Results of the First BioASQ Workshop

Ioannis Partalas, Eric Gaussier, and Axel-Cyrille Ngonga Ngomo

Abstract. The goal of the BioASQ project 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 bio-medicine. This goal is pursued by the organization of challenges. The first challenge consisted of two tasks: semantic indexing and question answering. 157 systems were registered by 12 different participants for the semantic indexing task, of which between 19 and 29 participated in each batch. The question answering task was tackled by 15 systems, which were developed by three different organizations. Between 2 and 5 of these systems addressed each batch. Overall, the best systems were able to outperform the strong baselines provided in the experiments in two out of three settings. 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 twofold. First, we aim to give an overview of the data issued during the first BioASQ challenge. In addition, we aim to present the systems that participated in the challenge and evaluate their performance w.r.t. to dedicated baseline systems. To this end, we begin by giving a brief overview of the tasks included in the challenge. Especially, we present the setup for the challenge, including the timing of the different tasks and the challenge data. Thereafter, we give an overview of the systems which participated in the challenge. We only provide descriptions for systems that provided us with an overview of the technologies they relied upon. Detailed descriptions of some of the systems are given in workshop 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 the workshop as well as striking findings.

2 Overview of the Tasks

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

Large-scale semantic indexing. In Task 1a the goal is to classify documents from the PubMed[1] digital library unto concepts of the MeSH[2] hierarchy. Here, new PubMed articles that are not yet annotated are collected on a daily 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 [12], and the Lowest Common Ancestor F-measure (LCA-F) from the family of hierarchical measures [4]. For completeness several other flat and hierarchical measures were reported [2]. In order to provide an on-line and large-scale scenario, the task was divided into three independent batches, where in each batch 6 test sets of biomedical articles were released consecutively. Each of these test sets were released in a weekly basis and the participants had 23 hours to provide their answers. Figure 1 gives an overview of the time plan of Task 1a.

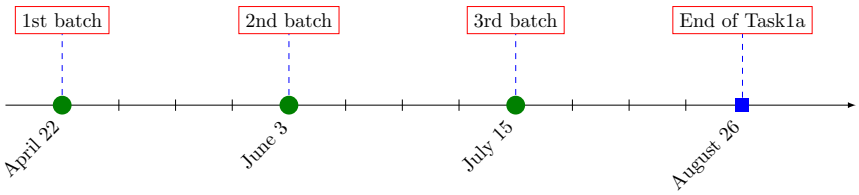

Fig. 1. The time plan of Task 1a.

Biomedical semantic QA. The goal here 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 1b comprised two phases: In phase A, BIOASQ released questions in English from the benchmark datasets and the participants had to respond with concepts (from specific terminologies and ontologies), snippets extracted from PubMed articles and RDF triples (from specific ontologies). In phase B, the released questions contained the correct answers for the elements (concepts, articles, snippets and RDF triples) 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 three 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 measure such as mean precision, mean recall, mean F-measure, mean average precision (MAP)

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

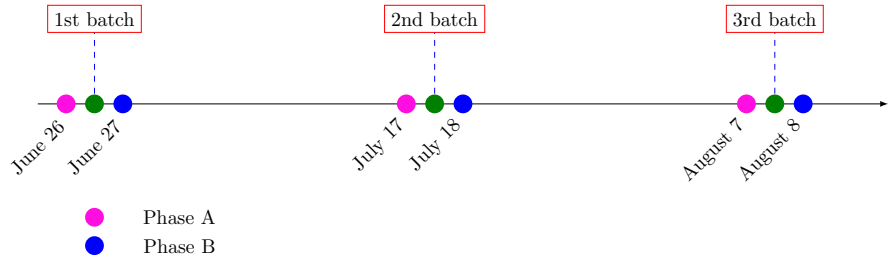

Fig. 2. The time plan of Task 1b. The two phases for each batch run in consecutive dates.

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 was carried out manually by biomedical experts on the ideal answers provided by the systems. For the sake of completeness, ROUGE [5] is also reported.

3 Overview of Participants

3.1 Task 1a

The participating systems in the semantic indexing task of the BioASQ challenge adopted a variety of approaches including hierarchical and flat algorithms as well as search-based approaches that relied on information retrieval techniques. In the rest of section we describe the proposed systems and stress their key characteristics.

In [9] the authors proposed two hierarchical approaches. The first approach, dubbed *Hierarchical Annotation and Categorization Engine* (HACE), follows a top-down hierarchical classification scheme [10] where for each node of the hierarchy a binary classifier is trained. For constructing the positive training examples for each node the authors employ a random method that selects a fixed amount of examples from the descendants of the current node and a method that is based on k-means which chooses the k closest examples to the centroid of the node. In both approaches the selected examples are fixed in order to create manageable datasets especially in the upper levels of the hierarchy. The second system (*Rebayct*) that has participated in the challenge was based on a Bayesian network which models the hierarchical relations among the labels as well as the training data (that is the terms in the abstracts ant titles). A major drawback of this system is that it cannot scale well to large classification problems with thousands of classes and millions of documents. For this reason, the authors reduced the training data to 10% and further split it into 5 disjoint parts in order to train five different models. During the testing phase, the models were aggregated with simple majority voting.

In [13] (*AUTH*) a flat classification approach was employed. This approach trains a binary SVM for each label that is present in the training data [11]. In

order to reduce the complexity of the problem the authors kept only the data that belong to the journals (1806 in total) from which the test sets were sampled during the testing phase of the challenge. The systems that were introduced in the challenge use a meta-model (called MetaLabeler [11]) for predicting the number of labels (N) of a test instance. During the prediction all the SVM classifiers are queried and the labels are sorted according to the corresponding confidence value. Finally, the system predicts the N top labels. While the proposed approach is relative simple it requires processing power for both the training and the testing procedures and also it has large storage requirements (the authors reported that the the size of the models for one of the systems was 406GB).

In [15] the authors follow two different approaches: a) one that relies in the results provided by the MetMap tool [1] and b) one that is based on the search engine Indri[3]. In the MetaMap-based approach, the title and abstract of the article of each test instance is used to query the MetaMap system. The returned results contain concepts and their corresponding confidence scores. The system calculates a final score by assigning weights the concepts that are obtained for the title and the abstract and exceed a predefined threshold for the confidence score. Finally, the system proposes the m top-ranked concepts, where m is a free parameter. In the search-based approach the authors index the training data using the engine Indri. For each test article a query q is generated and a score is calculated for each document d in the index. The concepts of the m top-ranked documents are assigned to the test article.

In the *Wishart* system [6] a typical flat classification approach and k-NN are used. In the flat approach, a binary SVM is trained for each label present in the training data. In the k-NN-based approach, the classifier is invoked for each test article to retrieve documents from a local index. Additionally, the NCBI Entrez system is queried in order to retrieve extra documents along with their labels. All the abstracts are ordered (first N - empirically set to 100) according to their distance and the top M (empirically set to 10) labels are retained. For the final prediction the two systems are combined by keeping the common predicted labels and the rest labels are ordered according to their confidence scores. The system predicts 10-15 labels for each test article.

A learning-to-rank method was used in the NCBI team [7]. The systems follow a three-stage approach: (1) first the k-nearest neighbors of the test article are retrieved from the Medline database, (2) next the labels are ordered using a learning to rank algorithm and (3) finally a cut-off method prunes the ordered list. It is interesting to note that in the definition of the features for the learning to rank problem the authors use the results of the MTIFL baseline system. More specifically, a binary feature indicates whether a specific label observed in the results of MTIFL.

Table 1 resumes the principal technologies that were employed by the participating systems and whether a hierarchical or a flat approach has been followed.

[3] http://www.lemurproject.org/indri.php

Reference	Approach	Technologies
[13]	flat	SVMs, MetaLabeler [11]
[9]	hierarchical	SVMs, Bayes networks
[15]	flat	MetaMap [1], information retrieval, search engines
[6]	flat	k-NN, SVMs
[7]	flat	k-NN, learning-to-rank

Table 1. Technologies used by participants in Task1a.

Baselines. During the first challenge two systems were served as baseline systems. The first one, dubbed BioASQ _Baseline, follows an unsupervised approach to tackle the problem and so it is expected that the systems developed by the participants will outperform it. The second baseline is a state-of-the-art method called Medical Text Indexer [8] which is developed by the National Library of Medicine[4] and serves as a classification system for articles of MEDLINE. MTI is used by curators in order to assist them in the annotation process. It is worth to note also that MTI is used in a few journals to fully automate the process of annotation. So, it is expected to be a hard baseline.

3.2 Task 1b

In the second task of the BioASQ challenge a total of three teams participated in both phases with 11 systems. Only two system descriptions were available when this paper was written[6].

For the phase A of Task 1b the Wishart system [6] invokes query processing and document ranking techniques. More specifically, each test question in natural language form is converted by extracting the noun phrases and reference them using a thesaurus of biomedical entities. Then the question is expanded by adding synonyms and relevant biomedical entities using the PolySearch tool[5]. The entities found by PolySearch are used to rank the retrieved set of concepts, articles, triples and snippets. In phase B of the task a similar approach to phase A is used in order to augment the set of given concepts. Extracted sentences from the retrieved documents are ranked according to the cosine similarity with respect to the augmented concepts. The top-ranked sentences are concatenated in order to provide an *ideal* answer.

The MCTeam system participated only in phase A [15]. In order to form an appropriate query the system first uses the test question to query MetaMap which responds with concept-related words. These words were used to form a query. In case where no concepts were returned by MetaMap, the final query formed by removing the stopwords from the test question. This query was used to retrieve the appropriate information from the BioASQ web services ans also from a local index of PubMed full-text articles[6]. The two lists of the retrieved results were then merged and formed the final results.

[4] http://ii.nlm.nih.gov/MTI/index.shtml
[5] http://wishart.biology.ualberta.ca/polysearch/
[6] The Indri search engine has been used for indexing the documents.

Baselines. Two baselines were used in phase A. The systems return the list of the top-50 and the top-100 entities respectively that may be retrieved using the keywords of the input question as a query to the BioASQ services. As a result, two lists for each of the main entities (concepts, documents, snippets, triples) are produced, of a maximum length of 50 and 100 items respectively.

For the creation of a baseline approach in Task 1B Phase B, three approaches were created that address respectively the answering of factoid and lists questions, summary questions, and yes/no questions [14]. The three approaches were combined into one system, and they constitute the BioASQ baseline for this phase of Task 1B. The baseline approach for the list/factoid questions utilizes and ensembles a set of scoring schemes that attempt to prioritize the concepts that answer the question by assuming that the type of the answer aligns with the lexical answer type (type coercion). The baseline approach for the summary questions introduces a multi-document summarization method using Integer Linear Programming and Support Vector Regression.

4 Results

4.1 Task 1a

During the evaluation phase of the Task1a, the participants submitted their results on a weekly basis to the online evaluation platform of the challenge[7]. The evaluation period was divided into three batches containing 6 test sets each. 11 teams were participated in the task with a total of 40 systems. 10,876,004 articles with 26,563 labels (22GB) were provided as training data to the participants. Table 2 shows the number of articles in each test set of each batch of the challenge.

Table 3 presents the correspondence of the systems for which a description was available and the submitted systems in Task 1a. The systems MTIFL, MTI and BioASQ _Baseline were the baseline systems used throughout the challenge. MTIFL and MTI refer to the NLM Medical Text Indexer system [8]. Systems that participated in less than 4 test sets in each batch are not reported in the results[8].

According to [3] 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 4 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[9]. The best ranked system

[7] `http://bioasq.lip6.fr`

[8] 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.

[9] http://bioasq.lip6.fr/general_information/Task1a/

Batch	Articles	Annotated Articles	Labels per article
1	1,942	1,543	10.00
	845	701	11.56
	793	706	10.87
	2,408	586	10.27
	6,742	4,194	11.70
	4,556	2,503	11.67
Subtotal	17,286	10,233	11.01
2	5,012	1,658	12.39
	5,590	1,658	11.48
	7,349	2,100	12.93
	4,674	1,552	12.37
	8,254	2,556	12.18
	8,626	2,284	13.20
Subtotal	39,505	11,808	12.42
3	7,650	2,002	12.58
	10,233	2,880	13.07
	8,861	2,274	12.44
	1,986	1,118	10.81
	1,750	1,024	10.70
	1,357	530	11.14
Subtotal	31,792	9,828	11.79
Total	88,628	31,869	12.01

Table 2. Statistics on the test datasets of Task1a.

Reference	Systems
[13]	system1, system2, system3, system4, system5
[9]	cole_hce1, cole_hce2, utai_rebayct, utai_rebayct_2
[15]	mc1, mc2, mc3, mc4, mc5
[6]	Wishart-*
[7]	RMAI, RMAIP, RMAIR, RMAIN, RMAIA
Baselines	MTIFL, MTI, bioasq_baseline

Table 3. Correspondence of reference and submitted systems for Task1a.

is highlighted with bold typeface. We can observe that during the first batch the MTIFL baseline achieved the best performance in terms of MiF measure exhibiting a state-of-the-art performance which is also evident in the other two batches. During the first batch RMAIP and system3 have the best performances in both measures. Interestingly, the ranking of the RMAIP according to the LCA-F measure is better than that based on MiF which shows that RMAIP is able to give answers in the neighborhood (as designated by the hierarchical relations among the classes) of the correct ones. In the other two batches the systems proposed in [13] ranked as the best performed ones occupying the first two places (system3 and system2 for the second batch and system1 and system 2 for the third batch). Recall that these systems follow a simple machine-learning approach which uses SVMs and the problem is treated as flat.

We note here the good performance of the learning-to-rank systems (RMAI, RMAIP, RMAIR, RMAIN, RMAIA), which are commonly used in information retrieval tasks.According to the available descriptions the only systems that

made of use of the MeSH hierarchy were the ones introduced by [9]. The top-down hierarchical systems, cole_hce1 and cole_hce2, achieved mediocre results. while the utai_rebayct systems had poor performances. For the systems based on a Bayesian network this behavior was expected as they cannot scale well to large problems. On the other hand the question that arises is whether the use of the MeSH hierarchy can be helpful for classification systems as the labels that are assigned by the curators to the PubMed articles do not follow the rule of the most specialized label. That is, an article may have been assigned a specific label in a deeper level of the hierarchy and in the same time a label in the upper hierarchy that is ancestor of the most specific one.

System	**Batch 1**		**Batch 2**		**Batch 3**	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
MTIFL	**1.25**	**1.75**	2.75	2.75	4.0	4.0
system3	2.75	2.75	**1.0**	**1.0**	2.0	2.0
system2	-	-	1.75	2.0	3.0	3.0
system1	-	-	-	-	**1.0**	**1.0**
MTI	-	-	-	-	3.25	3.0
RMAIP	2.50	**1.75**	5.0	4.5	5.25	5.5
RMAI	3.25	3.0	5.0	4.5	8.5	7.25
RMAIR	6.25	6.0	4.5	3.25	6.25	6.25
RMAIA	5.75	5.5	4.0	5.25	7.25	5.75
RMAIN	4.50	3.25	6.0	5.0	6.5	6.25
Wishart-S3-NP	8.75	9.0	14.25	15.0	-	-
Wishart-S1-KNN	8.75	9.25	12.25	12.5	-	-
Wishart-S5-Ensemble	9.5	8.0	9.50	10.25	-	-
mc4	14.75	14.25	21.0	21.0	21.5	21.25
mc3	11.0	11.25	19.75	19.75	22.0	21.5
mc5	11.25	10.0	15.0	14.75	17.0	17.0
cole_hce2	9.25	9.5	11.25	9.25	12.75	12.0
bioasq_baseline	14.0	14.0	17.75	16.75	20.75	
cole_hce1	13.5	13.5	14.75	14.0	16.0	14.75
mc1	8.75	8.25	13.75	13.25	13.0	13.5
mc2	11.25	11.5	17.75	18.25	14.25	15.75
utai_rebayct	15.5	16.0	16.75	17.5	19.25	21.5
Wishart-S2-IR	9.75	10.75	8.5	9.25	-	-
Wishart-S5-Ngram	-	-	10.5	9.75	-	-
utai_rebayct_2	-	-	-	-	18.25	18.5
TCAM-S1	-	-	-	-	11.25	12.25
TCAM-S2	-	-	-	-	12.25	12.25
TCAM-S3	-	-	-	-	12.5	12.5
TCAM-S4	-	-	-	-	12.0	12.75
TCAM-S5	-	-	-	-	12.75	12.0
FU_System	-	-	-	-	24.0	23.25

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

4.2 Task 1b

Phase A. Table 5 presents the statistics of the training and test data provided to the participants. As in Task 1a the evaluation included three test batches.

For the phase A of Task 1b the systems were allowed to submit responses to any of the corresponding categories, 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 [2]. The final ranking for each batch is calculated as the average of the individual rankings in the different categories. The detailed results for Task 1b phase A can be found in `http://bioasq.lip6.fr/results/1b/phaseA/`.

Batch	Size	# of documents	# of snippets	# of concepts	# of triples
training	29	10.31	14.00	4.82	3.67
1	100	14.89	19.89	8.30	21.87
2	100	14.66	20.24	7.58	5.56
3	82	14.47	17.06	6.24	4.50
total	311	14.28	18.70	7.11	9.00

Table 5. Statistics on the training and test datasets of Task 1b. All the numbers for the documents, snippets, concepts and triples refer to averages.

Table 6 presents the average ranking of each system in each batch of Task 1b phase A. It is evident from the results that the participated systems did not succeed in outperforming the two baselines that were used in phase A. Whether this ineffectiveness can be attributed to the inferior behavior of the participating systems is not clear as they seem to follow intuitive ways to construct the queries. We note also that the systems did not respond to all the categories. For example, the MCTeam systems did not submit snippets throughout the task.

System	Batch 1	Batch 2	Batch 3
Top 100 Baseline	**1.0**	**1.875**	**1.25**
Top 50 Baseline	2.5	2.375	1.75
MCTeamMM	3.625	4.5	3.5
MCTeamMM10	3.625	4.5	3.5
Wishart-S1	4.25	3.875	-
Wishart-S2	-	4.125	-

Table 6. Average ranks for each system for each batch of phase A of Task 1b. The MAP measure were used in order to rank the systems. A hyphenation symbol (-) is used whenever the system did not participate in the corresponding batch.

Focusing on the specific categories, (e.g., concepts) for the Wishart system we observe that it achieves a balanced behavior with respect to the baselines (Table 7). This is evident from the value of F-measure which is much higher that the values of the two baselines. This can be explained on the fact that the Wishart-S1 system responded with short lists while the baselines return always long lists (50 and 100 items respectively). Similar observations hold also for the other two batches.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
Top 100 Baseline	0.080	0.858	0.123	0.472	0.275
Top 50 Baseline	0.121	0.759	0.172	0.458	0.203
Wishart-S1	0.464	0.429	0.366	0.342	0.063
MCTeamMM	0.000	0.000	0.000	0.000	0.000
MCTeamMM10	0.000	0.000	0.000	0.000	0.000

Table 7. Results for batch 1 for concepts in phase A of Task1b.

Phase B. In the phase B of Task 1b 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 [2]. For reasons of completeness we report also the results of the systems for the exact answers. To do so, we average the individual rankings of the systems for the different types of questions, that is Yes/No, factoids and list.

Table 8 presents the average ranks for each system for the exact answers. In this phase we note that the Wishart system was able to outperform the BioASQ baselines.

System	Batch 1	Batch 2	Batch 3
Wishart-S1	**2.0**	**1.0**	-
Wishart-S2	**2.0**	-	-
Wishart-S3	**2.0**	-	-
Baseline1	4.66	**2.33**	**2.33**
Baseline2	4.33	4.0	2.66
main system	6.0	4.33	3.0
system 2	-	5.33	3.33
system 3	-	5.5	3.66
system 4	-	5.5	-

Table 8. Average ranks for each system and each batch of phase B of Task 1b. The final rank is calculated across the individual ranks of the systems for the different types of questions. A hyphenation symbol (-) is used whenever the system did not participate in the corresponding batch.

Table 9 presents the average scores[10] of the biomedical experts for each system across the batches. Note that the scores are between 1 and 5 and the higher it is the better the performance. According to the results the systems were able to provide comprehensible answers and in some cases, like in the second batch, high readable ones. Of course this depends on the difficulty of the question. This seems to be the case in the last batch were the averages scores are lower with respect to the other batches. Also, the calculated measures using ROUGE (the detailed results for Task 1b phase B can be found in `http://bioasq.lip6.fr/results/1b/phaseB/`) seem to be consistent with the

[10] Please consult the description of the evaluation measures used in the challenge for more information .

manual scores in the first two batches while the situation is inverted in the third batch.

System	Batch 1	Batch 2	Batch 3
Wishart-S1	**3.94**	**4.23**	-
Wishart-S2	**3.94**	-	-
Wishart-S3	**3.94**	-	-
Baseline1	2.86	-	**3.19**
Baseline2	2.73	-	3.17
main system	3.35	3.39	3.13
system 2	-	3.34	3.07
system 3	-	3.34	2.98
system 4	-	3.34	-

Table 9. Average scores for each system and each batch of phase B of Task 1b for the ideal answers. The final score is calculated as the average of the individual scores of the systems for the different evaluation criteria. A hyphenation symbol (-) is used whenever the system did not participate in the corresponding batch.

5 Conclusion

A large number of systems participated in Task 1A, the majority of which were able to cope with both the large scale of the problem as well as the on-line evaluation procedure with success. From the results we can draw three major conclusions: First, the majority of the systems were able to achieve good performance, as they were able to outperform the weak baseline throughout the batches. Second, the best systems were able to outperform even the strong baseline (MTIFL), which is the current state of the art for biomedical indexing. This is a very important achievement towards the goal of challenge and the development of accurate classification systems for large-scale problems. Finally, the wide variety of technologies used by the participants allowed us to asses them on a very large-scale scenario. Simple machine-learning approaches (see, e.g., [13]) were shown to achieve state-of-the-art results. Additionally, learning-to-rank approaches followed (see [7]) were shown to be effective for large-scale classification tasks. Interestingly, the hierarchical approach employed in [9] achieved moderate results revealing the fact that the MeSH hierarchy may not be appropriate for classification tasks.

The smaller number of participants in Task 1B and the poor results achieved by these systems suggest that this task is particularly challenging. As the systems seem to follow well principled ways to construct the queries we cannot conclude whether their low performance can be attributed to the use of low-performance methods. Other factors might have played a role, including the retrieval engines underlying the systems not being able to retrieve appropriate responses from the designated resources. Interestingly, a participant was still able to outperform the baselines in phase B (Wishart). The automatic measures that were used to asses the ideal answers seem to be in accordance with the manual scores assigned by

the BioASQ experts in the first two batches of the task while in the third one the measure have different behaviour. This discrepancy will be investigated in future work.

References

1. Alan R. Aronson and Franois-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17:229–236, 2010.
2. Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artieres, and Patrick Gallinari. Evaluation framework specifications. Project deliverable D4.1, 05/2013 2013.
3. Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.
4. Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *CoRR*, abs/1306.6802, 2013.
5. Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop 'Text Summarization Branches Out'*, pages 74–81, Barcelona, Spain, 2004.
6. Yifeng Liu. Bioasq system descriptions (wishart team). Technical report, 2013.
7. Yuqing Mao and Zhiyong Lu. Ncbi at the 2013 bioasq challenge task: Learning to rank for automatic mesh indexing. Technical report, 2013.
8. James Mork, Antonio Jimeno-Yepes, and Alan Aronson. The nlm medical text indexer system for indexing biomedical literature, 2013.
9. Francisco Ribadas, Luis de Campos, Victor Darriba, and Alfonso Romero. Two hierarchical text categorization approaches for bioasq semantic indexing challenge. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.
10. Carlos N. Silla, Jr. and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining Knowledge Discovery*, 22:31–72, 2011.
11. Lei Tang, Suju Rajan, and Vijay K. Narayanan. Large scale multi-label classification via metalabeler. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 211–220, New York, NY, USA, 2009. ACM.
12. Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining Multi-label Data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.
13. Grigorios Tsoumakas, Manos Laliotis, Nikos Markontanatos, and Ioannis Vlahavas. Large-scale semantic indexing of biomedical publications. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.
14. Dirk Weissenborn, George Tsatsaronis, and Michael Schroeder. Answering factoid questions in the biomedical domain. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.
15. Donhqing Zhu, Dingcheng Li, Ben Carterette, and Hongfang Liu. An incemental approach for medline mesh indexing. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.