BioASQ

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

George Paliouras and Anastasia Krithara

NCSR “D”

27th September 2013

BioASQ Workshop, Valencia

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

Outline

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 form **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.

Introduction
What is BioASQ

Some facts

- The challenge runs twice, in **two cycles** (two years).
- **1st cycle completed**.
- **2nd cycle** starts **March 2014**.
- **Participation** can be **partial** (any task, subtask, response type).
- **Prizes** for each task/subtask.
- BioASQ datasets, infrastructure, evaluation services etc. **available beyond the end of the project**.
- **Advisory board:** both academia and industry
  - NLM, NIST, CMU, IBM, MSR etc.

Presentation of the challenge

Tasks

Task A: Hierarchical text classification

- Organizers distribute **new unclassfied 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

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

Presentation of the challenge

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 1A

Hierarchical text classification

Basic statistics about the **training** data

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

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

Week	Batch 1	Batch 2	Batch 3
1	1,942 (1,532)	5,012 (1,466)	7,650 (1,974)
2	845 (681)	5,590 (1,604)	10,233 (2,777)
3	793 (694)	7,349 (1,968)	8,861 (2.179)
4	2,408 (585)	4,674 (1,433)	1986 (1,069)
5	6,742 (3,194)	8,254 (2,428)	1750 (885)
6	4,556 (1,703)	8,626 (2,194)	1357 (506)
Total	17,286 (8,389)	39,505 (11,093)	31,837 (9,390)

Task 1A

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-meData statistics for Task 1aasure (LCA-F)

*A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras and I. Androutsopoulos: Evaluation Measures for Hierarchical Classification: a unified view and novel approaches

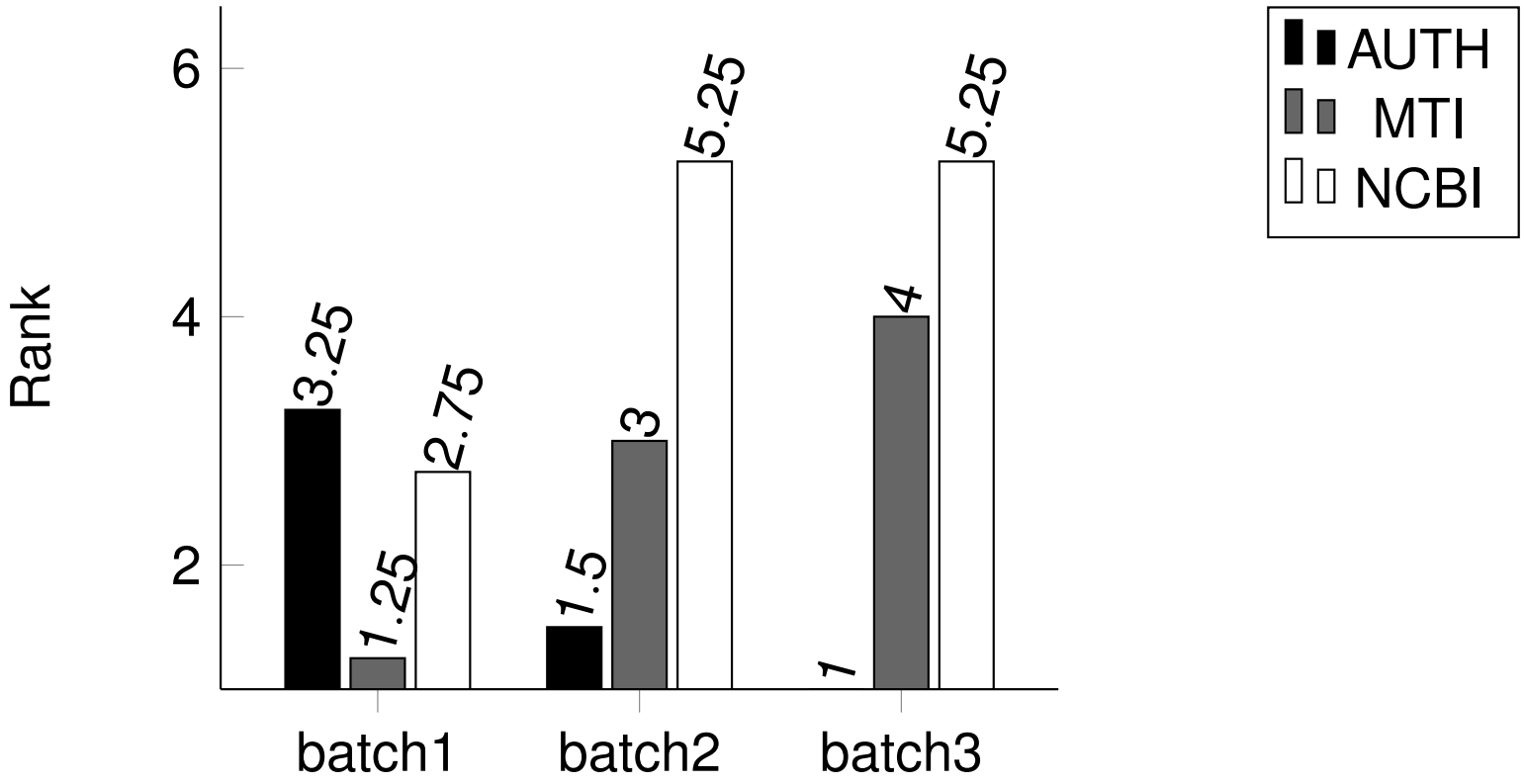
Task 1A

Participants

- **Baselines**:
  - BioASQ baseline (basic unsupervised system)
  - MTI default (as used by curators)
  - MTI First Line Index (biased on precision)
- **46 systems (11 teams)**:
  - Mayo Clinic, USA
  - University of Alberta, CANADA
  - Aristotle University of Thessaloniki, GREECE
  - University of Vigo, SPAIN
  - University of Colorado, USA
  - NCBI, NLM, USA
  - Universite de Rouen, FRANCE
  - Fudan University, CHINA
  - UCSD, USA
  - Toyota Technological Institute, JAPAN
  - Imran, PAKISTAN

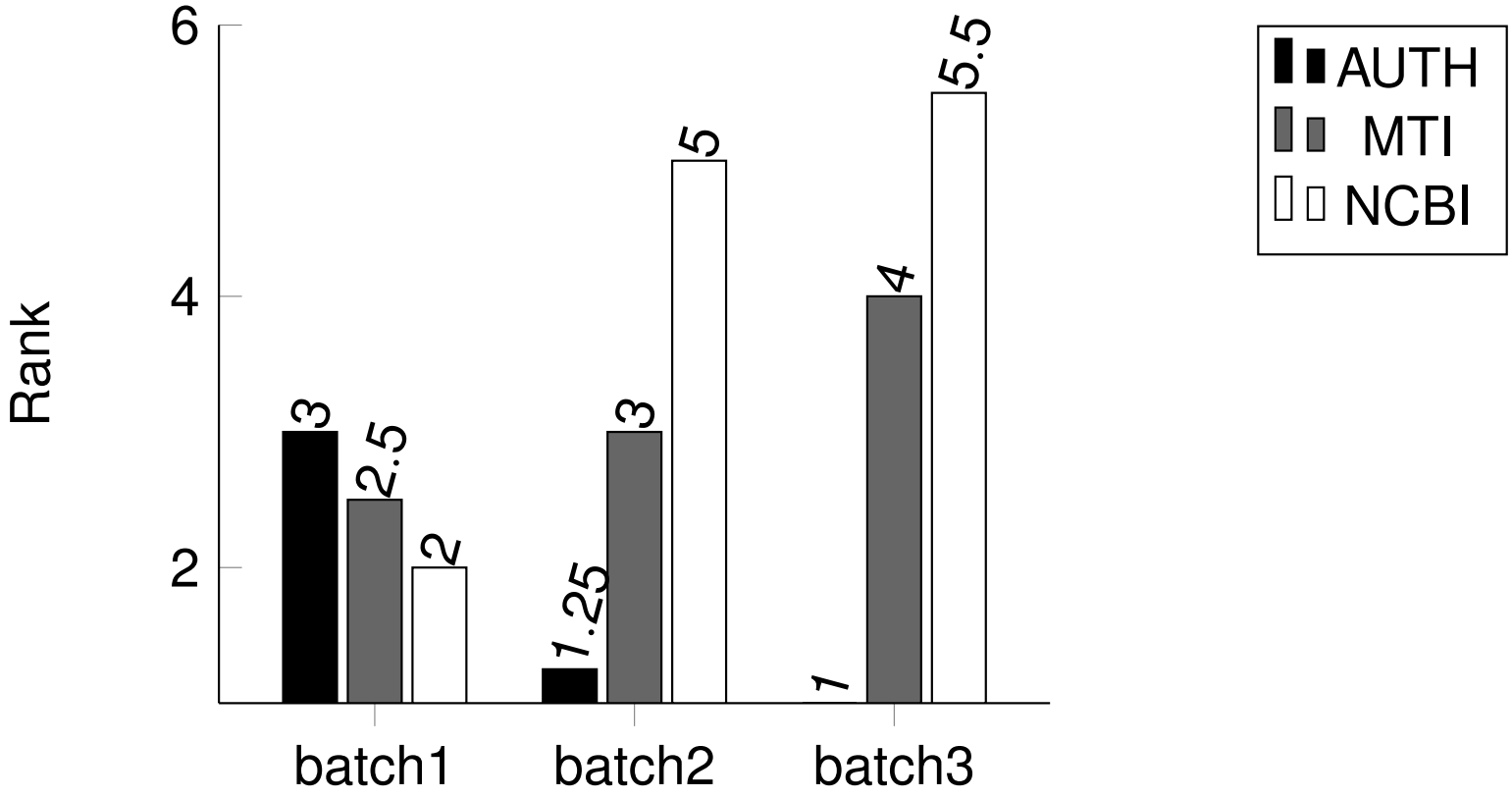
Task 1A

Results- MiF



Task 1A

Results - LCA-F



Task 1B

IR, QA, summarization

Dataset

- 311 **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)

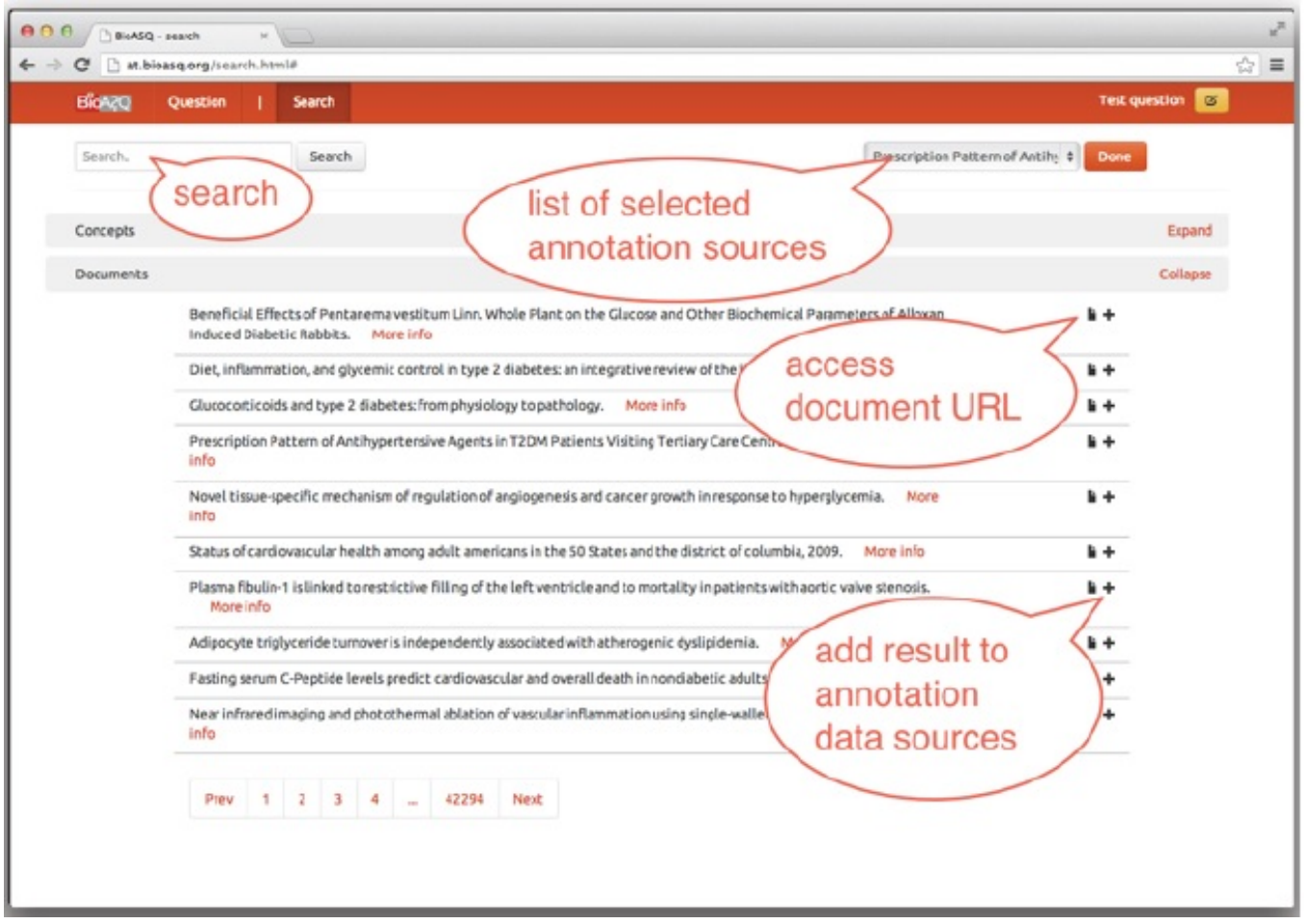
Task 1B

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 1B

Annotation tool for the creation of the data for QA



Task 1B

IR, QA, summarization

Basic statistics of the training and test data

	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	5.4
Avg #documents	10.3	11.4	12.1	12.9
Avg #snippets	14.0	17.1	17.4	15.97

Task 1B

Evaluation measures

- Evaluating **Phase A** (IR)

Retrieved items	Unordered retrieval measures	Ordered retrieval measures
concepts	mean Precision, Recall, F-Measure	MAP, **GMAP**
articles		
snippets		
triples		

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

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

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

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

*with the help of BioASQ Assessment tool.

Task 1B

Participating systems

Phase A

- **Baselines:**
  - Top 50/100 baseline
- **4 systems (2 teams)**:
  - Mayo Clinic, USA
  - University of Alberta, CANADA

Phase B

- **Baselines:**
  - BioASQ Baseline/ BioASQ Baseline 2
- **7 systems (2 teams)**:
  - University of Alberta, CANADA
  - Toyota Technological Institute, JAPAN

Task 1B

Statistics

Participation for Phases A and B of task 1B:

		Systems		Users	
	Size	**Phase A**	**Phase B**	**Phase A**	**Phase B**
Batch 1	100	3	4	2	2
Batch 2	100	4	5	2	2
Batch 3	82	2	3	1	1

Task 1B

Results - Phase B

Exact Answers (Yes/No) - Accuracy

System	Batch1	Batch2	Batch3
Wishart-S2	0.92	0.96	-
Baseline2	0.48	0.50	0.61
Baseline1	0.44	0.26	0.56
TRG	0.32	0.42	0.57
TRG2	-	0.42	0.57
TRG3	-	0.42	0.57

Task 1B

Results - Phase B

Exact Answers (List) - Mean F-measure

System	Batch1	Batch2	Batch3
Wishart-S2	0.23	0.33	-
Baseline2	0.02	0.08	0.03
Baseline1	0.02	0.08	0.03
TRG	0.0066	0.07	0.07
TRG2	-	0.04	0.06
TRG3	-	0.05	0.05

Task 1B

Results - Phase B

Exact Answers (Factoid) - MRR

System	Batch1	Batch2	Batch3
Wishart-S	0.31	0.30	-
Baseline2	0.10	0.07	0.04
Baseline1	0.02	0.08	0.04
TRG	0.03	0.03	0.03
TRG2	-	0.03	0.03
TRG3	-	0.04	0.03

Task 1B

Results - Phase B

Ideal Answers

System	Batch1	Batch2	Batch3
Wishart-S	**3.945**	**4.232**	-
TRG	3.352	3.397	**3.132**
TRG-2	-	3.342	3.075
TRG-3	-	3.340	2.987
Baseline1	2.862	-	3.195
Baseline2	2.732	-	3.175

Conclusions and Perspectives

Overall participation in BioASQ

- **117** registered users in BioASQ platform
- **73** systems registered, **46** different systems have submitted results for task 1A
- BioASQ website statistics:

Conclusions and Perspectives

What we've learnt

Main conclusions

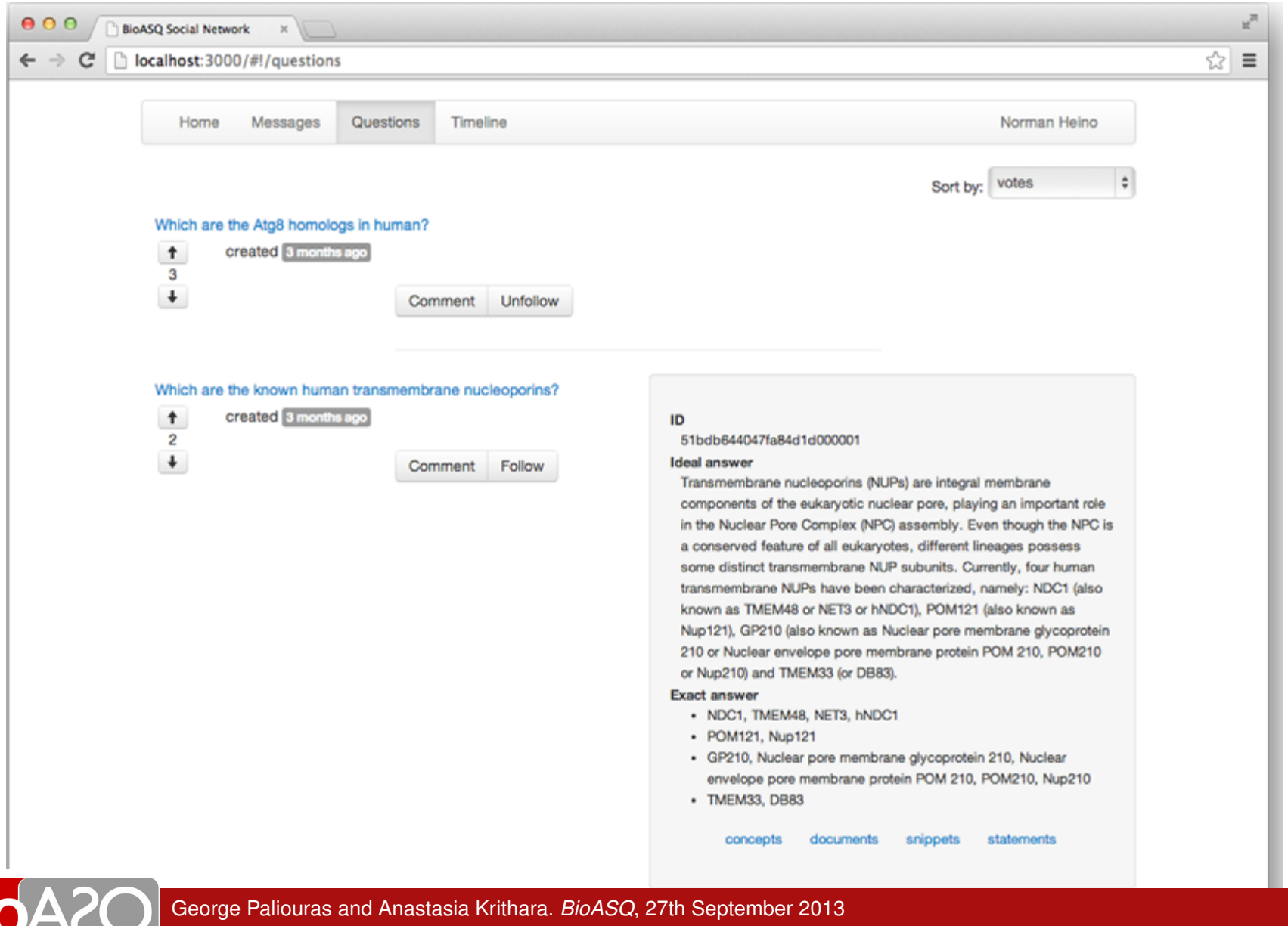
- Both tasks are challenging and interesting.
- It is difficult for humans to provide all required golden truth.
- Manual assessment and improvement of the data was necessary in task 1b.
- Evaluation is an open issue in both tasks.
- Satisfactory participation in the first year.

Goals and perspectives

- Aim to increase participation in the second year.
- Continue after the end of the project:
  - Task 1a continues running in non-challenge mode.
  - Oracle for continuous testing to be announced soon.
  - Social network for data creation and challenge set-up.

Social Network

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

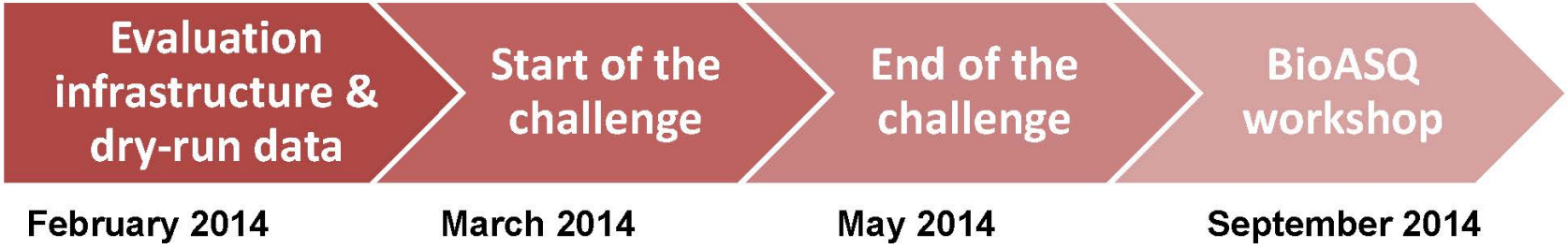


Stay Tuned!

BioASQ project

Visit www.bioasq.org
Follow @BioASQ

Please fill in the workshop survey
http://goo.gl/ncjzUH



Panel Discussion

BioASQ challenge next year and beyond the end of project

1. Evaluation measures: Which work best and how do we combine them?
2. What's the role of Semantic Indexing in QA?
3. Dissemination. How do we achieve:
   - 3.1 A better relation to other challenges such as BioNLP and BioCreative
   - 3.2 An increased participation
4. Would an oracle or open datasets be helpful? (What's the best evaluation format for QA? Online? Offline?)