第5节 分词与词性标注

1. 概述

- 词法分析的主要任务是**词性标注**和**词义标注**
- 词性是词汇的基本属性。进行词性标注通常有基于规则和基于统计的两种方法。
- 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
- 词义标注的重点就是解决如何确定多义词在具体语境中的义项问题。标注过程中，通常是先确定语境，再明确词义

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**

2. 英语的形态分析

基本任务：**单词识别+形态还原**

2. 1 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/.

2.1.1 英语中常见的特殊形式的单词识别

- prof., Mr., Ms. Co., Oct. 等放入词典
- Let's / let's ⇒ let + us
- I'am ⇒ I + am
- {it, that, this, there, what, where}'s ⇒ {it, that, this, there, what, where} + is
- can't => can + not; won't => will + not
- {is, was, are, were, has, have, had}n't ⇒ {is, was, are, were, has, have, had} + not
- X've => X + have; X'll => X + will; X're => X + are
- he's => he + is / has => ?; she's => she + is / has => ?
- X'd Y => X + would (如果 Y 为单词原型)
  => X + had (如果 Y 为过去分词)

2. 2 英语单词的形态还原

2.2.1 有规律变化单词的形态还原

示例：-ed结尾的动词过去式

*ed → *　　(e.g., worked → work)

*ed → *e　　(e.g., believed → believe)

*ied → *y　　(e.g., studied → study)

- **-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)

- **-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)

- **–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)

- **动词、名词、形容词、副词不规则变化单词的形态还原**

例：choose, chose, chosen

axis, axes

bad, worse, worst

…

3. 汉语分词

自动分词是汉语句法分析的基础

主要问题：

- **汉语中什么是词?**

  单字词？词与短语？

  如，花草，湖边，房顶，鸭蛋，小鸟，担水，一层，翻过

3.1 歧义切分字段

- **组合型歧义(Combinatorial ambiguities)**

组合歧义可以理解为汉字串AB满足A, B, AB同时为词

门把手弄坏了
门/ 把/ 手/ 弄/ 坏/ 了/
门/ 把手/ 弄/ 坏/ 了/

他将来我校讲学
他/ 将/ 来/ 我/ 校/ 讲学
他/ 将来/ 我/ 校/ 讲学

“将来”、“现在”、“才能”、“学生会”等，都是组合型歧义字段

- **歧义切分字段**
- **交集型歧义(Overlapped ambiguities)**

组合歧义可以理解为汉字串AXB满足AX, XB同时为词

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

“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，都是交集型歧义字段

3.2 汉语自动分词基本规则

- **未登录词的识别**

未登录词即没有被收录在分词词表中但必须切分出来的词，包括各类专有名词（人名、地名、企业名等）、缩写词、新增词汇等等。文本中的人名、地名、组织结构名等命名实体通常是不可能在词典中穷尽列出的。对于这些词的识别，称为**命名实体识别（Named Entity Recongition）**。

■ **合并原则**

语义或语类上无法由组合成分直接得到的应该合并为一个分词单位。

比如：好吃、好看、好听、进出口、或多或少、六月、邮递员、现代化...

■ **切分原则**

- 有明显分隔符标记的应该切分

上、下课 → 上/ 下课
洗了个澡 → 洗/ 了/ 个/ 澡

- 内部结构复杂、合并起来过于冗长的词尽量切分

喜欢/ 不/ 喜欢、参加/ 不/ 参加
看/ 清楚、讨论/ 完毕
太空/ 计划/ 室、塑料/ 制品/ 业

3.3 汉语自动分词基本算法

- 有词典切分 / 无词典切分
- 基于规则分析方法 / 基于统计方法

3.3.1 基于规则分析方法

- 最大匹配法（Maximum Matching, MM）

  有词典切分，机械切分，按匹配的方向分为：

  - 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)**

FMM算法思想

1. 假定词典中最长的单词长度为m，从左至右取待分词的前m个字符串作为匹配字段。

2. 查找字典，如果字典中存在和匹配字段相同的词语，则匹配成功，否则去掉匹配字段的最后一个字符重新匹配

3. 重复以上过程直到匹配全部完成

BMM是FMM的逆向思维，匹配不成功，将匹配字段的最前一个字符去掉重新匹配

双向最大匹配法是将FMM和BMM的到的结果进行比较，从而决定正确的分词方法。定义的匹配规则如下：

- 如果正反向匹配算法得到的结果相同，我们则认为分词正确，返回任意一个结果即可。
- 如果正反向匹配算法得到的结果不同，选择分词数量较少的结果

示例：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/。

优点：

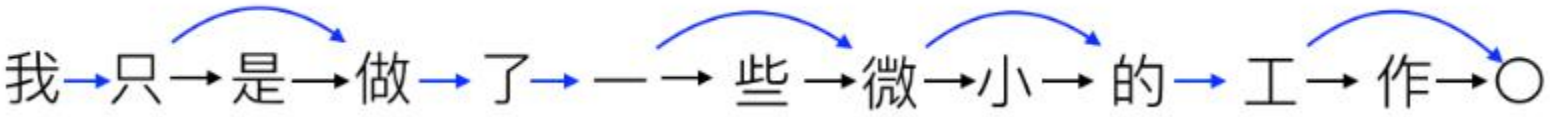
- 程序简单易行，开发周期短；
- 仅需要很少的语言资源（词表），不需要任何词法、句法、语义资源；

缺点：

- 切分歧义消解的能力差；
- 切分正确率不高，一般在95%左右

■ 最少分词法（最短路径法）

- 基于词典，每个句子将生成一个有向无环图, 每个字作为图的一个定点, 边代表可能的分词
- 若赋给相应的边长一个权值（该权值可以是常数，也可以是构成的词的属性值），然后针对该切分图，在起点到终点的所有路径中，求出最短路径，该最短路径上包含的词就是该句子的切分结果



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

缺点：

- 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

3.3.2 基于统计分析方法

- 每个字都是词的最小单元，如果相连的字在不同的文本中出现的频率越多，这就越有可能是一个词。因此我们可以用相邻字出现的频率来衡量组词的可能性，当频率高于某个阈值时，我们可以认为这些字可能会构成一个词。

- 主要统计模型： N元文法模型（N-gram），隐马尔可夫模型（Hidden Markov Model，HMM），最大熵模型（ME），条件随机场（Conditional Random Fields，CRF）等

HMM算法思想

可以把输入句子S作为HMM的输入；单词串S_w为状态的输出，即观察序列$S_w = w_1 w_2 \cdots w_n \quad (n \geq 1)$；词性序列$S_c$为状态序列，每个词性标记对应HMM中的一个状态$q_i$，$S_c = c_1 c_2 \cdots c_n$。

模型$\mu=(A,B,\pi)$中状态（词性）的数目为词性符号的个数N；从每个状态可能输出的不同符号（单词）的数目为词汇的个数M。

假设在统计意义上每个词性的概率分布只与上一个词的词性有关（即词性的二元语法），而每个单词的概率分布只与其词性相关。那么，我们就可以通过对已分词并做了词性标注的训练语料进行统计，得到如下三个矩阵：

(a) 初始状态（词性）的概率分布矩阵：

$$\pi_i = P(q_1 = c_i), \quad 1 \le i \le N$$

(b) 状态转移（词性到词性的转移）概率矩阵：

$$A = \{a_{ij}\}, \quad a_{ij} = P(q_t = c_j \mid q_{t-1} = c_i), \quad 1 \le i, j \le N$$

(c) 从状态（词性）观察到输出符号（单词）的概率分布矩阵：

$$B = \{b_j\}, \quad b_j(k) = P(S_{w_t} = w_k \mid q_t = c_j), \quad 1 \le j \le N, \quad 1 \le k \le M$$

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

缺点：

- 训练语料的规模和覆盖领域不好把握
- 计算量较大

4. 词性标注

词性（part-of-speech, POS）标注（tagging）的主要任务是消除词性兼类歧义。基本方法：

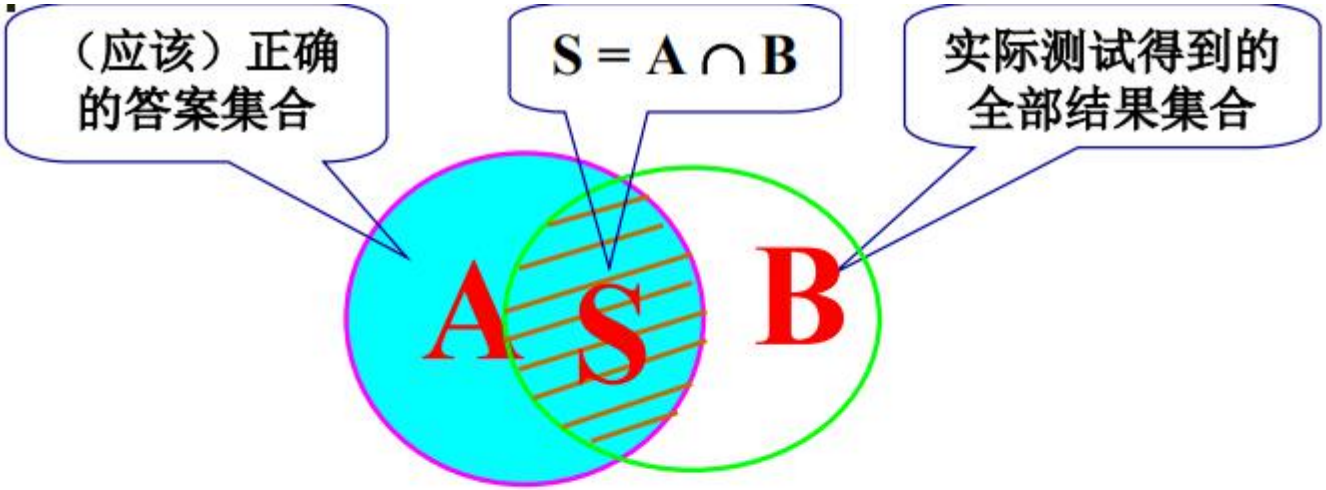
- 基于规则的词性标注方法（根据词性及词语结构定义规则）
- 基于统计模型的词性标注方法（n-gram, HMM...）
- 规则和统计方法相结合的词性标注方法
- 基于有限状态变换机的词性标注方法
- 基于神经网络的词性标注方法

5. 分词与词性标注结果评测

评测指标：

- 正确率(Correct ratio/Precision, C): 测试结果中正确结果的个数占系统所有输出结果的比例
- 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
- 测度值(F-Measure): 正确率与找回率的综合值

$$F-measure=\frac{(\beta^2+1)\times C\times R}{\beta^2\times C+R}\times 100\% \quad \text{一般地，} \beta=1$$



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

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

Thank you!