Probabilistic Global Models: CRFs

Motivation

- Problems of MEMMs and variants of chained sequential inference schemes with local classifiers **[Punyakanok et al., 2002; Giménez and Màrquez, 2003; Kudo and Matsumoto, 2001]**
  - ⋆ Training is local, without taking into account loss functions derived from global performance measures
  - ⋆ *Label bias problem*

Probabilistic Global Models: CRFs

Motivation

- Problems of MEMMs and variants of chained sequential inference schemes with local classifiers **[Punyakanok et al., 2002; Giménez and Màrquez, 2003; Kudo and Matsumoto, 2001]**
  - ⋆ Training is local, without taking into account loss functions derived from global performance measures
  - ⋆ *Label bias problem*
- The problems of generative models are also well-known (e.g., they cannot use arbitrary representations on the inputs)

Probabilistic Global Models: CRFs

Motivation

- Problems of MEMMs and variants of chained sequential inference schemes with local classifiers **[Punyakanok et al., 2002; Giménez and Màrquez, 2003; Kudo and Matsumoto, 2001]**
  - ⋆ Training is local, without taking into account loss functions derived from global performance measures
  - ⋆ *Label bias problem*
- The problems of generative models are also well-known (e.g., they cannot use arbitrary representations on the inputs)
- Conditional Random Fields **[Lafferty, McCallum, and Pereira 2001]** try to get the best of both worlds without any of the shortcomings

Conditional Random Fields

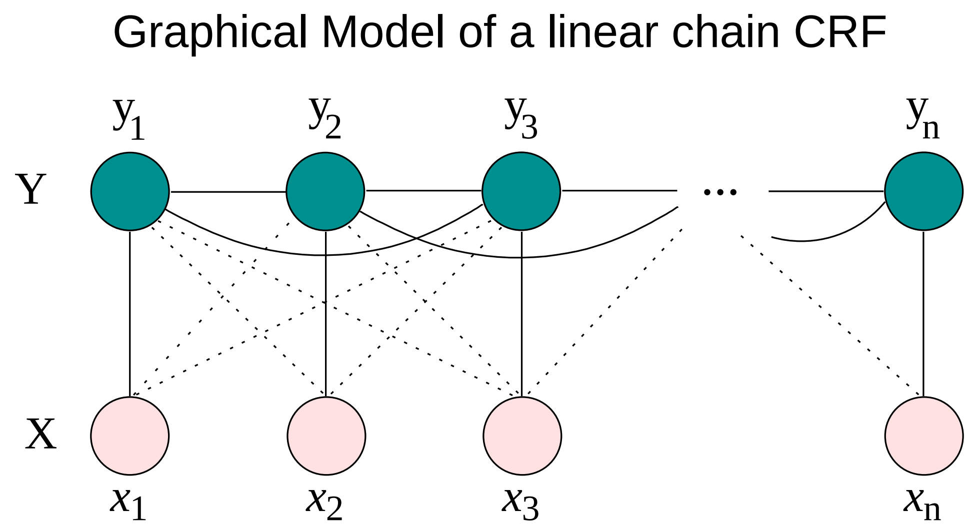
- CRF is a conditional model $p(\mathbf{y}|\mathbf{x})$
- It defines a single log-linear distribution over label structure ($\mathbf{y}$) given the observations ($\mathbf{x}$)
- CRF can be viewed as an undirected graphical model or Markov random field globally conditioned on $\mathbf{x}$

Conditional Random Fields

- CRF is a conditional model $p(\mathbf{y}|\mathbf{x})$
- It defines a single log-linear distribution over label structure ($\mathbf{y}$) given the observations ($\mathbf{x}$)
- CRF can be viewed as an undirected graphical model or Markov random field globally conditioned on $\mathbf{x}$
- A graphical model is a family of probability distributions that factorize according to an underlying graph.
- Represent the distribution over a large number of random variables by a product of local functions that each depend only on a small number of variables

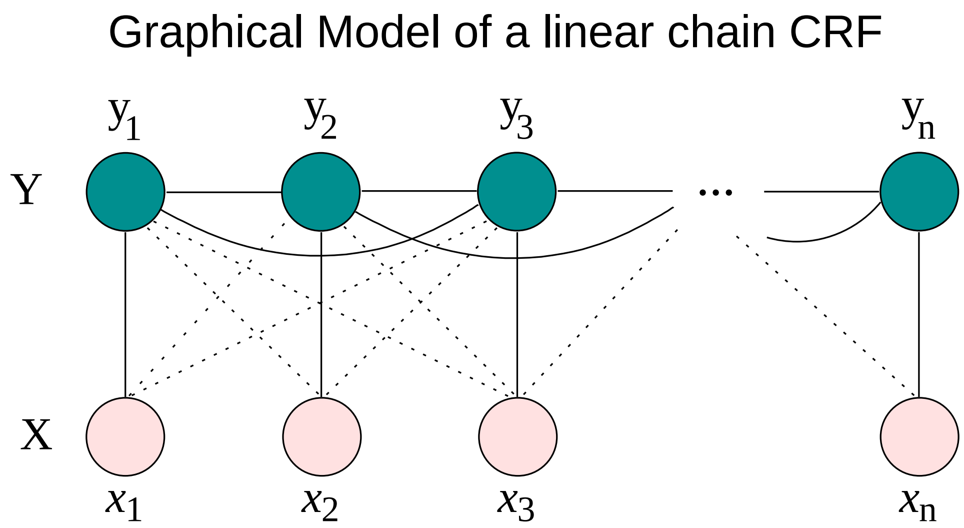
Conditional Random Fields

- The most common instantiation is the Linear-chain CRF model



Conditional Random Fields

- The most common instantiation is the Linear-chain CRF model



- Two types of dependencies: (y_{i-1}, y_i) and $(\mathbf{x}, y_i)$

- Training and decoding are efficient

- Direct application to all (NLP) sequential labeling problems

Linear-chain CRFs

- $p(\mathbf{y}|\mathbf{x})$ factorize in a normalized product of potential functions of the form: $\boxed{\exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, \mathbf{x}, i) + \sum_k \mu_k s_k(y_i, \mathbf{x}, i))}$
- $t_j(y_{i-1}, y_i, \mathbf{x}, i)$ is a transition feature function
- $s_k(y_i, \mathbf{x}, i))$ is a state feature function
- λ_j and μ_k are the parameters to be estimated from training data

Linear-chain CRFs

- $p(\mathbf{y}|\mathbf{x})$ factorize in a normalized product of potential functions of the form: $\boxed{\exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, \mathbf{x}, i) + \sum_k \mu_k s_k(y_i, \mathbf{x}, i))}$
- $t_j(y_{i-1}, y_i, \mathbf{x}, i)$ is a transition feature function
- $s_k(y_i, \mathbf{x}, i))$ is a state feature function
- λ_j and μ_k are the parameters to be estimated from training data

t_j and s_k are indicator functions. Example:

$$t_j(y_{i-1}, y_i, \mathbf{x}, i) = \begin{cases} 1 & \text{if } y_{i-1} = \text{IN and } y_i\text{=NNP and } x_i = \text{“September”} \\ 0 & \text{otherwise} \end{cases}$$

Linear-chain CRFs

- Expressing t_j and s_k as a general $f_j(y_{i-1}, y_i, \mathbf{x}, i)$
- ...and considering $F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i)$
- The probability of a label sequence **y** given an observation sequence **x** is

Linear-chain CRFs

- Expressing t_j and s_k as a general $f_j(y_{i-1}, y_i, \mathbf{x}, i)$
- ...and considering $F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i)$
- The probability of a label sequence **y** given an observation sequence **x** is

$$\boxed{p(\mathbf{y}|\mathbf{x}, \lambda) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_j \lambda_j F_j(\mathbf{y}, \mathbf{x})\right)}$$

- where $Z(\mathbf{x})$ is a normalization factor

Linear-chain CRFs

- Expressing t_j and s_k as a general $f_j(y_{i-1}, y_i, \mathbf{x}, i)$
- ...and considering $F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i)$
- The probability of a label sequence **y** given an observation sequence **x** is

$$\boxed{p(\mathbf{y}|\mathbf{x}, \lambda) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_j \lambda_j F_j(\mathbf{y}, \mathbf{x})\right)}$$

- where $Z(\mathbf{x})$ is a normalization factor
- This is a log-linear probability distribution similar to ME

CRFs: Parameter estimation

- Optimize the *conditional log-likelihood* of λ on the training set
- $l(\lambda) = \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}|\mathbf{y}^{(i)})$
- $l(\lambda) = \sum_{i=1}^{N} \sum_j \lambda_j F_j(\mathbf{y}^{(i)}, \mathbf{x}^{(i)})) - \sum_{i=1}^{N} \log Z(\mathbf{x}^{(i)}) - \sum_j \frac{\lambda_j^2}{2\sigma^2}$
- $\frac{1}{2\sigma^2}$ is a regularization parameter
- Several methods can be used for training
- Cost: $O(nM^2NG)$

CRFs: inference

- Decoding: $\boxed{\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \lambda)}$
- This can be done by using variants of the Viterbi dynamic programming for HMMs

CRFs: applications

- NLP: Text classification, POS tagging, chunking, named-entity recognition, semantic role labeling, etc.
  (See the survey by **[Sutton and McCallum, 2006]**)
- Bioinformatics
- Computer vision