Accepted Manuscript published in Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science, vol 13982. Springer, Cham. https://doi.org/10.1007/978-3-031-28241-6_66

BioASQ at CLEF2023: The eleventh edition of the Large-scale biomedical semantic indexing and question answering challenge

Anastasios Nentidis[1,2], Anastasia Krithara[1], Georgios Paliouras[1], Eulalia Farre-Maduell[3], Salvador Lima-Lopez[3], and Martin Krallinger[3]

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

Abstract. The large-scale biomedical semantic indexing and question-answering challenge (BioASQ) aims at the continuous advancement of methods and tools to meet the need of biomedical researchers and practitioners for efficient and precise access to the ever-increasing resources of their domain. With this purpose, during the last ten years a series of annual challenges have been organized with specific shared tasks on large-scale biomedical semantic indexing and question answering. Benchmark datasets have been concomitantly provided in alignment with the real needs of biomedical experts. BioASQ provides a unique common testbed where different teams around the world can investigate and compare new approaches for identifying and accessing biomedical knowledge. The eleventh version of the BioASQ Challenge will be held as an evaluation Lab within CLEF2023. In this version, three shared tasks will be presented: (i) the automated retrieval of relevant material for biomedical questions, and the generation of comprehensible answers. (ii) the synergistic retrieval of relevant material and generation of answers for open biomedical questions about developing topics, in collaboration with the experts posing the questions. (iii) the automated indexing of unlabelled clinical procedures-specific medical documents, primarily clinical case reports written in Spanish, with biomedical concepts and the extraction of human-interpretable evidence. As BioASQ rewards the methods that outperform the state of the art in these shared tasks, it pushes the research frontier towards approaches that accelerate access to biomedical knowledge.

Keywords: Biomedical Information · Semantic Indexing · Question Answering

1 Introduction

BioASQ[4] [20] is a series of international challenges and workshops on biomedical semantic indexing and question answering. Each edition of BioASQ is structured into distinct but complementary tasks and sub-tasks relevant to biomedical information access. As a result, the participating teams can focus on particular tasks of interest to their specific area of expertise, including but not limited to hierarchical text classification, machine learning, information retrieval, and multi-document query-focused summarization. The BioASQ challenge has been running annually since 2012, with the participation of more than 100 teams from 28 countries. The BioASQ workshop has been taking place at the CLEF conference till 2015. In 2016 and 2017 it took place in ACL, in conjunction with the BioNLP workshop [1]. In 2018 and 2019, it took place respectively in EMNLP and ECML as an independent workshop. Since 2020 the BioASQ workshop become again part of CLEF [15, 7, 5].

BioASQ allows multiple teams that work on biomedical information access systems around the world, to compete in the same realistic benchmark datasets and share, evaluate, and compare their ideas and approaches. Therefore, a key contribution of BioASQ are the benchmark datasets developed for its tasks, as well as the corresponding open-source infrastructure developed for running the challenges. In particular, as BioASQ consistently rewards the most successful approaches in each task and sub-task, it eventually pushes toward systems that outperform previous approaches. Such successful approaches for semantic indexing and question answering can eventually lead to the development of tools to support more precise access to valuable biomedical knowledge and to further improve health services.

Notably, the performance of MTI, the system developed by the National Library of Medicine (NLM) for assisting the manual semantic indexing of MEDLINE, has improved by almost 10% during the last 10 years, largely due to the adoption of ideas from the systems that compete in the large-scale biomedical semantic indexing task (*Task a*) of the BioASQ challenge [13, 22]. The high point for this task was the recent adoption of fully automated indexing by NLM in mid-2022 [5]. In short, ten years after its initial introduction, this task fulfilled its goal of facilitating the advancement of biomedical semantic indexing research. However, major advancement is still needed regarding biomedical question-answering, as well as the semantic indexing of other types of documents, such as clinical case reports and documents in languages beyond English.

2 BioASQ evaluation lab 2023

The eleventh BioASQ challenge (BioASQ11) will consist of three tasks that are central to biomedical knowledge access and the question-answering process: (i)

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

[5] https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html

Task b[6] on the processing of biomedical questions, the generation of answers, and the retrieval of supporting material, (ii) *Task Synergy* on biomedical question answering for developing problems under a scenario that promotes collaboration between biomedical experts and question-answering systems, and (iii) *Task MedProcNER* on text mining and semantic indexing of clinical procedures in medical documents in Spanish, including the annotation of concepts in unlabeled documents and the subsequent normalization of these concept annotations. *Task MedProcNER* can be considered as a follow-up task of the previous DisTEMIST task on disease mentions [10]. As all three tasks have also been organized in the context of previous editions of the BioASQ challenge, we respectively refer to their current version, in the context of BioASQ11, as *task 11b*, *task Synergy 11* and *task MedProcNER*.

2.1 Task 11b: Biomedical question answering

BioASQ *task 11b* takes place in two phases. In the first phase (Phase A), the participants are given questions in English formulated by biomedical experts. For each question, the participating systems have to retrieve relevant documents (from PubMed) and relevant snippets (passages) of the documents. Subsequently, in the second phase (Phase B) of *task 11b*, the participants are given some relevant documents and snippets that the experts themselves have identified (using tools developed in BioASQ [16]). In this phase, they are required to return 'exact' answers, such as names of particular diseases or genes, depending on the type of the question, and 'ideal' answers, which are paragraph-sized summaries of the most important information of the first phase for each question, regardless of its type. A training dataset of 4,721 biomedical questions will be available for participants of *task 11b* to train their systems and about 500 new biomedical questions, with corresponding golden annotations and answers, will be developed for testing the participating systems.

The evaluation of system responses is done both automatically and manually by the experts employing a variety of evaluation measures [8]. In phase A, the official evaluation for document retrieval is based on the Mean Average Precision (MAP) and for snippet retrieval with the F-measure. In phase B, for the exact answers, the official evaluation measure depends on the type of question. For yes/no questions the official measure is the macro-averaged F-Measure on questions with answers *yes* and *no*. The Mean Reciprocal Rank (MRR) is used for factoid questions, where the participants are allowed to return up to five candidate answers. For List questions, the official measure is the mean F-Measure. Finally, for ideal answers, even though automatic evaluation measures are provided and semi-automatic measures [19] are also considered, the official evaluation is still based on manual scores assigned by experts estimating the readability, recall, precision, and repetition of each response.

[6] Since the introduction of BioASQ, the task on large-scale biomedical semantic indexing is called *Task a*, and the task on biomedical question answering is called *Task b*, for brevity. Despite the completion of *Task a* last year, we keep this naming convention for *Task b*, for the sake of uniformity with previous versions.

2.2 Task Synergy 11: Question answering for developing topics

The original BioASQ *task b* is structured in a sequence of phases where the experts and the participating systems have minimal interaction. This is acceptable for research questions that have a clear, undisputed answer. However, for questions on developing topics, such as the COVID-19 pandemic, that may remain open for some time and where new information and evidence appear every day, a more interactive model is needed, aiming at a synergy between the automated question-answering systems and the biomedical experts.

In this direction, since 2020 we introduced the BioASQ *task Synergy* which is designed as a continuous dialog, that allows biomedical experts to pose unanswered questions for developing problems and receive the system responses to these questions, including relevant material (documents and snippets) and potential answers [5]. Next, the experts assess these responses, and provide feedback to the systems, in order to improve their responses. This process repeats iteratively with new feedback and new system responses for the same questions, as well as with new questions that may have arisen. In each round of this task, new material is also considered based on the current version of the resources. Initially, the task was focused on COVID-19 considering documents from the COVID-19 Open Research Dataset (CORD-19)[21]. This year, the topic of the questions in the *task Synergy 11* will be open to any developing problem considering documents from the current version of PubMed that will be designated for each round. As in previous versions of the task, the questions are not required to have definite answers and the answers to the questions can be more volatile.

The same evaluation measures used in *task 11b* are also employed in *task Synergy 11* for comparison. However, in order to capture the iterative nature of the task, only new material is considered for the evaluation of a question in each round, an approach known as *residual collection evaluation*[18]. In parallel, additional evaluation metrics are also examined in this direction. Through this task, we aim to facilitate the incremental understanding of new developing public health topics, such as COVID-19, and contribute to the discovery of new solutions.

2.3 Task MedProcNER: Medical Procedure Text Mining and Indexing Shared Task

Despite the importance of medical procedures for a diversity of topics such as health data mining, analytics and research, limited efforts have been made so far to automatically extract, index, or identify medical procedure mentions from clinical documents. Clinical procedures are a critical concept type for clinical coding systems and can be considered one of the most significant medical entity types to characterize medical tests and therapeutic or surgical aspects associated to patient care. Moreover, medical procedures are of uttermost importance for determining, measuring, or diagnosing a patient's condition and characterize clinical aspects of relevance for medical and surgical treatments of patients. Medical procedures also have a direct practical relevance regarding the use and safety of medical implants and devices. They are undoubtedly transversal medical entity types, of relevance for all medical specialties, including cardiology, oncology, psychiatry and surgery-related clinical specialties such as gynecology and urology.

Correct detection and normalization of medical procedure terms is critical for clinical coding and medical information retrieval systems. The novel *task MedProcNER* will focus on the recognition and indexing of medical procedures in clinical documents in Spanish, by posing subtasks on (1) indexing medical documents with controlled terminologies (2) automatic detection (indexing) of textual evidence, i.e. mentions of medical procedure entities, in text and (3) normalization of these medical procedure mentions to terminologies.

The BioASQ *task MedProcNER* will rely primarily on 1,000 clinical case report publications in Spanish (SciELO [17] full text articles) for indexing diseases with concept identifiers from SNOMED-CT [2], MeSH and ICD10 [7]. A large silver standard collection of additional case reports and medical abstracts will also be provided [9]. A silver standard can be described as a set of annotations provided automatically by state-of-the-art algorithms as opposed to manually annotated by experts. The evaluation of systems for this task will use flat evaluation measures following the *task a* [4] track (mainly micro-averaged F-measure, MiF).

2.4 BioASQ datasets and tools

During the ten years of BioASQ, hundreds of systems from research teams around the world have been evaluated on the indexing, retrieval, and analysis of hundreds of thousands of biomedical publications and on answering thousands of biomedical questions. In this direction, BioASQ has developed a lively ecosystem of tools that facilitate research, such as the BioASQ Annotation Tool [16] for question-answering dataset development and a range of evaluation measures for automated assessment of system performance in all tasks. All BioASQ software [8] and datasets [9] are publicly available.

In particular, for *task b* on biomedical question answering, BioASQ employs a team of trained biomedical experts who provide a set of about 500 questions on their specialized field of expertise annually for evaluating the performance of participating systems. For *task 11b*, a set of 4,721 realistic questions accompanied by answers, and supporting evidence (documents and snippets) is already available as a unique resource for the development of question-answering systems [6]. In addition, from previous versions of *task Synergy*, which took place in twelve rounds over the last two years, a dataset of 258 questions on COVID-19 is already available. These questions are incrementally annotated with different versions of exact and ideal answers, as well as documents and snippets assessed by the experts as relevant or irrelevant. During the *task Synergy 11* this set

[7] https://www.cdc.gov/nchs/icd/icd10cm.htm

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

[9] http://participants-area.bioasq.org/datasets

will be extended with more than fifty new open questions on COVID-19 and other developing health topics. Meanwhile, any existing questions that remain relevant may be enriched with more updated answers and more recent evidence (documents and snippets) [14].

In addition, for biomedical semantic indexing, a training dataset of more than 16.2 million articles and fifteen weekly test sets of around 6,000 articles each are available from the tenth edition of *task a* (*task 10a*) [14]. Even though the *task a* completed its life cycle as a BioASQ task in 2022, the corresponding resources are still useful for related tasks such as semantic indexing in other languages, other types of biomedical documents, or specific sets of labels. Similarly, the datasets from the previous versions of the *task MESINEP*, on medical semantic indexing in Spanish, are also available [3].

For the *task MedProcNER*, a new dataset of semantically annotated medical documents in Spanish labeled with text-bound evidence mentions of medical procedures together with concept identifiers for entity linking and semantic indexing will be released. This year, the dataset will additionally focus on high-impact clinical specialties, such as cardiology. In addition, a dataset of 1,000 clinical cases in Spanish is already available from the previous edition of *task DisTEMIST*, together with corresponding concept recognition and linking annotations, as well as a set of disease-relevant mentions from over 200,000 biomedical articles in Spanish [10]. This will allow the exploration of disease-medical procedure relations.

3 Conclusions

BioASQ facilitates the exchange and fusion of ideas, providing unique realistic datasets and evaluation services for research teams that work on biomedical semantic indexing and question answering. Therefore, it eventually accelerates progress in the field, as indicated by the gradual improvement of the scores achieved by the participating systems [13]. An illustrative example is the Medical Text Indexer (MTI) [12], which achieved significant improvements [13] largely due to the adoption of ideas from the systems that compete in the BioASQ challenge [11], eventually reaching a performance level that allows the adoption of fully automated indexing in NLM [10].

Similarly, we expect that the new version of BioASQ will allow the participating teams to bring further improvement to the open tasks of biomedical question answering (*task 11b*), answering open questions for developing topics (*task Synergy 11*), and clinical procedure text mining and semantic indexing of medical documents in Spanish (*task MedProcNER*). In conclusion, BioASQ aims to assist participating teams in their approach to the challenge's tasks, which represent key information needs in the biomedical domain.

[10] https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html

4 Acknowledgments

Google was a proud sponsor of the BioASQ Challenge in 2022. The eleventh edition of BioASQ is also sponsored by Atypon Systems inc. The *task MedProcNER* is supported by the Spanish Plan for the Advancement of Language Technologies (Plan TL), the 2020 Proyectos de I+D+i-RTI Tipo A (Descifrando El Papel De Las Profesiones En La Salud De Los Pacientes A Traves De La Mineria De Textos, PID2020-119266RA-I00). This project has received funding from the European Union Horizon Europe Coordination and Support Action under Grant Agreement No 101058779 (BIOMATDB) and DataTools4Heart - DT4H, Grant agreement No 101057849.

References

1. Cohen, K.B., Demner-Fushman, D., Ananiadou, S., Tsujii, J. (eds.): BioNLP 2017. Association for Computational Linguistics, Vancouver, Canada, (Aug 2017). https://doi.org/10.18653/v1/W17-23, https://aclanthology.org/W17-2300
2. Donnelly, K., et al.: SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics **121**, 279 (2006)
3. Gasco, L., Nentidis, A., Krithara, A., Estrada-Zavala, D., Murasaki, R.T., Primo-Peña, E., Bojo Canales, C., Paliouras, G., Krallinger, M., et al.: Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. CEUR Workshop Proceedings (2021)
4. 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)
5. Krallinger, M., Krithara, A., Nentidis, A., Paliouras, G., Villegas, M.: BioASQ at CLEF2020: Large-scale biomedical semantic indexing and question answering. In: European Conference on Information Retrieval. pp. 550–556. Springer (2020)
6. Krithara, A., Nentidis, A., Bougiatiotis, K., Paliouras, G.: BioASQ-QA: A manually curated corpus for Biomedical Question Answering. bioRxiv (2022)
7. Krithara, A., Nentidis, A., Paliouras, G., Krallinger, M., Miranda, A.: BioASQ at CLEF2021: Large-Scale Biomedical Semantic Indexing and Question Answering. In: European Conference on Information Retrieval. pp. 624–630. Springer (2021)
8. Malakasiotis, P., Pavlopoulos, I., Androutsopoulos, I., Nentidis, A.: Evaluation measures for task b. Tech. rep., Tech. rep. BioASQ (2018), http://participants-area.bioasq.org/Tasks/b/eval_meas_2018
9. Ménard, P.A., Mougeot, A.: Turning silver into gold: error-focused corpus reannotation with active learning. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 758–767 (2019)
10. Miranda-Escalada, A., Gascó, L., Lima-López, S., Farré-Maduell, E., Estrada, D., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2022)

11. Mork, J., Aronson, A., Demner-Fushman, D.: 12 years on–Is the NLM medical text indexer still useful and relevant? Journal of biomedical semantics **8**(1), 8 (2017)
12. Mork, J., Jimeno-Yepes, A., Aronson, A.: The NLM Medical Text Indexer System for Indexing Biomedical Literature (2013)
13. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Miranda-Escalada, A., Gasco, L., Krallinger, M., Paliouras, G.: Overview of BioASQ 2022: The Tenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13390 LNCS, pp. 337–361 (oct 2022). https://doi.org/10.1007/978-3-031-13643-6_22
14. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 10a, 10b and Synergy10 in CLEF2022. In: CEUR Workshop Proceedings. vol. 3180, pp. 171–178 (2022)
15. Nentidis, A., Krithara, A., Paliouras, G., Gasco, L., Krallinger, M.: BioASQ at CLEF2022: The Tenth Edition of the Large-scale Biomedical Semantic Indexing and Question Answering Challenge. In: European Conference on Information Retrieval. pp. 429–435. Springer (2022)
16. Ngomo, A.C.N., Heino, N., Speck, R., Ermilov, T., Tsatsaronis, G.: Annotation tool. Project deliverable D3.3 (02/2013 2013), http://www.bioasq.org/sites/default/files/PublicDocuments/2013-D3.3-AnnotationTool.pdf
17. Packer, A.L., Biojone, M.R., Antonio, I., Takenaka, R.M., García, A.P., Silva, A.C.d., Murasaki, R.T., Mylek, C., Reis, O.C., Delbucio, H.C.R.F.: SciELO: uma metodologia para publicação eletrônica. Ciência da informação **27**, nd–nd (1998)
18. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science **41**(4), 288–297 (jun 1990). https://doi.org/10.1002/(SICI)1097-4571(199006)41:4¡288::AID-ASI8¿3.0.CO;2-H
19. ShafieiBavani, E., Ebrahimi, M., Wong, R., Chen, F.: Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 905–914. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1077
20. 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
21. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al.: CORD-19: The COVID-19 open research dataset. ArXiv (2020), https://arxiv.org/abs/2004.10706v2
22. Zavorin, I., Mork, J.G., Demner-Fushman, D.: Using Learning-To-Rank to Enhance NLM Medical Text Indexer Results. ACL 2016 p. 8 (2016)