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

John Lafferty & Andrew McCallum & Fernando Pereira

Jing Zhang

Outline

- Motivation
- Conditional Random Fields (CRF)
- Parameter Estimation
- Experiments & Results
- Summary

Motivation – segment & label sequences

- Hidden Markov models (HMMs)

$$p(x, y) = p(y)\ p(x|y)$$

where $p(y) = \prod_{i=1}^{n} p(y_{i+1}|y_i)$

and $p(x|y) = \prod_{i=1}^{n} p(x_i|y_i)$

p(y) – Transition probability

P(x|y) – Emission probability

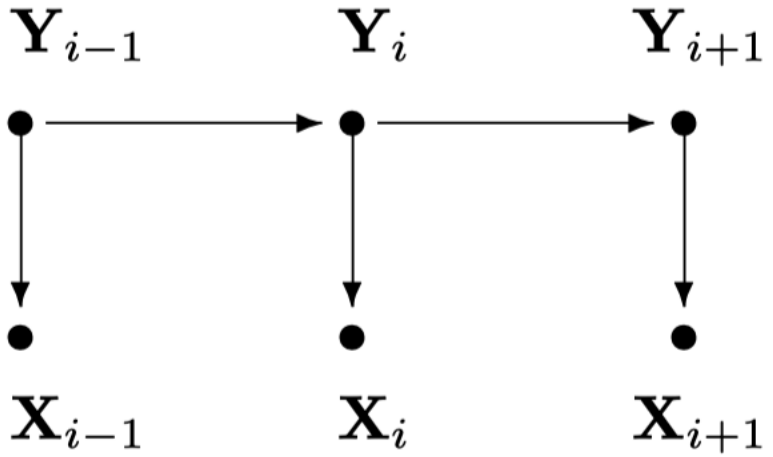


Fig 1. HMMs graphical structure [1]

Motivation – segment & label sequences

- Hidden Markov models (HMMs)

$$p(x, y) = p(y)\ p(x|y)$$

where $p(y) = \prod_{i=1}^{n} p(y_{i+1}|y_i)$

and $p(x|y) = \prod_{i=1}^{n} p(x_i|y_i)$

p(y) – Transition probability

P(x|y) – Emission probability

- Drawback

HMM models direct dependence between each state and **only** its corresponding observation, however, sequence tagging involves words, length, and context.

HMM learns the joint distribution p(x, y), the prediction is to find p(y|x).

Motivation – segment & label sequences

- Maximum entropy Markov models(MEMMs)

$$p(Y|X) = \prod_{i=1}^{n} p(y_i|y_{i-1}, x_i)$$

where $p(y_i|y_{i-1}, x_i) = \frac{\exp(\sum_{i=1} w_i f_i(y_i, y_{i-1}, x_i))}{Z(y_{i-1}, x_i)}$

and $Z(y_{i-1},\ x_i) = \sum_y \exp(\sum_{i=1} w_i f_i(y_i, y_{i-1}, x_i))$

f_i is features

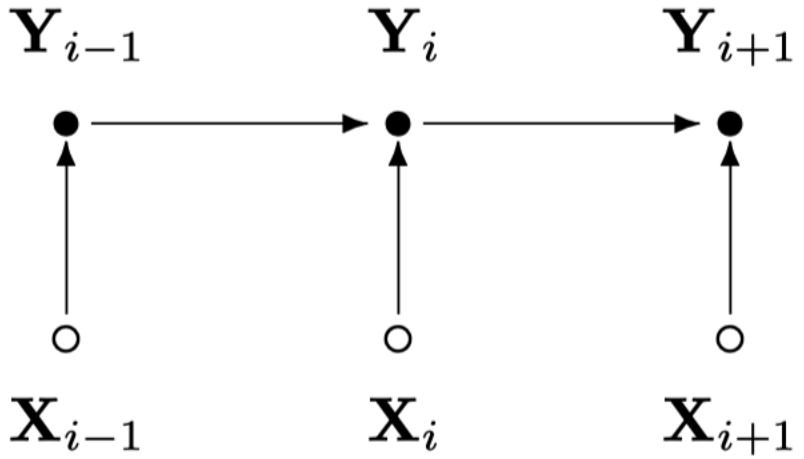


Fig 2. MEMMs graphical structure [1]

Motivation – segment & label sequences

- Maximum entropy Markov models(MEMMs)

$$p(Y|X) = \prod_{i=1}^{n} p(y_i|y_{i-1}, x_i)$$

where $p(y_i|y_{i-1}, x_i) = \frac{\exp(\sum_{i=1} w_i f_i(y_i, y_{i-1}, x_i))}{Z(y_{i-1}, x_i)}$

and $Z(y_{i-1},\ x_i) = \sum_y \exp(\sum_{i=1} w_i f_i(y_i, y_{i-1}, x_i))$

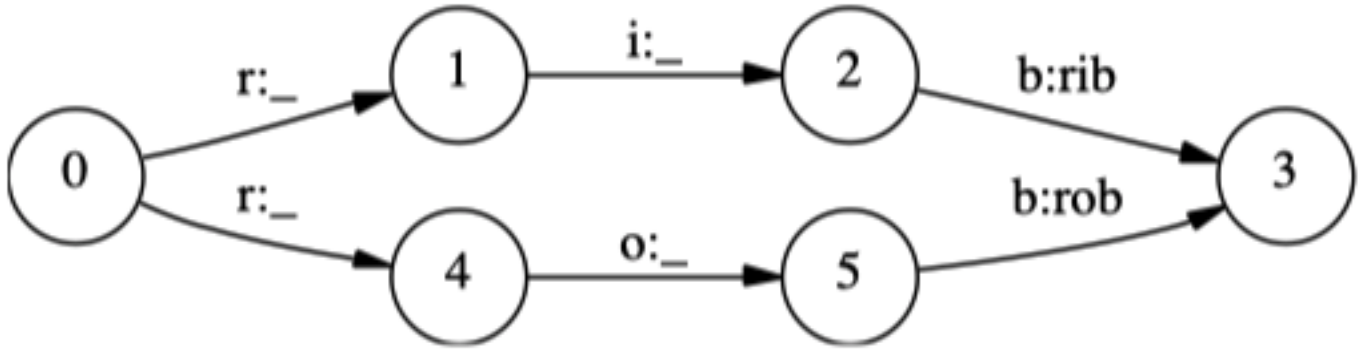
f_i is features

- Drawback

Computation costs larger than HMM.

Label bias problem.

Label bias problem

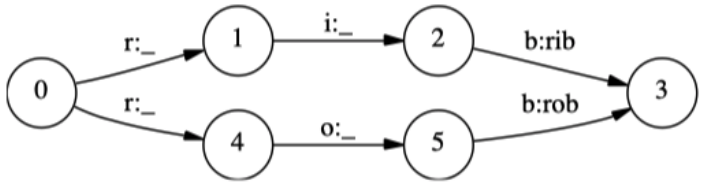


Observation: ‘rib’

If $p(1|r) < p(4|r)$, the bottom path will be chosen regardless of observation.

Label bias problem - solutions

- Determinization of the Finite State Machine
  - Not always possible
  - Leads to combinatorial explosion



- Start with a fully connected model and let the training procedure to find a good structure
  - Prior structural knowledge has proven to be valuable in information extraction tasks

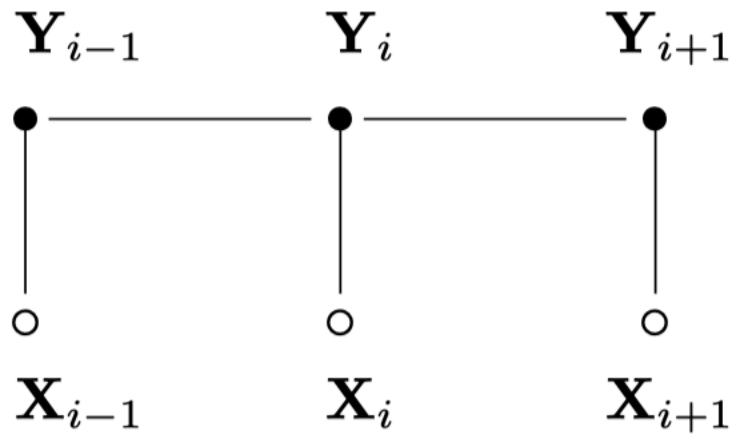
Conditional Random Fields (CRFs) - Definition

- Let G = (V, E) be a finite graph
- **Y** is indexed by the vertices of G
- Then (**X, Y**) is a conditional random field, if random variables $\boldsymbol{Y}_v$ conditioned on **X**, obey the Markov property w.r.t. the graph:

$$p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w \neq v) = p(\boldsymbol{Y}_v|\boldsymbol{X}, \boldsymbol{Y}_w, w \sim v),$$

where $w \sim v$ means that *w* and *v* are neighbors in G.

Conditional Random Fields (CRFs)



CRF is a undirected graphical model globally conditioned on the observation sequence

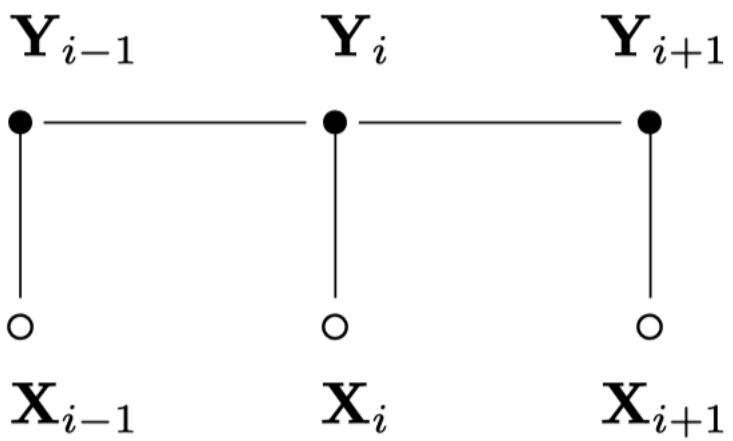
Conditional Random Fields (CRFs)

$$p(Y|X) = \frac{1}{Z(\boldsymbol{X})} \prod_c \psi_c(Y_c|X)$$

Finite state model with unnormalized transition probability

where $\psi_c(Y_c|X) = \exp(\sum_{e\in E,c} \lambda_c f_c(e, y|_e, x) + \sum_{v\in V,c} \mu_c g_c(v, y|_v, x))$

and $Z(X) = \sum_y \exp(\sum_{e\in E,c} \lambda_c f_c(e, y|_e, x) + \sum_{v\in V,c} \mu_c g_c(v, y|_v, x)))$



Parameter Estimation

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

$$\propto \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x}, \mathbf{y}) \log p_\theta(\mathbf{y} \mid \mathbf{x})$$

Log-likelihood objective function[1]

$$\mathcal{O}(\theta) = \log \prod_{x,y} p(x,y)^{\tilde{p}(x,y)}$$

$$= \sum_{x,y} \tilde{p}(x,y) \log p(x,y)$$

$$= \sum_{x,y} \tilde{p}(x,y) \log(\tilde{p}(x)p(y|x))$$

$$= \sum_{x,y} \tilde{p}(x,y) \log p(y|x) + \sum_{x,y} \tilde{p}(x,y) \log \tilde{p}(x)$$

constant

$$\theta = (\lambda_1, \lambda_2, \ldots; \mu_1, \mu_2, \ldots)$$

Parameter Estimation – Iterative Scaling

- Maximizes $\sum_{\mathbf{x},\mathbf{y}} \widetilde{p}(\mathbf{x},\mathbf{y}) \log p_\theta(\mathbf{y}\,|\,\mathbf{x})$ by iteratively updating

$$\lambda_k \leftarrow \lambda_k + \delta\lambda_k \qquad\qquad \mu_k \leftarrow \mu_k + \delta\mu_k$$

- Let an auxiliary function $\mathcal{A}(\theta', \theta) \leq \mathcal{O}(\theta') - \mathcal{O}(\theta)$, then
  - Initialize each λ_k
  - Do until convergence:
    - Solve $\frac{d\mathcal{A}(\theta',\theta)}{d\delta\lambda_k} = 0$ for each $\delta\lambda_k$
    - Update $\lambda_k \leftarrow \lambda_k + \delta\lambda_k$

Parameter Estimation – Iterative Scaling

When $\frac{d\mathcal{A}(\theta',\theta)}{d\delta\lambda_k} = 0$, we can get

$$\widetilde{E}[f_k] \stackrel{\text{def}}{=} \sum_{\mathbf{x},\mathbf{y}} \widetilde{p}(\mathbf{x},\mathbf{y}) \sum_{i=1}^{n+1} f_k(e_i, \mathbf{y}|_{e_i}, \mathbf{x})$$

$$= \sum_{\mathbf{x},\mathbf{y}} \widetilde{p}(\mathbf{x})\, p(\mathbf{y}\,|\,\mathbf{x}) \sum_{i=1}^{n+1} f_k(e_i, \mathbf{y}|_{e_i}, \mathbf{x})\, e^{\,\delta\lambda_k T(\mathbf{x},\mathbf{y})}$$

$T(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \sum_{i,k} f_k(e_i, \mathbf{y}|_{e_i}, \mathbf{x}) + \sum_{i,k} g_k(v_i, \mathbf{y}|_{v_i}, \mathbf{x})$ is the total feature count, which is a global property of (x,y)

Calculating the exponential sum is inefficient. They introduce Algorithm S with global slack feature, and define forward and backward variables, which makes similar $p_\theta(Y_i = y|x) = \frac{\alpha_i(y|x)\beta_i(y|x)}{Z_\theta(x)}$ like posterior as in HMM.

Experiment & Results – Modeling label bias

1. Data was generated from a simple HMM which encodes a noisy version of the finite-state network (“rib/ rob”)
2. Train both an MEMM and a CRF
3. The observation features are simply the identity of the observation symbols.
4. 2, 000 training and 500 test samples were used
5. Results:
   - CRF error: 4.6%
   - MEMM error: 42%
6. Conclusion:
   - MEMM fails to discriminate between the two branches and we get the label bias problem

Experiment & Results – Modeling mixed order sources

1. Data was generated from a mixed-order HMM with state transition probabilities given by

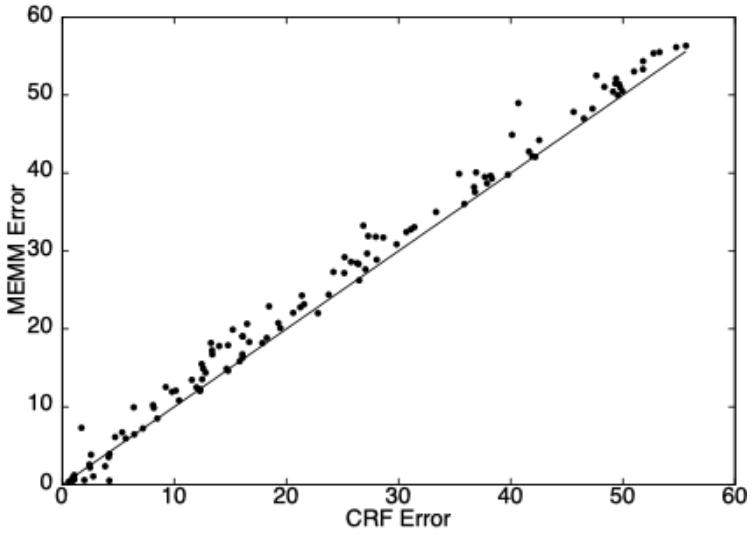
   $$p(y_i|y_{i-1}, y_{i-2}) = \alpha\, p_2(y_i|y_{i-1}, y_{i-2}) + (1-\alpha)p_1(y_i|y_{i-1})$$

2. Emission probabilities given by $p(x_i|y_i, x_{i-1}) = \alpha\, p_2(x_i|y_i, x_{i-1}) + (1-\alpha)p_1(x_i|y_i)$

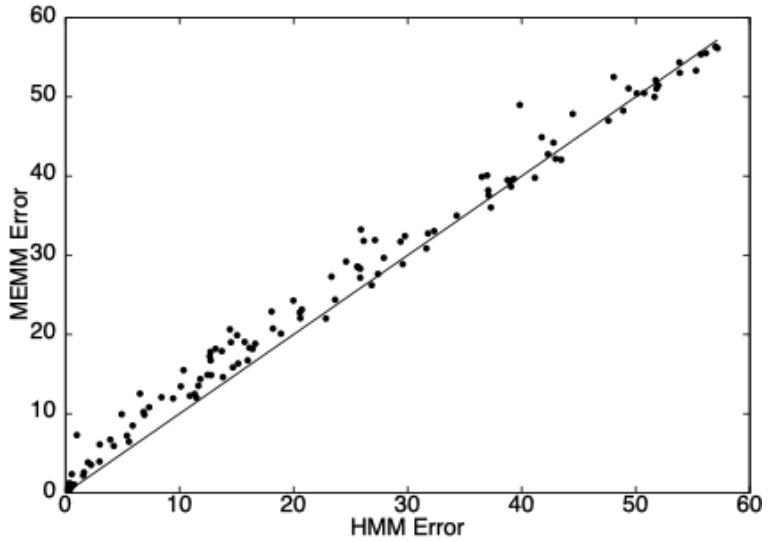
3. For each randomly generated model, a sample of 1,000 sequences of length 25 is generated for training and testing.

Experiment & Results – Modeling mixed order sources

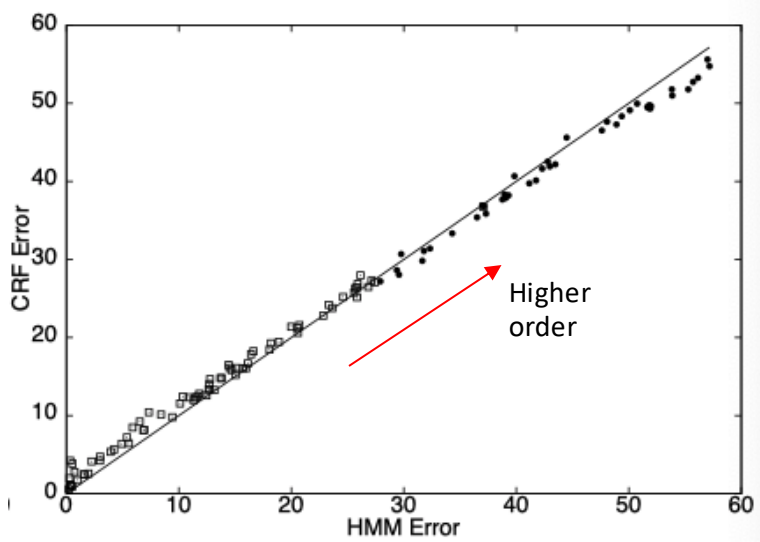
- Comparison of error rates on synthetic data



CRF outperforms MEMM



HMM outperforms MEMM



CRFs achieve the lowest error rate for higher order data

Experiment & Results – Parts of Speech tagging

45 syntactic tags

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

Using same set of features: CRF > HMM > MEMM
Using additional overlapping features: CRF > MEMM >> HMM

Summary

- Properties of CRFs:
  - Discriminatively trained models for sequence segmentation and labeling
  - Combination of arbitrary and overlapping observation features from both the past and future
  - Efficient training and decoding based on dynamic programming for a simple chain graph
  - Parameter estimation guaranteed to find the global optimum
- Disadvantage of CRFs:
  - Slow convergence during training
- Future directions:
  - More complex graph structures
  - Faster learning algorithms
  - Feature induction algorithms for CRFs.

Thanks!