CLTC 2022：汉语学习者文本纠错技术评测及研究综述

王莹莹[1], 孔存良[1], 刘鑫[1], 方雪至[1], 章岳[3], 梁念宁[2], 周天硕[3], 廖田昕[1],
杨麟儿[1]*, 李正华[3], 饶高琦[1], 刘正皓[4], 李辰[5], 杨尔弘[1], 张民[3], 孙茂松[2]
[1]北京语言大学
[2]清华大学
[3]苏州大学
[4]东北大学
[5]阿里巴巴达摩院

摘要

汉语学习者文本纠错（Chinese Learner Text Correction）评测比赛，是第21届中国计算语言学大会附属的第3个技术评测。针对汉语学习者文本，设置了中文拼写检查、中文语法错误检测、多维度汉语学习者文本纠错、多参考多来源汉语学习者文本纠错、语法纠错质量评估五个赛道，发布新的数据集，建立基于多参考答案的评价标准，构建基准评测框架，进一步推动汉语学习者文本纠错研究的发展。共有142支队伍报名参赛，最终14支队伍获得五个赛道的一至三等奖。

关键词： 学习者文本；语法纠错；拼写检查；质量评估

Overview of CLTC 2022 Shared Task: Chinese Learner Text Correction

Yingying Wang[1], Cunliang Kong[1], Xin Liu[1], Xuezhi Fang[1],
Yue Zhang[3], Nianning Liang[2], Tianshuo Zhou[4], Tianxin Liao[1],
Liner Yang[1], Zhenghua Li[3], Gaoqi Rao[1], Zhenghao Liu[4], Chen Li[5],
Erhong Yang[1], Min Zhang[3], Maosong Sun[2]
[1]Beijing Language and Culture University
[2]Tsinghua University
[3]Soochow University
[4]Northeastern University
[5]Alibaba Group

Abstract

Chinese Learner Text Correction (CLTC) is the third shared task attached to the 21st China National Conference on Computational Linguistics (CCL 2022). CLTC shared task sets up five tracks: Chinese Spelling Check, Chinese Grammatical Error Diagnosis, Multidimensional Chinese Learner Text Correction, Multi-reference Multi-source Chinese Learner Text Correction, and Quality Estimation. Moreover, the task released new datasets, established evaluation metrics based on multiple references, and constructed a benchmark evaluation framework for the technology of Chinese learner text correction. A total of 142 teams signed up for the task, and eventually 14 teams won the first to third prizes in the five tracks.

Keywords: Chinese learner text , grammatical error correction , spelling check , quality estimation

* 通讯作者

1 引言

近年来，中外交流合作逐步拓展和深化，中国在扩大开放中融入世界，全球汉语学习需求与日俱增。据教育部中外语言交流合作中心数据显示，目前全球共有180多个国家和地区开展汉语教育[0]，中国以外累计学习中文人数已达2亿[1]。日趋增多的汉语学习者给国际中文教育带来了机遇和挑战，同时也使得技术、方法、理念上的创新成为了迫切的需要。

随着科技的发展与进步，特别是人工智能技术的创新，智能计算机辅助语言学习 (Intelligent Computer-Assisted Language Learning, **ICALL**) 在国际中文教育中的作用越来越突出。其中，汉语学习者文本纠错就是一项重要的应用。

汉语学习者文本 (Chinses Learner Text) 指的是以汉语作为第二语言的学习者在说或写的过程中产出的文本。汉语学习者文本纠错 (Chinese Learner Text Correction, **CLTC**) 旨在通过智能纠错系统，自动检测并修改学习者文本中的标点、拼写、语法、语义等错误，从而获得符合原意的正确句子。

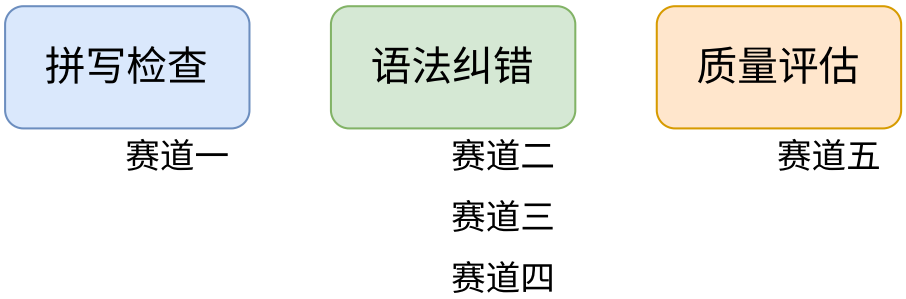

图 1: CCL2022-CLTC 评测任务与赛道

学界关于汉语学习者文本纠错已经开展了多方面、多角度的研究，如拼写检查 (Spell Check, SC)、语法纠错 (Grammatical Error Correction, GEC)、语法错误检测 (Grammatical Error Detection, GED) 等，也已发布有一些相关的评测任务。拼写检查任务关注句子中的字级别错误，例如由手写或 OCR 导致的错别字。SIGHAN 于2013-2015 年举办了三次拼写检查评测任务 (Wu et al., 2013; Yu et al., 2014b; Tseng et al., 2015)。NLP-TEA 也于 2017 年举办过评测任务 (Fung et al., 2017)。上述这些拼写检查任务均使用了繁体中文数据集。语法纠错任务受到关注较多，研究者众 (张生盛et al., 2021; Wang et al., 2021; Zhang et al., 2022; Yang et al., 2022)。2018 年，NLPCC 会议举办有中文语法纠错比赛 (Zhao et al., 2018)，吸引了许多研究者参与。这些研究与评测对中文语法纠错进行了不断的探索与推进。不过，这些工作使用的数据集、任务设置、评价指标等均存在差异，不利于各研究之间的横向对比。相比于语法纠错，语法检查只要求找到句子中出现错误的位置。自 2014 年起，NLP-TEA (Workshop on Natural Language Processing Techniques for Educational Applications，用于教育应用的自然语言处理技术) 已经举办了六次语法检测评测任务 (Yu et al., 2014a; Lee et al., 2015; Lee et al., 2016; Rao et al., 2017; Rao et al., 2018; Rao et al., 2020)，且自 2018 年始加入了进行语法纠错的任务要求。

延续上述汉语学习者文本纠错研究，我们在 CCL 2022 会议上举办了本次评测。如图 1 所示，本次评测设置有五个赛道，覆盖了拼写检查、语法纠错、质量评估三个任务。相比于之前的研究，本次评测有以下几点特色。

首先，就关注者最多的语法纠错任务，将现有资源整合汇聚于赛道二、三、四。其中，赛道二要求对留学生在汉语水平考试 (HSK) 作文中出现的错误进行检测、纠正，并首次公开了历年 CGED 评测数据[2]用于训练和开发。赛道三关注语法纠错中的多维度问题，即从最小改动 (Minimal Edit) 和流利提升 (Fluency Edit) 两个方面给出多种句子修改方案，使用 YACLC 数据集 (Wang et al., 2021) 用于开发和测试。赛道四关注文本纠错的多来源问题，考虑不同来源的文本中可能蕴含的不同类型的语法错误，使用 MuCGEC 数据集 (Zhang et al., 2022) 用于开发和测试。

其次，为进一步推进中文拼写检查研究，本次评测基于 YACLC 数据集 (Wang et al., 2021) 构建并公开了 YACLC-CSC 数据集，并作为赛道一的开发和测试数据。据我们所知，该数据集为首个简体中文拼写检查数据集。

[0]数据来源：http://www.chinese.cn/page/#/pcpage/article?id=714
[1]数据来源：http://www.chinese.cn/page/#/pcpage/article?id=352
[2]历年CGED 评测数据：https://github.com/blcuicall/cged_datasets

最后，扩展了文本纠错任务，首次将质量评估 (Quality Estimation, QE) 纳入评测任务。在使用不同语法纠错方法，或基于柱搜索 (Beam Search) 获得多种修改方案后，质量评估 (Liu et al., 2021b) 任务要求评估不同修改结果的质量。该任务可用于模型集成或其他情况下的结果重排序 (Re-Ranking)，可以在不改变模型的情况下明显提升修改效果。然而目前该任务所受关注较少，评测组织方倡议学界对这一任务多加重视。

本报告主要包含如下内容：第 2 节主要介绍拼写检查、语法纠错和质量评估任务，包括任务要求和各任务相关工作。第 3 节主要介绍五个赛道的具体设置，例如数据集、评价指标、基线模型等。第 4 节介绍本次评测的参赛和获奖情况。第 5 节分别展示各赛道参赛者所使用的方法进行总结分析。第 6 节对本次评测进行总结。

2 任务介绍

2.1 拼写检查

拼写检查 (Spelling Check, SC) 旨在检测并纠正文本中的拼写错误。由于汉字及汉语的特殊性，区别于英文拼写检查，中文拼写检查 (Chinese Spelling Check, CSC) 一般作为独立任务进行研究。如无特殊说明，本文中的“拼写检查”均指中文拼写检查任务。中文拼写错误一般指由汉字语音 (phonologically) 或形态 (visually) 上相近，导致将一个字误写为另一个相近字 (别字) 的错误。需要注意的是，单纯由语义相近导致的别字错误，而并非音近或形近，不纳入该任务的修改范围。表 1 中列举了一些拼写检查错误示例。

表 1: 拼写检查错误类型示例

句子	修改	错误类型
我每天六天半起床。	天 → 点	音近
教帅是一个高尚的职业。	帅 → 师	形近
街上青一色地全是小汽车。	青 → 清	音近 + 形近
好好调查一下自己作文中的错别字。	调 → 检	不属于拼写检查错误

SIGHAN 2013-2015 (Wu et al., 2013; Yu et al., 2014b; Tseng et al., 2015) 发布了三批繁体中文的拼写检查开源数据集，极大程度上推动了中文拼写检查任务的发展。根据汉字的音近、形近关系，SIGHAN 2013 还提供了一份混淆集，对每个收录其中的汉字分别给出了易混淆字。这些语料的来源为中国台湾的留学生/中小学生作文，修正语法、语义错误后得到源端文本，再经由人工对错别字位置、纠正结果进行标注后获得目标端文本。然而，由于原始语料为繁体中文，一般经由 OpenCC[3] 等工具转换为简体中文后使用，这一过程会引入较多噪声。例如，繁体字与简体字存在多对一的情况，原始语料中的错误字，简化后可能不再是错误的 (復習 → 复习；複習 → 复习)；繁体字可能存在的混淆情况，简化后也可能不再互相混淆 (那理，那裡)。从语音的角度看，受限于语料产生地区的主要方言 (闽南语和粤语)，语音相近的错误类型同普通话语音不一致。此外，SIGHAN 数据集的标注质量仍有待提高，存在一些漏标、误标的情况，例如未标注许多“的、地、得”混用的情况。2018年，Wang 等人 (2018) 提供了一份伪数据，语料来源自人民日报和开源的中文演讲数据集。总体而言，中文拼写检查任务仍然缺乏质量较高的简体中文数据集。因此，本次评测构建并发布了 YACLC-CSC 数据集，供研究者用于模型验证和测试。

随着深度学习的发展，许多工作将神经网络模型应用于中文拼写检查任务中。Li 等人 (2018) 首先将神经机器翻译模型应用在该任务中。Wang 等人 (2019) 提出利用指针网络以减少搜索空间，并提高纠错准确率。Hong 等人 (2019) 利用 BERT (Devlin et al., 2019) 作为降噪自编码器 (Denoised Auto-Encoder)，并提出了 FASPell 模型。近年来，代表性的中文拼写检查方法主要有以下三种。

先检测后纠错 Zhang 等人 (2020) 提出了 Soft-Masked BERT，将错误探测和错误纠正分为两个子任务。

在图神经网络中融入混淆信息 Cheng 等人 (2020) 提出了 SpellGCN，利用图神经网络编码汉字之间的音近、形近关系，并利用注意力机制对不同类型的关系进行加权结合。Nguyen 等人 (2019) 提出利用 TreeLSTM (Tai et al., 2015) 对汉字进行层次化嵌入表示。

[3]OpenCC: https://github.com/BYVoid/OpenCC

多模态信息融合 Liu 等人 (2021a) 提出在与训练过程中融入汉字混淆信息，利用 GRU 分别对拼音、笔画进行编码，使用 BERT 对语义进行编码。Huang 等人 (2021) 提出使用门控机制加权融合汉字的字音、字形及语义。Xu 等人 (2021) 采用了相似的门控机制，但字音部分采用单向 GRU 编码拼音序列，字形部分则采用了 ResNet (He et al., 2016) 进行编码汉字的不同书体。

2.2 语法自动纠错

相较于拼写检查任务中对错误类型的严格限制，语法自动纠错任务要求自动检测并修改出全部的错误，包括标点、拼写、词汇、语序、语法、语义等方面，从而获得符合原意的正确句子。该任务既可面向母语者所写文本，也可面向第二语言学习者在说或写的过程中产出的文本，即学习者文本。

汉语学习者文本较难采集，也仍需人工精标注偏误信息，因此现有的带偏误标注信息的汉语学习者语料库十分匮乏，可应用于语法错误检测和纠正任务的训练和评测数据集尤为稀少。从2014年开始，面向教育应用的自然语言处理技术 (atural Language Processing Techniques for Educational Applications，NLPTEA) 开始组织汉语语法错误检测 (hinese Grammatical Error Diagnosis) 的评测比赛(Yu et al., 2014b)，语料采集自参加汉语托福考试 (Test of Chinese as a Foreign Language，TOCFL)(Chang, 2013)的学生所写的繁体作文。当时的任务要求更偏向于判断句子的对错，每个句子中或无错误或包含一个错误。2015年的CGED比赛开始增加了判断偏误位置的任务(Lee et al., 2015)，2016年开始加入汉语水平考试 (Hanyu Shuiping Kaoshi，HSK) 的简体作文语料(Lee et al., 2016)，2017年开始仅提供HSK语料(Rao et al., 2017)，每年发布更新数据集。尤为重要的是，2018年，CGED比赛开始加入了纠错任务，要求在错误检测的基础上修改错误，这一改变也延续到了2020年(Rao et al., 2018; Rao et al., 2020)。2018年，NLPCC 举办了首次公开的中文语法自动纠错评测比赛(Zhao et al., 2018)，评测任务要求直接对句子进行错误纠正。该次比赛所使用的训练语料来自语言学习和写作平台Lang-8，测试语料来自北京大学汉语学习者语料库。

现有的上述评测数据仍存在如下关键问题：第一，语料来源较为固定，多为课堂、作业、考试场景，无法评测开放场景下对学习者文本的错误检测和纠正效果；第二，现有的评测数据集基本都是采用最小改动的标注方式，因此欠缺流利度维度的偏误纠正结果，继而无法评测纠错模型在流利提升这一真实写作需求下的应用效果；第三，现有的评测数据集中大部分句子仅提供一种修改结果。这种单一的修改结果，极易出现语法自动纠错模型修改正确但与答案不匹配的现象，进而出现模型学习困难以及评测结果不够精准的问题。因此本次评测设计了多个中文语法纠错赛道，采用多个各有侧重的评测数据集，多方面评价现有纠错系统的性能。

深度学习方法兴起之后，中文语法错误检测往往被作为序列标注任务进行研究。2016年起，多个研究使用双向长短期记忆网络结合条件随机场 (BiLSTM+CRF) 的方法检测语法错误的位置，并通过添加如词向量、分词、词性标注信息和N元特征等特征增强建模，如(Zheng et al., 2016; Shiue et al., 2017; Fu et al., 2018b)。在2020 年NLPTEA 的CGED 评测比赛(Zhao et al., 2018)中，预训练语言模型BERT大展身手，Wang等人(2020b)在Transformer 语言模型的基础上融入残差网络，增强输出层中每个输入字的信息；Cao等人(2020)使用BERT模型结合门控机制，融合了语义特征、输入序列的位置特征和基于评分的特征；Luo等人(2020)使用基于BERT 模型和图卷积网络的方法在多任务学习框架下结合序列标注和端到端模型来提高原始序列标注任务的性能；陈柏霖等人(2022)使用ELECTRA预训练语言模型(Clark et al., 2020)对文本进行表征，接着采用卷积神经网络提取文本的局部位置和语义信息，并引入了残差和门控机制，在CGED2020的评测集(Zhao et al., 2018)中上达到了目前最好结果。

自2016年神经机器翻译方法崭露头角，语法纠错任务往往被视作文本生成任务，使用序列到序列 (Seq2Seq) 的生成模型，尤其是Transformer模型成为主流趋势。在NLPCC 2018的中文语法纠错评测比赛中，Fu等人(2018a)提出一种分阶段纠正方案，先利用语言模型移除表层错误，再利用Transformer模型移除深层的复杂语法错误，并进行模型融合和纠错结果重排序。Zhou等人(2018)采用多模型平行结构，使用基于规则、基于统计和神经网络三大类模型，采用高、低两种不同的组合策略得到最终纠错结果。Ren等人(2018)将词语切分成子词单元，并采用了基于CNN的序列生成模型。随后的大多中文语法纠错研究都是针对NLPCC 2018数据集开展的，如王辰成等人(2020)采用提出一种动态残差结构来增强Transformer架构挖掘文本语义信息的能力，Zhao和Wang(2020)在训练过程中采用动态的词频、同音等替换策略作用于错误句子，从而得到更多的错误－正确句对来提高模型的泛化能力。

2019年开始，英文语法纠错任务的研究者们尝试将文本生成任务转换为文本编辑任务，即序列到编辑 (Seq2Edit) 模型，有效地提升了预测速度，如LaserTagger(Malmi et al., 2019)结合BERT编码器与一个自回归的Transformer解码器来预测编辑。PIE模型(Awasthi et al., 2019)可以并行迭代地输入编辑而非文本，GECToR模型(Omelianchuk et al., 2020)结合BERT编码器

与非自回归的线性变化层去预测Token级别的编辑。2020年，Liang等人(2020)首次将英文中的Seq2Edit模型GECToR(Omelianchuk et al., 2020)引入到中文语法纠错中。Hinson等(2020)结合了三个模型循环纠正包含语法错误的句子，三个模型分别为：基于Transformer 的Seq2Seq模型，基于LaserTagger(Malmi et al., 2019)的Seq2Edit模型和拼写检查模型。

为解决中文语法纠错数据匮乏的问题，现有工作往往从以下方面进行研究：(1) 融合外部资源，如拼音、字形等信息作为额外特征集成到模型中，在处理拼写错误时使用较多，如Wang等(2019)使用一个带有指针网络的生成模型利用混淆集解决拼写错误，Cheng等人(2020)提出的SpellGCN模型利用图卷积神经网络融合字符的音近形近信息，李嘉诚等人(2022)在序列到编辑的纠错模型上利用指针网络融入汉字之间的音近和形近知识；(2) 使用预训练语言模型，如孙邱杰等人(2022)通过BART(Bidirectional and Auto-Regressive Transformers)噪声器对输入样本引入噪声，并使用基于BERT 的中文预训练语言模型对编码器参数进行初始化；(3) 使用随机遮蔽、替换或回译的数据增强方法，如王辰成等人(2020)提出了一种基于腐化语料的单语数据增强方法，按照随机添加、删除、替换的简单规则对语法正确，汤泽成等人(2021)首先对文本纠错中出现的错误进行了字和词粒度的分类，在此基础上提出了融合字词粒度噪声的数据增强方法；(4) 使用迁移学习方法，如张生盛等人(2021)提出个性化文本纠错，通过迁移学习方法将一般的文本纠错系统适应到汉语学习者不同的领域。

2.3 质量评估

质量评估 (Quality Estimation, QE) 是一种评价语法纠错模型修改结果的有效办法 (Chollampatt and Ng, 2018)。该任务要求预测特定语句的多种语法纠错结果的质量评估分数 (QE Score) ，来衡量这些结果的质量，对结果中的冗余修改、错误修改及欠修改情况进行评估 (Liu et al., 2021b)。

质量评估任务已被广泛应用于机器翻译领域，通过预测结果的 BLEU 值进一步提升机器翻译的效果。Kreutzer 等人 (2015) 提出质量评估模型 QUETCH ，不使用任何先验语言知识，仅通过从质量评估数据中获取的语言特征训练模型，已经起到了明显的提升效果。Kim 等人(2016; 2017)提出了一种端到端的神经模型，由词预测模型和质量评估模型堆叠而成，通过将预测器中有用的语言学信息迁移到评估器中，克服了质量评估任务中缺少训练数据的难题。Ive 等人 (2018) 提出了收割文档级别的神经质量评估框架 deepQuest ，但同时也能使用到词和句子级别的质量评估。同时，预训练语言模型，尤其是多语言预训练语言模型也已引入质量评估方法 (Wang et al., 2020a; Ranasinghe et al., 2020; Lee, 2020; Hu et al., 2020; Nakamachi et al., 2020) ，如 Kim 等人 (2019) 还提出使用平行语料来微调跨语言预训练模型的方法。

语法纠错质量评估 (Liu et al., 2021b) 也借鉴了机器翻译质量评估的思路，根据质量评估分数可以对语法纠错系统生成的多个纠错结果进行重新排序，以期望通过整合多个语法纠错模型结果来提升语法纠错效果。Liu等人 (Liu et al., 2021b) 提出的语法纠错质量评估模型通过编码器-解码器架构 (Encoder-Decoder) 进一步地建立输入文本和语法纠错结果之间的交互，并通过预测语法纠错官方评估分数 $F_{0.5}$ 以评估语法纠错质量。此外，语法纠错质量评估还可以用于标注者标注质量的衡量。评测组织方认为，语法纠错质量评估任务对文本纠错任务中的语料库构建及模型性能提升有着重要的作用，是当前文本纠错领域应着力的重要研究问题之一。

3 赛道设置

本次评测面向汉语学习者文本纠错的拼写检查、语法纠错和质量评估三个任务设置了五个赛道，针对每个赛道提供评测数据集，包括供参赛队伍进行模型调优的开发集，以及评测参赛队伍的模型性能的封闭测试数据集。同时设计具有可比性的评测指标，提供统一的入口规整评测流程，精细化、标准化汉语学习者文本纠错任务的基准评测框架。

3.1 赛道一：中文拼写检查

中文拼写检查任务要求对于给定的一段输入文本，参赛队伍需给出拼写错误的位置及对应的修改结果，其中拼写错误包含：音近、形近、形音兼近三种。赛道一提供YACLC-CSC数据集，这是首个简体中文的拼写检查数据集，数据来源为汉语学习者文本多维标注数据集YACLC(Wang et al., 2021)。在拼写错误标注方面，YACLC-CSC继承前人的研究，规定只标注和修正“音近”和“形近”有关的错误。判定为“音近”或“形近”或“形音兼近”的依据来自相关的汉语语音学、文字学理论及对外汉语教学理论。标注过程采用多人标注再由专家审核的方式以保证标注质量。数据示例如表1，“14”“15”为两个错误位置，“印”“象”为对应位置的修改结果。如该句没有错误，则输出“(id=xxx) 0”即可。评测提交所需的结果文件格式为每行是对应一个原句的校对结果，每行内容为原句的id，错误位置及纠正结果。

表 2: 中文拼写检查任务示例

原句	(id=012) 我觉得春天给人留下清爽的好影响。
拼写错误检测及纠正	(id=012) 14,印,15,象

YACLC-CSC共有637个存在拼写错误的句子，每句平均存在1.16个拼写错误点。我们将存在错误的句子定义为正样本，无错误的句子为负样本，并从YACLC-CSC数据集中抽取了550条作为评测任务的封闭测试集的正样本，剩余87条作为开发集的正样本。同时，按照1：1的比例加入没有拼写错误的负样本。即，最终数据集规模为：封闭测试集1,100条，开发集174条。赛道一允许使用任意开源数据用于训练，同时提供一份整合了现有真实开源数据集和伪数据的训练数据集共参赛队伍使用。其中真实开源数据集有SIGHAN 2013-2015的评测数据，伪数据集有Wang等人(2018)提供的数据集。

表 3: 赛道一数据集统计

	原句数	错误点数量	错句比例
YACLC-CSC-Dev	174	100	50%
YACLC-CSC-Test	1,100	637	50%

为评价模型检测和纠正错别字的能力，中文拼写检查任务一般采用准确率 (Acc，Accuracy)，精确率 (P，Precision Score)、召回率 (R，Recall Score) 和F1值进行评测。并且，一般有两种层级的评价：一种是句级 (Sentence Level)，即一句输入文本中所有错别字都检测或纠正正确，则算作正例；一种是字级 (Character Level)，即不考虑当前句的限制，最终的评价是基于整个测试集所有汉字的错误检测或纠正结果确定。对每个级别来说，又分为错误检测 (Error Detection)和错误纠正 (Error Correction) 两个维度。错误检测评估的是错误位置的侦测效果，错误纠正评估的是对应位置错误修正的效果。赛道一使用这两个级别的评测方式，综合展现各参赛系统的性能表现。

3.2 赛道二：中文语法错误检测

中文语法错误检测任务目的是检测出中文文本中每一处语法错误的位置、类型。语法错误的类型分为赘余(Redundant Words，R)、遗漏(Missing Words，M)、误用(Word Selection，S)、错序(Word Ordering Errors，W)四类。针对M和S类错误，给出纠正结果。评测任务要求参加评测的系统输入句子 (群)，其中包含有零个到多个错误。参赛系统应判断该输入是否包含错误，并识别错误类型，标记出其在句子中的位置和范围，对缺失和误用给出修正答案。赛道二提供CGED-8数据集，数据来源为HSK动态作文语料库(张宝林, 2009)和全球汉语中介语语料库(张宝林and 崔希亮, 2022)。同时提供前六届提供的训练集、测试集共六万余个错误点用于训练。CGED-8共包括约1400个段落单元、3,000个错误。每个单元包含1-5个句子，每个句子都被标注了语法错误的位置、类型和修改结果。数据示例如表4，评测提交所需的结果文件格式见表2的任务示例。每行是对应一个原句的一处检测结果，每行内容为原句的id、错误位置、错误类型及纠正结果。

表 4: 中文语法错误检测任务示例

原句	(sid=00038800481) 我根本不能了解这妇女辞职回家的现象。 在这个时代，为什么放弃自己的工作，就回家当家庭主妇?
语法错误检测	00038800481, 6, 7, S, 理解 00038800481, 8, 8, R

赛道二在从五个方面以精确率、召回率和F1值对参赛系统性能进行评价：

假阳性 (False Positive，FPR)：正确句子被判包含错误的比例。

检测层 (Detective-level，DET)：对句子是否包含错误做二分判断。

识别层 (Identification-level，IDE)：给出错误点的错误类型。

定位层 (Position-level，POS)：对错误点的位置和覆盖范围进行判断，以字符偏移量计。

修正层 (Correction-level，COR)：提交针对字符串误用(S)和缺失(M)两种错误类型的修正词语。修正词语可以是一个词，也可以是一个词组。

综合打分 (Comprehensive Score，COM)：2022年CGED-8引入2.1-2.5这五项指标F1值的加权评价分数。计算公式为：

$$\mathrm{COM} = 0.25 * \mathrm{DET} + 0.25 * \mathrm{IDE} + 0.25 * \mathrm{POS} + 0.25 * \mathrm{COR} - 0.25 * \mathrm{FPR}$$

赛道二规定在所有错误定界中，均不再考虑词的边界问题，错误均已字定界。这也符合第二语言学习者的学习实际，即缺乏词观念。如对于S型错误，即便只有一个语素错误 (通常是一个字)，也不再将整个词判为误用。

3.3 赛道三：多维度汉语学习者文本纠错

同一个语法错误从不同语法点的角度可被划定为不同的性质和类型(张宝林, 2013)，也会因语言使用的场景不同、具体需求不同，存在多种正确的修改方案。赛道三的数据来源为汉语学习者文本多维标注语料库YACLC(Wang et al., 2021)，数据中提供针对一个句子的多个参考答案，并且从最小改动 (Minimal Edit，M) 和流利提升 (Fluency Edit，F) 两个维度对模型结果进行评测。最小改动维度要求尽可能好地维持原句的结构，尽可能少地增删、替换句中的词语，使句子符合汉语语法规则；流利提升维度则进一步要求将句子修改得更为流利和地道，符合汉语母语者的表达习惯。因此赛道三提供最小改动和流利提升两个维度的多参考数据集YACLC-Minimal、YACLC-Fluency。其中YACLC-Minimal属于最小改动维度，YACLC-Fluency和MuCGEC属于流利提升维度。我们从公开发布的YACLC 1.0中随机抽取了9,135句及对应的71,969句标注结果。其中的1,839句作为开发集YACLC-Minimal-Dev和YACLC-Fluency-Dev，平均参考答案的数量分别为8.67和1.81句。剩余的7,296句作为封闭测试集YACLC-Minimal-Test和YACLC-Fluency-Test，平均参考答案的数量分别为5.82和1.86句。数据示例如表6，在两个维度上都有多个参考答案。评测提交要求参赛系统针对两个维度的测试集分别提交一个结果文件，格式为每行对应一个原句的纠正结果，且每个原句仅需提供一个结果。赛道三的数据集统计信息如表6所示。赛道三提供的训练数据来源于NLPCC2018-GEC(Zhao et al., 2018)的训练数据，并经过我们再次处理，且仅允许参赛队伍使用此数据用于训练。

赛道三采用的评测指标为基于字的编辑级别的F0.5指标。分别在最小改动和流利提升两个维度上计算F0.5值，最终排序以平均值为准。

表 5: 多参考中文语法纠错任务示例

原句		因为我的中文没有好，我还要努力学汉语。
最小改动	参考答案1	因为我的中文~~没有~~不好，我还要在努力学汉语。
	参考答案2	因为我的中文~~没有~~不好，所以我还要努力学汉语。
流利提升	参考答案1	~~因为~~我的中文没有那么好，因此我还要努力学汉语。
	参考答案2	因为我的中文还没有学好，所以我还要更加努力地~~学汉语~~中文。

注：其中，红字表示替换字符，蓝字表示插入字符，~~删除线~~表示删除字符。

表 6: 赛道三数据集统计

	原句数	参考句数	平均参考句数	有修改的参考句数（比例）	原句平均字符数	参考句平均字符数
YACLC-Minimal-Dev	1,839	15,938	8.67	15,935 (99.98%)	25.85	27.22
YACLC-Minimal-Test	7,296	42,462	5.82	40,334 (94.99%)	21.19	23.25
YACLC-Fluency-Dev	1,839	3,332	1.81	3,332 (100.00%)	25.85	27.14
YACLC-Fluency-Test	5,515	10,237	1.86	8,604 (84.05%)	20.81	21.40

3.4 赛道四：多参考多来源汉语学习者文本纠错

不同来源的文本，其蕴含的语法错误类型也可能含有一定的差异。赛道四的数据来源自MuCGEC(Zhang et al., 2022)，一个多参考答案、多领域的中文语法纠错数据集，采用了基于流利度的直接改写标注方式。因此赛道四提供来自于三个不同文本源的中文学习者语法纠错评测数据，对于每一个句子提供多个遵循流利提升的修改答案，希望能够准确而全面地评估各参赛队伍

的纠错系统性能。从MuCGEC中随机抽取了1,125句作为开发集MuCGEC-Dev，平均参考答案数量为2.19句。剩余5,938句作为封闭测试集MuCGEC-Test，平均参考答案数量为2.21句。评测提交要求参赛系统针对测试集提交一个结果文件，格式为每行对应一个原句的纠正结果，且每个原句仅需提供一个结果。赛道四的评测指标与赛道三相同，也是采用基于字的编辑级别的F0.5指标。赛道四的数据集统计信息如表7所示。

表 7: 赛道三数据集统计

	原句数	参考句数	平均参考句数	有修改的参考句数（比例）	原句平均字符数	参考句平均字符数
MuCGEC-Dev	1,125	2,467	2.19	2,409 (97.65%)	44.02	45.17
MuCGEC-Test	5,938	13,119	2.21	12,584 (95.92%)	37.42	38.24

3.5 赛道五：语法纠错质量评估

赛道五提供的评测数据集基于赛道三提供的YACLC-Minimal和YACLC-Fluency进行构建，数据划分与赛道三相同。我们基于BART-large训练了基于Seq2seq结构的语法纠错模型，并将改语法纠错模型在柱搜索解码过程中排名前五的生成结果作为待进行质量评估的语法纠错候选方案，以此构建语法纠错质量评估的训练集、验证集以及测试集。同时训练集和验证集中给出了每个语法纠错方案的真实F0.5分值。数据示例如表8所示。赛道五要求参赛队伍在语法纠错结果重排序过程中只能对所提供的语法纠错候选进行重排序，不得混合其他语法纠错模型所提供的语法纠错结果。并且最终提供一个语法纠错质量评估结果，该结果可以由多个语法纠错质量评估模型整合得到。文件每一行有且仅有一个改错结果或分数。赛道五的训练数据使用要求同赛道三，只能使用处理过的Lang8训练数据集。

质量评估的评价指标通常从两个方面对模型性能进行评测：一是评价质量评估模型所生成的质量评估分数，计算质量评估分数与真实F0.5分数之间的皮尔逊相关系数 (Pearson Correlation Coefficient，PCC)，用以衡量语法纠错质量评估分数与真实F0.5分数之间的相关性；二是评价模型选取的最高分纠错结果，对所给定的语法纠错结果进行重新排序，并选取分数最高的语法纠错结果作为最终的语法纠错结果，计算与参考答案之间的F0.5。

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

表 8: 语法纠错质量评估任务示例

原句	他今天去田里干活，我不只道他何时从田返回回来。	质量评估分数
修改结果1	他今天去田里干活，我不知道他何时从田里返回~~回~~来。	1.0
修改结果2	他今天去田里干活，我不知道他何时从田里~~返~~回回来。	1.0
修改结果3	他今天去田里干活，我不知道他何时从田返回回来。	0.3846

注：其中，红字表示替换字符，蓝字表示插入字符，~~删除线~~表示删除字符。

4 参赛及获奖情况

本次比赛吸引了142支队伍报名参赛，包括清华、北大、中科院、北邮、苏大等高校和科研院所，达观、蜜度、好未来、视源等企业。综合各赛道参赛队伍的榜单成绩、代码完善和复现情况以及所提交的评测报告，本次比赛评选出五个赛道一至三等奖的参赛队伍，获奖队伍及单位如表9所示：

5 方法及分析

采用序列到序列或序列到编辑的神经网络模型，结合预训练语言模型是当前文本纠错技术的主流策略。各赛道获奖队伍的参赛系统也多是如此，同时加以多模型集成方式提高模型性能，采用基于规则或者混淆集的数据增强方法扩充训练数据规模。各赛道均提供基于预训练语言模型的基线模型，即模型架构仅使用预训练语言模型对整个输入进行编码，附加了一层线性层做为分类器，得到每个位置的字符的纠错结果。本节内容对各赛道的获奖队伍的方法进行分析。

表 9: 各赛道获奖情况

	赛道一	赛道二	赛道三	赛道四	赛道五
一等奖	哒哒 （达观数据）	NLP的未来 （好未来）	kk （北京大学）	啊对对对 （清华大学） 鱼饼啾啾 （北京大学）	CPIC (中国太平洋保险)
二等奖	iFunCun （方寸无忧）	一一 （达观数据）	改正带小助手 （苏州大学）	棒棒冰 (视源电子科技)	
三等奖	csc_runner (视源电子科技)	中国足球队 （蜜度）	BUPTCL (北京邮电大学)	后厂村9号 （海泰方圆）	

5.1 赛道一：中文拼写检查

赛道一共59支队伍在排行榜上做了有效提交，其中哒哒、iFunCun、csc_runner三支队伍获得了一至三等奖。三支队伍的参赛模型及基线模型BERT-base在句级和字级的错误纠正和检测维度上的F1值分数如下表。

表 10: 赛道一获奖队伍成绩

	句级		字级	
	C-F	D-F	C-F	D-F
哒哒	84.33	85.49	98.32	88.34
iFunCun	83.19	84.53	97.64	88.71
csc_runner	81.08	83.08	96.63	85.98
BERT-base	**50.05**	**56.52**	**84.67**	**61.94**

在模型架构上，三支队伍使用的都是序列到序列的纠错模型：一等奖哒哒队伍在Transformer输入端融合了拼音编码，使用多轮纠错的推理方式；二等奖iFunCun队伍集成多个基于BERT的纠错模型Macbert4csc(Xu, 2019)、ReaLiSe(Xu et al., 2021)、CRASpell(Liu et al., 2022)，融合了拼音、字形特征，基于困惑度对多模型纠错结果重排序；三等奖csc_runner队伍集成多个基于预训练语言模型 BERT (Devlin et al., 2019)、macbert(Cui et al., 2020)、RoBERTa(Liu et al., 2019)的纠错模型，通过平均模型输出概率获取纠错结果。

为解决中文拼写检查数据匮乏的问题，三支队伍都使用了数据增强策略以及多阶段的模型训练方式：一等奖哒哒队伍基于混淆集使用中文维基百科、微信和新闻语料获得超过200万的伪数据用于预训练，再使用SIGHAN和YACLC-CSC的真实训练数据以及Wang(2018)的伪数据进行微调；二等奖iFunCun队伍基于混淆集在SIGHAN和Wang(2018)切分出的训练集上进行字和词级别的替换，获取到14万伪平行句对用于基线模型的微调；三等奖csc_runner队伍基于混淆集和拼写错误实例在多来源的中文语料上进行字和词级别的替换，获取到1300万句对的伪数据用于预训练，再使用所有拼写纠错数据微调以及SIGHAN数据精调。

哒哒和csc_runner队伍还使用了基于语言模型或规则的方法对模型的过纠正问题进行后处理，包括对非音近形近的替换，以及成语和实体上的错纠漏纠。

综合来看，哒哒和iFunCun队伍的模型架构相较于csc_runner队伍融合了更多的字音字形信息，因此即使伪数据量较少时单模型也能达到比csc_runner队伍的多模型集成还要好的性能。同时，处理过纠正问题的模型后处理方案分别为哒哒和csc_runner队伍在句级别错误纠正维度上带来4.06和2.34的F1值提升。因此，单就拼写检查模型而言，性能仍有上升的空间。并且，由于汉字的特殊性，结合文字学知识以构建更合理的模型架构仍是最行之有效的研究方向。

5.2 赛道二：中文语法错误检测

赛道二共 26 支队伍在排行榜上做了有效提交，其中NLP的未来、一一、中国足球队三支队伍获得了一至三等奖。三支队伍的参赛模型及基线模型ELECTRA(Clark et al., 2020)、BERT(Devlin et al., 2019)和RoBERTa(Liu et al., 2019)在句级和字级的错误纠正和检测维度上的F1值分数如下表。

在模型架构上，三支队伍使用的都是多种序列到序列和序列到编辑的纠错模型集成，同时都针对拼写错误训练模型：一等奖NLP的未来队伍训练了基于BERT+CRF和BERT+BiLSTM+CRF的错误

表 11: 赛道二获奖队伍成绩

队伍名	COM	FPR	POS	COR
NLP的未来	48.50	14.9	39.85	28.31
一一	46.59	19.76	38.98	29.19
中国足球队	46.27	22.42	37.66	27.6
ELECTRA	37.34	29.2	30.97	19.69
BERT	37.1	30.24	30.13	19.64
RoBERTa	36.36	30.09	29.86	18.49

检测模型、基于GECToR的序列到编辑的纠错模型、基于Transformer融合指针网络的序列到序列的纠错模型、序列到动作的纠错模型以及两个基于BERT的拼写纠错模型，使用基于编辑级别投票的多模型集成方式；二等奖一一队伍训练了基于BERT的多任务拼写纠错模型和基于GECToR的序列到编辑的纠错模型，基于模型概率加权平均和困惑度集成多模型纠错结果；三等奖中国足球队队伍训练了针对不同类型错误的多个拼写纠错模型和基于GECToR的序列到编辑的纠错模型，从标签和编辑两个层面进行投票集成。三支队伍使用了混淆集基于规则进行字和词级别的数据增强方法。

综合来看，不同架构、不同针对性的模型各有所长，多种模型的集成是在中文语法错误检测任务上获得较高性能的关键，但仍有许多错误未被检测和纠正，模型性能仍有很大的提升空间。因此在保证模型效果的同时，降低模型复杂度、提升模型效率是中文语法错误检测任务未来主要的研究和应用方向。

5.3 赛道三：多维度汉语学习者文本纠错

赛道三共9支队伍在排行榜上做了有效提交，其中kk、改正带小助手、BUPTCL三支队伍获得了一至三等奖。三支队伍的参赛模型及基线模型BART在最小改动和流利提升两个维度以及平均的F0.5分数如下表。

表 12: 赛道三获奖队伍成绩

	平均	最小改动			流利提升		
	F0.5	F0.5	Prec	Rec	F0.5	Prec	Rec
kk	55.74	71.51	79.95	50.27	39.97	50.69	21.66
改正带小助手	50.41	64.42	69.99	48.86	36.4	43.01	22.54
BUPTCL	47.39	59.89	68.88	39.35	34.89	45.96	17.77
BART	41.98	54.43	60.1	39.52	29.52	37.62	15.86

在模型架构上，一等奖kk队伍使用基于BART的序列到序列的纠错模型，基于编辑级别投票对多个纠错结果进行集成；二等奖改正带小助手队伍使用基于BART的序列到序列的纠错模型和基于GECToR的序列到编辑的纠错模型,基于模型输出概率选择纠错结果；三等奖BUPTCL队伍使用基于GECToR的序列到编辑的纠错模型，基于模型输出概率平均和编辑级别投票进行模型集成。在数据增强方案上，kk队伍使用了基于规则的语料腐化方法生成伪数据，探索了动态和静态两种不同的加噪方式对模型性能的影响。同时，kk队伍还对纠错结果中的UNK问题进行了后处理。

综合赛道一、二三的参赛模型以及其他现有研究，中文文本的序列到序列的纠错模型范式逐渐从BERT转为BART模型，盖因BART模型使用了完整的Transformer架构，噪声自编码的预训练任务也与基于规则的数据增强技术类似。而中文文本的序列到编辑的纠错模型往往基于GECToR构建。同时，与赛道二类似，多模型集成是提升纠错性能的关键，模型后处理方案是针对特定类型错误的有效策略。但三支队伍都未针对最小改动和流利提升两个维度采用不同的模型或策略，对多参考答案的数据形式也并未采用特殊的表征方式。因此深度融合多维度、多参考的文本特征，以及适用于判错、改错、润色等多种应用需求的高性能、高效率的纠错模型设计是中文文本纠错任务的下一步研究方向。

5.4 赛道四：多参考多来源汉语学习者文本纠错

赛道四共17支队伍在排行榜上做了有效提交，其中啊对对对、鱼饼啾啾两支队伍并列获得了一等奖，棒棒冰队和后厂村9号队两支队伍分别获得了二和三等奖。四支队伍的参赛模型及基线模型在最小改动和流利提升两个维度以及平均的F0.5分数如下表。

表 13: 赛道四获奖队伍成绩

队伍名	F0.5
啊对对对	51.15
鱼饼啾啾	51.02
棒棒冰队	50.17
后厂村9号队	43.09
Base	39.62

四支队伍都训练了基于BART的序列到序列模型和基于GECToR的序列到编辑的模型，进行多模型集成。一等奖啊对对对与二等奖棒棒冰队伍使用了基于编辑级别的投票机制选择大多数模型提供的编辑作为最终保留的纠错操作；鱼饼啾啾队伍基于编辑级别进行加权和分层集成；三等奖后厂村9号队伍探索了串行和并行两种集成方式，实验发现串行集成可能会导致模型效果下降，而并行集成需要考虑融合基于编辑和整句困惑度。

针对多参考答案问题，啊对对对队伍提出了一个简单有效的数据清洗策略，基于编辑距离选择与源句最相似的参考答案进行训练，以消除多目标训练造成的概率质量稀释问题。评测数据的不同来源造成了错误分布的差异，赛道四的参赛队伍虽未针对多来源问题进行探究，但纠错模型的领域适用性仍是中文文本纠错任务的重要研究方向。

5.5 赛道五：语法纠错质量评估

赛道五共7支队伍在排行榜上做了有效提交，CPIC队伍获得了一等奖。

表 14: 赛道五获奖队伍成绩

	平均	最小改动		流利提升	
	F0.5	F0.5	PCC	F0.5	Prec
CPIC	47.91	61.17	0.2405	34.64	0.2437
BERT-QE	37.76	47.94	0.1018	27.58	0.0376
BERT-GQE	30.86	40.92	0.0538	20.79	0.0733

CPIC队伍在基线模型的基础上，替换不同的预训练模型，实验发现使用ELECTRA模型(Clark et al., 2020)能获取较好的结果。进而使用对抗训练的EMA策略(Miyato et al., 2016)利用滑动平均的参数来提高模型在测试数据上的健壮性，同时组合FGM策略(Miyato et al., 2016)提高模型应对恶意对抗样本时的鲁棒性，减少过拟合，提高模型泛化能力。同时，CPIC队伍还训练了判断是否包含错误的二分类模型和预测F0.5分数的回归模型，在多模型集成上尝试了基于模型输出概率平均和投票原则两种方式。与基线模型相比，CPIC队伍的平均F0.5得分提升了约117个百分点，在PCC得分上也得到了明显的提升，性能提升十分明显。

赛道五的参赛队伍更为关注提升最优纠错结果的评估效果，对识别不同纠错结果的评估质量有所忽视，因此设计同时提升评估效果和质量的模型架构，用以更全面、更可靠地评估现有文本纠错系统的性能，仍旧是中文文本纠错质量评估任务的重要研究方向。

6 总结

本次汉语学习者文本纠错评测比赛 (CLTC 2022)，由北京语言大学联合清华大学、苏州大学、东北大学、阿里巴巴达摩院共同组织，并附属于第21届中国计算语言学大会 (CCL 2022) 举办了评测研讨会。本次评测聚焦该研究领域中的前沿问题，整合了已有的文本纠错的相关评测数据和任务，发布了新的数据集，构建了汉语学习者文本纠错任务的基准评测框架，以设置多赛道、统一入口的方式开展比赛任务，开发了支持随时、长期进行评测的公共平台，旨在不断改进文本纠错数据及任务，充分发挥评测引领技术发展、推进研究进步的作用。在中文拼写检查、中文语法错误检测、多维度汉语学习者文本纠错、多参考多来源汉语学习者文本纠错、语法纠错质量评估这五个赛道上，共有142支来自各大高校、科研院所以及企业的队伍报名提交参赛系统，最终15支队伍获得奖项。相较于基线模型，参赛系统的性能有大幅提升，展现出了汉语学习者文本纠错任务上的现有水平。

参考文献

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. *CoRR*, abs/1910.02893.

Yongchang Cao, Liang He, Robert Ridley, and Xinyu Dai. 2020. Integrating BERT and score-based feature gates for Chinese grammatical error diagnosis. In *Proceedings of NLPTEA*.

Liping Chang. 2013. Tocfl作文语料库的建置与应用. 第二汉语中介语语料库建设与应用国际学术研讨会论文选集.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Shamil Chollampatt and Hwee Tou Ng. 2018. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of EMNLP 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Kai Fu, Jin Huang, and Yitao Duan. 2018a. Youdao's winning solution to the NLPCC-2018 task 2 challenge: A neural machine translation approach to chinese grammatical error correction. In *Natural Language Processing and Chinese Computing*.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2018b. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. In *Proceedings of NLPTEA*.

Gabriel Fung, Maxime Debosschere, Dingmin Wang, Bo Li, Jia Zhu, and Kam-Fai Wong. 2017. NLPTEA 2017 shared task – Chinese spelling check. In *Proceedings of NLPTEA*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Charles Hinson, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Heterogeneous recycle generation for Chinese grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. In *Proceedings of W-NUT*.

Chi Hu, Hui Liu, Kai Feng, Chen Xu, Nuo Xu, Zefan Zhou, Shiqin Yan, Yingfeng Luo, Chenglong Wang, Xia Meng, Tong Xiao, and Jingbo Zhu. 2020. The NiuTrans system for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*.

Li Huang, Junjie Li, Weiwei Jiang, Zhiyu Zhang, Minchuan Chen, Shaojun Wang, and Jing Xiao. 2021. PHMOSpell: Phonological and morphological knowledge guided Chinese spelling check. In *Proceedings of ACL-IJCNLP*.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. deepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of NAACL-HLT*.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*.

Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and Seung-Hoon Na. 2019. QE BERT: Bilingual BERT using multi-task learning for neural quality estimation. In *Proceedings of the Fourth Conference on Machine Translation*.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. QUality estimation from ScraTCH (QUETCH): Deep learning for word-level translation quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang. 2015. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. In *Proceedings of NLPTEA*.

Lung-Hao Lee, Gaoqi Rao, Liang-Chih Yu, Endong Xun, Baolin Zhang, and Li-Ping Chang. 2016. Overview of NLP-TEA 2016 shared task for Chinese grammatical error diagnosis. In *Proceedings of NLPTEA*.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*.

Chiao-Wen Li, Jhih-Jie Chen, and Jason Chang. 2018. Chinese spelling check based on neural machine translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Deng Liang, Chen Zheng, Lei Guo, Xin Cui, Xiuzhang Xiong, Hengqiao Rong, and Jinpeng Dong. 2020. BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis. In *Proceedings of NLPTEA*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arxiv:1907.11692.

Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. 2021a. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. In *Proceedings of ACL-IJCNLP*.

Zhenghao Liu, Xiaoyuan Yi, Maosong Sun, Liner Yang, and Tat-Seng Chua. 2021b. Neural quality estimation with multiple hypotheses for grammatical error correction. In *Proceedings of NAACL-HLT*.

Shulin Liu, Shengkang Song, Tianchi Yue, Tao Yang, Huihui Cai, TingHao Yu, and Shengli Sun. 2022. CRASpell: A contextual typo robust approach to improve Chinese spelling correction. In *Findings of ACL 2022*.

Yikang Luo, Zuyi Bao, Chen Li, and Rui Wang. 2020. Chinese grammatical error diagnosis with graph convolution network and multi-task learning. In *Proceedings of NLPTEA*.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of EMNLP-IJCNLP*.

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

Akifumi Nakamachi, Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. TMUOU submission for WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*.

Minh Nguyen, Gia H Ngo, and Nancy F Chen. 2019. Hierarchical character embeddings: Learning phonological and semantic representations in languages of logographic origin using recursive neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:461–473.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Gaoqi Rao, Baolin Zhang, Endong Xun, and Lung-Hao Lee. 2017. IJCNLP-2017 task 1: Chinese grammatical error diagnosis. In *Proceedings of the IJCNLP 2017, Shared Tasks*.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings of NLPTEA*.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of nlptea-2020 shared task for chinese grammatical error diagnosis. In *Proceedings of NLPTEA*.

Hongkai Ren, Liner Yang, and Endong Xun. 2018. A sequence to sequence learning for chinese grammatical error correction. In *NLPCC*.

Yow-Ting Shiue, Hen-Hsen Huang, and Hsin-Hsi Chen. 2017. Detection of Chinese word usage errors for non-native Chinese learners with bidirectional LSTM. In *Proceedings of ACL*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Zecheng Tang, Yixin Ji, Yibo Zhao, and Junhui Li. 2021. 基于字词粒度噪声数据增强的中文语法纠错. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to sighan 2015 bake-off for chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Dingmin Wang, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for Chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, Yimeng Chen, and Liangyou Li. 2020a. HW-TSC's participation at WMT 2020 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*.

Shaolei Wang, Baoxin Wang, Jiefu Gong, Zhongyuan Wang, Xiao Hu, Xingyi Duan, Zizhuo Shen, Gang Yue, Ruiji Fu, Dayong Wu, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2020b. Combining ResNet and transformer for Chinese grammatical error diagnosis. In *Proceedings of NLPTEA*.

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

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at sighan bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In *Findings of ACL-IJCNLP*.

Ming Xu. 2019. Pycorrector: Text error correction tool. https://github.com/shibing624/pycorrector.

Liner Yang, Chengcheng Wang, Yun Chen, Yongping Du, and Erhong Yang. 2022. Controllable data synthesis method for grammatical error correction. *Frontiers of Computer Science*, 16(4).

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014a. Overview of grammatical error diagnosis for learning chinese as a foreign language. In *Proceedings of NLPTEA*.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014b. Overview of sighan 2014 bake-off for chinese spelling check. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked bert. *arXiv preprint arXiv:2005.07421*.

Yueli Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. In *Proceedings of NAACL*.

Zewei Zhao and Houfeng Wang. 2020. Maskgec: Improving neural grammatical error correction via dynamic masking. In *Proceedings of AAAI*.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the nlpcc 2018 shared task: Grammatical error correction. In *CCF International Conference on Natural Language Processing and Chinese Computing*.

Bo Zheng, Wanxiang Che, Jiang Guo, and Ting Liu. 2016. Chinese grammatical error diagnosis with long short-term memory networks. In *Proceedings of NLPTEA*.

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. 2018. Chinese grammatical error correction using statistical and neural models. In *Natural Language Processing and Chinese Computing*.

孙邱杰, 梁景贵, and 李思. 2022. 基于bart噪声器的中文语法纠错模型. 计算机应用, 42(3):860–866.

张宝林and 崔希亮. 2022. “全球汉语中介语语料库” 的特点与功能. 世界汉语教学.

张宝林. 2009. “hsk 动态作文语料库” 的特色与功能. 汉语国际教育.

张宝林. 2013. 关于通用型汉语中介语语料库标注模式的再认识. 世界汉语教学, 27(1):128–140.

张生盛, 庞桂娜, 杨麟儿, 王辰成, 杜永萍, 杨尔弘, and 黄雅平. 2021. 面向汉语作为第二语言学习的个性化语法纠错. 中文信息学报, 35(12):28–35.

李嘉诚, 沈嘉钰, 龚晨, 李正华, and 张民. 2022. 基于指针网络融入混淆集知识的中文语法纠错. 中文信息学报, 36(4):29.

王辰成, 杨麟儿, 王莹莹, 杜永萍, and 杨尔弘. 2020. 基于transformer增强架构的中文语法纠错方法. 中文信息学报, 34(6):106.

陈柏霖, 王天极, 任丽娜, and 黄瑞章. 2022. 融合electra和文本局部信息的中文语法错误检测方法. 计算机工程.