

# Event Frame Extraction Based on a Gene Regulation Corpus

Yutaka Sasaki<sup>1</sup> Paul Thompson<sup>1</sup> Philip Cotter<sup>1</sup> John McNaught<sup>1,2</sup>  
Sophia Ananiadou<sup>1,2</sup>

<sup>1</sup> School of Computer Science, University of Manchester

<sup>2</sup> National Centre for Text Mining

MIB, 131 Princess Street, Manchester, M1 7DN, United Kingdom

Yutaka.Sasaki@manchester.ac.uk

## Abstract

This paper describes the supervised acquisition of semantic event frames based on a corpus of biomedical abstracts, in which the biological process of *E. coli gene regulation* has been linguistically annotated by a group of biologists in the EC research project "BOOTStrep". Gene regulation is one of the rapidly advancing areas for which information extraction could boost research. Event frames are an essential linguistic resource for extraction of information from biological literature. This paper presents a specification for linguistic-level annotation of gene regulation events, followed by novel methods of automatic event frame extraction from text. The event frame extraction performance has been evaluated with 10-fold cross validation. The experimental results show that a precision of nearly 50% and a recall of around 20% are achieved. Since the goal of this paper is event frame extraction, rather than event instance extraction, the issue of low recall could be solved by applying the methods to a larger-scale corpus.

## 1 Introduction

This paper describes the automatic extraction of linguistic event frames based on a corpus of MEDLINE abstracts that has been annotated with *gene regulation* events by a group of do-

main experts. Annotation is centred on both verbs and nominalised verbs that describe relevant events. For each event, semantic arguments that occur within the same sentence are marked and labelled with *semantic roles* and *named entity (NE)* types.

The focus of the paper is the extraction of *event frames* on the basis of the annotated corpus using machine learning techniques. Event frames are linguistic specifications concerning the behaviour of verbs and nominalised verbs, in terms of the number and types of semantic arguments with which they typically co-occur in texts. Our eventual goal is to exploit such information to improve information extraction. *Event frame extraction is different to event instance extraction (or template filling)*. Our event frames are destined for incorporation in the BOOTStrep *BioLexicon* to support identification of relevant event instances and discovery of event instance participants by NLP systems.

## 2 Background

There are several well-established, large-scale repositories of semantic frames for general language, e.g., VerbNet (Kipper-Schuler, 2005), PropBank (Palmer et al., 2005) and FrameNet (Rupenhoffer et al, 2006). These all aim to characterise verb behaviour in terms of the semantic arguments with which verbs occur but differ in how they represent semantic arguments and groupings of verbs.

In VerbNet, the semantic roles of arguments come from frame-independent roles, e.g. *Agent*, *Patient*, *Location* and *Instrument*.

In contrast, PropBank and FrameNet use a mixture of role types: some are common amongst a number of frames; others are specific to particular frames.

Whilst FrameNet and VerbNet differ in their treatment of semantic roles, they both specify

---

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (<http://creativecommons.org/licenses/by-nc-sa/3.0/>). Some rights reserved.

Table 1. Semantic Roles

Role Name	Description	Example ( <b>bold</b> = semantic argument, <i>italics</i> = focussed verb)
AGENT	Drives/instigates event	<b>The narL gene product</b> <i>activates</i> the nitrate reductase operon
THEME	a) Affected by/results from event b) Focus of events describing states	<b>recA protein</b> was <i>induced</i> by UV radiation The <b>FNR protein</b> <i>resembles</i> CRP
MANNER	Method/way in which event is carried out	cpxA gene <i>increases</i> the levels of csgA transcription by <b>dephosphorylation</b> of CpxR
INSTRUMENT	Used to carry out event	EnvZ <i>functions</i> through <b>OmpR</b> to control NP porin gene expression in Escherichia coli K-12.
LOCATION	Where <i>complete</i> event takes place	Phosphorylation of OmpR <i>modulates</i> expression of the ompF and ompC genes in <b>Escherichia coli</b>
SOURCE	Start point of event	A transducing lambda phage was <i>isolated</i> from a <b>strain</b> harboring a glpD' lacZ fusion
DESTINATION	End point of event	Transcription of gntT is activated by <i>binding</i> of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to a <b>CRP binding site</b>
TEMPORAL	Situates event in time w.r.t another event	The Alp protease activity is <i>detected</i> in cells <b>after introduction</b> of plasmids carrying the alpA gene
CONDITION	Environmental conditions/changes in conditions	Strains carrying a mutation in the crp structural gene fail to <i>repress</i> ODC and ADC activities in response to <b>increased cAMP</b>
RATE	Change of level or rate	marR mutations <i>elevated</i> inaA expression by <b>10- to 20-fold</b> over that of the wild-type.
DESCRIPTIVE-AGENT	Provides descriptive information about the AGENT of the event	It is likely that HyfR <i>acts</i> as a <b>formate-dependent regulator</b> of the hyf operon
DESCRIPTIVE-THEME	Provides descriptive information about the AGENT of the event	The FNR protein <i>resembles</i> <b>CRP</b> .
PURPOSE	Purpose/reason for the event occurring	The fusion strains were <i>used</i> <b>to study</b> the regulation of the cysB gene by assaying the fused lacZ gene product

semantic frames that correspond to *groups* of verbs with similar behaviour. However, frames in PropBank correspond to individual verbs.

Biology-specific extensions have been attempted both for PropBank (Wattarujeekrit et al., 2004) and FrameNet (Dolbey et al., 2006). However, to our knowledge, there has been no such attempt at extending VerbNet into the biological domain.

In common with VerbNet, our work is focused on producing event frames that use a set of *frame-independent* semantic roles. However, we adopt a smaller set of roles tailored to the domain. This use of frame-independent roles allows linguistic generalisations to be captured more easily (Cohen and Hunter, 2006). Also, the use of such roles is more suitable for direct exploitation by NLP systems (Zaphirain et al., 2008).

Unlike VerbNet, we aim to produce a set of frames that are *verb-specific* (rather than frames that apply to groups of verbs). Verb-specific frames are able to provide more detailed argument specifications—particularly important in the biomedical field, where phrases that identify information such as location, manner, timing and condition are essential for correct interpretation of events (Tsai et al., 2007).

### 3 Annotated corpus

To aid semantic event frame extraction, we need a corpus annotated with event-level information.

Several already exist for biology. Some target extraction of PropBank-style frames (e.g. Chou et al. (2006), Kulick et al. (2004)). The corpus produced by Kim et al. (2008) uses frame-independent roles. However, only a few semantic argument types are annotated.

The target of our event frame extraction is a set of semantic frames which specify *all* potential arguments of gene regulation events. For this purpose, we had to produce our own annotated corpus, using a larger set of event-independent semantic roles than Kim et al. (2008). Our roles had to cover sufficiently wide scope to allow annotation and characterization of all instantiated arguments of relevant events within texts. To our knowledge, this makes our scheme unique within the biomedical field.

In contrast to many other comparable resources, annotated events are centred on both verbs *and* nominalised verbs, such as *transcription* and *control*. Nominalised verbs play an important and possibly dominant role in biological texts (Cohen and Hunter, 2006). Our own corpus confirms this, in that the nominalised verb *expression* is the most commonly annotated word on which gene regulation events are centred. By annotating events centred on nominalised verbs in a similar way to verbs, it becomes possible to extract separate event frames for nominalised verbs. This enables their potentially idiosyncratic behaviour to be accounted for.

Table 2. Description of NE super-classes

NE class	Definition
DNA	Entities chiefly composed of nucleic acids and their structural or positional references. This includes the physical structure of all DNA-based entities and the functional roles associated with regions thereof.
PROTEIN	Entities chiefly composed of amino acids and their positional references. This includes the physical structure and functional roles associated with each type.
EXPERIMENTAL	Both physical and methodological entities, either used, consumed or required for a reaction to take place.
ORGANISMS	Entities representing individuals or collections of living things and their component parts.
PROCESSES	A set of <i>event</i> classes used to label biological processes described in text.

Our annotated corpus consists of 677 MEDLINE abstracts on *E. Coli*. Within them, a total of 4770 gene regulation events have been annotated.

### 3.1 Semantic Roles

Based on the observations of Tsai et al (2007) regarding the most important types of information specified for biomedical events, together with detailed examination of a large number of relevant events within our corpus, in discussion with biologists, we defined a set of 13 frame-independent semantic roles that are suitable for the domain.

Certain roles within the set are domain-independent, and are based on those used in VerbNet, e.g. AGENT, THEME, and LOCATION. To these, we have added a number of domain-dependent roles, e.g. CONDITION and MANNER. The size of the role set attempts to balance the need for a sufficiently wide-ranging set of roles with the need for one that is as small and general as possible, to reduce the burden on annotators, whilst also helping to ensure consistency across extracted verb frames. The full set of semantic roles used is shown in Table 1.

### 3.2 Named Entity Categorisation

Although our semantic roles are rather general, the annotation scheme allows more detailed information about semantic arguments to be encoded in the corpus through the assignment of named entity (NE) tags. Unlike other corpus projects, we do not annotate *all* entities within each abstract, but just those entities that occur as semantic arguments of annotated gene regulation events.

Table 3. Most commonly annotated verbs and nominalised verbs

Word	Count	Type
expression	409	NV
encode	351	V
transcription	125	NV
bind	110	V
require	100	V
express	93	V
regulate	91	V
synthesis	90	NV
contain	80	V
induce	78	V

Our set of NE tags goes beyond the traditional view of NEs, in that labelling is extended to include *events* represented by nominalised verbs (e.g. *repression*). A total of 61 NE classes have been defined as being relevant to the gene regulation field, which are divided into four entity-specific super-classes (DNA, PROTEIN, EXPERIMENTAL and ORGANISMS) and one event-specific super-class (PROCESSES). The NEs within each of these classes are hierarchically-structured. Table 2 provides definitions of each of these five super-classes. The NEs correspond to classes in the Gene Regulation Ontology (Splendiani et al, 2007), which has been developed as part of the BOOTStrep project in which this work has been carried out. The Gene Regulation Ontology integrates parts of other established bio-ontologies, such as Gene Ontology (Ashburner et al., 2000) and Sequence Ontology (Eilbeck,2005).

### 3.3 Annotation process

Annotation was carried out over a period of three months by seven PhD students with experience in gene regulation and with native or near-native competence in English.

Prior to annotation, each abstract was automatically processed. Firstly, linguistic pre-processing (i.e. morphological analysis, POS tagging and syntactic chunking)<sup>1</sup> was carried out.

Secondly, all occurrences from a list of 700 biologically relevant verbs were automatically marked. Annotators then considered each marked verb within an abstract. If the verb denoted a gene regulation event, annotators then:

- a. Identified all semantic arguments of the verb within the sentence
- b. Assigned a semantic role to each identified argument

<sup>1</sup> Each abstract to be annotated is first pre-processed with the GENIA tagger (Tsuruoka et al, 2005).

- c. If appropriate, assigned named entity categories to (parts of) the semantic argument span
- d. If the argument corresponded to a nominalised verb, repeated steps a–c to identify its own arguments.

Syntactic chunks were made visible to annotators. In conjunction with annotation guidelines, the chunks were used to help ensure consistency of annotated semantic arguments. For example, the guidelines state that semantic arguments should normally consist of complete (and preferably single) syntactic chunks. The annotation was performed using a customised version of WordFreak (Morton and LaCivita, 2003), a Java-based linguistic annotation tool.

### 3.4 Corpus statistics

The corpus is divided into 2 parts, i.e.

- 1) 597 abstracts, each annotated by a single annotator, containing a total of 3612 events,
- 2) 80 pairs of double-annotated documents, allowing checking of inter-annotator agreement and consistency, and containing 1158 distinct events.

In the corpus, 277 distinct verbs were annotated as denoting gene regulation events, of which 73 were annotated 10 times or more. In addition, annotation has identified 135 relevant nominalised verbs, of which 22 were annotated 10 times or more. The most commonly annotated verbs and nominalised verbs are shown in Table 3.

### 3.5 Inter-annotator agreement

Inter-annotator agreement statistics for the 80 pairs of duplicate-annotated abstracts are shown in Table 4.

The figures shown in Table 4 are direct agreement rates. Whilst the Kappa statistic is very familiar for calculating inter-annotator agreement, we follow Wilbur et al. (2006) and Pyysalo (2007) in choosing not to use it, because it is not appropriate or possible to calculate it for all of the above statistics. For instance:

1. For some tasks, like annotation of events and arguments spans, deciding how to calculate random agreement is not clear.
2. The Kappa statistic assumes that annotation categories are discrete and mutually exclusive. This is not the case for the NE categories, which are hierarchically structured.

Table 4. Inter-annotator agreement rates

AGREEMENT RATE	VALUE
Event identification	0.49
Argument identification (partial span match)	0.73
Semantic role assignment	0.78
NE identification (partial span match)	0.68
NE category assignment (exact)	0.62
NE category assignment (including parent)	0.65
NE category assignment (including ancestors)	0.73

Table 4 shows that, in terms of identifying events (i.e. determining which verbs denote gene regulation events), agreement between annotators is reached about half the time. The main reason for this relatively low figure is that reaching a consensus on the specific types of events to be annotated under the heading of “gene regulation” required a large amount of discussion. Thus, particularly towards the start of the annotation phase, annotators tended to either under- or over-annotate the events.

Greater amounts of consistency seem to be achievable for other sub-tasks of the annotation, with agreement rates for the identification and subsequent labelling of semantic arguments being achieved in around three quarters of cases.

Comparable, but slightly lower rates of agreement were achieved in the identification of NEs. In terms of assigning categories to them, the agreement rate for exact category matches is a little lower (62%). However, if we relax the matching conditions by exploiting the hierarchical structure of the NE categories (i.e. if we count as a match the cases where the category assigned by one annotator was the ancestor of the category assigned by the other annotator), then the agreement increases by around 11%.

The large number of NE categories (61), makes the decision of the most appropriate category rather complex; this was verified by the annotators themselves. Based on this, we will consider the use of a more coarse-grained scheme when carrying out further annotation of this type. However, in the current corpus, the hierarchical structuring of the NE categories means that it would be possible to use a smaller set of categories by mapping the specific categories to more general ones.

## 4 Corpus Format

For the purposes of event frame extraction, the annotations in the corpus were converted to an XML-style inline format consisting of three different types of element:

**EVENT** – surrounds text spans (i.e. verb phrases and nominalised verbs) on which events are centred.

**SLOT** – surrounds spans corresponding to semantic arguments (i.e. slots) of events. The head verb/nominalised verb of the event is also treated as a SLOT, with role type Verb. The eventid attribute links each slot with its respective event, whilst the Role attribute indicates the semantic role assigned to the slot.

**NE** – surrounds text spans annotated as named entities. The *cat* attribute stores the NE category assigned.

Where there are several annotations over some text span, elements are embedded inside each other. If more than one annotation begins at a particular offset, then the ordering of the embedding is fixed, so that SLOT elements are embedded inside EVENT elements, and that NE elements are embedded inside SLOT elements. An example of the annotation for the sentence "*TaqI restriction endonuclease has been subcloned downstream from an inducible phoA promoter*" is shown below:

```
<SLOT argid="4" eventid="5" Role="Theme">
<NE cat="ENZYME">TaqI restriction endonucle-
ase</NE></SLOT> <EVENT id="5">
has been <SLOT argid="6" eventid="5"
Role="Verb">subcloned </SLOT></EVENT>
<SLOT argid="8" eventid="5"
Role="Location">downstream from
<NE cat="PROMOTER">an inducible phoA pro-
moter</NE></SLOT>.
```

The EVENT created over the VP chunk *has been subcloned* has been annotated as having 2 semantic arguments (SLOTS), i.e. a THEME, *TaqI restriction endonuclease* and a LOCATION, i.e. *downstream from an inducible phoA promoter*. A 3<sup>rd</sup> SLOT element corresponds to the head verb in the VP chunk. Named entity tags have also been assigned to the THEME span and part of the LOCATION span.

## 5 Event Patterns and Event Frames

This section defines event patterns and event frames. Event patterns are syntactic patterns of sequences of surface words, NEs, and semantic roles, whilst event frames are the record-like data structures consisting of event slots and event slot values.

### 5.1 Event Patterns

Event patterns are fragments of event annotations in which semantic arguments are generalized to their semantic role and NE categories, if present.

An event pattern is extracted for each unique event id within an abstract. An event annotation span begins with the earliest SLOT span, and ends with the latest SLOT assigned to the event. An example event span is as follows:

```
<SLOT eventid="9" Role="Agent">
<NE cat="OPERON"> transfer operon</NE></SLOT>
<EVENT id="9"><SLOT eventid="9" Role="Verb">
expression </SLOT></EVENT></SLOT> of
<SLOT eventid="9" Role="Theme">
<NE cat="DNA_FRAGMENT"> F-like plasmids
</NE></SLOT>
```

For each event, each event span is generalized into an event pattern as follows:

- “Verb” role slots of the event are converted into a tuple consisting of the role type, part-of-speech and surface form, i.e., [Verb:POS:verb].
- Other semantic role slots and their NE slots for the event are generalized to tuples consisting of the role and NE super class, i.e., [role:NE\_super\_class].
- Other XML tags are removed.

The above example event span is thus generalized to the following event pattern:

```
[Agent:DNA] [Verb:NN:expression] of [Theme:DNA].
```

### 5.2 Event frames

Event frames are directly extracted from event patterns, and take the following general form:

```
event_frame_name(
    slot_name => slot_value,
    ...
    slot_name => slot_value).
```

where

- *event\_frame\_name* is the base form of the event verb or nominalized verb;
- *slot\_names* are the names of the semantic roles within the event pattern;
- *slot\_values* are NE categories, if present within the event pattern.

For example, the event frame corresponding to the event pattern shown in the previous section is as follows:

```
expression( Agent=>DNA,
            Theme=>DNA ).
```

## 6 Event Frame Extraction

Our event frame extraction is a fusion of sequential labelling based on Conditional Random Fields (CRF), and event pattern matching. Event frames are extracted in three steps. Firstly, a CRF-based Named Entity Recognizer (NER) assigns biological NEs to word sequences. Secondly, a CRF-based semantic role labeller determines the semantic roles of word sequences with NE labels. Thirdly, word sequences are compared with event patterns derived from the corpus. Only those event frames whose semantic roles, NEs, and verb POS satisfy event pattern conditions will be extracted.

### 6.1 Biological NER

Since it is costly and time-consuming to create a large-scale training corpus annotated by biologists, we need to concede to use coarse-grained biological NE categories. That is, the NER component is trained on the five NE super classes, i.e., *Protein*, *DNA*, *Experimental*, *Organisms*, and *Processes*.

The NER models are trained by CRFs (Lafferty et al., 2001) using the standard IOB2 labelling method. That is, the label “B-NE” is given to the first token of the target NE sequence, “I-NE” to each remaining token in the target sequence, and “O” to other tokens.

Features used are as follows:

- word feature

- orthographic features:

the first letter and the last four letters of the word form, in which capital letters in a word are normalized to “A”, lower case letters are normalized to “a”, and digits are replaced by “0”. For example, the word form “IL-2” is normalised to “AA-0”.

- postfix features: the last two and four letters

- POS feature

We applied first-order CRFs using the above features for the tokens within a window size of  $\pm 2$  of the current token.

### 6.2 Semantic Role Labelling

First of all, each NE token sequence identified by B and I labels is merged into a single token with the NE category name. Then, the semantic role labelling models are trained by CRFs in a similar way to NER. That is, the label “B-Role” is given to the first token of the target *Role* sequence, “I-

*Role*” to each remaining token in the target sequence, and “O” to other tokens.

Features used here are as follows:

- word feature
- base form feature
- POS feature
- NE feature

The window size was  $\pm 2$  of the current token.

### 6.3 Event pattern matching

When a new sentence is given, sequential labelling models decide NE and semantic role labels of tokenized input sentences. Then, the token sequences are converted into the following token sequences with POS, semantic role, and NE information (called *augmented token sequences*):

1. Each token sequence labelled by IOB semantic role labels is merged into a token labelled with the role.
2. Verbs and nominalized verbs are converted to [Verb:POS:surface\_form].
3. Tokens with semantic role label and NE super-class are converted into the form [Role:NE\_super\_class].
4. Other tokens with O label are converted to surface tokens.

Then, event patterns are generalized:

5. Event patterns are modified so that elements corresponding to verbs and nominalized verbs will match any words with the same POS, e.g., [Verb:POS:\*].

Finally, each event pattern is applied to augmented token sequences one by one:

6. By matching the generalized event patterns with augmented token sequences, i.e. when verbs or nominalized verbs and the surrounding semantic roles and NEs satisfy the event pattern conditions, then successfully unified event patterns are extracted as new event patterns.
7. The newly obtained event patterns are converted into event frames in the same way as described in Section 5.2.

## 7 Experimental Results

The aim of this section is to evaluate semantic frame extraction performance, given a set of annotated training data.

The annotated corpus was randomly separated into 10 document groups and their event patterns

Table 5. 10-fold cross validation results

	Score	#TP	#FN	#FP
Recall	0.186	165	730	
Precision	0.490	165		172

Table 6. NE identification performance

NE Type	Recall	Precision	F
DNA	0.627	0.660	0.643
Protein	0.525	0.633	0.574
Experimental	0.224	0.512	0.312
Processes	0.125	0.337	0.182
Organisms	0.412	0.599	0.488

and event frames were segmented into 10 groups according to the document separation.

We conducted 10-fold cross validation based on the 10 document groups. Named entity recognizers and semantic role labellers were trained using 9 groups of annotated documents. Event frames were then extracted from the remaining group of documents. Micro-average precision and recall for the set of event frames extracted from *all* the folds were evaluated.

Table 5 shows the event frame extraction performance. #TP, #FN, and #FP indicate the number of true positives, false negatives, and false positives, respectively.

Named entity recognition performance was also evaluated (Table 6). Since the training data size is small, the performance is between approximately 20-60% F-measure. However, this will not cause a problem for the event frame extraction task. This is because, if a particular event frame occurs multiple times in a corpus, it is sufficient to extract only a single occurrence of the event description. So, whilst the NE and semantic role labelling may not be successful for all occurrences of the event frame, there is a good chance that at least one occurrence of the event will be realized in the text in such a way as to allow the labelling to be carried out successfully, thus allowing the extraction of an appropriate event frame.

## 8 Discussion

Linguistic-level event annotation of biological events is an inherently difficult task. This is supported by the fact that the inter-annotator agreement level for the identification of events was 0.49 (see Table 4). Therefore, in terms of event extraction performance, a precision of 49.0% on 10-fold cross validation is almost comparable to human experts. The low recall of 18.6% may not be an issue, as the recall is likely to improve with the size of the target corpus.

The precision may additionally be underestimated in the evaluation due to inconsistencies in the annotation. We found that the *average* precision of our event frame extraction over 10 folds is around 30%, despite the fact that the precision of *all* event frames extracted from 10 folds is almost 50% compared with the annotated event frames in the whole corpus. This happens because some events not annotated in a particular fold are annotated in the rest of corpus. From this insight, our conjecture is that the true precision against the whole corpus would be somewhat higher (potentially 70-80%) if we were using an annotated corpus 10 times larger for the evaluation.

The automatic NER performance was also comparable to human annotators.

There are several approaches to the generation of information extraction patterns (e.g. Soderland et al., 1995; Califf et al., 1997; Kim and Moldovan, 1995). Our event patterns are similar to information extraction rules used in conventional IE systems. However, the goal of this paper is not event *instance* extraction but event (or semantic) frame extraction. We also combined CRF-based NER and semantic role labelling tuned for gene regulation with event extraction from sentences so that the clues of gene regulation event frames could be assigned automatically to un-annotated text.

## 9 Conclusion

This paper has presented linguistic annotation of gene regulation events in MEDLINE abstracts, and automatic event frame extraction based on the annotated corpus. Semantic event frames are linguistic resources effective in bridging between domain knowledge and text in IE tasks.

Although biological event annotations carried out by domain experts is a challenging task, experimental results on event frame extraction demonstrate a precision of almost 50%, which is close to the inter-annotator agreement rate of human annotators.

The extracted event frames will be included in the BOOTStrep BioLexicon, which will be made available for research purposes.

## Acknowledgement

This research is supported by EC IST project FP6-028099 (BOOTStrep), whose Manchester team is hosted by the JISC/BBSRC/EPSRC sponsored National Centre for Text Mining.

## References

- Califf, Mary E. and Raymond J. Mooney (1997). Relational Learning of Pattern-Match Rules for Information Extraction, *In Proceedings of the ACL-97 Workshop in Natural Language Learning*, pp 9–15.
- Chou, Wen-Chi., Richard T.H. Tsai, Ying-Shan Su, Wei Ku, Ting-Yi Sung and Wen-Lian Hsu (2006). A Semi-Automatic Method for Annotating a Biomedical Proposition Bank. *In Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pp 5–12.
- Cohen, K. Bretonnel and Laurence Hunter (2006). A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics* 7 (Suppl. 3), S5.
- Dolbey, Andrew, Michael Ellsworth and Jan Scheffczyk (2006). BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In O. Bodenreider (Ed.), *In Proceedings of KR-MED*, pp 87–94.
- Eilbeck, Karen, Suzanna E. Lewis., Christopher J. Mungall, Mark Yandell, Lincoln Stein, Richard Durbin and Michael Ashburner. (2005) The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* 6:R44
- Kim, Jin-Dong, Tomoko Ohta and Jun'ichi Tsujii (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9:10.
- Kim, Jun-Tae and Dan I. Moldovan (1995). Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. *IEEE Transaction on Knowledge and Data Engineering (IEEE TKDE)*, 7(5), pp.713–724.
- Kipper-Schuler, Karen (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. *PhD Thesis. Computer and Information Science Dept., University of Pennsylvania*. Philadelphia, PA.
- Kulick Seth, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar (2004) Integrated Annotation for Biomedical Information Extraction. *In HLT-NAACL 2004 Workshop: BioLink 2004, Linking Biological Literature, Ontologies and Databases*, pp 61–68.
- Lafferty John, Andrew McCallum and Fernando Pereira (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*, pp 282–289.
- Morton, Thomas and Jeremy LaCivita (2003). Word-Freak: an open tool for linguistic annotation. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp 17–18.
- Palmer Martha, Paul Kingsbury and Daniel Gildea (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), pp 71–106.
- Pyysalo, Sampo, 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.
- Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk (2006). FrameNet II: Extended Theory and Practice. Available online at <http://framenet.icsi.berkeley.edu/>
- Soderland, Steven, David Fisher, Jonathan Aseltine and Wendy Lenert (1995). CRYSTAL: Inducing a Conceptual Dictionary, *In Proceedings of The 13th International Joint Conference on Artificial Intelligence (IJCAI-95)*. pp.1314–1319.
- The Gene Ontology Consortium. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, pp 25–29.
- Tsai Richard T.H, Wen-Chi Chou, Ying-San Su, Yu-Chun Lin, Chen-Lung Sung, Hong-Jie Dai, Irene T.H Yeh, Wei Ku, Ting-Yi Sung and Wen-Lian Hsu (2007). BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features, *BMC Bioinformatics* 8:325
- Tsuruoka, Yoshimasa, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text, *In Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp 382–392.
- Wattarujeekrit, Tuangthong, Parantu K. Shah and Nigel Collier (2004). PASBio: predicate-argument structures for event extraction in molecular biology, *BMC Bioinformatics* 5:155.
- Wilbur, W. John, Andrey Rzhetsky, and Hagit Shatkay (2006). New Directions in Biomedical Text Annotations: Definitions, Guidelines and Corpus Construction. *BMC Bioinformatics*. 7:356
- Zapirain, Beñat, Eneko Agirre, Lluís Màrquez (2008). A Preliminary Study on the Robustness and Generalization of Role Sets for Semantic Role Labeling. In Alexander F. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, 9th International Conference, CICLing 2008.