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

1. 隊名、隊員(學校/科系)、指導教授(學校/科系)

隊名：NCUEE

隊員：國立中央大學/電機工程學系 陳昌浩、陳柏翰、鄭少鈞

指導教授：國立中央大學/電機工程學系 李龍豪 助理教授

2. 演算法說明

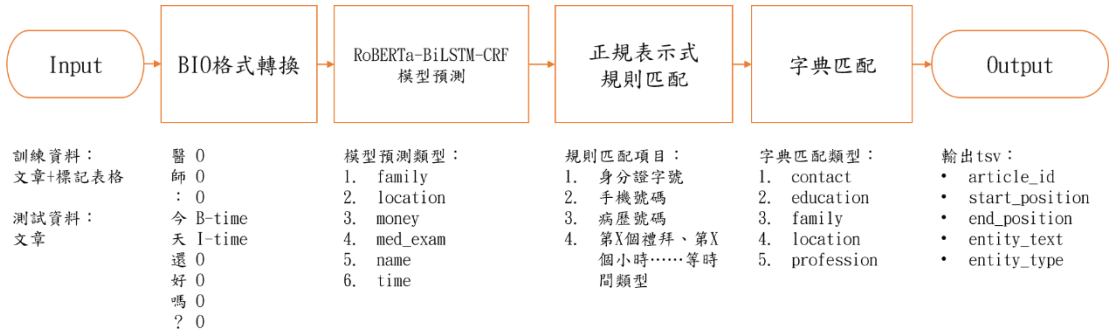

圖 1: NCUEE 系統流程圖

圖 1 為系統流程圖，對應的模組說明如下：

(1) RoBERTa-BiLSTM-CRF 模型預測

BERT 模型(Devlin et al., 2018)在各個自然語言處理任務中展現 state-of-art 的效果，這次使用的模為 RoBERTa-wwm-ext (Robustly optimized BERT approach using whole word masking and external data)。RoBERTa (Liu et al. 2019) 由 Facebook 和華盛頓大學於 2019 年 7 月發表，發表於 arxiv。作者認為 BERT 實際上是 undertrained，充分訓練以後可以反超之後發布的其他模型，在 GLUE、RACE、SQuAD 等指標性數據集上重新取得 state-of-art 的效果。從模型上來說，RoBERTa 基本沒有什麼太大創新，主要是在 BERT 基礎上做了幾點調整： 1）訓練時間更長，batch size 更大，訓練數據更多；2）移除了 next sentence prediction loss ； 3）max sequence length 訓練序列更長；4）動態調整 Masking 機制；5) larger Byte-level BPE 介於字級別和詞級別之間的編碼。

選用的模型 chinese-RoBERTa-wwm-ext-large 中使用的 Whole Word Masking (Cui et al., 2019) 是 google 在 2019 年 5 月 31 日發布的某一 BERT 的升級版本，主要更改了原預訓練階段的訓練樣本生成策略。Word Piece 的分詞方式會把一個

完整的詞切分成幾個子詞，在生成訓練樣本時，這些被分開的子詞會隨機被遮罩。在全詞 Mask 中，如果一個完整的詞的部分 Word Piece 子詞被遮罩，則同屬該詞的其他部分也會被遮罩，即全詞 Mask。

而 Google 官方發布的 BERT-base, Chinese 中，中文以字為粒度進行切分，沒有考慮到傳統 NLP 中的中文分詞（CWS）。RoBERTa-wwm-ext-large 將全詞 Mask 的方法應用在了中文中，使用了中文維基百科全書（包括簡體和繁體）進行訓練，並且使用了哈工大 LTP 作為分詞工具，即對組成同一個詞的漢字全部進行遮罩，作為抽取特徵用。

對於輸入的自然語言序列，可通過特徵工程的方法定義序列字元特徵，如詞性特徵、前後詞等，將其輸入模型。但現在多數情況下，可以直接選擇句中每個字元的字嵌入或詞嵌入向量，可以是事先訓練好的或是隨機初始化。對於中文，我個人傾向於將字元向量和其所屬的詞向量進行拼接，詞嵌入使用預訓練好的，字嵌入隨機初始化。

RoBERTa-BiLSTM-CRF 模型架構圖如圖 2。我們利用 RoBERTa 進行特徵抽取，將 RoBERTa 所輸出的向量，當作 BiLSTM 的輸入序列。雙向長短期記憶神經網路 (Bidirectional Long Short-Term Memory, BiLSTM)是由前 向 LSTM 與後向 LSTM 組合而成，適合做上下有關係的序列標註任務，因此在 NLP 中 常被用來建模上下文資訊。命名實體辨識屬於序列標記的多分類問題，傳統上在遇到多分類問題時，會採用 Softmax function 作為輸出函數，但在實際情況時，序列標註任務中的當前時刻的狀態， 均與當前時刻的前後狀態有所關連，因此條件隨機場域 (Condition Random Fields)取代了 Softmax function，成為了當前主流的架構。而目前較為常見的標記格式包 含 BIO 格式以及 BIOES 格式，在進行實體辨識時，正確的標記序列中標記 O 後面是不會接連著標記 I，因此在輸出層中採用條件 隨機場域 (Condition Random Fields, CRF)做為輸出層，以確保預測的標記是合理的。

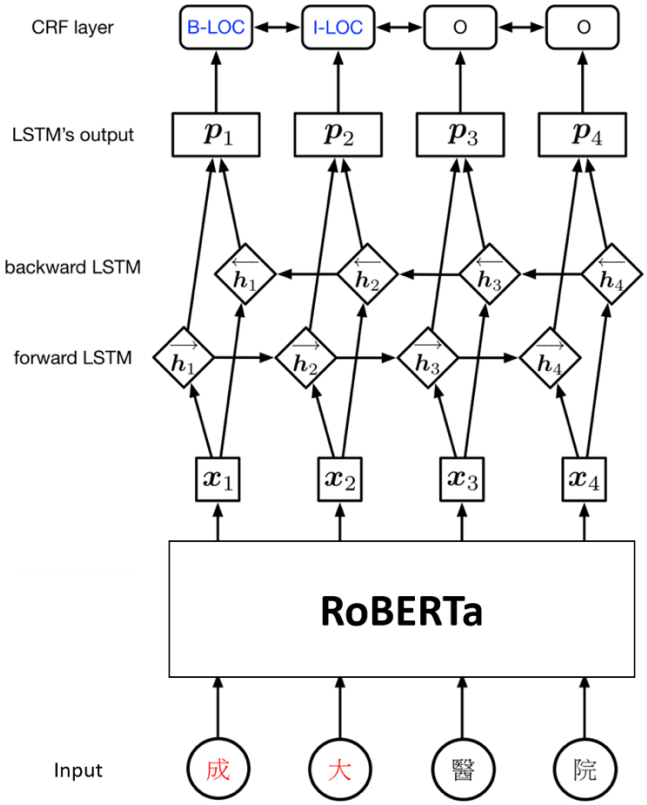

圖 2: RoBERTa-BiLSTM-CRF 模型架構圖

(2) 正規表示式規則匹配

正規表示式（英語：Regular Expression，常簡寫為 regex、regexp 或 RE），又稱正規表達式、正規表示法、規則運算式、常規表示法，是電腦科學的一個概念。正規表示式使用單個字串來描述、符合一系列符合某個句法規則的字串。在很多文字編輯器裡，正則表達式通常被用來檢索、替換那些符合某個模式的文字。

根據訓練資料上出現的標註文字發現，身份證字號應標為 ID、手機號碼應標為 contact，而這些屬於有規則的文字序列因此可以利用規則式將此種資料作標註。

(3) 字典匹配

字典匹配使用單純的字串比對。字典來源為訓練資料及網路，將一些專屬於該類特殊的詞彙使用字典匹配可以增進去識別化的標註效果。

3. 工具說明

- Python version 3.6
- tensorfow-gpu version 1.14
- BERT-BiLSTM-CRF-NER
https://github.com/swy0915/BERT-BiLSTM-CRF-NER/tree/master/BERT-BiLSTM-CRF-NER-master
- Chinese-BERT-wwm
https://github.com/ymcui/Chinese-BERT-wwm
- BERT
https://github.com/google-research/bert

4. 流程說明

Steps 1 資料轉換成 BIO 格式

Steps 2 BERT-BiLSTM-CRF 做六類命名實體(Family, money, name, med_exam, location, time) 模型預測

Steps3 模型輸出測試資料 BIO 答案

Steps4 規則匹配 (ID, contact, time)

Steps5 字典匹配 (family, profession, contact, education, location)

Steps6 Tsv 格式轉換

5. 組態說明 (e.g.環境設定、參數設定)

RoBERTa-BiLSTM-CRF 參數設定：

- init_checkpoint = chinese_roberta_wwm_large_ext_L-24_H-1024_A-16
- max_seq_length = 128
- train_batch_size = 32
- learning_rate = 2e-5
- num_train_epochs = 50
- droupout_rate = 0.4

6. 外部資源與參考文獻

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, Guoping Hu. Pre-Training with Whole Word Masking for Chinese BERT. arXiv:1906.08101

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805