

# Information Extraction from the World Wide Web

**Andrew McCallum**

*University of Massachusetts Amherst*

**William Cohen**

*Carnegie Mellon University*

# Example: The Problem

The image shows a screenshot of a Google search results page. At the top, the Google logo is on the left, and navigation links for 'Advanced Search', 'Preferences', 'Language Tools', and 'Search Tips' are on the right. Below the logo is a search input field containing the text 'baker job opening' and a 'Google Search' button. Underneath the search bar is a horizontal menu with tabs for 'Web', 'Images', 'Groups', 'Directory', and 'News-New!'. A blue banner below the menu reads 'Searched the web for **baker job opening**.' To the right of this banner, the word 'Results' is partially visible. The search results are listed below, each with a blue title link, a green URL, and a snippet of text. The results include: 1. 'Job Opening - Find ANY Job! - Search by Type, Industry & Geography' from www.careerbuilder.com. 2. 'Job Opening At Flipdog.Com' from www.FlipDog.com. 3. 'Softimage::Community::Discussion Groups::ds.archive.0004' with a snippet mentioning 'Le Rudulier', 'Ken Skaggs', and 'Martin Baker'. 4. Another 'Softimage::Community::Discussion Groups::ds.archive.0004' result with a snippet mentioning 'Philip Herring' and 'Martin Baker'. 5. 'CGI: Job Opening' from www.genomics.cornell.edu. 6. 'Information Activist Job Opening - May 2001' from www.igc.org. 7. 'Post an Employee Benefits Job Opening (Help Wanted) Ad' from www.benefitslink.com. 8. Another 'Post an Employee Benefits Job Opening (Help Wanted) Ad' from www.benefitslink.com. On the right side of the page, three orange callout boxes are positioned next to the search results. The top box points to the third result and contains the text 'Martin Baker, a person'. The middle box points to the fifth result and contains the text 'Genomics job'. The bottom box points to the seventh result and contains the text 'Employers job posting form'.

Advanced Search Preferences Language Tools Search Tips

Google™ baker job opening Google Search

Web Images Groups Directory News-New! Results

Searched the web for **baker job opening**.

[Job Opening - Find ANY Job! - Search by Type, Industry & Geography](#)  
www.careerbuilder.com Post Your RESUME Here to Reach Thousands of Employers - It's FREE!

[Job Opening At Flipdog.Com](#)  
www.FlipDog.com Fetch your next **job** at FlipDog.com!

[Softimage::Community::Discussion Groups::ds.archive.0004](#)  
... Le Rudulier; Drive space Ken Skaggs; Help about rendering denis.courtot; **JOB OPENING** ... Tony Cacciarelli; RE: ALE Karim Arbaoui; RE: omf to timeline Martin **Baker**; Re ...  
www.softimage.com/community/xsi/discuss/Archives/ds.archive.0004/default.htm - 49k - Cached - Similar pages

[Softimage::Community::Discussion Groups::ds.archive.0004](#)  
... Re: **JOB OPENING** Philip Herring - 2000/04/28 22:35. ... RE: omf to timeline Martin **Baker** - 2000/04/26 17:33; Re: omf to timeline adam - 2000/04/26 18:11. ...  
www.softimage.com/community/xsi/discuss/Archives/ds.archive.0004/ThreadIndex.htm - 50k - Cached - Similar pages  
[ More results from www.softimage.com ]

[CGI: Job Opening](#)  
www.genomics.cornell.edu/jobs/view\_job.cfm?id=10 - 15k - Cached - Similar pages

[Information Activist Job Opening - May 2001](#)  
www.igc.org/datacenter/job.html - 6k - Cached - Similar pages

[Post an Employee Benefits Job Opening \(Help Wanted\) Ad](#)  
... edit the ad to add a new **job opening** ... as possible when it is emailed to 2,985 **job** ... jobs/posthelpwanted.shtml  
- Webmaster: webmaster@BenefitsLink.com (Dave **Baker** ...  
www.benefitslink.com/jobs/posthelpwanted.shtml - 24k - Cached - Similar pages

[Post an Employee Benefits Job Opening \(Help Wanted\) Ad](#)  
Employee Benefits Jobs! Brought to you by BenefitsLink (tm) and its EmployeeBenefitsJobs.com (tm) division.  
www.benefitslink.com/jobs/pricinginfo.shtml - 7k - Cached - Similar pages  
[ More results from www.benefitslink.com ]

*Martin Baker, a person*

*Genomics job*

*Employers job posting form*

# Example: A Solution

job search find employment careers @ FlipDog.com free! - Microsoft Internet Explorer

Address <http://www.flipdog.com/home.html> Go File Edit View Favorites Tools Help Links

**FlipDog.com**

Home Find Jobs Your Account Resource Center Support Employers

Job Search at FlipDog.com: Employment & Career Management

**647,514**  
Job Opportunities  
from **53,641** Employers

**Find a Job!**

**Post Your Resume**

**Employers**  
click here for  
Products & Services

**Pigskin Places**

- Health Care in NY [2,770](#)
- Health Care in MD [1,262](#)
- Sales in NY [3,751](#)
- Sales in MD [958](#)
- Computing in NY [8,050](#)
- Computing in MD [4,114](#)

**Jobs for Sports Fans**

- [Head Football Coach](#)
- [Football Coach](#)
- [Asst. Football Coach](#)
- [High School Football Coach](#)
- [Univ. Asst. Football Coach](#)

**Showcase Jobs**

**MRI**  
Management Recruiters  
of Charlotte North

We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems.

[Learn More](#)

**Job Seeker Newsletter**

Enter your e-mail address:

[Sign Me Up!](#)

**Job Seekers: Find your dream job!**

- Check our 'Best Places to Find a Job' [January report](#).
- Open your [FREE account](#) and put your [resume online](#).
- Search 24x7 with our FREE automatic [JobHunters™](#).
- Research our database of over [50,000 employers](#).
- Get [expert advice](#) at our new [Resource Center](#).
- Access [salary surveys/calculators](#), [relocation tools](#), [networking opportunities](#), & [training/testing](#) tools.
- Use FlipDog.com to search jobs right from your desktop! Download [Snippets](#) today!

**PC** "Top 100 Web Sites"  
PC Magazine, Nov. 2000

**Media Metrix** "Top 10 Career Web Site"  
Media Metrix, Sept. 2000

**powered by WhizBang!**

**Top 10 Job Site**

Internet 12:12 AM

# Extracting Job Openings from the Web

OPUS International, Inc., an executive search firm focusing on the Food Science industry. - Microsoft Internet Explorer

File Edit View Favorites

Back Forward Stop

Address [http://www.foodscience.com/jobs\\_midwest.html#top](http://www.foodscience.com/jobs_midwest.html#top)

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

OPUS INTERNATIONAL INC.

About | Staff | Job

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorite

Address [http://www.foodscience.com/jobs\\_midwest.html#top](http://www.foodscience.com/jobs_midwest.html#top)

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

**Test Kitchen-Consumer Food Relations**

Major food manufacturer in Chicago area seeks a creative food professional to write recipes. Will make presentations; will be a key player in a cross-functional team. Requires a BS in human ecology, nutrition, Food Science, or related field with a minimum three years' and experience.

Contact: Moira: [e-mail](mailto:email)  
1-800-488-2611

**Ice Cream Guru**

If you dream of cold creamy chocolate or gooey, ooey cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.

Contact: Susan: [e-mail](mailto:email)  
1-800-488-2611

foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: [www.foodscience.com/jobs\\_midwest.htm](http://www.foodscience.com/jobs_midwest.htm)

OtherCompanyJobs: foodscience.com-Job1



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

**FlipDog.com** Fetch Your Next Job Here™

Home Find Jobs Your Account Resource Center

Return to Results | Modify Search | New Search

**The University Alliance** A RISK EDUCATION NETWORK Degrees Online  
Learn While You Earn **MBA, BA, AA** Degrees Online & **Project Mgt.**

Click here to e-mail your resume to 1000's of Head Hunters with **ResumeZapper.com**

Breakthrough ebook shows why most people are **WRONG** about how to apply for jobs.

1 - 25 of 47 jobs shown below 1 2 Next >

Search these results for:  GO! [Search tips](#) **Show Jobs Posted:**  For all time periods

View: [Brief](#) | [Detailed](#)

**Web Jobs:** FlipDog technology has found these jobs on thousands of employer Web sites.

<a href="#">Food Pantry Workers</a> at <a href="#">Lutheran Social Services</a>	October 11, 2002	<a href="#">Archbold, OH</a>
<a href="#">Cooks</a> at <a href="#">Lutheran Social Services</a>	October 11, 2002	<a href="#">Archbold, OH</a>
<a href="#">Bakers Assistants</a> at <a href="#">Fine Catering by Russell Morin</a>	October 11, 2002	<a href="#">Attleboro, MA</a>
<a href="#">Baker's Helper</a> at <a href="#">Bird-in-Hand</a>	October 11, 2002	United States
<a href="#">Assistant Baker</a> at <a href="#">Gourmet To Go</a>	October 11, 2002	<a href="#">Maryland Heights, MO</a>
<a href="#">Host/Hostess</a> at <a href="#">Sharis Restaurants</a>	October 10, 2002	<a href="#">Beaverton, OR</a>
<a href="#">Cooks</a> at <a href="#">Alta's Rustler Lodge</a>	October 10, 2002	<a href="#">Alta, UT</a>
<a href="#">Line Attendant</a> at <a href="#">Sun Valley Coporation</a>	October 10, 2002	<a href="#">Huntsville, UT</a>
<a href="#">Food Service Worker II</a> at <a href="#">Garden Grove Unified School District</a>	October 10, 2002	<a href="#">Garden Grove, CA</a>
<a href="#">Night Cook / Baker</a> at <a href="#">SONOCO</a>	October 10, 2002	<a href="#">Houma, LA</a>
<a href="#">Cooks/Prep Cooks</a> at <a href="#">GrandView Lodge</a>	October 10, 2002	<a href="#">Nisswa, MN</a>
<a href="#">Line Cook</a> at <a href="#">Lone Mountain Ranch</a>	October 10, 2002	<a href="#">Big Sky, MT</a>
<a href="#">Production Baker</a> at <a href="#">Whole Foods Market</a>	October 08, 2002	<a href="#">Willowbrook, IL</a>
<a href="#">Cake Decorator/Baker</a> at <a href="#">Mandalay Bay Hotel and Casino</a>	October 08, 2002	<a href="#">Las Vegas, NV</a>
<a href="#">Shift Supervisors</a> at <a href="#">Brueggers Bagels</a>	October 08, 2002	<a href="#">Minneapolis, MN</a>

# Data Mining the Extracted Job Information

Job Opportunity Index - Microsoft Internet Explorer provided by WhizBang! Labs

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print

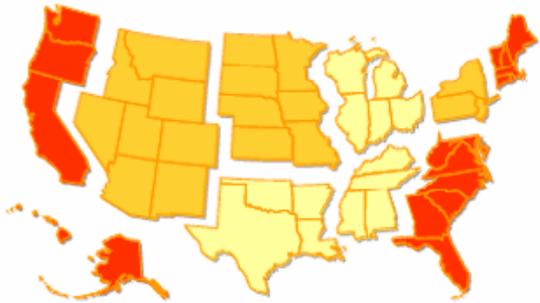
Address <http://joi.flipdog.com/joi/> Links >>

HOME SUBSCRIBE DOWNLOAD ABOUT JOI

FlipDog.com  
Job Opportunity Index®

November 2001

Welcome -- Tuesday, May 7, 2002



**U.S. JOB SUPPLY BY REGION**

- Above Average
- Average
- Below Average

**UNITED STATES**

**November 2001 JOI: 28.4** (October: 27.7)  
September Unemployment Rate: 5.4% (August: 4.9%)

**Click on a region to see individual reports.**

[See printable version](#)

**U.S. Job Supply Increases Amid Rising Unemployment**

The Job Opportunity Index™ (JOI) increased for the first time in three months in October – climbing 0.7 point to 28.4 and signifying a slight increase in U.S. job supply. However, numerous factors, including a dramatic half-point increase in the national unemployment rate, made October anything but normal.

[Subscribe now](#)

**Special Offer!** Find out how you can earn a free subscription to the *JOI Report on U.S. Labor Markets* through a limited-time JOI Subscriber Referral Program!

Done Internet

# What is “Information Extraction”

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

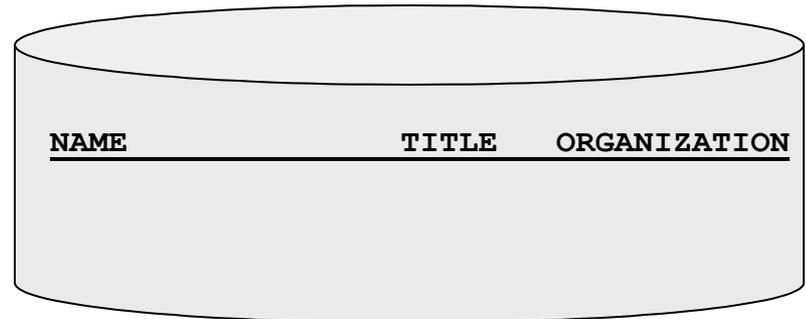
October 14, 2002, 4:00 a.m. PT

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.

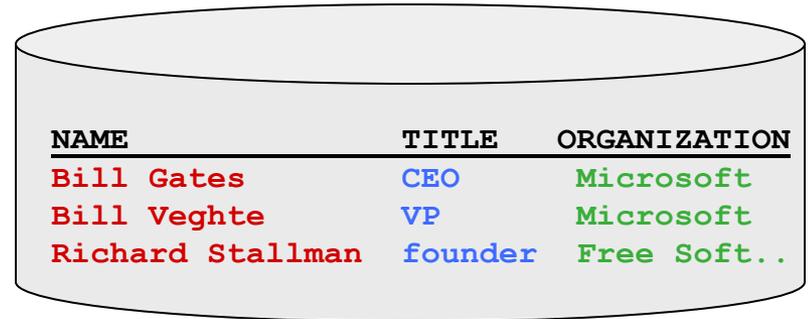
October 14, 2002, 4:00 a.m. PT

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...



<u>NAME</u>	<u>TITLE</u>	<u>ORGANIZATION</u>
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

# 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

# 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](#)

# 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](#)

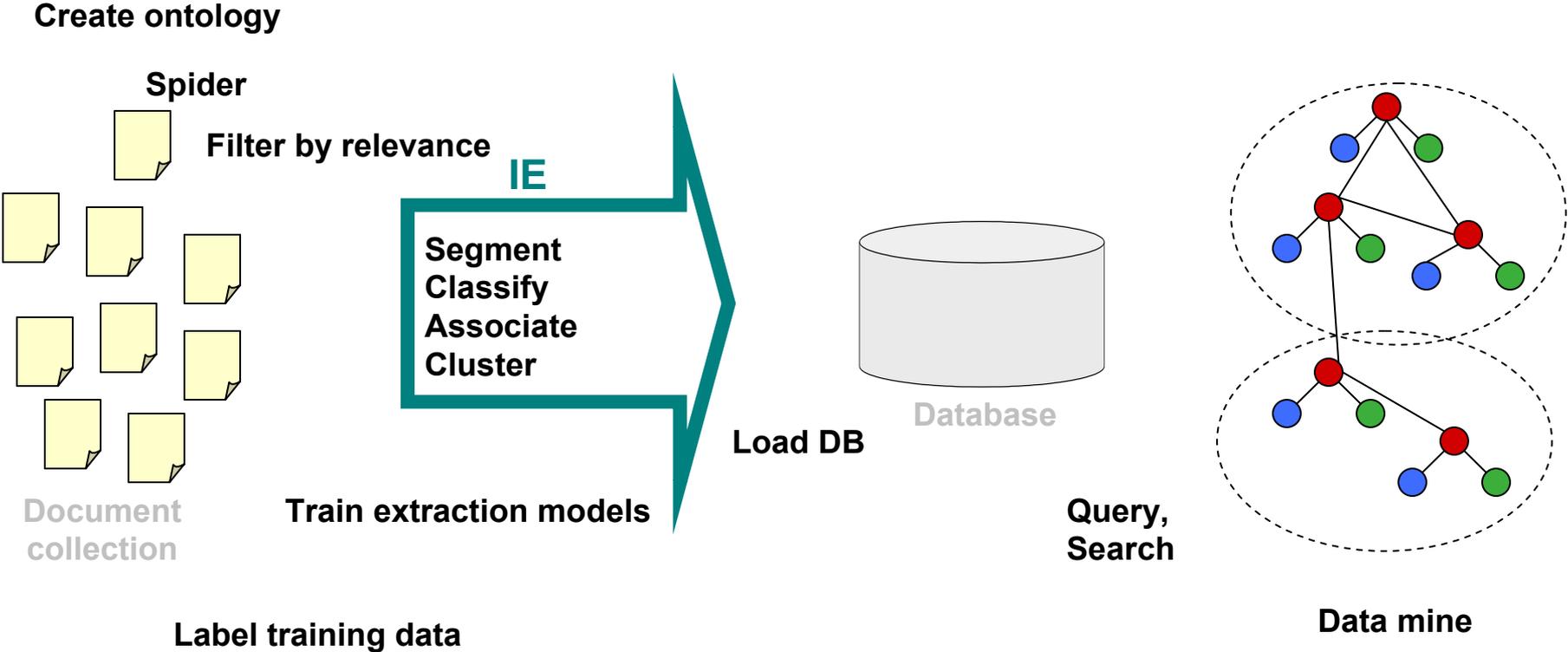
\* [Microsoft](#)  
[Bill Veghte](#)

\* [Microsoft](#)  
[VP](#)

[Richard Stallman](#)  
[founder](#)  
[Free Software Foundation](#)

NAME	TITLE	ORGANIZATION
<a href="#">Bill Gates</a>	<a href="#">CEO</a>	<a href="#">Microsoft</a>
<a href="#">Bill Veghte</a>	<a href="#">VP</a>	<a href="#">Microsoft</a>
<a href="#">Richard Stallman</a>	<a href="#">founder</a>	<a href="#">Free Soft...</a>

# IE in Context



# Why IE from the Web?

- Science
  - Grand old dream of AI: Build large KB\* and reason with it. IE from the Web enables the creation of this KB.
  - IE from the Web is a complex problem that inspires new advances in machine learning.
- Profit
  - Many companies interested in leveraging data currently “locked in unstructured text on the Web”.
  - Not yet a monopolistic winner in this space.
- Fun!
  - Build tools that we researchers like to use ourselves: Cora & CiteSeer, MRQE.com, FAQFinder,...
  - See our work get used by the general public.

\* KB = “Knowledge Base”

# Tutorial Outline

- IE History
- Landscape of problems and solutions
- Parade of models for segmenting/classifying:
  - Sliding window
  - Boundary finding
  - Finite state machines
  - Trees
- Overview of related problems and solutions
- Where to go from here

# IE History

## Pre-Web

- Mostly news articles
  - De Jong's *FRUMP* [1982]
    - Hand-built system to fill Schank-style “scripts” from news wire
  - *Message Understanding Conference (MUC)* DARPA ['87-'95], *TIPSTER* ['92-'96]
- Most early work dominated by hand-built models
  - E.g. SRI's *FASTUS*, hand-built FSMs.
  - But by 1990's, some machine learning: Lehnert, Cardie, Grishman and then HMMs: Elkan [Leek '97], BBN [Bikel et al '98]

## Web

- AAI '94 Spring Symposium on “Software Agents”
  - Much discussion of ML applied to Web. Maes, Mitchell, Etzioni.
- Tom Mitchell's WebKB, '96
  - Build KB's from the Web.
- Wrapper Induction
  - Initially hand-build, then ML: [Soderland '96], [Kushmeric '97],...

# What makes IE from the Web Different?

Less grammar, but more formatting & linking

## Newsire

### 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."

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

## Web

www.apple.com/retail

Coming Soon

[Millenia](#)  
Orlando, FL  
Grand Opening, October 19

Now Open

Arizona <a href="#">Chandler Fashion Center</a> Chandler	Florida <a href="#">The Falls</a> Miami	New York <a href="#">Crossgates</a> Albany
<a href="#">Biltmore</a> Phoenix	<a href="#">Wellington Green</a> Wellington	<a href="#">Palisades</a> West Nyack
	<a href="#">Roosevelt Field</a> Garden City	

In the News

[Jaguar Launch Event](#)  
All across the country, thousands of people came to Apple Stores for the nighttime Jaguar launch, lining up in anticipation of the release of Mac OS X v10.2. See what they wore and what they did on this special evening.

[Grand Opening at the Grove](#)  
See pictures from the grand opening weekend of The Grove, the new Apple store in Los Angeles.

www.apple.com/retail/soho

you to digital cameras, music, email and the Internet. Join us Saturday mornings for a free Getting Started Workshop for new Mac owners.

[Theater Events](#)

Address:  
SoHo  
103 Prince Street  
New York, NY 10012  
212-226-3126

Store Hours:  
Monday - Saturday  
10 a.m. to 8 p.m.  
Sunday  
11 a.m. to 6 p.m.

www.apple.com/retail/soho/theatre.html

Made on a Mac

Presentation	Presented By	Date	Time
Andy Milburn Filmmaker	Apple	Wed Oct 16	6:30 p.m.
Jean Miele Landscape Photographer	Apple	Thu Oct 17	6:30 p.m.
William Levin Cartoon Animator	Apple	Mon Oct 21	6:30 p.m.
David Chalk Photographer, Illustrator and Animator	Apple	Thu Oct 24	6:30 p.m.
Day in the Life of Africa David Cohen-Publisher David Turnley-Photographer Douglas Kirkland-Photographer	Apple	Thu Oct 29	6:30 p.m.

Theater

Presentation	Presented By	Date	Time
Getting Started on a Mac -Introduction and Basics -Advanced	Apple	Every Sat	9 a.m. 10 a.m.
Mac OS X v10.2 Jaguar Workshop	Apple	Every Sun	11:00 a.m.

In the News

**Made on a Mac**  
Eli Morgan Gesner,  
Creative Director  
Friday, Oct. 11  
6:30 p.m.

**Andy Milburn**  
Andy Milburn of the filmmaking partnership tomandandy discusses their groundbreaking audio technology called Q MIX. October 16, 6:30 p.m.

**Jean Miele**  
New York photographer Jean Miele discusses how he creates his large-scale black-and-white landscape photographs using his Power Mac G4, iBook, and three other Mac computers as replacements for the traditional darkroom. October 17, 6:30 p.m.

**William Levin**  
William "Macboy" Levin presents his animated Flash

# Landscape of IE Tasks (1/4): Pattern Feature Domain

## 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

**Dr. Steven Minton** - Founder/CTO  
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

- Press
- **Contact**
- General information
- Directions maps

**Frank Huybrechts** - COO  
Mr. Huybrechts has over 20 years of

## Non-grammatical snippets, rich formatting & links

## Tables

<b>Barto, Andrew G.</b> Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.	(413) 545-2109	<a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a>	CS276
<b>Berger, Emery D.</b> Assistant Professor.	(413) 577-4211	<a href="mailto:emery@cs.umass.edu">emery@cs.umass.edu</a>	CS344
<b>Brock, Oliver</b> Assistant Professor.	(413) 577-0334	<a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a>	CS246
<b>Clarke, Lori A.</b> Professor. Software verification, testing, and analysis; software architecture and design.	(413) 545-1328	<a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a>	CS304
<b>Cohen, Paul R.</b> Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.	(413) 545-3638	<a href="mailto:cohen@cs.umass.edu">cohen@cs.umass.edu</a>	CS278

8:30 - 9:30 AM	<b>Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty</b> <i>Joseph Y. Halpern, Cornell University</i>				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
<b>Cognitive Robotics</b>	<b>Logic Programming</b>	<b>Natural Language Generation</b>	<b>Complexity Analysis</b>	<b>Neural Networks</b>	<b>Games</b>
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Problem Solving through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van Nuffelen</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, Thomas Eiter, and Georg Gottlob</i>	179: Knowledge Extraction and Comparison from Local Function Networks <i>Kenneth McGarry, Stefan Wermter, and John MacIntyre</i>	71: Iterative Widening <i>Tristan Cazenave</i>
549: Online-Execution of ccGolog Plans <i>Henrik Grosskreutz</i>	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

# Landscape of IE Tasks (2/4): Pattern Scope

## Web site specific

### Formatting

### Amazon.com Book Pages

## Genre specific

### Layout

### Resumes

## Wide, non-specific

### Language

### University Names

amazon.com. VIEW CART

WELCOME YOUR STORE BOOKS ELECTRONICS DVD TOYS & GAMES

SEARCH BROWSE SUBJECTS

Get \$5 off

Machine Learning by Tom M. Mitchell

LOOK INSIDE! MACHIN LEARNING

NEW Super Saver Shipping FREE

Learning in Graphical Models by Michael Irwin Jordan (Editor)

LOOK INSIDE! Learning in Graphical Models

List Price: \$60.00 Price: \$60.00

This item ships for FREE with Super Saver Shipping

Availability: Usually ships within 2 to 3 days

Used & new from \$20.00

Edition: Paperback | All Editions

See more product details

Great Buy

Buy this book with Probabilistic Reasoning in Intelligent Systems Buy Together Today: \$128.95

Buy both now!

**Jason D. M. Rennie**

Massachusetts Institute of Technology  
MIT AI Lab NE43-733  
200 Technology Sq.  
Cambridge, MA 02139

jrennie@ai.mit.edu  
http://www.ai.mit.edu/people/jrennie  
(617) 253-5339

**Research Interests**

My main interests lie in the automated analysis of data for the purposes of classification, estimation and the acquiring of new knowledge. I have both interests in applying such techniques to real world problems, and in the analysis of existing algorithms and the creation of new ones.

**L. Douglas Baker**

Home Address available upon request  
Wean Hall, 8102

Office Address School of Computer Science  
Carnegie Mellon University  
5000 Forbes Avenue  
Pittsburgh, PA 15213

Office Phone (412) 683-6036  
Home Page http://www.cs.cmu.edu/~ldbapp

**Objective**

A position in a dynamic, highly-skilled applied research and development team using statistical machine learning to solve large-scale, real-world tasks such as Information Retrieval and Text Classification.

**Education**

<b>Carnegie Mellon University</b>	Pittsburgh, PA
Ph.D., Computer Science, in progress	
M.S., Computer Science, 1999	
<b>Technical University of Berlin</b>	Berlin, Germany
Exchange Fellow, 1992-1993	
<b>University of Michigan</b>	Ann Arbor, MI
M.S.E., Computer Science and Engineering, 1994 B.S.E., Computer Engineering, Summa Cum Laude, 1992	

**Research Experience**

<b>Carnegie Mellon University</b>	1994-present
-----------------------------------	--------------

I am currently pursuing my dissertation research: a hierarchical probabilistic model for novelty detection in text. This work is being done as part of the Topic Detection and Tracking project at CMU under the direction of Yimin Yann. The

8:30 - 9:30 AM	<b>Invited Talk: Plausibility Measures: A General Approach</b> <i>Joseph Y. Halpern, Cornell University</i>		
9:30 - 10:00 AM	Coffee Break		
10:00 - 11:30 AM	Technical Paper Sessions:		
<b>Cognitive Robotics</b>	<b>Logic Programming</b>	<b>Natural Language Generation</b>	<b>Complexity Analysis</b>
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Solving Problem through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, W</i>

**Dr. Steven Minton - Founder/CTO**

Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

**Frank Huybrechts - COO**

Mr. Huybrechts has over 20 years of

Press

Contact

- General information
- Directions maps

# Landscape of IE Tasks (3/4): Pattern Complexity

E.g. word patterns:

## Closed set

U.S. states

He was born in Alabama...

The big Wyoming sky...

## 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

## Regular set

U.S. phone numbers

Phone: (413) 545-1323

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

## 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): Pattern Combinations

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*

# Evaluation of Single Entity Extraction

## TRUTH:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

## PRED:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

$$\text{Precision} = \frac{\text{\# correctly predicted segments}}{\text{\# predicted segments}} = \frac{2}{6}$$

$$\text{Recall} = \frac{\text{\# correctly predicted segments}}{\text{\# true segments}} = \frac{2}{4}$$

$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

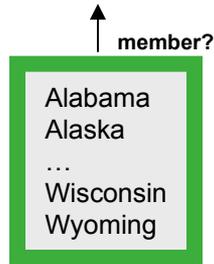
# State of the Art Performance

- Named entity recognition
  - Person, Location, Organization, ...
  - F1 in high 80's or low- to mid-90's
- Binary relation extraction
  - Contained-in (Location1, Location2)  
Member-of (Person1, Organization1)
  - F1 in 60's or 70's or 80's
- Wrapper induction
  - Extremely accurate performance obtainable
  - Human effort (~30min) required on each site

# Landscape of IE Techniques (1/1): Models

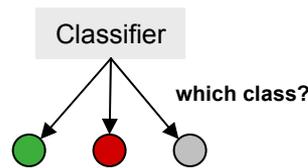
## Lexicons

Abraham Lincoln was born in Kentucky.



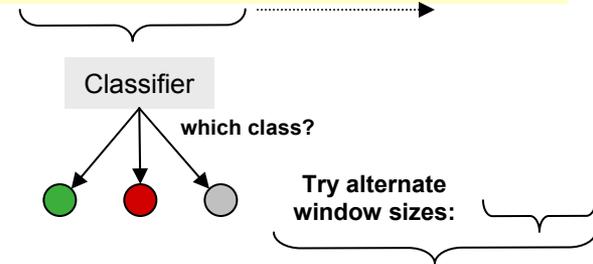
## Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.



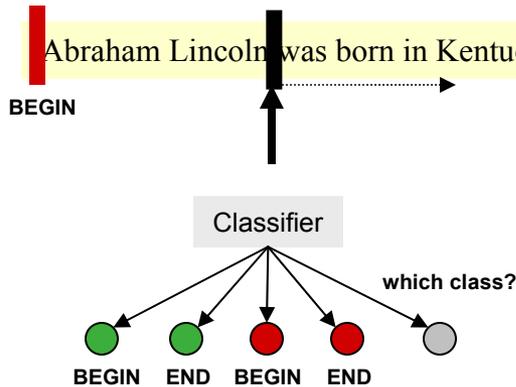
## Sliding Window

Abraham Lincoln was born in Kentucky.



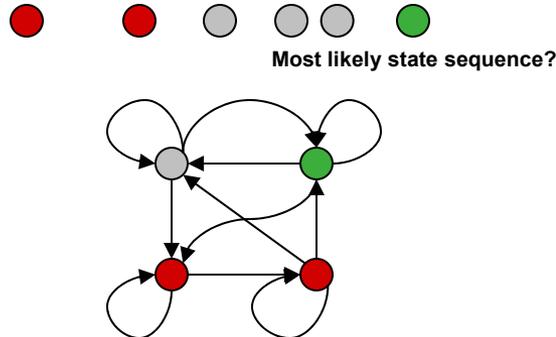
## Boundary Models

Abraham Lincoln was born in Kentucky.



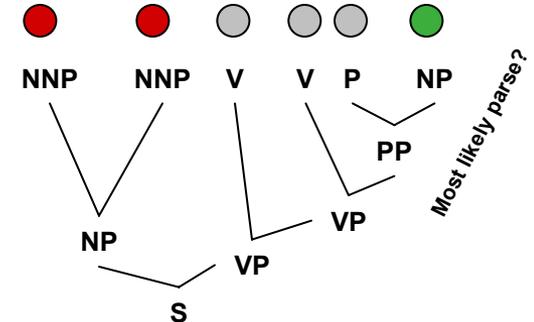
## Finite State Machines

Abraham Lincoln was born in Kentucky.



## Context Free Grammars

Abraham Lincoln was born in Kentucky.



**...and beyond**

Any of these models can be used to capture words, formatting or both.

# Sliding Windows

# Extraction by Sliding Window

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**E.g.  
Looking for  
seminar  
location**

**CMU UseNet Seminar Announcement**

# Extraction by Sliding Window

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**E.g.  
Looking for  
seminar  
location**

**CMU UseNet Seminar Announcement**

# Extraction by Sliding Window

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**E.g.  
Looking for  
seminar  
location**

**CMU UseNet Seminar Announcement**

# Extraction by Sliding Window

**E.g.  
Looking for  
seminar  
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

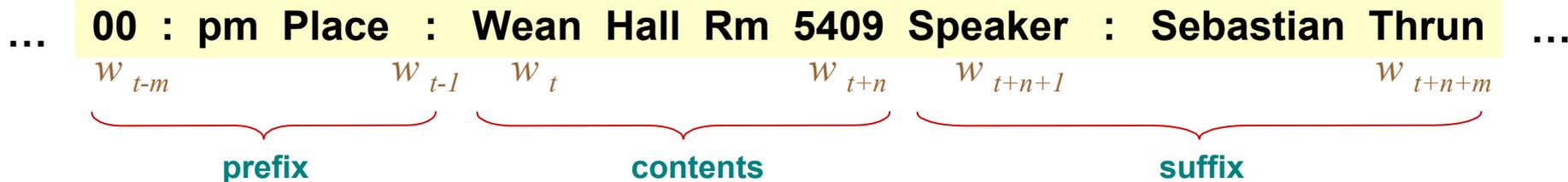


Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**CMU UseNet Seminar Announcement**

# A “Naïve Bayes” Sliding Window Model

[Freitag 1997]



$P(\text{“Wean Hall Rm 5409”} = \text{LOCATION}) =$

$$P(\text{bin}(t) | \theta_{\text{start}}) P(n | \theta_{\text{length}}) \prod_{i=t-m}^{t-1} P(w_i | \theta_{\text{prefix}, i-t}) \prod_{i=t}^{t+n} P(w_i | \theta_{\text{contents}}) \prod_{i=t+n+1}^{t+n+m} P(w_i | \theta_{\text{suffix}, i-t-n})$$

Prior probability  
of start position

Prior probability  
of length

Probability  
prefix words

Probability  
contents words

Probability  
suffix words

Try all start positions and reasonable lengths

Estimate these probabilities by (smoothed)  
counts from labeled training data.

If  $P(\text{“Wean Hall Rm 5409”} = \text{LOCATION})$  is above some threshold, extract it.

Other examples of sliding window: [Baluja et al 2000]  
(decision tree over individual words & their context)

# “Naïve Bayes” Sliding Window Results

## Domain: CMU UseNet Seminar Announcements

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell  
School of Computer Science  
Carnegie Mellon University

3:30 pm  
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

<u>Field</u>	<u>F1</u>
<b>Person Name:</b>	<b>30%</b>
<b>Location:</b>	<b>61%</b>
<b>Start Time:</b>	<b>98%</b>

# Problems with Sliding Windows and Boundary Finders

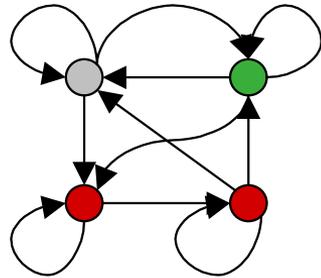
- Decisions in neighboring parts of the input are made independently from each other.
  - Naïve Bayes Sliding Window may predict a “seminar end time” before the “seminar start time”.
  - It is possible for two *overlapping* windows to both be above threshold.
  - In a Boundary-Finding system, left boundaries are laid down independently from right boundaries, and their pairing happens as a separate step.

# Finite State Machines

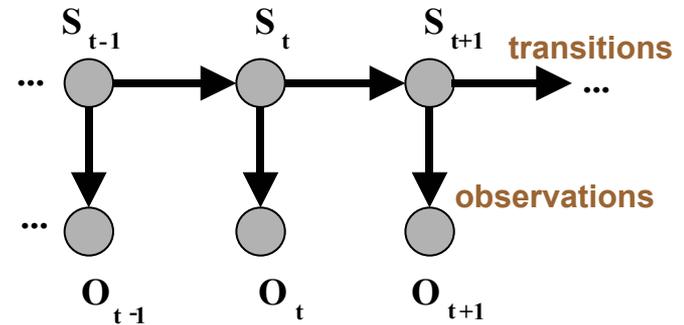
# Hidden Markov Models

HMMs are the standard sequence modeling tool in genomics, music, speech, NLP, ...

Finite state model



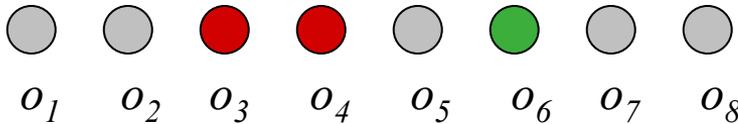
Graphical model



$$P(\vec{s}, \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t | s_{t-1}) P(o_t | s_t)$$

Generates:

State  
sequence  
Observation  
sequence



Parameters: for all states  $S = \{s_1, s_2, \dots\}$

Start state probabilities:  $P(s_t)$

Transition probabilities:  $P(s_t | s_{t-1})$

Observation (emission) probabilities:  $P(o_t | s_t)$  Usually a multinomial over atomic, fixed alphabet

Training:

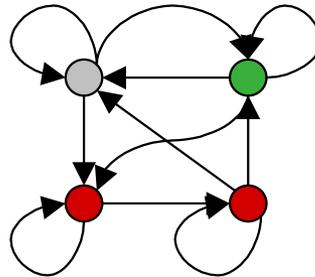
Maximize probability of training observations (w/ prior)

# IE with Hidden Markov Models

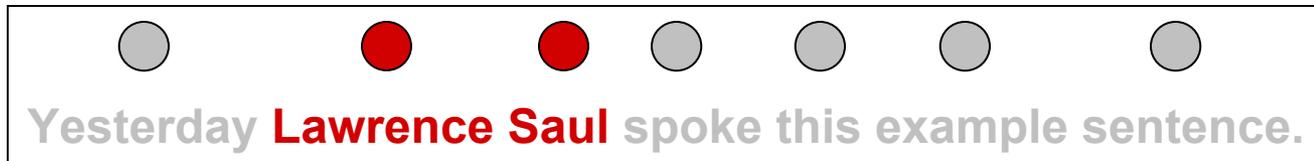
Given a sequence of observations:

Yesterday Lawrence Saul spoke this example sentence.

and a trained HMM:



Find the most likely state sequence: (Viterbi)  $\arg \max_{\bar{s}} P(\bar{s}, \bar{o})$



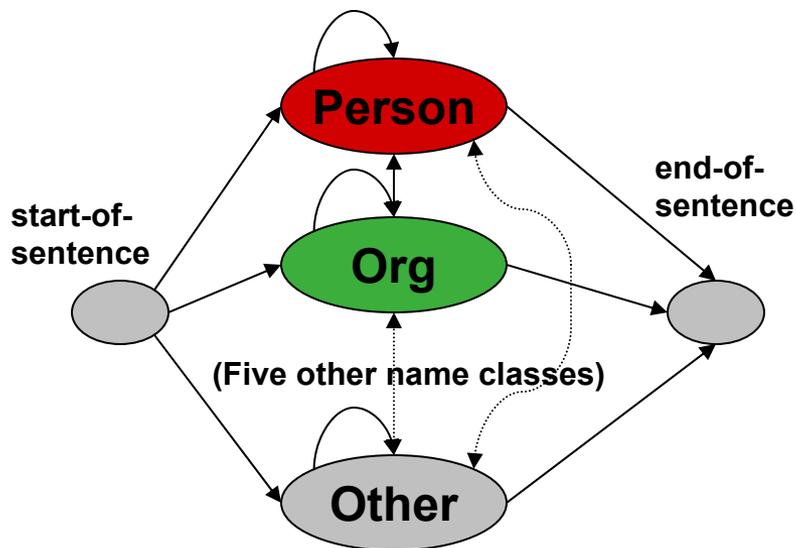
Any words said to be generated by the designated “person name” state extract as a person name:

Person name: Lawrence Saul

# HMM Example: “Nymble”

[Bikel, et al 1998],  
[BBN “IdentiFinder”]

Task: Named Entity Extraction



Train on 450k words of news wire text.

Results:

<u>Case</u>	<u>Language</u>	<u>F1 .</u>
Mixed	English	93%
Upper	English	91%
Mixed	Spanish	90%

# Regrets from Atomic View of Tokens

**Would like richer representation of text:  
multiple overlapping features, whole chunks of text.**

Example word features:

- identity of word
- is in all caps
- ends in “-ski”
- is part of a noun phrase
- is in a list of city names
- is under node X in WordNet or Cyc
- is in bold font
- is in hyperlink anchor
- *features of past & future*
- last person name was female
- next two words are “and Associates”

line, sentence, or paragraph features:

- length
- is centered in page
- percent of non-alphabets
- white-space aligns with next line
- containing sentence has two verbs
- grammatically contains a question
- contains links to “authoritative” pages
- *emissions that are uncountable*
- *features at multiple levels of granularity*

# Problems with Richer Representation and a Generative Model

- These arbitrary features are not independent:
  - Overlapping and long-distance dependences
  - Multiple levels of granularity (words, characters)
  - Multiple modalities (words, formatting, layout)
  - Observations from past and future
- HMMs are *generative* models of the text:  $P(\vec{s}, \vec{o})$
- Generative models do not easily handle these non-independent features. Two choices:
  - **Model the dependencies.** Each state would have its own Bayes Net. But we are already starved for training data!
  - **Ignore the dependencies.** This causes “over-counting” of evidence (ala naïve Bayes). Big problem when combining evidence, as in Viterbi!

# Conditional Sequence Models

- We would prefer a *conditional* model:  
 $P(\bar{s}|\bar{o})$  instead of  $P(\bar{s},\bar{o})$ :
  - Can examine features, but not responsible for generating them.
  - Don't have to explicitly model their dependencies.
  - Don't “waste modeling effort” trying to generate what we are given at test time anyway.
- If successful, this answers the challenge of integrating the ability to handle many arbitrary features with the full power of finite state automata.

# Experimental Data

## 38 files belonging to 7 UseNet FAQs

Example:

```
<head>           X-NNTP-Poster: NewsHound v1.33
<head>           Archive-name: acorn/faq/part2
<head>           Frequency: monthly
<head>
<question>       2.6) What configuration of serial cable should I use?
<answer>
<answer>         Here follows a diagram of the necessary connection
<answer>         programs to work properly.  They are as far as I know
<answer>         agreed upon by commercial comms software developers fo
<answer>
<answer>         Pins 1, 4, and 8 must be connected together inside
<answer>         is to avoid the well known serial port chip bugs.  The
```

**Procedure: For each FAQ, train on one file, test on other; average.**

# Features in Experiments

begins-with-number

begins-with-ordinal

begins-with-punctuation

begins-with-question-word

begins-with-subject

blank

contains-alphanum

contains-bracketed-number

contains-http

contains-non-space

contains-number

contains-pipe

contains-question-mark

contains-question-word

ends-with-question-mark

first-alpha-is-capitalized

indented

indented-1-to-4

indented-5-to-10

more-than-one-third-space

only-punctuation

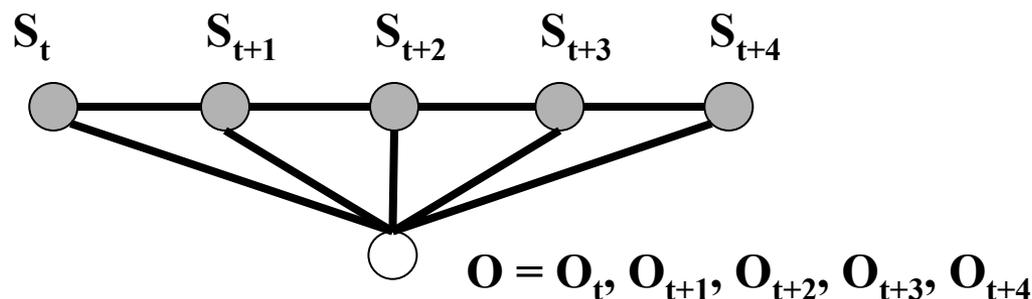
prev-is-blank

prev-begins-with-ordinal

shorter-than-30

# Conditional Random Fields (CRFs)

[Lafferty, McCallum, Pereira '2001]



Markov on  $s$ , conditional dependency on  $o$ .

$$P(\bar{s} | \bar{o}) \propto \frac{1}{Z_{\bar{o}}} \prod_{t=1}^{|\bar{o}|} \exp\left(\sum_k \lambda_k f_k(s_t, s_{t-1}, \bar{o}, t)\right)$$

Hammersley-Clifford-Besag theorem stipulates that the CRF has this form—an exponential function of the cliques in the graph.

Assuming that the dependency structure of the states is tree-shaped (linear chain is a trivial tree), inference can be done by dynamic programming in time  $O(|\bar{o}| |S|^2)$ —just like HMMs.

# General CRFs vs. HMMs

- More general and expressive modeling technique
- Comparable computational efficiency
- Features may be arbitrary functions of *any* or *all* observations
- Parameters need not fully specify generation of observations; require less training data
- Easy to incorporate domain knowledge
- State means only “state of process”, vs “state of process” and “observational history I’m keeping”

# Person name Extraction

[McCallum 2001,  
unpublished]

Press Release 1/18/99 - Microsoft Internet Explorer provided by WhizBang! Labs - [Working Offl...]

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print Copy Paste

Address C:\Documents and Settings\mccallum\Desktop\FromLinux\train30.s5.markedup\alloysilverstein.com Links >>

**GEORGE E. BARRETT, CPA, AWARDED CERTIFICATE OF EDUCATIONAL ACHIEVEMENT IN EMPLOYEE BENEFIT ADMINISTRATION**

Alloy, Silverstein, Shapiro, Adams, Mulford & Co., Cherry Hill, NJ, the 17th largest accounting firm with offices in the Philadelphia area, is pleased to announce that Associate Partner **George E. Barrett, CPA**, a Cherry Hill, NJ resident and 1983 graduate of Rutgers University, has been awarded a certificate of educational achievement in employee benefit administration from the Pennsylvania Institute of Certified Public Accountants. The certificate was awarded in recognition of **Mr. Barrett's** completion of a program which includes a series of seminars and comprehensive examinations.

Alloy, Silverstein, Shapiro, Adams, Mulford, & Co., which celebrates its 40th anniversary in 1999, provides a wide range of services including accounting, auditing, tax, management consulting, financial and estate planning, business valuations, litigation support and information technology.

For more information contact:

**Reynold P. Cicalese, CPA**  
Alloy, Silverstein, Shapiro, Adams, Mulford & Co.  
900 Kings Highway North  
Cherry Hill, NJ 08034-1561  
609.667.4100 extension 133

Done My Computer

# Person name Extraction

November 6&7 - PAWS on the Green Golf Tournament presented by M.A.B Paints - Microsoft Int...

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print Mail News RSS

Address [umanesocietyofbroward.com\nm-http+www.humanesocietyofbroward.com+pawongrengo.html](http://umanesocietyofbroward.com\nm-http+www.humanesocietyofbroward.com+pawongrengo.html) Links >>

After record success last year (more than \$119,000 was raised for the animals) all four co-persons decided to continue in their positions. The chairmen are **Katie Cunningham**, **Marti Huizenga** - HSBC Board Member, **Ursula Kekich** and **Barbara Weintraub**. This year's tournament promises to be even better with a new two-day format brought about by popular demand. Even though it is hoped the event will be dominated by eagles and birdies, it will literally be raining cats and dogs when arriving golfers are greeted by lots of furry friends, many of whom will melt the hearts of potential adopters.

In addition to the hard working Chairwomen of this event, the Committee Members are dedicated to making it a success and they are: **Joy Abbott**, **Meredith Bruder**, **Dianne Davant**, **Liz Ferayorni**, **Ann Gremillion**, **Madelaine Halmos**, **Elaine Heinrich**, **Celia Hogan**, **Paige Hyatt**, **Joanne Johnsen**, **Patty Kearns**, **Karin Kirschbaum**, **Carol McCarvill**, **Kay McFall**, **Annette Penrod**, **Tricia Rutsis**, **Caryl Sorensen**, **Kathie Stephensen** and **Marilyn Stull**.

For the second year, the tournament is presented by M.A.B Paints and sponsored by Cundy Insurance, AutoNation Inc, the Miami Dolphins, American Airlines, Barbara & **Michael Weintraub**, E-Z-Go South Florida, Merrill Lynch, Dianne Davant Interiors, Katz, Barron, Squitero and Faust, P.A.

The \$650 per-player entry fee will support the Humane Society of Broward County's many programs and services including: providing services for more than 20,000 animals each year, educating the community about respect for animals through partnerships with the Boy and Girl Clubs, the Girl Scouts of Broward County and

Done My Computer

# Features in Experiment

Capitalized	Xxxxx	Character n-gram classifier	Hand-built FSM person-name
Mixed Caps	XxXxxx	says string is a person	extractor says yes,
All Caps	XXXXXX	name (80% accurate)	(prec/recall ~ 30/95)
Initial Cap	X....	In stopword list	Conjunctions of all previous
Contains Digit	xxx5	(the, of, their, etc)	feature pairs, evaluated at
All lowercase	xxxx	In honorific list	the current time step.
Initial	X	(Mr, Mrs, Dr, Sen, etc)	Conjunctions of all previous
Punctuation	.,:;!(), etc	In person suffix list	feature pairs, evaluated at
Period	.	(Jr, Sr, PhD, etc)	current step and one step
Comma	,	In name particle list	ahead.
Apostrophe	'	(de, la, van, der, etc)	All previous features, evaluated
Dash	-	In Census lastname list;	two steps ahead.
Preceded by HTML tag		segmented by P(name)	All previous features, evaluated
		In Census firstname list;	one step behind.
		segmented by P(name)	
		In locations lists	
		(states, cities, countries)	
		In company name list	
		("J. C. Penny")	
		In list of company suffixes	
		(Inc, & Associates, Foundation)	

**Total number of features = ~200k**

# Training and Testing

- Trained on 65469 words from 85 pages, 30 different companies' web sites.
- Training takes 4 hours on a 1 GHz Pentium.
- Training precision/recall is 96% / 96%.
  
- Tested on different set of web pages with similar size characteristics.
- Testing precision is 92 – 95%,  
recall is 89 – 91%.

# Chinese Word Segmentation

*[McCallum & Feng,  
to appear]*

- Trained on 800 segmented sentences from UPenn Chinese Treebank.
- Training time: ~2 hours with L-BFGS.
- Training F1: 99.4%
- Testing F1: 99.3%
- Previous top contenders' F1: ~85-95%

# IE Resources

- Data

- RISE, <http://www.isi.edu/~muslea/RISE/index.html>
- Linguistic Data Consortium (LDC)
  - Penn Treebank, Named Entities, Relations, etc.
- <http://www.biostat.wisc.edu/~craven/ie>
- <http://www.cs.umass.edu/~mccallum/data>

- Code

- TextPro, <http://www.ai.sri.com/~appel/TextPro>
- MALLET, <http://www.cs.umass.edu/~mccallum/mallet>

- Both

- <http://www.cis.upenn.edu/~adwait/penntools.html>
- <http://www.cs.umass.edu/~mccallum/ie>

# References

- [Bikel et al 1997] Bikel, D.; Miller, S.; Schwartz, R.; and Weischedel, R. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP'97*, p194-201.
- [Califf & Mooney 1999], Califf, M.E.; Mooney, R.: Relational Learning of Pattern-Match Rules for Information Extraction, in *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*.
- [Cohen, Hurst, Jensen, 2002] Cohen, W.; Hurst, M.; Jensen, L.: A flexible learning system for wrapping tables and lists in HTML documents. *Proceedings of The Eleventh International World Wide Web Conference (WWW-2002)*
- [Cohen, Kautz, McAllester 2000] Cohen, W.; Kautz, H.; McAllester, D.: Hardening soft information sources. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000)*.
- [Cohen, 1998] Cohen, W.: Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity, in *Proceedings of ACM SIGMOD-98*.
- [Cohen, 2000a] Cohen, W.: Data Integration using Similarity Joins and a Word-based Information Representation Language, *ACM Transactions on Information Systems*, 18(3).
- [Cohen, 2000b] Cohen, W. Automatically Extracting Features for Concept Learning from the Web, *Machine Learning: Proceedings of the Seventeenth International Conference (ML-2000)*.
- [Collins & Singer 1999] Collins, M.; and Singer, Y. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
- [De Jong 1982] De Jong, G. An Overview of the FRUMP System. In: Lehnert, W. & Ringle, M. H. (eds), *Strategies for Natural Language Processing*. Lawrence Erlbaum, 1982, 149-176.
- [Freitag 98] Freitag, D: Information extraction from HTML: application of a general machine learning approach, *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*.
- [Freitag, 1999], Freitag, D. *Machine Learning for Information Extraction in Informal Domains*. Ph.D. dissertation, Carnegie Mellon University.
- [Freitag 2000], Freitag, D: Machine Learning for Information Extraction in Informal Domains, *Machine Learning* 39(2/3): 99-101 (2000).
- [Freitag & Kushmerick, 1999] Freitag, D; Kushmerick, D.: Boosted Wrapper Induction. *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*
- [Freitag & McCallum 1999] Freitag, D. and McCallum, A. Information extraction using HMMs and shrinkage. In *Proceedings AAAI-99 Workshop on Machine Learning for Information Extraction*. AAAI Technical Report WS-99-11.
- [Kushmerick, 2000] Kushmerick, N: Wrapper Induction: efficiency and expressiveness, *Artificial Intelligence*, 118(pp 15-68).
- [Lafferty, McCallum & Pereira 2001] Lafferty, J.; McCallum, A.; and Pereira, F., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *In Proceedings of ICML-2001*.
- [Leek 1997] Leek, T. R. *Information extraction using hidden Markov models*. Master's thesis. UC San Diego.
- [McCallum, Freitag & Pereira 2000] McCallum, A.; Freitag, D.; and Pereira, F., Maximum entropy Markov models for information extraction and segmentation, *In Proceedings of ICML-2000*
- [Miller et al 2000] Miller, S.; Fox, H.; Ramshaw, L.; Weischedel, R. A Novel Use of Statistical Parsing to Extract Information from Text. *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, p. 226 - 233.

# References

- [Muslea et al, 1999] Muslea, I.; Minton, S.; Knoblock, C. A.: *A Hierarchical Approach to Wrapper Induction*. Proceedings of Autonomous Agents-99.
- [Muslea et al, 2000] Muslea, I.; Minton, S.; and Knoblock, C. Hierarchical wrapper induction for semistructured information sources. *Journal of Autonomous Agents and Multi-Agent Systems*.
- [Nahm & Mooney, 2000] Nahm, Y.; and Mooney, R. A mutually beneficial integration of data mining and information extraction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 627--632, Austin, TX.
- [Punyakankok & Roth 2001] Punyakankok, V.; and Roth, D. The use of classifiers in sequential inference. *Advances in Neural Information Processing Systems 13*.
- [Ratnaparkhi 1996] Ratnaparkhi, A., A maximum entropy part-of-speech tagger, in *Proc. Empirical Methods in Natural Language Processing Conference*, p133-141.
- [Ray & Craven 2001] Ray, S.; and Craven, M. Representing Sentence Structure in Hidden Markov Models for Information Extraction. *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, WA. Morgan Kaufmann.
- [Soderland 1997]: Soderland, S.: Learning to Extract Text-Based Information from the World Wide Web. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*.
- [Soderland 1999] Soderland, S. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1/3):233-277.