CS544: Information Extraction

January 31, 2012

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

• My research is in
  – Large-scale Web Knowledge Acquisition
  – Ontology Construction and Population
  – Lexical Semantics and Paraphrases
  – Graph Algorithms
• Stepping stones towards complex NLP systems
  – IBM Watson
  – Siri
Today

- Information Extraction (IE)
  - What it is
  - Historic roots: MUC
  - Current state-of-art performance
  - Techniques for solving IE
Need a better solution: Web Site aggregator

Extract Jobs from Multiple Sites
What is “Information Extraction”?

- **Goal:** identify specific pieces of information from the content of unstructured or semi-structured textual documents.

- **Input:**
  - scenario of extraction (templates to be filled)
  - document collection

- **Output:**
  - a set of instantiated templates
Applications

- Apartment rental adds
- USC alert system
- Social event announcements
- Seminar announcements
- Conference call for papers
- Company information
- ...

Google Squared

http://www.google.com/squared
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a “cancer” that stifled technological innovation.

Today, Microsoft claims to “love” the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

“We can be open source. We love the concept of shared source,” said Bill Veghte, a Microsoft VP. “That’s a super-important shift for us in terms of code access.”

Richard Stallman, founder of the Free Software Foundation, countered saying...
What is “Information Extraction”?

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

What is “Information Extraction”?

As a technique: IE = segmentation + classification + association

“named entity extraction”

<table>
<thead>
<tr>
<th>NAME</th>
<th>TITLE</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>founder</td>
<td>Free Soft.</td>
</tr>
</tbody>
</table>

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…
IE is Different in Different Domains

Newswire

Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK—July 17, 2002—Apple’s first retail store in New York City will open in Manhattan’s SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple’s largest retail store to date and is a stunning example of Apple’s commitment to offering customers the world’s best computer shopping experience.

“Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week,” said Steve Jobs, Apple’s CEO. “We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles.”

The Web has less grammar, but more formatting & linking

IE Depends on Complexity of Extractions

Closed set
U.S. states

He was born in Alabama…

The big Wyoming sky…

Complex pattern
U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Regular set
U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-FAT-1299

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.
IE Depends on the Type of Extractions

<table>
<thead>
<tr>
<th>Single entity</th>
<th>Binary relationship</th>
<th>N-ary record</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Person</strong>: Jack Welch</td>
<td><strong>Relation</strong>: Person-Title</td>
<td><strong>Relation</strong>: Succession</td>
</tr>
<tr>
<td><strong>Person</strong>: Jeffrey Immelt</td>
<td><strong>Person</strong>: Jack Welch</td>
<td><strong>Company</strong>: General Electric</td>
</tr>
<tr>
<td><strong>Title</strong>: CEO</td>
<td><strong>Title</strong>: CEO</td>
<td><strong>Title</strong>: CEO</td>
</tr>
<tr>
<td><strong>Location</strong>: Connecticut</td>
<td><strong>Company</strong>: General Electric</td>
<td><strong>Out</strong>: Jack Welsh</td>
</tr>
<tr>
<td></td>
<td><strong>Location</strong>: Connecticut</td>
<td><strong>In</strong>: Jeffrey Immelt</td>
</tr>
</tbody>
</table>

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

IE with Single or Multiple Documents

- **Single-document IE**
  - extract facts from a specific document
  - cares what is reported in this particular story
  - usually only one opportunity to find a piece of information
  - must deal with the diversity of language

- **Multiple-document IE**
  - extract facts from a collection of documents (i.e. Web)
  - cares about the facts and does not care where the information comes from
  - has more opportunities to find the information
Genesis of IE

- DARPA funded IE in early to mid 1990’s
- Message Understanding Conference (MUC) was an annual competition where systems were evaluated
- Focused on extracting information from news articles, which are of interest to the intelligence community (CIA, NSA)

Message Understanding Conference

- Domains
    - 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
Message Understanding Conference

• Types of information that must be extracted:
  
  – Named Entities
    • Person, Organization, Location names
  
  – Co-reference
    • Clinton <-> President Bill Clinton
  
  – Template element
    • Perpetrator, Target
  
  – Template relation
    • Incident

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

<table>
<thead>
<tr>
<th>TIE-UP-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationship:</td>
</tr>
<tr>
<td>Entities:</td>
</tr>
<tr>
<td>Joint Venture Company:</td>
</tr>
<tr>
<td>Activity: ACTIVITY-1</td>
</tr>
<tr>
<td>Amount:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTIVITY-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity:</td>
</tr>
<tr>
<td>Company:</td>
</tr>
<tr>
<td>Product:</td>
</tr>
<tr>
<td>Start Date:</td>
</tr>
</tbody>
</table>
Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: “Bridgestone Sport Co.”
“a local concern”
“a Japanese trading house”
Joint Venture Company: “Bridgestone Sports Taiwan Co.”
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: Company:
Product:
Start Date:

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: “Bridgestone Sport Co.”
“a local concern”
“a Japanese trading house”
Joint Venture Company: “Bridgestone Sports Taiwan Co.”
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
“Bridgestone Sports Taiwan Co.”
Product:
“iron and ‘metal wood’ clubs”
Start Date: January 1990
IE Techniques

Lexicons
Abraham Lincoln was born in Kentucky.

Classify Pre-segmented Candidates
Abraham Lincoln was born in Kentucky.

Boundary Models
Abraham Lincoln was born in Kentucky.

Sliding Window
Abraham Lincoln was born in Kentucky.

Finite State Machines
Most likely state sequence?

Context Free Grammars
Most likely parse?

Three generations of IE systems

- Statistical Models [1997 – ]
Hand-Built Systems [1980s-]

- Rules written by hand
- Require experts who understand both the systems and the domain
- Iterative guess-test-tweak-repeat cycle

**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

Automatic Rule Extractors [1990s-]

- Rules are automatically learned from huge labeled corpora
- Typically one
  - starts with highly specialized patterns
  - iteratively generalize new extraction rules
  - stops when the set of patterns has been generated to sufficiently “cover” the training examples
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.
AutoSlog [Riloff 1993]

**Examples of learned patterns with AutoSlog:**

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

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>active-vp dobj</td>
<td>bombed &lt;target&gt;</td>
</tr>
<tr>
<td>infinitive dobj</td>
<td>to kill &lt;victim&gt;</td>
</tr>
<tr>
<td>active-vp infinitive dobj</td>
<td>tried to kill &lt;victim&gt;</td>
</tr>
<tr>
<td>passive-vp infinitive dobj</td>
<td>was hired to kill &lt;victim&gt;</td>
</tr>
<tr>
<td>subject auxiliary dobj</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>
Statistical Models [1997-]

- Use machine learning to learn which features indicate boundaries and types of entities.
- Learning usually supervised; may be partially unsupervised
- Trade-off: annotating texts vs. manual knowledge engineering
  - weeks vs. months
  - domain experts vs. computational linguists

Next Time
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*)

LaSIE Information Extraction System

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 northwestern part of San Salvador at 0650.
LaSIE Information Extraction System

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 northwestern part of San Salvador at 0650.
LaSIE Information Extraction System

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.

Why do we care?

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

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.
LaSIE Information Extraction System

Syntactico-semantic interpretation
- bottom-up chart parser
- cascade of NERC grammars (e.g. 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 part of_ San Salvador-loc** at-prep **0650-time**

NE2

LaSIE Information Extraction System

Syntactico-semantic interpretation
- 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-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

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)

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.

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”
LaSIE Information Extraction System

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

| Incident type: | bombing |
| Date: | March 11, 2010 |
| 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 |

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)

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

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

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

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 interpretation</td>
<td>.93</td>
<td>.89</td>
</tr>
</tbody>
</table>
Can I test an existing IE system?

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 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. 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.freep.com/article/20100310/BUSINESS/11031677/]
- 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 sudden-acceleration case caught on camera, its dealers accused General Motors of offering predatory incentives using federal money.

GM’s move was met with skepticism by other automakers. Last week, Toyota launched an incentive campaign of its own after a $50 drop in February sales.

"We understand why Toyota dealers would be frustrated, but they have to live with it," said GM president Kent O交通

WASHINGTON — An Toyota seeking to contain the fallout from a California sudden-acceleration case caught on camera, its dealers accused General Motors of offering predatory incentives using federal money.

GM’s move was met with skepticism by other automakers. Last week, Toyota launched an incentive campaign of its own after a $50 drop in February sales.

"We understand why Toyota dealers would be frustrated, but they have to live with it," said GM president Kent O交通