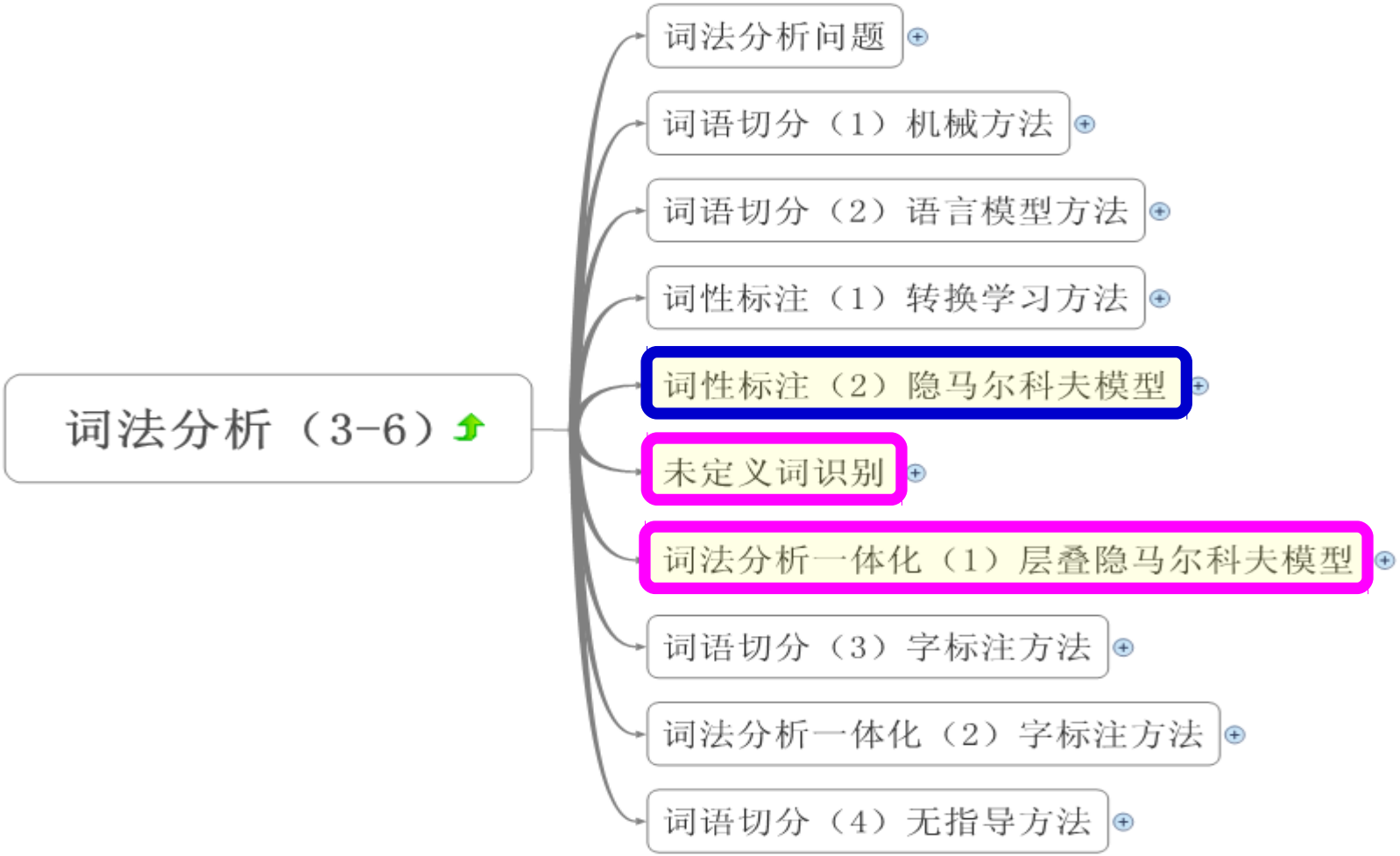
计算语言学

第 5 讲　词法分析（三）

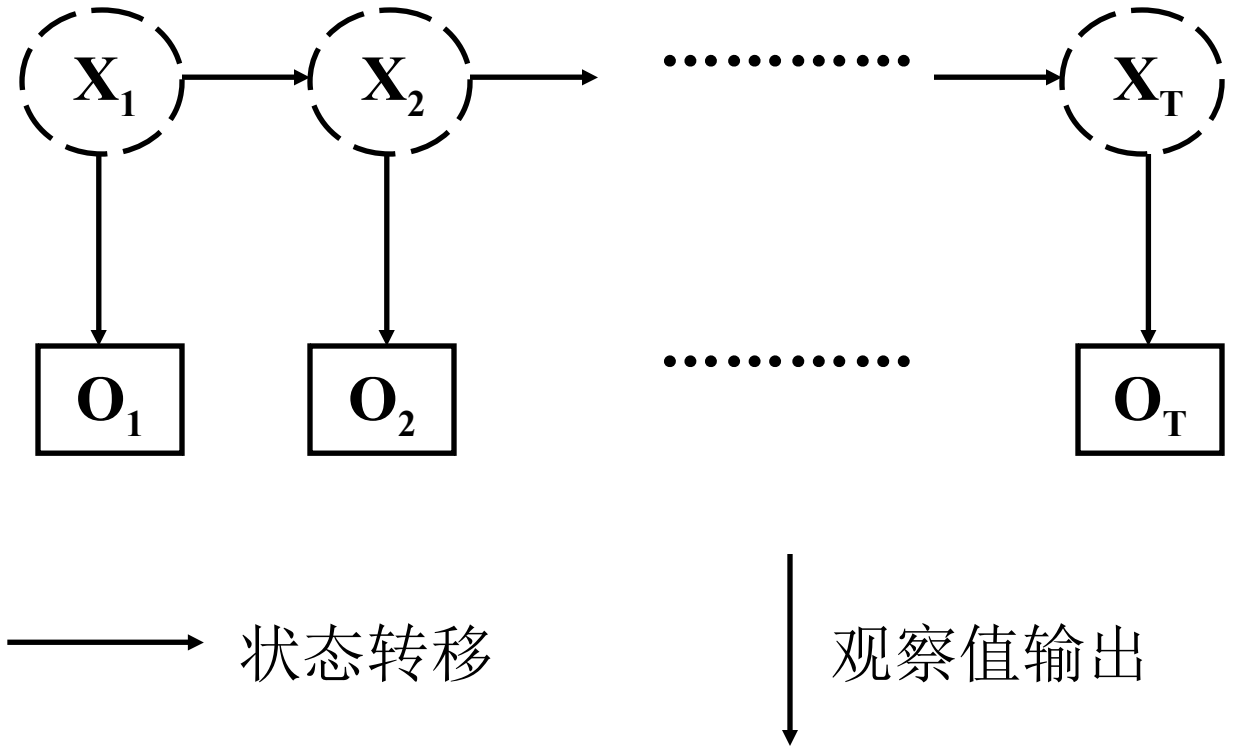
刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院 2012 年春季课程讲义

内容提要


词法分析（3-6）
词法分析问题
词语切分（1）机械方法
词语切分（2）语言模型方法
词性标注（1）转换学习方法
词性标注（2）隐马尔科夫模型
未定义词识别
词法分析一体化（1）层叠隐马尔科夫模型
词语切分（3）字标注方法
词法分析一体化（2）字标注方法
词语切分（4）无指导方法

隐马尔科夫模型—图示



隐马尔科夫模型—假设

对于一个随机事件，有一个观察值序列： $O_1,\ldots,O_T$

该事件隐含着一个状态序列： $X_1,\ldots,X_T$

假设 1 ：马尔可夫假设（状态构成一阶马尔可夫链）

$$p(X_i|X_{i-1}\ldots X_1) = p(X_i|X_{i-1})$$

假设 2 ：不动性假设（状态与具体时间无关）

$p(X_{i+1}|X_i) = p(X_{j+1}|X_j)$ ，对任意 i,j 成立

假设 3 ：输出独立性假设（输出仅与当前状态有关）

$$p(O_1,\ldots,O_T \mid X_1,\ldots,X_T) = \Pi\, p(O_t \mid X_t)$$

隐马尔科夫模型—定义

一个隐马尔可夫模型 (HMM) 是一个五元组：

$$(\Omega_X , \Omega_O, A, B, \pi)$$

其中：

$\Omega_X = \{q_1,...q_N\}$：状态的有限集合

$\Omega_O = \{v_1,...,v_M\}$：观察值的有限集合

$A = \{a_{ij}\}$， $a_{ij} = p(X_{t+1} = q_j \mid X_t = q_i)$：转移概率

$B = \{b_{ik}\}$， $b_{ik} = p(O_t = v_k \mid X_t = q_i)$：输出概率

$\pi = \{\pi_i\}$， $\pi_i = p(X_1 = q_i)$：初始状态分布

隐马尔科夫模型—例子

- 假设：某一时刻只有一种疾病，且只依赖于上一时刻疾病
  一种疾病只有一种症状，且只依赖于当时的疾病
- 症状 (观察值)：发烧，咳嗽，咽喉肿痛，流涕
- 疾病 (状态值)：感冒，肺炎，扁桃体炎
- 转移概率：从一种疾病转变到另一种疾病的概率
- 输出概率：某一疾病呈现出某一症状的概率
- 初始分布：初始疾病的概率
- 解码问题：某人症状为：咳嗽→咽喉痛→流涕→发烧
  请问：其疾病转化的最大可能性如何？

隐马尔科夫模型—例子（续）

- 转移概率

	感冒	肺炎	扁桃体炎
感冒	0.4	0.3	0.3
肺炎	0.2	0.6	0.2
扁桃体炎	0.1	0.1	0.8

- 输出概率

	发烧	咳嗽	咽喉痛	流涕
感冒	0.4	0.3	0.1	0.2
肺炎	0.3	0.5	0.1	0.1
扁桃体炎	0.2	0.1	0.6	0.1

- 初始分布

感冒	肺炎	扁桃体炎
0.5	0.2	0.3

隐马尔科夫模型—问题

令 $\lambda = \{A,B,\pi\}$ 为给定 HMM 的参数，

令 $\sigma = O_1,\ldots,O_T$ 为观察值序列，

隐马尔可夫模型（ HMM ）的三个基本问题：

1. 评估问题：对于给定模型，求某个观察值序列的概率 $p(\sigma|\lambda)$ ；（语言模型）
2. 解码问题：对于给定模型和观察值序列，求可能性最大的状态序列；
3. 学习问题：对于给定的一个观察值序列，调整参数 λ，使得观察值出现的概率 $p(\sigma|\lambda)$ 最大。

隐马尔科夫模型—算法

- 评估问题：向前算法
  - 定义向前变量
  - 采用动态规划算法，复杂度 $O(N^2T)$
- 解码问题：韦特比（ Viterbi ）算法
  - 采用动态规划算法，复杂度 $O(N^2T)$
- 学习问题：向前向后算法
  - EM 算法

HMM 评估问题

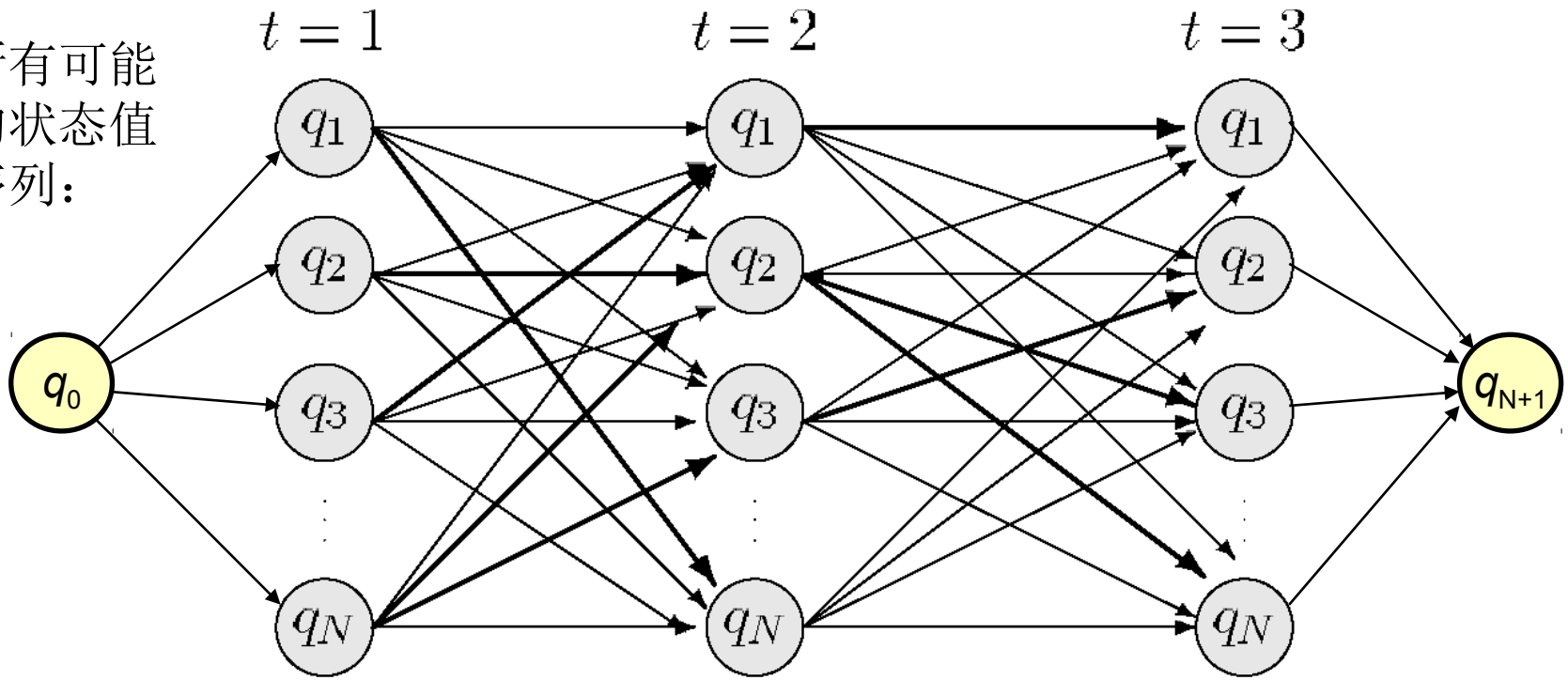
评估问题：对于给定模型，求某个观察值序列的概率 p(σ|λ) ；（语言模型）

$$
\begin{aligned}
P(O|\lambda) &= \sum_X P(O, X|\lambda) \\
&= \sum_X P(X|\lambda) P(O|X, \lambda) \\
&= \sum_X (\pi_{X_1} \prod_{i=2}^{T} a_{X_{i-1}X_i}) (\prod_{i=1}^{T} b_{X_i O_i}) \\
&= \sum_X (\pi_{X_1} b_{X_1 O_1} \prod_{i=2}^{T} a_{X_{i-1}X_i} b_{X_i O_i})
\end{aligned}
$$

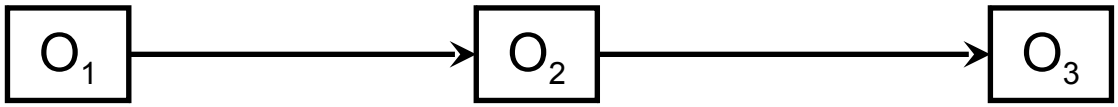
可能的状态序列有 N^T 种可能性，计算复杂度极高

HMM 评估问题

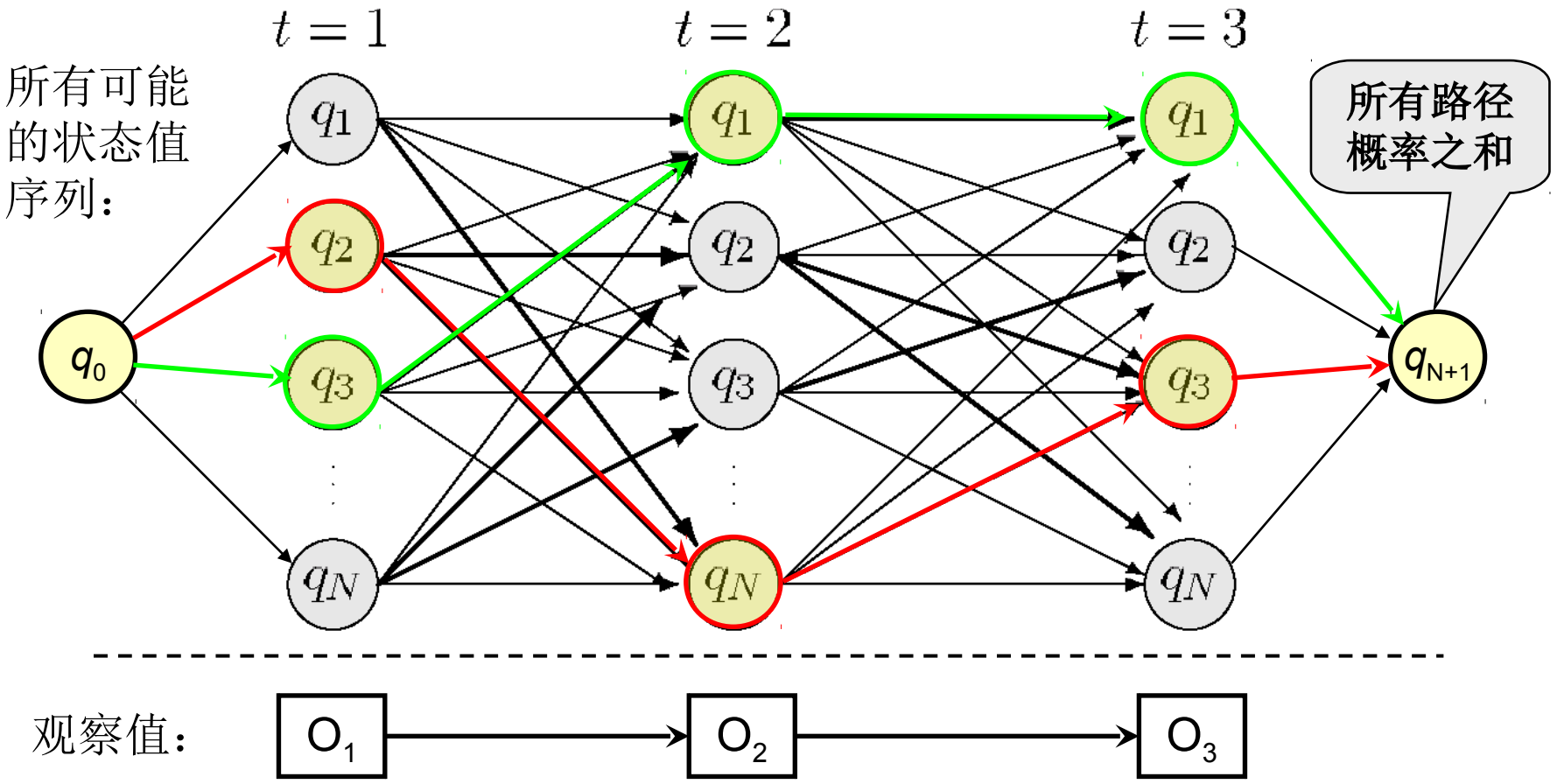
所有可能的状态值序列：



观察值：



HMM 评估问题



可能的状态序列有 N^T 种

HMM 评估问题—向前算法 (1)

定义前向变量为 HMM 在时间 t 输出序列 $O_1 \ldots O_t$，并且位于状态 X_t 的概率：

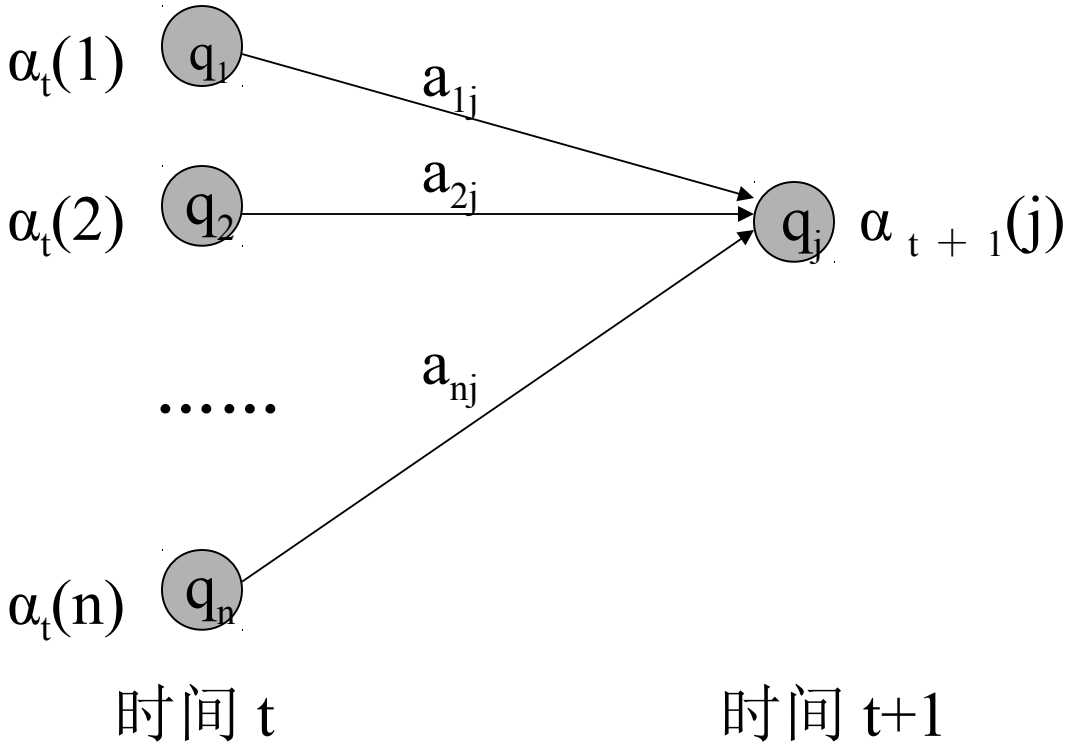
$$\alpha_t(i) = P(O_1 \ldots O_t, X_t = q_i | \lambda)$$

初始化：
$$\alpha_1(i) = \pi_i b_{iO_1}$$

迭代公式为：
$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_{jO_{t+1}}$$

最终结果为：
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

HMM 评估问题—向前算法 (2)



向前算法的时间复杂度： $O(N^2T)$

HMM 评估问题—向前算法 (3)

- **Forward Algorithm**
  - **Assign p(source_state)=1**
  - **For each observation *o* from source to destination**
    - **For each possible state *n* of observation *o***
      - **p(*n*)=0, previous_state(*n*)=ø**
      - **For each edge *e* directed to *n* from *n*'**
        - » **p'(*n*)=p(*n*')×transition_probability(*n*'|*n*)×output_probability(*o*|*n*)**
        - » **p(*n*)=p(*n*)+p'(*n*), previous_state(*n*)= *n*'**
  - **Return p(destination_state)**

HMM 评估问题—向后算法 (1)

定义后向变量为 HMM 在时间 t 并且位于状态 X_t 的情况下，输出序列 $O_{t+1}\ldots O_T$，：

$$\beta_t(i)=P(O_{t+1}\ldots O_T, X_t=q_i|\lambda)$$

初始化：

$$\beta_T(i)=1$$

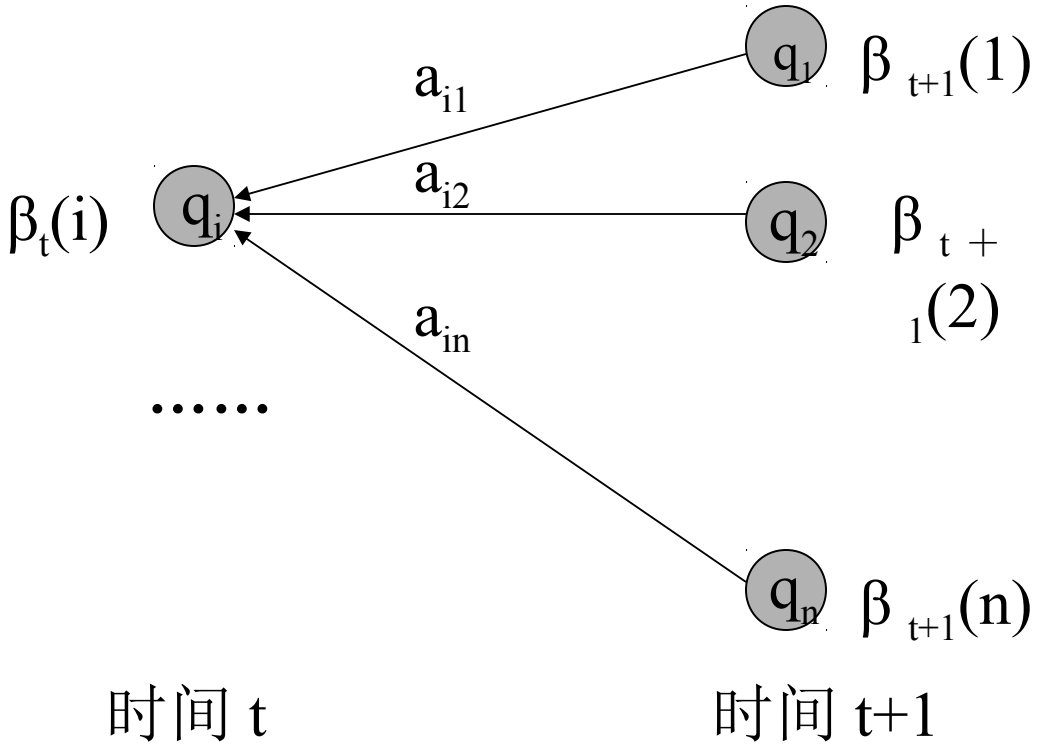
迭代公式为：

$$\beta_t(i)=\sum_{j=1}^{N}[a_{ij}b_{jO_{t+1}}\beta_{t+1}(j)]$$

最终结果为：

$$P(O|\lambda)=\sum_{i=1}^{N}\pi_i b_{iO_1}\beta_1(i)$$

HMM 评估问题—向后算法 (2)



向后算法的时间复杂度： $O(N^2T)$

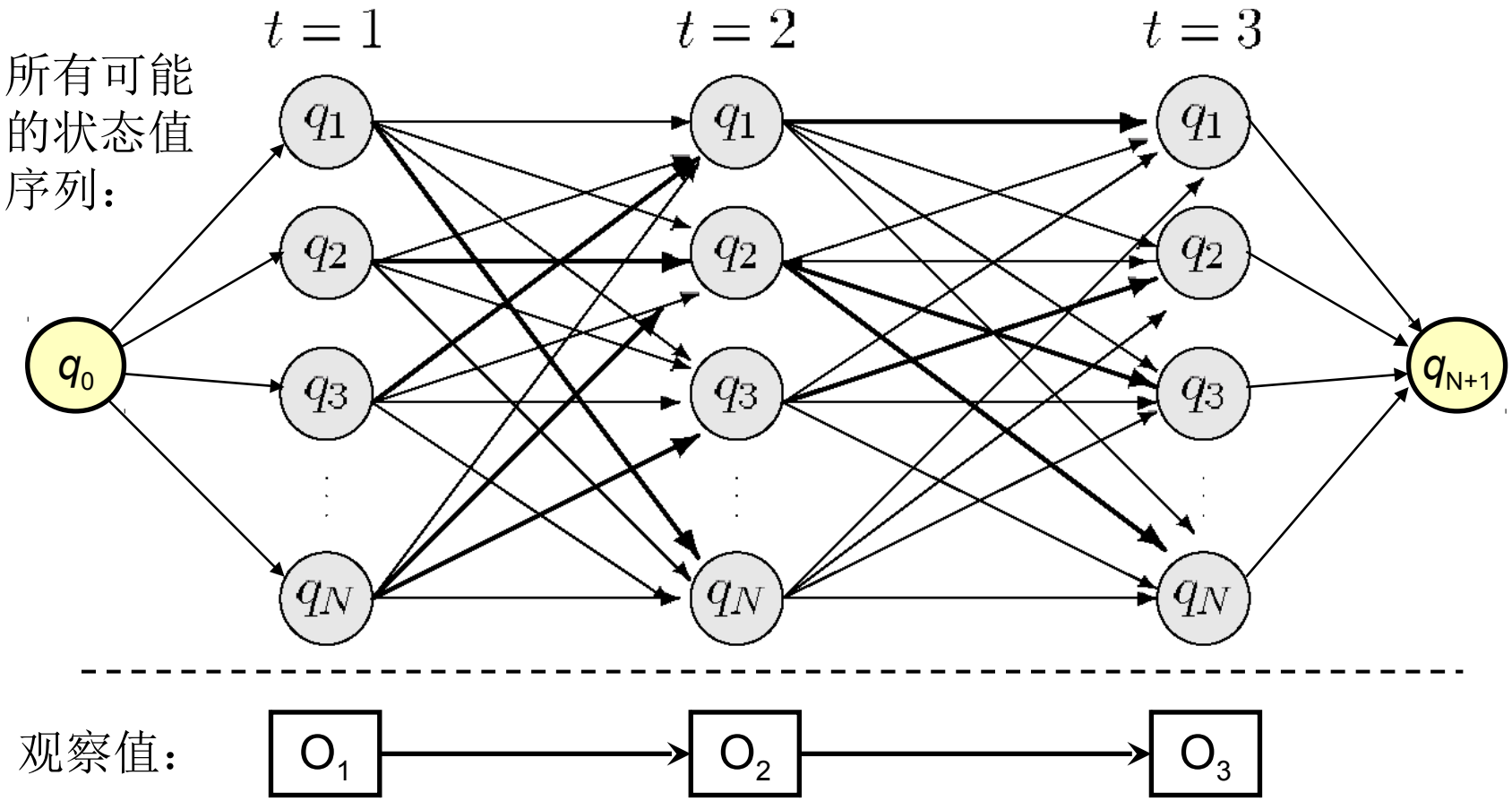
HMM 解码问题

解码问题：对于给定模型 λ 和观察值序列 O，
求可能性最大的状态序列 X；

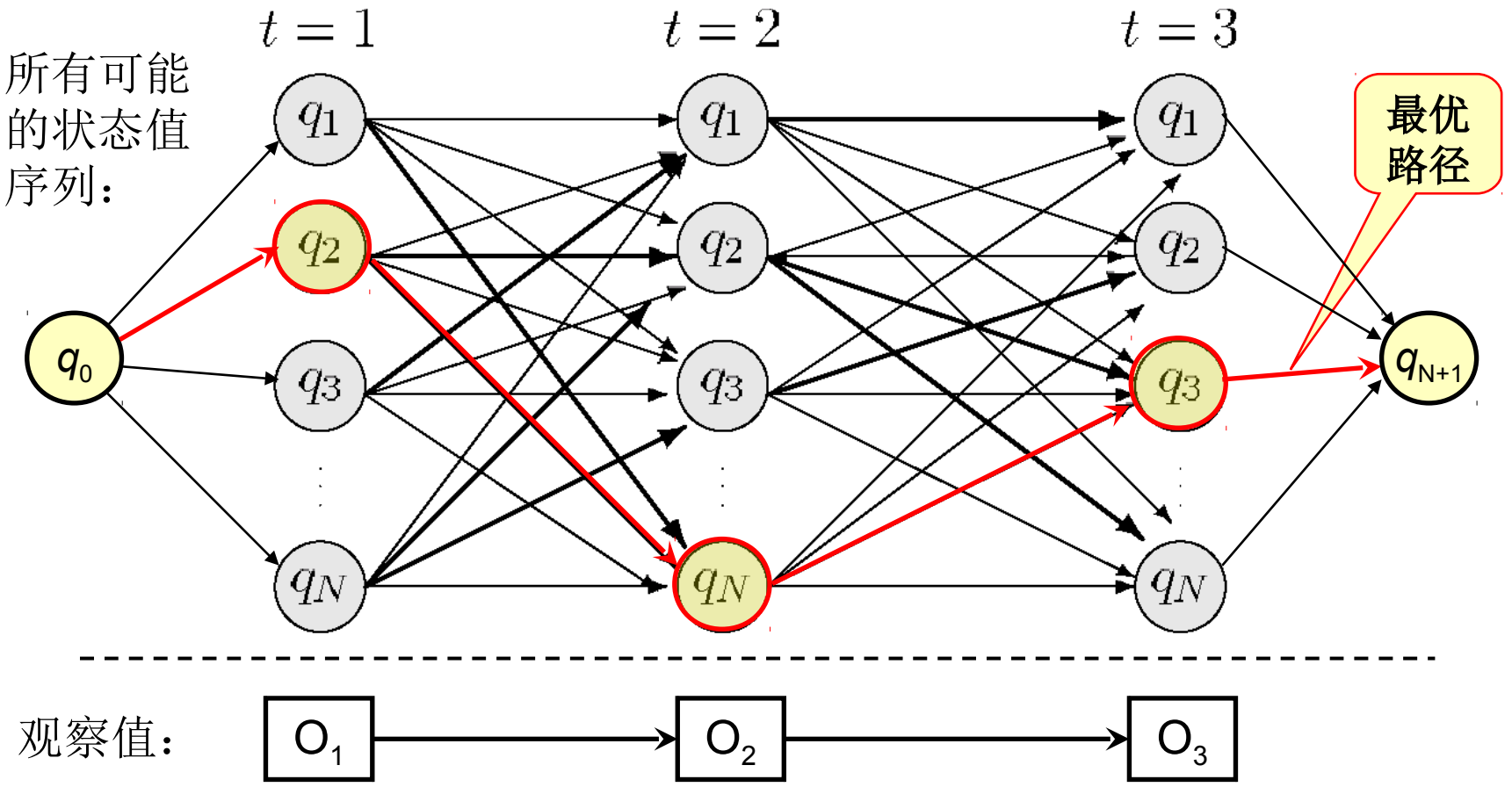
$$X^* = \arg\max_X P(X|O,\lambda)$$

如果要枚举所有的状态序列，时间复杂度是 $O(N^T)$

HMM 解码问题



HMM 解码问题



可能的状态序列有 N^T 种

HMM 解码问题— Viterbi 算法 (1)

定义 Viterbi 变量为 HMM 在时间 t 沿着某一条路径到达状态 q_i，且输出观察值 $O_1O_2...O_i$ 的最大概率

$$\delta_t(i) = \max_{X_1 X_2 \ldots X_{i-1}} P(X_1 X_2 \ldots X_t = q_i, O_1 O_2 \ldots O_t | \lambda)$$

HMM 解码问题— Viterbi 算法 (2)

- 初始化 $\delta_t(i)=\pi_i b_{iO_1}$
- 迭代计算 $\delta_t(j)=\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_{jO_t}$

  $2\le t\le T$

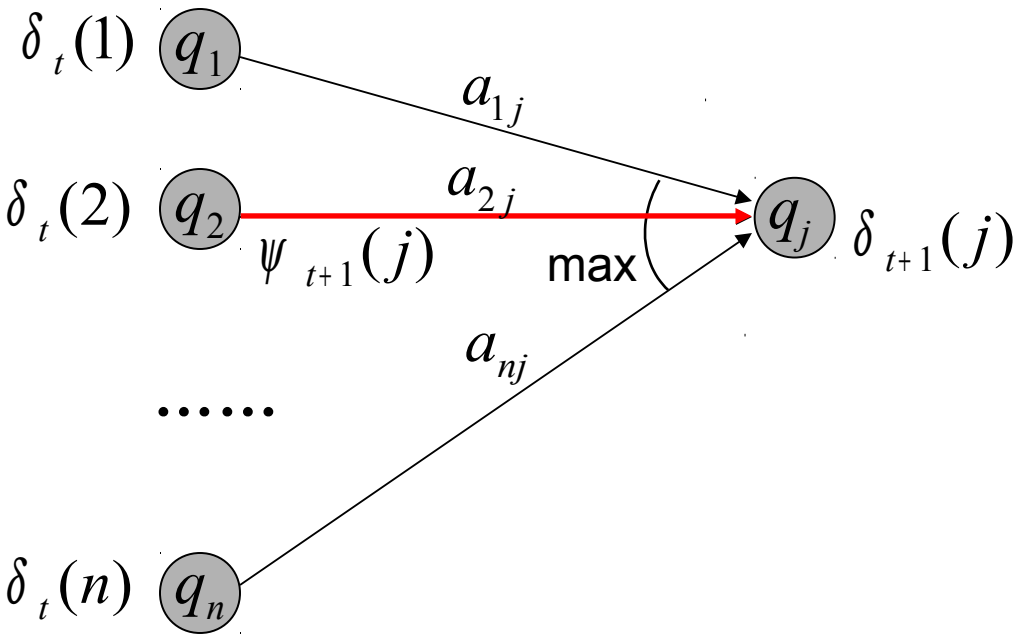
  $1\le j\le N$

  $\Psi_t(j)=\arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_{jO_t}$
- 取最优 $p^*=\max_{1\le i\le N}[\delta_T(i)]$

  $q_T^*=\arg\max_{1\le i\le N}[\delta_T(i)]$
- 路径回溯 $q_t^*=\Psi_{t+1}(q_{t+1}^*)$

  $2\le t\le T$

HMM 解码问题— Viterbi 算法 (3)



Viterbi 算法的时间复杂度： $O(N^2T)$

HMM 解码问题— Viterbi 算法 (4)

- **Viterbi Algorithm**
  - **Assign p(source_state)=1**
  - **For each observation *o* from source to destination**
    - **For each possible state *n* of observation *o***
      - **p(*n*)=0, previous_state(*n*)=ø**
      - **For each edge *e* directed to *n* from *n*'**
        - » **p'(*n*)=p(*n*')×transition_probability(*n*'|*n*)×output_probability(*o*|*n*)**
        - » **If p'(*n*) > p(*n*) then p(*n*)=p'(*n*), previous_state(*n*)= *n*'**
  - **Let *best_tag_sequence* is a empty array of states**
  - **Let state *n* is the destination state**
  - **Repeat until *n* is the source state**
    - **Push previous_state(*n*) to the head of *best_tag_sequence***
    - **assign *n*= previous_state(*n*)**
  - **Return *best_tag_sequence***

HMM 学习问题

- 已知观察序列 $O=O_1O_2\ldots O_T$
- 估计 λ 的参数： π_i, a_{ij}, b_{ik}

HMM 学习问题—最大似然估计

已知观察序列 O 对应的状态序列为（有指导学习）：

$$X = X_1 X_2 \ldots X_T$$

采用最大似然估计：

$$\bar{\pi}_i = \delta(X_1, q_i), \text{其中} \delta(x, y) = \begin{cases} 1, \text{如果} x = y \\ 0, \text{如果} x \neq y \end{cases}$$

$$\bar{a}_{ij} = \frac{X \text{中从状态} q_i \text{转移到状态} q_j \text{的次数}}{X \text{中从状态} q_i \text{转移到另一状态(含} q_j \text{)的次数}} = \frac{\sum_{t=1}^{T-1} \delta(X_t, q_i) \times \delta(X_{t+1}, q_j)}{\sum_{t=1}^{T-1} \delta(X_t, q_j)}$$

$$\bar{b}_{jk} = \frac{X \text{中从状态} q_j \text{输出到观察值} v_k \text{的次数}}{X \text{中到达状态} q_j \text{的次数}} = \frac{\sum_{t=1}^{T} \delta(X_t, q_j) \times \delta(O_t, v_k)}{\sum_{t=1}^{T} \delta(X_t, q_j)}$$

HMM 学习问题— Baum-Welch 算法 (1)

- 不知道 O 对应的状态序列：无指导学习
- 采用 Baum-Welch （又称向前向后算法）
- 是 EM （ Expectation-Maximization) 算法的一种实现
  - 初试化： λ_0
  - EM 步骤：循环执行以下步骤，直到 λ_i 收敛
    - E- 步骤：根据 λ_i 计算所有可能的状态序列
    - M- 步骤：根据状态序列和输出序列估计参数 λ_{i+1}

HMM 学习问题— Baum-Welch 算法 (2)

- 初试化：随机给 π_i, a_{ij}, b_{jk} 赋初始值

  需要满足以下归一化约束：

$$\sum_{i=1}^{N} \pi_i = 1$$

$$\sum_{j=1}^{N} a_{ij} = 1, \text{对于} 1 \le i \le N$$

$$\sum_{k=1}^{N} b_{jk} = 1, \text{对于} 1 \le j \le N$$

HMM 学习问题— Baum-Welch 算法 (3)

E- 步骤：已知观察序列 $O_1O_2\ldots O_T$ 和模型参数 π_i, a_{ij}, b_{jk}，估计：

- 在时间 t 和 t+1 分别位于状态 q_i， q_j 的概率：

$$\xi_t(i,j)=\frac{P(X_t=q_i, X_{t+1}=q_j, O|\lambda)}{P(O|\lambda)}$$

$$=\frac{\alpha_t(i)a_{ij}b_{jO_{t+1}}\beta_{t+1}(j)}{P(O|\lambda)}$$

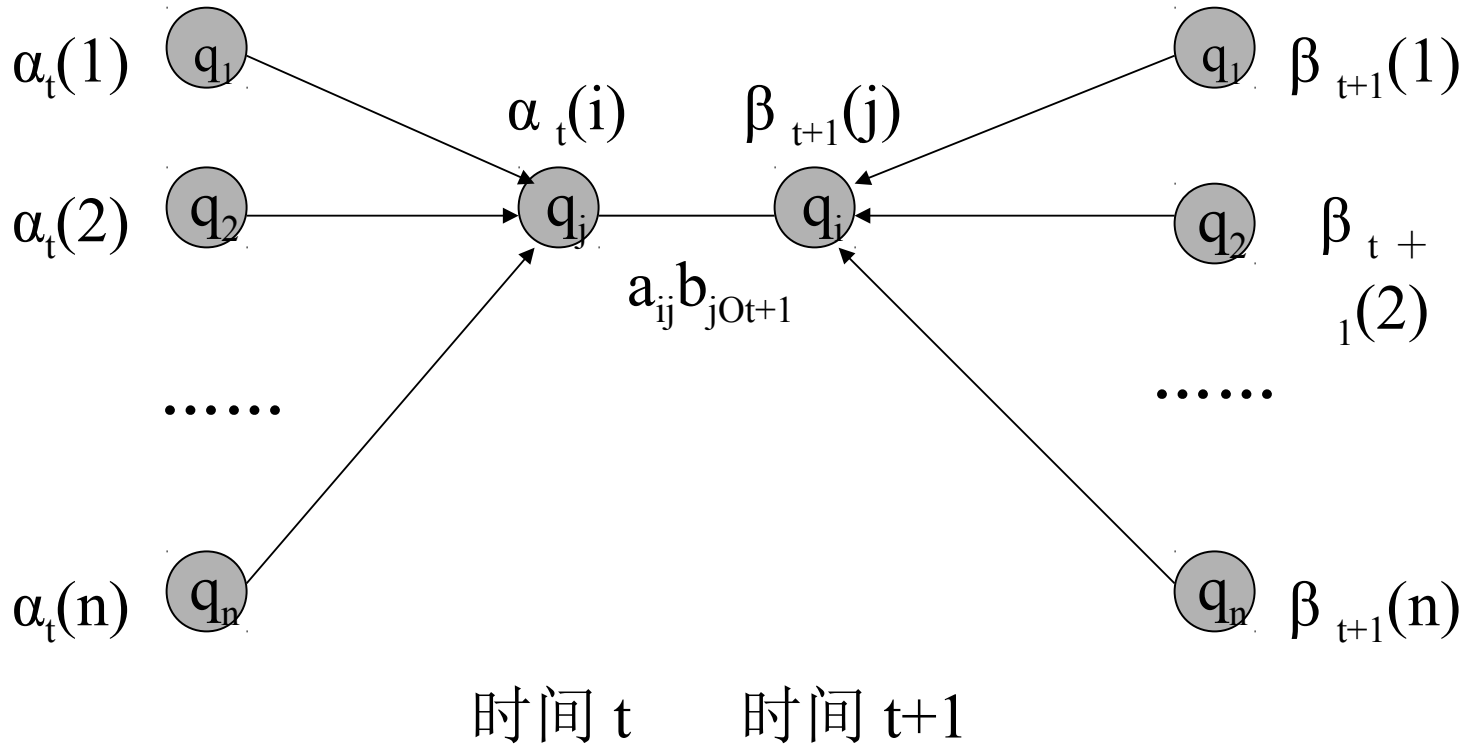
$$=\frac{\alpha_t(i)a_{ij}b_{jO_{i+1}}\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_{jO_{i+1}}\beta_{t+1}(j)}$$

- 在时间 t 位于状态 q_i 的概率：

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

HMM 学习问题— Baum-Welch 算法 (4)

$\xi_t(i, j)$ 的计算使用了前向变量和后向变量：



HMM 学习问题— Baum-Welch 算法 (5)

- M- 步骤：已知 $\xi_t(i,j)$ 和 $\gamma_t(i)$，估计模型 λ：

$$\bar{\pi}_i = X_1 \text{为} q_i \text{的概率} = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{X \text{中从状态} q_i \text{转移到状态} q_j \text{的次数}}{X \text{中从状态} q_i \text{转移到另一状态(含} q_j \text{)的次数}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma(i)}$$

$$\bar{b}_{jk} = \frac{X \text{中从状态} q_j \text{输出到观察值} v_k \text{的次数}}{X \text{中到达状态} q_j \text{的次数}} = \frac{\sum_{t=1}^{T} \gamma(j) \times \delta(O_t, v_k)}{\sum_{t=1}^{T} \gamma(j)}$$

HMM 学习问题— Baum-Welch 算法 (5)

迭代终止条件：

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

ε 是事先给定的阈值

HMM 学习问题— Baum-Welch 算法 (6)

- Baum-Welch 算法只能达到局部最优
- Baum-Welch 算法的结果依赖于初始值的设定

隐马尔科夫模型—应用

- 语音识别
- 音字转换
- 词性标注（ POS Tagging ）
- 组块分析
- 基因分析

隐马尔科夫模型—总结

- HMM 模型可以看作一种特定的 Bayes Net
- HMM 模型等价于概率正规语法或概率有限状态自动机
- HMM 模型可以用一种特定的神经网络模型来模拟
- 优点：研究透彻，算法成熟，效率高，效果好，易于训练

基于 HMM 进行词性标注 (1)

- 把**词汇序列**（记做 $W=w_1w_2\ldots w_n$ ）理解为**观察值**
- 把**词性标注序列**（记做 $T=t_1t_2\ldots t_n$ ）理解为隐含的**状态值**
- **状态转移概率**就是**基于词性的二元语法**
- **状态输出概率**就是**给定词性的词语概率分布**
- **词性标注问题**变成 HMM 中的**解码问题**
- 已知词串 W （观察序列）和模型 (λ) 情况下，求使得条件概率 P(T|W, λ) 值最大的那个 T’，一般记做：

$$T' = \arg\max_{T} P\left(T \middle| W, \lambda\right)$$

基于 HMM 进行词性标注 (2)

利用 Bayes 公式，可以进一步分解为：

$$\arg\max_T P(T|W) = \arg\max_T P(T)P(W|T)$$

其中：

$$P(T) = P(t_1|t_0)P(t_2|t_1,t_0)\ldots P(t_i|t_{i-1},t_{i-2},\ldots)$$

根据 HMM 假设，可得

$$P(T) \approx P(t_1|t_0)P(t_2|t_1)\ldots P(t_i|t_{i-1})$$

词性之间的转移概率可以从语料库中估算得到：

$$P(t_i|t_{i-1}) \approx \frac{\text{训练语料中}\ t_i\ \text{出现在}\ t_{i-1}\ \text{之后的次数}}{\text{训练语料中}\ t_i\ \text{出现的总次数}}$$

基于 HMM 进行词性标注 (3)

P(W|T) 是已知词性标记串，产生词串的条件概率：

$$P(W|T)=P(w_1|t_1)P(w_2|t_2,t_1,w_1)...P(w_i|t_i,t_{i-1},...,t_1,w_i,w_{i-1},...,w_1)$$

根据 HMM 假设，上面公式可简化为：

$$P(W|T)\approx P(w_1|t_1)P(w_2|t_2)...P(w_i|t_i)$$

已知词性标记下输出词语的概率可以从语料库中统计得到：

$$P(w_i|t_i)\approx \frac{\text{训练语料中}\ w_i\ \text{的词性被标记为}\ t_i\ \text{的次数}}{\text{训练语料中}\ t_i\ \text{出现的总次数}}$$

基于 HMM 进行词性标注示例

- 把 /?　这 /?　篇 /?　报道 /?　编辑 /?　一 /?　下 /?
  把 /q-p-v-n　这 /r　篇 /q　报道 /v-n　编辑 /v-n　一 /m-c　下 /f-q-v

P(T1|W) = P(q|$)P(把 |q)P(r|q)P(这 |r)...P(f|m)P(下 |f)

P(T2|W) = P(q|$)P(把 |q)P(r|q)P(这 |r)...P(q|m)P(下 |q)

P(T3|W) = P(q|$)P(把 |q)P(r|q)P(这 |r)...P(v|m)P(下 |v)

……

P(T96|W) = P(n|$)P(把 |n)P(r|q)P(这 |r)...P(v|c)P(下 |v)

从中选一个最大值

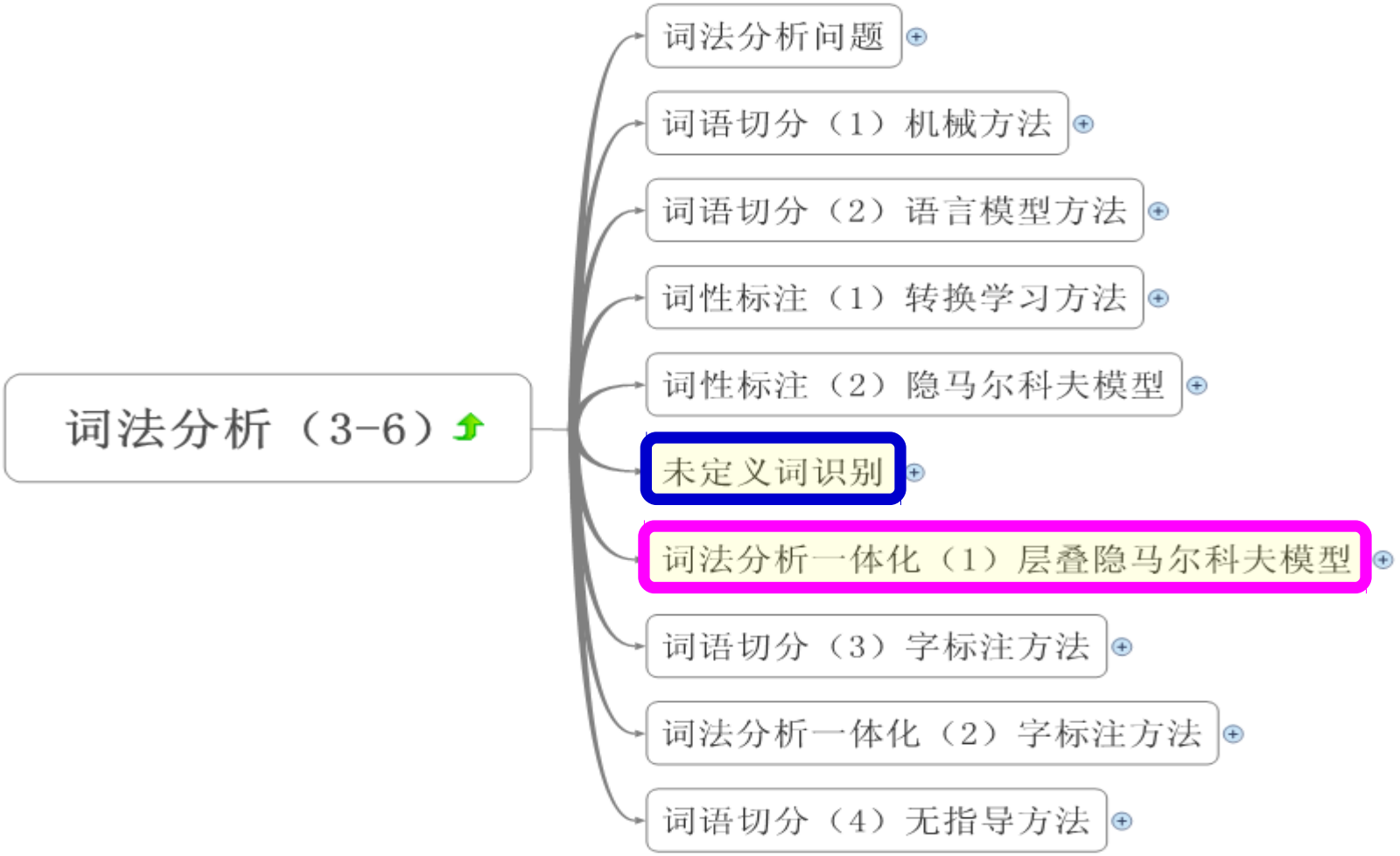
词性转移概率

词语输出概率

基于 HMM 进行词性标注 (4)

- 基于 HMM 的词性标注实际上对应着 HMM 的解码问题
- 采用 Viterbi 算法即可解决

内容提要


词法分析（3-6）
词法分析问题
词语切分（1）机械方法
词语切分（2）语言模型方法
词性标注（1）转换学习方法
词性标注（2）隐马尔科夫模型
未定义词识别
词法分析一体化（1）层叠隐马尔科夫模型
词语切分（3）字标注方法
词法分析一体化（2）字标注方法
词语切分（4）无指导方法

未定义词识别的一般方法

- 各种不同类型的未定义词识别方法思想大同小异，但实现时各有侧重
- 各种不同类型的未定义词识别都需要收集大量数据，建立不同的数据模型
- 常用的方法包括
  - 规则方法：人工总结或归纳出一些判别规则，并用程序实现
  - 统计方法：建立统计模型，通过人工标注语料库进行参数训练

将识别问题转化成标注问题 (1)

- 在统计方法中，未定义词识别的一种最通常的做法就是将识别问题转化成标注问题
- 对于输入句子中的每个汉字，定义四个标记：
  - 不属于未定义词 O
  - 未定义词首字 B
  - 未定义词尾字 E
  - 未定义词中间字 M

将识别问题转化成标注问题 (2)

- 如果能够把输入句子中的每个汉字都正确地按上述标记进行标注，那么未定义词的识别自然就解决了
- 汉字序列的标注问题可以采用隐马尔科夫模型（ HMM ）、最大熵（ ME ）、最大熵马尔科夫模型（ MEMM ）、条件随机场（ CRF ）等模型来解决

基于 HMM 的未定义词识别

- 输入文本：

  这 是 周 恩 来 、 邓 颖 超 生 前 居 住 的 地 方

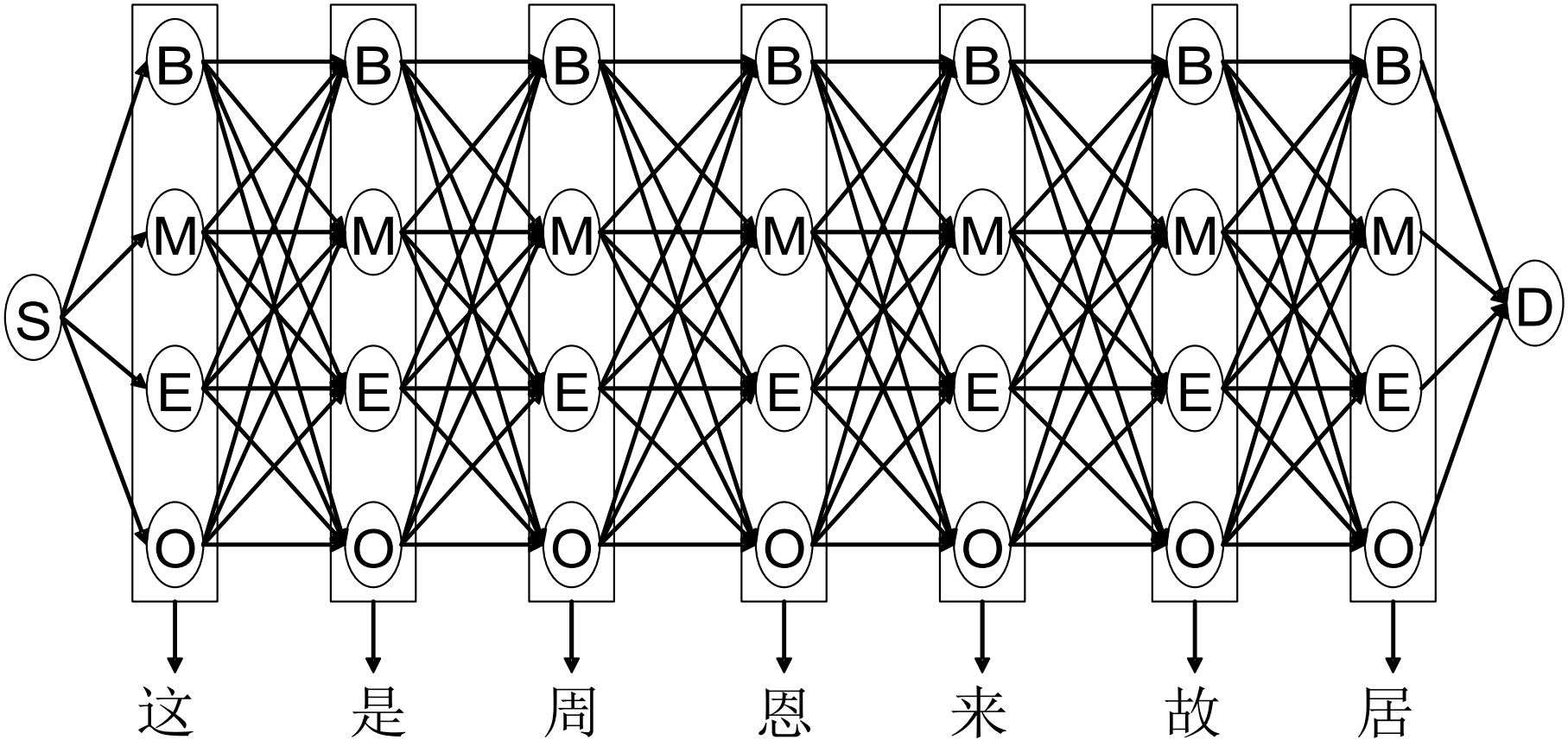
- 标注为：

  这 是 周 恩 来 、 邓 颖 超 生 前 居 住 的 地 方

  O O B M E O B M E O O O O O O O

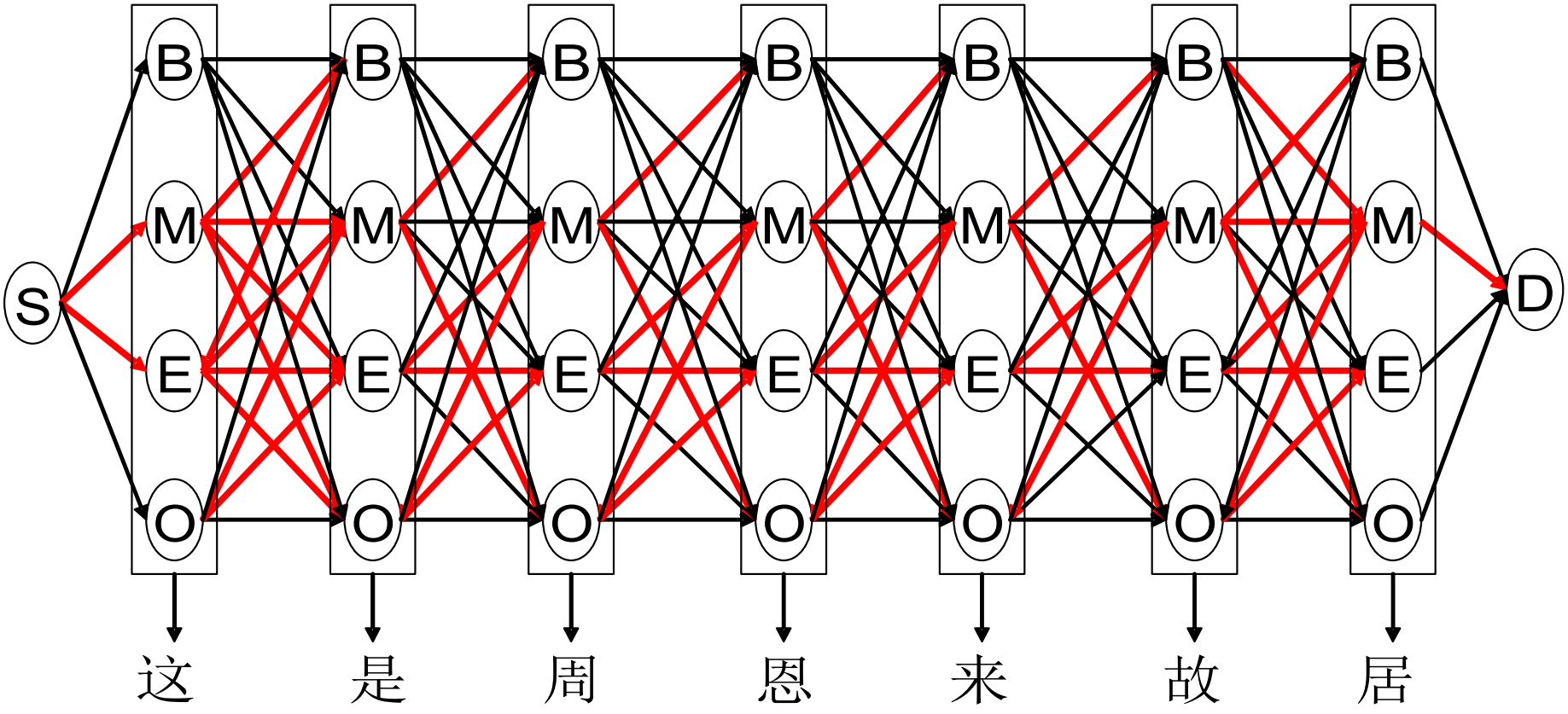
- 两处标注为 BME 的字串“周恩来”、“邓颖超”被识别为未定义词
- 训练语料库为已经标注未定义词的语料库

基于 HMM 的汉语人名识别



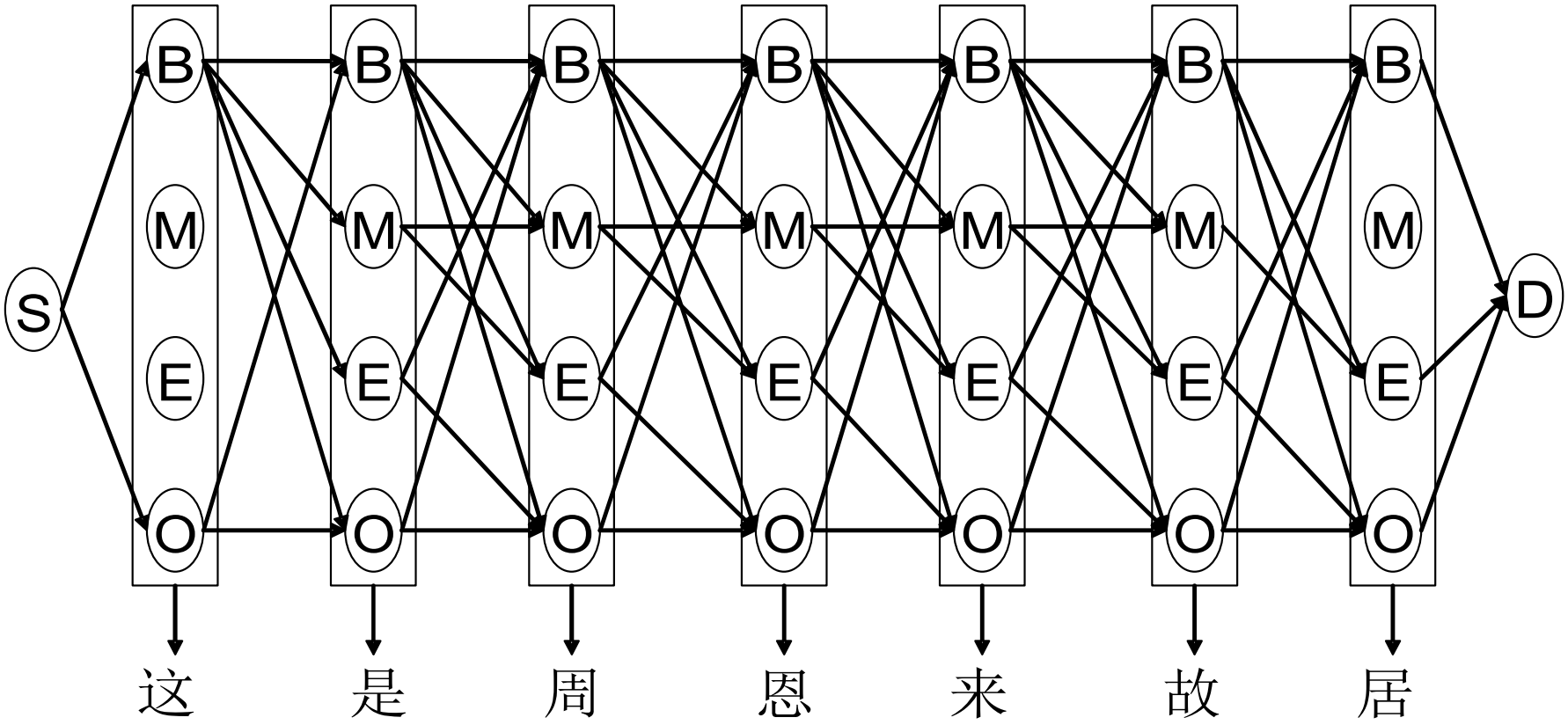
基于 HMM 的汉语人名识别

非法边：



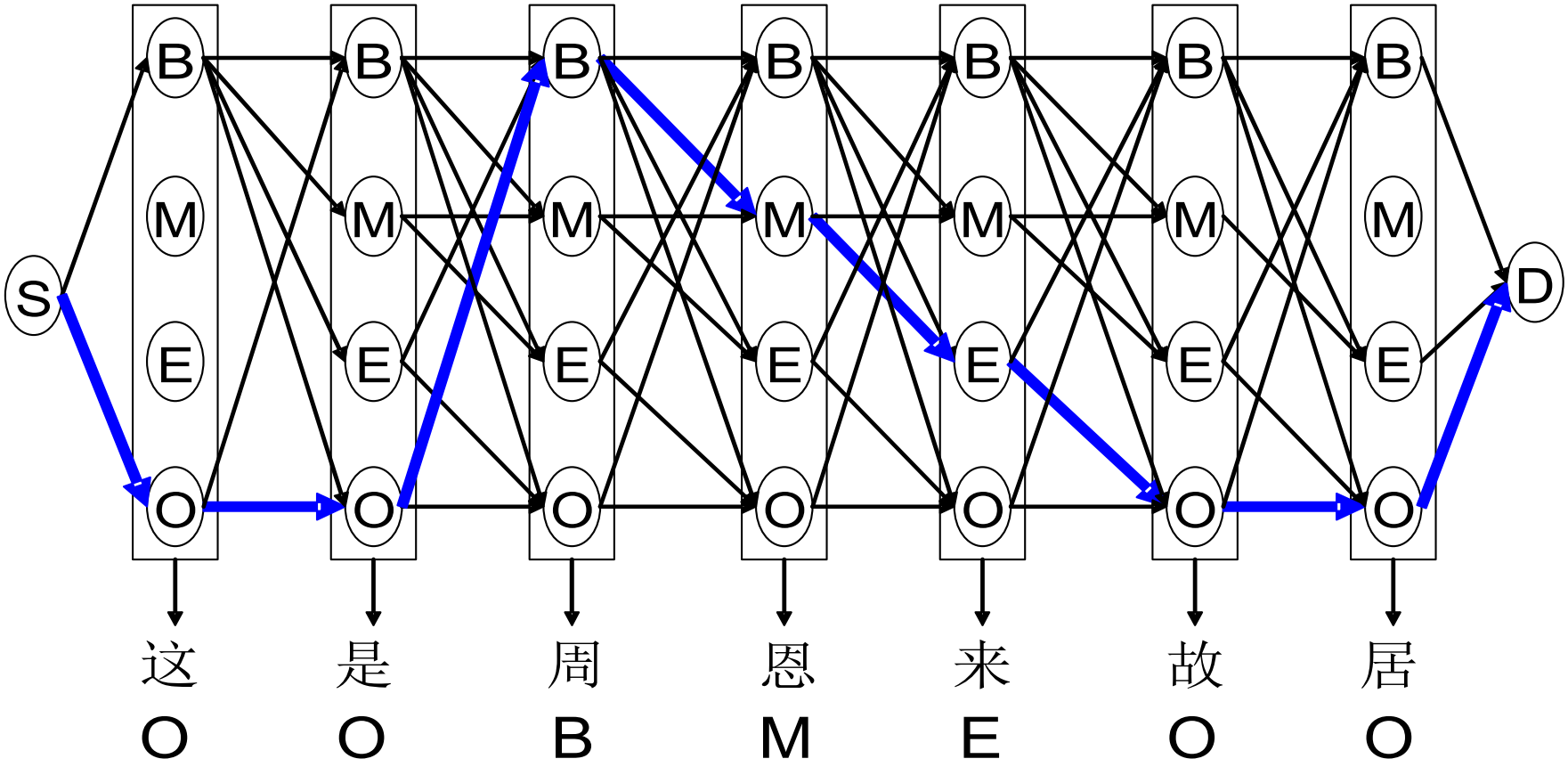
基于 HMM 的汉语人名识别

删除非法边：



基于 HMM 的汉语人名识别

搜索最优的标记路径：



基于 HMM 的汉语人名识别

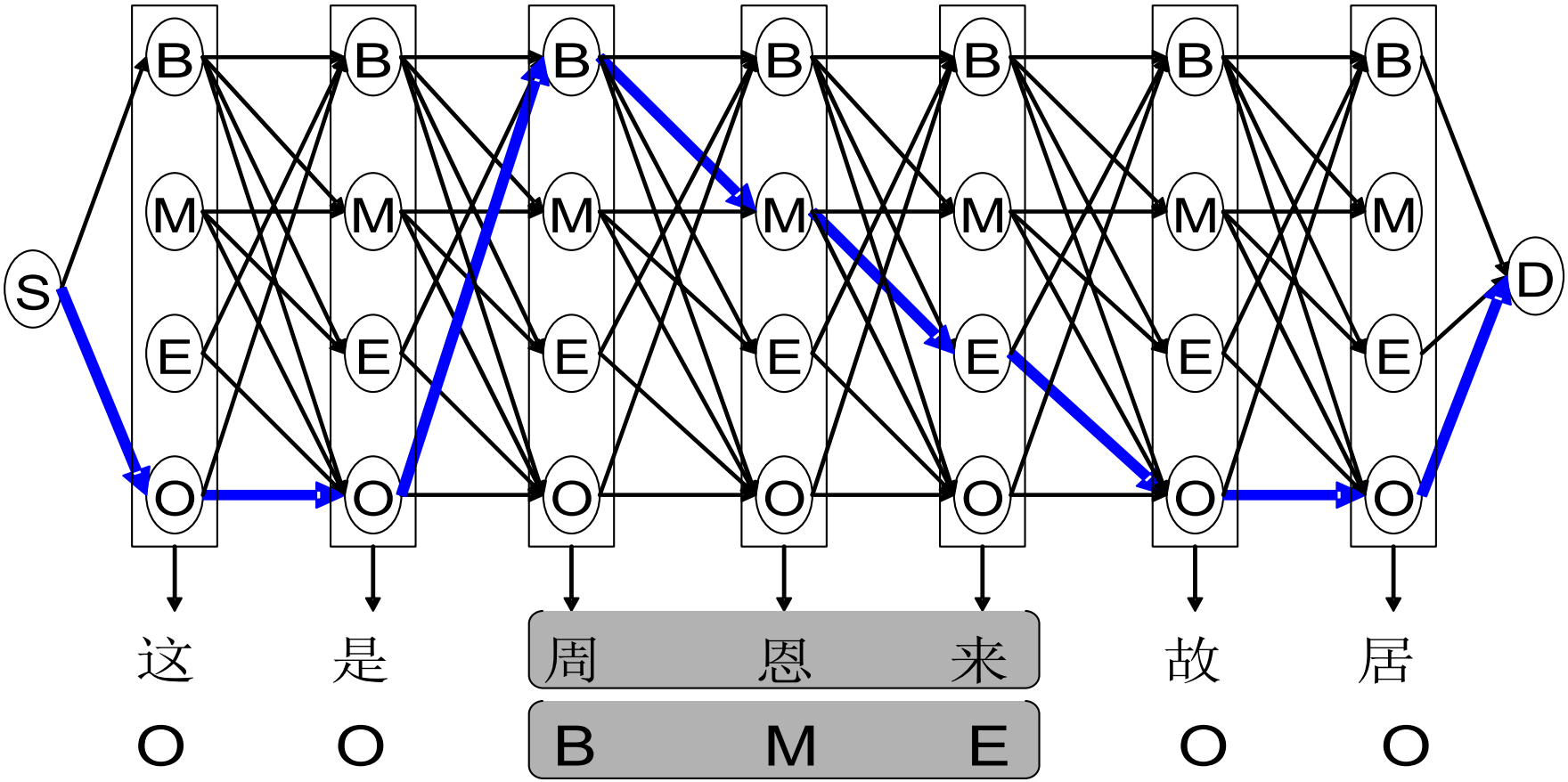
- 在最优路径上匹配以下标记片段：

  - B
  - BE
  - BME
  - BMME
  - ……

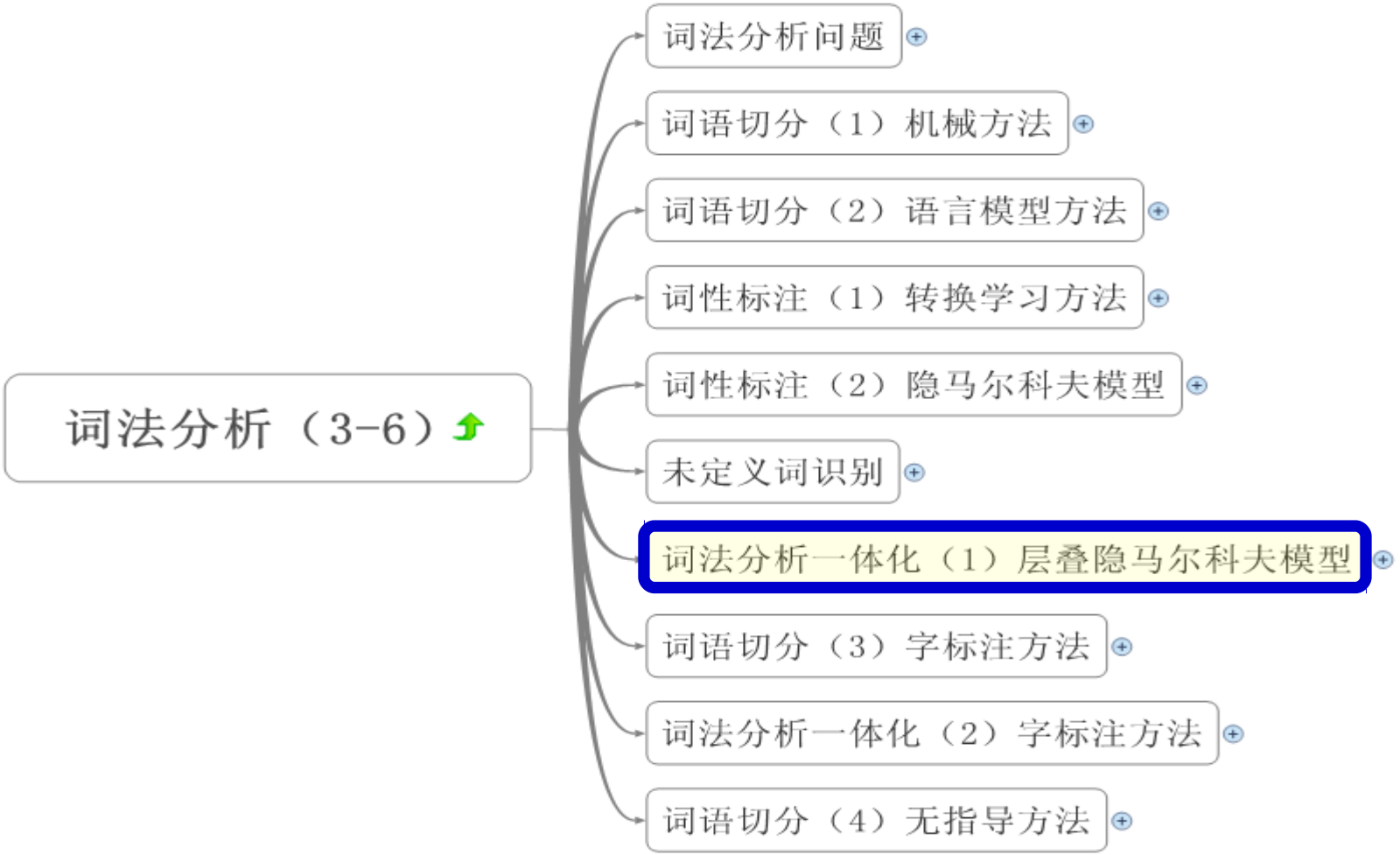
  } 人名标记序列

基于 HMM 的汉语人名识别

在标记序列上匹配找到人名片段：



内容提要


词法分析（3-6）
词法分析问题
词语切分（1）机械方法
词语切分（2）语言模型方法
词性标注（1）转换学习方法
词性标注（2）隐马尔科夫模型
未定义词识别
词法分析一体化（1）层叠隐马尔科夫模型
词语切分（3）字标注方法
词法分析一体化（2）字标注方法
词语切分（4）无指导方法

完整的汉语词法分析系统 (1)

- 任务
  - 形态变化：重叠词、前后缀、离合词
  - 词语切分
  - 未定义词识别
  - 词性标注
- 问题
  - 如何安排上述操作的顺序？
  - 如何消解产生的歧义？

完整的汉语词法分析系统 (2)

- 问题 1：未定义词识别与切分的顺序
  - 先切分后识别再切分（基于粗切分的未定义词识别）
  - 先识别后切分（基于字的未定义词识别）
  - 同时进行
- 问题 2：切分与标注的顺序
  - 先切分后标注
  - 同时进行
- 问题 3：建立统一的概率模型
  - 切分过程中变形词的概率计算
  - 切分过程中未定义词的概率计算
  - 切分与未定义词识别的统一概率模型
  - 切分与标注的统一概率模型

先识别未定义词，后切分

- 优点：切分不会对未定义词识别造成干扰
  - 未定义词内部成词
  - 未定义词本身又是其他词
  - 未定义词的首部与上文成词
  - 未定义词尾部与下文成词
- 缺点：无法利用切分所显现的上下文信息
  - 上下文词对未定义词识别有重要的提示作用，如职务词对人名识别的提示作用
  - 如果要利用这种提示作用，至少要构造基于字的三元语法，复杂程度远高于基于词的二元语法

先切分，后识别未定义词

- 优点：
  - 识别出的上下文词语对未定义词识别有重要的提示作用
- 缺点：
  - 识别本身对未定义词造成干扰
  - 未定义词识别完成后需要重新进行切分排歧

切分与未定义词识别同时进行

- 需要建立切分与未定义词识别的统一概率模型
  - 基于类的语言模型
  - 基于最大熵的切分模型
- 优点：一致的模型
- 缺点：搜索空间大

切分与标注的顺序问题

- 先切分后标注
  - 最好采用 N-Best 策略，否则切分的错误将无法挽回
  - 方法简单，效率较高
- 同时进行
  - 需要定义统一的概率模型
  - 搜索空间较大，时间复杂度高

词法分析中集成的概率模型

- 切分过程中变形词的概率计算
- 切分与未定义词识别的统一概率模型
- 切分与标注的统一概率模型

切分过程中变形词的概率计算

- 汉语词的变形
  - 重叠形式
  - 前后缀
  - 离合词
  - 缩略语：中国科学院→中科院
- 变形词的概率计算：没有合适的办法
  - 作为一个单独词计算：稀疏问题非常严重
  - 等同于原形词：不合理，变形词的分布特点与原形词不同，甚至词性都可能发生变化
  - 根据原形词的概率做某种变化处理：没有理论依据

切分过程中未定义词的概率计算

- 未定义词概率计算

$$P(\text{李素丽}|\text{人名})?$$

基于类语言模型的中文词法分析

- 孙健，基于统计方法的短语识别和句法结构歧义消解的研究，北京邮电大学博士论文，2002

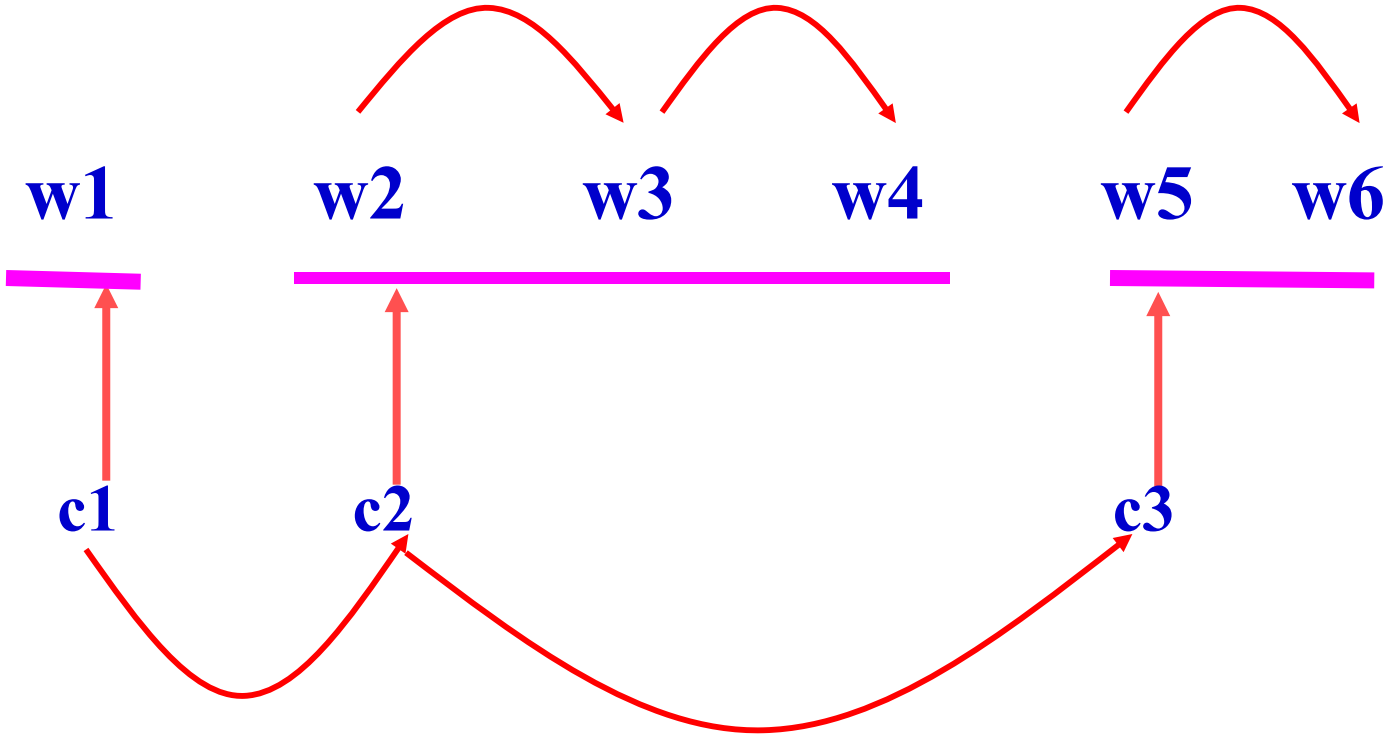
  “ 基于类的语言模型”

  “Class-Based Language Model”

基于类的语言模型 (1)

- 基于“类”的语言模型：
  - 未定义词划分成类
    - 中国人名
    - 外国人名
    - 中国地名
    - 机构名
  - 每个词典词单独作为一类
  - 识别与标注的过程就是将每个汉字归结到“类”的过程
  - 语境模型：类与类之间采用 N 元语法模型
  - 每一个未定义词内部也分别采用一部 N 元语法模型，分别构成人名模型、地名模型、机构名模型等等

基于类的语言模型 (2)



基于类的语言模型 (3)

$$C^* = \arg\max_C P(C \mid T)$$

$$= \arg\max_C [P(C) \times P(T \mid C)]$$

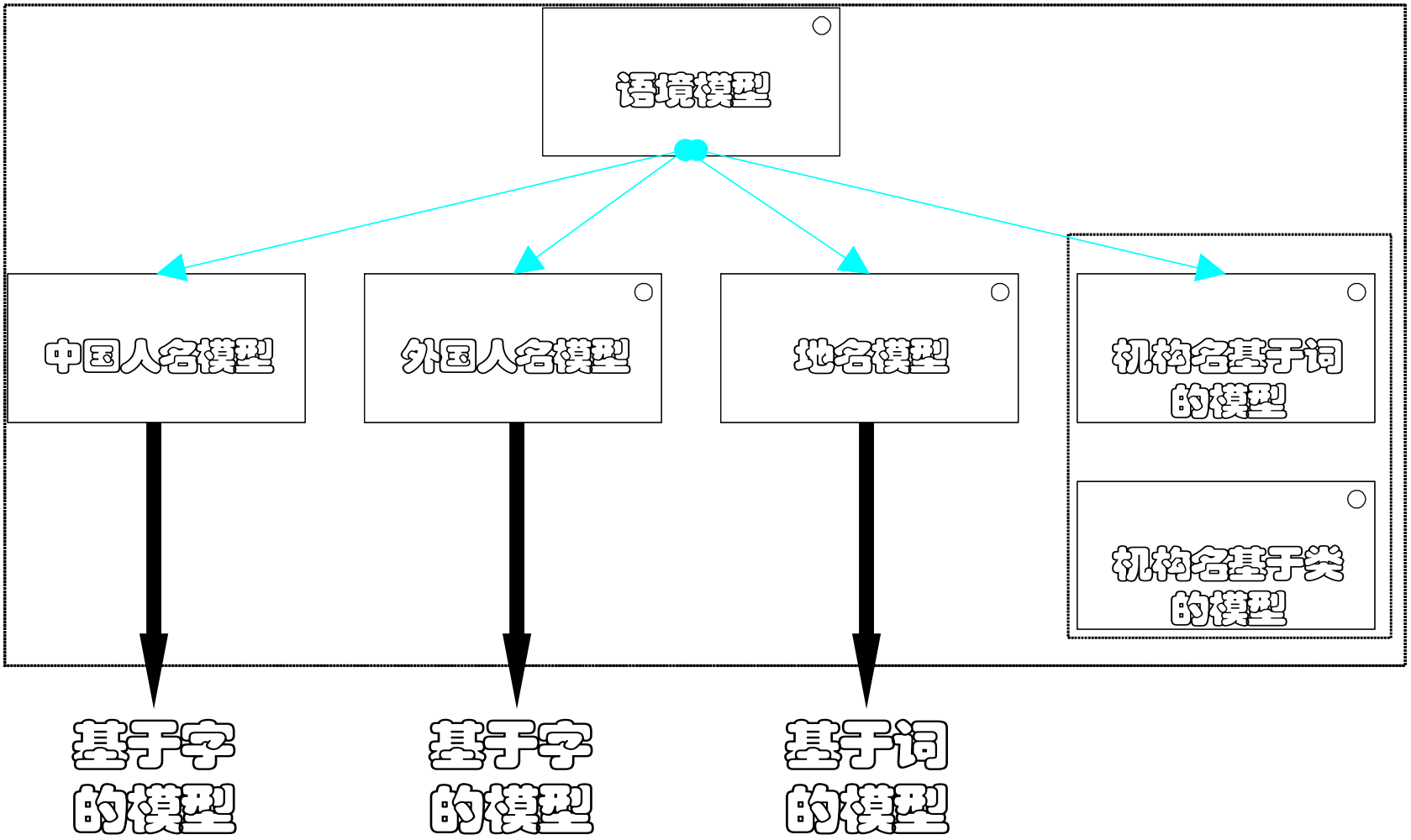
$$T = t_1 t_2 \ldots t_i \ldots t_n$$

$$C = c_1 c_2 \ldots c_i \ldots c_m$$

$P(C)$ --- class model　　C：类序列

$P(T \mid C)$ --- entity model　　T：汉字序列

基于类的语言模型 (4)


语境模型
中国人名模型
外国人名模型
地名模型
机构名基于词的模型
机构名基于类的模型
基于字的模型
基于字的模型
基于词的模型

基于类的语言模型 (5) 语料

Id	Domain	Named Entity Identification			(byte)
		Person	location	organization	
1	Army	65	202	25	19k
2	Computer	75	156	171	59k
3	Culture	548	639	85	138k
4	Economy	160	824	363	108k
5	Entertainment	672	575	139	104k
6	Literature	464	707	122	96k
7	Nation	448	1193	250	101k
8	People	1147	912	403	116k
9	Politics	525	1148	218	122k
10	Science	155	204	87	60k
11	Sports	743	1198	625	114k
	Total	5002	7758	2491	1037k

基于类的语言模型 (6) 结果

命名实体	准确率	召回率	F
人名	79.86	87.29	83.41
地名	80.88	82.46	81.66
机构名	76.63	56.54	65.07
3 类命名实体的综合	79.99	79.68	79.83

基于类的语言模型 (7) 总结

- 优点
  - 中文分词和命名实体识别结合在一起
  - 语境信息和实体内部信息有机结合在一起
  - 人名、地名和机构名这三类不同的命名实体识别纳入到统一的模型框架中
  - 启发式信息有机融入到语言模型
  - 能够识别嵌套的命名实体
- 缺点
  - 多层语言模型混合，模型之间组合的所有可能性都要考虑，搜索空间极大，时间复杂度高

基于层叠隐马尔科夫模型的汉语词法分析系统（ICTCLAS）

刘群，张华平，俞鸿魁，程学旗，基于层次隐马模型的汉语词法分析，计算机研究与发展， 2004.6

Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang and Hong-Kui Yu, Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, proceedings of 2nd SigHan Workshop, July 2003, pp. 63-70

张华平，刘群，基于角色标注的中国人名自动识别研究，计算机学报， 2004.1

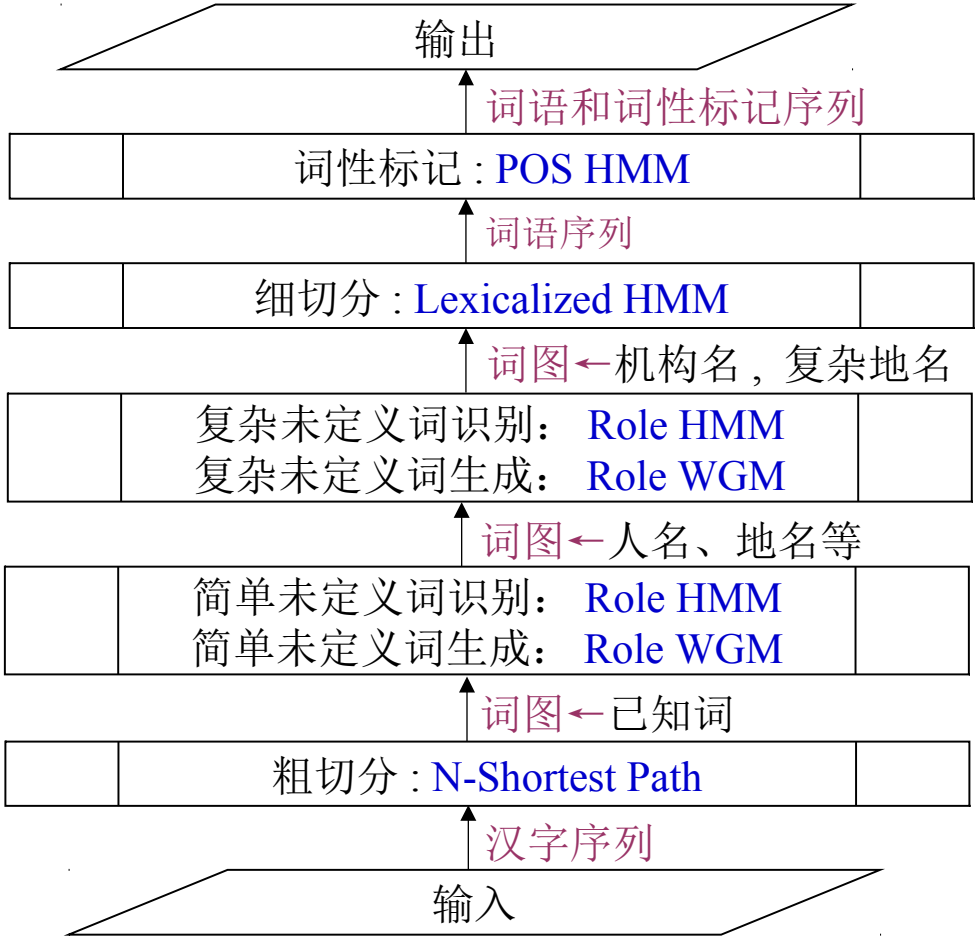
Hong-Kui Yu, Hua-Ping Zhang, Qun Liu, Recognition of Chinese Organization Name Based on Role Tagging（中文）, in Maosong Sun, Tianshun Yao, Chunfa Yuan, eds., Advances in Computation of Oriental Languages, Proceedings of 20th International Conference on Computer Processing of Oriental Languages, Tsinghua University Press, August 2003, pp. 79-87

张华平，刘群， 基于 N- 最短路径方法的中文词语粗分模型，《中文信息学报》第 16 卷第 5 期，2002 年

层叠隐马尔可夫模型 (1)

- 基本思想
  - 每个层次都是某种形式的隐马尔可夫模型
  - 层次之间通过概率信息的传递紧密耦合
  - 每个低层模型提供 N-Best 输入供高层模型选择
- 层次划分
  - 粗切分：基于 N- 最短路径的粗切分
  - 简单未定义词识别：基于角色标注的识别算法
  - 复合未定义词识别：基于角色标注的识别算法
  - 细切分：词汇化的隐马尔可夫模型
  - 词性标注：基于 HMM 的词性标注算法

层叠隐马尔可夫模型 (2)



层叠隐马尔可夫模型（ Cascaded HMM ）框架

基于 N 最短路径的粗切分 (1)

- N-Best 思想：在效率和性能之间的平衡
  - 每一阶段搜索时输出 N 个最好的结果而不是仅仅输出一个最好的结果
  - 对于分阶段的搜索是一种有效的做法
  - 既可以淘汰大部分不合理的结果，又不至于对最终结果的正确率造成太大损失
- 基于 N 最短路径的粗切分
  - N-Best 思想在汉语分词中的体现
  - 在未定义词识别之前进行粗切分
  - 粗切分采用基于 N 最短路径的一元语法，效率极高

基于 N 最短路径的粗切分 (2)

- 粗切分结果正确与否的判定
  - 粗分结果中除未登录词外的其它部分与参考结果必须完全一致；
  - 未登录词部分的字串必须可以组合成参考结果中对应的未登录词，即这部分的字串不能和其它部分组成词。
    - 错误：尉 健 行李 岚 清
    - 正确：尉 健 行 李 岚 清
- 粗切分的召回率
  - 粗切分产生的 N 个结果中包含一个正确结果的句子占所有句子的比例

基于 N 最短路径的粗切分 (3)

- 对 Viterbi 算法的改进：保留 N-Best 结果
- 算法基本思想：对于每一步，保留到达当前位置的 N-Best 结果（具体算法略）

基于 N 最短路径的粗切分 (4)

路径数（N）	粗分召回率（%）
1	92.88
2	98.55
3	99.33
4	99.67
6	99.80
7	99.86
8	99.89
9	99.91

基于角色标注的人名识别 (1)

Huaping Zhang, Qun Liu, Hao Zhang,Xueqi Cheng, Automatic Recognition of Chinese Unknown Words Based on Role Tagging, SigHan Workshop, attached with 19th International Conference on Computational Linguistics, Taipei, 2002.8.

基于角色标注的人名识别 (2)

中国人名识别的角色定义

角色	意义	例子
B	姓氏	*张*华平先生
C	双名的首字	张*华*平先生
D	双名的末字	张华*平*先生
E	单名	张*浩*说：“我是一个好人”
F	前缀	*老*刘、*小*李
G	后缀	王*总*、刘*老*、肖*氏*、吴*妈*、叶*帅*
K	人名的上文	又*来到*于洪洋的家。
L	人名的下文	新华社记者黄文*摄*
M	两个中国人名之间的成分	编剧邵钧林*和*稽道青说
U	人名的上文和姓成词	这里*有关*天培的壮烈
V	人名的末字和下文成词	龚学*平等*领导， 邓颖*超生*前
X	姓与双名的首字成词	*王国*维、
Y	姓与单名成词	*高峰*、*汪洋*
Z	双名本身成词	张*朝阳*
A	以上之外其他的角色	

基于角色标注的人名识别 (3)

- 例子
  - 人名识别前的切分结果：

    馆 / 内 / 陈列 / 周 / 恩 / 来 / 和 / 邓 / 颖 / 超生 / 前 / 使用 / 过 / 的 / 物品 / 。
  - 角色标注后的结果：

    馆 /A 内 /A 陈列 /K 周 /B 恩 /C 来 /D 和 /M 邓 /B 颖 /C 超生 /V 前 /A 使用 /A 过 /A 的 /A 物品 /A 。 /A

基于角色标注的人名识别 (4)

- 采用隐马尔科夫模型（ HMM ）进行角色标注
  - 将“角色”理解为 HMM 中的“状态”
  - 将人名识别前切分出的词理解为 HMM 中的“观察值”
  - 将已有的词标注语料库（《人民日报》语料库）转换为角色标注语料库
  - 利用角色标注语料库对 HMM 模型参数进行训练

基于角色标注的人名识别 (5)

- 语料库的转换
  - 《人民日报》语料库原始形式

    政务司 /n 司长 /n 陈 /nr 方 /nr 安生 /nr
    出任 /v 委员会 /n 主席 /n
  - 转换后的形式

    政务司 /A 司长 /K 陈 /B 方 /B 安 /C 生 /D
    出任 /L 委员会 /A 主席 /A
  - 统计得到状态转移矩阵和输出矩阵

基于角色标注的人名识别 (6)

- 角色分裂
  - 在人名识别之前，我们要对角色 U 和 V 进行分裂处理。相应地分裂为 KB 、 DL 或者 EL
  - 例子：
    - 馆 / 内 / 陈列 / 周 / 恩 / 来 / 和 / 邓 / 颖 / 超生 / 前 / 使用 / 过 / 的 / 物品 / 。
    - 分裂前： AAKBCDMBC*V*AAAAAA
    - 分裂后： AAKBCDMBC*DL*AAAAAA
- 模式匹配：得到人名
  - BBCD, BBE, BBZ, BCD, BEE,BE,BG,BXD,BZ,CD,EE,FB, Y,XD
  - 例子： 模式 BCD ：周恩来，邓颖超

基于角色标注的人名识别 (7)

类别	封闭测试语料 1	封闭测试语料 2	开放测试语料
来源：《人民日报》	98 年 1 月	98 年 2 月 1 日 -20 日	98 年 2 月 21 日 -28 日
语料库大小（字节）	8,621K	6,185K	2,605K
实际人名数	13360	7224	2967
识别出的人名数	17505	10929	4259
正确识别的人名数	13079	7106	2739
准确率	74.72%	65.02	64.32%
召回率	97.90%	98.37%	92.32%
F 值	84.75%	78.29%	75.81%

训练语料：《人民日报》 98 年 1 月 1 日～ 2 月 20 日

基于角色标注的人名识别 (8)

- 结果分析
  - 完全真实的测试环境：没有剔除不含人名的句子
  - 测试结果仅针对词典中未定义的人名（如果考虑词典中已有的人名，正确率和召回率都将达到 95 %以上）
  - 实验规模大
  - 还有改进的余地
  - 基于角色标注的方法可用于各种未定义词识别

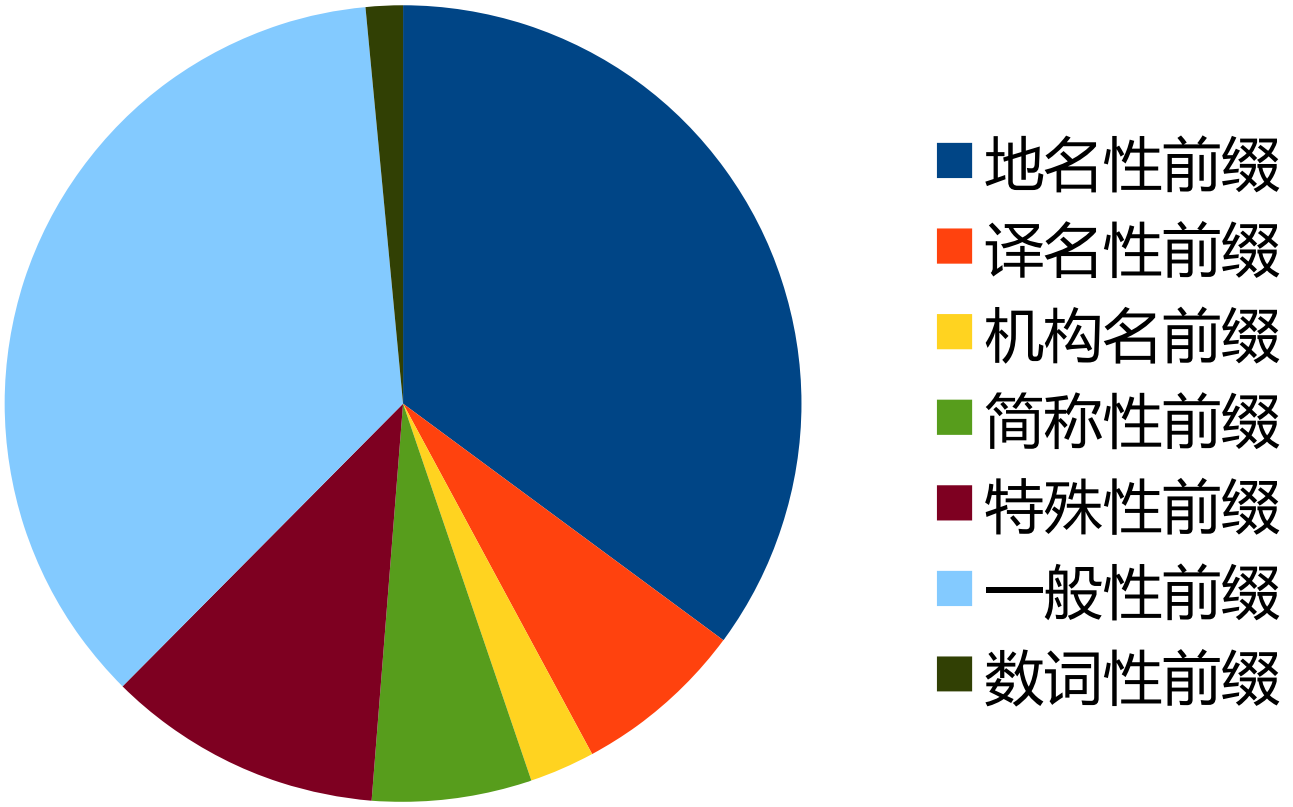
基于角色标注的音译名识别

代码	含义	例子
B	首字	*克* / 林 / 顿
C	中字	史 / *蒂* / *芬* /·/ *斯* / *皮* / *尔* / *伯* / 格
D	尾字	克 / 林 / *顿*
K	左邻（上文）	
L	右邻（下文）	
M	两个音译名之间的连接	
A	其他	
……		

基于角色标注的机构名识别 (1)

中文机构名称构成角色表		
角色	意义	例子
A	上文	*参与*亚太经合组织的活动
B	下文	中央电视台*报道*
X	连接词	北京电视台*和*天津电视台
C	特征词的一般性前缀	
F	特征词的译名性前缀	美国*摩托罗拉*公司
G	特征词的地名性前缀	交通银行*北京*分行
H	特征词的机构名前缀	*中共中央*顾问委员会
I	特征词的特殊性前缀	*中央*电视台
J	特征词的简称性前缀	
D	机构名的特征词	
Z	非机构名成份	

基于角色标注的机构名识别 (2)



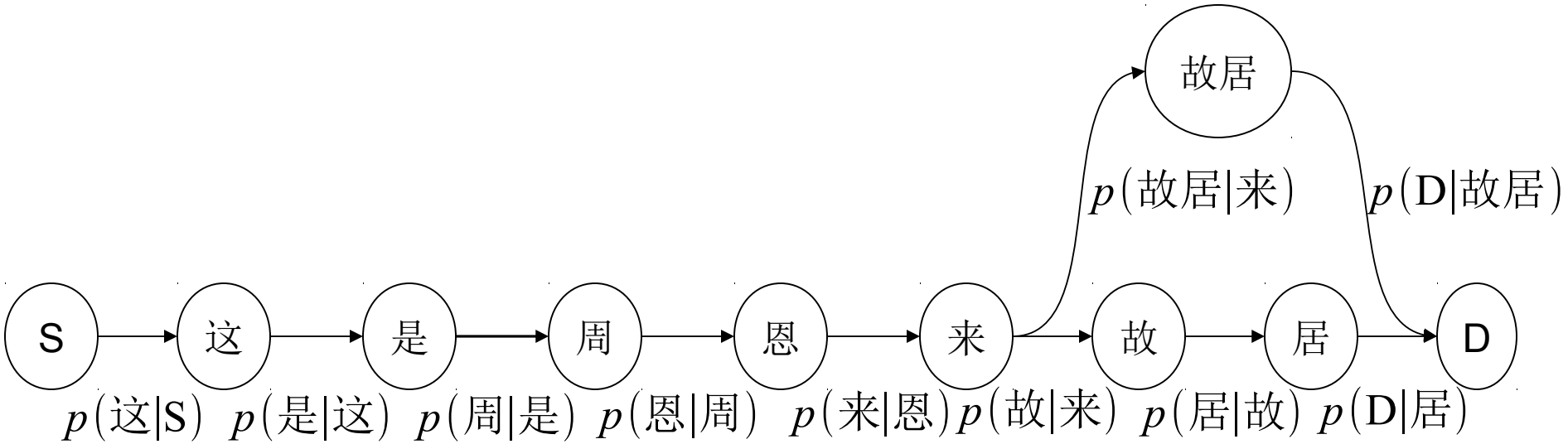
基于角色标注的机构名识别 (3)

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人民日报 1 月	7836	8476	6317	74.5	80.6	77.5
人民日报 6 月	9065	10216	7136	69.9	78.7	74.0

- 训练语料都是《人民日报》 1998 年 1 ～ 5 月语料
- 实验结果分析
  - 角色集合的选取对识别的结果至关重要，要反复尝试
  - 加入机构名识别后人名地名的识别正确率都有所提高

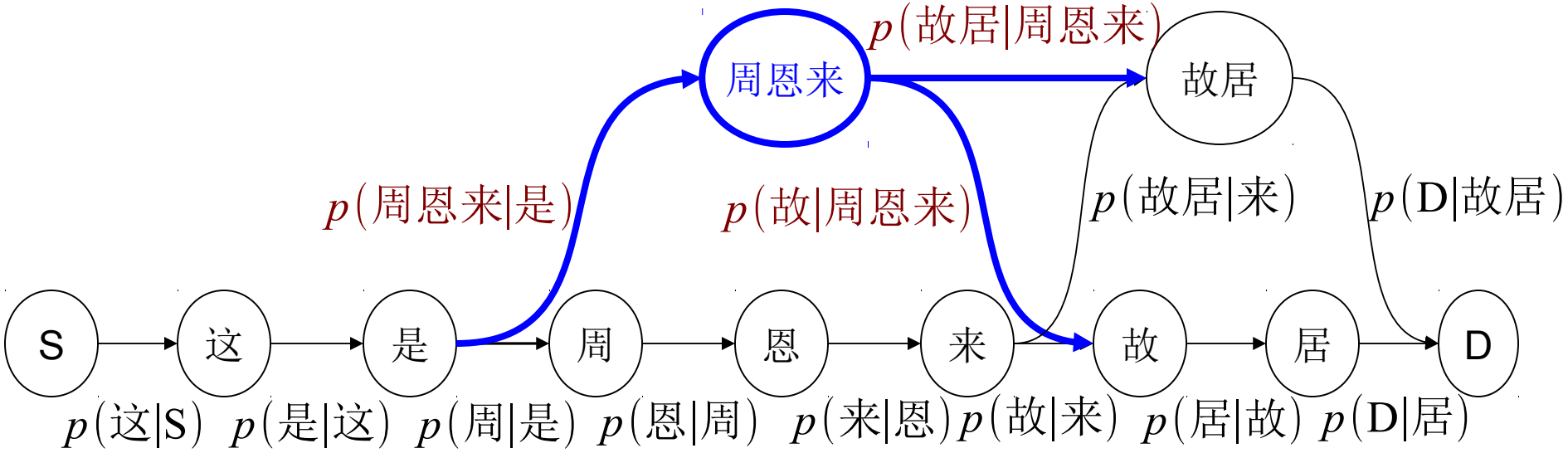
词语切分与未定义词识别的结合

原始词图：



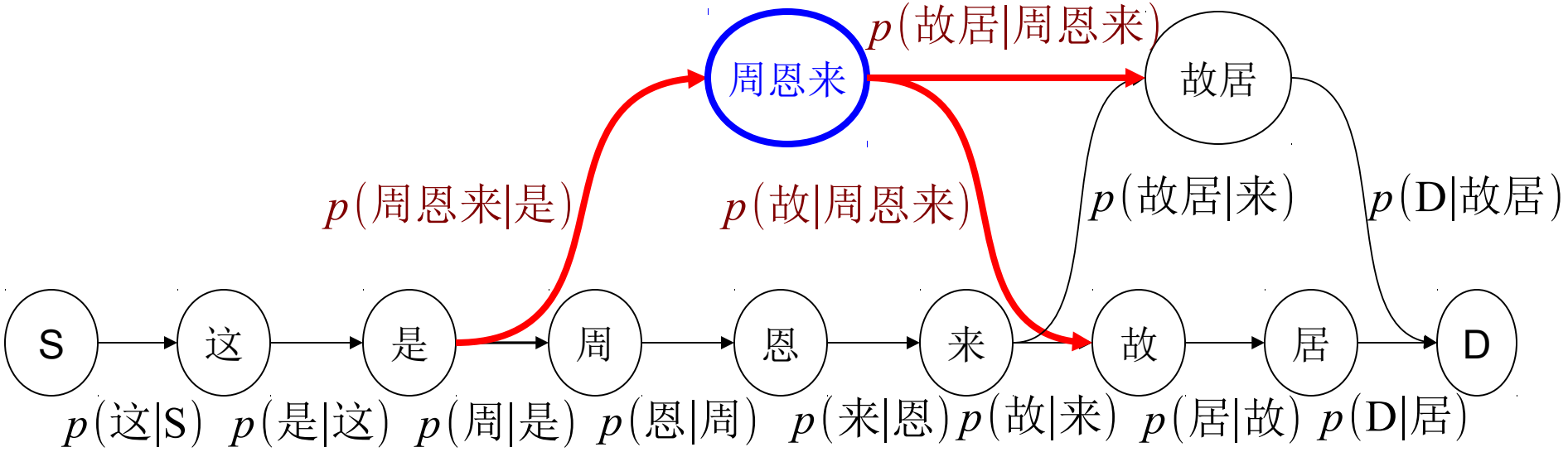
词语切分与未定义词识别的结合

词图上加上识别出的人名：



词语切分与未定义词识别的结合

问题：未定义词相关的 N 元语法概率如何计算？

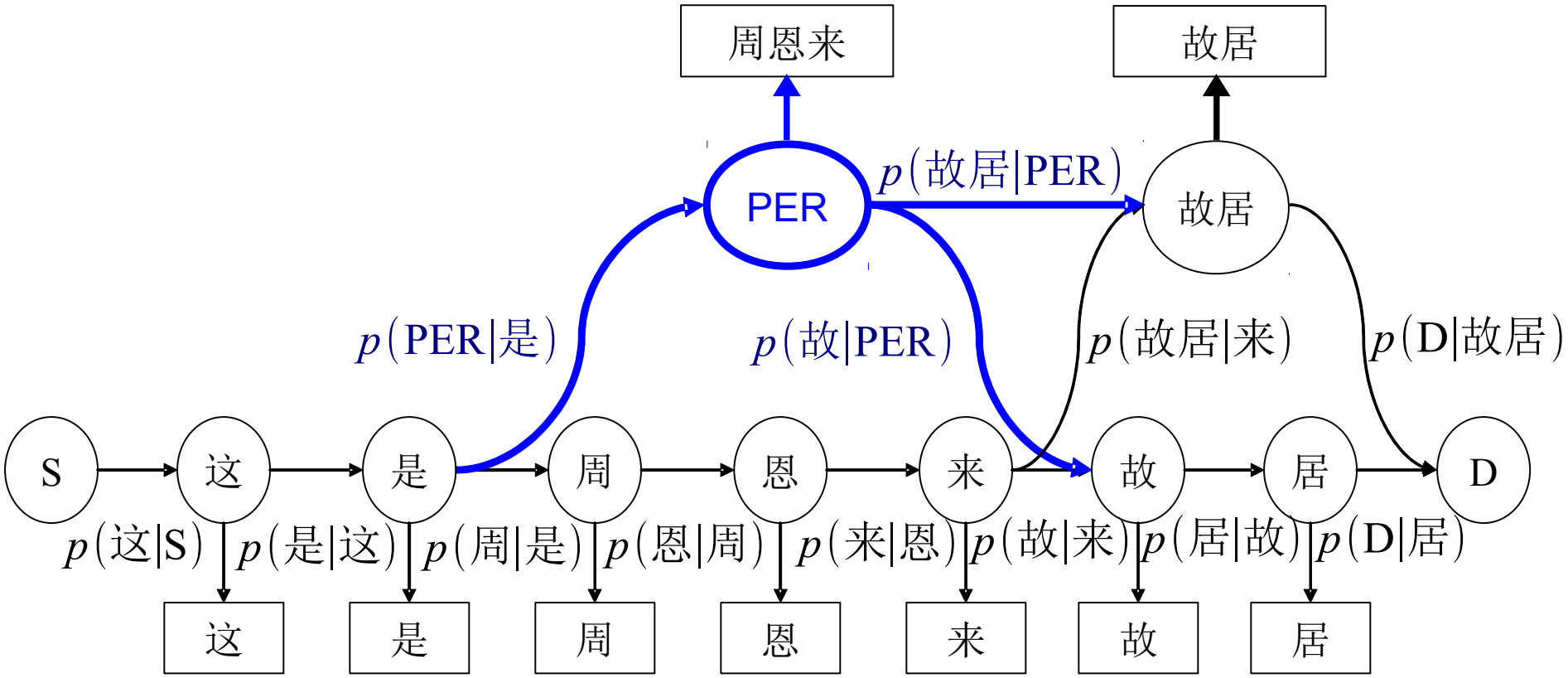


词语切分与未定义词识别的结合

- 解决办法：定义一个隐马尔科夫模型，将词语切分和未定义词识别结合起来
  - 观察值：句子中所有可能的词语，包括未定义词
  - 状态值：
    - 已知词：词语本身
    - 未定义词：未定义词的类型
      - PER: 人名
      - LOC: 地名
      - ORG: 机构名
      - ……

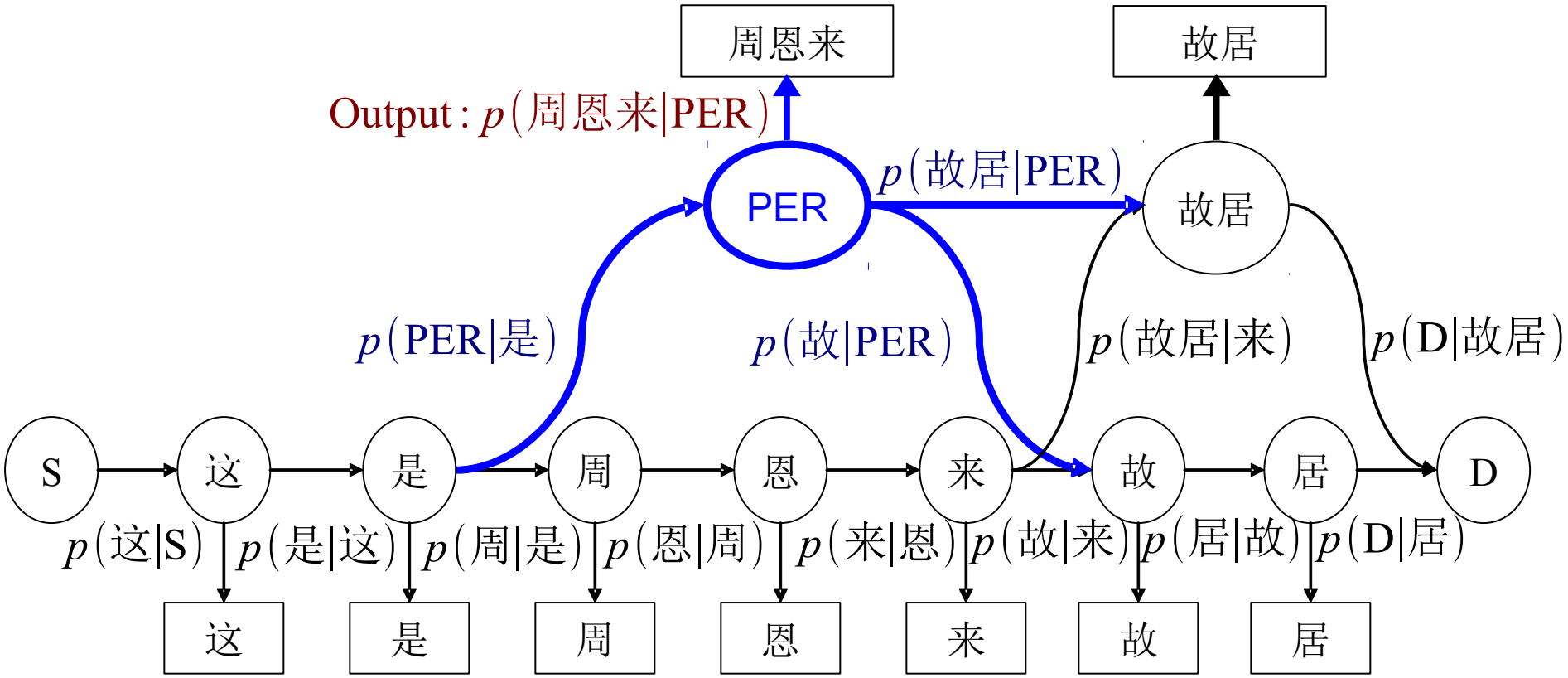
词语切分与未定义词识别的结合

好消息：转移概率可以从语料库统计得到



词语切分与未定义词识别的结合

坏消息：未定义词的输出概率无法从语料库中直接统计得到



基于角色的词语生成模型 (1)

- 未定义词的概率估计问题：如何在高层的 HMM 中估计低层 HMM 中识别出来的未定义词概率，如：*p*(陈鲁豫 | 中国人名)
- 角色隐马尔可夫模型仅仅给出了角色序列的概率，并没有给出未定义词的概率
- 未定义词概率的估计模型：
  基于角色的词语生成模型 (Role WGM)
- Role WGM 和 Role HMM 可以共享角色转移概率矩阵
- 复合未定义词的概率估计，需要用到嵌套的基于角色的词语生成模型（ Role WGM ）

基于角色的词语生成模型 (2)

基于角色的词语生成模型：

给定： $PatternSet(Type)=PT_1,PT_2,\ldots,PT_m$

根据输出独立性假设：

$$p(W|Type)=p(PT_k|Type)p(W|PT_k)=p(PT_k|Type)\prod_{i=1}^{l}p(w_i|t_{ki})$$

归一化： $\sum_{k=1}^{m}p(PT_k|Type)=1$

例子： $p(\text{陈鲁豫}|\text{PER})$

$$=p(\text{BBE}|\text{PER})\,p(\text{陈}|\text{B})\,p(\text{鲁}|\text{B})\,p(\text{豫}|\text{E})$$

词汇化的隐马尔可夫模型 (1)

词汇化的隐马尔可夫模型（ Lexicalized HMM ）

- 定义：一种特定的隐马尔可夫模型
  - 观察值：词语序列（包含词典词和未定义词）
  - 状态值：
    - 词典词：词语本身
    - 未定义词：未定义词的类别
- 搜索：从词图中找到最佳路径

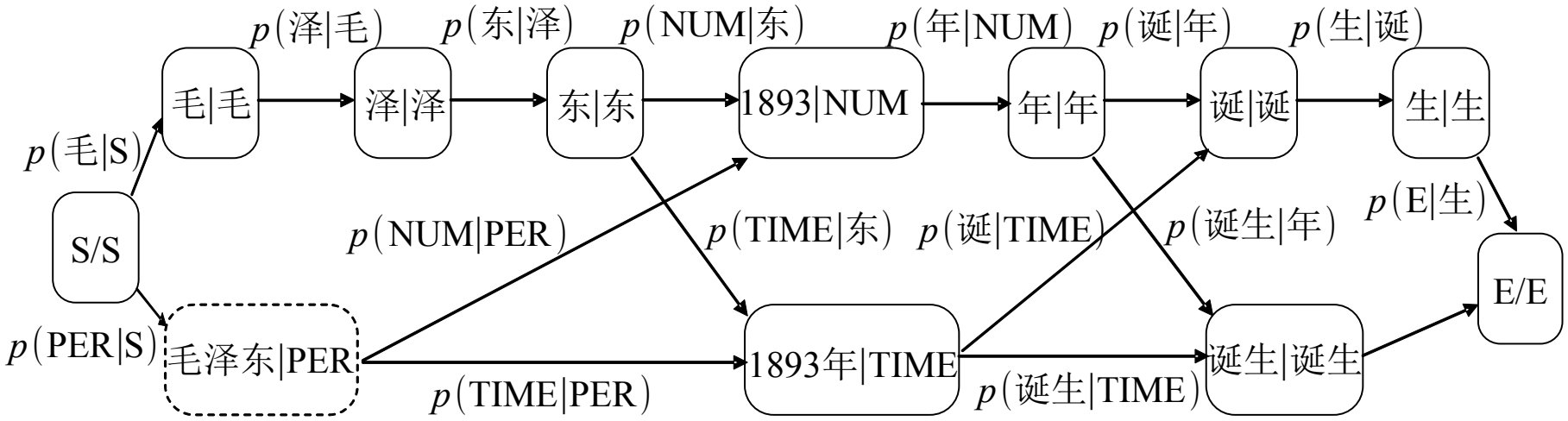
词汇化的隐马尔可夫模型 (2)

词汇化的隐马尔可夫模型的标记定义：

Tag	Description
W_i (the Word)	词典词
PER	人名
LOC	地名
ORG	机构名
NUM	数词
TIME	时间和日期
OTHER	其他未定义词
START	句首
END	句尾

词汇化的隐马尔可夫模型 (3)

词汇化的隐马尔可夫模型的路径搜索：



基于类的二元语法切分词图（原始字串：毛泽东 1893 年诞生）

说明：

1. 节点中表示的是“词语 / 类”（即 wi/ci), 节点的权值为类到词语的概率 p(wi|ci)；
2. 有向边的权值为相邻类的转移概率 p(ci|ci-1); S 为初始节点； E 为结束节点；
3. “毛泽东 /PER” 相关的虚线部分是人名识别 HMM 作用过之后产生的。

基于 HMM 的词性标注

$$\tilde{T} = \underset{T}{\operatorname{argmax}} P(T|W) = \underset{T}{\operatorname{argmax}} P(T)P(W|T)$$

其中：

$$P(T) \approx p(t_1|t_0)p(t_2|t_1)\ldots p(t_n|t_{n-1})$$

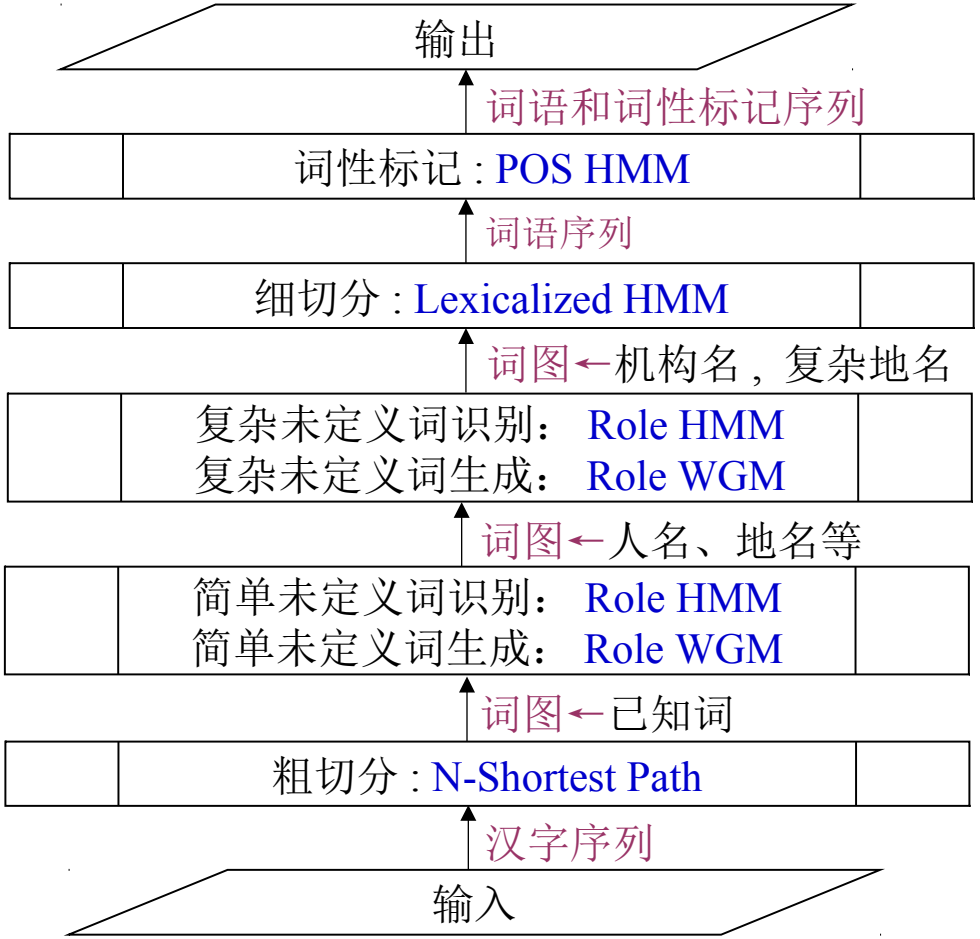
$$P(W|T) \approx p(w_1|t_1)p(w_2|t_2)\ldots p(w_n|t_n)$$

$P(t_i|t_{i-1})$ 直接用词性标记语料库进行估计

$P(w_i|t_i)$ 对于已定义词，直接用词性标记语料库进行估计

对于未定义词，采用前面介绍的 Role WGM 进行估计

层叠隐马尔可夫模型



层叠隐马尔可夫模型（ Cascaded HMM ）框架

ICTCLAS 实验结果 (1)

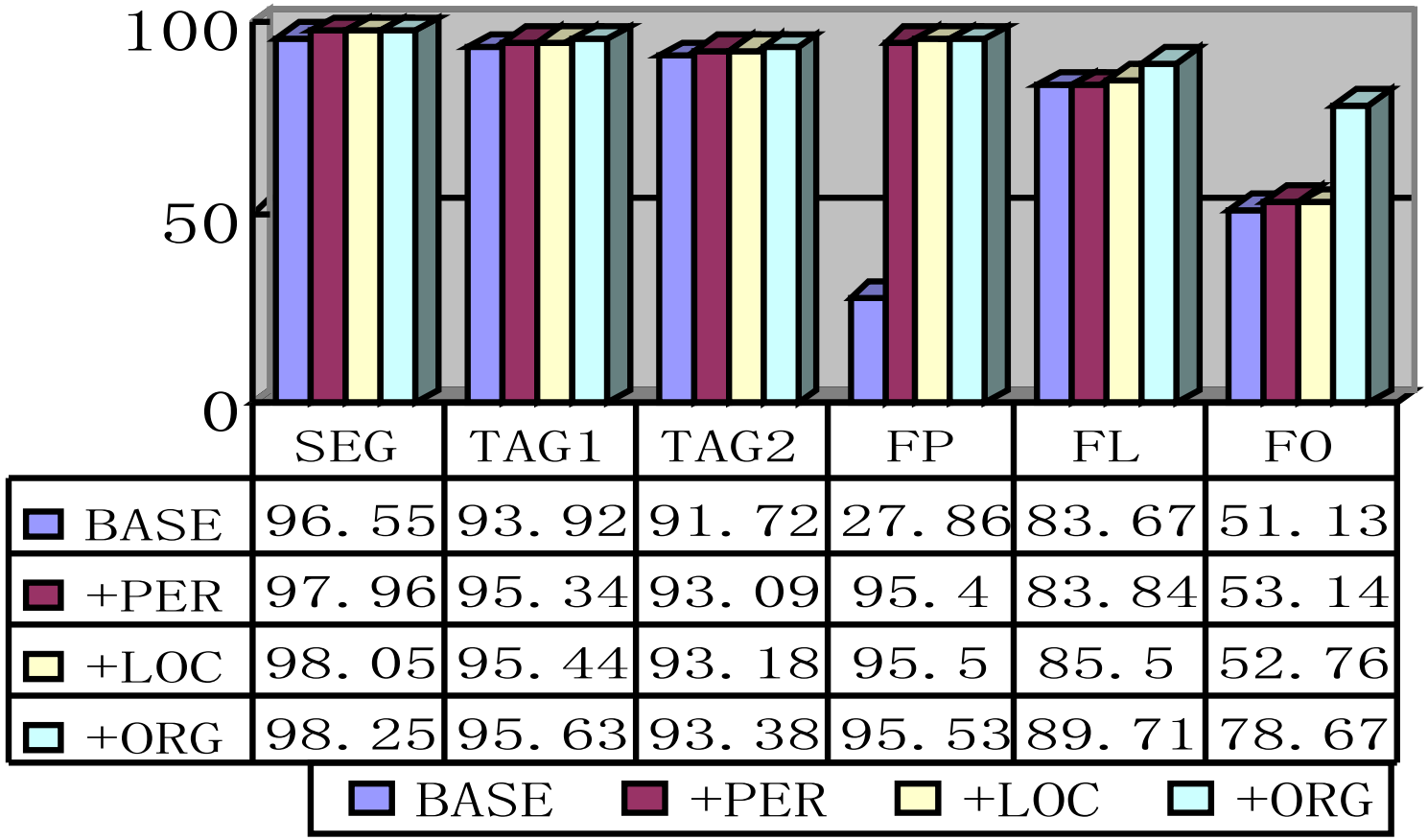
领域	词数	SEG	TAG1	RTAG
体育	33,348	97.01%	86.77%	89.31%
国际	59,683	97.51%	88.55%	90.78%
文艺	20,524	96.40%	87.47%	90.59%
法制	14,668	98.44%	85.26%	86.59%
理论	55,225	98.12%	87.29%	88.91%
经济	24,765	97.80%	86.25%	88.16%
总计	208,213	97.58%	87.32%	89.42%

1) 数据来源：国家 973 英汉机器翻译第二阶段评测的评测总结报告；

2) 标注相对正确率 RTAG=TAG1/SEG*100%

3) 由于我们采取的词性标注集和 973 专家组的标注集有较大出入，所以词性标注的正确率不具可比性。

ICTCLAS 实验结果 (2)



FP: 人名识别 F-Score FL: 地名识别 F-Score FO: 机构名识别 F-Score

ICTCLAS 实验结果 (3)

- 测试结果分析
  - 未定义词识别的准确率相当高
  - 随着人名识别、地名识别和机构名识别的加入，总体性能和各单项性能都稳步提高
  - 系统的各个模块之间达到了互相增益的效果。比如说：当我们在系统中加入机构名识别模块时，地名识别的 F1 值从 85.5% 提高到 89.7%
  - 对于汉语词法分析来说，层叠隐马尔可夫模型是一种有效统计模型

1st SIGHAN Bakeoff (2002)
第一次 SIGHAN 汉语分词比赛

- 完全的自动评测
- 唯一的标准答案
- 提供训练语料，不提供分词规范
- 仅有分词，没有任何标注信息
- 四类语料库：

Corpus	Abbrev.	Encoding	# Train. Words	# Test. Words
Academia Sinica	**AS**	Big Five (MS Codepage 950)	5.8M	12K
U. Penn Chinese Treebank	**CTB**	EUC-CN (GB 2312-80)	250K	40K
Hong Kong CityU	**HK**	Big Five (HKSCS)	240K	35K
Beijing University	**PK**	GBK (MS Codepage 936)	1.1M	17K

- 每一类训练语料分为两个任务
  - 开发训练（ Open ）任务：可以使用任何语料库进行训练
  - 封闭训练（ Closed ）任务：仅可以使用给定的语料库进行训练

ICTCLAS 参加 1st SIGHAN Bakeoff 评测结果（ PK closed ）

Site	word count	R	c_r	P	c_p	F	OOV	R_{oov}	R_{iv}
S01	17,194	0.962	±0.0029	0.940	±0.0036	0.951	0.069	0.724	0.979
S10	17,194	0.955	±0.0032	0.938	±0.0037	0.947	0.069	0.680	0.976
S09	17,194	0.955	±0.0032	0.938	±0.0037	0.946	0.069	0.647	0.977
S07	17,194	0.936	±0.0037	0.945	±0.0035	0.940	0.069	0.763	0.949
S04	17,194	0.936	±0.0037	0.942	±0.0036	0.939	0.069	0.675	0.955
S08	17,194	0.939	±0.0037	0.934	±0.0038	0.936	0.069	0.642	0.961
S06	17,194	0.933	±0.0038	0.916	±0.0042	0.924	0.069	0.357	0.975
S05	17,194	0.923	±0.0041	0.867	±0.0052	0.894	0.069	0.159	0.980

S01 是 ICTCLAS， R 是召回率， P 是准确率， OOV （ out-of-vocabulary ）是未定义词（在测试语料中出现但未在训练语料中出现的词）占测试语料词数的比例， Roov 是未定义词的召回率， Riv 是已定义词（ in-vocabulary ）的召回率。 C*r* 和 C*p* 分别召回率和准确率的 95% 置信区间

ICTCLAS 参加 1st SIGHAN Bakeoff 评测结果（ PK open ）

Site	word count	R	c_r	P	c_r	F	OOV	R_{oov}	R_{iv}
S10	17,194	0.963	±0.0029	0.956	±0.0031	0.959	0.069	0.799	0.975
S01	17,194	0.963	±0.0029	0.943	±0.0035	0.953	0.069	0.743	0.980
S08	17,194	0.939	±0.0037	0.938	±0.0037	0.938	0.069	0.675	0.959
S04	17,194	0.933	±0.0038	0.942	±0.0036	0.937	0.069	0.712	0.949
S03	17,194	0.940	±0.0036	0.911	±0.0043	0.925	0.069	0.647	0.962
S11	17,194	0.905	±0.0045	0.869	±0.0051	0.886	0.069	0.503	0.934

S01 是 ICTCLAS， R 是召回率， P 是准确率， OOV （ out-of-vocabulary ）是未定义词（在测试语料中出现但未在训练语料中出现的词）占测试语料词数的比例， Roov 是未定义词的召回率， Riv 是已定义词（ in-vocabulary ）的召回率。 *Cr* 和 *Cp* 分别召回率和准确率的 95% 置信区间

复习思考题

- 到“中文自然语言处理开放平台（ http://www.nlp.org.cn ）”上去下载 ICTCLAS 开放版本的源代码，研究并试图改进该系统。
- 研究汉语切分中变形词的概率计算问题，提出合理的解决办法。
- 研究如何利用未定义词的重复出现规律来提高未定义词识别的正确率和召回率。
- 研究汉语词语切分和词性标注的更有效的一体化概率模型。