

人工智能之NLP

CRF条件随机场

主讲人: GerryLiu

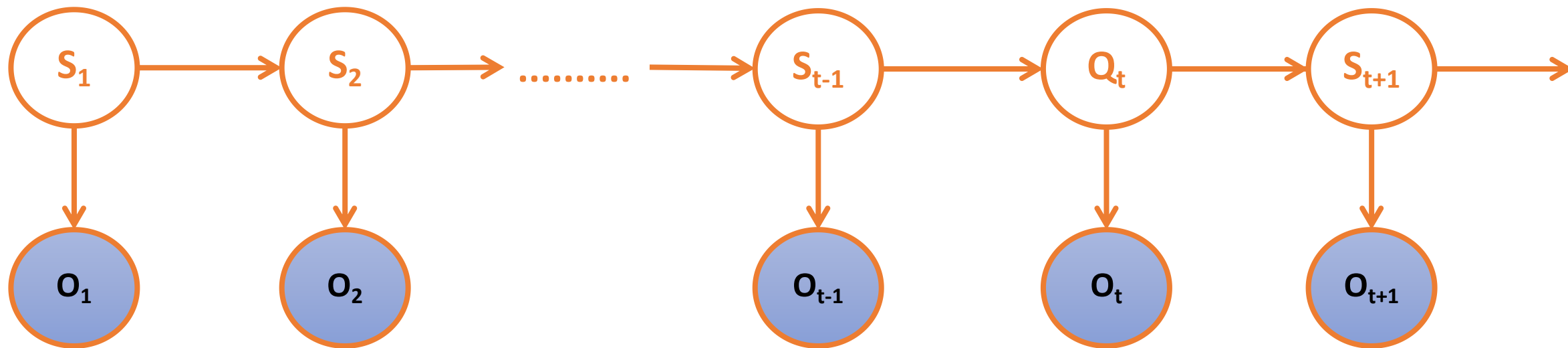
课程要求

- 课上课下“九字”真言
 - 认真听，善摘录，勤思考
 - 多温故，乐实践，再发散
- 四不原则
 - 不懒散惰性，不迟到早退
 - 不请假旷课，不拖延作业
- 一点注意事项
 - 违反“四不原则”，不推荐就业

课程内容

- HMM回顾
- CRF原理讲解
- CRF应用场景理解

HMM回顾



- HMM即隐马尔可夫模型，它是处理序列问题的统计学模型，描述的过程为：**由隐马尔科夫链随机生成不可观测的状态随机序列，然后各个状态分别生成一个观测，从而产生观测随机序列。**
- 在这个过程中，不可观测的序列称为状态序列(state sequence), 由此产生的序列称为观测序列(observation sequence)。

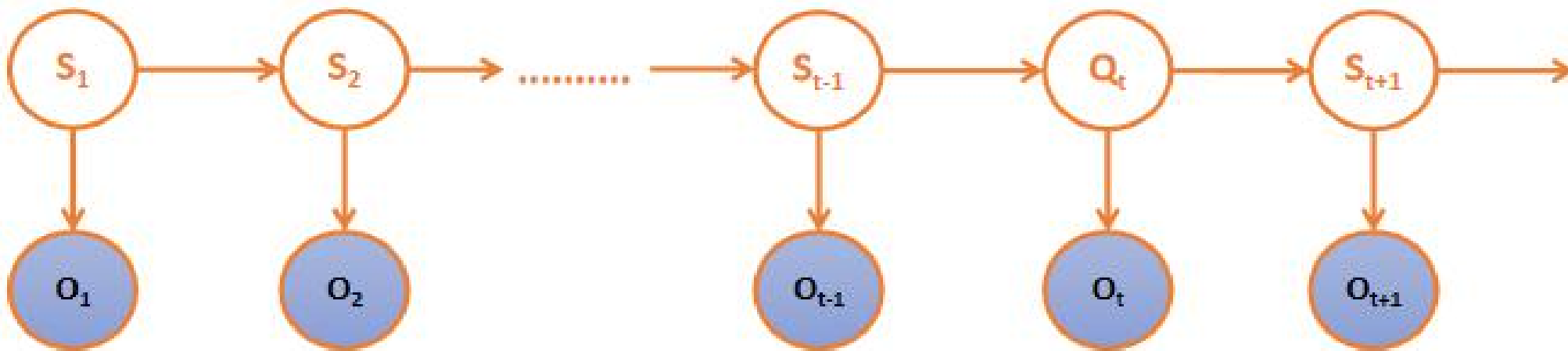
HMM回顾

- 隐马尔可夫模型 λ 可以由一下三个模型参数组成：
 - 初始概率分布：即初始的隐含状态概率分布，记为 π ；
 - 状态转移概率分布：即隐含状态之间的转移概率分布，记为 A ；
 - 观测概率分布：即由隐含状态生成观测状态的概率分布，记为 B 。

$$\lambda = (A, B, \pi)$$

HMM回顾

- 隐马尔可夫的三个基本问题：
 - 概率计算问题
 - 模型学习问题
 - 解码问题



HMM回顾

- Viterbi算法：用动态规划的思路求解HMM预测问题，求出概率最大的“路径”，每条“路径”对应一个状态序列。

$$\delta_t(i) = \max_{s_1, s_2, \dots, s_{t-1}} p(s_t = i, s_1, s_2, \dots, s_{t-1}, o_t, o_{t-1}, \dots, o_1; \lambda)$$

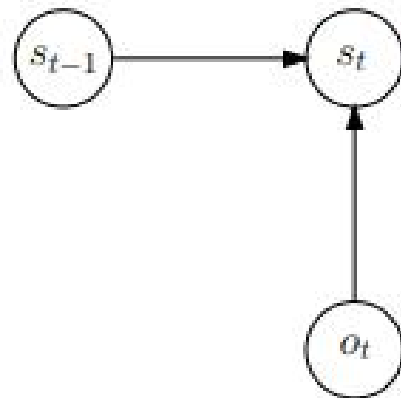
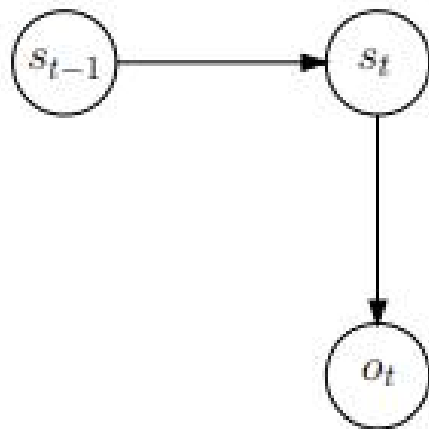
$$\delta_t(i) = \max_{1 \leq j \leq n} (\delta_{t-1}(j) a_{ji}) b_{io_t}$$

HMM回顾

- HMM假设**当前状态仅和前一个状态有关**，也就是具有一阶马尔可夫性质，但是在很多场景中，模型不仅仅需要考虑前一个状态的信息，也可能需要后一个状态的信息，因此需要模型提出更多的假设条件，也就是引入**图模型(当前状态和相连状态都有关)+条件模型(当前状态和观测值有关)=条件随机场**。比如：“我爱中国”，结构更加复杂，信息更加丰富。

MEMM

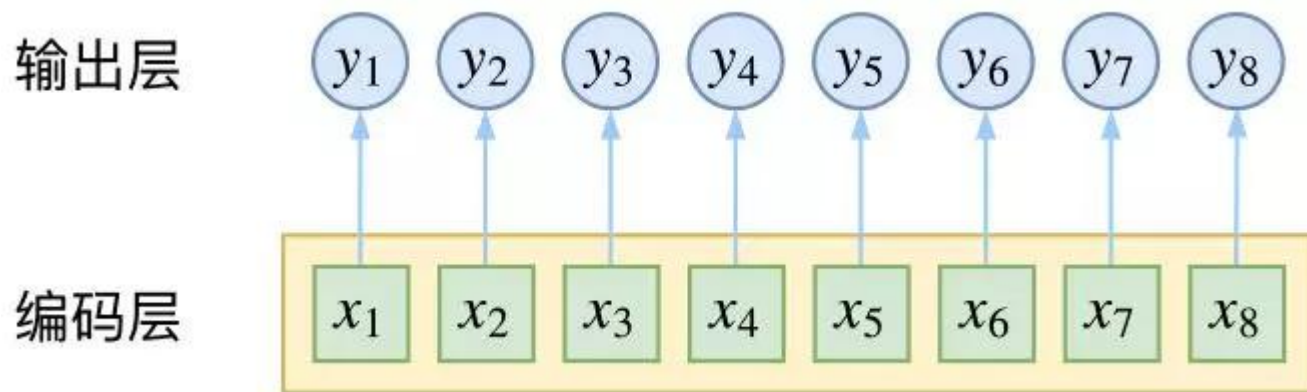
- Maximum Entropy Markov Model(MEMM, 最大熵马尔可夫模型)
 - 和HMM的区别:
 - 不通过联合概率建模, 而是直接使用条件概率。
 - HMM中当前时刻的状态值仅和上一个时刻的状态值有关, MEMM中当前时刻的状态值和上一个时刻的状态值以及当前时刻的观测值有关。
 - HMM中是状态影响观测值, 而在MEMM中是观测值影响状态值。



- 现在有小明同学一天内不同时段的照片，从小明起床到睡觉各个时间段都有。现在的任务是对这些照片进行分类。比如有的照片是吃饭，那就给它打上吃饭的标签；有的照片是跑步时拍的，那就打上跑步的标签；有的照片是开会时拍的，那就打上开会的标签。问题来了，如何进行标签给定呢？

CRF

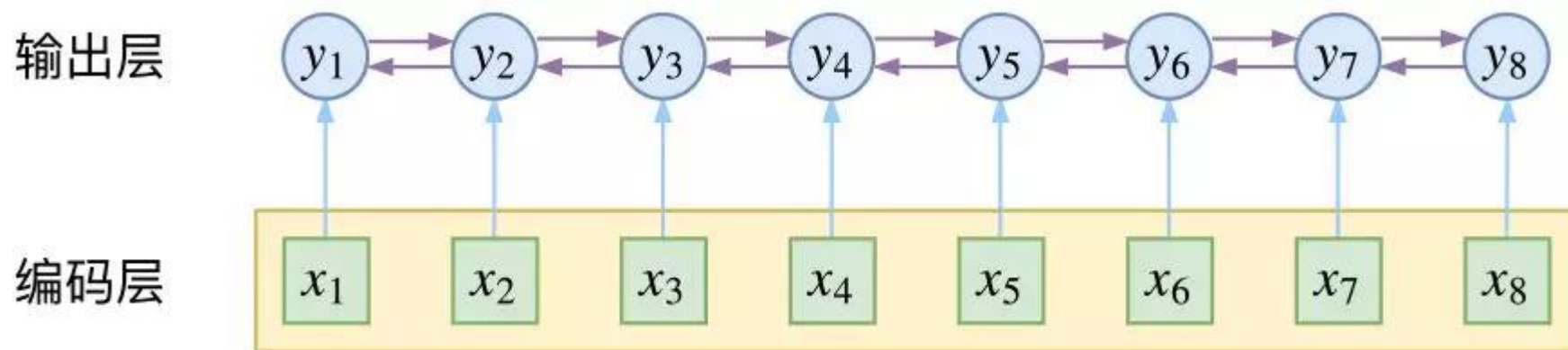
- 方式一：一个简单直观的办法就是，不管这些照片之间的时间顺序，想办法训练出一个多元分类器。就是用一些打好标签的照片作为训练数据，训练出一个模型，直接根据照片的特征来分类。例如，如果照片是早上6:00拍的，且画面是黑暗的，那就给它打上睡觉的标签;如果照片上有车，那就给它打上开车的标签。



- 问题：由于我们忽略了这些照片之间的时间顺序这一重要信息，我们的分类器会有缺陷的。举个例子，假如有一张小明闭着嘴的照片，怎么分类？显然难以直接判断，需要参考闭嘴之前的照片，如果之前的照片显示小明在吃饭，那这个闭嘴的照片很可能是小明在咀嚼食物准备下咽，可以给它打上吃饭的标签；如果之前的照片显示小明在唱歌，那这个闭嘴的照片很可能是小明唱歌瞬间的抓拍，可以给它打上唱歌的标签。

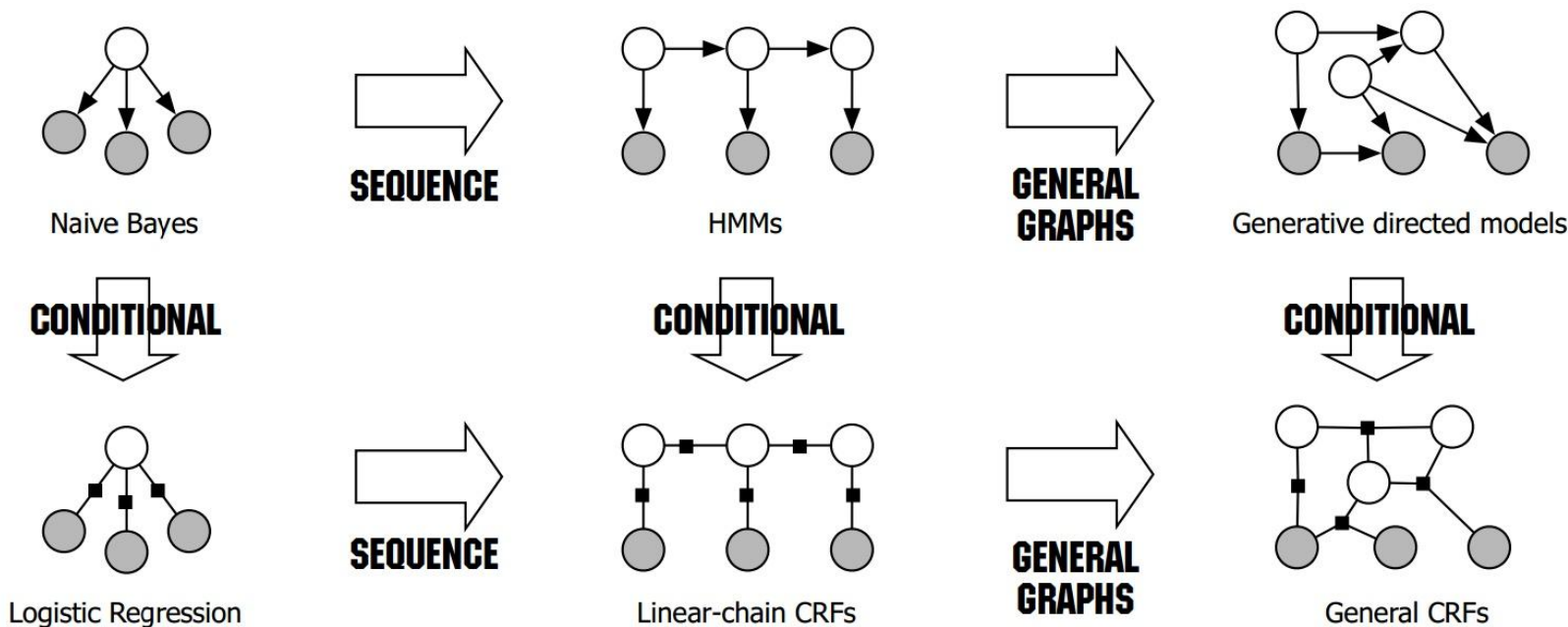
CRF

- 方式二：为了让我们的分类器能够有更好的表现，在为一张照片分类时，我们必须将与它相邻的照片的标签信息考虑进来。



CRF

- 条件随机场(Conditional Random Fields, 简称CRF)给定一组输入序列条件下另一组输出序列的条件概率分布模型，在自然语言处理中得到了广泛应用。最常见的形式为：**线性链(Linear Chain) CRF**。

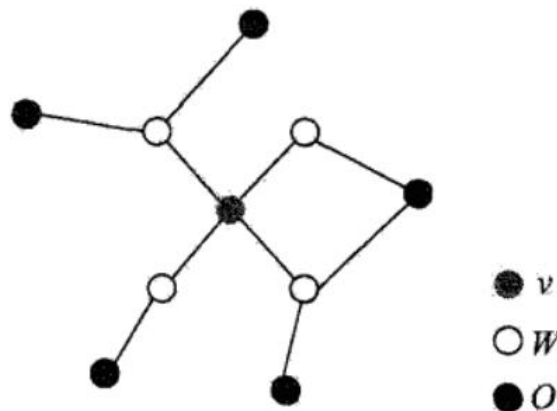


随机场

- 随机场是由若干个位置组成的整体，当给每一个位置中按照某种分布随机赋予一个值之后，其全体就叫做随机场。假如我们有一个十个词形成的句子需要做词性标注。这十个词每个词的词性可以在我们已知的词性集合（名词，动词...）中去选择。当我们为每个词选择完词性后，这就形成了一个随机场。

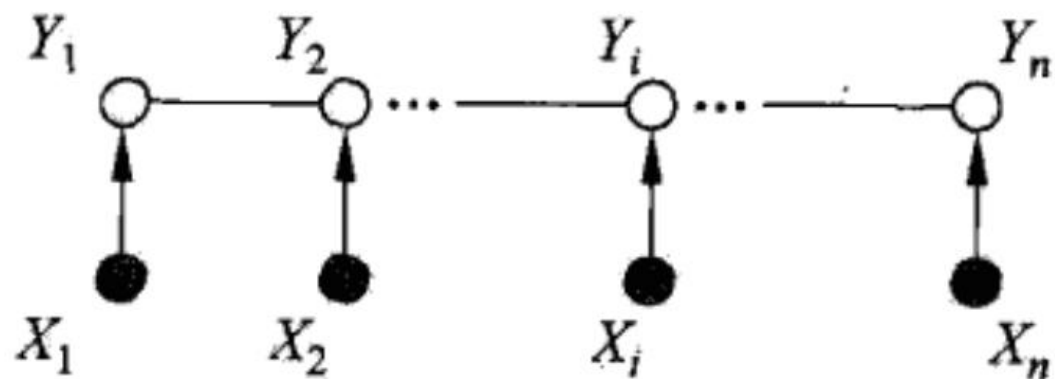
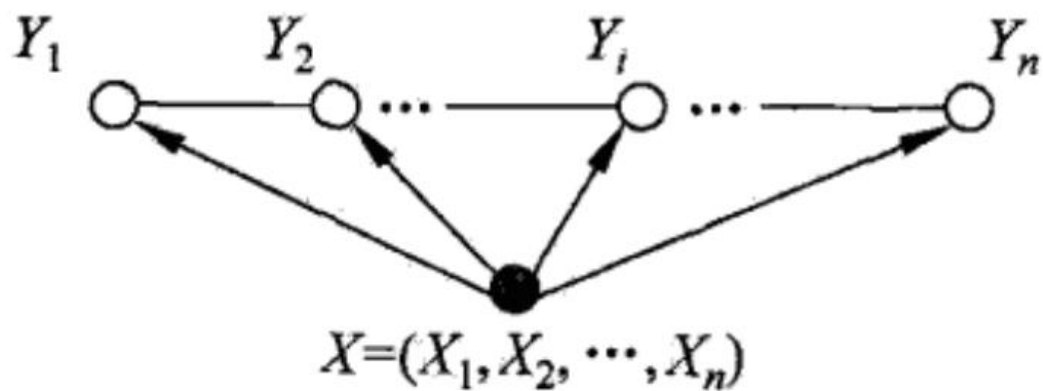
MRF

- 马尔可夫随机场(Markov Random Field)
 - 又称为概率无向图模型，是一个由无向图表示的联合概率分布。
 - 马尔可夫随机场是随机场的特例，假定随机场中某个位置的赋值仅仅和相邻的位置赋值有关，和与其不相邻的位置的赋值无关。
 - 比如假设所有词的词性和它相邻词的词性有关时，随机场就转换为马尔可夫随机场。比如第三个词的词性除了和自己本身的位置有关外，只与第二个词和第四个词有关。



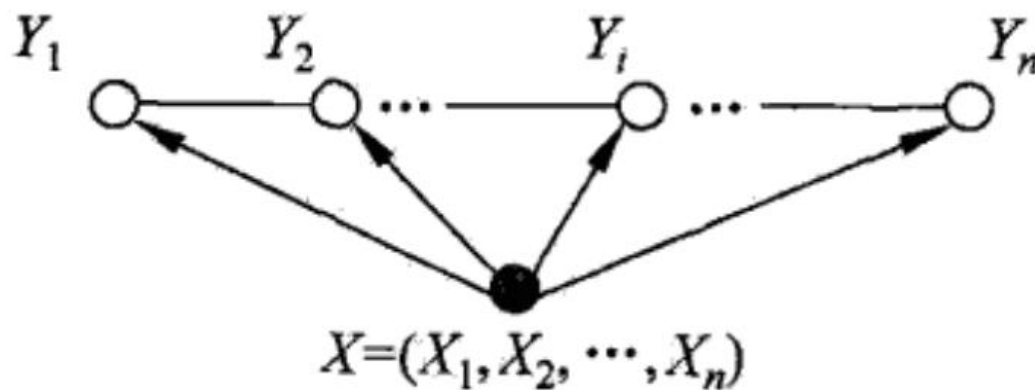
CRF

- 条件随机场(Conditional Random Fields, 简称CRF)是MRF的特例，在CRF中，假设马尔可夫随机场中只有 X 和 Y 两个变量， X 一般是给定的，而 Y 是当 X 给定的条件下模型的输出。



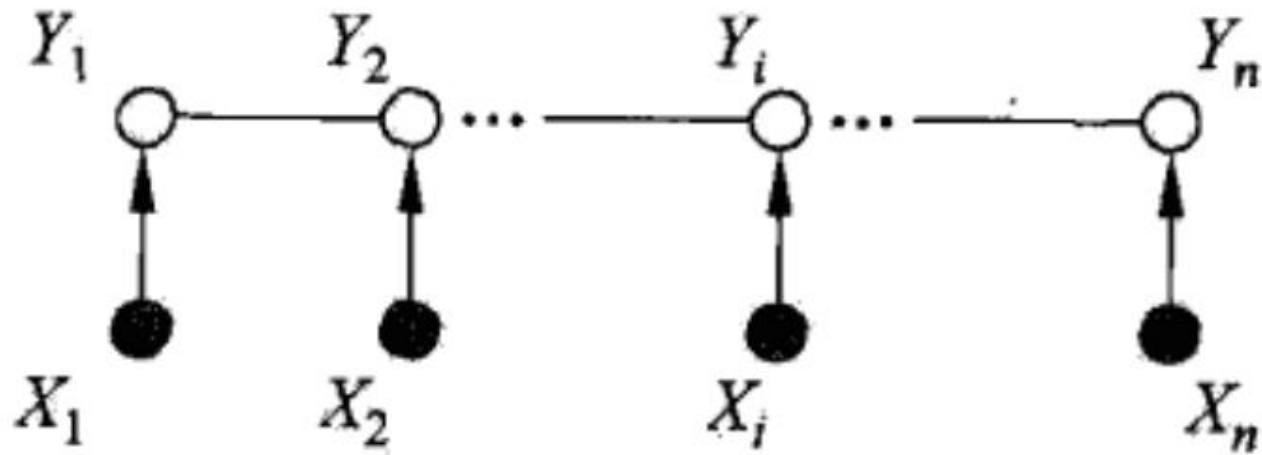
CRF

- 设 $X=(X_1, X_2, \dots, X_n)$ 和 $Y=(Y_1, Y_2, \dots, Y_n)$ 均为线性链表示的随机变量序列，在给定随机变量序列 X 的情况下，随机变量 Y 的条件概率分布 $P(Y|X)$ 就是条件随机场，即满足**马尔可夫性质**。



$$P(Y_i | X, Y_1, Y_2, \dots, Y_n) = P(Y_i | X, Y_{i-1}, Y_{i+1})$$

CRF



$$P(Y_i | X, Y_1, Y_2, \dots, Y_n) = P(Y_i | X_i, Y_{i-1}, Y_{i+1})$$

CRF

- 在Linear CRF中，特征函数分为两类；第一类是定义在Y节点上的节点特征函数，这个特征函数只和当前节点有关。

$$s_l(y_i, x, i), l = 1, 2, \dots, L$$

i 是当前节点在序列中的位置

L 是节点特征函数总数

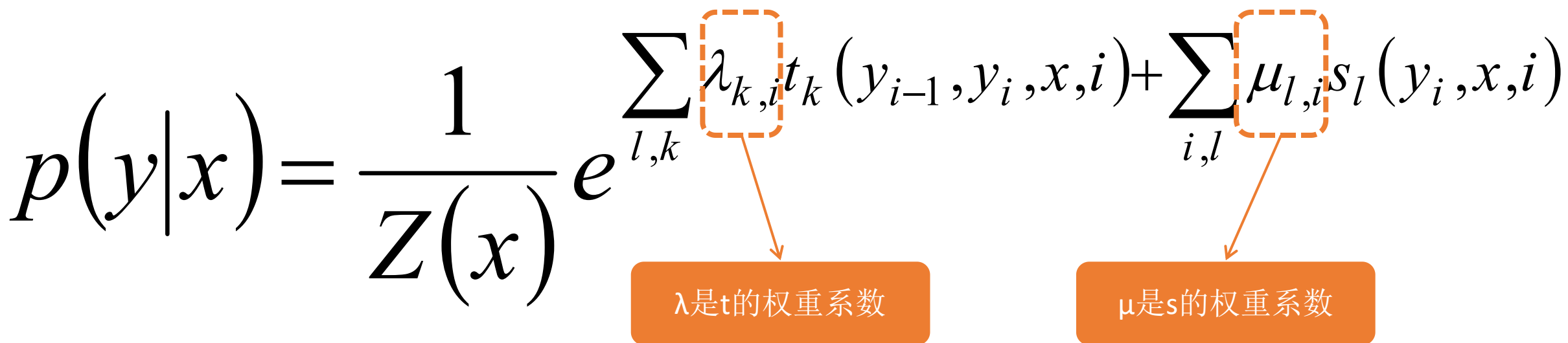
- 第二类是定义在Y上下文的局部特征函数，这类特征函数只和当前节点+上一个节点有关。

$$t_k(y_{i-1}, y_i, x, i), k = 1, 2, \dots, K$$

K 是局部特征函数总个数

CRF

- 无论是节点特征函数还是局部特征函数，它们的取值只能是0或者1。即满足特征条件或者不满足特征条件。同时，我们可以为每个特征函数赋予一个权值，用以表达我们对这个特征函数的信任度。从而得到Linear CRF的参数化形式如下：

$$p(y|x) = \frac{1}{Z(x)} e^{\sum_{l,k} \lambda_{k,i} t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_{l,i} s_l(y_i, x, i)}$$


λ是t的权重系数

μ是s的权重系数

$$p(y|x) = \frac{1}{Z(x)} e^{\sum_{i,k} \lambda_{k,i} t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_{l,i} s_l(y_i, x, i)}$$

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

$$s_l(y_i, x, i), l = 1, 2, \dots, L$$

- 假定Linear CRF在词性标注中的案例，简化词性类别: {1: 名词, 2: 动词}; 假定输入三个词的语句 $X=(x_1, x_2, x_3)$ ，输出标记记为 $Y=(y_1, y_2, y_3)$ ；假定通过模型训练最终得到如下内容：

t, k, λ	$i=2$	$i=3$
$y_{i-1}, y_i = 1, 1$	1, 0.6	0
$y_{i-1}, y_i = 1, 2$	1, 1.0	1, 1.0
$y_{i-1}, y_i = 2, 1$	1, 1.0	1, 1.0
$y_{i-1}, y_i = 2, 2$	0	1, 0.2

$$t_1(y_1 = 1, y_2 = 1, x, 2) = 1; \lambda_{1,2} = 0.6$$

$$t_2(y_{i-1} = 1, y_i = 2, x, i) = 1; \lambda_{2,2} = 1.0; \lambda_{2,3} = 1.0$$

$$t_3(y_{i-1} = 2, y_i = 1, x, i) = 1; \lambda_{3,2} = 1.0; \lambda_{3,3} = 1.0$$

$$t_4(y_2 = 2, y_3 = 2, x, 3) = 1; \lambda_{4,3} = 0.2$$

s, l, μ	$i=1$	$i=2$	$i=3$
$y_i = 1$	1, 1.0	1, 0.8	1, 0.8
$y_i = 2$	1, 0.5	1, 0.5	1, 0.5

$$s_1(y_i = 1, x, i) = 1; \mu_{1,1} = 1.0; \mu_{1,2} = 0.8; \mu_{1,3} = 0.8$$

$$s_2(y_i = 2, x, i) = 1; \mu_{2,1} = 0.5; \mu_{2,2} = 0.5; \mu_{2,3} = 0.5$$

CRF

s,μ	i=1	i=2	i=3
y_i = 1	1,1.0	1,0.8	1,0.8
y_i = 2	1,0.5	1,0.5	1,0.5

t,λ	i=2	i=3
y_{i-1},y_i = 1,1	1, 0.6	0
y_{i-1},y_i = 1,2	1, 1.0	1, 1.0
y_{i-1},y_i = 2,1	1, 1.0	1, 1.0
y_{i-1},y_i = 2,2	0	1, 0.2

- 求解标记为(1,2,2)的非规范概率值:

$$p(y|x) \propto \exp\left(\sum_{i,k} \lambda_{k,i} t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_{l,i} s_l(y_i, x, i)\right)$$

$$p(y_1 = 1, y_2 = 2, y_3 = 2|x)$$

$$\propto \exp((1.0 * 1 + 0.2 * 1) + (1.0 * 1 + 0.5 * 1 + 0.5 * 1)) = \exp(3.2)$$

- 为了便捷求解，将取值相同的当做一个特征函数，将式子转换为：

$$p(y|x) = \frac{1}{Z(x)} e^{\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)}$$

CRF

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

- 为了便捷求解，将取值相同的当做一个特征函数，将式子转换为：

t,λ	i=2	i=3
y _{i-1} ,y _i = 1,1	1, 0.6	0
y _{i-1} ,y _i = 1,2	1, 1.0	1, 1.0
y _{i-1} ,y _i = 2,1	1, 1.0	1, 1.0
y _{i-1} ,y _i = 2,2	0	1, 0.2

s,μ	i=1	i=2	i=3
y _i = 1	1,1.0	1,0.8	1,0.8
y _i = 2	1,0.5	1,0.5	1,0.5

$$t_1(y_1 = 1, y_2 = 1, x, 2) = 1; \lambda_{1,2} = 0.6$$

$$t_2(y_{i-1} = 1, y_i = 2, x, i) = 1; \lambda_2 = 1.0; i = 2, 3$$

$$t_3(y_{i-1} = 2, y_i = 1, x, i) = 1; \lambda_3 = 1.0; i = 2, 3$$

$$t_4(y_2 = 2, y_3 = 2, x, 3) = 1; \lambda_4 = 0.2$$

$$s_1(y_1 = 1, x, 1) = 1; \mu_1 = 1.0$$

$$s_2(y_i = 2, x, i) = 1; \mu_2 = 0.5; i = 1, 2, 3$$

$$s_3(y_i = 1, x, i) = 1; \mu_1 = 0.8; i = 2, 3$$

$$p(y_1 = 1, y_2 = 2, y_3 = 2 | x)$$

$$\propto \exp\left(\begin{aligned} &(0.6 * (0 + 0) + 1.0 * (1.0 + 0.0) + 1.0 * (0 + 0) + 0.2 * (0 + 1)) \\ &+ (1.0 * (1 + 0 + 0) + 0.5 * (0 + 1 + 1) + 0.8 * (0 + 0 + 0)) \end{aligned}\right) = \exp(3.2)$$

CRF

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

$$f_k(y, x) = \sum_i f_k(y_{i-1}, y_i, x, i)$$

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

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_k w_k f_k(y, x)\right) \quad Z(x) = \sum_k \exp\left(\sum_k w_k f_k(y, x)\right)$$

CRF

- 使用对数似然函数的相反数作为损失函数的值：

$$L(w) = \sum_{x,y} \bar{P}(x,y) \ln P_w(y|x)$$

先验分布概率

给定x的情况下
y的预测概率值

$$loss = -L(w)$$

CRF

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_k w_k f_k(y, x)\right) \quad Z(x) = \sum_k \exp\left(\sum_k w_k f_k(y, x)\right)$$

- 化简损失函数的形式(方便使用梯度下降)

$$\begin{aligned} loss &= -\sum_{x,y} \bar{P}(x, y) \ln P_w(y|x) \\ &= \sum_{x,y} \bar{P}(x, y) \ln Z_w(x) - \sum_{x,y} \bar{P}(x, y) \sum_k w_k f_k(x, y) \\ &= \sum_x \bar{P}(x) \ln Z_w(x) - \sum_{x,y} \bar{P}(x, y) \sum_k w_k f_k(x, y) \\ &= \sum_x \bar{P}(x) \ln \left(\sum_y \exp \left(\sum_k w_k f_k(x, y) \right) \right) - \sum_{x,y} \bar{P}(x, y) \sum_k w_k f_k(x, y) \end{aligned}$$

- 在损失函数中对 w 求导即可得到如下式子，也就表示可以使用梯度下降法进行优化操作。

$$\frac{\partial loss}{\partial w_k} = \sum_{x,y} \bar{P}(x) P_{w_k}(y|x) f_k(x,y) - \sum_{x,y} \bar{P}(x,y) f_k(x,y)$$

CRF

- 类似HMM的维特比算法，在CRF中，预测问题也是通过输入序列 x ，求解 $P(Y|X)$ 对应的概率最大的序列 Y 。

$$\delta_i(l) = \max_{1 \leq j \leq m} \left\{ \delta_{i-1}(j) + \sum_{k=1}^K w_k f_k(y_{i-1} = j, y_i = l, x, i) \right\}$$

$$y_n^* = \arg \max_{1 \leq j \leq m} \delta_n(j)$$

CRF

- 假定Linear CRF在词性标注中的案例，简化词性类别: {1: 名词, 2: 动词}; 假定输入三个词的语句 $X=(X1, X2, X3)$ ，输出标记记为 $Y=(Y1, Y2, Y3)$ ；求解最有可能的序列。

t, λ	$i=2$	$i=3$
$y_{i-1}, y_i = 1, 1$	1, 0.6	0
$y_{i-1}, y_i = 1, 2$	1, 1.0	1, 1.0
$y_{i-1}, y_i = 2, 1$	1, 1.0	1, 1.0
$y_{i-1}, y_i = 2, 2$	0	1, 0.2

s, μ	$i=1$	$i=2$	$i=3$
$y_i = 1$	1, 1.0	1, 0.8	1, 0.8
$y_i = 2$	1, 0.5	1, 0.5	1, 0.5

$$t_1(y_1 = 1, y_2 = 1, x, 2) = 1; \lambda_{1,2} = 0.6$$

$$t_2(y_{i-1} = 1, y_i = 2, x, i) = 1; \lambda_2 = 1.0; i = 2, 3$$

$$t_3(y_{i-1} = 2, y_i = 1, x, i) = 1; \lambda_3 = 1.0; i = 2, 3$$

$$t_4(y_2 = 2, y_3 = 2, x, 3) = 1; \lambda_4 = 0.2$$

$$s_1(y_1 = 1, x, 1) = 1; \mu_1 = 1.0$$

$$s_2(y_i = 2, x, i) = 1; \mu_2 = 0.5; i = 1, 2, 3$$

$$s_3(y_i = 1, x, i) = 1; \mu_1 = 0.8; i = 2, 3$$

- 第一个时刻的值:

$$\delta_1(l) = \sum_{k=1}^K w_k f_k(y_0 = start, y_1 = l, x, 1)$$

$$\delta_1(1) = 1.0 \quad \delta_1(2) = 0.8$$

CRF

t,λ	i=2	i=3
y _{i-1} ,y _i = 1,1	1, 0.6	0
y _{i-1} ,y _i = 1,2	1, 1.0	1, 1.0
y _{i-1} ,y _i = 2,1	1, 1.0	1, 1.0
y _{i-1} ,y _i = 2,2	0	1, 0.2

s,μ	i=1	i=2	i=3
y _i = 1	1,1.0	1,0.8	1,0.8
y _i = 2	1,0.5	1,0.5	1,0.5

$$t_1(y_1 = 1, y_2 = 1, x, 2) = 1; \lambda_{1,2} = 0.6$$

$$t_2(y_{i-1} = 1, y_i = 2, x, i) = 1; \lambda_2 = 1.0; i = 2, 3$$

$$t_3(y_{i-1} = 2, y_i = 1, x, i) = 1; \lambda_3 = 1.0; i = 2, 3$$

$$t_4(y_2 = 2, y_3 = 2, x, 3) = 1; \lambda_4 = 0.2$$

$$s_1(y_1 = 1, x, 1) = 1; \mu_1 = 1.0$$

$$s_2(y_i = 2, x, i) = 1; \mu_2 = 0.5; i = 1, 2, 3$$

$$s_3(y_i = 1, x, i) = 1; \mu_1 = 0.8; i = 2, 3$$

- 第二个时刻的值:

$$\delta_2(l) = \arg \max_{1 \leq j \leq 2} \left(\delta_1(j) + \sum_{k=1}^K w_k f_k(y_1 = j, y_2 = l, x, 2) \right)$$

$$\delta_2(1) = \arg \max_{1 \leq j \leq 2} (1.0 + (0.6 + 1.0 + 0.8), 0.8 + (1.0 + 0.5 + 0.8)) = 3.4$$

$$\delta_2(2) = \arg \max_{1 \leq j \leq 2} (0.8 + (1.0 + 1.0 + 0.5), 0.8 + (0 + 0.5 + 0.5)) = 3.5$$

- 第三个时刻的值:

$$\delta_3(l) = \arg \max_{1 \leq j \leq 2} \left(\delta_2(j) + \sum_{k=1}^K w_k f_k(y_2 = j, y_3 = l, x, 3) \right)$$

$$\delta_3(1) = \arg \max_{1 \leq j \leq 2} (3.4 + (0 + 0.8 + 0.8), 3.5 + (1.0 + 0.5 + 0.8)) = 5.8$$

$$\delta_3(2) = \arg \max_{1 \leq j \leq 2} (3.4 + (1.0 + 0.8 + 0.5), 3.5 + (0.2 + 0.5 + 0.5)) = 5.7$$

- 反向回溯:

$$\delta_1(1) = 1.0 \quad \delta_1(2) = 0.8$$

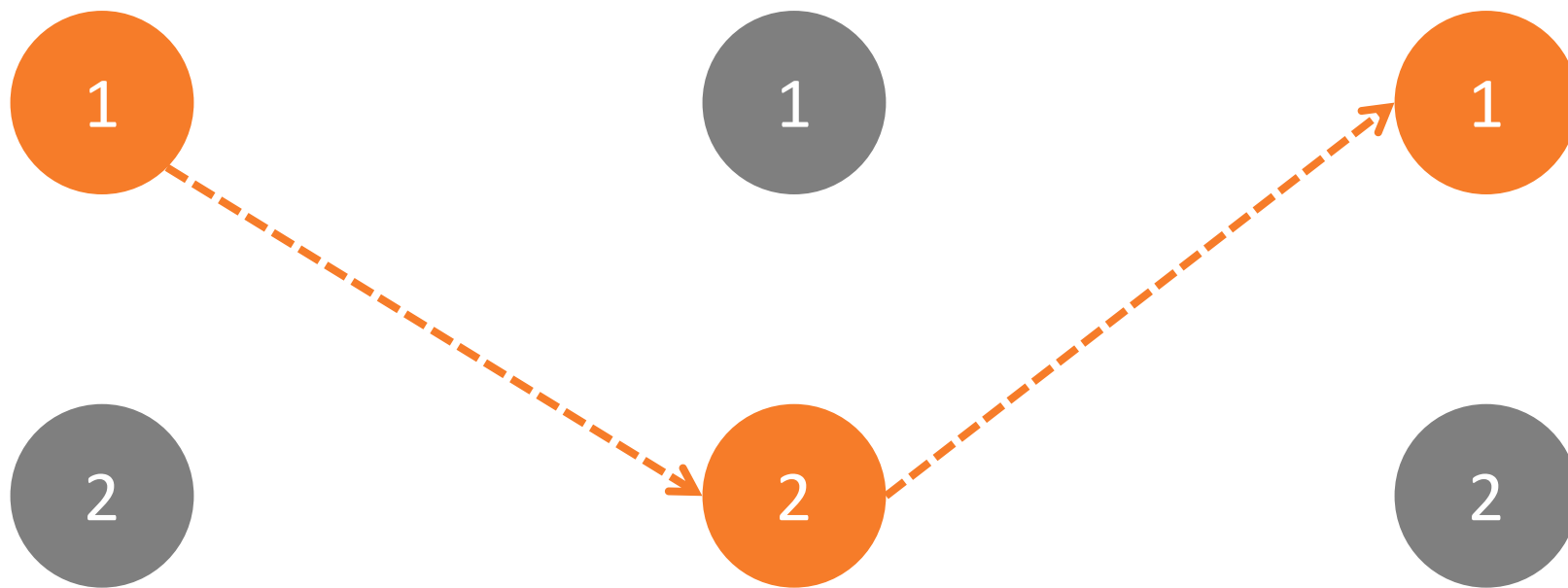
$$\delta_2(1) = \arg \max_{1 \leq j \leq 2} (1.0 + (0.6 + 1.0 + 0.8), 0.8 + (1.0 + 0.5 + 0.8)) = 3.4$$

$$\delta_2(2) = \arg \max_{1 \leq j \leq 2} (1.0 + (1.0 + 1.0 + 0.5), 0.8 + (0 + 0.5 + 0.5)) = 3.5$$

$$\delta_3(1) = \arg \max_{1 \leq j \leq 2} (3.4 + (0 + 0.8 + 0.8), 3.5 + (1.0 + 0.5 + 0.8)) = 5.8$$

$$\delta_3(2) = \arg \max_{1 \leq j \leq 2} (3.4 + (1.0 + 0.8 + 0.5), 3.5 + (0.2 + 0.5 + 0.5)) = 5.7$$

CRF



最有可能的序列为:1,2,1; 也就是名词、动词、名词

- CRF和HMM的主要区别：
 - HMM优化的是求解 $p(x,y)$ 联合概率，CRF优化的是求解 $p(y|x)$ 条件概率；
 - HMM中有向无环图，CRF中是无向图；

- CRF应用场景:
 - 中文分词
 - 词性标注
 - 命名实体识别
 - 语义角色标记
 - 事件提取
 - 等等.....

