Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011

Sampo Pyysalo* **Tomoko Ohta*** **Rafal Rak‡§** **Dan Sullivan†** **Chunhong Mao†**
Chunxia Wang† **Bruno Sobral†** **Jun'ichi Tsujii¶** **Sophia Ananiadou‡§**
*Department of Computer Science, University of Tokyo, Tokyo, Japan
†Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA
‡School of Computer Science, University of Manchester, Manchester, UK
§National Centre for Text Mining, University of Manchester, Manchester, UK
¶Microsoft Research Asia, Beijing, China
{smp,okap}@is.s.u-tokyo.ac.jp jtsujii@microsoft.com
{dsulliva,cmao,cwang,sobral}@vbi.vt.edu
{rafal.rak,sophia.ananiadou}@manchester.ac.uk

Abstract

This paper presents the preparation, resources, results and analysis of the Infectious Diseases (ID) information extraction task, a main task of the BioNLP Shared Task 2011. The ID task represents an application and extension of the BioNLP'09 shared task event extraction approach to full papers on infectious diseases. Seven teams submitted final results to the task, with the highest-performing system achieving 56% F-score in the full task, comparable to state-of-the-art performance in the established BioNLP'09 task. The results indicate that event extraction methods generalize well to new domains and full-text publications and are applicable to the extraction of events relevant to the molecular mechanisms of infectious diseases.

1 Introduction

The Infectious Diseases (ID) task of the BioNLP Shared Task 2011 (Kim et al., 2011a) is an information extraction task focusing on the biomolecular mechanisms of infectious diseases. The primary target of the task is event extraction (Ananiadou et al., 2010), broadly following the task setup of the BioNLP'09 Shared Task (BioNLP ST'09) (Kim et al., 2009).

The task concentrates on the specific domain of two-component systems (TCSs, or two-component regulatory systems), a mechanism widely used by bacteria to sense and respond to the environment (Thomason and Kay, 2000). Typical TCSs consist of two proteins, a membrane-associated sensor kinase and a cytoplasmic response regulator. The sensor kinase monitors changes in the environment while the response regulator mediates an adaptive response, usually through differential expression of target genes (Mascher et al., 2006). TCSs have many functions, but those of particular interest for infectious disease researchers include virulence, response to antibiotics, quorum sensing, and bacterial cell attachment (Krell et al., 2010). Not all TCS functions are well known: in some cases, TCSs are involved in metabolic processes that are difficult to precisely characterize (Wang et al., 2010). TCSs are of interest also as drugs designed to disrupt TCSs may reduce the virulence of bacteria without killing it, thus avoiding the potential selective pressure of antibiotics lethal to some pathogenic bacteria (Gotoh et al., 2010). Information extraction techniques may support better understanding of these fundamental systems by identifying and structuring the molecular processes underlying two component signaling.

The ID task seeks to address these opportunities by adapting the BioNLP ST'09 event extraction model to domain scientific publications. This model was originally introduced to represent biomolecular events relating to transcription factors in human blood cells, and its adaptation to a domain that centrally concerns both bacteria and their hosts involves a variety of novel aspects, such as events concerning whole organisms, the chemical environment of bacteria, prokaryote-specific concepts (e.g. regulons as elements of gene expression), as well as the effects of biomolecules on larger-scale processes involving hosts such as virulence.

Proceedings of BioNLP Shared Task 2011 Workshop, pages 26–35,
Portland, Oregon, USA, 24 June, 2011.

2 Task Setting

The ID task broadly follows the task definition and event types of the BioNLP ST'09, extending it with new entity categories, correspondingly broadening the scope of events, and introducing a new class of events, high-level biological processes.

2.1 Entities

The ID task defines five core types of entities: genes/gene products, two-component systems, regulons/operons, chemicals, and organisms. Following the general policy of the BioNLP Shared Task, the recognition of the core entities is not part of the ID task. As named entity recognition (NER) is considered in other prominent domain evaluations (Krallinger et al., 2008), we have chosen to isolate aspects of extraction performance relating to NER from the main task of interest, event extraction, by providing participants with human-created gold annotations for core entities. These annotations are briefly presented in the following.

Mentions of names of genes and their products (RNA and proteins) are annotated with a single type, without differentiating between subtypes, following the guidelines of the GENIA GGP corpus (Ohta et al., 2009). This type is named PROTEIN to maintain consistency with related tasks (e.g. BioNLP ST'09), despite slight inaccuracy for cases specifically referencing RNA or DNA forms. Two-component systems, consisting of two proteins, frequently have names derived from the names of the proteins involved (e.g. *PhoP-PhoR* or *SsrA/SsrB*). Mentions of TCSs are annotated as TWO-COMPONENT-SYSTEM, nesting PROTEIN annotations if present. Regulons and operons are collections of genes whose expression is jointly regulated. Like the names of TCSs, their names may derive from the names of the involved genes and proteins, and are annotated as embedding PROTEIN annotations when they do. The annotation does not differentiate between the two, marking both with a single type REGULON-OPERON.

In addition to these three classes relating to genes and proteins, the core entity annotation recognizes the classes CHEMICAL and ORGANISM. All mentions of formal and informal names of atoms, inorganic compounds, carbohydrates and lipids as well as organic compounds other than amino acid and nucleic acid compounds (i.e. gene/protein-related compounds) are annotated as CHEMICAL. Mentions of names of families, genera, species and strains as well as non-name references with comparable specificity are annotated as ORGANISM.

Finally, the non-specific type ENTITY[1] is defined for marking entities that specify additional details of events such as the binding site in a BINDING event or the location an entity moves to in a LOCALIZATION event. Unlike the core entities, annotations of the generic ENTITY type are not provided for test data and must be detected by participants addressing the full task.

2.2 Relations

The ID task involves one relation, EQUIV, defining entities (of any of the core types) to be equivalent. This relation is used to annotate abbreviations and local aliases and it is not a target of extraction, but provided for reference and applied in evaluation, where references to any of a set of equivalent entities are treated identically.

2.3 Events

The primary extraction targets of the ID task are the event types summarized in Table 1. These are a superset of those targeted in the BioNLP ST'09 and its repeat, the 2011 GE task (Kim et al., 2011b). This design makes it possible to study aspects of domain adaptation by having the same extraction targets in two subdomains of biomedicine, that of transcription factors in human blood cells (GE) and infectious diseases. The events in the ID task extend on those of GE in the inclusion of additional entity types as participants in previously considered event types and the introduction of a new type, PROCESS. We next briefly discuss the semantics of these events, defined (as in GE) with reference to the community-standard Gene Ontology (Ashburner et al., 2000). We refer to (Kim et al., 2008; Kim et al., 2009) for the ST'09/GE definitions.

[1] In terms of the GENIA ontology, ENTITY is used to mark e.g. PROTEIN DOMAIN OR REGION references. Specific types were applied in manual annotation, but these were replaced with the generic ENTITY in part to maintain consistency with BioNLP ST'09 data and to reduce the NER-related demands on participating systems by not requiring the assignment of detailed types.

Type	Core arguments	Additional arguments
GENE EXPRESSION	*Theme*(PROTEIN or REGULON-OPERON)	
TRANSCRIPTION	*Theme*(PROTEIN or REGULON-OPERON)	
PROTEIN CATABOLISM	*Theme*(PROTEIN)	
PHOSPHORYLATION	*Theme*(PROTEIN)	*Site*(ENTITY)
LOCALIZATION	*Theme*(Core entity)	*AtLoc*(ENTITY), *ToLoc*(ENTITY)
BINDING	*Theme*(Core entity)+	*Site*(ENTITY)+
PROCESS	*Participant*(Core entity)?	
REGULATION	*Theme*(Core entity / Event), *Cause*(Core entity / Event)?	*Site*(ENTITY), *CSite*(ENTITY)
POSITIVE REGULATION	*Theme*(Core entity / Event), *Cause*(Core entity / Event)?	*Site*(ENTITY), *CSite*(ENTITY)
NEGATIVE REGULATION	*Theme*(Core entity / Event), *Cause*(Core entity / Event)?	*Site*(ENTITY), *CSite*(ENTITY)

Table 1: Event types and their arguments. The type of entity allowed as argument is specified in parenthesis. "Core entity" is any of PROTEIN, TWO-COMPONENT-SYSTEM, REGULON-OPERON, CHEMICAL, or ORGANISM. Arguments that can be filled multiple times marked with "+", non-mandatory core arguments with "?" (all additional arguments are non-mandatory).

The definitions of the first four types in Table 1 are otherwise unchanged from the ST'09 definitions except that GENE EXPRESSION and TRANSCRIPTION extend on the former definition in recognizing REGULON-OPERON as an alternative unit of expression. LOCALIZATION, taking only PROTEIN type arguments in the ST'09 definition, is allowed to take any core entity argument. This expanded definition remains consistent with the scope of the corresponding GO term (GO:0051179). BINDING is similarly extended, giving it a scope largely consistent with GO:0005488 (binding) but also encompassing GO:0007155 (cell adhesion) (e.g. a bacterium binding another) and protein-organism binding. The three regulation types (REGULATION, POSITIVE REGULATION, and NEGATIVE REGULATION) likewise allow the new core entity types as arguments, but their definitions are otherwise unchanged from those in ST'09, that is, the GENIA ontology definitions. As in these resources, regulation types are used not only for the biological sense but also to capture statements of general causality (Kim et al., 2008). As in ST'09, all events of types discussed above require a *Theme* argument: only events involving an explicitly stated theme (of an appropriate type) should be extracted. All other arguments are optional.

The PROCESS type, new to ID, is used to annotate high-level processes such as virulence, infection and resistance that involve infectious organisms. This type differs from the others in that it has no mandatory arguments: the targeted processes should be extracted even if they have no explicitly stated participants, reflecting that they are of interest even without the further specification. When stated, the involved participants are captured using the generic role type *Participant*. Figure 1 shows an illustration of some of the the ID task extraction targets.

We term the first five event types in Table 1 taking exactly one *Theme* argument as their core argument *simple events*. In analysis we further differentiate *non-regulation events* (the first seven) and *regulation* (the last three), which is known to represent particular challenges for extraction in involving events as arguments, thus creating nested event structures.

2.4 Event modifications

The ID task defines two *event modification* extraction targets, NEGATION and SPECULATION. These modifications mark events as being explicitly negated (e.g. *virB is not expressed*) or stated in a speculative context (e.g. *virB may be expressed*). Both may apply simultaneously. The modification definitions are identical to the ST'09 ones, including the representation in which modifications (unlike events) are not assigned text bindings.

3 Data

The ID task data were newly annotated for the BioNLP Shared Task and are not based on any previously released resource. Annotation was performed by two teams, one in Tsujii laboratory (University of Tokyo) and one in Virginia Bioinformatics Institute (Virginia Tech). The entity and event annotation

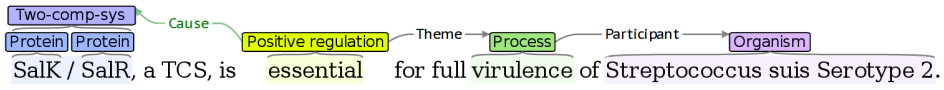

Figure 1: Example event annotation. The association of a TCS with an organism is captured through an event structure involving a PROCESS ("virulence") and POSITIVE REGULATION. Regulation types are used to capture also statements of general causality such as "is essential for" here. (Simplified from PMC ID 2358977)

Journal	#	Published
PLoS Pathogens	9	2006–2010
PLoS One	7	2008–2010
BMC Genomics	3	2008–2010
PLoS Genetics	2	2007–2010
Open Microbiology J.	2	2008–2010
BMC Microbiology	2	2008–2009
Other	5	2007–2008

Table 2: Corpus composition. Journals in which selected articles were published with number of articles (#) and publication years.

Entity type	prec.	rec.	F
PROTEIN	54.64	39.64	45.95
CHEMICAL	32.24	19.05	23.95
ORGANISM	90.38	47.70	62.44
TWO-COMPONENT-SYSTEM	87.69	47.24	61.40

Table 3: Automatic core entity tagging performance.

design was guided by previous studies on NER and event extraction in a closely related domain (Pyysalo et al., 2010; Ananiadou et al., 2011).

3.1 Document selection

The training and test data were drawn from the primary text content of recent full-text PMC open access documents selected by infectious diseases domain experts (Virginia Tech team) as representative publications on two-component regulatory systems. Table 2 presents some characteristics of the corpus composition. To focus efforts on natural language text likely to express novel information, we excluded tables, figures and their captions, as well as methods sections, acknowledgments, authors' contributions, and similar meta-content.

3.2 Annotation

Annotation was performed in two primary stages, one for marking core entities and the other for events and secondary entities. As a preliminary processing step, initial sentence segmentation was performed with the GENIA Sentence Splitter[2]. Segmentation errors were corrected during core entity annotation.

Core entity annotation was performed from the basis of an automatic annotation created using selected existing taggers for the target entities. The following tools and settings were adopted, with parameters tuned on initial annotation for two documents:

PROTEIN: NeMine (Sasaki et al., 2008) trained on the JNLPBA data (Kim et al., 2004) with threshold 0.05, filtered to only GENE and PROTEIN types.

ORGANISM: Linnaeus (Gerner et al., 2010) with "variant matching" for species names variants.

CHEMICAL: OSCAR3 (Corbett and Murray-Rust, 2006) with confidence 90%.

TWO-COMPONENT-SYSTEM: Custom regular expressions.

Initial automatic tagging was not applied for entities of the REGULON-OPERON type or the generic ENTITY type (for additional event arguments). All automatically generated annotations were at least confirmed through manual inspection, and the majority of the automatic annotations were revised in manual annotation. Table 3 summarizes the tagging performance of the automatic tools as measured against the final human-annotated training and development datasets.[3]

Annotation for the task extraction targets – events and event modifications – was created entirely manually without automatic annotation support to avoid any possible bias toward specific extraction methods or approaches. The Tsujii laboratory team orga-

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/~y-matsu/geniass/

[3] It should be noted that these results are low in part due to differences in annotation criteria (see e.g. (Wang et al., 2009)) and to data tagged using the ID task annotation guidelines not being applied for training; training on the newly annotated data is expected to allow notably more accurate tagging.

Item	Train	Devel	Test	Total
Articles	15	5	10	30
Sentences	2,484	709	1,925	5118
Words	74,439	21,225	57,489	153,153
Core entities	6,525	1,976	4,239	12,740
Events	2,088	691	1,371	4150
Modifications	95	45	74	214

Table 4: Statistics of the ID corpus.

nized the annotation effort, with a coordinating annotator with extensive experience in event annotation (TO) leading annotator training and annotation scheme development. Detailed annotation guidelines (Pyysalo et al., 2011) extending on the GENIA annotation guidelines were developed jointly with all annotators and refined throughout the annotation effort. Based on measurements of inter-annotator consistency between annotations independently created by the two teams, made throughout annotator training and primary annotation (excluding final corpus cleanup), we estimate the consistency of the final entity annotation to be no lower than 90% F-score and that of the event annotation to be no lower than 75% F-score for the primary evaluation criteria (see Section 4).

3.3 Datasets and statistics

Initial annotation was produced for the selected sections (see Section 3.1) in 33 full-text articles, of which 30 were selected for the final dataset as representative of the extraction targets. These documents were split into training, development and test sets of 15, 5 and 10 documents, respectively. Participants were provided with all training and development set annotations and test set core entity annotations. The overall statistics of the datasets are given in Table 4.

As the corpus consists of full-text articles, it contains a somewhat limited number of articles, but in other terms it is of broadly comparable size to the largest of the BioNLP ST corpora: the corpus word count, for example, corresponds to that of a corpus of approximately 800 PubMed abstracts, and the core entity count is comparable to that in the ST'09 data. However, for reasons that may relate in part to the domain, the event count is approximately a third of that for the ST'09 data. In addition to having less training data, the entity/event ratio is thus considerably higher (i.e. there are more candidates for each true target), suggesting that the ID data could be expected to provide a more challenging extraction task.

4 Evaluation

The performance of participating systems was evaluated in terms of events using the standard precision/recall/F-score metrics. For the primary evaluation, we adopted the standard criteria defined in the BioNLP'09 shared task. In brief, for determining whether a reference annotation and a predicted annotation match, these criteria relax exact matching for event triggers and arguments in two ways: matching of text-bound annotation (event triggers and ENTITY type entities) allows limited boundary variation, and only core arguments need to match in nested event arguments for events to match. For details of the matching criteria, please refer to Kim et al. (2009).

The primary evaluation for the task requires the extraction of all event arguments (both core and additional; see Table 1) as well as event modifications (NEGATION and SPECULATION). This is termed the *full task*. We additionally report extraction results for evaluation where both the gold standard reference data and the submission events are reduced to only core arguments, event modifications are removed, and resulting duplicate events removed. We term this the *core task*. In terms of the subtask division applied in the BioNLP'09 Shared Task and the GE task of 2011, the core task is analogous to subtask 1 and the full task analogous to the combination of subtasks 1–3.

5 Results

5.1 Participation

Final results to the task were successfully submitted by seven participants. Table 5 summarizes the information provided by the participating teams. We note that full parsing is applied in all systems, with the specific choice of the parser of Charniak and Johnson (2005) with the biomedical domain model of McClosky (2009) and conversion into the Stanford Dependency representation (de Marneffe et al., 2006) being adopted by five participants. Further, five of the seven systems are predominantly machine learning-based. These can be seen as extensions of trends that were noted in analysis of the BioNLP

Rank	Team	Org	NLP		Events				Other resources	
			Word	Parse	Trig.	Arg.	Group.	Modif.	Corpora	Other
1	FAUST	3NLP	CoreNLP, SnowBall	McCCJ + SD	(UMass+Stanford as features)				GE	word clusters
2	UMass	1NLP	CoreNLP, SnowBall	McCCJ + SD	Joint, dual dec.+MIRA 1-best			-	GE	-
3	Stanford	3NLP	CoreNLP	McCCJ + SD	MaxEnt	Joint, MSTParser		-	GE	word clusters
4	ConcordU	2NLP	-	McCCJ + SD	dict	rules	rules	rules	-	triggers and hedge words
5	UTurku	1BI	Porter	McCCJ + SD	SVM	SVM	SVM	SVM	-	hedge words
6	PNNL	1CS, 1NLP, 2BI	Porter	Stanford	SVM	SVM	rules	-	GE	UMLS, triggers
7	PredX	1CS, 1NLP	LGP	LGP	dict	rules	rules	-	-	UMLS, triggers

Table 5: Participants and summary of system descriptions. Abbreviations: Trig./Arg./Group./Modif.=event trigger detection/argument detection/argument grouping/modification detection, BI=Bioinformatician, NLP=Natural Language Processing researcher, CS=Computer scientist, CoreNLP=Stanford CoreNLP, Porter=Porter stemmer, Snowball=Snowball stemmer McCCJ=McClosky-Charniak-Johnson parser, LGP=Link Grammar Parser, SD=Stanford Dependency conversion, UMLS=UMLS resources (e.g. lexicon, metamap)

ST'09 participation. In system design choices, we note an indication of increased use of joint models as opposed to pure pipeline designs, with the three highest-ranking systems involving a joint model.

Several participants compiled dictionaries of event trigger words and two dictionaries of hedge words from the data. Four teams, including the three top-ranking, used the GE task corpus as supplementary material, indicating that the GE annotations are largely compatible with ID ones (see detailed results below). This is encouraging for future applications of the event extraction approach: as manual annotation requires considerable effort and time, the ability to use existing annotations is important for the feasibility of adaptation of the approach to new domains.

While several participants made use of supporting syntactic analyses provided by the organizers (Stenetorp et al., 2011), none applied the analyses for supporting tasks, such as coreference or entity relation extraction results – at least in cases due to time constraints (Kilicoglu and Bergler, 2011).

5.2 Evaluation results

Table 6 presents the primary results by event type, and Table 7 summarizes these results. The full task requires the extraction of additional arguments and event modifications and involves multiple novel challenges from previously addressed domain tasks including a new subdomain, full-text documents, several new entity types and a new event category.

Team	recall	prec.	F-score
FAUST	48.03	65.97	55.59
UMass	46.92	62.02	53.42
Stanford	46.30	55.86	50.63
ConcordU	49.00	40.27	44.21
UTurku	37.85	48.62	42.57
PNNL	27.75	52.36	36.27
PredX	22.56	35.18	27.49

Table 7: Primary evaluation results.

Nevertheless, extraction performance for the top systems is comparable to the state-of-the-art results for the established BioNLP ST'09 task (Miwa et al., 2010) as well as its repetition as the 2011 GE task (Kim et al., 2011b), where the highest overall result for the primary evaluation criteria was also 56% F-score for the FAUST system (Riedel et al., 2011). This result is encouraging regarding the ability of the extraction approach and methods to generalize to new domains as well as their applicability specifically to texts on the molecular mechanisms of infectious diseases.

We note that there is substantial variation in the relative performance of systems for different entity types. For example, Stanford (McClosky et al., 2011) has relatively low performance for simple events but achieves the highest result for PROCESS, while UTurku (Björne and Salakoski, 2011) results show roughly the reverse. This suggests further potential for improvement from system combinations.

	FAUST	UMass	Stanford	ConcordU	UTurku	PNNL	PredX	Size
GENE EXPRESSION	**70.68**	66.43	54.00	56.57	64.88	53.33	0.00	512
TRANSCRIPTION	69.66	68.24	60.00	**70.89**	57.14	0.00	53.85	77
PROTEIN CATABOLISM	**75.00**	72.73	20.00	66.67	33.33	11.76	0.00	33
PHOSPHORYLATION	64.00	**66.67**	40.00	54.55	60.61	64.29	40.00	69
LOCALIZATION	33.33	14.29	31.58	20.00	**66.67**	20.69	0.00	49
Simple event total	***68.47***	*63.55*	*52.72*	*56.78*	*62.67*	*43.87*	*18.18*	*740*
BINDING	31.30	34.62	23.44	**40.00**	22.22	20.00	28.28	156
PROCESS	65.69	62.26	**73.57**	67.17	41.57	51.04	53.27	901
Non-regulation total	***63.78***	*60.68*	*63.59*	*62.43*	*46.39*	*47.34*	*43.65*	*1797*
REGULATION	**35.44**	30.49	17.67	19.43	22.96	0.00	2.16	267
POSITIVE REGULATION	47.50	**49.49**	34.78	23.41	41.28	24.60	21.02	455
NEGATIVE REGULATION	58.86	**60.45**	44.44	47.96	52.11	25.70	9.49	260
Regulation total	***47.07***	*46.65*	*33.02*	*28.87*	*39.49*	*18.45*	*9.71*	*982*
Subtotal	***57.28***	*55.03*	*52.09*	*46.60*	*43.33*	*37.53*	*28.38*	*2779*
NEGATION	0.00	0.00	0.00	22.92	**32.91**	0.00	0.00	96
SPECULATION	0.00	0.00	0.00	3.23	**15.00**	0.00	0.00	44
Modification total	*0.00*	*0.00*	*0.00*	*11.82*	***26.89***	*0.00*	*0.00*	*140*
Total	***55.59***	*53.42*	*50.63*	*44.21*	*42.57*	*36.27*	*27.49*	*2919*

Table 6: Primary evaluation F-scores by event type. The "size" column gives the number of annotations of each type in the given data (training+development). Best result for each type shown in bold.

The best performance for simple events and for PROCESS approaches or exceeds 70% F-score, arguably approaching a sufficient level for user-facing applications of the extraction technology. By contrast, BINDING and regulation events, found challenging in ST'09 and GE, remain problematic also in the ID task, with best overall performance below 50% F-score. Only two teams, UTurku and ConcordU (Kilicoglu and Bergler, 2011), attempted to extract event modifications, with somewhat limited performance. The difficulty of correct extraction of event modifications is related in part to the recursive nature of the problem (similarly as for nested regulation events): to extract a modification correctly, the modified event must also be extracted correctly. Further, only UTurku predicted any instances of secondary arguments. Thus, teams other than UTurku and ConcordU addressed only the core task extraction targets. With the exception of ConcordU, all systems clearly favor precision over recall (Table 7), in many cases having over 15% point higher precision than recall. This a a somewhat unexpected inversion, as the ConcordU system is one of the two rule-based in the task, an approach typically associated with high precision.

The five top-ranking systems participated also in the GE task (Kim et al., 2011b), which involves a subset of the ID extraction targets. This allows additional perspective into the relative performance of the systems. While there is a 13% point spread in overall results for the top five systems here, in GE all these systems achieved F-scores ranging between 50–56%. The results for FAUST, UMass and Stanford were similar in both tasks, while the ConcordU result was 6% points higher for GE and the UTurku result over 10% points higher for GE, ranking third after FAUST and UMass. These results suggest that while the FAUST and UMass systems in particular have some systematic (e.g. architectural) advantage at both tasks, much of the performance difference observed here between the top three systems and those of ConcordU and UTurku is due to strengths or weaknesses specific to ID. Possible weaknesses may relate to the treatment of multiple core entity types (vs. only PROTEIN in GE) or challenges related to nested entity annotations (not appearing in GE). A possible ID-specific strength of the three top-ranking systems is the use of GE data for training: Riedel and McCallum (2011) report an estimated 7% point improvement and McClosky et al. (2011) a 3% point improvement from use of this data; McGrath et al. (2011) estimate a 1% point improvement from direct corpus combination. The integration strategies applied in training these systems

Team	recall	prec.	F-score	Δ
FAUST	50.62	66.06	57.32	1.73
UMass	49.45	62.11	55.06	1.64
Stanford	48.87	56.03	52.20	1.57
ConcordU	50.77	43.25	46.71	2.50
UTurku	38.79	49.35	43.44	0.87
PNNL	29.36	52.62	37.69	1.42
PredX	23.67	35.18	28.30	0.81

Table 8: Core task evaluation results. The Δ column gives the F-score difference to the corresponding full task (primary) result.

could potentially be applied also with other systems, an experiment that could further clarify the relative strengths of the various systems. The top-ranking five systems all participated also in the EPI task (Ohta et al., 2011), for which UTurku ranked first with FAUST having comparable performance for the core task. While this supports the conclusion that ID performance differences do not reflect a simple universal ranking of the systems, due to many substantial differences between the ID and EPI setups it is not straightforward to identify specific reasons for relative differences to performance at EPI.

Table 8 summarizes the core task results. There are only modest and largely consistent differences to the corresponding full task results, reflecting in part the relative sparseness of additional arguments: in the training data, for example, only approximately 3% of instances of event types that can potentially take additional arguments had at least one additional argument. While event modifications represent a further 4% of full task extraction targets not required for the core task, the overall low extraction performance for additional arguments and modifications limits the practical effect of these annotation categories on the performance difference between systems addressing only the core targets and those addressing the full task.

6 Discussion and Conclusions

We have presented the preparation, resources, results and analysis of the Infectious Diseases (ID) task of the BioNLP Shared Task 2011. A corpus of 30 full-text publications on the two-component systems subdomain of infectious diseases was created for the task in a collaboration of event annotation and domain experts, adapting and extending the BioNLP'09 Shared Task (ST'09) event representation to the domain.

Seven teams submitted final results to the ID task. Despite the novel challenges of full papers, four new entity types, extension of event scopes and the introduction of a new event category for high-level processes, the highest results for the full ID task were comparable to the state-of-the-art performance on the established ST'09 data, showing that the event extraction approach and present systems generalize well and demonstrating the feasibility of event extraction for the infectious diseases domain. Analysis of results suggested further opportunities for improving extraction performance by combining the strengths of various systems and the use of other event resources.

The task design takes into account the needs of supporting practical applications, and its results and findings will be adopted in future development of the Pathosystems Resource Integration Center[4] (PATRIC). Specifically, PATRIC will combine domain named entity recognition and event extraction to mine the virulence factor literature and integrate the results with literature search and retrieval services, protein feature analysis, and systems such as Disease View.[5] Present and future advances at the ID event extraction task can thus assist biologists in efforts of substantial public health interest.

The ID task will be continued as an open shared task challenge with data, supporting resources, and evaluation tools freely available from the shared task site, `http://sites.google.com/site/bionlpst/`.

Acknowledgments

This work was supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan). This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C, awarded to BWS Sobral.

[4] `http://patricbrc.org`

[5] See for example `http://patricbrc.org/portal/portal/patric/DiseaseOverview?cType=taxon&cId=77643`

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.

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.

M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25:25–29.

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.

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.

Peter Corbett and Peter Murray-Rust. 2006. High-throughput identification of chemistry in life science texts. *Computational Life Sciences II*, pages 107–118.

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.

Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85+, February.

Yasuhiro Gotoh, Yoko Eguchi, Takafumi Watanabe, Sho Okamoto, Akihiro Doi, and Ryutaro Utsumi. 2010. Two-component signal transduction as potential drug targets in pathogenic bacteria. *Current Opinion in Microbiology*, 13(2):232–239. Cell regulation.

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, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier, editors. 2004. *Introduction to the bio-entity recognition task at JNLPBA*, Geneva, Switzerland.

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, 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.

Tino Krell, Jess Lacal, Andreas Busch, Hortencia Silva-Jimnez, Mara-Eugenia Guazzaroni, and Juan Luis Ramos. 2010. Bacterial sensor kinases: Diversity in the recognition of environmental signals. *Annual Review of Microbiology*, 64(1):539–559.

Thorsten Mascher, John D. Helmann, and Gottfried Unden. 2006. Stimulus perception in bacterial signal-transducing histidine kinases. *Microbiol. Mol. Biol. Rev.*, 70(4):910–938.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event extraction as dependency parsing for bionlp 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

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

Liam McGrath, Kelly Domico, Courtney Corley, and Bobbie-Jo Webb-Robertson. 2011. Complex biological event extraction from full text using signatures of linguistic and semantic features. In *Proceedings of*

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

Makoto Miwa, Sampo Pyysalo, Tadayoshi Hara, and Jun'ichi Tsujii. 2010. Evaluating dependency representation for event extraction. In *Proceedings of COLING'10*, pages 779–787.

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, 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, 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. 2011. Annotation guidelines for infectious diseases event corpus. Technical report, Tsujii Laboratory, University of Tokyo. To appear.

Sebastian Riedel and Andrew McCallum. 2011. Robust biomedical event extraction with dual decomposition and minimal domain adaptation. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, and Chris Manning. 2011. Model combination for event extraction in bionlp 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Yutaka Sasaki, Yoshimasa Tsuruoka, John McNaught, and Sophia Ananiadou. 2008. How to make the most of NE dictionaries in statistical NER. *BMC bioinformatics*, 9 Suppl 11.

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.

Peter Thomason and Rob Kay. 2000. Eukaryotic signal transduction via histidine-aspartate phosphorelay. *J Cell Sci*, 113(18):3141–3150.

Yue Wang, Jin-Dong Kim, Rune Sætre, Sampo Pyysalo, and Jun'ichi Tsujii. 2009. Investigating heterogeneous protein annotations toward cross-corpora utilization. *BMC Bioinformatics*, 10(403).

Chunxia Wang, Jocelyn Kemp, Isabel O. Da Fonseca, Raymie C. Equi, Xiaoyan Sheng, Trevor C. Charles, and Bruno W. S. Sobral. 2010. Sinorhizobium meliloti 1021 loss-of-function deletion mutation in chvi and its phenotypic characteristics. *Molecular Plant-Microbe Interactions*, 23(2):153–160.