Refining Event Extraction
(Old-Fashion ‘Traditional’ IE)
Through Cross-document Inference

Heng Ji and Ralph Grishman
(hengji,grishman)@cs.nyu.edu

Computer Science Department
New York University
June, 2008
Outline

- Background: Event Extraction and Its Performance Limitation
- Motivation: Event Extraction Beyond Document Boundary
- Approach Overview
  - System Pipeline
  - Baseline Within-Sentence Event Extraction
  - Information Retrieval and Query Construction
- Global Confidence Estimation and Inference
- Experimental Results
  - Confidence Metric Thresholding
  - Overall Performance
- Conclusion and Future Work
Event Extraction: ‘Traditional’ IE

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Quit (a “Personnel/End-Position” event)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arguments</td>
<td></td>
</tr>
<tr>
<td>Role = Person</td>
<td>Barry Diller</td>
</tr>
<tr>
<td>Role = Organization</td>
<td>Vivendi Universal Entertainment</td>
</tr>
<tr>
<td>Role = Position</td>
<td>Chief</td>
</tr>
<tr>
<td>Role = Time-within</td>
<td>Wednesday</td>
</tr>
</tbody>
</table>

- Target: 33 different types of Automatic Content Extraction (ACE) events
IE Beyond Document Boundary

- Most event extraction systems operate a sentence a time; MUC-style Event Extraction hit the 60% ‘performance ceiling’

- Look back at the initial goal of IE
  - Create a database of relations and events from the *entire corpus*
  - Within-doc/Within-Sent IE was an artificial constraint to simplify the task and evaluation

- Many events will be reported multiple time in different forms
  - Get *background knowledge* from a *cluster of topically-related documents*
  - Favor *interpretation consistency* within each cluster

- Hypotheses
  - One Trigger Sense Per Cluster
  - One Argument Role Per Cluster
One Trigger Sense Per Cluster

Test Doc
It took a federal act -- by the Canadian government -- to give Martha Stewart permission to paddle a hollowed-out, 600-pound pumpkin across a lake in Windsor, Nova Scotia. Martha Stewart was planning to fire the pumpkin, write it a nice letter and let it appear on her daytime show...

Conflict_Attack or End_Position?

Training Corpora

<table>
<thead>
<tr>
<th>Conflict_Attack: 54%</th>
<th>End_Position: 7%</th>
<th>Other: 39%</th>
</tr>
</thead>
<tbody>
<tr>
<td>fire fire fire fire fire</td>
<td>fire fire</td>
<td>fire</td>
</tr>
</tbody>
</table>

Related Docs

<table>
<thead>
<tr>
<th>End_Position: 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>fire fire fire</td>
</tr>
</tbody>
</table>
Also held in Palu was Bali bomb suspect Ahmad Rohim, alias Saad. Police on the Indonesian resort island of Bali are deploying thousands of officers in the hope that the first Bali bombing trial will proceed smoothly.

JI suspect kills himself at Indonesian police HQ, six others held. Declared suspect in a string of church bombings in Indonesia in 2000 based on testimony by an alleged al-Qaida operative, Omar Al-Faruq, held by U.S. authorities.

Identified as Arrest Events by Classifier

-Arrest Event? Yes.
Vivendi earlier this week confirmed months of press speculation that it planned to shed its entertainment assets by the end of the year.

Vivendi has been trying to sell assets to pay off huge debt, estimated at the end of last month at more than $13 billion. Under the reported plans, Blackstone Group would buy Vivendi’s theme park division, including Universal Studios Hollywood, Universal Orlando in Florida.

Is a seller? Yes.
Cross-Sent/Cross-Doc Event Inference Architecture
Baseline Within-Sentence Event Extraction

1. Pattern matching
   - Build a pattern from each ACE training example of an event
     - British and US forces reported gains in the advance on Baghdad
     - PER report gain in advance on LOC

2. MaxEnt models
   ① Trigger Classifier
     - to distinguish event instances from non-events, to classify event instances by type
   ② Argument Classifier
     - to distinguish arguments from non-arguments
   ③ Role Classifier
     - to classify arguments by argument role
   ④ Reportable-Event Classifier
     - to determine whether there is a reportable event instance
Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment, the entertainment unit of French giant Vivendi Universal. Diller took on the "provisional" role at the top of Vivendi's US entertainment operations…
Global Confidence Estimation

- Within-Sentence IE system produces local confidence
- IR engine returns a cluster of related docs for each test doc

Document-wide and Cluster-wide Confidence
- Frequency weighted by local confidence
- \(XDoc-\text{Trigger-Freq}(\text{trigger, etype})\): The weighted frequency of string \text{trigger} appearing as the trigger of an event of type \text{etype} across all related documents
- \(XDoc-\text{Arg-Freq}(\text{arg, etype})\): The weighted frequency of \text{arg} appearing as an argument of an event of type \text{etype} across all related documents
- \(XDoc-\text{Role-Freq}(\text{arg, etype, role})\): The weighted frequency of \text{arg} appearing as an argument of an event of type \text{etype} with role \text{role} across all related documents
- \text{Margin} between the most frequent value and the second most frequent value, applied to resolve classification ambiguities
- ......
Cross-Sent/Cross-Doc Event Inference Procedure

- Remove triggers and argument annotations with local or cross-doc confidence lower than thresholds
  - *Local-Remove*: Remove annotations with low local confidence
  - *XDoc-Remove*: Remove annotations with low cross-doc confidence

- Adjust trigger and argument identification and classification to achieve document-wide and cluster-wide consistency
  - *XSent-Iden/XDoc-Iden*: If the highest frequency is larger than a threshold, propagate the most frequent type to all unlabeled candidates with the same strings
  - *XSent-Class/XDoc-Class*: If the margin value is higher than a threshold, propagate the most frequent type and role to replace low-confidence annotations
Experiments: Data and Setting

- Within-Sentence baseline IE trained from 500 English ACE05 texts (from March – May of 2003)
- Use 10 ACE05 newswire texts as development set to optimize the global confidence thresholds and apply them for blind test
- Blind test on 40 ACE05 texts, for each test text, retrieved 25 related texts from TDT5 corpus (278,108 texts, from April-Sept. of 2003)
Selecting Trigger Confidence Thresholds
to optimize Event Identification F-measure on Dev Set
Selecting Argument Confidence Thresholds
to optimize Argument Labeling F-measure on Dev Set
# Experiments: Trigger Labeling

<table>
<thead>
<tr>
<th>System/Human</th>
<th>Performance</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Sent IE (Baseline)</td>
<td></td>
<td>67.6</td>
<td>53.5</td>
<td>59.7</td>
</tr>
<tr>
<td>After Cross-Sent Inference</td>
<td></td>
<td>64.3</td>
<td>59.4</td>
<td>61.8</td>
</tr>
<tr>
<td>After Cross-Doc Inference</td>
<td></td>
<td>60.2</td>
<td>76.4</td>
<td>67.3</td>
</tr>
<tr>
<td>Human Annotator 1</td>
<td></td>
<td>59.2</td>
<td>59.4</td>
<td>59.3</td>
</tr>
<tr>
<td>Human Annotator 2</td>
<td></td>
<td>69.2</td>
<td>75.0</td>
<td>72.0</td>
</tr>
<tr>
<td>Inter-Adjudicator Agreement</td>
<td></td>
<td>83.2</td>
<td>74.8</td>
<td>78.8</td>
</tr>
</tbody>
</table>
Experiments: Argument Labeling

<table>
<thead>
<tr>
<th>Performance</th>
<th>System/Human</th>
<th>Argument Identification</th>
<th>Argument Classification Accuracy</th>
<th>Argument Identification +Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>WithSent IE</td>
<td>47.8</td>
<td>38.3</td>
<td>42.5</td>
</tr>
<tr>
<td></td>
<td>After Cross-Sent Inference</td>
<td>54.6</td>
<td>38.5</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>After Cross-Doc Inference</td>
<td>55.7</td>
<td>39.5</td>
<td>46.2</td>
</tr>
<tr>
<td></td>
<td>Human Annotator 1</td>
<td>60.0</td>
<td>69.4</td>
<td>64.4</td>
</tr>
<tr>
<td></td>
<td>Human Annotator 2</td>
<td>62.7</td>
<td>85.4</td>
<td>72.3</td>
</tr>
<tr>
<td></td>
<td>Inter-Adjudicator Agreement</td>
<td>72.2</td>
<td>71.4</td>
<td>71.8</td>
</tr>
</tbody>
</table>
Related Work

Yarowsky, 1995
Common: Trigger Labeling “One Sense Per Discourse”
Diff: Extend it to across related documents

Yangarber et al., 2005-2007
Common: Use Global Inference to correct local IE (location name for disease outbreak domain)
Diff: On more general domain

Mann, 2007
Common: Use Global Inference to correct local IE (MUC CEO succession event)
Diff: On more general event types

Our Work
Conclusion and Future Work

- Proposed a new approach to break the document boundaries for IE
- Gather together background knowledge from a set of related documents, and then apply inference to enhance traditional IE performance
- Recent work proved the same approach can improve relation extraction and social network extraction

Future Work
- Use results as seeds for unsupervised/open IE
- Develop Event-driven multi-document summarization
- Derive entailment rules from related events in different timeframes
- Long term: perform essential information reasoning and event prediction
Thank you
# One Trigger Sense Per Cluster

<table>
<thead>
<tr>
<th>Event Trigger</th>
<th>Event Type</th>
<th>In Training Corpora</th>
<th>In Test Doc</th>
<th>In Related Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>advance</strong></td>
<td>Movement_ Transport</td>
<td>31% of 16</td>
<td>50% of 12</td>
<td>88.9% of 27</td>
</tr>
<tr>
<td><strong>fire</strong></td>
<td>Personnel_ End-Position</td>
<td>7% of 81</td>
<td>100% of 2</td>
<td>100% of 10</td>
</tr>
<tr>
<td></td>
<td>Conflict_ Attack</td>
<td>54% of 81</td>
<td>100% of 3</td>
<td>100% of 19</td>
</tr>
<tr>
<td><strong>replace</strong></td>
<td>Personnel_ End-Position</td>
<td>5% of 20</td>
<td>100% of 1</td>
<td>83.3% of 6</td>
</tr>
<tr>
<td><strong>form</strong></td>
<td>Business_ Start-Org</td>
<td>12% of 8</td>
<td>100% of 2</td>
<td>100% of 23</td>
</tr>
<tr>
<td><strong>talk</strong></td>
<td>Contact_ Meet</td>
<td>59% of 74</td>
<td>100% of 4</td>
<td>100% of 26</td>
</tr>
</tbody>
</table>
Selecting Trigger Confidence Thresholds
to optimize Event Identification F-measure on Dev Set
Selecting Argument Confidence Thresholds
to optimize Argument Labeling F-measure on Dev Set