Information Extraction and Named Entity Recognition

- Slides from Christopher Manning
Goal of Information Extraction

Extracting structured information out of unstructured text
Information Extraction

• Information extraction (IE) systems
  • Goal: produce a structured representation of relevant information:
    • relations (in the database sense), a.k.a.,
    • a knowledge base
  • Objectives:
    • Organize information so that it is useful to people
    • Put information in a semantically precise form that allows further inferences to be made by computer algorithms
Information Extraction (IE)

IE systems extract clear, factual information

- Roughly: *Who did what to whom when?*
  - E.g.,
    - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
      - headquarters(“BHP Billiton Limited”, “Melbourne, Australia”)
    - Learn drug-gene product interactions from medical research literature
    - Extract citations from a research article
Low-level information extraction

- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing

- Often seems to be based on regular expressions and name lists
Low-level information extraction

Google search for "bhp billiton headquarters"

- Best guess for BHP Billiton Ltd. Headquarters is Melbourne, London
- Mentioned on at least 9 websites including wikipedia.org, bhpbilliton.com and bhpbilliton.com - Feedback

- BHP Billiton - Wikipedia, the free encyclopedia
  en.wikipedia.org/wiki/BHP_Billiton
  Merger of BHP & Billiton 2001 (creation of a DLC). Headquarters, Melbourne, Australia (BHP Billiton Limited and BHP Billiton Group) London, United Kingdom ...
  History - Corporate affairs - Operations - Accidents
Why is IE hard on the web?
How is IE useful?
Classified Advertisements (Real Estate)

Background:
- Plain text advertisements
- Lowest common denominator: only thing that 70+ newspapers using many different publishing systems can all handle

<ADNUM>2067206v1</ADNUM>
<Date>March 02, 1998</Date>
<ADTITLE>MADDINGTON $89,000</ADTITLE>
<ADTEXT>
OPEN 1.00 - 1.45
U 11 / 10 BERTRAM ST
NEW TO MARKET Beautiful
3 brm freestanding
villa, close to shops & bus
Owner moved to Melbourne
ideally suit 1st home buyer, investor & 55 and over.
Brian Hazelden 0418 958 996
R WHITE LEEMING 9332 3477
</ADTEXT>
Why doesn’t text search (IR) work?

What you search for in real estate advertisements:
- Town/suburb. You might think easy, but:
  - Real estate agents: Coldwell Banker, Mosman
  - Phrases: Only 45 minutes from Parramatta
  - Multiple property ads have different suburbs in one ad
- Money: want a range not a textual match
  - Multiple amounts: was $155K, now $145K
  - Variations: offers in the high 700s [but not rents for $270]
- Bedrooms: similar issues: br, bdr, beds, B/R
Named Entity Recognition (NER)

• A very important sub-task: find and classify names in text, for example:

• The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.
Named Entity Recognition (NER)

• A very important sub-task: find and classify names in text, for example:

• The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.
A very important sub-task: find and classify names in text, for example:

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.
**Named Entity Recognition (NER)**

The uses:

- Named entities can be indexed, linked off, etc.
- Sentiment can be attributed to companies or products
- A lot of IE relations are associations between named entities
- For question answering, answers are often named entities.

Concretely:

- Many web pages tag various entities, with links to bio or topic pages, etc.
- Reuters’ OpenCalais, Evri, AlchemyAPI, Yahoo’s Term Extraction, ...
- Apple/Google/Microsoft/... smart recognizers for document content
Evaluation of Named Entity Recognition

Precision, Recall, and the F measure; their extension to sequences
The 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not</td>
<td>fn</td>
<td>tn</td>
</tr>
<tr>
<td>not selected</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Precision and recall

- **Precision**: % of selected items that are correct
- **Recall**: % of correct items that are selected

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
The Named Entity Recognition Task

Task: Predict entities in a text

- Foreign: ORG
- Ministry: ORG
- spokesman: O
- Shen: PER
- Guofang: PER
- told: O
- Reuters: ORG

Standard evaluation is per entity, *not* per token.
Precision/Recall/F1 for IE/NER

• Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)

• The measure behaves a bit funny for IE/NER when there are boundary errors (which are common):
  • First Bank of Chicago announced earnings ...
  • This counts as both a fp and a fn
  • Selecting nothing would have been better
  • Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules)
Three standard approaches to NER (and IE)

1. Hand-written regular expressions
   - Perhaps stacked

1. Using classifiers
   - Generative: Naïve Bayes
   - Discriminative: Maxent models

1. Sequence models
   - HMMs
   - CMMs/MEMMs
   - CRFs
Hand-written Patterns for Information Extraction

- If extracting from automatically generated web pages, simple regex patterns usually work.
  - Amazon page
    - `<div class="buying"><h1 class="parseasinTitle"><span id="btAsinTitle" style="">(.*?)</span></h1>`
- For certain restricted, common types of entities in unstructured text, simple regex patterns also usually work.
  - Finding (US) phone numbers
    - `(?:(?![0-9]{3}\{3}\?[ -.]?[0-9]{3}\{3}\?[ -.]?[0-9]{4})`
Natural Language Processing-based Information Extraction

- For unstructured human-written text, some NLP may help
  - Part-of-speech (POS) tagging
  - Mark each word as a noun, verb, preposition, etc.
  - Syntactic parsing
  - Identify phrases: NP, VP, PP
  - Semantic word categories (e.g. from WordNet)
  - KILL: kill, murder, assassinate, strangle, suffocate
Rule-based Extraction Examples

Determining which person holds what office in what organization

- \([\text{person}] \text{,}\ [\text{office}]\ of\ [\text{org}]\)
- Vuk Draskovic, leader of the Serbian Renewal Movement

- \([\text{org}]\ (\text{name}d,\ \text{appointed},\ \text{etc.})\ [\text{person}]\ \text{Prep}\ [\text{office}]\)
- NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located

- \([\text{org}]\ \text{in}\ [\text{loc}]\)
- NATO headquarters in Brussels

- \([\text{org}]\ [\text{loc}]\ (\text{division},\ \text{branch},\ \text{headquarters},\ \text{etc.})\)
- KFOR Kosovo headquarters
Information extraction as text classification
Naïve use of text classification for IE

• Use conventional classification algorithms to classify substrings of document as “to be extracted” or not.
• In some simple but compelling domains, this naive technique is remarkably effective.
  • But do think about when it would and wouldn’t work!
From: Robert Kubinsky <robert@lousycorp.com>
Subject: Email update

Hi all - I’m moving jobs and wanted to stay in touch with everyone so....
My new email address is: robert@cubemedia.com
Hope all is well :) 

>>R
Change-of-Address detection
[Kushmerick et al., ATEM 2001]

1. Classification

```
From: Robert Kubinsky <robert@lousycorp.com>
Subject: Email update

Hi all - I'm moving jobs and wanted to stay in touch
with everyone so...
My new email address is: robert@cubemedia.com
Hope all is well :) >
```

大家都好... 我的新电子邮件地址是: robert@cubemedia.com 希望一切安好 :) >

From: Robert Kubinsky <robert@lousycorp.com> Subject: Email update Hi all - I’m

2. Extraction

```
"message"
naïve Bayes model
```

"address"
naïve-Bayes model

\[ P[\text{robert@lousycorp.com}] = 0.28 \]
\[ P[\text{robert@cubemedia.com}] = 0.72 \]
Change-of-Address detection results
[Kushmerick et al., ATEM 2001]

Corpus of 36 CoA emails and 5720 non-CoA emails

- Results from 2-fold cross validations (train on half, test on other half)
- Very skewed distribution intended to be realistic
- Note very limited training data: only 18 training CoA messages per fold
- 36 CoA messages have 86 email addresses; old, new, and miscellaneous

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Message classification</td>
<td>98%</td>
<td>97%</td>
<td>98%</td>
</tr>
<tr>
<td>Address classification</td>
<td>98%</td>
<td>68%</td>
<td>80%</td>
</tr>
</tbody>
</table>
The ML sequence model approach to NER

Training
1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing
1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities
## Encoding classes for sequence labeling

<table>
<thead>
<tr>
<th>Name</th>
<th>IO encoding</th>
<th>IOB encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred</td>
<td>PER</td>
<td>B-PER</td>
</tr>
<tr>
<td>showed</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Sue</td>
<td>PER</td>
<td>B-PER</td>
</tr>
<tr>
<td>Mengqiu</td>
<td>PER</td>
<td>B-PER</td>
</tr>
<tr>
<td>Huang</td>
<td>PER</td>
<td>I-PER</td>
</tr>
<tr>
<td>‘s</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>new</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>painting</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
Features for sequence labeling

• Words
  • Current word (essentially like a learned dictionary)
  • Previous/next word (context)

• Other kinds of inferred linguistic classification
  • Part-of-speech tags

• Label context
  • Previous (and perhaps next) label
Features: Word substrings

oxa

Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion
Features: Word shapes

- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

<table>
<thead>
<tr>
<th>Term</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varicella-zoster</td>
<td>Xx-xxx</td>
</tr>
<tr>
<td>mRNA</td>
<td>XXXX</td>
</tr>
<tr>
<td>CPA1</td>
<td>XXXd</td>
</tr>
</tbody>
</table>
Relation Extraction
Binary Relation Association as Binary Classification

Christos Faloutsos conferred with Ted Senator, the KDD 2003 General Chair.

Person-Role (Christos Faloutsos, KDD 2003 General Chair) → NO

Person-Role (Ted Senator, KDD 2003 General Chair) → YES
John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tuesday, May 29, 1917, at 3:00 pm,[7] the second son of Joseph P. Kennedy, Sr., and Rose Fitzgerald; Rose, in turn, was the eldest of John "Honey Fitz" Fitzgerald, a prominent Boston political figure who was the city's mayor and a three-term member of Congress. Kennedy lived in Brookline for ten years and attended Edward Devotion School, Noble and Greenough Lower School, and the Dexter School, through 4th grade. In 1927, the family moved to 5040 Independence Avenue in Riverdale, Bronx, New York City; two years later, they moved to 294 Pondfield Road in Bronxville, New York, where Kennedy was a member of Scout Troop 2 (and was the first Boy Scout to become President).[8] Kennedy spent summers with his family at their home in Hyannisport, Massachusetts, and Christmas and Easter holidays with his family at their winter home in Palm Beach, Florida. For the 5th through 7th grade,
Rough Accuracy of Information Extraction

<table>
<thead>
<tr>
<th>Information type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>90-98%</td>
</tr>
<tr>
<td>Attributes</td>
<td>80%</td>
</tr>
<tr>
<td>Relations</td>
<td>60-70%</td>
</tr>
<tr>
<td>Events</td>
<td>50-60%</td>
</tr>
</tbody>
</table>

- **IE tasks are hard!**
- Errors cascade (error in entity tag → error in relation extraction)
- These are very rough, actually optimistic, numbers
  - Hold for well-established tasks, but lower for many specific/novel IE tasks