基于BERT中文预训练模型的标准问分类器

本项目从实际问题出发，类似于在生活中我们所用到的智能客服，目标是输入多样化的问题输出统一的标准问，经过不断地学习和优化后，在结果准确率方面达到了88%左右，处于班级中游水平。

1. 问题描述

实现一个简单的问答系统的一种方法是：收到用户提出问题后，找到用户提问在已知回答的标准集合中对应的标准问，再利用标准问及其回答答复用户。

你的任务是：给定一个提问，例如：“预约取钱的有效时间是多久”, 预测它的标准问是什么，例如，这个问题对弈的标准问是：“预约取款的有效期限”。（所有可能的候选标准问是已知的）

将提供一份训练集文件（train.csv）和一份测试集文件（test.csv）。训练集的第一列为给定问题, 第二类为其对应的标准问。测试文件仅含有给定问题。

训练集含有2086个句子对，测试集含有1422 个给定问题。

1. 解决思路

根据以上描述，标准问全已知，为573种，故考虑将所有种类的标准问编码为数字（0-572）。调用transformers的AutoTokenizer加载“bert-base-chinese”预训练模型将query分词化，参数转换为pytorch张量后载入dataloader。为了验证训练精度，我考虑了将训练集划分开来，0.9为训练集，0.1为验证集。

模型调用huggingface中AutoModelForSequenceClassification序列分类器，加载“bert-base-chinese”预训练模型，optimizer选用AdamW优化器，并添加学习率调度器OneCycleLR更新学习率策略。

训练后保存在验证集上精度最高的模型参数用于预测结果。

1. 学习与改进

在个人汇报之前，我选择的是划分训练集与验证集为9：1，文本最大长度选择为35，经过各种超参数的调整，训练模型的预测效果（提交在kaggle上的结果精度）始终在84%左右徘徊。

我意识到训练集过于少，导致模型学习不够鲁棒，泛化能力不够强。于是我放弃了划分训练集，并且学习了其他同学的思路，将label本身也作为一个query进行预测，即query为“B股股票的成交原则是什么”，label也为“B股股票的成交原则是什么”，将训练集由2086个扩展到2659个，模型精度提升到了86%。

尔后我依然思考如何解决训练集的样本数量较少的问题，我考虑了交叉验证来更好地评估模型的性能和选择超参数，调用K-Fold进行多折训练（折数选择为5），但实际训练不仅大大延长了训练时间，多次调整参数后的训练效果也不尽人意，遂放弃。

最后考虑了通过数据增强来解决样本数量问题，我选用了nlpaug包中的ContextualWordEmbsAug增强器，依旧使用“bert-base-chinese”预训练模型，通过该模型来获取每个单词的上下文信息，并通过替换某些单词来增强文本数据，以达到保持句子的语义一致性的同时扩充训练数据集的效果。Epochs选择70时准确率达到88.46%，Epochs选择120时准确率达到88.88%。

1. 经验与总结

虽然本项目经过不断优化改进，准确率有所提升，但幅度不大，在排名靠前的几位优秀同学的结果面前更是相形见绌。具体分析还是本人水平有限，同时没有对文本进行具体的细节处理。文本有很大的操作空间但碍于有效方法积累的匮乏和评估难度较高，在尝试了几次处理之后效果不佳便放弃。

一学期深度学习课程使我一个非计科生对深度学习有了更深的认识，特别是在自然语言的处理方面有了一定的掌握。同时惊艳于两位同学训练的chatbot，也更让我笃定这门课将来一定会为我的学习工作提供支持。感谢崔老师与赵助教的付出！