CCL2022-CLTC赛道二：一种基于Seq2Edit模型的pipeline语法纠错方法

李根
上海蜜度信息技术有限公司 / 上海
ligen1@miduchina.com

王本强
上海蜜度信息技术有限公司 / 上海
wangbenqiang@midu.com

摘要

汉语学习者文本语法纠错任务旨在自动检测并修改汉语学习者文本中的赘余、遗漏、误用、错序，从而获得符合原意的正确句子。针对此类任务，我们提出了一种基于Seq2Edit模型的pipeline语法纠错方法。首先，针对赘余、遗漏、误用、错序错误，我们分别构造了错字、多字少字、乱序的单个纠错模型。其次，我们利用gector模型，采用不同的预训练方法训练多个统一模型。最后，我们采用多种集成方式对各个模型进行了集成和融合。在测试集上效果来看，我们的COM指标达到46.27，位于第三名，这表明我们的方法是有效的。

关键词： Seq2Edit ； pipeline ； 融合

1 引言

汉语学习者文本语法纠错任务旨在自动检测并修改汉语学习者文本中的赘余、遗漏、误用、错序，从而获得符合原意的正确句子。近年来，该任务越来越受到关注，也出现了一些有潜在商业价值的应用。

针对中国计算语言学大会举办的此类纠错任务[1]，我们提出了一种基于Seq2Edit模型的pipeline语法纠错方法。首先，针对赘余、遗漏、误用、错序错误，我们分别构造了错字、多字少字、乱序的单个纠错模型。其次，我们利用gector模型，采用不同的预训练方法训练多个统一模型。最后，我们采用多种集成方式对各个模型进行了集成和融合。

我们的代码开源在 https://github.com/wang-benqiang/DeepCTC

2 背景

2.1 评测内容

如图1所示，此次任务涉及四种类型的错误，缺失(M)、赘余(R)、误用(S)、乱序(W)。针对S和M类错误，需要给出纠正结果。

错误类型	原始句子	正确句子
缺失(M)	忍不住**跑**父亲的旁边。	忍不住**跑到**父亲的旁边。
赘余(R)	**对对**她的亲切我很感谢。	对她的亲切我很感谢。
误用(S)	因为电视节**且**给她带来了放松的感觉。	因为电视节**目**给她带来了放松的感觉。
乱序(W)	如果梦到掉了一颗牙齿那就糟糕了， 那就意味着**一有位**亲人要去世!	如果梦到掉了一颗牙齿那就糟糕了， 那就意味着**有一位**亲人要去世!

Table 1: 错误类型举例

参赛系统输入句子，其中包含零到多个错误，参赛系统应判断该输入是否包含错误，并识别错误类型，标记出其在句子中的位置和范围，对缺失(S)和误用(M)给出修正答案。

举例：

输入句子：1001 我根本不能了解这妇女辞职回家的现象。在这个时代，为什么放弃自己的工作，就回家当家庭主妇

输出结果：

1001, 6, 7, S，理解

1001, 8, 8, R

2.2 评测标准

从下述六个方面以精确率、召回率和 F1 值对系统性能进行评价：

- 假阳性（False Positive）：正确句子被判包含错误的比例。
- 侦测层（Detective-level）：对句子是否包含错误做二分判断。
- 识别层（Identification-level）：给出错误点的错误类型。
- 定位层（Position-level）：对错误点的位置和覆盖范围进行判断，以字符偏移量计。
- 修正层（Correction-level）：提交针对字符串误用（S）和缺失（M）两种错误类型的修正词语。修正词语可以是一个词，也可以是一个词组。
- 综合打分（Comprehensive Score）：2022 年 CGED-8 引入 1-5 这五项指标的加权平均分数作为综合打分。

3 系统

我们采用pipeline的方式进行纠错。流程图如下：

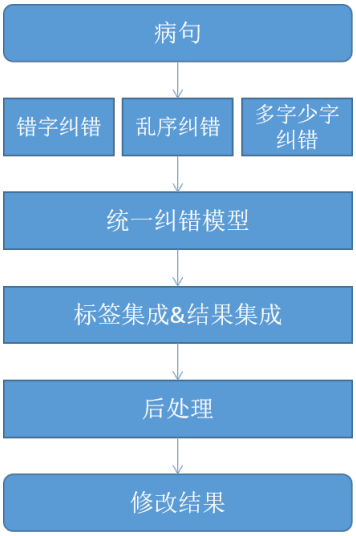

图 1: 模型工作流程图

3.1 单个模型

针对赘余、遗漏、误用、错序错误，我们分别构造了错字、乱序、多字少字的单个纠错模型。纠错顺序为乱序-错字-多字少字。具体模型如下：

3.1.1 错字

针对错字错误，我们采用realise文本模型进行纠错[2]，模型结构如下:

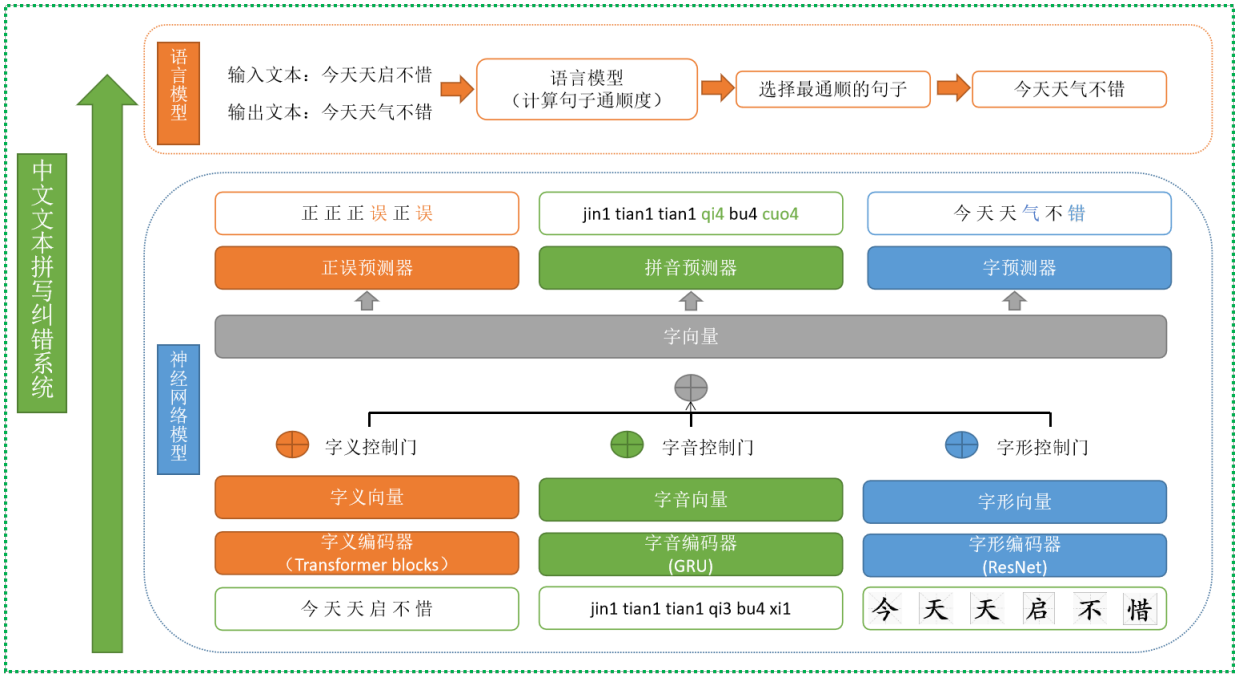

图 2: 错字词纠错模型图

3.1.2 乱序

针对乱序错误，我们采用BERT+CRF模型进行纠错,通过标注KEEP、LEFT、RIGHT标签来调整字词顺序。具体模型如下：

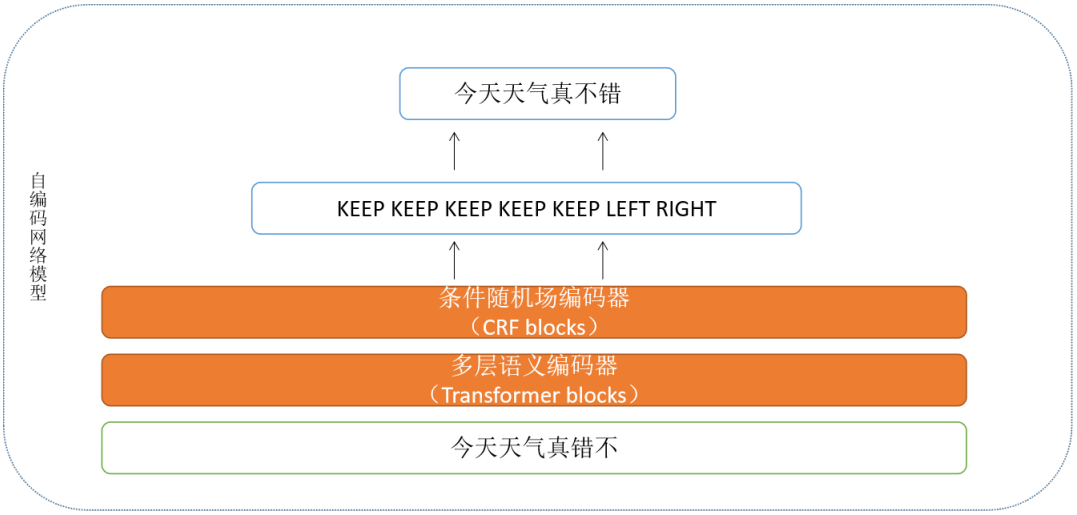

图 3: 乱序模型纠错模型图

- 仅标注KEEP、LEFT、RIGHT三个标签。
- 识别结束后，将最近连续的LEFT、RIGHT位置的字词进行替换纠正。

3.1.3 多字少字

针对多字少字错误，我们采用gector模型进行纠错[3]，具体模型如下：

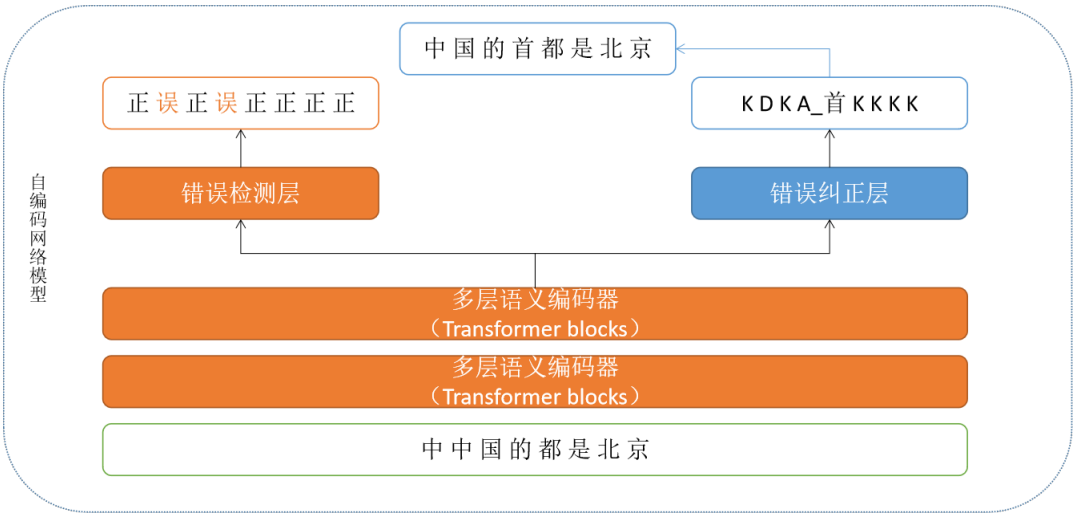

图 4: 多字少字纠错模型图

实验细节：

- 当错误检测层判断为“正”而错误纠正层判断为非“K”时，将错误检测结果的logit*0.4加到错误纠正层的“K”标签logit上；
- 当错误检测层判断为“误”而错误纠正层判断为“K”时，将错误纠正层的“K”标签上的logit减去错误检测结果的logit*0.4；

3.2 统一模型

由于单个模型的纠错结果具有片面性，因此，在单个模型纠错后，我们采用gector模型将各个

任务进行统一，融合了错字、乱序、多字少字各个任务，具体模型和多字少字模型相同。

3.3 集成

为了提高模型的准确性和泛化性，我们训练了多个模型进行融合，具体步骤如下：

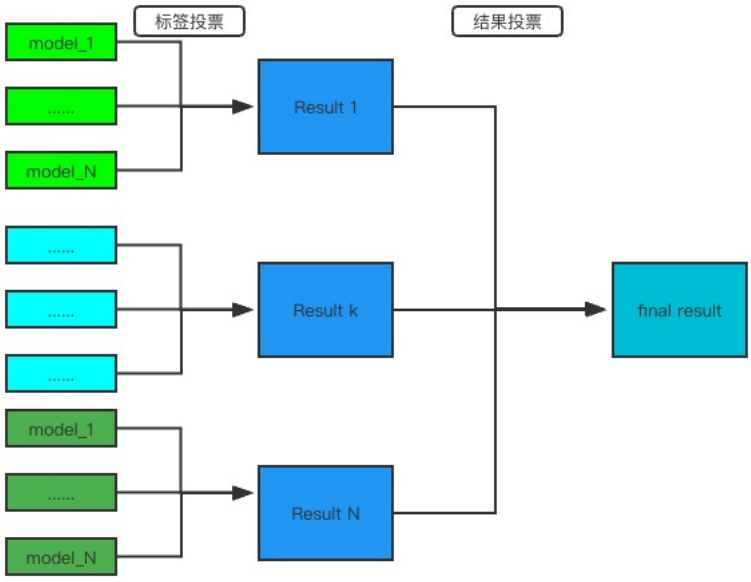

图 5: 模型集成流程图

融合分为两步：

- 标签层面的融合： 对每个iter预测的NER标签结果进行标签层面的投票， 选取标签最多的label作为本次iter的结果，并喂入下一个iter。
- 结果层面的融合：对各个模型的结果edits进行投票。

4 实验

4.1 数据

文本纠错是一个需要大量训练语料的复杂任务。因此，我们选用了多个开源数据，并构造了伪数据供训练。

模型用到的开源数据如下：

- CGED 历届数据
  地址：https://github.com/blcuicall/cged_datasets
- 中文 Lang8 数据集
  地址： http://yunpan.blcu.edu.cn:80/link/EDBB933F1FCD49C054F9AB7F65B0A746 密码：eSPB
- 中文拼写检查数据（SIGHAN+Wang271K）
  地址： http://yunpan.blcu.edu.cn:80/link/EF0963CBC2A979A71971BFECE8A34234 密码：lO42

- 文本智能校对大赛数据集
  地址：https://aistudio.baidu.com/aistudio/datasetdetail/157348

  伪数据构造方式如下：

- 针对字词的音型错误，利用混淆集构造伪数据，50%从音近混淆集中替换原数据，50%从型近混淆集中替换原数据。

- 针对乱序错误，随机交换字词位置构造伪数据。

- 针对多字少字错误，随机添加或删除字词，50%随机从大词林中添加近义词到当前字词的前面或者后面，50%删除字词。

4.2 其它trick

在此罗列一些模型的其它trick：

- 增加判断句子正误的二分类模型，用于数据后处理，缓解误报

- label smooth:缓解错误标签现象

- EMA：参数层面的集成

- 差分学习率：BERT层学习率2e-5；其他层学习率2e-3

- 参数初始化：模型其他模块与BERT采用相同的初始化方式

- 混合精度训练：提高训练速度

- 预训练模型采用macbert、macbertcsc、macbert-large

- 将繁体中文预先转为了简体

- 对非中文字符、”它、她、他”字符不做处理

4.3 实验结果

model	COM	COR	DET	FPR	IDE	POS
统一模型	39.77	23.04	83.66	39.23	57.20	34.40
单模型+统一模型	43.50	22.80	80.02	19.03	57.10	33.09
单模型+统一模型+标签层面集成	45.01	25.75	79.01	16.82	56.01	36.08
单模型+统一模型+标签和结果层面集成	46.27	27.60	83.62	22.42	58.63	37.66

Table 2: 实验结果

错例分析：

原始句子	修正结果
忍不住跑父亲的旁边。	忍不住跑到父亲的旁边。
对对她的亲切我很感谢。	对她的亲切我很感谢。
因为电视节且给她带来了放松的感觉。	因为电视节目给她带来了放松的感觉。
最近有会餐特别多。	最近回场特别多。
第四段：遇上了当地人，在她的热情帮助下，我坐吉普车。	第四段：遇上了当地人，在她的热情帮助下，我坐上了吉普车。
我不喜欢的动物是蛇，我很恐怕。	我不喜欢的动物是蛇，我觉得很恐怖。

Table 3: 错例分析

效果总结：

- 对于错字词，模型基本能够识别正确。
- 对于缺失或误用错误，当任务较难时，模型的识别结果效果一般；当缺失或误用的内容在原句中并不关键时，模型能够较好的进行纠正。

5 结论

本文提出了一种基于Seq2Edit模型的pipeline语法纠错方法。首先，针对赘余、遗漏、误用、错序错误，我们分别构造了错字、多字少字、乱序的单个纠错模型。其次，我们利用gector模型，采用不同的预训练方法训练多个统一模型。最后，我们采用多种集成方式对各个模型进行了集成和融合。在测试集上结果来看，我们的COM指标达到46.27，位于第三名，这表明我们的方法是有效的。

中文语法纠错一直是纠错中较为复杂的任务，且真实的纠错场景比比赛场景更为复杂。后续，我们会从效果、效率两方面继续优化纠错模型。

参考文献

[1] Wang, Yingying and Kong, Cunliang and Liu, Xin, et al. Overview of CLTC 2022 Shared Task : Chinese Learner Text Correction[C]. 2022.

[2] Xu H D, Li Z, Zhou Q, et al. Read, Listen, and See: Leveraging Multimodal Information Helps Chinese Spell Checking[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.

[3] Omelianchuk K, Atrasevych V, Chernodub A, et al. Grammatical Error Correction: Tag, Not Rewrite[J]. 2020.