Information Extraction

- Today
  - Learning approaches
    - Weakly supervised methods
    - Fully automatic methods for IE
  - Named entity identification

Syntactico-semantic patterns

The twister occurred without warning at approximately 7:15p.m. and destroyed two mobile homes.

Pattern:
- Trigger: “destroyed”
- Slot: Damaged-Object
- Position: direct-object
- condition: physical-object?

from Cardie [1997]

Pattern templates

Noun phrase extraction only

\[
\begin{align*}
&\text{<subject> <passive-verb>} \quad \text{<subject> <active-verb>} \\
&\text{<subject> <infinitive-verb>} \quad \text{<subject> <auxiliary-verb>}<\text{noun>}
\end{align*}
\]

*\text{<passive-verb> <dobj>}\quad \text{<active-verb> <dobj>} \\
\text{<infinitive> <dobj>} \quad \text{<passive-verb> <dobj>} \\
\text{<verb>+<infinitive> <dobj>} \quad \text{<verb>+<prepositional phrase>}

\[
\begin{align*}
&\text{<victim> was murdered} \\
&\text{<perpetrator> bombed} \\
&\text{<perpetrator> attempted to kill} \\
&\text{<victim> was victim}
\end{align*}
\]

Autoslog algorithm

- For each annotated “string fill”, $s$, in the training data
  - (Shallow) parse the sentence that contains $s$.
  - Apply the syntactic pattern templates in order. Execute the first one that applies to determine:
    - the trigger word
    - the triggering constraints (syntactic)
    - the position of phrase to be extracted (grammatical role)

- Determine slot type
  - The annotated slot type for $s$ in the training corpus

- Determine the semantic constraints
  - Defined a priori based on typical semantic class of fillers

- Create and save the extraction pattern
Results

- 1500 texts, 1258 answer keys
- 4780 slots (6 types)
- Autoslog generated 1237 patterns
- After human filtering: 450 patterns
- Compare to manually built patterns

<table>
<thead>
<tr>
<th>System/Data Set</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual/TST3</td>
<td>46</td>
<td>56</td>
<td>50.51</td>
</tr>
<tr>
<td>Autoslog/TST3</td>
<td>43</td>
<td>56</td>
<td>48.65</td>
</tr>
<tr>
<td>Manual/TST4</td>
<td>44</td>
<td>40</td>
<td>41.90</td>
</tr>
<tr>
<td>Autoslog/TST4</td>
<td>39</td>
<td>45</td>
<td>41.79</td>
</tr>
</tbody>
</table>

IE Example: Output Template

1. DATE                  10 NOV 88
2. LOCATION          CHILE: SANTIAGO (CITY)
3. TYPE                 MURDER
4. STAGE OF EXECUTION   ACCOMPLISHED
5. INCIDENT CATEGORY    TERRORIST ACT
6. PERP: INDIVIDUAL ID  "THEY"
7. PERP: CONFIDENCE     REPORTED AS FACT
8. HUM TGT: DESCRIPTION "BIRDS"
9. HUM TGT: TYPE        CIVILIAN: "BIRDS"
10. HUM TGT: NUMBER     2: "BIRDS"
11. EFFECT OF INCIDENT  DEATH: "BIRDS"
12. INSTRUMENT          STONE

IE Example: Input Text

SANTIAGO, 10 NOV 88 (QUE PASA) -- [TEXT] [CONTINUED]


Information Extraction

- Today
  - Learning approaches
    - Weakly supervised methods
    - Fully automatic methods for IE
  - Named entity identification
**Autoslog-TS**

- Largely unsupervised
- Two sets of documents: relevant, not relevant
- Apply pattern templates to extract every NP in the texts
- Compute **relevance rate** for each pattern $i$:

  $$Pr(\text{relevant text} \mid \text{text contains } i) = \frac{\text{freq of } i \text{ in relevant texts}}{\text{frequency of } i \text{ in corpus}}$$

- Sort patterns according to relevance rate and frequency

  $$\text{relevance rate} \times \log(\text{freq})$$

**Information Extraction**

- **Today**
  - Learning approaches
    - Weakly supervised methods
    - Fully automatic methods for IE
  - Named entity identification

**Covering algorithms**

- **E.g. Crystal** [Soderland et al. 1995]
  - Allows for more complicated patterns
    - Can test target NP or any constituent in its context for
      - presence of any word or sequence of words
      - semantic class of heads or modifiers
  - Crystal is a “covering” algorithm
  - Successively generalizes the patterns derived from input examples until the generalization produces errors

**Autoslog-TS**

- Human review of learned patterns is still required
- Also requires, for each pattern, the manual labeling of the semantic category of the extracted slot filler
Information Extraction

- Today
  - Learning approaches
    - Weakly supervised methods
    - Fully automatic methods for IE

Named entity extraction

NE Identification

- Identify all named locations, named persons, named organizations, dates, times, monetary amounts, and percentages.

Guidelines need to be specified

- *The Wall Street Journal*: artifact or organization?
- *White House*: organization or location?
- Is a street name a location?
- Should *yesterday* and *last Tuesday* be labeled as dates?
- Is *mid-morning* a time?

Examples

Figure 1.1 Examples. Examples of correct labels for English text and for Spanish text.

Figure 2.1 English Examples. Finding names ranges from the easy to the challenging. Company names are in boldface. It is crucial for any name-finder to deal with the underlined text.
### NE Results Using HMM’s

#### Table 5.1 F-measure Scores

This table illustrates Identifier’s performance as compared to the best reported scores for each category.

<table>
<thead>
<tr>
<th>Language</th>
<th>Best Rules</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed Case, English (WSJ)</td>
<td>96.4</td>
<td>94.9</td>
</tr>
<tr>
<td>Upper Case, English (WSJ)</td>
<td>89</td>
<td>93.6</td>
</tr>
<tr>
<td>Speech Form, English (WSJ)</td>
<td>74</td>
<td>90.7</td>
</tr>
<tr>
<td>Mixed Case, Spanish</td>
<td>93</td>
<td>90</td>
</tr>
</tbody>
</table>

### Class/Tag Values

- **B** – begins a PersonName, Loc, etc.
- **I** – inside a PersonName, Loc, etc.
- **O** – outside a PersonName, Loc, etc.

#### HMMs for entity detection

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td>B&lt;sub&gt;ORG&lt;/sub&gt;</td>
</tr>
<tr>
<td>Airlines</td>
<td>I&lt;sub&gt;ORG&lt;/sub&gt;</td>
</tr>
<tr>
<td>a</td>
<td>O</td>
</tr>
<tr>
<td>of</td>
<td>O</td>
</tr>
<tr>
<td>of AMR</td>
<td>B&lt;sub&gt;ORG&lt;/sub&gt;</td>
</tr>
<tr>
<td>Corp.</td>
<td>I&lt;sub&gt;ORG&lt;/sub&gt;</td>
</tr>
<tr>
<td>immediately</td>
<td>O</td>
</tr>
<tr>
<td>matched</td>
<td>O</td>
</tr>
<tr>
<td>the</td>
<td>O</td>
</tr>
<tr>
<td>move</td>
<td>O</td>
</tr>
<tr>
<td>plummeting</td>
<td>O</td>
</tr>
<tr>
<td>spokesman</td>
<td>O</td>
</tr>
<tr>
<td>Tim</td>
<td>B&lt;sub&gt;PER&lt;/sub&gt;</td>
</tr>
<tr>
<td>Wagner</td>
<td>I&lt;sub&gt;PER&lt;/sub&gt;</td>
</tr>
<tr>
<td>said</td>
<td>O</td>
</tr>
<tr>
<td>.</td>
<td>O</td>
</tr>
</tbody>
</table>

Figure, copyright J&M 2nd ed