Cross-document Event Extraction, Inference, Linking and Prediction

Heng Ji
hengji@cs.qc.cuny.edu
City University of New York
November 7, 2008
Motivation and Goal: Event Extraction Beyond Document Boundary

- Limitation of ‘Traditional’ IE: Within-document Event Extraction
  - Operate a sentence a time, generate large databases
    - Redundant: (half event instances can be pruned by simple merging)
    - Disconnect: poor event coreference; limited study on other relation types among events
    - Erroneous: state-of-the-art still hit ‘performance ceiling’
    - Incomplete: e.g. 2/3 event instances don’t include time arguments in text

- Recent Cross-doc IE Fashion
  - Researchers explored different angles but there is no standard task definition, annotated data and scoring metric for a fair comparison

- Our Goal
  - Respect INITIAL goals for IE and knowledge discovery: Create a database of facts from the entire corpus, allow further logical reasoning
  - Define a new cross-document event extraction task
  - Conduct event inference, linking, ranking and preliminary prediction
Project Architecture: More Specific Goals

**Within-document IE**

**Cross-document**

**Event Aggregation, Inference, Correction**

**Event Linking**

**Redundancy Removal**

**Event Ranking**

**Event Prediction**

**More Accurate**

"<SELLER>Vivendi</SELLER> has been trying to sell assets" +
"Blackstone would buy <SELLER>Vivendi</SELLER>'s theme park"
→"<SELLER>Vivendi</SELLER> earlier this week confirmed...that it planned to shed its entertainment assets by the end of the year."

**More Coherent**

"<SELLER>PLF</SELLER> attacked in boat in <SELLER>Italy</SELLER> in 1985" ↔
"<SELLER>Italy</SELLER> extradited <SELLER>Abbas</SELLER> in Apr 14, 2003"

**More Concise**

"2000 soldiers were killed in Lima Embassy" >>>"people died"

**More Complete**

"With marathon talks at the top world body failing late Thursday to reconcile French and Russian opposition to US-British war plans", "the United States upped its military presence, deploying more missile-firing warships for the war in the Red Sea". → "before Thursday"
Approach

- **Within-document IE**
  - Baseline to extract entities, relations and events (Grishman et al., 05; Florian et al., 06)

- **Cross-document Event Aggregation, Inference, Correction**
  - Cross-document inference using background knowledge (Ji and Grishman, 08; Yangarber, 06)
  - Apply Global Inference Learning (Roth and Yih, 04; Chambers and Jurafsky, 08) and Semi-supervised Learning (Ji and Grishman, 06)

- **Event Linking**
  - Time based Event Chains (Xue, 08; Chambers and Jurafsky, 08)
  - Person/Organization/Location Entity based Event Network
  - Causal Relation based Event Network (Bethard and Martin, 08; Girju, 03)

- **Redundancy Removal and Event Ranking**
  - Apply PageRank as used for IR and multi-document summarization (Nastase, 08), incorporating event and link weights

- **Event Prediction**
  - Perform probabilistic logical reasoning to predict missing events, focus on time argument prediction
Deliverables and Impact

- For NLP Research
  - A new definition of cross-document IE, annotation guideline, annotated corpus and scoring metrics
  - A high-performance cross-document IE system
  - A learning toolkit with semi-supervised learning based on confidence estimation
  - Direct Benefit other NLP tasks such as event-driven summarization, textual entailment and question-answering

- For NLP Education
  - Hands-on IE tutorials and implementation practices
  - Several PhD dissertations under supervision of proposers

- Pre-workshop Preparation
  - Task guideline, Data Annotation done
  - Pilot system done, will focus on improving event linking and prediction during the workshop
Proposers: Put All Recent Work Together

- **Senior Researchers**
  - Heng Ji and Ralph Grishman (ACL08, Cross-document Event Inference and Correction; English and Chinese IE systems)
  - Roman Yangarber (Cross-doc IE for Medical Domain)
  - Radu Florian (IBM IE System)
  - Nianwen Xue (EMNLP08, Temporal Inference)
  - Vincent Ng (ACL07, Semantics for Coreference)

- **Students**
  - Zheng Chen (PhD student, City University of New York)
  - Kazi Saidul Hasan (PhD student, University of Texas at Dallas)
  - Randolf Altmeyer (Undergraduate student, DFKI GmbH)

- **Other Potential Speakers**
  - Dan Jurafsky (ACL08, Narrative Event Chain Extraction)
  - Oren Etzioni (IJCAI05, Using Information Redundancy for IE)
  - Ellen Riloff (EMNLP07, Extracting patterns from Relevant Regions for IE)
<table>
<thead>
<tr>
<th>Tasks</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>11</td>
</tr>
<tr>
<td><strong>Pre-workshop</strong></td>
<td></td>
</tr>
<tr>
<td>Task Definition, Annotation Guideline</td>
<td></td>
</tr>
<tr>
<td>Data Annotation, start from ACE and Timebank</td>
<td></td>
</tr>
<tr>
<td>Tutorial Materials</td>
<td></td>
</tr>
<tr>
<td><strong>Pilot IE System</strong></td>
<td></td>
</tr>
<tr>
<td>Within-doc IE</td>
<td></td>
</tr>
<tr>
<td>Event Aggregation, Inference and Correction</td>
<td></td>
</tr>
<tr>
<td>Event Linking</td>
<td></td>
</tr>
<tr>
<td>Event Ranking</td>
<td></td>
</tr>
<tr>
<td>Event Prediction</td>
<td></td>
</tr>
<tr>
<td>Recruit Students</td>
<td></td>
</tr>
<tr>
<td><strong>During-workshop</strong></td>
<td></td>
</tr>
<tr>
<td>Tutorial Presentations</td>
<td></td>
</tr>
<tr>
<td>Event Linking</td>
<td></td>
</tr>
<tr>
<td>Event Prediction</td>
<td></td>
</tr>
<tr>
<td>Wrap-up and Report</td>
<td></td>
</tr>
</tbody>
</table>
Thank you
Related Work at This Meeting

- Time based Event Linking (Steven Bethard)
- Event Prediction (Noah Smith)
- Ngram Resources (Satoshi Sekine and Dekang Lin), case study to test effectiveness
- …
Output Examples: Person based Event Chain

Obama, Barack

Born: 1961  
Nationality: USA  
Occupation: Politician

- 1961 - Barack Obama was born on the 4th of August in Honolulu, Hawaii.
- 1963 - His parents got divorced. His mother married an Indonesian student, and when Obama was 6 years old they moved to Jakarta.
- 1971 - Obama returned to Hawaii to live with his maternal grandparents.
- 1979 - He was enrolled in the fifth grade at Punahou School where he graduated from high school at 18 years old.
- 1983 - He majored in political science in Columbia University, with a specialization in international relations.
- 1984 - He worked for a year at Business International Corporation before he moved to Chicago to a job with a non-profit organization local churches organize job training programs for residents of poor neighborhoods.
- 1991 - He was elected president of the Harvard Law Review.
- 1992 - Married to Michelle Obama on October 18th.
- 1996 - Elected to the Illinois State Senate representing the 13th District in the south side neighborhood of Hyde Park in Chicago.
- 2002 - He rededicated his efforts to the Illinois state Senate. In his campaign, he ran unopposed. Obama authored a law requiring police to videotape interrogations for crimes punishable by the death penalty. He also pushed through legislation that would force insurance companies to cover routine mammograms.
### Lehman Brothers

The following is a chronology of major events in the history of Lehman. (View [XML version](#)).

**Timeline anchors:** [1850, 1929, 2000]

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1914</td>
<td>Lehman Brothers established as a brokerage firm</td>
</tr>
<tr>
<td>1919</td>
<td>Lehman posts record-high net revenues</td>
</tr>
<tr>
<td>1929</td>
<td>Lehman shares sink 42 percent</td>
</tr>
<tr>
<td>1990</td>
<td>Lehman acquires Neuberger Berman and The Crossroads Group</td>
</tr>
<tr>
<td>2000</td>
<td>Lehman acquires Neuberger Berman and The Crossroads Group</td>
</tr>
</tbody>
</table>
### Wall Street's 8 brutal days

**Type:** Business  
**Significance Rank:** Four star  
**Source:** CNN  
**Location:** Wall Street, NY, US  
**Date:** 10/12/2008  
**Summary:** The Dow ended its worst week ever Friday and capped a staggering eight-session selloff that has seen the blue-chip index fall 2,400 points...[More]
Innovative Claims

- A new definition of Cross-doc IE
- Model cross-document event correction using machine learning
- Expand event linking work to more dimensions besides time
- First time to apply PageRank for IE
- First time to explore learning based event prediction
Time Correction and Prediction Example

- With marathon talks at the top world body failing late Thursday to reconcile French and Russian opposition to US-British war plans, the United States upped its military presence, deploying more missile-firing warships to the Red Sea.

→ “war” happened before Thursday.
Compare to previous causal relation detection and multi-document summarization

- Causal Relation Detection: we move from parsing to semantic based link detection
- Sentence linking and ranking for summarization: we don’t restrict to words and names, and we can even correct extracted facts while summarization has to stick to the original sentences (without re-writing)