

CCL2022-CLTC 赛道3: kk队评测报告

黄蓉

北京大学 软件与微电子学院, 北京

aroon@stu.pku.edu.cn

摘要

文章描述了本队伍在CCL 2022 汉语学习者文本纠错测评大赛的赛道三中提交的参赛系统。近年来, 汉语学习者文本纠错任务越来越受到关注, 但仍然存在语法错误复杂、纠错语料稀缺等特点。本队伍使用序列到序列的纠错模型, 采用两阶段训练策略, 第一阶段在训练集上混合动态加噪和静态加噪两种方式, 第二阶段使用验证集微调, 最终结合模型输出结果的特征完成后处理。系统测评中, 本队伍提交的结果在最小改动F0.5、流利提升F0.5、总得分三个指标上均位列第一。参赛模型可参考: <https://github.com/Sophiarong/CCL2022-MCLTC-kk.git>

关键词: 多维度文本纠错

CCL2022-CLTC Track3: Technical Reports of kk Team

Huang Rong

Peking University School of Software & Microelectronics, Beijing

aroon@stu.pku.edu.cn

Abstract

The report presents the model submitted by our team in Track 3 of the CCL 2022 Chinese Learners Text Correction Competition. In recent years, the task of text correction for Chinese learners has received increasing attention, but it has the problems of complex grammatical errors and scarce error correction corpus. Our team uses a sequence-to-sequence error correction model with a two-stage training strategy, mixing dynamic and static noise on the training set in the first stage, fine-tuning it using the validation set in the second stage. Finally, post-processing is completed based on the characteristics of the model output results. In the system evaluation, the results submitted by our team ranked first in the three indexes of minimal F0.5, fluency F0.5, and total score. The model can be found at: <https://github.com/Sophiarong/CCL2022-MCLTC-kk.git>

Keywords: Multidimensional Text Correction

1 引言

随着移动互联网的兴起，人人都能成为信息的生产者，由于信息生产的随意性和语音识别等环节的易错性，网络中含有语法错误的内容爆炸式增长。另外，国际中文教育需求旺盛，外国汉语学习者学习仍然面临在中文拼写以及语法方面易犯错误的问题。因此，中文文本纠错（Chinese Text Correction）的重要性逐渐凸显，它指的是自动检测句子中包含的语法错误，并纠正所检测出的错误。

目前，中文文本纠错数据集数量少、规模小，主要来自于Lang8平台的中介语数据和HSK动态合成语料库，这两个被广泛使用的语料库仅标注了汉字、词语和短语中的错误。而深度学习模型，尤其是序列到序列的模型的训练通常需要大规模的训练数据，受到数据集规模和改错方式的限制，现有模型往往倾向于修改小范围的助词错误和拼写错误等表层错误，未能够在保留原有语义的同时充分修正语法错误。

本次比赛的赛道三赛题为多维度汉语学习者文本纠错(Wang et al., 2022)，从最小改动和流利度提升两个维度测评模型。考虑到赛题设置和现有解决方案的缺陷，本队伍使用两阶段训练策略，第一阶段使用加噪后的Lang8语法纠错数据训练模型，第二阶段使用与测试集分布更相似的验证集继续训练，最终结合模型输出结果的特征完成后处理。相比于官方释出的基准模型，本队伍分别在最小改动F0.5和流利度提升F0.5两个指标上提升了17.08%和10.54%。

2 相关工作

语法错误纠正是自然语言处理的重要下层任务，英语语法错误纠正已经研究了很多年，并且取得了丰硕的成果，目前的研究方向主要分为三个部分：（1）基于序列标注的语法错误纠正方法、（2）基于文本生成的语法错误纠正方法、（3）含有语法错误的训练句对构造方法。

基于序列标注的方法(Awasthi et al., 2019; Malmi et al., 2019; Omelianchuk et al., 2020)具有可解释性强的特点，但对于语法操作标签的定义复杂严格。GECToR(Omelianchuk et al., 2020)详细定义了多种语法操作标签，标签空间达到了5000个，需要在预处理阶段根据错误-正确句对的编辑距离为错误句子的每个字进行标注以得到训练数据。

基于文本生成的方法与机器翻译的任务相似，都是基于原句得到目标句。考虑到语法错误纠正任务中的原句和目标句均属于同一种语言，Copy-augmented Model(Zhao et al., 2019)将copy机制引入语法错误纠正模型。BERT-fuse GED(Kaneko et al., 2020)在含有语法错误语料上微调序列标注的BERT模型，然后将微调阶段的输出作为额外特征，与原句共同输入由BERT初始化的语法错误纠正模型。

语法错误纠正的数据集主要由句对组成，是错误句子到正确句子的映射，对标注者的语言知识要求较高，因此数据集数量和规模有限，较难满足深度神经网络的训练需求。Kiyono et al. (2019)的研究分析了构造数据的种子语料、构造数据的方法以及训练策略对于语法错误纠正模型的影响。PoDA(Wang et al., 2019)对文本随机添加噪声，然后使用语法错误纠正模型复原加噪语料，最后将模型迁移到领域数据集上进行微调。Lichtarge et al. (2019)提出利用机器翻译系统将语料翻译成中间语言，再由中间语言译回源语言，原始语料和经过两轮翻译的语料可以组合成语法错误纠正的伪数据。

相对于英文语法错误纠正，中文文本纠错的相关资源和数据更为稀少，近年来举办的比赛包括NLPCC2018语法纠错评测(Zhao et al., 2018)、CGED2020评测(Rao et al., 2020)、CTC2021评测(Zhao et al., 2022)。Youdao(Fu et al., 2018)将中文语法错误纠正任务分成3个阶段，先移除句子中的表层错误，再将任务转化为机器翻译任务，最后集成第二阶段的若干模型。AliGM(Zhou et al., 2018)使用基于规则、基于统计和神经网络三大类模型，根据不同优先级组合模型。MaskGEC(Zhao and Wang, 2020)在每轮迭代的时候动态随机增加噪声，提升数据的泛化性。

3 任务介绍

多维度汉语学习者文本纠错（Multidimensional Chinese Learner Text Correction）任务旨在自动检测汉语学习者文本中的标点、拼写、语法、语义等错误，并且从最小改动（Minimal

Edit, M) 和流利提升 (Fluency Edit, F) 两个维度产生修正方案, 从而获得符合原意的正确句子。最小改动维度要求尽可能完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。本次比赛的任务示例如表1所示, 输入为可能含有错误的句子 $E = [e_1, e_2, \dots, e_m]$, 输出为符合原意的正确句子 $P = [p_1, p_2, \dots, p_n]$, 并期望输出结果尽可能分别贴近两个维度的参考答案。需要注意的是, 每个维度可能存在多个参考答案, $C_{minimal} = (C_{m.1}, C_{m.2}, \dots, C_{m.k})$ 表示最小改动维度存在 k 个参考答案, $C_{fluency} = (C_{f.1}, C_{f.2}, \dots, C_{f.g})$ 表示流利提升维度存在 g 个参考答案。

原句	他做梦的时候他是一个太空人了。
最小改动参考答案	他做梦梦到他是一个太空人。
	他做梦梦到他是一个太空人了。
	他梦见他是一个太空人了。
	他做梦的时候, 他梦到自己是一个太空人了。
	他做梦的时候他成为了一个太空人。
	他做梦的时候梦见他成为一个太空人了。
	他在梦里成为了一个太空人。
	梦中他成了一个太空人。
	他做梦的时候他变成一个太空人了。
	他做梦的时候梦到他变成一个太空人了。
流利提升参考答案	他梦到他是一个太空人。

表 1: 任务示例

数据集方面, 官方对NLPC2018-GEC发布的采集自Lang8 平台的中介语数据进行了处理, 提供约121万个错误-正确句对作为训练集。如表2所示, 一个包含语法错误的单句可能对应多个参考答案, 训练集中实际包含约71万句不重复的错误句, 平均每句话存在1.7个参考答案。

原句	参考答案
长成大人, 我盒饭做的很开心。	长大成人后, 我做盒饭做得很开心。
城市里的人能度过多方面的生活。	城市里的人能过丰富多彩的生活。
	城市里的人能过多种多样的生活。
	城市里的人能过多方面的生活。

表 2: 训练集示例

官方基于YACL数据集提供最小改动和流利提升两个维度上的验证集。最小改动维度的验证集包含1.5万个错误-正确句对, 其中共有1839句不重复的错误句, 平均每句话对应8.6个符合最小改动规则的参考答案。流利提升维度的验证集包含3322个错误-正确句对, 其中包含与最小改动维度完全相同的1839条不重复错误句, 平均每句话对应1.8个符合流利提升规则的参考答案。第一阶段测评包括两个维度的测试数据各900条, 第二阶段包括最小改动维度的测试数据6396条和流利提升维度的测试数据4615条, 每个阶段中两维度的测试数据内容不同。数据集整体构成如表3所示。

	训练集	验证集		第一阶段测试集		第二阶段测试集	
		minimal	fluency	minimal	fluency	minimal	fluency
错误的原句数量	712756	1839		900	900	6396	4615
错误-正确句对数量	1213457	15873	3322	\	\	\	\

表 3: 数据集构成

评价指标使用中文语法错误纠正评估工具ChERRANT(Zhang et al., 2022), 通过对比预测结果编辑和参考答案标准编辑, 计算预测结果的精确度、召回度、F0.5值。

4 模型

比赛中，本队伍采用两阶段训练策略，第一阶段使用加噪后的训练集训练模型，第二阶段使用与测试集分布更相似的验证集继续训练，最终结合模型输出结果的特征完成后处理。

4.1 数据预处理

对于神经网络模型，训练集的大小通常是影响模型性能的关键因素之一。汉语学习过程中，即使想表达相同的内容，不同的学习者所犯的语法错误也是多样化的。当两名学习者都想表达“我的成绩还不错”这个句意时，学习者一可能误表述出“我的成绩还不好”，学习者二可能误表述出“我的成绩错得少”。两个错误句子和正确句子间的差异不大，在模型训练过程中增加类似扰动可以提高模型的性能。受MaskGEC(Zhao and Wang, 2020)启发，本队伍采用了静态加噪和动态加噪两种方式。

基于BART-large自带的词典文件，本队伍筛选出在71万条原句中未出现的单字，构成词汇表 V 。对于给定的待纠错原句 E ，使用 v_i 替换 e_j ，其中 v_i 和 e_j 均由随机抽取得出。动态加噪过程如表4所示，在模型的第 t 次迭代中，从训练数据集 W 中随机选取数据量为 $\beta|W|$ 的子集合 $W^{(t)}$ ，其中 β 表示从训练集中被选取句子的比例， $|W|$ 表示训练集的大小。随后对子集合 $W^{(t)}$ 的每个数据对 (E, C) 的待纠错句子 E 添加噪音，获得含有更多语法错误的句子 \tilde{E} ，形成新的数据对 (\tilde{E}, C) 。对于每个被选中的句子，加噪次数重复 N 次，即每个被选中的句子在原来的基础上被替换了 N 个单字。上述过程结束后即可得到用于本轮训练模型的训练集 W^* 。相比动态加噪过程，静态加噪过程没有在每轮训练前重新加噪，而是在数据加载前完成该过程。两种加噪方式使用相同的加噪比例 β 和加噪次数 N 。

```

1 : 输入：数据集 $W$ ，加噪比例 $\beta$ ，加噪次数 $N$ 
2 : 输出：模型 $\Theta$ 
3 : 初始化模型
4 : for each training epoch  $t$  do
5 :      $W^* = W$ 
6 :     for  $i \leftarrow 0$  to  $N$  do
7 :          $\tilde{W}^{(t)} \leftarrow \emptyset$ ;
8 :          $W^{(t)} = \text{randomSelect}(W, k = \beta|W|)$ ; //从 $W^*$ 中随机选取 $k$ 个实例
9 :         for each  $(E, C) \in W^{(t)}$  do
10:             $(\tilde{E}, C) \leftarrow f((E, C))$ ; //对 $E$ 添加噪声，生成新数据对
11:             $\tilde{W}^{(t)} \leftarrow \tilde{W}^{(t)} \cup (\tilde{E}, C)$ ;
12:        end for
13:         $W^* = (W^* - W^{(t)}) \cup \tilde{W}^{(t)}$ ;
14:    Training model and update  $\Theta$  using  $W^*$ ;
15: end for

```

表 4: 动态加噪过程算法

4.2 模型设计

考虑到中文文本纠错任务与机器翻译任务的相似性，本队伍采用序列到序列模型结构。Katsumata and Komachi (2020)的研究表明，对于以Transformer作为基础架构的语法错误纠正模型，通过使用预训练语言模型BART(Lewis et al., 2019)能够获得较好的性能表现。不同于BERT(Devlin et al., 2018)，BART是更符合生成任务的预训练方法，兼具上下文语境信息和自回归特性。本队伍的模型搭建基于fairseq(Ott et al., 2019)工具包，模型详细信息如表5所示。另外，受预训练模型的词表限制，一些常见的中文标点或字符如上引号等在分词阶段会被[UNK]替代，导致模型输出也带有[UNK]标识。该标识在评测阶段会被当做错误的编辑操作。针对以上缺陷，本队伍在得到预测文本后，从原句-预测平行文本中提取字级别的编辑操作，忽略所有与未知标识符相关的操作，并将筛选过后的编辑操作复原为新的预测文本，以此作为最终输出。

配置属性	属性值
模型结构	BART
预训练模型	Chinese-BART-Large(Shao et al., 2021)
最大迭代轮数	20
最大批次长度	2048
优化器	Adam(Kingma and Ba, 2014)
学习率	1e-5
学习率调度器	inverse_sqrt
学习率预热步数	8000
梯度截断	0.1
梯度累计	4
损失函数	Label smoothed cross entropy (label-smoothing=0.1) (Szegedy et al., 2016)

表 5: 模型详细信息

	平均值		最小改动维度		流利提升维度		
	F0.5	F0.5	准确率	召回率	F0.5	准确率	召回率
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	36.40	43.01	22.54
BUPTCL	47.39	59.89	68.88	39.35	34.89	45.96	17.77
NLP的未来	47.12	59.35	65.10	43.84	34.89	42.44	20.39
BERT 4EVER	45.84	58.83	68.96	37.07	32.84	44.87	15.85
BART-base	41.98	54.43	60.10	39.52	29.52	37.62	15.80

表 6: 官方基线模型及前五名的队伍指标

4.3 模型集成

本队伍使用基于编辑的模型集成方法，将原句送入用于集成的 N 个模型中，获得 N 个原句-预测平行文本文件，提取得到所有编辑操作。当某个编辑出现次数大于等于阈值时，该编辑才会保留。

5 实验分析

5.1 实验设定

比赛中，本队伍采用两阶段训练策略，第一阶段直接使用官方提供的训练集和验证集分别作为训练集和验证集，第二阶段以9:1的比例分割官方提供的验证集并分别作为训练集和验证集。在训练过程中，最优模型的选择依据最小改动F0.5和流利提升F0.5的平均值。在模型的推理阶段，束搜索的大小设置为12。实验设备为NVIDIA GeForce RTX 3090。

5.2 评测结果

本次比赛的测试数据集采用封闭方式给出，仅给定原文本，由各个队伍提交结果文件至在线测评平台。在比赛的两个阶段，每支队伍每天均可提交5次测试集结果。表6展示了赛道三的第二阶段中指标超过官方基线模型的队伍，BART-base为官方基线模型，kk为本队队名。

5.3 加噪方式

本队伍探究了两个超参数 β 和 N 在第一阶段训练的加噪过程中的影响，图1所示为动态加噪情况下的验证集得分，且未添加模型后处理环节。加噪比例 β 分别取0.1、0.2、0.3、0.5和0.7，F0.5表示最小改动F0.5和流利提升F0.5的均值，当被加噪的句子占训练集句子的10%时，模型效果最好，当加噪比例持续增大时，噪声将会对模型表现产生负面影响。加噪次数的范围从2次至6次，从图中可知，当被加噪的句子有4个字符被随机替换时，

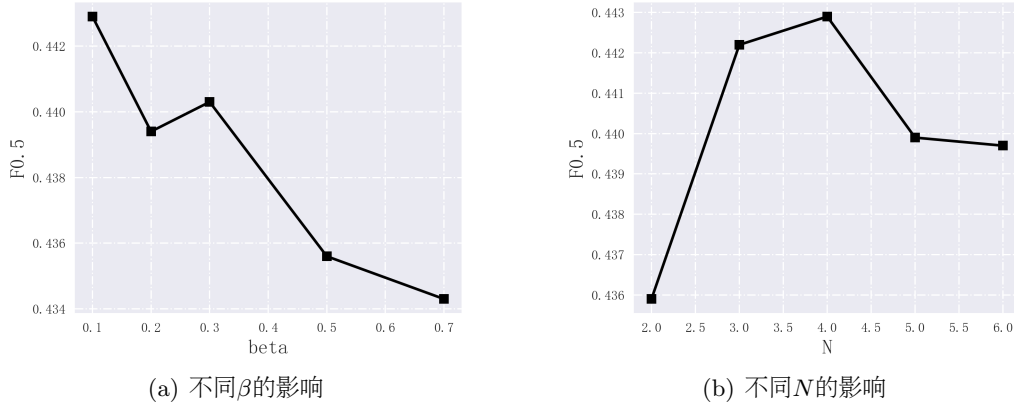


图 1: 加噪比例和加噪次数对模型的影响

	平均值		最小改动维度		流利提升维度		
	F0.5	F0.5	准确率	召回率	F0.5	准确率	召回率
BART-base	41.98	54.43	60.10	39.52	29.52	37.62	15.86
BART-large	44.89	57.72	59.60	50.47	32.26	36.82	21.58
+static noise	45.52	58.10	61.19	48.32	32.93	32.93	22.49
+dynamic noise	45.57	58.13	60.41	50.52	33.00	37.53	22.26
+continuous finetune	48.03	60.82	62.61	54.56	35.24	39.44	24.71
+post edit	50.90	65.56	71.26	49.67	36.23	44.38	20.89
Final single model	55.68	70.84	76.53	54.61	40.52	48.67	24.26

表 7: 第二阶段测试集评测结果

模型效果最好，加噪次数过大或过小都无法使模型在泛化性和鲁棒性之间达到平衡。另外，在同样的加噪比例 β 和加噪次数 N 下，动态加噪方式的模型表现略优于静态加噪方式，表7列出基线模型、使用静态加噪训练集的模型、使用动态加噪训练集的模型分别在第二阶段测试集的得分。

5.4 两阶段训练

考虑到本次比赛的训练集和测试集分别来自于两个数据集，存在分布不一致的问题，所以本队伍使用了两阶段训练策略，先在大量的原始训练集上训练，随后在数量较少的验证集上进行微调。由于验证集存在最小改动和流利提升两个版本，且最小改动验证集的句对数量约为流利提升验证集的5倍，所以最终使用最小改动验证集进行分割并投入训练。如表7所示，微调过后的模型在BART-large的基础上提升了3个点。

5.5 模型后处理

受预训练模型的词表限制，一些常见的中文标点或字符在分词阶段会被[UNK]替代，导致模型输出也带有[UNK]标识。通过分析模型预测结果可以发现，在此次比赛语料中，超出词表范围的字符有中文单双上下引号、生僻字“噫”等。其中，中文单双上下引号出现频率较高，训练集原句中共出现61306次，验证集原句中共出现2238次，第一阶段测试集中共出现112次，第二阶段测试集共出现1216次。如果直接评测模型输出结果，以第二阶段测试集为例，其中的1216次编辑操作都是将中文单双上下引号替换成[UNK]，这些编辑操作必然是错误的。所以本队伍在得到预测文本后，从原句-预测平行文本中提取字级别的编辑操作，忽略所有将原有字符替换为未知标识符的操作，如表7所示，该方法使模型在BART-large的基础上提升了6个点。

6 总结

在本次多维度汉语学习者文本纠错任务中，本队伍使用了基于文本生成的纠错模型。数据

方面，本队伍采用基于规则的动态和静态加噪方式，并且充分利用验证集进行两阶段训练。模型方面，本队伍针对任务特征提出有效的训练策略和后处理方式。评测结果表明，本队伍提出的多种方法均可以使模型性能得到明显的提升，最终在测试集上的总得分为55.74，位列第一。

但本次的系统依然存在一些不足。例如，原始中文BART词表中缺少常见的中文字符，未登录词问题仍待解决。此外，模型在流利提升维度的表现明显低于最小改动维度的表现。以流利提升原则进行文本纠错要求模型挖掘学习者想要表达的语义，并将实际表达的表面形式重组，未来可以研究如何利用汉语丰富的语义信息提升纠错质量。

参考文献

- Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. *arXiv preprint arXiv:1910.02893*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Kai Fu, Jin Huang, and Yitao Duan. 2018. Youdao’s winning solution to the nlpcc-2018 task 2 challenge: a neural machine translation approach to chinese grammatical error correction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 341–350. Springer.
- Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint arXiv:2005.00987*.
- Satoru Katsumata and Mamoru Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 827–832, Suzhou, China, December. Association for Computational Linguistics.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, 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.
- Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. *arXiv preprint arXiv:1904.05780*.
- Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. *arXiv preprint arXiv:1909.01187*.
- Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhashnyi. 2020. Gector-grammatical error correction: tag, not rewrite. *arXiv preprint arXiv:2005.12592*.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of nlptea-2020 shared task for chinese grammatical error diagnosis. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 25–35.
- 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*.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

- Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. 2019. Denoising based sequence-to-sequence pre-training for text generation. *arXiv preprint arXiv:1908.08206*.
- Yingying Wang, Cunliang Kong, Xin Liu, Xuezhi Fang, Yue Zhang, Nianning Liang, Tianshuo Zhou, Tianxin Liao, Liner Yang, Zhenghua Li, Gaoqi Rao, Zhenghao Liu, Chen Li, Erhong Yang, Min Zhang, and Maosong Sun. 2022. Overview of cltc 2022 shared task : Chinese learner text correction.
- Yue 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-HLT*, Online. Association for Computational Linguistics.
- Zewei Zhao and Houfeng Wang. 2020. Maskgec: Improving neural grammatical error correction via dynamic masking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1226–1233.
- 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*, pages 439–445. Springer.
- Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.
- Honghong Zhao, Baoxin Wang, Dayong Wu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2022. Overview of ctc 2021: Chinese text correction for native speakers. *arXiv preprint arXiv:2208.05681*.
- Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. 2018. Chinese grammatical error correction using statistical and neural models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 117–128. Springer.