CCL24-Eval任务7系统报告：中小学作文语法错误检测、病句改写与流畅性评级的自动化方法研究

田巍
北京慧点科技有限公司，北京，中国
tianweia@mail.taiji.com.cn

摘要

本研究旨在提高中小学生作文评改的质量和效率，通过引入先进的自然语言处理模型进行作文病句检测、纠正和流畅性评分，并分别针对三个具体的任务进行了模型构建。在任务一中，提出语法错误替换方法进行数据增强，接着基于UTC模型对语病类型进行识别。在任务二中，融合了预训练的BART模型和SynGEC策略进行文本纠错，充分利用了BART的生成能力和SynGEC的语法纠错特性。任务三中，基于TextRCNN-NEZHA模型进行作文流畅性的评级，构建了一个能够综合语义信息的分类器。经评测，本文提出的方法在任务一和任务二中均位列第一，任务三位列第二，即提出的方法可以有效地识别病句类型和纠正作文中的病句，并给出合理的作文流畅性评级。

关键词： UTC ； BART ； SynGEC ； TextRCNN-NEZHA

System Report for CCL24-Eval Task 7: Research on Automated Methods for Grammar Error Detection, Malapropism Revision, and Fluency Grading in Primary and Secondary School Compositions

Wei Tian
Beijing Smartdot Technology Co., Ltd, Beijing, China
tianweia@mail.taiji.com.cn

Abstract

This study aims to improve the quality and efficiency of essay grading for elementary and middle school students by introducing advanced natural language processing models to detect, correct malformations in sentences, and score fluency in essays. Also, models were built for three specific tasks. In Task 1, a method of grammatical error replacement was proposed for data augmentation, followed by the identification of types of language maladies based on the UTC model. In Task 2, the integration of the pretrained BART model and SynGEC strategy was implemented for text correction, fully utilizing BART's generative capabilities and SynGEC's grammatical correction features. In Task 3, the essay's fluency grading was conducted based on the TextRCNN-NEZHA model, building a classifier that integrates semantic information. Upon evaluation, the methods proposed in this paper ranked first in Tasks 1 and 2 and second in Task 3, demonstrating the effectiveness of the proposed methods in identifying and correcting malformations in sentences, as well as providing a reasonable fluency rating for essays.

Keywords: UTC , BART , SynGEC , TextRCNN-NEZHA

第二十三届中国计算语言学大会论文集，第278页-第284页，太原，中国，2024年7月25日至28日。
卷3：评测报告

1 引言

在教育领域不断发展和网络普及的背景下，作文评价面临着劳动成本高和效率低的挑战，从而促使研究人员和机构去探索使用计算机技术进行作文的自动评改。本任务目的在于帮助学生更加明确地识别自己写作中的问题，并为他们的作文修改提供具体的指导，评测任务包括以下三个具体方面：

（1）中小学作文病句类型识别：识别并分类作文中可能出现的各种病句类型，为病句的改正提供基础。

（2）中小学作文病句改写：通过自动化技术，针对识别出的病句进行有效的改写，以提升作文的流畅性和整体质量。

（3）中小学作文流畅性评级：综合作文的词汇使用、结构组织等方面，给出作文流畅性的综合评定，以指导学生如何提高写作能力。

为解决上述问题，文本在任务一中，基于UTC模型(Lou et al., 2023)进行两个阶段的训练；在任务二中，融合了预训练的BART模型(Lewis et al., 2019)和SynGEC策略(Zhang et al., 2022)进行文本纠错；在任务三中，基于TextRCNN-NEZHA模型进行作文流畅性的评级。

2 任务定义

任务一的病句识别，定义为一个多标签分类问题，即对于给定的一句话，预测其包含的病句错误类型。评测任务将病句错误分为两个层次：首先是5类粗粒度错误类型，它们分别为字符级错误，成分残缺型错误，成分赘余型错误，成分搭配不当型错误和不合逻辑错误；其次是14种细粒度错误类型，包括缺字漏字、错别字错误、缺少标点、错用标点、主语不明、谓语残缺、宾语残缺、其他成分残缺、主语多余、虚词多余、其他成分多余、语序不当、动宾搭配不当、其他搭配不当等，病句类型样例如表 1所示。

type	**text**
错别字错误	祥子再度沉论。
缺字漏字	虽然是件小事，可却能影孩子一生。
缺少标点	“聚如一团火，散是满天星”
错用标点	口说无凭，古人云：”以史为鉴“。
主语不明	现在在南京，已多年未见奶奶和姑姑了。
.....	.....
其他搭配不当	你还可以点开微信，无论多远，只要他有网，他都能收到。

Table 1: Task1 病句类型样例

任务二的病句改写，定义为文本生成任务，即给定一句话，在确保原文意图不发生改变的前提下，为中小学生作文中的错误句子提出最小化的修改方案，如表 2所示。自动对错误句子进行修正对于帮助学生理解写作中的问题、提升写作水平具有重要意义。

source	**target**
这些精神，往往会令人感到受益匪浅。	这些精神，往往会令人受益匪浅。
祥子再度沉论。	祥子再度沉沦。
我们秦谁可再会。	我们秦淮河再会。
她是多么地坚强啊！	她是多么的坚强啊！
驰车来到宏村，已是薄暮。	我驰车来到宏村，已是薄暮。

Table 2: Task2 病句改写样例

任务三的流畅性评价，定义为单标签分类问题，即给定一篇作文，根据作文的流畅度，进行打分，分数类别有优秀、一般和不及格，如表 3所示。这样可以方便教师进行最终评分，减

轻教师批改作文的负担，为教师提供一个更加高效和直观的方式来评估学生的写作能力，同时也使学生能够更清晰地了解自己在写作上的表现。

text	level
现实足此岸，理想是彼岸，中间隔着湍急的河流，...	优秀
她的个子不高，眼睛也不大，但是炯炯有神，...	中等
阳光是普遍的，如果没有阳光便就不可能，...	不及格

Table 3: Task3 流畅度评价样例

3 模型

3.1 UTC模型

UTC通过统一语义匹配方式USM（Unified Semantic Matching）来将标签和文本的语义匹配能力进行统一建模。这种方法可以更好的理解多类别文本数据，并从中筛选出正确的类别，如图 1所示。

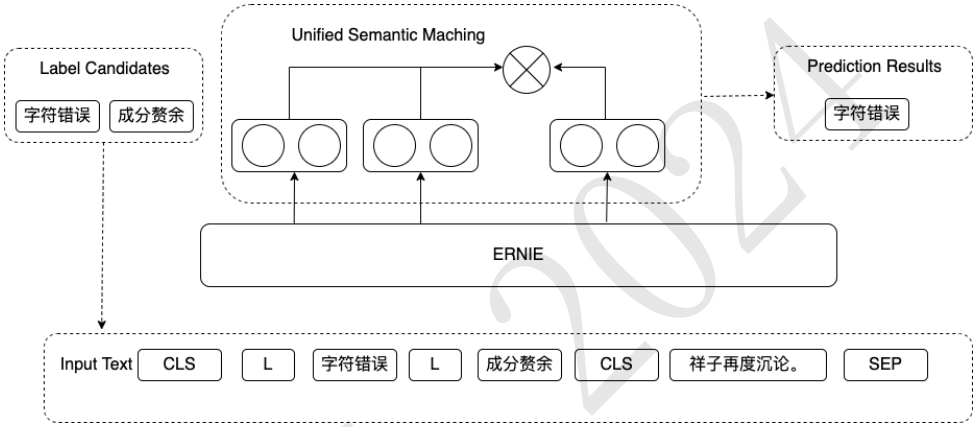

Figure 1: UTC模型架构

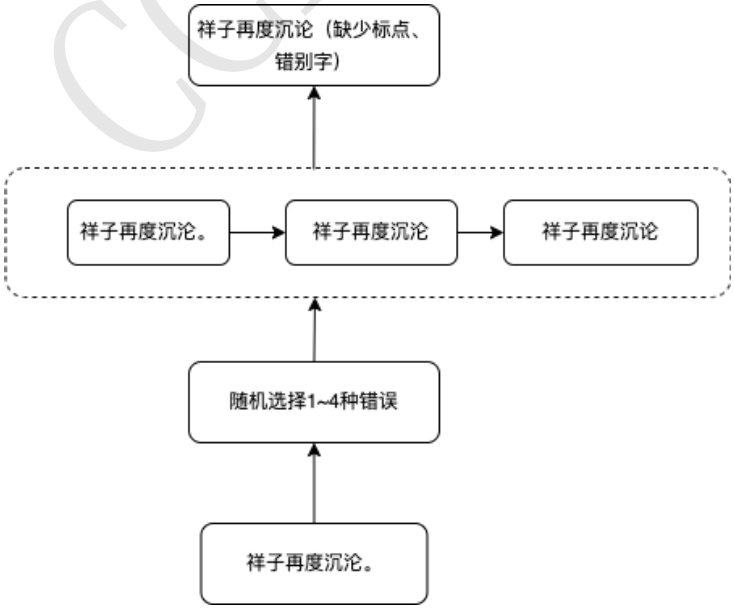

Figure 2: 语法错误替换方法

在Task1任务中，训练集和验证集的数量分别为1237句和104句，但测试集为9040句，在小样本的训练背景下，为增强模型的鲁棒性，我们提出一种语法错误替换方法，对正确语句进行随机的改错来使得模型具有好的泛化能力，如图 2所示，对“祥子再度沉沦。”这句话，从14种错误类型中，随机选择1-4种错误，比如错别字和缺少标点，然后依次进行错别字和缺少标点的改错，最终的语句为“祥子再度沉论”，通过上述的错误替换方法对正确的语句进行数据增强，使用UTC模型进行微调，可以在低资源情况下使得模型获取不错的纠错效果，将提出的语法错误替换方法开源在https://github.com/TW-NLP/CGED_DAT。

面对Task1的病句类型识别任务，本文采用两个阶段训练策略，第一个阶段是使用语法错误替换方法，对开源的公共数据集进行随机的改错，并基于UTC模型来进行微调。第二个阶段，使用提供的真实数据集来进行第二个阶段的微调，使得模型可以更好的检测出真实的应用场景中出现的语法错误。

3.2 BART-SynGEC模型

BART模型的工作机制概括为两个主要步骤，首先，通过对输入文本执行一系列的预处理操作（如删除部分文本、重排句子等），然后，模型试图使用这种破坏的版本来重建原始文本。因为BART模型的预训练任务与文本纠错任务较为相似，为此采用BART作为基础模型。为了有效的捕捉语法结构，因此融合SynGEC方法，有效的将输入句子的句法结构注入BART模型的编码器部分的方法，使得模型可以更好的生成无语法错误的句子，模型框架如图 3所示。

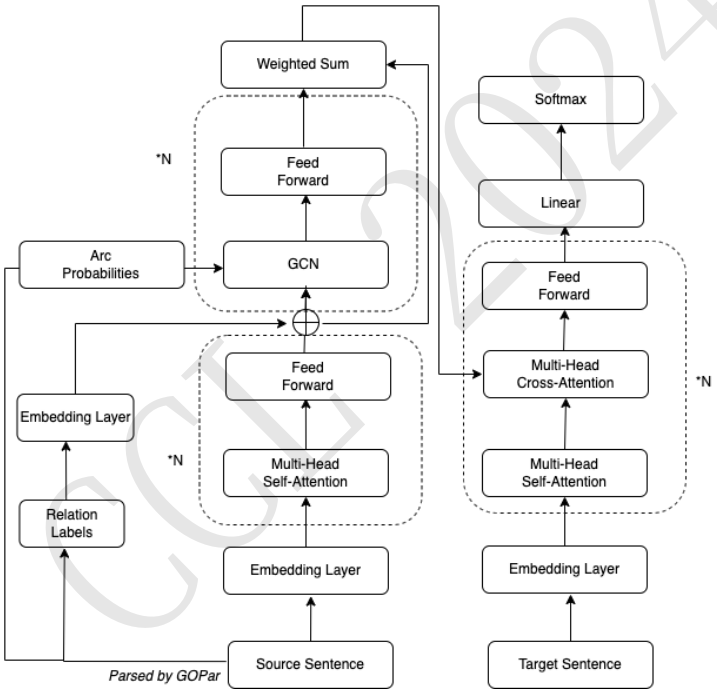

Figure 3: BART-SynGEC模型架构

面对Task2的病句改写任务，本文采用两个阶段的训练策略，第一阶段使用开源的Lang8数据集(Zhao et al., 2018)和HSK数据集(Zhang et al., 2009)进行模型的训练，其中Lang8包含1,220,906条数据，HSK有15,6870条数据，使用通用的数据进行训练，使得模型在通用领域拥有较好的纠错效果。第二阶段，使用Task2提供的训练集进行微调，更好的纠正在中小学生写作的场景下所出现的错误。

3.3 TextRCNN-NEZHA模型

NEZHA模型(Wei et al., 2019)是基于BERT进行了一系列经过验证的改进，其中包括相对位置编码、全字掩码策略、混合精度训练以及训练模型时的LAMB优化器(You et al., 2019)，为了更好的获取整个文章的语义信息，提出TextRCNN-NEZHA模型，如图 4所示。

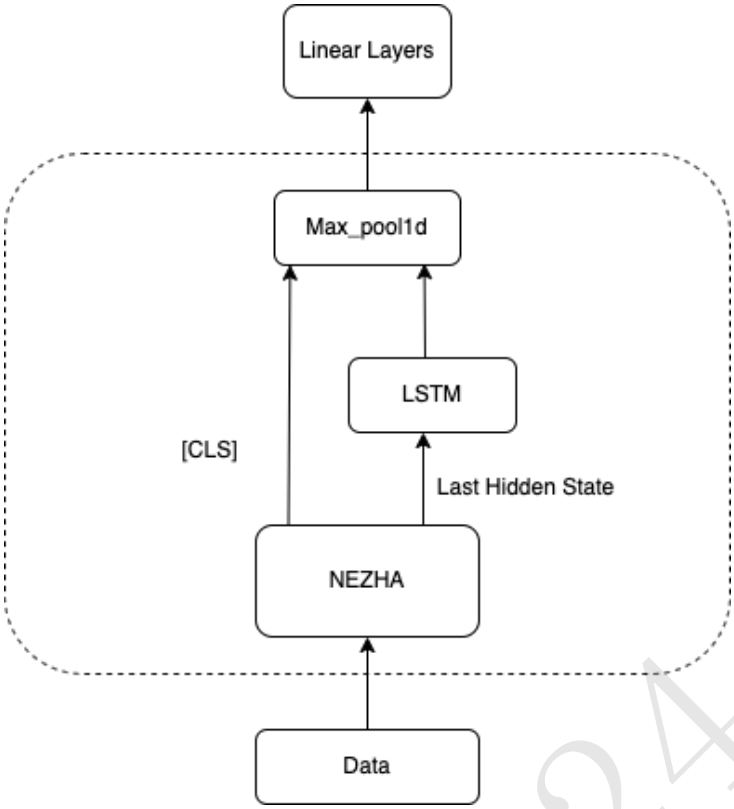

Figure 4: TextRCNN-NEZHA模型架构

面对Task3的流畅度评价任务，TextRCNN-NEZHA模型通过学习[CLS]和词的语义信息，更好的对句子进行语义编码，来捕捉句子流畅度信息。

4 实验

4.1 实验环境

实验所使用的是Ubuntu20.04.2LTS操作系统，用Python作为开发语言，采用Paddle和PyTorch深度学习开发框架。实验采用的CPU为Intel酷睿i7-9700F，GPU为NVIDIA V100，显存32G。

任务一的UTC模型在训练过程中选用AdamW作为优化器，在第一阶段训练的batch-size设置为32，学习率设定为2e-5，maxlen设置为512，epoch为20。第二个阶段训练的batch-size设置为8，学习率设置为1e-5,epoch为20。任务二的BART-SynGEC模型，采用Adam作为优化器，batch-size设置为4，学习率设置为1e-5，maxlen设置为512，epoch为10。任务三的TextRCNN-NEZHA模型，采用AdamW作为优化器，batch-size设置为4，学习率为3e-5，maxlen设置为512，epoch为15。

4.2 评测指标

任务一，粗粒度病句识别分数和细粒度病句识别分数，具体计算方式如下。

$$\text{Score}_{\text{track1}} = 0.5 * F_1^{\text{Course-grained}} + 0.5 * F_1^{\text{Fine-grained}} \quad (1)$$

任务二采用EM(Exact Match)、Bert PPL、与input的编辑距离、BLEU-4、BERTScore以及评估指标(参考MuCGEC)，最终实际排名将综合考虑上述所有指标得到AvgScore，计算方式如下。

$$Score_{track2} = (EM + BLEU + BERTScore)/4 - Levenshtein - PPL_{BERT} \quad (2)$$

任务三采用准确率（Accuracy，Acc）、精确率（Precision，P）、召回率（Recall，R）、F1值（Macro F1）、Quadratic weighted Kappa（QWK）来评估中小学作文流畅性评

级的分类效果，计算方式如下。

$$Score_{track3} = 0.5 * F_1 + 0.2 * QWK + 0.3 * ACC \tag{3}$$

4.3 任务一

针对中小学作文病句类型识别任务，首先基于BERT模型(Devlin et al., 2018)来完成文本分类任务，Micro-F1和Macro-F1的得分为46.36、22.88。为了更好的进行病句的识别，本文使用训练数据对UTC模型进行微调，Micro-F1和Macro-F1的得分为51.83、25.46，相比于BERT分别提升5.47和2.58。为了提升模型的鲁棒性，本文基于语法错误替换方法来进行两个阶段的微调，先在伪数据上使用UTC进行微调，然后用提供的训练数据进行第二阶段微调，最后Micro-F1和Macro-F1的得分为56.42、40.54，相比于BERT提升10.6和17.66，相比于UTC提升4.95和15.08，证明了本文基于UTC和语法错误替换方法的有效性，评测结果如表 4所示。

Model	Micro-F1	Macro-F1	Score
baseline（BERT-base-zh）	46.36	22.88	34.62
UTC	51.83	25.46	38.64
UTC+数据增强	56.42	40.54	48.48

Table 4: Task1实验结果

4.4 任务二

针对病句改写任务，首先基于BART模型在训练数据进行训练，最终得分为35.71，为增强模型的鲁棒性，使用伪数据和Nasgec数据集进行训练，并融合SynGEC进行语法纠错，最后在训练数据集进行微调，相比于BART模型，得分提高11.3，证明了所提方法的有效性，任务二评测结果如表 5所示。

Model	EM	PPL	LS	BLEU	Bert	Pre	Recall	F0.5	Score
BART	8.18	15.72	2.85	85.67	96.85	27.21	23.67	26.42	35.71
BART-SynGEC	19.32	15.68	1.63	91.28	97.86	55.15	33.45	48.81	47.01

Table 5: Task2实验结果

4.5 任务三

针对流畅性评价任务，首先基于BERT进行分类，由于训练集的数量为100条，且测试集为1855条，面对小样本分类数据，我们首先选用TextRCNN-NEZHA模型来作为基础模型，接着引入MulDrop策略(Inoue, 2019)，并使用DCE(Li et al., 2019)来进行损失的计算，评测结果如表 6所示。

Model	Acc	Pre	Recall	F1	QWK	AvgScore
Baseline(BERT-base-zh)	45.52	41.42	39.15	38.57	0.1543	44.48
TextRCNN-NEZHA	51.66	67.98	41.29	38.40	0.2008	46.71
TextRCNN-NEZHA+DCE	52.17	44.27	42.63	41.30	0.2330	48.63
TextRCNN-NEZHA+DCE+MulDrop	51.41	46.24	45.64	45.62	0.2330	50.56

Table 6: Task3实验结果

通过对比，本文提出的TextRCNN-NEZHA+DCE+MulDrop模型相比于Baseline，在F1和QWK分别提升7.05、0.0787，证明所提方法的有效性。

5 总结

针对CCL2024的Task7所提出的三个不同任务，本文提出三种不同的解决方案，在任务一中，采用语法错误替换方法进行数据的增强，并基于UTC进行模型的训练。

在任务二中，采用BART-SynGEC方法进行语法纠错。在任务三中，采用TextRCNN-NEZHA+DCE+MulDrop来提升模型的鲁棒性，经评测，所提出的方法可以有效地识别病句类型和纠正作文中的病句，并能给出合理的作文流畅性评级。

参考文献

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, number 11, pages 13318–13326.

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*.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022. SynGEC: Syntax-Enhanced Grammatical Error Correction with a Tailored GEC-Oriented Parser. *arXiv preprint arXiv:2210.12484*.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. NEZHA: Neural contextualized representation for Chinese language understanding. *arXiv preprint arXiv:1909.00204*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv:1911.02855*.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II 7*, pages 439–445.

Baolin Zhang. 2009. Features and functions of the HSK dynamic composition corpus. *International Chinese Language Education*, 4:71–79.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training BERT in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.