DOI:10.20033/j.1003-7241.(2025)07-0104-06

基于智能机器的英语翻译错误纠正系统设计

崇 宁
（西北大学现代学院基础部，陕西 西安 710130）

摘 要：为了帮助教师完成英语教学任务，提升学生的英语语法学习效率。研究创新性地采用双编码器的结构，来抽取英语句法和语义特征。通过双向门控循环单元实现句法分析，而双向的Transformer结合双向门控循环单元则用于深入挖掘语义层面的信息。并结合中国学生的典型语法错误模式，设计融合规则和概率的数据增强技术，来拓展学习者语料库。结果表明，研究方法有效提升了语法纠错系统的效果。针对大学英语考试作文，模型显示出90.4%的精确率、81.33%的召回率和85.2%的F1得分，验证了其作为自动作文批改工具的优越性。研究为跨学科研究提供了新的视角和贡献。

关键词：机器翻译；英语；纠错；Bi-GRU；BERT
中图分类号：TP391.2 **文献标识码：**A **文章编号：**1003-7241（2025）07-0104-06

Design of English Translation Error Correction System Based on Intelligent Machine

CHONG Ning
(Basic Department of Modern College, Northwest University, Xi'an 710130, China)

Abstract: To help teachers complete the task of English teaching and improve the efficiency of students' English grammar learning. This study innovatively uses a dual encoder structure to extract English syntactic and semantic features. The bidirectional gated loop unit is used for syntactic analysis, and the bidirectional Transformer combined with the bidirectional gated loop unit is used to dig deeper into the semantic level information. Combined with the typical grammatical error patterns of Chinese students, a data enhancement technique integrating rules and probability is designed to expand the learner corpus. The results show that the method can effectively improve the effect of the grammar error correction system. For college English test compositions, the model shows an accuracy rate of 90.4%, a recall rate of 81.33% and an F1 score of 85.2%, which verifies its superiority as an automatic essay correcting tool. Research provides new perspectives and contributions to interdisciplinary research.

Keywords: machine translation; english; error correction; bi-gated recurrent unit; BERT

0 引言

随着全球一体化的加速，语言翻译成为连接不同国家和文化的关键纽带。然而，由于语言本身的复杂性以及文化间的差异，在翻译过程中难免会出现差错。这些差错不仅可能导致信息理解错误和传递不精确，还有可能在一定程度上损害双方的交流关系[1-2]。因此，研究高效的翻译错误纠正技术变得极其重要。在此背景下，基于智能机器的翻译系统显示出了巨大的发展潜力。机器翻译方法主要分为基于统计机器翻译与基于神经机器翻译（neural machine translation，NMT）的方法[3]。对于英语翻译语法问题可以采用机器翻译任务的方法来处理，输入带有错误的语句，通过系统处理后输出语法无误的语句。国内外的差异主要在于技术发展水平、应用场景、语言处理能力等方面。例如Yanwen C等研究人员运用数据挖掘技术辨认与学生翻译失误相关的视觉要素，开发分析工具以提升识别效果。通过关联分析与模型分类精确捕捉翻译误差，并通过可理解性评价反馈改进学习[4]。尽管研究成果有效降低了译文错误率，但对于复杂或含糊的句子，机器翻译可能无法准确捕捉原文的意图和情感。目前，基于序列到序列的英语翻译语法纠错模型主要使用单一编码器，依赖于神经网络提取关键特征，但在利用句法和语义信息上还有提升空间[5]。而针对中国英语学习者的纠错模型，则面临公开语料库匮乏的问题[6]。为此，提出一种双编码器结构，结合句法和语义编码器，在解码器端应用混合注意力机制。此外，针对中国学生常见的语法错误，开发结合规则和概率的噪声注入技术，扩充学习者语料库，实现英语翻译错误纠正系统的设

*基金项目：陕西省教育科学“十四五”规划项目（SGH24Y2894）
收稿日期：2024-02-27

计。研究旨在设计一种更加高效、智能的翻译错误纠正系统，以此准确识别并纠正各种类型的翻译错误。

1 基于智能机器的英语翻译错误纠正系统设计

1.1 基于Bi-GRU与BERT的编码器结构设计

英语句子的语法纠错任务需综合考虑句法和语义两方面信息。借助辅助编码器概念，研究采纳双编码器架构，来分别提取句法与语义特征，进而有效获取与语法错误更正（grammatical error correction，GEC）任务相关的句法和语义信息[7]。研究所设计的句法编码器结构包括一个词性嵌入层及一个双向门控循环单元（bi-gated recurrent unit，Bi-GRU）层，其结构见图1。

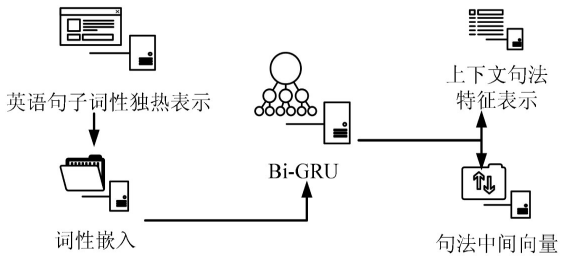


图1 句法编码器

图1中，句法编码器的词性嵌入层由前馈神经网络实现，位于Bi-GRU多层模型上方。该编码器的目的是从英语句子的词性独热编码中提取出关键的句法信息，生成句子的句法特征表示和句法中间向量，均为解码器后续处理的关键输入。首先，通过公式（1）计算得到句子的词性向量表示[8]。

$$\boldsymbol{S}_{\mathrm{p}}=\begin{bmatrix}0 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 1 & \cdots & 0\end{bmatrix} \tag{1}$$

式中，$\boldsymbol{S}_{\mathrm{p}}\in\mathbb{R}^{48\times n}$，$\boldsymbol{S}$表示一个英语句子中各个词性的独热编码。在该矩阵中，每一个列向量对应一个英文单词的词性独热编码表示。$\boldsymbol{S}_{\mathrm{p}}$将通过前馈神经网络在词性嵌入层中处理，生成新的词性向量表示$\boldsymbol{S}'_{\mathrm{p}}$，如式（2）所示。此过程将每个词性向量的维度转换为适合Bi-GRU层输入的格式。

$$\boldsymbol{S}'_{\mathrm{p}}=\tanh\left(W_{\mathrm{p}},\boldsymbol{S}_{\mathrm{p}}\right) \tag{2}$$

接下来，在Bi-GRU层的处理中，序列$\boldsymbol{S}'_{\mathrm{p}}$被输入，每个时刻$t$的隐藏状态$h_t$由正向GRU单元$h_t^+$的和反向GRU单元的$h_t^-$合并而成，如式（3）所示。

$$h_t=concat\left(h_t^+,h_t^-\right) \tag{3}$$

式中，$concat$函数为向量拼接操作。对于一个k层的Bi-GRU，假设第i层的隐层状态为h^i，则第i+1层的输入就是h^i，用$H_{\mathrm{syn}}=(h_1^{\mathrm{k}},h_2^{\mathrm{k}},\cdots,h_n^{\mathrm{k}})$表示最后一层的输出，其中$n$代表总计$n$个位置。序列$H_{\mathrm{syn}}$作为整个句子的句法特征表示。此外，在句法编码器中，通过拼接正向和反向GRU在最上层（第k层）最后时刻的隐藏状态，计算得到融合全句句法信息的中间向量c_{syn}，如式（4）所示。

$$c_{syn}=concat\left(h_t^{k+},h_1^{k-}\right) \tag{4}$$

在句法编码器完成计算后，输出包括句子的上下文句法特征表示$H_{\mathrm{syn}}=(h_{\mathrm{syn1}},h_{\mathrm{syn2}},\cdots,h_{\mathrm{synn}},)$和句法中间向量$c_{\mathrm{syn}}$，这些输出将被用作随后解码阶段的计算输入。研究设计的语义编码器如图2所示。

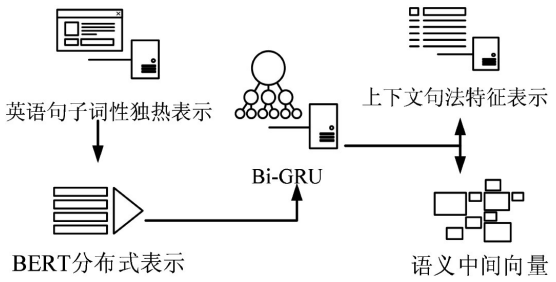


图2 语义编码器

图2中，语义编码器结合了BERT层和双向GRU层，以提取语义特征。首先，BERT模型为原文句子生成分布式向量表示，各个token向量不仅融入了BERT的预训练知识，还捕捉了语境语义信息。其次，Bi-GRU层对BERT输出进行深度语义特征提取，最终生成句子的上下文语义特征表示和语义中间向量。送入BERT模型前，英语句子首端加上起始标记[CLS]并在尾端附加结束符[SEP]。这些特殊符号源自BERT的预训练用途，在研究的编解码器结构中作为句首和句末标识使用。句子的输入格式由式（5）所示。

$$\begin{aligned}S_e=&\text{"[CLS] As we all know,}\\&\text{China develops rapidly in the past 20 years. [SEP]"}\end{aligned} \tag{5}$$

假定输入句子是S，BERT处理后的句子为S'，其中包含序列$\{x_1,x_2,\cdots,x_n\}$。每个成员x_i都是一个包含上下文语义信息且维度为768的向量，并融入了BERT预训练在广泛语料上获得的知识，见式（6）[9]。

$$S'=\mathrm{BERT}\left(S\right) \tag{6}$$

在语义编码器中，处理后的S'接着输入Bi-GRU层，该层的内部运算与句法编码器相同。通过计算，Bi-GRU顶层隐藏状态产出句子的上下文语义特征表示矩阵$\boldsymbol{H}_{\mathrm{senm}}=(h_{\mathrm{sem1}},h_{\mathrm{sem1}},\cdots,h_{\mathrm{semn}})$。同样，最后时间步的正向和反向隐藏状态拼接形成句子的语义中间向量C_{sem}。

1.2 基于混合注意力机制的解码器结构设计

为了更好地参照原句中上下文的句法和语义信息，研究设计一种混合注意力机制，利用单个查询向量，结合句法和语义编码器输出计算权重[10]。通过句法中间向量$\boldsymbol{c}_{\mathrm{syn}}$和语义中间向量$\boldsymbol{C}_{\mathrm{sem}}$形成查询向量$\boldsymbol{q}$，再用一个神经网络层提取关键信息，形成用于解码的混合表示，该过程见式（7）。

$$\boldsymbol{q}=concat\left(\boldsymbol{c}_{syn},\boldsymbol{c}_{sem}\right) \tag{7}$$

先对句法注意力进行计算，利用句法编码器生成的上下文向量矩阵$\boldsymbol{H}_{\text{senm}}=(h_{\text{sem1}},h_{\text{sem1}},\cdots,h_{\text{semn}})$，每一个位置$i$的$h_{\text{sem}i}$与$q$相结合，计算注意力权重。经过softmax标准化后进行加权累积，得到句法关注向量，该过程见式(8)。

$$\boldsymbol{a}_{\text{syn}}=\sum_{i=1}^{N}\alpha_{\text{i}}\cdot h_{\text{syni}}=\sum_{\text{i=1}}^{N}\text{softmax}\left(s\left(h_{\text{syni}},q\right)\right)h_{\text{syni}} \quad (8)$$

式中，$\boldsymbol{a}_{\text{syn}}$为最终的句法注意力向量，$\alpha_{\text{i}}$代表每一个位置$i$的注意力权重。$s$运算使用的是缩放点积的方法，公式如(9)所示[11]。

$$s\left(h,q\right)=\frac{h^{\text{T}}\cdot q}{\sqrt{D}} \quad (9)$$

式中，为了避免softmax在大输入值下的梯度消失问题，类似于Transformer的自注意力机制，D代表向量$\boldsymbol{h}$和$\boldsymbol{q}$的维度，研究对$\boldsymbol{h}$和$\boldsymbol{q}$的点积进行缩放处理。得到句法和语义注意力向量$\boldsymbol{a}_{\text{syn}}$与$\boldsymbol{a}_{\text{sem}}$后，研究将它们合并，并通过前馈网络进一步提取特征，形成混合注意力向量，见式(10)。

$$\boldsymbol{a}_{\text{final}}=\tanh\left(W_{\text{att}}\cdot\left(\text{concat}\left(\boldsymbol{a}_{\text{syn}},\boldsymbol{a}_{\text{sem}}\right)\right)\right) \quad (10)$$

通过计算得到了融合句法与语义的混合注意力向量$\boldsymbol{a}_{\text{final}}$，帮助后续Bi-GRU层解码。混合注意力层的内部结构如图3所示。

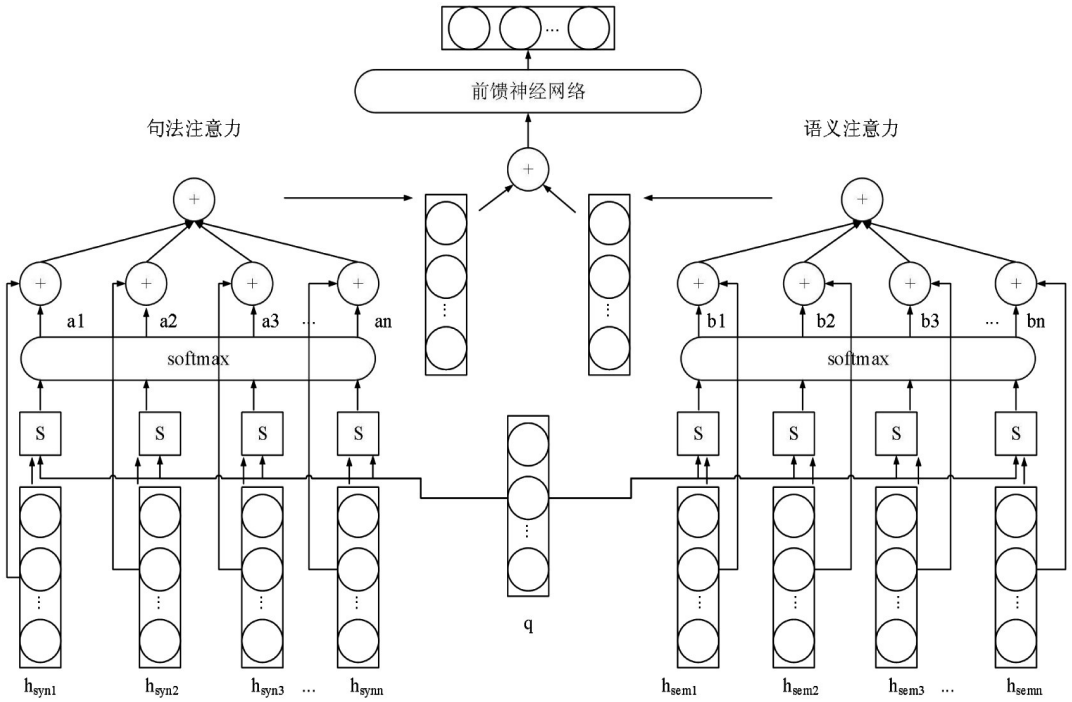


图3 混合注意力层内部结构

解码器核心采用单向门控循环单元(gated recurrent unit，GRU)架构。该设计考虑到解码过程仅涉及目前为止生成的单词和编码器提供的原始句子信息，无需双向网络结构。在每个时间点t，GRU单元接收上一时间点的预测词$\hat{\boldsymbol{y}}_{t-1}$作为输入$x_t$。同时，前一时刻的隐藏状态$h_{t-1}$经过注意力机制处理，作为查询向量$q$。该向量与句法编码器的句法特征和语义编码器的语义特征分别进行计算，形成句法和语义注意力。经过前馈网络处理后，生成最终的注意力向量a_{t-1}，该向量随后被用作GRU单元当前时刻的隐藏状态输入。每个GRU单元内部的计算过程见式(11)。

$$\begin{cases}r_t=\sigma\left(W_r\cdot\left[a_{t-1},x_t\right]\right)\\ \tilde{h}_t=\tanh\left(W_h\cdot\left[r_t\cdot a_{t-1},x_t\right]\right)\\ z_t=\sigma\left(W_z\cdot\left[a_{t-1},x_t\right]\right)\\ h_t=\left(1-z_t\right)\cdot a_{t-1}+z_t\cdot\tilde{h}_t\end{cases} \quad (11)$$

解码器的输出层结构由一个全连接层和softmax分类器组成。设在时刻t，GRU输出维度为M的向量$\boldsymbol{h}_t$。此向量经全连接层转换成维度为词汇量V的向量$\boldsymbol{o}_t$，其计算过程遵循式(12)。

$$\boldsymbol{o}_t=\tanh\left(W_h\cdot\boldsymbol{h}_t\right) \quad (12)$$

式中，$W_h\in\mathbb{R}^{\text{V}\times\text{M}}$。通过softmax函数，将向量$\boldsymbol{o}_t$转化为概率分布向量$\hat{\boldsymbol{y}}_t$，其中每个元素代表选取对应词汇的可能性。在模型训练阶段，选取概率最高的词作为该时刻输出，损失函数随后衡量预测准确度，并启动反向传播以优化参数。模型不断进化至最佳效能。交叉熵损失函数，通常用于评估和计算预测误差，计算见式(13)。

$$H\left(\hat{\boldsymbol{y}}_t,y_t\right)=-\sum_{i=1}^{n}\hat{\boldsymbol{y}}_t\left(x_i\right)\log\left(y_i\left(x_i\right)\right) \quad (13)$$

在某一时刻t，模型预测的概率分布表示为$\hat{\boldsymbol{y}}_t$，而y_t时刻的实际标签y_i则以单一词汇的独热编码形式出现。每

个向量元素由 x_i 标识。为了优化已训练好的语法纠错(GEC)模型的性能,在进行推理时必须实施精选的策略,以提高模型的准确性和有效性。

1.3 英语翻译语法自动纠错系统设计

对于英语文本的处理,首先通过预处理环节进行句子切分、词汇分解及词性标注。随后将这些信息转化为向量形式。接着,在编码阶段,系统将这些向量进行深入编码,提取句法和语义特征。解码器根据这些特征重构句子,进行语法纠正。最终依据特定筛选机制,系统挑选出最佳纠错结果,以确保输出的高质量和准确性。研究设计的英语翻译错误纠正系统整体结构如图4所示。

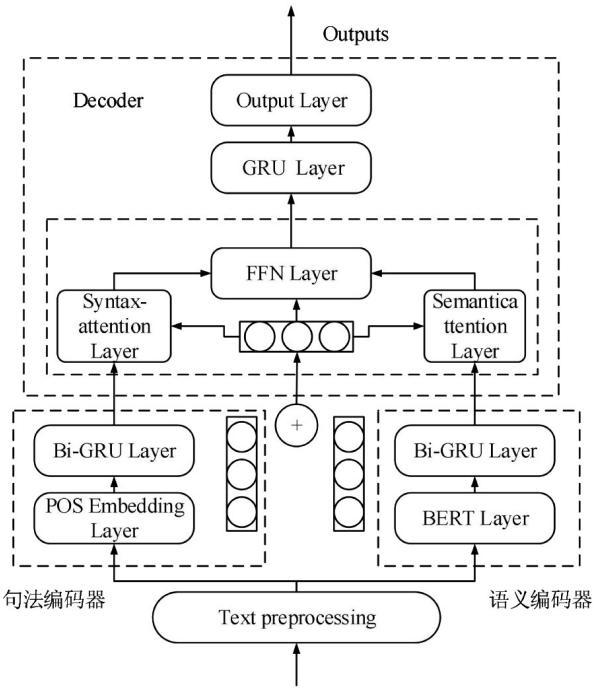


图4 英语翻译错误纠正系统

图4中,该系统由两个编码器和一个解码器组成。第一个编码器关注句法,利用词性嵌入和Bi-GRU来分析句子结构。第二个编码器基于BERT和Bi-GRU,分析句子的语义。这两部分的信息被送入解码器,后者通过混合注意力机制同时处理句法和语义信息,最终通过GRU层和输出层生成纠错结果。系统还采用集束搜索策略提高准确度。为适应中国学生的英语学习,该模型结合规则数据增强和基于概率的噪声添加技术,专门针对常见错误进行训练。在GEC模型的设计中,核心是生成一个由模型视为最优的文本序列。这涉及一个基于输入序列 X 的最大概率搜索问题,目标是找到一个单词序列 $\hat{Y}$,以使概率 $P(\hat{Y}|X)$ 最大化。这个搜索涉及的空间大小 $|V|^T$ 取决于词表 $|V|$ 的规模和模型预测生成的句子长度 T。研究采用了贪心搜索的改进方法集束搜索(beam search)[12]。假设在 t-1 时刻,存在B个候选序列的集合表示为 $\gamma_{t-1}=\{Y_{t-1}^1,\ldots,Y_{t-1}^B\}$。用集合 $S_t=\left\{\left(Y_{t-1}^b,y_t\right)\middle|\forall\left(Y_{t-1}^b\in\gamma_{t-1}\right)\wedge\left(y_t\in X\right)\right\}$ 来表示 t 时刻所有的序列组合。t 时刻的B个候选序列见式(14)。

$$\gamma_t=\arg\max_{Y_t^1,\ldots,Y_t^B\in S_t}\sum_{b=1}^{B}\log P\left(Y_t^b\middle|X\right) \tag{14}$$

集束搜索在各个时间步骤产生词语时,每个单词的概率位于0至1的区间内,随着序列构建,较长句子的整体生成概率逐渐降低,因累积计算结果趋小。这种机制可能使模型偏好于较短的句子。为了解决该问题,研究提出通过在计算公式中融入一个长度调节项T,来平衡句子的长度选择,修改后的公式如式(15)所示。

$$\gamma_t=\arg\max_{Y_t^1,\ldots,Y_t^B\in S_t}\sum_{b=1}^{B}\frac{1}{T^{\alpha}}\log P\left(Y_t^b\middle|X\right) \tag{15}$$

式中,$\alpha\in[0,1]$。当 α=0 时,则不进行长度惩罚,当 α=1 时,则直接使用句子长度 T 来进行惩罚。

2 实验结果分析

实验是在一台云服务平台和一台办公桌面计算机上完成的。表1中详细阐述所依赖的硬件配置和进行开发

表1 实验环境及相关参数设计

实验环境		实验参数	
操作系统	Windows 10	μ	0.3
开发环境	Jdk 1.8	词性向量维度	48维
开发平台	Eclipse	隐藏层向量维度	128维
CPU(服务器)	AMD EPYC 7402 @ 2.8 GHz	dropout	0.2
内存(服务器)	60 GB	激活函数	tanh
GPU(服务器)	NVIDIA GeForce RTX 3090	beam size	5
显存(服务器)	24 GB	α	0.5
CPU(桌面)	Intel(R) Core(TM)i5-6300HQ CPU @ 2.30 GHz	batch-size	32
内存(桌面)	16 GB	epoch	40
GPU(桌面)	NVIDIA GeForce GTX 960	学习率	0.005
显存(桌面)	4 GB	BERT模型	bert-base-cased

工作必备的软件系统。此外，还显示了在实验阶段所采用的模型配置及其相关参数的设定。

实验数据集划分为三个部分:训练集、开发集和测试集。训练集主要包括Lang-8、NUCLE、FCE以及CLEC语料库，这些均为当前语法纠错领域公认的学习者训练资源。在数据增强实验中，研究结合了规则和概率方法，对语料库进行扩充。原始训练集含有约200万句平行句对，而扩充规模设定为4 MB、8 MB、12 MB和16 MB，具体结果见图5。

由图5可知，随着数据规模增大，模型精度显著提高，精确率从0提升至73.2%，召回率亦从38.2%增至39.26%，证明研究提出的融合规则和概率的数据扩充方法能有效提升GEC模型性能。在两种测试集上进行测试，研究模型与4种基线模型的对比结果见图6。

图6(a)显示，在CoNLL-2014测试集上，研究模型在精确率、召回率和F0.5值方面显著超过Nested-GRU和

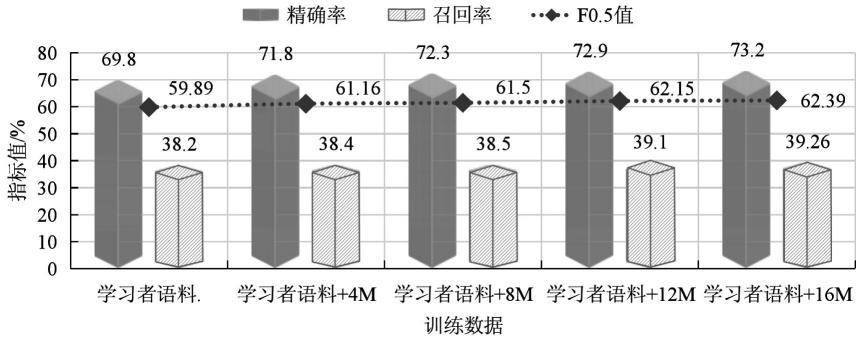
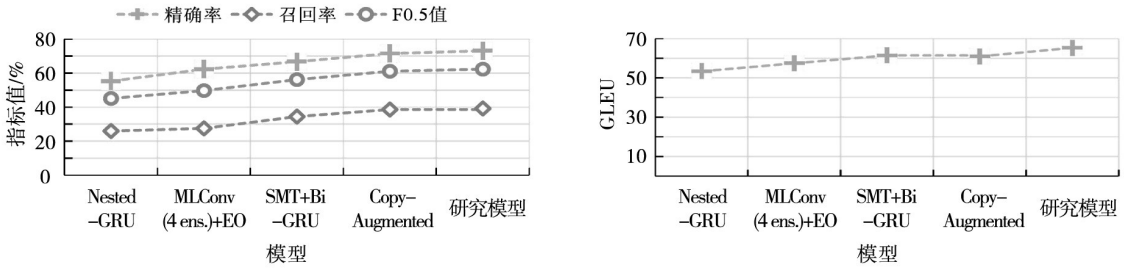
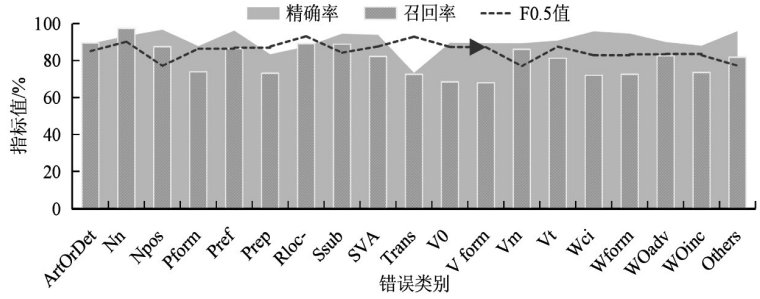


图5 数据扩充结果

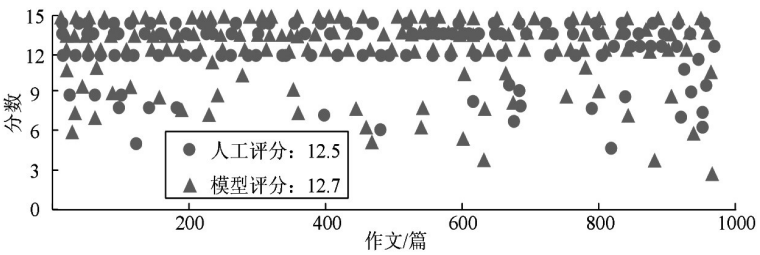


(a) CoNLL-2014测试集 (b) JFLEG测试集

图6 各模型在不同测试集上的性能结果



(a) 语法错误纠正情况分类统计



(b) 作文语法批改模型评分与人工评分对比

图7 研究模型对英语作文的纠错结果

MLConv(4 ens.)+EO模型，比SMT+Bi-GRU模型提高了6.42%、4.83%和6.26%。它也略胜于使用预训练的Copy-Augmented Transformer+Pretrain模型。图6(b)显示，研究模型通过双编码器和混合型注意力机制，在句法和语义上有效纠正文本，无须依赖额外的SMT系统就接近SMT+Bi-GRU的GLEU分数，表现优于其他NMT模型，显示出较高的纠错流畅性。为了进一步评估模型在中国学生英语作文语法批改方面的效能，实验从CLEC语料库中的SET3和SET4选取了500篇非英语专业大学生的四六级作文进行测试，结果如图7所示。

由图7(a)可知，研究模型在处理四六级英语作文时，平均语法错误的纠正精确率为90.4%，召回率为81.33%，F1值达85.2%。这些数据反映出模型不仅具备高效的纠错能力，而且在实际应用中显示出其优越的实用性和可靠性，能在一定程度上替代教师进行英语作文的语法批改。由图7(b)可知，人工语法评分平均为12.5分，而模型评分平均值为12.7分。这两者之间的误差仅为0.2分，这表明研究设计的英语翻译修正模型具备较高的实用性，并且在一定程度上能够替代教师对英语作文的语法校正工作。

3 结束语

为了能减轻教师的工作量，提升英语翻译错误纠正及学生英语语法学习的效率，研究设计了一种新型双编码器和混合注意力机制的英语翻译纠错模型，同时，设计了融合规则和概率的数据扩充方法。结果表明，研究方法有效提升了语法纠错系统的效果，其准确率达73.2%，召回率39.26%。在CoNLL-2014上，研究模型明显优于其他系统。与SMT+Bi-GRU模型相比，性能提升约6个百分点。在JFLEG集上，双编码器与混合注意力进一步增强了语言流畅性，且成绩接近顶尖模型。针对大学英语考试作文，模型显示出90.4%的精确率、81.33%的召回率和85.2%的F1得分，验证了其作为自动作文批改工具的优越性。虽然模型表现良好，但依赖更多合成数据提升有限，今后将优化数据扩充方法，进一步提高纠错能力。

参考文献：

[1] 高利利，马鹏霄. 基于语音识别的人机交互语言翻译系统设计[J]. 自动化与仪器仪表，2023(6)：175-178，183.

[2] 黄德根，刘俊鹏，刘欢，等. 一种融合特定语言适配器模块的多语言神经机器翻译方法[J]. 计算机科学，2022，49(1)：17-23.

[3] 龚龙超，郭军军，余正涛. 基于源语言句法增强解码的神经机器翻译方法[J]. 计算机应用，2022，42(11)：3386-3394.

[4] YANWEN C. Analyzing the design of intelligent English translation and teaching model in colleges using data mining[J]. Soft computing：A fusion of foundations，methodologies and applications，2023，27(19)：14497-14513.

[5] 孙晓东，王丕坤，杨东强. 基于反向翻译的英语语法纠错应用研究[J]. 计算机技术与发展，2022，32(10)：143-150.

[6] 刘宇宸，宗成庆. 跨模态信息融合的端到端语音翻译[J]. 软件学报，2023，34(4)：1837-1849.

[7] 张明，卢庆华，黄元忠，等. 自然语言语法纠错的最新进展和挑战[J]. 计算机工程与应用，2022，58(6)：29-41.

[8] 郭雨欣，陈秀宏. 融合BERT词嵌入表示和主题信息增强的自动摘要模型[J]. 计算机科学，2022，49(6)：313-318.

[9] 边陆，林少波，郭栋，等. 基于改进型深度学习算法的计算机数据分析[J]. 微型电脑应用，2023，39(9)：94-98.

[10] 刘建铭，陈伟侠，卢仲康. 联合注意力机制与目标点信息的车辆轨迹预测[J]. 计算机测量与控制，2023，31(11)：106-118.

[11] 姚冲，周晖. 基于时空图的行人多模态轨迹预测方法[J]. 计算机工程与设计，2022，43(10)：2918-2925.

[12] 白雯. 基于决策树的英文翻译软件缺陷检测方法[J]. 自动化技术与应用，2023，42(12)：108-111，176.

作者简介：崇宁(1980—)，女，硕士，副教授，研究方向：系统设计、英语语言文学，英语教学等。