Results of the fifth edition of the BioASQ Challenge

Anastasios Nentidis[1], **Konstantinos Bougiatiotis**[1], **Anastasia Krithara**[1], **Georgios Paliouras**[1] and **Ioannis Kakadiaris**[2]
[1]National Center for Scientific Research "Demokritos", Athens, Greece
[2]University of Houston, Texas, USA

Abstract

The goal of the BioASQ challenge is to engage researchers into creating cutting-edge biomedical information systems. Specifically, it aims at the promotion of systems and methodologies that are able to deal with a plethora of different tasks in the biomedical domain. This is achieved through the organization of challenges. The fifth challenge consisted of three tasks: semantic indexing, question answering and a new task on information extraction. In total, 29 teams with more than 95 systems participated in the challenge. Overall, as in previous years, the best systems were able to outperform the strong baselines. This suggests that state-of-the art systems are continuously improving, pushing the frontier of research.

1 Introduction

The aim of this paper is twofold. First, we aim to give an overview of the data issued during the BioASQ challenge in 2017. In addition, we aim to present the systems that participated in the challenge and evaluate their performance. To achieve these goals, we begin by giving a brief overview of the tasks, which took place from February to May 2017, and the challenge's data. Thereafter, we provide an overview of the systems that participated in the challenge. Detailed descriptions of some of the systems are given in workshop proceedings. The evaluation of the systems, which was carried out using state-of-the-art measures or manual assessment, is the last focal point of this paper, with remarks regarding the results of each task. The conclusions sum up this year's challenge.

2 Overview of the Tasks

The challenge comprised three tasks: (1) a large-scale semantic indexing task (Task 5a), (2) a question answering task (Task 5b) and (3) a funding information extraction task (Task 5c), described in more detail in the following sections.

2.1 Large-scale semantic indexing - 5a

In Task 5a the goal is to classify documents from the PubMed digital library into concepts of the MeSH hierarchy. Here, new PubMed articles that are not yet annotated by MEDLINE indexers are collected and used as test sets for the evaluation of the participating systems. In contrast to previous years, articles from all journals were included in the test data sets of task 5a. As soon as the annotations are available from the MEDLINE indexers, the performance of each system is calculated using standard flat information retrieval measures, as well as, hierarchical ones. As in previous years, an on-line and large-scale scenario was provided, dividing the task into three independent batches of 5 weekly test sets each. Participants had 21 hours to provide their answers for each test set. Table 1 shows the number of articles in each test set of each batch of the challenge. 12,834,585 articles with 27,773 labels were provided as training data to the participants.

2.2 Biomedical semantic QA - 5b

The goal of Task 5b was to provide a large-scale question answering challenge where the systems had to cope with all the stages of a question answering task for four types of biomedical questions: yes/no, factoid, list and summary questions (Balikas et al., 2013). As in previous years, the task comprised two phases: In phase A, BioASQ released 100 questions and participants were asked to respond with relevant elements from specific resources, including relevant MEDLINE articles, relevant snippets extracted from the articles, relevant concepts and relevant RDF triples. In phase B, the released questions were enhanced with relevant articles and snippets selected manu-

Proceedings of the BioNLP 2017 workshop, pages 48–57,
Vancouver, Canada, August 4, 2017. ©2017 Association for Computational Linguistics

Batch	Articles	Annotated Articles	Labels per Article
1	6,880	6,661	12.49
	7,457	6,599	12.49
	10,319	9,656	12.49
	7,523	4,697	11.78
	7,940	6,659	12.50
Total	40,119	34,272	12.39
2	7,431	7,080	12.40
	6,746	6,357	12.62
	5,944	5,479	12.87
	6,986	6,526	12.65
	6,055	5,492	12.41
Total	33,162	30,934	12.58
3	9,233	5,341	12.78
	7,816	2,911	12.58
	7,206	4,110	12.70
	7,955	3,569	12.17
	10,225	984	13.72
Total	42,435	21,323	12.68

Table 1: Statistics on test datasets for Task 5a.

ally and the participants had to respond with *exact answers*, as well as with summaries in natural language (dubbed *ideal answers*). The task was split into five independent batches and the two phases for each batch were run with a time gap of 24 hours. In each phase, the participants received 100 questions and had 24 hours to submit their answers. Table 2 presents the statistics of the training and test data provided to the participants. The evaluation included five test batches.

Batch	Size	Documents	Snippets
Train	1,799	11.86	20.38
Test 1	100	4.87	6.03
Test 2	100	3.93	5.13
Test 3	100	4.03	5.47
Test 4	100	3.23	4.52
Test 5	100	3.61	5.01
Total	2,299	10.14	17.09

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

2.3 Funding information extraction - 5c

Task 5c was introduced for the first time this year and the challenge at hand was to extract grant information from Biomedical articles. Funding information can be very useful; in order to estimate, for example, the impact of an agency's funding in the biomedical scientific literature or to identify agencies actively supporting specific directions in research. MEDLINE citations are annotated with information about funding from specified agencies[1]. This funding information is either provided by the author manuscript submission systems or extracted manually from the full text of articles during the indexing process. In particular, NLM human indexers identify the grant ID and the funding agencies can be extracted from the string of the grant ID[2]. In some cases, only the funding agency is mentioned in the article, without the grant ID.

In this task funding information from MEDLINE was used, as golden data, in order to train and evaluate systems. The systems were asked to extract grant information mentioned in the full text, but author-provided information is not necessarily mentioned in the article. Therefore, grant IDs not mentioned in the article were filtered out. This filtering also excluded grant IDs deviating from NLM's general policy of storing grant IDs as published, without any normalization. When an agency was mentioned in the text without a grant ID, it was kept only if it appeared in the list of agencies and abbreviations provided by NLM. Cases of misspellings or alternative naming of agencies were removed. In addition, information for funding agencies that are no longer indexed by NLM was omitted. Consequently, the golden data used in the task consisted of a subset of all funding information mentioned in the articles.

During the challenge, a training and a test dataset were prepared. The test set of MEDLINE documents with their full-text available in PubMed Central was released and the participants were asked to extract grant IDs and grant agencies mentioned in each test article. The participating systems were evaluated on (a) the extraction of grant IDs, (b) the extraction of grant agencies and (c) full-grant extraction, i.e. the combination of grant ID and the corresponding funding agency. Table 3 contains details regarding the datasets for training and test.

[1] https://www.nlm.nih.gov/bsd/grant_acronym.html

[2] https://www.nlm.nih.gov/bsd/mms/medlineelements.html#gr

Dataset	Articles	Grant IDs	Agencies	Time Period
Training	62,952	111,528	128,329	2005-13
Test	22,610	42,711	47,266	2015-17

Table 3: Dataset overview for Task 5c.

3 Overview of Participants

3.1 Task 5a

For this task, 10 teams participated and results from 31 different systems were submitted. In the following paragraphs we describe those systems for which a description was obtained, stressing their key characteristics. An overview of the systems and their approaches can be seen in Table 4.

System	Approach
Search system	search engine, UIMA ConceptMapper
MZ	tf-idf, LDA, BR classification
Sequencer	recurrent neural networks
DeepMesh	d2v, tf-idf, MESHlabeler
AUTH	d2v, tf-idf, LLDA, SVM, ensembles
Iria	bigrams, Luchene Index, k-NN, ensembles, UIMA ConceptMapper

Table 4: Systems and approaches for Task 5a. Systems for which no description was available at the time of writing are omitted.

The "*Search system*" and its variants were developed as a UIMA-based text and data mining workflow, where different search strategies were adopted to automatically annotate documents with MeSH terms. On the other hand, the "*MZ*" systems applied Binary Relevance (BR) classification, using TF-IDF features, and Latent Dirichlet allocation (LDA) models with label frequencies per journal as prior frequencies, using regression for threshold prediction. A different approach is adopted by the "*Sequencer*" systems, developed by the team from the Technical University of Darmstadt, that considers the task as a sequence-to-sequence prediction problem and use recurrent neural networks based algorithm to cope with it.

The "*DeepMeSH*" systems implement document to vector (*d2v*) and tf-idf feature embeddings (Peng et al., 2016), alongside the MESHLabeler system (Liu et al., 2015) that achieved the best scores overall, integrating multiple evidence using learning to rank (LTR). A similar approach, with regards to the d2v and tf-idf representations of the text, is followed by the "*AUTH*" team. Regarding the learning algorithms they've extended their previous system (Papagiannopoulou et al., 2016), improving the Labeled LDA and SVM base models, as well as introducing a new ensemble methodology based on label frequencies and multi-label stacking. Last but not least, the team from the University of Vigo developed the "*Iria*" systems. Building upon their previous approach (Ribadas et al., 2014) that uses an Apache Lucene Index to provide most similar citations, they developed two systems that follow a multilabel k-NN approach. They also incorporated token bigrams and PMI scores to capture relevant multiword terms through a voting ensemble scheme and the ConceptMapper annotator tool, from the Apache UIMA project (Tanenblatt et al., 2010), to match subject headings with the citation's abstract text.

Baselines: During the challenge, two systems served as baselines. The first baseline is a state-of-the-art method called Medical Text Indexer (MTI) (Mork et al., 2014) with recent improvements incorporated as described in (Zavorin et al., 2016). MTI is developed by the National Library of Medicine (NLM) and serves as a classification system for articles of MEDLINE, assisting the indexers in the annotation process. The second baseline is an extension of the system MTI, incorporating features of the winning system of the first BioASQ challenge (Tsoumakas et al., 2013).

3.2 Task 5b

The question answering task was tackled by 51 different systems, developed by 17 teams. In the first phase, which concerns the retrieval of information required to answer a question, 9 teams with 25 systems participated. In the second phase, where teams are requested to submit exact and ideal answers, 10 teams with 29 different systems participated. Two of the teams participated in both phases. An overview of the technologies employed by each team can be seen in Table 5.

The "*Basic QA pipeline*" approach is one of the two that participated in both Phases. It uses MetaMap for query expansion, taking into account

Systems	Phase	Approach
Basic QA pipeline	A, B	MetaMap, BM25
Olelo	A, B	NER, UMLS, SAP HANA, SRL
USTB	A	sequential dependence models, ensembles
fdu	A	MESHLabeler, Language model, word similarity
UNCC	A	Stanford Parser, Semantic Indexing
MQU	B	deep learning, neural nets, regression
Oaqa	B	agglomerative clustering, tf-idf, word embeddings, maximum margin relevance
LabZhu	B	PubTator, Standford POS tool, ranking
DeepQA	B	FastQA, SQuAD
sarrouti	B	UMLS, BM25, dictionaries

Table 5: Systems and approaches for Task 5b. Systems for which no information was available at the time of writing are omitted.

the text and the title of each article, and the BM25 probabilistic model (Robertson et al., 1995) in order to match questions with documents, snippets etc. The same goes for phase B, except for the exact answers, where stop words were removed and the top-k most frequent words were selected. "*Olelo*" is the second approach that tackles both phases of task B. It is built on top of the SAP HANA database and uses various NLP components, such as question processing, document and passage retrieval, answer processing and multi-document summarization based on previous approaches (Schulze et al., 2016) to develop a comprehensive system that retrieves relevant information and provides both exact and ideal answers for biomedical questions. Semantic role labeling (SRL) based extensions were also investigated.

One of the teams that participated only in phase A, is "*USTB*" who combined different strategies to enrich query terms. Specifically, sequential dependence models (Metzler and Croft, 2005), pseudo-relevance feedback models, fielded sequential dependence models and divergence from randomness models are used on the training data to create better search queries. The "*fdu*" systems, as in previous years (Peng et al., 2015), use a language model in order to retrieve relevant documents and keyword scoring with word similarity for snippet extraction. The "*UNCC*" team on the other hand, focused mainly on the retrieval of relevant concepts and articles using the Stanford Parser (Chen and Manning, 2014) and semantic indexing.

In Phase B, the Macquarie University (MQU) team focused on ideal answers (Molla, 2017), submitting different models ranging from a "trivial baseline" of relevant snippets to deep learning under regression settings (Malakasiotis et al., 2015) and neural networks with word embeddings. The Carnegie Mellon University team ("*OAQA*"), focused also on ideal answer generation, building upon previous versions of the "*OAQA*" system. They used extractive summarization techniques and experimented with different biomedical ontologies and algorithms including agglomerative clustering, Maximum Marginal Relevance and sentence compression. They also introduced a novel similarity metric that incorporates both semantic information (using word embeddings) and tf-idf statistics for each sentence/question.

Many systems used a modular approach breaking the problem down to question analysis, candidate answer generation and answer ranking. The "*LabZhu*" systems, followed this approach, based on previous years' methodologies (Peng et al., 2015). In particular, they applied rule-based question type analysis and used Standford POS tool and PubTator for candidate answer generation. They also used word frequencies for candidate answer ranking. The "*DeepQA*" systems focused on factoid and list questions, using an extractive QA model, restricting the system to output substrings of the provided text snippets. At the core of their system stands a state-of-the-art neural QA system, namely FastQA (Weissenborn et al., 2017), extended with biomedical word embeddings. The model was pre-trained on a large-scale open-domain QA dataset, SQuAD (Rajpurkar et al., 2016), and then the parameters were fine-tuned on the BioASQ training set. Finally, the "*sarrouti*" system, from Morocco's USMBA, uses among others a dictionary approach, term frequencies of UMLS metathesaurus' concepts and the BM25 model.

Baselines: For this challenge the open source

OAQA system proposed by (Yang et al., 2016) for BioASQ4 was used as a strong baseline. This system, as well as its previous version (Yang et al., 2015) for BioASQ3, had achieved top performance in producing exact answers. The system uses an UIMA based framework to combine different components. Question and snippet parsing is based on ClearNLP. MetaMap, TmTool, C-Value and LingPipe are used for concept identification and UMLS Terminology Services (UTS) for concept retrieval. In addition, identification of concept, document and snippet relevance is based on classifier components and scoring, ranking and reranking techniques are also applied in the final steps.

3.3 Task 5c

In this inaugural year for task c, 3 teams participated with a total of 11 systems. A brief outline of the techniques used by the participating systems is provided in table 6.

Systems	Approach
Simple	regions of interest, SVM, regular expressions, hand-made rules, char-distances, ensemble
DZG	regions of interest, SVM, tf-idf of bigrams, HMMs, MaxEnt, CRFs, ensemble
AUTH	regions of interest, regular expressions

Table 6: Overview of the methodologies used by the participating systems in Task 5c.

The Fudan University team, participated with a series of similar systems ("*Simple*" systems) as well as their ensemble. The general approach included the following steps: First, the articles were parsed and some sections, such as affiliation or references, were removed. Then, using NLP techniques, alongside pre-defined rules, each paragraph was split into sentences. These sentences were classified as *positive* (*i.e.* containing grant information) or not, using a linear SVM. The positive sentences were scanned for grant IDs and agencies through the use of regular expressions and hand-made rules. Finally, multiple classifiers were trained in order to merge grant IDs and agencies into suitable pairs, based on a wide range of features, such as character-level features of the grant ID, the agency in the sentence and the distance between the grant ID and the agency in the sentence.

The "*DZG*" systems followed a similar methodology, in order to classify snippets of text as possible grant information sources, implementing a linear SVM with tf-idf vectors of bigrams as input features. However, their methodology differed from that of Fudan in two ways. Firstly, they used an in-house-created dataset consisting of more than 1,600 articles with grant information in order to train their systems. Secondly, the systems deployed were based on a variety of sequential learning models namely conditional random fields (Finkel et al., 2005), hidden markov models (Collins, 2002) and maximum entropy models (Ratnaparkhi, 1998). The final system deployed was a pooling ensemble of these three approaches, in order to maximize recall and exploit complementarity between predictions of different models. Likewise, the AUTH team, with systems "*Asclepius*", "*Gallen*" and "*Hippocrates*" emphasized on specific sections of the text that could contain grant support information and extracted grant IDs and agencies using regular expressions.

Baselines: For this challenge a baseline was provided by NLM ("*BioASQ Filtering*") which is based on a two-step procedure. First, the system classifies snippets from the full-text, as possible grant support "zones" based on the average probability ratio, generated separately by Naive Bayes (Zhang et al., 2009) and SVM (Kim et al., 2009). Then, the system identified grant IDs and agencies in these selected grant support "zones", using mainly heuristic rules, such as regular expressions, especially for detecting uncommon and irregularly formatted grant IDs.

4 Results

4.1 Task 5a

Each of the three batches of task 5a was evaluated independently. The classification performance of the systems was measured using flat and hierarchical evaluation measures (Balikas et al., 2013). The micro F-measure (MiF) and the Lowest Common Ancestor F-measure (LCA-F) were used to choose the winners for each batch (Kosmopoulos et al., 2013).

According to (Demsar, 2006) 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,

System	Batch 1		Batch 2		Batch 3	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
auth1	8.88	8.25	10.50	9.75	10.25	9.75
auth2	7.25	6.50	7.63	7.50	8.88	9.75
auth3	6.75	8.25	7.50	10.25	6.50	7.00
auth4	-	-	7.38	8.25	9.63	9.75
auth5	-	-	7.50	7.00	8.50	7.50
DeepMeSH1	1.88	1.88	**1.00**	2.00	**1.00**	1.50
DeepMeSH2	**1.00**	**1.00**	3.00	3.00	2.50	2.75
DeepMeSH3	4.00	4.63	4.00	4.00	4.00	4.13
DeepMeSH4	5.00	4.38	5.00	5.50	4.88	5.63
DeepMeSH5	2.63	2.63	1.75	**1.00**	2.25	**1.25**
iria-1	-	-	13.75	13.75	12.75	12.75
iria-2	-	-	-	-	11.75	11.75
MZ1	10.75	10.75	-	-	-	-
Optimize Macro AUC	-	-	-	-	19.25	19.25
Optimize Micro AUC	-	-	-	-	15.75	18.25
Search system-1	12.25	12.25	-	-	13.75	13.25
Search system-2	13.25	13.25	-	-	14.75	14.25
Search system-3	16.25	16.25	-	-	18.50	17.50
Search system-4	15.25	15.25	-	-	16.75	16.25
Search system-5	14.25	14.25	-	-	15.75	15.25
Default MTI	7.50	6.25	8.75	6.00	7.50	6.75
MTI First Line Index	9.13	9.25	11.50	11.50	9.50	8.75

Table 7: Average system ranks across the batches of the Task 5a. A hyphenation symbol (-) is used whenever the system participated in fewer than 4 tests in the batch. Systems with fewer than 4 participations in all batches are omitted.

the second best rank 2.0 and so on. In case two or more systems tie, they all receive the average rank. Table 7 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.

On both test batches and for both flat and hierarchical measures, the *DeepMeSH* systems (Peng et al., 2016) and the AUTH systems outperform the strong baselines, indicating the importance of the methodologies proposed, including d2v and tf-idf transformations to generate feature embeddings, for semantic indexing. More detailed results can be found in the online results page [3].

4.2 Task 5b

Phase A: For phase A and for each of the four types of annotations: documents, concepts, snippets and RDF triples, we rank the systems according to the Mean Average Precision (MAP) measure. The final ranking for each batch is calculated as the average of the individual rankings in the different categories. In tables 8 and 9 some indicative results from batch 3 are presented. Full results are available in the online results page of task 5b, phase A[4].

It is worth noting that document and snippet retrieval for the given questions were the most popular part of the task. Moreover, for different evaluation metrics, there are different systems performing best, indicating that different approaches to the task may be preferable depending on the target

[3] http://participants-area.bioasq.org/results/5a/

[4] http://participants-area.bioasq.org/results/5b/phaseA/

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
testtext	0.1255	0.1789	0.1331	0.0931	**0.0017**
ustb-prir1	0.1306	0.1838	0.1372	0.0935	0.0016
ustb-prir4	0.1323	**0.2003**	**0.1412**	**0.1027**	0.0016
ustb-prir3	0.1307	0.1846	0.1376	0.0982	0.0015
ustb-prir2	0.1270	0.1832	0.1340	0.0975	0.0013
fdu	0.1551	0.1401	0.1286	0.0650	0.0005
fdu2	**0.1611**	0.1296	0.1185	0.0653	0.0005
Olelo	0.0702	0.1135	0.0764	0.0386	0.0003
HPI-S1	0.0475	0.1032	0.0593	0.0367	0.0003
KNU-SG	0.0678	0.0980	0.0702	0.0465	0.0003
c-e-50	0.0493	0.0662	0.0488	0.0345	0.0001
c-50	0.0520	0.0772	0.0530	0.0360	0.0001
c-idf-qe-1	0.0414	0.0574	0.0427	0.0326	0.0001
c-f-200	0.0485	0.0685	0.0484	0.0299	0.0001

Table 8: Results for snippet retrieval in batch 3 of phase A of Task 5b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
ustb-prir4	0.1707	0.4787	0.2200	0.1143	**0.0066**
ustb-prir1	0.1680	0.4750	0.2155	0.1108	0.0060
fdu2	0.1645	0.4628	0.2135	0.0976	0.0059
ustb-prir2	0.1737	0.4754	0.2220	0.1134	0.0059
ustb-prir3	0.1620	**0.4803**	0.2111	**0.1157**	0.0050
fdu	0.1615	0.4475	0.2120	0.1021	0.0049
testtext	0.1610	0.4690	0.2087	0.1138	0.0048
fdu4	0.1420	0.4310	0.1856	0.0926	0.0044
fdu3	0.1390	0.4098	0.1809	0.0976	0.0031
UNCC System 1	**0.2317**	0.3340	**0.2322**	0.0825	0.0009
fdu5	0.1060	0.2461	0.1298	0.0737	0.0007
Olelo	0.1327	0.2444	0.1481	0.0658	0.0005
HPI-S1	0.0823	0.2152	0.0997	0.0464	0.0005
KNU-SG	0.0730	0.2149	0.0967	0.0521	0.0005
c-e-50	0.0720	0.1921	0.0861	0.0547	0.0003
c-50	0.0720	0.1921	0.0861	0.0547	0.0003
c-idf-qe-1	0.0720	0.1921	0.0861	0.0547	0.0003
c-f-200	0.0720	0.1921	0.0861	0.0547	0.0003

Table 9: Results for document retrieval in batch 3 of phase A of Task 5b.

outcome. For example, one can see that the *UNCC System 1* performed the best on some unordered measures, namely mean precision and f-measure, however using MAP or GMAP to consider the order of retrieved elements, it is out preformed by other systems, such as the *ustb-prir*. Additionally, the combination of some of these approaches seem like a promising direction for future research.

Phase B: In phase B of Task 5b the systems were asked to produce exact and ideal answers. For ideal answers, the systems will eventually be ranked according to manual evaluation by the BioASQ experts (Balikas et al., 2013). Regarding exact answers[5], the systems were ranked according to accuracy for the yes/no questions, mean reciprocal rank (MRR) for the factoids and mean

[5]For summary questions, no exact answers are required

System	Yes/No	Factoid			List		
	Accuracy	Strict Acc.	Lenient Acc.	MRR	Precision	Recall	F-measure
Lab Zhu,Fudan Univer	0.5517	0.1818	0.3030	0.2298	**0.3608**	0.4231	**0.3752**
LabZhu,FDU	0.5517	0.2424	0.3636	0.2904	0.3608	0.4231	**0.3752**
LabZhu-FDU	0.5517	0.2727	0.3939	0.3207	0.3608	0.4231	**0.3752**
Deep QA (ensemble)	0.5517	**0.3030**	**0.4545**	**0.3606**	0.2833	0.3436	0.2927
Deep QA (single)	0.5517	0.2424	0.3939	0.2965	0.2254	0.3564	0.2419
Oaqa-5b	**0.6552**	0.1515	0.1818	0.1667	0.1252	**0.5353**	0.1909
Oaqa 5b	0.6207	0.0909	0.1212	0.1061	0.1165	0.4615	0.1792
Oaqa5b-tfidf	0.6207	0.0909	0.1212	0.1061	0.1165	0.4615	0.1792
LabZhu-FDU	0.5517	0.0909	0.1818	0.1313	0.1239	0.3077	0.1692
Lab Zhu ,Fdan Univer	0.5517	0.1212	0.2121	0.1591	0.1143	0.3077	0.1599
sarrouti	0.6207	0.0909	0.1212	0.0970	0.1077	0.2013	0.1369
Basic QA pipline	0.5517	0.0606	0.1818	0.1035	0.0769	0.1462	0.0967
SemanticRole Labeling	0.5517	0.0303	0.0606	0.0379	0.0846	0.1122	0.0943
fa1	0.5517	0.0909	0.1818	0.1187	0.0564	0.1333	0.0718
Olelo	0.5517	0.0000	0.0606	0.0253	0.0513	0.0513	0.0513
Olelo-GS	0.5172	-	-	-	0.0513	0.0513	0.0513
L2PS - Relations	0.5172	0.0303	0.0303	0.0303	0.0371	0.1667	0.0504
L2PS - DeepQA	0.5172	0.0000	0.0303	0.0061	0.0207	0.2423	0.0338
L2PS	0.5172	-	-	-	0.0192	0.0513	0.0280
Simple system	0.5517	-	-	-	-	-	-
fa2	0.5517	0.0303	0.0606	0.0404	-	-	-
fa3	0.5517	0.0303	0.0909	0.0465	-	-	-
Using NNR	0.5517	-	-	-	-	-	-
Using regression	0.5517	-	-	-	-	-	-
Trivial baseline	0.5517	-	-	-	-	-	-
BioASQ-Baseline	0.4828	0.0303	0.1212	0.0682	0.1624	0.4276	0.2180

Table 10: Results for batch 4 for exact answers in phase B of Task 5b.

F-measure for the list questions. Table 10 shows the results for exact answers for the fourth batch of task 5b. The symbol (-) is used when systems don't provide exact answers for a particular type of question. The full results of phase B of task 5b are available online[6].

From the results presented in Table 10, it can be seen that systems achieve high scores in the yes/no questions. This was especially in the first batches, where a high imbalance in yes-no classes leaded to trivial baseline solutions being very strong. This was amended in the later batches, as shown in the table for batch 4, where the best systems outperform baseline approaches.

On the other hand, the performance in factoid and list questions indicates that there is more room for improvement in these types of answer.

4.3 Task 5c

Regarding the evaluation of Task 5c and taking into account the fact that only a subset of grant IDs and agencies mentioned in the full text were included in the ground truth data sets, both for training and testing, micro-recall was the evaluation measure used for all three sub-tasks. This means that each system was assigned a micro-recall score for grant IDs, agencies and full-grants independently and the top-two contenders for each sub-

[6] http://participants-area.bioasq.org/results/5b/phaseB/

System	Grant ID MR	Grant Agency MR	Full-Grant MR
Simple-ML2	**0.9750**	0.9900	**0.9526**
Simple-ML	0.9702	**0.9907**	0.9523
simpleSystem	0.9684	0.9890	0.9505
Simple-Regex2	0.9550	0.9847	0.9416
Gallen	0.9498	0.9862	0.9412
Hippocrates	0.9491	0.9859	0.9409
Simple-Regex	0.9530	0.9844	0.9397
Asclepius	0.9472	0.9859	0.9390
DZG1	0.9232	0.9122	0.8443
DZG-agency	0.0000	0.8829	0.0000
DZG-grants	0.9235	0.0000	0.0000
BIOASQ Filtering	0.8167	0.8312	0.7174

Table 11: Micro Recall (MR) results on the test set of Task 5c.

task were selected as winners.

The results of the participating systems can be seen in Table 11. Firstly, it can be seen that the grant ID extraction task is harder compared to the agency extraction. Moreover, the overall performance of the participants was very good, and certainly better than the baseline system. This indicates that the currently deployed techniques can be improved and as discussed in section 3.3, this can be done through the use of multiple methodologies. Finally, these results, despite being obtained on a filtered subset of the data available, could serve as a springboard to enhance and redeploy the currently implemented systems.

5 Conclusion

In this paper, an overview of the fifth BioASQ challenge is presented. The challenge consisted of three tasks: semantic indexing, question answering and funding information extraction. Overall, as in previous years, the best systems were able to outperform the strong baselines provided by the organizers. This suggests that advances over the state of the art were achieved through the BioASQ challenge but also that the benchmark in itself is challenging. Consequently, we believe that the challenge is successfully towards pushing the research frontier in on biomedical information systems.

In future editions of the challenge, we aim to provide even more benchmark data derived from a community-driven acquisition process and design a multi-batch scenario for Task 5c similar to the other tasks. Finally, as a concluding remark, it is worth mentioning that the increase in challenge participation this year[7] highlights the healthy growth of the BioASQ community, gathering attention from different teams around the globe and constituting a reference point for biomedical semantic indexing and question answering.

Acknowledgments

The fifth edition of BioASQ is supported by a conference grant from the NIH/NLM (number 1R13LM012214-01) and sponsored by the Atypon Systems inc. BioASQ is grateful to NLM for providing baselines for tasks 5a and 5c and the CMU team for providing the baselines for task 5b. Finally, we would also like to thank all teams for their participation.

References

Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artieres, and Patrick Gallinari. 2013. Evaluation framework specifications. Project deliverable D4.1, UPMC.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*. pages 740–750.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pages 1–8.

[7] In BioASQ4, 6 teams participated in task 4a with 16 Systems and 11 teams in task 4b with 25 systems.

Janez Demsar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7:1–30.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 363–370.

Jongwoo Kim, Daniel X Le, and George R Thoma. 2009. Inferring grant support types from online biomedical articles. In *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on*. IEEE, pages 1–6.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. 2013. Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. *CoRR* abs/1306.6802. http://arxiv.org/pdf/1306.6802v2.

Ke Liu, Shengwen Peng, Junqiu Wu, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2015. Meshlabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics* 31(12):i339–i347.

Prodromos Malakasiotis, Emmanouil Archontakis, Ion Androutsopoulos, Dimitrios Galanis, and Harris Papageorgiou. 2015. Biomedical question-focused multi-document summarization: Ilsp and aueb at bioasq3. In *CLEF (Working Notes)*.

Donald Metzler and W Bruce Croft. 2005. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 472–479.

Diego Molla. 2017. Macquarie university at bioasq 5b query-based summarisation techniques for selecting the ideal answers. In *Proceedings BioNLP 2017*.

James G. Mork, Dina Demner-Fushman, Susan C. Schmidt, and Alan R. Aronson. 2014. Recent enhancements to the nlm medical text indexer. In *Proceedings of Question Answering Lab at CLEF*.

E Papagiannopoulou, Y Papanikolaou, D Dimitriadis, S Lagopoulos, G Tsoumakas, M Laliotis, N Markantonatos, and I Vlahavas. 2016. Large-scale semantic indexing and question answering in biomedicine. *ACL 2016* page 50.

Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2016. Deepmesh: deep semantic representation for improving large-scale mesh indexing. *Bioinformatics* 32(12):i70–i79.

Shengwen Peng, Ronghui You, Zhikai Xie, Yanchun Zhang, and Shanfeng Zhu. 2015. The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering. In *CEUR Workshop Proceedings*. CEUR Workshop Proceedings, volume 1391.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR* abs/1606.05250. http://arxiv.org/abs/1606.05250.

Adwait Ratnaparkhi. 1998. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania.

Francisco J Ribadas, Luis M De Campos, Vıctor M Darriba, and Alfonso E Romero. 2014. Cole and utai participation at the 2014 bioasq semantic indexing challenge. In *Proceedings of the CLEF BioASQ Workshop*. Citeseer, pages 1361–1374.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp* 109:109.

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

Michael A Tanenblatt, Anni Coden, and Igor L Sominsky. 2010. The conceptmapper approach to named entity recognition. In *LREC*.

Grigorios Tsoumakas, Manos Laliotis, Nikos Markontanatos, and Ioannis Vlahavas. 2013. Large-Scale Semantic Indexing of Biomedical Publications. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Fastqa: A simple and efficient neural architecture for question answering. *arXiv preprint arXiv:1703.04816* .

Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, and Eric Nyberg. 2015. Learning to answer biomedical factoid & list questions: Oaqa at bioasq 3b. In *CLEF (Working Notes)*.

Zi Yang, Yue Zhou, and Nyberg Eric. 2016. Learning to answer biomedical questions: Oaqa at bioasq 4b. *ACL 2016* page 23.

Ilya Zavorin, James G Mork, and Dina Demner-Fushman. 2016. Using learning-to-rank to enhance nlm medical text indexer results. *ACL 2016* page 8.

Xiaoli Zhang, Jie Zou, Daniel X Le, and George Thoma. 2009. A semi-supervised learning method to classify grant support zone in web-based medical articles. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, pages 72470W–72470W.