周报，2021年07月19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 上周计划执行情况

1. [×]

3 本周部分重点工作详述

3.1 标题-正文匹配实验更新

3.1.1 英文数据集

- 更新BERT生成模型：
  - 数据集：
    - 数据来源：bytecup_2018，英文作文
    - 数据分布，见表1
  - 指标更新，见表2

Bytecup_2018	数据量	正文平均句数	正文平均长度	标题平均长度
训练集	145821			
验证集	500	21	664	12
测试集	500			

Table 1: 生成模型数据分布

	F	**P**	**R**
Rouge-1	0.047	0.135	0.032
Rouge-2	0.001	0.005	8E-04
Rouge-l	0.049	0.151	0.031

Table 2: 生成模型指标

- 更新离题实验：
  - 测试集：
    - 数据来源：ICLE，13个主题，共830篇离题作文
    - 数据处理：去除3个主题（46篇，作文数太少/离题作文占比大），剩余784篇，按照3.0分划分，离题:不离题=51:733
    - 数据分布见表3
  - 指标更新见表4

Score	1	1.5	2	2.5	3	3.5	4	合计	主题数
原始数据	0	0	8	44	105	230	443	830	13
去除主题	0	0	8	43	103	215	415	784	10

Table 3: 英文离题数据分布

	R@10	R@20	R@50	P@1	P@5	P@10	spearman	ndcg	ndcg@10
TFIDF	0.3227	0.5486	0.8137	0.3000	0.3000	0.2100	0.2623	0.9799	0.8378
Doc2vec	0.3365	0.4018	0.8686	0.1000	0.1800	0.1700	0.1786	0.9755	0.8619
Skip-gram	0.4394	0.6212	0.8255	0.4000	0.3200	0.2500	0.3143	0.9811	0.8108
IDF-Embedding	0.2983	0.5751	0.8235	0.3000	0.2000	0.1600	0.1405	0.9774	0.8857
BERT-HT	0.4761	0.5378	0.7078	0.5000	0.4200	0.2600	0.2470	0.9802	0.8187

Table 4: 英文数据集指标更新

- 结论：
  - skip-gram指标最优
- BERTHT方案,参考作文1w

- 问题
  - ？BERTNSP方案
    - 明确任务
    - 训练数据构造

3.1.2 中实验更新

- 数据集：400篇作文，完全离题:部分离题:不离题=75:126:199
- 指标更新见表
  - 更新IDF-Embedding，（使用idf对word2vec表示加权）
  - 特征方案（未完成）
- 指标更新见表3
- 结论：
  - IDF-Embedding指标较Skip-gram提升
  - BERT-Gen HT方案指标最优

DataSet1		阈值	模型预测							随机Baseline	
		-	Acc	Marco-F1	跨二档率	相关系数	不离题F1	部分离题F1	完全离题F1	Acc	Marco-F1
人1-人2		-	0.6825	0.6219	0.0175	0.697	-	-	-	-	-
Baseline(线上系统)		0.80 / 0.95	0.385	0.3695	0.0775	0.2148	0.2879	0.4557	0.3649	0.3075	0.2668
标题-正文计算相似度	TFIDF	0.10 / 0.35	0.445	**0.4314**	0.1275	0.3557	0.5552	0.3077	0.4312	0.3025	0.2923
	Doc2vec	0.30 / 0.40	0.4125	0.3725	0.095	0.1659	0.511	0.3666	0.24	0.345	0.2944
	Skip-gram	0.55 / 0.70	0.4675	0.3847	0.1075	0.2065	0.6143	0.3033	0.2364	0.385	0.2939
	IDF-Embedding	0.40 / 0.55	0.425	0.3962	0.1125	0.2312	0.5459	0.2966	0.3462	0.3575	0.3184
BERT-Gen HT (Topk相似度取最大值，CLS，50W)		100 / 0.60 / 0.80	0.5025	**0.4367**	0.0925	0.3341	0.6716	0.3468	0.2917	0.345	0.2804
BERT-NSP			0.5275	0.4081	0.125	0.301	0.6851	0.2057	0.3333	0.39	0.3411
AS-Reader(+SPP,取平均, batch_size=16, lr=0.002)			0.505	0.3976	0.1375	0.2084	0.6531	0.2741	0.2655	0.4125	0.2943
AS-Reader(+SPP,取Last Time, batch_size=16)			0.5225	0.3386	0.1575	0.2491	0.6847	0.0303	0.3009	0.4525	0.2474

Table 5: 中文数据集指标更新

4 下周计划

1. [***] 讨论：
   - 任务明确(中英文任务对齐)
   - 方案、指标梳理
2. [***] 补充中英文实验