构建面向汉语二语学习者的手写作文语法纠错数据集：方法与应用

李春秋[1]，张潇晓[1]，梁远远[2]，袁丹[2]，兰韵诗[1]，王晔[1]
1. 华东师范大学数据科学与工程学院，上海 200062；
2. 华东师范大学国际汉语文化学院，上海 200062

摘要
针对中文语法纠错研究以印刷体文本数据集为主、缺少手写文本专用数据的现状，本研究分析了构建手写语法纠错数据集面临的关键挑战。据此，构建了面向汉语二语学习者的手写作文语法纠错数据集HCGEC，以真实手写作文为语料，使用优化OCR处理涂改重写等噪声，经过文本识别、标准化标注、图文对齐与质量校验，形成细粒度多模态的标注数据，共包含8 726条有效样本，覆盖599类细粒度错误类型，可为手写场景语法纠错、错误诊断及智能汉语教学提供数据支撑。

关键词
中文语法纠错；手写作文批改；汉语二语；数据集构建；光学字符识别

中图分类号：TP391.41　　文献标志码：A

Constructing Handwritten Grammatical Error Correction for CSL Learners: Methodology and Application

Abstract
Current research on Chinese grammatical error correction (CGEC) mainly relies on printed text datasets and lacks dedicated data for handwritten text. Accordingly, this paper analyzes the key challenges in constructing a handwritten CGEC dataset and develops HCGEC, a handwritten composition grammatical error correction dataset for Chinese as a Second Language (CSL) learners. Using real-world handwritten compositions as the source corpus, it adopts optimized OCR to handle noises such as crossing-out and rewriting, and forms fine-grained and multimodal annotated data through text recognition, standardized annotation, image-text alignment, and quality verification. The dataset contains 8, 726 valid samples covering 599 fine-grained error types, which can provide data support for grammatical error correction, error diagnosis, and intelligent Chinese language teaching in handwritten scenarios.

Key words
Chinese grammatical error correction, handwritten composition correction, Chinese as a second language, dataset construction, optical character recognition

0 引言

0.1 研究背景与意义

在大数据与人工智能技术深度融入教育领域的背景下，自然语言处理相关任务应用场景广泛，领域整体处于高速迭代、前沿活跃的发展与研究阶段。与此同时，智能教育、商业智能等各类实际应用场景，也为自然语言处理领域带来一系列全新技术难题，如何提升自然语言处理模型性能已成为科研工作者重点关注的课题。毋庸置疑，自然语言处理模型性能的提升高度依赖于高质量、场景适配的标注数据集，数据集的类型划分、研究对象界定及错误类型标注的优劣，直接决定了模型的训练效果与泛化能力。现阶段，数据要素市场化配置与教育数字化转型已成为国家重要战略方向，面向特定领域、特定场景的高质量数据集建设，是推动智能教育技术实用化、规模化的关键支撑。

随着国际汉语教育的快速发展，汉语学习者的规模持续扩大。写作能力作为语言综合运用能力的重要体现，在汉语教学中占据着核心地位。作文教学历来是汉语教学中的重点和难点，因为写作是一项高度综合性的语言技能，它要求学习者在掌握了汉字的书写规范、积累了一定的词汇量的基础上，准确运用语法语义，形成连贯有逻辑性的表达。对于汉语学习者而言，写作能力的培养无疑是跨越语言障碍的核心挑战，因而常被视为汉语教学的最终目标。

要实现写作能力的提升，及时、有效的反馈机制至关重要[1]。学习者只有清晰了解自己作文中的错误和不足、明确纠正的方向和方法后，才能在实践中不断进步。然而，在当前的汉语教学实践中，作文批改主要依赖教师人工完成。面对日益增长的留学生数量，教师需要投入大量时间逐字逐句审阅学生作文，标注错误并给出修改建议。这不仅耗费教师大量精力，导致反馈周期较长，而且受教师主观经验影响，批改标准难以统一。

语法纠错（Grammatical Error Correction, GEC）作为自然语言处理领域的重要任务，功能是对含噪声的输入语句进行自动识别与纠错，并给出修改后的结果[2, 3]，是实现智能作文批改的核心技术。在 GEC 领域，英文语法纠错（English Grammatical Error Correction, EGEC）的研究起步较早、发展成熟，目前已形成海量高质量标注数据集[4–10]。中文语法纠错（Chinese Grammatical Error Correction, CGEC）能够定位中文语句中的语法错误并提供修正建议[11]，将教师从繁重的批改任务中解放出来，使其能够更加专注于写作策略指导和个性化教学，是实现智能作文批改的核心技术。然而现有 CGEC 数据集多面向印刷体文本构建，缺乏针对手写文字的专用数据，尤其是面向汉语二语学习者的手写文字数据集。

手写作文是汉语教学中的核心载体，更贴近学习者真实书写习惯，但书写过程中的涂改、重写和字迹潦草等问题会导致通用 OCR 模型误识别与漏识别噪声。若直接将含 OCR 噪声的文本输入语法纠错模型，会使模型在学习真实语法错误之外，还被迫学习了由 OCR 错误产生的“伪错误”，导致模型训练偏离任务本质、性能大幅下降。因此，处理手写噪声、优化 OCR 并构建专用数据集，是提升纠错模型性能、使作文批改技术更好落地的关键。

为此，本研究构建面向汉语二语学习者的手写作文语法纠错数据集

（Handwriting Chinese Grammatical Error Correction, HCGEC），为手写场景语法纠错、错误诊断、汉语二语教学提供高质量数据支撑，推动中文智能教育技术落地应用。

1 研究现状与问题

中文语法纠错作为自然语言处理领域的重要基础任务，其模型性能的提升高度依赖于高质量、场景适配的标注数据集。数据集的类型划分、研究对象界定及错误类型标注，直接决定了模型的训练效果与泛化能力。目前，CGEC领域已涌现出多个具有代表性的公开数据集，涵盖母语者与汉语作为第二语言（Chinese as a Second Language, CSL）学习者两大核心场景。

（1）CCTC数据集

CCTC（Cross-Sentence Chinese Text Correction Dataset）由哈尔滨工业大学社会计算与信息检索研究中心（哈工大SCIR）王宝鑫、段兴义、车万翔等研究人员与科大讯飞研究院伍大勇、陈志刚、胡国平合作构建[12]。CCTC属于母语者中文语法纠错类数据集，其研究对象为汉语母语使用者。数据来源于互联网上母语者撰写的新闻、博客、日常创作、学术草稿等多领域文本，研究团队人工标注了1 500篇文章，包含30 811个句子及逾100万个中文汉字。核心聚焦于母语者书写过程中因疏忽、笔误或表达不规范产生的错误，区别于二语学习者的规则性错误。该数据集的错误类型相对集中，主要划分为四类：一是拼写错误，多为拼音输入或书写疏忽，这也是母语者最常见的错误类型；二是冗余错误，即句子成分多余；三是缺失错误，即句子核心成分缺失；四是乱序错误，即句子成分语序不当。与二语者数据集相比，CCTC的错误更具随机性与场景真实性，且需结合上下文跨句子纠错，填补了CGEC领域母语者真实纠错场景数据集的空白。

（2）NLPCC 2018 Shared Task

NLPCC 2018共享任务第七届CCF国际自然语言处理与中文计算会议（The 7th CCF International Conference on Natural Language Processing and Chinese Computing, NLPCC 2018）的官方评测任务（Shared Task），共设置8个共享任务，涵盖传统NLP问题和新兴应用问题，例如对话系统、问答、用户建模等。该任务训练集主要来源于语言学习社区Lang-8[13]，测试集取自北京大学中国学习者语料库（PKU Chinese Learner Corpus），经过了人工标注与校对，构建了错误句与对应正确句的平行语料[14]。该数据集以汉语二语学习者的写作为基础，覆盖词汇误用、成分缺失、语序不当、搭配错误等多种典型语法偏误类型。测试集规模约为2 000句，并采用多参考标注机制，即同一错误句可对应多个合理修改结果，从而更贴近真实语言教学中的多样化纠错需求。该数据集采用M^2 Scorer作为评价指标[15]，已成为中文语法纠错领域广泛使用的数据集之一。

（3）CGED数据集（Chinese Grammatical Error Diagnosis, CGED）

中文语法错误诊断（Chinese

Grammatical Error Diagnosis, CGED）是自然语言处理技术教育应用研讨会NLPTEA旗下的官方评测任务，作为中文语法纠错领域长期使用的权威基准数据集，其语料全部来自汉语二语学习者的真实写作文本，并经过语言学专业人员的人工标注与多轮校对，形成错误句子与正确句子对应的标准平行语料。该数据集聚焦留学生写作中的典型语法问题，统一界定拼写错误、词语冗余、词语缺失、语序不当四类核心错误，覆盖词汇、语法、搭配、虚词使用等常见偏误类型，能够全面评估模型对中文语法错误的识别、定位与诊断能力。CGED数据集采用多参考标注机制，同一错误句子可对应多个合理修正结果，更贴合真实教学中的批改习惯，长期以精确率、召回率及F1值作为官方评价指标，在中文语法纠错、偏误分析、智能教育评测等相关研究中被广泛采用，是该领域极具代表性的基准数据之一[16]。

除了以上数据集外，还有面向母语者的数据集MuCGEC[17]、YACLC[18]和NaCGEC[19]；另有数据集如FlaCGEC[20]和FCGEC[21]，专注于细粒度语法错误标注（例如，形容词误用、名词缺失）。但现有数据集多以印刷体文本为主要来源，手写文本数据较为匮乏，同时错误类型多为粗粒度标注、标注标准普遍不统一、缺乏图文对齐数据与书写痕迹信息，因此难以支撑精细化错误诊断，跨场景泛化能力有限。对于中文作为第二语言的学习者而言，手写作文是研究其语言偏误的宝贵资源。当前高质量数据集建设已有成熟方法论可循，相关研究从架构设计、构建流程等方面形成了完整路径[22]。而将这些手写材料数字化并构建高质量的CGEC数据集面临着以下几方面的严峻挑战：

首先，手写场景与规整的印刷体或工整的课堂听写不同，学习者的手写作文往往包含大量的修改痕迹，如图1所示，包括使用笔迹将错误内容完全覆盖或涂抹；用横线或斜线将错误内容划掉的痕迹；在划去内容的上方、下方或行间空白处插入修正后的文字；重写字迹与原划掉字迹在空间上存在重叠或紧邻。

其次，现有的通用OCR引擎，如Tesseract、百度OCR开放平台等，主要针对规整的文档图像进行优化，对于上述复杂的手写编辑模式识别能力有限。当面对涂改痕迹时，这些引擎往往会产生误识别和漏识别问题。误识别会将划掉的痕迹错误地识别为有效字符，引入虚假噪声。漏识别可能忽略位于非常规位置（如行间）的重写内容，导致文本缺失或位置错误。

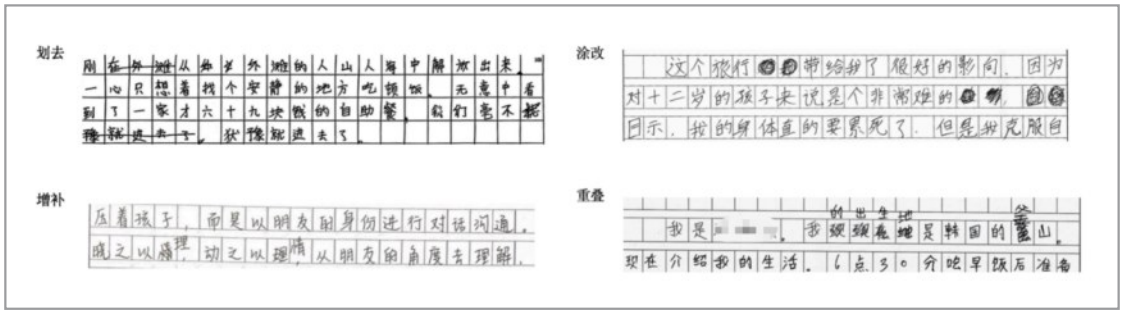

图1 修改痕迹分类

最后，由于真实手写作文普遍存在涂改、划去、重写、重叠等复杂编辑痕迹，现有通用OCR模型难以精准识别，易产生误识与漏识问题。若将含噪声的OCR结果直接输入语法纠错模型会产生严重的级联误差（Cascading Errors）[23]。纠错模型不仅需要学习真实的语法错误，还可能被迫去“纠正”那些由OCR识别错误产生的“伪错误”。这种错误的叠加效应会严重偏离语法纠错的任务本质，导致模型在实际应用中性能大幅下降。

因此，如何在充满噪声的手写图像中，准确恢复出作者最终意图表达的文本序列，是构建高质量的CGEC数据集的前提。

HCGEC（Handwriting Chinese Grammatical Error Correction）数据集构建的核心语料来源于外国留学生在大学汉语课程中的手写作文。这些留学生来自全球30余个国家，涵盖汉语水平从初级（HSK 3级）至高级（HSK 6级）[24]的多个阶段，确保了语料在语言能力维度上的多样性。作文体裁经过系统设计，全面覆盖记叙文（写人、记事）、议论文等常见写作类型，每类体裁下设多个具体话题，如“我的生活”“我的名牌观”“我经历过的营销”等，力求贴近留学生的真实表达需求与学习场景。

数据集的构建流程分为数据采集模块、标注工具模块、数据处理模块、质量验证模块四大核心模块，如图2所示。

2 数据集构建方法

2.1 方法流程

图2 数据集构建总体框架图

2.2 数据采集

原始数据的采集遵循严格流程。在课程教师的指导下，留学生在标准作文纸上完成手写作文，并上传至手写作文在线批改平台。由一线汉语教师及语言学背景的研究生助教在平台上进行批改，平台部署了优化后的OCR模型，在标注人员使用批改功能前，首先调用优化的OCR模型对原始作文图片进行识别，获取每个文字的文本内容及其在图片中的精确坐标（包括所在行的左右边界与上下边界）。

本研究所有手写作文数据均来自课程日常作业，已事先获得留学生书面知情同意，数据采集遵循自愿参与原则。研究过程中严格隐去姓名、学号等个人隐私信息，仅保留HSK等级、国籍等非识别性教学研究信息，符合学术研究伦理与数据安全规范。

2.3 标注工具

标注人员由具备汉语国际教育背景的一线教师与语言学专业研究生助教组成，经统一标注规范培训后开展标注工作，确保标注一致性与专业性。标注人员登录平台后，选择一份待批改的学生作文。整体标注流程如图3所示，首先对作文图片进行预处理（旋转、裁剪等），保留待批改作文部分图片，然后标注人员使用标注工具进行标注。

为采集准确、统一的纠错数据，并协调不同教师批改方式的差异，本研究设计并实现了一套标准化在线标注工具，用以规范批改流程。在熟悉使用后，该工具不仅帮助教师显著提升批改效率，也确保了批改结果的一致性。

通过调研和分析教师传统纸质批改方式，可将批改痕迹归纳为增加、修改和删除三种基本操作类型。据此，系统提供了三类标注工具：增加工具，用于在需要添加内容的位置绘制插入箭头；修改工具，用于在需修改的区域绘制矩形框；删除工具，用于在需删除内容处绘制删除横线。传统纸质批改中，多种批改符号叠加易造成卷面杂乱，影响学生理解。而基于网页的标准化批改工具使批改痕迹更加清晰可辨，有助于师生沟通，也为后续数据采集提供了规范基础。

标注绘制基于鼠标事件实现交互。当标注人员在图片上按下鼠标左键，系统记录当前坐标为起始点；拖动过程中持续捕获鼠标位置，更新终点坐标；松开鼠标后，系统根据当前激活的工具类型，结合起始与终点坐标生成对应的几何图形（如删除工具生成直线），并将标注显示在顶层画板。所有标注均基于标注人员实际拖动轨迹生成，保证了操作的直观性。

若要对已有的标注进行编辑或删除，系统提供了选择工具。标注人员激活选择工具后点击某标注区域，系统捕获点击坐标，遍历已有标注的覆盖范围，通过几何算法判断选中目标；若多个标注被命中，则选择面积最小的标注，以提升选中精度。系统采用双色编码增强标注状态的可视性：已保存的标注显示为红色，编辑中的标注显示为蓝色，便于标注人员区分操作状态与结果，如图4所示。

标注人员标记错误区域后，可在页面右侧填写详细标注信息，包括修改后的正确文本及对应的错误类型。每个句子可能包含多个错误，故支持多标签标注。随后，本研究构建了一套错误类型体系，将偏误分为汉字偏误、词汇偏误和语法偏误三大类，共计599条具体类型。每条类型包含五级分类信息，例如语法偏误中：“语法点错用/多用/漏用”—“词类”—“动词”—“能愿动词”—“想、要”。为便于标注人员选用，该系统将错误类型呈现为树形下拉菜单，支持按关键词搜索，每条类型附有例句参考，显著提升了批改的一致性与准确性。

在实际投入使用中发现，由于每个人在写作中的语言习惯具有高度一致性，同样的语法错误很可能在一篇作文中重复多次出现。针对这一特点，平台提供了批量标注功能，可以同时画出作文中的多处标注，再统一进行一次错误类型的选择。这

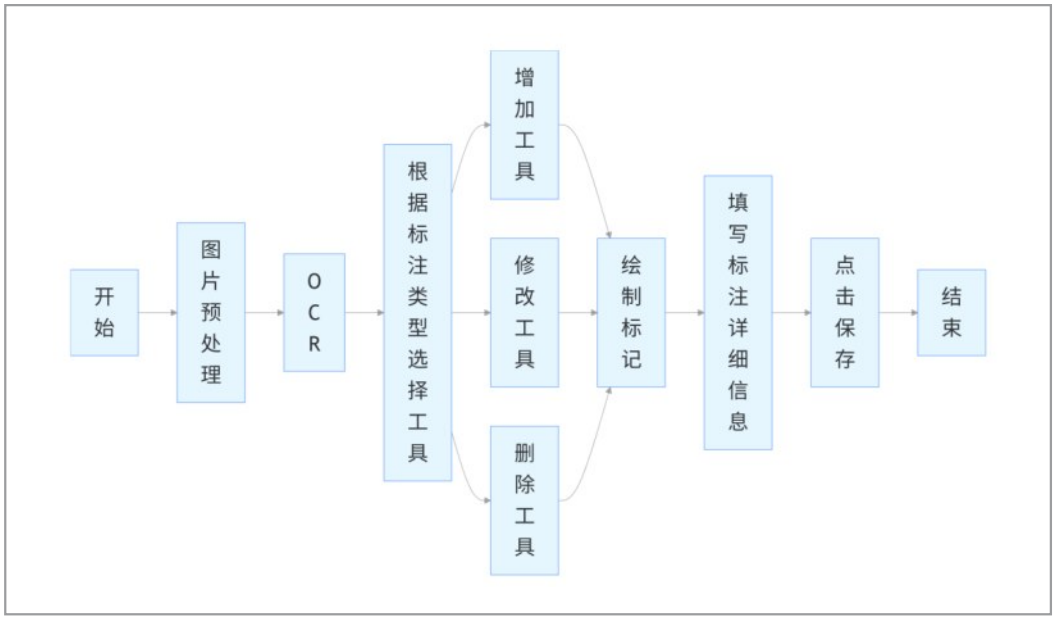

图 3　整体标注流程

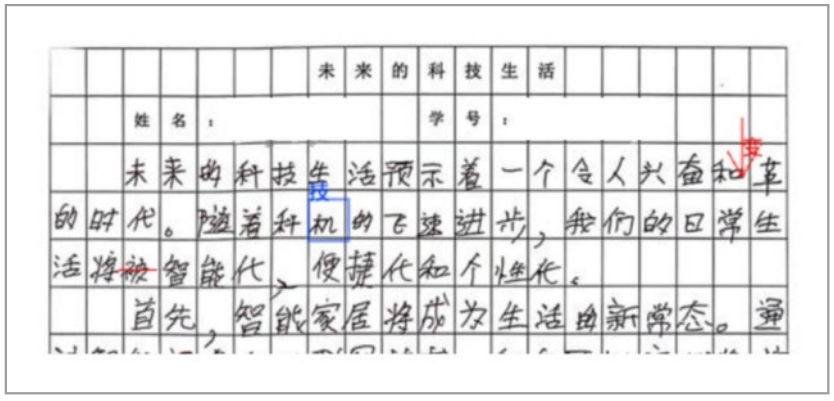

图 4　双色编码

种方式能有效提高人工批改效率，保证良好的用户体验。例如，当学生在一篇作文中多次出现“把”字句误用时，标注人员只需识别出所有相关位置，然后统一标注为“语法偏误-语序错误-把字句误用”，显著减少重复操作、降低标注成本。

2.4 数据处理

数据处理模块的核心目标是从含涂改、划去、重写等复杂修改痕迹的手写作文图像中，准确恢复学习者最终表达的文本序列，为后续语法纠错提供高质量文本输入。针对手写场景中修改痕迹噪声多、通用 OCR 识别精度低、易引入级联误差等问题，本研究基于 PaddleOCR 框架进行针对性改进，提出面向“划去重写”场景的注意力增强 OCR 方法。

本研究选用 PP-OCRv3 系列预训练模型作为改进基准，检测基准模型 ch_PP-

OCRv3_det_distill_train 采用基于可微分二值化（Differentiable Binarization, DB）的检测算法，并结合知识蒸馏（Distillation）策略训练，能够精准定位图像中的文本区域；识别基准模型 ch_PP-OCRv3_rec_train 采用 CRNN（Convolutional Recurrent Neural Network）架构，结合 CTC（Connectionist Temporal Classification）损失函数，能够将检测出的文本区域图像高效转换为字符序列。基于此基准，在文本检测阶段引入多尺度注意力增强机制，构建更适应手写修改痕迹的文本检测模型；在文本识别阶段，增加注意力解码分支，与CTC分支形成多任务学习框架，使模型能够动态聚焦有效书写区域，抑制划去、重叠等噪声干扰。

（1）基于多尺度注意力的文本检测优化

本研究在检测模型的特征金字塔网络（FPN）与路径聚合网络（PAN）的特征融合阶段，引入轻量级卷积块注意力模块（CBAM），构建多尺度注意力增强机制。该模块通过通道维度和空间维度的双向加权，精准捕捉有效文本特征响应，抑制划去痕迹等干扰，从而解决实例混淆并适配位置偏移。

改进后的检测流程如图5所示，在FPN生成多尺度特征图{P_2, P_3, P_4, P_5}并经PAN聚合得到{F_2, F_3, F_4, F_5}后，我们在每个尺度的特征图上串联接入CBAM模块。CBAM采用通道注意力和空间注意力的串行结构。

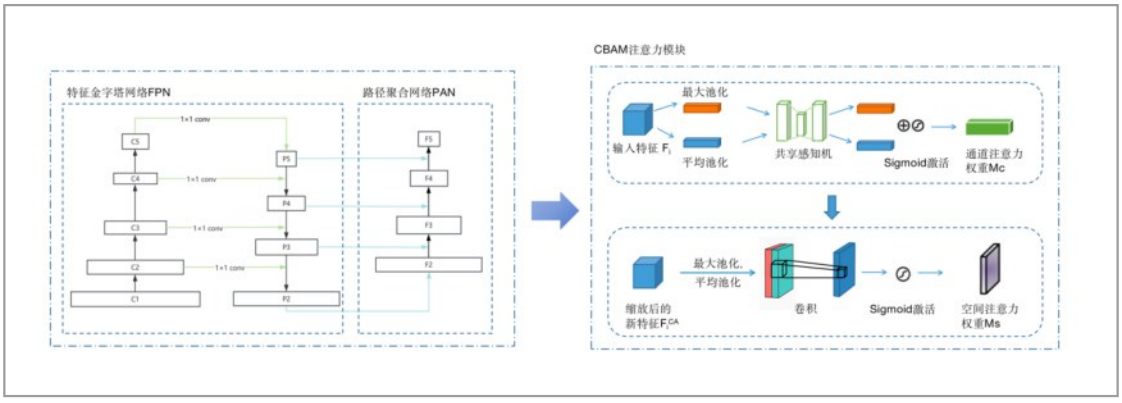

图5 引入多尺度空间注意力的检测模型架构图

（2）基于注意力解码的文本识别优化

传统CRNN+CTC架构缺乏对输入图像空间位置的显式关注，难以在“划去后重写”场景中动态聚焦有效区域。为此，本研究在CRNN的BiLSTM层后引入基于注意力机制的解码分支，与原有CTC分支构成多任务学习框架，如图6所示。

为验证上述方法在真实手写作文场景的有效性，本研究构建包含涂改、划去、重写等典型修改痕迹的专用测试集，从文本检测与文本识别两个维度开展对比实验。各模型在测试集上的性能对比如表1、表2所示。

从实验结果可以得出微调的有效性。无论是检测模型还是识别模型，对比模型A和模型B，在专用数据集上进行微调后，

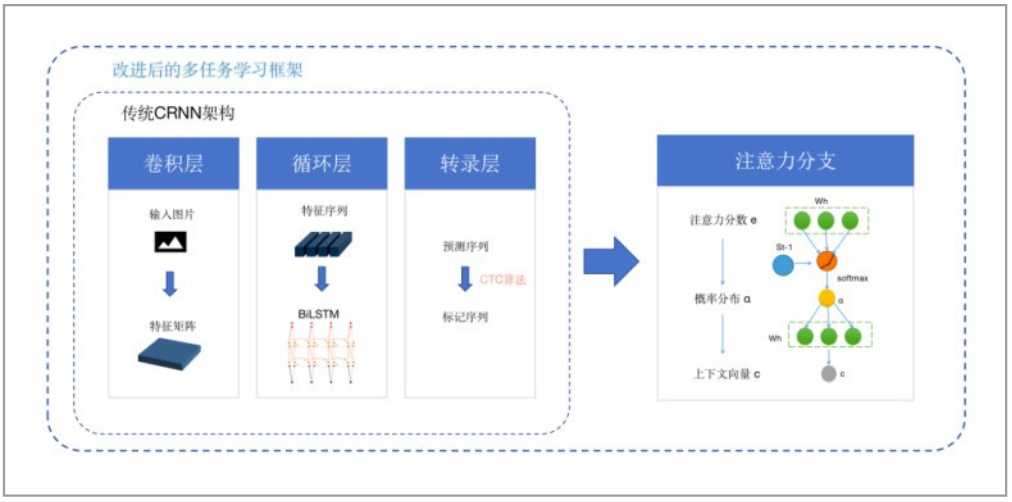

图 6　引入注意力分支的多任务学习框架图

模型的性能数据都有明显提升。这表明，即使不改变模型结构，让模型见过更多类似的复杂样本，也能显著提升其在特定场景下的表现。这验证了构建专用数据集的重要性。

注意力机制的有效性也得到了体现。对比检测模型模型 B 和模型 C，引入多尺度注意力机制后，精确率和召回率均有显著增长。识别模型的准确率和归一化编辑距离也有明显提升。这表明注意力机制确实帮助模型更好地定位和识别有效文本区域，同时减少了对干扰痕迹的误识别，证明了空间注意力机制在处理空间错位文本时的优越性。

表 1　不同检测模型在测试集上的性能对比

模型	精确率 (%)	召回率 (%)	F1 分数 (%)
模型 A（未微调基准）	68.35	61.83	64.92
模型 B（微调基准）	82.43	72.43	77.11
模型 C（本文方法）	89.32	82.18	85.60

表 2　不同识别模型在测试集上的性能对比

模型	准确率 (%)	归一化编辑距离 (%)
模型 A（未微调基准）	3.21	79.53
模型 B（微调基准）	9.37	85.31
模型 C（本文方法）	12.84	89.36

但整体而言，识别模型的准确率仍然较低，可能的原因是评估指标比较严苛。整句准确率要求预测序列与真实标注完全一致（Exact Match）。在留学生作文中，单行文本通常包含 15 至 22 个字符，而“划去重写”行为往往仅涉及其中 1 至 2 个字符的修正。如果模型正确识别了整行 95% 的字符，仅因未能完美处理 1 个被涂

改的字（例如：多识别了一个被划掉的旧字，或少识别了一个新字），该样本的整句准确率即被计为0。此外，留学生手写汉字的书写风格、潦草程度与通用OCR模型的预训练数据分布存在显著差异，进一步降低了整句识别准确率。归一化编辑距离（NED）能反映模型的实际可用性。模型C的NED较高，这意味着在绝大多数样本中，模型预测的字符序列与真实文本高度相似，平均每行仅存在极少量的字符误差。

综上所述，较低的整句准确率客观反映了“划去重写”场景下实现完美识别的极高难度，而高NED值与准确率的显著相对提升，则有效验证了本研究方法在字符级识别精度上的实质性突破。

借助优化后的OCR模型，系统可获取作文中每个字符的文本内容与精确坐标信息。标注人员完成纠错后，依据他们在图片上绘制的标注位置坐标，结合OCR识别得到的作文每个字符的文本和坐标信息，对于每个句子，通过逻辑判断（根据所在的行、起止字符），确定哪些标注是对这个句子进行纠错。再由标注的更正信息（如修改成“好”，插入“对”，删除坐标部分）就可以得到修正后的正确句子。在此处理过程中，系统记录错误原句子，错误原字词，并截取对应句子的图片切片，实现图文信息的对齐。

完成标注与句子的对齐，得到错误句子和正确句子后，进一步从数据库中提取相关详细信息，包括每条标注的错误类型（与修正方式对应）、学生信息（如国籍、HSK等级）等，最终输出结构化数据条目。为增强数据集的鲁棒性，还抽取了一部分无任何错误的句子作为负例加入数据集。

2.5 质量验证

通过上述自动化流程，初步获得11 475条原始数据。然而，由于OCR识别误差、坐标映射偏差或标注边界模糊等因素，部分数据质量难以保证。因此，还需进行严格的人工筛选环节，逐条校验自动化生成的数据，剔除图片模糊、OCR误识率过高、坐标明显错位或标注内容不清晰的样本，最终保留有效数据8 726条，保证数据集整体质量。

3 HCGEC数据集

3.1 数据集规模与构成

HCGEC数据集包含8 726条数据，每条数据为一个以img_id命名的目录，其中img为作文图片的唯一编号，id为该作文内错误文本切片的编号。因此，同一篇作文的不同错误具有相同的img编号和不同的id编号。每个目录内包含错误句子在原文中的截图和一个包含纠错信息的json文件，作文图片切片样例和json格式如图7、8所示。

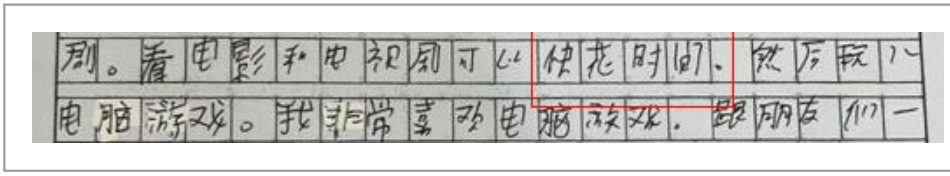

图7 作文图片切片样例

```
{
    "error_sentence": "看电影和电视剧可以快花时间，然后玩儿电脑游戏",
    "correct_sentence": "看电影和电视剧可以打发时间然后玩儿电脑游戏",
    "correct_edition": "打发时间",
    "error_type": ["602_词语偏误_误用"],
    "mark_type": "edit",
    "start_x": 328,
    "start_y": -3,
    "end_x": 443,
    "end_y": 42,
    "student_id": "12192720454"
}
```


图 8　json 格式

最终将数据集按 5:1:1 左右比例划分训练集、测试集和验证集：6 108 条作为训练集；1 308 条作为测试集；1 310 条作为验证集。该划分确保了各类错误类型在不同子集中的分布均衡，为后续模型训练与实验评估提供了可靠的数据基础。

3.2 错误类型分布

本研究将数据集的错误类型分为三大类：汉字偏误、词汇偏误和语法偏误，每个类型包含多级分类信息，其中汉字偏误设两级分类，一级 4 类、二级 14 类；词汇偏误设两级分类，一级 5 类、二级 7 类；语法偏误设五级分类，一级 6 类、二级 12 类、三级 37 类、四级 107 类、五级 578 类，共计 599 条具体的错误类型。

经数据清洗与人工校验后，本研究保留了 8 726 条有效手写句子样本。对样本内全部偏误进行频次统计，共计得到偏误项 39 763 个，平均每条句子有 4.6 个错误。每个错误类型末级分类下对应的偏误频次如图 9 所示，汉字偏误二级分类含错误项 8 479 个；词汇偏误二级分类含错误项 14 205 个；语法偏误五级分类含错误项 17 079 个。

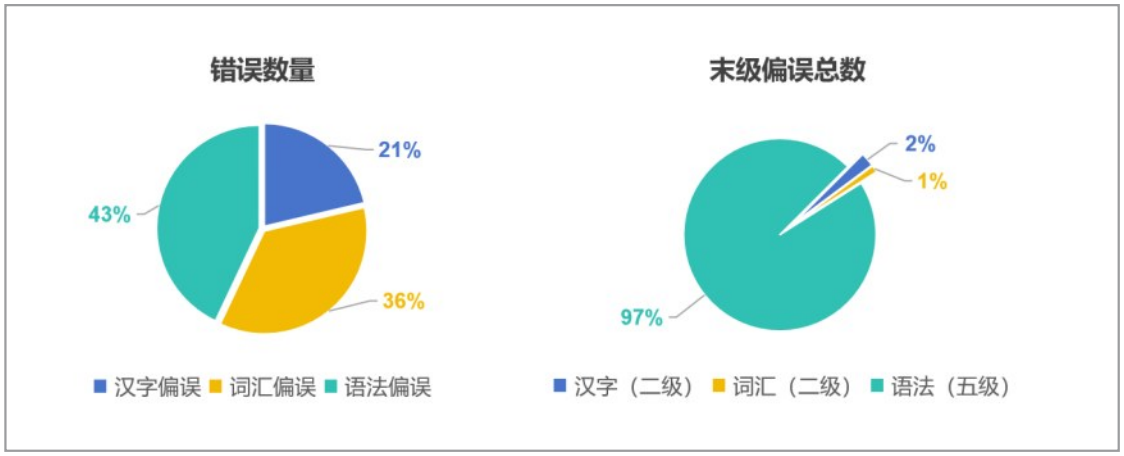

图 9　错误类型分布

3.3 数据集特点与优势

现有CGEC数据集多基于印刷体文本或通用语料，缺少面向留学生手写作文的专用数据集，难以适配国际汉语教育真实教学场景与学习者语言学习特征。与印刷体文本或母语者的书写不同，汉语二语学习者的手写作文具备高度独特性：一是学习者来自世界各地，母语背景差异带来了跨语言类型的书写与语法错误，区别于汉语母语者笔误类错误；二是学习者的汉语水平涵盖HSK多个阶段，初级至高级学习者的汉字书写熟练度与语法掌握程度差异明显，而汉语母语者语言能力成熟，水平差异不显著；三是手写形式保留了真实的书写痕迹，能够体现学习者学习和书写汉字时的思考与修正过程，能展示印刷体文本无法保留的学习过程数据。而本研究构建的HCGEC数据集聚焦二语学习者真实手写场景，填补了手写作文语法纠错数据集的空白，区别于MuCGEC、YACLC、CCTC等以印刷体文本为主的数据集。

HCGEC数据集的核心语料来源于外国留学生在大学汉语课程中的手写作文。这些留学生来自全球30余个国家，涵盖汉语水平从初级（HSK 3级）至高级（HSK 6级）的多个阶段，作文体裁全面覆盖记叙文、议论文等常见类型，话题贴近留学生生活与学习场景，更符合智能汉语教学的实际需求。手写作文更贴近学习者真实书写习惯，能反映汉字书写、涂改、重写等真实问题，是真实教学场景中的核心载体，对全面评估学习者汉语水平至关重要。

HCGEC具备多模态特性，包含手写图片切片、原文句子、修正句子、错误类型、字符坐标、学习者信息等多维结构，区别于传统纯文本数据集，更贴近真实教学场景。

4 数据集应用

4.1 语法纠错模型训练与评估

在数据集构建过程中，首先对原始作文图片进行文字识别与坐标提取，随后系统基于人工标注位置坐标进行逻辑判断，自动完成错误位置定位、修改方式识别及句子级图文对齐，最终获得8 726条高质量数据，为后续语法纠错大模型的微调与检索增强生成优化提供了可靠的数据基础。HCGEC数据集的提出同时也能满足教育、写作、内容审核等真实场景的应用需求，具有重要的研究与实际价值。

本文在构建完成HCGEC的基础上，以通义千问Qwen2.5-7B-Instruct为基础模型，采用LoRA参数微调结合多粒度检索增强RAG的方式，训练得到基于HCGEC数据集训练的语法纠错模型。为验证其有效性，本文将其与当前主流语法纠错模型进行横向对比，包括传统序列模型Seq2Seq、语法纠错经典模型GECToR、生成式模型BART，以及不同设置的Qwen2.5-7B大语言模型（零样本、少样本、LoRA微调）。实验在HCGEC测试集与公开数据集MuCGEC上采用语法纠错标准指标M^2进行评估。

各模型在HCGEC测试集和MuCGEC测试集上的详细M^2评估结果如表3所示。所有指标均通过官方M^2工具计算，采用精确率(P)、召回率(R)和$F_{0.5}$值进行评价，其中$F_{0.5}$值表明在CGEC任务中精确率两倍重要于召回率。

从实验结果可以看出大模型方法的优势，基于大模型的方法（Qwen系列）经过微调后整体优于传统序列模型。在HCGEC上，LoRA微调模型相比BART在精确率提升尤为显著，表明大模型在纠错准确性上具有明显优势。LoRA微调模型相比零样本和少样本都有所提升，表明针对CGEC任务的参数高效微调能够显著提

表 3 各模型在测试集上的 M^2 评估结果(%)

模型	HCGEC			MuCGEC		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Seq2Seq	42.56	28.18	38.62	38.42	23.51	34.10
GECToR	46.24	34.36	43.26	45.58	35.95	43.25
BART	39.18	27.42	36.08	37.02	24.40	33.55
Qwen2.5-7B(zero-shot)	32.43	30.87	32.10	28.76	28.34	28.67
Qwen2.5-7B(few-shot)	40.92	35.08	39.60	38.15	34.55	37.37
Qwen2.5-7B-LoRA	49.67	46.53	49.01	45.84	40.96	44.78
Ours(LoRA+多粒度RAG)	53.76	49.14	52.76	47.98	44.52	47.24

升模型对错误修正的准确性。引入 RAG 后，模型精确率和 $F_{0.5}$ 值都有所提升，表明多粒度设计在保持高精度的同时有效提升了错误召回能力。在外部测试集 MuCGEC 上，本文方法同样取得最优结果，相比基线模型提升显著。值得注意的是，所有模型在 MuCGEC 上的 $F_{0.5}$ 值均略低于 HCGEC，符合领域适应的预期，但本文方法的下降幅度最小，验证了方法的泛化能力。

4.2 错误类型识别与诊断

本数据集的错误类型体系较为完整，将偏误分为三大类，包含汉字偏误、词汇偏误和语法偏误，每条类型至多包含五级分类信息，例如语法偏误中："语法点错用/多用/漏用"—"词类"—"动词"—"能愿动词"—"想、要"。为便于标注人员使用，系统为错误类型设计了树形下拉菜单，支持按关键词搜索，每条类型附有例句参考，提升了批改的一致性与准确性。如图 10 所示，可以看到选中的标注有详细错误信息，包括错误类型信息、如何修改等。

图 10 详细错误信息

利用 HCGEC 数据集的细粒度错误标注，能够对数据集中标注的偏误进行量化分析，从而刻画不同母语背景、不同汉语水平学习者的偏误分布与典型偏误；统一

的错误分类标准能够支持不同作文样本、不同学习者群体之间的对比分析，让研究结论更客观，为国际汉语相关研究提供可靠的数据支撑。

4.3 个性化学习反馈生成

当标注人员完成对某篇作文的批改后，学生可登录平台查看已批改作文的详细反馈。平台支持两种反馈形式：一是可视化标注视图，即在作文原图之上叠加标注人员绘制的增加、修改、删除等标注图形，学生可直观看到错误位置及修正方式；二是结构化反馈列表，点击每个标注在页面右侧显示每个错误点的错误类型、修正方式。

从学生反馈来看，平台提供的详细批改结果和错误类型分析获得了积极评价。学生可以清晰地看到自己的错误位置、错误类型和正确表达方式，并能够查看历史批改记录，追踪自己的学习进步情况。从教师角度而言，若基于HCGEC数据集进行量化分析，可以帮助教师更直观地总结教学中的重难点，便于教师开展针对性教学、提供个性化学习反馈。

5 结论与展望

5.1 研究总结

本研究面向国际中文教育数字化需求，聚焦汉语二语学习者手写作文场景，构建了高质量、细粒度、图文对齐的HCGEC中文语法纠错数据集，填补手写场景数据集空白；研究以真实教学手写作文为语料来源，提出一套可复用的手写作文数据集构建流程，包括OCR优化、标准化标注、坐标对齐、图文配对等关键技术；建立包含599类细粒度错误类型的标注体系，形成包含8 726条有效样本的手写作文语法纠错数据集，具备图文对齐、细粒度标注、场景真实等特点，为语法错误诊断与教学分析提供标准化支撑；形成可用于模型训练、教学研究、语言分析的多模态数据集，具备广泛应用价值。

5.2 研究局限

HCGEC数据集虽然在标注粒度与手写场景上具有独特价值，但仍存在一定局限性。首先，数据集规模仅为8 726条，相较于大规模公开数据集仍偏小，在支撑超大规模模型训练方面存在不足，容易限制模型复杂错误的识别与纠错能力。其次，数据来源集中于高校留学生课程作文，体裁以记叙文、议论文为主，对应用文、说明文等其他体裁覆盖不足，领域泛化能力受限。此外，数据集中汉语水平以初中级为主，高级学习者作文样本相对较少，各层级样本分布不均衡，难以全面覆盖各水平学习者的错误特征，领域泛化能力受限。

尽管本研究制定了统一的标注规范与错误类型体系，但语法错误标注本身具有一定主观性，不同标注者对部分边缘错误的判定与修改方式可能存在差异。尽管经过多轮审核与筛选，仍无法完全消除标注主观性带来的影响，可能在一定程度上影响模型训练的效果。

5.3 未来工作

本研究已构建了在线的作文提交、批改的智能化平台，因此可以持续扩大HCGEC数据集规模，扩充作文体裁与学习者水平覆盖，增加不同母语背景学习者的样本，提升数据集的多样性与代表性。

本研究当前标注主要聚焦错误文本与

修正文本的对齐，尚未对作文中涂改、划去、重写等痕迹本身进行框选与标注。后续可在此基础上，对各类原始错误痕迹区域进行独立切片留存和标注，有助于研究二语学习者书写纠错的认知过程，丰富数据集的教育研究价值。

同时，当前数据集标注以单人审核校验为主，存在一定主观差异性。后续可采用多副本多人独立标注结合专家仲裁的标注机制，对同一句子同一错误由多名标注员独立标注，分歧样本交由专业教师统一裁定，有效降低标注主观偏差，提升数据集标注的一致性与规范性。

此外，后续可引入主动学习与半自动标注技术，利用模型预标注辅助人工标注，降低标注成本与主观性，提升数据集构建效率。基于数据集与纠错模型进一步开发面向教学的智能化应用，包括错误诊断报告、个性化学习推荐、针对性练习题生成等功能，推动数据集在国际汉语场景落地应用。

参考文献：

[1] Steiss J, Tate T, Graham S, et al. Comparing the quality of human and ChatGPT feedback of students' writing[J]. Learning and Instruction, 2024,91(000).

[2] Grundkiewicz R, Bryant C, Felice M. A crash course in automatic grammatical error correction: Proceedings of COLING: Tutorial Abstracts[C], 2020.

[3] Wang Y, Wang Y, Dang K, et al. A comprehensive survey of grammatical error correction[J]. ACM Transactions on Intelligent Systems and Technology, 2021, 12(5):1-51.

[4] Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading esol texts: Proceedings of ACL[C], 2011.

[5] Dahlmeier D, Ng H T, Wu S M. Building a large annotated corpus of learner english: The nus corpus of learner english: Proceedings of BEA@NAACL-HLT[C], 2013.

[6] Ng H T, Wu S M, Briscoe T, et al. The conll-2014 shared task on grammatical error correction: Proceedings of CoNLL: Shared Task[C], 2014.

[7] Napoles C, Sakaguchi K, Tetreault J. Jfleg: A fluency corpus and benchmark for grammatical error correction: Proceedings of EACL[C], 2017.

[8] Bryant C, Felice M, Andersen E, et al. The bea-2019 shared task on grammatical error correction: Proceedings of BEA@ACL[C], 2019.

[9] Flachs S, Lacroix O, Yannakoudakis H, et al. Grammatical error correction in low error density domains: A new benchmark and analyses: Proceedings of EMNLP[C], 2020.

[10] Napoles C, Nădejde M, Tetreault J. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses[J]. Transactions of the Association for Computational Linguistics, 2019,7:551-566.

[11] Yu L C, Lee L H, Chang L P. Overview of grammatical error diagnosis for learning Chinese as a foreign language: Proceedings of the 22nd International Conference on Computers in Education[C], 2014.

[12] WANG B, DUAN X, WU D. CCTC: A cross-sentence Chinese text correction dataset for native speakers[J]. Proceedings of the 29th International Conference on Computational Linguistics, 2024: 3342-3352.

[13] Zhao Y, Jiang N, Sun W, et al. Overview

of the nlpcc 2018 shared task: Grammatical error correction: CCF International Conference on Natural Language Processing and Chinese Computing[C], 2018.

[14] Wu Y, Zhang M. Overview of the NLPCC 2017 Shared Task: Chinese Word Semantic Relation Classification: Natural Language Processing and Chinese Computing[C], 2018.

[15] Dahlmeier D, Ng H T. Better Evaluation for Grammatical Error Correction: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies[C], Montr é al, Canada, 2012. Association for Computational Linguistics.

[16] Rao G, Yang E, Zhang B. Overview of NLPTEA-2020 Shared Task for Chinese Grammatical Error Diagnosis: Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications[C], 2020.

[17] Zhang Y, Li Z, Bao Z. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction: arXiv preprint[Z]. 2022.

[18] Wang Y, Kong C, Yang L. Yaclc: A chinese learner corpus with multidimensional annotation: arXiv preprint[Z]. 2021.

[19] Ma S, Li Y, Sun R. Linguistic rules-based corpus generation for native Chinese grammatical error correction: Findings of the Association for Computational Linguistics: EMNLP 2022[C], Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

[20] Du H, Zhao Y, Tian Q, et al. FlaCGEC: A Chinese Grammatical Error Correction Dataset with Fine-grained Linguistic Annotation[J]. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023:5321-5325.

[21] XU L, WU J, PENG J. FCGEC: Fine-grained corpus for Chinese grammatical error correction[J]. Findings of the Association for Computational Linguistics: EMNLP 2022, 2022:1900-1918.

[22] 姜春宇, 白玉真, 刘渊, 等. 构建企业级人工智能高质量数据集：方法与路径[J]. 大数据, 2025,11(06):47-56.
JIANG Chunyu, BAI Yuzhen, LIU Yuan, et al. Building high-quality datasets for enterprise-level artificial intelligence: methods and pathways[J]. BIG DATA RESEARCH,2025,11(06):47-56.

[23] Lopresti D. Optical character recognition errors and their effects on natural language processing[J]. International Journal on Document Analysis and Recognition, 2009,12:141-151.

[24] Xun E. Hsk dynamic composition corpus [EB/OL]. http://hsk.blcu.edu.cn/.

作者简介

李春秋，女，硕士在读，华东师范大学数据科学与工程学院，主要研究方向为智能教育。

张潇晓，女，硕士，华东师范大学数据科学与工程学院，主要研究方向为数据挖掘。

梁远远，男，博士，华东师范大学国际汉语文化学院，主要研究方向为自然语言处理、大语言模型以及国际中文教育。

袁丹，女，博士，华东师范大学国际汉语文化学院，副教授，主要研究方向为实验语音学、方言学、社会语言学以及二语习得。

兰韵诗，女，博士，华东师范大学数据科学与工程学院，副教授，主要研究方向为知识图谱，智能问答以及其他与自然语言处理相关的任务。

王晔，男，博士，华东师范大学数据科学与工程学院，专任研究员，主要研究方向为Web数据管理，海量数据挖掘，分布式系统。

收稿日期: XXXX-XX-XX
通信作者:
基金项目:
Foundation Items: