Overview of BioNLP Shared Task 2011

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

Sampo Pyysalo
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
smp@is.s.u-tokyo.ac.jp

Tomoko Ohta
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
okap@is.s.u-tokyo.ac.jp

Robert Bossy
National Institute for Agricultural Research
78352 Jouy en Josas, Cedex
Robert.Bossy@jouy.inra.fr

Ngan Nguyen
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
nltngan@is.s.u-tokyo.ac.jp

Jun'ichi Tsujii
Microsoft Research Asia
5 Dan Ling Street, Haiian District, Beijing
jtsujii@microsoft.com

Abstract

The BioNLP Shared Task 2011, an information extraction task held over 6 months up to March 2011, met with community-wide participation, receiving 46 final submissions from 24 teams. Five main tasks and three supporting tasks were arranged, and their results show advances in the state of the art in fine-grained biomedical domain information extraction and demonstrate that extraction methods successfully generalize in various aspects.

1 Introduction

The BioNLP Shared Task (BioNLP-ST, hereafter) series represents a community-wide move toward fine-grained information extraction (IE), in particular biomolecular event extraction (Kim et al., 2009; Ananiadou et al., 2010). The series is complementary to BioCreative (Hirschman et al., 2007); while BioCreative emphasizes the short-term *applicability* of introduced IE methods for tasks such as database curation, BioNLP-ST places more emphasis on the *measurability* of the state-of-the-art and *traceability* of challenges in extraction through an approach more closely tied to text.

These goals were pursued in the first event, BioNLP-ST 2009 (Kim et al., 2009), through *high quality benchmark data* provided for system development and *detailed evaluation* performed to identify remaining problems hindering extraction performance. Also, as the complexity of the task was high and system development time limited, we encouraged *focus on fine-grained IE* by providing gold annotation for named entities as well as various supporting resources. BioNLP-ST 2009 attracted wide attention, with 24 teams submitting final results. The task setup and data since have served as the basis for numerous studies (Miwa et al., 2010b; Poon and Vanderwende, 2010; Vlachos, 2010; Miwa et al., 2010a; Björne et al., 2010).

As the second event of the series, BioNLP-ST 2011 preserves the general design and goals of the previous event, but adds a new focus on *variability* to address a limitation of BioNLP-ST 2009: the benchmark data sets were based on the Genia corpus (Kim et al., 2008), restricting the community-wide effort to resources developed by a single group for a small subdomain of molecular biology. BioNLP-ST 2011 is organized as a joint effort of several groups preparing various tasks and resources, in which variability is pursued in three primary directions: *text types*, *event types*, and *subject domains*. Consequently, *generalization* of fine grained bio-IE in these directions is emphasized as the main theme of the second event.

This paper summarizes the entire BioNLP-ST 2011, covering the relationships between tasks and similar broad issues. Each task is presented in detail in separate overview papers and extraction systems in papers by participants.

Proceedings of BioNLP Shared Task 2011 Workshop, pages 1–6,
Portland, Oregon, USA, 24 June, 2011.

2 Main tasks

BioNLP-ST 2011 includes four main tracks (with five tasks) representing fine-grained bio-IE.

2.1 Genia task (GE)

The GE task (Kim et al., 2011) preserves the task definition of BioNLP-ST 2009, arranged based on the Genia corpus (Kim et al., 2008). The data represents a focused domain of molecular biology: *transcription factors in human blood cells*. The purpose of the GE task is two-fold: to measure the progress of the community since the last event, and to evaluate generalization of the technology to full papers. For the second purpose, the provided data is composed of two collections: the *abstract collection*, identical to the BioNLP-ST 2009 data, and the new *full paper collection*. Progress on the task is measured through the unchanged task definition and the abstract collection, while generalization to full papers is measured on the full paper collection. In this way, the GE task is intended to connect the entire event to the previous one.

2.2 Epigenetics and post-translational modification task (EPI)

The EPI task (Ohta et al., 2011) focuses on IE for protein and DNA modifications, with particular emphasis on events of epigenetics interest. While the basic task setup and entity definitions follow those of the GE task, EPI extends on the extraction targets by defining 14 new event types relevant to task topics, including major protein modification types and their reverse reactions. For capturing the ways in which different entities participate in these events, the task extends the GE argument roles with two new roles specific to the domain, *Sidechain* and *Contextgene*. The task design and setup are oriented toward the needs of pathway extraction and curation for domain databases (Wu et al., 2003; Ongenaert et al., 2008) and are informed by previous studies on extraction of the target events (Ohta et al., 2010b; Ohta et al., 2010c).

2.3 Infectious diseases task (ID)

The ID task (Pyysalo et al., 2011a) concerns the extraction of events relevant to biomolecular mechanisms of infectious diseases from full-text publications. The task follows the basic design of BioNLP-ST 2009, and the ID entities and extraction targets are a superset of the GE ones. The task extends considerably on core entities, adding to PROTEIN four new entity types, including CHEMICAL and ORGANISM. The events extend on the GE definitions in allowing arguments of the new entity types as well as in introducing a new event category for high-level biological processes. The task was implemented in collaboration with domain experts and informed by prior studies on domain information extraction requirements (Pyysalo et al., 2010; Ananiadou et al., 2011), including the support of systems such as PATRIC (`http://patricbrc.org`).

2.4 Bacteria track

The bacteria track consists of two tasks, BB and BI.

2.4.1 Bacteria biotope task (BB)

The aim of the BB task (Bossy et al., 2011) is to extract the habitats of bacteria mentioned in textbook-level texts written for non-experts. The texts are Web pages about the state of the art knowledge about bacterial species. BB targets general relations, *Localization* and *PartOf*, and is challenging in that texts contain more coreferences than usual, habitat references are not necessarily named entities, and, unlike in other BioNLP-ST 2011 tasks, all entities need to be recognized by participants. BB is the first task to target phenotypic information and, as habitats are yet to be normalized by the field community, presents an opportunity for the BioNLP community to contribute to the standardization effort.

2.4.2 Bacteria interaction task (BI)

The BI task (Jourde et al., 2011) is devoted to the extraction of bacterial molecular interactions and regulations from publication abstracts. Mainly focused on gene transcriptional regulation in *Bacillus subtilis*, the BI corpus is provided to participants with rich semantic annotation derived from a recently proposed ontology (Manine et al., 2009) defining ten entity types such as gene, protein and derivatives as well as DNA sites/motifs. Their interactions are described through ten relation types. The BI corpus consists of the sentences of the LLL corpus (Nédellec, 2005), provided with manually checked linguistic annotations.

Task	Text	Focus	#
GE	abstracts, full papers	domain (HT)	9
EPI	abstracts	event types	15
ID	full papers	domain (TCS)	10
BB	web pages	domain (BB)	2
BI	abstracts	domain (BS)	10

Table 1: Characteristics of BioNLP-ST 2011 main tasks. '#': number of event/relation types targeted. Domains: HT = human transcription factors in blood cells, TCS = two-component systems, BB = bacteria biology, BS = Bacillus subtilis

2.5 Characteristics of main tasks

The main tasks are characterized in Table 1. From the text type perspective, BioNLP-ST 2011 generalizes from abstracts in 2009 to full papers (GE and ID) and web pages (BB). It also includes data collections for a variety of specific subject domains (GE, ID, BB an BI) and a task (EPI) whose scope is not defined through a domain but rather event types. In terms of the target event types, ID targets a superset of GE events and EPI extends on the representation for PHOSPHORYLATION events of GE. The two bacteria track tasks represent an independent perspective relatively far from other tasks in terms of their target information.

3 Supporting tasks

BioNLP-ST 2011 includes three supporting tasks designed to assist in primary the extraction tasks. Other supporting resources made available to participants are presented in (Stenetorp et al., 2011).

3.1 Protein coreference task (CO)

The CO task (Nguyen et al., 2011) concerns the recognition of coreferences to protein references. It is motivated from a finding from BioNLP-ST 2009 result analysis: coreference structures in biomedical text hinder the extraction results of fine-grained IE systems. While finding connections between event triggers and protein references is a major part of event extraction, it becomes much harder if one is replaced with a coreferencing expression. The CO task seeks to address this problem. The data sets for the task were produced based on MedCO annotation (Su et al., 2008) and other Genia resources (Tateisi et al., 2005; Kim et al., 2008).

Event	Date	Note
Sample Data	31 Aug. 2010	
Support. Tasks		
Train. Data	27 Sep. 2010	7 weeks for development
Test Data	15 Nov. 2010	4 days for submission
Submission	19 Nov. 2010	
Evaluation	22 Nov. 2010	
Main Tasks		
Train. Data	1 Dec. 2010	3 months for development
Test Data	1 Mar. 2011	9 days for submission
Submission	10 Mar. 2011	extended from 8 Mar.
Evaluation	11 Mar. 2011	extended from 10 Mar.

Table 2: Schedule of BioNLP-ST 2011

3.2 Entity relations task (REL)

The REL task (Pyysalo et al., 2011b) involves the recognition of two binary part-of relations between entities: PROTEIN-COMPONENT and SUBUNIT-COMPLEX. The task is motivated by specific challenges: the identification of the components of proteins in text is relevant e.g. to the recognition of *Site* arguments (cf. GE, EPI and ID tasks), and relations between proteins and their complexes relevant to any task involving them. REL setup is informed by recent semantic relation tasks (Hendrickx et al., 2010). The task data, consisting of new annotations for GE data, extends a previously introduced resource (Pyysalo et al., 2009; Ohta et al., 2010a).

3.3 Gene renaming task (REN)

The REN task (Jourde et al., 2011) objective is to extract renaming pairs of *Bacillus subtilis* gene/protein names from PubMed abstracts, motivated by discrepancies between nomenclature databases that interfere with search and complicate normalization. REN relations partially overlap several concepts: explicit renaming mentions, synonymy, and renaming deduced from biological proof. While the task is related to synonymy relation extraction (Yu and Agichtein, 2003), it has a novel definition of renaming, one name permanently replacing the other.

4 Schedule

Table 2 shows the task schedule, split into two phases to allow the use of supporting task results in addressing the main tasks. In recognition of their higher complexity, a longer development period was arranged for the main tasks (3 months vs 7 weeks).

Team	GE	EPI	ID	BB	BI	CO	REL	REN
UTurku	1	1	1	1	1	1	1	1
ConcordU	1	1	1			1	1	1
UMass	1	1	1					
Stanford	1	1	1					
FAUST	1	1	1					
MSR-NLP	1	1						
CCP-BTMG	1	1						
Others	8	0	2	2	0	4	2	1
SUM	15	7	7	3	1	6	4	3

Table 3: Final submissions to BioNLP-ST 2011 tasks.

5 Participation

BioNLP-ST 2011 received 46 submissions from 24 teams (Table 3). While seven teams participated in multiple tasks, only one team, UTurku, submitted final results to all the tasks. The remaining 17 teams participated in only single tasks. Disappointingly, only two teams (UTurku, and ConcordU) performed both supporting and main tasks, and neither used supporting task analyses for the main tasks.

6 Results

Detailed evaluation results and analyses are presented in individual task papers, but interesting observations can be obtained also by comparisons over the tasks. Table 4 summarizes best results for various criteria (Note that the results shown for e.g. GEa, GEf and GEp may be from different teams).

The community has made a significant improvement in the repeated GE task, with an over 10% reduction in error from '09 to GEa. Three teams achieved better results than M10, the best previously reported individual result on the '09 data. This indicates a beneficial role from focused efforts like BioNLP-ST. The GEf and ID results show that generalization to full papers is feasible, with very modest loss in performance compared to abstracts (GEa). The results for PHOSPHORYLATION events in GE and EPI are comparable (GEp vs EPIp), with the small drop for the EPI result, suggesting that the removal of the GE domain specificity does not compromise extraction performance. EPIc results indicate some challenges in generalization to similar event types, and EPIf suggest substantial further challenges in additional argument extraction. The complexity of ID is comparable to GE, also reflected to their final results, which further indicate successful generalization to a new subject domain as well as to new argument (entity) types. The BB task is in part comparable to GEl and involves a representation similar to REL, with lower results likely in part because BB requires entity recognition. The BI task is comparable to LLL Challenge, though BI involves more entity and event types. The BI result is 20 points above the LLL best result, indicating a substantial progress of the community in five years.

Task	Evaluation Results
BioNLP-ST 2009 ('09)	*46.73 / 58.48 / 51.95*
Miwa et al. (2010b) (M10)	*48.62 / 58.96 / 53.29*
LLL 2005 (LLL)	*53.00 / 55.60 / 54.30*
GE abstracts (GEa)	50.00 / 67.53 / 57.46
GE full texts (GEf)	47.84 / 59.76 / 53.14
GE PHOSPHORYLATION (GEp)	79.26 / 86.99 / 82.95
GE LOCALIZATION (GEl)	37.88 / 77.42 / 50.87
EPI full task (EPIf)	52.69 / 53.98 / 53.33
EPI core task (EPIc)	68.51 / 69.20 / 68.86
EPI PHOSPHORYLATION (EPIp)	86.15 / 74.67 / 80.00
ID full task (IDf)	48.03 / 65.97 / 55.59
ID core task (IDc)	50.62 / 66.06 / 57.32
BB	45.00 / 45.00 / 45.00
BB PartOf (BBp)	32.00 / 83.00 / 46.00
BI	71.00 / 85.00 / 77.00
CO	22.18 / 73.26 / 34.05
REL	50.10 / 68.00 / 57.70
REN	79.60 / 95.90 / 87.00

Table 4: Best results for various (sub)tasks (recall / precision / f-score (%)). GEl: task 2 without trigger detection.

7 Discussion and Conclusions

Meeting with wide participation from the community, BioNLP-ST 2011 produced a wealth of valuable resources for the advancement of fine-grained IE in biology and biomedicine, and demonstrated that event extraction methods can successfully generalize to new text types, event types, and domains. However, the goal to observe the capacity of supporting tasks to assist the main tasks was not met. The entire shared task period was very long, more than 6 months, and the complexity of the task was high, which could be an excessive burden for participants, limiting the application of novel resources. There have been ongoing efforts since BioNLP-ST 2009 to develop IE systems based on the task resources, and we hope to see continued efforts also following BioNLP-ST 2011, especially exploring the use of supporting task resources for main tasks.

References

Sophia Ananiadou, Sampo Pyysalo, Junichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*.

Sophia Ananiadou, Dan Sullivan, William Black, Gina-Anne Levow, Joseph J. Gillespie, Chunhong Mao, Sampo Pyysalo, BalaKrishna Kolluru, Junichi Tsujii, and Bruno Sobral. 2011. Named entity recognition for bacterial type IV secretion systems. *PLoS ONE*, 6(3):e14780.

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

Robert Bossy, Julien Jourde, Philippe Bessières, Marteen van de Guchte, and Claire Nédellec. 2011. BioNLP Shared Task 2011 - Bacteria Biotope. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 33–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Julien Jourde, Alain-Pierre Manine, Philippe Veber, Karën Fort, Robert Bossy, Erick Alphonse, and Philippe Bessières. 2011. BioNLP Shared Task 2011 - Bacteria Gene Interactions and Renaming. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from 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, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. 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.

A.P. Manine, E. Alphonse, and Bessières P. 2009. Learning ontological rules to extract multiple relations of genic interactions from text. *International Journal of Medical Informatics*, 78(12):e31–38.

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.

Nédellec. 2005. Learning Language in Logic – Genic Interaction Extraction Challenge. In *Proceedings of 4th Learning Language in Logic Workshop (LLL'05)*, pages 31–37.

Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Overview of the Protein Coreference 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.

Tomoko Ohta, Sampo Pyysalo, Jin-Dong Kim, and Jun'ichi Tsujii. 2010a. A re-evaluation of biomedical named entity-term relations. *Journal of Bioinformatics and Computational Biology (JBCB)*, 8(5):917–928.

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

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

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.

Maté Ongenaert, Leander Van Neste, Tim De Meyer, Gerben Menschaert, Sofie Bekaert, and Wim Van Criekinge. 2008. PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Research*, 36(suppl_1):D842–846.

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

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*, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Han-Cheol Cho, Dan Sullivan, Chunhong Mao, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2010. Towards event extraction from full texts on infectious diseases. In *Proceedings of BioNLP'10*, pages 132–140.

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

Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. 2011b. Overview of the Entity Relations (REL) supporting 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.

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

Jian Su, Xiaofeng Yang, Huaqing Hong, Yuka Tateisi, and Jun'ichi Tsujii. 2008. Coreference Resolution in Biomedical Texts: a Machine Learning Approach. In *Ontologies and Text Mining for Life Sciences'08*.

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.

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

Cathy H. Wu, Lai-Su L. Yeh, Hongzhan Huang, Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhangzhi Hu, Panagiotis Kourtesis, Robert S. Ledley, Baris E. Suzek, C.R. Vinayaka, Jian Zhang, and Winona C. Barker. 2003. The Protein Information Resource. *Nucleic Acids Research*, 31(1):345–347.

H. Yu and E. Agichtein. 2003. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(suppl 1):i340.