Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data

Paper by John Lafferty, Andrew McCallum, and Fernando Pereira

ICML 2001

Presentation by Joe Drish

May 9, 2002

Main Goals

- Present a new framework for labeling sequence data: Conditional Random Fields (CRFs)

- Describe the label bias problem

- Motivate the use of CRFs to solve the label bias problem

- Define the structure and properties of CRFs

- [Describe two training procedures for learning CRF parameters]

- [Sketch a proof of the convergence of the two training procedures]

- Compare experimentally to hidden Markov models (HMMs) and maximum entropy Markov models (MEMMs)

Labeling Sequence Data

- Given observed data sequences $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
- A corresponding label sequence $\mathbf{y}_k$ for each data sequence $\mathbf{x}_{k,}$, and $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$

Prediction Task

Given a sequence $\mathbf{x}$ and model θ predict $\mathbf{y}$

Learning Task

Given training sets X and Y, learn the best model θ

Notation

set of sequences: X, uppercase
sequence of observations: $\mathbf{x}$, lowercase boldface, also called a data sequence
sequence of labels: $\mathbf{y}$, lowercase boldface, also called a sequence of states
single observation: x, lowercase, also called a data value

Example label and observation sequence

| label y		observation x
`<head>`		`X-NNTP-Poster: NewsHound v1.33`
`<head>`		
`<head>`		`Archive-name: acorn/faq/part2`
`<head>`		`Frequency: monthly`
`<head>`		
`<question>`		`2.6) What configuration of serial cable should`
`<question>`		`I use?`
`<answer>`		`Here follows a diagram of the necessary`
`<answer>`		`connections programs to work properly. They`
`<answer>`		`are as far as know agreed upon by commercial`
`<answer>`		`comms software developers fo`
`<answer>`		
`<answer>`		`Pins 1, 4, and 8 must be connected together`
`<answer>`		`is to avoid the well known serial port chip bugs.`

label sequence **y**   observation sequence **x**

4

Generative Modeling (HMMs)

Given training set X with label sequences Y:

- Train a model θ that maximizes $P(X, Y \mid \theta)$
- For a new data sequence **x**, the predicted label **y** maximizes $P(\mathbf{y} \mid \mathbf{x}) = P(\mathbf{y} \mid \mathbf{x}, \theta)P(\mathbf{x} \mid \theta)$
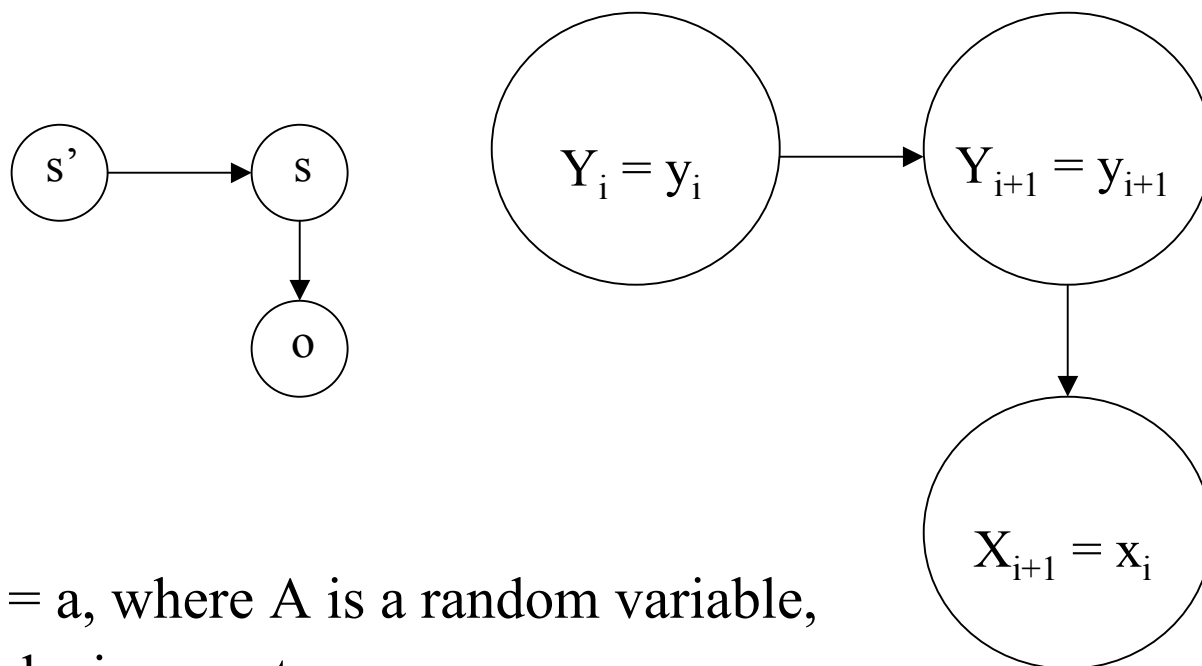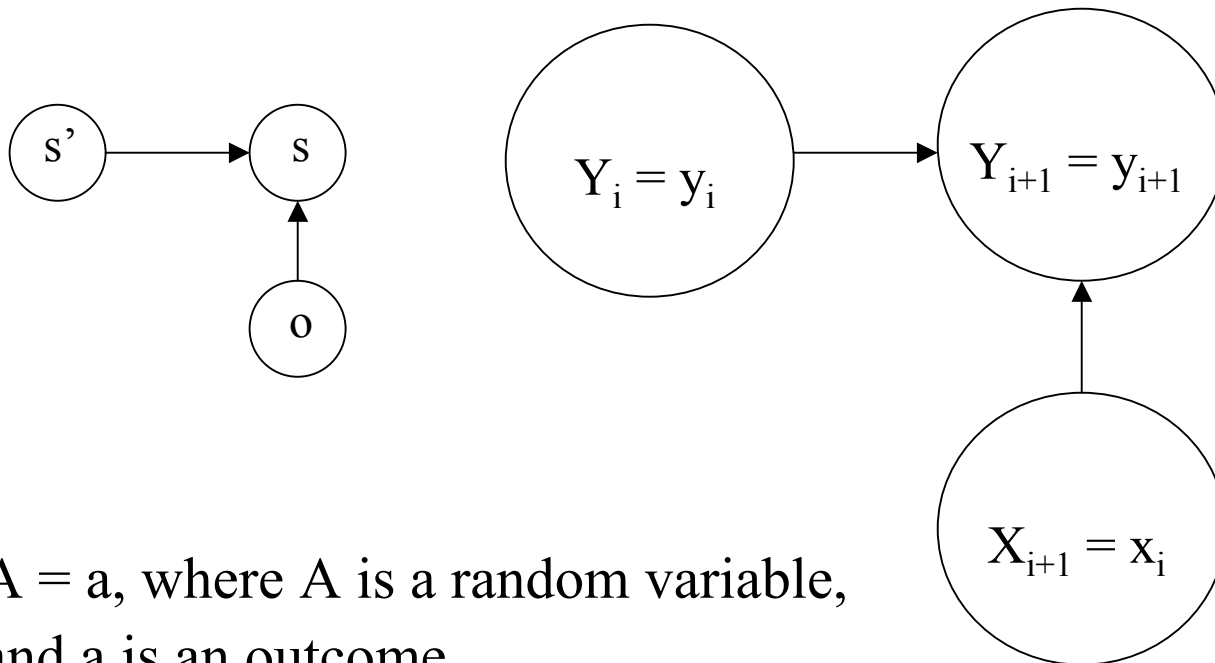

$A = a$, where A is a random variable, and a is an outcome

Discriminative Modeling (MEMMs)

Given training set X with label sequences Y:

- Train a model θ that maximizes $P(Y \mid X, \theta)$
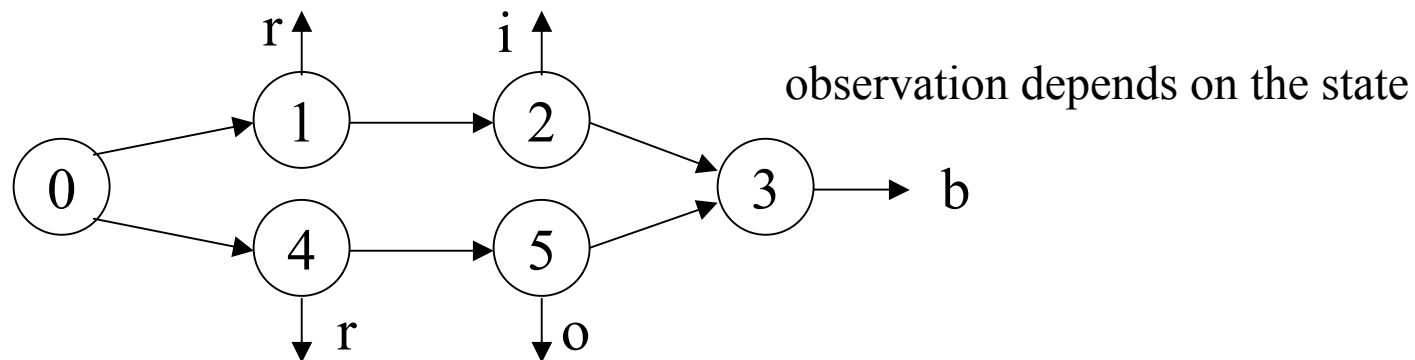- For a new data sequence **x**, the predicted label **y** maximizes $P(\mathbf{y} \mid \mathbf{x}, \theta)$

s' → s

o → s

$Y_i = y_i$ → $Y_{i+1} = y_{i+1}$

$X_{i+1} = x_i$ → $Y_{i+1} = y_{i+1}$

$A = a$, where A is a random variable, and a is an outcome

rib/rob models

Training data: {<rib, 123>, <rob, 453>}

HMM

observation depends on the state

MEMM
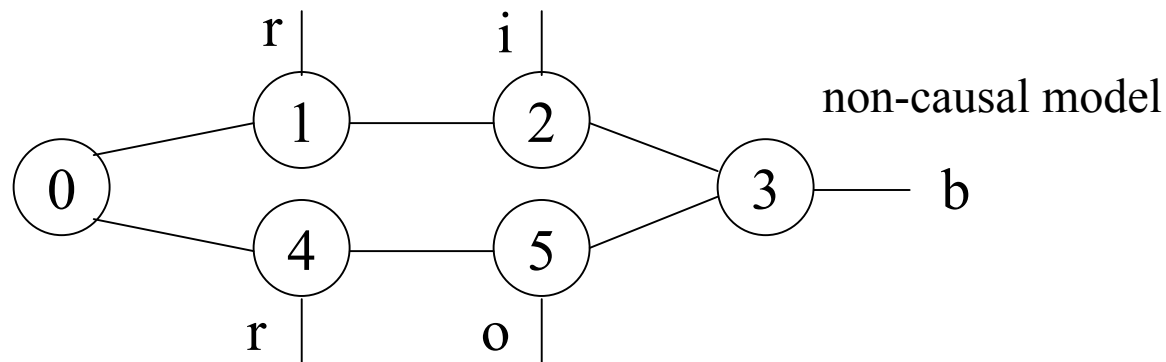
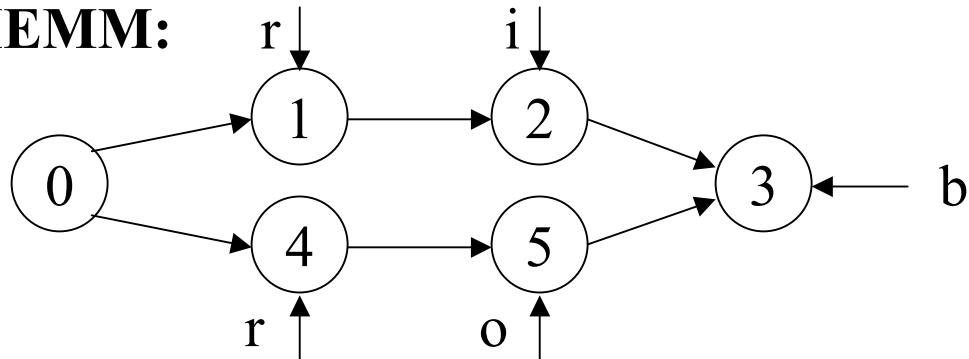state depends on the observation

CRF

non-causal model
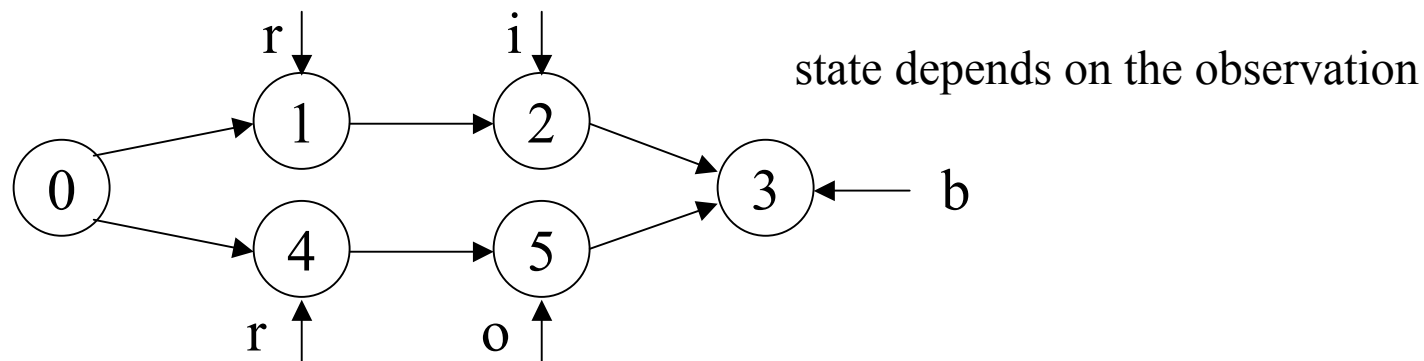
Label Bias Problem

Consider this MEMM:

The label sequence 1,2 should score higher when ri is observed compared to ro.
Or, we expect P(1 and 2 | ri) to be greater than P(1 and 2 | ro).

Mathematically,
P(1 and 2 | ro) = P(2 | 1 and ro)P(1 | ro) =   P(2 | 1 and o)P(1 | r)
P(1 and 2 | ri) =   P(2 | 1 and ri)P(1 | ri) =   P(2 | 1 and i)P(1 | r)

Since P(2 | 1 and x) = 1 for all x, P(1 and 2 | ro) = P(1 and 2 | ri)
In the training data, label value 2 is the only label value observed after label value 1
Therefore P(2 | 1) = 1, so P(2 | 1 and x) = 1 for all x

Example:

$$P(14 \text{ and } 2 \mid ri) = \quad P(2 \mid 14 \text{ and } ri)P(14 \mid ri) = \quad P(2 \mid 14 \text{ and } i)P(14 \mid r) = (1)(1) = 1$$
$$P(14 \text{ and } 2 \mid ro) = P(2 \mid 14 \text{ and } ro)P(14 \mid ro) = P(2 \mid 14 \text{ and } o)P(14 \mid r) = (0)(1) = 0$$

This is a solution to the label bias problem.

But, changing the set of states would be impractical.

Conditional Random Fields (CRFs)

Disadvantages of MEMMs

P($\mathbf{y}$ | $\mathbf{x}$) = product of factors, one for each label

Each factor can depend only on previous label, and not future labels

So, let

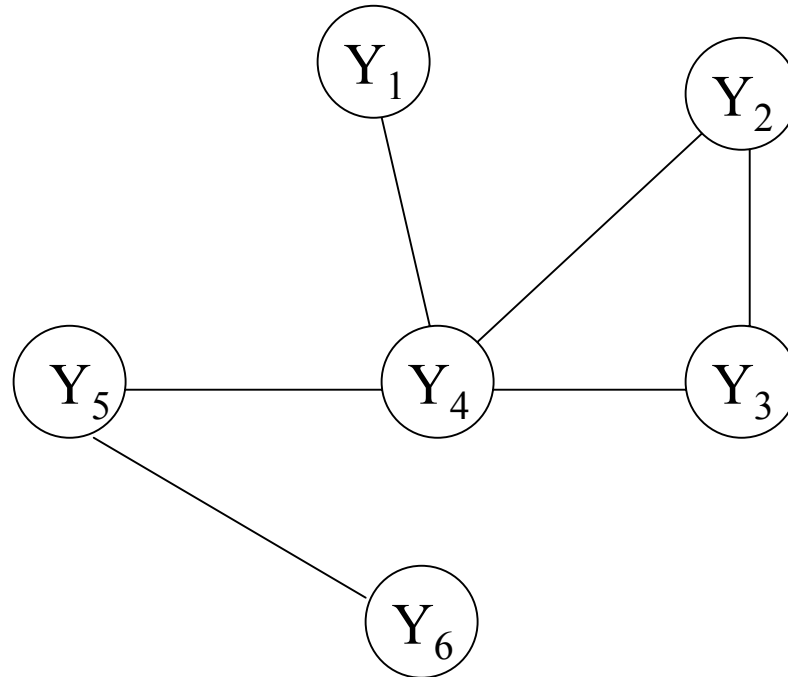$$P(\mathbf{y} \mid \mathbf{x}) = \exp\left(\sum_k f_k(\mathbf{y}, \mathbf{x})\right)$$

where each f_k is a property of part of $\mathbf{y}$ and $\mathbf{x}$

Example: $f_k(\mathbf{x}, \mathbf{y}) = 1$ is X_i is uppercase and label Y_i is a proper noun.

Random Field Example

Let G = (Y, E) be a graph where each vertex Y_v is a random variable
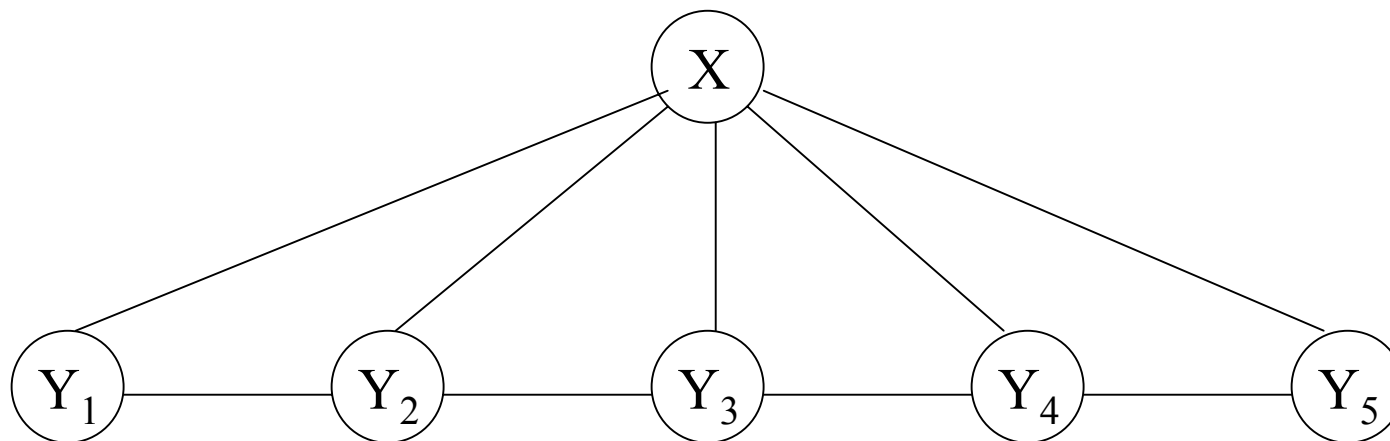Suppose $P(Y_v \mid$ all other $Y) = P(Y_v \mid$ neighbors$(Y_v))$ then Y is a random field

Example:

• $P(Y_5 \mid$ all other $Y) = P(Y_5 \mid Y_4, Y_6)$

Conditional Random Field Example

Suppose $P(Y_v \mid X$, all other $Y) = P(Y_v \mid X,$ neighbors$(Y_v))$
then X with Y is a **conditional** random field

- $P(Y_3 \mid X$, all other $Y) = P(Y_3 \mid X, Y_2 , Y_4)$
- Think of X as observations and Y as labels

Conditional Distribution

If Y is a tree, the distribution over the label sequence Y = y, given X = x, is:

$$p_\Theta(\mathbf{y} \mid \mathbf{x}) \propto$$

$$\exp\left(\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}\mid_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}\mid_v, \mathbf{x}) \right) \quad (1)$$

- $\mathbf{x}$ is a data sequence outcome
- $\mathbf{y}$ is a label sequence outcome
- v is a vertex from vertex set V = set of label random variables
- e is an edge from edge set E over V
- f_k and g_k are given and fixed features; each g_k is a property of x and a vertex v for the label random variable associated with v.
- k is the number of features
- $\Theta = (\lambda_1, \lambda_2, \ldots; \mu_1, \mu_2, \ldots)$; λ_k and μ_k are parameters to be estimated
- $y\mid_e$ is the components of y defined by edge e
- $y\mid_v$ is one component of y defined by vertex v

Matrix Random Variable

- Add special start and stop states $\mathbf{y}_0$ = start and $\mathbf{y}_{n+1}$ = start
- e_i is the edge with labels ($\mathbf{Y}_{i-1}$, $\mathbf{Y}_i$)
- v_i is the vertex with label $\mathbf{Y}_i$
- curly Y = set of possible label values
- The matrix random variable has the range |curly Y| ×|curly Y|
- If $Y_i = y_i$ then $y_i \in$ curly Y

Suppose that $p_\Theta(Y \mid X)$ is a CRF given by (1). Assume Y is a chain. For each position i in the observation sequence $\mathbf{x}$, we define a matrix random variable $M_i(\mathbf{x}) = [M_i(y', y \mid \mathbf{x})]$ as:

Y_{i-1} Y_i Y_{i+1}

X_{i-1} X_i X_{i+1}

$$M_i(y', y \mid \mathbf{x}) = \exp(\Lambda_i(y', y \mid \mathbf{x}))$$

$$\Lambda_i(y', y \mid \mathbf{x}) = \sum_k \lambda_k f_k(e_i, \mathbf{Y}\mid_{e_i}, Y\mid_{e_i} = (y', y), \mathbf{x}) +$$

$$\sum_k \mu_k g_k(v_i, \mathbf{Y}\mid_{v_i} = y, \mathbf{x})$$
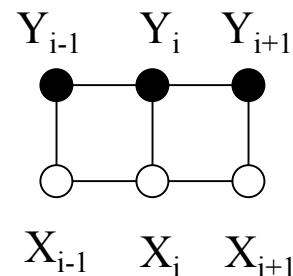
Conditional Probability for CRFs

The conditional probability of a label sequence **y** is

$$p_\Theta(\mathbf{y} \mid \mathbf{x}) = \frac{\prod_{i=1}^{n+1} M_i(\mathbf{y}_{i-1}, \mathbf{y}_i \mid \mathbf{x})}{\left(\prod_{i=1}^{n+1} M_i(\mathbf{x})\right)_{\mathrm{start,\,stop}}}$$

- n is length of sequence y = $y_1 \dots y_n$
- y_0 = start and y_{n+1} = stop

Parameter Estimation

Input: training data D = {(x$^{(i)}$, y$^{(i)}$)}, where i = 1…N with empirical distribution $\widetilde{p}\,(\,x\,,\,y\,)$

Output: parameters $\Theta = (\lambda_1, \lambda_2, …; \mu_1, \mu_2, …)$

Maximize: the log-likelihood objective function:

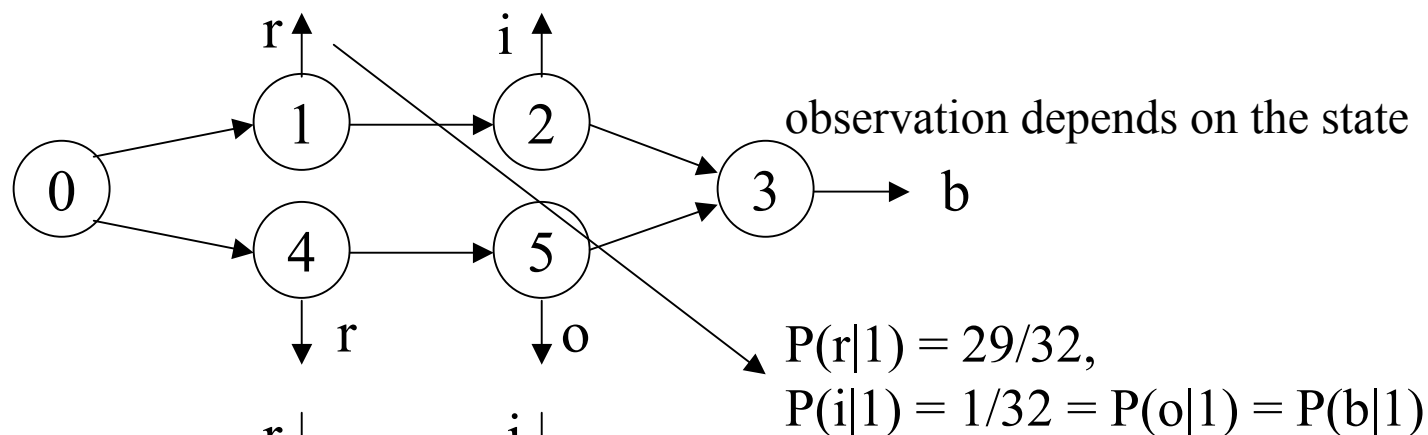$$O(\Theta) \quad = \quad \sum_{i=1}^{N} \log\, p_{\Theta}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$$

$$\propto \quad \sum_{\mathbf{x},\mathbf{y}} \widetilde{p}(\mathbf{x},\mathbf{y}) \log\, p_{\Theta}(\mathbf{y} \mid \mathbf{x})$$

- This is an expectation.

rib/rob models

Training data {<rib, 123>, <rob, 453>}

HMM

observation depends on the state

$P(r|1) = 29/32,$
$P(i|1) = 1/32 = P(o|1) = P(b|1)$

MEMM

state depends on the observation

CRF

non-causal model
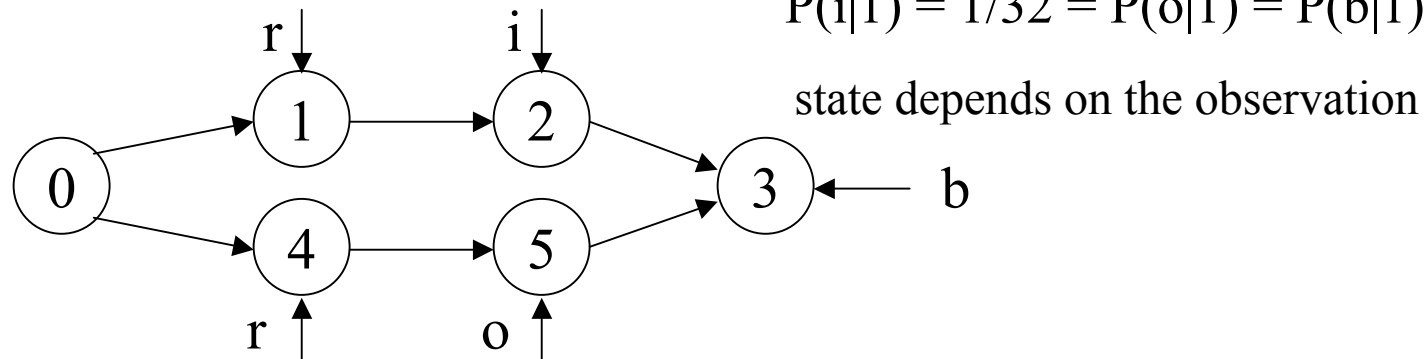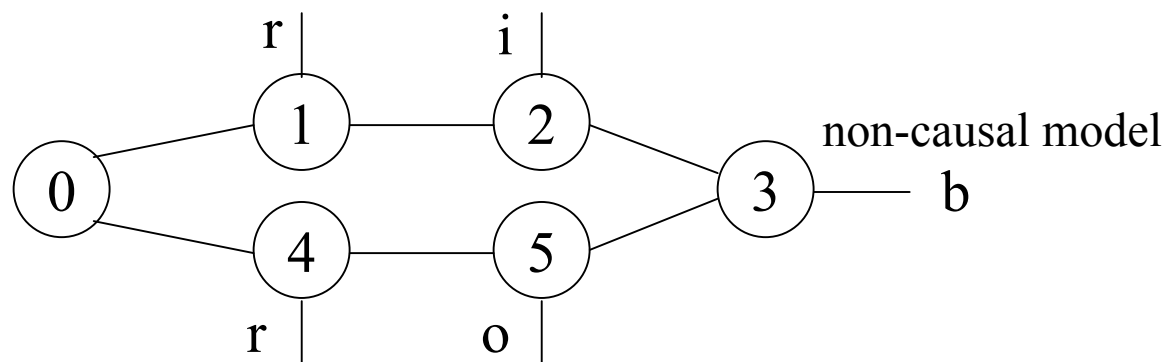
Modeling the label bias problem

- Each state emits its designated symbol with probability 29/32 and the other symbols with probability 1/32 (on picture)
- They train MEMM and CRF with the same topologies
- A run consists of 2,000 training examples and 500 test examples, trained to convergence
- CRF error is 4.6%, and MEMM error is 42%
- MEMM fails to discriminate between the two branches

Mixed-order HMM

State transition probabilities are given by

$$p_\alpha(\mathbf{y}_i \mid \mathbf{y}_{i-1}, \mathbf{y}_{i-2}) = \alpha p_2(\mathbf{y}_i \mid \mathbf{y}_{i-1}, \mathbf{y}_{i-2}) + (1-\alpha) p_1(\mathbf{y}_i \mid \mathbf{y}_{i-1})$$

Emission probabilities are given by

$$p_\alpha(\mathbf{x}_i \mid \mathbf{y}_i, \mathbf{x}_{i-1}) = \alpha p_2(\mathbf{x}_i \mid \mathbf{y}_i, \mathbf{x}_{i-1}) + (1-\alpha) p_1(\mathbf{x}_i \mid \mathbf{y}_i)$$

$\alpha = 0$ is a standard first-order HMM.

A first –order HMM has transition probabilities given by

$$p_\alpha(\mathrm{y}_i \mid \mathrm{y}_{i-1}, \mathrm{y}_{i-2}) = (1-\alpha) p_1(\mathrm{y}_i \mid \mathrm{y}_{i-1})$$

And emission probabilities given by

$$p_\alpha(\mathrm{x}_i \mid \mathrm{y}_i, \mathrm{x}_{i-1}) = (1-\alpha) p_1(\mathrm{x}_i \mid \mathrm{y}_i)$$

Modeling mixed-order sources

Experimental Setup

- For each randomly generated model, a sample of 1,000 sequences of length 25 is generated for training and testing

- On each randomly generated training set, a CRF is trained using Algorithm S, which is described in the paper

- On the same data an MEMM is trained using iterative scaling

- The Viterbi algorithm is used to label a test set

- *the advantages of the additional representational power of CRFs and MEMMs relative to HMMs.*
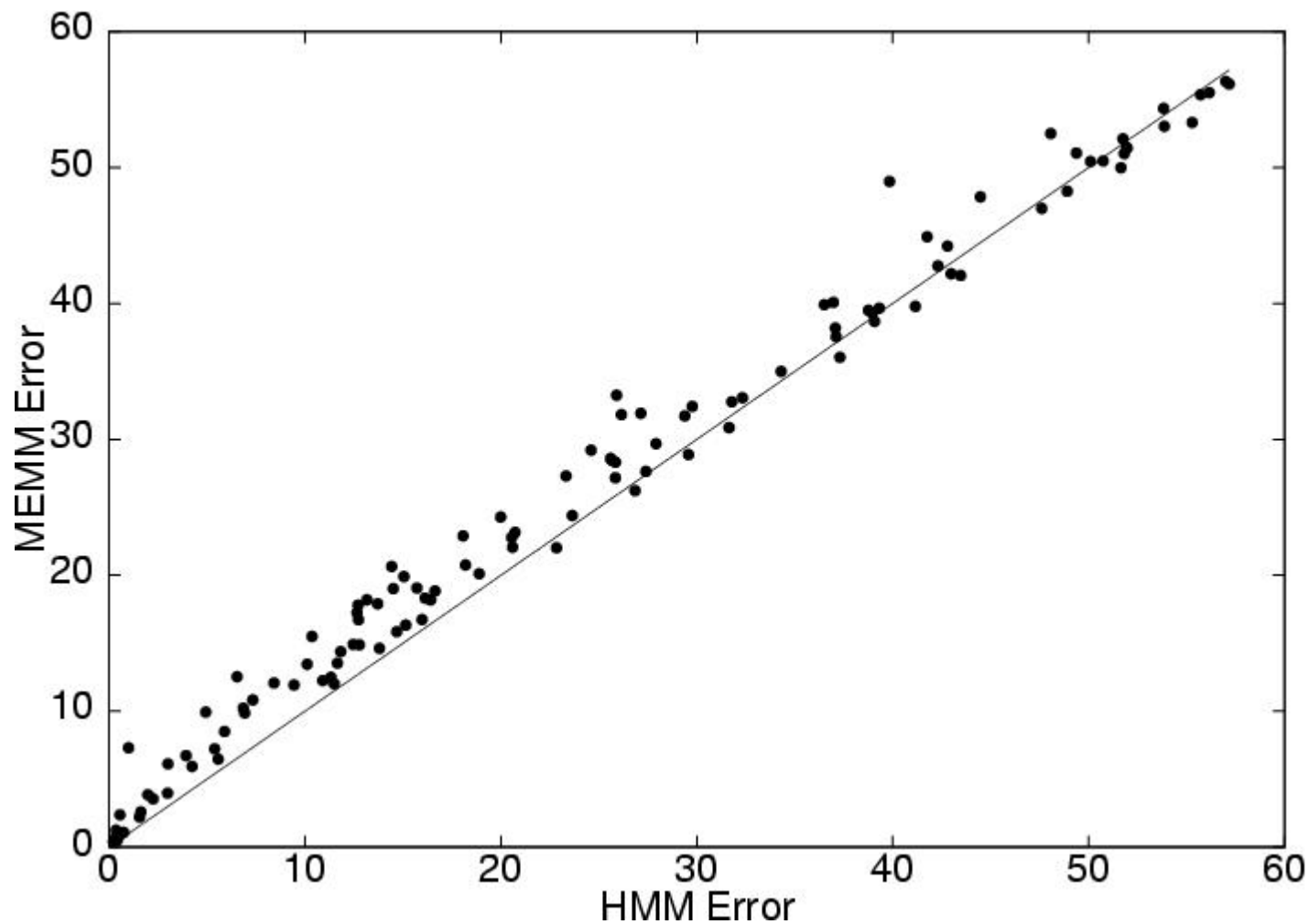
MEMM versus CRF

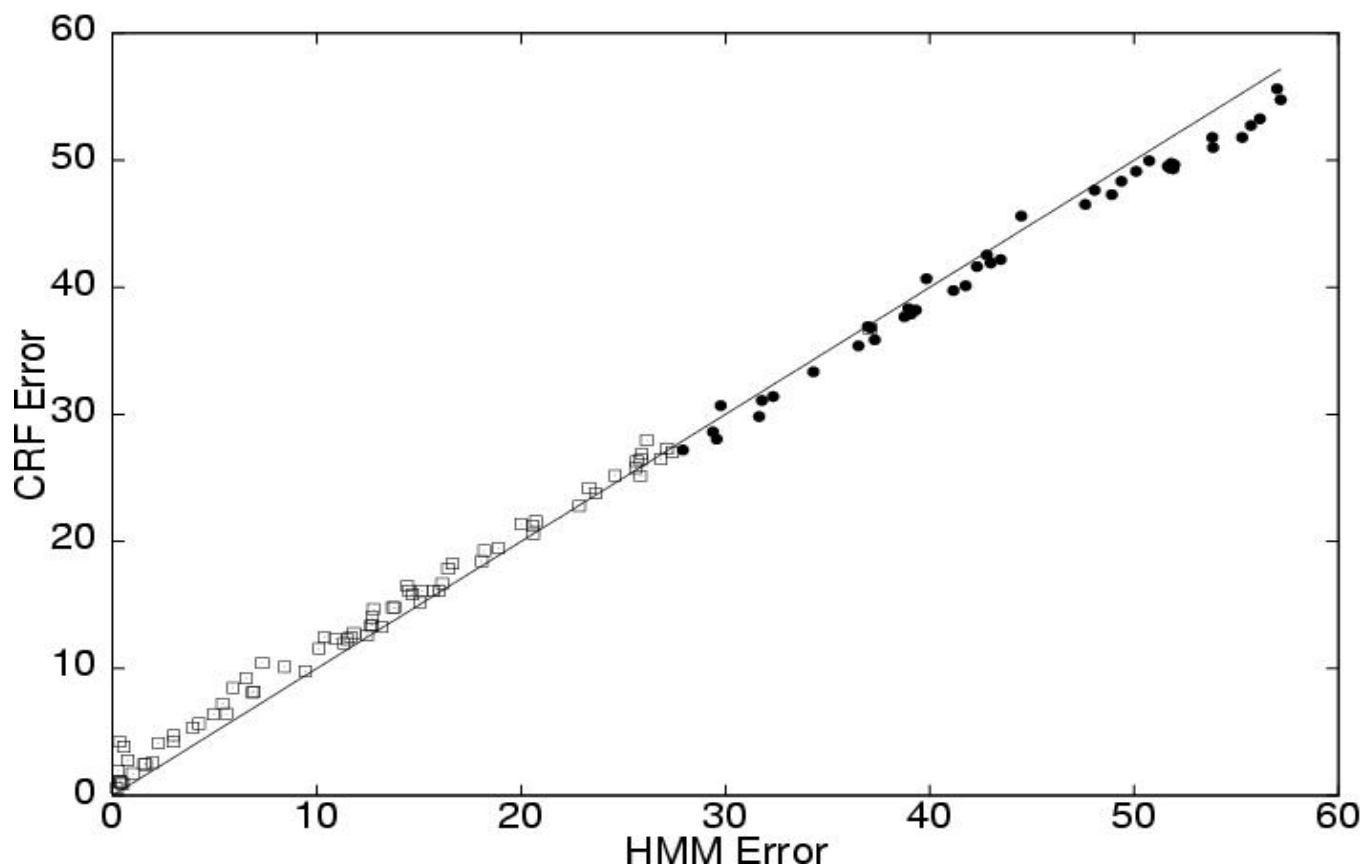- CRF usually outperforms the MEMM
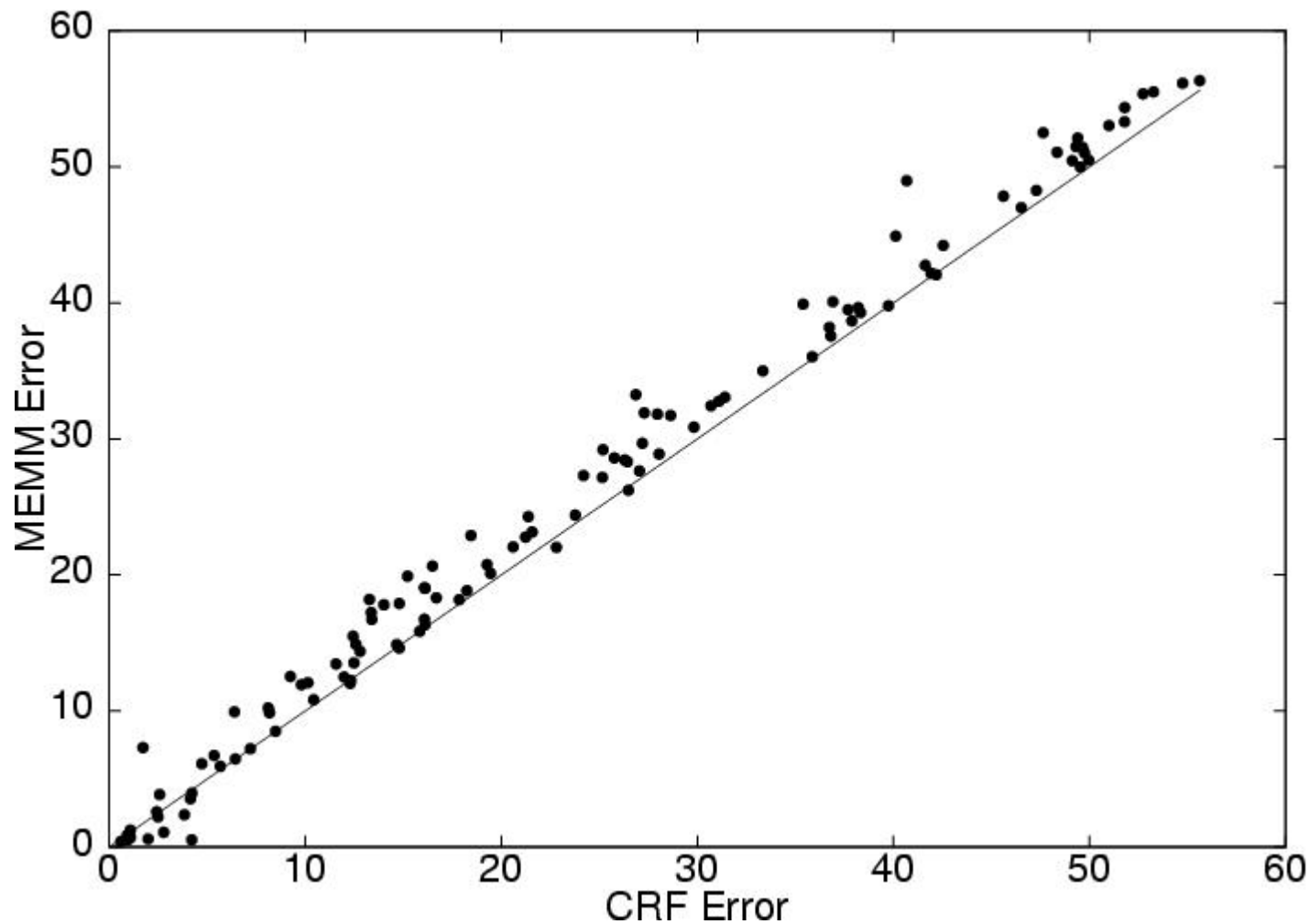
MEMM versus HMM

- The HMM outperforms the MEMM

CRF versus HMM

Each open square represents a data set with $\alpha < 1/2$, and a solid circle indicates a data set with $\alpha \geq 1/2$; When the data is mostly second order ($\alpha \geq 1/2$), the discriminatively trained CRF usually outperforms the HMM

UPenn Tagging Task

- 45 tags (syntactic), 1M words training

DT	NN	NN	NN	VBZ	RB	JJ
The	**asbestos**	**fiber ;**	**crocidolite ;**	**is**	**unusually**	**resilient**

IN	PRP	VBZ	DT	NNS	IN	RB	JJ	NNS
once	**it**	**enters**	**the**	**lungs ;**	**with ;**	**even**	**brief**	**exposures**

TO	PRP	VBG	NNS	WDT	VBP	RP	NNS	JJ
to	**it**	**causing**	**symptoms**	**that**	**show**	**up**	**decades**	**later ;**

NNS	VBD
researchers	**said**

POS tagging experiment 1

- Compared HMMs, MEMMs, and CRFs on Penn treebank POS tagging
- Trained first-order HMM, MEMM, and CRF models as in the synthetic data experiments
- Introduced parameters $\mu_{y,x}$ for each tag-word pari and $\lambda_{y'y}$ for each tag-tag pari in the training set
- oov = out-of-vocabulary, and are not observed in the training set

model	error	oov error
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%

POS tagging experiment 2

- Compared HMMs, MEMMs, and CRFs on Penn treebank POS tagging
- Each word in a given input sentence must be labeled with one of 45 syntactic tags
- Add a small set of orthographic features: whether a spelling begins with a number or upper case letter, whether it contains a hyphen, and if it contains one of the following suffixes: -ing, -ogy, -ed, -s, -ly, -ion, -tion, -ity, -ies

			using spelling features			
model	error	oov error	error	delta	oov error	delta
HMM	5.69%	45.99%				
MEMM	6.37%	54.61%	4.81%	-25%	26.99%	-50%
CRF	5.55%	48.05%	4.27%	-24%	23.76%	-50%

Presentation Message

- Discriminative models are prone to the label bias problem

- CRFs provide the benefits of discriminative models

- CRFs solve the label bias problem well, and demonstrate good performance