CCL2022-CLTC赛道一：csc_runner队评测报告

刘旺旺
视源电子科技有限公司/ 广州
liuwangwang@cvte.com

摘要

本文档描述了我们在CCL2022-CLTC赛道一，中文拼写检查任务中提交的参赛系统。该系统主要针对汉语学习者所产生的拼写错误进行检测纠正。目前，能获取到的汉语学习者拼写检查数据不多，评测数据可能出现连续错误，实体错误等，因此处理起来有一定难度。模型方面，我们集成了多个预训练模型，并且增加多个后处理模块。数据方面，我们使用基于规则的方法，利用混淆集构造了大量的数据，缓解了数据不足的问题。最后，我们模型在评测数据集上，句子级纠错F1值达到了81.08。

关键词： 中文拼写检查 ； 预训练模型 ； 数据增强

CCL2022-CLTC Track 1: csc_runner Team Report

Wangwang Liu
CVTE / Guangzhou
liuwangwang@cvte.com

Abstract

This document describes the system we submitted in the CCL2022-CLTC Track 1, Chinese Spelling Check task. The system mainly detects and corrects the spelling errors produced by Chinese learners. At present, there are not enough spelling check data produced by Chinese learners, and the evaluation data may have continuous errors and entity errors, so it is difficult to deal with them. About the model , we integrate multiple pre-trained models, and add matching error correction and rule processing modules. About training data, we use rule-based methods to construct large amounts of data by using the confusion set, which alleviates the problem of data shortage. Finally, our model achieves a sentence-level error correction F1 value of 81.08 on the evaluation data.

Keywords: Chinese Spelling Check , Pre-training Model , Data Augmentation

1 引言

中文拼写检查，即对输入文本中所包含的拼写错误进行检测和纠正，其中拼写错误包含：音近、形近、形音兼近三种。它对教育，出版，搜索引擎等领域都有着至关重要的作用。拼写错误不仅会影响阅读，而且可能完全改变文本传递的意义。赛道一主要关注汉语学习者所产生的拼写错误，这与以汉语为母语的人产生的错误分布差距比较大。

我们的参赛系统使用了3种预训练的模型，训练并预测输入文本每个位置正确的汉字，进行了最多三个阶段的训练，最后进行集成。对于集成后的结果，还进行了四个后处理：（1）为了能够修改更多的错误，尤其是连续错误，进行多次前向，迭代修改。（2）为了减少误修改，对非音近，形近的修改，进行还原。（3）为了能更好的处理成语错误，增加了成语匹配纠错模块，若能够收集到足够多的实体，该方法同样使用实体纠错。（4）为了能够更好的利用验证集，模型集成后的结果再经过验证集训练的模型进行纠错，相当于先改通用错误，再改领域错误。

我们的系统主要参考论文Li et al. (2021) 与CTC-2021中文纠错比赛第一名的报告[0]，具体的实现细节可查阅代码[1]

2 数据分析

该任务对汉语学习者所产生的拼写错误，进行检测纠正。给定输入句子，输出拼写错误的位置及对应的修改结果。如Table 1所示，“8”“11”为两个错误位置，“个”“逊”为对应位置的修改结果。如果该句没有错误，则输出“（YACLC-CSC-VALID-ID=0161）0”即可。

原句	(YACLC-CSC-VALID-ID=0161) 两个公司是同一的亚马孙，不是吗?
输出	(YACLC-CSC-VALID-ID=0161), 8, 个, 11, 逊

Table 1: 中文拼写检查示例

比赛给出了带标注的验证集170句，验证集的分布情况见Table 2，错误句子数占总数一半，平均每句话约有1个错误。使用ChERRANT工具，对验证集进行错误词性统计分析。共包含95个错误，其中单字错误个46，多字错误49个（可与相邻汉字组成词），词性分布如下Figure 1 所示。

句子数	170
有错误句子数占比	0.5
字数/句子数	18
错误数/ 错误句子数	1.16

Table 2: 验证集分布情况

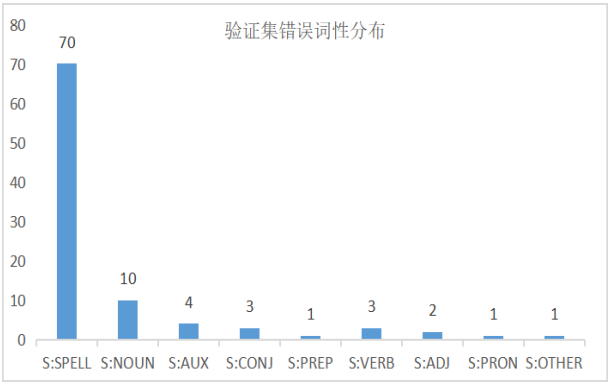

Figure 1: 验证集错误词性分布

[0]https://github.com/HillZhang1999/CTC-Report/blob/main/Report.pdf
[1]https://github.com/wangwang110/track1_csc_runner

比赛给出了无标注的测试集1100句，测试集最小长度8，最大长度71，平均长度8。该测试集有200条数据在NLPCC2018提供lang-8数据集[2]中，比赛提供了一份新的lang-8数据[3]，要求只能用该版本。更具体详细的评测数据以及任务相关信息，可参考CCL2022-CLTC评测概述Wang et al. (2022)。

3 系统

3.1 模型

我们使用的模型结构与比赛给出的baseline基本一致，使用bert Kenton and Toutanova (2019)系列的预训练模型作为编码器, 获得输入文本中的每个字符的向量表示，然后通过全连接层，预测每个位置正确的字符。与baseline不同之处在于，将用于mlm任务预训练的全连接层参数，也用于拼写纠错模型全连接层的参数初始化，如Figure 2 所示。

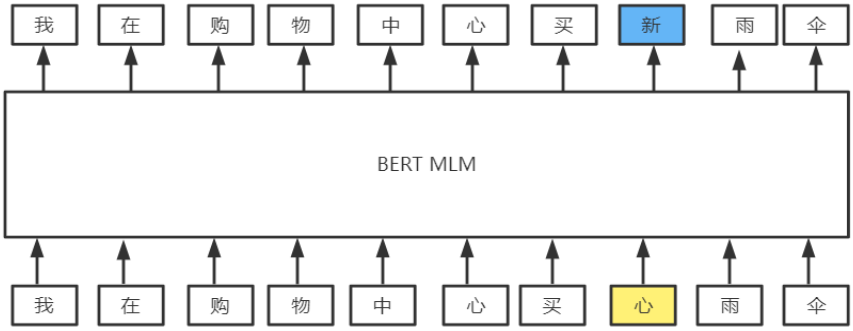

Figure 2: 模型结构

3.2 训练数据

中文拼写纠错训练语料比较少，我们获取了比赛提供的拼写纠错数据以及其他公开的中文语法纠错等多个来源的通用标注数据，并且人为构造了大量的数据。

拼写纠错数据:

（1）SIGHAN：2013到2015年间，台湾大学等组织联合举办的SIGHAN拼写纠错评测任务公开的数据集，这里使用的是比赛提供版本。

（2）Wang271K：论文Wang et al. (2018) 通过给ocr的输入和语音识别的输入加噪声，从而获得错误句子，与原始句子构成平行语料，这里使用的是比赛提供版本。

语法纠错数据:

（1）lang8：根据Lang-8语言学习网站中母语者对汉语学习者作文的修改记录生成，这里使用的是比赛提供版本。

（2）hsk：北京语言大学对汉语学习者参加汉语水平考试中写的作文组织人员进行了语病标注。

（3）cged[4]：中文句法错误诊断技术评测自2014-2022所发布的数据。

（4）ctc2021[5]：CTC2021中文文本纠错比赛公开数据集。

（5）wikideits[6]：中文维基百科编辑历史抽取并过滤得到的语料。

对于语法纠错数据，首先使用ChERRANT[7]工具进行标注，根据标注结果，选择Figure 1 中占比较高的类别S:SPELL，S:NOUN，S:AUX，S:VERB，S:CONJ并且修改之后未改变长度的错误。然后，根据选择的错误，将正确句子进行修改后作为原句。最后使用的各数据集统计

[2]http://tcci.ccf.org.cn/conference/2018/dldoc/trainingdata02.tar.gz
[3]https://github.com/blcuicall/CCL2022-CLTC/tree/main/datasets/track1
[4]https://github.com/blcuicall/cged_datasets
[5]https://github.com/destwang/CTC2021
[6]https://github.com/xueyouluo/wiki-error-extract
[7]https://github.com/HillZhang1999/MuCGEC/tree/main/scorers/ChERRANT

如Table 3，其中lang8，hsk，cged，ctc2021，wikideits经过去重处理，并且仅仅包含有错误句子对。

通用标注语料	全部数据量（句）	拼写纠错数据量（句）
SIGHAN	6461	6461
Wang271K	271281	271281
lang8	1076336	43327
hsk	91175	20604
cged	45245	3560
ctc2021	215852	87921
wikideits	4628049	324384
总计	6334399	757538

Table 3: 训练数据

人造伪数据

基于混淆集[8]，同音词集合[9]以及大量无标注语料，使用规则的方法生成。即选择无标注语料句子中的某个汉字或词，替换为对应的易混淆汉字或词，作为错误句子，原始句子作为正确句子。其中，无标注语料来自网上收集[10]以及拼写和语法纠错数据中正确的句子。此外，还利用拼写和语法纠错数据中的错误对原有的混淆集进行了扩充，最后混淆集对验证集的错误覆盖率为84.6%。最后的训练版本中，将验证集所有的错误对加入混淆集。易混淆字词示例如Table 4。我们以1300万句语料为基本数据，使用两种不同的策略，生成两份不同的数据。

混淆集名称	原字/词	混淆集内容
字级别混淆集-音似	放	防—访—犯—范—房—坊—仿—饭—芳—妨—方—返—纺
字级别混淆集-形似	假	佣—僻—暇—暇—儒—候—倾
同音词	放假	房价—放价—房家—放家—方佳—防夹

Table 4: 易混淆字词示例

策略一：以一定比例（0.25）随机选择句子中的汉字，若汉字在混淆集中则进行替换，不考虑词错误，具体替换规则如table 5.

80%	替换为音近汉字
15%	替换为形近汉字
5%	随机替换为词表中的某个汉字

Table 5: 构造数据策略一

策略二：每隔10个汉字，随机选择句子中的汉字或者词，进行替换，具体的替换规则如table 6.

3.3 模型的集成及后处理

模型集成：模型集成是提升比赛结果的重要方法，通过集成多个各方面表现差异较大的模型，可以大大提升系统的效果。对于深度模型，集成的方法也有多种，例如，投票法，平均checkponit，评分模型等。在本次评测中，我们将多个模型的词表保持一致，将全连接层输出的概率进行平均，相比与投票集成结果更好。

模型后处理：通过观察模型的输出，我们发现模型会出现如下问题，也分别设置了后处理策略：（1）模型对一些连续的错误，经常会漏修改。因此，将修改后的结果，再作为模型

[8]https://github.com/FDChongLi/TwoWaysToImproveCSC/blob/main/BERT/save/confusion.file
[9]https://github.com/LiangsLi/ChineseHomophones
[10]https://github.com/brightmart/nlp_chinese_corpus

43%（字）	84% 13% 3%	替换为音近汉字 替换为形近汉字 随机替换为词表中的某个汉字
57%（词）	43% 57%	替换为同音词 替换为统计到错误词对

Table 6: 构造数据策略二

的输入，多次前向，迭代修改。（2）为了增加模型的泛化能力，在预训练阶段，随机引入了部分非混淆汉字，这也造成了少量误修改。因此，设置规则，对非音近，形近的修改，进行还原。（3）模型对成语，实体等修改得不是很好。因此，增加了成语匹配纠错模块，首先，根据拼音相似，匹配到符合条件的成语，然后再利用语言模型选择最合适的成语或者保持不变。人名，地名，机构名等实体的收集的工作比较庞杂，这一块还没来得及处理。（4）模型对一些出现在比赛验证集的错误对，无法处理。因此，在预训练模型的基础上，使用验证集训练了一个模型。输入句子，首先经过集成模型进行纠错，纠错后的结果再输入到该模型中，进行再次纠错，也相当于先处理通用错误，再处理领域相关错误。

4 实验

4.1 训练设置

模型结构如3.1描述，代码使用pytorch编写，预训练模型部分使用transformers库构建，预训练模型分别使用了bert Cui et al. (2021)，roberta Cui et al. (2021)，macbert Cui et al. (2020)，具体训练流程如下：

两阶段训练：（1）基本数据为1300w，使用3.2所描述的两种策略，生成两份数据，分别进行预训练（2）使用hsk提取到的拼写纠错数据和SIGAHN进行微调

三阶段训练: (1）基本数据为1300w，使用3.2所描述的两种策略，生成两份数据，分别进行预训练（2）使用Table 3所示的所有拼写纠错数据进行微调（3）使用SIGHAN数据进行精调

因为验证集比较小，无法评估模型的效果，这里我们使用5-flod交叉验证法，将SIGHAN数据分成五份，每次用其中的四份做训练，一份做验证，选择平均结果最好的模型。

4.2 测试集结果

下面以预训练模型Roberta为例，给出不同训练设置下，测试集的结果。总体来说，策略二生成数据的方法相比策略一生成数据的方法略好。三阶段训练的结果要比两阶段训练更好。

(1) 两阶段训练-使用策略一生成的数据，在测试集上的结果见Table 7.

roberta	句子级检测	句子级纠正
预训练	52.4	47.69
微调	77.19	73.73

Table 7: roberta 两阶段训练-策略一

(2) 两阶段训练-使用策略二生成的数据，在测试集上的结果见Table 8.

roberta	句子级检测	句子级纠正
预训练	65.48	63.15
微调	76.51	74.26

Table 8: roberta 两阶段训练-策略二

(3) 三阶段训练-使用策略一生成的数据，在测试集上的结果见Table 9.

(4) 三阶段训练-使用策略二生成的数据，在测试集上的结果见Table 10.

roberta	检测	纠正
预训练	52.4	47.69
微调	75.79	72.25
精调	77.93	75.35

Table 9: roberta 三阶段训练-策略一

roberta	句子级检测	句子级纠正
预训练	65.48	63.15
微调	77.43	74.12
精调	79.2	76.0

Table 10: roberta 三阶段训练-策略二

最后选择使用单模型及其集成结果如Table 11所示，从单模型的效果来看，在拼写纠错任务上roberta >bert >macbert。这里我们集成了4个两阶段训练模型和3个三阶段训练模型，两阶段训练模型与三阶段训练模型相比，p值更高，r值更低，可以互相补充，最后的集成结果相比最好的单模型提升了2.74%。将集成模型的结果进行后处理操作，也有一定提升。其中，将集成后的结果再送入到验证集训练的模型进行纠正，提升最明显，约为1.08%.

roberta	句子级检测	句子级纠正
bert 两阶段训练-策略一	77.19	73.73
bert 两阶段训练-策略二	77.08	74.31
roberta 两阶段训练-策略二	76.51	74.26
macbert 两阶段训练-策略二	76.77	73.62
bert 三阶段训练-策略二	76.37	74.47
roberta 三阶段训练-策略二	79.2	76.0
macbert 三阶段训练-策略二	76.94	74.08
集成以上7个模型	80.57	78.74
+ 还原非音近形近修改	80.09	79.47
+ 迭代修改	81.44	79.79
+ 成语匹配纠错	82.2	80.0
+ 验证集训练模型	82.88	81.08

Table 11: roberta 集成模型及后处理结果

4.3 样例分析

Table 12给出了四种后处理策略能够辅助修正的样例。

目前，系统尚不能修改的样例如Table 13，其中实体错误无法修改的占比较大，这可能要通过大量收集专有名词字典，进行匹配纠错。

5 总结

在本次中文拼写检查任务中，主要难点在于汉语学习者拼写纠错数据不足，并且数据中存在很多实体错误，连续错误。我们的系统利用混淆集构造了大量的数据，缓解了数据不足的问题，并且通过概率集成了多个模型，而且还增加多个后处理模块。最终我们模型在评测数据集上，句子级纠错F1值达到了81.08.

从错误样例来看，目前还是会存在一些未修改或者修改错误的情况，主要原有有（1）混淆集对评测数据错误的覆盖度不够，（2）人工构造的数据还不够多，并且用于构造数据的来源并不是汉语学习者产生的，（3）实体错误修改比较差，需要专门的实体纠错模块。后面，我们也会尝试解决这些问题。

输入	我家附近有一家百**元**商店。
原始修改	我家附近有一家百**货**商店。
+ 还原非音近形近修改	我家附近有一家百**元**商店。
分析	“百货商店”更常见，但是“百元商店” 应该无错,“货”与“元”不是混淆字
输入	2008年是好**事怀**，变化太大。
原始修改	2008年是好**事坏**，变化太大。
+ 迭代修改	2008年是好**是坏**，变化太大。
错误原因	连续错误，很难一次修改完成，迭代修改先改“怀”，再改“事”
输入	他每次穿着**奇装衣服**，觉得很有意思。
原始修改	他每次穿着**奇装衣服**，觉得很有意思。
+ 成语匹配纠错	他每次穿着**奇装异服**，觉得很有意思。
分析	“奇装衣服” 模型没有改出，通过成语匹配纠错改出
输入	在端午节期间，我们一看就明白**那**个家有儿子， 因为有儿子的家在院子里会悬挂鲤鱼形状的旗。
原始修改	在端午节期间，我们一看就明白**那**个家有儿子， 因为有儿子的家在院子里会悬挂鲤鱼形状的旗。
+ 验证集训练模型	在端午节期间，我们一看就明白**哪**个家有儿子， 因为有儿子的家在院子里会悬挂鲤鱼形状的旗。
分析	“那-哪”错误字对在验证集出现过，经过验证集训练的模型可以改出

Table 12: 后处理可修正样例

参考文献

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 the Association for Computational Linguistics: EMNLP 2020*, pages 657–668.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Chong Li, Cenyuan Zhang, Xiaoqing Zheng, and Xuan-Jing Huang. 2021. Exploration and exploitation: Two ways to improve chinese spelling correction models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 441–446.

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*, pages 2517–2527.

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.

输入	家里中闻到有烤**白著**的香味儿。
预测	家里总闻到有烤**白猪**的香味儿。
标注	家里总闻到有烤**白薯**的香味儿。
错误原因	可能混淆集还不够全面
输入	我爸爸教我**个**我兄弟姐妹。
预测	我爸爸教我**跟**我兄弟姐妹。
标注	我爸爸教我**和**我兄弟姐妹。
错误原因	系统预测也是正确的
输入	**他**在证券公司工作。
预测	**她**在证券公司工作。
标注	**他**在证券公司工作。
错误原因	没有上下文，指代不明，系统过度修改
输入	还有，天**赋**罗卉，中华卉也有。
预测	还有，天**赋**罗卉，中华卉也有。
标注	还有，天**妇**罗卉，中华卉也有。
错误原因	实体错误，专有名词错误

Table 13: 系统不能修改样例