Haitai
海泰方圆

—数据安全·密码赋能—

# 基于PPL集成的中文语法纠错方法

Patrick

北京海泰方圆科技股份有限公司

二〇二二年十月

Haitai
海泰方圆

# 目录
# CONTENTS

# 评测任务背景-CLTC

## 1. 比赛介绍

汉语学习者文本纠错任务（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。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。为了推动这项研究的发展，研究者通过专家标注以及众包等形式构建一定规模的训练和测试数据，在语法检查以及语法纠错等不同任务上开展技术评测。同时，由于汉语学习者文本纠错任务相对复杂、各评测任务以及各数据集之间存在差异，在一定程度上限制了文本纠错的发展。因此，我们希望通过汇聚、开发数据集，建立基于多参考答案的评价标准，完善文本纠错数据及任务，聚焦该研究领域中的前沿问题，进一步推动汉语学习者文本纠错研究的发展。

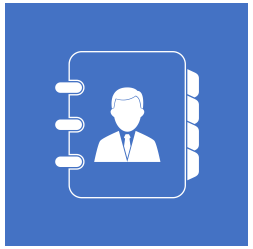
我们依托第二十一届中国计算语言学大会（CCL 2022），组织汉语学习者文本纠错评测。本次评测既整合了已有的相关评测数据和任务，又有新开发的数据集，以设置多赛道、统一入口的方式开展比赛任务。同时，我们研制了各赛道具有可比性的评测指标，立足于构建汉语学习者文本纠错任务的基准评测框架。

**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）。不同来源的文本，其蕴含的语法错误类型也可能含有一定的差异。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，对于每一个句子提供多个遵循流利提升的修改答案，希望能够准确而全面地评估各参赛队伍的纠错系统性能。

本次任务是汉语学习者文本纠错任务（CLTC）中的一项，即多参考、多来源汉语学习者文本纠错，包括了三种不同来源的文本，目标是纠错准确性和流利提升。

# 数据集

| 数据集 | 句子数 | 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(95.4%) | 29.7 | 2.5 | 2.5 |
| MuCGEC-CGED | 3125 | 2988(95.6%) | 44.8 | 4.0 | 2.3 |
| MuCGEC-Lang8 | 1942 | 1652(85.1%) | 37.5 | 2.8 | 2.1 |
| MuCGEC-ALL | 7063 | 6544(92.7%) | 38.5 | 3.2 | 2.3 |



训练集：nlpcc2018+hsk
来源：https://github.com/HillZhang1999/MuCGEC

开发集：MuCGEC（Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction），1137句（其中包含12句无法标注的句子）。
来源：主办方提供

测试集：MuCGEC（Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction），6000句（其中包含62句无法标注的句子）。
第一阶段：2000句（部分），第二阶段：6000句（完整）
来源：主办方提供

评测数据集由主办方提供，数据集来源多样，标注质量高，参考答案多样，平均每句标注2.3个基于流利度提升的不同参考答案。

# 方法研究

## 人工规则

基于规则的方法：虽然手工定义规则费时费力，实际应用会存在冲突，但在具体行业应用中，尤其是在医药、政务、金融、农业等**语句行业特征**比较明显的行业，手工定义的规则可作为辅助手段，还是有帮助的。

## 序列标注Seq2Edit

序列标注模型（Seq2Edit）：将纠错抽象为**有限空间的编辑动作**。通过预先定义一些编辑动作，包括复制、删除、增加、替换、变形等，理论上降低了解码空间。在增加/替换操作上的解码空间也非常大，实际上包含编辑操作空间也不小，每个编辑动作的训练数据并不充分。

## 序列到序列Seq2Seq

包括基于RNN、CNN、注意力的编解码方法。现在主流的是基于**自注意力机制**，在特征抽取和并行计算上有明显的优势。在实际应用中，也有一些很有意思的改进，比如针对纠错任务，大部分文本都是正确的这种特征，有**拷贝增强**机制。针对语言多样性，有多种**解码机制**。缺点是解码空间大，收敛性还需要继续研究。

根据任务、应用要求、数据分析来选择不同的方法，不拘泥于某一种特定的方法。

参考：

https://hillzhang1999.gitee.io/2021/03/31/yu-fa-jiu-cuo-jin-zhan-zong-shu/

http://blog.nghuyong.top/2021/05/26/NLP/text-corrector/

https://www.cnblogs.com/qftie/p/16807855.html

# 领域进展

**北京语言大学**总结的语法错误和纠正进展：错误类型主要包括多字，少字，选择错误，词序错误。子任务：

① 检测：判断是否有语法错误。

② 识别：判断具体错误类别。

③ 位置级：识别出start_pos,end_pos错误位置。

④ 改正：对选词错误和缺失词进行改正。

| Error Type | |
|---|---|
| #R | 769 (21.05%) |
| #M | 864 (23.65%) |
| #S | 1694 (46.36%) |
| #W | 327 (8.95%) |
| #Error | 3,654 (100%) |

Table 4: The distributions of error types in testing set.

参考：Overview of NLPTEA-2020 Shared Task for Chinese Grammatical Error Diagnos
Beijing Language and Culture University

| 模型 | NLPCC18-Official(m2socrer) | MuCGEC(ChERRANT) |
|---|---|---|
| seq2seq_lang8[Link] | 37.78/29.91/35.89 | 40.44/26.71/36.67 |
| seq2seq_lang8+hsk[Link] | 41.50/32.87/39.43 | 44.02/28.51/39.70 |
| seq2edit_lang8[Link] | 37.43/26.29/34.50 | 38.08/22.90/33.62 |
| seq2edit_lang8+hsk[Link] | 43.12/30.18/39.72 | 44.65/27.32/39.62 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 3 | 0.178 | 0.1536 | 0.1649 | 0.0934 | 0.2283 | 0.1325 |
| PCJG | 1 | 0.0492 | 0.0233 | 0.0307 | 0.0492 | 0.0223 | 0.0307 |
| TextCC-CloudPoineer | 1 | 0.1737 | 0.1247 | 0.1452 | 0.0983 | 0.1454 | 0.1173 |
| | 2 | 0.1696 | 0.1341 | 0.1498 | 0.0973 | 0.156 | 0.1198 |
| TMU-NLP | 1 | 0.2258 | 0.1032 | 0.1417 | 0.2258 | 0.1032 | 0.1417 |
| UNIPUS-Flaubert | 1 | 0.2848 | 0.1415 | **0.1891** | 0.2276 | 0.1595 | 0.1876 |
| | 2 | 0.2587 | 0.1372 | 0.1793 | 0.1582 | 0.1646 | 0.1613 |
| | 3 | 0.2014 | 0.1603 | 0.1785 | 0.1339 | 0.188 | 0.1564 |
| XHJZ | 1 | 0.1293 | 0.1763 | 0.1492 | 0.1293 | 0.1763 | 0.1492 |
| | 2 | 0.1465 | 0.1646 | 0.1550 | 0.1465 | 0.1646 | 0.1550 |
| | 3 | 0.1764 | 0.1646 | 0.1703 | 0.1764 | 0.1646 | 0.1703 |
| YD_NLP | 1 | 0.3238 | 0.1290 | **0.1845** | 0.2982 | 0.1372 | **0.1879** |
| | 2 | 0.3293 | 0.1263 | **0.1826** | 0.3132 | 0.1337 | **0.1874** |
| | 3 | 0.3386 | 0.1259 | **0.1836** | 0.3217 | 0.1333 | **0.1885** |
| ZZUNLP-HAN | 1 | 0.0027 | 0.0012 | 0.0017 | 0.0018 | 0.002 | 0.0019 |
| | 2 | 0.0009 | 0.0004 | 0.0006 | 0.0007 | 0.0008 | 0.0007 |

Table11. Results of CGED 2020 in Correction Level

好的")

# 本次评测验证方法

数据分析：

对训练集、开发集、测试集数据分析，包括数据特征检查，样本数量、长度、可能的错误类型等。

模型训练：

① 字表扩展，支持一些特殊的中文符号字符。

② 训练Seq2Edit模型，Seq2Seq模型。

③ 训练Plome模型。

模型集成：

① 备选模型

② 串行集成

③ 并行集成


数据处理：

① 按评测要求过滤不能参与训练的句子集。

② 过滤掉超长（单句大于128长度）。

③ 过滤掉重复出现的句子。

④ 中文字符处理，生成训练数据。

模型验证：

① 验证Seq2Edit模型。

② 验证Seq2Seq模型。

③ 验证Plome模型。

④ 验证其他模型：macBert,Ernie,T5。

⑤ 按准确率和召回率选择备选集成模型。

通过对序列到编辑、序列到序列以及Plome模型，macBert,Ernie,T5等模型验证，后面这几种模型效果一言难尽。这也说明，纠错也并非简单的通用任务，任务也是严重依赖对应的数据和数据特征。

# 评测效果

基于PPL的多模型并行集成方法，总体效果还不错，但对于特殊字符，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，会出现多个模型输出都很流畅，但不一定符合输入期望纠错的要求。

**模型集成：** 采用并行集成方法

| 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


**集成各模型贡献：** 命中率

| 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

**集成模型评分**

| 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

**错误分析**

| 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

# 行业应用落地探讨

数据、规则、单模型效果、计算性能、时延等是行业应用关注的重点。

## 行业落地应用可能的改进点

垂直行业领域的**数据特征**研究，如社交短文本，政务文本等。其次是**自动标注语料**错误研究。

**数据集**

01

单模型的准确率和召回率需要大幅提升，尤其是准确率。

**模型**

02

在单模型的效果大幅提升的基础上，集成效果会更明显，另外可以考虑人工规则辅助。

**模型集成**

03

多模型集成对**计算性能和响应时延**影响大，导致行业实际落地困难。

**性能优化**

04

# 让信息世界充满信任

## 谢谢聆听，敬请指导！