



第二讲 中文文本的自动分词

詹卫东

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



提纲

1. 从字符串到词串
2. 文本分词面对的问题
3. 文本分词的基本方法
4. 对文本分词质量的评价
5. 小结



1 从字符串到词串

学生人数多又能保证质量的才是好学校。

字符串

学 生 人 数 多 又 能 保 证 质 量 的 才 是 好 学 校 。

词串

Schools which can both enroll more students and ensure the quality of instruction are the successful ones.

关于中文分词的形式化定义，可参看马晏，1991，基于评价的汉语自动分词系统的研究与实现，载黄昌宁、夏莹编《语言信息处理专论》，清华大学出版社1996年版



从字符串到词串（续）

1. Dog's Let's
2. ad hoc and so on New York
3. strong stronger strongest
4. buy bought buying
5. eat ate eaten
6. try tried tries
7. treat treatment
8. news newspaper



从字符串到词串（续）

French: Paul ouvre le sac de pommes de terre et le pose sur la table.

English: Paul opens the bag of potatoes and puts it on the table

pommes de terre



apple of earth

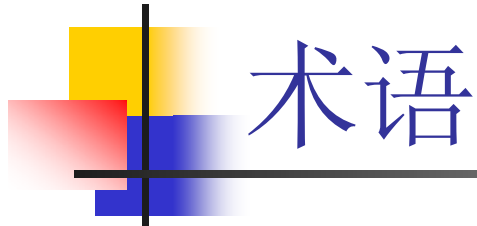


potato

le sac → the

le pose → it

Anne Abeillé, ed., 2003, Treebanks: Building and Using Parsed Corpora, Kluwer Academic Publishers. (Text, Speech and Language Technology Volume 20) Introduction



- Word Segmentation
- Word Tokenization
- Word Stemming/Lemmatization

Tokenization: 把字符串变为词串

I'm a student -> I 'm a student

Lemmatization: 对词进行内部结构和形式分析

took -> take + ed (past tense)



将字符串转换到词串的意义

文本分词是各个层次的自然语言处理任务的基础

1. 简繁转换 Automatic Conversion between traditional and simplified Chinese
2. 文语转换 Text-to-speech
3. 文本校对 Chinese Text Correction
4. 文本检索 Information retrieval
5. 词频统计、句法分析、机器翻译、



从字符串到词串的意义（续）

简繁转换 示例

- (1) 明成皇后，她是一个世纪前北韩王朝的最后_后一位皇_后。
明成皇后，她是一個世紀前北韓王朝的最後_後一位皇_后。
- (2) 负离子陶瓷烫_发机，内置负离子_发射器。
負離子陶瓷燙_髮機，內置負離子_發射器。
- (3) 每个战_斗单位只有一_斗米
每個戰_鬥單位只有一_斗米



从字符串到词串的意义（续）

文语转换 示例

(1) 为达到赢球的目的，一定要注意比赛时的情绪调动与心理调节
· · ·

(2) 他们村有三百多人种树
·

(3) 他喜欢展览馆门口播放的“小老鼠上灯台”这类欢快的儿歌
· · · · · ·



2 文本分词面对的问题

1. 分词歧义
2. 未登录词识别
3. 什么是中文的“词”



2.1 文本分词中的歧义

1. 张店区大学生不看重大城市的户口本

张店区 大学生 不 看 重大 城市 的 户口本

张店区 大学生 不 看重 大 城市 的 户口本

交集型
歧义

2. 你认为学生会听老师的吗

你 认为 学生会 听 老师 的 吗

你 认为 学生 会 听 老师 的 吗

组合型
歧义

3. 只有雷人才能吸引人

只有雷人 才能 吸引 人

只有雷人才 能吸引 人

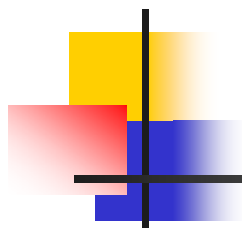
只有雷人 才 能 吸引 人

混合型
歧义



交集型歧义的链长

- 交集型歧义字段中含有交集字段的个数，称为链长。
 - 链长为1： 和尚未
 - 链长为2： 结合成分
 - 链长为3： 为人民工作
 - 链长为4： 中国产品质量
 - 链长为5： 鞭炮声响彻夜空
 - 链长为6： 努力学习语法规则
 - 链长为7： 中国企业主要求解决
 - 链长为8： 治理解放大道路面积水
 -



真实文本中分词歧义的分布情况

交集型歧义：组合型歧义 = 1: 22 语料规模：17,547字 [1]

语料规模：500万字新闻语料 [2]

链长 歧义 字段	1	2	3	4	5	6	7	8	总计
Token次数	47402	28790	1217	608	29	19	2	1	78248
比例%	50.58	47.02	1.56	0.78	0.04	0.02	0.00	0.00	100
Type种数	12686	10131	743	324	22	5	2	1	23914
比例%	53.05	42.36	3.11	1.35	0.09	0.02	0.01	0.01	100

[1] 刘挺、王开铸，1998，关于歧义字段切分的思考与实验。《中文信息学报》第2期，63-64页。

[2] 刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，65页。



汉语真实文本中的分词歧义情况（续）

- 真歧义

确实能在真实语料中发现多种切分形式

比如“应用于”、“地面积”、“解除了”

- 伪歧义

虽然有多种切分可能性，但在真实语料中往往取其中一种切分形式

比如“挨批评”、“市政府”、“太平淡”



汉语真实文本中的分词歧义情况（续）

78248个^[1]
交集型歧义字段

{	伪歧义：94%	{	多种切分均匀分布	12%	甲
	真歧义：6%		一种切分切分占优	88%	乙

甲 将信息技术/应用/于/教学实践
信息技术/应/用于/教学中的哪个方面

乙 上级/解除/了/他的职务
方程的/解/除了/零以外还有...

[1] 刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，66-67页。



汉语真实文本中的分词歧义情况（续）

在一个1亿字真实汉语语料库中抽取出的前4,619个高频交集型歧义切分覆盖了该语料库中全部交集型歧义切分的59.20%，其中4279个属伪歧义，占92.63%，如“和软件”、“充分发挥”、“情不自禁地”，这部分伪歧义类型的实例对语料的覆盖率高达53.35%。^[1]

[1] 孙茂松 等，1999，《高频最大交集型歧义切分字段在汉语自动分词中的作用》，载《中文信息学报》1999年第1期。



汉语真实文本中的分词歧义情况（续）

分词歧义的四个层级（语料规模：**50883**字）^[1]

- 词法歧义：84.1% （“用方块图形式加以描述”）
- 句法歧义：10.8% （“他一阵风似的跑了”）
- 语义歧义：3.4% （“学生会写文章”）
- 语用歧义：1.7% （“美国会采取措施制裁伊拉克”）

[1] 何克抗 等，1991，《书面汉语自动分词专家系统设计原理》，载《中文信息学报》，1991年第2期。



2.2 未登录词 (OOV)

- | | |
|----------------|-------------|
| 1. 汉族人名、地名 | 雪村 老张 中关村 |
| 2. 外族人名、地名 | 横路静二 突尼斯 |
| 3. 中外组织机构单位名称 | 联合国教科文组织 |
| 4. 商品品牌名 | 非常可乐 苹果iPad |
| 5. 专业术语 | 有限状态自动机 三分球 |
| 6. 新词语 | 秒杀 蚁族 羊羔体 |
| 7. 缩略语 | 人影办 两会 北医三院 |
| 8. 汉语重叠形式、离合词等 | 高高兴兴 幽了他一默 |



2.3 什么是“词”

- 语法学定义：能够独立运用的最小的音义结合体
分词规范：结合紧密，使用稳定
- 词表定义：枚举“词型”（type）
- 语料库定义：枚举“词例”（token）

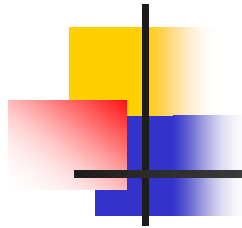


分词规范

刘源 等（1994）《信息处理用现代汉语分词规范及自动分词方法》，清华大学出版社、广西科学技术出版社，1994年版。

黄居仁、陈克健 等（1997）《信息处理用中文分词规范设计理念及规范内容》，载《语言文字应用》1997年第1期。

- 《信息处理用汉语分词规范》 GB/T13715-92，中国标准出版社，1993
- 《资讯处理用中文分词规范》 台湾中研院，1995
- 《人民日报》语料库词语切分规范 北大计算语言所，1999



不同的人对“词”的认识有差异

6人对100句（4372）字进行人工分词，然后两两比较认同率

	M2	M3	T1	T2	T3
M1	0.77	0.69	0.71	0.69	0.70
M2		0.72	0.73	0.71	0.70
M3			0.89	0.87	0.80
T1				0.88	0.82
T2					0.78

平均值
0.76

Sproat R. et al., 1996, A Stochastic Finite-state Word Segmentation Algorithm for Chinese.
Computational Linguistics, Vol.22, No.3, pp377-404.



待分词文本示例

不管你相不相信，美丽岛的贫困县人口中有五分之三已经超过半年领不到救济粮，其中散居全县各区的汉族人占到绝大多数。更糟糕的是，省气象部门预测即将到来的雨季降水量之大会让近半数的人无家可归。政府为此召开了紧急动员会。没想到的是，会上一些大型国有控股企业和东南沿海民营企业的代表纷纷表示会施以援手。正在当地集训的某大洋洲国家羽毛球队参加了此次还算说得过去的动员会。（未名新闻社胡编供稿）



3 文本分词的基本方法

基于词典的分词方法

最大匹配法
最短路径法
最大概率法
.....

“分”词

基于字序列标注的方法

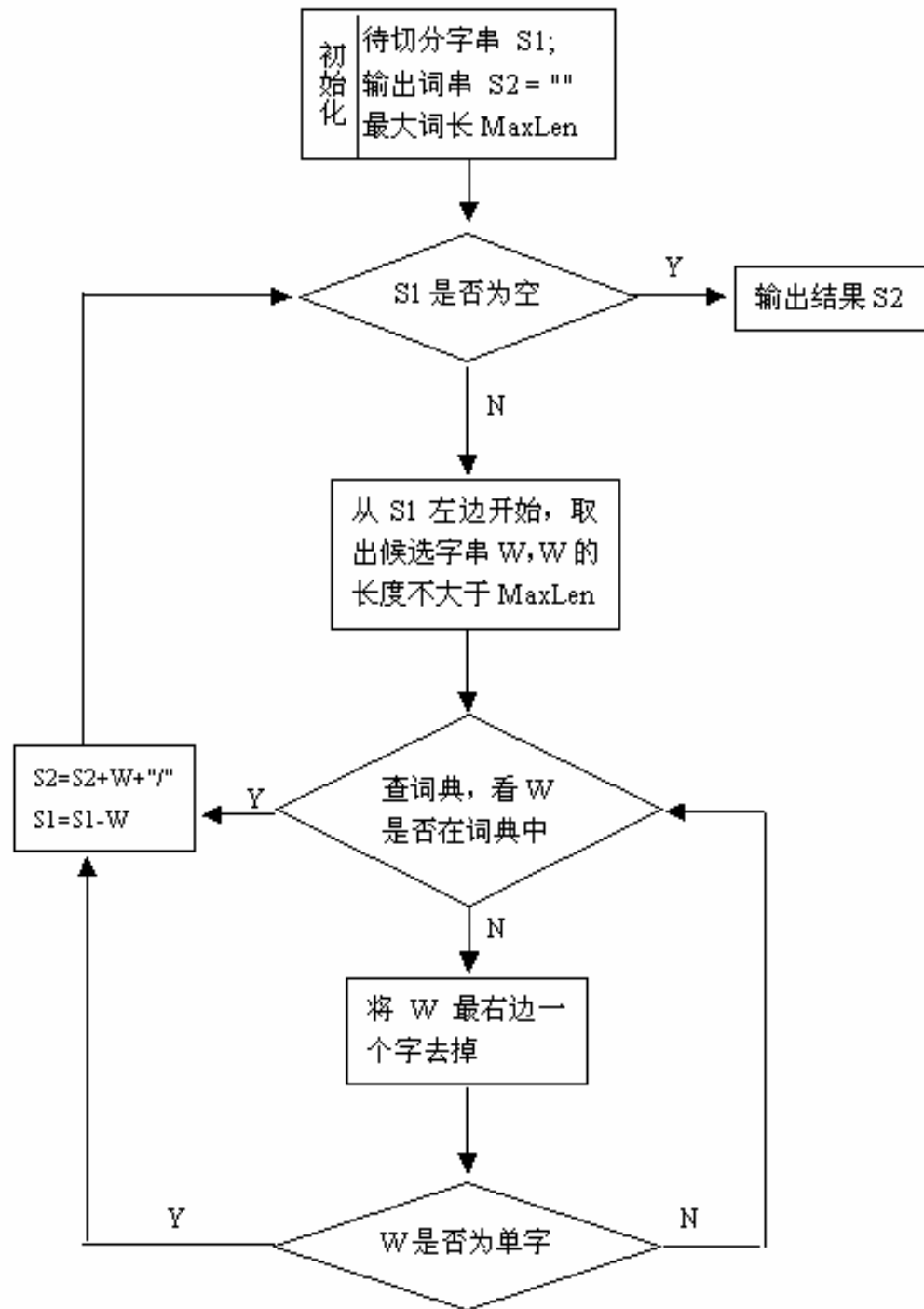
“合”词

最大匹配法

FMM（正向最大匹配）

BMM（逆向最大匹配）

长词优先原则





最大匹配法分词示例

输入：S1="计算语言学课程是三个课时"

输出：S2= " "

设定最大词长MaxLen = 5

W= 计算语言学

.....

词语
...
计算语言学
课程
课时
...

大规模真实语料中99%的词例（token）的长度都在5字以内^[1]

[1] 黄昌宁、赵海，2007，中文分词十年回顾，《中文信息学报》2007年第3期，8-19页。



最大匹配法的问题

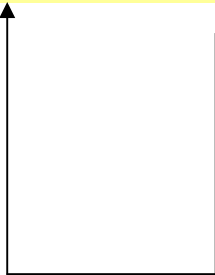
- 存在分词错误 → 增加知识，局部修改
- 无法发现分词歧义 → 从单向最大匹配改为双向最大匹配
 - A. 正向最大匹配和逆向最大匹配结果不同
 - FMM 有意/ 见/ 分歧/
 - BMM 有/ 意见/ 分歧/
 - B. 正向最大匹配和逆向最大匹配结果相同
 - FMM & BMM 原子/ 结合/ 成分/ 子时/



局部修改1：增加歧义词表，排歧规则

规则示例

IF $W = \text{"个人"}, W_{\text{Left}} = \text{数词}$ THEN $W = \text{"个/ 人/"}$ ENDIF



歧义词表
...
才能
个人
家人
马上
研究所
...



局部修改2：增加“回溯”

对于某些交集型歧义，可以通过增加回溯机制来修改最大匹配法的分词结果。

例如：“爱人民英雄”

顺向扫描的结果是：“爱人/ 民/ 英雄/”，

通过查词典知道“民”不在词典中，于是进行回溯，将“爱人”的尾字“人”取出与后面的“民”组成“人民”，再查词典，看“爱”，“人民”是否在词典中，如果在，就将分词结果调整为：“爱/ 人民/ 英雄/”

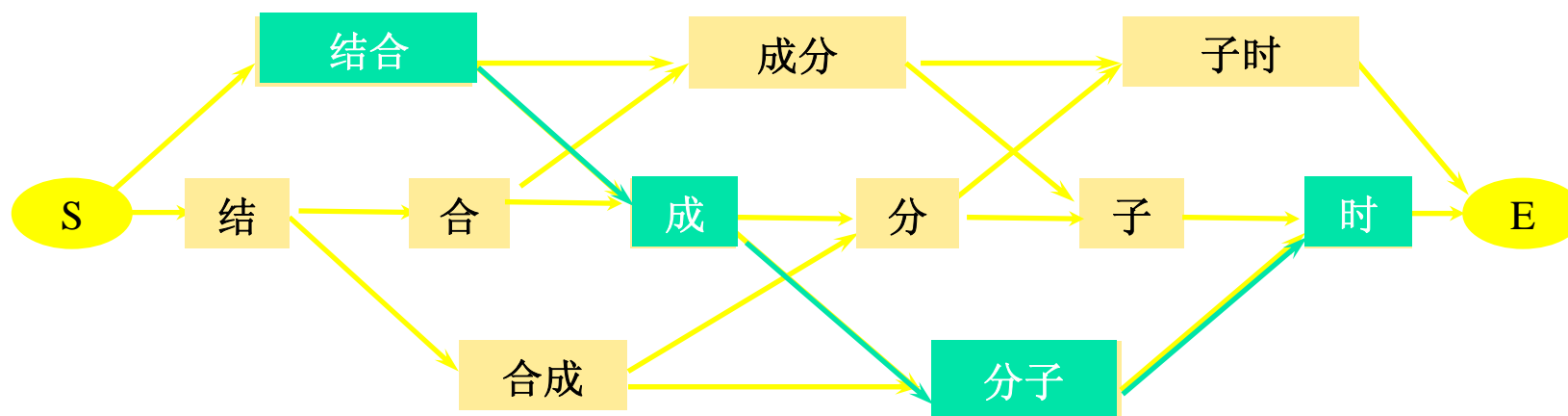


最大匹配法分词的问题（续）

- 双向最大匹配法可以发现链长为奇数的交集型歧义，但无法发现链长为偶数的交集型歧义
- 无法发现组合型歧义
- 在最大匹配法的基础上进行修改，如何给出“改错”的触发条件带有一定的主观性

需要更全面地考虑分词的改进办法

汉语词语切分的数据结构—词图



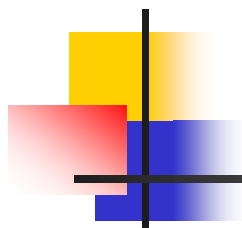
词图给出了一个字符串的全部切分可能性

分词任务：寻找一条起点S到终点E的最优路径



最短路径分词法

- 基本思想：在词图上选择一条词数最少的路径
- 优点：好于单向的最大匹配方法
 - 最大匹配：独立自主 和平 等 互利 的 原则 (6 words)
 - 最短路径：独立自主 和 平等互利 的 原则 (5 words)
- 缺点：同样无法解决大部分交集型歧义
 - 结合 成分 子时
 - 他 说 的 确 实 在 理 (都是最短路径)
他 说 的 确 实 在 理
他 说 的 确 实 在 理



半词法分词

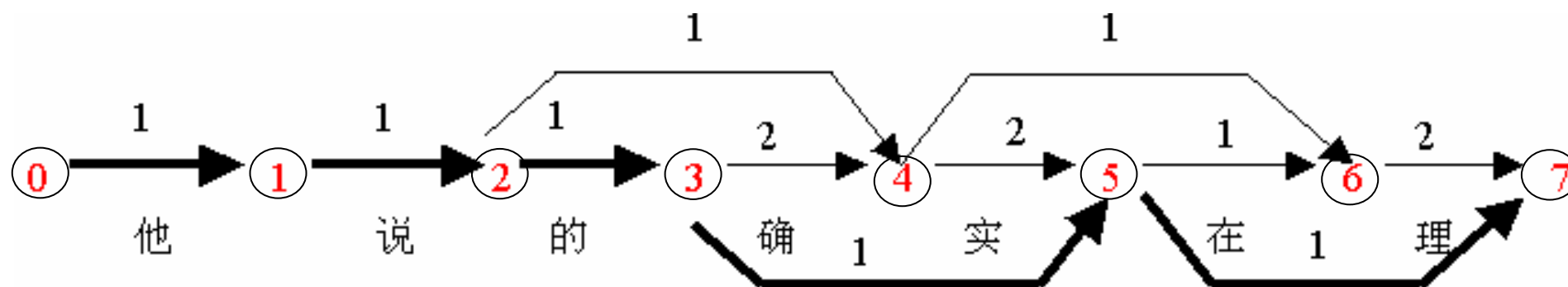
基本观察	大多数单字在语境里如果能组成合适的词就不倾向于单独使用。	
基本概念	半词	如果一个字不单独作为词使用，就是半词。半词既包含了成词语素，也包含了不成词语素，后者肯定是半词，比如“民”，前者则要看它作为语素的使用频度高，还是作为单字词的使用频度高，比如“见”。
	整词	如果一个字更倾向于自己成词而不倾向于和别的字组成词，这类“单字词”就称之为“整词”。这类词就是一般说的单字高频成词语素，比如“人、说、我”等。
基本思路	充分利用半词和整词的差别，尽量选择没有半词落单的分词方案。	



半词法分词的实现

- 在词图的路径优劣评判中引入罚分机制
- 罚分规则：
 - 1 每个词对应的边罚1分。
 - 2 每个半词对应的边加罚1分。
 - 3 一个分词方案的评分为它所对应的路径上所有边的罚分之和。
 - 4 最优路径就是罚分最低的分词路径。

半词法分词示例



他 说 的 确 实 在 理 (1+1+1+1+1 = 5分)

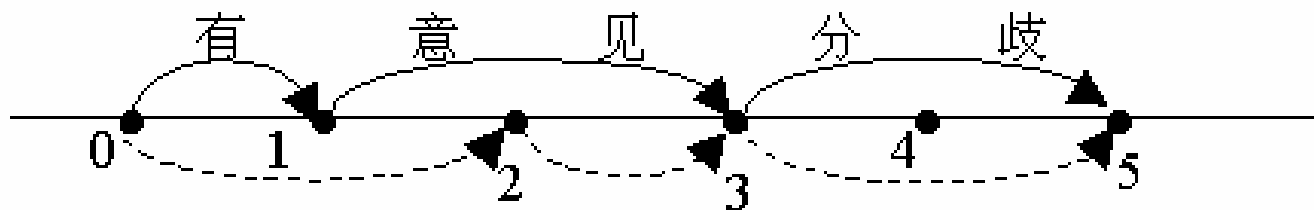
他 说 的 确 实 在 理 (1+1+1+2+1 = 6分)

他 说 的 确 实 在 理 (1+1+1+1+2 = 6分)

但是：仍然无法解决“有意见分歧”的问题

最大概率法分词

基本思想：在词图上选择概率最大的分词路经作为最优结果



路径1: 0—1—3—5

路径2: 0—2—3—5

该走哪条路呢？



最大概率法分词

输入 字符串S: 有意见分歧

输出 词串 W1: 有/ 意见/ 分歧/
词串W2: 有意/ 见/ 分歧/

$Max(P(W1/S), P(W2/S))$?

$$P(W | S) = \frac{P(S | W) \times P(W)}{P(S)} \approx P(W)$$

独立性假设, 一元语法

$$P(W) = P(w_1, w_2, \dots, w_i) \approx P(w_1) \times P(w_2) \times \dots \times P(w_i)$$

$$P(w_i) = \frac{w_i \text{在语料库中的出现次数} n}{\text{语料库中的总词数} N} = \frac{Freq(w_i)}{N}$$



最大概率法分词示例

词语	概率
...	...
有	0.0180
有意	0.0005
意见	0.0010
见	0.0002
分歧	0.0001
...	...

$$P(W1) = P(\text{有}) * P(\text{意见}) * P(\text{分歧})$$

$$= 1.8 \times 10^{-9}$$

$$P(W2) = P(\text{有意}) * P(\text{见}) * P(\text{分歧})$$

$$= 1.0 \times 10^{-11}$$

$$P(W1) > P(W2)$$



最大概率法分词的实现

—— 如何高效求解最优路径？

- 动态规划算法：最优路径中的第 i 个词 w_i 的累积概率等于它的左邻词 w_{i-1} 的累积概率乘以 w_i 自身的概率。

$$P'(w_i) = P'(w_{i-1}) \times P(w_i)$$

- 为方便计算，一般把概率转化为路径费用（代价）

$$C = -\log(P)$$

$$C'(w_i) = C'(w_{i-1}) + C(w_i) \quad \text{公式1}$$

↓ ↓
最小累积费用 最佳左邻词



最大概率分词法的实现：动态规划算法

- 1) 对于一个待分词的字串 S ，按照从左到右的顺序取出全部候选词 $w_1, w_2, \dots, w_i, \dots, w_n$ ；
- 2) 到词典中查出每个候选词的概率值 $P(w_i)$ ，转换为费用 $C(w_i)$ ，并记录每个候选词的全部左邻词；
- 3) 按照公式1计算每个候选词的累计费用，同时比较得到每个候选词的最佳左邻词；
- 4) 如果当前词 w_n 是字串 S 的尾词，且累计费用 $C'(w_n)$ 最大，则 w_n 就是 S 的终点词；
- 5) 从 w_n 开始，按照从右到左顺序，依次将每个词的最佳左邻词输出，即为 S 的分词结果。

最大概率法分词示例（续）

序号	候选词	费用	累计费用	最佳左邻
0	结	3.573	3.573	-1
1	结合	3.543	3.543	-1
2	合	3.518	7.091	0
3	合成	4.194	7.767	0
4	成	2.800	6.343	1
5	成分	3.908	7.451	1
6	分	2.862	9.205	4
7	分子	3.465	9.808	4
8	子	3.304	10.755	5
9	子时	6.000	13.451	5
10	时	2.478	12.286	7



最大概率法分词的问题

- 并不能解决所有的交集型歧义问题

“这事的确定不下来”

W1= 这/ 事/ 的确/ 定/ 不/ 下来/ $P(W1) < P(W2)$

W2= 这/ 事/ 的/ 确定/ 不/ 下来/

- 一般也无法解决组合型歧义问题

“做完作业才能看电视”

W1= 做/ 完/ 作业/ 才能/ 看/ 电视/ $P(W1) > P(W2)$

W2= 做/ 完/ 作业/ 才/ 能/ 看/ 电视/



识别未登录词的策略

- 1) 尽可能多地收集词汇，以降低碰到未登录词的机会；
- 2) 通过构词规则和上下文特征规则来识别；
“雪村先生创作了很多歌曲”
- 3) 通过统计的方法来猜测经过一般的分词过程后剩下的“连续单字词碎片”是人名、地名等的可能性，从而识别出未登录词。
- 4) 分而治之：对不同类的未登录词采用不同的办法识别



不同类别未登录词识别难度的差异

- 较成熟
 - 中国人名、译名
 - 中国地名
- 较困难
 - 商标字号
 - 机构名
- 很困难
 - 专业术语
 - 缩略语
 - 新词语



中国人名的内部构成规律 1

- 在汉语的未定义词中，中国人名是规律性最强，也是最容易识别的一类；
- 中国人名一般由以下部分组合而成：
 - 姓：张、王、李、刘、诸葛、西门、范徐丽泰
 - 名：李素丽，张华平，王杰、诸葛亮
 - 前缀：老王，小李
 - 后缀：王老，赵总
- 中国人名各组成部分用字比较有规律



中国人名的内部构成规律 2

- 根据统计，汉语姓氏大约有1000多个，
- 姓氏中使用频度最高的是“王”姓；
- “王, 陈, 李, 张, 刘”等5个大姓覆盖率达32%；
- 姓氏频度表中的前14个高频度的姓氏覆盖率为50%；
- 前400个姓氏覆盖率达99%。
- 人名的用字也比较集中。
- 频度最高的前6个字覆盖率达10.35%；
- 前10个字的覆盖率达14.936%；
- 前15个字的覆盖率达19.695%；
- 前400个字的覆盖率达90%。



中国人名的内部构成规律 3

- 中国人名内部各组成部分的组合规律

- 姓+名

- 姓

- 名

- 前缀+姓

- 姓+后缀

- 姓+姓+名（海外已婚妇女）

姓、名均可再分
“单字”“双字”



中国人名的上下文构成规律

中国人名外部特征：

- 身份词：
 - 前：工人、教师、影星、犯人
 - 后：先生、同志
 - 前后：校长、经理、主任、医生
- 地名或机构名：前：静海县大丘庄禹作敏
- 的字结构前：年过七旬的王贵芝
- 动作词
 - 前：批评，逮捕，选举
 - 后：说，表示，吃，结婚
-



中国人名识别的难点

- 一些高频姓名用字在非姓名中也是高频字
 - 姓氏：于，马，黄，张，向，常，高
 - 名字：周鹏和同学，周鹏和同学
- 人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词
 - [王国]维、[高峰]、[汪洋]、张[朝阻]、冯[胜利]
- 人名与其上下文组合成词
 - 这里[有关]天培的壮烈
 - 费孝[通向]人大常委会提交书面报告
 - 邓颖[超生]前使用过的物品
- 人名地名冲突
 - 河北省刘庄



中国人名识别的方法

- 不考虑上下文信息的识别方法：根据一个汉字串内部各汉字的特征计算该汉字串作为中文姓名的概率 [1]
- 考虑上下文信息的识别方法：把姓名及其上下文中的汉字标记不同的“角色”，将人名识别问题转换为汉字的角色标注问题 [2]

[1] Sproat R. et al., 1996, A Stochastic Finite-state Word Segmentation Algorithm for Chinese, *Computational Linguistics*, Vol.22, No.3, pp377-404.

[2] 张华平、刘群，2004，基于角色标注的中国人名自动识别研究，《计算机学报》Vol.27, No.1。



中文姓名识别模型示例

- 中文姓名的组合模型

r1: word -> name

r2: name -> 1-hanzifamily 2-hanzigiven

r3: name -> 1-hanzifamily 1-hanzigiven

r4: name -> 2-hanzifamily 2-hanzigiven

r5: name -> 2-hanzifamily 1-hanzigiven

r6: 1-hanzifamily -> hanzi_i

r7: 2-hanzifamily -> hanzi_i hanzi_j

r8: 1-hanzigiven -> hanzi_i

r9: 2-hanzigiven -> hanzi_i hanzi_j

- 例如：对于一个3字串 $C_0C_1C_2$ ，它构成中文姓名的概率为：

$$P(C_0C_1C_2) = P(r1) \cdot P(r2) \cdot P(r6) \cdot P(r9)$$

$$P(r1) = P(name | word)$$

$$P(r2) = P(1-hanzifamily, 2-hanzigiven | name)$$

$$P(r6) = P(C_0 | 1-hanzifamily)$$

$$P(r9) = P(C_1C_2 | 2-hanzigiven)$$

若 $P(C_0C_1C_2)$ 大于设定的阈值，则判为人名

Sproat R. et al., 1996, A Stochastic Finite-state Word Segmentation Algorithm for Chinese, *Computational Linguistics*, Vol.22, No.3, pp377-404.



音译名的识别 1

- 音译名用字非常集中 《英语姓名译名手册》中共收英语姓氏, 教名约4万个, 经计算机统计得出英语姓名译名用字表共476个:

“啊阿埃艾爱昂奥巴白柏拜班邦包保堡鲍北贝倍本比彼边别滨宾玻波博勃伯卜布采蔡藏策查察昌彻陈楚垂茨慈次聪存措达大戴代丹当道德得登邓迪底地蒂第帝丁东杜敦顿多厄恩耳尔法凡范方菲费芬丰冯佛夫福弗辅富盖甘冈高哥戈葛格各根贡古顾瓜圭郭果哈海罕翰汉杭豪赫黑亨洪侯胡华怀惠霍基吉季计嘉佳加贾简姜焦杰捷金津京久居喀卡开凯坎康考柯科可克肯孔扣寇库夸匡奎魁坤昆阔拉腊莱来赖兰朗劳勒乐雷黎理李里礼荔丽历利立莲连廉良列琳林霖龄留刘流柳龙隆卢鲁露路吕略伦萝罗洛玛马麦迈满曼芒茅梅门蒙孟米密敏明名摩莫墨默姆木穆拿娜纳乃奈南内嫩能妮尼年涅宁牛纽农努女诺欧帕派潘庞培佩彭蓬皮匹平泼朴普漆奇齐契恰钱强乔切钦琴青琼丘邱屈让热仁日荣茹儒瑞若撒萨塞赛三缮桑瑟森莎沙珊山尚绍舍申生盛圣施诗石什史士寿舒朔斯思丝松孙索所塔泰坦汤唐陶特藤提惕田铁汀廷亨通透图托脱娃瓦万旺威韦为维伟魏卫温文翁沃乌武伍西锡希悉席霞夏显香向晓肖歇谢欣辛兴幸姓雄休修雪逊雅亚延扬阳尧耀耶叶依易意因英永尤雨约宰赞早泽曾扎詹湛章张哲者珍真芝知智治朱卓兹子宗祖佐丕谟葆薇岑弼娅缪珀璫赉滕斐熙鳩窠艮麟黛”。

辛华编《英语姓名译名手册》商务印书馆1973年（修订版）

新华通讯社译名资料组编《英语姓名译名手册》商务印书馆1997年（第二次修订版）



音译名的识别 2

- 音译名内部很难划分出结构，但有一些常见音节，如“斯基、斯坦”等
- 不同语言的音译规律不尽相同，如法语、俄语、蒙古语译名用字与英语就有较大区别（蒙古人名举例：“那顺乌日图、青格勒图”），如果按不同的语言训练不同的模型可能会比使用统一的模型效果更好
- 音译名可以是人名、地名或其他专名，上下文规律差别较大
- 由于音译名用字比较集中，识别正确率较高



中国地名的识别

- 中国地名委员会编写了《中华人民共和国地名录》，收集了全国乡镇以上（含乡镇）各级行政区域的名称，以乡镇人民政府所在地为主的居民聚落名称，山、河、湖、海、岛、高原、盆地、沙溪等自然地理实体名称，名胜古迹、纪念地、古遗址、水库、桥梁、电站等名称。共收录地名10万多条。这个地名录中使用的汉字共2662个，频度最高的前65个汉字占总频度的50.22%，前622个汉字占总频度的90.01%，前1872个汉字占总频度的99%。
- 与人名的用字情况相比较，地名用字分散得多
- 地名内部也有一定的结构，右边界比左边界更容易识别



机构名的内部构成规律 1

- 机构名一般都是定中结构（如：教育部语信司）
- 机构名的后缀一般比较集中，识别相对容易
- 机构名左边界识别非常困难
- 机构名中含有大量的人名、地名、企业字号等专有名称。在这些专有名称中，地名所占的比例最大，其中未登录地名又占了相当一部分的比例。所以机构名识别应在人名、地名等其他专名识别之后进行，其他专名识别的正确率对机构名识别正确率有较大影响



机构名的内部构成规律 2

- 中文机构名用词非常广泛。通过对人民日报1998年1月中的10817个机构名所含的19986个词进行统计，共计27种词，其中名词最多（9941个），地名其次（5023个），以下依次为简称（1169个）、专有名词（1125个）、动词（848个）以及机构名（714个）等
- 机构名长度极其不固定
- 机构名很不稳定。随着社会的发展，新机构不断涌现，旧机构不断被淘汰、改组或更名



基于字序列标注的方法

- 分词可以看做是对字加“词位标记”的过程
- “字”的词位分类示例：

B	E	M	S
词首	词尾	词中	独立词
人 们	古 人	小 人 国	听 人 说
LL	RR	MM	LR

Nianwen Xue & Libin Shen, 2003, Chinese Word Segmentation as LMR Tagging, In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. Sapporo, Japan: July 11—12, 2003, pp.176—179.

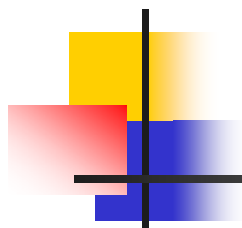
Nianwen Xue, 2003, Chinese Word Segmentation as Character Tagging, *Computational Linguistics and Chinese Language Processing*, Vol.8, No.1, pp.29-48.



基于字序列标注的方法

- 字标注的原理：根据字本身及其上下文的特征，来决定当前字的词位标注

特征模板示例	含义
C_0	当前字
C_{-2}, C_{-1}, C_1, C_2	当前字的左边第二字，第一字，右边第一字，第二字
$C_{-1}C_0, C_0C_1$ $C_{-2}C_{-1}, C_1C_2$	当前字跟其左边一个字，当前字跟其右边一个字 当前字的左边两个字，当前字的右边两个字
$C_{-1}C_1$	当前字的左边一个字加右边一个字
T_{-1} T_{-2}	左边第一个字的字位标注 左边第二个字的字位标注
Default feature	缺省特征



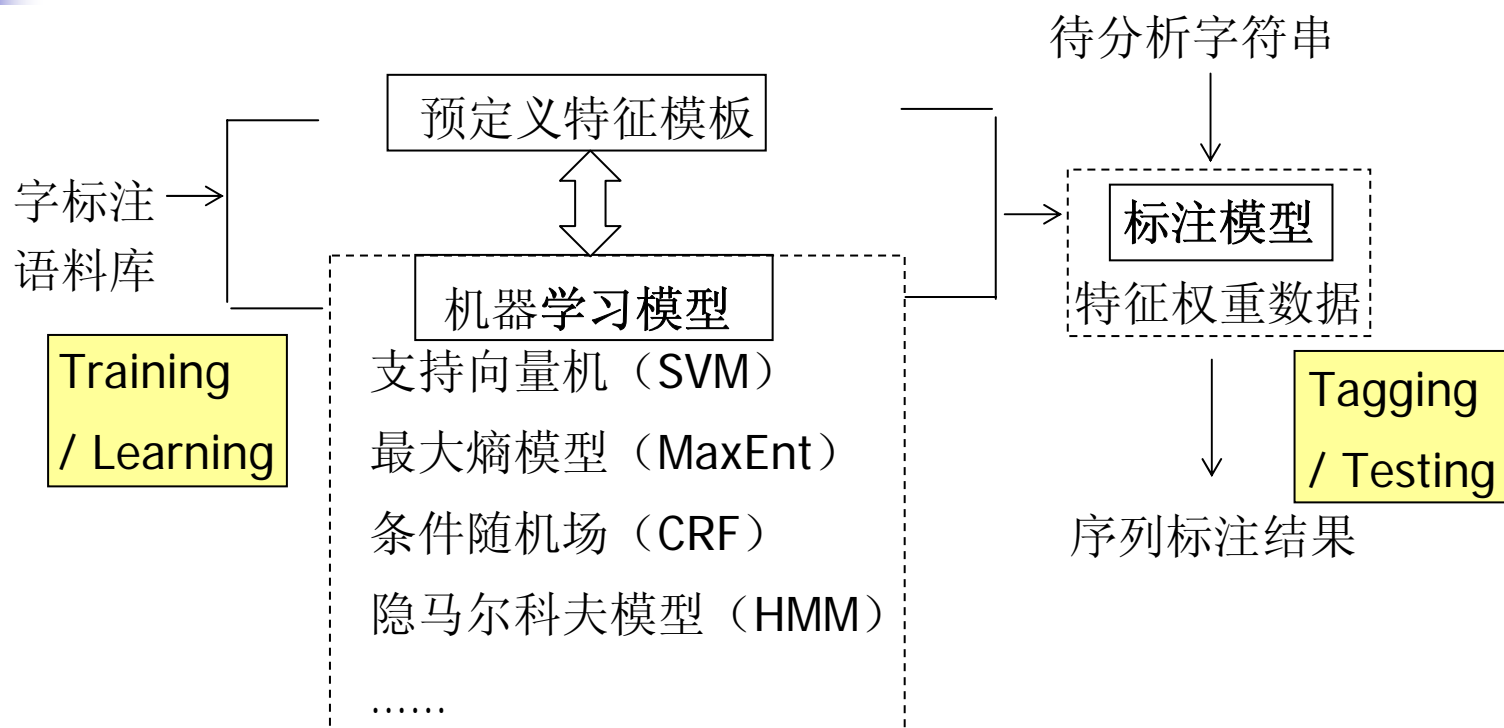
基于字序列标注的方法

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

C_0 生成的特征	C_1C_0 生成的特征	C_0C_1 生成的特征
和, S	的 和, S	和 尚, S
尚, B	和 尚, B	尚 未, B
未, E	尚 未, E	未 结, E
结, B	未 结, B	结 婚, B
婚, E	结 婚, E	婚 的, E
的, S	婚 的, S	的 都, S

.....

基于字序列标注的方法



- CRF 工具包 <http://crfpp.googlecode.com/svn/trunk/doc/index.html>
- Maximum Entropy 工具包 <https://github.com/lzhang10/maxent>
- SVM 工具包 <http://www.svms.org/software.html>



基于字序列标注的方法的优点

- 能够平衡地看待词表词和未登录词的识别问题。在这种分词技术中，文本中的词表词和未登录词都是用统一的字标注过程来实现的。在学习架构上，既可以不必专门强调词表词信息，也不用专门设计特定的未登录词(如人名、地名、机构名)识别模块。这使得分词系统的设计大大简化。在字标注过程中，所有的字根据预定义的特征进行词位特性的学习，获得一个概率模型。然后，在待分字串上，根据字与字之间的结合紧密程度，得到一个词位的标注结果。最后，根据词位定义直接获得最终的分词结果。总而言之，在这样一个分词过程中，分词成为字重组的简单过程。然而这一简单处理带来的分词结果却是令人满意的。



4 对文本分词质量的评价

- 计算分词正确率的不同标准
 - 以词数算
 - 以句数算
- 分词质量对NLP应用系统的影响
 - 分词质量对MT的影响
 - 分词质量对IR的影响
 - “移动电话”
 - 从合 —— 对翻译、校对有利
 - 从分 —— 对IR有利
 -



准确率、召回率、F-Score

- 准确率(precision)

$$\text{准确率 (P)} = \frac{\text{切分结果中正确分词数}}{\text{切分结果中所有分词数}} * 100\%$$

- 召回率(recall)

$$\text{召回率 (R)} = \frac{\text{切分结果中正确分词数}}{\text{标准答案中所有分词数}} * 100\%$$

- F-评价(F-measure 综合准确率和召回率的评价指标)

$$\text{F-指标} = \frac{2PR}{P+R}$$



中文分词效果评测

- 国内863计划，973计划，中文信息学会组织过多次评测
- 国际上SIGHAN bakeoff 2003 - 2007 <http://www.sighan.org/>

历届Sighan 在City-U 语料上评测结果F 值最好成绩

	Recall	Precision	F-score	Roov	训练语料词数	测试语料词数
2007	0.9526	0.9493	0.9510	0.7495	1.04M/43K	230K/23K
2006	0.9730	0.9720	0.9720	0.7870	1.6M/76K	220K/23K
2005	0.9410	0.9460	0.9430	0.6980	1.46M/69K	41K/9K
2003	0.9470	0.9340	0.9400	0.6250	240K	35K

刘群、钱跃良，2008，中文信息处理技术评测综述，《中国计算机学会通讯》2008 年第2期。



5 小结

- 歧义：交际型、组合型、链长
- 1. 分词的问题
 - 未登录词：不同类型的未登录词
 - 词的界定：规范+词表+语料库
- 2. 分词的方法
 - 从句到词：(a) 最大匹配 (b) 最佳路径
 - 从字到词：字序列标注法
- 3. 分词的评价
 - 指标：(1) 准确率P (2) 召回率R (3) F-Score
 - 评测：SIGHAN bakeoff等

分词方法的演进

① 最大匹配法 {正向、逆向、双向}

①.5 局部改进

- + 回溯机制
- + 歧义词表
- + 消歧规则

② 最优路径法

- 节点最少者最优
- 罚分最低者最优
- 概率最大者最优

③ 字位标注法

- 支持向量机模型
- 最大熵模型
- 条件随机场模型

分词歧
义问题

未登录
词问题



中文分词的意义

□ 词语破碎处，无物存在

—— 引自海德格尔《在通向语言的途中》，
商务印书馆1997年版

- 从字串到词串，存在着不确定性，分词的过程也就是一个降低不确定性的过程。为了降低不确定性，需要为计算机提供确定的“语言知识”，比如词典，规则，经过分词处理的语料库（可从中获取词语的各项统计数据）等知识形式。



进一步阅读文献

- 刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，第1—6章
- 赵铁军，2000，《机器翻译原理》，哈尔滨工业大学出版社，第3章
- 冯志伟，2001，《计算语言学基础》，商务印书馆，第2章
- 何克抗 等，1991，《书面汉语自动分词专家系统设计原理》，载《中文信息学报》，1991年第2期。
- 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。
- 孙茂松 等，1999，《高频最大交集型歧义切分字段在汉语自动分词中的作用》，载《中文信息学报》1999年第1期。
- 陈小荷，2000，《现代汉语自动分析》，北京语言文化大学出版社，第7章



复习思考题

- 1 什么是词？请谈谈你对“词”这个概念的认识。
- 2 汉语的自动分词面临哪些困难，请举例说明。
- 3 写出汉语词语重叠形式的分析规则。
- 4 归纳说明汉语产生新词的模式。
- 5 在互联网上找一篇字数在4000字左右的中文文章，进行人工分词，并列举、归纳碰到的问题。