Syntactic Dependency Based Heuristics for Biological Event Extraction

Halil Kilicoglu and Sabine Bergler

CLaC Lab
Department of Computer Science and Software Engineering
Concordia University

June 5, 2009

Summary

- BioNLP'09 Event Extraction Shared Task participation
  - Task 1: Core event extraction (3rd place)
  - Task 3: Speculation and negation recognition (1st place)
- Rule-based methodology
  - Statistical compilation of a dictionary of event triggers with limited expansion
  - Grammar for participant identification from dependency parse
  - Heuristics-based speculation and negation detection

Related Work

- Speculation recognition
  - Speculative sentences vs. non-speculative sentences
  - Supervised machine learning techniques [Light et al., 2004; Medlock and Briscoe, 2007; Szarvas, 2008]
  - Lexico-syntactic patterns [Kilicoglu and Bergler, 2008]
- Negation detection
  - Identification of negated terms or concepts
  - Machine learning techniques [Goldin and Chapman, 2003; Averbuch et al., 2004]
  - Pattern-based approaches [Mutalik et al., 2001; Chapman et al., 2001]
- Recent interest in negation and speculation scope identification
  - Memory-based classification [Morante et al., 2008]

Preprocessing

- Create an enriched XML representation of documents
  - Sentences, entities and their offset positions
  - Word information: tokens, POS tags, lemmas and indexes
  - Dependency parses of sentences
- Stanford Lexicalized Parser for word information and dependency parsing

Core Event Extraction Pipeline

1. Determine event triggers and their event types
    - Dictionary lookup and “goodness score”
2. Identify potential arguments
    - Participant identification grammar
3. Recursive sub-event identification for regulatory events
4. Post-processing rules for event trigger limitations

Event Triggers

- A dictionary of event triggers drawn from training data
- Event triggers: verbs, nouns and adjectives (predicates)
- Limited term expansion and filtering
  - Both hyphenated and non-hyphenated forms of prefixed triggers (e.g., *co*-, *down*-, *up*-, *trans*-)
  - Derivational forms (e.g., *dimerization* after *dimerize*)
  - Weak event trigger elimination (e.g., *absence* for `Negative_regulation`)
  - No multi-word event triggers
- Event trigger/event type "goodness score"
  - Maximum likelihood estimation
  - Used as a threshold

Event Participant Identification

- Based on rules using “collapsed” Stanford dependency relations
- Extracted and ranked the dependency paths between event triggers and participants
    - Frequency of occurrence

Dependency	**POS**	**Role**	**Event**	**Trigger**
dobj[1]	VB	Theme	*	*
nsubj[2]	JJ	Cause	[Regulatory]	[*responsible,sufficient,...*]
prep_to[3]	VB	Theme	Binding	*
prep_to	VB	Theme	Pos.regulation	[*lead,contribute*]
prep_between[4]	NN*	Theme	Binding	[*association,interaction*]

[1] direct object

[2] nominal subject

[3] prepositional phrase headed by *to*

[4] prepositional phrase headed by *between*

Special Considerations for Regulatory Events

- Events as event participants
  - Dependency path between the event trigger and the sub-event trigger
  - *We have examined the effect of LTB4 on the expression of . . .*
- Hyphenated participle modifiers
  - Additional rule without dependency relations
  - *. . . LPS-mediated TF expression . . .*
- Reversal of semantic roles for *require* and *involve*
- "Corrected" dependency paths for PP attachment errors
  - For certain trigger words (e.g., *effect, influence, role*)

dobj(examined,effect)
prep_on(examined,expression) ⇒ *prep_on*(effect,expression)

Coordination and Apposition

- Coordination
  - Both event trigger and participant coordination
  - Derived from dependency relations (*conj_**)
  - Additional rules for better resolution of coordinated entities
    - Separated by comma or semi-colon
    - Separated by a coordinating conjunction
    - Separated by parenthetical expression
    - Combination of above
  - *. . . interleukin-2 (IL-2) and IL-4 gene transcription . . .*
- Apposition
  - An entity and a word in an apposition construction are considered equivalent
  - Derived from dependency relations (e.g., *appos*, *abbrev*)
  - *. . . regulation of inflammatory cytokine genes including TNF,*
    prep_of(regulation,genes)
    prep_including(genes,TNF)

Postprocessing Rules

- Loosen the strict assumptions about event triggers
- Multi-word event triggers
  - (*positive* OR *negative*) + nominal `Regulation` trigger
    OR
    (*positively* OR *negatively*) + verbal `Regulation` trigger
    ⇒ `Positive_regulation` OR `Negative_regulation`
- "One trigger-one event" limitation
  - *overexpression*, *transfect*, etc.

Core Event Extraction Results

Event Class	**Recall**	**Precision**	**F-score**	**Rank**
Simple events	43.10	73.47	54.33	5
Regulatory events	27.47	49.89	35.43	2
TOTAL	34.98	61.59	44.62	3

- Our system favors precision over recall
- The results at empirically determined “goodness score” threshold of .08
- No special treatment of simple events
- Special focus on regulatory events
  - Led to better ranking

Core Event Extraction Errors

- Naive view of event triggers
  - "Once a trigger-always a trigger" (precision errors) (e.g., *oligomerization*)
  - Previously unseen triggers (recall errors)
- Errors in dependency relations
- Pattern incompleteness
- Anaphoric expressions
- Events spanning multiple sentences

Speculation Recognition

- Refinement of a speculation cue dictionary developed in prior work [Kilicoglu and Bergler, 2008]
  - Ignore some cue classes completely (modal verbs, epistemic adverbs)
  - Introduce a new speculative verb class (active cognition verbs)
    - *examine, evaluate, analyze, study, investigate*
    - Consider the nominal forms as well
- Speculation scope
  - A dependency path between the speculation cue and the event trigger
  - *. . . these data suggest that ETS1 may be involved in mediating the increased GM-CSF production . . .*
    ccomp(suggest,involved)

Negation Detection

- Lexical cues
- Syntactic dependency rules between lexical cues and event triggers

Cue	**Dependency**
lack, absence	*prep_of*
inability, failure	*infmod*
no, not, cannot	*det*

- Certain dependencies involving the event trigger (e.g., *neg*, *conj_negcc*)
- A negation cue word preceding the event trigger or an event participant (*no, not, cannot*)

Speculation and Negation Recognition Results

	Recall	**Precision**	**F-score**	**Rank**
Speculation	16.83	50.72	25.27	1
Negation	14.98	50.75	23.13	1
TOTAL	15.86	50.74	24.17	1

- Misidentified or missed base events
- Errors due to speculation and negation detection module
  - Speculation: 4 false positives (out of 39), 7 false negatives (out of 95)
  - Negation: 5 false positives (out of 31), 5 false negatives (out of 107)

Speculation and Negation Detection Errors

- False positive errors
  - Controversial false positive cases
    - *An unidentified Ets family protein binds to . . . and appears to negatively regulate the human IL-2R alpha promoter.*
    - Difficult to annotate correctly and consistently
  - The negation pattern involving negation cues in the token preceding the event trigger
    - Introduced to increase recall
- False negative errors
  - Complex and infrequent patterns
    - *. . . suggest a molecular mechanism for the inhibition of . . .*
    - *Galectin-3 is . . . and is expressed in many leukocytes, with the notable exception of B and T lymphocytes*

Conclusions and Future Work

- Dependency relations for biological event extraction as well as for speculation/negation detection
- Easy adaption of prior work, showing the portability and extensibility of a linguistically-oriented approach
- Difficulty of annotating speculation, necessity of developing annotation guidelines

- Anaphora resolution and multiple sentence spanning events
- Subcategorization information for event triggers
- Dependency relations extracted using constituent parses from different parsers (Charniak parser, etc.)