New Resources and Perspectives for Biomedical Event Extraction

Sampo Pyysalo[1], **Pontus Stenetorp**[2], **Tomoko Ohta**[1], **Jin-Dong Kim**[3] **and Sophia Ananiadou**[1]

[1]National Centre for Text Mining and University of Manchester,
Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, UK
[2]Tokyo University, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
[3]Database Center for Life Science, 2-11-16 Yayoi, Bunkyo-ku, Tokyo, Japan

Abstract

Event extraction is a major focus of recent work in biomedical information extraction. Despite substantial advances, many challenges still remain for reliable automatic extraction of events from text. We introduce a new biomedical event extraction resource consisting of analyses automatically created by systems participating in the recent BioNLP Shared Task (ST) 2011. In providing for the first time the outputs of a broad set of state-of-the-art event extraction systems, this resource opens many new opportunities for studying aspects of event extraction, from the identification of common errors to the study of effective approaches to combining the strengths of systems. We demonstrate these opportunities through a multi-system analysis on three BioNLP ST 2011 main tasks, focusing on events that none of the systems can successfully extract. We further argue for new perspectives to the performance evaluation of domain event extraction systems, considering a document-level, "off-the-page" representation and evaluation to complement the mention-level evaluations pursued in most recent work.

1 Introduction

Biomedical information extraction efforts are increasingly focusing on event extraction using structured representations that allow associations of arbitrary numbers of participants in specific roles (e.g. *Theme*, *Cause*) to be captured (Ananiadou et al., 2010). Domain event extraction has been advanced in particular by the BioNLP Shared Task (ST) events (Kim et al., 2011a; Kim et al., 2011b), which have introduced common task settings, datasets, and evaluation criteria for event extraction. Participants in these shared tasks have introduced dozens of systems for event extraction, and the resulting methods have been applied to automatically analyse the entire available domain literature (Björne et al., 2010) and applied in support of applications such as semantic literature search (Ohta et al., 2010; Van Landeghem et al., 2011b) and pathway curation support (Kemper et al., 2010).

It is possible to assess recent advances in event extraction through results for a task considered both in the BioNLP ST 2009 and 2011. By the primary evaluation criteria, the highest performance achieved in the 2009 task was 51.95% F-score, and a 57.46% F-score was reached in the comparable 2011 task (Kim et al., 2011b). These results demonstrate significant advances in event extraction methods, but also indicate that the task continues to hold substantial challenges. This has led to a call from task participants for further analysis of the data and results, accompanied by a proposal to release analyses from individual systems to facilitate such analysis (Quirk et al., 2011).

In this study, we explore new perspectives into the analyses and performance of event extraction methods. We build primarily on a new resource compiled with the support of the majority of groups participating in the BioNLP ST 2011, consisting of analyses from systems for the three main tasks sharing the text-bound event representation. We demonstrate the use of this resource through an evaluation focusing on events that cannot be extracted even by the union of combined systems, identifying particular remaining challenges for event extraction. We further propose and evaluate an alternate, document-level perspective to event extraction, demonstrating that when only unique events are considered for

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 100–108,
Montréal, Canada, June 8, 2012.

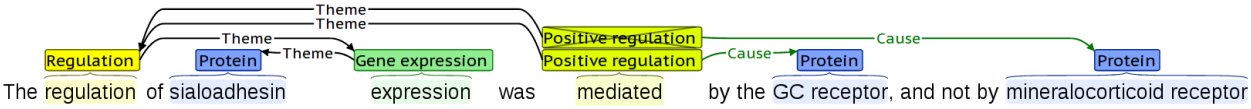


Figure 1: Example event annotations. The "crossed-out" event type identifies an event marked as negated. Event illustrations created using the STAV visualization tool (Stenetorp et al., 2011).

each document, the measured performance and even ranking of systems participating in the shared task is notably altered.

2 Background

In this work, we focus on the definition of the event extraction task first introduced in the BioNLP Shared Task 2009.[1] The task targets the extraction of *events*, represented as n-ary associations of participants (entities or other events), each marked as playing a specific *role* such as *Theme* or *Cause* in the event. Each event is assigned a *type* such as BINDING or PHOSPHORYLATION from a fixed, task-specific set. Events are further typically associated with specific *trigger* expressions that state their occurrence in text. As physical entities such as proteins are also identified in the setting with specific spans referring to the real-world entities in text, the overall task is "text-bound" in the sense of requiring not only the extraction of targeted statements from text, but also the identification of specific regions of text expressing each piece of extracted information. Events can further be marked with *modifiers* identifying additional features such as being explicitly negated or stated in a speculative context. Figure 1 shows an illustration of event annotations.

This BioNLP ST 2009 formulation of the event extraction task was followed also in three 2011 main tasks: the GE (Kim et al., 2011c), ID (Pyysalo et al., 2011a) and EPI (Ohta et al., 2011) tasks. A variant of this representation that omits event triggers was applied in the BioNLP ST 2011 bacteria track (Bossy et al., 2011), and simpler, binary relation-type representations were applied in three supporting tasks (Nguyen et al., 2011; Pyysalo et al., 2011b; Jourde et al., 2011). Due to the challenges of consistent evaluation and processing for tasks involving different representations, we focus in this work specifically on the three 2011 main tasks sharing a uniform representation: GE, ID and EPI.

[1] While far from the only formulation proposed in the literature, this specific task setting is the most frequently considered and arguably a *de facto* standard for domain event extraction.

3 New Resources for Event Extraction

In this section, we present the new collection of automatically created event analyses and demonstrate one use of the data through an evaluation of events that no system could successfully extract.

3.1 Data Compilation

Following the BioNLP ST 2011, the MSR-NLP group called for the release of outputs from various participating systems (Quirk et al., 2011) and made analyses of their system available.[2] Despite the obvious benefits of the availability of these resources, we are not aware of other groups following this example prior to the time of this publication.

To create the combined resource, we approached each group that participated in the three targeted BioNLP ST 2011 main tasks to ask for their support to the creation of a dataset including analyses from their event extraction systems. This suggestion met with the support of all but a few groups that were approached.[3] The groups providing analyses from their systems into this merged resource are summarized in Table 1, with references to descriptions of the systems used to create the included analyses. We compiled for each participant and each task both the final test set submission and a comparable submission for the separate development set.

As the gold annotations for the test set are only available for evaluation through an online interface (in order to avoid overfitting and assure the comparability of results), it is important to provide also development set analyses to permit direct comparison

[2] http://research.microsoft.com/bionlp/

[3] We have yet to hear back from a few groups, but none has yet explicitly denied the release of their data. Should any remaining group accept the release of their data, we will release a new, extended version of the resource.

Team	Task GE	EPI	ID	BB	BI	CO	REL	REN	System description
UTurku	1	1	1	1	1	1	1	1	Björne and Salakoski (2011)
ConcordU	1	1	1			1	1	1	Kilicoglu and Bergler (2011)
UMass	1	1	1						Riedel and McCallum (2011)
Stanford	1	1	1						McClosky et al. (2011)
FAUST	1	1	1						Riedel et al. (2011)
MSR-NLP	1	1							Quirk et al. (2011)
CCP-BTMG	1	1							Liu et al. (2011)
BMI@ASU	1								Emadzadeh et al. (2011)
TM-SCS	1								Bui and Sloot (2011)
UWMadison	1								Vlachos and Craven (2011)
HCMUS	1						1		Le Minh et al. (2011)
PredX		1							-
VIBGhent							1		Van Landeghem et al. (2011a)

Table 1: BioNLP ST 2011 participants contributing to the combined resource.

Task	Events Gold	FN	Recall
GE (task 1)	3250	1006	69.05%
EPI (CORE task)	601	129	78.54%
ID (CORE task)	691	183	73.52%

Table 2: Recall for the union of analyses from systems included in the combined dataset.

against gold annotations. The inclusion of both development and test set annotations also allows e.g. the study of system combination approaches where the combination parameters are estimated on development data for final testing on the test set (Kim et al., 2011a).

3.2 Evaluation

We demonstrate the use of the newly compiled dataset through a manual evaluation of GE, EPI and ID main task development set gold standard events that are not extracted by any of the systems for which analyses were available.[4] We perform evaluation on the GE subtask 1 and the EPI and ID task CORE subtasks, as all participating systems addressed the extraction targets of these subtasks.

We first evaluated each of the analyses against the development set of the respective task using the official shared task evaluation software, using options for the evaluation tools to list the sets of true positive (TP), false positive (FP) and false negative (FN) events. We then selected for each of the three tasks the set of events that were included in the FN list for all systems. This gives the results for the recall of the union of all systems shown in Table 2. The recall of the system union is approximately 30% points higher than that of any individual GE system (Kim et al., 2011c) and 25% points higher for EPI and ID (Ohta et al., 2011; Pyysalo et al., 2011a), suggesting potential remaining benefits from system combination. Nevertheless, a substantial fraction of the total set of gold events remains inaccessible also to this system union.

We then selected a random set of 100 events from each of the three sets of events that were not recovered by any system (i.e. 300 events in total) and performed a manual evaluation to identify frequent properties of these events that could contribute to extraction failures. In brief, we first performed a brief manual evaluation to identify common characteristics of these events, and then evaluated the 300 events individually to identify the set of these characteristics that apply to each event.

The results of the evaluation for common cases are shown in Table 3. We find that the most frequent property of the unrecoverable events is that they involve implicit arguments (Gerber and Chai, 2010), a difficult challenge that has not been extensively considered in domain event extraction. A closely related issue are events involving arguments in a sentence different from that containing the trigger ("cross-sentence"), connected either implicitly or through explicit coreference ("coreference"). Al-

[4]The final collection includes analyses from the systems of two groups that agreed to the release of their data after the completion of this analysis, but we expect the results to largely hold also for the final collection.

Type	GE	EPI	ID	Total
Implicit argument	18	33	15	66
Cross-sentence	14	40	4	58
Weak trigger	28	14	11	53
Coreference	12	20	18	50
Static Relation	6	28	6	40
Error in gold	17	4	9	30
Ambiguous type	2	9	11	22
Shared trigger	2	12	1	15

Table 3: Manual evaluation results for features of events that could not be recovered by any system.

though coreference was considered as as separate task in BioNLP ST 2011 (Nguyen et al., 2011), it is clear that it involves many remaining challenges for event extraction systems. Similarly, events where explicit arguments are connected to other arguments through "static" relations such as *part-of* (e.g. "*A binds the X domain of B*") represent a known challenge (Pyysalo et al., 2011b). These results suggest that further advances in event extraction performance could be gained by the integration of systems for the analysis of coreference and static relations, approaches for which some success has already been demonstrated in recent efforts (Van Landeghem et al., 2010; Yoshikawa et al., 2011; Miwa et al., 2012).

"Weak" trigger expressions that must be interpreted in context to determine whether they express an event, as well as a related class of events whose type must be disambiguated with reference to context ("ambiguous type") are comparatively frequent in the three tasks, while EPI in particular involves many cases where a trigger is shared between multiple events – an issue for approaches that assume each token can be assigned at most a single class. Finally, we noted a number of cases that we judged to be errors in the gold annotation; the number is broadly in line with the reported inter-annotator agreement for the data (see e.g. Ohta et al. (2011)).

While there is an unavoidable subjective component to evaluations such as this, we note that a similar evaluation performed following the BioNLP Shared Task 2009 using test set data reached broadly comparable results (Kim et al., 2011a). The newly compiled dataset represents the first opportunity for those without direct access to the test set data and submissions to directly assess the task results, as demonstrated here. We hope that this resource will encourage further exploration of both the data, the system analyses and remaining challenges in event extraction.

4 New Perspectives to Event Extraction

As discussed in Section 2, the BioNLP ST event extraction task is "text-bound": each entity and event annotation is associated with a specific span of text. Contrasted to the alternative approach where annotations are document-level only, this approach has a number of important benefits, such as allowing machine learning methods for event extraction to be directly trained on fully and specifically annotated data without the need to apply frequently error-prone heuristics (Mintz et al., 2009) or develop machine learning methods addressing the mapping between text expressions and document-level annotations (Riedel et al., 2010). Many of the most successful event extraction approaches involve direct training of machine learning methods using the text-bound annotations (Riedel and McCallum, 2011; Björne and Salakoski, 2011; McClosky et al., 2011). However, while the availability of text-bound annotations in data provided to task participants is clearly a benefit, there are drawbacks to the choice of exclusive focus on text-bound annotations in system output, including issues relating to evaluation and the applicability of methods to the task. In the following section, we discuss some of these issues and propose alternatives to representation and evaluation addressing them.

4.1 Evaluation

The evaluation of the BioNLP ST is instance-based and text-bound: each event in gold annotation and each event extracted by a system is considered independently, separating different mentions of the "same" real-world event. This is the most detailed (sensitive) evaluation setting permitted by the data, and from a technical perspective a reasonable choice for ranking systems performing the task.

However, from a practical perspective, this evaluation setting arguably places excessively strict demands on systems, and may result in poor correlation between measured performance and the practical value of systems. Our motivating observations are that specific real-world events tend to be men-

tioned multiple times in a single publication – especially the events that are of particular importance in the study – and that there are few practical applications for which it is necessary to find each such repeated mention. For example, in literature search for e.g. pathway or database curation support, one typical information need is to identify biomolecular reactions involving a specific protein. Event extraction can support such needs either by summarizing all events involving the protein that could be extracted from the literature (Van Landeghem et al., 2011b), or by retrieving documents (perhaps showing relevant text snippets) containing such events (Ohta et al., 2010). For the former to meet the information need, it may be sufficient that each different event is extracted once from the entire literature; for the latter, once from each relevant document. For uses such as these, there is no obvious need for, or, indeed, no very obvious benefit from the ability of extraction systems to separately enumerate every mention of every event in every publication. It is easy to envision other practical use cases where instance-level extraction performance is at best secondary and, we argue, difficult to identify ones where it is of critical importance.

For applications such as these, the important question is the reliability of the system at identifying events either on the level of documents or on the level of (a relevant subset of) the literature, rather than on the level of individual mentions. For a more complete and realistic picture of the practical value of event extraction methods, measures other than instance-level should thus also be considered.

4.2 Task setting

While applications can benefit from the ability of IE systems to identify a specific span of text supporting extracted information,[5] the requirement of the BioNLP ST setting that the output of event extraction systems must identify specific text spans for each entity and event makes it complex or impossible to address the task using a number of IE methods that might otherwise represent feasible approaches to event extraction.

[5]For example, for curation support tasks, this allows the human curator to easily check the correctness of extracted information and helps to select "evidence sentences", as included in many databases.

For example, Patwardhan and Riloff (2007) and Chambers and Jurafsky (2011) consider an IE approach where the extraction targets are MUC-4 style document-level templates (Sundheim, 1991), the former a supervised system and the latter fully unsupervised. These methods and many like them for tasks such as ACE (Doddington et al., 2004) work on the document level, and can thus not be readily applied or evaluated against the existing annotations for the BioNLP shared tasks. Enabling the application of such approaches to the BioNLP ST could bring valuable new perspectives to event extraction.

4.3 Alternative evaluation

We propose a new mode of evaluation that otherwise follows the primary BioNLP ST evaluation criteria, but incorporates the following two exceptions:

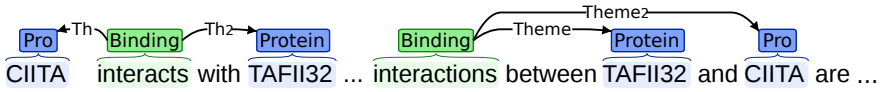
1. remove the requirement to match trigger spans
2. only require entity texts, not spans, to match

The first alternative criterion has also been previously considered in the GE task evaluation (Kim et al., 2011c); the latter has, to the best of our knowledge, not been previously considered in domain event extraction. We additionally propose to consider only the minimal set of events that are unique on the document level (under the evaluation criteria), thus eliminating effects from repeated mentions of a single event on evaluated performance. We created tools implementing this mode of evaluation with reference to the BioNLP ST 2011 evaluation tools.

While this type of evaluation has, to the best of our knowledge, not been previously applied specifically in biomedical event extraction, it is closely related (though not identical) to evaluation criteria applied in MUC, ACE, and the in-domain PPI relation extraction tasks in BioCreative (Krallinger et al., 2008).

4.4 Alternative representation

A true conversion to a document-level, "off the page" representation would require manual annotation efforts to identify the real-world entities and events referred to in text (Doddington et al., 2004). However, it is possible to reasonably approximate such a representation through an automatic heuristic conversion.



BioNLP Shared Task	*Document level*
T1 Protein 0 5 CIITA	T1 Protein CIITA
T2 Protein 21 28 TAFII32	T2 Protein TAFII32
T3 Binding 6 15 interacts	E1 Binding Theme:T1 Theme2:T2
E1 Binding:T3 Theme:T1 Theme2:T2	
T4 Protein 54 61 TAFII32	
T5 Protein 66 71 CIITA	
T6 Binding 33 45 interactions	
E2 Binding:T6 Theme:T4 Theme2:T5	

Figure 2: Illustration of BioNLP Shared Task annotation format and the proposed document-level ("off-the-page") format.

We first introduce a non-textbound annotation format that normalizes over differences in e.g. argument order and eliminates duplicate events. The format largely follows that of the shared task but removes any dependencies and references to text offsets (see Figure 2). The conversion process into this representation involves a number of steps. First, we merge duplicate pairs of surface strings and types, as different mentions of the same entity in different parts of the text are no longer distinguishable in the representation. In the original format, equivalence relations (Kim et al., 2011a) are annotated only for specific mentions. When "raising" the annotations to the document level, equivalence relations are reinterpreted to cover the full document by extending the equivalence to all mentions that share the surface form and type with members of existing equivalence classes. Finally, we implemented an event equivalence comparison to remove duplicate annotations from each document. The result of the conversion to this alternate representation is thus an "off-the-page" summary of the unique set of events in the document.

This data can then be used for training and comparison of methods analogously to the original annotations, but without the requirement that all analyses include text-bound annotations.

4.5 Experimental Results

We next present an evaluation using the alternative document-level event representation and evaluation, comparing its results to those for the primary shared task evaluation criteria. As comparatively few of the shared task participants attempted subtasks 2 and 3 for GE or the FULL task setting for EPI and ID, we consider only GE subtask 1 and the EPI and ID task CORE extraction targets in these experiments. We refer to the task overviews for the details of the subtasks and the primary evaluation criteria (Kim et al., 2011c; Pyysalo et al., 2011a; Ohta et al., 2011).

	Primary criteria			New criteria		
Group	Rec.	Prec.	F	Rec.	Prec.	F
FAUST	49.41	64.75	56.04	53.10	67.56	59.46
UMass	48.49	64.08	55.20	52.55	66.57	58.74
UTurku	49.56	57.65	53.30	54.23	60.11	57.02
MSR-NLP	48.64	54.71	51.50	53.55	58.24	55.80
ConcordU	43.55	59.58	50.32	47.42	60.85	53.30
UWMadison	42.56	61.21	50.21	46.09	62.50	53.06
Stanford	42.36	61.08	50.03	46.48	63.22	53.57
BMI@ASU	36.91	56.63	44.69	41.15	61.44	49.29
CCP-BTMG	31.57	58.99	41.13	34.82	66.89	45.80
TM-SCS	32.73	45.84	38.19	38.02	50.87	43.51
HCMUS	10.12	27.17	14.75	14.50	40.05	21.29

Table 4: Comparison of BioNLP ST 2011 GE task 1 results.

Tables 4, 5 and 6 present the results for the GE, EPI and ID tasks, respectively. For GE, we see consistently higher F-scores for the new criteria, in most cases reflecting primarily an increase in recall, but also involving increases in precision. The F-score differences range between 3-4% for most high-ranking systems, with more substantial increases for lower-ranking systems. Notable increases in precision are seen for some systems (e.g. HCMUS), indicating that the systems comparatively frequently extract correct information, but associated with the wrong spans of text.

	Primary criteria			New criteria		
Group	Rec.	Prec.	F	Rec.	Prec.	F
UTurku	68.51	69.20	68.86	74.20	69.14	71.58
FAUST	59.88	80.25	68.59	67.04	76.82	71.60
MSR-NLP	55.70	77.60	64.85	59.24	77.66	67.21
UMass	57.04	73.30	64.15	65.76	69.65	67.65
Stanford	56.87	70.22	62.84	62.74	67.12	64.86
CCP-BTMG	45.06	63.37	52.67	54.62	63.17	58.58
ConcordU	40.28	76.71	52.83	48.41	76.57	59.32

Table 5: Comparison of BioNLP ST 2011 EPI CORE task results.

	Primary criteria			New criteria		
Group	Rec.	Prec.	F	Rec.	Prec.	F
FAUST	50.84	66.35	57.57	50.11	65.33	56.72
UMass	49.67	62.39	55.31	49.34	60.98	54.55
Stanford	49.16	56.37	52.52	42.00	50.80	45.98
ConcordU	50.91	43.37	46.84	43.42	37.18	40.06
UTurku	39.23	49.91	43.93	48.03	51.84	49.86
PredX	23.67	35.18	28.30	20.94	30.69	24.90

Table 6: Comparison of BioNLP ST 2011 ID CORE task results.

For EPI (Table 5), we find comparable differences in F-score to those for GE, but there is a significant difference in the precision-recall balance: the majority of systems show over 5% points higher recall under the new criteria, but many show substantial losses in precision, while for GE precision was also systematically increased. This effect was not unexpected: we judge this to reflect primarily the increased number of opportunities to extract each unique event (higher recall) combined with the comparatively higher effect from errors from the reduction in the total number of unique correct extraction targets (lower precision). It is not clear from our analysis why a comparable effect was not seen for GE. Interestingly, most systems show a better precision/recall balance under the new criteria than the old, despite not optimizing for these criteria.

For ID (Table 6), we find a different effect also on F-score, with all but one system showing reduced performance under the new criteria, with some very clear drops in performance; the only system to benefit is UTurku. Analysis suggests that this effect traces primarily to a notable reduction in the number of simple PROCESS events that take no arguments[6] when considering unique events on the document level instead of each event mention independently.[7] Conversely, the Stanford system, which showed the highest instance-level performance in the extraction of PROCESS type events (see Pyysalo et al. (2011a)), shows a clear loss in precision.

[6]The ID task annotation criteria call for mentions of some high-level biological processes such as "infection" to be annotated as PROCESS even if no explicit participants are mentioned (Pyysalo et al., 2011a).

[7]It is interesting to note that there was an error in the UTurku system implementation causing it to fail to output any events without arguments (Jari Björne, personal communication), likely contributing to the effect seen here.

The clear differences in performance and the many cases in which the system rankings under the two criteria differ demonstrate that the new evaluation criteria can have a decisive effect in which approaches to event extraction appear preferred. While there may be cases for which the original shared task criteria are preferred, there is at the very minimum a reasonable argument to be made that the emphasis these criteria place on the extraction of each instance of simple events is unlikely to reflect the needs of many practical applications of event extraction.

While these experimental results demonstrate that the new evaluation criteria emphasize substantially different aspects of the performance of the systems than the original criteria, they cannot *per se* serve as an argument in favor of one set of criteria over another. We hope that these results and the accompanying tools will encourage increased study and discussion of evaluation criteria for event extraction and more careful consideration of the needs of specific applications of the technology.

5 Discussion and Conclusions

We have presented a new resource combining analyses from the systems participating in the GE, ID and EPI main tasks of the BioNLP Shared Task 2011, compiled with the collaboration of groups participating in these tasks. We demonstrated one use of the resource through an evaluation of development set events that none of the participating systems could recover, finding that events involving implicit arguments, coreference and participants in more than once sentence continue to represent challenges to the event extraction systems that participated in these tasks.

We further argued in favor of new perspectives to the evaluation of domain event extraction systems,

emphasizing in particular the need for document-level, "off-the-page" representations and evaluation to complement the text-bound, instance-level evaluation criteria that have so far been applied in the shared task evaluation. We proposed a variant of the shared task standoff representation for supporting such evaluation, and introduced evaluation tools implementing the proposed criteria. An evaluation supported by the introduced resources demonstrated that the new criteria can in cases provide substantially different results and rankings of the systems, confirming that the proposed evaluation can serve as an informative complementary perspective into event extraction performance.

In future work, we hope to further extend the coverage of the provided system outputs as well as their analysis to cover all participants of all tasks in the BioNLP Shared Task 2011. We also aim to use the compiled resource in further study of appropriate criteria for the evaluation of event extraction methods and deeper analysis of the remaining challenges in event extraction.

To encourage further study of all aspects of event extraction, all resources and tools introduced in this study are provided freely to the community from `http://2011.bionlp-st.org`.

Acknowledgments

We wish to thank the members of all groups contributing to the combined resource, and in particular the members of the MSR-NLP group for providing both the initial suggestion for its creation as well as the first publicly released analyses from their system. We would also like to thank the anonymous reviewers for their many insightful comments.

This work was funded in part by UK Biotechnology and Biological Sciences Research Council (BBSRC) under project Automated Biological Event Extraction from the Literature for Drug Discovery (reference number: BB/G013160/1), by the Ministry of Education, Culture, Sports, Science and Technology of Japan under the Integrated Database Project and by the Swedish Royal Academy of Sciences.

References

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*.

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, Maarten van de Guchte, and Claire Nédellec. 2011. BioNLP Shared Task 2011 - Bacteria Biotope. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 56–64.

Quoc-Chinh Bui and Peter. M.A. Sloot. 2011. Extracting biological events from text using simple syntactic patterns. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 143–146.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the ACL-HLT 2011*, pages 976–986.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of LREC*, volume 4, pages 837–840.

Ehsan Emadzadeh, Azadeh Nikfarjam, and Graciela Gonzalez. 2011. Double layered learning for biological event extraction from text. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 153–154.

Matthew Gerber and Joyce Chai. 2010. Beyond nombank: A study of implicit arguments for nominal predicates. In *Proceedings of ACL 2010*, pages 1583–1592.

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 BioNLP Shared Task 2011 Workshop*, pages 65–73.

Brian Kemper, Takuya Matsuzaki, Yukiko Matsuoka, Yoshimasa Tsuruoka, Hiroaki Kitano, Sophia Ananiadou, and Jun'ichi Tsujii. 2010. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26(12):i374–i381.

Halil Kilicoglu and Sabine Bergler. 2011. Adapting a general semantic interpretation approach to biological event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2011a. Extracting bio-molecular events from literature - the BioNLP'09 shared task. *Computational Intelligence*, 27(4):513–540.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011b.

Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task*, pages 1–6.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011c. Overview of the Genia Event task in BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, Alfonso Valencia, et al. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

Quang Le Minh, Son Nguyen Truong, and Quoc Ho Bao. 2011. A pattern approach for biomedical event annotation. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 149–150.

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 Shared Task 2011 Workshop*.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event extraction as dependency parsing. In *Proceedings of ACL-HLT 2011*, pages 1626–1635.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP 2009*, pages 1003–1011.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*.

Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Overview of BioNLP 2011 Protein Coreference Shared Task. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 74–82.

Tomoko Ohta, Takuya Matsuzaki, Naoaki Okazaki, Makoto Miwa, Rune Sætre, Sampo Pyysalo, and Jun'ichi Tsujii. 2010. Medie and info-pubmed: 2010 update. *BMC Bioinformatics*, 11(Suppl 5):P7.

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 Shared Task 2011 Workshop*.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of EMNLP-CoNLL 2007*, pages 717–727.

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 Shared Task 2011 Workshop*.

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 BioNLP Shared Task 2011 Workshop*, pages 83–88.

Chris Quirk, Pallavi Choudhury, Michael Gamon, and Lucy Vanderwende. 2011. MSR-NLP entry in BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 155–163.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proceedings of EMNLP 2011*, pages 1–12.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. *Machine Learning and Knowledge Discovery in Databases*, pages 148–163.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, and Chris Manning. 2011. Model combination for event extraction in BioNLP 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*.

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 Shared Task 2011 Workshop*.

Beth M. Sundheim. 1991. Third message understanding evaluation and conference (MUC-3): Phase 1 status report. In *Proceedings of the Speech and Natural Language Workshop*, pages 301–305.

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 BioNLP 2010*, pages 144–152.

Sofie Van Landeghem, Thomas Abeel, Bernard De Baets, and Yves Van de Peer. 2011a. Detecting entity relations as a supporting task for bio-molecular event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 147–148.

Sofie Van Landeghem, Filip Ginter, Yves Van de Peer, and Tapio Salakoski. 2011b. Evex: a pubmed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of BioNLP 2011 Workshop*, pages 28–37.

Andreas Vlachos and Mark Craven. 2011. Biomedical event extraction from abstracts and full papers using search-based structured prediction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 36–40.

Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, and Yuji Matsumoto. 2011. Coreference based event-argument relation extraction on biomedical text. *Journal of Biomedical Semantics*, 2(Suppl 5):S6.