周报，2021年02月22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. [✓×]

3 本周部分重点工作详述

3.1 doc2vec & seq2seq模型

3.1.1 doc2vec模型（使用gensim.models.doc2vec）

- 中文模型：
  - 数据集：讯飞初高中作文，共1210887篇
  - 参数设置：
    - vettor_size(向量维度): 200
    - alpha(初始学习率)：0.025
    - windows(窗口大小)：10
    - min_count(最小词频)：5
    - 训练轮数：100
- 英文模型：
  - 数据集：ICLE数据集，共9529篇
  - 参数设置：
    - vettor_size(向量维度): 256
    - alpha(初始学习率)：0.025
    - windows(窗口大小)：10
    - min_count(最小词频)：5
    - 训练轮数：200

3.1.2 seq2seq模型（正文生成题目）

- 中文模型：
  - 基于LSTM的(Abstractive Summarization)摘要生成模型
  - 代码地址：https://github.com/LowinLi/Text-Summarizer-Pytorch-Chinese
  - 数据集：乐乐课堂作文数据，数据分布见表1
  - 模型参数：
    - 词向量：使用预训练词向量初始化（488921*50）
    - 训练分为两部分：
      - 极大似然估计方法训练(MLE)：lr=0.001

	训练集	验证集	测试集	合计
作文量	68428	8553	8554	85535

Table 1: 中文seq2seq模型数据集

	Rouge_l
Seq2seq	0.2185

Table 2: 中文生成模型指标

· 强化学习(RL)：lr=0.0001

- 指标，见表2

- 英文模型：
  - 基于LSTM的(Abstractive Summarization)摘要生成模型
  - [**?**]
  - 代码地址：https://github.com/rohithreddy024/Text-Summarizer-Pytorch
  - 数据集：OpenNMT provided Gigaword dataset（https://github.com/harvardnlp/sent-summary），数据分布见表3

	训练集	验证集	测试集
作文数	3803957	135129	135100

Table 3: 英文生成模型数据集

  - 模型参数：词向量随机初始化，其他设置同中文模型
  - 指标，见表4

3.2 英文数据集离题实验更新

- 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（直接在**测试集**上调节阈值）：
  - 方案一：基于排序方法（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（动态调节阈值），指标见表6

4 下周计划

1. [***] 划分开发集进行调参

	Rouge_l
seq2seq	0.4263

Table 4: 英文生成模型指标

		离题			不离题		
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
doc2vec	**tr=1**	0.4615	0.0775	0.1327	0.5333	0.921	0.6755
habilstm	**tr=1**	0.4703	0.5736	0.5169	0.5391	0.4357	0.4819
seq2seq	**tr=1**	0.4774	0.9018	**0.6243**	0.6162	0.1377	0.2251

Table 5: 方案一指标

		离题			不离题		
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
doc2vec	**tr=0.45**	0.5897	0.416	0.4879	0.5943	0.7472	0.662
habilstm	**tr=0.5**	0.5417	0.4031	0.4622	0.5738	0.702	0.6315
seq2seq	**tr=0.9**	0.4842	0.6718	**0.5628**	0.5666	0.3747	0.4511

Table 6: 方案二指标