

Overview of BioASQ Tasks 10a, 10b and Synergy10 in CLEF2022

Anastasios Nentidis^{1,2}, Georgios Katsimpras¹, Eirini Vandorou¹, Anastasia Krithara¹ and Georgios Paliouras¹

¹NCSR Demokritos, Athens, Greece

²Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

BioASQ is a series of challenges focused on promoting methodologies and systems for large-scale biomedical semantic indexing and question answering. The BioASQ challenge is part of the Conference and Labs of the Evaluation Forum (CLEF) and includes a variety of tasks. This paper provides an overview of the tasks a, b, and Synergy of the tenth edition of BioASQ challenge. In the 2022 edition, 29 teams with more than 120 systems participated in these three tasks of the challenge, with 8 of them focusing on task 10a, 20 on task 10b, and 6 on task Synergy. Although the overall participation was decreased compared to previous versions, the high percentage of newly registered teams suggests that the interest of the community in large-scale biomedical semantic indexing and question answering is vigorous.

Keywords

Biomedical knowledge, Semantic Indexing, Question Answering

1. Introduction

This paper describes the shared tasks 10a, 10b and Synergy10 of the tenth edition of the BioASQ challenge in 2022. Additionally, details on the datasets that were used in each task are given. Section 2, gives an overview of tasks 10a and 10b, that took place from January to May 2022, task Synergy10, which took place from December 2021 to February 2022, as well as the corresponding datasets developed for training and testing the participating systems. Section 3, briefly outlines the participation in these three tasks. A detailed analysis of the methodologies followed by the participating systems will be available in the proceedings of the BioASQ lab. A brief discussion along with our conclusions are provided in the last section.

2. Overview of the Tasks


Overall, the 2022 BioASQ challenge consisted of four tasks: (1) a large-scale biomedical semantic indexing task (task 10a), (2) a biomedical question answering task (task 10b), (3) a task on biomedical question answering for the developing issue of COVID-19 (task Synergy10), all three

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ tasosnent@iit.demokritos.gr (A. Nentidis); gkatsibras@iit.demokritos.gr (G. Katsimpras); evandorou@iit.demokritos.gr (E. Vandorou); akrithara@iit.demokritos.gr (A. Krithara); paliourg@iit.demokritos.gr (G. Paliouras)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

considering documents in English, and (4) a new task on medical semantic indexing and disease text mining (task DisTEMIS), considering medical documents in Spanish. In this paper, we give a brief description of the first three established tasks 10a, 10b and Synergy10 with focus on differences from previous versions of the challenge [1]. In addition, a detailed description of 10a and 10b tasks can be found in [2], which also describes the general structure of BioASQ.

2.1. Large-scale semantic indexing - Task 10a

Table 1

Statistics on test datasets for Task 10a. Due to the early adoption of a new NLM policy for fully automated indexing, the third batch finally consists of a single testset.

Batch	Articles	Annotated Articles	Labels per Article
1	9659	9450	13.03
	4531	4512	12.00
	4291	4269	13.04
	4256	4192	12.81
	4862	4802	12.75
Total	27599	27225	12.72
2	8874	8818	12.70
	4071	3858	12.38
	4108	4049	12.60
	3193	3045	11.74
	3078	2916	12.07
Total	23324	22686	12.29
3	2376	1870	12.31
	28	0	-
Total	2404	1870	12.31

In Task 10a, participants are asked to classify articles from the PubMed/MedLine¹ digital library into concepts of the MeSH hierarchy. Specifically, new PubMed articles that are not yet annotated by the indexers at NLM are collected to build the test sets for the evaluation of the competing systems. However, NLM scaled-up its policy of fully automated indexing to all MEDLINE citations by mid-2022². In response to this change, the schedule of task 10a was shifted a few weeks earlier in the year and the task was completed in fewer rounds compared to previous years. The details of each test set are shown in Table 1. In consequence, we believe that, ten years after its initial introduction, the task has fulfilled its goal in facilitating the advancement of biomedical semantic indexing research and no new editions of this task are planned in the context of the BioASQ challenge.

The task is designed into three independent batches of 5 weekly test sets each. However, due to the new NLM policy the third batch finally consists of a single test set. A second testset has also been initially released in the context of the third batch, but due to its extremely small size and the fully automated annotation of all its articles by NLM, it was disregarded and no

¹<https://pubmed.ncbi.nlm.nih.gov/>

²https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html

results will be released for it. Overall, two scenarios are provided in this task: i) on-line and ii) large-scale. The test sets contain new articles from all available journals. Similar to previous versions of the task [3], standard flat and hierarchical information retrieval measures were used to evaluate the competing systems, as soon as the annotations from the NLM indexers were available. Moreover, for each test set, participants had to submit their answers in 21 hours. Additionally, a training dataset that consists of 16,218,838 articles with 12.68 labels per article, on average, and covering 29,681 distinct MeSH labels in total, was provided for Task 10a.

2.2. Biomedical semantic QA - Task 10b

Task 10b consists of a large-scale question answering challenge in which participants have to develop systems for all the stages of question answering in the biomedical domain. As in previous editions, the task examines four types of questions: “yes/no”, “factoid”, “list” and “summary” questions [3]. In this edition, the available training dataset, which the competing teams had to use to develop their systems, contains 4,234 questions that are annotated with relevant golden elements and answers from previous versions of the task. Table 2 shows the details of both training and testing sets for task 10b.

Table 2

Statistics on the training and test datasets of Task 10b. The numbers for the documents and snippets refer to averages per question.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	4,234	1148	816	1252	1018	9.22	12.24
Test 1	90	23	14	34	19	3.22	4.06
Test 2	90	18	15	34	23	3.13	3.79
Test 3	90	25	11	32	22	2.76	3.33
Test 4	90	24	12	31	23	2.77	3.51
Test 5	90	28	18	29	15	3.01	3.60
Test 6	37	6	15	6	10	3.35	4.78
Total	4,721	1272	901	1418	1130	3.92	5.04

Differently from previous challenges, task 10b was split into six independent bi-weekly batches. These include five official batches, as in previous versions of the task, and an additional sixth batch with questions posed by new biomedical experts. The motivation for this additional batch was to investigate whether biomedical experts that are not familiar with the BioASQ would find the responses of the systems interesting and useful. In particular, a collaborative schema was adopted for this additional batch, where the new experts posed their questions in the field of biomedicine and the experienced BioASQ expert team reviewed these questions to guarantee their quality. The test set of the sixth batch contains 37 questions developed by eight new experts.

Task 10b is also divided into two phases: (phase A) the retrieval of the required information and (phase B) answering the question, which run during two consecutive days for each batch. In each phase, the participants receive the corresponding test set and have 24 hours to submit the answers of their systems. This year, a test set of 90 questions, written in English, was released for phase A and the participants were expected to identify and submit relevant elements from

designated resources, including PubMed/MedLine articles and snippets extracted from these articles. Then, the manually selected relevant articles and snippets for these 90 questions were also released in phase B and the participating systems were 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.

2.3. Synergy10 Task

The Synergy task was first introduced in the previous edition of the BioASQ challenge[1] aiming at a synergy between the biomedical experts studying the developing issue of COVID-19 and the automated question answering systems participating in BioASQ. The experts assess the systems' responses and their assessment is fed back to the systems in order to help improving them, in a continuous iterative process.

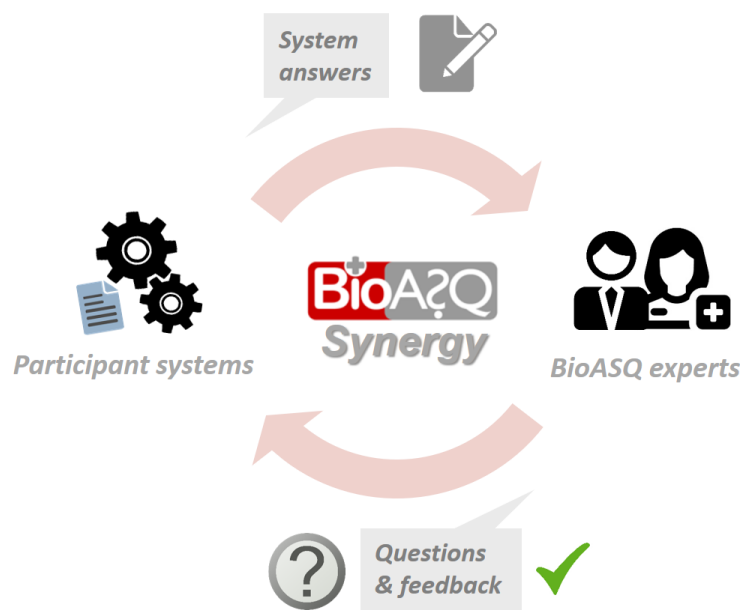


Figure 1: The iterative dialogue between the experts and the systems in the BioASQ Synergy task on question answering for COVID-19.

Fig. 1 sketches this procedure. The competing systems provide their initial answers for open questions on COVID-19 along with relevant documents and snippets, which are then assessed by the experts and fed back to the systems together with new or pending questions. This version of the Synergy task (Synergy10) is structured into four rounds, one every three weeks. In each round the system responses and expert feedback refer to the same questions, unless they have been closed by the experts for having received a full and definite answer that is not expected to change. This holds for questions from previous versions of the Synergy task, that remained open for updated material and answers in the light of new published knowledge. In addition

some new questions or new modified versions of some questions could be added into the test sets. Table 3 shows the details of the datasets used in task Synergy.

Table 3

Statistics on the datasets of Task Synergy. “Answer” stands for questions marked as having enough relevant material from previous rounds to be answered. “Feedback” stands for questions that already have some expert feedback from previous rounds.

Round	Size	Yes/No	List	Factoid	Summary	Answer	Feedback
1	72	21	20	13	18	13	26
2	70	20	19	13	18	25	70
3	70	20	19	13	18	41	70
4	64	18	19	10	17	47	64

In order to reflect the rapid developments in the field, each round of this task utilizes material from the current version of the COVID-19 Open Research Dataset (CORD-19) [4]. This year the time interval between two successive rounds was extended into three weeks, from two weeks in BioASQ9, to keep up with the release of new CORD-19 versions that were less frequent compared to the previous version of the task. In addition, apart from PubMed documents of the current CORD-19, CORD-19 documents from PubMed Central and ArXiv were also considered as additional resources of knowledge. Similar to task b, four types of questions are examined in Synergy: yes/no, factoid, list, and summary, and two types of answers, exact and ideal. Moreover, the assessment of the systems’ performance is based on the evaluation measures used in task 10b.

3. Overview of participation

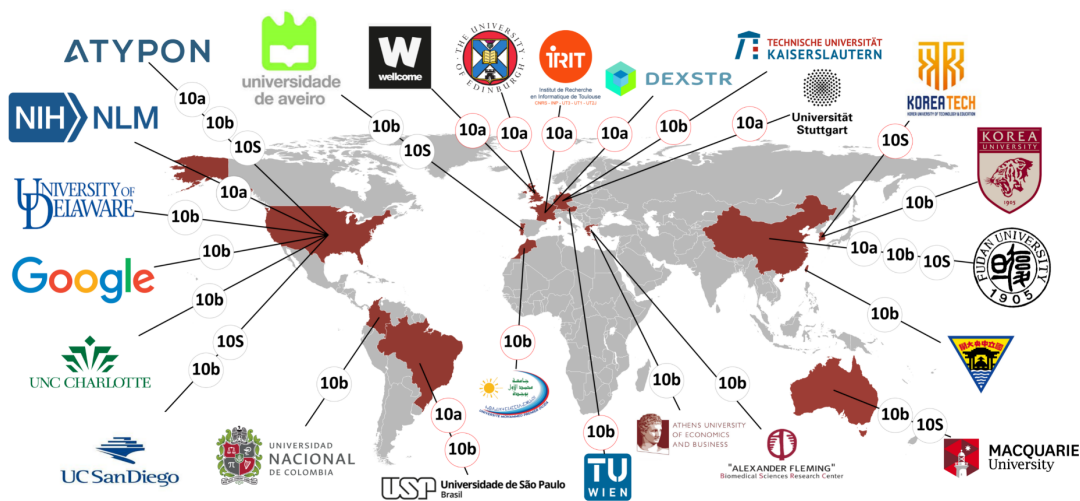


Figure 2: The world-wide distribution of teams participating in the tasks 10a, 10b and Synergy10 (10S), based on institution affiliations. A red circle indicates a newly registered team.

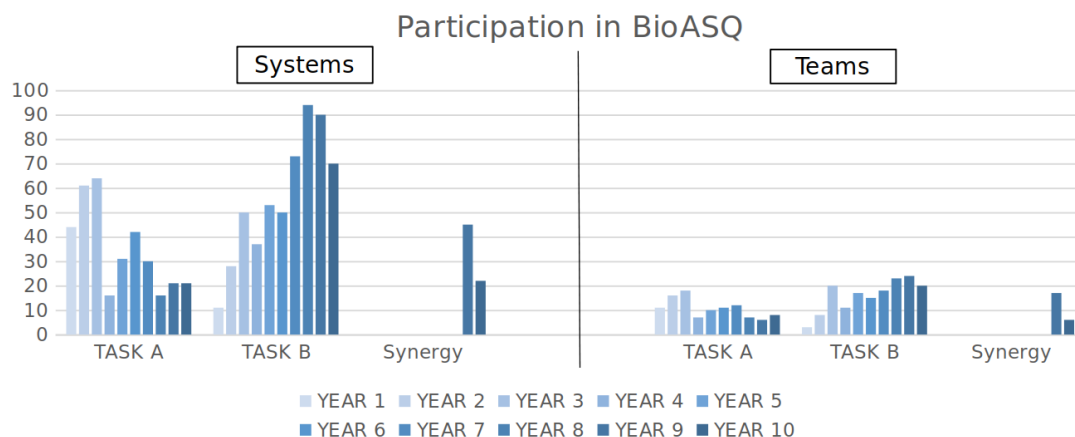


Figure 3: The evolution of participant teams in the BioASQ task a, b and Synergy in the ten years of BioASQ.

This year, 29 teams participated in the tasks 10a, 10b and Synergy10 of the challenge with more than 120 distinct systems, in total. Particularly, 8 of these teams submitted on task 10a, 20 on task 10b and 6 on task Synergy10. Furthermore, Fig. 2 illustrates the international interest in the challenge as the participating teams originate from various countries around the world.

As already observed in previous years of the challenge, the participation in task b is surpassing the participation in the other tasks. As shown in Fig. 3, the overall number of participating teams has been decreased this year, particularly for task Synergy. The fact that this year the task was running for only four rounds, instead of eight in BioASQ9, could be a reason for this decrease. However, the high percentage of teams that participated for the first time in the BioASQ challenge (red circles in Fig. 2), suggests that the interest of the community in large-scale biomedical semantic indexing and question answering is vigorous. In total, ten new teams participated in this year's editions of the tasks a, b and Synergy of the BioASQ challenge.

3.1. Task 10a

In task 10a, 8 teams competed this year with a total of 21 different systems. Teams that have already participated in previous versions of the task include the National Library of Medicine (NLM) team that submitted predictions with 5 different systems, and the Fudan University team that participated with 5 systems as well. On the other hand, 6 new teams competed for the first time, submitting results with 11 distinct systems, highlighting the interest of the research community in the task.

3.2. Task 10b

In task 10b, 20 teams competed this year with a total of 70 different systems for both phases A and B. In particular, 10 teams with 35 systems participated in phase A, while in phase B, the number of participants and systems were 16 and 49 respectively. Six teams engaged in both phases.

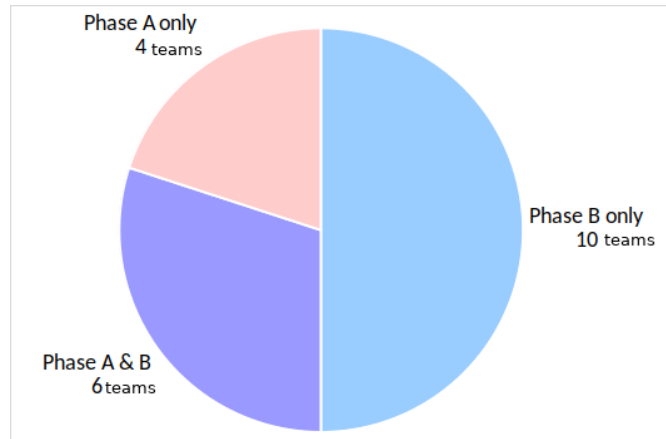


Figure 4: The distribution of participant teams in the BioASQ task 10b into phases.

3.3. Synergy Task

In task Synergy, 6 teams participated this year with a total of 22 distinct systems. As this task shares some common ideas with task b, some teams participated in both tasks. Specifically, 4 teams participated in both task 10b and Synergy as shown in Fig. 5.

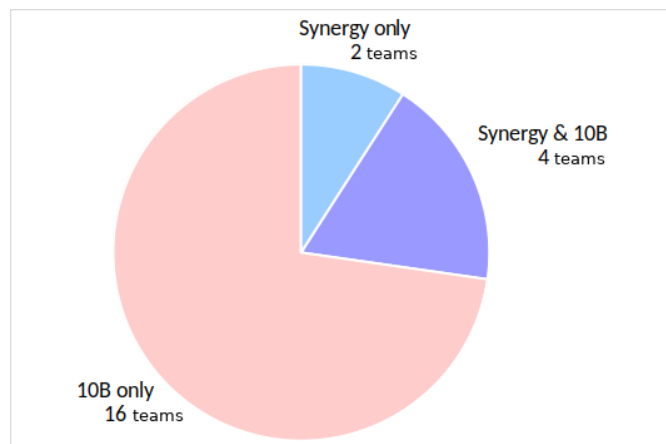


Figure 5: The overlap of participant teams in the BioASQ task 10b and Synergy.

4. Conclusions

In this paper, we presented the tenth version of the BioASQ tasks a, b and Synergy. Tasks 10a and 10b, are both already established through the previous nine years of the challenge, and Synergy task ran for the second year. Differently from previous years, the participation of teams was slightly decreased, on the other hand, we noticed a high number of newly registered

teams. Therefore, we consider that the challenge and the datasets developed for its tasks enhance the interest of the research community in large-scale biomedical semantic indexing and question answering and push towards the development of better solutions to aid the biomedical researchers' access to the abundance of biomedical knowledge.

Acknowledgments

Google was a proud sponsor of the BioASQ Challenge in 2021. The tenth edition of BioASQ is also sponsored by the Atypion Systems inc. BioASQ is grateful to NLM for providing the baselines for task 10a and to the CMU team for providing the baselines for task 10b. The Distemist 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 Distemist task.

References

- [1] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. Paliouras, 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, Springer, 2021, pp. 239–263.
- [2] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weisenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, *BMC Bioinformatics* 16 (2015) 138. doi:10.1186/s12859-015-0564-6.
- [3] G. Balikas, I. Partalas, A. Kosmopoulos, S. Petridis, P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, N. Baskiotis, E. Gaussier, T. Artieres, P. Gallinari, Evaluation Framework Specifications, Project deliverable D4.1, UPMC, 2013.
- [4] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al., CORD-19: The COVID-19 open research dataset, *ArXiv* (2020).