



## BioASQ

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

George Paliouras and Anastasia Krithara

NCSR "D"

16th September 2014

BioASQ Workshop, Sheffield



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

# Outline

Introduction

Presentation of the challenge

Task 2A

Task 2B

Challenge evaluation

Conclusions and Perspectives

Stay Tuned!

# Introduction

## What is BioASQ

### A competition funded by the European Union (FP7)

- ▶ 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.

# Presentation of the challenge

## Tasks

### Task A: Hierarchical text classification

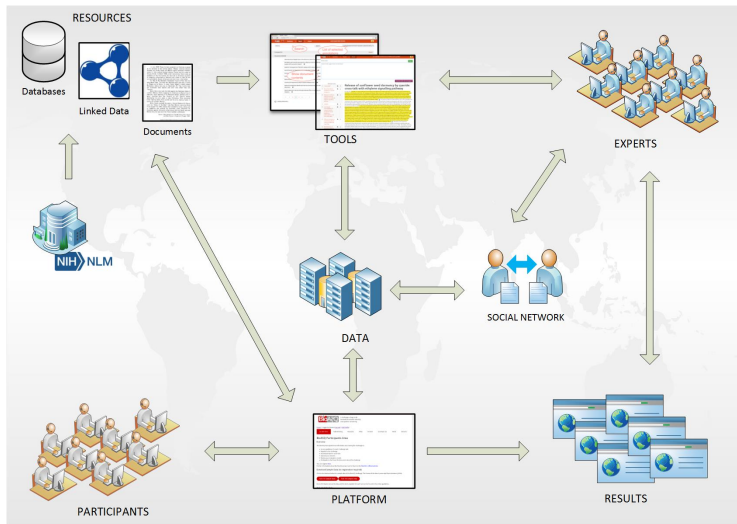
- ▶ Organizers distribute **new unclassified PubMed articles**.
- ▶ Participants assign **MeSH terms** to the articles.
- ▶ **Evaluation** based on annotations of **PubMed curators**.

### Task B: IR, QA, summarization

- ▶ Organizers distribute **English biomedical questions**.
- ▶ Participants provide: relevant **articles, snippets, concepts, triples, exact answers, summary answers**.
- ▶ **Evaluation:** both **automatic** (GMAP, MRR, Rouge etc.) and **manual** (by biomedical experts).

# Presentation of the challenge

## Behind the scenes



# Presentation of the challenge

## Resources

### Criteria for selecting the resources

- ▶ **Publicly available**
- ▶ **Coverage** of different biomedical subfields
- ▶ Widely **acceptable** and **usable** format (e.g. OWL, OBO)
- ▶ **Low degree of overlap** between them

### Selected resources

- ▶ Data sources include both text and structured info:
  - ▶ Task 1a: Medline articles and MeSH
  - ▶ Task 1b:
    - ▶ PubMed abstracts and PubMed Central articles
    - ▶ Gene Ontology, UniProt, Jochem, Disease Ontology

What makes **BioASQ** more challenging:

**LARGE SCALE** data and knowledge sources

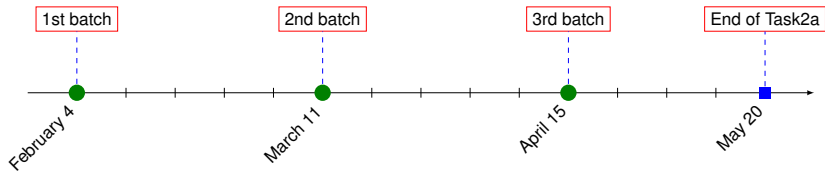
**REAL** questions and answers

of many different types

created by bio-medical experts

# Task 2A

## Schedule





# Task 2A

## Hierarchical text classification

Basic statistics about the **training** data

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

Number of articles for each **test** dataset in each batch.

<b>Week</b>	<b>Batch 1</b>	<b>Batch 2</b>	<b>Batch 3</b>
1	4440 (3319)	4085 (3422)	4342 (3009)
2	4721 (3734)	3496 (2788)	8840 (5883)
3	4802 (3884)	4524 (3274)	3702 (2860)
4	3579 (2431)	5407 (3923)	4726 (3252)
5	5299 (3693)	5454 (3666)	4533 (3252)
<b>Total</b>	<b>22,841 (17,061)</b>	<b>22,966 (17,073)</b>	<b>26,143 (18,256)</b>

# Task 2A

## Evaluation Measures

### Flat measures

---

- ▶ Accuracy (Acc.)
- ▶ Example Based Precision (EBP)
- ▶ Example Based Recall (EBR)
- ▶ Example Based F-Measure (EBF)
- ▶ Macro Precision/Recall/F-Measure (MaP, MaR, MaF)
- ▶ Micro Precision/Recall/F-Measure (MiP, MiR, MiF)

### Hierarchical measures

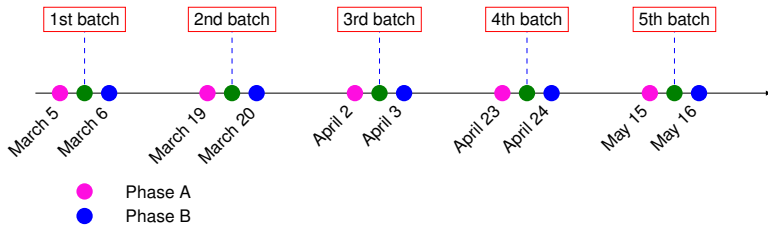
---

- ▶ Hierarchical Precision (HiP)
- ▶ Hierarchical Recall (HiR)
- ▶ Hierarchical F-Measure (HiF)
- ▶ Lowest Common Ancestor Precision (LCA-P)
- ▶ Lowest Common Ancestor Recall (LCA-R)
- ▶ Lowest Common Ancestor F-measure statistics for Task 1 (LCA-F)

A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras and I. Androutsopoulos: Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. Data Mining and Knowledge Discovery (To appear)

# Task 2B

## Schedule



# Task 2B

IR, QA, summarization

## Dataset

- ▶ 500 **Questions** and **gold reference answers** prepared by biomedical experts from around Europe.
  - ▶ Using tools/infrastructure developed by BioASQ.
- ▶ Four categories of questions:
  - ▶ Yes/No questions (both exact and ideal answer)
  - ▶ Factoids questions (both exact and ideal answer)
  - ▶ List questions (both exact and ideal answer)
  - ▶ Summary questions (ideal answer)

# Task 2B

## Examples of the different types of questions

- ▶ **Yes/No question:** *Is intense physical activity associated with longevity?*
- ▶ **Factoids question:** *Which is the protein (antigen) targeted by anti-Vel antibodies in the Vel blood group?*
- ▶ **List question:** *List the endoscopic diagnoses that have been reported in children with autism.*
- ▶ **Summary question:** *What is the role of thyroid hormone receptor alpha1 in insulin secretion?*

# Task 2B

## Annotation tool for the creation of the data for QA

The screenshot shows the BioASQ search results page. The interface includes a search bar at the top, a navigation menu with 'Question' and 'Search' tabs, and a 'Text question' button. Below the search bar, there are sections for 'Concepts' and 'Documents'. The 'Documents' section lists search results with titles, 'More info' links, and expand/collapse icons. Red callouts highlight specific features: 'search' points to the search bar; 'list of selected annotation sources' points to the 'Description Pattern of Antihi' dropdown menu; 'access document URL' points to the 'More info' links; and 'add result to annotation data sources' points to the expand/collapse icons.

search

list of selected annotation sources

access document URL

add result to annotation data sources

# Task 2B

**Social network** to help extend data, and set up new challenges

The screenshot shows a web browser window titled "BioASQ Social Network" with the URL "localhost:3000/#/questions". The page has a navigation bar with "Home", "Messages", "Questions", and "Timeline". The user "Norman Heino" is logged in. A "Sort by: votes" dropdown is visible. Two questions are listed:

- Question 1: "Which are the Atg8 homologs in human?" (created 2 months ago, 3 votes, buttons: Comment, Unfollow)
- Question 2: "Which are the known human transmembrane nucleoporins?" (created 2 months ago, 2 votes, buttons: Comment, Follow)

The answer for the second question is displayed in a grey box:

**ID**  
51bdb644047fa84d1d000001

**Ideal answer**  
Transmembrane nucleoporins (NUPs) are integral membrane components of the eukaryotic nuclear pore, playing an important role in the Nuclear Pore Complex (NPC) assembly. Even though the NPC is a conserved feature of all eukaryotes, different lineages possess some distinct transmembrane NUP subunits. Currently, four human transmembrane NUPs have been characterized, namely: NDC1 (also known as TMEM48 or NET3 or HNDC1), POM121 (also known as Nup121), GP210 (also known as Nuclear pore membrane glycoprotein 210 or Nuclear envelope pore membrane protein POM 210, POM210 or Nup210) and TMEM33 (or DBB3).

**Exact answer**

- NDC1, TMEM48, NET3, HNDC1
- POM121, Nup121
- GP210, Nuclear pore membrane glycoprotein 210, Nuclear envelope pore membrane protein POM 210, POM210, Nup210
- TMEM33, DBB3

At the bottom of the answer box are links for [concepts](#), [documents](#), [snippets](#), and [statements](#).

# Task 2B

## Statistics on datasets

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

The numbers for the documents, snippets, concepts and triples refer to averages



# Task 2B

## Evaluation measures

### ► Evaluating **Phase A** (IR)

Retrieved items	Unordered retrieval measures	Ordered retrieval measures
concepts	mean Precision, Recall, F-Measure	MAP, <b>GMAP</b>
articles		
snippets		
triples		

### ► Evaluating the '**exact**' answers for **Phase B** (Traditional QA)

Question type	Participant response	Evaluation measures
yes/no	'yes' or 'no'	<b>Accuracy</b>
factoid	up to 5 entity names	strict and lenient accuracy, <b>MRR</b>
list	a list of entity names	<b>mean</b> Precision, Recall, <b>F-measure</b>

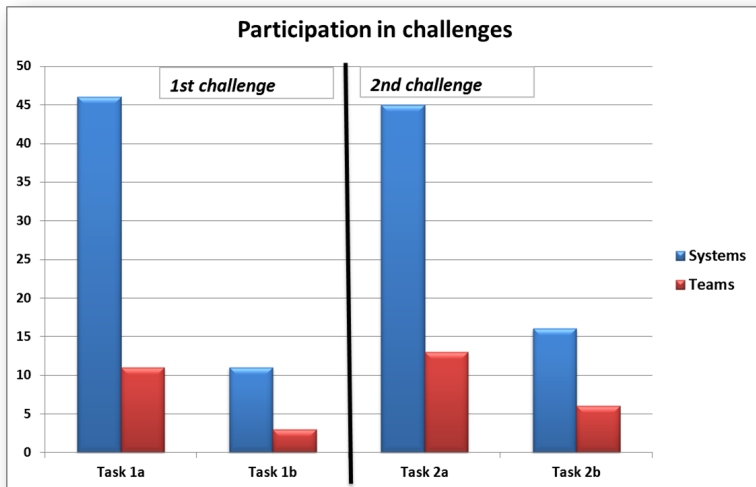
### ► Evaluating the '**ideal**' answers for **Phase B** (Query-focused Summarization)

Question type	Participant response	Evaluation measures
any	paragraph-sized text	ROUGE-2, ROUGE-SU4, <b>manual scores*</b> (Readability, Recall, Precision, Repetition)

\*with the help of BioASQ Assessment tool.

# Challenge evaluation

Comparison with first challenge participation



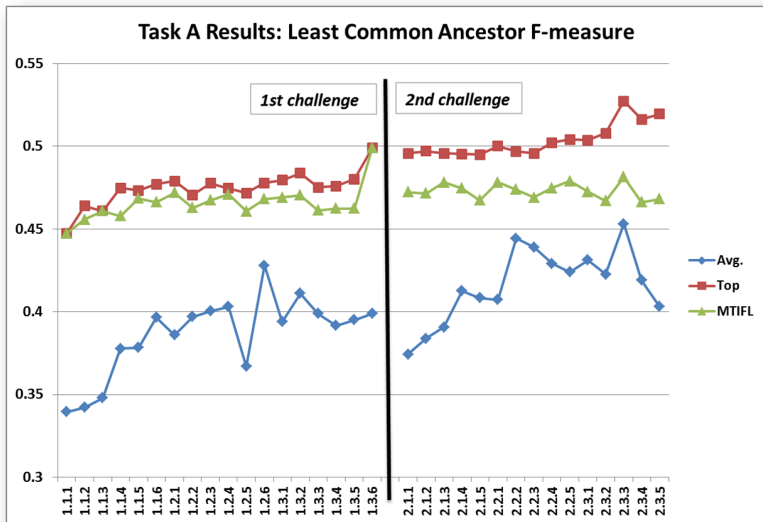
# Challenge evaluation

## Overall participation



# Challenge evaluation

## Results



# Conclusions and Perspectives

What we've learnt

## Main conclusions

- ▶ Both tasks are challenging and interesting.
- ▶ It is difficult for humans to provide all required golden truth.
- ▶ Manual assessment and improvement of the data was necessary in task 2b.
- ▶ Evaluation is an open issue in both tasks.
- ▶ Satisfactory participation in the both BioASQ challenges.

## Goals and perspectives

- ▶ Continue after the end of the project:
  - ▶ Task 2a continues running in non-challenge mode.
  - ▶ Oracle for continuous testing has been announced.
  - ▶ Social network for data creation and challenge set-up.
  - ▶ BioASQ 3 will run next year.

# Stay Tuned!

BioASQ project

Visit [www.bioasq.org](http://www.bioasq.org)

Follow [@BioASQ](https://twitter.com/BioASQ)

**Call for paper:**

Journal of Bio-Medical Semantics

**Supplement on Semantics-Enabled Biomedical Information  
Retrieval**

Deadline: 30 November 2014

**Stay tuned for BioASQ 3**