

周报，2021年05月24日

屈原斌  
首都师范大学  
ybqu@cnu.edu.cn

## 1 上周计划

1. 通用场景离题检测实验更新
2. 标题-正文匹配实验更新
2. 成语古诗文检错实验更新

## 2 上周计划执行情况

1. [×] 中文实验未更新
2. [√]
3. [√]

## 3 本周部分重点工作详述

### 3.1 通用场景离题检测实验

- 数据集：
  - ICLE数据集，共10个主题，离题:不离题=51:733(1:14)
- 实验策略：
  - 计算测试集中任意两篇作文之间的相似度
  - 对每篇作文，选择相似度最大的k个值取平均，作为当前作文评分
  - 对所有作文，按照评分从小到大排序
- 指标见表1
- 结论：
  -

### 3.2 标题-正文匹配实验

- 数据集：
  - 共341篇测试作文，符合题意:不符题意=47:294(1:6)
- 实验策略：
  - 计算当前作文与其他作文之间的相似度，取topk的作文标题
  - 使用doc2vec分别获取当前作文标题表示 $h$ 与topk作文标题表示 $h_n, n = 1, 2, \dots, k$ ，对topk的作文标题表示取平均 $h_{avg}$
  - 计算 $h$ 与 $h_{avg}$ 的相似度 $S$

		R@10	R@20	R@50	P@1	P@5	P@10	spearman	spearman@10	ndcg	ndcg@10
TFIDF	k=5	0.4392	0.6153	0.8314	0.5	0.32	0.24	0.2573	0.3817	0.9775	0.8244
	k=10	0.4785	0.6153	0.8314	0.5	0.32	0.26	0.2635	0.2963	0.978	0.8196
Skip-gram	k=5	0.5002	0.5902	0.7412	0.3	0.3	0.26	0.2143	0.166	0.9786	0.8248
	k=10	0.5002	0.6294	0.7412	0.3	0.32	0.26	0.2162	0.1327	0.9764	0.8231
Doc2vec	k=5	0.3791	0.6094	0.7471	0.5	0.28	0.2	0.1913	0.2551	0.9734	0.8511
	k=10	0.3791	0.5761	0.7471	0.5	0.26	0.2	0.2007	0.2213	0.9746	0.8423
Feature-vec	k=5	0.4308	0.6486	0.7863	0.4	0.3	0.24	0.1873	0.248	0.9753	0.8217
	k=10	0.4308	0.6486	0.7863	0.5	0.32	0.24	0.2112	0.2883	0.9771	0.8155
Bert-Classification	k=5	0.4059	0.6427	0.9529	0.2	0.22	0.23	0.2601	0.0341	0.9821	0.8434
	k=10	0.4059	0.5702	0.9137	0.2	0.22	0.23	0.2496	0.0518	0.9818	0.8412
Bert-Gen	k=5	0.3857	0.5702	0.7627	0.2	0.3	0.21	0.1728	0.0513	0.9723	0.863
	k=10	0.4191	0.5643	0.7627	0.2	0.3	0.22	0.1706	0.0061	0.9736	0.8624

Table 1: 离题实验更新结果

- 根据阈值判断是否符合题意
- 指标见表2
- 结论：
  - 使用Last1avg表示优于CLS
  - k值增大指标上升

			符合题意				不符题意		
最优阈值			Accuracy	precision	recall	F1-score	precision	recall	F1-score
BERT-Gen_CLS	k=5	0.50	0.7214	0.9028	0.7585	0.8244	0.2447	0.4894	0.3262
	k=10	0.60	0.6540	0.9190	0.6565	0.7659	0.2290	0.6383	0.3371
BERT-Gen_Last1avg	k=5	0.56	0.6686	0.9132	0.6803	0.7797	0.2295	0.5957	0.3314
	k=10	0.60	0.6452	0.9220	0.6429	0.7575	0.2279	0.6596	0.3388

Table 2: 标题-正文匹配实验结果

#### 4 下周计划

1. [\*\*\*\*\*] TFIDF/分布式表示验证实验
2. [\*\*\*] 标题-正文匹配测试集构建，更新三档实验结果
3. [\*\*\*] 补全中英文离题检测实验
4. [\*\*\*] 成语古诗文检错替换方案