周报，2021年02月18日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. [***] 实验更新

2 上周计划执行情况

1. [×] 英文生成模型未完成

3 本周部分重点工作详述

3.1 实验更新

3.1.1 中文数据集

- 题目生成实验更新（LSTM based Sequence-to-Sequence model for Abstractive Summarization）
  - 数据集：乐乐课堂数据
  - 数据分布见表1：
  - 指标见表2：

	训练集	验证集	测试集
作文数	68428	8553	8554

Table 1: 数据分布

	rouge_1	rouge_2	rouge_l
Seq2seq	0.2223	0.124	0.2185

Table 2: 生成模型指标更新

- 离题实验更新
  - 离题:切题=1542:59560(1:39)
  - 方案：基于排序方法（针对不同主题动态调节阈值），指标见表3

	离题			不离题		
	precision	recall	f1-score	precision	recall	f1-score
seq2seq	0.0393	0.4585	0.0724	0.9806	0.71	0.8236

Table 3: 中文数据集离题指标更新

3.1.2 英文数据集

- 标注数据分布，见表4(共13个主题，830篇作文)
- score=1.0-3.5作为离题作文，score=4.0作为切题作文，离题:切题=387:443（1:1.145）
- 离题实验更新：
  - 方案一：基于排序方法（针对不同主题动态调节阈值），指标见表5
    - 参考题目：筛选其他题目下作文数大于20篇的题目，共70个题目
  - 方案二：基于相似度（去除排序，直接根据相似度判断，若与主题相似度小于阈值，则判断为离题），指标见表6
- 结论：
  - 方案二指标优于方案一
  - 方案二中tr=0.4时指标最优（变化曲线见图1）

score	1	1.5	2	2.5	3	3.5	4
作文数	0	0	8	44	105	230	443

Table 4: 英文数据集标注数据分布

	离题			不离题		
	precision	recall	f1-score	precision	recall	f1-score
doc2vec	0.4762	0.0775	0.1333	0.5346	0.9255	0.6777

Table 5: 方案一指标更新

	离题			不离题		
	precision	recall	f1-score	precision	recall	f1-score
tr=0.25	0	0	0	0.5337	1	0.696
tr=0.4	0.6346	0.0853	0.1503	0.545	0.9571	0.6945
tr=0.5	0.4969	0.8165	0.6178	0.634	0.2777	0.3862

Table 6: 方案二指标更新

4 下周计划

1. [***] 完成英文seq2seq生成模型
2. [***] 完善英文数据实验

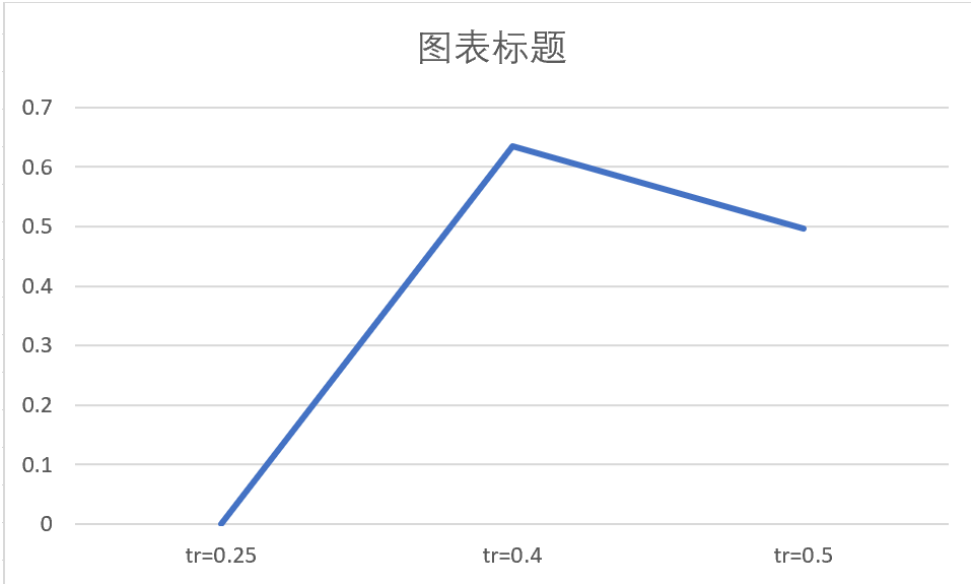

Figure 1: （方案二）不同阈值下F1变化