用BERT进行命名实体识别过程的报告

---- 以人民日报文本信息为例

李明聪

摘要：本文用标注后的人民日报信息作为数据集，使用BERT进行了命名实体识别。数据集共240万个字符，模型在GPU上训练耗时约50分钟，模型预测的f1 score为0.96。

**一、数据集概述**

1. 数据来源：人民日报1998年上半年的文本语料，进行了分词和词标注。

2. 数据集说明：

训练集共222万个字符，验证集共18万个字符，测试集共2362个字符。

2. 标注说明

(1) 实体位置的标注

|  |  |  |
| --- | --- | --- |
| 标注 | 含义 | 含义 |
| B-X | Begin | 代表实体X的开头 |
| I-X | Inside | 代表实体的内部 |
| O-X | outside | 代表不属于任何类型的 |

(2) 实体类型的标注：X

|  |  |
| --- | --- |
| LOC | 地名 |
| PER | 人民 |
| ORG | 机构名 |

4. 示例：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 《 | 北 | 京 | 文 | 物 | 保 | 存 | 保 | 管 | 状 | 态 | 之 | 调 | 查 | 报 | 告 | 》 |
| O | B-LOC | I-LOC | O | O | O | O | O | O | O | O | O | O | O | O | O | O |
| 调 | 查 | 范 | 围 | 涉 | 及 | 故 | 宫 | 、 | 历 | 博 | 、 | 古 | 研 | 所 |  |  |
| O | O | O | O | O | O | B-LOC | I-LOC | O | B-LOC | I-LOC | O | B-ORG | I-ORG | I-ORG |  |  |

**二、工作环境**

Google CoLab

**三、训练过程**

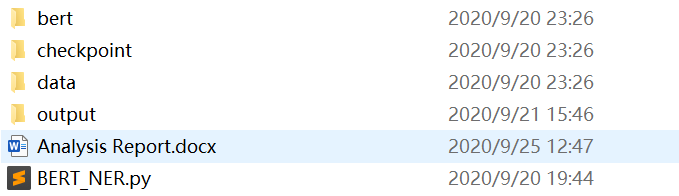
**1.准备好所有用到的文件。**

\* bert：从BERT官方GitHub仓库下载的代码

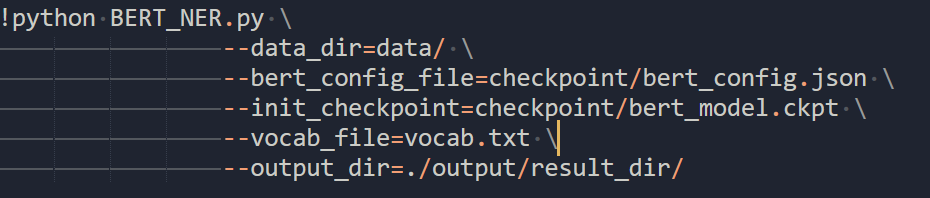
\* checkpoint：存放中文预训练模型

\* data：要用到的数据集

\* BERT\_NER：运行这个py文件进行学习与预测。



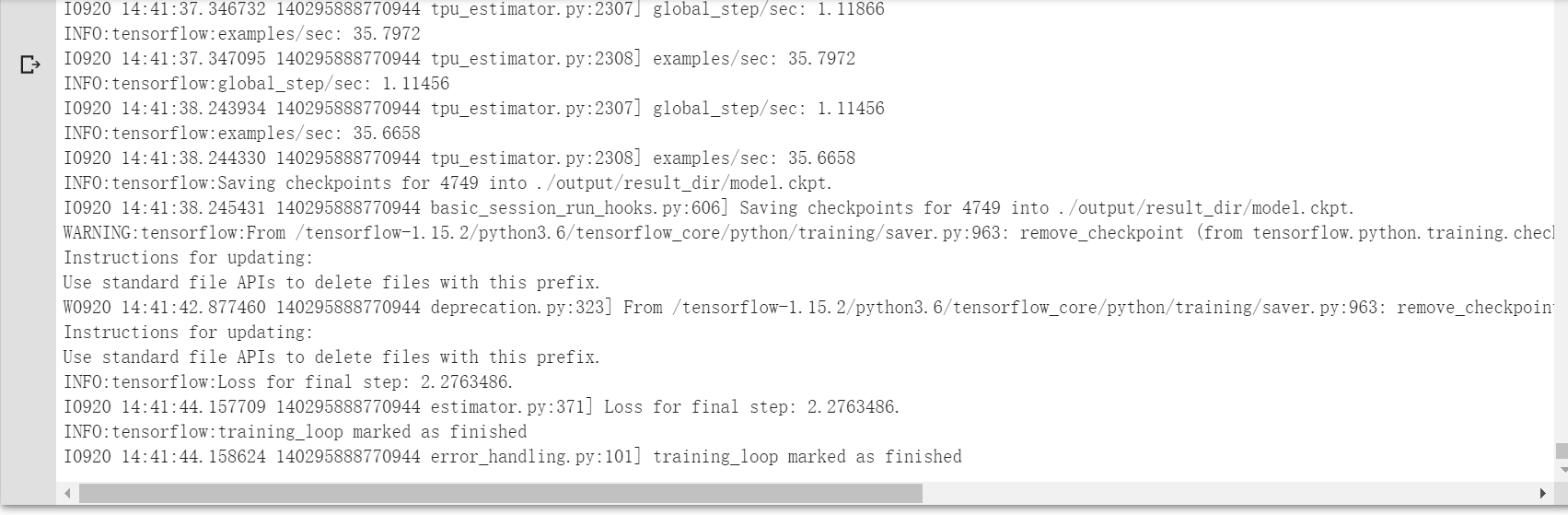
**2. 训练模型**



在CoLab里面可以通过Linux命令运行py文件。

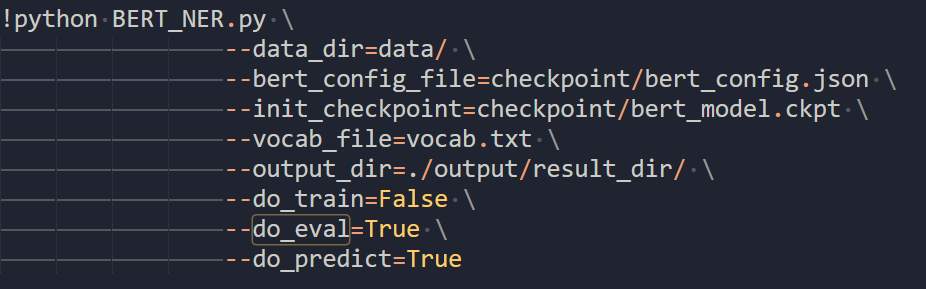
除第一行外，其他行均为传入的参数。

训练完毕的截图：

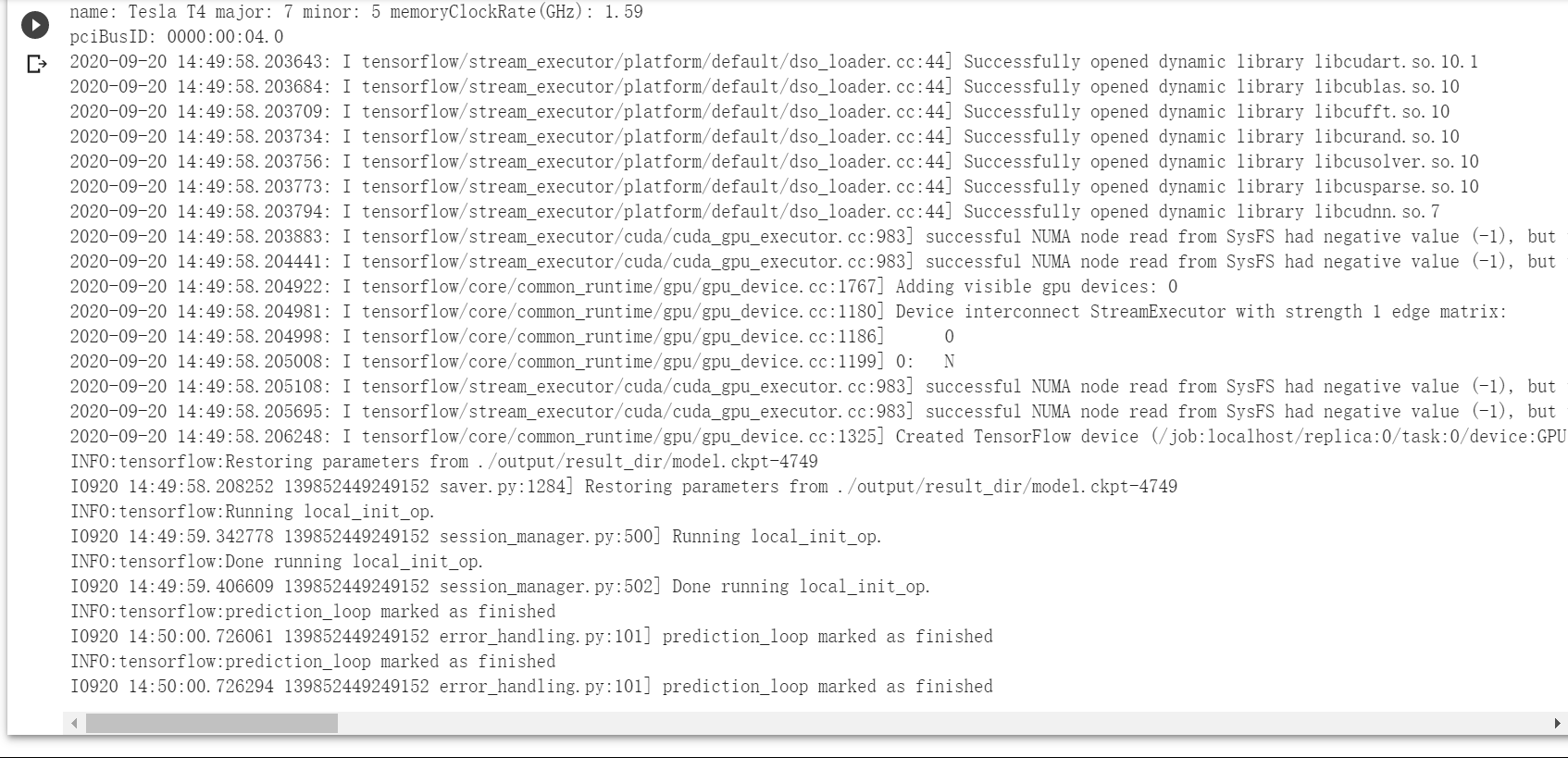


**3.预测**

相比训练模型时，增加三个参数，设置do\_predict=True，do\_train=False.



预测的完成的截图。



预测返回的结果：

|  |  |
| --- | --- |
|  |  |
| 测试集的token | 预测的label |

自动生成常见的评价指标。

