Overview of BioASQ Tasks 11b and Synergy11 in CLEF2023

Anastasios Nentidis[1,2], Georgios Katsimpras[1], Anastasia Krithara[1] and Georgios Paliouras[1]

[1]NCSR Demokritos, Athens, Greece
[2]Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

In this paper, we present an overview of the eleventh edition of the BioASQ challenge, which is part of the fourteenth Conference and Labs of the Evaluation Forum (CLEF). BioASQ is a series of challenges focused on promoting methodologies and systems for large-scale biomedical semantic indexing and question answering. This document provides an overview of the tasks b and Synergy in this year's BioASQ edition. Although fewer teams participated in this edition compared to previous ones, more than 80 systems were submitted by 22 teams for these two tasks. Task 11b was the focus of 19 teams while 5 teams participated in task Synergy. Like the previous year, the high percentage of newly registered teams suggests that the interest of the community in large-scale biomedical semantic indexing and question answering remains strong.

Keywords

Biomedical knowledge, Semantic Indexing, Question Answering

1. Introduction

In this paper, we describe the shared tasks b and Synergy of the eleventh edition of the BioASQ challenge in 2023, named 11b and Synergy11 respectively. Furthermore, we provide details on the datasets that were used in each task. Section 2, provides an overview of tasks 11b and Synergy11, that were held from March to May 2023, and from January to February 2023, respectively. Additionally, the corresponding datasets that were developed for training and testing the participating systems are described. Section 3, provides a brief overview of the participation in these two tasks. A detailed analysis of the methodologies followed by the participating systems will be available in the proceedings of the BioASQ lab and an overview of them in [1]. In the last section, we provide a brief discussion along with our conclusions.

2. Overview of the Tasks

In the eleventh edition of the BioASQ challenge, there were offered three tasks: (1) a biomedical question answering task (task b), (2) a task on biomedical question answering for open devel-

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
tasosnent@iit.demokritos.gr (A. Nentidis); gkatsibras@iit.demokritos.gr (G. Katsimpras); akrithara@iit.demokritos.gr (A. Krithara); paliourg@iit.demokritos.gr (G. Paliouras)

oping issues (task Synergy), both tasks considering documents in English, and (3) a new task focused on the detection, normalization, and indexing of clinical procedures (task MedProcNER), considering medical documents in Spanish [2]. In this paper, we give a description of the current version of the first two established tasks, task b and task Synergy, with a focus on differences from previous versions of the challenge [3, 4]. In particular, we use the names task 11b and task Synergy11, when referring to the current version of task b and task Synergy, in the context of the eleventh edition of BioASQ. A respective description of the task MedProcNER is provided in [5]. Additionally, a detailed introduction to the BioASQ challenge and the structure of its tasks in their initial versions can be found in [6].

2.1. Biomedical semantic QA - Task 11b

Task 11b consists of a large-scale question-answering challenge that involves developing 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 [7]. In this edition, the training dataset available to the competing teams contains 4,719 questions that are annotated with relevant golden elements and answers from previous versions of the task [8]. The teams had to use this dataset to develop their systems. The details of both training and testing sets for task 11b are shown in Table 1.

Table 1
Statistics on the training and test datasets of Task 11b. The numbers for the documents and snippets refer to averages per question.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	4,719	1,271	901	1,417	1,130	9.03	12.04
Test 1	75	24	12	19	20	2.48	3.28
Test 2	75	24	12	22	17	2.96	4.33
Test 3	90	24	18	26	22	2.66	3.77
Test 4	90	14	24	31	21	2.80	3.91
Total	5,049	1,357	967	1,515	1,210	8.62	11.5

Unlike previous challenges, task 11b was divided into four independent bi-weekly batches and the two phases for each batch run for two consecutive days. The two phases of task 11b consist of: (phase A) the retrieval of the required information and (phase B) answering the question, which run for 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, the first two test sets consisted of 75 questions each, and the remaining two test sets consisted of 90 questions each. For each test set, the respective questions, written in English, were 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 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.

2.2. Synergy11 Task

The Synergy task was first introduced in the ninth edition of the BioASQ challenge [9] with the goal of creating 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 improve them, in a continuous iterative process. Figure 1 sketches this procedure. The competing systems provide their initial answers for open questions on developing problems along with relevant documents and snippets. These answers are then evaluated by experts who then provide feedback to the systems and new or pending questions.

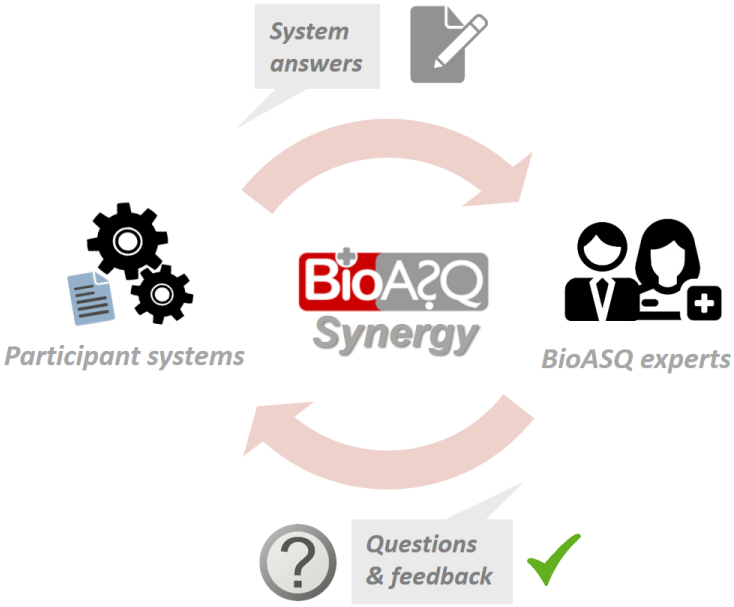

Figure 1: The iterative dialogue between the experts and the systems in the BioASQ Synergy task on question answering for open developing problems.

This version of the Synergy task (Synergy11) was structured into four rounds, one every two weeks, and was open to any developing problem considering documents from the current version of PubMed that was designated for each round. As in previous versions of the task, the questions were not required to have definite answers and the answers to the questions could be more volatile. A set of 311 questions on COVID-19 was also available from the previous versions of the Synergy task, together with respective expert feedback and answers, and was provided as a development set.

In each round of the Synergy task, 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. In Synergy11, in particular, a team of seven biomedical experts contributed 53 open biomedical questions in total and assessed the retrieved material (i.e. documents and snippets) and responses submitted by the participating systems in

each of the four rounds. Table 2 shows the details of the datasets used in task Synergy.

Table 2
Statistics on the datasets of Task Synergy. "Answer ready" stands for questions marked as having enough relevant material to be answered after the assessment of material submitted by the systems in the respective round. In round 2, ten questions were omitted from the test, as no feedback was available for them from the respective expert for the material retrieved by the systems in round 1. This feedback became available in round three, hence these questions were again included in the test set.

Round	Size	Yes/No	List	Factoid	Summary	Answer ready
1	53	12	17	11	13	14
2	43	11	14	7	11	32
3	53	12	17	11	13	37
4	53	12	17	11	13	42

Similar to task 11b, four types of questions are examined in Synergy11: 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 11b. After the completion of the Synergy11 task, enough relevant material was identified for providing an answer to about 79% of the questions. In addition, about 42% of the questions had at least one *ideal answer*, that had been submitted by the systems, which was considered satisfactory (ground truth) by the expert that posed the question.

3. Overview of participation

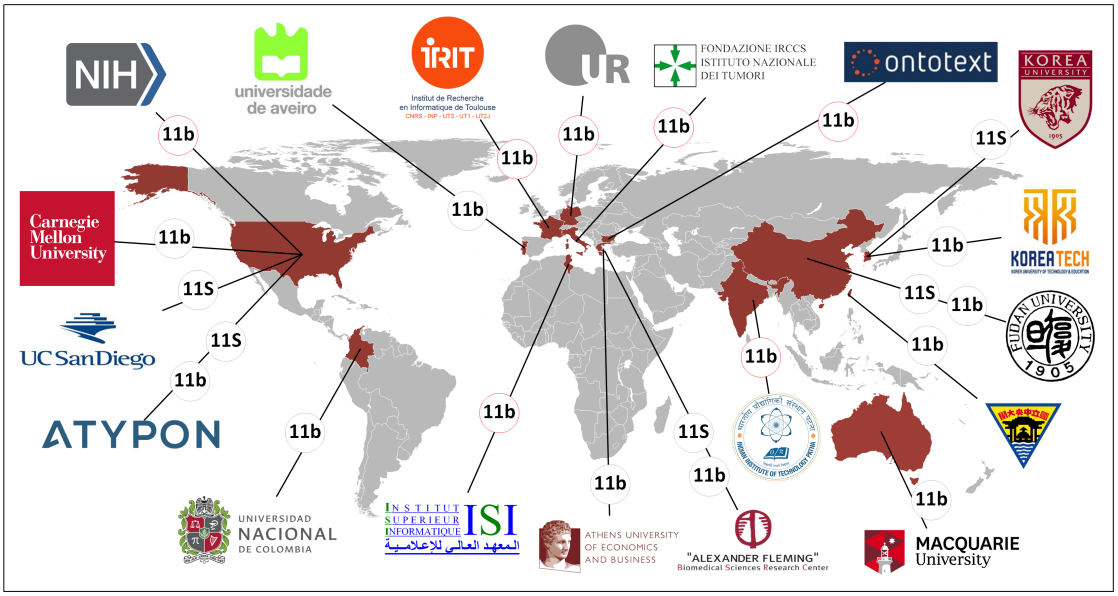

Figure 2: The world-wide distribution of teams participating in the tasks 11b and Synergy11 (11S), based on institution affiliations. A red circle indicates a newly registered team.

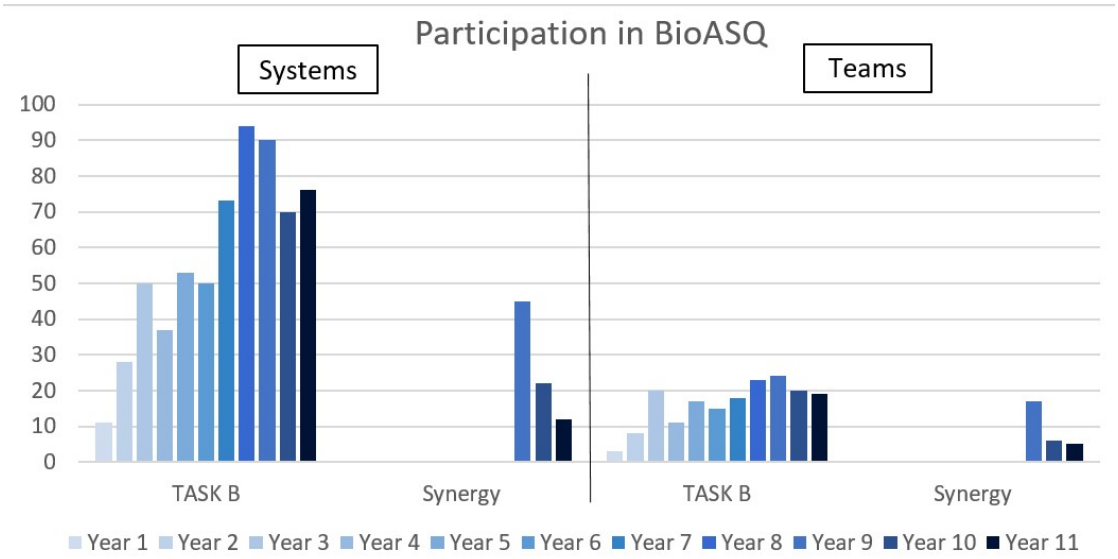

Figure 3: The evolution of participation in the BioASQ task b and Synergy in the eleven years of BioASQ.

This year's challenge saw more than 80 distinct systems participating in tasks 11b and Synergy11 with a total of 22 teams. Specifically, 19 of these teams submitted on task 11b and 5 on task Synergy11. Furthermore, Figure 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 of Synergy. The overall number of participating teams this year is similar to last year's with a slight decrease, as illustrated in Figure 3. However, the high percentage of teams that participated for the first time in the BioASQ challenge (red circles in Figure 2), suggests that the interest of the community in large-scale biomedical semantic indexing and question answering remains strong. A total of 7 new teams participated in this year's editions of tasks b and Synergy of the BioASQ challenge.

3.1. Task 11b

In task 11b, 19 teams competed this year with a total of 76 different systems for both phases A and B. In particular, 9 teams with 37 systems participated in phase A, while in phase B, the number of participants and systems were 16 and 59 respectively. 6 teams engaged in both phases.

3.2. Synergy Task

In task Synergy11, 5 teams participated this year with a total of 12 distinct systems. As this task shares some common ideas with task 11b, some teams participated in both tasks. Specifically, 2 teams participated in both task 11b and Synergy11 as shown in Figure 5. However, as already observed in previous versions of the tasks, less teams participate in Synergy11 than in task

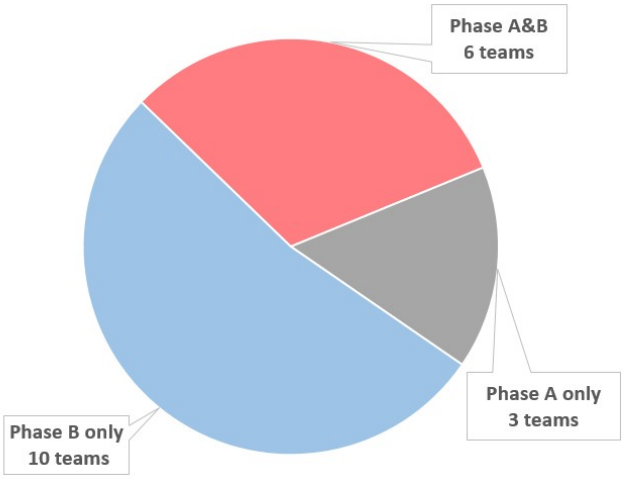

Figure 4: The distribution of participant teams in the BioASQ task 11b into phases.

11b. This could be due to the particularities of open questions in Synergy, such as the volatility of answers and the evolving nature of the relevant knowledge which make the task more challenging than traditional question answering.

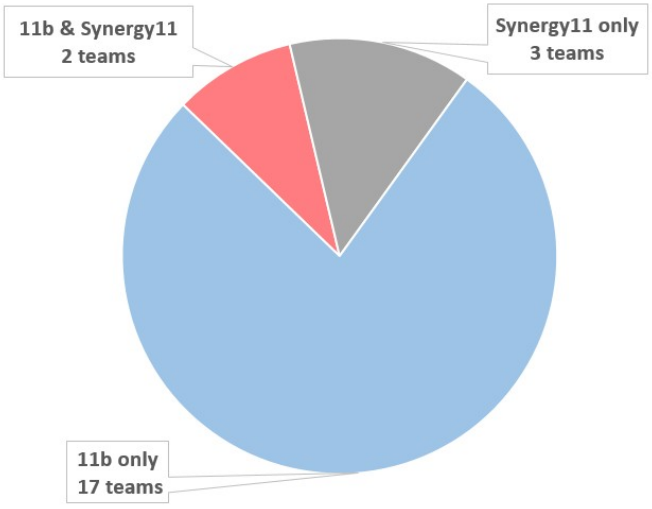

Figure 5: The overlap of participant teams in the BioASQ task 11b and Synergy11.

4. Conclusions

In this paper, we presented the eleventh version of the BioASQ tasks b and Synergy. Both tasks are already established through the previous versions of the challenge. The participation of

teams was comparable to last year's version of these tasks with a slight decrease. On the other hand, we noticed a high number of newly registered teams. Therefore, we believe that the challenge and the datasets developed for its tasks increase the research community's interest in question answering and encourage the development of better solutions to aid biomedical researchers' access to the abundance of biomedical knowledge.

Acknowledgments

Google was a proud sponsor of the BioASQ Challenge in 2022. The eleventh edition of BioASQ is also sponsored by Ovid. Atypon Systems Inc. is also sponsoring this edition of BioASQ. The MEDLINE/PubMed data resources considered in this work were accessed courtesy of the U.S. National Library of Medicine.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), ????

[2] A. Nentidis, A. Krithara, G. Paliouras, E. Farre-Maduell, S. Lima-Lopez, M. Krallinger, BioASQ at CLEF2023: The Eleventh Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 577–584.

[3] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, A. Miranda-Escalada, L. Gasco, M. Krallinger, G. Paliouras, 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), volume 13390 LNCS, 2022, pp. 337–361. doi:10.1007/978-3-031-13643-6_22. arXiv:2210.06852.

[4] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 10a, 10b and Synergy10 in CLEF2022, in: Proceedings of the 10th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, 2022. URL: https://ceur-ws.org/Vol-3180/paper-10.pdf.

[5] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[6] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, 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 se-

mantic indexing and question answering competition, BMC Bioinformatics 16 (2015) 138. doi:10.1186/s12859-015-0564-6.

[7] 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.

[8] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.

[9] 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.