NLP for IR/web search?

- It's a no-brainer that NLP should be useful and used for web search (and IR in general):
  - Search for 'jaguar'
    - Are you interested in big cats (scarce on the web), cars, a high-performance computer cluster, or yet other things (e.g., perhaps a molecule geometry package?)
  - Search for 'Michael Jordan'
    - The basketballer or the machine learning guy?
  - Search for laptop, don't find notebook
  - [Google used to not even stem:
    - Searching probabilistic model didn’t even match pages with probabilistic models – but it does now, though with different weightings]

NLP for IR/web search?

- But in practice it’s hard to win with an “NLP Search Engine”, because a lot of the problems are elsewhere
  - E.g., syntactic phrases should (and may) help, but people have been able to get most of the mileage with "statistical phrases" – which have been aggressively integrated into systems recently (covert phrases; proximity weighting)
  - What has worked well is a bottom up incorporation of just a little knowledge of language
  - Knowing about bigrams which should be treated as a collocation/unit (think, language models)
  - Context-sensitive substitution of synonyms think, what a MT phrase-table might learn)
  - Named entity knowledge … more on this soon

NLP for IR/web search?

- Much more progress has been made in link analysis, use of anchor text, clickstreams, etc.
- Anchor text gives human-provided synonyms
- Using human intelligence always beats artificial intelligence
- People can easily scan among results (on their 24" monitor) … if you’re above the fold
- Link or click stream analysis gives a form of pragmatics: what do people find correct or important (in a default context)
- Focus on short, popular queries, news, etc.

NLP for IR/web search?

- Methods which use rich ontologies, etc., can work very well for intranet search within a customer’s site (where anchor-text, link, and click patterns are much less relevant)
- But don’t really scale to the whole web
  - Conclusion: one should move up the food chain to tasks where finer-grained understanding of meaning is needed
- One possibility: information extraction
Named Entity Recognition

- Named entity recognition
  - Labeling names of things in web pages:
    - An entity is a discrete thing like "IBM Corporation"
    - But often extended in practice to things like dates, instances of products and chemical/biological substances that aren’t really entities…
    - "Named" means called "IBM" or "Big Blue" not "it"
  - E.g.,
    - Many web pages tag various entities
    - "Smart Tags" (Microsoft) inside documents
    - Reuters’ OpenCalais

Information Extraction

- Information extraction systems
  - Find and understand the limited relevant parts of texts
    - Clear, factual information (who did what to whom when?)
    - Produce a structured representation of the relevant information: relations (in the DB sense)
    - Combine knowledge about language and a domain
    - Automatically extract the desired information
  - E.g.,
    - Gathering earnings, profits, board members, etc. from company reports
    - Learn drug-gene product interactions from medical research literature
    - quarterlyProfit(Citigroup, 2010Q1, \$4.4x10^9)
    - lives(Chris, Palo Alto)

Product information/ Comparison shopping, etc.

- Need to learn to extract info from online vendors
- Can exploit uniformity of layout, and (partial) knowledge of domain by querying with known products
- Early e.g., Jango Shopbot (Etzioni and Weld 1997)
  - Gives convenient aggregation of online content
  - Early bug: originally not popular with vendors
  - Make personal agents rather than web services?
  - This seems to have changed
    - Now you definitely need to be able to be discovered by search engines

Commercial information…

- A book, not a toy
- Title
- Need this price
Low-level information extraction

- Is now available – and I think popular – in applications like Apple or Google mail
- Seems to be based on regular expressions and name lists

Background:
- Advertisements are plain text
- Lowest common denominator: only thing that 70+ newspapers with 20+ publishing systems can all handle

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
- 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)
Inconsistency: digital cameras

- Image Capture Device: 1.68 million pixel 1/2-inch CCD sensor
- Image Capture Device Total Pixels Approx. 3.34 million
- Image sensor Total Pixels: Approx. 2.11 million
- Imaging sensor Total Pixels: Approx. 2.11 million
- Effective Pixels: Approx. 3.24 million
- Recording Pixels: Approx. 3.145 million
- These all came off the same manufacturer’s website!
- And this is a very technical domain. Try sofa beds.

Using information extraction to populate knowledge bases

http://protege.stanford.edu/

Named Entity Extraction

- The task: find and classify names in text, for example:
  - The European Commission [ORG] said on Thursday it disagreed with German [MISC] advice.
  - Only France [LOC] and Britain [LOC] backed Fischler [PER]’s proposal.
  - “What we have to be extremely careful of is how other countries are going to take Germany’s lead”, Welsh National Farmers Union (NFU) chairman John Lloyd Jones [PER] said on BBC radio.

- The purpose:
  - … a lot of information is really associations between named entities.
  - … for question answering, answers are usually named entities.
  - … the same techniques apply to other slot-filling classifications.

CoNLL (2003) Named Entity Recognition task

Task: Predict semantic label of each word in text

<table>
<thead>
<tr>
<th>Foreign</th>
<th>Ministry</th>
<th>spokesman</th>
<th>Shen</th>
<th>Guofang</th>
<th>told</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>NNP</td>
<td>NN</td>
<td>NNP</td>
<td>NNP</td>
<td>VBD</td>
<td>NNP</td>
</tr>
<tr>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
<td>I-VP</td>
<td>I-NP</td>
</tr>
<tr>
<td>ORG</td>
<td>ORG</td>
<td>O</td>
<td>PER</td>
<td>PER</td>
<td>O</td>
<td>ORG</td>
</tr>
</tbody>
</table>

A combined measure: F

- Combined measure that assesses this tradeoff is F measure (weighted harmonic mean):
  \[ F = \frac{P \cdot R}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \]

- People usually use balanced F measure
  - i.e., with \( \beta = 1 \) (that is, \( \alpha = \frac{1}{2} \)): \( F = 2PR/(P+R) \)
- Harmonic mean is conservative average
  - See CJ van Rijsbergen, Information Retrieval

Precision and recall

- Precision: fraction of retrieved items that are relevant = P(correct|selected)
- Recall: fraction of relevant docs that are retrieved = P(selected|correct)

<table>
<thead>
<tr>
<th>Correct</th>
<th>Not Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected</td>
<td>tp</td>
</tr>
<tr>
<td>Not Selected</td>
<td>fn</td>
</tr>
</tbody>
</table>

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)
Quiz question

- What is the $F_1$ measure for the following 2 cases:
  - Precision = 90%, Recall = 30%
  - Precision = 50%, Recall = 50%

$$F = \frac{1}{\frac{1}{P} + \frac{1}{R} - \frac{1}{(1-\alpha)R} + \frac{1}{\beta^2P + R}}$$

- People usually use balanced $F_1$ measure
  - i.e., with $\beta = 1$ (that is, $\alpha = \frac{1}{2}$): $F = 2PR/(P+R)$

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 systems (e.g., MUC scorer) give partial credit (according to complex rules)

NER

- Three standard approaches
  - Hand-written regular expressions
  - Perhaps stacked
  - Using classifiers
  - Generative: Naïve Bayes
  - Discriminative: Maxent models
  - Sequence models
    - HMMs
    - CMMs/MEMMs
    - CRFs

Hand-written 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, simple regex patterns usually work.
    - Finding (US) phone numbers
    - `(?:\(?[0-9]{3}\)?[ -.]?[0-9]{3}[ -.]?[0-9]{4}`

Natural Language Processing-based Hand-written Information Extraction

- If extracting from more natural, 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
  - Extraction patterns can use POS or phrase tags.
  - Crime victim:
    - Prefix: [POS: V, Hypernym: KILL]
    - Filler: [Phrase: NNP]

MUC: the NLP genesis of IE

- DARPA funded significant efforts in IE in the early to mid 1990’s.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- Focused on extracting information from news articles:
  - Terrorist events
  - Industrial joint ventures
  - Company management changes
  - Information extraction is of particular interest to the intelligence community …
  - Though also to all other “information professionals”
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.

Example of IE from FASTUS (1993)

<table>
<thead>
<tr>
<th>TIE-UP-1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relationship:</strong> TIE-UP</td>
</tr>
<tr>
<td><strong>Entities:</strong></td>
</tr>
<tr>
<td>“Bridgestone Sports Co.”</td>
</tr>
<tr>
<td>“a local concern”</td>
</tr>
<tr>
<td>“a Japanese trading house”</td>
</tr>
<tr>
<td><strong>Joint Venture Company:</strong></td>
</tr>
<tr>
<td>“Bridgestone Sports Taiwan Co.”</td>
</tr>
<tr>
<td><strong>Activity:</strong> Activity: PRODUCTION Company:</td>
</tr>
<tr>
<td>“Bridgestone Sports Taiwan Co.”</td>
</tr>
<tr>
<td>“iron and ‘metal wood’ clubs”</td>
</tr>
<tr>
<td><strong>Start Date:</strong> DURING: January 1990</td>
</tr>
<tr>
<td><strong>Amount:</strong> NT$200000000</td>
</tr>
</tbody>
</table>

Example of IE: 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.

Example of IE: FASTUS(1993)

<table>
<thead>
<tr>
<th>TIE-UP-1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relationship:</strong> TIE-UP</td>
</tr>
<tr>
<td><strong>Entities:</strong></td>
</tr>
<tr>
<td>“Bridgestone Sports Co.”</td>
</tr>
<tr>
<td>“a local concern”</td>
</tr>
<tr>
<td>“a Japanese trading house”</td>
</tr>
<tr>
<td><strong>Joint Venture Company:</strong></td>
</tr>
<tr>
<td>“Bridgestone Sports Taiwan Co.”</td>
</tr>
<tr>
<td><strong>Activity:</strong> Activity: PRODUCTION Company:</td>
</tr>
<tr>
<td>“Bridgestone Sports Taiwan Co.”</td>
</tr>
<tr>
<td>“iron and ‘metal wood’ clubs”</td>
</tr>
<tr>
<td><strong>Start Date:</strong> DURING: January 1990</td>
</tr>
<tr>
<td><strong>Amount:</strong> NT$200000000</td>
</tr>
</tbody>
</table>

**FASTUS**

Based on finite state automata (FSA) transductions

1. Complex Words:
   - Recognition of multi-words and proper names
2. Basic Phrases:
   - Simple noun groups, verb groups and particles
3. Complex phrases:
   - Complex noun groups and verb groups
4. Domain Events:
   - Patterns for events of interest to the application
5. Merging Structures:
   - Templates from different parts of the texts are merged if they provide information about the same entity or event.

**Grep++ = Cascaded grepping**

Finite Automaton for Noun groups:
John's interesting book with a nice cover

**Rule-based Extraction Examples**

Determining which person holds what office in what organization
- [person], [office] of [org]
  - Vuk Draskovic, leader of the Serbian Renewal Movement
  - [org] (named, appointed, etc.) [person] P (office)
  - NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located
- [org] in [loc]
  - NATO headquarters in Brussels
  - [org] [loc] (division, branch, headquarters, etc.)
  - KFOR Kosovo headquarters

**Simple classification-based IE: Naive Bayes Classifiers**

Task: Classify a new instance based on a tuple of attribute values

\[
c_{MJP} = \arg\max_{c_i} P(c_i | x_1, x_2, \ldots, x_n)
\]

\[
c_{MJP} = \arg\max_{c_i} P(x_1, x_2, \ldots, x_n | c_i)P(c_i)
\]

\[
c_{MJP} = \arg\max_{c_i} P(x_1, x_2, \ldots, x_n | c_i)P(c_i)
\]
Naïve Bayes Classifier: Assumptions

- \( P(c_j) \)
  - Can be estimated from the frequency of classes in the training examples.
- \( P(x_1, x_2, \ldots, x_n | c_j) \)
  - Can only be estimated if a very, very large number of training examples was available.

Conditional Independence Assumption:

- Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities.

\[
P(X_1, \ldots, X_n | C) = P(X_1 | C) \cdot P(X_2 | C) \cdot \ldots \cdot P(X_n | C)
\]

Naïve integration of IE & text classification

- 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 think about when it would and wouldn’t work!

Kushmerick et al. 2001 ATEM: Change of Address Results

<table>
<thead>
<tr>
<th></th>
<th>Words</th>
<th>Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( P )</td>
<td>( R )</td>
</tr>
<tr>
<td>Message classification</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Address classification</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td>90%</td>
<td></td>
</tr>
</tbody>
</table>

36 CoA messages
86 addresses
55 old, 31 new
5720 non-CoA

"Change of Address" email

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

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