http://www.bioasq.org

Report on challenge operation and technical support 2

Georgios Balikas, Ioannis Partalas, Nicolas Baskiotis, Thierry Artieres, Eric Gaussier, Patrick Gallinari and Ioannis Vetsikas

Status: Final (Version 1.0)

September 2014

Project

Project ref.no.	FP7-318652
Project acronym	BioASQ
Project full title	A challenge on large-scale biomedical semantic indexing and question answering
Porject site	http://www.bioasq.org
Project start	October 2012
Project duration	2 years
EC Project Officer	Martina Eydner

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D4.8
Deliverable title	Report on challenge operation and technical support 2
Contractual date of delivery	M22 (July 2014)
Actual date of delivery	September 2014
Relevant Task(s)	WP4/Task 4.2
Partner Responsible	UPMC
Other contributors	UJF
Number of pages	16
Author(s)	Georgios Balikas, Ioannis Partalas, Nicolas Baskiotis, Thierry Artieres, Eric Gaussier, Patrick Gallinari and Ioannis Vetsikas
Internal Reviewers	Axel Ngonga Ngomo
Status & version	Final
Keywords	BioASQ, platform, challenge operation

Executive Summary

The deliverable describes the second year of the BIOASQ challenge as it was monitored in BIOASQ Participants Area[1]. The BIOASQ Participants Area is the online platform that was developed to provide the necessary functionality for data exchange, evaluation and participant support during the BIOASQ challenge.

The second BIOASQ challenge consisted of two tasks: semantic indexing (task 2a) and question answering (2b). A series of test datasets pertaining to both Task 2a[2] and Task 2b[3] were released on the platform. Participants downloaded the datasets and responded with the required answers and results that were produced by their systems. The system answers were evaluated against correct, human answers and based on this evaluation the winners of the second year of BIOASQ challenge were announced.

The participation in each task can be summarized as follows:

- 61 systems participated by 18 different participating teams for the semantic indexing task (task 2a), of which between 25 and 45 participated in each batch.

- The question annotation task (task 2b-phase A) was tackled by 22 systems, which were developed by 8 different organizations. Between 15 and 19 of these systems addressed each batch.

- The question answering task (task 2b-phase B) was tackled by 18 different systems, developed by 7 different organizations. Between 9 and 15 of these systems submitted results in each batch.

For task 2a, 75,950 articles were released, of which 42,170 have already been annotated (59%) For Task 2b, 5 test batches were released. Each batch consisted of 100 questions. In total, 500 questions were released.

This deliverable contains detailed information and statistics with respect on the following:

- The datasets for both tasks of the challenge for the second year of the challenge,

- The participation in both tasks of the challenge,

- The support that was provided from the BIOASQ team in the participants, and

---

[1] http://bioasq.lip6.fr

[2] The Task 2a datasets are available online in http://bioasq.lip6.fr/Tasks/2a/

[3] The Task 2b datasets are available online in http://bioasq.lip6.fr/Tasks/2b/phaseB/

- The problems that occurred in the operation as well as the actions that are proposed to overcome them.

Contents

List of Figures

List of Tables

1

Introduction

BIOASQ initiated a series of challenges on biomedical semantic indexing and question answering. The motivation behind the challenge is to push for solutions to the information access problem biomedical experts face and concerns their difficulty to synthesize and filter quickly, accurate and specialized information that comes from large and fast-growing sources.

The project organised two tasks within two concecutive years. During its second year (October 2013-September 2014) task 2a and task 2b were organised. For each task the BIOASQ team released test sets following a predefined and announced schedule. Participants were allowed to download the test sets and submit their results using the online BIOASQ Participants Area[1] (hereafter platform) within a limited time window. A short description of the tasks of the challenge follows. For a more detailed description, please visit `http://www.bioasq.org`.

Task 2a

Task 2a, entitled "Large scale online biomedical semantic indexing", deals with large scale classification of biomedical documents onto ontology concepts. It simulates the process that is followed in PubMed[2] by human curators. PubMed is a public, online database hosted in the US where new articles are uploaded in a daily basis and are annotated with concepts from the MeSH[3] hierarchy. The gap between the submission of an article in PubMed and its annotation is used by the BIOASQ team in order to release test sets that consist of non-annotated articles. Participants submit their system's estimations for the annotations of those articles. The evaluation of the participant systems is performed when the human annotations from PubMed become available. The articles and the deadlines for the test sets are selected in a way that prevents cheating and ensures a short annotation period.

There were 15 test sets released during the second year of the challenge, organized in 3 test batches. They were released as a continuation of the first year's test sets. Practically, since the beginning of task 1a the BIOASQ team never stopped releasing data for the semantic annotation task. In particular, after the end of the first year of the challenge, we continued releasing test sets in an off-challenge mode.

---

[1]The BIOASQ Participants Area is deployed under `http://bioasq.lip6.fr`
[2]`http://www.ncbi.nlm.nih.gov/pubmed`
[3]`http://www.ncbi.nlm.nih.gov/mesh`

Participants could use those test datasets in order to prepare and tune their systems for the second year of the challenge. The official part of the second year of the challenge (task 2a) begun on February 4th. The 15 test sets were released on a weekly basis. After the end of the official period of task 2a, we continue releasing test datasets for task 2a. Figure 1.1 shows the schedule of Task 2a. The date of the first test set of each test group is marked in the figure. The task's last official test set release was on the 20th of May, 2014.
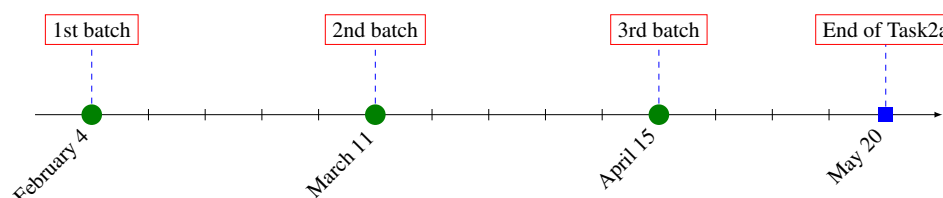

Figure 1.1: The time plan of Task 2a.

Task 2b

Task 2b, entitled "Biomedical Semantic Question Answering", examines the ability of the participating systems to annotate questions with concepts from relevant ontologies and return "exact" and paragraph sized, "ideal" answers. The network of ten experts around Europe, which was established during the first year of the project, created a benchmark dataset of 500 questions using an "Annotation Tool" developed from the BIOASQ consortium for this reason. The task was organised in two phases:

- Phase A: The BIOASQ team released questions from the benchmark datasets. The participating systems had to respond with relevant concepts from designated terminologies and ontologies, relevant articles in English from designated article repositories, relevant snippets from the relevant articles, and relevant RDF triples from designated ontologies.

- Phase B: The BIOASQ team released questions and gold (correct) relevant concepts, articles, snippets, and RDF triples from the benchmark datasets. The participating systems had to respond with exact answers (e.g., named entities in the case of factoid questions) and ideal answers (paragraph-sized summaries), both written in English. For the synthesis of the answers, using the provided gold annotations was sufficient. However, users were also allowed to use the annotations their systems estimated in Phase A.

The process of releasing test sets was repeated five times. The datasets for Phase B, were released after the expiration of Phase A. Figure 1.2 shows the schedule that was followed for Task 2b. For example, the first test dataset concerning phase A of Task 2b was released on 5th of March, 2014. More information on the process followed during Task 2b can be found in the guidelines in Androutsopoulos et al. (2013). More information about the "Annotation Tool" is available in Ngonga Ngomo et al. (2013) and Ngonga Ngomo et al. (2014).
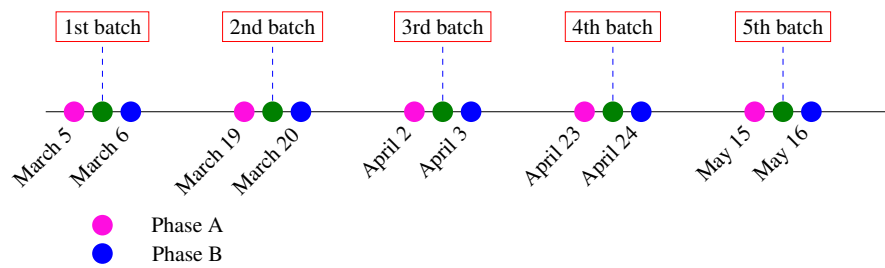
Figure 1.2: The time plan of Task 2b. The two phases for each batch run in consecutive days.

2

Challenge operation

The chapter contains a description of the challenge from the scope of data and participation as it was seen by the online platform. The platform integrates the necessary functionality the participants used during the challenge. In addition, it is where the test sets of both tasks of the challenge are released. As a result, monitoring its activity enables us to infer about the challenge operation. The chapter is organised as follows:

- The first section summarizes the functionality that is integrated in the platform,

- The second and the third sections describe the data and the participation in Task 2a,

- The fourth and the fifth sections describe the data and the participation in Task 2b and

- The sixth section provides information on the support of the users.

An overview of the traffic in the platform is presented in Figure 2.1 that comes from Google Analytics. There were more than 60 unique visitors in the platform on Mondays, with a maximum of around 100 unique visitors on 24th of March 2014. Please note that the spikes on Mondays, Wednesdays and Thursdays correspond to datasets releases.
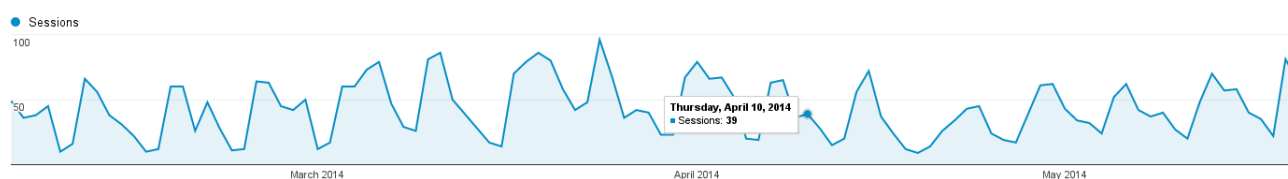

Figure 2.1: Google Analytics: Traffic in the platform during the second year of the challenge.

2.1 Functionality of the platform

The platform was designed to be user-friendly. A key concept during its development was to make it simple and user-friendly so that participants could find easily the information about the challenge

and the infrastructure that the BIOASQ team has integrated. During the second year of the challenge the BIOASQ team tried to apply further those principles and minor changes to the appearance of the platform were applied to make navigation easier. In addition, oracles for both tasks of the challenge have been developed and integrated in the platform as an extra feature. Oracles allow participants to submit results for previous BioASQ test sets in an off-challenge mode and receive the scores of the corresponding evaluation measures. For more information about the platform of the challenge and the oracles, please consult Balikas et al. (2014).

A short description of the provided functionality follows:

- Registration: Participants can register to the platform by filling a simple form. Registered users gain access to the datasets, to the forum, to the BIOASQ announcements and to the result submission forms and web services.

- BIOASQ datasets: The datasets include training sets for each task of the challenge and test sets that are released periodically. They are served as JSON[1] (JavaScript Object Notation) strings which are platform-independent and human-readable. The datasets can be accessed via links in the platform and via web services. The datasets remain available in the platform after their expiration but submitting results is disabled.

- Submissions of results: Participants can submit their results for the active testsets by using a simple form in the platform or by using web services. Participants can submit results more than once for a particular system before the dataset deadline.

- Support: There are detailed online guidelines that address the main points of the challenge. There is a BIOASQ Discussions Area (forum) where registered participants can discuss about the challenge. There is also a contact form that participants can use to contact the BIOASQ team directly.

- Evaluation results: Participants can browse tables that contain the evaluation measures for their systems and compare their system's performance with other users.

- Oracles: Participants can use the existing functionality to submit results for past test datasets of the challenge and receive immediate feedback about their performance.

More information on the platform functionality is provided in Balikas et al. (2013).

2.2 Datasets for Task 2a

In the context of Task 2a two different training datasets and 15 test datasets (organised in three test groups) were released. The difference between the two datasets released for 2014 lies in the sources and the publication dates of the articles: the articles of the small test set were published during the last three years in the set of journals[2] the BIOASQ team has selected for creating the official test set of the challenge. On the other hand, the articles in the big dataset come from the whole of PubMed. The training dataset released during 2013 is also available for reference reasons. The datasets are available both in text format and in a pre-processed format obtained with Apache Lucene framework[3]. More information on the pre-processed benchmark datasets and the format used for data exchange is available in Partalas et al. (2013). Table 2.1 provides information about the training datasets. The 2014-version

---

[1] www.json.org
[2] http://bioasq.lip6.fr/journals/
[3] http://lucene.apache.org/

of our training datasets include more than twelve million articles from MEDLINE published after 1949. Each article has 12.72 MeSH labels in average.

	version 2013	version 2014	version 2014(2)
Articles	10,876,004	12,628,968	4,458,300
Total labels	26,563	26,831	26,631
Labels per article	12.55	12,72	13,20
Size in GB	18	20,31	6,4

Table 2.1: Information on the training data for Task 2a.

The official test datasets of the BIOASQ challenge were released on Mondays starting from the 4th of February, 2014. The articles of the test sets come from 1,994 pre-selected journals. The list of the journals is available at http://bioasq.lip6.fr/journals/. The journals were selected by the BIOASQ team based on statistics about the annotation period of their articles. It was essential that the annotation period of an article was short, so that participants can receive feedback on their system's performance shortly after the submission of their results to improve it. Table 2.2 shows the size of each test, the number of the annotated articles when we were writing this deliverable and finally the average number of MeSH concepts that the annotators in MEDLINE gave in each of the annotated articles. The second test dataset for example consists of 4,721 articles. 3,716 out of 4,721 articles have been annotated with an average number of 13.13 MeSH concepts.

Batch	Articles	Annotated Articles	Labels per article
1	4,440	3,263	13.20
	4,721	3,716	13.13
	4,802	3,783	13.32
	3,579	2,341	13.02
	5,299	3,619	13.07
Subtotal	23,321	16,722	13.15
2	4,085	3,322	13.05
	3,496	2,752	12.28
	4,524	3,265	12.90
	5,407	3,848	13.23
	5,454	3,642	13.58
Subtotal	22,966	16,829	13.01
3	4,342	2,996	12.71
	8,840	5,783	13.37
	3,702	2,737	13.32
	4,726	3,225	13.90
	4,533	3,196	12.70
Subtotal	26,143	17,929	13.20
Total	72,430	51,480	13.12

Table 2.2: Statistics on the test datasets of Task 2a.

2.3 Participation in Task 2a

Table 2.3 shows the participation as it can be inferred from the submission of results for each test set. Participants of the BIOASQ challenge are allowed to participate in the challenge with a maximum of

five systems. This decision was made in order to help research teams participate with more than a system, since they usually test multiple implementations of an algorithm or several algorithms at the same time. For example, for the second test of the challenge 10 teams submitted results. However, there are 31 different result files, since most teams submit results with more than one system. In addition, the second test was downloaded by 18 unique teams since its release. A team usually downloads the data more than once, so we present the downloads w.r.t different teams because we believe it gives a better overview of the dynamics of the challenge. Otherwise, since downloading first the raw format and then their pre-processed description in Apache Lucene format counts as two downloads the results would be misleading. The enumeration of test sets in Table 2.3 follows the dates of release and horizontal lines are used to indicate the different batches.

Testset	Teams	Systems	# of downloads
1	9	25	33
2	10	31	18
3	10	28	20
4	8	26	18
5	10	36	18
6	10	35	17
7	10	40	14
8	11	44	13
9	12	45	14
10	11	43	14
11	10	36	13
12	11	39	10
13	10	39	13
14	11	39	12
15	13	39	12

Table 2.3: Statistics on the test datasets of Task 2a.

Table 2.4 shows the usernames and the affiliations of the teams that participated in Task 2a. In total, 18 different teams submitted results, at least once, for the challenge. Note that last year only 12 teams participated (increase=50%). For example, one of the teams that participated is "fribadas" and comes from University of Vigo. "fribadas" participated also in the first year of the challenge. Regarding the origin of the teams that participated in Task 2a, the majority of the teams come from Europe, whereas the majority of the teams during the first year of the challenge was from the U.S.. Figure 2.2 is a pie chart depicting the participation distribution over Europe, U.S. and Asia. Again, note that the participation from Europe has increased significantly and now the majority of the participating teams are Europe-based.

2.4   Datasets for Task 2b

The datasets for Task 2b of the BIOASQ challenge were created by a team of ten biomedical experts around Europe. More information on the team of the biomedical experts can be found in Polychronopoulos et al. (2013). The biomedical experts formulated questions depending on their field of specialization and analysed them by producing annotations, exact and ideal answers. The questions were created on-

Username	Institute	Region
ym	NCBI	North America (U.S)
bioasq	pierre curie	Europe (Germany)
chyc	Fudan University	Asia (China)
jgmork	U.S. National Library of Medicine	North America (U.S.)
tsoumakas	Aristotle University of Thessaloniki	Europe (Greece)
tnunes	Universidade de Aveiro	Europe (Portugal)
lvqian	Fudan University	Asia (China)
sana_amanat	UET	Asia (Pakistan)
antinomyra	Fudan University	Asia (China)
emilio_ortiz	Universidad Carlos III	Europe (Spain)
rapasric2	UC San Diego	North America (U.S.)
drameKhadim	ERIAS-ISPED	Europe (France)
wakeup06	Seoul National University	Asia (South Korea)
adamjo	Center For Spoken Language Understanding	North America (Canada)
fribadas	University of Vigo	Europe (Spain)
soblenes32	University of St Thomas	North America (U.S.)
ho2s	Holmes Semantic Solutions	Europe (France)
rapasric	University of California, San Diego	North America (U.S.)

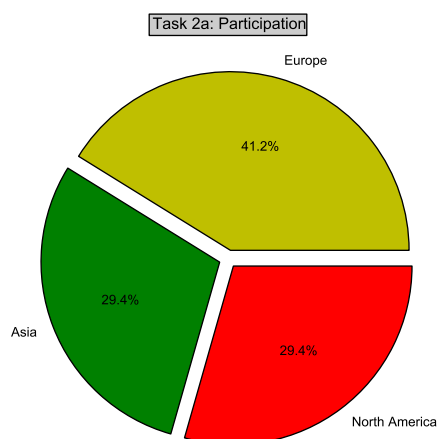Table 2.4: Affiliations of teams that participated in task 2a.

Figure 2.2: Distribution of the teams that participated in task 2a w.r.t their nationalities.

line using an "Annotation Tool" and an "Assessment Tool" the BIOASQ team developed for that reason. The production of the final version of the datasets that are used for the evaluation of the systems involves two steps:

- First, the experts formulated the questions they considered interesting and used the "Annotation Tool" to annotate them using documents, snippets, concepts and triples from designated ontologies. That was the first version of the golden dataset.

- After the first phase of the challenge where systems were requested to return documents, snippets, concepts and triples, the experts used the "Assessment Tool" to update the first version of the golden dataset. This final version is updated with annotations that the experts may had lost during the first round of annotation. During the assessment of the system responses of the first round, the experts also evaluated the ideal answers of the systems as described in Balikas et al. (2013) in order to be used for the official evaluation of the ideal answers.

More information on the process followed for the creation of the questions and on the "Annotation Tool" is available in Ngonga Ngomo et al. (2013). More information on the biomedical resources is available in Tsatsaronis et al. (2013). Table 2.5 provides information on the dataset the experts created for the task. The number of documents, snippets etc .refer to the average number of each type of annotation for each question of the corresponding batch.

Test batch	Size	# of documents	# of snippets	# of concepts	# of triples
training	310	14.28	18.70	7.11	9.00
1	100	7.89	9.64	6.50	24.48
2	100	11.69	14.71	4.24	204.85
3	100	8.66	10.80	5.09	354.44
4	100	12.25	14.58	5.18	58.70
5	100	11.07	13.18	5.07	271.68
total	810	11.83	14.92	5.93	116.30[4]

Table 2.5: Statistics on the training and test datasets of Task 2b. All the numbers for the documents, snippets, concepts and triples refer to averages.

Participants were given the golden dataset of the first year of the challenge as training dataset, which contains 310 questions. From Table 2.5 we can see that the experts produced 500 questions for the second year of the BIOASQ challenge. Those questions were semantically annotated with documents, snippets from those documents, concepts and triples. The average numbers of triples is based on the questions that have this type of annotations since experts could not locate them for every question. In contrast, documents, snippets and concepts were used to annotate every question. There are four types of questions released in the task 2b BIOASQ datasets: (i) yes/no questions, (ii) factoid questions, (iii) list questions and (iv) summary questions. Please, note that those numbers were produced using the first version of the golden dataset from the annotation process, not the final version that will be produced from the assessment process because it was not available when writing this document. Table 2.6 shows the distributions of those types of questions in each of the 5 released test batches.

	Training	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Total
Yes/No	85	32	28	36	32	24	152
Factoid	59	27	23	24	32	29	135
List	92	25	27	22	15	30	119
Summary	74	16	22	18	21	17	94
Total	*310*	*100*	*100*	*100*	*100*	*100*	*500*

Table 2.6: Distribution of questions w.r.t the four types of questions the BIOASQ team created.

2.5 Participation in Task 2b

Tables 2.7 and 2.8 show the participation in phase A and phase B of Task 2b respectively. Again, users could participate in the challenge with more than one systems to test and compare their methods without having to create different accounts in the platform. The participation in this task was lower than in Task 2a, since this task was more difficult and demanding. However, from the number of downloads of the first datasets we can infer that the task was interesting since many teams downloaded and inspected the data. In addition, there was a significant increase in the participation which shows that BIOASQ has started gaining momentum in the scientific community.

Testset	Users	Systems	# of Downloads
1	6	15	15
2	7	18	11
3	7	19	10
4	6	18	9
5	5	16	7

Table 2.7: Participation in phase A of Task 2b.

Testset	Users	Systems	# of Downloads
1	5	13	17
2	5	15	13
3	6	11	13
4	5	11	13
5	4	9	8

Table 2.8: Participation in phase B of Task 2b.

Table 2.9 and Table 2.10 show the affiliations of the users that participated in phase A and phase B of Task 2b respectively. From those teams, for example, "Wishart" participated in both phases of the Task 2b. Figures 2.3 and 2.4 are pie charts depicting the distribution of participants over Europe, U.S. and Asia for phase A and phase B of task 2b respectively. From the above-mentioned figures and tables we observe that participation in both phases has increased significantly. Note that in the first year of the challenge, we had no more than 3 teams submitting results in each phase which implies an increase of 130%. Furthermore, although still several teams are based in North America (37.5% and 33.3% for each

Username	Institute	Region
wishart	University of Alberta	North America (U.S.)
wakeup06	Seoul National University	Asia (South Korea)
rick70002	NCBI	North America (U.S.)
bioasq	upmc	Europe
mneves	Hasso-Plattner Institut	Europe (Germany)
UMass	University of Massachusetts Medical School	North America (U.S.)
wbc912	Fudan	Asia (China)
Kota	Toyota Technological Institute	Asia (Japan)

Table 2.9: Affiliations of teams that participated in phase A of task 2b.

phase) many Europe-based teams have entered the challenge. Again, note that in the first year of the challenge only the baseline systems of the challenge were from teams based in Europe and participants were from the U.S. and Canada. One of the most important outcomes is that during the second year of the challenge the participation in the BIOASQ challenges has increased and a lot of Europe-based teams have joined the challenge which was one of the major targets of the dissemination activities of the second year.
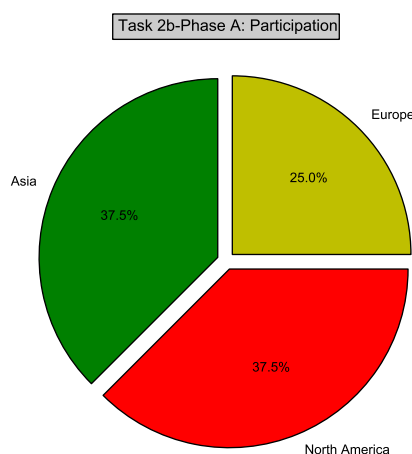

Figure 2.3: The distribution of the teams that participated in phase A of task 2b w.r.t their nationalities.

2.6 Providing support to the users

2.6.1 The Guidelines

In order to help participants understand the requirements and the process of the challenge the BIOASQ team improved the detailed guidelines that were first published on the first year of the BIOASQ challenge. The guidelines for participating in the challenge are available online, at `http://bioasq.lip6.fr`. No registration is required to access the guidelines and download samples of the datasets. The guidelines cover different topics concerning the participation in the challenge:

Username	Institute	Region
tsoumakas	Aristotle University of Thessaloniki	Europe (Greece)
wishart	University of Alberta	North America (U.S.)
wakeup06	Seoul National University	Asia (South Korea)
rick70002	NCBI	North America (U.S.)
bioasq	upmc	Europe
Kota	Toyota Technological Institute	Asia (Japan)

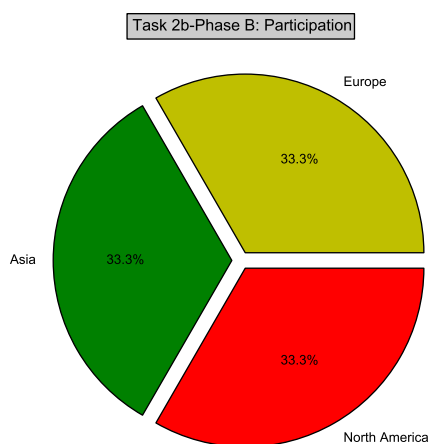Table 2.10: Affiliations of teams that participated in phase B of task 2b.

Figure 2.4: The distribution of the teams that participated in phase B of task 2b w.r.t their nationalities.

- Registration. The process is automated, after filling a form the participant receives a confirmation e-mail with an activation link. Clicking on the link registers the participant giving him access in the full functionality of the platform.

- The tasks of the challenge. There are separate guidelines with respect to the two tasks of the challenge. The provided information covers:

  - the schedule,

  - the data sources with statistics when there are available,

  - the evaluation process,

  - the provided tools,

  - the benchmark datasets that are released during the challenge and their format,

  - the format of the system answers and

  - code snippets written in Python, that implement the data exchange between a system and the BIOASQ platform using the provided web services.

2.6.2   Participation in the BIOASQ Discussions

BIOASQ Discussions is a forum integrated in the platform under http://bioasq.lip6.fr/forum/. It consists of a set of small forums where registered users can discuss the problems they are facing in the challenge, contact the organisers or ask for other participants' opinions. BIOASQ Discussions was organised under three topics of discussion:
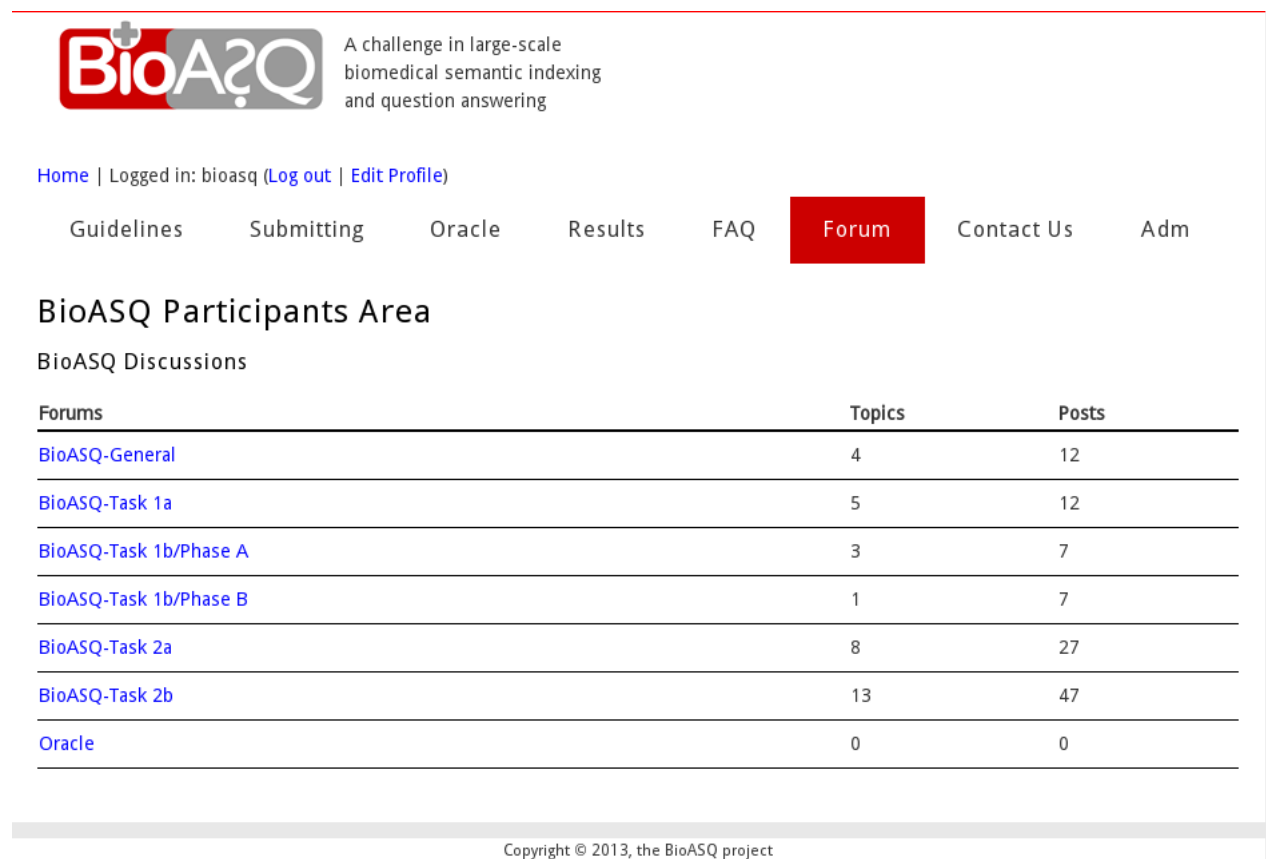
- BIOASQ-Task 2a,

- BIOASQ-Task 2b,

- BIOASQ-Oracles

Figure 2.5 shows the main page of the BIOASQ Discussions and the available forums. Apart from the discussions concerning the second year of the challenge, the discussion of the first year are also available for reference reasons. In the second year, 70 posts were made in the forums, mainly asking for clarifications about the schedule of the challenge, the format of the data, the schedule of the results announcements and the process followed when creating the data. Table 2.11 provides numeric details about those posts . The number of posts are given in parenthesis, while the numbers of topics the participants started are given outside the parenthesis. There was an increase in the discussions that took place in the forum revealing that more participants were dealing with the challenge and wanted to learn details and discuss about the process of the challenge. In every case, the BIOASQ team replied fast and after taking into account the feedback from the participants updated the guidelines or produced the necessary documents to clarify each aspect of the discussed topics.

2.6.3   The contact form

The contact form that was available in the platform was used, from users that had a problem while registering in the platform or from users that needed more help and instructions when manipulating the JSON files. The administrators of the challenge helped participants register successfully in the platform and access the available resources.

	Task A	Task B
1st year	5 (12)	4 (12)
2nd year	8 (27)	12 (43)

Table 2.11: Participation in the BioASQ forums. Outside the parenthesis we provide the number of new topics while in parenthesis we provide the number of posts.

Figure 2.5: The BIOASQ Discussion Area. There are four forums where participants can discuss about the challenge.

3

Lessons Learned

The goal of this chapter is to provide information about the problems the BIOASQ team faced in terms of results the participants provided. We also provide insight in the "lessons learned" during the second year and the infrastructure we will provide towards BIOASQ 3.

3.1   The format of the data

In order to help participants submit the correct format of data that would be the input in the automatic evaluation scripts we did the following:

- We released detailed guidelines describing the format of the answers the systems should provide. We insisted on covering all the possible cases that could cause questions to the participants.

- We released examples of result files for each of the cases above.

- We integrated the functionality to the platform that would catch the most common errors and would output informative messages so that participants could correct their data without extra help.

- We provided scripts in Python[1] to help participants with the JSON manipulation and encourage them to use the web services we developed.

However, especially in task 2b, a lot of extra effort had to be spent with many participants who had trouble formatting correctly their answers. Also, during the evaluation of the data, we went through the result files and wrote cleaning scripts to correct each of the submitted files. The fact that participants could not provide the correct format of the data would be a problem in the usage of the oracles and could cause problems in a completely automated system. Thus, we intend to further improve our guidelines and research the possibility of providing a JSON format validator to the participants of the challenge for each task.

---

[1] https://www.python.org/

Bibliography

I. Androutsopoulos, P. Malakasiotis, G. Tsatsaronis, M. Zschunke, and G. Balikas. Resources and Services for Task 1B. 2013.

G. Balikas, I. Partalas, A. Kosmopoulos, S. Petridis, P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, N. Baskiotis, E. Gaussier, T. Artieres, and P. Gallinari. Evaluation Framework Specifications. Technical Report D4.1, BioASQ Deliverable, 2013.

G. Balikas, I. Partalas, N. Baskiotis, E. Gaussier, T. Artieres, and P. Gallinari. Evaluation infrastructure software for the challenges 2nd version. Technical Report D4.7, BioASQ Deliverable, 2014.

A.-C. Ngonga Ngomo, N. Heino, R. Speck, T. Ermilov, and G. Tsatsaronis. Annotation Tool. Technical Report D3.3, BioASQ Deliverable, 2013.

A.-C. Ngonga Ngomo, N. Heino, and R. Speck. A tool suite for creating question answering benchmarks. 2014.

I. Partalas, G. Balikas, N. Baskiotis, E. Gaussier, T. Artieres, and P. Gallinari. Pre-processed benchmark set 1. Technical Report D 4.2, 2013.

D. Polychronopoulos, Y. Almirantis, A. Krithara, and G. Paliouras. Expert Team. 2013.

G. Tsatsaronis, M. Zschunke, M. R. Alvers, and C. Plonka. Report on existing and selected datasets. Technical Report D3.2, BioASQ Deliverable, 2013.