

Overview of BioASQ Tasks 9a, 9b and Synergy in CLEF2021

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

¹NCSR Demokritos, Athens, Greece

²Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

In this paper, we present an overview of the tasks a and b of the ninth edition of BioASQ challenge, together with a newly introduced task on question answering for developing problems called Synergy. All these tasks ran as part of the BioASQ challenge lab in the Conference and Labs of the Evaluation Forum (CLEF) 2021. The main focus of BioASQ is to promote methodologies and systems for large-scale biomedical semantic indexing and question answering. This is achieved through the organization of yearly challenges which enable the participation of teams around the world in developing and comparing their methods on the same benchmark datasets. This year, 42 teams with more than 170 systems participated in the four tasks of the challenge, with six of them focusing on task 9a, 24 on task 9b and 15 on task Synergy. Correspondingly to the previous years, the participation has increased, indicating the established presence of BioASQ challenge in the field.

Keywords

Biomedical knowledge, Semantic Indexing, Question Answering

1. Introduction


In this paper we give an overview of the shared tasks 9a and 9b of the ninth edition of the BioASQ challenge in 2021, as well as of the new task of BioASQ challenge called Synergy. In addition, we present in detail the datasets that were used in each task. In section 2, we provide an overview the shared tasks 9a and 9b, that took place from February to May 2021, the newly introduced task Synergy, which took place from December 2020 to February 2021 and from May to June 2021, as well as the corresponding datasets developed for training and testing the participating systems. In section 3, we summarize the participation in these three tasks. Detailed descriptions for some of the systems will be available in the proceedings of the BioASQ lab. Our conclusions are presented in section 4, along with a brief discussion about the ninth version of the BioASQ tasks a, b and Synergy.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ 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)



© 2021 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)

2. Overview of the Tasks

Taken together, in the ninth version of the BioASQ challenge were offered four tasks: (1) a large-scale biomedical semantic indexing task (task 9a), (2) a biomedical question answering task (task 9b), both considering documents in English, (3) a medical semantic indexing in Spanish (task MESINESP9), and (4) a new task on biomedical question answering (task Synergy). In this section, apart from providing a brief description of the two established tasks (9a and 9b) with focus on differences from previous versions of the challenge [1], we also concisely outline the Synergy task. For tasks 9a and 9b, a detailed overview of the initial tasks can be found in [2], which also describes the general structure of BioASQ.

2.1. Large-scale semantic indexing - Task 9a

Table 1

Statistics on test datasets for Task 9a.

Batch	Articles	Annotated Articles	Labels per Article
1	7967	7808	12.61
	10053	9987	12.40
	4870	4854	12.16
	5758	5735	12.34
	5770	5666	12.49
Total	34418	34050	12.42
2	6376	6374	12.39
	9101	6403	11.76
	7013	6590	12.15
	6070	5914	12.62
	6151	5904	12.63
Total	34711	31185	12.30
3	5890	5730	12.81
	10818	9910	13.03
	4022	3493	12.21
	5373	4005	12.62
	5325	2351	12.97
Total	31428	25489	12.71

Task 9a focuses on classifying articles from the PubMed/MedLine¹ digital library into concepts of the MeSH hierarchy. Specifically, the test sets for the evaluation of the competing systems consist of new PubMed articles that are not yet annotated by the indexers in NLM. A more detailed view of each test set can be seen in Table 1. Similarly to the previous years, the task is divided into three independent batches of 5 weekly test sets each. Two scenarios are provided: i) on-line and ii) large-scale. The test sets is a collection of new articles without any restriction on the journal published. To evaluate the participating systems we use standard flat and hierarchical information retrieval measures as in previous versions of the task [3]. In the case the annotations from the NLM indexers are available, hierarchical measures are used as

¹<https://pubmed.ncbi.nlm.nih.gov/>

well. As before, for each test set, participants are required to submit their answers in 21 hours. Also, there was a training dataset available for Task 9a that is composed of 15,559,157 articles with 12.68 labels per article, on average, and covering 29,369 distinct MeSH labels in total.

2.2. Biomedical semantic QA - Task 9b

The aim of Task 9b is to enable the competing teams to develop systems for all the stages of question answering in the biomedical domain by introducing a large-scale question answering challenge. Akin to the previous versions of the task, four types of questions are considered: “yes/no”, “factoid”, “list” and “summary” questions [3]. For this task, the available training dataset contains 3,742 questions which are annotated with golden relevant elements and answers from previous versions of the task. The dataset is used by the participating teams to develop their systems. The details of both training and testing sets are depicted in Table 2.

Table 2

Statistics on the training and test datasets of Task 9b. The numbers for the documents and snippets refer to averages per question.

Batch	Size	Yes/No	List	Factoid	Summary	Documents	Snippets
Train	3,743	1033	719	1092	899	9.43	12.32
Test 1	100	27	21	29	23	3.40	4.66
Test 2	100	22	20	34	24	3.43	4.88
Test 3	100	26	19	37	18	3.21	4.29
Test 4	100	25	19	28	28	3.10	4.01
Test 5	100	19	18	36	27	3.59	4.69
Total	4,243	1152	816	1256	1019	8.71	11.40

Task 9b is divided into two phases: (phase A) the retrieval of the required information and (phase B) answering the question. Moreover, it is split into five independent bi-weekly batches and the two phases for each batch run during two consecutive days. In each phase, the participants receive the corresponding test set and have 24 hours to submit the answers of their systems. More precisely, in phase A, a test set of 100 questions written in English is released and the participants are expected to identify and submit relevant elements from designated resources, including PubMed/MedLine articles, snippets extracted from these articles, concepts and RDF triples. In phase B, the manually selected relevant articles and snippets for these 100 questions are also released and the participating systems are 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. Synergy Task

The current BioASQ task B is structured in a sequence of phases. First comes the annotation phase; then with a partial overlap runs the challenge; and only when this is finished does the assessment phase start. This leads to minimal interaction between the experts and the participating systems, which is acceptable due to the nature of the questions that are generated. Namely, we are looking for interesting research questions that have a clear, undisputed answer.

This model is less suitable to developing biomedical research topics, such as the case of COVID-19, where new issues appear every day and most of them remain open for some time. A more interactive approach is needed for such cases, aiming at a synergy between the biomedical experts and the automated question answering systems. We envision such an approach as a continuous dialog, where experts issue open questions to the systems and the systems respond to the questions. Then, the experts assess the responses, and their assessment is fed back to the systems, in order to help improving them. Then, the process continues iteratively with new feedback and new system predictions.

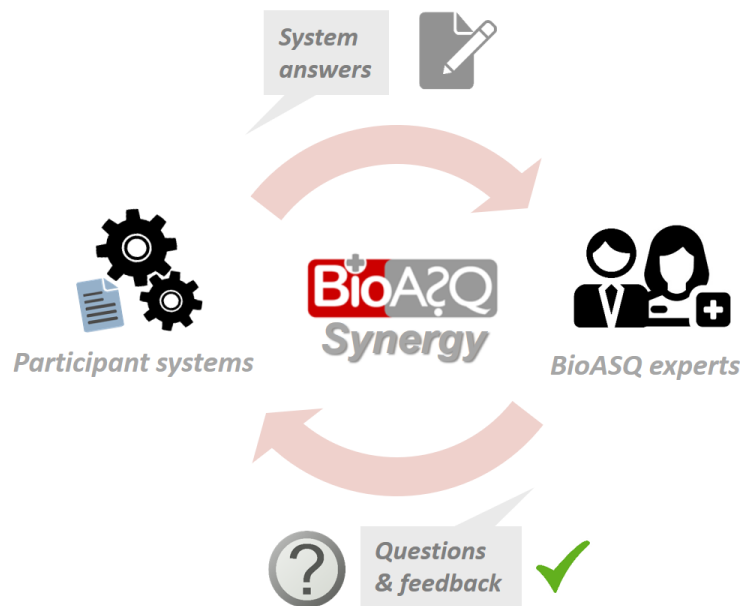


Figure 1: The iterative dialogue between the experts and the systems in the BioASQ Synergy task on question answering for COVID-19.

Fig. 1 sketches this vision, which motivates the new BioASQ Synergy task. This new task allows biomedical experts to pose unanswered questions for developing problems, such as COVID-19. Participating systems attempt to provide answers, together with supporting material (relevant documents and snippets), which in turn are assessed by the experts and fed back to the systems, together with new questions. At the same time, we are adapting the BioASQ infrastructure and expand the community to address new developing public health issues in the future. In this introductory year, Task Synergy took place in two versions. Each version was structured into four rounds, of systems responses and expert feedback for the same questions. However, some new questions or new modified versions of some questions could be added to the test sets. The details of the datasets used in task Synergy are available in Table 3.

Contrary to the task B, this task was not structured into phases, but both relevant material and answers were received together. However, for new questions only relevant material (documents and snippets) is required until the expert considers that enough material has been gathered

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”.

Version	Round	Size	Yes/No	List	Factoid	Summary	Answered	Feedback
1	1	108	33	22	17	36	0	0
1	2	113	34	25	18	36	53	101
1	3	113	34	25	18	36	80	97
1	4	113	34	25	18	36	86	103
2	1	95	31	22	18	24	6	95
2	2	90	27	22	18	23	10	90
2	3	66	17	14	18	17	25	66
2	4	63	15	14	17	17	33	63

during the previous round and mark the questions as “ready to answer”. When a question receives a satisfactory answer that is not expected to change, the expert can mark the question as “closed”, indicating that no more material and answers are needed for it.

In each round of this task, we consider material from the current version of the COVID-19 Open Research Dataset (CORD-19) [4] to reflect the rapid developments in the field. As in task B, four types of questions are supported, namely yes/no, factoid, list, and summary, and two types of answers, exact and ideal. The evaluation of the systems is based on the measures used in Task 9b. Nevertheless, for the information retrieval part we focus on new material. Therefore, material already assessed in previous rounds, available in the expert feedback, should not be re-submitted. Overall, through this process, we aim to facilitate the incremental understanding of COVID-19 and contribute to the discovery of new solutions.

3. Overview of participation

Overall, 37 teams from institutes around the world participated in the tasks 9a, 9b and Synergy of the challenge with more than 120 distinct systems. Particularly, six of these teams submitted on task 9a, 24 on task 9b and 15 on task Synergy. Furthermore, we can see from Fig. 2, that the participating teams in tasks 9a, 9b and Synergy are originating from various countries around the world, indicating the international interest in the challenge. We observe that a shift towards the most complex question answering task b, already observed in previous years of the challenge, is still apparent this year, as the number of participating teams is slightly increased as shown in Fig. 3. Detailed descriptions for some of the systems will be available at the proceedings of the workshop.

3.1. Task 9a

In this year’s Task 9a, 6 teams competed 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 five different systems, the Fudan University & Atypon team that participated with 4 systems, and the team from the University of Vigo and the University of A Coruña that participated with two systems. On the other hand, two new

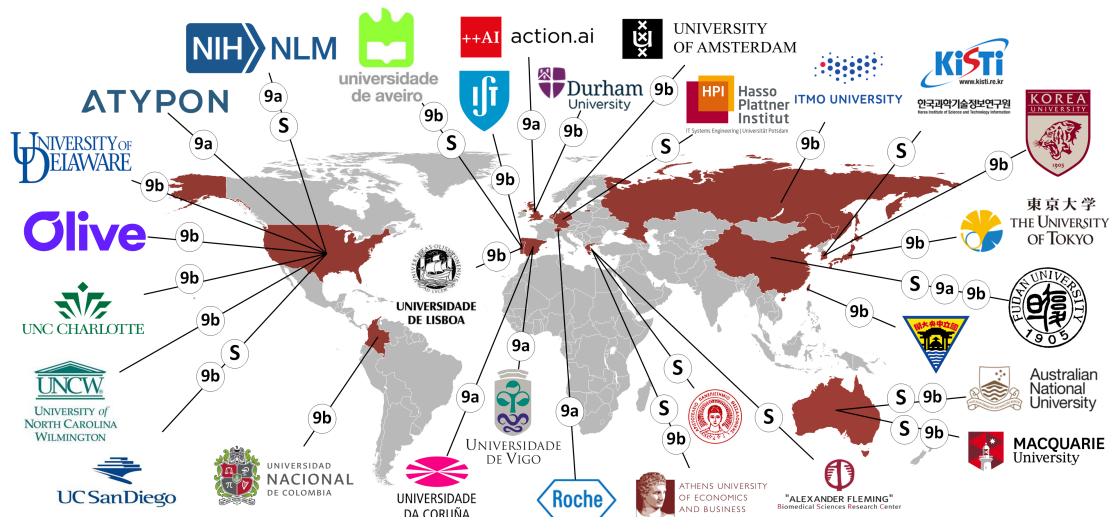


Figure 2: The world-wide distribution of teams participating in the tasks 9a, 9b and Synergy (S), based on institution affiliations.

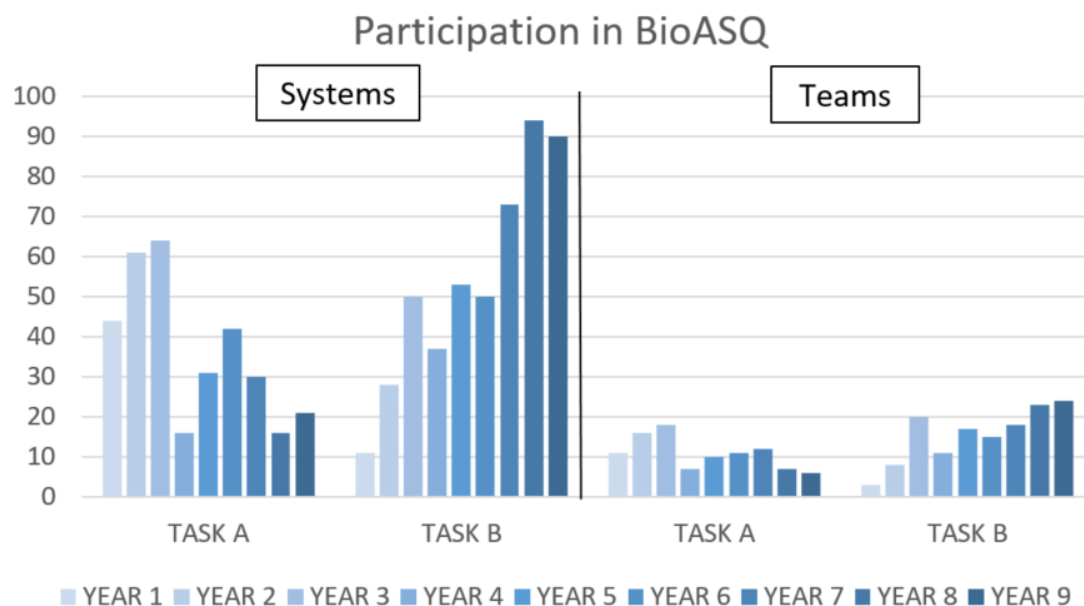


Figure 3: The evolution of participant teams in the BioASQ task a and b in the nine years of BioASQ.

teams from Roche and Atypon competed for the first time, submitting results with five and three systems respectively.

3.2. Task 9b

This year, 90 different systems have submitted predictions for Task 9b in total, for both phases A and B. These systems were developed by 24 teams. In phase A, 9 teams participated, submitting results from 34 systems. In phase B, the numbers of participants and systems were 20 and 70 respectively. There were only three teams that engaged in both phases.

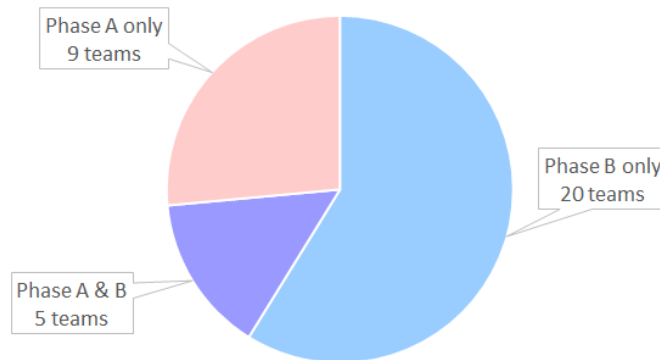


Figure 4: The distribution of participant teams in the BioASQ task 9b into phases.

3.3. Synergy Task

In the first two versions of the new task Synergy, introduced this year, 15 teams participated submitting the results from 39 distinct systems. Although significantly different from task b, this task is still about biomedical information retrieval and question answering, therefore the systems for both tasks are expected to share some common ideas and techniques. Therefore, some teams participated in both tasks. In particular, 8 teams participated both task 9b and Synergy as shown in Fig. 5.

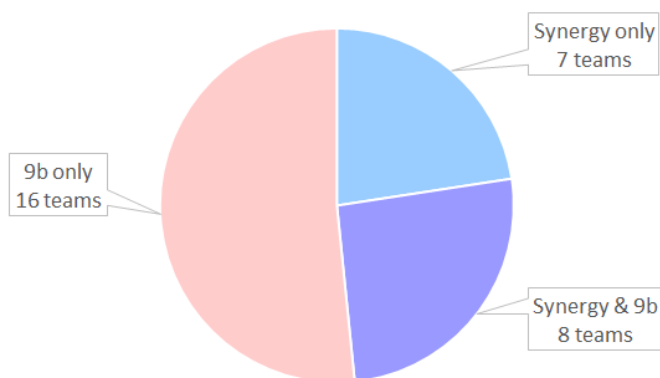


Figure 5: The overlap of participant teams in the BioASQ task 9b and Synergy.

4. Conclusions

This paper provides an overview of the ninth version of the BioASQ tasks a and b, along with the newly introduced task Synergy. Tasks 9a and 9b, are already established through the previous eight years of the challenge, and together with the MESINESP9 task on semantic indexing of medical content in Spanish, and the Synergy task, which ran for the first time, consisted the ninth edition of the BioASQ challenge.

Overall, the BioASQ challenge has been matured and established its presence through these years. Besides continuing the annual tasks a and b, this year we offered a new biomedical question answering task, Synergy. Similar to previous years, the participation of teams increased and therefore, we consider that the challenge keeps meeting its goal to push the research frontier in biomedical semantic indexing and question answering.

Acknowledgments

Google was a proud sponsor of the BioASQ Challenge in 2020. The ninth edition of BioASQ is also sponsored by the Atypon Systems inc. BioASQ is grateful to NLM for providing the baselines for task 9a and to the CMU team for providing the baselines for task 9b. The MESINESP 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 MESINESP task.

References

- [1] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, Overview of bioasq 8a and 8b: Results of the eighth edition of the bioasq tasks a and b (2020).
- [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).