Identification of Negated Regulation Events in the Literature: Exploring the Feature Space

Farzaneh Sarafraz[1], Goran Nenadic[1, 2]

[1]School of Computer Science, University of Manchester, Manchester, UK

[2]Manchester Interdisciplinary BioCentre, University of Manchester, UK

Email addresses: FS: sarafraf@cs.man.ac.uk; GN: g.nenadic@manchester.ac.uk

Abstract

Background. Regulation events are of critical importance to researchers trying to understand processes in living beings. These events are naturally complex and can involve both individual molecular entities and other biomedical events. Of equal importance is the ability to capture statements that refer to regulation events that do not take place. In this paper we explore the identification of negated regulation events in the literature using a number of features**.**

Results. We construe the problem as a classification task and apply support vector machines that use lexical, syntactic and semantic features associated with sentences that represent events. Lexical features include negation cues, part-of-speech tagging and surface distances, whereas syntactic features are engineered from constituency parse trees, the *command* relation between constituents and parse-tree distances. Semantic features include event sub-type and participant types. On a test dataset, best precision has been achieved by combing all features, while ignoring surface-level distances resulted in best recall. Overall, the best F-measure was 54%.

Conclusions. Syntactic features proved to be useful for improving recall, whereas semantic features proved useful for improving precision, demonstrating the potential and limits of task-specific feature engineering to negation detection. Contrasting statements are used frequently to express negated events and many false negatives were due to not capturing those events.

Background

Several efforts have been recently initiated in the text mining community that focus on the extraction of structured information about biomedical relations and events, including protein-protein interactions, gene expression, etc. [1, 2]. These efforts aim for both supporting data consolidation (population of curated databases) and knowledge exploration (e.g. hypothesis generation) [3, 4].

A topic that has been of particular interest in biology and medicine is the investigation of gene regulatory networks, which are of critical importance to researchers trying to understand regulatory mechanisms in living beings. There have been a number of databases developed to store knowledge about gene regulation in various model organisms (e.g. RegulonDB with regulation information in E. coli [5]), but populating such databases proved to be challenging given the pace of publications and complexity of the events. Regulatory events are particularly complex as their participants can be either entities (e.g. a protein) or other events. Therefore, regulation events can be recursively nested, and – given that regulations can be positive (facilitating a particular process) or negative (inhibiting a particular event) – they typically require complex linguistic expressions to report and explain regulation findings. In addition to affirmative findings, a number of events are also reported as negated (e.g. *However, NFATc.beta neither bound to the kappa3 element (an NFAT-binding site) in the tumor necrosis factor-alpha promoter nor activated the tumor necrosis factor-alpha promoter in cotransfection assays*). Detection of

negated events is of particular importance, as it affects the quality and the semantics of the extracted information. In recent years, several challenges and shared tasks have included the extraction of negations (e.g. the BioNLP'09 Shared Task 3 [2]).

In this paper we explore various features that can be used for identification of negated regulation events in a machine learning (ML) method. We examine features mainly engineered from a sentence parse tree with associated lexical and semantic cues.

Methods

Our target events are regulatory processes and causal relations between different biomedical entities and processes [2]. Each regulatory event expressed in text is identified by:

- **regulation type** – we consider three regulation sub-types: positive regulation and negative regulation, in addition to regulation events where there is no indication if it is positive or negative (e.g. *Involvement of mitogen-activated protein kinase pathways in interleukin-8 production by human monocytes*);
- **regulation theme** represents an entity and event that is regulated;
- **regulation cause** – a protein or event that causes regulation;
- **event trigger** – a token(s) that indicates presence of the event in the associated sentence.

Identification of these components is a challenging text mining task and has been discussed widely [2, 6, 7]. In order to focus on exploration of the complexity of negations, unaffected by automatic named entity recognition, event trigger detection, participant identification, etc., we use pre-identified events: given a sentence that describes an event, we assume that all event features (listed above) have been identified. Then, we construe the negation detection problem as a classification task: the aim is to classify the event as affirmative or negative. The method is based on features engineered from an event-representing sentence as follows.

Lexical features are based on a list of negation cues[1] and part-of-speech (POS) tagging of the associated sentence. We also consider the surface distance between the negation cue and trigger, theme and cause. More precisely, the lexical features include:

1. Whether the sentence contains a negation cue from the cue list;
2. The negation cue itself (if present);
3. The POS tag of the negation cue;
4. The POS tag of the trigger;
5. The POS tag of the theme; if the theme is another event, the POS tag of the trigger of that event is used;
6. The POS tag of the cause; if the cause is another event, the POS tag of the trigger of that event is used;
7. Surface distance between the trigger and cue;
8. Surface distance between the theme and cue;
9. Surface distance between the cause and cue;

Syntactic features are based on the results of constituency parsing of the associated sentence and the *command* relation. The concept of *command* has been introduced by Langacker in order to determine the scope within a sentence affected by an element [8]. If *a* and *b* are nodes in the constituency parse tree of a sentence, then *a* X-commands *b* iff the lowest ancestor of *a* with label X is also an ancestor of *b.* Langacker observed that when *a* S-commands *b*, then *a* affects the scope containing *b.* We hypothesise that if a negation cue X-commands an event trigger *or* participant, then the associated event is negated. We explored various types of X-command, including S-command (for sentence or sub-clause), NP-command (noun phrase), VP-command (verb phrase), PP-command (prepositional phrase), etc. We also consider the distance within the tree. Specifically, the syntactic features include:

---

[1] Negation cues include generic expressions (e.g. *no, not, none,* etc.) as well as 18 task-specific words (e.g. *unchanged, impaired, little, independent, except, exception*, etc.).

10. The type of the lowest common ancestor of the trigger and the cue (either S, VP, PP, NP, JJ or PP);
11. Whether or not the negation cue X-commands the trigger (X is S, VP, NP, JJ, PP)
12. Whether or not the negation cue X-commands the theme (X is S, VP, NP, JJ, PP)
13. Whether or not the negation cue X-commands the cause (X is S, VP, NP, JJ, PP)
14. The parse-tree distance between the event trigger and the negation cue.
15. The parse-tree distance between the theme and the negation cue.
16. The parse-tree distance between the cause and the negation cue.

Semantic features introduce known characteristics of the regulation participants and the sub-type of regulation (if known):

17. Regulation sub-type (positive, negative, none);
18. Theme type, which can be either a *protein* or one of the nine event types as defined in BioNLP'09: gene expression, transcription, protein catabolism, localization, phosphorylation, binding, regulation, positive regulation, and negative regulation;
19. Cause type is defined analogously to the theme type.

The above features have been used to train a series of binary SVM (support vector machine) classifiers that aim to identify negated regulation events. The standard metrics (precision, recall and F1-measure) were used to evaluate the results, where a true positive represents a correctly identified negated event; a false negative is a negated event reported incorrectly as affirmative.

Results

The data used in this study is provided by the BioNLP'09 challenge [2]. The training set contained a total of 4,870 regulation events, with 440 of these reported as negated; the test set contained 987 regulation events, of which 66 are negated. The associated sentences are annotated with event types (the nine types as specified above), textual trigger and participants. In addition, every event has been tagged as either affirmative (reporting a specific interaction) or negative (reporting that a specific interaction has not been observed). The training data was used for modelling and all the results refer to the methods applied on the development dataset using 10-cross validation. Constituency parse trees were produced the method described by [9].

Impact of lexical features. The results of using lexical features only are presented in Table 1. As expected, surface distances to the negation cue are not good indicators, and do not improve the performance of standard lexical and POS features – on the contrary, they reduce precision. Overall, precision is relatively high but recall is low.

Lexical features	**Precision**	**Recall**	**F1**
Features 1-6 (no surface distances)	75.00	22.73	34.88
All lexical features	71.43	22.73	34.48

Table 1. The results of using lexical features only

Impact of syntactic features. The results of using syntactic features only are presented in Table 2. As opposed to surface distances, parse-tree distances are more suitable features, improving the overall performance significantly (F1 improving from 11% to 36%). There were no significant differences in performance when different types of X-command relations are used (data not shown) – focusing only on S- and VP-command provides the same levels of accuracy.

Syntactic features	**Precision**	**Recall**	**F1**
Features 10-13 (no parse-tree distances)	80.00	6.06	11.27
All syntactic features	60.71	25.76	36.17

Table 2. The results of using syntactic features only

Impact of semantic features. Although there are no significant differences in the results when lexical or syntactic features are used, semantic features on their own resulted in very low performances, virtually missing all negated regulatory events (data not shown).

Combining features. Table 3 shows the results when features of various types are combined. Combining several feature types (lexical, syntactic and semantic) proved to be beneficial. Surface distances still reduce the overall precision, but overall improve recall. It is interesting that adding semantic features (which characterise the participants involved in the regulation) significantly improves precision (by 20% when compared to the lexical and syntactic feature sets). On the other hand, command relations improve recall (by almost 20%).

Features	Precision	Recall	F1
Lexical + syntactic	66.67	39.39	49.52
Lexical + semantic	50.00	15.15	23.26
Syntactic + semantic	72.22	19.70	30.95
All with no surface distances	73.68	**42.42**	**53.85**
All with no X-command on theme and cause	78.12	37.88	51.02
All features	**78.79**	39.39	52.53

Table 3. The results of combining different features

We note that some feature subsets (e.g. features 10-13, Table 2) do not provide a balance between precision and recall; depending on the application, the classification threshold could be adjusted to produce higher recall or precision.

Discussion

Several approaches have recently been suggested for the extraction of negated biomedical events. Kilicoglu and Bergler [6] and Hakenberg et al. [10] used a number of heuristic rules concerning the type of the negation cue and the type of the dependency relation. The former were the best performing negation detection approach in the BioNLP'09 shared task and reported recall of up to 15%, but with overall event detection sensitivity of 33% on a 'test' dataset different from that used in this study, and including simpler, non-regulatory events [6]. MacKinlay and colleagues also use ML, assigning a vector of complex deep parse features to every event trigger [7]. They achieved an F-score of 36% on a dataset containing both nested regulatory event and simpler events. Morante and Daelemans used machine learning to detect the negation scope in biomedical text, but have not separately addressed what could negate a biomedical event [11]. We have explored not only negation of triggers but also phrases in which regulation theme and cause have been negated (consider, for example, "*SLP-76*" in sentence "*In contrast, Grb2 can be coimmunoprecipitated with Sos1 and Sos2 but not with SLP-76*"). These have resulted in a slight improvement of both precision and recall.

Analysing the errors more closely shows that one of the recurring patterns contributing to a large portion of the false negative results were the contrasting patterns. It often happens that the authors express contrasting observations by describing one event and implying the other is the opposite. For example, consider this sentence

> *Unlike TNFR1, LMP1 can interact directly with receptor-interacting protein (RIP) and stably associates with RIP in EBV-transformed lymphoblastoid cell lines.*

In this example, a negated interaction is expressed, but there is no sign of a negation cue or negative sentence structure. Still, we can infer that TNFR1 *cannot* interact directly with RIP; it may also imply that TNFR1 *does not* stably associate with RIP in certain cell lines. The negation therefore can only be inferred by taking the following steps:

1. recognising the presence of a contrasting pattern in the sentence;
2. identifying the contrasting entities (in this example TNFR1 and LMP1);
3. extracting the explicitly stated event (*LMP1 interacts with RIP* in this case);
4. identifying the scope of contrast; this can be ambiguous, as in the above example it is not clear whether the two entities also contrast in "*stably associates with RIP*", or only in "*interact directly with RIP*".

Contrasting patterns are not uncommon. There are 125 phrases expressing contrast in the training data (in 800 abstracts) and 32 in the development data (150 abstracts) using only the

patterns "*unlike A, B*", "*B, unlike A*", and "*A; in contrast B*". In these cases, the negation is usually not linguistically explicit, and has to be inferred by analysing the contrasts. The future work will explore a *rule-based* framework that would identify contrasting patterns and entities, and treat such expressions separately from explicit negations, for which a ML approach could still be useful. We will also further explore the feature space by considering attributes and relations extracted from the constituency parse to provide a more comprehensive classification model.

Conclusions

Given the number of published articles, detection of negations is of particular importance for data consolidation and mining. Here we explored the identification of negated regulation events, given their triggers, themes and causes. A machine learning method that combines a set of lexical, syntactic and semantic features engineered from the associated sentence was used. Adding semantic features (which characterise the participants involved in the regulation) improved precision by 20%; similarly, adding syntactic relations improve recall by almost 20%. A number of false negatives originated from contrastive patterns that have been used to express both affirmative and negated statements in parallel. The results suggested that ML approaches could not learn from such examples, and that more complex syntactic or lexical patterns are needed to capture this kind of negations.

References

[1] Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** Genome Biol. 2008; 9(Suppl 2): S4.

[2] Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii Y: **Overview of BioNLP'09 shared task on event extraction**. BioNLP'09: Proceedings of the Workshop on BioNLP. 1-9.

[3] Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CW. **PreBIND and Textomy--mining the biomedical litera-ture for protein-protein interactions using a support vector machine**. BMC Bioinformatics 4: 11.

[4] Natarajan J, Berrar D., Dubitzky W, Hack C, Zhang Y, DeSesa C, Van Brocklyn JR, Bremer EG. **Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line**. BMC Bioinformatics. 7: 373.

[5] Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muñiz-Rascado L, Martínez-Flores I, Salgado H, Bonavides-Martínez C, Abreu-Goodger C, Rodríguez-Penagos C, Miranda-Ríos J, Morett E, Merino E, Huerta AM, Treviño-Quintanilla L, Collado-Vides J. **RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation**, Nucleic Acids Res. 2008 (Database issue):D120-4.

[6] Kilicoglu H, Sabine Bergler. **Syntactic dependency based heuristics for biological event extraction**. BioNLP'09: Proceedings of the Workshop on BioNLP. 119-127

[7] MacKinlay A, David Martinez, Timothy Baldwin. **Biomedical Event Annotation with CRFs and Precision Grammars**. BioNLP'09: Proceedings of the Workshop on BioNLP. 77-85.

[8] Langacker R. **On Pronominalization and the Chain of Command**. In D. Reibel and S. Schane (eds.), Modern Studies in English, Prentice-Hall, Englewood Cliffs, NJ. 160–186, 1969.

[9] McClosky D, Charniak E, Johnson M. **Effective Self-Training for Parsing. Proceedings of HLT/NAACL 2006**. 152-159.

[10] Hakenberg J, Solt I, Tikk D, Tari L, Rheinländer A, Ngyuen QL, Gonzalez G, Leser U. **Molecular event extraction from link grammar parse trees**. BioNLP'09: Proceedings of the Workshop on BioNLP. 86-94.

[11] Morante R, Daelemans W. **A Metalearning Approach to Processing the Scope of Negation**. CoNLL '09: Proceedings of the 13th Conference on Computational Natural Language Learning. 21-29.