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

Project **FP7-318652 / BioASQ**

Deliverable **D5.3**

Distribution **Public**

http://www.bioasq.org

Technology Overview Report 2

Ioannis Partalas, Georgios Balikas, Axel-Cyrille Ngonga Ngomo, Makis Malakasiotis, Anastasia Krithara, Eric Gaussier and Georgios Paliouras

Status: Final (Version 1.0)

September 2014

Project

Project ref.no.	FP7-318652
Project acronym	BioASQ
Project full title	A challenge on large-scale biomedical semantic indexing and question answering
Porject site	http://www.bioasq.org
Project start	October 2012
Project duration	2 years
EC Project Officer	Martina Eydner

Deliverable

Deliverabe type	Report
Distribution level	Public
Deliverable Number	D5.3
Deliverable title	Technology Overview Report 2
Contractual date of delivery	M22 (July 2014)
Actual date of delivery	September 2014
Relevant Task(s)	WP5/Task 5.1
Partner Responsible	UJF
Other contributors	NCSR “D”, ULEI, UPMC, AUEB-RC
Number of pages	26
Author(s)	Ioannis Partalas, Georgios Balikas, Axel-Cyrille Ngonga Ngomo, Makis Malakasiotis, Anastasia Krithara, Eric Gaussier and Georgios Paliouras
Internal Reviewers	
Status & version	Final
Keywords	BioASQ, technology overview, results analysis

Executive Summary

This deliverable reviews the systems that participated during the second BioASQ challenge and performs an analysis of the results. More specifically, in the deliverable a short description of each system is given providing also the key technologies that have been used. The objective of this deliverable is to identify the most promising approaches and to point out the progress made with the state-of-the-art.

The challenge comprised two tasks: a) large-scale online biomedical indexing (Task 2a) and b) biomedical semantic QA (Task 2b). Both tasks run in five consecutive batches.

In Task 2a 18 teams participated using 61 registered systems. The systems were evaluated in several performance measures and compared against two baseline systems. Most of them were able to cope with the large scale of the problem while three of them achieved to systematically outperform the state-of-the-art baseline (Medical Text Indexer). A variety of methods have been used like machine learning approaches or search-based ones and hierarchical or flat ones. Specifically, the best systems achieved to enlarge the margin of performance with the MTI system which also this year improved its performance.

In Task 2b 8 teams participated in both phases of the task with a total of 15 systems. In both phases the systems were able to achieve good performances and in most cases to achieve better results than the baselines.

Contents

List of Figures

List of Tables

1 Introduction

This deliverable reviews the systems that participated during the second BioASQ challenge and performs an analysis of the results. More specifically, in the deliverable a short description of each system is given providing also the key technologies that have been used. The objective of this deliverable is to identify the most promising approaches and to point out the progress made with the state-of-the-art.

The remainder of the deliverable is as follows:

- Chapter 1 describes briefly the BioASQ challenge providing also details of the evaluation procedure along with the corresponding time plans. Additionally, for each of the two tasks of the challenge, the total numbers of the participating systems and teams are reported.
- Chapter 2 reviews, for the two tasks, the systems that participated in the challenge. This review is based on the available descriptions provided by the participants. For each system, we present the key points of the proposed methods.
- Chapter 3 presents the results of the evaluation procedure available from the BioASQ evaluation platform[1].
- Chapter 4 presents the prizes awarded to the winners of each task.
- Chapter 5 concludes this deliverable by commenting on the advancement of the state-of-the-art in the biomedical semantic indexing and question answering domain. Also, it discusses the potential impact of the technologies on specialized search engines.

1.1 Challenge Description

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

[1] http://bioasq.lip6.fr

1.1.1 Large-scale semantic indexing

In Task 2a the goal is to classify documents from the PubMed[2] digital library unto concepts of the MeSH[3] 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 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 1.1 gives an overview of the time plan of Task 2a.

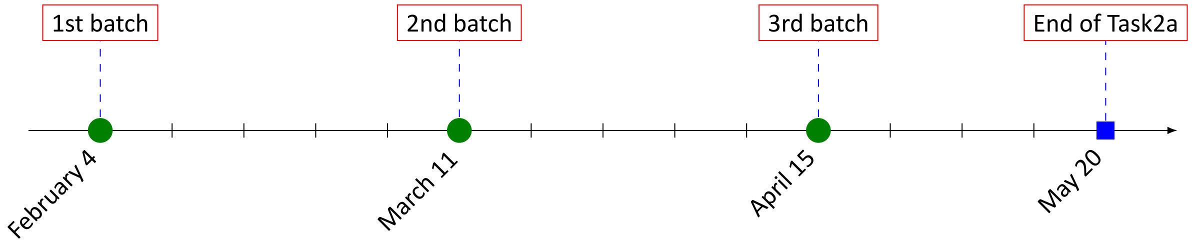

Figure 1.1: The time plan of Task 2a.

1.1.2 Biomedical semantic QA

The goal of task 2b 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 2b 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 and PubMedCentral[4] 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 (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 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 was carried out manually by biomedical experts on the ideal answers provided by the systems. For the sake of completeness, ROUGE (Lin, 2004) is also reported.

[2] http://www.ncbi.nlm.nih.gov/pubmed/
[3] http://www.ncbi.nlm.nih.gov/mesh/
[4] http://www.ncbi.nlm.nih.gov/pmc/

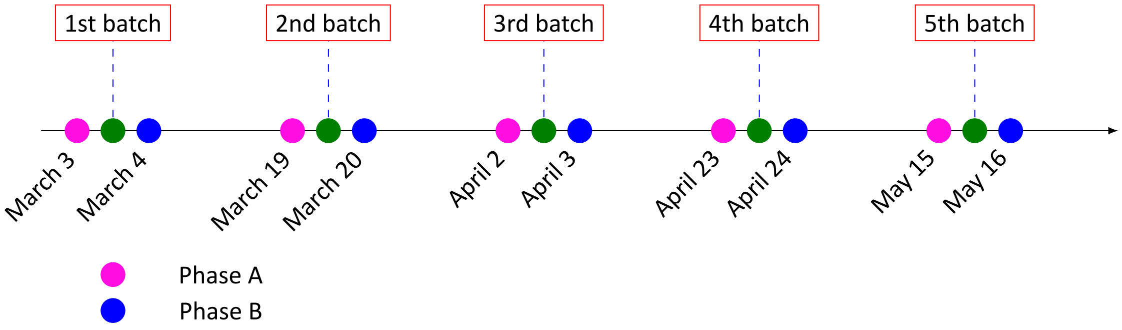

Figure 1.2: The time plan of Task 2b. The two phases for each batch run in consecutive days.

2

Technology Overview

2.1 Task 2a

2.1.1 Background and Related Work

Background. Task 1a deals with the semantic indexing of biomedical documents with concepts from the MeSH hierarchy. Typically the problem is tackled like a classification one where one should build classification models that assign classes from the designated hierarchy to documents. Under this setting, the training set can be represented by $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$. In the context of text classification, $\mathbf{x}^{(i)} \in \mathcal{X}$ denotes the vector representation of the i-th document in the input space $\mathcal{X} \subseteq \mathbb{R}^n$. Assuming that there are K classes denoted by the set $\mathcal{Y} = \{y_1 \ldots y_K\}$, the label $y^{(i)} \in \mathcal{Y}$ represents the class associated with the instance $\mathbf{x}^{(i)}$. In text classification the features (or terms) of the vector representation are the distinct words that occur in the training data. Each element x_k of the vector representation can be either a binary value (0/1), expressing the absence or the presence of the specific word in the document, or a real value calculated by statistical techniques. A simple approach (term frequency) is to calculate the number of occurrences of each word in the document. The most popular scheme is the $tf * idf$ (term-frequency inverse document frequency) where the tf is the term frequency of a specific term t and $idf = \ln \frac{m}{df_t}$ is the logarithm of the number of the documents in the collection divided by the number of documents that contain the term. The idf is a measure of the importance of a specific term in the collection. For example, very common words will have a low idf value. A standard chain for producing the vectors is the following: tokenization, stemming/lemmatization and stop-word removal.

Related work. There have been proposed several approaches for large-scale classification which either leverage the hierarchy information (a simple tree hierarchy is presented in Figure 2.1) by taking into account the parent-child relations among the classes (*hierarchical methods*) or they totally ignore this information (*flat measures*). Hierarchical methods suffer from the fact that the errors made at an upper level of the hierarchy are unrecoverable. On the other hand, flat methods are very slow in terms of training and testing compared to hierarchical methods (Babbar et al., 2013).

Some of the earlier works on exploiting hierarchy among target classes for the purpose of text classification has been studied in (Koller and Sahami, 1997). Parameter smoothing for Naive Bayes classifier along the

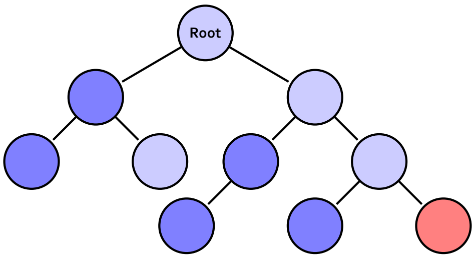

Figure 2.1: A simple tree hierarchy.

root to leaf path was explored by (McCallum et al., 1998). Maximum margin based approaches have been proposed in (Cai and Hofmann, 2004; Dekel et al., 2004), where the degree of penalization in mis-classification depends on the distance between the true and predicted class in the hierarchy tree. However, these approaches were applied to the datasets in which the number of categories were limited to a few hundreds. Liu et al. (2005) applied hierarchical SVM to the scale with over 100,000 categories in Yahoo! directory. More recently, other techniques for large scale hierarchical text classification have been proposed. Prevention of error propagation by applying *Refined Experts* trained on a validation was proposed in (Bennett and Nguyen, 2009). In this approach, bottom-up information propagation is performed by utilizing the output of the lower level classifiers in order to improve the classification of top-level classifiers. Deep Classification (Xue et al., 2008) proposes hierarchy pruning to first identify a much smaller subset of target classes. Prediction of a test instance is then performed by re-training Naive Bayes classifier on the subset of target classes identified from the first step. More recently, Bayesian modelling of large scale hierarchical classification has been proposed by Gopal et al. (2012) in which hierarchical dependencies between the parent-child nodes are modelled by centering the prior of the child node at the parameter values of its parent.

Hierarchy simplification by flattening entire layer in the hierarchy has been studied from an empirical view-point in (Wang and Lu, 2010; Malik, 2009). These strategies for taxonomy adaptation by flattening do not provide any theoretical justification for applying this procedure. Moreover, they offer no clear guidelines regarding which layer in the hierarchy one should flatten. Most of the existing approaches to large scale classification have focussed on the two extremes of flat or hierarchical classification. An approach based on taxonomy embedding has been proposed in (Weinberger and Chapelle, 2009), but this has been restricted to only small scale problems, wherein the target classes are of the order of few hundreds.

Apart from accuracy, other important factors while evaluating the classification strategies for large scale classification are training and prediction speed. The comparison of training time complexity for flat and hierarchical classification in the context of large taxonomies has been studied in (Liu et al., 2005). Learning the hierarchy tree from large number of classes in order to make faster prediction has also attained significance as explored in the recent works such as (Bengio et al., 2010; Beygelzimer et al., 2009; Gao and Koller, 2011). The aim in these approaches is to achieve better prediction speed while maintaining the same classification accuracy as flat classification. On the other end of the specturm are flat classification techniques such as employed in (Perronnin et al., 2012) which ignore the hierarchy structure. These strategies are likely to perform well for balanced hierarchies with sufficient training instances per target class and not so well in large scale taxonomies which suffer from the problem of rare classes.

2.1.2 Systems Overview

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 this section we describe the proposed systems and stress their key characteristics.

The new NCBI system (Yuqing Mao, 2014) for Task 2a is an extension of the work presented in 2013 and relies on the generic learning-to-rank approach presented in (Huang et al., 2011). This approach, differs from the previous approach in the following aspects: a) for each label a binary classifier is trained and the MeSH terms suggested by these classifiers are added in the candidate list of labels, b) the set of documents used as neighbor documents was reduced to documents indexed after 2009. Moreover, the score function for the selection of the number of features was changed from a linear to a logarithmic approach.

In (Papanikolaou et al., 2014) flat classification processes were employed for the semantic indexing task. In particular, the authors trained binary SVM classifiers for each label that was present in the data. In order to reduce the complexity they trained the SVMs in fractions of the data. They trained two systems on different corpus: Asclepios on 950 thousand documents and Hippocrates on 1.5 million documents. Those systems output a ranked lists with labels and a meta-model, namely MetaLabeler (Tang et al., 2009), is used to decide the number of labels that will be submitted for each document. The remaining three systems of the team employ ensemble learning methods. The approach that worked best was a combination of Hippocrates with a model of simple binary SVMs, which were trained by changing the weights parameter for positive instances (Lewis et al., 2004). During the training of a classifier with very few positive instances a false negative is penalized (a positive instance being misclassified) more than a false positive (a negative instance being misclassified). The proposed approaches, although they are relatively simple, require a lot of processing power and memory. For that reason they used a machine with 40 processors and 1TB RAM.

Ribadas-Pena et al. (2014) employ hierarchical models based on a top-down hierarchical classification scheme (Silla and Freitas, 2011) and a Bayesian network which models the hierarchical relations among the labels as well as the training data. The team participated in the first edition of the BioASQ challenge using the same technologies (Ribadas et al., 2013). In the current competition they focused on the pre-processing of the textual data while keeping the same classification models. More specifically, the authors employ techniques for identifying abbreviations in the text and expanding it afterwards in order to enrich the document. Also, a part of speech tagger is used in order to tokenize the text and identify noun, verbs, adjectives and unknown elements (not identified). Finally, a lemmatization step extracts the canonical forms of those words. Additionally, the authors extract word bigrams and keep only those that are identified as multiword terms. The rational is that multiword terms in a domain with complex terminology, like biomedicine, provide higher discriminant power.

In (Choi and Choi, 2014) the authors use a standard flat classification scheme, where a SVM is trained for each class label in MeSH. Different training set methodologies are used resulting in different trained classifiers. Due to computational issues only 50,000 documents were used for training. The selection of the best classification scheme is optimized on the precision at top k labels on a validation set.

In (Liu et al., 2014) the authors used the learning to rank (LTR) method for predicting MeSH headings. However, in addition to the information from similar citations, they also used the prediction scores from individual MeSH classifiers to improve the prediction accuracy. In particular, they trained a binary classifier (logistic regression) for each label in the training data. For a target citation, using the trained classifiers, they calculated the classification probability (score) of every MeSH heading. Then, using NCBI efetch[1],the system retrieves similar documents and their MeSH terms are used as candidate answers. The similarity scores of the target document and the documents retrieved are calculated and averaged over these documents. Finally, these two scores, together with the default results of NLM official solution MTI, were considered as features in the LTR framework. The LambdaMART (Burges, 2010) was used as the ranking method in the learning to rank framework.

Adams and Bedrick (2014) proposed a system which uses Latent Semantic Analysis to identify semantically similar documents in MEDLINE and then constructs a list of MeSH headers from candidates selected

[1] http://www.ncbi.nlm.nih.gov/books/NBK25499/

from the documents most similar to a new abstract.

Table 2.1 resumes the principal technologies that were employed by the participating systems and whether a hierarchical or a flat approach has been followed.

Reference	Approach	Technologies
Papanikolaou et al. (2014)	flat	SVMs, MetaLabeler Tang et al. (2009), Ensemble learning
Ribadas-Pena et al. (2014)	hierarchical	SVMs, Bayes networks
Choi and Choi (2014)	flat	SVMs
Liu et al. (2014)	flat	Logistic regression, learning-to-rank
Adams and Bedrick (2014)	flat	Latent Semantic Analysis
Yuqing Mao (2014)	flat	Learning-to-rank

Table 2.1: Technologies used by participants in Task 2a.

Baselines. During the first challenge, two systems were used as baseline systems. The first one, called BioASQ_Baseline, follows an unsupervised approach to tackle the problem; it is thus expected that the systems developed by the participants will outperform it. More specifically, the baseline implements Attribute Alignment Annotator (Doms, 2010). It is an unsupervised method, based on the Smith-Waterman sequence alignment algorithm (Smith and Waterman, 1981) and can recognizes terms from MeSH and Gene Ontology in a given text passage. The annotator first pre-processes both the ontology terms and the text by tokenizing them, removing the stop words and stemming the remaining terms (an in-house stop word list that is specific to the domain is used). Then the term stems are mapped onto the text stems using the local sequence alignment algorithms (Smith and Waterman, 1981). Insertions, deletions and gaps are penalized. The information value of terms calculated over the whole ontology is also taken into account during the alignment process, in a similar manner as the inverse document frequency score is used for the tf-idf weighting of terms.

The second baseline is a state-of-the-art method called Medical Text Indexer (James G. Mork, 2014) which is developed by the National Library of Medicine[2] 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 new annotator is an extension of the system presented in (Mork et al., 2013) with the approaches of the last year's winner (Tsoumakas et al., 2013). Consequently, we expected the baseline to difficult to beat.

2.2 Task 2b

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 8 teams with 22 systems participated. In the second phase, where team are requested to submit exact and paragraph-sized answers for the questions, 7 teams with 18 different systems participated.

The system presented in (Neves, 2014) relies on the Hana Database for text processing. It uses the Stanford CoreNLP package for tokenizing the questions. Each of the token is then sent to the BioPortal and to the Hana database for concept retrieval. The concepts retrieved from the two systems are finally merged to a single list that is used to retrieve relevant text passages from the documents at hand. To this end, four different types of queries are sent to the BioASQ services. Overall, the approach achieves between 0.18 and 0.23 F-measure.

In phase A, NCBI's framework (Yuqing Mao, 2014) used the cosine similarity between question and sentence to compute their similarity. The best scoring sentence from an abstract was chosen as relevant snippet for an answer. Concept retrieval was achieved by a customized dictionary lookup algorithm in combination

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

with MetaMap. For phase B, tailored approaches were used depending on the question types. For example, a manual set of rules was crafted to determine the answers to factoid and list questions based on the benchmark data for 2013. The system achieved an F-measure of up to betwen 0.2% (RDf triples) and 38.48% (concepts). It performed very well on Yes/No questions (up to 100% accuracy). Factoid and list questions led to an MRR of up to 20.57%.

In (Choi and Choi, 2014) the authors participated only in the document retrieval of phase A and in the generation of ideal answers in phase B. The Indri search engine is used to index the PubMed articles and different models are used to retrieve documents like pseudo-relevance feedback, sequential dependence model and semantic concept-enriched dependence model where the retrieved UMLS concepts in the query are used as additional dependence features for ranking documents. For the generation of ideal answers the authors retrieve sentences from documents and identify the common keywords. Then the sentences are ranked according to the number of times these keywords appear in each of them and finally the top ranked m are used to form the ideal answer. Despite the simplicity of the approach it achieves to perform well in both documents and ideal answers.

The authors of (Lingeman and Dietz, 2014) propose a method for the retrieval of relevant documents and snippets of task 2b. They develop a figure-inspired text retrieval method as a way of retrieving documents and text passages from biomedical publications. The method is based on the insight that for biomedical publications, the figures play an important role to the point that the captions can be used to provide abstract like summaries. The proposed approach uses an Information Retrieval perspective on the problem. In principle, the followed steps are: (i) the question is enriched by query expansion with information from UMLS, Wikipedia, and Figures, (ii) a ranking of full documents and snippets is retrieved from a corpus of PubMed Central Articles which is the set of full-text available articles, (iii) features are extracted for each document and snippet that provide proof of its relevance for the question and (iv) the documents/snippets are re-ranked with a learning-to-rank approach.

In the context of phase B of task 2b in (Papanikolaou et al., 2014), the authors attempted to replicate the work that already exists in literature and was presented in the BioASQ 2013 workshop (Weissenborn et al., 2013). They provided exact answers only for the factoid questions. Their 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 (Aronson and Lang, 2010) in order to extract candidate answers.

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 2B Phase B, three approaches were created that address respectively the answering of factoid and lists questions, summary questions, and yes/no questions (Weissenborn et al., 2013). The three approaches were combined into one system, and they constitute the BioASQ baseline for this phase of Task 2B. 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.

3

Setup and Results

3.1 Task 2a

3.1.1 Data and Setup

During the evaluation phase of the Task 2a, the participants submitted their results on a weekly basis to the online evaluation platform of the challenge[1]. The evaluation period was divided into three batches containing 5 test sets each. 18 teams were participated in the task with a total of 61 systems. 12,628,968 articles with 26,831 labels (20.31GB) were provided as training data to the participants. A reduced training dataset was also provided to the participants containing only the articles from the journals that the test sets are drawn. This dataset contained 4,458,300 documents using 26,631 MeSH terms. Figure 3.1 presents the category size distribution of this dataset. We can observe that a lot of categories have a few documents which is typical in large taxonomies (Yang et al., 2003; Babbar et al., 2014). Table 3.1 presents basic statistics on the provided training data.

	Training set 2013	Training set 2014	Reduced tr. set 2014
# of articles	10,876,004	12,628,968	4,458,300
Avrg. labels/article	12.55	12.72	13.20
MeSH labels	26,563	26,831	26,631
Size zip/unzip (raw)	5.1Gb/18Gb	6.2G/20.31Gb	1.9Gb/6.4Gb
Size zip/unzip (Lucene)	4.8Gb/6.2Gb	4.4G/6.2Gb	1.3Gb/1.9Gb

Table 3.1: Statistics of the training data provided to the participants for Task 2A. We also provide the statistics for the data of the first edition of the BIOASQ competition.

Table 3.2 shows the number of articles in each test set of each batch of the challenge. The articles were provided to the participants in their raw format (plain text) as well as in a pre-processed one (in a vector-

[1] http://bioasq.lip6.fr

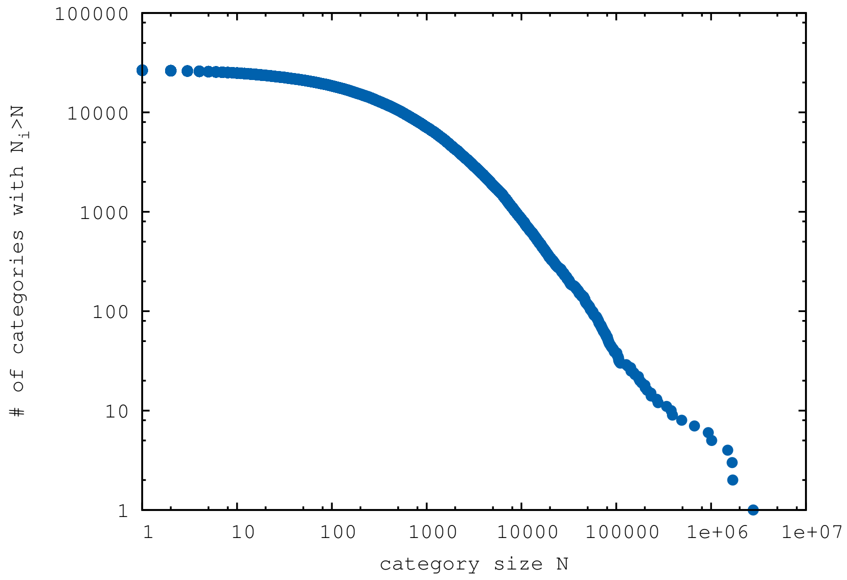

Figure 3.1: Category size vs. rank distribution for the training data in Task 2A.

ized format) under the Apache Lucene framework[2]. Lucene is an open-source library[3] dedicated to text search.Figure 3.2 presents an example of two articles extracted from the BIOASQ benchmark training data.

Batch	Articles	Annotated Articles	Labels per article
1	4,440	3,263	13.20
	4,721	3,716	13.13
	4,802	3,783	13.32
	3,579	2,341	13.02
	5,299	3,619	13.07
Subtotal	23,321	16,722	13.15
2	4,085	3,322	13.05
	3,496	2,752	12.28
	4,524	3,265	12.90
	5,407	3,848	13.23
	5,454	3,642	13.58
Subtotal	22,966	16,829	13.01
3	4,342	2,996	12.71
	8,840	5,783	13.37
	3,702	2,737	13.32
	4,726	3,225	13.90
	4,533	3,196	12.70
Subtotal	26,143	17,929	13.20
Total	72,430	51,480	13.12

Table 3.2: Statistics on the test datasets of Task 2a. The datasets were updated the 29th of June 2014.

Table 3.3 presents the correspondence of the systems for which a description was available and the submitted systems in Task 2a. The systems MTIFL, MTI-Default and BioASQ_Baseline were the baseline systems used throughout the challenge. MTIFL and MTI-Default refer to the NLM Medical Text Indexer system (James

[2] http://lucene.apache.org/

[3] Under the Apache Licence: http://www.apache.org/licenses/LICENSE-2.0.html

```
{
    ”abstractText”:”From the above it is seen that the [...]
    scientific guidance of which lies wholly
     in the hands of scientists.”,
    ”journal”:”Science (New York, N.Y.)”,
    ”meshMajor”:[”Biomedical Research”],
    ”pmid”:”17772322”,
    ”title”:”New Horizons in Medical Research.”,
    ”year”:”1946”
},
{
    ”abstractText”:”1. T antigens of group A hemolytic
     streptococci have been [...] T antigen in the intact
      streptococcus from which it was derived.”,
    ”journal”:”The Journal of experimental medicine”,
    ”meshMajor”:[”Antibodies”,”Antigens”,
    ”Immunity”,”Streptococcal Infections”,”Streptococcus”],
    ”pmid”:”19871581”,
    ”title”:”THE PROPERTIES OF T ANTIGENS EXTRACTED
    FROM GROUP A HEMOLYTIC STREPTOCOCCI.”,
    ”year”:”1946”
}
```

Figure 3.2: An extract from the training data of Task2a.

Reference	Systems
Papanikolaou et al. (2014)	Asclepius, Hippocrates, Sisyphus
Ribadas-Pena et al. (2014)	cole_hce1, cole_hce2, cole_hce_ne, utai_rebayct, utai_rebayct_2
Choi and Choi (2014)	SNUMedInfo*
Liu et al. (2014)	Antinomyra-*
Yuqing Mao (2014)	L2R*
Baselines	MTIFL, MTI-Default, bioasq_baseline

Table 3.3: Correspondence of reference and submitted systems for Task 2a.

G. Mork, 2014). Systems that participated in less than 4 test sets in each batch are not reported in the results[4].

Figure 3.3 presents the MiF measure for the best system in each test test against the MTI baseline as well as the average performance of all the systems participated in the task. For comparison reasons we also report the corresponding performances for last year competition (Task 1a). Interestingly, we first notice that the MTI baseline achieves a performance similar to that of the best system in last year's task. This is due to the accommodation of several features in the MTI baseline system from last year's top performed system which shows the impact of the technologies presented in the BIOASQ competition. Secondly, we observe clearly that this year the best system achieves a far better performance than the MTI baseline with the gap growing at the test sets. Finally, the average performance of the systems has also been improved which is an indication of the quality of the submitted systems this year. We observer a similar trend for the LCA-F

[4] 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.

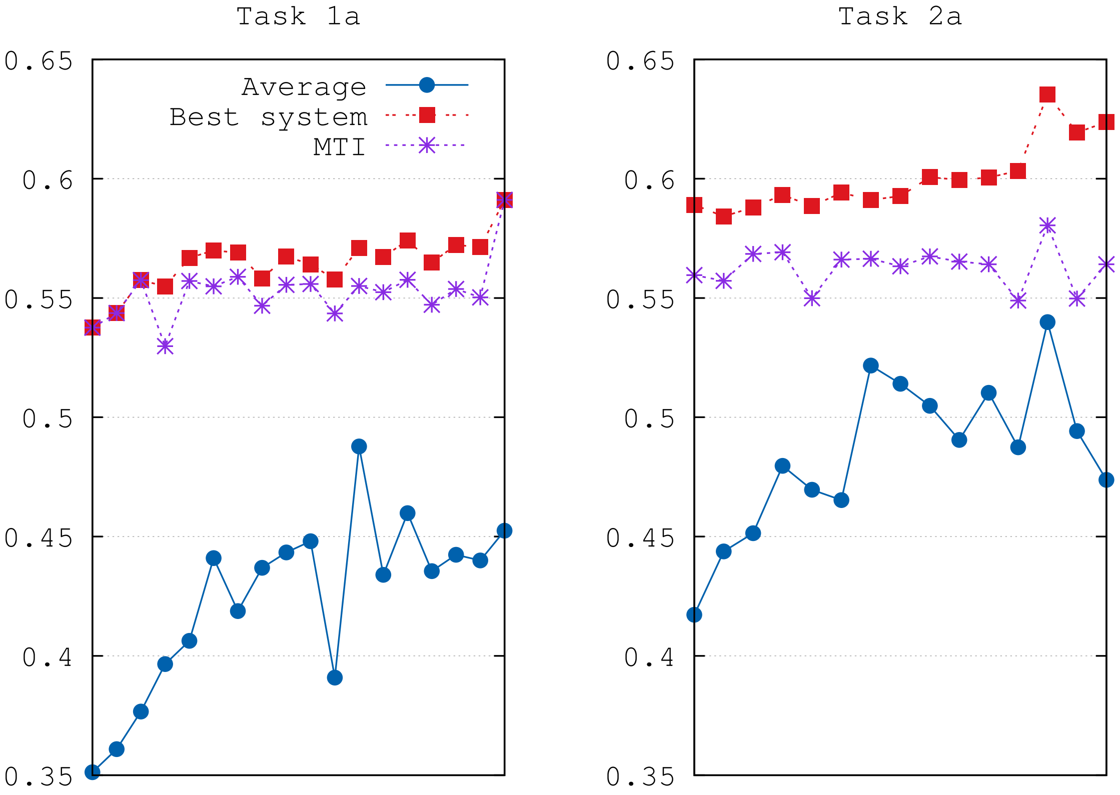

Figure 3.3: Comparison of the MiF measure for the best system in each test set against the MTI baseline and the average performance of all the systems participated in the task. The results for both versions (Task 1a and Task 2a) of the semantic indexing task are presented.

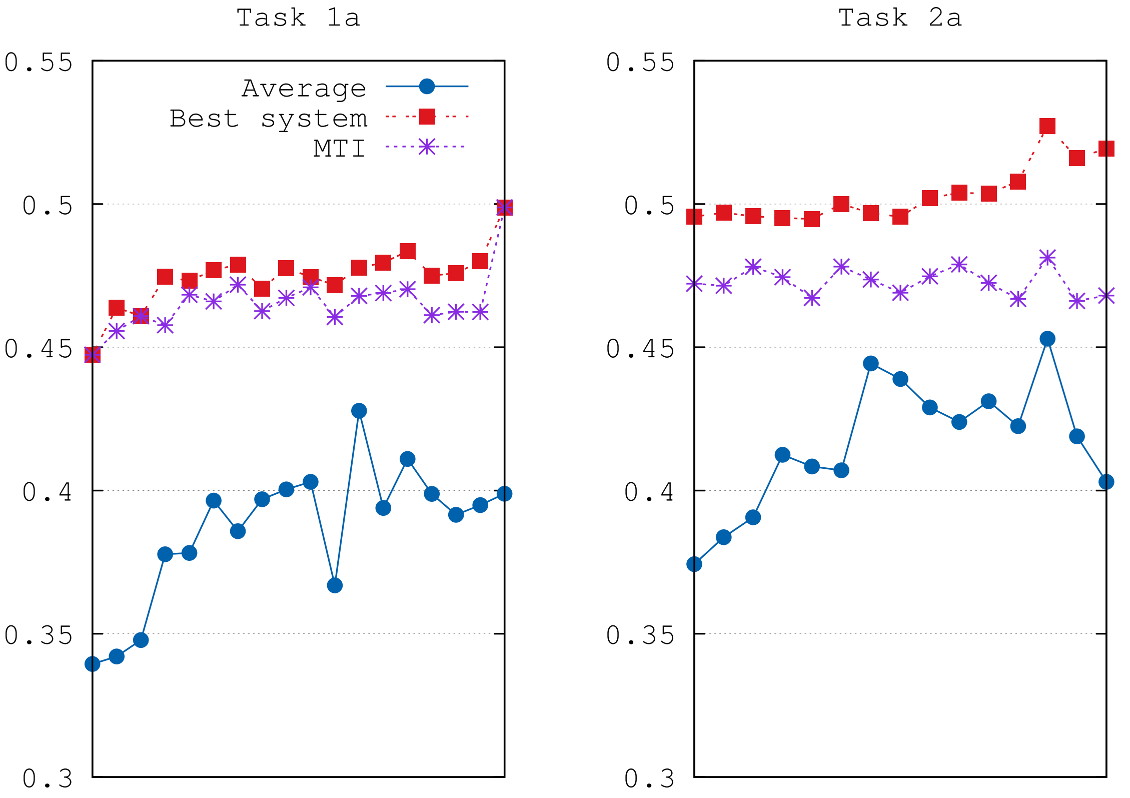

Figure 3.4: Comparison of the MiF measure for the best system in each test set against the MTI baseline and the average performance of all the systems participated in the task. The results for both versions (Task 1a and Task 2a) of the semantic indexing task are presented.

measure which is depicted in Figure 3.4.

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. Table 3.1.1 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[5]. The best ranked system at each batch and at each evaluation measure is highlighted with bold typeface.

First, we can observe that several systems outperforms the strong MTI baseline in terms of MiF and LCA measures exhibiting state-of-the-art performances. During the first batch the flat classification approach (Asclepius system) used in (Papanikolaou et al., 2014) tops the performance in the case of the MiF measure. This system follows a flat approach with linear complexity in the number of classes. Thus it requires large inference time in large-scale scenarios.

In the other two batches the learning-to-rank systems proposed by NCBI (L2R systems) and the Fudan University (Antinomyra systems) ranked as the best performed ones occupying the first two places in both measures. The Fudan team achieves the best performance in both batches for both evaluation measures. Note, that this systems uses the confidence values of the classifiers trained for each MeSH label as features

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

in the meta-learning problem.

We performed statistical tests among the best systems of each team in order to detect significant differences. More specifically, for both evaluation measures we performed a micro sign test (s-test) as proposed in (Yang and Liu, 1999) for each pair of the top systems. In all cases the tests reported significant differences for p-value$<$0.01.

According to the available descriptions the only systems that made of use of the MeSH hierarchy were the ones introduced by Ribadas et al. (2013). The top-down hierarchical systems, cole_hce1, cole_hce2 and cole_hce_ne achieved mediocre results. while the utai_rebayct systems had poor performances. On the other hand hierarchical systems are much faster in inference time than flat ones making them appealing for large-scale problems. For the systems based on a Bayesian network this behaviour 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. In this case the system that predicted the more specific label will be punished by the flat evaluation measures for not predicting the most general label, which is implied by the hierarchical relations.

System	Batch 1		Batch 2		Batch 3	
	MiF	LCA-F	MiF	LCA-F	MiF	LCA-F
Asclepius	**1.0**	3.25	7.75	7.75	-	-
L2R-n1	3.0	3.25	5.75	3.75	8.0	5.75
L2R-n5	4.25	5.75	4.5	4.5	7.75	8.75
L2R-n3	4.25	2.25	4.75	6.75	7.25	7.0
L2R-n2	2.75	**1.5**	4.75	4.0	6.0	4.25
L2R-n4	4.25	5.25	6.0	**3.5**	8.5	7.75
FU_System_t25	13.5	13.25	20.0	18.75	-	-
MTIFL	8.0	8.0	18.25	20.5	15.25	15.25
MTI-Default	6.25	5.5	13.0	10.75	14.25	14.25
FDU_MeSHIndexing_3	-	-	16.0	16.25	-	
FU_System_k25	15.75	15.25	19.75	19.25	-	-
FU_System_k15	15.50	13.75	17.75	15.0	-	-
FU_System_t15	14.50	13.0	19.5	17.75	-	-
Antinomyra0	-	-	**3.0**	**3.5**	**1.75**	5.0
Antinomyra1	-	-	8.75	7.75	2.0	3.25
Antinomyra3	9.50	12.25	5.0	5.25	3.5	**1.75**
Antinomyra2	-	-	6.0	7.25	2.0	2.5
Antinomyra4	12.75	14.0	8.5	7.25	4.25	3.25
FU_System	18.50	16.75	15.75	16.0	-	-
FDU_MeSHIndexing_1	-	-	14.25	13.75	-	-
FDU_MeSHIndexing_2	-	-	15.75	14.75	-	-
Micro	21.75	22.75	24.0	27.5	23.25	28.0
PerExample	21.75	21.75	26.5	26.5	25.25	26.0
Sisyphus	-	-	6.25	12.25	10.5	12.75
Hippocrates	-	-	6.2	6.75	11.5	9.5
Macro	25.00	24.5	32.75	30.75	32.25	30.5
Spoon	21.25	20.75	34.0	33.75	-	-
Accuracy	-	-	34.0	33.25	33.25	37.25
Fork	21.75	22.25	36.25	37.75	-	-
Spork	23.00	23.25	37.25	38.75	-	-
bioasq_baseline	23.75	23.25	39.5	36.0	36.75	34.25
ESIS1	-	-	35.75	34.25	18.0	18.5
ESIS	-	-	36.75	35.75	23.75	25.75
ESIS2	-	-	26.75	27.0	19.25	19.75
ESIS3	-	-	-	-	20.25	18.5
ESIS4	-	-	-	-	20.5	22.25
cole_hce1	-	-	24.5	23.75	25.5	20.25
cole_hce_ne	-	-	26.5	25.25	26.75	22.5
cole_hce2	-	-	27.25	25.75	28.0	22.25
SNUMedinfo3	-	-	32.0	33.5	19.5	24.75
SNUMedinfo4	-	-	32.75	32.0	21.75	23.5
SNUMedinfo1	-	-	33.50	34.75	25.25	28.0
SNUMedinfo5	-	-	33.75	32.75	20.5	22.5
SNUMedinfo2	-	-	34.25	35.5	19.75	23.75
utai_rebayct	-	-	38.50	38.75	34.75	34.25
utai_rebayct_2	-	-	36.50	34.75	31.75	28.5
vanessa-0	-	-	-	-	27.75	25.0
larissa-0	-	-	-	-	37.0	36.5

Table 3.4: 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.

3.2 Task 2b

3.2.1 Phase A

Table 3.5 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 2b 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 detailed results for Task 2b phase A can be found in `http://bioasq.lip6.fr/results/2b/phaseA/`.

Batch	Size	# of documents	# of snippets	# of concepts	# of triples
training	310	14.28	18.70	7.11	9.00
1	100	7.89	9.64	6.50	24.48
2	100	11.69	14.71	4.24	204.85
3	100	8.66	10.80	5.09	354.44
4	100	12.25	14.58	5.18	58.70
5	100	11.07	13.18	5.07	271.68
total	810	11.83	14.92	5.93	116.30[6]

Table 3.5: Statistics on the training and test datasets of Task 2b. All the numbers for the documents, snippets, concepts and triples refer to averages.

As only partial results are available (the golden data are revised by the experts considering the answers of the systems) in the following we present results of specific categories like concepts and documents.

Focusing on the specific categories, (e.g., concepts or documents) for the Wishart system we observe that it achieves a balanced behaviour with respect to the baselines (Table 3.7 and Table 3.6). 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 four batches, the results of which are available online.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
SNUMedinfo1	0.0457	0.5958	0.0826	0.2612	0.0520
SNUMedinfo3	0.0457	0.5947	0.0826	0.2587	0.0501
SNUMedinfo2	0.0451	0.5862	0.0815	0.2547	0.0461
SNUMedinfo4	0.0457	0.5941	0.0826	0.2493	0.0468
SNUMedinfo5	0.0459	0.5947	0.0829	0.2410	0.0449
Top 100 Baseline	0.2274	0.4342	0.2280	0.1911	0.0070
Top 50 Baseline	0.2290	0.3998	0.2296	0.1888	0.0059
main system	0.0413	0.2625	0.0678	0.1168	0.0015
Biomedical Text Ming	0.2279	0.2068	0.1665	0.1101	0.0014
Wishart-S2	0.1040	0.1210	0.0793	0.0591	0.0002
Wishart-S1	0.1121	0.1077	0.0806	0.0535	0.0002
UMass-irSDM	0.0185	0.0499	0.0250	0.0256	0.0001
Doc-Figdoc-UMLS	0.0185	0.0499	0.0250	0.0054	0.0001
All-Figdoc-UMLS	0.0185	0.0499	0.0250	0.0047	0.0001
All-Figdoc	0.0175	0.0474	0.0236	0.0043	0.0001

Table 3.6: Results for batch 1 for documents in phase A of Task2b.

System	Mean Precision	Mean Recall	Mean F-measure	MAP	GMAP
Wishart-S1	0.4759	0.5421	0.4495	0.6752	0.1863
Wishart-S2	0.4759	0.5421	0.4495	0.6752	0.1863
Top 100 Baseline	0.0523	0.8728	0.0932	0.5434	0.3657
Top 50 Baseline	0.0873	0.8269	0.1481	0.5389	0.3308
main system	0.4062	0.5593	0.4018	0.4006	0.1132
Biomedical Text Ming	0.1250	0.0929	0.0950	0.0368	0.0002

Table 3.7: Results for batch 1 for concepts in phase A of Task2b.

3.2.2 Phase B

In the phase B of Task 2b 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). For reasons of completeness we report also the results of the systems for the exact answers.

Table 3.8 shows the results for the exact answers for the first batch of task 2a. In case that systems 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://bioasq.lip6.fr/results/2b/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. Again, the system of Wishart (Wishart-S3) for example shows consistent performance as it manages to answer relatively well in all kinds of questions.

System	Yes/no		Factoid			List	
	Accuracy	Strict Acc.	Lenient Acc.	MRR	Precision	Recall	F-measure
Biomedical Text Ming	0.9375	0.1852	0.1852	0.1852	0.0618	0.0929	0.0723
system 2	0.9375	0.0370	0.1481	0.0926	-	-	-
system 3	0.9375	0.0370	0.1481	0.0926	-	-	-
Wishart-S3	0.8438	0.4074	0.4444	0.4259	0.4836	0.3619	0.3796
Wishart-S2	0.8438	0.4074	0.4444	0.4259	0.5156	0.3619	0.3912
main system	0.5938	0.0370	0.1481	0.0926	-	-	-
BioASQ_Baseline	0.5313	-	-	-	0.0351	0.0844	0.0454
BioASQ Baseline 2	0.5000	-	-	-	0.0351	0.0844	0.0454

Table 3.8: Results for batch 1 for exact answers in phase B of Task2b.

Table 3.9 presents the results in terms of the Rouge evaluation measures for ideal answers for the first batch of phase B for the Task 2B. According to the results, the systems were able to provide comprehensible answers, and in some cases like in the second batch, highly readable ones. Table 3.10 presents such an example for two questions for the SNUMedInfo1.

System	Rouge-2	Rouge-SU4
SNUMedinfo1	0.1529	0.1451
SNUMedinfo2	0.1497	0.1402
Biomedical Text Ming	0.1460	0.1476
SNUMedinfo4	0.1368	0.1286
Wishart-S3	0.1215	0.1132
Wishart-S2	0.1215	0.1132
SNUMedinfo3	0.1200	0.1097
SNUMedinfo5	0.1122	0.1035
system 2	0.0967	0.0884
system 3	0.0966	0.0883
main system	0.0965	0.0883
BioASQ_Baseline 2	0.0458	0.0466
BioASQ_Baseline	0.0449	0.0441

Table 3.9: Results for batch 1 for ideal answers in phase B of Task2b.

SNUMedInfo1	Golden answer
Overexpression of sirtuins (NAD(+)-dependent protein deacetylases) has been reported to increase lifespan in budding yeast (Saccharomyces cerevisiae)	Overexpression of sirtuins (NAD(+)-dependent protein deacetylases) has been reported to increase lifespan in budding yeast (Saccharomyces cerevisiae).
Catecholaminergic polymorphic ventricular tachycardia (CPVT) is a rare arrythmogenic disease characterized by exercise–or stress–induced ventricular tachyarrythmias, syncope, or sudden death, usually in the pediatric age group. Familial occurrence has been noted in about 30% of cases. Inheritance may be autosomal dominant or recessive, usually with high penetrance. The causative genes have been mapped to chromosome 1. Mutations of the cardiac ryanodine rece ptor gene (RyR2) have been identified in autosomal dominant pedigrees, while calsequestrin gene (CASQ2) mutations are seen in recessive cases. Several mutations in the genes encoding RyR1 and RyR2 have been identified in autosomal dominant diseases of skeletal and cardiac muscle, such as malignant hyperthermia (MH), central core disease (CCD), catecholaminergic polymorphic ventricular tachycardia (CPVT), and arrhythmogenic right ventricular dysplasia type 2 (ARVD2).	Autosomal dominant catecholaminergic polymorphic ventricular tachycardia (CPVT) was mapped to chromosome 1q42-43 with identificatio n of pathogenic mutations in RYR2.

Table 3.10: The ideal answers returned for two questions from the system SNUMedInfo along with the golden ones.

4

Prizes

Tables 4.1 presents the prizes that were awarded to the winners for Task 2A.

		1st place	**prize** (euros)	**2nd place**	**prize** (euros)
Batch 1	MiF	Auth	650	NCBI	350
	LCA-F	NCBI	650	Auth	350
Batch 2	MiF	Fudan	650	NCBI	350
	LCA-F	Fudan & NCBI	500 & 500	-	-
Batch 3	MiF	Fudan	650	NCBI	350
	LCA-F	Fudan	650	NCBI	350

Table 4.1: Prizes awarded for Task 2A.

The members of each team for task 2a were the following:

- **Auth**: Yannis Papanikolaou, Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, Ioannis Vlahavas
- **NCBI**: Yuqing Mao Chih-Hsuan Wei, Zhiyong Lu
- **Fudan**: Ke Liu, Junqiu Wu, Shengwen Peng, Chengxiang Zhai, Shanfeng Zhu

Tables 4.2 and 4.3 present the prizes that were awarded to the winners for Task 2B.
The members of each team for task 2b were the following:

- **Auth**: Dimitrios Dimitriadis, Grigorios Tsoumakas, Manos Laliotis, Nikos Markantonatos, Ioannis Vlahavas
- **NCBI**: Yuqing Mao Chih-Hsuan Wei, Zhiyong Lu
- **TTI**: Kota Makise, Yutaka Sasaki

		1st place	prize (euros)	2nd place	prize (euros)
Batch 1	Documents	SNU	200	TTI	100
	Concepts	ALBERTA	200	TTI	100
Batch 2	Documents	SNU	200	Fudan	100
	Concepts	ALBERTA	200	TTI	100
Batch 3	Documents	SNU	200	Fudan	100
	Concepts	ALBERTA	200	Fudan	100
Batch 4	Documents	SNU	200	NCBI	100
	Concepts	ALBERTA	200	NCBI	100
Batch 5	Documents	SNU	200	NCBI	100
	Concepts	ALBERTA	200	NCBI	100

Table 4.2: Prizes awarded for Task 2B -Phase A

		1st place	prize (euros)	2nd place	prize (euros)
Batch 1	Exact answer	ALBERTA	200	NCBI	100
	Ideal answer	SNU	200	NCBI	100
Batch 2	Exact answer	ALBERTA	200	NCBI	100
	Ideal answer	NCBI	200	SNU	100
Batch 3	Exact answer	ALBERTA	200	NCBI	100
	Ideal answer	SNU	200	NCBI	100
Batch 4	Exact answer	ALBERTA	200	AUTH	100
	Ideal answer	SNU	200	NCBI	100
Batch 5	Exact answer	NCBI	200	AUTH	100
	Ideal answer	NCBI	200	SNU	100

Table 4.3: Prizes awarded for Task 2B -Phase B

- **SNU**: Sungbin Choi, Jinwook Choi
- **ALBERTA**: Yifeng Liu
- **Fudan**: Beichen Wang, Shanfeng Zhu

5

Conclusions and Potential Impact

5.1 Task 2a

In the first task of BioASQ a large number of teams participated submitting a large number of systems. The majority of the systems were able to successfully cope with both the large scale of the problem as well as the on-line evaluation procedure. From the results, we can draw three main conclusions:

- The majority of the systems were able to achieve good performance being able to outperform the weak baseline throughout the batches. Interestingly, the average performance of the systems has greatly improved indicating that more high performance systems have participated in the competition.
- The best systems were able to outperform the strong baseline (MTI), thus pushing the state-of-the-art. More specifically, the systems achieved to enlarge the performance gap with the MTI baseline with respect to last year's results. We regard this as a very important achievement towards the goal of developing accurate classification systems for large-scale problems.
- A variety of methods have been used by the participants like pure machine learning approaches, search-based approaches and learning-to-rank approaches. The different technologies that were used allowed us to asses them on a very large-scale scenario. More specifically, the learning-to-rank approaches followed in (Liu et al., 2014; Yuqing Mao, 2014) showed that such systems can be effective for large-scale classification tasks. Also, even the hierarchical approach employed by Ribadas-Pena et al. (2014) achieved moderate results the low complexity of such approaches make them appealing for large-scale scenarios.

5.2 Task 2b

In the second task the participation has increased with respect to the first edition of the BioASQ challenge. In phase A the participating systems were able in most cases to outperform the baselines and they were to achieve good results indicating a participation of high quality systems.

Concerning phase B of the task, the participating systems were also able to obtain better performance than that of the baselines and provide comprehensible ideal answers.

5.3 Potential Impact of New Technologies

Firstly, we would like to point out the fact that this year's baseline system of MTI incorporated features from the best performed system in the first edition of BIOASQ competition (Tsoumakas et al., 2013; James G. Mork, 2014; Partalas et al., 2013). This resulted to an increase in the performance of the MTI system reflecting the impact of the technologies presented in the BIOASQ challenges in the state-of-art systems.

The top rated systems which were able to improve substantially over the MTI baseline follow different approaches. The first ranked systems followed a hybrid approach mixing an information retrieval phase and a learning-to-rank procedure (Liu et al., 2014; Yuqing Mao, 2014). The second best rated systems presented in (Papanikolaou et al., 2014) followed a pure machine learning approach employing flat classification schemes using SVMs and combining several systems with ensemble methods. Also the hierarchical approaches that presented in the competition achieved good results having low complexity due to the use of the hierarchical structure. While the former approaches are able to provide better results the latter enjoy faster training and inference times (very crucial for on-line search engines like GoPubMed). So, potentially both technologies could be used in order to boost the prediction capabilities of a search engine where the first can be employed in an off-line scenario for improving the annotations of the articles in the database.

The technologies of the learning-to-rank systems can be integrated in the front-end of the search engine in order to provide accurate and fast results to the users. In addition, the approaches developed and submitted in the framework of Task 1b, may be used as a basis to develop Q&A expansions of GoPubMed. Based on this observation, GoPubMed could be among the first search engines to launch a fully fledged Q&A for the biomedical domain in the search engine market. More details on the potential impact of the proposed approaches in BIOASQ challenges will be presented in the corresponding deliverable.

Bibliography

J. R. Adams and S. Bedrick. Automatic classification of pubmed abstracts with latent semantic indexing: Working notes. In *Proceedings of Question Answering Lab at CLEF*, 2014.

A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17:229–236, 2010.

R. Babbar, I. Partalas, E. Gaussier, and M.-R. Amini. On flat versus hierarchical classification in large-scale taxonomies. In *NIPS*, sep 2013.

R. Babbar, C. Metzig, I. Partalas, E. Gaussier, and M.-R. Amini. On power law distributions in large-scale taxonomies. *SIGKDD Explorations*, 2014.

G. Balikas, I. Partalas, A. Kosmopoulos, S. Petridis, P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, N. Baskiotis, E. Gaussier, T. Artieres, and P. Gallinari. Evaluation framework specifications. Project deliverable D4.1, 05/2013 2013. URL sites/default/files/PublicDocuments/BioASQ_D4.1-EvaluationFrameworkSpecification_final.pdf.

S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 163–171. 2010.

P. N. Bennett and N. Nguyen. Refined experts: improving classification in large taxonomies. In *Proceedings of the 32nd annual International ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, 2009.

A. Beygelzimer, J. Langford, Y. Lifshits, G. Sorkin, and A. Strehl. Conditional probability tree estimation analysis and algorithms. In *Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pages 51–58, Corvallis, Oregon, 2009. AUAI Press.

C. J. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, June 2010. URL http://research.microsoft.com/apps/pubs/default.aspx?id=132652.

L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *CIKM*, pages 78–87. ACM, 2004.

S. Choi and J. Choi. Classification and retrieval of biomedical literatures: Snumedinfo at clef qa track bioasq 2014. In *Proceedings of Question Answering Lab at CLEF*, 2014.

O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 27–35, 2004.

J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

A. Doms. *GoPubMed: Ontology-based literature search for the life sciences*. Phd thesis, Technische Universität Dresden, 2010.

T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2072–2079, 2011.

S. Gopal, Y. Yang, B. Bai, and A. Niculescu-Mizil. Bayesian models for large-scale hierarchical classification. In *Neural Information Processing Systems*, 2012.

M. Huang, A. N□v□ol, and Z. Lu. Recommending mesh terms for annotating biomedical articles. *JAMIA*, 18(5):660–667, 2011. URL http://dblp.uni-trier.de/db/journals/jamia/jamia18.html#HuangNL11.

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

D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, 1997.

A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *CoRR*, abs/1306.6802, 2013. URL http://arxiv.org/pdf/1306.6802v2.

D. D. Lewis et al. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

C.-Y. 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.

J. Lingeman and L. Dietz. UMass at BioASQ 2014: Figure-inspired text retrieval. In *2nd BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2014.

K. Liu, J. Wu, S. Peng, C. Zhai, and S. Zhu. The fudan-uiuc participation in the bioasq challenge task 2a: The antinomyra system. In *Proceedings of Question Answering Lab at CLEF*, 2014.

T.-Y. Liu, Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explor. Newsl.*, pages 36–43, 2005.

Y. Liu. Bioasq system descriptions (wishart team). Technical report, 2013.

H. Malik. Improving hierarchical svms by hierarchy flattening and lazy classification. In *1st Pascal Workshop on Large Scale Hierarchical Classification*, 2009.

Y. Mao and Z. Lu. Ncbi at the 2013 bioasq challenge task: Learning to rank for automatic mesh indexing. Technical report, 2013.

A. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 359–367, 1998.

J. Mork, A. Jimeno-Yepes, and A. Aronson. The nlm medical text indexer system for indexing biomedical literature, 2013.

M. Neves. Hpi in-memory-based database system in task 2b of bioasq. In *Proceedings of Question Answering Lab at CLEF*, 2014.

Y. Papanikolaou, D. Dimitriadis, G. Tsoumakas, M. Laliotis, N. Markantonatos, and I. Vlahavas. Ensemble Approaches for Large-Scale Multi-Label Classification and Question Answering in Biomedicine. In *Proceedings of Question Answering Lab at CLEF*, 2014.

I. Partalas, �ric Gaussier, and A.-C. N. Ngomo. Results of the first bioasq workshop. In *BioASQ@CLEF*, 2013.

F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *Computer Vision and Pattern Recognition*, pages 3482–3489, 2012.

F. Ribadas, L. de Campos, V. Darriba, and A. 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.

F. J. Ribadas-Pena, L. M. de Campos Ibanez, V. M. Darriba-Bilbao, and A. E. Romero. Cole and utai participation at the 2014 bioasq semantic indexing challenge. In *Proceedings of Question Answering Lab at CLEF*, 2014.

C. N. Silla, Jr. and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining Knowledge Discovery*, 22:31–72, 2011.

T. F. Smith and M. S. Waterman. Comparison of biosequences. *Advances in Applied Mathematics*, 2(4):482 – 489, 1981.

L. Tang, S. Rajan, and V. 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. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526738. URL http://doi.acm.org/10.1145/1526709.1526738.

G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining Multi-label Data. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.

G. Tsoumakas, M. Laliotis, N. Markontanatos, and I. Vlahavas. Large-scale semantic indexing of biomedical publications. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.

X. Wang and B.-L. Lu. Flatten hierarchies for large-scale hierarchical text categorization. In *Fifth IEEE International Conference on Digital Information Management*, pages 139–144, 2010.

K. Q. Weinberger and O. Chapelle. Large margin taxonomy embedding for document categorization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1737–1744. 2009.

D. Weissenborn, G. Tsatsaronis, and M. Schroeder. Answering factoid questions in the biomedical domain. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.

G.-R. Xue, D. Xing, Q. Yang, and Y. Yu. Deep classification in large-scale text hierarchies. In *Proceedings of the 31st annual International ACM SIGIR conference on Research and development in information retrieval*, pages 619–626, 2008.

Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the* 22^{nd} *annual International ACM SIGIR conference*, pages 42–49. ACM, 1999.

Y. Yang, J. Zhang, and B. Kisiel. A scalability analysis of classifiers in text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 96–103, 2003.

Z. L. Yuqing Mao, Chih-Hsuan Wei. Ncbi at the 2014 bioasq challenge task: large-scale biomedical semantic indexing and question answering. In *Proceedings of Question Answering Lab at CLEF*, 2014.

D. Zhu, D. Li, B. Carterette, and H. Liu. An incemental approach for medline mesh indexing. In *1st BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*, 2013.