, EAGli and Olelo in that it handles with a large amount of questions types including yes/no, factoid, list and summary questions. Third, the systematic and manual evaluations results demonstrated that SemBioNLQA is more effective as compared with the aforementioned systems.

In summary, biomedical QA is a very challenging task since it allows queries to be posed as natural language questions and provides precise answers instead of only presenting potentially relevant documents by integrating various resources. Therefore, no current system can always perform well on the myriad questions that can be asked of it. SemBioNLQA provides a practical and competitive alternative to help users find exact and ideal answers.

# Conclusion and future work

In this paper, we tackled a fully automatic QA system in the biomedical domain, SemBioNLQA, which has the ability to deal with four types of biomedical questions including yes/no questions, factoid questions, list questions, and summary questions. SemBioNLQA is currently able to provide exact answers (e.g., “yes”, “no”, a biomedical

|  |
| --- |
| **Fig. 5.** The SemBioNLQA output for the biomedical question “What medication were compared in the ROCKET AF Trial?” (identifier 56bb616dac7ad10019000008).    **Fig. 6.** The SemBioNLQA output for the biomedical question “Are mutations in the C9orf72 gene associated with macular degeneration?” (identifier |

58e11bf76fddd3e83e00000c).

**Table 7**

Comparison of the results of SemBioNLQA, EAGLi, AskHERMES and Olelo in terms of number of recognized questions and correct answers

|  |  |  |
| --- | --- | --- |
| Systems | Number of recognized questions | Number of correct answers |
| EAGLi [25] | 7/30 | 3/30 |
| AskHERMES [26] | 12/30 | 2/30 |
| Olelo [38] | 30/30 | 6/30 |
| SemBioNLQA | 30/30 | 18/30 |

entity, etc.) and paragraph-sized ideal answers (summaries of relevant information) for yes/no, factoid and list questions, whereas it only retrieves ideal answers for summary questions. The system relied on (1) handcrafted lexico-syntactic patterns and a machine learning approach for question classification, (2) PubMed search engine and UMLS similarity for document retrieval, (3) the BM25 model, stemmed words and UMLS concepts for passage retrieval, and (4) UMLS metathesaurus, BioPortal synonyms, sentiment analysis and term frequency metric for answer retrieval.

Thanks to an evaluation on standard benchmarks provided by the 2017 BioASQ challenge during our participation in phase B of Task 5b, we noted that SemBioNLQA achieved promising performances compared with the current integral biomedical QA system. Moreover, the manual evaluation results have shown that SemBioNLQA achieved good performances in terms of number of recognized questions and correct answers compared with the most current state-of-the-art system, as well as provides a practical and competitive alternative to help information seekers find exact and ideal answers for their biomedical questions.

To improve the performance of the SemBioNLQA system, three main perspectives are planned for future work:

* We are currently looking at several ways to incorporate deep learning into SemBioNLQA as deep learning models have been emerging as state-of-the-art for QA, especially in the open domain. However, deep learning models are usually applied for factoid-type QA. Moreover, such approaches require a large number of questionanswer pairs for the training phase.
* We are constantly working on some specific NLP methods to incorporate relationships between biomedical entities in order to improve ranking and better answers to currently supported question types.
* We intend to start working with full text (e.g., from PubMed Central) to see if we can improve SemBioNLQA performance with a focused look at the full text. MEDLINE which indexes more than 25 million references, contains a link to the free full text of the article archived in PubMed Central that contains 5.1 million articles.

# Author contribution

All authors designed the manuscript content, implemented the system, contribute to write the various iterations of the paper and give their final approval of the submitted manuscript.

**Competing interests**

The authors declare no conflict of interest.

# Summary points

* Development of an automatic biomedical question answering system named SemBioNLQA for retrieving exact and ideal answers to biomedical questions from peer-reviewed scientific articles.
* It is based on (1) handcrafted lexico-syntactic patterns and support vector machine (SVM) classifier for question classification, (2) PubMed and UMLS similarity for document retrieval, (3) the BM25 model, stemmed words and UMLS concepts for passage retrieval, and (4) UMLS metathesaurus, BioPortal synonyms, SENTIWORDNET and term frequency metric for answer extraction.
* It is able to accept a variety of natural language questions and to generate appropriate answers.
* It provides exact answers “yes” or “no” for yes/no questions, UMLS entities for factoid questions, and a list of UMLS entities for list questions. In addition to exact answers for these question types, SemBioNLQA also returns ideal answers (paragraph-sized summaries), while it retrieves only ideal answers for summary questions.
* Experimental evaluations performed on biomedical questions provided by the BioASQ challenge show that SemBioNLQA achieves good performances compared with the current integral state-of-theart systems.

# Availability

The SemBioNLQA source code is available at [https://github.com/ sarrouti/sembionlqa.](https://github.com/sarrouti/sembionlqa)

# Appendix A. List of BioASQ questions used for the manual evaluation

1. Which is the gene most commonly mutated in Tay-Sachs disease?(identifier 536e46f27d100faa09000012)
2. What medication were compared in the ROCKET AF Trial? (identifier 56bb616dac7ad10019000008)
3. Which enzyme does MLN4924 inhibit? (id

56ed03862ac5ed1459000004)

1. Where is the protein Pannexin1 located? (identifier 56af9f130a360a5e45000015)
2. RTS S AS01 vaccine was developed to prevent which disease? (identifier 56bc77a3ac7ad10019000015)
3. Which type of myeloma is ixazomib being evaluated for? (identifier56ed0ffe2ac5ed1459000008)
4. What enzyme is inhibied by Opicapone? (id

56c1d857ef6e394741000033)

1. Which gene has been implicated in Majeed Syndrome? (id 56f7c15a09dd18d46b000012)
2. Which gene is most commonly associated with severe congenitaland cyclic neutropenia? (identifier 5503133ae9bde6963400001d)
3. Which is the receptor for the immunosuppressive drug cyclosporinA (CsA)? (identifier 56f6c11109dd18d46b00000e)
4. Which proteins participate in the formation of the Notch transcriptional activation complex? (identifier 54fb6fb5d176fff445000004)
5. What is being measured with an accelerometer in back pain patients? (identifier 533f9df0c45e133714000016)
6. List inhibtors targeting the mitochondrial permeability transitionpore (identifier 5717cdd2070aa3d072000001)
7. List symptoms of the IFAP syndrome (identifier

56c1f038ef6e394741000051)

1. What is the functional role of the protein Drp1? (identifier 5717dbfe7de986d80d000001)
2. Which receptors are bound by Tasimelteon? (identifier 56c1f043ef6e394741000057)
3. Which disease phenotypes are associated to PRPS1 mutations? (identifier 5713b0a51174fb175500000e)
4. Which genes are thought to be involved in medulloblastoma development? (identifier 5539029cbc4f83e828000012)
5. Which miRNAs could be used as potential biomarkers for epithelialovarian cancer? (identifier 553fa78b1d53b76422000007)
6. Which interleukins are inhibited by Dupilumab? (identifier 56c1f005ef6e39474100003a)
7. Which are the genes thought to be regulated by EWS/FLI? (identifier 552faa43bc4f83e828000004)
8. Is the ACE inhibitor indicated for lung cancer treatment? (identifier530cf4fe960c95ad0c000005)
9. Is PTEN involved in follicular thyroid carcinoma? (identifier 55031650e9bde69634000026)
10. Is Fanconi anemia presented as a genetically and clinically heterogeneous disease entity? (identifier 54ede5-

c394afd61504000006)

1. Can the iPS cell technology be used in Fanconi anemia therapy?(identifier 54edef0594afd6150400000d)
2. Does surgery for ovarian endometriomas improve fertility? (identifier 54f088ee94afd61504000015)
3. Is irritable bowel syndrome more common in women with endometriosis? (identifier 54f08d4a94afd61504000016)
4. Is the regulation of Vsr endonuclease independent of the growthphase of bacteria? (identifier 553fbe9fe00431e071000001
5. Does TRIM37 gene mutation causes Mulibrey nanism? (identifier56c1f03cef6e394741000054)
6. Is the gene MAOA epigenetically modified by methylation? (identifier 56cf50253975bb303a00000b)

# References

1. Sarker A, Mollá D, Paris C. Query-oriented evidence extraction to support evidencebased medicine practice. Journal of Biomedical Informatics 2016;59:169–84. [https://doi.org/10.1016/j.jbi.2015.11.010. URL https://doi.org/10.1016%2Fj.jbi. 2015.11.010.](https://doi.org/10.1016/j.jbi.2015.11.010)
2. Ji X, Ritter A, Yen P-Y. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. Journal of Biomedical Informatics 2017;69:33–42. [https://doi.org/10.1016/j.jbi.2017.03.007. URL https://doi.org/ 10.1016%2Fj.jbi.2017.03.007.](https://doi.org/10.1016/j.jbi.2017.03.007)
3. Frasca M. Gene2disco: Gene to disease using disease commonalities. Artificial Intelligence in Medicine 2017;82:34–46. [https://doi.org/10.1016/j.artmed.2017.](https://doi.org/10.1016/j.artmed.2017.08.001)

[08.001. URL https://doi.org/10.1016%2Fj.artmed.2017.08.001.](https://doi.org/10.1016/j.artmed.2017.08.001)

1. Kopanitsa G. Integration of hospital information and clinical decision support systems to enable the reuse of electronic health record data. Methods of Information in Medicine 2017;56(3):238–47. [https://doi.org/10.3414/me16-01-0057. URL https://doi.org/10.3414%2Fme16-01-0057.](https://doi.org/10.3414/me16-01-0057)
2. Kropf S, Krücken P, Mueller W, Denecke K. Structuring legacy pathology reports by openEHR archetypes to enable semantic querying. Methods of Information in Medicine 2017;56(2):230–7. [https://doi.org/10.3414/me16-01-0073. URL https:// doi.org/10.3414%2Fme16-01-0073.](https://doi.org/10.3414/me16-01-0073)
3. Kannampallil TG, Patel VL. Special issue on cognitive informatics methods for interactive clinical systems. Journal of Biomedical Informatics 2017;71:207–10. [https://doi.org/10.1016/j.jbi.2017.06.003. URL https://doi.org/10.1016%2Fj.jbi. 2017.06.003.](https://doi.org/10.1016/j.jbi.2017.06.003)
4. Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier.

Artificial Intelligence in Medicine 2017;83:67–74. [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.artmed.2017.03.001)

[artmed.2017.03.001. URL https://doi.org/10.1016%2Fj.artmed.2017.03.001.](https://doi.org/10.1016/j.artmed.2017.03.001)

1. Wang L, Fiol GD, Bray BE, Haug PJ. Generating disease-pertinent treatment vocabularies from MEDLINE citations. Journal of Biomedical Informatics 2017;65:46–57. [https://doi.org/10.1016/j.jbi.2016.11.004. URL https://doi.org/ 10.1016%2Fj.jbi.2016.11.004.](https://doi.org/10.1016/j.jbi.2016.11.004)
2. Chen J, Yu H. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. Journal of Biomedical Informatics

2017;68:121–31. [https://doi.org/10.1016/j.jbi.2017.02.016. URL https://doi.org/ 10.1016%2Fj.jbi.2017.02.016.](https://doi.org/10.1016/j.jbi.2017.02.016)

1. Lamy J-B. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. Artificial Intelligence in Medicine 2017;80:11–28. [https://doi.org/10.1016/j.artmed.2017.](https://doi.org/10.1016/j.artmed.2017.07.002)

[07.002. URL https://doi.org/10.1016%2Fj.artmed.2017.07.002.](https://doi.org/10.1016/j.artmed.2017.07.002)

1. Goldstein A, Shahar Y, Orenbuch E, Cohen MJ. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. Artificial Intelligence in Medicine 2017;82:20–33. [https:// doi.org/10.1016/j.artmed.2017.09.001. URL https://doi.org/10.1016%2Fj.artmed. 2017.09.001.](https://doi.org/10.1016/j.artmed.2017.09.001)
2. D. Hristovski, D. Dinevski, A. Kastrin, T.C. Rindflesch, Biomedical question answering using semantic relations, BMC Bioinformatics 16 (1). doi:10.1186/s12859014-0365-3. URL https://doi.org/10.1186%2Fs12859-014-0365-3.
3. Ely JW. A taxonomy of generic clinical questions: classification study. BMJ 2000;321(7258):429–32. [https://doi.org/10.1136/bmj.321.7258.429. URL https://doi.org/10.1136%2Fbmj.321.7258.429.](https://doi.org/10.1136/bmj.321.7258.429)
4. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P, Mosbaek C, Kraemer D. Factors associated with success in searching medline and applying evidence to answer clinical questions. Journal of the American Medical Informatics Association 2002;9(3):283–93. [https://doi.org/10.1197/jamia.M0996. arXiv:http://jamia.oxfordjournals.org/content/9/3/283.full.pdf, URL http://jamia. oxfordjournals.org/content/9/3/283.](https://doi.org/10.1197/jamia.M0996)
5. [Lee M, Cimino JJ, Zhu HR, Sable C, Shanker V, Ely JW, Yu H. Beyond information retrieval-medical question answering. AMIA Annual Symposium Proceedings, Vol. 2006 2006:469–73.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0075)
6. Athenikos SJ, Han H. Biomedical question answering: A survey. Computer Methods and Programs in Biomedicine 2010;99(1):1–24. [https://doi.org/10.1016/j.cmpb.](https://doi.org/10.1016/j.cmpb.2009.10.003)

[2009.10.003. URL https://doi.org/10.1016%2Fj.cmpb.2009.10.003.](https://doi.org/10.1016/j.cmpb.2009.10.003)

1. Neves M, Leser U. Question answering for Biology. Methods 2015;74:36–46. [https://doi.org/10.1016/j.ymeth.2014.10.023. URL https://doi.org/10.1016%2Fj. ymeth.2014.10.023.](https://doi.org/10.1016/j.ymeth.2014.10.023)
2. Abacha AB, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. Information Processing & Management 2015;51(5):570–94. [https://doi.org/10.1016/j.ipm.2015.04.006.](https://doi.org/10.1016/j.ipm.2015.04.006)

[URL https://doi.org/10.1016%2Fj.ipm.2015.04.006.](https://doi.org/10.1016/j.ipm.2015.04.006)

1. Green BF, Wolf AK, Chomsky C, Laughery K. Baseball. Proceedings of the western joint IRE-AIEE-ACM computer conference, Los Angeles, May 9-11 1961. [https:// doi.org/10.1145/1460690.1460714. URL https://doi.org/. doi: 10.1145/1460690. 1460714.](https://doi.org/10.1145/1460690.1460714)
2. Woods WA. Progress in natural language understanding. Proceedings of national computer conference and exposition- AFIPS, June 4-8 1973. [https://doi.org/10. 1145/1499586.1499695. URL https://doi.org/. doi: 10.1145/1499586.1499695.](https://doi.org/10.1145/1499586.1499695)
3. Katz B, Felshin S, Yuret D, Ibrahim A, Lin J, Marton G, McFarland AJ, Temelkuran B. Omnibase: Uniform access to heterogeneous data for question answering. Natural Language Processing and Information Systems 2002. p. 230–4. [https://doi.org/10. 1007/3-540-36271-1\_23. URL https://doi.org/. doi: 10.1007/3-540-36271-1\_23.](https://doi.org/10.1007/3-540-36271-1_23)
4. Kaisser M. The QuALiM question answering demo. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language

Technologies Demo Session - HLT 08 2008. [https://doi.org/10.3115/1564144. 1564153. URL https://doi.org/. doi: 10.3115/1564144.1564153.](https://doi.org/10.3115/1564144.1564153)

1. Bauer MA, Berleant D. Usability survey of biomedical question answering systems. Human Genomics 2012;6(1):17. [https://doi.org/10.1186/1479-7364-6-17. URL https://doi.org/10.1186%2F1479-7364-6-17.](https://doi.org/10.1186/1479-7364-6-17)
2. [Cruchet S, Gaudinat A, Rindflesch T, Boyer C. What about trust in the question answering world. AMIA 2009 annual symposium 2009.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0120)
3. [Gobeill J, Patsche E, Theodoro D, Veuthey A-L, Lovis C, Ruch P. Question answering for biology and medicine. 9th International Conference on Information Technology and Applications in Biomedicine 2009:1–5.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0125)
4. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, Ely J, Yu H. AskHERMES: An online question answering system for complex clinical questions. Journal of Biomedical Informatics 2011;44(2):277–88. [https://doi.org/10.1016/j.jbi.2011.01.](https://doi.org/10.1016/j.jbi.2011.01.004)

[004. URL https://doi.org/10.1016/j.jbi.2011.01.004.](https://doi.org/10.1016/j.jbi.2011.01.004)

1. Neves M, Eckert F, Folkerts H, Uflacker M. Assessing the performance of olelo, a real-time biomedical question answering application. BioNLP 2017, Association for Computational Linguistics 2017. [https://doi.org/10.18653/v1/w17-2344. URL https://doi.org/10.18653%2Fv1%2Fw17-2344.](https://doi.org/10.18653/v1/w17-2344)
2. [Yang Z, Gupta N, Sun X, Xu D, Zhang C, Nyberg E. Learning to answer biomedical factoid & list questions: OAQA at BioASQ 3B. CLEF (Working Notes) 2015.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0140)
3. [Zhang Y, Peng S, You R, Xie Z, Wang B, Zhu S. The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering. In Working Notes for the Conference and Labs of the Evaluation Forum (CLEF) 2015.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0145)
4. [Choi S. Snumedinfo at CLEF QA track BioASQ 2015. Working Notes for the Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France 2015.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0150)
5. [Neves M. HPI question answering system in the BioASQ 2015 challenge. Working Notes for the Conference and Labs of the Evaluation Forum (CLEF) 2015.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0155)
6. Schulze F, Schuler R, Draeger T, Dummer D, Ernst A, Flemming P, Perscheid C, Neves M. HPI question answering system in BioASQ 2016. Proceedings of the

Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics 2016:38–44. [https://doi.org/10.18653/v1/w16-3105. URL https://doi. org/10.18653%2Fv1%2Fw16-3105.](https://doi.org/10.18653/v1/w16-3105)

1. M. Sarrouti, S.O. E. Alaoui, A machine learning-based method for question type classification in biomedical question answering, Methods of Information in Medicine 56 (3). doi:10.3414/me16-01-0116. URL https://doi.org/10.

3414%2Fme16-01-0116.

1. Sarrouti M, Alaoui SOE. A generic document retrieval framework based on UMLS similarity for biomedical question answering system. Intelligent Decision Technologies 2016 2016. p. 207–16. [https://doi.org/10.1007/978-3-319-39627-9\_ 18. URL https://doi.org/10.1007%2F978-3-319-39627-9\_18.](https://doi.org/10.1007/978-3-319-39627-9_18)
2. Sarrouti M, Alaoui SOE. A passage retrieval method based on probabilistic information retrieval and UMLS concepts in biomedical question answering. Journal of Biomedical Informatics 2017;68:96–103. [https://doi.org/10.1016/j.jbi.2017.03. 001. URL https://doi.org/10.1016%2Fj.jbi.2017.03.001.](https://doi.org/10.1016/j.jbi.2017.03.001)
3. Sarrouti M, Alaoui SOE. A biomedical question answering system in BioASQ 2017. BioNLP 2017, Association for Computational Linguistics 2017. [https://doi.org/10.](https://doi.org/10.18653/v1/w17-2337)

[18653/v1/w17-2337. URL https://doi.org/10.18653%2Fv1%2Fw17-2337.](https://doi.org/10.18653/v1/w17-2337)

1. Vicedo JL, Gómez J. TREC: Experiment and evaluation in information retrieval. Journal of the American Society for Information Science and Technology 2007;58(6):910–1. [https://doi.org/10.1002/asi.20583. URL https://doi.org/10.](https://doi.org/10.1002/asi.20583)

[1002%2Fasi.20583.](https://doi.org/10.1002/asi.20583)

1. Kraus M, Niedermeier J, Jankrift M, Tietbohl S, Stachewicz T, Folkerts H, Uflacker M, Neves M. Olelo: a web application for intuitive exploration of biomedical literature. Nucleic Acids Research 2017;45(W1):W478–83. [https://doi.org/10.1093/ nar/gkx363. URL https://doi.org/10.1093%2Fnar%2Fgkx363.](https://doi.org/10.1093/nar/gkx363)
2. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research 2013;41(W1):W518–22. [https://doi.org/10.](https://doi.org/10.1093/nar/gkt441)

[1093/nar/gkt441. URL https://doi.org/10.1093%2Fnar%2Fgkt441.](https://doi.org/10.1093/nar/gkt441)

1. [Erkan G, Radev DR. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 2004;22:457–79.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0200)
2. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, Almirantis Y,

Pavlopoulos J, Baskiotis N, Gallinari P, Artiéres T, Ngomo A-CN, Heino N, Gaussier E, Barrio-Alvers L, Schroeder M, Androutsopoulos I, Paliouras G. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 2015;16(1):1–28. [https://doi.org/10.1186/s12859-0150564-6. URL https://doi.org/10.1186/s12859-015-0564-6.](https://doi.org/10.1186/s12859-015-0564-6)

1. Moldovan D, Paşca M, Harabagiu S, Surdeanu M. Performance issues and error analysis in an open-domain question answering system. ACM Transactions on Information Systems (TOIS) 2003;21(2):133–54. [https://doi.org/10.1145/763693. 763694. URL https://doi.org/10.1145%2F763693.763694.](https://doi.org/10.1145/763693.763694)
2. Cao Y-g, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. Journal of biomedical informatics 2010;43(6):962–71. [https://doi.org/10.1016/j.jbi.2010.07.007. URL https://doi.org/10.1016%2Fj.jbi. 2010.07.007.](https://doi.org/10.1016/j.jbi.2010.07.007)
3. Patrick J, Li M. An ontology for clinical questions about the contents of patient notes. Journal of Biomedical Informatics 2012;45(2):292–306. [https://doi.org/10. 1016/j.jbi.2011.11.008. URL https://doi.org/10.1016/j.jbi.2011.11.008.](https://doi.org/10.1016/j.jbi.2011.11.008)
4. [Roberts K, Kilicoglu H, Fiszman M, Demner-Fushman D. Automatically classifying question types for consumer health questions. AMIA Annu Symp Proc 2014:1018–27.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0225)
5. Sarrouti M, Lachkar A. A new and efficient method based on syntactic dependency relations features for ad hoc clinical question classification. International Journal of Bioinformatics Research and Applications 2017;13(2):161–77. [https://doi.org/10.](https://doi.org/10.1504/ijbra.2017.10003490)

[1504/ijbra.2017.10003490. URL https://doi.org/10.1504%2Fijbra.2017. 10003490.](https://doi.org/10.1504/ijbra.2017.10003490)

1. [Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0235) information systems 1980;14(3):130–7. [https://doi.org/10.1108/eb046814. URL](https://doi.org/10.1108/eb046814) [MetaMap program. Proceedings of the AMIA Symposium. 2001. p. 17.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0235) [https://doi.org/10.1108/eb046814.](https://doi.org/10.1108/eb046814)
2. Bodenreider O. The unified medical language system (UMLS): integrating biome- [55] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent dical terminology. Nucleic Acids Research 2004;32(90001):267D–70D. [https://doi.](https://doi.org/10.1093/nar/gkh061) advances. Journal of the American Medical Informatics Association [org/10.1093/nar/gkh061. URL https://doi.org/10.1093%2Fnar%2Fgkh061.](https://doi.org/10.1093/nar/gkh061) 2010;17(3):229–36. [https://doi.org/10.1136/jamia.2009.002733. URL https://](https://doi.org/10.1136/jamia.2009.002733) [49] Hanauer DA, Wu DT, Yang L, Mei Q, Murkowski-Steffy KB, Vydiswaran VV, Zheng [doi.org/10.1136%2Fjamia.2009.002733.](https://doi.org/10.1136/jamia.2009.002733)

K. Development and empirical user-centered evaluation of semantically-based [56] Nentidis A, Bougiatiotis K, Krithara A, Paliouras G, Kakadiaris I. Results of the fifth query recommendation for an electronic health record search engine. Journal of edition of the BioASQ challenge. BioNLP 2017 2017. [https://doi.org/10.18653/v1/](https://doi.org/10.18653/v1/w17-2306) Biomedical Informatics 2017;67:1–10. [https://doi.org/10.1016/j.jbi.2017.01.013.](https://doi.org/10.1016/j.jbi.2017.01.013) [w17-2306. URL https://doi.org/10.18653%2Fv1%2Fw17-2306.](https://doi.org/10.18653/v1/w17-2306)

[URL https://doi.org/10.1016%2Fj.jbi.2017.01.013.](https://doi.org/10.1016/j.jbi.2017.01.013) [57] [Cohen KB, Demner-Fushman D, Ananiadou S, Tsujii J. Proceedings of the 16th](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0285)

1. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and [BioNLP workshop. The Association for Computational Linguistics; 2017. URL](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0285) statistical techniques. Computational Linguistics 2007;33(1):63–103. [https://doi.](https://doi.org/10.1162/coli.2007.33.1.63) [http://www.aclweb.org/anthology/W17-2300.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0285)

[org/10.1162/coli.2007.33.1.63. URL https://doi.org/10.1162%2Fcoli.2007.33.](https://doi.org/10.1162/coli.2007.33.1.63) [58] Sarrouti M, Alaoui SOE. A yes/no answer generator based on sentiment-word scores

[1.63.](https://doi.org/10.1162/coli.2007.33.1.63) in biomedical question answering. International Journal of Healthcare Information

1. [McInnes BT, Pedersen T, Pakhomov SV. UMLS-interface and UMLS-similarity: open](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0255) Systems and Informatics 2017;12(3):62–74. [https://doi.org/10.4018/ijhisi.](https://doi.org/10.4018/ijhisi.2017070104) [source software for measuring paths and semantic similarity. AMIA Annual](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0255) [2017070104. URL https://doi.org/10.4018%2Fijhisi.2017070104.](https://doi.org/10.4018/ijhisi.2017070104)

[Symposium Proceedings, Vol. 2009 2009:431.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0255) [59] [Baccianella S, Esuli A, Sebastiani F. Sentiwordnet 3.0: An enhanced lexical resource](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0295) [52] Otterbacher J, Erkan G, Radev DR. Biased LexRank: Passage retrieval using random [for sentiment analysis and opinion mining. LREC, Vol. 10 2010:2200–4.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0295)

walks with question-based priors. Information Processing & Management [60] [Balikas G, Partalas I, Kosmopoulos A, Petridis S, Malakasiotis P, Pavlopoulos I,](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0300) 2009;45(1):42–54. [https://doi.org/10.1016/j.ipm.2008.06.004. URL https://doi.](https://doi.org/10.1016/j.ipm.2008.06.004) [Androutsopoulos I, Baskiotis N, Gaussier E, Artiere T, Gallinari P. Evaluation fra](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0300)[org/10.1016%2Fj.ipm.2008.06.004.](https://doi.org/10.1016/j.ipm.2008.06.004) [mework specifications. 2013. project deliverable d4.1, 05/2013.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0300)

1. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The stanford [61] [Balikas G, Kosmopoulos A, Krithara A, Paliouras G, Kakadiaris I. Results of the](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0305) CoreNLP natural language processing toolkit. Proceedings of 52nd Annual Meeting [BioASQ tasks of the question answering lab at CLEF 2015. CLEF 2015:2015.](http://refhub.elsevier.com/S0933-3657(18)30275-6/sbref0305) of the Association for Computational Linguistics: System Demonstrations 2014. [62] Krithara A, Nentidis A, Paliouras G, Kakadiaris I. Results of the 4th edition of [https://doi.org/10.3115/v1/p14-5010. URL https://doi.org/. doi: 10.3115/v1/](https://doi.org/10.3115/v1/p14-5010) BioASQ challenge. Proceedings of the Fourth BioASQ workshop 2016. [https://doi.](https://doi.org/10.18653/v1/w16-3101) [p14-5010.](https://doi.org/10.3115/v1/p14-5010) [org/10.18653/v1/w16-3101. URL https://doi.org/10.18653%2Fv1%2Fw16-3101.](https://doi.org/10.18653/v1/w16-3101)
2. Porter M. An algorithm for suffix stripping. Program: electronic library and

1. <http://www.bioasq.org/> 2 <http://www.bioasq.org/participate/fifth-challenge-winners> [↑](#footnote-ref-1)
2. <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?> [↑](#footnote-ref-2)
3. <http://metamap.nlm.nih.gov/> [↑](#footnote-ref-3)
4. <http://data.bioontology.org/documentation> [↑](#footnote-ref-4)
5. <http://participants-area.bioasq.org/results/5b/phaseB/>13 <http://participants-area.bioasq.org/oracle/> [↑](#footnote-ref-5)
6. <https://sites.google.com/site/mouradsarrouti/datasets> [↑](#footnote-ref-6)
7. <https://sites.google.com/site/mouradsarrouti/datasets> [↑](#footnote-ref-7)
8. <http://www.bioasq.org/participate/fifth-challenge-winners> [↑](#footnote-ref-8)