简单分词模型之HMM&&CRF

本文主要介绍两个常见而又简单的分词模型: HMM(隐马尔科夫模型)和CRF(条件随机场模型)．之前我们介绍的是基于词的n元LM的分词方法,而本文介绍的是由字构词思想的汉语分词模型.

> 由字构词的分词方法认为每个字在构成一个特定的词语时都占据着一个确定的构词位置（词位）, 假如规定每个字只有 4 个词位，词首（B）,词中（M），词尾（E）,单独成词（S）,于是例子
> “你/现在/应该/去/幼儿园/了” 可以表达为：
> 你S现B在E应B该E去S幼B儿M园E了S
> 这样, 只要对一句话的每个字进行分类后, 就能得到分词结果了。

1.基本定义:

1.1 图:

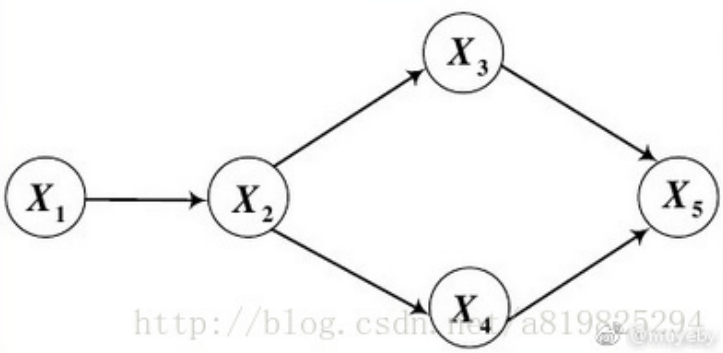
> 图是由结点和连接结点的边组成的集合。结点和边分别记作v和e，结点和边的集合分别记作V和E，图记作G=（V，E）。无向图是指边没有方向的图。

1.2 概率图模型（PGM）

> 概率图模型是一类用图的形式表示随机变量之间条件依赖关系的概率模型，是概率论与图论的结合。根据图中边有无方向，常用的概率图模型分为两类：有向图（贝叶斯网络、信念网络）、无向图（马尔可夫随机场、马尔可夫网络）。

(1) 有向概率图介绍

> 有向图的联合概率: $\boldsymbol{P(X_1, X_2 \dots X_N)} = \prod_{i=1}^{N} \boldsymbol{P(X_i|\pi(X_i))}$
> 其中，$\boldsymbol{\pi(X_i)}$是$\boldsymbol{X_i}$的父节点。



在上图中, 联合概率可以表示为

$$P(X_1, X_2 \dots X_5) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_2)P(X_5|X_3X_4)$$

(2) 无向概率图介绍

> 定义(概率无向图模型）设有联合概率分布$P(Y)$, 由无向图G=(V,E)表示，在图G中，结点v表示随机变量，边e表示随机变量之间的依赖关系。如果联合概率分布$P(Y)$满足*成对、局部或全局马尔可夫性*，就称此联合概率分布为*概率无向图模型*(probability undirected graphical model)，*或马尔可夫随机场*(Markov random field)。

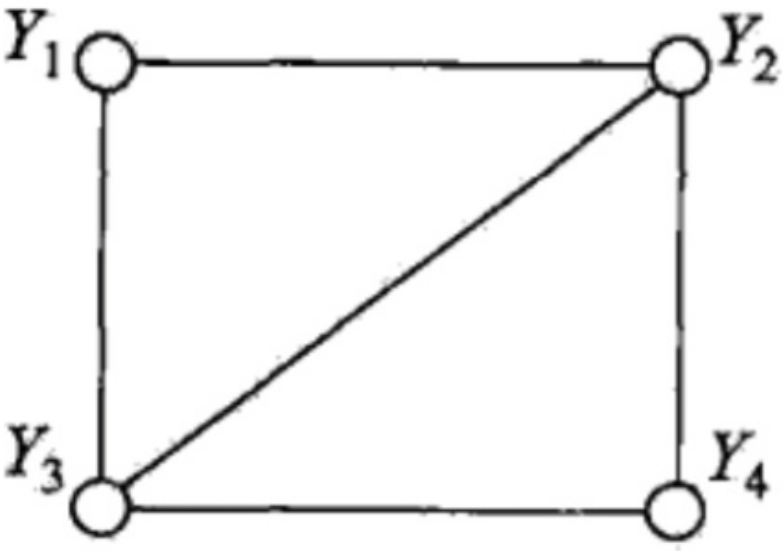
(2).1 概率无向图模型的因子分解

我们首先给出无向图中的团与最大团的定义。

> 定义(团与最大团）无向图G中任何两个结点均有边连接的结点子集称为*团*（clique)。若C是无向图G的一个团，并且不能再加进任何一个G的结点使其成为一个更大的团，则称此C为*最大团*（maximal clique)。

无向图中, 不能用条件概率密度对模型进行参数化, 而是使用一种成为*团势能(clique potentials)*的参数化因子. 所谓团势能又称势函数, 是定义在一个团上的非负实函数, 每个团都对应着一个势函数, 表示团的一个状态.

下图表示由4个结点组成的无向图。图中由2个结点组成的团有5个:{y1,y2},{y2,y3}，{y3,y4}和{y4,y2}，{y1,y3}。有2个最大团：{y1,y2,y3}和{y2,y3,y4}。而{y1,y2,y3,y4}不是一个团，因为y1和y4没有边连接。



将概率无向图模型的联合概率分布表示为其最大团上的随机变量的函数的乘积形式的操作，称为概率无向图模型的*因子分解（factorization)*。

给定概率无向图模型，设其无向图为G，C为G上的最大团，Y_C表示C对应的随机变量。那么概率无向图模型的联合概率分布$P(Y)$可写作图中所有最大团C上的函数$\Psi_c(Y_c)$的乘积形式，即

> $$P(Y) = \frac{1}{Z}\prod_C \Psi_c(Y_c)$$

其中，Z是规范化因子（normalization factor),由式

> $$Z = \sum_Y \prod_C \Psi_c(Y_c)$$

给出。规范化因子保证$P(Y)$构成一个概率分布。$\Psi_c(Y_c)$函数称为势函数(potential function)。这里要求势函数$\Psi_c(Y_c)$是严格正的，通常定义为指数函数：

> $$\Psi_c(Y_c) = exp\{-E(Y_c)\}$$

概率无向图模型的因子分解由下述定理来保证。

> 定理(**Hammersley-Clifford**定理）概率无向图模型的联合概率分布可以表示为如下形式：
>
> > $$P(Y) = \frac{1}{Z}\prod_C \Psi_c(Y_c)$$
> > $$Z = \sum_Y \prod_C \Psi_c(Y_c)$$
>
> 其中，C是无向图的最大团，Y_c是C的结点对应的随机变量，$\Psi_c(Y_c)$是C上定义的严格正函数，乘积是在无向图所有的最大团上进行的。

---

2. HMM模型分词

本节主要介绍HMM(隐马尔科夫模型)的基本原理(模型和解码过程)及其在分词任务上的应用.

2.1 HMM的概率图表示

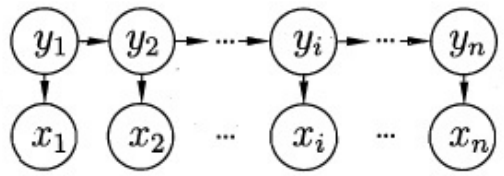


图 14.1 隐马尔可夫模型的图结构

显然上图是有向概率图,箭头表示依赖关系
顶点: $x_1, x_2 \dots x_n$是观测变量, 以分词为例就是指的未切分的句子；$y_1, y_2 \dots y_n$ 是隐变量,这里对应我们的$\{B, E, M, S\}$中的元素

边: 隐变量间的边表示条件转移概率 $P(y_{t+1} = s_j | y_t = s_j)$, 隐变量与观测变量间的边表示发射概率(观测概率)$P(x_t = o_j | y_t = s_j)$

HMM三要素:
一个隐马尔科夫模型是一个三元组（π, A, B） ,其中

> $\pi = p(\pi_i)$, 初始化概率向量；
> $A = (a_{ij})$, 状态转移矩阵；$P(y_{t+1} = s_j | y_t = s_j)$
> $B = (b_{ij})$, 混淆(观测)矩阵；$P(x_t = o_j | y_t = s_j)$

齐次马尔科夫假设**(**一阶马尔科夫假设**):** 假设隐藏的马尔科夫链在任意时刻t的状态只依赖于前一时刻的状态,与其他时刻的状态与观测无关,与时刻t无关
$P(y_t | y_{t-1}, x_{t-1}, \dots y_1, x_1) = P(y_t | y_{t-1})$

观测独立性假设**:** 任意时刻的观测只依赖于该时刻的马尔科夫链的状态,与其他观测及状态无关
$P(x_t | y_T, x_T, y_{T-1}, x_{T-1}, \dots y_{t+1}, x_{t+1}, y_t, y_{t-1}, x_{t-1} \dots y_1, x_1) = P(x_t | y_t)$

2.2 HMM三个问题

- 概率计算问题 :
  已知$\lambda = (A, B, \pi)$和观测序列$X = (x_1, x_2 \dots x_t)$, 求$P(X | \lambda)$
- 学习问题:
  已知观测序列, 估计模型$\lambda = (A, B, \pi)$, 使得在该模型下观测到X的概率$P(X | \lambda)$最大
- 预测问题(解码问题):
  已知$\lambda = (A, B, \pi)$和观测序列$X = (x_1, x_2 \dots x_t)$,求给定观测序列后,使条件概率$P(Y | X)$最大的状态序列$Y = (y_1, y_2 \dots y_t)$

我们这里主要关注解码问题.

2.3 HMM分词

Learning Stage—-easy here

HMM在中文分词任务上的参数可以通过监督(统计频率+平滑)的方法直接估计,得到$\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$

> 还有另一种无监督参数学习方法(Baum-Welch)算法

Decoding Stage

我们期望通过在隐状态解空间内搜索, 找到一个隐状态序列$\boldsymbol{Y} = (\boldsymbol{y_1}, \boldsymbol{y_2} \ldots \boldsymbol{y_t})$, 使得$\boldsymbol{Y}$可以最大化我们的 scoring rule (the probability of possible segmentations y over x) , 形式化表示为:

> $$\boldsymbol{Y} = \arg\max_{\boldsymbol{Y}} \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{\lambda}, \boldsymbol{X})$$

搜索的策略—-Viterbi (后面讲细节)

Scoring Rule

由条件概率公式, 最大化$\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{\lambda}, \boldsymbol{X})$等价于最大化$\boldsymbol{P}(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\lambda})$, 因为对于$\forall \boldsymbol{Y'}$分母都是一样的. 而

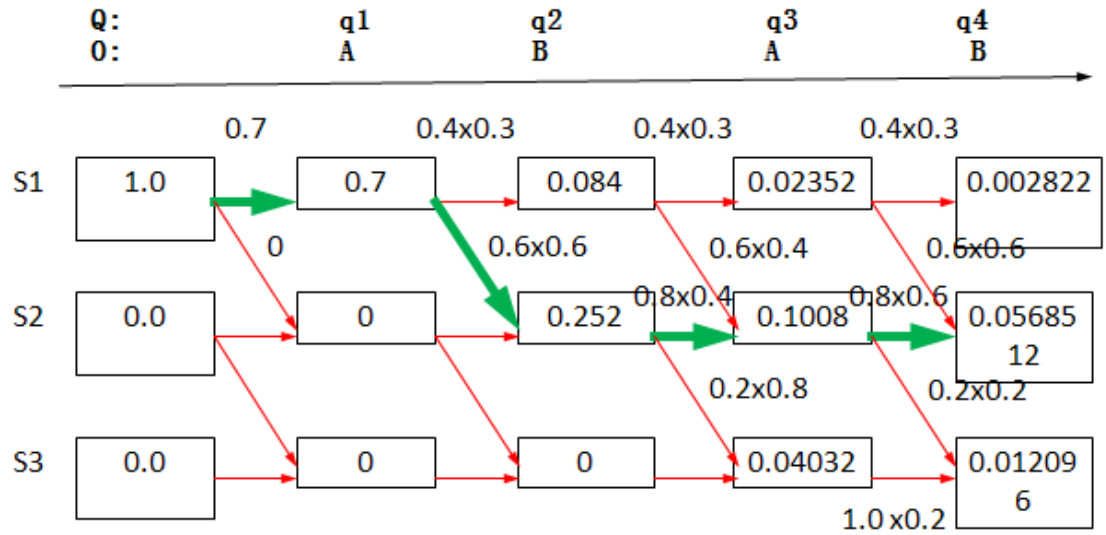
> $$\boldsymbol{P}(\boldsymbol{Y}, \boldsymbol{X}) = \boldsymbol{P}(\boldsymbol{y_1})\boldsymbol{P}(\boldsymbol{x_1}|\boldsymbol{y_1}) \prod_{\boldsymbol{i}=\boldsymbol{2}}^{\boldsymbol{t}} \boldsymbol{P}(\boldsymbol{y_i}|\boldsymbol{y_{i-1}})\boldsymbol{P}(\boldsymbol{x_i}|\boldsymbol{y_i})$$

这里的$\boldsymbol{P}(\boldsymbol{y_1}), \boldsymbol{P}(\boldsymbol{y_i}|\boldsymbol{y_{i-1}}), \boldsymbol{P}(\boldsymbol{x_i}|\boldsymbol{y_i})$就是在学习阶段得到的模型$\boldsymbol{\lambda}$的初始化概率,状态转移概率,观测概率.

> 根据上面的式子可以发现HMM对联合概率建模,是生成式模型

至此, 我们get了如何用隐马模型去评估一个切分结果的好坏, 下面就是根据一个搜索策略去遍历解空间, 找到最优解了.

HMM分词器解码过程



与上面图片不同的是,HMM分词器有四个隐含状态,需要填满一个4×t的二维数组.

3. CRF 分词

本节主要介绍CRF(条件随机场模型)的基本原理(模型和解码过程)及其在分词任务上的应用.

3.1 条件随机场的定义:

定义(条件随机场) 设X与Y是随机变量，$P(Y|X)$是在给定X的条件下Y的条件概率分布。若随机变量Y构成一个由无向图$G=(V,E)$表示的马尔可夫随机场，即

> $$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v)$$

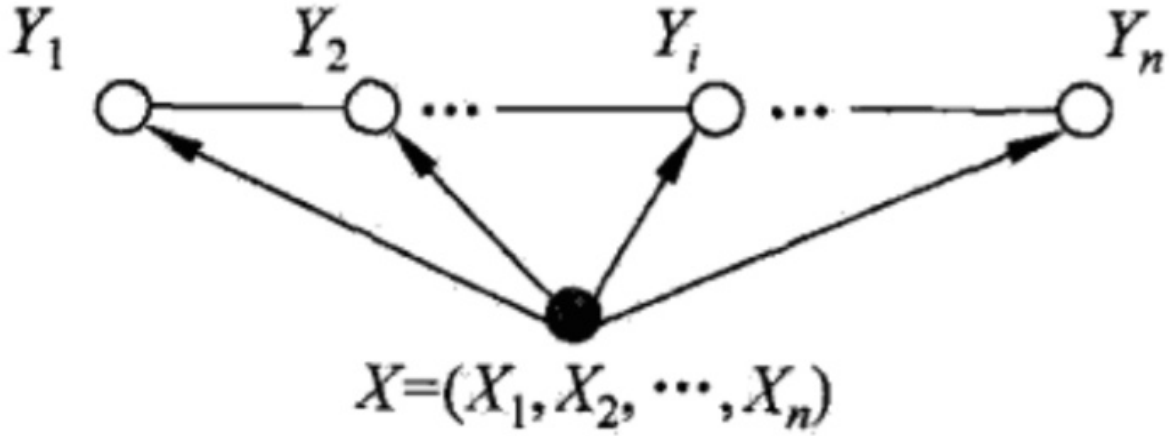
对任意结点v成立，则称条件概率分布$P(Y|X)$为条件随机场。
式中$w\neq v$表示结点v以外的所有结点，$w\sim v$表示在图$G=(V,E)$ 中与结点v有边连接的所有结点w, Y_v与Y_w为结点v,w对应的随机变量。

> 在条件随机场定义中并没有要求X和Y具有相同的结构。在分词任务中，一般假设X和Y有相同的图结构且主要考虑无向图为线性链的情况，即
>
> > $$G=(V=\{1,2,3...n\},E=\{(i,i+1)\}),\text{其中}i=1,2...n-1$$

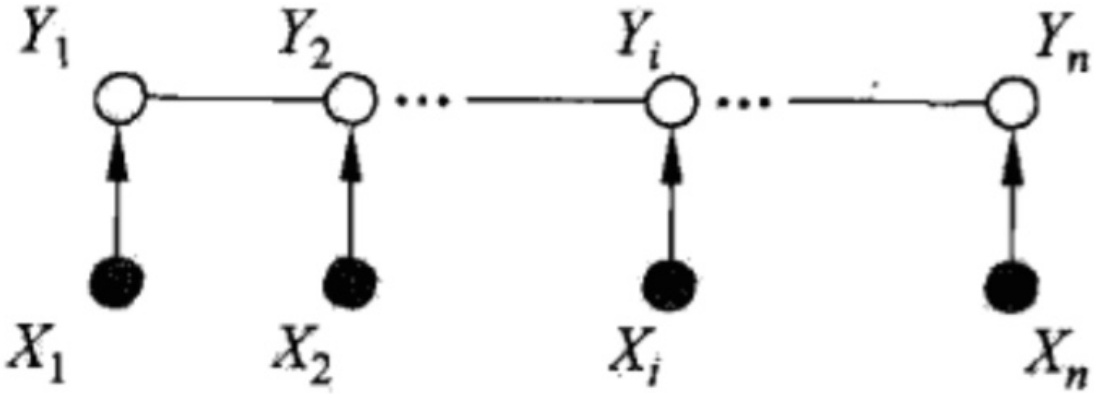
(线性链条件随机场) 设$X=(X_1,X_2...X_n),Y=(Y_1,Y_2...Y_n)$均为线性链表示的随机变量序列，若在给定随机变量序列X的条件下，随机变量序列Y的条件概率分布$P(Y|X)$构成条件随机场，即满足马尔可夫性

> $P(Y_i|X,Y_1,Y_2...Y_{i-1},Y_{i+1}...Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1})$
> $i=1,2,3...n$ (在i=1和n时只考虑单边)

则称$P(Y|X)$为线性链条件随机场。在标注问题中，X表示输入观测序列，Y表示对应的输出标记序列或状态序列.



线性链条件随机场



X和Y有相同的图结构的线性链条件随机场

线性条件随机场的参数化形式

根据Hammersley-Clifford定理,可以给出线性链条件随机场$P(Y|X)$因子分解式，各因子是定义在相邻两个结点(最大团)上的函数。

定理(线性链条件随机场的参数化形式）设$P(Y|X)$为线性链条件随机场，则在随机变量X取值为x的条件下，随机变量Y取值为y的条件概率具有如下形式:

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

其中,

$$Z(x)=\sum_y\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

式中，t_i和s_i是特征函数，λ_k和μ_k是对应的权值。$Z(x)$是规范化因子，求和是在所有可能的输出序列上进行的。(这里面的特征函数有些抽象，并且也不知道为什么有两项而且要加起来，i, k, l分别是什么？这些问题都不用急，下文会讲解。)

> CRF是对条件概率建模,所以是判别式模型

上面两个式子是线性链条件随机场模型的基本形式，表示给定输入序列x,对输出序列y预测的条件概率。

> 其中t_k是定义在边上(*最大团*)的特征函数，称为转移特征（t是transition的缩写，方便记忆），依赖于当前和前一个位置; s_l是定义在结点上的特征函数，称为状态特征（s是status的缩写），依赖于当前位置（无论哪种特征函数，都将当前可能的y_i作为参数）。t_k和s_l都依赖于位置，是局部特征函数。通常，特征函数取值为实数(这里是0,1), 当特征条件满足时取值为1,否则为0。

条件随机场完全由特征函数和对应的权值λ_k、μ_l确定。

举个特征函数的栗子:

> $$f_1(s, i, l_i, l_{i-1}) = 1$$
> 当l_{i-1}是B，l_i是E时，f_1 = 1，其他情况f_1 = 0。w_1也应当是正的，并且w_1越大，说明我们越认为B后面应当跟一个E。
> $$f_2(s, i, l_i, l_{i-1}) = 1$$
> 可以设置为如果l_i和l_{i-1}都是E，那么f_2等于1，其他情况f_2 = 0。这里，我们应当可以想到w_2是负的，并且w_2的绝对值越大，表示我们越不认可E后面还是E的标注序列。
>
> 这里只给出了转移特征函数的例子, 对于状态特征也是同理

线性链条件随机场也是对数线性模型（loglinear model)。

> 分子的形式可以看成是$exp(w_1^T \phi_1(x, y) + w_2^T \phi_2(x, y))$

线性条件随机场的简化形式

为了方便表示,我们给出条件随机场的简化表示形式.
为简便起见，首先将转移特征和状态特征及其权值用统一的符号表示。设有K_1个转移特征，K_2个状态特征，$K = K_1 + K_2$,记

$$f_k(y_{i-1}, y_i, x, i) = \begin{cases} t_k(y_{i-1}, y_i, x, i), & k = 1, 2, \cdots, K_1 \\ s_l(y_i, x, i), & k = K_1 + l;\ l = 1, 2, \cdots, K_2 \end{cases}$$

然后对转移与状态特征在各个位置i求和，记作:

$$f_k(y,x)=\sum_{i=1}^{n} f_k(y_{i-1},y_i,x,i),\quad k=1,2,\cdots,K$$

再用w_k示特征$f_k(y,x)$的权值，即

$$w_k=\begin{cases}\lambda_k, & k=1,2,\cdots,K_1\\ \mu_l, & k=K_1+l;l=1,2,\cdots,K_2\end{cases}$$

于是，条件随机场可表示为

$$P(y\mid x)=\frac{1}{Z(x)}\exp\sum_{k=1}^{K} w_k f_k(y,x)$$

$$Z(x)=\sum_{y}\exp\sum_{k=1}^{K} w_k f_k(y,x)$$

则条件随机场可以写成向量w与$F(y,x)$的内积的形式:

$$P_w(y\mid x)=\frac{\exp\left(w\cdot F(y,x)\right)}{Z_w(x)}$$

其中，

$$Z_w(x)=\sum_{y}\exp\left(w\cdot F(y,x)\right)$$

3.2 CRF的3个问题

- 概率计算问题：
  给定条件随机场$P(Y|X)$，输入序列x和输出序列y，计算条件概率$P(Y=y_i|X), P(Y_{i-1}=y_{i-1},Y_i=y_i|X)$以及相应的数学期望的问题。
- 学习问题:
  估计条件随机场模型参数的问题，即条件随机场的学习问题。条件随机场模型实际上是定义在时序数据上的对数线形模型，主要学习参数w，其学习方法包括极大似然估计和正则化的

极大似然估计。具体的优化实现算法有改进的迭代尺度法IIS、梯度下降法以及拟牛顿法。(具体细节不做展开)

- 预测问题(解码问题):——我们的关注点
  条件随机场的预测问题是给定条件随机场$P(Y|X)$和输入序列（观测序列）x,求条件概率最大的输出序列（标记序列）Y^*，即对观测序列进行标注，条件随机场的预测算法依旧是维特比算法。

3.3 CRF分词

Learning Stage

通过(正则化的)极大似然估计可以求得参数w, 也就知道了条件概率分布$P(Y|X)$

Decoding Stage

寻找Y^*,使得条件概率$P(Y|X)$最大.形式化表示为:

$$
\begin{aligned}
y^* &= \arg\max_y P_w(y \mid x) \\
&= \arg\max_y \frac{\exp(w \cdot F(y,x))}{Z_w(x)} \\
&= \arg\max_y \exp(w \cdot F(y,x)) \\
&= \arg\max_y (w \cdot F(y,x))
\end{aligned}
$$

为了求解最优路径(动态规划)，将$\max_y(w \cdot F(y,x))$写成如下形式：

$$
\max_y \sum_{i=1}^{n} w \cdot F_i(y_{i-1}, y_i, x)
$$

其中,

$$
F_i(y_{i-1}, y_i, x) = (f_1(y_{i-1}, y_i, x, i), f_2(y_{i-1}, y_i, x, i), \cdots, f_K(y_{i-1}, y_i, x, i))^{\mathrm{T}}
$$

由此, 可以由维特比算法解决:

状态转移方程:

$$
\delta_t(y_t = l) = \max_{1 \le j \le m} \{\delta_{t-1}(y_{t-1} = j) + w \cdot F_t(y_{t-1} = j, y_t = l, x)\}
$$

至此,CRF解码问题的简要介绍结束.

4 关于CRF的一点思考

4.1 CRF与逻辑回归

仔细观察CRF的条件概率:
$$P(Y|X) = \frac{exp(score(Y|X))}{\sum_{Y'} exp(score(Y'|X))} = \frac{exp(w \cdot F(Y,X))}{\sum_{Y'} exp(w \cdot F(Y',X))}$$
是不是有点逻辑回归的味道?

> **note:**
> 事实上，条件随机场是逻辑回归的序列化版本。逻辑回归是用于分类的对数线性模型，条件随机场是用于序列化标注的对数线性模型。

4.2 CRF与HMM的比较

对于分词问题, 之前介绍的HMM也能解决, HMM的思路是用生成办法，就是说，在已知要标注的句子X的情况下，去判断生成标注序列Y的概率，如下所示：
$$P(Y,X) = P(X_1)\prod_i P(Y_i|Y_{i-1})P(X_i|Y_i)$$

> 那么HMM与CRF相比如何呢?

我们从两个角度来分析这个问题:

1 . 数学角度

> 我们对上面的式子取对数, 可以得到
> $$logP(Y,X) = logP(X_1) + \sum_i logP(Y_i|Y_{i-1}) + \sum_i logP(X_i|Y_i)$$
> 我们把这个式子与CRF的式子进行比较：
> $$logP(Y|X) = \sum_{i,k}^{n} \lambda_k \cdot t_k(y_{i-1}, y_i, i, x) + \sum_{i,l}^{n} \mu_l \cdot s_l(y_i, i, x)$$
> 可以发现，如果我们把第一个HMM式子中的log形式的概率看做是第二个CRF式子中的特征函数的权重λ, μ的话，我们会发现，CRF和HMM具有相同的形式, 可以这样构造CRF的特征函数
> 对于HMM中的每一个转移概率$p(Y_i = y_i|Y_{i-1} = y_{i-1})$,我们可以定义这样的一个特征函数：
> $$t_k(X, Y_i, Y_i - 1, i) = 1$$
> 该特征函数仅当$Y_i = y_i, Y_{i-1} = y_{i-1}$时才等于1。这个特征函数的权重如下：
> $$\lambda_k = log(Y_i = y_i|Y_{i-1} = y_{i-1})$$
> 同理, 对于HMM中的发射概率，我们也都可以定义相应的特征函数s_l，并让该特征函数的权

> 重μ等于HMM中的log形式的发射概率。
> 可以发现, 用这些形式的特征函数和相应的权重计算出来的$p(Y|X)$和对数形式的HMM模型几乎是一样的！
> 所以:
> 每一个**HMM**模型都等价于某个**CRF**
> 每一个**HMM**模型都等价于某个**CRF**
> 每一个**HMM**模型都等价于某个**CRF**

2 . 图模型角度

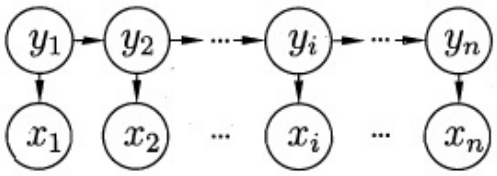
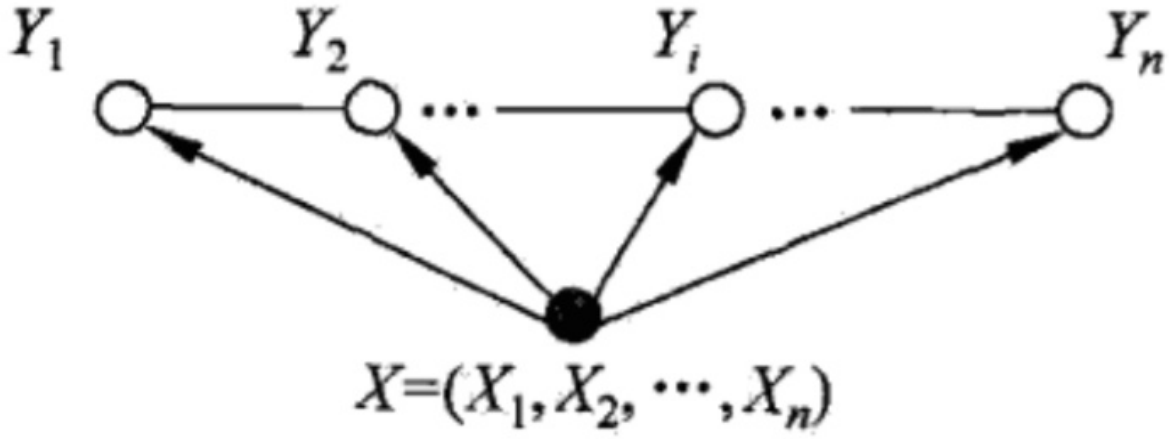
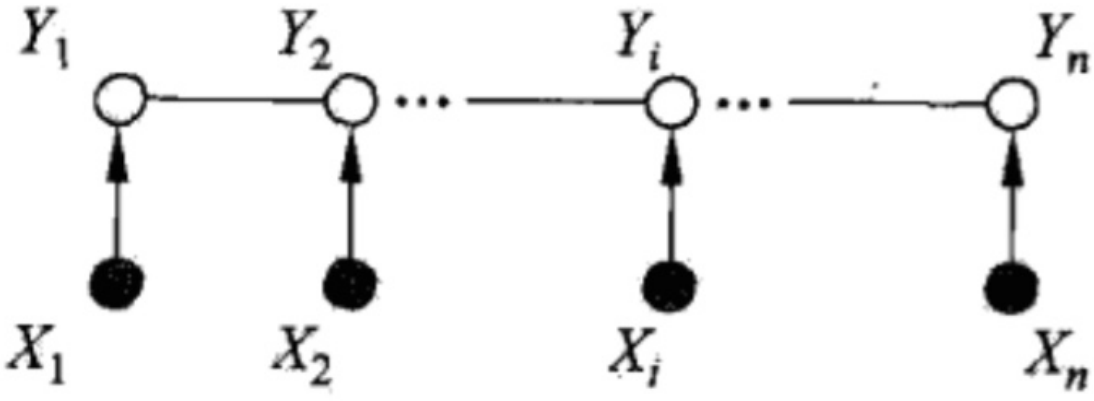


图 14.1 隐马尔可夫模型的图结构



线性链条件随机场



X和Y有相同的图结构的线性链条件随机场

> 从上图可以看出
> CRF 是无向图，每个节点既依赖于左邻节点，也依赖于右邻节点，而HMM是有向图，只考虑左侧节点的信息，因此HMM不如CRF.

4.3 本节的图模型与LM分词的比较

- 基于词的 n 元语法模型对于对于集内词（词典词）的处理可以获得比较好的性能表现，而对集外词（未登录词）的分词效果欠佳.
- 基于字（HMM,CRF）模型正好相反，它对集外词的处理拥有较好的鲁棒性，对集内词的处理却不是非常理想.

具体原因解释参见：
<<统计自然语言处理>> chap 7.2

5. 当下深度学习中文分词技术

- Pei 2014的模型与Zheng 2013 — Window Based NN分类器

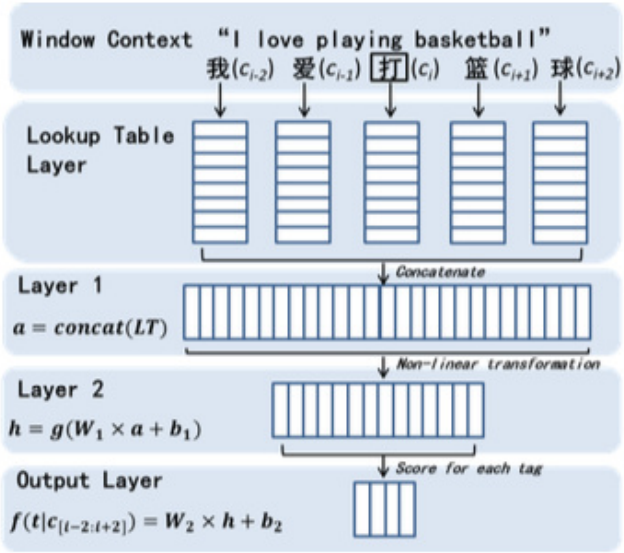


Figure 1: Conventional Neural Network

结果：

Model	PKU	MSRA
Best05(Chen et al., 2005)	95.0	96.0
Best05(Tseng et al., 2005)	95.0	96.4
(Zhang et al., 2006)	95.1	97.1
(Zhang and Clark, 2007)	94.5	97.2
(Sun et al., 2009)	95.2	97.3
(Sun et al., 2012)	95.4	97.4
(Zhang et al., 2013)	96.1	97.4
MMTNN	94.0	94.9
MMTNN + bigram	95.2	97.2

Table 6: Comparison with state-of-the-art systems

通过加入bigram特征终于做到了和传统特征工程comparable
2013年之前的都是非NN传统方法。截止2014年，NN方法没有取得显著成绩。

- sequence模型

> Chen等人2015年： GRU/LSTM ＋ bigram

Models	PKU	MSRA	CTB6
(Tseng et al., 2005)	95.0	96.4	-
(Zhang and Clark, 2007)	95.1	97.2	-
(Sun and Xu, 2011)	-	-	95.7
(Zhang et al., 2013)	96.1	97.4	-
This work	**96.5**	**97.4**	**96.0**

Table 6: Comparison of our model with state-of-the-art methods on three test sets.

- Multi-Criteria Joint Learning— 引入GAN + bigram

Models		MSRA	AS	PKU	CTB	CKIP	CITYU	NCC	SXU	Avg.
LSTM	P	95.13	93.66	93.96	95.36	91.85	94.01	91.45	95.02	93.81
	R	95.55	94.71	92.65	85.52	93.34	94.00	92.22	95.05	92.88
	F	95.34	94.18	93.30	**95.44**	92.59	94.00	91.83	95.04	93.97
	OOV	63.60	69.83	66.34	76.34	68.67	65.48	56.28	69.46	67.00
Bi-LSTM	P	95.70	93.64	93.67	95.19	92.44	94.00	91.86	95.11	93.95
	R	95.99	94.77	92.93	95.42	93.69	94.15	92.47	95.23	94.33
	F	**95.84**	94.20	93.30	95.30	**93.06**	**94.07**	92.17	95.17	**94.14**
	OOV	66.28	70.07	66.09	76.47	72.12	65.79	59.11	71.27	68.40
Stacked Bi-LSTM	P	95.69	93.89	94.10	95.20	92.40	94.13	91.81	94.99	94.03
	R	95.81	94.54	92.66	95.40	93.39	93.99	92.62	95.37	94.22
	F	95.75	**94.22**	**93.37**	95.30	92.89	94.06	**92.21**	**95.18**	94.12
	OOV	65.55	71.50	67.92	75.44	70.50	66.35	57.39	69.69	68.04
Multi-Criteria Learning										
Model-I	P	95.67	94.44	94.93	95.95	93.99	95.10	92.54	96.07	94.84
	R	95.82	95.09	93.73	96.00	94.52	95.60	92.69	96.08	94.94
	F	95.74	94.76	**94.33**	95.97	**94.26**	95.35	**92.61**	96.07	**94.89**
	OOV	69.89	74.13	72.96	81.12	77.58	80.00	64.14	77.05	74.61
Model-II	P	95.74	94.60	94.82	95.90	93.51	95.30	92.26	96.17	94.79
	R	95.74	95.20	93.76	95.94	94.56	95.50	92.84	95.95	94.94
	F	95.74	**94.90**	94.28	95.92	94.03	95.40	92.55	96.06	94.86
	OOV	69.67	74.87	72.28	79.94	76.67	81.05	61.51	77.96	74.24
Model-III	P	95.76	93.99	94.95	95.85	93.50	95.56	92.17	96.10	94.74
	R	95.89	95.07	93.48	96.11	94.58	95.62	92.96	96.13	94.98
	F	**95.82**	94.53	94.21	**95.98**	94.04	**95.59**	92.57	**96.12**	94.86
	OOV	70.72	72.59	73.12	81.21	76.56	82.14	60.83	77.56	74.34
Adversarial Multi-Criteria Learning										
Model-I+ADV	P	95.95	94.17	94.86	96.02	93.82	95.39	92.46	96.07	94.84
	R	96.14	95.11	93.78	96.33	94.70	95.70	93.19	96.01	95.12
	F	**96.04**	94.64	**94.32**	**96.18**	**94.26**	**95.55**	**92.83**	96.04	**94.98**
	OOV	71.60	73.50	72.67	82.48	77.59	81.40	63.31	77.10	74.96
Model-II+ADV	P	96.02	94.52	94.65	96.09	93.80	95.37	92.42	95.85	94.84
	R	95.86	94.98	93.61	95.90	94.69	95.63	93.20	96.07	94.99
	F	95.94	**94.75**	94.13	96.00	94.24	95.50	92.81	95.96	94.92
	OOV	72.76	75.37	73.13	82.19	77.71	81.05	62.16	76.88	75.16
Model-III+ADV	P	95.92	94.25	94.68	95.86	93.67	95.24	92.47	96.24	94.79
	R	95.83	95.11	93.82	96.10	94.48	95.60	92.73	96.04	94.96
	F	95.87	94.68	94.25	95.98	94.07	95.42	92.60	**96.14**	94.88
	OOV	70.86	72.89	72.20	81.65	76.13	80.71	63.22	77.88	74.44

6. 一点扩展

- 文中介绍的是HMM是基于一阶马尔科夫假设的, 扩展到二阶,三阶?

> 参照: TnT – A Statistical Part-of-Speech Tagger

- 关于HMM和CRF的学习算法

> 各种书籍，资源很多

- 最近几年的分词技术

> [1] Max-Margin Tensor Neural Network for Chinese Word Segmentation
> [2] Long Short-Term Memory Neural Networks for Chinese Word Segmentation
> [3] Neural Word Segmentation Learning for Chinese
> [4] Adversarial Multi-Criteria Learning for Chinese Word Segmentation
> [5] Fast and Accurate Neural Word Segmentation for Chinese