Extraction of Regulatory Events using Kernel-based Classifiers and Distant Supervision

Andre Lamurias[1]*, Miguel J. Rodrigues[1], Luka A. Clarke[2], Francisco M. Couto[1],
[1]LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2]BioISI: Biosystems & Integrative Sciences Institute,
Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract

This paper describes our system to extract binary regulatory relations from text, used to participate in the SeeDev task of BioNLP-ST 2016. Our system was based on machine learning, using support vector machines with a shallow linguistic kernel to identify each type of relation. Additionally, we employed a distant supervised approach to increase the size of the training data. Our submission obtained the third best precision of the SeeDev-binary task. Although the distant supervised approach did not significantly improve the results, we expect that by exploring other techniques to use unlabeled data should lead to better results.

1 Introduction

The SeeDev task of BioNLP-ST 2016 consisted in extracting relations between biomedical named entities on a set of texts about *Arabidopsis thaliana*(Chaix et al., 2016). These texts were manually annotated with entities and relations relevant to seed storage and reserve accumulation. Furthermore, the type of entities that could have a specific role on each type of relation was specified by the organization. There were two subtasks: the first task, binary relation extraction (SeeDev-binary), considered only relations between two arguments; the second, full event extraction, considered relations that could be composed by two to eight arguments. For both tasks, the evaluation criteria used consisted in comparing the type and arguments of each predicted relation to the gold standard. A total of 7 teams participated on this task. The best F-measure achieved was of 0.432, which is slightly lower than the best scores obtained for the comparable task on the 2013 edition of BioNLP-ST (Cancer Genetics task (Pyysalo et al., 2015): 0.554; Gene Regulation Network task (Bossy et al., 2015): 0.45; GENIA task (Kim et al., 2015): 0.489)

Our team has developed a system for the identification of chemical entities and interactions, based on Conditional Random Fields, kernel methods and domain knowledge. We have also adapted this system to other types of entities such as temporal expressions and clinical events. The SeeDev-binary subtask provided us with an opportunity to test our system on a new domain, which contains more types of entities and relations than the domains we had previously tested on.

We adapted the relation extraction module of our system to the types of relations considered by the SeeDev-binary subtask. For each type of relation, we trained a classifier with the shallow linguistic kernel. We used every sentence containing at least two entities of the types accepted by that relation type. Since there was no ontology readily available for this domain, we were not able to integrate domain knowledge. Alternatively, we experimented a distant supervision approach by using a large number of documents to find sentences containing pairs that were already present on the training corpus. Our system is available at `https://github.com/AndreLamurias/IBEnt`

The following sections describe the main methods used by our system (Section 2), the results obtained with our submission and post-challenge improvements (Section 3), and a discussion about these results (Section 4).

2 Methods

This section describes the methods used by our system. The pre-processing and relation extrac-

*Corresponding author: alamurias@lasige.di.fc.ul.pt

Proceedings of the 4th BioNLP Shared Task Workshop, pages 88–92,
Berlin, Germany, August 13, 2016.

tion steps were already part of our system, implemented for other biomedical domains. For this task, we tested a basic distant supervision approach.

2.1 Pre-processing

The first step of our system consisted in preprocessing the input text using the Genia Sentence Splitter (Sætre et al., 2007) and the Stanford CoreNLP pipeline (Toutanova and Manning, 2000). The latter tokenizes the text into word tokens and extracts the corresponding lemmas and part-of-speech, and named entity tags (proper noun, numerical and temporal entities). We implemented additional tokenization rules to separate words linked by dashes, dots and slashes because biomedical entities may be part of expressions containing these characters.

2.2 Relation extraction

Each of the 22 types of relations has two arguments, and each argument is restricted to a set of entity types specific to each relation type. These restrictions were established by the task organizers. The sentences that satisfied the entity type requirements were considered to train and test a classifier of that relation type. The tokens that comprise the relation arguments were replaced by a generic string in order to reduce the variability of the text. Furthermore, for the types "Has_Sequence_Identical_To" and "Is_Functionally_Equivalent_To", we considered only pairs with the same entity type.

The machine learning algorithm used to train the classifiers was a variation of Support Vector Machines, with the shallow linguistic kernel, as implemented by jSRE (Giuliano et al., 2006). Kernel methods rely on a kernel function which computes the inner product between every instance instead of a specific feature map. This kernel function in particular considers an instance as the sequence of tokens, lemmas, part-of-speech and named entities. The tokens that refer to each argument are identified, while the label of each instance was 0 if the pair was not a relation, or 1 if it was a relation. Each pair of entities that satisfied the argument type restrictions was considered a candidate pair. This kernel has been applied to biomedical text, for the extraction of relations between proteins (Tikk et al., 2010) and chemical compounds (Segura-Bedmar et al., 2011), obtaining positive results. The shallow linguistic kernel is a composite sequence kernel which uses both a local and global context window, which we set at 3 and 4, respectively. These are the only variable parameters of this kernel.

2.3 Distant supervision

The objective of this experiment was to find relations on PubMed abstracts which could increase the size of the training data, and therefore, improve the performance of the system. First, we retrieved the 10,000 most recent abstracts with the MeSH term "arabidopsis" from PubMed. Using the entity annotations from the gold standard, we trained Condition Random Fields (Lafferty et al., 2001) classifiers to recognize each type of entity on the abstracts. We have previously applied this approach to chemical entities, obtaining a F-measure of 0.847 (Lamurias et al., 2015b). We generated lists of the keywords most used in sentences where a relation is described, for each type of relations. To prevent common words from appearing on those lists, we also generated a list of the most used words on the corpus, and removed those words from each list. Our assumption was that if at least two keywords in the list were mentioned in the sentence, then the relation would be true. Since this approach produced mostly negative instances, we excluded some of those to maintain the same positive/negative ratio as the training data. This approach was based on the work of Thomas et al. (2011), where they used various filters to reduce the number of false positives. In this case, we used only instances of the 10 relations types that were least represented in the gold standard. Table 2.3 provides a comparison between the data set obtained with this technique (DS set) and the training set.

3 Results

To classify the test set, we trained with the documents of the gold standard. We present the results of our official submission, as well as the results obtained with the addition of distant supervised sentences (Table 3). More detailed results, as well as the results obtained by the other teams, are available at the task website [1]. After submitting the results, we found that, by mistake, we had trained the classifiers only with the training set. Therefore, we also present the results obtained with the

[1] http://2016.bionlp-st.org/tasks/seedev/seedev-evaluation

	Pairs		Ratio	
Pair type	DS set	Train	DS set	Train
Binds_To	4624	66	0.0449	0.0134
Composes_Primary_Structure	56	32	0.0003	0.0769
Composes_Protein_Complex	16	15	0.0042	0.1172
Exists_At_Stage	400	17	0.0074	0.0499
Is_Involved_In_Process	136	32	0.0127	0.0371
Occurs_In_Genotype	1312	34	0.0194	0.0804
Occurs_During	112	18	0.0032	0.0625
Regulates_Accumulation	5632	65	0.0112	0.0114
Regulates_Molecule_Activity	1664	16	0.0147	0.0015
Regulates_Tissue_Development	704	18	0.0788	0.0060

Table 1: Number of positive pair (Pairs) and positive/negative ratio (Ratio) for each of the relation types considered for the distant supervision approach. DS set refers to the data set generated using distant supervision while Train refers to the training set.

classifiers trained with both training and development sets.

Training	Recall	Precision	F1
Baseline	0.895	0.029	0.056
Train	0.256	0.379	0.306
Train + Dev	0.304	0.341	0.322
Train + Dev + DS	0.366	0.387	0.377

Table 2: SeeDev-binary test set results. Train refers to the training the classifiers with the training set, Dev to the development set and DS to the distant supervision set generated using distant supervision.

Table 3 also contains a baseline that we used during development of the system, to compare the performance of our system to a simple approach. In this case, the simple approach consisted in classifying every pair that satisfied the entity type requirements as a true relation. As expected, this baseline obtained high recall and low precision and F-measure. The reason why the recall is not 1 is because we only considered pairs of entities from the same sentence. This way, the recall of the baseline (0.895) is the maximum recall we could have obtained with our approach. We observed that with our system, the results obtained were better both in terms of precision and F-measure.

The main difference between training with just the training set and using both training and development was in the recall obtained. By increasing the number of training instances, the classifier was able to correctly identify more relations. Although it also decreased the precision, the difference in terms of F-measure was positive.

Using the distant supervision approach, we were able to use 6947 sentences as an additional data set (DS set). This approach improved the F-measure by 0.055, due to an increase in recall and precision.

4 Discussion

This task was a challenge for our system since it required the identification of 22 types of relations, while previously the system was tested only on one specific type of relation While we could optimize the system for one type of relation with domain knowledge, in this case we had to use a generic approach to various types.

Comparing with the other participants, our F-measure was the 5th best of the 7 participating teams, 0.126 points below the best. In terms of precision, our team was the 3rd best, 0.154 below the best. Our submitted results had higher precision because we used only the gold standard annotations to train the classifiers. This way, the output of the classifiers tended to be closer to the training corpora.

4.1 Error Analysis

In order to fairly compare our results with the other teams, we discuss only the errors of our official submission. There was a wide range of F-measure values within the different types of relations. The types "Has_Sequence_Identical_To" and "Is_Functionally_Equivalent_To" had a F-measure of 0.708 and 0.646, respectively. These

types obtain much higher scores possibly because the entity types of the two arguments had to be the same, reducing the number of candidate pairs. The most difficult relations were the ones less represented in the training data, such as "Is_Involved_In_Process" and "Is_Linked_To". In the case of the first type, no team was able to identify one of the 12 relation instances present in the test corpus, while with the second type, only one team was able to identify some relations. These results show that the performance of the techniques used for this task are dependent on the annotations of the training data.

Regarding the contribution of the distant supervision approach, we observed that the system predicted fewer relations of the less frequent relation types. Since we labeled each pair of entities automatically, it is possible that some relations were mislabeled. However, since we maintained the same positive/negative ratio as the training set (Table 2.3), this approach provided mostly negative instances.

4.2 Future Work

We intend to explore other techniques to use unlabeled data for distant supervision. A technique that has improved results on other domains consists of using a knowledge base to restrict which entities could constitute a relation (Bunescu and Mooney, 2007). By combining the knowledge base with the keyword based filter, we should obtain a set of instances with a high probability of being correctly labeled. These instances should then improve the quality of the classifiers by providing other ways to express a relation, and reduce the number of incorrect annotations.

Another technique to explore consists in applying semantic similarity measures (Couto and Pinto, 2013) to check if two entities are semantically related and therefore could constitute a relation (Lamurias et al., 2015a). Additionally, we intend to apply our distant supervision approach to improve the results of our biomedical question&answering system (WS4A) that participated in the BioASQ 2016 challenge(Rodrigues et al., 2016).

Acknowledgments

This work was supported by the Fundação para a Ciência e a Tecnologia (https://www.fct.mctes.pt/) through the PhD grant PD/BD/106083/2015 and LaSIGE Unit Strategic Project, ref. UID/CEC/00408/2013 (LaSIGE).

References

[Bossy et al.2015] Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec. 2015. Overview of the gene regulation network and the bacteria biotope tasks in bionlp'13 shared task. *BMC bioinformatics*, 16(Suppl 10):S1.

[Bunescu and Mooney2007] Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Annual meeting-association for Computational Linguistics*, volume 45, page 576.

[Chaix et al.2016] Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Delger, Pierre Zweigenbaum, Philippe Bessires, Loc Lepiniec, and Claire Ndellec. 2016. Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task workshop*, Berlin, Germany, August. Association for Computational Linguistics.

[Couto and Pinto2013] Francisco M Couto and H Sofia Pinto. 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of bioinformatics and computational biology*, 11(05):1371001.

[Giuliano et al.2006] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *EACL*, volume 18, pages 401–408. Citeseer.

[Kim et al.2015] Jin-Dong Kim, Jung-jae Kim, Xu Han, and Dietrich Rebholz-Schuhmann. 2015. Extending the evaluation of genia event task toward knowledge base construction and comparison to gene regulation ontology task. *BMC bioinformatics*, 16(Suppl 10):S3.

[Lafferty et al.2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[Lamurias et al.2015a] Andre Lamurias, João D Ferreira, and Francisco M Couto. 2015a. Improving chemical entity recognition through h-index based semantic similarity. *Journal of cheminformatics*, 7(1):1.

[Lamurias et al.2015b] Andre Lamurias, Manuel Lobo, Marta Antunes, Luka A Clarke, and Francisco M Couto. 2015b. Identifying chemical entities in patents using brown clustering and semantic similarity. In *5th BioCreative Challenge Evaluation*.

[Pyysalo et al.2015] Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC bioinformatics*, 16(Suppl 10):S2.

[Rodrigues et al.2016] Miguel J. Rodrigues, Miguel Falé, Andre Lamurias, and Francisco M. Couto. 2016. WS4A: a biomedical question and answering system based on public web services and ontologies. Poster session at the 4th BioASQ Workshop.

[Sætre et al.2007] Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. 2007. AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In *Proceedings of the Second BioCreative Challenge Workshop*, pages 209–212.

[Segura-Bedmar et al.2011] Isabel Segura-Bedmar, Paloma Martinez, and Cesar de Pablo-Sánchez. 2011. Using a shallow linguistic kernel for drug–drug interaction extraction. *Journal of biomedical informatics*, 44(5):789–804.

[Thomas et al.2011] Philippe Thomas, Illés Solt, Roman Klinger, and Ulf Leser. 2011. Learning to extract protein–protein interactions using distant supervision. *ROBUS 2011*, page 25.

[Tikk et al.2010] Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837.

[Toutanova and Manning2000] Kristina Toutanova and Christopher D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.