Information Extraction

Christopher Manning
CS224N – 2002
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 ... 14 months ago
  - Lots of 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    - iPhrase "Traditional keyword search technology is hopelessly outdated"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*

Information Food Chain II

Agents

Personal Assistants

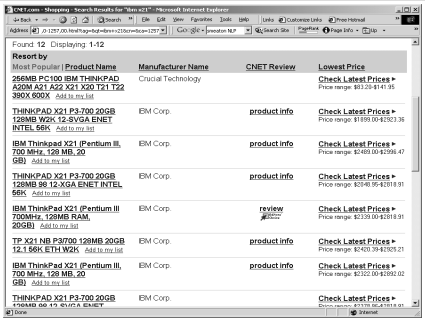
Indices, Directories

World Wide Web

Mass Services

Etzioni, Softbots 4

Product information



Product info

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

Inconsistency: digital cameras

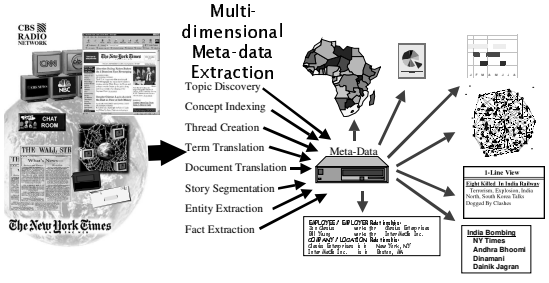
- Image Capture Device: 1.68 million pixel 1/2-inch CCD sensor
- Image Capture Device Total Pixels Approx. 3.34 million Effective Pixels Approx. 3.24 million
- Image sensor Total Pixels: Approx. 2.11 million-pixel
- Imaging sensor Total Pixels: Approx. 2.11 million 1,688 (H) x 1,248 (V)
- CCD Total Pixels: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

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

Knowledge Extraction Vision

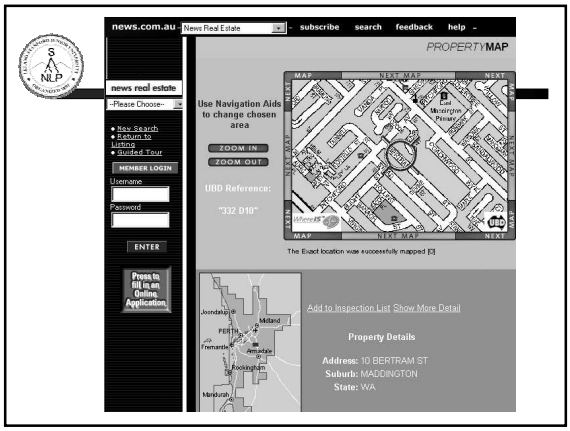


Classified Advertisements (Real Estate)

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

Example of IE from FASTUS (1993): it isn't easy!

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

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

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

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

Task: Wrapper Induction

- Wrapper Induction
  - Sometimes, the relations are structural.
    - Web pages generated by a database.
    - Tables, lists, etc.
  - Wrapper induction is usually regular relations which can be expressed by the *structure* of the document:
    - the item in bold in the 3rd column of the table is the price
- Handcoding a wrapper in Perl isn't very viable
  - sites are numerous, and their surface structure mutates rapidly (around 10% failures each month)
- Wrapper induction techniques can also learn:
  - If there is a page about a research project X and there is a link near the word 'people' to a page that is about a person Y then Y is a member of the project X.
    - [e.g, Tom Mitchell's Web->KB project]

Examples of Existing IE Systems

- Systems to summarize medical patient records by extracting diagnoses, symptoms, physical findings, test results, and therapeutic treatments.
- Gathering earnings, profits, board members, etc. from company reports
- Verification of construction industry specifications documents (are the quantities correct/reasonable?)
- Real estate advertisements
- Building job databases from textual job vacancy postings
- Extraction of company take-over events
- Extracting gene drug interactions from biomed texts

Three generations of IE systems

- Hand-Built Systems – Knowledge Engineering [1980s–]
  - Rules written by hand
  - Require experts who understand both the systems and the domain
  - Iterative guess-test-tweak-repeat cycle
- Automatic, Trainable Rule-Extraction Systems [1990s–]
  - Rules discovered automatically using predefined templates, using methods like ILP
  - Require huge, labeled corpora (effort is just moved!)
- Statistical Generative Models [1997 –]
  - One decodes the statistical model to find which bits of the text were relevant, using HMMs or statistical parsers
  - Learning usually supervised; may be partially unsupervised

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

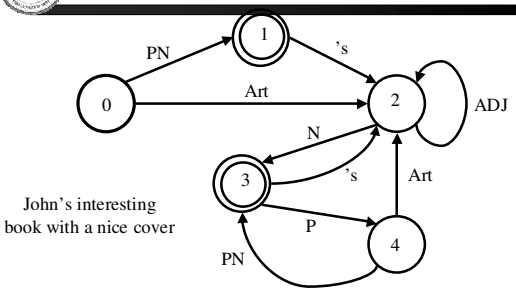
2.Basic Phrases:
Simple noun groups, verb groups and particles

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.

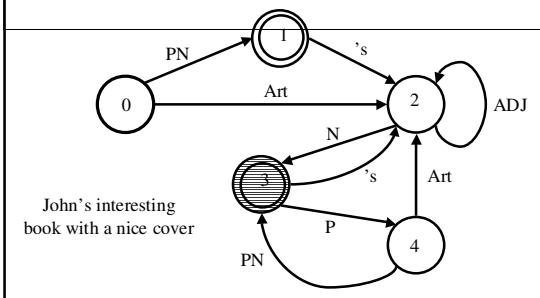
5. Merging Structures:
Templates from different parts of the texts are merged if they provide information about the same entity or event.

Grep++



Pattern-maching {PN 's | Art}(ADJ)* N (P Art (ADJ)* N)*

PN 's (ADJ)* N P Art (ADJ)* N



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 "metal wood" clubs a month.

1.Complex words

Attachment Ambiguities are not made explicit

2.Basic Phrases:

Bridgestone Sports Co.	: Company name
said	: Verb Group
Friday	: Noun Group
it	: Noun Group
had set up	: Verb Group
a joint venture	: Noun Group
in	: Preposition
Taiwan	: Location

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 "metal wood" clubs a month.

2.Basic Phrases:

Bridgestone Sports Co.	: Company name
said	: Verb Group
Friday	: Noun Group
it	: Noun Group
had set up	: Verb Group
a joint venture	: Noun Group
in	: Preposition
Taiwan	: Location

3.Complex Phrases

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

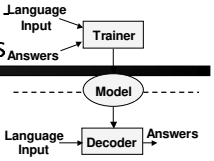
Determining where an organization is located

- [org] *in* [loc]
  - NATO headquarters in Brussels
- [org] [loc] (*division*, *branch*, *headquarters*, etc.)
  - KFOR Kosovo headquarters

Machine learning

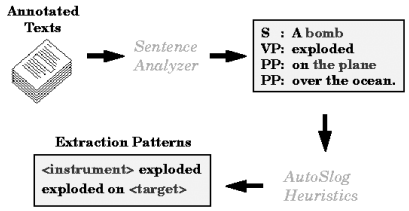
- To keep up with and exploit the web, you need to be able to *learn*
  - Discovery: How do you find new information sources *S*?
  - Extraction: How can you access and parse the information in *S*?
  - Semantics: How does one understand and link up the information in contained in *S*?
  - Pragmatics: What is the accuracy, reliability, and scope of information in *S*?
- Hand-coding just doesn't scale

Automatic Frame Learning Systems



- Pros:
  - Portable across domains
  - Tend to have broad coverage
  - Robust in the face of degraded input.
  - Automatically finds appropriate statistical patterns
  - System knowledge not needed by those who supply the domain knowledge.
- Cons:
  - Annotated training data, and lots of it, is needed.
  - Isn't necessarily better or cheaper than hand-built sol'n
- Classic e.g.: Riloff et al., AutoSlog (UMass) learns lexico-syntactic extraction patterns from templates

Autoslog



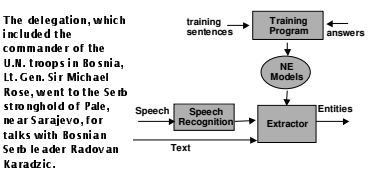
Information Extraction: Learning Lexico-Syntactic Patterns (Autoslog-TS, Riloff)

- Reduces greatly the needed resources: just classifications
- Start with known relevant articles; Extract noun phrases
- Create huge set of patterns of form:
  - **<subj> passive-verb** <victim> was murdered
  - **active-verb <dobj>** bombed <target>
- Use: key lexical items, syntactic frames, semantic sorts
- Compute relevance rate:
  - $Pr(\text{relevant}|\text{text has pattern}_i) = \text{rel-freq}_i / \text{total-freq}_i$
- Rank patterns in order of importance
  - relevance rate * $\log_2(\text{freq}_i)$, rr > 0.5
- Human judge reviewed top patterns
- Apply all patterns to each text. Worked almost as well

Statistical generative models

- Pros:
  - Well-understood underlying statistical model makes it easy to used wide range of tools from statistical decision theory
  - Portable across domains
  - Tend to have broad coverage, robustness, good recall
  - Data-driven
- Cons:
  - Range of features and patterns available may be limited
  - Not necessarily as good for complex multi-slot patterns
- Good current e.g.: Freitag & McCallum (CMU, JustSystems, Whizbang! Labs) statistically-learned models using Bayes classifiers, HMMs, etc.

Name Extraction via HMMs



- 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