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

Project **FP7-318652 / BioASQ**

Deliverable **D3.8 and D4.6**

Distribution **Restricted**

http://www.bioasq.org

Pre-processed benchmark set 2

Ioannis Partalas, Georgios Balikas, Nicolas Baskiotis, Dimitrios Polychronopoulos, Yannis Almirantis, Eric Gaussier, Thierry Artieres, Patrick Gallinari

Status: Final (Version 1.0)

March 31, 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	Other
Distribution level	Restricted
Deliverable Number	D3.8 and D4.6
Deliverable title	Pre-processed benchmark set 2
Contractual date of delivery	M18 (March 2014)
Actual date of delivery	March 31, 2014
Relevant Task(s)	WP3/Task 3.5, WP4/Task 4.1
Partner Responsible	UPMC
Other contributors	UJF, NCSR “D”
Number of pages	9
Author(s)	Ioannis Partalas, Georgios Balikas, Nicolas Baskiotis, Dimitrios Polychronopoulos, Yannis Almirantis, Eric Gaussier, Thierry Artieres, Patrick Gallinari
Internal Reviewers	George Tsatsaronis
Status & version	Final
Keywords	data, benchmark set

Contents

1 Executive Summary **1**

2 Introduction **2**
2.1 The Benchmark Data of the BIOASQ Challenge 2
2.2 Format of the Data 2

3 Data Description **5**
3.1 Task 2a 5
3.2 Task 2b 6

List of Figures

2.1 An extract from the training data of Task2a. 3
2.2 A sample of the training data of Task 2b. 4

3.1 The format of the training data of Task1b. 8

List of Tables

3.1 Description of the properties of the data for Task1a. 6
3.2 Basic statistics about the training data for Task 1a and Task 2a. 6
3.3 Number of articles for each test dataset in each batch. In parentheses the articles that have been annotated by the curators. 7
3.4 Basic statistics of the training and test data for Task 2b provided during the evaluation procedure. 7

1

Executive Summary

This report describes briefly the benchmark set which is released for the second edition of the challenge of the BIOASQ project. More specifically, this document details the type of data as well as its structure for both tasks of the second BIOASQ challenge. Note that this document accompanies deliverables D4.6 and D3.8, which are not report deliverables.

In the second edition of the challenge no major changes have been done in the level of the representation of the data. As in the first edition, the data are provided in a structured human readable format using the JavaScript Object Notation (JSON) format for both tasks of the challenge.

For the task of the large-scale biomedical semantic indexing (Task 2a), the data were aligned to the new version of the MESH taxonomy (2014). Also, a sample has been created from the full training set (12.5M of documents), containing only the documents that come from the selected journals used to draw the test sets during the evaluation procedure. In the question-answering task a development set of 310 questions has been released and during the evaluation procedure a set of 500 question will be released in five consecutive batches.

2

Introduction

This chapter briefly introduces the two tasks of the BIOASQ challenge and describes the requirements of the format of the data in the BIOASQ challenge. Then, it presents the selected format and describes its key properties.

2.1 The Benchmark Data of the BIOASQ Challenge

The second edition of the BIOASQ challenge consists of two tasks:

- **Task 2a**: Large-scale on-line biomedical semantic indexing.
- **Task 2b**: Biomedical Semantic Question-Answering.

In Task 2a the data that are available to the participants consist of biomedical articles indexed in PUBMED. Specifically, for each article in the training data, BIOASQ provides its abstract as it appears in PUBMED and the assigned labels to it. In the testing phase of the challenge the data contain only the abstract of the corresponding article without any further information. The articles are provided in their raw format (plain text) as well as in a pre-processed one (in a vectorized format). Figure 2.1 presents an example of two articles extracted from the BIOASQ benchmark training data.

Task 2b BIOASQ takes place in two phases. In the first phase, BIOASQ distributes a set of questions and the participants should respond with concepts, articles, snippets and triples. In the second phase BIOASQ distributes questions along with concepts, articles, snippets and triples and the participants respond with exact answers or summaries. The data for both phases of Task 2b are provided in a raw text format. An example of the representation of the data in Task 2b is presented in Figure 2.2 (one question from the development dataset).

2.2 Format of the Data

For the second version of the challenge no changes have been made in the format of the data. More specifically, as in the first year the JSON format is used for the representation of the data. To recall,

```
{
    "abstractText":"From the above it is seen that the [...]
    scientific guidance of which lies wholly
     in the hands of scientists.",
    "journal":"Science (New York, N.Y.)",
    "meshMajor":["Biomedical Research"],
    "pmid":"17772322",
    "title":"New Horizons in Medical Research.",
    "year":"1946"
},
{
    "abstractText":"1. T antigens of group A hemolytic
     streptococci have been [...] T antigen in the intact
      streptococcus from which it was derived.",
    "journal":"The Journal of experimental medicine",
    "meshMajor":["Antibodies","Antigens",
    "Immunity","Streptococcal Infections","Streptococcus"],
    "pmid":"19871581",
    "title":"THE PROPERTIES OF T ANTIGENS EXTRACTED
    FROM GROUP A HEMOLYTIC STREPTOCOCCI.",
    "year":"1946"
}
```


Figure 2.1: An extract from the training data of Task2a.

JSON[1] is a popular text-based format for data interchange. It is supported by the majority of the programming languages and cooperates well with web services[2].

[1] http://www.json.com/

[2] Please consult also deliverable (Partalas et al., 2013)

```
{ "body": "What is the action of molindone?",
 "documents": [
     "http://www.ncbi.nlm.nih.gov/pubmed/9577836",
     "http://www.ncbi.nlm.nih.gov/pubmed/9353417",
     "http://www.ncbi.nlm.nih.gov/pubmed/7656507",
     "http://www.ncbi.nlm.nih.gov/pubmed/7965768",
     "http://www.ncbi.nlm.nih.gov/pubmed/2895008",
     "http://www.ncbi.nlm.nih.gov/pubmed/6817377",
 ],
 "triples": [
     {
         "p": "http://www.w3.org/2004/02/skos/core#notation",
         "s": "http://linkedlifedata.com/resource/umls/label/A17796303",
         "o": "T43.505"
     },
     {
         "p": "http://www.w3.org/2008/05/skos-xl#literalForm",
         "s": "http://linkedlifedata.com/resource/umls/label/A17796303",
         "o": "Adverse effect of unspecified antipsychotics and neuroleptics"
     },
     {
         "p": "http://www.w3.org/2008/05/skos-xl#prefLabel",
         "s": "http://linkedlifedata.com/resource/umls/id/C2878523",
         "o": "http://linkedlifedata.com/resource/umls/label/A17809238"
     }
 ],
 "concepts": [
     "http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0042493",
     "http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0097332",
     "http://www.biosemantics.org/jochem#4249624",
     "http://www.biosemantics.org/jochem#4066301"
 ],
 "type": "summary",
 "id": "52bf1f5f03868f1b06000017",
 "snippets": [
     {
         "offsetInBeginSection": 404,
         "offsetInEndSection": 603,
         "text": " The antagonist molindone exhibits selectivity for  cortical
          serotonin-stimulated cyclase versus
           dopamine-stimulated cyclase and may prove useful for further elucidating
            the sites of lisuride action. ",
         "beginSection": "abstract",
         "document": "http://www.ncbi.nlm.nih.gov/pubmed/6249092",
         "endSection": "abstract"
     },
     {
         "offsetInBeginSection": 245,
         "offsetInEndSection": 928,
         "text": "Molindone in low intravenous doses (0.4-0.8 mg/kg) was found to
         reverse d-amphetamine and apomorphine induced depression of DA neurons and
         to block apomorphine induced depression of these cells. Molindone
         was also found to increase dopamine synthesis and
          dihydroxyphenylactic acid levels in the striatum and olfacotry tubercles.
          In all of these respects molindone behaves identically to most classical
          neuroleptics. However, unlike most antipsychotic drugs previously tested,
          molindone failed to increase the baseline firing rate of DA cells
          and blocked haloperidol induced increases in DA neuron activity. In this
          regard molindone most closely resembles thioridazine and clozapine. ",
         "beginSection": "abstract",
         "document": "http://www.ncbi.nlm.nih.gov/pubmed/1224004",
         "endSection": "abstract"
     }
```


Figure 2.2: A sample of the training data of Task 2b.

3

Data Description

This chapter details the data provided in the two tasks of the BIOASQ challenge and provides some descriptive statistics.

3.1 Task 2a

In the first task of the BIOASQ challenge, which concerns the classification of unlabelled articles of PUBMED, the data are provided in raw format as well as in a pre-processed format, in order to facilitate the participation of teams that would like to focus on the classification part of the task rather than on the pre-processing of the data.

The raw training data follow the JSON format where each article of PUBMED contains the fields presented in Table 3.1. The **pmid** field is a unique identifier that is used by PUBMED, while the **meshMajor** labels come from the Medical Subject Headings hierarchy which is the National Library of Medicine's thesaurus[1]. Figure 2.1 presents an example of the training data.

For the second year of the challenge, all the data were aligned to the version of 2014 of the MESH taxonomy. Additionally, a sample from the full training data is provided to the participants which contains only the articles that come from the journals that are used to draw the test sets during the evaluation.

Table 3.2 some basic statistics about the training data provided in the first edition of the challenge (first column) as well as for the second edition.

In each batch of Task 2a, 5 test datasets are given to the participants consecutively (one each week). Table 3.3 presents the number of articles of each test dataset in each batch of the evaluation procedure. For the third batch, the dataset is not available yet. The numbers in parentheses are those articles of the corresponding test dataset that have so far been annotated by the curators.

As mentioned above, the data are also available in a pre-processed format obtained with the Apache Lucene framework[2]. Lucene is an open-source library[3] dedicated to text search. It contains packages for scalable indexing as well as state-of-the-art search algorithms. The library is written in Java, so as to serve for cross-platform usage. In our case, the Lucene Core has been used for indexing the training data applying standard pre-processing procedures (stemming, stopword removal etc.).

[1] http://www.nlm.nih.gov/mesh/meshhome.html
[2] http://lucene.apache.org/
[3] Under the Apache Licence: http://www.apache.org/licenses/LICENSE-2.0.html

Field name	Description	JSON type
pmid	PubMed identifier	string
year	Year of publication	string
journal	Journal of publication	string
abstractText	Full text of the abstract	string
title	Title of the article	string
meshMajor	MESH labels of the article	array of strings

Table 3.1: Description of the properties of the data for Task1a.

	Task 1a	Task 2a	Task 2a (sample)
Articles	10,876,004	12,628,968	4,458,300
Unique labels	26,563	26831	26,631
Labels per article	12.55	12.72	13.20
Size in GB	18	20.31	6.4

Table 3.2: Basic statistics about the training data for Task 1a and Task 2a.

The pre-processed datasets are provided for both the full and the sample training data, resulting to two files of 6.2Gb and 1.9Gb respectively.

3.2 Task 2b

In Task 2b, the benchmark datasets contain development and test questions, in English, along with golden standard (reference) answers. The benchmark datasets have been constructed by a team of biomedical experts from around Europe (Malakasiotis et al., 2013; Polychronopoulos et al., 2012). No changes have been performed in the format of the data with respect to the first edition of the task. More detailed information can be found in the deliverable (Partalas et al., 2013), where the data of the first edition are described. Figure 3.1 presents the format of the data for Task 2b. An example, following this format is shown in Figure 2.2.

Table 3.4 presents descriptive statistics of the training and the test data. The development data for the second edition contains 310 questions while the test data 500 questions split in 5 batches.

Week	Batch 1	Batch 2	Batch 3
1	4,440 (2,784)	4,085 (2,458)	-
2	4,271 (3,083)	3,496 (1,922)	-
3	4,802 (3,210)	4,524 (1,075)	-
4	3,579 (2,026)	5,407 (574)	-
5	5,299 (2,953)	5,454 (0)	-
Total	22,391 (14,056)	22,966 (6029)	-

Table 3.3: Number of articles for each test dataset in each batch. In parentheses the articles that have been annotated by the curators.

	Training data	Test set 1	Test set 2
Questions	310	100	100
Yes/No	85	32	28
Factoid	59	27	27
List	92	25	23
Summary	74	16	22
Avg #concepts	7.1	6.5	4.2
Avg #documents	14.2	11.4	14.8
Avg #snippets	18.7	17.1	14.7
Avg #triples	9.0	102.0	125.3

Table 3.4: Basic statistics of the training and test data for Task 2b provided during the evaluation procedure.

```
{ "questions": [
        {
                "id": "the ID",
                "body": "the question?",(for the test )
                "type": "the type of the question",
                "concepts": [
                        "c1",
                        "c2",
                        ...
                        "cn"
                ],
                "documents": [
                        "d1",
                        "d2",
                        ...
                        "dn"
                ],
                "exact_answer": [
                        "ea1",
                        "ea2",
                        ...
                ],
                "ideal_answer": "the ideal answer",
                "snippets":[
                        {
                                "document": "dk",
                                "beginSection": "sections. #b",
                                "endSection": "sections.#e",
                                "offsetInBeginSection": number,
                                "offsetInEndSection": number,
                                "text": "the snippet"
                        }
                ],
                "triples": [
                        {
                                "o": "object",
                                "p": "predicate",
                                "s": "subject"
                        },
                        ...
                ]
        },
        ...
]
}
```

Figure 3.1: The format of the training data of Task1b.

Bibliography

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

I. Partalas, G. Balikas, N. Baskiotis, D. Polychronopoulos, Y. Almirantis, E. Gaussier, T. Artieres, and P. Gallinari. Pre-processed benchmark set 1. Technical Report D3.5 and 4.2, BioASQ Deliverable, 2013.

D. Polychronopoulos, Y. Almirantis, A. Krithara, and G. Paliouras. Expert Team. Technical Report D3.1, BioASQ Deliverable, 2012.