HPI in-memory-based database system in Task 2b of BioASQ

Mariana Neves[1]

[1]Hasso-Plattner-Institute at the University of Potsdam, Germany,
marianalaraneves@gmail.com

Abstract. We describe the participation of the Hasso-Plattner Institut (HPI) team in the BioASQ challenge and in particular in the Task 2b (Phase A), which consisted of providing results for potentially relevant concepts, documents and passages, given a certain question. Our systems relies on the in-memory based database (IMDB) technology and built-in text analysis provided by a IMDB database, as well as external resources, such as BioPortal and the pre-defined ontologies and terminologies to be used for the task. We present an evaluation of a preliminary version of our system on the training dataset (310 questions) and on the three of the test batches of 100 questions. Our results were particularly good for the passage retrieval, including a first position in one of the batches, which prove the feasibility of our approach for the question answering task.

Keywords: question answering, biomedicine, passage retrieval, document retrieval, concept extraction, in-memory database

1 Introduction

Question answering (QA) is the task of posing questions to search engines and receiving an exact answer in return [3]. It differs to information retrieval in two main aspects: (a) queries are presented as natural language questions (long input) instead of a set of keywords (short input); and (b) an exact and short answer (yes/no, a fact or a paragraph) is returned instead of a list of potential documents which might contain a answer. A variety of QA systems have been developed for the so-called open domain (e.g., START[1]) and the domain has received increasing attention from the scientific community recently since the IBM Watson system beat human participant in the Jeopardy TV show [2]. However, few systems currently exists for the biomedical domain and most previous approaches and datasets have focused on the medical domain [1].

The BioASQ challenge[2] is an EU-funded project which aims to foster research and solutions on the biomedical question answering area. A first challenge

[1] http://start.csail.mit.edu/index.php

[2] http://bioasq.org/

was run in 2013 as part of CLEF 2013 [5] and a new edition has been held in 2014 together with other QA and machine reading-related tasks in the CLEF Question Answering Task lab of CLEF 2014[3]. The BioASQ challenge consists of two main tasks: (2a) Large-Scale Online Biomedical Semantic Indexing and (2b) Biomedical Semantic QA, which focus on the QA itself and which is the focus of this article. The later is sub-divided in two phases: Phase A and Phase B.

In Phase A of task 2b, questions and their respective question type (yes/no, factoid, list or summary) were released and participants were requested to provide the following information:

- a list of relevant concepts belonging to five predefined ontologies and terminologies (GO, DO, MeSH, Jochem and Uniprot);
- a list of relevant articles from PubMed[4];
- a list of relevant snippets, including the PubMed document of origin, the start and end sections and offsets in the documents, and the text of the snippet;
- a list of relevant RDF triples.

In Phase B of task 2b, participants were provided with the questions and respective types released for Phase A as well as gold-standard information for Phase A, i.e, manually curated relevant concepts, documents, snippets and RDF triples. This time participants were requested to submit the following answers:

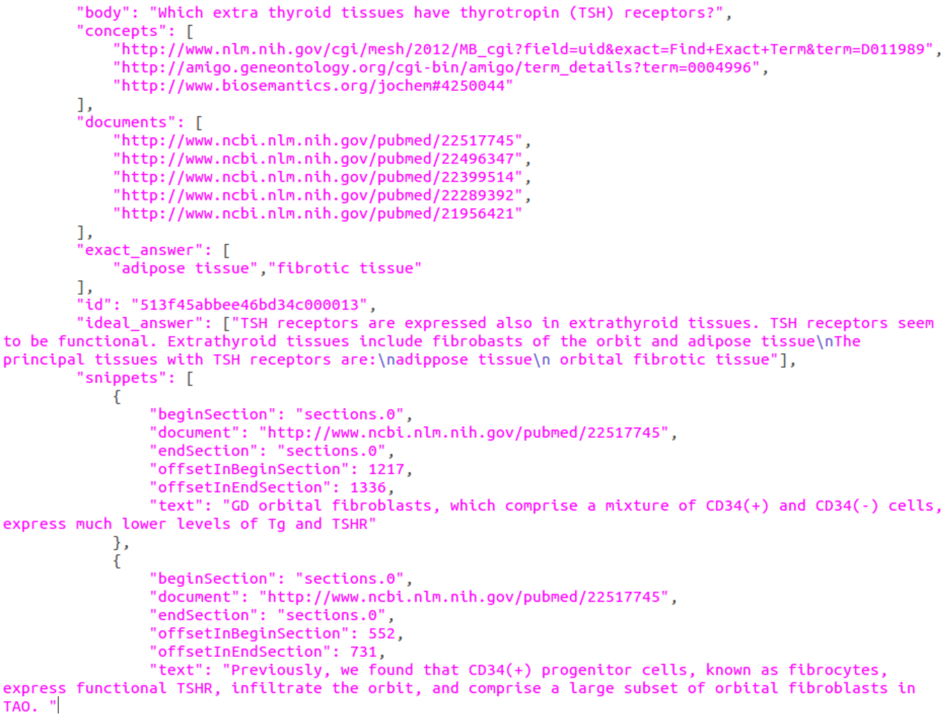
- an exact answer for the question: "Yes" or "No" for yes/no questions, and a single or a list of short answers for factoid and list questions, respectively.
- an ideal answer, which consists of a short paragraph for the summary questions as well as an extended answer to the yes/no, factoid and list questions.

A training dataset which includes of 310 questions and manually curated information for both Phases A and B above was released for the participants to allow training and/or evaluation during development of the system. And example of a question and the corresponding related information is shown in Figure 1.

We present the participation of the HPI team in Phase A of task 2b of the BioASQ challenge. We submitted results for the three above items required in Phase A: concepts, documents and snippets. Our system relies on the inmemory based technology (IMDB) [6] and on the built-in text analysis features provided by the SAP HANA database, such as sentences splitting, tokenization, dictionary-based named-entity recognition, full-text indexing and approximate string matching. An earlier version of this system has been recently applied for passage retrieval in multi-lingual question answering for three languages (English, German and Spanish) [4].

[3] `http://nlp.uned.es/clef-qa/`

[4] `http://www.ncbi.nlm.nih.gov/pubmed`



```
        "body": "Which extra thyroid tissues have thyrotropin (TSH) receptors?",
        "concepts": [
            "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D011989",
            "http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=0004996",
            "http://www.biosemantics.org/jochem#4250044"
        ],
        "documents": [
            "http://www.ncbi.nlm.nih.gov/pubmed/22517745",
            "http://www.ncbi.nlm.nih.gov/pubmed/22496347",
            "http://www.ncbi.nlm.nih.gov/pubmed/22399514",
            "http://www.ncbi.nlm.nih.gov/pubmed/22289392",
            "http://www.ncbi.nlm.nih.gov/pubmed/21956421"
        ],
        "exact_answer": [
            "adipose tissue","fibrotic tissue"
        ],
        "id": "513f45abbee46bd34c000013",
        "ideal_answer": ["TSH receptors are expressed also in extrathyroid tissues. TSH receptors seem
to be functional. Extrathyroid tissues include fibrobasts of the orbit and adipose tissue\nThe
principal tissues with TSH receptors are:\nadippose tissue\n orbital fibrotic tissue"],
        "snippets": [
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/22517745",
                "endSection": "sections.0",
                "offsetInBeginSection": 1217,
                "offsetInEndSection": 1336,
                "text": "GD orbital fibroblasts, which comprise a mixture of CD34(+) and CD34(-) cells,
express much lower levels of Tg and TSHR"
            },
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/22517745",
                "endSection": "sections.0",
                "offsetInBeginSection": 552,
                "offsetInEndSection": 731,
                "text": "Previously, we found that CD34(+) progenitor cells, known as fibrocytes,
express functional TSHR, infiltrate the orbit, and comprise a large subset of orbital fibroblasts in
TAO. "
```

Fig. 1. Extract of the JSON training data file. The factoid question "Which extra thyroid tissues have thyrotropin (TSH) receptors?" is shown followed by the manually curated information: the URIs for three concepts, the URLs for five PubMed documents, two exact answers (tissue names), the unique identifier of the question, an ideal answer which complement the exact answer and a list of relevant snippets (only two are shown) including details such as document, text and start and end section names and offsets.

The next section of this article describe details of the architecture of our system along with illustrative examples. Our results for both the training data and the test batches are presented in Section 3, followed by discussions on the performance of our system, error analysis and future improvements.

2 Architecture

We have developed a system which relies on the in-memory based technology and built-in text analysis provided by the SAP HANA database (hereafter called "HANA"). In this section we describe the architecture of our system and present each of its components in details.

Question answering systems usually include three main components [1]: question processing, passage retrieval and answer processing. Our architecture currently includes only the two first steps. A schema of the system is shown in Figure 2 and each of them its components are described in details below.

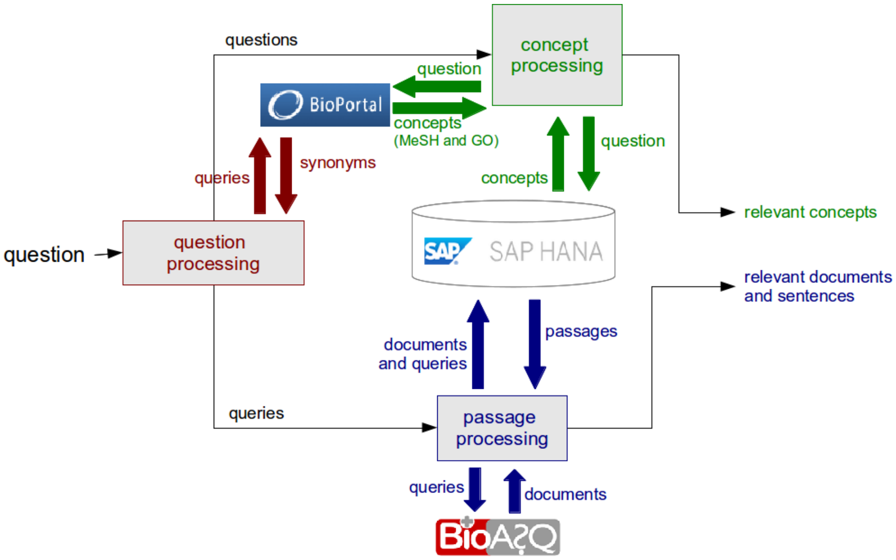


Fig. 2. Architecture of the system. The data work-flow for the three components are identified by distinct colors: "red" for question processing, "green" for concept recognition and "blue" for document and passage retrieval.

2.1 Question processing

The question processing component includes sentence splitting, tokenization, part-of-speech tagging and chunking using the Stanford CoreNLP package[5]. A query was generated from the original question based on both the tokens and the chunks. In this step, we ignored tokens which match any of the following: length less than 3, numerals (part-of-speech equals to "CD"), stopwords and symbols. For instance, for the question "Is Rheumatoid Arthritis more common in men or women?" (5118dd1305c10fae75000001) from the training dataset, six tokens (Rheumatoid, Arthritis, more, common, men, women) and four chunks (Rheumatoid Arthritis, more, common, men or women) are returned.

We performed query expansion based on the extracted tokens and by calling services from BioPortal[6]. We required exact match of the term to avoid potential wrong synonyms [7]. From the output returned by BioPortal, we considered all synonyms as well as definitions whose length are up to 20, which might also constitute potential synonyms. Synonyms were validated and we ignored those which contained any commas, parenthesis and other symbols, such as "Homo sapiens (living organism) [Ambiguous]" or "Human, Female". Examples of synonyms extracted for the tokens in the question above are "arthropathy" for "Arthritis", and "female", "femme" and "adult" for "women".

Weights were assigned for each term in the query according to the number of concepts which were returned by BioPortal, the higher the number of matched concepts, the lower the weight of the terms. Terms which did not match to any concept in BioPortal were assigned a weight of 0.5, i.e., an average weight. Otherwise, it was calculated based on the number of concepts which matched to this particular term (#MatchesToken) and the total number of concepts matched to all terms of the query (#MatchesTotal), according to the expression below:

$$weight = 1 - \frac{\#MatchesToken}{\#MatchesTotal} \tag{1}$$

2.2 Concepts retrieval

For each question, participants were required to return relevant concepts in five ontologies or terminologies: MeSH, GO, SwissProt, Jochem and DO. In our approach concepts were retrieved using two approaches: by matching previously compiled dictionaries to the text of the question using HANA and by making

[5] http://nlp.stanford.edu/software/corenlp.shtml

[6] http://data.bioontology.org/documentation0

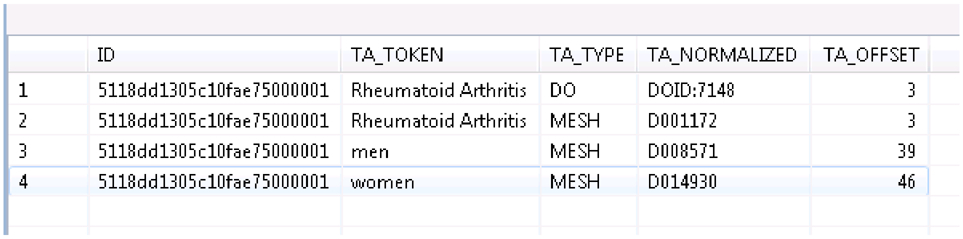
[7] e.g., http://data.bioontology.org/search?q=woman&apikey=7795d203-29ce-4f89-85aa-02c3555b21dd&exact_match=true

queries to the BioPortal web services.

The original ontologies and terminologies were retrieved from the respective web sites[8] and one dictionary was compiled for each of them. We considered the fields below for each of these resources:

- Jochem: lines identified with the codes "ID" (identifiers) and "TM" (terms);
- DO, MeSH and GO (OBO files): the fields "id" (identifiers), "name" (names) and "synonym" (synonyms);
- SwissProt: lines identified with the codes "ID" (identifiers), "DE" (description) and "GN" (gene names).

The dictionaries were converted to HANA database XML format, compiled and used for matching the terms in the text of the questions, which were previously loaded into a table in the database. We generated an index which includes the identifier of the document (question), the text span which was matched, the terminology/ontology, the respective identifier and the start offset with respect to the original text of the question, as shown in Figure 3. The terms recognized by HANA were retrieved from the database and we skipped those matches whose text length was less than 3 and which coincided with stopwords or Greek letters.



	ID	TA_TOKEN	TA_TYPE	TA_NORMALIZED	TA_OFFSET
1	5118dd1305c10fae75000001	Rheumatoid Arthritis	DO	DOID:7148	3
2	5118dd1305c10fae75000001	Rheumatoid Arthritis	MESH	D001172	3
3	5118dd1305c10fae75000001	men	MESH	D008571	39
4	5118dd1305c10fae75000001	women	MESH	D014930	46

Fig. 3. Concepts retrieved by the HANA database for the sentence "Is Rheumatoid Arthritis more common in men or women?" (5118dd1305c10fae75000001) from the training data.

Our second approach retrieved concepts of the MeSH and GO ontologies by making queries to the BioPortal Recommender[9]. Queries were created based on the full text of the question[10] and all returned concepts were considered.

[8] http://www.ncbi.nlm.nih.gov/mesh,http://www.geneontology.org/,http://www.uniprot.org/,http://www.biosemantics.org/,http://disease-ontology.org/

[9] http://data.bioontology.org/documentation

[10] e.g., http://data.bioontology.org/recommender?text=Is+Rheumatoid+Arthritis+more+common+in+men+or+women\%3F&ontologies=GO\%2CMESH&include_classes=true&apikey=7795d203-29ce-4f89-85aa-02c3555b21dd

The concepts retrieved from the HANA database and BioPortal were merged into a single list and votes were computed in regard to whether the concept was returned by one or both of them. The list was ranked by descending order of the number of votes.

2.3 Passages and documents retrieval

For each question, we perform four queries to the BioASQ PubMed service[11] according to whether considering query expansion or not (cf. "Question processing" above), and whether using "OR" or "AND" operators between the tokens. We retrieve up to the 500 top ranked documents and we only consider titles and abstracts. However, an analysis of the training dataset shows that these constitute about 90% of the relevant passages (5240 out of 5781 snippets).

The text of the titles and abstracts were inserted into the HANA database and a full text indexing was performed on them which include sentence splitting and tokenization. Queries were posed to the HANA database based on the terms of the query (including synonyms extracted during query expansion) and an approximate matching was performed by requiring at least 90% similarity. Sentences were ranked according to a score, which was calculated based on the similarity of the tokens, their weights in the query (cf. query processing above) and the total number of tokens which were matched. An example is shown in Table 1. For each question, passages (sentences) were retrieved only from those documents which were returned by the BioASQ services for it, although other documents (retrieved for other questions) are also included in out document collection. Sections were identified depending on whether the sentence came from the title or the abstract text and the offset are directly retrieved from the full text index generated by HANA. The top 100 sentences were retrieved for each question and the corresponding documents, i.e., usually less than 100, were returned as potential relevant documents.

3 Results

We performed experiments with the training/development dataset which contains 310 questions and their respective relevant concepts, documents and snippets. We calculated metrics of precision, recall and F-score for the concepts, documents and passages retrieval, but we did not considered their ranks in their respective lists. Concepts were evaluated based on their identifiers and documents based on the PubMed identifiers. Passages were evaluated based on the particular document and section they come from and we consider a true positive if there is an overlap of any length between the text of the gold standard and the one returned by our system. Results for the training dataset are presented in Table 2.

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

Table 1. Top 10 passages (sentences) which were retrieved for the question "Is Rheumatoid Arthritis more common in men or women?" (5118dd1305c10fae75000001).

PMID	Text of the passage
20685609	Rheumatoid arthritis (RA) is an autoimmune disease that is more common in women than in men.
21881200	Rheumatoid arthritis (RA) is an autoimmune disease that is more common in women than in men.
12192884	Rheumatoid arthritis (RA) is a chronic autoimmune disorder that, like most autoimmune diseases, is more common in women than in men.
10527397	Sex-specific linkage analysis may be of interest for rheumatoid arthritis on chromosome 3q since linkage of type 1 diabetes to IDDM9 derives predominantly from affected female sibpairs, and rheumatoid arthritis is more common in females than males.
6713801	Terminal phalangeal sclerosis was more common in females than in males and was more common in females with rheumatoid arthritis than in female controls.
7183585	Rheumatoid arthritis is three times more common in women and increasingly, over the last 40 years, women are working besides homemaking.
1616323	These results show that recurrent urinary tract infection is significantly more common in women with rheumatoid arthritis and secondary Sjögrens syndrome.
21977172	Rheumatoid arthritis (RA) is a systemic autoimmune disease whose main characteristic is persistent joint inflammation that results in joint damage and loss of function.Although RA is more common in females, extra-articular manifestations of the disease are more common in males.
19555469	INTRODUCTION: Rheumatoid arthritis (RA) is more common in females than males and sex steroid hormones may in part explain this difference.
7740304	Although RA is more common in women, rheumatoid lung disease occurs more frequently in men who have long-standing rheumatoid disease, positive rheumatoid factor and subcutaneous nodules.

Table 2. Results in terms of precision, recall and F-score for the training dataset.

Evaluation		Precision	Recall	F-Score
Concepts	HANA	0.28	0.18	0.22
	BioPortal	0.25	0.15	0.18
	HANA+BioPortal	0.26	0.21	0.23
Documents	w/o query expansion	0.022	0.11	0.037
	with query expansion	0.022	0.12	0.037
Passages	w/o query expansion	0.015	0.077	0.025
	with query expansion	0.015	0.078	0.025

The evaluation phase of the Phase A of task 2b consisted of five batches of questions which were released every 2/3 weeks. Participants had 24 hours to process the dataset, obtain the outputs for the corresponding information, build the JSON output file and submit it to the BioASQ web site.

Our system was under development while the BioASQ challenge was running and our submissions to the various batches of test questions varied accordingly. We did not submit runs for batches 1 and 5 and the only major change between the system used for batch 2 (HPI-S1) and batches 3 and 4 (HPI-S2) was that the first did not include synonyms for terms when performing queries to the BioASQ services for retrieving relevant documents. Predictions for concepts were only provided for batches 3 and 4. Table 3 presents the results for the three batches (as June 24th 2014) based on the metrics of mean precision, recall, f-measure and MAP which are described in details in the BioASQ guidelines[12].

Table 3. Results in terms of mean precision (P), recall (R), f-measure (FM) and MAP (as June 24th 2014), along with our position (Rank) in this batch with respect to the total number of runs. * indicates whether our position was ranked higher than the Top 100 and Top 50 baselines provided by the organizers. § indicates that no system outperformed any of the two baselines.

Documents	P	R	FM	MAP	Rank
Batch 2	0.0235	0.1341	0.0376	0.0733	10/18
Batch 3	0.0216	0.1773	0.0343	0.1016	11/19
Batch 4	0.0159	0.1399	0.0271	0.0558	10/18

Snippets	P	R	FM	MAP	Rank
Batch 2	0.0117	0.0746	0.0191	0.0521	1/10*
Batch 3	0.0126	0.0857	0.0195	0.0538	5/10*
Batch 4	0.0084	0.0882	0.0146	0.0339	6/12§

Concepts	P	R	FM	MAP	Rank
Batch 3	0.1134	0.1318	0.1034	0.0567	8/10§
Batch 4	0.1042	0.1080	0.0959	0.0522	8/8§

4 Discussion

Comparison of results between the training and test batches shows that the later have been only slight lower than what have been obtained in the training dataset, excepts for the concept retrieval whose results were far below. In general, except for the concept retrieval step, recall is much higher than the precision because we provided the top 100 snippets (and respective documents)

[12] http://bioasq.lip6.fr/Tasks/b/eval_meas/

for each question without specifying a minimum threshold score for that. Given that the MAP results were always higher than the precision for all document and snippets batches, we believe that we can indeed improve our precision (and consequently our F-Measure) by experimentally defining a minimum threshold in the passage retrieval step.

An error analysis for the concept extraction (training data) shows that some of the false positives could be considered as true positives and it is still not clear why they have not been included in the gold standard dataset. For instance, our system returned the concepts “D001172” (Arthritis, Rheumatoid) and “D014930” (Women) for the question “Is Rheumatoid Arthritis more common in men or women?” (5118dd1305c10fae75000001). On the other hand, we had concepts “D001171” (Arthritis, Juvenile Rheumatoid) and “D015535” (Arthritis, Psoriatic) returned as false negatives, which although related to the question, do not seem to be more relevant than the two false positives concepts shown above.

We had many false positives from Uniprot database due to the ambiguity of protein names, as we do not check which species name is being cited in the question, when any. For instance, we had the protein “P17276” (PH4H_DROME) from the Drosophila melanogaster returned for the question “Which are the most commonly reported pathological states associated with the formation of DNA G-quadruplexes?” (51600ab3298dcd4e51000036). The complexity of the gene/protein nomenclature complicates the retrieval of Uniprot concepts. For instance, from the phrase “prothymosin alpha c-terminal peptide” in the question “Describe the known functions for the prothymosin alpha c-terminal peptide?” (51be03c4047fa84d1d000004), we had false negatives for the identifier PTMA_HUMAN (synonym “Prothymosin alpha”) but false positives for matches to “C-terminal peptide” (e.g., PAHO_MOUSE). Finally, some questions cite gene/protein using synonyms which cannot be found, not even using an approximate matching approach, in the Uniprot database. For instance, in the question “Are there any DNMT3 proteins present in plants?” (511a16f9df1ebcce7d000005), the query “DNMT3” brings 12 hits in Uniprot, but none of them corresponds to the concepts assigned to this question in the gold standard (CMT1_ARATH, CMT2_ARATH, CMT3_ARATH) and a close look to these three entries does not clarify the reason for these associations.

Our results for document retrieval was tightly dependent on the query processing step, i.e., the conversion of the question to a appropriate query, to the performance of the BioASQ services, which was used for retrieving the documents, and to our passage retrieval step, from which the list of document was compiled. From a total of 3847 false positives in the training data, 2782 (72%) referred to documents which were not contained in our database and 368 (9.5%) to documents which were in the database but that were not directed linked to the question, i.e., the documents were retrieved by BioASQ for another question.

As future work on document and passage retrieval, we plan to query other services, such as the GeneView semantic browser [7] in order not to rely on only one service for document retrieval. Further, we also plan to have both the Medline collection of abstracts and the PubMed Central Open Access full texts indexed in our IMDB database and retrieve passages directly from them.

5 Conclusions

We have described our participation on the Task 2b (Phase A) of the BioASQ challenge for which we have developed a system based on in-memory database technology and that makes use of external resources, such as the ontologies and terminologies specified for the concept retrieval task, BioPortal web services for concept retrieval and query expansion and the BioASQ services for retrieval of potential relevant documents. We have obtained promising results for the passage retrieval task and we have described the future work which we plan to carry out to improve our system.

Acknowledgements MN would like to acknowledge funding from the HPI Research School.

References

1. Athenikos, S.J., Han, H.: Biomedical question answering: A survey. Computer Methods and Programs in Biomedicine 99(1), 1 – 24 (2010)
2. Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J.M., Schlaefer, N., Welty, C.A.: Building watson: An overview of the deepqa project. AI Magazine 31(3), 59–79 (2010)
3. Hirschman, L., Gaizauskas, R.: Natural language question answering: The view from here. Nat. Lang. Eng. 7(4), 275–300 (Dec 2001)
4. Neves, M., Herbst, K., Uflacker, M., Plattner, H.: Preliminary evaluation of passage retrieval in biomedical multilingual question answering. In: Proceedings of the Fourth Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2014) at Language Resources and Evaluation (LREC) 2014 (2014)
5. Partalas, I., Gaussier, E., Ngomo, A.C.N.: Results of the first bioasq workshop. In: 1st Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013) (nov 2013)
6. Plattner, H.: A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases. Springer, 1st edn. (2013)
7. Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., Leser, U.: Geneview: a comprehensive semantic search engine for pubmed. Nucleic Acids Research 40(W1), W585–W591 (2012)