本周，结合上周看的论文，我分别基于分支熵（BE），互信息（MI）和词频方差（TFDev）进行了实验，现将结果说明如下。

1. **分支熵实验**

前期由于采用了不合适的标准化方法，导致分词效果不佳，结合“Unsupervized Word Segmentation the Case for Mandarin Chinese”对分支熵标准化采用的方法，即

C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\c21263deb8ed43aa99e5fab12badc680\clipboard.png

其中的C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\7cb7b655060e4ae49ac5260869c42b8b\clipboard.png为原始分支熵；C:\Users\Administrator\AppData\Local\YNote\data\hao1386427@163.com\fd32b63f65c248eea9d4e64e506dfd12\clipboard.png为与x字串长度相等的所有字串的分支熵（右分支熵）的平均值。因为左右分支熵分别采用上述方法标准化后还需将他们融合为一个指标，融合方法我采用了三种方式（M1，M2，M3），

M1：左右分支熵相加；

M2：左右分支熵相乘；

M3：左右分支熵取较小值。

采用切分词图算法进行分词，在农业标注的5000句语料上测试每个句子的准确性，再取平均值，结果如下：

|  |  |
| --- | --- |
| 融合方法 | 平均准确率 |
| M1 | 0.824753332 |
| M2 | 0.813540165 |
| M3 | 0.813540165 |
| 中科院平均准确率 | 0.7867601 |

结果表明，相对中科院分词系统，基于分支熵的方法对分词结果有所提升，且融合方法中M1效果最佳。

1. **互信息（MI）实验**

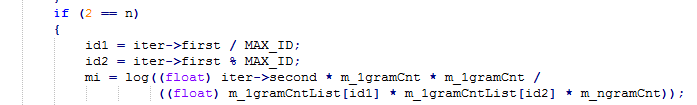
**2.1三元互信息的计算**

在计算三元的互信息时我有个疑问，也在网上找过ABC的互信息怎么计算，但是找到的信息很少，我记得老师当初给我看的一篇英文论文，他讲到了多元词汇（实际是一个英语单词整体的互信息）互信息的计算，即采用了木桶理论，将内部联系最薄弱处的互信息作为整体的互信息，即为：

MI(ABC)=min(MI(AB),MI(BC))

这样计算，其实质2gram和3gram的互信息全是基于2gram的MI进行计算，其值则具有可比性。

还有一个问题，老师当初写的计算MI的公式，我感觉是有问题的，以二元的MI计算为例，您的代码是这样写的



分子分母除的是各元词汇的数目，而我看《统计自然语言处理基础》第五章讲到的计算方式应该是这样的，



其中c代表词汇计数，代表语料库大小，也就是语料库一共多少个词。起初我是采用老师那种计算方式计算的，但是我发现很多词汇的MI与我们主观的感知不一样，于是我就修改为了这个计算公式，并用于实验。

但是，我不知道是否是公式的问题，个别MI计算出来是负值，按照书上说的MI的值应该永远为正才对。老师原来的公式计算的MI全是负的。

**2.1一元互信息的设定**

因为一元的互信息无法计算，老师当初给的解决方法一是指定一个固定值，方法二为a的互信息是ab,ac,ad(是否还需要考虑ba,ca,da的？)我感觉方法一指定默认值比较简单，方法二实现起来麻烦了些。于是我就实验寻找相对较优的一元词汇的互信息。

通过计算，2gram和3gram的互信息平均值分别为：Average(2gram)=3.5;Average(3gram)=2.8,对他们两个取平均值为：3.15。所以1gram默认值的取值范围我定在了3.15附近,通过不断尝试的方法，最终找到了一个比较好的默认值，default(1gram)=4,采用该默认值进行分词，平均准确率为81.0179%

1. **词频方差（TFDev）实验**

基于老师给的TFDev程序我做了下实验，统计了各元的TFEdv平均值。

|  |  |  |
| --- | --- | --- |
| TFDev | 元数 | 平均值 |
| 1gram | 0.833436 |
| 2gram | 0.001862 |
| 3gram | 0.000708 |

我观察了下各元TFEdv的分布和取值情况，发现1gram和其他大小差距比较大，而2gram和3gram数据的大小差距并不是很大，所以我考虑了不采用归一化方法和采用归一化方法两种方式进行实验，观察效果。

**3.1未采用归一化方法**

这时因为一元的TFDev比较大，所以需要针对TFDev（1gram）采取某种方式进行归一化或者指定某个合适的默认值。

基于观察，我首先测试了TFDev’（1gram）= TFDev’（1gram）/100这种方式对1gram的TFDev进行处理，分词结果平均准确率为：82.96%。

然后我也测试了给1gram的TFDev一个合适的默认值，最终经过实验和反复的尝试，较优的默认值为：TFDev’（1gram）=0.03，分词结果平均准确率为：82.43%。

**3.2采用归一化方法**

归一化方法也考虑了两种：一种类似分支熵采用的归一化方法，，即使用真实值减去平均值，这种方法实际测试的准确性为78.24%，相对中科院分词准确性78.67%没有提升

另一种归一化方法为：，这种方法实际测试的准确性为80.52%，相对中科院分词准确性78.67%有所提升。

但是对比上文采用的不使用归一化方法，一元TFEdv缩小100倍的方法来说，准确率还是相对较低，所以最后可以考虑上文不采取归一化的方法。