周报

屈原斌

2020 年 9 月 21 日

1 上周任务

- 数据处理；
- 分类实验；

1.1 任务详细

1.1.1 数据处理

- 数据：初中作文，共759个主题；
- 方法：
  - 根据中考作文评分标准（60分制）对应到百分制后对作文分数进行等级划分，各分数段划分具体见表1；
  - 各分类作文数统计结果见表2；

1.2 分类实验

- 实验方法：在原来的模型HBiLstm上使用bert词向量做初始化进行训练；
- 目前的问题：对模型调参后的效果都特别的差，模型基本什么都没有学到，然后使用原来的模型的方式将维度设置到768后进行测试对比，结果和之前的结果基本一致，都优于使用bert词向量的方法；
- 实验设置：

表 1: 初中作文评分标准

作文等级	60分制	对应百分制
一类卷	[60, 54]	[100, 90]
二类卷	(54, 48]	(90, 80]
三类卷	(48, 36]	(80, 60]
四类卷	(36, 18]	(60, 30]
五类卷	(18, 0]	(30, 0]

*百分制分数包含小数部分，进行了取整

表 2: 各等级作文数统计（759主题）

作文等级	一类卷	二类卷	三类卷	四类卷	五类卷	总计
主题数	611	757	757	742	688	-
作文数	38978	327143	648553	63326	13192	1091192

*各等级下主题数少于总的主题数（759）

  - 训练集：414各定题作文中每个主题抽取100篇做为训练集；
  - batch_size: 64;
  - lr: 0.001;

- 实验对比结果见表3：

表 3: 实验结果对比

模型	Acc.	P	R	F1
HBiLstm_400	0.737	0.754	0.737	0.739
HBiLstm_768	0.725	0.740	0.725	0.727
HBiLstm*_768	0.002	0.0	0.002	0.0

*HBiLstm*表示使用bert词向量做初始化，400和768表示词向量维度

2 本周计划

- 检查代码，使用BertTokenizer进行encode验证一下是否有影响；
- 和讯飞那边进行对接，先进行数据标注；
- 尝试使用段落信息，对每篇文章中的段落进行离题分析；