自然语言理解

第六章 词法分析与词性标注

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

No.95, Zhongguancun East Road
Beijing 100080, China

http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

第六章 词法分析与词性标注

6.1 概述

词是自然语言中能够独立运用的最小单位，是信息处理的基本单位。

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、类别和性质。

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。

6.1 概述

❑ 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，词根和词干语词的附加成分集合紧密。

词法分析：词的形态变化分析，即词的形态还原。

分析语（孤立语）（如，汉语）：分词。

黏着语（如，日语）：分词＋形态还原

6.2 英语的形态分析

- 基本任务
  - 单词识别
  - 形态还原

6.2 英语的形态分析

❑ 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/.

6.2 英语的形态分析

❑ 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 => 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

(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

6.2 英语的形态分析

(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**)

6.2 英语的形态分析

❑ 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)

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)

6.2 英语的形态分析

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)

4) -ly **结尾的副词**

*ly → * (e.g., hardly → hard)

… …

6.2 英语的形态分析

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'

6.2 英语的形态分析

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

6.2 英语的形态分析

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) 82th →**去掉** th **后，记录该数字为序数词；**

3) $200 → **去掉**$，**记录该数字为名词（**200**美圆）；**

4) 98.5% → 98.5**%作为一个数词**

6.2 英语的形态分析

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。**

6.2 英语的形态分析

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) **其它带连字符“-”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。**

6.2 英语的形态分析

❑ 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。

6.3 汉语自动分词概要

❑ 汉语自动分词的重要性

- ➢ 自动分词是汉语句法分析的基础
- ➢ 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

6.3 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

– 汉语中什么是词？两个不清的界限：

1）单字词与语素（词素）

2）词与短语

如，花草，湖边，房顶，鸭蛋，小鸟，担水，一层，翻过？

6.3 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

（2）门把手弄坏了。　　（组合型歧义）

门/ 把/ 手/ 弄/ 坏/ 了/ 。

门/ 把手/ 弄/ 坏/ 了/ 。

6.3 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、外国译名、术语等

  如：盛中国，张建国，蔡国庆，林徽因，彭太发生，平川三太郎，约翰·斯特朗，詹姆斯·埃尔德等。

  （2）个别俗语、方言等

6.3 汉语自动分词概要

❑ 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）

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）

6.3 汉语自动分词概要

（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）

i) 字串的语法功能不符合组合规律，如：好吃，好喝，好听，好看等

ii) 字串的内部结构不符合语法规律，如：游水等

6.3 汉语自动分词概要

❑ 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

6.3 汉语自动分词概要

❑ 汉语自动分词的辅助原则（续）

2) 附着性语（词）素和前后词合并为一个分词单位（合并原则）。

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。

6.3 汉语自动分词概要

❑ 汉语自动分词的辅助原则（续）

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）。

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）；

“春夏秋冬”、“轻重缓急”、“男女”（并列）；

“象牙”（名词偏正）；“暂不”、“毫不”、“不再”、“早已”（副词并列）等

6.3 汉语自动分词概要

❑ 汉语自动分词的辅助原则（续）

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）。

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”、“垃圾车、交通车、午餐车”、“立足点、共同点、着眼点”等。

6.3 汉语自动分词概要

❑ 汉语自动分词的辅助原则（续）

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。

6.3 汉语自动分词概要

❑ 汉语自动分词的辅助原则（续）

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）。

i) 词组带接尾词：太空/ 计划/ 室、塑料/ 制品/ 业

ii) 动词带双音节结果补语：看/ 清楚、讨论/ 完毕

iii) 复杂结构：自来水/ 公司、中文/ 分词/ 规范/ 研究/ 计划

iv) 正反问句：喜欢/ 不/ 喜欢、参加/ 不/ 参加

6.3 汉语自动分词概要

v) **动宾结构、述补结构的动词带词缀时：写信/ 给、取出/ 给、穿衣/ 去**

vi) **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**

vii) **专名带普通名词：胡/ 先生、京沪/ 铁路**

6.4 汉语自动分词基本算法

- 有词典切分 / 无词典切分
- 基于规则分析方法 / 基于统计方法

6.4 汉语自动分词基本算法

❑ 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）

句子：

$$S = c_1 c_2 \cdots c_n$$

假设词： $w_i = c_1 c_2 \cdots c_m$ m **为词典中最长词的字数。**

6.4 汉语自动分词基本算法

◆ FMM算法描述

0）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：

1）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，if n=1，转3)。否则，令 m=词典中最长单词的字数， if n < m, m = n；

2）从当前指针 p_i 起取m个汉字作为词 w_i，作如下判断：

i) 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 iii)；

ii) 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2)中的 i) 步；否则（即 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 iii)；

iii) 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3），否则， i=i+1，返回 1）；

3）输出切分结果，结束分词程序。

6.4 汉语自动分词基本算法

◆ **举例**

设词典中最长单词的字数为7。

输入字串：他是研究生物化学的。

切分过程：他是研究生物化学的。

p ↑ |

… …

他 || 是研究生物化学的。

p↑ |

FMM **切分结果：他 || 是 || 研究生 || 物化 || 学 || 的 ||。**

BMM **切分结果：他 || 是 || 研究 || 生物 || 化学 || 的 ||。**

6.4 汉语自动分词基本算法

◆ 评价

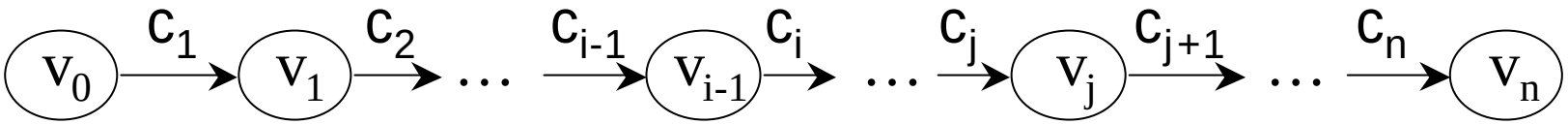
- ➢ 优点：
  - ？程序简单易行，开发周期短；
  - ？仅需要很少的语言资源（词表），不需要任何词法、句法、语义资源；
- ➢ 弱点：
  - 切分歧义消解的能力差；
  - 切分正确率不高，一般在95%左右。[李东，2003]

6.4 汉语自动分词基本算法

❑ 最少分词法（最短路径法）

◆ 基本思想

设待分字串 $S=c_1\ c_2\ldots\ldots c_n$，其中$c_i$（$i=1,2,\ldots n$）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，…，V_n。



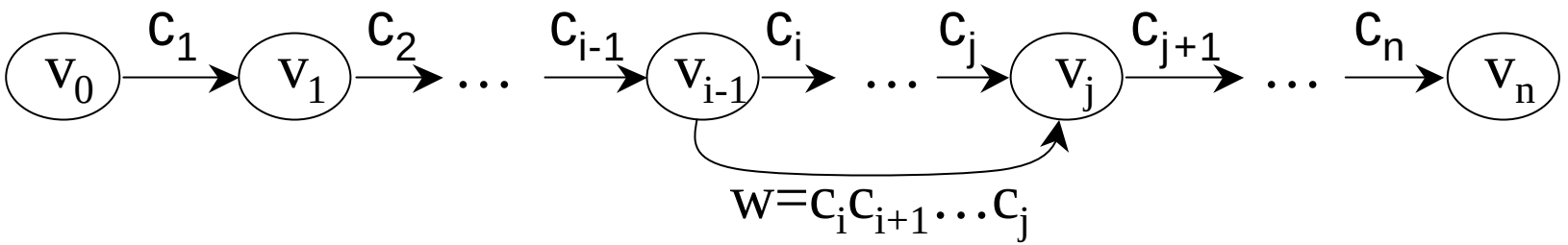
求N-最短路径：贪心法或简单扩展法。

6.4 汉语自动分词基本算法

◆ **算法：**

(1) **相邻节点** v_{k-1}, v_k **之间建立有向边** $<v_{k-1}, v_k>$，**边对应的词默认为** c_k (k =1, 2, …, n)。

(2) **如果** $w= c_i c_{i+1} \ldots c_j$ $(0<i<j\leq n)$ **是一个词，则节点**v_{i-1}, v_j **之间建立有向边** $<v_{i-1}, v_j>$，**边对应的词为** w。



(3) **重复上述步骤**(2)，**直到没有新的路径（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。**

6.4 汉语自动分词基本算法

◆ 举例

1) 输入字串：他只会诊断一般的疾病。

可能的输出：他||只会||诊断||一般||的||疾病。 (6)

他||只||会诊||断||一般||的||疾病。 (7)

… …

最终结果：他||只会||诊断||一般||的||疾病。

2) 输入字串：他说的确实在理。

可能的输出：他||说||的||确实||在理。 (5)

他||说||的确||实在||理。 (5)

… …

6.4 汉语自动分词基本算法

◆ 评价

➢ 优点：

- 采用的原则（切分出来的词数最少）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。

6.4 汉语自动分词基本算法

❑ 基于统计模型的分词方法

◆ **基本思想**：利用字与字之间和词与词之间的同现概率作为分词的依据。

◆ **方法描述**：

设待切分的汉字串为：$S = c_1\, c_2 \ldots\ldots c_n$，其中$c_i$（$i$ =1,2,…n）为单个的字，n 为字串的长度，$n \geq 1$。$W = w_1 w_2 \ldots\ldots w_k$ （$1 \leq k \leq n$）是一种可能的切分。

$$\hat{W} = \arg\max_W P(W \mid S) = \arg\max_W \frac{P(W)P(S \mid W)}{P(S)}$$

$$= \arg\max_W P(W) = \prod_{i=1}^{k} P(w_i \mid w_1, \cdots, w_{i-1})$$

6.4 汉语自动分词基本算法

- 评价
  - 优点：
    - 减少了很多手工标注知识库（语义词典、规则等）的工作
    - 在训练语料规模足够大和覆盖领域足够多的情况下，可以获得较高的切分正确率
  - 弱点：
    - 训练语料的规模和覆盖领域不好把握
    - 计算量较大

6.4 汉语自动分词基本算法

❑ 其它方法及多种方法相结合的混合分词方法

- ◆ 串频统计和词形匹配相结合的分词方法 [刘开瑛，2000]
  - ➢ 第一遍扫描：利用切分标记将文本切分成汉字短串；
  - ➢ 第二遍扫描：根据各短串的每个子串在上下文中的频度计算计算其权值，权值大的子串视为候选词；
  - ➢ 第三遍扫描：利用候选词集和一部分常用词词典对汉字短串进行切分。
  - ➢ 四个词典：单字数词词典，单字量词词典，静态常用词词典，临时词典
  - ➢ 两个库：切分标记库和词缀库

6.4 汉语自动分词基本算法

- 规则方法与统计方法相结合
- 最短路径法与统计方法相结合
- 词性标注与分词一体化的处理方法

6.5 汉语自动分词中的歧义消除方法

❑ 交叉歧义表现形式

交叉歧义又称交集型歧义，即汉字串 ABC 既可以切分成 AB||C 型，又可以切分成 A||BC型，即AB是词，BC也是词。如：上游泳课，上游，游泳

根据有关专家统计，平均每1000字有16次的交集型字段出现，其中，三字两词（链长为1）字段和四字三词（链长为2）字段的占了歧义字段总数的95%左右 [刘开瑛，2000；郑家恒，1997]。

6.5 汉语自动分词中的歧义消除方法

❑ 交叉歧义消除方法之一－统计方法

一般而言，交集型歧义字段中伪歧义字段的切分结果不随实际语境的变化而变化，仅用字段内部信息就可以作出判断。

1）链长为1的交集型歧义字段切分策略

（1）建立独立成词字段的频度库

（2）对于歧义字段ABC，如果AB的频度＋C的频度>A的频度＋BC的频度，则将ABC切分成AB和C，否则，切分成A和BC。

如：“本地区”：f(本)＝3，f(地区)＝600，f(本地)＝0，f(区)＝14。于是，“本||地区”。

6.5 汉语自动分词中的歧义消除方法

2）链长为2的交集型歧义字段切分策略

对于链长为2的歧义字段ABCD，直接切分成AB||CD。

如："献出自己的爱心"：献出||自己的爱心

"洪水已经过去"：洪水已经||过去

3）链长为3的交集型歧义字段切分策略

对于链长为3的歧义字段ABCDE，首先切分成ABC||DE，然后将ABC作为链长为 1 的交集型歧义字段处理。

如："奉献出自己的爱心"：奉献出||自己的爱心

6.5 汉语自动分词中的歧义消除方法

❑ 交叉歧义消除方法之二－词性方法

根据汉语词语之间词性的约束和搭配关系，确定切分位置。如，链长为1的交集型歧义字段中，

◆ A(介词)|| BC 型

从||中国/中央/军事/政治/…

为||人们/重点/主导/…

在||理论/世界/场院/…

把||关系/手术/手表/手掌/…

盲点：从小学、以北约、把握住等

6.5 汉语自动分词中的歧义消除方法

- AB|| C (介词) 型

  如：创建/存在/有利/应用/适用||于(介词)

盲点：从属于

对于切分中的特例，可以采用特殊规则处理 [侯敏，1995]。

6.5 汉语自动分词中的歧义消除方法

❑ 交叉歧义消除方法之三
—利用汉字间的二元关系 [孙茂松，1997]

◆ 定义：

对于有序汉字串xy，汉字x, y之间的互信息由如下估值公式计算：

$$I(x:y)=\log_2\frac{N\times r(x,y)}{r(x)\times r(y)}$$

其中，N 为训练语料的总字数，r(x, y) 为 x, y 邻接同现的次数，r(x), r(y) 分别为 x, y 独立出现的次数。

注意：这里定义的互信息与前面我们提到的信息论中互信息的定义不一样。

6.5 汉语自动分词中的歧义消除方法

- **方法：**

  如果歧义字段 xyz **有两种可能的切分：** Pt_1: xy||z, Pt_2: x||yz

  如果 $I(x:y) - I(y:z) \geq \alpha$，**则取** Pt_1 **切分；**

  $I(y:z) - I(x:y) \geq \alpha$，**则取** Pt_2 **切分。**

6.5 汉语自动分词中的歧义消除方法

❑ 组合歧义及其消除方法

◆ 组合歧义表现形式

汉字段 AB 既可以独立成词，又可以切分成 A 和 B 两个词。

如：1）他学会了解方程。

2）这不是一个人的问题。

3）学生会写文章。

4）他将来北京。

5）他现在一家公司上班。

6.5 汉语自动分词中的歧义消除方法

◆ 组合歧义消除方法

（1）真对性歧义消除规则

如：M + AB -> M + A(q) + B

(如果直接前趋词为数词时，量词A和B应分开，如：家人，个人，人为，多元，阵风等)

（2）针对性歧义消除规则与句法分析同时进行

（3）统计方法

6.6 汉语自动分词中的未登录词识别

- 专用名词
  - 中国人名
    - 中国地名、组织机构名称
    - 外国译名
- 其它新词

  如：非典（SARS），酷(cool)等。

6.6 汉语自动分词中的未登录词识别

❑ 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，三字姓 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：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，2002a, 2002b]

6.6 汉语自动分词中的未登录词识别

❑ 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，如，姓氏：于（介词），张（量词），江（名词）等；名字：建国，国庆，胜利，文革等。
- 缺乏可利用的启发标记

例如: (1) 祝贺老总百战百胜。

(2) 林徽因此时已是两个孩子的母亲。

(3) 赵微笑着走了。

(4) 南京市长江大桥。

6.6 汉语自动分词中的未登录词识别

❑ 中文姓名识别方法

◆ Step-1: **姓名库匹配，以姓氏作为触发信息，寻找潜在的名字**

◆ Step-2: **计算潜在姓名的概率估值及相应姓氏的姓名阈值，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。**

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：

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

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

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

6.6 汉语自动分词中的未登录词识别

字串 *Cname* **可能为姓名的概率估值：**

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

姓氏 X **构成姓名的最小阀值：**

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

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** b_X (threshold value)，**当** f **大于** b_X **时，该识别的汉字串确定为中文姓名。**

6.6 汉语自动分词中的未登录词识别

阈值 b_X **定义为：** $b_X = a_X \times \ln(T_{\min}(X))$

调整因子 $a_X = \dfrac{1+\sqrt{F(X)}}{2}$

显然， $0.5 < a_X \leq 1$

其含义为：对于文本中任意一个满足 $\mathrm{Cname} = \mathrm{X}m_1 m_2$ **的汉字串，如果** $\mathrm{F(X)}=100\%$ $(\alpha_X=1)$ **且** $\ln P(Cname) > \ln(T_{\min}(X))$ **则该字串被认定为姓名。如果** $\mathrm{F(X)}\neq 100\%$**，但** $\ln P(Cname) > a_X \times \ln(T_{\min}(X))$ **该汉字串也被认定为姓名。**

6.6 汉语自动分词中的未登录词识别

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。

- Step-3: **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。**

6.6 汉语自动分词中的未登录词识别

- Step-4: **校正潜在的姓名**

  依据：含有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。

6.6 汉语自动分词中的未登录词识别

❑ 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》（1994）收集88026个，不包括相当一部分街道、胡同、村庄等小地方名称。

➢ 真实语料中地名出现情况复杂。如地名简称、地名用词与其它普通词冲突、地名是其它专用名词的一部分，地名长度不一等。

—[刘开瑛，2000], [孙茂松，1995a]

6.6 汉语自动分词中的未登录词识别

- 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

6.6 汉语自动分词中的未登录词识别

- **基本方法**
  - **概率估值计算**
  - **通过训练语料选取阈值（threshold value）**
  - **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**

6.6 汉语自动分词中的未登录词识别

❑ 中文机构名称的识别方法

◆ 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。

6.6 汉语自动分词中的未登录词识别

- **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、专业合部门系统，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

6.6 汉语自动分词中的未登录词识别

- 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

6.6 汉语自动分词中的未登录词识别

❑ 英语译名的识别方法

- ◆ 英语译名用字资源 [孙茂松，1993]
  - ➢ 英语姓名译名用字表
  - ➢ 只能出现在译名首的用字表
  - ➢ 不能出现在译名首的用字表
  - ➢ 只能出现在译名用字尾的用字表
  - ➢ 不能出现在译名用字尾的用字表
- ◆ 潜在译名粗界定，首尾逼近分析

6.6 汉语自动分词中的未登录词识别

❑ 专名识别方法的改进

◆ 基于大规模真实语料的统计方法

- HMM-Based [Zhou, 2002]
- Maximum-Entropy [Collins, 2002]

上述方法主要真对英语名词实体（Named Entity, NE）的识别，用于汉语专名识别的尚不多见。

6.6 汉语自动分词中的未登录词识别

❑ 其它新词识别

- ◆ 利用统计和启发知识的方法 [Nie, 1995]
  - ➢ 基于分词词典和规则方法的 MM 法分词处理
  - ➢ 新词识别：1) n 元组生成
    2) 剔除噪声
    3) 消除 n 元重叠
- ◆ 统计信息与规则相结合 [Chang, 2002]
  - ➢ 词法、句法、语义、世界知识以及上下文信息
  - ➢ 基本思想：连续出现的两个关联紧密的词素合并
- ◆ 局部统计方法 [Chang, 2003], [沈达阳, 1997]

6.7 词性标注概要

词性标注（POS tagging）的主要任务是消除词性兼类歧义。词性兼类在任何一种自然语言中都普遍存在。

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%。

6.7 词性标注概要

◆ **汉语中的词性兼类现象：** [赵铁军，2001]

1）形同音不同，如：“好（hao3，形容词）、好（hao4，动词）”

2）同形、同音，但意义毫不相干，如：“会（会议，名词）、会（能够、动词）”

3）具有典型意义的兼类词，如：“典型（名词或形容词）”

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

6.8 词性标注集定义

❑ 标注集的确定原则：

不同语言的语法中，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体系；
- ➢ 兼容性：与已建立的资源标记尽量一致，或可转换；
- ➢ 可扩展性：扩充或修改；

6.8 词性标注集定义

◆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者借助于句法信息能够区分不同的词类；**
- **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]

6.8 词性标注集定义

◆ UPenn Treebank **中的英语词性标注集：**

NN	**名词（单数或不可数）**
NNS	**名词复数**
NNP	**专用名词（单数）**
NNPS	**专用名词（复数）**
VB	**动词**

… …

共计：30种标记符号

6.8 词性标注集定义

- **汉语词性标注集：**

  北大：名词n、时间词t、处所词s … …　　39 种标记符号

  山西大学：基本词类17类

  … …

  百花齐放

6.9 词性标注方法

❑ 基于规则的汉语词性标注方法

◆ 规则消歧法

- ➢ 非兼类词典
- ➢ 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 —[刘开瑛，2000]
  - ？并列鉴别规则

    如：体现了人民的<u>要求</u>（N/V ?）和愿望（N，非兼类）。

6.9 词性标注方法

- 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 ?）。

- 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。

- 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 ?）

6.9 词性标注方法

- 根据词语的结构建立词性标注规则
  - 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，…
    - 其它：年轻化，知识化，…{化}
      篮球赛，足球赛，…{赛}
      … …
  - 重叠词规则
    - 看看，瞧瞧，高高兴兴，热热闹闹，…

6.9 词性标注方法

- **基于错误驱动的词性标注方法**
  - **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**

—[Brill, 1992]

6.9 词性标注方法

❑ 基于统计模型的汉语词性标注方法

◆ 基于 *n*-gram 的语言模型

汉字串：*Text*

词性标注符号串：*Tags*

$$\widehat{Tags} = \arg\max_{Tags} P(Tags \mid Text) = \arg\max_{Tags} P(Tags)$$

应用系统：1）1983年 Mashall 提出的 LOB 语料库的标注系统: CLAWS (Constituent-Likelihood Automatic Word-tagging System)

2）DeRose 对 CLAWS 改进后 VOLSUNGA 系统 (bi-gram)。

6.9 词性标注方法

- **基于 HMM 的词性标注方法**
  - **状态集**
  - **输出符号**
  - **初始状态概率**
  - **状态转移概率**
  - **符号输出概率**

—[Manning, 2001] pp. 357-359:

. Jelink's Method

. Kupier's Method

6.9 词性标注方法

- **规则和统计相结合的汉语词性标注方法**

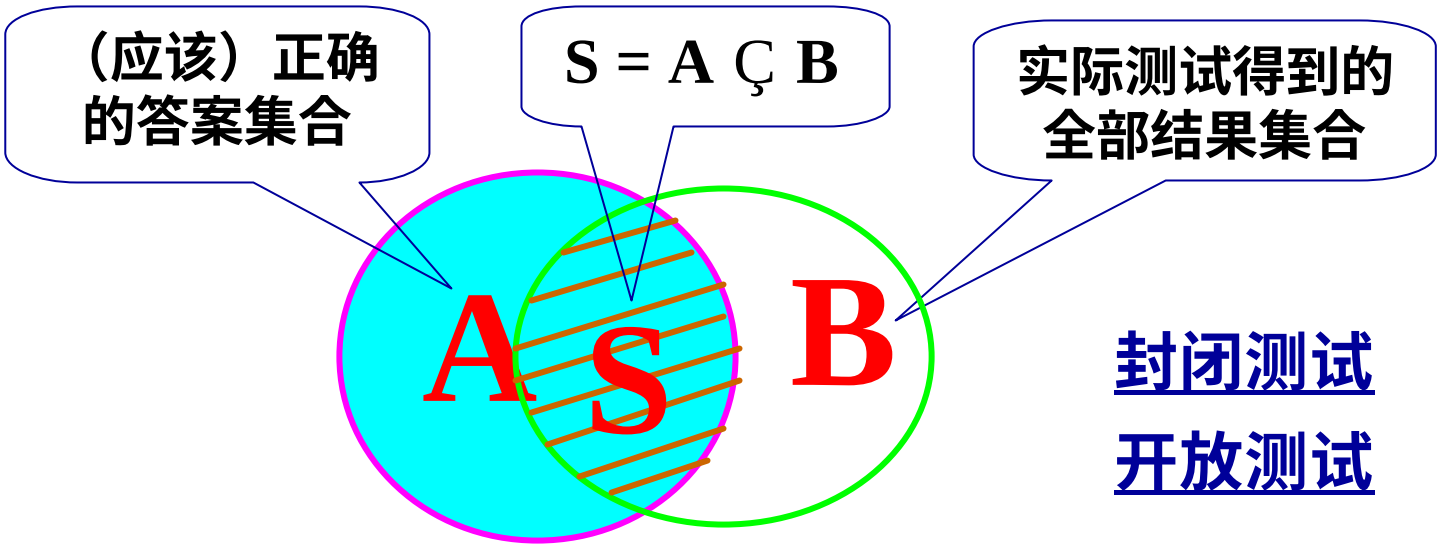
  规则消歧，统计概率引导

  或者统计方法赋初值，规则消歧

  —[周强，1995；张民，1998，荀恩东，1999]

6.10 汉语自动分词结果评测

❑ 正确率与召回率



封闭测试

开放测试

正确率(Correct ratio): $C = \frac{S}{B} \times 100\%$

召回率(Recall ratio): $R = \frac{S}{A} \times 100\%$

F-**测度**：

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

其中，$b = 1, \ldots$

本章小结

- 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
  - 基本原则合辅助原则
  - 几种基本方法（MM, 最少分词法，统计法）
- 汉语分词中的歧义消除策略
  - 交叉歧义消除（统计法，词性法，汉字间二元关系法）
  - 组合歧义消除

本章小结

- 汉语分词中的专名识别
  - 姓名（概率估值）
  - 地名
  - 机构名
  - 译名
  - 其它新词识别
- 词性标注的意义和复杂性
- 词性标注集确定的基本原则

本章小结

- 词性标注的基本方法
  - 规则方法
  - 统计方法
  - 规则和统计相结合的方法
- 分词及词性标注结果评价方法
  - 正确率
  - 召回率
  - F-测度

习题

1. 设计并实现算法用于还原英语动词。
2. 设计一个有限状态自动机用于识别缩写 {he, she}'s 是 he / she has 还是 he / she is，并编写程序实现该自动机。
3. 编写程序实现汉语逆向最大分词算法（可采用有限词表），并利用该程序对一段中文文本进行分词实验，校对切分结果，计算该程序分词的正确率、召回率及F-测度。
4. 设计并实现一个汉语未登录词的识别算法（可限定条件），并通过实验分析该算法的优缺点。

习题

5. 了解目前常见的几种汉语词性标注集，比较它们的差异，并阐述你个人的观点。
6. 掌握各种词性标注方法的要点，了解目前汉语词性标注的几种主要方法。
7. 试参考前人的工作，提出消除汉语自动分词中组合歧义的几点设想。
8. 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。

Thanks

谢谢!