Exponential Models for Sequential Data

John Lafferty
School of Computer Science
Carnegie Mellon University

Joint work with:
Doug Beeferman, Adam Berger
Andrew McCallum, Fernando Pereira

Acknowledgements:
Yoshua Bengio, Leon Bottou, Michael Collins,
Yann LeCun, Andrew Ng, Sebastian Thrun

Motivating Problem:

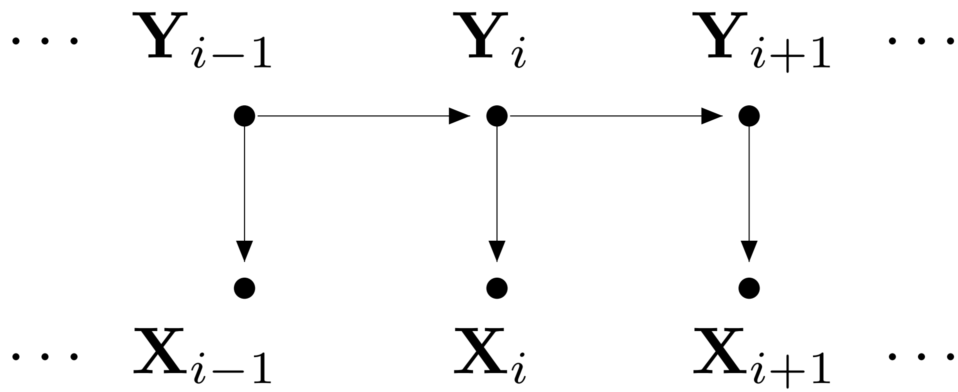
Segment & Annotate Data with Content Tags

Sequence Segmentation and Labeling

- *Goal*: mark up sequences with content tags
- Problem: *overlapping dependencies on context*
  - long-distance dependencies
  - multiple levels of granularity (e.g., words & characters)
  - aggregate properties (e.g., layout, html)
  - *past and future observations*
- *Generative* models that can represent such dependencies quickly become computationally intractable
- I'll focus on text, but similar problems in many other domains; e.g., biological sequence analysis

Modeling Sequences

Standard tool is the hidden Markov Model (HMM).



$$P(\mathbf{X}, \mathbf{Y}) \;=\; \prod_i P(\mathbf{X}_i \,|\, \mathbf{Y}_i)\, P(\mathbf{Y}_i \,|\, \mathbf{Y}_{i-1})$$

- Generative models, strong independence assumptions.
- Very widely used (genomics, natural language, information extraction...)

Conditional Models

- Model $p(\textit{label sequence } \mathbf{y} \,|\, \textit{observation sequence } \mathbf{x})$ rather than joint probability $p(\mathbf{y}, \mathbf{x})$
- Allow arbitrary dependencies on the observation sequence $\mathbf{x}$
- Still efficient (Viterbi, forward-backward) if dependencies within the state sequence $\mathbf{y}$ are constrained
- Do not need to use states to model dependency on past and future observations $\Rightarrow$ smaller state space, easier to design

Using Exponential Models (MEMMs)

- Represent probability $P(y' \mid x, y)$ of new state given observation and previous state as a product of "feature effects":

$$P(y' \mid y, x) = \frac{1}{Z(y, x)} \exp\left(\sum_k \underbrace{\lambda_k}_{\text{weight}} \underbrace{f_k(x, y, y')}_{\text{feature}} \right)$$

- Parameter estimation: Maximum likelihood or penalized (regularized) ML via iterative scaling

- Good empirical success for labeling and information extraction tasks (Rathnaparkhi, 1998; McCallum et al., 2000)

Outline

- Text Segmentation using Exponential Models
- The Label Bias Problem for State-Conditional Models
- Conditional Random Fields
- Experiments on Synthetic and Real Data

Text Segmentation

(BBL, 1999)

- Break up text stream into "semantically coherent" units
  - Not completely well-defined
  - Granularity depends on application
- Story segmentation: recover boundaries between "articles"
- Applications to video & audio retrieval
- Arises from temporal/sequential nature of data;
  analogous problems for DNA sequences, many other domains

Modeling the "Topic" Adaptively

Some doctors are more **skilled** at doing the **procedure** than others so it's **recommended** that **patients** ask **doctors** about their track record. People at high **risk** of **stroke** include those over age 55 with a family **history** or high **blood pressure**, **diabetes** and **smokers**. We urge them to be evaluated by their family **physicians** and this can be done by a very simple **procedure** simply by having them test with a **stethoscope** for symptoms of blockage.

An Adaptive Language Model (Generative)

- First construct a standard, static (stationary) backoff trigram model

$$p_{\mathsf{tri}}(w \mid w_{-2}, w_{-1})$$

- Use this as a prior/default in a family of conditional exponential models

$$p_{\mathsf{exp}}(w \mid H) = \frac{1}{Z(H)} \exp\left(\sum_i \lambda_i f_i(H, w)\right) p_{\mathsf{tri}}(w \mid w_{-2}, w_{-1})$$

where $H \equiv w_{-N}, w_{-W+1}, \ldots, w_{-1}$ is the word *history*.

An Adaptive Language Model (cont.)

- The features f_i depend both on the word history H and the word being predicted; assigned a weight λ_i.
- H is the previous 500-word context (sliding window)
- Here we use *trigger features*:

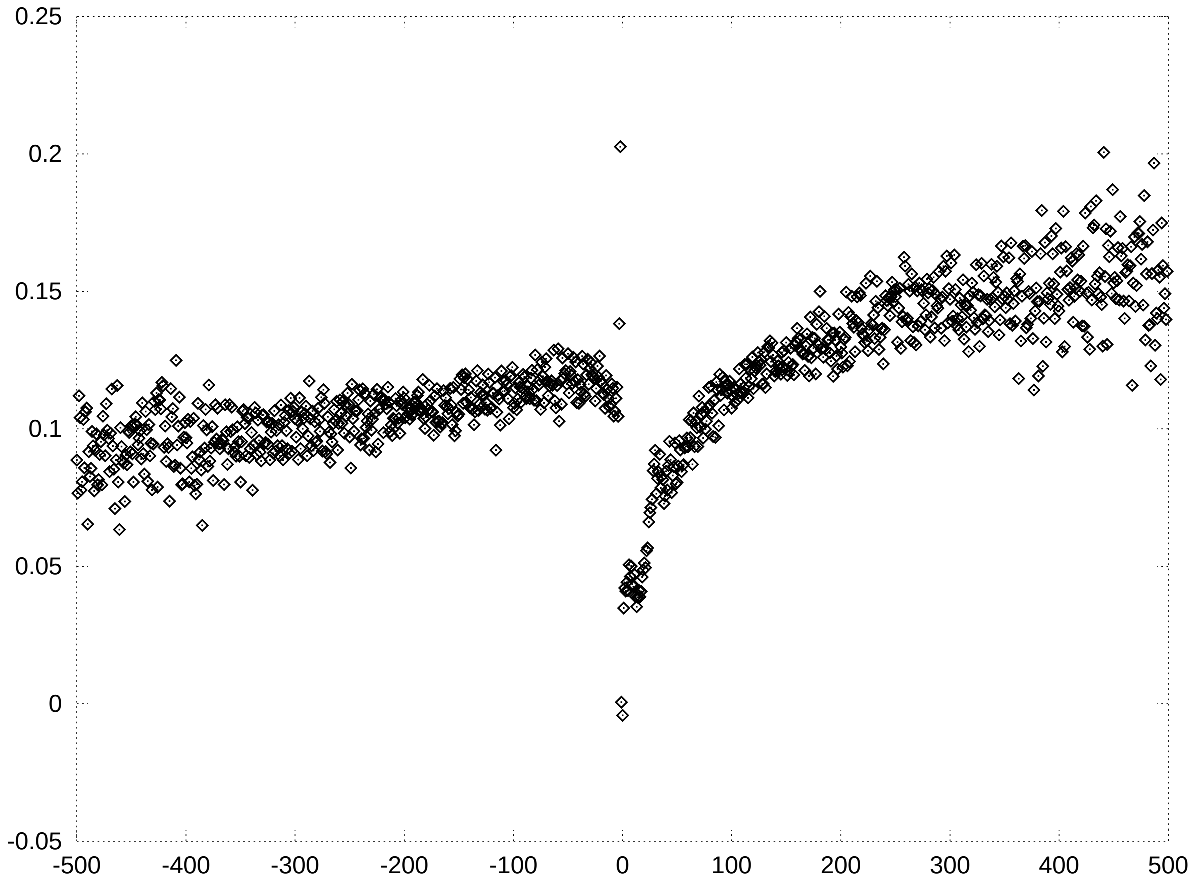
$$f_i(H, w) = \begin{cases} 1 & \text{if } s_i \in H \text{ and } w = t_i \\ 0 & \text{otherwise.} \end{cases}$$

Sample Triggers

(s,t)	e^{λ}
residues, carcinogens	2.3
Charleston, shipyards	4.0
microscopic, cuticle	4.1
defense, defense	8.4
tax, tax	10.5
Kurds, Ankara	14.8
Vladimir, Gennady	19.6
education, education	22.2
music, music	22.4
insurance, insurance	23.0
Pulitzer, prizewinning	23.6
Yeltsin, Yeltsin	23.7
Russian, Russian	26.1
sauce, teaspoon	27.1
flower, petals	32.3
casinos, Harrah's	42.8

(s,t)	e^{λ}
recent, recent	2.3
national, national	3.3
University, University	3.5
Doo, Chun	6.3
Soviet, Soviet	6.9
fraud, fraud	8.0
detergent, Tide	9.2
Carolco, Hoffman	9.7
Freddie, conventional	10.0
aluminium, smelter	10.4
officers, officers	11.0
records, records	11.5
merger, merger	11.6
proportionate, chances	15.6
nutrasweet, sweetener	18.4
waste, waste	20.7

Change Across Segment Boundaries



$$\mathsf{LR}_i = \log\left(\frac{p_{\mathsf{exp}}(w_i \mid H)}{p_{\mathsf{tri}}(w_i \mid w_{i-2}, w_{i-1})}\right)$$

Lexical Features

Broadcast news:

CNN'S RICHARD BLYSTONE IS HERE TO TELL US...

THIS IS WOLF BLITZER *reporting* LIVE FROM THE WHITE HOUSE.

Newswire:

A TEXAS AIR NATIONAL GUARD FIGHTER JET CRASHED *Friday* IN A REMOTE AREA OF SOUTHWEST TEXAS.

He WAS AT HOME WAITING FOR A LIMOUSINE TO TAKE HIM TO LOS ANGELES AIRPORT FOR A TRIP TO CHICAGO.

The Learning Paradigm: Feature Selection/Induction

- Goal: construct a probability distribution $q(b \,|\, \omega)$, where $b \in \{\text{YES}, \text{NO}\}$ is the value of a random variable describing the presence of a segment boundary in context ω.

- We consider distributions in the *exponential family*

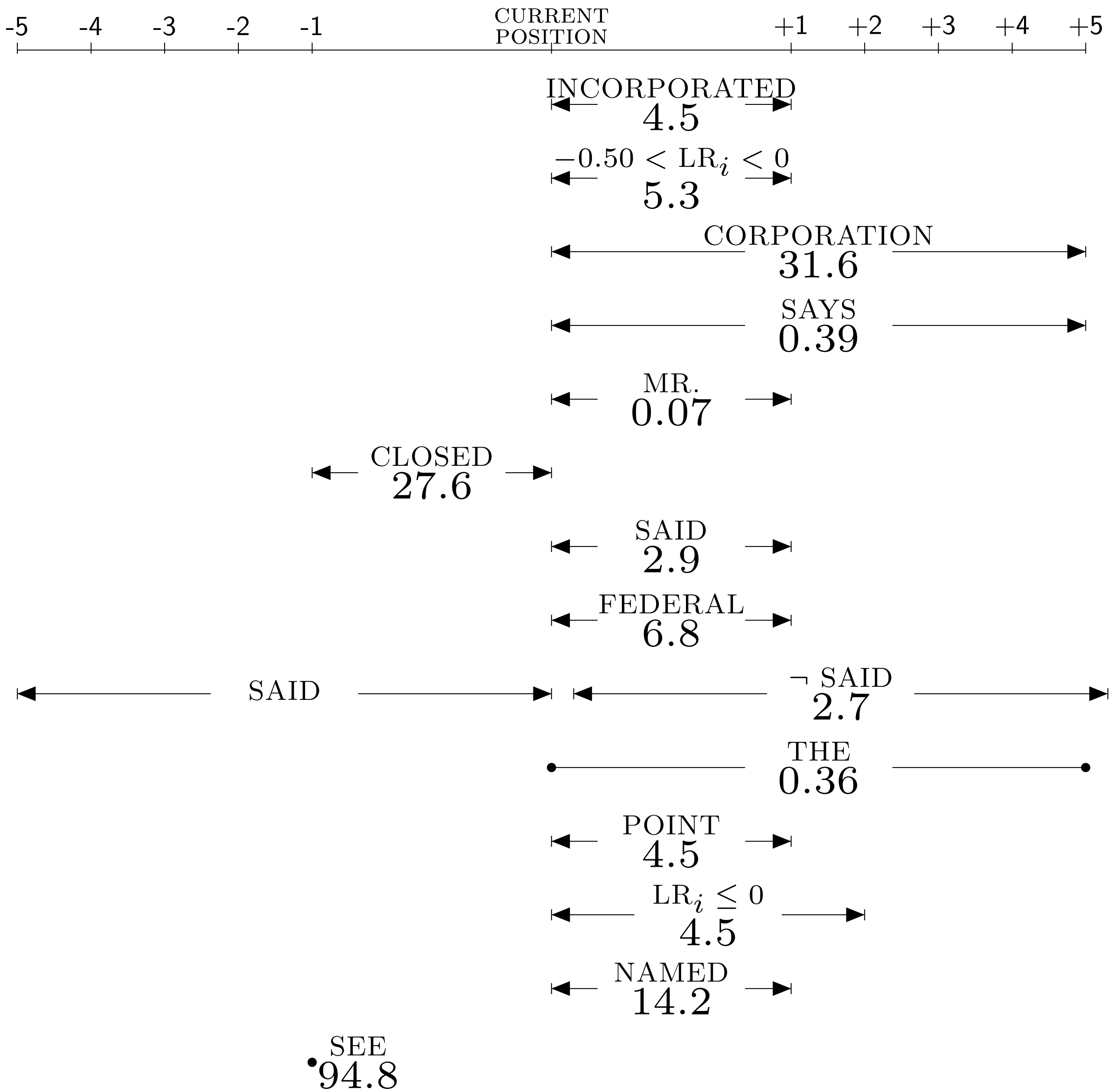
$$\mathcal{Q}(f, q_0) = \left\{ q(\cdot \,|\, \omega) \;:\; q(b \,|\, \omega) = \frac{1}{Z_\lambda(\omega)} e^{\lambda \cdot f(\omega)} \, q_0(b \,|\, \omega) \right\}$$

$$\lambda \cdot f(\omega) = \lambda_1 f_1(\omega) + \lambda_2 f_2(\omega) + \cdots \lambda_n f_n(\omega) \,.$$

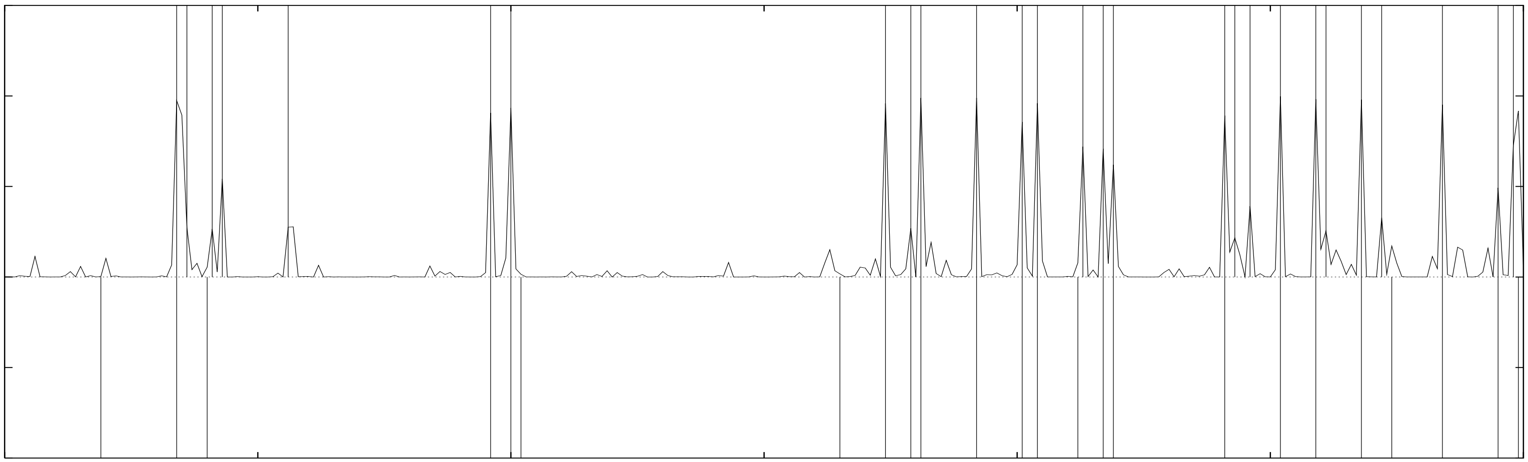
- The *gain* of the candidate feature g is defined to be

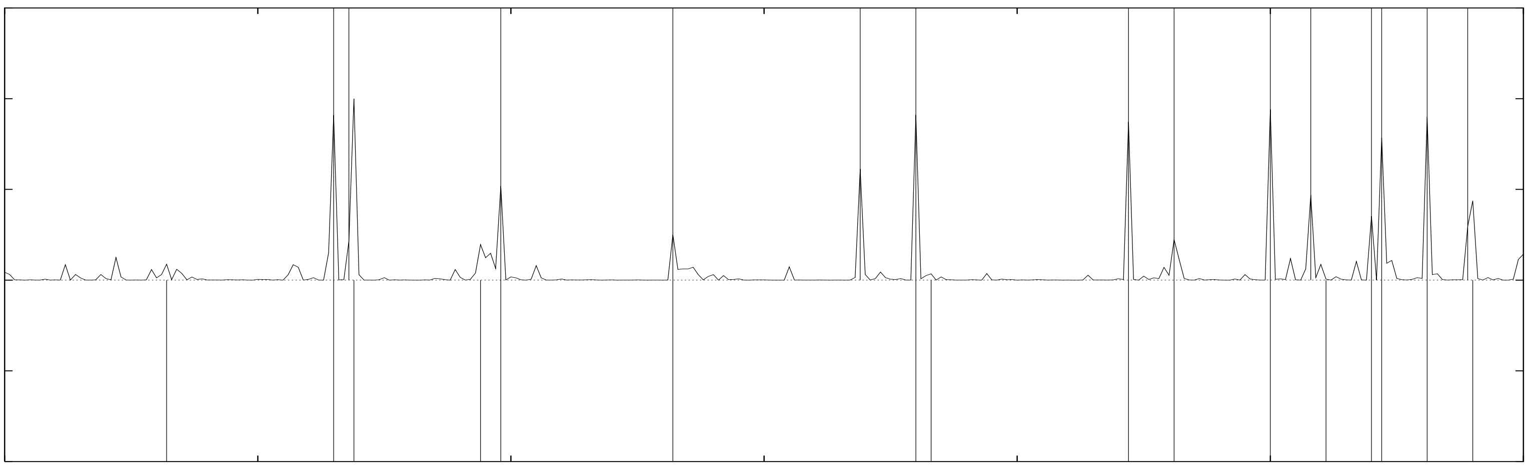
$$G_q(g) = \text{argmax}_\alpha \left(D(\tilde{p} \,\|\, q) - D(\tilde{p} \,\|\, q_{\alpha, f}) \right) \,.$$

First Features Selected for WSJ

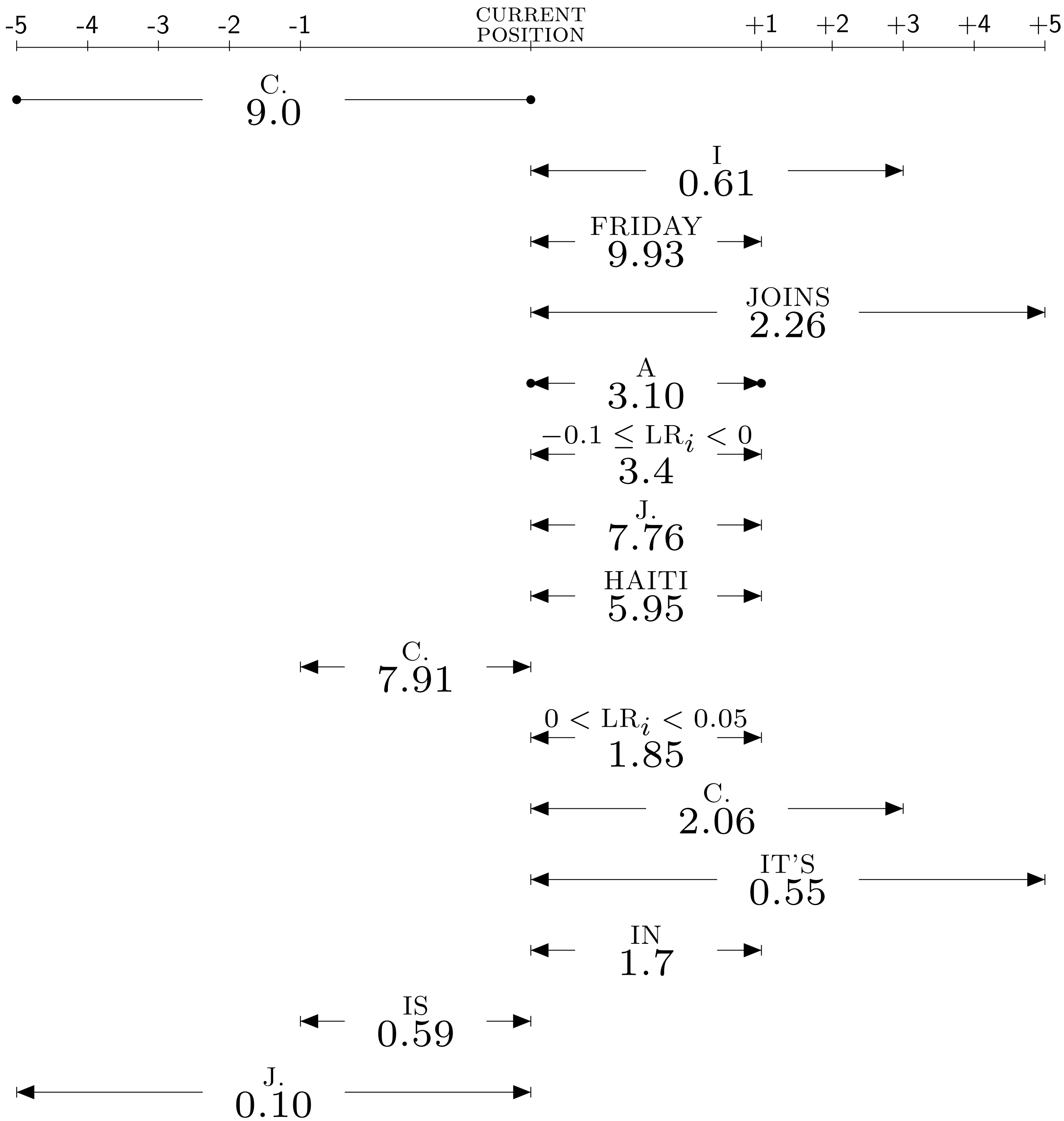


Sample Segmentations of Wall Street Journal

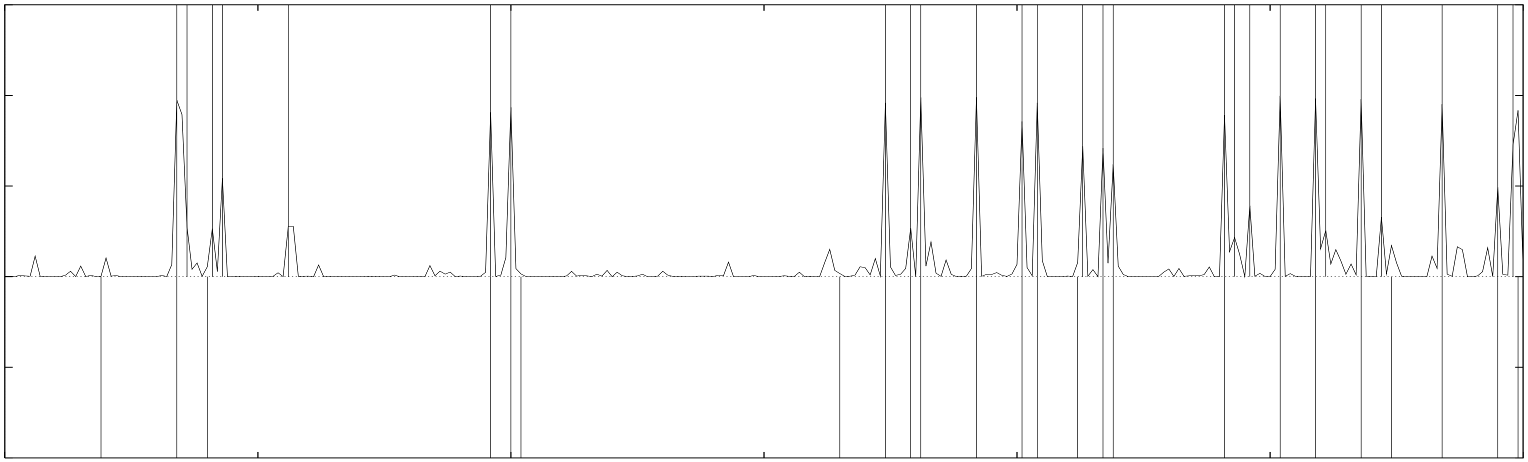


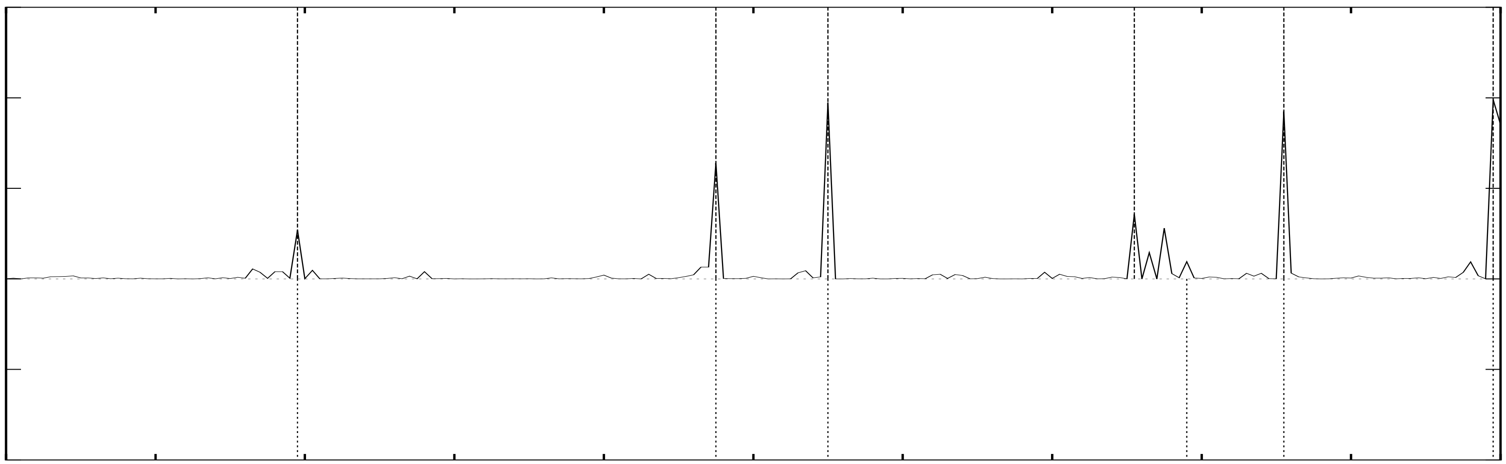


First Features Selected for CNN

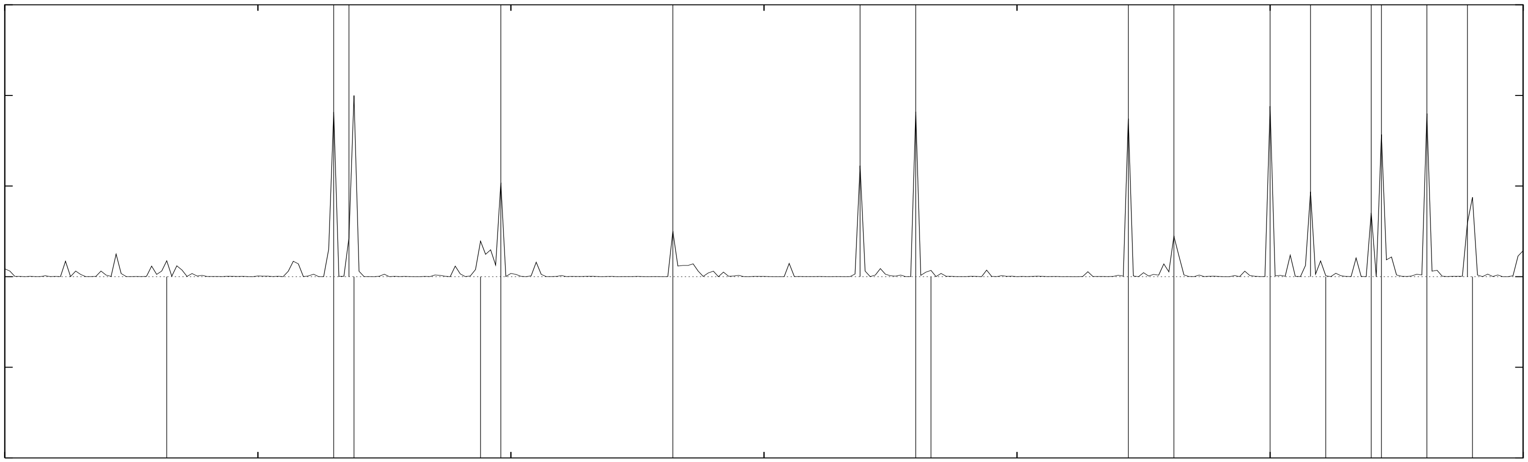


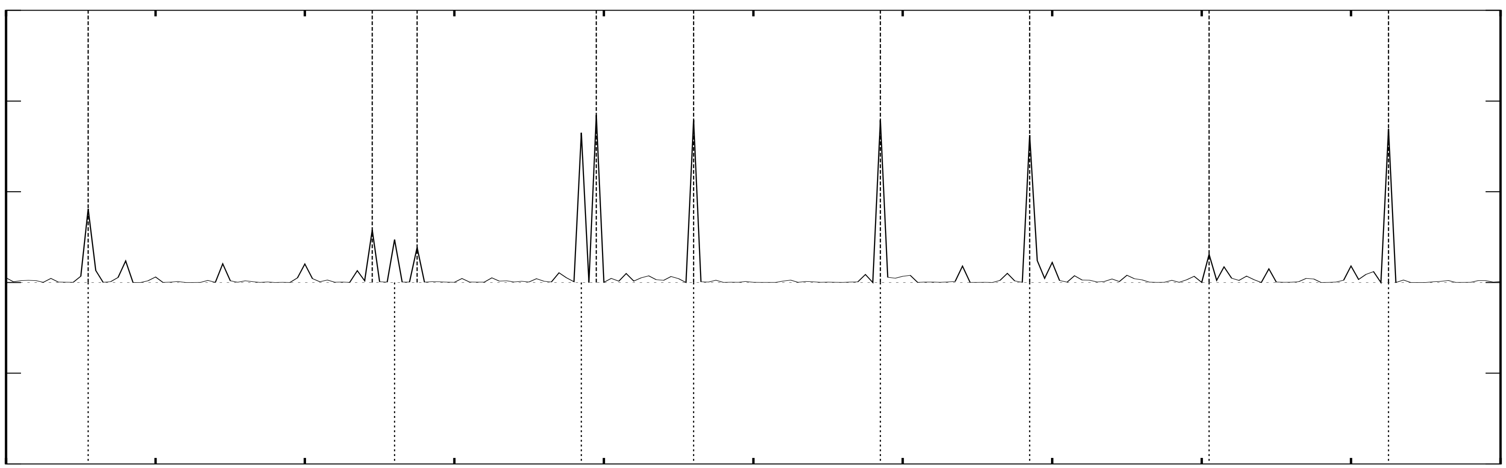
Sample Segmentations: WSJ/CNN





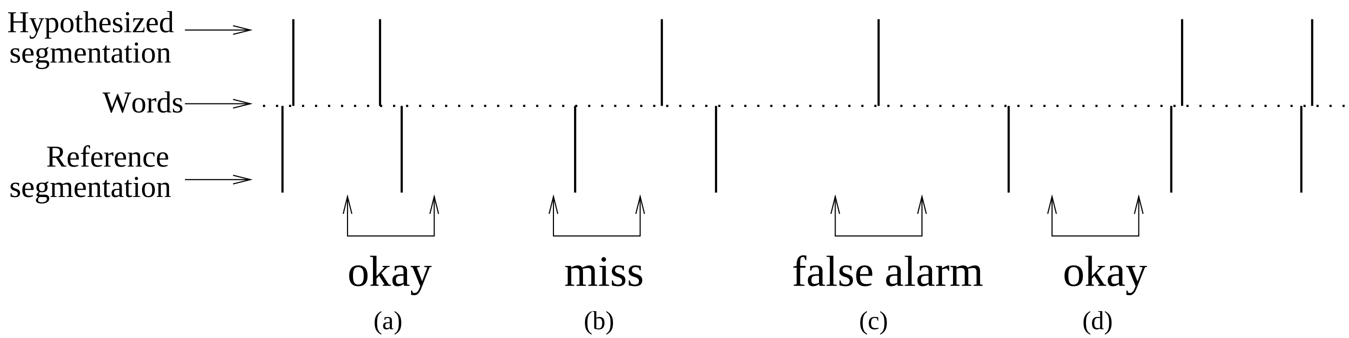
Sample Segmentations: WSJ/CNN





Evaluation: A Probabilistic Error Metric

Error is calculated as the probability P_μ that the reference and hypothesized segmentations disagree between two randomly chosen word positions:



Quantitative Segmentation Results

model	*reference segments*	*hypoth. segments*	P_μ	*precision*	*recall*	*F-meas.*
exp. model	9984	9543	0.12	60%	57%	58
random	9984	9984	0.32	12%	12%	12
all	9984	219,099	0.41	5%	100%	9
none	9984	0	0.57	0%	0%	—
even	9984	9980	0.26	14%	12%	13%

(Note: Have also compared to HMMs, decision trees, and some other methods.)

Modeling Temporal Structure

- This finesses the sequential/temporal nature of the problem: Viewed as series of classification problems with simple sequential decision rule

- Want explicit notion of time/state

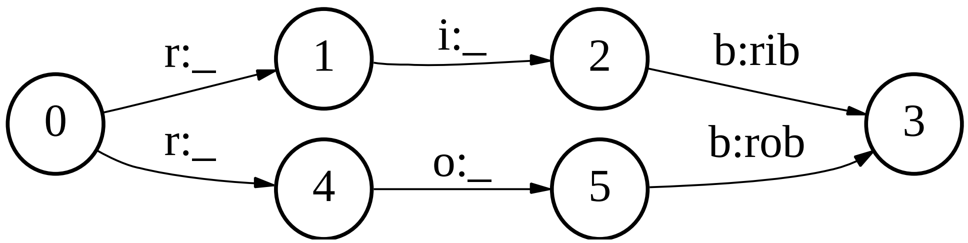
- Represent probability $P(y_i \,|\, x, y_{i-1})$ of new state given observation and previous state using features:

$$P(y_i \,|\, y_{i-1}, x) = \frac{1}{Z(y_{i-1}, x)} \exp(\sum_k \underbrace{\lambda_k}_{\text{weight}} \underbrace{f_k(x, y_{i-1}, y_i)}_{\text{feature}})$$

- However, potential problem...

The Label Bias Problem in Conditional Models

- Bias toward states with fewer outgoing transitions
- Example (after Bottou 91):



$$\begin{aligned} p(1,2\,|\,\texttt{ro}) &= p(1\,|\,\texttt{r})p(2\,|\,\texttt{o},1) \\ &= p(1,2\,|\,\texttt{ri}) \end{aligned}$$

- Per-state normalization does not allow the required $\text{score}(1,2\,|\,\texttt{ro}) \ll \text{score}(1,2\,|\,\texttt{ri})$

Experiments on Synthetic Data

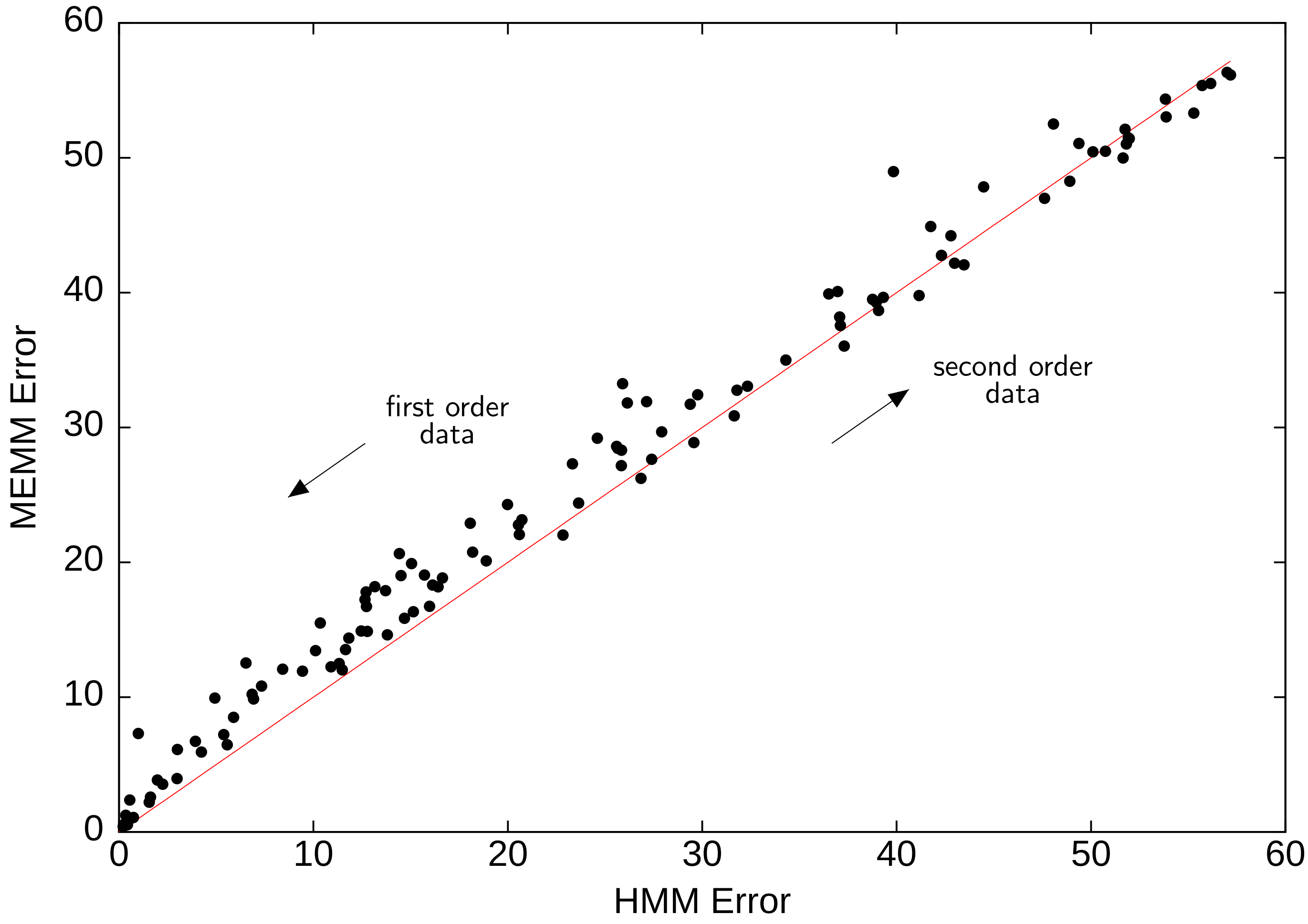
- Generate data according to mixture of first-order and second-order hidden Markov Model (5 states, 26 outputs)

$$p(\mathbf{x}, \mathbf{y}) \;=\; (1 - \alpha)\, p_1(\mathbf{x}, \mathbf{y}) + \alpha\, p_2(\mathbf{x}, \mathbf{y})$$

- Train *first-order* models parameterized in the same way.

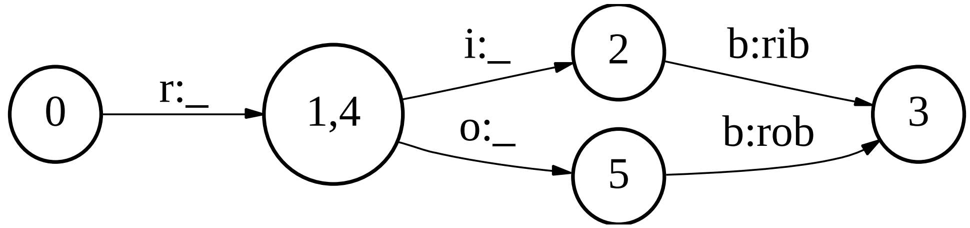
- As the data becomes more second order, the error rates increase, as first-order models fail to fit higher-order data.

MEMM vs. HMM



60
50
40
30
20
10
0
0
10
20
30
40
50
60
MEMM Error
HMM Error
first order
data
second order
data

Proposed Solutions

1. *Determinization*: 

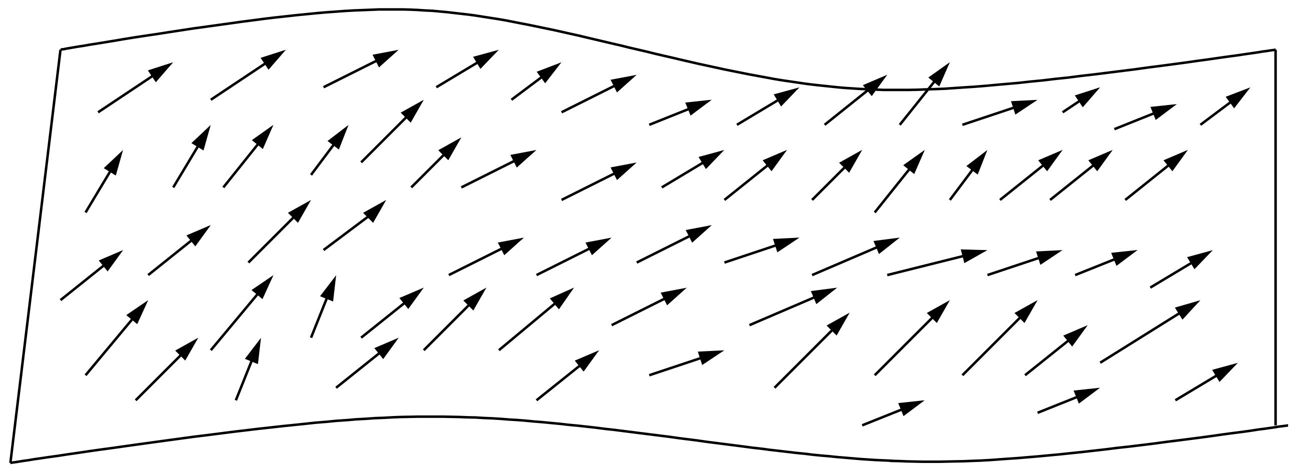
   - not always possible
   - state-space explosion

2. *Fully-connected models*: lack prior structural knowledge

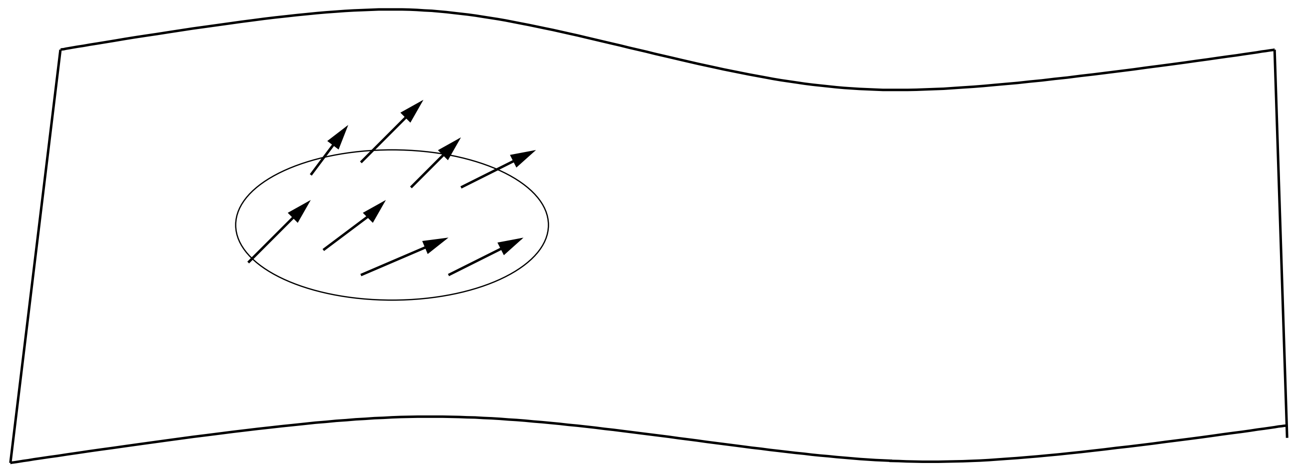
3. Our solution: *Conditional random fields* (CRFs):

   - Allow some transitions to "vote" more strongly than others in computing state sequence probability
   - *Whole sequence* rather than per-state normalization; conditioned on entire input sequence.
   - Convex likelihood function

Classical Notion of Random Field

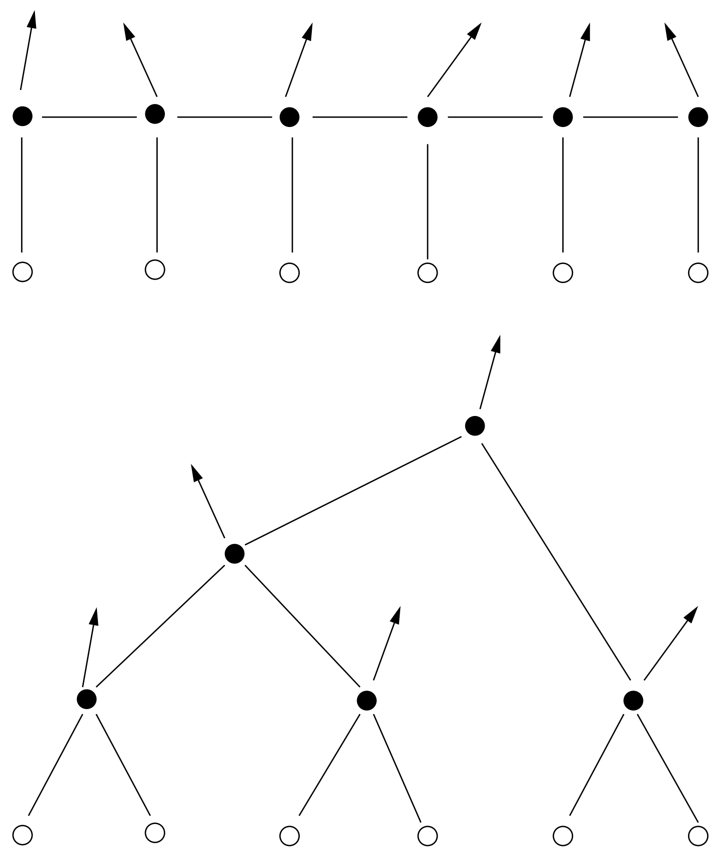


Markov Property



$$p(X_A \,|\, X_v, v \notin A) \;\; = \;\; p(X_A \,|\, X_v, v \in \partial A)$$

Random Fields on Sequences: Chains and Trees



Conditional Random Fields

Suppose there is a graphical structure to $\mathbf{Y}$; i.e., graph $G = (V, E)$ such that $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{|V|})$.

A distribution $p(\mathbf{Y} \,|\, \mathbf{X})$ is a *conditional random field* in case, when conditioned on $\mathbf{X}$, the random variables $\mathbf{Y}_v$ obey the Markov property with respect to the graph:

$$p(\mathbf{Y}_v \,|\, \mathbf{X}, \mathbf{Y}_w, w \neq v) \;\; = \;\; p(\mathbf{Y}_v \,|\, \mathbf{X}, \mathbf{Y}_w, (w, v) \in E)$$

Tree-based Models

Assume underlying graph is a tree. Hammersley-Clifford theorem says CRF is a Gibbs distribution:

$$p_\theta(\mathbf{y} \mid \mathbf{x}) \;\propto\; \exp\left(\sum_{e \in E, k} \lambda_k \, f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k \, g_k(v, \mathbf{y}|_v, \mathbf{x}) \right)$$

CRFs for Sequences

- The state sequence is a Markov random field *conditioned* on the observation sequence

- Model form: $p(\mathbf{y} \mid \mathbf{x}) \propto \exp \sum_{t=1}^{T} \left[\begin{array}{l} \sum_j \lambda_j f_j(y_t, y_{t-1} \mid \mathbf{x}, t) \\ + \sum_k \mu_k g_k(y_t \mid \mathbf{x}, t) \end{array} \right]$

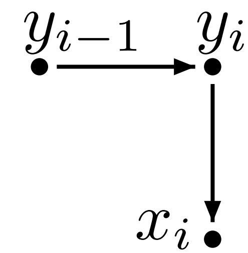
- Features:
  - f_j represent the interaction between successive states, conditioned on the observations
  - g_k represent the dependence of a state on the observations

- Dependence on entire observation sequence

A Special Case: From HMMs to CRFs

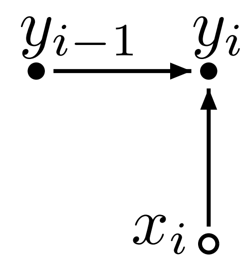
HMM:

$$p(\mathbf{y} \,|\, \mathbf{x}) \;\propto\; \prod_{t=1}^{T} p(y_t \,|\, y_{t-1}) p(x_t \,|\, y_t)$$



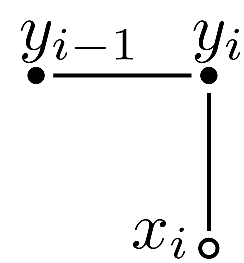
MEMM:

$$p(\mathbf{y} \,|\, \mathbf{x}) \;=\; \prod_{t=1}^{T} \frac{1}{Z_{y_{t-1},x_t}} \exp \left[\begin{array}{l} \sum_j \lambda_j f_j(y_t, y_{t-1}) \\ + \sum_k \mu_k g_k(y_t, x_t) \end{array} \right]$$



CRF:

$$p(\mathbf{y} \,|\, \mathbf{x}) \;=\; \frac{1}{Z_{\mathbf{x}}} \prod_{t=1}^{T} \exp \left[\begin{array}{l} \sum_j \lambda_j f_j(y_t, y_{t-1}) \\ + \sum_k \mu_k g_k(y_t, x_t) \end{array} \right]$$



Discriminative “Boltzmann chains” (Saul and Jordan; MacKay, 1996)

Efficient Estimation

Marginals and normalizing constant can be computed efficiently using dynamic programming

Matrix notation:

$$\begin{aligned} M_i(y', y \,|\, \mathbf{x}) &= \exp\left(\Lambda_i(y', y \,|\, \mathbf{x})\right) \\ \Lambda_i(y', y \,|\, \mathbf{x}) &= \textstyle\sum_k \lambda_k \, f_k(e_i, \mathbf{Y}|_{e_i} = (y', y), \mathbf{x}) + \\ &\quad \textstyle\sum_k \mu_k \, g_k(v_i, \mathbf{Y}|_{v_i} = y, \mathbf{x}) \end{aligned}$$

where e_i is the edge with labels $(\mathbf{Y}_{i-1}, \mathbf{Y}_i)$ and v_i is the vertex with label $\mathbf{Y}_i$.

Normalization (partition function):

$$Z_\theta(\mathbf{x}) = \left(M_1(\mathbf{x})\, M_2(\mathbf{x}) \cdots M_{n+1}(\mathbf{x})\right)_{\texttt{start},\texttt{stop}}$$

Forward-Backward Calculations

- Probability of label $\mathbf{Y}_i = y$, given observation sequence $\mathbf{x}$:

$$
\begin{aligned}
Prob_\theta(\mathbf{Y}_i = y \,|\, \mathbf{x}) &= \frac{\alpha_i(y \,|\, \mathbf{x})\, \beta_i(y \,|\, \mathbf{x})}{Z_\theta(\mathbf{x})} \\
\alpha_i(\mathbf{x}) &= \alpha_{i-1}(\mathbf{x})\, M_i(\mathbf{x}) \\
\beta_i(\mathbf{x})^\top &= M_{i+1}(\mathbf{x})\, \beta_{i+1}(\mathbf{x})
\end{aligned}
$$

- Training requires forward-backward ($unlike\ for\ HMMs$)

- Complexity same as standard Baum-Welch, even with "global" features.

Iterative Scaling

Update equations:

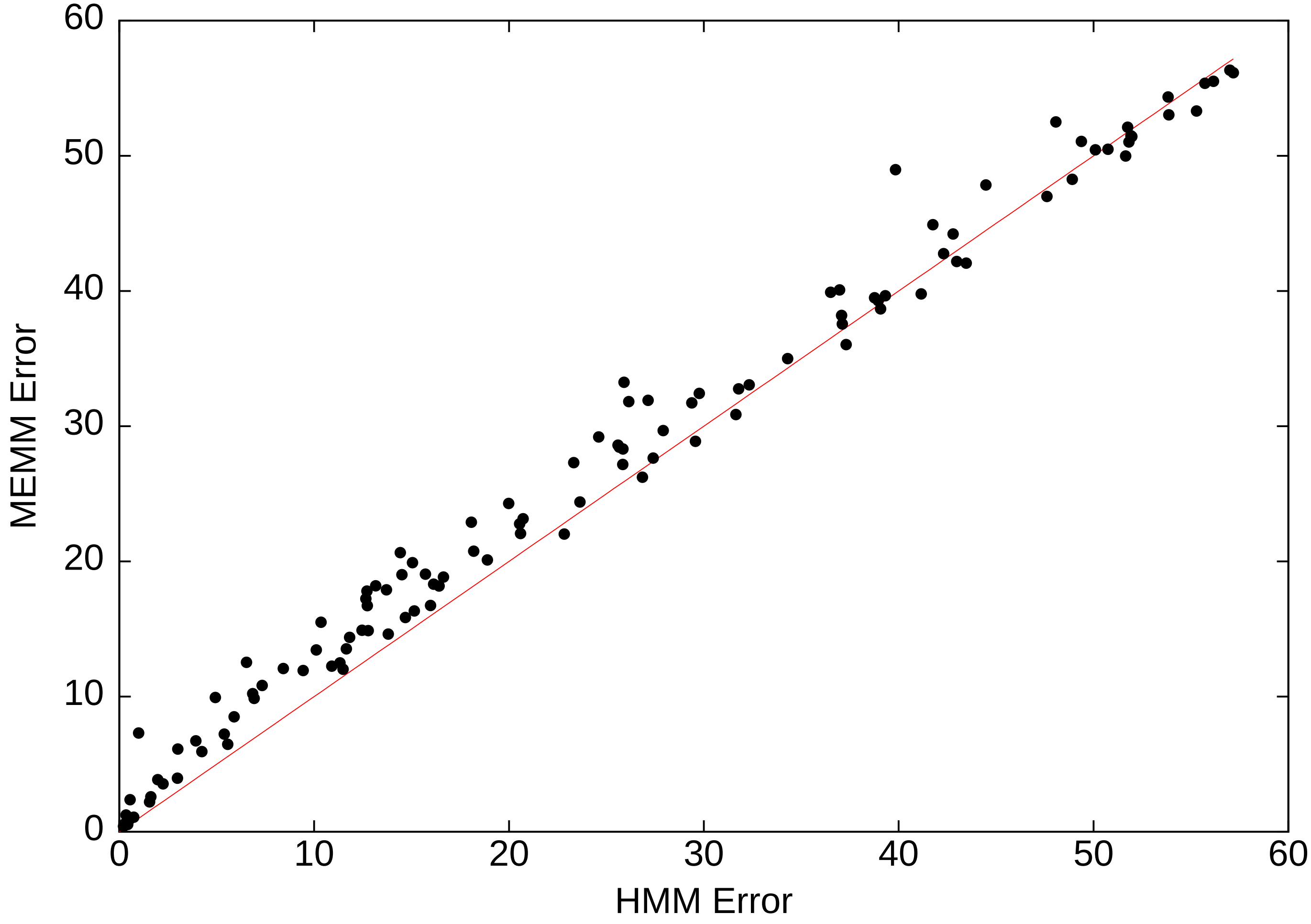
$$\delta\lambda_k = \frac{1}{S}\log\frac{\widetilde{E}f_k}{Ef_k}, \quad \delta\mu_k = \frac{1}{S}\log\frac{\widetilde{E}g_k}{Eg_k}$$

where

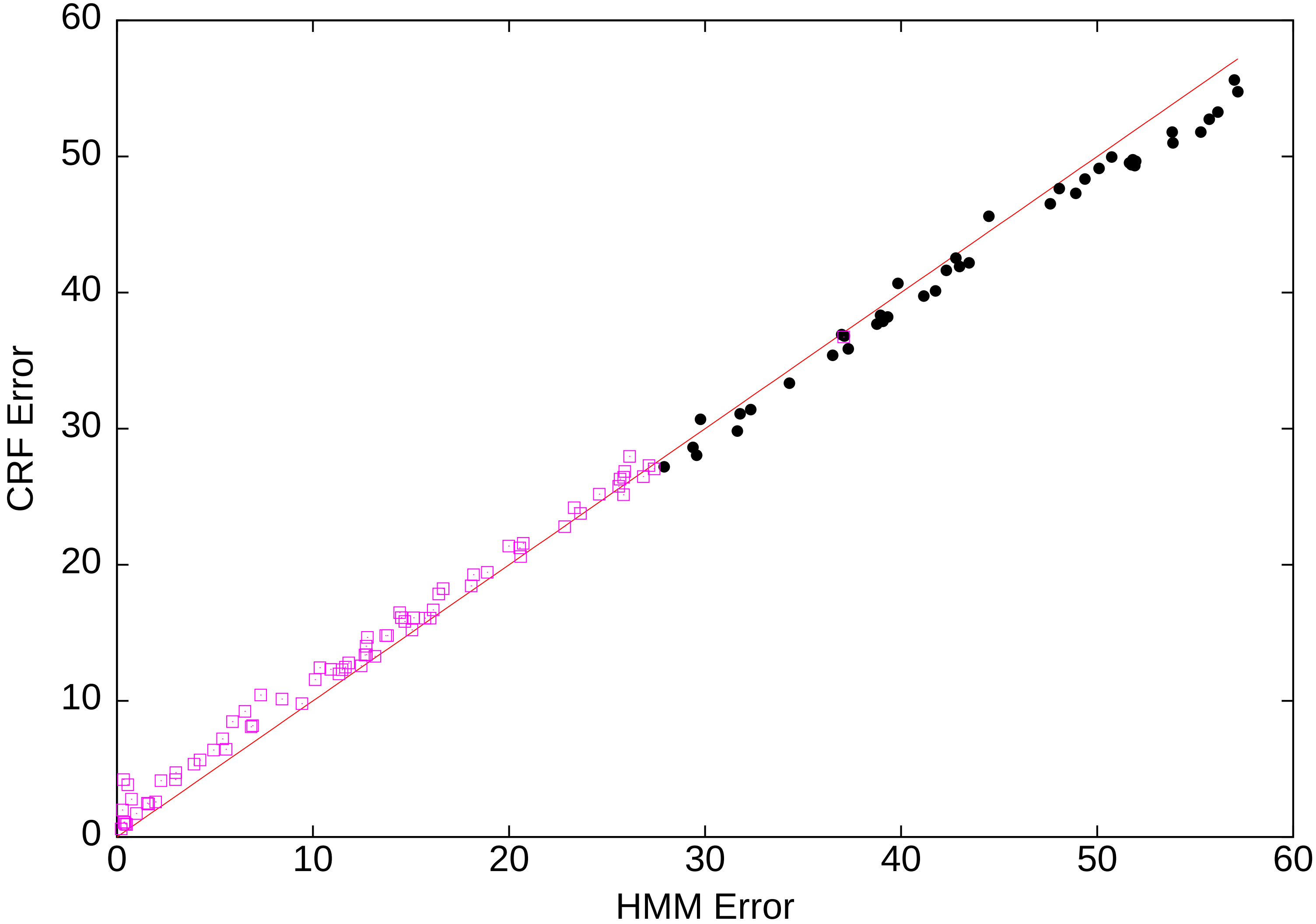
$$Ef_k = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x}) \sum_{i=1}^{n+1} \sum_{y',y} f_k(e_i, \mathbf{y}|_{e_i} = (y', y), \mathbf{x}) \times \frac{\alpha_{i-1}(y' \mid \mathbf{x})\, M_i(y', y \mid \mathbf{x})\, \beta_i(y \mid \mathbf{x})}{Z_\theta(\mathbf{x})}$$

(and similarly for Eg_k)

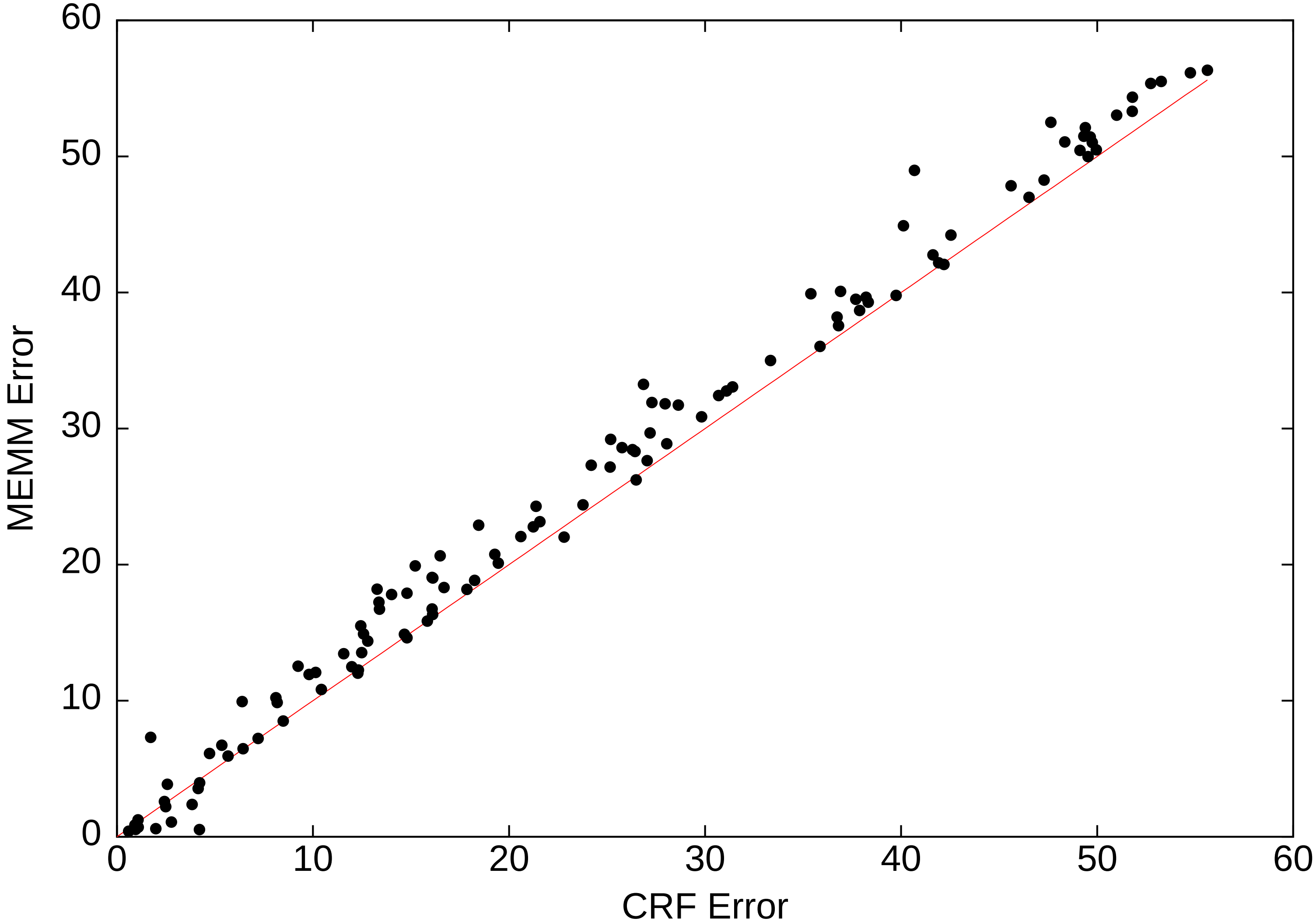
Recall: MEMM vs. HMM



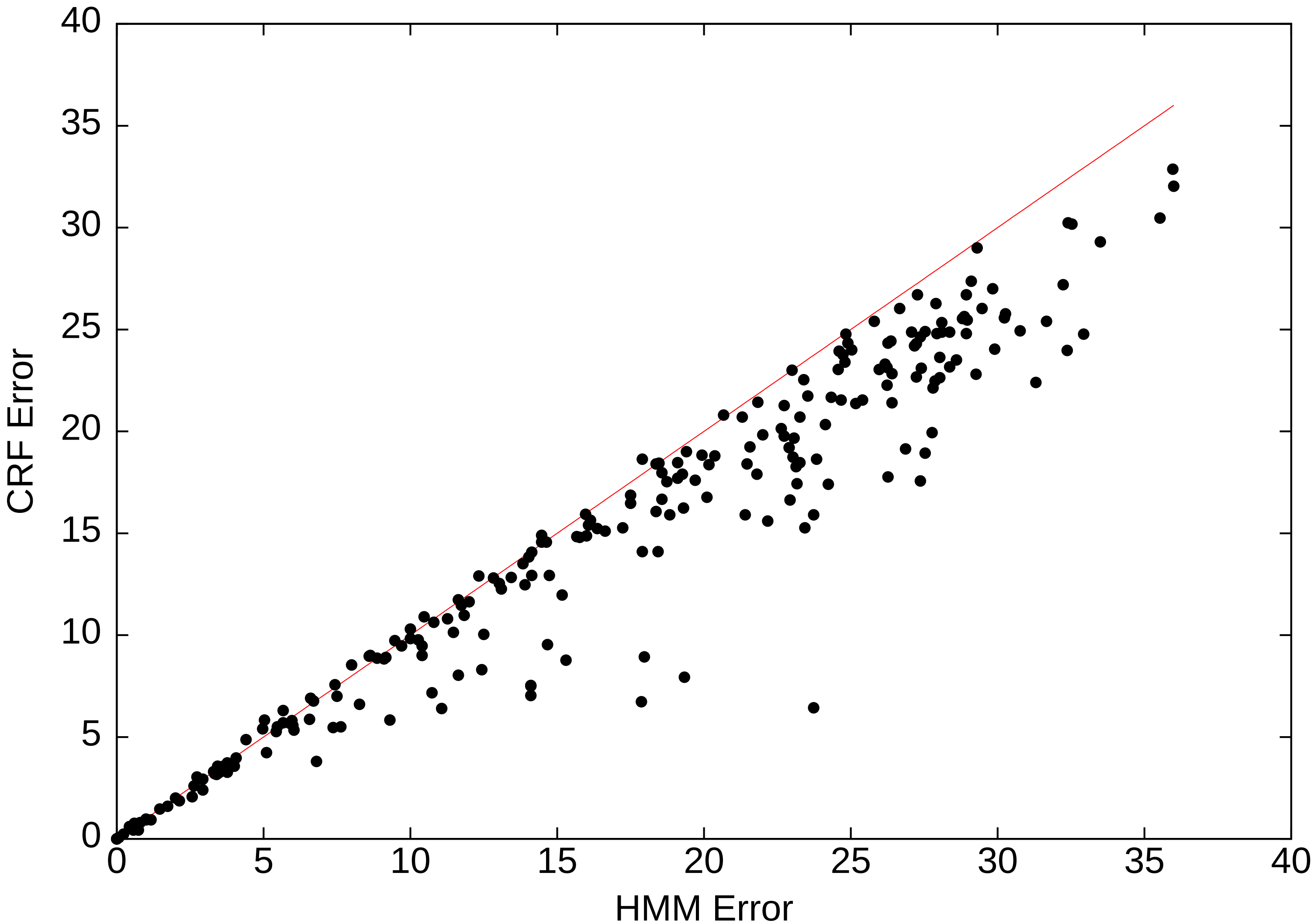
CRF vs. HMM



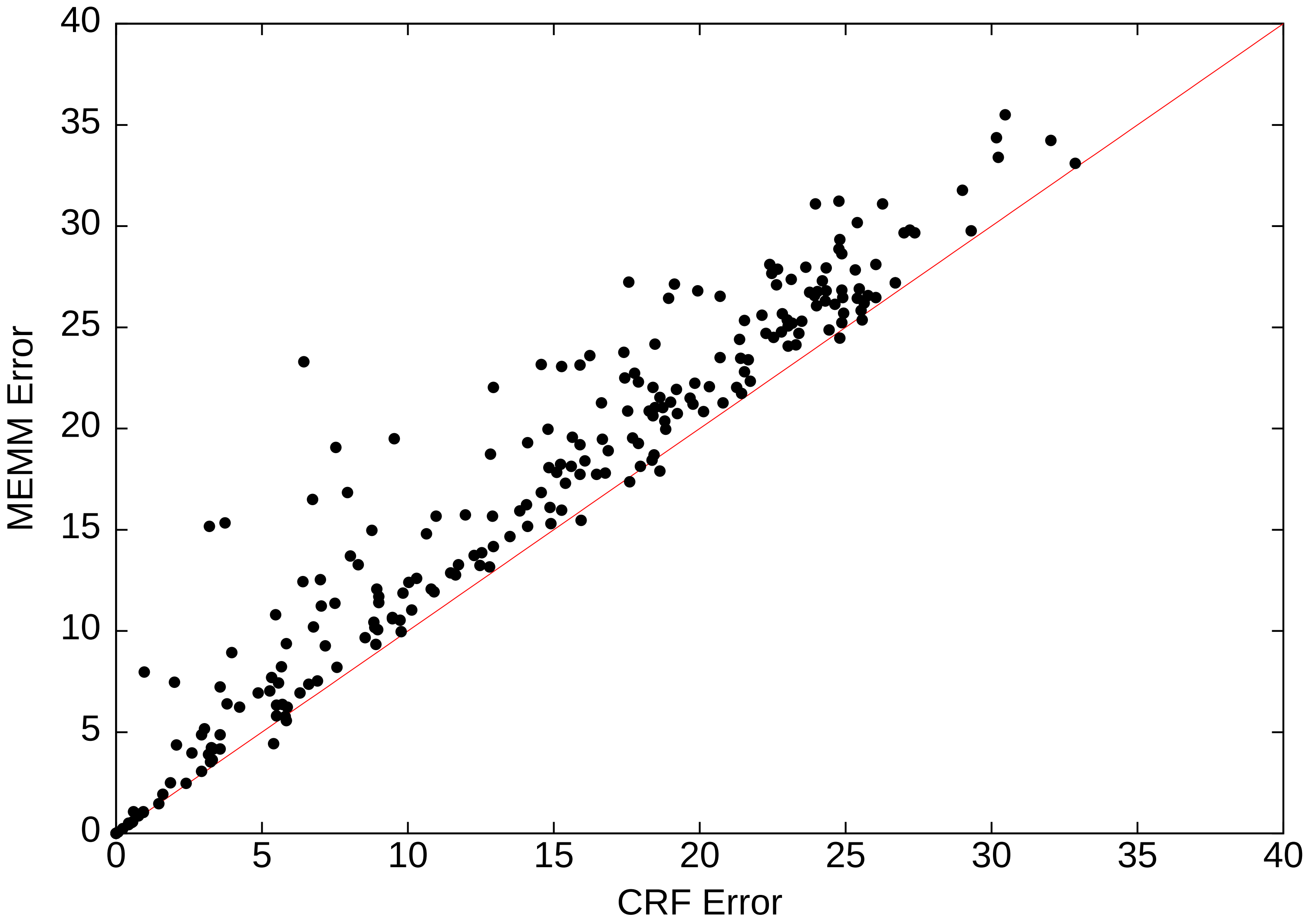
MEMM vs. CRF



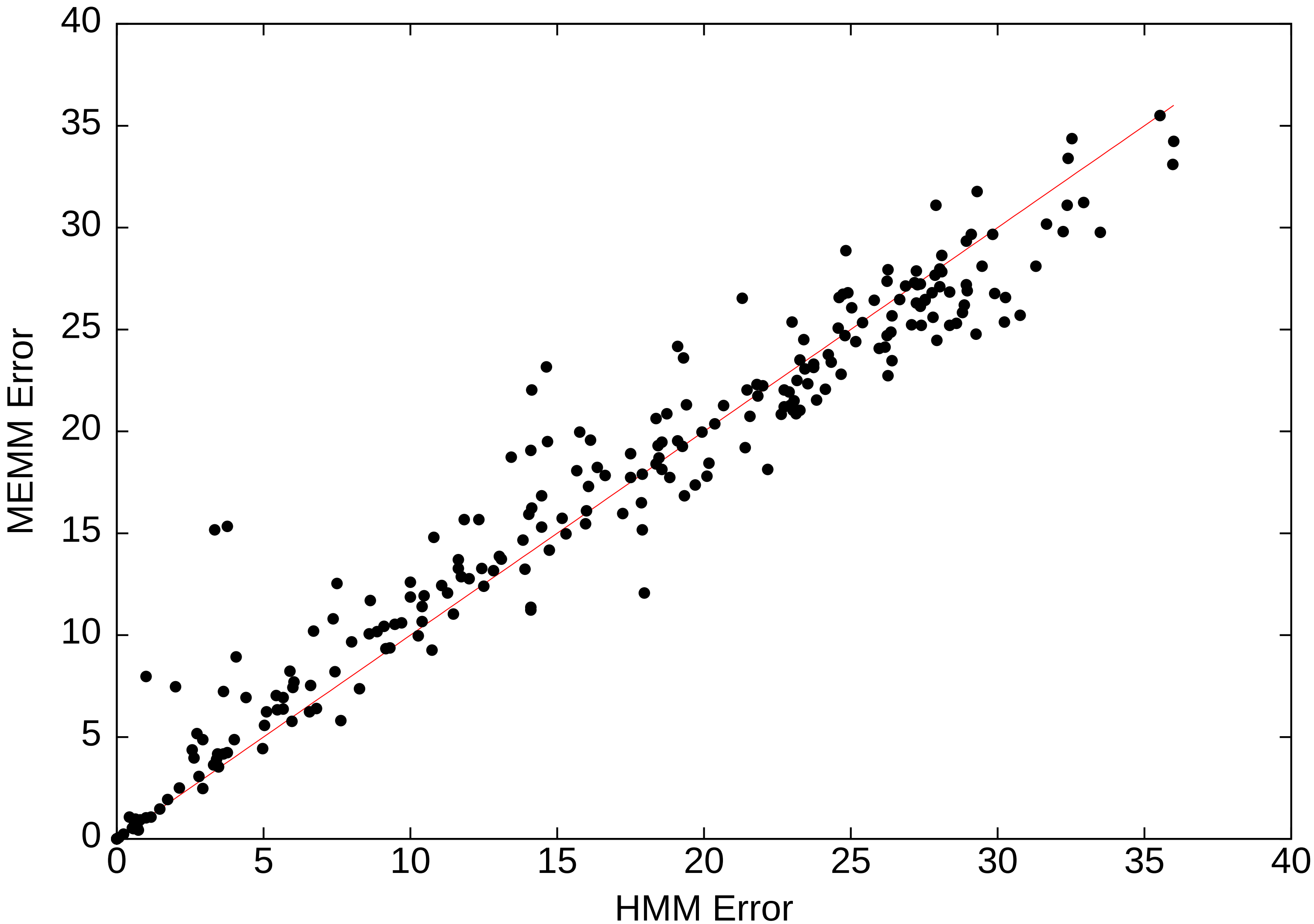
CRF vs. HMM



MEMM vs. CRF



MEMM vs. HMM



Experiments on Text

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**

Sample Results on Penn Data

	error	oov	oov error
HMM	5.69%	5.45%	45.99%
MEMM	6.37%	5.45%	54.61%
CRF	5.55%	5.45%	48.05%

Results with Spelling Features

			using spelling features			
	error	oov error	error	Δ	oov error	Δ
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%

Future Directions

- Tree-structured random fields for hierarchical parsing
- Feature selection and induction: automatically choose the f_k and g_k functions (efficiently)
- Train to maximize per-symbol likelihood $\prod_i Prob(y_i \mid \mathbf{x})$ (not pseudo-likelihood)
- Numerical methods to accelerate convergence (e.g. quasi-Newton, hybrid IS and conjugate gradient)
- Theoretical bounds on performance

Summary

- Conditional sequence models have the advantage of allowing complex dependencies among input features
- May be prone to the label bias problem
- CRFs are an attractive modeling framework that:
  - Discriminatively model sequence annotations
  - Allow non-local features
  - Avoid label bias through global normalization
  - Have efficient inference & estimation algorithms