Overview of Genia Event Task in BioNLP Shared Task 2011

Jin-Dong Kim
Database Center for Life Science
2-11-16 Yayoi, Bunkyo-ku, Tokyo
jdkim@dbcls.rois.ac.jp

Yue Wang
Database Center for Life Science
2-11-16 Yayoi, Bunkyo-ku, Tokyo
wang@dbcls.rois.ac.jp

Toshihisa Takagi
University of Tokyo
5-1-5 Kashiwa-no-ha, Kashiwa, Chiba
tt@k.u-tokyo.ac.jp

Akinori Yonezawa
Database Center for Life Science
2-11-16 Yayoi, Bunkyo-ku, Tokyo
yonezawa@dbcls.rois.ac.jp

Abstract

The Genia event task, a bio-molecular event extraction task, is arranged as one of the main tasks of BioNLP Shared Task 2011. As its second time to be arranged for community-wide focused efforts, it aimed to measure the advance of the community since 2009, and to evaluate generalization of the technology to full text papers. After a 3-month system development period, 15 teams submitted their performance results on test cases. The results show the community has made a significant advancement in terms of both performance improvement and generalization.

1 Introduction

The BioNLP Shared Task (BioNLP-ST, hereafter) is a series of efforts to promote a community-wide collaboration towards fine-grained information extraction (IE) in biomedical domain. The first event, BioNLP-ST 2009, introducing a bio-molecular event (bio-event) extraction task to the community, attracted a wide attention, with 42 teams being registered for participation and 24 teams submitting final results (Kim et al., 2009).

To establish a community effort, the organizers provided the task definition, benchmark data, and evaluations, and the participants competed in developing systems to perform the task. Meanwhile, participants and organizers communicated to develop a better setup of evaluation, and some provided their tools and resources for other participants, making it a collaborative competition.

The final results enabled to observe the state-of-the-art performance of the community on the bio-event extraction task, which showed that the automatic extraction of simple events - those with unary arguments, e.g. gene expression, localization, phosphorylation - could be achieved at the performance level of 70% in F-score, but the extraction of complex events, e.g. binding and regulation, was a lot more challenging, having achieved 40% of performance level.

After BioNLP-ST 2009, all the resources from the event were released to the public, to encourage continuous efforts for further advancement. Since then, several improvements have been reported (Miwa et al., 2010b; Poon and Vanderwende, 2010; Vlachos, 2010; Miwa et al., 2010a; Björne et al., 2010). For example, Miwa et al. (Miwa et al., 2010b) reported a significant improvement with binding events, achieving 50% of performance level.

The task introduced in BioNLP-ST 2009 was renamed to *Genia event (GE) task*, and was hosted again in BioNLP-ST 2011, which also hosted four other IE tasks and three supporting tasks (Kim et al., 2011). As the sole task that was repeated in the two events, the GE task was referenced during the development of other tasks, and took the role of connecting the results of the 2009 event to the main tasks of 2011. The GE task in 2011 received final submissions from 15 teams. The results show the community made a significant progress with the task, and also show the technology can be generalized to full papers at moderate cost of performance.

This paper presents the task setup, preparation, and discusses the results.

Proceedings of BioNLP Shared Task 2011 Workshop, pages 7–15,
Portland, Oregon, USA, 24 June, 2011.

Event Type	Primary Argument	Secondary Argument
Gene_expression	Theme(Protein)	
Transcription	Theme(Protein)	
Protein_catabolism	Theme(Protein)	
Phosphorylation	Theme(Protein)	Site(Entity)
Localization	Theme(Protein)	AtLoc(Entity), ToLoc(Entity)
Binding	Theme(Protein)+	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 task. The type of each filler entity is specified in parenthesis. Arguments that may be filled more than once per event are marked with "+".

2 Task Definition

The GE task follows the task definition of BioNLP-ST 2009, which is briefly described in this section. For more detail, please refer to (Kim et al., 2009).

Table 1 shows the event types to be addressed in the task. For each event type, the primary and secondary arguments to be extracted with an event are defined. For example, a *Phosphorylation* event is primarily extracted with the protein to be phosphorylated. As secondary information, the specific site to be phosphorylated may be extracted.

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 in Table 1 are classified as *simple event types*, requiring only unary arguments. The *Binding* and *Regulation* types are more complex: *Binding* requires detection of an arbitrary number of arguments, and *Regulation* requires detection of recursive event structure.

Based on the definition of event types, the entire task is divided to 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.

Task 1 serves as the backbone of the GE task and is mandatory for all participants, while the other two are optional.

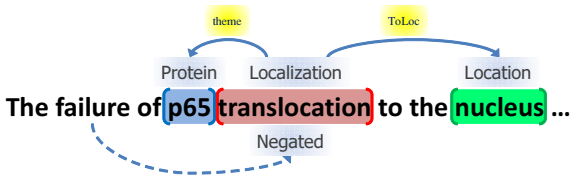

Figure 1: Event annotation example

Figure 1 shows an example of event annotation. The event encoded in the text is represented in a standoff-style annotation as follows:

```
T1  Protein       15 18
T2  Localization  19 32
T3  Entity        40 46
E1  Localization:T2 Theme:T1 ToLoc:T1
M1  Negation  E1
```

The annotation `T1` identifies the entity referred to by the string (*p65*) between the character offsets, *15* and *18* to be a *Protein*. `T2` identifies the string, *translocation*, to refer to a *Localization* event. Entities other than proteins or event type references are classified into a default class *Entity*, as in `T3`. `E1` then represents the event defined by the three entities, as defined in Table 1. Note that for Task 1, the entity, `T3`, does not need to be identified, and the event, `E1`, may be identified without specification of the secondary argument, `ToLoc:T1`:

```
E1'  Localization:T2 Theme:T1
```

Finding the full representation of `E1` is the goal of Task 2. In the example, the localization event, `E1`, is negated as expressed in *the failure of*. Finding the negation, `M1` is the goal of Task 3.

Item	Training		Devel		Test	
	Abs.	**Full**	**Abs.**	**Full**	**Abs.**	**Full**
Articles	*800*	*5*	*150*	*5*	*260*	*4*
Words	*176146*	*29583*	*33827*	*30305*	*57256*	*21791*
Proteins	9300	2325	2080	2610	3589	1712
Events	8615	1695	1795	1455	3193	1294
Gene_expression	1738	527	356	393	722	280
Transcription	576	91	82	76	137	37
Protein_catabolism	110	0	21	2	14	1
Phosphorylation	169	23	47	64	139	50
Localization	265	16	53	14	174	17
Binding	887	101	249	126	349	153
Regulation	961	152	173	123	292	96
Positive_regulation	2847	538	618	382	987	466
Negative_regulation	1062	247	196	275	379	194

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

3 Data preparation

The data sets are prepared in two collections: the abstract and the full text collections. The *abstract collection* includes the same data used for BioNLP-ST 2009, and is meant to be used to measure the progress of the community. The *full text collection* includes full papers which are newly annotated, and is meant to be used to measure the generalization of the technology to full papers. Table 2 shows the statistics of the annotations in the GE task data sets. Since the training data from the full text collection is relatively small despite of the expected rich variety of expressions in full text, it is expected that 'generalization' of a model from the abstract collection to full papers would be a key technique to get a reasonable performance.

A full paper consists of several sections including the title, abstract, introduction, results, conclusion, methods, and so on. Different sections would be written with different purposes, which may affect the type of information that are found in the sections. Table 3 shows the distribution of annotations in different sections. It indicates that event mentions, according to the event definition in Table 1, in *Methods* and *Captions* are much less frequent than in the other *TIAB*, *Intro.* and *R/D/C* sections. Figure 2 illustrates the different distribution of annotated event types in the five sections. It is notable that the *Methods* section (depicted in blue) shows very different distribution compared to others: while

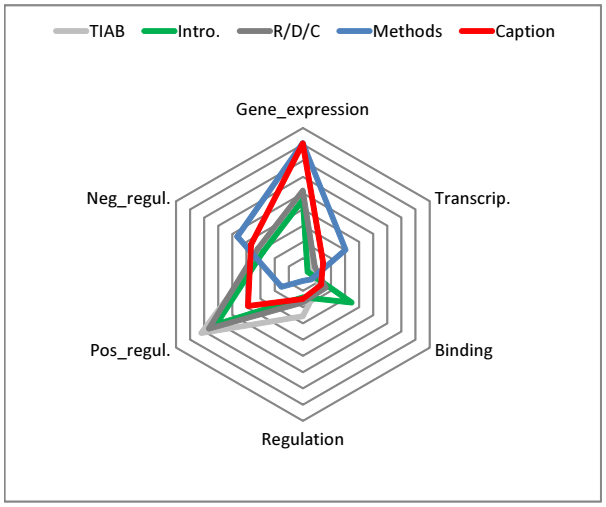

Figure 2: Event distribution in different sections

Regulation and *Positive_regulation* events are not as frequent as in other sections, *Negative_regulation* is relatively much more frequent. It may agree with an intuition that experimental devices, which will be explained in *Methods* sections, often consists of artificial processes that are designed to cause a negative regulatory effect, e.g. mutation, addition of inhibitor proteins, etc. This observation suggests a different event annotation scheme, or a different event extraction strategy would be required for *Methods* sections.

Item	Abstract	Full Paper					
		Whole	TIAB	Intro.	R/D/C	Methods	Caption
Words	*267229*	80962	3538	7878	43420	19406	6720
Proteins	*14969*	6580	336	597	3980	916	751
(Density: P / W)	*(5.60%)*	(8.13%)	(9.50%)	(7.58%)	(9.17%)	(4.72%)	(11.18%)
Events	*13603*	4436	272	427	3234	198	278
(Density: E / W)	*(5.09%)*	(5.48%)	(7.69%)	(5.42%)	(7.51%)	(1.02%)	(4.14%)
(Density: E / P)	*(90.87%)*	(67.42%)	(80.95%)	(71.52%)	(81.93%)	(21.62%)	(37.02%)
Gene_expression	*2816*	1193	62	98	841	80	112
Transcription	*795*	204	7	7	140	30	20
Protein_catabolism	*145*	3	0	0	3	0	0
Phosphorylation	*355*	137	12	12	101	10	2
Localization	*492*	47	3	15	22	7	0
Binding	*1485*	380	16	74	266	6	18
Regulation	*1426*	371	35	30	281	4	21
Positive_regulation	*4452*	1385	98	131	1087	15	54
Negative_regulation	*1637*	716	39	60	520	46	51

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).

4 Participation

In total, 15 teams submitted final results. All 15 teams participated in the mandatory Task 1, four teams in Task 2, and two teams in Task 3. Only one team, `UTurku`, completed all the three tasks.

Table 4 shows the profile of the teams, excepting three who chose to remain anonymous. A brief examination on the team organization (the **People** column) suggests the importance of a computer science background, `C` and `BI`, to perform the GE task, which agrees with the same observation made in 2009. It is interpreted as follows: the role of computer scientists may be emphasized in part due to the fact that the task requires complex computational modeling, demanding particular efforts in framework design and implementation and computational resources. The **'09** column suggests that previous experience in the task may have affected to the performance of the teams, especially in a complex task like the GE task.

Table 5 shows the profile of the systems. A notable observation is that four teams developed their systems based on the model of `UTurku09` (Björne et al., 2009) which was the winning system of BioNLP-ST 2009. It may show an influence of the BioNLP-ST series in the task. For syntactic analyses, the prevailing use of Charniak Johnson re-ranking parser (Charniak and Johnson, 2005) using the self-trained biomedical model from McClosky (2008) (*McCCJ*) which is converted to Stanford Dependency (de Marneffe et al., 2006) is notable, which may also be an influence from the results of BioNLP-ST 2009. The last two teams, `XABioNLP` and `HCMUS`, who did not use syntactic analyses could not get a performance comparable to the others, which may suggest the importance of using syntactic analyses for a complex IE task like GE task.

5 Results

5.1 Task 1

Table 6 shows the final evaluation results of Task 1. For reference, the reported performance of the two systems, `UTurku09` and `Miwa10` is listed in the top. `UTurku09` was the winning system of Task 1 in 2009 (Björne et al., 2009), and `Miwa10` was the best system reported after BioNLP-ST 2009 (Miwa et al., 2010b). Particularly, the latter made

Team	'09	Task	People	reference
FAUST	√	12-	3C	(Riedel et al., 2011)
UMASS	√	12-	1C	(Riedel and McCallum, 2011)
UTurku	√	123	1BI	(Bjrne and Salakoski, 2011)
MSR-NLP		1--	4C	(Quirk et al., 2011)
ConcordU	√	1-3	2C	(Kilicoglu and Bergler, 2011)
UWMadison	√	1--	2C	(Vlachos and Craven, 2011)
Stanford		1--	3C+1.5L	(McClosky et al., 2011)
BMI@ASU	√	12-	3C	(Emadzadeh et al., 2011)
CCP-BTMG	√	1--	3BI	(Liu et al., 2011)
TM-SCS		1--	1C	(Bui and Sloot, 2011)
XABioNLP		1--	4C	(Casillas et al., 2011)
HCMUS		1--	6L	(Minh et al., 2011)

Table 4: Team profiles: The **'09** column indicates whether at least one team member participated in BioNLP-ST 2009. In **People** column, C=Computer Scientist, BI=Bioinformatician, B=Biologist, L=Linguist

	NLP		Task			Other resources	
Team	**Lexical Proc.**	**Syntactic Proc.**	**Trig.**	**Arg.**	**group**	**Dictionary**	**Other**
FAUST	SnowBall, CNLP	McCCJ+SD	Stacking (UMASS + Stanford)				
UMASS	SnowBall, CNLP	McCCJ+SD	Joint infer., Dual Decomposition				
UTurku	Porter	McCCJ+SD	SVM	SVM	SVM	S. cues	
MSR-NLP	Porter	McCCJ+SD, Enju	SVM	MaxEnt	rules		Coref(Hobbs)
ConcordU	-	McCCJ+SD	dic	rules	rules	S./N. cues	
UWMadison	Morpha, Porter	MCCCJ+SD	Joint infer., SEARN				
Stanford	Morpha, CNLP	McCCJ+SD	MaxEnt	MSTParser			word clusters
BMI@ASU	Porter, WordNet	Stanford+SD	SVM	SVM	-		MeSH
CCP-BTMG	Porter, WordNet	Stanford+SD	Subgraph Isomorphism				
TM-SCS	Stanford	Stanford	dic	rules	rules		
XABioNLP	KAF	-	rules				
HCMUS	OpenNLP	-	dic, rules	rules			UIMA

Table 5: System profiles: SnowBall=SnowBall Stemmer, CNLP=Stanford CoreNLP (tokenization), KAF=Kyoto Annotation Format McCCJ=McClosky-Charniak-Johnson Parser, Stanford=Stanford Parser, SD=Stanford Dependency Conversion, S.=Speculation, N.=Negation

an impressive improvement with Binding events (44.41%→52.62%).

The best performance in Task 1 this time is achieved by the FAUST system, which adopts a combination model of UMass and Stanford. Its performance on the *abstract collection*, 56.04%, demonstrates a significant improvement of the community in the repeated GE task, when compared to both UTurku09, 51.95% and Miwa10, 53.29%. The biggest improvement is made to the Regulation events (40.11%→46.97%) which requires a complex modeling for recursive event structure - an event may become an argument of another event. The second ranked system, UMass, shows the best performance on the *full paper collection*. It suggests that what FAUST obtained from the model combination might be a better optimization to abstracts.

The ConcordU system is notable as it is the sole rule-based system that is ranked above the average. It shows a performance optimized for precision with relatively low recall. The same tendency is roughly replicated by other rule-based systems, CCP-BTMG, TM-SCS, XABioNLP, and HCMUS. It suggests that a rule-based system might not be a good choice if a high coverage is desired. However, the performance of ConcordU for simple events suggests that a high precision can be achieved by a rule based system with a modest loss of recall. It might be more true when the task is less complex.

This time, three teams achieved better results than Miwa10, which indicates some role of focused efforts like BioNLP-ST. The comparison between the

performance on abstract and full paper collections shows that generalization to full papers is feasible with very modest loss in performance.

5.2 Task 2

Tables 7 shows final evaluation results of Task 2. For reference, the reported performance of the task-winning system in 2009, `UT+DBCLS09` (Riedel et al., 2009), is shown in the top. The first and second ranked system, `FAUST` and `UMass`, which share a same author with `Riedel09`, made a significant improvement over `Riedel09` in the *abstract collection*. `UTurku` achieved the best performance in finding sites arguments but did not produce location arguments. In table 7, the performance of all the systems in *full text collection* suggests that finding secondary arguments in full text is much more challenging.

In detail, a significant improvement was made for *Location* arguments (36.59%→50.00%). A further breakdown of the results of *site* extraction, shown in table 8, shows that finding *site* arguments for *Phosphorylation*, *Binding* and *Regulation* events are all significantly improved, but in different ways. The extraction of protein sites to be phosphorylated is approaching a practical level of performance (84.21%), while protein sites to be bound or to be regulated remains challenging to be extracted.

5.3 Task 3

Table 9 shows final evaluation results of Task 3. For reference, the reported performance of the task-winning system in 2009, `Kilicoglu09`(Kilicoglu and Bergler, 2009), is shown in the top. Among the two teams participated in the task, `UTurku` showed a better performance in extracting negated events, while `ConcordU` showed a better performance in extracting speculated events.

6 Conclusions

The Genia event task which was repeated for BioNLP-ST 2009 and 2011 took a role of measuring the progress of the community and generalization IE technology to full papers. The results from 15 teams who made their final submissions to the task show that a clear advance of the community in terms of the performance on a focused domain and also generalization to full papers. To our disappointment, however, an effective use of supporting task results was not observed, which thus remains as future work for further improvement.

Acknowledgments

This work is supported by the "Integrated Database Project" funded by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

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

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at PubMed scale. *Bioinformatics*, 26(12):i382–390.

Jari Bjrne 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.

Quoc-Chinh Bui and Peter. M.A. Sloot. 2011. Extracting biological events from text using simple syntactic patterns. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Arantza Casillas, Arantza Daz de Ilarraza, Koldo Gojenola, Maite Oronoz, and German Rigau. 2011. Using Kybots for Extracting Events in Biomedical Texts. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

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.

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.

Ehsan Emadzadeh, Azadeh Nikfarjam, and Graciela Gonzalez. 2011. Double Layered Learning for Biological Event Extraction from Text. In *Proceedings*

of the BioNLP 2011 Workshop Companion Volume for Shared Task, Portland, Oregon, June. Association for Computational Linguistics.

Halil Kilicoglu and Sabine Bergler. 2009. Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 119–127, Boulder, Colorado, June. Association for Computational Linguistics.

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, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

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 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.

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.

Quang Le Minh, Son Nguyen Truong, and Quoc Ho Bao. 2011. A pattern approach for Biomedical Event Annotation . In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Makoto Miwa, Sampo Pyysalo, Tadayoshi Hara, and Jun'ichi Tsujii. 2010a. A comparative study of syntactic parsers for event extraction. In *Proceedings of BioNLP'10*, pages 37–45.

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

Hoifung Poon and Lucy Vanderwende. 2010. Joint inference for knowledge extraction from biomedical literature. In *Proceedings of NAACL-HLT'10*, pages 813–821.

Chris Quirk, Pallavi Choudhury, Michael Gamon, and Lucy Vanderwend. 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, Hong-Woo Chun, Toshihisa Takagi, and Jun'ichi Tsujii. 2009. A markov logic approach to bio-molecular event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 41–49, Boulder, Colorado, June. Association for Computational Linguistics.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, and Christopher 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.

Andreas Vlachos and Mark Craven. 2011. Biomedical Event Extraction from Abstracts and Full Papers using Search-based Structured Prediction. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Andreas Vlachos. 2010. Two strong baselines for the bionlp 2009 event extraction task. In *Proceedings of BioNLP'10*, pages 1–9.

Team		Simple Event	Binding	Regulation	All
UTurku09	A	*64.21 / 77.45 / 70.21*	*40.06 / 49.82 / 44.41*	*35.63 / 45.87 / 40.11*	*46.73 / 58.48 / 51.95*
Miwa10	A	*70.44*	*52.62*	*40.60*	*48.62 / 58.96 / 53.29*
FAUST	W	**68.47 / 80.25 / 73.90**	44.20 / 53.71 / 48.49	**38.02 / 54.94 / 44.94**	**49.41 / 64.75 / 56.04**
	A	**66.16 / 81.04 / 72.85**	**45.53 / 58.09 / 51.05**	**39.38 / 58.18 / 46.97**	**50.00 / 67.53 / 57.46**
	F	75.58 / 78.23 / 76.88	40.97 / 44.70 / 42.75	**34.99 / 48.24 / 40.56**	47.92 / 58.47 / 52.67
UMass	W	67.01 / 81.40 / 73.50	**42.97 / 56.42 / 48.79**	37.52 / 52.67 / 43.82	48.49 / 64.08 / 55.20
	A	64.21 / 80.74 / 71.54	43.52 / 60.89 / 50.76	38.78 / 55.07 / 45.51	48.74 / 65.94 / 56.05
	F	**75.58 / 83.14 / 79.18**	**41.67 / 47.62 / 44.44**	34.72 / 47.51 / 40.12	**47.84 / 59.76 / 53.14**
UTurku	W	68.22 / 76.47 / 72.11	42.97 / 43.60 / 43.28	38.72 / 47.64 / 42.72	49.56 / 57.65 / 53.30
	A	64.97 / 76.72 / 70.36	45.24 / 50.00 / 47.50	40.41 / 49.01 / 44.30	50.06 / 59.48 / 54.37
	F	78.18 / 75.82 / 76.98	37.50 / 31.76 / 34.39	34.99 / 44.46 / 39.16	48.31 / 53.38 / 50.72
MSR-NLP	W	68.99 / 74.30 / 71.54	42.36 / 40.47 / 41.39	36.64 / 44.08 / 40.02	48.64 / 54.71 / 51.50
	A	65.99 / 74.71 / 70.08	43.23 / 44.51 / 43.86	37.14 / 45.38 / 40.85	48.52 / 56.47 / 52.20
	F	78.18 / 73.24 / 75.63	40.28 / 32.77 / 36.14	35.52 / 41.34 / 38.21	48.94 / 50.77 / 49.84
ConcordU	W	59.99 / **85.53** / 70.52	29.33 / 49.66 / 36.88	35.72 / 45.85 / 40.16	43.55 / 59.58 / 50.32
	A	56.51 / **84.56** / 67.75	29.97 / 49.76 / 37.41	36.24 / 47.09 / 40.96	43.09 / 60.37 / 50.28
	F	70.65 / **88.03** / 78.39	27.78 / 49.38 / 35.56	34.58 / 43.22 / 38.42	44.71 / 57.75 / 50.40
UWMadison	W	59.67 / 80.95 / 68.70	29.33 / 49.66 / 36.88	34.10 / 49.46 / 40.37	42.56 / 61.21 / 50.21
	A	54.99 / 79.85 / 65.13	34.87 / 56.81 / 43.21	34.54 / 50.67 / 41.08	42.17 / 62.30 / 50.30
	F	74.03 / 83.58 / 78.51	15.97 / 29.87 / 20.81	33.11 / 46.87 / 38.81	43.53 / 58.73 / 50.00
Stanford	W	65.79 / 76.83 / 70.88	**39.92 / 49.87 / 44.34**	27.55 / 48.75 / 35.21	42.36 / 61.08 / 50.03
	A	62.61 / 77.57 / 69.29	**42.36 / 54.24 / 47.57**	28.25 / 49.95 / 36.09	42.55 / 62.69 / 50.69
	F	75.58 / 75.00 / 75.29	**34.03 / 40.16 / 36.84**	26.01 / 46.08 / 33.25	41.88 / 57.36 / 48.41
BMI@ASU	W	62.09 / 76.55 / 68.57	27.90 / 44.92 / 34.42	22.30 / 40.26 / 28.70	36.91 / 56.63 / 44.69
	A	58.71 / 78.51 / 67.18	26.22 / 47.40 / 33.77	22.99 / 40.47 / 29.32	36.61 / 57.82 / 44.83
	F	72.47 / 72.09 / 72.28	31.94 / 40.71 / 35.80	20.78 / 39.74 / 27.29	37.65 / 53.93 / 44.34
CCP-BTMG	W	53.61 / 75.13 / 62.57	22.61 / 49.12 / 30.96	19.01 / 43.80 / 26.51	31.57 / 58.99 / 41.13
	A	50.93 / 74.50 / 60.50	25.65 / 53.29 / 34.63	19.54 / 43.47 / 26.96	31.87 / 59.02 / 41.39
	F	61.82 / 76.77 / 68.49	15.28 / 37.29 / 21.67	17.83 / 44.63 / 25.48	30.82 / 58.92 / 40.47
TM-SCS	W	57.33 / 71.34 / 63.57	34.01 / 44.77 / 38.66	16.39 / 25.37 / 19.91	32.73 / 45.84 / 38.19
	A	53.65 / 71.66 / 61.36	36.02 / 49.41 / 41.67	18.29 / 27.07 / 21.83	33.36 / 47.09 / 39.06
	F	68.57 / 70.59 / 69.57	29.17 / 35.00 / 31.82	12.20 / 21.02 / 15.44	31.14 / 42.83 / 36.06
XABioNLP	W	43.71 / 47.18 / 45.38	05.30 / 50.00 / 09.58	05.79 / 26.94 / 09.54	19.07 / 42.08 / 26.25
	A	39.76 / 45.90 / 42.61	06.34 / 56.41 / 11.40	04.72 / 23.21 / 07.84	17.91 / 40.74 / 24.89
	F	55.84 / 50.23 / 52.89	02.78 / 30.77 / 05.10	08.18 / 33.89 / 13.17	21.96 / 45.09 / 29.54
HCMUS	W	24.82 / 35.14 / 29.09	04.68 / 12.92 / 06.88	01.63 / 10.40 / 02.81	10.12 / 27.17 / 14.75
	A	22.42 / 37.38 / 28.03	04.61 / 10.46 / 06.40	01.69 / 10.37 / 02.91	09.71 / 27.30 / 14.33
	F	32.21 / 31.16 / 31.67	04.86 / 28.00 / 08.28	01.47 / 10.48 / 02.59	11.14 / 26.89 / 15.75

Table 6: Evaluation results (recall / precision / f-score) of Task 1 in (W)hole data set, (A)bstracts only, and (F)ull papers only. Some notable figures are emphasized in bold.

Team		Sites (222)	Locations (66)	All (288)
UT+DBCLS09	A		*23.08 / 88.24 / 36.59*	*32.14 / 72.41 / 44.52*
FAUST	W	32.88 / 70.87 / 44.92	36.36 / 75.00 / 48.98	**33.68 / 71.85 / 45.86**
	A	43.51 / 71.25 / 54.03	**36.92 / 77.42 / 50.00**	**41.33 / 72.97 / 52.77**
	F	17.58 / 69.57 / 28.07	-	17.39 / 66.67 / 27.59
UMass	W	31.98 / 71.00 / 44.10	**36.36 / 77.42 / 49.48**	32.99 / 72.52 / 45.35
	A	42.75 / 70.00 / 53.08	36.92 / 77.42 / 50.00	40.82 / 72.07 / 52.12
	F	16.48 / 75.00 / 27.03	-	16.30 / 75.00 / 26.79
BMI@ASU	W	32.88 / 62.93 / 43.20	22.73 / 83.33 / 35.71	30.56 / 65.67 / 41.71
	A	37.40 / 67.12 / 48.04	23.08 / 83.33 / 36.14	32.65 / 70.33 / 44.60
	F	26.37 / 55.81 / 35.82	-	26.09 / 55.81 / 35.56
UTurku	W	**40.09 / 65.44 / 49.72**	00.00 / 00.00 / 00.00	30.90 / 65.44 / 41.98
	A	**48.09 / 69.23 / 56.76**	00.00 / 00.00 / 00.00	32.14 / 69.23 / 43.90
	F	**28.57 / 57.78 / 38.24**	-	**28.26 / 57.78 / 37.96**

Table 7: Evaluation results of Task 2 in (W)hole data set, (A)bstracts only, and (F)ull papers only

Team		Phospho. (67)	Binding (84)	Reg. (71)
Riedel'09	A	*71.43 / 71.43 / 71.43*	*04.76 / 50.00 / 08.70*	*12.96 / 58.33 / 21.21*
FAUST	W	71.64 / 84.21 / 77.42	05.95 / 38.46 / 10.31	**28.17 / 60.61 / 38.46**
	A	71.43 / 81.63 / 76.19	04.76 / 14.29 / 07.14	**29.63 / 66.67 / 41.03**
	F	**72.73 / 100.0 / 84.21**	06.35 / 66.67 / 11.59	23.53 / 44.44 / 30.77
UMass	W	76.12 / 79.69 / 77.86	04.76 / 36.36 / 08.42	22.54 / 64.00 / 33.33
	A	76.79 / 76.79 / 76.79	04.76 / 14.29 / 07.14	22.22 / 70.59 / 33.80
	F	**72.73 / 100.0 / 84.21**	04.76 / 75.00 / 08.96	**23.53 / 50.00 / 32.00**
BMI@ASU	W	52.24 / 97.22 / 67.96	20.24 / 53.12 / 29.31	29.58 / 43.75 / 35.29
	A	53.57 / 96.77 / 68.97	**09.52 / 22.22 / 13.33**	31.48 / 51.52 / 39.08
	F	45.45 / 100.0 / 62.50	23.81 / 65.22 / 34.88	23.53 / 26.67 / 25.00
UTurku	W	**76.12 / 91.07 / 82.93**	**21.43 / 51.43 / 30.25**	28.17 / 44.44 / 34.48
	A	**78.57 / 89.80 / 83.81**	09.52 / 18.18 / 12.50	31.48 / 54.84 / 40.00
	F	63.64 / 100.0 / 77.78	**25.40 / 66.67 / 36.78**	17.65 / 21.43 / 19.35

Table 8: Evaluation results of Site information for different event types in (A)bstracts

Team		Negation	Speculation	All
Kilicoglu09	A	*14.98 / 50.75 / 23.13*	*16.83 / 50.72 / 25.27*	*15.86 / 50.74 / 24.17*
UTurku	W	**22.87 / 48.85 / 31.15**	17.86 / 32.54 / 23.06	**20.30 / 39.67 / 26.86**
	A	**22.03 / 49.02 / 30.40**	19.23 / 38.46 / 25.64	**20.69 / 43.69 / 28.08**
	F	**25.76 / 48.28 / 33.59**	15.00 / 23.08 / 18.18	19.28 / 30.85 / 23.73
ConcordU	W	18.77 / 44.26 / 26.36	**21.10 / 38.46 / 27.25**	19.97 / 40.89 / 26.83
	A	18.06 / 46.59 / 26.03	**23.08 / 40.00 / 29.27**	20.46 / 42.79 / 27.68
	F	21.21 / 38.24 / 27.29	**17.00 / 34.69 / 22.82**	**18.67 / 36.14 / 24.63**

Table 9: Evaluation results of Task 3 in (W)hole data set, (A)bstracts only, and (F)ull papers only