CCL2022-CLTC：汉语学习者文本纠错评测总结报告

北京语言大学 语言监测与智能学习研究组

2022年10月30日

ICALL 智能语言学习

目录

一、评测背景

随着科技的发展与进步，特别是人工智能技术的创新，智能计算机辅助语言学习(Intelligent Computer-Assisted Language Learning, ICALL) 在国际中文教育中的作用越来越突出。其中，汉语学习者文本纠错就是一项重要的应用。

汉语学习者文本纠错(Chinese Learner Text Correction, CLTC) 旨在通过智能纠错系统，自动检测并修改汉语学习者文本(Chinses Learner Text) 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。

一、评测背景

達摩院
ALIBABA DAMO ACADEMY

二、评测任务与赛道简介

拼写检查	语法纠错	质量评估
赛道一	赛道二 赛道三 赛道四	赛道五

二、评测任务与赛道简介

赛道一

任务：中文拼写检查

训练集：本赛道允许使用任意开源数据用于训练。

开发与测试集：基于YACLC-CSC 数据集的开发集与测试集。

赛道二

任务：中文语法错误检测

训练集：

1.中文Lang8 数据集

2.CGED历年数据

测试集： CGED-8 数据集。

数据来源为 HSK动态作文语料库和全球汉语中介语语料库。

赛道三

任务：多维度汉语学习者文本纠错

训练集：处理后的Lang8中介语数据。

开发与测试集：

YACLC-Minimal

YACLC-Fluency

赛道四

任务：多参考多来源汉语学习者文本纠错

训练集：不提供官方训练数据集。

开发与测试集：基于流利提升的多参考数据集MuCGEC。

赛道五

任务：语法纠错质量评估

训练集：

1.中文Lang8 数据集

2.本赛道提供的带有语法纠错候选方案的数据

开发与测试集：

YACLC-Minimal

YACLC-Fluency

三、评测特色

- 首先，就关注者最多的语法纠错任务，将现有资源整合汇聚于三个赛道，从不同的侧重点分别进行评测。

- 其次，为进一步推进中文拼写检查研究，本次评测基于YACLC 数据集 (Wang et al., 2021) 构建并公开了YACLC-CSC 数据集，为首个简体中文拼写检查数据集。

- 最后，扩展了文本纠错任务，首次将质量评估(Quality Estimation, QE) 纳入评测任务。该任务可用于模型集成或其他情况下对多条纠错结果重排序(Re-Ranking)，可以在不改变模型的情况下明显提升修改效果。

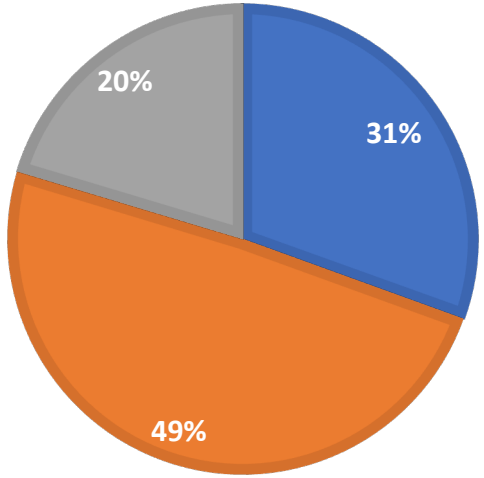
四、参赛情况

- 赛道一、二、三、四、五共142支队伍报名。
- 包括清华、北大、中科院、北邮、苏大等高校和科研院所，以及达观、蜜度、好未来、视源等企业。

四、参赛情况

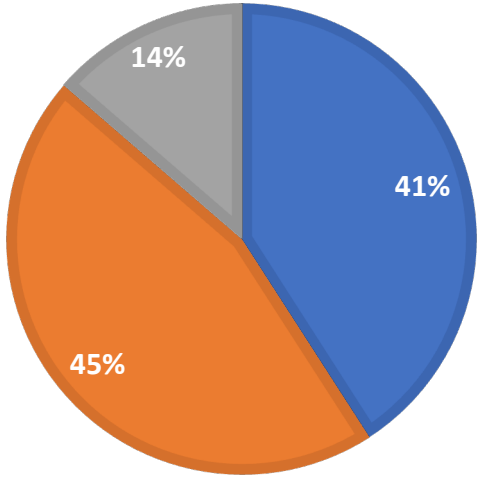
赛道一

企业：18 学校：29 其他：12 总数：59



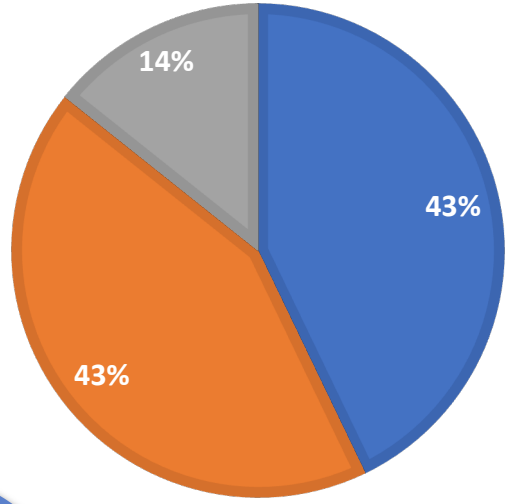
赛道二

企业：9 学校：10 其他：3 总数：22

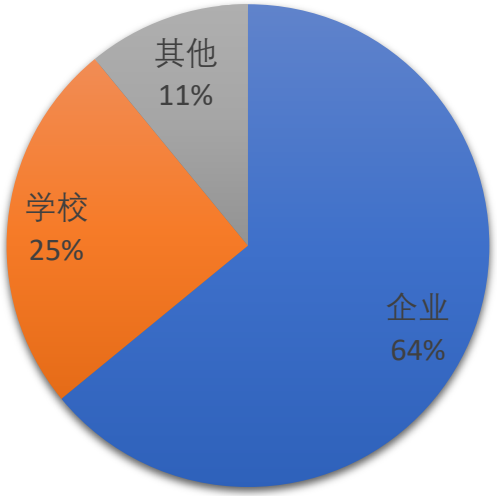


赛道三

企业：3 学校：3 其他：1 总数：7

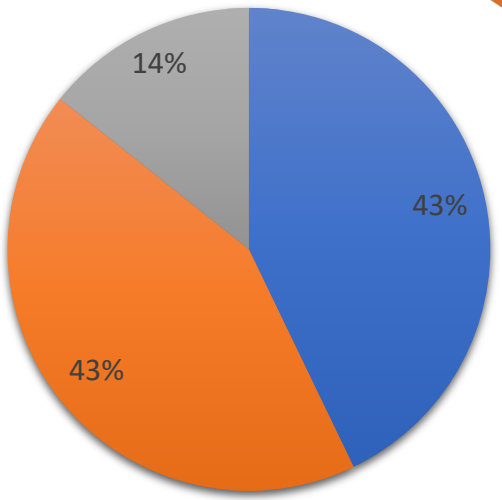


赛道四



企业 学校 其他 总数

赛道五



企业：3 学校：3 其他：1 总数：7

五、评测结果

	赛道一	赛道二	赛道三	赛道四	赛道五
一等奖	哒哒（达观数据）	NLP的未来（好未来）	kk（北京大学）	啊对对对（清华大学） 鱼饼啾啾（北京大学）	CPIC（中国太平洋保险）
二等奖	iFunCun（方寸无忧）	一一（达观数据）	改正带小助手（苏州大学）	棒棒冰（视源电子科技）	
三等奖	csc_runner（视源电子科技）	中国足球队（蜜度）	BUPTCL（北京邮电大学）	后厂村9号（海泰方圆）	

六、选手方案

	赛道一
模型方案	•纠错模型： •基于Transformer的序列到序列类纠错模型（前三名、第六名）； •模型集成： •基于困惑度的异构模型串行/并行集成（第一名，第二名）； •基于平均概率的异构模型并行集成（第三名、第六名）；
数据方案	•融合拼音编码（第一名） •数据增强：基于混淆集+规则的数据增强（前三名，第六名）；
后处理方案	•Ngram纠错（第一名，第二名，第六名） •实体纠错（第一名，第三名，第六名）

六、选手方案

	赛道二
模型方案	· 纠错模型： • 基于Transformer/BERT的序列到序列类纠错模型（第一名、第二名、第四名） • 基于GECToR的序列到编辑类纠错模型（第一名、第三名） · 模型集成： • 基于困惑度的异构模型串行/并行集成 • 基于编辑级别投票的异构模型并行集成
数据方案	数据增强：基于混淆集+规则的数据增强（前三名）；

六、选手方案

	赛道三
模型方案	•纠错模型： •基于GECToR的序列到编辑类纠错模型（第二、三名）； •基于Transformer/BART的序列到序列类纠错模型（第一、二名）； •模型集成： •有序的异构模型串行/并行集成（第二名）； •基于编辑级别投票的模型并行集成（第一、三名）； •基于模型输出概率平均投票的模型并行集成（第三名） •模型后处理： •忽略UNK（第一名）
数据方案	•数据增强：动态噪声（第一名）； •数据选择：开发集使用单一参考答案（第二名）；

六、选手方案

	赛道四
模型方案	•语法纠错模型 •基于GECToR的序列到编辑类纠错模型（前四名）； •基于Transformer/BART的序列到序列类纠错模型（前三名）； •PLMOE等基于多模态信息的拼写纠错模型（后厂村9号队）； •模型集成： •基于困惑度的异构模型串行/并行集成（后厂村9号队）； •基于编辑级别投票的异构模型并行集成（前三名）；
数据方案	•数据增强：基于混淆集+规则的数据增强（前三名）； •数据清洗：将多目标训练集转化为单目标训练集（啊对对对队）；
其余Tricks	•预训练模型； •推理阶段额外置信度（棒棒冰）； •切句预测（鱼饼啾啾）； •过滤非纠错修改（鱼饼啾啾）；

六、选手方案

	赛道五
模型方案	•语法纠错模型：Electra模型 •对抗训练：使用了EMA、FGM及二者组合的三种策略 •模型集成： •对所有模型的改正句分数求平均值，选取最高得分的句子作为修改质量最好的句子 •每个模型选出最优修改句，通过投票最终获得原句对应的最优修改句
数据方案	•数据预处理： •将数据集中的句子分数修改为0、1 •将少于10个改正句的原句，利用分数最低的改正句扩展到10句 •数据后处理：针对某一类错误，对模型输出结果做进一步的调整优化

总结和展望

- 本次评测整合了已有的文本纠错的相关评测数据和任务，发布了新的数据集，构建了汉语学习者文本纠错任务的基准评测框架，以设置多赛道、统一入口的方式开展比赛任务。

- 来自学界和企业界的多个队伍竞相参与，相较于基线模型，参赛系统的性能有大幅提升，展现出了汉语学习者文本纠错任务上的现有水平。

- 参赛队伍在模型和数据方面的工作相较此前无显著创新，但针对不同任务所采用的的模型集成、数据处理的策略和实践值得参考。

总结和展望

- 针对不同场景、需求和任务的语法纠错：
  - 如何设计更适合的模型架构?
  - 如何更好地表征多参考答案的训练数据?
  - 如何通过质量评估进行纠错后处理?
  - 如何降低自然标注数据中的噪声?
  - 如何大量生成更为逼真的伪数据?

长期评测榜单

YACLC

汉语学习者文本纠错评测

报名参赛　赛道一：中文拼写检查

http://cuge.baai.ac.cn/#/ccl_yaclc

赛事介绍　赛道一：中文拼写检查

刷新排行

排行榜

验证阶段排行榜

	Sentence Level		Character Level	
	C-F	D-F	C-F	D-F
哒哒	84.61	85.77	98.33	88.58
evaluation2	83.12	84.49	97.74	87.72
csc_runner	81.12	82.93	96.77	85.1
evaluation	81.12	82.93	96.77	85.1
黑boy	75.35	77.93	94.36	81.08
今天也有好好调试	51.51	55.89	87.76	61.75

谢谢!

关注我们

语言监测与智能学习公众号

语言监测与智能学习Github主页

语言监测与智能学习小组主页