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

John Lafferty
Andrew McCallum
Fernando Pereira

Goal: Sequence segmentation and labeling

- Computational biology
- Computational linguistics
- Computer science

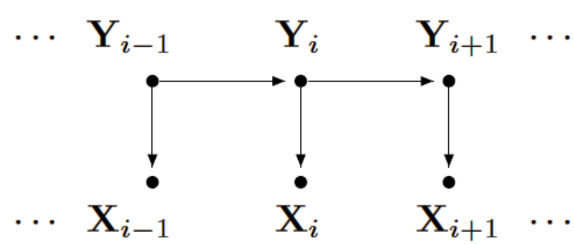
Overview

- Generative models
- Conditional models
- Label bias problem
- Conditional random fields
- Experiments

Generative Models

- HMMs and stochastic grammars
- Assign a joint probability to paired observation and label sequences
- Parameters are trained to maximize joint likelihood of training examples

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

- Need to enumerate all possible observation sequences
- To ensure tractability of inference problem, must make strong independence assumptions (*i.e.*, conditional independence given labels)

Conditional models

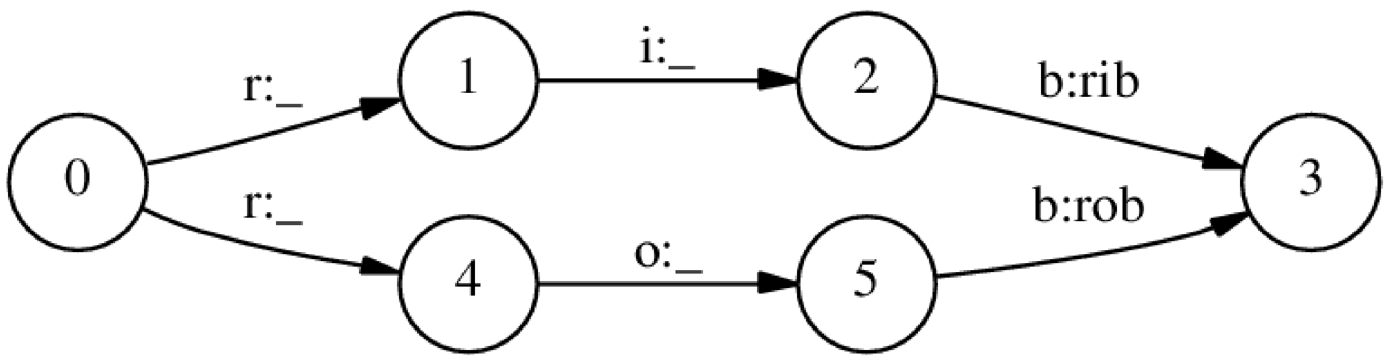
- Specify probabilities of label sequences given an observation sequence
- Does not expend modeling effort on the observations which are fixed at test time
- Conditional probability can dependent on arbitrary, non-independent features of the observation sequence

Example: MEMMs

- Maximum entropy Markov models
- Each source state has an exponential model that takes the observation feature as input and outputs a distribution over possible next states
- Weakness: Label bias problem

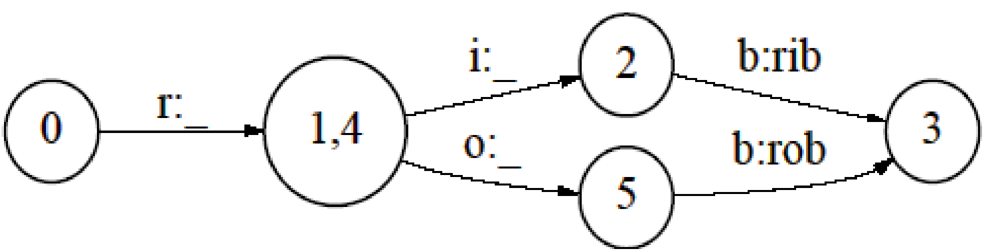
Label Bias Problem

- Per-state normalization of transition scores implies "conservation of score mass"
- Bias towards states with fewer outgoing transitions
- State with single outgoing transition effectively ignores observation



Solving Label Bias

- Collapse states, and delay branching until get a discriminating observation
  - Not always possible or may lead to combinatorial explosion



Solving Label Bias (cont'd)

- Start with fully-connected model and let training procedure figure out a good structure
  - Precludes use of prior structure knowledge

Overview

- Generative models
- Conditional models
- Label bias problem
- Conditional random fields
- Experiments

Conditional Random Fields

- Undirected graph (random field)
- Construct conditional model p(Y|X)
- Does not explicitly model marginal p(X)
- Assumption: graph is fixed
  - Paper concerns itself with chain graphs and sequences

CRFs: Distribution

weights

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

features

CRFs: Example Features

$$
\begin{aligned}
f_{y',y}\left(<u,v>, \mathbf{y}|_{<u,v>}, \mathbf{x}\right) &= \delta(\mathbf{y}_u, y')\,\delta(\mathbf{y}_v, y) \\
g_{y,x}\left(v, \mathbf{y}|_v, \mathbf{x}\right) &= \delta(\mathbf{y}_v, y)\,\delta(\mathbf{x}_v, x)
\end{aligned}
$$

- Corresponding parameters λ and μ similar to the (logarithms of the) HMM parameters p(y'|y) and p(x|y)

CRFs: Parameter Estimation

- Maximize log-likelihood objective function

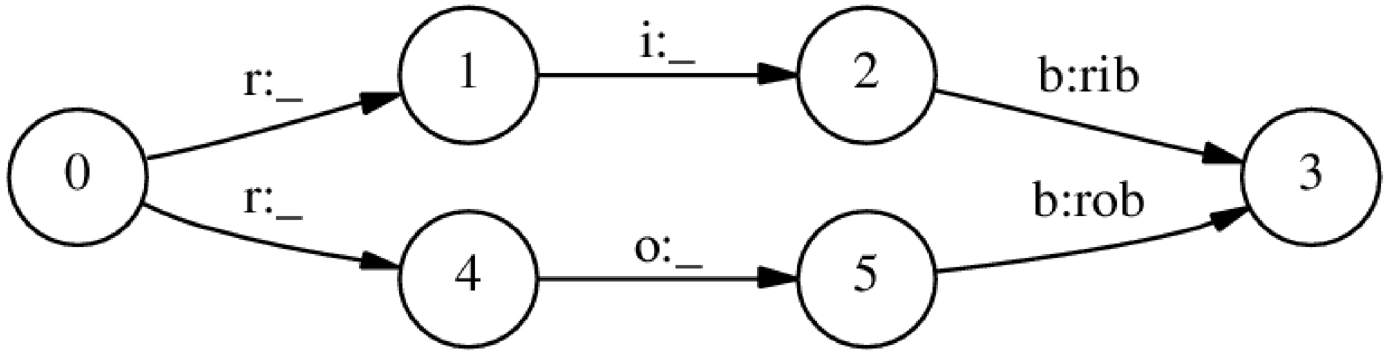
$$\mathcal{O}(\theta) = \sum_{i=1}^{N} \log p_\theta(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$$

- Paper uses iterative scaling to find optimal parameter vector

Overview

- Generative models
- Conditional models
- Label bias problem
- Conditional random fields
- Experiments

Experiment 1: Modeling Label Bias

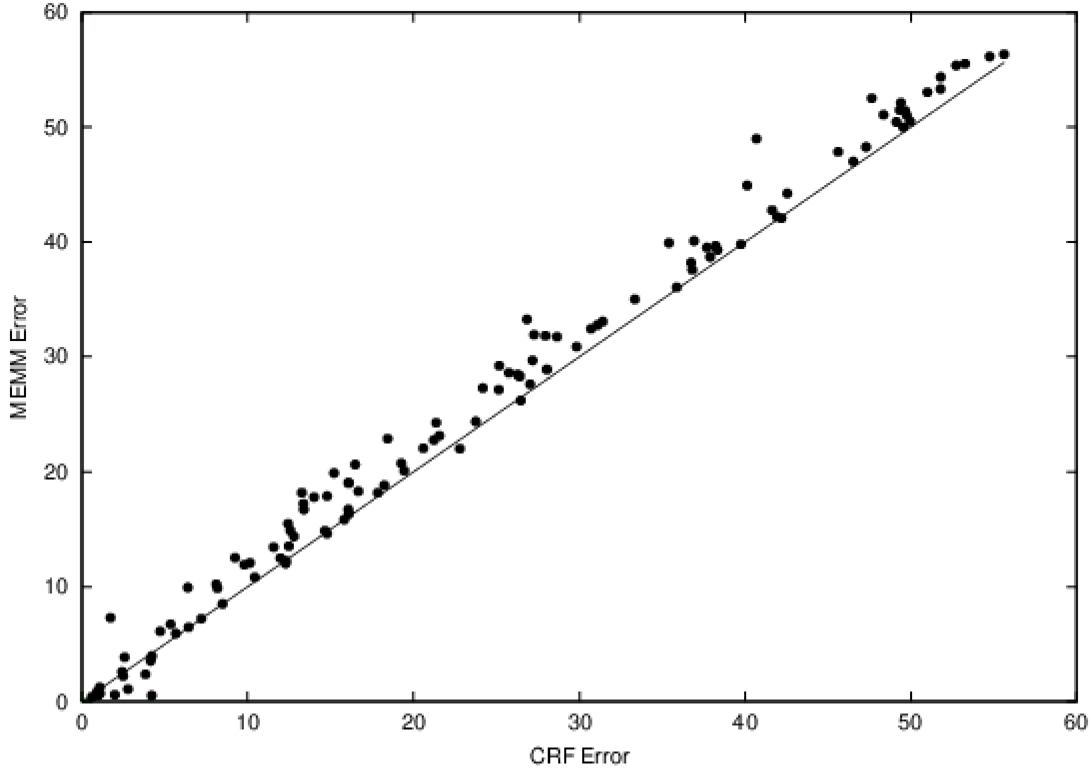


- Generate data from simple HMM that encodes noisy version of network
- Each state emits designated symbol with prob. 29/32
- 2,000 training and 500 test samples
- MEMM error: 42%; CRF error: 4.6%

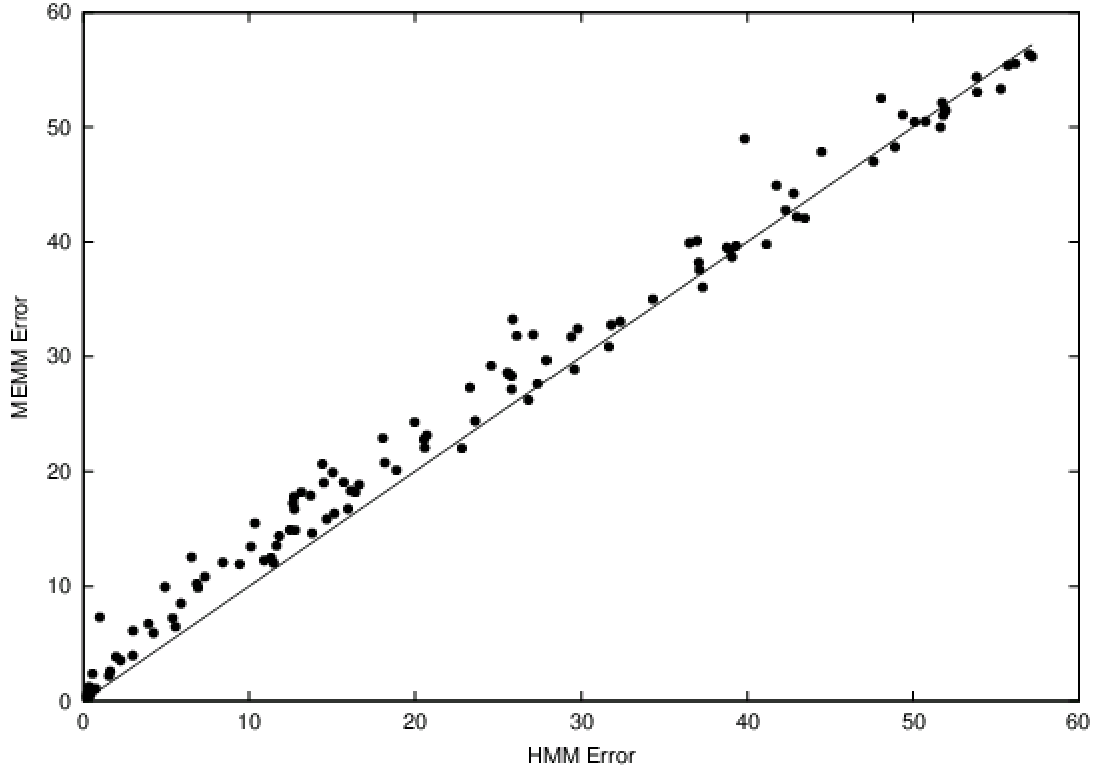
Experiment 2: More synthetic data

- Five labels: a – e
- 26 observation values: A – Z
- Generate data from a mixed-order HMM
- Randomly generate model
- For each model, generate sample of 1,000 sequences of length 25

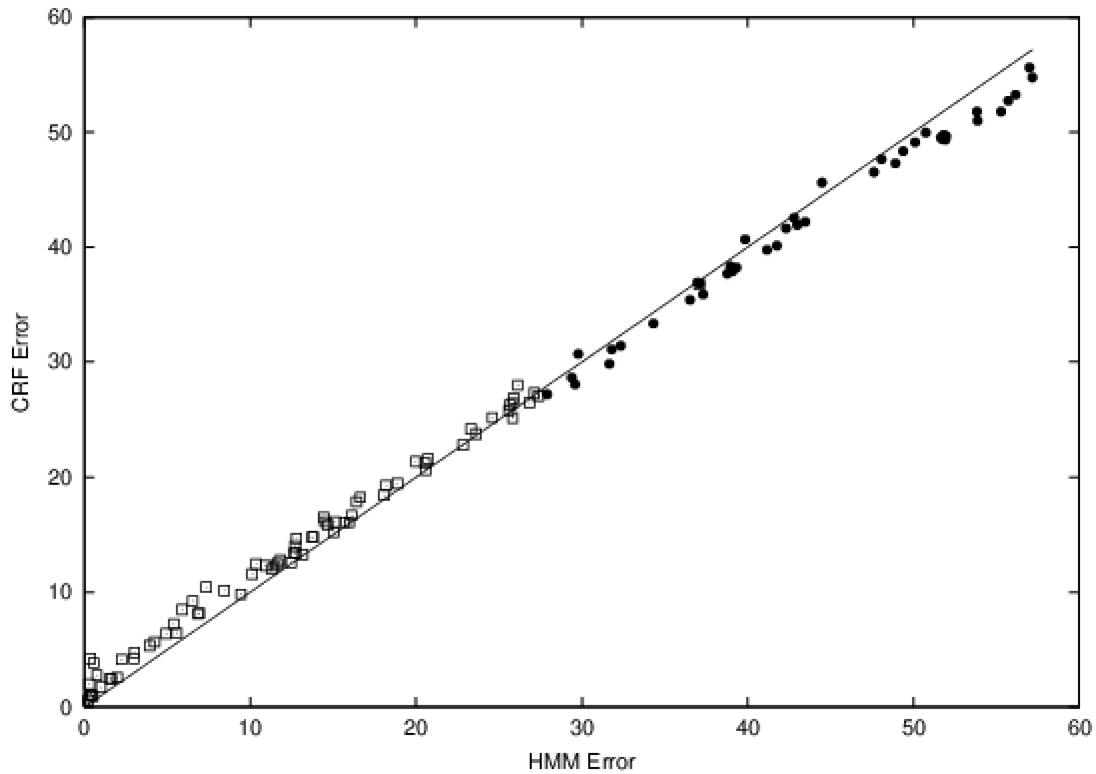
MEMM vs. CRF



MEMM vs. HMM



CRF vs. HMM



Experiment 3: Part-of-speech Tagging

- Each word to be labeled with one of 45 syntactic tags.
- 50%-50% train-test split
- out-of-vocabulary (oov) words: 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%
MEMM^{+}	4.81%	26.99%
CRF^{+}	4.27%	23.76%

$^{+}$Using spelling features

Part-of-speech Tagging

- Second set of experiments: add small set of orthographic features (whether word is capitalized, whether word ends in –ing, -ogy, -ed, -s, -ly …)
- Overall error rate reduced by 25% and oov error reduced by around 50%

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM$^+$	4.81%	26.99%
CRF$^+$	4.27%	23.76%

$^+$Using spelling features

Part-of-speech Tagging

- Usually start training with zero parameter vector (corresponds to uniform distribution)
- Use optimal MEMM parameter vector as starting point for training corresponding CRF
- $MEMM^+$ trained to convergence in around 100 iterations; CRF^+ took additional 1,000 iterations
- When starting from uniform distribution, CRF^+ had not converged after 2,000 iterations

Further Aspects of CRFs

- Automatic feature selection
  - Start from feature-generating rules and evaluate the benefit of the generated features automatically on data

Conclusions

- CRFs do not suffer from the label bias problem!
- Parameter estimation guaranteed to find the global optimum
- Limitation: Slow convergence of the training algorithm