CCL24-Eval任务7系统报告：基于大模型数据增强的作文流畅性评价方法

彭倩雯[1]，高延子鹏[1]，李晓青[1]，闵凡珂[1]，李明锐[1]，王志春[1,2]，刘天昀[3]
[1] 北京师范大学人工智能学院
[2] 智能技术与教育应用教育部工程研究中心
[3] 中国科学院信息工程研究所
{qwpeng,yanzipenggao,xiaoqingli,minfanke,mingruili}@mail.bnu.edu.cn
zcwang@bnu.edu.cn , liutianyun@iie.ac.cn

摘要

CCL2024-Eval任务7为中小学生作文流畅性评价（Chinese Essay Fluency Evaluation，CEFE），该任务定义了三项重要且富有挑战性的问题，包括中小学作文病句类型识别、中小学作文病句改写、以及中小学作文流畅性评级。本队伍参加了评测任务7的三项子任务，分别获得了45.19、43.90和45.84的得分。本报告详细介绍本队伍在三个子任务上采用的技术方法，并对评测结果进行分析。

关键词： 作文流畅性评价 ； 数据增强 ； 语言模型

System Report for CCL24-Eval Task 7: Essay Fluency Evaluation Method Based on Large Model Data Augmentation

Qianwen Peng[1],Yanzipeng Gao[1],Xiaoqing Li[1],Fanke Min[1],Mingrui Li[1]
Zhichun Wang[1,2],Tianyun Liu[3]
[1] School of Artificial Intelligence,Beijing Normal University
[2] National Engineering Laboratory for Cyberlearning and Intelligent Technology
[3] Institute of Information Engineering，Chinese Academy of Sciences
{qwpeng,yanzipenggao,xiaoqingli,minfanke,mingruili}@mail.bnu.edu.cn
zcwang@bnu.edu.cn , liutianyun@iie.ac.cn

Abstract

The CCL2024-Eval Task 7 focuses on Chinese Essay Fluency Evaluation (CEFE) for primary and secondary school students. This task encompasses three significant and challenging problems: identifying sentence errors in essays, rewriting erroneous sentences, and rating the fluency of the essays. Our team participated in all three sub-tasks (tracks) of Task 7, achieving scores of 45.19, 43.90, and 45.84. This report provides a detailed account of the technical methods employed by our team for each sub-task and analyzes the evaluation results.

Keywords: Essay Fluency Evaluation , Data Augmentation , Language Model

1 任务概述

作文流畅性指的是作文语句的通顺程度和语言使用的规范程度，是体现作文质量的重要方面。CCL24-Eval任务7为中小学生作文流畅性评价，设立三个子任务，包括：

（1）中小学作文病句类型识别：识别作文中不同的病句类型。

（2）中小学作文病句改写：改写作文中的病句使其成为正确句子。

（3）中小学作文流畅性评级：对作文流畅性作三等级评价。

评测任务提供了基于以汉语为母语的中小学生考试作文构建的测试数据，要求参赛者从词法、句法、语义等多角度对作文流畅性进行分析。

2 子任务一：中小学作文病句类型识别

2.1 任务描述

中小学作文病句类型识别任务本质上是一个多标签分类任务，其目标是针对给定的作文病句，预测该病句所属的一个或多个类型。病句类型包括4大类14小类，如表1所示。

Table 1: 中小学作文病句类型分类

粗粒度错误类型	细粒度错误类型
字符级错误	缺字漏字、错别字错误、缺少标点、错用标点
成分残缺型错误	主语不明、谓语残缺、宾语残缺、其他成分残缺
成分赘余型错误	主语多余、虚词多余、其他成分多余
成分搭配不当型错误	语序不当、动宾搭配不当、其他搭配不当

2.2 数据准备

为了训练模型实现对病句的准确分类，本队伍通过以下步骤构建训练数据集。

（1）开放作文数据获取：本队伍从中小学作文网[0]等网站，获取中小学生优秀作文234篇，包含句子共计10000句。

（2）基于GPT-3.5-Turbo的作文生成：构建如表2所示的提示，调用GPT-3.5-Turbo(OpenAI, 2023)模仿中小学生身份生成作文240篇。

Table 2: 中小学作文生成提示

类别	提示
叙事	你是一个初二年级的学生，现在请以“人生”为主题，写20篇叙事作文，要求内容充实，语句通顺。
记人	你是一个三年级的学生，现在请以“身边的英雄”为主题，写20篇记人作文，要求内容充实，语句通顺。

（3）基于大语言模型的病句生成：构建如表3的提示，调用GPT-3.5-Turbo 、文心一言3.5[1]等语言模型从给定的正确句子出发生成指定类型的病句；该步骤共生成病句19万条，其中由GPT-3.5-Turbo生成的病句占比60%，由文心一言生成的病句占比40%，覆盖表1中所有错误类型，并在各个细粒度错误类型上均匀分布。

Table 3: 提示示例

细粒度错误类型	提示
错别字错误	你是一个语文老师，你的任务是将句子中的一个字改成错别字，你可以从两个角度入手，一是把这个字改成它的形似字，二是把这个字改成它的同音字。注意不要原句返回也不要改变句子的长度，现在你试试句子：
错用标点	你是一个语文老师，你的任务是将句子中标点符号换成错误的，试试句子：
动宾搭配不当	现在你是一名语文不好的小学生，写出的句子经常出现中文语法错误。特别的，你只会写出“动宾搭配不当”的句子。动宾搭配不当指的是在句子中，动词（动作）和宾语（动作的对象）之间的搭配不合语法规则或语义习惯，导致句子意思不清或逻辑混乱。现在将给你一个句子，请修改为动宾搭配不当的句子。正确句子：

[0] http://www.zuowen.com/
[1] https://yiyan.baidu.com/

Table 4: 语言模型生成病句示例

细粒度错误类型	语言模型生成病句
错别字错误	她的手小小的，她的膊搏胖胖的，她的腿短短的，好像我家的布娃娃。
错用标点	今天是爸爸的生日；我决定亲手种一颗向日葵送给他，作为礼物。
动宾搭配不当	在家里和家人一起种植花的时候，我感觉特别开心和幸福。

（4）格式转换：将“原句-病句”的数据格式转换为“病句-错误类型标签”的数据格式，标签根据生成病句时使用的提示确定。

（5）数据增强：本队伍还使用了一种误差不变的数据增强方法，通过替换命名实体来增强数据的多样性。在本次测评任务中，这种数据增强方法可以提高模型的识别语病错误能力。

（6）数据清洗：由于数据集中存在噪声，需要对其进行清洗。其具体步骤如表5所示。

Table 5: 数据清洗步骤

步骤	实现
去除特殊字符和重复标点	正则表达式
统一使用中文标点	编写替换函数
去除输入长度与输出长度差异过大的文本	设置差异阈值5
去除长度过短的文本	设置最小长度阈值10

2.3 模型构建及训练

针对作文病句类型识别任务，本队伍构建了基于BERT的病句分类模型，使用训练数据和评测任务数据集对模型进行了微调训练。给定句子s，设$\text{BERT}_{\text{CLS}}(s)$为模型输出层[CLS]的隐式向量，病句分类模型可表示为：

$$f(s) = \text{MLP}(\text{BERT}_{\text{CLS}}(s)) \tag{1}$$

模型在微调训练过程中，采用了Adam优化器对BCEWithLogitsLoss损失函数进行优化。为提高模型分类性能，本队伍基于评测任务提供的验证集，搜索确定了每个标签分类概率的最优阈值；以0.001为步长，从0.5到0.6搜索并计算模型得分，选择得分最高的阈值作为模型各标签最终的分类阈值。基于上述模型及分类方法，本队伍在测试集上取得了45.19的得分。

2.4 对抗训练

本队伍尝试通过FGM对抗训练技巧 (Miyato et al., 2016)以提升模型的鲁棒性。其具体过程如算法1所示

Algorithm 1 FGM对抗训练流程

1: **for** 每个输入向量x **do**
2: 计算x 的前向损失、反向传播得到梯度$g1$
3: 根据嵌入矩阵的梯度计算出噪声扰动r，并加到当前嵌入上，相当于$x + r$
4: 计算$x + r$ 的前向损失，反向传播得到对抗的梯度，累加到$g1$上得到梯度$g2$
5: 将嵌入恢复为第一步时的值
6: 根据梯度$g2$对参数进行更新
7: **end for**

实验结果表明，FGM对抗训练在作文病句类型识别任务中对模型表现并没有明显增益效果。本队伍对此分析：对于作文病句类型识别任务，文本中存在大量噪声和对于分类并无帮助的内容。因此，通过对抗训练引入增强扰动，理论上应该能够增强模型的抗干扰性。但由于本队伍数据增强阶段已加入了有效的扰动，故使用FGM对抗训练技巧无法进一步提高模型的性能。

3 子任务二：中小学作文病句改写

3.1 任务描述

中小学作文病句改写任务本质上是一个文本生成任务，目标是针对输入的病句，输出改正后的句子。任务要求在保持语义不变的前提下，为中小学生作文中的错误句子提供最小化修改方案。

3.2 数据准备

为了训练模型实现对病句的准确修改，本队伍通过组合以下数据集构建混合训练数据。

（1）评测任务训练集：评测任务训练集来源于中小学作文数据，规模约1300句。

（2）大模型生成的病句数据集：在中小学作文病句类型识别任务中使用语言模型生成的数据的基础上，通过数据处理和数据清洗等步骤，得到适用于中小学作文病句改写任务的19万条句子。

（3）正确句子数据集：评测任务测试集中包含无语病的正确句子，这要求限制模型对正确句子的修改。为改善模型的过纠问题，本队伍构建了一个正确句子数据集，规模约10000句，数据来源于中小学作文病句类型识别任务中获取的开放数据集。

（4）开放语法纠错数据集：开放语法纠错数据集包括苏州大学和阿里巴巴达摩院联合发布的MuCGEC (Zhang et al., 2022b)中文语法纠错评测数据集、YACLC汉语学习者文本多维标注数据集 (Yingying Wang, 2021)等，规模约360万句。开放数据集和评测任务数据集相比，前者每一错误句子中包含的错误数目远少于后者。

综上，本队伍数据集整体构成如表6所示。

Table 6: 数据集信息

数据集	数量	平均长度
评测任务训练数据	1,336	46.33
生成病句数据	140,000	35.07
正确句子数据	10,000	36.53
开放纠错数据	3,648,481	30.84

3.3 模型构建及训练

针对中小学作文病句改写任务，本队伍采用大模型文本纠错方法，使用训练数据对基础模型进行了指令微调，构建了基于Qwen-14B(Bai et al., 2023)的病句改写模型。针对大模型普遍存在的过纠问题，本队伍通过设计具体的提示模版来强化模型的泛化能力，如表7所示。在训练过程中，本队伍探索了不同的参数设定，其在任务中的表现结果如表8所示。最终，本队伍在测试集上取得了43.90的成绩。

Table 7: 病句修改提示模板

类型	示例
身份假设	你是一个中小学语文老师
任务描述	你需要将下面的句子改为正确的、你需要修改下面的句子、你需要改正病句
具体要求	保证输出句子和原句之间较小的距离、请尽可能做较小的改动、保证修改后句子的流畅度、请尽可能让修改后的句子流畅

Table 8: 不同参数的实验结果

序列长度	学习率	迭代次数	BERT PPL	Levenshtein	BlEU-4	BERTscore
1024	5e-5	3.00	2.70	1.66	0.90	0.98
1024	5e-5	10.00	2.60	3.82	0.99	0.99
90	5e-4	10.00	2.59	3.61	0.98	0.99
1024	5e-4	10.00	2.59	3.76	0.99	0.99

3.4 大模型病句改写性能分析

本队伍对比了不同模型在病句改写方面的效果，包括Qwen-7B、Qwen-14B-Chat、Qwen1.5-14B-Chat 和基于BART (Lewis et al., 2019) 的序列标注模型。在纠错效果方面，本队伍观察到基于BART 的序列标注模型在处理字符级错误方面表现突出，而Qwen 模型则在涉及语义错误、如语序不当等复杂情况下表现更为出色，能够识别并修改BART 模型无法处理的错误。如表9所示，在处理病句‘这根本是一座不可能翻去的山。”’时，Qwen 系列模型能够根据句意将“翻去”修正为“翻越”，并且还能识别出句末的符号错误，而BART 则只能将“翻去”修改为“翻过去”。

Table 9: 大模型病句改写性能

原句	这根本是一座不可能翻去的山。”
Qwen-7B	这根本是一座不可能翻去的山。
Qwen-14B-Chat	这根本是一座不可能翻越的山。
Qwen1.5-14B-Chat	这根本是一座不可能翻越的山。
BART	这根本是一座不可能翻过去的山。”

在结果评分层面，由于大模型的幻觉问题，在处理细粒度纠错任务时会出现过纠现象，不必要的过度改写导致得分始终无法提升。但当考虑到大型模型的规模和参数量达到一定程度时，模型涌现出更多能力，具有更多的潜能来捕捉文本的语义信息和上下文关系。这种额外的能力使它们能够更准确地识别和纠正复杂错误，包括语义错误、逻辑错误等，而这些错误可能会超出传统文本纠错模型如BART的处理范围。因此，大型模型在病句改写任务中展现出了更为广阔的潜力，有望推动文本生成和修正领域的发展。

4 子任务三：中小学作文流畅性评级

4.1 任务描述

中小学作文流畅性评级任务本质上是一个多分类任务，目标是针对给定的作文，预测作文在流畅性方面所属的等级。本次评测任务共定义了三个流畅性等级：优秀、一般、不及格。

4.2 数据准备

考虑到存在语病的作文流畅性普遍较差，本队伍在中小学作文流畅性评级任务中额外训练了中小学作文病句识别模型。针对中小学作文病句识别模型，本队伍预先对数据进行处理，将评测任务训练集中无语病的句子作为正例，存在语病的句子作为负例。数据集规模为2600句，数据集格式：输入为作文句子，输出为有语病或无语病两种标签。

4.3 模型构建及训练

针对作文流畅性评级任务，本队伍构建了基于BERT的分类模型，使用训练数据对模型进行了微调训练。给定作文文段s，设$\mathrm{BERT}_{\mathrm{CLS}}(s)$为模型输出层[CLS]的隐式向量，作文流畅性评价模型可表示为：

$$f(s) = \mathrm{MLP}(\mathrm{BERT}_{\mathrm{CLS}}(s)) \tag{2}$$

本队伍在构建作文流畅性评级模型时，选择了中文BERT-WWM(Lewis et al., 2019)预训练模型来初始化模型的编码器部分，并利用官方数据集对模型进行了精细的训练。本队伍测试了不同学习率设置，发现过高或过低的学习率均会对模型的性能产生负面影响，结果如表10。

Table 10: 不同学习率的实验结果

学习率	Precision	Recall	F1 Score	Overall Score
5e-5	71.43	68.18	69.77	69.00
5e-4	50.00	45.45	47.62	47.44
5e-6	40.91	40.91	40.91	39.49

在解码过程中，模型有时会忽视文段中的句子语病。为了改善这一状况，本队伍针对性地设计了一套如算法2所示的解码后处理策略。通过本策略，模型能够更有效地纠正模型在解码过程中忽视语病的问题。

Algorithm 2 解码后处理

初始标签预测$\hat{y} = \mathcal{M}(s)$，检测语病数量n_e
if $\hat{y} =$ 优秀 $\wedge\, n_e > \tau$ **then**
    if $n_e \leq \tau_{\text{一般}}$ **then**
        最终标签$\hat{y} \leftarrow$ 一般
    else
        最终标签$\hat{y} \leftarrow$ 不及格
    end if
end if

具体而言，本队伍使用病句识别模型，判断文段中是否存在语病；使用基于BERT的分类模型，判断作文流畅性评级。若输入作文被初始评定为优秀，但其中存在语病，则根据存在语病的句子数量，将其评级降低为一般或不及格。为保证后处理策略的有效性，本队伍在验证集上对解码后处理模型的阈值进行搜索，确定其阈值为3。这意味着，如果一篇初始被评为优秀的作文中被识别的语法错误数量超过3，则会被调整评级为不及格；如果存在语法错误但数量未超过3，则会被调整评级为一般。基于上述模型及分类方法，本队伍在测试集上取得了45.84的得分。

5 结语

在CCL2024-Eval任务7-中小学作文流畅性评价评测任务中，针对三个子任务本队伍分别提出了中小学作文病句类型识别模型，中小学作文病句改写模型和中小学作文流畅性评级模型，并且提出了针对中小学作文数据集稀缺问题的大模型辅助数据生成方法，还尝试使用了一些额外的性能提升技术，例如对抗训练、多轮微调等。实验结果表明，本队伍提出的多种策略均可以使模型性能得到有效的提升，最终三个子任务得分分别为45.19、43.90、45.84。通过这三个子任务，本队伍实现了多维细粒度的自动作文评价，有效提高了作文评分的可解释性和反馈的丰富程度。如今，作文的自动流畅性评价具有明确的研究意义和应用场景。本队伍的方法推进了自动批改在课堂教学中的应用，可以有效辅助教师进行课堂教学。但是，要想真正实现让机器像人一样去欣赏和批判写作，包括对文章的立意思辨、篇章结构等方面进行评价依然是非常困难的。如何持续提高机器的对作文的审美能力和鉴别水平依然是开放问题。

6 致谢

资助本工作项目：科技创新2030 -“新一代人工智能”重大项目（2021ZD0113000），国家自然科学基金项目（62276026）。

参考文献

J. Achiam, S. Adler, S. Agarwal, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Izumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. *arXiv preprint arXiv:1909.00502*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online, April. Association for Computational Linguistics.

OpenAI. 2023. Gpt-3.5-turbo. Accessed: 2024-06-18.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51, Melbourne, Australia, July. Association for Computational Linguistics.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 25–35, Suzhou, China, December. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: Fine-grained corpus for Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918. Association for Computational Linguistics.

Liner Yang Yijun Wang Xiaorong Lu Renfen Hu Shan He Zhenghao Liu Yun Chen Erhong Yang Maosong Sun Yingying Wang, Cunliang Kong. 2021. Yaclc: A chinese learner corpus with multidimensional annotation. *arXiv preprint arXiv:2112.15043*.

Y. Zhang, H. Jiang, Z. Bao, et al. 2022a. Mining error templates for grammatical error correction. *arXiv preprint arXiv:2206.11569*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022b. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States, July. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

J. Zhou, C. Li, H. Liu, et al. 2018. Chinese grammatical error correction using statistical and neural models. *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II*, pages 117–128.