Large-Scale Semantic Indexing of Biomedical Publications at BioASQ

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

[1] Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
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. Automated annotation of scientific publications in real-world digital libraries requires dealing with challenges such as large number of concepts and training examples, multi-label training examples and hierarchical structure of concepts. BioASQ is a European project that contributes a large-scale biomedical publications corpus for working on these challenges. This paper documents the participation of our team to the large-scale biomedical semantic indexing task of BioASQ.

Keywords: multi-label learning, semantic indexing, biomedical literature, text mining

1 Introduction

The amount of scientific publications digitally available online is constantly increasing. New conference publications and journal articles are continuously added to digital libraries of publishers (e.g. Elsevier's sciencedirect, Springer's springerlink), scientific societies (e.g. ACM digital library, IEEE explore), search engines (e.g., Google Scholar) and open access repositories (e.g. arXiv.org, CiteSeerX). On top of this scientific knowledge, digital libraries strive to offer useful services, such as search, exploration, filtering, citation analysis and trend detection. Content-based services of digital libraries rely largely on publications being accompanied by semantic meta-data with all relevant concepts from the ontology of the corresponding domain, such as the Medical Subject Headings (MeSH) for Medicine and the ACM Computing Classification System for Computing.

Some libraries employ experts to manually annotate publications at the document level according to a domain's ontology. PubMed for example manually indexes its collection according to MeSH. However, this entails significant costs in time and money. An alternative solution is automatic indexing of publications by computer systems utilizing text categorization technology. Automatic indexing is important even for libraries that can afford manual annotation for

two reasons. Firstly, it may take a couple of months from the moment a publication enters the library to the moment it receives its annotation. For a publication with novel and important scientific results, this first period of its lifetime is quite important, yet it is this period that remains semantically invisible. Secondly, automatic indexing can serve as assistive technology to the human annotators by ranking the concepts according to predicted relevance to a document or filtering out large parts of the ontology that it predicts unrelated with high confidence.

While text mining research has progressed significantly in the last 10 years, the problem of automatic indexing of scientific publications in real-world digital libraries presents some unique challenges that remain largely unsolved. Real-world digital libraries curate ontologies composed of thousands of concepts and manage collections composed of millions of publications. Efficient yet accurate learning and inference with such large ontologies and training sets is non-trivial. The concepts in real-world ontologies are hierarchically structured as a directed acyclic graph indicating subsumption realtions among parent and child concepts. While some progress has been recently achieved on exploiting such relationships, it is not entirely clear when and how these relationships help accuracy. Each scientific document is typically annotated with more than a concept, rendering semantic indexing of scientific literature a multi-label learning task [1], which presents the additional challenge of exploiting label dependencies to improve accuracy. Finally, as domain ontologies evolve on par with the scienctific areas they describe, automatic indexing models must deal with changes in the ontology, both explicit (i.e. addition, deletion, merging of concepts) and implicit (i.e. altered semantics of concepts) ones.

BioASQ[4] is a timely European project that offers a perfect playground for researchers working on these challenges. It made available a training corpus consisting of approximately 11 million articles from MEDLINE, each one annotated on average with approximately 13 concepts from MeSH. It also organized a long-term real-world time-constrained benchmark. For 18 weeks, each Monday at 17:00 o' clock CET it released a batch of recent unannotated MEDLINE documents whose size ranged from 793 to 10,233 documents, and within 21 hours, it requested a set of concepts for each of these documents. This paper documents the participation of our team in this benchmark. Section 2 discusses the general approach that we used to deal with the problem and Section 3 describes the particular systems we used to submit annotations. Section 4 presents the results that we achieved and Section 5 mentions the open issues left for future research.

2 Our Approach

Starting from the 10,876,004 documents of the released training corpus, we initially removed duplicate entries based on the *pmid* of the documents, which is the unique number assigned to each PubMed record. This led to a reduced training corpus of 10,699,707 documents.

[4] http://bioasq.org/

These documents belong to 8,916 journals. The challenge organizers decided to sample test documents from 1,806 of these journals that are characterized by small annotation time, in order to avoid delays in the evaluation of systems. We therefore filtered the training corpus keeping only documents from these 1,806 journals, in order to make the distribution of the training documents as similar as that of the test documents, which is a core assumption in supervised machine learning. This was not straightforward, as the list of the 1,806 journals contained abbreviated titles, while the training corpus contained full titles. We fortunately managed to retrieve from the NLM catalog a text file containing both full and abbreviated titles of all PubMed journals (J_Medline.txt). This filtering process led to a reduced corpus of 3,950,721 documents.

The last 12,000 documents of this corpus was withheld as a test set, in order to simulate a cup of the challenge, as initial guidelines for this task mentioned that each batch would consist of approximately 2,000 documents. Note that the corpus was sorted chronologically, so these 12,000 documents were the most recent ones (from years 2012 and 2013).

We first extracted the title and the abstract of each document and tokenized the text using Stanford CoreNLP[5]. We then lower-cased the tokens and constructed a dictionary of unigrams and bigrams with at least 6 occurences in the training corpus. Tf-idf values were computed for each token and normalized to have unit length across each document.

Learnin was based on the meta-labeler approach [2], which learns one model for ranking labels according to their relevance with an instance and another model for predicting the number of labels related to an instance. Given a test instance, it selects the top N most relevant labels, where N is the prediction of the latter model. Our implementation of the meta-labeler approach is based on linear support vector machines (SVMs) using default parameters ($cost = 1$, $tolerance = 0.01$, $bias = 1$). For ranking the labels we train a binary classification model for each of the labels present in our training corpus, while for predicting the number of labels we train a regression model.

Our representation and learning approaches were chosen based on our experience with a similar learning problem for the past year, the main difference being the availability of the full text of publications. Within that project, we have investigated a variety of other approaches, including other thresholding strategies [3], such as SCut [4,5], class imbalance counterfeiting approaches [6] such as majority class undersampling and asymmetric bagging [7], and different representations, such as plain unigrams/bigrams, adding trigrams and BNS scaling [8], all with worse results compared to the one we described here. We have also unsuccessfully attempted to exploit the hierarchy information.

3 Particular Systems

We participated in the challenge using four variations of the main approach that was presented in Section 2. Systems 1 and 3 use respectively the last 800,000 and

[5] http://nlp.stanford.edu/software/corenlp.shtml

700,000 documents of the reduced corpus prior to the 12,000 documents withheld for testing. We did not use all 3,950,721 documents in order to reduce the time and space complexity of the approach. System 2 is a simple ensemble, which considers the output of the binary SVMs of System 3 for some of the labels and the output of the meta-labeler of System 3 for the rest. This was motivated from the observation that the meta-labeler performs worse than the binary SVMs for some labels, especially for highly frequent ones. The choice of model per label was tuned greedily based on the micro F-measure of the ensemble on the held-out test set. System 4 is an ensemble of three systems similar to systems 1 and 3, each one based on a different 500,000 document subset of the reduced corpus. In particular, we considered the last 1,500,000 documents prior to the 12,000 documents withheld for testing and distributed them round-robin to the three 500,000 document models. Combination of the output of these models was based on majority voting. Table 1 presents the number of examples, unigrams, bigrams and labels for each of the models.

Table 1. Number of examples, unigrams, bigrams and labels for each of our models.

system	examples	unigrams	bigrams	labels
1	800,000	215,133	1,908,088	25,625
2,3	700,000	197,590	1,720,818	25,509
4a	500,000	138,157	1,097,725	25,212
4b	500,000	138,230	1,097,289	25,215
4c	500,000	138,200	1,097,380	25,215

Experiments were run on an HP DL580R07 server featuring 4 10-core processors at 2.26 GHz, 1 Tb of RAM and 6 10k SAS disks of capacity 600 Gb each set up in RAID 5 for a total of 2.4 Tb storage. The server is running the Linux CentOS operating system. The largest computational challenge was training the thousands of binary SVM models. By utilizing parallelization at the label level and exploiting 40 threads, training required approximately between one and two days. Note that as predictions were required within 16 hours of test data release, serialization was used to store the trained binary models at disk. Storing the models of system 1 for example required 406 Gb. Parallelization at the label level was also used during prediction.

4 Results

Table 2 reports the performance of our systems as well as their relative position (in parentheses) at September 15th, 2013 in terms of Micro F-measure and Lowest Common Ancestor (LCA) F-measure [9]. Our systems topped the performance chart for most of the duration of the challenge.

During the first two weeks of the 1st cup, our systems were based on 565,134 documents corresponding to a small number of journals whose abbreviations

Table 2. Performance as well as relative position (in parentheses) of our systems at August 15th, 2013 in terms of Micro and LCA F-measure.

		Micro F-measure				Lowest Common Ancestor F-measure			
Cup	Week	System 1	System 2	System 3	System 4	System 1	System 2	System 3	System 4
1	1	0.415 (11)	0.432 (9)	0.424 (10)	-	0.344 (12)	0.0.356 (11)	0.381 (9)	-
	2	-	-	0.444 (12)	-	-	-	0.392 (12)	-
	3	-	-	-	-	-	-	-	-
	4	-	-	0.561 (1)	-	-	-	0.479 (1)	-
	5	-	0.566 (2)	0.573 (1)	-	-	0.466 (3)	0.477 (1)	-
	6	-	0.572 (2)	0.580 (1)	-	-	0.467 (2)	0.483 (1)	-
2	1	-	0.559 (2)	0.563 (1)	-	-	0.468 (3)	0.478 (1)	-
	2	-	0.567 (1)	0.565 (2)	-	-	0.472 (2)	0.476 (1)	-
	3	-	0.571 (2)	0.572 (1)	-	-	0.477 (2)	0.484 (1)	-
	4	0.568 (1)	0.566 (3)	0.567 (2)	-	0.478 (1)	0.470 (3)	0.474 (2)	-
	5	0.563 (1)	0.001 (23)	0.001 (22)	-	0.472 (1)	0.105 (23)	0.106 (22)	-
	6	0.578 (1)	0.572 (3)	0.576 (2)	-	0.484 (1)	0.473 (3)	0.483 (2)	-
3	1	0.573 (1)	0.571 (3)	0.572 (2)	-	0.484 (1)	0.475 (3)	0.481 (2)	-
	2	0.579 (1)	0.574 (3)	0.578 (2)	-	0.487 (1)	0.477 (3)	0.486 (2)	-
	3	0.578 (1)	0.574 (3)	0.575 (2)	-	0.487 (1)	0.480 (3)	0.486 (2)	-
	4	0.564 (1)	0.562 (2)	0.562 (3)	0.295 (24)	0.473 (1)	0.467 (3)	0.473 (2)	0.236 (28)
	5	0.567 (1)	0.562 (3)	0.563 (2)	0.320 (22)	0.476 (1)	0.465 (4)	0.471 (2)	0.257 (27)
	6	0.443 (21)	0.043 (33)	0.443 (19)	0.440 (20)	0.348 (25)	0.096 (33)	0.354 (21)	0.349 (24)

were equal to their title, such as *Nature* and *Gut*, as we were unaware of the fact that abbreviations were being used in the provided list of journals of test documents. Week 3 of the 1st cup coincided with the Orthodox Easter and we didn't manage to make a submission.

The correct version of system 3 was introduced at the 4th week of the 1st cup, while the correct version of system 2 followed one week later. System 3 was consistently better than System 2 in both evaluation measures, with the exception of the 2nd week of the 2nd cup and the 4th week of the 3rd cup in the case of Micro F-measure. This shows that the tuning we did for System 2 has most probably overfitted the evaluation set and that a more careful process for selecting the labels to be predicted directly by the corresponding binary SVMs must be devised. At the 5th week of the 2nd cup, a parsing issue led to erratic submissions for Systems 2 and 3. At the last week of the last cup, we accidentally submitted erratic models for systems 1-3.

System 1 was introduced at the 4th week of the 2nd cup and topped the performance tables since then, with the exception of the erratic submission in the very last week of the challenge. The actual values of the micro F-measure of this best system of the challenge are around 0.57, which is not a breathtaking performance, yet it surpassed the performance of *MTI First Line Index*, a baseline system from the National Library of Medicine. When contemplating absolute performance in this challenge, one should not forget the difficulties of the data (no availability of full text, large number of labels, complex relationships of labels, potentially noisy labelling).

System 4 was introduced at the 4th week of the 3rd cup. While its performance in terms of the two evaluation measures wasn't good, it was the system that outweighed the rest by far in the precision measures for the 3 weeks that

it participated in the contest. In particular its average performance in these 3 weeks was: 0.83 (example-based precision), 0.82/0.76 (micro/macro averaged precision), 0.91 (hierarchical precision) and 0.56 (LCA precision). This excellent performance shows that a majority voting ensemble can produce highly precise classifiers for this task that could potentially be used for partial yet accurate fully automatic indexing of biomedical literature.

5 Open Issues

One issue that we plan to explore in the near future concerns the temporal dimension of the data. We have already found that the frequency of the labels varies over time. We want to explore whether handling concept-drift [10] can bring predictive accuracy improvements. Other issues of interest are whether journal information can be exploited for improving predictive accuracy and whether text of the abstract and the title of a publication should be treated separately.

References

1. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In Maimon, O., Rokach, L., eds.: Data Mining and Knowledge Discovery Handbook. 2nd edn. Springer (2010) 667–685
2. 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
3. Ioannou, M., Sakkas, G., Tsoumakas, G., Vlahavas, I.: Obtaining bipartitions from score vectors for multi-label classification. In: 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2010), Los Alamitos, CA, USA, IEEE Computer Society (2010) 409–416
4. Yang, Y.: A study of thresholding strategies for text categorization. In: SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference, New York, NY, USA, ACM (2001) 137–145
5. Fan, R.E., Lin, C.J.: A study on threshold selection for multi-label classification. Technical report, National Taiwan University (2007)
6. Sun, A., Lim, E.P., Liu, Y.: On strategies for imbalanced text classification using svm: A comparative study. Decison Support Systems **48**(1) (2009) 191–201
7. Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(7) (2006) 1088–1099
8. Forman, G.: BNS feature scaling: an improved representation over tf-idf for svm text classification. In: Proceedings of the 17th ACM conference on Information and knowledge management. CIKM '08, New York, NY, USA, ACM (2008) 263–270
9. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation measures for hierarchical classification: a unified view and novel approaches (2013)
10. Xioufis, E.S., Spiliopoulou, M., Tsoumakas, G., Vlahavas, I.P.: Dealing with concept drift and class imbalance in multi-label stream classification. In Walsh, T., ed.: IJCAI, IJCAI/AAAI (2011) 1583–1588