词性标注

詹卫东
北京大学中文系
zwd@pku.edu.cn
http://ccl.pku.edu.cn/doubtfire

Outline

1. 词性标注任务概述
2. 词性标注的意义
3. 中文的词类系统
4. 词性标注的基本方法
5. 小结

1 词性标注任务概述

词类：　　对一种语言全部词汇的语法分类。
part of speech (pos)　word class

词性标注：　对文本中**兼类词**的实际词类归属进行判定。

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

- Time flies like an arrow.

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

c	连词
Det	冠词
f	方位词
m	数词
n	名词
p	介词
q	量词
r	代词
v	动词

2 词性标注的意义

- **句法分析**
- **词义排歧** 利用词类标记可以消除超过1/ 5 的汉语词义歧义 [1]
- **语音识别**
- **语音合成**
- **信息检索** 从文本中提取全部名词建索引 vs. 对所有词建索引 [2] [3]
- **信息抽取**

[1] 王惠，2003，机器翻译中基于语法、语义知识库的汉语词义消歧研究，《广西师范大学学报》（自然科学版）2003年第1期，pp.86-93

[2] Chowdhury A ,McCabe M,Improving information retrieval systems using part of speech tagging[R] . Technical Report TR 1998 - 48, Institute for Systems Research ,University of Maryland ,1998.

[3] 苏琪 等，2005，词性标注对信息检索系统性能的影响，《中文信息学报》2005年第2期。pp.58-65.

语音合成

多音词

1. 睡眠时间不够**长**，个子就**长**不高。
2. 他家的**地**是出了名**地**多。
3. 这家伙真不**地道**，不顾老人和孩子，自己先钻**地道**逃跑。
4. 这次出差，**同行**的四个人中没有一个**同行**。
5. 难怪你总是**倒数**第一，这么简单的**倒数**你都算不出来。
6. Some persons **object**ed that the proposed import duty would harm the **object** of the treaty.
7. They were very **content** with the **contents** of the training.

3 中文的词类系统

词类的划分标准：

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

词的分布（位置）

	组合关系					
	a	b	c	d	e	f
聚合关系	张三	把	杯子	摔	破	了
	工厂	把	水沟	填	平	了
	妈妈	把	饭	煮	糊	了
	……					

词的分布（位置）

聚合关系	组合关系						
	a	b	g	c	d	e	f
	张三	把	也	杯子	摔	破	了
	工厂	把	居然	水沟	填	平	了
	妈妈	把	已经	饭	煮	糊	了
	……						

词的分布（位置）

聚合关系	组合关系						
	a	g	b	c	d	e	f
	张三	也	把	杯子	摔	破	了
	工厂	居然	把	水沟	填	平	了
	妈妈	已经	把	饭	煮	糊	了
	……						

词类：

a = c	名词
b	介词
d	动词
e	形容词
f	助词
g	副词
……	

词的分布（位置）：合法位置 vs. 非法位置

	a	b
1	**他 迅速 销毁了密码本**	**他 立即 销毁了密码本**
2	他 销毁密码本 很 迅速	✗ 他 销毁密码本 很立即
	果断 坚决 突然 仔细 用力 大胆 ……	马上 悄悄 立刻 已经 必定 亲自 ……
	形容词	副词

划分词类的基本思路

1. **给出语言中基本组合（结构）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；**

汉语词类归属测试网页　　http://ccl.pku.edu.cn:8084/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（结构）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

体词性位置：中心语1、主语、宾语
谓词性位置：谓语、中心语2、述语1、述语2、补语、前谓、后谓
[修饰性位置]：　定语、状语

指称
陈述
[修饰]

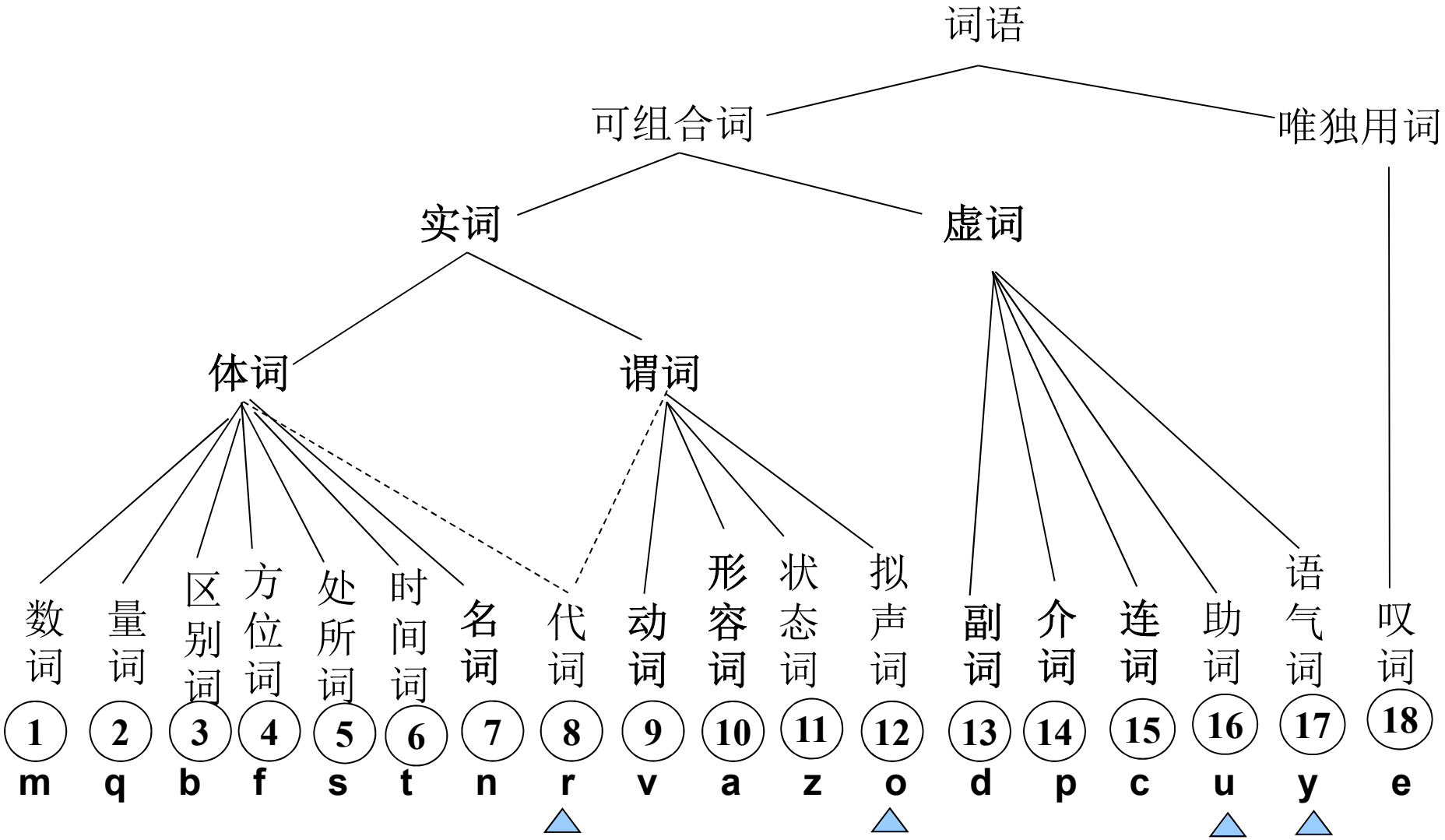
现代汉语的基本组合（结构）关系（续）

组合类型	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


词语
可组合词
唯独用词
实词
虚词
体词
谓词
数词 1 m
量词 2 q
区别词 3 b
方位词 4 f
处所词 5 s
时间词 6 t
名词 7 n
代词 8 r
动词 9 v
形容词 10 a
状态词 11 z
拟声词 12 o
副词 13 d
介词 14 p
连词 15 c
助词 16 u
语气词 17 y
叹词 18 e

实词 vs. 虚词

实词（content word）跟虚词（function word）的区别：

（1）功能：是否占据主要的句法结构位置

（2）意义：词汇意义还是语法意义

（3）自由与黏着：是否能单独使用

（4）位置：参与组合时位置是否固定

（5）开放与封闭：实词是**开放类（open class）**；虚词是**封闭类（closed class）**

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 _____

同类词的内部差异——观察分布从粗粒度到细粒度

例：“X心”的分布

	耐心	热心	信心	点心
有___	+	−	+	+
很___	+	+	−	−
很有___	+	−	+	−

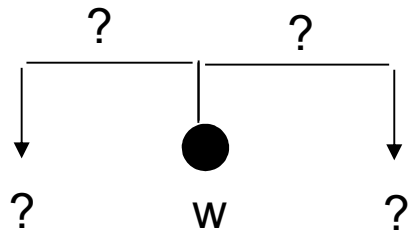
词类为句法结构分析提供的信息

（1）词（w）的组合方向：

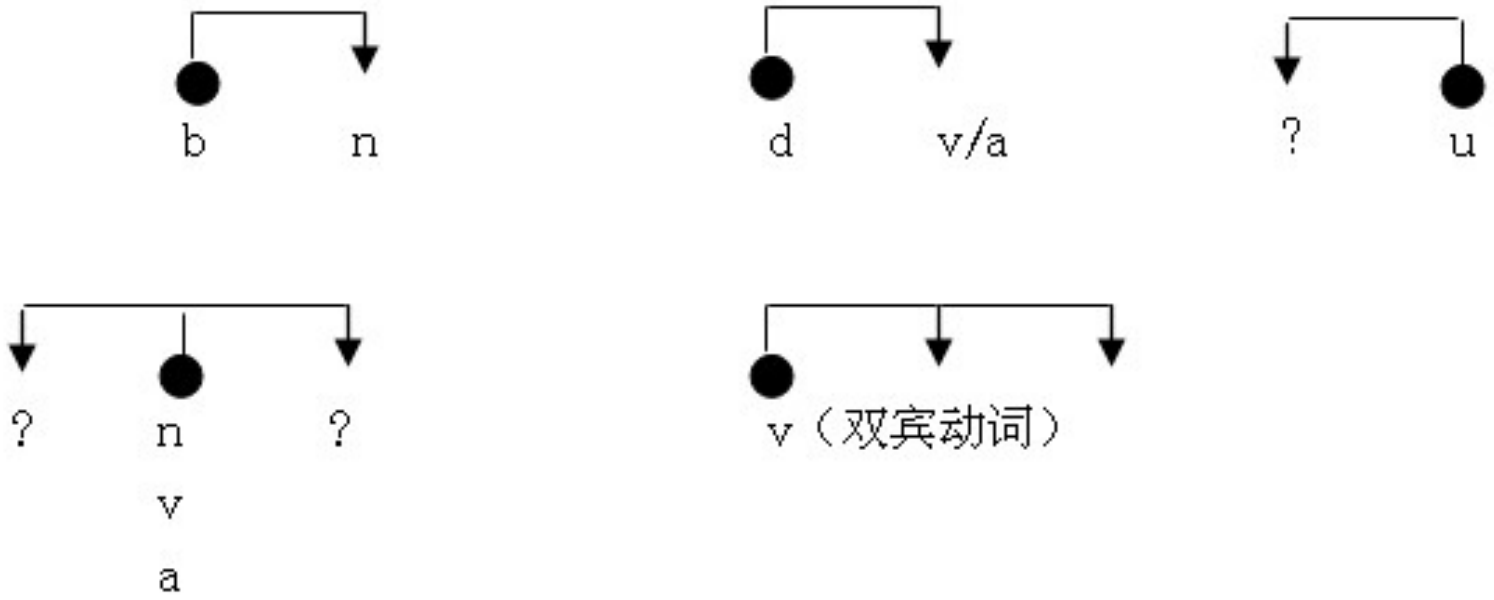
w在参与序列组合时朝哪个方向组合；

（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的信息（续）



b: 区别词　d：副词　u：助词　v：动词　a: 形容词　n：名词

词的兼类现象

- **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	兼类词	广义兼类词
2	两 把 锁	锁 两次		
3	一件 制服	制服 不了 小偷	同形词	
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	

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

span: 兼类词接续共现的个数。

例如：span = 2 表示两个兼类词接续共现

刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，182页。

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)

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

DeRose (1988)

数据来源：
Brown 语料库

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

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

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

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
			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

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** https://www.cis.upenn.edu/~treebank/
  - 45 tags
  - Used for UPenn treebank, Brown Corpus, WSJ Corpus
  - Brill tagger
- **UCREL's C5 tagset** http://ucrel.lancs.ac.uk/claws/
  - 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)

4 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%
……						

Jurafsky & Martin, 2000, *Speech and Language Processing*,: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall. Chapter 8, *Word Class and Part of Speech Tagging*

4.1 基于规则的词性标注

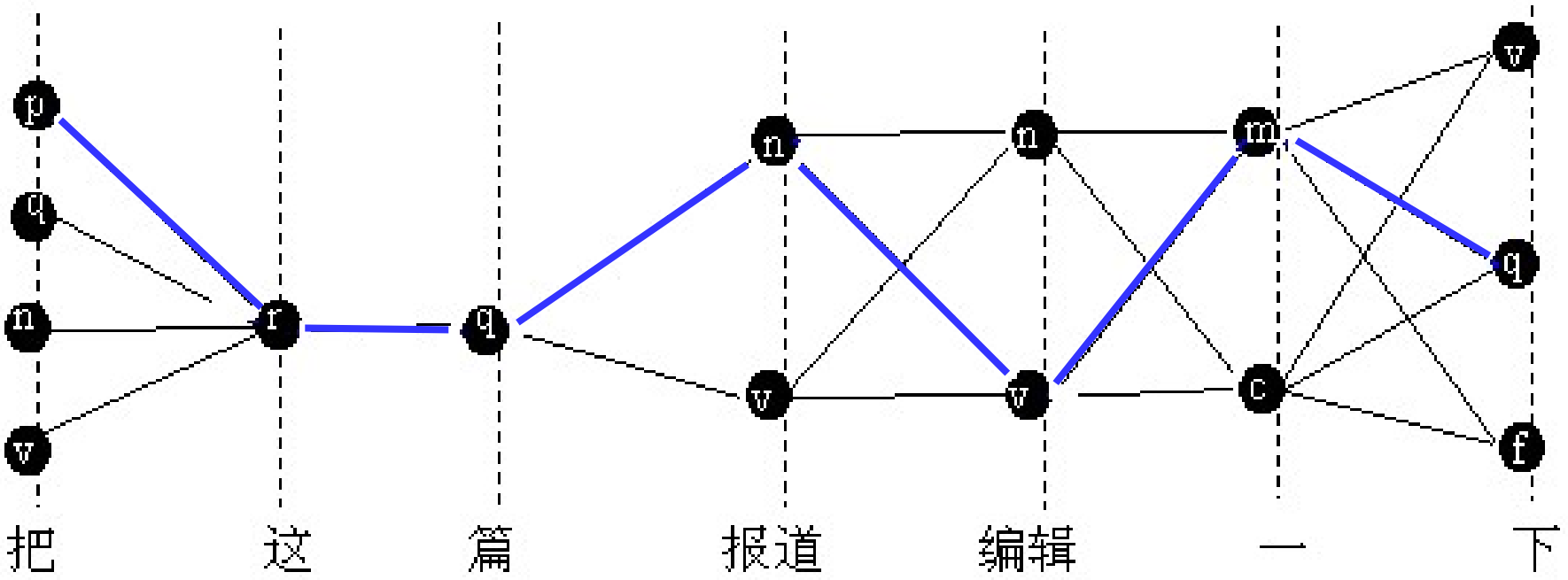
词性标注系统知识源：

（1） 词典：词条，词性标记
（2） 歧义消解规则库：上下文特征，词内部形式特征 → 确定词类标记

词性标注规则示例：

```
@@ 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
```

4.2 基于隐马尔可夫模型（HMM）的词性标注



4×1×1×2×2×2×3＝96种可能性，哪种可能性最大？

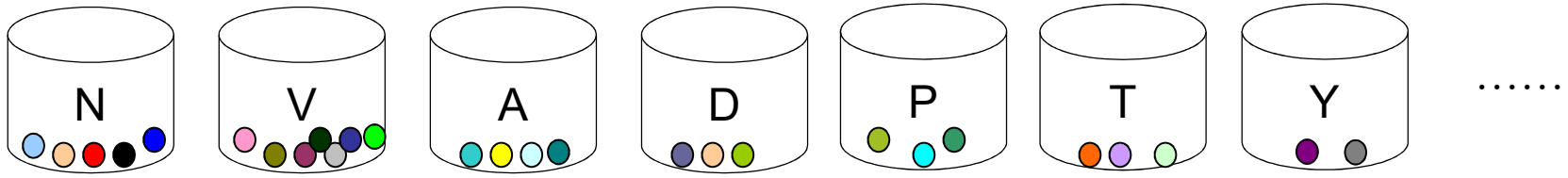
[1] Andrei Andreyevich Markov (1856-1922) http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Markov.html

[2] Daniel Jurafsky & James H. Martin, 2000, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall Inc.. Chapter 7, 8.

[3] 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

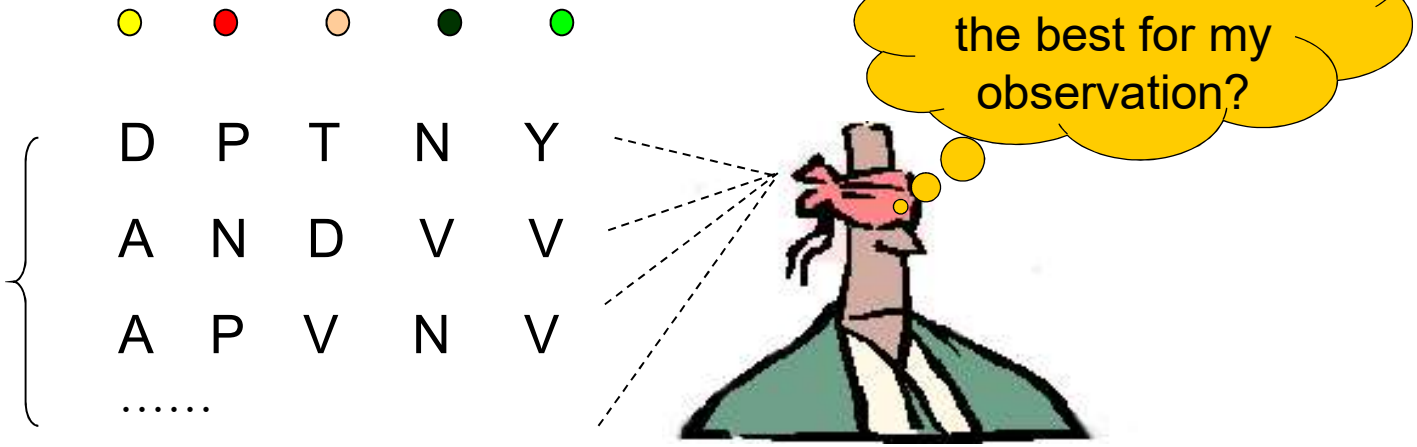
HMM概述

“词类罐子”：每个罐子（词类）中装了有限个词



可见的单词序列

隐藏的词类序列



HMM的形式描述

五元组：{S, O, A, B, π}　　也可简单地把 $\lambda = \{A, B, \pi\}$ 称为模型

1 **状态集合** $S=\{a_1,a_2,\ldots,a_N\}$;

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

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

对应着词性标注问题

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

基于HMM进行词性标注

- 两重随机过程：
  1 选择词罐子 —— 按照一定的转移概率随机地选择罐子
  2 选择每个罐子里的词 —— 按照一定的概率随机地从一个罐子中选择一个词输出

- 人只能看到词语序列（记做$W=w_1w_2...w_n$），需要去猜测词罐子序列（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：

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

基于HMM进行词性标注（续）

- **根据条件概率公式可得**

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

为简化描述，上式中 λ 省去

- **公式2可进一步简化为（根据Bayes公式）：**

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

基于HMM进行词性标注（续）

- **公式3可以进一步简化为：**

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

其中：

$$P(T) = P(t_1 \mid t_0)P(t_2 \mid t_1, t_0)...P(t_i \mid t_{i-1}, t_{i-2}, t_{i-3}, ...) \qquad \text{公式5}$$

根据一阶HMM的独立性假设，可得

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

词性之间的转移概率可以从语料库中估算得到：

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

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：

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

根据HMM的独立性假设，公式8可简化为：

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

$$P(w_i \mid t_i) = \frac{\text{训练语料中}w_i\text{的词性被标记为}t_i\text{的次数}}{\text{训练语料中}t_i\text{出现的总次数}}$$ 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K个词罐子，给定词串中有N个词，

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。

这意味着：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）**

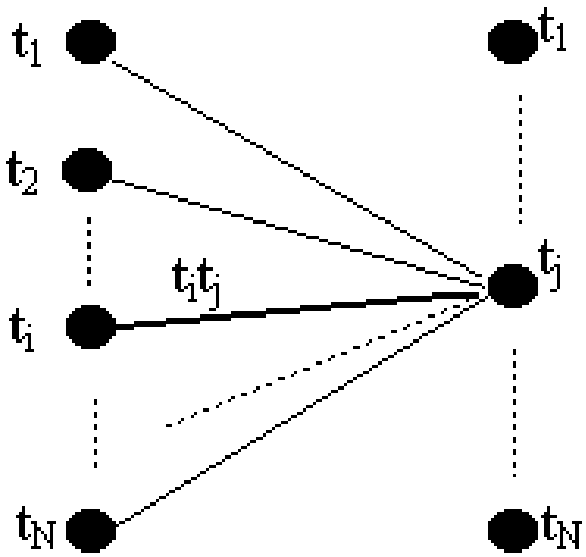
如果当前节点在最优路径上，那么，不管当前节点的后续路径如何，当前节点的来源路径必定是最优的。

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条路经中存在一条概率最大的路径，假定为t_it_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 $a_{ij} = P(t_j \mid t_i)$　　$1 \le i \le N;\ 1 \le j \le N$

   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-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

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

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

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

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） 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

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

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

（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K个词罐子，给定词串中有N个词，

考虑最坏的情况：每个词都有K个可能的词性标记，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。而扫描完整个词串（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算法的时间复杂性是线性的。

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

HMM 与 Pre-HMM的对比

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

$$\Pr ob(word \mid tag) \times \Pr ob(tag \mid previous \text{ n tags})$$ 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：Prob(NN|TO) = 0.021

TO→VB 转移概率： Prob(VB|TO) = 0.34

NN到race的输出概率：Prob(race|NN) = 0.000041

VB到race的输出概率： Prob(race|VB) = 0.00003

race标注NN的概率： Prob(NN|race) = 0.98

race标注VB的概率： Prob(VB|race) = 0.02

HMM vs. Pre-HMM

HMM

Prob(NN|TO) = 0.021
Prob(race|NN) = 0.00041

Prob(VB|TO) = 0.34
Prob(race|VB) = 0.00003

Prob(NN|race,TO) = 0.00041 * 0.021
= 0.00000861

Prob(VB|race,TO) = 0.00003 * 0.34
= 0.0000102 Winner

Pre-HMM

Prob(NN|TO) = 0.021
Prob(NN|race) = 0.98

Prob(VB|TO) = 0.34
Prob(VB|race) = 0.02

Prob(NN|race,TO) = 0.98 * 0.021
= 0.01932 Winner

Prob(VB|race,TO) = 0.02 * 0.34
= 0.0068

4.3 基于转换的错误驱动的词性标注方法

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

Eric Brill, 1995, 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

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i；

（2）当前词的前（后）面第二个词的词性标记是j；

（3）当前词的前（后）面两个词中有一个词的词性标记是k；

……

其中x，y，i, j, k是任意的词性标记代码。

激活环境 + 改写规则：

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

If t_{-1} = m, t_{-2}=q THEN v -> n

…

Brill(1995), 非词汇化标注器：转换规则模板 6 个。
词汇化标注器：转换规则模板新增8个，共14个。

根据模板可能学到的转换规则示例

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）；

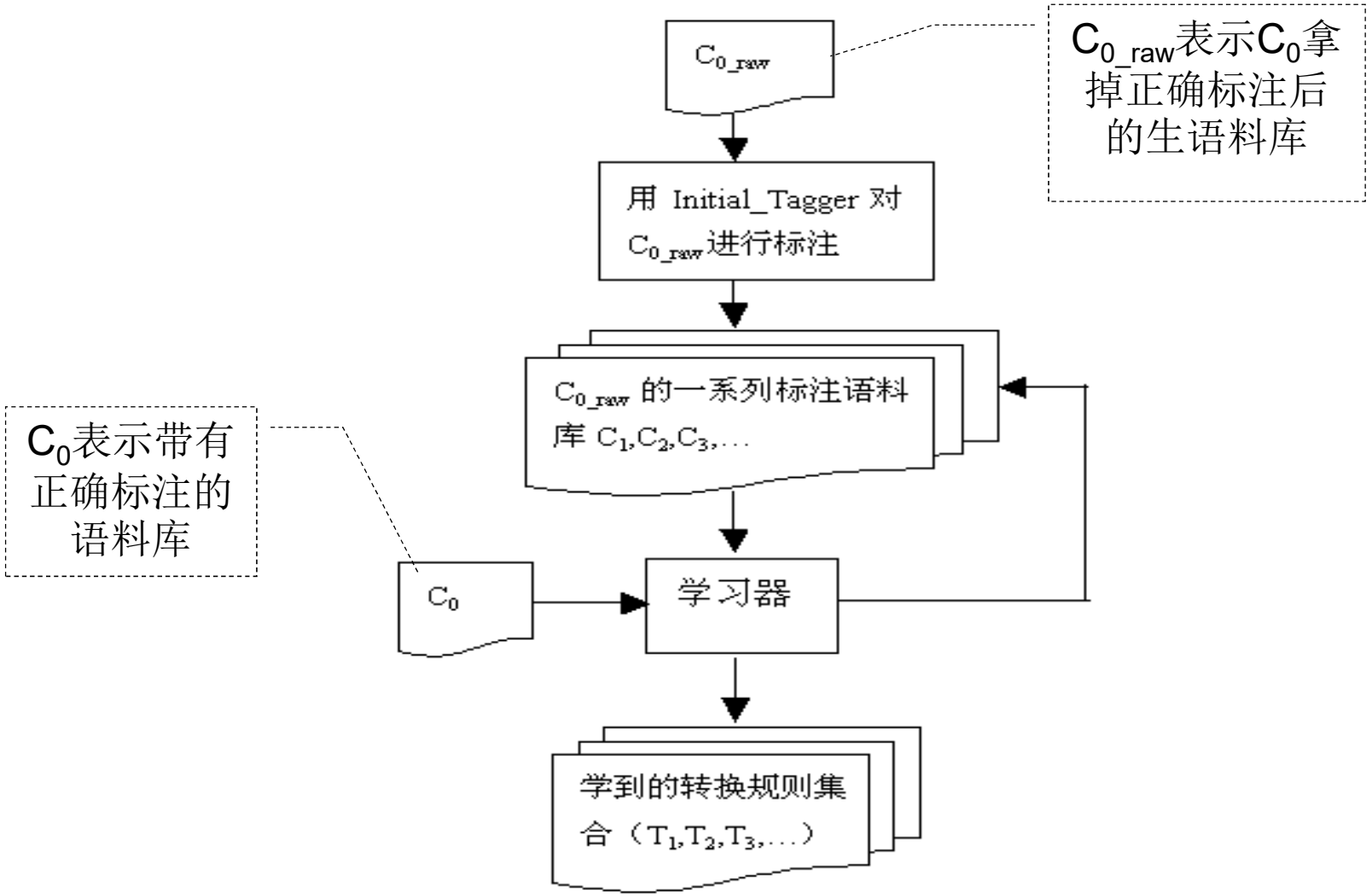
T_2：当前词的后一个词的词性标记是动词（v）时，将当前词的词性标记由动词（v）改为名词（n）；

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（a）改为数词（m）；

……

转换规则的学习流程


C_{0_raw}
C_{0_raw}表示C_0拿掉正确标注后的生语料库
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表示带有正确标注的语料库
C_0
学习器
学到的转换规则集合（$T_1, T_2, T_3, \ldots$）

转换规则学习器算法描述

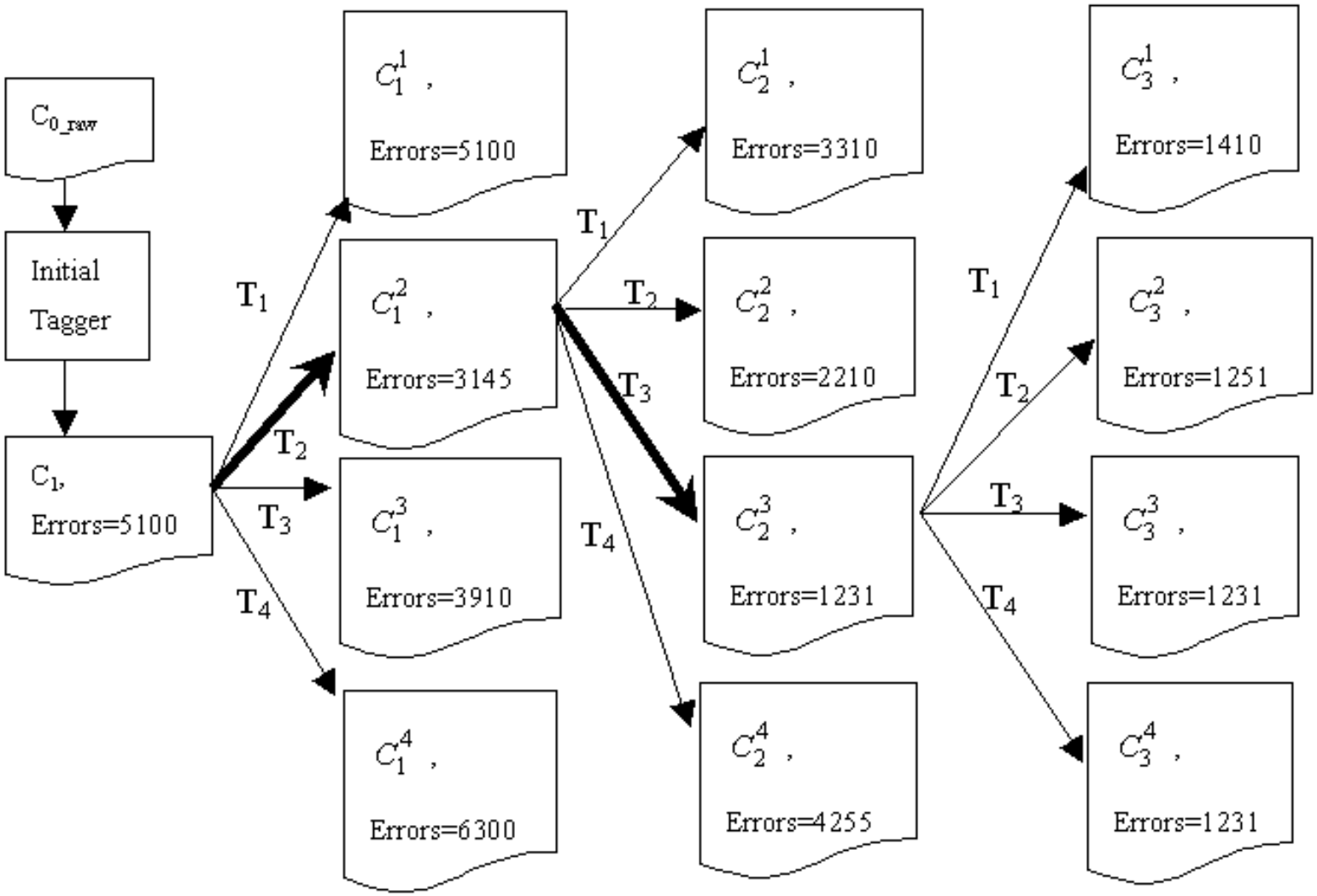
1) 首先用**初始标注器**对C_{0_raw}进行标注，得到带有词性标记的语料C_i (i =1)；

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$C_i \leftarrow C_i^j$。

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不多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, …}

转换规则学习过程示例



Eric Brill(1995)

4.4 分词和词性标注一体化

- 基于全切分（词图）的中文分词和词性标注一体化处理
- 基于字序列标注的中文分词和词性标注一体化处理

白栓虎，1995，《汉语词切分及标注一体化方法》，载陈力为、袁琦主编《计算语言学进展与应用》，清华大学出版社。pp. 56-61。

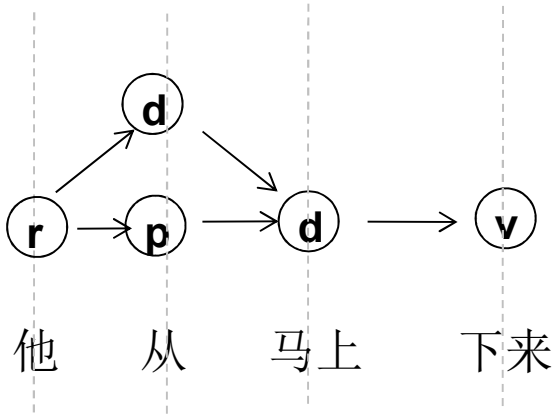
刘群 等，2004，基于层叠隐马模型的汉语词法分析，《计算机研究与进展》2004年第8期。pp.1421-1429。

朱聪慧 等，2010，基于无向图序列标注模型的中文分词词性标注一体化系统，《电子与信息学报》2010年第3期，pp. 700-704。

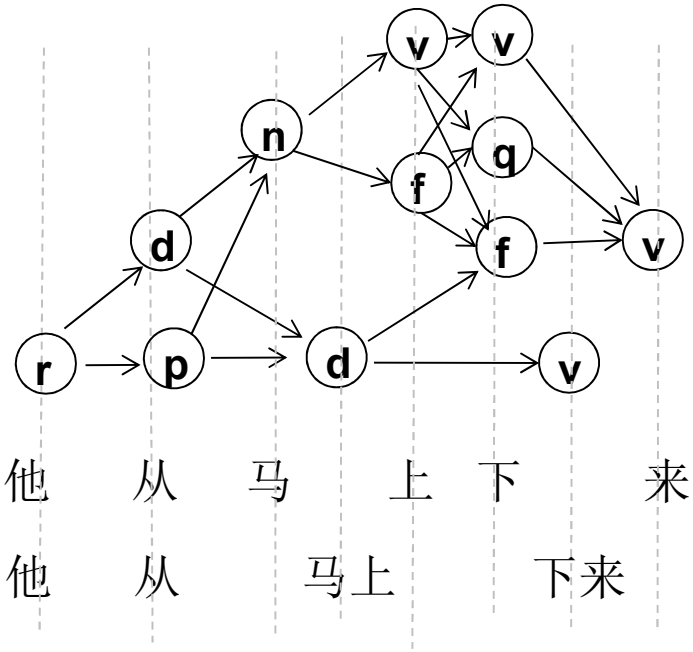
Hwee Tou Ng, Jin Kiat Low: Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? EMNLP 2004: 277-284.

分词和词性标注一体化词图示例

他从马上下来



先分词再标注



分词标注一体化

5 小结

- 词类及其划分原则
- 中文的词类系统
- 基于规则的词性标注方法
- 词性标注的HMM 模型
- 基于转换的错误驱动的标注方法

分词和词性标注的一体化

未登录词的词性标注

词性标注系统性能评测

汉语词类范畴缺乏形态标记。具体词语的词类归属存在“天然”的模糊性。

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.
- B.B.Greene and G.M. Rubin, 1971, Automatic Grammatical Tagging of English, Technical Report, Department of Linguistics, Brown University, Providence, Rhode Island
- 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
- Hwee Tou Ng, Jin Kiat Low: Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? EMNLP 2004: 277-284.
- 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。
- 刘群等，2004，基于层叠隐马模型的汉语词法分析，《计算机研究与进展》2004年第8期。

复习思考题

1. 找一篇3000字左右的中文文章，对其进行分词和词性标注，分析其中存在的问题。

2. 请谈谈你对现代汉语的词类系统的认识。