7.5 自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM) 词典切分, 机械切分.

- 正向 FMM
- 逆向 BMM
- 双向 Bi-directional MM

例: $S = c_1c_2 \cdots c_n$

某词: $w_i = c_1c_2 \cdots c_m$　m为最大字数 (最长词的)

简单易行, 无需词法等知识

正确率 ?%, 歧义消解的能力差.

m = 7, 6, 5 … 挨个去比对.

算法描述:

(1) 令 i=0, 当前指针 p_i 指向输入字串初始位置

(2) 计算 p_i 到字串末端的字数 n, n=1, 转(4).

否则 令 m = 阈值

(3) 从 p_i 起取 m 个汉字作为词 w_i, 判断

a) w_i 是词典中的词, 在 w_i 后添加一个切分标志, 转 (c).

b) w_i 不是词典中的词且 w_i 的长度大于1, 将 w_i 从右边去掉一个字, 转

a) 步, 否则 …

c) 根据 w_i 的长度修改 p_i 位置, 如 p_i 指向字串末尾转(4), 否则 i=i+1, 转 (2).

(4) 输出结果

2. 最少分词法 (最短路径法)　(词数最少的切分结果)

设待切分字串 $S = c_1c_2 \cdots c_n$, 其中 c_i 为单个的字, n 为串的长度, 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 $V_0, V_1, V_2, \ldots, V_n$

例:

求出路径中选择词数最少的

算法描述: (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 ($k=1, 2, \ldots, n$)

(2) $w = c_i c_{i+1} \cdots c_j$

V_{i-1}, V_j 建立有向边 $<V_{i-1}, V_j>$

(3) repeat (2) 直至无新词产生.

(4) 输出结果.

优: 符合汉语自然规律

缺: 歧义

长度相同的路径候选增加

第7章 词法分析与词性标注

词：NLP的基本单位

自动词法分析：分析形态等

词性/词类

- 曲折语：形态分析
- 分析语：分词
- 黏着语：分词+形态还原

7.2 英语的形态分析

7.2.1 单词识别 → 自动机 随着标准模板的增大，逐渐不再重要.

7.2.2 形态还原 → 查字典，按规则还原

7.3 汉语自动分词

问题
- 单字词与词素 / 词与短语 } — 汉语分词规范
- 交集型歧义 / 组合型歧义 } — 歧义切分 (121 (概率)) 链长：共多字
- 人名、机构名等 / 新出现的词汇 } — 未登录词的识别 ☆

原则
- 合并原则：语义上不可相加而得 / 无法合并得到(不符合组合规律
- 辅助原则(灵活性)
  - 切分原则：有明显分隔词
  - 合并
    - 附着语素
    - 缩略语的字串
    - 双音节+单音节的偏正式
    - 双音节的偏正式
  - 切分：内部完全，合并冗长

7.4 分词与词性标注结果评价方法

正确率(P)：$P = \frac{n}{N} \times 100\%$

召回率(R)：$R = \frac{n}{M} \times 100\%$

Roov RIV

集外词 集内词

n：正确结果

N：系统输出.

M：标准答案的个数.

F-测度值：$F\text{-measure} = \frac{(\beta^2+1) \times P \times R}{\beta^2 \times P + R} \times 100\%$ β取1

第八章　　语法理论

8.1 功能合一文法 (FUG)

提出动因：chomsky 短语结构的生成能力太强，产生许多不合语法/有歧义的句子

改进：单一形式的结构描述 | 合一运算

复杂特征集的定义：设 A 为一个功能描述 FD，当且仅当 A 可以表示为：

$$\begin{pmatrix} f_1 = v_1 \\ f_2 = v_2 \\ \cdots \\ f_n = v_n \end{pmatrix} \quad n \geq 1$$

f_i：特征名

v_i：特征值

$A(f_i) = v_i \quad (i = 1, \cdots, n)$

可以用复杂特征集描述词汇

- cat = N
- sem = 电脑 work
- lex = 计算机

短语

- $S \rightarrow NP + Verb$
- cat = S

句子

- cat = S
- subject =

特点：允许利用多个特征描述同一个语言单位。

数量、层次、运算约束。

合一运算：

若 A，B 均为复杂特征集，则 A，B 是相容的，当且仅当：

1. 如果 $A(f) = a$，$B(f) = b$，且 a，b 都是原子的，则必 $a = b$，B 是相容的。

当且仅当 $a = b$

2. $A(f)$，$B(f)$ 相容。

运算定义：1）a, b 均为原子，如果 a=b，则 $a \sqcup b = a$，否则 $a \sqcup b = \phi$

2）α, β 均为复杂特征集，则

(1) 若 $\alpha(f)=v$，但 $\beta(f)$ 的值未经定义，则 $f=v$ 属于 $\alpha \sqcup \beta$

(2) 反之亦然。

(3) $\alpha(f)=v_1$，$\beta(f)=v_2$，v_1 与 v_2 相容（不相矛盾），

则 $f=(v_1 \sqcup v_2)$ 属于 $\alpha \sqcup \beta$，否则为 ϕ。

8.2 词汇功能语法（LFG）

- 成分结构层次（C结构）（上下文无关文法）
- 功能结构层次 f结构
  - 简单符号
  - 语义形式
  - f结构
  - 上述三种集合

$S \rightarrow NP \quad VP$

$(\uparrow SUBJ)=\downarrow \quad \uparrow=\downarrow$

标注

$\uparrow \downarrow$：自指变元元变量

注：一个属性只能有一个值

由 c → f 代数

8.3 GB理论（管辖约束理论）

- 规则系统
  - 词汇规则
  - 句类和转换规则
  - 语音规则
  - 语义规则
- 原则系统
  - X理论
  - 格理论 → 可移位
  - 管辖理论
  - θ理论
  - 约束理论
  - 控制理论
  - 界限理论

8.4 树连接语法 TAG

$G = \langle V_N, V_T, S, I_0, A_0 \rangle$

- 替换　　立体 → 底平
- 附加　　else：扁平

8.5 广义短语结构语法

GPSG

句法规则 → 树形结构 ← 特征制约系统

语义解释 → 树形结构

语义解释系统。