Assignment 3 Named Entity Recognition on MSRA corpus

108502571 資工三 B 楊佳峻

1 Introduction

實體辨識(NER)任務是 Information Extraction 中很典型的任務，目的是要將文章中的專有名詞判斷出來。

2 Related work

相近於這次所用的 CRF 方法，還有 ME、MEMM、HMM、SVM 等等。若要辨識中文的工具，已經有人製作 HanLP、CRF++，可以幫助 IE 任務快速執行。

3 Method

由於要自己訓練一個可辨識的 NER 模型，我選擇使用 BERT BiLSTM CRF 模型，運用了 LSTM 記憶儲存先前資訊，並可關注前後項字詞的影響，BERT 部分則是使用 TF HUB 的預訓練模型，後續只需要針對 IOB 標籤的部分進行訓練就行。

4 Result

以下經由 conlleval.pl 評分，顯示自己訓練的模型能有較高較能。

4.1 My BERT BiLSTM CRF Model

F1 score: 86.16

```
processed 193754 tokens with 6208 phrases; found: 7471 phrases; correct: 5893.
accuracy:  98.01%; precision:  78.88%; recall:  94.93%; FB1:  86.16
              LOC: precision:  81.30%; recall:  94.57%; FB1:  87.44  3364
              ORG: precision:  62.84%; recall:  93.62%; FB1:  75.20  1986
           PERSON: precision:  90.05%; recall:  96.32%; FB1:  93.08  2121
```

4.2 CRF++

F1 score: 81.53

```
processed 19172 tokens with 9715 phrases; found: 9308 phrases; correct: 7755.
accuracy:  88.95%; precision:  83.32%; recall:  79.83%; FB1:  81.53
              ADJP: precision:  58.11%; recall:  25.00%; FB1:  34.96  74
              ADVP: precision:  65.44%; recall:  53.29%; FB1:  58.75  272
             CONJP: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
              INTJ: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
                NP: precision:  81.50%; recall:  79.54%; FB1:  80.51  4957
                PP: precision:  88.96%; recall:  95.27%; FB1:  92.01  2129
               PRT: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
              SBAR: precision:  84.62%; recall:  34.38%; FB1:  48.89  78
                VP: precision:  85.32%; recall:  80.61%; FB1:  82.90  1798
```

5 Conclusion

以 BERT 為原型的模型通常都可以達到較高的效能，不過缺點就是需要較多的運算時間與效能。另外 NER 任務也有可能會針對不同領域文章會有不同效能，所以需要看目標來使用需要有多少精度、什麼面向的訓練資料來幫助模型符合你的目標。

6 Reference

https://zhuanlan.zhihu.com/p/156914795