

周报，2021年06月21日

屈原斌  
首都师范大学  
ybqu@cnu.edu.cn

## 1 上周计划

1. 标题-正文匹配实验更新
2. 梳理技术方案图
3. 格式检查

## 2 上周计划执行情况

1. [×]
2. [×]
3. [×]

## 3 本周部分重点工作详述

### 3.1 标题-正文匹配实验更新

#### • 数据集：

- $Dataset_1$ : 同步作文，共59个主题，199篇原始作文，208篇构造数据，不离题:部分离题:完全离题=202:127:78
- $Dataset_2$ : 乐乐课堂随机替换题目，共2200篇，不离题：完全离题2000：200

#### • 实验策略：

##### 1. 策略一：

- \* 使用bert生成模型获取表示，检索相应标题，计算相似度
  - baseline:使用TFIDF和主题词计算相似度
- \* 参考作文：50W 乐乐课堂作文
- \* 指标见表1，2
- \* 结论：
  - 指标与参考作文数成正比
  - $Dataset_1$ CLS指标优于Last1avg,  $Dataset_2$ 相反
  - $Dataset_1$ 相似度取max指标最优， $Dataset_2$ 相似度取平均指标最优

##### 2. 策略二：

- \* 使用BertNSP做二分类任务
- \* 指标见表3
- \* 结论：
  -

		阈值	Acc	优		良		中		Marco-F1	跨二指标				
			P	R	F1-score	P	R	F1-score	P	R	F1-score				
BERT-Gen	baseline		0.75-0.85	0.4767	0.5547	0.7277	0.6296	0.3236	0.2362	0.2727	0.3469	0.2179	0.2677	0.39	0.1302
	topk表示取平均计算相似度	CLS	topk=50,0.50-0.60	0.516	0.5992	0.7624	0.671	0.4253	0.2913	0.3458	0.3016	0.2436	0.2695	0.4288	0.14
		Last1avg	topk=50,0.50-0.60	0.4963	0.5869	0.7525	0.6594	0.3735	0.2441	0.2952	0.2923	0.2436	0.2657	0.4068	0.14
	topk相似度取max	CLS	topk=50,0.60-0.80	0.4914	0.6685	0.6089	0.6373	0.3548	0.3465	0.3506	0.3333	0.4231	0.3729	0.4536	0.3729
		Last1avg	topk=100,0.60-0.80	0.5037	0.6635	0.6832	0.6732	0.3566	0.3622	0.3594	0.3	0.2692	0.2838	0.4388	0.0934
	topk相似度取平均	CLS	topk=20,0.35-0.50	0.4816	0.6444	0.5743	0.6073	0.3713	0.4882	0.4218	0.3	0.2308	0.2609	0.43	0.1007
		Last1avg	topk=20,0.35-0.45	0.4889	0.5926	0.7129	0.6472	0.3786	0.3071	0.3391	0.2623	0.2051	0.2302	0.4055	0.1376

Table 1: Dataset<sub>1</sub>指标更新

		阈值	优				中		
			Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
topk_取平均计算相似度	CLS	topk=15,0.55-0.65	0.8664	0.9677	0.8825	0.9231	0.375	0.705	0.4896
	Last1avg	topk=5,0.50-0.60	0.8614	0.9796	0.8655	0.919	0.3788	0.82	0.5182
topk相似度取max	CLS	topk=5,0.55-0.65	0.8673	0.9809	0.871	0.9227	0.3915	0.83	0.5321
	Last1avg	topk=5,0.55-0.65	0.8814	0.9801	0.8875	0.9315	0.4216	0.82	0.5569
topk相似度取平均	CLS	topk=50,0.65-0.75	0.9018	0.9822	0.9085	0.9439	0.4771	0.835	0.6073
	Last1avg	topk=100,0.70-0.80	0.9095	0.9818	0.9175	0.9486	0.5015	0.83	0.6262

Table 2: Dataset<sub>2</sub>指标更新

4 下周计划

- 1. [\*\*\*] 完成技术方案图
- 2. [\*\*\*] 标题-正文匹配更新attention方案

	优				中		
	Acc	Precision	Recall	F1-score	Precision	Recall	F1-score
Baseline	0.8814	0.9801	0.8875	0.9315	0.4216	0.82	0.5569
训练集15W	0.9718	0.9929	0.976	0.9844	0.7949	0.93	0.8571
训练集30W	0.9359	0.9952	0.934	0.9636	0.5913	0.955	0.7304
训练集60W	0.9227	0.9962	0.9185	0.9558	0.5421	0.965	0.6942
训练集75W	0.9291	0.9984	0.9235	0.9595	0.5629	0.985	0.7164

Table 3: 分类指标更新