周报，2021年07月05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. 代码封装
3. 更新Demo

2 上周计划执行情况

1. [×]
2. [×]
3. [×] Demo数据库问题未解决

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：
  - 构造三档数据
    - 构造方法，见图1
      1. 优 - 乐乐课堂/笔神作文
      2. 良 - 计算正文相似度，随机选择相似度较小作文，随机替换部分作文
      3. 中 - 计算正文相似度，随机选择相似度较小作文，替换正文
    - 数据分布，见表1

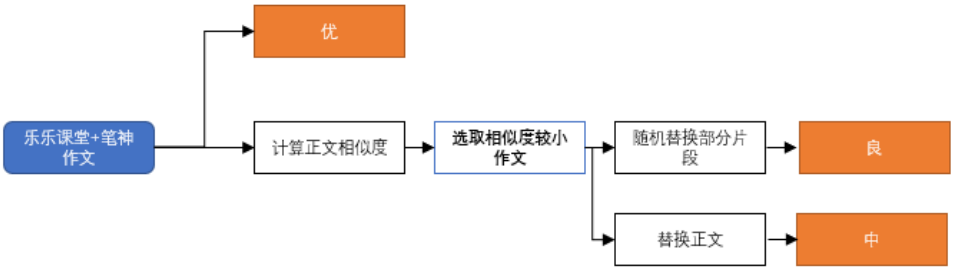

Figure 1: 数据集构造方法

- 实验策略：
  1. 策略一：BERT-Gen HT(无监督)
     - 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

		优	良	中	合计	数据来源
训练集		12000	3000	3000	18000	乐乐课堂，随机替换标题
验证集		500	50	50	600	
测试集	**Dataset_1**	202	127	78	407	同步作文（199原始数据+208篇构造数据，双标）
	Dataset_2	2000	200	200	2400	乐乐课堂，随机替换标题

Table 1: 数据集分布

2. 策略二：BERT-NSP
   * 使用BertNSP做二分类任务
3. 策略三：Attention sum reader
   * 使用SPP

		阈值	模型预测				优			良			中		
		-	**Acc**	**Marco-F1**	跨二档率	相关系数	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Baseline		0.75 / 0.85	0.4767	0.3900	0.1302	0.2546	0.5547	0.7277	0.6296	0.3236	0.2362	0.2727	0.3469	0.2179	0.2677
BERT-Gen(Topk相似度取最大值)	**CLS**	Topk=50 / 0.60 / 0.80	0.4939	**0.4571**	0.1081	0.3439	0.6685	0.6089	0.6373	0.3577	0.3465	0.3520	0.3400	0.4359	0.3820
	Last1avg	Topk=100 / 0.60 / 0.80	0.5037	0.4388	0.0934	0.3336	0.6635	0.6832	0.6732	0.3566	0.3622	0.3594	0.3000	0.2692	0.2838
BERT-NSP			0.4914	0.3136	0.1304	-	0.5184	0.9059	0.6595	0.2593	0.0551	0.0909	0.3704	0.1282	0.1905
ASReader			0.5184	0.3310	0.1400	-	0.5245	0.9554	0.6772	0.0000	0.0000	0.0000	0.5000	0.2308	0.3158

Table 2: $Dataset_1$指标更新

	阈值	模型预测			优			良			中		
	-	**Acc**	**Marco-F1**	跨二档率	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT-NSP		0.8650	0.6838	0.0804	0.9257	0.9215	0.9236	0.4588	0.3900	0.4216	0.6485	0.7750	0.7062
ASReader		0.8400	0.5258	0.1302	0.8880	0.9435	0.9149	0.4091	0.0900	0.1475	0.4805	0.5550	0.5151

Table 3: $Dataset_2$指标更新

4 下周计划

1. [***] 补充实验（尝试生成模型与NSP结合）
2. [***] 更新Demo