

Information Extraction

Craig Knoblock
University of Southern California

Thanks to Andrew McCallum and William Cohen for overview, sliding windows, and CRF slides. Thanks to Matt Michelson for slides on exploiting reference sets. Thanks to Fabio Ciravegna for slides on LP2.

What is “Information Extraction”

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

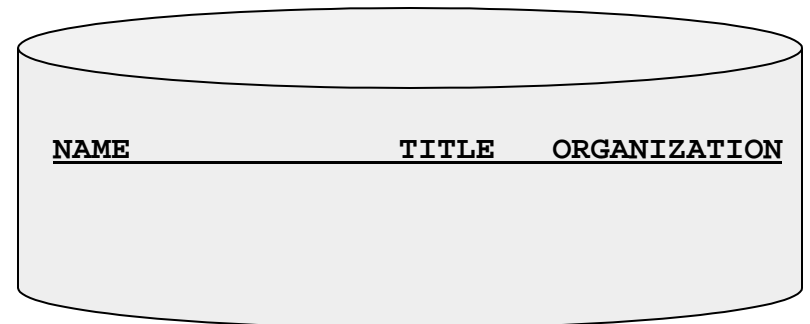
October 14, 2002, 4:00 a.m. PT

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

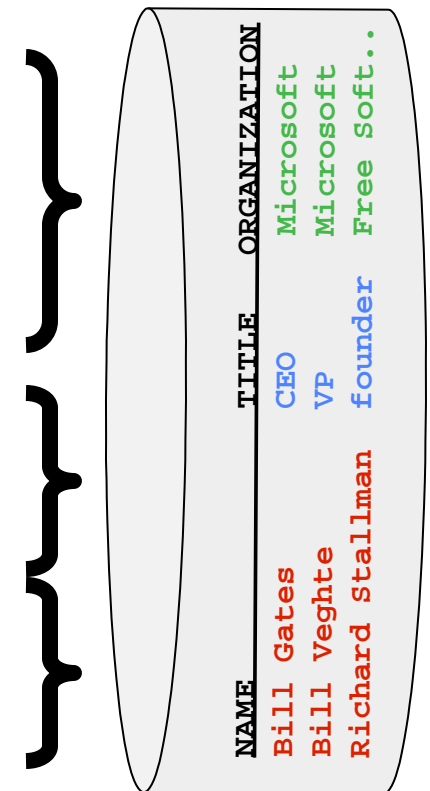
For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

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.

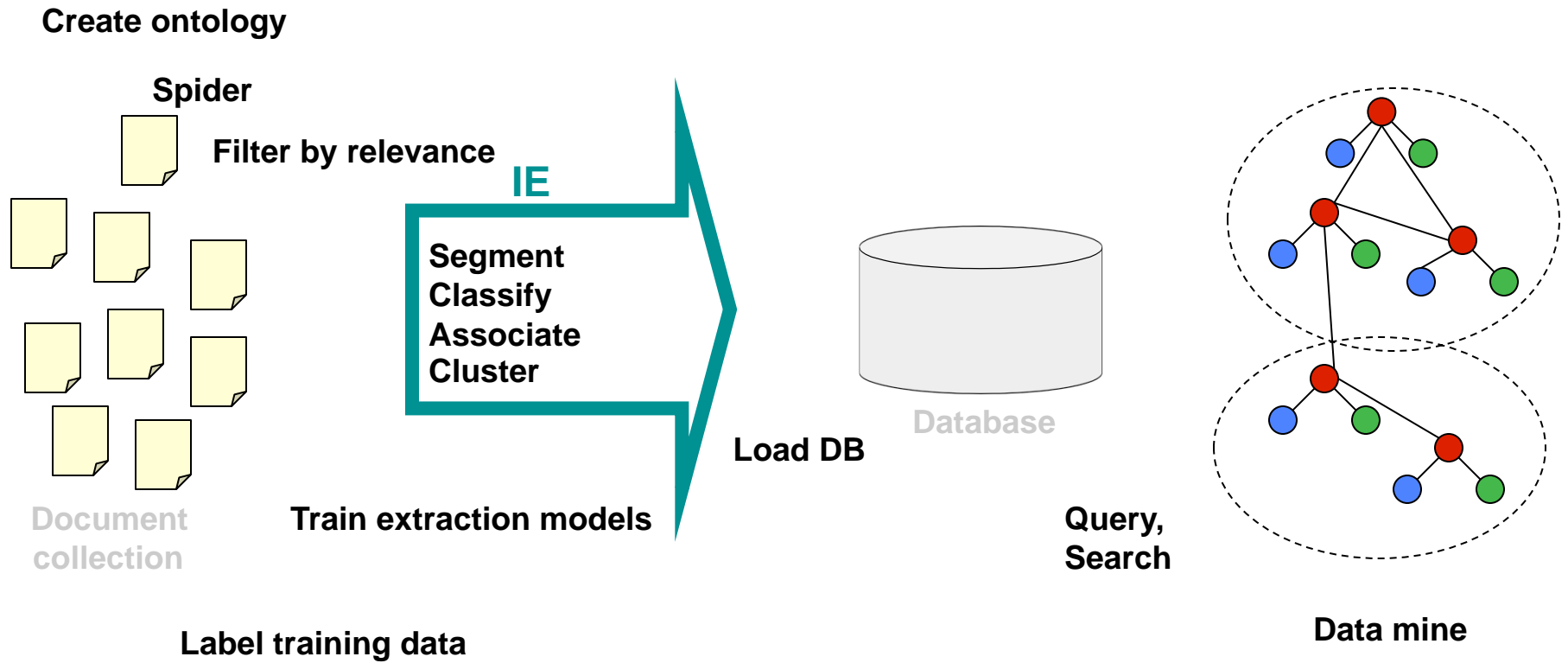
"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[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](#)



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”

Outline

- IE History
- Landscape of problems and solutions
- Models for segmenting/classifying:
 - Lexicons/Reference Sets
 - Sliding window
 - Boundary finding
 - Finite state machines

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 Chandler Fashion Center Chandler	Florida The Falls Miami	New York Crossgates Albany
Biltmore Phoenix	Wellington Green Wellington	Palisades West Nyack
	Roosevelt Field 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 -Introduction and Basics	Apple	Every Sun	11:00 a.m.
Digital Photography Workshop	Apple	Every Sun	3:00 p.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 cartoons and discusses the process of their creation. October 21, 6:45 p.m.

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.

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

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

Non-grammatical snippets, rich formatting & links

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

Tables

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty <i>Joseph Y. Halpern, Cornell University</i>				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis	Neural Networks	Games
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 and Gerhard Lakemeyer</i>	131: A Comparative Study of Logic Programs with Preference <i>Torsten Schaub and Kewen</i>	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation <i>Adnan Darwiche and Pierre Marquis</i>	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing

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

Web site specific

Formatting

Amazon.com Book Pages

The screenshot shows the Amazon.com interface for the book "Learning in Graphical Models" by Michael Irwin Jordan (Editor). The page features a navigation bar with categories like "WELCOME", "YOUR STORE", "BOOKS", "ELECTRONICS", "DVD", and "TOYS & GAMES". A prominent banner advertises "NEW Super Saver Shipping FREE" with a truck icon. The book cover is displayed with a "LOOK INSIDE!" feature. Pricing information shows a list price of \$60.00 and a current price of \$60.00. Availability is noted as "Usually ships within 2 to 3 days". A "Great Buy" section at the bottom suggests buying the book with "Probabilistic Reasoning in Intelligent Systems" for a total price of \$128.95.

Genre specific

Layout

Resumes

The screenshot displays two resumes. The first is for Jason D. M. Rennie, who is affiliated with the Massachusetts Institute of Technology (MIT AI Lab NE43-733). His research interests include the automated analysis of data for classification and knowledge acquisition. The second resume is for L. Douglas Baker, who has worked at Carnegie Mellon University and the Technical University of Berlin. His objective is to work in a dynamic, high-skilled research and development team using statistical machine learning for large-scale, real-world tasks like information retrieval and text classification. His education includes a Ph.D. from Carnegie Mellon University and a M.S. from the Technical University of Berlin.

Wide, non-specific

Language

University Names

The screenshot shows a university schedule and contact information. The schedule includes sessions for "Invited Talk: Plausibility Measures: A General Approach for Rep..." by Joseph Y. Halpern at Cornell University, a "Coffee Break", and "Technical Paper Sessions" in various categories like "Cognitive Robotics", "Logic Programming", "Natural Language Generation", "Complexity Analysis", and "Neural Networks". Below the schedule, there is a "Contact" section for Dr. Steven Minton, Founder/CTO of Fetch. It states that Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. He has worked at USC and Stanford. A "Press" section is also visible.

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 Welch
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

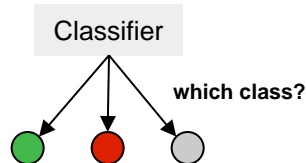
Lexicons

Abraham Lincoln was born in Kentucky.



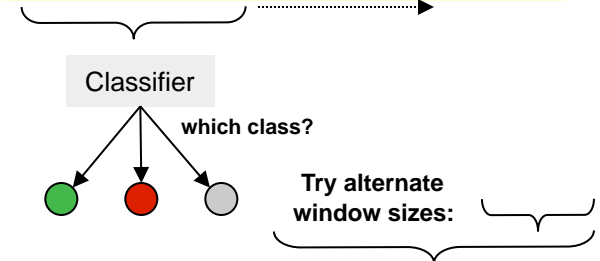
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.



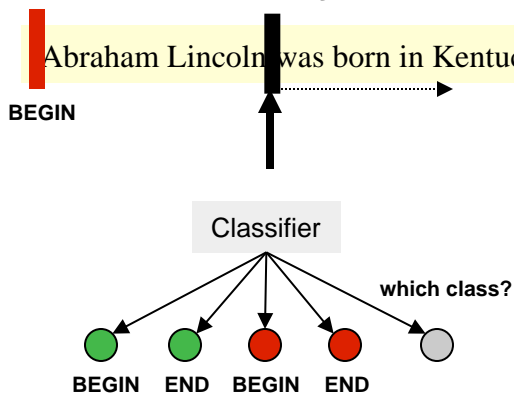
Sliding Window

Abraham Lincoln was born in Kentucky.



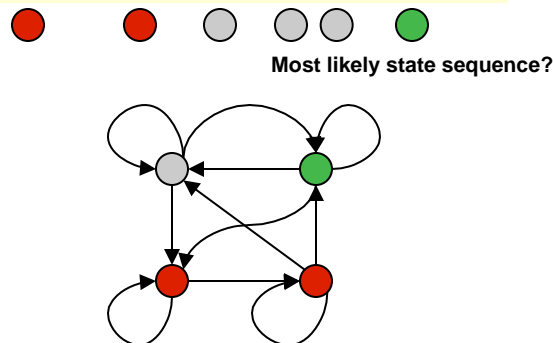
Boundary Models

Abraham Lincoln was born in Kentucky.



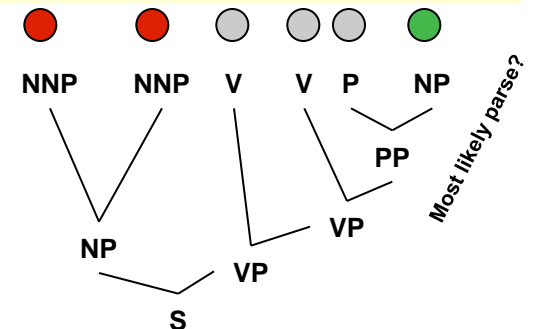
Finite State Machines

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.



Lexicons/Reference Sets

Outline

- Introduction
- Alignment
- Extraction
- Results
- Discussion

Ungrammatical & Unstructured Text

Page 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

	Topic	Replies	Last Comment	Started By
	 SACRAMENTO HOTEL LIST	0	11/21/04 9:56 pm	westcoastman
	3* Rancho Cordova Holiday Inn \$35, 1 nite (12/11)	1	12/9/04 12:37 am	future canadian
	3* Doubletree Sacto Arden 12/11 1 Night \$34	1	12/7/04 4:46 pm	OCTraveler
	4* Sacramento Failed Bid \$85 12/7	1	12/6/04 6:29 pm	Sheryl
	Failed bid Sacramento Downtown 12/6 for 1 night, 4*	13	12/6/04 6:25 pm	emaij
	2.5* Wingate Inn Rancho Cordova 5/10-5/13/05 \$32	0	12/4/04 7:11 pm	ego68
	3* DoubleTree Sacramento \$35 (12/04/04)	0	11/30/04 11:34 pm	shizzolator
	2.5* Rancho Cordova Wingate Inn \$32 (11/23-25)	1	11/27/04 12:19 pm	Profiler
	4* DT Hyatt 11/21 \$60 11/23 \$60; Sheraton Grand 11/25 \$55	0	11/22/04 1:22 pm	bonish
	3* Doubletree Arden/Sacramento \$37 11/19	1	11/20/04 1:53 am	ahallez
	2.5* Wingate Inn Rancho Cordova \$33 11/13	2	11/19/04 1:44 am	cykick42
	2.5* DT Hawthorne Suites \$40 (11/18-20)	0	11/18/04 10:08 pm	Colfax30
	Roseville 2.5*Larkspur \$72(11/22-24) 2* Fairfield \$80(11/24)	2	11/17/04 4:38 pm	mcrinca
	3* Rancho Cordova Holiday Inn \$32 (11/17)	0	11/16/04 10:20 pm	Colfax30
	3* Doubletree Sacramento \$40 (11/11)	2	11/16/04 11:05 am	OCTraveler
	3* Doubletree Sacramento Arden \$36 11/24	0	11/15/04 1:04 am	bomawin
	4* Sheraton Grand Sacramento 12/29 \$65	0	11/15/04 12:08 am	UnixJ
	2.5* Rancho Cordova Wingate Inn \$30 (11/12-17)	2	11/14/04 10:36 pm	Colfax30

Ungrammatical & Unstructured Text

For simplicity → “posts”

Goal:

<hotelArea>univ. ctr.</hotelArea>

Beware 2* at the airport!!!!	2	7/18/00 1:25 am
\$25 winning bid at holiday inn sel univ. ctr.	1	6/26/00 1:48 pm
3* Holiday Inn North-McKnight Rd, \$10+20, 1/19	3	1/27/01 6:34 pm

<price>\$25</price> <hotelName>holiday inn sel.</hotelName>

Wrapper based IE does not apply (e.g. Stalker, RoadRunner)

NLP based IE does not apply (e.g. Rapier)

Reference Sets

IE infused with outside knowledge

“Reference Sets”

- Collections of known entities and the associated attributes
- Online (offline) set of docs
 - CIA World Fact Book
- Online (offline) database
 - Comics Price Guide, Edmunds, etc.
- Build from ontologies on Semantic Web

Comics Price Guide Reference Set

Submit your books online and get **20% off**

CONTACT US
MEDIA KIT
ADMIN LOGIN
AD MANAGE

HOME
GRADING
MESSAGE BOARDS
STORE
CLASSIFIEDS
AUCTIONS
ISSUES SALES
FAQ

Login 131 users

Username:

Password:

Remember Me [Forgot Login](#) [Sign Up](#)

SEARCH BY PUBLISHER

SEARCH BY KEYWORDS

[Marvel](#)
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

FANTASTIC FOUR (1961-1996,2003-CURRENT)

255349 Total Searches

Add To Collection
books you do have

Add To Want List
books you must have

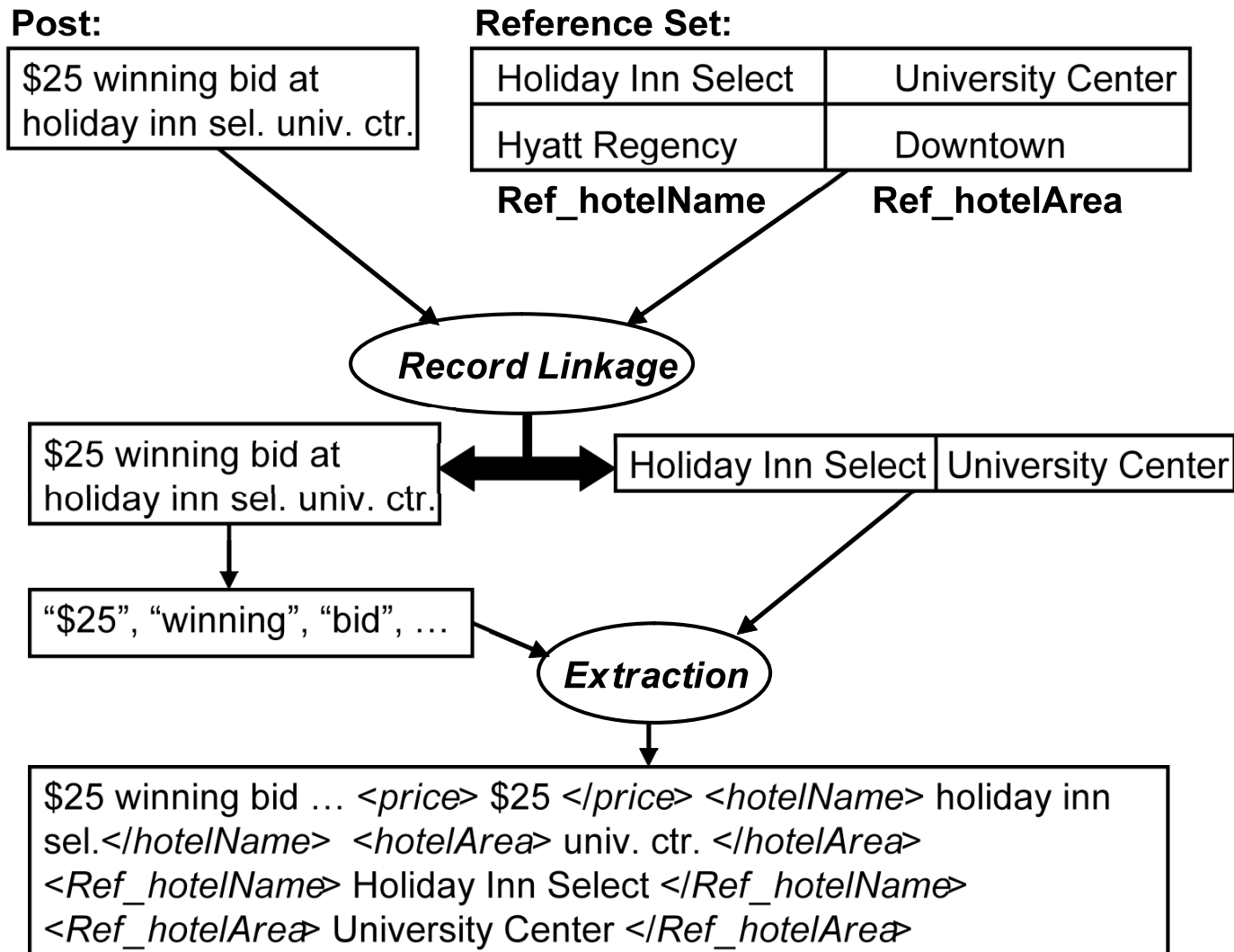
View Collection
see the issues you own

Print This
take home copy

Select All Page 1 2 3 4 5 6

Issue #	9.4 Value	9.4 CGC Graded	For Sale	Cover
<input type="checkbox"/> # 1	<u>\$32,000.00</u>	<u>\$192,000.00</u>		VIEW
First Appearance: Fantastic Four and The Mole Man				
<input type="checkbox"/> # 1A	<u>\$300.00</u>	<u>\$1,800.00</u>	SALE	VIEW
Golden Record Reprint Edition				
<input type="checkbox"/> # 1B	<u>\$200.00</u>	<u>\$1,200.00</u>		VIEW
Comic removed from album				
<input type="checkbox"/> # 2	<u>\$5,250.00</u>	<u>\$31,500.00</u>		VIEW
First Appearance: The Skrulls				
<input type="checkbox"/> # 3	<u>\$3,000.00</u>	<u>\$18,000.00</u>		VIEW
First Fantastic Four Costume				

Algorithm Overview – Use of Ref Sets



Outline

- Introduction
- Alignment
- Extraction
- Results
- Discussion

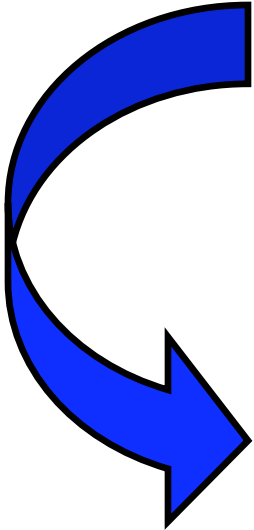
Our Record Linkage Problem

- *Posts not yet decomposed attributes.*
- *Extra tokens that match nothing in Ref Set.*

Post:

“\$25 winning bid at holiday inn sel. univ. ctr.”
hotel name hotel area

Reference Set:



Holiday Inn	Greentree
Holiday Inn Select	University Center
Hyatt Regency	Downtown

hotel name

hotel area

Our Record Linkage Solution

P = "\$25 winning bid at holiday inn sel. univ. ctr."

Record Level Similarity + Field Level Similarities

$$V_{RL} = \langle RL_scores(P, \text{"Hyatt Regency Downtown"}), \\ RL_scores(P, \text{"Hyatt Regency"}), \\ RL_scores(P, \text{"Downtown"}) \rangle$$

Binary Rescoring

SVM

Best matching member of the reference set for the post

Last Alignment Step

Return reference set attributes as annotation for the post

Post:

\$25 winning bid at holiday inn sel. univ. ctr.

<Ref_hotelName>Holiday Inn Select</Ref_hotelName>

<Ref_hotelArea>University Center</Ref_hotelArea>

... more to come in Discussion...

Outline

- Introduction
- Alignment
- Extraction
- Results
- Discussion

Extraction Algorithm

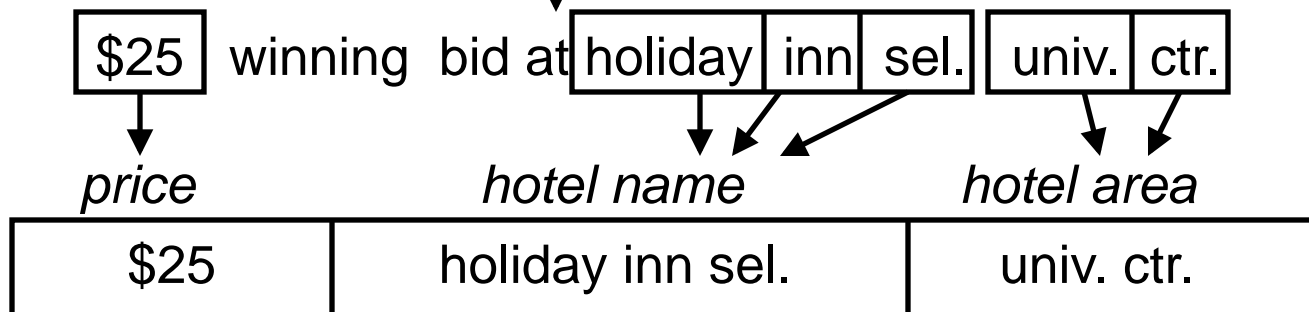
Post:

\$25 winning bid at holiday inn sel. univ. ctr.

Generate V_{IE}

Multiclass SVM

$V_{IE} = \langle \text{common_scores}(\text{token}),$
 $\text{IE_scores}(\text{token}, \text{attr1}),$
 $\text{IE_scores}(\text{token}, \text{attr2}),$
 $\dots \rangle$



Clean Whole Attribute

Common Scores

- Some attributes not in reference set
 - Reliable characteristics
 - Infeasible to represent in reference set
 - E.g. prices, dates
- Can use characteristics to extract/annotate these attributes
 - Regular expressions, for example
- These types of scores are what compose ***common_scores***

Outline

- Introduction
- Alignment
- Extraction
- Results
- Discussion

Experimental Data Sets

Hotels

- ***Posts***

- 1125 posts from www.biddingfortravel.com
 - Pittsburgh, Sacramento, San Diego
 - Star rating, hotel area, hotel name, price, date booked

- ***Reference Set***

- 132 records
- Special posts on BFT site.
 - Per area – list any hotels ever bid on in that area
 - Star rating, hotel area, hotel name

Experimental Data Sets

Comics

- ***Posts***

- 776 posts from EBay
 - “Incredible Hulk” and “Fantastic Four” in comics
 - Title, issue number, price, condition, publisher, publication year, description (1st appearance the Rhino)

- ***Reference Sets***

- 918 comics, 49 condition ratings
- Both come from ComicsPriceGuide.com
 - For FF and IH
 - Title, issue number, description, publisher

Comparison to Existing Systems

Record Linkage

- WHIRL
 - RL allows non-decomposed attributes

Information Extraction

- Simple Tagger (CRF)
 - State-of-the-art IE
- Amilcare
 - NLP based IE

Record linkage results

	Prec.	Recall	F-Measure
Hotel			
Phoebus	93.60	91.79	92.68
WHIRL	83.52	83.61	83.13
Comic			
Phoebus	93.24	84.48	88.64
WHIRL	73.89	81.63	77.57

10 trials – 30% train, 70% test

Token level Extraction results: Hotel domain

		Prec.	Recall	F-Measure	Freq
<i>Area</i>	Phoebus	89.25	87.50	88.28	809.7
	Simple Tagger	92.28	81.24	86.39	
	Amilcare	74.2	78.16	76.04	
<i>Date</i>	Phoebus	87.45	90.62	88.99	751.9
	Simple Tagger	70.23	81.58	75.47	
	Amilcare	93.27	81.74	86.94	
<i>Name</i>	Phoebus	94.23	91.85	93.02	1873.9
	Simple Tagger	93.28	93.82	93.54	
	Amilcare	83.61	90.49	86.90	
<i>Price</i>	Phoebus	98.68	92.58	95.53	850.1
	Simple Tagger	75.93	85.93	80.61	
	Amilcare	89.66	82.68	85.86	
<i>Star</i>	Phoebus	97.94	96.61	97.84	766.4
	Simple Tagger	97.16	97.52	97.34	
	Amilcare	96.50	92.26	94.27	

Not Significant

Token level Extraction results: Comic domain

		Prec.	Recall	F-Measure	Freq
<i>Condition</i>	Phoebus	91.8	84.56	88.01	410.3
	Simple Tagger	78.11	77.76	77.80	
	Amilcare	79.18	67.74	72.80	
<i>Descript.</i>	Phoebus	69.21	51.50	59.00	504.0
	Simple Tagger	62.25	79.85	69.86	
	Amilcare	55.14	58.46	56.39	
<i>Issue</i>	Phoebus	93.73	86.18	89.79	669.9
	Simple Tagger	86.97	85.99	86.43	
	Amilcare	88.58	77.68	82.67	
Price	Phoebus	80.00	60.27	68.46	10.7
	Simple Tagger	84.44	44.24	55.77	
	Amilcare	60.00	34.75	43.54	

Token level Extraction results: Comic domain (cont.)

		Prec.	Recall	F-Measure	Freq
<i>Publisher</i>	Phoebus	83.81	95.08	89.07	61.1
	Simple Tagger	88.54	78.31	82.83	
	Amilcare	90.82	70.48	79.73	
<i>Title</i>	Phoebus	97.06	89.90	93.34	1191.1
	Simple Tagger	97.54	96.63	97.07	
	Amilcare	96.32	93.77	94.98	
Year	Phoebus	98.81	77.60	84.92	120.9
	Simple Tagger	87.07	51.05	64.24	
	Amilcare	86.82	72.47	78.79	

Extraction results: Summary

	Token Level			<i>Hotel</i>	Field Level		
	Prec.	Recall	F-Mes.		Prec.	Recall	F-Mes.
Phoebus	93.60	91.79	92.68		87.44	85.59	86.51
Simple Tagger	86.49	89.13	87.79		79.19	77.23	78.20
Amilcare	86.12	86.14	86.11		85.04	78.94	81.88
	Token Level			<i>Comic</i>	Field Level		
	Prec.	Recall	F-Mes.		Prec.	Recall	F-Mes.
Phoebus	93.24	84.48	88.64		81.73	80.84	81.28
Simple Tagger	84.41	86.04	85.43		78.05	74.02	75.98
Amilcare	87.66	81.22	84.29		90.40	72.56	80.50

Results Discussion

3 attributes where Phoebus not max F-measure

- Hotel name – tiny difference
- Comic Title – low recall → lower F-measure
 - recall: missed tokens of titles not in ref. set
 - “The Incredible Hulk and Wolverine” → “The Incredible Hulk”
- Comic description
 - Simple Tagger learned internal structure of descriptions
 - High recall, low precision
 - Phoebus labels in isolation
 - Only meaningful tokens (like prop. Names) labeled
 - higher precision, lower recall → 2nd best F-measure

Outline

- Introduction
- Alignment
- Extraction
- Results
- Discussion

Summary extraction results

Expensive to label training data...

	Prec.	Recall	F-Mes.	# Train.
Hotel (30%)	93.6	91.79	92.68	338
Hotel (10%)	93.66	90.93	92.27	113
Comic (30%)	93.24	84.48	88.64	233
Comic (10%)	91.41	83.63	87.34	78

Token Level

Hotel (30%)	87.44	85.59	86.51
Hotel (10%)	86.52	84.54	85.52
Comic (30%)	81.73	80.84	81.28
Comic (10%)	79.94	76.71	78.29

Field Level

Reference Set Attributes as Annotation

- Standard query values
- Include info not in post
 - If post leaves out “Star Rating” can still be returned in query on “Star Rating” using ref. set annotation
- Perform better at annotation than extraction
 - Consider Rec. link results as field level extraction
 - E.g. no system did well extracting comic desc.
 - +20% precision, +10% recall using rec. link

Reference Set Attributes as Annotation

Then why do extraction at all?

- Want to see actual values
- Extraction can annotate when record linkage is wrong
 - Better in some cases at annotation than rec. link
 - If wrong rec. link, usually close enough record to get some extraction parts right
- Learn what something is not
 - Helps to classify things not in reference set
 - Learn which tokens to ignore better

Sliding Windows

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

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

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

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

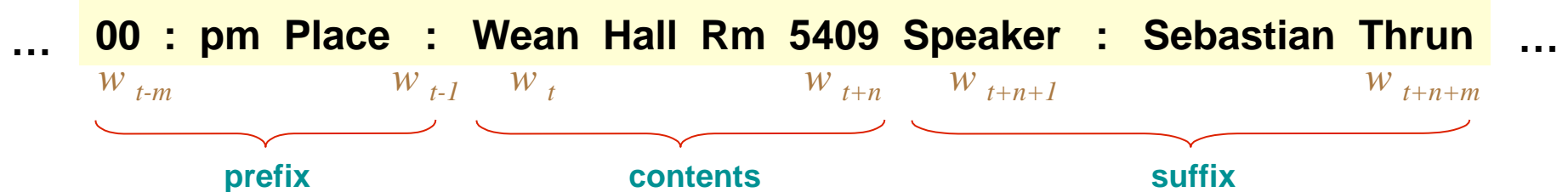
7500 Wean Hall

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>
Person Name:	30%
Location:	61%
Start Time:	98%