Results of the 4^{th} edition of BioASQ Challenge

Anastasia Krithara[1], **Anastasios Nentidis**[1], **George Paliouras**[1], and **Ioannis Kakadiaris**[2]

[1]National Center for Scientific Research "Demokritos", Athens, Greece
[2]University of Houston, Texas, USA

Abstract

The goal of this task is to push the research frontier towards hybrid information systems. We aim to promote systems and approaches that are able to deal with the whole diversity of the Web, especially for, but not restricted to, the context of biomedicine. This goal is pursued by the organization of challenges. The fourth challenge, as the previous challenges, consisted of two tasks: semantic indexing and question answering. 16 systems participated by 7 different participating teams for the semantic indexing task. The question answering task was tackled by 37 different systems, developed by 11 different teams. 25 of the systems participated in the phase A of the task, while 12 participated in phase B. 3 of the teams participated in both phases of the question answering task. Overall, as in previous years, the best systems were able to outperform the strong baselines. This suggests that advances over the state of the art were achieved through the BIOASQ challenge but also that the benchmark in itself is very challenging. In this paper, we present the data used during the challenge as well as the technologies which were at the core of the participants' frameworks.

1 Introduction

The aim of this paper is twofold. First, we aim to give an overview of the data issued during the BioASQ challenge in 2016. In addition, we aim to present the systems that participated in the challenge and for which we received system descriptions, as well as evaluate their performance. To achieve these goals, we begin by giving a brief overview of the tasks, including the timing of the different tasks and the challenge data. Thereafter, we give an overview of the systems which participated in the challenge and provided us with an overview of the technologies they relied upon. Detailed descriptions of some of the systems are given in lab proceedings. The evaluation of the systems, which was carried out by using state-of-the-art measures or manual assessment, is the last focal point of this paper. The conclusion sums up the results of this challenge.

2 Overview of the Tasks

The challenge comprised two tasks: (1) a large-scale semantic indexing task (Task 4a) and (2) a question answering task (Task 4b).

Large-scale semantic indexing. In Task 4a the goal is to classify documents from the PubMed[1] digital library into concepts of the MeSH[2] hierarchy. Here, new PubMed articles that are not yet annotated are collected on a weekly basis. These articles are used as test sets for the evaluation of the participating systems. As soon as the annotations are available from the PubMed curators, the performance of each system is calculated by using standard information retrieval measures as well as hierarchical ones. The winners of each batch were decided based on their performance in the Micro F-measure (MiF) from the family of flat measures (Tsoumakas et al., 2010), and the Lowest Common Ancestor F-measure (LCA-F) from the family of hierarchical measures (Kosmopoulos et al., 2013). For completeness several other flat and hierarchical measures were reported (Balikas et al., 2013). In order to provide an on-line and large-scale scenario, the task was divided into three independent batches. In each batch 5 test

[1]http://www.ncbi.nlm.nih.gov/pubmed/
[2]http://www.ncbi.nlm.nih.gov/mesh/

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1–7,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

sets of biomedical articles were released consecutively. Each of these test sets were released in a weekly basis and the participants had 21 hours to provide their answers. Figure 1 gives an overview of the time plan of Task 4a.

Biomedical semantic QA. The goal of task 4b was to provide a large-scale question answering challenge where the systems should be able to cope with all the stages of a question answering task, including the retrieval of relevant concepts and articles, as well as the provision of natural-language answers. Task 4b comprised two phases: In phase A, BIOASQ released questions in English from benchmark datasets created by a group of biomedical experts. There were four types of questions: "yes/no" questions, "factoid" questions,"list" questions and "summary" questions (Balikas et al., 2013). Participants had to respond with relevant concepts (from specific terminologies and ontologies), relevant articles (PubMed articles), relevant snippets extracted from the relevant articles and relevant RDF triples (from specific ontologies). In phase B, the released questions contained the correct answers for the required elements (articles and snippets) of the first phase. The participants had to answer with *exact* answers as well as with paragraph-sized summaries in natural language (dubbed *ideal* answers).

The task was split into five independent batches. The two phases for each batch were run with a time gap of 24 hours. For each phase, the participants had 24 hours to submit their answers. We used well-known measures such as mean precision, mean recall, mean F-measure, mean average precision (MAP) and geometric MAP (GMAP) to evaluate the performance of the participants in Phase A. The winners were selected based on MAP. The evaluation in phase B for the ideal answers was carried out manually by biomedical experts on the answers provided by the systems. For the sake of completeness, ROUGE (Lin, 2004) is also reported. For the exact answers, we used accuracy for the yes/no questions, mean reciprocal rank (MRR) for the factoids and mean F-measure for the list questions.

3 Overview of Participants

3.1 Task 4a

In this subsection we describe the proposed systems which have sent a description and stress their key characteristics.

In (Papagiannopoulou et al., 2016) flat classification processes were employed for the semantic indexing task. In particular, they used as a training set the last 1 million articles and kept the last 50 thousand as a validation set. Pre-processing of the articles was carried out by concatenated the abstract and the title. One-grams and bi-grams were used as features, removing stop-words and features with less than five occurrences in the corpus. The tf-idf representation has been used for the features. The proposed system includes several multi-label classifiers (MLC) that are combined in ensembles. In particular, they used the Meta-Labeler, a set of Binary Relevance (BR) models with Linear SVMs and a Labeled LDA variant, Prior LDA. All the above models were combined in an ensemble, using the MULE framework, a statistical significance multi-label ensemble that performs classifier selection.

The approach proposed by (Segura-Bedmar et al., 2016) is based on Elastic Search. They use ElasticSearch in order to index the training set provided by the BioASQ. Then, each document in the test set is translated into a query, that is fired against the index built from the training set, returning the most relevant documents and their MeSH categories. Finally, each MeSH category is ranked using a scoring system based on the frequency of the category and the similarity of relevant documents, which contain the category, with the test document to classify.

Baselines. During the challenge three systems were served as baseline systems. The first baseline is a state-of-the-art method called Medical Text Indexer (MTI) (Mork et al., 2014) which is developed by the National Library of Medicine[3] and serves as a classification system for articles of MEDLINE. MTI is used by curators in order to assist them in the annotation process. The second baseline is an extension of the system MTI with the approaches of the first BioASQ challenge's winner (Tsoumakas et al., 2013). The third one, dubbed BioASQ_Filtering (Zavorin et al., 2016) is

[3] http://ii.nlm.nih.gov/MTI/index.shtml

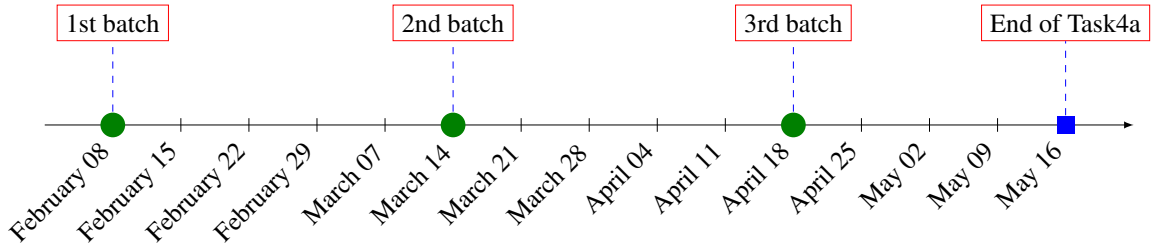

Figure 1: The time plan of Task 4a.

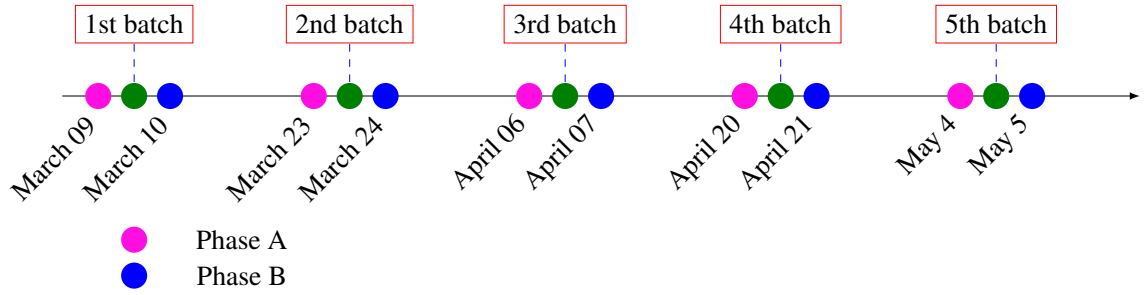

Figure 2: The time plan of Task 4b. The two phases for each batch run in consecutive days.

a new extension of the MTI system. In particular, Learning to Rank methodology is used as a boosting component of the MTI system. The improved system shows significant gains in both precision and recall for some specific classes of MeSH headings.

3.2 Task 4b

As mentioned above, the second task of the challenge is split into two phases. In the first phase, where the goal is to annotate questions with relevant concepts, documents, snippets and RDF triples 9 teams with 25 systems participated. In the second phase, where teams are requested to submit exact and paragraph-sized answers for the questions, 5 teams with 12 different systems participated.

The system presented in (Papagiannopoulou et al., 2016) is based on Indri search engine, and they use MetaMap and LingPipe to detect the biomedical concepts in local ontology files. For the relevant snippets, they calculate the semantic similarity between each one of the sentences and the query (expanded with synonyms) using a semantic similarity measure. Concerning phase B, They provided exact answers only for the factoid questions. Their system is based on their previous participation in BioASQ challenge (Papanikolaou et al., 2014). The system tries to extract the lexical answer type by manipulating the words of the question. Then, the relevant snippets of the question which are provided as inputs for this tasks are processed with the 2013 release of MetaMap in order to extract candidate answers. This year, they have extended their approach by expanding both the scoring mechanism, as well as the set of candidate answers.

The system presented in (Yang et al., 2016), extends the system in (Yang et al., 2015). In particular, they used TmTool (CH et al., 2016), in addition to MetaMap, to identify possible biomedical named entities, especially out-of-vocabulary concepts. In addition, they also extract frequent multi-word terms from relevant snippets to further improve the recall of concept and candidate answer text extraction. They also introduced a unified classification interface for judging the relevance of each retrieved concept, document, and snippet, which can combine the relevant scores evidenced by various sources. A supervised learning method is used to rerank the answer candidates for factoid and list questions based on the relation between each candidate answer and other candidate answers.

The system presented in (Schulze et al., 2016) relies on the Hana Database for text processing. It uses the Stanford CoreNLP package for tokenizing the questions. Each of the tokens is then sent to the BioPortal and to the Hana database for concept retrieval. The concepts retrieved from

the two stores are finally merged to a single list that is used to retrieve relevant text passages from the documents at hand. The second system relies on existing NLP functionality in the IMDB. They have extended it with new functions tailored specifically to QA.

The approach presented in (gu Lee et al., 2016) participated in phase A of task 4b. The main focus was the retrieval of relevant documents and snippets. The proposed system uses a clusterbased language model. Then, it reranks the retrieved top-n sentences using five independent similarity models based on shallow semantic analysis.

4 Results

4.1 Task 4a

During the evaluation phase of the Task 4a, the participants submitted their results on a weekly basis to the online evaluation platform of the challenge[4]. The evaluation period was divided into three batches containing 5 test sets each. 7 teams were participated in the task with a total of 16 systems. For measuring the classification performance of the systems several evaluation measures were used both flat and hierarchical ones (Balikas et al., 2013). The micro F-measure (MiF) and the Lowest Common Ancestor F-measure (LCA-F) were used to asses the systems and choose the winners for each batch (Kosmopoulos et al., 2013). $12,208,342$ articles with $27,301$ labels (19.4GB) were provided as training data to the participants. Table 1 shows the number of articles in each test set of each batch of the challenge.

Table 2 presents the correspondence of the systems for which a description was available and the submitted systems in Task 4a. The systems MTI First Line Index, Default MTI, BioASQ_Filtering were the baseline systems used throughout the challenge. Systems that participated in less than 4 test sets in each batch are not reported in the results[5].

According to (Demsar, 2006) the appropriate way to compare multiple classification systems over multiple datasets is based on their average rank across all the datasets. On each dataset the system with the best performance gets rank 1.0, the second best rank 2.0 and so on. In case that two or more systems tie, they all receive the average rank.

Tables 3 presents the average rank (according to MiF and LCA-F) of each system over all the test sets for the corresponding batches. Note, that the average ranks are calculated for the 4 best results of each system in the batch according to the rules of the challenge[6]. The best ranked system is highlighted with bold typeface.

Table 4: Statistics on the training and test datasets of Task 4b. All the numbers for the documents, snippets, concepts and triples refer to averages.

Batch	Size	# of documents	# of snippets
training	1307	13.00	17.86
1	100	4.56	6.41
2	100	5.25	6.98
3	100	4.79	6.46
4	100	4.90	7.25
5	97	3.93	6.10
total	1804	10.71	14.77

4.2 Task 4b

Phase A. Table 4 presents the statistics of the training and test data provided to the participants. The evaluation included five test batches. For the phase A of Task 4b the systems were allowed to submit responses to any of the corresponding types of annotations, that is documents, concepts, snippets and RDF triples. For each of the categories we rank the systems according to the Mean Average Precision (MAP) measure (Balikas et al., 2013). The final ranking for each batch is calculated as the average of the individual rankings in the different categories. In tables 6 and 7 some indicative results from batch 1 are presented. The detailed results for Task 4b phase A can be found in `http://participants-area.bioasq.org/results/4b/phaseA/`.

Phase B. In the phase B of Task 4b the systems were asked to report exact and ideal answers. The systems were ranked according to the manual evaluation of ideal answers by the BioASQ experts (Balikas et al., 2013), and according to automatic measures for the exact answers.

Table 7 shows the results for the exact answers for the first batch of task 4a. In case that systems

[4] `http://participants-area.bioasq.org/`

[5] According to the rules of BioASQ, each system had to participate in at least 4 test sets of a batch in order to be eligible for the prizes.

[6] `http://participants-area.bioasq.org/general_information/Task4a/`

Table 1: Statistics on the test datasets of Task 4a.

Batch	Articles	Annotated Articles	Labels per article
1	3,740	569	11.25
	2,872	714	12.01
	2,599	275	11.09
	3,294	520	13.72
	3,210	418	11.23
Subtotal	15,715	2,496	11.96
2	3,212	443	10.57
	3,213	371	11.37
	2,831	534	11.78
	3,111	541	10.67
	2,470	268	9.82
Subtotal	14,837	2,157	10.94
3	2,994	89	12.08
	3,044	353	11.79
	3,351	241	10.81
	2,630	93	9.77
	3,130	50	12.56
Subtotal	15,149	826	11.35
Total	45,701	5,479	11.42

Table 2: Correspondence of reference and submitted systems for Task 4a.

Reference	Systems
(Papagiannopoulou et al., 2016)	Auth1, Auth2
(Segura-Bedmar et al., 2016)	LABDA ElasticSearch, LargeElasticLABDA, LABDA baseline
Baselines ((Mork et al., 2013),(Zavorin et al., 2016))	MTI First Line Index, Default MTI, BioASQ_Filtering

Table 3: Average ranks for each system across the batches of the task 4a for the measures MiF and LCA-F. A hyphenation symbol (-) is used whenever the system participated in less than 4 times in the batch.

System	Batch 1		Batch 2		Batch 3	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
iria-1	-	-	9.0	9.0	-	-
LABDA ElasticSearch	-	-	-	-	-	-
d33p	-	-	-	-	-	-
auth1	2.75	3.25	3.75	3.75	-	-
Default MTI	4.0	3.0	5.0	4.5	-	-
auth2	-	-	6.0	6.25	-	-
MeSHLabeler	**1.25**	**1.25**	**1.25**	**1.25**	-	-
LargeElasticLABDA	-	-	-	-	-	-
LABDA baseline	-	-	-	-	-	-
BioASQ Filtering	4.5	4.75	5.75	5.5	-	-
MeSHLabeler-2	-	-	2.0	2.0	-	-
MeSHLabeler-1	1.75	1.75	-	-	-	-
MeSHLabeler-3	-	-	3.5	3.25	-	-
CSX-1	-	-	-	-	-	-
MTI First Line Index	5.5	5.75	5.75	6.25	-	-
UCSDLogReg	-	-	-	-	-	-

didn't provide exact answers for a particular kind of questions we used the symbol "-". The results of the other batches are available at `http://participants-area.bioasq.org/results/4b/phaseB/`. From those results we can see that the systems are achieving a very high ($> 90\%$ accuracy) performance in the yes/no questions. The performance in factoid and list questions is not as good indicating that there is room for improvements.

5 Conclusion

In this paper, an overview of the fourth BioASQ challenge is presented. As the previous challenges, the challenge consisted of two tasks: semantic indexing and question answering. Overall, as in previous years, the best systems were able to outperform the strong baselines provided by the organizers. This suggests that advances over the state of the art were achieved through the BIOASQ challenge but also that the benchmark in

Table 5: Results for batch 1 for documents in phase A of Task 4b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
testtext	0.169	0.5331	0.2276	0.0981	0.0128
ustb_prir2	0.158	0.5277	0.2164	0.0973	0.0119
ustb_prir4	0.165	0.5254	0.2224	0.0967	0.0109
fdu2	0.147	0.5011	0.2012	0.0885	0.0087
ustb_prir3	0.156	0.497	0.2114	0.0869	0.0095
fdu	0.153	0.5086	0.2081	0.0866	0.0095
ustb_prir1	0.155	0.4936	0.2097	0.0865	0.0088
fdu4	0.15	0.5057	0.205	0.0859	0.012
fdu3	0.154	0.5184	0.2112	0.0849	0.0109
fdu5	0.149	0.4971	0.2036	0.0823	0.01
KNU-SG Team_Korea	0.084	0.2258	0.1065	0.0486	0.0008
HPI-S1	0.1209	0.3266	0.1547	0.0474	0.0012
Auth001	0.069	0.1983	0.0914	0.0375	0.0004
WS4A	0.01	0.0134	0.011	0.0038	0
HPI-S2	0.005	0.0062	0.0054	0.0028	0

Table 6: Results for batch 1 for snippets in phase A of Task 4b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
HPI-S1	0.0822	0.1706	0.0917	0.0481	0.0005
KNU-SG Team_Korea	0.0482	0.0952	0.0534	0.0266	0.0002
ustb_prir2	0.0469	0.1135	0.0503	0.0216	0.0002
ustb_prir3	0.0452	0.1070	0.0482	0.0212	0.0002
ustb_prir1	0.0409	0.1080	0.0491	0.0211	0.0002
ustb_prir4	0.0449	0.1108	0.0477	0.0201	0.0002
testtext	0.0433	0.1098	0.0460	0.0188	0.0002

Table 7: Results for batch 3 for exact answers in phase B of Task 4b.

System	Yes/no	Factoid			List		
	Accuracy	Strict Acc.	Lenient Acc.	MRR	Precision	Recall	F-measure
fa1	0.9600	0.1154	0.1923	0.1442	0.2500	0.3000	0.2641
Lab Zhu ,Fdan Univer	0.9600	0.1923	0.2692	0.2192	0.1450	0.5929	0.2181
LabZhu,FDU	0.9600	0.1923	0.2692	0.2192	0.1444	0.6214	0.2176
LabZhu_FDU	0.9600	0.1923	0.2692	0.2192	0.1420	0.5929	0.2132
Lab Zhu,Fudan Univer	0.9600	0.1923	0.2692	0.2192	0.1455	0.5770	0.2185
oaqa-3b-3	0.5200	0.2308	0.2692	0.2436	0.5396	0.5008	0.4828
WS4A	0.2400	0.0385	0.0385	0.0385	0.1172	0.2817	0.1609
LabZhu-FDU	0.0400	0.1923	0.2692	0.2192	0.1420	0.5929	0.2132

itself is very challenging. Consequently, we regard the outcome of the challenge as a success towards pushing the research on bio-medical information systems a step further. In future editions of the challenge, we aim to provide even more benchmark data derived from a community-driven acquisition process.

Acknowledgments

The fourth edition of BioASQ is supported by a conference grant from the NIH/NLM (number 1R13LM012214-01) and sponsored by the Atypon company.

References

Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artieres, and Patrick Gallinari. 2013. Evaluation Framework Specifications. Project deliverable D4.1, 05/2013.

Wei CH, Leaman R, and Lu Z. 2016. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics*.

Janez Demsar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30.

Hyeon gu Lee, Minkyoung Kim, Juae Kim, Maengsik Choi Sunjae Kwon, Youngjoong Ko, Yi-Reun Kim, Jung-Kyu Choi, Harksoo Kim, and Jungyun Seo. 2016. KSAnswer: Question-answering System of Kangwon National University and Sogang University in the 2016 BioASQ Challenge . In *In Proceedings of the BioASQ Workshop, in ACL*.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. 2013. Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. *CoRR*, abs/1306.6802.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop 'Text Summarization Branches Out'*, pages 74–81, Barcelona, Spain.

James Mork, Antonio Jimeno-Yepes, and Alan Aronson. 2013. The NLM Medical Text Indexer System for Indexing Biomedical Literature. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*.

James G. Mork, Dina Demner-Fushman, Susan C. Schmidt, and Alan R. Aronson. 2014. Recent enhancements to the nlm medical text indexer. In *Proceedings of Question Answering Lab at CLEF*.

Eirini Papagiannopoulou, Yiannis Papanikolaou, Dimitris Dimitriadis, Sakis Lagopoulos, Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, and Ioannis Vlahavas. 2016. Large-Scale Semantic Indexing and Question Answering in Biomedicine. In *In Proceedings of the BioASQ Workshop, in ACL*.

Yannis Papanikolaou, Dimitrios Dimitriadis, Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, and Ioannis Vlahavas. 2014. Ensemble Approaches for Large-Scale Multi-Label Classification and Question Answering in Biomedicine. In *2nd BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*.

Frederik Schulze, Ricarda Schuler, Tim Draeger, Daniel Dummer, Alexander Ernst, Pedro Flemming, Cindy Perscheid, and Mariana Neves. 2016. HPI Question Answering System in BioASQ 2016. In *In Proceedings of the BioASQ Workshop, in ACL*.

Isabel Segura-Bedmar, Adrian Carruana, and Paloma Martnez. 2016. LABDA at the 2016 BioASQ challenge task 4a: Semantic Indexing by using ElasticSearch. In *In Proceedings of the BioASQ Workshop, in ACL*.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Mining Multi-label Data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US.

Grigorios Tsoumakas, Manos Laliotis, Nikos Markontanatos, and Ioannis Vlahavas. 2013. Large-Scale Semantic Indexing of Biomedical Publications. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*.

Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, and Eric Nyberg. 2015. Learning to answer biomedical factoid and list questions: Oaqa at bioasq 3b. In *CLEF*.

Zi Yang, Yue Zhou, and Eric Nyberg. 2016. Learning to answer biomedical questions: Oaqa at bioasq 4b. In *In Proceedings of the BioASQ Workshop, in ACL*.

Ilya Zavorin, James Mork, and Dina Demner-Fushman. 2016. Using Learning-To-Rank to Enhance NLM Medical Text Indexer Results. In *In Proceedings of the BioASQ Workshop, in ACL*.