实验一　中文分词及词性标注实验

一．实验目的：

1. 熟悉国内外汉语自动分词及词性标注的进展
2. 独立完成中文分词和词性标注的处理

二．实验原理：

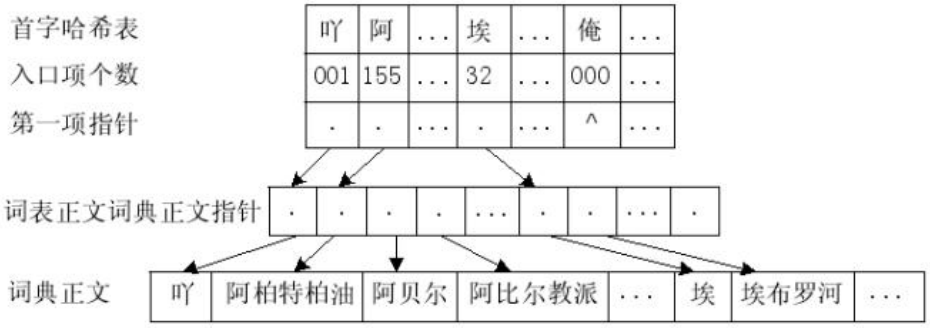
1. 建立高效快速的分词词典机制

我们采用的是基于 HASH 索引的分词词典。国际 GB2312 汉字编码表共收录了 6763 个汉字，为了对齐，里边加上有 5 个空白编码，共有 6768 个汉字。根据汉字机内码编码规律，汉字在编码表中的偏移量计算公式如下：

$$offset = (c1 - 0xB0) * 94 + (c2 - 0xA1)$$

其中，offset 代表某汉字在编码表中的位置，c1，c2 代表汉字的内部码。因此，每一个汉字都有自己唯一的偏移量。对词典中相同首字的词语的进行 Hash 表的索引，则词典中都有唯一的一项地址，表示以此字开头的词语的集合。

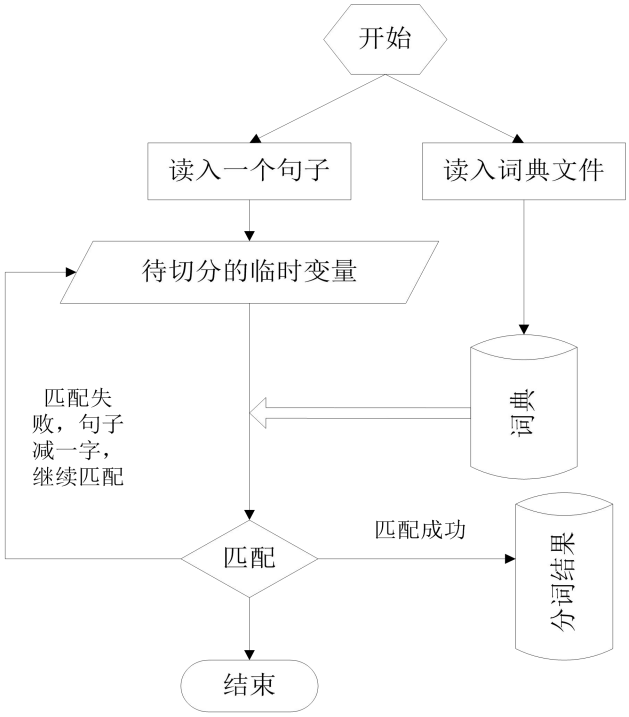
分词词典机制可以看作包含三个部分：首字 Hash 表、词索引表、词典正文。词典正文是以词为单位 txt 文件，词索引表是指向词典正文中每个词的指针表。通过首字 Hash 表的 Hash 定位和词索引表很容易确定指定词在词典正文中的可能位置范围，进而在词典正文中进行定位，匹配过程是一个全词匹配的过程。



词典结构

2. 基于字符串的匹配

- 最大匹配，即要求每一句的分词结果中的词汇量要最少。最大匹配减字法的流程如下。按照扫描方向的不同，字符串匹配算法又可以分为正向匹配和逆向匹配。



假设 MM 表示正向最大匹配方法，RMM 是逆向最大匹配方法，示例如下：

(1) MM 他/说/的确/实在/理

RMM 他/说/的/确实/在理

(2) MM 结合/成分/子时/有(…)

RMM 结/合成/分子/时有(…)

(3) MM 这个/项目/应用/于(…)

RMM 这个/项目/应/用于(…)

(4) MM 我/对/他/有意/见

RMM 我/对/他/有/意见

- 分词算法设计中的几个基本原则：

➢ 颗粒度越大越好：即单词的字数越多，所能表示的含义越确切。

➢ 切分结果中非词典词越少越好，单字字典词数越少越好。

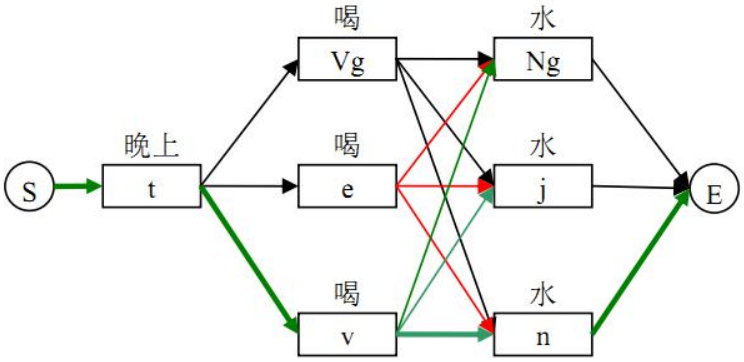
➢ 总体词数越少越好，在相同字数的情况下，总词数越少，说明语义单元越少，那么相对的单个语义单元的权重会越大，因此准确性会越高。

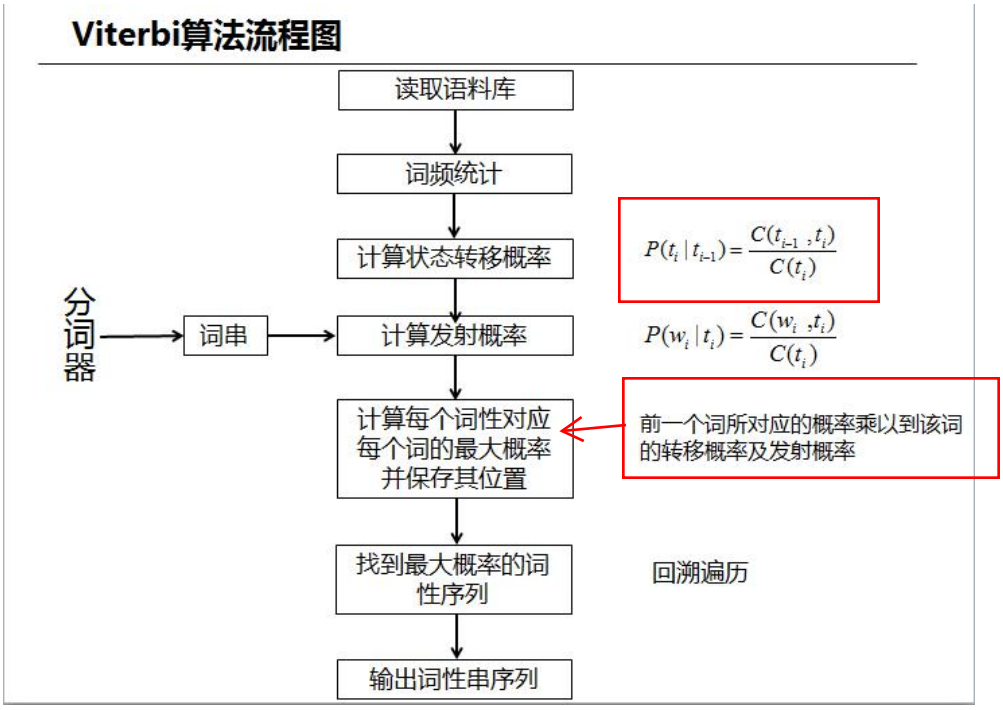
- 消除歧义

设 C = c1 c2 ...cm 表示输入的由 m 个汉字组成的歧义切分字段。W = w1 w2 ...wn ，V = v1 v2 ... vk 表示两种切分结果。frq(w)表示 w 的频率。

若 frq(w1)*frq(w2)*...*frq(wn) >frq(v1)*frq(v2)*...*frq(vn)，则选择切分结果 W 。

3. Viterbi 算法词性标记





三．实验步骤：

1. 终端输入一段文本，已知字典 dic.txt 文件，使用正向匹配和逆向匹配相结合的方法，对一段文字完成分词。

2. 把分词后的输出结果作为词性标注的输入，分别标注出每个词的词性。

3. 通过分析对于此分词处理所采用的时间，分词准确率和分词召回率，可以统计出本系统对于这段实验文本的分词效果。各指标定义如下所示：

- 分词正确率：表示切分出的词语中出现在标准结果中的词语比例。

$$\text{分词正确率} = \frac{\text{切分出的词语中出现在标准结果中的词语数}}{\text{切分出的词语总数}} \times 100\%$$

- 分词召回率：表示标准结果中被正确切分出的词语比例。

$$\text{分词召回率} = \frac{\text{切分出的词语中出现在标准结果中的词语数}}{\text{标准结果中的词语总数}} \times 100\%$$

- 分词速度：表示切分出的相应的单位词语所用的时间。

$$\text{分词速度} = \frac{\text{分词文件大小}}{\text{分词所用时间}} \times 100\%$$

四．实验要求：

1. 本次实验，两节课完成，交老师检查实验结果，并在一周内按时提交实验报告。

2. 实验报告统一格式：学号+姓名+第*次实验.pdf (doc)