CRF模型族及其在语音识别中的应用

张超

CSLT, RIIT, Tsinghua University

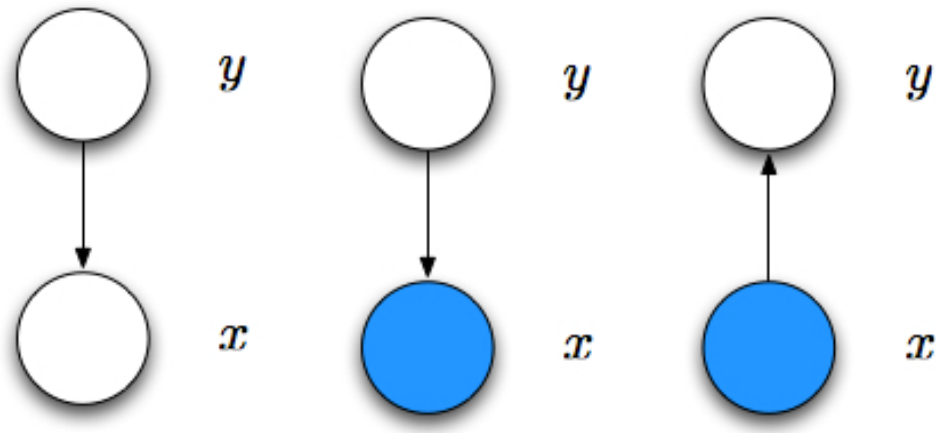
目录

图概率模型

- 节点(node, vertex, site): 随机变量
- 边(link, arc, edge): 随机变量间的概率约束关系
- 分类: 有向图模型, 无向图模型
- 例: $p(\omega_1, \omega_2, \omega_3) = p(\omega_1)p(\omega_2|\omega_1)p(\omega_3|\omega_2)$

Bayes定理的图表示

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



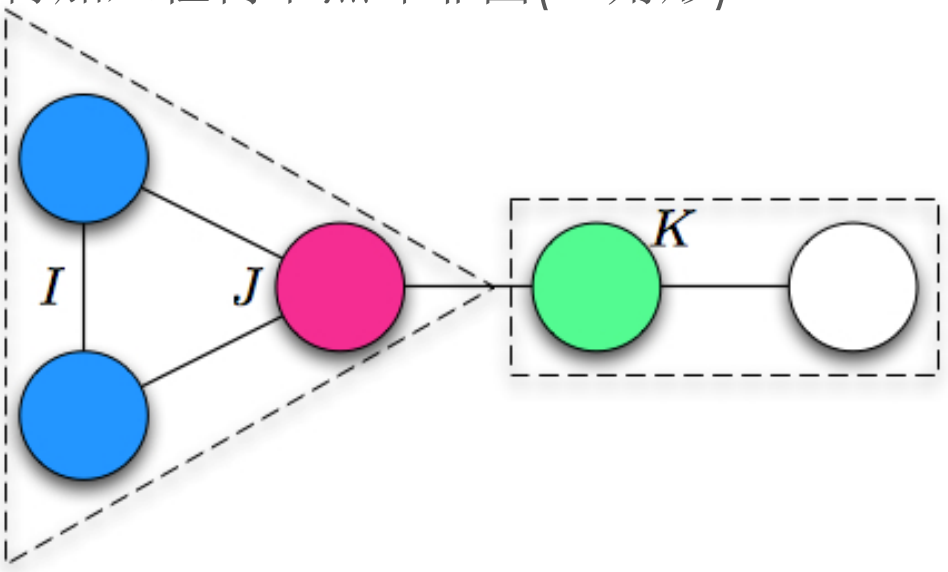
- 观测到 x，

  $p(x) = \sum_{y'} p(x|y')p(y')$

- $p(y|x) = \dfrac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')}$ → 推理

无向图的分解

- 条件独立: I 到 K 的路径必经过 J → $p(I, K|J) = p(I|J)p(K|J)$
- 团(clique): 任何两节点间有边连接的节点子集(三角形、矩形)
- 最大团: 再加入任何节点即非团(三角形)



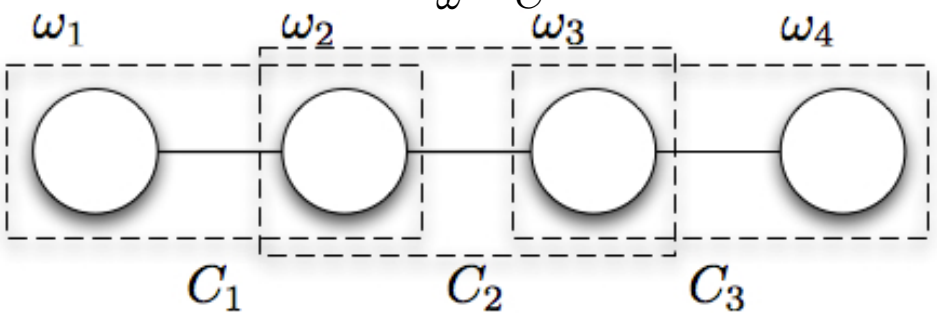
无向图的分解

- 设: 团C中的随机变量为$\vec{\omega_C} = (\vec{x}_C, \vec{y}_C)$, $\Psi_C(\vec{\omega_C})$是势函数(实值)
- 则由乘法原理, 联合分布可写作:

$$p(\vec{\omega}) = \frac{1}{Z}\prod_C \Psi_C(\vec{\omega_C})$$

$$= p(\vec{x}, \vec{y})$$

$$\text{其中 } Z = \sum_{\vec{\omega}}\prod_C \Psi_C(\vec{\omega_C})$$



无向图的分解

- 通常取$\Psi(\vec{\omega_C}) = \exp\{-E(\vec{\omega_C})\}$
  - 势函数 > 0, 出于方便取势函数为指数形式
  - $E(\vec{\omega_C})$能量函数; $\Psi(\vec{\omega_C})$玻尔兹曼分布
  - $p(\vec{\omega})$可视为每个子图上的能量之和
- Hammersley-Clifford定理
  - *UI*: 是与无向图中通过算法可得到的条件非独立性一致的分布组成的集合
  - *UF*: 是可看作 $p(\vec{\omega})$ 中某个因式的分布的集合
  - *UI*=*UF*=马尔可夫随机场(i.e.属于"完美图")
  - 什么是马尔可夫随机场?

随机场

- 随机场
  - 图染色问题: $\mathcal{X}$ 图中节点序列
  - $\Omega = \prod_{x \in \mathcal{X}} \Omega_x$, Ω_x是节点 x 的可能着色; $\mathcal{F}$是Ω的幂集
  - P 是$(\Omega, \mathcal{F})$上的一种概率测度; $P(\omega)$是概率密度($\omega \in \Omega$)
  - $(\Omega, \mathcal{F}, P)$是$\mathcal{X}$上的随机场(i.e.概率空间)
- 马尔可夫随机场
  - 马氏性: 邻接节点
  - 马尔可夫随机场: 即无向图模型(Markov Networks, Finite Lattice etc.)

随机场的最大熵分布

- $\because p(\vec{\omega}_C) = \frac{1}{Z}\prod_C \Psi_C(\vec{x_C}, \vec{y_C}) = \frac{1}{Z}\prod_C \exp\{-E_C(\omega_C)\}$

  由随机场的连续形式最大熵分布定理，该随机场的最大熵分布为：

  $$p(\vec{\omega}_C) = \frac{1}{Z}\prod_C \exp\{\sum_{k=1}^{K} \lambda_k f_k(\omega_C)\}$$

- 最大熵分布：对某类分布，若除了它属于某类C外对它一无所知，根据最大熵原则，默认应该选择最大熵的分布
  - 最大熵相当于最小化分布所固有的先验信息(信息：负熵)
  - 许多物理系统随着时间推移倾向于变为最大熵的形式

线性条件随机场(CRF)

- 线性图→一阶马尔可夫性 $\vec{x} = \{x_1, x_2, \ldots, x_T\}$, $\vec{y} = \{y_1, y_2, \ldots, y_T\}$
- 其最大熵分布为:

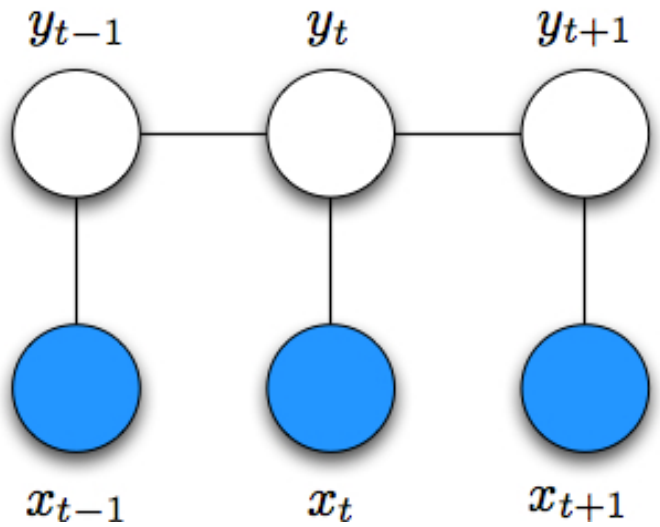
$$p(\vec{x}, \vec{y}) = \frac{1}{Z} \prod_{C} \Psi_C(\vec{x_C}, \vec{y_C}) = \frac{1}{Z} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \vec{x}_t)\}$$

- 则对应条件分布为:

$$p(\vec{y}|\vec{x}) = \frac{p(\vec{x}, \vec{y})}{\sum_{\vec{y}'} p(\vec{x}, \vec{y}')} = \frac{\exp\{\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \vec{x}_t)\}}{\sum_{\vec{y}'} \exp\{\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y'_t, y'_{t-1}, \vec{x}_t)\}}$$

线性条件随机场(CRF)

- 条件随机场(CRF): $p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \vec{x}_t)\}$
  - 其中$Z(\vec{x}) = \sum_{\vec{y}'} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t', y_{t-1}', \vec{x}_t)\}$



线性条件随机场的构造

- 训练集: $D = \{\vec{x}^{(i)}, \vec{y}^{(i)}\}_{i=1}^{N}, (\vec{x}^{(i)}, \vec{y}^{(i)}) = \begin{cases} \vec{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}\} \\ \vec{y}^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \ldots, y_T^{(i)}\} \end{cases}$
- 目标: 求出使得 $p(\vec{y}|\vec{x})$ 最大化的参数 $\theta = \{\lambda_k\}_{k=1}^{K}$
- 最大化对数似然(maximum likelihood):
  - $l(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_t^{(i)}, y_{t-1}^{(i)}, \vec{x}_t^{(i)}) - \sum_{i=1}^{N} \log Z(\vec{x}^{(i)})$
    - 线性函数
    - 过拟合(归一化因子过大时 $l(\theta)$ 不再是凸函数，无全局最优解)

线性条件随机场的构造

- 解决过拟合: 正则化 or k-折交叉验证
- 正则化: L2正则化+正则化参数$1/\delta^2$
- $$l(\theta)=\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_t^{(i)},y_{t-1}^{(i)},\vec{x}_t^{(i)})-\sum_{i=1}^{N}\log Z(\vec{x}^{(i)})-\sum_{k=1}^{K}\frac{\lambda_k^2}{2\delta^2}$$
  - 把 θ 看作0均值，方差为$\delta^2\mathbb{I}$的高斯分布
  - 正则化可看作对θ 做MAP估计
- 设 $C_k(\vec{y},\vec{x})=\sum_{t=1}^{T}f_k(y_t,y_{t-1},\vec{x}_t)$，为特征的计数器

线性条件随机场的构造

- $$\frac{\partial l}{\partial \lambda_k} = \sum_{i=1}^{N} C_k(\vec{y}^{(i)}, \vec{x}^{(i)}) - \sum_{i=1}^{N} \sum_{\vec{y}'} p(\vec{y}' | \vec{x}^{(i)}) C_k(\vec{y}', \vec{x}^{(i)}) - \frac{\lambda_k}{\delta^2}$$

  $$= N(E[f_k] - E_\theta[f_k])$$

  - 其中，$E[f_k]$为特征函数f_k的经验期望
  - $E_\theta[f_k]$为f_k在模型 θ 上的期望值

- L1正则化

线性条件随机场的构造

- $\frac{\partial l}{\partial \lambda_k}$ 梯度计算:
  - Newton's Method: Hessian矩阵计算目标函数二阶导数 → 过大的运算量
  - quasi-Newton's method(BFGS): 使用目标函数一阶导数近似Hessian矩阵 → 仍需计算二阶信息, 耗内存
  - Limited-memory BFGS(L-BFGS): 使用共轭梯度近似二阶信息 → 最常用
- 进行迭代优化, 收敛到全局最优解
- 其它方法: Precond. CG, Mixed CG, Plain CG, GIS etc.

线性条件随机场的构造

- 推理(inference):
  - 训练: 计算$p(\vec{y}|\vec{x})$→梯度: $Z(\vec{x})$
  - 解码: $\vec{y}^* = \arg\max_{\vec{y}} p(\vec{y}, \vec{x})$
- 边缘分布$p(y_t, y_t - 1|\vec{x})$:
  - 前向后向算法(forward-backward algorithm)
  - $\alpha_t(y) = \sum_{y'} \alpha_{t-1}(y') \exp(\sum_{k=1}^{K} \lambda_k f_k(y', y, \vec{x}_{t-1}))$
  - $\beta_t(y) = \sum_{y'} \beta_{t+1}(y') \exp(\sum_{k=1}^{K} \lambda_k f_k(y', y, \vec{x}_{t+1}))$
  - $p(y_t, y_t - 1|\vec{x}) \propto \alpha_{t-1}(y_{t-1}) \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \vec{x}_t)\} \beta_t(y_t)$

线性条件随机场的构造

- 分治函数$Z(\vec{x})$:
  - $Z(\vec{x}) = \sum_{\vec{y}'} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t', y_{t-1}', \vec{x}_t)\}$
  - 前向后向算法
  - 训练时需要; 解码时不需要(如需准确概率则需要)
- Viterbi解码:
  - $\delta_t(j) = \max_{i \in \mathcal{Y}} \exp\{\sum_{k=1}^{K} \lambda_k f_k(i, j, \vec{x}_t)\} \delta_{t-1}(i)$
  - $i, j \in \mathcal{Y}$, $\mathcal{Y}$是输出的所有可能状态(CRF中可能的输出与时间无关)
  - 动态规划

其它角度看CRF

- 有限自动机: WFA $A = (\Sigma, Q, q_s, F, E, \rho)$
  - WFA $A^{'}$, $L_A = L_{A'}, \forall \pi \in \Pi_A, \exists \pi^{'} \in \Pi_{A'}$, 标签 $l[\pi] = l[\pi^{'}]$
  - $A \circ A^{'}$: Intersection, $L_{A \circ A'}[\pi] = L_{A'}[\pi] \bigcap L_A[\pi]$
    $w_{A \circ A'}[\pi] = w_{A'}[\pi] + w_A[\pi]$, $l[\pi] = l[\pi_1] = l[\pi_2]$
  - 找使对数似然最大化的组合权值
- 对数线性模型
- Exponential Family
  - 可以用来估计任何概率密度
  - → Kernel Function

其它角度看CRF

- 机器学习理论:
  - $\mathfrak{B}$ 是有限图集合, $\mathfrak{g} \in \mathfrak{B}$, 图上的团的集合为 $C(\mathfrak{g})$
  - 学习函数 $h : \mathcal{X}(\mathfrak{B}) \to \mathcal{Y}(\mathfrak{B})$, $h(\mathfrak{g}, \vec{x}) \in \mathcal{Y}(\mathfrak{g})$
  - 定义损失函数: 最小化损失函数

$$\phi(\vec{y}, f(\mathfrak{g}, \vec{x})) = -\sum_{c \in \mathcal{C}(\mathfrak{g})} f_c(\vec{x}, \vec{y}_c) + \log \sum_{\vec{y}' \in \mathcal{Y}((g))} \exp(\sum_{c \in \mathcal{C}(\mathfrak{g})} f_c(\vec{x}, \vec{y}_c'))$$

  - CRF: $p(\vec{y}|\mathfrak{g}, \vec{x}) = Z^{-1}(\mathfrak{g}, \vec{x}, f) \exp(\sum_c f_c(\vec{x}, \vec{y}_c))$
- 逻辑回归推广
- Gibbs模型

CRF vs. HMM

- HMM:
  - 独立性假设
    - 每状态仅依赖其前一状态→转移$p(y_t|y_{t-1})$
    - 输入特征仅与当前状态输出有关→观测$p(x_t|y_t)$
  - $p(\vec{x},\vec{y}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t)$
- 从马尔可夫随机场看HMM:

$$p(\vec{x},\vec{y}) = \prod_{t=1}^{T} \frac{1}{Z} \exp\{ \sum_{i,j\in\mathscr{Y}} \lambda_{ij} \mathbf{1}_{y_t=i}\mathbf{1}_{y_{t-1}=j} + \sum_{i\in\mathscr{Y}}\sum_{o\in\mathscr{O}} \mu_{oi}\mathbf{1}_{y_t=i}\mathbf{1}_{x_t=o}\}$$

$$= \frac{1}{Z}\prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\}$$

CRF vs. HMM

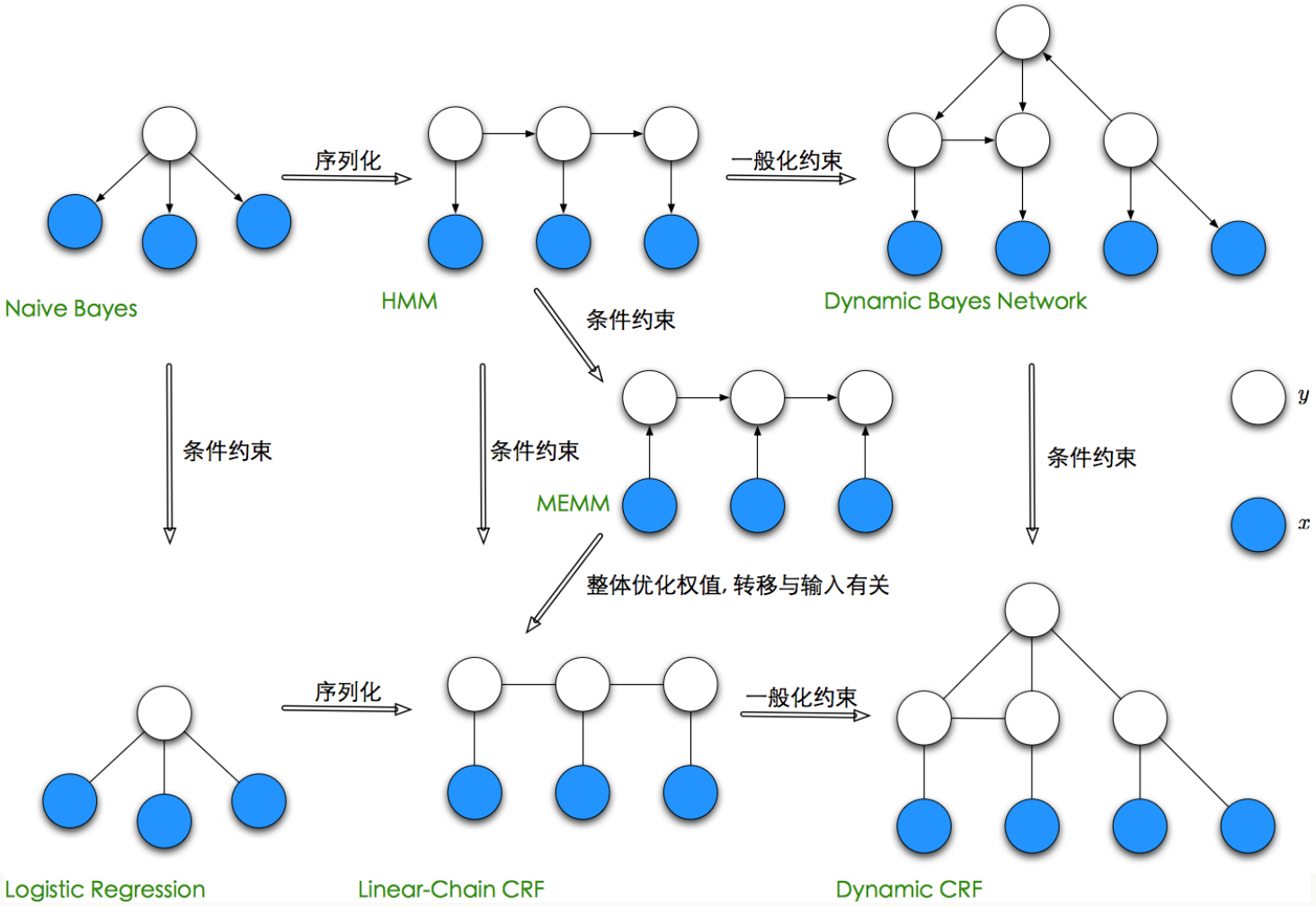
- CRF优于HMM:
  - $p(y_t|y_{t-1})$vs.$f_k(y_t, y_{t-1}, \vec{x}_t)$→ CRF转移可以依赖于相应输入 (Label Bias Problem)
  - $p(x_t|y_t)$vs.$f_k(y_t, y_{t-1}, \vec{x}_t)$→ CRF观测可以与前一状态有关
  - 任何单一时刻, CRF可以使用所有需要的输入$\vec{x}_t$
  - CRF可以综合所需各种冗余特征, 并自动学习对应scaling factor
  - CRF可达最优解, HMM容易陷入局部最优
- CRF劣于HMM:
  - HMM训练更快, 更省内存(无需计算归一化因子)
  - CRF冗余特征的选择缺乏明确指导
- 深层原因?

区分性模型VS生成性模型

- 生成式模型: $p(\vec{x}, \vec{y})$
  - 有向图模型(通常用有向图表达)
  - 输入$\vec{x}$ 不指向输出$\vec{y}$
  - 直接描述了如何产生输入$p(\vec{x}|\vec{y})$
- 区分性模型: $p(\vec{y}|\vec{x})$
  - 不需估计更容易整合冗余特征
  - 不需估计 $p(\vec{x})$ 通常包含难以估计的复杂依赖关系, 且存在未见数据
  - CRF说明了在 $\vec{y}$上考虑非独立性假设比在$\vec{x}$ 上合适(CRF无$p(\vec{x})$)

    → 退化的区分性模型比生成式模型更鲁棒

生成性/区分性图模型

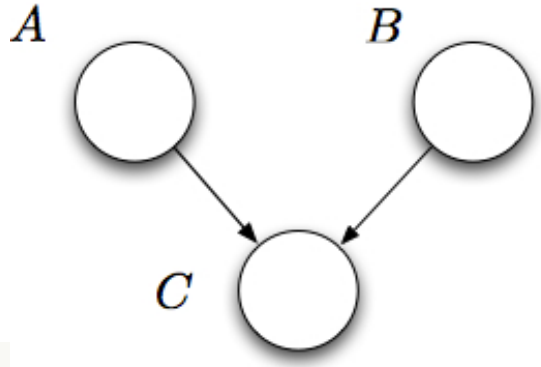

序列化
一般化约束
Naive Bayes
HMM
Dynamic Bayes Network
条件约束
条件约束
条件约束
条件约束
MEMM
y
x
整体优化权值，转移与输入有关
序列化
一般化约束
Logistic Regression
Linear-Chain CRF
Dynamic CRF

有向/无向图关系

- 有向图对应(节点不变)于无向图
  - 等价条件:

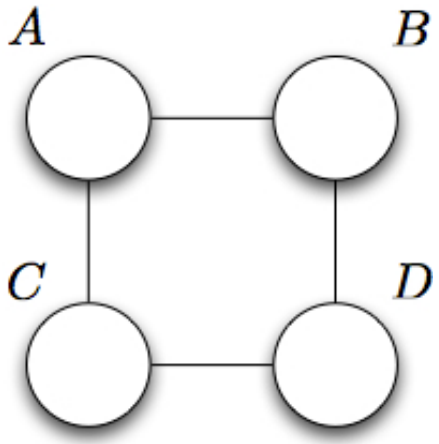
$$\begin{cases} \Psi_{1,2}(\omega_1, \omega_2) = p(\omega_1)p(x_2|x_1) \\ \Psi_{2,3}(\omega_2, \omega_3) = p(\omega_3|\omega_2) \end{cases}$$

- 有向图无对应无向图

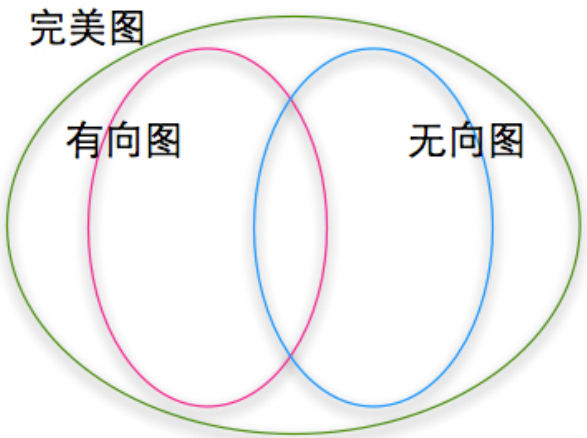


有向/无向图

- 无向图无对应有向图

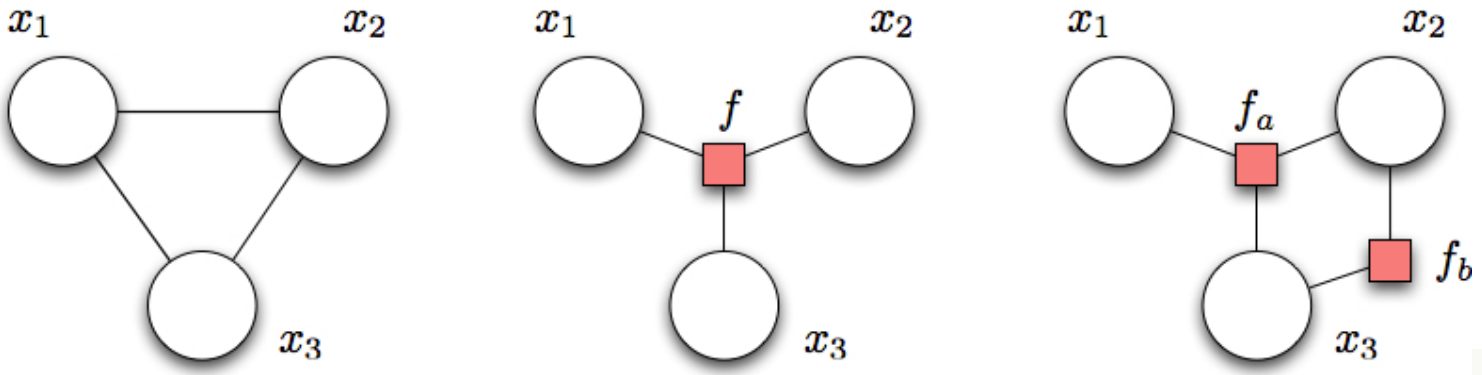


- 两种图的表述能力



更一般的CRF

- 任意图结构的CRF: $p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{C} \exp\{\sum_{k=1}^{K_C} \lambda_{Ck} f_{Ck}(\vec{x}_C, \vec{y}_C)\}$
  - 邻节点方式定义马尔可夫性
  - 用于表示复杂的依赖关系
  - 用于增加隐变量
- 复杂图结构→图分解(有向/无向图)
  - 显式增加分解因子图的节点



更一般的CRF

- 复杂图结构→团模板
  - $\mathcal{C} = \{C_p\}_{p=1}^{P}$
  - C_p 是参数共享的团模板
- 复杂图结构→推理
  - Belief Propagation(前向后向算法推广)与Viterbi算法: NP-hard
  - 随机初值加快收敛速度: MCMC
  - 近似推理: Loopy Belief Propagation (不确保收敛)
- 复杂图结构
  - 其它算法: TRP, BLP etc.

更一般的CRF

- 隐变量$\vec{w}$
  - $p(\vec{y},\vec{w}|\vec{x}) = \frac{1}{Z(\vec{x})}\prod_{C_p\in\mathcal{C}}\prod_{\Psi_{C\in C_p}}\Psi_C(\vec{x}_C,\vec{w}_C,\vec{y}_C;\theta_p)$
  - $l(\theta) = \log p(\vec{y}|\vec{x}) = \log\sum_{\vec{w}} p(\vec{y},\vec{w}|\vec{x})$
  - $p(\vec{w}|\vec{y},\vec{x}) = \frac{1}{Z(\vec{x},\vec{y})}\prod_{C_p\in\mathcal{C}}\prod_{\Psi_{C\in C_p}}\Psi_C(\vec{x}_C,\vec{w}_C,\vec{y}_C;\theta_p)$
  - $p(y|x) = \frac{p(\vec{y},\vec{w}|\vec{x})}{p(\vec{w}|\vec{y},\vec{x})} = \frac{Z(\vec{x},\vec{y})}{Z(\vec{x})}$
  - $Z(\vec{x},\vec{y}) = \sum_{\vec{w}}\prod_{C_p\in\mathcal{C}}\prod_{\Psi_{C\in C_p}}\Psi_C(\vec{x}_C,\vec{w}_C,\vec{y}_C;\theta_p),\ Z(\vec{x}) = \sum_{\vec{y}} Z(\vec{x},\vec{y})$

更一般的CRF

- 隐变量
  - $l(\theta)$ 非凸, 易陷入局部最优
  - 共轭梯度(L-BFGS, GIS), Generalized EM算法 etc.
- 隐变量的优点: 标注不足; 对问题认识不足

CRFs应用介绍

- 线性CRF
- 复杂图结构CRF: TCRF, Skip-Chain CRF, 2D CRF, Semi-Markov CRF, *SQL定义结构CRF
- 隐变量CRF: HCRF, GHCRF, LDCRF, FCRF
- 其它: *Kernel CRF

Linear-Chain CRF in ASR-(1)

- Linear-Chain CRF + Phonology Attribute + Phone (Eng.)
  - Monophone label

TABLE II
PHONOLOGICAL ATTRIBUTES EXTRACTED

Class	Output Attributes
SONORITY	vowel, obstruent, sonorant, syllabic, silence
VOICE	voiced, unvoiced, n/a
MANNER	fric., stop, closure, flap, nasal, approx., nasalflap, n/a
PLACE	lab., dent., alveolar, pal., vel., glot., lat., rhotic, n/a
HEIGHT	high, mid, low, lowhigh, midhigh, n/a
FRONT	front, back, central, backfront, n/a
ROUND	round, nonround, roundnonround, nonroundround, n/a
TENSE	tense, lax, n/a

Linear-Chain CRF in ASR-(1)

- 特征函数(特征, 转移)

$$f_{/b/,\text{voi}}(\vec{y},\vec{x},t)=\begin{cases}1, & \text{if } y_t=/b/\text{and voiced}(x_t)=\text{true}\\ 0, & \text{otherwise}\end{cases}$$

$$f_{/n/,/ah/\text{nas}}(\vec{y},\vec{x},t)=\begin{cases}1, & \text{if } y_{t-1}=/n/, y_t=/ah/\text{and nasal}(x_t)=\text{true}\\ 0, & \text{otherwise}\end{cases}$$

$$f_{/b/,\text{voi}}(\vec{y},\vec{x},t)=\begin{cases}\text{voiced}(x_t), & \text{if } y_t=/b/\\ 0, & \text{otherwise}\end{cases}$$

Linear-Chain CRF in ASR-(1)

- Feature探测器: MLP
- 其它技巧
  - CRF训练数据中包括reference与hypothesis
  - CRF迭代中进行数据的re-alignment
  - Frame-Level识别, 再坍缩为段级

Model	Feature Type	Number of Inputs	Number of Parameters	Core Accuracy	Enhanced Accuracy
Tandem Phone (16mix)	linear+KLT	61	1.7 million	69.28%	70.21%
Tandem All (16mix)	linear+KLT	105	2.8 million	69.85%	70.19%
CRF Phone	posteriors	61	5280	69.34%	70.40%
CRF All	posteriors	105	7392	69.94%	70.95%
CRF All	post. & linear+KLT	105	7392	70.74%	71.49%

Linear-Chain CRF in ASR-(2)

- Articulatory Features + Finals (Mandarin):
  - Monophone label

Attribute	HMMs
manner (Final)	Finals ended with static vocal tract, Finals ended with a vowel, Finals ended with a nasal
manner (Initial)	affricate, fricative, stop, nasal, lateral
Final onset type	/a/, /e/, /o/, /yi/, /yu/, /wu/, /eh/, N/A.
Final ending type	/ai/, /ei/, /yi/, /ao/, /ou/, /wu/, /an/, /en/, /ang/, /eng/, /yu/, /a/, /o/, /e/, /eh/, N/A
place	bilabial, labial-dental, front coronal, middle coronal, back coronal, dorsum, back (velar), N/A.
aspiration	aspiration, non-aspiration, N/A
voiced	voiced, unvoiced

Linear-Chain CRF in ASR-(2)

- 特征函数: CRF++类型(原子问题)

$$f_{/zh/,\text{bil.}}(\vec{y},\vec{x},t)=\begin{cases}1, & \text{if } y_t=/zh/\text{and bilabio}(x_t)=\text{true}\\ 0, & \text{otherwise}\end{cases}$$

$$f_{/zh/,/zh/,/ang/,\text{bil.}}(\vec{y},\vec{x},t)=\begin{cases}1, & \text{if } y_{t-1}=/zh/, y_t=/zh/, y_{t+1}=/ang/\\ & \text{and bilabio}(x_t)=\text{true}\\ 0, & \text{otherwise}\end{cases}$$

- Feature探测器: Right-CD HMM (Bi-phone)

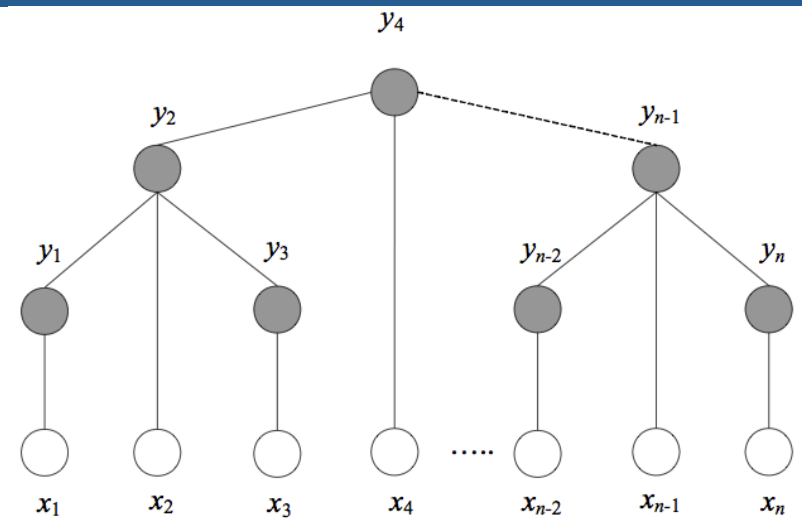
System	Correction (%)	Accuracy (%)
CI HMM	69.61	54.91
RCD HMM	71.84	44.12
HMM/CRF	61.60	58.25

Linear-Chain CRF in ASR-(3)

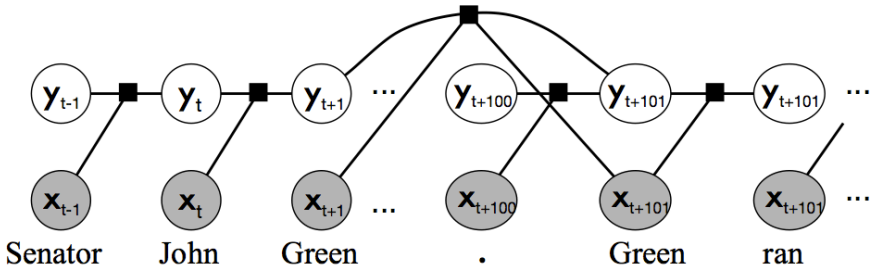
- 使用Linear-Chain CRF训练语言模型: 整合多种特征

复杂图结构CRF

- TCRF
  - Semantic Annotation

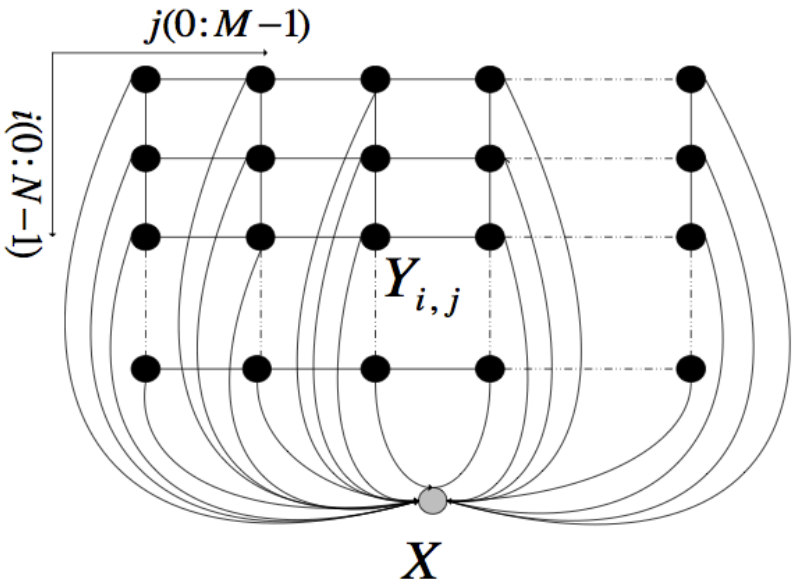


- Skip-Chain CRF
  - Name Entity Recognition



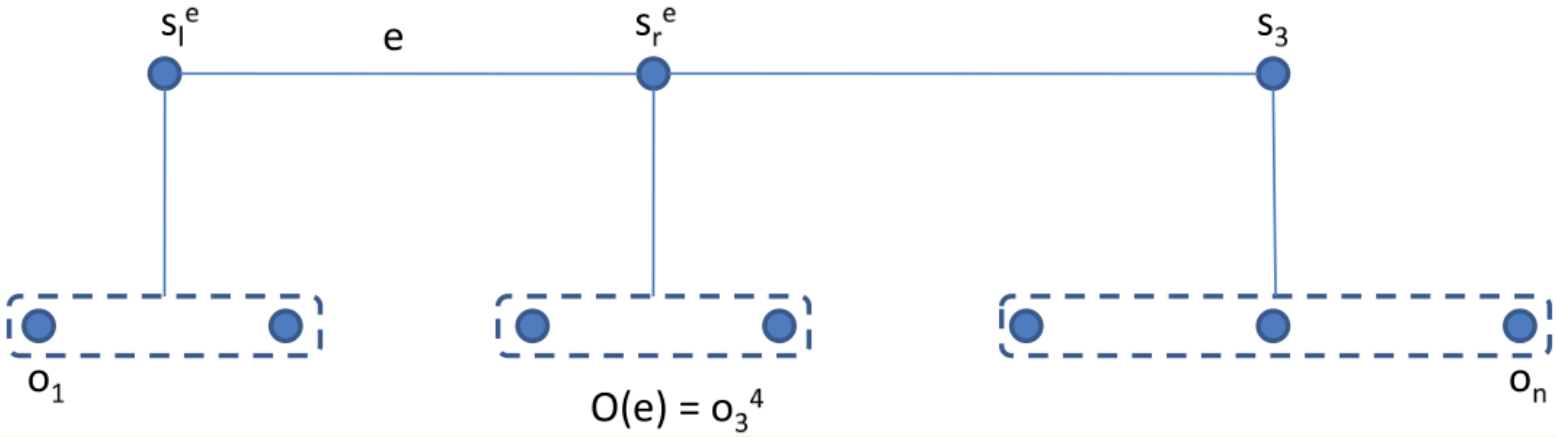
复杂图结构CRF

- 2D CRF
  - Web Information Extraction



Semi-Markov CRF

- 停时定理: 段级
- $p(\vec{y}|\vec{x}) = \frac{\sum_{s=1}^{|S|}\sum_{k=1}^{K}\lambda_k g_k(y_s, y_{s-1}, \vec{x}, st_s, et_s)}{Z(\vec{x})}$
  - S 是输出串, $|S|$ 是串长; st_s和et_s分别是第s个段的起/止时间
  - Name Entity Recognition



Semi-Markov CRF in ASR

- Word, Syllable, Phone, Articulatory, Duration etc.
- Word识别(应用于Bing移动搜索)
- 特征函数:
  - n-gram Existence Feature

$$f_u(y_s, y_{s-1}, o_{st}^{et}) = \delta(w(y_{s-1}) = u)\delta(u \in span(st, et)))$$

  - n-gram Expectation Feature: Hit, False Alarm, False Reject

$$f_u(y_s, y_{s-1}, o_{st}^{et}) = \delta(u \in pron(w(y_{s-1})))\delta(i \in span(st, et)))$$

$$f_u(y_s, y_{s-1}, o_{st}^{et}) = \delta(u \in pron(w(y_{s-1})))\delta(i \notin span(st, et)))$$

$$f_u(y_s, y_{s-1}, o_{st}^{et}) = \delta(u \notin pron(w(y_{s-1})))\delta(i \in span(st, et)))$$

Semi-Markov CRF in ASR

- Levenshtein Feature

f_u^{match} number of times u is matched

f_u^{sub} number of times u (in pronunciation) is substituted

f_u^{del} number of times u is deleted

f_u^{ins} number of times u is inserted

- Language Model Feature

$$f(y_s, y_{s-1}, o_{st}^{et}) = \mathrm{LM}(y_s, y_{s-1})$$

- Baseline Feature

$$f_b(y_s, y_{s-1}, o_{st}^{et}) = \begin{cases} +1 & \text{if } C(st, et) = 1 \text{ and } B(st, et) = w(y_{s-1}) \\ -1 & \text{otherwise.} \end{cases}$$

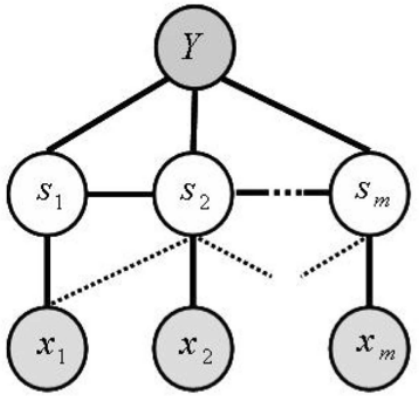
Semi-Markov CRF in ASR

- Feature探测器:
  - HMM (Bi-phone, Monophone)

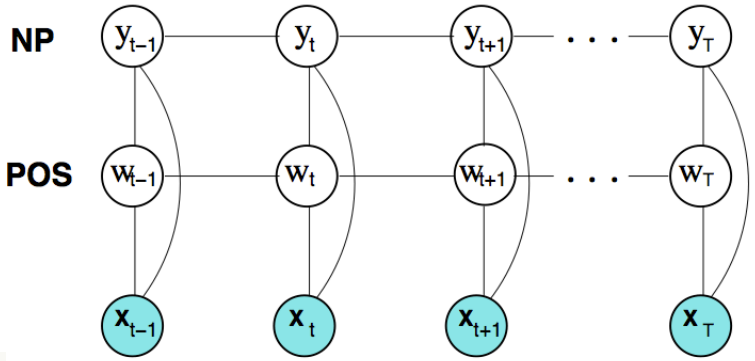
	System	SER
1	HMM (baseline feature)	37.1%%
2	+phone	36.2%
3	+multi-phone	35.7%
4	+phone +multi-phone	35.4%
5	+phone +multi-phone (3-best)	35.2%
6	+phone +multi-phone (3-best) +full LM	35.0%

隐变量CRF

- Hidden CRF
  - Object Recognition
  - Hidden: State

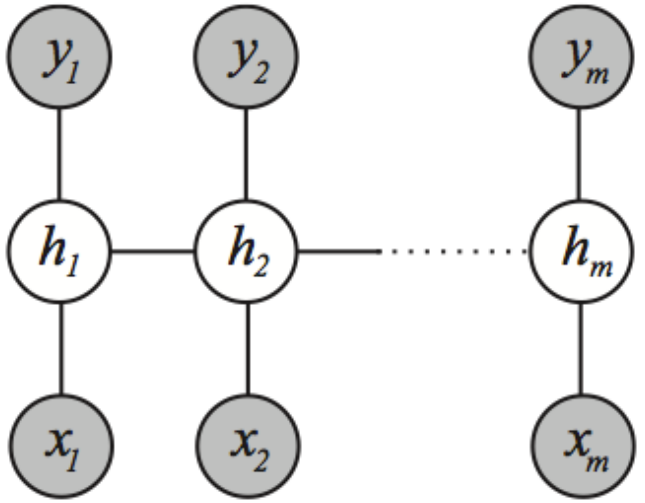


- Factorized CRF
  - Joint Noun Phrase Chunking and Part-of-Speech Tagging
  - Hidden: State(POS)



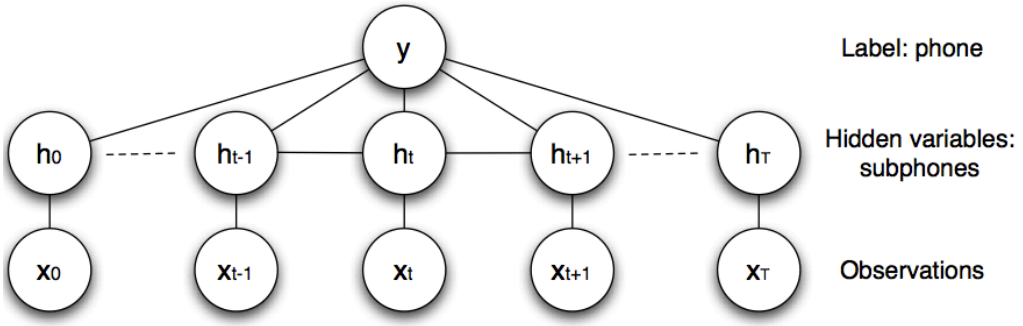
隐变量CRF

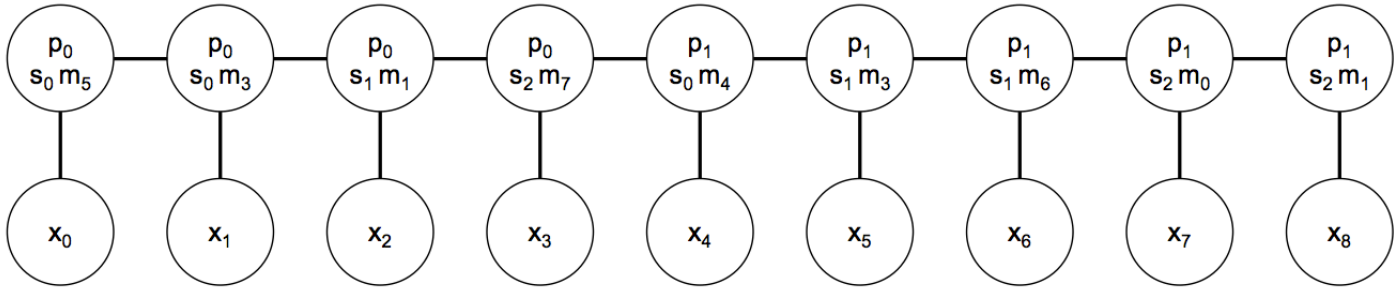
- Latent-Dynamic CRF
  - Shallow Parsing
  - Gesture Recognition
  - Hidden: State
  - Hidden State Number
  - BLP Inference (Linear)



(Gaussian) Hidden CRF in ASR

- HMM (GHMM)?
  - Hidden: GMM, State
  - CRF→GHMM?
  - HCRF→GHMM





(Gaussian) Hidden CRF in ASR

- MFCC or PLP etc. → monophone label
- 特征函数:

$$f_{y'}^{(\mathrm{LM})} = \delta(y = y') \qquad \forall y'$$

$$f_{y'ss'}^{(\mathrm{Tr})} = \sum_{t=1}^{T} \delta(y = y', s_{t-1} = s, s_t = s) \qquad \forall y', s, s'$$

$$f_{sm}^{(\mathrm{Occ})} = \sum_{t=1}^{T} \delta(s_t = s, m_t = m) \qquad \forall s, m$$

$$f_{sm}^{(\mathrm{M1})} = \sum_{t=1}^{T} \delta(s_t = s, m_t = m) x_t \qquad \forall s, m$$

$$f_{sm}^{(\mathrm{M2})} = \sum_{t=1}^{T} \delta(s_t = s, m_t = m) x_t^2 \qquad \forall s, m$$

(Gaussian) Hidden CRF in ASR

- Why named Gaussian HCRF?
- 若权值:

$$\lambda^{\mathrm{LM}}_{y'} = \log u_{y'} \qquad \forall y^{'}$$

$$\lambda^{\mathrm{Tr}}_{y' ss'} = \log a_{y' ss'} \qquad \forall y^{'}, s, s^{'}$$

$$\lambda^{\mathrm{Occ}}_{sm} = -\frac{1}{2}\sum_{d}(\log 2\pi\delta^2_{s,m,d} + \frac{\mu^2_{s,m,d}}{\delta^2_{s,m,d}}) \qquad \forall s, m$$

$$\lambda^{\mathrm{M2}}_{sm} = -\frac{1}{2\delta^2_{s,m,d}} \qquad \forall s, m$$

$$\lambda^{\mathrm{M1}}_{sm} = \frac{\mu_{s,m,d}}{\delta^2_{s,m,d}} \qquad \forall s, m$$

(Gaussian) Hidden CRF in ASR

- 此时GHCRF → Log-Linear GHMM

$$\exp\{\lambda^{\mathrm{LM}}_{y'}\} = u_{y'}$$

$$\exp\{\lambda^{\mathrm{Tr}}_{y's s'}\} = a_{y's s'}$$

$$\exp\{\lambda^{\mathrm{Occ}}_{sm} + \lambda^{\mathrm{M1}}_{sm} x_t + \lambda^{\mathrm{M2}}_{sm} x_t^2\} = \frac{1}{\sqrt{2\pi\delta^2_{s,m,d}}} \exp\{-\frac{1}{2}\frac{(\mu_{s,m,d} - x_{t,s,d})^2}{\delta^2_{s,m,d}}\}$$

- 区别: e.g. 语言模型与转移能够减弱某词或某转移的重要性
- 反向利用: 用已有HMM的对应参数做优化初值

(Gaussian) Hidden CRF in ASR

- Phone Classification

Mix Comp.s	HMM (ML)	HMM (MMI)	HCRF (L-BFGS)	HCRF (SGD)
10	28.1%	24.8%	23.7%	21.8%
20	26.8%	24.6%	23.2%	**21.7%**
40	26.4%	25.3%	23.3%	22.3%

- Phone Recognition

Comps	ML HMMs	MMI HMMs	MPE HMMs	HCRFs
8	35.9%	33.3%	32.1%	29.4%
16	33.5%	32.1%	31.2%	28.7%
32	31.6%	30.8%	30.5%	**28.3%**
64	31.1%	30.9%	31.0%	29.1%

(Gaussian) Hidden CRF in ASR

- 其它应用: MAP, MLLR自适应等

CRF特征选择

- Efficiently Inducing Features
- Relief
- Gaussian-Hidden CRF

CRF实现技巧

- 常用特征类型 → $f_{pk}(\vec{y}_c, \vec{x}_c) = \mathbf{1}_{\vec{y}_c = \vec{y}'_c} q_{pk}(\vec{x}_c)$
- 无内在顺序的离散观测 → 转换为二元特征
- 特征冗余有利于性能提高
- 前向后向算法/Belief Propagation计算得到的概率过小(误差)
  - 将前/后向因子之和归一化
  - 前/后向因子对数化