Biomedical Event Annotation with CRFs and Precision Grammars

Andrew MacKinlay, David Martinez, Timothy Baldwin
Motivation and Architecture

Our motivation: deep linguistic processing for detection of speculation and negation

Architecture:

- Task 1:
  - Trigger word detection: CRF and Lookup systems
  - Event-theme construction (hand-crafted rules)

- Task 3:
  - Deep parsing for semantic representation
  - Classification of events using Maximum Entropy
Trigger word detection with CRFs

- Conditional probability distribution over label sequences given a particular observation sequence
- CRF++ toolkit (Sha and Pereira, 2003)
- Tested features: word-form, lemma, POS, chunking marks, protein NER, grammatical dependencies (from Bikel parser and GDep)
- JULIE-Lab sentence splitter and Genia Tagger for pre-process
- Window sizes: ±3 and ±4
Best results (training data): Precision \sim 66\%, Recall \sim 30\%

- All features help except for grammatical dependencies
- \pm 3 window size
Decision list for each trigger string found in training data
  - Simply assign highest frequency class

Frequency cut-off

We can reach high recall (\(~ 77\%\)) but at the cost of precision (\(~ 13\%\))

Best f-score \(~ 36\%\) (\(~ 50\%\) recall)
Add all trigger words identified by CRF and look-up

Two approaches:
- Optimise per class (Optim)
- Always preference to CRF (All)
Event-theme construction

- Approach: assign closest events/proteins as themes (without crossing sentence boundaries)
- Basic events:
  - Single closest protein
- Binding events:
  - Closest proteins
  - Parameters: maximum distance and number of themes
- Regulation events
  - Single closest protein or event (give precedence to events)
  - Parameters: maximum distance and detect/ignore CAUSE
Table: Task 1 results with approximate span matching, recursive evaluation (our final submission is in bold)

<table>
<thead>
<tr>
<th>System</th>
<th>Rec.</th>
<th>Prec.</th>
<th>FSc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined (Optim.)</td>
<td>17.44</td>
<td>39.99</td>
<td>24.29</td>
</tr>
<tr>
<td>Combined (All)</td>
<td>24.36</td>
<td>30.87</td>
<td>27.23</td>
</tr>
<tr>
<td>CRF</td>
<td>12.23</td>
<td>62.24</td>
<td>20.44</td>
</tr>
<tr>
<td>CRF (+ synt feats)</td>
<td>12.01</td>
<td>61.91</td>
<td>20.11</td>
</tr>
<tr>
<td>Look-Up</td>
<td>22.88</td>
<td>29.67</td>
<td>25.84</td>
</tr>
<tr>
<td>Look-Up (freq &gt;= 20)</td>
<td>23.26</td>
<td>26.74</td>
<td>24.88</td>
</tr>
<tr>
<td>Look-Up (freq &gt;= 30)</td>
<td>21.37</td>
<td>30.50</td>
<td>25.13</td>
</tr>
</tbody>
</table>
English Resource Grammar (ERG): high-precision grammar in the HPSG framework

GENIA tagger to deal with named entities

72% of training sentences parsed
Semantic formalism: Robust Minimal Recursion Semantics

Elementary Predicates (EP): Predicates with their arguments

Relationships between trigger EP and lexical cues
  - Outscoping and shared-argument
Features for negation identification

- Pre-identify word lists:
  - Conjunctions: \_not\_c, \_but+not\_c, \_nor\_c
  - Other markers: \_only\_a, \_never\_a, \_not+as+yet\_a, \_not+as+yet\_a, \_unable\_a, neg\_rel

  - E.g. “...product was not (NEG-EP) able to bind (TRIG-EP) DNA and...”
    - NegOutscope neg\_rel = 1
    - NegOutscope not = 1
...product was not able to bind DNA and was recovered in cytoplasmic cellular extracts...

ERG analysis

- $l_8$: neg_rel(692 : 695)(e9, ARG1: h10)
- $l_{11}$: _able_a_.1(696 : 700)(e12, ARG1: x6, ARG2: h13)
- $l_{14}$: _bind_v_to_(704 : 708)(e17, ARG1: x6, ARG2: x16, ARG3: u15)
- h10 qeq $l_{11}$, h13 qeq $l_{14}$

Thus $l_8$ immediately outscopes $l_{11}$, and $l_{11}$ immediately outscopes $l_{14}$
Features for negation identification

- Negative conjunction: when trigger-EP is the argument (ARG0) of a negative conjunction EP
  - E.g. “...but not (NEG-EP) binding (TRIG-EP) DNA...”
- When trigger-EP is the argument (ARG0) of a negatively-outscoped EP
  - E.g. “...the product (TRIG-EP) was never (NEG-EP) considered...”
Features for speculation identification

- Pre-identify word lists:
  - Speculation verb short list: _investigate, _study, _examine, _test, _evaluate, _observe
  - Extended list: adding WordNet sisters
- SpecVOBJ: when verb part of “speculative-verbs” set, and object is a trigger word
  - E.g. “IkappaBalpha phosphorylation and degradation (TRIG-EP) was analyzed (SPEC-EP)”
    - SpecVObj2+WN-seed:examine = 1
    - SpecVObj2+wn-sister:_analyze_v_1(examine) = 1
    - SpecVObj2+wn-gen = 1
More features

- Speculation:
  - Modal verb outscopes trigger
  - ARG0 of trigger-EP occurs as argument of the word 'analysis'

- General features:
  - E.g. (Modifier adjective) “...Fas **upregulation** (TRIG-EP) is **central** (ADJ-EP) to the preservation...”
  - ’ModAdj:_central_a_1’ = 1
  - Trigger name, trigger POS, etc.
Negation/Speculation Classifiers

- Maximum Entropy classifier (Maxent Toolkit)
- Different feature combinations
- Baseline: bag of words
- Development phase:
  - Goldstandard events
  - 10-fold cross-validation
- Test phase:
  - Trained over goldstandard event extraction
  - Output of task-1 classifier as source of trigger words
Development results: Speculation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>22.1</td>
<td>47.7</td>
<td>30.2</td>
</tr>
<tr>
<td>Spec. + BOW</td>
<td>23.2</td>
<td>57.9</td>
<td>33.1</td>
</tr>
</tbody>
</table>

- Very low performance over automatic classification
- Linguistic features better than BOW
- Combination of features works best
Development results: Negation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>15.0</td>
<td>30.2</td>
<td>20.0</td>
</tr>
<tr>
<td>Neg. + BOW</td>
<td>24.3</td>
<td>68.4</td>
<td>35.9</td>
</tr>
</tbody>
</table>

- Bigger improvement over BOW
### Official results for Task 3

<table>
<thead>
<tr>
<th>TEAM</th>
<th>gold (match)</th>
<th>answer (match)</th>
<th>recall</th>
<th>prec.</th>
<th>fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConcordU</td>
<td>3617 ( 1182)</td>
<td>1943 ( 1182)</td>
<td>32.68</td>
<td>60.83</td>
<td>42.52</td>
</tr>
<tr>
<td>VIBGhent</td>
<td>3617 ( 1105)</td>
<td>2227 ( 1104)</td>
<td>30.55</td>
<td>49.57</td>
<td>37.80</td>
</tr>
<tr>
<td>ASU+HU+BU</td>
<td>3617 ( 710)</td>
<td>1185 ( 710)</td>
<td>19.63</td>
<td>59.92</td>
<td>29.57</td>
</tr>
<tr>
<td><strong>NICTA</strong></td>
<td>3617 ( 577)</td>
<td>1450 ( 575)</td>
<td>15.95</td>
<td>39.66</td>
<td><strong>22.75</strong></td>
</tr>
<tr>
<td>USzeged</td>
<td>3617 ( 722)</td>
<td>3113 ( 722)</td>
<td>19.96</td>
<td>23.19</td>
<td>21.46</td>
</tr>
<tr>
<td>CCP-BTMG</td>
<td>3617 ( 446)</td>
<td>777 ( 446)</td>
<td>12.33</td>
<td>57.40</td>
<td>20.30</td>
</tr>
</tbody>
</table>
Lessons learned

- Keyword detection suffers from data sparseness
- Rules for event construction are too naive
- Deep parsing better than lexical baseline, but there are coverage problems
- Combined approach (detect triggers and themes together) to be explored for task 1
Biomedical Event Annotation with CRFs and Precision Grammars

Andrew MacKinlay, David Martinez, Timothy Baldwin