Overview of BioASQ Tasks 12b and Synergy12 in CLEF2024

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

This paper presents an overview of the twelfth edition of BioASQ challenge, which is part of the Conference and Labs of the Evaluation Forum (CLEF) 2024. BioASQ serves as a key platform for advancing large-scale biomedical information retrieval and question-answering (QA) systems and includes a variety of tasks. In this paper, we present an overview of the QA tasks b and Synergy of the BioASQ 12 challenge. Notably, BioASQ 12 introduces an additional phase (Phase A+) for task b, further expanding the challenge's scope. This year, 27 teams with more than 100 systems participated in the two tasks of the challenge, with 26 of them focusing on task 12b, and 4 on task Synergy. While the total number of participating teams varies year-to-year, the increasing rate of new team participation, as observed in previous editions, highlights the impact of BioASQ in fostering robust biomedical QA solutions.

Keywords

Biomedical knowledge, Semantic Indexing, Question Answering

1. Introduction

This paper gives a brief overview of the twelfth edition of the BioASQ challenge (2024), focusing on shared tasks 12b and Synergy12. Furthermore, we describe the corresponding datasets used to train and evaluate participating systems. Details of tasks 12b and Synergy12, which ran from March to May and January to February 2024, respectively, are provided in Section 2. Section 3 provides a brief overview of the participation in these two tasks. A comprehensive analysis of the methodologies employed by participating systems will be included in the BioASQ workshop proceedings in [1]. We conclude the paper with a brief discussion and our key findings.

2. Overview of the Tasks

The twelfth edition of the BioASQ challenge consisted of four tasks: (1) a biomedical question answering task (task b), (2) a task on biomedical question answering for open developing issues (task Synergy), both tasks considering documents in English, (3) a new task focused on the automatic detection and normalization of mentions of four clinical entity types (task MultiCardioNER), considering cardiology clinical case documents in Spanish, English, and Italian, and (4) a new task on NLP challenges on biomedical nested named entity recognition (NER) systems for English and Russian languages (task BIONNE) [2].

In this paper, we describe the current versions of the first two established tasks, referring to them as Task 12b and Task Synergy12 within the context of the twelfth BioASQ edition. Detailed descriptions of the MultiCardioNER and BIONNE tasks can be found in [3] and [4], respectively. Additionally, a detailed introduction to the BioASQ challenge and its initial task structure is available in [5].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
tasosnent@iit.demokritos.gr (A. Nentidis); gkatsibras@iit.demokritos.gr (G. Katsimpras); akrithara@iit.demokritos.gr (A. Krithara); paliourg@iit.demokritos.gr (G. Paliouras)

2.1. Biomedical semantic QA - Task 12b

Task 12b introduces a comprehensive question-answering challenge in the biomedical field. Participants are required to create systems that address all stages of question-answering. Similar to previous editions, the task focuses on four question types: 'yes/no,' 'factoid,' 'list,' and 'summary' questions [6].

In the twelfth edition of the BioASQ Challenge, participating teams were provided with a new version of the BioASQ QA training dataset, containing 5,046 questions that had been annotated with relevant golden elements and answers from previous task versions [7]. These questions served as the basis for developing their systems. The details of both the training and testing sets for task 12b are outlined in Table 1. These statistics reveal that the average number of documents and snippets in training data is significantly larger than in the test batches. This can be attributed to two main factors. First, in the early years of BioASQ the annotation with relevant documents and snippets by the experts was exhaustive, in an attempt to identify as many relevant items as possible in the corpus. These questions are part of the training datasets affecting the average number of relevant items per question. Currently, only a sufficient number of relevant answers is required when the initial version of the data is developed. Still, when the participants submit their responses, the experts assess the submitted items and enrich the ground-truth data with potential additional relevant items detected by the participants. The numbers of relevant items for the test sets in Table 1 are preliminary, before the enrichment by the assessment process which is still in progress. The final evaluation of the participants will be against these enriched relevant items, ensuring that all the submitted items that are relevant are indeed handled as such.

Table 1
Statistics on the training and test datasets of Task 12b. The numbers for the documents and snippets refer to averages per question.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	5,049	1,357	967	1,515	1,210	9.06	11.91
Test 1	85	25	21	21	18	3.20	4.36
Test 2	85	26	18	19	22	2.72	3.69
Test 3	85	24	19	26	16	2.45	3.36
Test 4	85	27	22	19	17	2.18	3.44
Total	5,389	1,459	1,047	1,600	1,283	8.65	11.4

Unlike previous challenges, task 12b consisted of three phases. An additional phase (Phase A+) of submitting answers (exact and/or ideal), before the golden documents and snippets become available, i.e. answers based on documents identified by participant systems, was provided. The goal of this additional phase is to compare the performance of the competing systems with and without golden feedback. Task 12b was divided into four independent bi-weekly batches and the three phases for each batch run for two consecutive days. The three phases of task 12b consist of: (phase A) the retrieval of the required information, (phase A+) answering the question without golden feedback and (phase B) answering the question with golden feedback, 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. In the current year, the test sets comprised 85 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, these questions were also released in phase A+ 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. Finally, during phase B, manually selected relevant articles and snippets related to these questions were also made available, and participating systems were once again asked to provide *exact answers* and *ideal answers*.

2.2. Synergy12 Task

In the BioASQ challenge, the Synergy task was introduced in its ninth edition to foster collaboration between biomedical experts studying COVID-19 and automated question-answering systems participating in BioASQ. The goal is to create a synergy where experts assess system responses, and this feedback is used to iteratively improve the systems.

In the process depicted in Figure 1, competing systems provide their initial responses to open questions related to emerging problems. These responses, along with relevant documents and snippets, are evaluated by experts. Subsequently, the experts provide feedback to the systems and address any new or pending questions.

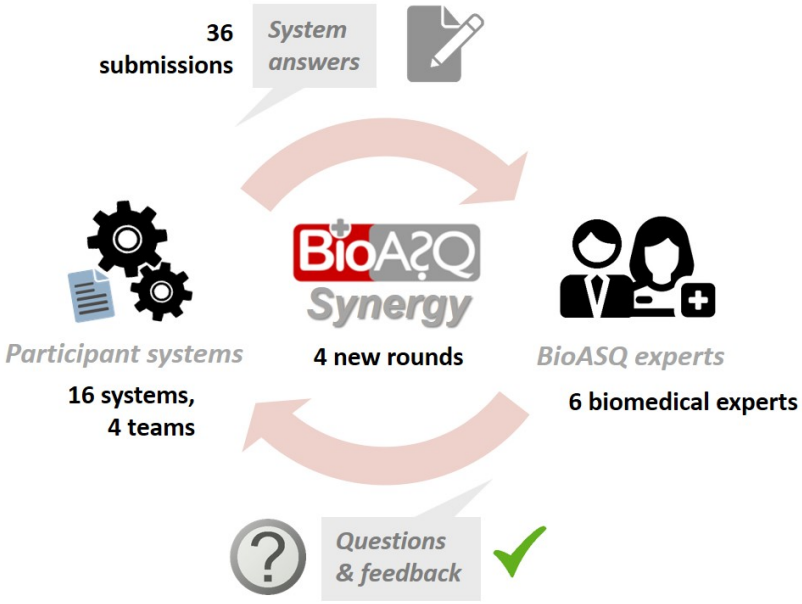

Figure 1: The iterative dialogue between the experts and the systems in the BioASQ Synergy12 task on question answering for open developing problems.

This version of the Synergy task (Synergy12) involved a series of four rounds, with a two-week interval between each round. The task focused on emerging issues, drawing from relevant documents in the current PubMed version. As with earlier versions, the questions posed were open-ended, allowing for dynamic responses.

In the Synergy task, during each round, the system responses and expert feedback address the same questions, unless those questions have already been closed by experts due to receiving a comprehensive and definite answer. Specifically, in Synergy12, a group of six biomedical experts contributed a total of 72 open biomedical questions. They evaluated the retrieved material (including documents and snippets) and the responses submitted by participating systems in all four rounds. Table 2 shows the details of the datasets used in task Synergy12.

Table 2
Statistics on the datasets of Task Synergy12. "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.

Round	Size	Yes/No	List	Factoid	Summary	Answer ready
1	72	11	29	17	15	33
2	72	11	29	18	14	46
3	64	10	24	16	14	50
4	64	10	24	17	13	57

Synergy12, similar to task 12b, explores four question types: yes/no, factoid, list, and summary, and two types of answers, exact and ideal. Moreover, the evaluation of systems relies on the same measures used in task 12b. Upon completing the Synergy12 task, relevant material was identified for answering roughly 78% of the questions. Additionally, around 51% of the questions had at least one ideal answer submitted by the systems, which was deemed satisfactory by the expert who posed the question.

3. Overview of participation

In this year's BioASQ challenge, over 100 distinct systems engaged in tasks 12b and Synergy12 with a total of 27 teams. Specifically, 26 of these teams submitted on task 12b and 4 on task Synergy12. Furthermore, Figure 2 demonstrates the global interest in the challenge, with participating teams representing various countries worldwide.

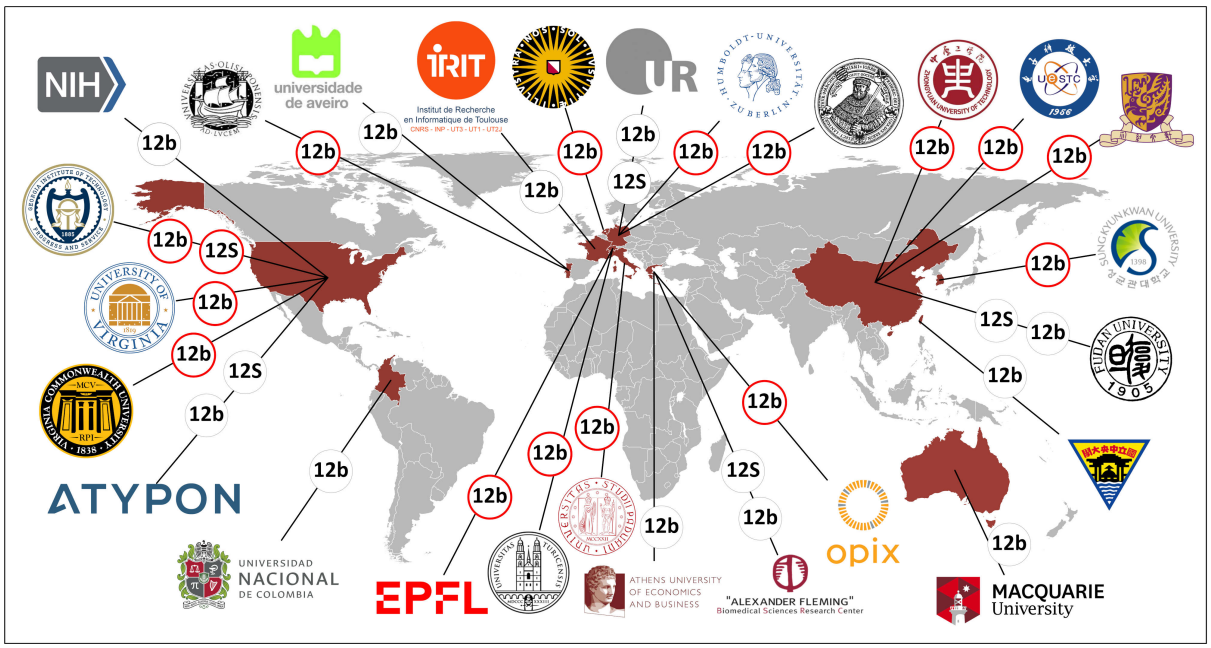

Figure 2: The world-wide distribution of teams participating in the tasks 12b and Synergy12, based on institution affiliations. A red circle indicates a newly registered team.

In line with previous years, task b attracted more participants than Synergy. Furthermore, Figure 3 illustrates a considerable increase in the total number of participating teams this year in comparison to last year. Additionally, the high percentage of teams joining the BioASQ challenge for the first time (indicated by red circles in Figure 2), indicates the enduring interest of the community in large-scale biomedical semantic indexing and question answering. Specifically, 16 new teams participated in this year's BioASQ tasks b and Synergy.

3.1. Task 12b

In task 12b, a total of 26 teams participated this year, contributing 89 different systems across all three phases A, A+, and B. Specifically, 18 teams with 64 systems competed in phase A, 8 teams with 34 systems participated in A+, and phase B saw 16 participants with 54 systems. Notably, 8 teams were involved in all three phases, as depicted in Figure 4.

3.2. Synergy Task

In task Synergy12, 4 teams participated this year contributing a total of 16 distinct systems. Since Synergy12 shares some common concepts with task 12b, a few teams participated in both tasks.

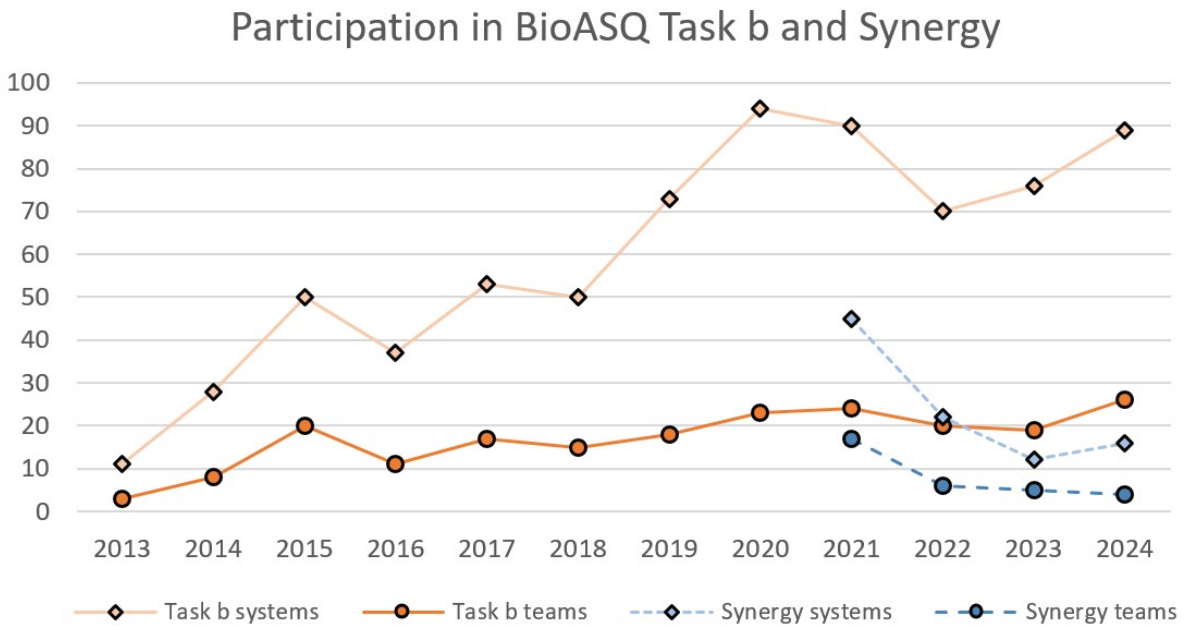

Figure 3: The evolution of participation in the BioASQ task b and Synergy in the twelve years of BioASQ.

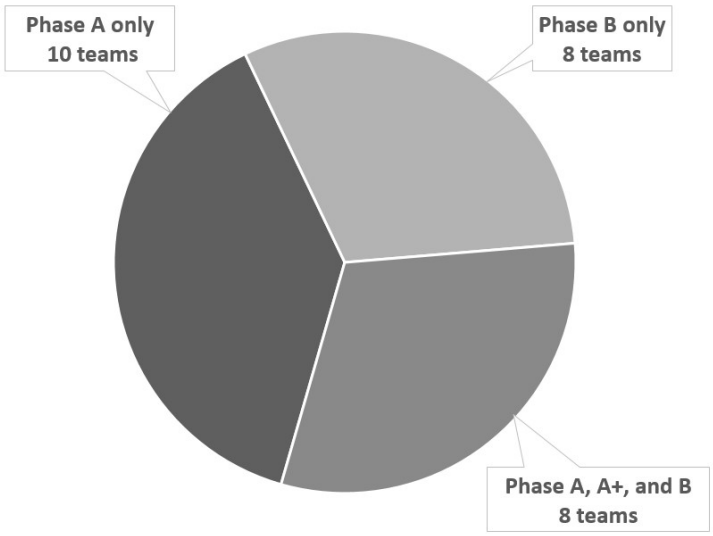

Figure 4: The distribution of participant teams in the BioASQ task 12b into phases.

Specifically, 3 teams engaged in both task 12b and Synergy12, as depicted in Figure 5. However, consistent with previous versions of the tasks, fewer teams participated in Synergy12 compared to task 12b. 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 pose greater challenges than traditional question answering.

4. Conclusions

In this paper, we introduced the twelfth 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

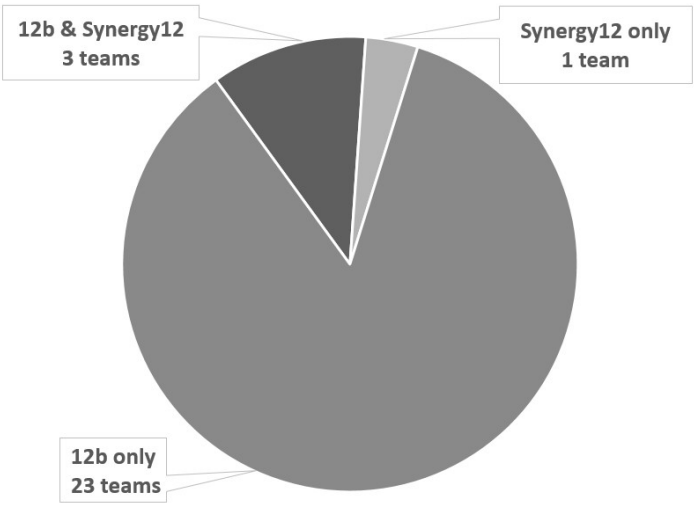

Figure 5: The overlap of participant teams in the BioASQ task 12b and Synergy12.

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

In this paper, we introduced the twelfth version of the BioASQ challenge, focusing on tasks b and Synergy. These tasks have been well-established through previous versions of the challenge. Notably, team participation has grown and we observed a significant increase in newly registered teams. As a result, we consider that the challenge, along with the associated datasets, has sparked greater interest within the research community and continues to advance the field of biomedical semantic indexing and question answering.

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 Atypon Systems inc. 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, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] A. Nentidis, A. Krithara, G. Paliouras, M. Krallinger, L. G. Sanchez, S. Lima, E. Farre, N. Loukachevitch, V. Davydova, E. Tutubalina, BioASQ at CLEF2024: The Twelfth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: European Conference on Information Retrieval, Springer, 2024, pp. 490–497.

[3] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Rodríguez-Ortega, L. Lilli, J. Lenkowicz, G. Ceroni, J. Kossoff, A. Shah, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.

[4] V. Davydova, N. Loukachevitch, E. Tutubalina, Overview of BioNNE Task on Biomedical Nested Named Entity Recognition at BioASQ 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

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

[6] 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.

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