Information Extraction & Named Entity Recognition

Christopher Manning
CS224N
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’
    • the computer should know or ask whether you’re interested in big cats [scarce on the web], cars, or, perhaps a molecule geometry and solvation energy package, or a package for fast network I/O in Java
  • 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.]
Word sense disambiguation technology generally works well (like text categorization)

- Synonyms can be found or listed
- Lots of people were into fixing this
  - Especially around 1999–2000
  - Lots of (ex-)startups:
    - LingoMotors
    - iPhrase “Traditional keyword search technology is hopelessly outdated”
NLP for IR/web search?

• But in practice it’s an idea that hasn’t gotten much traction
  • Correctly finding linguistic base forms is straightforward, but produces little advantage over crude stemming which just slightly over equivalence classes words
  • Word sense disambiguation only helps on average in IR if over 90% accurate (Sanderson 1994), and that’s about/above where we are
  • Syntactic phrases should 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)
NLP for IR/web search?

• Much more progress has been made in link analysis, the use of anchor text, 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

- Moral: it’s hard to beat keyword search for the task of general ad hoc document retrieval
- Conclusion: one should move up the food chain to tasks where finer-grained understanding of meaning is needed

- One possibility: information extraction
Information Food Chain II

- Agents
- Personal Assistants
- Indices, Directories
- Mass Services
- World Wide Web

Exoni, Softbots 4
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)
    - Gives convenient aggregation of online content
  - Bug: originally not popular with vendors
    - Make personal agents rather than web services?
- This seems to have changed (e.g., Froogle)
Gold Planet Micro Urban Rescue HW Set Hot Wheels
$11.78 - Yahoo! Auctions - Toy Cars & Vehicles
URBAN RESCUE, FIRE ENGINE, HOT WHEELS, HW PLANET MICRO
LIMITED EDITION GOLD ... RELINQUISHED EXODIA, PS2 PLAYSTATION
2, HEROCLIX, X-BOX, LEGO, MIGHTY MOOSE ...

LEGO Community Transport Set - 50 Pieces - SmarterKids
$51.99 - Shop.com - Blocks & Construction
... meaning of community helpers. Set includes 9 DUPLO
figures, a helicopter, fire engine, plane and
more. Developmental Area(s): Cognitive ...

LEGO Vehicles Set
$79.21 - homeschoolingsupply.com - Blocks & Construction
Special community vehicles include a crane, a police car,
a fire engine and 2 construction vehicles, and the set
includes 4 play mats with bases and special...

Peercn: Fire engine (#374-2)
$125.00 - www.peercn.com
Inventory for set #374-2: Fire engine Theme: LEGO LEGOLAND
/ Large Vehicle Year: 1971 Pcs: Figs: 0 MSR#: ? ...

Radio Flyer Red Fire Engine
$24.99 - www.raggedolls.com
Radio Flyer Red Fire Engine. Radio Flyer #309
Little Red Fire Engine SHIPPING INCLUDED! Radio
Flyer #309 ... Radio Flyer Red Fire Engine.
Commercial information...

A book, Not a toy

Title

Need this price

Luckys Collectors Guide To 20th Century Yo-Yos: History And Values
Author: Meisenheimer, Lucky J.; Editor: T Brown & Associates
Paperback
Published: October 1999
Lucky J’s Swim & Surf
ISBN: 0966761200

PRODUCT CODE: 0966761200
- USA/Canada: US$ 43.40
- Australia/NZ: A$ 124.50
- Other Countries: US$ 80.90

Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it
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
  - “Smart Tags” (Microsoft) inside documents
Classified Advertisements (Real Estate)

Background:
- Advertisements are plain text
- Lowest common denominator: only thing that 70+ newspapers with 20+ 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>
Use Navigation Aids to change chosen area

UBD Reference: "332 D10"

Property Details

Address: 10 BERTRAM ST
Suburb: MADDINGTON
State: WA
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)
Canonicalization:
Product information
# Canonicalization: Product Information

<table>
<thead>
<tr>
<th>Product</th>
<th>Features</th>
<th>Rating</th>
<th>Price</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IBM ThinkPad X31</strong></td>
<td>Provides a depth of features in a small package suited for serious travelers.</td>
<td>7.7 Good</td>
<td>$2004-$2235</td>
<td>Check price</td>
</tr>
<tr>
<td><strong>IBM ThinkPad X31 2672 - Pentium M 1.4 GHz - 12.1” TFT</strong></td>
<td>Specifications: 3.5 lbs, 1.4 GHz, Intel Pentium M, 256 MB DDR SDRAM, IDE, 140 GB Portable Internal, 12.1 in TFT active matrix, 1 Lithium Ion, Microsoft Windows XP Professional (Preinstalled).</td>
<td>4.9 Average</td>
<td>$2023-$2299</td>
<td>Check price</td>
</tr>
<tr>
<td><strong>IBM ThinkPad X31 2672 - Pentium M 1.3 GHz - 12.1” TFT</strong></td>
<td>Specifications: 3.5 lbs, 1.3 GHz, Intel Pentium M, 128 MB DDR SDRAM, Portable, 120 GB IDE Internal, TFT active matrix, 1.1 Lithium Ion, Microsoft Windows XP Professional (Preinstalled).</td>
<td>6.1 Average</td>
<td>$1808-$2054</td>
<td>Check price</td>
</tr>
<tr>
<td><strong>IBM ThinkPad X31 2672 - Pentium M 1.4 GHz - 12.1” TFT</strong></td>
<td>Specifications: 3.5 lbs, 1.4 GHz, Intel Pentium M, DDR SDRAM, 256 MB, Portable, 140 GB IDE Internal, TFT active matrix, 1.1 Lithium Ion.</td>
<td>4.9 Average</td>
<td>$1809-$2154</td>
<td>Check price</td>
</tr>
</tbody>
</table>

※ Email me when this product is available

[Image of a webpage showing IBM ThinkPad X31 products with ratings and prices]
Inconsistency: digital cameras

- Image Capture Device: 1.68 million pixel 1/2-inch CCD sensor
- Image Capture Device    Total Pixels Approx. 3.34 million  
  Effective Pixels Approx. 3.24 million
- Image sensor    Total Pixels: Approx. 2.11 million-pixel
- Imaging sensor  Total Pixels: Approx. 2.11 million  1,688 (H) x 1,248 (V)
- CCD     Total Pixels: Approx. 3,340,000 (2,140[H] x 1,560 [V] )
  - Effective Pixels: Approx. 3,240,000 (2,088 [H] x 1,550 [V] )
  - Recording Pixels: Approx. 3,145,000 (2,048 [H] x 1,536 [V] )
- *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/
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 [ORG] ( NFU [ORG] ) chairman John Lloyd Jones [PER] said on BBC [ORG] 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>NNP</th>
<th>I-NP</th>
<th>ORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ministry</td>
<td>NNP</td>
<td>I-NP</td>
<td>ORG</td>
</tr>
<tr>
<td>spokesman</td>
<td>NN</td>
<td>I-NP</td>
<td>O</td>
</tr>
<tr>
<td>Shen</td>
<td>NNP</td>
<td>I-NP</td>
<td>PER</td>
</tr>
<tr>
<td>Guofang</td>
<td>NNP</td>
<td>I-NP</td>
<td>PER</td>
</tr>
<tr>
<td>told</td>
<td>VBD</td>
<td>I-VP</td>
<td>O</td>
</tr>
<tr>
<td>Reuters</td>
<td>NNP</td>
<td>I-NP</td>
<td>ORG</td>
</tr>
</tbody>
</table>

Standard evaluation is per entity, not per token.
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></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>

- Precision $P = \frac{tp}{tp + fp}$
- Recall $R = \frac{tp}{tp + fn}$
A combined measure: F

- Combined measure that assesses this tradeoff is F measure (weighted harmonic mean):

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

- People usually use balanced F₁ measure
  - i.e., with \( \beta = 1 \) or \( \alpha = \frac{1}{2} \): \( F = 2PR/(P+R) \)
- Harmonic mean is conservative average
  - See CJ van Rijsbergen, *Information Retrieval*
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)
Natural Language Processing-based Hand-written Information Extraction

- If extracting from automatically generated web pages, simple regex patterns usually work.
- 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:
    - Prefiller: [POS: V, Hyponym: KILL]
    - Filler: [Phrase: NP]
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)

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:
  DURING: January 1990
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>
<th>ACTIVITY-1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relationship:</strong> TIE-UP</td>
<td><strong>Activity:</strong> PRODUCTION</td>
</tr>
<tr>
<td><strong>Entities:</strong> “Bridgestone Sport Co.”</td>
<td><strong>Company:</strong> “Bridgestone Sports Taiwan Co.”</td>
</tr>
<tr>
<td>“a local concern”</td>
<td>“iron and ‘metal wood’ clubs”</td>
</tr>
<tr>
<td>“a Japanese trading house”</td>
<td><strong>Start Date:</strong> DURING: January 1990</td>
</tr>
<tr>
<td><strong>Joint Venture Company:</strong></td>
<td></td>
</tr>
<tr>
<td>“Bridgestone Sports Taiwan Co.”</td>
<td><strong>Amount:</strong> NT$2000000000</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
   Basic templates are to be built.

5. Merging Structures:
   Templates from different parts of the texts are merged if they provide information about the same entity or event.

set up
new Taiwan dollars

a Japanese trading house
had set up

production of
20,000 iron and metal wood clubs

[company]
[set up]
[Joint-Venture]
with
[company]
Grep++ = Cascaded grepping

Finite Automaton for Noun groups:
John’s interesting book with a nice cover
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
Naive Bayes Classifiers

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

\[ \langle x_1, x_2, \ldots, x_n \rangle \]

\[
c_{MAP} = \underset{c_j \in C}{\text{argmax}} P(c_j \mid x_1, x_2, \ldots, x_n)
\]

\[
c_{MAP} = \underset{c_j \in C}{\text{argmax}} \frac{P(x_1, x_2, \ldots, x_n \mid c_j)P(c_j)}{P(c_1, c_2, \ldots, c_n)}
\]

\[
c_{MAP} = \underset{c_j \in C}{\text{argmax}} P(x_1, x_2, \ldots, x_n \mid c_j)P(c_j)
\]
Naïve Bayes Classifier:

Assumptions

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.

- $P(x_1, x_2, \ldots, x_n | c_j)$
  - $O(|X|^n \cdot |C|)$
  - Could 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.

$$P(X_1, \ldots, X_5 \mid C) = P(X_1 \mid C) \cdot P(X_2 \mid C) \cdot \cdots \cdot P(X_5 \mid C)$$
Naïve Bayes in NLP

- For us, the \( x_i \) are usually bags of occurring words
  - A class-conditional unigram language model!
    - *Different* from having a variable for each word type
- As usual, we need to smooth \( P(x_i|c_j) \)
  - Zero probabilities cannot be conditioned away, no matter what other evidence there is
    \[
    \hat{P}(x_{i,k} | c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}
    \]
- As before, multiplying lots of small numbers can cause floating-point underflow.
  - As \( \log(xy) = \log(x) + \log(y) \) and log is monotonic, it is faster and better to work by summing logs probabilities
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
**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 :)  

**2. Extraction**

From: Robert Kubinsky <robert@lousycorp.com> Subject: Email update  
Hi all - I’m  
everyone so.... My new email address is: robert@cubemedia.com Hope all is well :) >

- P[robert@lousycorp.com] = 0.28  
- P[robert@cubemedia.com] = 0.72
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>.96</td>
<td>.66</td>
</tr>
<tr>
<td>Address classification</td>
<td>.96</td>
<td>.62</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td>.96%</td>
<td></td>
</tr>
</tbody>
</table>

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