基于词语凝固度的分词算法

### 背景介绍

分词问题是自然语言处理中一个重要的问题，分词结果的优劣会直接影响语言模型的效果。在众多的分词算法中，*N-*最短路径方法是一种较为简单高效算法，但只能得到粗切分的结果，不能很好地解决歧义切分问题。

### 算法原理

鉴于上述问题，在粗切分的基础上，通过引入词语凝固度这一特征，可以较好地解决歧义切分问题。

词语的凝固度指一个文本片段成为一个词的概率。对一个文本片段*T*，它可以分解为几个单字或词语串接而成，串接的方式可能有一种或多种。设*T*的串接方式集合为*S*，其中某种串接*s*是由*k()*个词语（包括单字）*…*串接而成的。*P(T)*为*T*在语料中出现的概率，*P()*为在语料中出现的概率，则*T*在串接*s*下的凝固度

显而易见，对于*T*的不同串接方式，对应的凝固度有所不同。而T最合理的串接方式应该是最符合自然语言表达的方式，也就是*P()*在语料中出现概率较高的方式。因此，最合理的串接

因此*T*的凝固度为

可以用“图书馆”作为例子来具体说明上述定义。“图书馆”作为文本片段*T*，可以有“图 书 馆”、“图书 馆”和“图 书馆”三种串接方式。在大量语料统计中，易知*P(图书)\*P(馆) > P(图)\*P(书馆) > P(图)\*P(书)\*P(馆)*，所以

从上面的例子可以看出，在一般性的大语料中，*B(图书馆) > B(看图书)*。“图书馆”比“看图书”作为一个词更加合理，其凝固度也较高。由此可见，凝固度很好地描述了“图书馆”在语料中作为一个词的概率，能够很好地为“我们去看图书馆的开馆仪式了”这种带有歧义切分问题的句子分词。

### 操作方法



###### 预处理

无论是计算词语凝固度还是分词，都需要对词典和语料进行预处理，达到相应的标准。词典和语料中的中文编码格式必须一致，否则会导致相同的字符串无法匹配识别。建议中文编码格式统一为GBK，便于工具处理。词典格式保持每行一个词语，便于读取。语料要去除标点符号并分解成一行行的短句，每行最好50个字符以内，避免句子过长增加计算语言模型和分词算法的复杂度。

###### 生成一元模型

使用语言模型工具计算出词典和语料对应的一元语言模型，得到词典中每个词语出现的概率。为了便于计算，可以使用对数来表示。对于语料中未出现的词语，可以使用打折算法为其赋予一个较低的值。

###### 生成凝固度词典

设定由*n*个单字组成的词语称为*n*元词语。显然，对于*n>3*的情况，词语的凝固度的计算非常复杂。考虑到分词歧义多发生在二元和三元词语，而对于*n>3*的情况，依赖最短路径可以保证大部分高元词语得到正确的划分，因此可以将凝固度的计算限定在二元、三元词语。根据凝固度的定义，单字的凝固度都是1。高元词语的凝固度也设置为和单字相同。在实现中可以使用凝固度的对数来表示和计算。

在计算二元、三元词语的凝固度时，需要在一元语言模型中寻找构成词语的单字或二元词语。在词典和语言模型非常大时，可以使用哈希等数据结构提高查找的效率。在查找中可能会遇到找不到的情况，这可能是由于词典过小导致非常用单字找不到（如忐忑），或原本就不存在要查找的二元词语（如大别山），对于这种未找到的词语，可以给一个固定的低值。如在使用ngram工具生成的一元模型中，所有词语的频率值都在-10以上，那就可以将未找到的单字设置为-10，未找到的二元词语设置为-20，这样可以保证最合理的构词概率是最高的。

###### 使用凝固度消除歧义切分

在最短路径方法的非统计粗分模型中，“我们去看图书馆的开馆仪式了”可能会形成如下两种分词结果：

1、“我们 去 看 图书馆 的 开馆 仪式 了”

2、“我们 去 看图 书馆 的 开馆 仪式 了”

在最短路径的算法中，分析的结果如下图：



图中节点的值代表该节点距起始点的距离。在最短路径方法中，通过从终点逆向回溯最短路径来得到分词结果，逆向通过的边就作为一个词语。从图中可以看出，通过路径a和路径b都可以得到最短路径，产生“看 图书馆”和“看图 书馆”这种歧义切分。

在这种情况下，如果将凝固度用于计算边的权值，就能够解决歧义切分的问题。在一般语料中，很明显路径b的凝固度之和要高于路径a，由此可以得到符合日常的正确分词。

### 应用变例

词语凝固度是基于词典、语料和一元语言模型算法来得到的，而一元语言模型又直接受词典和语料的影响，因此保证词典和语料的广泛性是词语凝固度正确合理的关键。一般情况下，分词使用的词典可能只包含常用字词，要分词的语料也比较小，导致凝固度无法计算（如“忐忑”在词表中而两个单字都不在词表中）或凝固度与实际差异很大（语料太小）。对此，可以使用一个词汇丰富的大词表和内容涵盖广泛的大语料，计算出一个通用的凝固度词表。在使用小词表对小语料分词时，在凝固度词表中只抽出小词表中存在的词语，完成分词过程。