BioASQ at CLEF2020: Large-Scale Biomedical Semantic Indexing and Question Answering

Martin Krallinger[1], Anastasia Krithara[2], Anastasios Nentidis[2,3](✉),
Georgios Paliouras[2], and Marta Villegas[1]

[1] Barcelona Supercomputing Center, Barcelona, Spain
{martin.krallinger,marta.villegas}@bsc.es
[2] National Center for Scientific Research "Demokritos", Athens, Greece
{akrithara,tasosnent,paliourg}@iit.demokritos.gr
[3] Aristotle University of Thessaloniki, Thessaloniki, Greece
nentidis@csd.auth.gr

Abstract. This paper describes the eighth edition of the BioASQ Challenge, which will run as an evaluation Lab in the context of CLEF2020. The aim of BioASQ is the promotion of systems and methods for highly precise biomedical information access. This is done through the organization of a series of challenges (shared tasks) on large-scale biomedical semantic indexing and question answering, where different teams develop systems that compete on the same demanding benchmark datasets that represent the real information needs of biomedical experts. In order to facilitate this information finding process, the BioASQ challenge introduced two complementary tasks: (a) the automated indexing of large volumes of unlabelled data, primarily scientific articles, with biomedical concepts, (b) the processing of biomedical questions and the generation of comprehensible answers. Rewarding the most competitive systems that outperform the state of the art, BioASQ manages to push the research frontier towards ensuring that the biomedical experts will have direct access to valuable knowledge.

Keywords: Biomedical information · Semantic indexing · Question answering

1 Introduction

The availability of biomedical data increases rapidly with scientific literature being a major data resource for biomedical knowledge. MEDLINE/PubMed currently comprises more than 20 million articles with more than 2 new articles published in biomedical journals every minute[1]. This wealth of new knowledge is essential for scientific progress in biomedicine and can have a high impact on public health. However, ensuring that this knowledge is used in a timely manner by the biomedical experts is necessary for maximizing the benefit of the society.

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

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 550–556, 2020.
https://doi.org/10.1007/978-3-030-45442-5_71

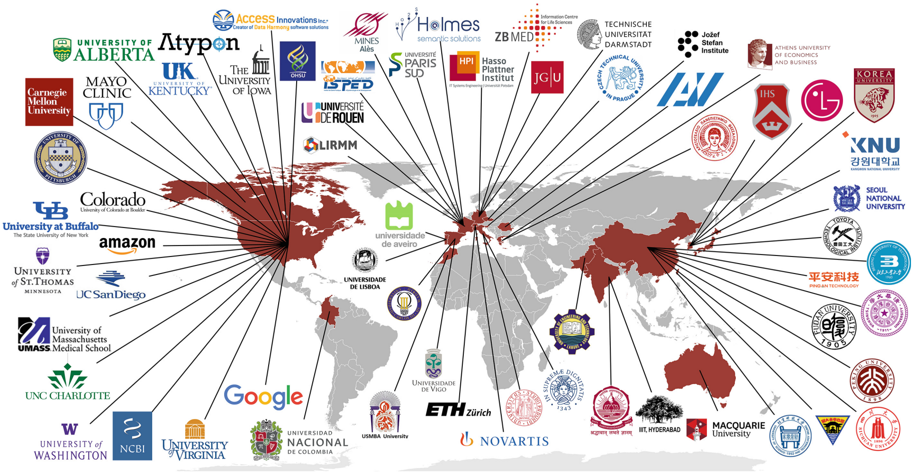
UNIVERSITY OF ALBERTA
Atypon
Access Innovations Inc.
OHSU
MINES Alès
Holmes semantic solutions
ZB MED Information Centre for Life Sciences
TECHNISCHE UNIVERSITÄT DARMSTADT
Jožef Stefan Institute
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
Carnegie Mellon University
MAYO CLINIC
UK UNIVERSITY OF KENTUCKY
THE UNIVERSITY OF IOWA
ISPED
UNIVERSITÉ PARIS SUD
HPI Hasso Plattner Institut
JG|U
CZECH TECHNICAL UNIVERSITY IN PRAGUE
IHS
KOREA UNIVERSITY
UNIVERSITÉ DE ROUEN
LIRMM
KNU 강원대학교
SEOUL NATIONAL UNIVERSITY
Colorado University of Colorado at Boulder
University at Buffalo The State University of New York
amazon
universidade de aveiro
UNIVERSIDADE DE LISBOA
TOYOTA TECHNOLOGICAL INSTITUTE
平安科技 PINGAN TECHNOLOGY
UNIVERSITY of ST.THOMAS MINNESOTA
UC San Diego
FUDAN UNIVERSITY 1905
University of Massachusetts UMASS Medical School
UNC CHARLOTTE
Google
UNIVERSIDADE DE VIGO
ETH Zürich
NOVARTIS
IIIT, HYDERABAD
MACQUARIE University
PEKING UNIVERSITY
W UNIVERSITY of WASHINGTON
NCBI
UNIVERSITY of VIRGINIA
UNIVERSIDAD NACIONAL DE COLOMBIA
USMBA University
SICHUAN UNIVERSITY

Fig. 1. Distribution of participating teams for the seven years of the BioASQ challenges

BioASQ[2] is a series of international challenges (shared tasks) and workshops focusing on biomedical semantic indexing and question answering. The BioASQ challenges [7] are structured into complementary tasks and sub-tasks so that participating teams can focus on tasks relevant to their area of expertise, including hierarchical text classification, machine learning, information retrieval and multi-document summarization amongst many other areas. BioASQ is also unique in requiring that the participating systems answer biomedical natural language questions by searching in both structured data (e.g. ontologies, databases) and unstructured data (e.g., biomedical articles).

As BioASQ consistently rewards highly precise biomedical information access systems developed by teams around the world, ensures that the biomedical experts eventually have more and more direct access to valuable knowledge that will help them avoid costly (sometimes even fatal) mistakes and provide high quality health services. The worldwide distribution of more than 60 teams from 20 counties participating in the challenges through the seven years of BioASQ is shown in Fig. 1. The BioASQ challenge has been running on an annual basis since 2012. The workshop has been taking place in the CLEF conference till 2015. In 2016 and 2017 it took place in ACL, in conjunction with BioNLP. In 2018 and 2019, it took place as an independent workshop in EMNLP and ECML-PKDD conferences respectively. In 2020, an independent BioASQ CLEF lab will run and the results will be presented in the context of CLEF 2020 conference.

[2] http://www.bioasq.org.

2 BioASQ Evaluation Lab 2020

The BioASQ challenge assesses the performance of information systems in supporting the following tasks that are central in the biomedical question answering process: (a) the indexing of large volumes of unlabeled data, primarily scientific articles, with biomedical concepts (in English and Spanish), (b) the processing of biomedical questions and the generation of answers and supporting material. Both these tasks have been running since the first year of BioASQ, but this is the first year for BioASQ to extend beyond the English language, by challenging the community to semantically index biomedical content in Spanish. Therefore, after the introduction of the new BioASQ task MESINESP in early 2020, the eighth BioASQ challenge will consist of the three tasks described in this section.

2.1 Task 8a: Large-Scale Biomedical Semantic Indexing

BioASQ task 8a requires systems to automatically assign MeSH terms to biomedical articles added to the MEDLINE database, thus assisting the indexing of biomedical literature. In effect, this is a classification task that requires documents to be automatically classified into a hierarchy of classes. A training dataset of about 14,9 million annotated articles is already available for task 8a and testsets with newly published articles will be released weekly, before the NLM curators annotate them. The systems will assign MeSH labels to them, which will be compared against the labels assigned by the curators.

Evaluation in Task 8a: As the manual annotations become gradually available, the scores of the systems are updated. In this manner, the evaluation of the systems participating in task 8a is fully automated on the side of BioASQ and thus can run on a weekly basis throughout the year. The performance of the systems taking part in task 8a is assessed with a range of different measures. Some of them are variants of standard information retrieval measures for multi-label classification problems (e.g. precision, recall, f-measure accuracy). Additionally, measures that use the MeSH hierarchy to provide a more refined estimate of the systems' performance are used. The official measures for identifying the winners of the task are micro-averaged F-measure (MiF) and the Lowest Common Ancestor F-measure (LCA-F) [2].

2.2 Task 8b: Biomedical Question Answering

BioASQ task 8b takes place in two phases. In the first phase, the participants are given English questions formulated by biomedical experts. For each question, the participating systems have to retrieve relevant MEDLINE documents, relevant snippets (passages) of the documents, relevant concepts (from five designated ontologies), and relevant RDF triples (from the Linked Life Data platform). This is also a classification task that requires questions to be classified into classes from multiple hierarchies. Subsequently, in the second phase of task

8b, the participants are given some relevant documents and snippets that the experts themselves have identified (using tools developed in BioASQ [6]), and they are required to return 'exact' answers (e.g., names of particular diseases or genes) and 'ideal' answers (a paragraph-sized summary of the most important information of the first phase per question). A training dataset of 3,243 biomedical questions is already available for participants of task 8b to train their systems and about 500 new biomedical questions, with corresponding golden annotations and answers, will be developed for the five testsets of task 8b.

Evaluation in Task 8b: The responses of the systems are evaluated both automatically and manually by the experts employing a variety of evaluation measures [3]. In phase A, on the retrieval of relevant material, both ordered and unordered measures are calculated but the official evaluation is based on the Mean Average Precision (MAP). For the exact answers in phase B, different evaluation measures are used depending on the type of the question. For yes/no questions the official evaluation measure is the macro-averaged F-Measure on questions with answers *yes* and *no*. For factoid questions, where the participants are allowed to return up to five answers, the Mean Reciprocal Rank (MRR) is used. For List questions, the official measure is the mean F-Measure. Finally, for ideal answers, the official evaluation is based on manual scores assigned by experts estimating the readability, recall, precision and repetition of each response provided by the participating systems.

2.3 Task MESINESP8: Medical Semantic Indexing in Spanish

Currently, most of the Biomedical NLP and IR research is being done on English documents, and only a few tasks have been carried out on non-English texts. Nonetheless, it is important to note that there is also a considerable amount of medically relevant content published in other languages than English and particularly clinical texts that are entirely written in the native language of each country, with a few exceptions. For example, there is a large subset of medical content published in Spanish each year. Resources like PubMed do only contain a fraction of the biomedical and medical literature originally published in Spanish, which is also stored in other resources such as IBECS[3], SCIELO[4] or LILACS[5].

In this task, the participants are asked to classify new IBECS and LILACS documents in Spanish. The classes come from the DeCS vocabulary[6] which was developed from the MeSH hierarchy. A dataset of more than 300,000 articles in Spanish with DeCS annotations retrieved from the Virtual Health Library (VHL) Portal[7] will be available for the participants to train their systems.

[3] http://ibecs.isciii.es/.

[4] https://scielo.org/en/.

[5] http://lilacs.bvsalud.org/en/.

[6] http://decs.bvs.br/I/decsweb2019.htm.

[7] https://bvsalud.org/en/.

Evaluation in Task MESINESP8: A set of about 1,500 articles in Spanish will be annotated by two human expert annotators with golden DeCS labels. Some of them will be provided to the participants as development dataset and some of them will be used as a testset to evaluate the performance of the systems. The responses of the systems in this task will be evaluated with the same variety of flat evaluation measures used for task 8a [2], with the micro-averaged F-measure (MiF) as the official one.

2.4 BioASQ Datasets and Tools

A major contribution of BioASQ is the development and maintenance of benchmark datasets for biomedical semantic indexing and question answering. The dataset for the semantic indexing task includes more than 14 millions articles from PubMed. This year, given the new MESINESP task, a new dataset of more than 300 thousands Spanish semantically indexed articles has been created. Furthermore, a set of 3,243 realistic questions and answers have been generated, constituting a unique resource for the development of question answering systems.

In addition, BioASQ has created a lively ecosystem, supported by tools and systems that facilitate research, such as the BioASQ Annotation Tool [6] for dataset development on question answering and a range of evaluation measures for automated assessment of system performance in all tasks. All software and data that are produced are open to the public[8]. It is worth mentioning, that this year we plan to create a repository, through which, several participating systems will also be available. This will allow new participants and teams, to build on existing models.

3 The Impact of BioASQ Results

BioASQ has reportedly had a very large impact, both in research and in industry; it has vastly helped advance the field of text mining in bioinformatics and has enabled researchers and practitioners to create novel computational models for life and health sciences. By bringing people together who work on the same benchmark data, BioASQ significantly facilitates the exchange and fusion of ideas and eventually accelerates progress in the field.

For example, the Medical Text Indexer (MTI) [5], which is developed by the NLM to assist in the indexing of biomedical literature, has improved its performance by almost 10% in the last 7 years (Fig. 2). NLM has announced that improvement in MTI is largely due to the adoption of ideas from the systems that compete in the BioASQ challenge [4]. Recently, MTI has reached a performance level that allows it to be used in the fully automated indexing of articles of specific types [1]. In general, a variety of biomedical semantic indexing and question answering systems has been developed based on the BioASQ datasets which are continuously maintained and extended. In addition, BioASQ keeps evolving considering the inclusion of new tasks and the development of new datasets.

[8] https://github.com/bioasq.

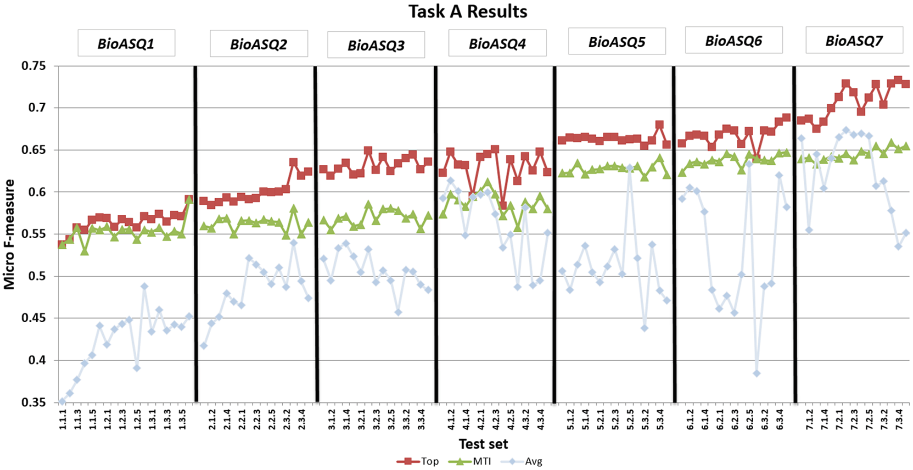

Fig. 2. Performance of the participating systems in task 8a, on semantic indexing. Each year, the participating systems push the state-of-the-art to higher levels

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. Incorporating values for indexing method in medline/pubmed xml. https://www.nlm.nih.gov/pubs/techbull/ja18/ja18_indexing_method.html. Accessed 1 Oct 2019
2. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Mining Knowl. Discov. **29**(3), 820–865 (2014). https://doi.org/10.1007/s10618-014-0382-x
3. Malakasiotis, P., Pavlopoulos, I., Androutsopoulos, I., Nentidis, A.: Evaluation measures for task b. Technical report, Technical report BioASQ (2018). http://participants-area.bioasq.org/Tasks/b/eval_meas_2018
4. Mork, J., Aronson, A., Demner-Fushman, D.: 12 years on–is the nlm medical text indexer still useful and relevant? J. Biomed. Semant. **8**(1), 8 (2017)
5. Mork, J., Jimeno-Yepes, A., Aronson, A.: The NLM medical text indexer system for indexing biomedical literature (2013)

6. Ngomo, A.C.N., Heino, N., Speck, R., Ermilov, T., Tsatsaronis, G.: Annotation tool. Project deliverable D3.3 (2013). http://www.bioasq.org/sites/default/files/PublicDocuments/2013-D3.3-AnnotationTool.pdf
7. 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