



# 第7章 自动分词 与命名实体识别

---



## 7.1 汉语自动分词概要

---

### 分词的意义

- 正确的机器自动分词是正确的中文信息处理的基础
  - 文本检索
    - 和服 | 务 | 于三日后裁制完毕，并呈送将军府中。
    - 王府饭店的设施 | 和 | 服务 | 是一流的。  
如果不分词或者“和服务”分词有误，都会导致荒谬的检索结果。
  - 文语转换
    - 他们是来 | 查 | 金泰 | 撞人那件事的。（“查”读音为cha）
    - 行侠仗义的 | 查金泰 | 远近闻名。（“查”读音为zha）



## 7.1 汉语自动分词概要

---

汉语文本是基于单字的，词与词之间没有显性的界限标志，因此分词是汉语文本分析处理中首先要解决的问题。

添加显性的**词语边界标志**，使得所形成的词串反映句子的本意，这个过程就是分词。



## 7.1 汉语自动分词概要

---

分词面临的主要难点：

- (1) 歧义切分问题
- (2) 未登录词识别问题



## 7.1 汉语自动分词概要

---

### ➤ 歧义切分处理

#### 1、中国人为了实现自己的梦想 (交集型歧义)

中国/ 人为/ 了/ 实现/ 自己/ 的/ 梦想

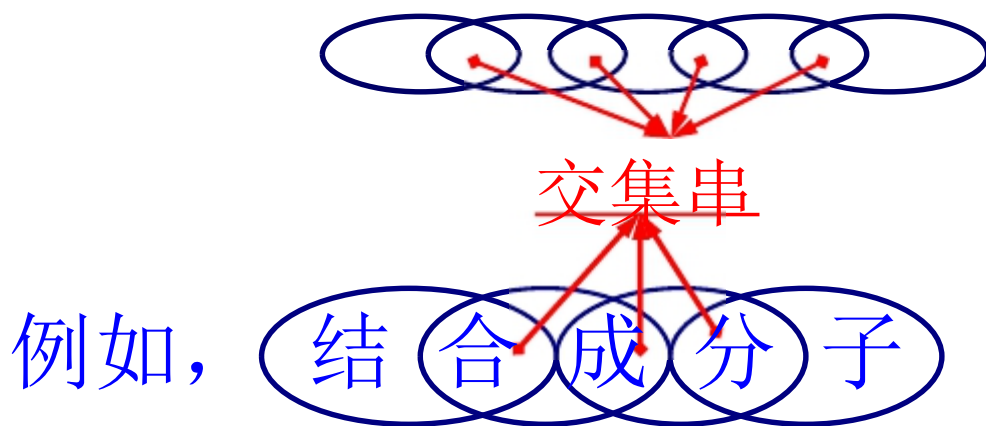
中国人/ 为了/ 实现/ 自己/ 的/ 梦想

中/ 国人/ 为了/ 实现/ 自己/ 的/ 梦想

例如：“大学生”、“研究生物”、“从小学起”、“为人民工作”、“中国产品质量”、“部分居民生活水平”等等

## 7.1 汉语自动分词概要

- 定义：链长 一个交集型切分歧义所拥有的交集串的集合称为交集串链，它的个数称为链长。



“结合”、“合成”、“成分”和“分子”均构成词，交集串的集合为{合，成，分}，因此，链长为3。



## 7.1 汉语自动分词概要

---

类似地，

(1) “为人民工作”

{人，民，工}，歧义字段的链长为 3；

(2) “中国产品质量”

{国，产，品，质}，歧义字段的链长为 4；

(3) “部分居民生活水平”

{分，居，民，生，活，水}，链长为 6。



## 7.1 汉语自动分词概要

---

2、门把手弄坏了。（组合型歧义）

门/ 把/ 手/ 弄/ 坏/ 了/ 。

门/ 把手/ 弄/ 坏/ 了/ 。

例如，“将来”、“现在”、“才能”、“学生会”等，都是组合型歧义字段。





## 7.1 汉语自动分词概要

---

### ➤ 未登录词的识别

#### 1、人名、地名、组织机构名等，例如：

盛中国，张建国，李爱国，蔡国庆，令计划；  
高升，高山，夏天，温馨，武夷山，时光；  
彭太发生，朱李月华；赛福鼎 艾则孜，爱新觉  
罗 溥仪；平川三太郎，约翰 斯特朗

#### 2、新出现的词汇、术语、个别俗语等，例如：

博客，非典，禽流感，恶搞，微信，给力，失联



## 7.1 汉语自动分词概要

---

例如：

- (1) 他还兼任何应钦在福州办的东路军军官学校的政治教官。
- (2) 大不列颠及北爱尔兰联合王国外交和英联邦事务大臣、议会议员杰克 斯特劳阁下在联合国安理会就伊拉克问题发言。
- (3) 坐落于江苏省南京市玄武湖公园内的夏璞墩是晋代著名的文学家、科学家夏璞的衣冠冢。

## 7.1 汉语自动分词概要

错误类型			错误数	比例(%)			例子
集外词	命名实体	人名	31	25.83	55.0	98.33	约翰·斯坦贝克
		地名	11	9.17			米苏拉塔
		组织机构名	10	8.33			泰党
		时间和数字	14	11.67			37万兆
	专业术语		4	3.33		脱氧核糖核酸	
	普通生词		48	40.00		致病原	
切分歧义			2	1.67			歌名为
合计			120	100			

互联网上随机摘取了418个句子，共含11,739个词，19,777个汉字



## 7.1 汉语自动分词概要

### ◆ 汉语自动分词的基本原则

《信息处理用现代汉语分词规范及自动分词方法》

- 1 二字、三字、四字词，以及结合紧密、使用稳定的：发展 红旗 对不起 自行车 青霉素 由此可见
- 2 四字成语一律为分词单位：胸有成竹 欣欣向荣
- 3 五字和五字以上的谚语、格言等，分开后如不违背原有组合的意义，应予切分：  
时间/就/是/生命/  
失败/是/成功/之/母



## 7.1 汉语自动分词概要

- 4 结合紧密、使用稳定的词组则不予切分:不管三七二十一
- 5 惯用语和有转义的词或词组:  
妇女能顶/半边天/  
他真小气, 象个/铁公鸡/
- 6 略语一律为分词单位:科技 奥运会 工农业
- 7 分词单位加形成儿化音的“儿”:花儿 悄悄儿  
玩儿
- 8 阿拉伯数字等, 仍保留原有形式:1234  
7890
- 9 现代汉语中其它语言的汉字音译外来词, 不予切分:巧克力 吉普



## 7.1 汉语自动分词概要

---

- 1 动词前的否定副词一律单独切分:不/写 不/能 没/研究 未/完成
- 2 动宾结构的词或结合紧密、使用稳定的:开会 跳舞 解决/吃饭/问题 孩子该/念书/了
- 3 动补结构的二字词或结合紧密、使用稳定的二字动补词组, 不予切分:打倒 提高 加长
- 4 偏正结构的词, 以及结合紧密的词不予切分:胡闹 瞎说 死记
- 5 多字动词无连词并列, 一律切分:调查/研究 宣传/鼓动



---

## 7.2 分词性能评价



# 7.1 分词性能评价

## ◆ 测试方法：

### ● 封闭测试 vs. 开放测试

封闭测试只允许使用固定训练语料学习，而开放测试可以使用任意资源

### ● 专项测试 vs. 总体测试



- ✓ 歧义字段切分能力
- ✓ 集外词(生词)处理能力
- ✓ 人名、地名、组织机构名等命名实体识别能力





## 7.1 分词性能评价

---

### ◆评价指标

- **正确率**(Correct ratio/Precision,  $P$ ): 测试结果中正确切分或标注的个数占**系统输出结果**的比例。假设系统输出 $N$ 个, 其中, 正确的结果为 $n$ 个, 那么,

$$P = \frac{n}{N} \times 100\%$$



## 7.1 分词性能评价

- **召回率** (Recall ratio,  $R$ ): 测试结果中正确结果的个数占**标准答案总数**的比例。假设系统输出 $N$ 个结果,其中正确的结果为 $n$ 个,而标准答案的个数为 $M$ 个,那么,

$$R = \frac{n}{M} \times 100\%$$

两种标记:  $R_{OOV}$  指**集外词**的召回率;  
 $R_{IV}$  指**集内词**的召回率。



## 7.1 分词性能评价

---

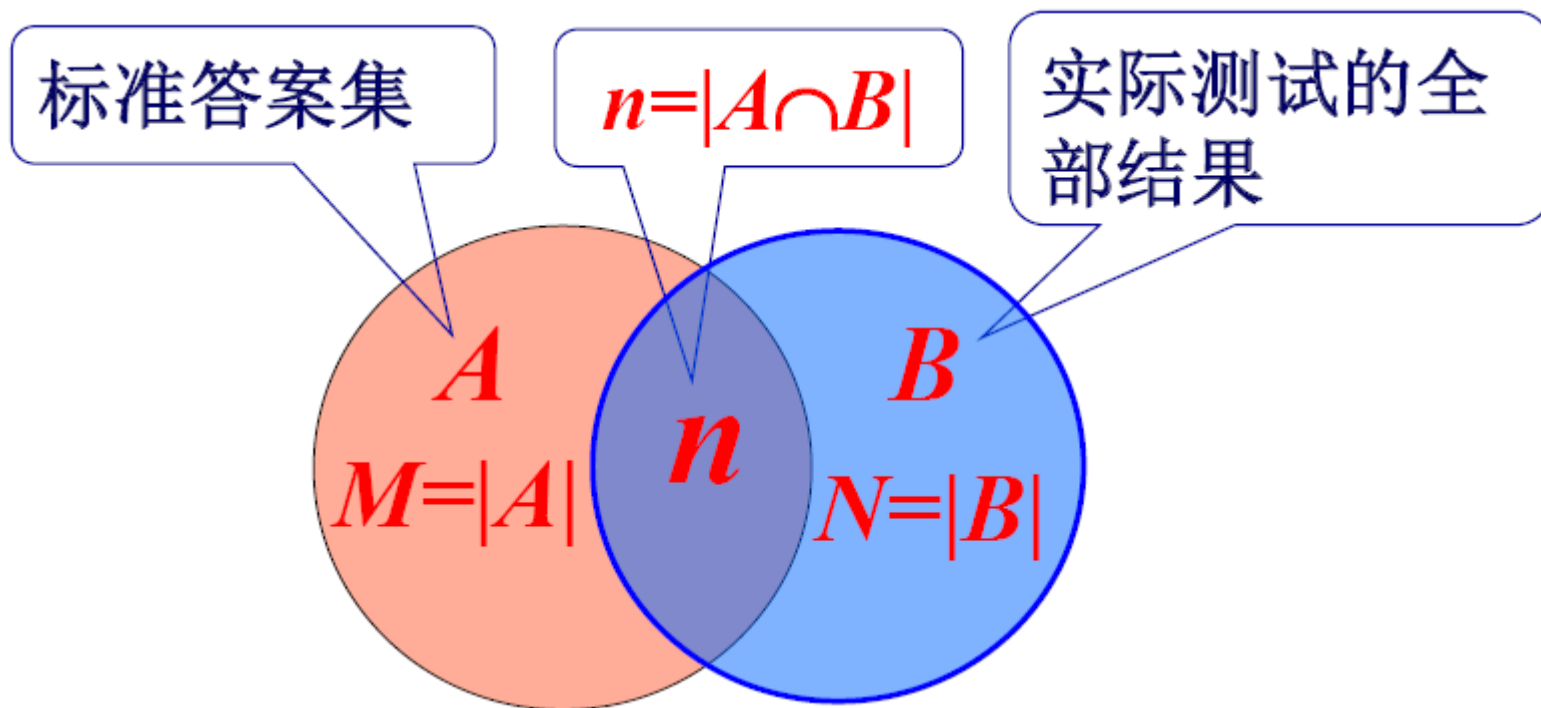
- **F-测度值**(F-Measure): 正确率与召回率的综合值。  
计算公式为:

$$F-measure = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times (P + R)} \times 100\%$$

一般地, 取  $\beta = 1$ , 即

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

## 7.1 分词性能评价



$$P = \frac{n}{N} \times 100\%$$

$$R = \frac{n}{M} \times 100\%$$



## 7.1 分词性能评价

假设某个汉语分词系统在一测试集上输出 **5260** 个分词结果，而标准答案是 **4510** 个词语，根据这个答案，系统切分出来的结果中有 **4120** 个是正确的。那么：

$$P = \frac{4120}{5260} \times 100\% = 78.33\%$$

$$R = \frac{4120}{4510} \times 100\% = 91.35\%$$

$$\begin{aligned} F1 &= \frac{2 \times P \times R}{P + R} \times 100\% \\ &= \frac{2 \times 78.33 \times 91.35}{78.33 + 91.35} \times 100\% \\ &= 84.34\% \end{aligned}$$



---

## 7.3 自动分词基本算法



## 7.3 汉语自动分词基本算法

---

◆ 有词典切分/ 无词典切分

◆ 基于规则的方法/ 基于统计的方法



## 7.3 汉语自动分词基本算法

### 1. 最大匹配法 (Maximum Matching, MM)

— 有词典切分，机械切分

- 正向最大匹配算法 (Forward MM, FMM)
- 逆向最大匹配算法 (Backward MM, BMM)
- 双向最大匹配算法 (Bi-directional MM)

假设句子:  $S = c_1c_2 \cdots c_n$

某一词:  $w_i = c_1c_2 \cdots c_m$ ,  $m$  为词典中最长词的字数。

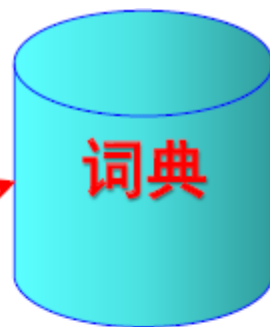
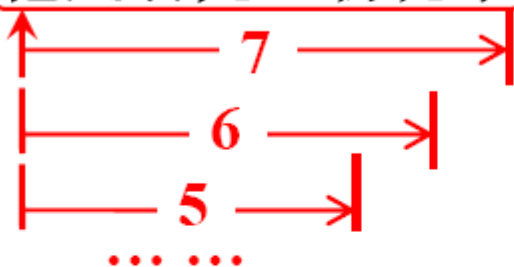


## 7.3 汉语自动分词基本算法

例：假设词典中最长单词的字数为 7。

输入字符串：他是研究生物化学的一位科学家。

切分过程：



他/ 是研究生物化学的一位科学家。



FMM 切分结果：他/ 是/ 研究生/ 物化/ 学/ 的/ 一/ 位 / 科学家/ 。

BMM 切分：他是研究生物化学的一位科学家。



BMM 切分结果：他/ 是/ 研究/ 生物/ 化学/ 的/ 一/ 位/ 科学家/ 。



## 7.3 汉语自动分词基本算法

### ➤ FMM 算法描述

- 1. 设自动分词词典中最长词条所含汉字个数为 $I$ ;
- 2. 取被处理材料当前字符串序数中的 $I$ 个字作为匹配字段，查找分词词典。若词典中有这样的—个 $I$ 字词，则匹配成功，匹配字段作为一个词被切分出来，转6;
- 3. 如果词典中找不到这样的—个 $I$ 字词，则匹配失败;
- 4. 匹配字段去掉最后一个汉字， $I--$ ;
- 5. 重复2-4，直至切分成功为止;
- 6.  $I$ 重新赋初值，转2，直到切分出所有词为止。



## 7.3 汉语自动分词基本算法

### ➤ 逆向最大匹配算法 (Backward MM, BMM)

- 分词过程与**FMM**方法相同，不过是从句子(或文章)末尾开始处理，每次匹配不成功时去掉的是前面的一个汉字
- “市场/中/国有/企业/才能/发展/”
- 实验表明：逆向最大匹配法比最大匹配法更有效，错误切分率为1 / 245



## 7.3 汉语自动分词基本算法

- 双向最大匹配算法 (Bi-directional MM)
  - 比较**FMM**法与**BMM**法的切分结果，从而决定正确的切分
  - 可以识别出分词中的交叉歧义



## 7.3 汉语自动分词基本算法

### ➤ 双向最大匹配算法 (Bi-directional MM)

分词消歧启发式规则：

1. 如果正、反向分词结果词数不同，则取分词数量较少的那个。
2. 如果分词结果词数相同
  - a. 分词结果相同，就说明没有歧义，可返回任意一个。
  - b. 分词结果不同，返回其中单字较少的那个。

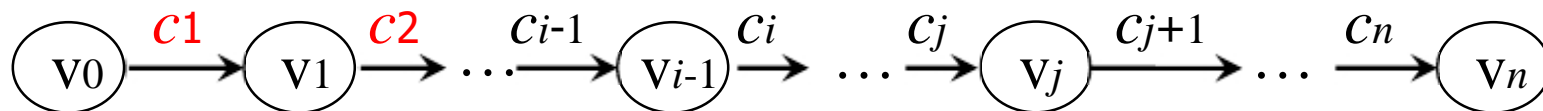
## 7.3 汉语自动分词基本算法

### 2. 最少分词法(最短路径法)

有词典切分

#### ➤ 基本思想

设待切分字符串  $S=c_1 c_2 \dots c_n$ , 其中  $c_i (i=1, 2, \dots, n)$  为单个的字。以字或词为边, 建立一有向无环图G:



目标: 寻找该图中含词量最少的路径。  
求最短路径: 贪心法。

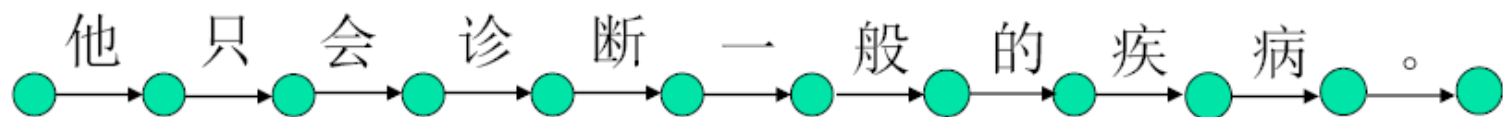
## 7.3 汉语自动分词基本算法

例：(1) 输入字串：他只会诊断一般的疾病。

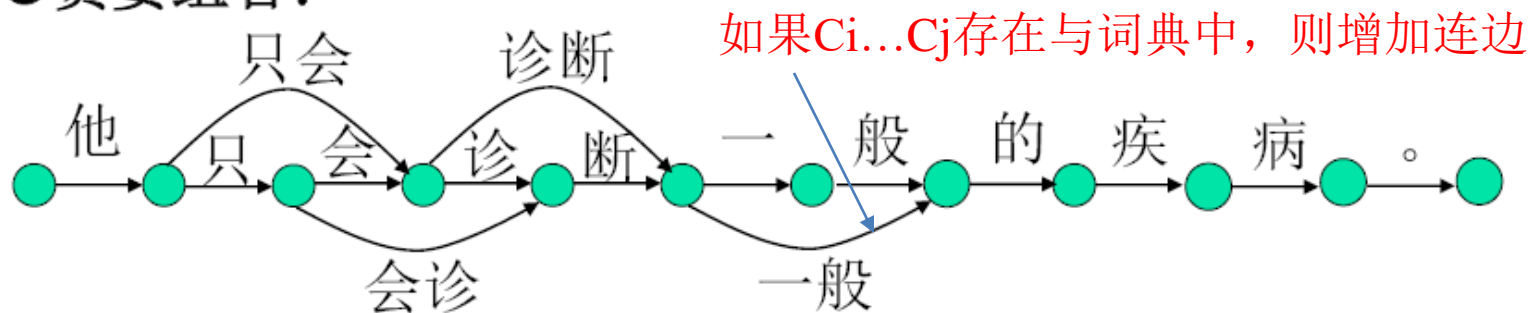
① 准备：词典



② 构建词图：



③ 贪婪组合：



输出候选： 他/ 只会/ 诊断/ 一般/ 的/ 疾病/。 (词个数： 7)

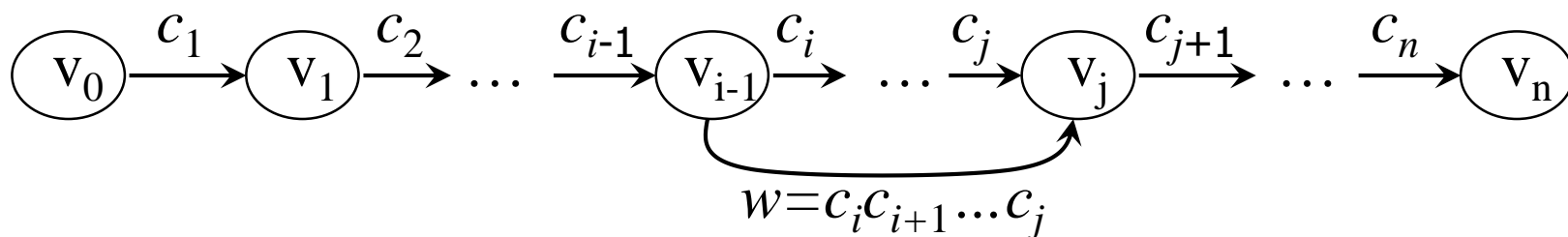
他/ 只/ 会诊/ 断/ 一般/ 的/ 疾病/。 (词个数： 8)

最终结果： 他/ 只会/ 诊断/ 一般/ 的/ 疾病/ 。

## 7.3 汉语自动分词基本算法

### ● 算法描述:

- (1) 相邻节点  $v_{k-1}, v_k$  之间建立有向边  $\langle v_{k-1}, v_k \rangle$ , 边对应的词默认为  $c_k$  ( $k=1, 2, \dots, n$ )。
- (2) 如果  $w = c_i c_{i+1} \dots c_j$  ( $0 < i < j \leq n$ ) 是一个词, 则节点  $v_{i-1}, v_j$  之间建立有向边  $\langle v_{i-1}, v_j \rangle$ , 边对应的词为  $w$ 。



- (3) 重复步骤(2), 直到没有新路径(词序列)产生。
- (4) 从产生的所有路径中, 选择**路径最短**的(词数最少的)作为最终分词结果。





## 7.3 汉语自动分词基本算法

### ➤ 缺点

- 同样对许多歧义词难以区分；

例 输入字串：他说的确实在理。

输出候选：他/ 说/ 的/ 确实/ 在理/ 。（词个数：5）

他/ 说/ 的确/ 实在/ 理/ 。（词个数：5）

- 字串长度较大时，长度相同的最短路径数可能急剧增加，选择正确的结果难度越大。

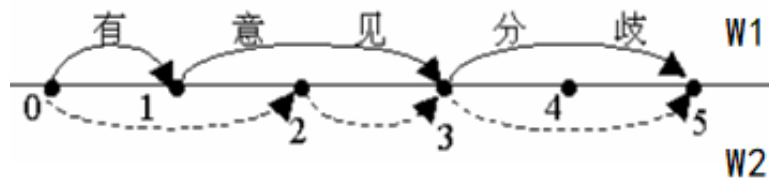
因此，最少分词法一般适合做粗分，以确定下一步切分消歧时的N种可能结果。

## 7.3 汉语自动分词基本算法

### 3. 基于语言模型的分词方法

#### ➤ 方法描述:

设对于待切分的句子 $S$ ,  $W = w_1w_2\dots w_k (1 \leq k \leq n)$  是一种可能的切分。



与最短路径法结合，可实现消歧

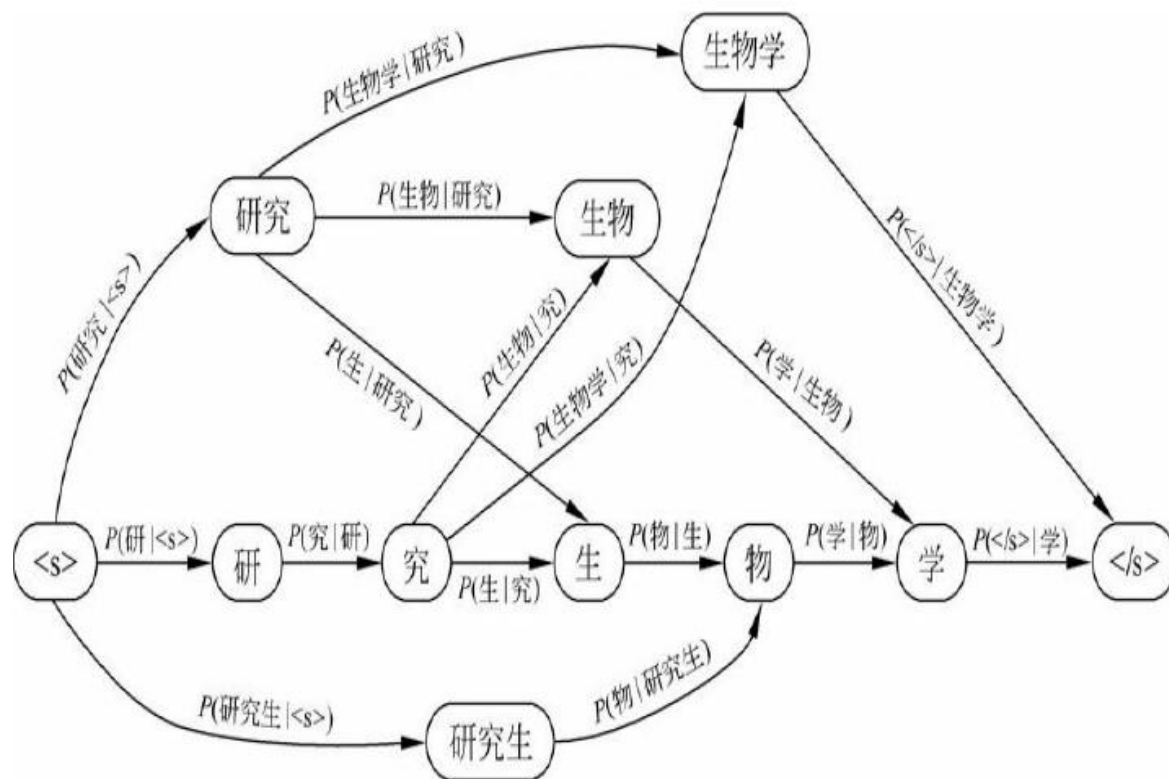
$$\begin{aligned} W^* &= \operatorname{argmax}_W p(W | S) \\ &= \operatorname{argmax}_W p(W) \times p(S | W) \end{aligned}$$

语言模型

## 7.3 汉语自动分词基本算法

### 3. 基于语言模型的分词方法

以“研究生物学”为例，构建基于二元文法的切分词图。





## 7.3 汉语自动分词基本算法

---

### ➤ 优点:

- 在训练语料规模足够大和覆盖领域足够多时，可以获得较高的切分正确率。

### ➤ 缺点:

- 模型性能依赖于训练语料的规模和质量。
- 计算量较大。
- 如果要处理未登录词，需要额外识别模块扩展。



## 7.3 汉语自动分词基本算法

### 4. 由字构词（基于字标注）的分词方法

- 基本思想：将分词过程看作是字的分类问题。该方法认为，每个字在构词中都有一个词位。假定每个字属于4个可能词位之一：

词首(B)、词中(M)、词尾(E)和单独成词(S)

如果将词位视为隐藏状态，则可以利用HMM模型进行分词。

## 7.3 汉语自动分词基本算法

### 4. 由字构词（基于序列标注）的分词方法

如：  
 $B$   $E$        $B$   $E$        $B$   $E$        $S$        $B$   $M$   $E$   
共 同      创 造      美 好      的      新 世 纪

(1) 隐藏状态有4个，即B,M,E,S，观测值为语句中的所有字。

(2) 状态转移概率矩阵

$$A = \begin{array}{c|cccc} & B & M & E & S \\ \hline B & 0.0 & p(M|B) & p(E|B) & 0.0 \\ M & 0.0 & p(M|M) & p(E|M) & 0.0 \\ E & p(B|E) & 0.0 & 0.0 & p(S|E) \\ S & p(B|S) & 0.0 & 0.0 & p(S|S) \end{array}$$

(3) 发射概率矩阵

$$B = \begin{bmatrix} p(v_1|B) & p(v_2|B) & \cdots & p(v_M|B) \\ p(v_1|M) & p(v_2|M) & \cdots & p(v_M|M) \\ p(v_1|E) & p(v_2|E) & \cdots & p(v_M|E) \\ p(v_1|S) & p(v_2|S) & \cdots & p(v_M|S) \end{bmatrix}$$



## 7.3 汉语自动分词基本算法

---

### ➤评价：

该方法的优势在于：它能够平衡地看待词表词和未登录词的识别问题，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。

不用设计未登录词识别模块，因此，大大地简化了分词系统的设计。



## 7.3 汉语自动分词基本算法

### 4. 由字构词（基于序列标注）的分词方法

#### ➤ 基于CRF的分词

构建特征模板：使用当前字的上下文特征辅助标注。

#### ➤ 深度学习

见7.4节

U00:%x[-2,0]

U01:%x[-1,0]

U02:%x[0,0]

U03:%x[1,0]

U04:%x[2,0]

U05:%x[-2,0]/%x[-1,0]/%x[0,0]

U06:%x[1,0]/%x[0,0]/%x[1,0]

U07:%x[0,0]/%x[1,0]/%x[2,0]

U08:%x[-1,0]/%x[0,0]

U09:%x[0,0]/%x[1,0]





## 7.4 未登录词识别



## 7.4 未登录词识别

---

未登录词识别包括：命名实体识别和其它新词识别。

### ◆命名实体(Named Entity, NE)

(专有名词)人名、地名、组织机构名、数字、日期、货币数量。

命名实体识别包括两方面任务：实体的边界，实体的类型。

### ◆其他新词

专业术语、新的普通词汇等。



## 7.4 未登录词识别

---

### ◆关于中文姓名

- 台湾出版的《中国姓氏集》收集姓氏 **5544**个，其中，单姓 **3410**个，复姓 **1990**个，3字姓 **144**个
- 中国目前仍使用的姓氏共 **737**个，其中，单姓 **729**个，复姓 **8**个



## 7.4 未登录词识别

### ◆ 中文姓名识别的难点

- ✧ 姓氏和名字都可以单独用于特指某一人。如:张[高文]
- ✧ 许多姓氏用字和名字用字(词)可以作为普通用字或词被使用。如:

姓氏为普通词: 于(介词), 张(量词), 江(名词)等;

名字为普通词: 建国, 国庆, 胜利, 文革, 计划等,

全名也是普通词, 如: 万里, 温馨, 高山, 高升, 高飞, 周密, 江山, 夏天等。

- ✧ 缺乏可利用的启发标记。

如: (1) 祝贺老总百战百胜。 (2) 林徽因此时已经离开了那里。



## 7.4 未登录词识别

---

### ◆ 中文姓名识别方法- 基于规则和词典的方法

- 以姓氏作为触发信息，从姓氏库和名字库中查找，匹配到潜在的名字。
- 计算潜在姓名的概率估计及相应姓氏的姓名阈值。
- 根据姓名概率评价函数和识别规则对潜在的姓名进行筛选。



## 7.4 未登录词识别

### ➤ 计算概率估计值

设姓名  $Cname = Xm_1m_2$ ，其中  $X$  表示姓， $m_1m_2$  分别表示名字首字和名字尾字。分别用下列公式计算姓氏和名字的使用频率：

$$F(X) = \frac{X \text{ 用作姓氏}}{X \text{ 出现的总次数}}$$

$$F(m_1) = \frac{m_1 \text{ 作为名字首字出现的次数}}{m_1 \text{ 出现的总次数}}$$

$$F(m_2) = \frac{m_2 \text{ 作为名字尾字出现的次数}}{m_2 \text{ 出现的总次数}}$$

## 7.4 未登录词识别

字串 *Cname* 可能为姓名的概率估值:

$$P(Cname) = \begin{cases} F(X) \times F(m_1) \times F(m_2) & \text{复名情况} \\ F(X) \times F(m_2) & \text{单名情况} \end{cases}$$

### ► 确定阈值

姓氏 *X* 构成姓名的最小阈值:

$$T_{\min}(X) = \begin{cases} F(X) \times \text{Min}(F(m_1) \times F(m_2)) & \text{复名情况} \\ F(X) \times \text{Min}(F(m_2)) & \text{单名情况} \end{cases}$$



## 7.4 未登录词识别

---

### ➤设计评估函数

姓名的评价函数：

$$f = -\ln P(Cname)$$

当 $f$ 大于 $\beta_x$ 时，该识别的汉字串确定为中文姓名。 $\beta_x$ 为姓氏 $x$ 从训练语料中得到的阈值。





## 7.4 未登录词识别

---

### ➤ 修饰规则:

如果姓名前是一个数字，或者姓氏与“。”字符的距离小于 **2** 个字符，则否定此姓名。

### ➤ 左界规则:

若潜在姓名前面是一称谓，或一标点符号，或者潜在姓名在句首，或者潜在的姓名的姓氏使用频率为100%，则姓名的左界确定。



## 7.4 未登录词识别

---

### ➤ 右界规则:

若姓名后面是一称谓，或者是一指界动词(如，说，是，指出，认为等)或标点符号，或者潜在的姓名在句尾，或者潜在姓名的尾字使用频率为100%，则姓名的右界确定。



## 7.4 未登录词识别

---

### ◆ 中文地名识别方法

#### ➤ 困难

- 地名数量大，缺乏明确、规范的定义。

《中华人民共和国地名录》收集88026个，不包括大部分街道、胡同、村庄等。

- 真实语料中地名出现情况复杂。

如地名简称、地名用词与其他普通词冲突（如走马）、地名是其他专用名词的一部分(如合川桃片)，地名长度不一等。



## 7.4 未登录词识别

---

### ➤地名识别资源

- 地名资源知识库

- 一 地名库、地名首字库、地名尾字库、地名中间字库

- 识别规则库

- 一 筛选规则、确认规则、否定规则



## 7.4 未登录词识别

---

### ➤ 基本识别方法

- 通过地名首、尾、中间字库，选出可能的地名
- 利用统计模型计算地名概率
- 通过训练语料选取阈值
- 地名初筛选
- 寻找可用的上下文信息
- 利用规则进一步确定地名



## 7.4 未登录词识别

---

### ◆ 中文机构名称的识别

#### ➤ 中文机构名称的构成（规则集）

- 词法角度：偏正式(修饰格式)的复合词  
{名词|形容词|数量词|动词} + 名词
- 句法角度：“定语 + 名词性中心语”型的名词短语(定名型短语)
- 中心语：机构称呼词，如：大学，学院，研究所，学会，公司等。



## 7.4 未登录词识别

---

### ➤ 机构名称识别方法

- 找到一机构称呼词
- 根据相应规则，往前逐个检查修饰词的合法性，直到发现非法词。
  - 如果修饰词 同 机构称呼词 构成一个合法的机构名称，则记录该机构名称。
  - 利用统计模型计算其概率。

## 7.4 未登录词识别

### ➤命名实体识别统一方法

通过CRF模型，将命名实体识别转化为序列标注问题。

假定标注类型为：人名(PER)、地名(LOC)、机构名(ORG)，则标注类别设计如下：

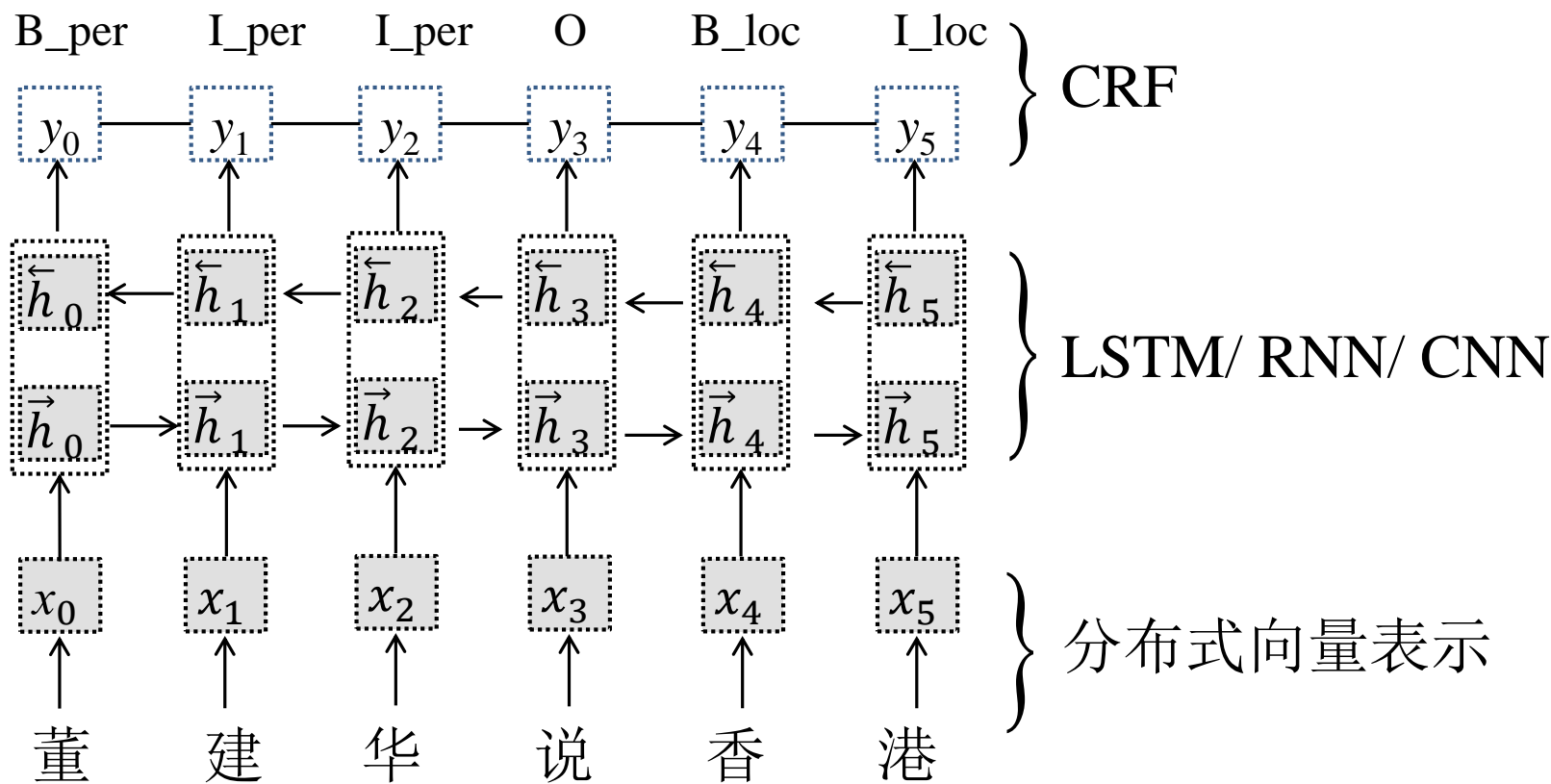
- (1) 每类实体都以B-开始，如地名的开始字由B-LOC表示；
- (2) 如果一个实体包括多个字，则后面的字标为I-XXX，如I-LOC。
- (3) 如果某个字不属于任何实体名，则标为O。

1.	谈	v	O
2.	到	v	O
3.	亚	ns	B_LOC
4.	洲	ns	I_LOC
5.	金	n	O
6.	融	n	O
7.	风	n	O
8.	波	n	O
9.	,	x	O
10.	董	nr	B_PER
11.	建	nr	I_PER
12.	华	nr	I_PER
13.	说	v	O
14.	:	x	O
15.	“	x	O
16.	香	ns	B_LOC
17.	港	ns	I_LOC
18.	将	d	O
19.	坚	i	O
20.	定	i	O
21.	不	i	O
22.	移	i	O
23.	地	uv	O



## 7.4 未登录词识别

### ◆ 基于神经网络的命名实体识别方法



**NER结果:** 董建华[人名]/ 说/ 香港[地名]



## 7.4 未登录词识别

- 除人名、地名、组织机构外，还有很多其它的命名体类别。
- 如医疗领域中，常见的命名体有**部位、疾病名、症状、药物、检查、手术**等。

患者4月前发现皮肤、巩膜黄染，伴食欲上降，晚餐后明显，时有阵发性腹痛、恶心，无腹泻、呕吐，时有胸闷、憋气、头晕，无头痛，无视物旋转，无发热、咳嗽，无胸痛、喘憋，大便颜色较前变浅。于我科住院，行**腹部增强MRI+MRCP**：胆总管末端占位，胆管癌可能性大，右肾下腺结节，考虑**腺瘤，右肾囊肿，两侧胸腔积液**。请肝胆外科会诊，认为患者无手术指征，于2017-4-24全麻上行**ERCP+胆总管金属支架植入**，术后患者恢复良好。并给以抗感染、抗肿瘤、免疫调节、平稳降糖、降压等对症治疗，全身及巩膜黄染逐渐减轻，化验指标逐渐改善。



---

## 7.5 词性标注概述



## 7.5 词性标注

---

词性标注是在已经切分好的文本中，给每一个词标注其所属的词类，例如动词、名词、代词、形容词。

词性标注对后续的句子理解有重要的作用。

◆ 面临的问题：词性兼类歧义。

(1) 形同音不同，如：“好(hao3, 形容词)、好(hao4, 动词)”

这个人什么都好，就是好酗酒。

(2) 同形、同音，但意义毫不相干，如：“会(会议, 名词)、会(能够、动词)”

每次他都会在会上制造点新闻。



## 7.5 词性标注

---

(3) 具有典型意义的兼类词，如：“典型(名词或形容词)”、“教育(名词或动词)”

用那种方式教育孩子，简直是对教育事业的侮辱。

(4) 上述情况的组合，如：“行(xing2, 动词/形容词; hang2, 名词/量词)”

每当他走过那行白杨树时，他都感觉好像每一棵树都在向他行注目礼。



## 7.5 词性标注

---

### ➤ **UPenn Treebank** 词性标注训练集

- **33 类**

- **NN** 名词、**NR** 专业名词、**NT** 时间名词、**VA** 可做谓语的形容词、**VC** “是”、**VE** “有”作为主要动词、**VV** 其他动词、**AD** 副词、**M** 量词，等等。



## 7.5 词性标注

---

### ➤ 北大计算语言学研究 词性标注训练集

- **26**个基本词类代码，**74**个扩充代码，标记集中共有**106**个代码。

名词(**n**)、时间词(**t**)、处所词(**s**)、方位词(**f**)、数词(**m**)、量词(**q**)、区别词(**b**)、代词(**r**)、动词(**v**)、形容词(**a**)、状态词(**z**)、副词(**d**)、介词(**p**)、连词(**c**)、助词(**u**)、语气词(**y**)、叹词(**e**)、拟声词(**o**)、成语(**i**)、习用语(**l**)、简称(**j**)、前接成分(**h**)、后接成分(**k**)、语素(**g**)、非语素字(**x**)、标点符号(**w**)。

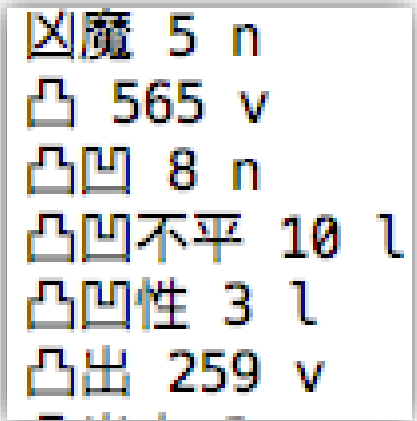
## 7.5 词性标注

### ➤ 词性标注方法

#### (1) 基于字符串匹配的字典查找算法

先对语句进行分词，然后从字典中查找每个词语的词性，对其进行标注即可。

这种方法比较简单，但是不能解决一词多词性的问题。



凸魔	5	n
凸	565	v
凸凹	8	n
凸凹不平	10	l
凸凹性	3	l
凸出	259	v





## 7.5 词性标注

---

### ➤词性标注方法

#### (2) 基于统计的词性标注算法

如Jieba分词就综合了两种算法，对于分词后识别出来的词语，直接从字典中查找其词性。

而对于未登录词，则采用隐马尔科夫模型和viterbi算法来识别。

观测序列即分词后的语句，隐藏序列为词性标注序列。



## 7.6 分词与词性标注系统

---

◆ 目前公开的分词与词性标注系统:

● <http://ictclas.nlpir.org/nlpir/>

中科院计算所(ICTCLASS)

● <https://gitee.com/tekin/fnlp/>

复旦大学

● <http://nlp.stanford.edu/software/tagger.shtml>

Stanford University

---



谢谢!