CS544: Information Extraction

January 18, 2011

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

Who is Jerry Hobbs?

Search Engine

Jerry R. Hobbs
Address: USC/ISI 4676 Admiralty Way

AMW.com

Jerry Hobbs - FugiDve
America's Most Wanted is a long-running American TV show produced by 20th Century Fox, and is the longest-running program of any kind in the history of the world.

Alleged Killer Dad Denied Bond, Services Set for Two LiVle Girld.

A judge had denied bond for Jerry Hobbs, who is accused of killing his 9-year-old in this undated...

Jerry Hobbs, who is accused of killing his 9-year-old, is a taekwondo teacher?

Dr. Jerry R. Hobbs (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, world martial arts games. Rank: 3rd Dan, Discipline: Ju-Jutsu, Issued By: Dentokan Martial Arts Association / Roy.
Named Entity Discrimination

- Discover the underlying meaning of a proper name
- The number of entities and their meaning is unknown

Knowledge Harvesting

- Given the name Jerry Hobbs, learn automatically from the Web semantic classes and members similar to it

Information Extraction
What is “Information Extraction”?  

- **Goal:** identify specific pieces of information from the content of unstructured or semi-structured textual documents.  
- **Input:**  
  - scenario of extraction (template schema to be filled)  
  - document collection  
- **Output:**  
  - a set of instantiated templates

---

A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported.  
According to unofficial sources, the bomb—allegedly detonated by urban guerrilla commandos—blew up a power tower in the northwestern part of San Salvador at 0650.

| Incident type: | bombing |
| Date:         | January 18, 2011 |
| Time:         | 0650 |
| Location:     | San Salvador (city) |
| Perpetrator:  | urban guerrilla commandos |
| Physical target: | power tower |
| Human target: | - |
| Effect on physical target: | destroyed |
| Effect on human target: | no injury or death |
| Instrument:   | bomb |

---

A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported.  
According to unofficial sources, the bomb—allegedly detonated by urban guerrilla commandos—blew up a power tower in the northwestern part of San Salvador at 0650.

| Incident type: | bombing |
| Date:         | January 18, 2011 |
| Time:         | 0650 |
| Location:     | San Salvador (city) |
| Perpetrator:  | urban guerrilla commandos |
| Physical target: | power tower |
| Human target: | - |
| Effect on physical target: | destroyed |
| Effect on human target: | no injury or death |
| Instrument:   | bomb |
MUC

- Message Understanding Conference (MUC) was an annual competition for IE systems funded by DARPA
    - Messages about naval operations
  - MUC-3 (1991) and MUC-4 (1992)
    - News articles about terrorist attacks
  - MUC-5 (1993)
    - News articles about joint ventures and microelectronics
  - MUC-6 (1995)
    - News articles about management changes
  - MUC-7 (1997)
    - News articles about space vehicle and missile launches
SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards, preferably Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards, preferably Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com
Two general approaches to IE

Pattern-Based Systems

Use patterns and rules which are applied to text.

Machine Learning

Use sequence tagging models to classify individual tokens as to whether or not they should be extracted.

A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb-allegedly detonated by urban guerrilla commandos blew up a power tower in the north western part of San Salvador at 0650.

<table>
<thead>
<tr>
<th>Incident type:</th>
<th>bombing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date:</td>
<td>January 18, 2011</td>
</tr>
<tr>
<td>Time:</td>
<td>0650</td>
</tr>
<tr>
<td>Location:</td>
<td>San Salvador (city)</td>
</tr>
<tr>
<td>Perpetrator:</td>
<td>urban guerrilla commandos</td>
</tr>
<tr>
<td>Physical target:</td>
<td>power tower</td>
</tr>
<tr>
<td>Human target:</td>
<td>-</td>
</tr>
<tr>
<td>Effect on physical target:</td>
<td>destroyed</td>
</tr>
<tr>
<td>Effect on human target:</td>
<td>no injury or death</td>
</tr>
<tr>
<td>Instrument:</td>
<td>bomb</td>
</tr>
</tbody>
</table>

LaSIE Information Extraction System

Gazetteers

Lexicon

Stratified grammar

Conceptual hierarchy

Gazetteer lookup

Sentence splitter

Brill tagger

Tagged Morph

bUCharT parser

Name matcher

Discourse interpreter

Template writer
LaSIE Information Extraction System

Tokenization - identify word boundaries in text
- white spaces indicate token boundaries
- full stops indicate sentences boundaries
  (not always true for example, 1. September; Nov. 1998)

Gazetteer Lookup – recognize phrases and keywords related to named entities which were previously stored in its lists (gazetteers)

A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb-allegedly detonated by urban guerrilla commandos blew up a power tower in the north western part of San Salvador at 0650.
LaSIE Information Extraction System

Gazetteer Lookup – recognize phrases and keywords related to named entities which were previously stored in its lists (gazetteers)
  • advantage – simple, fast, language independent as one just has to create the lists
  • disadvantage – collection and maintenance is time consuming, cannot deal with name variants, cannot resolve ambiguity

Sentence splitter - given a text, returns a list of strings where each element is a sentence.
  • uses a set of rules like the occurrences of "", "?" and "!" are indicators of sentence delimiters
  (not so simple, the "." in "B. Clinton" or "U.S." does not have this role)

Sentence1:
A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported.

Sentence2:
According to unofficial sources, the bomb-allegedly detonated by urban guerrilla commandos blew up a power tower in the north western part of San Salvador at 0650.
Sentence 1: A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported.

Sentence 2: According to unofficial sources, the bomb-allegedly detonated by urban guerrilla commandos blew up a power tower in the northwestern part of San Salvador at 0650.
According to unofficial sources, the bomb-allegedly detonated by urban guerrilla commandos blew up a power tower in the northwestern part of San Salvador at 0650.

NE1

NE2

Event(E1), detonate(E1,XX), urban_guerrilla_commandos(X), bomb(Y)

Event(E2), blow_up(E2,XX), power_tower(Z), location_of(Z,NE1), time_of(E2,NE2)
Name matcher – does not recognize new proper names, just adds identity relations between those found by the parser

• first token of the name matches the second name
  “Pepsi Cola” equals “Pepsi”
• one of the names is an acronym of the other
  “ISI” is equivalent to “Information Sciences Institute”
• one name is a reversal of the other
  “Defense Department” equals “Department of Defense”
• one name consists of concatenated contractions of the other
  “Pan America” equals “Pan Am”

Discourse interpreter – translates the semantic representations produced by the parser into

• representation of instances, their ontological classes and attributes
• coreference resolution

Output template generation
• procedure that writes the templates in the desired format
How well does this work?:

- Evaluate system’s performance on independent manually-annotated test data which was not used during system development
- IE systems are typically evaluated in terms of Precision (P) and Recall (R)

\[
P = \frac{\text{correctly extracted facts}}{\text{extracted facts}}
\]

\[
R = \frac{\text{correctly extracted facts}}{\text{correct facts}}
\]

\[
F = \frac{2PR}{P + R}
\]

LaSIE in MUC-6

<table>
<thead>
<tr>
<th>Task</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Named Entity</td>
<td>.94</td>
<td>.84</td>
</tr>
<tr>
<td>Co-reference resolution</td>
<td>.71</td>
<td>.51</td>
</tr>
<tr>
<td>Template Elements</td>
<td>.74</td>
<td>.66</td>
</tr>
<tr>
<td>Scenario Templates</td>
<td>.73</td>
<td>.37</td>
</tr>
</tbody>
</table>

LaSIE Named Entity

- Results for the Named Entity task over 30 texts
- Each setting indicates the contribution of LaSIE’s components

<table>
<thead>
<tr>
<th>No.</th>
<th>Setting</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gazetteer Look Up</td>
<td>.74</td>
<td>.37</td>
</tr>
<tr>
<td>2</td>
<td>1 + Parsing</td>
<td>.93</td>
<td>.80</td>
</tr>
<tr>
<td>3</td>
<td>2 + Name matching</td>
<td>.93</td>
<td>.88</td>
</tr>
<tr>
<td>4</td>
<td>3 + Discourse interpretaion</td>
<td>.93</td>
<td>.89</td>
</tr>
</tbody>
</table>
Can I test an existing IE system?

http://services.gate.ac.uk/annie/index.jsp

ANNIE is one of many Information Extraction systems that have been developed using GATE. It uses finite state algorithms and the JAPE language. This demo shows ANNIE recognising entities in texts.

Note: this demo uses a default set of components and IE resources; your mileage may vary! Also, complex HTML structures may prevent the system from being able to analyse the text they contain. This system does name recognition; see the IE User Guide for details of other forms of IE, and issues of domain-specificity and porting. Contact us about our cross-domain, multi-genre systems.

To use ANNIE, enter a URL in the box below. Select the types of entities that you would like to mark. GATE will then retrieve the document and extract the required information. This process may take a few seconds.

Enter a URL: [http://www.thespa.co.uk/poly/123](http://www.thespa.co.uk/poly/123)
- Person
- Location
- Organization
- Date
- Address
- Money
- Percent

Run analysis
Rule-based IE: Pros and Cons

**PROS:**
- clearly understood technology
- hand-written rules are relatively precise
- people can write rules with a reasonable amount of training

**CONS:**
- rules need to be written by hand
- requires experienced grammar developers
- difficult to port to a different domain

Can we automatically learn IE?

- In the mid-1990’s supervised IE systems were created.
- Supervised learning requires annotated training data.
- Trade-off: annotating texts vs. manual knowledge engineering
  - weeks vs. months
  - domain experts vs. computational linguists

Annotating Texts for IE

Alleged guerilla urban commandos launched two highpower bombs against a car dealership in downtown San Salvador this morning. A police report said that the attack set the building on fire but did not result any causalities.
Annotating Texts for IE

Alleged guerilla urban commandos launched two highpower bombs against a car dealership in downtown San Salvador this morning. A police report said that the attack set the building on fire but did not result any causalities.

Pattern Learning Algorithms

- A variety of different pattern/rule representations have been developed, but very commonly:
  - IE systems learn patterns by beginning with highly specialized patterns and iteratively generalizing them.
  - rule-learning stops when a set of patterns has been generated to sufficiently “cover” the training examples.

AutoSlog [Riloff 1993]

Annotated text

Parser

Syntactic Rules

Extraction Pattern: `<target> was bombed by <perpetrator>`

What happens if you apply on new text?

- Kuwait was bombed
- Newcastle was bombed
- Hiroshima was bombed
- Pearl Harbor was bombed
- ... Belgrade was bombed
Examples of learned patterns with AutoSlog:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;subject&gt; passive-vp</td>
<td>&lt;target&gt; was bombed</td>
</tr>
<tr>
<td>&lt;subject&gt; active-vp</td>
<td>&lt;perpetrator&gt; bombed</td>
</tr>
<tr>
<td>&lt;subject&gt; active-vp dobj</td>
<td>&lt;perpetrator&gt; threw dynamite</td>
</tr>
<tr>
<td>&lt;subject&gt; active-vp infinitive</td>
<td>&lt;perpetrator&gt; tried to kill</td>
</tr>
<tr>
<td>&lt;subject&gt; passive-vp infinitive</td>
<td>&lt;perpetrator&gt; was hired to kill</td>
</tr>
<tr>
<td>&lt;subject&gt; auxiliary dobj</td>
<td>&lt;victim&gt; was fatality</td>
</tr>
<tr>
<td>active-vp &lt;dobj&gt;</td>
<td>bombed &lt;target&gt;</td>
</tr>
<tr>
<td>infinitive &lt;dobj&gt;</td>
<td>to kill &lt;victim&gt;</td>
</tr>
<tr>
<td>active-vp infinitive &lt;dobj&gt;</td>
<td>tried to kill &lt;victim&gt;</td>
</tr>
<tr>
<td>passive-vp infinitive &lt;dobj&gt;</td>
<td>was hired to kill &lt;victim&gt;</td>
</tr>
<tr>
<td>subject auxiliary &lt;dobj&gt;</td>
<td>fatality was &lt;victim&gt;</td>
</tr>
<tr>
<td>passive-vp prep &lt;np&gt;</td>
<td>was killed by &lt;perpetrator&gt;</td>
</tr>
<tr>
<td>active-vp prep &lt;np&gt;</td>
<td>exploded in &lt;target&gt;</td>
</tr>
<tr>
<td>infinitive prep &lt;np&gt;</td>
<td>to kill with &lt;weapon&gt;</td>
</tr>
<tr>
<td>noun prep &lt;np&gt;</td>
<td>assassination of &lt;victim&gt;</td>
</tr>
</tbody>
</table>