HPI question answering system in the BioASQ 2015 challenge

Mariana Neves[1]

[1]Hasso-Plattner-Institute at the University of Potsdam, Germany,
mariana.neves@hpi.de

Abstract. I describe my participation on the 2015 edition of the BioASQ challenge in which I submitted results for the concept matching, document retrieval, passage retrieval, exact answer and ideal answer subtasks. My approach relies on a in-memory based database (IMDB) and its built-in text analysis features, as well as on PubMed for retrieving relevant citations, and on predefined ontologies and terminologies necessary for matching concepts to the questions. Although results are far below the ones obtained by other groups, I present an novel approach for answer extraction based on sentiment analysis.

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

1 Introduction

I describe my participation in the 2015 edition of the BioASQ challenge[1] [6] which took place in the scope of the CLEF initiative. This challenge aims to assess the current state of question answering systems and semantic indexing for biomedicine. The task 3b (Biomedical Semantic QA) is split in two phases and includes various sub-tasks such as concept mapping, information retrieval, question answering and text summarization. In phase A, participants receive a test set of questions along with their question type, i.e., “yes/no”, “factoid”, “list” or “summary”. Participants have 24 hours to submit predictions for relevant concepts, documents, passages and RDF triplets. When phase A is over, the organizers make available the test set for phase B containing the same questions of phase A, along with gold-standard annotations for concepts, documents, passages and RDF triples. This time, participants have 24 hours to submit predictions for the exact and ideal answers (short summaries). Exact answers are only required for “yes/no”, “factoid”, “list”, while ideal answers are expected to be returned for questions.

2 Architecture

I participated with a system developed on top of an in-memory database (IMDB) [5], the SAP HANA database, which is similar to the approach that I used during

[1] http://bioasq.org/

in the 2014 edition of the BioASQ challenge [2]. I participated in phases A and B of the task 3b of the 2015 edition of the BioASQ challenge and I have submitted predictions for potentially relevant concepts, documents, passages and answers.

Similar to previous QA systems [4], my system is composed of the following components: (a) question processing for construction of a query from the question; (b) concept mapping for performing concept recognition on the question; (c) document and passage retrieval for ranking and retrieval of relevant PubMed documents and passages; (d) answer extraction for building the short and long (summaries) answers. Figure 1 illustrates the architecture of the system and I describe the various steps in details below, including a short overview of the IMDB technology.

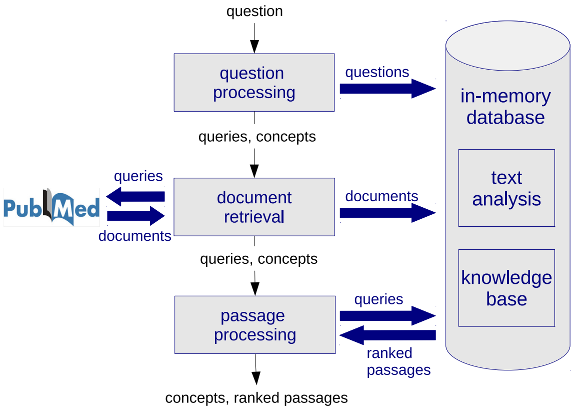


Fig. 1. Architecture of the system.

2.1 In-memory database

The SAP HANA database relies on IMDB technology [5] for fast access of data directly from main memory, in contrast to approaches which process data from files that reside on disk space and requires loading data into main memory. It also includes lightweight compression, i.e., a data storage representation that consumes less space than its original format, and built-in parallelization. The SAP HANA database comes with built-in text analysis which includes language detection, sentence splitting, tokenization, stemming, part-of-speech tagging, named-entity recognition based on pre-compiled dictionaries, information extraction based on manually crafted rules, document indexing, approximate searching and sentiment analysis.

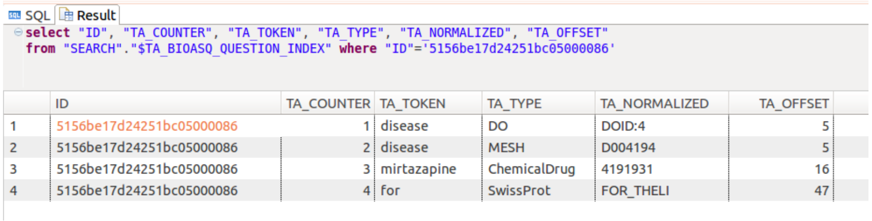
2.2 Question processing

In this step, the system processes the questions using the Standford CoreNLP [1] for sentence splitting, tokenization, part-of-speech tagging and chunking. The system constructed two queries for each question by selecting their more meaningful tokens. The first approach consists in removing all tokens which match a stopword list[2] and connecting them with the "OR", operator for more flexibility of the query. Both the document and passage retrieval steps as well as the answer extraction step made use of this high recall query for ranking documents and passages.

The second query aims on more precision and less recall and filters tokens further based on a list of the 5,000 most popular words of English[3] and uses the "AND" operator for connecting words. Only the document retrieval step used this high precision query for ranking relevant documents from PubMed. For instance, for the question "What disease is mirtazapine predominantly used for?", "disease OR mirtazapine OR predominantly OR used" is the resulting high recall query and "mirtazapine AND predominantly" is a higher precision query.

2.3 Concept mapping

The approach is the same that I used in the 2014 edition of the challenge [2]: I made use of the built-in named-entity recognition feature of the IMDB for mapping the questions to concepts from the five required terminologies and ontologies, which needed to be previously converted to dictionaries in an appropriate XML format. Given the dictionaries, the IMDB databases automatically matched terms to the words of the question, as illustrated in Figure 2.



```
select "ID", "TA_COUNTER", "TA_TOKEN", "TA_TYPE", "TA_NORMALIZED", "TA_OFFSET"
from "SEARCH"."$TA_BIOASQ_QUESTION_INDEX" where "ID"='5156be17d24251bc05000086'
```

	ID	TA_COUNTER	TA_TOKEN	TA_TYPE	TA_NORMALIZED	TA_OFFSET
1	5156be17d24251bc05000086	1	disease	DO	DOID:4	5
2	5156be17d24251bc05000086	2	disease	MESH	D004194	5
3	5156be17d24251bc05000086	3	mirtazapine	ChemicalDrug	4191931	16
4	5156be17d24251bc05000086	4	for	SwissProt	FOR_THELI	47

Fig. 2. Screen-shot of the entities recognized for the question "What disease is mirtazapine predominantly used for?".

[2] http://www.textfixer.com/resources/common-english-words.txt
[3] https://www.englishclub.com/vocabulary/common-words-5000.htm

2.4 Document and passage retrieval

The approach for retrieving relevant PubMed documents for each question is similar to the one described in my recently submitted paper [3]. It consisted in first posing the two generates queries to PubMed web services, retrieving up to 200 top ranked documents for each query and fetching the title and abstract for each PMID using the BioASQ web services. When querying PubMed, I restricted publication dates up to '2013/03/14' and I required citations to have an abstract available. This current approach differs from the one of my last year's participation [2] in terms that no I did not perform synonym expansion for the terms in the query, given the poor results obtained when relying on BioPortal for this purpose. Finally, titles and abstracts were inserted into a table in the IMDB.

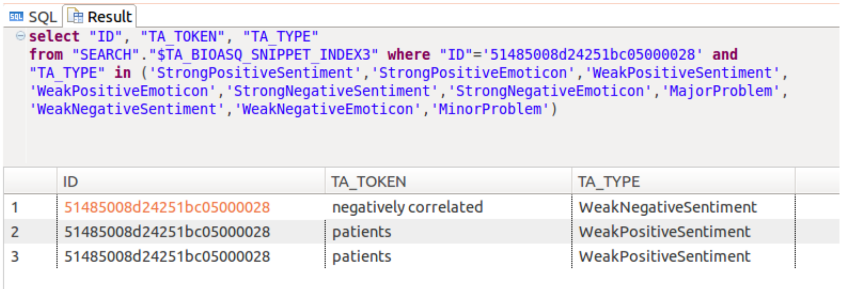
I retrieved passages using on the built-in information retrieval features available in the IMDB, which is based in approximated string similarity to match terms from the query to the words in the documents. The system proceeds ranks the passages (sentences) based on the TF-IDF metrics and I retrieve the top 10 sentences and corresponding documents as answers for the passage and document retrieval sub-tasks, respectively.

2.5 Answer extraction

I extracted both exact and ideal answers based on the gold-standard snippets that the organizers made available for phase B of task 3b. The process consisted in inserting the snippets into the IMDB database and I utilized built-in text analysis features for the extracting the answers, as described in details below for each question type.

Yes/No: Decision on either the answers “yes” or “no” was based on the sentiment analysis predictions provided by the IMDB. The assumption was that all snippets are somehow related to the question and that detection of sentiments in these passages could be used to distinguish between the two possible answers. Figure 3 shows the sentiments which were detected for a certain question.

The IMDB returns 10 types of sentiments, namely “StrongPositiveSentiment””, “StrongPositiveEmoticon”, “WeakPositiveSentiment”, “WeakPositiveEmoticon”, “StrongNegativeSentiment”, “StrongNegativeEmoticon”, “MajorProblem”, “WeakNegativeSentiment”, “WeakNegativeEmoticon” and “MinorProblem”. I merged some of these sentiment types into coarser categories according to simples rules (Table 1). The sentiments were first grouped into four coarse categories, i.e., “positiveStrong”, “positiveWeak”, “negativeStrong”, “negativeWeak”, and then into the three main sentiments “positive” or “negative”. For the rules shown in Table 1, I consider that the “positiveStrong” sentiment is stronger than the “negativeStrong” one, and therefore I assign the “positive” sentiment for such cases. Similarly, I consider “positiveWeak” weaker than “negativeWeak” when both are returned for the same question. Cases which did not match none of the rules for “positive” or “negative” sentiments are classified as “neutral”. Final



```sql
select "ID", "TA_TOKEN", "TA_TYPE"
from "SEARCH"."$TA_BIOASQ_SNIPPET_INDEX3" where "ID"='51485008d24251bc05000028' and
"TA_TYPE" in ('StrongPositiveSentiment','StrongPositiveEmoticon','WeakPositiveSentiment',
'WeakPositiveEmoticon','StrongNegativeSentiment','StrongNegativeEmoticon','MajorProblem',
'WeakNegativeSentiment','WeakNegativeEmoticon','MinorProblem')
```

	ID	TA_TOKEN	TA_TYPE
1	51485008d24251bc05000028	negatively correlated	WeakNegativeSentiment
2	51485008d24251bc05000028	patients	WeakPositiveSentiment
3	51485008d24251bc05000028	patients	WeakPositiveSentiment

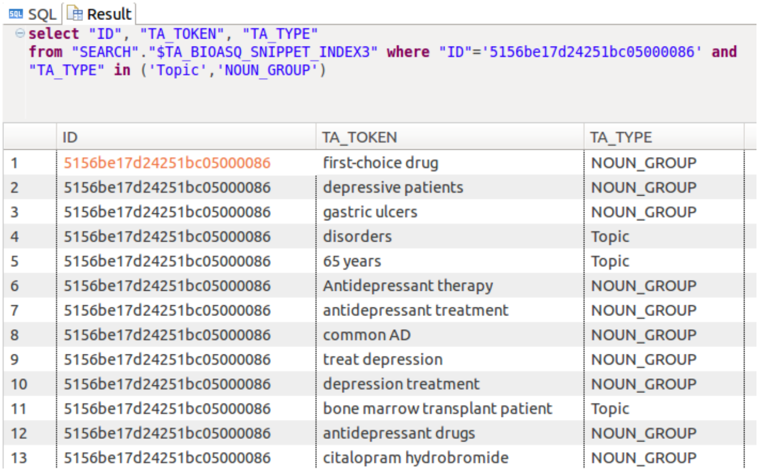
Fig. 3. Screen-shot of the sentiments detected from the gold-standard snippets for the question "Is miR-126 involved in heart failure?".

decision for the the answers "yes" or "no" was based on these three coarse sentiments. By default, I return the answer "no", unless I get "positive" or "neutral' as output from the above rules.

Table 1. Rules for merging fine-grained sentiments into coarser sentiments.

coarse sentiment	Rule
positiveStrong	StrongPositiveSentiment OR StrongPositiveEmoticon
positiveWeak	WeakPositiveSentiment OR WeakPositiveEmoticon
negativeStrong	StrongNegativeSentiment OR StrongNegativeEmoticon OR MajorProblem
negativeWeak	WeakNegativeSentiment OR WeakNegativeEmoticon OR MinorProblem
positive	(positiveStrong OR positiveWeak) AND (NOT(negativeStrong) AND NOT(negativeWeak))
positive	positiveStrong AND negativeStrong
positive	positiveStrong AND negativeWeak
negative	(negativeStrong OR negativeWeak) AND (NOT(positiveStrong) AND NOT(positiveWeak))
negative	positiveWeak AND negativeStrong
negative	positiveWeak AND negativeWeak

Factoid and list: I extracted factoid and list answers based also on built-in predictions provided by our IMDB, more specifically, on the annotations of noun phrases and topics, as presented in Figure 4. Given that no semantic processing was performed neither for the question nor for the snippets, in oder to tag named entities and to identify the entity type of expected answer, I choose the five top answers based on the order returned by the IMDB.



```
select "ID", "TA_TOKEN", "TA_TYPE"
from "SEARCH"."$TA_BIOASQ_SNIPPET_INDEX3" where "ID"='5156be17d24251bc05000086' and
"TA_TYPE" in ('Topic','NOUN_GROUP')
```

	ID	TA_TOKEN	TA_TYPE
1	5156be17d24251bc05000086	first-choice drug	NOUN_GROUP
2	5156be17d24251bc05000086	depressive patients	NOUN_GROUP
3	5156be17d24251bc05000086	gastric ulcers	NOUN_GROUP
4	5156be17d24251bc05000086	disorders	Topic
5	5156be17d24251bc05000086	65 years	Topic
6	5156be17d24251bc05000086	Antidepressant therapy	NOUN_GROUP
7	5156be17d24251bc05000086	antidepressant treatment	NOUN_GROUP
8	5156be17d24251bc05000086	common AD	NOUN_GROUP
9	5156be17d24251bc05000086	treat depression	NOUN_GROUP
10	5156be17d24251bc05000086	depression treatment	NOUN_GROUP
11	5156be17d24251bc05000086	bone marrow transplant patient	Topic
12	5156be17d24251bc05000086	antidepressant drugs	NOUN_GROUP
13	5156be17d24251bc05000086	citalopram hydrobromide	NOUN_GROUP

Fig. 4. Screen-shot of the noun phrases and topics detected from the relevant snippets for the question "What disease is mirtazapine predominantly used for?".

Summary: I also built summaries for the ideal answers based on the phrases which contain sentiments, as shown in Table 6. The assumption was that such phrases are more informative and relevant than the ones in which no sentiments were found. My approach consisted in concatenating the sentences up to a limit of 200 words, as specified in the challenge's guidelines.

3 Results and discussion

I submitted results for all five batches of test questions for task 3b: (a) phase A, i.e., concept mapping and document and passage retrieval, and (b) phase B, i.e., exact and ideal answers. Different from previous editions of the BioASQ challenge, when participants were allowed to submit up to 100 entries per question for each of the required sub-tasks, whether documents, concepts or exact answers, this year's edition limited concepts, documents and passages up to 10 per question and factoid answers up to 5. I present below the results I obtained as published by the organizers in the BioASQ Web site [4]. I do not show results for concept matching because the organizers seem not to have made them available yet.

Table 3 shows my results for document retrieval for each of the five test batches. As discussed in the methods section, I did not implement any specific approach for this task and documents were ranked based on the relevancy of the query to the passages and not to the documents (abstracts) themselves. In

[4] http://participants-area.bioasq.org/results/3b/phaseA/;http://participants-area.bioasq.org/results/3b/phaseB/

Table 2. Phases related to sentiments found in the gold-standard snippets of the question "What disease is mirtazapine predominantly used for?".

second-generation antidepressants (selective serotonin reuptake inhibitors, nefazodone, venlafaxine, and mirtazapine) in participants younger than 19 years with MDD, OCD, or non-OCD anxiety disorders.
patients 65 years or older with major depression.
A case report involving linezolid with citalopram and mirtazepine in the precipitation of serotonin syndrome in a critically ill bone marrow transplant patient is described in this article.
In 26 patients with FMS who completed a 6-week open study with mirtazapine, 10 (38%) responded with a reduction of at least 40% of the initial levels of pain, fatigue and sleep disturbances (Samborski et al 2004).
In general, drugs lacking strong cholinergic activity should be preferred.
Drugs blocking serotonin 5-HT2A or 5-HT2C receptors should be preferred over those whose sedative property is caused by histamine receptor blockade only.

this year's edition of the challenge, organizers required participants to submit up to 10 document, which is a hard assignment, given the millions of citations in PubMed. Indeed, results have been lower than the ones obtained by participants last year and it is unclear whether we (teams) performed better than the baseline systems as the organizers did not publish results for these systems yet.

Table 3. Results for document retrieval for the test set. The "Rank" column shows the position obtained by my system in relation to the total number of submissions.

test batch	Mean precision	Recall	F-measure	MAP	Rank
batch 1	0.1027	0.1250	0.0841	0.0464	17/18
batch 2	0.1164	0.1363	0.1009	0.0658	17/20
batch 3	0.1082	0.1139	0.0950	0.0634	17/20
batch 4	0.1354	0.1849	0.1283	0.0737	18/21
batch 5	0.1465	0.2810	0.1690	0.0700	16/19

Table 4 shows my results for passage retrieval for each of the five test batches. Few groups participated in this task, in comparison to the number of submissions for the document retrieval task. A task which is already very complex has been made even more difficult this year by the limitation of providing up to only 10 top passages.

Finally, tables 5 and 6 shows the results I obtained for the exact and ideal answers in phase B of task 3b.

Table 4. Results for passage retrieval for the test set. The "Rank" column shows the position obtained by my system in relation to the total number of submissions.

test batch	Mean precision	Recall	F-measure	MAP	Rank
batch 1	0.0545	0.0686	0.0501	0.0347	6/6
batch 2	0.0580	0.0493	0.0437	0.0355	7/7
batch 3	0.0542	0.0396	0.0391	0.0452	7/7
batch 4	0.0881	0.0981	0.0807	0.0624	8/8
batch 5	0.0859	0.1189	0.0883	0.0572	6/6

Table 5. Results for exact answers for the test set. The "Rank" column shows the position obtained by my system in relation to the total number of submissions.

test batch	Yes/No	Factoid			List			Rank
	Accuracy	Strict Acc.	Lenient Acc.	MRR	Mean precision	Recall	F-measure	
batch 1	0.6667	-	-	-	0.0292	0.0603	0.0364	7/9
batch 2	0.5625	-	-	-	0.0714	0.0161	0.0262	10/12
batch 3	0.6207	-	-	-	-	-	-	7/14
batch 4	0.5600	0.0345	0.0345	0.0345	0.1522	0.0473	0.0689	10/12
batch 5	0.3571	0.0909	0.0909	0.0909	0.0625	0.0292	0.0397	14/14

Table 6. Results for ideal answers for the test set. The "Rank" column shows the position obtained by my system in relation to the total number of submissions.

test batch	Rouge-2	Rouge-SU4	Rank
batch 1	0.1884	0.2008	15/15
batch 2	0.2026	0.2227	18/18
batch 3	0.1934	0.2189	17/17
batch 4	0.2504	0.2724	16/18
batch 5	0.1694	0.1790	18/18

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

References

1. Stanford core nlp, `http://nlp.stanford.edu/software/corenlp.shtml`
2. Neves, M.: HPI in-memory-based database system in task 2b of bioasq. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 1337–1347 (2014)
3. Neves, M.: In-memory database for passage retrieval in biomedical question answering. Journal Of Biomedical Semantics (submitted) (2015)
4. Neves, M., Leser, U.: Question answering for biology. Methods 74(0), 36 – 46 (2015), `http://www.sciencedirect.com/science/article/pii/S1046202314003491`
5. Plattner, H.: A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases. Springer, 1st edn. (2013)
6. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., et al.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics 16(1), 138 (2015)