

自然语言 处理与理解

赵云蒙

华东理工大学 信息科学与工程学院
能源化工过程智能制造教育部重点实验室
2023-2024 第二学期

第7章

词法分析与词性标注

7.1 概述

7.1 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。
- ★ 自动词法分析就是利用计算机对自然语言的形态(morphology)进行分析，判断词的结构和类别等。
- ★ 词性或称词类(Part-of-Speech, POS)是词汇最重要的特性，是连接词汇到句法的桥梁。

★ 不同语言的词法分析

✦ **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。

▲ 词法分析：词的形态分析（形态还原）。

✦ **分析语**（孤立语）（如：汉语）

▲ 词法分析：分词。

✦ **黏着语**（如：日语等）：

▲ 词法分析：分词+形态还原。

7.2 英语的形式分析

7.2 英语的形式分析

★ 基本任务

1、单词识别

2、形态还原

7.2 英语的形式分析

★ 英语单词的识别

★ 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/.

7.2 英语的形式分析

★ 英语中常见的特殊形式的单词识别

- (1) Prof., Mr., Ms. Co., Oct. 等放入词典;
- (2) Let's / let's \rightarrow let + us
- (3) I'm \rightarrow I + am
- (4) {it, that, this, there, what, where}'s \rightarrow
 {it, that, this, there, what, where} + is
- (5) can't \rightarrow can + not;
 won't \rightarrow will + not

7.2 英语的形式分析

(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)

7.2 英语的形式分析

★ 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)

7.2 英语的形式分析

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) -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)

7.2 英语的形式分析

4) -ly 结尾的副词

*ly → * (e.g., hardly → hard)

.....

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)

.....

7.2 英语的形式分析

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

.....

7) 名词所有格 X's, Xs'

7.2 英语的形式分析

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

7.2 英语的形式分析

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美元) ;
- 4) 98.5% → 98.5% 作为一个数词。

4. 合成词的形态还原

- 1) 基数词和序数词合成的分数词, e.g., 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, e.g., Human-computer, multi-engine, mixed-initiative, large-scale 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 + 形容词等组成的合成形容词, e.g., machine-readable, hand-coding, non-adjacent, context-free, rule-based, speaker-independent 等。

7.2 英语的形式分析

- 4) 名词 + 动词、形容词 + 动词、副词 + 动词构成的合成动词, e.g., job-hunt 等。
- 5) 其他带连字符 "-" 的合成词, e.g., co-operate, 7-color, bi-directional, inter-lingua, Chinese-to-English, state-of-the-art, part-of-speech, OOV-words, spin-off, top-down, quick-and-dirty, text-to-speech, semi-automatically, *i*-th 等。

7.2 英语的形式分析

★ 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。

7.3 汉语自动分词概要

7.3 汉语自动分词概要

★ 汉语自动分词的重要性

- ✦ 自动分词是汉语句子分析的基础
- ✦ 词语的分析具有广泛的应用（词频统计，词典编纂，文章风格研究等）
- ✦ 文献处理以词语为文本特征
- ✦ “以词定字、以词定音”，用于文本校对、同音字识别、多音字辨识、简繁体转换

7.3 汉语自动分词概要

★ 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

7.3 汉语自动分词概要

★ 歧义切分字段处理

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

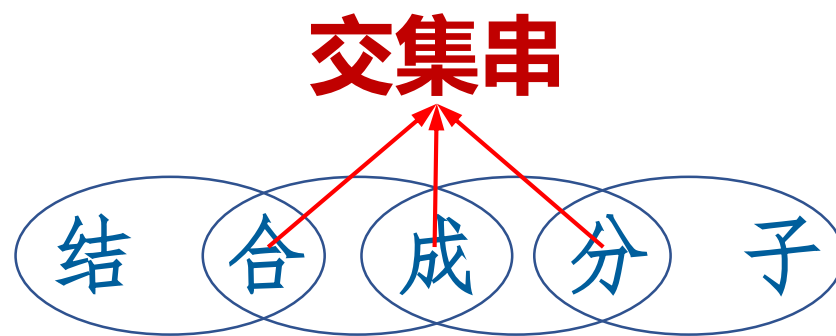
中/国人/为了/实现/自己/的/梦想

✦ 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

7.3 汉语自动分词概要

★ 链长

✦ 定义：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。

7.3 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。

7.3 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
- ✦ 梁南元（1987）曾经对一个含有48,092字的自然科学、社会科学样本进行了统计，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。

★ 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，蔡国庆，党政法，水皮，高升，高山，夏天，温馨，武夷山，陈冯富珍，平川三太郎，约翰·斯特朗，詹姆斯·埃尔德

2. 新出现的词汇、术语、个别俗语等，例如

新冠，懂王，睡王，奥力给，yyds，yygq，yysy，u1s1

7.3 汉语自动分词概要

★ 例如

- (1) 他还兼任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。

7.3 汉语自动分词概要

★ 一处统计：

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			

★ 汉语自动分词的基本原则

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。（合并原则）

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）

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。（合并原则）

- (1) 字串的语法功能不符合组合规律，如：好吃，好喝，
好听，好看
- (2) 字串的内部结构不符合语法规律，如：游水

★ 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。

1、有明显分隔符标记的应该切分之（切分原则）

✦ 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、附着性语/词素和前后词合并为一个分词单位（合并原则）

✦ 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、民营化等

3、使用频率高或共现率高的字串尽量合并为一个分词单位 (合并原则)

✦ 例如：

“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

4、双音节加单音节的偏正式名词尽量合并为一个分词单位 (合并原则)

✦ 例如：

“线、权、车、点”等所构成的偏正式名词：“国际线、分数线、三八线”、“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。

5、双音节结构的偏正式动词应尽量合并为一个分词单位 (合并原则)

✦ 本原则只适合少数偏正式动词，如：

“紧追其后”、“组建完成”等，不适合动宾及主谓式复合动词。

6、内部结构复杂、合并起来过于冗长的词尽量切分（切分原则）

(1) 词组带接尾词

太空/ 计划/ 室、塑料/ 制品/ 业

(2) 动词带双音节结果补语

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

7.3 汉语自动分词概要

(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

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、歌曲名等

鲸鱼/ 的/ 生/ 与/ 死、那/ 一/ 年/ 我们/ 都/ 很/ 酷

(7) 专名带普通名词

胡/ 先生、京沪/ 铁路

7.4 分词与词性标注结果 评价方法

7.4 分词与词性标注结果评价方法

★ 两种测试

★ 封闭测试 / 开放测试

★ 专项测试 / 总体测试

7.4 分词与词性标注结果评价方法

★ 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，

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

7.4 分词与词性标注结果评价方法

★ **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。

7.4 分词与词性标注结果评价方法

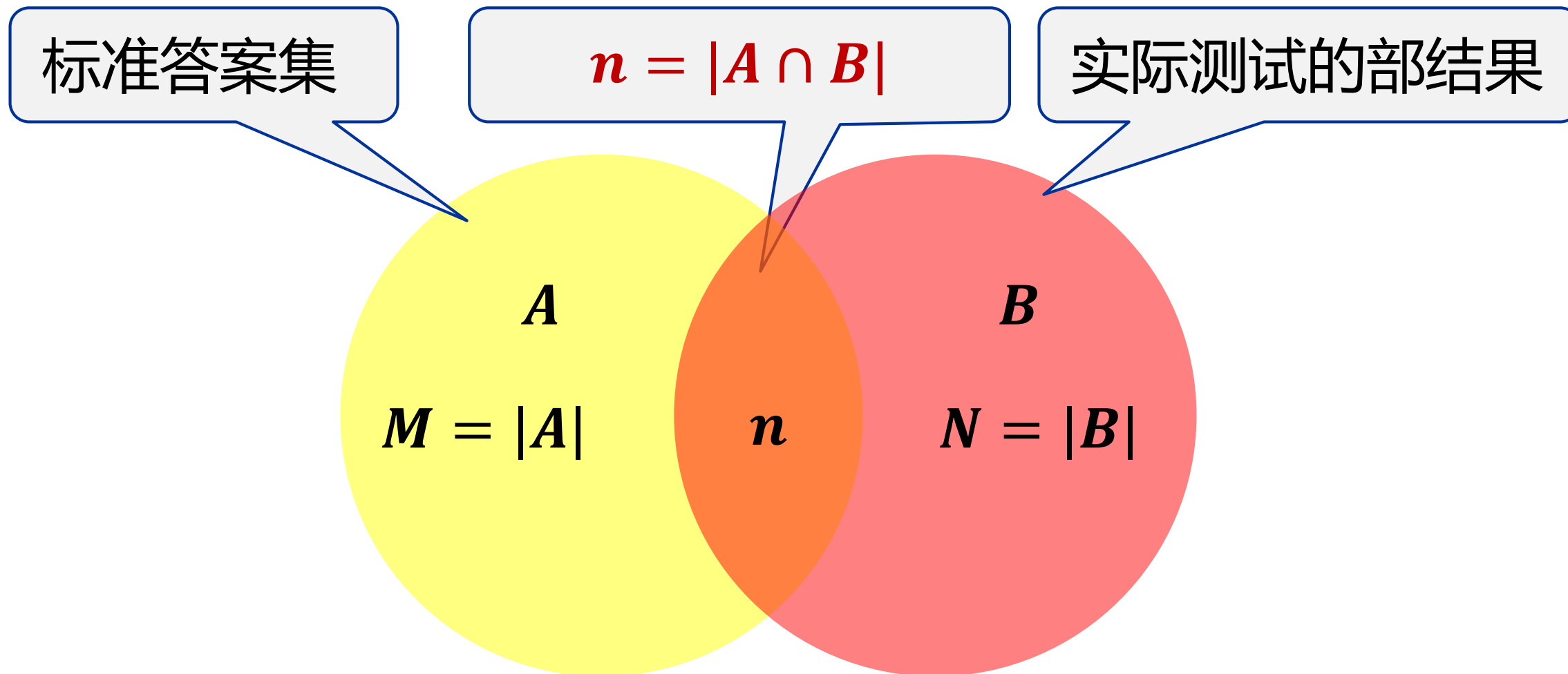
★ **F-测度值** (F-Measure) : 正确率与找回率的综合值。

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

一般地, 取 $\beta = 1$, 即

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

7.4 分词与词性标注结果评价方法



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

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

7.4 分词与词性标注结果评价方法

- ★ 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：

$$\begin{aligned} P &= \frac{4120}{5260} \times 100\% = 78.33\% & F1 &= \frac{2 \times P \times R}{P + R} \times 100\% \\ R &= \frac{4120}{4510} \times 100\% = 91.35\% & &= \frac{2 \times 78.33 \times 91.35}{78.33 + 91.35} \times 100\% \\ & & &= 84.34\% \end{aligned}$$

7.5 汉语自动分词基本算法

7.5 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

7.5 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

✦ 假设句子： $S = c_1 c_2 \cdots c_n$ ，某一词： $w_i = c_1 c_2 \cdots c_m$ ， m 为词典中最长词的字数。

★ 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)

7.5 汉语自动分词基本算法

★ FMM 算法描述

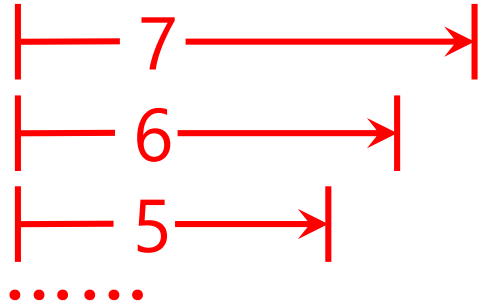
- (1) 令 $i = 0$ ，当前指针 p_i 指向输入字串的初始位置，执行下面的操作：
- (2) 计算当前指针 p_i 到字串末端的字数（即未被切分字串的长度） n ，如果 $n = 1$ ，转(4)，结束算法。否则，令 $m =$ 词典中最长单词的字数，如果 $n < m$ ，令 $m = n$ ；
- (3) 从当前 p_i 起取 m 个汉字作为词 w_i ，判断：
 - (a) 如果 w_i 确实是词典中的词，则在 w_i 后添加一个切分标志，转(c)；
 - (b) 如果 w_i 不是词典中的词且 w_i 的长度大于1，将 w_i 从右端去掉一个字，转(a)步；否则（ w_i 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。

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

7.5 汉语自动分词基本算法

★ 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。

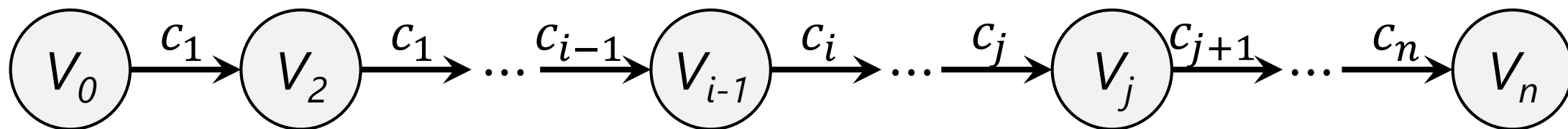
★ 弱点:

- ★ 歧义消解的能力差；
- ★ 切分正确率不高，一般在95%左右。

7.5 汉语自动分词基本算法

2. 最少分词法（最短路径法）

★ **基本思想：** 设待切分字符串 $S = c_1 c_2 \cdots c_n$ ，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_1, V_2, \dots, V_n 。

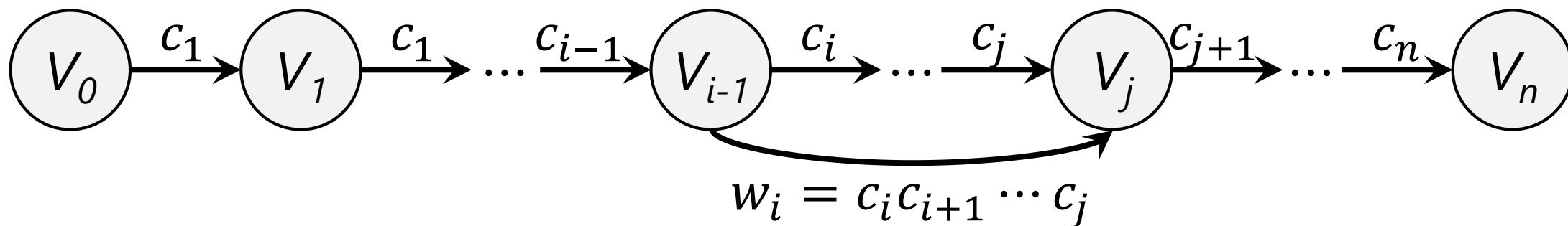


★ **求最短路径：** 贪心法或简单扩展法。

7.5 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 $w_i = c_i c_{i+1} \cdots c_j (0 < i < j \leq n)$ 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 ;



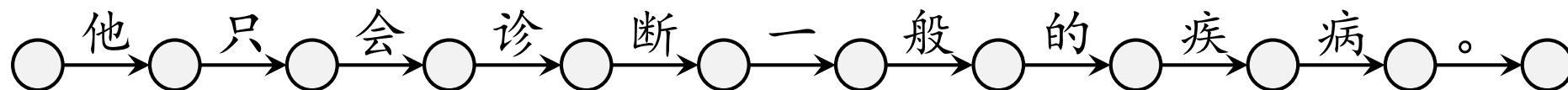
- (3) 重复步骤(2), 直到没有新路径 (词序列) 产生;
- (4) 从产生的所有路径中, 选择路径最短的 (词数最少的) 作为最终分词结果。

7.5 汉语自动分词基本算法

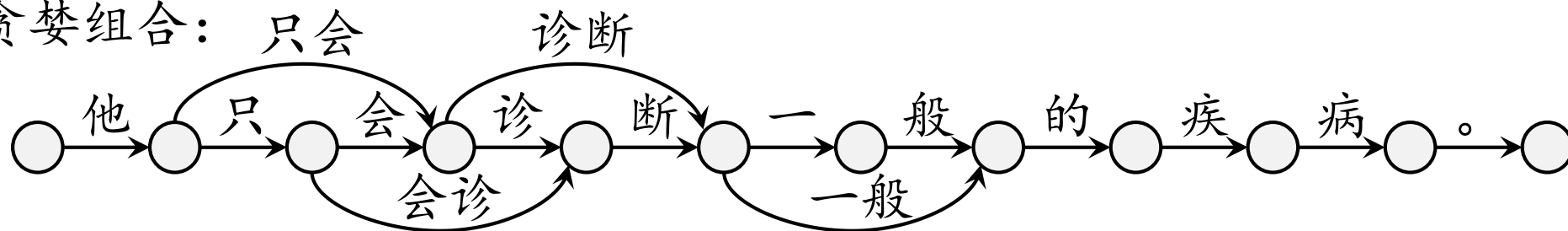
★ 例: (1) 输入字符串: 他只会诊断一般的疾病。

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/。

7.5 汉语自动分词基本算法

★ 例: (2) 输入字符串: 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)

★ 系统无法做正确性判断。

7.5 汉语自动分词基本算法

★ 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。

7.5 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 , $W = w_1 w_2 \cdots w_k (1 \leq k \leq n)$ 是一种可能的切分。

$$W^* = \operatorname{argmax}_W p(W|S)$$

$$= \operatorname{argmax}_W p(W) \times p(S|W)$$

语言模型

生成模型

详见第5章举例

7.5 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。

★ 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。

7.5 汉语自动分词基本算法

4. 基于HMM的分词方法

★ **基本思想**：把输入字符串（句子） S 作为HMM μ 的输入；切分后的单词串 S_w 为状态的输出，即观察序列 $S_w = w_1 w_2 \cdots w_n, n \geq 1$ 。词性序列 S_c 为状态序列，每个词性标记 c_i 对应HMM 中的一个状态 q_i , $S_c = c_1 c_2 \cdots c_n$ 。

$$\hat{S}_w = \operatorname{argmax}_{S_w} p(S_w | \mu)$$

详见第6章举例

7.5 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。

★ 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；
- ★ 模型实现复杂、计算量较大。

7.5 汉语自动分词基本算法

5. 由字构词（基于字标注）的分词方法 (Character-based tagging)

★ **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。

详见第6章举例

7.5 汉语自动分词基本算法

★ 评价:

该方法的重要优势在于，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。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]

6. 生成式方法与区分式方法的结合

★ 大部分基于**词**的分词方法采用的是**生成式模型**

(Generative model):

$$WSeq^* = \operatorname{argmax}_{WSeq} p(WSeq) = \operatorname{argmax}_{WSeq} p(WSeq|c_1^n)$$

✦ 使用 3-gram:

$$p(w_1^m) = \prod_{i=1}^m p(w_i|w_1^{i-1}) \approx \prod_{i=1}^m p(w_i|w_{i-2}^{i-1})$$

✦ 基于词的生成模型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。

7.5 汉语自动分词基本算法

★ 基于**字**的分词方法采用**区分式模型** (Discriminative model):

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

... .. c_{k-2} c_{k-1} c_k c_{k+1} c_{k+2}

↑
B, M, S, E

✦ 基于字的区分模型有利于处理集外词

7.5 汉语自动分词基本算法

★ 生成式模型与判别式模型的比较：

★ 生成（产生）式模型 (Generative Model)

- ✦ 假设 o 是观察值， q 是模型。
- ✦ 如果对 $p(o|q)$ 进行建模,就是生成式模型。
- ✦ **基本思想：** 首先建立样本的概率密度模型，再利用模型进行推理预测。要求已知样本无穷多或者尽可能地多。该方法一般建立在统计学和Bayes理论的基础之上。

7.5 汉语自动分词基本算法

- ✦ **主要特点：**从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度。
- ✦ **主要优点：**实际上所带的信息要比判别式模型丰富，研究单类问题比判别式模型灵活性强，模型可以通过增量学习得到，且能用于数据不完整 (missingdata) 情况。
- ✦ **主要缺点：**学习和计算过程比较复杂。

7.5 汉语自动分词基本算法

★ 判别（区分）式模型 (Discriminative Model)

- ✦ 如果对条件概率 $p(q|o)$ 进行建模，就是判别式模型。
- ✦ **基本思想是：**有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。表性理论为统计学习理论。
- ✦ **主要特点：**寻找不同类别之间的最优分类面，反映的是异类数据之间的差异。
- ✦ **主要优点：**判别式模型比生成式模型较容易学习。
- ✦ **主要缺点：**黑盒操作，变量间的关系不清楚，不可视。

7.5 汉语自动分词基本算法

- ★ 基于字的区分模型有利于处理集外词
- ★ 基于词的生成模型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

✦ 可以将两者的优势结合起来。

★ 结合方法1:

- ✦ 将待切分字符串的每个汉字用 $[c, t]_i$ 替代, 以 $[c, t]_i$ 作为基元, 利用语言模型选取全局最优(生成式模型)。

7.5 汉语自动分词基本算法

c_1	c_2	c_k	c_n
$[c_1, B]$	$[c_2, M]$	$[c_k, B]$	$[c_n, S]$
$[c_1, S]$	$[c_2, B]$	$[c_k, S]$	$[c_n, B]$
	$[c_2, S]$	$[c_k, E]$	$[c_n, M]$
	$[c_2, E]$	$[c_k, M]$	$[c_n, E]$

[上,B] [海,E] [计,B] [划,E] [到,S] [本,S] [世,B] [纪,E] ...

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

7.5 汉语自动分词基本算法

★ 实验结果:

[Ref. 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]

✦ 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: $P=95.0\%$

7.5 汉语自动分词基本算法

★ 分析:

✦ 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) 具有更好的鲁棒性;
- (2) 相对于生成模型, 容易处理更多的特征。

7.5 汉语自动分词基本算法

★ 结合方法2：插值法把两种方法结合起来

$Score(t_k)$

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

(0.0 ≤ α ≤ 1.0)

生成得分

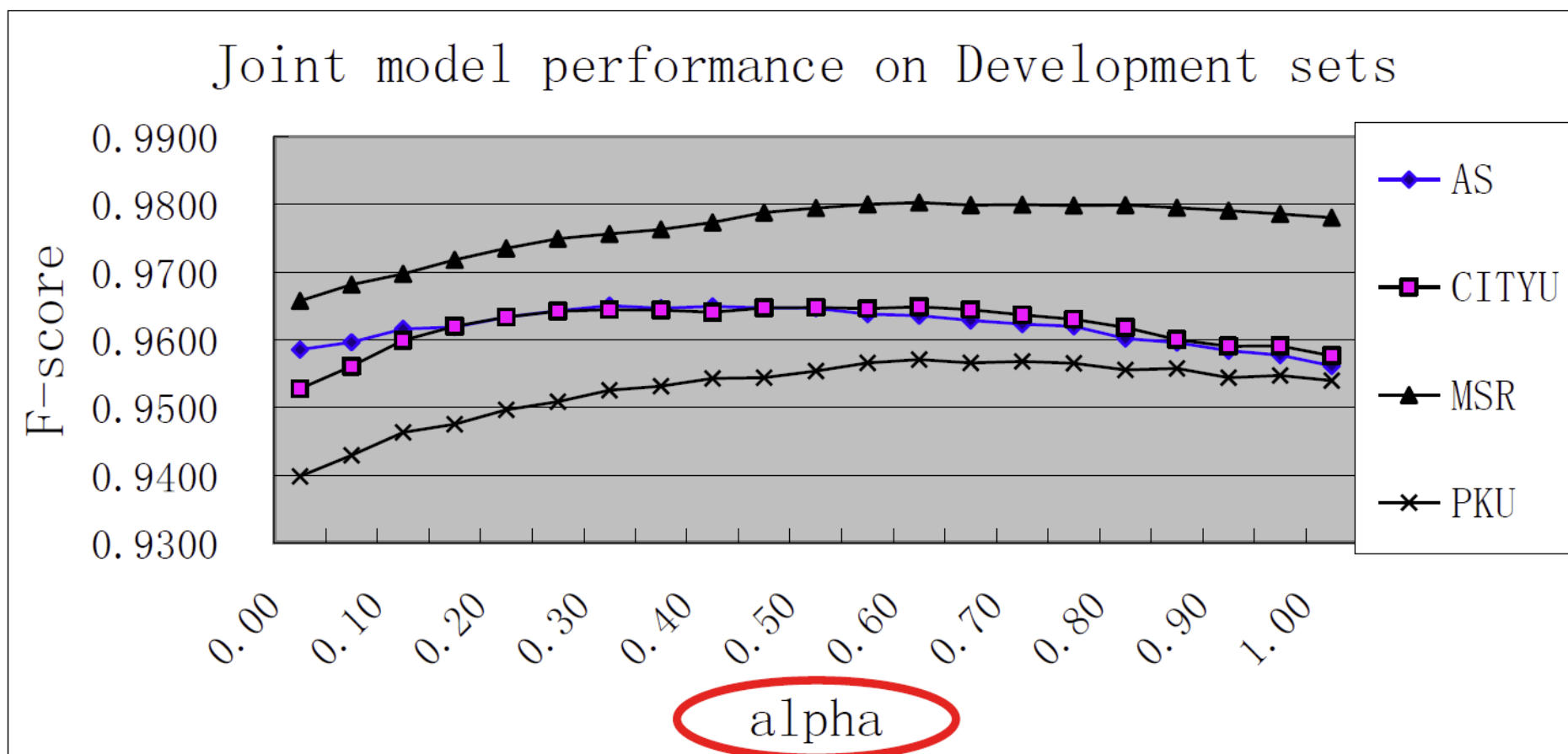
判别得分

✦ 优势：结合了基于字的生成模型和基于字的区分式模型的优点。

7.5 汉语自动分词基本算法

★ 性能测试

语料：2005 年 SIGHAN Bakeoff 语料，取少量做开发集。



7.5 汉语自动分词基本算法

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	0.95	0.956	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	0.954	0.691	0.979
MSR	Generative	0.974	0.967	0.97	0.561	0.985
	Discriminative	0.957	0.962	0.96	0.719	0.964
	Joint	0.974	0.971	0.972	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	0.62	0.941
	Joint	0.935	0.946	0.941	0.561	0.958

7.5 汉语自动分词基本算法

Corpus	Model	<i>R</i>	<i>P</i>	<i>F</i>	<i>R_{oov}</i>	<i>R_{IV}</i>
PKU converted (ucvt.) 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	0.956	0.616	0.966
Overall	Generative	0.953	0.946	0.95	0.511	0.973
	Discriminative	0.944	0.95	0.947	0.68	0.956
	Joint	0.957	0.955	0.956	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。

7.5 汉语自动分词基本算法

★ 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	0.94	0.94	0.757	0.974
Medicine	C	0.11	0.93	0.917	0.923	0.674	0.961
Finance	D	0.087	0.957	0.956	0.957	0.813	0.971

★ Ref

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

7.5 汉语自动分词基本算法

★ Urheen 汉语自动分词系统:

<http://www.nlpr.ia.ac.cn/cip/software.htm>

★ 其他分词方法

- ✦ 全切分方法
- ✦ 串频统计和词形匹配相结合的分词 方法
- ✦ 规则方法与统计方法相结合
- ✦ 多重扫描法
- ✦

★ 方法比较

- ① **最大匹配分词**算法是一种简单的基于词表的分词方法，有着非常广泛的应用。这种方法只需要最少的语言资源（仅需要一个词表，不需要任何词法、句法、语义知识），程序实现简单，开发周期短，是一个简单实用的方法，但对歧义字段的处理能力不够强大。

7.5 汉语自动分词基本算法

② **最短路径分词方法**的切分原则是使切分出来的词数最少。

这种切分原则多数情况下符合汉语的语言规律，但无法处理例外的情况，而且如果最短路径不止一条时，系统往往不能确定最优解。

③ **统计方法**具有较强的歧义区分能力，但需要大规模标注(或预处理)语料库的支持，需要的系统开销也较大。

7.6 未登录词识别

7.6 未登录词识别

★ 命名实体(Named Entity, NE)

✦ 专有名词 占95%

- ▲ 人名（中国人名和外国译名）

- ▲ 地名

- ▲ 组织机构名

- ▲ 数字、日期、货币数量

✦ 其他新词

- ▲ 专业术语、新的普通词汇等。

★ 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
- ✦ [曹文洁，2002] 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。

★ 中文姓名识别的难点

- ✦ 名字用字范围广，分布松散，规律不很明显。
- ✦ 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，
 - ▲ 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，计划等；
 - ▲ 全名也是普通词汇，如：许可，温馨，高山，高升，高飞，周密。
- ✦ 缺乏可利用的启发标记。
 - ▲ 例如：祝贺老总百战百胜。 林徽因此时已经离开了那里。

★ 中文姓名识别方法

- ✦ 姓名库匹配，以姓氏作为触发信息，寻找潜在的名字。
- ✦ 计算潜在姓名的概率估值及相应姓氏的姓名阈值(threshold value)，根据姓名概率评价函数和修饰规则对潜在的姓名进行筛选。

7.6 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。用下列公式计算姓和名的使用频率:

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

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

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

- ✦ 字符串 $Cname$ 可能为姓名的概率估值:

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

7.6 未登录词识别

★ 确定阈值

✦ 姓氏 X 构成姓名的最小阈值:

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

★ 设计评估函数:

✦ 姓名的评价函数:

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

✦ 对于特定的姓氏 X 通过训练语料得到一阈值 β_X , 当 f 大于 β_X 时, 该识别的汉字串确定为中文姓名。

★ 使用修饰规则：

✦ 否定潜在的姓名

- ▲ 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。

✦ 确定潜在的姓名边界

- ▲ 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。
- ▲ 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。

✦ 校正潜在的姓名

- ▲ 依据：含重合部分的潜在姓名不可能同时成立。利用各种规则消除冲突的潜在姓名。

★ 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》(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。

7.6 未登录词识别

★ 基本资源

- ✦ 建立地名资源知识库：地名库、地名用字库、地名用词库
- ✦ 建立识别规则库：筛选规则、确认规则、否定规则

★ 基本方法

- ✦ 统计模型
- ✦ 通过训练语料选取阈值
- ✦ 地名初筛选
- ✦ 寻找可以利用的上下文信息
- ✦ 利用规则进一步确定地名

★ 中文机构名称的识别

★ 中文机构名称的构成

- ✦ 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。

★ 中文机构名称的类型

- ✦ 地名，如：北京大学，武汉大学
- ✦ 人名，如：中山大学，哈佛大学
- ✦ 学科、专业和部门系统，如：公安部，教育委员会
- ✦ 研究、生产或经营等活动的对象，如：软件研究所，卫星制造厂
- ✦ 上述情况的综合，如：白求恩医科大学
- ✦ 大机构、团体、组织和职业的名称，如：中国人民解放军洛阳外国语学院，中国发明家学会等
- ✦ 专造的机构名，如：复旦大学，四通公司，微软研究院
- ✦ 创办、工作的方式，如：某某股份公司，中央电视大学

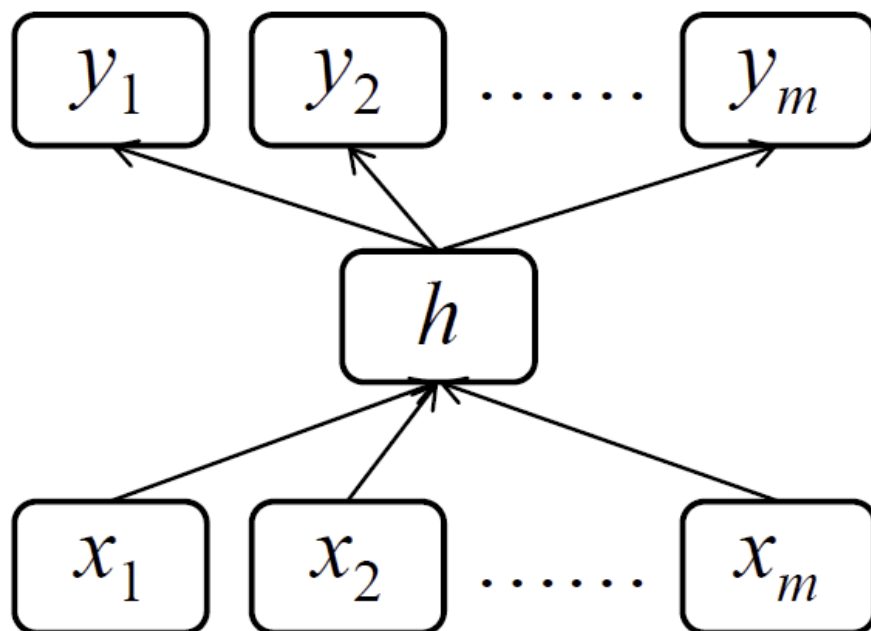
★ 机构名称识别方法

- ✦ 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
- ✦ 统计模型

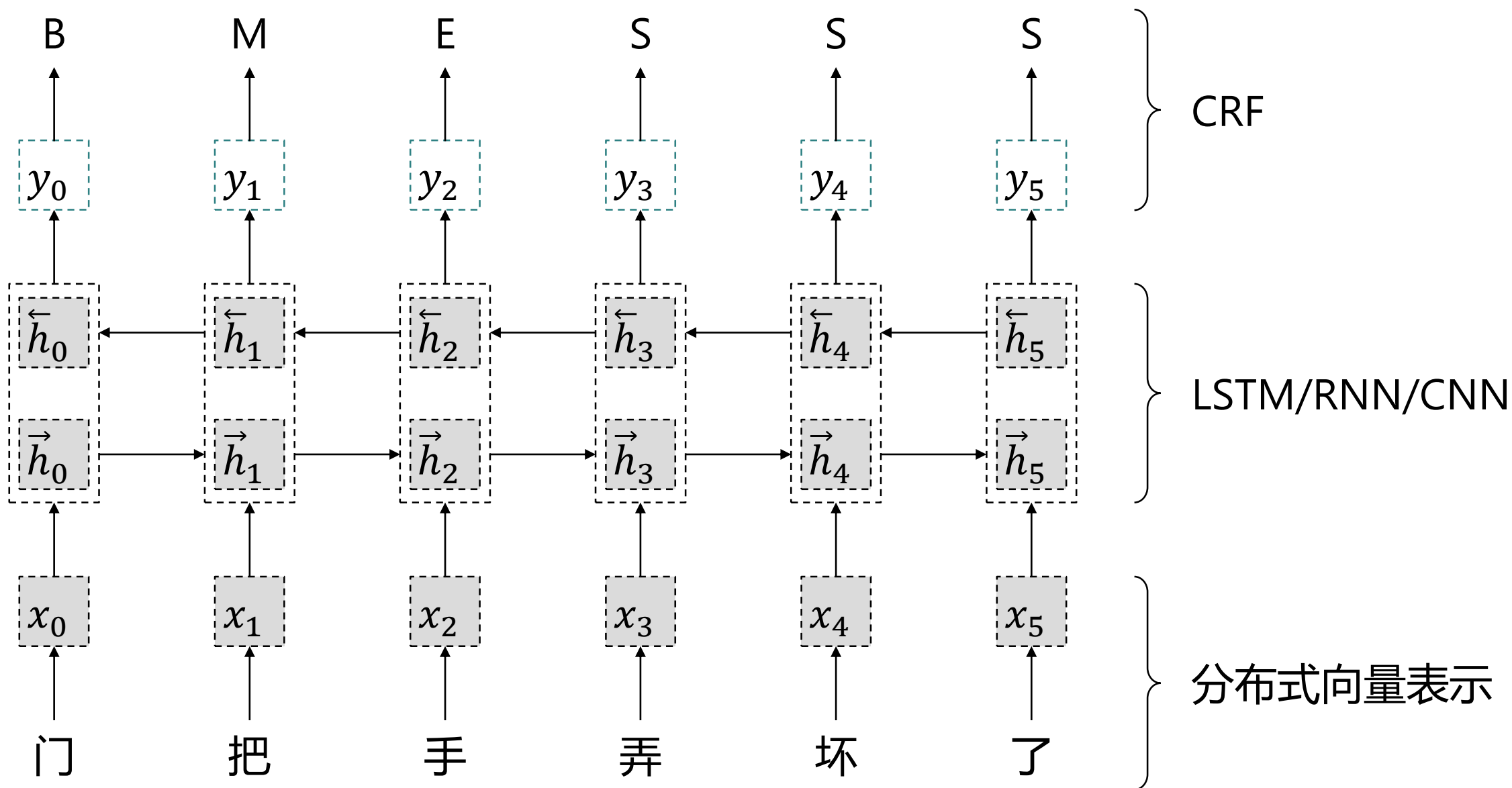
7.6 未登录词识别

★ 基于神经网络的命名实体识别方法

★ 把NER 看作序列标注任务，输入输出均为序列， many to many的对应关系。



7.6 未登录词识别



切分结果: 门把手 / 弄 / 坏 / 了

7.6 未登录词识别

★ 基于RNN的识别方法

- ✦ Y. Bengio, P. Simard, P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5, 157–166

★ 基于LSTM的识别方法

- ✦ S. Hochreiter, J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9, 1735–80 (1997)

★ 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。

7.7 词性标注概述

★ 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：

★ 在英语中

(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。汉语中常用词兼类现象严重，《现代汉语八百词》兼类占 22.5%。

7.7 词性标注概述

★ 在汉语中

(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) 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。

(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。

7.7 词性标注概述

★ 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。

7.7 词性标注概述

★ UPenn Treebank 的词性标注集确定原则:

- ✦ 可恢复性(recoverability): 从标注语料能恢复原词汇或借助于句法信息能区分不同词类;
- ✦ 一致性(consistency): 功能相同的词应该属于同一类;
- ✦ 不明确性(indeterminacy): 为了避免标注者在不明确的条件下任意决定标注类型, 允许标注者给出多个标记(限于一些特殊情况)。

[Marcus et al., 1993]

7.7 词性标注概述

★ 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。

7.7 词性标注概述

★ 北大计算语言学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)。

7.8 词性标注方法

7.8 词性标注方法

- ★ 基于规则的词性标注方法
- ★ 基于统计模型的词性标注方法
- ★ 规则和统计方法相结合的词性标注方法
- ★ 基于有限状态变换机的词性标注方法
- ★ 基于神经网络的词性标注方法

7.8 词性标注方法

★ 基于规则的词性标注方法

★ TAGGIT 词性标注系统 (Brown University)

- ✦ 86 种词性，3300 规则
- ✦ 手工编写词性歧义消除规则
- ✦ 机器自动学习规则

★ 山西大学的词性标注系统 [刘开瑛, 2000]

1、手工编写消歧规则

- ✦ 建立非兼类词典
- ✦ 建立兼类词典：词性可能出现的概率高低排列
- ✦ 构造兼类词识别规则

7.8 词性标注方法

(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：他们搞的这次大型(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 ?）

2、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， ...

7.8 词性标注方法

★ 基于错误驱动的机器学习方法 [E. Brill, 1992]

- ✦ 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

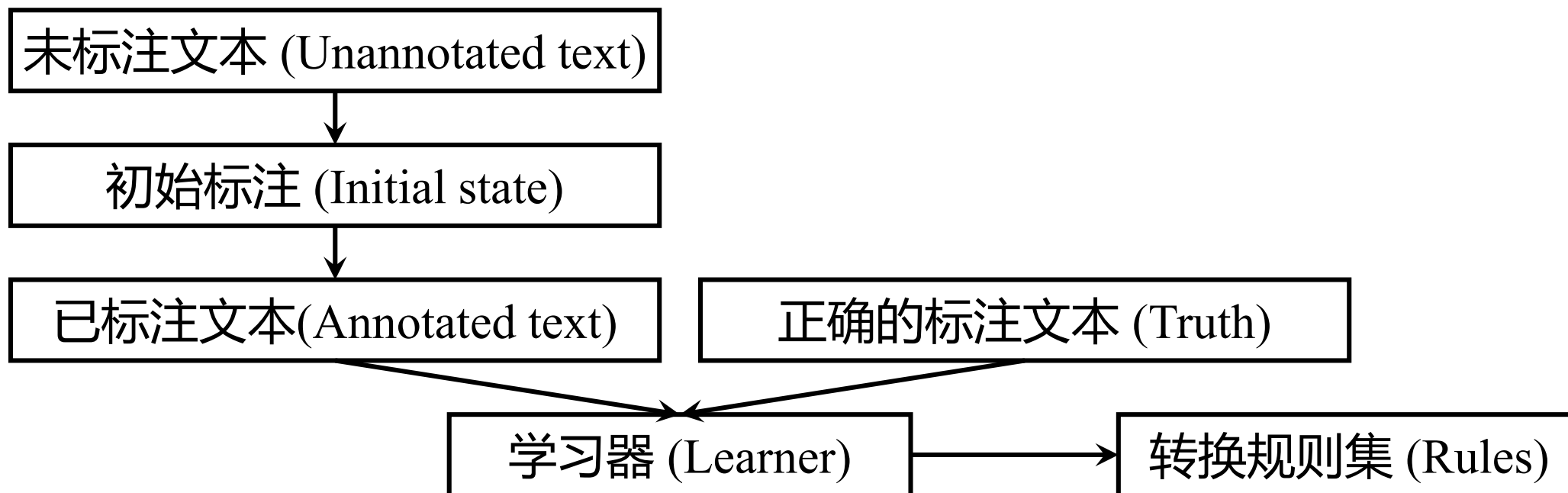


图. 基于转换规则的错误驱动的机器学习方法

7.8 词性标注方法

★ 基于 HMM 的词性标注方法

★ 规则和统计相结合的词性标注方法

- ✦ 规则消歧，统计概率引导
- ✦ 或者统计方法赋初值，规则消歧

7.9 下一步分词与词性标注 研究

7.9 下一步分词与词性标注研究

★ 当前分词技术存在的主要问题

- ✦ 分词模型过于依赖训练样本，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等对话文本、不同的专业领域(中医药、生物、化学、能源.....)。

★ 领域差异与陌生语言现象对现有方法提出巨大挑战

7.9 下一步分词与词性标注研究

★ 当前分词技术存在的主要问题

- ✦ 分词模型过于依赖训练样本，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等对话文本、不同的专业领域(中医药、生物、化学、能源.....)。

★ 领域差异与陌生语言现象对现有方法提出巨大挑战

7.9 下一步分词与词性标注研究

★ 举例：

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。

★ 分词准确率为：57.3% ~ 94.8%

★ **研究半监督学习、迁移学习等方法，解决领域的自适应问题，提高系统的鲁棒性和准确率。**

7.9 下一步分词与词性标注研究

★ 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

★ 根据对微博内容的统计，大约75%的内容为个人心情和感受方面的。

7.9 下一步分词与词性标注研究

★ 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)	1301320
微博用语词库	10330
网络用语大全	294
网络关键词以及词频数据	500000
人民日报微博词频统计	42315
百度百科对于网络用语的解释	1051
网络用语词典	541941 (经过合并筛选)
网络情感词典+传统情感词典 (情感词汇)	26207
词汇总数: 1753925 (经过合并筛选)	

7.9 下一步分词与词性标注研究

★ 补充后分词性能:

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

CWS: Chinese word segmentation based on ME model

★ 关于词性标注

- ✦ 进一步研究消歧方法，与其他技术相结合（如分词、句法分析等），提高性能
- ✦ 在有些任务或方法中，词性作用并不大，如基于词的统计机器翻译、目前的神经网络机器翻译等

★ 词法分析的任务（英语汉语有所不同）

★ 英语形态分析

- ✦ 单词识别
- ✦ 形态还原

★ 汉语自动分词

- ✦ 汉语分词中的主要问题

- ✦ 基本原则和辅助原则

- ✦ 几种基本方法：MM、最少分词法、统计法等

★ 未登录词识别

- ✦ 人名、地名、组织机构名、数字实体、特殊符号等

★ 词性标注

- ✦ 问题(兼类、标注集、规范)
- ✦ 方法(规则方法、统计方法、综合方法)

★ 分词与词性标注结果评测

- ✦ 正确率
- ✦ 找回率
- ✦ F-测度值

★ 分词与词性标注下一步努力的方向