CCL2022-CLTC赛道五：语法纠错质量评估

宋瑞林，任君翔，王思博，成臻，武悦娇，刘浪，尹文博，高依舟，张鹏涛

中国太平洋保险（集团）股份有限公司/ 上海
songruilin@cpic.com.cn

摘要

汉语文本纠错目的是检测汉语文本中的标点符号、汉字拼写、语义等错误。近些年来，纠错任务获得了越来越多的关注，并出现了具有商业价值的落地场景。今年主办方组织了汉语学习者文本纠错评测，共有5个赛道，本篇测评报告针对其中的第五赛道语法纠错质量评估，旨在评价语法纠错模型修改结果的质量，评测代码已开源在https://github.com/Babysong12/CCL2022-CLTC-Track5。

关键词： 文本纠错 ；质量评估

CCL2022-CLTC Track 5: Quality Estimation of Grammar Correction

Ruilin Song, Junxiang Ren, Sibo Wang, Zhen Cheng, Yuejiao Wu,
Lang Liu, Wenbo Yin, Yizhou Gao, Pengtao Zhang

China Pacific Insurance (Group) Co Ltd / Shanghai, China
songruilin@cpic.com.cn

Abstract

The purpose of Chinese text correction is to detect errors in punctuation, Chinese character spelling, and semantics in Chinese text. In recent years, the task of text correction has gained more and more attention, and landing scenarios with commercial value have emerged. This year, the organizers organized a text correction evaluation for Chinese learners, with five tracks. This evaluation report focuses on the fifth track of grammatical correction quality, aiming to evaluate the quality of grammatical correction model modification results. The code is available in https://github.com/Babysong12/CCL2022-CLTC-Track5.

Keywords: Text correction , Quality estimation

1 引言

语法纠错质量评估任务通过预测每个语法纠错结果的质量评估分数，评估语法纠错模型修改结果的质量，分数通过句子级别和词级别的质量评估分数得到 (王莹莹et al., 2022)。该任务有很高的应用价值，由于语言场景和具体需求不同，有很多种的正确修改方案，通过该质量评

估分数，可以对多个纠错结果进行分数排序，后续可以根据结果进一步提升纠错质量效果。尽管赛题中将皮尔逊相关系数作为评测标准之一，但是在最终结果中只关注分数最高的语法纠错结果的效果。

本赛道官方提供了两个baseline，都使用BERT-BASE-CHINESE (Devlin et al., 2018)预训练语言模型，其中一个使用Softmax方法计算改正句的质量评估分数，另外一个使用Sigmoid函数计算改正句的质量评估分数。本次评测主要使用了后者。

1.1 背景

本赛道的训练集提供了中文Lang8数据，并基于seq2seq使用BART-large训练了语法纠错模型。此后将该模型解码过程中每个原句排名前10（小于等于10）的修改结果作为质量评估的语法纠错方案，并给出了每个修改方案的真实F0.5分数。其中，带语法纠错方案的数据格式如下所示：

```
{
        "idx": 原始句子id,
        "src":原始句子,
        "hpys":
        [
                {
                        "idx": 修改句子id,
                        "text":修改句子1,
                        "p": 准确率,
                        "r":召回率,
                        "f05":f0.5分数
                }, {
                        "idx": 修改句子id,
                        "text":修改句子2,
                        "p": 准确率,
                        "r":召回率,
                        "f05":f0.5分数
                }
        ]
}
```

Figure 1: 数据集格式

验证集提供流利提升（fluency）和最小改动（minimal）两个维度的数据，其中最小改动指尽量维持原句结构的情况下，尽量少地增删、替换句子里的词语使句子符合汉语语法规则；流利提升指进一步修改句子使其更加地道和流畅，符合汉语母语者的表述习惯。验证集的数据格式与训练集相同。

测试集同样提供fluency和minimal两个维度的数据，包含原始句子及对应的修改句子，分别从两个维度对修改的质量进行评估，输入原句和修改句，得到每个句子对的F0.5分数，并对其排序，选取每个原句分数最高的修改句及其分数，最后提交评估结果。

本次评测基于数据集结构和Baseline，尝试了不同的优化方案，例如修改损失函数、判断原句是否需要纠错等，我们会在后续章节具体讨论优化方案。

2 方案思路

2.1 Baseline

评测时首先尝试了官方的baseline，使用了BERT预训练模型，将原句和修改句分别配对输入模型，通过sigmoid函数计算改正句的质量评估分数，并使用MSE作为损失函数。

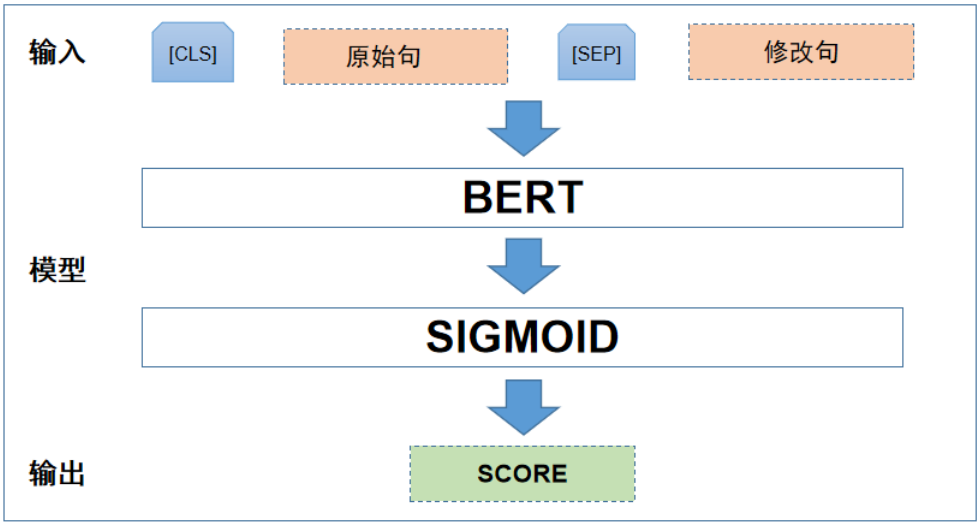

Figure 2: 模型框架

2.2 对抗训练

通过对embedding层添加微小扰动使用对抗训练提高模型的泛化能力，本次参赛主要使用了EMA、FGM (Miyato et al., 2016)及两者组合三种策略。

2.3 “原文+修改句”，标签=0/1

我们对每个原句对应的修改句分数进行处理，选择原句对应修改句分数最高的为1，其余分数为0，用BCE作为损失函数。

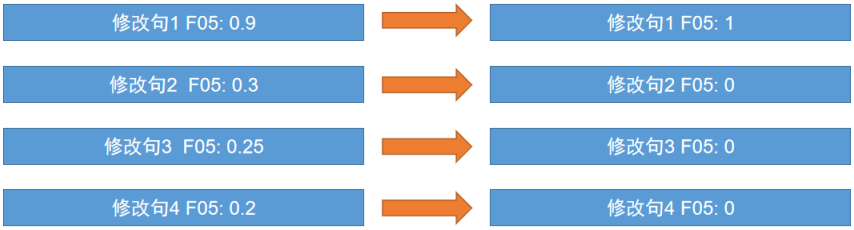

Figure 3: 标签调整

2.4 “原文+全部修改句”，标签=f05

由于之前的方案都只是单独计算了修改句的评估分数，因此我们打算将同一个原句的修改句之间的关系考虑在内。为方便计算，对于少于10句修改句的原句，用分数最低的修改句补齐到10句，继而输入到模型用softmax计算分数，并用交叉熵作为损失函数。

2.5 原句二分类模型

根据对数据集的观察和梳理，我们发现有部分原句本身就是正确的，其修改句的最高f05为1，且与原句完全相同，因此我们将原句正确的数据标签设置为1，其余则为0，训练一个二分类模型判断原句是否正确。

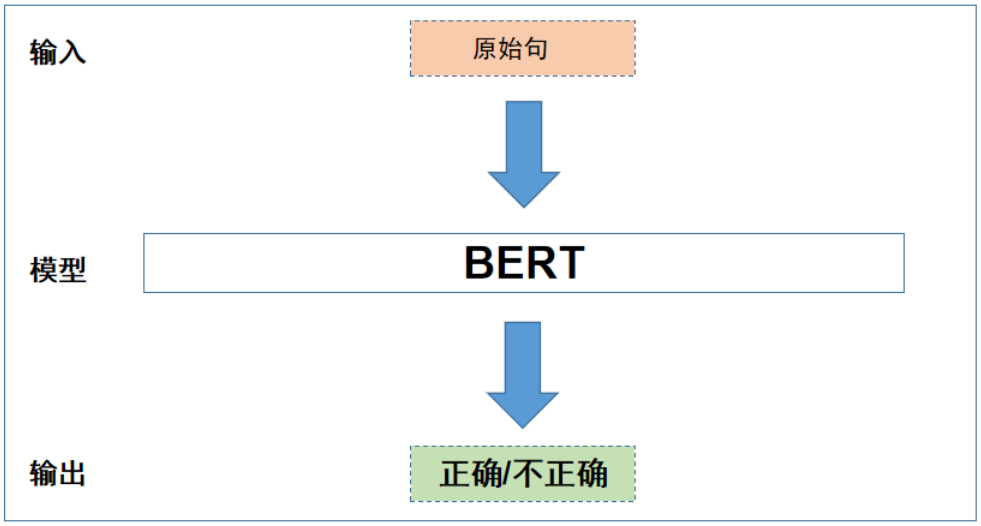

Figure 4: 原句二分类

2.6 集成融合

我们尝试了两种不同的融合策略，一种对所有修改句的F0.5求平均数，选出最高的作为修改质量最好的句子；另一种是先筛选出每个模型的原始句对应的最优修改句，再通过投票法得到每个原始句对应的最优修改句。

2.7 数据后处理

经过对数据结构的观察，我们发现部分错误是由于标点符号的替换，这部分用原来的模型非常难识别，所以选择了训练集中对标点符号进行替换的数据，单独训了一个分类模型。

以测试集的数据为例：对于原句中有逗号“，”无顿号“、”且选项中有顿号出现的情况，如果选项中仅有一个出现顿号，则赋予该选项最高分，很可能是将两个词语“A，B”改为“A、B”的情况。而对于出现多个顿号的选项，需要判断顿号的数量，如果大于1，则选择有“和”的选项，这种情况通常为多个词语并列，按照汉语母语者的表述习惯，最后会以“和”结束，例如：“我有A、B、C和D。”按照此类梳理的规则，可对模型输出后的结果进一步调整优化，提高模型效果。

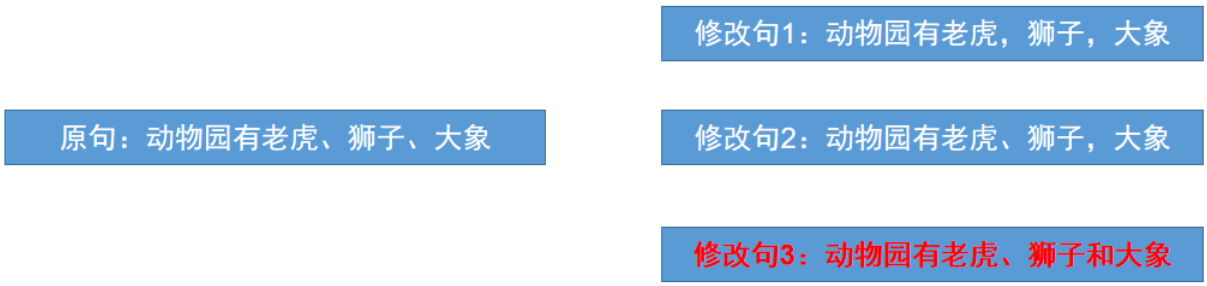

Figure 5: 数据处理

3 实验

本次参赛仅使用了主办方提供的训练集，并未使用额外的开源数据集。首先测试了官方的baseline及替换不同预训练模型的结果，实验结果见下表，表中的数据为各维度的F05值。

PTM	Fluency	Minimal	AVG
BERT-base	27.42	47.93	37.675
MacBERT-base	28.02	49.51	38.675
ERNIE	28.45	48.63	38.54
RoFormer	26.45	45.43	35.94
NEZHA	28.1	49.68	38.89
ELECTRA	27.76	52.16	39.96

Table 1: 不同预训练模型的实验结果

如上表所示，ELECTRA模型效果最好，因此我们决定采用ELECTRA (Kevin et al., 2020)预训练模型进行后续的实验。之后我们使用对抗训练，通过对embedding层添加微小扰动提高模型的泛化能力，使用EMA利用滑动平均的参数来提高模型在测试数据上的健壮性，及FGM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，主要使用了EMA、FGM及两者组合三种策略，实验效果如下：

对抗训练	Fluency	Minimal	AVG
FGM	28.45	48.63	38.54
EMA	27.76	52.4	40.08
FGM+EMA	28.45	52.16	40.305

Table 2: 不同对抗训练的实验结果

实验发现不同策略对模型效果有着不同程度的影响，其中FGM和EMA组合效果最好。后续我们采用了前一章节提及的不同方案思路，并更换随机种子，得到30个不同的单模，并使用前一章节的融合策略，尝试用投票法更换阈值测试结果，并进一步通过后处理，最终结果为47.905。

4 总结

本次参赛最终平均F05取得了47.905，仍有许多可以完善的地方。例如，很多模型PCC分数很高，但是F05值偏低，由于本次比赛排名只关注F05值，因此我们把注意力都集中于如何提升最优修改句的评估效果，而对于识别不同修改句的评估质量同样是很有帮助的，未来可以考虑把提升PCC作为优化目标，找到同时能提升F05和PCC两个指标的策略。

参考文献

Miyato T, Dai A M, Goodfellow I. 2016. *Adversarial Training Methods for Semi-Supervised Text Classification.*

Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning 2020. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.*

Devlin J, Chang M W, Lee K, et al. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding.*

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.