Complex Biological Event Extraction from Full Text using Signatures of Linguistic and Semantic Features

Liam R. McGrath and **Kelly Domico** and **Courtney D. Corley** and **Bobbie-Jo Webb-Robertson**
Pacific Northwest National Laboratory
902 Battelle BLVD, PO BOX 999
Richland, WA 99352
{liam | kelly.domico | court | bj}@pnl.gov

Abstract

Building on technical advances from the BioNLP 2009 Shared Task Challenge, the 2011 challenge sets forth to generalize techniques to other complex biological event extraction tasks. In this paper, we present the implementation and evaluation of a signature-based machine-learning technique to predict events from full texts of infectious disease documents. Specifically, our approach uses novel signatures composed of traditional linguistic features and semantic knowledge to predict event triggers and their candidate arguments. Using a leave-one out analysis, we report the contribution of linguistic and shallow semantic features in the trigger prediction and candidate argument extraction. Lastly, we examine evaluations and posit causes for errors in our complex biological event extraction.

1 Introduction

The BioNLP 2009 Shared Task (Kim et al., 2009) was the first shared task to address fine-grained information extraction for the bio-molecular domain, by defining a task involving extraction of event types from the GENIA ontology. The BioNLP 2011 Shared Task ((Kim et al., 2011)) series generalized this defining a series of tasks involving more text types, domains and target event types. Among the tasks for the new series is the Infection Disease task, proposed and investigated by (Pyysalo et al., 2011; Pyysalo et al., 2010; Bjorne et al., 2010).

Like the other tasks for the BioNLP Shared Task series, the goal is to extract mentions of relevant events from biomedical publications. To extract an event, the event trigger and all arguments must be identified in the text by exact offset and typed according to a given set of event and argument classes (Miwa et al., 2010). Entity annotations are given for a set of entity types that fill many of the arguments.

Here we describe Pacific Northwest National Laboratory's (PNNL) submission to the BioNLP 2011 Infectious Disease shared task. We describe the approach and then discuss results, including an analysis of errors and contribution of various features.

2 Approach

Our system uses a signature-based machine-learning approach. The system is domain-independent, using a primary task description vocabulary and training data to learn the task, but domain resources can be incorporated as additional features when available, as described here. The approach can be broken down into 4 components: an automated annotation pipeline to provide the basis for features, classification-based trigger identification and argument identification components, and a post-processing component to apply semantic constraints. The UIMA framework[1] is used to integrate the components into a pipeline architecture.

2.1 Primary Tasks

A definition of the events to be extracted is used to define candidates for classification and post-process the results of the classification. First a list of domain-specific entity classes is given. Entities of

[1] http://uima.apache.org/

Proceedings of BioNLP Shared Task 2011 Workshop, pages 130–137,
Portland, Oregon, USA, 24 June, 2011.

Event Class	Arguments
Gene_expression	Theme(Protein|Regulon-operon)
Transcription	Theme(Protein|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)*
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)?
Process	Participant(core_entity)?

Table 1: Summary of the target events. Type restrictions on fillers of each argument type are shown in parenthesis. Multiplicity of each argument type is also marked (+ = one-to-many, ? = zero-to-one, * = zero-to-many, otherwise = one).

these classes are assumed to be annotated in the data, as is the case for the Infectious Disease task. Then, each event class is given, with a list of argument types for each. Each argument is marked with its multiplicity, indicating how many of this argument type is valid for each event, either: one – exactly one is required, one-to-many – one or more is required, zero-to-one – one is optional, and zero-to-many – one or many are optional. Also, restrictions on the classes of entities that can fill each argument are given, by listing: one or more class names – indicating the valid domain-specific entity classes from the definition, core_entity – indicating that any domain-specific entity in the definition is valid, event – indicating that any event in the definition is valid, or entity – indicating that any span from the text is valid. Table 1 shows the summary of the event extraction tasks for the Infectious Disease track.

2.2 Annotation

Linguistic and domain annotations are automatically applied to the document to be used for trigger and argument identification in framing the tasks for classification and generating features for each instance. Linguistic annotations include sentence splits, tokens, parts of speech, tree parses, typed dependencies (deMarneffe et al., 2006; MacKinlay et al., 2009), and stems. For the Infectious Disease task, the parses from the Stanford Parser (Klein and Manning, 2003) provided by the Supporting Analysis (Stenetorp et al., 2011) was used to obtain all of these linguistic annotations, except for the stems, which were obtained from the Porter Stemmer (van Rijsbergen et al., 1980).

For the Infectious Disease task, two sets of domain specific annotations are included: known trigger words for each event class and semantic tags from the Unified Medical Language System (UMLS) (Bodenreider, 2004). Annotations for known trigger words are created using a dictionary of word stem-event class pairs created from annotated training data. An entry is created in the dictionary every time a new stem is seen as a trigger for an event class. When a word with one of these stems is seen during processing, it is annotated as a typical trigger word for that event class.

Semantic tags are calculating using MetaMap 2010 (Aronson and Lang, 2010). MetaMap provides semantic tags for terms in a document with up to three levels of specificity, from most to least specific: concept, type and group (Torii et al., 2011). Word sense disambiguation is used to identify the best tags for each term. For example, consider the tags identified by MetaMap for the phrase *Human peripheral B cells*:

Human
concept: Homo sapiens
type: Human
group: Living Beings
Peripheral
type: Spatial Concept
group: Concepts & Ideas
B-Cells
concept: B-Lymphocytes
type: Cell

group: Anatomy

In this example, semantic mappings were found for three terms: *Human*, *Peripheral* and *B-Cells*. *Human* and *B-Cells* were mapped to specific concepts, but *Peripheral* was mapped to a more general group.

Entities are also annotated at this point. For the Infectious Disease task, annotations for five entity types are given: Protein, Two-component system, Chemical, Organism, or Regulon/Operon.

2.3 Trigger Identification

Triggers are identified using an SVM classifier (Vapnik, 1995; Joachims, 1999). Candidate triggers are chosen from the words in the text by part-of-speech. Based on known triggers seen in the training data, all nouns, verbs, adjectives, prepositions and adverbs are selected as candidates. A binary model is trained for each event type, and candidate triggers are tested against each classifier.

The following features are used to classify candidate event triggers:

- **term** – the candidate trigger
- **stem** – the stem of the term
- **part of speech** – the part of speech of the term
- **capitalization** – capitalization of the term
- **punctuation** – individual features for the presence of different punctuation types
- **numerics** – the presence of a number in the term
- **ngrams** – 4-grams of characters from the term
- **known trigger types** – tags from list of known trigger terms for each event type
- **lexical context** – terms in the same sentence
- **syntactic dependencies** – the type and role (governor or dependent) of typed dependencies involving the trigger
- **semantic type** – type mapping from MetaMap
- **semantic group** – group mapping from MetaMap

For training data, both the Infectious Disease training set and the GENIA training set were used. Although the GENIA training set represents a different genre and is annotated with a slightly different vocabulary than the Infectious Disease task data, it is similar enough to provide some beneficial supervision. The Infectious Disease training data is relatively small at 154 documents so including the larger GENIA training set at 910 documents results in a much more larger training set. Testing on the Infectious Disease development data, a 1 point improvement in fscore in overall results is seen with the additional training data.

2.4 Argument Identification

Arguments are also identified using an SVM classifier. For each predicted trigger, candidate arguments are selected based on the argument types. For arguments that are restricted to being filled by some set of specific entity and event types, each annotated entity and predicted event is selected as a candidate. For arguments that can be filled by any span of text, each span corresponding to a constituent of the tree parse is selected as a candidate. Each pair of an event trigger and a candidate argument serves as an instance for the classification. A binary model is trained for each event type, and each pair is tested against each classifier.

Many of the features used are inspired by those used in semantic role labeling systems (Gildea and Jurafsky, 2002). Given an event trigger and a candidate argument, the following features are used to classify event arguments:

- **trigger type** – the predicted event type of the trigger
- **argument terms** – the text of the argument
- **argument type** – entity or event type annotation on the argument
- **argument super-type** – core entity or core argument
- **trigger and argument stems** – the stems of each
- **trigger and argument parts of speech** – the part of speech of each
- **parse tree path** – from the trigger to argument via least common ancestor in tree parse, as a list of phrase types
- **voice of sentence** – active or passive
- **trigger and argument partial paths** – from the trigger or argument to the least common ancestor in tree parse, as a list of phrase types

- **relative position of argument to trigger** – before or after
- **trigger sub-categorization** – representation of the phrase structure rule that describes the relationship between the trigger, its parent and its siblings.

The training data used is the same as for trigger identification: the Infectious Disease training set plus the Genia training set.

2.5 Post-processing

A post-processing component is used to turn output from the various classifiers into semantically valid output according to the target task. For each predicted trigger, the positive predictions for each argument model are collected, and the set is compared to the argument restrictions in the target task description.

For example, the types on argument predictions are compared to the argument restrictions in the target task, and non-conforming ones are dropped. Then the multiplicity of the arguments for each predicted event is checked against the task vocabulary. Where there were not sufficient positive argument predictions to make a full event, the best negative predictions from the model are tried. When a compliant set of arguments can not be created for a predicted event, it is dropped.

3 Results and Discussion

Results for the system on both the development data and the official test data for the task are shown in Table 2 and Table 5, respectively. For the development data, a system using gold-standard event triggers is included, to isolate the performance of argument identification. In all cases, the total fscore for non-regulation events were much higher than regulation events. On the official test data, the system performed the best in predicting Phosphorylation (fscore = 71.43), Gene Expression (fscore = 53.33) and Process events (fscore = 51.04), but was unable to find any Transcription and Regulation events. This is also evident in the results on the development data using predicted triggers; additionally, no matches were found for localization and binding events. The total fscore on the development data using gold triggers was 55.33, more than 13 points higher than when using predicted triggers. In the discussion that follows, we detail the importance of individual features and their contribution to evaluation fscores.

3.1 Feature Importance

The effect of each argument and trigger feature type on the Infectious Disease development data was determined using a leave-one-out approach. The argument and trigger feature effect results are shown in Table 3 and Table 4, respectively. In a series of experiments, each feature type is left out of the full feature set one-by-one. The difference in fscore between each of these systems and the full feature set system is the effect of the feature type; a high negative effect indicates a significant contribution to the system since the removal of the feature resulted in a lower fscore.

Features	fscore	effect
all features	41.66	
w/o argument terms	36.16	-5.50
w/o argument type	39.50	-2.16
w/o trigger partial path	40.65	-1.01
w/o argument part of speech	40.98	-0.68
w/o argument partial path	41.16	-0.50
w/o trigger sub-categorization	41.45	-0.21
w/o argument stem	41.48	-0.18
w/o argument super-type	41.63	-0.03
w/o trigger type	41.63	-0.03
w/o trigger part of speech	41.81	0.15
w/o trigger stem	41.81	0.15
w/o voice of sentence	41.85	0.19
w/o relative position	42.21	0.55
w/o parse tree path	42.67	1.01

Table 3: Effect of each argument feature type on Infectious Disease development data.

Within the argument feature set system, the parse tree path feature had a notable positive effect of 1.01. The features providing the greatest contribution were argument terms and argument type with effects of -5.50 and -2.16, respectively. Within the trigger feature set system, the lexical context and syntactic dependencies features showed the highest negative effect signifying positive contribution to the system. The text and known trigger types features showed a negative contribution to the system.

Event Class	Using Gold Triggers gold/ans./match	recall	prec.	fscore	Using Predicted Triggers gold/ans./match	recall	prec.	fscore
Gene_expression	134 / 110 / 100	74.63	90.00	81.60	134 / 132 / 85	64.18	64.39	64.29
Transcription	35 / 26 / 23	65.71	88.46	75.41	25 / 0 / 0	0.00	0.00	0.00
Protein_catabolism	0 / 0 / 0	0.00	0.00	0.00	0 / 0 / 0	0.00	0.00	0.00
Phosphorylation	13 / 13 / 13	100.00	100.00	100.00	13 / 14 / 13	100.00	92.86	96.30
Localization	1 / 1 / 0	0.00	0.00	0.00	1 / 10 / 0	0.00	0.00	0.00
Binding	17 / 6 / 0	0.00	0.00	0.00	17 / 3 / 0	0.00	0.00	0.00
Process	206 / 180 / 122	59.22	67.78	63.21	207 / 184 / 108	52.17	58.70	55.24
Regulation	81 / 61 / 20	24.69	32.79	28.17	80 / 0 / 0	0.00	0.00	0.00
Positive_regulation	113 / 91 / 36	31.86	39.56	35.29	113 / 42 / 13	11.50	30.95	16.77
Negative_regulation	90 / 71 / 32	35.56	45.07	39.75	90 / 42 / 11	12.22	26.19	16.67
TOTAL	690 / 559 / 346	50.14	61.72	55.33	680 / 427 / 230	33.97	53.86	41.66

Table 2: Results on Infectious Disease development data. The system is compared to a system using gold standard triggers to isolate performance of argument identification.

Features	fscore	effect
all features	41.66	
w/o lexical context	40.14	-1.52
w/o syntactic dependencies	40.28	-1.38
w/o ngrams	40.88	-0.78
w/o part of speech	41.48	-0.18
w/o capitalization	41.51	-0.15
w/o numerics	41.51	-0.15
w/o semantic group	41.55	-0.11
w/o punctuation	41.59	-0.07
w/o stem	41.74	0.08
w/o semantic type	41.82	0.16
w/o known trigger types	42.11	0.45
w/o text	42.31	0.65

Table 4: Effect of each trigger feature type on Infectious Disease development data.

3.2 Transcription and Regulation events

Lastly, we present representative examples of errors (e.g., false positive, false negative, poor recall) produced by our system in the Infectious Disease track core tasks. The discussion herein will cover evaluations where our system did not correctly predict (transcription and regulation) any events or partially predicted (binding and +/- regulation) event triggers and arguments. In the text examples that follow, triggers are underlined and arguments are italicized.

The following are transcription events from the document PMC1804205-02-Results-03 in the development data.

- In contrast to the phenotype of the pta ackA double mutant, *pbgP transcription* was reduced in the pmrD mutant (Fig. 3).
- Growth at pH 5.8 resulted in *pmrD transcript levels* that were approximately3.5-fold higher than in organisms grown at pH 7.7 (Fig. 4A).

In both the development and test data evaluations, our system did not predict any transcription events, resulting in a 0.0 fscore; however, the system achieved 75.41 fscore when the gold-standard triggers were provided to the evaluation. Because argument prediction performed well, the system will benefit most by improving transcription event trigger prediction.

The following are regulation events from the document PMC1804205-02-Results-01in the development data.

- . . . we grew Salmonella cells harbouring chromosomal lacZYA transcriptional fusions to the *PmrA-regulated* genes *pbgP*, *pmrC* and *ugd* (Wosten and Groisman, 1999) in N-minimal media buffered at pH 5.8 or 7.7.
- We determined that Chelex 100 was effective at chelating iron because expression of the *pmrA-independent* iron-repressed iroA gene . . .

Similar to the transcription task, our system did not predict any regulation events, resulting in a 0.0 fscore. Unlike transcription events though, our system performed poorly on both argument identification and trigger prediction. The system achieved a 28.17 fscore when gold-standard triggers were used

Event Class	gold	(match)	answer	(match)	recall	prec.	fscore
Gene_expression	152	80	148	80	52.63	54.05	53.33
Transcription	50	0	0	0	0.00	0.00	0.00
Protein_catabolism	5	1	12	1	20.00	8.33	11.76
Phosphorylation	16	10	12	10	62.50	83.33	71.43
Localization	7	4	22	4	57.14	18.18	27.59
Binding	56	7	14	7	12.50	50.00	20.00
Regulation	193	0	0	0	0.00	0.00	0.00
Positive_regulation	193	34	87	34	17.62	39.08	24.29
Negative_regulation	181	32	68	32	17.68	47.06	25.70
Process	516	234	401	234	45.35	58.35	51.04
TOTAL	1369	402	764	402	29.36	52.62	37.69

Table 5: Official results on Infectious Disease test data

in the evaluation. Hypotheses for poor performance on candidate argument prediction are addressed in the following sections.

We posit that false negative trigger identifications are due to the limited full text training data (i.e. transcription events) and the inability of our system to predict non-verb triggers (i.e. second transcription example above). The SVM classifier was unable to distinguish between true transcription event triggers and transcription-related terms and ultimately, did not predict any transcription event in the development or test evaluations. To improve transcription event prediction, immediate effort should focus on 1) providing additional training data (e.g., BioCreativečiteBioCreative) and 2) introduce a trigger word filter that defines a subset of event triggers that have the best hit rate in the corpus. The hit rate is the number of occurrences of the word in a sentence per event type, divided by the total count in the gold standard (Nguyen et al., 2010).

3.3 +/-Regulation and Binding

The following positive regulation event is from document PMC1874608-03-RESULTS-03 in the development data.

- Invasiveness for HEp-2 cells was reduced to 39.1% of the wild-type *level* by *mlc* mutation, whereas it was *increased* by 1.57-fold by *hilE* mutation (Figure 3B).

In the preceding example, our system correctly predicted the +regulation trigger and the theme hilE; however, the correct argument was a gene expression event, not the entity. Many errors in the positive and negative regulation events were of this type; the predicted argument was a theme and not an event.

Evaluation of our system's binding event predictions resulted in low recall (12.50 or 0.0) in the test and development evaluations. The proceeding binding events are from document PMC1874608-03-RESULTS-05 in the development data. In both of the examples, our system correctly predicted the trigger *binding*; however, no arguments were predicted. Evaluation on the development data with gold standard triggers also resulted in an fscore of 0.0; thus, further algorithm refinement is needed to improve binding scores.

- *Mlc* directly represses *hilE* by *binding* to the *P3 promoter*
- These results clearly demonstrate that *Mlc* can regulate directly the *hilE* P3 promoter by *binding* to the *promoter*.

The following binding event is from document PMC1874608-01-INTRODUCTION in the development data and is representative of errors across many of the tasks. Here, the trigger is correctly predicted; however, the candidate arguments did not match with the reference data. Upon closer look, the arguments were drawn from the entire sentence, rather than an independent clause. The syntactic parse feature was not sufficient to prevent over-predicting arguments for the trigger, a potential solution is to add the arguments syntactic dependency

to the trigger as a feature to the candidate argument selection.

- Using two-hybrid analysis, it has been shown that *HilE interacts* with *HilD*, which suggests that HilE represses **hilA** expression by inhibiting the activity of **HilD** through a protein-protein interaction (19,20).

4 Summary

This article reports Pacific Northwest National Laboratory's entry to the BioNLP Shared Task 2011 Infectious Disease track competition. Our system uses a signature-based machine-learning approach incorporating traditional linguistic features and shallow semantic concepts from NIH's METAMAP Thesaurus. We examine the contribution of each of the linguistic and semantic features to the overall fscore for our system. This approach performs well on gene expression, process and phosphorylation event prediction. Transcription, regulation and binding events each achieve low fscores and warrant further research to improve their effectiveness. Lastly, we present a performance analysis of the transcription, regulation and binding tasks. Future work to improve our system's performance could include pre-processing using simple patterns (Nguyen et al., 2010), information extraction from figure captions (Kim and Yu, 2011) and text-to-text event extraction. The last suggested improvement is to add semantic features to the candidate argument prediction algorithm in addition to using rich features, such as semantic roles (Torii et al., 2011).

Acknowledgements

The authors thank the Signature Discovery Initiative, part of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle for the U.S. Department of Energy under contract DE-ACO5-76RLO 1830.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3):229–36, May.

J Bjorne, F Ginter, S Pyysalo, J Tsujii, and T Salakoski. 2010. Complex event extraction at pubmed scale. *Bioinformatics*, 26(12):i382–i390, Jun.

O. Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267.

M.C. deMarneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

T. Joachims. 1999. Making large scale SVM learning practical. *Advances in Kernel Methods – Support Vector Learning*.

Daehyun Kim and Hong Yu. 2011. Figure text extraction in biomedical literature. *PLoS ONE*, 6(1):e15338, Jan.

JD Kim, T Ohta, S Pyysalo, Y Kano, and J Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. *Proceedings of the Workshop on BioNLP: Shared Task*, pages 1–9.

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

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

A MacKinlay, D Martinez, and T Baldwin. 2009. Biomedical event annotation with crfs and precision grammars. *Proceedings of the Workshop on BioNLP: Shared Task*, pages 77–85.

Makoto Miwa, Rune Saetre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *J. Bioinform. Comput. Biol.*, 8(1):131–46, Feb.

Quang Long Nguyen, Domonkos Tikk, and Ulf Leser. 2010. Simple tricks for improving pattern-based information extraction from the biomedical literature. *J Biomed Semantics*, 1(1):9, Jan.

S. Pyysalo, T. Ohta, H.C. Cho, D. Sullivan, C. Mao, B. Sobral, J. Tsujii, and S. Ananiadou. 2010. Towards event extraction from full texts on infectious diseases. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 132–140. Association for Computational Linguistics.

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

Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. BioNLP Shared Task 2011: Supporting Resources. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, Portland, Oregon, June. Association for Computational Linguistics.

Manabu Torii, Lanlan Yin, Thang Nguyen, Chand T Mazumdar, Hongfang Liu, David M Hartley, and Noele P Nelson. 2011. An exploratory study of a text classification framework for internet-based surveillance of emerging epidemics. *International Journal of Medical Informatics*, 80(1):56–66, Jan.

C.J. van Rijsbergen, S.E. Robertson, and M.F. Porter. 1980. New models in probabilistic information retrieval.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.