Information Extraction with Finite State Models and Scoped Learning

Andrew McCallum
WhizBang Labs & CMU

Joint work with John Lafferty (CMU), Fernando Pereira (UPenn), Dayne Freitag (Burning Glass), David Blei (UC Berkeley), Drew Bagnell (CMU) and many others at WhizBang Labs.
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

Ice Cream Guru

If you dream of cold creamy chocolate or ooey-gooey cookie dough ice cream, there’s a great opportunity for you to work with the major food manufacturer in the Chicago area. This job requires a BS in food science or a related field, and involves work in the test kitchen. Experience in product development and marketing is preferred. Contact us at 1-800-488-2611.
Step 1
Location:
Where do you want to work?

Step 2
Category:
What type of work?

Step 3
Employer:
Which employer?

Keywords:
Show recruiter & staffing agency listings

3,079 jobs found
1) Location: Pittsburgh Area, PA
2) Category: All Categories
3) Employer: All Employers

Tips
Click on a state to see the cities in that state where job opportunities exist. If you are interested in a particular city, select it. Then, select the jobs that interest you. If you have trouble, see the Search Tips.

Address:
http://www.flipdog.com/js/local.html?_requestid=1161768

Employers • Support

Start Over | Get Results

PA

Pennsylvania:
- Philadelphia Area
- Pine Grove
- Pittsburgh Area
- Pottsville
- Reading Area
- Reno
- Pikes Landing
- Robinson
- Rond
- Russell
- Sacramento
- Saint Clair
- Saint Thomas
- Schuykill Haven
- Scotland
Step 1
Location:
Where do you want to work?

Step 2
Category:
What type of work?

- Health Care
- Human Resources
- Manufacturing/Business Operations
- Marketing/Advertising
- Media
- Other
- Professional Services
- Sales

Step 3
Employer:
Which employer?
Show recruiter & staffing agency listings

Keywords:

28 jobs found

1) Location: Pittsburgh Area, PA
2) Category: Human Resources
3) Employer: All Employers

Tips
Select a category to see a list of functions that contain jobs. To select or deselect multiple categories or
1 - 25 of 29 jobs shown below

Search within results for:  

Premium Postings

Partner Consultant at Profiles International.com
Independent Consultant calling on business and industry marketing 21st century Assessment Instruments to help with hiring, developing and managing of human capital needs. We are a business service operating in 38+ countries. This

Web Directory

Coordinating Interviewer at University of Pittsburgh  
Director of Employee Relations at Macromedia  
COMPENSATION & BENEFITS MANAGER at Chelsea Building Products, Inc.  
National Fleet Safety Manager at Western Pennsylvania Chapter, American Society of Safety Engineers  
Drafters at Oxford Technology  
Career Services Student Counselor at University of Pittsburgh
An HR office

Jobs, but not HR jobs

Microsoft Great Plains Business Solutions: Human Resources & Payroll
www.greatplains.com  Manage employee information, benefits and payroll efficiently

University of Pittsburgh Office of Human Resources 100 Craig Hall...
University of Pittsburgh Office of Human Resources 100 Craig Hall
Pittsburgh, PA 15260 Telephone: (412) 624-8150. ...
www.hr.pitt.edu/employment/default.htm - 11k - Cached - Similar pages

New Page 1
www.hr.pitt.edu/employ/employ.htm - 1k - Cached - Similar pages
[ More results from www.hr.pitt.edu ]

Pittsburgh jobs and job listings from Pittsburgh.com
SEARCH: The Web Yellow Pages, HOME, Job Search: Find Pittsburgh jobs
Keyword: City: ... Browse Pittsburgh Job Postings by Category. ...
www.realpittsburgh.com/shared/jobs/ - 27k - Cached - Similar pages

Pittsburgh.com: Human Resources Job Search
SEARCH: The Web Yellow Pages, ... Your Human Resources Job Search Find a Human Resources job; ... Exclude National & Regional Jobs. Salary range (per year): ...
www.realpittsburgh.com/shared/jobs/thform09.html - 24k - Cached - Similar pages
[ More results from www.realpittsburgh.com ]

Carnegie Library of Pittsburgh:--Working at CLP
... This page is maintained by the Human Resources Department at the Carnegie Library
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!
Extracting Continuing Education Courses

Introduction to Medical Insurance Billing
NCR 9131
Sat. - 2 p.m. - 5 p.m.
9/23/2000. 1 meeting
Cal Poly, TBA
Fee $99 (includes course materials). 8 CEU
Registration Deadline: 9/25/2000

As the baby boomer generation ages, health care will continue to be one of the fastest-growing sectors of the U.S. economy. Medical insurance billers can work in a variety of settings, including physicians’ offices, clinics, hospitals, medical supply firms, and even home offices. In this one-day class, you will be introduced to the concepts of coding, medical terminology, and how to fill out and submit an insurance claim form. Instruction will include explanations and exercises in how to bill government programs like Medicare, MediCal and Champus, private insurance (such as Blue Cross, Blue Shield, and other private carriers), workers’ compensation, and managed care organizations (including HMOs, POSs, IPAs and how they work).

You will receive a certificate of completion.

Peter Winston, M.H.A., has worked in management and administrative positions in hospitals and medical groups in Nevada and California since 1982. He is currently administrator of an organization that owns and operates clinics and independent practice associations (IPA).

College Prep

Maximizing College Entrance Potential: SAT I Prep Course
NCR 91634

Description
Become a Notary Public in One Day NCR 9139 Sat., 9 a.m. - 2 p.m., 9/23/2000. 1 meeting at Cal Poly, TBA. Registration Deadline: 9/18/2000. This is a one-day intensive course designed to provide you with every...
From http://www.calpoly.edu/~exted/COURSES/Courses.html

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Access for Office</td>
<td>Maximize College Entrance Potential: SAT I Prep Course NCR 91634.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number</th>
<th>Cost</th>
<th>Meeting time</th>
<th>Meeting time</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCR 9131</td>
<td>$99</td>
<td>Sat., 9 a.m. - 2 p.m., 9/23/2000. 1 meeting at Cal Poly, TBA.</td>
<td></td>
</tr>
<tr>
<td>NCR 91634</td>
<td>$190</td>
<td>Mon., 5:30 - 9:30 p.m., 10/16/2000 - 11/6/2000. 4 meetings</td>
<td></td>
</tr>
</tbody>
</table>
Are you looking for career satisfaction and personal growth? Do your employees need new skills? You’ve come to the right place.

- Search for training & education resources.
- Explore financial aid options.
- Post course offerings in our online provider marketplace.

Search Type
Keyword
Subject

Enter a Keyword

OR

Choose a Subject

Administration and Management
Basic Skills
Business and Finance
Clerical, Secretarial, Administrative Support
You are searching "Courses" for keyword "bioinformatics" delivered via "all delivery type" delivered via in the state of "Maryland"

### Search Results: 3 Courses [1 - 3]

<table>
<thead>
<tr>
<th>Compare</th>
<th>Course Title</th>
<th>State</th>
<th>City</th>
<th>Cost</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Biotechnology and Genetic Engineering</td>
<td>MD</td>
<td>Adelphi</td>
<td></td>
<td>University of Maryland University College</td>
</tr>
<tr>
<td></td>
<td>Techniques in Biotechnology I</td>
<td>MD</td>
<td>Catonsville</td>
<td></td>
<td>Community College of Baltimore County</td>
</tr>
<tr>
<td></td>
<td>Techniques in Biotechnology II</td>
<td>MD</td>
<td>Catonsville</td>
<td></td>
<td>Community College of Baltimore County</td>
</tr>
</tbody>
</table>

Check up to 4: [Compare](#) [List](#)

Enter additional search criteria [New Search](#)
**Biotechnology and Genetic Engineering**

<table>
<thead>
<tr>
<th>Provider</th>
<th>University of Maryland University College [More details on this provider]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>(Formerly GNSC 398B. Science background not required.) An introduction to the basic principles and applications of biotechnology and genetic engineering to medicine, agriculture, and industry. Topics include gene therapy, cloning, the identification and isolation of genes involved in human health and disease, diagnostic and forensics testing, the human genome project, bioremediation, microbial and plant bioengineering, and bioinformatics. Students may receive credit for only one of the following courses: BIOL 398I or GNSC 398B.</td>
</tr>
<tr>
<td>Type of offering</td>
<td>Course</td>
</tr>
<tr>
<td>Available starting</td>
<td>01/12/2002</td>
</tr>
<tr>
<td>Available until</td>
<td>01/19/2002</td>
</tr>
<tr>
<td>Location(s) of offering</td>
<td>College Park</td>
</tr>
<tr>
<td>Method of delivery</td>
<td>Classroom</td>
</tr>
<tr>
<td>Credit earned</td>
<td>1 credits</td>
</tr>
<tr>
<td>Provider's course code</td>
<td>BIOL398I4075</td>
</tr>
</tbody>
</table>

For registration or more information:
This took place in ‘99

Not in Maryland

This took place in ‘99

Courses from all over the world
Why prefer “knowledge base search” over “page search”

• *Targeted, restricted universe of hits*
  – Don’t show resumes when I’m looking for job openings.

• **Specialized queries**
  – Topic-specific
  – Multi-dimensional
  – Based on information spread on multiple pages.

• **Get correct granularity**
  – Site, page, paragraph

• **Specialized display**
  – Super-targeted hit summarization in terms of DB slot values

• **Ability to support sophisticated data mining**
Issues that arise

**Application issues**
- Directed spidering
- Page classification
- Information extraction
- Record association
- De-duplication

**Scientific issues**
- Learning more than 100k parameters from limited and noisy training data
- Taking advantage of rich, multi-faceted features and structure
- Leveraging local regularities in training and test data
- Clustering massive data sets
Issues that arise

**Application issues**
- Directed spidering
- Page classification
- Information extraction
- Record association
- De-duplication

**Scientific issues**
- Learning more than 100k parameters from limited and noisy training data
- Taking advantage of rich, multi-faceted features and structure
- Leveraging local regularities in training and test data
- Clustering massive data sets
Mining the Web for Research Papers

[McCallum et al ‘99]

Reinforcement Learning: A Survey

Leslie Pack Kaelbling
Michael L. Littman
Computer Science Department, Box 1910, Brown University
Providence, RI 02912-1910 USA

Andrew W. Moore
Smith Hall 261, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

This paper surveys the field of reinforcement learning from a computer-science perspective. It is written to be accessible to researchers familiar with machine learning. Both the historical basis of the field and a broad selection of current work are summarized. Reinforcement learning is the problem faced by an agent that learns behavior through trial-and-error interactions with a dynamic environment. The work described here has resemblance to work in psychology, but differs considerably in the details and in the use of the word “reinforcement.” The paper discusses central issues of reinforcement learning including trading off exploration and exploitation, establishing the foundations of the field via Markov decision theory, learning from delayed reinforcement, constructing empirical models to accelerate learning, making use of generalization and hierarchy, and coping with hidden state. It concludes with a survey of some implemented systems and an assessment of the practical utility of current methods for reinforcement learning.

1. Introduction

Reinforcement learning dates back to the early days of cybernetics and work in art psychology, neuroscience, and computer science. In the last five to ten years, it has attained a rapidly increasing interest in the machine learning and artificial intelligence communities. Its promise is beguiling—a way of programming agents by reward and punishment without needing to specify how the task is to be achieved. But there are formidable computational obstacles to fulfilling the promise.

This paper surveys the historical basis of reinforcement learning and some of the current work from a computer science perspective. We give a high-level overview of the field, taste of some specific approaches. It is, of course, impossible to mention all of the important work in the field; this should not be taken to be an exhaustive account.
Information Extraction with HMMs

Parameters = $P(s_t|s_{t-1})$, $P(o_t|s_t)$ for all states in $S=\{s_1, s_2, \ldots\}$

Emissions = word

Training = Maximize probability of training observations (+ prior).

For IE, states indicate “database field”.

[Seymore & McCallum ‘99]
[Freitag & McCallum ‘99]
Regrets with HMMs

1. Would prefer richer representation of text: multiple overlapping features, whole chunks of text.

   - Example word features:
     - identity of word
     - word is in all caps
     - word ends in “-ski”
     - word is part of a noun phrase
     - word is in bold font
     - word is on left hand side of page
     - word is under node X in WordNet
     - features of past and future

   - Example line or paragraph features:
     - length
     - is centered
     - percent of non-alphabetics
     - total amount of white space
     - contains two verbs
     - begins with a number
     - grammatically contains a question
     - agglomerative features of sequence

2. HMMs are generative models of the text: $P(\{s\ldots\},\{o\ldots\})$. Generative models do not handle easily overlapping, non-independent features. Would prefer a conditional model: $P(\{s\ldots\}|\{o\ldots\})$. 
Solution: conditional sequence model

[McCallum, Freitag, Pereira ‘2000]

Old graphical model

*Traditional HMM*

```
... S_{t-1} -> S_t -> O_t ...
```

- $P(o_t|s_t)$
- $P(s_t|s_{t-1})$

New graphical model

*Maximum Entropy Markov Model*

```
... S_{t-1} -> S_t -> O_t ...
```

- $P(s_t|o_t, s_{t-1})$

Standard belief propagation: forward-backward procedure. Viterbi and Baum-Welch follow naturally.
Exponential Form for “Next State” Function

\[
P(s_t \mid s_{t-1}, o_t) \Rightarrow P_{s_{t-1}}(s_t \mid o_t) = \frac{1}{Z(o_t, s_{t-1})} \exp \left( \sum_k \lambda_k f_k(o_t, s_t) \right)
\]

Recipe:
- Labeled data is assigned to transitions.
- Train each state’s exponential model by maximum likelihood (iterative scaling).
Experimental Data

38 files belonging to 7 UseNet FAQs

Example:

X-NNTP-Poster: NewsHound v1.33
Archive-name: acorn/faq/part2
Frequency: monthly

2.6) What configuration of serial cable should I use?

Here follows a diagram of the necessary connection programs to work properly. They are as far as I know agreed upon by commercial comms software developers fo

Pins 1, 4, and 8 must be connected together inside

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>Contains question mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Begins with ordinal</td>
<td>Contains question word</td>
</tr>
<tr>
<td>Begins with punctuation</td>
<td>Ends with question mark</td>
</tr>
<tr>
<td>Begins with question word</td>
<td>First alpha is capitalized</td>
</tr>
<tr>
<td>Begins with subject</td>
<td>Indented</td>
</tr>
<tr>
<td>Blank</td>
<td>Indented 1-to-4</td>
</tr>
<tr>
<td>Contains alphanum</td>
<td>Indented 5-to-10</td>
</tr>
<tr>
<td>Contains bracketed-number</td>
<td>More than one-third space</td>
</tr>
<tr>
<td>Contains http</td>
<td>Only punctuation</td>
</tr>
<tr>
<td>Contains non-space</td>
<td>Prev is blank</td>
</tr>
<tr>
<td>Contains number</td>
<td>Prev begins with ordinal</td>
</tr>
<tr>
<td>Contains pipe</td>
<td>Shorter than 30</td>
</tr>
</tbody>
</table>
Models Tested

• **ME-Stateless**: A single maximum entropy classifier applied to each line independently.

• **TokenHMM**: A fully-connected HMM with four states, one for each of the line categories, each of which generates individual tokens (groups of alphanumeric characters and individual punctuation characters).

• **FeatureHMM**: Identical to TokenHMM, only the lines in a document are first converted to sequences of features.

• **MEMM**: The maximum entropy Markov model described in this talk.
## Results

<table>
<thead>
<tr>
<th>Learner</th>
<th>Segmentation precision</th>
<th>Segmentation recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME-Stateless</td>
<td>0.038</td>
<td>0.362</td>
</tr>
<tr>
<td>TokenHMM</td>
<td>0.276</td>
<td>0.140</td>
</tr>
<tr>
<td>FeatureHMM</td>
<td>0.413</td>
<td>0.529</td>
</tr>
<tr>
<td>MEMM</td>
<td>0.867</td>
<td>0.681</td>
</tr>
</tbody>
</table>
Label Bias Problem in Conditional Sequence Models

- Example (after Bottou ‘91):

```
start
   ↓
  r   o   b
    ↓    ↓
   r   i   b
```

“rob”

“rib”

- Bias toward states with few “siblings”.
- Per-state normalization in MEMMs does not allow “probability mass” to transfer from one branch to the other.
Proposed Solutions

• Determinization:
  – not always possible
  – state-space explosion

• Use fully-connected models:
  – lacks prior structural knowledge.

• Our solution: *Conditional random fields* (CRFs):
  – Probabilistic conditional models generalizing MEMMs.
  – Allow some transitions to *vote* more strongly than others in computing state sequence probability.
  – *Whole sequence* rather than per-state normalization.
From HMMs to MEMMs to CRFs

\[ \vec{s} = s_1, s_2, \ldots, s_n \quad \vec{o} = o_1, o_2, \ldots, o_n \]

**HMM**

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

**MEMM**

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

\[ \propto \prod_{t=1}^{\vec{o}} \frac{1}{Z_{s_{t-1}, o_t}} \exp \left( \sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, x_i) \right) \]

**CRF**

\[ P(\vec{s} | \vec{o}) \propto \frac{1}{Z_{\vec{o}}} \prod_{t=1}^{\vec{o}} \exp \left( \sum_j \lambda_j f_j(s_t, s_{t-1}) + \sum_k \mu_k g_k(s_t, x_i) \right) \]

(A special case of MEMMs and CRFs.)
Conditional Random Fields

Markov on $O$, conditional dependency on $S$.

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

Assuming that the dependency structure of the states is tree-shaped, Hammersley-Clifford-Besag theorem stipulates that the CRF has this form—an exponential function of the cliques in the graph.

Set parameters by maximum likelihood and Conjugate Gradient. Convex likelihood function; guaranteed to find optimal solution!
General CRFs vs. HMMs

- More general and expressive modeling technique
- Comparable computational efficiency
- Features may be arbitrary functions of any / 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”
MEMM & CRF Related Work

• Maximum entropy for language tasks:
  – Language modeling [Rosenfeld ‘94, Chen & Rosenfeld ‘99]
  – Part-of-speech tagging [Ratnaparkhi ‘98]
  – Segmentation [Beeferman, Berger & Lafferty ‘99]

• HMMs for similar language tasks
  – Part of speech tagging [Kupiec ‘92]
  – **Named entity recognition** [Bikel et al ‘99]
  – Information Extraction [Leek ‘97], [Freitag & McCallum ‘99]

• Serial Generative/Discriminative Approaches
  – Speech recognition [Schwartz & Austin ‘93]
  – Parsing [Collins, ‘00]

• Other conditional Markov models
  – Non-probabilistic local decision models [Brill ‘95], [Roth ‘98]
  – Gradient-descent on state path [LeCun et al ‘98]
  – Markov Processes on Curves (MPCs) [Saul & Rahim ‘99]
Part-of-speech Tagging

45 tags, 1M words training data

The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said.

Using spelling features*

<table>
<thead>
<tr>
<th></th>
<th>Error</th>
<th>oov error</th>
<th>error</th>
<th>( \Delta )</th>
<th>oov error</th>
<th>( \Delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>5.69%</td>
<td>45.99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF</td>
<td>5.55%</td>
<td>48.05%</td>
<td>4.27%</td>
<td>-24%</td>
<td>23.76%</td>
<td>-50%</td>
</tr>
</tbody>
</table>

* use words, *plus* overlapping features: capitalized, begins with #, contains hyphen, ends in -ing, -ogy, -ed, -s, -ly, -ion, -tion, -ity, -ies.
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
Person name Extraction

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 Gremillion, Madelaine Halmos, Elaine Henrik, 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 Type</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capitalized</td>
<td>Xxxxx</td>
<td>Character n-gram classifier says string is a person name (80% accurate)</td>
</tr>
<tr>
<td>Mixed Caps</td>
<td>XxXxxx</td>
<td>In stopword list (the, of, their, etc)</td>
</tr>
<tr>
<td>All Caps</td>
<td>XXXXX</td>
<td>In honorific list (Mr, Mrs, Dr, Sen, etc)</td>
</tr>
<tr>
<td>Initial Cap</td>
<td>X...</td>
<td>In person suffix list (Jr, Sr, PhD, etc)</td>
</tr>
<tr>
<td>Contains Digit</td>
<td>xxx5</td>
<td>In name particle list (de, la, van, der, etc)</td>
</tr>
<tr>
<td>All lowercase</td>
<td>xxxx</td>
<td>In Census lastname list; segmented by P(name)</td>
</tr>
<tr>
<td>Initial</td>
<td>X</td>
<td>In Census firstname list; segmented by P(name)</td>
</tr>
<tr>
<td>Punctuation</td>
<td>..;.;;!(), etc</td>
<td>In locations list (states, cities, countries)</td>
</tr>
<tr>
<td>Period</td>
<td>.</td>
<td>In company name list (&quot;J. C. Penny&quot;)</td>
</tr>
<tr>
<td>Comma</td>
<td>,</td>
<td>In list of company suffixes (Inc, Associates, Foundation)</td>
</tr>
<tr>
<td>Apostrophe</td>
<td>‘</td>
<td></td>
</tr>
<tr>
<td>Dash</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Preceded by HTML tag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hand-built FSM person-name extractor</td>
<td></td>
<td>says yes, (prec/recall ~ 30/90)</td>
</tr>
<tr>
<td>Conjunctions of all previous feature pairs</td>
<td>evaluated at the current time step.</td>
<td></td>
</tr>
<tr>
<td>Conjunctions of all previous feature pairs</td>
<td>evaluated at current step and one step ahead.</td>
<td></td>
</tr>
<tr>
<td>All previous features, evaluated two steps ahead.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All previous features, evaluated one step behind.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Training and Testing

- Trained on 65469 words from 85 pages, 30 different companies’ web sites.
- Training takes about 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 0.92 - 0.95, recall is 0.89 - 0.91.
<table>
<thead>
<tr>
<th>Match</th>
<th>by immersing gamers in total game environments says Adam Bosnjak, VP of Marketing for interSense, &quot;Everywhere.</th>
</tr>
</thead>
<tbody>
<tr>
<td>personname</td>
<td>January 2000 issue of PC World article writer Joel Strauch reports: &quot;Alex Garden, CEO of...</td>
</tr>
<tr>
<td>Match</td>
<td>World article writer Joel Strauch reports: &quot;Alex Garden, CEO of Relic Entertainment and creator of...</td>
</tr>
<tr>
<td>personname</td>
<td>all, our minds.&quot; • • • Joel Strauch • • • FC World January 2000 Out-of-</td>
</tr>
<tr>
<td>Match</td>
<td>based 3D games, such as Turok and Jedi Knight, bringing new levels of immersion and excitement...</td>
</tr>
<tr>
<td>personname</td>
<td>number of projects utilizing CAVE enabled technology. Dr. Ronald D. Kriz, Associate Professor in the Department of Engineering...</td>
</tr>
<tr>
<td>Match</td>
<td>new alloys with increased strength and ductility says Diana Farkas, Materials Science and Engineering. The CAVE...</td>
</tr>
<tr>
<td>personname</td>
<td>prototyping of interior components or entire environments says Joan McLain-Kark, Principal Investigator. Our particular interests...</td>
</tr>
<tr>
<td>Match</td>
<td>see while walking through the cathedral. says Dennis Jones, Principal Investigator. With architectural 3-...</td>
</tr>
<tr>
<td>personname</td>
<td>simulation of a robot arm for NASA. Professor Mike Deisenroth and Jeff Sugar, M.S.</td>
</tr>
<tr>
<td>Personname</td>
<td>Match</td>
</tr>
<tr>
<td>personname</td>
<td>Match</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Match</th>
<th>ext: ? in the role of Florence Taggart</th>
<th>Florence Taggart, Florence Taggart is the mother of Det, Marcus Taggart</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Match</th>
<th>ref: Florence Taggart is the mother of Det.</th>
<th>Marcus Taggart (Real Andrews). Moody will be</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Miss</th>
<th>ref: the mother of Det. Marcus Taggart (Real Andrews). Moody will be introduced on the</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Match</th>
<th>ref: GUIDING LIGHT Joining ? Guiding Light ? is</th>
<th>Jordi Vilasuso, who has been cast as Tony Santos</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Match</th>
<th>ref: Jordi Vilasuso, who has been cast as</th>
<th>Tony Santos. Santos is the reckless brother of Father</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Spurious</th>
<th>ref: Santos, Santos is the reckless brother of</th>
<th>Father Ray (George Alvarez)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Match</th>
<th>ref: is the reckless brother of Father Ray</th>
<th>George Alvarez, ? Guiding Light ? has made</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Partial Short</th>
<th>ref: have the character of Matt, played by</th>
<th>Kurt Robin McKinney, leave the show following the departure of</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Match</th>
<th>ref: leave the show following the departure of</th>
<th>Maeve Kinkaid, Kinkaid, who played McKinney ? s</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>personname</th>
<th>Match</th>
<th>ref: Also exiting ? Guiding Light ? is</th>
<th>Robert Lupone (Leo Flynn). Lupone left to</th>
</tr>
</thead>
</table>
## Person name Extraction

<table>
<thead>
<tr>
<th>Position</th>
<th>Name</th>
<th>E-mail Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>Gina Williams</td>
<td>gina <a href="mailto:oleksinski@prudential.com">oleksinski@prudential.com</a></td>
</tr>
<tr>
<td>VP Prof. Activities</td>
<td>Jennifer Bendetti</td>
<td><a href="mailto:jennifer_bendetti@hotmail.com">jennifer_bendetti@hotmail.com</a></td>
</tr>
<tr>
<td>VP Membership</td>
<td>Robert Bendetti</td>
<td><a href="mailto:robert_bendetti@hotmail.com">robert_bendetti@hotmail.com</a></td>
</tr>
<tr>
<td>Treasurer</td>
<td>Gary Wagner</td>
<td><a href="mailto:porterassociates@mindspring.com">porterassociates@mindspring.com</a></td>
</tr>
<tr>
<td>Secretary</td>
<td>Jeff Blanchard</td>
<td><a href="mailto:jeff_blanchard@wachovia.com">jeff_blanchard@wachovia.com</a></td>
</tr>
<tr>
<td>Dir. - Communications</td>
<td>Amy Becker</td>
<td><a href="mailto:abecker@yahoo.com">abecker@yahoo.com</a></td>
</tr>
<tr>
<td>Dir. - Community Service</td>
<td>WANTED</td>
<td>WANTED</td>
</tr>
<tr>
<td>Dir. - Fund Raising</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dir. - Social</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dir. - Social (brunches)</td>
<td>Bylli Daniels</td>
<td><a href="mailto:bylli_daniels@yahoo.com">bylli_daniels@yahoo.com</a></td>
</tr>
<tr>
<td>Web Chair</td>
<td>Carrie Sagel</td>
<td><a href="mailto:csagel@aaaofga.com">csagel@aaaofga.com</a></td>
</tr>
<tr>
<td>Dir. at Large</td>
<td>Van Phan</td>
<td><a href="mailto:vphan@dlalaw.com">vphan@dlalaw.com</a></td>
</tr>
<tr>
<td>Past President</td>
<td>T.J. Shriver</td>
<td><a href="mailto:t.j.dspek421@mindspring.com">t.j.dspek421@mindspring.com</a></td>
</tr>
<tr>
<td>District Dir.</td>
<td>Velvet Simmons</td>
<td><a href="mailto:velvet_simmons@nas.adp.com">velvet_simmons@nas.adp.com</a></td>
</tr>
</tbody>
</table>

To make any changes, additions, or deletions to the information above click here.
Local features, like formatting, exhibit regularity on a particular subset of the data (e.g. web site or document). Note that future data will probably not have the same regularities as the training data.

Global features, like word content, exhibit regularity over an entire data set. Traditional classifiers are generally trained on these kinds of features.
Scoped Learning
Generative Model

1. For each of the $D$ docs or sites:
   a) Generate the multinomial formatting feature parameters $\phi$ from $p(\phi|\alpha)$

2. For each of the $N$ words in the document:
   a) Generate the $n$th category $c_n$ from $p(c_n)$.
   b) Generate the $n$th word (global feature) from $p(w_n|c_n, \theta)$
   c) Generate the $n$th formatting feature (local feature) from $p(f_n|c_n, \phi)$

$$p(\phi, c, w, f) = p_\alpha(\phi) \prod_{n=1}^{N} p(c_n) p_\theta(w_n|c_n) p(f_n|c_n, \phi)$$
Inference

Given a new web page, we would like to classify each word resulting in \( \mathbf{c} = \{c_1, c_2, \ldots, c_n\} \)

\[
p(c|\mathbf{w}, \mathbf{f}) = \frac{\int \prod_{n=1}^{N} p(w_n|c_n)p(f_n|c_n, \phi)p(c_n)p(\phi) d\phi}{\int \prod_{n=1}^{N} \sum_{c_n} p(w_n|c_n)p(f_n|c_n, \phi)p(c_n)p(\phi) d\phi}
\]

This is not feasible to compute because of the integral and sum in the denominator. We experimented with two approximations:
- MAP point estimate of \( \phi \)
- Variational inference
MAP Point Estimate

If we approximate $\phi$ with a point estimate, $\hat{\phi}$, then the integral disappears and $c$ decouples. We can then label each word with:

$$\hat{c}_n = \arg \max_{c_n} p(w_n|c_n)p(f_n|c_n, \hat{\phi})p(c_n)$$

A natural point estimate is the posterior mode: a maximum likelihood estimate for the local parameters given the document in question:

$$\hat{\phi} = \arg \max_{\phi} p(\phi|f, w)$$

E-step:

$$p^{(t+1)}(c_n|w_n, f_n; \phi) \propto p^{(t)}(f_n|c_n; \phi)p(w_n|c_n)p(c_n)$$

M-step:

$$\hat{\phi}_{c,f} = p^{(t+1)}(f|c; \phi) \propto \sum_{\{n:c_n=c, f_n=f\}} p^{(t)}(c_n|f_n, w_n)$$
Job Title Extraction

Family Services Director
Creative, energetic and enjoy working with people? Seeking director for program development, implementation and administration. Must possess a Bachelor’s Degree in Recreation, Family Studies or related field. Strong interpersonal and organizational skills a must. Excellent benefits. Send resumes to Jane Kim, Dir of Camping and Family Services, North Suburban YMCA, 2705 Techny Road, Northbrook, IL 60062.

Massage Therapist - Male
The North Suburban YMCA is seeking a certified massage therapist to work part time in our men's program center. Flexible hours, y membership, on-site child care available if needed. Please contact Harlan Stritchko by email or call at 847-272-7250.

Starbucks Server
Early day, evening and weekend shifts available for in-house cafe serving the Starbucks product line. An exciting opportunity and membership is included! Contact Sarah Tucker at 847-272-7250 x213.

Teacher for ChildCare Center
Part-time 2-6 pm, Monday through Friday. Minimum requirements are 60 college credit hours in Early childhood or Education or similar subject. At least one year experience working with 2-5 year olds. Contact Helen at (847) 272-7250 x222 and fax resume to (847) 272-7587.

Art Coordinator
Creative? Enjoy working with children? the North suburban Y is looking for an art coordinator for the summer. Call Jane at (972) 272-7250 for more information.

Teachers
Seeking part-time early childhood teachers for summer or all year. 2-3 mornings per week from 9am-11:15am. Free child care on-site while you work. Free YMCA membership. College degree required in education or related field. Pick up an application at the front desk or call Caryn Shulman, Child Development Coordinator at (847) 272-7250 x232.

Group Exercise
Personal Training
Interested individuals with proper certification may contact Myleen Stanovict at (847) 272-7250 x 217

Customer Service Rep
OVERQUALIFIED APPLY HERE! Hone your skills by working in a friendly environment. The front desk is looking for part time staff to work flexible shifts for early weekday mornings, day and evening shifts. Benefits include YMCA membership and babysitting during your shift. Please contact Sarah Tucker or Cheryl Stewart at (847) 272-7250 x213.

Lifeguards and Swim Instructors
Love to swim? Love kids? Put the two together and make a difference. The North Suburban YMCA is looking for qualified and experienced swim instructors and lifeguards. Visit the main office at 9am-9pm daily or call Cheryl Stewart at (847) 272-7250 x213.
Job Title Extraction

**Family Services Director**
Creatively energetic and enjoy working with people? Seeking a director for program development, implementation, and administration. Must possess a Bachelor's Degree in Recreation, Family Studies, or related field. Strong interpersonal and organizational skills a must. Excellent benefits. Send resumes to Jane Kim, Dir of Camps and Family Services, North Suburban YMCA, 2705 Techny Road, Northbrook, IL 60062.

**Massage Therapist - Male**
The North Suburban YMCA is seeking a certified massage therapist to work part-time in our Men's program center. Flexible hours, y membership, on-site child care available if needed. Please contact Harlan Storchko by email or call at 847-272-7250.

**Starbucks Server**
Early day, evening and weekend shifts available for in-house cafe serving the Starbucks' product line. An exciting opportunity and membership is included. Contact Sarah Tucker at 847-272-7250 x 213.

**Teacher for ChildCare Center**
Part-time 2-6 pm, Monday through Friday. Minimum requirements are 60 college credit hours in Early Childhood Education or similar subject. At least one year experience working with 2-5 year olds. Contact Helen at (847) 272-7250 x 222 and fax resume to (847) 272-7587.

**Art Coordinator**
Creative? Enjoy working with children? The North suburban YM is looking for an art coordinator for the summer. Call Jane at (847) 272-7250 for more information.

**Teachers**
Seeking part-time early childhood teachers for summer or all year. 2-3 mornings per week from 9am-11:15am. Free child care on-site while you work. Free YMCA membership. College degree required in education or related field. Pick up an application at the front desk or call Caryn Shulman, Child Development Coordinator at (847) 272-7250 x 232.

**Group Exercise Personal Training**
Interested individuals with proper certification may contact Mylene Stanovit at (847) 272-7250 x 217.

**Customer Service Rep**
OVERQUALIFIED APPLY HERE! Hone your skills by working in a friendly environment. The front desk is looking for part-time staff to work flexible shifts for early week day mornings, day and evening shifts. Benefits include YMCA membership and babysitting during your shift. Please contact Sarah Tucker or Cheryl Stewart at (847) 272-7250 x 213.

**Lifeguards and Swim Instructors**
Love to swim? Love kids? Put the two together and make a difference. The North Suburban YMCA is looking for qualified and experienced swim instructors and lifeguards. For information contact Maryann Croce, Aquatics Director at (847) 272-7250 x 203.
Scoped Learning Related Work

- **Co-training [Blum & Mitchell 1998]**
  - Although it has no notion of scope, it also has an independence assumption about two independent views of data.

- **PRMs for classification [Taskar, Segal & Koller 2001]**
  - Extends notion of multiple views to multiple kinds of relationships
  - This model can be cast as a PRM where each locale is a separate group of nodes with separate parameters. Their inference corresponds our MAP estimate inference.

- **Classification with Labeled & Unlabeled Data [Nigam et al, 1999, Joachims 1999, etc.]**
  - Particularly *transduction*, in which the unlabeled set is the test set.
  - However, we model locales, represent a difference between local and global features, and can use locales at training time to learn hyper-parameters over local features.

- **Classification with Hyperlink Structure [Slattery 2001]**
  - Adjusts a web page classifier using ILP and a hubs & authorities algorithm.
Future Directions

• Feature selection and induction: automatically choose the $f_k$ functions (efficiently).

• Tree-structured Markov random fields for hierarchical parsing.

• Induction of finite state structure.

• Combine CRFs and Scoped Learning.

• Data mine the results of information extraction, and integrate the data mining with extraction.

• Create a text extraction and mining system that can be assembled and trained to a new vertical application by non-technical users.
Summary

• Conditional sequence models have the advantage of allowing complex dependencies among input features. (Especially good for extraction from the Web.)

• But they seemed to be prone to the label bias problem.

• CRFs are an attractive modeling framework that:
  – avoids label bias by moving from state normalization to global normalization,
  – preserves the ability to model overlapping and non-local input features,
  – has efficient inference & estimation algorithms,
  – converges to the global optima, because the likelihood surface is convex.

Papers on MEMMs, CRFs, Scoped Learning and more available at  http://www.cs.cmu.edu/~mccallum
End of talk