探索性数据分析（EDA）

train: 1893

dev: 631

若将任务定义为序列分类：

可以构造样本（句子，xxx(其余特征)，被告人in被告人集合，是否关联）

train: 8539

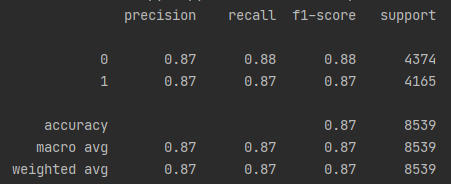
dev: 2863

train 正类(4165)：负类(4374)

直接判别法：遍历被告人集合，若被告人在字段“句子”当中，则将其判定为有关联的被告人

train: macro-F1=0.8735

dev: macro-F1=0.8909



bad case分析：

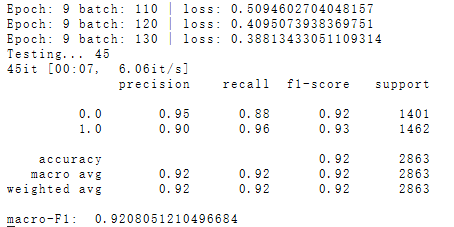
使用直接判别法对训练样本进行分类，存在1079个错分样本，其中FP包含517例，FN包含562例。

假正类（FP）：被告人在句子中，但与要素无关联

['本院认为，被告人杨鹏、姚敦启、尹德华共谋，利用杨鹏作为国家司法工作人员的职务便利，徇私枉法，为非法持有毒品的涉案人员姚某1提供虚假的立功材料，使罪重的姚某1受到轻判', '姚敦启']

假负类（FN）：被告人不在句子中，但与要素有关联

['其中标示为辉瑞制药有限公司生产，批号为29658183（盒内批号为B314833201）的“万艾可”（枸橼酸西地那非片），经长春市食品药品监督管理局药品稽查分局认定为按假药论处', '吴某某']



训练环境：

GPU: NVIDIA GeForce RTX 3090 显卡内存: 24 GB

Python3.7, CUDA 11.1, cuDNN 8.0.5, Pytorch 1.8.1, Ubuntu 18.04

参数设置

|  |  |
| --- | --- |
| Parameter | Value |
| Batch\_size | 64 |
| Epoch\_size | 10 |
| Pad\_size | 256 |
| Learngin\_rate | 5e-5 |
| Weight\_decay | 1e-2 |

调参过程：（BERT fine tune）

调整输出层；加入“##\*”

BERT + FC(L) + “##”: 0.9042

BERT + FC(L): 0.9150

BERT + FC(S): 0.9208

FC(L):

Linear(768, 384)

relu

Linear(384, 1)

dropout(0.5)

FC(S):

Linear(768, 1)

dropout(0.5)