Detecting Entity Relations as a Supporting Task for Bio-Molecular Event Extraction

Sofie Van Landeghem[1,2], **Thomas Abeel**[1,2,3], **Bernard De Baets**[4] **and Yves Van de Peer**[1,2]
1. Dept. of Plant Systems Biology, VIB, Belgium
2. Dept. of Plant Biotechnology and Genetics, Ghent University, Belgium
3. Broad Institute of MIT and Harvard, Cambridge, MA, USA
4. Dept. of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium
yves.vandepeer@psb.ugent.be

Abstract

Recently, the focus in the BioNLP domain has shifted from binary relations to more expressive event representations, largely owing to the international popularity of the BioNLP Shared Task (ST) of 2009. This year, the ST'11 provides a further generalization on three key aspects: text type, subject domain, and targeted event types. One of the supporting tasks established to provide more fine-grained text predictions is the extraction of entity relations. We have implemented an extraction system for such non-causal relations between named entities and domain terms, applying semantic spaces and machine learning techniques. Our system ranks second of four participating teams, achieving 37.04% precision, 47.48% recall and 41.62% F-score.

1 Introduction

Understanding complex noun phrases with embedded gene symbols is crucial for a correct interpretation of text mining results (Van Landeghem et al., 2010). Such non-causal relations between a noun phrase and its embedded gene symbol are referred to as *entity relations*. As a supporting task for the BioNLP ST'11, we have studied two types of such entity relations: Subunit-Complex and Protein-Component. These relationships may occur within a single noun phrase, but also between two different noun phrases. A few examples are listed in Table 1; more details on the datasets and definitions of entity relations can be found in (Pyysalo et al., 2011).

Valid entity relations involve one GGP (gene or gene product) and one domain term (e.g. "promoter") and they always occur within a single sentence. In the first step towards classification of entity relations, we have calculated the semantic similarity between domain terms (Section 2). Supervised learning techniques are then applied to select sentences likely to contain entity relations (Section 3). Finally, domain terms are identified with a novel rule-based system and linked to the corresponding GGP in the sentence (Section 4).

2 Semantic analysis

To fully understand the relationship between a GGP and a domain term, it is necessary to account for synonyms and lexical variants. We have implemented two strategies to capture this textual variation, grouping semantically similar words together.

The first method takes advantage of manual annotations of semantic categories in the GENIA event corpus. This corpus contains manual annotation of various domain terms such as promoters, complexes and other biological entities in 1000 PubMed articles (Kim et al., 2008).

The second method relies on statistical properties of nearly 15.000 articles, collected by searching PubMed articles involving *human transcription factor blood cells*. From these articles, we have then calculated a semantic space using latent semantic analysis (LSA) as implemented by the S-Space Package (Jurgens and Stevens, 2010). The algorithm results in high-dimensional vectors that represent word contexts, and similar vectors then refer to semantically similar words. We have applied the Markov Cluster algorithm (MCL) (van Dongen, 2000) to group semantically similar terms together.

Proceedings of BioNLP Shared Task 2011 Workshop, pages 147–148,
Portland, Oregon, USA, 24 June, 2011.

Type of relation	Examples
Subunit-Complex	"the c-fos content of [AP-1]" / "c-jun, a component of the transcription factor [AP-1]"
Protein-Component	"the [IL-3 promoter]" / "the activating [ARRE-1 site] in the IL-2 promoter"

Table 1: Examples of entity relations. GGPs are underlined and domain terms are delimited by square brackets.

3 Machine learning framework

Our framework tries to define for each GGP in the data whether it is part of any of the two entity relations, by analysing the sentence context. To capture the lexical information for each sentence, we have derived bag-of-word features. In addition, 2- and 3-grams were extracted from the sentence. Finally, the content of the gene symbol was also used as lexical information. All lexical information in the feature vectors has undergone generalization by blinding the gene symbol with "protx" and all other co-occurring gene symbols with "exprotx". Furthermore, terms occurring in the semantic lexicons described in Section 2 were mapped to the corresponding cluster number or category. For each generalization, a blinded and a non-blinded variant is included in the feature vector.

Dependency graphs were further analysed for the extraction of grammatical patterns consisting of two nodes (word tokens) and their intermediate edge (grammatical relation). For the nodes, the same generalization rules as in the previous paragraph are applied. Finally, similar patterns are generated with the nodes represented by their part-of-speech tag.

The final feature vectors, representing sentences with exactly one tagged gene symbol, are classified using an SVM with a radial basis function as kernel. An optimal parameter setting (C and $gamma$) for this kernel was obtained by 5-fold cross-validation on the training data.

4 Entity detection

Once a sentence with a gene symbol is classified as containing a certain type of entity relation, it is necessary to find the exact domain term that is related to that gene symbol. To this end, we have designed a pattern matching algorithm that searches within a given window (number of tokens) around the gene symbol. The window size is increased to a predefined maximum as long as a maximal number of domain terms was not found.

Within the search window, a rule-based algorithm decides whether a given token qualifies as a relevant domain term, employing first a high-precision dictionary and then high-recall dictionaries.

5 Results

Our system achieves a global performance of 37.04% precision, 47.48% recall and 41.62% F-score, coming in second place after the university of Turku who obtained an F-score of 57.71%, and ranking before Concordia University who scores 32.04%. It remains an open question why the final results of the top ranked systems differ so much.

Acknowledgments

SVL and TA would like to thank the Research Foundation Flanders (FWO) for funding their research. TA is a post doctoral fellow of the Belgian American Education Foundation. The authors thank Jari Björne for his help with the manuscript.

References

David Jurgens and Keith Stevens. 2010. The S-Space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 30–35.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. 2011. Overview of the Entity Relations (REL) supporting task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, June.

Stijn van Dongen. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

Sofie Van Landeghem, Sampo Pyysalo, Tomoko Ohta, and Yves Van de Peer. 2010. Integration of static relations to enhance event extraction from text. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 144–152.