Web信息处理与应用大作业

胡景文

2019101400

1. 数据集

本次实验采用后羿采集器对豆瓣最受欢迎的影评进行了采集，总共采集了10596条评论以及评论者对电影的打分。具体操作界面如下图所示，后羿采集器采集了每一个模块中的电影短评和后面的打分，打分分为五个等级，很差、较差、还行、推荐、力荐。在数据集的预处理过程中，对短评进行了去停用词以及去掉一些符号的操作，并且将标签根据等级映射为数字标签0到4。爬取的数据中部分仅有短评，没有标签，在预处理的过程中删除此类数据，剩余数据共9729条。



1. 模型

2.1 LSTM+Attention

模型分为三层，第一层为embedding层，包含一个vocab\_size\*embedding\_dim的矩阵，输入为句子中每一个词的one-hot向量组成的矩阵，输出为每一个词对应的特征表示，第二层为LSTM层，输入是一个句子中词的特征表示的集合，输出为所有time-step的隐层特征集合，第三层为fc层，利用attention机制将LSTM层的隐层输出进行加权求和后作为fc层的输入，输出是n个类的概率分布。第一层的embedding矩阵使用预训练好的特征矩阵进行初始化。

2.2 Bert

Bert是目前比较火的用来处理nlp问题的模型，在本次实验中基于bet-base-chinese预训练模型进行fine-tune，在bert模型后加上fc层，将模型进行句子级别的encode之后的特征作为fc层的输入，输出为n个类的概率分布。

1. 实验分析

|  |  |  |  |
| --- | --- | --- | --- |
|  | recall | precision | F1-score |
| LSTM+Attention | 54.56 | 55.21 | 54.95 |
| Bert | 59.7 | 60.6 | 60.3 |

表1.LSTM+Attention模型和Bert模型对影评数据集预测的recall, precision, F1-score结果

表1中为第二部分所述的两个模型在影评数据集上对评价预测的结果，可以看到Bert模型的结果还是优于LSTM+Attention模型的结果。

1. 结论

针对影评数据集，本次实验尝试了两种主流的nlp方法，结果表明bert的结果会优于LSTM+Attention，两个模型均能比较好的解决基于影评预测评论者对这篇文章打分的结果。