**同济大学计算机系**

**中文信息处理课程实验报告**

****

**学 号 1552215**

**姓 名 田甘迅**

**专 业 计算机**

**授课老师 卫志华**

## 一、实验内容

基于统计的分词算法

基于统计的方法理解一句话，就是要从一句话的每个位置全部候选字中选择一个最可能的句子，我们假定一个字只与前面的N-1个字有关，若没有N-1个音就补以空音。这种N-1阶的马尔可夫模型，在语音理解中又特别地被称为N元文法模型(N-gram Model)。

N元统计计算语言模型的思想是：一个单词的出现与其上下文环境中出现的单词序列密切相关，第n个词的出现只与前面n-1个词相关，而与其它任何词都不相关。

n-gram 是常用的统计语言模型，其中尤以二元文法（Bigram）和三元文法（Trigram）的应用最多。n-gram 模型统计各种语言的邻接共现规律，反映句子的局部范围是否符合语言规范，被广泛用于自动校对中的错误查找。在英文的自动校对中使用了词和词的Trigram，把以字、词、词性为共现元素的Bigram 应用于了中文自动校对中，取得了较好的效果。

由于局部错误在文本的错误中占多数，所以选择合适的查找局部错误的方法对自动校对来说至关重要。我们知道，对正确的语言现象，字词之间的邻接共现概率较高，对一些不符合语法规范的错误，字词之间的邻接共现概率较低。所以，可应用反映字词之间邻接共现规律的n-gram 模型，对句子进行局部分析，查找文本中的局部错误。

方法1：对训练语料先进行自动分词，然后统计二元词邻接共现的规律，相邻的两个词wiwi+1 记为一个词单元ui，下一单元ui+1= wi+1wi+2，该方法统计句子中词单元的绝对次数Count(ui)，将其作为模型的参数；

方法2：使用方法1 的训练数据，以相邻三个词wiwi+1 wi+2 作为一个词单元ui，然后统计三元词邻接共现的绝对次数Count(ui)，扩大词共现的邻接范围

方法3：由于三元词的共现存在很大程度的数据稀疏，这里，对方法2的统计结果进行平滑处理，采用的是插值平滑方法。

## 二、统计分词规则

1.我们针对汉字分词，我们规定分词规则是这样的：

1）.对所有的一个字的词我们都当作是一个词，我们不做分析

2）.针对两个字词，我们按照频率大小和最左优先原则规定他是一个词。

3）.针对三个往上的词，我们只做频度大于一的词的分词，其余按照两个字组成的词那样处理。

2.如何判定当前词是否是一个适当的词：

例如 BC--3 AB--3

针对两个字组成的词，如果当前词（BC）的第一个字(B)是另一个词(AB)的最后一个字(B)并且当前（BC）概率和另一个词(AB)的概率相同并且(AB)的下一个字是(C)，那么我们将(AB)作为一个更合适的词，如果(AB)和(BC)概率不同并且(AB)下一个字是(C),那么谁的概率大，谁将是更合适的词。否则，如果(AB)的下一个字不是(C)，那么他们是两个分开的词。（优先匹配原则）考虑一个特例：ABC，ABC，BC，BC，BC，BC。此时违反我们上述的算法规则，那我们就需要遍历《BC》得前一个字看是否是《A》，如果不是，则按两个词组计算，如果是则按照上述算法分词。经过分析，我们可以同时处理两个字和两个字往上组成的词。比如现在有这样的一个例子：原文出现四个 ABCD,ABCD,ABCD,ABCD。经过初次的统计之后，我们发现AB,BC,CD,ABC,BCD,ABCD都各自出现了四次。现在，我们知道我们需要的词是ABCD而不是其他的组合，此时其他的组合都是ABCD的子集。那么，此时我们就应该按照ABCD的频度和AB,BC,CD,ABC,BCD的频度比较。如果子集的频度小于等于ABCD的频度，则删除其子集，否则按照不同词处理。如果有原文：ABCDE, ABCDE, ABCDE, ABCDE此时统计出来的结果将是：ABCD 4 BCDE 4 另外加上其子集。按照两个字的处理方式，先读到《ABCD》该词组，取其从第二个字开始往后组词，即《BCD》，《CD》，《D》。如果是《BCD》则将判断《BCD》，《BCDE》是否是合适的词，按频度判断。如果是《CD》则将判断《CD》，《CDE》，《CDEF》是否是合适的词。如果是《D》则判断《DE》，《DEF》，《DEFG》是否是合适的词语。

3.对于查找，如果我们规定要对 M 个字组成的词进行分词，则我们首先一次读 M 个字默认它为一个词。这时我们在word\_count中查找该词，如果有，我们直接写入到输出文件并使用一个空格来分隔其他的词；如果当前词不在word\_count中，此时应该查找 M个 字的子集，看谁是一个词并按照上述规定写入到输出文件。例子：现在要查ABCD，如果word\_count中有该词，则将它直接写入到输出文件中并将source\_string的下标移到D的后面；如果没有找到 ABCD，此时前后顺序组合他的所有子集即 AB，BC，CD，ABC，BCD。这个时候，如果 AB 没有找到，则在 A 字的后面空格分开，如果 BC 没有找到，则继续在 B 的后面加上空格分隔。以此类推，将ABCD的所有子集查找，在这个过程中也必须时刻的移动source\_string的下标（指针）。

4.由于map是按照字典树的数据结构来存储数据的，那么我们可以根据此特点来快速的查找。将方法一逆向方式查找就是另一种方法，即先查找子集，如果子集存在，我们探索多一个的子集，一直到我们要的 M 个字的词，此时按照贪心算法，该词是一个合适的词。如果子集不存在，则表明他们是单个字，而不是一个合适的词。

## 三、主要函数

main

SEGarticiple::SEG\_word\_analyze

SEGarticiple::SEG\_word\_print

SEGarticiple::SEG\_word\_wback

## 测试样例

、

## 心得体会

基于统计的分词算法的核心是统计，掌握n-gram模型即可掌握统计的大体思路。对分词和查找的规定则是影响分词结果的重要因素，良好的规则是必不可少的。