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

Project **FP7-318652 / BioASQ**

Deliverable **D3.5 and D4.2**

Distribution **Restricted**

http://www.bioasq.org

Pre-processed benchmark set 1

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

Status: Final (Version 1.0)

June 30, 2013

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	D3.5 and D4.2
Deliverable title	Pre-processed benchmark set 1
Contractual date of delivery	M9 (June 2013)
Actual date of delivery	June 30, 2013
Relevant Task(s)	WP3/Task 3.5, WP4/Task 4.1
Partner Responsible	UPMC
Other contributors	UJF, NCSR “D”
Number of pages	11
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

List of Figures

List of Tables

1

Executive Summary

This report describes briefly the benchmark set which is released for the first 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 first BIOASQ challenge. Note that this document accompanies deliverables D4.2 and D3.5, which are not report deliverables.

The remainder of this document is structured as follows:

- Chapter 2 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.
- Chapter 3 details the data provided in the two tasks of the BIOASQ challenge and provides some descriptive statistics.

2

Introduction

2.1 The Benchmark Data of the BIOASQ Challenge

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

- **Task 1a**: Large-scale on-line biomedical semantic indexing.
- **Task 1b**: Introductory biomedical semantic Question-Answering.

In Task 1a the data that are available to the participants consist of biomedical articles published 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 1b 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 1b are provided in a raw text format. An example of the representation of the data in Task 1b is presented in Figure 2.2 (one question from the development dataset).

2.2 Format of the Data

The format of the data is crucial as it affects the implementation decisions for the evaluation platform of the challenge. Additionally, the participants interact directly with the data which means that they should be provided with comprehensible and easy to manipulate data. The choice of the data format was based on the following criteria:

- To be in a simple and structured format.
- To follow an open standard format[1].

[1] http://en.wikipedia.org/wiki/Open_standard

```
{
    "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 Task1a.

- To be human readable.
- To be programming language-independent.

A popular format that fulfils the above criteria is the JavaScript Object Notation[2] (JSON) which is a text-based format for data interchange. It is supported by the majority of the programming languages and cooperates well with web services used in the BIOASQ evaluation platform. The JSON format is structured and comprehensible by the humans. Finally, JSON is one of the most widely used data formats in the Web (e.g. the API of twitter uses JSON format for data exchange).

The JSON format can support the following data types:

- Number (double precision floats)
- String (double quoted)
- Boolean (true, false)
- Array (the values maybe of different types)
- Object, which defines an unordered collection of *key:value* pairs
- Null, which corresponds to an empty entry

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

```
{ "questions": [
{
"id": "5118dd1305c10fae75000001",
"body": "Is Rheumatoid Arthritis more common in men or women?",
"type": "factoid",
"concepts": [
   "http://www.disease-ontology.org/api/metadata/DOID:7148",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D001171",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D012217",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D013167",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D015535"
],
"documents": [
   "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
   "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
   "http://www.ncbi.nlm.nih.gov/pubmed/21340496",
   "http://www.ncbi.nlm.nih.gov/pubmed/20889597",
   "http://www.ncbi.nlm.nih.gov/pubmed/20810033",
   "http://www.ncbi.nlm.nih.gov/pubmed/19158113",
   "http://www.ncbi.nlm.nih.gov/pubmed/18759162",
   "http://www.ncbi.nlm.nih.gov/pubmed/17965425",
   "http://www.ncbi.nlm.nih.gov/pubmed/16418123",
   "http://www.ncbi.nlm.nih.gov/pubmed/15083883",
   "http://www.ncbi.nlm.nih.gov/pubmed/12723987",
   "http://www.ncbi.nlm.nih.gov/pubmed/1563036"
],
"exact_answer": [
        "Women"
],
"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.",
"snippets":[
    {
     "beginSection": "sections.0",
     "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
     "endSection": "sections.0",
     "offsetInBeginSection": 591,
     "offsetInEndSection": 678,
     "text": "Our results show a high prevalence of RA in LAC women with a ratio of 5.2 women
     per man"
    },
    {
     "beginSection": "sections.0",
     "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
     "endSection": "sections.0",
     "offsetInBeginSection": 1140,
     "offsetInEndSection": 1394,
     "text": "RA in LAC women is not only more common but presents with some clinical
     characteristics that differ from RA presentation in men. Some of those characteristics
     could explain the high rates of disability and worse prognosis observed in
      women with RA in LAC"
    }
  ]
}
]
}
```


Figure 2.2: A sample of the training data of Task1b.

Figure 2.3 presents an abstract example of a JSON file with all the aforementioned elements. Note, that the structures can be nested (e.g. object with nested objects and arrays).

```
anObject={
     "attr1": "value1",
     "attr2": 23.45,
     "attr3": true,
     "array": [1, 2, 3]
     "array2": [
              {
                  "attrOfNestedObject": "value",
                  "anArray": ["string1", "string2"]
              },
              {
                  "attrOfNestedObject2": "value",
                  "anArray2": ["string1", "string2"]
              }
     ]
}
```

Figure 2.3: An abstract example of a JSON file.

3

Data Description

3.1 Task 1a

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 and Table 3.2 some basic statistics about the training data.

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.

In each batch of Task1a 6 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 only three test datasets are so far available. 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

[1] http://www.nlm.nih.gov/mesh/meshhome.html

Articles	10,876,004
Unique labels	26,563
Labels per article	12.55
Size in GB	22

Table 3.2: Basic statistics about the training data for Task1a.

Week	Batch 1	Batch 2	Batch 3
1	1,942 (1,256)	5,012 (1,153)	7,605 (783)
2	845 (607)	5,590 (1,133)	10,233
3	793 (465)	7,349 (1,553)	8,861
4	2,408 (555)	4,674 (1,034)	-
5	6,742 (1,516)	8,254 (1,617)	-
6	4,556 (1,231)	8,626 (1,319)	-
Total	17,286 (5,630)	39,505 (7,809)	26,699 (783)

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

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 that serves 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.) resulting to a file of 6.2GB.

3.2 Task 1b

In Task 1b, 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).

As in Task 1a the datasets follow the JSON format. More specifically, each dataset (development and test sets) contains an array of questions where each question (represented as an object in the JSON format) is constructed as follows:

- **id**: a unique id.
- **body**: the actual question.
- **type**: the type of the question:
  - Yes/No: these questions require strictly yes or no answers.
  - Factoid: these questions require a short expression as an answer like an entity name or a number.
  - List: these questions require a list of short expressions.

[2]http://lucene.apache.org/

[3]Under the Apache Licence: http://www.apache.org/licenses/LICENSE-2.0.html

  - Summary: these questions ask for a text summary from relevant resources.

- **concepts**: an array of concepts where each concept is a unique identifier that comes from the following terminologies and ontologies: a) Medical Subject Heading, b) Gene Ontology[4], c) Universal Protein Resource[5], d) Joint Chemical Dictionary[6] and e) Disease Ontology.[7]

- **documents**: an array of documents where each document is represented by its unique identifier (URL). The documents are retrieved from PUBMED.

- **exact_answer**: the exact answer format depends on the type of the question. For yes/no questions the exact answer will be a string (either 'yes' or 'no'). For factoid questions it will be an array of short expressions (limited to 5). In the case of list questions the format is an array of arrays where each inner array contains a number of short expressions. The total number of expressions is limited to 100 of no more 100 characters each. Finally, for summary questions no exact answer is required.

- **ideal_answer**: a string which summarizes the relevant information retrieved for the specific question. Ideal answers are valid for all types of question and are restricted to 200 words.

- **snippets**: an array of snippet objects from the returned documents. Note, that a snippet is defined as a continuous sequence of words. Each object in the array must contain the following information:

  - **document**: a unique string identifier of the document from where the snippet is extracted.
  - **beginSection**: a string that identifies the section of the document in which the snippets starts. In the JSON format, this is defined as "section.#b" where #b is the corresponding sequential number of the section.
  - **endSection**: a string that identifies the section of the document in which the snippets ends. It is formed in the same manner as the **beginSection** field.
  - **offsetInBeginSection**: an integer that represents the offset of the first character of the snippet with respect to the first character of the **beginSection**.
  - **offseInEndSection**:an integer that represents the offset of the first character of the snippet in the **endSection**, with respect to the first character of the **endSection**.[8]
  - **text**: the text of the snippet.

- **triples**: an array of triple objects where each object represents an RDF triple. The triples come from the Linked Life Platform[9]. The triple is constructed as follows:

  - **o**: a string for the object.
  - **p**: a string for the predicate.
  - **s**: a string for the subject.

Figure 3.1 presents the format of the data for Task 1b. An example, following this format is shown in Figure 2.2. Table 3.4 presents descriptive statistics of the training and the test data.

[4] http://www.geneontology.org/
[5] http://www.uniprot.org/
[6] http://www.biosemantics.org/index.php?page=Jochem
[7] http://disease-ontology.org/
[8] More information on offsets can be found in (Balikas et al., 2013).
[9] http://linkedlifedata.com/

```
{ "questions": [
        {
                "id": "the ID",
                "body": "the question?",
                "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.

	Training data	Test set 1	Test set 2	Test set 3
Questions	29	100	100	82
Yes/No	8	25	26	26
Factoid	5	18	20	16
List	8	31	31	23
Summary	8	26	23	17
Avg #concepts	4.8	5.3	6.0	12.9
Avg #documents	10.3	11.4	12.1	5.4
Avg #snippets	14.0	17.1	17.4	15.97

Table 3.4: Basic statistics of the training and test data for Task 1b provided during the evaluation procedure.

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.