The Genia Event Extraction Shared Task, 2013 Edition - Overview

Jin-Dong Kim and **Yue Wang** and **Yamamoto Yasunori**
Database Center for Life Science (DBCLS)
Research Organization of Information and Systems (ROIS)
{jdkim|wang|yy}@dbcls.rois.ac.jp

Abstract

The Genia Event Extraction task is organized for the third time, in BioNLP Shared Task 2013. Toward knowledge based construction, the task is modified in a number of points. As the final results, it received 12 submissions, among which 2 were withdrawn from the final report. This paper presents the task setting, data sets, and the final results with discussion for possible future directions.

1 Introduction

Among various resources of life science, literature is regarded as one of the most important types of knowledge base. Nevertheless, lack of explicit structure in natural language texts prevents computer systems from accessing fine-grained information written in literature. *BioNLP Shared Task (ST)* series (Kim et al., 2009; Kim et al., 2011a) is one of the community-wide efforts to address the problem. Since its initial organization in 2009, BioNLP-ST series has published a number of fine-grained information extraction (IE) tasks motivated for bioinformatics projects. Having solicited wide participation from the community of natural language processing, machine learning, and bioinformatics, it has contributed to the production of rich resources for fine-grained BioIE, e.g., TEES[1] (Björne and Salakoski, 2011), SBEP[2] (McClosky et al., 2011) and EVEX[3] (Van Landeghem et al., 2011).

The Genia Event Extraction (GE) task is a seminal task of BioNLP-ST. It was first organized as the sole task of the initial 2009 edition of BioNLP-ST. The task was originally designed and implemented based on the Genia event corpus (Kim et al., 2008b) which represented domain knowledge around NFκB proteins. There were also some efforts to explore the possibility of literature mining for pathway construction (Kim et al., 2008a; Oda et al., 2008). The GE task was designed to make such an effort a community-driven one by sharing available resources, e.g., benchmark data sets, and evaluation tools, with the community.

In its second edition (Kim et al., 2011b) organized in BioNLP-ST 2011 (Kim et al., 2011a), the data sets were extended to include full text articles. The data sets consisted of two collections. The *abstract collection*, that had come from the first edition, was used again to measure the progress of the community between 2009 and 2011 editions, and the *full text collection*, that was newly created, was used to measure the generalization of the technology to full text papers.

In its third edition this year, while succeeding the fundamental characteristics from its previous editions, the GE task tries to evolve with the goal to make it a more "real" task toward knowledge base construction. The first design choice to address the goal is to construct the data sets fully with recent full papers, so that the extracted pieces of information can represent up-to-date knowledge of the domain. The abstract collection, that had been already used twice (in 2009 and 2011), is removed from official evaluation this time[4]. Second, GE task subsumes the coreference task which has long been considered critical for improvement of event extraction performance. It is implemented by providing coreference annotation in integration with event annotation in the data sets.

The paper explains the task setting and data sets, presents the final results of participating systems, and discusses notable observations with conclusions.

[1] https://github.com/jbjorne/TEES/wiki
[2] http://nlp.stanford.edu/software/eventparser.shtml
[3] http://www.evexdb.org/

[4] However, if necessary, the online evaluation for the previous editions of GE task may be used, which is available at http://bionlp-st.dbcls.jp/GE/.

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 8–15,
Sofia, Bulgaria, August 9 2013.

Event Type	Primary Argument	Secondary Argument
Gene_expression	Theme(Protein)	
Transcription	Theme(Protein)	
Localization	Theme(Protein)	Loc(Entity)?
Protein_catabolism	Theme(Protein)	
Binding	Theme(Protein)+	Site(Entity)*
Protein_modification	Theme(Protein), Cause(Protein/Event)?	Site(Entity)?
Phosphorylation	Theme(Protein), Cause(Protein/Event)?	Site(Entity)?
Ubiquitination	Theme(Protein), Cause(Protein/Event)?	Site(Entity)?
Acetylation	Theme(Protein), Cause(Protein/Event)?	Site(Entity)?
Deacetylation	Theme(Protein), Cause(Protein/Event)?	Site(Entity)?
Regulation	Theme(Protein/Event), Cause(Protein/Event)?	Site(Entity)?, CSite(Entity)?
Positive_regulation	Theme(Protein/Event), Cause(Protein/Event)?	Site(Entity)?, CSite(Entity)?
Negative_regulation	Theme(Protein/Event), Cause(Protein/Event)?	Site(Entity)?, CSite(Entity)?

Table 1: Event types and their arguments for Genia Event Extraction task. The type of each filler entity is specified in parenthesis. Arguments that may be filled more than once per event are marked with "+", and optional arguments are with "?".

2 Task setting

This section explains the task setting of the 2013 edition of the GE task with a focus on changes to previous editions. For comprehensive explanation, readers are referred to Kim et al. (2009).

The changes made to the task setting are threefolds, among which two are about event types to be extracted. Table 1 shows the event types and their arguments targeted in the 2013 edition. First, four new event types are added to the target of extraction; the `Protein_modification` type and its three sub-types, `Ubiquitination`, `Acetylation`, `Deacetylation`. Second, The `Protein_modification` types are modified to be directly linked to causal entities, which was only possible through `Regulation` events in previous editions.

The modifications were made based on analysis on preliminary annotation during preparation of the data sets: in recent papers on *NFκB*, discussions on protein modification were observed with non-trivial frequency. However, in the end, it turned out that the influence of the above modifications was trivial in terms of the number of annotated instances in the final data sets, as shown in section 3, after filtering out events on non-individual proteins, e.g., protein families, protein complexes.

Third change made to the task setting is addition of coreference and part-of annotations to the data sets. It is to address the observation from 2009 edition that coreference structures and entity relations often hide the syntactic paths between event triggers and their arguments, restricting the performance of event extraction. In 2011, the *Protein coreference task* and *Entity Relation* were organized as sub-tasks, to explicitly address the problem, but this time, coreference and part-of annotations are integrated in the GE task, to encourage an integrative use of them for event extraction. Figure 1 shows an example of annotation with coreference and part-of annotations[5]. Note that the event representation in the figure is relation centric[6], which is different from the event centric representation of the default BioNLP-ST format. The two representations are interchangeable, and the GE task provides data sets in both formats, together with an automatic converter between them. Below is the corresponding annotation in the BioNLP-ST format:

```
T8  Protein  933 938   TRAF1
T9  Protein  940 945   TRAF2
T10 Protein  947 952   TRAF3
T11 Protein  958 963   TRAF6
T12 Protein  1038 1042 CD40
T41 Anaphora 1058 1072 These proteins
T48 Binding  1112 1119 binding
T49 Entity   1127 1143 cytoplasmic tail
T13 Protein  1147 1151 CD40
R1  Coreference Subject:T41 Object:T8
R2  Coreference Subject:T41 Object:T9
R3  Coreference Subject:T41 Object:T10
R4  Coreference Subject:T41 Object:T11
E4  Binding:T48 Theme:T8 Theme2:T13 Site2:T49
E5  Binding:T48 Theme:T9 Theme2:T13 Site2:T49
E6  Binding:T48 Theme:T10 Theme2:T13 Site2:T49
E7  Binding:T48 Theme:T11 Theme2:T13 Site2:T49
```

In the example, the event trigger, *binding*, denotes four binding events, in which the four proteins, *TRAF1*, *TRAF2*, *TRAF3*, and *TRAF6*, bind to the protein, *CD40*, respectively, through the site, *cytoplasmic tail*. The links between the four

[5]The example is taken from the file, PMC-3148254-01-Introduction.

[6]PubAnnotation (http://pubannotation.org) format.

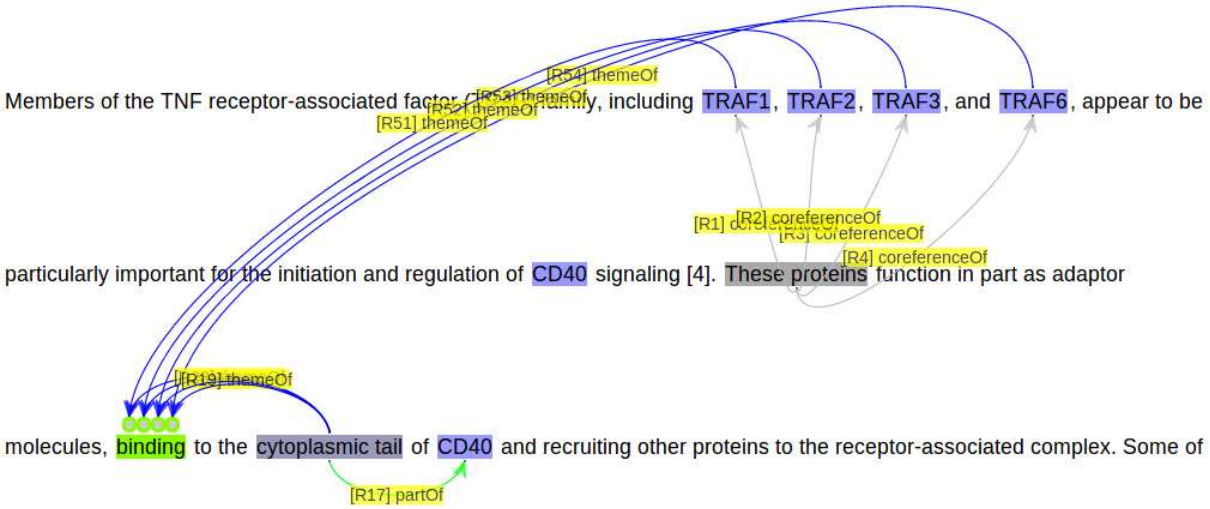

Figure 1: Annotation example with coreferences and part-of relationship

proteins and the event trigger are however very hard to find, without being bridged by the demonstrative noun phrase (NP), *These proteins*. In the case, if the link between the demonstrative NP, *These proteins* and its four antecedents, *TRAF1*, *TRAF2*, *TRAF3*, and *TRAF6*, can be somehow detected, the remaining link, between the demonstrative NP and the trigger, may be detected by their syntactic connection. A key point here is the different characteristics of the two step links: detecting the former is rather semantic or discoursal while the latter may be a more syntactic problem. Then, solving them using different processes would make a sense. To encourage an exploration into the hypothesis, the coreference annotation is provided in the training and development data sets.

Based on the definition of event types, the entire task is divided into three sub-tasks addressing event extraction at different levels of specificity:

Task 1. Core event extraction addresses the extraction of typed events together with their primary arguments.

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

Task 3. Negation/Speculation detection addresses the detection of negations and speculations over the extracted events.

For more detail of the subtasks, readers are referred to Kim et al. (2011b).

Item	Training	Devel	Test
Articles	10	10	14
Words	54938	57907	75144
Proteins	3571	4138	4359
Entities	121	314	327
Events	2817	3199	3348
Gene_expression	729	591	619
Transcription	122	98	101
Localization	44	197	99
Protein_catabolism	23	30	14
Binding	195	376	342
Protein_modification	8	1	1
Phosphorylation	117	197	161
Ubiquitination	4	2	30
Acetylation	0	3	0
Deacetylation	0	5	0
Regulation	299	284	299
Positive_regulation	780	883	1144
Negative_regulation	496	532	538
Coreferences	178	160	197
to Protein	152	123	169
to Entity	5	6	6
to Event	18	27	13
to Anaphora	3	4	9

Table 2: Statistics of annotations in training, development, and test sets

3 Data Preparation

As discussed in section 1, for the 2013 edition, the data sets are constructed fully with full text papers. Table 2 shows statistics of three data sets for training, development and test. The data sets consist of 34 full text papers from the Open Access subset of PubMed Central. The papers were retrieved using lexical variants of the term, "*NFκB*" as primary keyword, and "*pathway*" and "*regulation*" as secondary keywords. The retrieved papers were given to the annotators with higher priority

Item	TIAB	Intro.	R/D/C	Methods	Caption	all
Words	10483	25543	125172	59612	29085	263133
Proteins	816	1507	9060	1797	2169	16427
(Density: P / W)	(7.78%)	(5.90%)	(7.24%)	(3.01%)	(7.46%)	(6.24%)
Prot. Coreferences	18	89	267	5	33	445
(Density: C / P)	(2.21%)	(5.91%)	(2.95%)	(0.28%)	(1.52%)	(2.71%)
Events	510	902	6391	311	892	9364
(Density: E / W)	(4.87%)	(3.53%)	(5.11%)	(0.52%)	(3.07%)	(3.56%)
(Density: E / P)	(62.50%)	(59.85%)	(70.54%)	(17.31%)	(41.12%)	(57.00%)
Gene_expression	101	152	1265	125	220	1939
Transcription	10	18	209	36	47	321
Localization	19	47	191	8	41	340
Protein_catabolism	0	3	49	0	8	67
Binding	29	158	572	15	92	913
Protein_modification	1	1	7	0	0	10
Phosphorylation	27	38	347	19	35	475
Ubiquitination	0	2	8	0	10	36
Acetylation	0	3	0	0	0	3
Deacetylation	0	5	0	0	0	5
Regulation	67	76	625	7	66	882
Positive_regulation	167	286	2045	19	203	2807
Negative_regulation	89	113	1073	69	170	1566

Table 3: Statistics of annotations in different sections of text: the *Abstract* column is of the abstraction collection (1210 titles and abstracts), and the following columns are of full paper collection (14 full papers). *TIAB* = title and abstract, *Intro.* = introduction and background, *R/D/C* = results, discussions, and conclusions, *Methods* = methods, materials, and experimental procedures. Some minor sections, supporting information, supplementary material, and synopsis, are ignored. *Density* = relative density of annotation (P/W = Protein/Word, E/W = Event/Word, and E/P = Event/Protein).

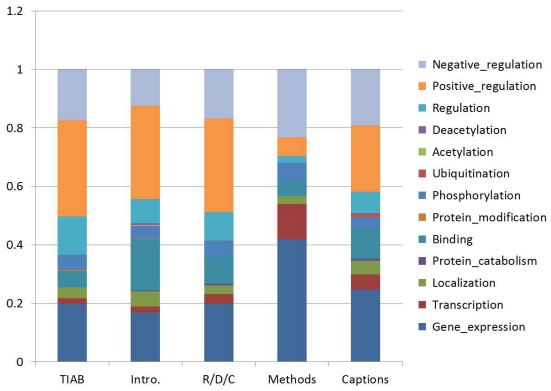

Figure 2: Event distribution in different sections

to newer ones. Note that among 34 papers, 14 were from the *full text collection* of 2011 edition data sets, and 20 were newly collected this time. The annotation to the all 34 papers were produced by the same annotators who also produced annotations for the previous editions of GE task.

The annotated papers are divided into the training, development, and test data sets; 10, 10, and 14, respectively. Note that the size of the training data set is much smaller than previous editions, in terms of number of words and events, while the size of the development and test data sets are comparable to previous editions. It is the consequence of a design choice of the organizers with the notion that (1) relevant resources are substantially accumulated through last two editions, and that (2) therefore the importance of training data set may be reduced while the importance of development and test data sets needs to be kept. Instead, participants may utilize, for example, the abstract collection of the 2011 edition, of which the annotation was produced by the same annotators with almost same principles. As another example, the data sets of the EPI task (Ohta et al., 2011) also may be utilized for the newly added protein modification events.

Table 3 shows the statistics of annotated event types in different sections of the full papers in the data sets. For the analysis, the sections are classified to five groups as follows:

- The `TIAB` group includes the titles and abstracts. In the GE-2011 data sets, the corresponding files match the pattern, `PMC-*TIAB*.txt`.
- The `Intro` group includes sections for introduction, and background. The corresponding files match the pattern, `PMC-*@(-|._)@(I|Back)*.txt`.

Team	'09	'11	Task	Expertise
EVEX	UTurku		123	2C+2BI+1B
TEES-2.1	UTurku		123	2BI
BioSEM		TM-SCS	1--	1C+1BI
NCBI	CCP-BTMG		1--	3BI
DlutNLP			1--	3C
HDS4NLP			1--	3C
NICTANLM	CCP-BTMG		1-3	6C
USheff			1--	2C
UZH	UZurich		1--	6C
HCMUS		HCMUS	1--	4C

Table 4: Team profiles: The **'09** and **'11** columns show the predecessors in 2009 and 2011 editions. In **Expertise** column, C=Computer Scientist, BI=Bioinformatician, B=Biologist, L=Linguist

- The `R/D/C` group includes sections on results, discussions, and conclusions. The files match the pattern, `PMC-*@(-|._)@(R|D|Conc)*.txt`
- The `Methods` group includes sections on methods, materials, and experimental procedures. The files match the pattern, `PMC-*@(-|._)@(Met|Mat|MAT|E)*.txt`
- The `Caption` group includes the captions of tables and figures. The corresponding files math the pattern, `PMC-*aption*.txt`.

Figure 2 illustrates the different distribution of annotated event types in the five section groups. It shows that the `Methods` group has significantly different distribution of annotated events, confirming a similar observation reported in Kim et al. (2011b).

4 Participation

The GE task received final submissions from 12 teams, among which 2 were withdrawn from final report. Table 4 summarizes the teams. Unfortunately, the subtasks 2 and 3 did not met a large participation.

Table 5 profiles the participating systems. The systems are roughly grouped into SVM-based pipeline (`EVEX`, `TEES-2.1`, and `DlutNLP`), rule-based pipeline (`BioSEM` and `UZH`), mixed pipeline (`USheff` and `HCMUS`), joint pattern matching (`NCBI` and `NICTANLM`), and joint SVM (`HDS4NLP`) systems. In terms of use of external resources, 5 teams (`EVEX`, `TEES-2.1`, `NCBI`, `DlutNLP`, and `USheff`) utilized data sets from 2011 edition, and two teams (`HDS4NLP` and `NICTANLM`) utilized independent resources, e.g., UniProt (Bairoch et al., 2005), IntAct (Kerrien et al., 2012), and CRAFT (Verspoor et al., 2012).

5 Results and Discussions

Table 6 shows the final results of subtask 1. Overall `EVEX`, `TEES-2.1`, and `BioSEM` show the best performance only with marginal difference between them. In detail, the performance of `BioSEM` is significantly different from `EVEX` and `TEES-2.1`: (1) while `BioSEM` show the best performance with *Binding* and *Protein_modification* events, `EVEX` and `TEES-2.1` show the best performance with *Regulation* events which takes the largest portion of annotation in data sets; and (2) while the performance of `EVEX` and `TEES-2.1` is balanced over recall and precision, `BioSEM` is biased for precision, which is a typical feature of rule-based systems. It is also notable that `BioSEM` has achieved a near best performance using only shallow parsing. Although it is not shown in the table, `NCBI` is the only system which produced `Ubiquitination` events, which is interpreted as a result of utilizing 2011-EPI data sets (Ohta et al., 2011) for the system development.

Table 7 shows subtask 1 final results only within *TIAB* sections. It shows that the systems developed utilizing previous resources, e.g., 2011 data sets, and EVEX, perform better for titles and abstracts, which makes sense because those resources are title and abstract-centric.

Tables 8 and 9 show evaluation results within *Methods* and *Captions* section groups, respectively. All the systems show their worst performance in the two section groups. Especially the drop of performance with regulation events is huge. Note the two section groups also show significantly different event distribution compared to other section groups (see section 3). It suggests that language expression in the two section groups may be quite different from other sections, and an extensive examination is required to get a reasonable performance in the sections.

Table 10 and 11 show final results of Task 2 (Event enrichment) and 3 (Negation/Speculation detection), respectively, which unfortunately did not meet a large participation.

6 Conclusions

In its third edition, the GE task is fully changed to a full text paper centric task, while the online evaluation service on the abstract-centric data sets

	NLP		Task			Other resources	
Team	**Lexical Proc.**	**Syntactic Proc.**	**Trig.**	**Arg.**	**group**	**Dic.**	**Other**
EVEX	Porter	McCCJ	SVM	SVM	SVM	S. cues	EVEX
TEES-2.1	Porter	McCCJ	SVM	SVM	SVM	S. cues	
BioSEM	OpenNLP, LingPipe	OpenNLP(shallow)	dic	rules	rules		
NCBI	MedPost, BioLemm	McCCJ	Subgraph Isomorphism		rules		2011 GE / EPI
DlutNLP	Porter, GTB-tok	McCCJ	SVM	SVM	rules		2011 GE
HDS4NLP	CNLP, Morpha	McCCJ	SVM		SVM		UniProt, IntAct
NICTANLM		ClearParser	Subgraph Isomorphism		rules		CRAFT, EVEX
USheff	Porter, LingPipe	Stanford	dic	SVM	SVM, rules		2011 GE
UZH	Porter, Morpha, LingPipe	LTT2, Pro3Gres	dic. MaxEnt	rules	rules		
HCMUS	SnowBall	McCCJ	dic, SVM	rules, SVM	rules		

Table 5: System profiles: SnowBall=SnowBall Stemmer, CNLP=Stanford CoreNLP (tokenization), McCCJ=McClosky-Charniak-Johnson Parser, Stanford=Stanford Parser, S.=Speculation, N.=Negation

Team	Simple Event	Binding	Prot-Mod.	Regulation	All
EVEX	73.83 / 79.56 / 76.59	41.14 / 44.77 / 42.88	61.78 / 69.41 / 65.37	**32.41 / 47.16 / 38.41**	**45.44 / 58.03 / 50.97**
TEES-2.1	74.19 / 79.64 / 76.82	42.34 / 44.34 / 43.32	63.87 / 69.32 / 66.49	**33.08 / 44.78 / 38.05**	**46.17 / 56.32 / 50.74**
BioSEM	67.71 / 86.90 / 76.11	**47.45 / 52.32 / 49.76**	**69.11 / 80.49 / 74.37**	28.19 / 49.06 / 35.80	**42.47 / 62.83 / 50.68**
NCBI	72.99 / 72.12 / 72.55	37.54 / 41.81 / 39.56	64.92 / 77.02 / 70.45	24.74 / 55.61 / 34.25	40.53 / 61.72 / 48.93
DlutNLP	69.15 / 80.56 / 74.42	40.84 / 44.16 / 42.43	62.83 / 77.42 / 69.36	26.49 / 43.46 / 32.92	40.81 / 57.00 / 47.56
HDS4NLP	**75.27 / 83.27 / 79.07**	41.74 / 33.74 / 37.32	70.68 / 75.84 / 73.17	16.67 / 30.86 / 21.64	37.11 / 51.19 / 43.03
NICTANLM	73.59 / 57.67 / 64.66	32.13 / 31.10 / 31.61	42.41 / 72.97 / 53.64	21.60 / 47.14 / 29.63	36.99 / 50.68 / 42.77
USheff	54.50 / 80.07 / 64.86	31.53 / 46.88 / 37.70	39.79 / 92.68 / 55.68	21.14 / 52.69 / 30.18	31.69 / 63.28 / 42.23
UZH	60.26 / 77.47 / 67.79	22.22 / 28.03 / 24.79	62.30 / 70.83 / 66.30	11.06 / 31.02 / 16.31	27.57 / 51.33 / 35.87
HCMUS	67.47 / 60.24 / 63.65	38.74 / 26.99 / 31.81	64.92 / 57.67 / 61.08	19.60 / 19.93 / 19.76	36.23 / 33.80 / 34.98

Table 6: Evaluation results (recall / precision / f-score) of Task 1. Some notable figures are emphasized in bold.

is kept maintained. Unfortunately, the coreference annotation, which has been integrated in the event annotation in the data sets, was not exploited by the participants, during the official shared task period. An analysis shows that the performance of systems significantly drops in the *Methods* and *Captions* sections, suggesting for an extensive examination in the sections.

As usual, after the official shared task period, the GE task is maintaining an online evaluation that can be freely accessed by anyone but with a time limitation; once in 24 hours per a person. With a few new features that are introduced in 2013 editions but are not fully exploited by the participants, the organizers solicit participants to continuously explore the task using the online evaluation. The organizers are also planning to provide more resources to the participants, based on the understanding that interactive communication between organizers and participants is important for progress of the participating systems and also the task itself.

References

Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. 2005. The universal protein resource (uniprot). *Nucleic Acids Research*, 33(suppl 1):D154–D159.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 183–191, Portland, Oregon, USA, June. Association for Computational Linguistics.

Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C. Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. 2012. The intact molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846.

Jin-Dong Kim, Tomoko Ohta, Kanae Oda, and Jun'ichi Tsujii. 2008a. From text to pathway: corpus annotation for knowledge acquisition from biomedical literature. In *Proceedings of the 6th Asia Pacific Bioinformatics Conference, Series on Advances in Bioin-*

formatics and Computational Biology, pages 165–176. Imperial College Press.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008b. 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. 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.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1635, Portland, Oregon, USA, June. Association for Computational Linguistics.

Kanae Oda, Jin-Dong Kim, Tomoko Ohta, Daisuke Okanohara, Takuya Matsuzaki, Yuka Tateisi, and Jun'ichi Tsujii. 2008. New challenges for text mining: Mapping between text and manually curated pathways.

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*, Portland, Oregon, June. Association for Computational Linguistics.

Sofie Van Landeghem, Filip Ginter, Yves Van de Peer, and Tapio Salakoski. 2011. Evex: A pubmed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of BioNLP 2011 Workshop*, pages 28–37, Portland, Oregon, USA, June. Association for Computational Linguistics.

Karin Verspoor, Kevin Cohen, Arrick Lanfranchi, Colin Warner, Helen Johnson, Christophe Roeder, Jinho Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William Baumgartner, Michael Bada, Martha Palmer, and Lawrence Hunter. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(1):207.

Team	Simple Event	Binding	Prot-Mod.	Regulation	All
EVEX	91.67 / 88.00 / 89.80	55.56 / 62.50 / 58.82	85.71 / 75.00 / 80.00	**51.18 / 59.09 / 54.85**	**62.83 / 68.18 / 65.40**
TEES-2.1	91.67 / 88.00 / 89.80	55.56 / 62.50 / 58.82	85.71 / 75.00 / 80.00	**51.18 / 57.02 / 53.94**	**62.83 / 66.67 / 64.69**
NCBI	81.25 / 79.59 / 80.41	55.56 / 45.45 / 50.00	85.71 / 66.67 / 75.00	37.01 / 67.14 / 47.72	50.79 / 69.78 / 58.79
BioSEM	83.33 / 88.89 / 86.02	**66.67 / 66.67 / 66.67**	85.71 / 75.00 / 80.00	35.43 / 54.22 / 42.86	50.79 / 66.90 / 57.74
DlutNLP	87.50 / 93.33 / 90.32	44.44 / 50.00 / 47.06	**85.71 / 85.71 / 85.71**	37.01 / 51.09 / 42.92	51.83 / 65.13 / 57.73
USheff	81.25 / 88.64 / 84.78	44.44 / 57.14 / 50.00	71.43 / 71.43 / 71.43	29.13 / 56.06 / 38.34	44.50 / 68.55 / 53.97
NICTANLM	93.75 / 57.69 / 71.43	22.22 / 25.00 / 23.53	42.86 /100.00 / 60.00	29.92 / 49.35 / 37.25	46.07 / 53.01 / 49.30
HDS4NLP	**93.75 / 90.00 / 91.84**	66.67 / 54.55 / 60.00	**85.71 / 85.71 / 85.71**	19.69 / 31.65 / 24.27	42.93 / 55.78 / 48.52
HCMUS	93.75 / 69.23 / 79.65	33.33 / 27.27 / 30.00	71.43 / 41.67 / 52.63	27.56 / 25.36 / 26.42	46.07 / 38.94 / 42.21
UZH	72.92 / 79.55 / 76.09	44.44 / 57.14 / 50.00	71.43 / 71.43 / 71.43	11.02 / 32.56 / 16.47	30.37 / 57.43 / 39.73

Table 7: Evaluation results (recall / precision / f-score) of Task 1 in titles and abstracts. Some notable figures are emphasized in bold.

Team	Simple Event	Binding	Prot-Mod.	Regulation	All
BioSEM	**70.83 / 90.44 / 79.44**	**48.24 / 53.93 / 50.93**	**74.17 / 82.41 / 78.07**	28.74 / 51.25 / 36.83	**42.97 / 64.90 / 51.70**
EVEX	73.51 / 83.26 / 78.08	43.72 / 47.80 / 45.67	66.67 / 66.12 / 66.39	**32.79 / 46.79 / 38.56**	45.29 / 58.05 / 50.88
TEES-2.1	74.09 / 83.37 / 78.46	43.72 / 47.80 / 45.67	66.67 / 65.04 / 65.84	**33.24 / 44.48 / 38.04**	45.70 / 56.34 / 50.46
NCBI	74.28 / 75.59 / 74.93	38.19 / 45.24 / 41.42	67.50 / 81.82 / 73.97	24.69 / 55.46 / 34.17	40.01 / 63.56 / 49.11
DlutNLP	70.06 / 84.49 / 76.60	39.20 / 44.32 / 41.60	67.50 / 74.31 / 70.74	27.78 / 43.23 / 33.83	41.01 / 56.70 / 47.60
NICTANLM	75.24 / 57.14 / 64.95	35.68 / 41.76 / 38.48	52.50 / 76.83 / 62.38	22.33 / 46.83 / 30.24	37.73 / 52.30 / 43.84
USheff	56.81 / 80.43 / 66.59	32.66 / 48.15 / 38.92	45.00 / 94.74 / 61.02	21.67 / 53.55 / 30.85	32.27 / 63.93 / 42.89
HDS4NLP	**76.20 / 84.65 / 80.20**	41.21 / 38.14 / 39.61	75.83 / 75.21 / 75.52	16.58 / 30.16 / 21.40	36.19 / 51.26 / 42.42
UZH	63.53 / 78.25 / 70.13	23.12 / 28.75 / 25.63	66.67 / 74.07 / 70.18	10.61 / 29.39 / 15.59	27.36 / 50.89 / 35.58
HCMUS	67.18 / 62.84 / 64.94	38.19 / 28.15 / 32.41	67.50 / 61.83 / 64.54	19.45 / 20.11 / 19.78	35.09 / 33.95 / 34.51

Table 8: Evaluation results (recall / precision / f-score) of Task 1 in *Methods* section group. Some notable figures are emphasized in bold.

Team	Simple Event	Binding	Prot-Mod.	Regulation	All
TEES-2.1	76.67 / 67.65 / 71.88	53.19 / 46.30 / 49.50	60.61 / 76.92 / 67.80	**22.68 / 39.29 / 28.76**	**43.41/53.74 / 48.02**
BioSEM	60.00 / 78.26 / 67.92	**68.09 / 58.18 / 62.75**	**69.70 / 82.14 / 75.41**	23.20 / 34.35 / 27.69	42.31/54.42 / 47.60
EVEX	76.67 / 67.65 / 71.88	53.19 / 46.30 / 49.50	48.48 / 72.73 / 58.18	21.13 / 39.81 / 27.61	41.48/53.74 / 46.82
DlutNLP	70.00 / 67.02 / 68.48	55.32 / 48.15 / 51.49	57.58 / 79.17 / 66.67	18.04 / 46.67 / 26.02	39.29/57.89 / 46.81
NCBI	80.00 / 58.54 / 67.61	40.43 / 41.30 / 40.86	66.67 / 70.97 / 68.75	14.95 / 44.62 / 22.39	39.01/53.58 / 45.15
HDS4NLP	**78.89 / 78.02 / 78.45**	48.94 / 29.49 / 36.80	66.67 / 68.75 / 67.69	06.19 / 14.63 / 08.70	35.16/45.23 / 39.57
UZH	57.78 / 68.42 / 62.65	23.40 / 26.19 / 24.72	69.70 / 74.19 / 71.88	12.89 / 43.10 / 19.84	30.49/53.62 / 38.88
USheff	47.78 / 74.14 / 58.11	36.17 / 45.95 / 40.48	30.30 /100.00 / 46.51	13.40 / 45.61 / 20.72	26.37/59.26 / 36.50
NICTANLM	75.56 / 53.12 / 62.39	40.43 / 27.94 / 33.04	18.18 / 54.55 / 27.27	11.34 / 36.67 / 17.32	31.59/43.07 / 36.45
HCMUS	73.33 / 52.80 / 61.40	53.19 / 25.51 / 34.48	63.64 / 53.85 / 58.33	15.46 / 17.96 / 16.62	39.01/33.10 / 35.81

Table 9: Evaluation results (recall / precision / f-score) of Task 1 in *Captions* section group. Some notable figures are emphasized in bold.

Team	Site-Binding	Site-Phosphorylation	Loc-Localization	Total
TEES-2.1	31.37 / 56.14 / 40.25	37.21 / 82.05 / 51.20	36.67 / 78.57 / 50.00	22.03 / 61.90 / 32.50
EVEX	31.37 / 56.14 / 40.25	32.56 / 80.00 / 46.28	36.67 / 78.57 / 50.00	20.90 / 61.67 / 31.22

Table 10: Evaluation results (recall / precision / f-score) of Task 2

Team	Negation	Speculation	Total
TEES-2.1	21.68 / 36.84 / 27.30	18.46 / 33.96 / 23.92	19.53 / 35.59 / 25.22
EVEX	20.98 / 38.03 / 27.04	18.46 / 32.73 / 23.61	19.82 / 34.41 / 25.15
NICTANLM	15.38 / 32.76 / 20.94	14.36 / 34.15 / 20.22	14.79 / 33.57 / 20.54

Table 11: Evaluation results (recall / precision / f-score) of Task 3