CPLM-CSC：基于单字级别预训练语言模型的中文错别字纠正方法*

谢海华[1]，李奥林[1]，李亚博[1]，陈志优[1]，程静[1]，吕肖庆[2]，汤帜[2]

（1. 北大方正集团有限公司数字出版技术国家重点实验室，北京市 100871；2. 北京大学计算机科学技术研究所，北京市 100871）

摘要：由于汉语语义表达的多样性和复杂性，中文错别字自动纠正目前存在很多挑战。现有的错别字纠正算法的性能普遍不够理想，而且需要大量高质量的语料进行训练。本文提出一种基于预训练语言模型的错别字纠正方法，CPLM-CSC，能够显著地提高纠错性能。CPLM-CSC 采用基于单字级别预训练语言模型来进行错别字检测，并采用掩字语言模型来进行错别字纠正。为了提高纠正性能，CPLM-CSC 采用音近形近字判断等多种筛选纠正结果的方法，并针对一些典型且特殊的错误，例如：“的地得”误用，采取了专门的数据增强方法。CPLM-CSC 在 SIGHAN 2015 的评测数据集上进行了测试，并取得了 0.654 的 F1 值，性能优于其他模型。

关键词：错别字纠正；预训练语言模型；单字级别模型；掩字语言模型；文档审校

中图分类号：TP391　　**文献标识码：**A

CPLM-CSC: Char-based Pre-trained Language Model based Approach for Chinese Spelling Checking and Correction

Haihua XIE[1], Aolin LI[1], Yabo LI[1], Zhiyou CHEN[1], Jing CHENG[1], Xiaoqing LYU[2], Zhi TANG[2]

(1. State Key Laboratory of Digital Publishing Technology (Peking University Founder Group Co. LTD.), Beijing, 100871, China; 2. Institute of Computer Science and Technology of Peking University, Peking University, Beijing, 100871, China)

Abstract: Due to the variability and complexity of Chinese semantic expression, Chinese spelling checking and correction is a challenging task. The current models of Chinese spelling correction normally do not perform well when dealing with complex grammatical and semantic errors, and they require a large-scale high-quality corpus for training. This paper proposes an approach based on pre-trained language models for Chinese spelling checking and correction, named as CPLM-CSC, which significantly improves the correction performance. In CPLM-CSC, the char-based pre-trained language model is employed for spelling checking, and a masked language model is designed for spelling correction. To enhance the correction performance, CPLM-CSC employs several ways of final result filtering and applies data enhancement means for certain special errors such as misuse of “的”, “地” and “得”. CPLM-CSC was tested on the dataset of SIGHAN 2015 and achieved the state-of-the-art performance, with F1 score of 0.654.

Key words: Chinese spelling checking and correction; char-based model; pre-trained language model.

1 引言

中文错别字检测和纠正（Chinese spelling checking and correction, CSC）的目标是检测并

* **收稿日期：2019.06.15**　　**定稿日期：2019.08.10**

基金项目：国家自然科学基金（No. 61472014, No. 61573028, No. 61432020）；北京市自然科学基金（No. 4142023）；北京市科技新星计划（XX2015B010）

作者简介：谢海华（1983 年生），男，博士，高级工程师，主要研究方向为知识图谱、自然语言处理；李奥林（1989 年生），男，硕士，主要研究方向为自然语言处理；李亚博（1990 年生），男，硕士，主要研究方向为自然语言处理；陈志优（1990 年生），男，硕士，主要研究方向为自然语言处理；程静（1994 年生），女，硕士，主要研究方向为自然语言处理；吕肖庆（1967 年生），男，博士，副研究员，主要研究方向为图形识别与检索；汤帜（1965 年生），男，博士，研究员，主要研究方向：文档处理技术，模式识别，数字版权保护技术。

纠正中文语句里的字词误用情况。与英文等拼音文字的拼写检查类似，CSC 可以分为两个类别：词错误（word error）纠正和非词错误（non-word error）纠正。词错误指的是两个正常词之间的误用，例如英文中的 quantity 和 quality 互相误用，中文里的“权力”和“权利”互相误用。非词错误指的是把一个正常词误写为一个非词（不在词典中的词），例如：把 quantity 误写为 quatity，把“进入”写成“进人”。

错别字通常发生在音近字（读音相近或相同的字）之间或者形近字（形状相似的字）之间。一般地，错别字纠正的方法分成两步。第一步是错别字的检测，即发现句子中的错别字并确定它的位置。通常基于语言模型的原理来设计错别字检测算法。第二步是错别字纠正，通常基于一个大规模的混淆集（Confusion Set）来给出正确字的候选集，并评估每个候选正确字的概率，概率最高且大于一定阈值的候选字被认为是纠正的结果。评估候选正确字的概率的方法也主要基于语言模型的原理来设计[1]。

汉语的某些特征使得中文错别字纠正具有很大的难度。首先，汉语中有错误的词，但是没有错误的字，而很多错别字发生在单字词之间（例如：“在”和“再”）。单字词的用法种类很多，需要结合上下文语义才能判断它们是否误用。其次，由于中文的词之间没有分隔符，非词出现在句子中不一定是错误的，例如：“进人”在句子“化合物被注射进人体”中就没有错误。另外，发生错别字的词，在语句分词时不能被正确地切分。因此，不适宜以词为单位进行错别字的纠正。

在当前的错别字纠正模型中，非词错误纠正是主要的研究内容，但是目前的准确率也只有 70%左右[2]。而词错误纠正的准确率更低，因为词错误主要发生在易混淆词之间，例如：“无需”和“无须”，不仅仅读音相同而且意思相近，需要基于语义分析才能准确地检测和纠正[3]。除了词错误和非词错误的纠错困难，错别字纠正工作还存在一个问题，即很多非错别字被误报。误报情况主要发生在训练数据或者词典中很少出现的词语当中，例如：专业名词、人名地名、网络用语等。

针对上述问题，我们提出一种基于单字级别预训练语言模型（CPLM-CSC）的方法来进行错别字纠正。由于错词典的构建极为复杂，而且错别字层出不穷，无法穷尽，CPLM-CSC 的错别字检测模块和纠正模块，都基于预训练语言模型来设计，以避免词典构建，并减少训练数据规模。由于词级别纠错模型的固有缺点，即：发生错别字的词无法被准确地切分，而且错别字纠正的目标是找出错字而非错词，CPLM-CSC 采用单字级别的模型来进行纠错，即：模型的输入和输出都是字序列，以避免分词的错误。在错别字纠正阶段，CPLM-CSC 采用掩字模型的方式，即：将检测出来的错别字掩盖，然后基于语言模型给出候选正确字的列表，并计算每个候选字的概率以选择其中最优一个作为输出。CPLM-CSC 还采取了语言模型微调、数据增强、排名阈值和概率阈值等方法，来提升错别字纠正的性能。在 SIGHAN 2015 的数据集上进行测试，CPLM-CSC 取得了 68.95%的准确率和 62.19%的召回率，F1 值为 0.654，性能优于其他模型。

2 相关工作

在现有的中文错别字纠正（CSC）算法和系统当中，语言模型理论、词切分、混淆集、错词典等是常用的技术和工具。CSC 的流程主要包括两个步骤：错别字检测和错别字纠正。错别字检测的目标是判断语句中哪个位置出现了错别字，而错别字纠正的目标是纠正识别出来的错别字。

错别字检测是 CSC 的重点和难点。Chang[4]提出了一个简单的做法，假设语句中所有字都是可能错误的，然后对每个字进行纠正。为了减少计算量，Lin and Chu[5]提出了一个优化的方案，假设语句分词之后所有的单字都是可能错误的。除此之外，Hsieh 等[6]使用基于未

知词检测和语言模型校验的方法，以及基于由混淆集生成的词典来识别错别字。最后，将各种方法产生的结果结合起来形成最终的识别结果。其中，混淆集里面包含所有汉字以及每个字对应的音近字和形近字。由混淆集生成的词典，即错词典，包含的是正确词与它对应的可能错词，而一个正确词对应的错词是基于音近字和形近字生成的。Yang 等[7]提出了一个高性能的模式识别器来增强分词后候选词识别的效果。Zhao 等[8]为每个句子构建了一个有向无环图，并运用单源最短路径算法识别一般的错别字情况。

在错别字纠正方面，Chang[4]基于混淆集，用音近字或形近字替换掉语句中的可能错字，然后使用二元语言模型来计算原始句子的概率和所有更改的句子的概率。尽管该方法能够有效地识别错别字，但是它的计算时间长，而且有非常多的漏报。Chiu 等[9]运用了统计机器翻译模型来将含有错别字的句子翻译成正确的句子，并选择拥有最高的翻译概率的句子为最终正确的句子。Yang 等[11]使用 ePMI 矩阵统计字词之间的共现情况，并基于 ePMI 矩阵，选择与前后文经常共现的字词作为错别字的纠正候选。Fu 等[11]将含有错别字的语句“翻译”成语法正确的语句，并从翻译后的句子中找出错别字纠正的结果。

上述方法大多需要构建大规模的混淆集和错词典，并基于大规模语料训练语言模型和分类器，总体上效率低下，而且性能不佳。另外，词切分是一项基础工作。然而，词切分本身带来的错误，会对错别字的检测和纠正带来非常大的影响。

3 算法基本流程

图 1 是 CPLM-CSC 错别字纠正的基本流程。

图 1　CPLM-CSC 错别字检测和纠正的基本流程

Fig.1　The workflow of Chinese spelling checking and correction in CPLM-CSC

模型的输入是待检查的中文段落（可包含多个语句），而输出是经过错别字纠正之后的段落。流程中间三个步骤的简单介绍如下：

1) 词性标注和长度一致化调整

在训练和预测阶段，对输入模型的原始中文语句，需要进行如下的词性标注和长度一致化调整。

（a） 单字级别词性标注。词性特征对有些错别字的检测非常有用，例如：“的”字之后的词一般是名词，如果其后是动词的话，则“的”字用错了。由于输入 CPLM-CSC 的是字序列，即输入序列中每个单元是一个字。对字序列进行词性标注的流程如下：

i. 对语句进行分词并标注词性；

ii. 基于每个词的词性，按照 BIES 方式给语句中的每个字标注词性。BIES 的含义如下：

- B：词的起始字
- I：词的中间字
- E：词的结尾字
- S：单字词

例如：“哈尔滨”的词性为地名（ns），其中每个字的词性为：哈（ns-B），尔（ns-I），滨（ns-E）。

（b） 长度一致化调整。过短的语句很可能没有包含完整的文本信息，而过长的语句在模型中容易发生信息丢失。我们设计了基于规则的文本切割方法，并通过实验设置了语句长度范围，效果明显优于简单地按照终结符进行语句切分。例如，设置语句的长度范围为 15~30 字。对于长度大于 30 字的语句，则截取 30 字以内的分句。对于长度小于 15 的句子，则根据情况拼到前后句子中。具体的算法如下。

ALGORITHM 1: sentence segmentation

Input: list of sentences S, upper limit max
Output: list of segmented sentences $\mathscr{H}$

$\mathscr{H} \leftarrow []$
for s in S **do**
  if length(s) > max **then**
    $L \leftarrow$ split s by (‘,’ | ‘;’ | ‘--’ | ‘:’)
    for s' in L **do**
      if length(s') > max **then**
        $L' \leftarrow$ split s' by “、”
        append every element in L' to $\mathscr{H}$
      else
        append s' to $\mathscr{H}$
      end
    end
  else
    append s to $\mathscr{H}$
  end
end

ALGORITHM 2: sentence merging

Input: list of sentences $\mathscr{H}$, lower threshold min
Output: list of merged sentences $\mathscr{R}$

$\mathscr{R} \leftarrow []$
while $\mathscr{H}$ not empty **do**
  $s_1 \leftarrow$ pop first sentence of $\mathscr{H}$
  while length(s_1) < min **do**
    $s_2 \leftarrow$ pop first sentence of $\mathscr{H}$
    $s_3 \leftarrow$ pop first sentence of $\mathscr{R}$
    if length(s_2) < length(s_3) **or**
      s_3 ends with {‘。’, ‘!’, ‘?’, ‘…’} **do**
      $s_1 \leftarrow s_1 + s_2$
      append s_3 to $\mathscr{R}$
    else
      $s_1 \leftarrow s_3 + s_1$
      push s_2 back to $\mathscr{H}$
    end
  end
  append s_1 to $\mathscr{R}$
end

2) 基于预训练语言模型错别字检测

CPLM-CSC 的错别字检测模块基于单字级别预训练语言模型来设计（详细描述见 4.1 节）。单字级别预训练语言模型（例如：BERT[12]，ERNIE[13]），接受字序列作为输入。由于预训练语言模型是运用大规模语料训练的，因此 CPLM-CSC 的错别字检测模块不需要大量语料来训练，只需要少量数据进行模型微调。

3) 基于掩字语言模型错别字纠正

CPLM-CSC 的错别字纠正模块依然基于预训练语言模型设计以减少训练语料（详细描述见 4.2 节）。在输入的语句中，由错别字检测模块识别出来的错别字被掩去，并在输出端产生错别字的候选纠正结果。针对现有方法误报较多的问题，为了提升错别字纠正性能，CPLM-CSC 采取概率阈值、排名阈值、音近形近判断等方式，以筛除不合适的纠正结果，并选择最优的结果输出。

4 算法的具体步骤

4.1 基于单字级别预训练语言模型的错别字检测

图 2 显示的是 CPLM-CSC 错别字检测的基本框架。输入模型的是中文语句字序列，以及单字级别词性序列。以 x 表示语句中的一个汉字，pos(x)表示它的词性标注，CPLM 表示单字级别预训练语言模型，POSM 表示词性编码矩阵。

x 经过 CPLM 编码，输出的结果为：

$$\mathrm{v}'(x) = \mathrm{CPLM}(x) \tag{1}$$

输出结果$\mathrm{v}'(x)$是结合深度上下文信息的字表征，维度为 1*768。

pos(x)是以 one_hot 模式表示的 x 的词性，维度为 1*144（36 种词性标签，以及 B/I/E/S 这四种单字级别标签）。pos(x)经过词性编码矩阵的输出为：

$$\mathrm{pos}'(x) = \mathrm{POSM}(\mathrm{pos}(x)) \tag{2}$$

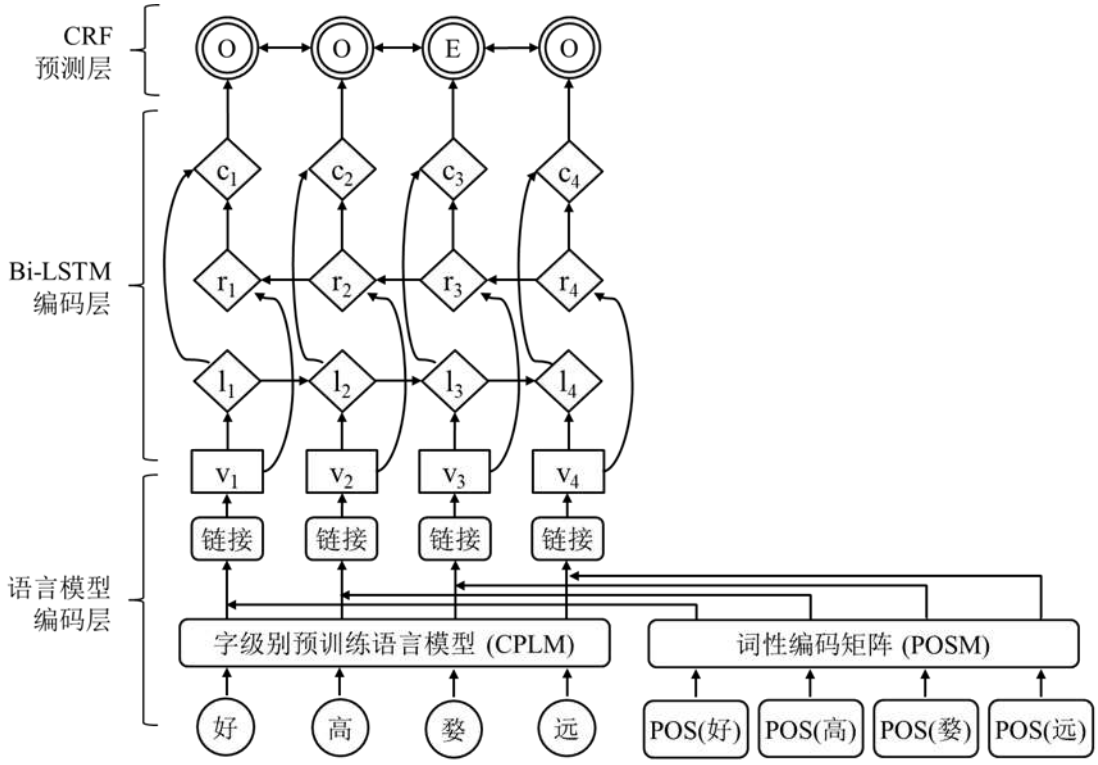

图 2　CPLM-CSC 错别字检测模块的基本框架

Fig.2　The framework of Chinese spelling checking in CPLM-CSC

POSM 是一个 144*72 的矩阵，在训练之前随机初始化，并在训练后确定矩阵元素值。$\mathrm{pos}'(x)$是维度为 1*72 的向量。字编码和词性编码在链接之后（见公式（3））作为语言模型编码层的输出，并输入到 Bi-LSTM 层。

$$\mathrm{v}(x) = \mathrm{catenation}(\mathrm{v}'(\mathrm{x}), \mathrm{pos}'(\mathrm{x})) \tag{3}$$

$\mathrm{v}(x)$是维度为 1*840 的向量。经过 Bi-LSTM 层，输出为：

$$\mathrm{w}(x) = \mathrm{BLSTM}(\mathrm{v}(x)) \tag{4}$$

$\mathrm{w}(x)$的维度为 256，输入到 CRF 预测层，以给出每个汉字对应的标签。标签为‘E’（表示错别字）或者‘O’（表示正确字）。CRF 预测层会计算每个可能的标签序列的概率，并将概率最大的序列作为输出。标签序列$L^i = \{l_1^i, \ldots, l_K^i\}$的概率计算如公式（5）。$K$ 表示输入汉字序列的长度，l_m^i表示第 m 个字的标签。

$$\mathrm{P}(L^i) = \sum_{k=1}^{K} \{\mathbb{H}_{(l_{k-1}^i, l_k^i)} + \varphi(l_k^i, k)\} \tag{5}$$

$\mathbb{H}$是表达标签之间转移概率的矩阵，由于标签只有‘E’和‘O’，因此$\mathbb{H}$是 2*2 的矩阵。

$\mathbb{H}_{(l_{k-1}^i, l_k^i)}$表示$l_{k-1}^i$转移至$l_k^i$的概率。$\varphi(l_k^i, k)$表示语句中第 k 个元素被标注为l_k^i的得分。φ是 2*V 的矩阵，其中 V 是字符集（包含汉字，标点及其他字符）的基数，2 是标签集（O 和 E）的大小。$\mathbb{H}$和φ的元素值随机初始化生成，并在训练后确定。

基于错别字检测模块的输出，标注为‘E’的汉字被认为是可能的错字，需要在下一步进行纠正。而标注为‘O’的字，不需要进行处理。

4.2 基于掩字语言模型的错别字纠错模块

CPLM-CSC 采用掩字模型进行错别字的纠正。错别字纠正模块的框架见图 3。

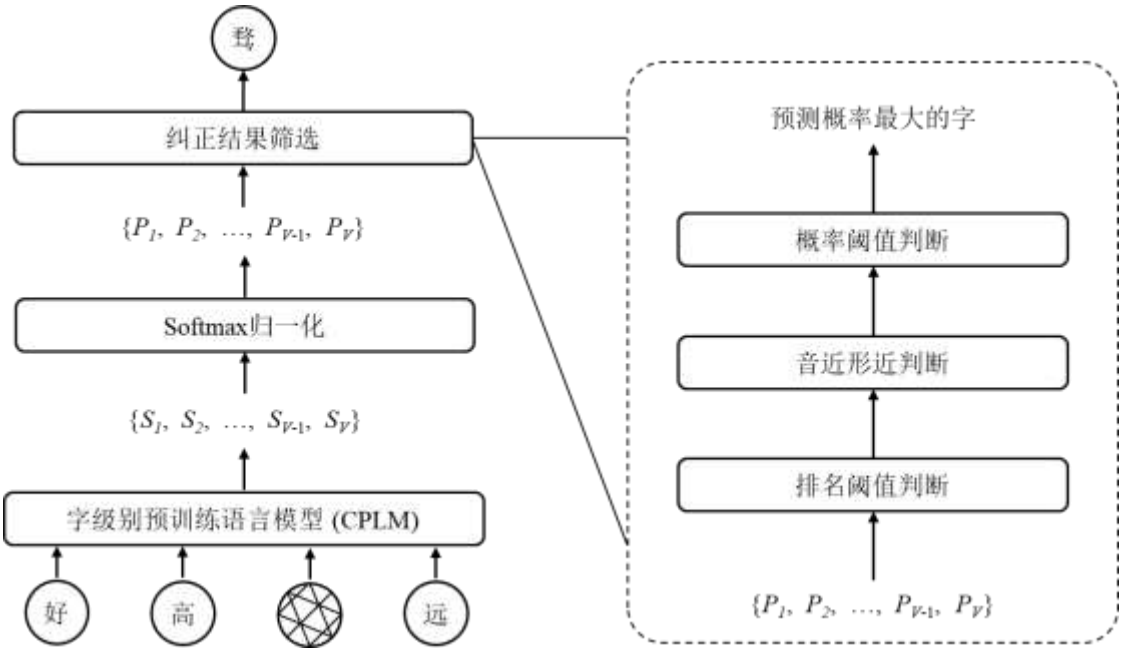

图 3　CPLM-CSC 错别字纠正模块的基本框架

Fig.3　The framework of Chinese spelling correction in CPLM-CSC

模型的输入是一个中文语句，在错别字检测阶段中被识别出来的错别字被掩盖。经过 CPLM 计算之后，输出为一个维度为 1*V 的向量{P_1, P_2, …, $P_{V\text{-}1}$, P_V}（V 是字符集的基数），向量中的元素值表示相应字符是正确字的概率（称为正确字置信度）。可以将置信度最高的字作为模型输出，但是效果并不理想。为了提升最终输出结果的准确率，本文采用了以下三种方法来处理正确字置信度向量，以筛选最优的输出结果。

1) 排名阈值判断。如果被掩盖的可能错字（例如图 3 中的“婺”）的正确字置信度在所有字符的正确字置信度的排名大于一定阈值（例如：100），说明它很有可能是正确的用法。这样，它被选为最终输出，即该字不做纠正。
2) 音近形近判断。由于绝大多数错别字出现在音近字或形近字之间，因此首先排除音不近及形不近的候选字。基于公开的音近形近字混淆集（Confusion Set），按照正确字置信度从大到小的顺序，筛掉与疑似错字（即被掩盖的字）音不近且形不近的字，直到出现一个音相近或者形相近的字。
3) 概率阈值判断。在音近形近判断之后，如果剩余的字符里面正确字置信度最大者（即置信度最大的音近字或形近字）的置信度大于一定阈值，则将该字输出。否则，由于该字的置信度不高，纠正的把握不大，因此不做纠正，输出（被掩盖的）疑似错字。

基于以上处理，最终的输出结果可能是原来的字，即不做修改，或者是满足排名阈值和

概率阈值，与识别结果读音相近或者字形相近，而且正确字置信度最大的字。

错别字纠错模块的参数不需要经过训练来确定。掩字语言模型直接采用公开的、已经训练过的预训练预研模型，而概率阈值和排名阈值则由人工来确定和更新。

5 实验与分析

5.1 评测数据集

实验采用的数据是 SIGHAN 2015 数据集[3]，其中的数据采集自初学汉语的外国人写的作文，并由汉语母语者进行标注。SIGHAN 2015 数据集包含词错误和非词错误两种类型，分成以下两个部分。

1) 训练集：包含 970 篇中文作文，3143 个错别字错误。
2) 测试集：包含 1100 个中文段落，其中一半至少有一个错别字错误，而另一半没有错误。

SIGHAN 评测规定，参赛者可以使用任意的语言和计算资源。大多数参赛者使用了往年的 SIGHAN 评测的数据集来训练模型。为使得对比公平，我们的实验使用了 SIGHAN 2013[2] 和 SIGHAN 2014[14]的训练集。SIGHAN 2013 的数据是从中学生的作文中收集，错误类型较为单一，只有非词错误，共有 2000 个语句，1621 个错别字错误。SIGHAN 2014 的数据也是从外国人的作文中收集，包含非词错误，单字词错误、和双字词错误，共有 1363 篇作文，6076 个错别字错误。

结果的评测分成以下两个方面。

1) 错别字检测: 语句中的错别字被正确地识别出来。
2) 错别字纠正: 语句中的错别字被正确地识别并且纠正。

错别字纠正结果采用准确率（Acc.），精确率（Pre.），误报率（FPR）和召回率（Rec.）进行评价。按照 SIGHAN 规定，TP 表示错别字被正确纠正的次数；FP 表示非错别字被误报为错别字的次数；TN 表示不含错别字而且没有被误报的语句数量；FN 表示错别字未能被正确检测的次数。性能评测指标的计算方式如下。

- $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$
- $\mathrm{Acc.} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})$
- $\mathrm{Pre.} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$
- $\mathrm{Rec.} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$
- $\mathrm{F1} = 2 * \mathrm{Pre.} * \mathrm{Rec.}/(\mathrm{Pre.} + \mathrm{Rec.})$

针对一些特殊的错误类型，实验采用了数据增强的方法来扩充训练数据集。例如：把数据集里“的”、“地”和“得”进行随机替换，将“在”和“再”进行随机替换等。

5.2 性能分析

CPLM-CSC 以及其他模型的错别字检测和纠正性能见表 1。实验中，CPLM-CSC 里采用的预训练语言模型是 BERT[12]，排名阈值是 150，正确字置信度是 0.5。表 1 中，CAS-Run1 和 CAS-Run2 是 SIGHAN 2015 评测中结果最好的两个模型。同时，为了更加全面地分析模型，我们对 CPLM-CSC 模型进行了多种修改，分别进行实验，并将结果列举在表 1 当中。

表 1　各种模型的错别字检测和纠正性能

Table 1　Performance of various models for Chinese spelling checking and correction

方法		FPR	错别字检测				错别字纠正			
			Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
CAS-Run1		**0.1164**	0.6891	**0.8095**	0.4945	0.614	0.68	**0.8037**	0.4764	0.5982
CAS-Run2		0.1309	**0.7009**	0.8027	0.5327	0.6404	**0.6918**	0.7972	0.5145	0.6254
CPLM-CSC		0.28	0.6864	0.6998	**0.6527**	**0.6754**	0.6709	0.6895	**0.6219**	**0.654**
w/o 预训练语言模型		0.3145	0.6464	0.6588	0.6072	0.6319	0.6245	0.6418	0.5636	0.6002
w/o 词性特征		0.2964	0.6536	0.6707	0.6036	0.6354	0.6418	0.6618	0.58	0.6182
w/o 概率/排名/音形判断 [1]		0.28	0.6864	0.6998	0.6527	0.6754	0.5982	0.6298	0.524	0.572
不同阈值	概率阈值(0.4)	0.28	0.6864	0.6998	0.6527	0.6754	0.65	0.6744	0.58	0.6236
	排名阈值(100)	0.28	0.6864	0.6998	0.6527	0.6754	0.6545	0.6778	0.5891	0.6303

1.　由于概率/排名/音形判断只作用在错别字纠正阶段，因此最后三行里的错别字检测的性能与最终模型没有区别。

我们对结果进行了深入分析，并统计了各种类型错误的纠正效果。对于某一种类型错误，例如“非词错误”，我们首先计算含有“非词错误”的语句数量，然后用前文所述的各项指标的计算方法来计算纠正性能。需要注意的是，有些语句包含多种类型错误，它们会在不同类型错误的纠正效果的统计当中被重复计算。统计结果见表 2。

表 2　CPLM-CSC 针对不同错误类型纠正的性能分析

Table 2　Performance of CPLM-CSC for different spelling error types

方法	FPR	错别字检测				错别字纠正			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
非词错误	0.3411	0.7318	0.599	0.854	0.7041	0.71	0.5819	0.7956	0.6722
词错误	0.2488	0.7155	0.4215	0.5977	0.4944	0.7045	0.4017	0.5508	0.4646
单字词错误 [1]	0.1871	0.7809	0.4625	0.6514	0.5409	0.7709	0.4426	0.6009	0.5097
“的地得”误用	0.0971	0.8882	0.5463	0.7887	0.6455	0.8782	0.5206	0.7113	0.6012

1.　单字词错误是词错误的一种，而“的地得”错误是单字词错误的一种。

从上表可以看出，非词错误的纠正效果明显好于词错误纠正，而采用了数据增强的“的地得”误用的纠正效果比单字词错误的平均纠正效果要好。从这些结果中，可以看出：（1）词错误依然是错别字纠正的难点；（2）训练数据对纠正效果有较大的影响。

6 总结

本文针对中文错别字的检测和纠正问题，提出了一种基于预训练语言模型的错别字纠正方法，以利用从大规模语料学习到的语法和语义知识，节省训练语料规模，并提高纠正的性能。本文提出的方法采用单字级别语言模型，可以消除中文分词错误对错别字识别的影响。另外，在错别字纠正阶段，本方法采用基于掩字方式构建的语言模型，并利用上下文信息计算出错别字的纠正候选字。在 SIGHAN 2015 评测数据集上进行测试，本文提出的方法取得了 0.654 的 F1 值，性能明显优于其他模型。

基于测试结果的分析，本方法在非词错误，单字词错误（特别是“的地得”误用，“再在”误用等）等错误上表现较好，但是在词错误上表现不好。由于词错误经常发生在一些易混淆词之间，需要结合上下文语义分析进行判断，因此它一直是错别字检测的难点。另外，本方

法的误报和漏报情况依然很多，如果借助一些特殊的数据和手段，例如错别字集和专业词汇集，可以降低误报率和提高召回率。

针对上述的问题，以下列举了一些可以改进模型性能的方法。

1) 针对易混淆词之间的误用情况，依靠语言模型无法得到很好地解决。可以设计出更长距离依赖的模型，以解决这种需要进行深度上下文语义分析才能识别的错误。
2) 借助混淆集来查找音近字和形近字，存在效率低和准确率低的缺点，毕竟混淆集包含的情况有限。设计算法来评估两个字的音近和形近程度，并给出相应得分，可以更加有效地辅助错别字纠正。
3) 由于错别字的类型较多，针对每个类型的错误，采用单独的训练数据，设计独特的数据增强方法，甚至设计单独的纠正模型，可以显著地提高纠错性能。不过这里需要大量细致的工作。
4) 有些字词的用法有固定的语法规则，需要应用语言学知识才能更好地判断它们的用法是否正确。结合语言学知识进行错别字纠正，也是一个可行的方向。

参考文献

[1] Kuan-Yu Chen, Hung-Shin Lee, Chung-Han Lee, et al. A study of language modeling for Chinese spelling check. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 2013: 79–83

[2] Shih-Hung Wu, Chao-Lin Liu, Lung-Hao Lee. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 2013: 35–42

[3] Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, et al. Introduction to SIGHAN 2015 Bake-off for Chinese Spelling Check. In Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing, 2015: 32–37

[4] Chao-Huang Chang. A new approach for automatic Chinese spelling correction. In Proceedings of the Natural Language Processing Pacific Rim Symposium, 1995: 278–283

[5] Chuan-Jie Lin, Wei-Cheng Chu. NTOU Chinese spelling check system in SIGHAN Bake-off 2013. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 2013: 102–107

[6] Yu-Ming Hsieh, Ming-Hong Bai, Keh-Jiann Chen. Introduction to CKIP Chinese spelling check system for SIGHAN Bakeoff 2013 evaluation. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 2013: 59–63

[7] Ting-Hao Yang, Yu-Lun Hsieh, Yuh-Suan Chen, et al. Sinica-IASL Chinese spelling check system at SIGHAN-7. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 2013: 93–96

[8] Hai Zhao, Deng Cai, Yang Xin, et al. A Hybrid Model for Chinese Spelling Check. ACM Transactions on Asian and Low-Resource Language Information Processing, 2017, 16(3):1-22

[9] Hsun-Wen Chiu, Jian-Cheng Wu, Jason S. Chang. Chinese spelling checker based on statistical machine translation. In Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 2013: 49–53

[10] Yi Yang, Pengjun Xie, Jun Tao, et al. Alibaba at IJCNLP-2017 Task 1: Embedding Grammatical Features into LSTMs for Chinese Grammatical Error Diagnosis Task. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Shared Tasks, 2017: 41–46

[11] Ruiji Fu, Zhengqi Pei, Jiefu Gong, et al. Chinese Grammatical Error Diagnosis using Statistical and Prior Knowledge driven Features with Probabilistic Ensemble Enhancement. In Proceedings of the Fifth Workshop on Natural Language Processing Techniques for Educational Applications, 2018: 52–59

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR abs/1810.04805, 2018

[13] Yu Sun, Shuohuan Wang, Yukun Li, et al. ERNIE: Enhanced Representation through Knowledge Integration. CoRR abs/1904.09223, 2019

[14] Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, et al. Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check. In Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, 2014: 126–132