Information Extraction

Christopher Manning

CS224N – 2005

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 used to not even *stem:*
    - Searching *probabilistic model* didn't even match pages with *probabilistic models* – but it does now.]

NLP for IR/web search?

- 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    - iPhrase "Traditional keyword search technology is hopelessly outdated"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?

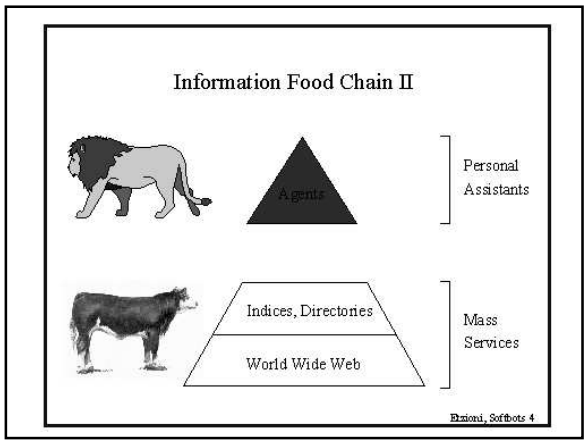
- But in practice it's an idea that hasn't gotten much traction
  - Correctly finding linguistic base forms is straightforward, but produces little advantage over crude stemming which just slightly over equivalence classes words
  - Word sense disambiguation only helps on average in IR if over 90% accurate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

NLP for IR/web search?

- Much more progress has been made in link analysis, and 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 21" 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.

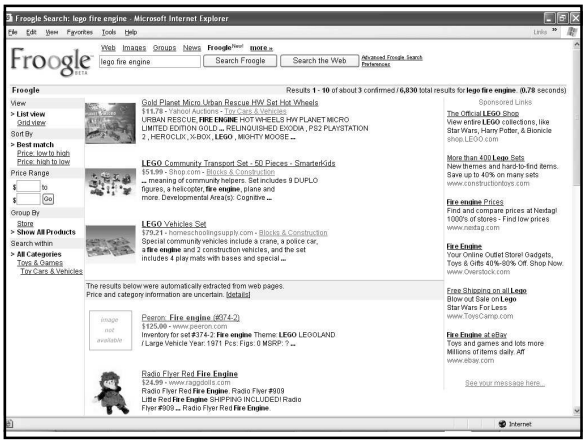
NLP for IR/web search?

- Methods which use 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*
- One possibility: information extraction

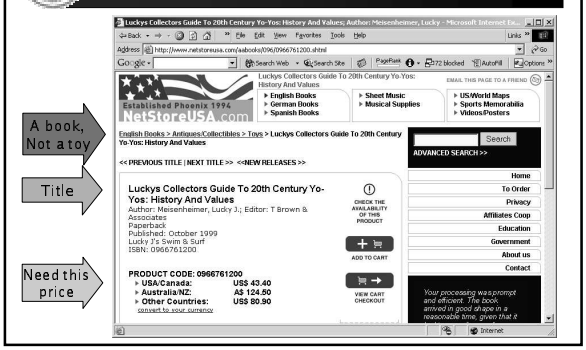


Product information/ Comparison shopping, etc.

- 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)



Commercial information…



Information Extraction

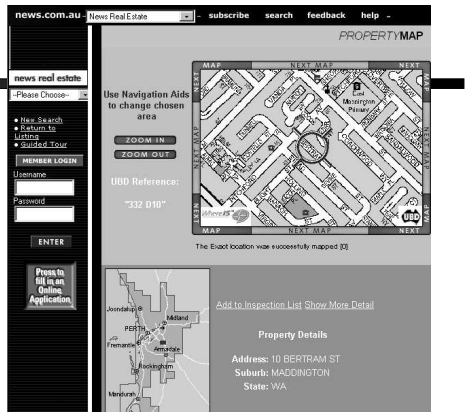
- Information extraction systems
  - Find and understand the limited relevant parts of texts
    - Clear, factual information (*who did what to whom when?*)
  - Produce a structured representation of the relevant information: *relations* (in the DB sense)
  - Combine knowledge about language and a domain
  - 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<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>
```

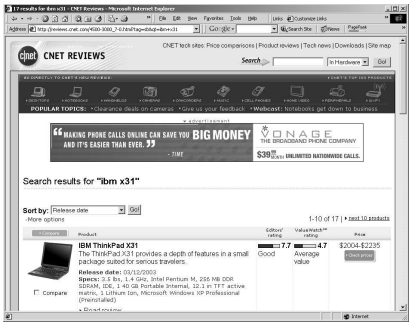


Why doesn't text search (IR) work?

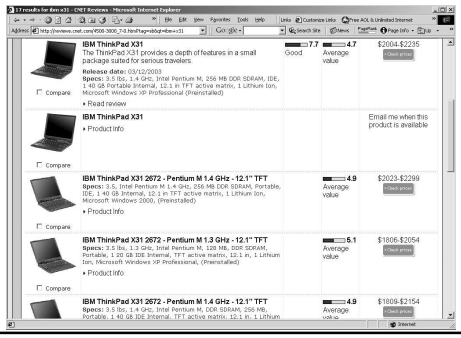
What you search for in real estate advertisements:

- 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



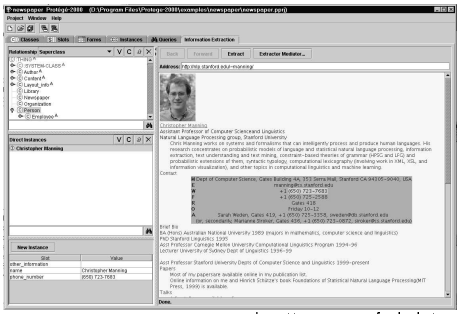
Canonicalization: Product information



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/

Statistical sequence models for Information Extraction

- There are other techniques for information extraction (template/wrapper learning, hand-coded rules)
- But statistical sequence models (Hidden Markov Models, MaxEnt markov models) are good methods 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:
  - Not necessarily as good for complex multi-slot patterns

Named Entity Extraction

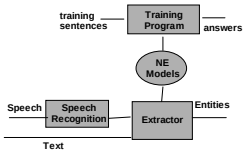
- The task: find and classify names in text, for example:

```
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 [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.

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

CoNLL (2003) Named Entity Recognition task

Task: Predict semantic label of each word in text

Foreign	NNP	I-NP	ORG
Ministry	NNP	I-NP	ORG
spokesman	NN	I-NP	O
Shen	NNP	I-NP	PER
Guofang	NNP	I-NP	PER
told	VBD	I-VP	O
Reuters	NNP	I-NP	ORG
:	:	:	:

} 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)

	Correct	Not Correct
Selected	tp	fp
Not Selected	fn	tn

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

Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget….

- People doing information retrieval want to find *something* and have a certain tolerance for junk

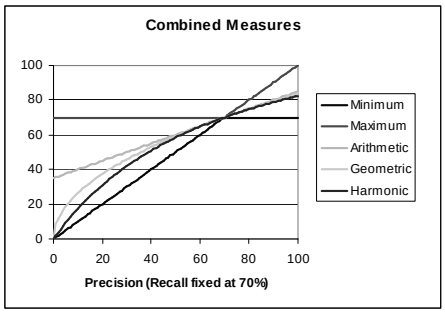
A combined measure: F

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

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

- People usually use balanced F_1 measure
  - i.e., with $\beta = 1$ or $\alpha = ½$: $F = 2PR/(P+R)$
- Harmonic mean is conservative average
  - See CJ van Rijsbergen, *Information Retrieval*

F_1 and other averages

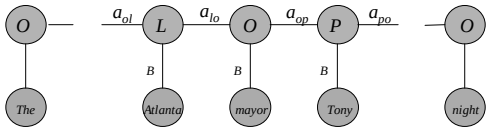


Precision/Recall/F1 for IE

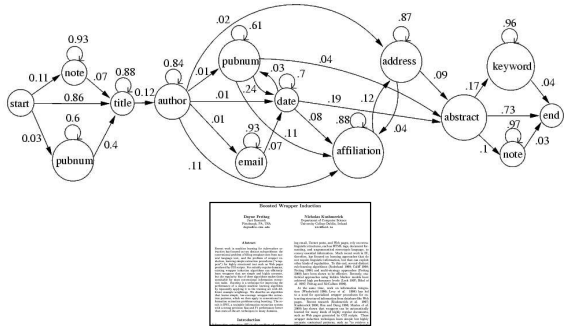
- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funnily for IE/NER when there are *boundary errors* (which are *common*):
  - First Bank of Chicago announced earnings …
- This counts as both 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)

Applying HMMs to IE
(Leek 1997, Freitag and McCallum 2000)

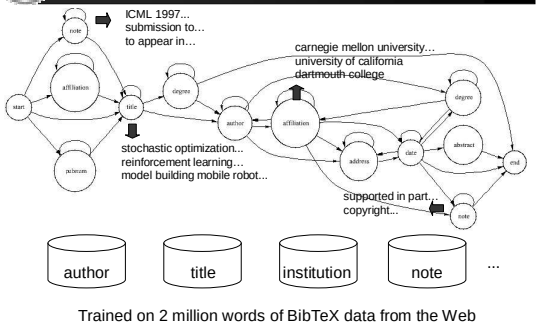
- **Document** ⇒ generated by a stochastic process
- **Observation** ⇒ word
- **State** ⇒ "reason/explanation" for a given token
  - '*Background*' state emits tokens like '*the*', '*said*', …
  - '*Money*' state emits tokens like '*million*', '*euro*', …
  - '*Organization*' state emits tokens like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

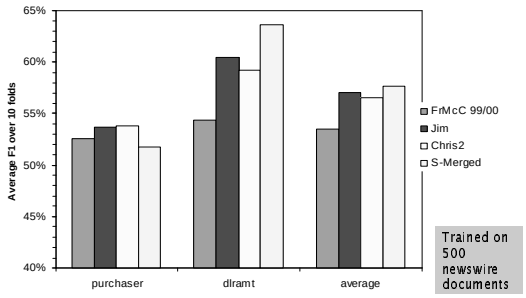


Trained on 2 million words of BibTeX data from the Web

Freitag and McCallum (2000) IE with HMMs details

- Partly fixed structure, partly hidden (constrained EM using remote supervision)
  - Class HMM (also used in comp. bio.)
- Parameter tying and shrinkage smoothing techniques
  - Better just to use a good unknown model?
- Structure learning of transition structure
  - Why not just plain EM?
- Results good on semi-structured data
  - Still rather modest on free form text
    - Need richer model class?

HMM IE results (F_1) on Freitag and McCallum Acquisitions data

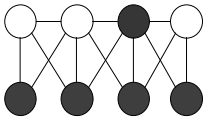


Incorporating Global Knowledge Sources: Using MaxEnt

- Occurrence Patterns:
  - Person names often referred to by the final word:
    - **Anton Schutte, an official with …**
    - **Schutte later added …**
  - Organization names often referred to by the first word:
    - **Scorpion Minerals Inc, a junior gold exploration company …**
    - **But Scorpion was raising a lot of eyebrows …**
- Other-cased instances:
  - **Defender Hassan Abbas rose to intercept**
  - **Final Partizan v Olympiakos.**
- HMMs are not able to naturally incorporate these features.

A Maximum-Entropy Model

- Klein et al. (2003)'s model chains together a sequence of maximum-entropy (logistic regression) classifiers.
- Each classifier makes a local choice of type based on:
  - Surrounding words
  - Surrounding character types
  - Surrounding part-of-speech tags
  - Previous (and/or next) classifications
  - Word substrings
  - Word occurrence patterns



$$P(c \mid f_1 \ldots f_n) \propto e^{\sum_i \lambda_{i\, fi(c)}}$$

- The benefit of the system comes not only from the basic model type, but from the use of better features (and better smoothing).

The Discriminative Advantage

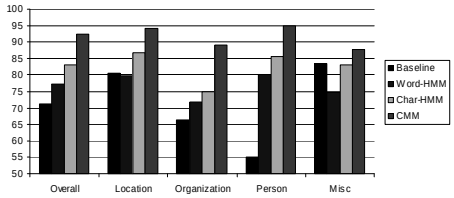
- Issues with multiple evidence sources:

  … morning at Grace Road

  - Evidence can be contradictory:
    - **Grace** looks like a person name.
    - **Road** looks like a location word.
  - Evidence can be redundant:
    - **Grace** and **Grac-** both usually indicate person names, but they aren't independent evidence.
- Unlike generative models (like HMMs), discriminative models (like maxent) take these issues into account when setting parameters.

Results: Discriminative Model

- Increases from better features, a better classification model.



CoNLL 2003 Shared Task: English NER