

# 计算语言学

## 第7讲 词法分析II

刘群

中国科学院计算技术研究所

liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

## 关于汉语词性 1

- 英语的词性
  - 英语的词性主要是由词的变化形式决定的
  - 英语的词性与词的句法功能存在着比较明确的一一对应关系
- 汉语的词性
  - 汉语词几乎没有形态变化
  - 汉语词性与所充当的语法功能不存在明确的一一对应关系
- 问题：如何确定汉语的词性？

## 关于汉语词性 2

- 问题1：词性判定的依据是什么？
  - 句法功能
  - 语义 ×
    - 金、银 ⇔ 铜、铁、锡
    - 红、黑、紫、灰、粉 ⇔ 红色、咖啡色
    - 战争 ⇔ 打仗

## 关于汉语词性 3

- 问题2：同一个词在充当不同句子成分时词性是否发生变化？
  - 我们调查了这件事。
  - 调查很重要。
  - 这件事很困难。
  - 我们不怕困难。
  - 去是有道理的。
  - 暂时不去是有道理的。

## 关于汉语词性 4

- 问题1的答案：词性判定的依据只能是句法功能；
- 问题2的答案：同一个词在充当不同句子成分时词性不发生变化。

参考文献：朱德熙，《语法答问》，商务印书馆，1985

## 汉语词性标记集

- 几个典型的词性标记集
  - 北京大学《人民日报》语料库标记集
  - 清华大学《汉语树库》词性标记集
  - 语用所《信息处理用现代汉语词类及词性标记集规范》
  - 宾州树库规范
  - 计算所词性标记集（V3.0）
- 参考：词性标记集对照表

# 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-a 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 ( 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/标注项目	标记集	方法，特点	处理语料规模	精确率
1	Klein&Simmons (1963)	30	手工规则	百科全书小样本	90%
2	TAGGIT (Greene&Rubin, 1971)	86	人工规则(3300条)	Brown语料库	77%
3	CLAWS (Marshall,1983; Booth, 1985)	130	概率方法 效率低	LOB语料库	96%
4	VOLSUNGA (DeRose,1988)	97	概率方法 效率高	Brown语料库	96%
5	Eric Brill's tagger (1992-94)	48	机器规则(447条) 效率高	UPenn WSJ语料库	97%

## 规则方法进行词性标注示例

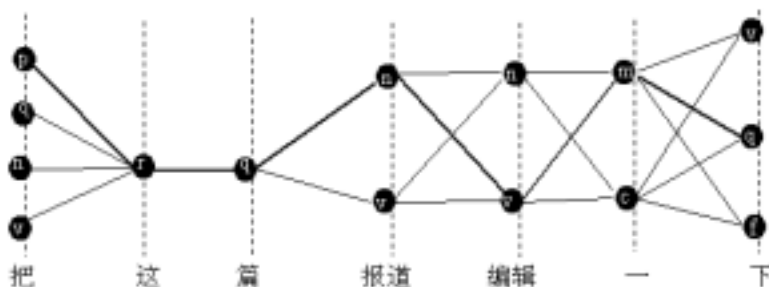
@@ 信(n-v)

```
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



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

## 基于HMM进行词性标注 1

- 把词汇序列（记做  $W=w_1w_2\ldots w_n$ ）理解为观察值
- 把词性标注序列（记做  $T=t_1t_2\ldots t_n$ ）理解为隐含的状态值
- 词性标注问题变成HMM中的解码问题
- 已知词串  $W$ （观察序列）和模型  $\lambda$  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 2

根据Bayes公式：

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

由于 $P(W|\lambda)$ 是常数：

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

## 基于HMM进行词性标注 3

公式3可以进一步分解为：

$$\arg \max_T P(T | W) = \arg \max_T P(T)P(W | T) \quad \text{公式4}$$

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HMM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 4

$P(W|T)$ 是已知词性标记串，产生词串的条件概率：

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

根据HMM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：

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

## 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

$$P(T1|W) = P(q|\$)P(\text{把}|q)P(r|q)P(\text{这}|r) \dots P(f|m)P(\text{下}|f)$$

$$P(T2|W) = P(q|\$)P(\text{把}|q)P(r|q)P(\text{这}|r) \dots P(q|m)P(\text{下}|q)$$

$$P(T3|W) = P(q|\$)P(\text{把}|q)P(r|q)P(\text{这}|r) \dots P(v|m)P(\text{下}|v)$$

.....

$$P(T96|W) = P(n|\$)P(\text{把}|n)P(r|q)P(\text{这}|r) \dots P(v|c)P(\text{下}|v)$$

从中选  
一个最  
大值

词性转移概率

词语输出概率

# 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) 让机器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。

## 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( $n$ )；

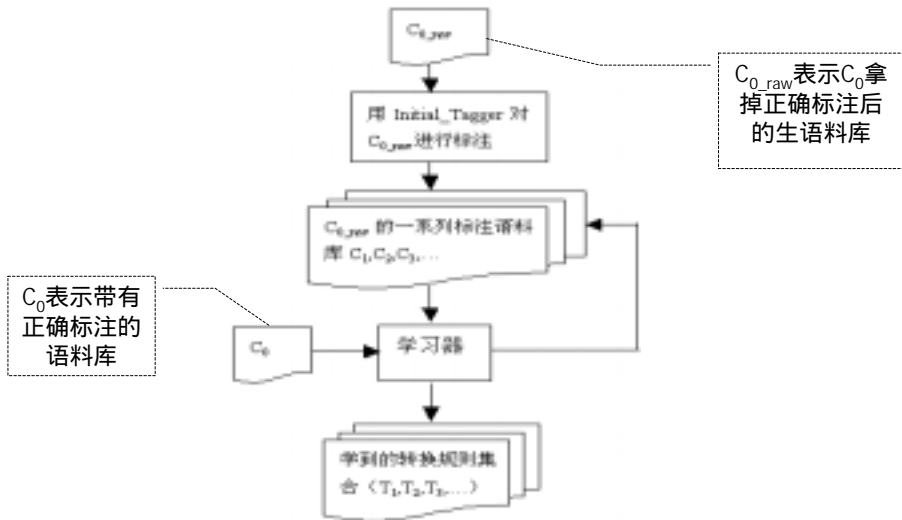
$T_2$ ：当前词的后一个词的词性标记是动词( $v$ )时，将当前词的词性标记由动词( $v$ )改为名词( $n$ )；

$T_3$ ：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, 并产生一个可应用的候选规则集合;
- 3) 依次从候选规则中取出一条规则  $T_m (m=1, 2, \dots)$ , 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  $C_i^m (m=1, 2, 3, \dots)$ , 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  $C_i^j$  (且  $C_i^j$  中的错误数少于  $C_i$  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  $C_2^k, C_3^l, C_4^m, \dots$  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  $\{T_a, T_b, T_c, \dots\}$

# 完整的汉语词法分析系统 1

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

# 完整的汉语词法分析系统 2

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

# 词法分析集成问题一

- 未定义词识别与切分的顺序
  - 先切分后识别再切分
  - 先识别后切分
  - 同时进行

## 先识别未定义词，后切分

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

## 先切分，后识别未定义词

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

## 基于N最短路径的粗切分 1

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

## 基于N最短路径的粗切分 2

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

## 基于N最短路径的粗切分 3

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



## 基于N最短路径的粗切分 4

Length No.	粗分召回率 ( % )
1	92.88
2	98.55
3	99.33
4	99.67
6	99.80
7	99.86
8	99.89
9	99.91

## 切分与未定义词识别同时进行

- 需要建立切分与未定义词识别的统一概率模型
- 搜索空间较大，时间复杂度高

## 词法分析集成问题二

- 切分与标注的顺序
  - 先切分后标注
  - 同时进行

## 先切分，后标注

- 切分只产生一个结果：不可取
  - 切分错误将无法更改
- 切分保留N-Best结果：可取
  - 切分错误有可能在标注阶段得以恢复

## 切分与标注同时进行

- 需要建立切分与标注的统一概率模型
- 搜索空间较大，时间复杂度高

## 词法分析集成问题三

- 建立统一的概率模型
  - 切分过程中变形词的概率计算
  - 切分与未定义词识别的统一概率模型
  - 切分与标注的统一概率模型

# 切分过程中变形词的概率计算

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

# 切分过程中未定义词的概率计算 1

- 基于N元语法的未定义词概率计算

$$\begin{aligned}P(\text{李素丽} \mid \text{人名}) &= P(\text{李} \mid \text{人名首字}) \\ &\quad \times P(\text{素} \mid \text{李}, \text{人名}) \\ &\quad \times P(\text{丽} \mid \text{素}, \text{人名})\end{aligned}$$

## 切分过程中未定义词的概率计算 2

- 基于概率上下文无关语法的概率计算

单名规则：人名 = 姓 + 名

双名规则：人名 = 姓 + 名<sub>1</sub> + 名<sub>2</sub>

前缀规则：人名 = 前缀 + 姓

后缀规则：人名 = 姓 + 后缀

..... ???

$$P(\text{李素丽} \mid \text{人名}) = P(\text{双名} \mid \text{人名}) \times P(\text{李} \mid \text{姓}) \\ \times P(\text{素} \mid \text{名}_1) \times P(\text{丽} \mid \text{名}_2)$$

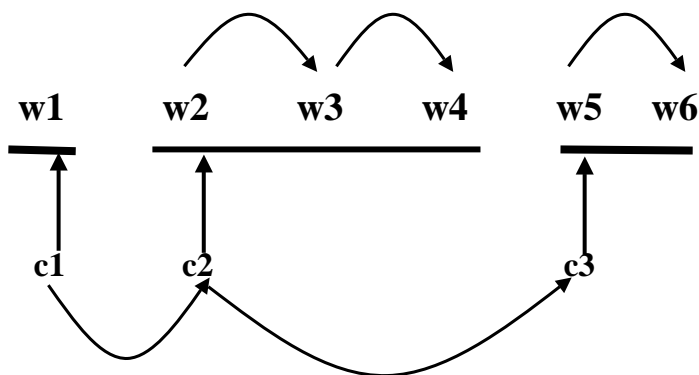
## 切分与未定义词识别的 统一概率模型

- 孙健，基于统计方法的短语识别和句法结构歧义消解的研究，北京邮电大学博士论文，2002  
“基于类的语言模型”  
“Class-Based Language Model”

# 基于类的语言模型 1

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

# 基于类的语言模型 2



## 基于类的语言模型 3

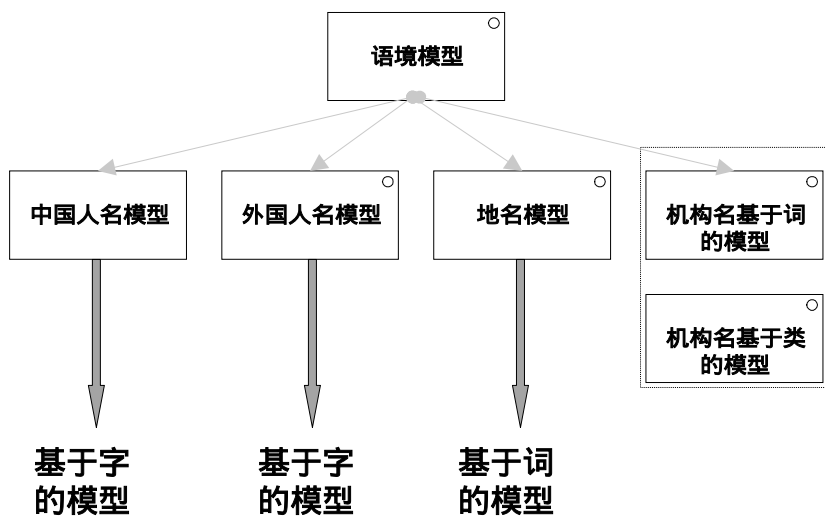
--- class model

C : 类序列

--- entity model

T : 汉字序列

## 基于类的语言模型 4



## 基于类的语言模型 5 语料

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

## 基于类的语言模型 6 结果

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



## 基于类的语言模型 7 总结

- 优点
  - 中文分词和命名实体识别结合在一起
  - 语境信息和实体内部信息有机结合在一起
  - 人名、地名和机构名这三类不同的命名实体识别纳入到统一的模型框架中
  - 启发式信息有机融入到语言模型
  - 能够识别嵌套的命名实体
- 缺点
  - 两层Viterbi搜索，搜索空间大，时间复杂度高

## 切分与标注的统一概率模型

高山 等，基于三元统计模型的汉语分词标注一体化研究，全国第五届计算语言学联合学术会议（JSCL-2001）

- 切分和标注混合的统计模型
- 切分模型采用三元语法
- 标注模型采用三元HMM
- 取切分模型和标注模型的加权平均

# 切分标注的混合统计模型 1

切分模型

$$P(W) = \prod_{i=1}^n p(w_i | w_{i-1}w_{i-2})$$

标注模型

$$P(T, W) = P(T) \times P(W | T) = \prod_{i=1}^n p(w_i | t_i) p(t_i | t_{i-1}t_{i-2})$$

混合模型

$$P^*(T, W) = \prod_{i=1}^n p(w_i | t_i) p(t_i | t_{i-1}t_{i-2}) + \lambda \prod_{i=1}^n p(w_i | w_{i-1}w_{i-2})$$

取 $\lambda$  = 词典中词  $w$  的个数 / 词性  $t$  的种类

# 切分标注的混合统计模型 2

- 实验结果

文本长度 (词数)	分词 正确率	二级标注 正确率	一级标注 正确率
719	99.1%	93.2%	96.8%
4644	98.3%	92.5%	96.1%
5627	97.9%	93.3%	96.5%
13166	98.0%	93.1%	96.3%

- 分析

- 结果尚可，不过实验规模偏小
- 是一个经验公式，理论上说服力不强
- 没有处理未定义词

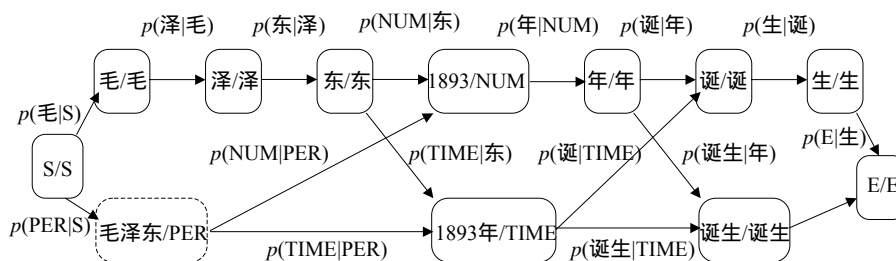
# 汉语词法分析系统ICTCLAS 1

- 基于N-Best策略的多层次扫描算法
  - 查词典与词形变化处理
  - 粗切分：基于N-最短路径的粗切分
  - 简单未定义词识别：基于角色标注的识别算法
  - 嵌套未定义词识别：基于角色标注的识别算法
  - 细切分：基于类的二元语法的切分算法
  - 词性标注：基于HMM的词性标注算法

# 汉语词法分析系统ICTCLAS 2

- 细节
  - 精心定制的词性标记集ICTPOS3.0：99个标记
  - 每一阶段都采用N-Best策略，最大限度减少阶段性错误的传递
  - 细切分中未定义词的概率计算：基于概率上下文无关语法
  - 基于角色标注的未定义词识别：精心调试的角色集合

# 汉语词法分析系统ICTCLAS 3



基于类的二元语法切分词图（原始字串为“毛泽东1893年诞生”）

说明：

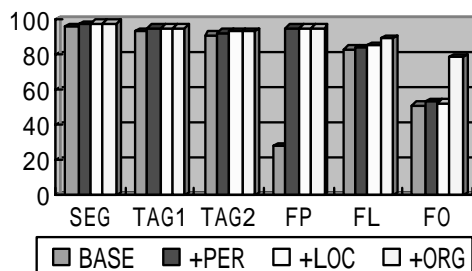
1. 节点中表示的是“词语/类”（即  $w_i / c_i$ ），节点的权值为类到词语的概率  $p^*(w_i | c_i)$ ；
2. 有向边的权值为相邻类的转移概率  $p^*(c_i | c_{i-1})$ ；S为初始节点；E为结束节点；
3. “毛泽东/PER”相关的虚线部分是人名识别HMM作用过之后产生的。

# 汉语词法分析系统ICTCLAS 4

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

- 1) 数据来源：国家973英汉机器翻译第二阶段评测的评测总结报告；
- 2) 标注相对正确率  $RTAG = TAG1 / SEG * 100\%$
- 3) 由于我们采取的词性标注集和973专家组的标注集有较大出入，所以词性标注的正确率不具可比性。

# 汉语词法分析系统ICTCLAS 5



FP：人名识别F-Score

FL：地名识别F-Score

FO：机构名识别F-Score

可以看到，随着人名识别、地名识别和机构名识别的加入，总体性能和各单项性能都稳步提高

## 复习思考题

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