Biomedical Event Detection using Rules, Conditional Random Fields and Parse Tree Distances

Farzaneh Sarafraz*, James Eales*, Reza Mohammadi♦, Jonathan Dickerson♣, David Robertson♣, Goran Nenadic*

*School of Computer Science, University of Manchester
♣Faulty of Life Sciences, University of Manchester
♦Dept. of Mathematics and Computer Science, Sharif University of Technology
sarafraf@cs.man.ac.uk, g.nenadic@manchester.ac.uk

Abstract

This paper reports on a system developed for the BioNLP'09 shared task on detection and characterisation of biomedical events. Event triggers and types were recognised using a conditional random field classifier and a set of rules, while event participants were identified using a rule-based system that relied on relative distances between candidate entities and the trigger in the associated parse tree. The results on previously unseen test data were encouraging: for non-regulatory events, the F-score was almost 50% (with precision above 60%), with the overall F-score of around 30% (49% precision). The performance on more complex regulatory events was poor (F-measure of 7%). Among the 24 teams submitting the test results, our results were ranked 12th for the overall F-score and 8th for the F-score of non-regulation events.

1 Introduction

The aim of the BioNLP'09 shared task 1 was to characterise molecular events being reported in a Medline abstract by identifying the textual trigger, event type and participating entities (Kim et al. 2009). Nine event types were considered: *gene expression*, *transcription*, *protein catabolism*, *localisation*, *phosporylation*, *binding*, *regulation*, *positive regulation*, and *negative regulation*. Depending on the event type, the task included the identification of either one (for the first five event types mentioned above) or more (e.g. for *binding*) participating proteins. Information requested for regulatory events was more complex: in addition to one theme (a protein or another event), these events could also have a cause (a protein or another event) that needed to be identified.

The organisers have distributed a training dataset of 800 abstracts, with gene and gene product mentions pre-annotated in text. In addition, a development set (150 abstracts) was provided to assess the quality of the extractions during the training and development phases.

2 Methods

The system developed for the challenge consists of three main modules: (1) event trigger and type detection, (2) event participant detection, and (3) post-processing of the results.

2.1 Event Trigger and Type Detection

Our view of the event trigger and type detection subtask was that each token in a sentence needed to be tagged either as a trigger for one of the nine event types, or as a non-trigger/event token. We therefore decided to identify event types and triggers in a single step by training a conditional random field (CRF) classifier that assigned one of ten (nine types plus non-trigger) tags to each token. CRFs have been shown to be particularly suitable for tagging sequential data such as natural language text, because they take into account features and tags of neighbouring tokens when evaluating the probability of a tag for a given token.

Tokens and their part-of-speech (POS) tags were recognised using the Genia Tagger (Tsuruoka et al. 2005). Each stemmed token was represented using a feature vector consisting of the following features:

- A binary feature indicating whether the token is a protein;
- A binary feature indicating whether the token is a known protein-protein interaction word (we used a pre-complied dictionary of

Proceedings of the Workshop on BioNLP: Shared Task, pages 115–118,
Boulder, Colorado, June 2009.

such words collected from previous studies (Fu et al. 2008; Yang et al. 2008);
- The token's POS tag;
- The log-frequencies of the token being a trigger for each event type in the training data (nine features);
- The number of proteins in the given sentence.

Other features (e.g. separating the known interaction words according to the nine event types) were explored during the development phase, but were not included in the final feature list since they increased the sparseness of the data and did not improve the overall results. The CRF parameters were adjusted for maximum performance, including the choice of training algorithms, the number of training steps, the size of the window within which the tokens can affect any certain token, and the number of training abstracts used in each training step. It was interesting to notice that there were no significant improvements in the performance after training on 100, 400 or 800 abstracts from the training set (data not shown).

2.2 Locating Event Themes

After detecting potential triggers and associated event types, the next task was to locate possible participants (i.e. 'themes' and 'causes') for each event. It was obvious that participants did not have to be the nearest to the trigger on the surface level, so our approach was based on distances within the parse trees associated with the sentences containing candidate events. Parse tree distances have been studied previously in clustering and automatic translation tasks (Emms 2008), so we hypothesised that we could use them to identify the most likely participants. The training data was analysed for the proximities between the triggers and the (correct) event participants in the parse tree of the sentence.[1] Figure 1 gives a detailed density function of these distances (ignoring non-protein nodes). The analysis showed that a theme was usually amongst the nearest proteins to the trigger in terms of parse tree distances: for example, in 60% of all single theme events (e.g. *localisation*, *phosphorylation*) the correct protein participant was the trigger's nearest or second nearest protein in the parse tree. A further analysis demonstrated that it was more likely for a theme to appear in the sub-tree of the corresponding trigger, with 70% of all single theme events having a theme which appeared in the sub-tree of the trigger. Furthermore, specific analyses of the parse trees associated to the *binding* events (which can have more than one theme) suggested a linear relationship between the parse tree distance and binding event participant number ($participant_1$ is the nearest, $participant_2$ is the second nearest, etc.).

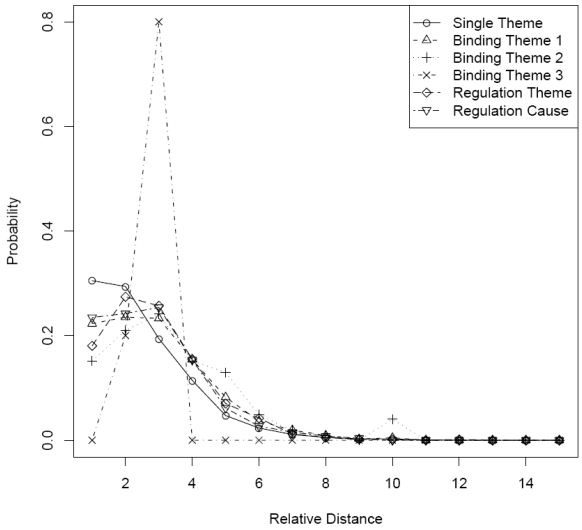

Figure 1: Probability density function of the distance between the trigger and the theme in the parse tree (ignoring the tokens that are not proteins)

We used this distributional analysis (derived from the training data) to design a rule-based method for the identification of participating themes. The rules were manually derived for each of the nine event classes, by defining:

- a threshold for the maximum distance to the trigger in the sub-tree for the given event type;
- a threshold for the difference between the maximum distance in the whole tree and the given sub-tree for the given event type;
- the number of nearest proteins to be reported for each trigger.

All entities that satisfied a distance-based rule for a given trigger were selected as the corresponding theme(s). For example, if the event type is *binding*, then up to the second closest protein in the sub-tree, and the first closest protein in the rest of the tree are reported as themes.

[1] The parse trees were produced by the GDep parser (Sagae and Tsujii, 2007) and supplied by the challenge organisers.

Figure 2 provides an example of the method applied to a sentence with multiple events. *Regulates* and *secretion* are correctly identified as triggers for a regulation and a localization event in the first phase. Using the rules for localization, the themes for two localization events are correctly recognised as proteins *T2* and *T3*, whereas *T1* was ignored since it did not appear in the trigger's sub-tree.

Engineering and applying rules for non-regulatory events was relatively straightforward. However, regulatory events can have different kinds of participants (a protein or an event). In the case of an event, we were trying to locate the nearest trigger for the event (being regulated) in the parse tree. For example, in Figure 2, the nearest option to the regulation trigger (*secretion*) was the trigger of the two localization events, and both events should be (correctly) reported as the themes of two regulation events. Therefore, we require a number of recursions in the application of the rules to represent higher-order regulatory dependences. For the purposes of this challenge, only regulations up to the second "order" were detected, allowing other events to act as themes and causes as well as proteins. Attempts to find more complicated regulatory events using this method resulted in a decreased precision and/or F-score.

2.3 Post-processing Event Profiles

The performance of the first two phases was studied on the development dataset: we noted a number of false-positive and false-negative results that were mostly due to a set of recurring triggers. We therefore decided to perform a post-processing step to improve the identification of event triggers and associated types. In the first step (improving the event trigger and type detection), the output of the CRF was overridden in cases where the triggers appeared in a list of negatively discriminated trigger words which was collected after the manual analysis of the false positive results on the training and development data. Similarly, in cases where the CRF missed a highly indicative trigger (from a manually collected set) for a given event type, the trigger was added as part of post-processing. In the latter case, the sentence was then processed for the event theme detection (as described in 2.2).

In the second step of the pre-processing phase, we forced highly indicative regulation triggers (if not previously identified) to be associated with an

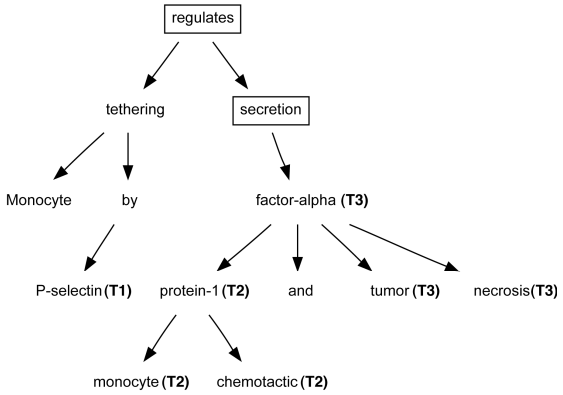

Figure 2: The parse tree of sentence *Monocyte tethering by P-selectin regulates monocyte chemotactic protein-1 and tumor necrosis factor-alpha secretion.* The triggers are shown in boxes, and the entities are numbered.

event by assigning proteins appearing in the sentence to them, even when no protein in the sentence satisfied the theme or cause criteria described in Section 2.2. This was aimed at improving the extremely low recall for regulatory events.

Finally, since triggers could consist of more than one consecutive token, a set of simple rules were applied to remove typical false-negative constituents identified by the CRF as part of triggers (e.g. sometimes linking words appeared within triggers).

3 Results and discussion

The task 1 assessment was based on the output of the system when applied to the test dataset of 260 previously unseen abstracts. An event was counted as a true positive if its type, trigger and all participants had been correctly identified. The overall F-score for our system was 30.35% with 48.61% precision (approximate span matching, see Table 1). The best performing event types were *phosphorylation* (the best F-score and the best recall) and *gene expression* (the best precision with a reasonably good F-measure). While the results for non-regulatory events were encouraging, they were low for regulatory events. Among the 24 teams submitting the test results, our results were ranked 12th for the overall F-score and 8th for the F-score of non-regulation events.

A preliminary analysis of the results was performed on the development data (as the test data is not available), which had around 5% higher overall F-score than the test data (9% for non-regulation events, see Table 2 for details).

Event Class	#Gold	R	P	F-score
Localisation	174	44.83	53.06	48.60
Binding	347	12.68	40.37	19.30
Gene expression	722	52.63	**69.34**	59.84
Transcription	137	15.33	67.74	25.00
Protein catabolism	14	42.86	50.00	46.15
Phosphorylation	135	**78.52**	53.81	**63.86**
Non-reg total	1529	41.53	60.82	49.36
Regulation	291	3.09	19.15	5.33
Positive regulation	983	1.12	8.87	1.99
Neg. regulation	379	12.4	20.52	15.46
Regulatory total	1653	4.05	16.75	6.53
All total	3182	22.06	48.61	30.35

Table 1: Evaluation of the test data (260 abstracts), (approximate span matching; #Gold = the number of examples in the gold standard)

Event Class	#Gold	R	P	F-score
Localisation	53	67.92	46.75	55.38
Binding	312	21.47	63.81	32.13
Gene expression	356	64.61	76.33	69.98
Transcription	82	53.66	**89.80**	67.18
Protein catabolism	21	90.48	67.86	**77.55**
Phosphorylation	47	**91.49**	53.09	67.19
Non-reg total	871	50.4	68.44	58.05
Regulation	172	5.23	33.33	9.05
Positive regulation	632	3.48	21.36	5.99
Neg. regulation	201	9.45	15.08	11.62
Regulatory total	1005	4.98	19.53	7.93
All total	1876	26.07	54.46	35.26

Table 2: Evaluation of the development data (150 abstracts) (approximate span matching; #Gold as in Table 1)

In order to assess the effects of different steps in our approach, we evaluated the performance of the event trigger and event participant detection steps separately. The results presented in Table 3 indicated that the performance of the CRF module was not much better than the overall performance of the system (an F-score of 43% vs. 35%), suggesting that the CRF part was mostly responsible for the errors, by both missing triggers and falsely reporting them. This was particularly the case with non-regulatory events (even for binding). Conversely, when considering only those events whose triggers were correctly identified, their participants were also correctly recognised in most cases. Overall, the analysis suggested that the parse tree distance method performed reasonable well, despite a reduction in recall of approximately 12%.

There are a number of possibilities for improvements. We believe applying the CRF model in two stages would be a better approach to detect events: first identify triggers and then link them to event classes. In addition, the rules employed for determining themes need to be more specific to reflect both event type and grammatical structure. In the case of regulatory events, however, significantly better results were noticed in the trigger detection part when compared to the overall scores, indicating that it was difficult to identify regulatory participants, as any of those participants could be either a protein or another event.

Overall, the results achieved by our system suggest that combining parse tree results, rules and CRFs is a promising approach for the identification of non-regulatory events in the literature, while more work would be needed for regulatory events.

Event Class	#Gold	R	P	F-score
Localisation	40	77.50	47.69	59.05
Binding	180	33.33	54.55	41.38
Gene expression	282	76.60	58.54	66.36
Transcription	68	58.82	18.60	28.27
Protein catabolism	19	84.21	**88.89**	86.49
Phosphorylation	40	**97.50**	81.25	**88.64**
Non-reg total	629	63.91	48.73	55.30
Regulation	138	13.04	62.07	21.56
Positive regulation	462	13.85	54.24	22.07
Neg. regulation	153	29.41	45.92	35.86
All total	1382	38.28	49.44	43.15

Table 3: Trigger-only evaluation of the development data

References

Emms M. 2008. *Tree-distance and some other Variants of evalb*. Proc. of LREC 2008, pp 1373-1379.

Fu W. *et al.* 2008. *Human Immunodeficiency Virus type 1, Human protein interaction database at NCBI*, Nucleic Acid Research 2008, D417-D422

Kim JD *et al.* 2009. *Overview of BioNLP'09 Shared Task on Event Extraction***,** Proc. of BioNLP NAACL 2009 Workshop (to appear)

Sagae K, Tsujii J. 2007. *Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles*. Proc. of the CoNLL 2007 Shared Task, 1044-50

Tsuruoka Y. *et al.* 2005. *Developing a Robust Part of-Speech Tagger for Biomedical Text*. Advances in Informatics, 382–392.

Yang H. et al. 2008. *Identification of Transcription Factor Contexts in Literature using Machine Learning Approaches*. BMC Bioinformatics, Vol. 9(3):S11.