基于互信息改进算法和 t-测试差的壮文分词算法研究

覃 俊[1],林叶川[1],易云飞[2,*]

(1 中南民族大学 计算机科学学院,武汉 430074;2 河池学院 计算机与信息工程学院,宜州 546300)

摘 要 针对传统的壮文分词方法将单词之间的空格作为分隔标志,在多数情况下,会破坏多个单词关联组合而成的语义词所要表达的完整且独立的语义信息,在借鉴前人使用互信息 MI 方法来度量相邻单词间关联程度的基础上,首次采用互信息改进算法 MI^k 和 t-测试差对壮文文本分词,并结合两者在评价相邻单词间的静态结合能力和动态结合能力的各自优势,提出了一种 MI^k 和 t-测试差相结合的 TD-MI^k 混合算法对壮文文本分词,并对互信息改进算法 MI^k、t-测试差、TD-MI^k 混合算法三种方法的分词效果进行了比较. 使用人民网壮文版上的文本集作为训练及测试语料进行了实验,结果表明: 三种分词方法都能够较准确而有效地提取文本中的语义词,并且 TD-MI^k 混合算法的分词准确率最高.

关键词 壮文分词; MI 改进算法; t-测试差; 混合算法; 语义词

中图分类号 TP391 **文献标识码** A **文章编号** 1672-4321(2017)04-0100-06

Research on Zhuangwen Word Segmentation Algorithm Based on Mutual Information Improved Algorithm and t-test Difference

Qin Jun[1], *Lin Yechuan*[1], *Yi Yunfei*[2]

(1 College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China;
2 College of Computer and Information Engineering, Hechi University, Yizhou 546300, China)

Abstract The traditional method of Zhuangwen word segmentation is to use the space between words as a separation mark. But in most cases, the word segmentation method will destroy multiple words association combination of semantic words which express the complete and independent semantic information. For the first time we use the mutual information to improve algorithm MI^k and t-test difference in Zhuangwen text word segmentation that based on the use of mutual information MI method to measure the degree of correlation between adjacent words, and combine with the two in the evaluation of adjacent words' static binding ability and dynamic binding ability, a TD-MI^k hybrid algorithm based on the MI^k and t-test difference is proposed. The segmentation effects of MI^k, t-test difference and TD-MI^k hybrid algorithm are compared. We use the text set on the People´s network in Zhuangwen as a training and test corpus to do the experiments. The experimental results show that the three segmentation methods can extract the semantic words in text accurately and efficiently, and TD-MI^k hybrid algorithm has the highest accuracy of word segmentation .

Keywords zhuangwen word segmentation; MI improved algorithm; t-test difference; hybrid algorithm; semantic word

壮语是汉藏语系壮侗语族壮傣语支的一种语言,目前存世的壮族文字有古壮文和现代壮文,其中,现代壮文(简称壮文)是一种拼音文字[1,2]. 在互联网发展迅速的时代,专注于壮文交流的有人民网壮文版、壮族在线、鼓歌壮族、壮族娱乐网等. 分词是文本信息处理的第一道“工序”,在自然语言处理的许多应用领域,如机器翻译、文本分类、信息检索等扮演着极其重要的角色. 目前,在分词这个研究领域,汉文分词的研究成果已经很多也相对成熟,汉文分词技术发展到今天,大体上可以将其分为这几类: 基于词典的分词方法(又称机械分词)、基于统计的分词方法、基于理解的分词方法、词典与统计相结合的分词方法[3]. 汉文分词的难点主要有两个: 消除歧义和未登录词. 对于消除歧义问题,需要结合上下

收稿日期 2017-09-01 * **通讯作者** 易云飞,研究方向: 智能计算,E-mail: gxyiyf@163.com
作者简介 覃 俊(1968-),女,教授,博士,研究方向: 智能优化、数据挖掘,E-mail: 498011695@qq.com
基金项目 国家科技支撑计划项目子课题(2015BAD29B01);中南民族大学研究生学术创新基金项目(2017sycxjj051)

文语境,因此,基于理解的分词方法能够较好的克服这个难点,而基于统计的分词可以较好地克服未登录词这个难点[4]. 壮文分词与汉文有很大的相似性,但是,对壮文的文本信息处理的研究成果较少:2011 年南宁平方软件开发了一款壮汉翻译试验系统,翻译结果的可理解率大约在 40%;同年,广西民族大学的学生开发了一款基于短语的汉壮统计机器翻译系统[2];2016 年由中国民族语文翻译局与东北大学自然语言处理实验室共同研发的壮文与汉文智能翻译系统正式上线运行,这是国内首套壮文与汉文双向翻译系统;除此之外,对于壮文文本信息处理的研究成果基本没有较为详实的公开报道. 因此,对于壮文分词算法的研究具有重要意义.

壮文文本是一种以空格分隔开的词序列,传统的壮文分词只需把空格标志作为分词方法即可. 但在很多情况下,壮文中的多个单词的关联组合模式,也是一种不可分割的独立语言单元,这种多个单词的组合将会表达一个具体而独立的语义信息,用空格隔开的分词方法将会丧失这种单词组合所要表达的完整语义信息. 这样获取的单词就难以在文本标引中发挥词的作用,相应的各种文本处理效果也很不理想[5,6].

在壮文的机器翻译中,按传统的分词方法翻译,例如"raemx hawq raen bya",这是一个固定的词组,汉文意思是"水落石出",如果以空格分隔,借助 Stoneman、honghlaj 等制作的 Sawloih Cuengh-Gun 壮汉词典,按每个单词去翻译,意思就变为"水干见山(石)". 这样就很难正确表达出一个单词组合所要表达的语义信息,大大降低了翻译的准确性.

在信息检索中,用传统的分词方法检索出来的文本信息并不全是与搜索关键字相关的结果. 例如关键字"sevei cujyi"(社会主义),传统分词会把它分成"sevei"和"cujyi". 用该关键字在百度上进行测试,返回了相关结果约 6440 个,其中返回的第一个页面的 10 个结果中,有 4 个结果是与该关键字不相关的(2017 年 5 月 10 日测试),结果不令人满意.

此外,传统的空格分词方法还会在文本主题词提取、文本分类及聚类等文本处理中产生负面的影响. 因此,研究一种能够从壮文文本中获取结构稳定、语义完整而独立的壮文单词的组合将对壮文的各种文本处理效果起到重要的积极作用. 本文从汉文分词和其他一些少数民族语言(如维吾尔文)的相关研究工作中得到启发,使用互信息的方法作为壮文相邻单词间关联程度的度量,切分壮文文本中能够独立表达完整的语义信息的单词组,并提出一种基于互信息改进算法 MI^k 和 t-测试差相结合的 TD-MI^k 混合算法,更准确而有效地提取文本中的语义词.

1　基于互信息的壮文分词

1.1　互信息 MI 的壮文分词思路及分词过程

根据互信息原理,对于以空格分隔的壮文词串序列 $W_i\ W_{i+1}$、单词 W_i 和 W_{i+1} 之间的互信息 MI(Mutual Information)定义如下:

$$\mathrm{MI}(W_i,W_{i+1}) = \log_2 \frac{P(W_i,W_{i+1})}{P(W_i)\,P(W_{i+1})}, \quad (1)$$

其中,$P(W_i,W_{i+1})$ 表示词串序列 $W_i\ W_{i+1}$ 在文本集中出现的概率;$P(W_i)$ 表示单词 Wi 在文本集中出现的概率;$P(W_{i+1})$ 表示单词 W_{i+1} 在文本集中出现的概率. 互信息 $\mathrm{MI}(W_i,W_{i+1})$ 反映了相邻单词 W_i 和 W_{i+1} 之间的关联程度:若 $\mathrm{MI}(W_i,W_{i+1}) \geqslant 0$,则 $W_i\ W_{i+1}$ 间是强关联的,当 $\mathrm{MI}(W_i,W_{i+1})$ 大于给定的一个阈值时,认为 $W_i\ W_{i+1}$ 可以构成一个不可分割的独立语言单元;若 $\mathrm{MI}(W_i,W_{i+1}) \approx 0$,则 $W_i\ W_{i+1}$ 间是弱关联的,表示 $W_i\ W_{i+1}$ 很难构成一个独立语言单元;若 $\mathrm{MI}(W_i,W_{i+1}) < 0$,则 $W_i\ W_{i+1}$ 间是互斥的,表示 $W_i\ W_{i+1}$ 间基本不能构成一个独立语言单元.

基于互信息 MI 的壮文分词过程如下:每次从句子文本集 S 中依次读取一个句子 $S_i(1 \leqslant i \leqslant n)$,并从 S_i 中的第一个单词开始,从左到右依次扫描词串,以两个单词为一组,统计单词 W_i,W_{i+1} 以及它们的组合 $W_i\ W_{i+1}$ 在训练文本集 Ds 中的频度,并根据公式(1)计算 $W_i\ W_{i+1}$ 间的互信息 $\mathrm{MI}(W_i,W_{i+1})$. 若 $\mathrm{MI}(W_i,W_{i+1}) \geqslant T_i$($T_i$ 为给定的阈值),则认为当前组合可以构成一个独立的语义词. 然后把 $W_i\ W_{i+1}$ 看成一个新的单词,并与下一个单词 W_{i+2} 组合(本文限制最大词串的单词数为 4),同样地计算它们之间的互信息;依此类推,将问题始终简化为计算相邻两个单词之间的互信息,判断它们是否能构成独立的语义词. 组词过程如图 1 所示.

互信息的计算公式始终不变,在组词过程中,当计算得到的互信息小于对应的阈值时,说明后续新加入单词更不可能构成词. 所以,应把当前加入的新的单词作为第一个单词,开始新一轮的组词. 考虑到组词长度(单词个数)的影响,根据组词长度的增加,其对应的阈值也相应地变小($T_1 > T_2 > T_3 > \cdots > T_{n-1}$).

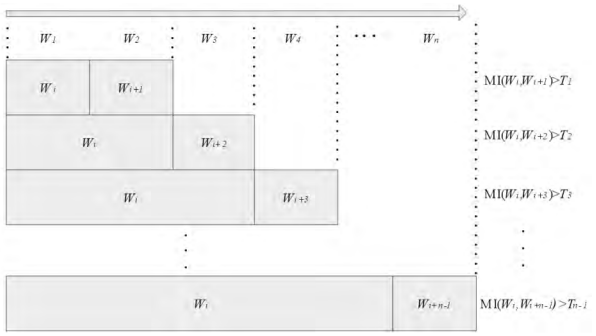

图1 组词过程图

Fig. 1 The graph of Lexical process

1.2 改进互信息算法 MI^k

基于互信息MI的壮文分词的特点是算法简单、速度快. 但在分词过程中发现,存在部分两个低频单词且总是一起出现的词串,MI方法会过高地估计包含这些低频词串的结合强度. 例如, "daihheiq mokfaenx"(大气雾霾)、"canzciz ginhyinz"(残疾军人)等这些低频词在训练语料中低频且总是相邻出现,这些词串的MI值非常高,导致包含这些低频词的垃圾词串相应的MI值也非常高,例如"gij daihheiq mokfaenx"(的大气雾霾),明显不符合分词要求. 为了过滤掉这些垃圾词串,克服高估低频词串结合强度这个缺点,我们首次采用改进互信息MI^k算法对壮文进行分词,MI^k算法是在MI方法基础上引进一个或者多个字串A与B的联合概率因子$P(A,B)$[7,8]. MI^k算法的定义如下:

$$MI^k(A,B) = \log_2 \frac{P^k(A,B)}{P(A)P(B)}; k \in N^+, \quad (2)$$

其中$P(A)$、$P(B)$分别表示单词A、B的概率,$P(A,B)$表示词串A B的联合概率,$MI^k(A,B)$表示词串A B之间的相关度,也称MI^k值. 特殊地,当$k=1$时,MI^k算法即为MI算法[9,10]. 我们分别对k值在1~10之间的取值做了实验,得到当$k=2$或3时,分词效果有比较明显地提高,k的取值对分词准确率的影响见第3节实验.

MI^k算法的壮文分词思路: 对于每一个由四个单词组成的壮文词串序列x A B y,计算中间两个单词串A B和前面两个单词串x A的MI^k值的平均值$average1$以及中间两个单词串A B和后面两个单词串B y的MI^k值的平均值$average2$. 计算公式如下:

$$average1 = \frac{MI^k(A,B) + MI^k(x,A)}{2}, \quad (3)$$

$$average2 = \frac{MI^k(A,B) + MI^k(B,y)}{2}, \quad (4)$$

对于词串序列x A B y,如果满足:

$$MI^k(A,B) > MI^k(x,A) + average1, \quad (5)$$

$$MI^k(A,B) > MI^k(B,y) + average2, \quad (6)$$

则认为词串A B能构成独立语义词或者是独立语义词组成的一部分的概率较大; 否则,认为单词A B是各自成词或者是词的边界的概率较大.

1.3 改进互信息算法 MI^k 的优势及局限性

从互信息的计算公式(1)和(2)可知,互信息考虑的是相邻单词之间的内部结合强度,与上下文无关,反映了单词之间的静态结合能力. 实验发现,改进的互信息方法不仅能够有效过滤掉低频词的垃圾词串,还能够更好地保证由多个单词组成的长词串语义词不被遗漏,像"it rangh it loh"(一带一路)等这样的新词串能够更好地提取出来,但同时也存在不足的地方,在连接词的切分准确度较低. 例如"caenleix caeuq fazcanj"(真理和发展),MI^k算法把这三个单词组合当作一个独立的语义词,因为这三个单词组合的频度较高,它只考虑了单词间的静态结合,但没有考虑上下文单词之间结合趋势. 而正确的分词结果应该以中间的连词"caeuq"为边界切分成三个词. 因此,如果能有一种能够衡量上下文单词间的动态结合趋势的方法,来弥补互信息这种静态结合的局限性,分词的准确性将会提高.

2 TD-MI^k 混合算法的壮文分词

2.1 t-测试差

为寻找一种能够衡量壮文上下文单词间的动态结合趋势的方法,我们首次采用t-测试差对壮文进行分词. Church[11]等首次引入t-测试,以度量一个英文单词A与其它任意两个单词x和y的结合紧密程度. 根据定义,壮文单词串x y z的t-测试值如下公式所示:

$$t_{x,z}(y) = \frac{p(z \mid y) - p(y \mid x)}{\sqrt{\sigma^2(p(z \mid y)) - \sigma^2(p(y \mid x))}}, \quad (7)$$

其中$p(y \mid x)$,$p(z \mid y)$分别是y关于x,z关于y的条件概率,$\sigma^2(p(y \mid x))$,$\sigma^2(p(z \mid y))$代表各自的方差. 由t-测试值的定义可知: 若$t_{x,z}(y) > 0$,则y与后继z的结合强度大于与前驱x的结合强度,此时y应与x分开,而与z组词; 若$t_{x,z}(y) = 0$,则无法判断y要与哪个分开或者组词; 若$t_{x,z}(y) < 0$,则y与前驱x的结合强度大于与后继z的结合强度,此时y应与z分开,而与x组词.

t-测试是基于字的统计量,而不是基于字间位置,因此为了能够在汉文分词中直接用来计算相邻

字间连断概率,清华大学孙茂松教授等人提出了 t-测试差的概念[12]. 根据定义,对于壮文单词串 $x\ A\ B\ Y$,相邻单词 A,B 之间的 t-测试差值计算如下所示:

$$\mathrm{TD}(A,B) = t_{x,B}(A) - t_{A,y}(B). \quad (8)$$

当 $\mathrm{TD}(A,B) > T$(T 为阈值) 时,$A\ B$ 的单词间位置更倾向于连,反之倾向于断. 与互信息不同,t-测试差考虑的是单词之间的相对结合强度,是一个单词与上下文的结合趋向,反映了相邻单词之间的动态结合能力. 从分词的结果发现,t-测试差方法在连接词的切分准确度更高,例如"caenleix caeuq fazcanj"(真理和发展),t-测试差能够将这个词串分成 3 个词.

2.2　TD-MI^k 混合算法

从前面的分析中可以知道,互信息反映的是单词之间的静态结合能力,而 t-测试差反映的是单词之间的动态结合能力,两种方法在壮文的分词中各有优势,但各自又存在局限性. 例如: 改进互信息 MI^k 方法能够准确提取"it rangh it loh"(一带一路)等这样的新词串,而 t-测试差方法不能; t-测试差方法能够准确地把"caenleix caeuq fazcanj"(真理和发展)切分成三个词,而改进互信息 MI^k 方法不能. 因此,如果能够把互信息和 t-测试差这两个统计原理相结合,起到互补效果的可行性极大. 鉴于此,我们将改进的互信息方法 MI^k 与 t-测试差相结合发现,该方法能够在一定程度起到互补的作用,既能提取"it rangh it loh",又能将"caenleix caeuq fazcanj"正确切分成三个词. MI^k 与 t-测试差组合的 TD-MI^k 的混合算法的计算公式如下:

$$\text{TD-MI}^k(A,B) = \alpha^* \mathrm{TD}(A,B) + \beta^* \mathrm{MI}^k(A,B), \quad (9)$$

其中,α、β 分别是 t-测试差和 MI^k 算法的权重因子,它们的和为 1,具体取值见下一节实验判断合适的取值.

TD-MI^k 混合算法的分词思路: 对于壮文单词串 $x\ A\ B\ y$,计算中间词串 $A\ B$ 的 TD-MI^k 的值,当 TD-MI$^k(A,B) > T$(T 为阈值)时,则认为词串 $A\ B$ 能构成独立语义词或者是独立语义词组成的一部分的概率较大; 否则,认为词串 $A\ B$ 是各自成词或者是词的边界的概率较大. TD-MI$^k(A,B)$ 既能在 $\mathrm{TD}(A,B)$ 和 $\mathrm{MI}^k(A,B)$ 两者判断一致时保持判断不变,又能在两者判断不一致时,在一定程度上得到互补. 例如"caeuq fazcanj"的 MI^k 值为 -5.97,判断为连,而 TD 值为 -9.78,判断为断,混合后的 TD-MI^k 值为 -8.26,判断为断,把两个单词切分开.

3　实验与分析

3.1　实验数据集

为了验证算法的可行性及准确率,从人民网壮文版搜集所有壮文文本,随机选取一组文章作为测试文本,并以中国民族语文翻译局的翻译系统为辅助工具,对测试文本做人工标记. 壮文文本训练语料的大小约为 2.8MB,主要为政府工作报告文章及政治新闻文章.

3.2　评价指标

本文采用准确率、召回率和 F 值 3 个指标来衡量分词算法的性能,计算公式如下:

$$\text{准确率} = \frac{\text{切分结果正确的词数}}{\text{切分结果的总词数}} \times 100\%, \quad (10)$$

$$\text{召回率} = \frac{\text{切分结果正确的词数}}{\text{分词后应得到的总词数}} \times 100\%, \quad (11)$$

$$F = \frac{\text{准确率} \times \text{召回率} \times 2}{\text{准确率} + \text{召回率}} \times 100\%, \quad (12)$$

其中,切分结果正确的词数(C_1)是指测试文本根据分词算法切分后切分正确的词数; 切分结果的总词数(C_2)是指测试文本根据分词算法切分后得到的总词数; 分词后应得到的总词数(C_3)是指测试文本人工切分后得到的总词数; F 值反应的是根据准确率和召回率得出的算法的综合性能指标.

3.3　实验及结果分析

使用 C++ 语言,Visual Studio 2015 为实验工具,对壮文进行分词实验. 对于改进互信息 MI^k 算法中,k 值的选取与准确率变化的趋势图如图 2 所示.

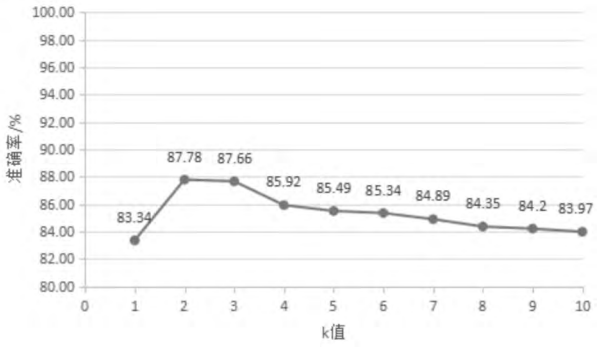

图 2　随着 k 值变化的准确率变化趋势

Fig. 2　The change tendency of accuracy with the change of k value

从图 2 中可以看出,当 $k=2$ 或 3 时,分词准确率有比较明显的提高.

在 TD-MI^k 混合算法中 α、β 的取值实验如图 3 所示.

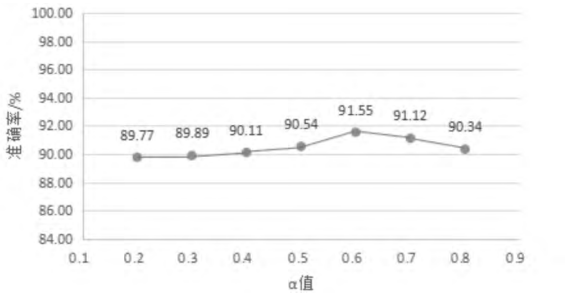

图3 随着 α 值变化的准确率变化趋势

Fig. 3 The change tendency of accuracy with the change of alpha value

图3中 α 是 t-测试差的权重因子,相应的 MI^k 算法的权重因子 β 值为 $1-\alpha$. 从图中可看出当 $\alpha=0.6$ 时为图中曲线波峰,分词的准确率最高.

分别用传统分词方法、改进互信息的 MI^k 算法、基于 t-测试差算法以及 TD-MI^k 混合算法对测试文本进行切分得到四种词汇表,如表1所示. 其中,测试文本分词后应得到的总词数即人工切分后得到的总词数 C_3 为9384.

表1 不同分词方法对应切分结果词汇表

Tab. 1 The different segmentation methods correspond to segmentation results

分词方法	词汇类型及个数					
	单词	双词	三词	四词	总词数(C_2)	正确词数(C_1)
传统分词	11590	–	–	–	11590	7560
改进互信息 MI^k 分词算法($k=2$)	8164	935	228	218	9545	8382
基于 t-测试差分词算法	8150	967	218	213	9548	8445
TD-MI^k 混合算法	8310	911	206	210	9637	8823

从表1可以看出传统的分词方法将文本集中大约25%的单词组合而成的语义词拆分成单个单词,而这种多个单词的组合将会表达一个具体而独立的语义信息,包括一些人名、成语、专有名词等,拆分成多个单词将会丧失这种单词组合所要表达的完整语义信息. 因此,用传统分词方法获取的单词就难以在文本标引中发挥词的作用,相应的各种文本处理效果也会受到消极影响. 相比较而言,使用改进互信息的 MI^k 算法、基于 t-测试差算法以及 TD-MI^k 混合算法作为分词算法所获得的词语不止有单个词,还有双词、三词、四词,这样的词就能更好的表达完整的语义信息,构成更能表达文本信息的语义词,这将有助于提升各种文本处理的效果. 例如,在文本的主题词提取中,一般的主题词都是由多个单词组合而成的,这样就有效地提高了主题词提取的准确性. 表2中列举了2个例句的分词实验结果.

表2 分词实验结果举例

Tab. 2 The examples of segmentation results

例句	实验结果
Doicaenh hwnqguh rangh dieg ginghci diuz roen seicouz.	Doicaenh $ hwnqguh $ rangh dieg ginghci $ diuz roen seicouz $ 推进 $ 建设 $ 经济带 $ 丝绸之路 $
Siz Ginbingz cujsiz daengj bouxdaeuz guekgya okbae fangjvwn lai guek.	Siz Ginbingz cujsiz $ daengj $ bouxdaeuz guekgya $ okbae fangjvwn $ lai guek $ 习近平主席 $ 等 $ 国家领导人 $ 出访 $ 多国 $

表2的第一个例句中的“diuz roen seicouz”(丝绸之路)由3个单词组成一个专有名词,是不可分割的,实验结果表明本文使用的分词算法能够正确地把这单词组合切分出来. 此外,壮文在某些方面与汉文有很大的相似性,汉文方面,两个单一的字组合成一个词语,例如“建设”是一个词语;在壮文方面,第二个例句中的“lai guek”(多国),也是由两个单词组合成一个词语,把这样的单词组合切分出来对后续的机器翻译或者信息检索等都有很大的促进作用. 显然,本文使用的分词算法很好的切分出壮文句子中独立而完整的语义词,且切分效果也是令人满意的.

本文使用的几种分词方法由于分词的策略不同,在分词的效果上也有差异. 不同分词方法的分词效果差异对比如表3所示.

表3 分词方法效果对比

Tab. 3 The comparison map of segmentation method

分词方法	准确率/%	召回率/%	F值/%
改进互信息 MI^k 分词算法($k=2$)	87.78	89.32	88.54
基于 t-测试差分词算法	88.45	89.99	89.21
TD-MI^k 混合算法	91.55	94.02	92.77

从表3可以看出,本文使用的分词方法都能得

到较高的准确率和召回率,对应的 F 值也较高. 互信息反映的是单词之间的静态结合能力,而 t-测试差反映的是单词之间的动态结合能力,两者都各有优势,t-测试差的分词准确率相比改进互信息 MI^k 方法稍微好一些,而 TD-MI^k 混合算法则结合两者的特点,分词的准确率和召回率分别提高了 3.77% 和 4.7%.

4　结语

本文分析了壮文文本中多个单词组合所表达的完整而独立的语义信息,以及传统壮文分词方法对这种完整而独立的语义信息的破坏,同时在各种文本信息处理中所获得的结果不令人满意. 为了能够更好地提取文本中的这种能够更好的表达完整的语义信息的语义词,在使用互信息 MI 方法来度量壮文相邻单词间关联程度的基础上,提出一种基于互信息改进算法 MI^k 和 t-测试差相结合的 TD-MI^k 混合算法对壮文文本分词,并用准确率、召回率和 F 值对分词结果进行评价. 实验表明本文的分词算法的分词结果得到较高的准确率和召回率,能够较准确而有效地提取文本中的语义词,提出的 TD-MI^k 混合算法也有效地提高了分词的准确率. 另外,由于目前网上的壮文文本大都是政府工作报告文章和政治类新闻文章,所以本文的分词算法在对在政府工作报告和政治类壮文文本的分词效果相对较好,该分词算法同样也适用于壮文的其他各类文本.

参　考　文　献

[1] 韦景云,覃晓航. 状语通论 [M]. 北京: 中央民族大学出版社,2006: 3-110.

[2] 刘连芳,顾　林,黄家裕,等. 壮文与壮文信息处理 [J]. 中文信息学报,2011,25(6): 175-182.

[3] 赵秦怡,王丽珍. 一种基于互信息的串扫描中文文本分词方法 [J]. 情报杂志,2010,29(7): 161-162.

[4] Min K, Ma C, Zhao T, et al. BosonNLP: An ensemble approach for word segmentation and POS tagging [C] // Springer. The 4th CCF Conference on Natural Language Processing and Chinese Computing (NLPCC2015). Berlin: Springer, 2015: 520-526.

[5] 吐尔地·托合提,艾克白尔·帕塔尔,艾斯卡尔·艾木都拉. 基于互信息的维吾尔文自适应组词算法 [J]. 计算机应用研究,2013,30(2): 429-431.

[6] 吐尔地·托合提,艾克白尔·帕塔尔,艾斯卡尔·艾木都拉. 语义词特征提取及其在维吾尔文文本分类中的应用 [J]. 中文信息学报,2014,28(4): 140-144.

[7] Bouma G. Normalized (pointwise) mutual information in collocation extraction [C] // UIMA. Proceedings of German Society for Computational Linguistics (GSCL 2009), Potsdam: UIMA, 2009: 31-40.

[8] Pazienza M, Pennacchiotti M, Zanzotto F. Terminology extraction: an analysis of linguistic and statistical approaches [J]. Springer Berlin Heidelberg, 2005, 185: 255-279.

[9] 杜丽萍,李晓戈,于　根,等. 基于互信息改进算法的新词发现对中文分词系统改进 [J]. 北京大学学报(自然科学版),2016,52(1): 35-40.

[10] 杜丽萍,李晓戈,周元哲,等. 互信息改进方法在术语抽取中的应用 [J]. 计算机应用,2015,35(4): 996-1000.

[11] Church K W, Gale W, Hanks P, et al. Using Statistics in Lexical Analysis [M]. Hillsdale NJ : Lawrence Erlbaum Associates, 1991: 115-164.

[12] 孙茂松,肖　明,邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词 [J]. 计算机学报,2004,27(6): 736-742.