Results of the fifth edition of the BioASQ Challenge

A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras and I. Kakadiaris

NCSR “Demokritos”, University of Houston

4th of August 2017

BioNLP Workshop, Vancouver
Introduction
What is BioASQ

A competition

- BioASQ is a series of challenges on biomedical semantic indexing and question answering (QA).
- Participants are required to semantically index 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 MEDLINE articles.
- Participants have 21 hours to assign MeSH terms to the articles.
- Evaluation based on annotations of MEDLINE curators.
Presentation of the challenge

Tasks

Task B: IR, QA, summarization

- Organizers distribute **English biomedical questions**.
- Participants have 24 hours to provide: relevant **articles**, **snippets**, **concepts**, **triples**, **exact answers**, **ideal answers**.
- **Evaluation**: both **automatic** (GMAP, MRR, Rouge etc.) and **manual** (by biomedical experts).
Presentation of the challenge
New task

Task C: Funding Information Extraction

- Organizers distribute **PMC full-text articles**.
- Participants have 48 hours to extract: **grant-IDs, funding agencies, full grants** (i.e. the combination of a grant-ID and the corresponding funding agency).
- **Evaluation** based on annotations of **MEDLINE curators**.
Presentation of the challenge

BioASQ ecosystem

Data sources include both **text** and **structured** info.
- PubMed articles.
- MeSH, Gene Ontology, UniProt, Jochem, Disease Ontology.

**Task A**: 12,800,000 articles with MeSH headings available for training. New test batches available every week.

**Task B**: 2,300 English questions, plus gold relevant documents, snippets, concepts, triples, "exact", and "ideal" answers.

**Task C**: 80,000 articles with corresponding funding information.

Online platform that enables participants to enter the BioASQ challenge.

Annotation & Assessment tool (Open Source).

Biomedical experts, responsible for the formulation of questions, the identification of gold relevant documents, snippets, concepts, triples, the composition of gold "exact" and "ideal" answers, and the manual evaluation of system responses.

A social network allows biomedical experts to follow and comment upon benchmark questions and gold answers formulated by their peers.

Presentation of the challenge
BioASQ ecosystem

**Task A:** 12,800,000 articles with MeSH headings available for training. New test batches available every week.

**Task B:** 2,300 English questions, plus gold relevant documents, snippets, concepts, triples, "exact", and "ideal" answers.

**Task C:** 80,000 articles with corresponding funding information.

Online platform that enables participants to enter the E
Presentation of the challenge

Per task

![Bar chart showing participants 2017](image)
Task 5A
Hierarchical text classification

▶ **Training data**

<table>
<thead>
<tr>
<th></th>
<th>version 2015</th>
<th>version 2016</th>
<th>version 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>11,804,715</td>
<td>12,208,342</td>
<td>12,834,585</td>
</tr>
<tr>
<td>Total labels</td>
<td>27,097</td>
<td>27,301</td>
<td>27,773</td>
</tr>
<tr>
<td>Labels per article</td>
<td>12.61</td>
<td>12.62</td>
<td>12.66</td>
</tr>
<tr>
<td>Size in GB</td>
<td>19</td>
<td>19.4</td>
<td>20.5</td>
</tr>
</tbody>
</table>

▶ **Test data**

<table>
<thead>
<tr>
<th>Week</th>
<th>Batch 1</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Batch 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Batch 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6,880 (6,661)</td>
<td>7,431 (7,080)</td>
<td>9,233 (5,341)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7,457 (6,599)</td>
<td>6,746 (6,357)</td>
<td>7,816 (2,911)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10,319 (9,656)</td>
<td>5,944 (5,479)</td>
<td>7,206 (4,110)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7,523 (4,697)</td>
<td>6,986 (6,526)</td>
<td>7,955 (3,569)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7,940 (6,659)</td>
<td>6,055 (5,492)</td>
<td>10,225 (984)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40,119 (34,272)</td>
<td>33,162 (30,934)</td>
<td>42,435 (21,323)</td>
<td></td>
</tr>
</tbody>
</table>

The numbers in parentheses are the annotated articles for each test dataset.
Task 5A
System approaches

- **Feature Extraction**: Representing each abstract
  - *tf-idf* of words and bi-words
  - *doc2vec* embeddings of paragraphs

- **Concept Matching**: Finding relevant MeSH labels
  - *k-NN* between article-vector representations
  - *Linear SVM* binary classifiers for each MESH label
  - *Recurrent Neural Networks* for sequence-to-sequence prediction
  - *UIMA-ConceptMapper* and *MeSHLabeler* tools for boosting NER and Entity-to-MeSH matching
  - *Latend Dirichlet Allocation and Labeled LDA* utilizing topics found in abstracts
  - *Ensemble* methodologies and stacking
## 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 (LCA-F)

---

Task 5A results
Evaluation

- Systems ranked using **MiF** (flat) and **LCA-F** (hierarchical).
- Results, in all batches and for both measures:
  1. **Fudan**
  2. **AUTH-Atypon**
Task 5A results

Task A Results: Micro F-measure

- 2013
- 2014
- 2015
- 2016
- 2017

Data points for Top, MTI, and Avg across years from 2013 to 2017.
### Task 5B
Statistics on datasets

<table>
<thead>
<tr>
<th>Batch</th>
<th>Size</th>
<th># of documents</th>
<th># of snippets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1,799</td>
<td>11.86</td>
<td>20.38</td>
</tr>
<tr>
<td>Test 1</td>
<td>100</td>
<td>4.87</td>
<td>6.03</td>
</tr>
<tr>
<td>Test 2</td>
<td>100</td>
<td>3.49</td>
<td>5.13</td>
</tr>
<tr>
<td>Test 3</td>
<td>100</td>
<td>4.03</td>
<td>5.47</td>
</tr>
<tr>
<td>Test 4</td>
<td>100</td>
<td>3.23</td>
<td>4.52</td>
</tr>
<tr>
<td>Test 5</td>
<td>100</td>
<td>3.61</td>
<td>5.01</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>2,299</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The numbers for the documents and snippets refer to averages.
Task 5B
Training Dataset Insights

▶ **1799** Questions
  ◦ 500 yes/no
  ◦ 486 factoid
  ◦ 413 list
  ◦ 400 summary

▶ **13** Experts

▶ ≈ **3450** unique biomedical concepts

---

![Bar chart showing the average of items per question for Concepts, Documents, and Snippets from 2013 to 2016. The chart indicates an increase in the average number of items per question over the years.](image)

---

G. Paliouras. *Results of the fifth edition of the BioASQ Challenge, 4th of August 2017*
Task 5B
Training Dataset Insights

- Broad terms (e.g. proteins, syndromes)
- More specific terms (e.g. cancer, heart, thyroid)
Task 5B
Training Dataset Insights

- Number of questions related to **cancer** vs **thyroid** per year
- The numbers on top of the bars denote the contributing experts
## Task 5B

### Evaluation measures

#### Evaluating **Phase A** (IR)

<table>
<thead>
<tr>
<th>Retrieved items</th>
<th>Unordered retrieval measures</th>
<th>Ordered retrieval measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>concepts</td>
<td>Mean Precision, Recall, F-Measure</td>
<td>MAP, GMAP</td>
</tr>
<tr>
<td>articles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>snippets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>triples</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Evaluating the ‘**exact**’ answers for **Phase B** (Traditional QA)

<table>
<thead>
<tr>
<th>Question type</th>
<th>Participant response</th>
<th>Evaluation measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes/no</td>
<td>‘yes’ or ‘no’</td>
<td>Accuracy</td>
</tr>
<tr>
<td>factoid</td>
<td>up to 5 entity names</td>
<td>strict and lenient accuracy, MRR</td>
</tr>
<tr>
<td>list</td>
<td>a list of entity names</td>
<td><strong>Mean</strong> Precision, Recall, <strong>F-measure</strong></td>
</tr>
</tbody>
</table>

#### Evaluating the ‘**ideal**’ answers for **Phase B** (Query-focused Summarization)

<table>
<thead>
<tr>
<th>Question type</th>
<th>Participant response</th>
<th>Evaluation measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>any</td>
<td>paragraph-sized text</td>
<td>ROUGE-2, ROUGE-SU4, <strong>manual scores</strong> <em>(Readability, Recall, Precision, Repetition)</em></td>
</tr>
</tbody>
</table>

*with the help of BioASQ Assessment tool.

---

G. Paliouras. *Results of the fifth edition of the BioASQ Challenge, 4th of August 2017*
Task 5B
System approaches

- **Question analysis**: Rule-based, regular expressions, ClearNLP, Semantic role labeling (SRL), Stanford Parser, tf-idf, SVD, word embeddings.
- **Query expansion**: MetaMap, UMLS, sequential dependence models, ensembles, LingPipe.
- **Document retrieval**: BM25, UMLS, SAP HANA database, Bag of Concepts (BoC), statistical language model.
- **Snippet selection**: Agglomerative Clustering, Maximum Marginal Relevance, tf-idf, word embeddings.
- **Exact answer generation**: Standford POS, PubTator, FastQA, SQuAD, Semantic role labeling (SRL), word frequencies, word embeddings, dictionaries, UMLS.
- **Ideal answer generation**: Deep learning (LSTM, CNN, RNN), neural nets, Support Vector Regression.
- **Answer ranking**: Word frequencies.

Task 5B Results

- Our experts are currently assessing systems’ responses
- The results will be announced in autumn
Task 5C
Statistics on datasets

Grant ID distribution in training data set

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>62,952</td>
<td>22,610</td>
</tr>
<tr>
<td>Grant IDs</td>
<td>111,528</td>
<td>42,711</td>
</tr>
<tr>
<td>Agencies</td>
<td>128,329</td>
<td>47,266</td>
</tr>
<tr>
<td>Time Period</td>
<td>2005-13</td>
<td>2015-17</td>
</tr>
</tbody>
</table>

- 104 unique agencies
- 92,437 unique grant IDs
Task 5C
Statistics on datasets

Number of articles per agency in training dataset

- Wellcome Trust
- Public Health Service
- National Institute on Drug Abuse
- National Institute on Deafness and other Communication Disorders
- National Institute on Alcohol Abuse and Alcoholism
- National Institute on Aging
- National Institute of Neurological Disorders and Stroke
- National Institute of Mental Health
- National Institute of General Medical Sciences
- National Institute of Environmental Health Sciences
- National Institute of Diabetes and Digestive and Kidney Diseases
- National Institute of Dental and Craniofacial Research
- National Institute of Child Health and Human Development
- National Institute of Arthritis and Musculoskeletal and Skin Diseases
- National Institute of Allergy and Infectious Diseases Extramural Activities
- National Human Genome Research Institute
- National Heart, Lung, and Blood Institute
- National Eye Institute
- National Cancer Institute
- Medical Research Council
- Howard Hughes Medical Institute
- Division of Epidemiology and Clinical Applications
- Cancer Research UK
- Canadian Institutes of Health Research
- British Heart Foundation
- Biotechnology and Biological Sciences Research Council

G. Paliouras. Results of the fifth edition of the BioASQ Challenge, 4th of August 2017
Task 5C
Evaluation measures

- A **subset** of the Grant IDs and Agencies mentioned in full text are available in ground truth data ⇒ **Micro-Recall**
  - Each Grant ID (or lone Agency) must exist verbatim in the text
- Different scores for each subtask:
  - Grant IDs
  - Agencies
  - Full Grants
Task 5C
System approaches

- **Grant Support Sentences**: Identifying sentences containing grant information
  - Features: $tf-idf$ of n-grams
  - Techniques: SVM and *Naive Bayes* for scoring, specific XML fields considered

- **Grant Information Extraction**: Detecting Grant-IDs and Agencies
  - Manually crafted *Regular Expressions*
  - *Heuristic Rules*
  - *Sequential Learning Models*, such as *Conditional Random Fields*, *Hidden Markov Models*, *Max Entropy Models*
  - Ensemble of classifiers for pairing Grant-IDs to Agencies

---

Task 5C

Results

<table>
<thead>
<tr>
<th>Agency</th>
<th>Micro-Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fudan</td>
<td>0.975</td>
</tr>
<tr>
<td>AUTH</td>
<td>0.95</td>
</tr>
<tr>
<td>DZG</td>
<td>0.924</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agency</th>
<th>Micro-Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fudan</td>
<td>0.991</td>
</tr>
<tr>
<td>AUTH</td>
<td>0.986</td>
</tr>
<tr>
<td>DZG</td>
<td>0.912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agency</th>
<th>Micro-Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fudan</td>
<td>0.953</td>
</tr>
<tr>
<td>AUTH</td>
<td>0.941</td>
</tr>
<tr>
<td>DZG</td>
<td>0.844</td>
</tr>
</tbody>
</table>

Results of the fifth edition of the BioASQ Challenge, 4th of August 2017
Conclusions and Perspectives

Goals and perspectives

- BioASQ will run in 2018.
- Continuous development of benchmark datasets.

BioASQ DATASETS
12,800,000 articles in task A
2,300 questions in task B
80,000 articles in task C

BioASQ PLATFORM
36,898 users
48,307 sessions
91,243 pageviews

BioASQ WEB SITE
16,020 users
35,804 sessions
92,195 pageviews
Conclusions and Perspectives

Oracle for continuous testing

Task: Task A
Test: Task 4a: Test batch 3, Week 5
Your system: bioasq_baseline

Select the task you are submitting results for.
Specify the test set by choosing one from the drop down menu. The tests sets for both tasks can be downloaded from here and are those that been already used for the BioASQ challenge.
Select one of your systems that will be used in the "Oracle Results" tab.
Select a file to upload that contains a JSON string with the answers of a test. The format of the JSON is described in the online guidelines of each task, e.g. here.

Attention: Calculating the evaluation results takes several minutes. Please, do not refresh the content.

Results
Annotated documents: 627 out of 3130.
Please, take a look at the results below and fill the following form:

Keep my results visible: ☑️
If enabled, your uploaded results will be visible in the oracle to any registered user. Otherwise, it will be visible only to you.

Save my score: ☑️
If enabled, it will replace the previous score for the selected system and testset in the BioASQ database.

Flat Measures

<table>
<thead>
<tr>
<th>System</th>
<th>MiF</th>
<th>Acc.</th>
<th>EBP</th>
<th>EBR</th>
<th>EBF</th>
<th>MaP</th>
<th>MaR</th>
<th>MaF</th>
<th>MiP</th>
<th>MiR</th>
</tr>
</thead>
<tbody>
<tr>
<td>auth1</td>
<td>0.5954</td>
<td>0.4247</td>
<td>0.5887</td>
<td>0.6133</td>
<td>0.5793</td>
<td>0.5659</td>
<td>0.4776</td>
<td>0.4593</td>
<td>0.5948</td>
<td>0.5959</td>
</tr>
<tr>
<td>Current Submission</td>
<td>0.5817</td>
<td>0.4091</td>
<td>0.5843</td>
<td>0.5994</td>
<td>0.5641</td>
<td>0.5481</td>
<td>0.4821</td>
<td>0.4634</td>
<td>0.5794</td>
<td>0.5841</td>
</tr>
<tr>
<td>d33p</td>
<td>0.5746</td>
<td>0.3978</td>
<td>0.6150</td>
<td>0.5473</td>
<td>0.5507</td>
<td>0.5626</td>
<td>0.3897</td>
<td>0.3811</td>
<td>0.6143</td>
<td>0.5397</td>
</tr>
<tr>
<td>Default MTI</td>
<td>0.5854</td>
<td>0.4165</td>
<td>0.6036</td>
<td>0.5934</td>
<td>0.5711</td>
<td>0.5369</td>
<td>0.5173</td>
<td>0.4960</td>
<td>0.5967</td>
<td>0.5745</td>
</tr>
</tbody>
</table>
Collaborations

- NLM
  - Task A design and baselines
  - Task C design and baselines
- CMU
  - OAQA Baselines for task B
- DBCLS
  - BioASQ and PubAnnotation: Using linked annotations in biomedical question answering (BLAH3)
- iASiS
  - Question answering over big heterogeneous biomedical data for precision medicine
Grateful to the BioASQ consortium

BioASQ started as a European FP7 project, with the following partners:

- National Centre for Scientific Research “Demokritos” (GR)
- Transinsight GmbH (DE)
- Universite Joseph Fourier (FR)
- University Leipzig (DE)
- Universite Pierre et Marie Curie Paris 6 (FR)
- Athens University of Economics and Business Research Centre (GR)
Sponsors

PLATINUM SPONSOR

NIH NLM

SILVER SPONSOR

Atypon
Stay Tuned!

Visit www.bioasq.org
Follow @BioASQ

BioASQ 6 to be announced soon!