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
最終報告

一、簡介

隊名：
我們是天使吧

隊員：
郭宇雋（政治大學 資訊管理學系）
江岳憫（政治大學 資訊管理學系）
洪忻柔（政治大學 資訊管理學系）

指導教授：
黃瀚萱（政治大學 資訊科學系）

二、演算法說明

BERT Embedding + 2-Layer BiLSTM + CRF + Rule Based Filters

三、工具說明

語言：
Python3

套件：
1) 資料處理
Numpy
Pandas
Matplotlib
Seaborn
…

2) 機器學習

Sklearn
Pytorch
Keras
Transformers
…

四、流程說明

Step1. 以 BERT Vector 做為文章中每個字元的 Input Feature

Step2. 搭建並訓練 2-Layer BiLSTM 進行 NER 預測任務

Step3. 將 2-Layer BiLSTM 的 Output 作為進入下一階段模型的特徵

Step4. 搭建 CRF Model，此模型的 Input Feature 為上一階段每個字元的 Output Vector，加上各字元的 Character, Bigram 與 Trigram 特徵，做為 CRF Model 的 Input

Step5. CRF 模型得到的結果，會再透過 Rule based 演算法進行進一步過濾與優化，其中包含透過外部中文 NER 套件進行 NER 任務，抓出如時間等常見的特徵標籤；以及透過正則表達式，抓出某些格式固定的類別標籤等

五、組態說明

開發環境：
Python3

LSTM Model:
2-Layer BiLSTM with 128 units in each layer
Drop out rate = 0.2 in each layer
Recurrent drop out rate = 0.2 in each layer
Optimizer use *Adam* with learning rate = 6e-4
Training Epochs = 25, batch size = 16

CRF Model:
Algorithm = lbfgs
c1 = 0.1
c2 = 0.1
Max Iterations = 100

六、外部資源與參考文獻

外部資源：
Transformers
CKIP

參考文獻：
1) 工業界如何解決 NER 問題
http://blog.itpub.net/69946223/viewspace-2707407/