CCL2022-CLTC赛道四：MuCGEC多参考多来源汉语学习者文本纠错技术研究

张阳[1] 张君牧[2] 左俊祥[1] 孙烁[1]

1.北京海泰方圆公司,北京市海淀区中关村软件园9号
2.北京101中学,北京市海淀区颐和园路11号

摘要

摘要：本文描述了在CCL2022-CLTC（第二十一届中国计算语言学大会-汉语学习者文本纠错任务）赛道四多参考多来源汉语学习者文本纠错（Multi-reference Multi-source Chinese Learner Text Correction）提交的参赛技术报告。针对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，对每一个句子提供多个遵循流利提升的修改答案，我们基于专家经验、多语言模型、模型训练方法改进、PPL集成提供一套融合纠错系统方案。在最终的比赛结果中名列第四。并对基于seq2seq串行多轮集成及并行集成等方法进行了研究分析，对后续的改进提出建议。

关键词： 文本纠错 ； PPL集成

1 引言

1.1 参赛任务描述

中文语法纠错（Chinese Grammatical Error Correction, CGEC）技术旨在对其中存在的拼写、词法、语法等各类错误进行自动纠正。现有的CGEC评测数据集存在着数据量小，领域单一，参考答案数目少的缺陷。针对上述问题，苏州大学、阿里巴巴达摩院联合发布了MuCGEC（Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction）多源多参考中文语法纠错评测数据集，并依托第二十一届中国计算语言学大会（CCL 2022）在阿里云天池平台举办了评测任务。该技术在教育、新闻、通讯乃至搜索等领域都拥有着广阔的应用空间。

1.2 问题和方法总结

文本错误类型不仅包括操作基本的字/词粒度错误，如缺失、冗余、替换、词序错误，还包括语言学级别的错误，如拼写和词序错误等。即错误类型并非单一的拼写或者语法错误。本次参赛系统综合利用文本纠错专家经验、多语言模型、模型训练方法优化改进、PPL集成等方法对问题进行了研究和验证，并对后续的研究提出了改进意见，有助于该技术的演进和在实际行业应用中落地。

1.3 主要经验

在实际研究和验证中发现，不同的模型针对不同的错误效果差异比较大，seq2edit模型不同的Edit标签无论是训练过程还是在实际应用推理，表现都不错，但某些标签如APPEND、REPLACE上的类别过多，导致训练和推理效果欠佳。seq2seq模型训练和推理效果优于seq2edit，尤其在部分类型错误上表现更优，比如乱序、冗余等错误，但经常容易出现过纠的问题。基于拼写的模型在乱序、冗余的错误上表现不佳，但在拼写错误上还不错。另外，具体模型的训练方法上，推荐分两阶段甚至三阶段的方法，效果会有部分提升。

2 背景

2.1 数据集

评测任务数据集由主办方提供，主要的特点：1）参考答案多样：平均每句标注了2.3个基于流利度提升的不同参考答案；2）来源多样：从三种数据源采样数据进行标注；3）标注质量高：采用了严格的三人标注加审核专家审核的标注方式，并且制定了详尽的标注规范。

MUCGEC数据集主要来自汉语学习者，分别采样自以下数据集：NLPCC18测试集（来自于NLPCC18-shared Task2评测任务）、CGED测试集（来自于CGED18-20评测任务）以及中文Lang8训练集（来自于NLPCC18-shared Task2评测任务）。数据的整体统计如下所示：

数据集	句子数量	错误句子数比例	平均字数	平均编辑数	平均答案数
MuCGEC-NLPCC18	1996	1904	29.7	2.5	2.5
MuCGEC-CGED	3125	2988	44.8	4.0	2.3
MuCGEC-Lang8	1942	1652	37.5	2.8	2.1
MuCGEC-ALL	7063	6544	38.5	3.2	2.3

Table 1: MUCGEC数据集

2.2 研究进展

CTC 2021中文文本纠错比赛共有16支队伍参加，包括苏州大学、清华大学、北京理工大学、南京大学等高校，也包括阿里巴巴、上海蜜度、蚂蚁金服、拓尔思、中原银行、华夏银行、人民网等企业。包括了检错和纠错两个任务，分数从5.68到67.32。检错的分数高于纠错的分数。

基于CTC 2021中文文本纠错比赛排名第一的苏州大学、阿里巴巴达摩院联合发布了本次MuCGEC的数据集和baseLine模型，单个模型基于lang8和HSK数据集的精确度/召回率/F0.5值如下表。显然，使用lang8和hsk两个数据集的效果好于单独使用lang8数据

模型	**NLPCC18-Official(m2socrer)**	**MuCGEC(ChERRANT)**
seq2seq$_l ang8$	37.78/29.91/35.89	40.44/26.71/36.67
seq2seq$_l ang8 + hsk$	41.50/32.87/39.43	44.02/28.51/39.70
seq2edit$_l ang8$	37.43/26.29/34.50	38.08/22.90/33.62
seq2edit$_l ang8 + hsk$	43.12/30.18/39.72	44.65/27.32/39.62

Table 2: 模型效果

集。seq2seq和seq2edit模型差异不大。

3 系统

3.1 方法综述

在现有研究的基础上，就此问题的解决来看，技术研究和实际应用不同。针对实际场景应用，为了满足实际应用中实时性的要求，通常考虑单个模型加一些人工规则的方法。针对技术研究，比如本次评测，主要解决的问题是提升准确性，而不是考虑处理的时延，可以考虑多个模型集成的方法。具体到模型集成，可以采用的方法有两大类，一类是串行集成，即依次执行不同的模型，最后输出纠错结果。另外一类是并行模型，即使用多个模型同时计算，使用条件判断来对多个模型输出的结果进行判断是否采纳。本次技术研究对串行和并行集成都进行了验证。

3.2 数据处理

数据处理包括常规数据处理和模型训练词表处理。常规数据处理包括数据格式转换、分句、切词、异常字符处理。模型训练词表处理是指针对中文文本及本次评测的数据，对中文的一些特殊字符进行扩展和补充，保证训练的模型能正确处理待处理文本中的中文特色字符。比

如“”,‘’,【】,《》等只有中文才有的一些符号。针对本次评测，按评测要求，还增加了过滤处理。将可能影响最终评测结果的数据从训练集中删除，避免这些评测数据提前被模型认识和学习，假装识别了这些错误，虚假地提升了模型的评分。

3.3 seq2edit

seq2edit称为sequence-to-edit模型，通常用于中文语法纠错（CGEC）。对于CGEC通常方法有端对端纠错和检错-排序-召回两种思路。seq2edit属于端对端纠错。本方法实际上是将纠错问题抽象为KEEP、DELETE、APPEND、REPALCE等编辑操作进行纠错，从前面研究看，此方法速度快，效果不错。但在实际应用中也有问题，其一是训练时KEEP很容易收敛，DELETE、APPEND、REPALCE这几个操作很难收敛；其二是实际应用KEEP准确性较高，DELETE、APPEND、REPALCE这几个操作准确性欠佳。直觉上看，是这几个操作的数据很不平衡，KEEP操作的数据比例最高，也符合实际数据分布，但DELETE、APPEND、REPLACE这几个操作的数据分布极为不均。尤其是APPEND、REPLACE因为每个字符/词可能被APPEND或者REPLACE的字符/词操作会比较多，同一个模型实际上很难一视同仁的处理其loss。

3.4 seq2seq

seq2Seq称为sequenc-to-sequence模型，属于中文语法纠错（CGEC）的另外一种主流的端对端纠错方法。本方法类似机器翻译模型，可以简单理解为将一个句子（病句）翻译或者生成一个正确句子（纠错之后的句子）。基于文本生成的特点和优势，本方法长处在于解决乱序，缺失等问题，尤其是长距离调序的语法错误。在实际研究和技术验证过程中，对于本次评测数据，seq2Seq模型相对于seq2edit模型，表现更好。

3.5 plome模型

plome模型是Pre-training with Misspelled Knowledge for Chinese Spelling Correction，是腾讯公司研究人员针对中文拼写错误提出的纠错模型。模型解决音近和形近两种类型错误。模型通过将汉字的拼音序列和笔画序列也作为模型输入，并分别用两个子网络来计算其表示向量，让模型可以自动学习任意两个汉字在读音和字形上的相似度。另外，为了让模型更好地学习读音和汉字之间的关系，预训练阶段同时优化汉字预测和拼音预测两个任务。从实际使用效果看，此模型可以作为seq2seq模型的补充，新增的纠错占总的输出0.05左右。

3.6 其他模型

除了以上模型，还对macBert,Ernie,T5等一些模型进行了研究和技术验证。这些模型不能对评测结果实现进一步的效果提升。在最终的集成方案中排除了这些方案。其他方法包括构建困惑集方法、主办方开源的基于句子模板的方法。因为没有足够的时间去构建困惑集，也没有时间增加相应的模板，在最终的集成方案中也被排除掉。

3.7 串行集成方法

所谓串行集成，就是将有错误的病句依次输入不同的模型，得到最终的模型预测结果作为正确的输出句子。本次技术评测，选中的模型包括seq2seq ，plome模型。串行集成方法一，首先将病句输入seq2seq模型，然后将seq2seq的输出结果输入到plome模型，plome模型的输出作为最终结果。串行集成方法二，先将病句输入到plome模型，然后将plome模型的输出结果输入到seq2seq模型，将seq2seq模型作为最终结果。串行集成方法三，先将病句输入到seq2seq模型，然后将seq2seq模型的输出结果输入到第二个seq2seq模型，将第二个seq2seq模型作为最终结果。串行集成方法四，先将病句输入到plome模型，然后将plome模型的输出结果输入到第二plome模型，将第二个plome模型作为最终结果。实际效果如下：seq2seq+plome和plome+seq2seq相比较单个模型都有2-3个百分点的提升。plome+plome模型和单个plome模型相比，几乎不变。而seq2seq+seq2seq模型效果比单个seq2seq模型还差一些。这个可能和单模型的准确率不超过0.5相关。

集成方法	效果
seq2seq+plome	提升
plome+seq2seq	提升
seq2seq+seq2seq	变坏
plome+plome	不变

Table 3: 串行集成方法效果

3.8 并行集成方法

并行集成方法，是选择多个模型对病句进行预测，然后采用判定策略对各模型输出结果进行判断，选择其中某个模型的输出作为输出结果。首先需要选择模型，从测试集数据看，既包括拼写错误，也包括语法错误，所以备选模型需要包括支持语法纠错的模型，也需要包括支持拼写错误的模型。当然，支持同一种错误的模型可以选择多个。其次需要选择判定策略。判定策略可以是针对某个单一错误的字/词策略，也可以是针对整个句子的策略。因为Baseline中提供了针对单一字/词策略错误的策略，本次技术研究采用了针对整个句子的策略。句子判定策略可以有很多，比如针对拼写纠错，可以有句子长度判定策略。针对语法纠错，可以有句子流畅度判定策略。

4 实验

4.1 实验结果

根据对现有研究的了解，对数据集的分析，考虑时间的限制，选择以seq2seq模型,plome模型为基础模型，以lang8,hsk为训练数据和验证数据。处理步骤包括筛选、过滤等数据处理，单个模型训练与验证，根据单个模型的训练表现选择备选模型，对比串行和并行集成方法选择集成方法，综合处理得到输出结果。根据lang8,hsk数据分析，对数据进行过滤，对于超长、同一个样本多次改动、比赛放明确要求过滤的数据进行过滤，最终得到852016个训练样本。

训练分两次进行，训练20轮，调优20轮，然后根据验证数据的表现选择4个备选模型。选择基于句子困惑度作为整句判定策略，4个模型输出的结果，依次计算PPL，选择最小的PPL作为最终的输出结果。单个模型30+分，经过4个模型的并行集成，在最终6000个样本的测试集上的分数为43.09，提升了约7到10个百分点。将集成结果再次输入到seq2seq模型，召回率大幅上升到38.16，但准确率从47.71下降到35.59，导致得分下降到36.08。

模型	F0.5	准确率	召回率
s2s+plome并行集成	43.09	47.71	31.06
s2s+plome并行集成-s2s串行集成	36.08	35.59	38.16

Table 4: 集成模型得分

4.2 定量分析

2个s2s模型，2个plome模型集成得分为43.09，其中各备选模型的纠错贡献如下表所示：

模型	命中句子数量	占比
rule	634	0.11
s2s1	3747	0.62
s2s2	1295	0.22
plome1	246	0.04
plome2	78	0.01
total	6000	1.00

Table 5: 定量分析-各模型贡献

4.3 错误分析

在实际使用时，采用PPL整句判定策略总体表现还不错，但也有部分判断错误的情况。如下表所示，模型2正确进行了预测，但PPL高于模型1。错误示例：

模型	预测	ppl分数
原句	每天会有不少的毒气体泄漏从工厂里出来。	49.419
模型1	每天会有不少的有毒气体泄漏从工厂里出来。	40.412
模型2	每天会有不少的毒气体从工厂里泄漏出来。	41.626
模型3	每天会有不少的毒气体泄漏从工厂里出来。	49.419
模型4	每天会有不少的毒气体泄漏从工厂里出来。	49.419

Table 6: 模型集成-错误示例

其他PPL判断错误，比如，输入为“北纬37º33’，东经127º为暖温代气候。”，模型输出“北纬37—33—，东经127—为暖温带气候。”的PPL比模型输出“北纬37º33’，东经127º为暖温带气候。”低，导致判断错误。输入“我14～15岁的时候有热的歌星。”，输出为“我14—5岁的时候有个热门的歌星。”的PPL比输出“我14～15岁的时候有热的歌星。”的PPL低，导致判断错误。

可以看出，基于PPL的判定总体效果不错，但也严重依赖于PPL的语言模型，对于特殊字符，专用名词等识别不一定准确。另外，单独依赖PPL做判断只考虑了输出，没有考虑输入，会出现两个模型输出都很流畅，但不一定符合输入期望纠错的要求。比如，输入“这样，你就会尝到泰国人死爱的味道。”，模型2输出“这样，你就会尝到泰国人最爱的味道。”，模型3输出“这样，你就会尝到泰国人喜爱的味道。”两个输出都很流畅，但根据输入，应该选择模型3的输出，由于模型3的PPL高于模型2的PPL导致PPL出错。

5 总结

总的来看，针对多源多参考中文语法纠错任务，目前的纠错水平还比较低，准确率0.40左右，召回率低于0.35，很难达到实际商用要求。还需要有新的技术突破，才能在行业应用落地。

从研究和实际验证来看，可能的改进路径包括几方面，第一是数据集，需要有针对性的选择数据集解决不同的任务。比如针对格式相对规范的政务文本、社交短文本、外国人学习中文的数据差异明显，不适合用社交短文本的数据来训练政务文本的模型。在MUCGEC数据集上F0.5分数在0.36的模型应用于社交短文本模型F0.5的分数仅有0.16。第二，模型，尤其是单模型的效果要得到明显提升，首先要提升的是准确率，单模型的准确率如果低于0.5，无法采用串行集成的方法提升总体效果。目前主流包括seq2edit和seq2seq两种模型。其中，seq2edit模型改进应该是在标签太多的APPEND和REPLACE，尤其是APPEND需要改进，可以考虑从数据构造，标签构造和LOSS计算方法进行改进；seq2seq模型改进点应该在灵活度的控制上。第三，模型集成，如果单模型的效果没有明显提升，使用串行集成方法可能会导致模型效果下降。对并行集成方法，可以考虑结合单一字/词错误和整句PPL方法融合，并需要考虑计算PPL的语言模型的调优以及考虑输入数据错误点的权重。第四，针对行业专用领域，需要考虑领域内的专用名词/短语被误纠的问题。第五，使用Bart模型进行训练，2张16GB的GPU卡，训练85万数据，每轮需要6-7小时，训练速度过慢，可以考虑优化训练速度。

参考文献

Wang, Yingying and Kong, Cunliang and Liu, Xin and Fang, Xuezhi and Zhang, Yue and Liang, Nianning and Zhou, Tianshuo and Liao, Tianxin and Yang, Liner and Li, Zhenghua and Rao, Gaoqi and Liu, Zhenghao and Li, Chen and Yang, Erhong and Zhang, Min and Sun, Maosong. 2022. *Overview of CLTC 2022 Shared Task : Chinese Learner Text Correction*

王莹莹,孔存良,刘鑫,方雪至,章岳,梁念宁,周天硕,廖田昕,杨麟儿,李正华,饶高琦,刘正皓,李辰,杨尔弘,张民,孙茂松. 2022. *CLTC 2022*: 汉语学习者文本纠错技术评测及研究综述

Zhang Yue, Li Zhenghua, et al. 2022. *MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,pages:3118–3130 2022-mucgec.

Shulin Liu, Tao Yang, et al. 2021. *PLOME: Pre-training with Misspelled Knowledge for Chinese Spelling Correction Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*,pages:2991–3000