Towards Event Extraction from Full Texts on Infectious Diseases

Sampo Pyysalo[*] **Tomoko Ohta**[*] **Han-Cheol Cho**[*] **Dan Sullivan**[†]
Chunhong Mao[†] **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
{smp,okap,priancho,tsujii}@is.s.u-tokyo.ac.jp
{dsulliva,cmao,sobral}@vbi.vt.edu
Sophia.Ananiadou@manchester.ac.uk

Abstract

Event extraction approaches based on expressive structured representations of extracted information have been a significant focus of research in recent biomedical natural language processing studies. However, event extraction efforts have so far been limited to publication abstracts, with most studies further considering only the specific transcription factor-related subdomain of molecular biology of the GENIA corpus. To establish the broader relevance of the event extraction approach and proposed methods, it is necessary to expand on these constraints. In this study, we propose an adaptation of the event extraction approach to a subdomain related to infectious diseases and present analysis and initial experiments on the feasibility of event extraction from domain full text publications.

1 Introduction

For most of the previous decade, biomedical Information Extraction (IE) efforts have focused primarily on tasks that allow extracted information to be represented as simple pairs of related entities. This representation is applicable to many IE targets of interest, such as gene-disease associations (Chun et al., 2006) and protein-protein interactions (Nédellec, 2005; Krallinger et al., 2007). However, it has limited applicability to advanced applications such as semantic search, Gene Ontology term annotation, and pathway extraction, tasks for which and relatively few resources or systems (e.g. (Rzhetsky et al., 2004)) have been introduced. A number of recent studies have proposed more expressive representations of extracted information, introducing resources supporting advanced IE approaches (Pyysalo et al., 2007; Kim et al., 2008; Thompson et al., 2009; Ananiadou et al., 2010a). A significant step in the development of domain IE methods capable of extracting this class of representations was taken in the BioNLP'09 shared task on event extraction, where 24 teams participated in an IE task setting requiring the extraction of structured representations of multi-participant biological events of several types (Kim et al., 2009).

While the introduction of structured event extraction resources and methods has notably advanced the state of the art in biomedical IE representations, the focus of event extraction studies carries other limitations frequently encountered in domain IE efforts. Specifically, resources annotated for biomedical events contain exclusively texts from publication abstracts, typically further drawn from small subdomains of molecular biology. These choices constrain not only the types of texts but also the types of events considered, restricting the applicability of event extraction. This paper presents results from one ongoing effort to extend an event extraction approach over these boundaries, toward event extraction from full text documents in the domain of infectious diseases.

In this study, we consider the subdomain related to Type IV secretion systems as a model subdomain of interest within the broad infectious diseases domain. Type IV secretion systems (T4SS) are mechanisms for transferring DNA and proteins across cellular boundaries. T4SS are found in a broad range of Bacteria and in some Archaea. These translocation systems enable gene transfer across cellular membranes thus contributing to the spread of antibiotic resistance and viru-

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 132–140,
Uppsala, Sweden, 15 July 2010.

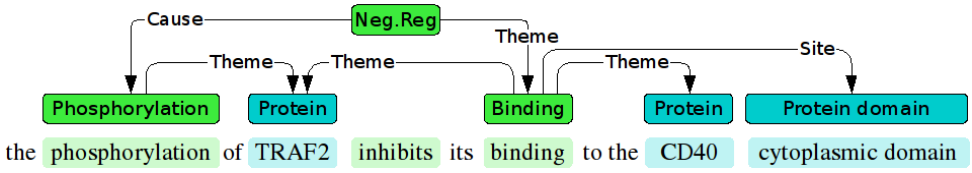

Figure 1: Event representation example. Inhibition of binding caused by phosphorylation is represented using three events. The shaded text background identifies the text bindings of the events and entities.

lence genes making them an especially important mechanism in infectious disease research (Juhas et al., 2008). Type IV secretion systems are found in plant pathogens, such as Agrobacterium tumefaciens, the cause of crown gall disease as well as in animal pathogens, such as Helicobacter pylori, a cause of severe gastric disease. The study of T4SS has been hampered by the lack of consistent terminology to describe genes and proteins associated with the translocation mechanism thus motivating the use of natural language processing techniques to enhance information retrieval and information extraction from relevant literature.

2 Event Extraction for the T4SS Domain

This section presents the application of an event extraction approach to the T4SS domain.

2.1 Event Extraction

We base our information extraction approach on the model introduced in the BioNLP'09 shared task on event extraction. Central to this approach is the event representation, which can capture the association of multiple participants in varying roles and numbers and treats events as primary objects of annotation, thus allowing events to be participants in other events. Further, both entities and events are text-bound, i.e. anchored to specific expressions in text (Figure 1).

The BioNLP'09 shared task defined nine event types and five argument types (roles): *Theme* specifies the core participant(s) that an event affects, *Cause* the cause of the the event, *Site* a specific domain or region on a participant involved in the event, and *ToLoc* and *AtLoc* locations associated with localization events (Table 1). Theme and Cause arguments may refer to either events or gene/gene product entities, and other arguments refer to other physical entities. The Theme argument is always mandatory, while others can be omitted when a relevant participant is not stated.

The event types were originally defined to capture statements of biologically relevant changes in the state of entities in a target subdomain involving transcription factors in human blood cells. In adapting the approach to new domains, some extension of the event types is expected to be necessary. By contrast, the argument types and the general design of the representation are intended to be general, and to maintain compatibility with existing systems we aim to avoid modifying these.

Event type	Args	Example
Gene expression	T	5-LOX is *coexpressed*
Transcription	T	IL-4 *transcription*
Protein catabolism	T	IkB-A *proteolysis*
Localization	T,L	*translocation* of STAT6
Phosphorylation	T,S	NF90 was *phosphorylated*
Binding	T+,S+	Nmi *interacts* with STAT
Regulation	T,C,S	IL-4 gene *control*
Positive regulation	T,C,S	IL-12 *induced* binding
Negative regulation	T,C,S	*suppressed* dimerization

Table 1: Event types targeted in the BioNLP'09 shared task and their arguments, with minimal examples of each event type. Arguments abbreviate for (T)heme, (C)ause, (S)ite and L for ToLoc/AtLoc, with "+" identifying arguments than can occur multiple times. The expression marked as triggering the event shown in italics.

2.2 T4SS Domain

A corpus of full-text publications relating to the T4SS subdomain of the infectious diseases domain annotated for biological entities and terms of interest to domain experts was recently introduced by (Ananiadou et al., 2010b). In the present study, we use this corpus as a reference standard defining domain information needs. In the following we briefly describe the corpus annotation and the view it provides of the domain.

The T4SS corpus annotation covers four classes of tagged entities and terms: Bacteria, Cellular components, Biological Processes, and Molecular functions. The latter three correspond to the three Gene Ontology (GO) (Ashburner et al., 2000) top-level sub-ontologies, and terms of these types were annotated with reference to both GO and relevance to the interests of domain experts, with guidelines

Bacterium		Cell component		Biological process		Molecular function	
A. tumefaciens	32.7%	T4SS	5.2%	virulence	14.1%	nucleotide-binding	20.3%
H. pylori	20.0%	Ti plasmid	5.1%	conjugation	7.9%	ATPase activity	17.3%
L. pneumophila	16.3%	outer membrane	4.2%	localization	6.1%	NTP-binding	14.7%
E. coli	12.3%	membrane	3.5%	nuclear import	5.8%	ATP-binding	12.2%
B. pertussis	3.0%	genome	3.4%	transfer	5.1%	DNA-binding	9.1%

Table 2: Most frequently tagged terms (after normalization) and their relative frequencies of all tagged entities of each of the four types annotated in the T4SS corpus.

Type	Annotations
Bacteria	529
Cellular component	2237
Biological process	1873
Molecular function	197

Table 3: Statistics for the existing T4SS corpus annotation.

requiring that marked terms be both found in GO and associated with T4SS. These constraints assure that the corpus is relevant to the information needs of biologists working in the domain and that it can be used as a reference for the study of automatic GO annotation. In the work introducing the corpus, the task of automatic GO annotation was studied as facilitating improved information access, such as advanced search functionality: GO annotation can allow for search by semantic classes or co-occurrences of terms of specified classes. The event approach considered in this study further extends on these opportunities in introducing a model allowing e.g. search by specific associations of the concepts of interest.

The previously created annotation of the T4SS corpus covers 27 full text publications totaling 15143 pseudo-sentences (text sentences plus table rows, references, etc.) and 244942 tokens.[1] A total of nearly 5000 entities and terms are annotated in these documents; Table 2 shows the most frequently tagged terms of each type after basic normalization of different surface forms, and Table 3 gives the per-class statistics. Domain characteristics are clearly identifiable in the first three tagged types, showing disease-related bacteria, their major cellular components, and processes related to movement, reproduction and infection. The last term type is dominated by somewhat more generic binding-type molecular functions.

In addition to the four annotated types it was recognized during the original T4SS corpus annotation that genes and gene products are centrally important for domain information needs, but their annotation was deferred to focus on novel categories. As part of the present study, we introduce annotation for gene/gene product (GGP) mentions (Section 3.2), and in the following discussion of applying an event extraction approach to the domain the availability of this class annotation as an additional category is assumed.

2.3 Adaptation of the Event Model

The event model involves two primary categories of representation: physical entities such as genes and proteins are elementary (non-structured) and their mentions annotated as typed spans of text,[2] and events and processes ("things that happen") are represented using the structured event representation described in Section 2.1. This division applies straightforwardly to the T4SS annotations, suggesting an approach where bacteria and cell components retain their simple tagged-term representation and the biological processes and molecular functions are given an event representation. In the following, we first analyze correspondences between the latter two classes and BioNLP'09 shared task events, and then proceed to study the event arguments and their roles as steps toward a complete event model for the domain.

Molecular functions, the smallest class tagged in the T4SS corpus, are highly uniform: almost 75% involve binding, immediately suggesting representation using the Binding class of events defined in the applied event extraction model. The remaining functions are *ATPase activity*, together with its exact GO synonyms (e.g. *ATP hydrolase activity*) accounting for 19% of the terms, the general type *hydrolysis* (4.5%), and a small number of rare other functions. While these have no correspondence with previously defined event types,

[1] While the document count is modest compared to that of abstract-based corpora, we estimate that in terms of the amount of text (tokens) the corpus corresponds to over 1000 abstracts, comparable in size to e.g. the GENIA event corpus (Kim et al., 2008).

[2] Normalization identifying e.g. the Uniprot entry corresponding to a protein mention may also be necessary, but here excluded from consideration an independent issue.

Class	Category	Freq
Location	Transfer	27.6%
	Localization	15.6%
	Import/export	14.5%
High-level process	Virulence	14.1%
	Assembly	8.7%
	Conjugation	8.3%
	Secretion	8.1%
(Other)		1.8%

Table 4: Categorization of T4SS corpus biological processes and relative frequency of mentions of each category of the total tagged.

their low overall occurrence counts make them of secondary interest as extraction targets.

The biological processes are considerably more diverse. To identify general categories, we performed a manual analysis of the 217 unique normalized terms annotated in the corpus as biological processes (Table 4). We find that the majority of the instances (58%) relate to location or movement. As related types of statements are annotated as Localization events in the applied model, we propose to apply this event type and differentiate between the specific subtypes on the basis of the event arguments. A further 39% are of categories that can be viewed as high-level processes. These are distinct from the events considered in the BioNLP'09 shared task in involving coarser-grained events and larger-scale participants than the GGP entities considered in the task: for example, conjugation occurs between bacteria, and virulence may involve a human host.

To analyze the role types and arguments characteristic of domain events, we annotated a small sample of tagged mentions for the most frequent types in the broad classification discussed above: Binding for Molecular function, Transfer for Location-related, and Virulence for High-level process. The statistics of the annotated 65 events are shown in Tables 5, 6 and 7. For Binding, we find that while an estimated 90% of events involve a GGP argument, the other participant of the binding is in all cases non-GGP, most frequently of Nucleotide type (e.g. NTP/ATP). While only GGP Binding arguments were considered in the shared task events, the argument structures are typical of multi-participant binding and this class of expressions are in scope of the original GENIA Event corpus annotation (Kim et al., 2008). Event annotations could thus potentially be derived from existing data. Localization event arguments show substantially greater variety and

Freq	Arguments
78%	Theme: GGP, Theme: Nucleotide
5.5%	Theme: GGP, Theme: DNA
5.5%	Theme: GGP, Theme: Sugar
5.5%	Theme: Protein family, Theme: DNA
5.5%	Theme: Protein, Theme: Nucleotide

Table 5: Binding event arguments.

Freq	Arguments
16%	Theme: DNA, From/To: Organism
16%	Theme: DNA
16%	Theme: Cell component
12%	Theme: DNA, To: Organism
8%	Theme: Protein family, From/To: Organism
4%	Theme: GGP
4%	Theme: GGP, To: Organism
4%	Theme: GGP, From: Organism
4%	Theme: Protein family, From: Organism
4%	Theme: Protein family
4%	Theme: Organism, To: Cell component
4%	Theme: DNA From: Organism, To: Cell component
4%	(no arguments)

Table 6: Localization (Transfer) event arguments.

Freq	Arguments
64%	Cause: GGP
16%	Theme:Organism, Cause: GGP
8%	Cause: Organism
8%	(no arguments)
4%	Cause: Protein family

Table 7: Process (Virulence) arguments.

some highly domain-specific argument combinations, largely focusing on DNA and Cell component (e.g. phagosome) transfer, frequently involving transfer between different organisms. While the participants are almost exclusively of types that do not appear in Localization events in existing annotations, the argument structures are standard and in our judgment reasonably capture the analyzed statements, supporting the applicability of the general approach. Finally, the argument analysis shown in Table 7 supports the previous tentative observation that the high-level biological processes are notably different from previously considered event types: for over 80% of these processes no overtly stated *Theme* could be identified. We take this to indicate that the themes – the core participants that the processes concern – are obvious in the discourse context and their overt expression would be redundant. (For example, in the context *virulence* obviously involves a host and *conjugation* involves bacteria.) By contrast, in the corpus the entities contributing to these processes are focused: a participant we have here analyzed as *Cause* is stated in over 90% of cases. This

	Sentences	Tokens
Abstracts	150	3789
Full texts	448	13375
Total	598	17164

Table 8: Statistics for the selected subcorpus.

	GGP	GGP/sentence
Abstracts	124	0.82
Full texts	394	0.88
Total	518	0.87

Table 9: Statistics for the GGP annotation.

novel pattern of event arguments suggests that the event model should be augmented to capture this category of high-level biological processes. Here, we propose an event representation for these processes that removes the requirement for a Theme and substitutes instead a mandatory Cause as the core argument. In the event annotation and experiments, we focus on this newly proposed class.

3 Annotation

This section describes the new annotation introduced for the T4SS corpus.

3.1 Text Selection

The creation of exhaustive manual annotation for the full T4SS corpus represents a considerable annotation effort. Due to resource limitations, for this study we did not attempt full-scope annotation but instead selected a representative subset of the corpus texts. We aimed to select texts that provide good coverage of the text variety in the T4SS corpus and can be freely redistributed for use in research. We first selected for annotation all corpus documents with at least a freely available PubMed abstract, excluding 3 documents. As the corpus only included a single freely redistributable Open Access paper, we extended full text selection to manuscripts freely available as XML/HTML (i.e. not only PDF) via PubMed Central. While these documents cannot be redistributed in full, their text can be reliably combined with standoff annotations to recreate the annotated corpus.

In selected full-text documents, to focus annotation efforts on sections most likely to contain reliable new information accessible to natural language processing methods, we further selected the publication body text, excluding figures and tables and their captions, and removed Methods and Discussion sections. We then removed artifacts such as page numbers and running heads and cleaned remaining errors from PDF conversion of the original documents. This selection produced a subcorpus of four full-text documents and 19 abstracts. The statistics for this corpus are shown in Table 8.

3.2 Gene/Gene Product Annotation

As gene and gene product entities are central to domain information needs and the core entities of the applied event extraction approach, we first introduced annotation for this entity class. We created manual GGP annotation following the annotation guidelines of the GENIA GGP Corpus (Ohta et al., 2009). As this corpus was the source of the gene/protein entity annotation provided as the basis of the BioNLP shared task on event extraction, adopting its annotation criteria assures compatibility with recently introduced event extraction methods. Briefly, the guidelines specify tagging for minimal continuous spans of specific gene/gene product names, without differentiating between DNA/RNA/protein. A "specific name" is understood to be a a name that allows a domain expert to identify the entry in a relevant database (Entrez gene/Uniprot) that the name refers to. Only GGP names are tagged, excluding descriptive references and the names of related entities such as complexes, families and domains.

The annotation was created on the basis of an initial tagging created by augmenting the output of the BANNER tagger (Leaman and Gonzalez, 2008) by dictionary- and regular expression-based tagging. This initial high-recall markup was then corrected by a human annotator. To confirm that the annotator had correctly identified subdomain GGPs and to check against possible error introduced through the machine-assisted tagging, we performed a further verification of the annotation on approx. 50% of the corpus sentences: we combined the machine- and human-tagged annotations as candidates, removed identifying information, and asked two domain experts to identify the correct GGPs. The two sets of independently produced judgments showed very high agreement: holding one set of judgments as the reference standard, the other would achieve an f-score of 97% under the criteria presented in Section 4.2. We note as one contributing factor to the high agreement that the domain has stable and systematically applied GGP naming criteria. The statistics of the full GGP annotation are shown in Table 9.

	Events	Event/sentence
Abstracts	15	0.1
Full texts	5	0.01
Additional	80	2.2
Total	100	0.16

Table 10: Statistics for the event annotation.

3.3 Event Annotation

Motivated by the analysis described in Section 2.3, we chose to focus on the novel category of associations of GGP entities in high-level processes. Specifically, we chose to study biological processes related to virulence, as these are the most frequent case in the corpus and prototypical of the domain. We adopted the GENIA Event corpus annotation guidelines (Kim et al., 2008), marking associations between specific GGPs and biological processes discussed in the text even when these are stated speculatively or their existence explicitly denied. As the analysis indicated this category of processes to typically involve a single stated participant in a fixed role, annotations were initially recorded as (GGP, process) pairs and later converted into an event representation.

During annotation, the number of annotated GGP associations with the targeted class of processes in the T4SS subcorpus was found to be too low to provide material for both training and testing a supervised learning-based event extraction approach. To extend the source data, we searched PubMed for cases where a known T4SS-related protein co-occurred with an expression known to relate to the targeted process class (e.g. *virulence*, *virulent*, *avirulent*, *non-virulent*) and annotated a further set of sentences from the search results for both GGPs and their process associations. As the properties of these additional examples could not be assured to correspond to those of the targeted domain texts, we used these annotations only as development and training data, performing evaluation on cases drawn from the T4SS subcorpus.

As the annotation target was novel, we performed two independent sets of judgments for all annotated cases, jointly resolving disagreements. Although initial agreement was low, for a final set of judgments we measured high agreement, corresponding to 93% f-score when holding one set of judgments as the gold standard. The statistics of the annotation are shown in Table 10. Annotations are sparse in the T4SS subcorpus and, as expected, very dense in the targeted additional data.

4 Experiments

4.1 Methods

For GGP tagging experiments, we applied a state-of-the-art tagger with default settings as reference and a custom tagger for adaptation experiments. As the reference tagger, we applied a recent release of BANNER (Leaman and Gonzalez, 2008) trained on the GENETAG corpus (Tanabe et al., 2005). The corpus is tagged for gene and protein-related entities and its texts drawn from a broad selection of PubMed abstracts. The current revision of the tagger[3] achieves an f-score of 86.4% on the corpus, competitive with the best result reported in the BioCreative II evaluation (Wilbur et al., 2007), 87.2%. The custom tagger[4] follows the design of BANNER in both the choice of Conditional Random Fields (Lafferty et al., 2001) as the applied learning method and the basic feature design, but as a key extension can further adopt features from external dictionaries as both positive and negative indicators of tagged entities. Tagging experiments were performed using a document-level 50/50 split of the GGP-annotated subcorpus.

For event extraction, we applied an adaptation of the approach of the top-ranking system in the BioNLP'09 shared task (Björne et al., 2009): all sentences in the input text were parsed with the McClosky-Charniak (2008) parser and the resulting phrase structure analyses then converted into the Stanford Dependency representation using conversion included in the Stanford NLP tools (de Marneffe et al., 2006). Trigger recognition was performed with a simple regular expression-based tagger covering standard surface form variation. Edge detection was performed using a supervised machine learning approach, applying the LibSVM (Chang and Lin, 2001) Support Vector Machine implementation with a linear kernel and the feature representation of Björne et al˙ (2009), building largely around the shortest dependency path connecting a detected trigger with a candidate participant. The SVM regularization parameter was selected by a sparse search of the parameter space with evaluation using cross-validation on the training set. As the class of events targeted for extraction in this study are of a highly restricted type, each taking only of a single mandatory Cause argument, the construction of events from detected

[3] http://banner.sourceforge.net

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/NERsuite/

	Precision	Recall	F-score
Abstracts	68.1%	89.5%	77.3%
Full texts	56.9%	80.7%	66.7%
Total	59.4%	82.8%	69.2%

Table 11: Initial GGP tagging results.

triggers and edges could be implemented as a simple deterministic rule.

4.2 Evaluation Criteria

For evaluating the performance of the taggers we apply a relaxed matching criterion that accepts a match between an automatically tagged and a gold standard entity if the two overlap at least in part. This relaxation is adopted to focus on true tagging errors. The GENETAG entity span guidelines differ from the GENIA GGP guidelines adopted here in allowing the inclusion of e.g. head nouns when names appear in modifier position, while the annotation guidelines applied here require marking only the minimal name.[5] When applying strict matching criteria, a substantial number of errors may trace back to minor boundary differences (Wang et al., 2009), which we consider of secondary interest to spurious or missing tags. Overall results are microaverages, that is, precision, recall and f-score are calculated from the sum of true positive etc. counts over individual documents.

For event extraction, we applied the BioNLP'09 shared task event extraction criteria (Kim et al., 2009) with one key change: to make it possible to evaluate the extraction of the high-level process participants, we removed the requirement that all events must define a Theme as their core argument.

4.3 Gene/Gene Product Tagging

The initial GGP tagging results using BANNER are shown in Table 11. We find that even for the relaxed overlap matching criterion, the f-score is nearly 10% points lower than reported on GENETAG in the evaluation on abstracts. For full texts, performance is lower yet by a further 10% points. In both cases, the primary problem is the poor precision of the tagger, indicating that many non-GGPs are spuriously tagged.

To determine common sources of error, we performed a manual analysis of 100 randomly selected falsely tagged strings (Table 12). We find

Category	Freq	Examples
GGP family or group	34%	VirB, tmRNA genes
Figure/table	26%	Fig. 1B, Table 1
Cell component	10%	T4SS, ER vacuole
Species/strain	9%	E. coli, A348deltaB4.5
Misc.	9%	step D, Protocol S1
GGP domain or region	4%	Pfam domain
(Other)	8%	TrIP, LGT

Table 12: Common sources of false positives in GGP tagging.

	Precision	Recall	F-score
Abstracts	90.5%	95.7%	93.1%
Full texts	90.0%	93.2%	91.6%
Total	90.1%	93.8%	91.9%

Table 13: GGP tagging results with domain adaptation.

that the most frequent category consists of cases that are arguably correct by GENETAG annotation criteria, which allow named protein families of groups to be tagged. A similar argument can be made for domains or regions. Perhaps not surprisingly, a large number of false positives relate to features common in full texts but missing from the abstracts on which the tagger was trained, such as figure and table references. Finally, systematic errors are made for entities belonging to other categories such as named cell components or species.

To address these issues, we applied a domain-adapted custom tagger that largely replicates the features of BANNER, further integrating information from the UMLS Metathesaurus,[6] which provides a large dictionary containing terms covering 135 different semantic classes, and a custom dictionary of 1081 domain GGP names, compiled by (Ananiadou et al., 2010b). The non-GGP UMLS Metathesaurus terms provided negative indicators for reducing spurious taggings, and the custom dictionary positive indicators. Finally, we augmented the GENETAG training data with 10 copies[7] of the training half of the T4SS GGP corpus as in-domain training data.

Table 13 shows the results with the domain-adapted tagger. We find dramatically improved performance for both abstracts and full texts, showing results competitive with the state of the art performance on GENETAG (Wilbur et al., 2007). Thus, while the performance of an unadapted tagger falls short of both results reported

[5] GENETAG annotations include e.g. *human ets-1 protein*, whereas the guidelines applied here would require marking only *ets-1*.

[6] http://www.nlm.nih.gov/research/umls/

[7] As the GENETAG corpus is considerably larger than the T4SS GGP corpus, replication was used to assure that sufficient weight is given to the in-domain data in training.

	Precision	Recall	F-score
Co-occurrence	65%	100%	78%
Machine learning	81%	85%	83%

Table 14: Event extraction results.

on GENETAG and levels necessary for practical application, adaptation addressing common sources of error through the adoption of general and custom dictionaries and the use of a small set of in-domain training data was successful in addressing these issues. The performance of the adapted tagger is notably high given the modest size of the in-domain data, perhaps again reflecting the consistent GGP naming conventions of the subdomain.

4.4 Event Extraction

We performed an event extraction experiment following the training and test split described in Section 3.3. Table 14 shows the results of the applied machine learning-based method contrasted with a co-occurrence baseline replacing the edge detection with a rule that extracts a Cause edge for all trigger-GGP combinations co-occurring within sentence scope. This approach achieves 100% recall as the test data was found to only contain events where the arguments are stated in the same sentence as the trigger.

The results show that the machine learning approach achieves very high performance, matching the best results reported for any single event type in the BioNLP'09 shared task (Kim et al., 2009). The very high co-occurrence baseline result suggests that the high performance largely reflects the relative simplicity of the task. With respect to the baseline result, the machine-learning approach achieves a 21% relative reduction in error.

While this experiment is limited in both scope and scale, it suggests that the event extraction approach can be beneficially applied to detect domain events represented by novel argument structures. As a demonstration of feasibility the result is encouraging for both the applicability of event extraction to this specific new domain and for the adaptability of the approach to new domains in general.

5 Discussion and Conclusions

We have presented a study of the adaptation of an event extraction approach to the T4SS subdomain as a step toward the introduction of event extraction to the broader infectious diseases domain. We applied a previously introduced corpus of subdomain full texts annotated for mentions of bacteria and terms from the three top-level Gene Ontology subontologies as a reference defining domain information needs to study how these can be met through the application of events defined in the BioNLP'09 Shared Task on event extraction. Analysis indicated that with minor revision of the arguments, the Binding and Localization event types could account for the majority of both biological processes and molecular functions of interest. We further identified a category of "high-level" biological processes such as the *virulence* process typical of the subdomain, which necessitated extension of the considered event extraction model.

Based on argument analysis, we proposed a representation for high-level processes in the event model that substitutes Cause for Theme as the core argument. We further produced annotation allowing an experiment on the extraction of the dominant category of virulence processes with gene/gene product (GGP) causes, annotating 518 GGP mentions and 100 associations between these and the processes. Experiments indicated that with annotated in-domain resources both the GGP entities and their associations with processes could be extracted with high reliability.

In future work we will extend the model and annotation proposed in this paper to the broader infectious diseases domain, introducing annotated resources and extraction methods for advanced information access. All annotated resources introduced in this study are available from the GENIA project homepage.[8]

Acknowledgments

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan), the National Institutes of Health, grant number HHSN272200900040C, and the Joint Information Systems Committee (JISC, UK).

References

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

[8] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

Sophia Ananiadou, Dan Sullivan, Gina-Anne Levow, Joseph Gillespie, Chunhong Mao, Sampo Pyysalo, Jun'ichi Tsujii, and Bruno Sobral. 2010b. Named entity recognition for bacterial type IV secretion systems. (manuscript in review).

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, 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.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. 2006. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'06)*, pages 4–15.

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.

Mario Juhas, Derrick W. Crook, and Derek W. Hood. 2008. Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. *Cellular microbiology*, 10(12):2377–2386.

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.

Martin Krallinger, Florian Leitner, and Alfonso Valencia. 2007. Assessment of the Second BioCreative PPI task: Automatic Extraction of Protein-Protein Interactions. In L. Hirschman, M. Krallinger, and A. Valencia, editors, *Proceedings of Second BioCreative Challenge Evaluation Workshop*, pages 29–39.

John D. Lafferty, Andrew McCallum, and Fernando C . N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

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 and Eugene Charniak. 2008. Self-Training for Biomedical Parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT'08)*, pages 101–104.

Claire Nédellec. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge. In J. Cussens and C. Nédellec, editors, *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, pages 31–37.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, and Jun'ichi Tsujii. 2009. Incorporating GENETAG-style annotation to GENIA corpus. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 106–107, Boulder, Colorado. Association for Computational Linguistics.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Duboué, Wubin Weng, W. John Wilbur, Vasileios Hatzivassiloglou, and Carol Friedman. 2004. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53.

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

Paul Thompson, Syed Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1):349.

Yue Wang, Jin-Dong Kim, Rune Saetre, Sampo Pyysalo, and Jun'ichi Tsujii. 2009. Investigating heterogeneous protein annotations toward cross-corpora utilization. *BMC Bioinformatics*, 10(1):403.

John Wilbur, Lawrence Smith, and Lorraine Tanabe. 2007. BioCreative 2. Gene Mention Task. In L. Hirschman, M. Krallinger, and A. Valencia, editors, *Proceedings of Second BioCreative Challenge Evaluation Workshop*, pages 7–16.