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'
    - 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]

2

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    - iPhrase "Traditional keyword search technology is hopelessly outdated"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

3

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

4

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.

5

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

6

Named Entity Recognition

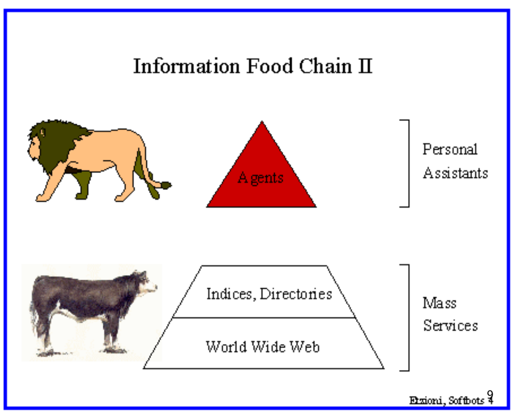
- Named entity recognition
  - Labeling names of things in web pages:
    - An entity is a discrete thing like "IBM Corporation"
      - But often extended in practice to things like dates, instances of products and chemical/biological substances that aren't really entities…
    - "Named" means called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

7

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)

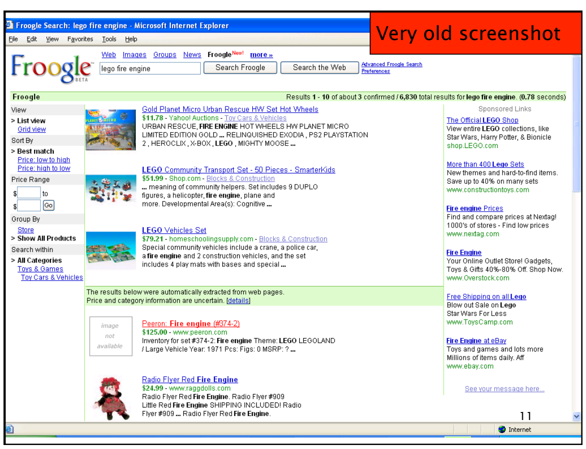
8



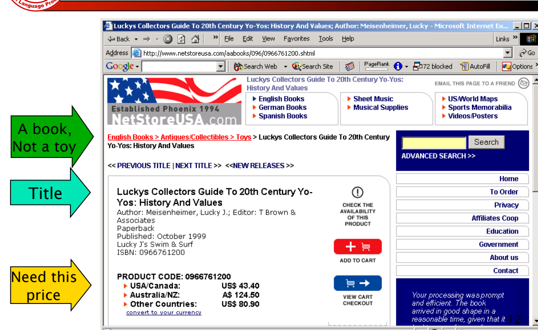
Product information/ Comparison shopping, etc.

- 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

10



Commercial information…



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

Jason invited you to "Danse Libre performance at the Palo Alto JCC" on Sunday, April 26 at 3:30pm.

Event: Danse Libre performance at the Palo Alto JCC
"A free ~1 hour performance of Victorian and Ragtime era dances."
What: Performance
Host: The Academy of Danse Libre
Start Time: Sunday, April 26 at 3:30pm
End Time: Sunday, April 26 at 5:00pm
Where: Cubberley Community Center A

Create New iCal Event...
Show This Date in iCal

13

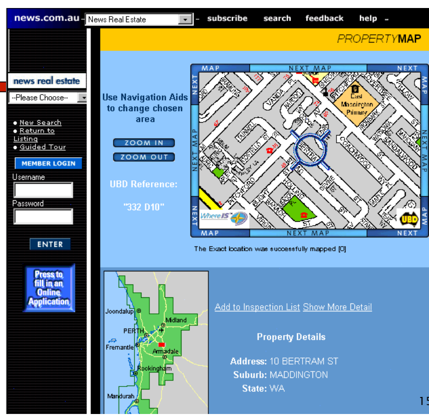
Classified Advertisements (Real Estate)

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>

14



15

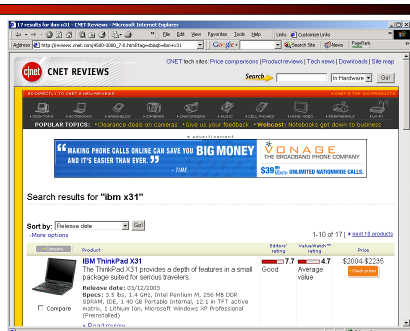
Why doesn't text search (IR) work?

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)

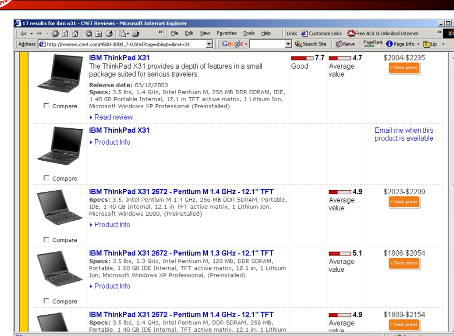
16

Canonicalization: Product information



17

Canonicalization: Product information



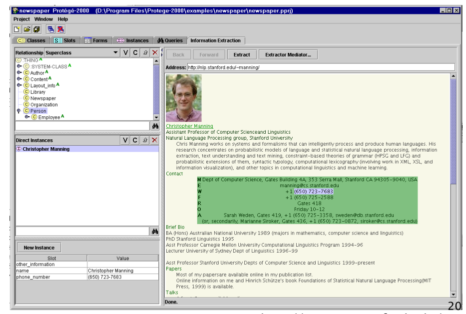
18

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.

19

Using information extraction to populate knowledge bases



20

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 [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.

21

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

22

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)

23

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 $\beta = 1$ (that is, $\alpha = ½$): $F = 2PR/(P+R)$
- Harmonic mean is conservative average
  - See CJ van Rijsbergen, *Information Retrieval*

24

Quiz question

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

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

- People usually use balanced F_1 measure
  - i.e., with $\beta = 1$ (that is, $\alpha = ½$): $F = 2PR/(P+R)$

25

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 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)

26

NER

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

27

Hand-written Information Extraction

- If extracting from automatically generated web pages, simple regex patterns usually work.
  - Amazon page
  - <div class="buying"><h1 class="parseasinTitle">(.*?)</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}

28

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:
    - Prefiller: [POS: V, Hypernym: KILL]
    - Filler: [Phrase: NP]

29

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"

30

Example of IE from 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.

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

31

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.

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

32

FASTUS

Based on finite state automata (FSA) transductions

set up
new Taiwan dollars

1.Complex Words:
Recognition of multi-words and proper names

a Japanese trading house
had set up

2.Basic Phrases:
Simple noun groups, verb groups and particles

production of 20, 000 iron and metal wood clubs

3.Complex phrases:
Complex noun groups and verb groups

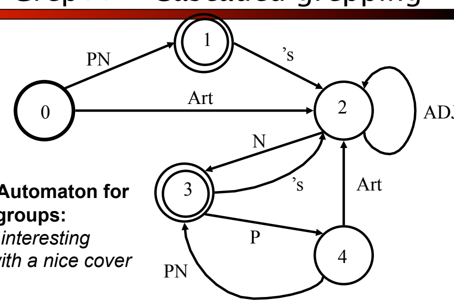
[company]
[set up]
[Joint-Venture]
with
[company]

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.

33

Grep++ = Cascaded grepping



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

34

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

35

Simple classification-based IE: Naive Bayes Classifiers

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

$$\langle x_1, x_2, \ldots, x_n \rangle$$

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

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

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

36

Naïve Bayes Classifier: Assumptions

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1,x_2,\ldots,x_n|c_j)$
  - $O(|X|^n \bullet |C|)$
  - 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.

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

Naïve Bayes in NLP

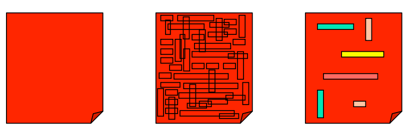
- For us, the x_i are usually bags of occurring words
  - A class-conditional unigram language model!
    - *Different* from having a variable for each word type!!
- As usual, we need to smooth $P(x_i|c_j)$
  - Zero probabilities cannot be conditioned away, no matter what other evidence there is

$$\hat{P}(x_{i,k} \mid c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}$$

- 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

Naive integration of IE & text classification

- Use conventional classification algorithms to classify substrings of document as "*to be extracted*" or not.



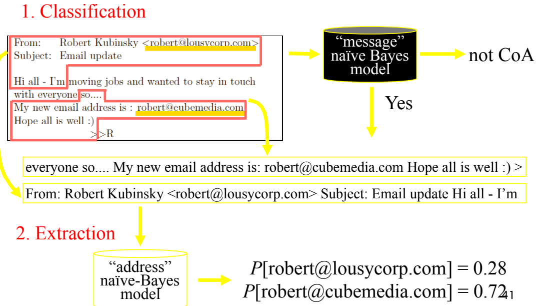
- In some simple but compelling domains, this naive technique is remarkably effective.
  - But think about when it would and wouldn't work!

'Change of Address' email

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

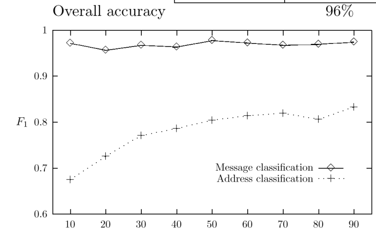
Hi all - I'm moving jobs and wanted to stay in touch
with everyone so....
My new email address is robert@cubemedia.com
Hope all is well :)
>>R

CoA: Details



Kushmerick et al. 2001 *ATEM*: Change of Address Results

	Words			Phrases		
	P	R	F_1	P	R	F_1
Message classification	.96	.66	.78	.98	.97	.98
Address classification	.96	.62	.76	.98	.68	.80
Overall accuracy					96%	



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