实验报告

PB17000163 周天尧
PB16000160 黄梓旭

实验目的

本实验要求以给定的英文文本数据集为基础，实现一个信息抽取系统

实验过程

本实验我们使用了两种不同的方法，分别是svm分类器和基于BERT的分类器。我们对比了它们的正确率，最终选取了正确率较高的一种方案提交。

SVM分类方案：

首先记录所有出现过单词的数量k，使用一个k维向量作为文本的特征向量，分别尝试了词频数，tfidf值作为对应元素，使用sklearn中LinearSVC进行分类。最终准确率在55%左右。后来我们尝试进行优化，首先是降维，不降维的数据有30000维左右，我们尝试了降到了3000-25000维，但是效果不明显甚至正确率有所下降。后来我们尝试计算特征与结果的相关系数，删去相关系数低于某个阈值的特征，我们设置了多个阈值，0.1，0.01，0.001等进行实验，最终在取0.01时正确率达到最高，为59%。

BERT方案：

使用huggingface transformers中的BERTForClassification模型，在此基础上继续训练。在训练过程中，测试集的正确率不高，分析每一类的判断成另一类的频率，发现Other类的错误率最高，考虑到Other类型的数据数量是比其它类多至多两倍，可能是由于样本数量不均匀导致整体正确率偏低，在测试平台上的正确率只有约68%。

于是我们对数据集进行简单复制的Over sampling处理，在从训练数据中划分出的测试集上正确率由78%上升到85%。最终在测试平台上的正确率是69%，相对没有Over sampling的提升不大。

实验总结

本次实验中，我们探索了传统SVM和深度学习在关系分类上的使用，发现在不做什么处理的情况下，使用预训练BERT的方法要好于SVM。

但是通过查阅文献发现，SVM方法的正确率和深度学习的方法五五开，SVM的方法使用了例如Position of speech等多种特征，才达到这个效果，可见SVM结合特征工程也可以做得很好，不需要遇事不决，机器学习。