Intelligent Information Management
Targeted Competition Framework
ICT-2011.4.4(d)

Project **FP7-318652 / BioASQ**

Deliverable **Technical report**

Distribution **Public**

http://www.bioasq.org

Resources and Services for Task 11b

Anastasios Nentidis, Eirini Vandorou, Prodromos Malakasiotis, Ion Androutsopoulos, Matthias Zschunke, George Tsatsaronis and George Balikas

Status: Final (Version 1.0)

December 2021

1

Introduction

BIOASQ Task 11b will take place in two phases:

Phase A (retrieve relevant articles, snippets): In this phase, participants will be provided with biomedical questions written in English and will be asked to retrieve relevant articles and text snippets from designated article repositories. The system responses of Phase A will be automatically compared against gold responses constructed by the BIOASQ team of biomedical experts; consult Malakasiotis et al. (2013) for further information on the gold responses.

Phase B (find and report 'exact' and 'ideal' answers): In this phase, the questions and gold responses of Phase A (correct articles and snippets) will be provided as input. The participants will be asked to report 'exact answers' (e.g., named entities in the case of factoid questions) and 'ideal answers' (paragraph-sized summaries). The 'exact' and 'ideal' answers of the systems will be automatically compared against gold 'exact' and 'ideal' answers constructed by the BIOASQ team of biomedical experts; again, consult Malakasiotis et al. (2013) for further information on the gold 'exact' and 'ideal' answers. All the 'ideal' answers of the systems will also be manually evaluated by the biomedical experts.

The test dataset of Task 11b will be released in four batches, each containing approximately 90 questions. For each batch, first only the questions of the batch will be released, and the participants will have to submit their answers for Phase A (articles, snippets) within 24 hours; then gold articles and snippets for the questions of the batch will also be provided, and the participants will again have 24 hours to submit their answers for Phase B ('exact' and 'ideal' answers). Consult the online guidelines of Task 11b in the participants area of BIOASQ for further information[1]. The evaluation measures that will be used in Phase A and Phase B are discussed in Chapter 4 of Balikas et al. (2013); a document describing the evaluation measures of Task B is also available online in the participants area.

Please note:

- The gold responses of Phase A that will be provided to the participants of Phase B will contain the relevant articles and snippets that the biomedical experts who prepared the questions managed to identify in the designated resources.

[1] http://participants-area.bioasq.org/

- There will be enough information in the provided gold responses of Phase A to find the ‘exact’ answers and to formulate ‘ideal’ answers. However, there may be (and most probably, will be) additional correct (relevant) articles and snippets, which will not be included in the provided gold responses of Phase A.

- When producing the ‘exact’ and ‘ideal answers in Phase B, the participants are allowed to use both the provided gold responses of Phase A and any other resource (e.g., the articles and snippets they retrieved in Phase A, in addition to the gold responses of Phase A).

- Before announcing the official evaluation results of Phase B, biomedical experts will examine the lists of articles and snippets returned by the participating systems, as well as the ‘exact’ and ‘ideal’ answers of the participating systems, in order to enhance the gold lists of concepts, articles, snippets, triples, and the gold ‘exact’ and ‘ideal’ answers that will be used as references for evaluation purposes, as discussed in Chapter 4 of Balikas et al. (2013).

The rest of this document is organized as follows. Chapter 2 lists the designated resources of Phase A and explains how to access and search them. Chapter 3 explains how to download batches of the test sets of Task 11b (both for Phase A and Phase B) and submit results. Chapter 4 provides important licensing information. The remainder of this document assumes that the reader has already studied the online guidelines of Task 11b.

2

How to access and search the resources of Task 11b Phase A

2.1 Designated resources of Phase A

In Phase A of Task 11b relevant articles are to be retrieved from PUBMED, using the Annual Baseline Repository for 2022[1]. Their unique identifiers are their URLs in PUBMED. For example, "`http://www.ncbi.nlm.nih.gov/pubmed/23687640`".[2]

The relevant snippets will have to be parts of relevant articles.

For a more detailed description of the resources used in BIOASQ, consult Tsatsaronis et al. (2013).

2.2 Accessing and searching the designated resources of Phase A

The indices of the above resources are accessible through Javascript Object Notation (JSON) based web services.[3] More specifically, the user opens a session–based URL to send HTTP-POST requests and receive corresponding responses, both in JSON format. Table 2.1 lists the URL of the web service, and Table 2.2 lists the technical specifications required to use them.

Resource	URL
PUBMED	`http://bioasq.org:8000/pubmed`

Table 2.1: Access points of the web services to search the indices of the resources of Phase A.

2.2.1 Searching for articles

Figure 2.1 describes the JSON objects that are used to search for articles. As with the service for searching concepts, the request consists of three basic elements:

[1] `https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/`.

[2] `http://www.ncbi.nlm.nih.gov/pubmed`.

[3] JSON is an easy to use lightweight data interchange format. `http://json.org/` for more information.

Exchange format	JSON
Character encoding	UTF-8
Communication method	HTTP-POST with parameter “json”.
Content type	“application/x-www-form-urlencoded” or “multipart/form-data” with “application/json” as part type.
Time out	10 minutes. On expire an exception is returned. The user should then open a new session.
Return value	JSON object with key name “result” upon success and “exception” on failure. “result” is followed by a JSON object containing the returned data. “exception” is followed by a string message describing the cause of the exception and a time stamp. After an exception a new session should be opened for further requests.

Table 2.2: Technical specifications of the web services to search the indices of the resources of Phase A.

keywords: The query to perform the search. The full potential of PUBMED queries is supported.[4] For instance, the following is a valid query:

- “Rheumatoid Arthritis” [Title] AND (“gender” [Title] OR “male” [Title] OR “female” [Title])’

page, articlesPerPage: These fields have the same role as when searching for concepts. For instance:

- *page*=0, *articlesPerPage*=10 will retrieve the first 10 articles,
- *page*=1, *articlesPerPage*=10 will retrieve the next 10 articles.

The retrieved JSON object has the following structure:

result: A JSON object containing the search results with the following fields:

articlesPerPage: The same value as in the JSON object used to post a request.

documents: A list of JSON objects representing the retrieved articles, each object with the following structure:

documentAbstract: The abstract of the article.

fulltextAvailable: This value states whether the full text of the article is available.

journal: The full title of the journal the article was published in.

meshAnnotations: This field is always *null*. This used to contain a list of GOPUBMED based MESH annotations of the article with the following structure:

termLabel: The MESH term label.

uri: The URI of the MESH term with the following fields:

id: The MESH id of the term.

namespace: This is always set to be “MeSH”.

meshHeading: A list of MESH headings of the article.

pmid: The PUBMED id of the article.

sections: This would normally contain the full text (if any) of the article, but due to licensing issues the full text cannot be distributed. Thus, this field is always *null*.s

title: The title of the article.

[4] http://www.ncbi.nlm.nih.gov/books/NBK3827/#_pubmedhelp_Search_Field_Descrip_

year: The publication year of the article.

fullPubmedQuery: The PubMed queries are automatically subjected to the following restrictions:
`ownernlm, not hascommenton, not hasretractionof, not haserratumfor, not haspartialretractionof, not hasrepublishedin, not hasupdatein, not Editorial[PT], not Comment[PT], not Letter[PT], not News[PT], not Review[PT], not pubmednotmedline[sb], not indatareview[sb].`
fullPubmedQuery contains the expanded query to comply with these restrictions.

keywords: The same as *keywords* in the request object.

maxDate: The latest allowed publication date. All the articles retrieved were published on that date or earlier.

page: The same as *page* in the request object.

size: The number of articles that satisfy the query.

timeMS: The time (in msec) it took to process the request.

Important note: Even when the full text of an article is available, it is not included in the returned results of the web service, due to licensing issues.

2.3 Example runs of the search services

Below we give some short examples of how the search web services of the previous sections should be used. Example source code is also provided in the directory "services_examples" that accompanies this document. The examples were implemented in Java, using JSON–simple 1.1 to deal with JSON objects, and Apache Commons HTTP Client 3.0.1 for the HTTP-POST requests.

Let us assume that we wish to perform a search in PubMed. The following steps are to be followed:

1. Create a session:

   Call: `http://bioasq.org:8000/pubmed`

   Response: `http://bioasq.org:8000/pubmed/2?-4fd1740f%3A1531863bc8b%3A-7ff2`

2. Perform a search:

   (a) Call: `http://bioasq.org:8000/pubmed/2?-4fd1740f%3A1531863bc8b%3A-7ff2`

   (b) HTTP-POST Parameter: json={"findEntities": ["nitric oxide synthase", 0, 10]}

   (c) Response: A JSON object with the results.

The session URL obtain during step 1 can be used to perform more than one searches. If, however, an "exception" is returned, the user should obtain a new session URL, by repeating step 1.
The corresponding code examples and JSON responses can be found in:

- "examples/PubMedSearchServiceCall.java", and
- "examples/PubMedSearchServiceCall.example.response.json"

Request

```
{
  "api": {
    "findPubMedCitations": [
      "keywords": String,
      "page": Number,
      "articlesPerPage": Number
    ]
  }
}
```

Response

```
{
  "result":{
  "articlesPerPage": Number,
  "documents": [
    {
      "documentAbstract": String,
      "fulltextAvailable": Boolean,
      "journal": String,
      "meshAnnotations": [
        {
          "termLabel": String,
          "uri":
           {
             "id": String,
             "namespace": "MeSH:"
           }
        },
        ...
      ],
      "meshHeading": [
        String,
        String,
        ...
      ],
      "pmid": String,
      "sections": null,
      "title": String,
      "year": String,
    },
    ...
  ],
  "fullPubmedQuery": String,
  "keywords": String,
  "maxDate": "2013/03/14",
  "page": Number,
  "size": Number,
  "timeMS": Number
  }
}
```

Figure 2.1: JSON objects used to search for articles.

3

How to download test sets and submit results for Task 11b

This chapter describes how the participants can download batches of the test dataset of Task 11b (for both Phases A and B) and submit their results. Examples are also provided in the accompanying directory "submission_examples".

3.1 Downloading test batches

Test batches for both Phase A and B of Task 11b can be downloaded using either a web interface or an API. In particular:

Web interface: In the menus 'Submitting/Task 11b-Phase A' and 'Submitting/Task 11b-Phase B' of the BIOASQ participants area, the participants can find lists with the released test batches of phase A and phase B, respectively [1]. The test batches can be downloaded by clicking on their links.

API: Participants can also download the test batches by sending an HTTP-POST request to:

- http://participants-area.bioasq.org/Tasks/11b/phaseA/api/*number*/ for Phase A, and
- http://participants-area.bioasq.org/Tasks/11b/phaseB/api/*number*/ for Phase B.

replacing *num* by the number of the test set they wish to download. To obtain the test batch, they should post a JSON string with their username and password, i.e.,

```
{"username":"your_username",
 "password":"your_password"}
```

[1] See http://participants-area.bioasq.org/

For example, the following command would download the first test batch of Phase A to the standard output of the console of the registered participant “bill”: >curl -v -H ``Accept: application/json'' -H ``Content-type: application/json'' -X POST -d '{``username'':``bill'', ``password'':``XXXX''}' http://participants-area.bioasq.org /Tasks/11b/phaseA/api/1/

3.2 Submitting results

Participants can submit their results using either the web interface or an API. In particular:

Web interface:

Phase A: In the menu ‘Submitting/Task 11b-Phase A’ of the BIOASQ participants area, there will be a form for submitting results. The form will be available if there is an active test batch. Note that it is not possible for more than one test batches to be active at the same time. The form will have two fields:

- “Select a file”, where the participant selects the file from his/her computer that contains the test results in the JSON format discussed below, and
- “System name”, where the participant selects from a drop-down menu one of his/her systems. Users can see only their own systems in the drop-down menu.

The JSON format of the file with the results should be the same as the JSON format of the development data; consult the on-line guidelines of Task 11b in the participants area. More precisely, for each question, only the ‘id’ field (question id) should be present, along with any (and ideally both) of the fields: ‘documents’, ‘snippets’; any other fields will be ignored. An example follows:

```
{"questions":[
  {"id":"5118dd1305c10fae750000010",
   "documents": [
   "http://www.ncbi.nlm.nih.gov/pubmed/12723987",
   ...],
   "snippets":[
   {"document": "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
    "text": "The expression and clinical course of RA are influenced
    by gender. In developed countries the prevalence of RA is 0,5 to
    1.0%, with a male:female ratio of 1:3.",
    "offsetInBeginSection": 559,
    "offsetInEndSection": 718,
    "beginSection": "sections.0"
    "endSection": "sections.0"},
    ...],
   },
   ...]}
```

Phase B: The same functionality as in Phase A will be available. The sumbission form will be in ‘Submitting/Task 11b-Phase B’. Again, the JSON format of the results file to be uploaded should be the same as the JSON format of the development data; consult the on-line guidelines of Task 11b. More precisely, for each question, only the ‘id’ field (question id) should be present, along with any (ideally both) of the fields: ‘exact_answer’ and ‘ideal_answer’; any other fields will be ignored. For example:

```
{"questions":[
  {"id":"5118dd1305c10fae750000010",
   "ideal_answer": "Disease patterns in RA vary between the sexes;
   the condition is more commonly seen in women, who exhibit a more
   aggressive disease and a poorer long-term outcome.",
   "exact_answer": ["Women"]},
  ...]}
```

API: Participants can submit their results by sending an HTTP-POST request to:

- `http://participants-area.bioasq.org/Tasks/b/phaseA/submit/`*number*`/` for Phase A, and
- `http://participants-area.bioasq.org/Tasks/b/phaseB/submit/`*number*`/` for Phase B.

replacing *number* by the number of the active test batch that the user wants to submit results for. In the HTTP-POST request, users have to post a JSON string with their username, their password, their system name, and their results, i.e.,

```
{"username":"your_username",
 "password":"your_password",
 "system":"your_system",
 "questions":[...]}
```

where the ‘questions’ field is as when using the web interface.

Assuming that user “bill” has saved the results of ‘system1’ for the first (and active) test batch of Phase A in the file “results.json’, he would be able to submit them using the following command:

```
>curl -v -H ``Accept:  application/json'' -H ``Content-type:
application/json'' -X POST --data-binary @results.json
bioasq.lip6.fr/Tasks/b/phaseA/submit/1/
```

4

Licensing issues

The resources and services provided by BIOASQ should be used for the purposes of BIOASQ only; for any other purpose, please contact the BIOASQ organizers. Furthermore, the participants should ensure that they comply with the licenses of the third-party resources of Table 4.1, where the URLs of both the resources and their licenses are listed.

Resource	**URL**
PUBMED abstracts	http://www.ncbi.nlm.nih.gov/pubmed/ http://www.nlm.nih.gov/databases/license/license.pdf
PUBMEDCENTRAL full text documents	http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/ http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ http://www.ncbi.nlm.nih.gov/pmc/about/copyright/

Table 4.1: Third party resources used by BIOASQ and their licenses.

Bibliography

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.

P. Malakasiotis, I. Androutsopoulos, Y. Almirantis, D. Polychronopoulos, and I. Pavlopoulos. Tutorials and Guidelines. Technical Report D3.4, BioASQ Deliverable, 2013.

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