智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于数据增广和复制的中文语法错误纠正方法

汪权彬， 谭营

引用本文:
汪权彬, 谭营. 基于数据增广和复制的中文语法错误纠正方法[J]. 智能系统学报, 2020, 15(1): 99–106.
WANG Quanbin, TAN Ying. Chinese grammatical error correction method based on data augmentation and copy mechanism[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(1): 99–106.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202001014

DOI: 10.11992/tis.202001014

基于数据增广和复制的中文语法错误纠正方法

汪权彬，谭营
（北京大学 信息科学技术学院，北京 100871）

摘　要：中文作为一种使用很广泛的文字，因其同印欧语系文字的天然差别，使得汉语初学者往往会出现各种各样的语法错误。本文针对初学者在汉语书写中可能出现的错别字、语序错误等，提出一种自动化的语法纠正方法。首先，本文在自注意力模型中引入复制机制，构建新的 C-Transformer 模型。构建从错误文本序列到正确文本序列的文本语法错误纠正模型，其次，在公开数据集的基础上，本文利用序列到序列学习的方式从正确文本学习对应的不同形式的错误文本，并设计基于通顺度、语义和句法度量的错误文本筛选方法；最后，还结合中文象形文字的特点，构造同形、同音词表，按词表映射的方式人工构造错误样本扩充训练数据。实验结果表明，本文的方法能够很好地纠正错别字、语序不当、缺失、冗余等错误，并在中文文本语法错误纠正标准测试集上取得了目前最好的结果。

关键词：自注意力机制；复制机制；序列到序列学习；中文；语法错误纠正；神经网络；文本生成；通顺度
中图分类号：TP389.1　　**文献标志码：**A　　**文章编号：**1673−4785(2020)01−0099−08

中文引用格式：汪权彬，谭营. 基于数据增广和复制的中文语法错误纠正方法 [J]. 智能系统学报，2020，15(1)：99–106.
英文引用格式：WANG Quanbin，TAN Ying. Chinese grammatical error correction method based on data augmentation and copy mechanism[J]. CAAI transactions on intelligent systems，2020，15(1)：99–106.

Chinese grammatical error correction method based on data augmentation and copy mechanism

WANG Quanbin，TAN Ying
(School of Electronics Engineering and Computer Science，Peking University，Beijing 100871，China)

Abstract: Chinese is a widely used language. However，due to its natural difference between Indo-European languages，Chinese learners tend to make various grammatical errors. This article proposes an automatic grammar correction method for those who will make errors like typos or improper words order. First，we built the C-Transformer model that adopts copy mechanism in the self-attention model to translate wrong text sequence to the correct one. Second，based on the public data set，a pure sequence to sequence method is utilized to generate wrong text corresponding to the correct one，and an error text filter is designed based on fluency，semantic，and syntactic measurements. Finally，since Chinese words are pictographic，based on the collected homographs and homophones dictionaries，some error samples are artificially constructed to expand training data. The experimental results show that our method can well correct typos，improper word order，missing，redundancy and other errors，and achieved the state-of-the-art performance on the standard test set of Chinese text grammatical error correction.

Keywords: self-attention mechanism; copy mechanism; sequence to sequence learning; Chinese; grammatical error correction; neural networks; text generation; fluency

收稿日期：2020−01−09.
基金项目：国家重点研发计划资助项目 (2018AAA0100300、2018AAA0102301)；国家重点基础研究发展计划项目 (2015CB352302)；国家自然科学基金项目 (61673025、61375119)；北京市自然科学基金项目 (4162029).
通信作者：谭营. E-mail：ytan@pku.edu.cn.

汉语现在是在全球范围内被广泛使用的语言之一，许多外国人开始学习中文，并在各种场景下使用普通话进行交流。但是汉语因为其独特性，许多以印欧语系为母语的学习者在学习和使

用中文时总会出现各种各样的语法错误，一个自动化的中文文本语法错误纠正系统对于将汉语作为第二语言的学习者，将很好地辅助他们进行汉语学习；此外，这样的自动化纠错系统，对于国内的婴幼儿在学习中文时，也能起到正确的指引作用，从小养成正确的汉语表述和书写习惯，避免经常出现的误用和乱用。基于此，我们在之前英文语法错误纠正的工作基础上，进行了中文语法错误纠正的相关研究，根据中文语法和汉字的特殊性，提出了一种利用同形异义词和同音异义词的错误文本构造方法和基于通顺度、语法和句法相似度的人造数据筛选方案，并将这些构造的错误文本扩充到公开数据集中进行实验，首次在中文文本语法错纠正问题上使用自注意力机制[1]和复制机制[2]，在标准的中文语法错误纠正测试集上取得了最好的结果。

1　文本语法错误纠正

文本语法错误的自动化纠正作为一个充满挑战性的自然语言处理任务，一直以来备受关注。其目的是对人们书写或者其他系统，比如语音识别、机器翻译，产生的文本进行错误的检测和纠正。近几年来，随着深度学习的兴起，文本语法错误纠正从之前基于规则的方法逐步转变为基于学习的方式，传统的先进行错误检测，再对存在错误的文本进行纠正的两段式方案已被取代，纠正阶段也不需要针对不同的错误设计特定的纠正规则，而是将语法错误纠正看作一个错误文本到正确文本的翻译过程，利用深度学习时代自然语言处理领域的经典模型——序列到序列学习进行端到端的文本语法错误纠正。这种方式需要提供有经验的对应语言使用者标注好的训练数据，原始错误文本和对应的正确文本组成成对的一条样本，如表 1 所示。

表 1　训练样本示例

Table 1　Some examples in the training data

模型输入	模型输出
我在家里一个人学习中文。	我在家里自学中文。
一直下雨而点儿冷。	一直下雨而且有点儿冷。 一直下雨所以有点儿冷。
我的小狗正在出牙。	我的小狗正在长牙。

2　相关工作

因为英语作为被全世界广泛使用的第二语言，针对英文文本语法错误纠正的研究工作相对于中文文本语法错误纠正要多很多。最早基于规则的语法错误纠正工作可以追朔到 20 世纪 80 年代，Macdonald 等[3]提出了一种自动化语法错误检查工具，依据预先设计的规则进行错误匹配和纠正。但是这种基于规则的错误纠正方法的性能很大程度上取决于制定的规则的好坏和文本预处理工具的鲁棒性，所以这一类人工制定规则的方法很快被基于数据自动化抽取规则方法取代，其中比较具有代表性是 1997 年由 Mangu 等[4-5]从 Brown 数据集自动抽取出的自动化拼写错误纠正规则系统。

随着错误文本语料的进一步扩大和计算能力的提升，统计机器学习方法被广泛应用于文本语法错误纠正，比如 Cahill 等[6]在 2013 年提出的基于分类模型的介词错误纠正模型，以及大量在统计机器翻译领域被证明有效的方法的也被大量引入文本语法错误纠正这个任务中[7-8],这一类方法进一步将文本语法错误纠正方法在英文标准测试集上提升到了一个新的高度。随着深度学习近十几年来在人工智能领域的绝对主导地位以及神经机器翻译方法较统计翻译方法在机器翻译任务上的优势，大量研究学者也在探索基于深度神经网络的自动化文本语法错误纠正。其中最早的工作是 2014 年 Xie 等[10]提出的一种基于序列到序列学习的字母级别文本语法错误纠正系统。这一类的方法基本都采用这种序列到序列学习的框架，然后从各个角度去进行改进，以提升模型的效果，比如引入编辑操作信息指导错误纠正操作[11]，通过数据增广和通顺度度量进行选择的迭代型错误纠正方法[12]以及引入与训练任务和解码中引入复制机制的语法错误纠正模型[13]等。

与英文文本语法错误纠正的广泛研究不同，中文文本语法错误纠正的研究工作相对较少，2018 年之前的工作主要聚焦于文本错误的检测[14]，也有一些工作在研究繁体中文的拼写错误，比如 SIGHAN 系列的繁体中文拼写错误检查比赛[15]。2018 年，国际自然语言处理和中文计算会议 (NLPCC) 举办了简体中文语法错误纠正比赛[16]，这次比赛不在局限于繁体中文拼写错误，而是针对所有不同类型的错误进行纠正，并在标准测试集上进行模型性能对比。其中，阿里巴巴的团队提出了一种基于统计机器翻译模型和神经机器翻译模型的层次化混合模型[17];来自有道的 Fu 等人提出一种分阶段纠正方案，他们针对不同的错误类型设计了不同的纠正模型[18]，然后利用集成学习的思想和候选解排序的后处理等方式，取得了

标准测试集上的最好结果; 此外, 不同于其他队伍广泛采用的长短时记忆网络 LSTM(Long Short Term Memory)[19],Ren 等[21] 还提出了一种基于卷积神经网络 CNN(Convolutional Neural Network) 的序列到序列纠正模型, 也取得了非常不错的结果。

3　C-Transformer 模型

本文使用的模型是经典的带注意力机制的序列到序列学习模型[22] 结构,所不同的是, 本文中采用的是目前应用更为广泛的基于自注意力机制的 Transformer 模型[1], 自注意力机制已经在很多场景下被证明能取得比文献 [22] 中采用的软注意力机制更好的性能, 基于此的 Transformer 模型也因其天然的并行性, 在自然语言处理任务中被广泛使用。

此外, 鉴于文本语法错误纠正这个任务与机器翻译等其他任务的特异性, 需要纠正的只是整个文本很少的几个词, 其余大部分的文本直接复制到目标文本中即可, 所以与文献 [12] 中的工作类似, 本文在中文文本语法错误纠正任务中引入复制机制[2],由模型判断直接从原始文本中复制还是从词表空间生成。C-Transformer(Copy Transformer) 模型结构如图 1 所示。

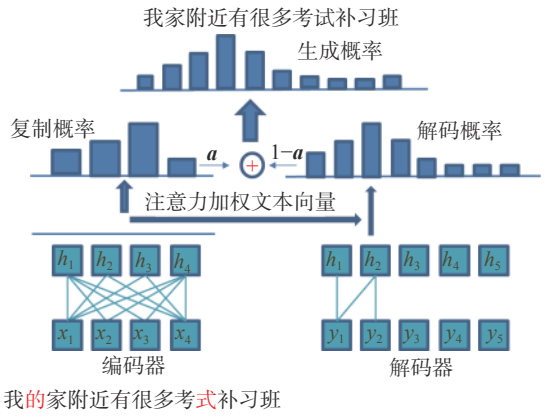

图 1　C-Transformer 模型结构

Fig. 1　The model architecture of C-Transformer

3.1　序列到序列学习

序列到序列学习框架在包括机器翻译、自动问答等不同的自然语言处理任务上取得了巨大的成功。给定一段输入文本 $\boldsymbol{x}$, 序列到序列学习模型首先对输入进行编码, 学习输入文本的向量化表示, 然后基于这个输入的向量化表示解码其对应的目标输出文本 $\boldsymbol{y}$, 这种学习通常利用成对的训练数据去最大化对数似然来实现的, 如式 (1):

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in T} \log p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) \tag{1}$$

式中: T 表示训练数据集, 本文中指错误−正确文本对; $\boldsymbol{\theta}$ 表示采用的模型的参数, 本文中指使用的自注意力机制模型的参数。

训练好模型之后, 对于给定的一段未知是否存在错误的文本, 序列到序列学习的模型通过自回归的解码方式生成累积概率最大候选纠正文本, 如式 (2):

$$p(\boldsymbol{y}|\boldsymbol{x};\widehat{\boldsymbol{\theta}}) = \prod_{i=1}^{L} p(y_i|\boldsymbol{x},\boldsymbol{y}<i;\widehat{\boldsymbol{\theta}}) \tag{2}$$

3.2　Transformer 模型

以 LSTM 和 GRU[23] 为代表的循环神经网络在机器翻译等序列建模问题上一直都处于统治地位, 2017 年基于自注意力机制的 Transformer 模型[1] 的提出打破了这一垄断, 目前很多地自然语言处理任务的最好结果都是基于这种模型结构。Transformer 不再依赖循环或者卷积的网络结构, 完全依赖于注意力机制进行序列数据的建模, 并能很好的捕捉长距离文本之间的依赖关系; 此外这种结构具备天然的并行性, 可以避免循环神经网络结构的序列前后依赖导致的串行缺点。下面介绍一下 Transformer 中的几个组件:

1) 缩放点积注意力 (scaled dot-product attention) 缩放点积注意力机制的核心如式 (3) 所示:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

式中: $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别表示多个维度为 d_k 的问题、关键词和值向量组合而成的矩阵。这种注意力机制对于一个给定的问题, 需要计算其与所有关键词的点积, 并除以向量维度以消除维度过高带来影响, 然后利用这个点积值作为对每一关键词对应的值向量的注意力度量, 对值向量进行加权求和, 得到最终的基于注意力权重的向量化表示。

对应到自然语言处理中的序列建模, 以本文的文本错误纠正问题为例, 在编码时 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 对应的都是同一批训练样本中的可能存在错误的输入文本, 计算每个样本的每一个词向量与本身其他所有词之间的点积值, 并作为权重去计算每一个词的注意力加权的向量化表示, 这也是自注意力机制中这个“自”的意思, 每一个词在学习其表示时都需要注意到所在文本的其他词的含义。

2) 多头注意力 (multi-head attention)

文献 [1] 中的实验结果表明, 与进行单一的缩放点积相比, 将问题、关键词和值向量分别进行多次映射之后分别进行缩放点积注意力计算之后, 再拼接起来映射回原始维度的效果会更好, 如式 (4) 所示;

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\text{head}_2,\cdots,\text{head}_h)\boldsymbol{W}^O$$
$$\text{head}_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{4}$$

式中：$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$ 和 $\boldsymbol{W}^O$ 都表示可学习的权重矩阵；h 表示“头”的数量，实际使用时需要注意这些矩阵的维度。值得说明的是，虽然引入多头机制的缩放点积注意力机制，但可以通过维度变化，使得这些操作实际的计算复杂度与单一缩放点积注意力计算是一致的。

3) 编码 (Positional Encoding)

因为 Transformer 中没有对序列的顺序依赖和前后关系的建模，为了能让模型对利用序列中词的先后关系的信息，在计算词向量时引入了位置信息编码，将序列中词的相对位置和绝对位置信息进行编码表示，如式 (5) 所示：

$$\begin{aligned}\text{PE}(\text{pos},2i)&=\sin(\text{pos}/10\,000^{2i/d})\\ \text{PE}(\text{pos},2i+1)&=\cos(\text{pos}/10\,000^{2i/d})\end{aligned} \tag{5}$$

式中：pos 表示词的绝对位置；i 表示位置向量维度 d 中的第几位；分别用正弦函数和余弦函数计算位置向量的奇数位和偶数位的值。

实验表明，这种方式的位置信息编码和学习得到的位置信息编码取得的效果基本一致。

3.3　复制机制

复制机制现在也是在自然语言处理问题中被广泛使用的一种方法，在文本摘要[24]和语义解析[25]等问题上都被证明了其有效性。文献 [13] 首次将复制机制应用到了英文文本的语法错误纠正问题，并在句子级别的子任务预训练上对复制机制进行与训练，取得了非常好的效果。基于此，本文将复制机制首次应用在中文文本错误纠正问题上，带复制机制的正确文本解码过程如式 (6) 所示：

$$p_t(w)=(1-a_t)\times p_t^*(w)+a_t\times p_t'(w) \tag{6}$$

式中：$p_t^*(w)$ 表示 t 时刻利用常规注意力机制的词表解码生成概率；a_t 表示从词表生成和从原文复制的一个平衡因子，计算方式见式 (7)：

$$a_t=\text{sigmoid}(\boldsymbol{W}^{\mathrm{T}}\sum((\boldsymbol{q}_t^{\mathrm{T}}\boldsymbol{K})^{\mathrm{T}}\cdot\boldsymbol{V})) \tag{7}$$

$p_t'(w)$ 为 t 时刻解码的词是从原文复制的概率，从公式 8 计算得到：

$$p_t'(w)=\text{softmax}(\boldsymbol{q}_t^{\mathrm{T}}\boldsymbol{K}) \tag{8}$$

式中：$q_t,\boldsymbol{K},\boldsymbol{V}=h_t\boldsymbol{W}_q^{\mathrm{T}},\boldsymbol{H}^s\boldsymbol{W}_k^{\mathrm{T}},\boldsymbol{H}^s\boldsymbol{W}_v^{\mathrm{T}}$，分别表示对 t 时刻解码隐状态的映射和对输入编码隐状态的两种不同变换。每个 $\boldsymbol{W}$ 都表示各种情况下可学习的参数矩阵。

整个模型的训练过程就是最大化式 (1) 的似然函数，式 (1) 的 p 就对应这里 $p_t(w)$ 对所有时刻的累计。

4　数据增广

4.1　数据集

实验所使用的数据集是在 NLPCC 比赛公开的数据集基础上进行扩充得到的，原始数据的统计属性如表 2 所示。

表 2　实验数据统计信息

Table 2　The statistical information of the data

数据	类别	数量	最大长度	最小长度	平均长度	不同字数
训练集	原始S	651 336	212	1	17.80	7 782
	原始T		213	1	18.20	7 721
	生成1S	1 608 754	36	2	16.26	7 782
	生成1T		36	2	16.67	7 721
	生成2S	1 552 687	36	2	16.26	8 506
	生成2T		36	2	16.67	7 721
测试集	S	2 000	247	7	29.66	2 214

之前英文文本语法错误纠正工作所获得的经验表明，对于这个任务数据集的扩充带来的纠正效果的提升会非常显著，这是因为错误文本中的错误非常稀疏，给定的一份训练语料中，错误出现的概率的只有 10% 左右，但是错误的种类非常多，所以通过一些方式扩充数据，使得训练数据中能出现更多的错误类型，能在一定程度上提升模型的纠正能力。

针对中文这种象形文字的特殊性，本文不仅仅使用了正确到错误的逆序生成方式生成错误文本，而且还整理了一份同形异义和同音异义词表；并基于这个映射词表，进行基于规则的错误文本生成。扩充的训练数据部分统计信息见表 2。

4.2　正确到错误逆序生成

类似于本文解决的文本错误纠正任务，本文同样采用序列到序列学习模型，进行正确文本到

错误文本的生成过程的这种方法在机器翻译中非常常见, 被称为反向翻译; 可以很容易地应用到错误文本的生成中, 利用错误–正确文本对, 直接训练一个正确文本到存在错误的文本的逆向生成模型, 然后利用这个模型通过大量的正确文本去生成大量的存在错误的文本。这里判断生成的文本是否存在错误, 采用的是文献 [12] 中提出的通顺度度量方案, 利用大规模的正确文本语料预训练了一个语言模型, 计算文本在这个语言模型的生成概率作为通顺度度量。

此外, 为了保证生成的错误文本与原始正确文本的相关性, 还利用语义相似性和编辑距离两种方式对生成的错误文本进行进一步综合筛选。其中语义相似性的度量, 本文采用的是一个在大规模正确文本语料上预训练的 3 层 LSTM 网络模型, 通过这个模型进行句子级别语义向量计算, 取最后一个时刻的模型输出作为输入文本的句子向量, 并计算对用正确–错误文本对的语义向量的余弦相似性。

最后对通过正确到错误逆序生成的所有错误文本进行筛选, 保留通顺度低于原始正确文本且语义相似度大于等于 0.9, 编辑距离小于 5 的“错误”文本, 与原始正确文本构成成对训练数据, 扩充到训练数据中。

4.3　词表映射生成

针对中文象形文字的特点, 本文还提出了一种新型的错误文本生成方案——基于同形异义词和同音异义词表[26] 的错误文本生成, 主要针对中文中特别容易出现的错字别字类型的错误。需要指出的是, 文献 [26] 中提供的相似字集合是繁体的, 而本文针对的是简体中文错误纠正, 所以利用 OpenCC[27] 这个开源工具对这个集合进行转化, 集合中部分示例如表 3 所示。

表 3　混淆集中的同音或同形字示例

Table 3　Some examples of the confusion set with similar pronunciation or shape

原词	同音/同形词
兄	凶汹匈胸熊雄讻芎
呐	纳那钠讷娜捺那拿哪蜡剌腊辣落
己	已忌厄泛妃改杞凹卮犯危
甲	匣押呷柙胛钾申狎伸坤呻

在词表映射错误文本生成的过程中, 对每一句正确文本, 随机地将其中的 1~3 个字替换成其对应的随机一个同形异义或同音异义字, 若选中的字不存在相似字, 则不进行替换, 这样生成的文本与原始正确文本构成一个新的成对样本添加到训练数据中。为了保证多样性, 这个生成过程重复 5 次, 最后对所有生成的文本对进行去重, 构成词表映射生成的扩充训练数据。

本文最终使用的训练数据的统计信息见表 2, 其中原始表示公开数据集, 生成 1 表示正确到错误的序列到序列逆序生成方法, 生成 2 表示基于相似字的词表映射生成方式, S 和 T 分别表示模型输入错误文本和需要的正确输出文本, 我们只保留文本长度大于 2, 小于等于 36, 错误–正确文本对的编辑距离小于等于 5 的文本对构成最终的数据集。

5　实验结果与分析

5.1　实验参数设置

实验所用的 Transformer 模型的编码和解码都为 6 层, “头”的数量为 8, 词向量维度为 512, 序列最大长度为 36, 全连接层节点数为 4 096, 初始学习率为 0.001, 在连续 2 代性能无改进时按 0.95 的比率进行线性衰减, 连续 6 代无改进则提前终止训练, 保留在验证集上最好的模型用于测试, 不同实验配置下, 随着训练的进行, 验证集上的损失函数变化曲线如图 2 所示。

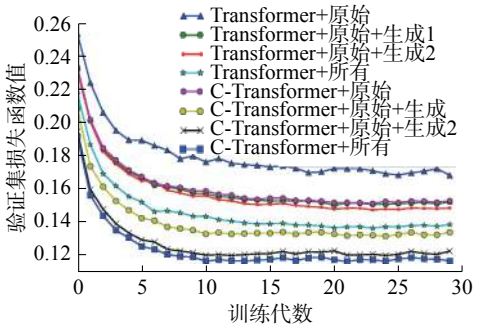

图 2　不同配置下模型验证集损失函数值

Fig. 2　The loss on validation set with different settings

5.2　评价指标

文本语法错误纠正的标准评价指标是最大匹配分数 (M^2-Scorer), 计算的是模型的输出修改和标准修改在字、词或短语级别的最大覆盖, 本文采用的是字粒度的度量, 对模型的输出 和标准的修改, 按式 (9) 计算精确度、召回率和 $F_{0.5}$。

$$P=\frac{\sum_{i=1}^{n}|e_i\cap g_i|}{\sum_{i=1}^{n}|e_i|}\qquad R=\frac{\sum_{i=1}^{n}|e_i\cap g_i|}{\sum_{i=1}^{n}|g_i|}$$
$$F_{0.5}=\frac{5PR}{P+4R} \tag{9}$$

式中: $e_i\cap g_i=\{e\in e_i|\exists g\in g_i, e==g\}$; e_i 和 g_i 分别表示

模型输出和标准答案对应的修改的集合。

5.3　实验结果

本文利用图 1 所示的模型，在第 3 章中介绍的数据集上进行了充分的实验，在 2 000 条标准测试集上进行取得标准评价指标结果如表 4 所示。

需要说明的是，为了避免分词工具带来的干扰，所有的实验都是在字粒度上进行的，所以将标准测试集中的人工标注的标签进行了字节处理。此外，对长度大于 36 的输入文本进行了切分，按标点符号切分成长度数段长度小于等于 36 的子串，纠正后再将结果拼接成原始文本。实验结果表明，本文提出的数据增广方式和复制机制能显著提升中文文本语法错误纠正的效果，在标准测试集上取得目前最好的结果。

表 4　实验结果对比表

Table 4　The results of our method on the standard dataset

模型	数据	P	R	$F_{0.5}$
AliGM[17]	原始	**41.00**	13.75	29.36
YouDao[18]	原始	35.24	18.64	29.91
Transformer	原始	32.27	18.58	27.65
	原始+生成1	34.76	21.55	30.96
	原始+生成2	34.92	21.88	31.20
	所有	35.03	**23.87**	32.03
C-Transformer	原始	36.49	19.15	30.89
	原始+生成1	37.09	21.57	32.42
	原始+生成2	37.34	22.74	33.09
	所有	**38.22**	23.72	**34.05**

5.4　分析与讨论

从表 4 的实验结果可以看出，单一的 Transformer 模型在同样的数据上效果无法与集成的模型或者多模型混合相比，但在补充了人造数据之后，无论是基于序列生成的数据扩充还是基于词表映射生成扩充，在测试集上的效果均好于目前最好的结果，两种数据扩充方式分别提升+1.05、+1.29，且扩充数据带来的影响要大于复制机制的影响。一般情况下，在文本纠错领域，利用数据扩充往往能带来比模型改进更明显的效果提升，这一结论与英文文本错误纠正一致。

进一步可以发现，不管是否使用复制机制，基于中文象形文字特点而提出基于同形异义/同音异义词表映射的错误文本生成构造的样本，能带来比通用错误文本生成更明显的性能提升。在使用复制机制之后，在数据扩充之后，性能得到更进一步的提升，最好的结果为 34.05，较基准提升 4.14。这一方面是因为更多的数据使得模型的学习更加充分，更能表征不同错误到正确的映射关系；另一方面，人造错误样本引入了更多的错误类型，模型在训练时能感知到更多的错误形式，提升了泛化性能。

此外，模型性能的提升主要在召回率指标上，精确度没有明显的改善,低于最高的精度值。这是因为人造数据中出现的错误会比正常样本中错误的占比高，使得模型更倾向于对文本进行修改，所以显著提升了召回率；而复制机制使得模型学习到在某大部分时候从输入直接复制文本到目标文本，能一定程度上缓解这种过多修改的现象，因为对于错误文本而言，一般倾向于保守修改，以免引入不必要的错误，这也是 $F_{0.5}$ 这个指标的初衷。

最后，从表 5 中的一些修改示例中，能看出我们的方法能处理中文文本常见同形/同音错别字、语序、缺失、冗余、连接词等错误都具备比较好的纠正能力。

表 5　模型修改样本示例

Table 5　Some examples of the corrected sentences

序号	模型输入	模型输出
1	我**的**家附近有很多考**式**补习班。	我家附近有很多考**试**补习班。
2	她已经**毕业了大学**，专业是视觉设计。	她已经**大学毕业了**，专业是视觉设计。
3	每星期二，老师**让**我们布置作业。	每**个**星期二，老师**都会给**我们布置作业。
4	我刚洗完**了**米**而**把米放在电饭锅。	我刚洗完米**，**把米放在电饭锅里。
5	诊断**化**谢异常综合征	诊断**代**谢异常综合征
6	中医诊断:项痹病，气滞血**淤**	中医诊断:项痹病，气滞血**瘀**
7	在中国，不**官**是谁，都抽烟。	在中国，不**管**是谁，都抽烟。
8	对我们国家来说，帮挨饿的人是当然**做**的事情。	对我们国家来说，帮挨饿的人是当然**要做**的事情。

续表 5

序号	模型输入	模型输出
9	随着生活的提高, 人们的要求也越来越**提**高。	随着生活**水平**的提高, 人们的要求也越来越高。
10	这种材料表示吸烟引起了**人们**多么大的损害。	这种材料表明吸烟引起了多么大的损害。
11	父母**也**一般在看电视很少说话。	父母一般在看电视**时也**很少说话。
12	为了说好中文, 经常找路人聊天	为了说好中文, **我**经常找路人聊天
13	农药会有害于人的健康, **甚至**孕妇吃了以后会生产不正常的孩子。	农药会有害于人的健康, 孕妇吃了以后**甚至**会生产不正常的孩子。
14	随着中国经济突飞猛**近**, 建造工业与日俱增。	随着中国经济突飞猛**进**, 建造工业与日俱增。
15	那些空气污染也**没有**助于人**生**的身体**建**康。	那些空气污染也**无**助于人的身体**健**康。
16	这些**问题**其实是不该**范**的, 写文章**再**要细心, 重视细节。	这些**错误**其实是不该**犯**的, 写文章要细心, 重视细节。
17	我把弟弟醒过来后, **一起开始**找答案。	我把弟弟**叫**醒过来后, **开始一起**找答案。
18	但是, 互联网对我们有利还是有弊都**由**我们如何使用。	但是, 互联网对我们有利还是有弊都**在于**我们如何使用它。
19	我**被**交通事故了。	我**遇到**交通事故了。
20	你来北京的时候, 去看长城、故宫和颐和园。	你来北京的时候, **一定要**去看长城、故宫和颐和园。
21	他们**特别热情对我**, 因为我爸爸是老师。	他们**对我特别热情**, 因为我爸爸是老师。
22	大家都很**满足**我们的活动。	大家都很**喜欢**我们**办**的活动。
23	没有小孩的**二**个人, 都把她**看成**亲生女儿**对待**。	没有小孩的**两**个人, 都把她**当作**亲生女儿。
24	结果那一天不吃**了午饭**。	结果那一天不吃**午饭了**。
25	问题从**占**菜的时候, 就出现了。	问题从**点**菜的时候, 就出现了。

6 结束语

本文针对中文文本语法错误纠正问题, 从数据扩充和模型改进两个角度进行了研究。首先利用序列到序列学习的方式生成错误样本, 并基于通顺度、语义相似度和编辑距离等指标进行生成数据的筛选; 然后利用中文象形文字的特点, 提出一种基于同音/同形异义词表的映射方法构造错误样本, 这两种方式扩充训练数据集之后, 基准模型的性能有显著提升。此外, 因为错误在自然文本中出现的比例较低, 纠正过程中大部分的文本只需要从原文本中直接复制即可这一特点, 首次在中文文本语法错误纠正中引入复制机制, 进一步提升了模型的效果。我们还发现, 复制机制对由于生成的语料中错误占比较高, 使得模型倾向于进行更多的修改, 导致误报率较高的问题, 也有一定的矫正能力, 保证召回率的同时提升精确度。

参考文献:

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998–6008.

[2] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, USA, 2015: 2692–2700.

[3] MACDONALD N, FRASE L, GINGRICH P, et al. The writer's workbench: computer aids for text analysis[J]. IEEE transactions on communications, 1982, 30(1): 105–110.

[4] FRANCIS W N, KUCERA H. A standard corpus of present-day edited American English, for use with digital computers[R]. Providence, RI: Department of Linguistics, Brown University, 1979.

[5] MANGU L, BRILL E. Automatic rule acquisition for spelling correction[C]//Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, USA, 1997: 187–194.

[6] CAHILL A, MADNANI N, TETREAULT J, et al. Robust systems for preposition error correction using Wikipedia revisions[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA, 2013: 507–517.

[7] BROCKETT C, DOLAN W B, GAMON M. Correcting ESL errors using phrasal SMT techniques[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for

Computational Linguistics. Stroudsburg, USA, 2006: 249–256.

[8] JUNCZYS-DOWMUNT M, GRUNDKIEWICZ R. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA, 2016: 1546–1556.

[9] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, USA, 2014: 3104–3112.

[10] XIE Z, AVATI A, ARIVAZHAGAN N, et al. Neural language correction with character-based attention[J]. arXiv preprint arXiv: 1603.09727, 2016.

[11] WANG Quanbin, TAN Ying. Automatic grammatical error correction based on edit operations information[C]// Proceedings of 25th International Conference on Neural Information Processing. Siem Reap, Cambodia, 2018: 494–505.

[12] GE Tao, WEI Furu, ZHOU Ming. Fluency boost learning and inference for neural grammatical error correction[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 1055–1065.

[13] ZHAO Wei, WANG Liang, SHEN Kewei, et al. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data[J]. arXiv preprint arXiv: 1903.00138, 2019.

[14] RAO Gaoqi, ZHANG Baolin, XUN Endong, et al. IJCNLP-2017 Task 1: Chinese grammatical error diagnosis[C]//Proceedings of the IJCNLP 2017. Taipei, China, 2017: 1–8.

[15] WU S H, LIU Chaolin, LEE L H. Chinese spelling check evaluation at SIGHAN Bake-off 2013[C]//Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing. Nagoya, Japan, 2013: 35–42.

[16] ZHAO Yuanyuan, JIANG Nan, SUN Weiwei, et al. Overview of the NLPCC 2018 shared task: grammatical error correction[C]//Proceedings of 7th CCF International Conference on Natural Language Processing and Chinese Computing. Hohhot, China, 2018: 439–445.

[17] ZHOU Junpei, LI Chen, LIU Hengyou, et al. Chinese grammatical error correction using statistical and neural models[C]//Proceedings of 7th CCF International Conference on Natural Language Processing and Chinese Computing. Hohhot, China, 2018: 117–128.

[18] FU Kai, HUANG Jin, DUAN Yitao. Youdao's winning solution to the NLPCC-2018 Task 2 challenge: a neural machine translation approach to Chinese grammatical error correction[C]//Proceedings of 7th CCF International Conference on Natural Language Processing and Chinese Computing. Hohhot, China, 2018: 341–350.

[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[20] LECUN Y, BENGIO Y. Convolutional networks for images, speech, and time-series[M]//ARBIB M A. The Handbook of Brain Theory and Neural Networks. Cambridge, USA: MIT Press, 1995: 3361.

[21] REN Honghai, YANG Liner, XUN Endong. A sequence to sequence learning for Chinese grammatical error correction[C]//Proceedings of 7th CCF International Conference on Natural Language Processing and Chinese Computing. Hohhot, China, 2018: 401–410.

[22] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409.0473, 2014.

[23] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv: 1412.3555, 2014.

[24] SEE A, LIU P J, MANNING C D. Get to the point: summarization with pointer-generator networks[J]. arXiv preprint arXiv: 1704.04368, 2017.

[25] JIA R, LIANG P. Data recombination for neural semantic parsing[J]. arXiv preprint arXiv: 1606.03622, 2016.

[26] LIU CHAOLIN, LAI MINHUA, TIEN K W, et al. Visually and phonologically similar characters in incorrect Chinese words: analyses, identification, and applications[J]. ACM transactions on Asian language information processing, 2011, 10(2): 10.

[27] A project for conversion between traditional and simplified Chinese[EB/OL]. [2019-12-20].https://github.com/BYVoid/OpenCC.

作者简介：

汪权彬，博士研究生，主要研究方向为机器学习、深度神经网络、自然语言处理。

谭营，教授，博士生导师，主要研究方向为智能科学、计算智能与群体智能、机器学习、人工神经网络、群体机器人、大数据挖掘。烟花算法发明人，出版学术专著 12 部，发表学术论文 330 余篇。