Towards Exhaustive Protein Modification Event Extraction

Sampo Pyysalo* **Tomoko Ohta*** **Makoto Miwa*** **Jun'ichi Tsujii†**
*Department of Computer Science, University of Tokyo, Tokyo, Japan
†Microsoft Research Asia, Beijing, China
{smp,okap,mmiwa}@is.s.u-tokyo.ac.jp, jtsujii@microsoft.com

Abstract

Protein modifications, in particular post-translational modifications, have a central role in bringing about the full repertoire of protein functions, and the identification of specific protein modifications is important for understanding biological systems. This task presents a number of opportunities for the automatic support of manual curation efforts. However, the sheer number of different types of protein modifications is a daunting challenge for automatic extraction that has so far not been met in full, with most studies focusing on single modifications or a few prominent ones. In this work, aim to meet this challenge: we analyse protein modification types through ontologies, databases, and literature and introduce a corpus of 360 abstracts manually annotated in the BioNLP Shared Task event representation for over 4500 mentions of proteins and 1000 statements of modification events of nearly 40 different types. We argue that together with existing resources, this corpus provides sufficient coverage of modification types to make effectively exhaustive extraction of protein modifications from text feasible.

1 Introduction

In the decade following the sequencing of the human genome, the critical role of protein modifications in establishing the full set of protein functions from forms transcribed from the fixed DNA is increasingly appreciated, reflected in the rise of proteomics as an extension and complement to genetics in efforts to understand gene and protein functions. The mapping of the space of modifications of specific proteins is a formidable undertaking: the number of known *types* of post-translational modifications (PTMs) is as high as 300 (Witze et al., 2007) with new types identified regularly (e.g. (Brennan and Barford, 2009)), and the number of specific molecular variants of proteins in cells may be several orders of magnitude larger than that encoded in the genome; up to millions for humans (Walsh, 2006). Automatic extraction of protein modifications from the massive literature on the topic could contribute significantly to addressing these challenges.

Biomedical information extraction (IE) has advanced substantially in recent years, shifting from the detection of simple binary associations such as protein-protein interactions toward resources and methods for the extraction of multiple types of structured associations of varying numbers participants in specific roles. These IE approaches are frequently termed *event extraction* (Ananiadou et al., 2010). While protein modifications have been considered in numerous IE studies in the domain (e.g. (Friedman et al., 2001; Rzhetsky et al., 2004; Hu et al., 2005; Narayanaswamy et al., 2005; Saric et al., 2006; Yuan et al., 2006; Lee et al., 2008; Ohta et al., 2010), event extraction efforts have brought increased focus also on the extraction of protein modifications: in the BioNLP Shared Task series that has popularized event extraction, the 2009 shared task (Kim et al., 2009) involved the extraction of nine event types including one PTM, and in the 2011 follow-up event (Kim et al., 2011) the Epigenetics and Post-translational modifications (EPI) task (Ohta et al., 2011) targeted six PTM types, their re-

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 114–123,
Portland, Oregon, USA, June 23-24, 2011.

verse reactions, and statements regarding their catalysis. The results of these tasks were promising, suggesting that the single PTM type could be extracted at over 80% F-score (Buyko et al., 2009) and the core arguments of the larger set at nearly 70% F-score (Björne and Salakoski, 2011).

The increasing availability of systems capable of detailed IE for protein modifications, their high performance also for multiple modifications types, and demonstrations of the scalability of the technology to the full scale of the literature (Björne et al., 2010) are highly encouraging for automatic extraction of protein modifications. However, previous efforts have been restricted by the relatively narrow scope of targeted modification types. In the present study, we seek to address the task in full by identifying all modifications of substantial biological significance and creating an annotated resource with effectively complete type-level coverage. We additionally present preliminary extraction results to assess the difficulty of exhaustive modification extraction.

2 Event representation

To be able to benefit from the substantial number of existing resources and systems for event extraction, we apply the event representation of the BioNLP Shared Task (ST) for annotating protein modifications. Specifically, we directly extend the approach of the BioNLP ST 2011 EPI task (Ohta et al., 2011). In brief, in the applied representation, each event is marked as being expressed by a specific span of text (the *event trigger*) and assigned a type from a fixed ontology defining event types. Events can take a conceptually open-ended number of participants, each of which is similarly bound to a specific textual expression and marked as participating in the event in a specific role. In this work, we apply three roles: *Theme* identifies the entity or event that is affected by the event (e.g. the protein that is modified), *Cause* its cause, and *Site* specifies a specific part on the *Theme* participant that is affected, i.e. the modification site or region. Further, events are primary objects of annotation and can thus in turn be participants in other events as well as being marked as e.g. explicitly negated ("is not phosphorylated") or stated speculatively ("may be phosphorylated"). An event annotation example is shown in Figure 1.

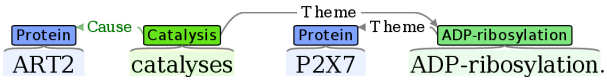

Figure 1: Illustration of the event representation. An event of type ADP-RIBOSYLATION (expressed through the text "ADP-ribosylation") with a PROTEIN ("P2X7") participant in the *Theme* role is in turn the *Theme* of a CATALYSIS event with another PROTEIN ("ART2") as its *Cause*.

3 Protein Modifications

We next present our selection of protein modification types relevant to event annotation and an extended analysis of their relative prominence.

3.1 Protein Modifications in Ontologies

For mapping and structuring the space of protein modification concepts, we primarily build on the community-standard Gene Ontology (GO) (Ashburner et al., 2000). GO has substantial representation of protein modifications: the sub-ontology rooted at `protein modification process` (*GO:0006464*) in the GO biological process ontology contains 805 terms[1] (including both leaf and internal nodes). This set of terms is the starting point for our selection of modifications types to target.

First, many specific GO terms can be excluded due to the different approach to semantic representation taken in event annotation: while GO terms represent detailed concepts without explicit structure (see e.g. (Ogren et al., 2004)), the event representation is structured, allowing more general terms to be applied while capturing the same information. For example, many GO modification terms have child nodes that identify the target (substrate) of modification, e.g. `protein phosphorylation` has the child `actin phosphorylation`. In the event representation, the target of modification is captured through the *Theme* argument. Similarly, GO terms may identify the site or region of modification, which becomes a *Site* argument in the event representation (see Figure 2). To avoid redundancy, we exclude GO terms that differ from a more general included term only in specifying a substrate or modification site. We similarly exclude terms that specify a catalyst or refer to regulation of modifi-

[1] GO structure and statistics from data retrieved Dec. 2010.

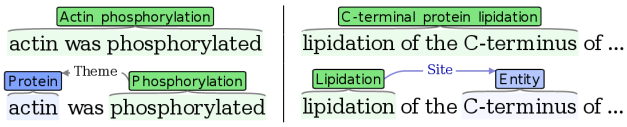

Figure 2: Comparison of hypothetical text-bound GO annotation with specific terms (top) and event annotation with general GO terms (bottom).

cation, as these are captured using separate events in the applied representation, as illustrated in Figure 1. For an analogous reason, we do not separately include type-level distinctions for "magnitude" variants of terms (e.g. `monoubiquitination`, `polyubiquitination`) as these can be systematically modeled as aspects that can mark any event (cf. the low/neutral/high *Manner* of Nawaz et al. (2010)).

Second, a number of the GO terms identify reactions that are in scope of previously defined (non-modification) event types in existing resources. To avoid introducing redundant or conflicting annotation with e.g. the GENIA Event corpus (Kim et al., 2008) or BioNLP ST resources, we excluded terms that involve predominantly (or exclusively) non-covalent binding (included in the scope of the event type BINDING) and terms involving the removal of or binding between the amino acids of a protein, including protein maturation by peptide bond cleavage (annotated – arguably somewhat inaccurately – as PROTEIN CATABOLISM in GENIA/BioNLP ST data). By contrast, we do differentiate between reactions involving the addition of chemical groups or small proteins and those involving their removal, including e.g. PALMITOYLATION and DEPALMITOYLATION as distinct types. To preserve the ontology structure, we further include also internal nodes appearing in GO for the purposes of structuring the ontology (e.g. `small protein conjugation or removal`), although we only apply more specific leaf nodes in event annotation.

This selection, aiming to identify the maximal subset of the protein modification branch of the GO ontology relevant to event annotation, resulted in the inclusion of 74 terms, approximately 9% of the branch total. Table 1 shows the relevant part of the GO `protein modification` subontology term structure, showing each term only once[2] and excluding very rare terms for space. (A detailed description of other information in the table is given in the following sections.)

In addition to GO, we consider protein modifications in the MeSH ontology,[3] used to index PubMed citations with concepts relevant to them. Further, for resolving cases not appearing in GO, we refer to the Uniprot controlled vocabulary of posttranslational modifications[4] and the Proteomics Standards Initiative Protein Modification Ontology[5] (PSI-MOD) (Montecchi-Palazzi et al., 2008).

3.2 Protein Modifications in Databases

A substantial number of databases tracking protein modifications from a variety of perspectives exist, and new ones are introduced regularly. The databases range from the specific (e.g. (Gupta et al., 1999; Diella et al., 2004; Zhang et al., 2010)) to the broad in scope (Lee et al., 2005; Li et al., 2009). Information on protein modifications is also found in general protein knowledge resources such as Swiss-Prot (Boeckmann et al., 2003) and PIR (Wu et al., 2003). The relative number of entries relevant to each protein modification in such resources is one possible proxy for the biological significance of the various modifications. We apply two such estimates in this work.

One of the primary applications of GO is the use of the ontology terms to annotate gene products, identifying their functions. These annotations, provided by a variety of groups in different efforts (e.g. (Camon et al., 2004)), are readily available in GO and used in various GO tools as a reflection of the prominence of each of the ontology concepts. As GO is a community standard with wide participation and a primary source in this work, we give these annotation numbers priority in introducing an additional filter: we exclude from detailed analysis any term that has no gene product association annotations, taking this as an indication that the modifica-

[2] GO allows multiple inheritance, and e.g. `protein palmitoylation` occurs under both `protein lipidation` and `protein acylation` reflecting the biological definition.

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

[4] `http://www.uniprot.org/docs/ptmlist`

[5] `http://www.psidev.info/MOD`

Term	GO ID	GPA	SysPTM	PubMed	GENIA	Ohta'10	EPI	This study
phosphorylation	GO:0006468	8246	24705	93584	546	3	130	85
small protein conj./removal	GO:0070647							
small protein conjugation	GO:0032446							
ubiquitination	GO:0016567	1724	439	4842	6	-	340	52
sumoylation	GO:0016925	121	260	886	-	-	-	101
neddylation	GO:0045116	66	2	100	-	-	-	52
ufmylation	GO:0071569	33	-	1	-	-	-	-
urmylation	GO:0032447	16	-	7	-	-	-	-
pupylation	GO:0070490	11	-	15	-	-	-	-
small protein removal	GO:0070646							
deubiquitination	GO:0016579	360	-	206	0	-	17	2
deneddylation	GO:0000338	45	-	39	-	-	-	8
desumoylation	GO:0016926	20	-	45	-	-	-	3
dephosphorylation	GO:0006470	1479	121	8339	28	-	3	1
glycosylation	GO:0006486	1145	2982	12619	-	122	347	62
acylation	GO:0043543	1	-	1728	-	-	-	71
acetylation	GO:0006473	522	2000	4423	7	90	337	17
palmitoylation	GO:0018345	49	198	1009	-	-	-	187
myristoylation	GO:0018377	27	150	895	-	-	-	34
octanoylation	GO:0018190	4	-	11	-	-	-	-
palmitoleylation	GO:0045234	3	-	0	-	-	-	-
alkylation	GO:0008213	0						
methylation	GO:0006479	552	499	9749	-	90	374	18
lipidation	GO:0006497	34	51	258	-	-	-	16
prenylation	GO:0018342	64	111	822	-	-	-	71
farnesylation	GO:0018343	19	-	118	-	-	-	48
geranylgeranylation	GO:0018344	26	-	79	-	-	-	30
deacylation	GO:0035601	1	-	331	-	-	-	1
deacetylation	GO:0006476	320	6	1056	1	-	50	4
depalmitoylation	GO:0002084	9	-	81	-	-	-	9
ADP-ribosylation	GO:0006471	261	9	3113	-	-	-	52
cofactor linkage	GO:0018065							
lipoylation	GO:0009249	53	-	49	-	-	-	14
FAD linkage	GO:0018293	46	-	6	-	-	-	-
pyridoxal-5-phosphate linkage	GO:0018352	6	-	0	-	-	-	-
dealkylation	GO:0008214	0						
demethylation	GO:0006482	116	-	1465	-	-	13	1
deglycosylation	GO:0006517	22	1	1204	-	-	27	0
ISG15-protein conjugation	GO:0032020	20	-	3	-	-	-	-
arginylation	GO:0016598	20	-	46	-	-	-	-
hydroxylation	GO:0018126	20	226	2948	-	103	139	3
sulfation	GO:0006477	18	132	960	-	-	-	37
carboxylation	GO:0018214	17	7	595	-	-	-	34
nucleotidylation	GO:0018175	0						
adenylylation	GO:0018117	16	-	116	-	-	-	-
uridylylation	GO:0018177	1	-	105	-	-	-	-
polyglycylation	GO:0018094	17	-	14	-	-	-	-
de-ADP-ribosylation	GO:0051725	16	-	7	-	-	-	5
nitrosylation	GO:0017014	14	-	670	-	-	-	-
glutathionylation	GO:0010731	11	-	279	-	-	-	-
biotinylation	GO:0009305	8	-	1247	-	-	-	4
deglutathionylation	GO:0080058	3	-	42	-	-	-	-
delipidation	GO:0051697	3	-	303	-	-	-	-
oxidation	GO:0018158	3	475	23413	-	-	-	21
phosphopantetheinylation	GO:0018215	3	-	26	-	-	-	-
tyrosinylation	GO:0018322	2	-	2	-	-	-	-
deamination	GO:0018277	1	-	840	-	-	-	-
esterification	GO:0018350	1	-	1180	-	-	-	-
glucuronidation	GO:0018411	1	-	705	-	-	-	-
polyamination	GO:0018184	1	-	13	-	-	-	-

Table 1: Protein modifications and protein modification resources. GO terms shown abbreviated, mostly by removing "protein" (e.g. "acylation" instead of "protein acylation"). Terms with 0 GPA not shown except when required for structure. Columns: GPA: number of Gene Product Associations for each term in GO (not including counts of more specific child nodes), SysPTM: number of SysPTM modification entries (excluding sites), PubMed: PubMed query matches (see Section 3.3), GENIA: GENIA corpus (Kim et al., 2008), Ohta'10: corpus introduced in Ohta et al. (2010), EPI: BioNLP ST'11 EPI task corpus (Ohta et al., 2011) (excluding test set).

tion is not presently established as having high biological significance.[6]

In addition to the GO associations, we include an estimate based on dedicated protein modification databases. We chose to use the integrated SysPTM resource (Li et al., 2009), which incorporates data from five databases, four webservers, and manual extraction from the literature. In its initial release, SysPTM included information on "nearly 50 modification types" on over 30,000 proteins. The columns labeled *GPA* and *SysPTM* in Table 1 show the number of gene product associations for each selected type in GO and entries per type in SysPTM, respectively.

3.3 Protein Modifications in domain literature

As a final estimate of the relative prominence of the various protein modification types, we estimated the relative frequency with which they are discussed in the literature through simple PubMed search, querying the Entrez system for each modification in its basic nominalized form (e.g. *phosphorylation*) in a protein-related article. Specifically, for each modification string MOD we searched Entrez for

"MOD"[TIAB] AND "protein"[TIAB]

The modifier *[TIAB]* specifies to search the title and abstract. The literal string "protein" is included to improve the estimate by removing references that involve the modification of non-proteins or related concepts that happen to share the term.[7] While this query is far from a perfect estimate of the actual number of protein modifications, we expect it to be a useful as a rough indicator of their relative frequencies and more straightforward to assess than more involved statistical analyses (e.g. (Pyysalo et al., 2010)). The results for these queries are given in the *PubMed* column of Table 1.

[6]We are also aware that GO coverage of protein modifications is not perfect: for example, citrullination, eliminylation, sialylation, as well as a number of reverse reactions for addition reactions in the ontology (e.g. demyristoylation) are not included at the time of this writing. As for terms with no gene product associations, we accept these omissions as indicating that these modifications are not biologically prominent.

[7]For example, search for only *dehydration* – a modification with zero GPA in GO – matches nearly 10 times as many documents as search including *protein*, implying that most of the hits for the former query likely do not concern protein modification by dehydration. By contrast, the majority of hits for *phosphorylation* match also *phosphorylation AND protein*.

3.4 Protein Modifications in Event Resources

The rightmost four columns of Table 1 present the number of annotations for each modification type in previously introduced event-annotated resources following the BioNLP ST representation as well as those annotated in the present study. While modification annotations are found also in other corpora (e.g. (Wu et al., 2003; Pyysalo et al., 2007)), we only include here resources readily compatible with the BioNLP ST representation.

Separating for the moment from consideration the question of what level of practical extraction performance can be supported by these event annotations, we can now provide an estimate of the upper bound on the coverage of relevant modification statements for each of the three proxies (GO GPA, SysPTM DB entries, PubMed query hits) simply by dividing the sum of instances of modifications for which annotations exist by the total. Thus, for example, there are 8246 GPA annotations for `Phosphorylation` and a total of 15597 GPA annotations, so the BioNLP ST'09 data (containing only PHOSPHORYLATION events) could by the GPA estimate cover 8246/15597, or approximately 53% of individual modifications.[8]

For the total coverage of the set of types for which event annotation is available given the corpus introduced in this study, the coverage estimates are: GO GPA: 98.2%, SysPTM 99.6%, PubMed 97.5%. Thus, we estimate that correct extraction of the included types would, depending on whether one takes a gene association, database entry, or literature mention point of view, cover between 97.5% to 99.6% of protein modification instances – a level of coverage we suggest is effectively exhaustive for most practical purposes. We next briefly describe our annotation effort before discarding the assumption that correct extraction is possible and measuring actual extraction performance.

4 Annotation

This section presents the entity and event annotation approach, document selection, and the statistics of the created annotation.

[8]The remarkably high coverage for a single type reflects the Zipfian distribution of the modification types; see e.g. Ohta et al. (2010).

4.1 Entity and Event Annotation

To maximize compatibility with existing event-annotated resources, we chose to follow the general representation and annotation guidelines applied in the annotation of GENIA/BioNLP ST resources, specifically the BioNLP ST 2011 EPI task corpus. Correspondingly, we followed the GENIA gene/gene product (Ohta et al., 2009) annotation guidelines for marking protein mentions, extended the GENIA event corpus guidelines (Kim et al., 2008) for the annotation of protein modification events, and marked CATALYSIS events following the EPI task representation. For compatibility, we also marked event negation and speculation as in these resources. We followed the GO definitions for individual modification types, and in the rare cases where a modification discussed in text had no existing GO definition, we extrapolated from the way in which protein modifications are generally defined in GO, consulting other domain ontologies and resources (Section 3.1) as necessary.

4.2 Document Selection

As the distribution of protein modifications in PubMed is extremely skewed, random sampling would recover almost solely instances of major types such as phosphorylation. As we are interested also in the extraction of very rare modifications, we applied a document selection strategy targeted at individual modification types. We applied one of two primary strategies depending on whether each targeted modification type had a corresponding MeSH term or not. If a MeSH term specific to the modification exists, we queried PubMed for the MeSH term, thus avoiding searches for specific forms of expression that might bias the search. In cases where no specific MeSH term existed, we searched the text of documents marked with the generic MeSH term `protein processing, post-translational` for mentions of likely forms of expression for the modification.[9] Finally, in a few isolated instances we applied custom text-based PubMed searches with broader coverage. Then, as many of the modifications are not limited to protein substrates, to select documents relating specifically to *protein* modification we proceeded to tagged a large random sample of selected documents with the BANNER named entity tagger (Leaman and Gonzalez, 2008) trained on the GENETAG corpus (Tanabe et al., 2005) and removed documents with fewer than five automatically tagged gene/protein-related entities. The remaining documents were then randomly sampled for annotation.[10]

[9] Specifically, we applied a regular expression incorporating the basic form of modification expression and allowing variance through relevant affixes and inflections derived from an initial set of annotations for documents for which MeSH terms were defined.

Item	Count
Abstract	360
Word	76806
Protein	4698
Event type	37
Event instance	1142

Table 2: Annotation statistics.

4.3 Corpus Statistics

We initially aimed to annotate balanced numbers of modification types in order of their estimated prominence, with particular focus on previously untargeted reaction types involving the addition of chemical groups or small proteins. However, it became apparent in the annotation process that the extreme rarity of some of the modifications as well as the tendency for more frequent modifications to be discussed in texts mentioning rare ones made this impossible. Thus, while preserving the goal of establishing broadly balanced numbers of major new modifications, we allowed the number of rare reactions to remain modest.

Table 2 summarizes the statistics of the final corpus, and the rightmost column of Table 1 shows per-type counts. We note that as reactions involving the removal of chemical groups or small proteins were not separately targeted, only few events of such types were annotated. We did not separately measure inter-annotator agreement for this effort, but note that this work is an extension of the EPI corpus annotation, for which comparison of independently created event annotations indicated an F-score of 82% for the full task and 89% for the core targets (see Section 5.1) (Ohta et al., 2011).

[10] This strategy, including MeSH-based search, was applied also in the BioNLP Shared Task 2011 EPI task document selection.

5 Experiments

To assess actual extraction performance, we performed experiments using a state-of-the art event extraction system.

5.1 Experimental Setup

We first split the corpus into a training/development portion and a held out set for testing, placing half of the abstracts into each set. The split was stratified by event type to assure that relatively even numbers of each event type were present in both sets. All development was performed using cross-validation on the visible portion of the data, and a single final experiment was performed on the test dataset.

To assure that our results are comparable with those published in recent event extraction studies, we adopted the standard evaluation criteria of the BioNLP Shared Task. The evaluation is event instance-based and uses the standard precision/recall/F_1-score metrics. We modified the shared task evaluation software to support the newly defined event types and ran experiments with the standard *approximate span matching* and *partial recursive matching* criteria (see (Kim et al., 2009)). We further follow the EPI task evaluation in reporting results separately for the extraction of only *Theme* and *Cause* arguments (*core* task) and for the *full* argument set.

5.2 Event extraction method

We applied the EventMine event extraction system (Miwa et al., 2010a; Miwa et al., 2010b), an SVM-based pipeline system using an architecture similar to that of the best-performing system in the BioNLP ST'09 (Björne et al., 2009); we refer to the studies of Miwa et al. for detailed description of the base system. For analysing sentence structure, we applied the mogura 2.4.1 (Matsuzaki and Miyao, 2007) and GDep beta2 (Sagae and Tsujii, 2007) parsers.

For the present study, we modified the base EventMine system as follows. First, to improve efficiency and generalizability, instead of using all words as trigger candidates as in the base system, we filtered candidates using a dictionary extracted from training data and expanded by using the UMLS specialist lexicon (Bodenreider, 2004) and the "hypernyms" and "similar to" relations in WordNet (Fellbaum, 1998). Second, to allow generalization across argument types, we added support for solving a single classification problem for event argument detection instead of solving multiple classification problems separated by argument types. Finally, to facilitate the use of other event resources for extraction, we added functionality to incorporate models trained by other corpora as reference models, using predictions from these models as features in classification.

5.3 Experimental results

We first performed a set of experiments to determine whether models can beneficially generalize across different modification event types. The EventMine pipeline has separate classification stages for event trigger detection, event-argument detection, and the extraction of complete event structures. Each of these stages involves a separate set of features and output labels, some of which derive directly from the involved event types: for example, in determining whether a specific entity is the *Theme* of an event triggered by the string "phosphorylation", the system by default uses the predicted event type (PHOSPHORYLATION) among its features. It is possible to force the model to generalize across event types by replacing specific types with placeholders, for example replacing PHOSPHORYLATION, METHYLATION, etc. with MODIFICATION.

In preliminary experiments on the development set, we experimented with a number of such generalizations. Results indicated that while some generalization was essential for achieving good extraction performance, most implementation variants produced broadly comparable results. We chose the following generalizations for the final test: in the trigger detection model, no generalization was performed (allowing specific types to be extracted), for argument detection, all instances of event types were replaced with a generic type (EVENT), and for event structure prediction, all instances of specific modification event types (but not CATALYSIS) were replaced with a generic type (MODIFICATION). Results comparing the initial, ungeneralized model to the generalized one are shown in the top two rows of Table 3. The results indicate that generalization is clearly beneficial: attempting to learn each of the event types in isolation leaves F-score results approximately 4-5% points lower than when general-

	Core	Full
Initial	39.40 / 46.36 / 42.60	31.39 / 38.88 / 34.74
Generalized	39.02 / 61.18 / 47.65	31.07 / 51.89 / 38.87
+Model	41.28 / 61.28 / 49.33	33.66 / 53.06 / 41.19
+Ann	38.46 / 66.99 / 48.87	32.36 / 59.17 / 41.84
+Model +Ann	41.84 / 66.17 / 51.26	33.98 / 56.00 / 42.30
Test data	45.69 / 62.35 / 52.74	38.03 / 54.57 / 44.82

Table 3: Experimental results.

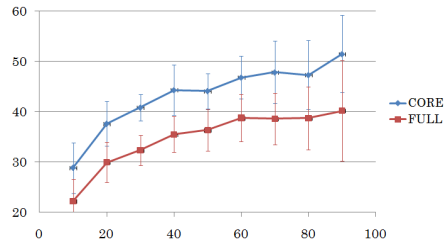

Figure 3: Learning curve.

izing across types. A learning curve for the generalized model is shown in Figure 3. While there is some indication of decreasing slope toward use of the full dataset, the curve suggests performance could be further improved through additional annotation efforts.

In a second set of experiments, we investigated the compatibility of the newly introduced annotations with existing event resources by incorporating their annotations either directly as training data (+Ann) or indirectly through features from predictions from a model trained on existing resources (+Model), as well as their combination. We performed experiments with the BioNLP Shared Task 2011 EPI task corpus[11] and the generalized setting. The results of these experiments are given in the middle rows of Table 3. We find substantial benefit from either form of existing resource integration alone, and, interestingly, an indication that the benefits of the two approaches can be combined. This result indicates that the newly introduced corpus is compatible with the EPI corpus, a major previously introduced resource for protein modification event extraction. Evaluation on the test data (bottom row of Table 3) confirmed that development data results were not overfit and generalized well to previously unseen data.

6 Discussion and Conclusions

We have presented an effort to directly address the challenges involved in the exhaustive extraction of protein modifications in text. We analysed the Gene Ontology `protein modification process` subontology from the perspective of event extraction for information extraction, arguing that due largely to the structured nature of the event representation, 74 of the 805 ontology terms suffice to capture the general modification types included. Through an analysis of the relative prominence of protein modifications in ontology annotations, domain databases, and literature, we then filtered and prioritized these types, estimating that correct extraction of the most prominent half of these types would give 97.5%-99.6% coverage of protein modifications, a level that is effectively exhaustive for practical purposes.

To support modification event extraction and to estimate actual extraction performance, we then proceeded to manually annotate a corpus of 360 PubMed abstracts selected for relevance to the selected modification types. The resulting corpus annotation marks over 4500 proteins and over 1000 instances of modification events and more than triples the number of specific protein modification types for which text-bound event annotations are available. Experiments using a state-of-the-art event extraction system showed that a machine learning method can beneficially generalize features across different protein modification event types and that incorporation of BioNLP Shared Task EPI corpus annotations can improve performance, demonstrating the compatibility of the created resource with existing event corpora. Using the best settings on the test data, we found that the core extraction task can be performed at 53% F-score.

The corpus created in this study is freely available for use in research from `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA`.

Acknowledgments

We would like to thank Yo Shidahara and Yoshihiro Okuda of NalaPro Technologies for their efforts in creating the corpus annotation. This work was supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

[11] When combining EPI annotations directly as additional training abstracts, we filtered out abstracts including possible "missing" annotations for modification types not annotated in EPI data using a simple regular expression.

References

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25:25–29.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*.

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 BioNLP'09 Shared Task*, pages 10–18.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Scaling up biomedical event extraction to the entire pubmed. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 28–36.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–70.

B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, 31(1):365.

D.F. Brennan and D. Barford. 2009. Eliminylation: a post-translational modification catalyzed by phosphothreonine lyases. *Trends in biochemical sciences*, 34(3):108–114.

Ekaterina Buyko, Erik Faessler, Joachim Wermter, and Udo Hahn. 2009. Event extraction from trimmed dependency graphs. In *Proceedings of the BioNLP'09 Shared Task*, pages 19–27, Boulder, Colorado, June. Association for Computational Linguistics.

Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucl. Acids Res.*, 32(suppl 1):D262–266.

Francesca Diella, Scott Cameron, Christine Gemund, Rune Linding, Allegra Via, Bernhard Kuster, Thomas Sicheritz-Ponten, Nikolaj Blom, and Toby Gibson. 2004. Phospho.elm: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5(1):79.

C. Fellbaum. 1998. Wordnet: an electronic lexical database. In *International Conference on 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.

Ramneek Gupta, Hanne Birch, Kristoffer Rapacki, Sren Brunak, and Jan E. Hansen. 1999. O-glycbase version 4.0: a revised database of o-glycosylated proteins. *Nucleic Acids Research*, 27(1):370–372.

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.

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. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: An executable survey of advances in biomedical named entity recognition. In *Proceedings of PSB'08*, pages 652–663.

Tzong-Yi Lee, Hsien-Da Huang, Jui-Hung Hung, Hsi-Yuan Huang, Yuh-Shyong Yang, and Tzu-Hao Wang. 2005. dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Research*, 34(suppl 1):D622–D627.

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.

Hong Li, Xiaobin Xing, Guohui Ding, Qingrun Li, Chuan Wang, Lu Xie, Rong Zeng, and Yixue Li. 2009. Sysptm: A systematic resource for proteomic research on post-translational modifications. *Molecular & Cellular Proteomics*, 8(8):1839–1849.

Takuya Matsuzaki and Yusuke Miyao. 2007. Efficient HPSG parsing with supertagging and CFG-filtering. In *In Proceedings of IJCAI-07*, pages 1671–1676.

Makoto Miwa, Sampo Pyysalo, Tadayoshi Hara, and Jun'ichi Tsujii. 2010a. Evaluating dependency representations for event extraction. In *Proceedings of Coling'10*, pages 779–787.

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

Luisa Montecchi-Palazzi, Ron Beavis, Pierre-Alain Binz, Robert Chalkley, John Cottrell, David Creasy, Jim Shofstahl, Sean Seymour, and John Garavelli. 2008. The PSI-MOD community standard for representation of protein modification data. *Nature Biotechnology*, 26:864–866.

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.

R. Nawaz, P. Thompson, J. McNaught, and S. Ananiadou. 2010. Meta-Knowledge Annotation of Bio-Events. *Proceedings of LREC 2010*, pages 2498–2507.

P.V. Ogren, K.B. Cohen, G.K. Acquaah-Mensah, J. Eberlein, and L. Hunter. 2004. The compositional structure of Gene Ontology terms. In *Pacific Symposium on Biocomputing*, page 214.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. 2009. Incorporating GENETAG-style annotation to GENIA corpus. In *Proceedings of BioNLP'09*, pages 106–107.

Tomoko Ohta, Sampo Pyysalo, Makoto Miwa, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction for post-translational modifications. In *Proceedings of BioNLP'10*, pages 19–27.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*.

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).

Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. 2010. An analysis of gene/protein associations at pubmed scale. In *Proceedings of the fourth International Symposium for Semantic Mining in Biomedicine (SMBM 2010)*.

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.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of EMNLP-CoNLL'07*, pages 1044–1050.

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.

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

Christopher Walsh. 2006. *Posttranslational modification of proteins: expanding nature's inventory*. Roberts & Company Publishers.

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.

X. Yuan, ZZ Hu, HT Wu, M. Torii, M. Narayanaswamy, KE Ravikumar, K. Vijay-Shanker, and CH Wu. 2006. An online literature mining tool for protein phosphorylation. *Bioinformatics*, 22(13):1668.

Yan Zhang, Jie Lv, Hongbo Liu, Jiang Zhu, Jianzhong Su, Qiong Wu, Yunfeng Qi, Fang Wang, and Xia Li. 2010. Hhmd: the human histone modification database. *Nucleic Acids Research*, 38(suppl 1):D149–D154.