Maximum Entropy Markov Models for Information Extraction and Segmentation

Andrew McCallum

formerly *Just Research,* now *WhizBang Labs - Research*

Dayne Freitag

formerly *Just Research,* now *Burning Glass, Inc.*

Fernando Pereira

AT&T Labs - Research

Thanks also to

Kamal Nigam, John Lafferty and Michael Collins.

Summary

- Overall goal: Extracting information from text.
- Specific task: segment FAQ's into question/answer pairs.
- Problem: Sequential structure matters, but standard sequence models, such as HMMs, are not appropriate.
- Solution: A new probabilistic sequence model, closely related to HMMs, and based on maximum entropy.

Task: Extract *Speaker*, *Title*, *Location*, *Time*, *Date* from Email Seminar Announcement

Dr. Gibbons is spending his sabbatical from Bell Labs with us.
His work bridges databases, datamining and theory,
with several patents and applications to commercial DBMSs.

Christos

Date: Monday, March 20, 2000
Time: 3:30-5:00 (Refreshments provided)
Place: 4623 Wean Hall

Phil Gibbons
Carnegie Mellon University

The Aqua Approximate Query Answering System

In large data recording and warehousing environments, providing an exact answer to a complex query can take minutes, or even hours, due to the amount of computation and disk I/O required. Moreover, given the current trend towards data analysis over gigabytes, terabytes, and even petabytes of data, these query response times are increasing despite improvements in

Task: Extract name of *Acquired* and *Acquiring Company* from WSJ article about corporate acquisition

Sara Lee to Buy 30% of DIM

Chicago, March 3 - Sara Lee Corp said it agreed to buy a 30 percent interest in Paris-based DIM S.A., a subsidiary of BIC S.A., at cost of about 20 million dollars. DIM S.A., a hosiery manufacturer, had sales of about 2 million dollars.

The investment includes the purchase of 5 million newly issued DIM shares valued at about 5 million dollars, and a loan of about 15 million dollars, it said. The loan is convertible into an additional 16 million DIM shares, it noted.

The proposed agreement is subject to approval by the French government, it said.

Task: Extract *question/answer pairs* from FAQ

```
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 connections for common terminal
programs to work properly.  They are as far as I know the informal standard
agreed upon by commercial comms software developers for the Arc.

   Pins 1, 4, and 8 must be connected together inside the 9 pin plug.  This
is to avoid the well known serial port chip bugs.  The modem's DCD (Data
Carrier Detect) signal has been re-routed to the Arc's RI (Ring Indicator)
most modems broadcast a software RING signal anyway, and even then it's
really necessary to detect it for the model to answer the call.

2.7) The sound from the speaker port seems quite muffled.
     How can I get unfiltered sound from an Acorn machine?

   All Acorn machine are equipped with a sound filter designed to remove
high frequency harmonics from the sound output.  To bypass the filter, hook
into the Unfiltered port.  You need to have a capacitor.  Look for LM324 (chip
39) and and hook the capacitor like this:
```

Task: Extract *Title*, *Author*, *Institution* & *Abstract* from CS research paper

Journal of Artificial Intelligence Research 4 (1996) 237-285 Submitted 9/95; published 5/96

Reinforcement Learning: A Survey

Leslie Pack Kaelbling LPK@CS.BROWN.EDU
Michael L. Littman MLITTMAN@CS.BROWN.EDU
Computer Science Department, Box 1910, Brown University
Providence, RI 02912-1910 USA

Andrew W. Moore AWM@CS.CMU.EDU
Smith Hall 221, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

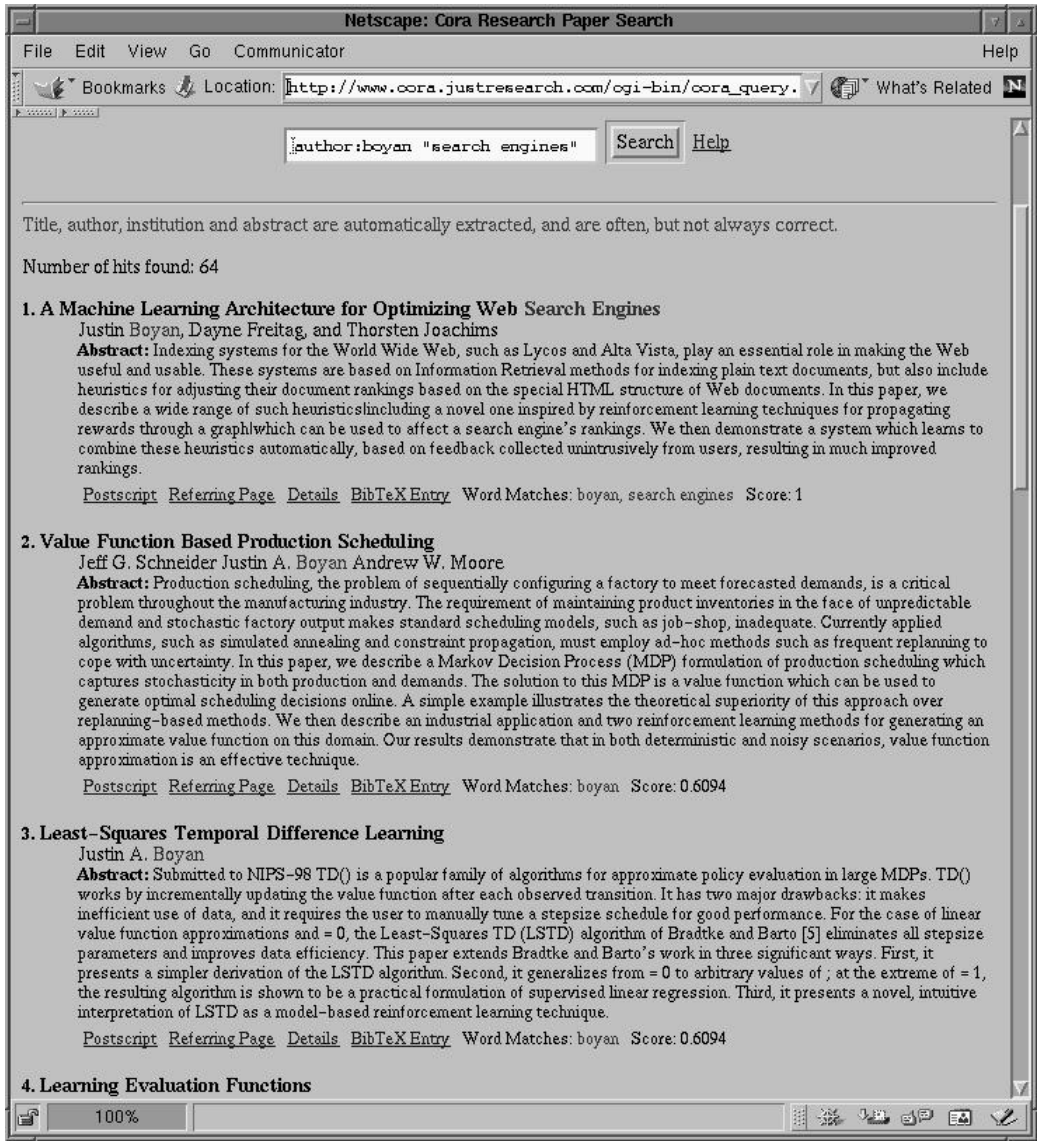
Abstract

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 a 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

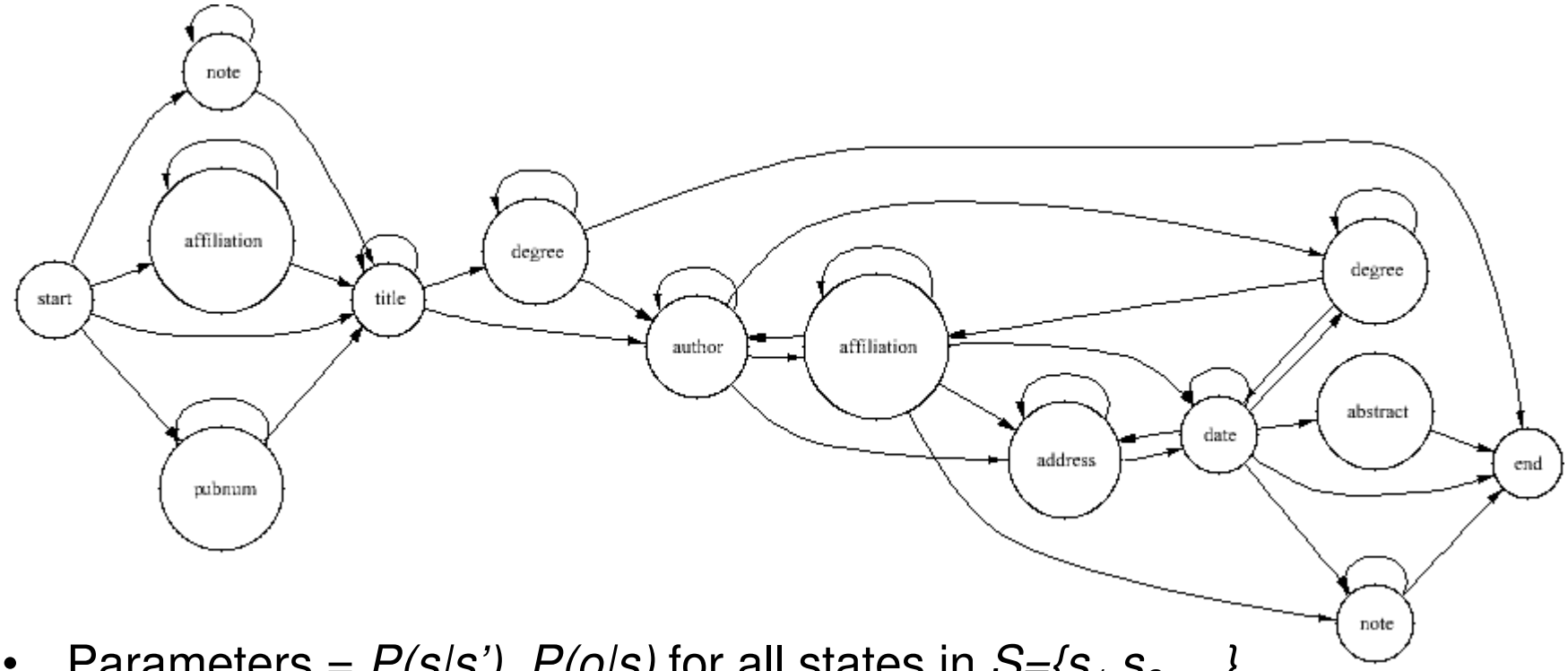
Reinforcement learning dates back to the early days of cybernetics and work in statistics, psychology, neuroscience, and computer science. In the last five to ten years, it has attracted rapidly increasing interest in the machine learning and artificial intelligence communities. Its promise is beguiling—a way of programming agents by reward and punishment without

author:boyan title:”search engines” abstract:HTML reference:”reinforcement learning”



Information Extraction with HMMs

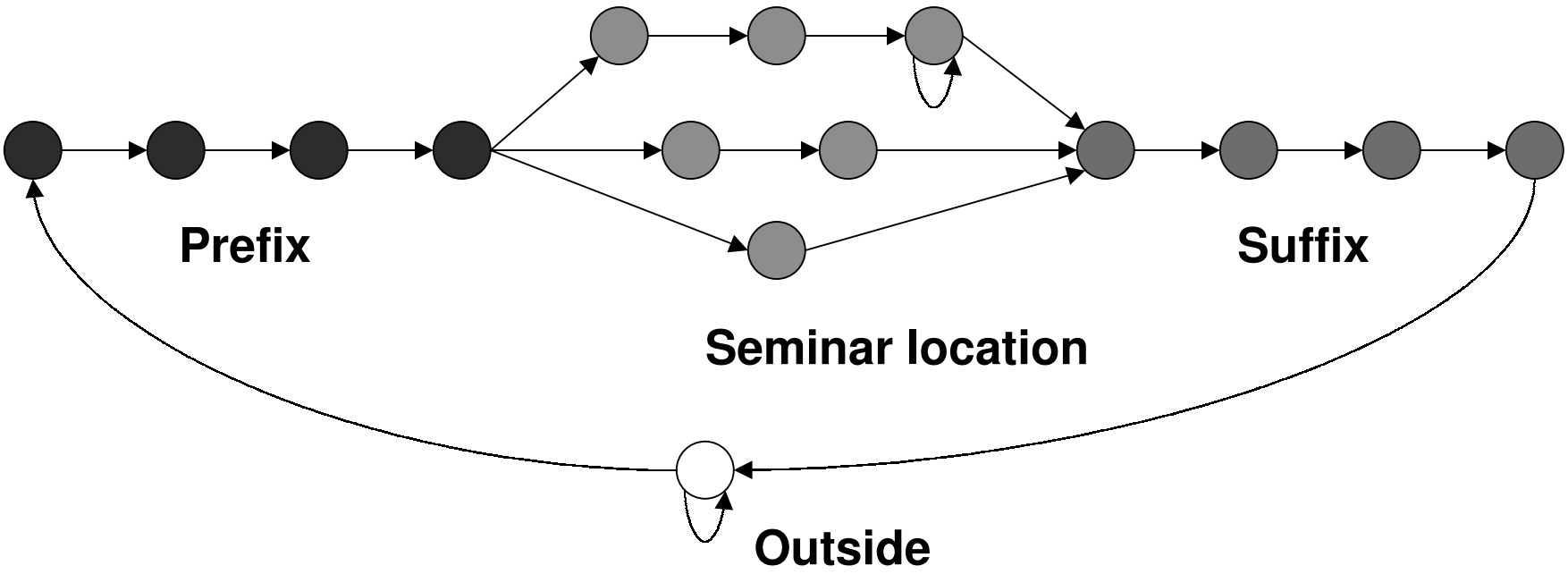
[Seymore & McCallum '99]



- Parameters = *P(s|s'), P(o|s)* for all states in $S=\{s_1,s_2,\ldots\}$
- Emissions = word
- Training = Estimate these parameters to maximize the probability of the training data.
- For IE, states indicate "field", or "database slot".

HMM to Extract Seminar Location

[Freitag & McCallum '99]



Regret #1

- HMMs are *generative* models of the material under analysis. (They calculate the probability of the model producing the sequence of words and states.) P({s…},{o…})

- But for IE, at test time, the sequence of words is *given*. P({s…}|{o…})

  We want a *conditional model*.

Regret #2

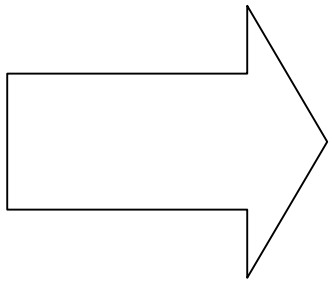
- Would prefer richer representation of text: multiple overlapping features, whole chunks of text.
- Example word features:
  - word identity
  - word is capitalized
  - word is in all caps
  - word ends in “-tion”
  - word is part of a noun phrase
  - word is in bold font
  - word is on left side of page
  - word is under node X in WordNet
- Example line features:
  - length of line
  - line is indented
  - line is centered
  - percent of non-alphabetics
  - total amount of white space
  - line contains two verbs
  - line begins with a number
  - line is grammatically a question
- Not compatible with generative model.

Solution:
New probabilistic sequence model

Traditional HMM

P(s|s')

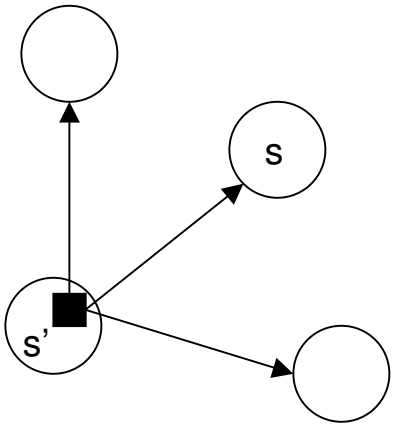
P(o|s)



Maximum Entropy Markov Model

P(s|o,s')

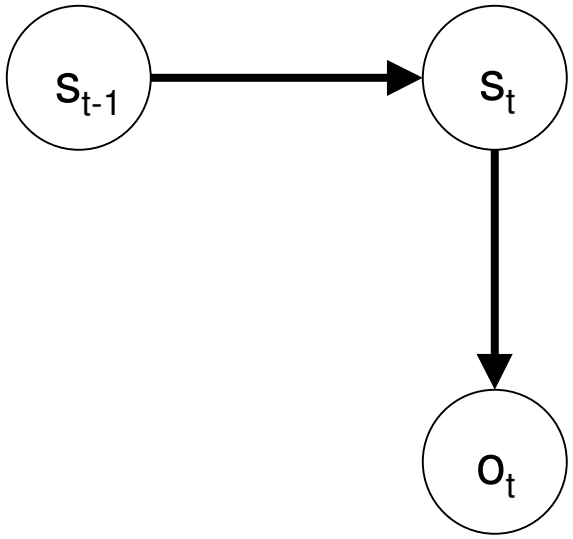
(represented by exponential model fit by maximum entropy)



For the time being, capture dependency on *s'* with /*S*/ independent functions, $P_{s'}(s|o)$

Each state contains a "next-state classifier" black box, that, given the next observation, will produce a probability distribution over possible next states, $P_{s'}(s|o)$.

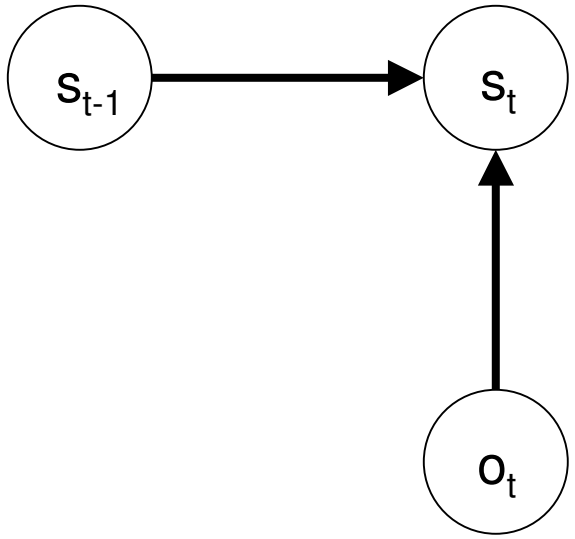
Old graphical model



P(o|s)

P(s|s')

New graphical model

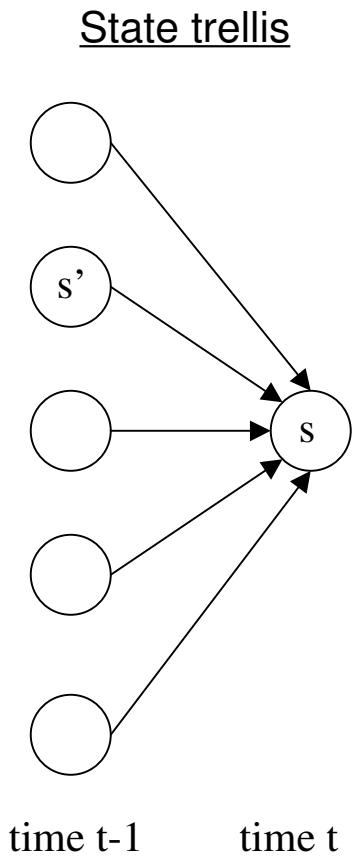


P(s|o,s')

Old "Forward Algorithm"

$\alpha_t(s)$ is the probability of being in state *s* at time *t* **and** producing the observation sequence up to time *t*.

$$\alpha_t(s) = \sum_{s' \in S} \alpha_{t-1}(s') \cdot P(s \mid s') \cdot P(o_t \mid s)$$



New "Forward Algorithm"

$\alpha_t(s)$ is the probability of being in state *s* at time *t* **given** the observation sequence up to time *t*.

$$\alpha_t(s) = \sum_{s' \in S} \alpha_{t-1}(s') \cdot P_{s'}(s \mid o_t)$$

"Backward Algorithm" analogous; Viterbi and Baum-Welch follow naturally.

State Transition Probabilities based on Overlapping Features

Model $P_{s'}(s|o)$ in terms of multiple arbitrary overlapping (binary) features.

Example observation feature tests:
- *o* is the word “apple”
- *o* is capitalized
- *o* is on a left-justified line

Actual feature, *f*, depends on both a binary observation feature test, *b*, and a destination state, *s*.

$$f_{<b,s>}(o_t, s_t) = \begin{cases} 1 & \text{if } b(o_t) \text{ is true and } s = s_t \\ 0 & \text{otherwise} \end{cases}$$

$P_{s'}(s|o)$ is some function of these features:

$$P_{s'}(s \mid o) = Q(f_{<b_1,s_1>}(o,s), f_{<b_1,s_2>}(o,s), f_{<b_2,s_1>}(o,s), f_{<b_2,s_2>}(o,s), \ldots)$$

Maximum Entropy Principle

Maximum entropy is based on the principle that the best model for the data is the one that is consistent with certain constraints derived from the training data, but otherwise makes the fewest possible assumptions.

Constraints we choose: that each feature, $f_{<b,s>}$, have the same expected value in the learned distribution as it does in the training data:

$$\overbrace{\frac{1}{m_{s'}}\sum_{k=1}^{m_{s'}} f_{<b,s>}(o_{t_k}, s_{t_k})}^{E_{data}[f_{<b,s>}]} = \overbrace{\frac{1}{m_{s'}}\sum_{k=1}^{m_{s'}}\sum_{s\in S} P_{s'}(s \mid o_{t_k}) f_{<b,s>}(o_{t_k}, s)}^{E_{P_{S'}(s|o)}[f_{<b,s>}]}$$

$$\text{where } t_1,...,t_{m_{s'}} \text{ are the time steps with } s_{t_k} = s'$$

"Fewest possible assumptions" in our probabilistic framework means the distribution "that is closest to the uniform distribution", i.e., the distribution that has maximum entropy.

Exponential Form

When constraints are imposed in this way, the constraint-satisfying probability distribution that has maximum entropy is guaranteed to be:

[Della Pietra, Della Pietra, Lafferty, '97]

(1) unique
(2) the same as the maximum likelihood solution
(3) in exponential form:

$$P_{s'}(s \mid o) = \frac{1}{Z(o, s')} \exp\left(\sum_{<b,s>} \lambda_{<b,s>} f_{<b,s>}(o, s) \right)$$

(Analogous to naïve Bayes, where

$$\lambda_{<b,s>} \approx \log(P(b \mid s)) \approx \log(P(w \mid c))$$

$$f_{<b,s>}(o, s) \approx \text{the number of times word } b \text{ occurs in test document.}$$

but exp(λ)'s don't have to sum to one!)

Generalized Iterative Scaling

[Darroch & Ratcliff '72]

A standard algorithm for finding feature weights λ_a in the exponential model.

For each state, learn the parameters of its "black box" next state function, $P_{s'}(s|o)$, as follows:

1. Calculate the relative frequency of each feature in the training data to get the expected value of each feature "according to the training data":

$$F_{<b,s>} = \frac{1}{m_s} \sum_{k=1}^{m_s} f_{<b,s>}(o_{t_k}, s_{t_k})$$

2. Start iteration 0 of GIS arbitrary parameter values, say $\lambda_{<b,s>}^{(0)} = 1$

3. At iteration *j*, use current $\lambda_{<b,s>}^{(j)}$ values in $P_{s'}^{(j)}(s \mid o)$ to calculate expected value of each feature "according to the model".

$$E_{<b,s>}^{(j)} = \frac{1}{m_s} \sum_{k=1}^{m_s} \sum_{s \in S} P_{s'}^{(j)}(s \mid o) f_{<b,s>}(o_{t_k}, s)$$

4. Update parameters to make the expected value of each feature "according to the model" be closer to the expected value "according to the training data"

$$\lambda_{<b,s>}^{(j+1)} = \lambda_{<b,s>}^{(j)} + \frac{1}{C} \log\left(\frac{F_{<b,s>}}{E_{<b,s>}^{(j)}} \right)$$

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 part, 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

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 entopy Markov model described in this talk.

Results

Learner	*Segmentation precision*	*Segmentation recall*
ME-Stateless	0.038	0.362
TokenHMM	0.276	0.140
FeatureHMM	0.413	0.529
MEMM	0.867	0.681

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]*
- Other conditional Markov models
  - Markov Processes on Curves (MPCs) *[Saul & Rahim '99]*

Variation #1:
Factored State Representation

The problem: Certain kinds of memories result in a multiplicative explosion in the number of states.

A solution: Factored HMMs [e.g. Ghahramani & Jordan '96].

Our variation: We can do the same in a less complex way:
Don't split P(s|s',o) into |S| different functions $P_{s'}(s|o)$; instead, represent state as a collection of features, just like the observations.

Example state features:
- "we have already consumed the start-time extraction field"
- "we haven't yet exited the preamble of the document"
- "the subject of the previous sentence is female"
- "the last paragraph was an question"

Variation #2:
Observations in states instead of transitions

The problem: Sparse training data---especially since MEMMs put observations "in the transitions" instead of "in the states".

Our solution: Put observations back "in the states". Have a single maximum entropy classifier, instead of a separate one for each source state.

$$P(s \mid s', o) = P(s \mid s') \frac{1}{Z(o, s')} \exp\left(\sum_{<b,s>} \lambda_{<b,s>} f_{<b,s>}(o, s) \right)$$

(This method of "correcting" a simple multinomial or prior by adding extra features with maximum entropy has been used previously [e.g. Rosenfeld 94].)

Variation #3:
Add actions to make a POMDP

The problem: The advantages of overlapping observation features and the desire for factored state representation also appear in reinforcement learning.

Our solution: Add features for actions also. Additionally, can have a factored representation of actions.

$$P(s \mid s', o, a)$$

Conclusions

- Presented a new probabilistic sequence model based on maximum entropy.
  - Conditional model
  - Handles arbitrary overlapping features
- Shown positive experimental results on FAQ segmentation.
- Shown variations for factored state, reduced complexity model, and reinforcement learning.