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

刘亮亮[1] 曹存根[2]
(江苏科技大学计算机科学与工程学院 镇江 212003)[1]
(中国科学院计算技术研究所智能信息重点实验室 北京 100190)[2]

摘 要 中文的真词错误类似于英文的真词错误,指一个中文词错成另一个词典中的词。提出一种基于混淆集的真词错误发现方法,通过对目标词的局部特征的提取,形成局部左邻接二元、右邻接二元及3个三元特征,然后通过和目标词对应的混淆集中的混淆词来估计二元概率和三元概率。最后提出一种多特征融合的模型,然后利用规则来判断中文文本中的真词错误。将查错结果分为标记错误和更改错误两种类型,采用18组混淆集,构造2万行的测试语料进行实验。实验表明,该方法能有效地发现中文文本中的真词错误,并且能给出真词错误的修改建议。该方法是一种集自动查错和自动纠错于一体的中文文本自动校对方法。

关键词 真词错误,混淆集,上下文特征,NGram模型

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.005

Chinese Real-word Error Automatic Proofreading Based on Combining of Local Context Features

LIU Liang-liang[1] CAO Cun-gen[2]
(School of Computer Science & Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)[1]
(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract Similar to the English context-sensitive spelling correction, real-word error in Chinese refers to the error that a Chinese word is misused to another Chinese Word. In the paper, a Chinese real word error detection and correction method based on confusion sets was proposed. This method extracts local feature around the aim word which forms left adjacent bigram, right adjacent bigram and three trigrams. The probability of bigram and trigram are computed with the confusion words in the aim word's confusion set. A model based on multi-feature fusion was proposed and rules was used to find the real-word errors. We classified the result into two types, marking the errors and rewriting the errors. In the experiment, we used 18 group confusion sets and built 20000 sentences corpus to validate the algorithm. The results show that the proposed method can find the real-word errors in Chinese texts and give the correction lists. The proposed method combines automatic error-detecting and automatic error-correction.

Keywords Real-word error, Confusion set, Context feature, NGram model

1 引言

英文拼写错误主要分为两种,一种是"非词错误",另外一种是"真词错误"[1]。英文的非词错误是指一个英文单词写错成一个不是词典中的词的英文串,例如"the"错写成"teh"。而英文的真词错误又称为上下文相关错误,是指一个单词写成词典中的另外一个单词,例如"a peace of cake"中的"peace"就是真词错误。汉字无法像英文那样通过键盘输入到计算机,需要通过形输入法或音输入法输入到计算机。现在越来越多的人采用拼音输入法来输入汉语,首先输入汉语的拼音,通过拼音输入法转换成汉字,然后选择对应的词,因此中文不会出现英文的"非词错误",通过键盘输入的汉字不会出现"非字/词错误",中文的"非字错误"一般只会出现在手写文本或OCR识别中。而中文存在真词错误,即句子中的词用错成词典中的另外一个词,例如"他接收总经理的邀请参加会议"中的"接收"是一个真词错误。由于人们的粗心选择以及对汉语词语之间的区别的认知不足,汉语文本中出现了很多的真词错误。例如,在汉语文本中,"直播"与"直拨"、"一下"与"以下"、"接收"与"接受"等词经常相互混淆。

本文提出一种方法来识别汉语文本中的真词错误,首先根据汉语拼音的音相似和形相似生成真词混淆集(Confusion Set),例如集合{直播,直拨}就是一个真词混淆集,这两个词具有相同的汉语拼音"zhibo"。本文把真词错误识别问题转换成从混淆集中选择适合于当前上下文环境的正确的词。

到稿日期:2015-08-27 返修日期:2015-12-07 本文受国家自然科学基金项目(91224006,61173063,61035004,61203284,30973713),国家社科基金重点项目(10AYY003)资助。

刘亮亮(1979-),男,博士,讲师,主要研究领域为自然语言理解、知识工程与知识获取,E-mail:lingyun79626@126.com;曹存根(1964-),研究员,主要研究领域为知识工程。

2 相关研究

英文的非词错误自动发现方法一般采用查字典法，而现在英文拼写错误的研究主要面向于真词错误。随着自然语言理解技术的发展，一些统计方法被应用到真词拼写错误检查中，包括词的 n-gram 模型[2,3]、词性标注[4-6]、贝叶斯分类[7]、决策列表[8]、贝叶斯混合方法[9]、词性和贝叶斯组合方法[6]、潜在语义分析[10]。目前，英文真词错误研究主要分为以下几类：基于机器学习的方法、基于语义信息的方法，以及基于概率统计的方法。

2.1 基于机器学习的方法

基于机器学习的方法将真词拼写纠错看成是词的歧义消解问题，这种方法必须依靠预先定义的混淆集，混淆集是由容易混淆的单词组成，例如“dessert”与“desert”。通过学习来获得混淆集中每个词的上下文特征，然后判断在特定的上下文中混淆集中的哪个词更合适。这种方法的最大的缺点是依赖于混淆集，对于不在预先定义的混淆集中的词没有办法检查其正确性。

Golding 等于 1995 年利用决策列表从混淆集中选择合适的词，实验表明组合证据会获得更好的结果，即其不仅仅是考虑一些强有力的证据，而是利用所有可用的证据的组合[9]。他利用贝叶斯分类方法运行相同的实验，但是结果比决策列表提高较少。Golding 于 1996 年提出了对贝叶斯分类方法的扩展，首先对文本进行词性标注，如果混淆集中的词的词性都不同，在一个上下文环境中词性标记器会选择最有可能的标记，从而得到最有可能的词。如果混淆集中的词不能通过词性来进行区分，此时则利用原先的贝叶斯分类方法来进行判断。结果表明这种混合的方法比原先单独使用贝叶斯分类方法的效果要好得多。

Golding 等于 1999 年提出了基于 Winnow 的真词拼写纠错方法 WinSpell，Winnow 是一种乘权更新的算法，在处理大规模特征时会取得很高的准确率。这种方法可以根据不同的权值学到大量的特征集合。Winnow 方法的权值训练方法比贝叶斯分类等方法表现更有效[11]。Golding 于 1999 年采用 Brown 语料和 21 组混淆集分别对他在 1995 年提出的 BaySpell 方法、Mangu 于 1997 年提出的 RuleS 方法、Jones 于 1997 年提出的 LSA 方法进行对比实验，实验结果表明，LSA 方法的精度是 82.8%，RuleS 方法的为 88.5%，BaySpell 方法的是 93.8%，WinSpell 达到 96.4%。从实验结果可知，WinSpell 方法取得了最好的效果。另一种基于机器学习进行真词拼写校对的方法是 Carlson 等提出的 Snow 方法[13]，该方法采用 SNow 方法进行学习[12,13]，在 265 个混淆集中的实验结果表明其平均精度达到 99%[13]。

2.2 基于语义信息的方法

基于语义信息的方法不需要预先定义的混淆集。这种方法的本质是基于正确的词与其周围的词之间满足某种语义的联系，而真词拼写错误的词不满足这种语义联系。Hirst 等首先提出基于语义信息方法来发现真词错误[14]；Hirst 进一步将该方法发展，通过使用 WordNet 来计算词与词之间的语义距离，如果词与其上下文语义距离较远，则认为其是错误的，与上下文语义距离近的词可能就是其对应正确的词[15]。采用人工构造的错误的语料进行实验，获得了 23%～50%的召回率，精度为 15%～25%。

2.3 基于概率统计的方法

最初利用概率统计方法来识别真词错误的是采用词和词性的 n-gram 统计语言模型的方法[16-18]。n-gram 统计方法基于大规模语料统计词的 n-gram 序列，通过 n-gram 概率来发现真词错误，一般采用 2-gram 或 3-gram。n-gram 方法一般认为低概率序列是错误的，高概率序列是纠错建议的候选列表。例如“The cat mowed loudly”中 P(mower | Thecat)概率非常低，意味着这里可能会出现错误，在与词“mowed”编辑距离很小的词中，P(meowed | The cat)概率比较大，这意味着“meowed”是这个句子中最可能的词。

Mays 等采用 3-gram 的方法进行查错，通过统计来查错和纠错，实验结果表明该方法可以发现 76%的错误，同时可以纠正 73%的错误[2]。

Islam 等采用 Google Web 1T 3-gram 数据集和标准的及改进的最长公共子序列的字符串匹配算法来进行真词错误的查错和校对[19]。该工作主要目的在于提高查错召回率和纠错召回率，并且保持适当的精度。实验数据采用 Wilcox-O'Hearn[15]中的测试数据，结果表明该方法取得了 89%的查错召回率和 76.3%的纠错召回率，比相同数据在 Wilcox-O'Hearn 及 Hirst[15]的测试结果要好。

这些方法最大的问题是需要大规模的语料来训练n-gram模型，并且在运行时需要非常巨大的 n-gram 表，其另一个缺点是无法获得长距离的词之间的依赖关系。基于概率统计的方法的优点之一是不需要依赖于预先定义好的混淆集。词的 n-gram 模型的主要问题是数据稀疏，即使有相当大的训练数据量。研究表明 Berlinsky[3]用词的二元模型比三元模型的效果要好，很可能也是因为数据稀疏问题。

目前，中文文本自动校对没有像英文文本自动校对那样严格地划分为“非词错误”和“真词错误”，中文文本自动校对目前仍然处于字词级的查错阶段，并且主要以自动查错为主。啄木鸟系统[20]是国内出现得比较早的中文自动校对系统，该系统的自动查错方法的出发点是文本中的绝大多数错误都会导致切分后的单字词。张照煌等根据 4 种汉字的相似类型：1)同音或近音；2)字形相近；3)字义相近；4)输入编码相近，生成各个汉字的相似子集，对每个句子中的每个汉字用其相似子集中的汉字依次替换，然后用词间字二元模型和词性的二元模型对各字串进行评分，最后选出得分最高的字串[21]。张磊等[22]提出基于特征和 Winnow 学习模型的中文自动校对方法，首先定义字或词的混淆集，然后提取目标串的二元接续关系、词性类的三元接续关系、上下文语义类、词性类邻接字 4 种特征集，根据 Winnow 方法进行特征学习，然后利用这些上下文特征对目标词混淆集中的词进行选择。马金山等[23]构建了一种多方法融合的中文自动校对模型，该模型以三元模型为基础，对文本进行局部分析，以查找文本中的局部错误；同时利用依存文法的特点对句子进行依存分析，在查找全局错误的研究中提出了新的思路和方法。张仰森于 2006 年提出一种基于规则和统计相结合的方法[24]，根据正确文本分词后单字词的出现规律，提出一组错误发现规则，并与针对分词后单字散串建立的字二元、三元统计模型和词性二元、三元统计模型相结合，建立了文本自动查错模型与实现算法。吴林等[25]针对中文文本中的字词级错误、语法级错误和语义级错误 3 个层次的错误，提出基于知识库的多层级中文文本查错推理模型，构建了一个综合查错系统。该方法取得了

85.62%的召回率。刘亮亮等[26]提出了一种基于散串合并与统计验证的方法来发现中文文本的错别字，该方法对领域问答系统日志进行分词，对分词中的多字词和合并的串进行相似词串聚类，对相似词串的上下文语境进行统计分析，从中自动获取错别字对。该系统获得了71.32%的召回率和82.6%的准确率。

3 本文的方法

本文的方法首先利用汉字拼音和形状建立汉字词的混淆集，然后利用上下文特征计算混淆集中每个混淆词的得分排名，再利用得分排名来发现句子中的真词错误，并且对其进行自动校对，给出修改建议。本文提出的算法是一种集自动查错与自动纠错于一体的方法。

3.1 中文词混淆集的构建

一个中文词 W_i 的混淆集 $CSet(W_i)$ 是指中文词典中的一组词与 W_i 音相似或形相似或意相似；而在人们的使用过程中，W_i 与 $CSet(W_i)$ 的词常常容易混淆，从而在使用过程中出现错误，可以表示为：

$$CSet(W_i)=\{W_i^1,W_i^2,\cdots,W_i^k\}$$

例如："直播"与"直拨"的汉语拼音相同，人们在使用"直播"的时候，常常错用成"直拨"，反过来亦然。因此，$CSet$(直播)=｛直拨｝，$CSet$(直拨)=｛直播｝。

中文词是由汉字构成的，两个词是混淆词是由于其对应位置的汉字混淆，因此本文通过施恒利等构造的汉字混淆集[27]和汉语词典构造中文词的混淆集。文中汉字混淆集是指一个汉字由于和其它汉字音相似或形相似导致混淆，例如：$CSet$(拨)=｛拔，播｝，中文词混淆集的构成是对中文词中的每个汉字利用其汉字混淆字进行替换而成，替换后如果仍然是词典中的词，则加入到该词的混淆集中。例如词"直拨"，用$CSet$(拨)中的"拔"、"播"替换得到"直拔"、"直播"，由于"直拔"不是词典中的词，因此"直播"是"直拨"的混淆词，将其加入到"直拨"的混淆集中，即 $CSet$(直拨)=｛直播｝。

3.2 上下文特征

通过混淆集来判断句子中的真词错误，相当于从混淆集中选择一个最适合当前上下文的词，这是一个排歧的过程，评价特定上下文对特定目标的支持度，选择支持度最高的词。如果支持度最高的词与原句子中的词是相同的，那么该句子中的词是正确的；如果不一致，则该句子中的目标词可能是一个真词错误。

从混淆集中选择目标词的线索之一是目标词的上下文特征，其中二元和三元表示词之间的局部的接续关系[23]，因此可以通过NGram模型来检查局部的真词错误。上下文特征是从目标词的上下文中提取的，通过上下文特征来判断混淆集中的哪个词适合当前的上下文特征。例如，对于"直播"与"直拨"，"直拨"的周围一般出现"电话"、"手机"等，而"直播"的周围出现"比赛"、"赛事"等比赛类词。

假设句子分词后，$S=W_1W_2\cdots W_n$，对于目标词 W_i，本文提取以下特征。

词的左二元接续关系：这个特征记录目标词的左边邻接词 W_{i-1}，形成特征模板为"W_{i-1} *"，* 表示目标词。该特征表示目标词与左边邻接的词的接续关系，如果目标词是句子的首词，本文用"＃Begin＃"表示句首标识。

词的右二元接续关系：这个特征记录目标词的右边邻接词 W_{i+1}，形成特征模板为"* W_{i+1}"，其中 * 表示目标词，该特征表示目标词与右边邻接的词的接续关系。同样地，如果目标词是句子的尾词，那么用"＃End＃"表示句尾标识。

词的三元接续关系：这个特征记录目标词左右邻接关系，抽取其左右邻接词，形成三元接续关系特征模板 W_{i-1} * W_{i+1}，该特征表示词的三元接续关系。

3.3 N-Gram概率估计

对于句子 $S=W_1W_2\cdots W_n$，引入以下假设：一个词 W_i 出现的概率只与其上文词或下文词有关，独立于其它词。本文采用极大似然估计来计算NGram概率。

左邻接二元的概率估计为：

$$P_{left}=(W_i|W_{i-1})=\frac{Count(W_{i-1}W_i)}{Count(W_{i-1}W_i)+\sum_{k=1}^{|CSet(W_i)|}Count(W_{i-1}W_i^k)} \quad (1)$$

右邻接二元的概率估计为：

$$P_{right}=(W_i|W_{i+1})=\frac{Count(W_iW_{i+1})}{Count(W_iW_{i+1})+\sum_{k=1}^{|CSet(W_i)|}Count(W_i^kW_{i+1})} \quad (2)$$

邻接三元的概率估计为：

$$P_{left\_tri}=(W_i|W_{i-2}W_{i-1})=\frac{Count(W_{i-2}W_{i-1}W_i)}{Count(W_{i-2}W_{i-1}W_i)+\sum_{k=1}^{|CSet(W_i)|}Count(W_{i-2}W_{i-1}W_i^k)} \quad (3)$$

$$P_{tri}=(W_i|W_{i-1}W_{i+1})=\frac{Count(W_{i-1}W_iW_{i+1})}{Count(W_{i-1}W_iW_{i+1})+\sum_{k=1}^{|CSet(W_i)|}Count(W_{i-1}W_i^kW_{i+1})} \quad (4)$$

$$P_{right\_tri}=(W_i|W_{i+1}W_{i+2})=\frac{Count(W_iW_{i+1}W_{i+2})}{Count(W_iW_{i+1}W_{i+2})+\sum_{k=1}^{|CSet(W_i)|}Count(W_i^kW_{i+1}W_{i+2})} \quad (5)$$

其中，$W_i^k\in CSet(W_i)$，$Count(W_{i-1}W_i)$ 表示 $W_{i-1}W_i$ 在训练语料中的共现频次，$Count(W_iW_{i+1})$ 表示 W_iW_{i+1} 在训练语料中的共现频次，$Count(W_{i-1}W_iW_{i+1})$ 表示 $W_{i-1}W_iW_{i+1}$ 在训练语料中的共现频次，$Count(W_{i-1}W_i^k)$ 表示 $W_{i-1}W_i^k$ 在语料中的共现频次，$Count(W_i^kW_{i+1})$ 表示 $W_i^kW_{i+1}$ 在语料中的共现频次，$Count(W_{i-1}W_i^kW_{i+1})$ 表示 $W_{i-1}W_i^kW_{i+1}$ 在训练语料中的共现频次。

通过大规模语料进行统计，根据式(1)－式(5)的定义，可以得到：

$$\sum_{k=0}^{|CSet(W_i)|}P_{left}(W_i^k|W_{i-1})=1$$

$$\sum_{k=0}^{|CSet(W_i)|}P_{right}(W_i^k|W_{i+1})=1$$

$$\sum_{k=0}^{|CSet(W_i)|}P_{left\_tri}(W_i^k|W_{i-2}W_{i-1})=1$$

$$\sum_{k=0}^{|CSet(W_i)|}P_{tri}(W_i^k|W_{i-1}W_{i+1})=1$$

$$\sum_{k=0}^{|CSet(W_i)|}P_{right\_tri}(W_i^k|W_{i+1}W_{i+2})=1$$

3.4 纠错模型与算法

根据NGram模型的定义，将 $P_{left}(W_i|W_{i-1})$，$P_{right}(W_i|W_{i+1})$，$P_{left\_tri}(W_i|W_{i-2}W_{i-1})$，$P_{tri}(W_i|W_{i-1}W_{i+1})$，$P_{right\_tri}$

$(W_i|W_{i+1}W_{i+2})$特征融合在下面的 Score 函数中，来判断混淆集中哪个词最适合。

$$Score(W_i)=P_{left}(W_i|W_{i-1})+P_{right}(W_i|W_{i+1})+P_{tri}(W_i|W_{i-1}\ W_{i+1})+P_{left\_tri}(W_i|W_{i-2}\ W_{i-1})+P_{right\_tri}(W_i|W_{i+1}W_{i+2}) \quad (6)$$

同时利用二元和三元特征来进行判断比单独利用二元或三元特征进行判断的效果要好。阶数越高的 NGram 需要更多的上下文特征，但是数据稀疏；而低阶的 NGram 只能表现有限的上下文，但是其数据稀疏没有高阶 NGram 那么严重。因此将权重引入到式(6)中，高阶 NGram 的权重比低阶 NGram 的权重大。

$$Score(W_i)=\alpha_1 * P_{left}(W_i|W_{i-1})+\alpha_2 * P_{right}(W_i|W_{i+1})+\alpha_3 * P_{tri}(W_i|W_{i-1}W_{i+1})+\alpha_4 * P_{left\_tri}(W_i|W_{i-2}W_{i-1})+\alpha_5 * P_{right\_tri}(W_i|W_{i+1}W_{i+2}) \quad (7)$$

其中，$\alpha_1+\alpha_2+\alpha_3+\alpha_4+\alpha_5=1$。

根据上述模型(7)设计算法来自动发现文本中的真词错误，并且进行自动校对。首先对混淆集中的每一个混淆词计算 $Score(W_i)$，然后对 $Score(W_i)$进行排序。本文将查错状态分为 3 种：1)$Ok(W)$状态——认为词 W 是正确的；2)$Mark(W)$状态——对该词 W 进行标记错误，但是不知道其对应正确的词；3)$Rewrite(W_i,\langle W_j\rangle)$状态，表示 W_i 是错误的词，但是 $\langle W_j\rangle$是其对应的正确的词的集合，表示其正确的修改建议。

判断规则：

(1)如果 $Score(W_i)=0$，同时如果在 W_i 的混淆集中存在混淆词 W_i^k 满足 $Score(W_i^k)>0$，则将 W_i^k 加入到其修改建议列表中，置查错状态为 Rewrite；如果不存在混淆词 W_i^k，使得 $Score(W_i^k)>0$，则对 W_i 进行错误标记，置查错状态为 Mark；

(2)如果 $Score(W_i)>0$，同时如果在 W_i 的混淆集中存在混淆词 W_i^k 满足 $Score(W_i)<\beta * Score(W_i^k)$，则对 W_i 标记错误，同时将满足条件的 W_i^k 加入到修改建议列表中，置查错状态为 Rewrite。β表示 W_i 错误的概率，一般而言，一个词用错成另外一个词的概率较小，一般不超过 0.01，因此在实验中 $\beta=0.01$；否则，认为 W_i 为正确的词，对 W_i 标记 Ok 状态。

具体的算法如下。

算法 1 Real Word Spelling Correction

输入：$S=W_1W_2\cdots W_n$

输出：对错误的真词进行标记

1. Begin
2. 依次扫描 W_i
3. 计算 $Score(W_i)$与 $Score(W_i^k)$(W_i^k 是混淆集中的每一个混淆词)
4. if $Score(W_i)=0$
5. 则标记 W_i 为一个真词错误，同时 $Score(W_i^k)>0$ 的 W_i^k 为其对应正确的词
6. Else if $Score(W_i)>0$
7. if ($Score(W_i^k)<Score(W_i)$)
8. W_i 为正确的；
9. Else if $Socre(W_i>0)$&$Score(W_i)<0.01 * Score(W_i^k)$
10. W_i 为错误的，同时满足 $Score(W_i)<0.01 * Score(W_i^k)$，$W_i^k$ 即为其对应正确的词
11. End

对于混淆词“树立”与“竖立”：

例 1 …如何竖立正确的政绩观政绩观…

通过计算 $Score$(竖立)$=1.7326017903551835E-4$，$Score$(树立)$=0.999653479641929$。根据判断，句中的“竖立”是错误的，对其标记 Rewrite，其对应正确的是“树立”。

例 2 …无限的憧憬和抱负。但作为新领导该如何竖立威信，让下属信服，…

根据计算，$Score$(竖立)$=0$，$Score$(树立)$=1$，对其进行标记 Rewrite，其对应正确的是“树立”。

例 3 三个月的宝宝树立抱着好吗？

根据计算，$Score$(树立)$=0$，$Score$(竖立)$=0$，则对句子中的“树立”标记 Mark。

例 4 作为当代大学生，应当树立怎样的世界观、人生观、价值观…

根据计算，$Score$(树立)$=1$，$Score$(竖立)$=0$，则对“树立”标记 Ok，表示句子中的“树立”是正确的。

4 实验与分析

混淆集：本文通过汉字的音相似和形相似选择了 18 组中文词混淆集来进行真词错误实验。18 组混淆集如表 1 所列。

表 1 混淆集

混淆集
直拨 直播
接收 接手 接受
监查 监察 检查 检察
鼎立 鼎力
无限 无线
标明 表明
复式 复试
资费 自费
工夫 功夫
起用 启用
亲身 亲生
学历 学力
震动 振动
原型 原形
增殖 增值
树立 竖立
反映 反应
相应 响应

测试集：为了评估本文方法的有效性，利用 18 组混淆集，从百度知道和某领域问答系统中抽取 2 万个包含混淆集中的词的句子构成测试句子，其中每组 1000 个句子，人工对 2 万行句子进行标注。标记格式如表 2 所列。

表 2 测试集标记示例

标记示例
原句：DDD-/n 国内/s 长途/b 直播/v 电话/n 业务/n ，/w 用户/n 利用/v…
标记：DDD-/n 国内/s 长途/b ＜直播/v|直拨＞电话/n 业务/n ，/w 用户/n 利用/v…
原句：卫星/n 直播/v 业务/n 使/v 星/n 上/v 电视/n 节目/n 直接/a 到/v 户/n
标记：卫星/n 〈直播/v|OK〉业务/n 使/v 星/n 上/v 电视/n 节目/n 直接/a 到/v 户/n
原句：怎样/r 写/v 一/m 篇/q 转变/v 工作/v 态度/n 竖立/v 正确/a 心态/n
标记：怎样/r 写/v 一/m 篇/q 转变/v 工作/v 态度/n 〈竖立/v|树立〉正确/a 心态/n
原句：为/p 自己/r 竖立/v 一个/m 明确/a 的/u 奋斗/v 目标/n ，/w 是/v 每/r 一个/m 希望/v 人生/n…
标记：为/p 自己/r 〈竖立/v|树立〉一个/m 明确/a 的/u 奋斗/v 目标/n ，/w 是/v 每/r 一个/m 希望/v 人生/n…
……

注：标记中“〈〉”表示该词是混淆集中的词，“\|”后面表示人工标记，用“Ok”表示。

评价指标：本文的方法分为两个过程，一个是自动发现测试集中的真词错误。自动查错采用两个指标来评价，即查错召回率 Recall 与查错准确率 Precision。定义如下。

$$\text{Recall}=\frac{\text{正确发现错误的总数}}{\text{文本中的错误总数}}\times 100\%$$

$$\text{Precision}=\frac{\text{正确发现错误的总数}}{\text{发现错误的总数}}\times 100\%$$

另一个是对发现的错误进行自动校对，给出修改建议，利用纠正率，表示正确纠错的比例。定义如下：

$$\text{Correct\_rate}=\frac{\text{正确纠错的总数}}{\text{文本中的错误总数}}\times 100\%$$

根据以上两个阶段的评价指标的计算方法，得到实验结果的评价指标值如表 3 所列。

表 3　实验结果指标

查错召回率	查错准确率	纠正率
74.9%	75.8%	70%

部分实验结果如表 4 所列。

表 4　部分实验结果

标记示例
示例 1：结构化系统开发方法，原形法和面向对象开发方法的优缺点和适用场合/v…
查错结果：错误词：原形(0.0)[1] 标记类型：Rewrite 修改建议：原型(0.55)
示例 2：excel 中的检察拼写在哪里噢？
查错结果：错误词：检察(0.0036218553097548635)标记类型：Rewrite 修改建议：检查(0.45677229453157336)
示例 3：电脑开机时显示请检察信号线路怎么办
查错结果：错误词：检察(0.0) 标记类型：Rewrite 修改建议：监察(0.0012300123001230013)|检查(0.44876998769987697)
示例 4：作为当代大学生，应当树立怎样的世界观、人生观、价值观…
查错结果：正确词：树立(1.0)标记类型：Ok
示例 5：乘风破浪前程广，鼎立创新步步高是什么意思
查错结果：错误词：鼎立(0.0) 起始位置：4 终止位置：4 查错类型：Mark
示例 6：我们应该接收大家在会议上提出的意见，改正工作中的缺点和错误
查错结果：正确词：接收(0.65)标记类型：Ok
……

对实验结果进行分析可知，局部特征的数据稀疏是本方法漏报和错报的主要原因，如表 4 中的示例 5，由于“鼎力”与局部上下文特征“创新”和“步步高”共现都为 0，从而对其产生了误报。混淆词都不满足局部特征，无法区分目标词与混淆词，这是漏报的一个主要原因。表 4 中示例 6 的“接收”满足局部上下文特征，从而产生了漏报，而发现这个真词错误需要用到远距离的搭配特征“意见”，“接受…意见”是一个固定搭配，将搭配特征引入到真词错误检查中也是下一步的研究方向。

结束语　本文介绍了一种基于混淆集的方法来发现汉语真词错误，利用目标词的局部上下文特征来判断目标词是否是一个真词错误，采用的局部上下文特征包括：左邻接二元、右邻接二元以及 3 个邻接三元，通过将多个上下文特征融合起来进行真词错误发现，并且对真词错误进行校对，生成修改建议。实验表明，本文的方法能有效地发现汉语真词错误。同时，本文提出的方法比单独使用二元或三元方法的查错效果要好，引入邻接二元与三元组合可以有效地解决三元的数据稀疏带来的问题。下一步的工作将引入远距离的搭配特征，来解决由于数据稀疏导致的误判以及由于发生远距离搭配错误导致的真词错误。

[1] 原形(0.0)：括号中的数值为词“原形”的 Score 得分。

参 考 文 献

[1] Kuckich K. Techniques for automatically correcting words in text[J]. ACM Computing Surveys (CSUR), 1992, 24(4): 377-439

[2] Mays E, Damerau F J, Mercer R L. Context based spelling correction[J]. Information Processing & Management, 1991, 27(5): 517-522

[3] Berlinsky-Schine A. Context-based detection of real word typographical errors using markov models[R]. Cornell University, Ithaca, NY, 2004

[4] Marshall I. Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus[J]. Computers and the Humanities, 1983, 17(3): 139-150

[5] Garside R, Sampson G, Leech G. The computational analysis of English: A corpus-based approach[J]. Lingua, 1991, 85(4): 365-367

[6] Golding A R, Schabes Y. Combining trigram-based and feature-based methods for context-sensitive spelling correction[C]// Proceedings of the 34th annual meeting on Association for Computational Linguistics. 1996: 71-78

[7] Gale W A, Church K W, Yarowsky D. A method for disambiguating word senses in a large corpus[J]. Computers and the Humanities, 1992, 26(5/6): 415-439

[8] Yarowsky D. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French[C]// Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. 1994: 88-95

[9] Golding A R. A Bayesian hybrid method for context-sensitive spelling correction[C]// Proceedings of the Third Workshop on Very Large Corpora. 1995, 3: 39-53

[10] Jones M P, Martin J H. Contextual spelling correction using latent semantic analysis[C]// Proceedings of the Fifth Conference on Applied Natural Language Processing. 1997: 166-173

[11] Golding A R, Roth D. A winnow-based approach to context-sensitive spelling correction[J]. Machine Learning, 1999, 34(1-3): 107-130

[12] Roth D, Zelenko D. Part of speech tagging using a network of linear separators[C]// Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2. 1998: 1136-1142

[13] Carlson A, Cumby C, Rosen J, et al. The SNoW learning architecture[R]. Technical Report UIUCDCS, 1999

[14] Hirst G, St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms[M]// WordNet: An Electronic Lexical Database, 1997: 305-332

[15] Hirst G, Budanitsky A. Correcting real-word spelling errors by restoring lexical cohesion[J]. Natural Language Engineering, 2005, 11(1): 87-111

[16] Atwell E, Elliott S. Dealing with ill-formed English text[M]// The Computational Analysis of English: A Corpus-Based Approach, 1987: 120-138

[17] Gale W A, Church K W. Estimation procedures for language context: poor estimates are worse than none[M]// Compstat. 1990: 69-74

[18] Church K W, Gale WA. Probability scoring for spelling correc-

tion[J]. Statistics and Computing,1991,1(2):93-103

[19] Islam A, Inkpen D. Real-word spelling correction using Google Web IT 3-grams[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Volume 3. 2009:1241-1249

[20] Shi De-sheng, Wang Liang-zhi, Chen Zhi-da, et al. A Statistics-based Approache for Automatic Detecting Errors in Chinese Text[J]. Computer and Communications,1992,8:19-26(in Chinese)
施得胜,王良志,陈志达,等. 基于统计的中文错字侦测法[J]. 电脑与通讯,1992,8:19-26

[21] Zhang Zhao-huang. Automatic Error Detection and Correction of ChineseText[J]. Communications of COLIPS, 1994, 4(2): 143-149(in Chinese)
张照煌. 中文错别字自动订正方法初探[J]. Communications of COLIPS,1994,4(2):143-149

[22] Zhang L, Zhou M, Huang C, et al. Multifeature-based approach to automatic error detection and correction of Chinese text[C]// Proceedings of the First Workshop on Natural Language Processing and Neural Networks. 1999

[23] Ma Jin-shan, Zhang Yu, Liu Ting, et al. Detecting Chinese Text Errors Based on Trigram and Dependency Parsing[J]. Journal of the China Society for Scintific and Technical Information, 2005, 23(6):723-728(in Chinese)
马金山,张宇,刘挺,等. 利用三元模型及依存分析查找中文文本错误[J]. 情报学报,2005,23(6):723-728

[24] Zhang Yang-sen, Cao Yuan-da, Yu Shi-wen. 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(in Chinese)
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

[25] Wu Lin, Zhang Yang-sen. Reasoning Model of Multi-level Chinese Text Error-detecting Based on Knowledge Bases[J]. Computer Engineering, 2012, 38(20): 21-25(in Chinese)
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]. Computer Engineering,2012,38(20):21-25

[26] Liu Liang-liang, Wang Shi, Wang Dong-sheng, et al. Automatic Text Error Detection in Domain Question Answering[J]. Journal of Chinese Information Processing, 2013, 27(3): 77-83(in Chinese)
刘亮亮,王石,王东升,等. 领域问答系统中的文本错误自动发现方法[J]. 中文信息学报,2013,27(3):77-83

[27] Shi Heng-li, Liu Liang-liang, Wang Shi, et al. Research on Method of Constructing Chinese Character Confusion Set[J]. Computer Science, 2014, 41(8): 229-232, 253(in Chinese)
施恒利,刘亮亮,王石,等. 汉字种子混淆集的构建方法研究[J]. 计算机科学,2014,41(8):229-232,253

(上接第 7 页)

[75] Kalal Z, Mikolajczyk K, Matas J. Tracking Learning Detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1409-1422

[76] Zhou Xin, Qian Qiu-meng, Ye Yong-qiang, et al. Improved TLD visual target tracking algorithm [J]. Journal of Image and Graphics, 2013, 18(9): 1115-1123(in Chinese)
周鑫,钱秋朦,叶永强,等. 改进后的 TLD 视频目标跟踪方法[J]. 中国图象图形学报,2013,18(9):1115-1123

[77] Dong Yong-kun, Wang Chun-xiang, Xue Lin-ji, et al. Pedestrian Detection and Tracking Based on TLD Framework[J]. J. Huazhong Univ. of Sci. & Tech. (Natural Science Edition), 2013, 41(S): 226-228(in Chinese)
董永坤,王春香,薛林继,等. 基于 TLD 框架的行人检测和跟踪[J]. 华中科技大学学报(自然科学版),2013,41(S):226-228

[78] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297

[79] Avidan S. Support Vector Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(8): 1064-1072

[80] Song Hua-jun. Study on Target Tracking Method Based on SVM[D]. Changchun: Changchun Institute of Optics Fine Mechanics and Physics, Aacademia Sinica, 2006: 30-42(in Chinese)
宋华军. 基于支持向量机的目标跟踪技术研究[D]. 长春:中国科学院研究生院(长春光学精密机械与物理研究所),2006:30-42

[81] Tian Q, Hong P, Huang T S. Update relevant image weights for content-based image retrievalusing support vector machines[C]// IEEE International Conference on Multimedia and Expo. 2000: 1199-1202

[82] Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines [C] // Advances in Kernel Methods—Suport Vector Learning. 1998: 212-223

[83] Keerthi S S, Shevade S K, Bhattacharyya C, et al. Improvements to Platt's SMO algorithm for SVM classifier design[J]. Neural Computation, 2001, 13(3): 637-649

[84] 机器学习 10 大经典算法[OL]. http://www.360doc.com/content/11/1102/14/4404107_161074278.shtml Ten Classical algorithms in Machine Learning. http:// www.360doc.com/content/11/1102/14/4404107_161074278.shtml

[85] Zhang H, Sheng S. Learning weighted naive Bayes with accurate ranking [C] // Proceedings of the Fourth IEEE International Conference on Data Miningg. IEEE Computer Society, 2004: 567-570

[86] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2/3): 131-163

[87] Laskey K B, Pradeeds H. Comparing Bayesian network classifiers[C]// Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1999: 101-108

[88] Mai Hua-an. Research on Object Detection and Tracking Algorithm Based on Bayesian Framework [D]. Guangzhou: South China University of Technology, 2013: 11-21(in Chinese)
麦华岸. 基于贝叶斯框架的目标检测跟踪算法研究[D]. 广州:华南理工大学,2013:11-21

[89] Xu Jing, Wang Xiao-feng. Recognition method of moving target using Bayesian probability theory[J]. Journal of Nanjing University of Science and Technology, 2013, 37(1): 76-80(in Chinese)
许敬,王晓锋. 基于贝叶斯概率的运动目标识别方法[J]. 南京理工大学学报,2013,37(1):76-80

[90] Xia Shuang-zhi, Liu Hong-wei, Jiu Bo. A Method of Relay of Tracking Based on Bayesian Theory[J]. Journal of Electronics & Information Technology, 2011, 33(3): 652-658(in Chinese)
夏双志,刘宏伟,纠博. 基于贝叶斯理论的一种接力跟踪方法[J]. 电子与信息学报,2011,33(3):652-658