实习成果汇报-第二周

张益宁

5.22-5.28 实习成果汇报

日程规划

|  |  |  |
| --- | --- | --- |
| **周** | **任务** | **完成情况** |
| 5.11~5.14 | 确定选题 | 完成 |
| 了解命名实体识别基础知识 | 完成 |
| 下载并初步分析数据集 | 完成 |
| 制定日程规划 | 完成 |
| 5.15~5.21 | 了解基于规则和词典的传统方法 | 完成 |
| 了解深度学习相关方法 | 完成 |
| 手动编写代码框架 | 完成 |
| 5.22~5.28 | 对代码框架进行传统深度学习适配 | 完成 |
| 训练调参 | 完成 |

上周回顾

本周主要是继续完成代码的优化和调参工作，上周完成了框架的搭建，但是训练的时候出现了不能收敛的问题。

本周一一解决了这些问题，并做了大量的实验：

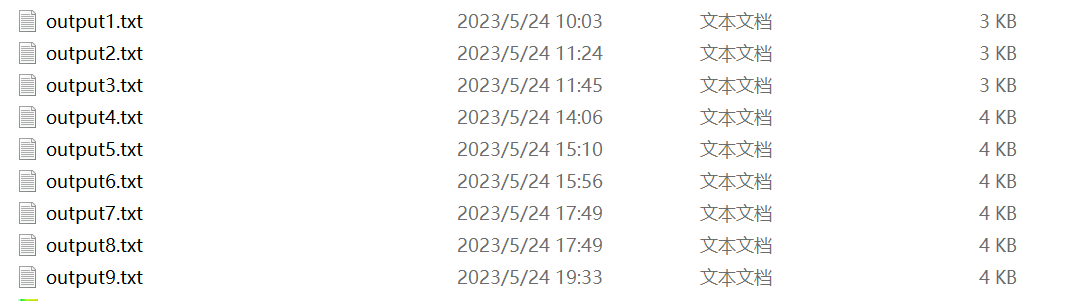


Figure 5.24当天实验内容

成果：

* 困难——模型难以收敛
* 原因——数据集分布严重不合理，比如对地点进行识别的时候100w个字符的数据集中只有1w多个B-LOC
* 操作1——先进行数据集清洗和预处理，我选择如果这句话中（两个空行之间算一句话）没有B-LOC或者I-LOC，那么就有一定的概率将其删去。
* 操作2——训练的时候一个batch里面如果发现这个真实的标签是O，那我就有一定的概率不让这个temp\_loss加到总体的loss中。
* 结果——还不错。

下面是一个实验记录：

参数：

MLP：1500\*256\*3

lr = 1e-3 # 学习率

batch\_size = 500 # batch

epoch = 10000 # 训练次数

val\_time = 500 # 每训练多少次验证一次

weight = 0.175 # 有多大的概率保留O的loss

预处理删除0.9的全O

基本信息统计：

训练集大小:481408 验证集大小：112189 测试集大小：223834

-------------train-------------

0/10000

loss=120387.4375

hit=1298 p\_sum=52630 r\_sum=4532

precision=0.024662739882196465 recall=0.28640776699029125

f1-measure=0.04541478604667436

……

10000/10000

loss=67642.0546875

hit=3048 p\_sum=7259 r\_sum=4532

precision=0.41989254718280755 recall=0.6725507502206531

f1-measure=0.5170044949537783

-------------test-------------

loss=134715.609375

hit=5645 p\_sum=14006 r\_sum=8606

precision=0.4030415536198772 recall=0.6559377178712527

f1-measure=0.4992924111091456

可以看到在测试集上的表现f1为0.5，对于一个MLP模型来说还挺成功的。

下周计划

接着调参训练，考虑到考试周的迫近可能还需要多兼顾一下课内的内容。

下周主要读读论文，了解一下目前自然语言处理的比较优秀的模型

BiLSTM

LSTM

CRF

LSTM+CRF

BERT

等等

下周按计划主要是学习新技术，读相关的论文，所以可能代码成果不会比较多，但是感觉读读新技术有助于目前我自己写的这个简单NER处理的代码。