Results of the Seventh Edition of the BioASQ Challenge

Anastasios Nentidis[1,2(✉)], Konstantinos Bougiatiotis[1,3], Anastasia Krithara[1], and Georgios Paliouras[1,4]

[1] National Center for Scientific Research "Demokritos", Athens, Greece
{tasosnent,bogas.ko,akrithara,paliourg}@iit.demokritos.gr
[2] Aristotle University of Thessaloniki, Thessaloniki, Greece
[3] National and Kapodistrian University of Athens, Athens, Greece
[4] University of Houston, Houston, TX, USA

Abstract. The results of the seventh edition of the BioASQ challenge are presented in this paper. The aim of the BioASQ challenge is the promotion of systems and methodologies through the organization of a challenge on the tasks of large-scale biomedical semantic indexing and question answering. In total, 30 teams with more than 100 systems participated in the challenge this year. 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.

Keywords: Semantic indexing · Question answering · Biomedical knowledge

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 2019. 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 2019, 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 the 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 biomedical semantic indexing task (Task 7a) and (2) a biomedical question answering task (Task 7b). In this section a brief description of the tasks is provided focusing on differences from previous years and updated statistics about the corresponding datasets. A complete overview of the tasks and the challenge is presented in [58].

P. Cellier and K. Driessens (Eds.): ECML PKDD 2019 Workshops, CCIS 1168, pp. 553–568, 2020.
https://doi.org/10.1007/978-3-030-43887-6_51

Table 1. Statistics on test datasets for Task 7a.

Batch	Articles	Annotated articles	Labels per article
1	7,358	7,194	11.67
	7,166	7,021	12.95
	11,019	10,831	13.04
	5,566	5,482	12.32
	6,729	6,353	12.96
Total	37,838	36,881	12.31
2	6,380	6,098	12.51
	6,785	6,621	12.75
	6,207	5,927	12.75
	7,382	7,079	13.00
	7,240	6,756	12.65
Total	33,994	32,481	12.27
3	6,266	5,835	12.58
	11,455	10,386	12.86
	4,750	3,947	12.67
	7,338	5,021	12.70
	6,920	4,554	12.63
Total	36,729	29,743	12.14

2.1 Large-Scale Semantic Indexing - 7a

In Task 7a 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. Similarly to task 5a and 6a, articles from all journals were included in the test data sets of task 7a. 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 h 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. 14,200,259 articles with 12.69 labels per article, on average, were provided as training data to the participants.

2.2 Biomedical Semantic QA - 7b

The goal of Task 7b 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 [5]. 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 h. In each phase, the participants received 100 questions and had 24 h 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.

Table 2. Statistics on the training and test datasets of Task 7b. All the numbers for the documents and snippets refer to averages.

Batch	Size	Documents	Snippets
Train	2,747	11.14	13.91
Test 1	100	3.07	3.93
Test 2	100	2.64	3.22
Test 3	100	3.08	4.05
Test 4	100	2.78	3.71
Test 5	100	2.39	2.62
Total	3,247	9.85	12.31

3 Overview of Participants

3.1 Task 7a

For this task, 12 teams participated and results from 30 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 National Library of Medicine (NLM) team, in its "*ceb*" systems [48], adopts an end-to-end deep learning architecture with Convolutional Neural Networks (CNN) [27] to improve the results of the Medical Text Indexer (MTI) [35]. In particular, they combine text embeddings with journal information. They also consider information about the years of publication and indexing, to capture concept drift and variations in the MeSH vocabulary respectively. They also experiment with an ensemble of independently trained DL models.

The Fudan University team builds upon their previous "*DeepMeSH*" systems, which are based on document to vector (*d2v*) and tf-idf feature embeddings [43], the MESHLabeler system [28] and learning to rank (LTR). This year,

Table 3. Systems and approaches for Task 7a. Systems for which no description was available at the time of writing are omitted.

System	Approach
ceb	CNN, embeddings, ensembles
DeepMesh	d2v, tf-idf, MESHlabeler, attention scheme, PLT
Iria	bigrams, Luchene Index, k-NN, ensembles, UIMA ConceptMapper
MeSHProbeNet-P	Bidirectional RNN (GRU), attention scheme, encoder-decoder architecture
Semantic NoSQL KE	UIMA ConceptMapper, par2vec, DeepLearning4j[a]

[a]https://deeplearning4j.org/ Accessed June 2019

they incorporate AttentionXML [66], a deep-learning-based extreme multi-label text classification model, in the "*DeepMeSH*" framework. In particular, AttentionXML combines a multi-label attention mechanism, to capture label-specific information, with a shallow and wide probabilistic label tree (PLT) [18], for improved efficiency.

The "*Iria*" systems [52] are based on the same techniques used by their systems for the previous version of the challenge which are summarized in Table 3 and described in the corresponding challenge overview [38].

The "*MeSHProbeNet-P*" systems are upgraded versions of MeSHProbeNet [61], which participated in BioASQ6 with the name "*xgx*". Their approach is based on an end-to-end deep learning model with an encoder-decoder architecture. The encoder consists of a recurrent neural network with multiple attentive MeSH probes to extract different aspects of biomedical knowledge from each input article. In "*MeSHProbeNet-P*" the attentive MeSH probes are also personalized for each biomedical article, based on the domain of each article as expressed by the journal where it has been published.

Finally, the "*Semantic NoSQL KE*" system variants [37] were developed extending previous year's "*SNOKE*" systems. The systems are based on the ZB MED Knowledge Environment [36], utilizing the Snowball Stemmer [1] and the UIMA [56] ConceptMapper to find matches between MeSH terms and words in the title and abstract of each target document, adopting different matching strategies. Paragraph Vectors [24] trained on the BioASQ corpus are used to rank and filter all the MeSH headings suggested by the UIMA-based framework for each document.

Similarly to the previous year, two systems developed by NLM to assist the indexers in the annotation of MEDLINE articles, served as baselines for the semantic indexing task of the challenge. MTI [35] with some enchantments introduced in [67] and an extension of it, incorporating features of the winning system of the first BioASQ challenge [59].

3.2 Task 7b

The question answering task was tackled by 73 different systems, developed by 18 teams. In the first phase, which concerns the retrieval of information required to answer a question, 6 teams with 23 systems participated. In the second phase, where teams are requested to submit exact and ideal answers, 13 teams with 52 different systems participated. An overview of the technologies employed by each team can be seen in Table 4.

Table 4. Systems and approaches for Task7b. Systems for which no information was available at the time of writing are omitted.

Systems	Phase	Approach
AUTH	A, B	MetaMap, BeCAS, Lucene Index, ElasticSearch, Wordnet, ELMo, SentiWordnet, w2vec, BiLSTM
AUEB	A	BM25, w2vec, BERT, DL (BCNN, PACRR, PDRMM)
MindLab	A	ElasticSearch, BM25, QuickUMLS, w2vec, WMD, DL (CNN)
_sys	A	Word and Sentence embeddings, Pseudo Relevance Feedback, BM25, LSI
BJUTNLP	B	SQUAD, GloVe, BiLSTM, Pointer Network
BIOASQ_VK	B	ELMo, DMN attention mechanisms, NLTK-VADER
DMIS	B	BioBERT, SQUAD, transfer learning
google	B	BERT, CoQA, Natural Questions
L2PS	B	SQUAD, Quasar-T, DRQA (RNN, LSTM), PSPR (LSTM), BioBERT
LabZhu	B	PubTator, Stanford POS tool, SPARQL
MQU	B	w2vec, tf-idf, DL (LSTM), Reinforcement Learning
UNCC	B	BioBERT, SQUAD, Stanford POS tool, AllenNLP entailment
unipi-quokka-QA	B	ELMo, ELMo-PubMed, BERT, BioBERT, SciSpacy

The "*AUTH*" team participated in both phases of Task 7B, with focus on phase B. For the document retrieval task, they experimented with approaches based on the BioASQ search services and ElasticSearch, querying with the conjunction of words in each question for the top 10 documents. In Phase B, for factoid and list questions they used updated versions of their BioASQ6 system [11], based on word embeddings, MetaMap [3], BeCAS [40] and WordNet. For yes/no questions they experiment with different deep learning methods, based on ELMo embeddings [46], SentiWordnet [12] and similarity matrices to represent the question/answer pairs and use them as input for different BiLSTM architectures [11].

The "*AUEB*" team participated in Phase A on document and snippet retrieval tasks yielding great results. They built upon their BioASQ6 document retrieval systems [6,29], which they modify to yield a relevance score for each sentence and experiment with BERT and PACRR [30] for this task. For snippet retrieval, they utilize a BCNN [64] model and a model based on POSIT-DRMM (PDRMM) [30]. They also introduce JPDRMM, a novel deep learning approach for joint document and snippet ranking, based on PDRMM [42].

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 [47]. For the document retrieval they use the BM25 model [53] and ElasticSearch [15] for efficiency, along with a Word Mover's Distance [22] based re-ranking scheme. For snippet retrieval, as in the previous approach, they utilized a very large collection of PubMed articles to train a CNN with similarity matrices of question-answer pairs. More specifically, they employ the BioNLPLab[1] w2vec embeddings that take into account the Part of Speech of each word. Also, they deploy the QuickUMLS [55] tool to create a cui2vec embedding for each snippet.

The "*_sys*" systems also participated in Phase A of Task 7B. These systems filter the queries, using stop-word lists and regular expressions, and expand them using word embeddings and pseudo-relevance feedback. Relevant documents are retrieved, utilizing Query Likelihood with bigrams and BM25, and reranked, based on Latent Semantic Indexing (LSI) and document vectors. In particular, document vectors based on averaging sentence embeddings are adopted. Finally, different lists of documents are merged to form the final result, considering the position of the documents in each list.

In phase B, most systems focused on using embeddings and deep learning methodologies to tackle the tasks. For example the "*BJUTNLP*" system utilizes the SQUAD Dataset for pre-training. The system uses both GloVe embeddings [45] (fine tuned during training) and character-level word embeddings (through a 1-dimensional CNN) as input to a BiLSTM model and for each question a Pointer Network [54] is finally responsible for pinpointing the exact start and end position of the answer in the relevant snippets.

The "*BIOASQ_VK*" systems were based on BioBERT [25], but with novel modifications to allow the model to cope with yes/no, factoid and list questions [41]. They pre-trained the model on the SQUAD dataset (for factoid and list questions) and SQUAD2 (for yes/no questions) to leverage the small size of the BioASQ dataset and by exploiting different pre-/post-processing techniques they obtained great results on all subtasks.

The "*DMIS*" systems focused on the importance of the information (words, phrases and sentences) for a given question [65]. To this end, sentence level embeddings based on ELMo embeddings [46] and attention mechanisms facilitated by Dynamic Memory Networks (DMN) [21] are deployed. Moreover, sentiment analysis is performed on yes/no questions to guide the classification (positive corresponds to yes) using the NLTK-VADER [17] tool.

[1] http://bio.nlplab.org Accessed June 2019.

The "*google*" systems [16], focus on factoid questions and are based on BERT based models [9], specifically the one in [2] trained on the Natural Questions [23] dataset, while also utilizing the CoQA [50] and the BioASQ datasets. They experiment with different input to the models, including the abstracts of relevant articles, the provided gold snippets and predicted relevant snippets. In particular, they focus on error propagation in end-to-end information retrieval and question answering systems, reaching the interesting conclusion that the information retrieval part is a bottleneck for such end-to-end QA systems.

Interesting results come from the "*L2PS*" team where they quantify the importance of pre-training and fine-tuning models for question answering and view the task under different regimes, namely Reading Comprehension (RC) and Open QA [19]. For the RC regime they use DRQA's document reader [7] while for the Open QA they utilize the PSPR model [26]. They experiment with different datasets (SQUAD [49] for RC and Quasar-T [10] for Open QA) for fine-tuning the models, as well as BioBert [25] embeddings to gain insights on the effect of the context length in this task.

The "*LabZhu*" [44] systems improved upon their systems from BioASQ6, with focus on exact answer generation. In particular, for factoid and list questions they developed two distinct approaches. One based on traditional information retrieval approaches, involving candidate answer generation and ranking, and one Knowledge-Graph based approach. In the latter approach, the answer type and the topic entity of the question are predicted and a SPARQL query is generated based on them and used to retrieve some results from the Knowledge Graph. Finally, the results of the two approaches are combined for the final answer of the question.

The Macquarie University ("*MQU*") team focused on ideal answers and approached the task under a classification approach for snippet relevance [33]. Extending their previous work [31,32] the snippets are marked as summary relevant or not, utilizing w2vec embeddings and tf-idf vectors of the question-sentence pairs, showcasing that a classification scheme is more appropriate than a regression one. Also, based on their previous work [34], they conduct experiments using reinforcement learning towards the ROUGE score of the ideal answers and a correlation analysis between various ROUGE metrics and the BioASQ human evaluation scores, observing poor correlation of the ROUGE-Recall score with human evaluation.

The "*UNCC*" team focused on factoid, list and yes/no questions [57]. Their work is based on the BioBERT [25] embeddings fine-tuned on previous years of BioASQ. They also utilize the SQUAD dataset for factoid answers and incorporated the Lexical Answer Type (LAT) [13] and POS-tags along with hand made rules to address specific errors of the system. Furthermore, they incorporated the entailment of the candidate sentences in yes/no questions using the AllenNLP library [14].

Finally, the "*unipi-quokka-QA*" system tackled all the different question types in phase B [51]. Their work focused on experimenting with different Transformer models and embeddings, namely: ELMo, ELMo-Pumbed, BERT and BioBERT.

They used different strategies depending on the question type, such as ensembles on yes/no questions, biomedical named entity extraction (using SciSpacy [39]) on list questions and different pre-/post-processing procedures.

In this challenge too, the open source OAQA system proposed by [63] 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 [60], C-Value and LingPipe [4] 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 7a

Each of the three batches of Task 7a were evaluated independently. The classification performance of the systems were measured using flat and hierarchical evaluation measures [5]. The micro F-measure (MiF) and the Lowest Common Ancestor F-measure (LCA-F) were used to choose the winners for each batch [20].

According to [8] 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 7a show that in all test batches and for both flat and hierarchical measures, some systems outperform the strong baselines. In particular, The "*MeSHProbeNet-P*" systems achieve the best performance in the first batch, outperformed by the "*DeepMeSH*" systems in the last two batches. More detailed results can be found in the online results page[2]. 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 Fig. 1.

4.2 Task 7b

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

[2] http://participants-area.bioasq.org/results/7a/.

Table 5. Average system ranks across the batches of the Task 7a. 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.

System	Batch 1		Batch 2		Batch 3	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
DeepMeSH5	-	-	1,00	1,00	1	1
DeepMeSH4	-	-	9,50	9,50	2,25	1,75
DeepMeSH3	8,25	8,50	3,50	5,00	2,5	2,75
DeepMeSH1	5,00	6,25	2,00	2,63	3,75	4,13
DeepMeSH2	7,25	7,25	3,50	4,50	4,75	4,38
MeSHProbeNet-P2	2,63	2,63	4,63	5,88	6,5	8,25
MeSHProbeNet-P1	3,25	2,13	6,38	4,25	6,88	6,5
MeSHProbeNet-P3	5,00	4,63	8,38	7,25	7,5	7,38
MeSHProbeNet-P	2,38	3,25	7,00	4,38	8,13	7,75
MeSHProbeNet-P0	1,50	1,25	6,25	5,63	8,75	7,88
ceb 1 ensemble	-	-	-	-	11	11
Default MTI	9,75	8,75	12,00	11,75	12,25	12,25
ceb1	8,75	9,25	11,00	11,25	12,25	13,5
MTI First Line Index	11,50	11,25	13,00	12,50	13,25	12
iria-mix	-	-	14,00	14,00	14,5	14,75
Semantic NoSQL KE 2	-	-	-	-	16	16
Semantic NoSQL KE 1	-	-	-	-	17	17,75

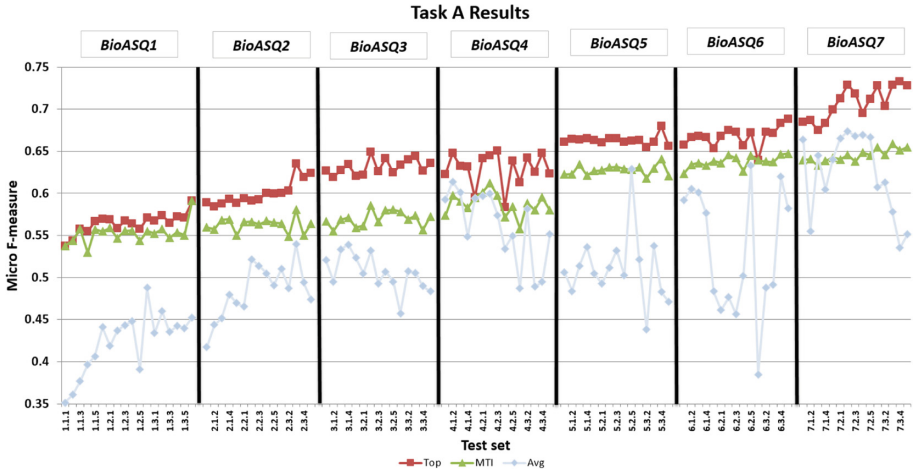

Fig. 1. The micro f-measure achieved by systems across different years of the BioASQ challenge. For each test set the micro F-measure is presented for the best performing system (Top) and the MTI, as well as the average micro f-measure of all the participating systems (Avg).

Table 6. Results for snippet retrieval in batch 4 of phase A of Task 7b.

System	Mean precision	Mean recall	Mean F-measure	MAP	GMAP
aueb-nlp-2	0.2060	0.4039	0.2365	**0.2114**	0.0075
aueb-nlp-1	0.2124	0.4083	0.2440	0.2086	0.0065
aueb-nlp-5	**0.2157**	**0.4235**	**0.2467**	0.1821	**0.0098**
MindLab QA Reloaded	0.1587	0.2760	0.1723	0.1527	0.0013
Deep ML methods for	0.1331	0.2692	0.1589	0.1234	0.0009
MindLab Red Lions++	0.1371	0.2538	0.1535	0.1187	0.0014
aueb-nlp-3	0.1488	0.3427	0.1779	0.1149	0.0053
MindLab QA System ++	0.1288	0.2049	0.1364	0.1136	0.0010
aueb-nlp-4	0.1520	0.3237	0.1791	0.1116	0.0056
MindLab QA System	0.1297	0.2536	0.1478	0.1094	0.0016
lh_sys1	0.0399	0.0810	0.0478	0.0178	0.0001
lh_sys3	0.0233	0.0437	0.0266	0.0151	0.0001
lh_sys5	0.0233	0.0437	0.0266	0.0151	0.0001
lh_sys4	0.0233	0.0437	0.0266	0.0148	0.0001
lh_sys2	0.0182	0.0281	0.0193	0.0051	0.0001

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

System	Mean precision	Mean recall	Mean F-measure	MAP	GMAP
aueb-nlp-4	0.1750	**0.6266**	0.2471	**0.1199**	0.0151
aueb-nlp-2	0.1740	0.6139	0.2449	0.1121	0.0156
aueb-nlp-5	**0.3599**	0.6128	**0.4034**	0.1102	**0.0164**
aueb-nlp-1	0.1700	0.5912	0.2380	0.1041	0.0118
auth-qa-1	0.2675	0.3896	0.2894	0.1033	0.0018
aueb-nlp-3	0.1600	0.5806	0.2266	0.0986	0.0104
lh_sys4	0.1420	0.5490	0.2081	0.0920	0.0069
Ir_sys1	0.1410	0.5365	0.2059	0.0907	0.0059
lh_sys1	0.1420	0.5449	0.2076	0.0881	0.0063
MindLab QA Reloaded	0.1330	0.5288	0.1950	0.0863	0.0062

as the average of the individual rankings in the different categories. In Tables 6 and 7 some indicative results from batches 3 and 4 are presented. Full results are available in the online results page of Task 7b, phase A[3]. These results are

[3] http://participants-area.bioasq.org/results/7b/phaseA/.

Table 8. Results for batch 5 for exact answers in phase B of Task 7b. Only the top-10 systems are presented along with the BioASQ baseline.

System	Yes/No		Factoid			List		
	Acc.	F1	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
BioBERT-DMIS-3	**0.8286**	**0.8250**	**0.2857**	0.4286	0.3452	**0.5653**	0.4131	**0.4619**
BioBERT-DMIS	0.8000	0.7822	0.2571	0.4571	0.3224	0.5236	0.3714	0.4202
unipi-quokka-QA-5	0.8000	0.7939	0.0857	0.1714	0.1152	0.1713	**0.5873**	0.2537
BioBERT-DMIS-2	0.7429	0.7200	0.2571	0.4571	0.3271	0.5486	0.3992	0.4468
BioBERT-DMIS-4	0.7429	0.7351	0.2286	0.4571	0.3238	0.5069	0.3575	0.4051
google-gold-input-ab	0.7143	0.6941	0.2286	0.2857	0.2571	0.1774	0.4175	0.2415
unipi-quokka-QA-4	0.7143	0.6941	0.0857	0.1714	0.1152	0.1713	**0.5873**	0.2537
unipi-quokka-QA-3	0.6857	0.6578	0.0857	0.1714	0.1152	0.1713	**0.5873**	0.2537
google-gold-input	0.6571	0.6023	**0.2857**	0.3714	0.3167	0.2159	0.4452	0.2824
DMIS	0.6571	0.6023	**0.2857**	**0.5143**	**0.3638**	0.5050	0.3714	0.4124
BioASQ_Baseline	0.4857	0.4643	0.0571	0.1429	0.0867	0.2127	0.3619	0.2573

preliminary. The final results for Task 7b, phase A will be available after the manual assessment of the system responses.

Phase B: In phase B of Task 7b 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 [5]. Regarding exact answers[4], 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 last batch of Task 7b. These results are preliminary. The full results of phase B of Task 7b are available online[5]. The final results for Task 7b, phase B will be available after the manual assessment of the system responses.

The results presented in Fig. 2 show that this year the performance of systems in the yes/no questions, has clearly improved. In batch 5 for example, presented in Table 8, some systems outperformed the strong baseline based on previous versions of the OAQA system, with the top system achieving almost double the score of the baseline. Some improvement is also observed in the performance of the top systems for factoid and list questions in the preliminary results. However, there is even more room for improvement in these types of question as can be seen in Fig. 2.

[4] For summary questions, no exact answers are required.
[5] http://participants-area.bioasq.org/results/7b/phaseB/.

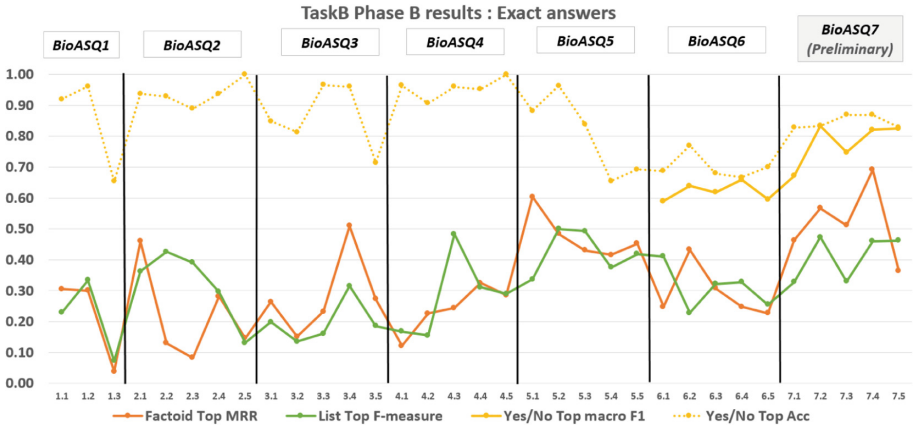

Fig. 2. The performance achieved by systems in exact answer generation part of Task B, Phase B, across different years of the BioASQ challenge. For each test set the performance of the best performing system (Top) is presented based on the official evaluation measures. Since BioASQ6 the macro-averaged F1 score (macro F1) is the official measure for Yes/No questions, but accuracy (Acc), the former official measure, is also presented. The results for BioASQ7 are preliminary. The final results for Task 7b, phase B will be available after the manual assessment of the system responses.

5 Conclusions

In this paper, an overview of the seventh 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, the shift towards systems that incorporate ideas based on deep learning models observed in the previous year, is even more clear. Novel ideas have been tested and state-of-the-art deep learning methodologies have been adapted to biomedical question answering with great results. Specifically, the breakthroughs in different NLP tasks using clever techniques with the advent of new language-models, such as BERT and gpt-2, gave birth to new approaches that significantly boost the performance of the systems. In the future, we expect novel methodologies, such as the newly proposed XLNet [62], to further cultivate research in the biomedical information systems field. Consequently, we believe that the challenge is successfully pushing the research frontier of this domain. In future editions of the challenge, we aim to provide even more benchmark data derived from a community-driven acquisition process.

Acknowledgments. Google was a proud sponsor of the BioASQ Challenge in 2018. The seventh edition of BioASQ is also sponsored by the Atypon Systems inc. BioASQ is grateful to NLM for providing baselines for task 7a and to the CMU team for providing the baselines for task 7b. Finally, we would also like to thank all teams for their participation.

References

1. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: Proceedings of the Fifth ACM Conference on Digital Libraries, DL 2000, pp. 85–94. ACM, New York (2000). https://doi.org/10.1145/336597.336644
2. Alberti, C., Lee, K., Collins, M.: A BERT baseline for the natural questions. arXiv preprint arXiv:1901.08634 (2019)
3. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. **17**, 229–236 (2010)
4. Baldwin, B., Carpenter, B.: Lingpipe (2003). Available from World Wide Web. http://alias-i.com/lingpipe
5. Balikas, G., et al.: Evaluation framework specifications. Project deliverable D4.1, UPMC, May 2013 (2013)
6. Brokos, G.I., Liosis, P., McDonald, R., Pappas, D., Androutsopoulos, I.: AUEB at BioASQ 6: Document and Snippet Retrieval, September 2018. http://arxiv.org/abs/1809.06366
7. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading Wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051 (2017)
8. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dhingra, B., Mazaitis, K., Cohen, W.W.: Quasar: datasets for question answering by search and reading. arXiv preprint arXiv:1707.03904 (2017)
11. Dimitriadis, D., Tsoumakas, G.: Word embeddings and external resources for answer processing in biomedical factoid question answering. J. Biomed. Inform. **92**, 103118 (2019). https://doi.org/10.1016/j.jbi.2019.103118
12. Esuli, A., Sebastiani, F.: SENTIWORDNET: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation, LREC 2006, pp. 417–422 (2006)
13. Ferrucci, D., et al.: Building Watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
14. Gardner, M., et al.: AllenNLP: a deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640 (2017)
15. Gormley, C., Tong, Z.: Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. O'Reilly Media, Inc., Sebastopol (2015)
16. Hosein, S., Andor, D., Mcdonald, R.: Measuring domain portability and error propagation in biomedical QA. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
17. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
18. Jain, H., Prabhu, Y., Varma, M.: Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2016, pp. 935–944. ACM Press, New York (2016). https://doi.org/10.1145/2939672.2939756

19. Kamath, S., Grau, B., Ma, Y.: How to pre-train your model? Comparison of different pre-training models for biomedical question answering. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
20. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Min. Knowl. Disc. **29**(3), 820–865 (2015)
21. Kumar, A., et al.: Ask me anything: dynamic memory networks for natural language processing. In: International Conference on Machine Learning, pp. 1378–1387 (2016)
22. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning, pp. 957–966 (2015)
23. Kwiatkowski, T., et al.: Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguist. **7**, 453–466 (2019). https://www.mitpressjournals.org/doi/full/10.1162/tacl_a_00276
24. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents, May 2014. http://arxiv.org/abs/1405.4053
25. Lee, J., et al.: BioBERT: pre-trained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746 (2019)
26. Lin, Y., Ji, H., Liu, Z., Sun, M.: Denoising distantly supervised open-domain question answering. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1736–1745 (2018)
27. Liu, J., Chang, W.C., Wu, Y., Yang, Y.: Deep learning for extreme multi-label text classification. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 115–124. ACM (2017)
28. Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., Zhu, S.: MeSHLabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. Bioinformatics **31**(12), i339–i347 (2015)
29. McDonald, R., Brokos, G.I., Androutsopoulos, I.: Deep relevance ranking using enhanced document-query interactions, September 2018. http://arxiv.org/abs/1809.01682
30. McDonald, R., Brokos, G.I., Androutsopoulos, I.: Deep relevance ranking using enhanced document-query interactions. arXiv preprint arXiv:1809.01682 (2018)
31. Molla, D.: Macquarie University at BioASQ 5B query-based summarisation techniques for selecting the ideal answers. In: Proceedings BioNLP 2017 (2017)
32. Molla, D.: Macquarie University at BioASQ 6B: deep learning and deep reinforcement learning for query-based summarisation. In: Proceedings of the 6th BioASQ Workshop. A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, pp. 22–29 (2018)
33. Molla, D., Jones, C.: Classification betters regression in query-based multi-document summarisation techniques for question answering. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
34. Mollá-Aliod, D.: 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, pp. 103–107 (2017)
35. Mork, J.G., Demner-Fushman, D., Schmidt, S.C., Aronson, A.R.: Recent enhancements to the NLM medical text indexer. In: Proceedings of Question Answering Lab at CLEF (2014)

36. Müller, B., Poley, C., Pössel, J., Hagelstein, A., Gübitz, T.: LIVIVO – the vertical search engine for life sciences. Datenbank-Spektrum **17**(1), 29–34 (2017). https://doi.org/10.1007/s13222-016-0245-2
37. Mller, B., Rebholz-Schuhmann, D.: Selected approaches ranking contextual term for the BioASQ multi-label classification (Task6a and 7a). In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
38. Nentidis, A., Bougiatiotis, K., Krithara, A., Paliouras, G., Kakadiaris, I.: Results of the fifth edition of the BioASQ challenge. In: BioNLP 2017, pp. 48–57 (2017)
39. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327. Association for Computational Linguistics, Stroudsburg (2019). https://doi.org/10.18653/v1/W19-5034, https://www.aclweb.org/anthology/W19-5034
40. Nunes, T., Campos, D., Matos, S., Oliveira, J.L.: BeCAS: biomedical concept recognition services and visualization. Bioinformatics **29**(15), 1915–1916 (2013). https://doi.org/10.1093/bioinformatics/btt317
41. Oita, M., Vani, K., Oezdemir-Zaech, F.: Semantically corroborating neural attention for biomedical question answering. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
42. Pappas, D., McDonald, R., Brokos, G.I., Androutsopoulos, I.: AUEB at BioASQ 7: document and snippet retrieval. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
43. Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., Zhu, S.: DeepMeSH: deep semantic representation for improving large-scale mesh indexing. Bioinformatics **32**(12), i70–i79 (2016)
44. Peng, S., You, R., Xie, Z., Zhang, Y., Zhu, S.: The Fudan participation in the 2015 BioASQ challenge: large-scale biomedical semantic indexing and question answering. In: CEUR Workshop Proceedings, vol. 1391. CEUR Workshop Proceedings (2015)
45. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
46. Peters, M.E., et al.: Deep contextualized word representations, February 2018. http://arxiv.org/abs/1802.05365
47. Pineda-Vargas, M., Rosso-Mateus, A., Gonzlez, F., Montes-Y-Gmez, M.: A mixed information source approach for biomedical question answering: MindLab at BioASQ 7B. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
48. Rae, A., Mork, J., Demner-Fushman, D.: Convolutional neural network for automatic MeSH indexing. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
49. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
50. Reddy, S., Chen, D., Manning, C.D.: CoQA: a conversational question answering challenge. Trans. Assoc. Comput. Linguist. **7**, 249–266 (2019)
51. Resta, M., Arioli, D., Fagnani, A., Attardi, G.: Transformer models for question answering at BioASQ 2019. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)

52. Ribadas-Pena, F.J., de Campos, L.M., Bilbao, V.M.D., Romero, A.E.: 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, 8–11 September 2015 (2015). http://ceur-ws.org/Vol-1391/84-CR.pdf
53. Robertson, S.E., Jones, K.S.: Relevance weighting of search terms. J. Am. Soc. Inf. Sci. **27**(3), 129–146 (1976)
54. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368 (2017)
55. Soldaini, L., Goharian, N.: QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR (2016)
56. Tanenblatt, M.A., Coden, A., Sominsky, I.L.: The conceptmapper approach to named entity recognition. In: LREC (2010)
57. Telukuntla, S.K., Kapri, A., Zadrozny, W.: UNCC biomedical semantic question answering systems. BioASQ: Task-7B, Phase-B. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
58. Tsatsaronis, G., et al.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics **16**, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6
59. Tsoumakas, G., Laliotis, M., Markontanatos, N., Vlahavas, I.: Large-scale semantic indexing of biomedical publications. In: 1st BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2013)
60. Wei, C.H., Leaman, R., Lu, Z.: Beyond accuracy: creating interoperable and scalable text-mining web services. Bioinformatics (Oxford, England) **32**(12), 1907–1910 (2016). https://doi.org/10.1093/bioinformatics/btv760
61. Xun, G., Jha, K., Yuan, Y., Wang, Y., Zhang, A.: MeSHProbeNet: a self-attentive probe net for MeSH indexing. Bioinformatics, 1–8 (2019). https://doi.org/10.1093/bioinformatics/btz142
62. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. CoRR abs/1906.08237 (2019). http://arxiv.org/abs/1906.08237
63. Yang, Z., Zhou, Y., Eric, N.: Learning to answer biomedical questions: OAQA at BioASQ 4B. In: ACL 2016, p. 23 (2016)
64. Yin, W., Schütze, H., Xiang, B., Zhou, B.: ABCNN: attention-based convolutional neural network for modeling sentence pairs. Trans. Assoc. Comput. Linguist. **4**, 259–272 (2016)
65. Yoon, W., Lee, J., Kim, D., Jeong, M., Kang, J.: Pre-trained language model for biomedical question answering. In: Seventh BioASQ Workshop: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering (2019)
66. You, R., Dai, S., Zhang, Z., Mamitsuka, H., Zhu, S.: AttentionXML: extreme multi-label text classification with multi-label attention based recurrent neural networks, pp. 1–16, November 2018. http://arxiv.org/abs/1811.01727
67. Zavorin, I., Mork, J.G., Demner-Fushman, D.: Using learning-to-rank to enhance NLM medical text indexer results. In: ACL 2016, p. 8 (2016)