# 周报,2020年1月2日

屈原斌 首都师范大学 ybqu@cnu.edu.cn

### 1 上周计划

- 1. [\*\*\*] 完成数据集的构建。
- 2. [\*\*\*] 跟进数据标注。
- 3. [\*\*\*] 抽取作文进行分析。
- 4. [\*\*] 使用BERT在随机构建数据集进行测试。
- 5. [\*] 配置数据标注环境doccano。
- 2 上周计划执行情况
  - 1. [√] 数据集构建完成。
  - 2. [√×] 数据标注这部分目前标注了1/3,标注过程中的问题会实时进行反馈。
  - 3. [√] 完成抽取数据的分析,但统计还不够完善。
  - 4. [√×] 代码有一点问题,需要修改一下。
  - 5. [√×] 配置一直有问题,添加用户不能登录/无法创建项目等,需要重新配置一下。
- 3 本周部分重点工作详述
- 3.1 数据集构建
  - DataSet<sub>1</sub>: 人工标注35个题目;
  - $DataSet_2$ :
    - 数据集大小: 30个题目×50篇作文;
    - 方法:
      - \* 正样本:
        - · 从当前题目下一类卷+二类卷抽取20篇作文;
        - · 随机生成(从当前题目下抽取作文随机打乱段落/句子顺序);
      - \* 负样本:
        - · 从其他题目下一类卷+二类卷抽取20篇作文作为负样本;
        - · 随机生成 (对一段或者一句话重复生成/打乱段落活句子顺序);
        - ·添加一些著名作者的优美散文;

#### 3.2 数据分析

- 数据: 2个题目, 每个题目下每类卷抽取20篇, 每个题目共100篇;
- 结论:
  - 作文分类可参考表1;
  - 离题作文主要分布在四、五类卷中;
  - 五类卷中存在乱写和抄袭(背诵)的作文较多;
  - 半命题作文中作文题目存在的问题较多:

Table 1. 数据分析结果

Table 1. Why will have												
		离题										
	不离题	部分离题					完全离题					
		正常	流水账作文	抄袭(背诵)	转写错误	正常	流水账作文	抄袭(背诵)	乱写	转写错误		
记住这一天	62	5	7	2	0	12	3	7	2	0		
你的样子,真美丽	57	16	3	5	0	4	5	8	2	0		

# 4 存在问题

- 问题1 代码问题,需要修改一下;
- 问题2 数据标注环境doccano需要重新配置;

# 5 下周计划

- 1. [\*\*\*] 完善数据分析部分。
- 2. [\*\*\*] 跟进数据标注工作。
- 3. [\*\*\*] 修改代码,完成第二个数据集上的实验。
- 4. [\*\*] 阅读论文。

## A 附录

## A.1 本周阅读论文列表

论文名称	会议/期刊	状态
RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering	Arxiv 2020	略读
Constrained multi-task learning for automated essay scoring.	ACL 2016	未看完

Table 2: 本周论文阅读列表