Overview of BioNLP Shared Task 2013

Claire Nédellec
MIG INRA UR1077
F-78352 Jouy-en-Josas cedex
claire.nedellec@jouy.inra.fr

Robert Bossy
MIG INRA UR1077
F-78352 Jouy-en-Josas cedex
robert.bossy@jouy.inra.fr

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

Jung-jae Kim
Nanyang Technological University
Singapore
jungjae.kim@ntu.edu.sg

Tomoko Ohta
National Centre for Text Mining and
School of Computer Science
University of Manchester
tomoko.ohta@manchester.ac.uk

Sampo Pyysalo
National Centre for Text Mining and
School of Computer Science
University of Manchester
sampo.pyysalo@gmail.com

Pierre Zweigenbaum
LIMSI-CNRS
F-91403 Orsay
pz@limsi.fr

Abstract

The BioNLP Shared Task 2013 is the third edition of the BioNLP Shared Task series that is a community-wide effort to address fine-grained, structural information extraction from biomedical literature. The BioNLP Shared Task 2013 was held from January to April 2013. Six main tasks were proposed. 38 final submissions were received, from 22 teams. The results show advances in the state of the art and demonstrate that extraction methods can be successfully generalized in various aspects.

1 Introduction

The BioNLP Shared Task (BioNLP-ST hereafter) series is a community-wide effort toward fine-grained biomolecular event extraction, from scientific documents. BioNLP-ST 2013 follows the general outline and goals of the previous tasks, namely BioNLP-ST'09 (Kim *et al.*, 2009) and BioNLP-ST'11 (Kim *et al.*, 2011). BioNLP-ST aims to provide a common framework for the comparative evaluation of information extraction (IE) methods in the biomedical domain. It shares this common goal with other tasks, namely BioCreative (Critical Assessment of Information Extraction in Biology) (Arighi *et al.*, 2011), DDIExtraction (Extraction of Drug-Drug Interactions from biomedical texts) (Segura-Bedmar *et al.*, 2011) and i2b2 (Informatics for Integrating Biology and the Bedside) Shared-Tasks (Sun *et al.*, 2013).

The biological questions addressed by the BioNLP-ST series belong to the molecular biology domain and its related fields. With the three editions, the series gathers several groups that prepared various tasks and resources, which represent diverse themes in biology. As the two previous editions, this one measures the progress accomplished by the community on complex text-bound event extraction. Compared to the other initiatives, the BioNLP-ST series proposes a linguistically motivated approach to event representation that enables the evaluation of the participating methods in a unifying computer science framework. Each edition has attracted an

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 1–7,
Sofia, Bulgaria, August 9 2013. ©2013 Association for Computational Linguistics

increasing number of teams with 22 teams submitting 38 final results this year. The task setup and the data serve as a basis for numerous further studies, released event extraction systems, and published datasets.

The first event in 2009 triggered active research in the community on a specific fine-grained IE task called Genia event extraction task. Expanding on this, the second BioNLP-ST was organized under the theme *Generalization*, where the participants introduced numerous systems that could be straightforwardly applied to different tasks. This time, the BioNLP-ST goes a step further and pursues the grand theme of *Knowledge base construction*. There were five tasks in 2011, and this year there are 6.

- [GE] Genia Event Extraction for NFkB knowledge base
- [CG] Cancer Genetics
- [PC] Pathway Curation
- [GRO] Corpus Annotation with Gene Regulation Ontology
- [GRN] Gene Regulation Network in Bacteria
- [BB] Bacteria Biotopes

The grand theme of *Knowledge base construction* is addressed in various ways: semantic web (GE, GRO), pathway (PC), molecular mechanism of cancer (CG), regulation network (GRN) and ontology population (GRO, BB).

In the biology domain, BioNLP-ST 2013 covers many new hot topics that reflect the evolving needs of biologists. BioNLP-ST 2013 broadens the scope of the text-mining application domains in biology by introducing new issues on cancer genetics and pathway curation. It also builds on the well-known previous datasets GENIA, LLL/BI and BB to propose tasks closer to the actual needs of biological data integration.

As in previous events, manually annotated data are provided to the participants for training, development and evaluation of the information extraction methods. According to their relevance for biological studies, the annotations are either bound to specific expressions in the text or represented as structured knowledge. Linguistic processing support was provided to the participants in the form of analyses of the dataset texts produced by state-of-the art tools.

This paper summarizes the BioNLP-ST 2013 organization, the task characteristics and their relationships. It gives synthetic figures on the participants and discusses the participating system advances.

2 Tasks

The BioNLP-ST'13 includes six tasks from four groups: DBCLS, NaCTeM, NTU and INRA. As opposed to the last edition, all tasks were main extraction tasks. There were no supporting tasks designed to assist the extraction tasks.

All tasks share the same event-based representation and file format, which is similar to the previous editions. This makes it easier to reuse the systems across tasks. Five kinds of annotation types are defined:

- `T`: text-bound annotation (entity/event trigger)
- `Equiv`: entity aliases
- `E`: event
- `M`: event modification
- `R`: relation
- `N`: normalization (external reference)

The normalization type has been introduced this year to represent the references to external resources such as dictionaries for GRN or ontologies for GRO and BB. The annotations are stand-off: the texts of the documents are kept separate from the annotations that refer to specific spans of texts through character offsets. More detail and examples can be found on the BioNLP-ST'13 web site.

2.1 Genia Event Extraction (GE)

Originally the design and implementation of the GE task was based on the Genia event corpus (Kim *et al.*, 2008) that represents domain knowledge of NFκB proteins. It was first organized as the sole task of the initial 2009 edition of BioNLP-ST (Kim *et al.*, 2009). While in 2009 the data sets consisted only of Medline abstracts, in its second edition in 2011 (Kim *et al.*, 2011b), it was extended to include full text articles to measure the generalization of the technology to full text papers. For its third edition this year, the GE task is organized with the goal of making it a more "real" task useful for knowledge base construction. The first design choice is to construct the data sets with recent full papers only, so that the extracted pieces of information could represent up-to-date knowledge of the domain. Second, the co-reference annotations are integrated into the event annotations, to encourage the use of these co-reference features in the solution of the event extraction.

2.2 Cancer Genetics (CG)

The CG task concerns the extraction of events relevant to cancer, covering molecular foundations, cellular, tissue, and organ-level effects, and organism-level outcomes. In addition to the domain, the task is novel in particular in extending event extraction to upper levels of biological organization. The CG task involves the extraction of 40 event types involving 18 types of entities, defined with respect to community-standard ontologies (Pyysalo *et al.*, 2011a; Ohta *et al.*, 2012). The newly introduced CG task corpus, prepared as an extension of a previously introduced corpus of 250 abstracts (Pyysalo *et al.*, 2012), consists of 600 PubMed abstracts annotated for over 17,000 events.

2.3 Pathway Curation (PC)

The PC task focuses on the automatic extraction of biomolecular reactions from text with the aim of supporting the development, evaluation and maintenance of biomolecular pathway models. The PC task setting and its document selection protocol account for both signaling and metabolic pathways. The 23 event types, including chemical modifications (Pyysalo *et al.*, 2011b), are defined primarily with respect to the Systems Biology Ontology (SBO) (Ohta *et al.*, 2011b; Ohta *et al.*, 2011c), involving 4 SBO entity types.

The PC task corpus was newly annotated for the task and consists of 525 PubMed abstracts, chosen for the relevance to specific pathway reactions selected from SBML models registered in BioModels and PANTHER DB repositories (Mi and Thomas, 2009). The corpus was manually annotated for over 12,000 events on top of close to 16,000 entities.

2.4 Gene Regulation Ontology (GRO)

The GRO task aims to populate the Gene Regulation Ontology (GRO) (Beisswanger *et al.*, 2008) with events and relations identified from text. The large size and the complex semantic representation of the underlying ontology are the main challenges of the task. Those issues, to a greater extent, should be addressed to support full-fledged semantic search over the biomedical literature, which is the ultimate goal of this work.

The corpus consists of 300 MEDLINE abstracts, prepared as an extension of (Kim *et al.*, 2011c). The analysis of the inter-annotator agreement between the two annotators shows Kappa values of 43%-56%, which might indicate the difficulty of the task.

2.5 Gene Regulation Network in Bacteria (GRN)

The Gene Regulation Network task consists of the extraction of the regulatory network of a set of genes involved in the sporulation phenomenon of the model organism *Bacillus subtilis*. Participant system predictions are evaluated with respect to the target regulation network, rather than the text-bound relations. The aim is to assess the IE methods with regards to the needs of systems biology and predictive biology studies.

The GRN corpus is a set of sentences from PubMed abstracts that extends the BioNLP-ST 2011 BI (Jourde *et al.,* 2011) and LLL (Nedellec, 2005) corpora. The additional sentences cover a wider range of publication dates and complement the regulation network of the sporulation phenomenon. It has been thoroughly annotated with different levels of biological abstraction: entities, biochemical events, genic interactions and the corresponding regulation network.

The network prediction submissions have been evaluated against the reference network using an original metric, the Slot Error Rate (Makhoul *et al.*, 1999) that is more adapted to graph comparison than the usual Recall, Precision and F-score measures.

2.6 Bacteria Biotopes (BB)

The Bacteria Biotope (BB) task concerns the extraction of locations in which bacteria live and the categorization of these habitats with concepts from OntoBiotope,[1] a large ontology of 1,700 concepts and 2,000 synonyms. The association between bacteria and their habitats is essential information for environmental biology studies, metagenomics and phylogeny.

In the previous edition of the BB task, participants had to recognize bacteria and habitat entities, to categorize habitat entities among eight broad types and to extract localization relations between bacteria and their habitats (Bossy *et al.*, 2011). The BioNLP-ST 2013 edition has been split into 3 sub-tasks in order to better assess the performance of the predictive systems for each step. The novelty of this task is mainly the more comprehensive and fine-grained categorization. It addresses the critical problem of habitat normalization necessary for the

[1] http://bibliome.jouy.inra.fr/MEM-OntoBiotope

automatic exploitation of bacteria-habitat databases.

2.7 Task characteristics

Task features are given in Table 1. Three different types of text were considered: the abstracts of scientific papers taken from PubMed (CG, PC, GRO and GRN), full-text scientific papers (GE) and scientific web pages (BB).

Task	Documents	# types	# events
GE	34 Full papers	2	13
CG	600 Abstracts	18	40
PC	525 Abstracts	4	23
GRO	300 Abstracts	174	126
GRN	201 Abstracts	6	12
BB	124 Web pages	563	2

Table 1. Characteristics of the BioNLP-ST 2013 tasks.

The number of relations or events targeted greatly varies with the tasks as shown in column 3. The high number of types and events reflect the increasing complexity of the biological knowledge to be extracted. The grand theme of *Knowledge base construction* in this edition has been translated into rich knowledge representations with the goal of integrating textual data with data from sources other than text. These figures illustrate the shared ambition of the organizers to promote fine-grained information extraction together with an increasing biological plausibility. Beyond gene and protein interactions, they include many complex biological phenomena and environmental factors.

3 BioNLP-ST'13 organization

BioNLP-ST'13 was split in three main periods. During thirteen weeks from mid-January to the first week of April, the participants prepared their systems with the training data. Supporting resources were delivered to participants during this period. Supporting resources were provided by the organizers and by three external providers after a public call for contribution. They range from tokenizers to entity detection tools, mostly focusing on syntactic parsing (Enju (Miyao and Tsujii, 2008), Stanford (Klein and Manning, 2002), McCCJ (Charniak and Johnson, 2005)). The test data were made available for 10 days before the participants had to submit their final results using on-line services. The evaluation results were communicated shortly after and published on the ST site. The descriptions of the tasks and representative sample data have been available since October 2012 so that the participants could become acquainted with the task goals and data formats in advance. Table 2 shows the task schedule.

Date	Event
23 Oct. 2012	Release of sample data sets
17 Jan 2013	Release of the training data sets
06 Apr. 2013	Release of the test data sets
16 Apr. 2013	Result submission
17 Apr. 2013	Notification of the evaluation results

Table 2: Schedule of BioNLP-ST 2013.

The BioNLP-ST'13 web site and a dedicated mailing-list have kept the participant informed about the whole process.

4 Participation

	GE 1-2-3			CG	PC	GRO	GRN	BB 1 - 2-3		
EVEX	•	•	•				•			
TEES-2.1	•	•	•	•	•	•	•		•	•
BioSEM	•									
NCBI	•									
DlutNLP	•									
HDS 4NLP	•									
NICTA	•		•							
USheff	•									
UZH	•									
HCMUS	•									
NaCTeM				•	•					
NCBI				•						
RelAgent				•						
UET-NII				•						
ISI				•						
OSEE						•				
U. of Ljubljana							•			
K.U. Leuven							•			
IRISA-TexMex							•	•	•	
Boun								•	•	
LIPN								•		
LIMSI								•	•	•

Table 3: Participating teams per task.

BioNLP-ST 2013 received 38 submissions from 22 teams (Table 3). One third, or seven teams, participated in multiple tasks. Only one team, UTurku, submitted final results with TEES-2.1 to all the tasks except one – entity categorization. This broad participation resulted from the growing capability of the systems to be applied to various tasks without manual tuning. The remaining 15 teams participated in one single task.

5 Results

Table 4 summarizes the best results and the participating systems for each task and sub-task. They are all measured using F-scores, except when it is not relevant, in which case SER is used instead. It is noticeable that the TEES-2.1 system that participated in 9 of the 10 tasks and sub-tasks achieved the best result in 6 cases. Most of the participating systems applied a combination of machine learning algorithms and linguistic features, mainly syntactic parses, with some noticeable exceptions.

Tasks	Evaluation results
GE *Core event extraction*	TEES-2.1, EVEX, BioSEM: 0.51
GE 2 *Event enrichment*	TEES2.1: 0.32
GE 3 *Negation/Speculation*	TEES-2.1, EVEX: 0.25
CG	TEES-2.1: 0.55
PC	NaCTeM: 0.53
GRO	TEES-2.1: 0.22 (events), 0.63 (relations)
GRN	U. of Ljubljana: 0.73 (SER)
BB 1 *Entity detection and categorization*	IRISA: 0.46 (SER)
BB 2 *Relation extraction*	IRISA: 0.40
BB 3 *Full event extraction*	TEES-2.1: 0.14

Table 4. Best results and team per task (F-score, except when SER).

Twelve teams submitted final results to the GE task. The performance of highly ranked systems shows that the event extraction technology is applicable to the most recent full papers without drop of performance.

Six teams submitted final results to the CG task. The highest-performing systems achieved results comparable to those for established molecular level extraction tasks (Kim *et al.*, 2011). The results indicate that event extraction methods generalize well to higher levels of biological organization and are applicable to the construction of knowledge bases on cancer.

Two teams successfully completed the PC task, and the highest F-score reached 52.8%, indicating that event extraction is a promising approach to support pathway curation efforts.

The GRN task attracted five participants. The best SER score was 0,73 (the higher, the worse), which shows their capability of designing regulatory network, but handling modalities remains an issue.

Five teams participated to the 3 BB subtasks with 10 final submissions. Not surprisingly, the systems achieved better results in relation extraction than habitat categorization, which remains a major challenge in IE.

One team participated in the GRO task, and their results were compared with those of a preliminary system prepared by the task organizers. An analysis of the evaluation results leads us to study issues such as the need to consider the ontology structure and the need for semantic analysis, which are not seriously dealt with by current approaches to event extraction.

6 Organization of the workshop

The BioNLP Shared Task 2013 (BioNLP-ST) workshop was organized as part of the ACL BioNLP 2013 workshop. After submission of their system results, participants were invited to submit a paper on their systems to the workshop. Task organizers were also invited to present overviews of each task, with analyses of the participant system features and results. The workshop was held in August 2013 in Sofia (Bulgaria). It included overview presentations on tasks, as well as oral and poster presentations by Shared Task participants.

7 Discussion and Conclusion

This year, the tasks has significantly gained in complexity to face the increasing need for Systems Biology knowledge from various textual sources. The high level of participation and the quality of the results show that the maturity of the field is such that it can meet this challenge. The innovative and various solutions applied this year will without doubt be extended in the future. As for previous editions of BioNLP-ST, all tasks maintain an online evaluation service that is publicly available. This on-going challenge will contribute to the assessment of the evolving information extraction field in the biomedical domain.

References

Auhors. 2013. Title. In *Proceedings of the BioNLP 2013 Workshop Companion Volume for Shared Task,* Sofia, Bulgaria. Association for Computational Linguistics.

Arighi, C., Lu, Z., Krallinger, M., Cohen, K., Wilbur, W., Valencia, A., Hirschman, L. and Wu, C. 2011. Overview of the BioCreative III Workshop. *BMC Bioinformatics*, 12, S1.

E Beisswanger, V Lee, JJ Kim, D Rebholz-Schuhmann, A Splendiani, O Dameron, S Schulz, U Hahn. Gene Regulation Ontology (GRO): Design principles and use cases. Studies in Health Technology and Informatics, 136:9-14, 2008.

BioNLP-ST'13 web site: https://2013.bionlp-st.org

Robert Bossy, Julien Jourde, Philippe Bessières, Maarten van de Guchte, Claire Nédellec. 2011. BioNLP shared Tasks 2011 - Bacteria Biotope. In *Proceedings of BioNLP 2011 Workshop*, pages 65-73. Association for Computational Linguistics, Portland, USA, 2011.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics.

Julien Jourde, Alain-Pierre Manine, Philippe Veber, Karen Fort, Robert Bossy, Erick Alphonse, Philippe Bessières. 2011. BioNLP Shared Task 2011 - Bacteria Gene Interactions and Renaming. In *Proceedings of BioNLP 2011 Workshop*, pages 65-73. Association for Computational Linguistics, Portland.

Jin-Dong Kim, Tomoko Ohta and Jun'ichi Tsujii, 2008, Corpus annotation for mining biomedical events from literature, 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 the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1-9.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP 2011 Workshop*, pages 1-6. 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.

Jung-Jae Kim, Xu Han and Watson Wei Khong Chua. 2011c. Annotation of biomedical text with Gene Regulation Ontology: Towards Semantic Web for biomedical literature. Proceedings of LBM 2011, pp. 63-70.

Dan Klein and Christopher D Manning. 2002. Fast ex act inference with a factored model for natural language parsing. *Advances in neural information processing systems*, 15(2003):3–10.

John Makhoul, Francis Kubala, Richard Schwartz and Ralph Weischedel. 1999. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop,* Herndon, VA, February.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80.

Huaiyu Mi and Paul Thomas. 2009. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. In *Protein Networks and Pathway Analysis*, pages 123–140. Springer.

Claire Nédellec. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge. In *Proceedings of the Learning Language in Logic (LLL05) workshop* joint to ICML'05. Cussens J. and Nedellec C. (eds). Bonn, August.

Tomoko Ohta, Sampo Pyysalo, Sophia Ananiadou, and Jun'ichi Tsujii. 2011b. Pathway curation support as an information extraction task. Proceedings of *LBM* 2011.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011c. From pathways to biomolecular events: opportunities and challenges. In *Proceedings of BioNLP 2011 Workshop*, pages 105–113. Association for Computational Linguistics.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of DSSD 2012*, pages 27–36.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575-i581.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, and Jun'ichi Tsujii. 2011b. Towards exhaustive event extraction for protein modifications. In *Proceedings of the BioNLP 2011 Workshop*, pp.114-123, Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Jun'ichi Tsujii and Sophia Ananiadou. 2011a. Anatomical Entity Recognition with Open Biomedical Ontologies. In *proceedings of LBM 2011*.

Isabel Segura-Bedmar, Paloma Martinez, and Daniel Sanchez-Cisneros. 2011. The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. In *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011, SEPLN 2011 satellite workshop*. Huelva, Spain, September 7.

Weiyi Sun, Anna Rumshisky, Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc*.