

周报，2020年12月14日

屈原斌  
首都师范大学  
ybqu@cnu.edu.cn

## 1 上周计划

1. [\*\*] 数据处理。
2. [\*\*\*] 离题检测实验。
3. [\*\*\*] 开题报告和ppt。

## 2 上周计划执行情况

1. [✓] 根据比例构建 $DataSet_1$ 。
2. [✓×] 在新的数据集上进行测试，更新方案，完成部分实验。
3. [✓×] 未修改报告。

## 3 本周部分重点工作详述

### 3.1 数据处理

- 重新构建 $DataSet_1$ 
  - 按照四五类卷抽取比例添加一二三类卷，共分成两个数据集，第一个数据集只包含标注双评一致数据score=1和score=4的作文，第二个数据集包含标注双评一致全部作文（将score=1、2、3看作离题作文）
  - 两个数据集数据分布：

	离题	不离题作文	总计
作文数	341	17330	17641

Table 1:  $DataSet_{11}$ 数据集分布

	离题			不离题作文 (score=4)	总计
	完全离题 (score=1)	部分离题 (score=2)	部分离题 (score=3)		
作文数	341	205	238	28237	29021

Table 2:  $DataSet_{12}$ 数据集分布

### 3.2 离题检测实验

- 实验方案
  - 方案一：取topk, k=3
  - 方案二：取范文表示平均作为题目表示
  - 方案三：取全部一二类卷表示平均作为题目表示

		方案一, strategy='topk', k=3				方案二				方案三			
		最优Threshold	P	R	F1	最优Threshold	P	R	F1	最优Threshold	P	R	F1
HABiLstm_C	离题 不离题	66	0.2121 0.9847	0.2258 0.9835	0.2188 0.9841	53	0.1699 0.9852	0.2551 0.9754	0.204 0.9803	70	0.2528 0.9843	0.1994 0.9884	0.223 0.9863
HABiLstm_W	离题 不离题	62	0.25 0.9861	0.2962 0.9825	0.2711 0.9843	73	0.2884 0.9856	0.2698 0.9869	0.2788 0.9863	45	0.3041 0.9863	0.305 0.9862	0.3045 0.9863
Bert_CLS(微调全部层)	离题 不离题	48	0.1971 0.9856	0.2757 0.9779	0.2298 0.9817	36	0.162 0.9863	0.3196 0.9674	0.215 0.9768	35	0.2029 0.9851	0.2463 0.9809	0.2225 0.983
Bert_CLS(微调9-11层)	离题 不离题	64	0.2755 0.9865	0.3167 0.9836	0.2947 0.985	64	0.2752 0.9867	0.3284 0.9829	0.2995 0.9848	30	0.2709 0.9873	0.3607 0.9809	0.3094 0.9841
HABiLstm_C(+64)	离题 不离题	52	0.2519 0.9861	0.2991 0.9825	0.2735 0.9843	35	0.2348 0.9875	0.3724 0.9761	0.288 0.9817	38	0.2701 0.9863	0.305 0.9838	0.2865 0.985
HABiLstm_W(+64)	离题 不离题	61	0.2768 0.9867	0.3255 0.9832	0.2992 0.9849	81	0.3299 0.9858	0.2786 0.9888	0.3021 0.9873	50	0.3262 0.9864	0.3109 0.9873	0.3183 0.9869
Bert_CLS(微调9-11层+64)	离题 不离题	39	0.4778 0.9854	0.2522 0.9946	0.3301 0.99	42	0.4911 0.9852	0.2434 0.995	0.3255 0.9901	13	0.3344 0.9863	0.305 0.988	0.319 0.9872

Figure 1: *DataSet*<sub>11</sub> (341:17330) 实验结果

		方案一, strategy='topk', k=3				方案二				方案三			
		最优Threshold	P	R	F1	最优Threshold	P	R	F1	最优Threshold	P	R	F1
HABiLstm_C	离题 不离题	29	0.1063 0.9777	0.2194 0.9488	0.1432 0.963	27	0.0994 0.9775	0.2168 0.9454	0.1363 0.9612	34	0.1382 0.9765	0.1569 0.9728	0.147 0.9747
HABiLstm_W	离题 不离题	41	0.1591 0.9777	0.2015 0.9704	0.1778 0.974	46	0.1812 0.9771	0.1747 0.9781	0.1779 0.9776	29	0.1942 0.9775	0.1888 0.9783	0.1915 0.9779
Bert_CLS(微调全部层)	离题 不离题	29	0.1298 0.9776	0.2054 0.9618	0.1591 0.9696	33	0.1388 0.9777	0.2066 0.9644	0.1661 0.971	24	0.1427 0.9765	0.1569 0.9738	0.1495 0.9752
Bert_CLS(微调9-11层)	离题 不离题	55	0.193 0.9772	0.176 0.9796	0.1841 0.9784	43	0.1783 0.9778	0.2015 0.9742	0.1892 0.976	27	0.1955 0.9775	0.1875 0.9786	0.1914 0.978
HABiLstm_C(+64)	离题 不离题	32	0.1677 0.9777	0.2015 0.9722	0.1831 0.975	25	0.1653 0.978	0.213 0.9701	0.1862 0.974	12	0.1464 0.9782	0.227 0.9632	0.178 0.9707
HABiLstm_W(+64)	离题 不离题	36	0.1818 0.978	0.2092 0.9739	0.1945 0.9759	52	0.2184 0.9772	0.1722 0.9829	0.1926 0.98	18	0.1902 0.9788	0.2423 0.9713	0.2131 0.9751
Bert_CLS(微调9-11层+64)	离题 不离题	6	0.2082 0.978	0.2079 0.978	0.208 0.978	5	0.21 0.9782	0.2143 0.9776	0.2121 0.9779	7	0.2346 0.9774	0.1798 0.9837	0.2036 0.9805

Figure 2: *DataSet*<sub>12</sub> (784:28237) 实验结果

– 归一化（未完成）

- 实验结果：见图1、2
  - 增加离题/不离题作文比例整体指标都有下降
  - 去掉部分离题数据后指标上升
  - 方案三指标在大部分模型上均有提升

#### 4 下周计划

1. [\*\*\*] 补充分析实验结果。
3. [\*\*\*] 修改报告。
4. [\*\*] 补充仲裁数据。