Event Extraction as Dependency Parsing

David McClosky, Mihai Surdeanu, and Christopher D. Manning
Stanford University
{mcclosky,mhais,manning}@stanford.edu

Motivation

Leading event extraction systems use pipelines which ignore joint information. Casting event extraction as a parsing problem allows use of standard parsing tools which can balance decisions more globally.

Approach

Above event structures converted to dependencies

Convert event structures to dependency graphs, use reranking parser to predict event structure.

Parser

Dependency graph: nodes are words, edges are semantic relations (Theme, Cause, etc.)

Use MSTParser to handle n-best non-projective parses

Features: original MST features plus new event-specific features which use “full” sentence and its syntactic parse

Multiple decoders available, capture different views (e.g. 2P = second order edge features, projective)

Anchor Detection

- Cast as token classification problem (9 event types)
- Multiword anchors reduced to their syntactic head (shown as dashed boxes)
- Features include word and lemma of nearby words, syntactic context, gazetteer, and nearby entities
- Token-level logistic regression applied
- Skewed towards high recall (parser can ignore extraneous event anchors)

Results

<table>
<thead>
<tr>
<th>AD</th>
<th>Parse</th>
<th>RR</th>
<th>Conv</th>
<th>R</th>
<th>P</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>45.9</td>
<td>61.8</td>
<td>52.7</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>48.7</td>
<td>59.3</td>
<td>53.5</td>
</tr>
<tr>
<td>G</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>68.9</td>
<td>77.1</td>
<td>72.7</td>
</tr>
<tr>
<td>G</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>68.5</td>
<td>78.2</td>
<td>73.1</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>✓</td>
<td>✓</td>
<td>81.6</td>
<td>93.4</td>
<td>87.1</td>
</tr>
</tbody>
</table>

Ablation study between system components

G indicates gold, ✓ indicates component was used.

Current system, Parse and Reranker (RR), perform well with predicted triggers (53.5%) and gold (73.1%).

Conclusions

- Final system performs competitively (48.6% BioNLP F-score on final test set)
- Would have been 2nd place in BioNLP 2009. Best current results are 53.3% (Miwa et al., 2010) and 52.0% (Björne et al., 2009).
- Joint inference and global features in reranker can capture more structure than standard pipeline approaches.
- Minimal domain-specific tuning: almost no domain dependent features were used thus easily adaptable.