AUTH-Atypon at BioASQ 3: Large-Scale Semantic Indexing in Biomedicine

Yannis Papanikolaou[1], Grigorios Tsoumakas[1], Manos Laliotis[2], Nikos Markantonatos[3], and Ioannis Vlahavas[1]

[1] Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
{ypapanik, dndimitr, greg, vlahavas}@csd.auth.gr
[2] Atypon, 5201 Great America Parkway Suite 510, Santa Clara, CA 95054, USA
elalio@atypon.com
[3] Atypon Hellas, Dimitrakopoulou 7, Agia Paraskevi 15341, Athens, Greece
nikos@atypon.com

Abstract. In this paper we present the methods and the approaches employed in terms of our participation to the BioASQ Challenge 2015 and more specifically in task 3a, concerning the automatic semantic annotation of scientific abstracts. Based on the successful approaches of the previous years we considered a variety of ensembles, incorporated journal-specific semantic information and developed an approach to handle the concept drift within the BioASQ corpus. The official results demonstrate a consistent advantage of our approaches against the BioASQ and the National Library of Medicine (NLM) baselines. Specifically, the systems proposed by our team ranked among the top tier ones along the competition, obtaining the second place in 10 out of 15 weeks.

Keywords: semantic indexing · multi-label learning · bio-medicine · BioASQ

1 Introduction

The BioASQ project [1] aims to provide a challenge framework for researchers dealing with classification (semantic indexing) and natural language processing (question answering) tasks in the field of bio-medicine. The challenge, similar to the previous two years, is divided in two tasks: automated semantic indexing (3a) and question answering (3b). In Task 3a participants are given a set of new, unannotated articles and are required to automatically predict the relevant MeSH terms for each one of them in a given time. For each article only the abstract along with some meta-information is provided (journal, year and title). This task is particularly difficult, as the MeSH taxonomy comprises of a large number of labels ($\sim$ 27000), with the label set following a power-law similar distribution. Furthermore the terms are subject to a significant concept drift along time.

A number of different approaches have been pursued along the previous challenges, in order to automatically annotate new articles. The NLM Medical Text

Indexer (MTI/MTIFL) [2], is a system that incorporates multiple rule-based and machine learning methods in order to effectively provide MeSH label recommendations for new articles. Other approaches include Learning-to-Rank methods [3][4], hierarchical classification [5] or multi-label ensemble approaches [6].

In this work we build on the previous year's methods [6], employing ensemble techniques for the semantic indexing task. The rest of the paper is organized as follows. In Section 2, we present the methods used throughout the semantic indexing part of the challenge. Section 3 shows the relevant results. Final considerations and conclusions are drawn in Section 4.

2 Methods

In this section we present the methods that we used for the semantic indexing task. We first provide a brief description of those approaches that were used also in our previous challenge participation [6] and then provide the various extensions of our work with more detail.

In this year's participation, we used as a training set the last 1 million articles and reserved the last 20 thousand as a validation set. For pre-processing of the articles, a similar pipeline was used as in the previous years; the abstract and the title were concatenated, one-grams and bi-grams were used as features and stop-words as well as features with less than five occurrences in the corpus were removed. Following the above steps we obtained 257,197 one-grams and 478,533 bi-grams. The *tf-idf* representation was used for the features. Also, zoning of the features belonging to the title and those equal to a MeSH label was performed; specifically, we increased the *tf-idf* value of features that belonged to the title by $log2$ and those being equal to a label by $log1.25$.

The above features were used in order to train several multi-label learning models. We used the Meta-Labeler [7], a set of Binary Relevance (BR) models with Linear SVMs (both tuned and with default parameters) and a Labeled LDA variant, Prior LDA [8]. Specifically for the SVM models, we used different values for the C parameter and handled class imbalance by penalizing more heavily false negative errors than false positive ones by adjusting properly the weight parameter [9].

2.1 Rule-Based Journal Model

Along with the previously mentioned models, we developed a rule-based model, exploiting the journal-specific distributions of labels. The BioASQ corpus contains scientific papers from more than 5000 journals, that cover diverse scientific domains and topics and therefore we expect the MeSH terms distributions to greatly vary among them. Furthermore, articles belonging to a specific journal may contain one or more MeSH labels particular to that journal, e.g. we expect an article belonging to the journal "Pediatrics", to contain the MeSH terms "Infant" or "Infant, Newborn" with a rather high probability.

Given the above observations, we first studied the label distributions among different journals and we observed that specific labels appear with very high probabilities (≥ 0.75) in every journal. Subsequently, we implemented a rule-based journal model in which labels are divided in two categories, frequent (for instance with more than 100,000 appearances out of the entire corpus of 4.2 million documents) and non-frequent. Then, each instance, according to the journal it belongs, is assigned automatically a frequent label if it has a probability of more than 0.95 in that journal and similarly a non-frequent label if the relevant probability is greater than 0.75. Naturally, multiple frequent and non-frequent labels can be assigned to the same instance. The above values were heuristically chosen, based on small-scale experiments.

2.2 Ensembles

The systems used throughout the challenge, were mainly based on ensemble methods, similar to the previous year participation. We used the MULE framework [6] and further experimented on voting systems. In the following, we describe the details.

MULE MULE [6] is a statistical significance multi-label ensemble that performs classifier selection. The key idea is to combine a set of multi-label classifiers aiming to optimize a selected measure (for the purpose of this challenge, we are mainly interested in the micro-F measure) and validate this combination through a statistical significance test; McNemar's test. This way, each label of the multi-label problem is predicted with a specific component model, the one that (a) contributes to the greatest improvement to the evaluation metric of interest and (b) is validated from the statistical test to indeed produce the aforementioned improvement. If the null hypothesis of the statistical test is not rejected for a given label (i.e. if the improvement for a specific component model is not statistically significant), we predict that label with the globally optimal model.

Voting ensembles We further considered three voting ensembles, which decide whether to assign a label to an article or not based on the votes of the component models. The first voting ensemble relied on the majority vote, while the others on two and three votes respectively.

2.3 Full-Text retrieval

In the PubMed interface[4], for a number of journals, open-access to the full text of the articles is available, through the PubMed Central (PMC) web page. The motivation is that the full text of an article will provide more semantic information and more features in order to learn MeSH terms, especially those occurring more rarely. After retrieval of a total of 160,691 entire articles (out of

[4] http://www.ncbi.nlm.nih.gov/pubmed

the entire corpus which consisted of 4.2 million abstracts) and having trained a Meta-Labeler model on the new data set, we used the model for prediction of new articles for which the full-text was also available.

In order to study the effect of including the full-text to learn a model, we considered the following strategies:

- FF: stands for use of the full text for both training and testing documents.
- FA: stands for using the full text only for documents in the training set (for documents in the test set we use only the abstract).
- AA: stands for using only the abstract for both training and testing
- AF: for using full text only for the test set documents

Table 1 shows the relevant results. We can easily see that including the full text of an article yields an improvement in Micro-F but not necessarily in the Macro-F measure. Also, the model does not seem to benefit from the respective combinations (AF, FA) In short, we would propose using the full-text on a similar semantic indexing task, only if the full text is available for both the training and the testing documents. Finally, we note that as the training data set for the full-text model was a lot smaller that the default abstracts data set (1m), the performance for these particular instances was significantly worse so this approach was not further considered during the BioASQ challenge.

Table 1. Results for four different scenarios of full text use for a training set of the first 150k and a test set of the remaining (10k) BioASQ documents. The Meta-Labeler was used to train all models.

	Micro-F	Macro-F
FF	**0.50958**	0.57187
FA	0.46018	0.57808
AA	0.49006	**0.58547**
AF	0.43992	0.55408

2.4 Strategies against the Concept Drift

The BioASQ corpus extends over a period of almost 70 years (1946-2015) and thus we expect significant changes in the meaning and the context of concepts (i.e. MeSH terms). For instance, a disease in 1970 and in 2000 can be connected to totally different causes. Furthermore, the MeSH ontology is subject to changes and additions of new terms every year. The above factors affect the MeSH - word distributions and consequently a machine learning model performance. In order to handle this phenomenon, we trained classifiers with variable training sizes and extending across various time periods (2012-2014, 2010-2014, 2007-2014) and combined them through the MULE framework (Sect. 2.2) along with the rest of the models. In this manner, we managed to use the useful semantic

information across large portions of the corpus, at the same time smoothing out the effect of the concept drift in the model's performance. A more detailed study of the temporal aspects of the data along with their effect on performance can be found in [10].

3 Results

In this section we present and discuss some key aspects with respect to the official challenge results (http://participants-area.bioasq.org/results/3a/, concerning our systems.

Table 2. Official results for the first batch of the BioASQ challenge 2015 among the AUTH-Atypon, NCBI and NLM teams.

Micro-F					
System	week 1	week 2	week 3	week 4	week 5
Auth	**0.5957**	**0.5959**	**0.6019**	**0.6094**	**0.5880**
MTI (NLM)	0.5709	0.5614	0.5724	0.5743	0.5627
MeSH NOW (NCBI)	0.5491	0.4370	0.5983	0.6012	**0.5880**
LCA-F					
Auth	**0,4942**	**0,4931**	**0,5024**	**0,5070**	0,4904
MTI (NLM)	0,4852	0,4775	0,4803	0,4881	0,4777
MeSH NOW(NCBI)	0,4582	0,3822	0,4982	0,5066	**0,4921**

We made submissions in 14 out of a total of 15 weeks. During the first and the third batch, we steadily obtained the second place for both Micro-F and LCA-F metrics (9 out of 10 weeks) while in the second batch we ranked in the third place. Table 2 shows the results in terms of the Micro-F and the LCA-F measures, for the best performing model of each of the AUTH-Atypon, NLM and NCBI teams, during the first batch (the respective results for the two other batches are available at the BioASQ challenge website). Results are shown for the already annotated articles, as of May, 11. In total, we outperformed both NLM's (MTI/MTIFL) and NCBI's (MeSH Now BF/HR) systems in terms of Micro-F, while in terms of LCA-F, we outperformed the NCBI systems in 4 out of 5 weeks and NLM systems throughout the batch.

In order to indicate the improvement of our systems with respect to last year, in Table 3, we additionally compare the mean results of this year's challenge (BioASQ 3) to the respective ones from last year (BioASQ 2). Results are shown for each year's data sets in terms of the mean performance in terms of Micro-F and LCA-F, for the top performing model across all weeks. We can observe a significant improvement between our two participations, mainly related to our use of a wider variety of component models, as well as to different parameterizations of the MULE ensembles.

Table 3. Mean performance for our best performing system across BioASQ challenges 2014 and 2015.

Micro-F			
	Batch 1	Batch 2	Batch 3
BioASQ 2 (2014)	$0,58982 \pm 0,00291$	$0,59388 \pm 0,00489$	$0,59528 \pm 0,01007$
BioASQ 3 (2015)	$0,59818 \pm 0,00798$	$0,60826 \pm 0,01056$	$0,62400 \pm 0,01237$
LCA-F			
	Batch 1	Batch 2	Batch 3
BioASQ 2 (2014)	$0,49495 \pm 0,00095$	$0,49716 \pm 0,00367$	$0,49752 \pm 0,00614$
BioASQ 3 (2015	$0,49742 \pm 0,00698$	$0,50090 \pm 0,00613$	$0,51840 \pm 0,00790$

4 Conclusions

In this paper we presented the participation of the AUTH-Atypon team in the BioASQ challenge 2015. Building on the successful approaches in the past two challenges, we further extended our line of work to improve the performance of our systems, employing ensemble techniques for a number of component models. The official challenge results demonstrate a clear advantage of our methods over the BioASQ baseline as well as the NLM and NCBI systems.

References

1. Balikas, G., Partalas, I., Ngomo, A.N., Krithara, A., Paliouras, G.: Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. (july 2014) 1181–1193
2. Mork, J.G., Demner-Fushman, D., Schmidt, S.C., Aronson, A.R.: Recent enhancements to the NLM medical text indexer. In: Working Notes for CLEF 2014 Conference, Sheffield, UK. (2014) 1328–1336
3. Mao, Y., Wei, C., Lu, Z.: NCBI at the 2014 BioASQ Challenge Task: Large-scale Biomedical Semantic Indexing and Question Answering. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. (2014) 1319–1327
4. Liu, K., Wu, J., Peng, S., Zhai, C., Zhu, S.: The Fudan-UIUC Participation in the BioASQ Challenge Task 2a: The Antinomyra system. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. (2014) 1311–1318
5. Ribadas-Pena, F.J., de Campos Ibañez, L.M., Bilbao, V.M.D., Romero, A.E.: CoLe and UTAI Participation at the 2014 BioASQ Semantic Indexing Challenge. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. (2014) 1361–1374
6. Papanikolaou, Y., Dimitriadis, D., Tsoumakas, G., Laliotis, M., Markantonatos, N., Vlahavas, I.P.: Ensemble Approaches for Large-Scale Multi-Label Classification and Question Answering in Biomedicine. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. (2014) 1348–1360
7. Tang, L., Rajan, S., Narayanan, V.K.: Large scale multi-label classification via metalabeler. In: WWW '09: Proceedings of the 18th international conference on World wide web, New York, NY, USA, ACM (2009) 211–220

8. Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical Topic Models for Multi-label Document Classification. Mach. Learn. **88**(1-2) (July 2012) 157–208
9. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: RCV1: A New Benchmark Collection for Text Categorization Research. J. Mach. Learn. Res. **5** (2004) 361–397
10. Papanikolaou, Y., Tsoumakas, G., Laliotis, M., Markantonatos, N., Vlahavas, I.: Large-scale semantic indexing of biomedical papers via a statistical significance multi-label ensemble. Journal of Biomedical Semantics (2015) Manuscript accepted for publication.