

# 周报，2020年11月23日

屈原斌  
首都师范大学  
ybqu@cnu.edu.cn

## 1 上周计划

1. [\*\*\*] Baseline实验。
2. [\*\*\*] 数据集处理。

## 2 上周计划执行情况

1. [✓×] 尝试一种方案，其他方案未实现。
2. [✓×] 根据重标数据更新 $DataSet_1$ ，范文和一二类卷中转写问题未处理。

## 3 本周部分重点工作详述

### 3.1 $DataSet_1$ 更新

- 根据重标数据更新 $DataSet_1$
- 数据集：
  - 标注数据：离题作文341篇，不离题作文308篇；
  - 补充数据：添加一类卷+二类卷，每个题目下作文补充至50篇；
  - 数据集大小：30个题目×50篇作文；
  - 数据分布见表1：

	作文数
离题作文	341
不离题作文	1409

Table 1:  $DataSet_1$ 数据分布

### 3.2 Baseline实验

- 实验方案：
  - 初始方案：和题目下范文表示分别做相似度计算
  - 更新方案：对范文表示取平均作为题目的表示后进行相似度表示
- 实验设置：
  - threshold=20
- 模型说明：
  - HBiLstm\*：字向量模型
  - HBiLstm：词向量模型

Threshold=20, strategy='aver'					Threshold=20, strategy='max'					Threshold=20, strategy='topk', topk=3							
		accuracy	precision	recall	f1			accuracy	precision	recall	f1			accuracy	precision	recall	f1
HBiLstm*	离题作文	-	0.43	0.55	0.48	HBiLstm*	离题作文	-	0.46	0.51	0.48	HBiLstm*	离题作文	-	0.48	0.5	0.49
	不离题作文	-	0.88	0.82	0.85		不离题作文	-	0.88	0.86	0.87		不离题作文	-	0.88	0.87	0.87
	macro avg	0.77	0.66	0.69	0.67		macro avg	0.79	0.67	0.68	0.68		macro avg	0.8	0.68	0.68	0.68
HBiLstm	离题作文	-	0.39	0.64	0.49	HBiLstm	离题作文	-	0.44	0.55	0.49	HBiLstm	离题作文	-	0.45	0.57	0.51
	不离题作文	-	0.9	0.76	0.82		不离题作文	-	0.88	0.83	0.86		不离题作文	-	0.89	0.83	0.86
	macro avg	0.74	0.64	0.7	0.65		macro avg	0.77	0.66	0.69	0.67		macro avg	0.78	0.67	0.7	0.68
Bert_CLS	离题作文	-	0.54	0.49	0.51	Bert_CLS	离题作文	-	0.72	0.39	0.5	Bert_CLS	离题作文	-	0.71	0.4	0.51
	不离题作文	-	0.88	0.9	0.9		不离题作文	-	0.87	0.96	0.91		不离题作文	-	0.87	0.96	0.91
	macro avg	0.82	0.71	0.69	0.7		macro avg	0.85	0.79	0.68	0.71		macro avg	0.85	0.79	0.68	0.71

Figure 1: 初始方案-*DataSet<sub>1</sub>*实验结果

DataSet1, Threshold=20					
		accuracy	precision	recall	f1
HBiLstm*	离题作文	-	0.51	0.44	0.47
	不离题作文	-	0.87	0.9	0.88
	macro avg	0.81	0.69	0.67	0.68
HBiLstm	离题作文	-	0.5	0.54	0.52
	不离题作文	-	0.89	0.87	0.88
	macro avg	0.8	0.69	0.7	0.7
Bert_CLS	离题作文	-	0.5	0.67	0.57
	不离题作文	-	0.91	0.84	0.87
	macro avg	0.8	0.71	0.75	0.72

Figure 2: 更新方案-*DataSet<sub>1</sub>*实验结果

- Bert\_CLS：微调后模型，取模型输出第一维作为表示

#### • 实验结果：

- 见图表1、2、3
- 增大阈值结果变好；
- 词向量模型优于字向量模型；
- bert模型的结果最优；

## 4 下周计划

1. [\*\*\*] 清洗数据集。
2. [\*\*\*] 完成其他方案实验。

DataSet2, Threshold=20					
		accuracy	precision	recall	f1
HBiLstm*	离题作文	-	0.91	0.86	0.89
	不离题作文	-	0.87	0.91	0.89
	macro avg	0.89	0.89	0.89	0.89
HBiLstm	离题作文	-	0.88	0.88	0.88
	不离题作文	-	0.88	0.87	0.88
	macro avg	0.88	0.88	0.88	0.88
Bert_CLS	离题作文	-	0.78	0.92	0.84
	不离题作文	-	0.9	0.74	0.81
	macro avg	0.83	0.84	0.83	0.83

Figure 3: 更新方案-*DataSet<sub>2</sub>*实验结果