

Mayank Kejriwal
University of Southern California

Information Extraction - II

Why IE from the Web?

- Science
 - Build large knowledge base 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 people can use
 - Cora & CiteSeer (papers), MRQE.com (movie reviews), FAQFinder,...

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], [Kushmerick '97],...

What makes IE from the Web Different?

Newswire

Apple to Open Its First Retail Store in New York City

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

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

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.

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

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.

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</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

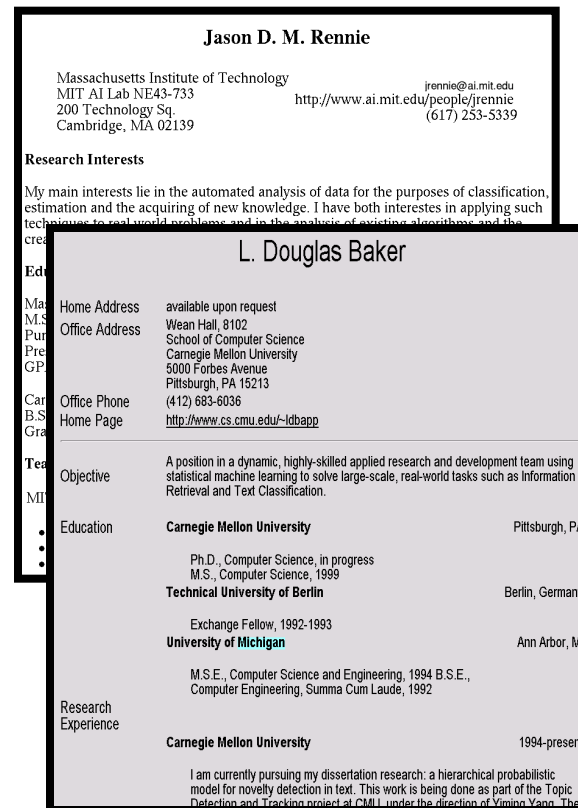
Amazon.com Book Pages



Genre specific

Layout

Resumes



Wide, non-specific

Language

University Names

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</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, McGarry, Stefan Wermter, and</i>	179: Knowledge Extraction and Comparison from Local Function Networks <i>Kenneth McGarry, Stefan Wermter, and</i>	71: Iterative Widening <i>Tristan Cazenave</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.					353: Temporal Difference Learning Applied to a High Performance Game-Playing
Frank Huybrechts - COO Mr. Huybrechts has over 20 years of					ation-Guided Learning for strained mutations in ral-Network e-Series

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.

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

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

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

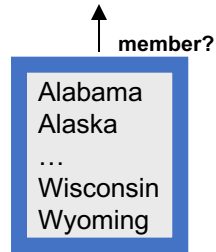
State of the Art Performance

- 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

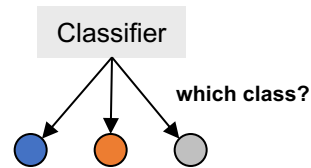
Lexicons

Abraham Lincoln was born in Kentucky.



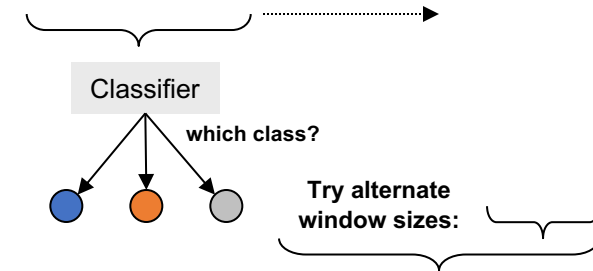
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.



Sliding Window

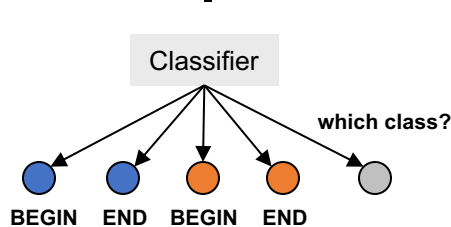
Abraham Lincoln was born in Kentucky.



Boundary Models

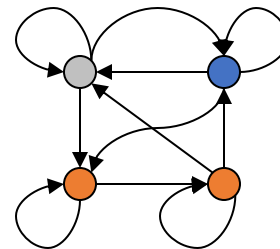
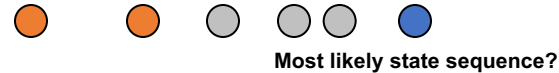
Abraham Lincoln was born in Kentucky.

BEGIN



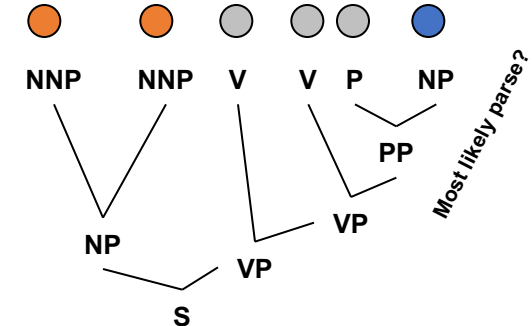
Finite State Machines

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.