

第三讲 词性标注

詹卫东

<http://ccl.pku.edu.cn/course/nlp>



提纲

1. 汉语的词类划分及标注问题
2. 词类自动标注的方法
 - 隐马尔可夫模型（HMM）+ Viterbi算法
 - 基于转换的错误驱动的词性标注方法
3. 小结



1 汉语的词类划分及标注问题

- 词类： 一个语言全部词汇的子类划分。
Part of speech, word class

那么，依据什么标准对词进行分类呢？

1. 意义 事物、动作行为、性质状态、数量、.....
2. 形态 -ed -ing -s -er -ly
3. 分布 词语能出现和不能出现的位置



词语的分布示例 (I)

| | a | b | c |
|---|-----------|--------------------------|-----------|
| 1 | 很 多 学 生 | 很 多 高 个 子 学 生 | 很 多 高 个 子 |
| 2 | 高 个 子 学 生 | 高 个 子 很 多 学 生 | 高 个 子 很 多 |

“很多”和“高个子”的分布（位置）不同！



词语的分布示例 (II)

- | | a | b |
|---|-----------|--------------|
| 1 | 悠久 历史 | 历史 悠久 |
| 2 | ✗ 悠久 不 历史 | 历史 不 悠久 |
| 3 | 两千年的 悠久历史 | ✗ 两千年的 历史 悠久 |

“历史”和“悠久”的分布（位置）不同！



词语的分布示例 (III)

a

1 他 迅速 销毁了密码本

2 他 销毁密码本 很 迅速

果断

坚决

突然

仔细

用力

大胆

.....

b

他 立即 销毁了密码本

✗ 他 销毁密码本 很立即

马上

悄悄

立刻

已经

必定

亲自

.....



划分词类的基本思路

1. 给出语言中基本组合关系的清单——确定句法位置；
2. 根据词语进入基本组合关系中不同位置的能力，对词语进行**功能分类**；

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>



现代汉语的基本组合关系

| 组合类型 | 句法结构成分（位置） | 实例 |
|------|------------|----------------|
| 主谓结构 | 主语 + 谓语 | 老张 去 机器 很重 |
| 述宾结构 | 述语1 + 宾语 | 修理 桌子 学习 语法 |
| 述补结构 | 述语2 + 补语 | 看 清楚 站 稳 |
| 定中结构 | 定语 + 中心语1 | 木头 桌子 汉语 语法 |
| 状中结构 | 状语 + 中心语2 | 赶快 出发 非常 了解 |
| 连谓结构 | 前谓 + 后谓 | 走路 去 回家 休息 |
| 联合结构 | 前项 + 后项 | 长江 黄河 唱歌 跳舞 |



现代汉语的基本组合关系（续）

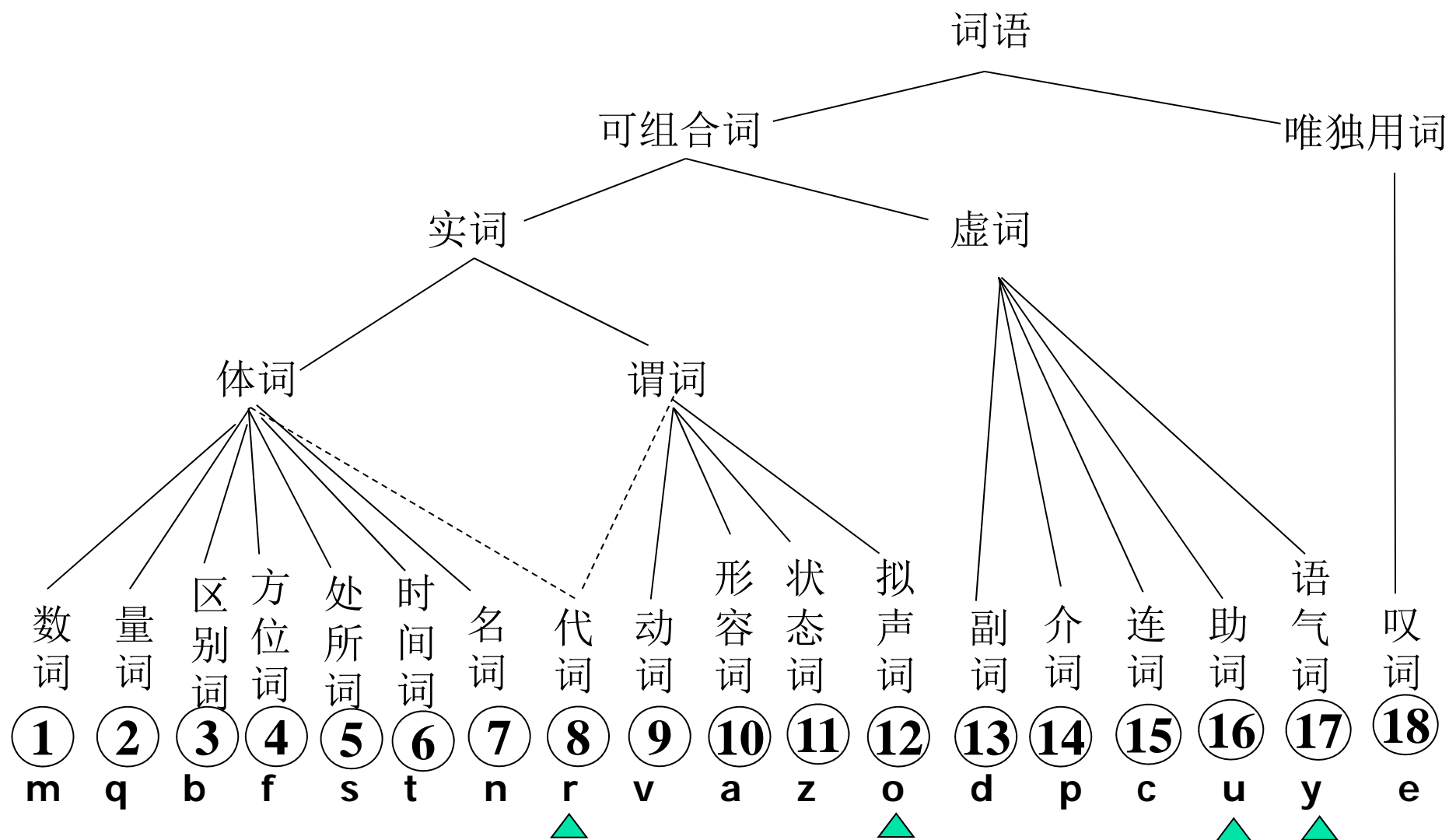
| 结构类型 | 句法结构成分（位置） | 实例 |
|--------|-------------------|-------------------------|
| “的”字结构 | X + 的 | 听话 的 张三 的 |
| “地”字结构 | X + 地 | 悄悄 地 高兴 地 |
| “所”字结构 | 所 + X X + 所 + Y | 所 提 (条件) 学校 所 需要 (的) |
| 介宾结构 | 介词 + 宾语 | 把 大家 向 窗外 |
| 方位结构 | 时间 处所 + 方向 相对位置 | 春节 以前 桌子 上 |
| 数量结构 | 数词 + 量词 | 三十二 本 两 批 |



根据词语占据的结构位置划分词类

- 数词：可枚举。在量词前
- 量词：可枚举。在数词后
- 名词：数+量+ _____
- 动词：主+ _____ _____+宾 _____+补
- 形容词：主+ _____ ~~_____~~+宾 很 _____
- 区别词：_____定+中 ~~_____~~状+中 很 ~~_____~~
- 副词：~~_____~~定+中 _____状+中 很 ~~_____~~
- 状态词：_____定+中 _____状+中 很 ~~_____~~
-

现代汉语的词类系统





词的分布 – 词类 – 包含什么信息？

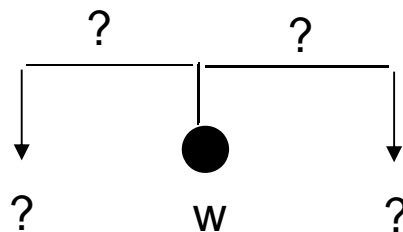
(1) 词 (w) 的组合方向：

w 在参与序列组合时朝哪个方向组合；

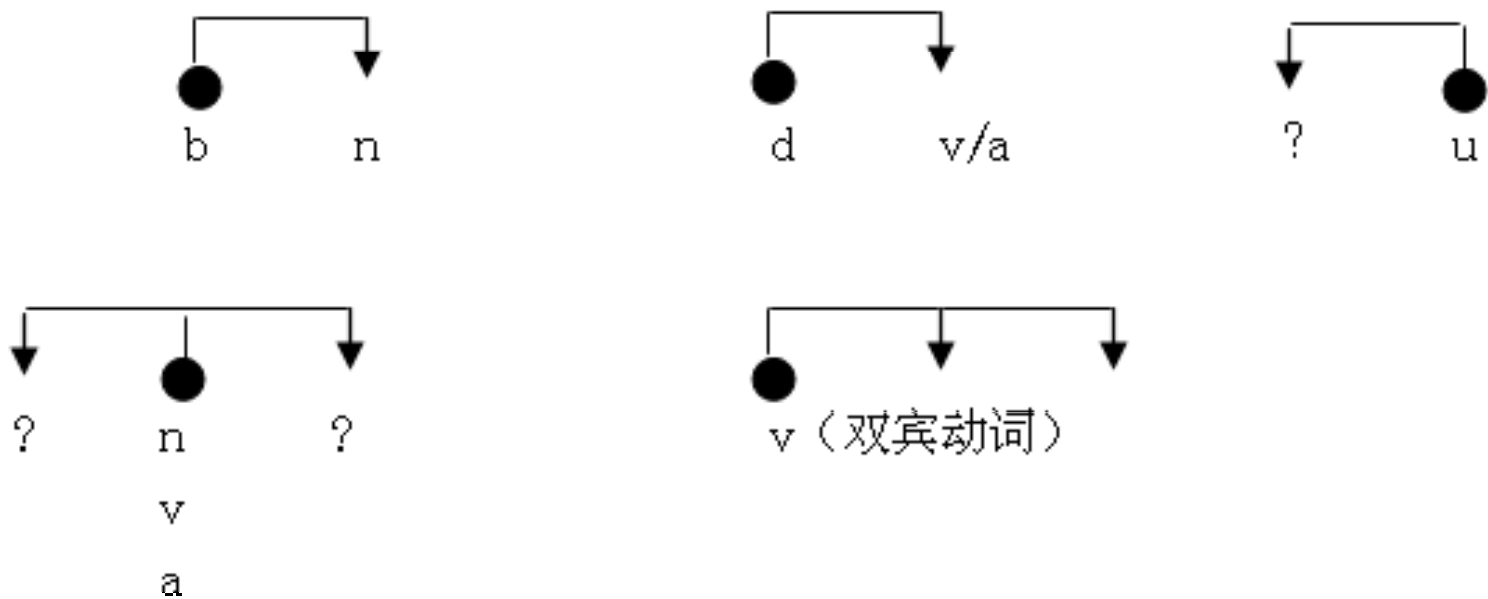
(2) 词 (w) 的组合对象：

w 要求跟几个成分组合；

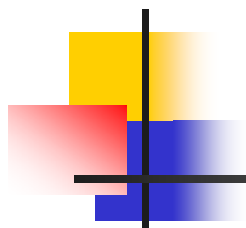
w 要求跟什么类型的语言成分组合。



“词类”传递的信息



b: 区别词 d: 副词 u: 助词 v: 动词 a: 形容词 n: 名词



更细粒度的分布测试

例：“X心”的分布

| | 耐心 | 热心 | 信心 | 手心 |
|--------|----|----|----|----|
| 有____ | + | - | + | + |
| 很____ | + | + | - | - |
| 很有____ | + | - | + | - |

4种分布，几个词类？



词类区别举例：形容词与状态词

白 — 雪白

香 — 喷香/香喷喷

1. 很 _____

2. 张三的脸 比 李四的脸 _____

3. 这碗饭 比 那碗饭 _____



词类的典型成员和非典型成员

词类划分无法做到严格的集合划分的要求：

对内一致性，对外排他性

A. 电脑 风衣 风水 文笔 战争 看法

B. 般配 一样 矮小 安全 耐烦 方便

A 组中不大典型的名词是 _____

B 组中不大典型的形容词是 _____



词语的兼类现象

- 一个词的分布位置兼具两类词的分布特征，则该词兼属两个词类。

| | a | b |
|---|-------|----------|
| 1 | 自动 步枪 | 自动 回复 |
| 2 | 两 把 锁 | 锁 两次 |
| 3 | 一件 制服 | 制服 不了 小偷 |
| 4 | 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 | 例子 | 说明 |
|-----|------|-------|-------------|---------------------------------------|
| 5 | 3 | 0.01% | 和 c-n-p-q-v | 我和你；三数之和；姓和；我和上级说；两和(huo4)药；和(huo4)稀泥 |
| 4 | 20 | 0.04% | 光 a-d-n-v | 磨得很光；光说不做；一丝光透进来；光着膀子 |
| 3 | 126 | 0.23% | 画 n-q-v | 一幅画；四画字；画一幅画 |
| 2 | 1475 | 2.67% | 锁 n-v | 锁不了门；三把锁 |
| 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



汉语中兼类词的比例（续）

| 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 | |



英语词的兼类现象

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

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)

Degree of ambiguity

Total frequency (39,440)

1 tag 35,340

2-7 tags 4,100

2 3,760

3 264

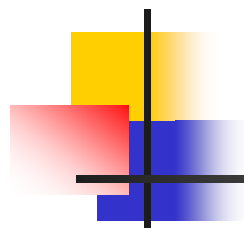
4 61

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（汉语）

| 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>



词类标注问题

- 对文本中兼类词的具体词类归属进行判定
 - 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
 - Time flies like an arrow.
Time/n-v flies/v-n like/p-v an/Det arrow/n

| 标记 | ex. | 描述 | 标记 | ex. | 描述 |
|----|-----|------|----|-----|------|
| Ag | 孤 | 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 | 甚 | 副语素 | r | | 代词 |
| d | | 副词 | s | | 处所词 |
| e | | 叹词 | Tg | 昨 | 时语素 |
| f | | 方位词 | t | | 时间词 |
| g | | 语素 | u | | 助词 |
| h | | 前接成分 | Vg | 育 | 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 | 汉 | 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

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。

ad : a用作d vd : v用作d
an : a用作n vn : v用作n

| | 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 | ... | ... | ... |

北大计算语言所分 词和词性标注语料 库分级词性标记集

1999, 2002, 2003

ex. 胜利召开、循环使用

ex. 外交工作、巨大进展

ex. 能、可以、应

ex. 进行、予以



英语词性标记集举例

- Brown corpus tagset
 - 87 tags
 - Used for Brown Corpus (1-million-word, 1963-1964, Brown University)
 - TAGGIT program
- UPenn treebank tagset
 - 45 tags
 - Used for UPenn treebank, Brown Corpus, WSJ Corpus
 - Brill tagger
- UCREL's C5 tagset
 - 61 tags
 - Used for British National Corpus (BNC)
 - Lancaster CLAWS tagger

| Tag | Description | Example | Tag | Description | Example |
|------|-----------------------|------------------------|------|-----------------------|-------------------------|
| CC | Coordin. Conjunction | <i>and, but, or</i> | SYM | Symbol | <i>+, %, &</i> |
| 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> | VCN | 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 | <i>\$</i> |
| NNPS | Proper noun, plural | <i>Carolinas</i> | # | Pound sign | <i>#</i> |
| PDT | Predeterminer | <i>all, both</i> | “ | Left quote | <i>(‘ or “)</i> |
| POS | Possessive ending | <i>’s</i> | ” | Right quote | <i>(’ or ”)</i> |
| PP | Personal pronoun | <i>I, you, he</i> | (| Left parenthesis | <i>([, { , <)</i> |
| PP\$ | Possessive pronoun | <i>your, one’s</i> |) | Right parenthesis | <i>(],), }, >)</i> |
| RB | Adverb | <i>quickly, never</i> | , | Comma | <i>,</i> |
| RBR | Adverb, comparative | <i>faster</i> | . | Sentence-final punc | <i>(. ! ?)</i> |
| RBS | Adverb, superlative | <i>fastest</i> | : | Mid-sentence punc | <i>(: ; ... – -)</i> |
| RP | Particle | <i>up, off</i> | | | |

UPenn
treebank
POS
tagset
(45 tags)



2 词类自动标注的方法

| 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% |
| | | | | | | |



用规则方法进行词性标注示例

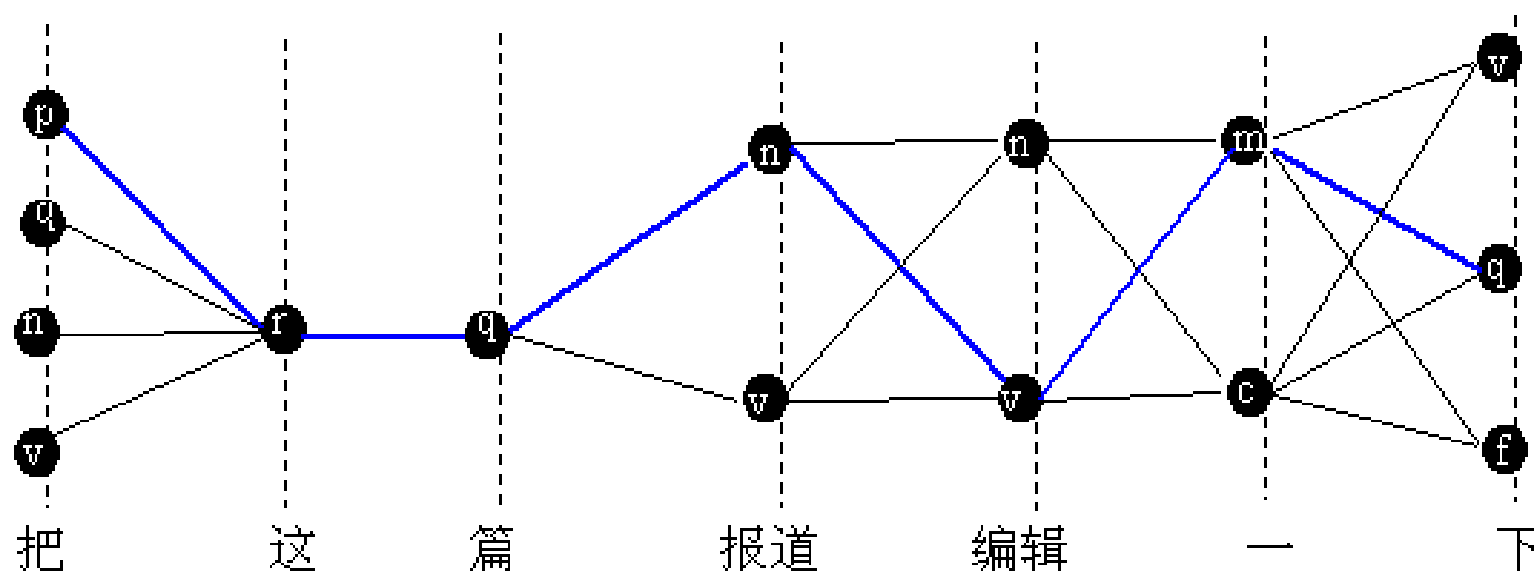
@@ 信(n-v)

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

@@ 一边(c-s)

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

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？

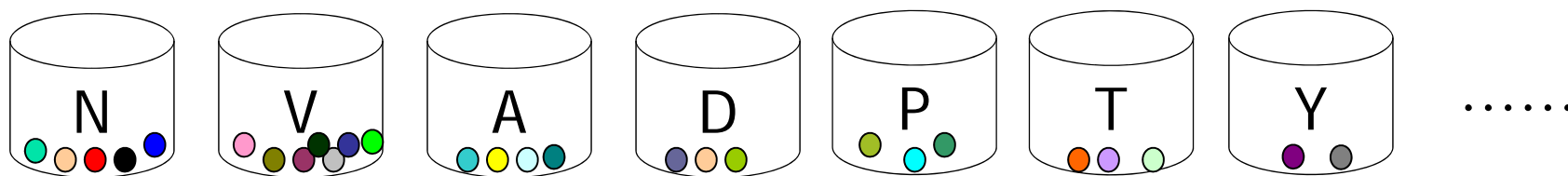


2.1 隐马尔可夫模型 (Hidden Markov Model)

- 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 (2000), 第1章;
 翁富良 (1998), 第8章第2节;

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，



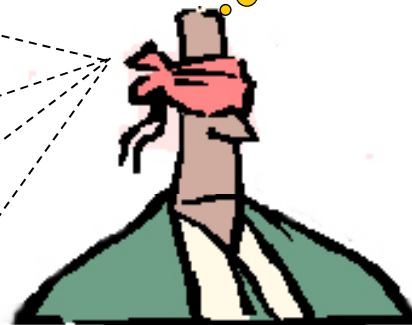
可观察序列



猜测隐藏在幕
后的罐子序列:
Markov序列

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

Which one
is the best?





HMM的形式描述

- 1 状态集合 $S=\{a_1, a_2, \dots, a_N\}$, 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:

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

- 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:

$$b_j(k) = P(O_t = v_k | q_t = S_j), 1 \leq j \leq N, 1 \leq K \leq M \quad b_j(k) \geq 0 \quad \sum_{k=1}^M b_j(k) = 1$$

- 5 初始状态概率分布, 一般记做 $\pi = \{ \pi_i \}$, 其中:

$$\pi_i = P(q_1 = S_i), 1 \leq i \leq N \quad \pi_i \geq 0 \quad \sum_{i=1}^N \pi_i = 1$$



HMM的三个基本问题

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

- 1 给定一个观察序列 $O=O_1O_2...O_T$ 和模型 λ ，如何计算给定模型 λ 下观察序列 O 的概率 $P(O|\lambda)$
- 2 给定一个观察序列 $O=O_1O_2...O_T$ 和模型 λ ，如何计算状态序列 $Q=q_1q_2...q_T$ ，使得该状态序列能“最好地解释”观察序列
- 3 给定一个观察序列 $O=O_1O_2...O_T$ ，如何调节模型 λ 的参数值，使得 $P(O|\lambda)$ 最大

对应着词性标注问题



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 $W=w_1w_2...w_n$ ），需要去猜测罐子序列（隐藏在幕后的词性标记序列，记做 $T=t_1t_2...t_n$ ）
- 已知词串 W （观察序列）和模型 λ 情况下，求使得条件概率 $P(T|W, \lambda)$ 值最大的那个 T' ，一般记做：

$$T' = \arg \max_T P(T | W, \lambda) \quad \text{公式1}$$



基于HMM进行词性标注（续）

- 根据条件概率公式可得

$$P(T | W, \lambda) = \frac{P(T, W | \lambda)}{P(W | \lambda)} \quad \text{公式2}$$

- 公式2可进一步简化为（根据Bayes公式）：

$$P(T | W) = \frac{P(T, W)}{P(W)} = \frac{P(T)P(W | T)}{P(W)} \quad \text{公式3}$$



基于HMM进行词性标注（续）

- 公式3可以进一步简化为：

$$P(T | W) \approx P(T)P(W | T) \quad \text{公式4}$$

其中：

$$P(T) = P(t_1 | t_0)P(t_2 | t_1, t_0) \dots P(t_i | t_{i-1}, t_{i-2}, \dots) \quad \text{公式5}$$

根据一阶HMM的独立性假设，可得

$$P(T) \approx P(t_1 | t_0)P(t_2 | t_1) \dots P(t_i | t_{i-1}) \quad \text{公式6}$$

词性之间的转移概率可以从语料库中估算得到：

$$P(t_i | t_{i-1}) = \frac{\text{训练语料中 } t_i \text{ 出现在 } t_{i-1} \text{ 之后的次数}}{\text{训练语料中 } t_{i-1} \text{ 出现的总次数}} \quad \text{公式7}$$



基于HMM进行词性标注（续）

- $P(W|T)$ 是已知词性标记串，产生词串的条件概率：

$$P(W | T) = P(w_1 | t_1)P(w_2 | t_2, t_1, w_1) \dots P(w_i | t_i, t_{i-1}, \dots, t_1, w_{i-1}, \dots, w_1) \quad \text{公式8}$$

- 根据HMM的独立性假设，公式8可简化为：

$$P(W | T) \approx P(w_1 | t_1)P(w_2 | t_2) \dots P(w_i | t_i) \quad \text{公式9}$$

- 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：

$$P(w_i | t_i) = \frac{\text{训练语料中 } w_i \text{ 的词性被标记为 } t_i \text{ 的次数}}{\text{训练语料中 } t_i \text{ 出现的总次数}} \quad \text{公式10}$$

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

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

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

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

.....

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

从中选
一个最
大值

词性转移概率

词语输出概率



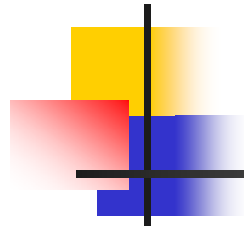
效率问题

假定有 K 个词性标记（罐子），给定词串中有 N 个词（彩球）

考虑最坏的情况：每个词都有 K 个可能的词性标记，则可能的状态序列有 K^N 个

随着 N （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.....



Veterbi算法 —— 提高效率之道

- Veterbi算法是一种动态规划方法（dynamic programming）

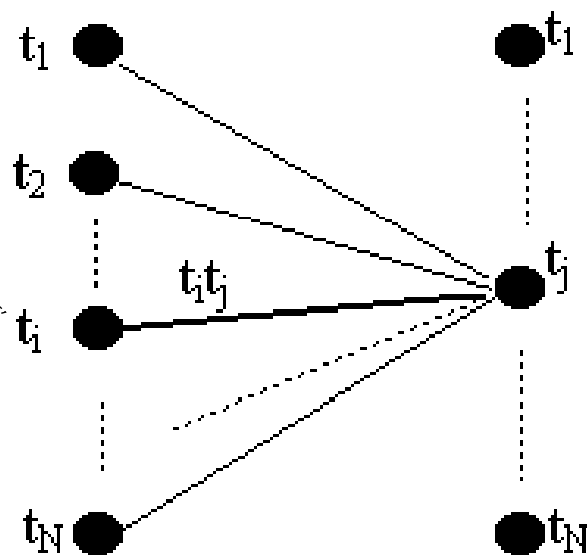
如果当前节点在最优路径上，那么，不管当前节点的后续路径如何，当前节点的来源路径必定是最优的。

最优路径的求解可以迭代进行。

Viterbi, A., 1967, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory Vol.13 No.2, pp260–269

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 ... w_m w_{m+1} ...

假定共有 N 个词性标记，给定词串有 M 个词



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 $a_{ij} = P(t_j | t_i)$ $1 \leq i \leq N$; $1 \leq j \leq N$

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. 第 m 个词 (w_m) 的各个词性标记取词语 w_m 的条件概率可以记作 $b_i(w_m) = P(w_m | t_i)$ $1 \leq i \leq N$
3. 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

$$\delta_m(i) = \max_{t_1, t_2, \dots, t_{m-1}} P(t_1, t_2, \dots, t_m = i, w_1, w_2, \dots, w_m | \lambda) \quad 1 \leq m \leq M; \quad 1 \leq i \leq N$$



定义与记号（续）

- 4 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在前一个状态（第m-1个词）时的最大概率来求得，即Viterbi变量可以递归求值

$$\delta_m(j) = [\max_{1 \leq i \leq N} \delta_{m-1}(i) a_{ij}] \times b_j(w_m) \quad 1 \leq m \leq M, \quad 1 \leq j \leq N$$

- 5 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

$$\Delta_m(j) = \arg \max_{1 \leq i \leq N} [\delta_{m-1}(i) a_{ij}] \times b_j(w_m) \quad 2 \leq m \leq M, \quad 1 \leq j \leq N$$



Veterbi算法

(1) 初始化:

$$\delta_1(i) = \pi_i b_i(w_1) \quad 1 \leq i \leq N \quad \Delta_1(i) = 0$$

(2) 迭代计算通向每个词(w_m)的每个词性标记(t_j)的最佳路径

$$\delta_m(j) = [\max_{1 \leq i \leq N} \delta_{m-1}(i) a_{ij}] \times b_j(w_m) \quad 2 \leq m \leq M, \quad 1 \leq i \leq N$$

$$\Delta_m(j) = \arg \max_{1 \leq i \leq N} [\delta_{m-1}(i) \times a_{ij}] \times b_j(w_m) \quad 2 \leq m \leq M, \quad 1 \leq i \leq N$$

(3) 到达最后一个词(w_M)时, 计算这个词的最佳词性标记

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

(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

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



Veterbi算法的复杂度

假定有 K 个词性标记（罐子），给定词串中有 N 个词（彩球）

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（ K 个）到当前词的各个词性标记（ K 个），有 $K \times K = K^2$ 条路经，即 K^2 次运算，扫描完整个词串（长度为 N ），计算次数为 K^2 个 N 相加，即 $K^2 \times N$ 。

对于确定的词性标注系统而言， K 是确定的，因此，随着 N 长度的增加，计算时间以线性方式增长。也就是说，Veterbi算法的计算复杂性是线性的。



用Viterbi算法进行词性标注示例

词性转移矩阵（用于估算转移概率）

| Tag \ 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 |
|----|----|-------|--|----|----|-------|
| 把 | p | 9877 | | 编辑 | n | 243 |
| 把 | q | 290 | | 编辑 | v | 100 |
| 把 | n | 2 | | 一 | m | 20672 |
| 把 | v | 208 | | 一 | c | 2229 |
| 这 | r | 21990 | | 下 | f | 6313 |
| 篇 | q | 706 | | 下 | q | 161 |
| 报道 | v | 4040 | | 下 | v | 2271 |
| 报道 | n | 420 | | | | |



词性频度表

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

词性标记总频次：
7284443



Veterbi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

把 -> 这

$$\text{Delta(这/r)}_1 = a_{12}(\text{把/p} \rightarrow \text{这/r}) * b_2(\text{这/r}) = (27249 / 269186) * (21990 / 214942) = 0.01036$$

$$\text{Delta(这/r)}_2 = a_{12}(\text{把/q} \rightarrow \text{这/r}) * b_2(\text{这/r}) = (760 / 155374) * (21990 / 214942) = 5e-4$$

$$\text{Delta(这/r)}_3 = a_{12}(\text{把/n} \rightarrow \text{这/r}) * b_2(\text{这/r}) = (9769 / 1539367) * (21990 / 214942) = 6.49e-4$$

$$\text{Delta(这/r)}_4 = a_{12}(\text{把/v} \rightarrow \text{这/r}) * b_2(\text{这/r}) = (46697 / 1193317) * (21990 / 214942) = 0.004$$



Veterbi算法词性标注过程示例(续)

篇 -> 报道

Delta(篇) 只有一个，略去。

$$\text{Delta(报道/n)}_1 = a_{12}(\text{篇/q} \rightarrow \text{报道/n}) * b_2(\text{报道/n}) = (52310 / 155374) * (420 / 1539367) = 9.1857e-5$$

$$\text{Delta(报道/v)}_1 = a_{12}(\text{篇/q} \rightarrow \text{报道/v}) * b_2(\text{报道/v}) = (13288 / 155374) * (4040 / 1193317) = 2.8954e-4$$



Veterbi算法词性标注过程示例(续)

报道 -> 编辑

$$\begin{aligned}\text{Delta}(\text{编辑}/n)1 &= \text{Delta}(\text{报道}/n)1 * a23(\text{报道}/n \rightarrow \text{编辑}/n) * b3(\text{编辑}/n) \\ &= 9.1857e-5 * (277181 / 1539367) * (243 / 1539367) = 2.6e-9\end{aligned}$$

$$\begin{aligned}\text{Delta}(\text{编辑}/n)2 &= \text{Delta}(\text{报道}/v)1 * a23(\text{报道}/v \rightarrow \text{编辑}/n) * b3(\text{编辑}/n) \\ &= 2.8954e-4 * (221796 / 1193317) * (243 / 1539367) = 8.49e-9\end{aligned}$$

$$\begin{aligned}\text{Delta}(\text{编辑}/v)1 &= \text{Delta}(\text{报道}/n)1 * a23(\text{报道}/n \rightarrow \text{编辑}/v) * b3(\text{编辑}/v) \\ &= 9.1857e-5 * (221776 / 1539367) * (100 / 1193317) = 1.1e-9\end{aligned}$$

$$\begin{aligned}\text{Delta}(\text{编辑}/v)2 &= \text{Delta}(\text{报道}/v)1 * a23(\text{报道}/v \rightarrow \text{编辑}/v) * b3(\text{编辑}/v) \\ &= 2.8954e-4 * (191967 / 1193317) * (100 / 1193317) = 3.9e-9\end{aligned}$$



Veterbi算法词性标注过程示例(续)

编辑 -> 一

$$\begin{aligned}\text{Delta}(\text{一}/\text{m})1 &= \text{Delta}(\text{编辑}/\text{n})2 * a34(\text{编辑}/\text{n} \rightarrow \text{一}/\text{m}) * b4(\text{一}/\text{m}) \\ &= 8.49\text{e-}9 * (27918 / 1539367) * (20672 / 270381) = 1.18\text{e-}11\end{aligned}$$

$$\begin{aligned}\text{Delta}(\text{一}/\text{m})2 &= \text{Delta}(\text{编辑}/\text{v})2 * a34(\text{编辑}/\text{v} \rightarrow \text{一}/\text{m}) * b4(\text{一}/\text{m}) \\ &= 3.9\text{e-}9 * (70914 / 1193317) * (20672 / 270381) = 1.77\text{e-}11\end{aligned}$$

$$\begin{aligned}\text{Delta}(\text{一}/\text{c})1 &= \text{Delta}(\text{编辑}/\text{n})2 * a34(\text{编辑}/\text{n} \rightarrow \text{一}/\text{c}) * b4(\text{一}/\text{c}) \\ &= 8.49\text{e-}9 * (55177 / 1539367) * (2229 / 168350) = 4\text{e-}12\end{aligned}$$

$$\begin{aligned}\text{Delta}(\text{一}/\text{c})2 &= \text{Delta}(\text{编辑}/\text{v})2 * a34(\text{编辑}/\text{v} \rightarrow \text{一}/\text{c}) * b4(\text{一}/\text{c}) \\ &= 3.9\text{e-}9 * (13715 / 1193317) * (2229 / 168350) = 5.9\text{e-}13\end{aligned}$$



Veterbi算法词性标注过程示例(续)

一 -> 下

$$\begin{aligned}\text{Delta(下/v)1} &= \text{Delta(一/m)2} * a_{45}(\text{一/m} \rightarrow \text{下/v}) * b_5(\text{下/v}) \\ &= 1.77\text{e-}11 * (13778 / 270381) * (2271 / 1193317) = 1.7\text{e-}15\end{aligned}$$

$$\begin{aligned}\text{Delta(下/v)2} &= \text{Delta(一/c)1} * a_{45}(\text{一/c} \rightarrow \text{下/v}) * b_5(\text{下/v}) \\ &= 4\text{e-}12 * (40148 / 168350) * (2271 / 1193317) = 1.8\text{e-}15\end{aligned}$$

$$\begin{aligned}\text{Delta(下/q)1} &= \text{Delta(一/m)2} * a_{45}(\text{一/m} \rightarrow \text{下/q}) * b_5(\text{下/q}) \\ &= 1.77\text{e-}11 * (139653 / 270381) * (161 / 155374) = 9.47\text{e-}15\end{aligned}$$

$$\begin{aligned}\text{Delta(下/q)2} &= \text{Delta(一/c)1} * a_{45}(\text{一/c} \rightarrow \text{下/q}) * b_5(\text{下/q}) \\ &= 4\text{e-}12 * (53 / 168350) * (161 / 155374) = 1.3\text{e-}18\end{aligned}$$

$$\begin{aligned}\text{Delta(下/f)1} &= \text{Delta(一/m)2} * a_{45}(\text{一/m} \rightarrow \text{下/f}) * b_5(\text{下/f}) \\ &= 1.77\text{e-}11 * (1470 / 270381) * (6313 / 110878) = 5.47\text{e-}15\end{aligned}$$

$$\begin{aligned}\text{Delta(下/f)2} &= \text{Delta(一/c)1} * a_{45}(\text{一/c} \rightarrow \text{下/f}) * b_5(\text{下/f}) \\ &= 4\text{e-}12 * (700 / 168350) * (6313 / 110878) = 9.47\text{e-}16\end{aligned}$$



HMM 与 Pre-HMM的对比

$\text{Pr ob}(tag | word) \times \text{Pr ob}(tag | previous\ n\ tags)$ Pre-HMM

$\text{Pr ob}(word | tag) \times \text{Pr ob}(tag | previous\ n\ tags)$ HMM

Pre-HMM: CLAWS, Marshall(1983)

VOLSUNGA, DeRose(1988)

PARTS, Church(1988)

Jurafsky & Martin, 2000, *Speech and Language Processing*, Prentice Hall. Chapter 8, *Word Class and Part of Speech Tagging* 中译本《自然语言处理综论》pp.190-191, 199-200, 电子工业出版社2005年版



“race”的词性标注示例

Secretariat is expected to race tomorrow
NNP VBZ VBN TO ? NN
NN | VB

TO→NN 转移概率: $\text{Prob}(\text{NN}|\text{TO}) = 0.021$

TO→VB 转移概率: $\text{Prob}(\text{VB}|\text{TO}) = 0.34$

NN到race的输出概率: $\text{Prob}(\text{race}|\text{NN}) = 0.000041$

VB到race的输出概率: $\text{Prob}(\text{race}|\text{VB}) = 0.00003$

race标注NN的概率: $\text{Prob}(\text{NN}|\text{race}) = 0.98$

race标注VB的概率: $\text{Prob}(\text{VB}|\text{race}) = 0.02$



HMM vs. Pre-HMM

HMM

$\text{Prob}(\text{NN}|\text{TO}) = 0.021$
 $\text{Prob}(\text{race}|\text{NN}) = 0.00041$
 $\text{Prob}(\text{VB}|\text{TO}) = 0.34$
 $\text{Prob}(\text{race}|\text{VB}) = 0.00003$

$$\begin{aligned}\text{Prob}(\text{NN}|\text{race}, \text{TO}) &= 0.00041 * 0.021 \\ &= 0.00000861\end{aligned}$$

$$\begin{aligned}\text{Prob}(\text{VB}|\text{race}, \text{TO}) &= 0.00003 * 0.34 \\ &= 0.0000102\end{aligned}$$

Winner

Pre-HMM

$\text{Prob}(\text{NN}|\text{TO}) = 0.021$
 $\text{Prob}(\text{NN}|\text{race}) = 0.98$
 $\text{Prob}(\text{VB}|\text{TO}) = 0.34$
 $\text{Prob}(\text{VB}|\text{race}) = 0.02$

$$\begin{aligned}\text{Prob}(\text{NN}|\text{race}, \text{TO}) &= 0.98 * 0.021 \\ &= 0.01932\end{aligned}$$

$$\begin{aligned}\text{Prob}(\text{VB}|\text{race}, \text{TO}) &= 0.02 * 0.34 \\ &= 0.0068\end{aligned}$$

Winner



2.2 基于转换的错误驱动的词性标注方法

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, 然后用学习得到转换规则进行词性标注

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



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 T_1

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 T_1 ↓

S1: 他/r 做/v 了/u 一/m 个/q 报告/n



转换规则的模板 (template)

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 t1 = z THEN x -> y
```

```
If t1 = m, t2=q, THEN v -> n
```

```
...
```

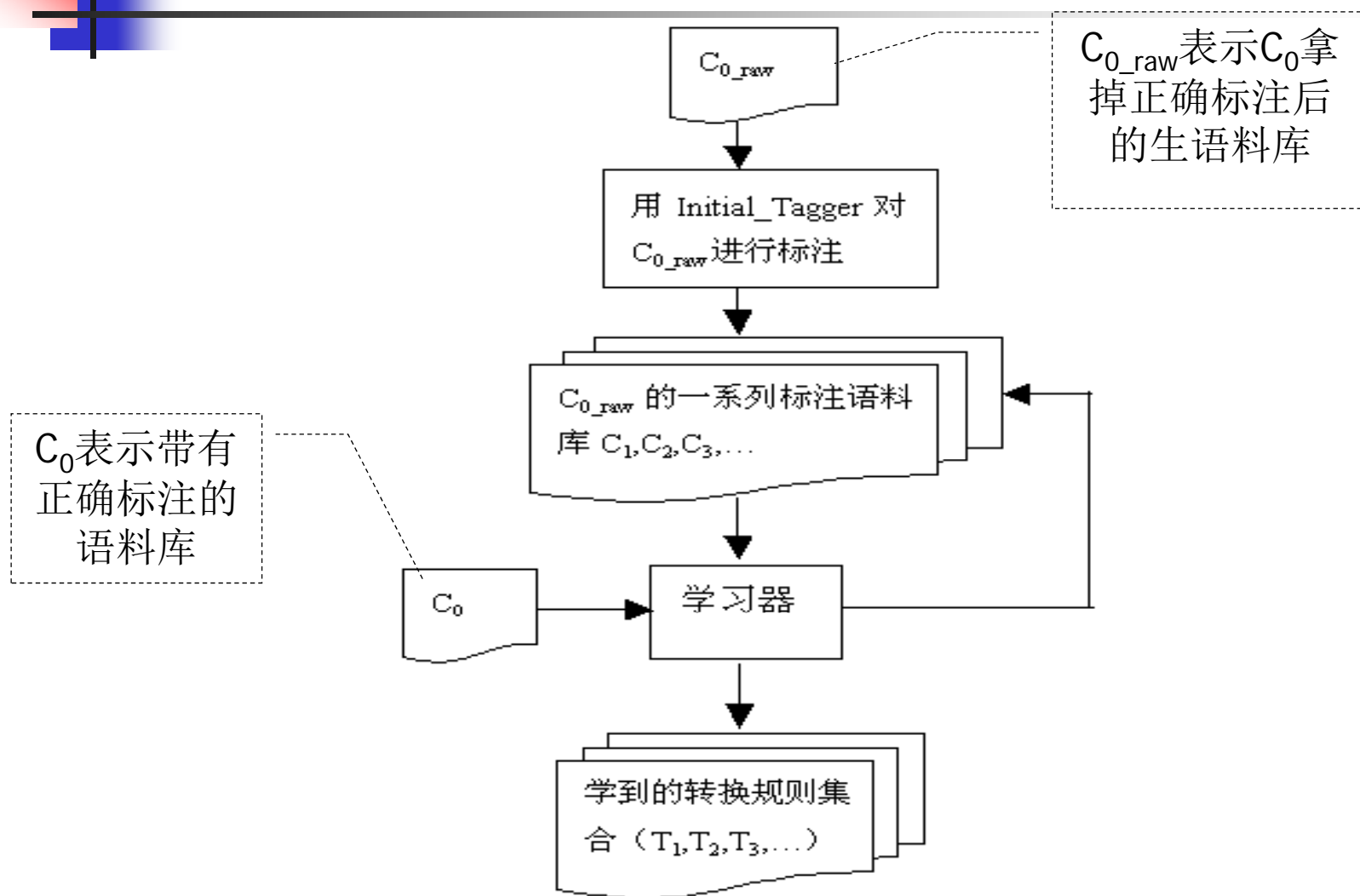


根据模板可能学到的转换规则

- T_1 : 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);
- T_2 : 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);
- T_3 : 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);
- T_4 : 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

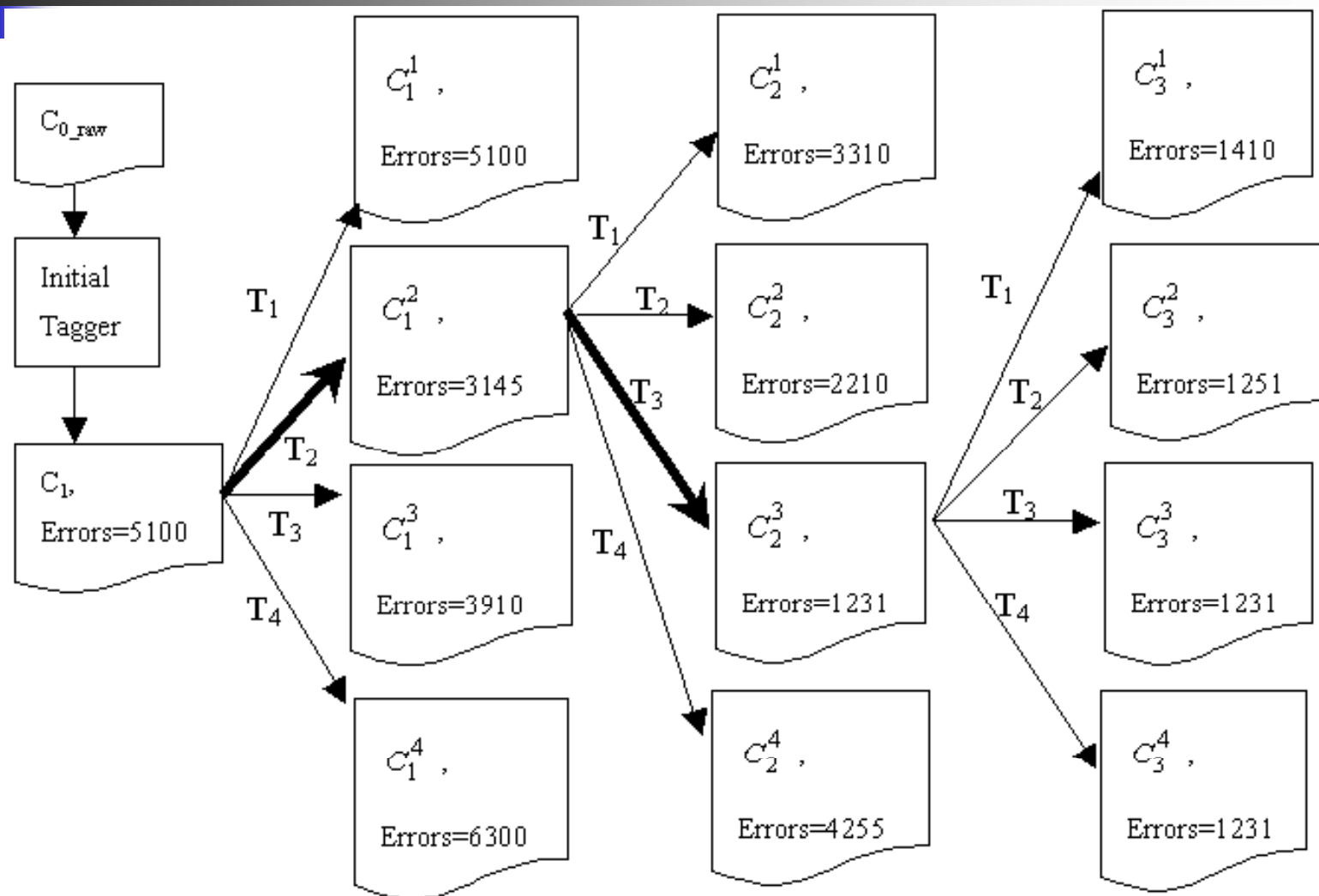




转换规则学习器算法描述

- 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





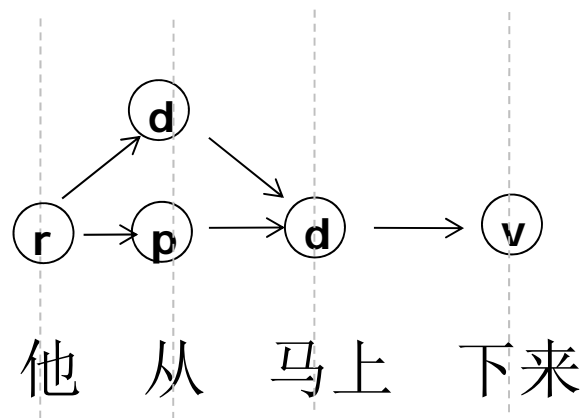
3 小结

1. 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，神经元模型，最大熵模型，SVM.....
 - 不同统计方法的融合
 - 词性标注与分词过程的融合

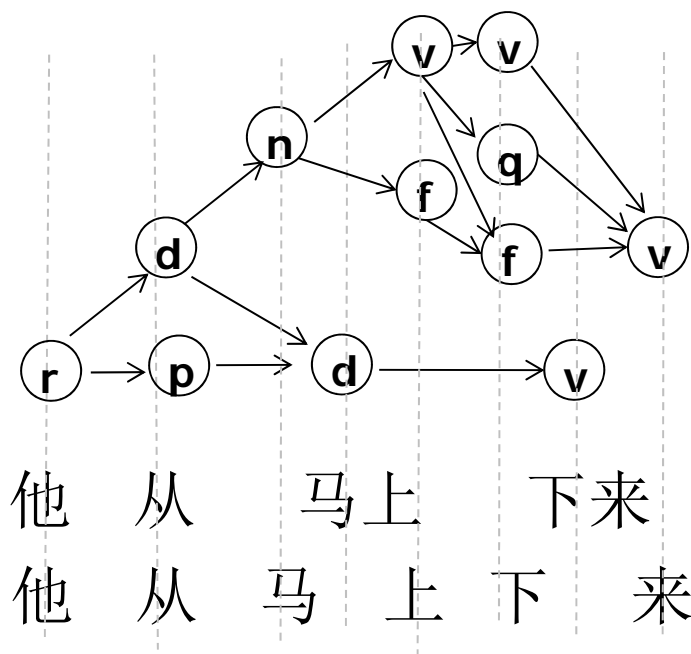
刘群 等，2004，基于层叠隐马模型的汉语词法分析，《计算研究与发展》2004年第8期。

分词和词性标注一体化示例

他从马上下来



先分词再标注



分词标注一体化



小结（续）

2. 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。

从NLP的应用需求出发，对词语进行“聚类”，或许是比“分类”更好的一个视角。

少 ←———— 词性标记 —————→ 多

粗 ←———— 语法特征 —————→ 细

简 ←———— 语法规则 —————→ 繁



进一步阅读文献

- Klein, Sheldon & R. F. Simmons. 1963. A Computational Approach to Grammatical Coding of English Words. *Journal of the Association for Computing Machinery*, Vol. 10, No. 3: 334-347.
- Marshall, I. (1983). Choice of Grammatical Word-class without Global Syntactic Analysis: Tagging Words in the LOB Corpus. *Computers and the Humanities* 17, 139-50.
- Garside, R. (1987). The CLAWS Word-tagging System. In: R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.
- Steven J. DeRose, (1988) "Grammatical Category Disambiguation by Statistical Optimization," *Computational Linguistics* Vol.14, No.1: 31-39.
- Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, December 1995. Vol.21, No.4
- L.R.Rabiner, 1989, A Tutorial on Hidden Markov Models and selected applications in Speech Recognition, *Proceedings of IEEE* vol.77, no.2, pp257-286
- 刘开瑛, 2000, 《中文文本自动分词和标注》, 商务印书馆, 第7章。
- 陈小荷, 2000, 《现代汉语自动分析》, 北京语言文化大学出版社, 第10章
- 杨尔弘 等, 2006, 汉语自动分词和词性标注评测, 《中文信息学报》2006年第1期。



复习思考题

1. 试构造一个汉语词性标注的实例，说明用Viterbi算法进行词性标注的过程。
2. 对汉语中的兼类词进行分析，撰写词的兼类问题的分析报告，尝试给出词类判定的语言学规则