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

If you dream of cold creamy chocolate or ooey-gooey cookie there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream business. You will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.
Job Openings:
Category = Food Services
Keyword = Baker
Location = Continental U.S.

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

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
<th>Location</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food Pantry Workers</td>
<td>Lutheran Social Services</td>
<td></td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Cooks</td>
<td>Lutheran Social Services</td>
<td></td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Bakers Assistants</td>
<td>Fine Catering by Russell Morin</td>
<td></td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Baker's Helper</td>
<td>Bird-in-Hand</td>
<td></td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Assistant Baker</td>
<td>Gourmet To Go</td>
<td></td>
<td>October 11, 2002</td>
</tr>
<tr>
<td>Host/Hostess</td>
<td>Sharis Restaurants</td>
<td></td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Cooks</td>
<td>Alta's Rustler Lodge</td>
<td></td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Line Attendant</td>
<td>Sun Valley Corporation</td>
<td></td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Food Service Worker II</td>
<td>Garden Grove Unified School District</td>
<td></td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Night Cook / Baker</td>
<td>SONOCO</td>
<td></td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Cooks/Prep Cooks</td>
<td>GrandView Lodge</td>
<td></td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Line Cook</td>
<td>Lone Mountain Ranch</td>
<td></td>
<td>October 10, 2002</td>
</tr>
<tr>
<td>Production Baker</td>
<td>Whole Foods Market</td>
<td></td>
<td>October 08, 2002</td>
</tr>
<tr>
<td>Cake Decorator/Baker</td>
<td>Mandalay Bay Hotel and Casino</td>
<td></td>
<td>October 08, 2002</td>
</tr>
<tr>
<td>Shift Supervisors</td>
<td>Brueggers Bagels</td>
<td></td>
<td>October 08, 2002</td>
</tr>
</tbody>
</table>
U.S. Job Supply Increases Amid Rising Unemployment

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

Special Offer! Find out how you can earn a free subscription to the JOI Report on U.S. Labor Markets through a limited-time JOI Subscriber Referral Program!
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…
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...
What is “Information Extraction”

Information Extraction = segmentation + classification + association + clustering

As a family of techniques:

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 + 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...
IE in Context

Create ontology

Spider

Filter by relevance

IE

Segment
Classify
Associate
Cluster

Load DB

Train extraction models

Document collection

Label training data

Database

Query, Search

Data mine
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”
• 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

Coming Soon

Millenia
Orlando, FL
Grand Opening, October 19

New Open

Arizona
Chandler Fashion Center

Florida
The Falls
Miami

New York
Crossgates
Albany

Billmore
Pennsylvania

Wellington Green
West Nyack

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 at The Grove, the new Apple store in Los Angeles.

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

Theater Events

Address:
Soho
108 Prince Street
New York, NY 10012
212-226-8126

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

Presentation
Made on a Mac

Presented By
Andy Milburn

Date
Every Sat.
9 a.m.

Time
10 a.m.
11:00 a.m.

Made on a Mac
Andy Milburn of the filmmaking partnership
temporarily discusses their groundbreaking video

technology called Q-NIX.

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 "Madboy" Levin
presents his animated Flash

Day in the Life of Africa
David Coleman-Publisher

David Turnley-Photographer

Kirkland-Photographer

Presentation
Getting Started on a Mac
-Introduction and Basics

-Advanced

Workshop
Mac OS X v10.2 Jaguar

Presented By
Apple

Date
Every Sat.
9 a.m.

Time
10 a.m.
11:00 a.m.

Andy Milburn
Director of Creative

David Chalk
Photographer, Illustrator

Jean Miele
Landscape Photographer

William Levin
Cartoon Illustrator

made on a mac

Made on a Mac
Andy Milburn

Jean Miele

William Levin

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mac

made on a mc
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.
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
Landscape of IE Tasks (3/4): Pattern Complexity

E.g. word patterns:

<table>
<thead>
<tr>
<th><strong>Closed set</strong></th>
<th><strong>Regular set</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. states</td>
<td>U.S. phone numbers</td>
</tr>
<tr>
<td>He was born in Alabama…</td>
<td>Phone: (413) 545-1323</td>
</tr>
<tr>
<td>The big Wyoming sky…</td>
<td>The CALD main office can be reached at 412-268-1299</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Complex pattern</strong></th>
<th><strong>Ambiguous patterns,</strong></th>
<th><strong>Person names</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. postal addresses</td>
<td>needing context and many sources of evidence</td>
<td>...was among the six houses sold by Hope Feldman that year.</td>
</tr>
<tr>
<td>University of Arkansas</td>
<td></td>
<td>Pawel Opalinski, Software Engineer at WhizBang Labs.</td>
</tr>
<tr>
<td>P.O. Box 140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hope, AR 71802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Headquarters:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1128 Main Street, 4th Floor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cincinnati, Ohio 45210</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

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

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

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

Any of these models can be used to capture words, formatting or both.
Sliding Windows
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.
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.
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
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.
A “Naïve Bayes” Sliding Window Model

[Freitag 1997]

Try all start positions and reasonable lengths

Prior probability of start position
Prior probability of length
Probability prefix words
Probability contents words
Probability suffix words

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)
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.
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:

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) \( \arg \max_s P(\tilde{s}, \tilde{o}) \)

Yesterday Lawrence Saul spoke this example sentence.

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

Person name: Lawrence Saul
HMM Example: “Nymble”

Task: Named Entity Extraction

![Diagram of entity classification]

Train on 450k words of news wire text.

Results:

<table>
<thead>
<tr>
<th>Case</th>
<th>Language</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed</td>
<td>English</td>
<td>93%</td>
</tr>
<tr>
<td>Upper</td>
<td>English</td>
<td>91%</td>
</tr>
<tr>
<td>Mixed</td>
<td>Spanish</td>
<td>90%</td>
</tr>
</tbody>
</table>

[BBN “IdentiFinder”]

[Bikel, et al 1998]
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: $P(\tilde{s}, \tilde{d})$
- 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: $P(s|o)$ instead of $P(s,o)$:
  – 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> 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> 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

<table>
<thead>
<tr>
<th>Begins with number</th>
<th>Begins with ordinal</th>
<th>Begins with punctuation</th>
<th>Begins with question word</th>
<th>Begins with subject</th>
<th>Blank</th>
<th>Contains alphanum</th>
<th>Contains bracketed number</th>
<th>Contains http</th>
<th>Contains non-space</th>
<th>Contains number</th>
<th>Contains pipe</th>
<th>Contains question mark</th>
<th>Contains question word</th>
<th>Ends with question mark</th>
<th>First-alpha-is-capitalized</th>
<th>Indented</th>
<th>Indented-1-to-4</th>
<th>Indented-5-to-10</th>
<th>More-than-one-third-space</th>
<th>Only-punctuation</th>
<th>Prev-is-blank</th>
<th>Prev-begins-with-ordinal</th>
<th>Shorter-than-30</th>
</tr>
</thead>
</table>
Markov on $s$, conditional dependency on $o$.

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

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 $O(|\tilde{o}| |S|^2)$—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”
GEORGE E. BARRETT, CPA, AWARDED CERTIFICATE OF EDUCATIONAL ACHIEVEMENT IN EMPLOYEE BENEFIT ADMINISTRATION

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

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

For more information contact:

Reynold P. Cicalese, CPA
Alloy, Silverstein, Shapiro, Adams, Mulford & Co.
900 Kings Highway North
Cherry Hill, NJ 08034-1561
609.667.4100 extension 133
After record success last year (more than $119,000 was raised for the animals) all four co-persons decided to continue in their positions. The chairmen are Katie Cunningham, Marti Huizenga - HSBC Board Member, Ursula Kekich and Barbara Weintraub. This year’s tournament promises to be even better with a new two-day format brought about by popular demand. Even though it is hoped the event will be dominated by eagles and birds, it will literally be raining cats and dogs when arriving golfers are greeted by lots of furry friends, many of whom will melt the hearts of potential adopters.

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

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

The $650 per-player entry fee will support the Humane Society of Broward County’s many programs and services including: providing services for more than 20,000 animals each year, educating the community about respect for animals through partnerships with the Boys and Girls Clubs, the Girl Scouts of Broward County and...
## Features in Experiment

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capitalized</td>
<td>Character n-gram classifier says string is a person name (80% accurate)</td>
</tr>
<tr>
<td>Mixed Caps</td>
<td>In stopword list</td>
</tr>
<tr>
<td>All Caps</td>
<td>(the, of, their, etc)</td>
</tr>
<tr>
<td>Initial Cap</td>
<td>In honorific list</td>
</tr>
<tr>
<td>Contains Digit</td>
<td>(Mr, Mrs, Dr, Sen, etc)</td>
</tr>
<tr>
<td>All lowercase</td>
<td>In person suffix list</td>
</tr>
<tr>
<td>Initial</td>
<td>(Jr, Sr, PhD, etc)</td>
</tr>
<tr>
<td>Punctuation</td>
<td>In name particle list</td>
</tr>
<tr>
<td>Period</td>
<td>(de, la, van, der, etc)</td>
</tr>
<tr>
<td>Comma</td>
<td>In Census lastname list; segmented by P(name)</td>
</tr>
<tr>
<td>Apostrophe</td>
<td>In Census firstname list; segmented by P(name)</td>
</tr>
<tr>
<td>Dash</td>
<td>In locations lists</td>
</tr>
<tr>
<td>Preceded by HTML tag</td>
<td>(states, cities, countries)</td>
</tr>
<tr>
<td></td>
<td>In company name list</td>
</tr>
<tr>
<td></td>
<td>(“J. C. Penny”)</td>
</tr>
<tr>
<td></td>
<td>In list of company suffixes</td>
</tr>
<tr>
<td></td>
<td>(Inc, &amp; Associates, Foundation)</td>
</tr>
</tbody>
</table>

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

• Trained on 800 segmented sentences from UPenn Chinese Treebank.
• Training time: ~2 hours with L-BFGS.

• Training F1: 99.4%
• Testing F1: 99.3%

• Previous top contenders’ F1: ~85-95%

[McCallum & Feng, to appear]
IE Resources

• Data
  – 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

• Both
  – http://www.cis.upenn.edu/~adwait/penntools.html
  – http://www.cs.umass.edu/~mccallum/ie
References