Overview of BioASQ 8a and 8b: Results of the eighth edition of the BioASQ tasks a and b

Anastasios Nentidis[1,2], Anastasia Krithara[1], Konstantinos Bougiatiotis[1,3], and Georgios Paliouras[1]

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

Abstract. In this paper, we present an overview of the eighth edition of the tasks a and b of the BioASQ challenge, which ran as a lab in the Conference and Labs of the Evaluation Forum (CLEF) 2020. BioASQ aims at promoting methodologies and systems for large-scale biomedical semantic indexing and question answering through the organization of yearly challenges since 2012. These shared tasks offer to teams around the world the opportunity to develop and compare their methods on the same benchmark datasets that represent the demanding information needs of biomedical experts. This year, apart from introduction of a new task on medical semantic indexing in Spanish (MESINESP8), the eighth versions of the two established BioASQ tasks on semantic indexing (8a) and question answering (8b) in English were also offered. In total, 34 teams with more than 100 systems participated in the three tasks of the challenge, with seven of them focusing on task 8a and 23 on task 8b. As in previous versions of the tasks, the evaluation of system responses reveals some participating systems managed to outperform the strong baselines, indicating that continuous advancements in state-of-the-art systems keep pushing the frontier of research leading to performance improvements.

Keywords: Biomedical knowledge · Semantic Indexing · Question Answering

1 Introduction

This paper presents the shared tasks 8a and 8b of the eighth edition of the BioASQ challenge in 2020, the corresponding datasets and the approaches and achieved results of the participating systems. A detailed description of the new task on medical indexing in Spanish is offered in the MESINESP task overview. A condensed BioASQ 2020 Lab overview [2] is also available, describing the

---

eighth edition the BioASQ challenge as a whole, in the context of the Conference and Labs of the Evaluation Forum (CLEF) 2020. Towards this direction, in section 2 we provide an overview of the shared tasks 8a and 8b, that took place from February to May 2020, as well as the corresponding datasets developed for training and testing the participating systems. In section 3, we briefly overview the participating systems and the approaches proposed by the corresponding teams for these two tasks. Detailed descriptions for some of the systems are also available in the proceedings of the BioASQ lab. In section 4, we present the results of the evaluation of participating systems, based on manual assessment or state-of-the-art evaluation measures, depending on the nature of the required system response. Finally, we conclude and discuss the eighth version of the BioASQ tasks a and b in section 5.

2 Overview of the Tasks

In the eighth version of the BioASQ challenge were offered three tasks: (1) a large-scale biomedical semantic indexing task (task 8a), (2) a biomedical question answering task (task 8b), both considering documents in English, and (3) a new task on medical semantic indexing in Spanish (task MESINESP). In this section we provide a brief description of the two established tasks (8a and 8b) with focus on differences from previous versions of the challenge [32]. A detailed overview of the initial versions of the tasks and the general structure of BioASQ is also already available [46].

2.1 Large-scale semantic indexing - Task 8a

In Task 8a the aim is to classify articles from the PubMed/MedLine[4] digital library into concepts of the MeSH hierarchy. In particular, new PubMed articles that are not yet annotated by the indexers in NLM are gathered to form the test sets for the evaluation of the participating systems. Some basic details about each test set and batch are provided in Table 1. As done in previous versions of the task, the task is divided into three independent batches of 5 weekly test sets each, providing an on-line and large-scale scenario, and the test sets consist of new articles without any restriction on the journal published. The performance of the participating systems is calculated using standard flat information retrieval measures, as well as, hierarchical ones, when the annotations from the NLM indexers become available. As usual, participants have 21 hours to provide their answers for each test set. However, as it has been observed that new MeSH annotations are released in PubMed earlier that in previous years, we shifted the submission period accordingly to avoid having some annotations available from NLM while the task is still running. For training, a dataset of 14,913,939 articles with 12.68 labels per article, on average, was provided to the participants.

---

[4] https://pubmed.ncbi.nlm.nih.gov/

Batch	Articles	Annotated Articles	Labels per Article
1	6510	6487	12.49
	7126	7074	12.27
	10891	10789	12.55
	6225	6182	12.28
	6953	6887	12.75
Total	37705	37419	0.99
2	6815	6787	12.49
	6485	6414	12.52
	7014	6975	11.92
	6726	6647	12.90
	6379	6246	12.45
Total	33419	33069	0.99
3	6842	6601	12.70
	7212	6456	12.37
	5430	4764	12.59
	6022	4858	12.33
	5936	3999	12.21
Total	31442	26678	0.85

Table 1. Statistics on test datasets for Task 8a.

2.2 Biomedical semantic QA - Task 8b

Task 8b aims at providing a realistic large-scale question answering challenge offering to the participating teams the opportunity to develop systems for all the stages of question answering in the biomedical domain. Four types of questions are considered in the task: "yes/no", "factoid", "list" and "summary" questions [4]. A training dataset of 3,243 questions annotated with golden relevant elements and answers is provided for the participants to develop their systems. Table 2 presents some statistics about the training dataset as well as the five test sets.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	3,243	881	644	941	777	10.15	12.92
Test 1	100	25	20	32	23	3.45	4.51
Test 2	100	36	14	25	25	3.86	5.05
Test 3	100	31	12	28	29	3.35	4.71
Test 4	100	26	17	34	23	3.23	4.38
Test 5	100	34	12	32	22	2.57	3.20
Total	3,743	1033	719	1092	899	9.23	11.78

Table 2. Statistics on the training and test datasets of Task 8b. The numbers for the documents and snippets refer to averages per question.

As in previous versions of the challenge, the task is structured into two phases that focus on the retrieval of the required information (phase A) and answering the question (phase B). In addition, the task is split into five independent bi-weekly batches and the two phases for each batch run during two consecutive days. In each phase, the participants receive the corresponding test set and have 24 hours to submit the answers of their systems. In particular, in phase A, a test set of 100 questions written in English is released and the participants are expected to identify and submit relevant elements from designated resources, including PubMed/MedLine articles, snippets extracted from these articles, concepts and RDF triples. In phase B, the manually selected relevant articles and snippets for these 100 questions are also released and the participating systems are asked to respond with *exact answers*, that is entity names or short phrases, and *ideal answers*, that is natural language summaries of the requested information.

3 Overview of participation

Fig. 1. The world-wide distribution of teams participating in the tasks 8a and 8b, based on institution affiliations.

This year, 34 teams from institutes around the world participated in the three tasks of the challenge with more than 100 distinct systems. Seven of these teams focused on task 8a and 23 on task 8b. As presented in fig. 1, the institutions hosting the teams that participated in tasks 8a and 8b are distributed around the world highlighting the international interest in the tasks. Compared to previous versions of the challenge, we observe a shift towards the most complex question answering task b, where the number of participating teams and systems is increasing during the last years as shown in Fig. 2.

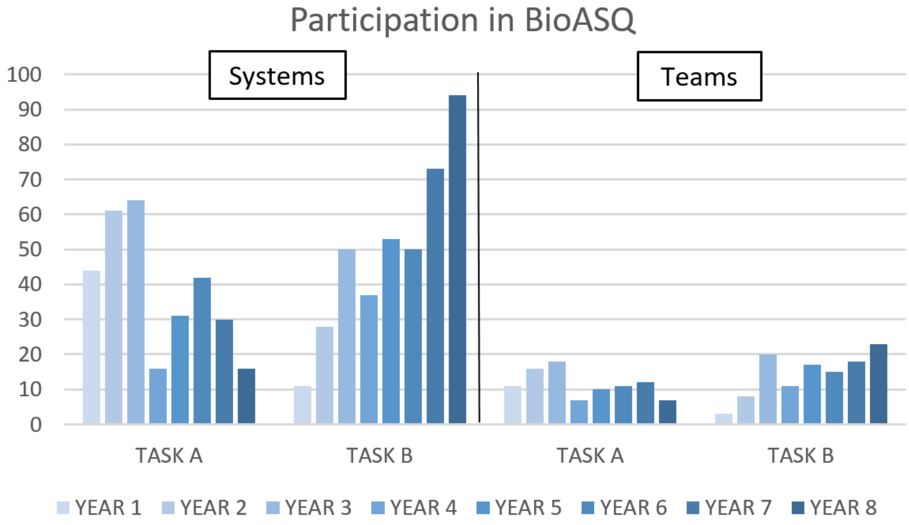

Fig. 2. The evolution of participation in the BioASQ tasks a and b until their current eighth version.

3.1 Task 8a

This year, 7 teams participated in the eighth edition of task a, submitting predictions from 16 different systems in total. Here, we provide a brief overview of those systems for which a description was available, stressing their key characteristics. A summing-up of the participating systems and corresponding approaches is presented in Table 3.

System	Approach
X-BERT BioASQ	X-BERT, Transformers, ELMo, MER
NLM CNN	SentencePiece, CNN, embeddings, ensembles
dmiip_fdu	d2v, tf-idf, SVM, KNN, LTR, DeepMeSH, AttentionXML, BERT, PLT, BERTMeSH
Iria	Lucene Index, k-NN, stem bigrams, ensembles, UIMA ConceptMapper

Table 3. Systems and approaches for Task 8a. Systems for which no description was available at the time of writing are omitted.

This year, the LASIGE team from the University of Lisboa, in its "X-BERT BioASQ" system propose a novel approach for biomedical semantic indexing combining a solution based on Extreme Multi-Label Classification (XMLC) with a Named-Entity-Recognition (NER) tool. In particular, their system is based on X-BERT [8], an approach to scale BERT [14] to XMLC, combined with the use of the MER [12] tool to recognize MeSH terms in the abstracts of the articles. The system is structured into three steps. The first step is the semantic indexing of the labels into clusters using ELMo [39]; then a second step matches the indices using a Transformer architecture; and finally, the third step focuses on ranking the labels retrieved from the previous indices.

Other teams, improved upon existing systems already participating in previous versions of the task. Namely, the National Library of Medicine (NLM) team, in its "*NLM CNN*" system enhance the previous version of their "*ceb*" systems [40], based on an end-to-end Deep Learning (DL) architecture with Convolutional Neural Networks (CNN), with SentencePiece tokenization [24]. The Fudan University team also builds upon their previous "*AttentionXML*" [55] and "*DeepMeSH*" [38] systems as well their new "*BERTMeSH*" [54] system, which are based on document to vector (d2v) and tf-idf feature embeddings, learning to rank (LTR) and DL-based extreme multi-label text classification, Attention Mechanisms and Probabilistic Label Trees (PLT) [18]. Finally, this years versions of the "*Iria*" systems [43] are also based on the same techniques used by the systems in previous versions of the challenge which are summarized in Table 3.

Similarly to the previous versions of the challenge, two systems developed by NLM to facilitate the annotation of articles by indexers in MedLine/PubMed, where available as baselines for the semantic indexing task. MTI [31] as enhanced in [56] and an extension based on features suggested by the winners of the first version of the task [47].

3.2 Task 8b

This version of Task b was tackled by 94 different systems in total, developed by 23 teams. In particular, 8 teams participated in the first phase, on the retrieval of relevant material required for answering the questions, submitting results from 30 systems. In the second phase, on providing the exact and ideal answers for the questions, participated 18 teams with 72 distinct systems. Three of the teams participated in both phases. An overview of the approaches, technologies and datasets used by the teams is provided in Table 4 and a graphical representation of them as a word cloud, weighted by their frequency in logarithmic scale, is also provided in Fig. 3. Only systems for which a description was available is included in this section. Detailed descriptions for some of the systems are available at the proceedings of the workshop.

The "*ITMO*" team participated in both phases of the task experimenting in its "pa" systems with differing solutions across the batches. In general, for document retrieval the systems follow an approach with two stages. First, they identify initial candidate articles based on BM25, and then they re-rank them using variations of BERT [14], fine-tuned for the binary classification task with the BioASQ dataset and pseudo-negative documents. They extract snippets from the top documents and rerank them using biomedical Word2Vec based on cosine similarity with the question. To extract exact answers they use BERT fine-tuned on SQUAD [41] and BioASQ datasets and employ a post-processing to split

Systems	**Phase**	**Approach**
pa	A (documents, snippets) B (exact, ideal)	BM25, BERT, Word2Vec, SQuAD, PubMedQA, BioMed-RoBERTa
bio-answerfinder	A (documents, snippets) B (exact, ideal)	Bio-AnswerFinder, LSTM, ElasticSearch, BERT, Electra, BioBERT, SQuAD, wRWMD
AUEB	A (documents, snippets) B (exact)	BM25, Word2Vec, Graph-Node Embeddings, SciBERT, DL (JPDRMM)
bioinfo	A (documents, snippets)	BM25, ElasticSearch, distant learning, DeepRank
Google	A (documents)	BM25, BioBERT, Synthetic Query Generation, BERT, reranking
KU-DMIS	B (exact, ideal)	BioBERT, NLI, MultiNLI, SQuAD, BART, beam search, BERN, language_check
NCU-IISR	B (exact, ideal)	BioBERT, logistic regression, LTR
UoT	B (exact)	BioBERT, multi-task learning, BC2GM
BioNLPer	B (exact)	BioBERT, multi-task learning, NLTK, ScispaCy
LabZhu	B (exact)	BERT, BioBERT, XLNet, SpanBERT, transfer learning, SQuAD, ensembling
umass_czi	B (exact)	BioBERT, SciBERT, BioSentVec, Pubtator, SQuAD, PubMedQA, transfer learning
MQ	B (ideal)	Word2Vec, BERT, LSTM, Reinforcement Learning (PPO)
DAIICT	B (ideal)	textrank, lexrank, UMLS
sbert	B (ideal)	Sentence-BERT, BioBERT, SNLI, MutiNLI, multi-task learning, MQU

Table 4. Systems and approaches for Task8b. Systems for which no information was available at the time of writing are omitted.

the answer for list questions and additional fine-tuning on PubMedQA [19] for yes/no questions. Finally, for ideal answers they generate some candidates from the snippets and their sentences and rerank them using the model used for phase A. In the last batch, they also experiment with generative summarization, developing a model based on BioMed-RoBERTa [17] to improve the readability and consistency of the produced ideal answers.

Another team participating in both phases of the task is the "*UCSD*" team with its "bio-answerfinder" system. In particular, for phase A they rely on previously developed Bio-AnswerFinder system [34], which is also used as a first step in phase B, for re-ranking the sentences of the snippets provided in the test set. For identifying the exact answers for factoid and list questions they experimented on fine-tuning Electra [10] and BioBERT [25] on SQuAD and BioASQ datasets combined. The answer candidates are then scored considering classification probability, the top ranking of corresponding snippets and number

Fig. 3. A word cloud of the approaches, techniques and datasets used by task 8b participating teams, weighted by their frequency in logarithmic scale.

of occurrences. Finally a normalization and filtering step is performed and, for list questions, and enrichment step based on coordinated phrase detection. For yes/no questions they fine-tune BioBERT on the BioASQ dataset and use majority voting. For summary questions, they employ hierarchical clustering, based on weighted relaxed word mover's distance (wRWMD) similarity [34] to group the top sentences, and select the sentence ranked highest by Bio-AnswerFinder to be concatenated to form the summary.

The "*AUEB*" team also participated in both tasks focusing on phase A and briefly experimenting with Phase B. Working on extending their previous top-performing model [36], they experimented with graph-node embeddings generated from a biomedical entity co-occurrence graph from publications [23]. Moreover, they experimented with new ways to encode and retrieve relevant snippets, but concluded that conventional BM25 pre-fetching was more efficient. For Phase B, they worked with exact answer extraction. To this end, they experimented with a SciBERT-based model [5] modelled for cloze-style biomedical machine reading comprehension [37] (MRC). However, their initial results indicated that the MRC task differs greatly from the exact answer extraction task and they did not pursue this research direction further.

In phase A, the team from the University of Aveiro participated with its "bioinfo" systems, which consists of a fine-tuned BM25 retrieval model based on ElasticSearch [16], followed by a neural reranking step. For the latter, they use an interaction-based model inspired on the DeepRank [35] architecture building upon previous versions of their system [1]. The focus of the improvements was on the sentence splitting strategy, on extracting of multiple relevance signals, and the independent contribution of each sentence for the final score. The "*Google*" team also participated in phase A, with four distinct systems for document retrieval based on different approaches. In particular, they used a BM25 retrieval model, a neural retrieval model, initialized with BioBERT and trained on a large set of questions developed through Synthetic Query Generation (QGen), and a hybrid retrieval model [5] based on a linear blend of BM25 and the neural model [28]. In addition, they also used a reranking model, rescoring the results of the hybrid model with a cross-attention BERT rescorer [36].

In phase B, this year the "*KU-DMIS*" team participated on both exact and ideal answers. For exact answers, they build upon their previous BioBERT-based systems [53] and try to adapt the sequential transfer learning of Natural Language Inference (NLI) to biomedical question answering. In particular, they investigate whether learning knowledge of entailment between two sentence pairs can improve exact answer generation, enhancing their BioBERT-based models with alternative fine-tuning configurations based on the MultiNLI dataset [50]. For ideal answer generation, they develop a deep neural abstractive summarization model based on BART [26] and beam search, with particular focus on pre-processing and post-processing steps. In particular, alternative systems were developed either considering the answers predicted by the exact answer prediction system in their input or not. In the post-processing step, the generated candidate ideal answers for each question where scored using the predicted exact answers and some grammar scores provided by the language_check tool[6]. For factoid and list questions in particular, the BERN [21] tool was also employed to recognize named entities in the candidate ideal answers for the scoring step.

The "*NCU-IISR*" team also participated in both parts of phase B, constructing two BioBERT-based models for extracting the exact answer and ranking the ideal answers respectively. The first model is fine-tuned on the BioASQ dataset formulated as a SQuAD-type QA task that extracts the answer span. For the second model, they regard the sentences of the provided snippets as candidate ideal answers and build a ranking model with two parts. First, a BioBERT-based model takes as input the question and one of the snippet sentences and provides their representation. Then, a logistic regressor, trained on predicting the similarity between a question and each snippet sentence, takes this representation and outputs a score, which is used for selecting the final ideal answer.

The "*UoT*" team participated with three different DL approaches for generating exact answers. In their first approach, they fine-tune separately two distinct BioBERT-based models extended with an additional neural layer depending on the question type, one for yes/no and one for factoid and list questions together. In their second system, they use a joint-learning setting, where the same BioBERT layer is connected with both the additional layers and jointly trained for all types of questions. Finally, in their third system they propose a multi-task model to learn recognizing biomedical entities and answers to questions simulta-

---

[5] https://ai.googleblog.com/2020/05/an-nlu-powered-tool-to-explore-covid-19.html

[6] https://pypi.org/project/language-check/

neously, aiming at transferring knowledge from the biomedical entity recognition task to question answering. In particular, they extend their joint BioBERT-based model with simultaneous training on the BC2GM dataset [45] for recognizing gene and protein entities.

The "*BioNLPer*" team also participated in the exact answers part of phase B, focusing on factoids. They proposed 5 BioBERT-based systems, using external feature enhancement and auxiliary task methodologies. In particular, in their "factoid qa model" and "Parameters retrained" systems they consider the prediction of answer boundaries (start and end positions) as the main task and the whole answer content prediction as an auxiliary task. In their "Features Fusion" system they leveraged external features including NER and part-of-speach (POS) extracted by NLTK [27] and ScispaCy [33] tools as additional textual information and fused them with the pre-trained language model representations, to improve answer boundary prediction. Then, in their "BioFusion" system they combine the two methodologies together. Finally, their "BioLabel" system employed the general and biomedical domain corpus classification as the auxiliary task to help answer boundary prediction.

The "LabZhu" systems participated in phase B as well, with focus on exact answers for the factoid and list questions. They treat answer generation as an extractive machine comprehension task and explore several different pretrained language models, including BERT, BioBERT, XLNet [51] and SpanBERT [20]. They also follow a transfer learning approach, training the models on the SQuAD dataset, and then fine-tuning them on the BioASQ datasets. Finally, they also rely on voting to integrate the results of multiple models. The "umass_czi" team also focused on the exact answer part of phase B, experimenting with unsupervised representation learning approaches in the context of Biomedical QA. In particular, they considered pretrained representations based on BioBERT, SciBERT, and BioSentVec [9] and experimented with transferring knowledge from the SQuAD and PubMedQA datasets in to the BioASQ 8b QA task. Finally, they also develop a new pre-training method based on a self-supervised de-noising approach. In this method, they first generate a QA dataset randomly replacing entities automatically recognized by PubTator [48] in PubMed abstracts. Then, train their model on extracting the span of the new entities given the original ones as a queries.

The "*MQ*" team, as in past years, focused on ideal answers, approaching the task as query-based summarisation. In some of their systems the retrain their previous classification and regression approaches [30] in the new training dataset. In addition, they also employ reinforcement learning with Proximal Policy Optimization (PPO) [44] and two variants to represent the input features, namely Word2Vec-based and BERT-based embeddings. The "*DAIICT*" team also participated in ideal answer generation, using the standard extractive summarization techniques textrank [29] and lexrank [15] as well as sentence selection techniques based on their similarity with the query. They also modified these techniques investigating the effect of query expansion based on UMLS [6] for sentence selection and summarization.

Finally, the "*sbert*" team, also focused on ideal answers. They experimented with different embedding models and multi-task learning in their systems, using parts from previous "*MQU*" systems for the pre-processing of data and the prediction step based on classification and regression [30]. In particular, they used a Universal Sentence Embedding Model [11] (BioBERT-NLI [7]) based on a version of BioBERT fine-tuned on the the SNLI [7] and the MultiNLI datasets as in Sentence-BERT [42]. The features were fed to either a single logistic regression or classification model to derive the ideal answers. Additionally, in a multi-task setting, they trained the model on both the classification and regression tasks, selecting for the final prediction one of them.

In this challenge too, the open source OAQA system proposed by [52] served as baseline for phase B exact answers. 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 [49], C-Value and LingPipe [3] 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 and finally ranking techniques.

4 Results

4.1 Task 8a

System	**Batch 1**		**Batch 2**		**Batch 3**	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
deepmesh_dmiip_fdu	**1.25**	**2.25**	1.875	3.25	2.25	2.25
deepmesh_dmiip_fdu_	2.375	3.625	**1.25**	**1.25**	1.75	2
attention_dmiip_fdu	3	**2.25**	3.5	3.125	3	3.25
Default MTI	4.75	3.75	6	5.25	6	5.5
MTI First Line Index	5.5	4.5	6.75	5.875	5.75	5.25
dmiip_fdu	-	-	2.375	1.625	**1.5**	**1.25**
NLM CNN	-	-	5	6.75	5.5	7
iria-mix	-	-	-	-	8.25	8.25
iria-1	-	-	-	-	9.25	9.25
X-BERT BioASQ	-	-	-	-	10.75	10.75

Table 5. Average system ranks across the batches of the task 8a. A hyphenation symbol (-) is used whenever the system participated in fewer than 4 test sets in the batch. Systems participating in fewer than 4 test sets in all three batches are omitted.

In Task 8a, each of the three batches were independently evaluated as presented in Table 5. Standard flat and hierarchical evaluation measures [4] were

---

[7] https://huggingface.co/gsarti/biobert-nli

used for measuring the classification performance of the systems. In particular, the micro F-measure (MiF) and the Lowest Common Ancestor F-measure (LCA-F) were used to identify the winners for each batch [22]. As suggested by Demšar [13], the appropriate way to compare multiple classification systems over multiple datasets is based on their average rank across all the datasets. In this task, the system with the best performance in a test set gets rank 1.0 for this test set, the second best rank 2.0 and so on. In case two or more systems tie, they all receive the average rank. Then, according to the rules of the challenge, the average rank of each system for a batch is calculated based on the four best ranks of the system in the five test sets of the batch. The average rank of each system, based on both the flat MiF and the hierarchical LCA-F scores, for the three batches of the task are presented in Table 5.

The results in Task 8a show that in all test batches and for both flat and hierarchical measures, the best systems outperform the strong baselines. In particular, the "*dmiip_fdu*" systems from the Fudan University team achieve the best performance in all three batches of the task. More detailed results can be found in the online results page[8]. Comparing these results with the corresponding results from previous versions of the task, suggests that both the MTI baseline and the top performing systems keep improving through the years of the challenge, as shown in Figure 4.

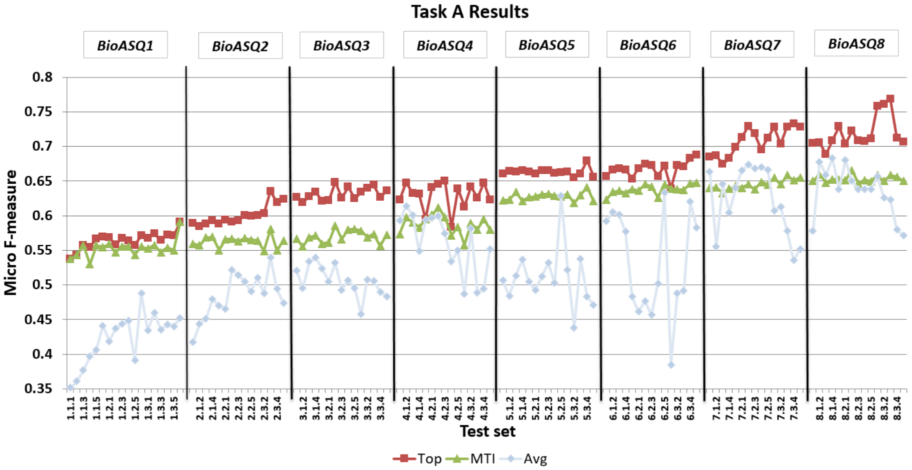

Fig. 4. The micro f-measure (MiF) achieved by systems across different years of the BioASQ challenge. For each test set the MiF score is presented for the best performing system (Top) and the MTI, as well as the average micro f-measure of all the participating systems (Avg).

---

[8] http://participants-area.bioasq.org/results/8a/

4.2 Task 8b

Phase A: In the first phase of Task 8b, the systems are ranked according to the Mean Average Precision (MAP) measure for each of the four types of annotations, namely documents, snippets, concepts and RDF triples. This year, the calculation of Average Precision (AP) in MAP for phase A was reconsidered as described in the official description of the evaluation measures for Task 8b[9]. In brief, since BioASQ3, the participant systems are allowed to return up to 10 relevant items (e.g. documents), and the calculation of AP was modified to reflect this change. However, the number of golden relevant items in the last years have been observed to be lower than 10 in some cases, resulting to relatively small AP values even for submissions with all the golden elements. For this reason, this year, we modified the MAP calculation to consider both the limit of 10 elements and the actual number of golden elements. In Tables 6 and 7 some indicative preliminary results from batch 2 are presented. The full results are available in the online results page of Task 8b, phase A[10]. The results presented here are preliminary, as the final results for the task 8b will be available after the manual assessment of the system responses by the BioASQ team of biomedical experts.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
pa	**0.1934**	0.4501	**0.2300**	**0.3304**	0.0185
AUEB-System1	0.1688	**0.4967**	0.2205	0.3181	0.0165
bioinfo-3	0.1500	0.4880	0.2027	0.3168	**0.0223**
bioinfo-1	0.1480	0.4755	0.1994	0.3149	0.0186
bioinfo-4	0.1500	0.4787	0.2002	0.3120	0.0161
AUEB-System2	0.1618	0.4864	0.2126	0.3103	0.0149
bioinfo-2	0.1420	0.4648	0.1914	0.3084	0.0152
bioinfo-0	0.1380	0.4341	0.1830	0.2910	0.0117
AUEB-System5	0.1588	0.4549	0.2057	0.2843	0.0116
Ir_sys4	0.1190	0.4179	0.1639	0.2807	0.0056
Google-AdHoc-MAGLEV	0.1310	0.4364	0.1770	0.2806	0.0109
Ir_sys2	0.1190	0.4179	0.1639	0.2760	0.0055
Google-AdHoc-BM25	0.1324	0.4222	0.1758	0.2718	0.0088
AUEB-System3	0.1688	**0.4967**	0.2205	0.2702	0.0146
Ir_sys3	0.1325	0.3887	0.1730	0.2678	0.0045

Table 6. Results for document retrieval in batch 2 of phase A of Task 8b. Only the top-15 systems are presented.

Phase B: In the second phase of task 8b, the participating systems were expected to provide both exact and ideal answers. Regarding the ideal answers, the systems will be ranked according to manual scores assigned to them by the BioASQ experts during the assessment of systems responses [4]. For the

---

[9] http://participants-area.bioasq.org/Tasks/b/eval_meas_2020/

[10] http://participants-area.bioasq.org/results/8b/phaseA/

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
AUEB-System1	**0.1545**	0.2531	**0.1773**	**0.6821**	0.0015
AUEB-System2	0.1386	0.2260	0.1609	0.6549	0.0011
pa	0.1348	**0.2578**	0.1627	0.3374	**0.0047**
bioinfo-4	0.1308	0.2009	0.1413	0.2767	0.0016
bioinfo-1	0.1373	0.2103	0.1461	0.2721	0.0016
bioinfo-2	0.1299	0.2018	0.1408	0.2637	0.0011
bioinfo-3	0.1321	0.2004	0.1404	0.2607	0.0014
MindLab QA System	0.0811	0.1454	0.0916	0.2449	0.0005
MindLab Red Lions++	0.0830	0.1469	0.0932	0.2394	0.0005
AUEB-System5	0.0943	0.1191	0.0892	0.2217	0.0011
MindLab QA Reloaded	0.0605	0.1103	0.0691	0.2106	0.0002
Deep ML methods for	0.0815	0.0931	0.0811	0.2051	0.0001
bioinfo-0	0.1138	0.1617	0.1175	0.1884	0.0009
MindLab QA System ++	0.0639	0.0990	0.0690	0.1874	0.0001
AUEB-System3	0.0966	0.1285	0.0935	0.1556	0.0011
bio-answerfinder	0.0910	0.1617	0.1004	0.1418	0.0008
AUEB-System4	0.0080	0.0082	0.0077	0.0328	0.0000

Table 7. Results for snippet retrieval in batch 2 of phase A of Task 8b.

exact answers, which are required for all questions except the summary ones, the measure considered for ranking the participating systems depends on the question type. For the yes/no questions, the systems were ranked according to the macro-averaged F1-measure on prediction of no and yes answer. For factoid questions, the ranking was based on mean reciprocal rank (MRR) and for list questions on mean F1-measure. Some indicative results for exact answers for the third batch of Task 8b are presented in Table 8. The full results of phase B of Task 8b are available online[11]. These results are preliminary, as the final results for Task 8b will be available after the manual assessment of the system responses by the BioASQ team of biomedical experts.

Figure 5 presents the performance of the top systems for each question type in exact answers during the eight years of the BioASQ challenge. The diagram reveals that this year the performance of systems in the yes/no questions keeps improving. For instance, in batch 3 presented in Table 8, various systems manage to outperform by far the strong baseline, which is based on a version of the OAQA system that achieved top performance in previous years. Improvements are also observed in the preliminary results for list questions, whereas the top system performance in factoid questions is fluctuating in the same range as done last year. In general, Figure 5 suggests that for the latter types of question there is still more room for improvement.

---

[11] http://participants-area.bioasq.org/results/8b/phaseB/

System	Yes/No		Factoid			List		
	Acc.	F1	Str. Acc.	Len. Acc.	MRR	Prec.	Rec.	F1
Umass_czi_5	**0.9032**	0.8995	0.2500	0.4286	0.3030	**0.7361**	0.4833	**0.5229**
Umass_czi_1	0.8065	0.8046	0.2500	0.3571	0.2869	0.6806	0.4444	0.4683
Umass_czi_2	0.8387	0.8324	0.2500	0.3571	0.2869	0.6806	0.4444	0.4683
pa-base	**0.9032**	0.8995	0.2500	0.4643	0.3137	0.5278	0.4778	0.4585
pa	**0.9032**	0.8995	0.2500	0.4643	0.3137	0.5278	0.4778	0.4585
Umass_czi_4	**0.9032**	0.9016	**0.3214**	0.4643	0.3810	0.6111	0.4361	0.4522
KU-DMIS-1	**0.9032**	**0.9028**	**0.3214**	0.4286	0.3601	0.6583	0.4444	0.4520
KU-DMIS-4	0.8387	0.8360	0.2857	0.4286	0.3357	0.6167	0.4444	0.4490
KU-DMIS-5	**0.9032**	**0.9028**	**0.3214**	0.4643	0.3565	0.6167	0.4444	0.4490
KU-DMIS-2	0.8710	0.8697	**0.3214**	0.4286	0.3446	0.6028	0.4444	0.4467
KU-DMIS-3	0.8387	0.8360	0.2500	0.4643	0.3357	0.6111	0.4444	0.4431
UoT_allquestions	0.5806	0.3673	**0.3214**	0.3929	0.3423	0.5972	0.4111	0.4290
UoT_baseline	0.5806	0.3673	**0.3214**	0.3929	0.3512	0.4861	0.4056	0.4214
Best factoid	0.5806	0.4732	0.2857	0.3929	0.3333	0.5208	0.4056	0.4107
bio-answerfinder	0.8710	0.8640	**0.3214**	0.4286	0.3494	0.3884	**0.5083**	0.4078
FudanLabZhu2	0.7419	0.6869	**0.3214**	0.5357	**0.3970**	0.5694	0.3583	0.3988
FudanLabZhu3	0.7419	0.6869	**0.3214**	0.4643	0.3655	0.5583	0.3472	0.3777
FudanLabZhu4	0.7419	0.6869	0.2857	**0.5714**	0.3821	0.5583	0.3472	0.3777
FudanLabZhu5	0.7419	0.6869	**0.3214**	0.4286	0.3690	0.5583	0.3472	0.3777
UoT_multitask_l.	0.5161	0.3404	**0.3214**	0.4286	0.3643	0.5139	0.3556	0.3721
BioASQ_Baseline	0.5161	0.5079	0.0714	0.2143	0.1220	0.2052	0.4833	0.2562

Table 8. Results for batch 3 for exact answers in phase B of Task 8b. Only the performance of the top-20 systems and the BioASQ Baseline are presented.

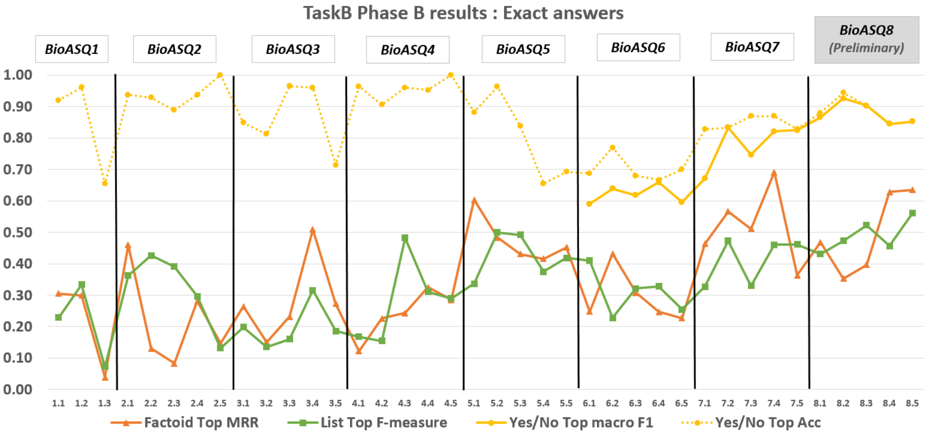

Fig. 5. The official evaluation scores of the best performing systems in Task B, Phase B, exact answer generation, across the eight years of the BioASQ challenge. Since BioASQ6 the official measure for Yes/No questions is the macro-averaged F1 score (macro F1, but accuracy (Acc) is also presented as the former official measure. The results for BioASQ8 are preliminary, as the final results for Task 8b will be available after the manual assessment of the system responses.

5 Conclusions

This paper provides an overview of the eighth version of the BioASQ tasks a and b, on biomedical semantic indexing and question answering in English respectively. These tasks, already established through the previous seven years of the challenge, together with the new MESINESP task on semantic indexing of medical content in Spanish, which ran for the first time, consisted the eighth edition of the BioASQ challenge.

The overall shift of participant systems towards deep neural approaches, already noticed in the previous years, is even more apparent this year. State-of-the-art methodologies have been successfully adapted to biomedical question answering and novel ideas have been investigated. In particular, most of the systems adopted neural embedding approaches, notably based on BERT and BioBERT models, for both tasks. In the QA task in particular, different teams attempted transferring knowledge from general domain QA datasets, notably SQuAD, or from other NLP tasks such as NER and NLI, also experimenting with multi-task learning settings. In addition, recent advancements in NLP, such as XLNet [51], BART [26] and SpanBERT [20] have also been tested for the tasks of the challenge.

Overall, as in previous versions of the tasks, the top preforming systems were able to advance over the state of the art, outperforming the strong baselines on the challenging shared tasks offered by the organizers. Therefore, we consider that the challenge keeps meeting its goal to push the research frontier in biomedical semantic indexing and question answering. The future plans for the challenge include the extension of the benchmark data though a community-driven acquisition process.

6 Acknowledgments

Google was a proud sponsor of the BioASQ Challenge in 2019. The eighth edition of BioASQ is also sponsored by the Atypon Systems inc. BioASQ is grateful to NLM for providing the baselines for task 8a and to the CMU team for providing the baselines for task 8b. The MESINESP task is sponsored by the Spanish Plan for advancement of Language Technologies (Plan TL) and the Secretaría de Estado para el Avance Digital (SEAD). BioASQ is also grateful to LILACS, SCIELO and Biblioteca virtual en salud and Instituto de salud Carlos III for providing data for the BioASQ MESINESP task.

References

1. Almeida, T., Matos, S.: Calling attention to passages for biomedical question answering. In: European Conference on Information Retrieval. pp. 69–77. Springer (2020)

2. Anastasios, N., Anastasia, K., Konstantinos, B., Martin, K., Carlos, R.P., Marta, V., Georgios, P.: Overview of bioasq 2020: The eighth bioasq challenge on large-scale biomedical semantic indexing and question answering. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), Thessaloniki, Greece, September 22–25, 2020, Proceedings. vol. 12260. Springer (2020)
3. Baldwin, B., Carpenter, B.: Lingpipe. Available from World Wide Web: http://alias-i. com/lingpipe (2003)
4. Balikas, G., Partalas, I., Kosmopoulos, A., Petridis, S., Malakasiotis, P., Pavlopoulos, I., Androutsopoulos, I., Baskiotis, N., Gaussier, E., Artieres, T., Gallinari, P.: Evaluation framework specifications. Project deliverable D4.1, UPMC (05/2013 2013)
5. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
6. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
7. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
8. Chang, W.C., Yu, H.F., Zhong, K., Yang, Y., Dhillon, I.: X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers. arXiv preprint arXiv:1905.02331 (2019)
9. Chen, Q., Peng, Y., Lu, Z.: Biosentvec: creating sentence embeddings for biomedical texts. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). pp. 1–5. IEEE (2019)
10. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
11. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
12. Couto, F.M., Lamurias, A.: MER: a shell script and annotation server for minimal named entity recognition and linking. Journal of Cheminformatics **10**(1), 58 (dec 2018). https://doi.org/10.1186/s13321-018-0312-9
13. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**, 1–30 (2006)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**(Mlm), 4171–4186 (oct 2018), http://arxiv.org/abs/1810.04805
15. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research **22**, 457–479 (2004)
16. Gormley, C., Tong, Z.: Elasticsearch: The definitive guide: A distributed real-time search and analytics engine. "O'Reilly Media, Inc." (2015)
17. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 (2020)
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 '16. pp. 935–944. ACM Press, New York, New York, USA (2016). https://doi.org/10.1145/2939672.2939756
19. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: Pubmedqa: a dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146 (2019)
20. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2020)
21. Kim, D., Lee, J., So, C.H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., Kang, J.: A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access **7**, 73729–73740 (2019)
22. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Mining and Knowledge Discovery **29**(3), 820–865 (2015)
23. Kotitsas, S., Pappas, D., Androutsopoulos, I., McDonald, R., Apidianaki, M.: Embedding biomedical ontologies by jointly encoding network structure and textual node descriptors. arXiv preprint arXiv:1906.05939 (2019)
24. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Stroudsburg, PA, USA (2018). https://doi.org/10.18653/v1/D18-2012
25. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: pre-trained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746 (2019)
26. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
27. Loper, E., Bird, S.: Nltk: the natural language toolkit. arXiv preprint cs/0205028 (2002)
28. Ma, J., Korotkov, I., Yang, Y., Hall, K., McDonald, R.: Zero-shot neural retrieval via domain-targeted synthetic query generation. arXiv preprint arXiv:2004.14503 (2020)
29. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp. 404–411 (2004)
30. Mollá, D., Jones, C.: Classification betters regression in query-based multi-document summarisation techniques for question answering. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 624–635. Springer (2019)
31. 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)
32. Nentidis, A., Bougiatiotis, K., Krithara, A., Paliouras, G.: Results of the seventh edition of the bioasq challenge. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 553–568. Springer (2019). https://doi.org/10.1007/978-3-030-43887-6_51
33. Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: Fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669 (2019)
34. Ozyurt, I.B., Bandrowski, A., Grethe, J.S.: Bio-answerfinder: a system to find answers to questions from biomedical texts. Database **2020** (2020)

35. Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J., Cheng, X.: Deeprank: A new deep architecture for relevance ranking in information retrieval. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 257–266 (2017)
36. 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)
37. Pappas, D., Stavropoulos, P., Androutsopoulos, I., McDonald, R.: Biomrc: A dataset for biomedical machine reading comprehension. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. pp. 140–149 (2020)
38. 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)
39. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. Proceedings of the Conference on Empirical Methods in Natural Language Processing pp. 31–40 (feb 2018), http://arxiv.org/abs/1802.05365
40. 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)
41. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
42. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
43. Ribadas, F.J., De Campos, L.M., Darriba, V.M., Romero, A.E.: CoLe and UTAI at BioASQ 2015: Experiments with similarity based descriptor assignment. CEUR Workshop Proceedings **1391** (2015)
44. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
45. Smith, L., Tanabe, L.K., nee Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., et al.: Overview of biocreative ii gene mention recognition. Genome biology **9**(S2), S2 (2008)
46. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: 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
47. 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)
48. Wei, C.H., Kao, H.Y., Lu, Z.: Pubtator: a web-based text mining tool for assisting biocuration. Nucleic acids research **41**(W1), W518–W522 (2013)
49. Wei, C.H., Leaman, R., Lu, Z.: Beyond accuracy: creating interoperable and scalable text-mining web services. Bioinformatics (Oxford, England) **32**(12), 1907–10 (2016). https://doi.org/10.1093/bioinformatics/btv760
50. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426 (2017)

51. 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
52. Yang, Z., Zhou, Y., Eric, N.: Learning to answer biomedical questions: Oaqa at bioasq 4b. ACL 2016 p. 23 (2016)
53. 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)
54. You, R., Liu, Y., Mamitsuka, H., Zhu, S.: Bertmesh: Deep contextual representation learning for large-scale high-performance mesh indexing with full text. bioRxiv (2020). https://doi.org/10.1101/2020.07.04.187674, https://www.biorxiv.org/content/early/2020/07/06/2020.07.04.187674
55. You, R., Zhang, Z., Wang, Z., Dai, S., Mamitsuka, H., Zhu, S.: Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. arXiv preprint arXiv:1811.01727 (2018)
56. Zavorin, I., Mork, J.G., Demner-Fushman, D.: Using learning-to-rank to enhance nlm medical text indexer results. ACL 2016 p. 8 (2016)