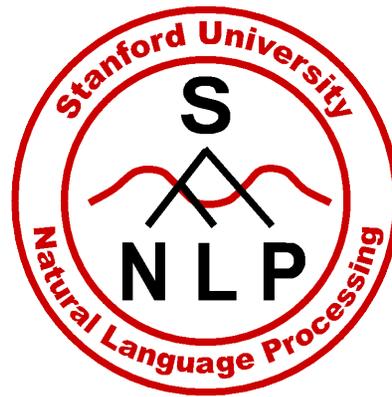


Information Extraction & Named Entity Recognition



Bill MacCartney

CS224N

[Based on slides by Chris 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'
 - 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?

- Word sense disambiguation technology generally works well (like text categorization)
- Synonyms can be found or listed
- OMG, there should be an NLP search engine!!!
 - Especially around 1999–2000
 - Lots of (ex-)startups:
 - LingoMotors
 - iPhrase “Traditional keyword search technology is hopelessly outdated”
 - PowerSet



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

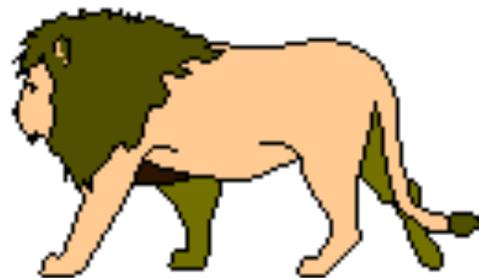
- Named entity recognition
 - Labeling names of things in web pages:
 - An entity is a discrete thing like “IBM Corporation”
 - But often extended in practice to things like dates, instances of products and chemical/biological substances that aren’t really entities...
 - “Named” means called “IBM” or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



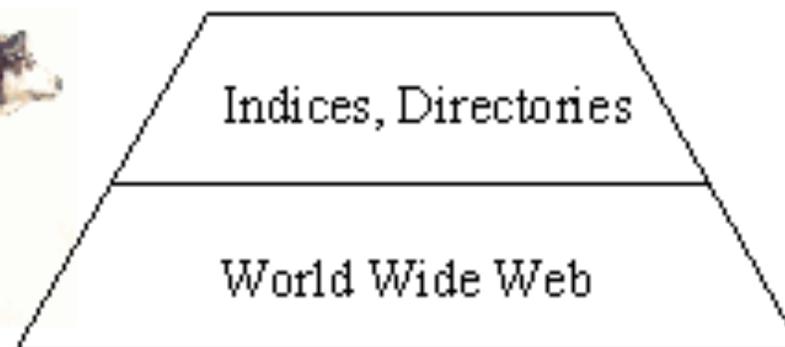
Information Extraction

- 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, $4.4x109)`
 - `lives(Chris, Palo Alto)`

Information Food Chain II



Personal
Assistants



Mass
Services



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 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

Very old screenshot



Web Images Groups News Froogle ^{New!} more »

lego fire engine

Search Froogle

Search the Web

Advanced Froogle Search Preferences

Froogle

Results 1 - 10 of about 3 confirmed / 6,830 total results for lego fire engine. (0.78 seconds)

View

- > List view
- Grid view

Sort By

- > Best match
- Price: low to high
- Price: high to low

Price Range

\$ to \$

Group By

- Store
- > Show All Products

Search within

- > All Categories
- Toys & Games
- Toy Cars & Vehicles



[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 ...

The results below were automatically extracted from web pages. Price and category information are uncertain. [\[details\]](#)



[Peeron: Fire engine \(#374-2\)](#)
\$125.00 - www.peeron.com
 Inventory for set #374-2: Fire engine Theme: LEGO LEGOLAND / Large Vehicle Year: 1971 Pcs: Figs: 0 MSRP: ? ...



[Radio Flyer Red Fire Engine](#)
\$24.99 - www.raggdolls.com
 Radio Flyer Red Fire Engine. Radio Flyer #909 Little Red Fire Engine SHIPPING INCLUDED! Radio Flyer #909 ... Radio Flyer Red Fire Engine.

Sponsored Links

[The Official LEGO Shop](#)
 View entire LEGO collections, like Star Wars, Harry Potter, & Bionicle [shop.LEGO.com](#)

[More than 400 Lego Sets](#)
 New themes and hard-to-find items. Save up to 40% on many sets [www.constructiontoys.com](#)

[Fire engine Prices](#)
 Find and compare prices at Nextag! 1000's of stores - Find low prices [www.nextag.com](#)

[Fire Engine](#)
 Your Online Outlet Store! Gadgets, Toys & Gifts 40%-80% Off. Shop Now. [www.Overstock.com](#)

[Free Shipping on all Lego](#)
 Blow out Sale on Lego Star Wars For Less [www.ToysCamp.com](#)

[Fire Engine at eBay](#)
 Toys and games and lots more Millions of items daily. Aff [www.ebay.com](#)

[See your message here...](#)



Commercial information...

A book,
Not a toy

Title

Need this
price

The screenshot shows a Microsoft Internet Explorer browser window displaying a product page on NetStoreUSA.com. The browser's address bar shows the URL: <http://www.netstoreusa.com/aabooks/096/0966761200.shtml>. The page features a navigation menu with categories like English Books, German Books, Spanish Books, Sheet Music, Musical Supplies, US/World Maps, Sports Memorabilia, and Videos/Posters. The main content area displays the product title, author (Meisenheimer, Lucky J.), editor (T Brown & Associates), and publication details (October 1999). A price table is visible, showing the product code 0966761200 and prices for USA/Canada (US\$ 43.40), Australia/NZ (A\$ 124.50), and Other Countries (US\$ 80.90). A 'CHECK THE AVAILABILITY OF THIS PRODUCT' button is present, along with 'ADD TO CART' and 'VIEW CART CHECKOUT' buttons. A sidebar on the right contains a search bar, an 'ADVANCED SEARCH >>' link, and a list of navigation links including Home, To Order, Privacy, Affiliates Coop, Education, Government, About us, and Contact. A testimonial at the bottom right states: 'Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it'.

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

[convert to your currency](#)

ADD TO CART

VIEW CART CHECKOUT

Home

To Order

Privacy

Affiliates Coop

Education

Government

About us

Contact

Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it



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



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

17 results for ibm x31 - CNET Reviews - Microsoft Internet Explorer

Address: http://reviews.cnet.com/4500-3000_7-0.html?tag=sb&qt=ibm+x31

CNET tech sites: Price comparisons | Product reviews | Tech news | Downloads | Site map

CNET REVIEWS Search [] In Hardware [] Go!

GO DIRECTLY TO CNET'S NEW REVIEWS: CNET'S TOP 100 PRODUCTS

DESKTOPS | NOTEBOOKS | HANDHELD | CAMERAS | CAMCORDERS | MUSIC | CELL PHONES | HOME VIDEO | PERIPHERALS | WI-FI

POPULAR TOPICS: Clearance deals on cameras | Give us your feedback | Webcast: Notebooks get down to business

advertisement

“MAKING PHONE CALLS ONLINE CAN SAVE YOU **BIG MONEY** AND IT'S EASIER THAN EVER.” - TIME

VONAGE THE BROADBAND PHONE COMPANY
\$39⁹⁹ MONTH UNLIMITED NATIONWIDE CALLS.

Search results for "ibm x31"

Sort by: Release date [] Go! More options

1-10 of 17 | next 10 products

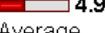
Product	Editors' rating	ValueWatch™ rating	Price
 <p>IBM ThinkPad X31 The ThinkPad X31 provides a depth of features in a small package suited for serious travelers. Release date: 03/12/2003 Specs: 3.5 lbs, 1.4 GHz, Intel Pentium M, 256 MB DDR SDRAM, IDE, 1 40 GB Portable Internal, 12.1 in TFT active matrix, 1 Lithium Ion, Microsoft Windows XP Professional (Preinstalled) Read review</p>	<p>7.7 Good</p>	<p>4.7 Average value</p>	<p>\$2004-\$2235 Check prices</p>



Canonicalization: Product information

17 results for ibm x31 - CNET Reviews - Microsoft Internet Explorer

Address: http://reviews.cnet.com/4500-3000_7-0.html?tag=sb&qt=ibm+x31

 <input type="checkbox"/> Compare	IBM ThinkPad X31 The ThinkPad X31 provides a depth of features in a small package suited for serious travelers. Release date: 03/12/2003 Specs: 3.5 lbs, 1.4 GHz, Intel Pentium M, 256 MB DDR SDRAM, IDE, 1 40 GB Portable Internal, 12.1 in TFT active matrix, 1 Lithium Ion, Microsoft Windows XP Professional (Preinstalled) Read review	 7.7 Good	 4.7 Average value	\$2004-\$2235 Check prices
 <input type="checkbox"/> Compare	IBM ThinkPad X31 Product Info			Email me when this product is available
 <input type="checkbox"/> Compare	IBM ThinkPad X31 2672 - Pentium M 1.4 GHz - 12.1" TFT Specs: 3.5, Intel Pentium M 1.4 GHz, 256 MB DDR SDRAM, Portable, IDE, 1 40 GB Internal, 12.1 in TFT active matrix, 1 Lithium Ion, Microsoft Windows 2000, (Preinstalled) Product Info		 4.9 Average value	\$2023-\$2299 Check prices
 <input type="checkbox"/> Compare	IBM ThinkPad X31 2672 - Pentium M 1.3 GHz - 12.1" TFT Specs: 3.5 lbs, 1.3 GHz, Intel Pentium M, 128 MB, DDR SDRAM, Portable, 1 20 GB IDE Internal, TFT active matrix, 12.1 in, 1 Lithium Ion, Microsoft Windows XP Professional, (Preinstalled) Product Info		 5.1 Average value	\$1806-\$2054 Check prices
 <input type="checkbox"/> Compare	IBM ThinkPad X31 2672 - Pentium M 1.4 GHz - 12.1" TFT Specs: 3.5 lbs, 1.4 GHz, Intel Pentium M, DDR SDRAM, 256 MB, Portable, 1 40 GB IDE Internal, TFT active matrix, 12.1 in, 1 Lithium		 4.9 Average value	\$1809-\$2154 Check 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

newspaper Protégé-2000 (D:\Program Files\Protege-2000\examples\newspaper\newspaper.pprj)

Project Window Help

Classes Slots Forms Instances Queries Information Extraction

Relationship Superclass

- THING
- SYSTEM-CLASS
- Author
- Content
- Layout_info
- Library
- Newspaper
- Organization
- Person
- Employee

Direct Instances

- Christopher Manning

New Instance

Slot	Value
other_information	
name	Christopher Manning
phone_number	(650) 723-7683

Address: <http://nlp.stanford.edu/~manning/>



Christopher Manning
 Assistant Professor of Computer Science and Linguistics
 Natural Language Processing group, Stanford University

Chris Manning works on systems and formalisms that can intelligently process and produce human languages. His research concentrates on probabilistic models of language and statistical natural language processing, information extraction, text understanding and text mining, constraint-based theories of grammar (HPSG and LFG) and probabilistic extensions of them, syntactic typology, computational lexicography (involving work in XML, XSL, and information visualization), and other topics in computational linguistics and machine learning.

Contact

M Dept of Computer Science, Gates Building 4A, 353 Serra Mall, Stanford CA 94305-9040, USA
E manning@cs.stanford.edu
W +1 (650) 723-7683
F +1 (650) 725-2588
R Gates 418
O Friday 10-12
A Sarah Weden, Gates 419, +1 (650) 725-3358, sweden@db.stanford.edu
 (or, secondarily, Marianne Siroker, Gates 436, +1 (650) 723-0872, siroker@cs.stanford.edu)

Brief Bio
 BA (Hons) Australian National University 1989 (majors in mathematics, computer science and linguistics)
 PhD Stanford Linguistics 1995
 Asst Professor Carnegie Mellon University Computational Linguistics Program 1994-96
 Lecturer University of Sydney Dept of Linguistics 1996-99

Asst Professor Stanford University Depts of Computer Science and Linguistics 1999-present

Papers
 Most of my papers are available online in my publication list.
 Online information on me and Hinrich Schütze's book Foundations of Statistical Natural Language Processing (MIT Press, 1999) is available.

Talks

Done.



Named Entity Recognition



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.



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:** % of selected items that are correct
= $P(\text{selected \& correct} \mid \text{selected})$
- **Recall:** % of correct items that are selected
= $P(\text{selected \& correct} \mid \text{correct})$

	correct	not correct
selected	tp	fp
not selected	fn	tn

$$\text{Precision } P = \text{tp} / (\text{tp} + \text{fp})$$

$$\text{Recall } R = \text{tp} / (\text{tp} + \text{fn})$$



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



Quiz question

Which of the following combinations of precision and recall yields the highest F_1 ?

- (a) $P = 35\%$ $R = 100\%$
- (b) $P = 45\%$ $R = 80\%$
- (c) $P = 55\%$ $R = 55\%$
- (d) $P = 65\%$ $R = 50\%$
- (e) $P = 75\%$ $R = 45\%$



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)



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">(.*?)</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:
 - Prefiller: [POS: V, Hypernym: 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"

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

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



FASTUS

Based on finite state automata (FSA) transductions

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]

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

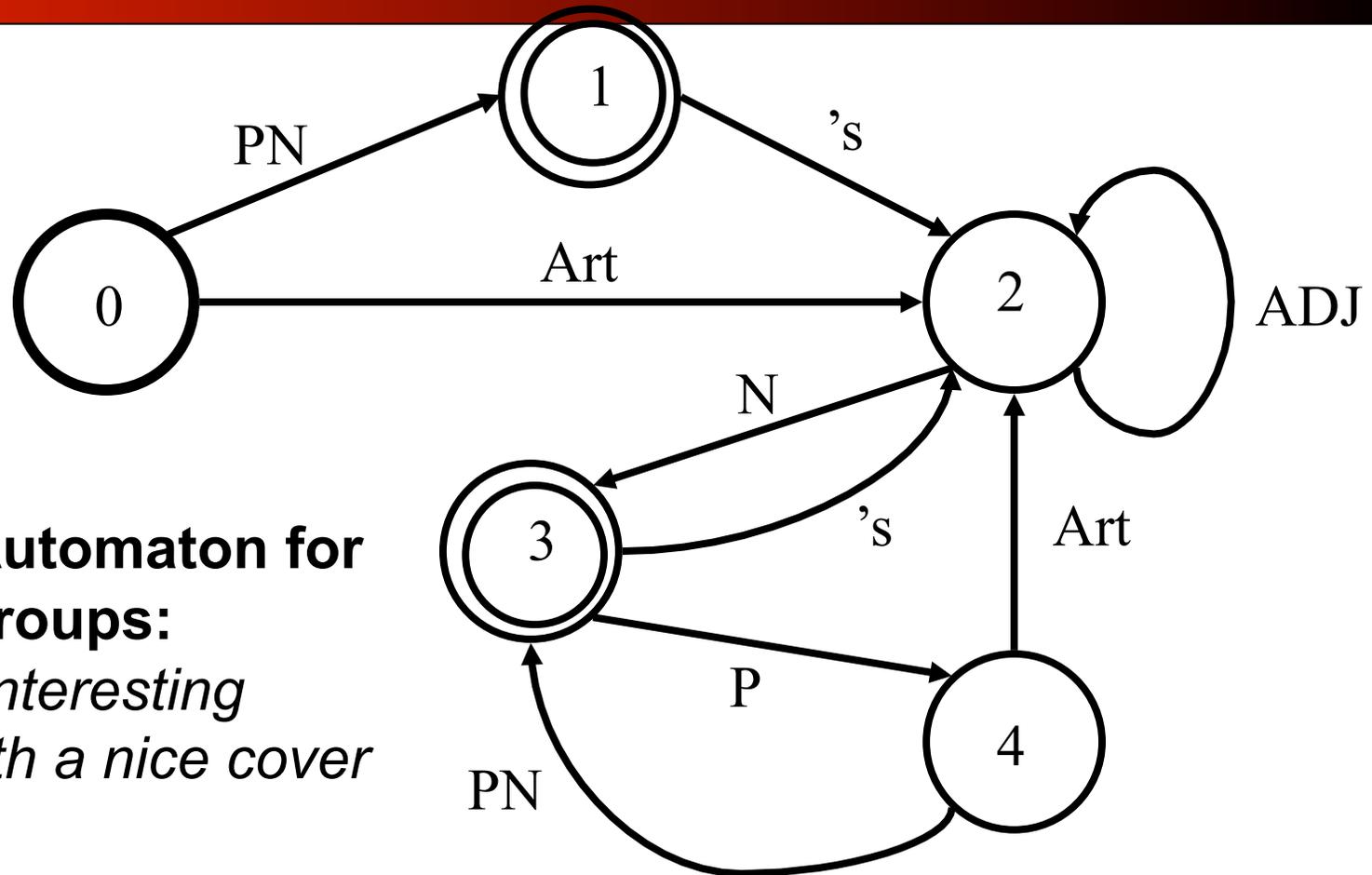
Basic templates are to be built.

5. Merging Structures:

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



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

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

$$c_{MAP} = \operatorname{argmax}_{c_j \in C} P(c_j | x_1, x_2, \dots, x_n)$$

$$c_{MAP} = \operatorname{argmax}_{c_j \in C} \frac{P(x_1, x_2, \dots, x_n | c_j) P(c_j)}{P(c_1, c_2, \dots, c_n)}$$

$$c_{MAP} = \operatorname{argmax}_{c_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j)$$



Naïve Bayes Classifier: Assumptions

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

$$P(X_1, \dots, X_5 | C) = P(X_1 | C) \cdot P(X_2 | C) \cdot \dots \cdot P(X_5 | C)$$



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

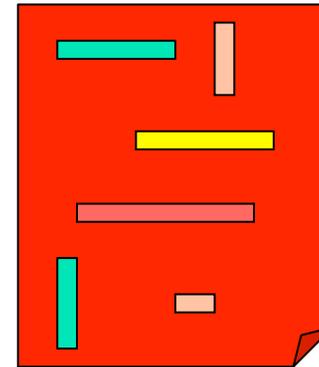
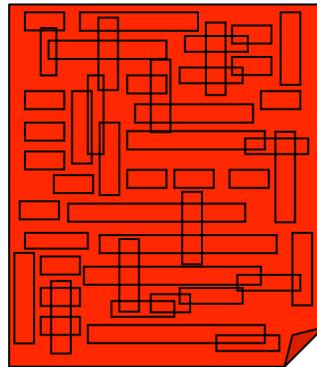
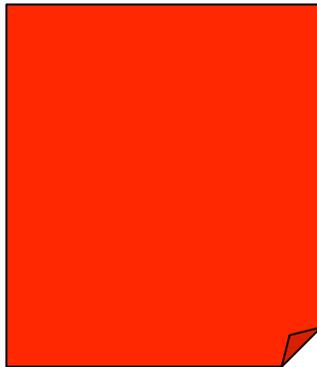
$$\hat{P}(x_{i,k} | c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}$$

- 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



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 :)

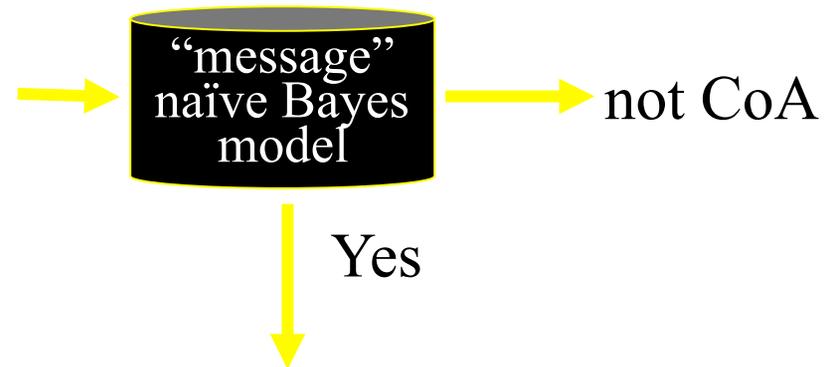
>>R



CoA: Details

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 :)
>>R



everyone so.... My new email address is: robert@cubemedia.com Hope all is well :) >

From: Robert Kubinsky <robert@lousycorp.com> Subject: Email update Hi all - I'm

2. Extraction

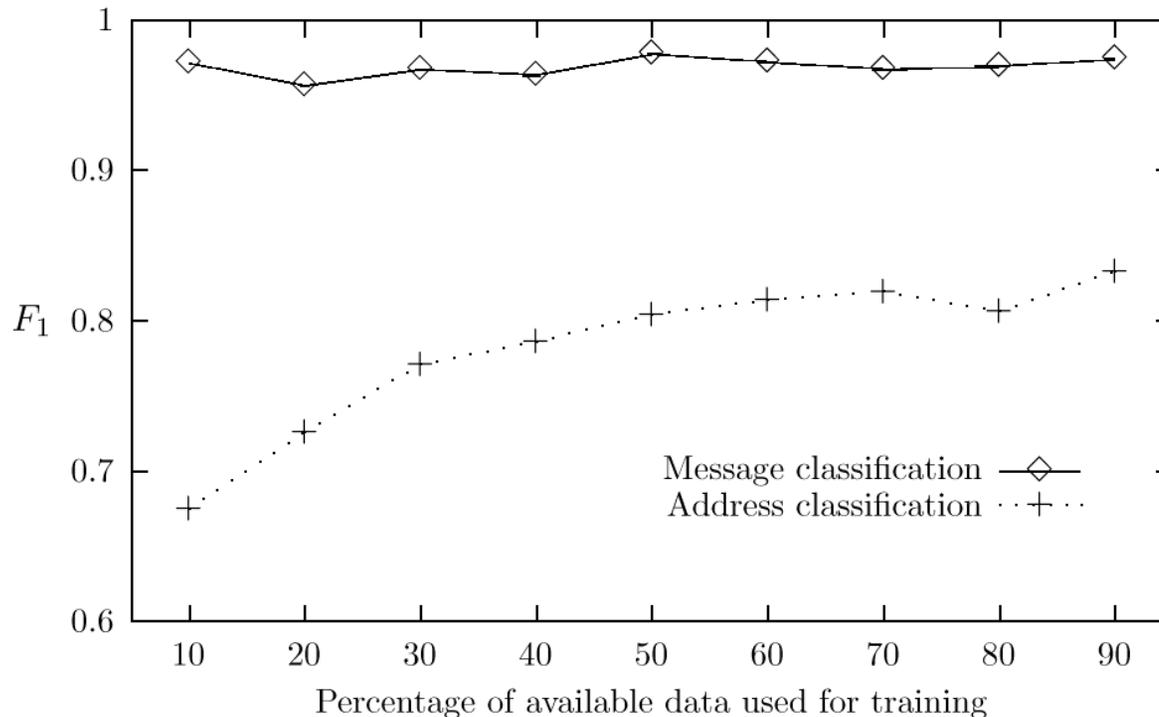


$$P[\text{robert@lousycorp.com}] = 0.28$$
$$P[\text{robert@cubemedia.com}] = 0.72_3$$



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

	Words			Phrases		
	<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> 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