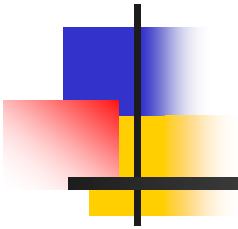
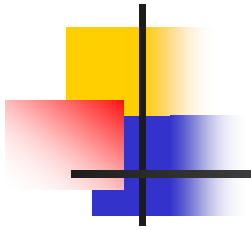


第七章 词汇分析（二）

—— 从词串到词性标记串

詹卫东

http://ccl.pku.edu.cn/doubtfire/



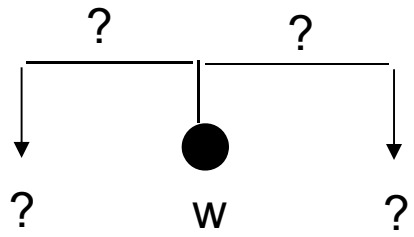
提纲

- 1 词性标注与兼类词
- 2 隐马尔可夫模型（HMM）
- 3 Viterbi算法
- 4 基于转换的错误驱动的词性标注方法
- 5 小结

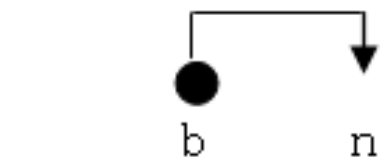
词类：为句法结构分析提供组合信息

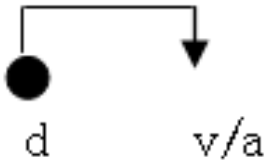
（1）词（w）的组合方向：　　w在参与序列组合时朝哪个方向组合；

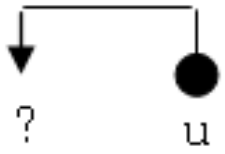
（2）词（w）的组合对象：　　w要求跟几个成分组合；
w要求跟什么类型的语言成分组合。

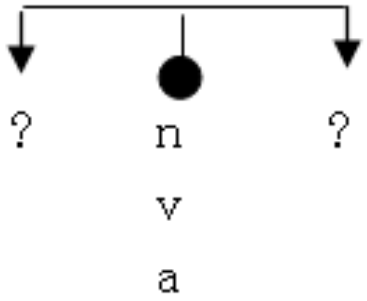


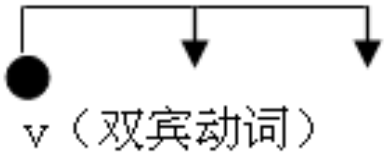
词类：为句法结构分析提供组合信息





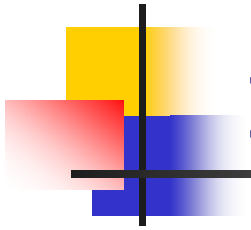






b: 区别词　d：副词　u：助词　v：动词　a: 形容词　n：名词

词类（分布）信息 = 组合方向 + 对组合对象的约束



1 词性标注（pos tagging）

- 语法体系 —— 词性标记集的确定
- 一词多类现象
  - Time flies like an arrow.

    Time/n-v flies/v-n like/p-v an/Det arrow/n

  - 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

词的兼类现象

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

词的兼类现象（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

词的兼类现象（续）

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	

词的兼类现象（English data, from Brown corpus）

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

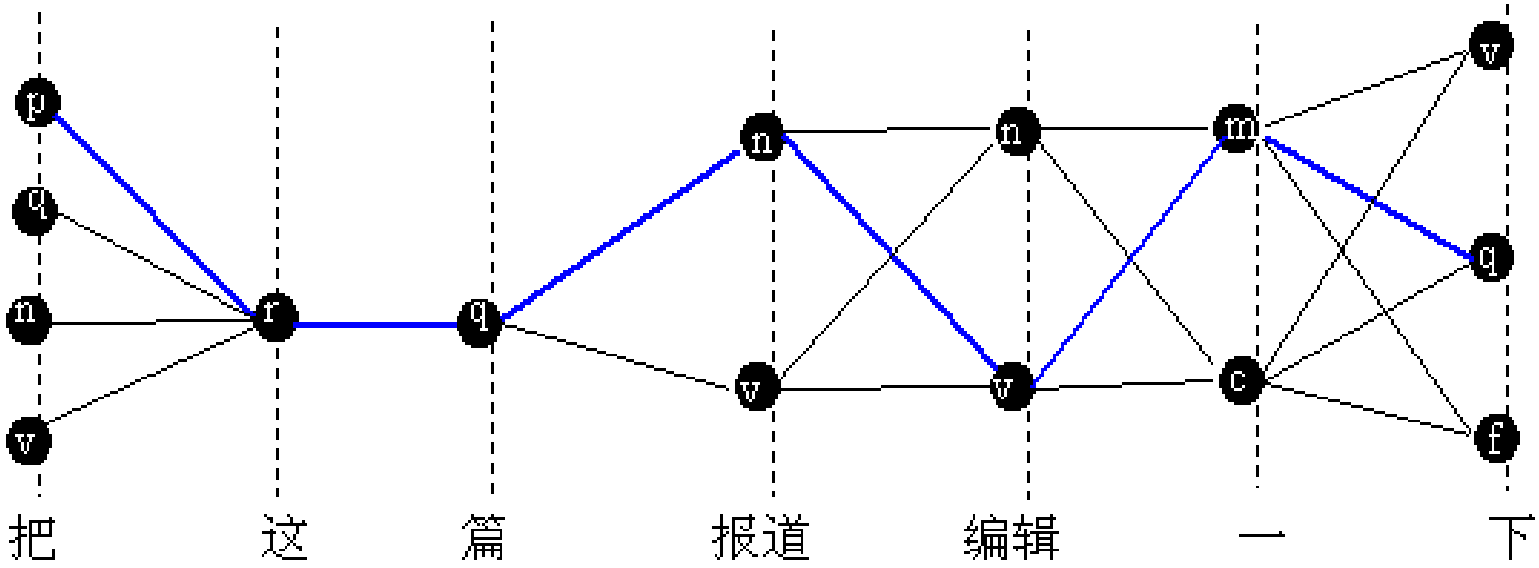
词性标注方法回顾

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%
……						

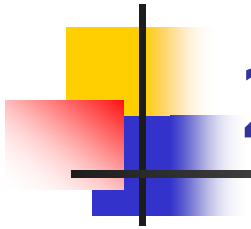
规则方法进行词性标注示例

```
@@ 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



4×1×1×2×2×2×3＝96种可能性，哪种可能性最大？

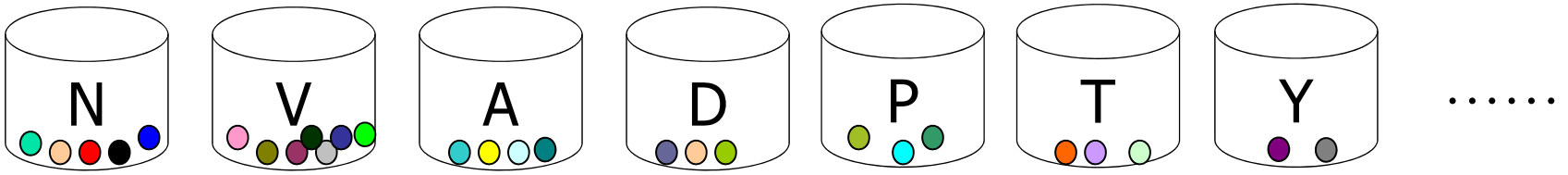


2 隐马尔可夫模型（Hidden Markov Model）

- 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

D P T N Y

A N D V V

A P V N V

……

HMM的形式描述

1 状态集合 $S=\{a_1,a_2,\ldots,a_N\}$，一般以q_t表示模型在t时刻的状态；

2 输出符号集合 $O=\{O_1,O_2,\ldots O_M\}$；

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：

$$a_{ij} = P(q_{t+1} = j \mid q_t = i), 1 \le i, j \le N \qquad a_{ij} \ge 0 \qquad \sum_{j=1}^{N} a_{ij} = 1$$

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：

$$b_j(k) = P(O_t = v_k \mid q_t = S_j), 1 \le j \le N, 1 \le K \le M \qquad b_j(k) \ge 0 \qquad \sum_{k=1}^{M} b_j(k) = 1$$

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

$$\pi_i = P(q_1 = S_i), 1 \le i \le N \qquad \pi_i \ge 0 \qquad \sum_{i=1}^{N} \pi_i = 1$$

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$O=O_1O_2\ldots O_T$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$O=O_1O_2\ldots O_T$和模型λ，如何计算状态序列$Q=q_1q_2\ldots q_T$，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$O=O_1O_2\ldots O_T$，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$W=w_1w_2\ldots w_n$），需要去猜测罐子序列（隐藏在幕后的词性标记序列，记做$T=t_1t_2\ldots t_n$）
- 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, λ)值最大的那个T′，一般记做：

$$T' = \arg\max_{T} P(T \mid W, \lambda) \qquad \text{公式1}$$

基于HMM进行词性标注（续）

- 根据条件概率公式可得

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

- 公式2可进一步简化为（根据Bayes公式）：

$$P(T \mid W) = \frac{P(T, W)}{P(W)} = \frac{P(T)P(W \mid T)}{P(W)} \qquad \text{公式3}$$

基于HMM进行词性标注（续）

- 公式3可以进一步简化为：

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

其中：

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

根据一阶HMM的独立性假设，可得

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

词性之间的转移概率可以从语料库中估算得到：

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

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：

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

- 根据HMM的独立性假设，公式8可简化为：

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

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

$P(T_1|W) = P(q|\$)P(把|q)P(r|q)P(这|r)\ldots P(f|m)P(下|f)$

$P(T_2|W) = P(q|\$)P(把|q)P(r|q)P(这|r)\ldots P(q|m)P(下|q)$

$P(T_3|W) = P(q|\$)P(把|q)P(r|q)P(这|r)\ldots P(v|m)P(下|v)$

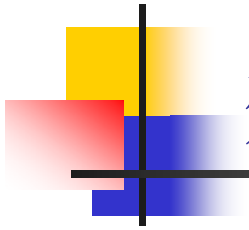
……

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

从中选一个最大值

词性转移概率

词语输出概率



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……

3 Veterbi算法 —— 提高效率之道

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

  安德鲁·维特比（Andrew J. Viterbi）于1967年提出

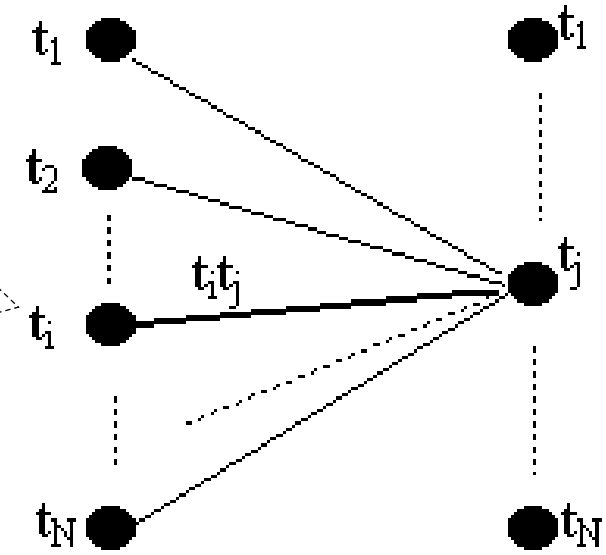
  https://en.wikipedia.org/wiki/Viterbi_algorithm

  Viterbi A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*. April 1967, 13 (2): 260–269. doi:10.1109/TIT.1967.1054010. (note: the Viterbi decoding algorithm is described in section IV.) Subscription required.

  http://ieeexplore.ieee.org/document/1054010/?arnumber=1054010

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条路经中存在一条概率最大的路径，假定为t_it_j



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

   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 $b_i(w_m) = P(w_m \mid t_i)$ $1 \le i \le N$

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，必有一条路经使得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

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

定义与记号（续）

4 HMM的状态从第m个词转移到第m+1个词，整个路径的概率可以通过HMM在前一个状态（第m个词）时的最大概率来求得，即Viterbi变量可以递归求值

$$\delta_{m+1}(j) = [\max_{1\le i\le N} \delta_m(i)a_{ij}] \times b_j(w_{m+1}) \qquad 1\le \mathrm{m} \le \mathrm{M\text{-}1},\ \ 1\le \mathrm{j} \le \mathrm{N}$$

5 当扫描过第m个词，状态转移到第m+1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

$$\Delta_m(i) = \arg\max_{1\le j\le N}[\delta_{m-1}(j)a_{ji}] \times b_i(w_m) \qquad 2\le \mathrm{m} \le \mathrm{M},\ \ 1\le \mathrm{i} \le \mathrm{N}$$

Veterbi算法

（1）初始化：

$$\delta_1(i) = \pi_i b_i(w_1) \quad 1 \le \mathrm{i} \le \mathrm{N} \qquad \Delta_1(i) = 0$$

（2） 递归计算通向每个词(w_m)的每个词性标记(t_j)的最佳路径

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

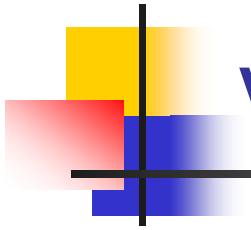
$$\Delta_m(j) = \arg\max_{1 \le i \le N} [\delta_{m-1}(i) \times a_{ij}] \times b_j(w_m) \; 2 \le \mathrm{m} \le \mathrm{M}, \quad 1 \le \mathrm{i} \le \mathrm{N}$$

（3） 到达最后一个词(w_M)时，计算这个词的最佳词性标记

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

（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

$$t_m^* = \Delta_{m+1}(t_{m+1}^*) \qquad \mathrm{m} = \mathrm{M\text{-}1, M\text{-}2, \ldots, 2,1}$$



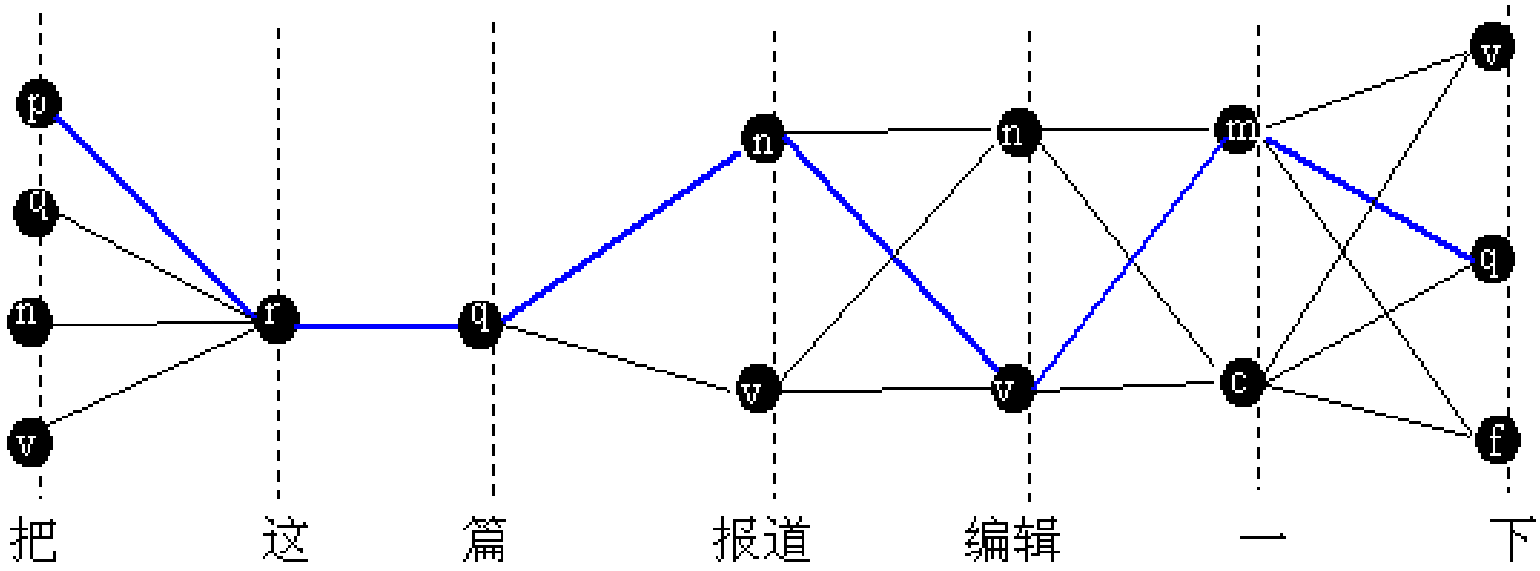
Veterbi算法的复杂度

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，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，扫描完整个词串（长度为M），计算次数为M个N^2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，Veterbi算法的计算复杂性是线性的。

用Veterbi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

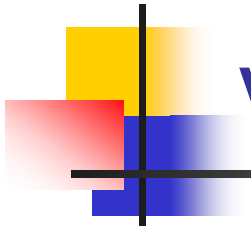
词语/词性频度表（用于估算输出概率）

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

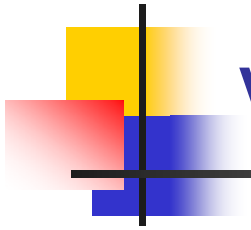
把 -> 这

Delta(这/r)1 = a12(把/p -> 这/r) * b2(这/r) = (27249 / 269186) * (21990 / 214942) = 0.01036

Delta(这/r)2 = a12(把/q -> 这/r) * b2(这/r) = (760 / 155374) * (21990 / 214942) = 5e-4

Delta(这/r)3 = a12(把/n -> 这/r) * b2(这/r) = (9769 / 1539367) * (21990 / 214942) = 6.49e-4

Delta(这/r)4 = a12(把/v -> 这/r) * b2(这/r) = (46697 / 1193317) * (21990 / 214942) = 0.004



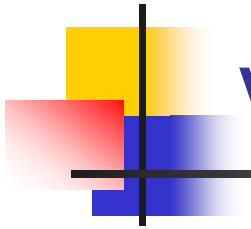
Veterbi算法词性标注过程示例(续)

篇 -> 报道

Delta(篇) 只有一个，略去。

Delta(报道/n)1 = a12(篇/q -> 报道/n) * b2(报道/n) = (52310 / 155374) * (420 / 1539367) = 9.1857e-5

Delta(报道/v)1 = a12(篇/q -> 报道/v) * b2(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

Delta(编辑/n)1 = Delta(报道/n)1 * a23(报道/n -> 编辑/n) * b3(编辑/n)

= 9.1857e-5 * (277181 / 1539367) * (243 / 1539367) = 2.6e-9

Delta(编辑/n)2 = Delta(报道/v)1 * a23(报道/v -> 编辑/n) * b3(编辑/n)

= 2.8954e-4 * (221796 / 1193317) * (243 / 1539367) = 8.49e-9

Delta(编辑/v)1 = Delta(报道/n)1 * a23(报道/n -> 编辑/v) * b3(编辑/v)

= 9.1857e-5 * (221776 / 1539367) * (100 / 1193317) = 1.1e-9

Delta(编辑/v)2 = Delta(报道/v)1 * a23(报道/v -> 编辑/v) * b3(编辑/v)

= 2.8954e-4 * (191967 / 1193317) * (100 / 1193317) = 3.9e-9

Veterbi算法词性标注过程示例(续)

编辑 -> 一

Delta(一/m)1 = Delta(编辑/n)2 * a34(编辑/n -> 一/m) * b4(一/m)

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

Delta(一/m)2 = Delta(编辑/v)2 * a34(编辑/v -> 一/m) * b4(一/m)

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

Delta(一/c)1 = Delta(编辑/n)2 * a34(编辑/n -> 一/c) * b4(一/c)

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

Delta(一/c)2 = Delta(编辑/v)2 * a34(编辑/v -> 一/c) * b4(一/c)

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

Veterbi算法词性标注过程示例(续)

一 -> 下

Delta(下/v)1 = Delta(一/m)2 * a45(一/m -> 下/v) * b5(下/v)

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

Delta(下/v)2 = Delta(一/c)1 * a45(一/c -> 下/v) * b5(下/v)

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

Delta(下/q)1 = Delta(一/m)2 * a45(一/m -> 下/q) * b5(下/q)

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

Delta(下/q)2 = Delta(一/c)1 * a45(一/c -> 下/q) * b5(下/q)

= 4e-12 * (53 / 168350) * (161 / 155374) = 1.3e-18

Delta(下/f)1 = Delta(一/m)2 * a45(一/m -> 下/f) * b5(下/f)

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

Delta(下/f)2 = Delta(一/c)1 * a45(一/c -> 下/f) * b5(下/f)

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

估算HMM的参数（Baum-Welch算法）

从生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 π_i ；

（2）词性转移概率 $a_{ij} = P(t_j \mid t_i)$ ；

（3）已知词性条件下词语的输出概率 $b_i(w_m) = P(w_m \mid t_i)$ ；

参看：翁富良，1998，第8章第2节

L.Rabiner, 1989

4 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://research.microsoft.com/~brill/

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；

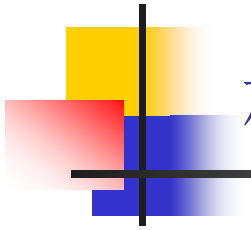
……

其中x，y，z是任意的词性标记代码。

If t_{-1} = z THEN x -> y

If t_{-1} = m, t_{-2}=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）；

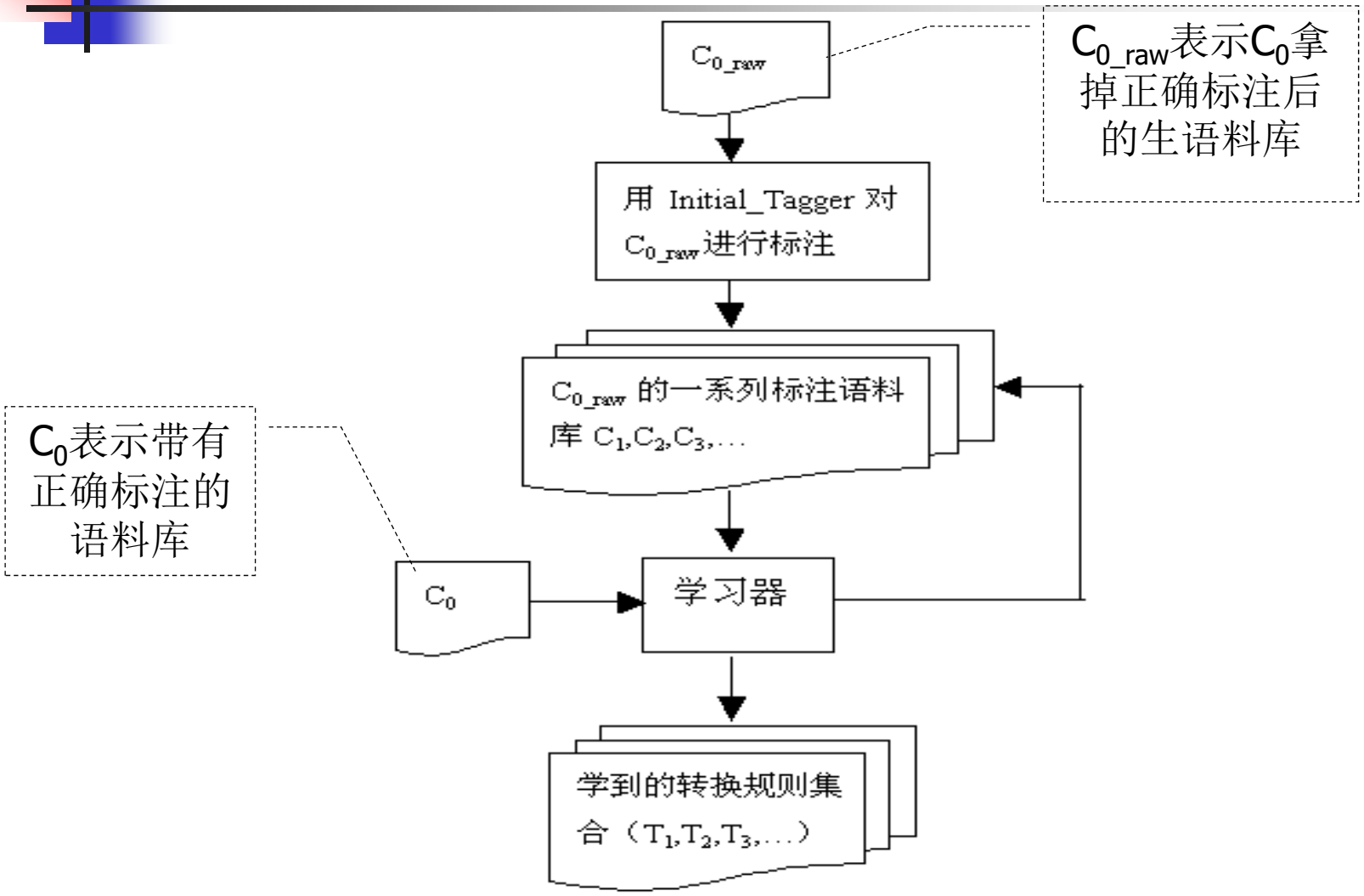
T_2：当前词的后一个词的词性标记是动词（v）时，将当前词的词性标记由动词（v）改为名词（n）；

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


C_{0_raw}
用 Initial_Tagger 对 C_{0_raw} 进行标注
C_{0_raw} 的一系列标注语料库 $C_1, C_2, C_3, \ldots$
C_0
学习器
学到的转换规则集合（$T_1, T_2, T_3, \ldots$）
C_{0_raw}表示C_0拿掉正确标注后的生语料库
C_0表示带有正确标注的语料库

转换规则学习器算法描述

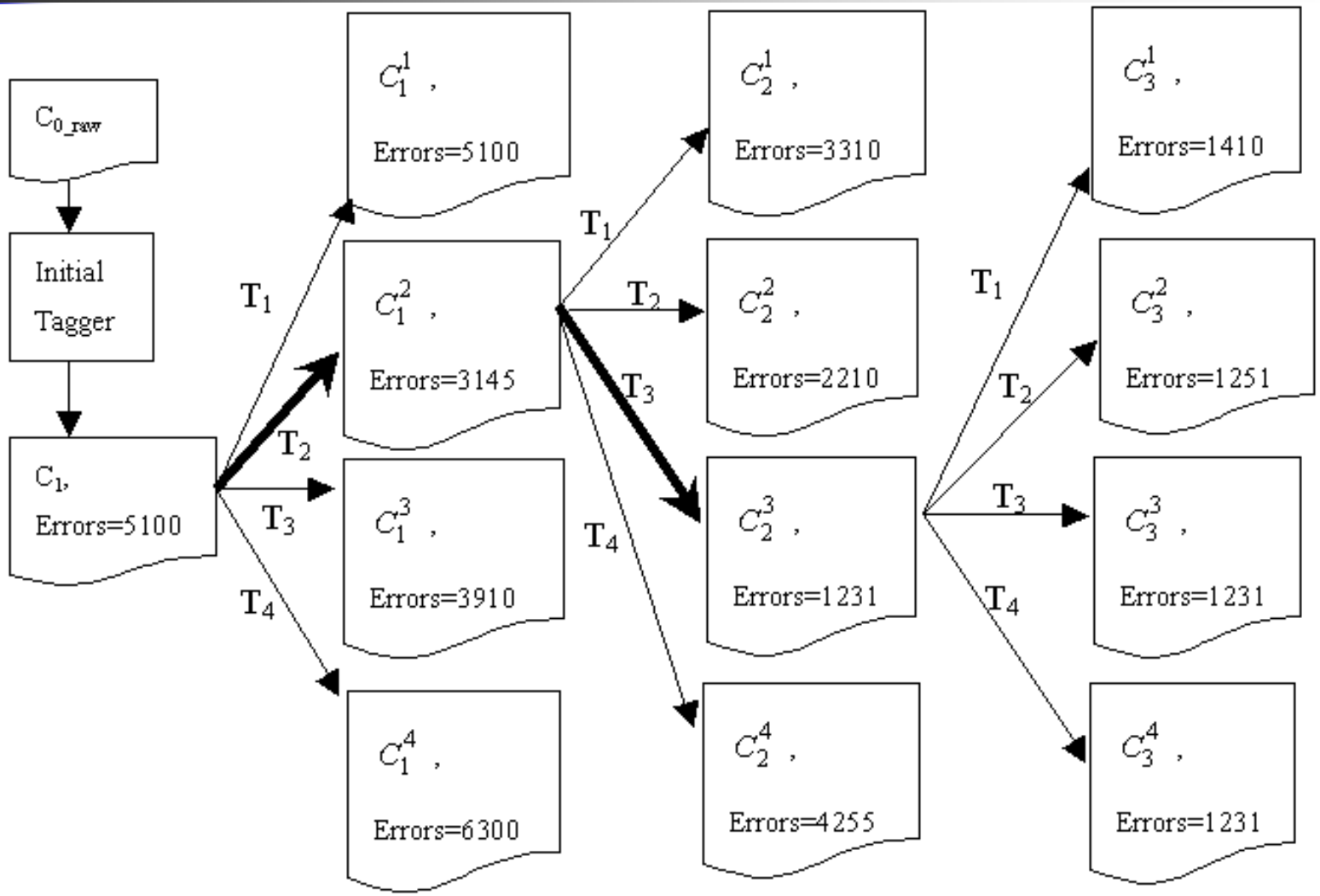
1) 首先用初始标注器对C_{0_raw}进行标注，得到带有词性标记的语料C_i (i =1)；

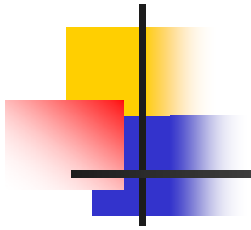
2) 将C_i跟正确的语料标注结果C_0比较，可以得到C_i中总的词性标注错误数；

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


C_{0_raw}
Initial Tagger
C_1, Errors=5100
C_1^1, Errors=5100
C_1^2, Errors=3145
C_1^3, Errors=3910
C_1^4, Errors=6300
C_2^1, Errors=3310
C_2^2, Errors=2210
C_2^3, Errors=1231
C_2^4, Errors=4255
C_3^1, Errors=1410
C_3^2, Errors=1251
C_3^3, Errors=1231
C_3^4, Errors=1231
T_1
T_2
T_3
T_4



小结

1. 统计方法与机器学习规则的方法在词性标注中有比较显著的优势
2. 人自身对汉语词性的认识还有不够清晰的地方，“兼类”、“依句辨品”？
3. 从语言学的角度看，词性标记集有粗细差异；
   从实际应用要求看，标注集也可以有多种选择；
   对词性标记集的选取和评价都是很重要的问题。

词类划分与词性判定的困难

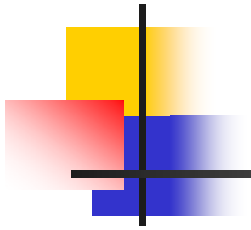
- 胡明扬主编（1996），《词类问题考察》，北京语言学院出版社
  胡明扬（2004），《词类问题考察续集》，同上

- 一手抓**教育**，一手抓……
  他很会**教育**孩子
  教育方法是非常重要的

  航空/？ 邮件/n　　双面/？ 打印/v

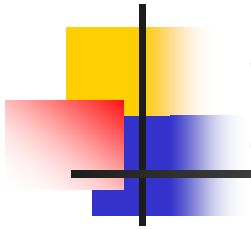
分类目的？　分类标准？　判定方法？

代词？　连词？　副词？　……



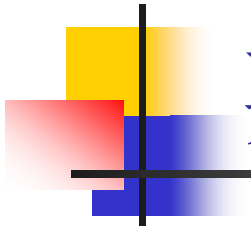
词性标记集的比较和评价

- 孙宏林（2000）《利用遗传算法实现词类标记集的优化》，发表于《多语言信息处理国际会议2000‘ICMIP论文集》，2000年8月，新疆. 乌鲁木齐（文章附录含有一个汉语语料库用到的词性标记集，107个标记）
  电子版下载
  http://icl.pku.edu.cn/research/papers/chinese/collection-4/15yichuan.htm
- 若干家汉语词性标记集（刘群提供）



附录一：词性标记集

- Upeen词性标记集（英语）
  - http://icl.pku.edu.cn/doubtfire/NLP/Lexical_Analysis/Word_Segmentation_Tagging/POS_Tagging_Overview/treebank-tags.html
- 北大计算语言所词性标记集（汉语）
  - http://www.icl.pku.edu.cn/research/corpus/addition.htm



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
- 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

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

UPenn treebank

POS tagset

(45 tags)

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

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（可能是比“词”大的单位）；也包含一些标记，用语标注语素、前接成份、后接成份等比词小的单位。

	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
			nrf	姓
			nrg	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

- 俞士汶 主编，1999， 现代汉语语料库加工——词语切分与词性标注规范与手册, 课题技术报告 http://icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm
- 俞士汶、段慧明、朱学锋、孙斌, 2002, 北京大学现代汉语语料库基本加工规范, 《中文信息学报》2002年第5期。Vol.15, No.5, pp.51-66
- 俞士汶、段慧明、朱学锋、孙斌、常宝宝，2003，北大语料库加工规范：切分·词性标注·注音，Journal of Chinese Language and Computing[新加坡], Vol.13, No.2, pp.121-158

附录二：词性标记歧义在语料中的分布

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

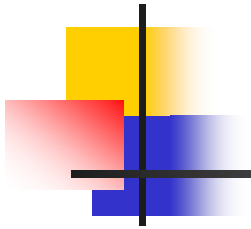
引自刘开瑛2000，p182

附录二：词性标记歧义在语料中的分布（续）

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

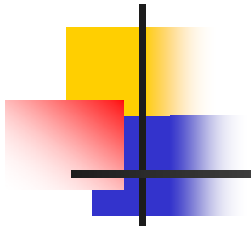
数据来源：
Brown 语料库

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



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



复习思考题

1. 试构造一个汉语词性标注的实例，说明用Veterbi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
3. 在互联网上找一篇3000到5000字之间的中文文章，手工进行分词和词性标注，归纳碰到的问题。