2021年03月23日进度汇报

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. [***] 构建新的中文离题数据集，更新离题指标
2. [***] 验证生成模型效果(kmeans)
3. [***] 谚语/名人名言数据爬取
4. [***] 成语古诗文检错任务对接

2 工作详述

2.1 构建中文离题数据集（$Dataset_2$）

- 共25个题目，每个题目下50篇作文，离题:切题=1：1，数据构成见表1：
- 生成模型离题指标，见表2：
- 结论：
  - 在简单数据集上可以偏向于判定为离题作文

切题作文	25篇一二类卷作文
离题作文	25篇乐乐课堂作文

Table 1: $Dataset_2$单个题目下作文构成

		离题			不离题		
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
方案一	开发集	0.6140	0.8000	0.6587	0.4340	0.3600	0.3637
	测试集	0.4732	0.6782	0.5447	0.3048	0.2345	0.2524
方案二	开发集	0.7000	0.8800	0.7406	0.4200	0.4533	0.4297
	测试集	0.6945	0.7418	0.6514	0.3399	0.4836	0.3991

Table 2: $Dataset_2$ - Bert生成模型离题指标更新

2.2 生成模型效果对比

- 使用kmeans方法，对所以切题作文ht进行聚类
- 指标见表3：
- 结论：
  - 增加训练集后离题效果提升

	purity_score
bertabs_6w	0.2043
bertabs_80w	0.2743

Table 3: Bert生成模型取[CLS]表示效果对比