Results of the BioASQ Track of the Question Answering Lab at CLEF 2014

George Balikas, Ioannis Partalas, Axel-Cyrille Ngonga Ngomo, Anastasia Krithara, Eric Gaussier, and George Paliouras

Abstract. The goal of this task is to push the research frontier towards hybrid information systems. We aim to promote systems and approaches that are able to deal with the whole diversity of the Web, especially for, but not restricted to, the context of bio-medicine. This goal is pursued by the organization of challenges. The second challenge consisted of two tasks: semantic indexing and question answering. 61 systems participated by 18 different participating teams for the semantic indexing task, of which between 25 and 45 participated in each batch. The semantic indexing task was tackled by 22 systems, which were developed by 8 different organizations. Between 15 and 19 of these systems addressed each batch. The question answering task was tackled by 18 different systems, developed by 7 different organizations. Between 9 and 15 of these systems submitted results in each batch. Overall, the best systems were able to outperform the strong baselines provided by the organizers.

1 Introduction

The aim of this paper is twofold. First, we aim to give an overview of the data issued during the BioASQ track of the Question Answering Lab at CLEF 2014. In addition, we aim to present the systems that participated in the challenge and for which we received system descriptions. In particular, we aim to evaluate their performance w.r.t. to dedicated baseline systems. To achieve these goals, we begin by giving a brief overview of the tasks included in the track, including the timing of the different tasks and the challenge data. Thereafter, we give an overview of the systems which participated in the challenge and provided us with an overview of the technologies they relied upon. Detailed descriptions of some of the systems are given in lab proceedings. The evaluation of the systems, which was carried out by using state-of-the-art measures or manual assessment, is the last focal point of this paper. The conclusion sums up the results of the track.

2 Overview of the Tasks

The challenge comprised two tasks: (1) a large-scale semantic indexing task (Task 2a) and (2) a question answering task (Task 2b).

Large-scale semantic indexing. In Task 2a the goal is to classify documents from the PubMed[1] digital library unto concepts of the MeSH[2] hierarchy. Here, new PubMed articles that are not yet annotated are collected on a weekly basis. These articles are used as test sets for the evaluation of the participating systems. As soon as the annotations are available from the PubMed curators, the performance of each system is calculated by using standard information retrieval measures as well as hierarchical ones. The winners of each batch were decided based on their performance in the Micro F-measure (MiF) from the family of flat measures [23], and the Lowest Common Ancestor F-measure (LCA-F) from the family of hierarchical measures [9]. For completeness several other flat and hierarchical measures were reported [3]. In order to provide an on-line and large-scale scenario, the task was divided into three independent batches. In each batch 5 test sets of biomedical articles were released consecutively. Each of these test sets were released in a weekly basis and the participants had 21 hours to provide their answers. Figure 1 gives an overview of the time plan of Task 2a.

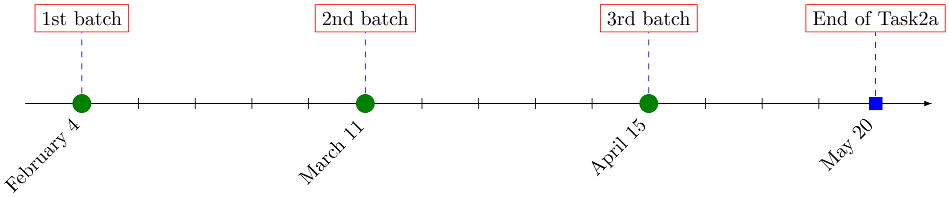

Fig. 1. The time plan of Task 2a.

Biomedical semantic QA. The goal of task 2b was to provide a large-scale question answering challenge where the systems should be able to cope with all the stages of a question answering task, including the retrieval of relevant concepts and articles, as well as the provision of natural-language answers. Task 2b comprised two phases: In phase A, BioASQ released questions in English from benchmark datasets created by a group of biomedical experts. There were four types of questions: "yes/no" questions, "factoid" questions,"list" questions and "summary" questions [3]. Participants had to respond with relevant concepts (from specific terminologies and ontologies), relevant articles (PubMed and PubMedCentral[3] articles), relevant snippets extracted from the relevant articles and relevant RDF triples (from specific ontologies). In phase B, the released questions contained the correct answers for the required elements (concepts, articles, snippets and RDF triples) of the first phase. The participants had to answer with *exact* answers as well as with paragraph-sized summaries in natural language (dubbed *ideal* answers).

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

[2] `http://www.ncbi.nlm.nih.gov/mesh/`

[3] `http://www.ncbi.nlm.nih.gov/pmc/`

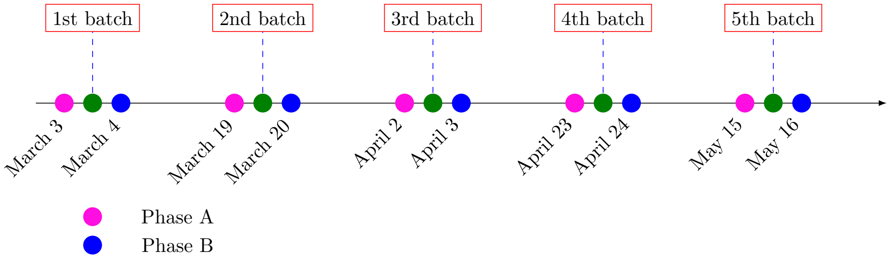

Fig. 2. The time plan of Task 2b. The two phases for each batch run in consecutive days.

The task was split into five independent batches. The two phases for each batch were run with a time gap of 24 hours. For each phase, the participants had 24 hours to submit their answers. We used well-known measures such as mean precision, mean recall, mean F-measure, mean average precision (MAP) and geometric MAP (GMAP) to evaluate the performance of the participants in Phase A. The winners were selected based on MAP. The evaluation in phase B was carried out manually by biomedical experts on the ideal answers provided by the systems. For the sake of completeness, ROUGE [11] is also reported.

3 Overview of Participants

3.1 Task 2a

The participating systems in the semantic indexing task of the BioASQ challenge adopted a variety of approaches including hierarchical and flat algorithms as well as search-based approaches that relied on information retrieval techniques. In the rest of section we describe the proposed systems and stress their key characteristics.

The new NCBI system [26] for Task 2a is an extension of the work presented in 2013 and relies on the generic learning-to-rank approach presented in [7]. This novel approach, dubbed LAMBDA-MART, differs from the previous approach in the following aspects: First, the set of features has been extended to include binary classifier results. In addition, the set of documents used as neighbor documents was reduced to documents indexed after 2009. Moreover, the score function for the selection of the number of features was changed from a linear to a logarithmic approach. Overall, the novel approach achieves an F-measure between 0 (RDF triples) and 0.38 (concepts).

In [18] flat classification processes were employed for the semantic indexing task. In particular, the authors trained binary SVM classifiers for each label that was present in the data. In order to reduce the complexity they trained the SVMs in fractions of the data. They trained two systems on different corpus: Asclepios on 950 thousand documents and Hippocrates on 1.5 million. Those systems output a ranked lists with labels and a meta-model, namely MetaLabeler [22], is used to decide the number of labels that will be submitted for each document. The remaining three systems of the team employ ensemble learning methods. The approach that worked best was a combination of Hippocrates with a model of simple binary SVMs, which were trained by changing the weights parameter for positive instances [10]. During the training of a classifier with very few positive instances they can chose to penalize a false negative (a positive instance being misclassified) more than a false positive (a negative instance being mis-classified). The proposed approaches, although they are relatively simple, require a lot of processing power and memory. For that reason they used a machine with 40 processors and 1TB RAM.

Ribadas et al. [20] employ hierarchical models based on a top-down hierarchical classification scheme [21] and a Bayesian network which models the hierarchical relations among the labels as well as the training data. The team participated in the first edition of the BioASQ challenge using the same technologies [19]. In the current competition they focused on the pre-processing of the textual data while keeping the same classification models. More specifically, the authors employ techniques for identifying abbreviations in the text and expanding it afterwards in order to enrich the document. Also, a part of speech tagger is used in order to tokenize the text and identify noun, verbs, adjectives and unknown elements (not identified). Finally, a lemmatization step extracts the canonical forms of those words. Additionally, the authors extract word bigrams and keep only those that are identified as multiword terms. The rational is that multiword terms in a domain with complex terminology, like biomedicine, provide higher discriminant power.

In [5] the authors use a standard flat classification scheme, where a SVM is trained for each class label in MeSH. Different training set methodologies are used resulting in different trained classifiers. Due to computational issues only 50,000 documents were used for training. The selection of the best classification scheme is optimized on the precision at top k labels on a validation set.

In [13] the authors used the learning to rank (LTR) method for predicting MeSH headings. However, in addition to the information from similar citations, they also used the prediction scores from individual MeSH classifiers to improve the prediction accuracy. In particular, they trained a binary classifier (logistic regression) for each label (MeSH heading). For a target citation, using the trained classifiers, they calculated the annotation probability (score) of every MeSH heading. Then, using NCBI efetch[4],they retrieved similar citations for the neighbor scores. Finally, these two scores, together with the default results of NLM official solution MTI, were considered as features in the LTR framework. The LambdaMART [4] was used as the ranking method in the learning to rank framework.

In [1], they proposed a system which uses Latent Semantic Analysis to identify semantically similar documents in MEDLINE and then constructs a list of MeSH headers from candidates selected from the documents most similar to a new abstract.

[4] http://www.ncbi.nlm.nih.gov/books/NBK25499/

Table 1 resumes the principal technologies that were employed by the participating systems and whether a hierarchical or a flat approach has been followed.

Table 1. Technologies used by participants in Task 2a.

Reference	Approach	Technologies
[18]	flat	SVMs, MetaLabeler [22]
[18]	flat	Ensemble Learning, SVMs
[19]	hierarchical	SVMs, Bayes networks
[27]	flat	MetaMap [2], information retrieval, search engines
[14]	flat	k-NN, SVMs
[15]	flat	k-NN, learning-to-rank
[13]	flat	Logistic regression, learning-to-rank
[1]	flat	LSA
[26]	flat	Learning-to-rank

Baselines. During the first challenge two systems were served as baseline systems. The first one, dubbed BioASQ _Baseline, follows an unsupervised approach to tackle the problem and so it is expected that the systems developed by the participants will outperform it. The second baseline is a state-of-the-art method called Medical Text Indexer [8] which is developed by the National Library of Medicine[5] and serves as a classification system for articles of MEDLINE. MTI is used by curators in order to assist them in the annotation process. The new annotator is an extension of the system presented in [16] with the approaches of the last year's winner [24]. Consequently, we expected the baseline to difficult to beat.

3.2 Task 2b

As mentioned above, the second task of the challenge is split into two phases. In the first phase, where the goal is to annotate questions with relevant concepts, documents, snippets and RDF triples 8 teams with 22 systems participated. In the second phase, where team are requested to submit exact and paragraph-sized answers for the questions, 7 teams with 18 different systems participated.

The system presented in [17] relies on the Hana Database for text processing. It uses the Stanford CoreNLP package for tokenizing the questions. Each of the token is then sent to the BioPortal and to the Hana database for concept retrieval. The concepts retrieved from the two stores are finally merged to a single list that is used to retrieve relevant text passages from the documents at hand. To this end, four different types of queries are sent to the BioASQ services. Overall, the approach achieves between 0.18 and 0.23 F-measure.

The approach proposed by NCBI [26] for Task 2b can be used in combination with the approach by the same group for Task 2a. In phase A, NCBI's framework used the cosine similarity between question and sentence to compute their

[5] http://ii.nlm.nih.gov/MTI/index.shtml

similarity. The best scoring sentence from an abstract was chosen as relevant snippet for an answer. Concept recognition was achieved by a customized dictionary lookup algorithm in combination with MetaMap. For phase B, tailored approaches were used depending on the question types. For example, a manual set of rules was crafted to determine the answers to factoid and list questions based on the benchmark data for 2013. The system achieved an F-measure of up to betwen 0.2% (RDf triples) and 38.48% (concepts). It performed very well on Yes/No questions (up to 100% accuracy). Factoid and list questions led to an MRR of up to 20.57%.

In [5] the authors participated only in the document retrieval of phase A and in the generation of ideal answers in phase B. The Indri search engine is used to index the PubMed articles and different models are used to retrieve documents like pseudo-relevance feedback, sequential dependence model and semantic concept-enriched dependence model where the recognised UMLS concepts in the query are used as additional dependence features for ranking documents. For the generation of ideal answers the authors retrieve sentences from documents and identify the common keywords. Then the sentences are ranked according to the number of times these keywords appear in each of them and finally the top ranked m are used to form the ideal answer.

The authors of [12] propose a method for the retrieval of relevant documents and snippets of task 2b. They develop a figure-inspired text retrieval method as a way of retrieving documents and text passages from biomedical publications. The method is based on the insight that for biomedical publications, the figures play an important role to the point that the captions can be used to provide abstract like summaries. The proposed approach uses an Information Retrieval perspective on the problem. In principle, the followed steps are: (i) the question in enriched by query expansion with information from UMLS, Wikipedia, and Figures, (ii) a ranking of full documents and snippets is retrieved from a corpus of PubMed Central Articles which is the set of full-text available articles, (iii) features are extracted for each document and snippet that provide proof of its relevance for the question and (iv) the documents/snippets are re-ranked with a learning-to-rank approach.

In the context of phase B of task 2b in [18], the authors attempted to replicate the work that already exists in literature and was presented in the BioASQ 2013 workshop [25]. They provided exact answers only for the factoid questions. Their system tries to extract the lexical answer type by manipulating the words of the question. Then, the relevant snippets of the question which are provided as inputs for this tasks are processed with the 2013 release of MetaMap [2] in order to extract candidate answers.

Baselines. Two baselines were used in phase A. The systems return the list of the top-50 and the top-100 entities respectively that may be retrieved using the keywords of the input question as a query to the BioASQ services. As a result, two lists for each of the main entities (concepts, documents, snippets, triples) are produced, of a maximum length of 50 and 100 items respectively.

For the creation of a baseline approach in Task 2B Phase B, three approaches were created that address respectively the answering of factoid and lists questions, summary questions, and yes/no questions [25]. The three approaches were combined into one system, and they constitute the BioASQ baseline for this phase of Task 2B. The baseline approach for the list/factoid questions utilizes and ensembles a set of scoring schemes that attempt to prioritize the concepts that answer the question by assuming that the type of the answer aligns with the lexical answer type (type coercion). The baseline approach for the summary questions introduces a multi-document summarization method using Integer Linear Programming and Support Vector Regression.

4 Results

4.1 Task 2a

During the evaluation phase of the Task 2a, the participants submitted their results on a weekly basis to the online evaluation platform of the challenge[6]. The evaluation period was divided into three batches containing 5 test sets each. 18 teams were participated in the task with a total of 61 systems. 12,628,968 articles with 26,831 labels (20.31GB) were provided as training data to the participants. Table 2 shows the number of articles in each test set of each batch of the challenge.

Table 2. Statistics on the test datasets of Task 2a.

Batch	Articles	Annotated Articles	Labels per article
1	4,440	3,263	13.20
	4,721	3,716	13.13
	4,802	3,783	13.32
	3,579	2,341	13.02
	5,299	3,619	13.07
Subtotal	23,321	16,722	13.15
2	4,085	3,322	13.05
	3,496	2,752	12.28
	4,524	3,265	12.90
	5,407	3,848	13.23
	5,454	3,642	13.58
Subtotal	22,966	16,829	13.01
3	4,342	2,996	12.71
	8,840	5,783	13.37
	3,702	2,737	13.32
	4,726	3,225	13.90
	4,533	3,196	12.70
Subtotal	26,143	17,929	13.20
Total	72,430	51,480	13.12

[6] `http://bioasq.lip6.fr`

Table 3. Correspondence of reference and submitted systems for Task 2a.

Reference	**Systems**
[18]	Asclepius, Hippocrates, Sisyphus
[20]	cole_hce1, cole_hce2, cole_hce_ne, utai_rebayct, utai_rebayct_2
[5]	SNUMedInfo*
[13]	Antinomyra-*
[26]	L2R*
Baselines	MTIFL, MTI-Default, bioasq_baseline

Table 3 presents the correspondence of the systems for which a description was available and the submitted systems in Task 2a. The systems MTIFL, MTI-Default and BioASQ _Baseline were the baseline systems used throughout the challenge. MTIFL and MTI-Default refer to the NLM Medical Text Indexer system [16]. Systems that participated in less than 4 test sets in each batch are not reported in the results[7].

According to [6] the appropriate way to compare multiple classification systems over multiple datasets is based on their average rank across all the datasets. On each dataset the system with the best performance gets rank 1.0, the second best rank 2.0 and so on. In case that two or more systems tie, they all receive the average rank. Tables 4 presents the average rank (according to MiF and LCA-F) of each system over all the test sets for the corresponding batches. Note, that the average ranks are calculated for the 4 best results of each system in the batch according to the rules of the challenge[8]. The best ranked system is highlighted with bold typeface.

First, we can observe that several systems outperforms the strong MTI baseline in terms of MiF and LCA measures exhibiting state-of-the-art performances. During the first batch the flat classification approach (Asclepius system) used in [18]. In the other two batches the learning-to-rank systems proposed by NCBI (L2R systems) and the Fudan University (Antinomyra systems) ranked as the best performed ones occupying the first two places in both measures.

According to the available descriptions the only systems that made of use of the MeSH hierarchy were the ones introduced by [19]. The top-down hierarchical systems, cole_hce1, cole_hce2 and cole_hce_ne achieved mediocre results. while the utai_rebayct systems had poor performances. For the systems based on a Bayesian network this behavior was expected as they cannot scale well to large problems.

4.2 Task 2b

Phase A. Table 5 presents the statistics of the training and test data provided to the participants. The evaluation included five test batches. For the phase A of Task 2b the systems were allowed to submit responses to any of the corresponding

[7] According to the rules of BioASQ, each system had to participate in at least 4 test sets of a batch in order to be eligible for the prizes.

[8] http://bioasq.lip6.fr/general_information/Task1a/

Table 4. Average ranks for each system across the batches of the challenge for the measures MiF and LCA-F. A hyphenation symbol (-) is used whenever the system participated in less than 4 times in the batch.

System	**Batch 1**		**Batch 2**		**Batch 3**	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
Asclepius	**1.0**	3.25	7.75	7.75	-	-
L2R-n1	3.0	3.25	5.75	3.75	8.0	5.75
L2R-n5	4.25	5.75	4.5	4.5	7.75	8.75
L2R-n3	4.25	2.25	4.75	6.75	7.25	7.0
L2R-n2	2.75	**1.5**	4.75	4.0	6.0	4.25
L2R-n4	4.25	5.25	6.0	3.5	8.5	7.75
FU_System_t25	13.5	13.25	20.0	18.75	-	-
MTIFL	8.0	8.0	18.25	20.5	15.25	15.25
MTI-Default	6.25	5.5	13.0	10.75	14.25	14.25
FDU_MeSHIndexing_3	-	-	16.0	16.25	-	
FU_System_k25	15.75	15.25	19.75	19.25	-	-
FU_System_k15	15.50	13.75	17.75	15.0	-	-
FU_System_t15	14.50	13.0	19.5	17.75	-	-
Antinomyra0	-	-	**3.0**	**3.5**	**1.75**	5.0
Antinomyra1	-	-	8.75	7.75	2.0	3.25
Antinomyra3	9.50	12.25	5.0	5.25	3.5	**1.75**
Antinomyra2	-	-	6.0	7.25	2.0	2.5
Antinomyra4	12.75	14.0	8.5	7.25	4.25	3.25
FU_System	18.50	16.75	15.75	16.0	-	-
FDU_MeSHIndexing_1	-	-	14.25	13.75	-	-
FDU_MeSHIndexing_2	-	-	15.75	14.75	-	-
Micro	21.75	22.75	24.0	27.5	23.25	28.0
PerExample	21.75	21.75	26.5	26.5	25.25	26.0
Sisyphus	-	-	6.25	12.25	10.5	12.75
Hippocrates	-	-	6.2	6.75	11.5	9.5
Macro	25.00	24.5	32.75	30.75	32.25	30.5
Spoon	21.25	20.75	34.0	33.75	-	-
Accuracy	-	-	34.0	33.25	33.25	37.25
Fork	21.75	22.25	36.25	37.75	-	-
Spork	23.00	23.25	37.25	38.75	-	-
bioasq_baseline	23.75	23.25	39.5	36.0	36.75	34.25
ESIS1	-	-	35.75	34.25	18.0	18.5
ESIS	-	-	36.75	35.75	23.75	25.75
ESIS2	-	-	26.75	27.0	19.25	19.75
ESIS3	-	-	-	-	20.25	18.5
ESIS4	-	-	-	-	20.5	22.25
cole_hce1	-	-	24.5	23.75	25.5	20.25
cole_hce_ne	-	-	26.5	25.25	26.75	22.5
cole_hce2	-	-	27.25	25.75	28.0	22.25
SNUMedinfo3	-	-	32.0	33.5	19.5	24.75
SNUMedinfo4	-	-	32.75	32.0	21.75	23.5
SNUMedinfo1	-	-	33.50	34.75	25.25	28.0
SNUMedinfo5	-	-	33.75	32.75	20.5	22.5
SNUMedinfo2	-	-	34.25	35.5	19.75	23.75
utai_rebayct	-	-	38.50	38.75	34.75	34.25
utai_rebayct_2	-	-	36.50	34.75	31.75	28.5
vanessa-0	-	-	-	-	27.75	25.0
larissa-0	-	-	-	-	37.0	36.5

types of annotations, that is documents, concepts, snippets and RDF triples. For each of the categories we rank the systems according to the Mean Average Precision (MAP) measure [3]. The final ranking for each batch is calculated as the average of the individual rankings in the different categories. The detailed results for Task 2b phase A can be found in `http://bioasq.lip6.fr/results/2b/phaseA/`.

Table 5. Statistics on the training and test datasets of Task 2b. All the numbers for the documents, snippets, concepts and triples refer to averages.

Batch	Size	# of documents	# of snippets	# of concepts	# of triples
training	310	14.28	18.70	7.11	9.00
1	100	7.89	9.64	6.50	24.48
2	100	11.69	14.71	4.24	204.85
3	100	8.66	10.80	5.09	354.44
4	100	12.25	14.58	5.18	58.70
5	100	11.07	13.18	5.07	271.68
total	810	11.83	14.92	5.93	116.30[9]

Focusing on the specific categories, (e.g., concepts or documents) for the Wishart system we observe that it achieves a balanced behavior with respect to the baselines (Table 7 and Table 6). This is evident from the value of F-measure which is much higher that the values of the two baselines. This can be explained on the fact that the Wishart-S1 system responded with short lists while the baselines return always long lists (50 and 100 items respectively). Similar observations hold also for the other four batches, the results of which are available online.

Table 6. Results for batch 1 for documents in phase A of Task2b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
SNUMedinfo1	0.0457	0.5958	0.0826	0.2612	0.0520
SNUMedinfo3	0.0457	0.5947	0.0826	0.2587	0.0501
SNUMedinfo2	0.0451	0.5862	0.0815	0.2547	0.0461
SNUMedinfo4	0.0457	0.5941	0.0826	0.2493	0.0468
SNUMedinfo5	0.0459	0.5947	0.0829	0.2410	0.0449
Top 100 Baseline	0.2274	0.4342	0.2280	0.1911	0.0070
Top 50 Baseline	0.2290	0.3998	0.2296	0.1888	0.0059
main system	0.0413	0.2625	0.0678	0.1168	0.0015
Biomedical Text Ming	0.2279	0.2068	0.1665	0.1101	0.0014
Wishart-S2	0.1040	0.1210	0.0793	0.0591	0.0002
Wishart-S1	0.1121	0.1077	0.0806	0.0535	0.0002
UMass-irSDM	0.0185	0.0499	0.0250	0.0256	0.0001
Doc-Figdoc-UMLS	0.0185	0.0499	0.0250	0.0054	0.0001
All-Figdoc-UMLS	0.0185	0.0499	0.0250	0.0047	0.0001
All-Figdoc	0.0175	0.0474	0.0236	0.0043	0.0001

Phase B. In the phase B of Task 2b the systems were asked to report exact and ideal answers. The systems were ranked according to the manual evaluation of ideal answers by the BioASQ experts [3]. For reasons of completeness we report also the results of the systems for the exact answers.

Table 7. Results for batch 1 for concepts in phase A of Task2b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
Wishart-S1	0.4759	0.5421	0.4495	0.6752	0.1863
Wishart-S2	0.4759	0.5421	0.4495	0.6752	0.1863
Top 100 Baseline	0.0523	0.8728	0.0932	0.5434	0.3657
Top 50 Baseline	0.0873	0.8269	0.1481	0.5389	0.3308
main system	0.4062	0.5593	0.4018	0.4006	0.1132
Biomedical Text Ming	0.1250	0.0929	0.0950	0.0368	0.0002

Table 8 shows the results for the exact answers for the first batch of task 2a. In case that systems didn't provide exact answers for a particular kind of questions we used the symbol "-". The results of the other batches are available at `http://bioasq.lip6.fr/results/2b/phaseB/`. From those results we can see that the systems are achieving a very high (> 90% accuracy) performance in the yes/no questions. The performance in factoid and list questions is not as good indicating that there is room for improvements. Again, the system of Wishart (Wishart-S3) for example shows consistent performance as it manages to answer relatively well in all kinds of questions.

Table 8. Results for batch 1 for concepts in phase A of Task2b.

System	Yes/no	Factoid			List		
	Accuracy	Strict Acc.	Lenient Acc.	MRR	Precision	Recall	F-measure
Biomedical Text Ming	0.9375	0.1852	0.1852	0.1852	0.0618	0.0929	0.0723
system 2	0.9375	0.0370	0.1481	0.0926	-	-	-
system 3	0.9375	0.0370	0.1481	0.0926	-	-	-
Wishart-S3	0.8438	0.4074	0.4444	0.4259	0.4836	0.3619	0.3796
Wishart-S2	0.8438	0.4074	0.4444	0.4259	0.5156	0.3619	0.3912
main system	0.5938	0.0370	0.1481	0.0926	-	-	-
BioASQ_Baseline	0.5313	-	-	-	0.0351	0.0844	0.0454
BioASQ Baseline 2	0.5000	-	-	-	0.0351	0.0844	0.0454

5 Conclusion

The participation to the second BioASQ challenge signalizes an uptake of the significance of biomedical question answering in the research community. We monitored an increased participation of both Tasks 2a and 2b. The baseline that we used this year in Task 2a incorporated techniques from last year's winning system. Although we had more data and thus more possible sources of errors (but also more training data), the best system in the first challenge clearly outperformed the baseline. This suggest an improvement of large-scale classification systems over the last year. The results achieved in Task 2b also suggest that the state of the art was pushed a step further. Consequently, we regard the outcome of the challenge as a success towards pushing the research on bio-medical information systems a step further. In future editions of the challenge, we aim to provide even more benchmark data derived from a community-driven acquisition process.

References

1. Joel Robert Adams and Steven Bedrick. Automatic classi
cation of pubmed abstracts with latent semantic indexing: Working notes. In *Proceedings of Question Answering Lab at CLEF*, 2014.
2. Alan R. Aronson and Franois-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17:229–236, 2010.
3. Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artieres, and Patrick Gallinari. Evaluation Framework Specifications. Project deliverable D4.1, 05/2013 2013.
4. Christopher J.C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, June 2010.
5. Sungbin Choi and Jinwook Choi. Classification and retrieval of biomedical literatures: Snumedinfo at clef qa track bioasq 2014. In *Proceedings of Question Answering Lab at CLEF*, 2014.
6. Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.
7. Minlie Huang, Aurlie Nvol, and Zhiyong Lu. Recommending mesh terms for annotating biomedical articles. *JAMIA*, 18(5):660–667, 2011.
8. Susan C. Schmidt Alan R. Aronson James G. Mork, Dina Demner-Fushman. Recent enhancements to the nlm medical text indexer. In *Proceedings of Question Answering Lab at CLEF*, 2014.
9. Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. *CoRR*, abs/1306.6802, 2013.
10. David D. Lewis et al. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.
11. Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop 'Text Summarization Branches Out'*, pages 74–81, Barcelona, Spain, 2004.
12. Jessa Lingeman and Laura Dietz. UMass at BioASQ 2014: Figure-inspired text retrieval. In *2nd BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2014.
13. Ke Liu, Junqiu Wu, Shengwen Peng, Chengxiang Zhai, and Shanfeng Zhu. The fudan-uiuc participation in the bioasq challenge task 2a: The antinomyra system. In *Proceedings of Question Answering Lab at CLEF*, 2014.
14. Yifeng Liu. BioASQ System Descriptions (Wishart team). Technical report, 2013.
15. Yuqing Mao and Zhiyong Lu. NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic MeSH Indexing. Technical report, 2013.
16. James Mork, Antonio Jimeno-Yepes, and Alan Aronson. The NLM Medical Text Indexer System for Indexing Biomedical Literature, 2013.
17. Mariana Neves. Hpi in-memory-based database system in task 2b of bioasq. In *Proceedings of Question Answering Lab at CLEF*, 2014.

18. Yannis Papanikolaou, Dimitrios Dimitriadis, Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, and Ioannis Vlahavas. Ensemble Approaches for Large-Scale Multi-Label Classification and Question Answering in Biomedicine. In *2nd BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2014.
19. Francisco Ribadas, Luis de Campos, Victor Darriba, and Alfonso Romero. Two hierarchical text categorization approaches for BioASQ semantic indexing challenge. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.
20. Francisco J. Ribadas-Pena, Luis M. de Campos Ibanez, Victor Manuel Darriba-Bilbao, and Alfonso E. Romero. Cole and utai participation at the 2014 bioasq semantic indexing challenge. In *Proceedings of Question Answering Lab at CLEF*, 2014.
21. Jr. Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining Knowledge Discovery*, 22:31–72, 2011.
22. Lei Tang, Suju Rajan, and Vijay K. Narayanan. Large scale multi-label classification via metalabeler. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 211–220, New York, NY, USA, 2009. ACM.
23. Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining Multi-label Data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.
24. Grigorios Tsoumakas, Manos Laliotis, Nikos Markontanatos, and Ioannis Vlahavas. Large-Scale Semantic Indexing of Biomedical Publications. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.
25. Dirk Weissenborn, George Tsatsaronis, and Michael Schroeder. Answering Factoid Questions in the Biomedical Domain. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.
26. Zhiyong Lu Yuqing Mao, Chih-Hsuan Wei. Ncbi at the 2014 bioasq challenge task: large-scale biomedical semantic indexing and question answering. In *Proceedings of Question Answering Lab at CLEF*, 2014.
27. Donhqing Zhu, Dingcheng Li, Ben Carterette, and Hongfang Liu. An Incemental Approach for MEDLINE MeSH Indexing. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.