一、从HMM到CRF

1、任务类型：Sequence Labling 输入和输出都是一个序列

$x_1 \quad x_2 \quad \cdots \quad x_L$ (seq)

$\downarrow \quad \downarrow \quad \cdots \quad \downarrow$

$y_1 \quad y_2 \quad \cdots \quad y_L$ (seq)

2、案例：POS Tagging (词性标注)

3、HMM步骤

① 首先构建词性(隐藏状态)的转移矩阵&初始概率，

$P(PN\ V\ D\ N) = P(PN \mid start)\, P(V \mid PN)\, P(V \mid PN)\, P(D \mid V)\, P(N \mid D)$

计算所有可能的词性序列(联合概率)

② 然后计算联合概率的前提下，由发射矩阵，确定取当前字词的条件概率

$P(John\ saw\ the\ saw \mid PN\ V\ D\ N) = P(John \mid PN) * P(saw \mid V) * P(the \mid D) * P(saw \mid N)$

③ 令 $P(PN\ V\ P\ N) = p(y)$

$P(John\ saw\ the\ saw \mid PN\ V\ P\ N) = p(x \mid y)$

所以：$P(x, y) = p(y)\, p(x \mid y)$

公式：

step 1：$p(y) = P(y_1 \mid start) \times \prod_{L=1}^{L-1} P(y_{L+1} \mid y_L) \times P(end \mid y_L)$ 转移概率

step 2：$p(x \mid y) = \prod_{L=1}^{L} P(x_L \mid y_L)$ 发射概率

4、HMM问题求解：$y = \arg\max P(y \mid x) = \arg\max \frac{P(x, y)}{P(x)}$

$P(x)$已知 $y = \arg\max P(x, y)$

① 穷举所有的 y，找到 $p(x,y)$ 最大的解
② 维特比算法

5、问题：

HMM的的矩阵是基于语料库的，对于未出现在语料库的组合HMM会统计为一个值，但这个值不一定很小

↓

所以这一情况存在数据噪很小的情况，HMM效果一般会更好

↓

原因是：转移矩阵和发射矩阵之间是独立的，要分别 model

↓

在使用上述矩阵的前提下，解决HMM的问题

6、CRF：

① 模型假设 $p(x,y) \propto \exp(w \cdot \phi(x,y))$

- $\phi(x,y)$ 特征矩阵
- w 权重（通过训练学习的）
- $\exp(w \cdot \phi(x,y))$：确保值大于1

$$p(x,y) = \frac{\exp(w \cdot \phi(x,y))}{R}$$

$$p(y|x) = \frac{p(x,y)}{\sum_{y'} p(x,y')} = \frac{\exp(w \cdot \phi(x,y)) \times R}{\sum_{y'} \exp(w \cdot \phi(x,y')) \times R}$$

$$= \frac{\exp(w \cdot \phi(x,y))}{\sum_{y'} \exp(w \cdot \phi(x,y'))} \quad (\textstyle\sum_{y'} \text{所以}y\text{的可能})$$

↓

归一化系数，总和为1有关

$$= \exp(w \cdot \phi(x,y)) / Z(x)$$

② 样本说明：假设一个 Batch N 个样本，一个序列长为 n

上标 i：第 i 个样本　下标 t：第 t 个样本

$(l_1, l_2, \cdots)$ $(s_1, s_2, \cdots)$

$x_1^1, x_2^1, \cdots, x_n^1 \longrightarrow y_1^1, y_2^1, \cdots, y_n^1$
$x_1^2, x_2^2, \cdots, x_n^2 \longrightarrow y_1^2, y_2^2, \cdots, y_n^2$
$\vdots$
$x_1^N, x_2^N, \cdots, x_n^N \longrightarrow y_1^N, y_2^N, \cdots, y_n^N$

③ 标签说明：$P(y|x) = \frac{e^{w \cdot \phi(x,y)}}{\sum_y e^{w \cdot \phi(x,y)}}$ 这一共2个单词，S个标签

- 单词和标签之间组合：$|L| \times |S|$
- 标签之间组合：$|S| \times |S|$
- start $\to s$ 和 $s \to$ end：$2 \times |S|$

所以 $\phi_k(x,y)$ 是一个计数值为 $N_{s,t}(x,y)$，表示特点样本 (x,y) 中发生的次数，

已知HMM $P(x,y) = P(y)\,P(x|y)$

$$P(y_1|\text{start}) \times \prod_{l=1}^{L-1} P(y_{l+1}|y_l) \times P(\text{end}|y_L) \times \prod_{l=1}^{L} P(x_l|y_l)$$

$$\log P(x,y) = \log P(y_1|\text{start}) + \sum_{l}^{L-1} \log P(y_{l+1}|y_l) + \log P(\text{end}|y_L) + \sum_{l=1}^{L} \log P(x_l|y_l)$$

比如：x The dog ate the homework
y D N V D N

这里 (The, D) 出现2次，所以 $\log P(x,y)$ 可以按次数进行统计：

$$\log P(x,y) = \sum_s \log P(s|\text{start}) \times N_{\text{start},s}(x,y) + \quad (\text{tags } s \text{ 转 tags } s')$$
$$\sum_{s,s'} \log P(s'|s) \times N_{s,s'}(x,y) +$$
$$\sum_s \log P(\text{end}|s) \times N_{s,\text{end}}(x,y) +$$
$$\sum_{s,t} \log P(t|s) \times N_{s,t}(x,y)$$
(t: word 的 tags)

所以 $\log P(x,y)$ 可以看作是两个向量相乘：

$$\begin{bmatrix} \vdots \\ \log P(H|S) \\ \vdots \\ \log P(S|start) \\ \vdots \\ \log P(S'|S) \\ \vdots \\ \log P(end|S) \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} \vdots \\ N_{S,H}(x,y) \\ \vdots \\ N_{start,S}(x,y) \\ \vdots \\ N_{S,S'}(x,y) \\ \vdots \\ N_{S,end}(x,y) \\ \vdots \end{bmatrix} = w \cdot \phi(x,y)$$

不同样本 (x^i, y^i) 的 $\phi(x^i, y^i)$ 不同，但 w 训练参数共享

④ CRF++ 区别：

1°、没有 start 和 end 标签（更灵活）

2°、特征使用模板拼接，可以关注词之间的组合

↓

相同的词的模板不同，它的词缀也不同