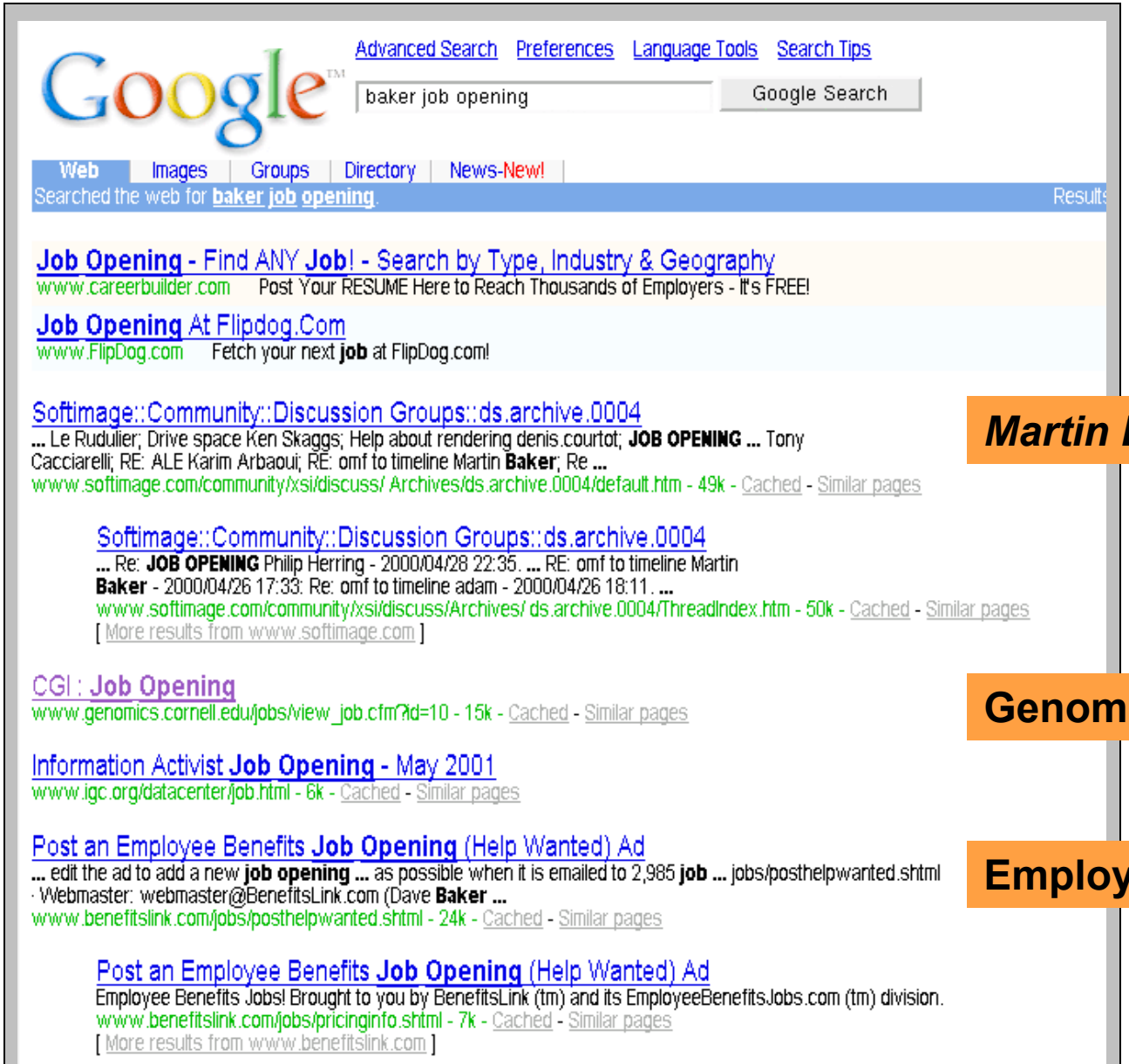
Information Extraction from the World Wide Web

Andrew McCallum
University of Massachusetts Amherst

William Cohen
Carnegie Mellon University

Example: The Problem



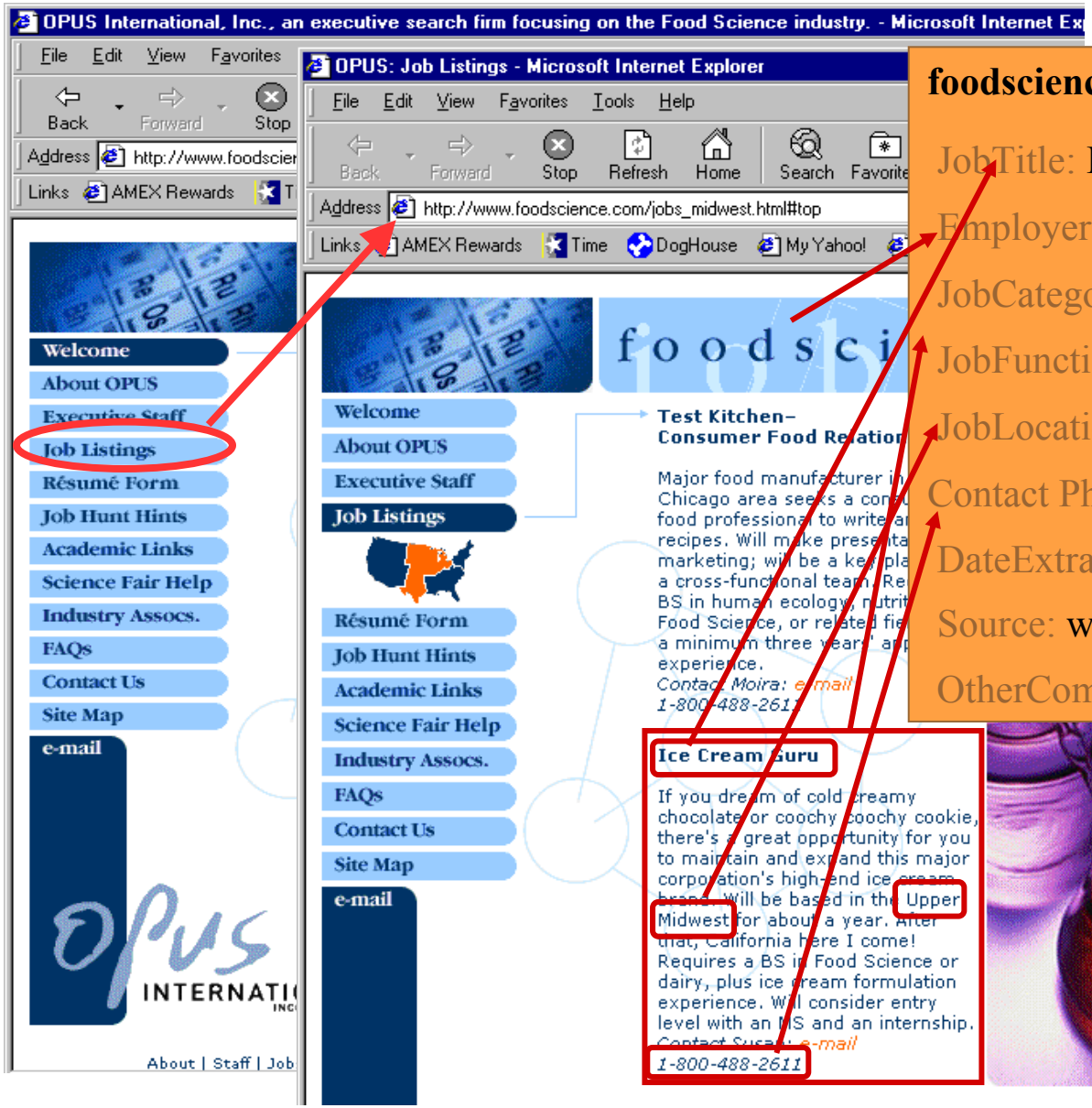
***Martin Baker*, a person**

Genomics job

Employers job posting form

Example: A Solution

Extracting Job Openings from the Web



foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

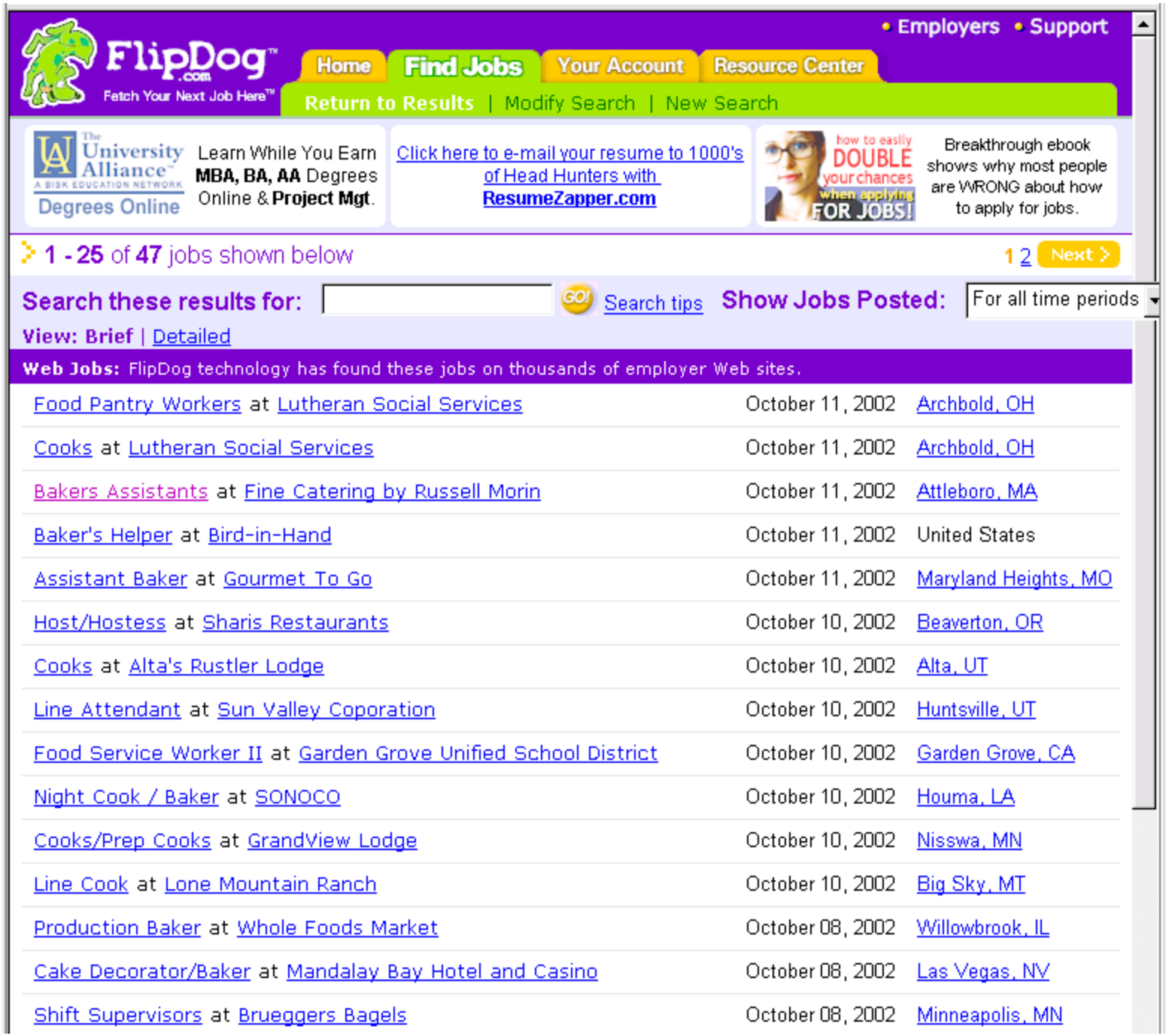
Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.htm

OtherCompanyJobs: foodscience.com-Job1

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



FlipDog.com
Fetch Your Next Job Here™

Employers • Support

Home | Find Jobs | Your Account | Resource Center

Return to Results | Modify Search | New Search

1 - 25 of **47** jobs shown below

1 2 Next

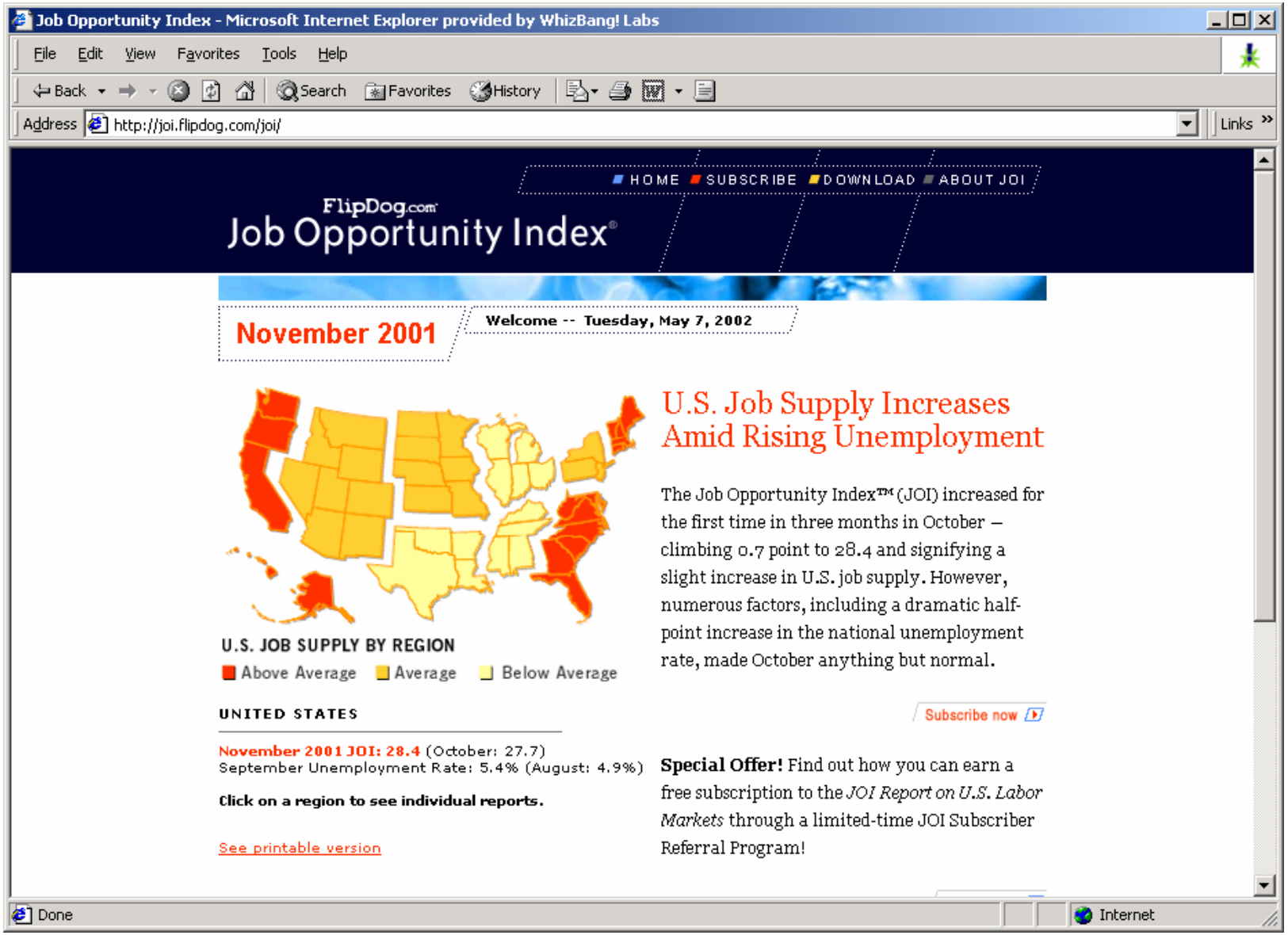
Search these results for: GO! Search tips **Show Jobs Posted:** For all time periods

View: Brief | Detailed

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

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

Data Mining the Extracted Job Information



What is “Information Extraction”

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

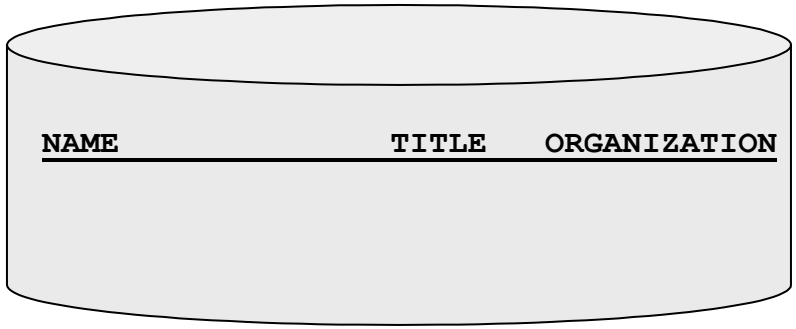
October 14, 2002, 4:00 a.m. PT

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

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

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

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



What is “Information Extraction”

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

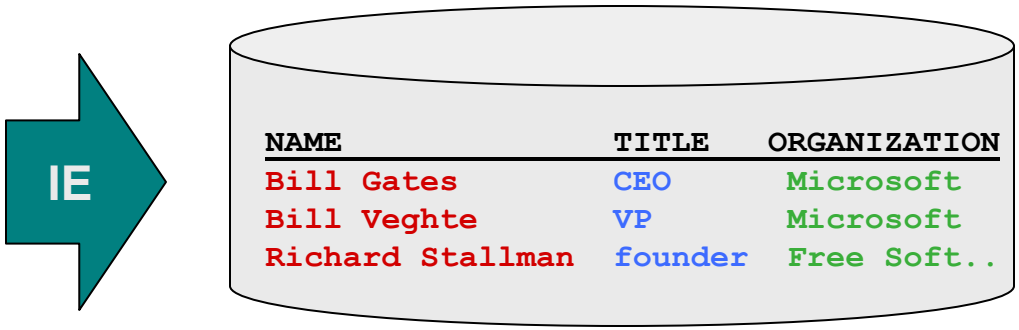
October 14, 2002, 4:00 a.m. PT

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

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

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

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



What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + clustering + association

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

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

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

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

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

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

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…

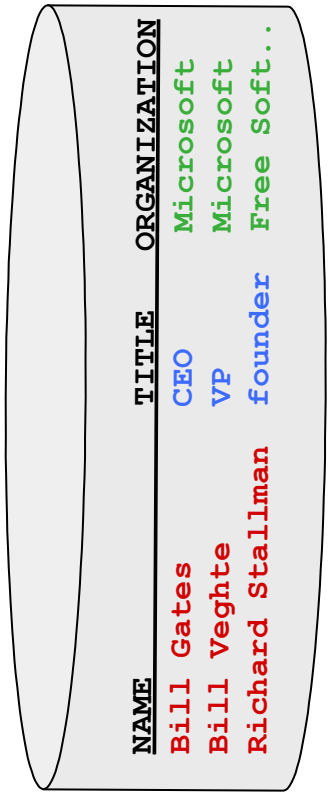
* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

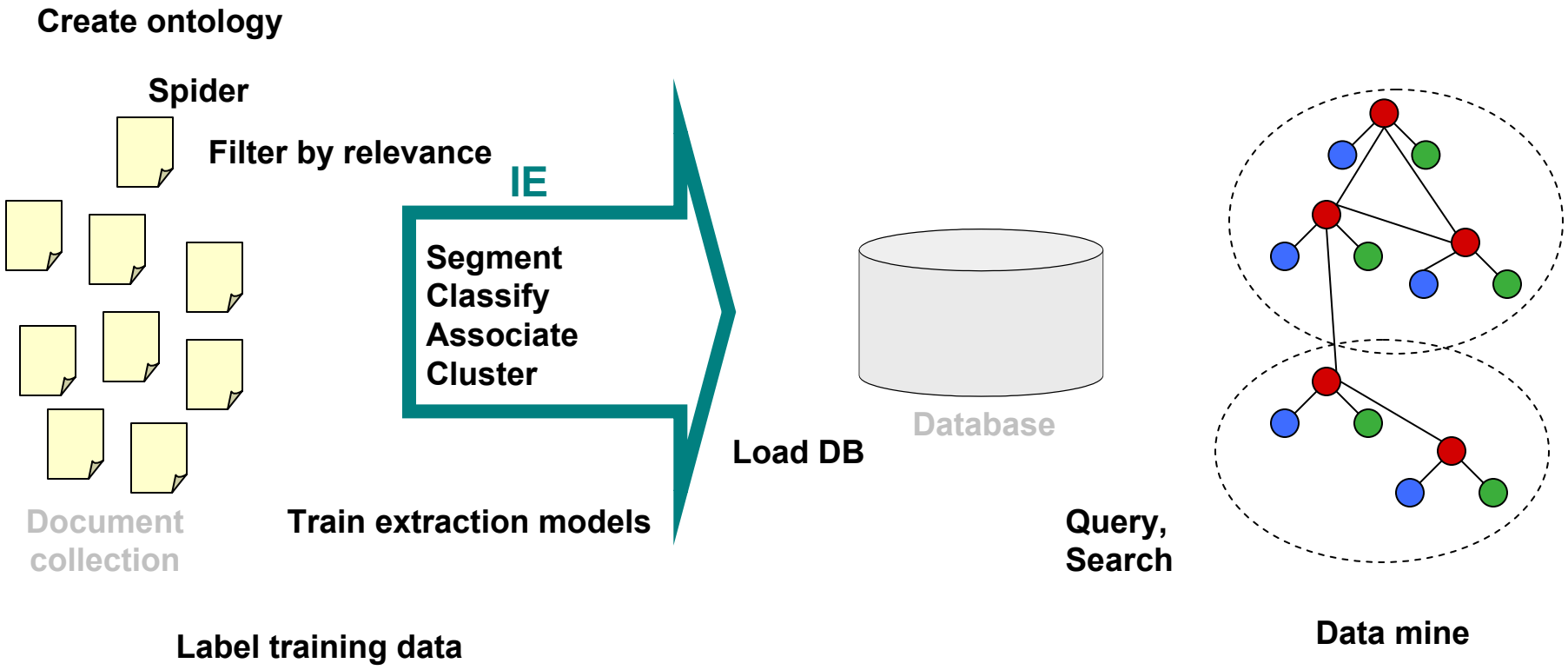
* Microsoft

* 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"

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

- AAAI '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

Newswire

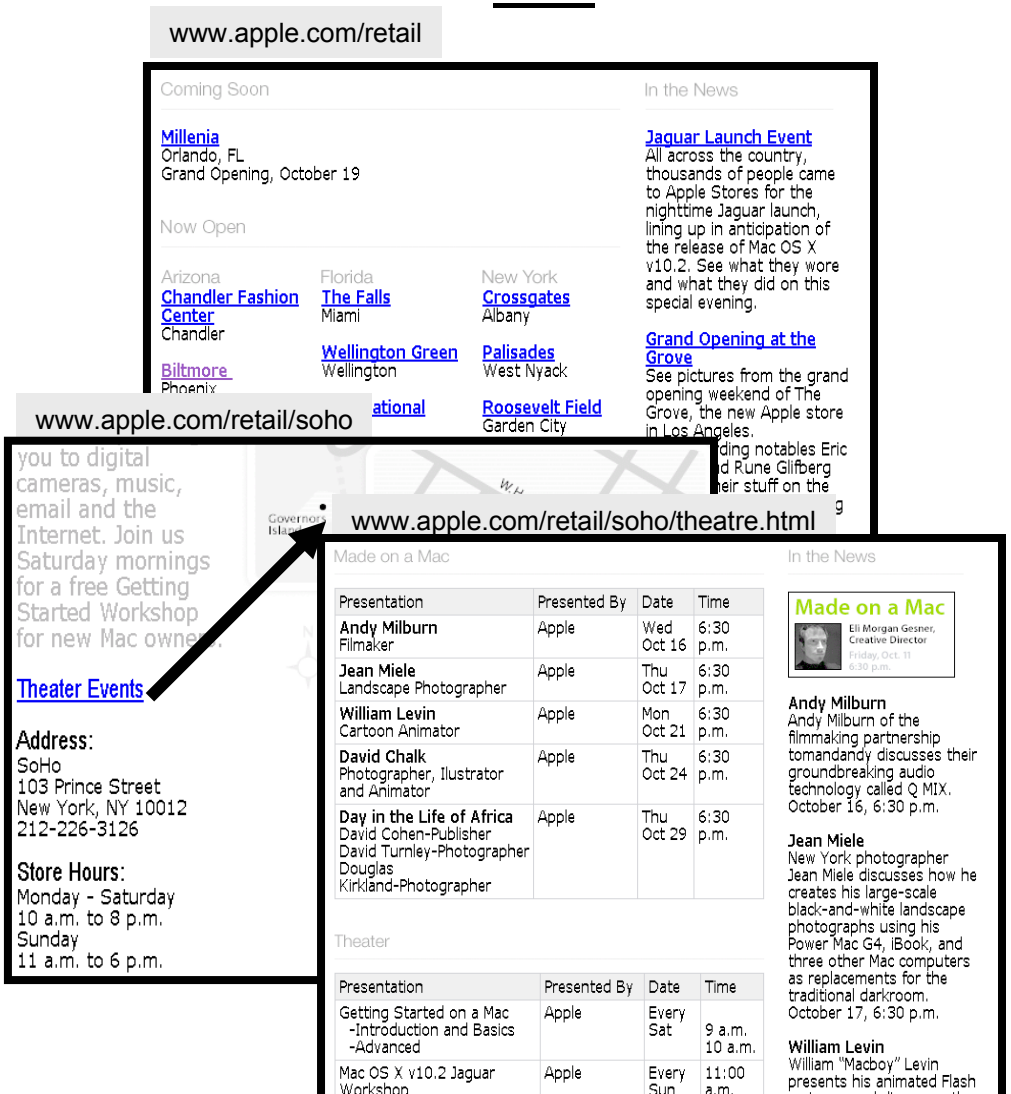
Apple to Open Its First Retail Store in New York City

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



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

Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*				
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 *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	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

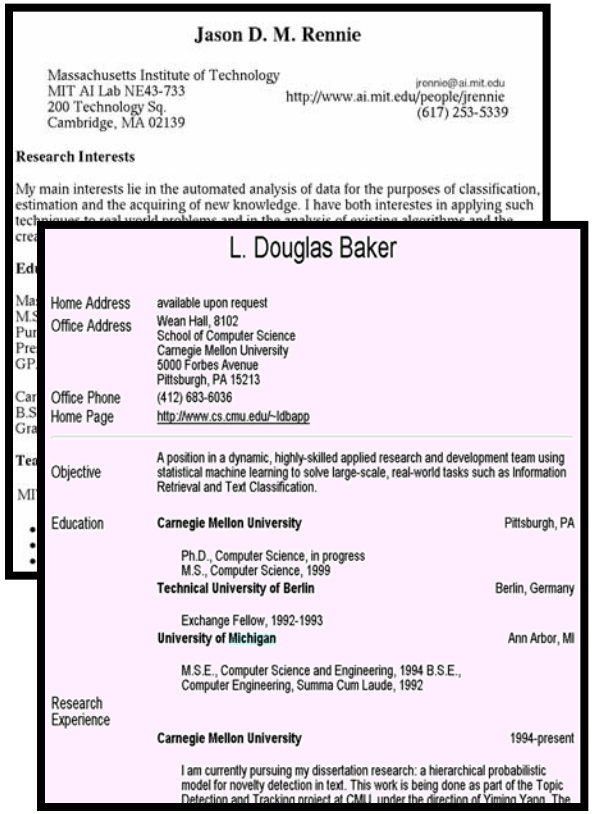
Formatting

Amazon.com Book Pages

Genre specific

Layout

Resumes



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 interestes in applying such

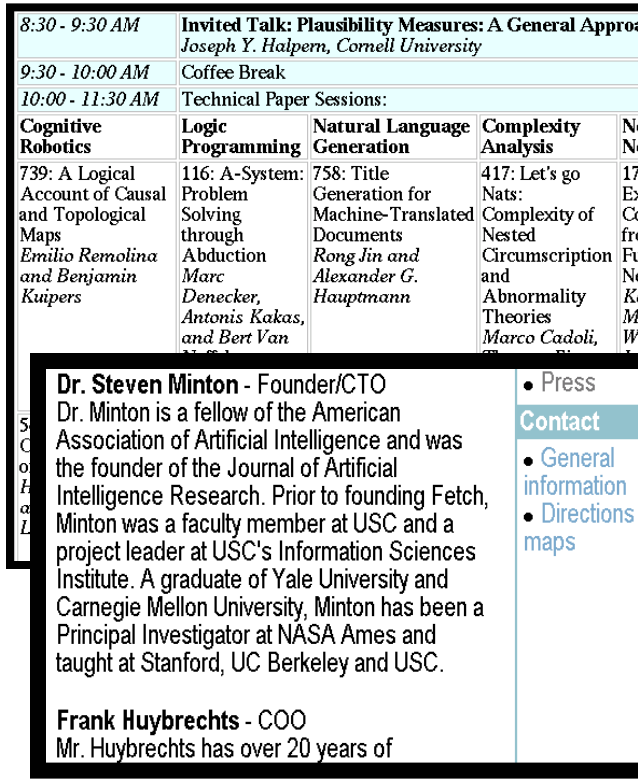
L. Douglas Baker

Home Address	available upon request
Office Address	Wean Hall, 8102 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	**Carnegie Mellon University** Pittsburgh, PA Ph.D., Computer Science, in progress M.S., Computer Science, 1999 **Technical University of Berlin** Berlin, Germany Exchange Fellow, 1992-1993 **University of Michigan** Ann Arbor, MI M.S.E., Computer Science and Engineering, 1994 B.S.E., Computer Engineering, Summa Cum Laude, 1992
Research Experience	**Carnegie Mellon University** 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

Wide, non-specific

Language

University Names



8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approa** *Joseph Y. Halpern, Cornell University*
9:30 - 10:00 AM	Coffee Break
10:00 - 11:30 AM	Technical Paper Sessions:

Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis	N
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli,*	17 Ex Co fr Fu Ne K M W

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…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

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

Complex pattern

U.S. postal addresses

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

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

Ambiguous patterns, needing context and many sources of evidence

Person names

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

Pawel Opalinski, Software Engineer at WhizBang Labs.

Landscape of IE Tasks (4/4): 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

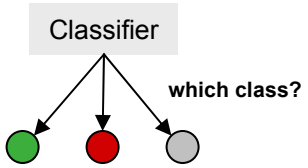
Landscape of IE Techniques (1/1): Models

Lexicons

Abraham Lincoln was born in Kentucky.

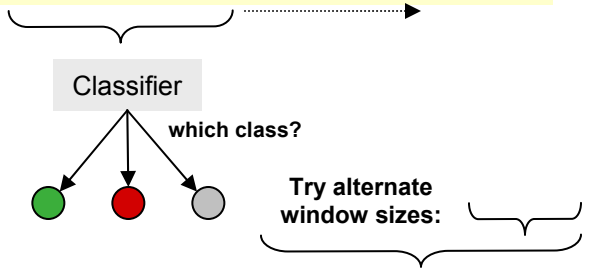
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.



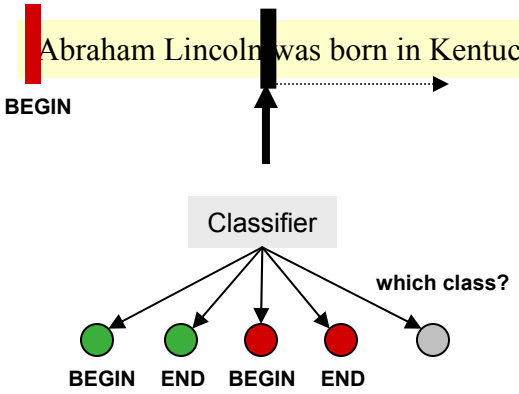
Sliding Window

Abraham Lincoln was born in Kentucky.



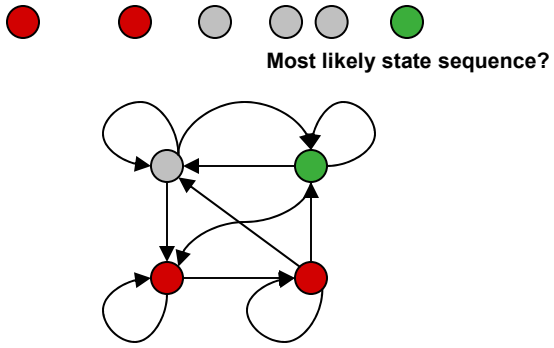
Boundary Models

Abraham Lincoln was born in Kentucky.



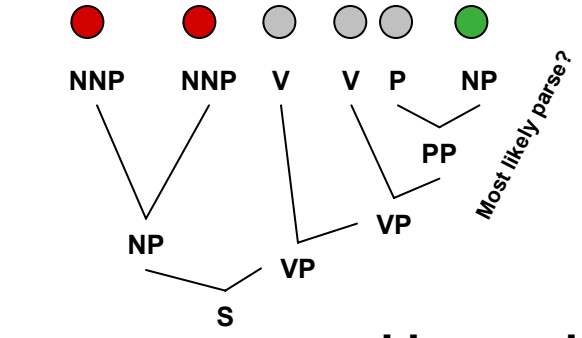
Finite State Machines

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

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]

… 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun …

w_{t-m} … w_{t-1} w_t … w_{t+n} w_{t+n+1} … w_{t+n+m}

prefix contents suffix

P("Wean Hall Rm 5409" = LOCATION) =

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

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("Wean Hall Rm 5409" = 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.
```

Field	F1
Person Name:	**30%**
Location:	**61%**
Start Time:	**98%**

Problems with Sliding Windows and Boundary Finders

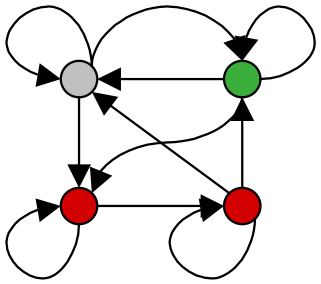
- Decisions in neighboring parts of the input are made independently from each other.
  - Naïve Bayes Sliding Window may predict a "seminar end time" before the "seminar start time".
  - It is possible for two *overlapping* windows to both be above threshold.
  - In a Boundary-Finding system, left boundaries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

Finite State Machines

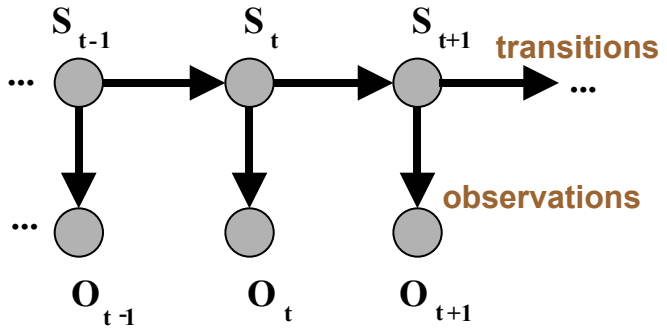
Hidden Markov Models

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

Finite state model



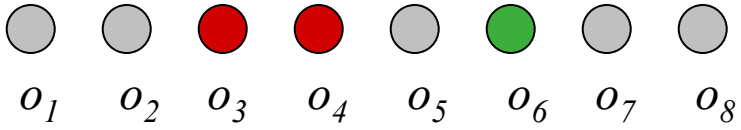
Graphical model



Generates:

State sequence

Observation sequence



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

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

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:

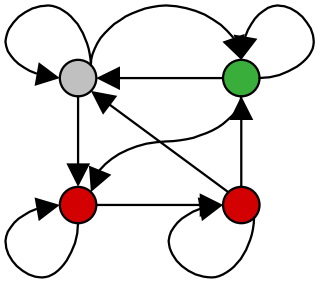
Maximize probability of training observations (w/ prior)

IE with Hidden Markov Models

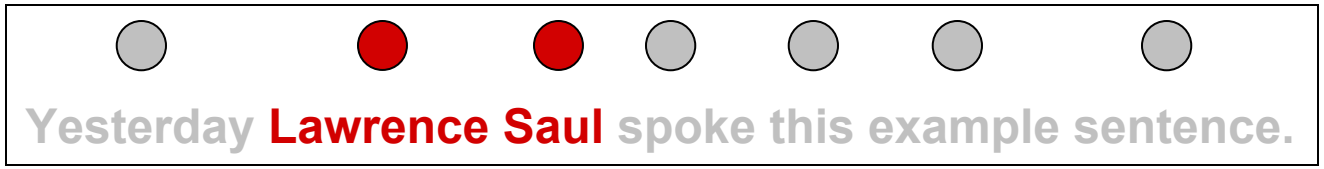
Given a sequence of observations:

Yesterday Lawrence Saul spoke this example sentence.

and a trained HMM:



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



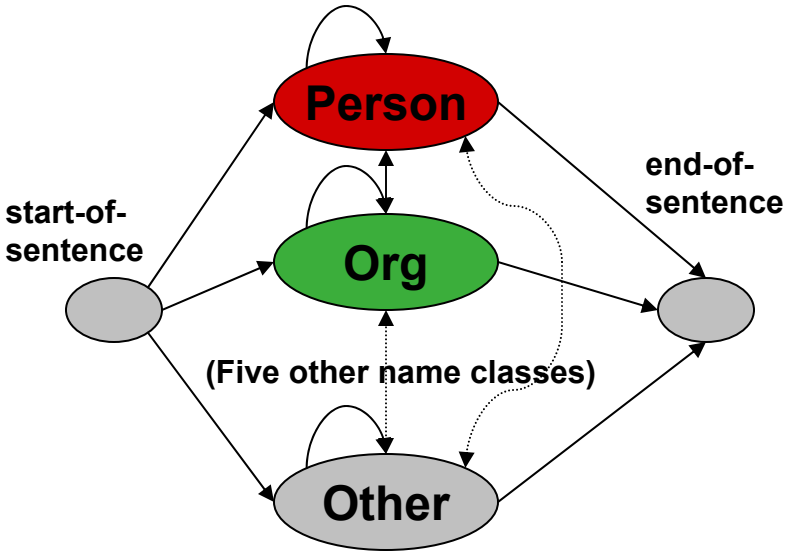
Any words said to be generated by the designated "person name" state extract as a person name:

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:

Case	Language	F1 .
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-alphabetics
- 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: $P(\overline{s}|\overline{o})$ instead of $P(\overline{s},\overline{o})$:
  - 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
```

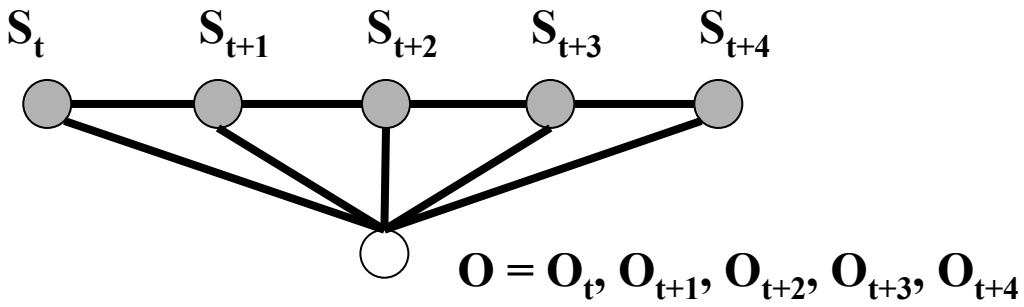
Procedure: For each FAQ, train on one file, test on other; average.

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*.

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

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

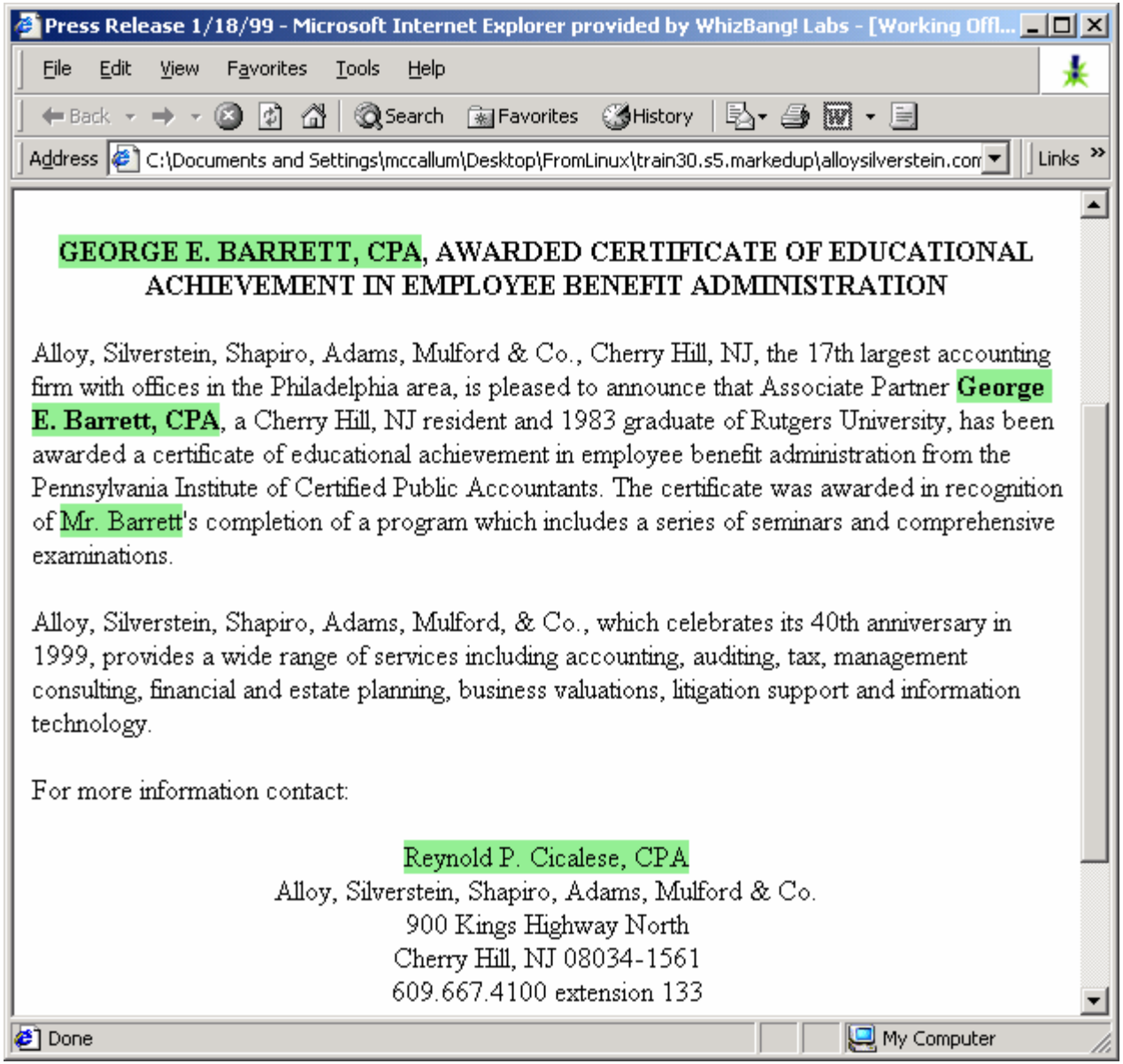
Assuming that the dependency structure of the states is tree-shaped (linear chain is a trivial tree), inference can be done by dynamic programming in time $O(|\vec{o}|\ |S|^2)$—just like HMMs.

General CRFs vs. HMMs

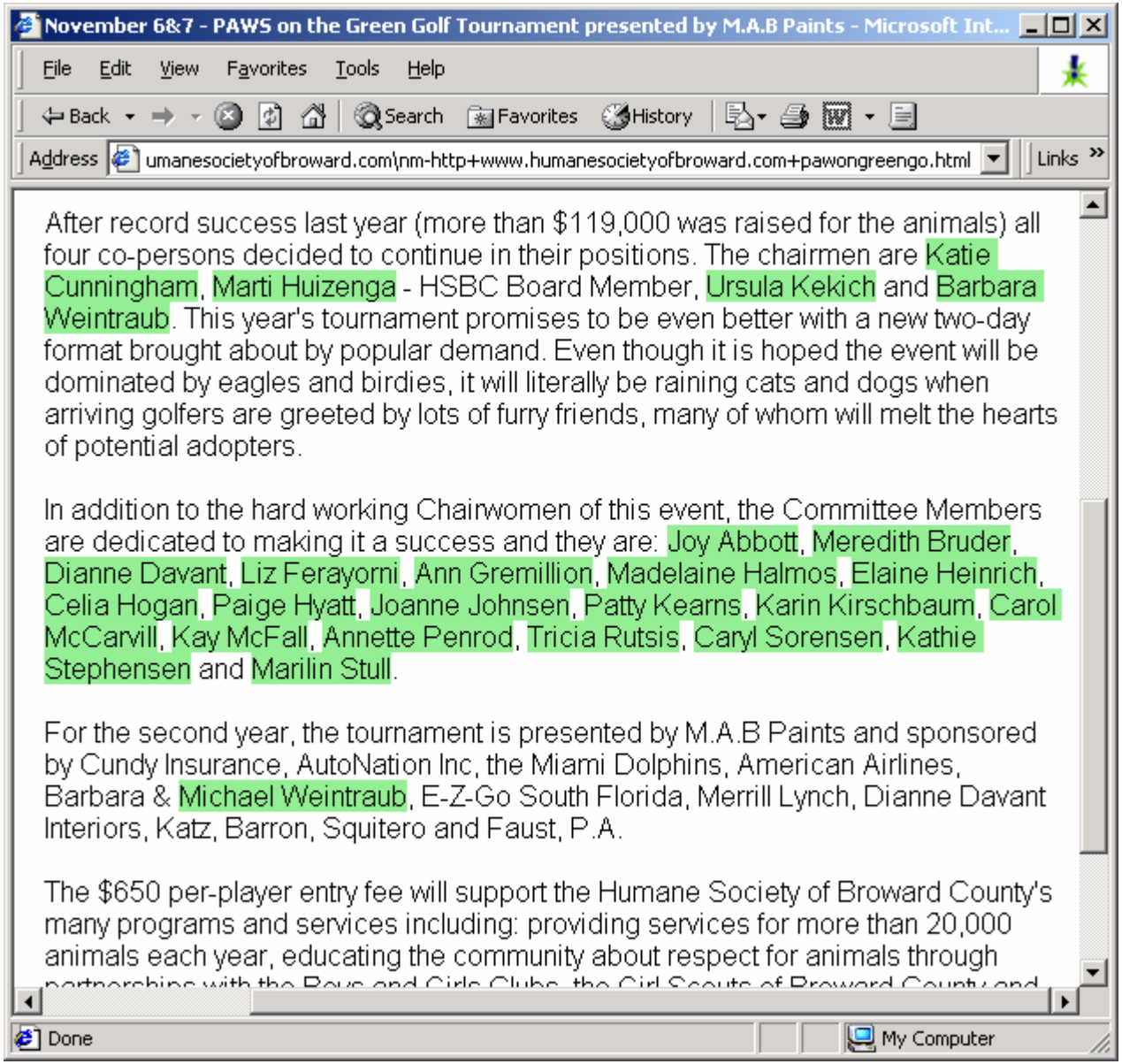
- More general and expressive modeling technique
- Comparable computational efficiency
- Features may be arbitrary functions of *any* or *all* observations
- Parameters need not fully specify generation of observations; require less training data
- Easy to incorporate domain knowledge
- State means only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

Person name Extraction

[McCallum 2001, unpublished]



Person name Extraction



Features in Experiment

Capitalized	Xxxxx
Mixed Caps	XxXxxx
All Caps	XXXXX
Initial Cap	X….
Contains Digit	xxx5
All lowercase	xxxx
Initial	X
Punctuation	.,:;!(), etc
Period	.
Comma	,
Apostrophe	‘
Dash	-
Preceded by HTML tag	

Character n-gram classifier says string is a person name (80% accurate)

In stopword list (the, of, their, etc)

In honorific list (Mr, Mrs, Dr, Sen, etc)

In person suffix list (Jr, Sr, PhD, etc)

In name particle list (de, la, van, der, etc)

In Census lastname list; segmented by P(name)

In Census firstname list; 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)

Hand-built FSM person-name extractor says yes, (prec/recall ~ 30/95)

Conjunctions of all previous feature pairs, evaluated at the current time step.

Conjunctions of all previous feature pairs, evaluated at current step and one step ahead.

All previous features, evaluated two steps ahead.

All previous features, evaluated one step behind.

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 contendors' 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/~appelt/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 Seventeeth 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*. Larence 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 shrinakge. 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] Musclea, I.; Minton, S.; and Knoblock, C. Hierarhical 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.
- [Punyakanok & Roth 2001] Punyakanok, 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, MI.* 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.