基于动态文本窗口和权重动态分配的中文文本纠错方法

黄改娟[1,2]，　王匆匆[1]，　张仰森[1,2]

(1. 北京信息科技大学 智能信息处理研究所　北京 100101；2. 网络文化与数字传播北京市重点实验室　北京 100101)

摘要：提出一种基于动态文本窗口的中文文本查错方法,依靠窗口的不断滑动来检测文本错误。当中文文本有疑似错误时,采用聚类词集平滑数据稀疏问题,然后采用权重动态分配的纠错词集进行纠错,若纠错结果仍不符合检错规则,则用缩小文本窗口法和拓展窗口法来检查具体错误。构建纠错词集则采用基于最小编辑距离和权重动态分配的方法。实验结果表明,基于动态文本窗口查错方法的 F 值达到了 77.9%;再结合权重动态分配的纠错方法,纠错准确率达到 78.1%,相较黑马校对系统和基于平均权重的纠错策略,准确率分别提升了 9.7%和 15.8%。

关键词：语义搭配；数据稀疏；动态文本窗口；权重动态分配

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1671-6841(2020)03-0009-06

DOI：10.13705/j.issn.1671-6841.2019393

0　引言

随着人们从纸质书写转变为键盘输入,写作效率不断提升,但产生的中文文本错误影响了人们对语言的理解效率。在中文文本错误中,虽然大多数人能根据自己的经验去推测原来正确文本的含义,但是对于那些刚开始学习汉字的人群以及对文字准确度要求很高的出版社来说,这种细微的错误不仅体现了人们对工作的态度,还体现出人们对文字的重视程度。因此,中文文本的纠错不但具有丰富的需求背景,而且它还是自然语言处理领域中重要的基础问题。目前,依靠人工纠错文本不但费时费力,还可能导致二次错误。实现中文文本的自动纠错将会节约编辑或审稿人的时间,从而实现更精准的人工文本纠错,对教育出版行业以及办公自动化将产生深远影响。在文本错误类型的分类上,当前研究者们主要从字词的拼写错误、词语搭配不当以及文本语义准确度入手,文本的自动纠错方法则是基于 N 元文法、散串和语义检测的纠错方法。本文模拟人工纠错时的阅读过程,提出了基于动态文本窗口的文本自动检错算法。此外,在已有词典的基础上,使用基于最小编辑距离并具有权重动态分配功能的纠错算法体现了纠错词语选取的合理性。

1　相关工作

1.1　中文文本错误类型

中文文本错误类型可分为字词层面的错误[1]、语法层面的错误、语义层面的错误。在这三个层面中,字词层面最为基础,语法层面其次,语义层面最接近语言的思维表达形式。另外,这三个层面的联系可体现为:字词层面错误会导致整个词语的词性变动或者产生错词,从而使语句在语法层面不符合语法规则,进而引起语义模糊的问题。因此,字词层面的错误也会牵扯到语法和语义层面,可见字词层面的纠错是文本纠错的基础。文献[2]将字词层面的错误分为错字、多字和漏字三类。此后,文献[3]定义了“真多字词错误”和“非

收稿日期:2019-08-29

基金项目:国家自然科学基金项目(61772081);国家重点研发计划项目(2018YFB1402901);科技创新服务能力建设-科研基地建设-北京实验室-国家经济安全预警工程北京实验室项目(PXM2018_014224_000010)。

作者简介:黄改娟(1964—),女,山西临猗人,高级实验师,主要从事自然语言处理研究,E-mail:hgj@bistu.edu.cn;通信作者:王匆匆(1998—),男,安徽阜阳人,硕士研究生,主要从事自然语言处理研究,E-mail:837482294@qq.com。

多字词错误”。“真多字词错误”的定义源自于语句的分词结果中没有任何错词,但是这些词语会与相邻词语产生搭配不当的关系;“非多字词错误”即分词结果中包含了由单字组成的词语所引起的文本错误。

1.2 中文文本纠错方法

1.2.1 基于N元文法的文本纠错 基于N元文法的纠错方法根据Markov随机过程理论,针对字词的参数空间过大而产生的数据稀疏问题[4]简化N的大小,即当前状态仅与前面有限数量的状态有关,以此类推,判断整个语句的通顺程度。由于中文句子的书写规范常常受字词之间搭配习惯的影响[5],因此基于N元文法的纠错方法主要利用了字词的前后接续程度[6]、互信息[7]、共现频率[8]等文本的相关信息提供纠错建议。

1.2.2 基于散串的文本纠错 散串是指文本经分词后连续出现的零散字符串,文献[9]把散串区间作为错误的初始范围。文献[10]统计了基于散串的查错效果,指出错误文本被分词后所产生的单字、双字以及多字串的概率为90.3%。文献[11]使用散串纠错文本,纠错过程分为错误定位和错误矫正两步。错误定位采用了统计字词搭配度和统计句子语法架构的方法定位错误位置,其中统计搭配度的错误定位方法主要针对五笔字型和拼音两种类型的输入法,统计语法架构的错误定位方法则是将统计法获取的错误位置再使用语法规则进行精准定位。错误矫正则采用传统的人机交互法,由系统提供候选纠错词供用户选择。

1.2.3 基于语义检测的文本纠错 文本语义错误的自动纠错属于文本纠错研究领域较难解决的问题[11]。在语义错误查找方面,一般方法是通过词语搭配知识库进行查错。文献[12]统计了大量语义搭配实例,并构建语义搭配规则库用以检查语义级错误。文献[13]首先以《现代汉语实词搭配词典》和知网为基础建立了语义搭配知识库,然后结合证据理论解决了纠错词选取的模糊性问题。文献[14]认为,真词错误存在的原因主要是词语语义表达不符合词语搭配习惯,因此提出了专门面向真词错误的易错词集、在特定词语语境下词语的泛化模型和常用语言模型结合词语使用规则的真词错误自动纠错方法。

2 基于动态窗口的文本查错方法

2.1 基于动态窗口的文本查错思路

人们在阅读文本时,注意力的视野集中范围是不断向后移动的,尤其是在精读的情况下大都在看文本中的几个字,有时还会向前面观察一遍再联系后面的文字体会文本的语义。依据这种阅读特点,首先采用动态文本窗口的方法来模拟人们在阅读时注意力的变化范围,然后在此范围中,从中心向前展开计算文本搭配的合理程度,再将文本窗口不断地向后移动以检测后面的文本。在文本查错过程中,距离更近的字词则是依靠缩小窗口的方法检测字词方面的正误。随着窗口的缩小,文本里真正的错误又可能是由后面文本所导致的,于是采用拓展窗口的方法确定具体错误的位置。

2.2 动态窗口的相关定义

2.2.1 动态文本窗口 假设文本语句可被$T=\{c_1,c_2,\cdots,c_i,\cdots,c_{n-1},c_n\}$表示为一种文本的窗口,语句长度为$N$,最小为1。以某个字$c_i(i=1,2,\cdots,n)$为中心,在它左右相邻的$n$个字符上建立一个宽度为$2n-1$的动态文本窗口,表示为$Win_n(c_i)=\{c_{i-n+1},\cdots,c_{i-1},c_i,c_{i+1},\cdots,c_{i+n-1}\}$。当$c_i$的$i=0$时,$Win_n(c_i)=\{c_i,c_{i+1},\cdots,c_{i+n-1}\}$;当$c_i$的$i=n$时,$Win_n(c_i)=\{c_{i-n+1},\cdots,c_{i-1},c_i\}$。

2.2.2 动态窗口缩小 当窗口大小从N缩小为$N-1$时,表示为$Win_{n-1}(c_i)=\{c_{i-n+2},\cdots,c_{i-1},c_i,c_{i+1},\cdots,c_{i+n-2}\}$。当$c_i$的$i=0$时,$Win_{n-1}(c_i)=\{c_i,c_{i+1},\cdots,c_{i+n-2}\}$;当$c_i$的$i=n$时,$Win_{n-1}(c_i)=\{c_{i-n+2},\cdots,c_{i-1},c_i\}$。

2.2.3 动态窗口后移 窗口为N,中心字符为c_i时,窗口后移表示为令$c_i=c_{i+1}$。

2.2.4 动态窗口拓展 窗口为N,中心字符为c_i时,经拓展后的窗口表示为$Win_nExt(c_i)=\{c_{i-1},c_i,c_{i+1},c_{i+2}\}$。

2.2.5 动态窗口包含文字的前项概率 窗口为N,中心字符为c_i时,c_i的前项概率表示为

$$PF_0^{-n-1}(c_{i-n+1},\cdots,c_{i-1},c_i)=\frac{N(c_{i-n+1},\cdots,c_{i-1},c_i)}{N(x_{i-n+1},\cdots,x_{i-1},x_i)},$$

式中:$N(c_{i-n+1},\cdots,c_{i-1},c_i)$表示在训练语料中字符串$\{c_{i-n+1},\cdots,c_{i-1},c_i\}$出现的次数;$N(x_{i-n+1},\cdots,x_{i-1},x_i)$为任意$n$长度的字符串$\{x_{i-n+1},\cdots,x_{i-1},x_i\}$在训练语料中出现的次数。

2.2.6 动态窗口包含文字的后项概率 窗口为N,中心字符为c_i时,c_i的后项概率表示为

$$PB_0^{-n-1}(c_i,c_{i+1},\cdots,c_{i+n-1})=\frac{N(c_i,c_{i+1},\cdots,c_{i+n-1})}{N(x_i,x_{i+1},\cdots,x_{i+n-1})},$$

式中:$N(c_i,c_{i+1},\cdots,c_{i+n-1})$ 表示在训练语料中字符串 $\{c_i,c_{i+1},\cdots,c_{i+n-1}\}$ 出现的次数;$N(x_i,x_{i+1},\cdots,x_{i+n-1})$ 为任意 n 长度的字符串 $\{x_i,x_{i+1},\cdots,x_{i+n-1}\}$ 在训练语料中出现的次数。

2.3　基于动态窗口的文本查错算法

基于动态窗口的文本查错算法的具体步骤如下。

Step 1:将文本分句,去除标点及特殊符号。

Step 2:若仍有句子未被检测,算法继续。否则,算法结束。

Step 3:以待查字符 c 为中心,构造大小为 3 的动态文本查错窗口 $Win_3(c_i)=\{c_{-2},c_{-1},c,c_1,c_2\}$。

Step 4:设字符 c 的动态文本窗口 $Win_3(c)$ 的左侧相邻字符串 $\{c_{-2},c_{-1},c\}$ 的共现概率为 $PF_0^{-2}(c)$,字符 c 的动态文本窗口 $Win_3(c)$ 的右侧相邻字符串 $\{c,c_1,c_2\}$ 的共现概率为 $PB_0^{-2}(c)$。$threadF_3$ 表示 $Win_3(c)$ 的前项概率判定的阈值,$threadB_3$ 表示 $Win_3(c)$ 的后项概率判定的阈值。

Step 5:如果 $-\log(PF_0^{-2}(c))<threadF_3$,则 $c\leftarrow c_1$,转向 Step 2。否则,转向 Step 6。

Step 6:如果左侧相邻字符串 $\{c_{-2},c_{-1},c\}$ 存在易混淆词库中,则选择易混淆词集处理,然后转向 Step 5。否则,转向 Step 7。

Step 7:使用大小为 2 的窗口检查 $\{c_{-2},c_{-1}\}$ 是否存在错误。如果是,则输出错误字符,转向 Step 8。否则,转向 Step 10。

Step 8:如果 $-\log(PB_0^{-2}(c))\leqslant threadB_3$,则选择聚类词平滑策略进行处理,转向 Step 10。

Step 9:如果 $-\log(PB_0^{-2}(c))>threadB_3$,则认为 $\{c,c_1\}$ 是一个可疑错误区间,发出一个错误警告。执行 $c\leftarrow c_1$,转向 Step 2。

Step 10:缩小查错文本窗口大小为 2,$Win_2(c_i)=\{c_{-1},c,c_1\}$,重复 Step 11~14。

Step 11:令 $PF\_BI_0^{-1}(c)$ 表示 $Win_2(c)$ 的左侧部分 $\{c_{-1},c\}$ 的共现概率,$threadB_2$ 是后项概率判定的阈值。$PF\_BI_0^{1}(c)$ 表示 $Win_2(c)$ 的右侧部分 $\{c,c_1\}$ 的共现概率,$threadF_2$ 是前项概率判定的阈值。

Step 12:如果 $-\log(PF\_BI_0^{1}(c))>threadF_2$ 且 $-\log(PF\_BI_0^{-1}(c))\leqslant threadB_2$,拓展查错窗口,即在 $Win_2(c)$ 的右侧增加一个字符 c_2 变为 $Win_2Ext(c)=\{c_{-1},c,c_1,c_2\}$,设它的概率判定的阈值为 $thread\_expend$。

Step 13:如果 $-\log(\frac{N(c_1)N(c_2)}{N(c_1)+N(c_2)})>thread\_expend$,则认为 $\{c_1,c_2\}$ 是一个可疑错误区间,发出错误警告,$c\leftarrow c_1$,转向 Step 2。

Step 14:如果 $-\log(\frac{N(c_1)N(c_2)}{N(c_1)+N(c_2)})\leqslant thread\_expend$,则不认为 $\{c_1,c_2\}$ 是一个可疑错误区间,$c\leftarrow c_1$,转向 Step 2。

2.4　用于查错的聚类词平滑策略

聚类词的概念是把近义词、同义词和一些可以互换使用的字词作为同一个词来对待。将这类可以在同样语境下使用的词语当作同一种词,不但符合文本的常用表达方式,而且减少了查错模型因数据稀疏而导致的误报率。表 1 列出了部分聚类词。聚类词库的使用方法为:假设待检测动态文本窗口的文本出现次数小于常用文本阈值,并且该动态文本窗口内的文本包含某些在聚类词库的词语,于是将词库中与待检测词语聚为一类的词在当前窗口内文本语境下所出现的概率累加到待检测动态文本窗口,然后再次检测此处动态文本窗口内是否包含错误。

表 1　部分聚类词

Table 1　Partial clustering words

词类	词语
人称代词	同乡、老乡、乡亲、故乡人、父老乡亲
数词	巨额、千万、万万、亿万、大量、大批、成批
量词	把、支、辆、对、头、件、个、斤、批、打
指代词	这、那、此、彼、某

3 基于权重动态分配的纠错方法

3.1 纠错词集的构建思路

考虑到语料库中99.48%的字为常用中文字符,约有3 700多个,因此本文构建了基于常用高频字集的候选纠错词集。假设动态文本窗口的大小 N,考虑到计算纠错词集所需的时间复杂度和空间复杂度较高,因此设定动态文本窗口长度为3,采取促使局部文本出现概率最大化的方式获取纠错词。在纠错词集的权重分配方面,采用动态文本窗口内疑似错误和前后文本的词语组合长度作为纠错词的权重,建立基于最小编辑距离的纠错词集。图1为基于权重动态分配的纠错策略示意图。首先对文本中的疑似错误词语建立相应的纠错词集,然后计算每个纠错词与系统词典的最小编辑距离,将系统词典中与纠错词编辑距离最小的系统词的长度作为纠错建议的权重。

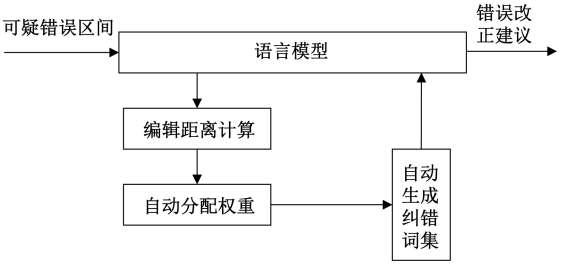

图1 基于权重动态分配的纠错策略示意图

Figure 1 Schematic diagram of error correction strategy based on dynamic weight allocation

3.2 基于纠错词权重动态分配的纠错算法

基于纠错词权重动态分配的纠错算法的具体步骤如下。

Step 1: 设动态窗口的文本为 $T=\{\cdots,c_{-2},c_{-1},c_0,c_1,c_2,\cdots\}$, 系统词典为 D。

Step 2: 令 $T_f=\{\cdots,c_{-1}\}$ 为文本中字符 c 的左侧相邻的字符串, $T_b=\{\cdots,c_n\}$ 为文本中字符 c 的右侧相邻的字符串。令 $T_f^i=\{c_{-i},\cdots,c_{-1}\}$ 为文本中字符 c 的左侧相邻长度为 i 的字符串,令 $T_b^j=\{c_1,\cdots,c_j\}$ 为文本中字符 c 的右侧相邻长度为 j 的字符串。

Step 3: 设 $W=\{T_f^i,c,T_b^j\}$, 若 $W\in D$,将满足 $ed(\{T_f^i,c,T_b^j\},W)=1$ 且 $length(\{T_f^i,c,T_b^j\})=length(W)$ 的词作为纠错词,权重为 $length(W)$。

Step 4: 若 $W\notin D$, 则纠错词由多个系统词构成,将纠错词长设为 $length(\{T_f^i,c,T_b^j\})$ 的词长,依照 $ed(\{T_f^i,c,T_b^j\},\{w_1,\cdots,w_{j-i+1}\})=1$ 来构建纠错词集,权重为 $length(\{T_f^i,c,T_b^j\})$。

4 实验结果与分析

4.1 中文文本错误的查错

从日常报纸杂志的100篇文章中收集了500个人工校对好的标准句子,其中包含220个文本错误。表2为多种文本检错策略的实验结果。当仅使用窗口大小为3的检错方案时,检测文本错误的召回率为68.2%,准确率为58.5%。当添加窗口缩小策略改进检错方案后,检错的召回率下降了8.2%,准确率提高了12.4%。主要原因是对文本采取更加细粒度的检错方案降低了原本较大窗口产生的文本数据稀疏的出现概率。当使用添加窗口缩小和窗口拓展策略的检错方案时,检错的准确率又上升了13.7%。主要原因是窗口下文的内容也被转移进来,相当于进一步考虑了上下文的语句搭配合理度。最后,综合前两种策略和聚类词平滑策略后,检错的准确率再次上升了3.3%,F 值达到77.9%。实验结果表明,基于动态窗口的文本查错算法在很大程度上提高了文本自动检错的效率和可行性。

表2 多种文本检错策略的实验结果

Table 2 Experiment results of multiple strategies for error detection

检错结果	窗口大小为3的检错策略	添加窗口缩小策略	添加窗口缩小和窗口拓展策略	添加窗口缩小、窗口拓展及聚类词策略
准确率	0.585	0.709	0.846	0.879
召回率	0.682	0.600	0.673	0.700
F 值	0.629	0.649	0.749	0.779

4.2　中文文本错误的纠错

在本文构建的数据集下比较了三种自动纠错方法,分别为本文的基于权重动态分配的纠错模型、传统的黑马校对系统以及文献[15]所提出的权重均分的纠错模型,通过人工鉴定这些方法所生成的纠错建议的合理度,并统计合理纠错建议的数量,纠错性能结果比较如表 3 所示。可以看出,传统的黑马校对系统首先检测出 178 个文本错误,被人工鉴定为合理的纠错建议有 169 个;文献[15]模型检测出 170 个文本错误,被人工鉴定为合理的纠错建议有 155 个;本文模型发现了 193 个文本错误,被人工鉴定为合理的纠错建议有 177 个,纠错准确率最高且耗用资源比最低,其中纠错准确率达到 78.1%,比黑马校对系统和文献[15]模型分别提升了 9.7%和 15.8%,在实际应用上表现良好。

表 3　纠错性能结果比较

Table 3　Comparison on error correction performance results

纠错方法	文本错误的个数	纠错建议合理的个数	纠错准确率	耗用资源比
黑马校对系统	178	169	0.684	0.403
文献[15]模型	170	155	0.623	0.366
本文模型	193	177	0.781	0.256

5　结束语

本文以动态文本窗口对人们阅读文本时的注意力集中范围进行建模,实现了注意力集中范围的放大与缩小,并以中文词语的构词长度自动分配纠错建议的权重,实现了文本自动纠错模型。虽然目前的文字编辑工作中使用文本自动纠错技术能有效地降低一些文本错误所产生的不良影响,但是在文本的自动查错与纠错方面仍然需要进行更加深入的研究。在检错时,一方面还可以考虑语法和语义知识,另一方面还需要考虑语义知识库的建立与使用方法。除了基于知识库的方法外,还可以建立自动获取错误文本特征的模型。纠错建议生成方法可以全面地从篇章、段落、整句上使用自然语言推理技术推理出原文内容,这样能更加体现出纠错方法的科学性。

参考文献:

[1]　柏晓鹏.汉语中介语文本词语级错误的自动查错研究及其实现:AECIT[D].南京:南京师范大学,2007.
BAI X P. Research and realization of automatic error detection of word level errors in Chinese intermediary texts: AECIT[D]. Nanjing: Nanjing Normal University, 2007.

[2]　王虹,张仰森.基于词性预测的中文文本自动查错研究[J].贵州师范大学学报(自然科学版),2001,19(2):72-75.
WANG H, ZHANG Y S. The research of Chinese text automatic error-checking method based on the port-of-speech (pos) of words[J]. Journal of Guizhou normal university (natural science), 2001, 19(2): 72-75.

[3]　张仰森,曹元大,俞士汶.基于规则与统计相结合的中文文本自动查错模型与算法[J].中文信息学报,2006,20(4):1-7.
ZHANG Y S, CAO Y D, YU S W. A hybrid model of combining rule-based and statistics-based approaches for automatic detecting errors in Chinese text[J]. Journal of Chinese information processing, 2006, 20(4): 1-7.

[4]　张仰森.统计语言建模与中文文本自动校对技术[M].北京:科学出版社,2017.
ZHANG Y S. Statistical language modeling and Chinese text automatic proofreading technology[M]. Beijing: Science Press, 2017.

[5]　葛诗利.自动作文评分中词汇接续错误自动识别研究[J].外语电化教学,2010(4):15-20.
GE S L. The automatic detection of language errors of binary adjacent word pairs in automated essay scoring[J]. Foreign language audiovisual teaching, 2010(4): 15-20.

[6]　张仰森,丁冰青.基于二元接续关系检查的字词级自动查错方法[J].中文信息学报,2001,15(3):36-43.
ZHANG Y S, DING B Q. Automatic errors detecting of Chinese texts based on the bi-neighborship[J]. Journal of Chinese information processing, 2001, 15(3): 36-43.

[7]　吴林,张仰森.基于知识库的多层级中文文本查错推理模型[J].计算机工程,2012,38(20):21-25.
WU L, ZHANG Y S. Reasoning model of multi-level Chinese text error-detecting based on knowledge bases[J]. Computer engi-

neering,2012,38(20):21-25.
[8] 刘亮亮,曹存根.基于局部上下文特征的组合的中文真词错误自动校对研究[J].计算机科学,2016,43(12):30-35.
LIU L L,CAO C G. Chinese real-word error automatic proofreading based on combining of local context features[J]. Computer science,2016,43(12):30-35.
[9] 郇政永.基于OCR的中文文本校对研究[D].北京:北方工业大学,2011.
HUAN Z Y. Research in Chinese text proofreading based on OCR[D]. Beijing: North China University of Technology,2011.
[10] 朱金金.中文文本自动查错与纠错模型的构建及实现[D].北京:北京信息科技大学,2010.
ZHU J J. Construction and implementation of Chinese text automatic error detection and correction model[D]. Beijing:Beijing Information Science and Technology University,2010.
[11] 贾玉祥,李育光,昝红英.基于MDL的汉语语义选择限制自动获取[J].郑州大学学报(理学版),2018,50(1):66-71.
JIA Y X,LI Y G,ZAN H Y. Acquiring Chinese selectional preferences using the MDL principle[J]. Journal of Zhengzhou university(natural science edition),2018,50(1):66-71.
[12] 骆卫华,罗振声,龚小谨.中文文本自动校对的语义级查错研究[J].计算机工程与应用,2003,39(12):115-118.
LUO W H, LUO Z S,GONG X J. Study of semantic errors checking in automatic proofreading for Chinese text[J]. Computer engineering and applications, 2003,39(12):115-118.
[13] 张仰森,郑佳.中文文本语义错误侦测方法研究[J].计算机学报,2017,40(4):911-924.
ZHANG Y S, ZHENG J. Study of semantic error detecting method for Chinese text[J]. Chinese journal of computers,2017, 40(4):911-924.
[14] 顾德之.中文真词错误自动校对方法研究[D].镇江:江苏科技大学,2017.
GU D Z. Research on Chinese real-word error automatic detection and correction[D]. Zhenjiang: Jiangsu University of Science and Technology, 2017.
[15] MAYS E, DAMERAU F J, MERCER R L. Context based spelling correction[J]. Information processing and management, 1991, 27(5): 517-522.

Chinese Text Error Correction Method Based on Dynamic Text Window and Weighted Dynamic Allocation

HUANG Gaijuan[1,2], WANG Congcong[1], ZHANG Yangsen[1,2]
(1. *Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing* 100101, *China*;
2. *Beijing Key Laboratory of Internet Culture Digital Dissemination Research, Beijing* 100101, *China*)

Abstract: A Chinese text error checking method based on dynamic text window was proposed, which relied on the continuous sliding window to detect errors in text. When the text was suspected to be wrong, the data sparse problem was smoothed by the clustering word set, and the error correction would be carried out by using the word set assigned dynamically. If the error correction results still could not conform to the error detection rules, the reduced window method and the extended window method would be used to check the specific errors. The error correction word set was constructed by a method which based on the minimum edit distance and the weighted dynamic allocation. The experimental results showed that the *F*-score of the dynamic text window error checking method was 77.9%. Combined with the error correction method of the weighted dynamic allocation, the error correction accuracy was 78.1%, which was 9.7% and 15.8% higher than black horse proofreading system and average weighted error correction strategy, respectively.

Key words: semantic collocation; data sparse; dynamic text window; weighted dynamic allocation

(责任编辑:孔 薇)