Integration of Static Relations to Enhance Event Extraction from Text

Sofie Van Landeghem[1,2], **Sampo Pyysalo**[3], **Tomoko Ohta**[3], **Yves Van de Peer**[1,2]
1. Dept. of Plant Systems Biology, VIB, Gent, Belgium
2. Dept. of Plant Biotechnology and Genetics, Ghent University, Gent, Belgium
3. Department of Computer Science, University of Tokyo, Tokyo, Japan
yves.vandepeer@psb.vib-ugent.be

Abstract

As research on biomedical text mining is shifting focus from simple binary relations to more expressive event representations, extraction performance drops due to the increase in complexity. Recently introduced data sets specifically targeting static relations between named entities and domain terms have been suggested to enable a better representation of the biological processes underlying annotated events and opportunities for addressing their complexity. In this paper, we present the first study of integrating these static relations with event data with the aim of enhancing event extraction performance. While obtaining promising results, we will argue that an event extraction framework will benefit most from this new data when taking intrinsic differences between various event types into account.

1 Introduction

Recently, biomedical text mining tools have evolved from extracting simple binary relations between genes or proteins to a more expressive event representation (Kim et al., 2009). Furthermore, new data sets have been developed targeting relations between genes and gene products (GGPs) and a broader category of entities, covering terms that can not be annotated as named entities (NEs) but that are still highly relevant for biomedical information extraction (Ohta et al., 2009b). In contrast to relations involving change or causality, the annotation for this data covers relations such as *part-of*, here termed "static relations" (SR) (Pyysalo et al., 2009).

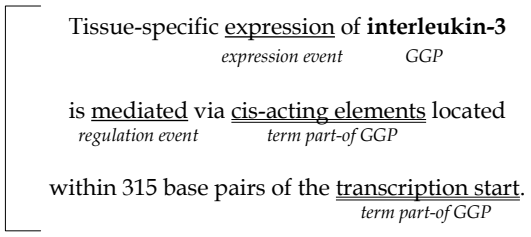

Figure 1: A sentence from PMID:8662845, showing how the event data set (single line) and the SR data set (double line) offer complementary information, enabling a more precise model of the biological reality.

As an example, Figure 1 depicts a sentence containing complementary annotations from the event data set and the SR data. The event annotation indicates an expression event involving the GGP "interleukin-3". Furthermore, regulation of this expression event is stated by the trigger word "mediated". In addition, the SR annotation marks two terms that refer to parts of the GGP, namely "cis-acting elements" and "transcription starts". These two terms provide more detailed information on the regulation event. Thus, by combining the two types of annotation, a text mining algorithm will be able to provide a more detailed representation of the extracted information. This would be in particular beneficial in practical applications such as abstract summarization or integration of the predictions into complex regulatory pathways.

In addition to providing enhanced representation of biological processes, the SR data set also offers interesting opportunities to improve on event extraction. As an example, consider the sentence presented in Figure 2, in which "c-Rel" and "p50" are both annotated as being subunits of the

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 144–152,
Uppsala, Sweden, 15 July 2010.

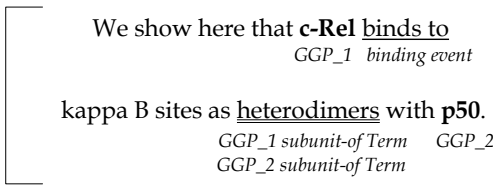

Figure 2: A sentence from PMID:1372388, showing how SR data (double line) can provide strong clues for the extraction of biomolecular events (double line) from text.

term "heterodimers". The SR data thus provides strong clues for the extraction of a Binding event involving both c-Rel and p50.

During the last few years, event extraction has gained much interest in the field of natural language processing (NLP) of biomedical text (Pyysalo et al., 2007; Kim et al., 2008; Kim et al., 2009). However, owing to the more complex nature of this task setting, performance rates are lower than for the extraction of simple binary relations. The currently best performing framework for event extraction obtains 53.29% F-score (Miwa et al., 2010), which is considerably lower than the performance reported for extraction of protein-protein interaction relations, ranging between 65% and 87% depending on the data set used for evaluation (Miwa et al., 2009).

In this paper, we will study how data on static relations can be applied to improve event extraction performance. First, we describe the various data sets (Section 2) and the text mining framework that was applied (Section 3). The main contributions of this paper are presented in Section 4, in which we study how static relation information can be integrated into an event extraction framework to enhance extraction performance. Finally, Section 5 presents the main conclusions of this work.

2 Data

In this section, we provide an overview of the two main data sets used in this work: event annotation (Section 2.1) and static relation annotation (Section 2.2).

2.1 Event Data

The BioNLP'09 Shared Task data, derived from the GENIA Event corpus (Kim et al., 2008), defines nine types of biomolecular events and is divided into three data sets: training data, development data and final test data, covering 800, 150 and 260 PubMed abstracts respectively. The event types and their statistics in the three data sets are shown in Table 1.

Event type	Args	Train	Devel	Test
Gene expression	T	1738	356	722
Transcription	T	576	82	137
Protein catabolism	T	110	21	14
Localization	T	265	53	174
Phosphorylation	T	169	47	139
Binding	T+	887	249	349
Regulation	T, C	961	173	292
Positive regulation	T, C	2847	618	987
Negative regulation	T, C	1062	196	379
TOTAL	-	8615	1795	3193

Table 1: BioNLP ST events, primary argument types and data statistics. Arguments abbreviate for (T)heme and (C)ause, with + marking arguments that can occur multiple times for an event. We refer to the task definition for details.

In the shared task setting, participants were provided with the gold annotations for Gene/Gene Product (GGP) named entities, and for all three data sets the texts of the abstracts and the gold GGP annotations are publicly available. However, while full gold event annotation is available for the training and development data sets, the shared task organizers have chosen not to release the gold annotation for the test data set. Instead, access to overall results for system predictions is provided through an online interface. This setup, adopted in part following a similar design by the organizers of the LLL challenge (Nédellec, 2005), is argued to reduce the possibility of overfitting to the test data and assure that evaluations are performed identically, thus maintaining comparability of results.

For the current study, involving detailed analysis of the interrelationships of two classes of annotations, the lack of access to the gold annotations of the test set rules this data set out as a potential target of study. Consequently, we exclude the blind test data set from consideration and use the development set as a test set.

To simplify the analysis, we further focus our efforts in this study on simple events involving only the given GGPs as participants. In the full shared task, events of the three Regulation types may take events as arguments, resulting in recursive event structures. These event types were found to be the most difficult to extract in the

SR type	Examples
term variant-of GGP	[RFX5 fusion protein], [Tax mutants], [I kappa B gamma isoforms]
term part-of GGP	[murine B29 promoter], [c-fos regulatory region], [transactivation domain] of Stat6, the nearby [J element] of the human DPA gene, the [consensus NF-kappa B binding site] of the E-selectin gene
GGP member-of term	The [Epstein-Barr virus oncoprotein] latent infection membrane protein 1, [Ikaros family members], PU.1 is a transcription factor belonging to the [Ets-family]
GGP subunit-of term	the [NF-kappa B complex] contains both RelA and p50, Human TAFII 105 is a cell type-specific [TFIID] subunit, [c-Rel/p65 heterodimers]

Table 2: Training examples of some of the SR types, including both noun phrase relations as well as relations between nominals. GGPs are underlined and terms are delimited by square brackets.

shared task evaluation (Kim et al., 2009). Furthermore, their inclusion introduces a number of complications for evaluation as well as analysis, as failure to extract a referenced event implies failure to extract events in which they appear as arguments. We note that even with the limitations of considering only the smallest of the three data sets and excluding Regulation events from consideration, the ST data still contains over 800 development test events for use in the analysis.

2.2 Static Relation Data

The data on relations is drawn from two recently introduced data sets. Both data sets cover specifically static relations where one of the participants is a GGP and the other a non-GGP term. The GGPs are drawn from the data introduced in (Ohta et al., 2009a) and the terms from the GENIA corpus term annotation (Kim et al., 2003), excluding GGPs. The first data set, introduced in (Pyysalo et al., 2009), covers static relations involving GENIA corpus terms that are annotated as participants in the events targeted in the BioNLP'09 shared task. The second data set, introduced in (Ohta et al., 2009b), contains annotation for relations holding between terms and GGPs embedded in those terms. In this study, we will use the non-embedded relations from the former data set, referring to this data as RBN for "Relations Between Nominals" in recognition of the similarity of the task setting represented by this data set and the task of learning semantic relations between nominals, as studied e.g. in SemEval (Girju et al., 2007; Hendrickx et al., 2009). We use all of the latter data set, below referred to as NPR for "Noun Phrase Relations". The NPR data set extends on the embedded part of the data introduced by (Pyysalo et al., 2009), increasing the coverage of terms included and the granularity of the annotated event types. While RBN only differentiates between a domain-specific *Variant* relation and four different part-whole relations, in NPR these are refined into more than 20 different types.

To apply these data sets together in a single framework, it was necessary to resolve the differences in the annotated relation types. First, as the finer-grained NPR types are organized in a hierarchy that includes the four part-whole relations of the RBN categorization as intermediate types (see Fig. 1 in Ohta et al. (2009b)), we collapsed the subtypes of each into these supertypes. While this removes some potentially useful distinctions, many of the finer-grained types are arguably unnecessarily detailed for the purposes of the event extraction task which, for example, makes no distinctions between events involving different gene components. Furthermore, the NPR annotations also define an Object-Variant class with multiple subtypes, but as these were judged too diverse to process uniformly, we did not collapse these subtypes as was done for part-whole relations. Rather, we divided them into "near" and "far" variants by a rough "functional distance" to the related GGP, as suggested by Ohta et al. (2009b). The relations *GGP-Modified Protein*, *GGP-Isoform* and *GGP-Mutant* were accepted into the "near" set, expected to provide positive features for inclusion in events, and the remaining subtypes into the "far" set, expected to provide negative indicators.

In addition to the primary annotation covering static relations, the RBN annotation only recognizes a mixed "other relation/out" category, used to annotate both GGP-term pairs for which the stated relation is not one of the targeted types (e.g. a causal relation) and pairs for which no relation is stated. Due to the heterogeneity of this category,

it is difficult to make use of these annotations, and we have chosen not to consider them in this work.

By contrast, the NPR annotation also subdivides the "other relation" category into five specific types, providing an opportunity to also use the part of the data not strictly involving static relations. We judged the classes labeled *Functional*, *Experimental Method* and *Diagnosis and Therapeutics* to involve terms where contained GGP names are unlikely to be participants in stated events and thus provide features that could serve as potentially useful negative indicators for event extraction. As an example, the Functional category consists of GGP-term pairs such as *GGP inhibitor* and *GGP antibody*, where the term references an entity separate from the GGP, identified through a functional or causal relation to the GGP. As such terms occur in contexts similar to ones stating events involving the GGP, explicit marking of these cases could improve precision. Consider, for example, GGP_1 *binds* GGP_2, GGP_1 *binds* GGP_2 *promoter*, GGP_1 *binds* GGP_2 *inhibitor* and GGP_1 *binds* GGP_2 *antagonist*: a binding event involving GGP_1 and GGP_2 should be extracted for the first two statements but not the latter two.

Table 2 lists some interesting examples of static relation grouped by type, including both noun phrase relations as well as relations between nominals. The consolidated data combining the two static relations - related data sets are available at the GENIA project webpage.[1]

3 Methods

The text mining tool used for all analyses in this paper is based on the event extraction framework of Van Landeghem et al. (2009), which was designed specifically for participation in the BioNLP'09 Shared Task. In this framework, triggers are discovered in text by using automatically curated dictionaries. Subsequently, candidate events are formed by combining these triggers with an appropriate number of GGPs co-occurring in the same sentence. For each distinct event type, a classifier is then built using all training examples for that specific type. Final predictions are merged for all types, forming a complex interaction graph for each article in the test set.

To distinguish between positive instances and negatives, the framework extracts rich feature vectors by analyzing lexical and syntactic information from the training data. Subsequently, a support vector machine (SVM) is built with these training patterns. The patterns include trigrams, bag-of-word features, vertex walks and information about the event trigger. As part of the current study discusses the extension and generalization of these feature patterns (Section 4.4), we will briefly discuss the various types in this section.

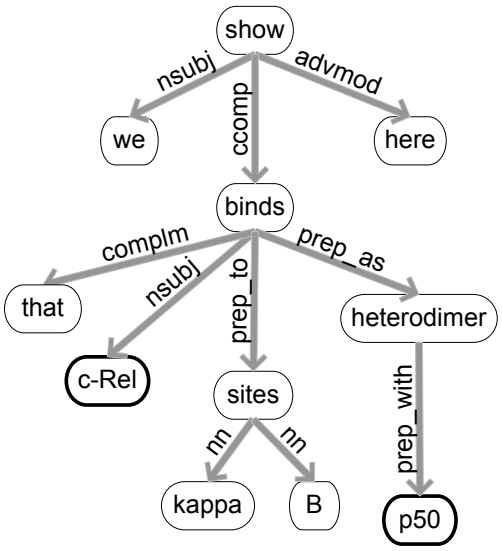

Figure 3: Dependency graph for the sentence "We show here that c-Rel binds to kappa B sites as heterodimers with p50". Words of the sentence form the nodes of the graph, while edges denote their syntactic dependencies.

To derive syntactic patterns, dependency parsing is applied using the Stanford parser (Klein and Manning, 2003; De Marneffe et al., 2006). Specifically, for each candidate event, the smallest subgraph is built including the relevant nodes for the trigger and the GGP names. Each edge in this subgraph then gives rise to a pattern including the information from the connecting nodes (or vertices) in combination with the syntactic relation specified by the edge. Trigger words and GGP names are blinded by replacing their text with the strings *protx* and *trigger* (respectively), resulting in highly general features.

Figure 3 depicts an exemplary dependency graph. For the Binding event between c-Rel and p50, the following vertex walks would be extracted: "trigger *nsubj* protx", "trigger *prep-as* heterodimer" and "heterodimer *prep-with* protx".

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

Events	Training		Dev. test	
Pos. SR data	1190	32%	227	28%
Neg. SR data	841	22%	207	26%
All SR data	1635	44%	350	43%

Table 3: Number of events that can be linked to at least one static relation, including explicitly annotated "near miss" negative annotations, also showing percentage of all gold-standard events.

Furthermore, lexical information is provided by bag-of-word (BOW) features and trigrams. BOW features incorporate all words occurring as nodes in the dependency sub-graph. They include highly informative words such as "promoter". Trigrams are formed by combining three consecutive words in the sub-sentence delimited by the trigger and GGP offsets in text. They are capable of capturing common phrases such as "physical association with".

Finally, the lexical tokens of the event trigger are highly relevant to determine the plausibility of the event being a correct one. For example, "secretion" points to a Localization event, but more general words often lead to false candidate events, such as "present". The part of speech tags of the trigger words are also included as separate features.

During feature generation, all lexical patterns are stemmed using the Porter stemming algorithm (Porter, 1980), creating even more general features and reducing sparseness of the feature vectors.

4 Experiments

This section describes a thorough study on how data on static relations can be integrated into an event extraction framework. First, we will analyze the amount of useful complementary annotations across both data sets (Section 4.1). Next, we describe the generation and evaluation of new candidate events using terms involved in static relations, in an effort to boost recall of the event predictions (Section 4.2). To additionally improve on precision, we have implemented a false positive filter exploiting SR annotations of GGPs involved in relations judged to serve as negative indicators, such as "GGP inhibitor" (Section 4.3). Finally, Section 4.4 details experiments on the creation of more extensive features for event extraction by including static relation data.

	Predicted instances	Percentage of data set
Gene expression	63	17.70%
Transcription	34	41.46%
Protein catabolism	4	19.05%
Phosphorylation	20	42.55%
Localization	4	7.55%
Binding	73	29.44%
All events	198	24.54%

Table 4: Maximal recall performance of event instances involving at least one non-NE term as argument. These terms are functioning as aliases for the GGPs they are positively associated with.

4.1 Analysis of complementary data across the two data sets

To assess the usability of the SR data set for event extraction, we first analyze the amount of complementary annotations across the two data sets. On the document level, the static relations data contains some annotation for 87.6% of all training set articles and for 94.67% of the development test set, including both positive static relations as well as explicitly negated ones. Most articles from the event data set thus involve terms at least potentially involved in static relations.

Analyzing the overlap in more detail, we determined the number of events that could benefit from adding SR data by counting the number of events for which at least one GGP is also involved in a static relation (either a positive or a negative one). Table 3 shows the results of this evaluation. In the training data, 1635 events involve at least one GGP with SR annotation, which is 44% of all events in the gold-standard annotation. For the development test set, the number is 350 out of the 808 gold standard events, i.e. 43% of events. These development set events in particular will be the subject of this study.

4.2 Terms as aliases for related GGPs

Our first application of static relations in an event extraction framework involves the use of non-NE terms appearing in the SR data set as aliases for the GGPs they are positively associated with. In the event extraction framework, new candidate events can thus be formed by treating the terms as GGPs, and mapping them back to the real GGPs after classification. This procedure is motivated by the definition of the various SR types and the underlying biological processes. For example, if a complex is known to activate the expression of a cer-

	Recall	Precision	F-score
Gene expression	11.24%	81.63%	19.75%
Transcription	20.73%	89.47%	33.66%
Protein catabolism	19.05%	100.00%	32.00%
Phosphorylation	36.17%	100.00%	53.12%
Localization	3.77%	25.00%	6.56%
Binding	12.50%	45.59%	19.62%
All events	13.75%	67.27%	22.84%

Table 5: Performance of event instances involving at least one non-NE term as argument. These terms are functioning as aliases for the GGPs they are positively associated with.

tain target GGP, then the various subunits of this complex can be annotated as participants in that event.

Obviously, this approach has some intrinsic limitations as not all GGPs occurring as arguments in events have a corresponding term that could be used as alias. However, from Table 3 it is clear that it should still be possible to extract 227 gold standard cases. To test the limitation, we have used the event extraction framework detailed in Section 3, removing the SVM classifier from the pipeline and simply labeling all candidate events as positive predictions. The result indicates that the framework is capable of retrieving 198 of the 227 gold standard cases (Table 4). The 29 missing events are due to trigger words not appearing (frequently) in the training set and thus missing from the dictionary, preventing the event to be formed as a candidate in the framework.

Our results thus show that nearly 25% of all events are potentially retrievable by using non-NE terms as aliases for GGPs. However, the analysis also indicates that in this approach, some event types are much easier to extract than others. For example, less than 8% of Localization events can be found with this setup, while maximal recall for Phosphorylation events is over 40%. These results reflect the intrinsic differences between event types and the ways in which they are typically expressed, and suggest that it should be beneficial for event extraction to take these differences into account when incorporating static relations.

Having established an upper bound for recall, a subsequent experiment involves treating the newly created instances as normal candidate events. For classification, we use an SVM trained on regular candidate events involving GGPs, as this ensures sufficient training material.

Both lexical and syntactic patterns are expected

	Baseline predictions	Merged predictions
Gene expression	77.01%	77.56%
Transcription	63.41%	64.24%
Protein catabolism	86.36%	86.36%
Phosphorylation	70.10%	76.47%
Localization	80.00%	76.77%
Binding	38.69%	40.52%
All events	64.71%	65.33%
All events (precision)	69.11%	67.19%
All events (recall)	60.84%	63.57%

Table 6: Performance of the event extraction framework. First column: using only normal events involving GGPs ("baseline"). Second column: merging the new predictions (Table 5) with the first ones. All performance rates indicate F-score, except for the last two rows.

to be similar for events involving either non-NE terms or GGPs. To test this hypothesis, we have run the event-extraction pipeline for these new instances. Evaluation is performed with the standard evaluation script provided by the BioNLP'09 Shared Task organizers, which measures the percentage of true events amongst all predictions (precision), the percentage of gold-standard events recovered (recall) and the harmonic mean of these two metrics (F-score). The results are detailed in Table 5. While we have already established that recall is subject to severe limitations (Table 4), we note in particular the high precision rates of the predictions. In particular, four out of six event types achieve a precision rate higher than 80%.

To allow for a meaningful comparison, these results should be put into perspective by merging the new predictions with the predictions of a baseline extractor and comparing against this baseline (Table 6). This analysis reveals interesting results: while overall performance increases slightly from 64.71% to 65.33% F-score, this trend is not common to all event types. For instance, prediction of Localization drops 3.23% points F-score. Considering the maximum recall results, this is not entirely surprising and confirms the hypothesis that the prediction of Localization events will not benefit from static relation data in this approach.

However, we do observe a considerable increase in performance for Phosphorylation (6.37% points F-score) events and some increase for Binding events (1.83% points F-score). This performance boost is mainly caused by an increase in recall (10.64% and 4.43% points, respectively). When considering all protein events, recall is increased

from 60.84% to 63.57% (Table 6, last row). These results clearly indicate that the inclusion of static relations can improve recall while retaining and even slightly improving general performance.

4.3 Using static relations to filter false positive events

To further improve event extraction performance, we have designed a false-positive (FP) filter using specific categories of relations serving as negative indicators for event extraction. In particular, we have used the "far variants" and *Functional* relation annotations, as described in Section 2.2. For each such relation, we add the GGP involved to the FP filter, as the GGP should not participate in any event. Thus, for example, the GGP in "GGP antibodies" would be filtered as the GGP is considered too far removed from the containing term to be a participant in any event in the context.

In the development test set, this strategy has automatically identified 24 relevant GGP mentions that should not be annotated as being involved in any event. Even though this number is relatively small, we aim at designing a high specificity FP filter while relying on the SVM classifier to solve more ambiguous cases.

Applying the FP filter to the baseline result detailed in Table 6, we find that 3 events are discarded from the set of predictions. All three instances represented false positives; two of them were Binding events and one a Gene expression event. Overall precision and F-score increased by 0.30% points and 0.13% points, respectively.

4.4 Extended feature representation incorporating information on static relations

The last type of experiment aims to boost both precision and recall by substantially extending the feature generation module for event extraction using the newly introduced SR data. Table 3 shows that such an enhanced feature representation could influence 1190 events in the training data (1635 events including negative annotations) and 227 events in the development test data (350 including negative), covering a significant part of the data set.

Building further on the feature generation module described in Section 3, we have added a range of new features to the feature vectors while also providing enhanced generalization of existing features. Generalization is crucial for the text mining framework as it enables the extraction of relations from new contexts and forms of statements.

First, for each term involved in a static relation with a GGP, the string of the term is included as a separate feature. This generates relation-associated features such as "tyrosine", which is strongly correlated with Phosphorylation events. For terms spanning multiple tokens, we additionally include each token as a separate feature, capturing commonly used words such as "promoter" or "receptor". Each distinct feature is linked to its specific relation type, such as Part-of or Member-collection (Section 2.2). To make use of annotation for "near-miss" negative cases, we generate features also for these relations, marking each feature to identify whether it was derived from a positive or negative annotation.

Additionally, we introduced a new feature type expressing whether or not the trigger of the event is equal to a term related to one or more GGPs involved in the event. As an example, suppose the candidate event is triggered by the word "homodimer". If the GGP involved is annotated as being a subunit of this homodimer, this provides a strong clue for a positive event. Similarly, the explicit negation of the existence of any static relation indicates a negative event.

Apart from these new features, we have also investigated the use of static relations to create more general lexical patterns. In particular, we have adjusted the lexical information in the feature vector by blinding terms involved in relevant relations, depending on the specific type of relation. For each such term, the whole term string is replaced by one word, expressing the type of the static relation and whether the relation is positive or negative. This results in more general patterns such as "inhibit prep-to partx" (vertex walk) or "activ in nonpartx" (trigram). In Figure 3, "heterodimer" would be blinded as "complexx" as both c-Rel and p50 are members of this complex.

Initial experiments with the extended feature representation showed that an increase in performance could be obtained on the development test set, achieving 61.34% recall, 69.58% precision and 65.20% F-score. However, it also became clear that not all event types benefit from the new features. Surprisingly, Binding is one such example. We hypothesize that this is mainly due to the intrinsic complexity of Binding events, requiring an even more advanced feature representation.

	Baseline predictions	New predictions
Gene expression	77.01%	78.06%
Transcription	63.41%	63.80%
Protein catabolism	86.36%	86.36%
Phosphorylation	70.10%	76.29%
Localization	80.00%	84.21%
Binding	38.69%	38.34%
All events	64.71%	65.73%
All events (precision)	69.11%	69.99%
All events (recall)	60.84%	61.96%

Table 7: Performance of the event extraction framework. First column: using the baseline feature representation. Second column: using the extended feature representation. All performance rates indicate F-score, except for the last two rows.

To take the inherent differences between various event types into account, we selected the optimal set of features for each type. In a new experiment, the feature generation step thus depends on the event type under consideration. Table 7 details the results of this optimization: an overall F-score of 65.73% is achieved. Similar to the experiments in Section 4.2, the F-score for the prediction of Phosphorylation events increases by 6.19% points. Additionally, in this experiment we obtain an increase of 4.21% points in F-score for Localization events, even though we were unable to improve on them when using terms as aliases for additional candidate events (Section 4.2). Additional experiments suggested the reason to be that while the Localization event type in general does not benefit from positive static relations, negative static relations seem to provide strong clues to the SVM classifier.

5 Conclusion

We have presented the first study on the applicability of static relations for event prediction in biomedical texts. While data on static relations can offer a more detailed representation of biomolecular events, it can also help to boost the performance of event prediction. We have performed three sets of experiments to investigate these opportunities. First, we have designed new candidate events by treating non-NE terms as aliases for the GGPs they are associated with. By augmenting the normal event predictions with predictions for these new candidates, we have established a considerable increase in recall. Next, we have implemented a false positive filter to improve precision, by exploiting annotation for relations judged to imply only distant associations of the GGP and the enclosing term. Finally, the last type of experiment involves integrating complementary data on static relations to obtain more informative feature vectors for candidate events. Results show that both recall and precision can be increased slightly by this last, more complex configuration.

During the experiments, it has become clear that there are important differences between the data sets of distinct event types. For example, we have found that Phosphorylation events benefit the most from added static relations data, while Localization events can be enhanced using only features of negative static relation annotations. For some event types, such as Protein catabolism, the current techniques for integration of static relations do not generate a performance boost. However, our findings pave the way for experiments involving more detailed representations, taking the intrinsic properties of the various event types into account and combining the various ways of integrating the new information. We regard these opportunities as promising future work.

Finally, having established the potential added value offered by data on static relations in an event extraction framework, additional future work will focus on the automatic extraction of the static relations. Similar relations have been considered in numerous recent studies, and while challenges to reliable prediction remain, several methods with promising performance have been proposed (Girju et al., 2007; Hendrickx et al., 2009). By integrating predictions from both static relations and events instead of using gold standard relation annotations, we will be able to study the effect of the relation information on new data, including the shared task test set. Such experiments are key to establishing the practical value of static relations for biomolecular event extraction.

Acknowledgments

SVL would like to thank the Research Foundation Flanders (FWO) for funding her research. The work of SP and TO was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

References

M. De Marneffe, B. Maccartney, and C. Manning. 2006. Generating typed dependency parses from

phrase structure parses. In *Proceedings of LREC-06*, pages 449–454.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, Colorado, June. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1):i180–i182.

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

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July. Association for Computational Linguistics.

Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. 2009. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 121–130, Morristown, NJ, USA. Association for Computational Linguistics.

Makoto Miwa, Rune Saetre, Jin-Dong D. Kim, and Jun'ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *Journal of bioinformatics and computational biology*, 8(1):131–146, February.

Claire Nédellec. 2005. Learning language in logic - genic interaction extraction challenge. In *Proceedings of the Learning Language in Logic Workshop (LLL'05)*.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. 2009a. Incorporating genetag-style annotation to genia corpus. In *Proceedings of the BioNLP 2009 Workshop*, pages 106–107, Boulder, Colorado, June. Association for Computational Linguistics.

Tomoko Ohta, Sampo Pyysalo, Kim Jin-Dong, and Jun'ichi Tsujii. 2009b. A re-evaluation of biomedical named entity - term relations. In *Proceedings of LBM'09*.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):50+.

Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the BioNLP 2009 Workshop*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Sofie Van Landeghem, Yvan Saeys, Bernard De Baets, and Yves Van de Peer. 2009. Analyzing text in search of bio-molecular events: a high-precision machine learning framework. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 128–136, Morristown, NJ, USA. Association for Computational Linguistics.