Information Extraction and Integration

All slides from:

William Cohen
Andrew McCallum
Eugene Agichtein
Sunita Sarawagi

The Value of Text Data

- "Unstructured" text data is the primary source of human-generated information
  - Citeseer, comparison shopping, PIM systems, web search, data warehousing
- Managing and utilizing text: information extraction and integration
- Scalability: a bottleneck for deployment
- Relevance to data mining community

Example: A Solution

Extracting Job Openings from the Web

Job Openings:
Category = Food Services
Keyword = Baker
Location = Continental U.S.

What is "Information Extraction"

As a task: **Filling slots in a database from sub-segments of text.**

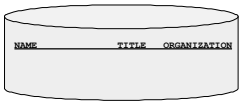
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

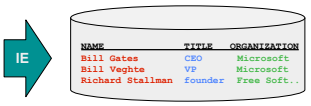
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



What is "Information Extraction"

As a family of techniques: Information Extraction = segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka "named entity extraction"

What is "Information Extraction"

As a family of techniques: Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is "Information Extraction"

As a family of techniques: Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

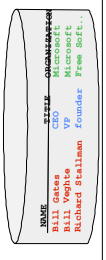
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

* Microsoft Corporation
CEO
Bill Gates
* Microsoft
Gates
* Microsoft
* Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation



IE is different in different domains!

Example: on web there is less grammar, but more formatting & linking

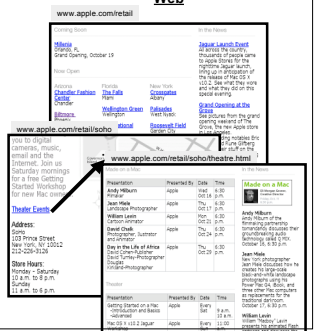
Newswire

Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

Web



The directory structure, link structure, formatting & layout of the Web is its own new grammar.

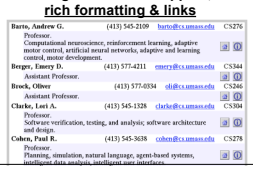
Landscape of IE Tasks (1/4):
Degree of Formatting

Text paragraphs without formatting

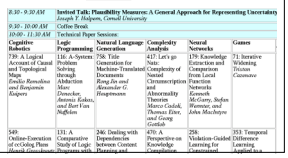
Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences and some formatting & links

Non-grammatical snippets, rich formatting & links



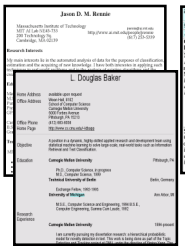
Tables



Landscape of IE Tasks (2/4):
Intended Breadth of Coverage

Web site specific
Formatting
Amazon.com Book Pages

Genre specific
Layout
Resumes



Wide, non-specific
Language
University Names

Landscape of IE Tasks (3/4):
Complexity

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Landscape of IE Tasks (4/4):
Single Field/Record

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: Jack Welch

Person: Jeffrey Immelt

Location: Connecticut

Binary relationship

Relation: Person-Title
Person: Jack Welch
Title: CEO

Relation: Company-Location
Company: General Electric
Location: Connecticut

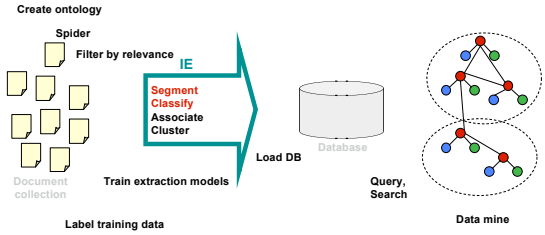
N-ary record

Relation: Succession
Company: General Electric
Title: CEO
Out: Jack Welsh
In: Jeffrey Immelt

"Named entity" extraction

Broader View

Up to now we have been focused on segmentation and classification



Hand Coded Rule Example: Conference Name

```
# These are subordinate patterns
$wordOrdinals="(?:first|second|third|fourth|fifth|sixth|seventh|eighth|ninth|tenth|eleventh|twelfth|thirtee
rteenth|fifteenth)";
my $numberOrdinals="(?:\\d?(?:1st|2nd|3rd|1th|2th|3th|4th|5th|6th|7th|8th|9th|0th))";
my $ordinals="(?:$wordOrdinals|$numberOrdinals)";
my $confTypes="(?:Conference|Workshop|Symposium)";
my $words="(?:[A-Z]\\w+\\s*)"; # A word starting with a capital letter and ending with 0 or more spaces
my $confDescriptors="(?:international\\s+[A-Z]+\\s+)"; # .e.g "International Conference ..." or the confe
name for workshops (e.g. "VLDB Workshop ...")
my $connectors="(?:on|of)";
my $abbreviations="(?:\\([A-Z]\\w\\w+[\\W\\s]*?(?:\\d\\d+)?\\))"; # Conference abbreviations like "(SIGMO
# The actual pattern we search for. A typical conference name this pattern will find is
# "3rd International Conference on Blah Blah Blah (ICBBB-05)"
my
$fullNamePattern="((?:$ordinals\\s+$words*|$confDescriptors)?$confTypes(?:\\s+$connectors\\s+.*?|\\
breviations?)(?:\\n|\\r|\\.|<)";
########################################## ##############################
# Given a <dbworldMessage>, look for the conference pattern
##################################################################
lookForPattern($dbworldMessage, $fullNamePattern);
##################################################################
# In a given <file>, look for occurrences of <pattern>
# <pattern> is a regular expression
##################################################################
sub lookForPattern {
    my ($file,$pattern) = @_;
```

Machine Learning Methods

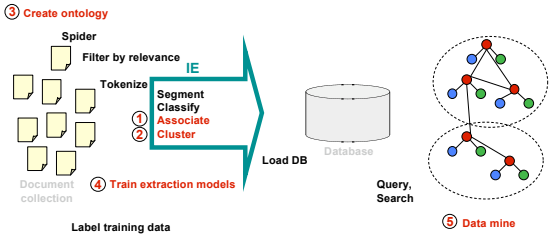
- Sequence models: HMMs, CMMs/MEMMs, CRFs
- Can work well when training data is easy to construct and is plentiful
- Can capture complex patterns that are hard to encode with hand-crafted rules
  - e.g., determine whether a review is positive or negative
  - extract long complex gene names

> ***The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300."***

- Can be labor intensive to construct training data
  - Question: how much training data is sufficient?

Broader View

Now touch on some other issues



Relation Extraction: Disease Outbreaks

- **E**xtract structured relations from text

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly **Ebola** epidemic in **Zaire**, is finding itself hard pressed to cope with the crisis...

Information Extraction System (e.g., NYU's Proteus)

Disease Outbreaks in ***The New York Times***

Date	*Disease Name*	*Location*
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Example: Protein Interactions

„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."

CBF-A <interact / complex> CBF-C

CBF-B <associates> CBF-A-CBF-C complex

Relation Extraction

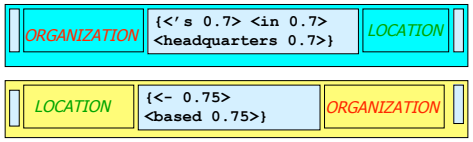
- Typically require Entity Tagging as preprocessing
- Knowledge Engineering
  - Rules defined over lexical items
    - "<company> located in <location>"
  - Rules defined over parsed text
    - "((Obj <company>) (Verb located) (*) (Subj <location>))"
  - Proteus, GATE, …
- Machine Learning-based
  - Learn rules/patterns from examples
    Dan Roth 2005, Cardie 2006, Mooney 2005, …
  - Partially-supervised: bootstrap from "seed" examples
    Agichtein & Gravano 2000, Etzioni et al., 2004, …
- Recently, hybrid models [Feldman2004, 2006]

Example Extraction Rule [NYU Proteus]

```
;;; For <company> appoints <person> <position>

(defpattern appoint
   "np-sem(C-company)?  rn?  sa?  vg(C-appoint) np-sem(C-person) ','?
    to-be?  np(C-position) to-succeed?:
    company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attributes
    position-at=8.attributes |
...
(defun when-appoint (phrase-type)
    (let ((person-at (binding 'person-at))
       (company-entity (entity-bound 'company-at))
       (person-entity (essential-entity-bound 'person-at 'C-person))
       (position-entity (entity-bound 'position-at))
       (predecessor-entity (entity-bound 'predecessor-at))
       new-event)
     (not-an-antecedent position-entity)
     ;; if no company is specified for position, use agent
...
```

Example Extraction Patterns: Snowball [AG2000]



(1) Association as Binary Classification

Christos Faloutsos conferred with Ted Senator, the KDD 2003 General Chair.

Person Person Role

Person-Role (Christos Faloutsos, KDD 2003 General Chair) → NO

Person-Role (Ted Senator, KDD 2003 General Chair) → YES

Do this with SVMs and tree kernels over parse trees.

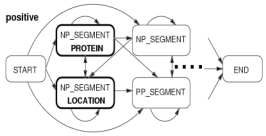
[Zelenko et al, 2002]

(1) Association with Finite State Machines

[Ray & Craven, 2001]

… This enzyme, UBC6, localizes to the endoplasmic reticulum, with the catalytic domain facing the cytosol. …

DET	this
N	enzyme
N	ubc6
V	localizes
PREP	to
ART	the
ADJ	endoplasmic
N	reticulum
PREP	with
ART	the
ADJ	catalytic
N	domain
V	facing
ART	the
N	cytosol

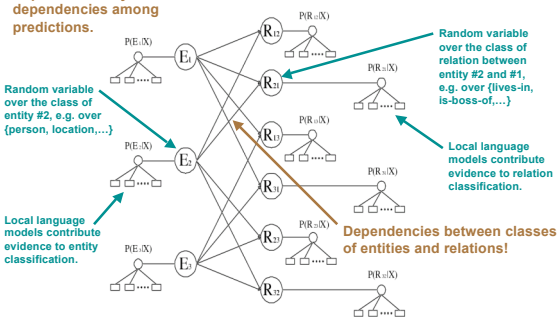


Subcellular-localization (UBC6, endoplasmic reticulum)

(1) Association with Graphical Models

[Roth & Yih 2002]

Capture arbitrary-distance dependencies among predictions.

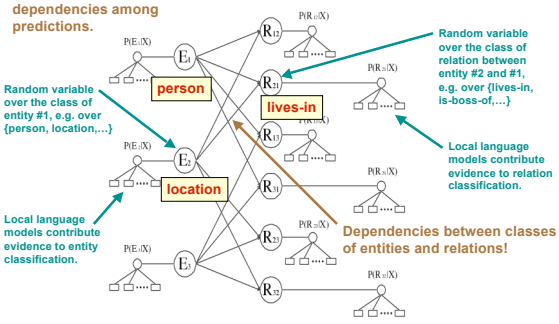


Inference with loopy belief propagation.

(1) Association with Graphical Models

[Roth & Yih 2002]

Also capture long-distance dependencies among predictions.



Inference with loopy belief propagation.

Accuracy of Information Extraction

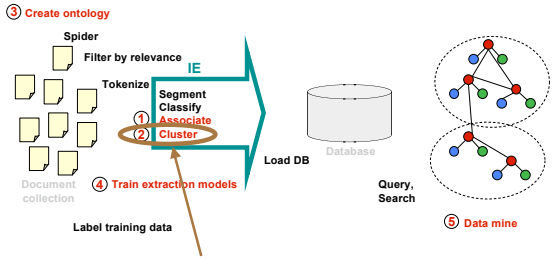
Information Type	Accuracy
Entities	90-98%
Attributes	80%
Facts	60-70%
Events	50-60%

[Feldman, ICML 2006 tutorial]

- Errors cascade (error in entity tag → error in relation extraction)
- This estimate is optimistic:
  - Holds for well-established tasks
  - Many specific/novel IE tasks exhibit lower accuracy

Broader View

Now touch on some other issues



When do two extracted strings refer to the same object?

Extracted Entities: Resolving Duplicates


Document 1: The Justice Department has officially ended its inquiry into the assassinations of **John F. Kennedy** *and Martin Luther King Jr., finding ``no persuasive evidence'' to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that* **Kennedy** *was ``probably'' assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the Warren Commission 's belief that Lee Harvey Oswald acted alone in Dallas on Nov. 22, 1963.*

Document 2: In 1953, Massachusetts **Sen. John F. Kennedy** *married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate* **John F. Kennedy** *confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, ``I do not speak for my church on public matters, and the church does not speak for me.''*

Document 3: **David Kennedy** *was born in Leicester, England in 1959. ...***Kennedy** *co-edited The New Poetry (Bloodaxe Books 1993), and is the author of New Relations: The Refashioning Of British Poetry 1980-1994 (Seren 1996).*

[From Li, Morie, & Roth, AI Magazine, 2005]

Important Problem

- Appears in numerous real-world contexts
- Plagues many applications
  - Citeseer, DBLife, AliBaba, Rexa, etc.

(2) Information Integration

[Minton, Knoblock, et al 2001], [Doan, Domingos, Halevy 2001], [Richardson & Domingos 2003]

Goal might be to merge results of two IE systems:

Name:	Introduction to Computer Science
Number:	CS 101
Teacher:	M. A. Kludge
Time:	9-11am
Name:	Data Structures in Java
Room:	5032 Wean Hall

Title:	Intro. to Comp. Sci.
Num:	101
Dept:	Computer Science
Teacher:	Dr. Klüdge
TA:	John Smith
Topic:	Java Programming
Start time:	9:10 AM