

# 2021年04月25日进度汇报

屈原斌  
首都师范大学  
ybqu@cnu.edu.cn

## 1 今日进度

### 1. 更新离题检测实验

## 2 工作详述

### • 实验数据：

– ICLE数据集，11个主题，共827篇作文，离题:切题=51:776 (1:15)

### • 实验方案：

#### – 聚类方案一：

- \* 计算聚类结果中可能离题的类与范文类的相似度
- \* distance\_threshold从最小到最大距离进行调参，example\_threshold从1到10进行调参
- \* 实验结果见表1
- \* 结论：
  - doc2vec表示时指标最优

#### – 聚类方案二（one-class）：

- \* 计算全部作文与它们的质心的相似度，指标见表2
- \* 结论：
  - bert分类模型指标最优

#### – 聚类方案三（Prompt-independent）

- \* 五折交叉验证，数据划分见表3
- \* 指标见表4
- \* 结论：
  - bert生成模型指标最优
- ！开发集调参后测试集部分主题聚类后没有小类

		R@10	R@15	R@20	R@50	P@1	P@5	P@10	spearman
baseline		0.5334	0.5464	0.6676	0.7321	0.2727	0.3091	0.2727	0.1355
tfidf		0.4350	0.4836	0.5291	0.7861	0.4545	0.3273	0.2364	0.2260
doc2vec		0.6646	0.6646	0.6797	0.7464	0.1818	0.2364	0.2909	0.1698
分类模型	lstm	0.4150	0.4204	0.5173	0.6318	0.3636	0.2000	0.2000	0.1257
	bert	0.6146	0.6146	0.6449	0.7661	0.2727	0.2545	0.2818	0.1790
生成模型	lstm	0.5348	0.5478	0.5629	0.7562	0.0909	0.2000	0.2273	0.0137
	bert	0.5355	0.5485	0.5485	0.6576	0.2727	0.2364	0.2636	0.0657

Table 1: 聚类方案一指标更新

		<b>R@10</b>	<b>R@15</b>	<b>R@20</b>	<b>R@50</b>	<b>P@1</b>	<b>P@5</b>	<b>P@10</b>	<b>spearman</b>
	<b>baseline</b>	0.3917	0.4631	0.5594	0.7148	0.4545	0.2909	0.2182	0.1429
	<b>tfidf</b>	0.4350	0.4533	0.5041	0.7558	0.4545	0.3091	0.2364	0.2437
	<b>doc2vec</b>	0.3865	0.4381	0.5291	0.7041	0.5455	0.2727	0.2182	0.1865
分类模型	<b>lstm</b>	0.3728	0.4062	0.4479	0.7398	0.4545	0.2727	0.1909	0.1535
	<b>bert</b>	0.3029	0.4480	0.5419	0.7344	0.2727	0.2182	0.1818	0.0960
生成模型	<b>lstm</b>	0.1263	0.2463	0.3676	0.7041	0.0000	0.0727	0.0818	-0.0039
	<b>bert</b>	0.5259	0.5389	0.5389	0.7504	0.3636	0.2364	0.2636	0.1266

Table 2: 聚类方案二（one-class）指标更新

	主题数	作文数
一折	1	237
二折	2	144
三折	2	151
四折	3	141
五折	3	154

Table 3: 五折划分数据

		<b>R@10</b>	<b>R@15</b>	<b>R@20</b>	<b>R@50</b>	<b>P@1</b>	<b>P@5</b>	<b>P@10</b>	<b>spearman</b>
	<b>baseline</b>	0.1524	0.1676	0.1676	0.3155	0.1818	0.1273	0.0909	0.0855
	<b>tfidf</b>	0.3138	0.3624	0.3775	0.5437	0.3636	0.2727	0.2000	0.1711
	<b>doc2vec</b>	0.2098	0.2205	0.2259	0.3328	0.0909	0.1273	0.1091	0.0517
分类模型	<b>lstm</b>	0.1023	0.1380	0.1683	0.2645	0.0909	0.0364	0.0455	-0.0061
	<b>bert</b>	0.1679	0.3497	0.3800	0.4566	0.1818	0.1818	0.1091	0.1551
生成模型	<b>lstm</b>	0.2194	0.2816	0.3270	0.6067	0.0000	0.0545	0.1091	-0.0386
	<b>bert</b>	0.4241	0.4847	0.5052	0.6630	0.3636	0.3091	0.2182	0.1149

Table 4: Prompt-independent方案指标更新