Refining Information Extraction Through Unsupervised Cross-document Inference

Heng Ji and Ralph Grishman

Computer Science Department
New York University
April, 2008
Information Extraction and Its Role in GALE Distillation

Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment.

<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>

- 33 different types of ACE events
- 12%-52% improvement for mono-lingual GALE question answering (Schiffman et al., HLT07; Hakkani-Tür et al., Interspeech07; Levit et al., ASRU07)
- 12.7% improvement for cross-lingual GALE document retrieval (Hakkani-Tür et al., RANLP07)
Also held in Palu was Bali bomb suspect Ahmad Rohim, alias Saad.

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.
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

1. **Test Doc**
   - **Within-Sent Event Tagger**
   - **Cross-Sent Event Inference**

2. **INDRI IR**
   - **Cluster of Related Docs**
   - **Candidate Events & Confidence**

3. **Cross-Doc Inference**
   - **Related Events & Confidence**
   - **Refined Events**

4. **Within-Sent Event Tagger**
   - **Cross-Sent Inference**
Cross-Sent/Cross-doc Event Inference

- Within-Sentence IE system produces local confidence

- Document-wide and Cluster-wide Confidence
  - Frequency weighted by local confidence
  - Count frequency of trigger with a particular event type
  - For each argument and its coreferred names, count frequency of event type
  - For each argument and its coreferred names, count frequency of event type and role

- Inference Actions
  - Remove triggers and arguments with low local or global confidence
  - Adjust trigger and argument identification and classification to achieve global consistency
Experiments: ACE Event Extraction

- Training from 500 English ACE 05 texts, test on 40 texts
- For each test text, retrieved 25 related texts from TDT5
- Confidence thresholds optimized based on dev set
- Performance (Ji and Grishman, ACL08)

<table>
<thead>
<tr>
<th>System/Human</th>
<th>Performance</th>
<th>Trigger Labeling</th>
<th>Argument Labeling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Within-Sent IE</td>
<td></td>
<td>67.6</td>
<td>53.5</td>
</tr>
<tr>
<td>After Cross-Sent Inference</td>
<td></td>
<td>64.3</td>
<td>59.4</td>
</tr>
<tr>
<td>After Cross-Doc Inference</td>
<td></td>
<td>60.2</td>
<td>76.4</td>
</tr>
<tr>
<td>Human Annotator 1</td>
<td></td>
<td>59.2</td>
<td>59.4</td>
</tr>
<tr>
<td>Human Annotator 2</td>
<td></td>
<td>69.2</td>
<td>75.0</td>
</tr>
</tbody>
</table>
Experiments: GALE “Arrest” Template

- Test on 17 Y1 queries, 5812 snippets
- Evaluate whether a snippet sentence includes an arrest event or not
- Key was created by LDC + post corrections

Graph:
- Within-Sent Event Extraction
- +Cross-Sent/Cross-Doc Inference

Precision (%) vs. Recall (%)
- 91.5%
- 79.0%