Preemptive Information Extraction using Unrestricted Relation Discovery

Yusuke Shinyama
Satoshi Sekine
New York University
Quick Recipe for IE

- Specify the information you want.
  - e.g. “hiring and firing event”

- Create a system to extract the information.

- Get the results.

<table>
<thead>
<tr>
<th>Person</th>
<th>Position</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>President</td>
<td>PQR</td>
</tr>
<tr>
<td>Jones</td>
<td>Vice President</td>
<td>XYZ</td>
</tr>
</tbody>
</table>
IE Gone Wild

- We want to discover all the IE tasks that are potentially feasible.
- Create an IE system for every scenario *preemptively*.
- Specifying the information
  ⇒ Searching a relevant IE system (table).
Prior Work

• Pattern acquisition for predefined relation:
  - (Riloff, 96; Brin, 98; Yangarber et al, 00; Agichtein et al, 00)

• Unsupervised relation discovery:
  - (Hasegawa, 04; Chen et al, 05; Zhang et al, 05)
Preemptive IE - How?

- Overall procedure:
  1. Find a pair (or triple) of entities that have a similar relation in articles.
     - ex. Relation between “Katrina”-“New Orleans” in Article A is similar to relation between “Longwang”-“Taiwan” in Article B.
  2. Cluster relations based on their similarity.
  3. A cluster of similar relations = table.

<table>
<thead>
<tr>
<th>Article A</th>
<th>Katrina</th>
<th>New Orleans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article B</td>
<td>Longwang</td>
<td>Taiwan</td>
</tr>
</tbody>
</table>
Find Similar Relations

• Consider a simple case.
  - Named Entities (NE) are tagged.

  **Article A**
  Katrina **headed for** New Orleans.

  **Article B**
  Longwang **headed for** Taiwan.

Common expression
Find Similar Relations

- Consider a simple case.
  - Named Entities (NE) are tagged.

```
Article A
Katrina headed for New Orleans.
```
```
Article B
Longwang headed for Taiwan.
```

Common expression between NEs = They have similar relation.
Find Similar Relations

- Actual expressions are varied, but...

**Article A**

Katrina headed for ... New Orleans was hit ...
Katrina threatened ... New Orleans residents ...
Katrina is category 5 ... evacuated New Orleans

**Article B**

Longwang hit ... Taiwan’s coast ...
Longwang headed for ... Taiwan was pounded...
Longwang is swirling ... Taiwan was hit...
Find Similar Relations

• There might be common expressions.

**Article A**

Katrina headed for ... New Orleans was hit ...
Katrina threatened ... New Orleans residents ...
Katrina is category 5 ... evacuated New Orleans

Common expression
(Disjoint, but close)

**Article B**

Longwang hit ... Taiwan ’s coast ...
Longwang headed for ... Taiwan was pounded ...
Longwang is swirling ... Taiwan was hit ...
Find Similar Relations

- Parallel correspondence of expressions = similar relations.

**Article A**

Katrina headed for New Orleans was hit...
Katrina threatened ...
New Orleans residents ...
Katrina is category 5 ...
Evacuated New Orleans

**Article B**

Longwang hit ...
Taiwan ’s coast ...
Longwang headed for ...
Taiwan was pounded ...
Longwang is swirling ...
Taiwan was hit ...

Parallel correspondence of expressions = similar relations.
Find Similar Relations

• Unrestricted Relation Discovery:

1. Collect all the expressions (patterns) that modify each NE.

2. For every pair of articles:
   - Try to find a parallel correspondence of patterns between multiple NEs.
   - Exist? → These NEs must have similar relations.
   - Cluster similar relations.

- We have a bunch of similar relations:
  - Don’t know what they are, but show them as a table.
Multiplying Patterns

• Reality:
  - The number of patterns is not enough.

• Solution:
  1. Use a cluster of articles reporting the same event instead of a single article. (*basic cluster*)
  2. Cluster basic clusters that have similar relations. (*metacluster*)
    - Metaclusters = IE tables.
Multiplying Patterns

- Use multiple news sources and coreference resolution.

- **Newspaper A1**
  - Katrina headed for...
  - New Orleans was hit...

- **Newspaper A2**
  - Katrina threatened...

- **Newspaper B1**
  - Longwang hit...
  - Taiwan coastal...

- **Newspaper B2**
  - Taiwan was pounded...

- **Basic Cluster A**

- **Basic Cluster B**

- **Article B3**

- **corref**
Pattern Format

- GLARF: predicate argument structure. (Meyers et al. 01)
  - Graph representation, regularized.
  - Pattern: subgraph which spans NE nodes.

\texttt{PER+SBJ:hit}

“Katrina hit Louisiana’s coast.”
Metaclustering

Metacluster :

Basic cluster

Pattern

Pattern

Pattern

...
Implementation

• Obtain news articles from multiple news sites on the Web.
• Basic Cluster: Vector space model.
• NE tagging & Coref. Resolution: Use an in-house NE tagger and coreference resolver.
• Generate GLARF structures; weight obtained patterns with ICF.
Implementation

• System Overview
  1. Crawling the Web
  2. Clustering articles *(basic clusters)*
  3. Parsing & NE tagging
  4. Coreference Resolution...

- Crawling
- Clustering
- Parsing
- NE & Coref.

Newspapers
Implementation

- System Overview (cont.)
  5. GLARF generation (patterns)
  6. Metaclustering
    - Use patterns as features of a basic cluster.

Katrina hit Louisiana’s coast.

Longwang hit Taiwan.
# Experiments

<table>
<thead>
<tr>
<th>Source articles (2 months)</th>
<th>28,009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic clusters (≥ 3 articles)</td>
<td>5,543</td>
</tr>
<tr>
<td>Patterns - token</td>
<td>643,767</td>
</tr>
<tr>
<td>Patterns - type</td>
<td>7,990</td>
</tr>
<tr>
<td>Metaclusters</td>
<td>302</td>
</tr>
<tr>
<td>Metaclusters (≥ 3 events)</td>
<td>101</td>
</tr>
</tbody>
</table>

We used 12 news sites in U.S.
Evaluation

• We evaluated 48 tables manually.
  
  - Try to find a sentence that explains at least 2/3 of the rows in the table (”a consistent table”).
  
  - For consistent tables, count how many rows fit the explanation.

<table>
<thead>
<tr>
<th>Consistent</th>
<th>36 (75%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rows fitted</th>
<th>118 (73%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows not fitted</td>
<td>43</td>
</tr>
<tr>
<td>Total</td>
<td>161</td>
</tr>
</tbody>
</table>

Tables          Rows

20/24
## Evaluation

<table>
<thead>
<tr>
<th>Table Description</th>
<th>Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storm PERSON probably affected GPE.</td>
<td>8/16</td>
</tr>
<tr>
<td>Nominee PERSON must be confirmed by ORG.</td>
<td>4/7</td>
</tr>
<tr>
<td>PERSON urges GPE to make changes.</td>
<td>4/6</td>
</tr>
<tr>
<td>GPE launched an attack in GPE.</td>
<td>3/5</td>
</tr>
<tr>
<td>PERSON ran against PERSON in an election.</td>
<td>4/5</td>
</tr>
<tr>
<td>PERSON visited GPE on a diplomatic mission.</td>
<td>2/4</td>
</tr>
<tr>
<td>PERSON beat PERSON in golf.</td>
<td>4/4</td>
</tr>
<tr>
<td>GPE soldier(s) were killed in GPE.</td>
<td>3/3</td>
</tr>
<tr>
<td>PERSON ran for governor of GPE.</td>
<td>2/3</td>
</tr>
<tr>
<td>Boxer PERSON fought boxer PERSON.</td>
<td>3/3</td>
</tr>
</tbody>
</table>
Error Analysis

- Incorrect rows (out of 10)
  - 4 rows: coreference resolution error (referring to wrong NEs.)
  - 4 rows: patterns are distant to each other.
  - 1 row: parse error.
  - 1 row: obscure.

  "Hamas’ most senior leader, Mahmoud al-Zahar, announced ... Sharon faced a vote in his party ..."
Conclusion

- We proposed Preemptive Information Extraction.
  - Find parallel correspondence of patterns between multiple entities.
  - Use clustering to build tables.
- Obtained a reasonable result.
  - Possible to prepare many tables offline: Elections, Sports results, M&As, ...
Future Work

• Questions:
  - How NE types and pattern format could limit types of news articles?
  - How much does each stage affect the overall performance?

• Improvements:
  - Use various NE types:
    ■ DATE, CURRENCY, NUMBER, ...
  - Decent evaluation (ACE event corpus).