1 Publishable summary

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

European project BioASQ sets a challenge to push research in biomedical information retrieval and question answering

Vision and objectives

Every day, approximately 3000 new articles are published in biomedical journals. That averages to more than 2 articles every minute! Managing this large amount of data is a challenge in itself. Yet, ensuring that this wealth of knowledge is used for the sake of the patients in a timely manner is an even more demanding task for both computer scientists and biomedical experts. The BioASQ project, which started on October 1st, 2012 and is running for 2 years, aims to push research in information technology towards highly precise biomedical information retrieval systems. The project will achieve this goal through a competition (challenge or shared task), in which systems from teams around the world compete. BioASQ provides the data, software, hardware and the evaluation infrastructure for the challenge. By these means, the project will ensure that the biomedical experts of the future can rely on software tools to identify, process and present the fragments of the huge space of biomedical resources that address their personal questions.

The tasks included in the BioASQ challenges will help advance the state of the art in two fields. First, the automatic classification of biomedical documents will be improved. Here, systems will be required to tag large numbers of scientific biomedical articles with terms from a predefined biomedical vocabulary. Additionally, the challenge will evaluate how well systems identify text fragments in scientific articles, and related data in public knowledge bases, in order to answer questions set by the biomedical expert team of BioASQ. Further results of the project include a set of open-source tools and a social network that support experts in setting up similar challenges, beyond the end of the project.

The BioASQ team combines researchers with complementary expertise from 6 organisations in 3 countries: the Greek National Center for Scientific Research “Demokritos” (coordinator), participating with its Institutes of ‘Informatics & Telecommunications’ and ‘Biosciences & Applications’, the German IT company Transinsight GmbH, the French University Joseph Fourier, the German research Group for Agile Knowledge Engineering and Semantic Web at the University of Leipzig, the French University Pierre et Marie Curie-Paris 6 and the Research Center of the Athens University of Economics and Business in Greece. Moreover, biomedical experts from several countries assist in the creation of the evaluation data and a number of key players in the industry and academia from around the world participate in the advisory board of the project.

Progress so far

1st BioASQ Challenge

The first BioASQ challenge was successfully completed in August 2013. It comprised two tasks:

BioASQ Task 1a: Large-scale online biomedical semantic indexing

This task was based on the standard process followed by MedLine to index journal abstracts. The participants were asked to classify new MedLine documents, written in English, as they became available online, before MedLine curators annotated (in effect, classified) them manually. The classes came from the MeSH hierarchy; they were the subject headings that are currently used to manually index the abstracts. As new manual annotations became available, they were used to

evaluate the classification performance of participating systems (that classified articles before they were manually annotated), using standard IR measures (e.g., precision, recall, accuracy), as well as hierarchical variants of them. The participants were able to train their classifiers, using the whole history of manually annotated abstracts.

Task 1a ran for three consecutive periods (batches) of 6 weeks each. The first batch started in April 2013. On Monday each week, a new test set was released and the participants were asked to provide the classes of the test documents within 21 hours. The evaluation results were calculated gradually, as MedLine curators provided manual classification.

In the duration of the 18 weeks that it ran, 88,628 documents were provided for testing. 12 teams from various countries in three continents (Europe, North America and Asia) participated, representing academic (e.g. University of Alberta in Canada and Aristotle University of Thessaloniki in Greece), as well as industrial research (e.g. Toyota Technological Institute in Japan and Mayo Clinic in the USA). Competition was particularly intense, with each team participating with more than one systems (up to 5 were allowed). The MTI system of NLM was used as one of the baseline systems and was particularly hard to beat, as it is used to recommend MeSH terms to the MedLine curators, in order to speed up their work.

The evaluation of the systems was based on both established measures used for flat classification, as well as novel hierarchical measures, proposed by the BioASQ consortium (see report at http://arxiv.org/abs/1306.6802). Separate winners were announced for each batch and were awarded the corresponding prizes. The winning teams used various advanced text mining techniques and we were positively surprised to see one of the systems to almost consistently outperform the highly optimised MTI baseline. More details about the winners of the first BioASQ challenge are available at http://www.bioasq.org/participate/first-challenge-winners.

BioASQ Task 1b: Introductory biomedical semantic QA

Task 1b used benchmark datasets containing development and test questions, in English, along with gold standard (reference) answers. The benchmark datasets were constructed by a team of biomedical experts from around Europe. Task 1b ran in two phases:

- Phase A: This phase corresponds effectively to the information retrieval stage of question answering. In this phase, BioASQ released questions from the benchmark datasets and the participants had to respond with relevant concepts (from designated terminologies and ontologies), relevant articles (in English, from PubMed), relevant snippets (from the relevant articles), and relevant RDF triples (from designated knowledge bases).
- Phase B: This is the phase of the construction of the final answer, taking one of the four following forms, according to the type of the question:
    - 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 only)

  The term 'ideal answer' refers to a paragraph-sized text that is considered as the ideal response by human experts. Clearly, a range of advanced language technologies are required to address this hard task, such as multi-document text summarization and natural language generation from ontologies. In this phase, BioASQ released questions and gold (correct) relevant concepts, articles, snippets, and RDF triples from the benchmark datasets. The participants had to respond with exact answers (e.g., named entities in the case of factoid questions) and ideal answers (paragraph-sized summaries), both in English.

The test dataset of Task 1b was released in three batches, each containing approximately 100 questions. In each batch, first only the questions of the batch were released, and the participants had to submit their answers for Phase A (concepts, articles, snippets, RDF triples) within 21 hours; then the gold concepts, articles, snippets, and RDF triples for the questions of the batch were also provided, and the participants again had 21 hours to submit their answers for Phase B.

Due to the complexity of the task, both participation and evaluation of the results was particularly demanding. Three teams with long experience and infrastructure for Question Answering participated in the task, representing again both academic (University of Alberta in Canada) and industrial research (Toyota Technological Institute in Japan and Mayo Clinic in the USA). Automated evaluation results were provided for all aspects of the challenge, including intermediate and final results e.g., mean average precision in Phase A; accuracy, mean reciprocal rank, mean F-measure for exact answers to yes/no, factoid, and list questions in Phase B; ROUGE for ideal answers). However, in addition to the automated scores, the BioASQ biomedical expert team was asked to provide manual scores (for readability, information recall and precision, lack of repetitions) on the final 'ideal' answer that each system produced in Phase B. Despite the complexity of the task and the short time that the participants had for preparing their systems, the BioASQ experts seemed particularly happy about the result that the participants produced, judging from the manual scores that they provided. More details about the winners of the first BioASQ challenge are available at http://www.bioasq.org/participate/first-challenge-winners.

Infrastructure, tools and benchmark data

In order to support the challenge, BioASQ has built powerful and agile infrastructure for developing benchmark data sets for biomedical semantic indexing and question answering, as well as for using these data to evaluate participating systems, either automatically or manually. Most of the software produced by BioASQ is provided as open source and the data are provided free of charge for research use.

The basic tool provided for benchmark data generation is the **BioASQ annotation tool**, which can be used by biomedical experts to create questions and answering material of the form used in task 1b of the first BioASQ challenge. The tool is publicly available as an open-source project at https://github.com/AKSW/BioASQ-AT. It has been already tested and improved based on feedback collected by the BioASQ biomedical expert team and a new version of it will be released in the second year of the BioASQ project.

Based on the annotation tool, an assessment tool was created in order to help the experts manually assess the created corpus. The **BioASQ assessment tool** has the same design as the annotation tool, in order to minimise the required training time for its biomedical expert users. It allows the experts to assess the results of systems that participate in the challenge and improve the benchmark data set, based on the system responses. The BioASQ assessment tool will also be provided as an open-source project. Figure 1 provides sample screenshots of the two tools.

In addition to the tools provided for the biomedical experts, BioASQ has constructed a platform for setting up and managing the evaluation campaign. The platform is available online at http://bioasq.lip6.fr/ (BioASQ Participants' Area). The functionality of the online platform includes: (i) the unit that enables users to register in the platform, (ii) the Web services and the Web interface that enable users to upload/download data, (iii) the evaluation function that calculates the evaluation measures (iv) the discussion forum (v) the detailed online documentation and guidelines for both tasks (vi) an e-mail help desk that is publicly accessible. The source code of the platform will be made openly available at the end of the project, in order to be used in the future to set up new biomedical QA challenges, possibly based on new benchmarks produced by the BioASQ expert network and/or with additional challenge tasks.

Using the tools and infrastructure developed by BioASQ, as well as several knowledge sources, two benchmark datasets have been created in the first year of the project, one for each task.

In particular, for task 1a, the dataset consists of biomedical abstracts published on Medline, together with MeSH headings, manually assigned to them. The initial dataset provided for training purposes, contained more than 10,000,000 articles with more than 26,000 unique labels.

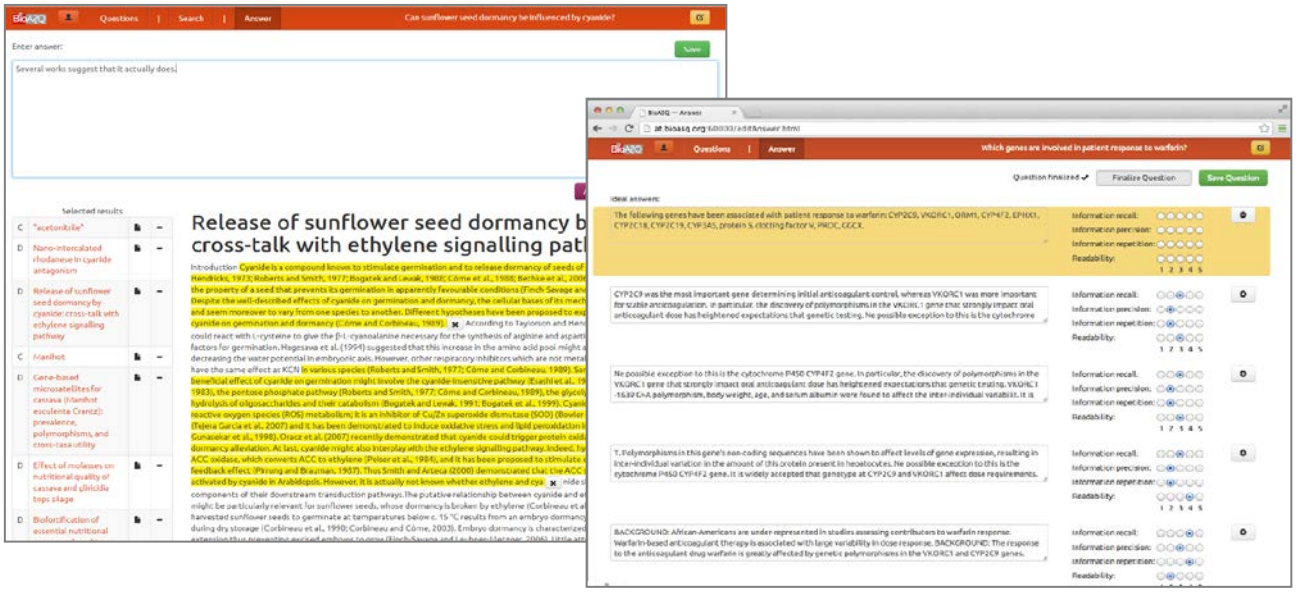

Figure 1: Screenshots of the BioASQ Annotation and the BioASQ Assessment tools

Additionally, the test data that is produced weekly since April 22, 2013 is available for testing purposes through the BioASQ Participants' Area. The generation of weekly test data continued beyond the end of the challenge in August 2013, in order to help people who are interested in this task tune and prepare their systems. So far, more than 100,000 test documents have been provided.

For task 1b, 311 real questions and gold reference answers, as well as related concepts, documents and snippets were prepared by biomedical experts from around Europe. The emphasis was on realistic and interesting (from the experts point of view) questions, as well as on the generation of rich accompanying material (concepts, triples, documents, snippets, and the actual answers). This material, which is unique in the domain of biomedical QA will be made available openly for research purposes. This will give the opportunity also to develop and train the systems that will participate in the second BioASQ challenge and beyond.

Community presence

A particularly important goal of the BioASQ project is to raise the awareness of the community on the importance of the semantic indexing and question answering task and explain how the BioASQ challenge can help in promoting research in this area. In order to achieve that goal, BioASQ was presented in a variety of very diverse fora and organisations. Among these activities the organisation of the first BioASQ workshop, as a post-conference event of CLEF 2013, was clearly the most important one.

The first BioASQ workshop took place on September 27th, 2013 in UPV, Valencia, Spain. The goal of the workshop was to further the interaction with the wider community of biomedical semantic indexing and question answering. The workshop was attended by more than 30 people, who contributed to a lively full-day programme, comprising 9 talks and a panel discussion. The workshop presentations included an overview of the first-year BioASQ challenge, details of some of the most competitive systems that participated in the challenge and two invited talks by Dr. Alan Roy Aronson from the NLM and Dr. Jennifer Chu-Caroll from IBM Research. In addition to the two invited speakers, two other prominent researchers in the field joined our discussion panel: Prof. Udo Hahn from the University of Jena, Germany and Dr. Rebholz-Schuhmann from the University of Zurich, Switzerland. Information about the workshop and its proceedings are available at http://www.bioasq.org/workshop.

In its efforts to raise awareness, BioASQ is happy to be assisted by an **advisory board** of key international figures in the area. The BioASQ advisory board consists of 29 international experts from the areas of bio-informatics, computational biology and medical informatics, who are assisting


Figure 2: Pictures from the first BioASQ workshop, in Valencia, Spain.

BioASQ significantly in its decision-making process and in informing the community about the BioASQ challenge. Details of the members of the BioASQ Advisory Board are available at http://www.bioasq.org/project/advisory-board

The establishment of the BioASQ advisory board was the result of an active effort to establish collaborations with related organisations and projects around the world. Among these, the collaboration with personnel of the US National Library of Medicine is worth noting, as it has helped BioASQ greatly in setting up and running task 1a. Naturally, there is a potential mutual benefit, as the technology developed for this task of the BioASQ challenge can be used to increase the efficiency and effectiveness of the MedLine content indexing process. Other notable collaborations are with the European research projects Visceral and Khrsemoi, with the IBM Watson research team, who have become known worldwide for the success of their automated QA system in the Jeopardy TV show, as well as with the related CLEF challenges QA4MRE and QALD, with which a joined QA track will be organised in the context of CLEF 2014 (see http://nlp.uned.es/clef-qa/).

Additionally, BIOASQ has been presented in various high-attendance conferences (such as WWW, ESWC, ICML, NAACL, ECIR, BioNLP, ICNB/ECCB 2013, AAAI Fall Symposium 2012) as well as companies (e.g. IBM Research and Brazil's largest media company Globo), who have showed particular interest in the potential results of the BioASQ effort.

One of the most recent activities of BioASQ, was the establishment of the **social network** of biomedical experts, which is available at http://sn.bioasq.org/. The purpose of the social network is to provide a common forum for interested biomedical experts to access and provide feedback on the benchmark data that is produced for the BioASQ challenges. The BioASQ social network is an open source project (available at https://github.com/AKSW/BioASQ-SN) and will continue to exist after the end of the project, providing a platform for maintaining and extending the BioASQ benchmarks, based on contributions and evaluation by peers. Figure 3 provides an illustration of the BioASQ social network.

BioASQ' s expected impact

The main long-term goal of BioASQ is to push significantly the research in information systems and methods that aim in turn at improved access to biomedical information. The potential impact of such a development is enormous and affects the biomedical experts, companies providing services in this industry, including information technology providers, and eventually everyone who will benefit from improved biomedical processes.

On the way to this big goal, BioASQ is set to facilitate a number of significant intermediate results, among which:

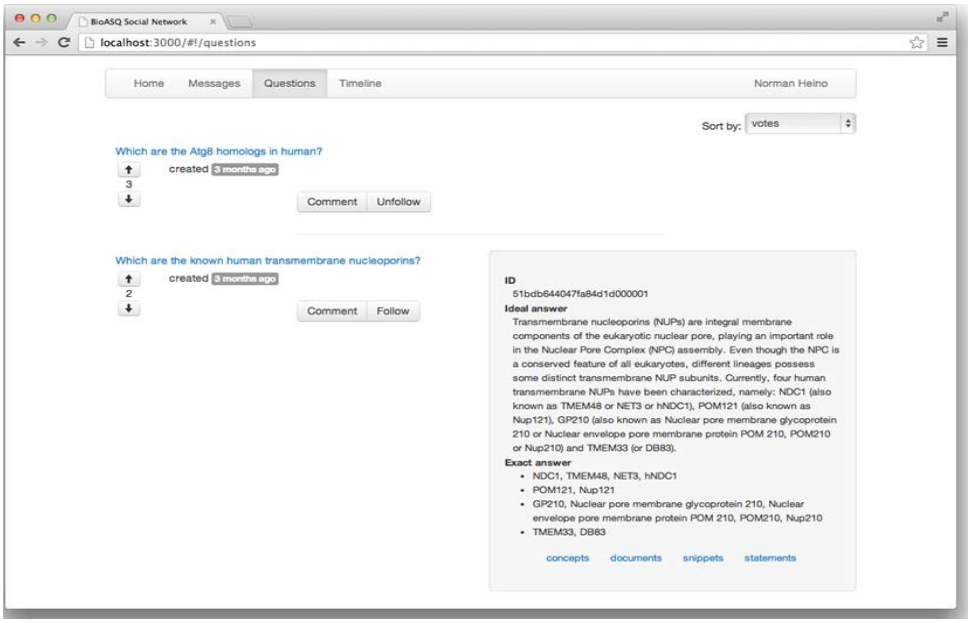

Figure3: A screenshot of the BioASQ Social Network

- Better understanding of the current semantic indexing and question answering technologies and their limitations.
- Improved awareness of the biomedical community about the possibility of significant improvement of their work, using intelligent information systems.
- Establishment of bridges among information technologists and biomedical experts, with the common goal of creating challenging tasks for current information systems.
- A number of tools, infrastructure and benchmark data that facilitate the organisation of BioASQ challenges, beyond the end of the project.

Plans for the rest of the project

The main target for the second year of the project is to successfully organize the second challenge and attract as many as possible of the key players to participate in the tasks. The ultimate goal of the project is making the challenge ***viable***, at ***very low cost*** after the end of the project. For this purpose, additional effort is planned in the following directions:

- Improving the quality and increasing even more the quantity of the data for task 2b.
- Making training data available as early as possible.
- Advertising the BioASQ in the context of the QA-track in CLEF 2014.
- Increasing the involvement of the biomedical community, through the social network.

Further information:

http://www.bioasq.org/ http://www.linkedin.com/groups/BioASQ-4801043

http://twitter.com/bioasq http://plus.google.com/104709672946762321818

Project Co-ordinator: Georgios Paliouras, NCSR “DEMOKRITOS”

paliourg@iit.demokritos.gr http://www.iit.demokritos.gr/~paliourg/