Results of the BioASQ tasks of the Question Answering Lab at CLEF 2015

Georgios Balikas, Aris Kosmopoulos, Anastasia Krithara, Georgios Paliouras, Ioannis Kakadiaris

Results of the BioASQ tasks of the Question Answering Lab at CLEF 2015

Georgios Balikas[2], Aris Kosmopoulos[1], Anastasia Krithara[1], Georgios Paliouras[1⋆], and Ioannis Kakadiaris[3]

[1] NCSR "Demokritos", Greece
[2] Laboratoire d' Informatique de Grenoble, France
[3] University of Houston, USA

Abstract. The goal of the BioASQ challenge is to push research towards highly precise biomedical information access 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 biomedicine. The third challenge consisted of two tasks: semantic indexing and question answering. 59 systems by 18 different teams participated in the semantic indexing task (Task 3a). The question answering task was further subdivided into two phases. 24 systems from 9 different teams participates in the annotation phase (Task 3b-phase A), while 26 systems of 10 different teams participated in the answer generation phase (Task 3b-phase B). Overall, the best systems were able to outperform the strong baselines provided by the organizers. In this paper, we present the data used during the challenge as well as the technologies which were used by the participants.

1 Introduction

The aim of this paper is to present an overview of the BioASQ challenge in CLEF 2015. The overview provides information about:

1. the two BioASQ tasks of the Question Answering Lab at CLEF 2015,
2. the data provided during the BioASQ tasks,
3. the systems that participated in the challenge, according to the system descriptions that we have received; detailed descriptions of some of the systems are given in the lab proceedings which we cite,
4. evaluation results about the performance of the participating systems and compare them to dedicated baseline systems.

2 Overview of the Tasks

The challenge comprised two tasks: (1) a large-scale semantic indexing task (Task 3a) and (2) a question answering task (Task 3b). Information about the challenge and the nature of the data it provides is available at [21, 2].

⋆ contact email: paliourg@iit.demokritos.gr

Large-scale semantic indexing. In Task 3a the goal is to classify documents from the MEDLINE[4] digital library unto concepts of the MeSH[5] hierarchy. Here, new MEDLINE 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 MEDLINE curators, the performance of each system is assessed using standard information retrieval measures as well as hierarchical ones. The winners of each batch are decided based on their performance in the Micro F-measure (MiF) from the family of flat measures [22], and the Lowest Common Ancestor F-measure (LCA-F) from the family of hierarchical measures [11]. For completeness several other flat and hierarchical measures are 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 following a pre-announced schedule. The test sets were released on a weekly basis (on Monday 17.00 CET) and the participants were asked to provide their system's answers within 21 hours. Figure 1 gives an overview of the time plan of Task 3a.

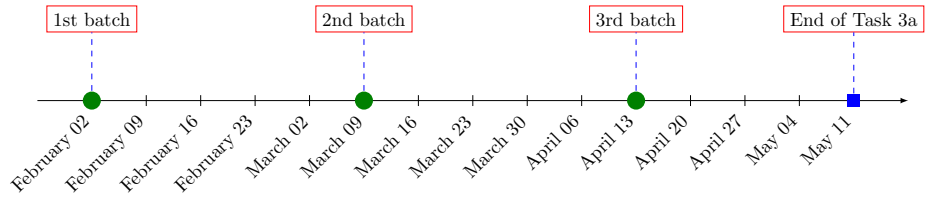

Fig. 1. The time plan of Task 3a.

Biomedical semantic QA. The goal of Task 3b was to assess the performance of participating systems in different stages of the question answering process, ranging from the retrieval of relevant concepts and articles, to the generation of natural-language answers. Task 3b 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 question: "yes/no" questions, "factoid" questions,"list" questions and "summary" questions [3]. Participants were asked to respond with relevant concepts (from specific terminologies and ontologies), relevant articles (PubMed and PubMedCentral[6] articles), relevant snippets extracted from the relevant articles and relevant RDF triples (from specific ontologies). In phase B, the released questions were accompanied by the correct answers for a subset of the required elements of phase A; namely documents and

[4] http://www.ncbi.nlm.nih.gov/pubmed/
[5] http://www.ncbi.nlm.nih.gov/mesh/
[6] http://www.ncbi.nlm.nih.gov/pmc/

snippets.[7] The participants had to answer with *exact* answers as well as with paragraph-sized summaries in natural language (dubbed *ideal* answers).

The task was split into five independent batches (see Fig. 2). 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 [12] was also reported.

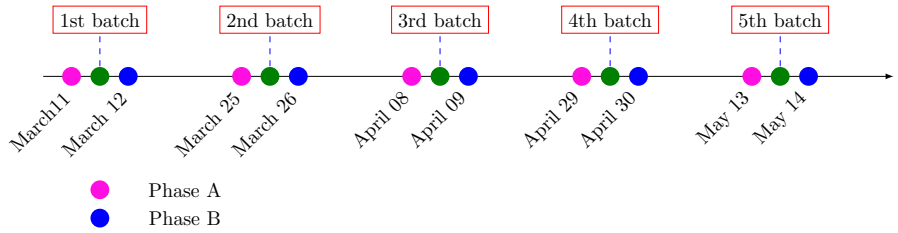

Fig. 2. The time plan of Task 3b. The two phases for each batch ran in consecutive days.

3 Technology Overview of the Participating Systems

3.1 Task 3a

The systems that participated in the semantic indexing task of the BioASQ challenge adopted a variety of approaches based mostly on flat classification. In the rest of section we describe the participating systems and stress their key characteristics.

The *NCBI* system [14], called MeSH Now, was contributed as a baseline system for the semantic indexing task of 2015. This allowed other participants to use its predictions, in order to improve their own results. The system is very similar to that developed by *NCBI* for the BioASQ2 challenge, based on the generic learning-to-rank approach presented in [9]. The main improvements were the addition of new training data from the third iteration of the challenge and the submission of two separate runs each week, one favoring high F_1 and one favoring high recall. Improvements were also done on the scalability of the system which now runs in parallel using a computer cluster.

[7] In the first two editions of the BioASQ challenge, the datasets released for Phase B contained relevant articles, snippets, concepts and RDF triples for each question.

The *AUTH-Atypon* system [16] also adopted a flat classification approach. The approach is based on binary linear SVM models for each class. A Meta-Labeler [20] is used to predict the number of classes that an instance should be assigned to. An ensemble of such classifiers, trained on variable training data sizes and different time periods, is then used in order to deal with the problem of Concept Drift.

A domain-independent k-nearest-neighbor approach is adopted by the *IIIT* team [1]. Initially the system uses k-NN in order to find the most relevant MeSH headings. Then a series of procedures are used, based on POS-tagging, IDF computation and SVM-rank in order to assign some extra classes to each test instance and improve the recall of the initial k-NN results. In the final step, tree-based classifiers (one versus all) are used (*FastXML*), which actually take in to account the hierarchical relations between the MeSH terms.

Another k-nearest-neighbor approach is that of *USI*[8], which does not take into account the hierarchy. The authors claim that the method is generic since it does not take into account the domain or use any NLP, although they believe that an NLP module would boost their performance. Given an instance the system finds the k nearest instances in the training corpus and then uses the labels of these instances for annotating it by computing semantic similarities. During the challenge they experimented with various parameters of their system, such as the value of k and they also took into account the predictions of the baselines in order to improve their results.

The *CoLe* and *UTAI* [18] teams introduce a new approach, compared to their approach during the previous challenges. This year they use only conventional information retrieval tools, such as Lucene, combined with k-NN methods. The authors also experimented with several approaches of index term extraction ranging from simple to more complex ones requiring the use of *NLP*.

The *ESIS** systems used the Lucene index in order to find useful features for each of the MeSH classes separately. In this direction, they selected words that co-occur often with a particular class, as well as the most common terms excluding stop words. The decision function follows an k-nearest-neighbor approach, where for each test instance and given the feature extraction process they find in the Lucene index the most common training examples that decide the class of the test instance. Intuitively, the probability of a class increases if a term that is strongly associated with it is present and decreases if a frequent term is absent.

The *Fudan* system [17] uses a learning to rank (LTR) method for predicting MeSH headings. The *MeSHLabeler* algorithm consists of two components. The first component, called *MeSHRanker*, returns an ordered list of MeSH headings for each test instance. The ranking is determined by a combination of (a) binary classifiers, one for each MeSH heading, (b) the most similar citations to the test instance, (c) pattern matching between the MeSH headings and the title of the abstract and (d) the prediction of the MTI system. The second component, called *MeSHNumber*, predicts the actual number of MeSH heading that must be assigned to each test instance.

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

Table 1. Technologies used by participants in Task 3a.

Team	Approach	Technologies
NCBI[14]	flat	k-NN, learning-to-rank
AUTH-Atypon[16]	flat	SVMs, MetaLabeler [20], Ensembles
IIIT[1]	hierarchical	k-NN, POS-tagging, SVM-rank, FastXML
USI[8]	flat	k-NN, semantic similarities, used Baseline
CoLe and *UTAI*[18]	flat	k-NN, Lucene
Fudan[17]	flat	Logistic regression, learning-to-rank, used Baseline

Baselines. Five systems have served as baseline systems for BioASQ task 3a. The first one, dubbed BioASQ_Baseline, follows a simplistic unsupervised approach to the problem and is thus easy to beat. The rest of the systems are implementations of state-of-the-art methods: the *Medical Text Indexer* (MTI) and the *MTI First Line Index* [10] were developed and are maintained by the National Library of Medicine (NLM). [8] They serve as classification systems for articles of MEDLINE and are actively used by the MEDLINE curators in order to assist them in the annotation process. Furthermore, *MeSH Now BF* and *MeSH Now HR* were developed by NCBI and were among the best-performing systems in the second edition of the BioASQ challenge [14]. Consequently, we expected these baselines to be hard to beat.

3.2 Task 3b

As mentioned above, the second task of the challenge is further divided into two phases. In the first phase, where the goal is to annotate questions with relevant concepts, documents, snippets and RDF triples, 9 teams with 24 systems participated. In the second phase, where team are requested to submit exact and paragraph-sized answers for the questions, 10 teams with 26 different systems participated.

The *OAQA* system described in [23] focuses on learning to answer factoids and list questions. The participants trained three supervised models, using factoid and list questions of the previous editions of the task. The first is an answer type prediction model, the second assigns a score to each predicted answer while the third is a collective re-ranking model. Although the system also participated in phase A of Task 3b its performance was much better in the factoid and list questions of phase B.

In contrast, the *USTB* system [25] participated only in phase A of the challenge. This approach initially uses a sequential dependence model for document retrieval. It then uses Word Embeddings (specifically the *Word2Vec* tool) to rank

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

the results and improve the document retrieval of the previous step. In the final step, biomedical concepts and corresponding RDF triples are extracted, using concept recognition tools, such as *MetaMap* and *Banner*.

Another system that focused on phase A is by the *IIIT* team and is described in [24]. The authors relied on the PubMed search engine to retrieve relevant documents. They then applied their own snippet extraction methods, which is based on the similarity of the top 10 sentences of the retrieved documents and the query.

The *HPI* system [15] participated in both phases of Task 3b. The system relies on in-memory based database technology, in order to map the given questions to concepts. The Stanford CoreNLP package is used for question tokenization and the BioASQ services are used for relevant document retrieval. The selection of snippets from the retrieved documents is performed using string similarity between terms of the question and words of the documents. Exact and ideal answers are both extracted using the gold-standard snippets that were provided to the participants.

The *Fudan* system [17] also participated in the second task of challenge. For phase A a language model is used in order to retrieve relevant documents. For snippet extraction, the retrieved documents are searched for query keywords by giving extra credit to terms that appear close to the query keywords. Regarding exact and ideal answers, the system is split into three main components: question analysis, candidate answer generation and candidate answer ranking.

In the system of *ILSP* and *AUEB* [13] a different approach for question answering is presented based on multi-document summarization from relevant documents. The system first uses an *SVR* in order to assign scores to each sentence of the relevant documents. The most relevant sentences are then combined to form an answer. In order to avoid redundancy, two main approaches are examined, the use of an *ILP* model and the use of a more greedy strategy. Several versions of the system were examined, which differ on the features and training data that was used.

The *YodaQA* system, described in [4], is a pipeline question answering system that was altered in order to make it compatible with the BioASQ task. The system first extracts natural language features from the questions and then searches its knowledge base for existing answers. It then either directly provides these passages as answers or performs passage analysis in order to produce answers from the extracted texts. Each answer is evaluated using a logistic regression classifier and those with the highest scores are provided as a final answer. The initial system was designed to answer only factoid questions, so modifications were necessary in order to be able to answer list questions.

The final system is the *SNUMedinfo* described in [5]. Regarding Phase A, the system participated only in the document retrieval task. The approach was based on the *Indri* search engine [19] and the semantic concept-enriched model presented in [6]. In phase B, the system participated only in the ideal answer generation subtask, where it ranked each passage from the provided list, based on the unique keywords it contained. A set of m (parameter of the system)

passages were selected, in rank order, by selecting only passages that contain a minimum proportion of new tokens compared the already selected ones.

Table 2 describes the principal technologies that were employed by the participating systems and in which phase (A and/or B) have participated.

Table 2. Technologies used by participants in Task 3b.

Reference	Phase	Technologies
OAQA [23]	A,B	supervised learning, collective re-ranking model
USTB [25]	A	Word Embeddings, MetaMap, Banner
IIIT [24]	A	PubMed search engine, sentence similarity
HPI [15]	A, B	Stanford CoreNLP, string similarity
Fudan [17]	A, B	language model, word similarity, ranking
ILSP-AUEB[13]	A, B	multi-document summarization, ILP model, greedy strategy
YodaQA [4]	A, B	natural language features, logistic regression
SNUMedinfo [5]	A, B	Indri search engine, semantic concept-enriched model

Baselines. The BioASQ baseline of Task 3b phase B is a system similar to [13]. It applies a multi-document summarization method using Integer Linear Programming and Support Vector Regression.

4 Results

4.1 Task 3a

During the evaluation phase of the Task 3a, the participants submitted their results on a weekly basis to the online evaluation platform of the challenge.[9] The evaluation period was divided into three batches containing 5 test sets each. 18 teams participated in the task with a total of 59 systems. Two training datasets were provided: the first contains 11,804,715 articles that cover 27,097 MeSH labels; the second is a subset containing 4,607,922 articles and covers 26,866 MeSH labels. The latter dataset focuses on the journals that appear also in the test sets. The uncompressed size of those training sets in text format is 19Gb and 7.4Gb respectively. Table 3 shows the number of articles in each test set of each batch of the challenge.

Table 4 presents the correspondence of the system names in the BioASQ Participants Area Leaderboard for Task 3a and the system description submitted in the track's working notes. Systems that participated in less than 4 test sets in each batch are not reported in the results.[10]

According to [7] the appropriate way to compare multiple classification systems over multiple datasets is based on their average rank across all the datasets.

[9] `http://participants-area.bioasq.org/`

[10] According to the rules of BioASQ, each system had to participate in at least 4 test sets of a batch in order to be able to win the batch.

Table 3. Statistics on the test datasets of Task 3a.

Batch	Articles	Annotated Articles	Labels per article
1	21014	14,145	13.03
	4,435	3,338	13.27
	3,638	2,906	13.29
	2,153	1,625	13.27
	5,725	4,223	13.10
Subtotal	36,965	26,237	13.12
2	3,617	2,634	12.60
	4,725	3,020	12.97
	4,861	3,342	13.41
	2,902	2,254	12.89
	4,059	2,911	12.67
Subtotal	20,164	14,161	12.93
3	3,902	2,937	13.40
	4,027	2,822	13.49
	3,162	2,116	13.29
	3,621	2,299	13.56
	3,842	2,362	12.82
Subtotal	18,554	12,536	13.32
Total	72,430	52,934	13.11

Table 4. Correspondence between the public names of the participating teams on the BioASQ Participants Area leaderboard and their submissions on the lab working notes.

Reference	Systems
[14]	MeSH Now HR, MeSH Now BF
[16]	auth*
[1]	qaiiit system *
[8]	Abstract framework, USI 20 neighbours, USI baseline, USI 10 neighbours
[18]	iria-*
[17]	MeSHLabeler-*

On each dataset the system with the best performance gets rank 1.0, the second best rank 2.0 and so on. In case two or more systems tie, they all receive the average rank. Table 5 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[11]. The best ranked system is highlighted with bold typeface. As it can be noticed, on all three batches and for both flat and hierarchical measures, the Fudan system [17] clearly outperforms other approaches. The AUTH-Atypon system [16] managed to score second in two out of three batches, while the MeSH-UK0 scored second in one of the batches.

[11] `http://participants-area.bioasq.org/general_information/Task3a/`

Table 5. Average ranks for each system across the batches of task 3a for the measures MiF and LCA-F. A dash (-) is used whenever the system participated in less than 4 times in the batch. Systems that didn't participate in the challenge regularly, i.e. they didn't submit results for at least four test sets in at least one of the three batches, were excluded from the table.

System	Batch 1		Batch 2		Batch 3	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
auth1	7.5	7.0	10.5	8.5	10.0	8.0
qaiiit system 1	-	-	25.0	25.0	-	-
TextCategorisation5	8.5	9.5	-	-	-	-
MeSH-UK2	-	-	-	-	9.0	7.75
Dexstr system	-	-	24.25	23.5	-	-
USI 20 neighbours	15.25	13.0	15.25	14.75	16.75	16.75
iria-1	16.5	16.0	21.25	20.75	17.25	17.25
pseudo n-grams	-	-	-	-	27.25	27.0
iria-4	-	-	-	-	24.25	24.5
auth2	7.5	8.75	7.0	9.0	7.0	7.5
test_unibitri	-	-	-	-	20.0	20.5
auth3	4.25	3.75	5.25	6.5	5.0	4.75
it is a test submit	22.5	22.0	26.0	25.5	28.0	28.0
MeSHLabeler-3	2.25	3.25	**1.0**	**1.0**	2.25	2.5
MeSH-UK0	-	-	8.0	7.25	7.75	10.75
MeSHLabeler-1	2.5	**1.75**	2.5	3.0	**2.0**	**2.0**
fork-fork	17.75	18.0	16.75	17.75	-	-
TextCategorisation3	8.75	11.0	-	-	-	-
MeSH-UK4	-	-	6.25	7.75	10.5	11.5
TextCategorisation1	11.25	12.75	-	-	-	-
testLee15	-	-	-	-	27.0	26.5
spoon-spoon	16.25	16.5	14.5	16.0	-	-
IMI-KOI	-	-	-	-	30.0	30.5
auth4	10.0	10.5	8.0	9.75	4.5	4.5
MeSHLabeler-4	**1.0**	2.25	2.75	2.5	2.5	3.5
MeSH-UK3	-	-	7.75	11.0	10.25	12.0
BioASQ_Baseline	24.25	24.25	27.75	27.5	29.25	29.0
MeSHLabeler-2	4.5	3.75	1.75	1.75	3.0	**2.0**
Default MTI	12.0	9.5	14.0	13.0	15.75	13.75
MeSHLabeler	3.5	2.25	2.5	2.25	3.75	3.25
Abstract framework	17.5	18.75	17.25	18.0	19.5	20.25
iria-2	-	-	21.0	20.0	21.75	22.0
MeSH Now BF	8.25	7.75	11.25	7.75	13.0	9.75
MeSH Now HR	23.75	23.75	20.75	20.75	31.25	31.75
USI 10 neighbours	18.5	17.75	18.25	17.5	20.5	19.75
IMI-KOI_R	-	-	-	-	31.25	30.75
iria-3	-	-	20.0	19.75	22.0	22.25
iria-mix	-	-	-	-	14.0	13.0
MTI First Line Index	16.0	12.75	16.0	15.25	18.5	18.0
USI baseline	6.25	6.0	11.5	9.25	14.5	13.5
TextCategorisation4	8.5	9.5	-	-	-	-
IIIT system 2	-	-	18.75	19.0	-	-
MeSH-UK1	-	-	4.5	4.5	9.25	9.75
TextCategorisation2	10.0	11.25	-	-	-	-

4.2 Task 3b

Phase A. Table 6 presents the statistics of the test data that were provided to the participants. The evaluation included five test batches. For phase A of Task 3b the systems were allowed to submit up to 10 responses per question to any of the corresponding type of annotation; that is documents, concepts, snippets and RDF triples. For each of the categories we rank the systems according to the

Table 6. Statistics on the test datasets of Task 3b. The numbers concerning the documents and snippets refer to averages.

Batch	Size	# of documents	# of snippets	Yes/No	List	Factoid	Summary
1	100	11.27	13.33	33	22	26	19
2	100	10.96	12.95	16	28	32	24
3	100	9.3	10.98	28	17	26	29
4	97	9.37	11.97	29	23	25	20
5	100	5.84	8.53	28	24	22	26
total	497	9.35	11.55	134	114	131	118

Table 7. Results for batch 1 for documents in phase A of Task3b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
SNUMedinfo1	0.2430	0.3055	0.2220	0.1733	0.0117
SNUMedinfo2	0.2440	0.3076	0.2231	0.1731	0.0115
SNUMedinfo4	0.2420	0.3062	0.2220	0.1724	0.0117
fdu3	0.2320	0.3275	0.2232	0.1719	0.0071
fdu2	0.2290	0.3242	0.2201	0.1703	0.0066
SNUMedinfo3	0.2340	0.2900	0.2117	0.1695	0.0076
fdu4	0.2320	0.3290	0.2242	0.1695	0.0078
ustb_prir3	0.2430	0.3092	0.2245	0.1687	0.0120
testtext	0.2410	0.3042	0.2226	0.1681	0.0124
ustb_prir4	0.2430	0.3088	0.2241	0.1666	0.0105
ustb_prir1	0.2370	0.3045	0.2194	0.1663	0.0105
fdu	0.2200	0.3045	0.2091	0.1590	0.0067
SNUMedinfo5	0.2240	0.2854	0.2050	0.1569	0.0070
qaiiit system 1	0.1957	0.1757	0.1559	0.1099	0.0006
fa1	0.1385	0.0888	0.0935	0.0489	0.0001
ilsp.aueb.1	0.1264	0.1103	0.0922	0.0485	0.0001
HPI-S2	0.1027	0.1250	0.0841	0.0464	0.0005
fdu5	0.0370	0.0314	0.0276	0.0138	0.0000

Table 8. Results for batch 1 for snippets in phase A of Task3b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
ustb_prir3	0.0845	0.0967	0.0785	0.0570	0.0004
ustb_prir1	0.0829	0.0970	0.0774	0.0546	0.0003
qaiiit system 1	0.0616	0.0697	0.0511	0.0545	0.0002
testtext	0.0887	0.0948	0.0797	0.0529	0.0004
ustb_prir4	0.0772	0.0882	0.0706	0.0513	0.0003
HPI-S2	0.0545	0.0686	0.0501	0.0347	0.0002

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. Tables 7 and 8 present the scores of the participating systems for document and snippet retrieval in the first batch of Phase A .[12] Note that systems are allowed to participate in any or all four parts of the task e.g., *SNUMedinfo** retrieved only

[12] In contrast to the first two editions of the challenge, the biomedical experts of BioASQ were not asked to produce golden concepts and triples prior to the challenge. The ground truth for concepts and snippets will be constructed by the experts on the basis of the material provided by the systems.

documents. It is worth noting, that document retrieval for the given questions was the most popular aspect of the task; far fewer systems returned document snippets, concepts and RDF triples. The detailed results for Task 3b phase A can be found in `http://participants-area.bioasq.org/results/3b/phaseA/`.

Phase B. In phase B of Task 3b, the systems were asked to generate exact and ideal answers. The systems will be 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. In contrast to the previous editions of the BioASQ challenge, the test files of Phase B included only relevant documents and snippets for each question instead of relevant documents, snippets, concepts and RDF triples. As a result, the participating systems had less information available in order to construct the exact and the ideal answers.

Table 9 shows the results for the exact answers in the first batch of task 3b. For systems that didn't provide exact answers for a particular kind of question we use the dash symbol "-". The results of the other batches are available at `http://participants-area.bioasq.org/results/3b/phaseB/`. They are not reproduced here in the interest of space. From those results we can see that some of the systems are achieving a very high ($> 80\%$ accuracy) performance in the yes/no questions. The performance in factoid and list questions is not as good, indicating that there is room for improvement. On the other hand, the performance on ideal answers has improved compared to the previous years [2], which in combination with the increase of participation leads us to believe that a significant amount of effort was invested by the participants and that the task is gaining attention. It is to be noted that those conclusions are based only on the automated evaluation measures; the manual assessment was still in progress at the time of writing this document.

Table 9. Results for batch 1 for exact and ideal answers in phase B of Task3b.

System	**Yes/no**	**Factoid**			**List**			**Ideal Answers**	
	Acc.	Strict Acc.	Lenient Acc.	MRR	Prec.	Recall	F-meas	Rouge2	Rouge-SU4
fa1	.8485	.0769	.0769	.0769	-	-	-	-	-
fdu	.8485	.0769	.1538	.1038	.0477	.2362	.0756	.2634	.2648
fdu2	.8485	.0769	.1538	.1038	.0477	.2362	.0756	.2669	.2781
fdu3	.8485	.0769	.1538	.1038	.0477	.2362	.0756	.2760	.2973
fdu4	.8485	.1154	.1923	.1423	.0379	.2957	.0650	.2760	.2973
main system	.8485	.1154	.3077	.1936	.1311	.1802	.1362	.2934	.3066
HPI-S2	.6667	-	-	-	.0292	.0603	.0364	.1884	.2008
BioASQ Baseline 2	.5455	-	-	-	-	-	-	.3604	.3787
BioASQ_Baseline	.4545	-	-	-	-	-	-	.4033	.4217
SNUMedinfo1	-	-	-	-	-	-	-	.2929	.3069
SNUMedinfo2	-	-	-	-	-	-	-	.2940	.3071
SNUMedinfo3	-	-	-	-	-	-	-	.2894	.3034
SNUMedinfo4	-	-	-	-	-	-	-	.2567	.2703
SNUMedinfo5	-	-	-	-	-	-	-	.2650	.2784
ilsp.aueb.1	-	-	-	-	-	-	-	.3829	.4052
ilsp.aueb.2	-	-	-	-	-	-	-	.4050	.4213

5 Conclusions

The third edition of the BioASQ challenge has led to a number of interesting results by the participating systems. Despite them being quite advanced systems, the baselines that we provided have been beaten by the best systems. Both tasks have attracted an increasing number of participants and the number of submissions to the workshop has also increase. Therefore, we believe that the third edition of the challenge has been another contribution towards better biomedical information systems.This encourages us to continue the effort and establish BioASQ as a reference point for research in the area. In future editions of the challenge, we aim to provide even more benchmark data derived from a community-driven acquisition process.

Acknowledgments

The third edition of BioASQ is supported by a conference grant from the NIH/NLM (number 1R13LM012214-01) and sponsored by the companies Viseo and Atypon.

References

1. Kamineni Avinash, Fatma Nausheen, Das Arpita, Shrivastava Manish, and Chinnakotla Manoj. Extreme Classification of PubMed Articles usingMeSH Labels. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
2. George Balikas, Ioannis Partalas, Axel-Cyrille Ngonga Ngomo, Anastasia Krithara, Eric Gaussier, and George Paliouras. Results of the bioasq track of the question answering lab at clef 2014. *Results of the BioASQ Track of the Question Answering Lab at CLEF*, 2014:1181–93, 2014.
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. Petr Baudis and Jan Sedivy. Biomedical Question Answering using the YodaQA System: Prototype Notes. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
5. Sungbin Choi. SNUMedinfo at CLEF QA track BioASQ 2015. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
6. Sungbin Choi, Jinwook Choi, Sooyoung Yoo, Heechun Kim, and Youngho Lee. Semantic concept-enriched dependence model for medical information retrieval. *Journal of biomedical informatics*, 47:18–27, 2014.
7. Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.
8. Nicolas Fiorini, Sylvie Ranwez, Sébastien Harispe1, Jacky Montmain, and Vincent Ranwez. USI at BioASQ 2015: a semantic similarity-based approach for semantic indexing. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.

9. Minlie Huang, Aurlie Nvol, and Zhiyong Lu. Recommending mesh terms for annotating biomedical articles. *JAMIA*, 18(5):660–667, 2011.
10. Susan C. Schmidt Alan R. Aronson James G. Mork, Dina Demner-Fushman. Recent enhancements to the nlm medical text indexer. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, volume 1180, Sheffied, UK, 2014.
11. 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.
12. 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.
13. Prodromos Malakasiotis, Emmanouil Archontakis, Ion Androutsopoulos, Dimitrios Galanis, and Harris Papageorgiou. Biomedical question-focused multi-document summarization: ILSP and AUEB at BioASQ3. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
14. Yuqing Mao and Zhiyong Lu. NCBI at the 2015 BioASQ challenge task: Baseline results from MeSH Now. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
15. Mariana Neves. HPI question answering system in the BioASQ 2015 challenge. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
16. Yannis Papanikolaou, Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, and Ioannis Vlahavas. AUTH-Atypon at BioASQ 3: Large-Scale Semantic Indexing in Biomedicine. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
17. Shengwen Peng, Ronghui You, Zhikai Xie, Yanchun Zhang, and Shanfeng Zhu. The Fudan participation in the 2015 BioASQ Challenge: Large-scale Biomedical Semantic Indexing and Question Answering. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
18. Francisco J. Ribadas, Luis M. de Campos, Víctor M. Darriba1, and Alfonso E. Romero. CoLe and UTAI at BioASQ 2015: experiments with similarity based descriptor assignment. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
19. Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6, 2005.
20. 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.
21. George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.
22. 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.
23. Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, and Eric Nyberg. Learning to Answer Biomedical Factoid and List Questions OAQA at BioASQ 3B. In *Work-*

ing Notes for the Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France, 2015.

24. Harish Yenala, Avinash Kamineni, Manish Shrivastava, and Manoj Chinnakotla. BioASQ 3b Challange 2015: Bio-Medical Question Answering System. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.
25. Zhi-Juan Zhang, Tian-Tian Liu, Bo-Wen Zhang, Yan Li, Chun-Hua Zhao, Shao-Hui Feng, Xu-Cheng Yin, and Fang Zhou. A generic retrieval system for biomedical literatures: USTB at BioASQ2015 Question Answering Task. In *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF)*, Toulouse, France, 2015.