半周报

屈原斌

2020 年 9 月 17 日

1 上周任务

- 使用BERT词向量作为模型embedding做分类实验；
- 统计处理数据；
- 转写任务；

1.1 分类实验

- 实验设置有一点问题，改过来了，结果还没有跑出来；

1.2 数据处理

- 重新给我了一批数据，在原有的基础上增加了200个主题，目前主要统计了各分数段的作文数信息（表1和表2），完成了所有作文和对应的分数的抽取，数据处理还没做完；
- 目前数据处理的思路：
  - 减小主题数量：从414各主题中抽取部分主题不相关（不相似）的主题，比如只抽取50个主题；
  - 减小训练集和测试集的大小：目前的单个主题下作文数分布不平均且数据量过大，可以根据分数信息抽取部分高分作文和部分低分作文；
- 问题：

– 单纯的根据分数信息来划分离题程度是不准确的，所以我觉得还是需要人工进行标注；

表 1: 各分数段作文数统计

分数段	[0, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100]	总计
作文数	104227	211624	612072	416573	46855	1391351

表 2: 0-60分数段作文数统计

分数段	0	(0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)	[50, 60)	总计
作文数	10521	4070	5371	6073	9675	17635	50882	104227

2 本周计划

- 完成分类部分的实验；
- 完成数据的处理；