CCL2022-CLTC赛道[三]：[双模型结合的多维度文本纠错系统]

宋思琦，耿磊，吕奇，艾春辉，闫旭，曹自强*
苏州大学，人工智能研究院
{sqsong, lgeng, aopolinqlv, chai, xuyanlp}@stu.suda.edu.cn

摘要

本文描述了我们在CCL2022赛道三多维度汉语学习者文本纠错比赛中提交的参赛系统。多维度汉语学习者文本纠错要求在最小改动和流利提升两个维度对汉语学习者文本进行语法纠错，具有一定的难度。我们提出了一个基于序列到序列和序列到编辑的双模型结合的纠错系统，来互相补充的解决文本中的语法错误。除此之外，我们探究了多参考数据和单一参考数据作为训练集和开发集的不同表现。最终，我们提交的系统在测试集上的最小改动维度F0.5、流利提升维度F0.5和总得分均排列第二。

关键词： 序列到序列 ； 序列到编辑 ； 文本纠错

CCL2022-CLTC Track [3]: [A Multi-dimensional Textual Error Correction System Combining Dual Models]

Siqi Song, Lei Geng ,Qi Lv, Chunhui Ai, Xu Yan, Ziqiang Cao*
Soochow University, Institute of Artificial Intelligence
{sqsong, lgeng, aopolinqlv, chai, xuyanlp}@stu.suda.edu.cn

Abstract

This paper describes the system we submitted for the CCL2022 Track 3 Multidimensional Chinese Learner Text Correction Competition. Multidimensional Chinese learner text error correction requires grammatical error correction of Chinese learner text in both the minimal change and fluency enhancement dimensions, which is challenging. We propose an error correction system based on a combination of sequence-to-sequence and sequence-to-edit dual models to complement each other in solving grammatical errors in the text. In addition, we explored the different performance of multiple reference data and single reference data as training and development sets. In the end, our submitted system ranked second in the test set in terms of minimum change dimension F0.5, fluency improvement dimension F0.5 and overall score.

Keywords: Sequence-to-sequence , Sequence-to-edit , Text Correction

1 任务介绍

*代表指导老师
评测系统代码 https://github.com/47777777/ccl-track3

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。

Figure 1: 多参考中文语法纠错任务示例

汉语学习者文本纠错任务 [1]（Chinese Learner Text Correction，CLTC）旨在自动检测并修改汉语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。该任务属于综合性的自然语言处理研究方向，能够比较全面体现自然语言处理的技术水平，近年来越来越受到关注，也出现了一些有潜在商业价值的应用。

赛道三的多维度汉语学习者文本纠错（Multidimensional Chinese Learner Text Correction）是汉语学习者文本纠错任务下的一个分支，其重点在于多维度。由于同一个语法错误能从语法的不同角度划分为不同的性质和类型 [2]，并且随着语言使用场景和具体需求的变化，同一个语法错误存在着多种正确的修改答案。针对这种多参考答案的情况，多维度汉语学习者文本纠错任务在最小改动（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。具体样例如图1所示，原句在两个维度均有多个语法纠错的参考答案，其中，加粗表示替换字符，斜体表示插入字符，删除线表示删除字符。

赛道三的输入为可能含有错误的句子，输出和样例不同，在最小改动和流利提升两个维度各只需要提供一种修改后的结果即可。

2 任务数据

本次比赛，赛道三官方提供了经过处理的NLPCC2018-GEC12 [6] 发布的采集自Lang8 平台的中介语数据(1213457条)作为训练集。参赛者仅允许使用该数据用于训练。

还提供了两个维度的多参考数据集YACLC-Minimal8(1839条)、YACLC-Fluency8(1839条)作为开发集。其中YACLC-Minimal 属于最小改动维度，YACLC-Fluency属于流利提升维度。开发集来源为汉语学习者文本多维标注数据集YACLC [5]。它是一个大规模、高质量、篇章级别、多维度、多参考的中文语法纠错数据集。标注实践中采用众包策略，在搭建的可供多人同时使用的在线标注平台上分组、分任务、分阶段地进行标注和审核工作。

训练集和开发集都是多参考数据集，我们对开发集进行了分析，发现数据中包含多种错误类型，大致分为冗余、缺失、替换、乱序四种，复杂的是一句话中往往存在多种错误，如表1所示，这给纠错任务带来了很大的困难。同时这种一个句子中包含多种错误的情况，也让我们无法具体统计出这些错误的个数。

原句	修改后的句子	包含的错误类型
她是个中国人，出生在东北。	她是个中国人，出生在东北。	替换、冗余
-我认识有些字文章里。	-我认识文章里的一些字。	乱序、缺失、替换

Table 1: 修改示例及错误类型

综上1、2节所述，我们认为本次比赛的难点在于：(1)和以往的汉语学习者文本纠错评测不同，需要综合考虑Minimal和Fluency两个维度。(2) 参赛者仅允许使用赛道官方提供的数据用于训练，所以不能进行预训练。（3）从数据中并无法清晰的分析出，不同错误类型具体的数量情况。（4）一条错误文本中可能含有多种错误类型。

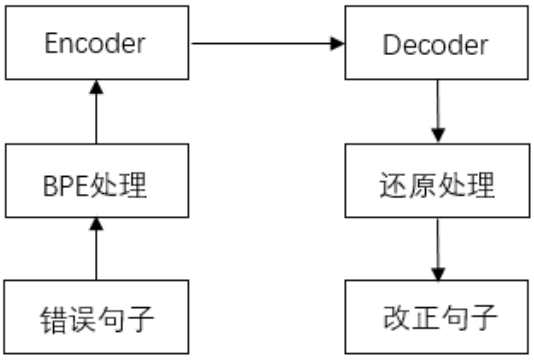

Figure 2: 基于序列到序列的语法纠错模型推理流程

3 参赛系统

本次比赛，我们采用了双模型纠错的方法，在限定训练集的条件下，尽可能的提升修改效果：1.我们首先使用基于序列到序列的语法纠错模型进行纠错；2.然后我们使用基于序列到编辑的语法纠错模型进行纠错；3.将两种纠错结果进行一定策略的合并。

3.1 基于序列到序列的语法纠错模型

虽然没有办法对数据集错误类型的分布进行具体的统计，但是通过观察可以发现绝大多数错误都涉及到缺失和冗余等语法错误，因此我们使用语法纠错模型对其进行更改，我们首先使用了基于序列到序列（seq2seq）的语法纠错模型，对于一个用于语法改错任务的seq2seq模型，其基本的训练数据为一个由原始句子和正确句子所组成的改错句对 [3]。输入为错误句子，使用encoder-decoder 模型，直接输出改正后的句子。这里用到的是赛道官方提供的baseline版本。为了能够最大限度的提升效果，我们使用了bart-large作为预训练模型。在句子输入模型之前，需要先做BPE处理，同理输出的句子也是BPE形式的，需要还原处理成原来的形式，我们使用的方法是将BPE结果文件中多余的空格删除，由于我们忽略了一些句子中的词并不在词表里，因此有少部分修改后的句子里包含[UNK] 和## ，对于这种情况，我们没有对其进行修改，直接让修改句子等于原始句子。推理流程如图2所示

3.2 基于序列到编辑的语法纠错模型

受到训练数据的限制，只使用基于序列到序列这一种语法纠错模型不能很好的进行纠错，很多错误并不能被发现出来从而纠正，因此我们使用了第二种语法纠错模型，目前基于序列到编辑（seq2edit）的最优语法纠错模型GECToR [4]，目的是对第一种模型没有纠正的错误进行补充纠正。GECToR模型的思路是将语法纠错任务转化成序列标签任务，给每一个token打标签，这些标签类型如表2，由于训练数据只是错误-正确的句子对，因此首先需要把输入转换成对应的变换标签，然后送入模型。GECToR的模型结构就是类似BERT的Transformer模型，在最上面加两个全连接层和一个softmax。通过标签的变换，可实现插入、删除、替换等纠错操作，也可进行多轮迭代打标签，直到没有发现新的错误，输出最后的结果，推理流程如图3所示。

标签类型	说明	个数
$KEEP	当前token保持不变	1
$DELETE	删除当前token	1
$APPEND	在当前token后增加其他token	10268
$REPLACE	把当前token替换成其他token	10268

Table 2: GECToR模型的标签类型

3.3 双模型的纠错结果合并

我们的集成方法是将序列到序列模型的修改方案作为第一选择，GECToR模型的修改方案

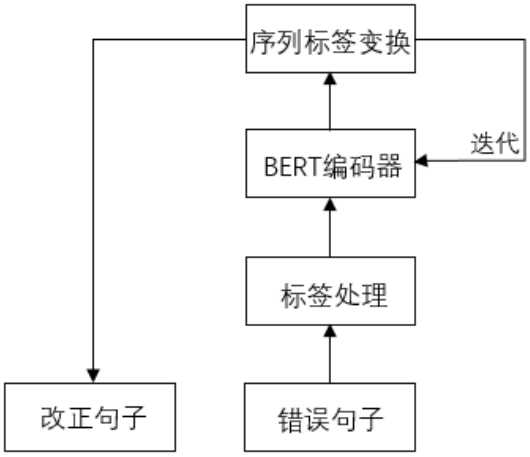

Figure 3: GECToR模型推理流程

作为第二选择，这是因为单独使用seq2seq模型时的结果要比单独使用GECToR模型时的结果好。将得到的两种选择相加，即第一选择进行修改了句子则令其作为最后的结果，否则令第二选择作为最后的结果。我们观察到这种直接相加的集成方法一定程度上会提高改正的结果，但是不能忽视的是，在某些句子中，seq2seq模型的修改方案是错误的，而GECToR模型的修改方案是正确的。为了尽可能减少这种情况，我们提高了seq2seq模型的修改方案的正确率的方法,，在seq2seq模型的修改结果中，根据输出分数的高低，对于分数小于-0.45分的，使用seq2edit模型的修改方案。

4 实验

4.1 训练设置

两种语法纠错模型如3.1，3.2节所述，具体训练流程为：（1）对于基于序列到序列的纠错模型，我们使用bart-large作为预训练模型进行初始化，并根据其输出的分数，只保留大于-0.45分的修改结果，其余的句子不改变，得到纠错结果A。（2）对于基于基于序列到编辑的纠错模型，我们使用macbert作为预训练模型进行初始化，得到纠错结果B。（3）将纠错结果A作为第一选择方案，纠错结果B作为第二选择方案，如果A相对于原句进行了修改，则最后结果把A中的句子作为修改结果，否则选择B中的句子为修改结果。

在给定的训练集和开发集基础上，我们进行了多种尝试，发现：1.在训练集加入句子的不同修改结果，相对于只有一条修改结果，在最小改动和流利提升两个维度的结果都会提升，这也说明了，给定多种修改方案对于拼写纠错的任务是有帮助的。2. 在开发集加入句子的不同修改结果，相对于只有一条修改结果，在最小改动和流利提升两个维度的结果都会下降。因此我们最后的方案是训练集使用多种参考答案，开发集只使用一种答案。

4.2 测试集结果

我们首先分别用bart-base初始化的seq2seq纠错模型(s2s-bart-base)和bert初始化的GECToR纠错模型(gec-bert)进行了测试，然后将其结果相加后又进行了测试(s2s-bb_gec-b)，发现将两种模型的纠错结果相加后分数有很大上升，随后我们将bart-base换为bart-large，将bert换为macbert，进行了测试(s2s-bl_gec-m)。最后又根据seq2seq模型的输出分数作为超参数，分别在-0.5和-0.45这两个值上进行了测试，不同方案的测试集结果如表3所示。我们选取(s2s-bl_gec-m_-0.45)作为最后的提交方案。最终模型在测试集上的Minimal F0.5、Fluency F0.5和总F0.5指标上排名第二。

方案	Minimal F0.5	Fluency F0.5	average F0.5
s2s-bl_gec-m_-0.45	64.42	36.4	50.41

Table 3: 测试集结果

5 总结

在本次CCL2022赛道三多维度汉语学习者文本纠错的评测任务中，我们使用了基于序列到序列和基于序列到编辑，两种语法纠错模型相结合的纠错方案，并探索了多参考答案数据和单一参考答案数据作为训练集和开发集的不同表现。实验表明我们提出的两种模型相结合的方法能在限定训练集的条件下，使纠错的效果得到有效提升，最终提交的系统在测试集上的总分为50.41，排在第二名。

但是，本次的提交系统还存在着许多不足。例如超参数的选择，我们只尝试了-0.5和-0.45，并没有更多尝试其他的值。此外，两种模型纠错结果的结合方式，相加的方式还是过于粗糙了，我们观察到赛道五的任务是帮助在多种纠错方案中选择最优的方案，也许有选择性的结合方式能够达到更优的效果，也是未来我们准备努力的方向。

参考文献

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.

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, 01:128-140.

Ge T , Wei F , Zhou M . Fluency Boost Learning and Inference for Neural Grammatical Error Correction[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, Oleksandr Skurzhanskyi. GECToR – Grammatical Error Correction: Tag, Not Rewrite. BEA 2020.

Yingying Wang, Cunliang Kong, Liner Yang, Yijun Wang, Xiaorong Lu, Renfen Hu, Shan He, Zhenghao Liu, Yun Chen, Erhong Yang, and Maosong Sun. 2021. YACLC: A Chinese Learner Corpus with Multidimensional Annotation. arXiv preprint arXiv:2112.15043.

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 (NLPCC), pages 439–445.