Event Extraction for Post-Translational Modifications

Tomoko Ohta* **Sampo Pyysalo*** **Makoto Miwa*** **Jin-Dong Kim*** **Jun'ichi Tsujii*†‡**
*Department of Computer Science, University of Tokyo, Tokyo, Japan
†School of Computer Science, University of Manchester, Manchester, UK
‡National Centre for Text Mining, University of Manchester, Manchester, UK
{okap,smp,mmiwa,jdkim,tsujii}@is.s.u-tokyo.ac.jp

Abstract

We consider the task of automatically extracting post-translational modification events from biomedical scientific publications. Building on the success of event extraction for phosphorylation events in the BioNLP'09 shared task, we extend the event annotation approach to four major new post-transitional modification event types. We present a new targeted corpus of 157 PubMed abstracts annotated for over 1000 proteins and 400 post-translational modification events identifying the modified proteins and sites. Experiments with a state-of-the-art event extraction system show that the events can be extracted with 52% precision and 36% recall (42% F-score), suggesting remaining challenges in the extraction of the events. The annotated corpus is freely available in the BioNLP'09 shared task format at the GENIA project homepage.[1]

1 Introduction

Post-translational-modifications (PTM), amino acid modifications of proteins after translation, are one of the posterior processes of protein biosynthesis for many proteins, and they are critical for determining protein function such as its activity state, localization, turnover and interactions with other biomolecules (Mann and Jensen, 2003). Since PTM alter the properties of a protein by attaching one or more biochemical functional groups to amino acids, understanding of the mechanism and effects of PTM are a major goal in the recent molecular biology, biomedicine and pharmacology fields. In particular, epigenetic ("outside conventional genetics") regulation of gene expression has a crucial role in these fields and PTM-like modifications of biomolecules are a burning issue. For instance, tissue specific or context dependent expression of many proteins is now known to be controlled by specific PTM of histone proteins, such as *Methylation* and *Acetylation* (Jaenisch and Bird, 2003). This *Methylation* and *Acetylation* of specific amino acid residues in histone proteins are strongly implicated in unwinding the nucleosomes and exposing genes to transcription, replication and DNA repairing machinery.

The recent BioNLP'09 Shared Task on Event Extraction (Kim et al., 2009a) (below, BioNLP shared task) represented the first community-wide step toward the extraction of fine-grained event representations of information from biomolecular domain publications (Ananiadou et al., 2010). The nine event types targeted in the task included one PTM type, *Phosphorylation*, whose extraction involved identifying the modified protein and, when stated, the specific phosphorylated site. The results of the shared task showed this PTM event to be single most reliably extracted event type in the data, with the best-performing system for the event type achieving 91% precision and 76% recall (83% F-score) in the extraction of phosphorylation events (Buyko et al., 2009). The results suggest both that the event representation is well applicable to PTM and that current extraction methods are capable of reliable PTM extraction. Most of the proposed state-of-the-art methods for event extraction are further largely machine-learning based. This suggest that the coverage of many existing methods could be straightforwardly extended to new event types and domains by extending the scope of available PTM annotations and retraining the methods on newly annotated data. In this study, we take such an annotation-based approach to extend the extraction capabilities of state of the art event extraction methods for PTM.

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

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 19–27,
Uppsala, Sweden, 15 July 2010.

Term	Count	
Phosphorylation	172875	50.90%
Methylation	49780	14.66%
Glycosylation	36407	10.72%
Hydroxylation	20141	5.93%
Acetylation	18726	5.51%
Esterification	7836	2.31%
Ubiquitination	6747	1.99%
ADP-ribosylation	5259	1.55%
Biotinylation	4369	1.29%
Sulfation	3722	1.10%
...		
TOTAL	339646	100%

Table 1: PTM mentions in PubMed. The number of citations returned by the PubMed search engine for each PTM term shown together with the fraction of the total returned for all searches. Searches were performed with the terms as shown, allowing MeSH term expansion and other optimizations provided by the Entrez search.

2 Corpus Annotation

We next discuss the selection of the annotated PTM types and source texts and present the representation and criteria used in annotation.

2.1 Event Types

A central challenge in the automatic extraction of PTMs following the relatively data-intensive BioNLP shared task model is the sheer number of different modifications: the number of known PTM types is as high as 300 and constantly growing (Witze et al., 2007). Clearly, the creation of a manually annotated resource with even modest coverage of statements of each of the types would be a formidable undertaking. We next present an analysis of PTM statement occurrences in PubMed as the first step toward resolving this challenge.

We estimated the frequency of mentions of prominent PTM types by combining MeSH ontology[2] PTM terms with terms occurring in the `post-translational protein modification` branch of the Gene Ontology (The Gene Ontology Consortium, 2000). After removing variants (e.g. *polyamination* for *amination* or *dephosphorylation* for *phosphorylation*) and two cases judged likely to occur frequently in non-PTM contexts (*hydration* and *oxidation*), we searched PubMed for the remaining 31 PTM types. The results for the most frequent types are shown in Table 1. We find a power-law - like distribution with *phosphorylation* alone accounting for over 50% of the total, and the top 6 types together for over 90%. By contrast, the bottom ten types together represent less than a percent of total occurrences.

This result implies that fair coverage of individual PTM event mentions can be achieved without considering even dozens of different PTM event types, let alone hundreds. Thus, as a step toward extending the coverage of event extraction systems for PTM, we chose to focus limited resources on annotating a small selection of types so that a number of annotations sufficient for supervised learning and stable evaluation can be provided. To maximize the utility of the created annotation, the types were selected based on their frequency of occurrence.

2.2 Text Selection

Biomedical domain corpora are frequently annotated from selections of texts chosen as a sample of publications in a particular subdomain of interest. While several areas in present-day molecular biology are likely to provide ample source data for PTM statements, a sample of articles from any subdomain is unlikely to provide a well-balanced distribution of event types: for example, the most frequent PTM event type annotated in the GENIA event corpus occurs more than 10 times as often as the second most frequent (Kim et al., 2008). Further, avoiding explicit subdomain restrictions is not alone sufficient to assure a balanced distribution of event types: in the BioInfer corpus, for which sentences were selected on the basis of their containing mentions of protein pairs known to interact, the most frequent PTM type is again annotated nearly four times as often as the second most frequent (Pyysalo et al., 2007).

To focus annotation efforts on texts relevant to PTM and to guarantee that the annotation results in relatively balanced numbers of PTM events of each targeted type, we decided to annotate a targeted set of source texts instead of a random sample of texts for a particular subdomain. This type of targeted annotation involves a risk of introducing bias: a badly performed selection could produce a corpus that is not representative of the

[2] `http://www.nlm.nih.gov/mesh/meshhome.html`

PTM type	AB	FT
Acetylation	103	128
Glycosylation	226	336
Methylation	72	69
Phosphorylation	186	76
Hydroxylation	71	133

Table 2: Number of abstracts (AB) and full-text articles (FT) tagged in PIR as containing PTM statements.

statements expressing PTMs in text and thus poor material for either meaningful evaluation or for training methods with good generalization performance.[3] To avoid such bias, we decided to base our selection of the source texts on an independently annotated PTM resource with biological (as opposed to textual) criteria for inclusion. Owing in part to the recent interest in PTMs, there are currently a wealth of resources providing different levels of annotation for PTMs.

Here, we have chosen to base initial annotation on corpora provided by the Protein Information Resource[4] (PIR) (Wu et al., 2003). These corpora contain annotation for spans with evidence for five different PTM types (Table 2), corresponding to the five PTMs found above to occur in PubMed with the highest frequency. A key feature setting this resource apart from others we are aware of is that it provides text-bound annotations identifying the statement by which a PTM record was made in the context of the full publication abstracts. While this annotation is less specific and detailed than the full BioNLP shared task markup, it could both serve as an initial seed for annotation and assure that the annotation agrees with relevant database curation criteria. The PIR corpora have also been applied in previous PTM extraction studies (e.g. (Hu et al., 2005; Narayanaswamy et al., 2005)).

We judged that the annotated Phosphorylation events in the BioNLP shared task data provide sufficient coverage for the extraction of this PTM type, and chose to focus on producing annotation for the four other PTM types in the PIR data. As the high extraction performance for phosphorylation events in the BioNLP shared task was

[3]One could easily gather PTM-rich texts by performing protein name tagging and searching for known patterns such as "[PROTEIN] methylates [PROTEIN]", but a corpus created in this way would not necessarily provide significant novelty over the original search patterns.

[4]http://pir.georgetown.edu

Protein	Site	PTM	Count
collagen	lysine	Hydroxylate	44
myelin	arginine	Methylate	17
M protein	N-terminal	Glycosylate	2
EF-Tu	lysine	Methylate	1
Actobindin	NH2 terminus	Acetylate	0

Table 3: Example queried triples and match counts from Medie.

achieved with annotated training data containing 215 PTM events, in view of the available resources we set as an initial goal the annotation of 100 events of each of the four PTM types. To assure that the annotated resource can be made publicly available, we chose to use only the part of the PIR annotations that identified sections of PubMed abstracts, excluding full-text references and non-PubMed abstracts. Together with the elimination of duplicates and entries judged to fall outside of the event annotation criteria (see Section 2.4), this reduced the number of source texts below our target, necessitating a further selection strategy.

For further annotation, we aimed to select abstracts that contain specific PTM statements identifying both the name of a modified protein and the modified site. As for the initial selection, we further wished to avoid limiting the search by searching for any specific PTM expressions. To implement this selection, we used the Medie system[5] (Ohta et al., 2006; Miyao et al., 2006) to search PubMed for sentences where a specific protein and a known modified site were found together in a sentence occurring in an abstract annotated with a specific MeSH term. The (protein name, modified site, MeSH term) triples were extracted from PIR records, substituting the appropriate MeSH term for each PTM type. Some examples with the number of matching documents are shown in Table 3. As most queries returned either no documents or a small number of hits, we gave priority to responses to queries that returned a small number of documents to avoid biasing the corpus toward proteins whose modifications are frequently discussed.

We note that while the PIR annotations typically identified focused text spans considerably shorter than a single sentence and sentence-level search was used in the Medie-based search to increase the likelihood of identifying relevant statements, after selection all annotation was performed to full abstracts.

[5]http://www-tsujii.is.s.u-tokyo.ac.jp/medie/

Event type	Count
Protein_modification	38
Phosphorylation	546
Dephosphorylation	28
Acetylation	7
Deacetylation	1
Ubiquitination	6
Deubiquitination	0

Table 4: GENIA PTM-related event types and number of events in the GENIA event corpus. Type names are simplified: the full form of e.g. the *Phosphorylation* type in the GENIA event ontology is *Protein_amino_acid_phosphorylation*.

Event type	Arguments	Count
Protein_modification	Theme	31
Phosphorylation	Theme	261
Phosphorylation	Theme, Site	230
Phosphorylation	Site	20
Phosphorylation	Theme, Cause	14
Dephosphorylation	Theme	16

Table 5: GENIA PTM-related event arguments. Only argument combinations appearing more than 10 times in the corpus shown.

2.3 Representation

The employed event representation can capture the association of varying numbers of participants in different roles. To apply an event extraction approach to PTM, we must first define the targeted representation, specifying the event types, the mandatory and optional arguments, and the argument types – the roles that the participants play in the events. In the following, we discuss alternatives and present the representation applied in this work.

The GENIA Event ontology, applied in the annotation of the GENIA Event corpus (Kim et al., 2008) that served as the basis of the BioNLP shared task data, defines a general *Protein_modification* event type and six more specific modification subtypes, shown in Table 4. While the existing *Acetylation* type could thus be applied together with the generic *Protein_modification* type to capture all the annotated PTMs, we believe that identification of the specific PTM type is not only important to users of extracted PTM events but also a relatively modest additional burden for automatic extraction, owing to the unambiguous nature of typical expressions used to state

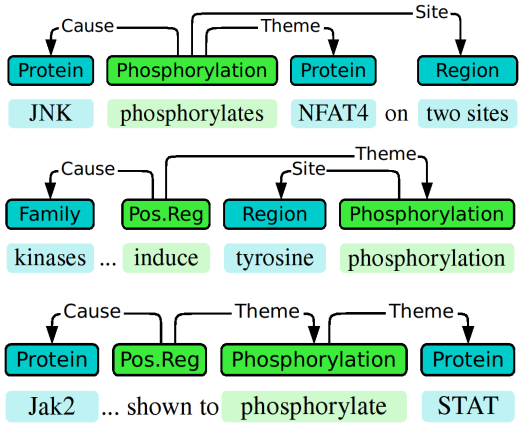

Figure 1: Alternative representations for PTM statements including a catalyst in GENIA Event corpus. PTM events can be annotated with a direct Cause argument (top, PMID 9374467) or using an additional *Regulation* event (middle, PMID 10074432). The latter annotation can be applied also in cases where there is no expression directly "triggering" the secondary event (bottom, PMID 7613138).

PTMs in text. We thus chose to introduce three additional specific modification types, *Glycosylation*, *Hydroxylation* and *Methylation* for use in the annotation.

The GENIA Event corpus annotation allows PTM events to take Theme, Site and Cause arguments specifying the event participants, where the Theme identifies the entity undergoing the modification, Site the specific region being modified, and Cause an entity or event leading to the modification. Table 5 shows frequent argument combinations appearing in the annotated data. We note that while Theme is specified in the great majority of events and Site in almost half, Cause is annotated for less than 5% of the events. However, the relative sparsity of Cause arguments in modification events does not imply that e.g. catalysts of the events are stated only very rarely, but instead reflects also the use of an alternative representation for capturing such statements without a Cause argument for the PTM event. The GENIA event annotation specifies a *Regulation* event (with *Positive_regulation* and *Negative_regulation* subtypes), used to annotate not only regulation in the biological sense but also statements of general causality between events: *Regulation* events are used generally to connect entities or events stated to other events that they are stated to cause. Thus, PTM

events with a stated cause (e.g. a catalyst) can be alternatively represented with a Cause argument on the PTM event or using a separate *Regulation* event (Figure 1). The interpretation of these event structures is identical, and from an annotation perspective there are advantages to both. However, for the purpose of automatic extraction it is important to establish a consistent representation, and thus only one should be used.

In this work, we follow the latter representation, disallowing Cause arguments for annotated PTM events and applying separate Regulation events to capture e.g. catalyst associations. This choice has the benefits of providing an uniform representation for catalysis and inhibition (one involving a Positive regulation and the other a Negative regulation event), reducing the sparseness of specific event structures in the data, and matching the representation chosen in the BioNLP shared task, thus maintaining compatibility with existing event extraction methods. Finally, we note that while we initially expected that glycosylation statements might frequently identify specific attached side chains, necessitating the introduction of an additional argument type to accurately capture all the stated information regarding Glycosylation events, the data contained too few examples for either training material or to justify the modification of the event model. We adopt the constraints applied in the BioNLP shared task regarding the entity types allowed as specific arguments. Thus, the representation we apply here annotated PTM events with specific types, taking as Theme argument a gene/gene product type entity and as Site argument a physical (non-event) entity that does not need to be assigned a specific type.

2.4 Annotation criteria

To create PTM annotation compatible with the event extraction systems introduced for the BioNLP shared task, we created annotation following the GENIA Event corpus annotation criteria (Kim et al., 2008), as adapted for the shared task. The criteria specify that annotation should be applied to statements that involve the occurrence of a change in the state of an entity – even if stated as having occurred in the past, or only hypothetically – but not in cases merely discussing the state or properties of entities, even if these can serve as the basis for inference that a specific change has occurred. We found that many of the spans annotated in PIR as evidence for PTM did not fulfill the criteria for event annotation. The most frequent class consisted of cases where the only evidence for a PTM was in the form of a sequence of residues, for example

> Characterization [...] gave the following sequence, Gly-Cys-Hyp-D-Trp-Glu-Pro-Trp-Cys-NH2 where Hyp = 4-trans-hydroxyproline. (PMID 8910408)

Here, the occurrence of hydroxyproline in the sequence implies that the protein has been hydroxylated, but as the hydroxylation event is only implied by the protein state, no event is annotated.

Candidates drawn from PIR but not fulfilling the criteria were excluded from annotation. While this implies that the general class of event extraction approaches considered here will not recover all statements providing evidence of PTM to biologists (per the PIR criteria), several factors mitigate this limitation of their utility. First, while PTMs implied by sequence only are relatively frequent in PIR, its selection criteria give emphasis to publications initially reporting the existence of a PTM, and further publications discussing the PTM are not expected to state it as sequence only. Thus, it should be possible to extract the corresponding PTMs from later sources. Similarly, one of the promises of event extraction approaches is the potential to extract associations of multiple entities and extract causal chains connecting events with others (e.g. *E catalyzes the hydroxylation of P, leading to* ...), and the data indicates that the sequence-only statements typically provide little information on the biological context of the modification beyond identifying the entity and site. As such non-contextual PTM information is already available in multiple databases, this class of statements may not be of primary interest for event extraction.

2.5 Annotation results

The new PTM annotation covers 157 PubMed abstracts. Following the model of the BioNLP shared task, all mentions of specific gene or gene product names in the abstracts were annotated, applying the annotation criteria of (Ohta et al., 2009). This new named entity annotation covers 1031 gene/gene product mentions, thus averaging more than six mentions per annotated abstract. In total, 422 events of which 405 are of the novel PTM

Event type	Count
Glycosylation	122
Hydroxylation	103
Methylation	90
Acetylation	90
Positive reg.	12
Phosphorylation	3
Protein modification	2
TOTAL	422

Table 6: Statistics of the introduced event annotation.

Arguments	Count
Theme, Site	363
Theme	36
Site	6

Table 7: Statistics for the arguments of the annotated PTM events.

types were annotated, matching the initial annotation target in number and giving a well-balanced distribution of the specific PTM types (Table 6). Reflecting the selection of the source texts, the argument structures of the annotated PTM events (Table 7) show a different distribution from those annotated in the GENIA event corpus (Table 5): whereas less than half of the GENIA event corpus PTM events include a Site argument, almost 90% of the PTM events in the new data include a Site. PTM events identifying both the modified protein and the specific modified site are expected to be of more practical interest. However, we note that the greater number of multi-argument events is expected to make the dataset more challenging as an extraction target.

3 Evaluation

To estimate the capacity of the newly annotated resource to support the extraction of the targeted PTM events and the performance of current event extraction methods at open-domain PTM extraction, we performed a set of experiments using an event extraction method competitive with the state of the art, as established in the BioNLP shared task on event extraction (Kim et al., 2009a; Björne et al., 2009).

3.1 Methods

We adopted the recently introduced event extraction system of Miwa et al. (2010). The system applies a pipeline architecture consisting of three supervised classification-based modules: a trigger detector, an event edge detector, and an event detector. In evaluation on the BioNLP shared task test data, the system extracted phosphorylation events at 75.7% precision and 85.2% recall (80.1% F-score) for Task 1, and 75.7% precision and 83.3% recall (79.3% F-score) for Task 2, showing performance comparable to the best results reported in the literature for this event class (Buyko et al., 2009). We assume three preconditions for the PTM extraction: proteins are given, all PTMs have Sites, and all arguments in a PTM co-occur in sentence scope. The first of these is per the BioNLP shared task setup, the second fixed based the corpus statistics, and the third a property intrinsic to the extraction method, which builds on analysis of sentence structure.[6] In the experiments reported here, only the four novel PTM event types with Sites in the corpus are regarded as a target for the extraction.

The system extracted PTMs as follows: the trigger detector detected the entities (triggers and sites) of the PTMs, the event edge detector detected the edges in the PTMs, and the event detector detected the PTMs. The evaluation setting was the same as the evaluation in (Miwa et al., 2010) except for the threshold. The thresholds in the three modules were tuned with the development data set.

Performance evaluation is performed using the BioNLP shared task primary evaluation criteria, termed the "Approximate Span Matching" criterion. This criterion relaxes the requirements of strict matching in accepting extracted event triggers and entities as correct if their span is inside the region of the corresponding region in the gold standard annotation.

3.2 Data Preparation

The corpus data was split into training and test sets on the document level with a sampling strategy that aimed to preserve a roughly 3:1 ratio of occurrences of each event type between training and test data. The test data was held out during system development and parameter selection and only applied in a single final experiment. The event extraction system was trained using the 112 abstracts of the training set, further using 24 of the abstracts

[6] We note that in the BioNLP shared task data, all arguments were contained within single sentences for 95% of events.

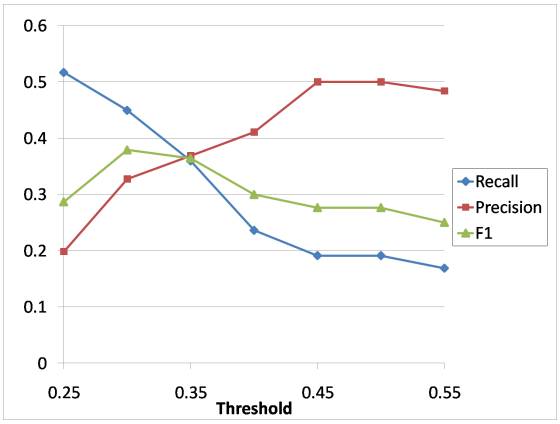

Figure 2: Performance of PTM extraction on the development data set.

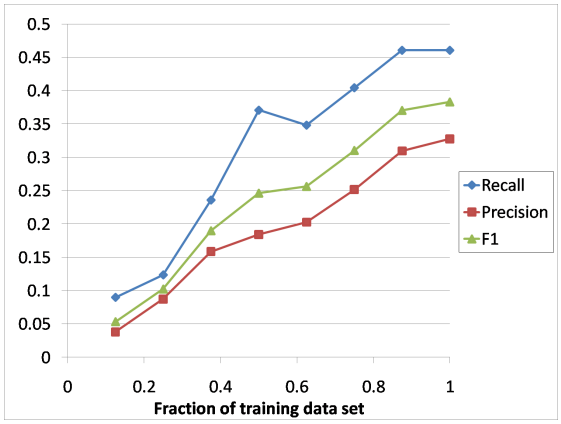

Figure 3: Learning curve of PTM extraction on the development data set.

Event type	Prec	Rec	F
Acetylation	69.6%	36.7%	48.1%
Methylation	50.0%	34.2%	40.6%
Glycosylation	36.7%	42.5%	39.4%
Hydroxylation	57.1%	29.3%	38.7%
Overall	52.1%	35.7%	42.4%

Table 8: Event extraction results on the test set.

as a development test set.

3.3 Results

We first performed parameter selection, setting the machine learning method parameter by estimating performance on the development data set. Figure 2 shows the performance of PTM extraction on the development data set with different values of parameter. The threshold value corresponding to the best performance (0.3) was then applied for an experiment on the held-out test set.

Performance on the test set was evaluated as 52% precision and 36% recall (42% F-score), matching estimates on the development data. A breakdown by event type (Table 8) shows that *Acetylation* is most reliably extracted with extraction for the other three PTM types showing similar F-scores despite some variance in the precision/recall balance. We note that while these results fall notably below the best result reported for Phosphorylation events in the BioNLP shared task, they are comparable to the best results reported in the task for Regulation and Binding events (Kim et al., 2009a), suggesting that the dataset allows the extraction of the novel PTM events with Theme and Site arguments at levels comparable to multi-argument shared task events.

Further, a learning curve (Figure 3) plotted on the development data suggests roughly linearly increasing performance over most of the curve. While the increase appears to be leveling off to an extent when using all of the available data, the learning curve indicates that performance can be further improved by increasing the size of the annotated dataset.

4 Discussion

Post-translational modifications have been a focus of interest in the biomedical text mining community, and a number of resources and systems targeting PTM have been proposed. The GENIES and GeneWays systems (Friedman et al., 2001; Rzhetsky et al., 2004) targeted PTM events such as phosphorylation and dephosphorylation under the more general *createbond* and *breakbond* types. Hu et al. (2005) introduce the RLIMS-P rule-based system for mining the substrates and sites for phosphorylation, which is extended with the capacity to extract intra-clausal statements by Narayanaswamy et al. (2005). Saric et al. (2006) present an extension of their rule-based STRING-IE system for extracting regulatory networks to capture phosphorylation and dephosphorylation events. Lee et al. (2008) present E3Miner, a tool for automatically extracting information related to ubiquitination, and Kim et al. (2009b) present a preliminary study adapting the E3Miner approach to the mining of acetylation events.

It should be noted that while studies targeting single specific PTM types report better results than found in the initial evaluation presented here (in many cases dramatically so), different

extraction targets and evaluation criteria complicate direct comparison. Perhaps more importantly, our aim here is to extend the capabilities of general event extraction systems targeting multiple types of structured events. Pursuing this broader goal necessarily involves some compromise in the ability to focus on the extraction of individual event types, and it is expected that highly focused systems will provide better performance than retrained general systems.

The approach to PTM extraction adopted here relies extensively on the availability of annotated resources, the creation of which requires considerable effort and expertise in understanding the target domain as well as the annotation methodology and tools. The annotation created in this study, performed largely on the basis of partial existing annotations drawn from PIR data, involved an estimated three weeks of full-time effort from an experienced annotator. As experiments further indicated that a larger corpus may be necessary for reliable annotation, we can estimate that extending the approach to sufficient coverage of each of hundreds of PTM types without a partial initial annotation would easily require several person-years of annotation efforts. We thus see a clear need for the development of unsupervised or semisupervised methods for PTM extraction to extend the coverage of event extraction systems to the full scale of different PTM types. Nevertheless, even if reliable methods for PTM extraction that entirely avoid the need for annotated training data become available, a manually curated reference standard will still be necessary for reliable estimation of their performance. To efficiently support the development of event extraction systems capable of capturing the full variety of PTM events, it may be beneficial to reverse the approach taken here: instead of annotating hundreds of examples of a small number of PTM types, annotate a small number of each of hundreds of PTM types, thus providing both seed data for semisupervised approaches as well as reference data for the evaluation of broad-coverage PTM event extraction systems.

5 Conclusions and Future Work

We have presented an event extraction approach to automatic PTM recognition, building on the model introduced in the BioNLP shared task on event extraction. By annotating a targeted corpus for four prominent PTM types not considered in the BioNLP shared task data, we have created a resource that can be straightforwardly used to extend the capability of event extraction systems for PTM extraction. We estimated that while systems trained on the original shared task dataset could not recognize more than 50% of PTM mentions due to their types, the introduced annotation increases this theoretical upper bound to nearly 90%. An initial experiment on the newly introduced dataset using a state-of-the-art method indicated that straightforward adoption of the dataset as training data to extend coverage of PTM events without specific adaptations of the method is feasible, although the measured performance indicates remaining challenges for reliable extraction. Further, while the experiments were performed on a dataset selected to avoid bias toward e.g. a particular subdomain or specific forms of event expressions, it remains an open question how extraction performance generalizes to biomedical literature beyond the selected sample. As experiments indicated clear remaining potential for the improvement of extraction performance from more training data, the extension of the annotated dataset is a natural direction for future work. We considered also the possiblity of extending annotation to cover small numbers of each of a large variety of PTM types, which would place focus on the challenges of event extraction with little or no training data for specific event types.

The annotated corpus covering over 1000 gene and gene product entities and over 400 events is freely available in the widely adopted BioNLP shared task format at the GENIA project homepage.[7]

Acknowledgments

We would like to thank Goran Topic for automating Medie queries to identify target abstracts. This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Japan-Slovenia Research Cooperative Program (JSPS, Japan and MHEST, Slovenia).

References

Sophia Ananiadou, Sampo Pyysalo, Junichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*. (to appear).

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

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 10–18, Boulder, Colorado, June. Association for Computational Linguistics.

Ekaterina Buyko, Erik Faessler, Joachim Wermter, and Udo Hahn. 2009. Event extraction from trimmed dependency graphs. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 19–27, Boulder, Colorado, June. Association for Computational Linguistics.

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl. 1):S74–S82.

Z. Z. Hu, M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker, and C. H. Wu. 2005. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, 21(11):2759–2765.

Rudolf Jaenisch and Adrian Bird. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33:245–254.

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

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009a. 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.

Youngrae Kim, Hodong Lee, and Gwan-Su Yi. 2009b. Literature mining for protein acetylation. In *Proceedings of LBM'09*.

Hodong Lee, Gwan-Su Yi, and Jong C. Park. 2008. E3Miner: a text mining tool for ubiquitin-protein ligases. *Nucl. Acids Res.*, 36(suppl.2):W416–422.

Matthias Mann and Ole N. Jensen. 2003. Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21:255–261.

Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *Journal of Bioinformatics and Computational Biology (JBCB)*, 8(1):131–146, February.

Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. 2006. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of COLING-ACL 2006*, pages 1017–1024.

M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. 2005. Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics*, 21(suppl.1):i319–327.

Tomoko Ohta, Yusuke Miyao, Takashi Ninomiya, Yoshimasa Tsuruoka, Akane Yakushiji, Katsuya Masuda, Jumpei Takeuchi, Kazuhiro Yoshida, Tadayoshi Hara, Jin-Dong Kim, Yuka Tateisi, and Jun'ichi Tsujii. 2006. An Intelligent Search Engine and GUI-based Efficient MEDLINE Search Tool Based on Deep Syntactic Parsing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 17–20.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, and Jun'ichi Tsujii. 2009. Incorporating GENETAG-style annotation to GENIA corpus. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 106–107, Boulder, Colorado. Association for Computational Linguistics.

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(50).

Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Duboué, Wubin Weng, W. John Wilbur, Vasileios Hatzivassiloglou, and Carol Friedman. 2004. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53.

Jasmin Saric, Lars Juhl Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. 2006. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, 22(6):645–650.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Eric S Witze, William M Old, Katheryn A Resing, and Natalie G Ahn. 2007. Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, 4:798–806.

Cathy H. Wu, Lai-Su L. Yeh, Hongzhan Huang, Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhangzhi Hu, Panagiotis Kourtesis, Robert S. Ledley, Baris E. Suzek, C.R. Vinayaka, Jian Zhang, and Winona C. Barker. 2003. The Protein Information Resource. *Nucl. Acids Res.*, 31(1):345–347.