Automatically Constructing a Dictionary for Information Extraction Tasks

Ellen Riloff

Overview

- Problem: Domain specific dictionary
  - Portability
  - Practicality
- Solution: AutoSlog
- Automatic Training Approach
  - Representative corpus
  - Terrorist event descriptions
Information Extraction

- Selective Concept Extraction
- CIRCUS: Sentence Analyzer
- Domain-specific Dictionary
- Concept node definition
  - Enabling conditions
  - Slots (e.g., victims, perpetrators, ...)
    - Syntactic expectation
    - Hard and soft constraints
Extraction Procedure

- Sentence
- CIRCUS
- Concept Nodes
- Concept Node Dictionary
Observations

News reports follow certain stylistic conventions.

- Assumption: the first reference to a targeted piece of information is most likely where the relationship between that information and the event is made explicit.

The immediate linguistic context surrounding the targeted information usually contains the words or phrases that describe its role in the events.
### Conceptual Anchor Points

- A conceptual anchor point is a word that should activate a concept.
  - **CIRCUS:** trigger word

<table>
<thead>
<tr>
<th>Linguistic Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;subject&gt;</code> passive-verb</td>
<td><code>&lt;victim&gt;</code> was murdered</td>
</tr>
<tr>
<td><code>&lt;subject&gt;</code> active-verb</td>
<td><code>&lt;perpetrator&gt;</code> bombed</td>
</tr>
<tr>
<td><code>&lt;subject&gt;</code> verb infinitive</td>
<td><code>&lt;perpetrator&gt;</code> attempted to kill</td>
</tr>
<tr>
<td><code>&lt;subject&gt;</code> auxiliary noun</td>
<td><code>&lt;victim&gt;</code> was victim</td>
</tr>
<tr>
<td>passive-verb <code>&lt;dobj&gt;</code></td>
<td>killed <code>&lt;victim&gt;</code></td>
</tr>
<tr>
<td>active-verb <code>&lt;dobj&gt;</code></td>
<td>bombed <code>&lt;target&gt;</code></td>
</tr>
<tr>
<td>infinitive <code>&lt;dobj&gt;</code></td>
<td>to kill <code>&lt;victim&gt;</code></td>
</tr>
<tr>
<td>verb infinitive <code>&lt;dobj&gt;</code></td>
<td>threatened to attack <code>&lt;target&gt;</code></td>
</tr>
<tr>
<td>gerund <code>&lt;dobj&gt;</code></td>
<td>killing <code>&lt;victim&gt;</code></td>
</tr>
<tr>
<td>noun auxiliary <code>&lt;dobj&gt;</code></td>
<td>fatality was <code>&lt;victim&gt;</code></td>
</tr>
<tr>
<td>noun prep <code>&lt;np&gt;</code></td>
<td>bomb against <code>&lt;target&gt;</code></td>
</tr>
<tr>
<td>active-verb prep <code>&lt;np&gt;</code></td>
<td>killed with <code>&lt;instrument&gt;</code></td>
</tr>
<tr>
<td>passive-verb prep <code>&lt;np&gt;</code></td>
<td>was aimed at <code>&lt;target&gt;</code></td>
</tr>
</tbody>
</table>

**13 Heuristics**
Automated Dictionary Construction

**Input**
- A set of training text
- Answer keys

**Output:** a set of concept node definitions

**Procedure:** how to generate a concept node definition?
A Good Concept Node Def.

Id: DEV-MUC4-0657  Slot filler: “public buildings”
Sentence: (in la oroya, junin department, in the central peruvian mountain range, public buildings were bombed and a car-bomb was detonated.)

CONCEPT NODE
Name: target-subject-passive-verb-bombed
Trigger: bombed
Variable Slots: (target (*S* 1))
Constraints: (class phys-target *S*)
Constant Slots: (type bombing)
Enabling Conditions: ((passive))
Domain Specifications

- A set of mappings from template slots to concept node slots
- Hard and soft constraints for each type of concept node slot
- A set of mappings from template types to concept node types
Another Good Concept Node Def.

**Id:** DEV-MUC4-0071

**Slot filler:** “guerrillas”

**Sentence:** (the salvadoran guerrillas on mar. 12 89, today, threatened to murder individuals involved in the mar. 19 88 presidential elections if they do not resign from their posts.)

**CONCEPT NODE**

**Name:** perpetrator-subject-verb-infinitive-threatened-to-murder

**Trigger:** murder

**Variable Slots:** (perpetrator (*S* 1))

**Constraints:** (class perpetrator *S*)

**Constant Slots:** (type perpetrator)

**Enabling Conditions:** ((active)

(trigger-preceded-by? ’to ’threatened))
A Bad Concept Node Def.

**Id:** DEV-MUC4-1192  
**Slot filler:** “gilberto molasco”

**Sentence:** (they **took** 2-year-old gilberto molasco, son of patricio rodriguez, and 17-year-old andres argueta, son of emimesto argueta.)

**CONCEPT NODE**

- **Name:** victim-active-verb-dobj-took
- **Trigger:** **took**
- **Variable Slots:** (victim (*DOBJ* 1))
- **Constraints:** (class victim *DOBJ*)
- **Constant Slots:** (type kidnapping)
- **Enabling Conditions:** ((active))
Reasons of Bad Definitions

- When a sentence contains the targeted string but does not describe the event.
- When a heuristic proposes the wrong conceptual anchor point.
- When CIRCUS incorrectly analyzes the sentence.
Evaluation

- Terrorist Event Description
- MUC-4 Corpus
  - 1500 texts
  - 1258 answer keys
- A human (5 hours)
  - Keeps: 450
  - 4280 string fillers
  - 1237 concept node definitions
### Comparative Results

<table>
<thead>
<tr>
<th>System/Test Set</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC-4/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>MUC-4/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>
Conclusions

- Achieve Above 90% performance of the hand-crafted dictionary (1500 person hours)
- Improve performance
- Create dictionary from scratch
- No explicit domain theory is needed
- Need an annotated training corpus
- Make progress for scalable and portable IE systems