AI Cup 期末專題報告

1. 問題、分析方法與模型介紹

(1)問題：識別出的隱私類型，如 name、location、time、family 等。

(2)分析方法：

a. 詞分割分析：對於句子切成詞的方式去做分析。

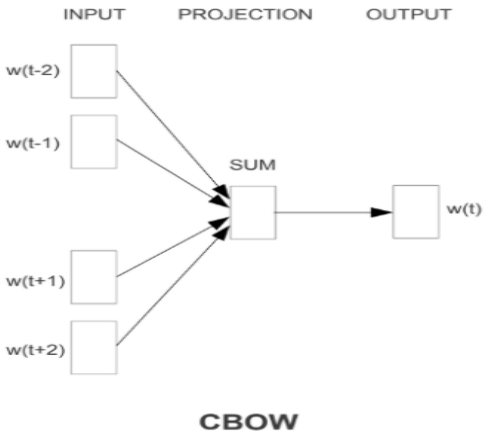
b. 資料清理：有發現資料有非常多無意義的詞如齁、啦等語助詞，將其刪除以做清洗。

c. 模型參數調整：利用 earlystopping 去找該參數下的模型狀態。

d. 模型種類選取：本專題只用到二種模型。

(3)用過二種模型：word2vec+jieba+CRF 和 bert-bilstm-crf-ner

a. 第一種模型是用 word2vec 的 CBOW 與 jieba 去做前處理，之後再用 CRF 去做 train。



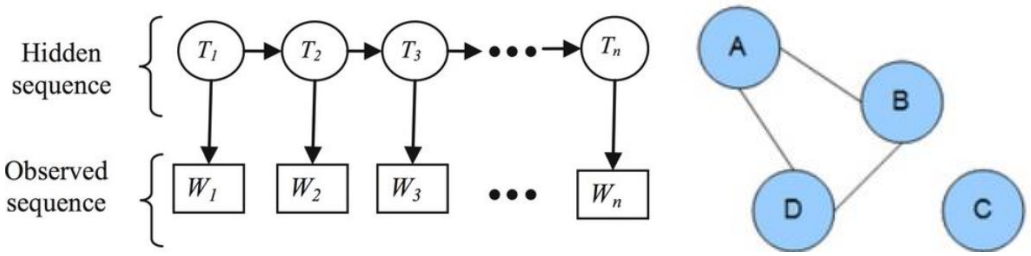
「我們 去 六福村 玩 雲霄飛車 好不好」
「我們 去 麗寶樂園 玩 雲霄飛車 好不好」
「我們 去 九族文化村 玩 雲霄飛車 好不好」

CBOW 是給定上下文，來預測輸入的字詞。若以右圖為例，就是將六福村、麗寶樂園和雲霄飛車歸類為同一類，也就是只要上下文中出現

「去」和「玩」，有極大機率中間填入的詞為六福村、麗寶樂園和雲霄飛車。

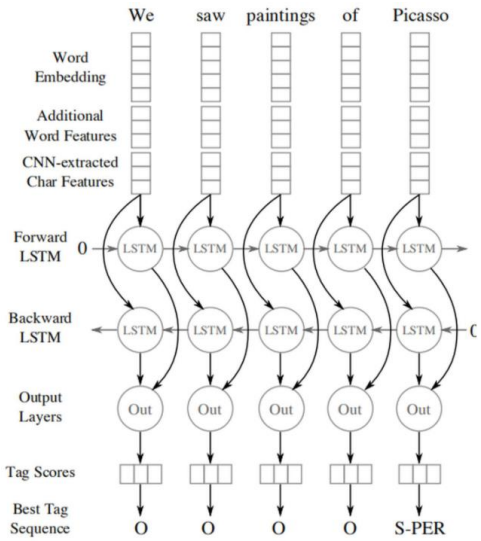
['醫師', ':', '那', '我們', '還是', '有', '，', '就', '幫', '我', '一下', '這樣', '好不好', '？',

另外 jieba，就是將句子中的詞做分割，以達到能訓練 word2vec 的 training data。



還有 CRF 採用的是馬爾科夫隨機場，它是一個無向圖模型。無向圖中的(如右圖)團的界定，在 CRF 中的作用巨大。不像左圖的 HMM 是用某一個觀察值僅僅依賴於對應節點的 weight 變化，而 CRF 將衡量隨機變數之間的相關關係，然後簡化條件概率模型。以此作業來說就是以隨機變數去找詞之間的關聯性。

b. 第二種模型是用 bert-bilstm-crf-ner



此模型分成 3 個部分，先經過 bert 的 pretrained 語言模型獲得詞向量，再將其輸入到 bilstm 去做處理，最後用 CRF 去做預測。

2. 實驗環境與結果 (含改進過程與最後分數)

(1) baseline code

由於手邊有的環境是在 google colaboratory 上跑程式，因此 RAM 最大為 25GB，此 512 維的 cna.cbow.cwe_p.tar_g.512d.0.txt 則太大直接讓 RAM 爆掉。

結果：F1_score = 0

分析：接下來的實作皆會以如何能在 colab 上跑去做努力。

(2) Word2vec 的降維

直接拿 512 維的 cna.cbow.cwe_p.tar_g.512d.0.txt 裡頭的向量資料直接降維，希望以此能達到與 512 維 model 相同學習的能力。

報導	-0.746177	0.598292	-1.730891	-0.250867	-1.273854	-0.258893	0.146908
地區	1.280977	-0.575436	-0.246468	1.419103	1.032942	-0.261977	-0.823605
仍	-0.737807	-0.431719	1.832474	-0.032166	0.851099	-0.499760	0.658231

結果：F1_score = 0.001

分析：由於此模型是用上下文關係去找詞向量關係，因此像我用詞去 train 是沒有意義的。

(3) 自己訓練 word2vec

為了解決 RAM 爆掉的問題，我使用的方法是自己訓練 word2vec，並嘗試過的維度有 200、256 和 400，training data 則是從醫學的報章

雜誌去找對話來訓練，藉此提高模型對於醫病資訊對話的強度。

結果：F1_score = 0.3345

分析：我認為問題有兩個，一個是找的報章雜誌大多雖然也是以醫療為主，但大部分都不像本專題是醫療診斷的對話。另一個問題是 training data 中的'O'label 太多了，以至於模型會一直誤認為是此 label。

(4) bert-bilstm-crf-ner 模型使用

裝不起來，因此沒有使用到。

分析：只好用有限的方式去 train

(5) 將 training data 中的'O'label 全部刪掉

用不要判斷這麼多 label 的方式讓他去 train，效果還不錯。

結果：F1_score = 0.4836

分析：確實解決 overfitting 的問題，而準確度也有一部份的提升。

3. Dataset bug report

(1) Label 並沒有全部用上，實際上 model 並不用判斷這麼多 label

account
belonging_mark
biomarker
special_skills
unique_treatment
others

這幾個都沒有用上

4. 心得

這次專題對於沒碰過 ner 的我來說，的確有點吃力，但至少有計畫的方式

去 train 是有辦法將結果提升的。雖然常常看到自己的基本功不佳，但只要看到人家的方法，一直去做嘗試，終究會是有收穫的。另外會希望之後如果組員退光的話，能不能並到其他組去，自己一個人做還是會超吃力的。

5. 資料來源：

(1)An Introduction to Conditional Random Fields：

https://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf

(2)Efficient Estimation of Word Representations in Vector Space：

https://arxiv.org/pdf/1301.3781.pdf