第 60 卷　第 3 期
2024 年 5 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 60, No. 3
May, 2024

DOI:10.13232/j.cnki.jnju.2024.03.009

基于语境与文本结构融合的中文拼写纠错方法

刘昌春[1],张　凯[1,2*],包美凯[2],刘　烨[2],刘　淇[1,2]

(1.中国科学技术大学计算机科学与技术学院,合肥,230027;2.中国科学技术大学大数据学院,合肥,230027)

摘　要:在中文拼写纠错任务的处理中往往存在对句子的语义理解不够且对于汉字的语音和视觉信息利用较少的问题,针对这一问题,提出一种基于语境置信度和汉字相似度的纠错方法(ECS).该方法基于深度学习的理论,融合汉字的视觉相似度、汉字的语音相似度以及微调过的预训练BERT模型,能自动提取句子语义并利用汉字的相似性.具体地,通过对预训练的中文BERT模型进行微调,使之能适应下游的中文拼写纠错任务;同时,利用表意文字描述序列获取汉字的树形结构作为视觉信息,采用汉字的拼音序列作为语音信息;最后,利用编辑距离得出汉字的视觉和语音相似度,并将这些相似度数据与微调过的BERT模型融合,以实现纠错任务.在SIGHAN标准数据集上的测试结果显示,和基准模型相比,提出的ECS方法其$F1$-score提升巨大,在检错层面上提升2.1%,在纠错层面上提升2.8%,也验证了将汉字的语境信息、视觉信息与语音信息融合用于中文拼写纠错任务的适用性.

关键词:中文拼写纠错,BERT,汉字语音相似度,汉字视觉相似度,预训练模型

中图分类号:TP391.1　　**文献标志码:**A

Research on Chinese spelling correction based on the integration of context and text structure

Liu Changchun[1], Zhang Kai[1,2*], Bao Meikai[2], Liu Ye[2], Liu Qi[1,2]

(1. School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230027, China;
2. School of Data Science, University of Science and Technology of China, Hefei, 230027, China)

Abstract: In Chinese Spelling Correction (CSC) tasks, there are often problems such as insufficient semantic understanding of sentences and less use of phonetic and visual information of Chinese characters. Addressing these issues, we propose a novel error correction method based on context confidence and Chinese character similarity for Chinese spelling error correction (ECS). Based on deep learning principles, this approach integrates visual similarity of Chinese characters, and phonetic similarity of Chinese characters, and a fine-tuned pre-trained BERT model, which automatically extracts sentence semantics and exploits the similarity of Chinese characters. Specifically, we fine-tune the pre-trained Chinese BERT model to adapt to downstream Chinese spelling correction tasks. Then, we use the ideographic description sequence to capture the tree structure of Chinese characters as visual information and the phonetic sequence of Chinese characters as phonetic information. Finally, combining the visual and phonetic similarity (calculated by Levenshtein distance) of Chinese characters with the fine-tuned BERT model, we achieve the completion of the correction task. Experimental results on SIGHAN benchmark datasets show that the proposed ECS method has a huge improvement in $F1$-score compared with the baseline model, which is 2.1% higher on the error detection level and 2.8% higher on the error correction level, verifying the applicability of the fusion of context information, visual information and phonetic information for Chinese spelling correction tasks.

Key words: Chinese spelling correction, BERT, phonological similarity of Chinese characters, visual similarity of Chinese characters, pretrained model

基金项目:国家重点研发计划(2021YFF0901003)

收稿日期:2023－12－28

*通讯联系人,E-mail:kkzhang08@ustc.edu.cn

中国互联网络信息中心2023年3月2日发布的第51次《中国互联网络发展状况统计报告》指出,中国网民规模已达10.67亿,互联网普及率高达75.6%,移动网络终端连接总数为35.28亿户,移动物联网连接数增长至18.45亿户.在自媒体时代,网民的数量庞大,每个人都能在互联网上发布消息,互联网因此充满了来自短视频平台、微博、新闻网站、公众号等多元化渠道的海量文本信息.由于个人在发布信息时往往无法确保信息的准确性和严谨性,因此,这些文本信息存在大量的不规范甚至错误的字词和语句.

值得注意的是,随着大数据和人工智能的飞速发展,这些文本不仅被用于人与人之间的信息传递,还更广泛地被用于推荐系统、问答系统和精确搜索等领域.准确无误的文本信息不仅能避免错误含义的传播,还能提高相关系统的性能,提升用户在使用推荐系统、问答系统或精确搜索等应用时的体验.因此,无论是为了确保信息传播的严谨性和准确性,还是为了在人工智能应用中提供更好的用户体验,纠正文本中的错误至关重要.然而,对于互联网上非常庞大的数据量,人工纠错几乎无法实现,因此,研究并实施自动的中文拼写纠错(Chinese Spelling Correction,CSC)方法的需求日益迫切.

作为自然语言处理领域的重要分支,拼写纠错的研究已有相当长的历史.20世纪90年代中国即开始研究中文拼写纠错的方法,和国际上的相关研究相比,起步较晚.并且,由于中英文在语言特征上的显著差异,如中文不采用空格分割词语,需要依赖上下文语境来进行分词,且中文的语义与语法高度依赖于上下文,因此,英文的纠错方法在中文纠错场景中往往无法直接应用.

在多年的发展历程中,中文拼写纠错的研究方向大致分几种:基于规则的方法、基于机器学习的方法以及当前主流的基于深度学习的方法和利用大语言模型进行纠错的方法.

基于规则的方法主要采用预设规则来识别并纠正文本中的错误.例如,徐连诚和石磊[1]设计了一种方法,通过比较考生输入的有序字符集和标准答案并利用基于动态规划的最长公共子序列长度求解算法,解决计算机考试中文本输入的自动评分问题.龚小谨等[2]提出一种基于两阶段检测的方法来检测中文文本中的语法错误:在第一阶段对全文进行扫描,并根据从错误语料库中总结的搭配错误规则,通过模式匹配进行搭配错误的检测;在第二阶段采用句元分析的方式,对全文进行再次扫描,以检测与句元有关的语法错误.这种两阶段的检测可以有效地发现并纠正文本中的各种类型错误.尽管基于规则的中文拼写纠错方法能在一定程度上提高纠错的准确性,但也存在一些明显的缺点.首先,这种方法需要大量的语言学知识,而非语言领域的人员在这方面的知识常常是缺乏的,这意味着学习或编写相关规则需要投入大量的时间和金钱成本.其次,即使投入巨大的成本去编写相关规则,也难以保证规则在纠错任务中的完备性.随着计算机技术的飞速发展和计算能力的显著提升,这种基于规则的方法已经慢慢淡出了人们的视野.

基于机器学习的方法是利用大规模语料库中的数据来识别和修复文本中的错误,该方法的核心思想是基于文本的频率分布统计来推断错误单词的正确形式,并通过推测来修正原始文本.例如,陈笑蓉等[3]采用词性邻接关系检查、词语接续关系检查等策略进行错误查找,并最终选择第一候选项作为纠错结果.马金山等[4]使用3-gram来进行局部分析和查错,同时利用依存句法进行全局分析,从而揭示了词之间的依存关系,并发现了远距离错误.张仰森等[5]根据正确文本分词后单字词的出现规律以及“非多字词错误”的概念,提出一组错误发现规则,对分词后的单字散串,结合字二元、三元统计模型和词性二元、三元统计模型,建立了文本自动查错模型及实现算法.于勐和姚天顺[6]融合了语言模型和词语匹配两种方法,基于最长模式匹配的规则进行中文分词,然后用语言模型计算词与词之间的共现频率,将低于阈值的词语标记为可疑错字.但这些方法有一定的限制.首先,机器学习的性能往往受限于训练数据的规模和质量,需要足够的数据来进行可靠的推断;其次,其在处理复杂的非线性关系时表现较差,因其通常基于线性模型或者简单的概率模型,对于复杂的数据关系可能难以建模.

随着计算机计算能力的不断提升和大数据时

代下数据的不断丰富,以深度学习为代表的各种模型算法层出不穷,和传统机器学习相比,这些模型大大提升了对特定任务的效果,并逐渐成为主流的机器学习算法. 例如,Li et al[7]利用Transformer设计了检错网络和纠错网络,仅纠正被检错网络遮盖的字. Hong et al[8]利用预训练的BERT,采用特殊的数据遮罩模式对数据进行遮罩,并对模型进行微调,将模型输出的置信度与汉字的视觉与语音相似度相结合,拟合二者曲线来寻找最优解. Wang et al[9]提出一种基于指针网络的复制机制来减少搜索空间并提高生成正确字的概率,同时,还利用混淆集来确保生成的字只能在混淆集中,而不是在整个词表中. Cheng et al[10]利用图卷积网络来提取汉字的语义,并将图卷积网络的输出作为预训练模型输出结果的分类器.

综上,统计学方法往往要求手动设计大量的规则以描述语言特性并且难以有效地提取长距离的依赖关系,即其对语义的理解能力不够;同时,深度神经网络的鲁棒性通常较差,对微小的改动很敏感,而中文拼写纠错任务中,输入和输出的句子,除了错误的部分外其余部分保持一致,模型难以理解和分辨其中错误的拼写,也无法利用汉字的相似特性,因此表现效果不佳. 总体上,目前存在两个挑战. 第一,难以识别部分正确但不合适的词语,即对句子的语义理解不够,例如,“太阳从地平线西边缓缓升起”中的“西边”是一个正确的方位,但结合上下文信息会发现“东边”才是正确的语境. 第二,拼写错误的字往往具有较强的相似性,识别难度高,例如,“中国的首都是北京式”,这里的“式”和正确的“市”具有相同的读音.

对此,本文提出基于语境置信度和汉字相似度的纠错方法(ECS). 针对第一个挑战,通过引入预训练中文BERT来引入预训练过程中的背景知识,并让输入的句子在此基础上进行编码,从而融入当前的语境特征;针对第二个挑战,通过对文本结构,即汉字的相似度进行建模,帮助模型识别相似的错别字.

具体地,本文使用预训练中文BERT作为模型的基础,并对预训练模型在中文拼写纠错任务上进行微调,之后将微调后BERT模型输出的结果作为每一个位置不同汉字的置信度;同时,利用表意文字描述序列和拼音序列来计算汉字之间的相似度;最后,将相似度与置信度进行融合,综合评价每一个位置最可能的汉字. 将本文提出的方法在SIGHAN13/14/15的测试集上进行测试,测试结果证明了该方法的有效性.

本文提出的基于语境置信度和汉字相似度的纠错方法的主要贡献如下.

(1)提出一种全新的基于语境置信度和汉字相似度的中文拼写纠错方法,可以修改正确但不合适的词语以及错误且相似的词语.

(2)验证了融合汉字的语境信息、视觉信息与语音信息对中文拼写纠错任务的适用性.

(3)在SIGHAN数据集上的实验表明,和基准方法相比,提出的模型在检错层面和纠错层面的性能有较大的提升.

1　相关工作

以深度学习为代表的各种模型算法已经逐渐成为主流的机器学习算法. 一般地,利用深度学习来进行中文拼写纠错任务的大致步骤如下.

(1)数据预处理:收集相关数据,并对汉字进行繁体简体转换、分词、随机遮罩等操作.

(2)构建模型:模型通常由编码器和解码器组成. 编码器将输入的文本转换成一组特征向量,解码器则将特征向量映射为纠正后的文本. 通常采用循环神经网络、卷积神经网络、自注意力机制等网络结构来构建模型.

(3)模型训练:利用处理好的数据来训练神经网络模型,并设置合适的损失函数、学习率和优化算法等参数.

(4)模型测试:将待纠错的文本输入训练好的神经网络模型,并将纠正后的文本与正确的句子相比较,计算相关评价指标.

近年来,对中文拼写纠错的研究主要是基于深度学习的方法,具体如下.

Li et al[11]提出一种基于字符替换的方法,鼓励模型探索看不见的拼写错误,创建大型伪数据来训练模型. Zhang et al[12]发现了预训练模型检错能力的不足,并提出检错网络与纠错网络架构,利用Bi-GRU (Gate Recurrent Unit)作为检错网络,将soft-mask的检错结果输入BERT组成的纠

错网络. Liu et al[13]利用GRU获得汉字的语音与视觉编码,加入词嵌入并利用预训练中文模型RoBERTa来进行预测,最后通过预测拼音和汉字并进行联合优化来进行纠错. Wang et al[14]将语音加入词嵌入并利用动态规划算法和语音相似性来解决以往预测中预测词语不连贯的问题. Liu et al[15]为了解决错别字给句子带来的噪声,在预训练时给每个训练样本构造一个噪声上下文,然后修正模型被迫产生相似的基于噪声和原始上下文的输出. 为了解决过度纠正问题还提出了复制机制,如Zhang et al[16]将语音相似性融入预训练过程,用语音相似的单词替换原单词,并利用嵌入的拼音和预训练模型一起改正错误.

Guo et al[17]针对以往方法没有考虑句子中的错误词对句子表征及纠错任务的影响,训练了一个针对纠错任务的预训练语言模型BERT_CRS,提出全局注意力解码器(Global Attention Decoder,GAD)方法来学习输入句子中正确字符和错误字符候选集之间的关联关系,减轻局部错误的上下文影响. Bao et al[18]提出基于组块解码的方法,以单字、多字词、短语、成语为组块,利用全局优化修改单字和多字词错误,结合发音、字形、语义混淆集来处理多种不同的错误. Xu et al[19]使用文本、声音、视觉三个编码器来学习信息表示,采用BERT作为语义编码器的主干来捕获文本信息,使用分层编码器处理字符级和句子拼音字母,使用ResNet对图像进行分块编码,构建了多通道字符图像作为图形特征,得到字符图形标识. Huang et al[20]分别利用VGG19,TTS和BERT模型来获取汉字的视觉、语音和语义信息,并联合三者信息进行纠错. Ji et al[21]同样利用视觉和语音信息,分别采用部首偏旁信息,将整个拼音视为整体而非序列,并利用图网络进行信息融合.

Wang and Shang[22]针对连续错误会改变语义导致无法有效提取语义的问题,提出基于混淆集生成多个候选句子来改正错误. Zhu et al[23]基于Transformer和BERT设计一个多任务的网络修正模型,保留错误字的信息,使预测更准确. Yang and Yu[24]通过复制机制来解决BERT过度纠正的问题,并将注意力机制用于提取汉字的语音与形状信息. Bao et al[25]利用Bi-GRU编码汉字语音信息,利用表意文字序列(Ideographic Description Characters,IDS)编码汉字结构信息,多通道融合进行纠错. Zhao et al[26]利用视觉转换器、GRU+CNN和BERT分别提取字形、拼音和语义特征,并对三种嵌入进行平均来生成最终的多模态表示,送入Transfomer进行纠正. Li et al[27]认为错误的拼写会导致错误的分词,提出在嵌入层加入分词信息并加以训练来提高分词正确率,提升模型效果. Li et al[28]引入汉语语音预测辅助任务来改进汉语语音识别,并首次系统讨论了该辅助任务的自适应性和粒度问题. Sun et al[29]提出一种错误检测方法,引导模型更多地关注编码过程中可能出现的错误标记,并引入一个新的损失函数来整合错误混淆集,使模型能区分容易被误用的词语. Li[30]提出一个名为uChecker的框架来进行无监督拼写错误检测和纠正,还提出一种混淆集引导掩蔽策略来精细训练掩蔽语言模型,以进一步提高无监督检测和校正的性能. Jiang et al[31]定义了医学领域中文拼写纠正(Medical CSC,MCSC)的任务,并提出MCSCSet,这是一个包含大约20万个样本的大规模专家注释数据集,还论证了开放领域和医学领域的拼写纠正之间存在的显著差距. Sun et al[32]提出一种新的基于知识图谱的纠错方法,从知识图谱中查询三元组,并将查询到的三元组作为领域知识注入句子,使模型具有推理能力和常识.

2 问题形式化

给定初始句子$S_u=\{c_1,c_2,\cdots,c_n\}$,其中,$n$表示句子的长度,目标是判断给定的句子中是否存在错误的字,如果有则找到并纠正句子中错误的字c_i,使更正后的句子与拼写正确的目标句子$S_t=\{c_1',c_2',\cdots,c_n'\}$一致. 具体地,从检错和纠错两个层面进行研究.

3 基于语境置信度和汉字相似度的纠错方法

基于语境置信度和汉字相似度的纠错方法(ECS)首先利用中文拼写纠错任务常用数据集(SIGHAN13/14/15和Wang271K[33])来微调中文

BERT模型,使其适应下游的中文拼写纠错任务,获得当前字的语境置信度;然后,利用表意文字描述序列获取汉字的视觉信息,利用拼音序列获取汉字的语音信息,利用编辑距离综合视觉信息和语音信息获得汉字相似度;最终将语境置信度和汉字相似度进行融合,从而对句子进行检错和纠错. 本文提出的模型如图1所示.

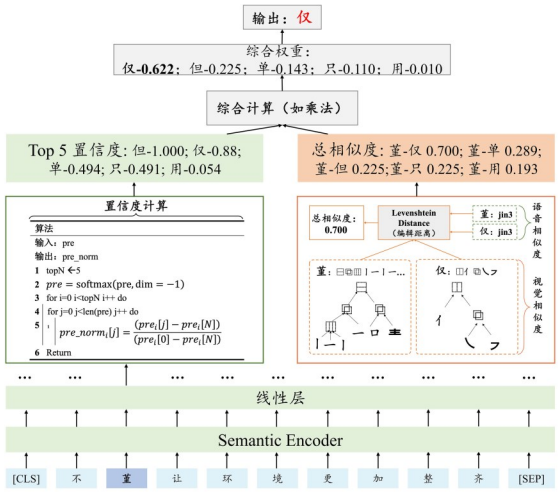

图1　ECS方法的主体框架

Fig. 1　The main framework of ECS

3.1　语境融合的置信度建模方法　由于BERT等预训练模型在预训练时使用了预测[MSAK]词的方法,因此具有从上下文推测缺失词语的能力. 本文尝试通过中文BERT对各个字的上下文语境进行融合,实现语境融合的置信度建模.

为了让预训练模型更好地了解任务,首先对模型进行微调,中文BERT微调的具体过程如下. 假设一个句子 $src=\{c_1, c_2, \cdots, c_n\}$,长度为 n,其目标句子为 $tgt=\{c'_1, c'_2, \cdots, c'_n\}$,设置句子的最大长度为 max _*length*,预训练BERT模型的词表大小 $vocab\_size=21128$. 将句子分词,加入特殊标记,补全到 max _*length* 长度并转换为词表索引后,可以得到模型需要的输入 $src'=[101, t_1, t_2, \cdots, t_n, 0, 0, \cdots, 0, 102]$和对应的标签 $tgt'=[101, t'_1, t'_2, \cdots, t'_n, 0, 0, \cdots, 0, 102]$,其中,101和102分别为[CLS]和[SEP]在词表中的索引,0对应[PAD]的索引,$t'_i, t_i \in [0, 21127]$是词对应词表中的索引.

设模型为BERT,模型的输出为BERT的最后一个隐层的向量,其维度为 $n\times 768$,n 为句子长度. 为了得到预测词在词表中的概率,还需对每一个位置的输出连接一个大小为 768×21128 的线性层Linear,则其输出 *pre* 可以表示为式(1):

$$pre=Linear\left(BERT\left(src'\right)\right) \tag{1}$$

然后,将预测值转换为0～1的概率 *pre*′,对各个词进行打分,将其作为置信度,如式(2)所示:

$$pre'=Softmax\left(pre, dim=-1\right) \tag{2}$$

其中,*Softmax* 为归一化指数函数. *pre*′的形状是 $n\times 21128$,pre'_i 代表模型预测的第 i 个词的概率分布. 由此可得模型优化的目标,如式(3)所示:

$$Loss=-\sum_{i=1}^{n}\sum_{j=0}^{21127}\lg P\left(t'_i \mid pre'_i\right) \tag{3}$$

利用上述方法对BERT模型在中文拼写纠错数据集上进行微调,通过优化目标函数来不断修改BERT模型参数,进而从上下文语义融合的角度完成对各个词的置信度建模.

3.2　汉字的视觉相似度　文本纠错中一类常见的错误就是视觉相似的错误,常常由于汉字的形状结构过于接近导致OCR识别或者汉语初学者犯错,故识别并改正这些视觉相似的汉字是提升模型性能的关键.

获取汉字的视觉信息有很多种方法,如Wang et al[33]将汉字的笔画顺序作为汉字的视觉信息,但这个方法存在不足,即对于笔画相同但结构不同的汉字,这个方法没有办法区分其不同. 比如这种方法无法区分汉字“由”和“田”,故其对于汉字视觉相似性的描述不够精确.

因此,本文采用表意文字描述序列(Ideographic Description Sequence,IDS)来表示字符的形状. Unicode组织从版本3.0开始支持中日韩越四种语言的统一表意文字(CJKV Unified Ideographs)的IDS,如表1所示,其利用十二种组合字符来描述文字内部构字部件的相对位置,从而精确地表示文字结构.

利用上述字符和汉字的基本笔画就可以画出基于汉字结构的树状图,而树状结构的先序遍历所得的序列可以作为汉字的视觉信息. 图2展示了汉字“田”和“由”的树状结构以及先序遍历获得的IDS. 可以看出,汉字“田”的IDS为“⿴ ⿴ 丨 冂

表 1 IDS 结构字符

Table 1 Ideographic description characters

编码	字符	例字
U+2FF0	⿰	相
U+2FF1	⿱	吉
U+2FF2	⿲	滩
U+2FF3	⿳	京
U+2FF4	⿴	回
U+2FF5	⿵	同
U+2FF6	⿶	凶
U+2FF7	⿷	匠
U+2FF8	⿸	局
U+2FF9	⿹	戒
U+2FFA	⿺	超
U+2FFB	⿻	巫

⿱⿻一丨一”,汉字“由”的 IDS 为“⿻⿰丨丁⿱⿻一丨一”. 虽然它们在笔画上无法区分,但在 IDS 中却有不同的序列,意味着其可以被区分. 同时,“田”和“由”的 IDS 非常相似,表明“田”和“由”有很强的视觉相似性. 这个例子从一定程度上体现了 IDS 和汉字笔画相比的优越性.

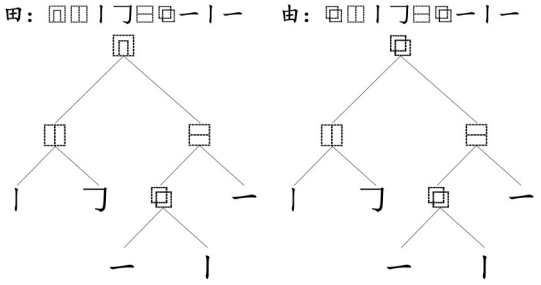

图 2 汉字“田”和“由”的树状结构及其 IDS

Fig. 2 The tree structure of Chinese characters “田”, “由” and their IDS

3.3 汉字的语音相似度 文本纠错中比视觉相似的错误更常见的是语音相似性的错误,通常是汉字读音过于接近或相同导致 ASR 或汉语初学者出错,故识别这些读音相似的汉字也同样是改正错误的关键.

由于汉字的拼音可以直接诠释汉字的读音,故利用汉字的拼音序列可以判断汉字的语音相似性,本文将中文汉字转换为它们对应的拼音,再进行后续的语音相似度计算.

pypinyin 是一个 Python 库,可以将中文汉字转换为它们对应的拼音. 它是一个轻量级的库,提供简单易用的 API 和多种转换选项,能方便地集成到中文自然语言处理应用程序中. 具体地,本文使用 lazy_pinyin 函数将汉字转换为 Style. 表 2 展示了汉字转成拼音序列的 TONE3 风格示例,序列由“拼音+声调(由数字代表)”组成.

表 2 将汉字转换成拼音序列的示例

Table 2 Examples of converting Chinese characters into pinyin sequences

汉字	拼音序列
田	tian2
由	you2

3.4 汉字的相似度 使用编辑距离(Levenshtein Distance)来计算两个字符串之间的相似度,其衡量了将一个字符串转换为另一个字符串所需的最少编辑次数,编辑操作包括插入、删除和替换字符. 这个字符串可以是 IDS 也可以是拼音序列,并且,Levenshtein Distance 越小的字符串之间的相似度更高,进而具有相似的形状或者读音. 计算 Levenshtein Distance 一般使用动态规划算法,其状态转移方程如式(4)所示:

$$lev_{a,b}(i,j)=\begin{cases}\max(i,j) & \text{if } \min(i,j)=0\\ \min\begin{cases}lev_{a,b}(i-1,j)+1\\ lev_{a,b}(i,j-1)+1\\ lev_{a,b}(i-1,j-1)+1_{(a_i\neq b_j)}\end{cases} & \text{otherwise}\end{cases} \tag{4}$$

式(4)第一行是算法的初始化方法;$(a_i\neq b_j)$是指示函数,$a_i\neq b_j$时为 1,$a_i=b_j$时为 0. 假设 a,b 是两个字符串且 a 的字符串长度为$|a|$,b 的字符串长度为$|b|$,则字符串 a 与 b 的编辑距离就是 $lev_{a,b}(|a|,|b|)$.

为了消除不同字符串长度对结果的影响,并使数值越大相似度越高,所以对计算结果进行归一化处理,如式(5)所示:

$$lev'_{a,b}=1-\frac{lev_{a,b}(|a|,|b|)}{\max(|a|,|b|)} \tag{5}$$

最终,将计算得到的语音相似度和视觉相似度按一定权重进行加权求和,得出汉字相似度.

3.5　语境置信度和汉字相似度结合　对于模型的输出pre,可知每个字的预测结果pre_i,将pre_i的预测值从大到小排序,选择最大的N个值(实验中设置为5),这些值称为模型预测的置信度.这N个值代表模型认为的在i处最有可能的N个字,将这N个字作为输出的候选.计算并利用这N个字和原字的相似度以及置信度就可以进一步提高模型的性能,关键是对它们的综合利用.

由于测试时模型输出的值没有经过$Softmax$归一化,但相似度已经归一化了,所以,为了量纲统一,需要将pre_i归一化.和$Softmax$归一化相比,式(6)只考虑了前$N+1$个值,没有考虑词表中的所有词.这样做是为了拉开置信度差距,因为根据实验结果,发现pre_i的前几个候选值非常接近,按照$Softmax$归一化的方法进行操作会使归一化后的值仍然非常接近,使模型预测的置信度的作用大大降低,最终会只按字的相似度进行输出,所以通过这种方法可以使被考虑的词之间的置信度均匀分布.此外,因为最终要考虑N个候选,归一化时就必须将第$N+1$个词作为最小值,否则第N个候选的置信度将为0,这是毫无意义的.综上,使用式(6)的方法进行归一化:

$$pre\_norm_i[j]=\frac{\left(pre_i[j]-pre_i[N]\right)}{\left(pre_i[0]-pre_i[N]\right)} \tag{6}$$

此外,实验发现,两个字的相似度如果很高,则其视觉和语音相似度必须也很高,而这种情况的可能性是很小的.这就带来一个问题,假如某一候选字和输入的字一模一样的话,两者之间的相似度就是1.0,这大大提高了其作为最终输出的可能性,但忽略了其余相似度也很高且置信度大于这个字的候选项,这本质上是因为前文描述的相似度计算方法得出的值存在断层现象.针对这个问题,本文对字的相似度作如下修正,如果候选字和原字相同,则用此字的置信度代替此字的相似度,如式(7)所示:

$$total'_{\text{similarity}_{d(t_j),d(id(j))}}=\begin{cases}pre_{\text{norm}\,i}[j] & \text{if } t_j=id(pre_i)\\ total_{\text{similarity}_{d(t_j),d(id(j))}} & \text{otherwise}\end{cases} \tag{7}$$

最终,将语境置信度和汉字相似度进行结合来完成中文拼写纠错的任务.综合考虑置信度和相似度有很多种方法,如采用二者相乘的方法来进行综合评判,如式(8)所示:

$$value[j]=total'_{\text{similarity}_{d(t_j),d(id(j))}}\times pre_{\text{norm}\,i}[j] \tag{8}$$

此外,本文还尝试使用线性加权的方法来融合相似度和置信度,如式(9)所示:

$$value[j]=total'_{\text{similarity}_{d(t_j),d(id(j))}}\times W+pre_{\text{norm}\,i}[j]\times(1-W) \tag{9}$$

最后,选择最大的$value$对应的索引作为最后的输出$\bar{t}_i$,将$\bar{t}_i$和正确的字t_i'进行对比来计算相关指标,如式(10)所示:

$$\bar{t}_i=id\left(argmax(value)\right) \tag{10}$$

4　实验

4.1　数据集介绍　SIGHAN数据集是中国台湾学者公开的用于中文文本纠错任务的数据集,其测试集已成为中文拼写纠错领域中判断模型好坏的基准数据集.SIGHAN目前包含三个版本,分别是SIGHAN13,SIGHAN14,SIGHAN15,其数据量如表3所示.但其存在一些问题,首先,这个数据集中的句子都由繁体字组成,而目前主流的中文拼写纠错任务是基于简体字来完成的,所以本文利用Open Chinese Convert(OpenCC)工具包将其转换为简体字.此外,SIGHAN数据集还有很多不一致的地方,包括编码问题、句子编号对不上、标签问题等,但这些错误的量不大,较易修复.为了解决SIGHAN数据集中的不一致问题,可以根据其特点,使用基于规则的方法进行简单修复,方便后面的实验.

表3　SIGHAN数据集的数据量

Table 3　The amount of data in SIGHAN dataset

SIGHAN	句子数量	平均长度	错字数量
13(训练集)	700	41.8	343
14(训练集)	3437	49.6	5122
15(训练集)	2339	31.3	3037
13(测试集)	1000	74.3	1224
14(测试集)	1062	50.0	771
15(测试集)	1100	30.6	703

4.2　实验评价指标　一般常用检错能力和纠错能力两个层面来评价模型的中文文本纠错性能,每个层面都有四个指标,即精确率(Accuracy,

Acc)、准确率(Precision,P)、召回率(Recall,R)、$F1$指数($F1$-score),具体如表4所示,计算如式(11)~(14)所示:

$$Acc=\frac{TP+TN}{TP+FP+TN+FN} \tag{11}$$

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R=\frac{TP}{TP+FN} \tag{13}$$

$$F1\text{-score}=\frac{2\times P\times R}{P+R} \tag{14}$$

表4 实验评价指标的细则

Table 4 Detailed rules for experimental evaluation indicators

检错层面		
	预测有错	预测无错
句子本身有错	*TP*:该纠的字都纠了,纠没纠对不管;不该纠的字都没纠	*FN*:该纠但未纠或把不该纠的字纠了
句子本身没错	*FP*:不该纠但纠了	*TN*:不该纠也没纠
纠错层面		
	预测有错	预测无错
句子本身有错	*TP*:该纠且纠对了	*FN*:该纠但未纠或纠错了
句子本身没错	*FP*:不该纠但纠了或纠了但没纠对(原句是否有错不管)	*TN*:不该纠也没纠

4.3 对比实验方法 为了更全面地评估基于语境置信度和汉字相似度的纠错方法在中文纠错任务中的效果,本文引入了三个基准中文纠错模型,即LMC[34],SL[33]和PN[9],并与本文提出的模型在检错和纠错两个方面进行比较,以此更好地了解本文的方法在纠错任务中的优势和局限性.

Xie et al[34]引入一种基于n-gram的语言模型方法(LMC),利用混淆集替换字符,然后通过结合2-gram和3-gram语言模型对修改后的句子进行评估,并利用动态规划算法加快效率.

Wang et al[33]针对Wang271K数据集,提出一种基于LSTM(Long Short-Term Memory)的序列标注(SL),使用序列标注模型BiLSTM来检查并标记句子中的错误,同时使用3-gram和混淆集来纠正检测出的错误.

Wang et al[9]提出基于指针网络的复制机制方法(PN),同时利用混淆集使生成的字只能在混淆集中而不是在整个词表中,从而减少搜索空间并提高生成正确字的概率.

大语言模型已是自然语言处理领域中绕不开的话题,探究大语言模型在自然语言处理任务上的性能非常重要.本文使用ChatGPT大语言模型进行中文拼写纠错,并与提出的方法进行比较.大语言模型解决下游任务的主流方法是使用提示工程(Prompt Engineering),即设计和优化用于训练AI模型的Prompt,通过清晰、简洁和具有针对性的Prompt来优化模型的性能和效果.直接使用ChatGPT等大语言模型进行文本纠错的效果很差,因为模型存在同义改写、字符转换、句式变换等问题,纠错后的句子虽然语义可能不变,但与文本纠错的目标相违背.故可以通过提示工程来找到适合于文本纠错的Prompt,进而提高其在文本纠错任务上的性能.具体地,本文利用ChatGPT接口在SIGHAN15数据集上进行纠错.

4.4 实验配置 对于基于语境置信度和汉字相似度的纠错方法模型,设置模型的*batch_size*为64,*epoch*为10,*learning_rate*为1e−5,Top n为5.对于上文的三个基准方法,使用原论文中的初始配置.ChatGPT使用的Prompt是:"请纠正下列文本中的错字,纠正时仅输出纠正后的结果,尽量保证和原句长度相同,纠正字与原句读音相近或相同:{ }".

4.5 实验结果 表5展示了融合BERT置信度和汉字相似度的纠错方法(ECS)及三个基准方法在SIGHAN测试集上的效果,表中黑体字表示最大的数值,表6展示了ChatGPT在SIGHAN15测试集上的表现.

和LMC,SL和PN方法相比,ECS在检错层面的$F1$-score高24.9%,13.7%和2.1%,在纠错层面的$F1$-score高25.0%,16.7%和2.8%,即在检错层面和纠错层面均有很大提升.

以上结果表明,ECS方法能更准确地检测和纠正文本中的错误.因为融合BERT置信度和汉字相似度,ECS能更好地理解和处理文本中的语义和语法信息,提高了纠错的准确性.同时还可以看到,ChatGPT在中文拼写纠错任务上还有较

大的提升空间.

表5　ECS和对比的基准方法在SIGHAN数据集上的测试结果

Table 5　Test results of ECS and benchmark method on the SIGHAN dataset

模型	检错层面				纠错层面			
	Acc	*P*	*R*	*F*1-score	*Acc*	*P*	*R*	*F*1-score
LMC[34]	(—)	73.3%	37.0%	49.2%	(—)	72.1%	34.9%	46.8%
SL[33]	(—)	54.2%	68.3%	60.4%	(—)	(—)	(—)	55.1%
PN[9]	(—)	62.3%	**82.3%**	72.0%	(—)	76.8%	62.6%	69.0%
ECS	**77.6%**	**80.2%**	68.8%	**74.1%**	**76.3%**	**77.7%**	**66.7%**	**71.8%**

表6　ChatGPT在SIGHAN15上的纠错效果

Table 6　Test result of ChatGPT on SIGHAN15

	Acc	*P*	*R*	*F*1-score
检错层面	29.7%	20.1%	34.9%	25.5%
纠错层面	26.4%	16.2%	28.0%	20.5%

4.6　消融实验　表7展示了基于语境置信度和汉字相似度的纠错模型ECS在SIGHAN测试集上的效果,并与三个已经提出的模型进行了对比,表中黑体字表示最大的数值.其中,BERT表示直接使用预训练模型BERT进行中文拼写纠错,ECS w/o finetuning表示使用ECS方法中未经微调的预训练模型,ECS w/o sim表示使用按照本文方法进行微调的模型但去除了本文汉字的语音相似度和视觉相似度的方法.

由表可见,无论BERT模型是否经过微调,只要在测试中加入针对汉字相似度的判断就能使模型效果有一定提升,在检错和纠错层面使*F*1-score提升2%～8%.因为BERT模型预训练的任务是完形填空,其中就有对[MASK]对应的内容进行预测,这种训练方法可以学习到句子的上下文语义特征.然而,仅仅语义正确在中文拼写纠错领域是完全不够的,因为中文拼写纠错的目标是猜测句子中错误汉字的真正原字,不是在语义大致不变的基础上替换成一个新字而使句子没有错误.如前所述,中文拼写纠错的错误类型中有很大一部分是读音和视觉相似的汉字混淆造成的,除了利用BERT学习到的语义特征外,还要结合原字和候选字的相似度,才能使改正的字不仅使句子正确,还真正还原了错字代表的原字.

此外,BERT模型在训练集的微调使其*F*1-score在检错和纠错层面提升22%～32%.这是由于BERT模型是基于大规模无监督预训练的通用语言模型,对于特定的任务,对可以学习到的通用的语言表示仅仅使用预训练模型来进行推理是不够的,在目标任务上对预训练模型进行微调可以使模型更好地适应该任务的数据和特征,提高模型在该任务上的泛化能力和性能.

上述分析和实验结果进一步证明了本文提出的计算汉字相似度的方法和在BERT上进行微调的方法对于中文拼写纠错的有效性.

4.7　参数实验　结合模型特征进行参数实验有助于得到更好的结果,本文主要从候选项数量、相似度与置信度结合权重两个方面进行实验,还对训练过程的稳定性进行了实验.

首先是候选项数量对实验结果的影响.在Top *n*为2～9时进行测试,测试结果的精确率和

表7　ECS和其他对比模型在SIGHAN数据集上的测试结果

Table 7　Results of ablation experiment of ECS and other models on the SIGHAN dataset

模型	检错层面				纠错层面			
	Acc	*P*	*R*	*F*1-score	*Acc*	*P*	*R*	*F*1-score
ECS	**77.6%**	**80.2%**	68.8%	**74.1%**	**76.3%**	**77.7%**	**66.7%**	**71.8%**
BERT	58.7%	57.6%	41.9%	48.5%	50.5%	39.8%	28.9%	33.5%
ECS w/o finetuning	60.3%	66.4%	40.8%	50.5%	55.7%	54.7%	33.6%	41.6%
ECS w/o sim	75.8%	70.7%	**69.6%**	70.1%	73.4%	66.9%	65.9%	66.4%

$F1$-score 如表 8 所示. 由表可见,不同候选值对实验结果影响不大,因为微调后的模型其本身具有较强的利用语境进行纠错的能力,所以 BERT 模型给出的前几个候选项已经包含了正确答案,增加候选项对结果的影响不大;此外,BERT 给出的置信度差距较大,不太可能的选项因为置信度极低,所以无论相似度多高都无法影响实验结果.

本文还探索了线性加权方法下不同权重对实验结果的影响. 实验结果如表 9 所示,表中黑体字表示最大的数值,实验折线图如图 3 所示. 可以看出,权重为 0.4 时效果最好,权重过大或过小时性能都有下降. 但是,在相似度占权重过大的情况下算法性能没有下降太多,可能是因为大部分地方没有错误,而输出原字的相似度与置信度都较高,被选中的概率较大.

表 8 不同个数候选项对实验结果的影响

Table 8 The impact of different top n on experimental results

候选个数		2	3	4	5	6	7	8	9
检错层面	P	78.0%	77.8%	77.8%	77.6%	77.6%	77.5%	77.5%	77.4%
	$F1$-score	74.1%	74.1%	74.2%	74.1%	74.0%	73.9%	74.0%	74.8%
纠错层面	P	76.3%	76.3%	76.4%	76.3%	76.1%	76.0%	76.1%	76.0%
	$F1$-score	71.3%	71.7%	71.9%	71.8%	71.6%	71.4%	71.5%	71.4%

表 9 不同线性加权权重对实验结果的影响

Table 9 The impact of different linear weights on experimental results

权重		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
检错层面	P	77.2%	78.5%	**79.0%**	**79.0%**	78.1%	77.8%	77.2%	77.2%	75.6%
	$F1$-score	71.8%	73.9%	75.2%	**75.4%**	74.3%	74.0%	73.4%	71.8%	71.9%
纠错层面	P	75.0%	76.6%	77.2%	**77.4%**	76.7%	76.3%	76.3%	75.0%	68.5%
	$F1$-score	68.2%	70.8%	72.3%	**72.7%**	72.0%	71.5%	71.5%	68.2%	59.7%

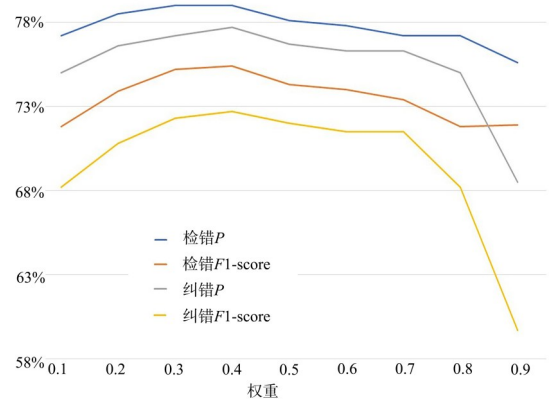

图 3 线性加权的权重对实验结果的影响

Fig. 3 The influence of linear weights on experimental results

最后还对训练过程的稳定性进行了实验,图 4 展示了模型的训练集损失随迭代次数的变化. 图中,迭代次数为 2k 时损失为 0.1451,在迭代次数在 2k~4k 时损失有一个急速下降,随后下降速度趋于稳定直至最后收敛,证明提出的模型能稳定地进行训练.

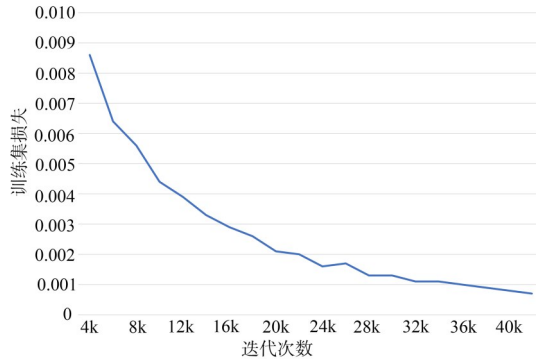

图 4 微调模型时损失的变化趋势

Fig. 4 The trend of loss changes during model fine-tuning

4.8 案例分析 选用 SIGHAN 测试集中的一句话进一步论证汉字相似度方法的有效性. 如图 5 所示,输入句子“不堇让环境更加整齐”,如果没有相似度计算,将输出 BERT 的第一候选,即“不但让环境更加整齐”;加上相似度计算后,排在第二

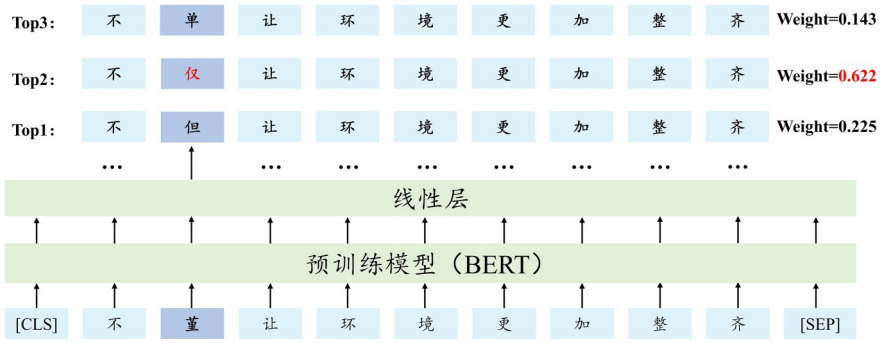

图5　选用SIGHAN测试集中的一句话进行案例分析

Fig. 5　A sentence from the SIGHAN test set for case analysis

的候选“不仅让环境更加整齐”的权重更大,被选中输出.对比这两句话,虽然都是正确的句子,但其错误原因,“不堇－不但”很难分析其错误原因,而“不堇－不仅”明显是语音相似的错误,更贴合原字.和无法追溯原因的错误相比,有原因的错误更可能在现实中发生,而SIGHAN测试集中将“不仅让环境更加整齐”作为正确的句子也证明了这一点.这个例子同样论证了添加汉字相似度计算可以更好地提高中文文本的纠错性能.

5　结论

本文首先介绍了中文拼写纠错任务、相关的数据集以及之前的研究方法等,然后提出一种基于语境置信度与汉字相似度融合的中文拼写纠错方法.首先,使用SIGHAN和Wang271K数据集在预训练中文BERT模型上进行微调,获得当前字的语境置信度,然后,将提出的计算汉字相似度的方法与语境置信度相结合,形成基于BERT的语境置信度、汉字语音相似度和视觉相似度的中文拼写纠错方法.通过广泛的实验,验证了所提出方法的有效性,也验证了融合汉字的语境信息、视觉信息与语音信息对中文拼写纠错任务的适用性.

参考文献

[1] 徐连诚,石磊.自动文字校对动态规划算法的设计与实现.计算机科学,2002,29(9):149－150.(Xu L C, Shi L. The design and application of a dynamic program algorithm in automatic text collating. Computer Science,2002,29(9):149－150.)

[2] 龚小谨,罗振声,骆卫华.中文文本自动校对中的语法错误检查.计算机工程与应用,2003,39(8):98－100,127.(Gong X J, Luo Z S, Luo W H. Automatically detecting syntactic errors in Chinese texts. Computer Engineering and Applications,2003,39(8):98－100,127.)

[3] 陈笑蓉,秦进,汪维家,等.中文文本校对技术的研究与实现.计算机科学,2003,30(11):53－55.(Chen X R, Qin J, Wang W J, et al. Research and implementation of Chinese text proofreading. Computer Science,2003,30(11):53－55.)

[4] 马金山,刘挺,李生.基于n-gram及依存分析的中文自动查错方法.北京:清华大学出版社,2003:597－603.

[5] 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.中文信息学报,2006,20(4):1－7,55.(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. Journal of Chinese Information Processing,2006,20(4):1－7,55.)

[6] 于勐,姚天顺.一种混合的中文文本校对方法.中文信息学报,1998,12(2):31－36.(Yu M, Yao T S. A hybrid method for Chinese text collation. Journal of Chinese Information Processing, 1998, 12(2):31－36.)

[7] Li J, Wu G S, Yin D F, et al. DCSpell: A detector-corrector framework for Chinese spelling error correction//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Online: ACM,2021:1870－1874.

[8] Hong Y Z, Yu X G, He N, et al. FASPell: A fast,

adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm//Proceedings of the 5th Workshop on Noisy User-Generated Text. Hong Kong, China: Association for Computational Linguistics, 2019: 160—169.

[9] Wang D M, Tay Y, Zhong L. Confusionset-guided pointer networks for Chinese spelling check// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 5780—5785.

[10] Cheng X Y, Xu W D, Chen K L, et al. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, WA, USA: Association for Computational Linguistics, 2020: 871—881.

[11] Li C, Zhang C Y, Zheng X Q, et al. Exploration and exploitation: Two ways to improve Chinese spelling correction models//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online: Association for Computational Linguistics, 2021: 441—446.

[12] Zhang S H, Huang H R, Liu J C, et al. Spelling error correction with soft-masked BERT//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Compu-tational Linguistics, 2020: 882—890.

[13] Liu S L, Yang T, Yue T C, et al. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online: Association for Computational Linguistics, 2021: 2991—3000.

[14] Wang B X, Che W X, Wu D Y, et al. Dynamic connected networks for Chinese spelling check// Findings of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021: 2437—2446.

[15] Liu S L, Song S K, Yue T C, et al. CRASpell: A contextual typo robust approach to improve Chinese spelling correction//Findings of the Association for Computational Linguistics. Dublin, Ireland: Association for Computational Linguistics, 2022: 3008—3018.

[16] Zhang R Q, Pang C, Zhang C Q, et al. Correcting Chinese spelling errors with phonetic pre-training// Findings of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021: 2250—2261.

[17] Guo Z, Ni Y, Wang K Q, et al. Global attention decoder for Chinese spelling error correction// Findings of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021: 1419—1428.

[18] Bao Z Y, Li C, Wang R. Chunk-based Chinese spelling check with global optimization//Findings of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 2031—2040.

[19] Xu H D, Li Z L, Zhou Q Y, et al. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking//Findings of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2021: 716—728.

[20] Huang L, Li J J, Jiang W W, et al. PHMOSpell: Phonological and morphological knowledge guided Chinese spelling check//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online: Association for Computational Linguistics, 2021: 5958—5967.

[21] Ji T, Yan H, Qiu X P. SpellBERT: A lightweight pretrained model for Chinese spelling check// Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics, 2021: 3544—3551.

[22] Wang S, Shang L. Improve Chinese spelling check by reevaluation//Proceedings of the 26th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2022: 237—248.

[23] Zhu C X, Ying Z Q, Zhang B Y, et al. MDCSpell: A multi-task detector-corrector framework for Chinese

spelling correction//Findings of the Association for Computational Linguistics: ACL. Dublin, Ireland: Association for Computational Linguistics, 2022: 1244－1253.

[24] Yang S J, Yu L. CoSPA: An improved masked language model with copy mechanism for Chinese spelling correction//Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence. Eindhoven, The Netherlands: PMLR, 2022: 2225－2234.

[25] Bao L J, Chen X S, Ren J W, et al. PGBERT: Phonology and glyph enhanced pre‐training for Chinese spelling correction//Proceedings of the 11th CCF International Conference on Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2022: 16－28.

[26] Zhao G C, Guo Y, Xia F L, et al. A multimodal method for Chinese spelling correction//2022 International Joint Conference on Neural Networks. Padua, Italy: IEEE, 2022: 1－7.

[27] Li F F, Shan Y R, Duan J W, et al. WSpeller: Robust word segmentation for enhancing Chinese spelling check//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 1179－1188.

[28] Li J H, Wang Q, Mao Z D, et al. Improving Chinese spelling check by character pronunciation prediction: The effects of adaptivity and granularity// Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 4275－4286.

[29] Sun R, Wu X Y, Wu Y F. An error-guided correction model for Chinese spelling error correction//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 3800－3810.

[30] Li P J. uChecker: Masked pretrained language models as unsupervised Chinese spelling checkers// Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022: 2812－2822.

[31] Jiang W J, Ye Z H, Ou Z J, et al. MCSCSet: A specialist‐annotated dataset for medical‐domain Chinese spelling correction//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta, GA, USA: ACM, 2022: 4084－4088.

[32] Sun X M, Zhou J, Wang S, et al. Chinese spelling error detection and correction based on knowledge graph//Database Systems for Advanced Applications. DASFAA 2022 International Workshops: BDMS, BDQM, GDMA, IWBT, MAQTDS, and PMBD. Springer Berlin Heidelberg, 2022: 149－159.

[33] Wang D M, Song Y, Li J, et al. A hybrid approach to automatic corpus generation for Chinese spelling check//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 2517－2527.

[34] Xie W J, Huang P J, Zhang X R, et al. Chinese spelling check system based on N‐gram model// Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing. Beijing, China: Association for Computational Linguistics, 2015: 128－136.

（责任编辑　杨可盛）