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（中文结果未跑完）
2. 成语古诗文检错指标评估/方案讨论
3. 标题正文匹配测试集

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和附件1
    - 结论：
      - doc2vec表示时指标最优

		R@10	R@15	R@20	R@50
baseline		0.5334	0.5464	0.5767	0.6412
tfidf		0.4350	0.4533	0.4988	0.5437
doc2vec		**0.6646**	0.6646	0.6797	0.7464
分类模型	**habilstm**	0.4150	0.4204	0.4567	0.5712
	bert	0.6146	0.6146	0.6449	0.7661
生成模型	**lstmabs**	0.5348	0.5478	0.5629	0.6956
	bertabs	0.5355	0.5485	0.5485	0.6576

Table 1: 聚类方案实验指标更新