哈尔滨工业大学 2018 年秋季学期

计算机学院本科生
“中文信息处理”课
实验报告（三）

报 告 题 目：拼音汉字转换
姓 名：肖松
学 号：1160300527
学 生 专 业：计算机科学与技术
任 课 教 师：刘秉权

2018 年 12 月 29 日

报 告 正 文

1.实验内容

1. 利用统计语言模型实现拼音汉字转换；
2. 输入：拼音串，输出：对应的汉字串；
3. 给定 10000 字的测试语料，测试音字转换的准确率；
4. 针对音字转换结果中存在的问题给出具体分析；
5. 以图表的形式表示上述结果；

2. 实验要求和目的

1. 自己准备词表；
2. 自己准备语料，规模应在一千万字以上；
3. 编程环境、汉字编码不限。

3.实验环境

win10

Python 3.7.0

pypinyin 0.34.0

4.程序主要算法

HMM(隐马尔科夫模型):

马尔科夫假设：

随机过程中各个状态 S_t 的概率分布，只与它的前一个状态 S_{t-1} 有关，即 $P(S_t|S_1,S_2,S_3,\ldots,S_{t-1}) = P(S_t|S_{t-1})$。

符合马尔可夫假设的随机过程称为马尔可夫过程，也称为马尔可夫链。在隐马尔科夫模型中，含有两条马尔科夫链：

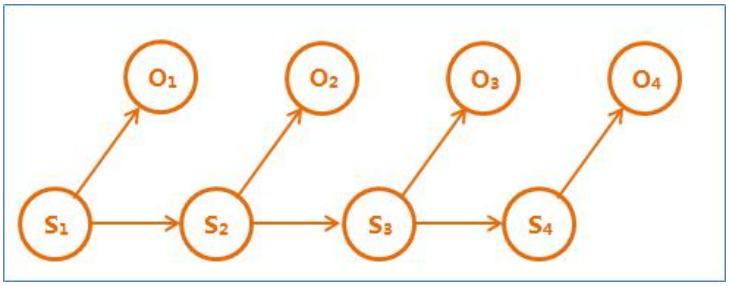


图 1.隐马尔科夫模型

其中，S_1,S_2,S_3,S_4 为隐含状态，O_1,O_2,O_3,O_4 为观察到的序列

在拼音转汉字实验中，拼音为观察的输出序列，而对应的汉字则为产生该输出的状态序列。拼音转汉字过程是寻找一个拼音序列所对应的汉字序列，并且该汉字序列的产生概率最大。我们可以将其描述为以下过程：给定一个模型和某个特定的输出序列，如何找到最可能产生这个输出的状态序列。该过程可以用维特比算法进行求解。

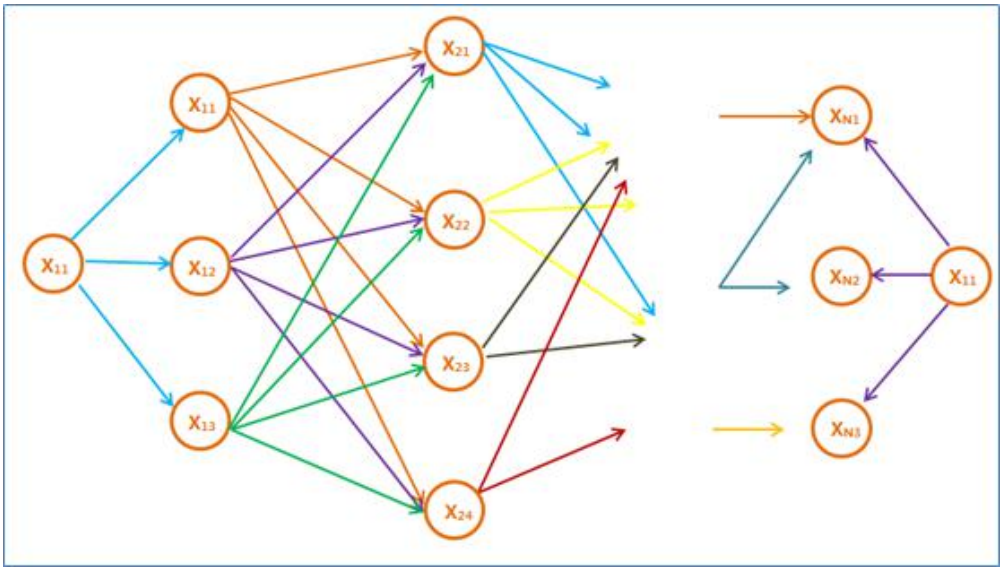
Viterbi 算法：

维特比算法运用了动态规划的思想，算法过程描述如下：

（1）如果概率最大的路径 P（或叫最短路径）经过某个点，比如下图中的 X_{22}，那么这条路径上从起始点 S 到 X_{22} 的这一段子路径 Q，一定是 S 到 X_{22} 之间的最短路径。否则，用 S 到 X_{22} 的最短路径 R 替代 Q，便构成了一条比 P 更短的路径，这显然是矛盾的。

（2）从 S 到 E 的路径必定经过第 i 时刻的某个状态，假定第 i 时刻有 k 个状态，那么如果记录了从 S 到第 i 个状态的所有 k 个节点的最短路径，最终的最短路径必经过其中的一条。这样，在任何时刻，只需要考虑非常有限条最短路径即可。

（3）结合上述两点，假定当我们从状态 i 进入状态 i+1 时，从 S 到状态 i 上各个节点的最短路径已经找到，并且记录在这些节点上，那么在计算从起点 S 到前一个状态 i 所有的 k 个结点的最短路径，以及从这 k 个节点到 X_{i+1}，j 的距离即可。



Viterbi 算法可以有效求出隐马尔科夫模型的最短路径（最大概率），且时间复杂度为$O(N*D^2)$(在拼音转汉字中，N 为拼音个数，D 为每个拼音对应的汉字个数，由于每个拼音对应的汉字个数是一定的，可视为常数，因此时间复杂度为 O(N)，满足要求）

5. 实验过程

1. 训练模型

训练集与测试集语料的格式如下：

```
六一八左右
它使自己成为现代观众必须接受的事实
当水利事业实行的时候
而怀着一种好奇的异想的外国人所能明了的
有条件放牧的要放牧
日本帝国主义企图强占全中国的最明显表示
优质名牌产品和新产品比重上升
并且对膜的化学成分进行分析
手淫的危害及其防治
由于十三个人吃
总有几位从北京去的军中的高级干部
学音乐教材中
印第安人的
日结束的
其中一个跑到老斤面前
在苗木进入高生长速生期
经营不让陶朱富
大学者
形形色色的趣闻轶事
咸通三年三月
包干指标便可能突破
上书郑板桥题的四个大字
细则
一队民工走来
```

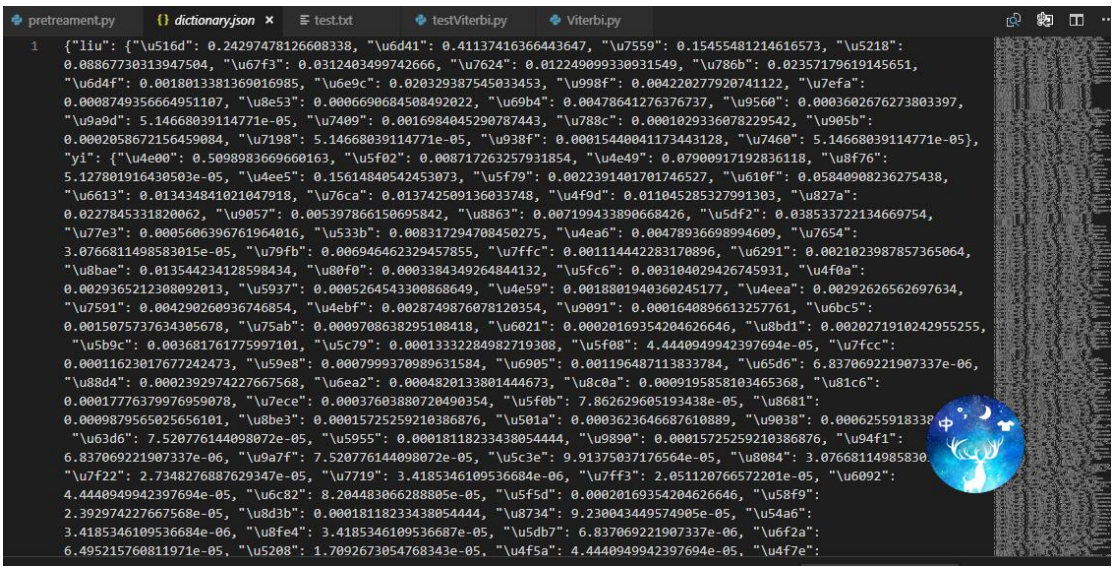
每一行为单独的一句，不含有汉字以外的符号。在训练过程中，我们需要先得到每个字对应的拼音，然后根据对应关系获取字典，转移矩阵以及各汉字的发射概率，计算过程如下：

1. 通过 pypinying，获取每个句子对应的拼音

```
1.  PinYin = lazy_pinyin(words,errors="ignore")
```

2. 字典生成及发射概率计算：统计每个拼音所对应的所有汉字，并统计其出现频率，除以该拼音对应的汉字的总数，并存为 json 文件

生成的字典如下;



3. 若在句子 S 中，汉字 A 与 B 满足， $AB \subseteq S$,则 A 的转移转移矩阵中必包含 B。统计 A，得到所有的 B，并根据其各自出现的频率计算转移矩阵中对应的概率大小，并存为 json 文件.

生成的转移矩阵文件格式如下：

训练模型时，读取的训练的语料为 resource 文件夹下的 sentence.txt 文件，运行 pretreament.py 文件，会将生成的字典存入 resource 文件夹下的 dictionary.json 文件中，将生成的转移矩阵存入 resource 文件夹下的 transfermatrix.json 文件中

2.根据得到的模型完成拼音到汉字的转换

加载 dictionary.json 与 transfermatrix.json，获取字典与对应发射概率，以及转移矩阵。

并通过维特比算法完成求解，求得具有最大概率的中文序列（具体代码见 Viterbi.py 文件代码）

运行 Viterbi.py 文件可以查看效果

（也可以手动输入测试用例，如 python Viterbi.py “wo shi shei”）

3.读取测试样本，生成对应的拼音，并调用 Viterbi.py 中的 viterbi 方法获取每个拼音对应的汉字，并与原本的汉字进行对比，计算准确率。

运行 testViterbi.py 文件，可以完成测试，默认的测试样本为 resource 文件夹下的 test.txt 文件，也可以通过参数指定测试样本文件。如:python testViterbi.py “resouce/test.txt”

4. 当拼音有多组汉字与之对应时，可以显示多个预选项。（按概率从大到小选择）

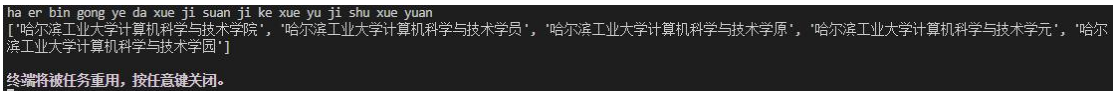
6.实验结果

1.拼音转汉字测试：

拼音：ha er bin gong ye da xue ji suan ji ke xue yu ji shu xue yuan

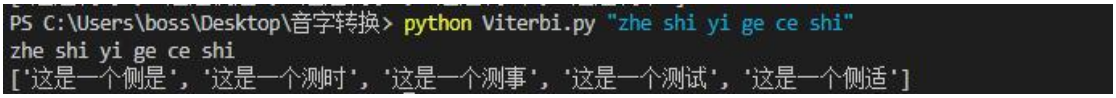
汉字

:['哈尔滨工业大学计算机科学与技术学院', '哈尔滨工业大学计算机科学与技术学员', '哈尔滨工业大学计算机科学与技术学原', '哈尔滨工业大学计算机科学与技术学元', '哈尔滨工业大学计算机科学与技术学园']



拼音：zhe shi yi ge ce shi

汉字：['这是一个侧是', '这是一个测时', '这是一个测事', '这是一个测试', '这是一个侧适']



拼音：pin yin shu ru fa

汉字：['拼音数如发', '拼音数如法', '拼音数入罚', '玭铟鉢洳乏', '玭铟鉢洳阀']

```
PS C:\Users\boss\Desktop\音字转换> python Viterbi.py "pin yin shu ru fa"
pin yin shu ru fa
['拼音数如发', '拼音数如法', '拼音数入罚', '玭铟鉢洳乏', '玭铟鉢洳阀']
```

利用测试样本进行准确率测试（所有备选项中转换正确最多的个数）：

备选项个数	准确率
1	0.7831163317294084
2	0.8022219077271441
3	0.8088734786300594
4	0.8121992640815171

当我们增加备选项时，可以看见，准确率有小幅度提高，因此，我们应当合理加入备选项。

结果分析：

利用隐马尔科夫模型进行拼音转汉字是可行的，但是转换结果受到训练样本影响较大，语料的覆盖范围需要很广泛，否则在某些情况下结果较差。

在输入”zhe shi yi ge ce shi”时，由于训练样本中，“测试”一词出现的频率较低，因此，导致在将”zhe shi yi ge ce shi”转为汉字时，“这是一个测试”出现的概率也就偏低，与实际情况相反。因此，训练所用语料在选择时应当考虑全面。

在输入”pin yin shu ru fa”时，由于训练样本中没有“音输”两字相连的情况，因此，训练出的模型同样无法进行正确的转换。

7.实验结论和体会

在实现拼音转汉字时，所得结果受语料的影响较大。语料的覆盖应当要全面。同时，在面对不同需求时，我们应当用不同倾向的语料来进行训练，可以有效提高使用时的效率。