

# 第8章 词语切分与词性标注

宗成庆

中国科学院自动化研究所

[cqzong@nlpr.ia.ac.cn](mailto:cqzong@nlpr.ia.ac.cn)

# 本章内容

---



1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题
8. 附录

# 1. 概述

- ◆ 词是自然语言中能够独立运用的最小单位，是自然语言处理的基本单位。
- ◆ 不同的语言在词法层面需要完成不同的分析任务
  - 曲折语(如英语、德语、俄语等): 用词的形态变化表示语法关系，一个形态成分可以表示若干种不同的语法意义，词根和词干与语词的附加成分结合紧密。曲折语词法分析的任务就是词的形态分析(形态还原)(morphology analysis)。
  - 分析语(孤立语)(如汉语、越南语、苗语): 词语切分。
  - 黏着语(如日语、韩语、土耳其语等): 词语切分+形态还原。

本章主要关注汉语词语的切分、子词压缩和词性标注方法。

# 1. 概述

- ◆ 词性或称词类(Part-of-Speech, POS)是词汇最重要的特性，是语言中词的**语法分类**，具有相同句法功能、能够出现在同样的组合位置中的词，聚合在一起所形成的范畴。词类连接词汇到句法的桥梁。

如在汉语中，词类分为两大类：实词(content words)和虚词(functional words)，实词包括体词、谓词，体词又包括名词、代词等，谓词包括动词、形容词等。

词性标注的任务是让系统自动对词汇标注词性标记。

# 1. 概述

## ◆ 汉语自动分词和词性标注的重要性

- 词语切分是句子结构分析的基础
- 词语的分析具有广泛的应用，如词频统计，词典编纂，文章风格研究，文献处理，文本校对，简繁体转换等
- 即使在数据驱动的自然语言处理中，包括统计学习方法和神经网络方法，通常情况下基于词（具有较好的切分准确率）建立的模型性能优于以字或子词建立的模型
- 词性是反映句法结构信息的重要特征
- 词性在众多NLP任务中（如文本分类、情感分类、自动文摘等）具有重要作用

# 本章内容

1. 概述

 2. 汉语分词要点

3. 汉语分词方法

4. 命名实体识别

5. 子词压缩

6. 词性标注

7. 习题

8. 附录

## 2. 汉语分词要点

### ◆ 汉语自动分词中的主要问题

- 汉语分词规范问题（《信息处理用限定汉语分词规范（GB13715）》）

— 汉语中什么是词？两个不清的界限：

(1) 单字词与词素，如：新华社25日[讯](#)

(2) 词与短语，如：花草，湖边，房顶，鸭蛋，小鸟，担水，一层，翻过？

## 2. 汉语分词要点

### ● 歧义切分字段处理

#### (1) 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

中/ 国人/ 为了/ 实现/ 自己/ 的/ 梦想

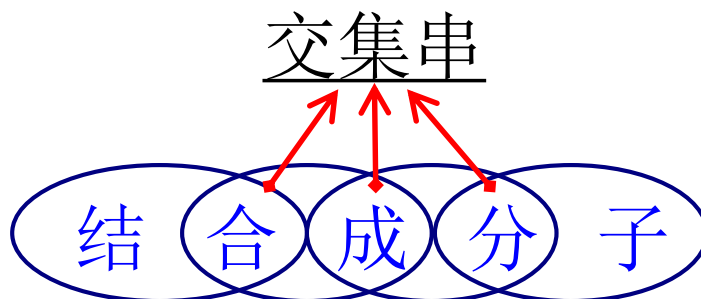
例如：“大学生”、“研究生物”、“从小学起”、“为人民工作”、“中国产品质量”、“部分居民生活水平”等等

➤ **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。



## 2. 汉语分词要点

例如：



“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。

类似地，“为人民工作”中的公共交集字为：{人，民，工}，歧义字段的链长为 3；“中国产品质量”中的交集字为：{国，产，品，质}，歧义字段的链长为 4；“部分居民生活水平”中的交集字为：{分，居，民，生，活，水}，链长为 6。

## 2. 汉语分词要点

### (2) 组合型歧义

门把手弄坏了。

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。

## 2. 汉语分词要点

梁南元 曾于1987年对一个含有48,092字的自然科学、社会科学领域的样本进行了统计，结果交集型切分歧义有**518**个，多义组合型切分歧义有**42**个。据此推断，中文文本中切分歧义的出现频度约为**1.2次/100字**，交集型切分歧义与多义组合型切分歧义的出现比例约为**12:1**。

## 2. 汉语分词要点

### ● 未登录词的识别

(1)人名、地名、组织机构名等命名实体，例如：

盛中国，张建国，李爱国，蔡国庆；

高升，高山，夏天，温馨，温泉，武夷山，时光，程序；

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，内卷，新冠

## 2. 汉语分词要点

我们的统计结果：

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				

从互联网上随机摘取了418个句子，共含11,739个词，19,777个汉字（平均每个句长约为28个词，每个词约含1.68个汉字）。

## 2. 汉语分词要点

### ◆ 汉语自动分词的基本原则

- 合并原则1：语义上无法由组合成分直接相加而得到的字串应该合并为一个分词单位。

例如：不管三七二十一(成语)，或多或少(副词片语)，十三点(定量结构)，六月(定名结构)，谈谈(重叠结构，表示尝试)，辛辛苦苦(重叠结构，加强程度)，进出口(合并结构)

- 合并原则2：语类无法由组合成分直接得到的字串应该合并为一个分词单位。

(a)字串的语法功能不符合组合规律，如：好吃，好喝，好听，好看等

(b)字串的内部结构不符合语法规律，如：游水等

## 2. 汉语分词要点

### ◆ 汉语自动分词的辅助原则

操作性原则，富于弹性，不是绝对的。

- 切分原则1：有明显分隔符标记的应该切分之。

分隔标记指标点符号或一个词。如：

上、下课 → 上/ 下课

洗了个澡 → 洗/ 了/ 个/ 澡

## 2. 汉语分词要点

- 切分原则2：结构复杂、合并起来过于冗长的词尽量切分。

(1) 词组带接尾词，如：太空/ 计划/ 室、塑料/ 制品/ 业

(2) 动词带双音节结果补语，如：看/ 清楚、讨论/ 完毕

(3) 复杂结构：自来水/公司；中文/分词/规范/研究/计划

(4) 正反问句：喜欢/ 不/ 喜欢、参加/ 不/ 参加

(5) 动宾结构、述补结构的动词带词缀时。如：写信/ 给；  
取出/ 给；穿衣/ 去

(6) 词组或句子的专名，多见于书面语，戏剧名、 歌曲名。

如：鲸鱼/的/生/与/死；那/一/年/我们/都/很/酷

(7) 专名带普通名词。如：胡/ 先生、京沪/ 铁路



## 2. 汉语分词要点

- **合并原则1**：附着性语(词)素与前后词合并为一个单位。

例如：“吝”是一个附着语素，“不吝”、“吝于”等合并成一个词；“员”：检查员、邮递员、技术员等；“化”：现代化、合理化、多变化、年轻化、民营化等。

- **合并原则2**：使用频率高或共现率高的字串尽量合并。

如：“进出”、“收放”（**动词并列**）；“大笑”、“改称”（**动词偏正**）；“关门”、“洗衣”、“卸货”（**动宾结构**）；“春夏秋冬”、“轻重缓急”、“男女”（**并列结构**）；“象牙”（**名词偏正**）；“暂不”、“毫不”、“不再”、“早已”（**副词并列**）等。

## 2. 汉语分词要点

- 合并原则4：双音节加单音节的偏正式名词尽量合并。

如：“线、权、车、点”等所构成的偏正式名词：“国际线、分数线、贫困线”、“领导权、发言权、知情权”、“垃圾车、交通车、午餐车”、“立足点、共同点、着眼点”等。

- 合并原则5：双音节结构的偏正式动词应尽量合并。

这条原则只适合于少数偏正式的动词，如：“紧追其后”、“组建完成”等，不适合动宾及主谓式复合动词。

## 2. 汉语分词要点

### ◆ 分词结果测试

- 封闭测试 vs. 开放测试
- 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等

## 2. 汉语分词要点

### ◆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所有输出结果的比例。假设系统输出 $N$ 个, 其中, 正确的结果为 $n$ 个, 那么,

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

- 召回率(找回率) (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。

## 2. 汉语分词要点

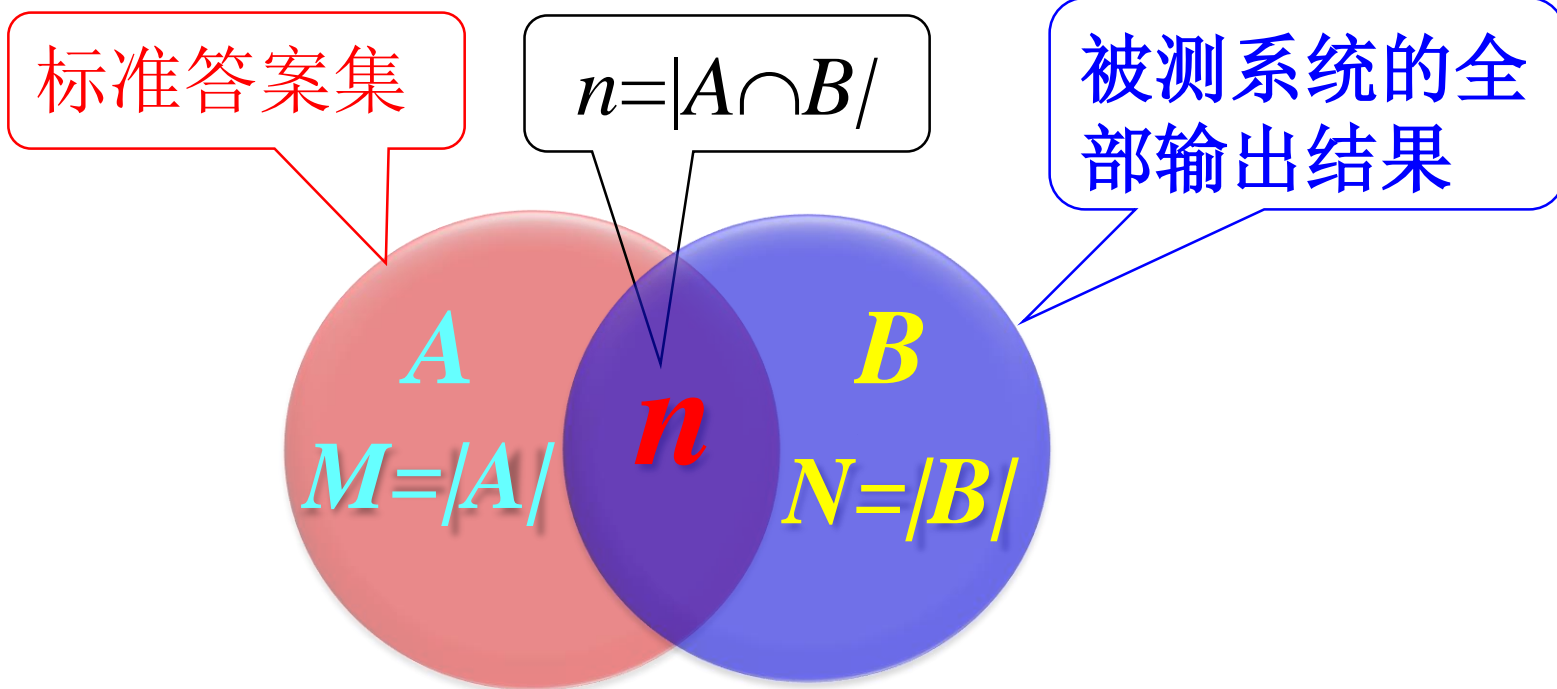
- F-测度值(F-Measure): 正确率与找回率的综合值。计算公式为:

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

一般地, 取  $\beta = 1$ , 即:

$$F1 = \frac{2 \times P \times R}{P + R}$$

## 2. 汉语分词要点



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

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

## 2. 汉语分词要点

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：


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

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

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

# 本章内容

---

1. 概述
2. 汉语分词要点
-  3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题
8. 附录



# 3. 汉语分词方法

---

- ◆ 有词典切分 vs. 无词典切分
- ◆ 基于规则的方法 vs. 基于统计的方法

# 3. 汉语分词方法

## ① 最大匹配法(Maximum Matching, MM)

是一种有词典的切分方法，也称机械切分方法。

按照切分方向分为：

- 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)

### ● 基本思路：

给定字串  $S = c_1 c_2 \dots c_n$ ，某一词  $w_i = c_1 c_2 \dots c_m$ ， $m$  为词典中最长词的字数。假设  $m = 7$ 。

# 3. 汉语分词方法

输入字符串: 他是研究生物化学的一位科学家。

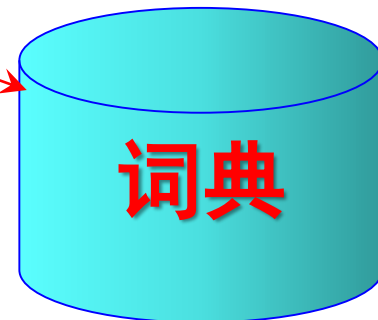
切分过程:

↑ — 7 — → |

| — 6 — → |

| — 5 — → |

...



他/ 是研究生物化学的一位科学家。

↑ — 7 — → |

...

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。

... .. ← 7 — ↑

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

# 3. 汉语分词方法

## ● FMM 算法描述

- (1) 令  $i=0$ , 当前指针  $p_i$  指向输入字符串初始位置, 执行以下操作:
- (2) 计算当前指针  $p_i$  到字符串末端的字数  $n$ , 如果  $n=1$ , 转(4), 结束算法。  
否则, 令  $m$ =词典中最长单词的字数, 如果  $n<m$ , 令  $m=n$ ;
- (3) 从当前  $p_i$  起取  $m$  个汉字作为词  $w_i$ , 判断:
  - (a) 如果  $w_i$  是词典中的词, 则在  $w_i$  后添加一个切分标志, 转(c);
  - (b) 如果  $w_i$  不是词典中的词且  $w_i$  的长度大于1, 将  $w_i$  从右端去掉一个字, 转(a)步; 否则( $w_i$  的长度等于1), 则在  $w_i$  后添加一个切分标志, 将  $w_i$  作为单字词添加到词典中, 执行 (c)步;
  - (c) 根据  $w_i$  的长度修改指针  $p_i$  的位置, 如果  $p_i$  指向字符串末端, 转(4), 否则,  $i=i+1$ , 返回 (2);
- (4) 输出切分结果, 结束分词程序。

# 3. 汉语分词方法

## ● 方法评价

### ➤ 优点：

- 程序简单易行，开发周期短；
- 仅需要很少的语言资源（词表），不需要任何词法、句法、语义资源。

### ➤ 弱点：

- 歧义消解的能力差；
- 切分正确率不高，一般在95%左右。

# 3. 汉语分词方法

## ② 基于语言模型的分词方法

无词典切分

### ● 基本思路

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

$$W^* = \arg \max_W p(W | S)$$

$$= \arg \max_W p(W) \times p(S | W)$$

语言模型

生成模型

详见第4章举例

# 3. 汉语分词方法

## ● 方法评价

### ➤ 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。

### ➤ 弱点:

- 模型性能较多地依赖于训练语料的规模和质量，训练语料的规模和覆盖领域不好把握；
- 计算量较大。

# 3. 汉语分词方法

## ③ 由字构词的分词方法(Character-based tagging) (或称“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，每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。假定每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)，那么，每个字归属一特定的词位。

该方法最早由N. Xue (薛念文) 和 S. Converse 提出，首篇论文发表在2002年第一届ACL SIGHAN (汉语特别兴趣小组)组织的汉语分词评测研讨会上[Xue and Converse, 2002]。

**条件随机场(CRFs)是广泛使用的序列标注模型。**



# 3. 汉语分词方法

## ● 条件随机场的提出

在NLP和图像处理中有一类问题是进行序列标注和结构划分，而 $n$ -gram和HMM都是利用当前时刻 $t$ 之前已经发生的事件信息。J. Lafferty 等人于2001年提出了条件随机场 (conditional random fields, CRFs)这一概率化结构模型。

## ➤ 基本思想

给定观察序列  $X$ ，输出标识序列  $Y$ ，通过计算  $P(Y|X)$  求解最优标注序列。

# 3. 汉语分词方法

## ➤ 定义

设  $G=(V, E)$  为一个无向图,  $V$  为结点集合,  $E$  为无向边的集合,  $Y = \{ Y_v | v \in V \}$ , 即  $V$  中每个结点对应于一个随机变量  $Y_v$ , 其取值范围为可能的标记集合  $\{y\}$ 。如果以观察序列  $X$  为条件, 每个随机变量  $Y_v$  都满足以下马尔可夫特性:

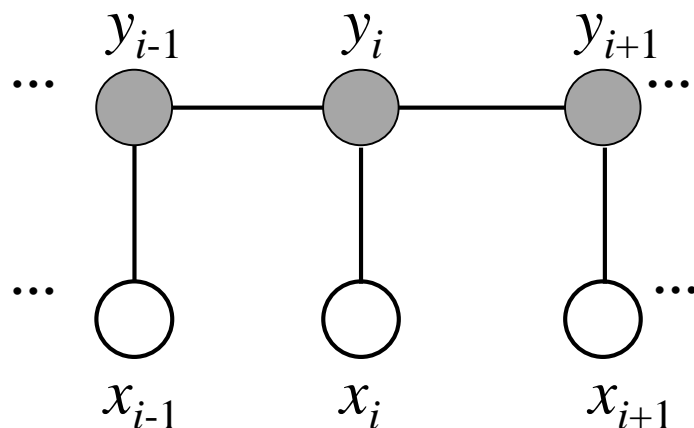
$$p(Y_v / X, Y_w, w \neq v) = p(Y_v / X, Y_w, w \sim v) \quad \dots (8-1)$$

其中,  $w \sim v$  表示两个结点在图中是邻近结点。那么,  $(X, Y)$  为一个条件随机场。

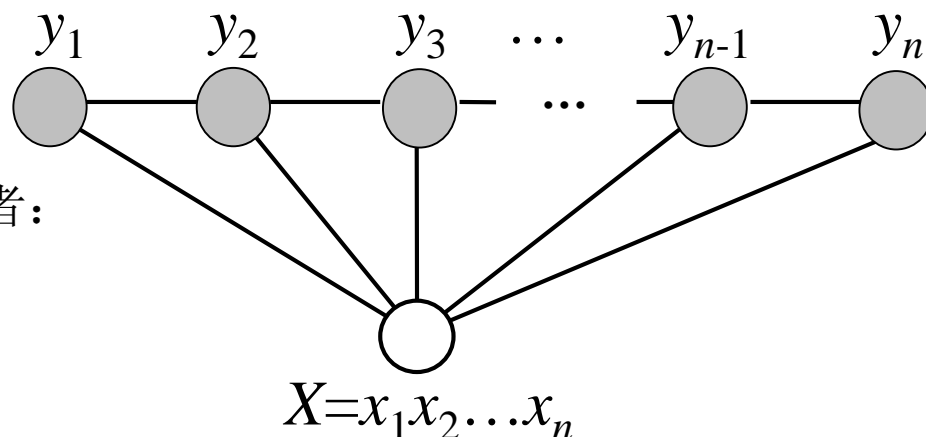
# 3. 汉语分词方法

图示：

$$p(Y_v / X, Y_w, w \neq v) = p(Y_v / X, Y_w, w \sim v)$$



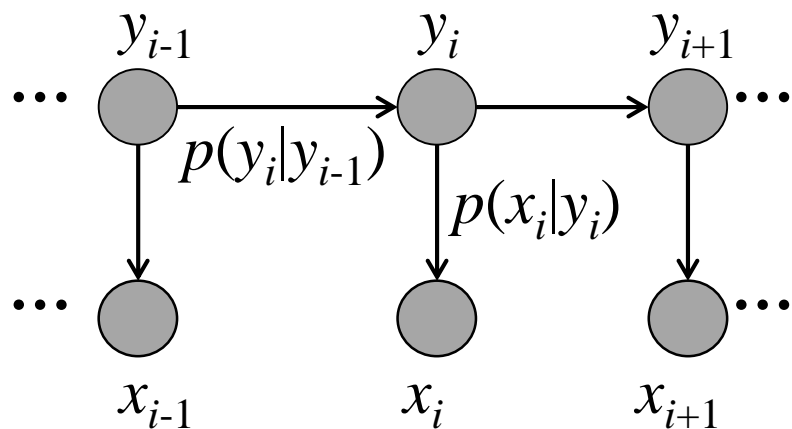
或者：



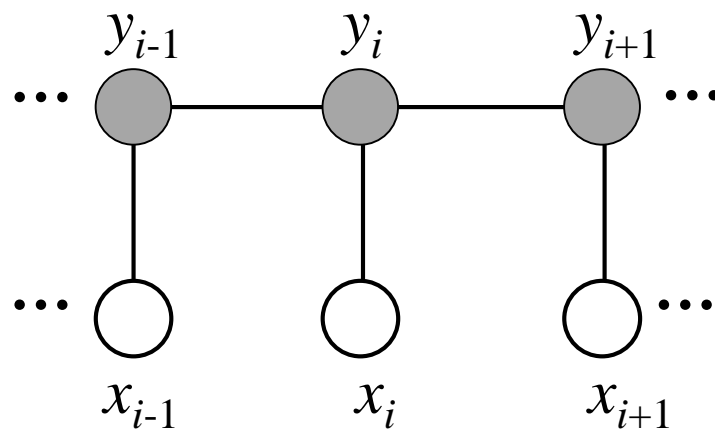
序列标注问题可以建模为简单的链式结构图，结点对应标记序列 $Y$ 中的元素。理论上，只要在标记序列中描述一定的条件独立性， $G$ 的图结构可以任意的。

# 3. 汉语分词方法

## ✧ CRFs 与 HMM 的对比



HMM



CRFs

CRFs 中的节点  $x$  并不是由模型生成的。

# 3. 汉语分词方法

在链式CRFs中, 给定观察序列 $x$ 时, 某个特定标记序列 $y$ 的概率可以定义为:

$$p(y | x) = \frac{1}{Z(x)} \exp \left( \sum_{i,j} \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_{i,k} \mu_k s_k(y_i, x, i) \right) \dots (8-2)$$

其中,  $t_j(y_{i-1}, y_i, x, i)$ 是转移函数, 表示对于观察序列 $x$ 的标注序列在 $i-1$ 和 $i$ 位置上标记的转移概率, 通常称作二元特征。

$s_k(y_i, x, i)$ 是状态函数, 表示观察序列 $x$ 在 $i$ 位置的标记概率, 通常称作一元特征。

$\lambda_j$ 和 $\mu_k$ 分别是  $t_j$ 和 $s_k$  的权重, 需要从训练样本中估计出。

$Z(x)$ 为归一化因子:

$$Z(x) = \sum_y \exp \left( \sum_{i,j} \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_{i,k} \mu_k s_k(y_i, x, i) \right) \dots (8-3)$$

# 3. 汉语分词方法

定义一组关于观察序列的 $\{0, 1\}$  二值特征  $b(x, i)$ , 表示训练样本中某些特征的分布, 如

$$b(x, i) = \begin{cases} 1 & \text{如果 } x \text{ 的 } i \text{ 位置为某个特定的词} \\ 0 & \text{否则} \end{cases}$$

转移函数可以定义为如下形式:

$$t_j(y_{i-1}, y_i, x, i) = \begin{cases} b(x, i) & \text{如果 } y_{i-1} \text{ 和 } y_i \text{ 满足某种搭配条件} \\ 0 & \text{否则} \end{cases}$$

也可以把状态函数写成如下形式:

$$s_k(y_i, x, i) = s_k(y_{i-1}, y_i, x, i)$$

# 3. 汉语分词方法

一元函数和二元函数可以统一为一种形式： $f(y_{i-1}, y_i, x, i)$ 。  
 对于一个长度为 $n$ 的观察序列，假设有 $K_1$ 个转移函数， $K_2$ 个状态函数，共计有  $K=K_1+K_2$  个特征函数。那么，

$$f_j(y_{i-1}, y_i, x, i) = \begin{cases} t_j(y_{i-1}, y_i, x, i) & j = 1, 2, \dots, K_1 \\ s_k(y_i, x, i) & j = K_1 + k, k = 1, 2, \dots, K_2 \end{cases}$$

对于一个长度为 $n$ 的序列，某个特征函数 $f_j(\bullet)$ （泛指转移函数或状态函数）可能在多个位置上都存在，因此在该序列中应对所有成立的情况求和：

$$f_j(y, x) = \sum_{i=1}^n f_j(y_{i-1}, y_i, x, i), \quad j = 1, 2, \dots, K$$

# 3. 汉语分词方法

例如:

start	y:	S	S	S	B	E	S	S	B	M	E	S	S	S	S	S
	x:	他	写	完	作	业	后	把	乒	乓	球	拍	子	卖	了	。
	i =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

$n = 15$

$$t_1(y_0, y_1, x, 1) = \begin{cases} 1 & y_0 = \text{start} \text{ 且 } y_1 = S \\ 0 & \text{否则} \end{cases}$$

简化后:  $t_1(\text{start}, S, x, 1)$

$$s_1(y_1, x, 1) = \begin{cases} 1 & \text{当 } y_1 = S \text{ 且 } x_1 = \text{他} \\ 0 & \text{否则} \end{cases}$$

简化后:  $s_1(S, x, 1)$

$$t_2(S, S, x, i), i = 2, 3, 7, 12, \dots, 15$$

$$t_3(S, B, x, i), i = 4, 8$$

$$t_4(B, E, x, 5)$$

$$t_5(E, S, x, i), i = 6, 11$$

$$t_6(B, M, x, 9)$$

$$t_7(M, E, x, 10)$$

$K_1=7$

$$K = K_1 + K_2 = 11$$

$K_2=4$

$$s_1(S, x, i), i = 1, 2, 3, 6, 7, 11, \dots, 15$$

$$s_2(B, x, i), i = 4, 8$$

$$s_3(E, x, i), i = 5, 10$$

$$s_4(M, x, 9)$$



# 3. 汉语分词方法

采用同样的方式，用权重 $w_j$ 表示 $f_j(\bullet)$ 的权重：

$$w_j = \begin{cases} \lambda_j & j = 1, 2, \dots, K_1 \\ \mu_k & j = K_1 + k, k = 1, 2, \dots, K_2 \end{cases}$$

所以，(8-2)式和(8-3)式被简化成：

$$p(y | x) = \frac{1}{Z(x)} \exp \left( \sum_{j=1}^K w_j f_j(y, x) \right) \quad \dots(8-4)$$

$$Z(x) = \sum_y \exp \left( \sum_{j=1}^K w_j f_j(y, x) \right) \quad \dots(8-5)$$

# 3. 汉语分词方法

将整个序列中的所有特征用向量  $F(y, x)$  表示，即：

$$F(y, x) = (f_1(y, x), f_2(y, x), \dots, f_K(y, x))^T$$

用  $\mathbf{w}$  表示权重向量：  $\mathbf{w} = (w_1, w_2, \dots, w_K)^T$ ,

那么，(8-4)式和(8-5)式被进一步简化为：

$$p(y | x) = \frac{1}{Z(x)} \exp(\mathbf{w} \cdot F(y, x)) \quad \dots(8-6)$$

或者

$$p(y | x, \mathbf{w}) = \frac{1}{Z(x)} \exp(\mathbf{w} \cdot F(y, x)) \quad \dots(8-7)$$

$$Z(x) = \sum_y \exp(\mathbf{w} \cdot F(y, x)) \quad \dots(8-8)$$

# 3. 汉语分词方法

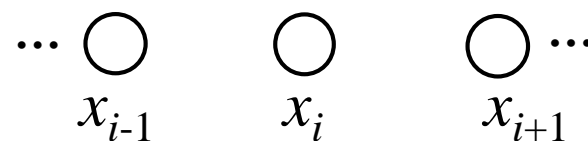
## ✧ CRFs 与 ME 模型的对比

### \*相同点:

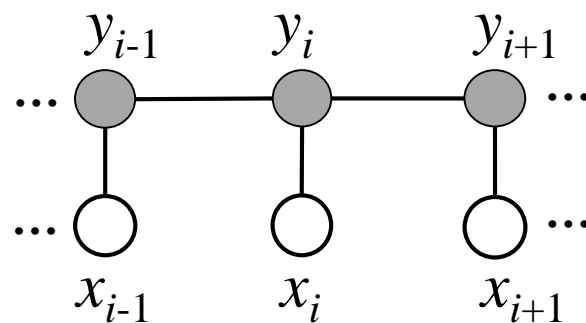
- 都是通过特征函数计算概率，模型形式也一样；
- 可以采用同样的特征选择方法和参数训练方法。

### \*不同点:

- 基于 ME 的分类器对给定输入  $x$  的整体（作为一个单位）或局部点进行分类；
- CRFs 模型是对给定输入  $x$  进行序列标注，最终求解的是全局最优标注序列  $y$ 。



(a) ME



(b) CRFs

# 3. 汉语分词方法

## ✧ CRFs 与 Softmax 的对比

$$p(y | x) = \frac{\exp(\mathbf{w} \cdot F(y, x))}{\sum_y \exp(\mathbf{w} \cdot F(y, x))}$$

CRFs

$$p(y = c | \mathbf{x}) = \frac{\exp(\mathbf{w}_c^T \mathbf{x})}{\sum_{c'=1}^C \exp(\mathbf{w}_{c'}^T \mathbf{x})}$$

Softmax( $\mathbf{w}_c^T \mathbf{x}$ )

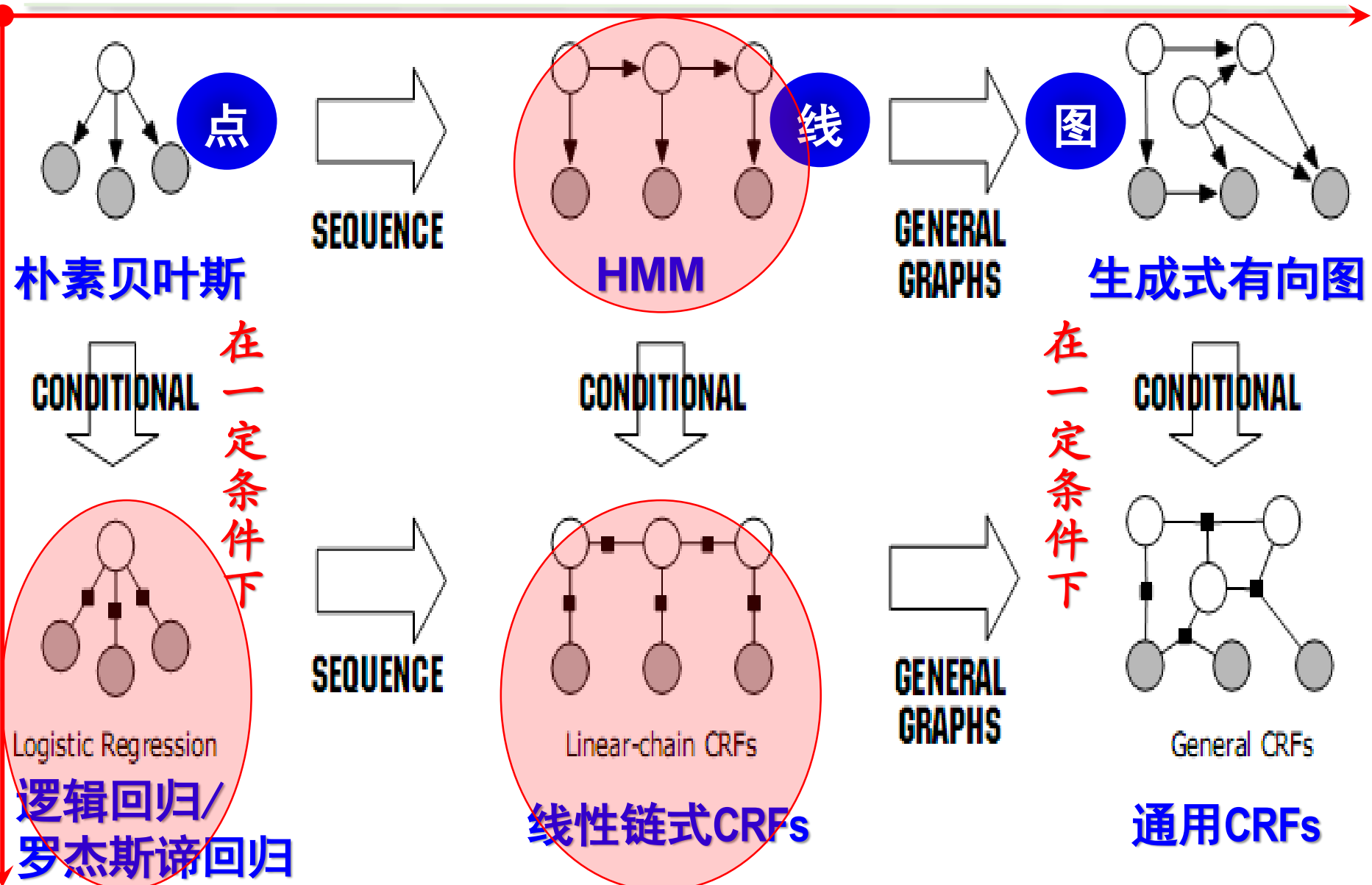
- Softmax 是多项(Multinomial)或多类(Multi-Class)的 Logistic 回归, 可以看作是一种条件最大熵模型。假设类别标签  $y \in \{1, 2, \dots, C\}$ , 给定一个样本  $\mathbf{x}$ ,  $\mathbf{w}_c$  是第  $c$  类的权重向量,

决策函数:

$$\hat{y} = \arg \max_{c=1}^C p(y = c | \mathbf{x}) = \arg \max_{c=1}^C \mathbf{w}_c^T \mathbf{x}$$

$$\mathbf{w}_c = \begin{bmatrix} w_{c1} \\ w_{c2} \\ \vdots \\ w_{cn} \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

# 3. 汉语分词方法



# 3. 汉语分词方法

## ➤ CRFs 模型实现

需要解决的三个问题：

- ①特征选取
- ②参数训练
- ③解码

# 3. 汉语分词方法

## ● 应用举例

例如：乒乓球拍卖完了。

(1) 乒乓球/ 拍/ 卖/ 完/ 了/ 。/

(2) 乒乓球/ 拍卖/ 完/ 了/ 。/

(3) 乒/B 乓/M 球/E 拍/S 完/S 了/S 。/S

在字位标注过程中，对所有的字根据预定义的特征进行字位特征学习，获得一个概率模型，然后在待切分字符串上，根据字与字之间的结合紧密程度，得到一个词位的分类结果，最后根据字位定义直接获得最终的分词结果。

# 3. 汉语分词方法

乒/B 乓/M 球/E 拍/S 卖 完 了 。



B, E, M, S ?

- 当前字的前后  $n$  个字
- 当前字左边字的标记
- 当前字在词中的位置
- .....



# 3. 汉语分词方法

## ①特征选取

- 一元特征（状态函数）：当前字、当前字的前一个字、当前字的后一个字

$$s_1(y_i, x, i) = \begin{cases} 1 & \text{如果当前字是“拍”，当前字的标记} y_i \text{是S} \\ 0 & \text{否则} \end{cases}$$

$$s_2(y_i, x, i) = \begin{cases} 1 & \text{如果当前字是“拍”，当前字的标记} y_i \text{是E} \\ 0 & \text{否则} \end{cases}$$

.....

乒/B 乒/M 球/E 拍/S 卖/? 完了。

# 3. 汉语分词方法

## ➤ 二元特征（转移函数）：

$$t_1(y_{i-1}, y_i, x, i) = \begin{cases} 1 & \text{如果前一个字的标记} y_{i-1} \text{是B, 当前字的标记} y_i \text{是M} \\ 0 & \text{否则} \end{cases}$$

$$t_2(y_{i-1}, y_i, x, i) = \begin{cases} 1 & \text{如果前一个字的标记} y_{i-1} \text{是M, 当前字的标记} y_i \text{是M} \\ 0 & \text{否则} \end{cases}$$

.....

乒/B 乒/M 球/E 拍/S 卖/? 完了。

# 3. 汉语分词方法

## ②参数训练

通过训练语料估计特征权重 $w_j$ ，使其在给定一个观察序列 $x$ 的条件下，找到一个最有可能的标记序列 $y$ ，即条件概率 $p(y|x)$ 最大。

条件概率可由 (8-7)(8-8)式计算得出：

$$p(y | x, \mathbf{w}) = \frac{1}{Z(x)} \exp(\mathbf{w} \cdot F(y, x))$$

$$Z(x) = \sum_y \exp(\mathbf{w} \cdot F(y, x))$$

# 3. 汉语分词方法

训练特征权重 $\mathbf{w}$ ，可以采用改进的广义迭代缩放算法(GIS)，也可以采用梯度下降法或拟牛顿法(quasi-Newton method)。假设采用梯度下降法，需要计算模型的损失和梯度，由梯度更新 $\mathbf{w}$ ，直到 $\mathbf{w}$ 收敛。损失函数可定义为负对数似然函数：

$$L(\mathbf{w}) = -\sum_{x,y} \log p(y|x, \mathbf{w}) + \frac{\varepsilon}{2} \|\mathbf{w}\|^2 \quad (\varepsilon \text{取值范围: } 10^{-6} \sim 10^{-3})$$

损失函数的梯度为：

$$\frac{\partial L(\mathbf{w})}{\partial w_j} = \sum_{x,y} \left( \frac{\partial \log Z(x)}{\partial \lambda_j} - F(y, x) \right) + \varepsilon w_j \quad \dots(8-9)$$

$$w_{j+1} = w_j - \ell \frac{\partial L(\mathbf{w})}{\partial w_j} \quad (\ell \text{为学习率, 经验值, 可设为0.1等。})$$

# 3. 汉语分词方法

## ③解码

条件随机场解码的过程就是根据模型求解的过程，通常由维特比(Viterbi)搜索算法完成，通过动态规划，局部路径成为整体最优路径的一部分。

# 3. 汉语分词方法

例句：乒 乓 球 拍 卖 完 了

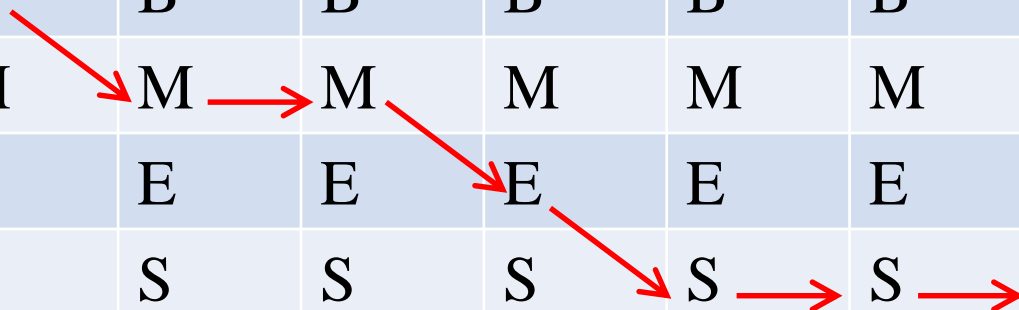
维特比算法就是在下面由标记组成的矩阵中搜索一条最优的路径。

乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

分词结果：乒/B 乓/M 球/M 拍/E 卖/S 完/S 了/S

# 3. 汉语分词方法

乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S



到达每个标记的分数由以下三部分组成：

- **标记的一元特征权重  $W$** ：分别用  $W_1^B$  表示第一个字被标记为B的权重， $W_1^S$  表示第一个字被标记为S的权重，等等。
- **标记的路径得分  $R$** ：分别用  $R_2^B$  表示第二个字被标记为B时的路径得分， $R_2^E$  表示第二个字被标记为E的路径得分，等等。
- **前一个字的标记到当前字标记转移的特征权重  $T$** ：用  $T_{BM}$  表示由标记B到M的转移特征权重。类似地，其他转移特征权重分别记为： $T_{BE}$ 、 $T_{MM}$ 、 $T_{ME}$ 、 $T_{EB}$ 、 $T_{ES}$ 、 $T_{SB}$  和  $T_{SS}$  等。

# 3. 汉语分词方法

乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

Diagram illustrating the sequence of labels (B, M, E, S) for the sentence "乒乓球拍卖了完了". Red arrows indicate transitions between labels: B to M, M to M, M to E, E to S, S to S, and S to S.

- 利用下式迭代计算每一字被标记为某一种标记的分数：

$$R_{i+1}^B = \max \{ T_{EB} \times R_i^E, T_{SB} \times R_i^S \} \times W_{i+1}^B$$

$$R_{i+1}^E = \max \{ T_{BE} \times R_i^B, T_{ME} \times R_i^E \} \times W_{i+1}^E$$

$$R_{i+1}^S = \max \{ T_{ES} \times R_i^E, T_{SS} \times R_i^S \} \times W_{i+1}^S$$

.....



# 3. 汉语分词方法

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

**第1步：** 计算第1个字“乒”的标记分数(以标记B为例)。由于不存在转移特征，故路径权重 $R_1^B$ 为：

$$R_1^B = W_1^B = \lambda_1 \times f(\text{start}, \text{乒}, B) + \lambda_2 \times f(\text{乒}, B) + \lambda_3 \times f(\text{乒}, B, \text{乒})$$

$f(\bullet)$ 表示特征，其中 $f(\text{start}, \text{乒}, B)$ 表示当前字“乒”被标记为B，前一个字为空； $f(\text{乒}, B)$ 表示当前字“乒”被标记为B； $f(\text{乒}, B, \text{乒})$ 表示当前字“乒”被标记为B，且后一个字为“乒”。特征的权重 $\lambda_1$ 、 $\lambda_2$ 和 $\lambda_3$ 都可以从训练中得到(参数训练部分)。

# 3. 汉语分词方法

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

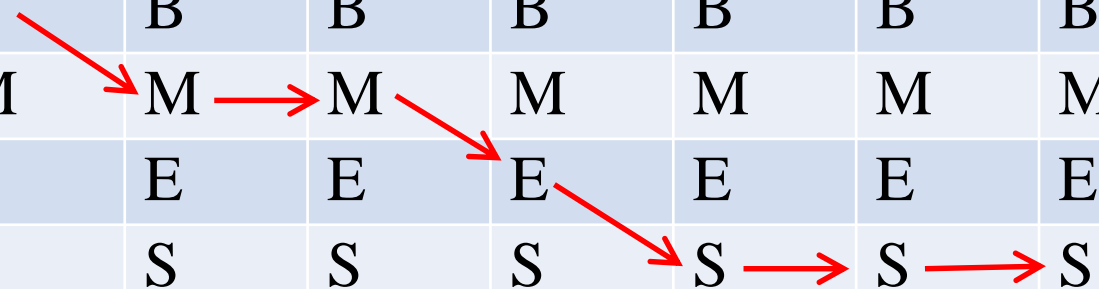
**第2步：** 计算第2个字“乓”的标记分数（以标记B为例）。  
 首先计算一元权重 $W_2^B$ ，继而由上一个字的路径权重计算当前路径权重 $R_2^B$ 为：

$$R_2^B = \max\{T_{EB} \times R_1^E, T_{SB} \times R_1^S\} \times W_2^B$$

同样，对于“乓”字的标记S、M和E分别计算 $R_2^M$ 、 $R_2^E$ 和 $R_2^S$ 。

# 3. 汉语分词方法

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S



**第3步：**同第2步，迭代计算，直至最后一个“了”字，分别得到  $R_7^E$  和  $R_7^S$  两条路径的分值。比较后确定最优路径，然后以该路径的标记点为起始点回溯，得到整个句子的路径标记序列。

解码完毕。

# 3. 汉语分词方法

条件随机场模型的开源代码:

- CRF++ (C++版):

<http://crfpp.googlecode.com/svn/trunk/doc/index.html>

- CRFSuite (C语言版):

<http://www.chokkan.org/software/crfsuite/>

- MALLET (Java版, 通用的自然语言处理工具包, 包括分类、序列标注等机器学习算法):

<http://mallet.cs.umass.edu/>

- NLTK (Python版, 通用的自然语言处理工具包, 很多工具是从MALLET中包装转成的Python接口): <http://nltk.org/>

# 3. 汉语分词方法

关于 **CRFs** 的经典文献:

- [1]J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc.ICML'2001*, pages 282-289
- [2]H. M. Wallach. Conditional Random Fields: An Introduction. *CIS Technical Report MS-CIS-04-21*, Univ. of Penn., 2004

# 3. 汉语分词方法

## ● 方法评价

### ➤ 优点:

- 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。在学习构架上，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识别模块，因此，大大地简化了分词系统的设计[黄昌宁，2006]。

### ➤ 弱点:

- 对于集内词的处理能力不如基于语言模型的分词方法。

# 3. 汉语分词方法

## ④生成式方法与区分式方法的结合

基于 $n$ -gram 的分词方法属于生成式模型(Generative model):

$$WSeq^* = \arg \max_{WSeq} p(WSeq | c_1^n)$$

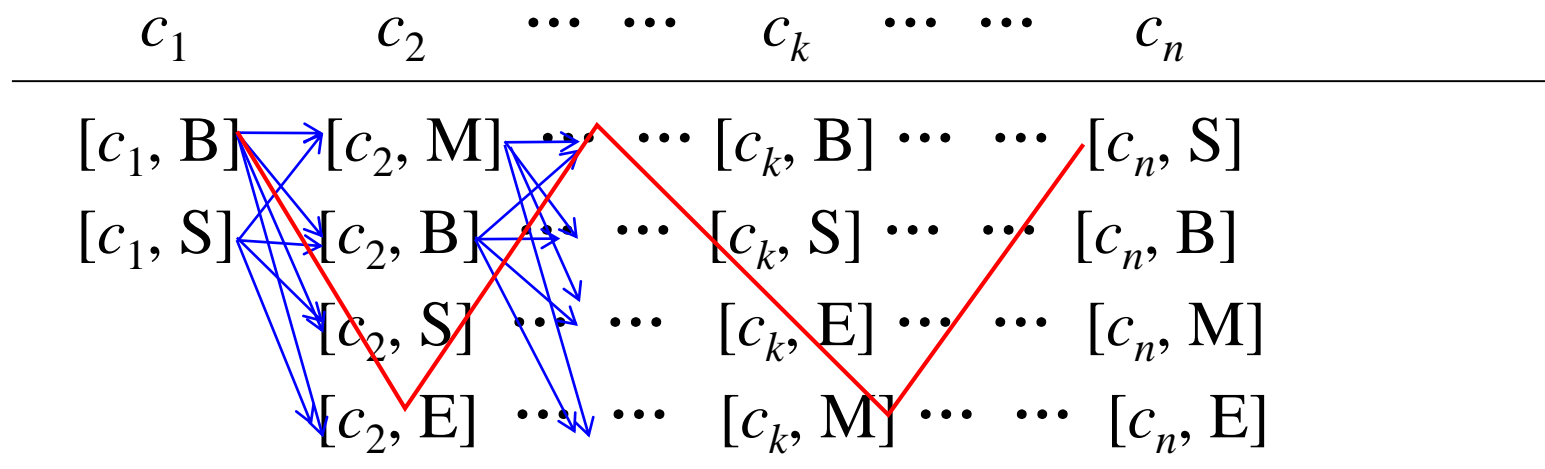
更多地考虑了词汇之间以及词汇内部字与字之间的依存关系, 对于集内词处理有较好的性能, 而由字构词的分词方法是区分式模型(Discriminative model):

$$P(t_1^n | c_1^n) = \prod_{k=1}^n P(t_k | t_1^{k-1}, c_1^n) \approx \prod_{k=1}^n P(t_k | c_{k-2}^{k+2})$$

有利于处理集外词。因此, 可以将两者的优势结合起来。

# 3. 汉语分词方法

✧ **结合方法1:** 将待切分字符串的每个汉字用 $[c, t]_i$ 替代, 以 $[c, t]_i$ 作为统计基元, 利用 $n$ -gram模型选取全局最优(生成式模型)。



[上, B] [海, E] [计, B] [划, E] [到, S] [本, S] [世, B] [纪, E]  $\dots$

$$P([c, t]_1^n) \approx \prod_{i=1}^n P([c, t]_i | [c, t]_{i-k}^{i-1})$$



# 3. 汉语分词方法

## ● 实验结果:

- 利用第二届 SIGHAN Bakeoff 评测语料(2005)
- 4种语料: 北大、台湾中研院、香港城大、微软
- 分词正确率( $P$ ):
  - (1) 基于词的 3-gram:  $P=89.8\%$
  - (2) 基于字的 CRF:  $P=94.3\%$
  - (3) 融合方法 3-gram:  $P=95.0\%$

K. Wang, C. Zong, and K. Su. Which is More Suitable for Chinese Word Segmentation, the Generative Model or the Discriminative One? In *Proc. PACLIC-23*. 3-5 Dec. 3-5, 2009, HK. pp. 827-834

# 3. 汉语分词方法

## 进一步分析：

### ➤ 上述方法的优点：

- (1)充分考虑了相邻字之间的依存关系进行建模；
- (2)相对于区分模型，对集内词(IV)有较好的鲁棒性。

### ➤ 弱点：

难以利用后续的上下文信息。

### ➤ 回顾—基于字的区分式模型的优点：

- (1)相对于基于词的方法，对集外词(OOV)有更好的鲁棒性；
- (2)相对于生成模型，容易处理更多的特征。

# 3. 汉语分词方法

✧ **结合方法2:** 插值法把两种方法结合起来。

$$Score(t_k) = \alpha \times \log(P([c, t]_k \mid [c, t]_{k-2}^{k-1})) + (1 - \alpha) \times \log(P(t_k \mid c_{k-2}^{k+2}))$$

(0.0 ≤ α ≤ 1.0)

Generative score

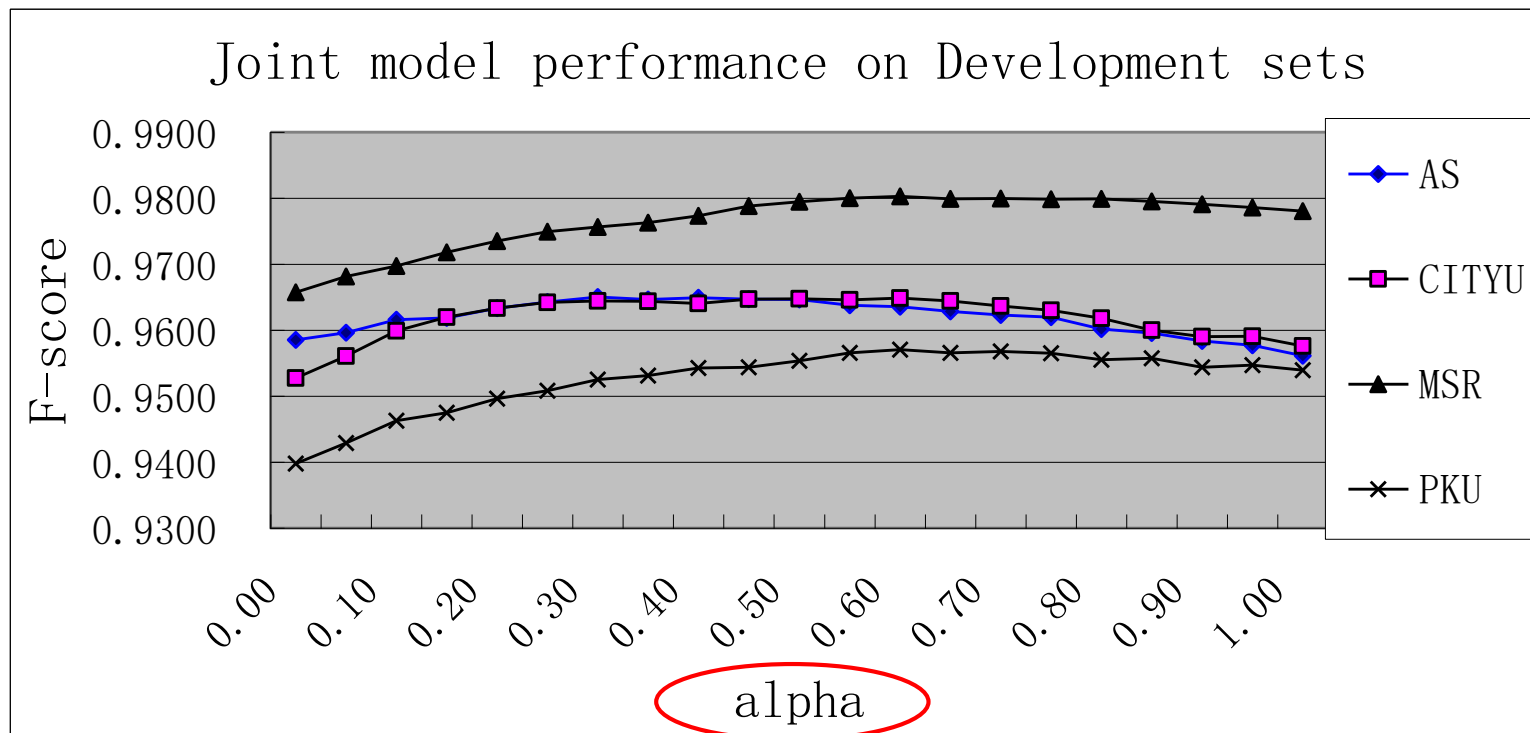
Discriminative score

这样做的优点：充分结合了基于字的生成模型和基于字的区分式模型的优点。

# 3. 汉语分词方法

## ● 实验测试

语料：2005 年 SIGHAN Bakeoff 语料，取少量做开发集。



alpha

# 3. 汉语分词方法

Corpus	Model	$R$	$P$	$F$	$R_{Oov}$	$R_{IV}$
AS	Generative	0.958	0.938	0.948	0.518	0.978
	Discriminative	0.955	0.946	0.951	0.707	0.967
	Joint	0.962	0.950	<b>0.956</b>	0.679	0.975
CITYU	Generative	0.951	0.937	0.944	0.609	0.978
	Discriminative	0.941	0.944	0.942	0.708	0.959
	Joint	0.957	0.951	<b>0.954</b>	0.691	0.979
MSR	Generative	0.974	0.967	0.970	0.561	0.985
	Discriminative	0.957	0.962	0.960	0.719	0.964
	Joint	0.974	0.971	<b>0.972</b>	0.659	0.983
PKU unconverted (ucvt.) case	Generative	0.929	0.933	0.931	0.435	0.959
	Discriminative	0.922	0.941	0.932	0.620	0.941
	Joint	0.935	0.946	<b>0.941</b>	0.561	0.958

### 3. 汉语分词方法

Corpus	Model	$R$	$P$	$F$	$R_{OOV}$	$R_{IV}$
PKU converted (cvt.) case	Generative	0.952	0.951	0.952	0.503	0.968
	Discriminative	0.940	0.951	0.946	0.685	0.949
	Joint	0.954	0.958	<b>0.956</b>	0.616	0.966
Overall	Generative	0.953	0.946	0.950	0.511	0.973
	Discriminative	0.944	0.950	0.947	0.680	0.956
	Joint	0.957	0.955	<b>0.956</b>	0.633	0.971

**总体性能：**相对错误率比区分式模型减少 21%，比生成式模型减少14%。

注：‘(cvt.) case’指已将测试集中的数字、西文字母等编码转换，使其与训练集中的编码一致，‘(ucvt.) case’指未做转换。

# 3. 汉语分词方法

2010 CIPS-SIGHAN 评测结果：

Domain	Mark	OOV Rate	$R$	$P$	$F1$	$R_{OOV}$	$R_{IV}$
Literature	A	0.069	0.937	0.937	0.937	0.652	0.958
Computer	B	0.152	0.941	0.940	0.940	0.757	0.974
Medicine	C	0.110	0.930	0.917	0.923	0.674	0.961
Finance	D	0.087	0.957	0.956	0.957	0.813	0.971

请参阅：

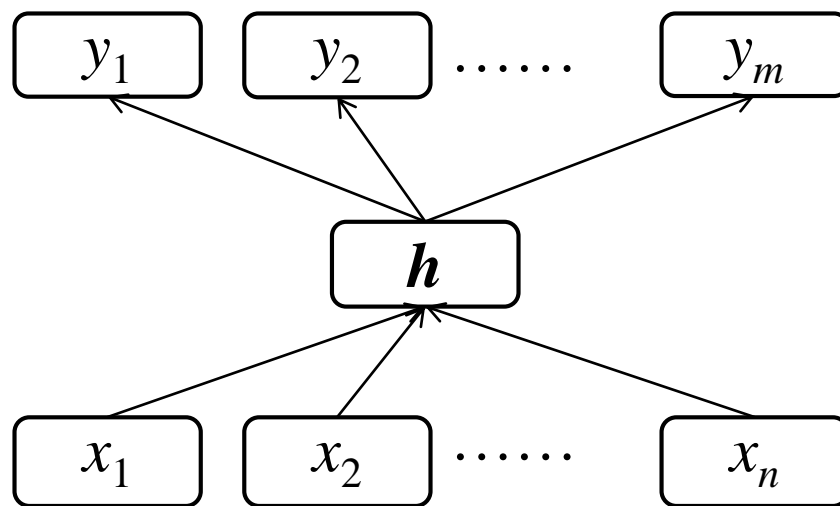
- [1]K. Wang et al. A Character-Based Joint Model for Chinese Word Segmentation. *Proc. COLING 2010*, Aug. 23-27, 2010, pp. 1173-1181
- [2]K. Wang et al. A Character-Based Joint Model for CIPS-SIGHAN Word Segmentation Bakeoff 2010. *Proc. CLP2010*, 2010, pages 245-248
- [3]K. Wang et al. Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation. *ACM TALIP*, Vol. 11, No.2, 2012

Urheen 汉语自动分词系统： <http://www.nlpr.ia.ac.cn/cip/software.htm>

# 3. 汉语分词方法

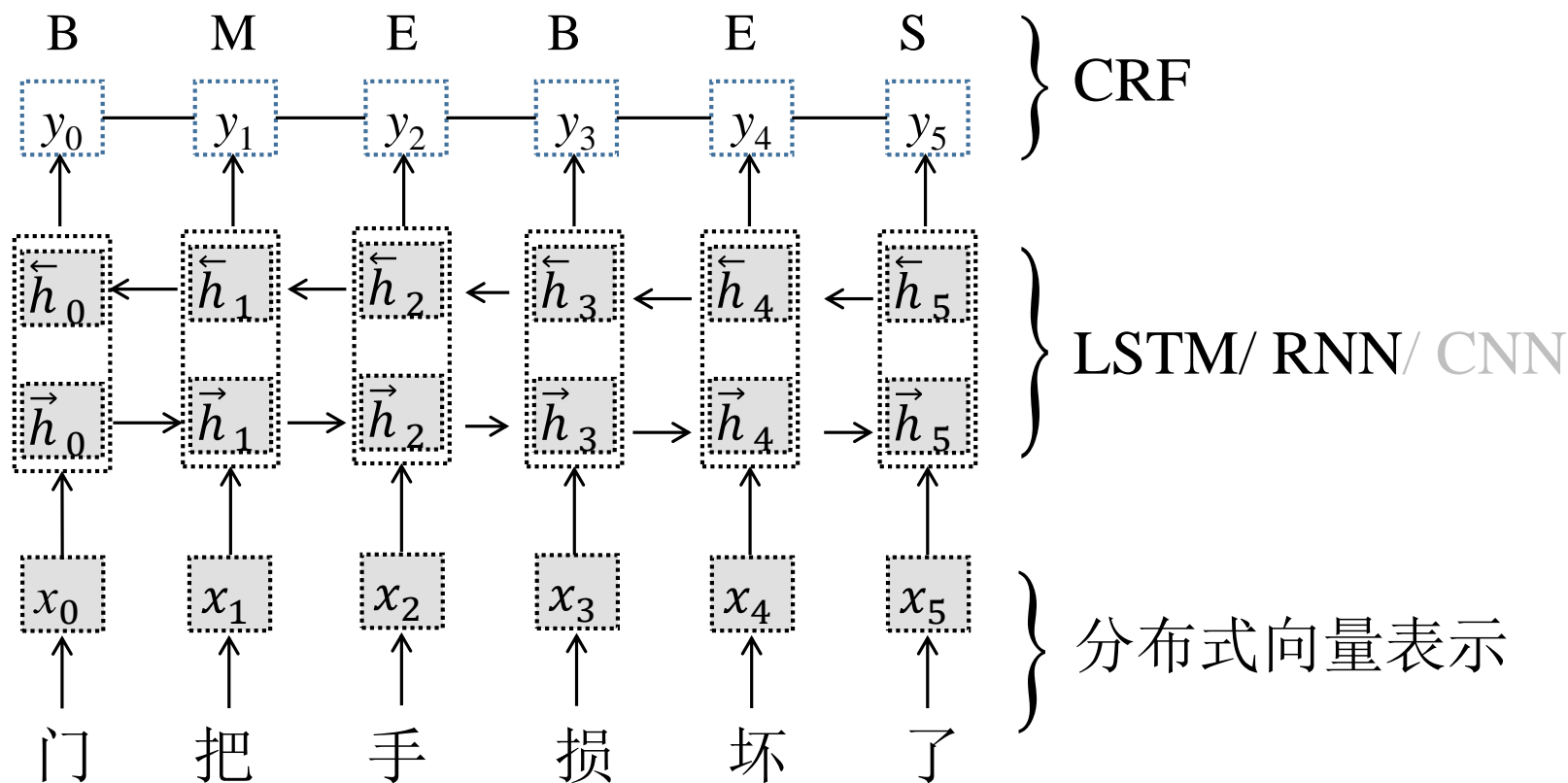
## ⑤ 基于神经网络的分词方法

把分词看作序列标注任务，输入输出均为序列， $n:m$  的对应关系。





# 3. 汉语分词方法



切分结果：门把手/ 损坏/ 了

# 3. 汉语分词方法


## ● 实验结论:

- RNN与CRF相比，CRF取词的窗口作为输入，特征只在窗口范围内选取，而神经网络可以学习长距离关系，但是RNN难以训练，存在梯度消失/爆炸现象；
- 在序列标注任务上，RNN(LSTM)优于CNN；
- LSTM无需使用外部词表资源，效果依然很好；可同时应用到多种语言，多种序列标注任务上；但是，LSTM变种结构多、参数多、调参过程困难。

### 分词工具:

- (1)FastHan: <https://github.com/fastnlp/fastHan> (BERT+CRF)
- (2)WMSeg: <https://github.com/SVAIGBA/WMSeg> (ZEN + CRF)

# 本章内容

1. 概述
2. 汉语分词要点
3. 汉语分词方法
-  4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题
8. 附录

# 4. 命名实体识别

## ◆ 命名实体(Named Entity, NE)

- 通常指：人名、地名、组织机构名、数字、日期、货币和数量。
- 在特定领域，如医学领域，有时也包含专业术语，如疾病名称、药物名称、化学成分等。

命名实体识别(named entity recognition, NER)被简称为NER任务。

# 4. 命名实体识别

## ◆关于汉语人名

- 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。
- 根据我们收集的 300 万个人名统计，姓氏有974个，其中，单姓 952个，复姓 23 个，300万人名中出现汉字4064个。

[曹文洁，2002]

- ✧名字用字范围广，分布松散，规律不很明显，没有标记。
- ✧姓氏和名字都可以单独使用用于特指某一人。
- ✧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。

## 4. 命名实体识别

### ◆关于汉语地方名

- 《中华人民共和国地名录》(1994)收集88026个，不包括相当一部分街道、胡同、村庄等小地方的名称。
- 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、地名是其他专用名词的一部分，地名长度不一等。

# 4. 命名实体识别

## ◆NER与汉语分词的关系

- 在汉语分词的基础上以词为单位使用规则、统计、神经网络等各种方法
- 以汉字为单位使用序列标注方法

## ◆NER方法

- 基于规则的识别方法
- 统计学习方法( $n$ -gram/ CRFs等)
- CRFs + 神经网络
- 利用规则方法进行识别后校正 (可选步骤)

# 4. 命名实体识别

## ● 用于人名识别后的校正规则

### ➤ 修饰规则

如果姓名前是一个数字，或者与“.”字符的距离小于2个字节，则否定此姓名。

### ➤ 边界规则

- 左界规则：若潜在姓名前是一称谓或者标点，或者潜在姓名在句首，或者姓氏使用频率为100%，则左界确定。
- 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。



## 4. 命名实体识别

- 用于机构名识别后的校正规则
  - 找到一机构称呼词
  - 根据相应规则往前逐个检查名词作为修饰名词的合法性，直到发现非法词
  - 如果所接受的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

## 4. 命名实体识别

推荐参阅:

- [1] Y. Chen et al. A Joint Model to Simultaneously Identify and Align Bilingual Named Entities. *Computational Linguistics*, 39(2): 229-266
- [2] Y. Chen et al. On Jointly Recognizing and Aligning Bilingual Named Entities. *Proc. ACL'2010*, pp. 631-639
- [3] C. Dong et al. December. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. *Proc. NLPCC'2016*, pp. 239-250. [Radical-LSTM:在 LSTM-CRF结构基础上, 对中文汉字做偏旁部首级别的 LSTM 变换。]

# 4. 命名实体识别

## ◆分词与NER存在的主要问题

- 过于依赖训练样本，而标注大规模训练样本费时费力，且仅局限于个别领域，由此导致分词和NER系统对新词的识别能力差，往往在与训练样本差异较大的测试集上性能大幅度下降。
- 现有的训练样本主要在新闻领域，而实际应用千差万别：网络新闻、微博/ 微信/ QQ 等非规范文本、不同的专业领域(中医药、生物、化学、能源 ..... )。

**领域差异和生词识别是分词和NER面临的最大挑战**

## 4. 命名实体识别

### ● 举例1

李时珍（约1518～1593），字东璧，晚号濒湖山人，蕲州（今湖北蕲春）人。世业医，父言闻，有医名。幼习儒，三次应乡试不中。自嘉靖三十一年（1552年）至万历六年（1578年），历时二十七载，三易其稿，著成《本草纲目》五十二卷，初刊于金陵。

公开的分词系统切分准确率为：57.3%～94.8%

# 4. 命名实体识别

## ● 举例2

类别	类别描述
事件报道	特定事件/具体事件
新闻内容	新闻消息/格式较规范
观点传播	观点词汇多/日常闲谈/观点评论
信息共享	分享的信息或者链接/为他人提供的建议
私人会话	帖子开头有“@某人”/日常闲谈
交易信息	帖子中出现金钱、比例词汇

根据对2011年微博内容的统计，大约75%的内容为个人心情和感受方面的。

# 4. 命名实体识别

补充词汇:

词典来源	词语数量
维基百科+常用在线词典(普通词汇)	<b>1301320</b>
(1)微博用语词库	10330
(2)网络用语大全	294
(3)网络关键词以及词频数据	500000
(4)《人民日报》微博词频统计	42315
(5)百度百科对于网络用语的解释	1051
<b>上述五项经合并、筛选后形成的网络用语词典</b>	<b>541941</b>
网络情感词典+传统情感词典(情感词汇)	<b>26207</b>
<b>词汇总数: 1753925 (经过合并筛选)</b>	

# 4. 命名实体识别

分词性能:

分词方法	准确率(%)	召回率(%)	F1值(%)
<b>Stanford</b>	80.40	76.52	78.41
<b>Urheen</b>	80.46	77.43	78.92
<b>ICTCLAS(+微博处理)</b>	82.62	83.52	83.07
<b>CWS</b>	80.12	73.24	76.52
<b>CWS(+词典+符号处理)</b>	<b>90.52</b>	<b>90.73</b>	<b>90.62</b>

**CWS:** Chinese word segmentation based on ME model

# 本章内容

---

1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
- ➡ 5. 子词压缩
6. 词性标注
7. 习题
8. 附录



# 5. 子词压缩

句子表示和生成的基本粒度：单词和字符/字。

## ◆ 以单词为基本粒度的缺点：

- “长尾”分布，低频词的表示较差
- 词表的单词较多，计算复杂性高

单词1	↑ 单词
单词2	
...	
单词29999	↓ 集外词
<unk>	

## ◆ 以字符为基本粒度的缺点：

- 字符的歧义性较高
- 字符的序列长度较大

字符1	↑ 字符/字
字符2	
...	
字符2999	↓ 集外字
<unk>	

## ◆ 寻找一种介于单词和字符之间的粒度：子词(sub-word)

综合单词和字符粒度的优势，使其在句子表示和生成中最好地发挥作用。

# 5. 子词压缩

## ● 基本思路

- 对于汉语文本，如果有很好的分词工具，先对文本进行词语切分，在切分结果的基础上利用双字节编码算法(Byte Pair Encoding, BPE)算法进行单字压缩，合并有着最大次数的相邻字符。
- 对于英语等屈折语文本，可直接用BPE算法进行字符压缩。

# 5. 子词压缩

## ● BPE算法

- ① 对邻近的两个字符(汉字)合并，统计被合并的两个邻近字符(汉字)在整个文本中出现的次数 $\alpha$ ；
- ② 将 $\alpha$ 最大的两个邻近字符(汉字)用原文本中不存在的符号替换(压缩)，重复进行上面的操纵。直到没有被合并的字符(汉字)为止，或者达到限定合并的次数。

例1: aaabdaaababc China

① a a a b d a a a b a b c  ② XabdXababc  ③ XYdXYXYc

aa 出现4次  
ab 出现3次  
其他出现1次

X=aa

ab出现次数  
最多: 3次  
Y=ab

XY出现次数最多:  
2次。Z=XY

④ ZdZYc

aaab@@ d@@ aaab@@ ab@@ c

还原, 标记

# 5. 子词压缩

例2:

这 / 本 / 童 话 书 / 已 / 翻 译 / 成 / 中 文 / 了  
 请 / 用 / 中 文 / 复 述 / 这 / 篇 / 故 事  
 中 文 / 的 / “ / 危 机 / ” / 分 / 为 / 两 / 个 / 字  
 春 / 因 / 繁 花 / 而 / 美 丽 /  
 繁 花 似 锦 / 的 / 深 圳 / 洋 溢 / 着 / 欢 乐 / 的 / 气 氛



第一次迭代, 合并“中文”(3次)

这 / 本 / 童 话 书 / 已 / 翻 译 / 成 / 中文 / 了  
 请 / 用 / 中文 / 复 述 / 这 / 篇 / 故 事  
 中文 / 的 / “ / 危 机 / ” / 分 / 为 / 两 / 个 / 字  
 春 / 因 / 繁 花 / 而 / 美 丽 /  
 繁 花 似 锦 / 的 / 深 圳 / 洋 溢 / 着 / 欢 乐 / 的 / 气 氛



第二次迭代, 合并“繁花”(2次)

这 / 本 / 童 话 书 / 已 / 翻 译 / 成 / 中文 / 了  
 请 / 用 / 中文 / 复 述 / 这 / 篇 / 故 事  
 中文 / 的 / “ / 危 机 / ” / 分 / 为 / 两 / 个 / 字  
 春 / 因 / 繁花 / 而 / 美 丽 /  
 繁花 似 锦 / 的 / 深 圳 / 洋 溢 / 着 / 欢 乐 / 的 / 气 氛

“/” 为分词标记

循环修改过程, 直到: 1)达到最大迭代次数; 或者2)双字符的最大出现次数为1 (约定数)。

## 5. 子词压缩

在机器翻译中，WMT14 的训练语料为450万英德双语对照的平行句对，采用子词压缩合并之后，抽取出的词表为3.2万个子词（源语言端和目标语言大都是这个数目）。

### 参考文献：


Rico Sennrich, Barry Haddow, and Alexandra Birch, 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of ACL 2016, pages 1715–1725.

### 开源代码：

<https://github.com/rsennrich/subword-nmt>

# 本章内容

---

1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
-  6. 词性标注
7. 习题
8. 附录

# 6. 词性标注

## ◆ 面临的问题

词性(part-of-speech, POS)标注(tagging)的主要任务是消除词性兼类歧义。在任何一种自然语言中，词性兼类问题都普遍存在。例如：

(1) Time flies like an arrow.

(2) I want you to web our annual report.

对 Brown 语料库的统计，55%词次兼类。根据《现代汉语八百词》，兼类词占 22.5%。

# 6. 词性标注

## ● 在汉语中

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

例句：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)”。例句：每次他都**会**在**会**上制造点新闻。

(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)”。例句：用那种方式**教育**孩子，简直是对**教育**事业的嘲笑。

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



# 6. 词性标注

## ◆ 标注集的确定原则

不同语言中，词性划分基本上已经约定俗成。  
自然语言处理中对词性标记要求相对细致。

### ● 一般原则：

- 标准性：普遍使用和认可的分类标准和符号集；
- 兼容性：与已有资源标记尽量一致，或可转换；
- 可扩展性：扩充或修改。

# 6. 词性标注

## ● UPenn Treebank 的词性标注集

- **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。

## ● 北大计算语言研究所的词性标注集

- **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)。

# 6. 词性标注

## ◆ 标注方法

- 基于规则/有限状态机的词性标注方法
- 基于统计模型的词性标注方法
  - $n$ -gram model
  - HMM model
  - CRFs model
- 规则和统计方法相结合的词性标注方法

## ◆ 性能评价指标：准确率

# 6. 词性标注

- 基于规则的方法

- 手工编写消歧规则

- ① 建立非兼类词典
- ② 建立兼类词典一词性可能出现的概率高低排列
- ③ 构造兼类词识别规则

# 6. 词性标注

## (1) 并列鉴别规则

如：体现了人民的要求(N/V ?)和愿望(N, 非兼类)。

## (2) 同境鉴别规则

如：一个优秀的企业必须具备一流的产品(名词, 非兼类)、一流的管理(N/V ?)和一流的服务(N/V ?)。

## (3) 区别词鉴别规则(区别词只能直接修饰名词)

如：这次大型(鉴别词, 非兼类) 调查(V/N ?)历时半年。

## (4) 唯名形容词鉴别规则(有些形容词只能直接修饰名词)

如：重大（唯名形容词）损失（N/V ?）

巨大（唯名形容词）影响（N/V ?）

# 6. 词性标注

## ➤根据词语的结构建立词性标注规则

### (1) 词缀（前缀、后缀）规则

- 形容词：蓝茵茵，绿油油，金灿灿...
- 数量词：一片片，一次次，一回回...
- 人名简称：李总，张工，刘老...
- 其他：年轻化，知识化，...{化}  
           篮球赛，足球赛，...{赛} ...

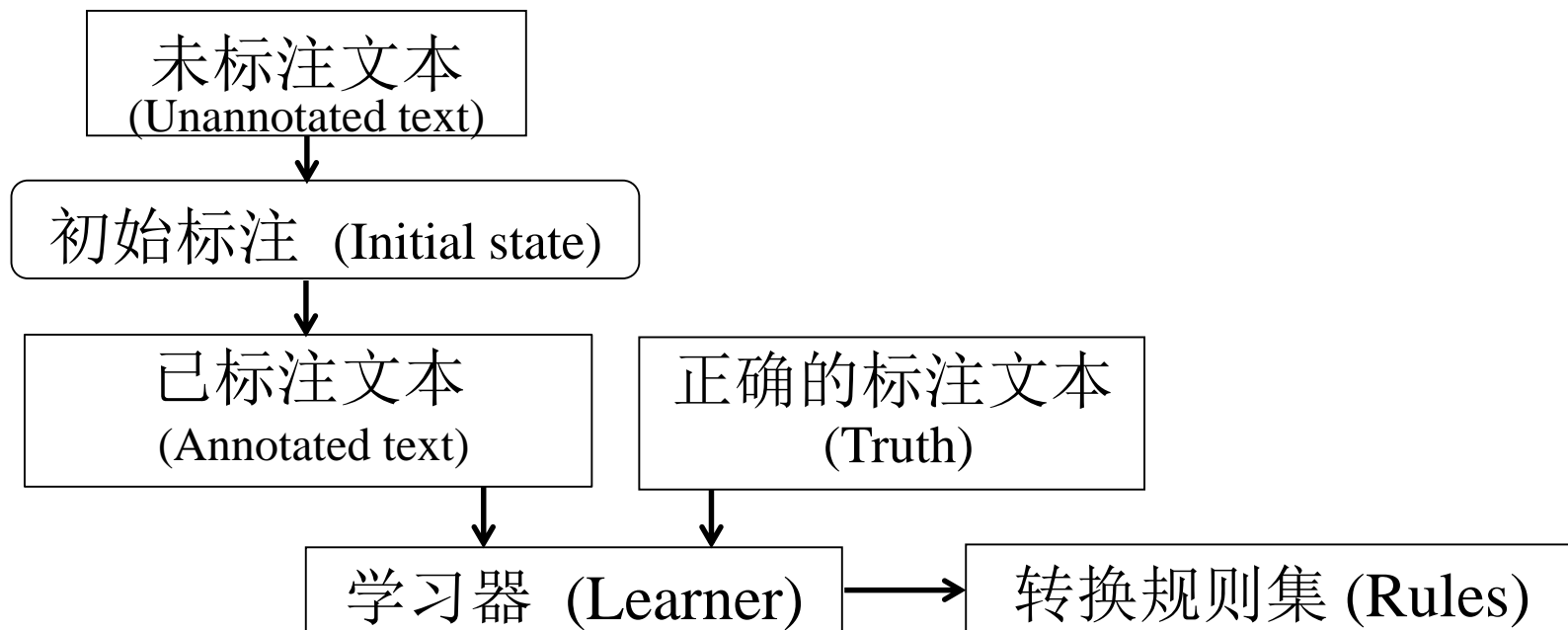
### (2) 重叠词规则

- 看看，瞧瞧，高高兴兴，热热闹闹...

# 6. 词性标注


## ● 机器学习方法：序列标注

- 初始词性赋值
- 对比正确标注的句子，自动学习结构转换规则
- 利用转换规则调整初始赋值



# 本章内容

---

1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注
-  7. 习题
8. 附录



# 7. 习题

1. 阅读《信息处理用现代汉语分词规范》(中华人民共和国国家标准 GB13715), 了解规范的基本内容。
2. 利用已经学过的技术方法和北京大学标注的《人民日报》分词和词性标注语料, 设计实现一个多种方法相结合的汉语自动分词系统。然后利用不同类型的网络文本测试你的分词系统, 分析不同分词方法在不同测试样本上性能的变化。
3. 在第2题得到的分词结果的基础上, 进行子词压缩。
4. 设计实现一个人名识别系统(针对中英文均可)。
5. 设计实现一个组织机构名识别系统(针对中英文均可)。
6. 设计实现一个基于CRFs模型的汉语词性标注系统。

**第2题和第3题做作业。**

# 本章小结

## ◆汉语分词要点

- 汉语分词中的主要问题
- 两种歧义
- 切分原则：基本原则和辅助原则
- 性能评价方法

## ◆分词方法

- MM、最少分词法、统计法、由字构词法（CRFs）等

## ◆命名实体识别：人名、地名、组织机构名识别

## ◆子词压缩

## ◆词性标注：标注原则、规范、兼类消除

# 本章内容

---

1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题

 8. 附录

# 8. 附录

◆ 英语形态分析



◆ 汉语最少分词方法

- 基本任务

- 单词识别

- 形态还原

## 8. 附录

### ◆ 单词识别

例 (1) Mr. Green is a good English teacher.

(2) I'll see prof. Zhang home after the concert.

识别结果:

(1) Mr./ Green/ is/ a/ good/ English/ teacher/.

(2) I/ will/ see/ prof./ Zhang/ home/ after/ the/ concert/.

## 8. 附录

### ◆ 英语中常见的特殊形式的单词识别：词典+规则

(1) prof., Mr., Ms. Co., Oct. 等放入词典；

(2) Let's / let's => let + us

(3) I'am => I + am

(4) {it, that, this, there, what, where}'s =>  
{it, that, this, there, what, where} + is

(5) can't => can + not;

won't => will + not

## 8. 附录

- (6) {is, was, are, were, has, have, had}n't =>  
      {is, was, are, were, has, have, had} + not
- (7) X've => X + have;  
      X'll=> X + will;   X're => X + are
- (8) he's => he + is / has => ?  
      she's => she + is / has => ?
- (9) X'd Y => X + would   (如果 Y 为单词原型)  
      => X + had       (如果 Y 为过去分词)

# 8. 附录

## ◆ 英语单词的形态还原

### (1) 有规律变化单词的形态还原

1) -ed 结尾的动词过去时，去掉 ed;

\*ed → \* (e.g., worked → work)

\*ed → \*e (e.g., believed → believe)

\*ied → \*y (e.g., studied → study)

2) -ing 结尾的现在分词，

\*ing → \* (e.g., developing → develop)

\*ing → \*e (e.g., saving → save)

\*ying → \*ie (e.g., die → dying)

3) -ly 结尾的副词: \*ly → \* (e.g., hardly → hard)



# 8. 附录

4) -s 结尾的动词单数第三人称;

\*s → \* (e.g., works → work)

\*es → \* (e.g., discuss → discusses)

\*ies → \*y (e.g., studies → study)

5) -er/est 结尾的形容词比较级、最高级

\*er → \* (e.g., cold → colder)

\*ier → \*y (e.g., easier → easy) .....

6) s/ses/xes/ches/shes/oes/ies/ves 结尾的名词复数, ies/ves 结尾的名词还原时做相应变化:

bodies → body, shelves → shelf,

boxes → box, etc.

7) 名词所有格 X's, Xs'

## 8. 附录

### (2)动词、名词、形容词、副词不规则变化单词的形态还原

—建立不规则变化词表

例: choose, chose, chosen

axis, axes

bad, worse, worst

### (3)对于表示年代、时间、百分数、货币、序数词的数字形态还原

1) 1990s → 1990, 标明时间名词;

2) 87th → 去掉 th 后, 记录该数字为序数词;

3) \$20 → 去掉\$, 记录该数字为名词(20美圆);

4) 98.5% → 98.5% 作为一个数词。

## 8. 附录

### (4)合成词的形态还原

- 1)基数词和序数词合成的分数词，如 one-fourth 等。
- 2)名词+名词、形容词+名词、动词+名词等组成的合成名词，如 Human-computer, multi-engine, mixed-initiative 等。
- 3)形容词+名词+ed、形容词+现在分词、副词+现在分词、名词+过去分词、名词+形容词等组成的合形成形容词，如 machine-readable, hand-coding, non-adjacent, context-free 等。
- 4)名词+动词、形容词+动词、副词+动词构成的合成动词，如 job-hunt 等。
- 5)其他带连字符“-”的合成词，如 co-operate, 7-color, inter-lingua, Chinese-to-English, state-of-the-art, *i*-th 等。

# 8. 附录

## ◆形态分析的一般方法

- (1) 查词典，如果词典中有该词，直接确定该词的原形；
- (2) 根据不同情况查找相应规则对单词进行还原处理，如果还原后在词典中找到该词，则得到该词的原形；如果找不到相应变换规则或者变换后词典中仍查不到该词，则作为未登录词处理；
- (3) 进入未登录词处理模块。

# 8. 附录

---

◆ 英语形态分析

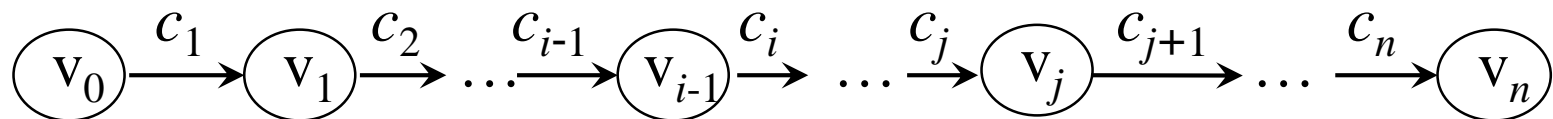
◆ 汉语最少分词方法

又称最短路径法，属于有词典的切分方法。

# 8. 附录

## ◆ 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， $n$  为串的长度， $n \geq 1$ 。建立一个节点数为  $n+1$  的切分有向无环图  $G$ ，各节点编号依次为  $V_0, V_1, V_2, \dots, V_n$ 。



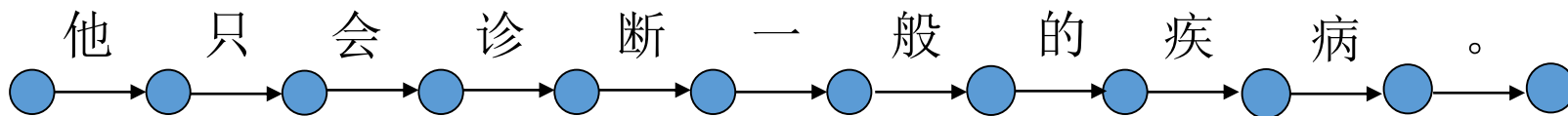
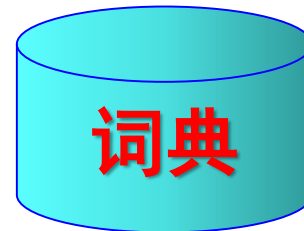
求最短路径：贪心法或简单扩展法。

## 8. 附录

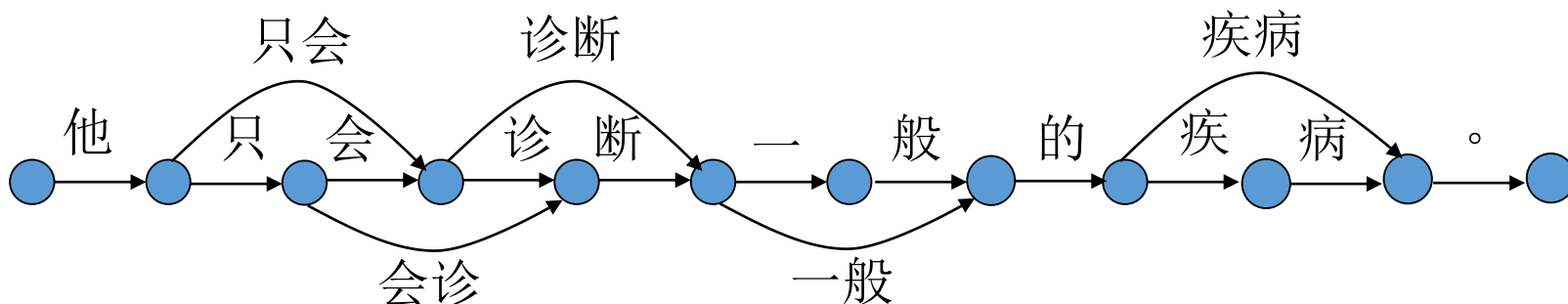
给定输入字符串：他只会诊断一般的疾病。

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)

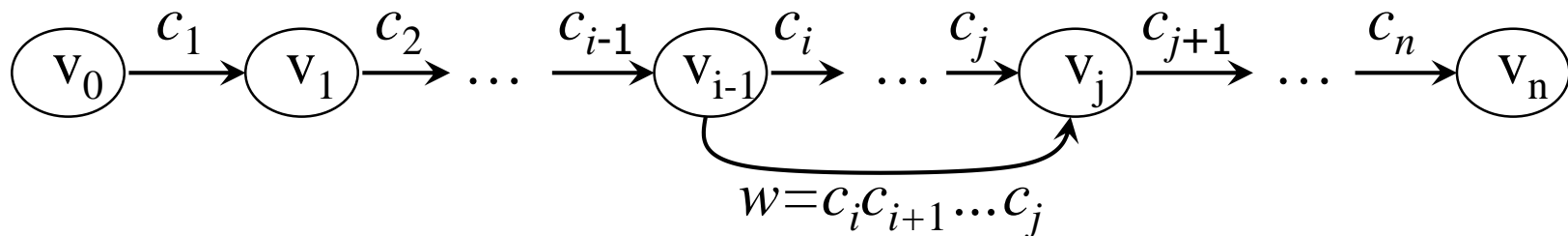
他/ 只/ 会诊/ 断/ 一般/ 的/ 疾病/。 (词个数: 8)

最终结果: 他/ 只会/ 诊断/ 一般/ 的/ 疾病/。

# 8. 附录

## ● 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。



## 8. 附录

例如：给定字符串：他说的确实在理。

输出候选：他/ 说/ 的/ 确实/ 在理/ 。（词个数：6）

他/ 说/ 的确/ 实在/ 理/ 。（词个数：6）

多个切分结果词数相同时系统无法做正确的判断。

# 8. 附录

## ● 方法评价

### ➤ 优点:

- 切分原则符合汉语自身规律;
- 需要的语言资源 (词表) 也不多。

### ➤ 弱点:

- 对许多歧义字段难以区分, 最短路径有多条时, 选择最终的输出结果缺乏应有的标准;
- 字串长度较大和选取的最短路径数增大时, 长度相同的路径数急剧增加, 选择最终正确的结果困难越来越大。

---

谢谢!

*Thanks!*

