BioASQ

A challenge on large-scale biomedical semantic indexing and question answering

Thierry Artières, Patrick Gallinari

LIP6, Universite Pierre et Marie Curie Paris 6

February 28, 2014

Web-Scale Classification Workshop, WSDM 2014

SEVENTH FRAMEWORK PROGRAMME

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

Vision

What is BioASQ

- BioASQ initiates a series of **challenges** on **biomedical semantic indexing** and **question answering (QA)**.
- Participants are required to index semantically content from **large-scale** biomedical resources (e.g. MEDLINE) and/or
- to assemble data from **multiple heterogeneous sources** (e.g. scientific articles, knowledge bases, databases)
- to compose **informative answers** to biomedical natural language questions.

Challenge objectives

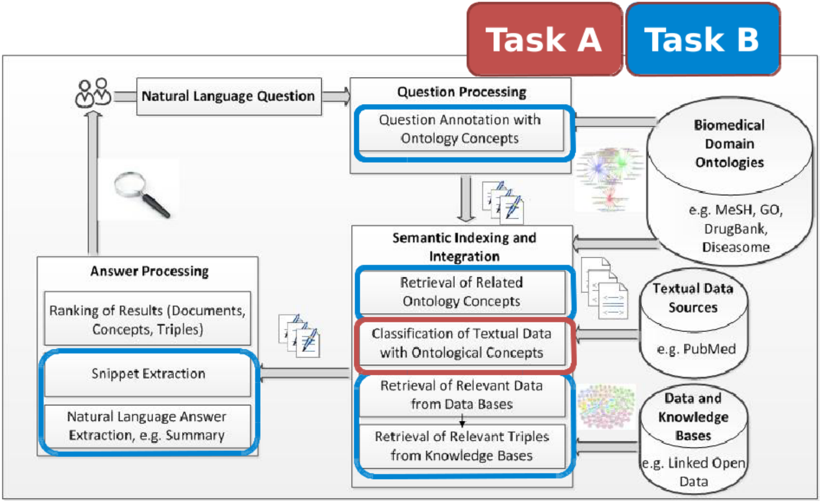
Scientific tasks

Task A

- **Large-scale classification of biomedical documents** onto ontology concepts (semantic indexing)

Task B

- **Classification of biomedical questions** onto relevant concepts
- **Retrieval** of relevant document snippets, concepts and knowledge base triples
- **Summarization** of the retrieved information in a concise and user-understandable form

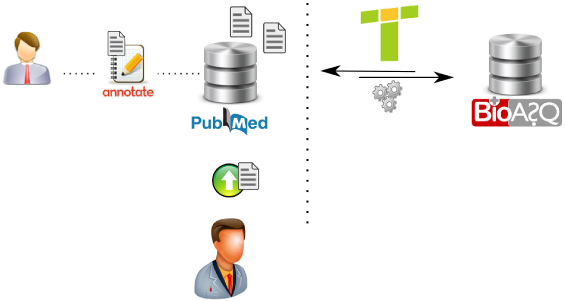


Task A
Task B
Natural Language Question
Question Processing
Question Annotation with Ontology Concepts
Biomedical Domain Ontologies
e.g. MeSH, GO, DrugBank, Diseasome
Semantic Indexing and Integration
Retrieval of Related Ontology Concepts
Classification of Textual Data with Ontological Concepts
Retrieval of Relevant Data from Data Bases
Retrieval of Relevant Triples from Knowledge Bases
Textual Data Sources
e.g. PubMed
Data and Knowledge Bases
e.g. Linked Open Data
Answer Processing
Ranking of Results (Documents, Concepts, Triples)
Snippet Extraction
Natural Language Answer Extraction, e.g. Summary

Challenges

Task A: Large-scale online biomedical semantic indexing

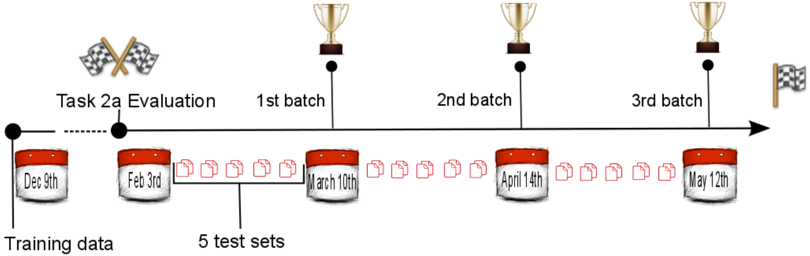
Hierarchical text classification



- Organizers distribute **new unclassfied PubMed articles.**
- Participants assign **MeSH terms** to the articles.
- **Automatic evaluation** based on annotations of **PubMed curators** (∼ 15 terms per article).

Challenges

Task A : Data and Schedule



Schedule :

- New test set : each week at Monday, 17.00 CET ($\sim$ 5k articles)
- Participants have to upload answers until Tuesday, 14.00 CET
- Challenge divided on 3 batches of 5 weeks/dataset
- A leaderboard by batch : enter when you want !

Challenges

Task B: Introductory biomedical semantic QA, IR, Summarization

Type of questions

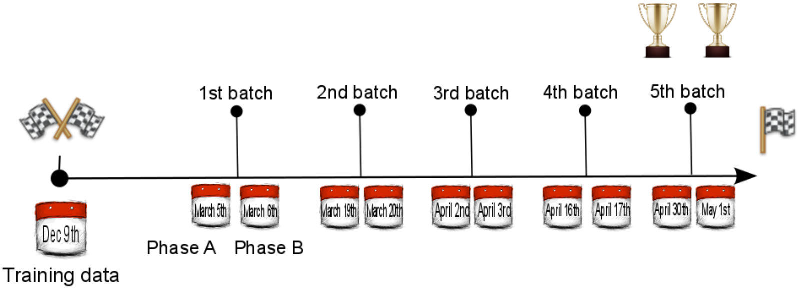
- **Yes/No** : *Do CpG islands colocalise with transcription start sites ?*
- **Factoid** : *Which virus is best known as the cause of infectious mononucleosis?*
- **List** : *Which are the Raf kinase inhibitors?*
- **Summary** : *What is the treatment of infectious mononucleosis ?*

Type of answers

- **exact** answers : **Yes/No**, relevant **concepts** and **triples** from ontologies, relevant **articles** from PubMed, relevant **text snippets**, . . .
- **ideal** answers : text summary explaining the exact answer

Challenges

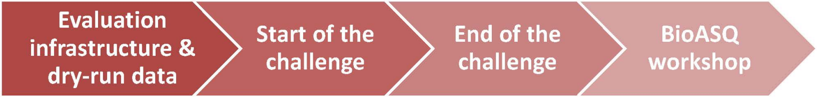
Task B : Data and schedule



- New test set each 2 weeks ($\sim$ 100 questions), 24h to upload results
- Two phases :
  - Phase A (IR phase) : upload only relevant concepts, articles, snippets, . . .
  - Phase B (QA, Summarization) : upload exact or/and ideal answers

Challenges

- Both tasks run twice, in **two cycles** (two years).
- **1st cycle completed**.



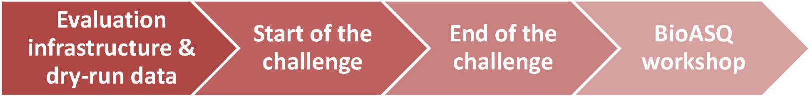
Evaluation infrastructure & dry-run data — March 2013

Start of the challenge — June 2013

End of the challenge — August 2013

BioASQ workshop — September 2013

- **2nd cycle** started **February 2014**.



Evaluation infrastructure & dry-run data — February 2014

Start of the challenge — March 2014

End of the challenge — May 2014

BioASQ workshop — September 2014

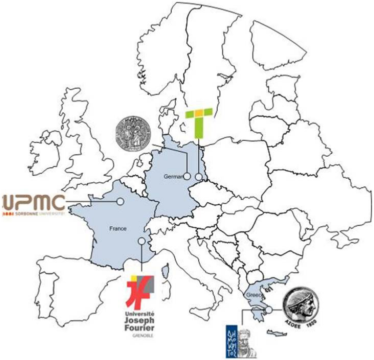
Challenges

Logistics

- **Regular announcement** of questions - **time-limits** on answers.
- Easy submission of results through **Web services**.
- Large **hardware infrastructure** (a cluster of 5000 cores) available for those who want to use it.
- **Partial participation** (any task, subtask, response type).
- **Prizes** to the best performing systems for each task/subtask.
- **Journal special issue** for outstanding methods.

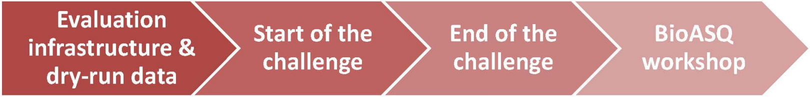
Project consortium

- National Centre for Scientific Research “Demokritos” - NCSR “D” (EL)
- Transinsight GmbH - TI (D)
- Universite Joseph Fourier -UJF (F)
- University Leipzig - ULEI (D)
- Universite Pierre et Marie Curie Paris 6 - UPMC (F)
- Athens University of Economics and Business-Research Centre - AUEB-RC (EL)



Thank you

Visit www.bioasq.org
Follow @BioASQ



Evaluation infrastructure & dry-run data — February 2014

Start of the challenge — March 2014

End of the challenge — May 2014

BioASQ workshop — September 2014