Tagging Funding Agencies and Grants in Scientific Articles using Sequential Learning Models

Subhradeep Kayal **Zubair Afzal** **George Tsatsaronis** **Sophia Katrenko**
Pascal Coupet **Marius Doornenbal** **Michelle Gregory**
Content and Innovation Group
Operations Division
Elsevier B.V.
The Netherlands
{d.kayal, m.afzal.1, g.tsatsaronis, s.katrenko, p.coupet, m.doornenbal, m.gregory}@elsevier.com

Abstract

In this paper we present a solution for tagging funding bodies and grants in scientific articles using a combination of trained sequential learning models, namely conditional random fields (*CRF*), hidden markov models (*HMM*) and maximum entropy models (*MaxEnt*), on a benchmark set created in-house. We apply the trained models to address the *BioASQ* challenge 5c, which is a newly introduced task that aims to solve the problem of funding information extraction from scientific articles. Results in the dry-run data set of *BioASQ* task 5c show that the suggested approach can achieve a micro-recall of more than 85% in tagging both funding bodies and grants.

1 Introduction and Description of the BioASQ Task 5c

The scientific research and development market is a $136bn industry in the US alone, with a 5-year growth of 2.3%, as recorded in 2017[1]. Within this economy, organizations which fund research need to ensure that they are awarding funds to the right research teams and topics so that they can maximize the impact of the associated available funds. As a result, institutions and researchers are required to report on funded research outcomes, and acknowledge the funding source and grants. In parallel, funding bodies should be in a position to trace back these acknowledgements and justify the impact and results of their research allocated funds to their stakeholders and the taxpayers alike. Researchers should also be able to have access to such information, which can help them make better educated decisions during their careers, and help them discover appropriate funding opportunities for their scientific interests, experience and profile. This situation creates unique opportunities for the affiliated industry, to coordinate and develop low-cost, or cost-free, solutions that can serve funding agencies and researchers. A fundamental problem that needs to be addressed is, however, the ability to automatically extract the funding information from scientific articles, which can in turn become searchable in bibliographic databases.

In this work we address this problem of automating the extraction of funding information from text, using machine learning techniques. We evaluate and combine several state-of-the-art sequential learning approaches, to accept a scientific article as a raw text input and provide the detected funding agencies and associated grant IDs as output.

In order to test our approach, we have participated in the *BioASQ* challenge 5c[2], which is a part of the larger *BioASQ* challenge. *BioASQ* organizes challenges which include tasks relevant to hierarchical text classification, machine learning, information retrieval, QA from texts and structured data, multi-document summarization and many other areas (Tsatsaronis et al., 2015). In this particular task (challenge 5c), the participants are asked to extract grant and funding agency information from full text documents available in PubMed Central[3]. Annotations from PubMed are used to evaluate the information extraction performance of participating systems, with the evaluation criterion being micro-recall. Furthermore, the agencies to be reported must be in a predetermined list as provided by the National Library of Medicine

[1] https://www.ibisworld.com/industry/default.aspx?indid=1430

[2] http://participants-area.bioasq.org/general_information/Task5c/

[3] https://www.ncbi.nlm.nih.gov/pmc/

Proceedings of the BioNLP 2017 workshop, pages 216–221,
Vancouver, Canada, August 4, 2017.

(NLM)[4].

2 Background Literature

2.1 Named Entity Recognition

Named entity recognition (*NER*) locates units of information, such as names of organizations, persons and locations and numeric expressions, from unstructured text. Each such unit of information is then known as a *named entity*. In the context of this paper, the named entities that are identified are either *Funding Agencies* (*FA*) or *Grant IDs* (*GR*). As an example, given a text of the form: "*This work was supported by the Funding Organization with grant No. 1234*", the *NER* task is to label "*Funding Organization*" in text as *FA* and "*1234*" as *GR*. In principle, effective *NER* systems usually employ rule-based (Farmakiotou et al., 2000; Cucerzan and Yarowsky, 1999; Chiticariu et al., 2010), gazetteer (Ritter et al., 2011; Torisawa, 2007) and machine learning approaches (Chieu, 2002; McCallum and Li, 2003; Florian et al., 2003; Zhou and Su, 2002). In this work we utilize several sequential learning (Dietterich, 2002) machine learning approaches for *NER*, which are discussed next. A detailed survey of *NER* techniques for further reading may be found in the work of Nadeau et al. (2007).

2.1.1 Sequential Learning Approaches

Sequential learning approaches model the relationships between nearby data points and their class labels, and can be classified into *generative* or *discriminative*. In the context of *NER*, *Hidden Markov Models* (*HMMs*) are generative models that learn the joint distribution between words and their labels (Bikel et al., 1999; Zhou and Su, 2002). A *HMM* is a *Markov chain* with hidden states, and in *NER* the observed states are words while the hidden states are their labels. Given labelled sentences as training examples, *NER HMMs* find the maximum likelihood estimate of the parameters of the joint distribution, a problem for which many algorithmic solutions are known (Rabiner, 1990). *Conditional Random Fields* (*CRFs*) are discriminative, in contrast to *HMMs*, and find the most likely sequence of labels or entities given a sequence of words. The relationship between the labels is modelled by a *Markov Random Field*. *Linear chain CRFs* are well suited to sequence analysis and have been applied succssfully in the past in parts-of-speech tagging (Lafferty et al., 2001), shallow parsing (Sha and Pereira, 2003) and *NER* (McCallum and Li, 2003). Finally, another way of modelling data for *NER* is *Maximum Entropy* (*MaxEnt*) models, which select the probability distribution that maximizes entropy, thereby making as little assumptions about the data as possible. Following the seminal work of Berger et al. (1996), maximum entropy estimation has been successfully applied to *NER* in many works (Chieu, 2002; Bender et al., 2003). Essentially, *CRFs* are also maximum entropy models working over the entire sequence, whereas *MaxEnt* models make decisions for each state independently of the other states.

2.1.2 State-of-the-art Open-source Toolkits

Several open-source toolkits implement one or more of the learning approaches mentioned in the previous section. This section discusses three of them in particular, which have been found to be efficient, scalable and robust in practice, and which are used as base approaches in the current work.

The *Stanford CoreNLP toolkit*[5] is a *JVM*-based text annotation framework whose *NER* implementation is based on enhanced *CRFs* with long-distance features to capture more of the structure in text (Finkel et al., 2005). An important feature of the toolkit is the ability to use distributional similarity measures, which assume that similar words appear in similar contexts (Curran, 2003). The toolkit is released with a well-engineered feature extractor, as well as pre-trained models for recognizing *persons*, *locations* and *organizations*.

LingPipe[6] is another *Java*-based *NLP* toolkit, whose efficient *HMM* implementation includes n-gram features. The toolkit has been successfully applied in the past in gene recognition in text (Carpenter, 2007).

Finally, in this work we also use the *Apache OpenNLP*[7] toolkit, which implements *NER* either by using discriminative trained *HMMs* (Collins, 2002), or by training *MaxEnt* models (Ratnaparkhi, 1998).

[4] https://www.nlm.nih.gov/bsd/grant_acronym.html

[5] http://stanfordnlp.github.io/CoreNLP/

[6] http://alias-i.com/lingpipe/demos/tutorial/read-me.html

[7] https://opennlp.apache.org/

2.2 Related Work

To the best of our knowledge, this is the first piece of research work that systematically explores the concept of extracting funding information from the full text of scientific articles. The next closest category of related published research works mostly aims at extracting names of organizations from affiliation strings, e.g., the works of Jonnalagadda et al. (2010), and Yu et al. (2007), both of which aim at extracting names of organizations from the metadata of published scientific articles. There are, however, several initiatives that started recently and are aiming at a similar direction to the current work, such as the *ERC* project "*Extracting funding statements from full text research articles in the life sciences*"[8].

3 Methodology

3.1 Overview

The suggested approach receives as input a text chunk, e.g., the raw full text of a scientific article, and annotates the input text with entities corresponding to *Funding Agencies* (*FAs*) and *Grant IDs* (*GRs*), where present. A two-step search strategy for finding *FA* and *GR* entities in text has been implemented. The process starts by splitting the input text into paragraphs, which are in turn given sequentially as input to a binary text classifier that identifies only those paragraphs which may contain any funding information. *NER* is performed next, only on the said filtered text paragraphs, to annotate them with *FA* and *GR* labels. This design enjoys several benefits; primarily it minimizes the execution time of the approach, as the most costly component, which is the *NER* part, is only executed in a small selection of paragraphs in which the binary text classifier has detected evidence of funding information. In parallel, it reduces significantly the false positives of the approach, as there are many text segments in a scientific full text article that contain strings which a *NER* component could potentially annotate falsely as *FA*, e.g., the organisation names in the affiliation information of the authors.

3.2 Training Data Gathering

For this task, we have created a "*Gold*" set for training, i.e., a manually curated and annotated set of scientific articles with *FA* and *GR* labels. Such a gold set was created, even though *BioASQ* task 5c provides a training set, as several discrepancies were observed in the said training set, the most important being the absence of entity offsets. The "*gold*" set was created with journal articles from a large number of scientific publishers, and comprises $1,950$ articles annotated by three professional annotators, who were provided with comprehensive guidelines explaining the process and the entities. A harmonization process then merged the annotations of the three experts; when all three agreed, annotations were automatically harmonized, whilst the disagreements between the annotators were resolved manually by a subject matter expert (*SME*). From the $1,950$ articles, $1,682$ contained at least one funding-related annotation. As for the individual entities, a total of $3,428$ *FA* and $2,592$ *GR* annotations exist in the set. Pairwise averaged *Cohen*'s *kappa* (Cohen, 1960) was used to calculate the inter-annotators agreement, which for this set was measured at 0.89, suggesting a high-quality dataset. The "*gold*" set was used for two purposes: (i) to train the binary text classifier that detects the paragraphs of text which contain funding information; the number of positive samples were found to be $1,682$, while the number of negative samples had a much higher value at $47,565$, constituting a highly imbalanced set for the task, and, (ii) to train the *NER* components that detect *FA* and *GR* entities.

3.3 Detecting Text with Funding Information

The first step is to separate the parts of the text which contain funding information from the parts which do not. To address this problem, we have used *Support Vector Machines* (*SVMs*), which are known to perform favourably on text classification problems (Joachims, 1998). More precisely, an *L2 regularized linear SVM* has been used, operating on *TF-IDF* vectors extracted from the segments of each input text, based on a bigram bag-of-words representation. The *SVM* was trained on the examples of positive and negative segments, i.e., paragraphs with and without funding information, which could be found in the "gold" set described in the previous section. The regularization parameter for the *SVM* was found to be $C = 2$ based on cross-validation experiments to maximize the final recall.

[8] http://cordis.europa.eu/result/rcn/186297_en.html

3.4 Training and Using Sequential Learning Models

As described in section 2.1.1 and 2.1.2, we have employed a variety of complementary techniques to best extract the described entities from text. All of the individual models, namely, a *CRF* implementation from the *Stanford CoreNLP*, a *LingPipe* based enhanced *HMM*, and an *OpenNLP* implementation of the *MaxEnt* tagger, were trained on the said "gold" set using the default hyperparameter settings, as provided by their respective implementations.

Additionally, word clusters were provided to the *Stanford CoreNLP* toolkit, which has the ability to utilize distributional similarity features. The clustering was performed by first extracting word-embedding vectors from the "gold" set, using the unsupervised *Word2Vec* algorithm by Mikolov et al. (2013), followed by performing k-means clustering to create the clusters, based on the cosine-similarity of the word vectors.

For the specific purpose of *BioASQ* challenge 5c, keeping in mind that it is evaluated on micro-recall, the unique outputs of the various models were pooled in, to create the final list of named entities to be provided as output.

3.5 Task Specific Post-processing Detected Entities

In order to perform well on *BioASQ* 5c, some additional post-processing steps were performed.

Extraction of Funding Agency from Grant ID Usually grant IDs contain an acronym from which the corresponding funding agencies can be inferred. As an example, a fictitious grant of the form "MRC123A" would contain the acronym "MRC", signifying that it has been sanctioned by the "Medical Research Council". For task 5c of *BioASQ*, NLM provides a dictionary of acronyms mapped to the respective agency[9], which has been used to retrieve funding agencies from the detected grant IDs.

Corrections to Grants In some cases the prefix of grant numbers was incorrectly published with a letter 'O' rather than the numeric '0'. For example, RO1/AI45338-04 instead of R01/AI45338-04. As NLM has corrected these in their annotations, so did we in a post-processing step.

[9] https://www.nlm.nih.gov/bsd/grant_acronym.html

Method	FA μR	GR μR
HMM	80.4	82.3
MaxEnt	81.1	83.9
CRF-distsim	83.3	86.1
Pooled	**85.2**	**86.2**

Table 1: Percentage Micro-recall results for the identification of *Funding Agencies* (*FA*) and *Grant IDs* (*GR*) from the dry-run dataset of *BioASQ* task 5c.

4 Results

As the aforementioned models are trained on a entirely different manually curated "gold" set, evaluations could be made in one pass on the entire dry-run data set of *BioASQ* task 5c, which consisted of $15,205$ documents from PubMed.

Table 1 presents the micro-recall results of the trained models being evaluated on the dry-run dataset. The models listed as *HMM* and *MaxEnt* are self-explanatory, while *CRF-distsim* is the *Stanford CoreNLP* toolkit based *CRF* model which also utilizes distributional similarities, as described in section 3.4. *Pooling* represents the meta-model created by pooling in all the outputs from the individual models. In each case, the outputs undergo the same post-processing step, as described in the previous section.

The table shows that the *CRF* model performs extremely well and is complemented by the other models, all of which better the micro-recall of the *pooled* meta-model, which performs 1.9 percentage points better than the *CRF* in detecting *FA* entities, while performing comparably for *GR* annotations.

5 Conclusions

In this paper we have tackled the problem of funding information extraction from scientific articles, in the context of the *BioASQ* challenge 5c. We have tested and combined state-of-the-art sequential learning models, along with creating a benchmark dataset for training. The results on the dry-run dataset of the challenge indicate the good performance of *Conditional Random Fields* as well as the complementary performance of the other models, whose combination is evaluated at an overall best micro-recall of 85.2% for *Funding Agencies* and 86.2% for *Grant IDs*.

References

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum entropy models for named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. pages 148–151.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–71.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning* 34(1-3):211–231.

Bob Carpenter. 2007. Lingpipe for 99.99% recall of gene mentions. In *Proceedings of the 2nd BioCreative Workshop*.

Hai Leong Chieu. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 2002 International Conference on Computational Linguistics*. pages 190–196.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. pages 1002–1012.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement* 20(1):37–46.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. pages 1–8.

Silviu Cucerzan and David Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC*. pages 90–99.

James R. Curran. 2003. *From distributional to semantic similarity*. Ph.D. thesis, University of Edinburgh.

Thomas G. Dietterich. 2002. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*. pages 15–30.

Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D. Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000*. pages 75–78.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. pages 363–370.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. pages 168–171.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*. pages 137–142.

Siddhartha Jonnalagadda and Philip Topham. 2010. Nemo: Extraction and normalization of organization names from pubmed affiliation strings. *Journal of Biomedical Discovery and Collaboration* 5:50–75.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. pages 282–289.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. pages 188–191.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. pages 3111–3119.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3–26.

Lawrence R. Rabiner. 1990. In *Readings in Speech Recognition*, chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–296.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 1524–1534.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. pages 134–141.

Kentaro Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pages 698–707.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* 16(1).

Wei Yu, Ajay Yesupriya, Anja Wulf, Junfeng Qu, Marta Gwinn, and Muin J. Khoury. 2007. An automatic method to generate domain-specific investigator networks using pubmed abstracts. *BMC Medical Informatics and Decision Making* 7(1).

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. pages 473–480.