Semantic role labeling of gene regulation events: preliminary results

Roser Morante
CLiPS - University of Antwerp
Prinsstraat 13, B-2000 Antwerpen, Belgium
Roser.Morante@ua.ac.be

Abstract

This abstract describes work in progress on semantic role labeling of gene regulation events. We present preliminary results of a supervised semantic role labeler that has been trained and tested on the GREC corpus.

1 Introduction

Semantic role labeling (SRL) is a natural language processing task that consists of identifying the arguments of predicates within a sentence and assigning a semantic role to them. This task can support the extraction of relations from biomedical texts. Recent research has produced a rich variety of SRL systems to process general domain corpora. However, only a few systems have been developed to process biomedical corpora (Tzong-Han Tsai et al, 2007; Bethard et al., 2008). In this abstract, we present preliminary results of a new system that is trained on the GREC corpus (Thompson et al., 2009).

The GREC corpus consists of 240 MEDLINE abstracts, in which gene regulation events have been annotated with different types of information, like the span of the event and of its arguments, and the semantic role of the arguments. Events can be verbs (58%) and nominalised verbs (42%). The corpus is divided into two species-specific subcorpora: *E. coli* (167 abstracts, 2394 events) and human (73 abstracts, 673 events).

2 System description

We perform two preprocessing steps. First, we extract the text and parse it with the GDep parser (Sagae and Tsujii, 2007) and then we convert the corpus from xml into CoNLL format. Table 1 shows a preprocessed sentence. The system performs argument identification and semantic role assignment in a single step, assuming gold standard event identification. It consists of one classifier that classifies an instance into one of the semantic role classes or the NONE class. An instance represents a combination of an event and a potential argument (PA). In order to generate the PAs, the system relies on information from the dependency syntax tree, which means that errors in the syntactic tree influence directly the performance of the system. We consider that the following tokens or combinations of tokens can be PAs: main verbs, nouns, adjectives, pronouns and adverbs; main verbs, nouns, adjectives, pronouns and adverbs with their modifiers to the left in the string of words; main verbs, nouns, adjectives, pronouns, adverbs, prepositions and relative pronouns with their modifiers to the left and to the right in the string of words.

The features extracted to perform the classification task are the following:

- About the event and the PA: chain of words, lemmas, POS, and dependency labels of all the tokens; lemma, POS and dependency label of head token, first token and last token; lemma and POS of syntactic father of head; lemma, POS, and dependency label of previous and next three tokens in the string of words; even type.
- About the dependency tree: feature indicating who is the ancestor (event, PA, other); lemma, POS, and dependency label of the first common ancestor of event and PA, if there is one; chain of dependency labels and chain of POS from event to common ancestor, and from PA to common ancestor, if there is one; chain of dependency labels and chain of POS from PA to event, if event is ancestor of PA; chain of dependency labels and chain of POS from event to PA, if PA is ancestor of event; chain of dependency labels and POS from event to ROOT and from PA to ROOT.
- Normalised distance in number of tokens between event an potential argument in the string of words.

We use an IB1 memory–based algorithm as implemented in TiMBL (version 6.1.2) [1](Daelemans et al., 2009), a memory-based classifier based on the k-nearest neighbor rule. The IB1 algorithm was parameterised by using Jeffrey divergence as the similarity metric, gain ratio for feature weighting, using 5 k-nearest neighbors, and weighting

[1]TiMBL: http://ilk.uvt.nl/timbl

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 126–127,
Uppsala, Sweden, 15 July 2010.

#	WORD	LEMMA	CHUNK	POS	DEP	LABEL	#E	TYPE	ROLES		
1	Lrp	Lrp	B-NP	NN	2	SUB	_	_	B-Agent	B-Agent	B-Agent
2	binds	bind	B-VP	VBZ	0	ROOT	E1	GRE	_	_	_
3	to	to	B-PP	TO	2	VMOD	_	_	_	_	_
4	two	two	B-NP	CD	5	NMOD	_	_	_	_	_
5	regions	region	I-NP	NNS	3	PMOD	_	_	_	_	_
6	in	in	B-PP	IN	5	NMOD	_	_	_	_	_
7	the	the	B-NP	DT	10	NMOD	_	_	B-Destination	_	_
8	dadAX	dadAX	I-NP	NN	10	NMOD	_	_	I-Destination	_	_
9	promoter	promoter	I-NP	NN	10	NMOD	_	_	I-Destination	_	_
10	region	region	I-NP	NN	6	PMOD	_	_	I-Destination	_	_
11	of	of	B-PP	IN	10	NMOD	_	_	_	_	_
12	Escherichia	Escherichia	B-NP	FW	13	NMOD	_	_	_	_	_
13	coli	coli	I-NP	FW	11	PMOD	_	_	_	_	_
14	to	to	B-VP	TO	15	VMOD	_	_	_	_	_
15	repress	repress	I-VP	VB	13	NMOD	E2	Gene_Repression	_	_	_
16	and	and	I-VP	CC	15	VMOD	_	_	_	_	_
17	activate	activate	I-VP	VB	15	VMOD	E3	Gene_Activation	_	_	_
18	transcription	transcription	B-NP	NN	17	OBJ	_	_	_	B-Theme	B-Theme
19	directly	directly	B-ADVP	RB	17	VMOD	_	_	_	B-Manner	B-Manner
20	.	.	O	.	2	P	_	_	_	_	_

Table 1: Sentence 1 from abstract 10216857 in E. coli corpus. Column # contains the token number; WORD, the word; LEMMA to LABEL contain information provided by the GDEP parser; #E, the event number; TYPE, the type of event, and ROLES contains columns with argument labels for each event following textual order, i.e., the first column corresponds to the first event in #E, the second column to the second event, etc.

the class vote of neighbors as a function of their inverse distance.

3 Preliminary results

We provide 5 fold cross-validation (CV) and cross-domain (CD) results in Table 2. The CV results are obtained by training and testing on different partitions of the same corpus. The CD results are obtained by training on one corpus and testing on the other. Although we cannot directly compare this results with results of other systems on exactly the same corpus, Sasaki et al. (2008) report CV results on a corpus of 677 MEDLINE abstracts on E. Coli gene regulation events. The precision achieved by their system is 49.00 and the recall 18.60. We consider that the results of our system are encouraging to proceed with further research.

Corpus	Precision	Recall	F1
E coli CV	59.72	32.29	41.92
E coli CD	49.87	18.07	26.53
Human CV	47.98	22.43	30.57
Human CD	56.57	25.90	35.53

Table 2: F1, precision and recall for argument identification and labeling.

4 Future work

Future work will deal with incorporating domain specific knowledge and with improving the machine learning techniques. We will experiment with other algorithms, like Conditional Random Fields, which are well known sequence labelers. Additionally, we will implement also a constraint satisfaction algorithm.

Acknowledgments

This preliminary study was made possible through financial support from the University of Antwerp (GOA project BIOGRAPH).

References

S. Bethard, Z. Lu, J.H. Martin, and L. Hunter. 2008. Semantic role labeling for protein transport predicates. *BMC Bioinformatics*, 9:277.

W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2009. TiMBL: Tilburg memory based learner, version 6.2, reference guide. Technical Report Series 09-01, ILK, Tilburg, The Netherlands.

K. Sagae and J. Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proc. of CoNLL 2007: Shared Task*, pages 82–94, Prague, Czech Republic.

Y. Sasaki, P. Thompson, Ph. Cotter, J. McNaught, and S. Ananiadou. 2008. Event frame extraction based on a gene regulation corpus. In *Proc. of Coling 2008*, pages 761–768, Manchester, UK.

P. Thompson, S. A Igbal, J. McNaught, and S. Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349.

R. Tzong-Han Tsai et al. 2007. BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, 8:325.