# 中文词法分析系统
# 工程报告

课程：自然语言理解

## 一、实验背景

词是自然语言中能够独立运用的最小单位，是自然语言处理的基本单位。自动词法分析就是利用计算机对自然语言的形态 (morphology) 进行分析，判断词的结构和类别等。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。

中文信息处理是指自然语言处理的分支，是指用计算机对中文进行处理。和大部分西方语言不同，书面汉语的词语之间没有明显的空格标记，句子是以字串的形式出现。因此对中文进行处理的第一步就是进行自动分词，即将字串转变成词串。

自动分词的重要前提是以什么标准作为词的分界。词是最小的能够独立运用的语言单位。词的定义非常抽象且不可计算。给定某文本，按照不同的标准的分词结果往往不同。词的标准成为分词问题一个很大的难点，没有一种标准是被公认的。但是，换个思路思考，若在同一标准下，分词便具有了可比较性。因此，只要保证了每个语料库内部的分词标准是一致的，基于该语料库的分词技术便可一较高下。

分词的难点在于消除歧义，有些歧义无法在句子内部解决，需要结合篇章上下文。词的颗粒度选择问题是分词的一个难题。研究者们往往把"结合紧密、使用稳定"视为分词单位的界定准则，然而人们对于这种准则理解的主观性差别较大，受到个人的知识结构和所处环境的很大影响。选择什么样的词的颗粒度与要实现具体系统紧密相关。例如在机器翻译中，通常颗粒度大翻译效果好。比如"联想公司"作为一个整体时，很容易找到它对应的英文翻译 Lenovo，如果分词时将其分开，可能翻译失败。然而，在网页搜索中，小的颗粒度比大的颗粒度好。比如"清华大学"如果作为一个词，当用户搜索"清华"时，很可能就找不到清华大学。

目前中文分词的方式有很多，主要的有最大匹配法，复杂最大匹配法，基于

字标注的分词法。我们在上课主要介绍了 n-gram 模型，前后向最长匹配，隐马尔科夫，条件随机场等模型，通过训练 1998-01-2003.txt 的语料库对新的句子进行分词以及词性标注。

## 二、模型方法

隐马尔可夫模型是马尔可夫链的一种，它的状态不能直接观察到，但能通过观测向量序列观察到，每个观测向量都是通过某些概率密度分布表现为各种状态，每一个观测向量是由一个具有相应概率密度分布的状态序列产生。所以，隐马尔可夫模型是一个双重随机过程----具有一定状态数的隐马尔可夫链和显示随机函数集。自 20 世纪 80 年代以来，HMM 被应用于语音识别，取得重大成功。到了 90 年代，HMM 还被引入计算机文字识别和移动通信核心技术“多用户的检测”。HMM 在生物信息科学、故障诊断等领域也开始得到应用。

隐马尔可夫模型（HMM）可以用五个元素来描述，包括 2 个状态集合和 3 个概率矩阵：

1. 隐含状态 S

这些状态之间满足马尔可夫性质，是马尔可夫模型中实际所隐含的状态。这些状态通常无法通过直接观测而得到。（例如 S1、S2、S3 等等）

2. 可观测状态 O

在模型中与隐含状态相关联，可通过直接观测而得到。(例如 O1、O2、O3 等等，可观测状态的数目不一定要和隐含状态的数目一致。）

3. 初始状态概率矩阵 π

表示隐含状态在初始时刻 t=1 的概率矩阵，（例如 t=1 时， P(S1)=p1、P(S2)=P2、P(S3)=p3，则初始状态概率矩阵 π=[ p1 p2 p3 ].

4. 隐含状态转移概率矩阵 A。

描述了 HMM 模型中各个状态之间的转移概率。

其中 Aij = P( Sj | Si ),1≤i,,j≤N.

表示在 t 时刻、状态为 Si 的条件下，在 t+1 时刻状态是 Sj 的概率。

5. 观测状态转移概率矩阵 B （英文名为 Confusion Matrix，直译为混淆矩阵不太易于从字面理解）。

令 N 代表隐含状态数目，M 代表可观测状态数目，则：

Bij = P( Oi | Sj ), 1≤i≤M,1≤j≤N.

表示在 t 时刻、隐含状态是 Sj 条件下，观察状态为 Oi 的概率。

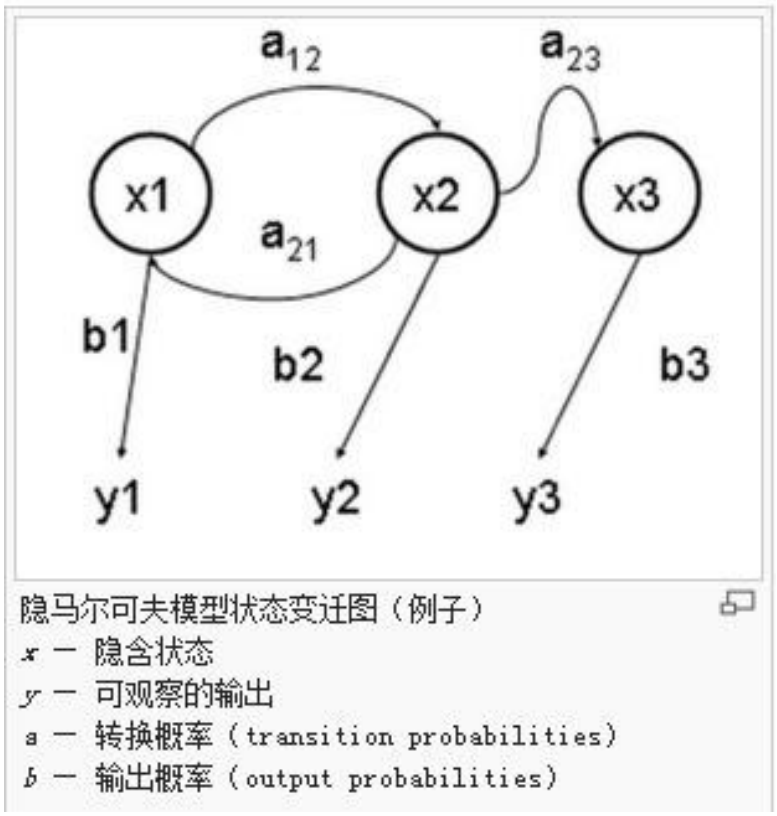


图 1

总结：一般的，可以用 λ=(A,B, π)三元组来简洁的表示一个隐马尔可夫模型。隐马尔可夫模型实际上是标准马尔可夫模型的扩展，添加了可观测状态集合和这些状态与隐含状态之间的概率关系。

1. 评估问题。

给定观测序列 O=O1O2O3…Ot 和模型参数 λ=(A,B, π)，怎样有效计算某一观测序列的概率，进而可对该 HMM 做出相关评估。例如，已有一些模型参数各异的 HMM，给定观测序列 O=O1O2O3…Ot，我们想知道哪个 HMM 模型最可能生成该观测

序列。通常我们利用 forward 算法分别计算每个 HMM 产生给定观测序列 O 的概率，然后从中选出最优的 HMM 模型。

这类评估的问题的一个经典例子是语音识别。在描述语言识别的隐马尔科夫模型中，每个单词生成一个对应的 HMM，每个观测序列由一个单词的语音构成，单词的识别是通过评估进而选出最有可能产生观测序列所代表的读音的 HMM 而实现的。

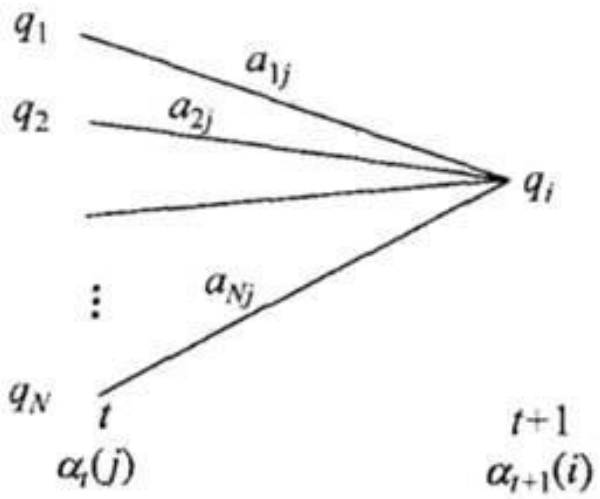


图 2

2.解码问题

给定观测序列 O=O1O2O3…Ot 和模型参数 λ=(A,B, π)，怎样寻找某种意义上最优的隐状态序列。在这类问题中，我们感兴趣的是马尔科夫模型中隐含状态，这些状态不能直接观测但却更具有价值，通常利用 Viterbi 算法来寻找。

这类问题的一个实际例子是中文分词，即把一个句子如何划分其构成才合适。例如，句子“发展中国家”是划分成“发展-中-国家”，还是“发展-中国-家”。这个问题可以用隐马尔科夫模型来解决。句子的分词方法可以看成是隐含状态，而句子则可以看成是给定的可观测状态，从而通过建 HMM 来寻找出最可能正确的分词方法。

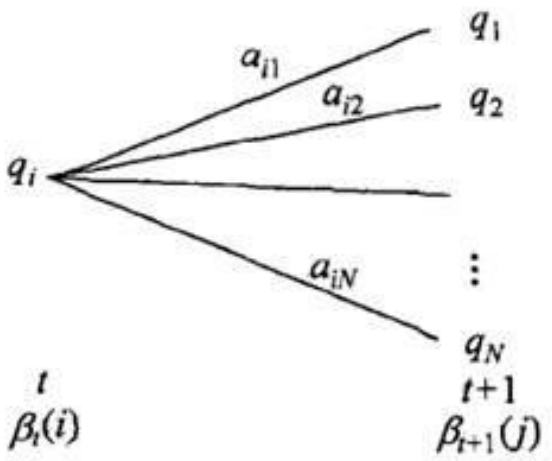


图 3

3. 学习问题。

即 HMM 的模型参数 λ=(A,B, π)未知，如何调整这些参数以使观测序列 O=O1O2O3…Ot 的概率尽可能的大。通常使用 Baum-Welch 算法以及 Reversed Viterbi 算法解决。

## 三、系统设计

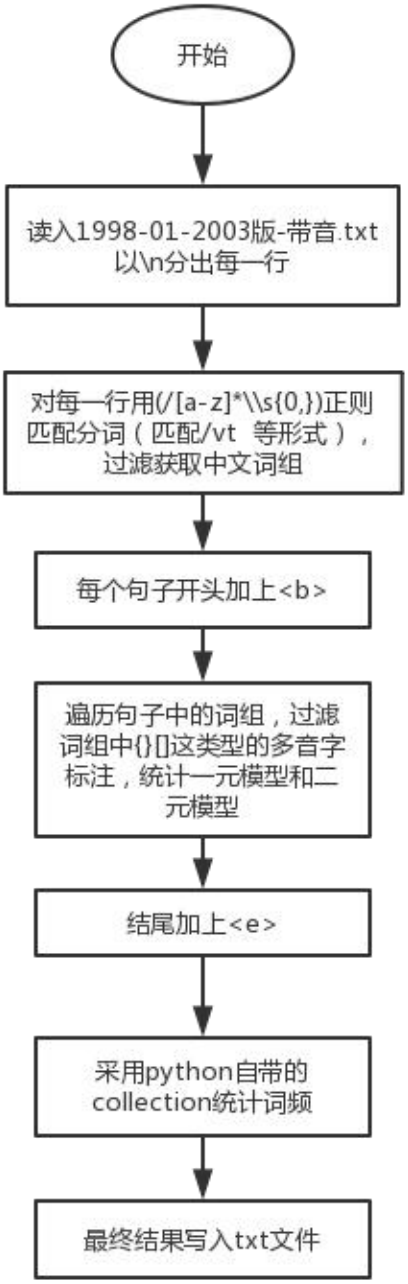
开始
读入1998-01-2003版-带音.txt
以\n分出每一行
对每一行用(/[a-z]*\\s{0,})正则
匹配分词（匹配/vt 等形式），
过滤获取中文词组
每个句子开头加上<b>
遍历句子中的词组，过滤
词组中{}[]这类型的多音字
标注，统计一元模型和二
元模型
结尾加上<e>
采用python自带的
collection统计词频
最终结果写入txt文件

图 4

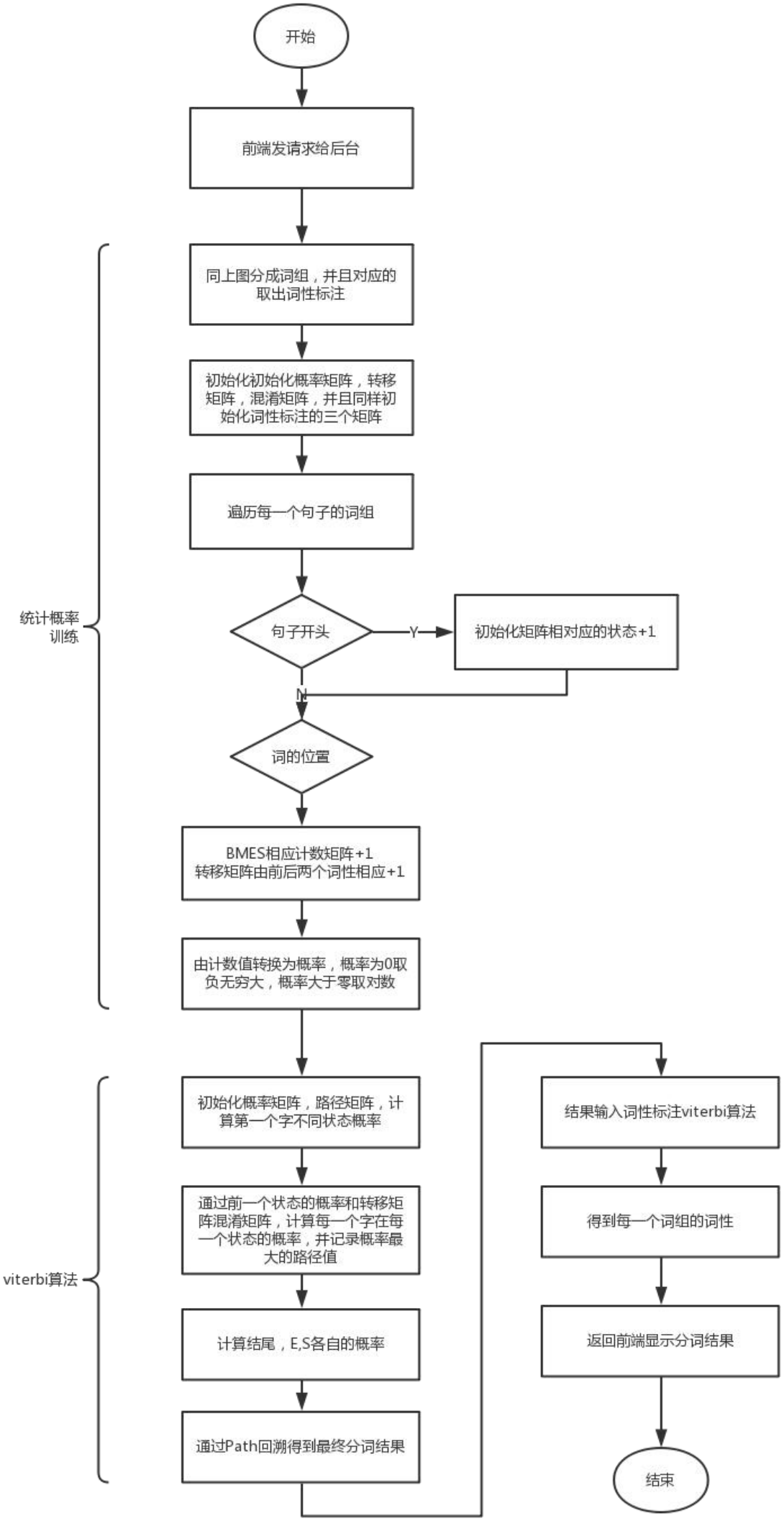
开始
前端发请求给后台
同上图分成词组，并且对应的取出词性标注
初始化初始化概率矩阵，转移矩阵，混淆矩阵，并且同样初始化词性标注的三个矩阵
遍历每一个句子的词组
句子开头
Y
初始化矩阵相对应的状态+1
N
词的位置
BMES相应计数矩阵+1
转移矩阵由前后两个词性相应+1
由计数值转换为概率，概率为0取负无穷大，概率大于零取对数
统计概率训练
初始化概率矩阵，路径矩阵，计算第一个字不同状态概率
通过前一个状态的概率和转移矩阵混淆矩阵，计算每一个字在每一个状态的概率，并记录概率最大的路径值
计算结尾，E,S各自的概率
通过Path回溯得到最终分词结果
viterbi算法
结果输入词性标注viterbi算法
得到每一个词组的词性
返回前端显示分词结果
结束

图 5

实验采用隐马尔科夫模型，图 4 描述了统计一元模型和二元模型的过程，图 5 描述了采用隐马尔科夫模型的过程，首先训练统计整篇文档的转移矩阵和混淆矩阵，统计出所有的数据之后，对于待分词的句子采用 viterbi 算法得到到概率最大的路径，即最终的分词结果。

## 四、系统演示与分析

实验的分词结果如下：

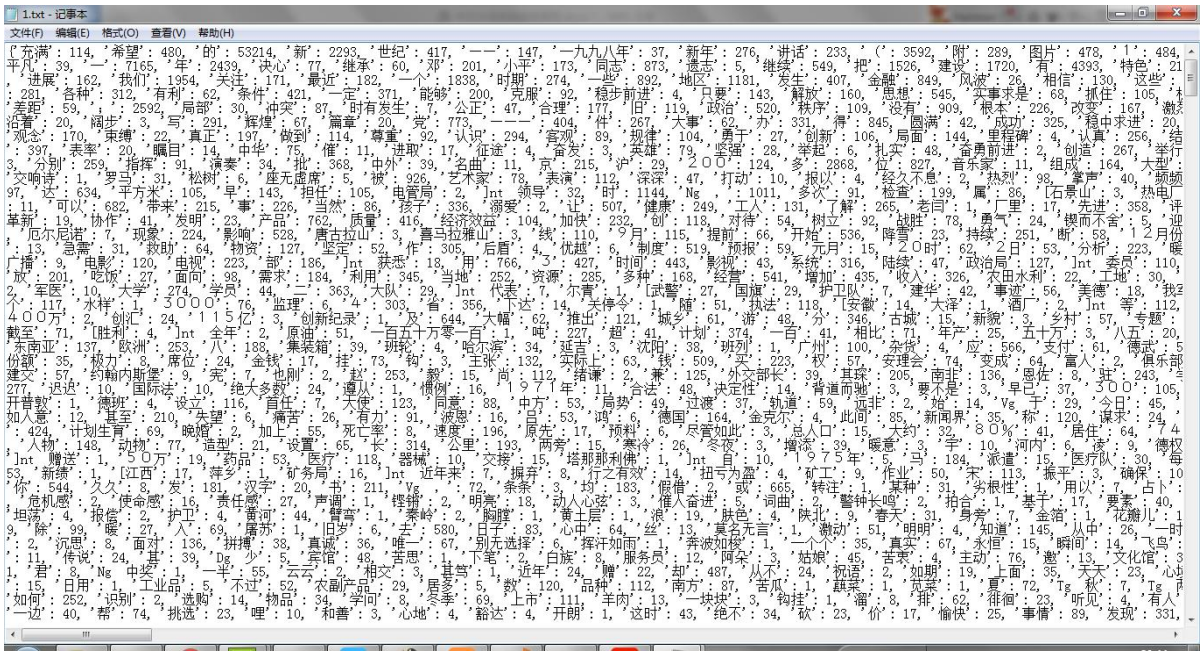


图 6（一元模型）

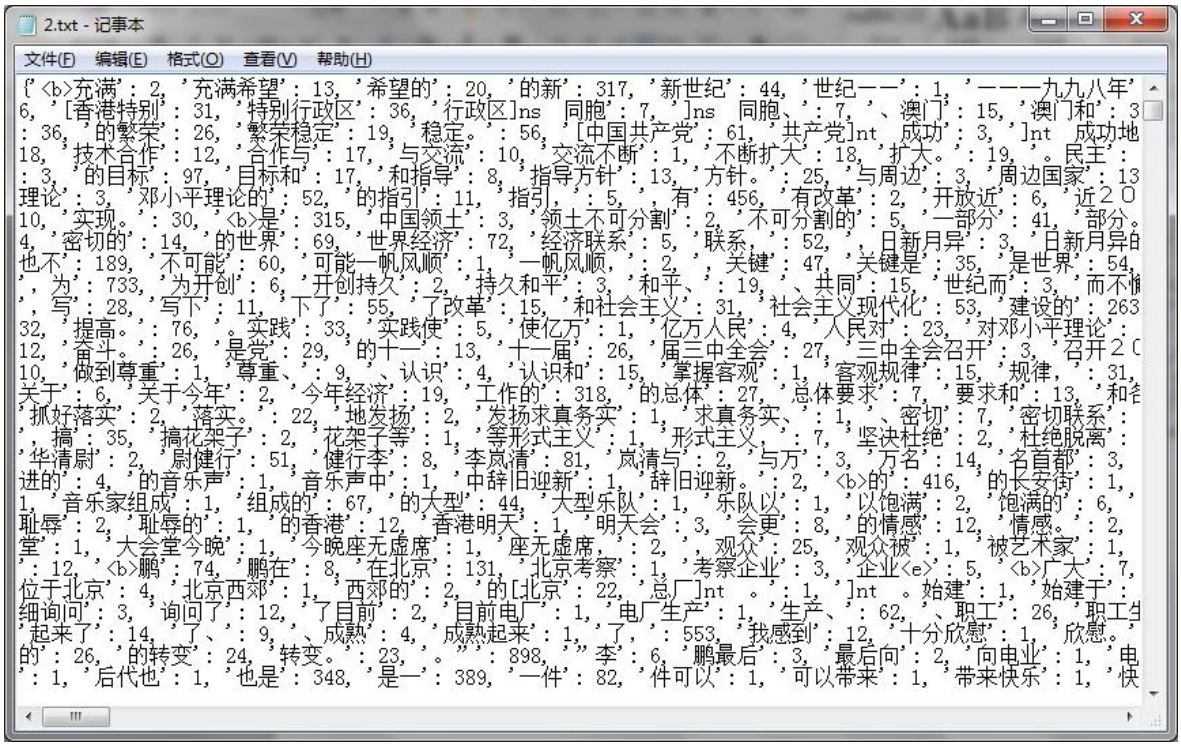


图 7（二元模型）

最终采用隐马尔科夫模型，将结果用 html 展示，通过 django 连接服务，页面上通过输入待分词的句子，点击提交之后将句子传到服务端，通过 viterbi 算法进行计算，得到最后的分词结果，并且将分词结果重新用词性标注的隐马尔科夫模型进行计算得到最后的带有词性标注的分词结果。


图 8（我是一个学生）


图 9（中国强大了）


图 10（欢迎广大新老师生前来就餐）


图 11（这是一个没有登录词语的长句子）

图 8 到 11 展示了一些句子的分词效果，可以看出对于图 8 和图 9 而言分词效果不错，并且词性标注的效果也不错，但观察图 10，显而易见分词的结果不是很理想，即本系统在分析歧义较多的语句上的能力还有待加强。

图 11 为显示标注的词性，原因是在隐马尔科夫模型分词的基础上在套用隐马尔科夫词性标注的时候，因为词性标注是以一个词语为单位，当句子中出现了由预料得到的词典中没有的词组的时候，词性标注中转移矩阵和混淆矩阵都会出

现问题，即未登录词的词性标注未能很好的解决。

通过资料的查询发现了可以通过在词性标注的隐马尔科夫模型上通过前后向最大匹配方式得到分词和词性标注的结果。

同时对于隐马尔科夫模型由于生成模型定义的是联合概率，必须列举所有观察序列的可能值，这对多数领域来说是比较困难的。基于观察序列中的每个元素都相互条件独立。即在任何时刻观察值仅仅与状态（即要标注的标签）有关。对于简单的数据集，这个假设倒是合理。但大多数现实世界中的真实观察序列是由多个相互作用的特征和观察序列中较长范围内的元素之间的依赖而形成的。

总的来说 CRF 优于 HMM 的地方在于，它可以引入更多的特征，包括词语本身特征和词语所在上下文的特征，而非单词本身。从某种角度讲，它结合了 HMM 和最大熵方法。

**五、对本门课程的感想**

在本次课程初步认识到了自然语言的一些基础知识，了解了自然语言处理方面的一些最基础的处理方式，在处理中体现了通过计算机与计算机进行通信，最大的困难体现自身在语言文本和对话的各个层次上广泛存在的各种各样的歧义性或多义性，通过老师的介绍我们对自然语言有了一个初步大致的了解，并且在本次的两个实验中，简单的实现了字词的统计和频率的统计，在这两个小型的宋词生成系统和中文分词系统的实现中，感受到了处理自然语言的困难，因为中文语义的复杂性，使得我们的处理遇到了很大的问题，通过本次课程，也加深了我对自然语言处理的兴趣。