Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011

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

Abstract

This paper presents the preparation, resources, results and analysis of the Epigenetics and Post-translational Modifications (EPI) task, a main task of the BioNLP Shared Task 2011. The task concerns the extraction of detailed representations of 14 protein and DNA modification events, the catalysis of these reactions, and the identification of instances of negated or speculatively stated event instances. Seven teams submitted final results to the EPI task in the shared task, with the highest-performing system achieving 53% F-score in the full task and 69% F-score in the extraction of a simplified set of core event arguments.

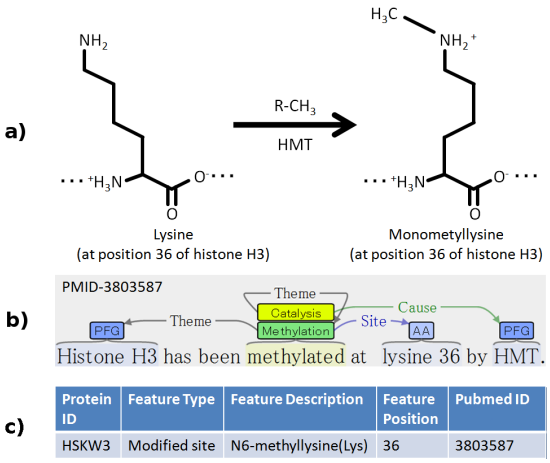
1 Introduction

The Epigenetics and Post-translational Modifications (EPI) task is a shared task on event extraction from biomedical domain scientific publications, first introduced as a main task in the BioNLP Shared Task 2011 (Kim et al., 2011a).

The EPI task focuses on events relating to epigenetic change, including DNA methylation and histone methylation and acetylation (see e.g. (Holliday, 1987; Jaenisch and Bird, 2003)), as well as other common protein post-translational modifications (PTMs) (Witze et al., 2007). PTMs are chemical modifications of the amino acid residues of proteins, and DNA methylation a parallel modification of the nucleotides on DNA. While these modifications are chemically simple reactions and can thus be straightforwardly represented in full detail, they have a crucial role in the regulation of gene expression and protein function: the modifications can alter the conformation of DNA or proteins and thus control their ability to associate with other molecules, making PTMs key steps in protein biosynthesis for introducing the full range of protein functions. For instance, protein phosphorylation – the attachment of phosphate – is a common mechanism for activating or inactivating enzymes by altering the conformation of protein active sites (Stock et al., 1989; Barford et al., 1998), and protein ubiquitination – the post-translational attachment of the small protein ubiquitin – is the first step of a major mechanism for the destruction (breakdown) of many proteins (Glickman and Ciechanover, 2002).

Many of the PTMs targeted in the EPI task involve modification of histone, a core protein that forms an octameric complex that has a crucial role in packaging chromosomal DNA. The level of methylation and acetylation of histones controls the tightness of the chromatin structure, and only "unwound" chromatin exposes the gene packed around the histone core to the transcriptional machinery. Since histone modification is of substantial current interest in epigenetics, we designed aspects of the EPI task to capture the full detail in which histone modification events are stated in text. Finally, the DNA methylation of gene regulatory elements controls the expression of the gene by altering the affinity with which DNA-binding proteins (including transcription factors) bind, and highly methylated genes are not transcribed at all (Riggs, 1975; Holliday and Pugh, 1975). DNA methylation can thus "switch off" genes, "removing" them from the genome in a way that is reversible through DNA demethylation.

Proceedings of BioNLP Shared Task 2011 Workshop, pages 16–25,
Portland, Oregon, USA, 24 June, 2011.



Protein ID	Feature Type	Feature Description	Feature Position	Pubmed ID
HSKW3	Modified site	N6-methyllysine(Lys)	36	3803587

Figure 1: Three views of protein methylation. a) chemical formula b) event representation c) modification database entry.

The BioNLP'09 Shared Task on Event Extraction (Kim et al., 2009), the first task in the present shared task series, involved the extraction of nine event types including one PTM type, PHOSPHORYLATION. The results of the shared task showed this PTM event to be the single most reliably extracted event type in the task, with the best-performing system for the type achieving 91% precision and 76% recall (83% F-score) in its extraction (Buyko et al., 2009). The results suggest both that the event representation is well applicable to PTM extraction and that current extraction methods are capable of reliable PTM extraction. The EPI task follows up on these opportunities, introducing specific, strongly biologically motivated extraction targets that are expected to be both feasible for high-accuracy event extraction, relevant to the needs of present-day molecular biology, and closely applicable to biomolecular database curation needs (see Figure 1) (Ohta et al., 2010a).

2 Task Setting

The EPI task is an *event extraction task* in the sense popularized by a number of recent domain resources and challenges (e.g. (Pyysalo et al., 2007; Kim et al., 2008; Thompson et al., 2009; Kim et al., 2009; Ananiadou et al., 2010)). In broad outline, the task focuses on the extraction of information on statements regarding change in the state or properties of (physical) entities, modeled using an *event representation*.

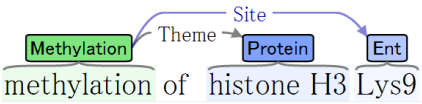


Figure 2: Illustration of the event representation. An event of type METHYLATION (expressed through the text "methylation") with two participants of the types PROTEIN ("histone H3") and ENTITY ("Lys9"), participating in the event in *Theme* and *Site* roles, respectively.

In this representation, events are typed n-ary associations of participants (entities or other events) in specific roles. Events are bound to specific expressions in text (the *event trigger* or *text binding*) and are primary objects of annotation, allowing them to be marked in turn e.g. as negated or as participants in other events. Figure 2 illustrates these concepts.

In its specific formulation, EPI broadly follows the definition of the BioNLP'09 shared task on event extraction. Basic modification events are defined similarly to the PHOSPHORYLATION event type targeted in the '09 and the 2011 GE and ID tasks (Kim et al., 2011b; Pyysalo et al., 2011b), with the full task extending previously defined arguments with two additional ones, *Sidechain* and *Contextgene*.

2.1 Entities

The EPI task follows the general policy of the BioNLP Shared Task in isolating the basic task of named entity recognition from the event extraction task by providing task participants with manually annotated gene and gene product entities as a starting point for extraction. The entity types follow the BioNLP'09 Shared Task scheme, where genes and their products are simply marked as PROTEIN.[1]

In addition to the given PROTEIN entities, some events involve other entities, such as the modification *Site*. These entities are not given and must thus be identified by systems targeting the full task (see Section 4). In part to reduce the demands of this entity recognition component of the task, these additional entities are not given specific types but are generically marked as ENTITY.

[1] While most of the modifications targeted in the task involve proteins, this naming is somewhat inaccurate for the *Theme*s of DNA METHYLATION and DNA DEMETHYLATION events and for *Contextgene* arguments, which refer to genes. Despite this inaccuracy, we chose to follow this naming scheme for consistency with other tasks.

Type	Core arguments	Additional arguments
HYDROXYLATION	*Theme*(PROTEIN)	*Site*(ENTITY)
PHOSPHORYLATION	*Theme*(PROTEIN)	*Site*(ENTITY)
UBIQUITINATION	*Theme*(PROTEIN)	*Site*(ENTITY)
DNA METHYLATION	*Theme*(PROTEIN)	*Site*(ENTITY)
GLYCOSYLATION	*Theme*(PROTEIN)	*Site*(ENTITY), *Sidechain*(ENTITY)
ACETYLATION	*Theme*(PROTEIN)	*Site*(ENTITY), *Contextgene*(PROTEIN)
METHYLATION	*Theme*(PROTEIN)	*Site*(ENTITY), *Contextgene*(PROTEIN)
CATALYSIS	*Theme*(Event), *Cause*(PROTEIN)	

Table 1: Event types and their arguments. The type of entity allowed as argument is specified in parenthesis. For each event type except CATALYSIS, the reverse reaction (e.g. DEACETYLATION for ACETYLATION) is also defined, with identical arguments. The total number of event types in the task is thus 15.

2.2 Relations

The EPI task does not define any explicit relation extraction targets. However, the task annotation involves one relation type, EQUIV. This is a binary, symmetric, transitive relation between entities that defines two entities to be equivalent (Hoehndorf et al., 2010). The relation is used in the gold annotation to mark local aliases such as the full and abbreviated forms of a protein name as referring to the same real-world entity. While the '09 task only recognized equivalent PROTEIN entities, EPI extends on the scope of EQUIV annotations in allowing entities of any type to be marked equivalent. In evaluation, references to any of a set of equivalent entities are treated identically.

2.3 Events

While the EPI task entity definition closely follows that of the previous shared task, the task introduces considerable novelty in the targeted events, adding a total of 14 novel event types and two new participant roles. Table 1 summarizes the targeted event types and their arguments.

As in the BioNLP'09 shared task, *Theme* arguments identify the entity that the event is *about*, such as the protein that is acetylated in an acetylation event. A *Theme* is always mandatory for all EPI task events. *Site* arguments identify the modification site on the *Theme* entity, such as a specific residue on a modified protein or a specific region on a methylated gene. The *Sidechain* argument, specific to GLYCOSYLATION and DEGLYCOSYLATION among the targeted events, identifies the moiety attached or removed in the event (in glycosylation, the sugar).[2] Finally, the *Contextgene* argument, specific to ACETYLATION and METHYLATION events and their reverse reactions, identifies the gene whose expression is controlled by these modifications. This argument applies specifically for histone protein modification: the modification of the histones that form the nucleosomes that structure DNA are key to the epigenetic control of gene expression. The *Site*, *Sidechain* and *Contextgene* arguments are not mandatory, and should only be extracted when explicitly stated.

For CATALYSIS events, representing the catalysis of protein or DNA modification by another protein, both *Theme* and *Cause* are mandatory. While CATALYSIS is a new event type, it is related to the '09 POSITIVE_REGULATION type by a class-subclass relation: any CATALYSIS event is a POSITIVE_REGULATION event in the '09 task terms (but not vice versa).

2.4 Event modifications

In addition to events, the EPI task defines two *event modification* extraction targets: NEGATION and SPECULATION. Both are represented as simple binary "flags" that apply to events, marking them as being explicitly negated (e.g. *H2A is not methylated*) or stated in a speculative context (e.g. *H2A may be methylated*). Events may be both negated and speculated.

[2]Note that while arguments similar to *Sidechain* could be defined for other event types also, their extraction would provide no additional information: the attached molecule is always acetyl in acetylation, methyl in methylation, etc.

3 Data

The primary EPI task data were annotated specifically for the BioNLP Shared Task 2011 and are not based on any previously released resource. Before starting this annotation effort, we performed two preparatory studies using in part previously released related datasets: in (Ohta et al., 2010a) we considered the extraction of four protein post-translational modifications event types with reference to annotations originally created for the Protein Information Resource[3] (PIR) (Wu et al., 2003), and in (Ohta et al., 2010b) we studied the annotation and extraction of DNA methylation events with reference to annotations created for the PubMeth[4] (Ongenaert et al., 2008) database. The corpus text selection and annotation scheme were then defined following the understanding formed in these studies.

3.1 Document selection

The texts for the EPI task corpus were drawn from PubMed abstracts. In selecting the primary corpus texts, we aimed to gather a representative sample of all PubMed documents relevant to selected modification events, avoiding bias toward, for example, specific genes/proteins, species, forms of event expression, or subdomains. We primarily targeted DNA methylation and the "prominent PTM types" identified in (Ohta et al., 2010a). We defined the following document selection protocol: for each of the targeted event types, 1) Select a random sample of PubMed abstracts annotated with the MeSH term corresponding to the target event (e.g. `Acetylation`) 2) Automatically tag protein/gene entities in the selected abstracts, removing ones where fewer than a specific cutoff are found 3) Perform manual filtering removing documents not relevant to the targeted topic (optional).

MeSH is a controlled vocabulary of over 25,000 terms that is used to manually annotate each document in PubMed. By performing initial document retrieval using MeSH terms it is possible to select relevant documents without bias toward specific expressions in text. While search for documents tagged with e.g. the `Acetylation` MeSH term is sufficient to select documents relevant to the modification, not all such documents necessarily concern specifically protein modification, necessitating a filtering step. Following preliminary experiments, we chose to apply the BANNER named entity tagger (Leaman and Gonzalez, 2008) trained on the GENETAG corpus (Tanabe et al., 2005) and to filter documents where fewer than five entities were identified. Finally, for some modification types this protocol selected also a substantial number of non-relevant documents. In these cases a manual filtering step was performed prior to full annotation to avoid marking large numbers of non-relevant abstracts.

This primary corpus text selection protocol does not explicitly target reverse reactions such as deacetylation, and the total number of these events in the resulting corpus was low for many types. To be able to measure the extraction performance for these types, we defined a secondary selection protocol that augmented the primary protocol with a regular expression-based filter removing documents that did not (likely) contain mentions of reverse reactions. This protocol was used to select a secondary set of test abstracts enriched in mentions of reverse reactions. Performance on this secondary test set was also evaluated, but is not part of the primary task evaluation. Due to space considerations, we only present the primary test set results in this paper, referring to the shared task website for the secondary results.

3.2 Annotation

Annotation was performed manually. The gene/protein entities automatically detected in the document selection step were provided to annotators for reference for creating PROTEIN annotations, but all entity annotations were checked and revised to conform to the specific guidelines for the task.[5] For the annotation of PROTEIN entities, we adopted the GENIA gene/gene product (GGP) annotation guidelines (Ohta et al., 2009), adding one specific exception: while the primary guidelines require that only specific individual gene or gene product names are annotated, we allowed also the annotation of mentions of groups of histones or

[3] `http://pir.georgetown.edu`

[4] `http://www.pubmeth.org/`

[5] This revision was substantial: only approximately 65% of final PROTEIN annotations exactly match an automatically predicted one due to differences in annotation criteria (Wang et al., 2009).

the entire histone protein family to capture histone modification events also in cases where only the group is mentioned.

All event annotations were created from scratch without automatic support to avoid bias toward specific automatic extraction methods or approaches. The event annotation follows the GENIA event corpus annotation guidelines (Kim et al., 2008) as they apply to protein modifications, with CATALYSIS being annotated following the criteria for the POSITIVE_REGULATION event type with the additional constraints that the *Cause* of the event is a gene or gene product entity and the form of regulation is catalysis of a modification reaction.

The manual annotation was performed by three experienced annotators with a molecular biology background, with one chief annotator with extensive experience in domain event annotation organizing and supervising the annotator training and the overall process. After completion of primary annotation, we performed a final check targeting simple human errors using an automatic extraction system.[6] This correction process resulted in the revision of approximately 2% of the event annotations. To evaluate the consistency of the annotation, we performed independent event annotation (taking PROTEIN annotations as given) for a random sample of 10% of the corpus documents. Comparison of the two manually created sets of event annotations under the primary task evaluation criteria gave an F-score of 82% for the full task and 89% for the core task.[7] We found that CATALYSIS events were particularly challenging, showing just 65% agreement for the core task.

Table 2 shows the statistics of the primary task data. We note that while the corpus is broadly comparable in size to the BioNLP'09 shared task dataset (Kim et al., 2009) in terms of the number of abstracts and annotated entities, the number of annotated events in the EPI corpus is approximately 20% of that in the '09 dataset, reflecting the more focused event types.

[6]High-confidence system predictions differing from gold annotations were provided to a human annotator, not used directly to change corpus data. To further reduce the risk of bias, we only informed the annotator of the entities involved, not of the predicted event structure.

[7]Due to symmetry of precision/recall and the applied criteria, this score was not affected by the choice of which set of annotations to consider as "gold" for the comparison.

Item	Training	Devel	Test
Abstract	600	200	400
Word	127,312	43,497	82,819
Protein	7,595	2,499	5,096
Event	1,852	601	1,261
Modification	173	79	117

Table 2: Statistics of the EPI corpus. Test set statistics shown only for the primary test data.

4 Evaluation

Evaluation is instance- and event-oriented and based on the standard precision/recall/F-score[8] metrics. The primary evaluation criteria are the same as in the BioNLP'09 shared task, incorporating the "approximate span matching" and "approximate recursive matching" variants to strict matching. In brief, under these criteria text-bound annotations (event triggers and entities) in a submission are considered to match a corresponding gold annotation if their span is contained within the (mildly extended) span of the gold annotation, and events that refer to other events as arguments are considered to match if the *Theme* arguments of the recursively referred events match, that is, non-*Theme* arguments are ignored in recursively referred events. For a detailed description of these evaluation criteria, we refer to (Kim et al., 2009).

In addition to the primary evaluation criteria, we introduced a new relaxed evaluation criterion we term *single partial penalty*. Under the primary criteria, when a predicted event matches a gold event in some of its arguments but lacks one or more arguments of the gold event, the submission is arguably given a double penalty: the predicted event is counted as a false positive (FP), and the gold event is counted as a false negative (FN). Under the single partial penalty evaluation criterion, predicted events that match a gold event in all their arguments are not counted as FP, although the corresponding gold event still counts as FN (the "single penalty"). Analogously, gold events that partially match a predicted event are not counted as FN, although the corresponding predicted event with "extra" arguments counts as FP. This criterion can give a more nuanced view of performance for partially correctly predicted events.

[8]Specifically F_1. F is used for short throughout.

			NLP		Events				Other resources	
Rank	Team	Org	word	parse	trigger	arg	group	modif.	corpora	other
1	UTurku	1BI	Porter	McCCJ + SD	SVM	SVM	SVM	SVM	-	hedge words
2	FAUST	3NLP	CoreNLP, SnowBall	McCCJ + SD	(UMass+Stanford as features)			-	-	word clusters
3	MSR-NLP	1SDE, 3NLP	Porter, custom	McCCJ + SD, Enju	SVM	SVM	SVM	-	-	triggers, word clusters
4	UMass	1NLP	CoreNLP, SnowBall	McCCJ + SD	Joint, dual decomposition			-	-	-
5	Stanford	3NLP	custom	McCCJ + SD	MaxEnt	Joint, MSTParser		-	-	word clusters
6	CCP-BTMG	3BI	Porter, WN-lemma	Stanford + SD	Graph extraction & matching			-	-	-
7	ConcordU	2NLP	-	McCCJ + SD	Dict	Rules	Rules	Rules	-	triggers and hedge words

Table 3: Participants and summary of system descriptions. Abbreviations: BI=Bioinformatician, NLP=Natural Language Processing researcher, SDE=Software Development Engineer, CoreNLP=Stanford CoreNLP, Porter=Porter stemmer, Snowball=Snowball stemmer, WN-lemma=WordNet lemmatization, McCCJ=McClosky-Charniak-Johnson parser, Charniak=Charniak parser, SD=Stanford Dependency conversion, Dict=Dictionary

The full EPI task involves many partially independent challenges, incorporating what were treated in the BioNLP'09 shared task as separate subtasks: the identification of additional non-*Theme* event participants (Task 2 in '09) and the detection of negated and speculated events (Task 3 in '09). The EPI task does not include explicit subtasks. However, we specifies minimal *core* extraction targets in addition to the *full* task targets. Results are reported separately for core targets and full task, allowing participants to choose to only extract core targets. The full task results are considered the primary evaluation for the task e.g. for the purposes of determining the ranking of participating systems.

5 Results

5.1 Participation

Table 3 summarizes the participating groups and the features of their extraction systems. We note that, similarly to the '09 task, machine learning-based systems remain dominant overall, although there is considerable divergence in the specific methods applied. In addition to domain mainstays such as support vector machines and maximum entropy models, we find increased application of joint models (Riedel et al., 2011; McClosky et al., 2011; Riedel and McCallum, 2011) as opposed to pure pipeline systems (Björne and Salakoski, 2011; Quirk et al., 2011) . Remarkably, the application of full parsing together with dependency-based representations of syntactic analyses is adopted by all participants, with the parser of Charniak and Johnson (2005) with the biomedical domain model of McClosky (2009) is applied in all but one system (Liu et al., 2011) and the Stanford Dependency representation (de Marneffe et al., 2006) in all. These choices may be motivated in part by the success of systems using the tools in the previous shared task and the availability of the analyses as supporting resources (Stenetorp et al., 2011).

Despite the availability of PTM and DNA methylation resources other than those specifically introduced for the task and the PHOSPHORYLATION annotations in the GE task (Kim et al., 2011b), no participant chose to apply other corpora for training. With the exception of externally acquired unlabeled data such as PubMed-derived word clusters applied by three groups, the task results thus reflect a closed task setting in which only the given data is used for training.

5.2 Evaluation results

Table 4 presents a the primary results by event type, and Table 5 summarizes these results. We note that only two teams, UTurku (Björne and Salakoski, 2011) and ConcordU (Kilicoglu and Bergler, 2011), predicted event modifications, and only UTurku predicted additional (non-core) event arguments (data not shown). The other five systems thus addressed

	UTurku	FAUST	MSR-NLP	UMass	Stanford	CCP-BTMG	Con-cordU	Size
HYDROXYLATION	**42.25**	10.26	10.20	12.80	9.45	12.84	6.32	139
DEHYDROXYLATION	-	-	-	-	-	-	-	1
PHOSPHORYLATION	**67.12**	51.61	50.00	49.18	40.98	47.06	44.44	130
DEPHOSPHORYLATION	0.00	0.00	0.00	0.00	0.00	**50.00**	0.00	3
UBIQUITINATION	**75.34**	72.95	67.88	72.94	67.44	70.87	69.97	340
DEUBIQUITINATION	**54.55**	40.00	0.00	31.58	0.00	42.11	14.29	17
DNA METHYLATION	**60.21**	31.21	34.54	23.82	31.02	15.65	8.22	416
DNA DEMETHYLATION	**26.67**	0.00	0.00	0.00	0.00	0.00	0.00	21
Simple event total	***63.05***	*45.17*	*44.97*	*43.01*	*40.96*	*40.62*	*37.84*	*1067*
GLYCOSYLATION	**49.43**	41.10	38.87	40.00	37.22	25.62	25.94	347
DEGLYCOSYLATION	**40.00**	35.29	0.00	38.10	30.00	35.29	26.67	27
ACETYLATION	**57.22**	40.00	41.42	40.25	35.12	37.50	38.19	337
DEACETYLATION	**54.90**	28.00	31.82	29.17	21.74	24.56	27.27	50
METHYLATION	**57.67**	24.82	19.57	23.67	18.54	16.99	15.50	374
DEMETHYLATION	**35.71**	0.00	0.00	0.00	0.00	0.00	0.00	13
Non-simple event total	***54.36***	*33.86*	*31.85*	*33.07*	*29.28*	*25.06*	*25.10*	*1148*
CATALYSIS	7.06	6.58	**7.75**	5.00	2.84	7.58	1.74	238
Subtotal	***55.02***	*36.93*	*36.17*	*35.30*	*32.85*	*30.58*	*28.92*	*2453*
NEGATION	18.60	0.00	0.00	0.00	0.00	0.00	**26.51**	149
SPECULATION	**37.65**	0.00	0.00	0.00	0.00	0.00	6.82	103
Modification total	***28.07***	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*16.37*	*252*
Total	***53.33***	*35.03*	*34.27*	*33.52*	*31.22*	*28.97*	*27.88*	*2705*
Addition total	***59.33***	*40.27*	*39.05*	*38.65*	*36.03*	*32.75*	*31.50*	*2038*
Removal total	***44.29***	*22.41*	*15.73*	*22.76*	*14.41*	*23.53*	*17.48*	*132*

Table 4: Primary evaluation F-scores by event type. The "size" column gives the number of annotations of each type in the given data (training+development). Best result for each type shown in bold. For DEHYDROXYLATION, no examples were present in the test data and none were predicted by any participant.

Team	recall	prec.	F-score
UTurku	52.69	53.98	53.33
FAUST	28.88	44.51	35.03
MSR-NLP	27.79	44.69	34.27
UMass	28.08	41.55	33.52
Stanford	26.56	37.85	31.22
CCP-BTMG	23.44	37.93	28.97
ConcordU	20.83	42.14	27.88

Table 5: Primary evaluation results

only the core task. For the full task, this difference in approach is reflected in the substantial performance advantage for the UTurku system, which exhibits highest performance overall as well as for most individual event types.

Extraction performance for simple events taking only *Theme* and *Site* arguments is consistently higher than for other event types, with absolute F-score differences of over 10% points for many systems. Similar notable performance differences are seen between the *addition* events, for which ample training data was available, and the *removal* types for which data was limited. This effect is particularly noticeable for DEPHOSPHORYLATION, DNA DEMETHYLATION and DEMETHYLATION, for which the clear majority of systems failed to predict any correct events. Extraction performance for CATALYSIS events is very low despite a relatively large set of training examples, indicating that the extraction of nested event structures remains very challenging. This low performance may also be related to the fact that CATALYSIS events are often triggered by the same word as the catalysed modification (e.g. Figure 1b), requiring the assignment of multiple event labels to a single word in typical system architectures.

Table 6 summarizes the full task results with the addition of the single partial penalty criterion. The F-scores for the seven participants under this crite-

Team	recall	prec.	F-score	Δ
UTurku	54.79	58.42	56.55	3.22
FAUST	28.88	72.05	41.24	6.21
MSR-NLP	27.79	66.72	39.24	4.97
UMass	28.08	63.28	38.90	5.38
Stanford	26.56	56.83	36.20	4.98
CCP-BTMG	23.44	50.79	32.08	3.11
ConcordU	20.83	60.55	30.99	3.11

Table 6: Full task evaluation results for primary criteria and with single partial penalty. The Δ column gives F-score difference to the primary results.

Team	recall	prec.	F-score	Δ_1	Δ_2
UTurku	68.51	69.20	68.86	15.53	12.31
FAUST	59.88	80.25	68.59	33.56	27.35
MSR-NLP	55.70	77.60	64.85	30.58	25.61
UMass	57.04	73.30	64.15	30.63	25.25
Stanford	56.87	70.22	62.84	31.62	26.64
ConcordU	40.28	76.71	52.83	24.95	21.84
CCP-BTMG	45.06	63.37	52.67	23.70	20.59

Table 7: Core task evaluation results. The Δ_1 column gives F-score difference to primary full task results, Δ_2 to full task results with single partial penalty.

rion are on average over 4% points higher than under the primary criteria, with the most substantial increases seen for high-ranking participants only addressing the core task: for example, the precision of the FAUST system (Riedel et al., 2011) is nearly 30% higher under the relaxed criterion. These results provide new perspective deserving further detailed study into the question of what are the most meaningful criteria for event extraction system evaluation.

Table 7 summarizes the core task results. While all systems show notably higher performance than for the full task, high-ranking participants focusing on the core task gain most dramatically, with the FAUST system core task F-score essentially matching that of the top system (UTurku). For the core task, all participants achieve F-scores over 50% – a result achieved by only a single system in the '09 task – and the top four participants average over 65% F-score. These results confirm that current event extraction technology is well applicable to the core PTM extraction task even when the number of targeted event types is relatively high and may be ready to address the challenges of exhaustive PTM extraction (Pyysalo et al., 2011a). The best core tasks results, approaching 70% F-score, are particularly encouraging as the level of performance is comparable to or better than state-of-the-art results for many reference resources for protein-protein interaction extraction (see e.g. Tikk et al. (2010))) using the simple untyped entity pair representation, a standard task that has been extensively studied in the domain.

6 Discussion and Conclusions

This paper has presented the preparation, resources, results and analysis of the BioNLP Shared Task 2011 Epigenetics and Post-translational modifications (EPI) main task. The results demonstrate that the core extraction target of identifying statements of 14 different modification types with the modified gene or gene product can be reliably addressed by current event extraction methods, with two systems approaching 70% F-score at this task. Nevertheless, challenges remain in detecting statements regarding the catalysis of these events as well as in resolving the full detail of such modification events, a task attempted by only one participant and at which performance remains at somewhat above 50% in F-score.

Detailed evaluation showed that the highly competitive participating systems differ substantially in their relative strengths, indicating potential for further development at protein and DNA modification event detection. The task results are available in full detail from the shared task webpage, `http://sites.google.com/site/bionlpst/`.

In the future, we will follow the example of the BioNLP'09 shared task in making the data and resources of the EPI task open to all interested parties to encourage further study of event extraction for epigenetics and post-translational modification events, to facilitate system comparison on a well-defined standard task, and to support the development of further applications of event extraction technology in this important area of biomolecular science.

Acknowledgments

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

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.

D. Barford, A.K. Das, and M.P. Egloff. 1998. The structure and mechanism of protein phosphatases: insights into catalysis and regulation. *Annual review of biophysics and biomolecular structure*, 27(1):133–164.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Ekaterina Buyko, Erik Faessler, Joachim Wermter, and Udo Hahn. 2009. Event extraction from trimmed dependency graphs. In *Proceedings of BioNLP Shared Task 2009*, pages 19–27.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of ACL'05*, pages 173–180.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 449–454.

M.H. Glickman and A. Ciechanover. 2002. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiological reviews*, 82(2):373.

R. Hoehndorf, A.C.N. Ngomo, S. Pyysalo, T. Ohta, A. Oellrich, and D. Rebholz-Schuhmann. 2010. Applying ontology design patterns to the implementation of relations in GENIA. In *Proceedings of the Fourth Symposium on Semantic Mining in Biomedicine SMBM 2010*.

Robin Holliday and JE Pugh. 1975. Dna modification mechanisms and gene activity during development. *Science*, 187:226–232.

Robin Holliday. 1987. The inheritance of epigenetic defects. *Science*, 238:163–170.

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

Halil Kilicoglu and Sabine Bergler. 2011. Adapting a general semantic interpretation approach to biological event extraction. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(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. 2011a. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011b. Overview of the Genia Event task in BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

R. Leaman and G. Gonzalez. 2008. Banner: an executable survey of advances in biomedical named entity recognition. *Proceedings of the Pacific Symposium on Biocomputing (PSB'08)*, pages 652–663.

Haibin Liu, Ravikumar Komandur, and Karin Verspoor. 2011. From graphs to events: A subgraph matching approach for information extraction from biomedical text. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event extraction as dependency parsing for bionlp 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

David McClosky. 2009. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Department of Computer Science, Brown University.

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. 2010a. Event extraction for post-translational modifications. In *Proceedings of BioNLP'10*, pages 19–27.

Tomoko Ohta, Sampo Pyysalo, Makoto Miwa, and Jun'ichi Tsujii. 2010b. Event extraction for dna methylation. In *Proceedings of SMBM'10*.

Maté Ongenaert, Leander Van Neste, Tim De Meyer, Gerben Menschaert, Sofie Bekaert, and Wim

Van Criekinge. 2008. PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucl. Acids Res.*, 36(suppl_1):D842–846.

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, Makoto Miwa, and Jun'ichi Tsujii. 2011a. Towards exhaustive protein modification event extraction. In *Proceedings of the BioNLP 2011 Workshop*, Portland, Oregon, June. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2011b. Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Chris Quirk, Pallavi Choudhury, Michael Gamon, and Lucy Vanderwende. 2011. Msr-nlp entry in bionlp shared task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Sebastian Riedel and Andrew McCallum. 2011. Robust biomedical event extraction with dual decomposition and minimal domain adaptation. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, and Chris Manning. 2011. Model combination for event extraction in bionlp 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

A.D. Riggs. 1975. X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14:9–25.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. BioNLP Shared Task 2011: Supporting Resources. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

JB Stock, AJ Ninfa, and AM Stock. 1989. Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiology and Molecular Biology Reviews*, 53(4):450.

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.

Paul Thompson, Syed Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1):349.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837, 07.

Yue Wang, Jin-Dong Kim, Rune Sætre, Sampo Pyysalo, and Jun'ichi Tsujii. 2009. Investigating heterogeneous protein annotations toward cross-corpora utilization. *BMC Bioinformatics*, 10(403), Dec. ISSN: 1471-2105.

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.