CS544: Information Extraction, Named Entity Recognition and Classification

March 11, 2010

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva
Named Entity Recognition and Classification

- Identify mentions in text and classify them into a predefined set of categories of interest:
  - Person Names: Prof. Jerry Hobbs, Jerry Hobbs
  - Organizations: Hobbs corporation, FbK
  - Locations: Ohio
  - Date and time expressions: February 2010
  - E-mail: mkg@gmail.com
  - Web address: www.usc.edu
  - Names of drugs: paracetamol
  - Names of ships: Queen Marry
  - Bibliographic references:
  - ...

Prof. Jerry Hobbs taught CS544 during February 2010.
Jerry Hobbs killed his daughter in Ohio.
Hobbs corporation bought FbK.

Named Entity Discrimination

- Discover the underlying meaning of a proper name

Prof. Jerry Hobbs taught CS544 during February 2010.
Jerry Hobbs killed his daughter in Ohio.
Hobbs corporation bought FbK.

- The number of clusters and their meaning is unknown
Hypernym and Hyponym Learning

- Given an instance (e.g. Jerry Hobbs), learn automatically from the Web corresponding hypernyms and hyponyms

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

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
    • 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
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: March 11, 2010
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
### Today

**http://www.factual.com/**

<table>
<thead>
<tr>
<th>Video Game</th>
<th>Platform</th>
<th>Publisher</th>
<th>Developer</th>
<th>Genre</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 on 3 NHL Arcade</td>
<td>XBox 360</td>
<td>EA Sports</td>
<td>EA Canada</td>
<td>Sports</td>
<td>Hockey</td>
</tr>
<tr>
<td>4x4 Evo 2</td>
<td>XBox</td>
<td>7K Games</td>
<td>Terminal Reality, Inc.</td>
<td>Sports: Driving/Racing</td>
<td></td>
</tr>
<tr>
<td>2K: The Game</td>
<td>PlayStation 2</td>
<td>2K Games</td>
<td>SCI Studio Cambridge</td>
<td>Action</td>
<td>Espionage</td>
</tr>
<tr>
<td>2K to Life</td>
<td>PlayStation 2</td>
<td>Eidos Interactive</td>
<td>Avalanche Software LLC, Inc</td>
<td>Action: Shooter</td>
<td>Crime</td>
</tr>
<tr>
<td>50 Cent: Bulletproof</td>
<td>XBox</td>
<td>Vivendi Universal</td>
<td>Genuine Games Ltd.</td>
<td>Action: Shooter</td>
<td>Crime</td>
</tr>
<tr>
<td>1942: Joint Strike</td>
<td>XBox 360</td>
<td>Backbone Entertainment</td>
<td>Curt-Op Classics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abuse</td>
<td>PC</td>
<td>Alive Medisoft; Jungle Stud</td>
<td>Crack dot Com</td>
<td>Action; Shooter</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>AC/DC Live: Rock Band Track Pak</td>
<td>XBox 360</td>
<td>Electronic Arts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ace Combat 5: The Unsung War</td>
<td>PlayStation 2</td>
<td>Namco</td>
<td>Namco Bandai Games Inc.</td>
<td>Action; Simulation; Flight Sim</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Ace Combat X: Skies of Deception</td>
<td>PlayStation Portable</td>
<td>Namco Games</td>
<td>Namco Bandai Games Inc.</td>
<td>Action; Simulation; Flight Sim</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Ace Combat Zero: The Belkan War</td>
<td>PlayStation 2</td>
<td>Namco</td>
<td>Namco Bandai Games Inc.</td>
<td>Action; Simulation; Flight Sim</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Activision Anthology</td>
<td>PlayStation 2</td>
<td>Activision</td>
<td>Activation</td>
<td>Action; Strategy; Sports; Drink</td>
<td></td>
</tr>
<tr>
<td>Act of War: Direct Action</td>
<td>PC</td>
<td>QF Russia; Atari, Inc.</td>
<td>Eupen Systems</td>
<td>Strategy; Real-Time Strategy</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Advance Guardian Heros</td>
<td>Game Boy Advance</td>
<td>UBI Soft</td>
<td>Treasure</td>
<td>Action; Fighting</td>
<td></td>
</tr>
<tr>
<td>Advance Wars: Days of Ruin</td>
<td>Nintendo DS</td>
<td>Nintendo</td>
<td>Intelligent Systems Co., Ltd.</td>
<td>Strategy</td>
<td>Post-Apocalyptic</td>
</tr>
<tr>
<td>Advance Wars: Dual Strike</td>
<td>Nintendo DS</td>
<td>Nintendo</td>
<td>Intelligent Systems Co., Ltd.</td>
<td>Strategy</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Advent Rising</td>
<td>PC</td>
<td>Majesco Entertainment</td>
<td>GryphA Games, LLC</td>
<td>Action; Shooter: Adventure</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Advent Rising</td>
<td>XBox</td>
<td>Majesco Entertainment</td>
<td>GryphA Games, LLC</td>
<td>Action; Shooter: Adventure</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Angeles Wing</td>
<td>XBox 360</td>
<td>Microsoft Game Studios</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aeon Flux</td>
<td>XBox</td>
<td>Majesco Sales Inc.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Afro Samurai       | XBox 360 | Namco                       | Namco Bandai Games Inc.    | Platformer; Action            | Martial Arts; Sci-
| Age of Booty       | XBox 360 | Certain Affinity            |                            |                               | Fi                 |
| Age of Empires II | PC       | Microsoft Game Studios; Mat | Ensemble Studios          | Strategy; Real-Time Strategy  | Alternate History  |
| Age of Empires: The Age of Kings | Nintendo DS | Majesco Sales Inc. | Backbone Entertainment | Strategy                  | Fantasy            |

### Other Applications

- Job postings
- Seminar announcements
- Conference call for papers
- Company information
- Apartment rental adds
- Social event announcements
- ...
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; preferable 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 or rules that are applied to text.

- Sequence tagging models classify individual tokens as to whether or not they should be extracted.
LaSIE Information Extraction System

Tokenization - identify word boundaries in text
- white spaces indicate token boundaries
- full stops indicate sentences boundaries
  (not always true e.g. 1. September)

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, but the occurrence of “.” in
  “B. Clinton” or “U.S.” does not have this role
LaSIE Information Extraction System

**Gazetteers**
- Gazetteer lookup
- Sentence splitter
- Brill tagger
- Brill Morph parser
- Name matcher
- Discourse interpreter
- Template writer

**Lexicon**
- Conceptual hierarchy
- Stratified grammar

**Part-of-speech tagging** – identify and mark up the words in a text with the corresponding part of speech such as noun, verb, adjective, adverb etc.

According to -adv unofficial-adj source[s]-n, the-det bomb-n allegedly-adv detonate[ed]-v by-prep urban-adj guerrilla-n commando[s]-n blow_up-v a-det power_tower-n in-prep the-det northwestern-adj part-n of-prep San Salvador-loc at-prep 0650-time

LaSIE Information Extraction System

**Gazetteers**
- Gazetteer lookup
- Sentence splitter
- Brill tagger
- Brill Morph parser
- Name matcher
- Discourse interpreter
- Template writer

**Lexicon**
- Stratified grammar
- Conceptual hierarchy

**Syntactico-semantic interpretation**
- Bottom-up chart parser
- Cascade of NERC grammars (eg. aircraft, person, money, time)

According to -adv unofficial-adj source[s]-n, the-det bomb-n allegedly-adv detonate[ed]-v by-prep urban-adj guerrilla-n commando[s]-n blow_up-v a-det power_tower-n in-prep the-det northwestern-adj part-n of-prep San Salvador-loc at-prep 0650-time

San Salvador-loc NE1

San Salvador-loc NE2
LaSIE Information Extraction System

```
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

Syntactico-semantic interpretation
- bottom-up chart parser
- cascade of NERC grammars (e.g., aircraft, person, money, time)
- cascade of partial grammars (NPs, PPs, complex NP, VPs, complex VPs, RelClauses, Sentence)

S(According_to-adv NP(unofficial-adj source[s]-n) , NP(the-det bomb-n) allegedly-adv VP(detonate[ed]-v) PP(by-prep NP(urban-adj guerrilla-n commando[s]-n)) - VP(blow_up-ν-v) NP(a-det power_tower-n) PP(in-prep NP (the-det NE1-loc)) PP(at-prep NP(NE2-time)))
```

LaSIE Information Extraction System

```
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

Syntactico-semantic interpretation
- bottom-up chart parser
- cascade of NERC grammars (e.g., aircraft, person, money, time)
- cascade of partial grammars (NPs, PPs, complex NP, VPs, complex VPs, RelClauses, Sentence)
- logic form

Event(E1), detonate(E1,Y,X), urban_guerrilla_comando(X), bomb(Y),
Event(E2), blow_up(E2,Y,Z), power_tower(Z), location_of(Z,NE1), time_of(E2,NE2)
```
LaSIE Information Extraction System

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

Event(E1), detonate(E1,Y,X), urban_guerrilla_comando(X), bomb(Y),
Event(E2), blow_up(E2,Y,Z), power_tower(Z), location_of(Z,NE1), time_of(E2,NE2)
LaSIE Information Extraction System

Output template generation
- procedure that writes the templates in the desired format

<table>
<thead>
<tr>
<th>Incident type</th>
<th>bombing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>March 11, 2010</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>

How well does this work\(^1\)?

- Evaluate system’s performance on independent manually-annotated test gold data which was not used during system development

- IE systems are typically evaluated in terms of Precision (P) and Recall (R)

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

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

\[
F_1 = \frac{2PR}{P + R}
\]

\(^1\)http://www-nlpir.nist.gov/related_projects/muc/proceedings/st_score_report.html
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 interpretation</td>
<td>.93</td>
<td>.89</td>
</tr>
</tbody>
</table>
ANNIE Demo
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 JAFÉ 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. The 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.freeep.com/article/20100310/BUSINESS1/10031067/]

Person
Location
Organization
Date
Address
Money
Percent

California Prius incident probed; GM offers criticized
WASHINGTON -- As Toyota sought to contain the fallout from a California safety probe over sudden, unneeded acceleration, its dealers accused General Motors of offering predatory incentives using federal money.

WASHINGTON -- As Toyota sought to contain the fallout from a California safety probe over sudden, unneeded acceleration, its dealers accused General Motors of offering predatory incentives using federal money.

The Japanese automaker and the National Highway Traffic Safety Administration disagreed on the investigation in San Diego to analyse the recent Toyota Prius involving in Farmer Sims, Jr. More called the California Highway Patrol on Monday evening reporting that his Prius was accelerating of its own accord, hitting speeds of up to 90 mph.

“I pushed the gas pedal to pass a car and it did something kind of funny,” Sims said at a news conference. “It jumped and it just stuck there.”

An officer pulled alongside the Prius, and over a loudspeaker told Sims to pull the emergency brake and press the regular brake hard. Sims was able to slow the car and stop it in about 10 seconds.

The incident happened a few miles from the site of the recent crash last month that spurred Toyota to recall 5.8 million vehicles for mechanical defects that could lead to sudden acceleration -- including 31,000 Prius, which was involved in the rear-end crash.

But Sims said he had taken his Prius to his Toyota dealer and was told it wasn’t covered. Toyota said it was a statement that the invoice for the Prius had not yet been sent to dealers. The automaker had told owners to remove the driver’s-side floor mat until the repairs could be made, but Sims had left the floor mat in his vehicle.

Toyota recalled 31,000 Prius for the Prius was under recall. It had said last year that it could delay service for all of the mandated vehicles to be fixed.

In another sign of the pressure facing Toyota, its national dealer panel passed a resolution of using “appropriate dollars to fund … a nationwide advertising campaign.” Shortly after Toyota began its recalls last month, GM began running a campaign that includes non-personal financing and up to $2,500 cash back.

It is outrageous the GM is using our taxpayer dollars against us, making new and old Toyota dealers pay to undermine our own businesses,” said Paul Mitchell, president of Toyota’s U.S. dealer council.

GM’s move was met with furtherens. Last week, Toyota launched an incentive campaign of its own after a big drop in February sales.

“We understand the Toyota dealers would be frustrated, but that is not true,” said GM president Kelly. 
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.

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]

SUBJ: The World Trade Center (target)
VP: was bombed
PP: by terrorists

Syntactic Rules

Extraction Pattern: <target> was bombed

Examples of learned extraction patterns by AutoSlog:

- <subject> active-vp <perpetrator> bombed
- <subject> active-vp dobj <perpetrator> threw dynamite
- <subject> active-vp infinitive <perpetrator> tried to kill
- passive-vp infinitive <dobj> was hired to kill <victim>
- subject auxiliary <dobj> fatality was <victim>
- passive-vp prep <np> was killed by <perpetrator>