

周报，2020年11月16日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. [***] 数据重标。
2. [***] baseline实验。

2 上周计划执行情况

1. [✓×] 完成重标数据分析。
2. [✓×] 重新构建数据集，修改实验方案，完成部分实验指标。

3 本周部分重点工作详述

3.1 重构数据集

- *DataSet₁*
 - 原始数据：共35个题目，603篇作文（不离题作文275篇，离题作文328篇）；
 - 补充数据：每个题目下添加一类卷+二类卷，使每个题目下50篇作文；
 - 数据量：35个题目×50篇作文（不离题作文1422篇，离题作文328篇）；
- *DataSet₂*
 - 数据量：30个题目×50篇作文（不离题作文750篇，离题作文750篇）；
 - 单个题目下数据：25篇离题作文，25篇不离题作文

3.2 实验结果

- 实验说明：
 - BertModel:原始模型，取[CLS]表示
 - topk=3
 - 阈值取1
- 实验结果：见下图
 - 相对比之前的指标整体有所下降
 - 不离题作文recall指标很低
 - Bert原始模型的效果较差
 - 词向量模型的结果很差 (?)
- 分析：
 - 330个题目中的范文在分类模型训练集中抽取，得到的表示都偏向于同一类，对于测试作文模型可能会偏向某一类，造成与对应分类下范文的相似度值很高，尝试调高阈值试试；
 - check一下代码和相似度计算结果

		取average				取max				取top3			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
BertModel	离题作文	-	0.20	0.98	0.33	-	0.20	0.97	0.33	-	0.20	0.97	0.33
	不离题作文	-	0.95	0.10	0.19	-	0.93	0.09	0.17	-	0.94	0.11	0.20
	总体(macro avg)	0.27	0.57	0.54	0.26	0.26	0.57	0.53	0.25	0.27	0.57	0.54	0.27
HBiLstm 字向量	离题作文	-	0.22	0.84	0.34	-	0.23	0.84	0.36	-	0.24	0.81	0.37
	不离题作文	-	0.89	0.29	0.44	-	0.90	0.36	0.51	-	0.90	0.39	0.55
	总体(macro avg)	0.40	0.55	0.57	0.39	0.45	0.57	0.60	0.44	0.47	0.57	0.60	0.46
HBiLstm 词向量	离题作文	-	0.19	1.00	0.08	-	0.19	1.00	0.32	-	0.19	1.00	0.32
	不离题作文	-	1.00	0.00	0.01	-	0.86	0.00	0.01	-	1.00	0	0.01
	总体(macro avg)	0.19	0.59	0.50	0.16	0.19	0.52	0.50	0.16	0.19	0.59	0.50	0.16

Figure 1: $DataSet_1$ 实验结果

	取average				取max				取top3			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
离题作文	-	0.59	1.00	0.74	-	0.60	0.99	0.74	-	0.63	0.99	0.77
不离题作文	-	0.99	0.31	0.47	-	0.97	0.33	0.49	-	0.97	0.41	0.58
总体(macro avg)	0.65	0.79	0.65	0.61	0.66	0.78	0.66	0.62	0.70	0.80	0.70	0.67
其他题目下作文	-	0.54	1.00	0.70	-	0.54	0.99	0.70	-	0.57	0.99	0.72
随机打乱	-	0.15	0.99	0.26	-	0.15	0.98	0.27	-	0.17	0.98	0.28
优美散文	-	0.10	0.98	0.18	-	0.10	0.95	0.18	-	0.11	0.98	0.20

Figure 2: HBiLstm字向量模型 $DataSet_2$ 实验结果

	取average				取max				取top3			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
离题作文	-	0.50	1.00	0.67	-	0.50	0.99	0.67	-	0.50	1.00	0.67
不离题作文	-	0.75	0.00	0.01	-	0.43	0.00	0.01	-	0.60	0.00	0.01
总体(macro avg)	0.50	0.63	0.50	0.34	0.50	0.46	0.50	0.34	0.50	0.55	0.50	0.34
其他题目下作文	-	0.44	1.00	0.61	-	0.44	1.00	0.62	-	0.44	1.00	0.61
随机打乱	-	0.11	1.00	0.19	-	0.11	1.00	0.19	-	0.11	1.00	0.19
优美散文	-	0.07	1.00	0.14	-	0.07	0.98	0.14	-	0.07	0.98	0.14

Figure 3: HBiLstm词向量模型 $DataSet_2$ 实验结果

	取average				取max				取top3			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
离题作文	-	0.52	0.99	0.68	-	0.52	0.99	0.68	-	0.53	0.99	0.69
不离题作文	-	0.93	0.07	0.14	-	0.93	0.07	0.14	-	0.94	0.10	0.19
总体(macro avg)	0.53	0.72	0.53	0.41	0.53	0.72	0.53	0.41	0.55	0.73	0.55	0.44
其他题目下作文	-	0.39	0.99	0.56	-	0.39	0.99	0.56	-	0.40	0.99	0.57
随机打乱	-	0.30	1.00	0.46	-	0.30	1.00	0.46	-	0.31	1.00	0.47
优美散文	-	0.10	1.00	0.18	-	0.10	0.98	0.17	-	0.10	0.98	0.18

Figure 4: Bert模型 $DataSet_2$ 实验结果

4 存在问题

- 词向量模型结果不正常

5 下周计划

1. [***] check代码和相似度计算结果。
2. [***] 梳理方案，进行验证。