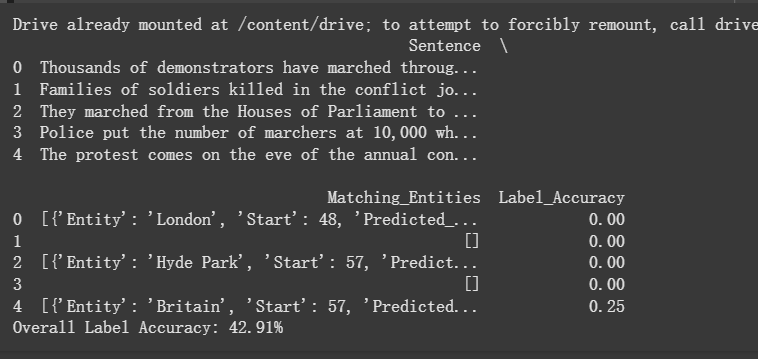
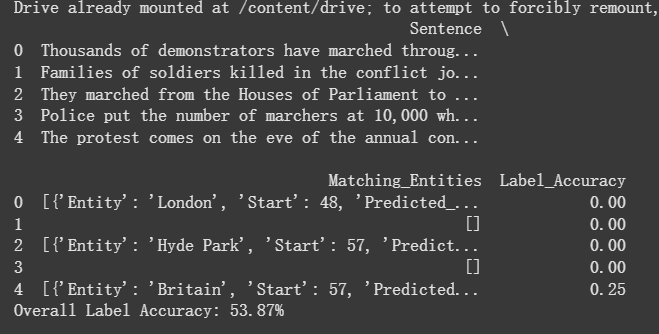
## **1.1探究spacy在数据集0准确度低的原因：**

由于数据集和spacy的标准不一致，所以需要建立一个map，将数据集的entity label转换成spacy的 entity label.但出现了一些问题，比如数据集仅仅对geo和gpe进行了区分，可以分别映射到LOC和GPE。但spacy会将某些地理名词识别为NORP标签(nationality, religion, political group)，难以进行统一映射。

以下是将NORP计入准确度计算（实体加入总样本计数）：



NORP不计入：



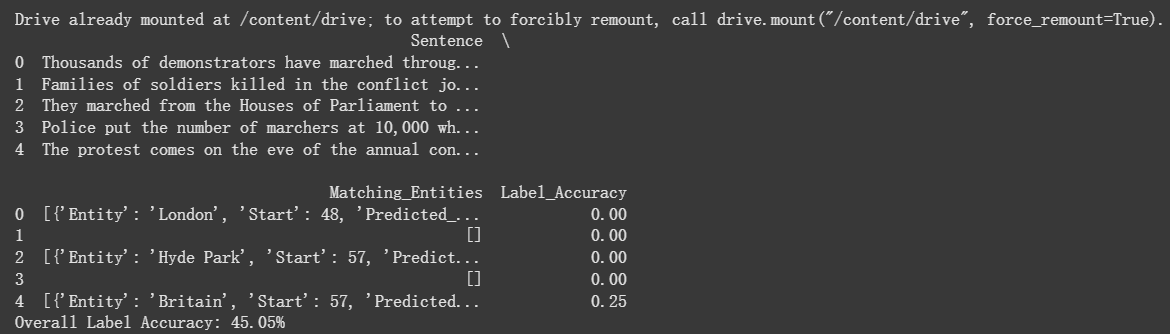
Spacy输出的位置不是token而是字符的位置（计入空格和标点），这和大部分数据集不一致，需要一些额外的处理，在目前的实验中，位置准确度并未得到计算。

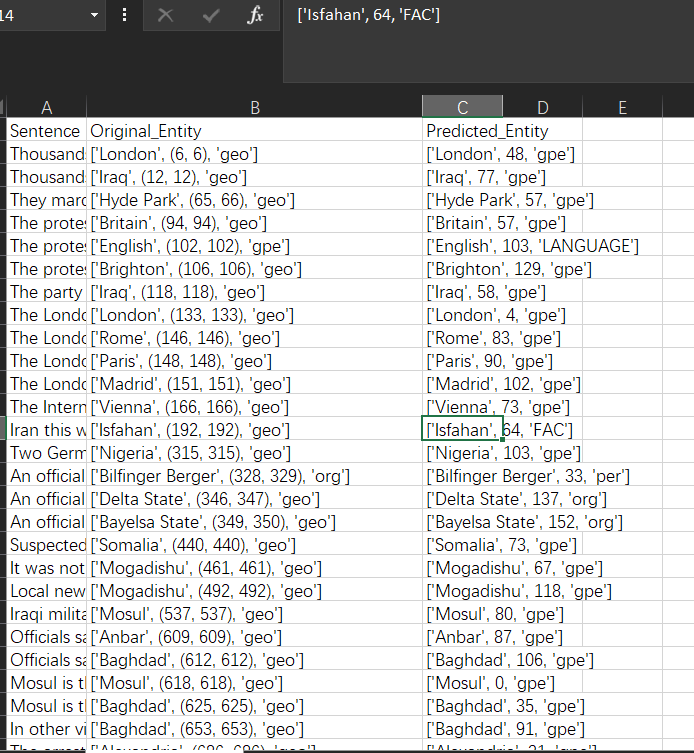
**1.2探究使用预训练管道结合特定数据集进行微调**

截取10000行数据进行训练测试集拆分(0.7:0.3)。在节省时间进行实验的情况下，训练均采用2 epoch.

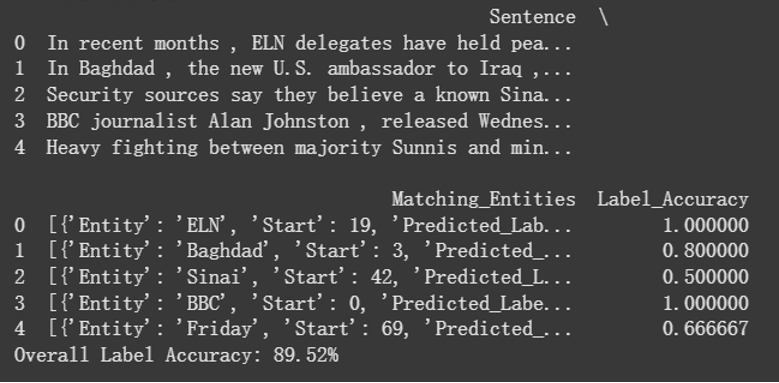
1.2.1使用空白管道进行训练

出现大量geo gpe的识别错误，如图

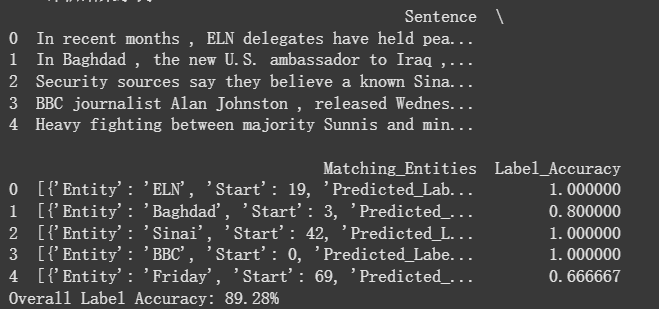


1.2.2载入预训练管道进行训练

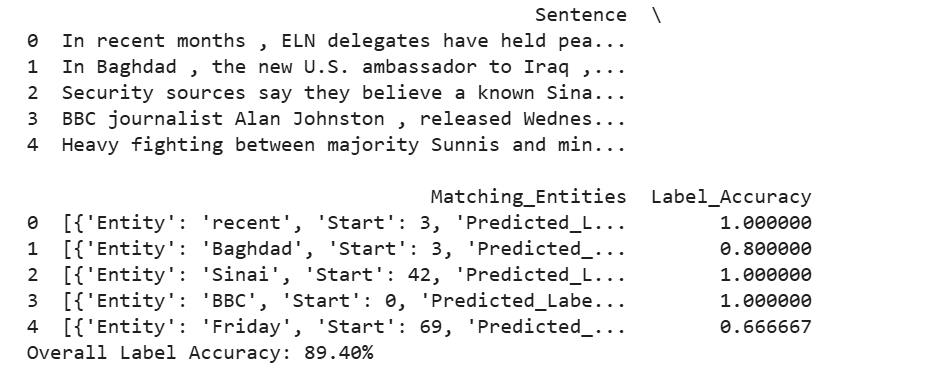
en\_core\_web\_sm



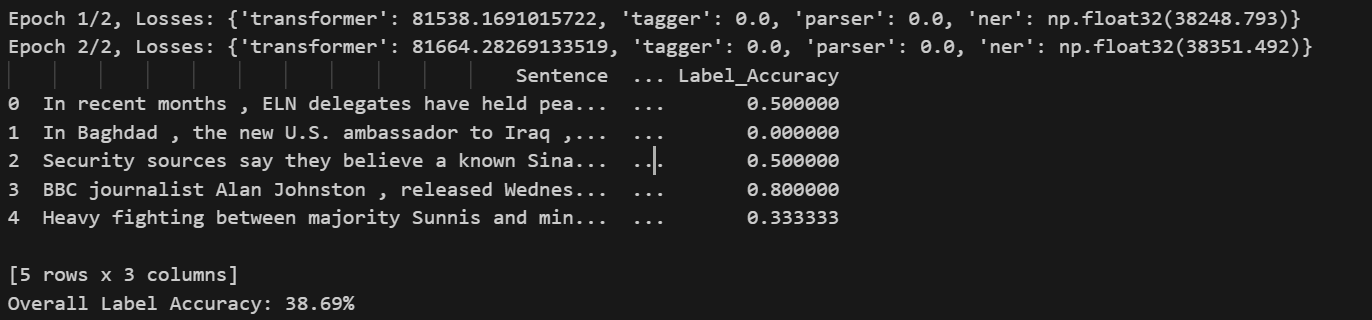
en\_core\_web\_md

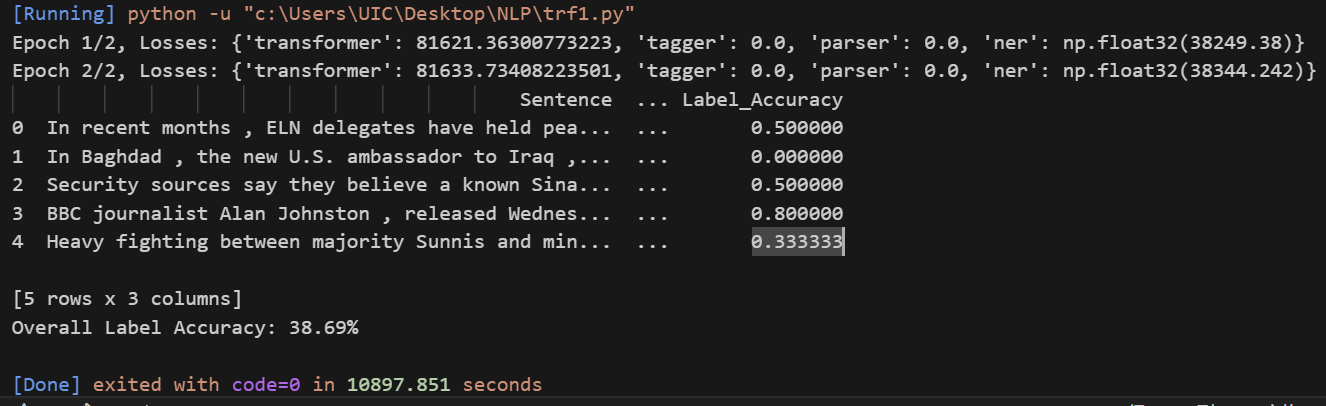


en\_core\_web\_lg

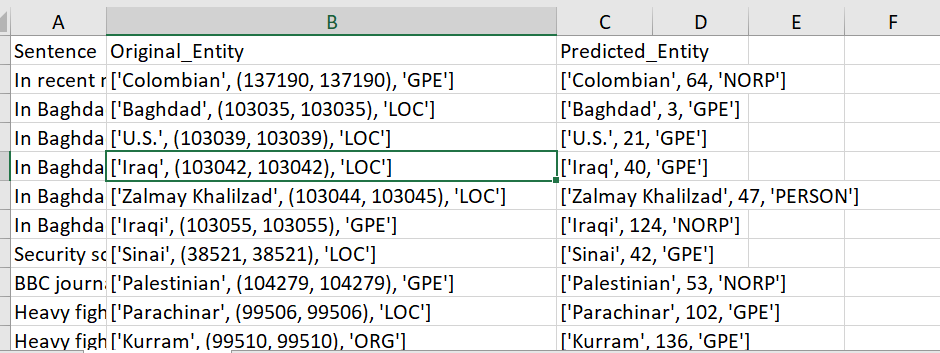


en\_core\_web\_trf





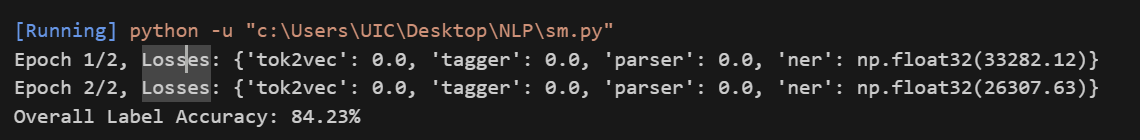
以上两个结果是在重构的trf测试代码和之前其他管道代码仅改写spacy.load()函数中管道的结果。这意味着代码有错误的可能很低，很有可能是trf管道不符合本次的实验条件，根据error\_df

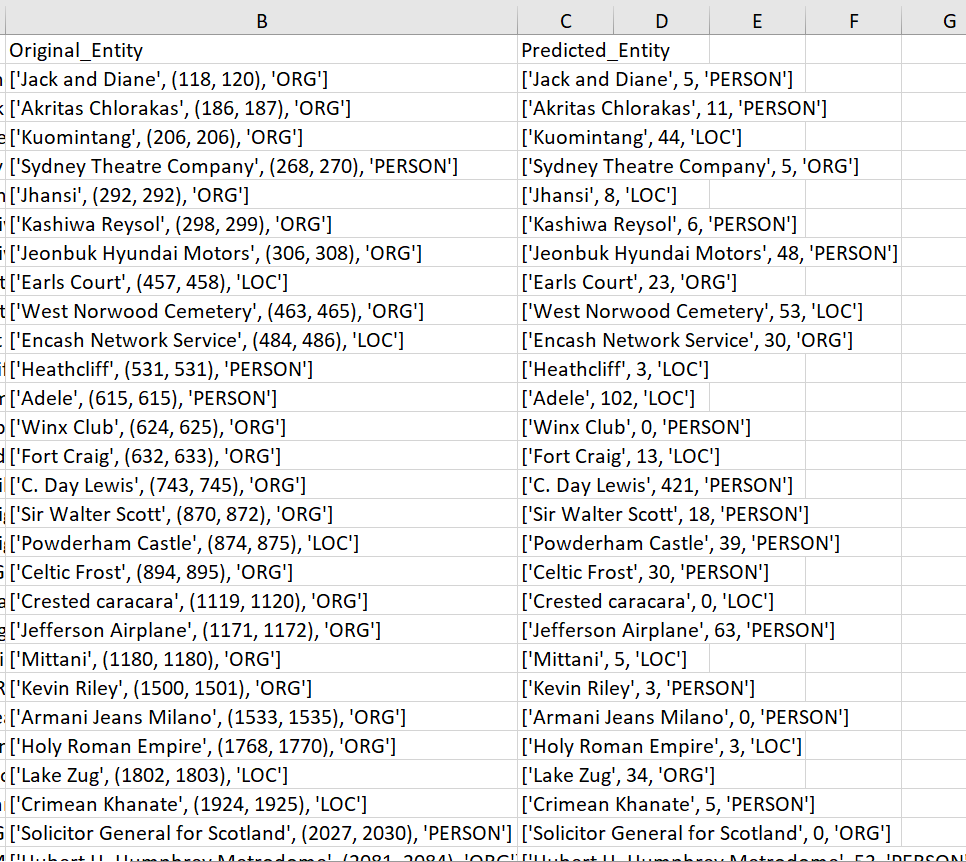
排查错误原因可能为trf的参数太多，训练epoch少和训练量少导致没有完全更新，trf的输出没有遵循要求的规则。

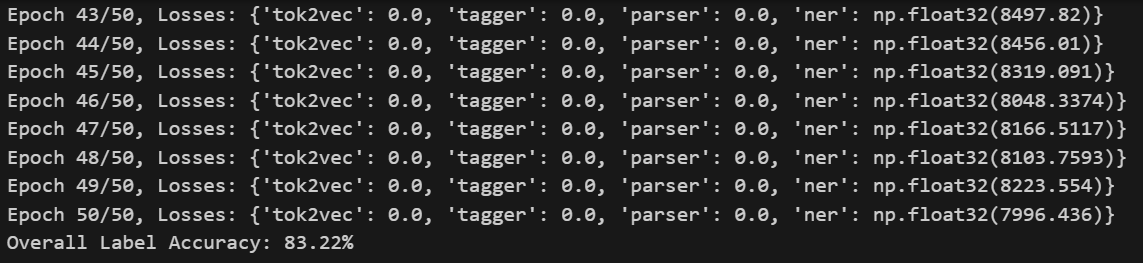
1.2.3 对微调管道进行更深入的研究

使用数据集1进行训练与测试，训练集和测试集使用原始数据集的train.conllu和test.conllu文件格式化而成。

在数据集1(数据标签为PER,LOC和ORG)中spacy的sm管道正确率，（2epouch）





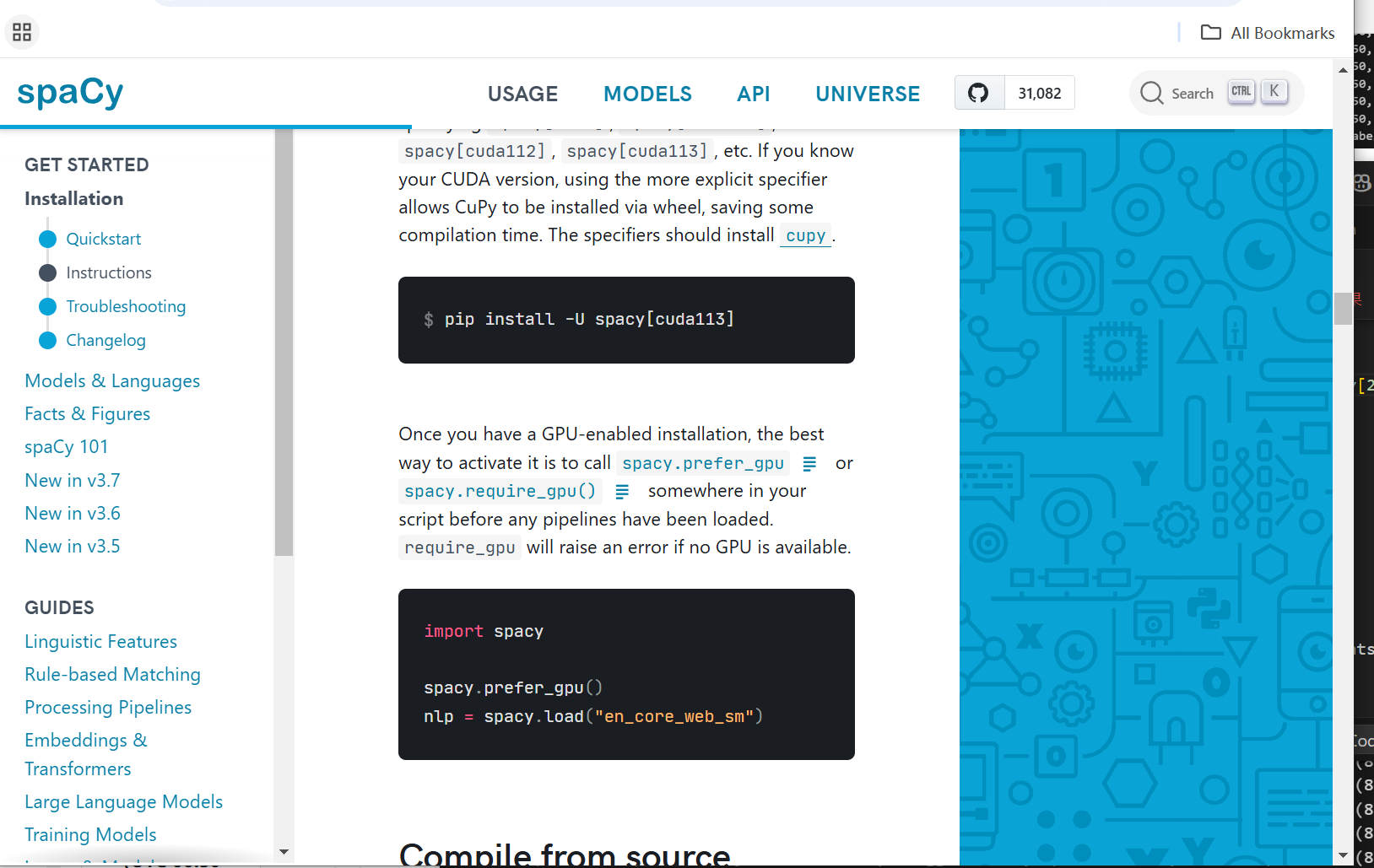
50epoch：

## **2.待解决问题：**

2.1 Deepseek

Deepseek api返回速度过慢，除了极长的思考时间，tokens的生成速度也过慢，约20tokens/s及以下。数据集每行tokens不统一，可能会造成溢出；目前采用比较保守的输入策略，造成测试较慢，同时api的返回不完整，会出现断行的情况。目前采用对api返回的sentence和数据集的sentence的前十个字符进行匹配的策略，依旧处理非常缓慢。

2.2 spacy

Trf管道的训练似乎是进行了封装，不能调用torch的相关cuda加速函数。在训练中出现了设备不统一的问题，主要出现在 spaCy 的 transformer 管道内部，由于该部分代码高度封装且缺乏公开的接口，很难通过外部修改来保证所有数据在训练时都统一在 gpu上。尽管spacy的官方文档中有相关使用gpu进行训练加速的说明

但未能解决问题。使用cpu训练这个管道过于缓慢。