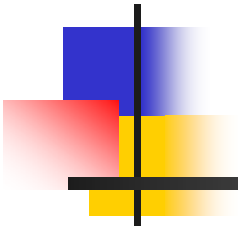
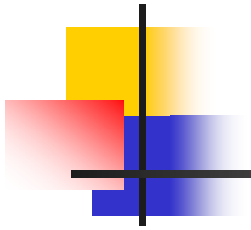


第七章 词汇分析（一）

—— 找出字符串中的“词”

詹卫东

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



提纲

- 1 从字符串到词串
- 2 英语词汇处理
  - 2.1 Tokenization
  - 2.2 Lemmatization
- 3 汉语词汇处理
  - 3.1 汉语自动分词面临的困难
  - 3.2 汉语自动分词的基本方法
  - 3.3 对分词质量的评价
- 4 小结

1 从字符串到词串

从字符串到词串，是一个降低不确定性的过程

- 简繁转换
  - 後面，皇后 —— 后
  - 松树，鬆开 —— 松
- 文语转换
  - 目的，调节（多音字）
  - 小雨伞　法语语法（连读变调）
- 文本校对（ 改错别字）
  - 抛妻别字 —— 抛妻别子 （字音编码输入）
  - 于预 —— 干预 （字形编码输入）
- 文本检索
  - 检索“人为”，可能输出的结果：人为因素、人为什么活着、以人为本、……

词汇分析是许多NLP应用系统的基础。

从字符串到词串（续）

- 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
- 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
- 我家门前的小河很难过

……

从字符串到词串，存在着不确定性

从字符串到词串（续）

Dog’s - Let’s

ad hoc - and so on - New York

strong – stronger – strongest

buy – bought

eat – ate - eaten

try - tried – tries

treat – treatment

在“字符串”这个层次上，“eat”跟“ate”是相同字母的不同顺序形式，

在“词串”这个层次上，“eat”跟“ate”是同一个词的不同表现形式

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

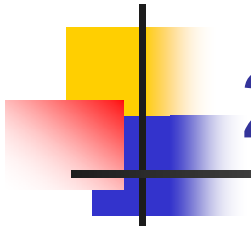
↓

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



2 英语的词汇处理

- 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)

2.1 Tokenization

1）数字：123,456.78[1]　90.7%　3/8　11/20/2000

2）缩略（包含不同的情况）：

a. 字母－点号－字母－点号组成的序列，比如：U.S.　i.e. 等等；

b. 字母开头，最后以点号结束，比如：A.　b.　Mr.　eds.　prof.；

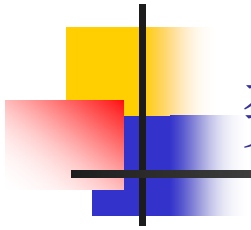
3）包含非字母字符，比如：AT&T　Micro$oft

4）带杠的词串，比如：three-years-old，one-third，so-called

5）带瞥号的词串，比如：I'm　can't　dog's　let's

6）带空格的词串，比如："and so on"，"ad hoc"

Note 1: 不同语言书写数字的习惯可能有较大差别，比如法语文本中这个数字就写成：123 456,78



数字的识别（正则表达式/regular expression）

a. 识别分数，日期的正则表达式：
[0-9] + (/ [0-9] +) +　　e.g. 12/21　5/13/2002

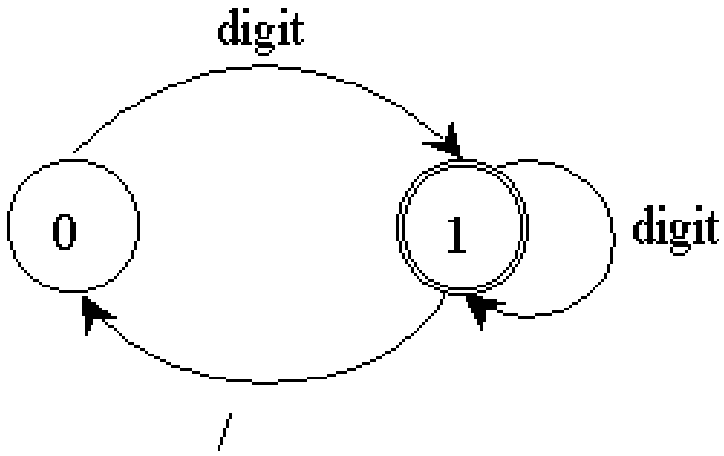
b. 识别百分数的正则表达式：
[\+ | \ -] ? [0-9]+ . ? [0-9] * %　　e.g. - 5.9%　91%

c. 识别十进制数字的正则表达式：
([0-9] + , ?) + (. [0-9] + | [0-9] +) *　e.g. 12,345

+表示出现1到无穷次，\ 表示转义，?表示不出现或只出现1次，
*表示出现0到无穷次, [ ] 表示单个字符，( )表示任意个字符

有关正则表达式的更多知识，可参看 杜淑敏 等编著《编译程序设计原理》，北京大学出版社1990年版，pp51-55

数字的识别（有限状态转移网络）



digit = 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

Tokenization算法的一般过程

（1）对一个待分析的字符串（S），从左到右进行扫描，读入当前字符（char）到候选词数组（W[i]），并将指针（pointer）前移，i=i+1；

（2）看char是否为词分隔符（事先可以预定义空格以及一般标点均为词分隔符）；

（3）如果char是词分隔符，并且W不是空格，将W中从起始位置到i-1位置的字符作为一个词汇单位输出，同时将S中的W部分删去，然后清空W，转入（1），如果char是词分隔符，且W是空格，将S中的W部分删去，清空W，转入（1）；

（4）如果不是词分隔符，看指针是否已经指到字符流尾部；

（5）如果指针已经指到字符流尾部，将当前W从起始位置到i-1位置的字符作为一个词汇单位输出，结束。

（6）如果不是字符流尾部，转入（1）；

Tokenization示例

字符流尾部标记

Char = “.” i=2, W=“Mr.” 输出: Mr， 这时S的格局为：

……

最后输出： Mr Wu was ill

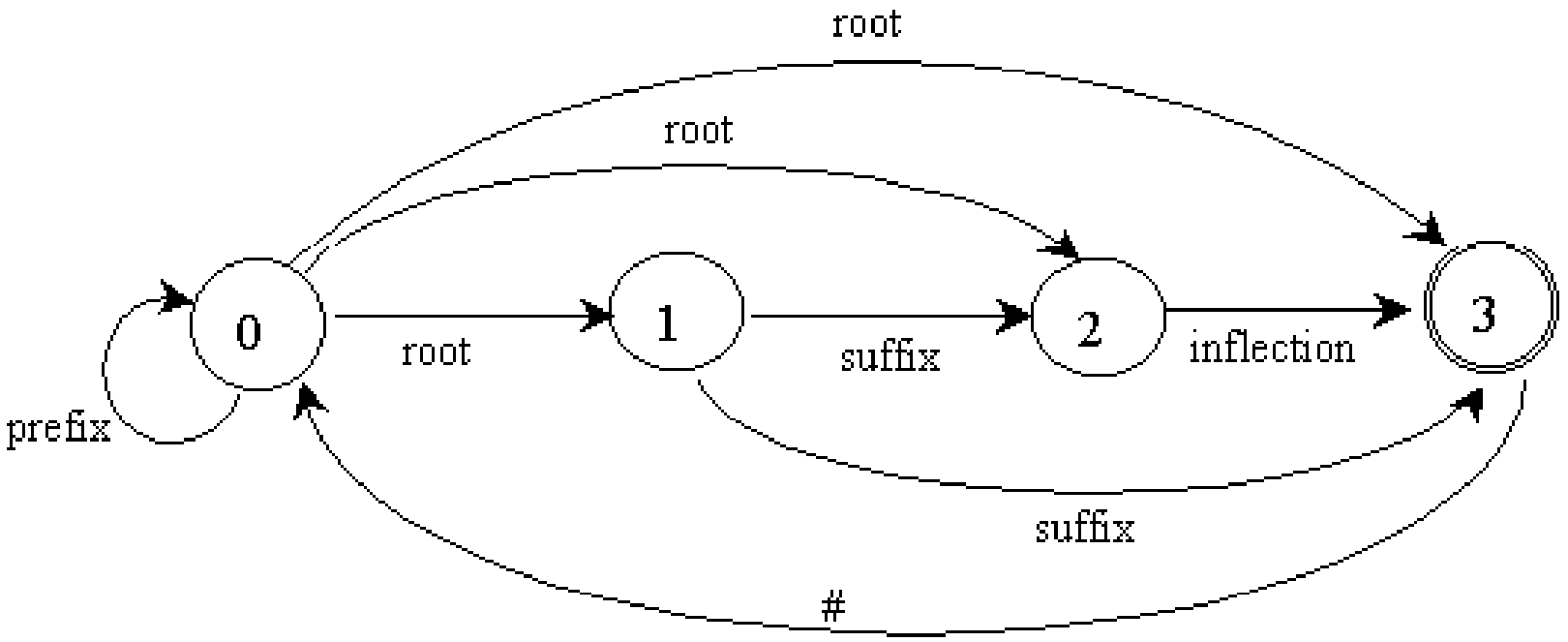
要得到“Mr.”，需要构造一个词典，收录这一类词

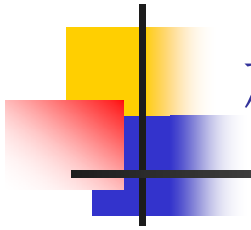
真实文本中还有http://icl.pku.edu.cn/， doubtfire@263.net之类的串需要识别!!!

2.2 Lemmatization

词干

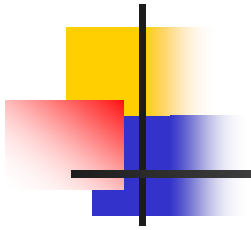
- 英语构词模式：
  {前缀}＋ {词根} ＋ {后缀} ＋ [词尾]





构词示例

- boy-s（复数词尾）
- happi-ness（名词后缀）
- im-possible　in-correct（前缀） -im = -in
- re-im-port （多个前缀）
- strongest（形容词最高级词尾）
- housewarmings（复合词＋复数词尾）



构词分析需要的知识库

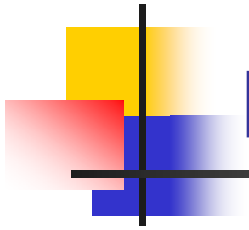
- 词典（Dict）
- 前缀表（PrefixList）
- 后缀表（SuffixList）
- 有关屈折词尾变形的规则（Rules）

比如可以构造下面这样的规则：

```
ies -> i        // 词尾字符串"ies"替换为i
i -> y          // 替换词尾"i"为"y“
s ->            // 词尾字符"s"替换为空
```

tries → try

boys → boy



Lemmatization算法的一般过程

（1）初始化：待分析的词形＝W，d=W的字符数，i=1，设输出串R="";

（2）到Dict中查找W，如果找到，R=W，转入（8）；

（3）如果i≤ (d/2)，执行（4）到（7）步，否则转入（8）；

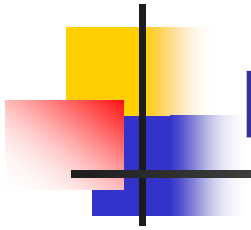
（4）从W中取出i个尾字字符，W分成W1 + W2（W2为取出的尾字符串）；

（5）到SuffixList中查找W2，如果查到，调用规则，对W1进行处理，得到W1';

（6）到Dict中查找W1'，如果找到，R=W1' + " " + W2，转入（8）；

（7）如果没有找到，i=i+1，转入（3）；

（8）输出R，结束；



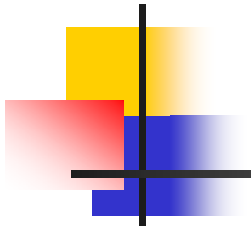
Lemmatization示例

- 待分析的词形 W=“boys” , d= 4, i=1, R=“”
- W不在词典中，从W中取出1个尾字符， “boy” + “s”
- W2=“s”，W1’=“boy”
- 输出： “boy” + “s”

Lemmatization容易碰到的问题

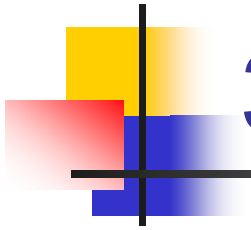
- 不规则词形变化：
  child — children
- 歧义问题：
  1）是词缀 还是 词根中的字符，有时不易判断
  比如：分析副词词尾 “ly”的规则：
  （1）将串尾字符 “y”去掉；
  （2）如果剩下的字符串以 “ll”结尾，将 “ll”变为 “le”
  - wholly → whol → whole
  - fully → ful → fule
  - only，inform，……

  2）不同的词根原形，相同的词形变化
  best <- good / well?



Lemmatization要做到何种程度

- 词干层。如：impossibilities→impossibility+ies
- 词根层。如：impossibilities→im+poss+ibil+it+ies
- 分析程度取决于自然语言处理系统的深度：
  - 不解决未定义词，分析到词干层
  - 解决未定义词，要分析到词根层。

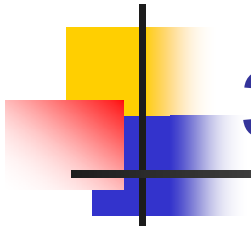


3 汉语词汇处理

3.1 汉语自动分词面临的困难

3.2 汉语自动分词的基本方法

3.3 对分词质量的评价



3.1 汉语自动分词面临的困难

1. 分词歧义

2. 未登录词识别

3. 分词规范：什么是中文的“词”？

3.1.1 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%
- 真歧义：6%
  - 多种切分均匀分布 12% 甲
  - 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。

3.1.2 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

3.1.3 分词规范：什么是“词”？

例：“联合国教科文组织”是不是一个词？

- 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）

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。

一些有代表性的分词规范

- 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

认同率
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.

3.2 汉语自动分词的基本方法

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

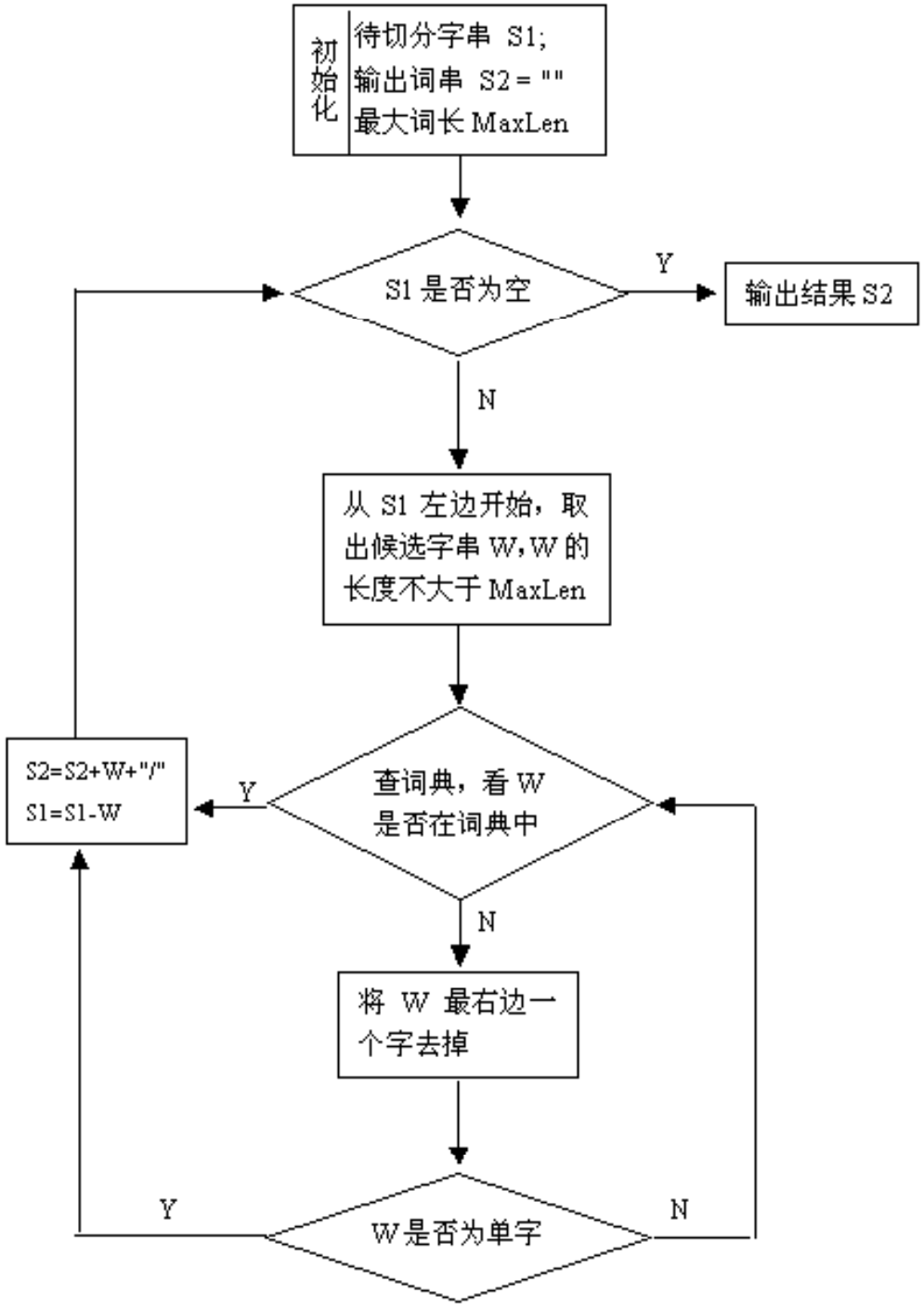
汉语分词的形式化定义，可参看马晏1991，基于评价的汉语自动分词系统的研究与实现，载黄昌宁、夏莹编《语言信息处理专论》，清华大学出版社1996年版

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

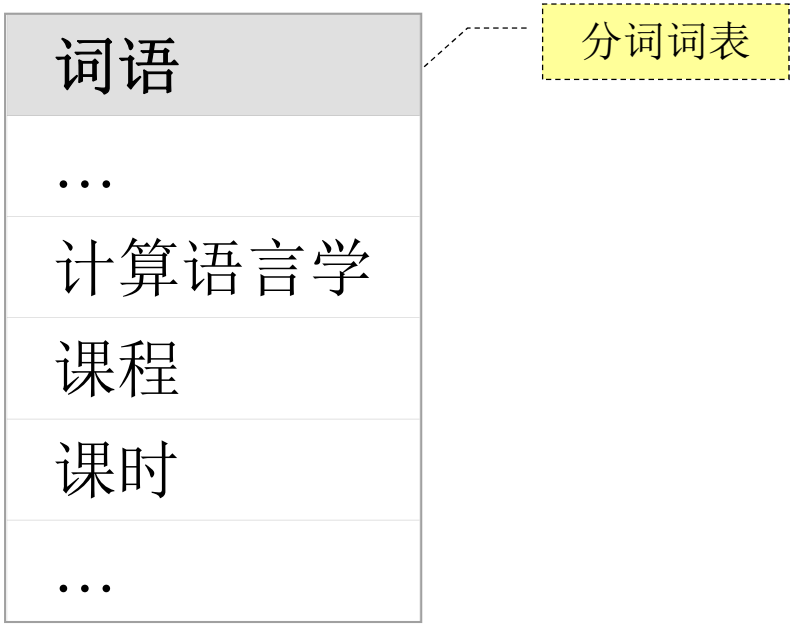


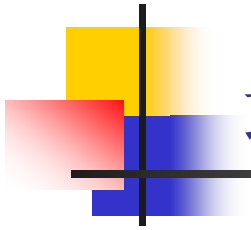
最大匹配法分词示例

S1="计算语言学课程是三个课时"

设定最大词长
MaxLen = 5

S2= " "





最大匹配法分词示例（续）

（1）S2=""；S1不为空，从S1左边取出候选子串W="计算语言学"；

（2）查词表，“计算语言学”在词表中，将W加入到S2中，S2=“计算语言学/ ”，并将W从S1中去掉，此时S1="课程是三个课时"；

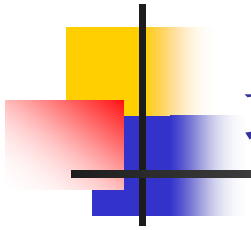
（3）S1不为空，于是从S1左边取出候选子串W="课程是三个"；

（4）查词表，W不在词表中，将W最右边一个字去掉，得到W="课程是三"；

（5）查词表，W不在词表中，将W最右边一个字去掉，得到W="课程是"；

（6）查词表，W不在词表中，将W最右边一个字去掉，得到W="课程"

（7）查词表，W在词表中，将W加入到S2中，S2=“计算语言学/ 课程/ ”，并将W从S1中去掉，此时S1="是三个课时"；



最大匹配法分词示例（续）

（8）S1不为空，于是从S1左边取出候选子串W="是三个课时"；

（9）查词表，W不在词表中，将W最右边一个字去掉，得到W="是三个课"；

（10）查词表，W不在词表中，将W最右边一个字去掉，得到W="是三个"；

（11）查词表，W不在词表中，将W最右边一个字去掉，得到W="是三"

（12）查词表，W不在词表中，将W最右边一个字去掉，得到W=“是”，这时W是单字，将W加入到S2中，S2=“计算语言学/ 课程/ 是/ ”，并将W从S1中去掉，此时S1="三个课时"；

（13）S1不为空，从S1左边取出候选子串W="三个课时"；

（14）查词表，W不在词表中，将W最右边一个字去掉，得到W="三个课"；

（15）查词表，W不在词表中，将W最右边一个字去掉，得到W="三个"；

（16）查词表，W不在词表中，将W最右边一个字去掉，得到W=“三”，这时W是单字，将W加入到S2中，S2=“计算语言学/ 课程/ 是/ 三/ ”，并将W从S1中去掉，此时S1="个课时"；

最大匹配法分词示例（续）

（17）S1不为空，从S1左边取出候选子串W="个课时"；

（18）查词表，W不在词表中，将W最右边一个字去掉，得到W="个课"；

（19）查词表，W不在词表中，将W最右边一个字去掉，得到W=“个”，这时W是单字，将W加入到S2中，S2=“计算语言学/ 课程/ 是/ 三/ 个/ "，并将W从S1中去掉，此时S1="课时"；

（20）S1不为空，从S1左边取出候选子串W="课时"；

（21）查词表，W在词表中，将W加入到S2中，S2=“计算语言学/ 课程/ 是/ 三/ 个/ 课时/ "，并将W从S1中去掉，此时S1=""。

（22）S1为空，输出S2作为分词结果，分词过程结束。

最大匹配法分词的问题

- 最大词长的确定

  （1）词长过短，长词就会被切错（“中华人民共和国”）

  （2）词长过长，效率就比较低

- 掩盖了分词歧义

A.“有意见分歧”　（正向最大匹配和逆向最大匹配结果不同）

有意/　见/　分歧/

有/　意见/　分歧/

B.“结合成分子时”　（正向最大匹配和逆向最大匹配结果相同）

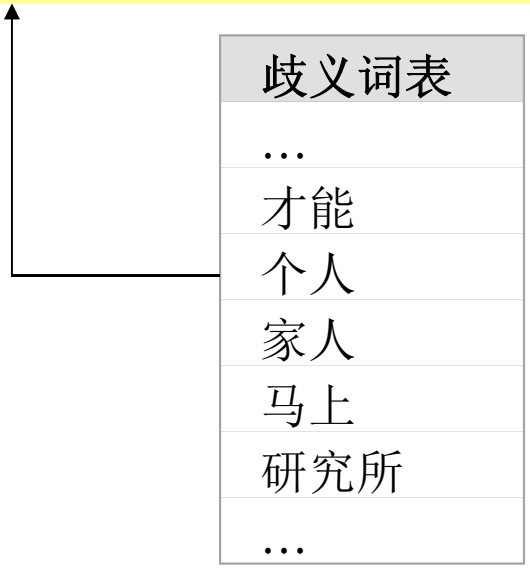
结合/　成分/　子时/

应对策略：增加语言知识，局部修改分词错误

局部修改1：增加歧义词表，排歧规则

规则示例

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



局部修改2：增加“回溯”机制

对于某些交集型歧义，可以通过增加回溯机制来改进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/”，

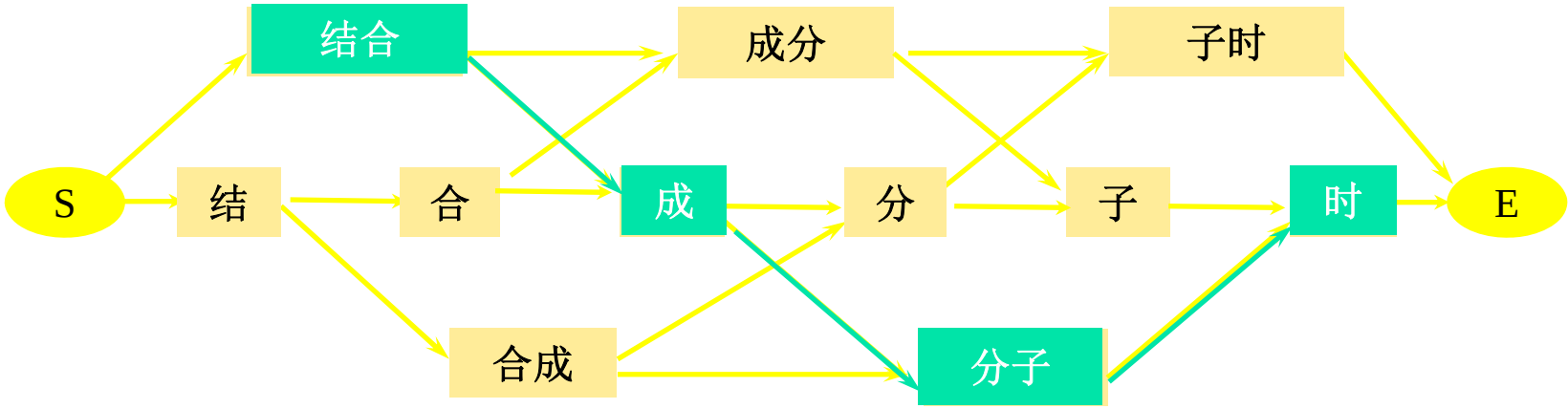
通过查词典知道“民”不在词典中，于是进行回溯，将“爱人”的尾字“人”取出与后面的“民”组成“人民”，再查词典，看“爱”，“人民”是否在词典中，如果在，就将分词结果调整为：“爱/ 人民/ 英雄/”

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

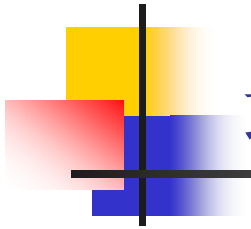
需要更全面地考虑分词的改进办法

汉语词语切分的数据结构—词图



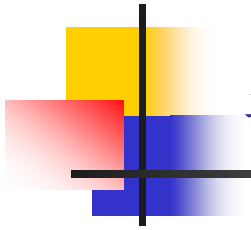
词图给出了一个字符串的**全部切分**可能性

分词任务：寻找一条起点S到终点E的**最优路径**



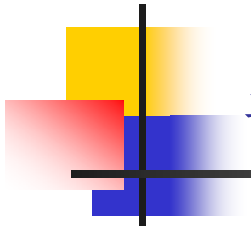
最短路径分词法

- 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：

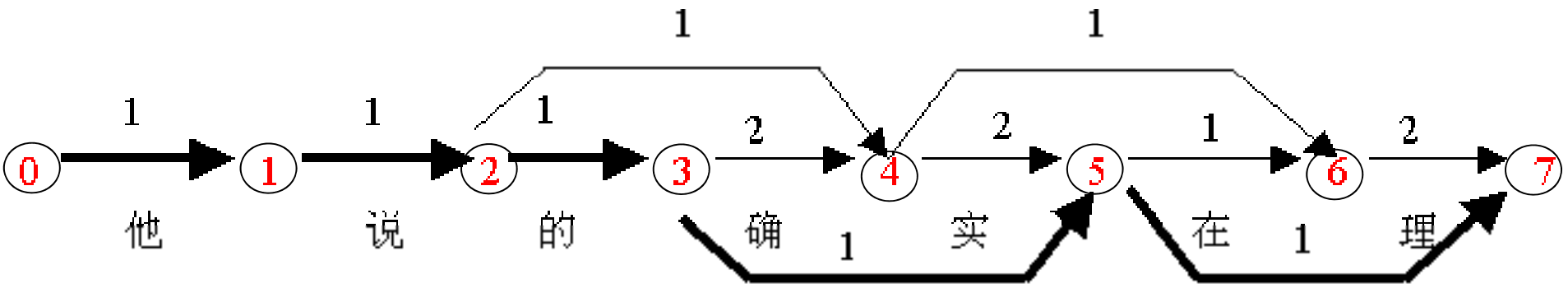
  1 每个词对应的边罚1分。

  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）

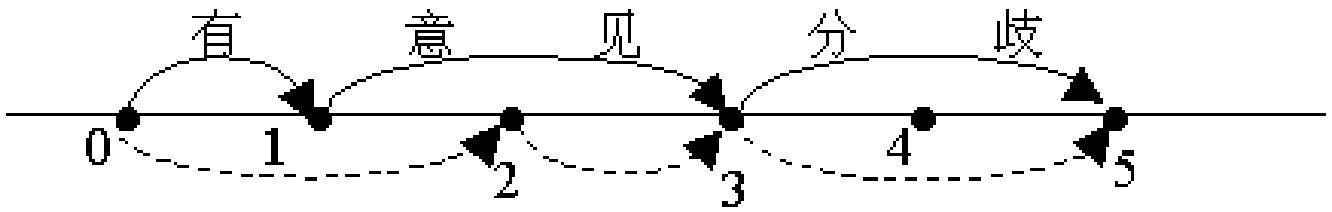
但是：仍然无法解决 “有意见分歧”的问题

最大概率法

基本思想是：

（1）一个待切分的汉字串可能包含多种分词结果

（2）将其中概率最大的那个作为该字串的分词结果



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

- S: 有意见分歧
  - W1: 有/ 意见/ 分歧/
  - W2: 有意/ 见/ 分歧/

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

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

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

独立性假设，一元语法

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

最大概率法

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

$P(W1) = P(有) * P(意见) * P(分歧)$

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

$P(W2) = P(有意) * P(见) * P(分歧)$

$= 1 \times 10^{-11}$

$P(W1) > P(W2)$

提高计算效率

如何尽快找到概率最大的词串（路径）？

到达候选词w_i时的累计概率

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

$$P'(\text{意见}) = P'(\text{有}) \times P(\text{意见})$$

$$P'(\text{有}) = P(\text{有})$$

提高计算效率（续）

- 左邻词

  假定对字串从左到右进行扫描，可以得到 $w_1, w_2, \ldots, w_{i-1}, w_i, \ldots$ 等若干候选词，如果 w_{i-1} 的尾字跟 w_i 的首字邻接，就称 w_{i-1} 为 w_i 的左邻词。比如上面例中，候选词“有”就是候选词“意见”的左邻词，“意见”和“见”都是“分歧”的左邻词。字串最左边的词没有左邻词。

- 最佳左邻词

  如果某个候选词 w_i 有若干个左邻词 $w_j, w_k, \ldots$ 等等，其中累计概率最大的候选词称为 w_i 的最佳左邻词。比如候选词“意见”只有一个左邻词“有”，因此，“有”同时也就是“意见”的最佳左邻词；候选词“分歧”有两个左邻词“意见”和“见”，其中“意见”的累计概率大于“见”累计概率，因此“意见”是“分歧”的最佳左邻词

最大概率法的实现

1) 对一个待分词的字串 S，按照从左到右的顺序取出全部候选词 $w_1, w_2, \ldots, w_i, \ldots, w_n$ ；
2) 到词典中查出每个候选词 的概率值$P(w_i)$ ，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概率，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，且累计概率$P'(w_n)$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

（1）对“有意见分歧”，从左到右进行一遍扫描，得到全部候选词：“有”，“有意”，“意见”，“见”，“分歧”；

（2）对每个候选词，记录下它的概率值，并将累计概率赋初值为0；

（3）顺次计算各个候选词的累计概率值，同时记录每个候选词的最佳左邻词：

P'(有)=P(有)，

P'(有意) = P(有意)，

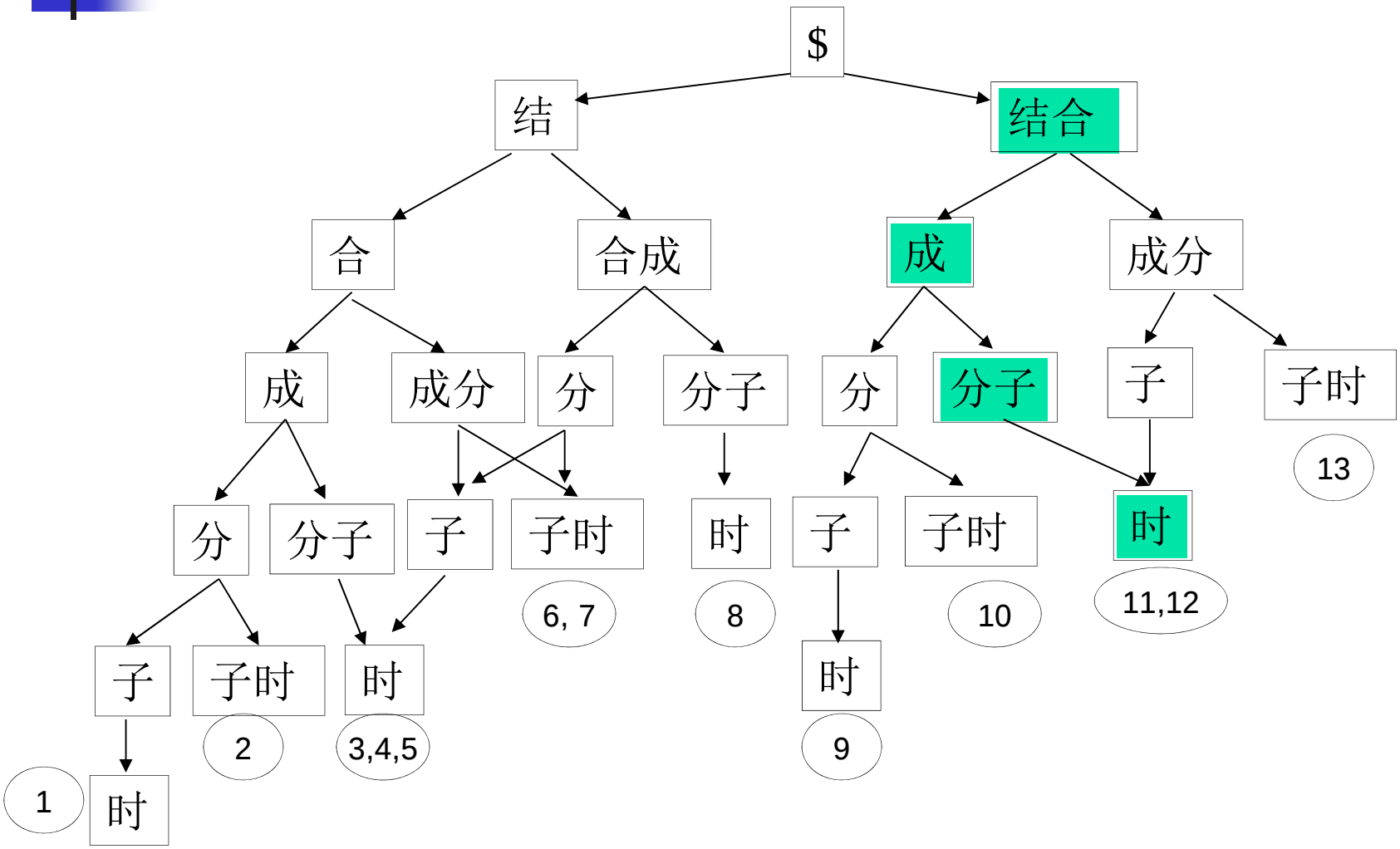
P‘(意见)=P’(有) × P(意见)，（“意见”的最佳左邻词为“有”）

P‘(见)=P’(有意) × P(见)，（“见”的最佳左邻词为“有意”）

P'(意见)>P'(见)

（4）“分歧”是尾词，“意见”是“分歧”的最佳左邻词，分词过程结束，输出结果：有/ 意见/ 分歧/

“结合成分子时”的全部分词路径示意



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)

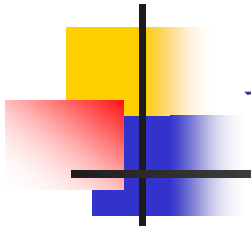
  W2= 这/ 事/ 的/ 确定/ 不/ 下来/

- 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；

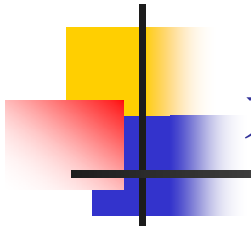
2) 通过构词规则和上下文特征规则来识别；
“雪村先生创作了很多歌曲”

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



未登录词识别的一般方法

- 每一类未定义词都要构造专门的识别算法
- 识别依据
  - 内部构成规律（用字规律）
  - 外部环境（上下文）
  - 重复出现规律

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

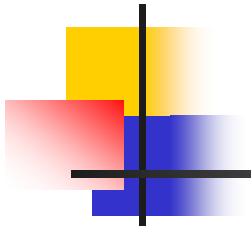
- 中国人名内部各组成部分的组合规律
  - 姓＋名
  - 姓
  - 名
  - 前缀＋姓
  - 姓＋后缀
  - 姓＋姓＋名（海外已婚妇女）

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[朝阳]、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
- 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：

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

$$P(r1) = P(name \mid word)$$

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

$$P(r6) = P(C_0 \mid 1-hanzifamily)$$

$$P(r9) = P(C_1C_2 \mid 2-hanzigiven)$$

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.

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

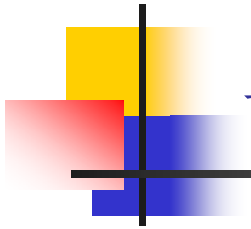
音译名的识别 1

- 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亭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瑙赉滕斐熙鸠窦艮麟黛”。

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（蒙语人名举例：“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

机构名的内部构成规律 1

- 机构名一般都是定中结构
- 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：

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

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

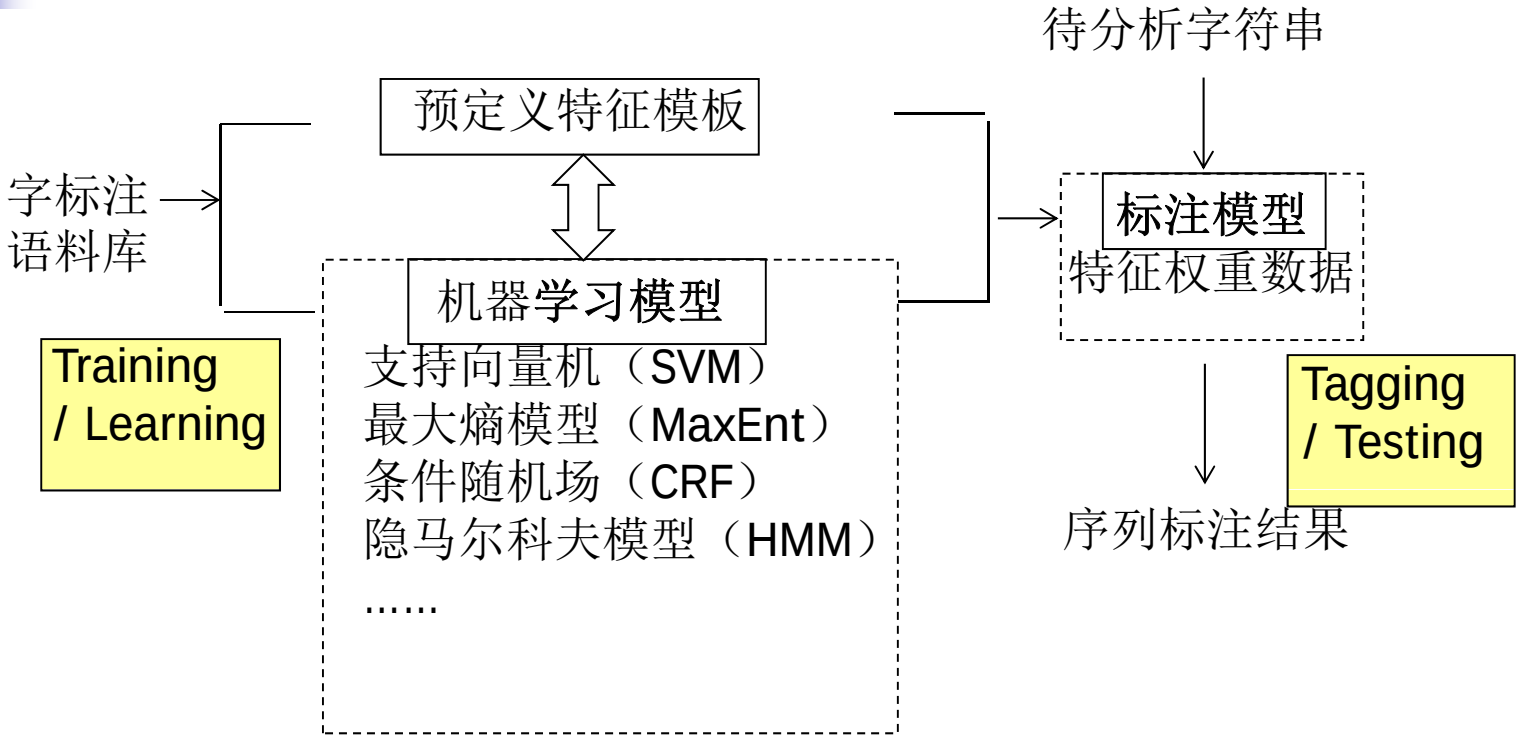
基于字序列标注的方法

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	$C_{-1}C_0$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。

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。

3.3 对分词质量的评价

- 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)

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

- 召回率(recall)

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

- F-评价(F-measure 综合准确率和召回率的评价指标)

$$F\text{-}指标 = \frac{2PR}{P+R}$$

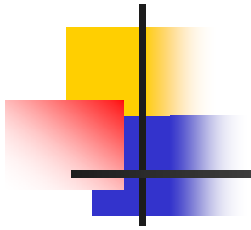
中文分词效果评测

- 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

1. 分词的问题

分词歧义：交际型、组合型、链长
未登录词：未登录词的类型
词的界定：规范+词表+语料库

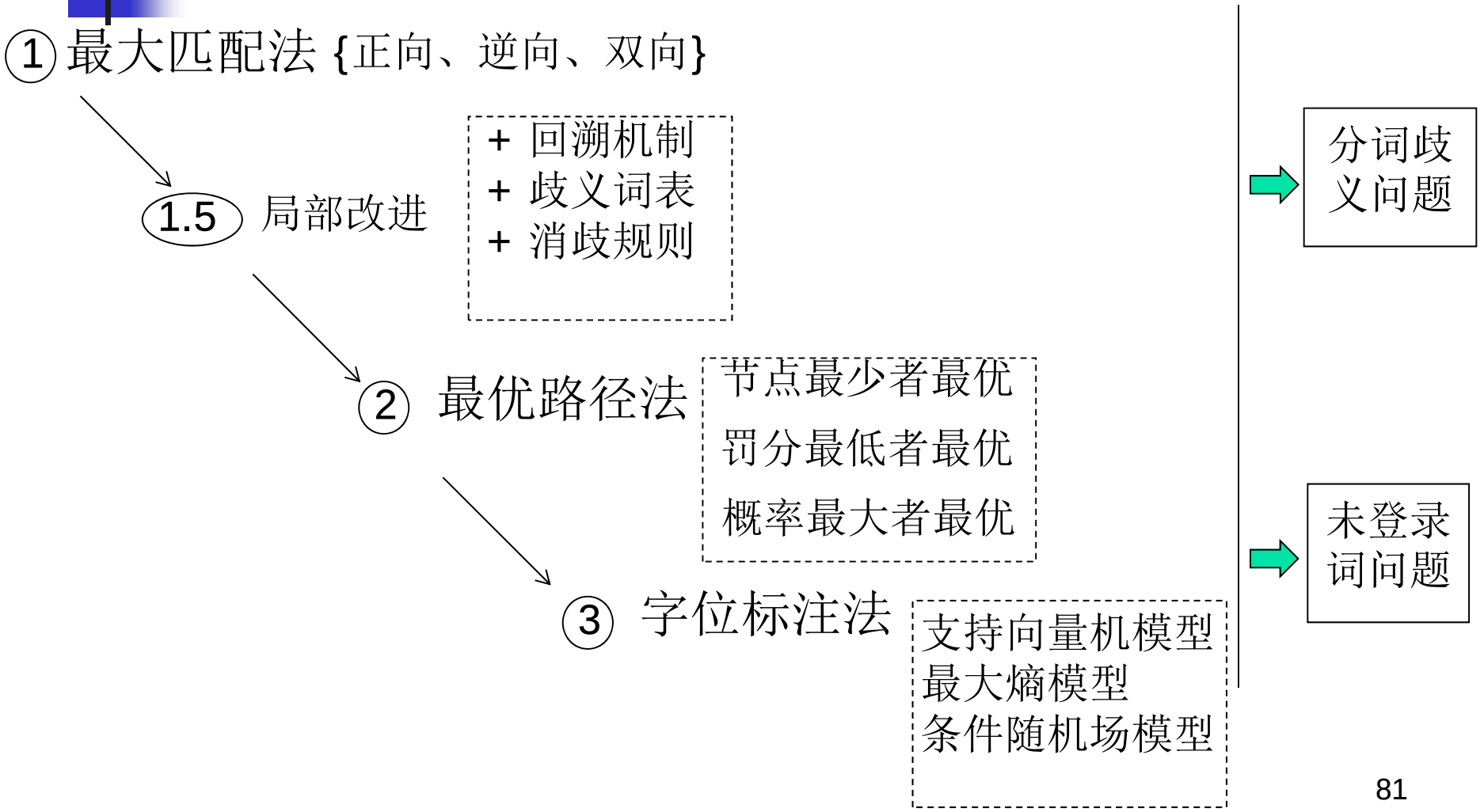
2. 分词的方法

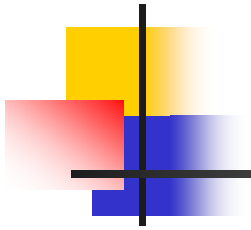
从句到词：(a) 最大匹配 (b) 最佳路径
从字到词：字序列标注法

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

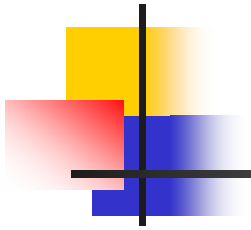


小结（续）

- 词语破碎处，无物存在

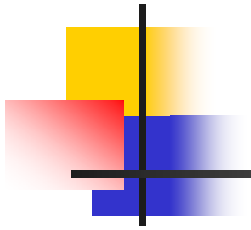
  —— 引自海德格尔《在通向语言的途中》，商务印书馆1997年版

- 从字串到词串，存在着多种可能性（不确定性），因而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”，这种知识可以是词典形式的，可以是规则形式的，也可以是统计数据形式的。



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
- 语言学光标网站之词法分析板块



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。