



北京航空航天大學  
BEIHANG UNIVERSITY

# 自然语言处理

人工智能研究院

主讲教师 沙磊

# Contents

- 生成模型与判别模型
- 隐马尔科夫模型 HMM
- 条件随机场模型 CRF
- 词性标注

# 生成模型与判别模型

- Generative Models (Two-step)
  - Infer class-conditional densities  $p(x|C_k)$  and priors  $p(C_k)$
  - then use Bayes theorem to determine posterior probabilities.

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

- Discriminative Models (One-step)
  - Directly infer posterior probabilities  $p(C_k|x)$

# 生成模型

- Given  $M$  variables,  $\mathbf{x} = (x_1, \dots, x_M)$ , class variable  $y$  and joint distribution  $p(x,y)$  we can
  - Marginalize 
$$p(y) = \sum_x p(x, y)$$
  - Condition 
$$p(y|x) = \frac{p(x, y)}{p(x)}$$
    - By conditioning the joint pdf we form a classifier
- Huge need for samples
  - If  $x_i$  are binary, need  $2^M$  values to specify  $p(x,y)$

# ML方法分类

- Generative Methods
  - “Generative” since sampling can generate synthetic data points
  - Popular models
    - Naïve Bayes, Mixtures of multinomials
    - Mixtures of Gaussians, Hidden Markov Models
    - Bayesian networks, Markov random fields
- Discriminative Methods
  - Focus on given task – better performance
  - Popular models
    - Logistic regression, SVMs
    - Traditional neural networks, Nearest neighbor
    - Conditional Random Fields (CRF)



# 隐马尔科夫 模型

# 隐马尔科夫模型

- 隐马尔科夫模型(Hidden Markov Model, HMM)是对马尔科夫模型的一种扩充。
- 隐马尔科夫模型的基本理论形成于上世纪60年代末期和70年代初期。(L.E.Baum)
- 70年代, CMU的J.K.Baker以及IBM 的F.Jelinek 等把隐马尔科夫模型应用于语音识别。
- 隐马尔科夫模型在计算语言学中有着广泛的应用。例如隐马尔科夫模型在词类自动标注中的应用。

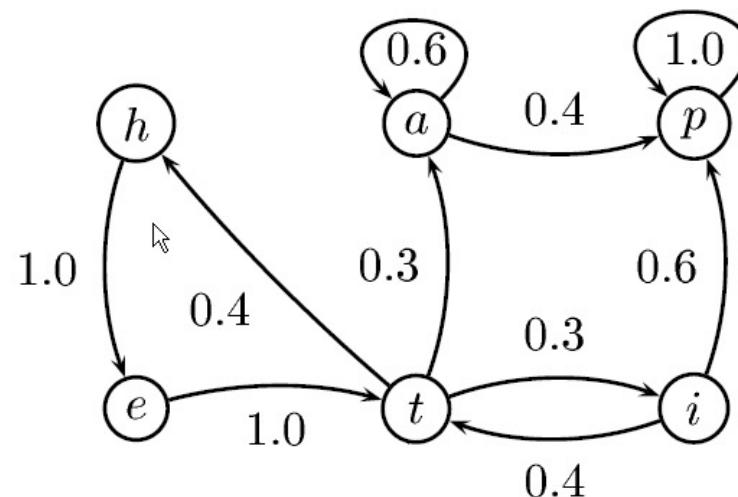
# 马尔科夫模型

- 马尔科夫模型是由Andrei A. Markov于1913年提出的。
- 设 $S$ 是一个由有限个状态组成的集合。
  - $S=\{1, 2, 3, \dots, n-1, n\}$
- 随机序列 $X$ 在 $t$ 时刻所处的状态为 $q_t$ , 其中 $q_t \in S$ , 若有:
  - $P(q_t = j | q_{t-1} = i, q_{t-2} = k, \dots) = P(q_t = j | q_{t-1} = i)$
  - 则随机序列 $X$ 构成一个一阶马尔科夫链。 (Markov Chain)
- 令  $P(q_t = j | q_{t-1} = i) = P(q_s = j | q_{s-1} = i)$ , 则对于所有的 $i, j$  有下面的关系成立:

$$a_{ij} = P(q_t = j | q_{t-1} = i) \quad 1 \leq i, j \leq n \quad \sum_{j=1}^n a_{ij} = 1 \quad a_{ij} \geq 0$$

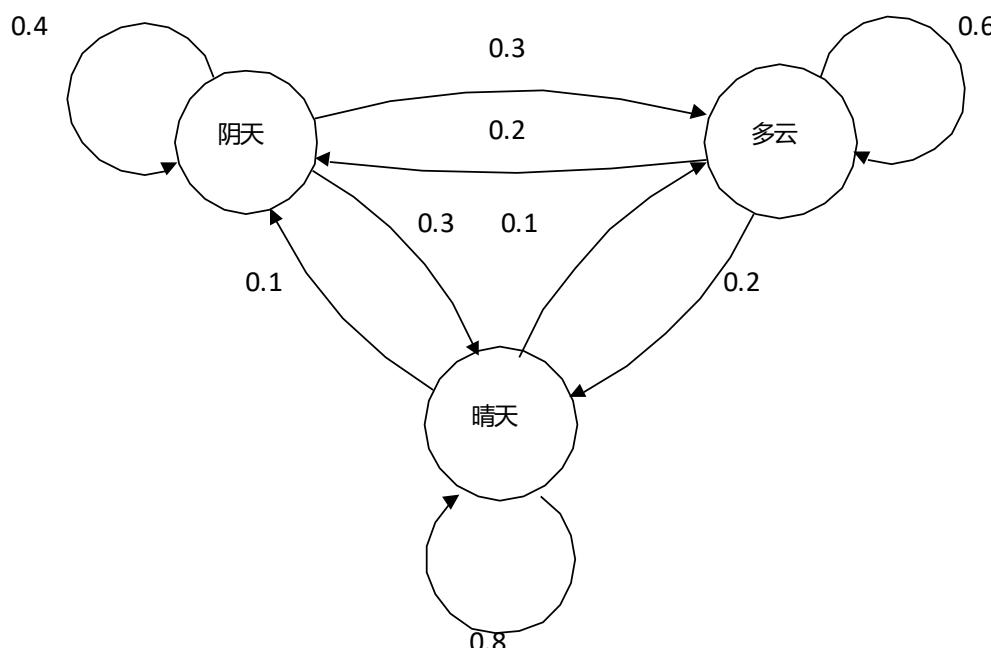
# 马尔科夫模型

- 一阶马尔科夫模型可以描述为一个二元组(  $S, A$  )，  $S$ 是状态的集合，而 $A$ 是所有状态转移概率组成的一个 $n$ 行 $n$  列的矩阵，其中每一个元素 $a_{ij}$ 为从状态 $i$ 转移到状态 $j$ 的 概率。
- 同有限状态自动机类似，状态转移关系也可以用状态转换图来表示。



# 马尔科夫模型举例

- 天气的变化，三种状态{1(阴天), 2(多云), 3(晴天)}。
- 今天的天气情况仅和昨天的天气状况有关。
- 根据对历史数据的观察得到下列状态转移关系。



$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

# 马尔科夫模型

- 如果把晴天称为状态3的输出，阴天称为状态1 的输出，多云称为状态2的输出。 因为状态和输出是一对一的关系，所以根据观察到的输出序列就可以决定模型中的状态转换序列。
- 对于马尔科夫模型，给定了观察序列，同时也确定了状态转换序列。

例如有关天气状况的观察序列。

(晴 晴 晴 阴 阴 晴 云 晴)

则状态转换序列为

(3, 3, 3, 1, 1, 3, 2, 3 )

# 坛子与小球

- 在一个房间中，假定有  $N$  个坛子，每个坛子中都装有不同颜色的小球，并且假定总共有  $M$  种不同颜色的小球。
- 一个精灵在房间中首先随机地选择一个坛子，再从这个坛子中随机选择一个小球，并把小球的颜色报告给房间外面的人员记录下来作为观察值。
- 精灵然后把球放回到坛子中，以当前的坛子为条件再随机选择一个坛子，从中随机选择一个小球，并报告小球的颜色，然后放回小球，如此继续...，随着时间的推移，房间外的人会得到由这个过程产生的一个小球颜色的序列。

# 坛子与小球

- 如果令每一个坛子对应与一个状态，令小球颜色对应状态的输出。
- 可以用一个一阶马尔科夫过程来描述坛子的选择过程。
- 在马尔科夫过程中，每个状态只有一个输出，但在坛子和小球的问题中。可以从每个坛子中拿出不同颜色的小球。也就是每个状态能按照特定的概率分布产生多个输出，状态和输出之间不再是一一对应关系。
- 在坛子与小球问题中，如果给定一个观察序列(不同颜色的小球序列)，不能直接确定状态转换序列(坛子的序列)，因为状态转移过程被隐藏起来了。所以这类随机过程被称为隐马尔科夫过程。

# 隐马尔科夫模型

- 隐马尔可夫模型 $\lambda$ 可以表示为一个五元组( $S, V, A, B, \pi$ )
  - ❖  $S$ 是一组状态的集合。
  - ❖  $S = \{1, 2, 3, \dots, N\}$  (状态 $n$ 对应坛子 $n$ )
  - ❖  $V$ 是一组输出符号组成的集合。
  - ❖  $V = \{v_1, v_2, v_3, \dots, v_M\}$  ( $v_1$ 对应红色小球)
  - ❖  $A$ 是状态转移矩阵， $N$ 行 $N$ 列。
  - ❖  $A = [a_{ij}]$
  - ❖  $a_{ij} = P(q_{t+1}=j \mid q_t=i), 1 \leq i, j \leq N$

# 隐马尔科夫模型

- ❖  $B$  是输出符号的概率分布。
- ❖  $B = \{ b_j(k) \}$   $b_j(k)$  表示在状态  $j$  时输出符号  $v_k$  的概率
- ❖  $b_j(k) = P(v_k | j), 1 \leq k \leq M, 1 \leq j \leq N$
  
- ❖  $\pi$  是初始状态概率分布  $\pi = \{ \pi_i \}$
- ❖  $\pi_i = P(q_1 = i)$  表示时刻 1 选择某个状态的概率。
  
- 隐马尔可夫过程是一个双重随机过程，其中一重随机过程不能直接观察到，通过状态转移概率矩阵描述。另一重随机过程输出可以观察到的观察符号，这由输出概率来定义。

# 利用隐马尔科夫模型生成观察序列

- 可以把隐马尔可夫模型看做一个符号序列的生成装置，按照一定的步骤，隐马尔可夫模型可以生成下面的符号序列：
- $O = (o_1 o_2 o_3 \dots o_T)$

1. 令  $t = 1$ ，按照初始状态概率分布  $\pi$  选择一个初始状态  $q_1 = i$ 。
2. 按照状态  $i$  输出符号的概率分布  $b_i(k)$  选择一个输出值  $o_t = v_k$ 。
3. 按照状态转移概率分布  $a_{ij}$  选择一个后继状态  $q_{t+1} = j$ 。
4. 若  $t < T$ ，令  $t = t + 1$ ，并且转移到算法第2步继续执行，否则结束。

# 抛掷硬币

- 三枚硬币，随机选择一枚，进行抛掷，记录抛掷结果。可以描述为一个三个状态的隐马尔科夫模型 $\lambda$ 。

- $\lambda = (S, V, A, B, \pi)$ , 其中

- $S = \{1, 2, 3\}$

- $V = \{H, T\}$

$A$  如下表所示

	1	2	3
1	0.9	0.05	0.05
2	0.45	0.1	0.45
3	0.45	0.45	0.1

$B$  如下表所示

	1	2	3
H	0.5	0.75	0.25
T	0.5	0.25	0.75

$$\pi = \{1/3, 1/3, 1/3\}$$

# 抛掷硬币

- 问题一：

给定上述模型，观察到下列抛掷结果的概率是多少？

$$O = (H H H H T H T T T)$$

- 问题二：

给定上述模型，若观察到上述抛掷结果，最可能的硬币选择序列  
(状态转换序列)是什么？

- 问题三：

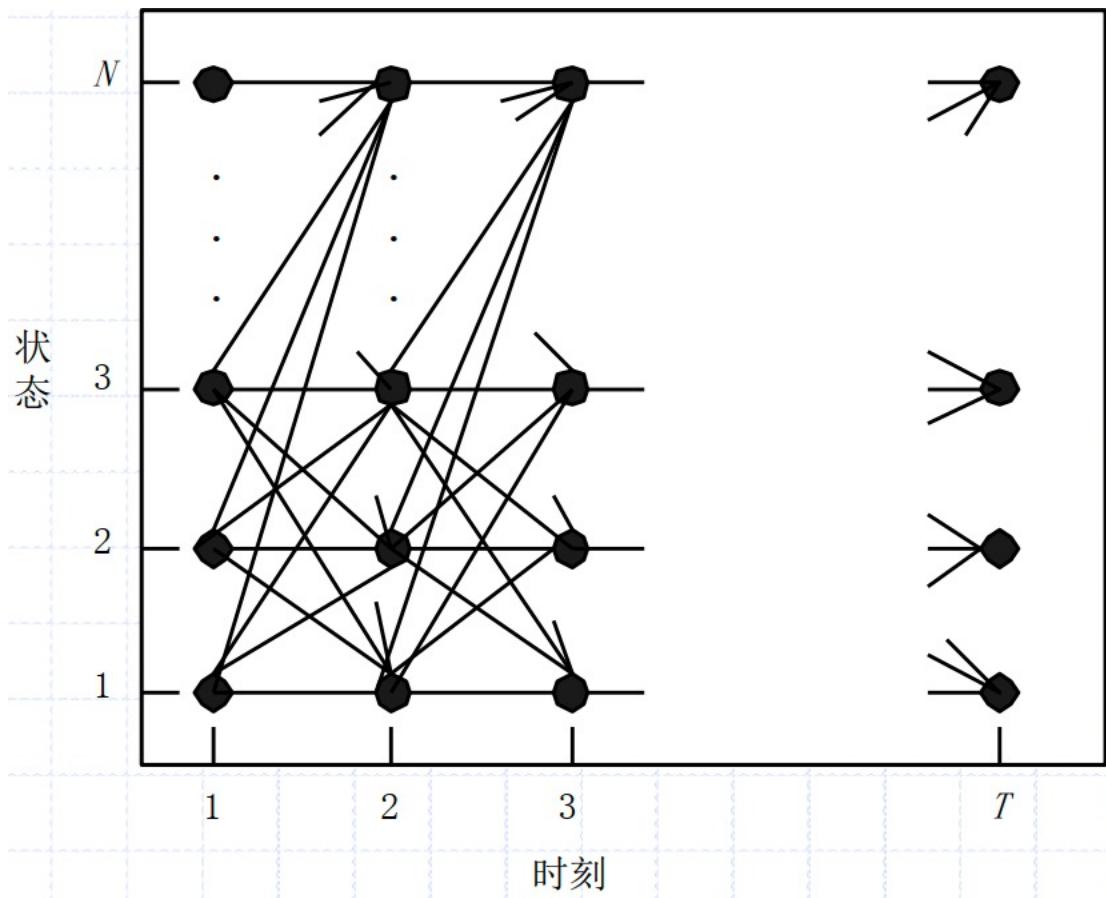
若上述模型中的状态转移矩阵  $A$ 、状态输出概率  $B$  和初始状态分布  $\pi$  均未知，如何根据观察序列得到它们？

# 隐马尔科夫模型的三个问题

- 给定HMM  $\lambda = (A, B, \pi)$ , 给定观察序列  $O = (o_1 o_2 o_3 \dots o_T)$ , 如何有效地计算出观察序列的概率, 即  $P(O|\lambda)$ ? (估算问题) (另一种语言模型)
- 给定HMM  $\lambda = (A, B, \pi)$ , 给定观察序列  $O = (o_1 o_2 o_3 \dots o_T)$ , 如何寻找一个状态转换序列  $q = (q_1 q_2 q_3 \dots q_T)$ , 使得该状态转换序列最有可能产生上述观察序列? (解码问题)
- 在模型参数未知或不准确的情况下, 如何根据观察序列  $O = (o_1 o_2 o_3 \dots o_T)$  求得模型参数或调整模型参数。按照 MLE 的原则, 即如何确定一组模型参数, 使得  $P(O|\lambda)$  最大? (学习问题或训练问题)

# 问题1：估算观察序列概率

- 对隐马尔可夫模型而言，状态转换序列是隐藏的，一个观察序列可能由任何一种状态转换序列产生。因此要计算一个观察序列的概率值，就必须考虑所有可能的状态转换序列。
- 右图表示了产生观察序列 $O = (o_1 o_2 o_3 \dots o_T)$ 的所有可能的状态转换序列。



# 估算观察序列概率

- 给定  $\lambda$ , 以及状态转换序列  $q = (q_1 q_2 q_3 \dots q_T)$  产生观察序列  $O = (o_1 o_2 o_3 \dots o_T)$  的概率可以通过下面的公式计算:

$$P(O|q, \lambda) = b_{q_1}(o_1)b_{q_2}(o_2)\dots b_{q_T}(o_T)$$

- 给定  $\lambda$ , 状态转换序列  $q = (q_1 q_2 q_3 \dots q_T)$  的概率可以通过下面的公式计算:

$$P(q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T}$$

- 则  $O$  和  $q$  的联合概率为:  $P(O, q|\lambda) = P(O|q, \lambda)P(q|\lambda)$

- 考虑所有的状态转换序列, 则

$$P(O|\lambda) = \sum_q P(O, q|\lambda) = \sum_{q_1 \dots q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \dots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

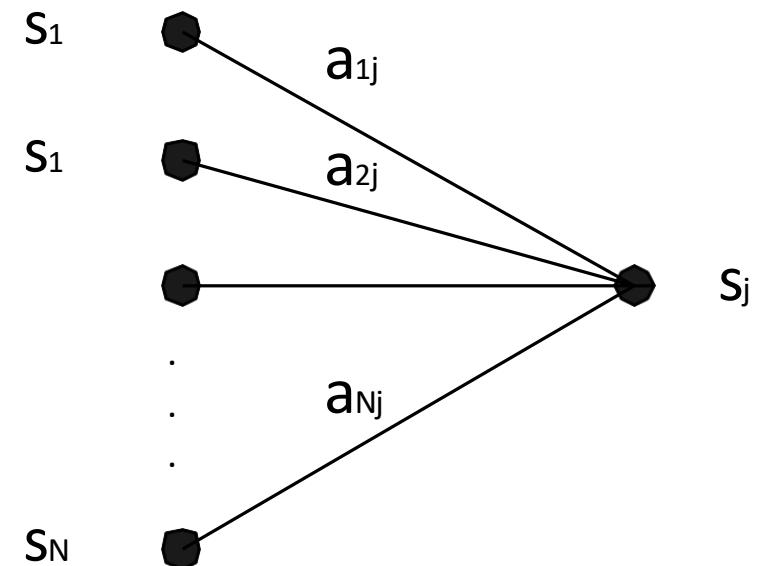
# 估算观察序列概率

- 理论上，可以通过穷举所有状态转换序列的办法计算观察序列 $O$ 的概率。
- 实际上，这样做并不现实。
  - 可能的状态转换序列共有 $N^T$ 个。
  - 需要做 $(2T-1)N^T$ 次乘法运算， $N^T-1$  次加法运算。
  - 若 $N=5$ ,  $T=100$ , 则  $(2 \times 100-1) \times 5^{100} \approx 10^{72}$
- 需要寻找更为有效的计算方法。

# 向前算法(Forward Algorithm)

- 向前变量  $\alpha_t(i)$   $\alpha_t(i) = P(o_1 o_2 o_3 \dots o_t, q_t = i | \lambda)$
- $\alpha_t(i)$  的含义是，给定模型  $\lambda$ ，时刻  $t$ ，处在状态  $i$ ，并且部分观察序列  $o_1 o_2 o_3 \dots o_t$  的概率。
- 显然有  $\alpha_1(i) = \pi_i b_i(o_1) (1 \leq i \leq N)$
- 若  $\alpha_t(i) (1 \leq i \leq N)$  已知，如何计算  $\alpha_{t+1}(i)$ ？

$$\alpha_{t+1}(j) = \left[ \sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), 1 \leq t \leq T-1, 1 \leq j \leq N$$



# 向前算法

1. 初始化  $\alpha_1(i) = \pi_i b_i(o_1) (1 \leq i \leq N)$

2. 迭代计算

$$\alpha_{t+1}(j) = \left[ \sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), 1 \leq t \leq T-1, 1 \leq j \leq N$$

3. 终止

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i)$$

## ◆ 计算量

- $N(N+1)(T-1)+N$  次乘法
- $N(N-1)(T-1)$  次加法
- 若  $N=5, T=100$ , 则  
大约需要 5000 次运算

# 计算实例

- 抛掷硬币问题，计算观察到( $HHT$ )的概率。

$\alpha_t(i)$	$H$	$H$	$T$	$P(HHT   \lambda)$
1	0.16667	0.15000	0.08672	
2	0.25000	0.05312	0.00684	0.11953
3	0.08333	0.03229	0.02597	

# 向后算法 (Backward Algorithm)

- 向后变量  $\beta_t(i) \quad \beta_t(i) = P(o_{t+1} o_{t+2} \dots o_T | q_t = i, \lambda)$
- $\beta_t(i)$  的含义是，在给定模型  $\lambda$ ，时刻  $t$ ，处在状态  $i$ ，并且部分观察序列为  $o_{t+1} o_{t+2} \dots o_T$  的概率。

$$\beta_T(i) = 1$$

- 若  $\beta_{t+1}(j) (1 \leq j \leq N)$  已知，如何计算  $\beta_t(i)$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), 1 \leq t \leq T-1, 1 \leq j \leq N$$

# 向后算法

1. 初始化  $\beta_T(i) = 1 (1 \leq i \leq N)$

2. 迭代计算

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), 1 \leq t \leq T-1, 1 \leq j \leq N$$

3. 终止

$$P(O|\lambda) = \sum_{i=1}^N \pi_i b_i(o_1) \beta_1(i)$$

# 计算实例

- 抛掷硬币问题，计算观察到 $(H\ H\ T)$ 的概率。

	$H$	$H$	$T$		$P(H\ H\ T \lambda)$
$\beta_t(i)$	$\pi_i b_i(H) \beta_1(i)$	$\beta_1(i)$	$\beta_2(i)$	$\beta_3(i)$	
1	0.04203	0.25219	0.50000	1.00000	
2	0.05074	0.20297	0.58750	1.00000	0.11953
3	0.02676	0.32109	0.41250	1.00000	

# 求解最佳状态转换序列

- 隐马尔可夫模型的第二个问题是计算出一个能最好解释观察序列的状态转换序列。
- 理论上，可以通过枚举所有的状态转换序列，并对每一个状态转换序列 $q$ 计算 $P(O, q | \lambda)$ ，能使 $P(O, q | \lambda)$ 取最大值的状态转换序列 $q^*$ 就是能最好解释观察序列的状态转换序列，即：

$$q^* = \arg \max_q P(O, q | \lambda)$$

- 同样，这不是一个有效的计算方法，需要寻找更好的计算方法。

# 韦特比算法(Viterbi Algorithm)

- 韦特比变量  $\delta_t(i)$

$$\delta_t(i) = \max_{q_1 \dots q_{t-1}} P(q_1 q_2 \dots q_{t-1}, q_t = i, o_1 o_2 \dots o_t | \lambda)$$

- $\delta_t(i)$  的含义是，给定模型  $\lambda$ ，在时刻  $t$  处于状态  $i$ ，观察到  $o_1 o_2 o_3 \dots o_t$  的最佳状态转换序列为  $q_1 q_2 \dots q_t$  的概率。

$$\delta_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N$$

- 若  $\delta_t(i) (1 \leq i \leq N)$  已知，如何计算  $\delta_{t+1}(i)$ ？

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(o_{t+1})$$

- 如何记录路径？设定  $T$  个数组  $\psi_1(N), \psi_2(N), \dots, \psi_T(N)$

$\psi_t(i)$  记录在时刻  $t$  到达状态  $i$  的最佳状态转换序列  $t-1$  时刻的最佳状态。

# 韦特比算法

1. 初始化  $\delta_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N$   
 $\psi_1(i) = 0$

## 2. 迭代计算

$$\delta_t(j) = [\max_{1 \leq i \leq N} \delta_{t-1}(i)a_{ij}]b_j(o_t) \quad \psi_t(j) = \arg \max_{1 \leq i \leq N} \delta_{t-1}(i)a_{ij}]$$

## 3. 终止

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \quad q_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)]$$

## 4. 求解最佳路径

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T - 1, T - 2, \dots, 1$$

# 计算实例

- 抛掷硬币问题，观察到 $(HHT)$ ，寻找产生该观察序列的最佳路径以及最佳路径的概率。

$\delta_t(i)$	$H$	$H$	$T$	$P^*$
1	0.16667	0.07500	0.03375	
2	0.25000	0.02812	0.00316	0.03375
3	0.08333	0.02812	0.00949	

$\psi_t(i)$	$\psi_1(i)$	$\psi_2(i)$	$\psi_3(i)$	$q^*$
1	0	1	1	
2	0	3	3	1
3	0	2	2	

- 最佳状态转换序列为1 1 1

# 参数学习

- 隐马尔科夫模型的第三个问题是根据观察序列  $O = (o_1 o_2 o_3 \dots o_T)$  求得模型参数或调整模型参数，即如何确定一组模型参数使得  $P(O|\lambda)$  最大？
- 隐马尔科夫模型的前两个问题均假设模型参数已知，第三个问题是模型参数未知，求最佳模型的问题，是三个问题中最为困难的问题。

# 有指导的参数学习 (supervised learning)

- 在模型( $\lambda$ )未知的情况下，如果给定观察序列的同时，也给定了状态转换序列，此时可以通过有指导的学习方法学习模型参数。例如给定下面的训练数据，可以通过最大似然估计法估计模型参数：

$H/1 H/1 T/1 T/2 H/3 T/5 \dots$

$T/2 H/1 T/2 H/3 H/3 H/1 \dots$

- 参数学习非常简单，在训练数据足够大的前提下，效果不错。
- 缺点，状态信息未知时无法使用。或者要由人工标注状态信息，代价高。
- 在NLP中，在无指导学习效果不佳时，需要采用有指导学习。

# 无指导的参数学习 (unsupervised learning)

- 在模型( $\lambda$ )未知的情况下，如果仅仅给定了观察序列，此时学习模型的方法被称做无指导的学习方法。
- 对于隐马尔科夫模型而言，采用无指导学习方法，没有解析方法。通常要首先给定一组不准确的参数，再通过反复迭代逐步求精的方式调整模型参数，最终使参数稳定在一个可以接受的精度。
- 利用无指导的学习方法估计隐马尔科夫模型参数时，并不能一定保证求得最优模型，一般能得到一个局部最优模型。

# 直观的想法

- 给定一组初始参数( $A$   $B$   $\pi$ )
- 由于没有给定状态转换序列，无法计算状态转移的频率、状态输出的频率以及初始状态频率。
- 假定任何一种状态转换序列都可能。
- 对每种状态转换序列中的频次加权处理，计算状态转移、状态输出、以及初始状态的期望频次
- 利用计算出的期望频次更新 $A$ 、 $B$ 和 $\pi$

# 直观的想法

- 权值如何选择

对状态转换序列 $q$ 而言，选择 $P(q|O, \lambda)$

- 理论上可行，现实不可行

要考虑所有的状态转移路径

需要多次迭代，问题更为严重

- 需要更为有效的算法，即Baum-Welch算法

# Baum-Welch Algorithm

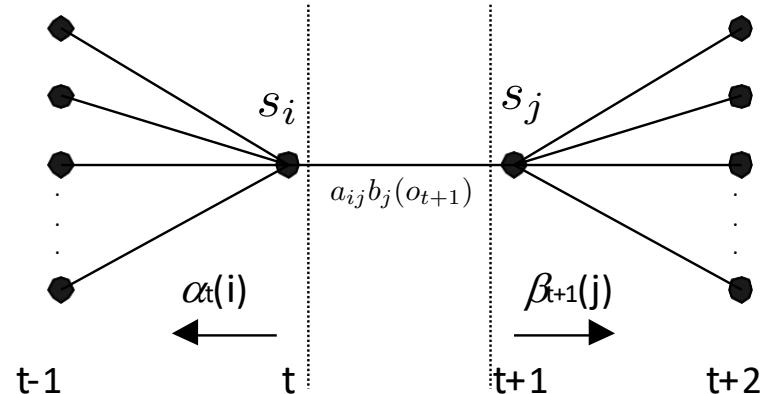
- 定义变量  $\xi_t(i, j)$

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j | O, \lambda)$$

- $\xi_t(i, j)$  含义是，给定模型  $\lambda$  和观察序列  $O$ ，在时刻  $t$  处在状态  $i$ ，时刻  $t+1$  处在状态  $j$  的期望概率。

- $\xi_t(i, j)$  可以进一步写成：

$$\begin{aligned}
 \xi_t(i, j) &= \frac{P(q_t = i, q_{t+1} = j, O | \lambda)}{P(O | \lambda)} \\
 &= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O | \lambda)} \\
 &= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}
 \end{aligned}$$



# Baum-Welch Algorithm

- 定义变量  $\gamma_t(i)$ , 令其表示在给定模型以及观察序列的情况下,  $t$  时刻处在状态  $i$  的概率, 则有:

$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j)$$

- 观察序列  $O$  中从状态  $i$  出发的转换的期望概率

$$\sum_{t=1}^{T-1} \gamma_t(i)$$

- 观察序列  $O$  中从状态  $i$  到状态  $j$  的转换的期望概率

$$\sum_{t=1}^{T-1} \xi_t(i, j)$$

# Baum-Welch Algorithm

- 关于  $\pi, A, B$ , 一种合理的估计方法如下

$$\bar{\pi}_i = \gamma_1(i)$$

在  $t = 1$  时处在状态  $i$  的期望概率

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

从状态  $i$  到状态  $j$  的转换的期望概率除以从状态  $i$  出发的转换的期望概率

$$\bar{b}_j(k) = \frac{\sum_{t=1}^T \gamma_t(j) \delta(x_t, v_k)}{\sum_{t=1}^T \gamma_t(j)}$$

当  $o_t = v_k$  时 ,  $\delta(o_t, v_k) = 1$   
当  $o_t \neq v_k$  时 ,  $\delta(o_t, v_k) = 0$

在状态  $j$  观察到  $v_k$  的期望概率

处在状态  $j$  的期望概率

# Baum-Welch Algorithm

- 利用上述结论，即可进行模型估算
- 选择模型参数初始值，初始值应满足隐马尔科夫模型的要求，即：

$$\sum_{i=1}^N \pi_i = 1 \quad \sum_{j=1}^N a_{ij} = 1, 1 \leq i \leq N \quad \sum_{k=1}^M b_j(k) = 1, 1 \leq j \leq N$$

- 将初始值代入前面的公式中，计算一组新的参数  $\bar{\pi}, \bar{A}, \bar{B}$
- 再将新的参数代入，再次计算更新的参数。
- 如此反复，直到参数收敛。

# Baum-Welch Algorithm

- Baum-Welch算法是一种EM算法。
- E-step:
  - 计算 $\xi_t(i, j)$ 和 $\gamma_t(i)$
- M-step:
  - 估计模型 $\lambda$
- 终止条件

$$\left| \log(P(O|\lambda_{i+1})) - \log(P(O|\lambda_i)) \right| < \epsilon$$

# Baum-Welch Algorithm

- Baum等人证明要么估算值  $\bar{\lambda}$  和估算前的参数值  $\lambda$  相等，要么估算值  $\bar{\lambda}$  比估算前的参数值  $\lambda$  更好的解释了观察序列  $O$ 。
- 参数最终的收敛点并不一定是一个全局最优值，但一定是一个局部最优值。

L.R.Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech recognition, Proc. IEEE, 77(2): 257-286, 1989

# 隐马尔科夫模型的实现

- 浮点溢出问题
  - 对于韦特比算法，采用取对数的方式
  - 对于Baum-Welch算法，采用放大因子
  - 对于向前算法采用放大因子以及取对数的方式。

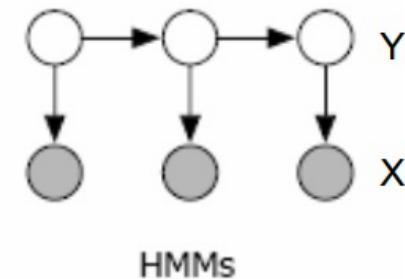


CRF

# 序列模型

- Hidden Markov Model (HMM)

$$p(y, x) = \prod_{t=1}^T p(y_t | y_{t-1}) p(x_t | y_t)$$

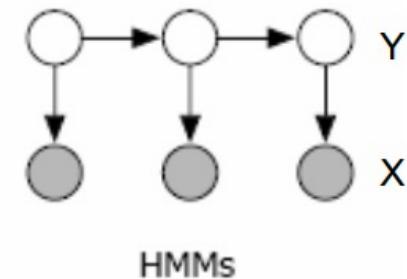


- 独立性假设：
  - 每个状态只直接依赖于其前一个
  - 每个观察变量只依赖于当前状态
- 局限性：
  - 观察变量X之间存在强独立性假设。
  - 建模联合概率  $p(y, x)$  引入大量参数，这需要建模分布  $p(x)$

# 序列模型

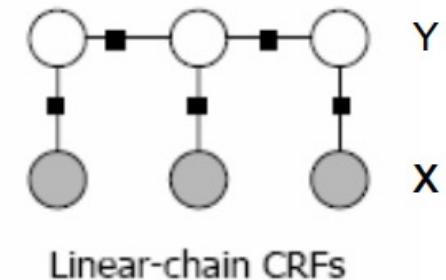
- Hidden Markov Model (HMM)

$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^T p(y_t | y_{t-1}) p(x_t | y_t)$$



- Conditional Random Fields (CRF)

$$p(\mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

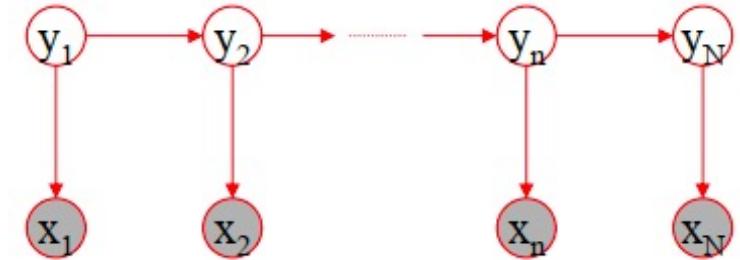


- 条件随机场 (CRF) 的一个重要优势是它们具有很大的灵活性，可以包含各种任意的、非独立的观测特征。

# Generative Model: HMM

- X is observed data sequence to be labeled,
- Y is the random variable over the label sequences
- HMM is a distribution that models  $p(Y, X)$
- Joint distribution is

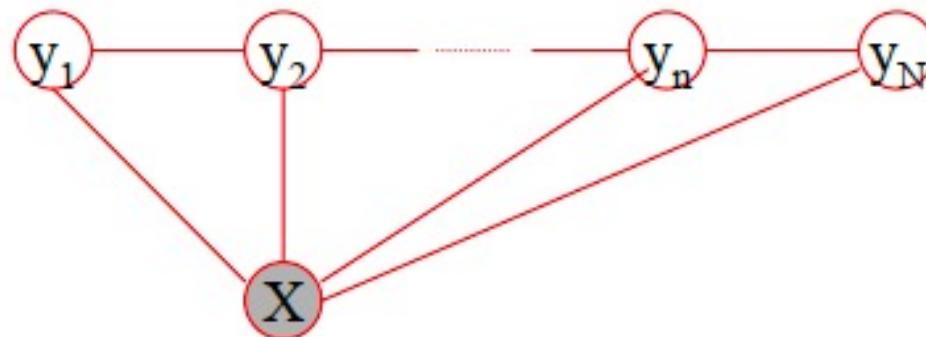
$$p(\mathbf{Y}, \mathbf{X}) = \prod_{n=1}^N p(y_n | y_{n-1}) p(x_n | y_n)$$



- Highly structured network indicates conditional independences,
  - Past states independent of future states
  - Conditional independence of observed given its state.

# 针对序列模型的判别模型

- CRF模型建模了给定观测值 $X$ 条件下的条件分布 $p(Y|X)$
- CRF是一个随机场，全局地以观测值 $X$ 为条件
- 从联合分布 $p(Y,X)$ 中得出的条件分布 $p(Y|X)$ 可以被重写为一个马尔可夫随机场。



# Markov Random Field (MRF)

- 也称为无向图模型
- 变量集 $x$ 的联合分布由一个无向图定义

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C)$$

- 其中 $C$ 是最大团 (clique) (每个节点与每个其他节点相连)
  - $x_C$ 是该团中的变量集,  $\psi_C$ 是潜在函数potential function (或局部函数 local function) 或兼容函数(compatibility function)) , 满足  $\psi_C(x_C) > 0$ , 通常  $\psi_C(x_C) = \exp\{-E(x_C)\}$ , 而 $Z$ 是用于归一化的配分函数。
  - 模型是指一组分布, 而场则指一个具体的分布。

$$Z = \sum_x \prod_C \psi_C(x_C)$$

# MRF with Input-Output Variables

- $X$ 是一组被观测到的输入变量
  - $X$ 的元素用 $x$ 表示
- $Y$ 是一组我们要预测的输出变量
  - $Y$ 的元素用 $y$ 表示
- $A$ 是 $X \cup Y$ 的子集
  - $A$ 中属于 $A \cap X$ 的元素用 $x_A$ 表示
  - $A$ 中属于 $A \cap Y$ 的元素用 $y_A$ 表示
- 那么无向图模型的形式为

$$p(x,y) = \frac{1}{Z} \prod_A \Psi_A(x_A, y_A)$$

where

$$Z = \sum_{x,y} \prod_A \Psi_A(x_A, y_A)$$

# MRF Local Function

- 假设每个局部函数的形式为

$$\Psi_A(x_A, y_A) = \exp \left\{ \sum_m \theta_{Am} f_{Am}(x_A, y_A) \right\}$$

- 其中  $\theta_A$  是一个参数向量，  $f_A$  是特征函数，  $m=1,..M$  是特征下标。

# From HMM to CRF

- HMM 中

$$p(Y, X) = \prod_{n=1}^N p(y_n | y_{n-1}) p(x_n | y_n)$$

- 可以被写作：

分布参数： $\theta = \{\lambda_{ij}, \mu_{oi}\}$

$$p(Y, X) = \frac{1}{Z} \exp \left\{ \sum_n \sum_{i,j \in S} \lambda_{ij} \mathbb{I}_{y_n=i} \mathbb{I}_{y_{n-1}=j} + \sum_n \sum_{i \in S} \sum_{o \in O} \mu_{oi} \mathbb{I}_{y_n=i} \mathbb{I}_{x_n=o} \right\}$$

- 进一步写作：

$$p(Y, X) = \frac{1}{Z} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

特征函数有如下形式： $f_m(y_n, y_{n-1}, x_n)$   
对每个状态转换  $i \rightarrow j$  需要一个特征

$$f_{ij}(y, y', x) = \mathbb{I}(y = i) \mathbb{I}(y' = j)$$

对每个状态-观察对也需要一个特征

$$f_{io}(y, y', x) = \mathbb{I}(y = i) \mathbb{I}(x = o)$$

# From HMM to CRF

$$p(Y, X) = \frac{1}{Z} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- 进一步

$$p(Y|X) = \frac{p(y, x)}{\sum_{y'} p(y', x)} = \frac{\exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}}{\sum_{y'} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y'_n, y'_{n-1}, x_n) \right\}}$$

# CRF definition

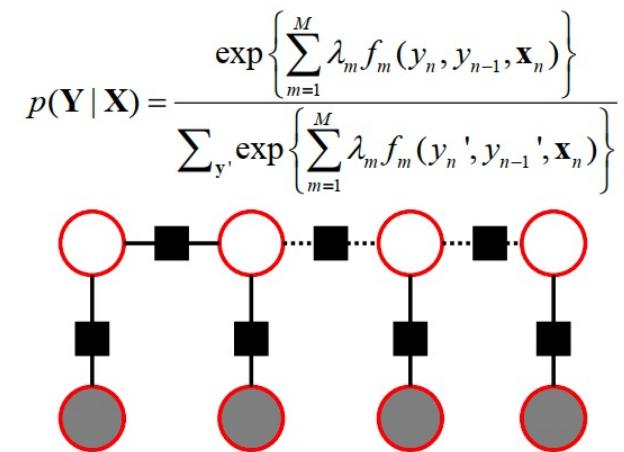
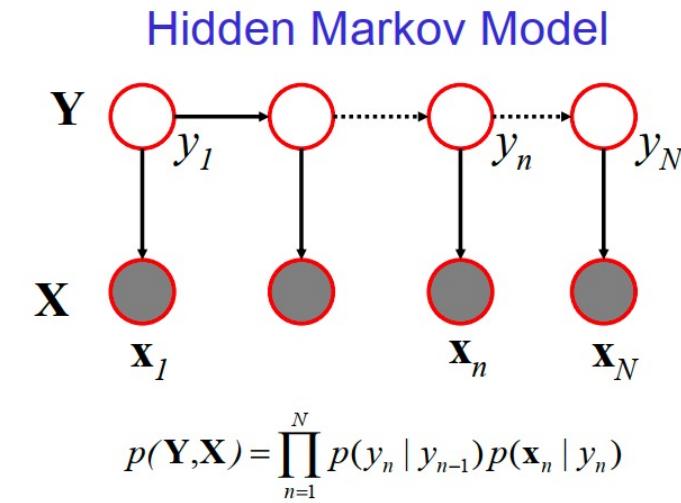
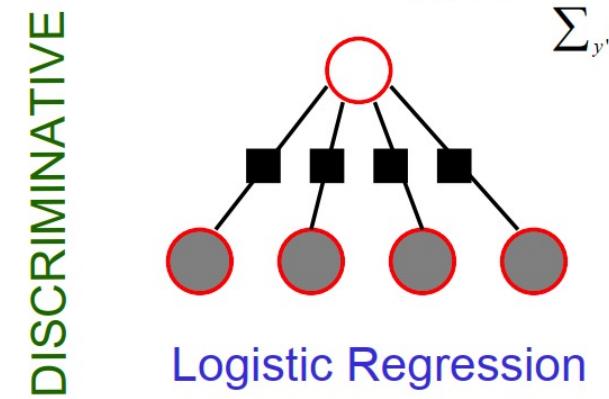
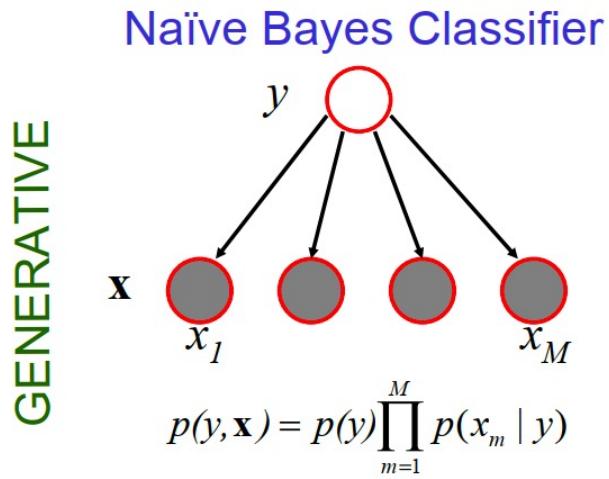
- 线性链CRF定义为分布 $p(Y|X)$ ,其形式如下:

$$p(Y|X) = \frac{1}{Z(X)} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- 其中 $Z(X)$ 是一个实例特定的归一化函数。

$$Z(X) = \sum_{y'} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y'_n, y'_{n-1}, x_n) \right\}$$

# 功能模型



**Conditional Random Field**

# CRF的优势

- CRF放松了对于给定标签的观测数据的条件独立性的假设
- CRF可以包含任意的特征函数
  - 每个特征函数可以使用整个输入数据序列。 观测数据片段的标签概率可能取决于任何过去或未来的数据片段。
- CRF可以避免其他具有偏向后继状态较少的状态的判别性马尔可夫模型的限制

# CRF 优化

$$p(Y|X) = \frac{1}{Z(X)} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- 目标函数（极大似然法）

$$\max_{\lambda_m \in \mathbb{R}^+} \prod_{x,y} p(y|x; \lambda)^{\tilde{P}(x,y)}$$

- Why not set the loss as this?

$$\tilde{P}(x, y) = \tilde{P}(y|x)\tilde{P}(x)$$

$$\max_{\lambda_m \in \mathbb{R}^+} \prod_{x,y} p(y|x; \lambda)^{\tilde{P}(y|x)}$$

# CRF 优化

- 实际依然采用对数值进行优化

$$\min_{\lambda \in \mathbb{R}^+} f(\lambda) = - \sum_{x,y} \tilde{P}(x,y) \log p(y|x; \lambda)$$

- 优化方法：
  - 梯度下降
  - 拟牛顿法
  - L-BFGS

# CRF 解码-- Viterbi

$$p(Y|X) = \frac{1}{Z(X)} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, x_n) \right\}$$

- 韦特比变量  $\delta_t(i)$
- $\delta_t(i)$  的含义是，给定模型  $\lambda$ ，在时刻  $t$  处于状态  $i$ ，观察到  $o_1 o_2 o_3 \dots o_t$  的最佳状态转换序列为  $q_1 q_2 \dots q_t$  的概率。

$$\delta_t = \prod_{i=1}^t \exp \left[ \sum_k \lambda_k f_k(y_{i-1}, y_i, x) \right]$$

- 递推公式：

$$\delta_{t+1} = \delta_t \cdot \exp \left[ \sum_k \lambda_k f_k(y_t, y_{t+1}, x) \right]$$

# CRF 词性标注

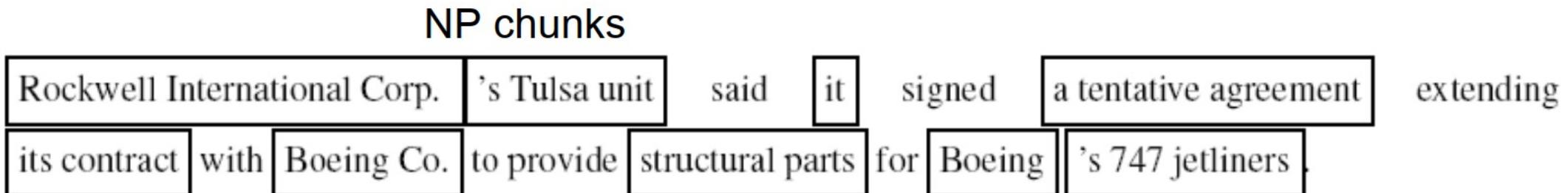
- w = The quick brown fox jumped over the lazy dog
- s = DET VERB ADJ NOUN-S VERB-P PREP DET ADJ NOUN-S
- Baseline is already 90%
  - Tag every word with its most frequent tag
  - Tag unknown words as nouns

Model	Error
HMM	5.69%
CRF	5.55%

# Shallow Parsing

- 完整的parsing或信息提取的前身。
  - 识别文本中各种短语类型的非递归核心。
- 输入：带有POS tag单词的句子
- 任务：为每个单词打上标签，指示单词是否在短语块(chunk)之外(O)，是否开始一个短语块(B)，或者是否继续一个短语块(I)。
- CRF 在标准评估数据集上击败了所有单一模型的 NP 分块结果。

Model	F score
CRF	94.38%
Generalized winnow	93.89%
Voted perceptron	94.09%
MEMM	93.70%





# 词性标注

# 词的分类依据

- 形态标准

*Words that function similarly with respect to the **affixes** they take (their **morphological properties**) are grouped into classes.*

- 分布标准

*Words that function similarly with respect to what can occur **nearby** (their “**syntactic distributional properties**”) are grouped into classes.*

- 意义标准(×)

*While word classes do have tendencies toward **semantic coherence** (nouns do in fact often describe “people, places or things”, and adjectives often describe properties), this is not necessarily the case, and in general we don’t use the **semantic coherence** as a definition criterion for part-of-speech.*

# 英语中词的分类

- 英语词类
  - *preposition, determiner, pronoun, conjunction, nouns, verbs, adjectives, adverbs, numeral, interjection*
- *closed class and open class*
  - *Closed classes are those that have relatively fixed membership, in which new words are rarely coined.*
  - *function word and content word*

# 英语中词的分类

- 词类的子类
  - noun
    - ◆ proper noun *eg. Beijing, IBM*
    - ◆ common noun
      - countable noun *eg. book, table*
      - mass noun *eg. communism, salt*
  - adverb
    - ◆ directional adverb *eg. downhill, home*
    - ◆ degree adverb *eg. somewhat, extremely, very*
    - ◆ manner adverb *eg. slowly, delicately*
    - ◆ temporal adverb *eg. yesterday, tomorrow*

# 汉语中词的分类

- 汉语中词的分类依据
  - 汉语缺乏形态，形态特征不能用作分类依据。
  - 汉语中词分类的依据主要是词的分布特征，或者说主要依据词的语法功能。
  - 同英语一样，汉语中划分词类也不用意义作为分类依据。概念相近的词，语法性质未必相同，例：战争(名词)、战斗(动词)
- 词的语法功能主要指词在句法结构里所能占据的语法位置。
  - 词在句法结构中充当句法成分的能力
  - 词与某类词或某些词组合成短语的能力
- 虽然不能根据意义对词进行分类，但按照分布特征同属一类的词意义上也常有共性。
- 名词通常表示事物的名称、动词通常表示动作和行为、形容词表示事物的性质和状态。

# 汉语中词的分类

- 实词和虚词

- 从功能上看，实词可以充当主语、谓语和宾语。虚词则不可以。
- 从意义上讲，实词有实在的意义，表示事物、动作、行为、变化、性质、状态、处所、时间等。虚词基本只起语法作用，本身多无实在意义。
- 从数量上看，实词多为开放类，虚词多为封闭类。

- 体词和谓词

- 实词通常可进一步分成体词和谓词。体词可以做主语和宾语。谓词主要做谓语。

# 汉语中词的分类

- 体词
  - 名词(1)、处所词(2)、方位词(3)、时间词(4)、区别词(5)、数词(6)、量词(7)、代词(8)
- 谓词
  - 动词(9)、形容词(10)
- 虚词
  - 副词(11)、介词(12)、连词(13)、助词(14)、语气词(15) 拟声词(16)、感叹词(17)

# 汉语中词的分类

- 为什么说一个词是形容词？
  - 可以用作主谓结构中的谓语，但不能带真宾语。
    - 例：长江比黄河长、长三角
  - 可以受“很”这类程度副词修饰。例：很长、很雄伟、很安静
  - 可以作述补结构中的补语。例：洗干净、捆结实
  - 直接或加“地”后作状中结构中的状语。例：迅速提高
  - 直接或加“的”后作定中结构中的定语。例：美丽人生
  - 可以用“a + 不 + a”的形式提问。例：舒服不舒服？
  - .....

# 汉语中词的分类

- 对汉语词类问题有兴趣，可进一步参考有关书籍。
- 由于汉语缺乏形态，词的类别不如英语等西方语言那样易于判别。汉语语言学家曾在汉语词类划分问题上有过不同意见，并经过长期争议，至今仍然存在多种看法。
- 利用计算机处理语言，词语的语法分类及其代码化不可缺少。面向信息处理用汉语词类体系的建立和大规模词语归类实践必须进行。

# 兼类问题

- 如果同一个词具有不同词类的语法功能，则认为这个词兼属不同的词类，简称兼类。

- 例一

- |                      |                      |
|----------------------|----------------------|
| (1a) <b>共同完成一些任务</b> | (1b) 我们的 <b>共同愿望</b> |
| (2a) <b>自动控制这个开关</b> | (2b) 方便的 <b>自动步枪</b> |
| (3a) <b>定期检查机器</b>   | (3b) 一笔 <b>定期存款</b>  |
- 在(a)组中，是副词、在(b)组中是区别词。

- 例二

- |                      |                      |
|----------------------|----------------------|
| (4a) 买了一 <b>束花</b>   | (4b) 花了 <b>很多时间</b>  |
| (5a) 开了一个 <b>会</b>   | (5b) <b>会拉小提琴</b>    |
| (6a) 桌子上有两封 <b>信</b> | (6b) 别 <b>信他的话</b>   |
| (7a) 选举他当 <b>代表</b>  | (7b) 他 <b>代表我们发言</b> |
- 在(a)组中是名词，在(b)组中是动词。

# 兼类问题

- English data, from Brown corpus:
  - 11.5 percent of the lexicon is ambiguous as to part-of-speech (types)
  - 40 percent of the words in the Brown corpus are ambiguous (tokens)
- Degree of ambiguity (No. tags per word)
  - 1 tags 35340
  - 2-7 tags 4100 total: 39440
  - 2 tags 3760
  - 3 tags 264
  - 4 tags 61
  - 5 tags 12
  - 6 tags 2
  - 7 tags 1

# 兼类问题

- 《现代汉语语法信息词典》 数据(1997年版)

■ 总词数	55191	
■ 2-5 tags	1624	2.94%
■ 2 tags	1475	2.67%
■ 3 tags	126	0.23%
■ 4 tags	20	0.04%
■ 5 tags	3	0.01%

- 例：

■ 和	c-n-p-q-v
■ 光	a-d-n-v

# 英语词类标记集

- *Brown corpus tagset*
  - 87 tags
  - Used for Brown Corpus (1-million-word, 1963-1964, Brown University)
  - T<sub>AGGIT</sub> program
- *Penn treebank tagset*
  - 45 tags
  - Used for Penn treebank, Brown Corpus, WSJ Corpus
  - Brill tagger
- *UCREL's C5 tagset*
  - 61 tags
  - Used for British National Corpus (BNC)
  - Lancaster CLAWS tagger

# 英语词类标记集

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction <i>and, but, or</i>		SYM	Symbol	+%, &
CD	Cardinal number	<i>one, two, three</i>	TO	“to”	<i>to</i>
DT	Determiner	<i>a, the</i>	UH	Interjection	<i>ah, oops</i>
EX	Existential ‘there’	<i>there</i>	VB	Verb, base form	<i>eat</i>
FW	Foreign word	<i>mea culpa</i>	VBD	Verb, past tense	<i>ate</i>
IN	Preposition/sub-conj	<i>of, in, by</i>	VBG	Verb, gerund	<i>eating</i>
JJ	Adjective	<i>yellow</i>	VBN	Verb, past participle	<i>eaten</i>
JJR	Adj., comparative	<i>bigger</i>	VBP	Verb, non-3sg pres	<i>eat</i>
JJS	Adj., superlative	<i>wildest</i>	VBZ	Verb, 3sg pres	<i>eats</i>
LS	List item marker	<i>1, 2, One</i>	WDT	Wh-determiner	<i>which, that</i>
MD	Modal	<i>can, should</i>	WP	Wh-pronoun	<i>what, who</i>
NN	Noun, sing. or mass	<i>llama</i>	WP\$	Possessive wh-	<i>whose</i>
NNS	Noun, plural	<i>llamas</i>	WRB	Wh-adverb	<i>how, where</i>
NNP	Proper noun, singular	<i>IBM</i>	\$	Dollar sign	\$
NNPS	Proper noun, plural	<i>Carolinas</i>	#	Pound sign	#
PDT	Predeterminer	<i>all, both</i>	“	Left quote	(‘ or “)
POS	Possessive ending	's	”	Right quote	(’ or ”)
PP	Personal pronoun	<i>I, you, he</i>	(	Left parenthesis	( [ , {, <
PP\$	Possessive pronoun	<i>your, one's</i>	)	Right parenthesis	( ] ), }, > )
RB	Adverb	<i>quickly, never</i>	,	Comma	,
RBR	Adverb, comparative	<i>faster</i>	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	<i>fastest</i>	:	Mid-sentence punc	(: ; ... --)
RP	Particle	<i>up, off</i>			

*Penn treebank  
POS tagset  
(45 tags)*

# 汉语词类标记集

- 北京大学《人民日报》语料库词类标记集
  - 规范2001
    - ◆ 约40个词类标记
    - ◆ 用于标注《人民日报》语料库
    - ◆ 词类标记规范参见
      - [http://www.icl.pku.edu.cn/icl\\_groups/corpus/corporus-annotation.htm](http://www.icl.pku.edu.cn/icl_groups/corpus/corporus-annotation.htm)
  - 规范2003
    - ◆ 扩充至106个词类标记
- 国家语委语用所词类标记集
  - ??个词类标记
  - 参见语委用用所《信息处理用现代汉语词类及词性标记集规范》
- 其它词类标记集

# 汉语词类标记集

标记	描述	标记	描述
Ag	形语素	ns	地名
a	形容词	nt	机构团体
ad	副形词	nz	其他专名
an	名形词	o	拟声词
b	区别词	p	介词
c	连词	q	量词
Dg	副语素	r	代词
d	副词	s	处所词
e	叹词	Tg	时语素
f	方位词	t	时间词
g	语素	u	助词
h	前接成分	Vg	动语素
i	成语	v	动词
j	简称略语	vd	副动词
k	后接成分	vn	名动词
l	习用语	w	标点符号
m	数词	x	非语素字
Ng	名语素	y	语气词
n	名词	z	状态词
nr	人名		

北大《人民日报》标注语料库词类标记集(40+ tags)

为了处理真实语料，汉语词类标记集中通常包含一些非功能分类的标记，例如：成语、习用语、简称略语；也包含一些语素、前接成份、后接成份等比词小的标记。

# 词类自动标注

- 词类自动标注的任务
  - 判定自然语言句子中的每个词的词类并给每个词赋以词类标记。
  - 例如：
    - ◆ book that flight.
    - book/VB that/DT flight/NN ./.
    - ◆ 这份特区政府的报告长达 20 页。
    - 这/r 份/q 特区/n 政府/n 的/u 报告/n 长/a 达/v 20/m 页/q 。/w
  - 对于兼类词，词类标注程序应根据上下文确定 兼类词在句子中最合适的词类标记。(难点所在)

# 词类自动标注

- 词类自动标注是深层语言分析的基础
  - 句法分析
- 词类标注程序判定依据
  - 要标注的词的不同词类的分布
    - can MD-VB-NN (大部分情形是MD)
    - dumb tagger 统计每个兼类词的词类概率分布，并给每个词赋概率最大的词类。
      - 对英语而言，试验结果 90%
  - 上下文中其它词的词类信息
    - 英语中，词类串DT JJ NN比DT JJ VBZ更加可能。

# 词类自动标注

- 基本方法
  - 基于规则的词类标注
  - 基于统计的词类标注
  - 统计规则相结合的词类标注

# 基于规则的词类标注方法

- 早期的词类标注方法多为基于规则的方法
  - 70年代初 TAGGIT 标注程序
  - 约 3300 条人工总结的规则
  - 标注Brown语料库（87 tags），准确率约77%
- 目前的基于规则的词类标注程序性能远远好于TAGGIT标注程序
- 基于规则的词类标注程序工作过程
  1. 查词典，给句中各词标记所有可能的词类标记。
  2. 应用规则，逐步删除错误的标记，最终只留下正确的标记。

# 基于规则的词类标注

- 以 EngCG tagger (1995) 为例
- 查词典，给句中各词标记可能的词类标记以及特征信息(形态特征、次范畴化框架特征等)

Pavlov had shown that salivation ...



Pavlov	<b>PAVLOV N NOM SG PROPER</b>
had	<b>HAVE V PAST VFIN SVO</b>
	HAVE PCP2 SVO
shown	<b>SHOW PCP2 SVOO SVO SV</b>
that	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	<b>CS</b>
salivation	<b>N NOM SG</b>
	...

# 基于规则的词类标注

- 规则(constraint)示例

ADVERBIAL-THE THAT RULE

**Given input:** "that"

**if**

(+1 A/ADV/QUANT); /\* if next word is adj, adverb, or quantifier \*/  
(+2 SENT-LIM); /\* and following which is a sentence boundary, \*/  
(NOT -1 SVOC/A); /\* and the previous word is not a verb like \*/  
/\* 'consider' which allows adjs as object complements \*/

**then** eliminate non-ADV tags

**else** eliminate ADV tag

- 规则用以删除和上下文环境不相容的标记。

- *it isn't that odd.*
- *I consider that odd.*

# 基于隐马尔科夫模型的词类标注

- HMM状态集                  词类标记集
- HMM输出符号集    词表
- 如何根据观察到的词串(句子), 求解最可能的词类标记序列(状态转换序列)。                  维特比算法
- 模型参数
  - $p(t_i | t_{i-1})$                   词类转移概率
  - $p(w_i | t_i)$                   词类 $t_i$ 生成词 $w_i$ 的概率
  - $p(t)$                   词类 $t$ 出现在句首的概率

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n | w_1^n) = \operatorname{argmax}_{t_1^n} \prod_{i=1}^n p(w_i | t_i) p(t_i | t_{i-1})$$

# 基于隐马尔科夫模型的词类标注

- 参数学习
  - 多采用有指导的学习方法
  - 需要预先准备带词类标记的语料库
    - ◆ 例如，1998年1月《人民日报》标注语料库
  - 也可以采用无指导学习，例如用Baum-Welch算法
- 最大似然估计

$$p(t_i \mid t_{i-1}) = \frac{c(t_{i-1}, t_i)}{c(t_{i-1})}$$

特殊标记 $\langle BOS \rangle$   $\langle EOS \rangle$

$$p(w_i \mid t_i) = \frac{c(t_i, w_i)}{c(t_i)}$$

# 汉语词类标注实例

- 1998年1月《人民日报》标注语料
- 作为动词的“报告”(30次)
  - 1...5岁3岁的福塞特向总部报告说，负责热气球...
  - 2...将刘青山、张子善的严重犯罪事实报告党中央，...
  - 3...有关矿产资源情况，要每周向中央主要领导报告。
- 作为名词的“报告”(200次)
  - 1...在党的十五大报告中，江主席再次郑重地...
  - 2...报告认为，虽然日本政府为减少限制性贸易...
  - 3...国际金融协会发表资金流动报告...
- ?发生交通事故时，当事人应当迅速报告公安机关，听候处理...

# 汉语词类标注实例

$c(t_i, t_{i+l})$	...	a	ad	an	n	v	vn	...	$\Sigma$
...	...		...	...	...	...	...	...	...
a		800	8	127	10923	942	2267		34473
ad	...	76	34	0	3	5533	2	...	5933
an	...	10	5	47	238	257	218	...	2837
n	...	4047	1273	440	42491	32933	12508	...	312263
v	...	6924	855	735	42671	27142	4735	...	229776
vn	...	284	113	54	16021	2677	3165	...	42734
...	...	...	...	...	...	...	...	...	...

$c(w_i, t_i)$	...	a	ad	an	n	v	vn	...	$\Sigma$
...	...	...	...	...	...	...	...	...	...
当事人	...	0	0	0	25	0	0	...	25
应当	...	0	0	0	0	340	0	...	340
迅速	...	50	116	1	0	0	0	...	167
报告	...	0	0	0	200	30	4	...	234
公安	...	0	0	0	188	0	0	...	188
机关	...	0	0	0	354	0	0	...	354
...	...	...	...	...	...	...	...	...	...

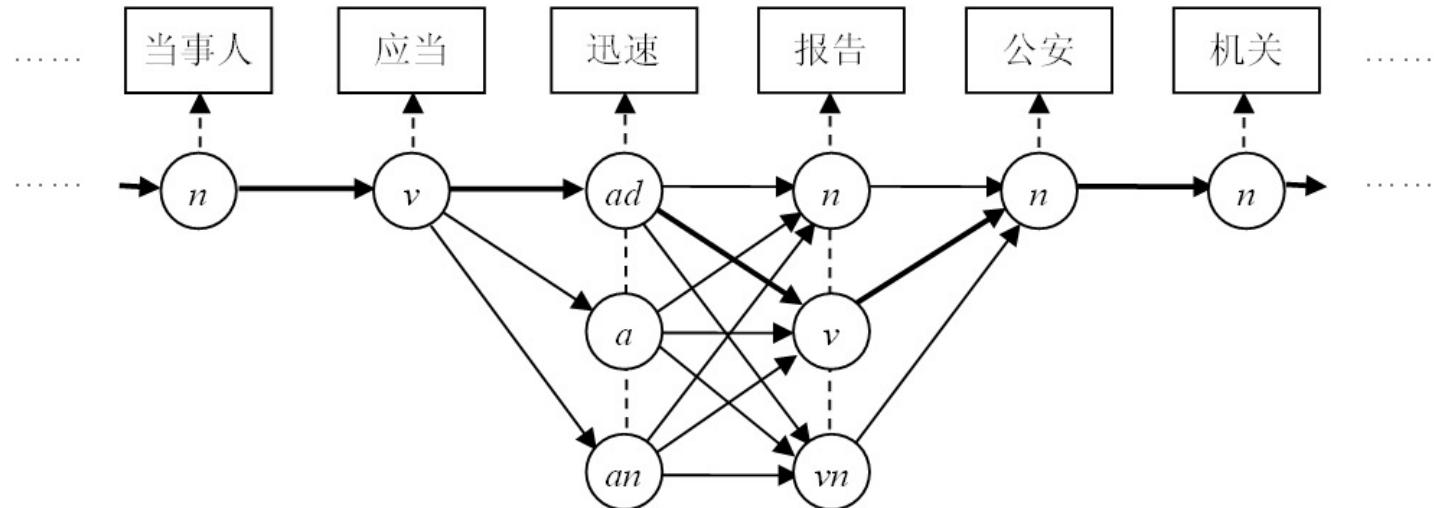
# 汉语词类标注实例

$p(t_{i+1} t_i)$	...	a	ad	an	n	v	Vn	...	$\Sigma$
...	...		...	...	...	...	...	...	...
a		0.0232065676	0.0002320657	0.0036840426	0.3168566704	0.0273257332	0.0657616105		
ad	...	0.0128097084	0.0057306590	0	0.0005056464	0.9325804820	0.0003370976	...	1
an	...	0.0035248502	0.0017624251	0.0165667959	0.0838914346	0.0905886500	0.0768417342	...	1
n	...	0.0129602290	0.0040766918	0.0014090686	0.1360743988	0.1054655851	0.0400559785	...	1
v	...	0.0301336954	0.0037210152	0.0031987675	0.1857069494	0.1181237379	0.0206070260	...	1
Vn	...	0.0066457622	0.0026442645	0.0012636308	0.3749005476	0.0626433285	0.0740628071	...	1
...	...		...		...	...	...	...	...

# 汉语词类标注实例

$p(w_i t_i)$	...	a	ad	an	n	v	vn	...
...	...	...	...	...	...	...	...	...
当事人	...	0	0	0	0.0000800607	0	0	...
应当	...	0	0	0	0	0.0014797019	0	...
迅速	...	0.0014504105	0.0195516602	0.0003524850	0	0	0	...
报告	...	0	0	0	0.0006404857	0.0001305619	0.0000936023	...
公安	...	0	0	0	0.0006020566	0	0	...
机关	...	0	0	0	0.0011336598	0	0	...
...	...	...	...	...	...	...	...	...
$\Sigma$	...	1	1	1	1	1	1	...

# 汉语词类标注实例



$P(\dots \text{n} \text{ v} \text{ ad} \text{ n} \text{ n} \text{ n} \dots | \dots \text{ 当事人} \text{ 应当} \text{ 迅速} \text{ 报告} \text{ 公安} \text{ 机关} \dots)$

$$\begin{aligned} &= \dots \times p(\text{当事人}|n) \times p(v|n) \times p(\text{应当}|v) \times p(ad|v) \times p(\text{迅速}|ad) \times p(n|ad) \\ &\quad \times p(\text{报告}|n) \times p(n|n) \times p(\text{公安}|n) \times p(n|n) \times p(\text{机关}|n) \times \dots \end{aligned}$$

# 汉语词类标注实例

$T$	$\prod p(w_i t_i) p(t_i t_{i-1})$	$P(T \dots)$ 应当 迅速 报告 公安 ...)
...v ad n n...	$\dots p(ad v) \times p(\text{迅速} ad) \times p(n ad) \times p(\text{报告} n) \times p(n n) \dots$	3.2061059e-12
...v ad v n...	$\dots p(\mathbf{ad v}) \times p(\text{迅速} \mathbf{ad}) \times p(\mathbf{v ad}) \times p(\text{报告} v) \times p(n v) \dots$	<b>1.64503834e-9</b>
...v ad vn n...	$\dots p(ad v) \times p(\text{迅速} ad) \times p(vn ad) \times p(\text{报告} vn) \times p(n vn) \dots$	8.6060396e-13
...v a n n...	$\dots p(a v) \times p(\text{迅速} a) \times p(n a) \times p(\text{报告} n) \times p(n n) \dots$	1.20695769e-9
...v a v n...	$\dots p(a v) \times p(\text{迅速} a) \times p(v a) \times p(\text{报告} v) \times p(n v) \dots$	2.8957414e-11
...v a vn n...	$\dots p(a v) \times p(\text{迅速} a) \times p(vn a) \times p(\text{报告} vn) \times p(n vn) \dots$	1.0085986e-10
...v an n n...	$\dots p(an v) \times p(\text{迅速} an) \times p(n an) \times p(\text{报告} n) \times p(n n) \dots$	8.2437876e-12
...v an v n...	$\dots p(an v) \times p(\text{迅速} an) \times p(v an) \times p(\text{报告} v) \times p(n v) \dots$	2.4765193e-12
...v an vn n...	$\dots p(an v) \times p(\text{迅速} an) \times p(vn an) \times p(\text{报告} vn) \times p(n vn) \dots$	3.0403464e-12

... 当事人/n 应当/v 迅速/ad 报告/v 公安/n 机关/n ...

# 改进基于HMM的词类标注

- 暗含两个假设：
  - (1) 句中某个词是否出现只和该词的词类标记有关。和句中的其他词以及其它词的词类标记无关。
  - (2) 句中某个词的词类只和该词前面一个词的词类有关。而和句中其它词类无关。(词类的bigram模型)
- 可以扩充基于隐马尔科夫模型的词类标注模型，考虑更多的上下文，把词类的bigram模型改作trigram模型。

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n \mid w_1^n) = \operatorname{argmax}_{t_1^n} \prod_{i=1}^n p(w_i \mid t_i) p(t_i \mid t_{i-1}, t_{i-2})$$

# 改进基于HMM的词类标注

- 最大似然估计

$$p(t_i|t_{i-1}, t_{i-2}) = \frac{c(t_{i-2}, t_{i-1}, t_i)}{c(t_{i-2}, t_{i-1})}$$

- 数据稀疏问题、应用平滑技术(线形插值)

$$\hat{p}(t_i|t_{i-1}, t_{i-2}) = \lambda_1 p(t_i|t_{i-1}, t_{i-2}) + \lambda_2 p(t_i|t_{i-1}) + \lambda_3 p(t_i)$$

- 输出概率平滑

$$p(w|t) = \frac{c(t, w) + 1}{c(t) + T_w}$$

# 基于转换的词类标注

- Eric Brill提出(1995)
- 特点(兼具规则和统计两个方面的特性)
  - 应用规则进行标注，规则称为转换。
  - 规则不是人工总结，而是应用机器学习的办法学习得到。使用的机器学方法通常称作基于转换的学习 (Transformation-Based Learning or TBL)。

Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, Computational Linguistics, Vol. 21, No. 4, 1995, pp.543-565

# 基于转换的词类标注

- 什么是一个转换(transformation)?
  - 激发环境(triggering environment)
    - 描述了应用该转换需要满足的条件
  - 重写规则(rewriting rule)
    - 描述了应用规则所要进行的动作
    - 重写规则形如  $t_1 \rightarrow t_2$ , 含义是把词类标记 $t_1$ 改作 $t_2$
    - 注意重写规则与一般意义上的重写规则的区别
- 转换举例

 $if t_{-1} = \text{TO} \text{ then } \text{NN} \rightarrow \text{VB}$

含义 : Change NN to VB when the previous tag is TO

# 基于转换的词类标注

- 转换的应用
  - *race* NN VB
    - ... is expected to **race** tomorrow ...
    - ...the **race** for outer space ...
  - 初标注结果
    - ... is/VBZ expected/VBN to/TO race/NN tomorrow/NN ...
    - ... the/DT race/NN for/IN outer/JJ space/NN ...
  - 应用转换规则
    - ... is/VBZ expected/VBN to/TO race/VB tomorrow/NN ...
    - ... the/DT race/NN for/IN outer/JJ space/NN ...
  - 转换规则可以视为一种纠错规则
    - ◆ 在转换规则使用前，待标注的句子已经进行过初步标注，转换规则负责改正其中的错误标注

# 基于转换的词类标注

激发环境：当前词前面一个词的词类是副形词(ad)

重写规则：把当前词的词类从名词(n)改作动词(v)

... 当事人/n 应当/v 迅速/ad 报告/n 公安/n 机关/n ...

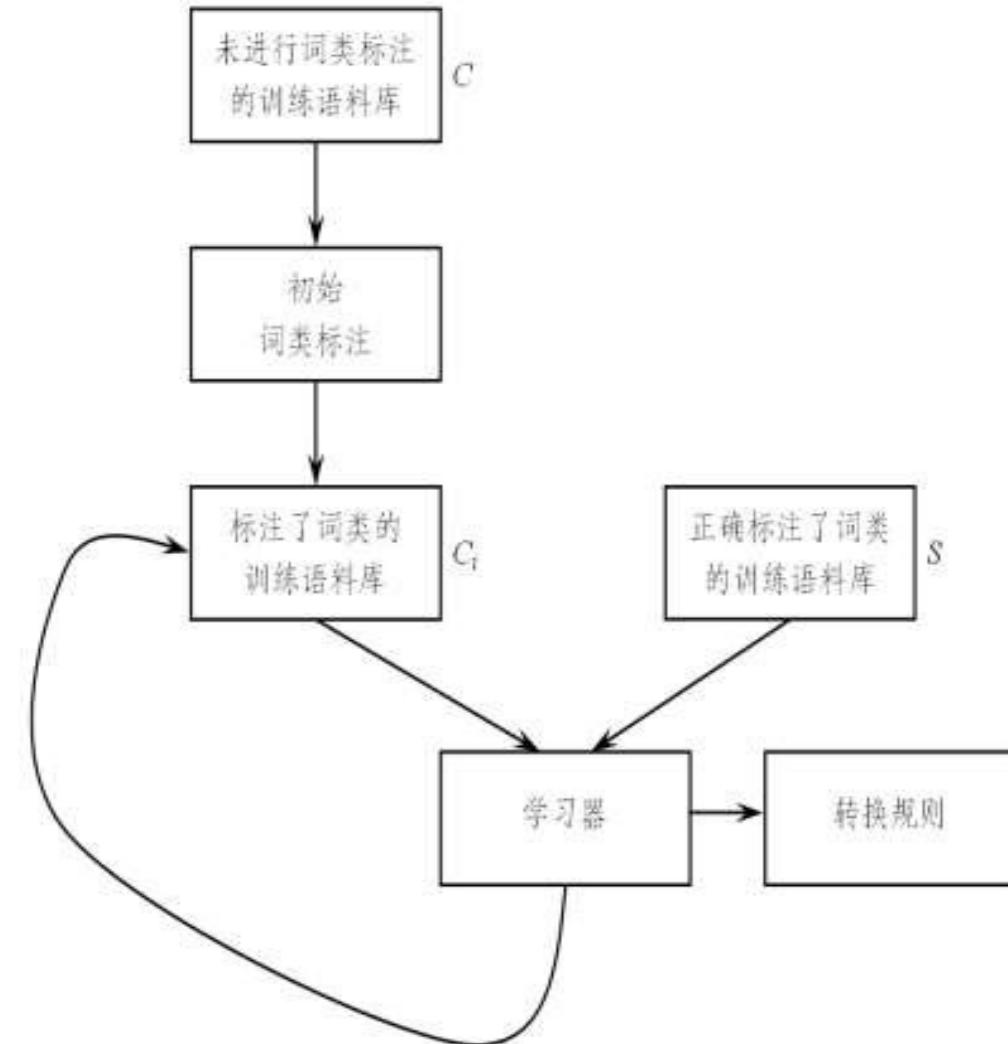
... 当事人/n 应当/v 迅速/ad 报告/v 公安/n 机关/n ...

... 发动/v 全区/n 企事业/n 单位/n 积极/ad. 文化/n 扶贫/vn ...

... 发动/v 全区/n 企事业/n 单位/n 积极/ad. 文化/v 扶贫/vn ...

# 基于转换的词类标注

- 学习转换的基本思想
  - 准备未加标注的训练语料  $C$
  - 对训练语料进行初标注，形成语料  $C_0$
  - 对  $C_0$  进行人工校对，形成正确标注的语料库  $S$
  - 将  $C_0$  与  $S$  进行对比，学习转换规则
  - 评价学到的转换规则，选择能最大限度地降低  $C_0$  错误率的规则  $\tau$
  - 对  $C_0$  应用转换  $\tau$ ，产生语料  $C_1$
  - 对比  $C_1$  与  $S$ ，按照上述过程继续学习、应用转换规则，直到错误率不再有明显降低为止



# 基于转换的词类标注

**PROCEDURE**  $TBLearner( S, T )$  **begin**

$C \leftarrow$  删除  $S$  中的词类信息形成的未标注词类的语料库;

$C_0 \leftarrow$  基于初始标注程序对  $C$  进行标注形成的标注语料库;

**for**  $k \leftarrow 0$  **step 1 do**

$\tau \leftarrow$  可使  $E(u_i(C_k))$  取最小值的转换  $u_i$ ;

**if** ( $E(C_k) - E(\tau(C_k)) < \varepsilon$ ) **then break**

$C_{k+1} \leftarrow \tau(C_k)$ ;

$T_{k+1} \leftarrow \tau$ ;

**end**

**end.**

# 基于转换的词类标注

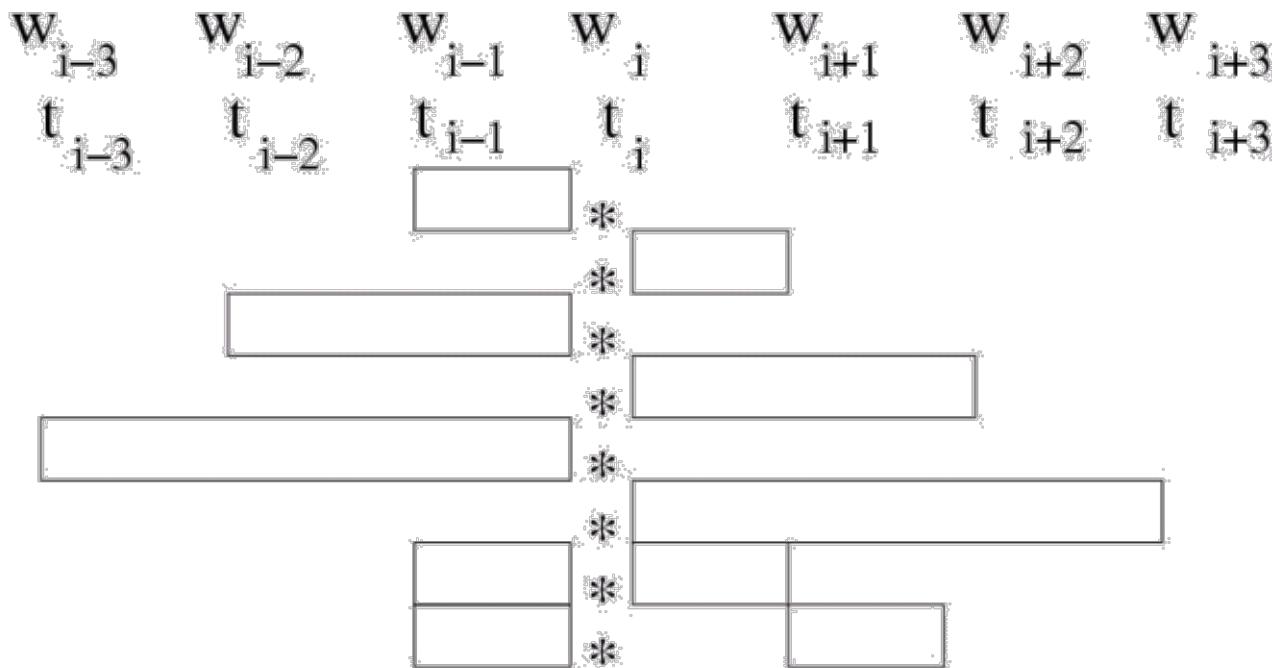
- 初标注器的选择
  - 学习到的转换规则和初标注器有关，选择不同的初标注器学习到的转换规则不同
  - 用dumb tagger进行初始标注
    - 用基于规则的词类标注器进行标注
    - 用基于隐马模型的词类标注器进行标注
    - ...
  - 用学到的规则进行词类标注时，应保证和学习规则时用相同初标注器。
- 转换规则的排列顺序是有意义的
  - 先学到的转换规则先使用，后学到的规则后使用，后学到的规则的作用对象是先学到的规则的处理结果
  - 先学到的规则效果明显、后学到的规则对错误率的改进较小
  - 规则的使用过程类似于创作油画

# 基于转换的词类标注

- 激发环境的选择
  - 激发环境的选择确定了利用的上下文知识的多少
    - ◆ 前文例子中，激发环境仅考虑了标注词前一个词的词类信息(即  $t_{-1} = \text{TO}$ )
  - 理论上，利用的上下文知识越多性能越好
  - 对激发环境不加限制，导致学习效率严重下降，需进行权衡
  - Brill使用激发环境模板来限制可以使用的环境

Change tags				
#	From	To	Condition	Example
1	NN	VB	Previous tag is TO	to/TO race/NN → VB
2	VBP	VB	One of the previous 3 tags is MD	might/MD vanish/VBP → VB
3	NN	VB	One of the previous 2 tags is MD	might/MD not reply/NN → VB
4	VB	NN	One of the previous 2 tags is DT	
5	VBD	VBN	One of the previous 3 tags is VBZ	

# 基于转换的词类标注



*Brill Tagger* 中使用的激发环境模板

# 未登录词

- 未登录词
  - 视作兼类词，可能是任何一个词类，均匀分布
  - 依照出现一次的词(hapax legomenon)的规律处理
    - 更可能是名词 不大可能是限定词等
    - 将出现一次的词的分布平均作为未登录词的分布
  - 对于英文等语言可以利用形态特性(词缀)、拼写特性判定(首字母大小写)
  - 未登录词的词性标注是难点

# 其他方法

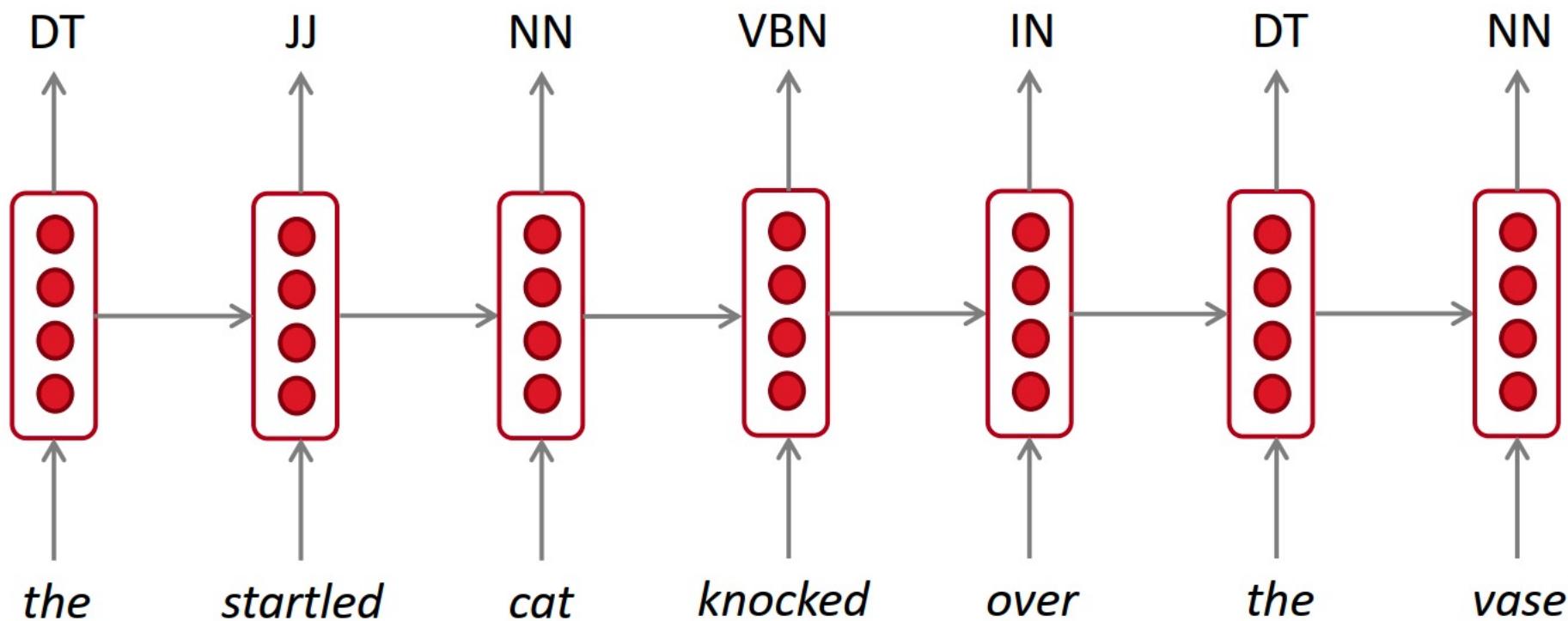
• 在上世纪90年代，词类标注问题得到了持续关注，不断有新的方法和模型提出，除我们介绍的方法外，下列方法也取得了较好的标注效果：

- 基于决策树(Schmid 1994)
- 基于神经网络(Benello et al. 1989)
- 基于最大熵原则(Ratnaparkhi 1996)

1. Schmid,H., Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. 1994.
2. Benello, J., et al. Syntactic category disambiguation with neural networks. Computer Speech and Language, 3, 1989.
3. Ratnaparkhi, A., A Maximum Entropy Part of Speech Tagger. In conference of Empirical Methods in Natural Language Processing, University of Pennsylvania, 1996.

# 现代方法

- POS tag, NER



**Thank you!**