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

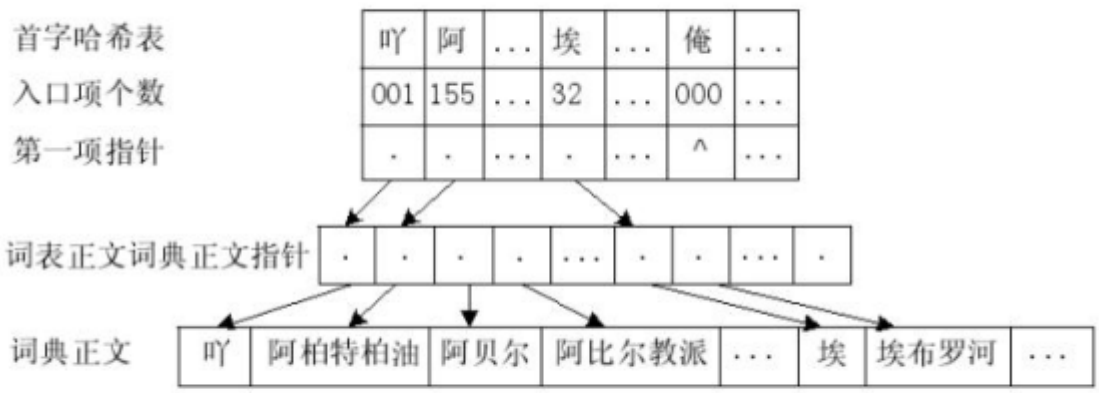
二、实验原理

1.建立高效快速的分词词典机制

采用基于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：

$offset = (c1 - 0xB0) * 94 + (c2 - 0xA1)$ c1,c2为汉字内码。

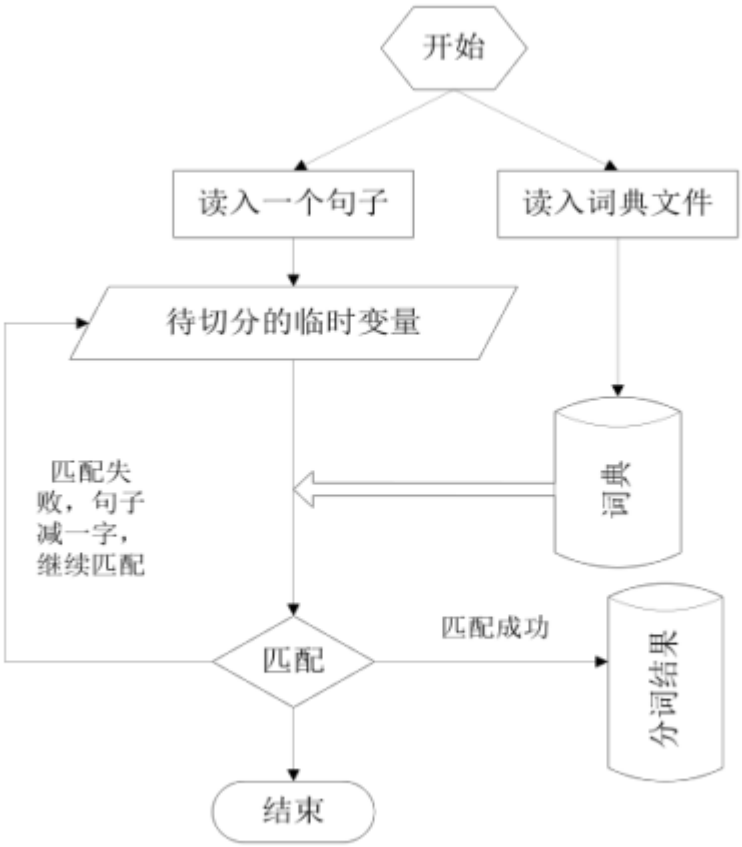
分词词典机制可以看作包含三个部分：首字 Hash 表、词索引表、词典正文。词典正文是以词为单位 txt 文件，词索引表是指向词典正文中每个词的指针表。通过首字Hash表的Hash定位和词索引表很容易确定指定词在词典正文中的可能位置范围，进而在词典正文中进行定位，匹配过程是一个全词匹配的过程。



词典结构

2.基于字符串的匹配

本实验采用的是正向最大匹配(MM)，以及逆向最大匹配法(RMM)。主要流程如下图所示



正向匹配法：采用贪婪的思想，尽可能找到字数比较多的词作为一个整体来切分，方向由前往后。

1.取可能的最大的词长度(匹配窗口大小)（$max(\text{句子长度，词典最大单词长度})$）

2.将从前往后得到词，在词典中进行查询匹配。

3.若没有查询到，则将选择词长度-1（最小长度为1，长度为1时，自动分词），重复2步骤。若有查询到，则进行一次分词，将句子去掉此部分，并重新对新的句子进行分词，直到分词结束。

逆向匹配法：与正向匹配法，区别只是在于分词方向改为了从后向前。

最少切分：使每一句中切出的词数最小。（这里可以使用动态规划的思想，划分成局部问题）

双向最大匹配算法：大部分情况下，前向匹配和后向匹配得到的结果是相同的，但是也有不同的情况。双向最大匹配就是同时进行前行和后向最大匹配，然后分析两种的结果。如果两种结果一致，则认为不存在歧义现象；如果不一致，则需要定位到歧义字段处理。优点是提高了分词的准确率，消除了部分歧义现象。缺点是算法执行要做双向扫描，时间复杂度会有所增加。

同时自己查阅资料也了解到除了基于字典的分词算法还有**基于统计的分词算法**比较常见

是在给定大量已经分词的文本的前提下，利用统计机器学习模型学习词语切分的规律（称为训练），从而实现对未知文本的切分。例如最大概率分词方法和最大熵分词方法等。基于统计的分词方法从一定程度上可以考虑到语义问题，得到最好的分词结果，主要的统计模型有：**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）**，**条件随机场模型（Conditional Random Fields，CRF）**等。这样的方法一般分为两个步骤：

1.找出句子的所有分词结果；（大致就是递归求解的方法）

2.在所有的分词结果中找到最好的那一个。（基于各种统计模型）

3.分词算法的几个原则及消除歧义

- 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。

本次实验中提到的消除歧义方法主要为**词频消歧**。即：

设 $C = c1c2...cm$表示输入的由 m 个汉字组成的歧义切分字段。$W = w1w2...wn$，$V = v1v2...vk$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 。

更多消除歧义的方法了解：

消除歧义可以大致分为

1．分词的消歧，这是很常见的一个例子(南京　市　长　江大桥)

2．多义词的具体词义

3．词性的判断

常用的算法有：

一、常用的算法

1、监督学习算法

a.确定词表和释义表，如目标词“bass”，有两个释义：乐器-贝斯，鱼类-鲈鱼；

b.获取语料：Google、百度

c.特征提取：一般先设定一个窗口，只关心这个窗口的词。

d.分类器选择：朴素贝叶斯、逻辑回归、SVM、KNN、神经网络

例如：基于贝叶斯分类

任何的多义词含义都跟上下文语境相关。假设语境（context）为c，语义为s则

P(s|c)=p(c|s)*p(s)/p(c)，可以根据大量的语料统计，从而计算得到结果。

2、半监督学习算法

当目标词没有足够的语料的时候，从少量手动标注启动，按照同一共现释义中，不同词出现频率进行扩展，如bass的鲈鱼解释一般与fish共现；乐器贝斯解释一般与play共现。因此可以标注所有<fish, bass> 和 <play, bass> 的语句。

3、无监督学习算法

一种贝叶斯分类器，参数估计不是基于有标注的训练语料，而是先随机初始化参数p(v|s),根据EM算法重新估计概率值，对每个词义上下文计算得到p(c|s)，不断迭代从而得到最终分类的模型，最终利用余弦相似性计算得到结果。

4、其他方法

如基于语义角色标注、依存句法分析，可以对某些问题得到一个比较好的结果。

4.Viterbi算法词性标记

大致思想为，根据语料库，获得一个词性的转移矩阵，以及根据观察到的分词结果建立发射矩阵，词性的初始概率集，词性作为结束词的概率集。

然后计算出观测结果中的所有词性组合中概率最大的那一种。

最直接的暴力求解法时间复杂度为$O(m^n)$,m为分词出的个数，n为可能的词性。这种方法时间复杂度太高，可以根据数据的特点而进行简化，简化的算法即为Viterbi算法。

这里的词性标记其实是隐马尔可夫链的第三个问题（求观察序列的最可能的标注序列）。而Viterbi即为此问题的一个实现方式。

Viterbi算法其实就是多步骤每步多选择模型的最优选择问题，其在每一步的所有选择都保存了前续所有步骤到当前步骤当前选择的最小总代价（或者最大价值）以及当前代价的情况下前继步骤的选择。依次计算完所有步骤后，通过回溯的方法找到最优选择路径。

我自己实现的算法大致步骤如下：

1.读取语料库，获取词性转移矩阵，观测单词的的发射矩阵，初始词词性的概率集，结束词词性的概率集。

2.对得到的矩阵进行+1平滑处理。（假设转移矩阵为trans，发射矩阵为emit）(trans(i,j)表示从词性i转移到j的概率，emit(i,k)表示第k个词转换为词性i的概率)

3.计算第一个单词的词性概率，单词对应词性 i 的发射概率*此词性作为初始词的概率，将计算值赋值给

V(i,0),记录最大的可能词性到路径矩阵中。

4.然后从第二个单词依次到最后一个单词。进行如下计算：(假设当前为第k个词)

计算从第k-1个词到第k个词从词性i转换到j的各种可能组合,并记录最大值到v矩阵中，即

v(j,k) = max{ v(i,k-1) * trans(i,j) * emit(j,k) }

并记录该词选择为词性 **j** 时的上一个词性选择 $choose_{j,k} = i$ 。

5.根据计算得到的最后一个单词n的各种词性可能性状态中$v(i,n)$选择出概率最大的一个词性（这里可以考虑结合结束词词性概率集进行最后一个词的词性选择），假设为maxn。

然后根据$choose_{maxn,n}$ 找到上一个词的词性。再重复此步骤，依次回溯直到确定所有词的词性。然后输出。

动图过程如下：

more about 词性标注

词性标注常见方法

基于规则的词性标注方法

基于规则的词性标注方法是人们提出较早的一种词性标注方法，其基本思想是按兼类词搭配关系和上下文语境建造词类消歧规则。早期的词类标注规则一般由人工构建。

随着标注语料库规模的增大，可利用的资源也变得越来越多，这时候以人工提取规则的方法显然变得不现实，于是乎，人们提出了基于机器学习的规则自动提出方法。

基于统计模型的词性标注方法

统计方法将词性标注看作是一个序列标注问题。其基本思想是：给定带有各自标注的词的序列，我们可以确定下一个词最可能的词性。

现在已经有隐马尔可夫模型（HMM）或条件随机域（CRF）等统计模型了，这些模型可以使用有标记数据的大型语料库进行训练，而有标记的数据则是指其中每一个词都分配了正确的词性标注的文本。

基于统计方法与规则方法相结合的词性标注方法

理性主义方法与经验主义义相结合的处理策略一直是自然语言处理领域的专家们不断研究和探索的问题，对于词性标注问题当然也不例外。

这类方法的主要特点在于对统计标注结果的筛选，只对那些被认为可疑的标注结果，才采用规则方法进行歧义消解，而不是对所有情况都既使用统计方法又使用规则方法。

基于深度学习的词性标注方法

可以当作序列标注的任务来做，目前深度学习解决序列标注任务常用方法包括LSTM+CRF、BiLSTM+CRF等。

三、实验步骤

1.分词：

1.读取字典；

2.读取待分词句子

3.正向最大匹配，逆向最大匹配分词

4.根据分词原则或者根据消歧方法选择最终结果

5.查看性能

2.词性标注

1.读取语料库

2.根据读取内容建立计算所需要的各矩阵，转移矩阵，发射矩阵，初始概率等

3.viterbi算法

4.输出结果

实验代码

四、实验结果

```
"C:\Program Files (x86)\Java\jdk1.8.0_181\bin\java.exe" ...
[安徽省, 合肥市, 长江路, 悬挂, 起, 3, 3, 0, 0, 盏, 大红, 灯笼, 为, 节日, 营造, 出, 千, 盏灯, 笼, 凌空, 舞, 十里, 长街, 别样红, 的, 欢乐祥和, 气氛]
[安, 徽, 省, 合肥市, 长江路, 悬, 挂起, 3, 3, 0, 0, 盏, 大, 红灯笼, 为, 节日, 营, 造出, 千, 盏, 灯笼, 凌空, 舞, 十里, 长街, 别样红, 的, 欢乐, 祥和气氛]
[安徽省, 合肥市, 长江路, 悬挂, 起, 3, 3, 0, 0, 盏, 大红, 灯笼, 为, 节日, 营造, 出, 千, 盏灯, 笼, 凌空, 舞, 十里, 长街, 别样红, 的, 欢乐祥和, 气氛]
finished
ns ns n v v v m m m q b n v n v v m q n v v u n v u n n
分词准确率：0.7407407407407407
分词召回率：0.6666666666666666

Process finished with exit code 0
```