醫病訊息決策與對話語料分析競賽 - 秋季賽：醫病資料去識別化

隊名：ken19980727

隊員：

隊員一 : 高雄科技大學 電機工程系 李佑謙

隊員二 : 高雄科技大學 電機工程系 張祐晨

隊員三 : 高雄科技大學 電機工程系 陳冠穎

隊員四 : 高雄醫學大學 學士後醫學系 王慶雲

指導教授：高雄科技大學 戴鴻傑 副教授

演算法說明：

在2018年，已推出許多出色的遷移學習網路，如Allen AI的ELMO[[1]](#footnote-1)、Open AI的Open-GP[[2]](#footnote-2)，還有我們本次使用的模型──Transformer系列的BERT[[3]](#footnote-3)(Bidirectional Encoder Representations from Transformers)，如圖一、圖二。

BERT是一種理解上下文的語言模型（Language Model, LM），屬於Transformers中的encoder。LM做的事情為給定一些詞彙的前提下，去估計下一個詞彙出現的機率分佈。

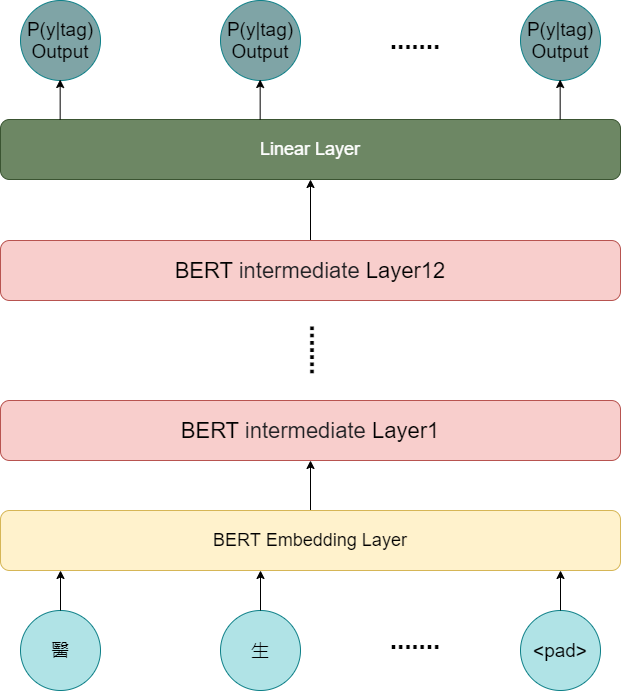


圖 二、BERT模型架構

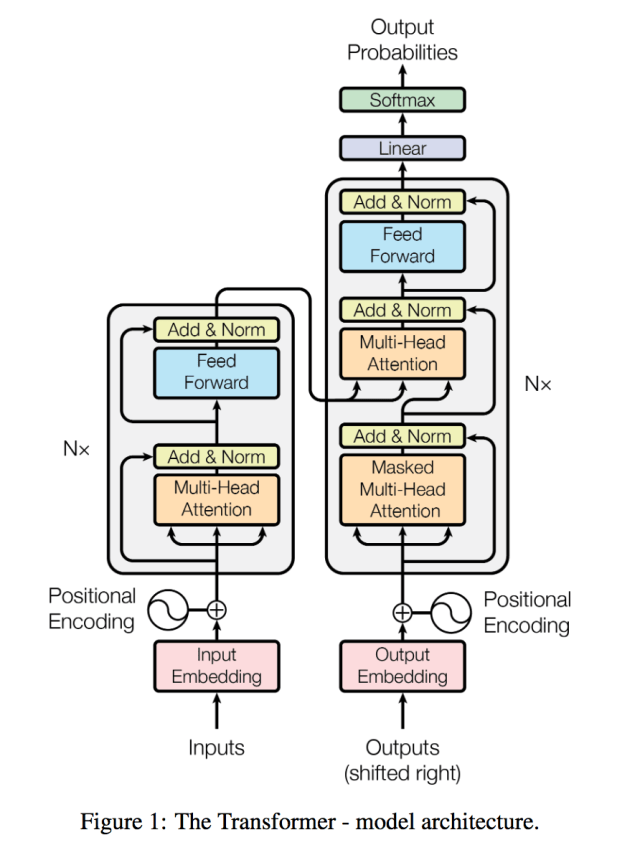


圖 一、Transformer 模型架構

工具說明：

Programing Language: python3.7

Pytorch: 深度學習用

Transformer: 深度學習用

Jason: 儲存Corpus

Pandas: 資料清理、視覺化

Seqeval: 驗證算分用

流程說明 :

我們以BERT作為模型的基礎，再對其指定的任務進行 fine-tune。

由於Transformer系列的模型對於語言的編碼以及資訊都已經有相當好的參數，剩下的就是針對指定的資料集、任務進行參數的微調，而不需要重新訓練，大大減少的使用者在訓練上的時間，並且能得到相當好的成效。

Loss Function部分，Hugging Face的套件會自動把Loss算出來，包含各層的梯度，接著只需用Pytorch的語法，即可達成向後傳播的功能。

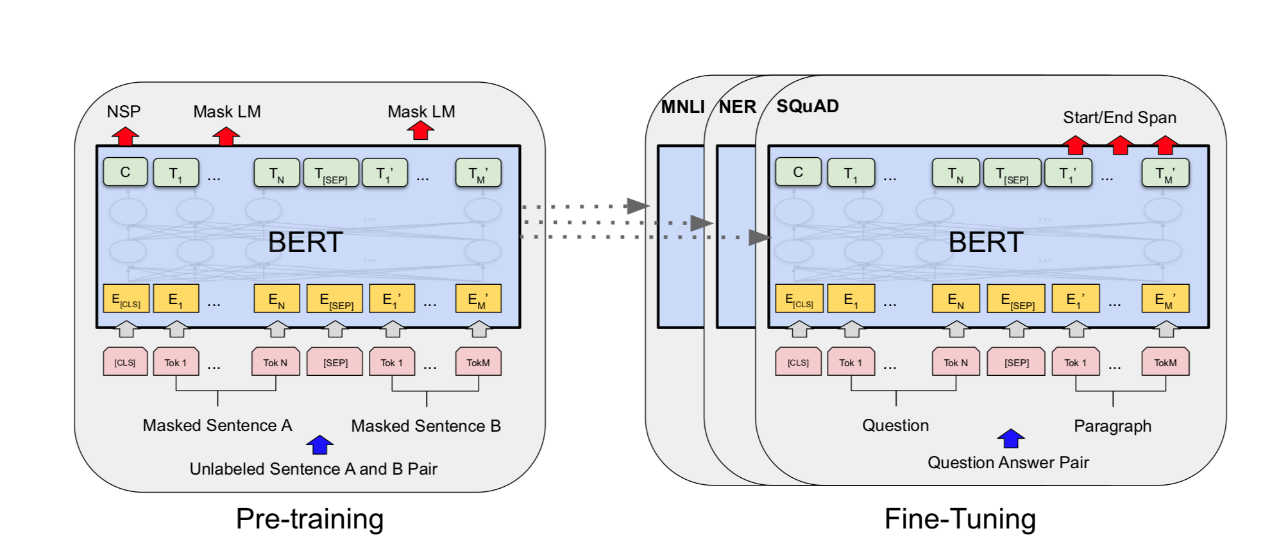
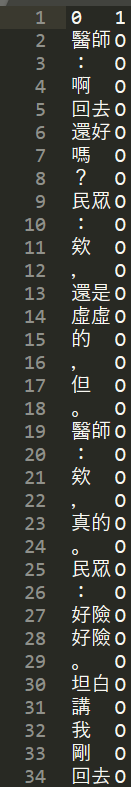
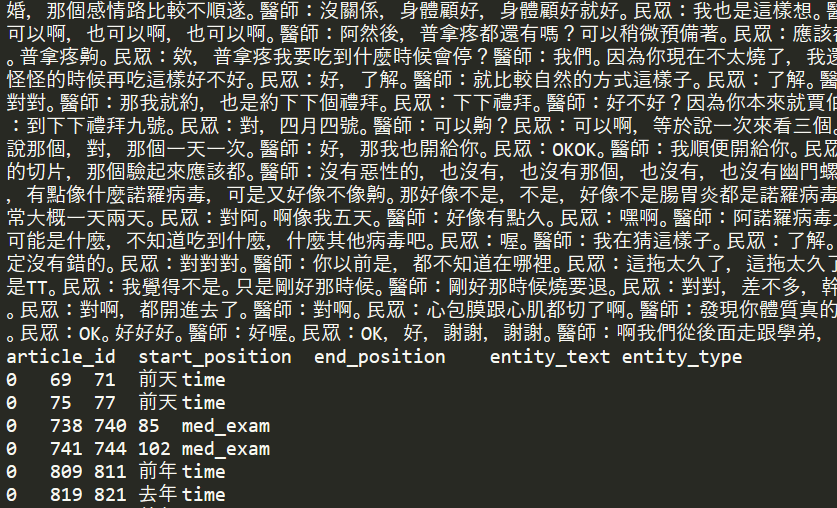
流程圖請看圖三。

圖 三、fine-tune示意圖

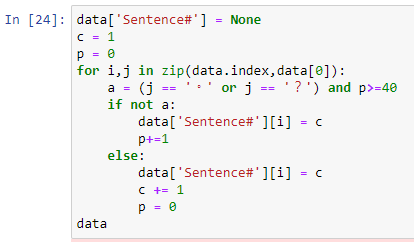
前處理 : 對話轉成NER格式

File Name : baseline.ipynb & Data\_processing\_ex.ipynb

先利用官方提供的baseline.ipynb將資料轉成NER的格式

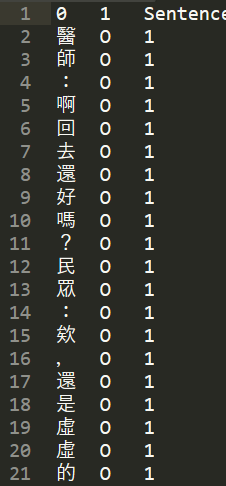


這裡我們使用 Data\_processing\_ex.ipynb 決定每個句子的長度



將每個句子長度控制在40上下 (實測過後這個range對我們的模型效益最好)

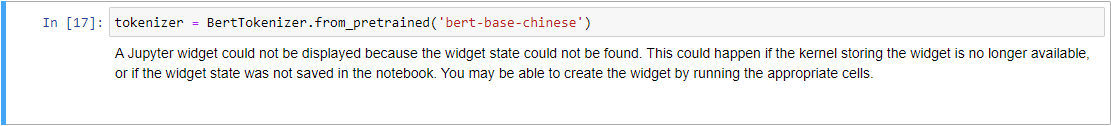
整理完後，文件將呈現成這樣 : (訓練集與測試集都是這樣)



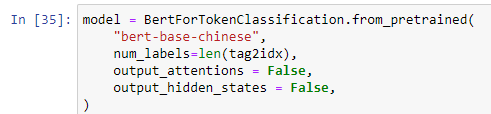
訓練過程 : 訓練NER格式的資料集

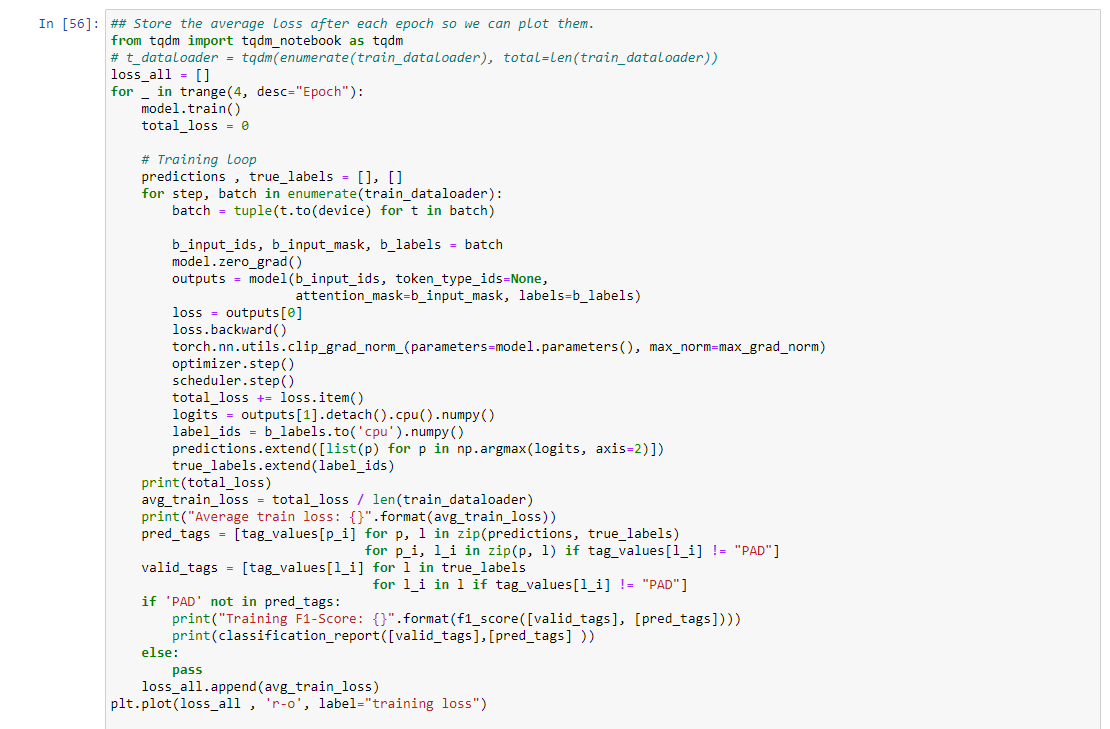
File Name : Bert.ipynb

訓練過程我們使用bert-fine tune，這也是為什麼我們使用字元作為我們的切詞格式(原本我們使用ckip)，因為bert-chinese-base是以字元做為tokenize以及模型的embedding(我們也嘗試過roberta)

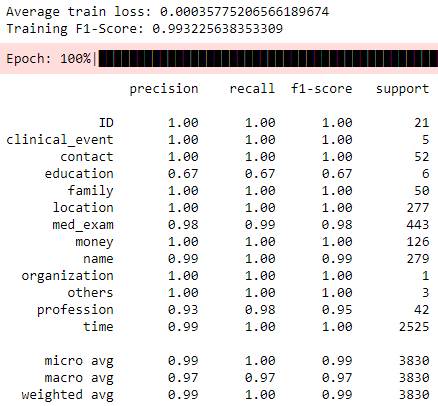






以上為training過程

超參數(最好的那一次)

Epoch : 50

Max\_norm : 2.9

Batch\_size : 8

Sentence\_padding : 500

Optimizer : AdamW

Learning-Rate : 3e-5

訓練過後，我們會做兩件事情，一個是將模型預測測試集的結果輸出，另一個是會檢查在訓練集中的混淆矩陣，而我們發現FN特別多，尤其是time類別的



這時候就需要用後處理將time的FN拉高，不過此舉動會讓FP提高進而導致Precision下降，不過整體來說分數是會上升的

正規表示法後處理 :

File Name : Post\_Process.ipynb

我們會利用正規表示法 : 把我們在檢查FN的地方不足的部分將其補上，例如[月|年]底這種類別，完整的類別可以看Post\_Process.ipynb，補了不少項目，補完後再輸出成Aidea上傳格式

組態說明

* 環境設定
  + CPU: Intel i9-10980XE
  + Ram: 128G
  + GPU: RTX3090
* 組態說明
  + Model: Bert-base-chinese
  + Epochs: 50
  + Optimizer: AdamW
  + Sentence Length (including padding): 500
  + Batch Size: 8
  + Max\_grad\_norm: 2.9
  + Learning Rate: 3e-5

外部資源與參考文獻

<https://www.depends-on-the-definition.com/tags/named-entity-recognition/>

<https://pytorch.org/tutorials/>

<https://github.com/huggingface/transformers>

<https://www.mdeditor.tw/pl/pE92/zh-tw>

1. <https://allennlp.org/elmo> [↑](#footnote-ref-1)
2. <https://github.com/OpenGP/OpenGP> [↑](#footnote-ref-2)
3. <https://github.com/google-research/bert> [↑](#footnote-ref-3)