Results of the sixth edition of the BioASQ Challenge

Anastasios Nentidis[1,2], **Anastasia Krithara**[1], **Konstantinos Bougiatiotis**[1], **Georgios Paliouras**[1,3] and **Ioannis Kakadiaris**[3]
[1]National Center for Scientific Research "Demokritos", Athens, Greece
[2]Aristotle University of Thessaloniki, Thessaloniki, Greece
[3]University of Houston, Texas, USA

Abstract

This paper presents the results of the sixth edition of the BioASQ challenge. The BioASQ challenge aims at the promotion of systems and methodologies through the organization of a challenge on two tasks: semantic indexing and question answering. In total, 26 teams with more than 90 systems participated in this year's challenge. 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 2018. 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 2018, 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 two tasks: (1) a large-scale semantic indexing task (Task 6a) and (2) a question answering task (Task 6b).

2.1 Large-scale semantic indexing - 6a

In Task 6a 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 6a. 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. 13,486,072 articles with 12.69 labels per article, on average, were provided as training data to the participants.

2.2 Biomedical semantic QA - 6b

The goal of Task 6b was to provide a large-scale question answering challenge where the systems had to cope with all 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 manually 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 ques-

Proceedings of the 2018 EMNLP Workshop BioASQ: Large-scale Biomedical Semantic Indexing and Question Answering, pages 1–10
Brussels, Belgium, November 1st, 2018.

Batch	Articles	Annotated Articles	Labels per Article
1	7,240	6,639	11.67
	7,678	7,499	12.95
	10,488	10,319	13.04
	6,225	6,073	12.32
	6,617	6,486	12.96
Total	38,248	37,016	12.65
2	6,239	6,118	12.51
	7,152	6,803	12.75
	7,113	6,575	12.75
	5,833	5,412	13.00
	7,379	6,606	12.65
Total	33,716	31,514	12.73
3	6,469	5,768	12.58
	6,544	5,501	12.86
	6,743	5,467	12.67
	8,487	5,615	12.70
	7,478	4,038	12.63
Total	35,721	26,389	12.69

Table 1: Statistics on test datasets for Task 6a.

tions 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	2,251	12.01	14.72
Test 1	100	4.06	6.02
Test 2	100	3.77	5.03
Test 3	100	3.97	4.80
Test 4	100	3.39	4.03
Test 5	100	3.94	5.07
Total	2,751	10.52	12.95

Table 2: Statistics on the training and test datasets of Task 6b. All the numbers for the documents and snippets refer to averages.

3 Overview of Participants

3.1 Task 6a

For this task, 11 teams participated and results from 42 different systems were submitted. In the following paragraphs we describe those systems for which a description was available, stressing their key characteristics. An overview of the systems and their approaches can be seen in Table 3.

The "*SNOKE*" system variants were developed as an UIMA (Tanenblatt et al., 2010) text and data mining workflow, combined with a heterogeneous database architecture, where different search strategies were adopted to automatically select probable MeSH terms. More specifically, the system is based on the ZB MED Knowledge Environment (Müller et al., 2017), while also utilizing the Snowball Stemmer (Agichtein and Gravano, 2000), to find matches between MeSH terms and words in the title and abstract of each target document.

System	Approach
AttentionMeSH	RNN, w2v, attention scheme
AUTH	d2v, tf-idf, LLDA, SVM, ensembles
DeepMesh	d2v, tf-idf, MESHlabeler
Iria	bigrams, Luchene Index, k-NN, ensembles, UIMA ConceptMapper
SNOKE	search engine, UIMA ConceptMapper

Table 3: Systems and approaches for Task 6a. Systems for which no description was available at the time of writing are omitted.

The "*AttentionMeSH*" systems utilize deep learning and attention mechanisms which enable the models to associate textual evidence with annotations, thus providing interpretability at the word level. Firstly, they use a bidirectional gated recurrent unit to derive word representations with contextual information (Cho et al., 2014), to represent each document. At the same time, all MeSH terms are embedded using a technique that takes into account co-occuring MeSH terms in textually similar articles and finally an attention matrix (Mullenbach et al., 2018) is created based on the MeSH and word representations, leading to MeSH-specific article representations. This procedure allows the model to provide local interpretations of the predicted MeSH terms in relation to words of a specific article, raising the interesting subject of how explanations of an automatic MeSH indexer could further help human annotators in this task.

Other participating systems, including the "*DeepMeSH*" systems (Peng et al., 2016), the systems of the "*AUTH*" team (Papagiannopoulou et al., 2016) and the "*Iria*" systems (Ribadas-Pena

et al., 2015) are based on the same techniques used by theirs systems for the previous version of the challenge which are summarized in Table 3 and described in the corresponding challenge overview (Nentidis et al., 2017). Similarly to the previous year, two systems developed by the National Library of Medicine (NLM) to assist the indexers in the annotation of MEDLINE articles, served as baselines for the semantic indexing task of the challenge. The Medical Text Indexer (MTI) (Mork et al., 2014) with some enchantments introduced in (Zavorin et al., 2016) and an extension of it, incorporating features of the winning system of the first BioASQ challenge (Tsoumakas et al., 2013).

3.2 Task 6b

The question answering task was tackled by 50 different systems, developed by 15 teams. In the first phase, which concerns the retrieval of information required to answer a question, 9 teams with 27 systems participated. In the second phase, where teams are requested to submit exact and ideal answers, 10 teams with 27 different systems participated. Four of the teams participated in both phases. An overview of the technologies employed by each team can be seen in Table 4.

Systems	Phase	Approach
Olelo	A, B	SRL toolkits (BioKIT, BioSmile, PathLSTM)
AUTH	A, B	MetaMap, LingPipe, Lucene Index, Stanford Parser
AUEB	A	BM25, w2v , DL (PACRR, DRMM, ABCNN)
USTB	A	Sequential Dependence Models, Ensembles
MindLab	A	ElasticSearch, BM25, POS-Tags, w2v, DL (CNN)
MQU	B	DL (LSTM), w2v, Regression models, Reinforcement Learning
Oaqa	B	Maximum Margin Relevance, w2v, Block Ordering, ILP
LabZhu	B	PubTator, Standford POS tool, ranking
UNCC	B	Metamap, Lexical Chaining
L2PS	B	SQUAD, DRQA (RNN, LSTM), GloVe

Table 4: Systems and approaches for Task 6b. Systems for which no information was available at the time of writing are omitted.

The "*AUEB*" team that participated only in Phase A, used novel extensions of deep learning models for retrieving question-relevant documents and snippets. Firstly, they pre-trained word embeddings (Mikolov et al., 2013) on a very large collection of articles from MEDLINE/PubMed, while also implementing some pre-processing steps (stop-word removal, stemming (Krovetz, 1993), tokenization etc.). Then, for the document retrieval task they focused on the PACRR model of (Hui et al., 2017) and the DRMM model (Guo et al., 2016), while for snippets retrieval they utilized the ABCNN model (Yin et al., 2015). Alongside the extensions made on these models, they also deployed a clever post-processing scheme for snippet retrieval, as well as a model for initial document-retrieval based on BM25 (Robertson and Jones, 1976) for efficiency purposes.

Another approach based on deep learning methodologies for Phase A, focusing again on document and snippet retrieval, was proposed by the "*MindLaB*" team from the National University of Colombia. While for the document retrieval they use the BM25 model and ElasticSearch for efficiency, they train a Convolutional Neural Network (CNN) for snippet retrieval. As in the previous approach, they utilized a very large collection of PubMed Articles to train the CNN with similarity matrices of question-answer pairs. More specifically, they deploy similar pre-processing steps (tokenization, lowercasing, skip-gram embeddings (Moen and Ananiadou, 2013)) for the question and the document texts, however they also apply Part of Speech tagging to extract syntactical information regarding the terms. Based on the idea that not all terms are equally informative (Dong et al., 2015), they deploy a salience weighting scheme focusing on verbs, nouns and adjectives. Another interesting extension is the way final rankings of the snippets are generated based on a pseudo-relevance-feedback re-ranking step (Riezler et al., 2007).

In Phase B, the Macquarie University ("*MQU*") team focused on ideal answers and explored ideas of reinforcement learning on deep learning mod-

els. Extending their previous work (Molla, 2017), they implemented different models under a regression setting for finding similar sentences to a question, based on the corresponding word2vec embeddings of the question-sentence pairs. They also experimented with different ways of utilizing these embeddings, notably using a bidirectional Recurrent Neural Networks with LSTM cells (Hochreiter and Schmidhuber, 1997) to equip the model with knowledge regarding the sentence position. Moreover, they also run interesting experiments using reinforcement learning towards the ROUGE score of the ideal answers, based on their previous work (Mollá-Aliod, 2017), but the results did not advocate for the use of such models.

The Carnegie Mellon University team ("*OAQA*"), focused also on ideal answer generation, building upon previous versions of the "*OAQA*" system (Chandu et al., 2017). They experimented with ways to improve the generated answer by extracting the most relevant non-redundant sentences from multiple documents and then re-ordering and fusing them to make the resulting text more human-readable and coherent. To this end, they tried different ordering algorithms for sentences and also made various improvements in different stages of the candidate sentences expansion, fusion and filtering procedure that was already used by their model. Among the notable additions is the use of an Integer Linear Program (ILP) module that is capable of fusing repeated content and simplifying complicated sentences, thus improving human readability.

Another system deployed by the same team focuses on answer generation using a knowledge graph and a neural learning-to-rank approach, combined with different summarization techniques. One of the novelties introduced is the creation of an ontology-based retrieval module for relevant snippets, through the relation extraction between biomedical entities found in the abstracts' texts (Abacha and Zweigenbaum, 2015). Also, different learning-to-rank approaches were explored (Qin et al., 2010; Cao et al., 2006, 2007) alongside both extractive (Allahyari et al., 2017) and abstractive (See et al., 2017) summarization techniques for the ideal answers generation.

An interesting approach comes from the "*L2PS*" team where they use an open-domain model (Chen et al., 2017), pre-trained on the SQUAD (Rajpurkar et al., 2016) dataset, and fine-tuned to the biomedical domain. An interesting difference with other deep learning approaches is the fact that the GloVe embeddings (Pennington et al., 2014) were the best amongst the ones tried. Moreover, they raise interesting questions regarding the effects of non-normalized answers (synonyms, abbreviations, multi-word answers) in the evaluation of different systems.

The "*UNCC*" team participated in Phase B, deploying lexical chaining techniques (Reeve et al., 2006) for sentence similarity and ranking to extract summaries from related snippets and efficiently fuse them in an ideal answer. They take advantage of the MetaMap tool (Aronson and Lang, 2010) for biomedical entity recognition and they also present a way to extend their methodology to factoid/list question answering in Phase A as well.

"*Olelo*" is one of the approaches that tackles both phases of the question answering task. More specifically, in Phase A Semantic Role Labeling (SRL) approaches for QA systems were utilized. These focus on the automatic extraction of predicate-argument structures (PAS) from both questions and document text, aimed at finding semantically related PAS between associated pairs. For Phase B, the system 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.

Other systems, including the "*USTB*" (Jin et al., 2017) and the "*LabZhu*" (Peng et al., 2015) systems employed the same techniques used by their systems for the previous version of the challenge, as summarized in Table 4 and described in the previous challenge overview (Nentidis et al., 2017). In this challenge too, the open source OAQA system proposed by (Yang et al., 2016) served as baseline for phase B. The system which achieved among the highest performances in previous versions of the challenge remains a strong baseline for the exact answer generation task. The system is developed based on the UIMA framework. ClearNLP is employed for question and snippet parsing. MetaMap, TmTool (Wei et al.,

System	Batch 1		Batch 2		Batch 3	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
AttentionMeSH	-	-	12.75	13	10	12.875
AttentionMeSH2	-	-	13.25	13.5	9.125	11.625
AttentionMeSH3	-	-	11.875	12	8.625	10.625
AttentionMeSH4	-	-	10.625	10.75	7.375	11.375
AttentionMeSH5	-	-	9.875	11	7.25	11
DeepMeSH1	3.75	4.75	4	5	9.75	10.75
DeepMeSH2	1.875	1.875	2	2	7.25	7.5
DeepMeSH3	2.625	2.625	3	3	5.75	6
DeepMeSH4	**1**	**1**	**1**	**1**	7.25	7.75
Default MTI	4.875	3.75	5	3.75	10.5	5.25
iria-1	9.75	9.75	13	13	15.75	15.75
iria-2	-	-	-	-	18.75	18.75
MeSHmallow-1	-	-	-	-	24	24
MeSHmallow-2	-	-	-	-	24	24
MeSHmallow-3	-	-	-	-	24	24
MTI First Line Index	6	6	8.25	7.5	13.75	9.75
Semantic NoSQL KE 1	-	-	16.25	16	-	-
Semantic NoSQL KE 2	-	-	15.75	17	-	-
Semantic NoSQL KE 3	-	-	19.5	20	-	-
Semantic NoSQL KE 4	-	-	17.5	18	-	-
Semantic NoSQL KE 5	-	-	18.5	19	-	-
UMass Amherst T2T	-	-	-	-	19.25	19.25
xgx	8.5	8.5	5.75	6	5.25	4
xgx0	-	-	8.5	7	**3.25**	**2**
xgx1	-	-	-	-	4.5	2.375
xgx2	-	-	-	-	3.5	3.875
xgx3	-	-	-	-	4.75	4.25

Table 5: Average system ranks across the batches of the Task 6a. 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.

2016), C-Value and LingPipe (Baldwin and Carpenter, 2003) are used for concept identification and UMLS Terminology Services (UTS) for concept retrieval. The final steps include identification of concept, document and snippet relevance, based on classifier components and scoring, ranking and reranking techniques.

4 Results

4.1 Task 6a

Each of the three batches of Task 6a were evaluated independently. The classification performance of the systems were 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, 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.

The results in Task 6a show that in all test batches and for both flat and hierarchical measures, some systems outperform the strong baselines. The "*DeepMeSH*" systems achieve the best performance in the first two batches, outperformed only by "*xgx*" systems in the third batch. More detailed results can be found in the online results page[1]. Comparison of these results with corresponding system results from previous years reveals the improvement of both the baseline and the top performing systems through the years of the competition as shown in Figure 1.

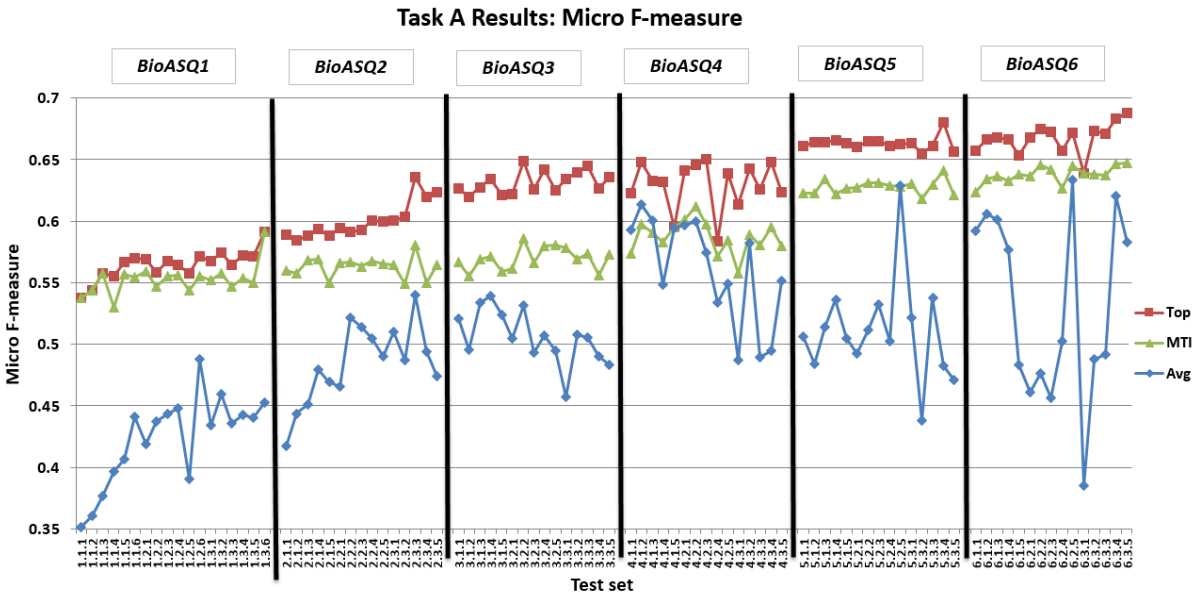

Figure 1: The micro f-measure achieved by systems across different years of the BioASQ challenge. For each test set the micro F-measure is presented for the best performing system (Top) and the MTI, as well as the average micro f-measure of all the participating systems (Avg).

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
aueb-nlp-5	**0.2551**	**0.3412**	**0.2744**	**0.2314**	**0.0068**
MindLab QA Reloaded	0.1614	0.2657	0.1877	0.1344	0.0014
aueb-nlp-1	0.1384	0.2288	0.1563	0.1331	0.0046
aueb-nlp-3	0.1341	0.2263	0.1526	0.1294	0.0038
aueb-nlp-4	0.1325	0.2252	0.1519	0.1293	0.0038
aueb-nlp-2	0.1308	0.2204	0.1494	0.1262	0.0034
MindLab QA System	0.1542	0.2754	0.1833	0.1156	0.0023
MindLab Red Lions++	0.1406	0.2346	0.1636	0.1150	0.0013
MindLab QA System ++	0.1325	0.2252	0.1559	0.1148	0.0001
testtext	0.1802	0.2331	0.1831	0.1124	0.0035

Table 6: Results for snippet retrieval in batch 3 of phase A of Task 6b. Only the top-10 systems are presented.

4.2 Task 6b

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 6 and 7 some indicative results from batch 3 are presented. Full results are available in the online results page of Task 6b, phase A[2]. These results are preliminary. The final results for Task 6b, phase A will be available after the manual assessment of the system responses.

Phase B: In phase B of Task 6b 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

[1] http://participants-area.bioasq.org/results/6a/

[2] http://participants-area.bioasq.org/results/6b/phaseA/

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
ustb_prir2	0.1660	**0.5674**	0.2186	**0.1281**	**0.0113**
ustb_prir3	**0.2007**	0.5609	**0.2496**	0.1259	0.0106
testtext	**0.2007**	0.5609	**0.2496**	0.1254	0.0106
ustb_prir4	0.1620	0.5601	0.2136	0.1253	0.0105
ustb_prir1	0.1700	0.5559	0.2203	0.1217	0.0100
aueb-nlp-2	0.1877	0.5352	0.2345	0.1147	0.0108
aueb-nlp-4	0.1877	0.5399	0.2345	0.1137	0.0106
aueb-nlp-3	0.1877	0.5429	0.2350	0.1135	0.0109
aueb-nlp-1	0.1877	0.5399	0.2345	0.1122	0.0101
sdm/rerank	0.1810	0.5422	0.2301	0.1061	0.0087

Table 7: Results for document retrieval in batch 3 of phase A of Task 6b. Only the top-10 systems are presented.

System	Yes/No		Factoid			List		
	Acc.	F1	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
Oaqa-5b	**0.6667**	**0.6592**	0.0606	0.2121	0.1313	0.0867	0.2722	0.1299
fa2	0.6296	0.3864	0.2121	**0.3030**	**0.2475**	0.2511	0.3889	0.2955
fa4	0.6296	0.3864	0.2121	**0.3030**	0.2434	0.2800	0.3889	0.3131
fa1	0.6296	0.3864	0.2121	0.2727	0.2374	0.1600	0.4333	0.2290
fa3	0.6296	0.3864	0.2121	0.2727	0.2283	0.1800	**0.4778**	0.2564
Lab Zhu ,FDU	0.6296	0.3864	0.0909	0.1212	0.1061	0.1657	0.2833	0.1663
MQ-1	0.6296	0.3864	-	-	-	-	-	-
MQ-2	0.6296	0.3864	-	-	-	-	-	-
MQ-3	0.6296	0.3864	-	-	-	-	-	-
MQ-4	0.6296	0.3864	-	-	-	-	-	-
MQ-5	0.6296	0.3864	-	-	-	-	-	-
fa5	0.6296	0.5559	0.2121	**0.3030**	0.2434	0.2800	0.3889	0.3131
Lab Zhu,FDU	0.6296	0.3864	0.2121	0.2424	0.2273	0.2944	0.3444	0.2934
LabZhu,FDU	0.6296	0.3864	**0.2424**	0.2424	0.2424	**0.4130**	0.3389	**0.3312**
BioASQ_Baseline	0.4815	0.475	0.0606	0.1212	0.0859	0.1774	0.3944	0.2236

Table 8: Results for batch 4 for exact answers in phase B of Task 6b.

exact answers[3], the systems were ranked according to accuracy, F1 score on prediction of yes answer, F1 on prediction of no and macro-averaged F1 score for the yes/no questions, mean reciprocal rank (MRR) for the factoids and mean F-measure for the list questions. Table 8 shows the results for exact answers for the fourth batch of Task 6b. 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 6b are available online[4]. These results are preliminary. The final results for Task 6b, phase B will be available after the manual assessment of the system responses.

The results presented in Table 8 show that evaluation of system performance in the yes/no questions using the macro averaged F1 measure this year is useful to identify systems that achieve good performance regardless of any dataset imbalance in the yes-no classes. In batch 4 for example, two systems outperformed the strong baseline based on previous versions of the OAQA system, which is not clear considering only the accuracy. Regarding factoid and list questions, the performance achieved by the systems indicates that there is even more room for improvement in these types of question.

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

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

5 Conclusions

In this paper, an overview of the sixth BioASQ challenge is presented. The challenge consisted of two tasks: semantic indexing and question answering. 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. Moreover, a clear shift towards the use of systems that incorporate ideas based on deep learning models can be seen, with respect to previous years. Novel ideas have been tested and state-of-the-art deep learning methodologies have been adapted to biomedical question answering with great results. Consequently, we believe that the challenge is successfully pushing the research frontier in biomedical information systems. In future editions of the challenge, we aim to provide even more benchmark data derived from a community-driven acquisition process.

Acknowledgments

The sixth 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 task 6a and the CMU team for providing the baselines for task 6b. Finally, we would also like to thank all teams for their participation.

References

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, 51(5):570–594.

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Alan R. Aronson and Franois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17:229–236.

Breck Baldwin and Bob Carpenter. 2003. Lingpipe. *Available from World Wide Web: http://alias-i.com/lingpipe*.

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.

Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. 2006. Adapting ranking svm to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM.

Khyathi Chandu, Aakanksha Naik, Aditya Chandrasekar, Zi Yang, Niloy Gupta, and Eric Nyberg. 2017. Tackling biomedical text summarization: Oaqa at bioasq 5b. *BioNLP 2017*, pages 58–66.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

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

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 260–269.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. Pacrr: A position-aware neural ir model for relevance matching. *arXiv preprint arXiv:1704.03940*.

Zan-Xia Jin, Bo-Wen Zhang, Fan Fang, Le-Le Zhang, and Xu-Cheng Yin. 2017. A multi-strategy query processing approach for biomedical question answering: Ustb_prir at bioasq 2017 task 5b. *BioNLP 2017*, pages 373–380.

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.

Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pages 39–43.

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

Diego Mollá-Aliod. 2017. Towards the use of deep reinforcement learning with global policy for query-based extractive summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 103–107.

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*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *CoRR*, abs/1802.05695.

Bernd Müller, Christoph Poley, Jana Pössel, Alexandra Hagelstein, and Thomas Gübitz. 2017. Livivo – the vertical search engine for life sciences. *Datenbank-Spektrum*, 17(1):29–34.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis Kakadiaris. 2017. Results of the fifth edition of the bioasq challenge. *BioNLP 2017*, pages 48–57.

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*, volume 1391. CEUR Workshop Proceedings.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. 2010. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Lawrence Reeve, Hyoil Han, and Ari D Brooks. 2006. Biochain: lexical chaining methods for biomedical text summarization. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 180–184. ACM.

Francisco J. Ribadas-Pena, Luis M. de Campos, Víctor Manuel Darriba Bilbao, and Alfonso E. Romero. 2015. Cole and UTAI at bioasq 2015: Experiments with similarity based descriptor assignment. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 464–471.

Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

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.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

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*.

Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2016. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics (Oxford, England)*, 32(12):1907–10.

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

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193*.

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.