

BioASQ at CLEF2024: The twelfth edition of the Large-scale biomedical semantic indexing and question answering challenge

Anastasios Nentidis^{1,2}, Anastasia Krithara¹, Georgios Paliouras¹, Martin Krallinger³, Luis Gasco Sanchez³, Salvador Lima³, Eulalia Farre³, Natalia Loukachevitch⁴, Vera Davydova⁵, and Elena Tutubalina^{5,6,7}

¹ National Center for Scientific Research “Demokritos”, Athens, Greece

{`tasosnent, akrithara, paliourg`}@iit.demokritos.gr

² Aristotle University of Thessaloniki, Thessaloniki, Greece

³ Barcelona Supercomputing Center, Barcelona, Spain

{`martin.krallinger, eulalia.farre, salvador.limalopez, lgasco`}@bsc.es

⁴ Moscow State University, Russia

⁵ Sber AI, Russia

⁶ Artificial Intelligence Research Institute, Russia

⁷ Kazan Federal University, Russia

tutubalinaev@gmail.com

Abstract. The large-scale biomedical semantic indexing and question-answering challenge (BioASQ) aims at the continuous advancement of methods and tools to meet the needs of biomedical researchers and practitioners for efficient and precise access to the ever-increasing resources of their domain. With this purpose, during the last eleven 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, providing a unique common testbed where different teams around the world can investigate and compare new approaches for accessing biomedical knowledge.

The twelfth version of the BioASQ Challenge will be held as an evaluation Lab within CLEF2024 providing four shared tasks: (i) *Task b* on the information retrieval for biomedical questions, and the generation of comprehensible answers. (ii) *Task Synergy* the information retrieval and generation of answers for open biomedical questions on developing topics, in collaboration with the experts posing the questions. (iii) *Task MultiCardioNER* on the automated annotation of clinical entities in medical documents in the field of cardiology, primarily in Spanish, English, Italian and Dutch. (iv) *Task BioNNE* on the automated annotation of biomedical documents in Russian and English with nested named entity annotations. 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

BioASQ⁸ [27] is a series of international challenges and workshops on biomedical semantic indexing and question answering (QA). 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 machine learning, information retrieval, information extraction, and multi-document query-focused summarization. The BioASQ challenge has been running since 2012, with the participation of more than 100 teams from 30 countries, and a BioASQ workshop is organized annually [22, 23, 7, 4]. This year, the BioASQ workshop is part of the fifteenth CLEF conference⁹.

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 is the benchmark datasets and the open-source infrastructure developed for running its tasks [6]. 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 QA can eventually lead to the development of tools to support more precise access to biomedical knowledge and to further improve health services.

2 BioASQ evaluation lab 2024

The twelfth BioASQ challenge (BioASQ12) will consist of four tasks that are central to biomedical knowledge access and the question-answering process: (i) *Task b*¹⁰ on the processing of biomedical questions, the generation of answers, and the retrieval of supporting material, (ii) *Task Synergy* on biomedical QA for developing problems under a scenario that promotes collaboration between biomedical experts and question-answering systems, (iii) *Task MultiCardioNER* on text mining and semantic indexing of diseases and medications in cardiology clinical case report documents in Spanish, including the annotation of concepts in unlabeled documents and the subsequent normalization of these concept annotations. *Task MultiCardioNER* can be considered as a follow-up task of the previous *MedcProcNER* [8] and *DisTEMIST* [13] tasks on mentions of diseases and medical procedures. (iv) *Task BioNNE* on the automated annotation of unlabelled biomedical documents, written in Russian and English, with nested named entities. As *Task b* and *Task Synergy* have also been organized in the context of previous editions of the BioASQ challenge [16, 19], we refer to their

⁸ <http://www.bioasq.org>

⁹ <https://clef2024.imag.fr/>

¹⁰ Since the first BioASQ, the task on large-scale biomedical semantic indexing is called *Task a*, and the task on biomedical QA is called *Task b*, for brevity. After the completion of *Task a* [5], we keep this naming convention for *Task b* for consistency.

current version, in the context of BioASQ12, as *task 12b*, *task Synergy 12* respectively.

2.1 Task 12b: Biomedical question answering

BioASQ *task 12b* takes place in two phases. In the first phase (Phase A), the participants are given questions in English formulated by biomedical experts and their systems have to retrieve relevant documents (from PubMed) and snippets (passages) of the documents. In the second phase (Phase B), some relevant documents and snippets identified by the experts (using the BioASQ tools [24]) are also provided, and ‘exact’ and ‘ideal’ answers are required by the systems. Depending on the type of question, the ‘exact’ answer can be a *yes* or *no* (yes/no), an entity name, such as a disease or gene (factoid), or a list of entity names (list). The ‘ideal’ answer is a paragraph-sized summary of the most important information for each question, regardless of its type.

About 500 new biomedical questions annotated with golden documents, snippets, and answers (‘exact’ and ‘ideal’), will be developed by the BioASQ team of trained biomedical experts for testing and assessing the performance of the participating systems. In addition, a training set of about 5,050 biomedical questions, accompanied by answers, and supporting evidence (documents and snippets), will be available from previous versions of the tasks, as a unique resource for the development of question-answering systems [6]. The evaluation in *task 12b* is done manually by the experts and automatically by employing a variety of evaluation measures [12] as done in the previous version of the task [16]. The official measure for document retrieval is the Mean Average Precision and for snippet retrieval the F-measure. For the exact answers, the official measure depends on the type of question. For yes/no questions we use the macro-averaged F-Measure. For factoid questions, where up to five candidate answers can be submitted, we use the Mean Reciprocal Rank. For List questions, we use the mean F-Measure. Finally, for ideal answers, the official evaluation is based on manual scores assigned by experts assessing the readability, recall, precision, and repetition of each response.

2.2 Task Synergy 12: Question answering for developing topics

In an effort to use the BioASQ question-answering ecosystem in order to promote research in developing biomedical topics, such as COVID-19, we introduced the BioASQ *task Synergy* in 2020 [7]. Contrary to the original *task b*, *task Synergy* is designed as a continuous dialog, that allows biomedical experts to pose unanswered questions for developing problems, for which they do not know beforehand whether a definite answer can be provided. The experts receive relevant material (documents and snippets) and potential answers to these questions and assess them, providing 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 original document resource [18]. Since 2023, this evolving document resource is PubMed [16].

A training dataset of about 300 questions on COVID-19 and other developing topics is already available from previous versions of *task Synergy*, which took place in sixteen rounds over the last three years [17, 21, 20]. 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 not. During the *task Synergy 12* this set will be extended with more than fifty new open questions on developing health topics. Meanwhile, any existing questions that remain relevant may be enriched with more updated answers and more recent evidence. In *task Synergy 12* we use the same evaluation measures with *task 12b*, considering only new material for the information retrieval part, an approach known as *residual collection evaluation* [26]. However, the focus of this task is to aid the experts in contributing to the incremental understanding of new developing health topics and the discovery of new solutions.

2.3 Task MultiCardioNER: Multiple clinical entity detection in multilingual medical content

The extraction of clinical variables from unstructured content such as clinical case reports or EHRs is key to enable medical data analytics. Due to the highly specialized medical language, with considerable variation depending on the medical specialty or document type, more specialized automatic semantic annotation resources are needed, not only for English but also other languages. This is particularly true for clinical content related to the cardiovascular diseases (CVDs), which represent the leading cause of death globally, responsible for approximately 17.9 million death/year. Previous efforts to recognize clinical concepts, for instance in Spanish, have focused typically only on a single or limited number of entity types, using a general collection medical document, or focusing on clinical content written in a single language. This resulted in valuable datasets and resources, such as the DISTEMIST [13], SYMPTEMIST, PharmaCoNER [2], and Medprocner [9] corpora and systems, but the interplay and complementarity of multi-label entity extraction approaches were not targeted and evaluated. To address all these issues the novel *task MultiCardioNER* will focus on the automatic recognition, entity linking and indexing of key clinical variables or concept types, namely diseases and medications.

The *task MultiCardioNER* will focus on the recognition and indexing of these clinical entity types in cardiology clinical case documents in Spanish and other languages, by posing three subtasks on (1) automatic detection of mentions of diseases and medications in cardiology clinical case texts written in Spanish, (2) normalization of clinical entity mentions in Spanish to controlled vocabularies (clinical entity linking and concept normalization) and (3) automatic detection of mentions of diseases and medications in cardiology clinical case texts in English, Italian and Dutch.

The BioASQ *task MultiCardioNER* will rely on a collection of 1000 clinical case reports in Spanish annotated with disease and medication mentions, as well

as a collection on 600 cardiology clinical case reports annotated also by clinical experts with diseases and medications. These mentions were manually mapped to their corresponding concept identifiers from SNOMED-CT. Moreover, a multilingual collection of 600 cardiology case reports translated into English, Italian and Dutch where concept mentions had been manually corrected by clinical experts will be used for the third sub-track. As a training set 1000 clinical case reports together with an additional collection of 400 cardiology case reports will be used, while the remaining 200 cardiology case reports will serve as a test set collection. The evaluation of systems for this task will use flat evaluation measures following the *task a* [3] track (mainly micro-averaged F-measure, MiF).

2.4 Task BioNNE: Nested NER in Russian and English

Most biomedical datasets and named entity recognition (NER) methods have been designed to capture flat (non-nesting) mention structures. To facilitate ongoing research on nested NER, we introduce a BioNNE shared task on nested NER in PubMed abstracts in Russian and English. The evaluation framework will be divided into three broad tracks (bilingual or language-oriented). The train/dev sets are based on the NEREL-BIO dataset [11] which extends an annotation scheme of the general-domain NEREL [10] and includes annotated mentions of disorders, anatomical structures, chemicals, diagnostic procedures, and biological functions. The NEREL-BIO dataset includes over 700 annotated PubMed abstracts in the Russian language and a small set of annotated English abstracts (app. 100). We strongly encourage participants to apply cross-language (Russian to English) and cross-domain (existing bio NER corpora to the BioNNE set) techniques in this task. Participants will be able to run their systems on novel test sets on the Codalab platform. The span-level macro-averaged F1 will be used as an evaluation metric.

In particular, the *Task BioNNE* will be structured into three tracks: **Track 1 - Bilingual:** participants in this track are required to train a single multi-lingual NER model using training data for both Russian and English languages. The model should be used to generate prediction files for each language’s dataset. Please note that predictions from any mono-lingual model are not allowed in this track. **Tracks 2 & 3 - Language-oriented:** Participants in these tracks are required to train a model that works for one language. Participants are allowed to train any model architecture on any publicly available data in order to achieve best performance for the language of their choice.

2.5 BioASQ datasets and tools

During the eleven 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 [24] for question-answering dataset development and a range of evaluation measures for

automated assessment of system performance in all tasks. All BioASQ software ¹¹ and datasets ¹² are publicly available.

In particular, beyond the unique datasets described above for the four tasks running this year, BioASQ also provides: i) a training dataset of more than 16.2 million articles on large-scale biomedical semantic indexing with MeSH topics (*task a*) [5], ii) the *task MESINEP* datasets [1, 25] of more than 300K articles, on medical semantic indexing in Spanish, and iii) the *task DisTEMIST* [13] and *task MedProcNER* [8] datasets of 1,000 clinical cases in Spanish, on semantic indexing with SNOMED-CT entities for diseases and medical procedures.

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 [16, 19]. An illustrative example is the Medical Text Indexer (MTI) [15], which achieved significant improvements [19] largely due to the adoption of ideas from the systems that compete in the BioASQ challenge [14], eventually reaching a performance level that allows the adoption of fully automated indexing in NLM [5].

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 12b*), answering open questions for developing topics (*task Synergy 12*), multi-lingual named entity recognition in cardiology (*task MultiCardioNER*), and nested named entity recognition in Russian and English (*task BioNNE*). 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.

4 Acknowledgments

Google was a proud sponsor of the BioASQ Challenge in 2023. The twelfth edition of BioASQ is also sponsored by Ovid Technologies, Inc., Elsevier, and Atypion Systems inc. The *task MultiCardioNER* 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 Profesionales En La Salud De Los Pacientes A Traves De La Minería 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. The work on the BioNNE task was supported by the Russian Science Foundation [grant number 23-11-00358].

¹¹ <https://github.com/bioasq>

¹² <http://participants-area.bioasq.org/datasets>

References

1. 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)
2. Gonzalez-Agirre, A., Marimon, M., Intxaurreondo, A., Rabal, O., Villegas, M., Krallinger, M.: Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. pp. 1–10 (2019)
3. 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)
4. 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)
5. Krithara, A., Mork, J.G., Nentidis, A., Paliouras, G.: The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. *Frontiers in Research Metrics and Analytics* **8** (2023). <https://doi.org/10.3389/frma.2023.1250930>
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. Lima-López, S., Farré-Maduell, E., Gascó, L., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of MedProcNER Task on Medical Procedure Detection and Entity Linking at BioASQ 2023. In: CEUR Workshop Proceedings (2023)
9. Lima-López, S., Farré-Maduell, E., Gascó, L., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023. Working Notes of CLEF (2023)
10. Loukachevitch, N., Artemova, E., Batura, T., Braslavski, P., Denisov, I., Ivanov, V., Manandhar, S., Pugachev, A., Tutubalina, E.: NEREL: A Russian dataset with nested named entities, relations and events. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 876–885. INCOMA Ltd., Held Online (Sep 2021), <https://aclanthology.org/2021.ranlp-1.100>
11. Loukachevitch, N., Manandhar, S., Baral, E., Rozhkov, I., Braslavski, P., Ivanov, V., Batura, T., Tutubalina, E.: Nerel-bio: a dataset of biomedical abstracts annotated with nested named entities. *Bioinformatics* **39**(4), btad161 (2023)
12. 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
13. 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)
14. 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)
 15. Mork, J., Jimeno-Yepes, A., Aronson, A.: The NLM Medical Text Indexer System for Indexing Biomedical Literature (2013)
 16. Nentidis, A., Katsimpras, G., Krithara, A., Lima López, S., Farré-Maduell, E., Gasco, L., Krallinger, M., Paliouras, G.: Overview of BioASQ 2023: The Eleventh BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. pp. 227–250. Springer Nature Switzerland, Cham (2023)
 17. Nentidis, A., Katsimpras, G., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 11b and Synergy11 in CLEF2023. In: *CEUR Workshop Proceedings* (2023)
 18. Nentidis, A., Katsimpras, G., Vadorou, E., Krithara, A., Gasco, L., Krallinger, M., Paliouras, G.: Overview of BioASQ 2021: The Ninth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. pp. 239–263. Springer (2021)
 19. Nentidis, A., Katsimpras, G., Vadorou, 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
 20. Nentidis, A., Katsimpras, G., Vadorou, E., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 9a, 9b and Synergy in CLEF2021. In: *Proceedings of the 9th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*. CEUR Workshop Proceedings (2021), <http://ceur-ws.org/Vol-2936/paper-10.pdf>
 21. Nentidis, A., Katsimpras, G., Vadorou, 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)
 22. Nentidis, A., Krithara, A., Paliouras, G., Farre-Maduell, E., Lima-Lopez, S., Krallinger, M.: BioASQ at CLEF2023: The Eleventh Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) *Advances in Information Retrieval*. pp. 577–584. Springer Nature Switzerland, Cham (2023)
 23. 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)
 24. 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>
 25. Rodriguez-Penagos, C., Nentidis, A., Gonzalez-Agirre, A., Asensio, A., Armengol-Estapé, J., Krithara, A., Villegas, M., Paliouras, G., Krallinger, M.: Overview of

- MESINESP8, a Spanish Medical Semantic Indexing Task within BioASQ 2020 (2020)
26. 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](https://doi.org/10.1002/(SICI)1097-4571(199006)41:4<288::AID-ASI8;3.0.CO;2-H)
 27. 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>