Overview of the Entity Relations (REL) supporting task of BioNLP Shared Task 2011

Sampo Pyysalo[*] **Tomoko Ohta**[*] **Jun'ichi Tsujii**[†]
[*]Department of Computer Science, University of Tokyo, Tokyo, Japan
[†]Microsoft Research Asia, Beijing, China
{smp,okap}@is.s.u-tokyo.ac.jp, jtsujii@microsoft.com

Abstract

This paper presents the Entity Relations (REL) task, a supporting task of the BioNLP Shared Task 2011. The task concerns the extraction of two types of part-of relations between a gene/protein and an associated entity. Four teams submitted final results for the REL task, with the highest-performing system achieving 57.7% F-score. While experiments suggest use of the data can help improve event extraction performance, the task data has so far received only limited use in support of event extraction. The REL task continues as an open challenge, with all resources available from the shared task website.

1 Introduction

The BioNLP Shared Task 2011 (BioNLP ST'11) (Kim et al., 2011a), the follow-up event to the BioNLP'09 Shared Task (Kim et al., 2009), was organized from August 2010 (sample data release) to March 2011. The shared task was divided into two stages, with supporting tasks carried out before the main tasks. The motivation for this task setup drew in part from analysis of the results of the previous shared task, which suggested that events that involve coreference or entity relations represent particular challenges for extraction. To help address these challenges and encourage modular extraction approaches, increased sharing of successful solutions, and an efficient division of labor, the two were separated into independent supporting tasks on Coreference (CO) (Nguyen et al., 2011) and Entity Relations in BioNLP ST'11. This paper presents the Entity Relations (REL) supporting task.

2 Task Setting

In the design of the REL task, we followed the general policy of the shared task in assuming named entity recognition (NER) as a given starting point: participants were provided with manually annotated gold standard annotations identifying gene/protein names in all of the training, development, and final test data. By limiting effects due to NER performance, the task remains more specifically focused on the key challenge studied.

Following the results and analysis from previous studies (Pyysalo et al., 2009; Ohta et al., 2010), we chose to limit the task specifically to relations involving a gene/protein named entity (NE) and one other entity. Fixing one entity involved in each relation to an NE helps assure that the relations are "anchored" to real-world entities, and the specific choice of the gene/protein NE class further provides a category with several existing systems and substantial ongoing efforts addressing the identification of those referents through named entity recognition and normalization (Leaman and Gonzalez, 2008; Hakenberg et al., 2008; Krallinger et al., 2008; Morgan et al., 2008; Wermter et al., 2009). The recognition of biologically relevant associations of gene/protein NEs is a key focus of the main event extraction tasks of the shared task. By contrast, in the REL task setting, only one participant in each binary relation is a gene/protein NE, while the other can be either a non-name reference such as *promoter* or the name of an entity not of the gene/protein type (e.g. a complex).[1] Motivated in part by the relatively limited number of existing methods for the detec-

[1] Pronominal references are excluded from annotation scope.

Proceedings of BioNLP Shared Task 2011 Workshop, pages 83–88,
Portland, Oregon, USA, 24 June, 2011. ©2011 Association for Computational Linguistics

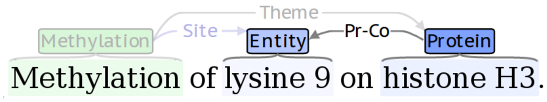

Figure 1: Simple REL annotation example showing a PROTEIN-COMPONENT (PR-CO) relation between "histone H3" and "lysine 9". An associated METHYLATION event and its arguments (shaded, not part of the REL task targets) shown for context.

Item	Training	Devel	Test
Abstract	800	150	260
Word	176,146	33,827	57,256
Protein	9,297	2,080	3,589
Relation	1,857	480	497
PROTEIN-COMPONENT	1,302	314	334
SUBUNIT-COMPLEX	555	166	163

Table 1: REL dataset statistics.

tion of such entity references, their detection is included in the task: participants must recognize these secondary entities in addition to extracting the relations they participate in. To limit the demands of this NER-type task, these entities are not assigned specific types but rather the generic type ENTITY, and exact matching of their boundaries is not required (see Section 4).

The general task setting encompasses a rich set of potential relation extraction targets. For the task, we aimed to select relations that minimize overlap between the targets of other tasks while maintaining relevance as a supporting goal. As the main tasks primarily target events ("things that happen") involving change in entities, we chose to focus in the REL task on what we have previously termed "static relations" (Pyysalo et al., 2009), that is, relations such as part-of that hold between entities without necessary implication of causality or change. A previous study by Van Landeghem et al. (2010) indicated that this class of relations may benefit event extraction. We based our choice of specific target relation on previous studies of entity relations domain texts (Pyysalo et al., 2009; Ohta et al., 2010), which indicated that part-whole relations are by far the most frequent class of relevant relations for the task setting and proposed a classification of these relations for biomedical entities. We further found that – in terms of the taxonomy of Winston et al. (1987) – object-component and collection-member relations account for the the great majority of part-of relations relevant to the domain. For REL, we chose to omit collection-member relations in part to minimize overlap with the targets of the coreference task. Instead, we focused on two specific types of object-component relations, that holding between a gene or protein and its part (domain, regions, promoters, amino acids, etc.) and that between a protein and a complex that it is a subunit of. Following the biological motivation and the general practice in the shared task to term genes and gene products PROTEIN for simplicity, we named these two relations PROTEIN-COMPONENT and SUBUNIT-COMPLEX. Figure 1 shows an illustration of a simple relation with an associated event (not part of REL). Events with *Site* arguments such as that shown in the figure are targeted in the GE, EPI, and ID tasks (Kim et al., 2011b; Ohta et al., 2011; Pyysalo et al., 2011) that REL is intended to support.

3 Data

The task dataset consists of new annotations for the GENIA corpus (Kim et al., 2008), building on the existing biomedical term annotation (Ohta et al., 2002), the gene and gene product name annotation (Ohta et al., 2009) and the syntactic annotation (Tateisi et al., 2005) of the corpus. The general features of the annotation are presented by Pyysalo et al. (2009), describing a previous release of a subset of the data. The REL task annotation effort extended the coverage of the previously released annotation to all relations of the targeted types stated within sentence scope in the GENIA corpus.

For compatibility with the BioNLP ST'09 and its repeat as the GE task in 2011 (Kim et al., 2011b), the REL task training/development/test set division of the GENIA corpus abstracts matches that of the BioNLP ST'09 data. The statistics of the corpus are presented in Table 1. We note that both in terms of training examples and the data available in the given development set, the number of examples of the PROTEIN-COMPONENT relation is more than twice that for SUBUNIT-COMPLEX. Thus, at least for methods based on machine learning, we might generally expect to find higher extraction performance for the former relation.

			NLP		Extraction		Other resources	
Rank	Team	Org	Word	Parse	Entities	Relations	Corpora	Other
1	UTurku	1BI	Porter	McCCJ + SD	SVM	SVM	-	-
2	VIBGhent	1NLP, 1ML, 1BI	Porter	McCCJ + SD	SVM	SVM	GENIA, PubMed	word similarities
3	ConcordU	2NLP	-	McCCJ + SD	Dict	Rules	-	-
3	HCMUS	6L	OpenNLP	OpenNLP	Dict	Rules	-	-

Table 2: Participants and summary of system descriptions. Abbreviations: BI=Bioinformatician, NLP=Natural Language Processing researcher, ML=Machine Learning researcher, L=Linguist, Porter=Porter stemmer, McCCJ=McClosky-Charniak-Johnson parser, SD=Stanford Dependency conversion, Dict=Dictionary

	UTurku	VIBGhent	ConcordU	HCMUS
Protein-Component	50.90 / 68.57 / **58.43**	47.31 / 36.53 / 41.23	23.35 / 52.05 / 32.24	20.96 / 21.63 / 21.29
Subunit-Complex	48.47 / 66.95 / **56.23**	47.85 / 38.12 / 42.43	26.38 / 39.81 / 31.73	4.91 / 66.67 / 9.14
Total	50.10 / 68.04 / **57.71**	47.48 / 37.04 / 41.62	24.35 / 46.85 / 32.04	15.69 / 23.26 / 18.74

Table 3: Primary evaluation results for the REL task. Results given as recall / precision / F-score.

4 Evaluation

The evaluation of the REL task is relation-based and uses the standard precision/recall/F_1-score metrics. Similarly to the BioNLP'09 ST and most of the 2011 main tasks, the REL task relaxes the equality criteria for matching text-bound annotations: for a submission entity to match an entity in the gold reference annotation, it is sufficient that the span of the submitted entity (i.e. its start and end positions in text) is entirely contained within the span of the gold annotation. This corresponds largely to the *approximate span matching* criterion of the 2009 task (Kim et al., 2009), although the REL criterion is slightly stricter in not involving testing against an extension of the gold entity span. Relation matching is exact: for a submitted relation to match a gold one, both its type and the related entities must match.

5 Results

5.1 Participation

Table 2 summarizes the participating groups and approaches. We find a remarkable number of similarities between the approaches of the systems, with all four utilizing full parsing and a dependency representation of the syntactic analysis, and the three highest-ranking further specifically the phrase structure parser of Charniak and Johnson (2005) with the biomedical domain model of McClosky (2009), converted into Stanford Dependency form using the Stanford tools (de Marneffe et al., 2006). These specific choices may perhaps be influenced by the success of systems building on them in the 2009 shared task (e.g. Björne et al. (2009)). While UTurku (Björne and Salakoski, 2011) and VIBGhent (Van Landeghem et al., 2011) further agree in the choice of Support Vector Machines for the recognition of entities and the extraction of relations, ConcordU (Kilicoglu and Bergler, 2011) and HCMUS (Le Minh et al., 2011) pursue approaches building on dictionary- and rule-based extraction. Only the VIBGhent system makes use of resources external to those provided for the task, extracting specific semantic entity types from the GENIA corpus as well as inducing word similarities from a large unannotated corpus of PubMed abstracts.

5.2 Evaluation results

Table 3 shows the results of the REL task. We find that the four systems diverge substantially in terms of overall performance, with all pairs of systems of neighboring ranks showing differences approaching or exceeding 10% points in F-score. While three of the systems notably favor precision over recall, VIBGhent shows a decided preference for recall, suggesting a different approach from UTurku in design details despite the substantial similarities in overall system architecture. The highest-performing system, UTurku, shows an F-score in the general range of state-of-the-art results in the main event extraction task, which could be taken as an indication that the reliability of REL task analyses created with presently available methods may not be high enough for direct use as a building block for the main tasks. However, the emphasis of the UTurku system on precision is encouraging for such applications: nearly 70% of the entity-relation pairs that the system predicts are correct. The two top-ranking systems show similar precision and recall results for the two relation types. The submission of HCMUS shows a decided advantage for PROTEIN-COMPONENT relation extraction as tentatively predicted from the relative numbers of training examples (Section 3 and Table 1), but their rule-based approach suggests training data size is likely not the decisive factor. While the limited amount of data available prevents strong conclusions from being drawn, overall the lack of correlation between training data size and extraction performance suggests that performance may not be primarily limited by the size of the available training data.

6 Discussion

The REL task was explicitly cast in a support role for the main event extraction tasks, and REL participants were encouraged to make their predictions of the task extraction targets for the various main task datasets available to main task participants. The UTurku team responded to this call for supporting analyses, running their top-ranking REL task system on all main task datasets and making its output available as a supporting resource (Stenetorp et al., 2011). In the main tasks, we are so far aware of one application of this data: the BMI@ASU team (Emadzadeh et al., 2011) applied the UTurku REL predictions as part of their GE task system for resolving the *Site* arguments in events such as BINDING and PHOSPHORYLATION (see Figure 1). While more extensive use of the data would have been desirable, we find this application of the REL analyses very appropriate to our general design for the role of the supporting and main tasks and hope to see other groups pursue similar possibilities in future work.

7 Conclusions

We have presented the preparation, resources, results and analysis of the Entity Relations (REL) task, a supporting task of the BioNLP Shared Task 2011 involving the recognition of two specific types of part-of relations between genes/proteins and associated entities. The task was run in a separate early stage in the overall shared task schedule to allow participants to make use of methods and analyses for the task as part of their main task submissions.

Of four teams submitting finals results, the highest-performing system, UTurku, achieved a precision of 68% at 50% recall (58% F-score), a promising level of performance given the relative novelty of the specific extraction targets and the short development period. Nevertheless, challenges remain for achieving a level of reliability that would allow event extraction systems to confidently build on REL analyses to address the main information extraction tasks. The REL task submissions, representing four independent perspectives into the task, are a valuable resource for further study of both the original task data as well as the relative strengths and weaknesses of the participating systems. In future work, we will analyse this data in detail to better understand the challenges of the task and effective approached for addressing them.

The UTurku team responded to a call for supporting analyses by providing predictions from their REL system for all BioNLP Shared Task main task datasets. These analyses were adopted by at least one main task participant as part of their system, and we expect that this resource will continue to serve to facilitate the study of the position of part-of relations in domain event extraction. The REL task will continue as an open shared challenge, with all task data, evaluation software, and analysis tools available to all interested parties from `http://sites.google.com/site/bionlpst/`.

Acknowledgments

We would like to thank the UTurku team for their generosity with their time and tools in providing REL task analyses for all the BioNLP Shared Task 2011 main task datasets. This work was supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

References

Jari Björne 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.

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.

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.

J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez. 2008. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24(16):i126.

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, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(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.

M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. 2008. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome biology*, 9(Suppl 2):S1.

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.

R. Leaman and G. Gonzalez. 2008. Banner: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663.

David McClosky. 2009. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Department of Computer Science, Brown University.

A.A. Morgan, Z. Lu, X. Wang, A.M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, et al. 2008. Overview of BioCreative II gene normalization. *Genome biology*, 9(Suppl 2):S3.

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, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. 2002. GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference (HLT'02)*, pages 73–77.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. 2009. Incorporating GENETAG-style annotation to GENIA corpus. In *Proceedings of BioNLP'09*, pages 106–107.

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

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.

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, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. 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.

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.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotation for the GENIA corpus. In *Proceedings of IJCNLP'05*, pages 222–227.

Sofie Van Landeghem, Sampo Pyysalo, Tomoko Ohta, and Yves Van de Peer. 2010. Integration of static relations to enhance event extraction from text. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 144–152.

Sofie Van Landeghem, Thomas Abeel, Bernard De Baets, and Yves Van de Peer. 2011. Detecting entity relations as a supporting task for bio-molecular event extraction. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

J. Wermter, K. Tomanek, and U. Hahn. 2009. High-performance gene name normalization with GeNo. *Bioinformatics*, 25(6):815.

Morton E. Winston, Roger Chaffin, and Douglas Herrmann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11.