Assessing the performance of Olelo, a real-time biomedical question answering application

Mariana Neves, Fabian Eckert, Hendrik Folkerts, Matthias Uflacker
Hasso Plattner Institute at University of Potsdam
August Bebel Strasse 88, Potsdam 14482 Germany
mariana.neves@hpi.de

Abstract

Question answering (QA) can support physicians and biomedical researchers to find answers to their questions in the scientific literature. Such systems process large collections of documents in real time and include many natural language processing (NLP) procedures. We recently developed Olelo, a QA system for biomedicine which includes various NLP components, such as question processing, document and passage retrieval, answer processing and multi-document summarization. In this work, we present an evaluation of our system on the the fifth BioASQ challenge. We participated with the current state of the application and with an extension based on semantic role labeling that we are currently investigating. In addition to the BioASQ evaluation, we compared our system to other on-line biomedical QA systems in terms of the response time and the quality of the answers.

1 Introduction

Question answering (QA) is the task of automatically answering questions posed by users (Jurafsky and Martin, 2013). As opposed to information retrieval (IR), input is in the form of natural language, e.g., English, instead of keywords, and answers are provided as short answers, instead of presenting a list of relevant documents. Therefore, QA systems need to rely on various natural language processing (NLP) components, such as question understanding, named-entity recognition (NER), document and passage retrieval, answer extraction and multi-document summarization, among others. QA systems have been developed for many domains, including biomedicine (Athenikos and Han, 2010; Neves and Leser, 2015). Given the large collection of biomedical documents, e.g., in PubMed, researchers and physicians need to obtain answers for their various questions in a timely manner.

Much research has been published in the past for biomedical QA (Athenikos and Han, 2010), but focus was previously mainly on clinical documents. QA for biomedicine has recently gained importance owing to the BioASQ challenges (Tsatsaronis et al., 2015), for which the organizers created comprehensive datasets of questions, answers and intermediate results. The BioASQ challenge considers four types of questions: (i) yes/no, (ii) factoid, (iii) list and (iv) summary. For yes/no questions, a system should return either of the two answers, factoid and list questions expect one or more short answers, e.g., a gene name, while a short paragraph should be generated as answer for summary questions. Despite the accessibility of these datasets to support development and evaluation of QA systems for biomedicine, few QA applications are currently available on-line.

We recently developed Olelo[1], a QA system for biomedicine (Kraus et al., 2017). It relies on a local index of the Medline documents, includes domain terminologies and implements algorithms specifically designed for biomedical QA. Previous versions of our system were evaluated in the last three editions of the BioASQ challenges (Schulze et al., 2016; Neves, 2015, 2014).

In this work, we perform a comprehensive evaluation of our application, both automatically, during participation in the fifth edition of the BioASQ challenge, as well as manually, by checking our answers against the gold standard ones from BioASQ benchmarks. We also present re-

[1] http://hpi.de/plattner/olelo

sults for a new extension based on semantic role labeling (SRL) that we are considering for our application. Finally, we performed a comparison of Olelo to the other on-line biomedical QA tools.

2 Related work

We are only aware of three other biomedical QA services, as surveyed in (Bauer and Berleant, 2012), namely, askHERMES (Cao et al., 2011), EAGLi (Gobeill et al., 2015) and HONQA (Cruchet et al., 2009). However, none of the these systems performs robustly to most question types. Further, as far as we know, they have not been recently evaluated on comprehensive biomedical QA benchmarks, as the ones provided by BioASQ.

askHermes extracts answers from various sources, e.g., PubMed and Wikipedia, and presents answers as a cluster of terms, a ranked list or clustered by content, along with the corresponding relevant passages. However, the result page tends to be very long and contains more information than most users can deal with. The methods behind askHermes include regular expressions for question understanding, classification into 12 topics and keyword identification, both based on machine learning approaches, and the use of the MetaMap system for concept recognition. Document indexing is based on the BM25 model and passage ranking is based on the longest common subsequence (LCS) score.

EAGLi extracts the answers exclusively from PubMed abstracts and returns a list of concepts as answers. When no answer is found, the system returns a list of potential relevant publications, along with selected passages. The system locally indexes Medline with the Terrier information retrieval platform and uses the Okapi BM25 as weighting scheme to rank documents. EAGLi provides answers based on the Gene Ontology (GO) concepts.

Finally, HONQA relies on certified websites from the Health On The Net (HON), from which it extracts the answers, and considers a variety of question types. Questions can also be posed in French and Italian. The system relies on UMLS to detect the type of the expected answer and it follows the typical architecture of QA systems, but no details are presented in the publication.

3 Methods

In this section, we briefly describe the current methods behind Olelo as well as its extension for answer extraction based on SRL.

3.1 Olelo QA application

Olelo relies on the typical three steps of QA workflow (Athenikos and Han, 2010), namely, question processing, document/passage retrieval and answer extraction. Details of our methods have been previously published (Kraus et al., 2017; Schulze and Neves, 2016), but we give an overview of these below.

The application is built on top of an in-memory database (IMDB) that accounts for data storage (question, documents and terminologies) and text analysis. The latter are based both on built-in text processing features from the database, namely, sentence splitting, tokenization and part-of-speech tagging, as well as custom implemented SQL procedures for some QA components, such as question understanding, multi-document summarization and answer extraction. The database also includes an NER component based on custom dictionaries that we compiled based on concepts from MeSH and UMLS. Our document collection currently includes Medline abstracts and full text from PubMed Central Open Access.

When a question is posed to the system, its type (e.g., factoid or summary) is extracted via regular expressions. Further, in the case of factoid or list questions, the expected semantic types are detected, e.g., whether a gene or disease name. A query is then generated for the question based on the detected named entities (from the NER component) and other keywords from the question. Relevant documents and passages are then retrieved based on some simple heuristics that consider keywords and named entities from the question. For the answer extraction, different approaches are considered depending on the question type. For summary questions, a custom summary is generated based on the relevant sentence and corresponding named entities. Our approach is based on a graph-based approach for sentence selection (Schulze and Neves, 2016). In the case of factoid questions, and given the set of potentially relevant sentences, the system returns the corresponding MeSH concepts which belong from the same types of the expected type.

Contrary to BioASQ, our application does not distinguish between factoid and list questions, thus, more that one exact answer can be returned for a factoid question. Further, it does not yet support yes/no questions. Finally, Olelo supports definition questions, e.g., “What is zika virus?”, a type not supported in BioASQ. For these cases, the system returns the respective MeSH definition.

3.2 Semantic role labeling for answer extraction

We currently investigate an extension to our system based on SRL whose goal is to improve both the question understanding and answer extraction steps. Our aim is to find correct answers by identifying semantic conformities between a question and its snippets. As a first investigative step, we experimented with the BioKIT SRL tool (Dahlmeier and Ng, 2010) and used it to label the datasets from the first three years of the BioASQ challenge. We propose an initial rule-based proof-of-concept approach to investigate if SRL could improve Olelo QA system. Therefore, we put focus on finding suitable rules for all question types supported in the BioASQ challenge, i.e., yes/no, factoid, list and summary. In our experiments, we relied only on the gold-standard snippets provided by BioASQ, instead of the ones retrieved by Olelo.

Yes/no questions. When experimenting with yes/no questions, we soon observed problems with the skewed nature of the training data. As more than 4 out of 5 correct answers had to be ”yes”, the challenge was more about finding out in which rare cases to answer ”no”, instead of whether the answer was ”yes” or ”no”. The latter approach would regularly lead to worse results than the approach of simply answering ”yes” to every question. Initially, we were motivated by the idea that SRL could help us to be more confident when an answer was ”yes”. This could be achieved by finding a semantically matching answer to a question. However, due to the characteristics of the BioASQ data, being confident of when to output ”yes” was not helpful to improve results. Hence, we investigated if we could find out whether an answer is ”no” by applying a similar strategy.

We investigated the detection of negation. Looking at specific cases of the training data, we created rules which include negation terms or the occurrence of certain domain-specific terms. If multiple answer snippets matched, we calculated an overall score for taking the yes/no decision. For this score, answer snippets were weighted differently depending on the strength of their match.

Factoid and list questions. Factoid and list questions demand slightly different approaches. For both categories, we implemented a rule-based priority queue on answer candidates. The highest priority was given to answers where question and answer snippet contained the same predicate and for which the argument type of the answer matched the argument type of the question word of the question (e.g. ”what”). The next highest priority was given to answers which were somehow related to the matching predicate. Hereby, the argument types ”Arg0” and ”Arg1” have higher priority. For factoid questions, the top five answers were selected for the submission. For list questions, the maximum number of answers to be listed decreased depending on how low the priority levels got. This should ensure that we do not leave out a high-priority answer with a high probability to be correct in our model. Additionally, too many low-priority answers should be avoided to keep an acceptable precision level. Besides the SRL-based priority queue, we introduced a rule for the list question approach. As of the essence of list questions, we consider enumerations by detecting symbols like commas or the conjunction ”and”.

Summary questions and ideal answers. We also investigated SRL for the summarization task. For summary questions, out of the given sets of answer snippets, the system selects the ones with the largest semantic conformities. Similar to factoid and list questions, the semantic conformity is determined by the degree to which question and answer snippet contain similar predicate argument structures or vocabulary. The same, previously described priority queue is applied. The ideal answers for yes/no, factoid and list questions were retrieved by selecting the whole answer snippet that included the highest priority answer to the corresponding question. If no answer could be determined, we followed the same procedure as for summary questions.

4 Results

In this section we present an evaluation of Olelo based on two aspects: (a) an automatic evaluation of its QA components and the SRL extension approach on the fifth edition of the BioASQ challenge; and (b) its comparison to other on-line QA

Batch	System	Doc. retr.	Pass. retr.
1	Olelo	0.0465	0.0441
2	Olelo	0.0318	0.0246
3	Olelo	0.0658	0.0386
4	Olelo	0.0449	0.0347
5	Olelo	0.0381	0.0386
top results		[0.0874,0.1157]	[0.0467,0.0898]

Table 1: Results for mean average precision (MAP) for Olelo in BioASQ task 5b phase A, i.e., for document retrieval and passage retrieval. Range of top results in all batches are presented in the last row.

systems, in terms of processing time and quality of the answers.

4.1 BioASQ test sets

We participated in Task 5b (Biomedical Semantic QA) of the fifth edition of the BioASQ challenge. This task is composed of two phases: (a) Phase A, which includes submission of results for relevant concepts, documents, snippets and RDF triples. (b) Phase B, which includes submission of results for exact and ideal answers. A new batch of questions is released every two weeks and participants have 24 hours to submit results. For each batch of Phase A, the organizers release a JSON file which includes 100 questions and their corresponding type and identifier. After the end of phase A (24 hours), phase B starts after the release of an extended version of the JSON file which includes the gold standard concepts, documents, passages and RDF snippets, i.e., the answers for Phase A. Therefore, predictions for phase B can rely on this gold standard information, which we indeed used in some of our runs.

4.2 Evaluation on BioASQ task 5b

In this section we present results for both Olelo and the SRL approach. These are the official results that were made available and based on the official metrics that are described in the guidelines[2].

Table 1 presents the results for phase A based on mean average precision (MAP). For this phase, we provide results only for document and passage retrieval. We simply provide the top 10 documents and passages as returned by Olelo for each question, following the maximum number of documents and snippets which is specified in the BioASQ's guidelines.

Table 2 presents results of Olelo and the SRL approach for the exact answers of Phase B. We only provide results for yes/no questions using the SRL approach as this question type is not supported by Olelo. For the first batch, we had two submissions for SRL. SRL2 considers the detection of enumerations for list questions and fixes some minor bugs regarding the retrieval of ideal answers. The results in batch 1, SRL2 shows a significant improvement for list questions. Given that SRL2 was an improvement of SRL, we did not submit the latter from the second batch on.

For yes/no questions we did not measure any achievements in comparison to the approach of just saying ”yes” to any question. The training data from recent years was very ”yes”-biased and subsequently was our system. The results imply that this must have changed for the 4th and 5th batch.

The results for factoid questions based on SRL were constantly lower than the Olelo system, but they both reached a similar magnitude, which indicates a potential for a combination of both.

For list questions, the SRL approach achieved much higher F-Measure scores than Olelo. However, it should be noted that the Olelo QA system was performing its own passage retrieval and was not simply relying on the gold standard snippets provided by the challenge.

Finally, Table 3 presents our results for Olelo and the SRL approach for the ideal answers, i.e., custom summaries. These summaries should be provided for all questions, independent of their type. The difference between the Olelo and the Olelo-GS submissions is that the later relies on the gold standard (GS) snippets, instead of the ones retrieved by the system.

As expected, the Olelo-GS submissions usually obtained a higher score than the Olelo ones, but difference was lower than our expectations. The SRL-based approaches obtained much lower scores than Olelo runs. All Rouge metrics for the SRL approach were below 10%, which can be explained by the fact that it was basically just an answer snippet selection approach.

4.3 Comparison to other on-line QA applications

We compare the time response provided by our system to three other biomedical QA

[2] http://participants-area.bioasq.org/general_information/Task5b/

Batch	System	Yes/No	Factoid	List
1	Olelo	-	0.0400	0.0240
	Olelo-GS	-	0.0400	0.0477
	SRL	0.8824	-	0.0038
	SRL2	0.8824	-	0.1183
2	Olelo	-	0.0430	0.0281
	Olelo-GS	-	0.0323	0.0287
	SRL2	0.9630	0.0129	0.1123
3	Olelo	-	0.0192	0.0408
	Olelo-GS	-	0.0192	0.0549
	SRL2	0.8065	0.0128	0.1715
4	Olelo	-	0.0253	0.0513
	Olelo-GS	-	0.0513	0.0513
	SRL2	0.5517	0.0379	0.0943
5	Olelo	-	-	0.0202
	Olelo-GS	-	-	0.0379
	SRL2	0.4615	0.0286	0.2870
top results		[0.8387,0.9630]	[0.3606,0.5713]	[0.3358,0.5001]

Table 2: Results for Olelo and the SRL approach in the BioASQ task 5b phases B (exact answers). Results for yes/no questions are in terms of accuracy, MRR for factoid questions and f-measure for list questions. Range of top results in all batches are presented in the last row.

Batch	System	Rouge-2	Rouge-SU4
1	Olelo	0.2222	0.2710
	Olelo-GS	0.2958	0.3243
	SRL	0.0467	0.0510
	SRL2	0.0833	0.0870
2	Olelo	0.2751	0.2976
	Olelo-GS	0.2048	0.2500
	SRL2	0.0425	0.0418
3	Olelo	0.3426	0.3604
	Olelo-GS	0.2891	0.3262
	SRL2	0.0411	0.0416
4	Olelo	0.2261	0.2696
	Olelo-GS	0.3460	0.3516
	SRL2	0.0796	0.0740
5	Olelo	0.3418	0.3536
	Olelo-GS	0.2117	0.2626
	SRL2	0.0406	0.0413
top results		[0.5153,0.6891]	[0.5182,0.6789]

Table 3: Results for ideal answers (summaries) in terms of Rouge metrics for Olelo and the SRL approach. Range of top results in all batches are presented in the last row.

systems[3], namely AskHermes (Cao et al., 2011), EAGLi (Gobeill et al., 2015) and HONQA (Cruchet et al., 2009). However, we did not obtain any answer for none of the questions posed to HONQA, instead, only the following message: “A problem has occurred. Try later”.

We randomly selected ten factoid questions from the BioASQ dataset and posed these to the three systems - AskHermes, EAGLi and our application. 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 questions given that this type of answer is easier to check manually than summaries. Table 4 shows the list of questions.

We manually recorded the time response using a stopwatch. Time record started when clicking on the search button and stopped when any results was shown. All experiments were carried out from a laptop using the Chrome browser installed in the Ubuntu operating system. Further, it was carried out from home, i.e., not in the network of our institution, in order not to favor lower response times from Olelo. We manually and carefully checked the output provided by each system to look for the gold standard answer as provided by BioASQ. This ranged from short titles, as returned by EAGLi, short summaries returned by Olelo and even three long pages of text, as in the case of AskHermes. Table 5 summarizes the re-

[3] respectively, http://www.askhermes.org/; http://eagl.unige.ch/EAGLi/oldindex.htm; http://www.hon.ch/QA/

Number	Question
1	Which is the protein (antigen) targeted by anti-Vel antibodies in the Vel blood group?
2	Where in the cell do we find the protein Cep135?
3	Which enzyme is involved in the maintenance of DNA (cytosine-5-)-methylation?
4	Which is the most widely used model for the study of multiple sclerosis (MS)?
5	Which medication should be administered when managing patients with suspected acute opioid overdose?
6	What is the lay name of the treatment for CCSVI (chronic cerebro-spinal venous insufficiency) in multiple sclerosis?
7	What is the percentage of responders to tetrabenazine treatment for dystonia in children?
8	Intact macromolecular assemblies are analysed by advanced mass spectrometry. How large complexes (in molecular weight) have been studied?
9	Which is the most important prognosis sub-classification in Chronic Lymphocytic Leukemia?
10	What disease is mirtazapine predominantly used for?

Table 4: List of ten factoid questions that we considered for manual evaluation.

Systems	Output	Answers	Time
AskHermes	7/10	1/10	10.1 [2.09,19.74]
EAGLi	10/10	2/10	58.6 [21.41,107.72]
Olelo	10/10	4/10	8.84 [3.35,28.12]

Table 5: Results in terms of number of correct answers and response time for the on-line QA applications.

sults that we obtained. All output pages (or answers) returned by the systems are available for download[4].

5 Discussion

In this section we discuss our performance in the current edition of the BioASQ challenge and present an error analysis based on datasets from the previous years, given that gold standard results for this year's challenge are not yet available. We also provide a discussion on the comparison of Olelo to other on-line biomedical QA systems.

5.1 Olelo's performance in BioASQ task 5b

Although we have been participating in BioASQ in the last years, the development of our application did not have the challenge as goal. Thus, we still do not use any of past challenge datasets for training data. The system is not tuned to obtain best performance in BioASQ, except for the Olelo-GS submissions. As discussed above (cf. Section 3), Olelo does not distinguish between factoid and list questions, and we might have provided multiple results even for factoid questions.

The methods behind Olelo are constantly being improved. Currently, besides the approach based on SRL that we presented here, we also evaluated a new approach based on neural networks for the extraction of exact answers (Wiese et al., 2017), which obtained top results for factoid and list questions. We plan to integrate this new component into Olelo soon.

5.2 Error analysis based on previous data

In order to analyze the errors returned by our application, we carried out an evaluation on the test datasets of the two last editions (2015 and 2016) of the BioASQ challenge. We evaluated our exact answers using the BioASQ Oracle system, an on-line system that allows uploading JSON result files and obtaining evaluations at any time. We considered only the questions identified as ”factoid” and ”list” in the BioASQ dataset. We obtained a MAP that ranged from 0.0000 (no single match) to 0.0909 for factoid questions and a MAP from 0.0010 to 0.1000 for list questions.

This automatic evaluation is based solely on automatic matching procedures and results shown here are for the strict accuracy, i.e., an exact matching should apply. However, as described in our methods, our answers are derived from MeSH terms, while the gold standard answers in BioASQ are mostly based on the text spans as they appear in the document. For instance, for one of the questions, we returned the disease name ”Hirschsprung Disease”, while the gold standard consists of the text ”Aganlionic megacolon or Hirschsprung disease”. Indeed, during the BioASQ challenge, the organizers carry out a manual evaluation of all submitted answers, besides performing the automatic evaluation. Finally, our system does have some limitations on the exact answers that it is able to return, given its dependency to the MeSH terms. For instance, it performs particularly bad on questions which require gene/protein names in return, given that these entity types are poorly represented in MeSH. Indeed, almost 30% of the

[4] https://hpi.de//en/plattner/projects/in-memory-natural-language-processing/olelo.html

questions in BioASQ expect a gene/proteins in return, as pointed by (Neves and Kraus, 2016).

We manually checked our exact answers for all factoid and list questions. Unfortunately, the BioASQ Oracle system only returns a score for each batch of questions but does not give any information regarding true positives (TP), false negatives (FN) and false positives (FP). In our manual evaluation, we did not simply consider any overlap as a TP. For instance, we did not consider ”Receptors, Notch” and ”Notch intracellular domain (NICD)” as a match. However, we did record as TP those cases in which our answers were correct, e.g., ”Ethambutol” and ”Rifampin”, even though they did not match exactly the gold standard answer, which is the case of the following very long answer (sentence): ”Rifampin 10 mg/kg daily, ciprofloxacin 500 mg twice daily, clofazimine 100 mg every day, and ethambutol 15 mg/kg orally daily for 24 weeks, [...]”.

For a total of 502 factoid and list questions, our application was able to provide a total of 116 TPs, and at least one correct answer for a total of 71 questions. However, we missed many correct answers (FNs) as well as provided many false answers (FPs), sometimes even more than 20 FPs for a question.

Olelo did not return any results for many questions, and we believe that these might have been recognized as summary questions. As discussed above, our system still fails to return answers for concepts not properly covered by the MeSH ontology, but results are promising given the complexity of the task. More importantly, the manual evaluation shows that the user could receive at least one correct answers for 14% of the questions, while some answers could also have been found in the summary, for those questions for which only a summaries were returned.

5.3 Performance of semantic role labeling experiments

As of the date of the BioASQ submissions, our experiments on SRL were still in a preliminary phase. For the specific case of list questions, we could already show how a biomedical QA system could benefit from SRL. However, in general, we got the impression that SRL should not be used to design a QA system from scratch (as we tried in our experiments) but to improve our existing approaches. A major problem of our SRL approach was its coverage: if no matching labels for a question were found, we need an alternative approach to it. Otherwise, the recall will be too low, as experienced in our experiments. For list questions, considering enumerations as a baseline approach was very helpful.

For yes/no questions, more sophisticated detection strategies based on negation might be applied to find out when the answer is ”no” with higher precision. There might be further potential when analyzing occurrences of double negation or other sophisticated contextual information. A less ”yes”-biased training dataset in the BioASQ challenge could also produce further insights. At least having training data with more ”no”-samples might be desirable and allow more sophisticated approaches like machine learning. As stated before, the answer snippet selection strategy for the summarization task was not meant to be very promising. Nevertheless, the strategy could be combined with the current approach in the olelo system.

5.4 QA performance in a real-time scenario

Given the comparison of our systems to the other three available biomedical QA applications (cf. Section 4), we now present a discussion on the performance of the systems.

Olelo displayed high response times (19.85 seconds and 28.12 seconds) only for two questions, namely: ”Where in the cell do we find the protein Cep135?” and ”Intact macromolecular assemblies are analysed by advanced mass spectrometry. How large complexes (in molecular weight) have been studied?”. The second question is indeed longer than usual questions in BioASQ. Even though AskHermes outperformed Olelo in both minimum and maximum time, our application has on average a lower response time, besides being able to return an answer to all questions (cf. below). Further, three of the questions with response time under 10 seconds in AskHermes were those which returned no results, which suggests that the processing might have been interrupted. Finally, processing in EAGLi takes far too much time.

We manually analyzed the answers provided for the questions by each system. For all questions, Olelo returned a summary as answer, and in four of these questions, the summaries contained at least one of the correct answers for the question, as provided in the BioASQ benchmark. For instance, the following sentence is the first one in the summary that the system returned: "Cep135 is a 135-kDa, coiled-coil centrosome protein important for microtubule organization in mammalian cells." (PubMed 14983524). It contains the answer (centrosome) for the question "Where in the cell do we find the protein Cep135?".

In contrast, AskHermes extracted the correct answer only for one question. Nevertheless, the answer was indeed given as the top ranked. EAGLi could not provide exact answers for none of the questions, instead, only relevant documents (titles) and their corresponding single selected passages were presented. Two of these top passages indeed contained the correct answer to the question. Some of the snippets that contained the answer, as returned by AskHermes and EAGLi, appeared at the far end of a very long results page. However, these were too far from the top ranked answers (or passages) to be read by the average user, in our opinion. Finally, we should notice that EAGLi restricts the size of the question up to 80 characters, which could result in some questions not being properly processed by the system.

Even though Olelo was not able to detect that the questions were of the factoid type, and thus generated summaries for all questions, these summaries contain a maximum of five sentences (default value). Thus, we believe that most users could indeed find those four correct answers by reading through the short paragraphs. Changes on our question processing component could allow our system to output more short answers, instead of summaries, for questions that are in fact of the the factoid type. Currently, 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.

6 Conclusions

In this work, we presented an assessment of our Olelo QA system for biomedicine. We considered both the current online state of the system as well as a future extension based on semantic role labeling. We presented an evaluation both in terms of response time and robustness, in comparison to other online QA systems, as well as automatic and manual evaluation of the exact answers based on the BioASQ dataset. Our results are promising, given the complexity of the QA task, and future work will focus on the improvement of our current methods, integration of additional terminologies (e.g., for gene/proteins names) and support for additional question types (e.g., yes/no questions).

References

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

MichaelA Bauer and Daniel Berleant. 2012. Usability survey of biomedical question answering systems. *Human Genomics* 6(1):1–4.

Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont D. Antieau, Andrew S. Bennett, James J. Cimino, John W. Ely, and Hong Yu. 2011. Askhermes: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics* 44(2):277–288.

Sarah Cruchet, Arnaud Gaudinat, Thomas Rindflesch, and Celia Boyer. 2009. What about trust in the question answering world? In *Proceedings of the AMIA Annual Symposium*. San Francisco, USA, pages 1–5.

Daniel Dahlmeier and Hwee Tou Ng. 2010. Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics* 26(8):1098.

Julien Gobeill, Arnaud Gaudinat, Emilie Pasche, Dina Vishnyakova, Pascale Gaudet, Amos Bairoch, and Patrick Ruch. 2015. Deep question answering for protein annotation. *Database* 2015:bav081.

Daniel Jurafsky and James H. Martin. 2013. *Speech and Language Processing*. Prentice Hall International, 2 revised edition.

Milena Kraus, Julian Niedermeier, Marcel Jankrift, Sören Tietböhl, Toni Stachewicz, Hendrik Folkerts, Matthias Uflacker, and Mariana Neves. 2017. Olelo: a web application for intuitive exploration of biomedical literature. *Nucleic Acids Research gkx363* .

Mariana Neves. 2014. HPI in-memory-based database system in task 2b of bioasq. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*. pages 1337–1347.

Mariana Neves. 2015. HPI question answering system in the bioasq 2015 challenge. In *In Working Notes for CLEF 2015 Conference, Toulouse, France*.

Mariana Neves and Milena Kraus. 2016. BioMedLAT corpus: Annotation of the lexical answer type for biomedical questions. In *Open Knowledge Base and Question Answering Workshop at the 26th International Conference on Computational Linguistics (Coling)*.

Mariana Neves and Ulf Leser. 2015. Question answering for biology. *Methods* 74(0):36 – 46.

Frederik Schulze and Mariana Neves. 2016. Entity-supported summarization of biomedical abstracts. In *Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining at the 26th International Conference on Computational Linguistics (Coling)*.

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

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

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural question answering at bioasq 5b. In *In Biomedical Natural Language Processing (BioNLP) workshop at ACL*.