Overview of BioNLP'09 Shared Task on Event Extraction

Jin-Dong Kim[*] **Tomoko Ohta**[*] **Sampo Pyysalo**[*] **Yoshinobu Kano**[*] **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
{jdkim,okap,smp,kano,tsujii}@is.s.u-tokyo.ac.jp

Abstract

The paper presents the design and implementation of the BioNLP'09 Shared Task, and reports the final results with analysis. The shared task consists of three sub-tasks, each of which addresses bio-molecular event extraction at a different level of specificity. The data was developed based on the GENIA event corpus. The shared task was run over 12 weeks, drawing initial interest from 42 teams. Of these teams, 24 submitted final results. The evaluation results are encouraging, indicating that state-of-the-art performance is approaching a practically applicable level and revealing some remaining challenges.

1 Introduction

The history of text mining (*TM*) shows that shared tasks based on carefully curated resources, such as those organized in the MUC (Chinchor, 1998), TREC (Voorhees, 2007) and ACE (Strassel et al., 2008) events, have significantly contributed to the progress of their respective fields. This has also been the case in *bio-TM*. Examples include the TREC Genomics track (Hersh et al., 2007), JNLPBA (Kim et al., 2004), LLL (Nédellec, 2005), and BioCreative (Hirschman et al., 2007). While the first two addressed *bio-IR* (information retrieval) and *bio-NER* (named entity recognition), respectively, the last two focused on *bio-IE* (information extraction), seeking relations between bio-molecules. With the emergence of NER systems with performance capable of supporting practical applications, the recent interest of the bio-TM community is shifting toward IE.

Similarly to LLL and BioCreative, the BioNLP'09 Shared Task (the BioNLP task, hereafter) also addresses bio-IE, but takes a definitive step further toward finer-grained IE. While LLL and BioCreative focus on a rather simple representation of relations of bio-molecules, i.e. protein-protein interactions (PPI), the BioNLP task concerns the detailed behavior of bio-molecules, characterized as bio-molecular events (*bio-events*). The difference in focus is motivated in part by different applications envisioned as being supported by the IE methods. For example, BioCreative aims to support curation of PPI databases such as MINT (Chatr-aryamontri et al., 2007), for a long time one of the primary tasks of bioinformatics. The BioNLP task aims to support the development of more detailed and structured databases, e.g. pathway (Bader et al., 2006) or Gene Ontology Annotation (GOA) (Camon et al., 2004) databases, which are gaining increasing interest in bioinformatics research in response to recent advances in molecular biology.

As the first shared task of its type, the BioNLP task aimed to define a bounded, well-defined bio-event extraction task, considering both the actual needs and the state of the art in *bio-TM* technology and to pursue it as a community-wide effort. The key challenge was in finding a good balance between the utility and the feasibility of the task, which was also limited by the resources available. Special consideration was given to providing evaluation at diverse levels and aspects, so that the results can drive continuous efforts in relevant directions. The paper discusses the design and implementation of the BioNLP task, and reports the results with analysis.

Proceedings of the Workshop on BioNLP: Shared Task, pages 1–9,
Boulder, Colorado, June 2009.

Type	Primary Args.	Second. Args.
Gene_expression	T(P)	
Transcription	T(P)	
Protein_catabolism	T(P)	
Phosphorylation	T(P)	Site
Localization	T(P)	AtLoc, ToLoc
Binding	T(P)+	Site+
Regulation	T(P/Ev), C(P/Ev)	Site, CSite
Positive_regulation	T(P/Ev), C(P/Ev)	Site, CSite
Negative_regulation	T(P/Ev), C(P/Ev)	Site, CSite

Table 1: Event types and their arguments. The type of the filler entity is specified in parenthesis. The filler entity of the secondary arguments are all of *Entity* type which represents any entity but proteins: T=Theme, C=Cause, P=Protein, Ev=Event.

The failure of p65 translocation to the nucleus ...	
T3	(Protein, 40-46)
T2	**(Localization, 19-32)**
E1	**(Type:T2, Theme:T3**, ToLoc:T1**)**
T1	(Entity, 15-18)
M1	(Negation E1)

Figure 1: Example event annotation. The protein annotation T3 is given as a starting point. The extraction of annotation in bold is required for Task 1, T1 and the ToLoc:T1 argument for Task 2, and M1 for Task 3.

2 Task setting

To focus efforts on the novel aspects of the event extraction task, is was assumed that named entity recognition has already been performed and the task was begun with a given set of gold protein annotation. This is the only feature of the task setting that notably detracts from its realism. However, given that state-of-the-art protein annotation methods show a practically applicable level of performance, i.e. 88% F-score (Wilbur et al., 2007), we believe the choice is reasonable and has several advantages, including focus on event extraction and effective evaluation and analysis.

2.1 Target event types

Table 1 shows the event types addressed in the BioNLP task. The event types were selected from the GENIA ontology, with consideration given to their importance and the number of annotated instances in the GENIA corpus. The selected event types all concern protein biology, implying that they take proteins as their theme. The first three types concern protein metabolism, i.e. protein production and breakdown. Phosphorylation is a representative protein modification event, and Localization and Binding are representative fundamental molecular events. Regulation (including its sub-types, Positive and Negative_regulation) represents regulatory events and causal relations. The last five are universal but frequently occur on proteins. For the biological interpretation of the event types, readers are referred to Gene Ontology (GO) and the GENIA ontology.

As shown in Table 1, the theme or themes of all events are considered primary arguments, that is, arguments that are critical to identifying the event. For regulation events, the entity or event stated as the *cause* of the regulation is also regarded as a primary argument. For some event types, other arguments detailing of the events are also defined (*Secondary Args.* in Table 1).

From a computational point of view, the event types represent different levels of complexity. When only primary arguments are considered, the first five event types require only unary arguments, and the task can be cast as relation extraction between a predicate (event trigger) and an argument (Protein). The Binding type is more complex in requiring the detection of an arbitrary number of arguments. Regulation events always take a Theme argument and, when expressed, also a Cause argument. Note that a Regulation event may take another event as its theme or cause, a unique feature of the BioNLP task compared to other event extraction tasks, e.g. ACE.

2.2 Representation

In the BioNLP task, events are expressed using three different types of entities. *Text-bound entities* (*t-entities* hereafter) are represented as text spans with associated class information. The t-entities include event triggers (*Localization*, *Binding*, etc), protein references (*Protein*) and references to other entities (*Entity*). A t-entity is represented by a pair, (*entity-type*, *text-span*), and assigned an id with the prefix "T", e.g. T1–T3 in Figure 1. An *event* is expressed as an n-tuple of typed t-entities, and has a id with prefix "E", e.g. E1. An *event modification* is expressed by a pair, (*predicate-negation-or-speculation*, *event-id*), and has an id with prefix "M", e.g. M1.

Item	Training	Devel.	Test
Abstract	800	150	260
Sentence	7,449	1,450	2,447
Word	176,146	33,937	57,367
Event	8,597 / 8,615	1,809 / 1,815	3,182 / 3,193

Table 2: Statistics of the data sets. For events, Task1/Task2 shown separately as secondary arguments may introduce additional differentiation of events.

2.3 Subtasks

The BioNLP task targets semantically rich event extraction, involving the extraction of several different classes of information. To facilitate evaluation on different aspects of the overall task, the task is divided to three sub-tasks addressing event extraction at different levels of specificity.

Task 1. Core event detection detection of typed, text-bound events and assignment of given proteins as their primary arguments.

Task 2. Event enrichment recognition of secondary arguments that further specify the events extracted in Task 1.

Task 3. Negation/Speculation detection detection of negations and speculation statements concerning extracted events.

Task 1 serves as the backbone of the shared task and is mandatory for all participants. Task 2 involves the recognition of *Entity* type t-entities and assignment of those as secondary event arguments. Task 3 addresses the recognition of negated or speculatively expressed events without specific binding to text. An example is given in Fig. 1.

3 Data preparation

The BioNLP task data were prepared based on the GENIA event corpus. The data for the training and development sets were derived from the publicly available event corpus (Kim et al., 2008), and the data for the test set from an unpublished portion of the corpus. Table 2 shows statistics of the data sets.

For data preparation, in addition to filtering out irrelevant annotations from the original GENIA corpus, some new types of annotation were added to make the event annotation more appropriate for the purposes of the shared task. The following sections describe the key changes to the corpus.

3.1 Gene-or-gene-product annotation

The named entity (NE) annotation of the GENIA corpus has been somewhat controversial due to differences in annotation principles compared to other biomedical NE corpora. For instance, the NE annotation in the widely applied GENETAG corpus (Tanabe et al., 2005) does not differentiate proteins from genes, while GENIA annotation does. Such differences have caused significant inconsistency in methods and resources following different annotation schemes. To remove or reduce the inconsistency, GENETAG-style NE annotation, which we term gene-or-gene-product (GGP) annotation, has been added to the GENIA corpus, with appropriate revision of the original annotation. For details, we refer to (Ohta et al., 2009). The NE annotation used in the BioNLP task data is based on this annotation.

3.2 Argument revision

The GENIA event annotation was made based on the GENIA event ontology, which uses a loose typing system for the arguments of each event class. For example, in Figure 2(a), it is expressed that the binding event involves two proteins, TRAF2 and CD40, and that, in the case of CD40, its cytoplasmic domain takes part in the binding. Without constraints on the type of theme arguments, the following two annotations are both legitimate:

(Type:Binding, Theme:*TRAF2*, Theme:*CD40*)
(Type:Binding, Theme:*TRAF2*,
Theme:*CD40 cytoplasmic domain*)

The two can be seen as specifying the same event at different levels of specificity[1]. Although both alternatives are reasonable, the need to have consistent training and evaluation data requires a consistent choice to be made for the shared task.

Thus, we fix the types of all non-event primary arguments to be proteins (specifically GGPs). For GENIA event annotations involving themes other than proteins, additional argument types were introduced, for example, as follows:

[1] In the GENIA event annotation guidelines, annotators are instructed to choose the more specific alternative, thus the second alternative for the example case in Fig. 2(a).

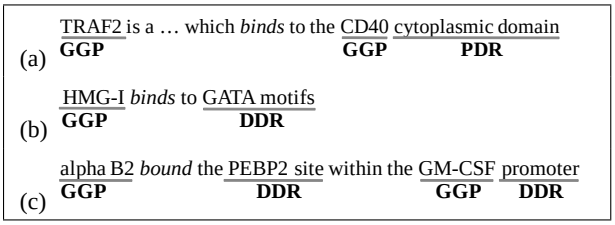

Figure 2: Entity annotation to example sentences from (a) PMID10080948, (b) PMID7575565, and (c) PMID7605990 (simplified).

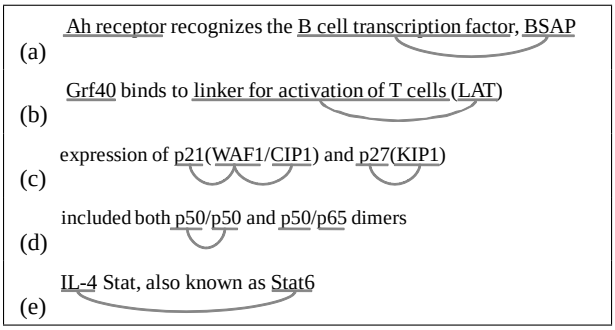

Figure 3: Equivalent entities in example sentences from (a) PMID7541987 (simplified), (b) PMID10224278, (c) PMID10090931, (d) PMID9243743, (e) PMID7635985.

(Type:Binding, Theme1:*TRAF2*, Theme2:*CD40*,
Site2:*cytoplasmic domain*)

Note that the protein, CD40, and its domain, cytoplasmic domain, are associated by argument numbering. To resolve issues related to the mapping between proteins and related entities systematically, we introduced partial static relation annotation for relations such as Part-Whole, drawing in part on similar annotation of the BioInfer corpus (Pyysalo et al., 2007). For details of this part of the revision process, we refer to (Pyysalo et al., 2009).

Figure 2 shows some challenging cases. In (b), the site *GATA motifs* is not identified as an argument of the binding event, because the protein containing it is not stated. In (c), among the two sites (*PEBP2 site* and *promoter*) of the gene *GM-CSF*, only the more specific one, *PEBP2*, is annotated.

3.3 Equivalent entity references

Alternative names for the same object are frequently introduced in biomedical texts, typically through apposition. This is illustrated in Figure 3(a), where the two expressions *B cell transcription factor* and *BSAP* are in apposition and refer to the same protein. Consequently, in this case the following two annotations represent the same event:

(Type:Binding, Theme:*Ah receptor*,
Theme:*B cell transcription factor*)
(Type:Binding, Theme:*Ah receptor*, Theme:*BSAP*)

In the GENIA event corpus only one of these is annotated, with preference given to shorter names over longer descriptive ones. Thus of the above example events, the latter would be annotated. However, as both express the same event, in the shared task evaluation either alternative was accepted as correct extraction of the event. In order to implement this aspect of the evaluation, expressions of equivalent entities were annotated as follows:

Eq (*B cell transcription factor*, *BSAP*)

The equivalent entity annotation in the revised GENIA corpus covers also cases other than simple apposition, illustrated in Figure 3. A frequent case in biomedical literature involves use of the slash symbol ("/") to state synonyms. The slash symbol is ambiguous as it is used also to indicate dimerized proteins. In the case of *p50/p50*, the two *p50* are annotated as equivalent because they represent the same proteins at the same state. Note that although rare, also explicitly introduced aliases are annotated, as in Figure 3(e).

4 Evaluation

For the evaluation, the participants were given the test data with gold annotation only for proteins. The evaluation was then carried out by comparing the annotation predicted by each participant to the gold annotation. For the comparison, equality of annotations is defined as described in Section 4.1. The evaluation results are reported using the standard recall/precision/f-score metrics, under different criteria defined through the equalities.

4.1 Equalities and Strict matching

Equality of events is defined as follows:

Event Equality equality holds between any two events when (1) the event types are the same, (2) the event triggers are the same, and (3) the arguments are fully matched.

A full matching of arguments between two events means there is a perfect 1-to-1 mapping between the two sets of arguments. Equality of individual arguments is defined as follows:

Argument Equality equality holds between any two arguments when (1) the role types are the same, and (2-1) both are t-entities and equality holds between them, or (2-2) both are events and equality holds between them.

Due to the condition (2-2), event equality is defined recursively for events referring to events. Equality of t-entities is defined as follows:

T-entity Equality equality holds between any two t-entities when (1) the entity types are the same, and (2) the spans are the same.

Any two text spans $(beg1, end1)$ and $(beg2, end2)$, are the same iff $beg1 = beg2$ and $end1 = end2$. Note that the event triggers are also t-entities thus their equality is defined by the t-entity equality.

4.2 Evaluation modes

Various evaluation modes can be defined by varying equivalence criteria. In the following, we describe three fundamental variants applied in the evaluation.

Strict matching The *strict matching* mode requires exact equality, as defined in section 4.1. As some of its requirements may be viewed as unnecessarily precise, practically motivated relaxed variants, described in the following, are also applied.

Approximate span matching The *approximate span matching* mode is defined by relaxing the requirement for text span matching for t-entities. Specifically, a given span is equivalent to a gold span if it is entirely contained within an extension of the gold span by one word both to the left and to the right, that is, $beg1 \geq ebeg2$ and $end1 \leq eend2$, where $(beg1, end1)$ is the given span and $(ebeg2, eend2)$ is the extended gold span.

Approximate recursive matching In strict matching, for a regulation event to be correct, the events it refers to as theme or cause must also be be strictly correct. The *approximate recursive matching* mode is defined by relaxing the requirement for recursive event matching, so that an event can match even if the events it refers to are only partially correct. Specifically, for partial matching, only Theme arguments are considered: events can match even if referred events differ in non-Theme arguments.

Event	Release date
Announcement	Dec 8
Sample data	Dec 15
Training data	Jan 19 → 21, Feb 2 (rev1), Feb 10 (rev2)
Devel. data	Feb 7
Test data	Feb 22 → Mar 2
Submission	Mar 2 → Mar 9

Table 3: Shared task schedule. The arrows indicate a change of schedule.

5 Schedule

The BioNLP task was held for 12 weeks, from the sample data release to the final submission. It included 5 weeks of *system design period* with sample data, 6 weeks of *system development period* with training and development data, and a 1 week *test period*. The system development period was originally planned for 5 weeks but extended by 1 week due to the delay of the training data release and the revision. Table 3 shows key dates of the schedule.

6 Supporting Resources

To allow participants to focus development efforts on novel aspects of event extraction, we prepared publicly available BioNLP resources readily available for the shared task. Several fundamental BioNLP tools were provided through U-Compare (Kano et al., 2009)[2], which included tools for tokenization, sentence segmentation, part-of-speech tagging, chunking and syntactic parsing.

Participants were also provided with the syntactic analyses created by a selection of parsers. We applied two mainstream Penn Treebank (PTB) phrase structure parsers: the Bikel parser[3], implementing Collins' parsing model (Bikel, 2004) and trained on PTB, and the reranking parser of (Charniak and Johnson, 2005) with the self-trained biomedical parsing model of (McClosky and Charniak, 2008)[4]. We also applied the GDep[5], native dependency parser trained on the GENIA Treebank

[2] http://u-compare.org/

[3] http://www.cis.upenn.edu/∼dbikel/software.html

[4] http://www.cs.brown.edu/∼dmcc/biomedical.html

[5] http://www.cs.cmu.edu/∼sagae/parser/gdep/

Team	Task	Org	NLP			Task		Ext. Resources
			Word	Chunking	Parsing	Trigger	Argument	
UTurku	1--	3C+2BI	Porter		MC	SVM	SVM (SVMlight)	
JULIELab	1--	1C+2L+2B	OpenNLP Porter	OpenNLP	GDep	Dict+Stat	SVM(libSVM) ME(Mallet)	UniProt, Mesh, GOA, UMLS
ConcordU	1-3	3C	Stanford		Stanford	Dict+Stat	Rules	WordNet, VerbNet, UMLS
UT+DBCLS	12-	2C	Porter		MC CCG	Dict	MLN(thebeast)	
VIBGhent	1-3	2C+1B	Porter,		Stanford	Dict	SVM(libSVM)	
UTokyo	1--	3C	GTag		GDep, Enju	Dict	ME(liblinear)	UIMA
UNSW	1--	1C+1B			GDep	CRF	Rules	WordNet, MetaMap
UZurich	1--	3C	LingPipe, Morpha	LTChunk	Pro3Gres	Dict	Rules	
ASU+HU+BU	123	6C+2BI	Porter		BioLG, Charniak	Dict	Rules Rules	Lucene
Cam	1--	3C	Porter		RASP	Dict	Rules	
UAntwerp	12-	3C	GTag		GDep	MBL	MBL(TiMBL) Rules	
UNIMAN	1--	4C+2BI	Porter GTag		GDep	Dict, CRF	SVM Rules	MeSH, GO
SCAI	1--	1C					Rules	
UAveiro	1--	1C+1L	NooJ	NooJ			Rules	BioLexicon
USzeged	1-3	3C+1B	GTag			Dict, VSM	C4.5(WEKA) Rules	BioScope
NICTA	1-3	4C	GTag		ERG	CRF(CRF++)	Rules	JULIE
CNBMadrid	12-	2C+1B	Porter, GTag	GTag			CBR Rules	
CCP-BTMG	123	7C	LingPipe	LingPipe	OpenDMAP	LingPipe, CM	Rules	GO, SO, MIO, UIMA
CIPS-ASU	1--	3C	MontyTagger	Custom	Stanford	CRF(ABNER)	Rules, NB(WEKA)	
UMich	1--	2C	Stanford		MC	Dict	SVM(SVMlight)	
PIKB	1--	5C+2B				MIRA	MIRA	
KoreaU	1--	5C	GTag		GDep	Rules, ME	ME	WSJ

Table 4: Profiles of the participants: GTag=GENIAtagger, MLN=Markov Logic Network, UMLS=UMLS SPECIALIST Lexicon/tools, MC=McClosky-Charniak, GDep=Genia Dependency Parser, Stanford=Stanford Parser, CBR=Case-Based Reasoning, CM=ConceptMapper.

(Tateisi et al., 2005), and a version of the C&C CCG deep parser[6] adapted to biomedical text (Rimell and Clark, 2008).

The text of all documents was segmented and tokenized using the GENIA Sentence Splitter and the GENIA Tagger, provided by U-Compare. The same segmentation was enforced for all parsers, which were run using default settings. Both the native output of each parser and a representation in the popular Stanford Dependency (SD) format (de Marneffe et al., 2006) were provided. The SD representation was created using the Stanford tools[7] to convert from the PTB scheme, the custom conversion introduced by (Rimell and Clark, 2008) for the C&C CCG parser, and a simple format-only conversion for GDep.

7 Results and Discussion

7.1 Participation

In total, 42 teams showed interest in the shared task and registered for participation, and 24 teams submitted final results. All 24 teams participated in the obligatory Task 1, six in each of Tasks 2 and 3, and two teams completed all the three tasks.

Table 4 shows a profile of the 22 final teams, excepting two who wished to remain anonymous. A brief examination on the team organization (the *Org* column) shows a computer science background (C) to be most frequent among participants, with less frequent participation from bioinformaticians (BI), biologists (B) and liguists (L). This may be attributed in part to the fact that the event extraction task required complex computational modeling. The role of computer scientists may be emphasized in part due to the fact that the task was novel to most participants, requiring particular efforts in framework design and implementation and computational resources. This also suggests there is room for improvement from more input from biologists.

7.2 Evaluation results

The final evaluation results of Task 1 are shown in Table 5. The results on the five event types involv-

[6] http://svn.ask.it.usyd.edu.au/trac/candc/wiki

[7] http://nlp.stanford.edu/software/lex-parser.shtml

Team	Simple Event	Binding	Regulation	All
UTurku	**64.21 / 77.45 / 70.21**	**40.06 / 49.82 / 44.41**	**35.63 / 45.87 / 40.11**	46.73 / 58.48 / 51.95
JULIELab	**59.81 / 79.80 / 68.38**	**49.57 / 35.25 / 41.20**	**35.03 / 34.18 / 34.60**	45.82 / 47.52 / 46.66
ConcordU	**49.75 / 81.44 / 61.76**	20.46 / 40.57 / 27.20	**27.47 / 49.89 / 35.43**	34.98 / 61.59 / 44.62
UT+DBCLS	**55.75 / 72.74 / 63.12**	23.05 / 48.19 / 31.19	**26.32 / 41.81 / 32.30**	36.90 / 55.59 / 44.35
VIBGhent	**54.48 / 79.31 / 64.59**	**38.04 / 38.60 / 38.32**	17.36 / 31.61 / 22.41	33.41 / 51.55 / 40.54
UTokyo	45.69 / 72.19 / 55.96	**34.58 / 50.63 / 41.10**	14.22 / 34.26 / 20.09	28.13 / 53.56 / 36.88
UNSW	45.85 / 69.94 / 55.39	23.63 / 37.27 / 28.92	16.58 / 28.27 / 20.90	28.22 / 45.78 / 34.92
UZurich	44.92 / 66.62 / 53.66	30.84 / 37.28 / 33.75	14.82 / 30.21 / 19.89	27.75 / 46.60 / 34.78
ASU+HU+BU	45.09 / 76.80 / 56.82	19.88 / 44.52 / 27.49	05.20 / 33.46 / 09.01	21.62 / 62.21 / 32.09
Cam	39.17 / 76.40 / 51.79	12.68 / 31.88 / 18.14	09.98 / 37.76 / 15.79	21.12 / 56.90 / 30.80
UAntwerp	41.29 / 65.68 / 50.70	12.97 / 31.03 / 18.29	11.07 / 29.85 / 16.15	22.50 / 47.70 / 30.58
UNIMAN	50.00 / 63.21 / 55.83	12.68 / 40.37 / 19.30	04.05 / 16.75 / 06.53	22.06 / 48.61 / 30.35
SCAI	43.74 / 70.73 / 54.05	28.82 / 35.21 / 31.70	12.64 / 16.55 / 14.33	25.96 / 36.26 / 30.26
UAveiro	43.57 / 71.63 / 54.18	13.54 / 34.06 / 19.38	06.29 / 21.05 / 09.69	20.93 / 49.30 / 29.38
Team 24	41.29 / 64.72 / 50.41	22.77 / 35.43 / 27.72	09.38 / 19.23 / 12.61	22.69 / 40.55 / 29.10
USzeged	47.63 / 44.44 / 45.98	15.27 / 25.73 / 19.17	04.17 / 18.21 / 06.79	21.53 / 36.99 / 27.21
NICTA	31.13 / 77.31 / 44.39	16.71 / 29.00 / 21.21	07.80 / 18.12 / 10.91	17.44 / 39.99 / 24.29
CNBMadrid	50.25 / 46.59 / 48.35	33.14 / 20.54 / 25.36	12.22 / 07.99 / 09.67	28.63 / 20.88 / 24.15
CCP-BTMG	28.17 / 87.63 / 42.64	12.68 / 40.00 / 19.26	03.09 / 48.11 / 05.80	13.45 / 71.81 / 22.66
CIPS-ASU	39.68 / 38.60 / 39.13	17.29 / 31.58 / 22.35	11.86 / 08.15 / 09.66	22.78 / 19.03 / 20.74
UMich	52.71 / 25.89 / 34.73	31.70 / 12.61 / 18.05	14.22 / 06.56 / 08.98	30.42 / 14.11 / 19.28
PIKB	26.65 / 75.72 / 39.42	07.20 / 39.68 / 12.20	01.09 / 30.51 / 02.10	11.25 / 66.54 / 19.25
Team 09	27.16 / 43.61 / 33.47	03.17 / 09.82 / 04.79	02.42 / 11.90 / 04.02	11.69 / 31.42 / 17.04
KoreaU	20.56 / 66.39 / 31.40	12.97 / 50.00 / 20.59	00.67 / 37.93 / 01.31	09.40 / 61.65 / 16.31

Table 5: Evaluation results of Task 1 (recall / precision / f-score).

Team	All	Site for Phospho.(56)	AtLoc & ToLoc (65)	All Second Args.
UT+DBCLS	**35.86 / 54.08 / 43.12**	**71.43 / 71.43 / 71.43**	**23.08 / 88.24 / 36.59**	**32.14 / 72.41 / 44.52**
UAntwerp	21.52 / 45.77 / 29.27	00.00 / 00.00 / 00.00	01.54 /100.00 / 03.03	06.63 / 52.00 / 11.76
ASU+HU+BU	19.70 / 56.87 / 29.26	00.00 / 00.00 / 00.00	00.00 / 00.00 / 00.00	00.00 / 00.00 / 00.00
Team 24	22.08 / 38.28 / 28.01	**55.36 / 93.94 / 69.66**	**21.54 / 66.67 / 32.56**	**30.10 / 76.62 / 43.22**
CCP-BTMG	13.25 / 70.97 / 22.33	30.36 /**100.00** / 46.58	00.00 / 00.00 / 00.00	08.67 /100.00 / 15.96
CNBMadrid	25.02 / 18.32 / 21.15	**85.71 / 57.14 / 68.57**	**32.31 / 47.73 / 38.53**	50.00 / 09.71 / 16.27

Table 6: Evaluation results for Task 2.

ing only a single primary theme argument are shown in one merged class, "Simple Event". The broad performance range (31% – 70%) indicates even the extraction of simple events is not a trivial task. However, the top-ranked systems show encouraging performance, achieving or approaching 70% f-score.

The performance ranges for Binding (5% – 44%) and Regulation (1% – 40%) events show their extraction to be clearly more challenging. It is interesting that while most systems show better performance for binding over regulation events, the systems [ConcordU] and [UT+DBCLS] are better for regulation, showing somewhat reduced performance for Binding events. This is in particular contrast to the following two systems, [ViBGhent] and [UTokyo], which show far better performance for Binding than Regulation events. As one possible explanation, we find that the latter two differentiate binding events by their number of themes, while the former two give no specific treatment to multi-theme binding events. Such observations and comparisons are a clear benefit of a community-wide shared task.

Table 6 shows the evaluation results for the teams who participated in Task 2. The "All" column shows the overall performance of the systems for Task 2, while the "All Second Args." column shows the performance of finding only the secondary arguments. The evaluation results show considerable differences between the criteria. For example, the system [Team 24] shows performance comparable to the top ranked system in finding secondary arguments, although its overall performance for Task 2 is more limited. Table 6 also shows the three systems, [UT+DBCLS], [Team 24] and [CNBMadrid],

Team	Negation	Speculation
ConcordU	**14.98 / 50.75 / 23.13**	**16.83 / 50.72 / 25.27**
VIBGhent	**10.57 / 45.10 / 17.13**	08.65 / 15.79 / 11.18
ASU+HU+BU	03.96 / 27.27 / 06.92	06.25 / 28.26 / 10.24
NICTA	05.29 / 34.48 / 09.17	04.81 / 30.30 / 08.30
USzeged	05.29 / 01.94 / 02.84	12.02 / 03.88 / 05.87
CCP-BTMG	01.76 / 05.26 / 02.64	06.73 / 13.33 / 08.95

Table 7: Evaluation results for Task 3.

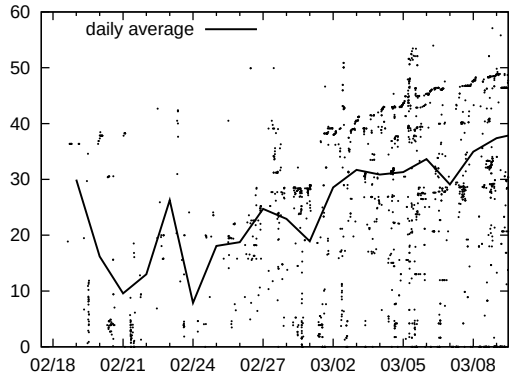

Figure 4: Scatterplot of the evaluation results on the development data during the system development period.

show performance at a practical level in particular in finding specific sites of phosphorylation.

As shown in Table 7, the performance range for Task 3 is very low although the representation of the task is as simple as the simple events. We attribute the reason to the fact that Task 3 is the only task of which the annotation is not bound to textual clue, thus no text-bound annotation was provided.

Figure 4 shows a scatter plot of the performance of the participating systems during the system development period. The performance evaluation comes from the log of the online evaluation system on the development data. It shows the best performance and the average performance of the participating systems were trending upwards up until the deadline of final submission, which indicates there is still much potential for improvement.

7.3 Ensemble

Table 8 shows experimental results of a system ensemble using the final submissions. For the experiments, the top 3–10 systems were chosen, and the output of each system treated as a weighted vote[8]. Three weighting schemes were used; "Equal" weights each vote equally; "Averaged" weights each vote by the overall f-score of the system; "Event Type" weights each vote by the f-score of the system for the specific event type. The best score, 55.96%, was obtained by the "Event Type" weighting scheme, showing a 4% unit improvement over the best individual system. While using the final scores for weighting uses data that would not be available in practice, similar weighting could likely be obtained e.g. using performance on the development data. The experiment demonstrates that an f-score better than 55% can be achieved simply by combining the strengths of the systems.

[8]We used the 'ensemble' function of U-Compare.

Ensemble	Equal	Averaged	Event Type
Top 3	53.19	53.19	54.08
Top 4	54.34	54.34	55.21
Top 5	54.77	55.03	55.10
Top 6	**55.13**	**55.77**	**55.96**
Top 7	54.33	55.45	55.73
Top 10	52.79	54.63	55.18

Table 8: Experimental results of system ensemble.

8 Conclusion

Meeting with the community-wide participation, the BioNLP Shared Task was successful in introducing fine-grained event extraction to the domain. The evaluation results of the final submissions from the participants are both promising and encouraging for the future of this approach to IE. It has been revealed that state-of-the-art performance in event extraction is approaching a practically applicable level for simple events, and also that there are many remaining challenges in the extraction of complex events. A brief analysis suggests that the submitted data together with the system descriptions are rich resources for finding directions for improvements. Finally, the experience of the shared task participants provides an invaluable basis for cooperation in facing further challenges.

Acknowledgments

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Genome Network Project (MEXT, Japan).

References

Gary D. Bader, Michael P. Cary, and Chris Sander. 2006. Pathguide: a Pathway Resource List. *Nucleic Acids Research.*, 34(suppl_1):D504–506.

Daniel M. Bikel. 2004. Intricacies of Collins' Parsing Model. *Computational Linguistics*, 30(4):479–511.

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.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180.

Andrew Chatr-aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Research*, 35(suppl 1):D572–574.

Nancy Chinchor. 1998. Overview of MUC-7/MET-2. In *Message Understanding Conference (MUC-7) Proceedings*.

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.

William Hersh, Aaron Cohen, Ruslenm Lynn, , and Phoebe Roberts. 2007. TREC 2007 Genomics track overview. In *Proceeding of the Sixteenth Text REtrieval Conference*.

Lynette Hirschman, Martin Krallinger, and Alfonso Valencia, editors. 2007. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. CNIO Centro Nacional de Investigaciones Oncológicas.

Yoshinobu Kano, William Baumgartner, Luke McCrohon, Sophia Ananiadou, Kevin Cohen, Larry Hunter, and Jun'ichi Tsujii. 2009. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*. To appear.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 70–75.

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

David McClosky and Eugene Charniak. 2008. Self-Training for Biomedical Parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT'08)*, pages 101–104.

Claire Nédellec. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge. In J. Cussens and C. Nédellec, editors, *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, pages 31–37.

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*. To appear.

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, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static Relations: a Piece in the Biomedical Information Extraction Puzzle. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*. To appear.

Laura Rimell and Stephen Clark. 2008. Porting a lexicalized-grammar parser to the biomedical domain. *Journal of Biomedical Informatics*, To Appear.

Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

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.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax Annotation for the GENIA corpus. In *Proceedings of the IJCNLP 2005, Companion volume*, pages 222–227.

Ellen Voorhees. 2007. Overview of TREC 2007. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*.

John Wilbur, Lawrence Smith, and Lorraine Tanabe. 2007. BioCreative 2. Gene Mention Task. In L. Hirschman, M. Krallinger, and A. Valencia, editors, *Proceedings of Second BioCreative Challenge Evaluation Workshop*, pages 7–16.