Information Extraction

Christopher Manning
CS224N – 2003
http://nlp.stanford.edu/~manning/

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 doesn't even *stem:*
    - Search for *probabilistic model*, and you don't even match pages with *probabilistic models.*

NLP for IR/web search?

- Word sense disambiguation technology generally works well (like text categorization)
- Synonyms can be found or listed
- Lots of people have been into fixing this
  - e-Cyc had a beta version with Hotbot that disambiguated senses, and was going to go live in 2 months ... 26 months ago
  - Lots of (ex-)startups:
    - LingoMotors
    - iPhrase "Traditional keyword search technology is hopelessly outdated"

NLP for IR/web search?

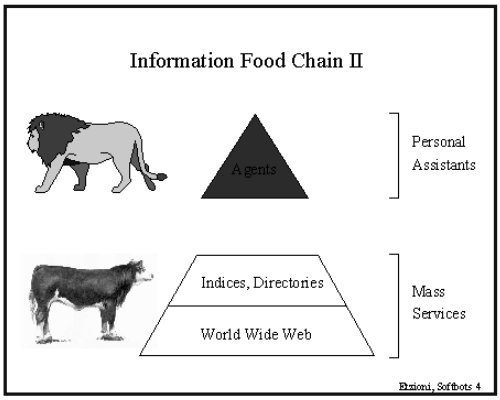
- But in practice it's an idea that hasn't gotten much traction
  - Correctly finding linguistic base forms is straightforward, but produces little advantage over crude stemming which just slightly over equivalence classes words
  - Word sense disambiguation only helps on average in IR if over 90% accurate (Sanderson 1994), and that's about 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

NLP for IR/web search?

- People can easily scan among results (on their 21" monitor) ... if you're above the fold
- Much more progress has been made in link analysis, and use of anchor text, etc.
- Anchor text gives human-provided synonyms
- 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.
- Using human intelligence always beats artificial intelligence

NLP for IR/web search?

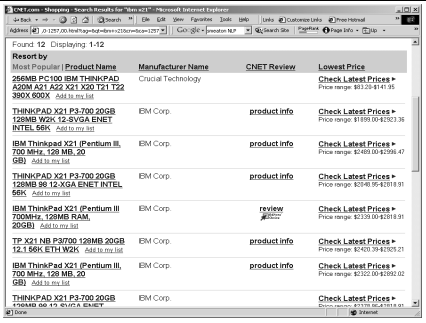
- Methods which use of rich ontologies, etc., can work very well for intranet search within a customer's site (where anchor-text, link, and click patterns are much less relevant)
  - But don't really scale to the whole web
- *Moral: it's hard to beat keyword search for the task of general ad hoc document retrieval*
- *Conclusion: one should move up the food chain to tasks where finer grained understanding of meaning is needed*



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
- E.g., Jango Shopbot (Etzioni and Weld)
  - Gives convenient aggregation of online content
- Bug: not popular with vendors
  - A partial solution is for these tools to be personal agents rather than web services

Product information



Product info

- C-net markets this information
- How do they get most of it?
  - Phone calls
  - Typing.

Inconsistency: digital cameras

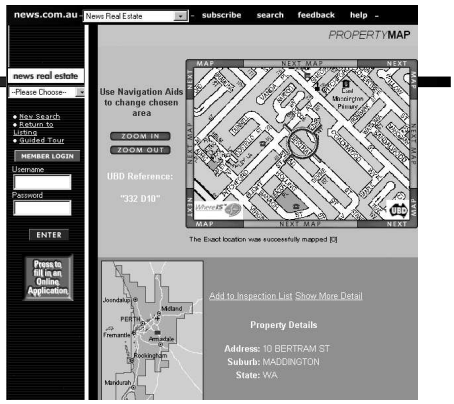
- Image Capture Device: 1.68 million pixel 1/2-inch CCD sensor
- Image Capture Device Total Pixels Approx. 3.34 million Effective Pixels Approx. 3.24 million
- Image sensor Total Pixels: Approx. 2.11 million-pixel
- Imaging sensor Total Pixels: Approx. 2.11 million 1,688 (H) x 1,248 (V)
- CCD Total Pixels: 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.

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<BR>
U 11 / 10 BERTRAM ST<BR>
NEW TO MARKET Beautiful<BR>
3 brm freestanding<BR>
villa, close to shops & bus<BR>
Owner moved to Melbourne<BR>
ideally suit 1st home
   buyer,<BR>
investor & 55 and over.<BR>
Brian Hazelden 0418 958 996<BR>
R WHITE LEEMING 9332 3477
</ADTEXT>
```



Why doesn't text search (IR) work?

What you search for in real estate advertisements:

- Suburbs. 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)

Task: Information Extraction

Suppositions:

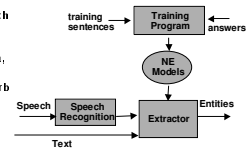
- A lot of information that *could* be represented in a structured semantically clear format isn't
- It may be costly, not desired, or not in one's control (screen scraping) to change this.
- Goal: being able to answer semantic queries (a.k.a. "database queries") using "unstructured" natural language sources
- Caveats: need clear, factual information; answers in small text snippets, some errors can be tolerated

HMMs for IE

- There are other techniques for information extraction (we'll discuss them more next time)
- But Hidden Markov Models are a powerful method for sequence-based information extraction
- Pros:
  - Well-understood underlying statistical model makes it easy to used wide range of tools from statistical decision theory
  - Portable, broad coverage, robust, good recall
- Cons:
  - Range of features and patterns usable may be limited
  - Not necessarily as good for complex multi-slot patterns

Name Extraction via HMMs

The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.



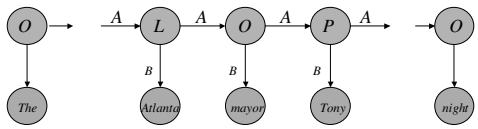
The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.

- Locations
- Persons
- Organizations

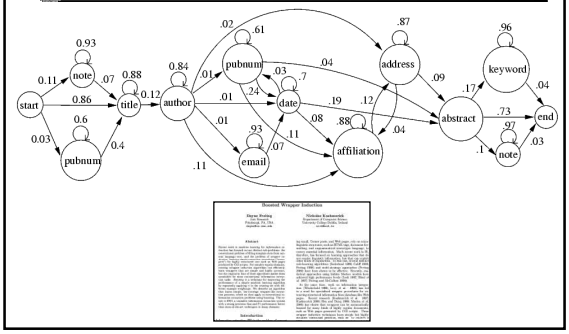
An easy but successful application:
- Prior to 1997 - no learning approach competitive with hand-built rule systems
- Since 1997 - Statistical approaches (BBN, NYU, MITRE, CMU/JustSystems) achieve state-of-the-art performance

Applying HMMs to IE

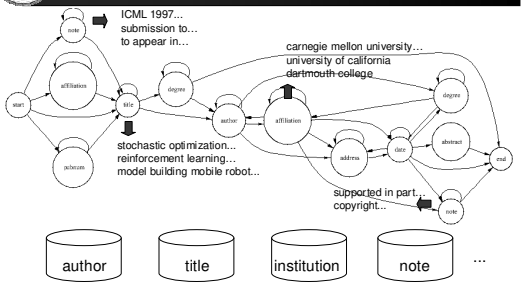
- **Document** ⇒ generated by a stochastic process
- **Observation** ⇒ word
- **State** ⇒ "reason/explanation" for a given token
  - '*Background*' state emits tokens like '*the*', '*said*', …
  - '*Money*' state emits tokens like '*million*', '*euro*', …
  - '*Organization*' state emits tokens like '*university*', '*company*', …
- **Extraction**: via the Viterbi algorithm



HMM for research papers: transitions [Seymore *et al.*, 99]



HMM for research papers: emissions [Seymore *et al.*, 99]



Trained on 2 million words of BibTeX data from the Web