2021年04月10日进度汇报

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 今日进度

1. 使用中文测试集微调生成模型
2. 成语古诗文检错
3. 中文测试集处理

2 工作详述

2.1 微调中文生成模型

- 数据：全部测试集数据

2.2 中文测试集处理

- 乱写作文：
  - 使用乱写检测模型检测
  - 共抛出561篇乱写作文，全部处理为完全离题作文(score=1)
- 补充数据处理（目前的处理思路）
  - 分别计算当前题目下离题作文与补充作文的相似度
  - 去除相似度≥ 0.95的作文，剩余作文补充为切题
- 处理后测试集规模：35个题目，共21300篇作文，离题:切题=1971:19329(1:9.8)
- 各档作文分布，见表1：

	完全离题（score=1）	部分离题（score=2）	部分离题（score=3）	切题（score=4）
作文数	1015	470	486	19329

Table 1: 处理后测试集作文分布