

第三章 词性标注

王峰

华东师大计算机系

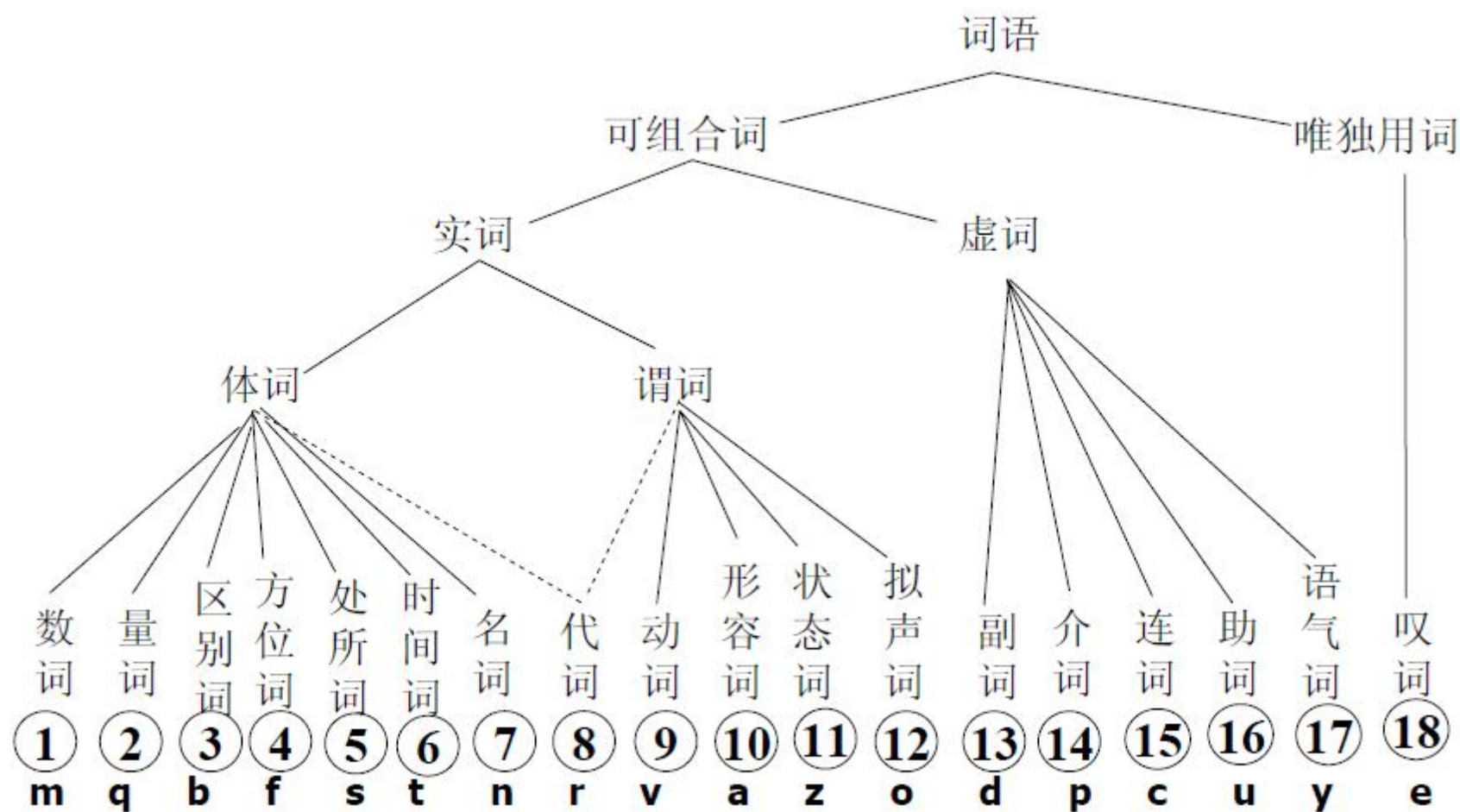
主要内容

- 词性标注概述
- 词性自动标注的方法
 - 基于规则
 - 基于统计
- 小结

1. 词性标注概述

- 分词：判断词的结构，对自然语言形态分析
- 词类：一个语言全部词汇的子类划分
 - Part of speech (POS), Word class
- 词性标注 (POS Tagging): 判断词的类别，是词汇最重要的特性，是连接词汇到句法的桥梁

现代汉语的词类系统



词性标注概述

- 词性标注的主要任务是消除 *词性兼类歧义*。
- 在任何一种自然语言中，词性兼类问题都普遍存在。

例如：(p:介词； q:量词； m:数词； c:连词； f:方位词； r:代词)

1) Time flies like an arrow.

Time/*n-v* flies/*v-n* like/*p-v* an/*det* arrow/*n*.

2)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*

3) 两 把 锁 锁 两次

一件 制服 制服 不了小偷

两 朵 花 花 时间

汉语中兼类词的比例

兼类数	兼类词数	百分比	例词及词性标记
5	3	0.01%	和 c-n-p-q-v
4	20	0.04%	光 a-d-n-v
3	126	0.23%	画 n-q-v
2	1475	2.67%	锁 n-v
合计	1624	2.94%	总词数: 55191

数据来源：北大计算语言所《现代汉语语法信息词典》1997年版

兼类	词数	百分比	例词
n-v	613	42%	爱好，把握，报道
a-n	74	5%	本分，标准，典型
a-v	217	15%	安慰，保守，抽象
b-d	103	7%	长期，成批，初步
n-q	64	4%	笔，刀，口
a-d	30	2%	大，老，真
合计	1101	75%	兼两类词数: 1475

兼类词在实际语料中分布示例

词	词性1: 概率	词性2: 概率	词性3: 概率	词性4: 概率
把	p: 0.96	q: 0.03	v: 0.01	m: 0.00
被	p: 1.00	Ng: 0.00		
并	c: 0.86	d: 0.14		
次	q: 1.00	Bg: 0.00		
从	p: 1.00	Vg: 0.00		
大	a: 0.92	d: 0.08		
到	v: 0.80	p: 0.20		
得	u: 0.76	v: 0.24	e: 0.00	
等	u: 0.98	v: 0.02	q: 0.00	
地	u: 0.89	n: 0.11		
对	p: 0.98	v: 0.01	q: 0.01	a: 0.00
就	d: 0.87	p: 0.13	c: 0.00	
以	p: 0.84	c: 0.11	j: 0.05	
由	p: 1.00	v: 0.00		
在	p: 0.95	d: 0.02	v: 0.02	

兼类词在实际语料中分布

SPAN	1	2	3	4	5	6	7	8	9	10	11
#	2043	898	377	202	83	39	21	10	2	1	
%	55.58	24.43	10.26	5.50	2.26	1.06	0.57	0.27	0.05	0.05	0.03
+%	55.58	80.01	90.27	95.77	98.03	99.09	99.66	99.93	99.98	100.0	100.0

刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，182页

- 兼类词在汉语词汇中所占比例较小；
- 常用词兼类比例较高；
- 大部分兼类词所兼词类是使用频度较高的主要词性。

英语词的兼类现象

10.4 percent of the lexicon is ambiguous as to part-of-speech (types).
40 percent of the words in the Brown corpus are ambiguous (tokens).

引自 <http://www.cs.columbia.edu/~becky/cs4999/04mar.html>

Degree of ambiguity	Total frequency (39440)
1 tag	35340
2	3760
3	263
4	61
5	12
6	2
7	1

数据来源：
Brown 语料库

兼类词串在语料中的分布统计（英语）

Span	Frequency	Span	Frequency
3	397,111	11	382
4	143,447	12	161
5	60,224	13	58
6	26,515	14	29
7	11,409	15	14
8	5,128	16	6
9	2,161	17	1
10	903	18	0
		19	1

数据来源：
Brown 语料库

引自 <http://www.cs.columbia.edu/~becky/cs4999/span-lengths.html>

词性标记集

- 标注集的确定原则
 - 不同语言中，词性划分基本上已经约定俗成。
 - 自然语言处理中对词性标记要求相对细致。
- 一般原则：
 - 标准性: 普遍使用和认可的分类标准和符号集；
 - 兼容性: 与已有资源标记尽量一致，或可转换；
 - 可扩展性: 扩充或修改。

汉语词类归属测试（调查）网页 <http://ccl.pku.edu.cn:8080/pos/>

现代汉语语法基础知识网页

<http://ccl.pku.edu.cn/course/xdhyjs/question.asp>

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》 标注语料库词性标记集

标记集中共有**106**个代码

✓ **26**个基本词类代码，

✓ **74**个扩充代码。

在处理真实语料的时候，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、简称略语等比词大的单位；

也包含一些标记，用于标注语素、前接成份、后接成份等比词小的单位。

分级词性标记集

	第一级	第二级	第三级	说明
数量	26	48	106	
标记	a			
	b			
	c			
	...			
	n	n	n	名词
		nr	nr	人名
			nr ^f	姓
			nr ^g	名
		ns		...
		nt		
		nz		
	...			
	v	v	v	动词
		vd	vd	副动词
		vn	vn	名动词
			vu	助动词
			vx	形式动词
	
	z

语法学界对汉语词类的认识还有不够清晰的地方。汉语词类的划分标准，词类数量的多寡，词类之间的关系等等，都还存在争议。

少 ← 词性标记 → 多
粗 ← 语法特征 → 细
简 ← 语法规则 → 繁

2. 词性自动标注

词性自动标注的方法

序号	作者	标记集	方法	标注效率	处理语料规模	精确率
1	Klein&Simmons (1963)	30	人工规则	-	百科全书样本	90%
2	TAGGIT (Greene&Rubin, 1971)	86	人工规则	-	Brown语料库	77%
3	CLAWS (Marshall,1983; Booth, 1985)	130	概率法	低	LOB语料库	96%
4	VOLSUNGA (DeRose,1988)	97	概率法	高	Brown语料库	96%
5	Eric Brill's tagger (1992-94)	48	机器规则	高	Upenn WSJ语 料库	97%
.....						

词性自动标注的方法

- 基于规则的词性标注方法
 - 人工规则
 - 机器学习规则
- 基于统计模型的词性标注方法
- 规则和统计方法相结合的词性标注方法
- 基于有限状态变换机的词性标注方法
- 基于神经网络的词性标注方法
-

基于规则的词性标注方法

- **TAGGIT 词性标注系统 (Brown University)**
 - 86 种词性，3300 规则
 - 手工编写词性歧义消除规则
- **山西大学的词性标注系统 [刘开瑛，2000]**
 - 手工编写消歧规则
 - 建立非兼类词典
 - 建立兼类词典
 - 词性可能出现的概率高低排列
 - 构造兼类词识别规则

基于规则的词性标注示例

@@ 信($n-v$)

- **CONDITION FIND(L,NEXT,X){%X.yx=的|封|写|看|读}**
SELECT n
- OTHERWISE SELECT v**

@@ 一边($c-s$)

- **CONDITION FIND(LR,FAR,X) {%X. yx=一边} SELECT c**
OTHERWISE SELECT s

基于规则的词性标注示例

(1) 并列鉴别规则

如：体现了人民的要求(n/v?)和愿望(n, 非兼类)

(2) 同境鉴别规则

如：一个优秀的企业必须具备一流的产品(n, 非兼类)、一流的管理(n/v?)和一流的服务(n/v?)。

(3) 区别词鉴别规则(区别词只能直接修饰名词)

如：他们搞的这次大型(区别词, 非兼类) 调查(v/n?)历时半年。

(4) 唯名形容词鉴别规则(有些形容词只能直接修饰名词)

如：重大(唯名形容词) 损失(n/v?)

巨大(唯名形容词) 影响(n/v?)

机器学习规则的词性标注方法

- 基于转换的错误驱动的机器学习方法
- **Eric Brill (1992,1995), Transformation-based error-driven part of speech tagging**

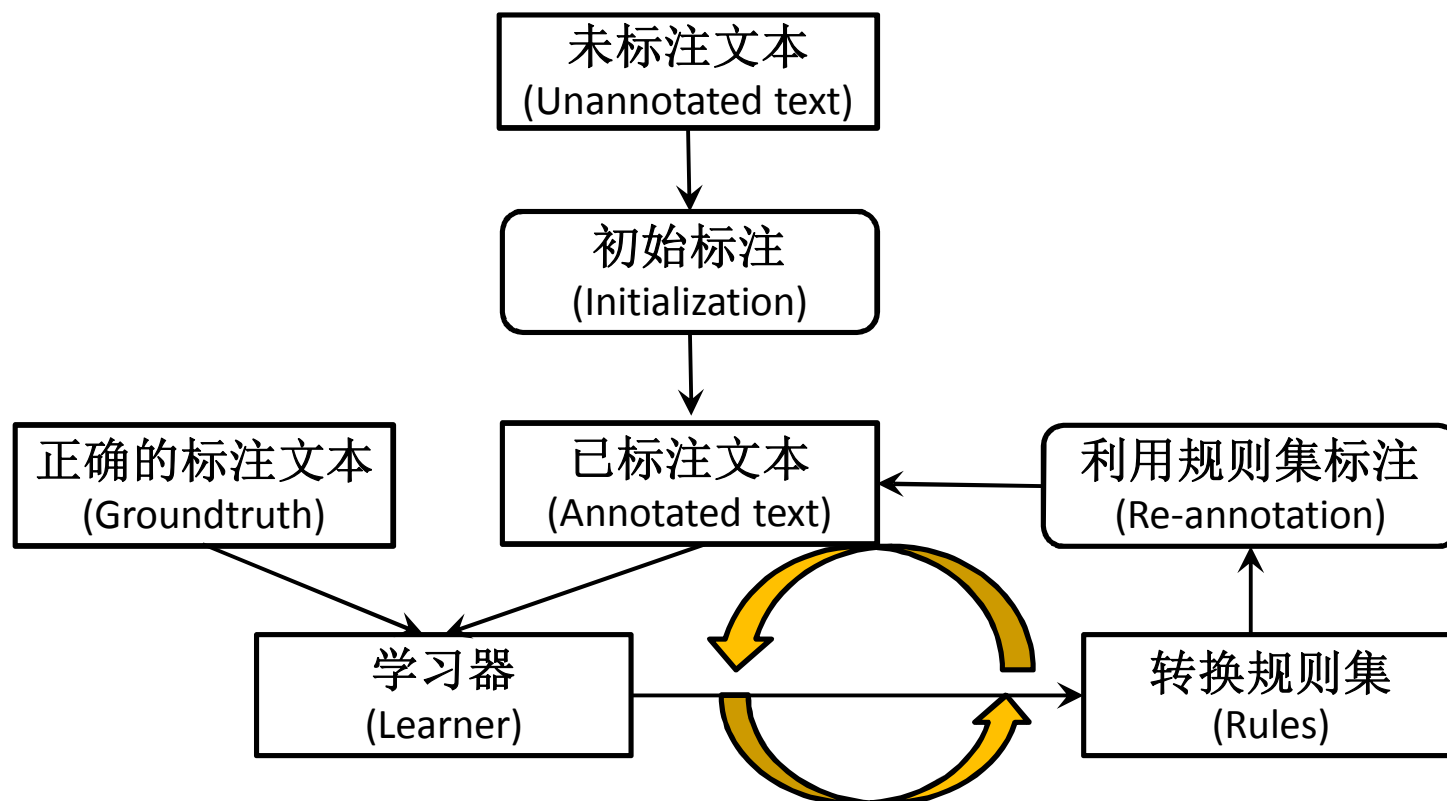
基本思想：

1. 正确结果是通过不断修正错误得到的
2. 修正错误的过程是有迹可循的
3. 让计算机学习修正错误的过程，这个过程可以用转换规则（**transformation**）形式记录下来，然后用学习得到的转换规则进行词性标注

下载Brill's tagger: http://en.wikipedia.org/wiki/Brill_tagger

基于转换的错误驱动的词性标注方法

- 错误驱动的机器学习方法
 - 初始词性赋值
 - 对比正确标注的句子，自动学习结构转换规则
 - 利用转换规则调整初始赋值



转换规则的形式

- 转换规则由两部分组成
 - 改写规则（**rewriting rule**）
 - 激活环境（**triggering environment**）
 - 一个例子：转换规则**T1**
 - 改写规则：将一个词的词性从动词（**v**）改为名词（**n**）；
 - 激活环境：
 - 该词左边第一个紧邻词的词性是量词（**q**），
 - 第二个词的词性是数词（**m**）
- S0: 他/r 做/v 了/u 一/m 个/q 报告/v**
(运用T1)
S1: 他/r 做/v 了/u 一/m 个/q 报告/n

转换规则的模板

- 改写规则：将词性标记 x 改写为 y
- 激活环境：
 - (1) 当前词的前（后）面一个词的词性标记是 z ；
 - (2) 当前词的前（后）面第二个词的词性标记是 z ；
 - (3) 当前词的前（后）面两个词中有一个词的词性标记是 z ；

.....

（其中 x ， y ， z 是任意的词性标记代码）

If $t_{-1}==z$, then $x \rightarrow y$;

If $t_{-1}==q$, $t_{-2}==m$, then $v \rightarrow n$;

.....

根据模板可能学到的转换规则

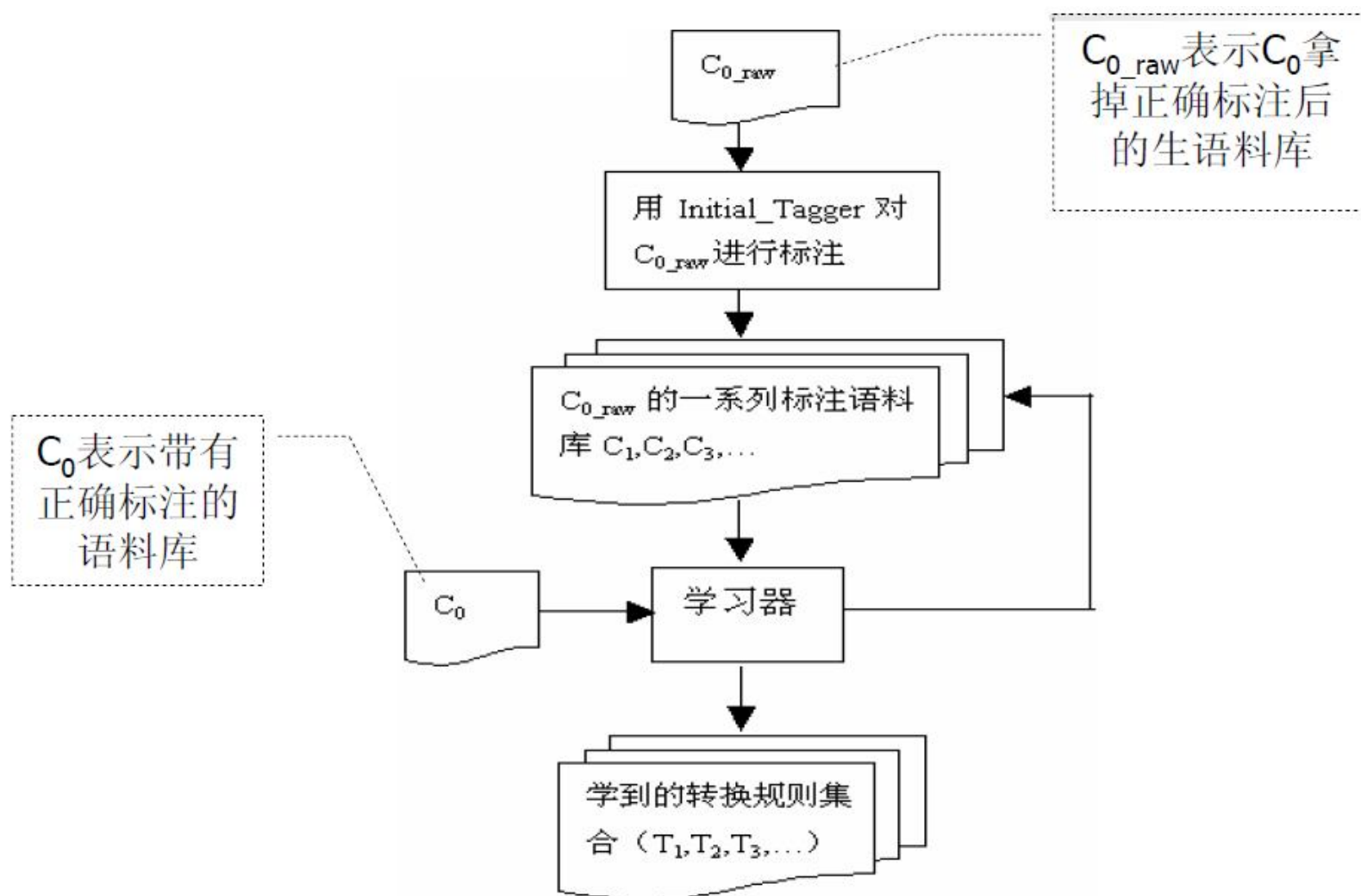
T1: 当前词的前一个词的词性标记是量词（**q**）时，将当前词的词性标记由动词（**v**）改为名词（**n**）；

T2: 当前词的后一个词的词性标记是动词（**v**）时，将当前词的词性标记由动词（**v**）改为名词（**n**）；

T3: 当前词的后一个词的词性标记是形容词（**a**）时，将当前词的词性标记由动词（**v**）改为名词（**n**）；

T4: 当前词的前面两个词中有一个词的词性标记是名词（**n**）时，将当前词的词性标记由形容词（**v**）改为数词（**m**）；

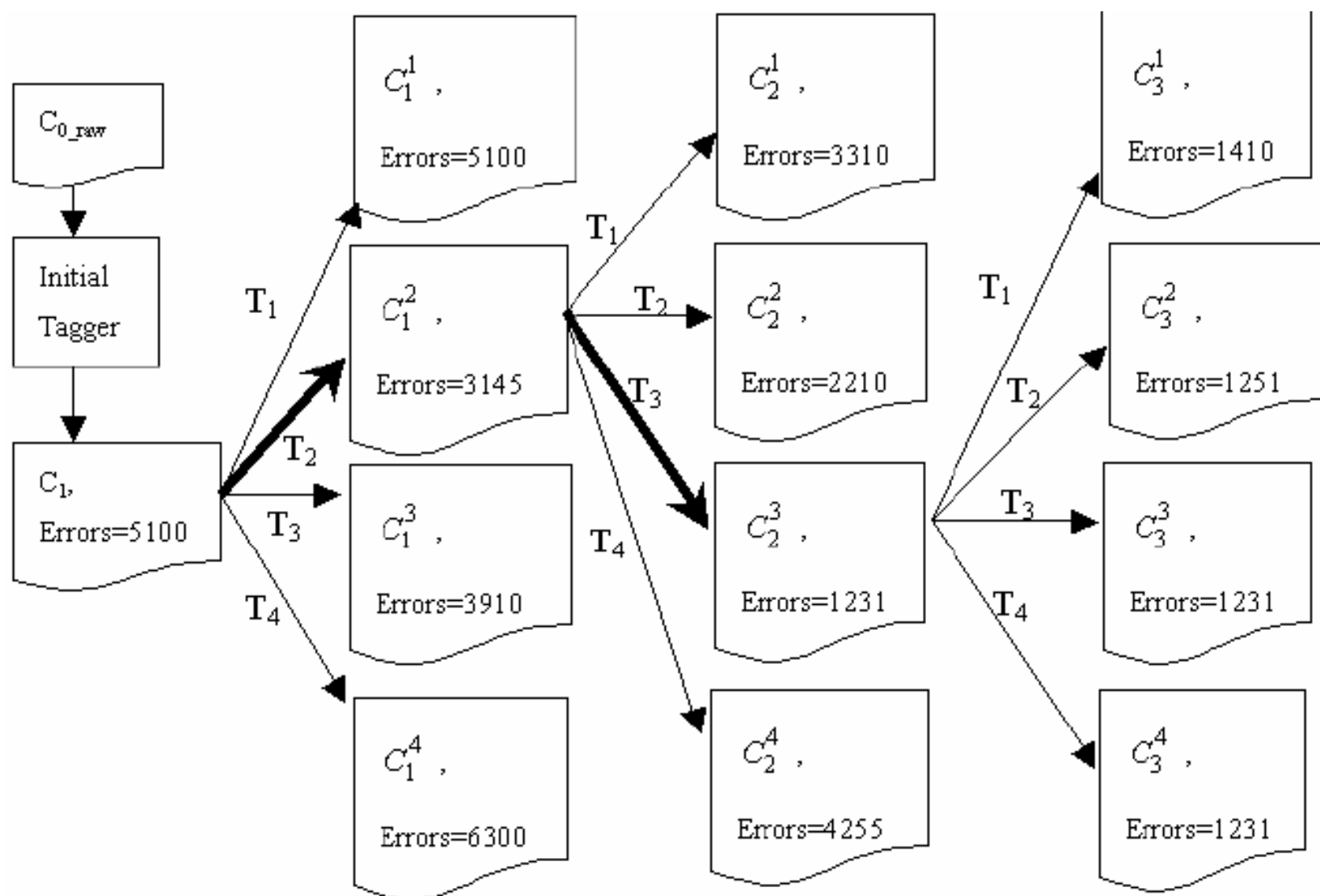
转换规则的学习流程



转换规则学习器算法描述

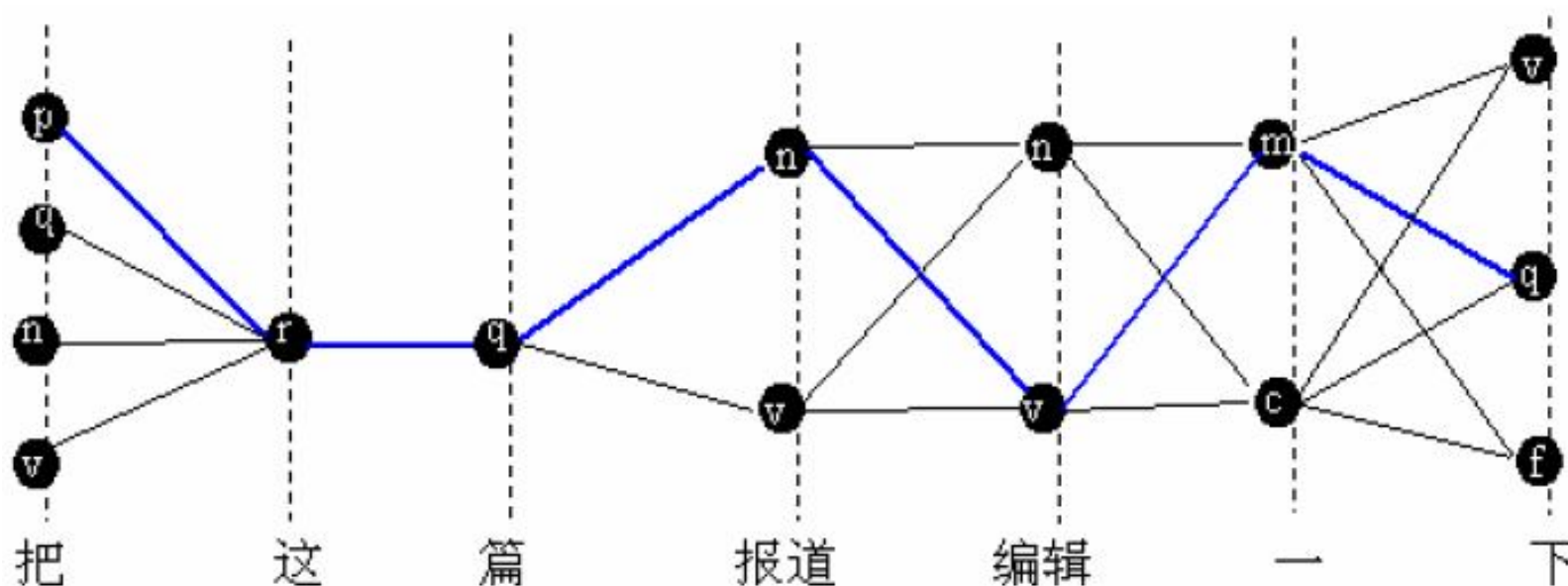
- 1) 首先用初始标注器对 C_{0_raw} 进行标注, 得到带有词性标记的语料 $C_i (i=1)$;
- 2) 将 C_i 跟正确的语料标注结果 C_0 比较, 可以得到 C_i 中总的词性标注错误数;
- 3) 依次从候选规则中取出一条规则 $T_m (m=1,2,...)$, 每用一条规则对 C_i 中的词性标注结果进行一次修改, 就会得到一个新版本的语料库, 不妨记做 $C_i^m (m=1,2,3,...)$, 将每个 C_i^m 跟 C_0 比较, 可计算出每个 C_i^m 中的词性标注错误数。假定其中错误数最少的那个是 C_i^j (可预期 C_i^j 中的错误数一定少于 C_i 中的错误数), 产生它的规则 T_j 就是这次学习得到的转换规则; 此时 C_i^j 成为新的待修改语料库, 即 $C_i = C_i^j$ 。
- 4) 重复第3步的操作, 得到一系列的标注语料库 $C_2^k, C_3^l, C_4^m, ...$ 后一个语料库中的标注错误数都少于前一个中的错误数, 每一次都学习到一条令错误数降低最多的转换规则。直至运用所有规则后, 都不能降低错误数, 学习过程结束。这时得到一个有序的转换规则集合 $\{T_a, T_b, T_c, ... \}$

转换规则学习示例



基于统计模型的词性标注方法

词性标注问题：寻找最优路径



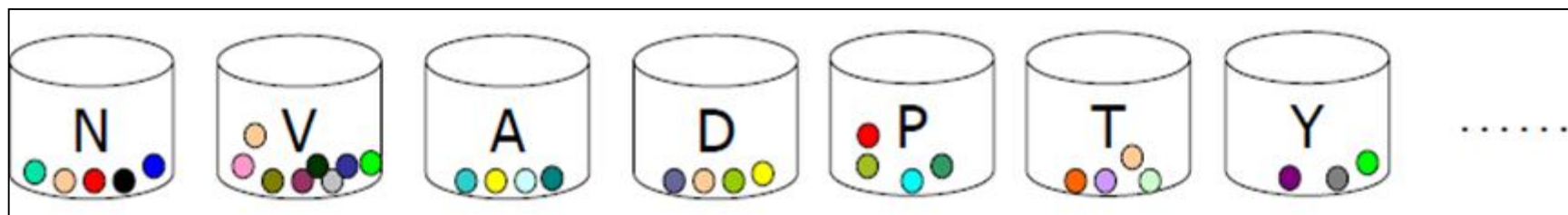
$4 \times 1 \times 1 \times 2 \times 2 \times 2 \times 3 = 96$ 种可能性，哪种可能性最大？

隐马尔可夫模型 (Hidden Markov Model)

- Andrei Andreyevich Markov (1856-1922)
<http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Markov.html>
- 有关马尔可夫过程(Markov Process)、隐马尔可夫模型 (Hidden Markov Model) 更详细的介绍，参见：
 - 陈小荷 《现代汉语自动分析》，北京语言文化大学出版社，2000，第10章。

HMM的罐子比喻 (L.R.Rabiner, 1989)

放有彩色球的罐子，每个罐子都有编号，上帝随机地从罐子中摸出彩球



可观察序列

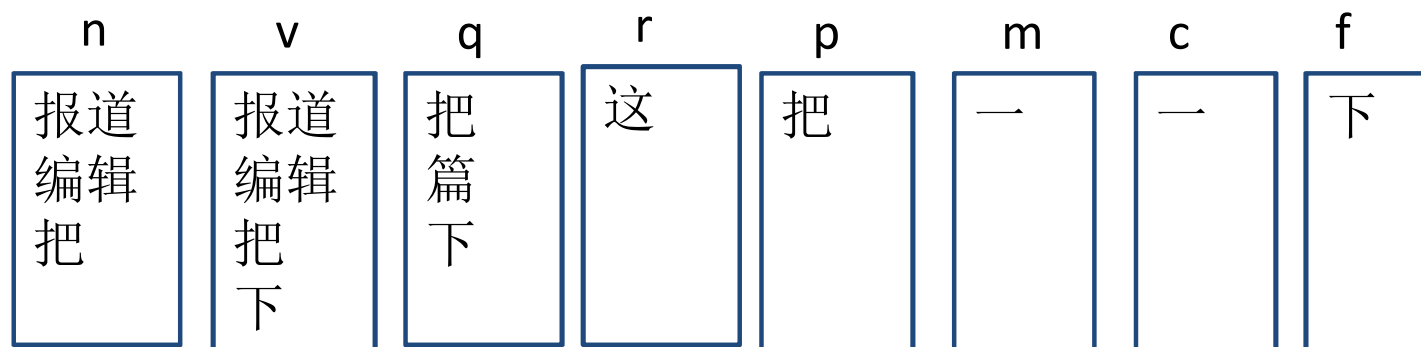


猜测隐藏在
幕后的罐子
序列

D	P	T	N	Y
A	N	D	V	V
A	P	V	N	V
.....				



HMM的罐子比喻



- 观察到的序列: 把/ 这/ 篇/ 报道/ 编辑/ 一/ 下/
- 隐藏的罐子序列:
n - r - q - n - n - m - q
n - r - q - v - v - c - q
v - r - q - v - v - c - f
q - r - q - v - n - c - f
.....

基于HMM进行词性标注

两个随机过程：

1. 选择罐子 — 上帝按照一定的转移概率随机地选择罐子
2. 选择彩球 — 上帝按照一定的概率随机地从一个罐子中选择一个彩球输出

人只能看到彩球序列（词汇序列，记做 $\mathbf{W} = w_1 w_2 \cdots w_n$ ），需要去猜测罐子序列（隐藏在幕后的词性标记序列，记做 $\mathbf{T} = t_1 t_2 \cdots t_n$ ）

已知词串 \mathbf{W} （观察序列）情况下，求使得条件概率 $P(\mathbf{T}|\mathbf{W})$ 值最大的那个 \mathbf{T}' ，一般记做：

$$\mathbf{T}' = \arg \max_{\mathbf{T}} P(\mathbf{T}|\mathbf{W})$$

基于HMM进行词性标注

- 根据条件概率公式可得

$$P(T|W) = \frac{P(T,W)}{P(W)} = \frac{P(T)P(W|T)}{P(W)} \quad (1)$$

- 可以进一步简化为:

$$P(T|W) \approx P(T)P(W|T) \quad (2)$$

- 其中

$$\begin{aligned} P(T) &= P(t_1, t_2, \dots, t_n) \\ &= P(t_1)P(t_2|t_1) \cdots P(t_n|t_{n-1}, t_{n-2}, \dots, t_1) \end{aligned} \quad (3)$$

- 根据一阶HMM的独立性假设, 可得

$$P(T) \approx P(t_1)P(t_2|t_1) \cdots P(t_n|t_{n-1}) \quad (4)$$

- 词性之间的转移概率可以从语料库中估算得到:

$$P(t_i|t_{i-1}) = \frac{\text{训练语料中 } t_i \text{ 出现在 } t_{i-1} \text{ 之后的次数}}{\text{训练语料中 } t_{i-1} \text{ 出现总次数}} \quad (5)$$

基于HMM进行词性标注

- **$P(W|T)$** 是已知词性标记串，产生词串的条件概率

$$\begin{aligned} P(W|T) &= P(w_1, w_2, \dots, w_n | t_1, t_2, \dots, t_n) \\ &= P(w_1 | t_1) P(w_2 | t_2, t_1, w_1) \cdots P(w_n | t_n, \dots, t_1, w_{n-1}, \dots, w_1) \end{aligned} \quad (6)$$

- 根据**HMM**的独立性假设，公式**6**可简化为：

$$P(W|T) \approx P(w_1 | t_1) P(w_2 | t_2) \cdots P(w_n | t_n) \quad (7)$$

- 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：

$$P(w_i | t_i) = \frac{\text{训练语料中 } w_i \text{ 的词性被标记为 } t_i \text{ 的次数}}{\text{训练语料中 } t_i \text{ 出现的总次数}} \quad (8)$$

基于HMM进行词性标注示例

$$P(T|W) \approx P(T)P(W|T) \\ = P(t_1)P(w_1|t_1) * P(t_2|t_1)P(w_2|t_2) \cdots P(t_n|t_{n-1})P(w_n|t_n)$$

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

$$P(T_1|W) = P(q)P(\text{把}|q)*P(r|q)P(\text{这}|r)...P(f|m)P(\text{下}|f)$$

$$P(T_2|W) = P(q)P(\text{把}|q)*P(r|q)P(\text{这}|r)...P(q|m)P(\text{下}|q)$$

$$P(T_3|W) = P(q)P(\text{把}|q)*P(r|q)P(\text{这}|r)...P(v|m)P(\text{下}|v)$$

.....

$$P(T_{96}|W) = P(n)P(\text{把}|n)*P(r|q)P(\text{这}|r)...P(v|c)P(\text{下}|v)$$

从中选出一个最大值

初始状态概率分布

词语输出概率

词性转移概率

估算HMM的参数

- 从语料库估算用于词性标注的HMM的参数：
 - (1) 初始状态的概率分布
 - (2) 词性转移概率
 - (3) 已知词性条件下词语的输出概率

词性转移矩阵(用于估算转移概率)

Tag	c	f	m	n	p	q	r	v
c	736	700	3971	43250	9253	53	7776	40148
f	900	475	4569	7697	2968	278	1290	26951
m	547	1470	17505	46001	1722	139653	305	13778
n	55177	50571	27918	277181	43023	404	9769	221776
p	47	2664	14131	78251	3363	142	27249	36807
q	732	7845	4506	52310	2451	176	760	13288
r	2055	1225	12820	43953	11229	7681	3572	53391
V	13715	14843	70914	221796	44651	3226	46697	191967

词性	频次
c	168350
f	110878
m	270381
n	1539367
p	269186
q	155374
r	214942
v	1193317

$$a(c \rightarrow f) = P(f | c) = \frac{P(cf)}{P(c)} = \frac{700}{168350}$$

词语 | 词性频度表 (用于估算输出概率)

词语	词性	频次		词语	词性	频次	词性	频次
把	p	9877		编辑	n	243	c	168350
把	q	290		编辑	v	100	f	110878
把	n	2		一	m	20672	m	270381
把	v	208		一	c	2229	n	1539367
这	r	21990		下	f	6313	p	269186
篇	q	706		下	q	161	q	155374
报道	v	4040		下	v	2271	r	214942
报道	n	420					v	1193317

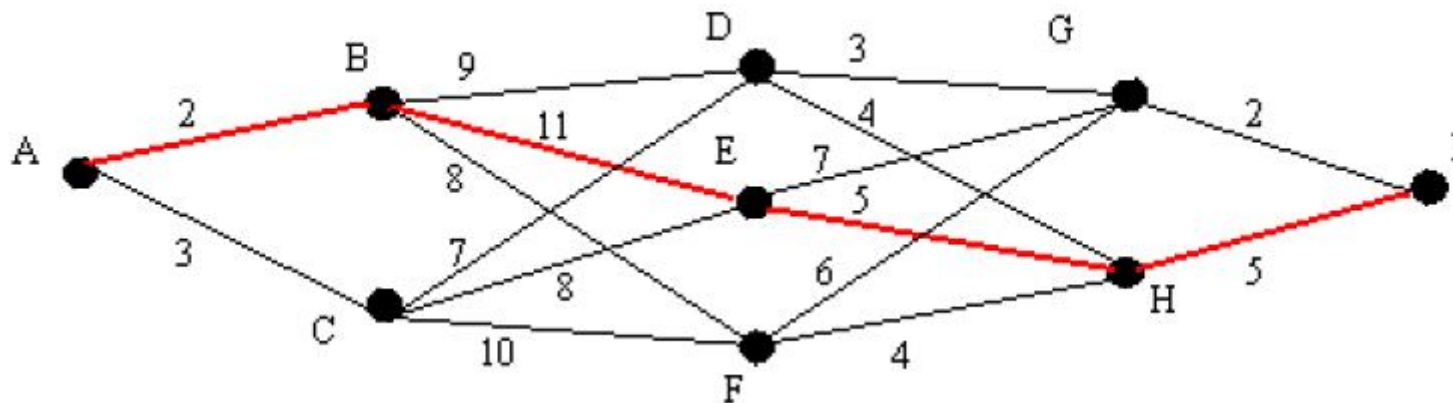
$$(把|p) = \frac{9877}{269186}$$

效率问题

- 假定有 N 个词性标记（罐子），给定词串中有 M 个词（彩球） 考虑最坏的情况：每个词都有 N 个可能的词性标记，则可能的状态序列有 N^M 个
- 随着 M （词串长度）的增加，需要计算的可能路径数目以指数方式增长，即算法复杂性为指数级
- 需要寻找更有效的算法.....

Viterbi算法 — 提高效率

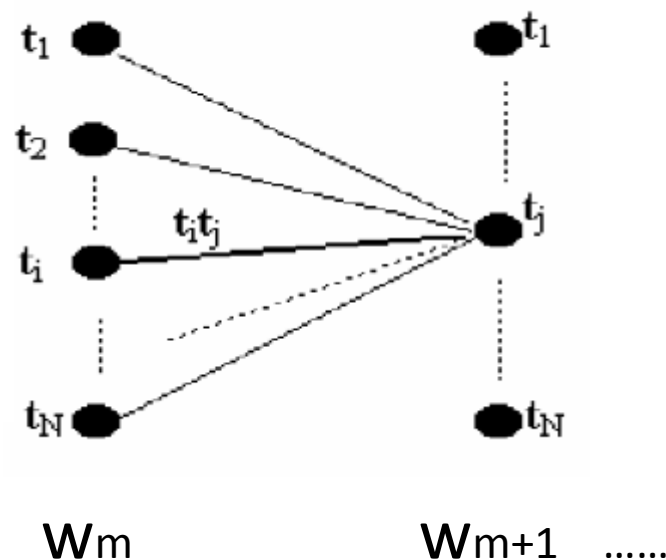
- Viterbi算法是一种动态规划方法（dynamic programming）
- 如果当前节点在最优路径上，那么，不管当前节点的后续路径如何，当前节点的来源路径必定是最优的。
- 最优路径的求解可以迭代进行。
- **12条路径**（ $1 \times 2 \times 3 \times 2 \times 1 = 12$ ），哪条路权重最重？



动态规划示例

词性标记局部路径示意

- 假定一个词串 W 中每个词都有 N 个词性标记，那么从词串中第 m 个词 (w_m) 到第 $m+1$ 个词 (w_{m+1}) 的第 j 个词性标记就有 N 条可能的路径。这 N 条路径中存在一条 概率最大的路径，假定为 $t_i t_j$



$W:$ w_1 w_2

定义与记号

1. 从第 m 个词 w_m 的各个词性标记向第 $m+1$ 个词 w_{m+1} 的各个词性标记转移的概率，可以记作 $a_{ij} = a(t_i \rightarrow t_j) = P(t_j|t_i)$ 。第1个词 w_1 前面没有词， w_1 的各个词性标记也满足一定的概率分布，可以记作 π_i 。
2. 从起点词到第 m 个词的第 i 个词性标记的各种可能路径（即各种可能的词性标记串）中，必有一条路径使得 w_m 概率最大，可以用一个变量来对这一过程加以刻画，这个变量即Viterbi变量，记作
$$\delta_m(i) = \max_{t_1, t_2, \dots, t_{m-1}} P(t_1, t_2, \dots, t_m = i, w_1, w_2, \dots, w_m)$$

定义与记号

3. HMM的状态从第m-1个词转移到第m个词，整个路径的概率可以通过HMM在第m-1个词时的最大概率来求得，即Viterbi变量可以递归求值

$$\delta_m(j) = \max_{1 \leq i \leq N} \delta_{m-1}(i) a_{ij} P(w_m | t_j) \quad 2 \leq m \leq M, \quad 1 \leq j \leq N$$

4. 当扫描过第m-1个词，状态转移到第m个词时，需要有一个变量记录已经走过的路径中，哪一条是最佳路径，即记住该路径上 w_m 的最佳词性标记，这个变量可以记作 $\Delta_m(j)$

Viterbi算法

1. 初始化:

$$\delta_1(i) = \pi_i P(w_1 | t_i), \quad 1 \leq i \leq N$$

2. 迭代计算通向每个词(w_m)的每个词性标记的最佳路径

$$\delta_m(j) = \max_{1 \leq i \leq N} \delta_{m-1}(i) a_{ij} P(w_m | t_j), \quad 2 \leq m \leq M, \quad 1 \leq j \leq N$$

$$\Delta_m(j) = \operatorname{argmax}_{1 \leq i \leq N} \delta_{m-1}(i) a_{ij} P(w_m | t_j)$$

3. 到达最后一个词(w_M)时, 计算这个词的最佳词性标记

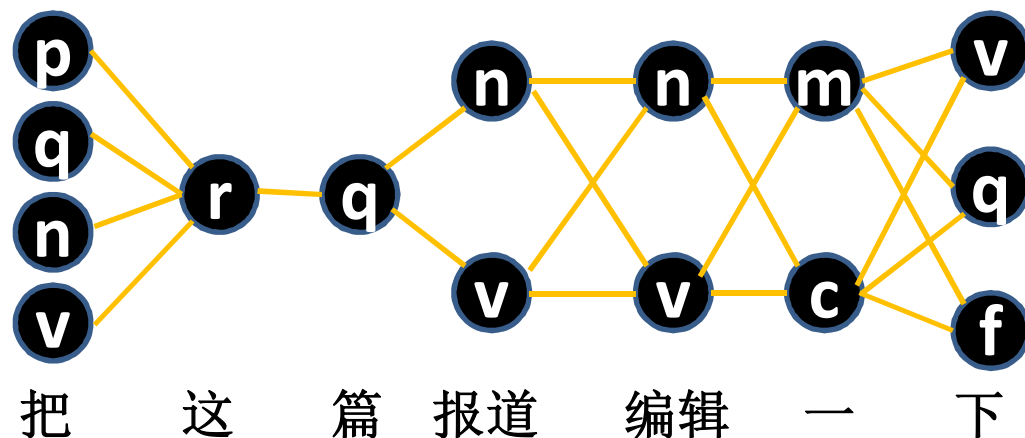
$$P = \max_{1 \leq i \leq N} \delta_M(i), \quad t_M = \operatorname{argmax}_{1 \leq i \leq N} \delta_M(i)$$

4. 从 w_M 的最佳词性标记开始, 顺次取得每个词的最佳词性标记

$$t_m^* = \Delta_{m+1}(t_{m+1}^*), \quad m = M-1, M-2, \dots, 1$$

Viterbi算法词性标注过程示例

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

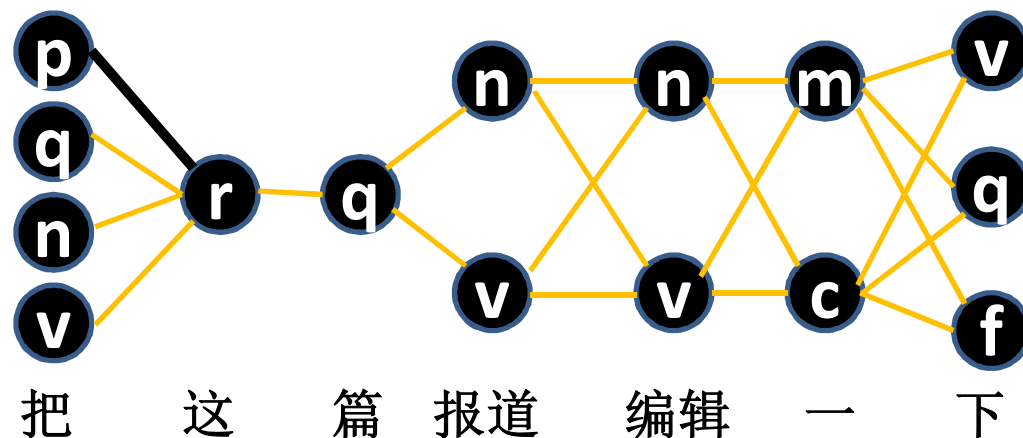


$$\delta(\text{把}/p) = \pi(p) * p(\text{把} | p) = \pi(p) * (9877/269186) \approx 0.0367$$

$$\delta(\text{把}/q) = \pi(q) * p(\text{把} | q) = \pi(q) * (290/155374) \approx 0.00187$$

$$\delta(\text{把}/n) = \pi(n) * p(\text{把} | n) = \pi(n) * (2/1539367) \approx 1.299e-6$$

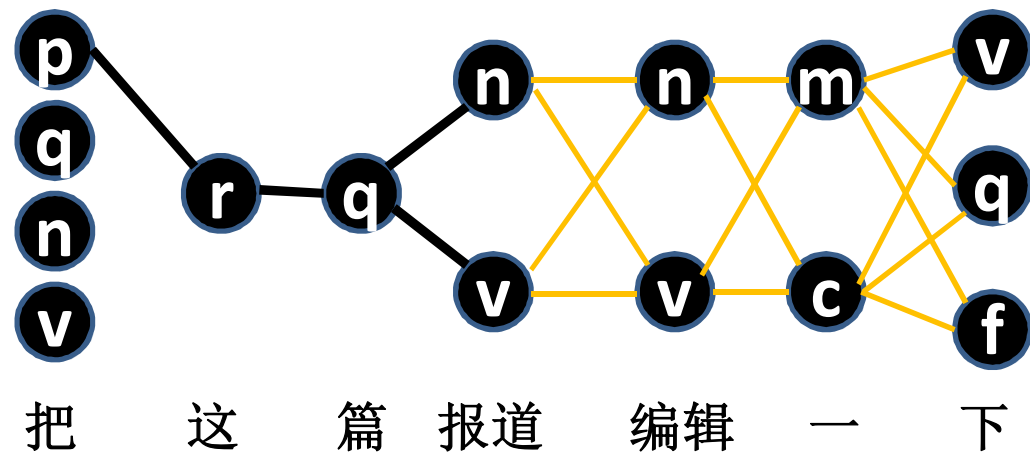
$$\delta(\text{把}/v) = \pi(v) * p(\text{把} | v) = \pi(r) * (208/1193317) \approx 1.743e-4$$



把->这

$$\begin{aligned}
 \sqrt{\delta(\text{这}/r)1} &= \delta(\text{把}/p) * a(p \rightarrow r) * p(\text{这} | r) \\
 &= 0.0367 * (27249/269186) * (21990/214942) = 3.8e-4 \\
 \delta(\text{这}/r)2 &= \delta(\text{把}/q) * a(q \rightarrow r) * p(\text{这} | r) \\
 &= 0.00187 * (760/155374) * (21990/214942) = 9.35e-7 \\
 \delta(\text{这}/r)3 &= \delta(\text{把}/n) * a(n \rightarrow r) * p(\text{这} | r) \\
 &= 1.299e-6 * (9769/1539367) * (21990/214942) = 8.43e-10 \\
 \delta(\text{这}/r)4 &= \delta(\text{把}/r) * a(v \rightarrow r) * p(\text{这} | r) \\
 &= 1.743e-4 * (46697/1193317) * (21990/214942) = 6.972e-7
 \end{aligned}$$

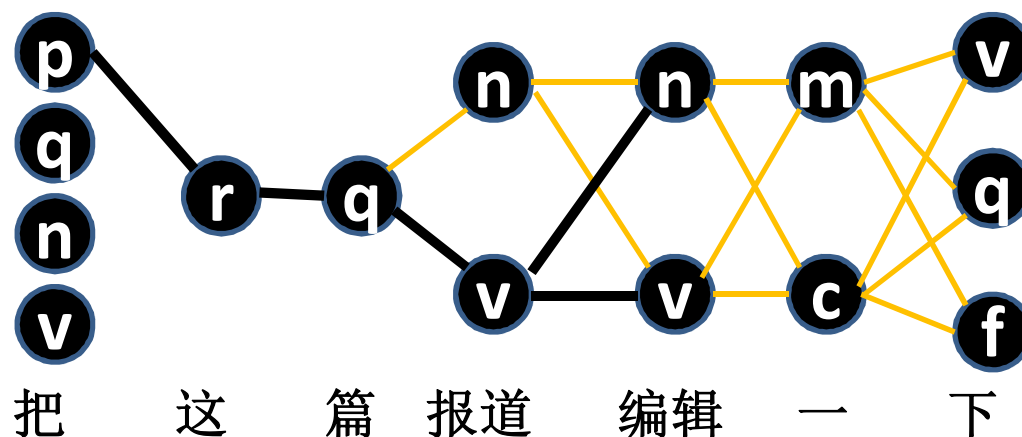
Viterbi算法词性标注过程示例



篇->报道

$$\begin{aligned} \sqrt{\delta(\text{报道}/n)1} &= \delta(\text{篇}/q) * a(q \rightarrow n) * p(\text{报道} | n) \\ &\approx (52310/155374) * (420/1539367) = 9.1857e-5 \end{aligned}$$

$$\begin{aligned} \sqrt{\delta(\text{报道}/v)1} &= \delta(\text{篇}/q) a(q \rightarrow v) * p(\text{报道} | v) \\ &\approx (13288/155374) * (4040/1193317) = 2.8954e-4 \end{aligned}$$



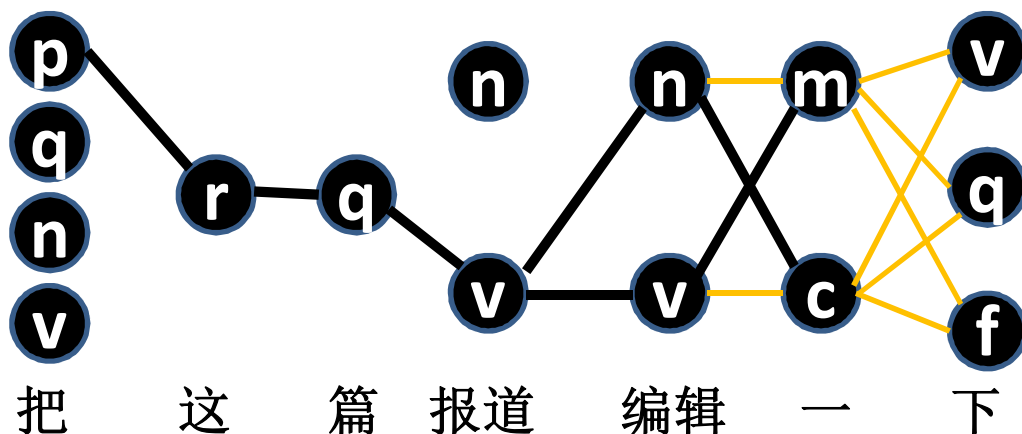
报道->编辑

$$\begin{aligned}\delta(\text{编辑}/n)_1 &= \delta(\text{报道}/n)_1 * a(n \rightarrow n) * p(\text{编辑} | n) \\ &= 9.1857e-5 * (277181/1539367) * (243/1539367) = 2.6e-9\end{aligned}$$

$$\begin{aligned}\sqrt{\delta(\text{编辑}/n)_2} &= \delta(\text{报道}/v)_1 * a(v \rightarrow n) * p(\text{编辑} | n) \\ &= 2.8954e-4 * (221796/1193317) * (243/1539367) = 8.49e-9\end{aligned}$$

$$\begin{aligned}\delta(\text{编辑}/v)_1 &= \delta(\text{报道}/n)_1 * a(n \rightarrow v) * p(\text{编辑} | v) \\ &= 9.1857e-5 * (221776/1539367) * (100/1193317) = 1.1e-9\end{aligned}$$

$$\begin{aligned}\sqrt{\delta(\text{编辑}/v)_2} &= \delta(\text{报道}/v)_1 * a(v \rightarrow v) * p(\text{编辑} | v) \\ &= 2.8954e-4 * (191967/1193317) * (100/1193317) = 3.9e-9\end{aligned}$$



编辑->一

$$\delta(\text{一}/m)_1 = \delta(\text{编辑}/n)_2 * a(n \rightarrow m) * p(\text{一} | m)$$

$$= 8.49e-9 * (27918/1539367) * (20672/270381) = 1.18e-11$$

$$\sqrt{\delta(\text{一}/m)_2 = \delta(\text{编辑}/v)_2 * a(v \rightarrow m) * p(\text{一} | m)}$$

$$= 3.9e-9 * (70914/1193317) * (20672/270381) = 1.77e-11$$

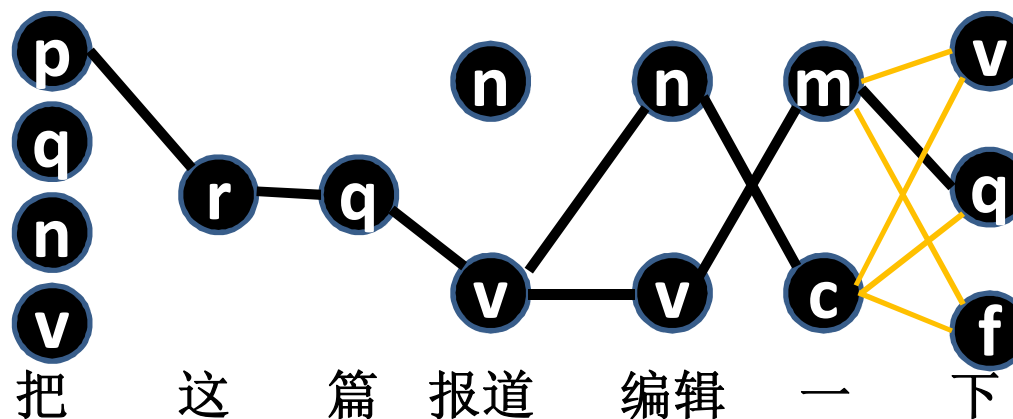
$$\sqrt{\delta(\text{一}/c)_1 = \delta(\text{编辑}/n)_2 * a(n \rightarrow c) * p(\text{一} | c)}$$

$$= 8.49e-9 * (55177/1539367) * (2229/168350) = 4e-12$$

$$\delta(\text{一}/c)_2 = \delta(\text{编辑}/v)_2 * a(v \rightarrow c) * p(\text{一} | c)$$

$$= 3.9e-9 * (13715/1193317) * (2229/168350) = 5.9e-13$$

一->下



$$\delta(\text{下}/v)_1 = \delta(\text{一}/m)_2 * a(m \rightarrow v) * p(\text{下} | v)$$

$$= 1.77e-11 * (13778/270381) * (2271/1193317) = 1.7e-15$$

$$\delta(\text{下}/v)_2 = \delta(\text{一}/c)_1 * a(c \rightarrow v) * p(\text{下} | v)$$

$$= 4e-12 * (40148/168350) * (2271/1193317) = 1.8e-15$$

$$\sqrt{\delta(\text{下}/q)_1 = \delta(\text{一}/m)_2 * a(m \rightarrow q) * p(\text{下} | q)}$$

$$= 1.77e-11 * (139653/270381) * (161/155374) = 9.47e-15$$

$$\delta(\text{下}/q)_2 = \delta(\text{一}/c)_1 * a(c \rightarrow q) * p(\text{下} | q)$$

$$= 4e-11 * (53/168350) * (161/155374) = 1.3e-18$$

$$\delta(\text{下}/f)_1 = \delta(\text{一}/m)_2 * a(m \rightarrow f) * p(\text{下} | f)$$

$$= 1.77e-11 * (1470/270381) * (6313/110878) = 5.47e-15$$

$$\delta(\text{下}/f)_2 = \delta(\text{一}/c)_1 * a(c \rightarrow f) * p(\text{下} | f)$$

$$= 4e-12 * (700/168350) * (6313/110878) = 9.47e-16$$

Viterbi算法的复杂度

- 假定有 N 个词性标记，给定词串中有 M 个词
- 考虑最坏的情况，扫描到每一个词时，从前一个词的各个词性标记（ N 个）到当前词的各个词性标记（ N 个），有 $N \times N = N^2$ 条路经，即 N^2 次运算，扫描完整整个词串（长度为 M ），计算次数为 N^2 个 M 相加，即 $N^2 \times M$ 。
- 对于确定的词性标注系统而言， N 是确定的，因此，随着 M 长度的增加，计算时间以线性方式增长。也就是说，Viterbi算法的计算复杂性是线性的。

HMM的形式描述

一个HMM可以记做 $\lambda = (S, O, A, B, \pi)$:

1. 状态集合 $S = \{s_1, s_2, \dots, s_N\}$, 一般以 q_t 表示模型在 t 时刻的状态;
2. 输出符号集合 $O = \{o_1, o_2, \dots, o_M\}$;
3. 状态转移矩阵 $A = \{a_{ij}\}$ (a_{ij} 是从 i 状态转移到 j 状态的概率), 其中 $a_{ij} = P(q_{t+1} = j | q_t = i)$, $1 \leq i, j \leq N$, $a_{ij} \geq 0$, $\sum_{j=1}^N a_{ij} = 1$;
4. 可观察符号的概率分布 $B = b_j(k)$, 表示在状态 j 时输出符号 v_k 的概率, 其中:
 $b_j(k) = P(o_t = v_k | q_t = s_j)$, $1 \leq j \leq N$, $1 \leq k \leq M$, $b_j(k) \geq 0$, $\sum_{k=1}^M b_j(k) = 1$;
5. 初始状态概率分布, 一般记做 $\pi = \{\pi_i\}$, 其中: $\pi_i = P(q_1 = s_i)$, $1 \leq i \leq N$, $\pi_i \geq 0$, $\sum_{i=1}^N \pi_i = 1$.

HMM的三个基本问题

- 给定一个观察序列 $O = o_1 o_2 \cdots o_T$ 和模型 λ ，如何计算给定模型 λ 下观察序列 O 的概率 $P(O|\lambda)$
- 给定一个观察序列 $O = o_1 o_2 \cdots o_T$ 和模型 λ ，如何计算状态序列 $Q = q_1 q_2 \cdots q_T$ ，使得该状态序列能“最好地解释”观察序列 (对应词性标注问题)
- 给定一个观察序列 $O = o_1 o_2 \cdots o_T$ ，如何调节模型 λ 的参数值，使得 $P(O|\lambda)$ 最大

HMM的应用

序列(sequential data)分类、标注模型

- 基于字标注的分词

B	E	M	S
词首	词尾	词中	独立词

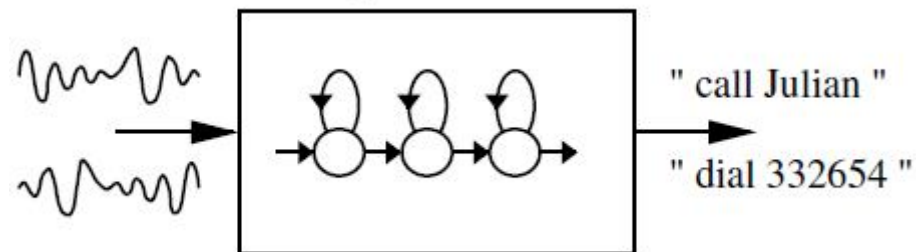
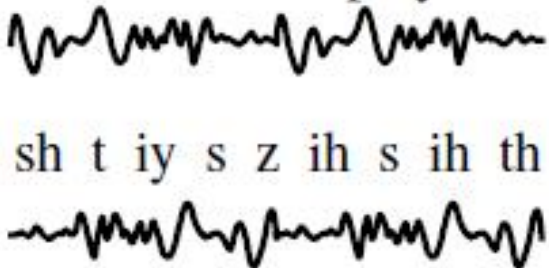
自然句形式	已结婚的和尚未结婚的都应该到计生办登记
词切分结果	已/ 结婚/ 的/ 和/ 尚未/ 结婚/ 的/ 都/ 应该/ 到/ 计生办/ 登记/
字标注结果	已 结 婚 的 和 尚 未 结 婚 的 都 应 该 到 计 生 办 登 记 S B E S S B E B E S S B E S B M E B E

- 条件随机场 (Conditional Random Fields)

HMM的应用

- 语音识别
- 手语识别
- 行为分析

th ih s ih s p iy t sh
sh t iy s z ih s ih th



HMM的应用

- 图像内容标注
- 视频



Sky
Animal
Grass



Sky
Building
Bus
Ground



Sky
Water, Sand
People
Building

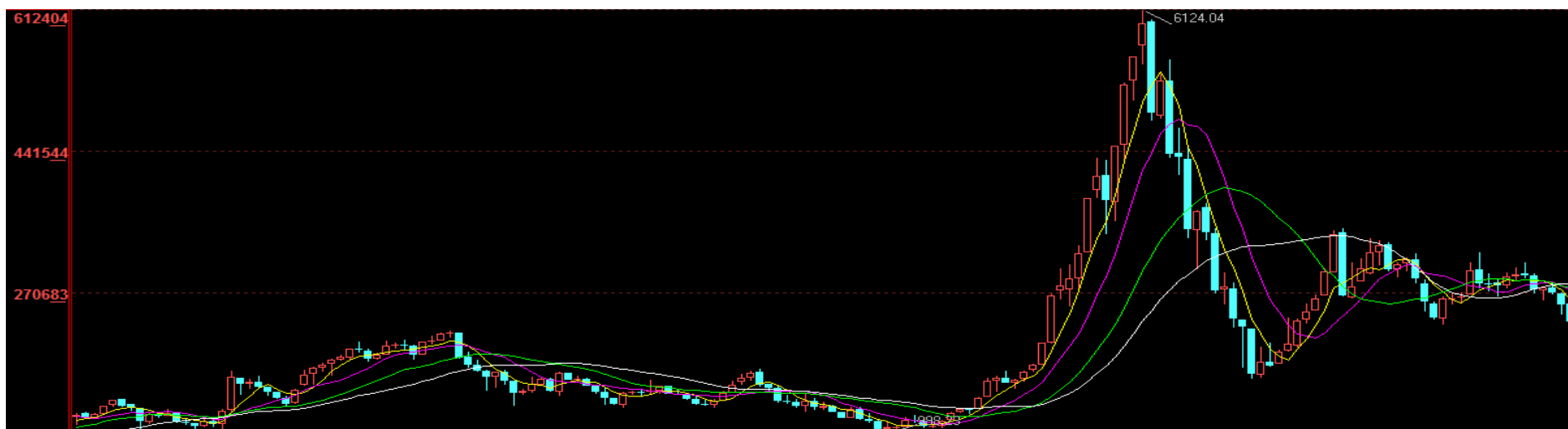
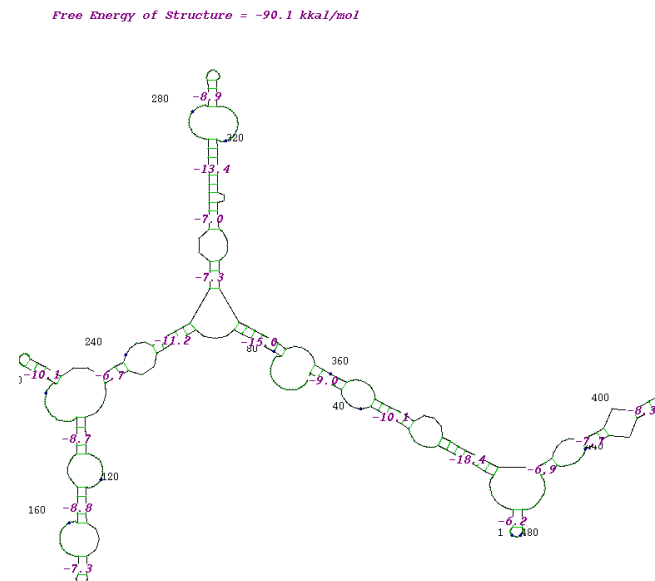


Sky
Mountain
Snow



HMM的应用

- 生物序列分析
- 金融市场预测
- 网络安全、信息抽取



规则和统计相结合的词性标注方法

- 规则消歧，统计概率引导
- 或者统计方法赋初值，规则消歧

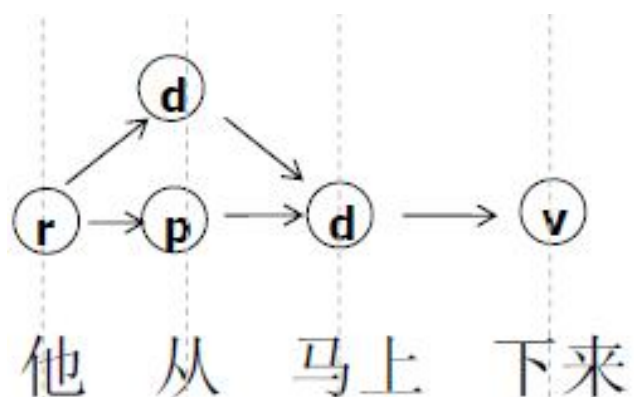
例：HMM分词结果

把/p 这/r 篇/q 报道/v 编辑/v 一/m 下/q

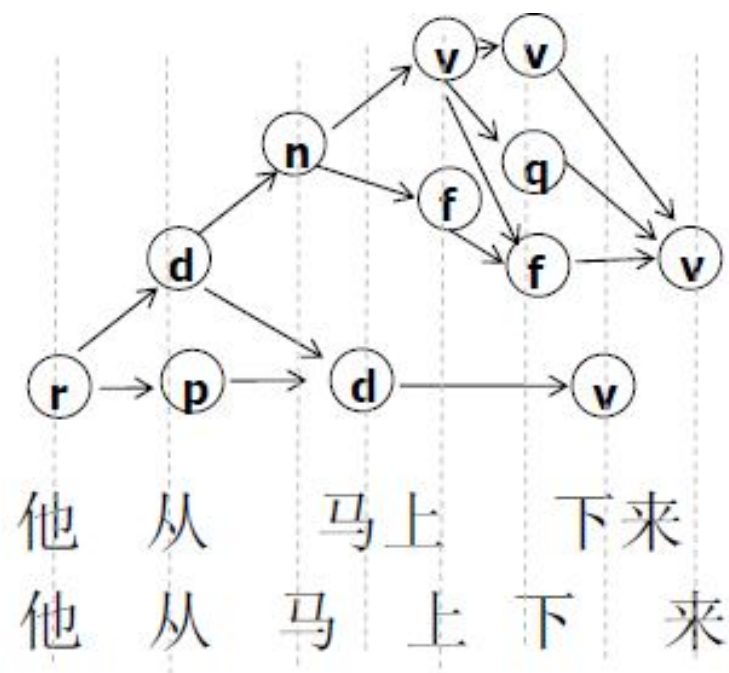
规则T1：当前词的前一个词的词性标记是量词（q）时，将当前词的词性标记由动词（v）改为名词（n）；

最后结果：把/p 这/r 篇/q 报道/n 编辑/v 一/m 下/q

分词和词性标注一体化示例



先分词再标注



分词标注一体化

3 小结

□统计方法、机器学习改错规则等基于语料库的方法在词性标注中有比较显著的优势。

- 统计模型的多样性

 - 决策树模型，最大熵模型， **SVM**，神经网络.....

- 不同统计方法的融合

- 词性标注与分词过程的融合