**自然语言处理课程大作业**

**大作业题目：**三种中文分词方法的性能对比与评分

**目录**

[1.选题背景及意义 5](#_Toc59046013)

[2.研究现状分析 5](#_Toc59046014)

[3.本文算法 6](#_Toc59046015)

[3.1 算法概述 7](#_Toc59046016)

[3.2 算法公式、文字描述 8](#_Toc59046017)

[3.3 算法细节 12](#_Toc59046018)

[4.实验结果 14](#_Toc59046019)

[5.讨论和分析 16](#_Toc59046020)

[5.1结果展示和对比 16](#_Toc59046021)

[5.2分析 16](#_Toc59046022)

[6.结论 19](#_Toc59046023)

[7.学习体会和建议 19](#_Toc59046024)

[8.小组成员贡献 19](#_Toc59046025)

**中文摘要**

随着网络信息的急剧增长给人们搜索信息带来一定的困难，搜索引擎的出现及时地解决了这个问题，而在搜索引擎中，其最核心的部分之一便是中文分词算法，中文分词算法的分词效率在一定程度上影响着检索词条的速度。在这个互联网信息快速发展的时代，效率无疑是在商场上竞争的核心要素。

本学期，我们在自然语言处理课上学习了多种中文分词算法，在本次大作业中，我们选择了其中的三个算法：最大匹配的三种算法--正向、逆向、双向；基于统计的Uni-Gram模型；隐马尔可夫（HMM）统计模型。首先我们将会根据上课所学内容实现这三种模型的代码，在PKU词典的正确分词级集、测试集和训练集的基础上，针对他们三种算法的召回率、F1 score、准确率这三项指标进行比较，最终将结果打印出来，选出一个最优的算法作为实验结果。

关键词：中文分词算法、比较、准确率、召回率、F1 score

Abstract

With the rapid growth of network information, people have some difficulties in searching information. The emergence of search engine has solved this problem in time. In search engine, one of the most core parts is Chinese word segmentation algorithm. The efficiency of Chinese word segmentation algorithm affects the speed of retrieval entries to a certain extent. In this era of rapid development of Internet information, efficiency is undoubtedly the core element of competition in shopping malls.

This semester, we learned a variety of Chinese word segmentation algorithms in natural language processing class. In this assignment, we selected three of them: three algorithms of maximum matching: forward, backward and bidirectional; uni gram model based on statistics; hidden Markov model (HMM). First of all, we will implement the codes of the three models according to the content learned in class. On the basis of the correct word segmentation level set, test set and training set of PKU dictionary, we will compare the recall rate, F1 score and accuracy rate of the three algorithms. Finally, we will print out the results and select an optimal algorithm as the experimental result.

Key word:Chinese word segmentation algorithm, Comparison, Accuracy, Recall rate, F1 score

# 1.选题背景及意义

存在中文[分词技术](https://baike.baidu.com/item/%E5%88%86%E8%AF%8D%E6%8A%80%E6%9C%AF" \t "https://baike.baidu.com/item/%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D/_blank)，是由于中文在基本文法上有其特殊性，具体表现在：

1．与英文为代表的拉丁语系语言相比，英文以空格作为天然的分隔符，而中文由于继承自古代汉语的传统，词语之间没有分隔。古代汉语中除了连绵词和人名地名等，词通常就是单个汉字，所以当时没有分词书写的必要。而现代汉语中双字或多字词居多，一个字不再等同于一个词。

2．在中文里，“词”和“词组”边界模糊

现代汉语的基本表达单元虽然为“词”，且以双字或者多字词居多，但由于人们认识水平的不同，对词和短语的边界很难去区分。

例如：“对随地吐痰者给予处罚”，“随地吐痰者”本身是一个词还是一个短语，不同的人会有不同的标准，同样的“海上”“酒厂”等等，即使是同一个人也可能做出不同判断，如果汉语真的要分词书写，必然会出现混乱，难度很大。

中文分词的方法其实不局限于中文应用，也被应用到英文处理，如手写识别，单词之间的空格就不很清楚，中文分词方法可以帮助判别英文单词的边界。

随着网络信息的急剧增长给人们搜索信息带来一定的困难，中文分词到底对搜索引擎有多大影响？对于搜索引擎来说，最重要的并不是找到所有结果，因为在上百亿的网页中找到所有结果没有太多的意义，没有人能看得完，最重要的是把最相关的结果排在最前面，这也称为相关度排序。中文分词的准确与否，常常直接影响到对搜索结果的相关度排序。

# 2.研究现状分析

在[自然语言处理](http://baike.baidu.com/view/18784.htm" \t "https://blog.csdn.net/u010551621/article/details/_blank)技术中，中文处理技术比西文处理技术要落后很大一段距离，许多西文的处理方法中文不能直接采用，就是因为中文必需有分词这道工序。中文分词是其他[中文信息处理](http://baike.baidu.com/view/638201.htm" \t "https://blog.csdn.net/u010551621/article/details/_blank)的基础，搜索引擎只是中文分词的一个应用。其他的比如[机器翻译](http://baike.baidu.com/view/21352.htm" \t "https://blog.csdn.net/u010551621/article/details/_blank)（MT）、[语音合成](http://baike.baidu.com/view/549184.htm" \t "https://blog.csdn.net/u010551621/article/details/_blank)、自动分类、自动摘要、自动校对等等，都需要用到分词。中文分词算法现在一般分为三类：基于字符串匹配，基于理解，基于统计的分词。 基于字符串匹配分词：机械分词算法。将待分的字符串与一个充分大的机器词典中的词条进行匹配。实际应用中，将机械分词作为初分手段，利用语言信息提高切分准确率。优先识别具有明显特征的词，以这些词为断点，将原字符串分为较小字符串再机械匹配，以减少匹配错误率；或将分词与词类标注结合。基于理解分词：分词同时进行句法语义分析等模拟人对句子的理解，包括分词子系统，句法语义系统，总控部分。总控部分协调下，分词字系统可以获得有关词，句子等的句法和语义信息对分词歧义进行判断。需要大量的语言知识信息。基于统计：相邻的字同时出现的次数越多，越有可能构成一个词语，对语料中的字组频度进行统计，不需要切词字典，但错误率很高。颜敏、程婷、张庆轩[1]以大数据为背景，研究如何快速有效地建立地点名称文字描述与地理坐标对应关系；蒋驷驹、卢章平、李明珠[2]运用大数据处理工具提取赛珍珠文化元素，探究在大数据环境下，大数据挖掘理念在文创产品设计中应用的可行性；董富江、张文学[3]研究主题舆情采集与分析关键技术，探讨分布式计算环境下的主题舆情采集与中文分词数据分析；杨鹏、张利强、贺斯慧[4]将垃圾信息细化成单独的个体,也就是对信息进行分词；张恒源[5]基于Hash机制的Trie树在实现最大长度搜索的时候具有较高的查找效率；金欣[6]针对传统的非结构化大数据深度挖掘方法结果不准确等问题,提出基于蚁群算法的非结构化大数据深度挖掘方法；王勇、周松、邢策梅[7]为了解决中文地名地址的解析和空间化普遍存在语义理解不准确、空间化精度不高、匹配效率不理想等问题、构建适用于智慧城市时空大数据平台的地名地址时空数据引擎,引入HMM构建了一种地名地址语义解析及地址空间化的方法;李岩、郭凤英、翟兴、陈晓倩、佟金铎[8]以在线医疗网站患者对医生的评价数据为基础,基于jieba分词技术从医生基本信息、医生所在医院、患者评价信息3个维度构建用户画像；陈楠、张标[9]针对点状地图注记长度过长需要换行处理的问题,提出一种地图注记的自动换行方法。在分析地图注记内容和评价准则的基础上,梳理出地图注记换行处理的原则,将基于词典的中文分词技术引入到地图注记的换行处理中；徐嘉泽、潘长在、贺莉丽、王宏甦、张力伟、邓柯[10]利用Top WORDS方法的基本原理、工作流程以及特点和优点,并将其应用于《汉书》和《明史》这两部古籍的词语发现,且以相关古籍的人名及地名索引为标准对Top WORDS方法抓取古文献中专有名词的能力进行了量化评估和比较；马学海[11]探讨了神经网络算法在中文分词中的研究和应用,利用误差反向传播的理念,设计出一种基于神经网络的分词系统。

**中文分词存在问题：**计算机与人对歧义的定义无法统一；

字典中未收录的词识别率低。

**分词系统判断标准（待解决问题）：**歧义识别，新词（未登录词）识别等。

**分词模型：**N元模型

最大熵隐马尔可夫模型

匹配法分词

基于决策树和有向图

# 3.本文算法

由于本文共涉及三个算法（最大正向、逆向、双向匹配算法、Uni-Gram模型以及隐式马尔可夫模型（2-Gram）），故这里将三个算法拆开，在每个标题里均用三个小标题来分别阐述。

## 3.1 算法概述

**3.1.1 最大匹配算法**

最大匹配算法主要包括正向最大匹配算法、逆向最大匹配算法、双向匹配算法等。 其主要原理都是切分出单字串，然后和词库进行比对，如果是一个词就记录下来， 否则通过增加或者减少一个单字，继续比较，一直还剩下一个单字则终止，如果该单字串无法切分，则作为未登录处理。而双向匹配算法则是根据正向和逆向的切分结果以及中文分词一般原则，来选出更好的那个。中文分词的一般原则为：切分出词的数量越少越好，切分出的单字数量越少越好。

**3.1.2 Uni-Gram模型**

1. gram模型的介绍：

**N-gram是自然语言处理中常见一种基于统计的语言模型。它的基本思想是将文本里面的内容按照字节进行大小为N的滑动窗口操作，形成了长度是N的字节片段序列。每一个字节片段称为gram，在所给语句中对所有的gram出现的频数进行统计。再根据整体语料库中每个gram出现的频数进行比对可以得到所给语句中每个gram出现的概率。N-gram在判断句子合理性、句子相似度比较、分词等方面有突出的表现。**

N-gram是基于一个假设：

第n个词出现与前n-1个词相关，而与其他任何词不相关。整个句子出现的概率就等于各个词出现概率的乘积。各个词的概率通过语料中的统计计算得到。

假设句子T是有词序列w1,w2,w3...wn组成，用公式表示N-Gram语言模型如下：

P(T)=P(W1W2…Wn)=P(W1)P(W2|W1)P(W3|W1W2)…P(Wn|W1W2…Wn-1)

当n = 1时，即为Uni-Gram模型，公式如下：

P(T) =P(W1)P(W2)P(W3)…P(Wn)

该模型只考虑了每个单独的标注出最可能出现的情况，并未考虑一个标注的上下文对其的影响。

**3.1.3 HMM隐式马尔可夫**

维特比算法(Viterbi algorithm)是一种动态规划算法，它用于寻找最可能产生观测到的事件的序列，这个序列是隐含状态序列，也叫维特比路径(Viterbi path)。最典型的应用场景是马尔科夫信息源上下文和隐马尔科夫模型(HMM)。

该算法广泛应用于通信技术中的解卷积码，例如：CDMA、GSM数字蜂窝网、拨号调制解调器、卫星、深空通信以及802.11无线网。当前，该算法也经常用于语音识别、语音合成、说话人分割（diarization, 姑且这么翻译，它的意思是将同一段语音中不同人的话语归为一类）、关键词识别、计算语言学、生物信息学中。例如，在语音转文本中（语音识别），声音信号作为观察到的事件序列，而文本序列作为声音信号的隐含原因。对于给定的声音信号，维特比算法找出最可能的文本序列。

## 3.2 算法公式、文字描述

**3.2.1 最大匹配算法**

正向匹配算法：正向最大匹配是从左到右扫描字符串，在一个给定的词典中寻找词的最大匹配。具体思路文字描述如下：

1.输入字符串

2.令指针 指向 的初始位置

3.**repeat**

4.计算当前指针 到字串末端的字数(即未被切分字串的长度)

5.令 = 词典中最长单词的字数，如果 , 令 =

6.从当前 起往右取 个字作为词

7.**\*if \*** 在词典中

8.then 在 后添加一个切分标志，根据 的长度修改指针

9.**\* else \***

10.从 右端去掉一个字

11.**until** 指向字串末尾

12.最后输出添加切分标志后的字符串

逆向匹配算法：逆向匹配算法思路与正相匹配算法相差无几，只不过扫描字符串变成从右倒左扫描。具体思路文字描述如下：

1.输入字符串

2.令指针 指向 的末尾位置

3.**repeat**

4.计算当前指针 到字串开头的字数(即未被切分字串的长度)

5.令 = 词典中最长单词的字数，如果 , 令 =

6.从当前 起往左取 个字作为词

7.**\*if \*** 在词典中

8.then 在 后添加一个切分标志，根据 的长度修改指针

9.**\* else \***

10.从 左端去掉一个字

11.**until** 指向字串开头

12.最后输出添加切分标志后的字符串

双向匹配算法则是将两种算法分别运行后的结果来进行比较，取更好的一种来作为分词结果。

**3.2.2** **Uni-Gram模型**

从统计的角度来看，自然语言中的一个句子s可以由任何词串构成，不过概率 P(s)有大有小。例如：

s1= 我刚吃过晚饭；s2= 刚我过晚饭吃；

显然，对于中文而言 s1是一个通顺而有意义的句子，而s2则不是，所以对于中文来说，P(s1)>P(s2)。但不同语言来说，这两个概率值的大小可能会反转。其次，另外一个例子是，如果我们给出了某个句子的一个节选，我们其实可以能够猜测后续的词应该是什么，例如：

1.the large green \_\_ . Possible answer may be “mountain” or “tree” ?  
2.Kate swallowed the large green \_\_ . Possible answer may be “pill” or “broccoli” ?

显然，如果我们知道这个句子片段更多前面的内容的情况下，我们会得到一个更加准确的答案。这就告诉我们，前面的信息越多，对后面未知信息的约束就越强。如果我们有一个由 m个词组成的序列（或者说一个句子），我们希望算得概率

P(w1,w2,⋯,wm)

根据链式规则，可得

P(w1,w2,⋯,wm)=P(w1)P(w2|w1)P(w3|w1,w2)⋯P(wm|w1,⋯,wm−1)

这个概率显然并不好算，不妨利用马尔科夫链的假设，即当前这个词仅仅跟前面几个有限的词相关，因此也就不必追溯到最开始的那个词，这样便可以大幅缩减上诉算式的长度。即

P(wi|w1,⋯,wi−1)=P(wi|wi−n+1,⋯,wi−1)

特别地，对于 n取得较小值的情况当 n=1, 一个一元模型（unigram model)即为

P(w1,w2,⋯,wm)=∏i=1mP(wi)

当 n=2, 一个二元模型（bigram model)即为

P(w1,w2,⋯,wm)=∏i=1mP(wi|wi−1)

当 n=3, 一个三元模型（trigram model)即

P(w1,w2,⋯,wm)=∏i=1mP(wi|wi−2wi−1)

接下来的思路就比较明确了，可以利用最大似然法来求出一组参数，使得训练样本的概率取得最大值。

**3.2.3 HMM隐式马尔可夫**

隐含马尔科夫模型在中文分词中的应用是将分词作为字在字符串中的序列标注任务来实现的：每个字在构成一个特定的词语时都会占据着一个确定的构词位置（词位），这里每个字只有四个位置：即B（词首）、M（词中）、E（词尾）、S（单独成词）。例如：

应该分为：小明/硕士/毕业于/中国/科学院/计算/所

对应标注：BE/BE/BME/BE/BME/BE/S

这也是我们最终要求的标注序列。

数学抽象为：用代表输入的句子，n代表句子的长度，表示字，代表输出得标注结果，那么此时我们要求的就是，可以理解为输入一个句子，我们要求的是一组输出标注，使得此时的条件概率最大。这里的即为B,M,E,S这4种标记之一。

这里如果直接计算上述的条件概率，是关于2n个变量的条件概率，且n不固定，因此计算有一定的难度，这个时候就引入了观测独立性假设，此时就得到：

这个时候通过观测独立性假设，目标问题得到了极大的简化，然而该方法完全没有考虑到上下文，且会出现不合理的标注，比如很可能会得到：BBB、BEM等。这个时候就要引入HMM：

我们一直期望的是，这时通过贝叶斯公式得到：

此时最大化 就等价于最大化：

这时引入上述的HMM输出独立性假设：

同时对 有：

这里引入HMM的另一个假设：有限历史性假设，每一个输出仅与上一个输出有关，得到：

这里其实是一个二元语言模型，当每个输出与前两个有关时，就变成了三元语言模型，当然复杂度也随之变大。

到这里，通过贝叶斯公式，HMM的三个基本假设，得到了：

这里面的 称为发射概率， 称为转移概率。通过设置 =0可以排除BBB、EM等不合理的情况。

而在HMM中，求解的常用方法就是维特比算法。核心思想是：如果最终的最优路径经过某个，那么从初始节点到点的路径也是一个最优路径（因为每一个节点只会影响前后两个和。

## 3.3 算法细节

**3.3.1 最大匹配算法的算法细节**

在最大匹配算法中，通过测试我们发现一个问题：文本中有大量的日期和数字没能正确分词，因为语料库本身不包含。所以，我们在最大匹配算法的基础上，人工定义了一个规则：将所有数字单独分成一个词，并且每当数字末尾有“年”“月”“日”时，使其和数字合并成一个词。通过加入了这个规则

改进效果如下：

11月 9日 , 经过 200 多 位 专家 学者 历时 5年 的 努力 , “ 夏商周 断代 工程 ” 正式 公布 《 夏商周 年表 》 , 把 我国 的 历史 纪年 由 西周 晚期 的 共和 元年 , 即 公元前 841年 向前 廷伸 了1200多年。

**3.3.2 HMM隐马尔可夫**

输入:观察空间 ,

状态 ,

观察序列 若在t时间观察值为 ,则 ,

大小为的转移矩阵A , 为从状态到的转移概率,

大小为的放射矩阵B ,为状态观察到的概率,

初始概率数组 of size K ,为的概率,

输出:最有可能的隐含状态序列

*A01* **function** : X

*A02* **for** each state **do**

*A03*

*A04*

*A05* **end for**

*A06* **for** **do**

*A07* **for** each state **do**

*A08*

*A09*

*A10* **end for**

*A11* **end for**

*A12*

*A13*

*A14* **for** **do**

*A15*

*A16*

*A17* **end for**

*A18* **return** X

*A19* **end function**

# 4.实验结果

数据集：以下实验使用的数据集为Bakeoff 2005的PKU数据集。

性能指标：准确率，召回率，F1Score，处理速度。

实验中发现有大量日期和数字没有能够正确分词，因此采用了人工定义规则的方法去改进，改进后评分得到挺大提升。

人工定义规则：所有数字单独分做一个词。若数字末尾有"年","月","日",和其合并成一个词。缺点:不能覆盖所有情况,以文字出现的日期和数字情况多样,难以用规则描述。如,上千,一两等等

结果：最大匹配法，Unigram以及隐式马尔可夫模型如下列各表所示。

最大匹配法结果：

**表1 最大匹配法结果展示**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **Precision** | **Recall** | **F1** | **速度（字/s）** |
| **前向** | **0.851** | **0.915** | **0.882** | **56000+** |
| **后向** | **0.852** | **0.917** | **0.884** | **19000+** |
| **双向** | **0.953** | **0.918** | **0.884** | **15000+** |

**特殊字符处理**

分析五个模型的分词结果，我们发现不管是基于词典的最大匹配法，基于统计的隐式马尔可夫模型，还是二者混合的 Unigram 模型，均无法有效的将时间，数字，人名和地名准确的切分处理。其原因在于最大匹配是基于字典的切分方式，当遇到字典中未出现的词语时，最大匹配法无法正确的切分。

解决方法：因为数字和日期的出现规律较为单一，所以采用人工规定义规则的方式，增加了基于规则的数字，日期匹配算法。将所有的数字单独分做一个词。如果数字末尾有年、月、日，则和其合并为一个词。

缺点是不能覆盖所有的情况，以文字出现的日期和数字情况多样，难以用规则描述。如：上千、一两等等。

通过实验，我们取得了 3%的 F1 值提升。 具体结果如下表所示。

**表3 加入日期数字后最大匹配法结果展示**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **Precision** | **Recall** | **F1** | **速度（字/s）** |
| **前向** | **0.907** | **0.931** | **0.919** | **9300+** |
| **后向** | **0.909** | **0.933** | **0.921** | **17000+** |
| **双向** | **0.910** | **0.933** | **0.921** | **5700+** |

**Uni-Gram模型**

**表2 n-gram结果展示**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **Precision** | **Recall** | **F1** | **速度（字/s）** |
| **普通** | **0.844** | **0.922** | **0.881** | **8400+** |
| **识别数字/日期** | **0.892** | **0.937** | **0.914** | **3600+** |

**HMM隐马尔可夫**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **Precision** | **Recall** | **F1** | **速度（字/s）** |
| **隐马尔可夫** | **0.777** | **0.792** | **0.785** | **20000+** |

# 5.讨论和分析

## 5.1结果展示和对比

数据集：以下实验使用的数据集为Bakeoff 2005的PKU数据集。

结果：最大匹配法，Unigram以及隐式马尔可夫模型如下列各表所示：

## 5.2分析

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | Recall | F1 | 字/s |
| 隐式马尔可夫 | 0.792 | 0.785 | 20000+ |

**实例展示：**

**实例1：最大匹配法与Unigram对比**

Unigram:人类 的 生活 会 变 得 怎样 ?☑

前向: 人类 的 生活会 变 得 怎样 ? ✖

后向: 人类 的 生活会 变 得 怎样 ? ✖

分析：最大匹配法只考虑的词语是否出现在字典中，没有考虑整个句子是否正确。在这种情况下unigram的表现更好。

**实例2：采用特殊字符判断的前后对比**

没有处理之前，通过最大正向匹配的结果如下所示，可以发现在对于数字和日期都无法正确切分，而是分为单个字。因此加入了上文所提到的规则方法去进行改进。改进之后可以正确切分。

（ 新华社 北京 1 2 月 3 1 日 电 ）

图片 ：

中共中央 总书记 、 国家 主席 、 中央军委 主席 江 泽民 发表 新年 贺词 。

（ 新华社 记者 兰 红光 摄 ）

为 新世纪 欢呼 （ 图片 ）

2 0 0 1 年 1 月 1 日 零时 ， 随着 新世纪 钟声 的 响起 ， 北京 中华 世纪 坛 礼花 齐 放 ， 万民 欢腾 。 （ 本报 记者 徐 烨 摄 ）

北京 新年 音乐会 展现 经典 魅力 尉 健行 李 岚清 与 数千 首都 观众 一起 欣赏

本报 北京 1 2 月 3 1 日 讯 新华社 记者 秦 杰 、 本报 记者 赵 川东 报道 ： 乐 音 飞扬 ， 弦 歌 阵阵 。 党 和 国家 领导人 尉 健行 、 李 岚清 今晚 在 人民 大会堂 与 数千 名 首都 观众 一起 ， 欣赏 世纪 音乐 盛 典 ——— 2 0 0 1 年 北京 新年 音乐会 。

处理之后：

（ 新华社 北京 12月 31日 电 ）

图片 ：

中共中央 总书记 、 国家 主席 、 中央军委 主席 江 泽民 发表 新年 贺词 。

（ 新华社 记者 兰 红光 摄 ）

为 新世纪 欢呼 （ 图片 ）

2001年 1月 1日 零时 ， 随着 新世纪 钟声 的 响起 ， 北京 中华 世纪 坛 礼花 齐 放 ，万 民 欢腾 。 （ 本报 记者 徐 烨 摄 ）

北京 新年 音乐会 展现 经典 魅力 尉 健行 李 岚清 与 数千 首都 观众 一起 欣赏

本报 北京 12月 31日 讯 新华社 记者 秦 杰 、 本报 记者 赵 川东 报道 ： 乐 音 飞扬 ， 弦 歌 阵阵 。 党 和 国家 领导人 尉 健行 、 李 岚清 今晚 在 人民 大会堂 与 数千 名 首都 观众 一起 ， 欣赏 世纪 音乐 盛 典 ——— 2001年 北京 新年 音乐会 。

**5.2.2 隐式马尔可夫模型**

从结果来看，隐式马尔可夫模型的表现较差。

不管是精确率，召回率还是 F1 值都显著低于最大匹配法和 Unigram 模型。

经过分析，我们认为可能有以下几个原因：

1. 隐式马尔可夫模型是基于字标注的模型，它没能很好的利用词典的信息进行标注。

2. 隐式马尔可夫模型对于其发射矩阵，转移矩阵，初始矩阵的空值较为敏感，简单的平滑处理不能很好的体现真实情况。

3. 训练语料与测试中的句子长度也可能是其表现较差的原因。

因此我们认为：尝试整合词典，采用更为复杂的平滑方法，长句化为几个短句进行处理，均可能提升隐式马尔可夫模型的分词效果。不过由于时间关系，我们还未进行实际的实验。

# 6.结论

（1）在分词上，简单的最大匹配法无法较好的处理未登录的词的问题。针对特殊字符，制定相应的匹配规则能够解决部分未登录词的问题。而整合基于统计的 n-gram 模型，则能够解决部分歧义的问题。

（2）单纯的隐式马尔可夫模型无法在分词上取得非常好的效果，其训练需要较为细致的编码调节以及大量的数据。

# 7.学习体会和建议

在这次以中文分词为基调进而分析比较三种不同算法之间正确率、召回率以及F1 score的自然语言处理大作业当中，我们深深地感受到了中文分词算法这个领域的深奥和奇妙，以及中文分词算法技术广泛应用于社会中的各个领域，特别是在搜索引擎上。这几年随着网络信息的急剧增长给人们搜索信息带来一定的困难，搜索引擎的出现及时地解决了这个问题，而在搜索引擎中，其最核心的部分之一便是中文分词算法，中文分词算法的分词效率在一定程度上决定着检索词条的速度。本次我们选题便是从这个角度出发，针对最大匹配算法、Uni-Gram模型、Hmm隐马尔可夫模型这三个算法，进行效率上的指标对比，找出不同情况下每一种算法的应用利与弊。

在最开始组队讨论的时候，我们展开了激烈的讨论，有组员上网查资料提出做CRF条件随机场算法、有组员提议研究中文分词+词性标注、还有组员提议比较算法的优劣。最终经过投票决定，我们敲定了大致的讨论方向，从比较不同算法优劣程度这个方向入手，展开我们大作业的篇章。在第一次讨论过程中，我们分配好了任务，每个组员便按部就班回去负责搜索自己负责部分的资料，我们专门建了一个微信群进行大作业的讨论与进度对接，在资料收集完毕后我们再一起集中进行了第二次的讨论会，内容是就第一次资料收集确定研究方向的拍案以及代码实现方面难点的互帮互助。在后面对数据集进行处理时，我们发现算法在日期和数字上识别有困难，于是我们对算法进行了字符串上的优化处理，能更精确识别出日期和数字。

其实在这次自然语言处理大作业过程中回过头想一想，中文分词算法已经深刻地渗透到了我们每一个人的生活当中来了。它在信息检索、智能输入、自动摘要、中外文翻译等各个领域中都有应用。汉语的历史悠久与博大精深决定了它的复杂性,提高分词算法的划分精度和划分速度,具有重要的理论意义和现实意义。中文分词能够从大量繁杂的数据中获取隐含中其中的不同含义的信息，也能把我们需要搜索的信息快速及时地进行搜索并反馈回给人们；另一方面，国外的计算机处理技术要想进入[中国市场](http://baike.baidu.com/view/1869794.htm" \t "https://blog.csdn.net/u010551621/article/details/_blank)，首先也是要解决中文分词问题，在中文研究方面，相比外国人来说，中国人有十分明显的优势。我们完全可以认为，中文分词技术在国内外领域拥有着卓绝的不可替代性。因此，作为计算机信息专业的学生，我们更应努力钻研，将自然语言处理这门专业课和其他相关的计算机技术的基础课程学好学精，成为日后进入该行业有力的叩门砖。

在调研的过程中，除了对自然语言处理领域有着更深刻的理解以外，我们摸索着各种调研方法，通过各种收集资料的途径，筛选整合海量的资料，渐渐学会如何更为有效的收集和整理资料；除此之外，这是一份小组合作的大作业，大家在每次小组讨论上都能够各抒己见，我们在积极发表自己的意见的同时，也认真地听取小组其他成员的意见，做到“取其精华，去其糟粕”，批判继承每个人的意见，最终统一小组各成员的思想观点，然后分工完成调研报告，最后整合全文。

在这次完成大作业的过程中，我们收获的远不止是一份大作业的成果，更是一次深入了解认识自然语言处理的机会，一份写调研报告的心得，一次相互合作的经历。

**参考文献：**

[1]颜敏,程婷,张庆轩.中文分词名称匹配与数据分层更新方法探讨[J].测绘地理信息,2020,45(06):129-131.

[2]蒋驷驹,卢章平,李明珠.基于大数据挖掘的赛珍珠文化元素提取与应用[J/OL].包装工程:1-11[2020-12-17].http://kns.cnki.net/kcms/detail/50.1094.TB.20201112.2003.004.html.

[3]董富江,张文学.分布式主题舆情采集与分析系统设计[J].软件导刊,2020,19(11):116-119.

[4]杨鹏,张利强,贺斯慧.基于Word的中文词频分析系统设计与实现[J].企业科技与发展,2020(10):70-72.

[5]张恒源.基于Trie树的最大长度匹配分词的Python实现[J].电子制作,2020(18):56-58.

[6]金欣.基于蚁群算法的非结构化大数据深度挖掘仿真[J].计算机仿

真,2020,37(11):329-333.

[7]王勇,周松,邢策梅.HMM的地名地址时空数据引擎构建方法[J].测绘科学,2020,45(10):161-167.

[8]李岩,郭凤英,翟兴,陈晓倩,佟金铎.基于jieba中文分词的在线医疗网站医生画像研究[J].医学信息学杂志,2020,41(07):14-18.

[9]陈楠,张标.基于中文分词的点状地图注记自动换行研究[J].测绘工程,2020,29(04):21-26.

[10]徐嘉泽,潘长在,贺莉丽,王宏甦,张力伟,邓柯.基于TopWORDS方法的古文献专名识别——以《汉书》和《明史》为例[J].数字人文,2020(02):90-105.

[11]马学海.基于神经网络的中文分词技术研究[J].科学技术创新,2019(32):82-83.