命名實體識別：結合預訓練模型及對抗訓練的解決方案
CrowNER at Rocling 2022 Shared Task: NER using MacBERT and Adversarial Training

張秋霞 **Qiu-Xia Zhang***
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
r10922164@ntu.edu.tw

戚得郁 **Te-Yu Chi***
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
d09922009@ntu.edu.tw

楊德倫 **Te-Lun Yang***
國立政治大學資訊科學系
Department of Computer Science
National Chengchi University
111971029@nccu.edu.tw

張智星 **Jyh-Shing Roger Jang**
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
jang@mirlab.org

摘要

本研究使用「ROCLING 2022 中文健康照護命名實體辨識任務」(Lee et al., 2022) 中的訓練與驗證資料來進行建模。建模過程採用了資料擴增與資料後處理等技術，並使用 MacBERT 預訓練模型來建立一個專用中文醫療領域的 NER 辨識器。在微調過程中，我們也加入對抗式訓練的方法，如 FGM 和 PGD，最後調適所得的模型成效接近於任務評測最佳團隊。此外，藉由引入混合精度訓練，我們也大幅降低了訓練所需時間成本。

Abstract

This study uses training and validation data from the "ROCLING 2022 Chinese Health Care Named Entity Recognition Task" for modeling. The modeling process adopts technologies such as data augmentation and data post-processing, and uses the MacBERT pre-training model to build a dedicated Chinese medical field NER recognizer. During the fine-tuning process, we also added adversarial training methods, such as FGM and PGD, and the results of the final tuned model were close to the best team for task evaluation. In addition, by introducing mixed-precision training, we also greatly reduce the time cost of training.

關鍵字： MacBERT、條件隨機場域、命名實體辨識、對抗訓練

Keywords: MacBERT, Conditional Random Field, Name Entity Recognition, Adversarial Training

1 緒論

自然語言處理 (Natural Language Processing, NLP) 的持續發展，使機器逐漸能夠以人類大腦思考的方式來理解與解析語意，降低人類與

*These authors contributed equally to this work.

機器之間溝通的鴻溝，將人類常用的語言轉換成機器可以理解的格式，藉以進行文字上的分類、預測、推論等與自然語言理解 (Natural Language Understanding, NLU) 相關的任務。自然語言理解與語言學 (Linguistics) 有著密不可分的關係，它逐漸發展成包括人工智慧、計算機科學等領域的一門學科 (Bates, 1995)。近年來，隨著神經網路與機器學習技術的進步，以及網際網路上大量文字語料的取得，自然語言理解相關的理論與實務操作，得到了廣泛的應用。

機器在簡單閱讀理解任務上的表現，已經可以逐漸接近 (Rajpurkar et al., 2016) 甚至超越人類 (Yu et al., 2018)，然而在眞實世界的應用上，卻還是有較大的效能差距 (Zheng et al., 2019)，會有這樣的差異，在於人類具有了解實際情況並且作出回應的能力，此能力易於將閱讀得到的文字資訊，自動地建立關聯，並賦予意義，雖然機器能夠將非結構性的文字透過斷詞技術 (word segmentation)，將不同的文句切割成字詞，但字詞之間並沒有辦法直接建立有意義的關聯，於是需要透過系統性的標註方式 (labelling)，讓機器理解上下文、段落、文句和字詞之類的關係，將重要的資訊提取出來，進而得到不同領域的知識，例如閱讀一則衛教 (Health Education) 文章，文章中會提及哪些人 (Who) 可能會得到什麼樣的疾病 (What)，通常這些疾病好發於什麼時間 (When) 如季節、月份等，以及爲什麼會得到這些疾病 (Why) 和如何治療 (How)，這些標註可以幫助機器更好地理解字詞之間的關係、順序和意義，掌握字詞的特徵，以便於了解文章整體的重點，這種資訊擷取 (Information Extraction) 的方法，稱之爲命名實體辨識 (Name Entity Recognition, NER)，是自然語言處理的基本任務之一。

本研究使用 ROCLING 2022 中文健康照護命名實體辨識 (Chinese Healthcare Named Entity Recognition) 任務 (Lee et al., 2022) 所提供的訓練與驗證資料，結合資料擴增 (Data Augmentation) 與資料後處理 (Post-Processing)，以 BERT 爲基礎的預訓練語言模型 MacBERT 進行微調訓練 (Fine-tune)，在評估語言模型的成效以後，計算出 F1-score 的結果爲 0.7796。而後，在既有的語言模型上加入了條件隨機場域 (Conditional Random Field, CRF) 等模型提升方法，計算出 F1-score 的結果爲 0.8076，提升了 2.8% 的效能，與任務評測最佳的系統，有著類似的成效與水平，並且藉由引入混合精度訓練，大幅減少訓練所需時間成本。

下一章將簡要地進行文獻回顧，第三章說明本研究所執行的步驟，包括採用的神經網路架構、資料處理的方式，第四章呈現系統展示的結果，並說明改善的方法。

2 相關研究

©uo ?N | ¦

Bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) 是由 Google 於 2018 年提出的 NLP 預訓練模型，用以解決 NLP 在各種領域下游所遭遇的問題，例如：意圖分類問題、情緒分類問題、前後文預測問題等。

BERT 基於 Transformer (Vaswani et al., 2017) 的 Encoder 作爲其雙向訓練的架構；由於相同的詞 (word) 在上下文 (context) 中可能表示不同意義，BERT 將詞轉換爲 contextual word embedding，投射到一向量空間以表示其特徵並作爲 Transformer 的輸入；爲了讓文字的序列是有意義的，BERT 同時將 Position Encoding 作爲輸入傳入 Transformer 中。Transformer 的 Multi-head attention 則是在透過 self-attention 進行平行運算以獲得每個詞的上下關係。Transformer 的架構如圖 1。

BERT 訓練分爲兩個階段，分別爲預訓練 (Pre-training) 及微調 (Fine-tuning)。預訓練所使用的語料庫由 BooksCorpus (800M) 及英文維基百科 (2,500M) (僅取文字內容部分) 所擷取的詞所組成。預訓練共有兩個任務，分別爲 Masked LM (MLM) 及 Next Sentence Prediction (NSP)。MLM 任務藉由隨機遮蔽 15% 的詞 (替換爲 [MASK] 標籤) 並進行 Mask 的值預測；NSP 任務則是輸入兩個句子，藉由 [CLS] 標籤置於句首以識別進行分類，並在語句之間放置 [SEP] 表示斷句以進行上下文的預測。

MacBERT (Cui et al., 2021) 延伸自 BERT，主要優化的部分在於 BERT 在預訓練的 MLM 任務隨機將詞替換爲 [MASK]，然而實際上 [MASK] 並不出現於下游任務，MacBERT 將 MLM 任務更換爲 MLM as correction 任務，基於 word2vec 演算法計算詞的相似度，藉由相似詞取代 [MASK]，同時引入 Whole Word Mask (WWM) 及 N-Gram masking 技術，針對需要對 N-Gram 進行 Mask 時進行相似詞的查找替換，若無相似詞則使用隨機詞進行替換。另外 MacBERT 相較 BERT 有一大優勢即 MacBERT 的預訓練語料爲中文，可解決 BERT 無法應用於中文分類的缺陷。本研究最終使用 MacBERT 作爲主要的訓練模型。

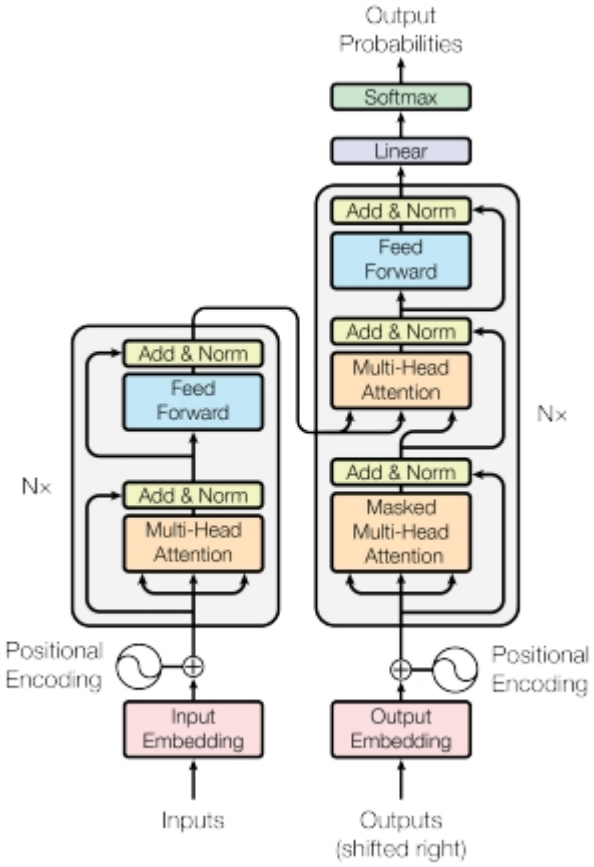

圖 1. Transformer 架構.

條件隨機場域 (Conditional Random Field, CRF) 是一種圖形結構的機率模型，用於分割與標註序列資料 (Lafferty et al., 2001)，成效優於傳統的隱形馬可夫模型 (Rabiner and Juang, 1986) 以及最大化熵馬可夫模型 (McCallum et al., 2000)。

隱形馬可夫模型 (Hidden Markov Model, HMM) 的觀察值 (Observation) 之間相互獨立，同時狀態 (State) 之間具有方向性，在狀態移轉的過程中，僅與前一個狀態有關，無法考慮序列之間的前後關係，限制其特徵選擇，實際上序列資料標註的品質，與了解字詞、文句、段落長度，以及上下文之間，有著很大的關係。如圖 2 所示，{$y1, y2, ..., yn$} 爲狀態變數，即對應的序列標註，{$x1, x2, ..., xn$} 爲觀察變數，即待標註的文本序列資料，在 $y1$ 移轉至 $y2$ 的過程中，$x1$ 的值僅依賴於當前的 $y1$，同時 $y2$ 值由 $y1$ 決定，不依賴其它變數，形成 $x1$ 到 xn 之間彼此獨立的現象。

最大化熵馬可夫模型 (Maximum-entropy Markov model, MEMM) 的狀態移轉過程同樣具有方向性，卻解決了 HMM 的觀察值獨立問題，對相鄰狀態之間的依賴關係與整個觀察序列加以考量，可以任意選擇特徵，然而 MEMM 在狀態移轉的過程中，進行了局部歸一化，僅求出局部的最佳結果，傾向於選擇更少移轉的狀態，此種作法容易產生標註偏差的問題 (Label Bias Problem)，造成語料當中未曾或鮮少出現的字詞，容易被忽略。如圖3 所示，$y2$ 的值，是根據前一個狀態 $y1$ 與當前的觀察值 $x2$ 得出，每一個狀態移轉的過程，都要服從最大化熵的模型計算結果，形成局部歸一化，容易會有標註偏差問題的產生。

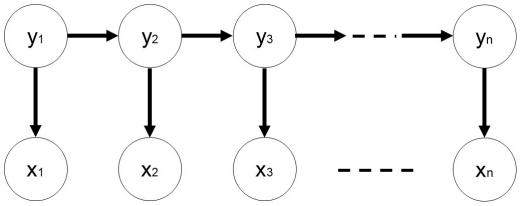

圖 2. Hidden Markov Model.

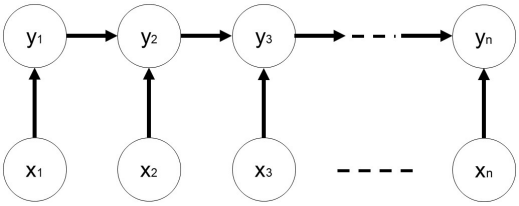

圖 3. Maximum-entropy Markov model.

CRF 能夠對所有觀察序列加以考量，且狀態移轉不具有方向性，代表不需要在每一個狀態移轉的情況下，各別進行局部歸一化，而是能夠將所有特徵進行全域性的了解，再進行歸一化，讓序列標註過程中的每一個狀態，都能與當前全部狀態有所關聯，也因此能夠得到最好的序列標註成效。如圖4所示，隨機輸入的 x 將會求出對應的 y，而 y 值的計算，是透過動態規劃 (Dynamic Programming) 的演算方式得知，試圖從鄰近的 y 預測出所有組合，找出最有可能的標註結果。本研究將使用標準的 CRF 模型。

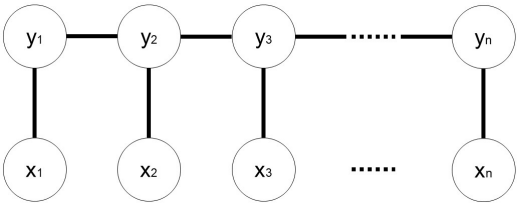

圖 4. Conditional Random Field.

©u¨ 對抗訓練

對抗訓練 (Adversarial Training) 由 Ian Goodfellow 等人於 2015 年提出 (Goodfellow et al., 2014)，該文設計一方法 FGSM (Fast Gradient Sign Method)，有效在高維的線性空間中將輸入資料上加入少量的擾動使得輸出結果預測錯誤 (圖5)；同時藉由所產生的資料作爲輸入樣本進行訓練以提高預測的準確率。公式如 (1)：

$$= sign(r\,xJ(\ ;x;y)): \qquad (1)$$

其中 爲模型參數，x 爲輸入資料，y 爲輸出目標，$J(\ ;x;y)$ 則爲損失函數 (Cost function)。FGM (Fast Gradient Method) 同樣由

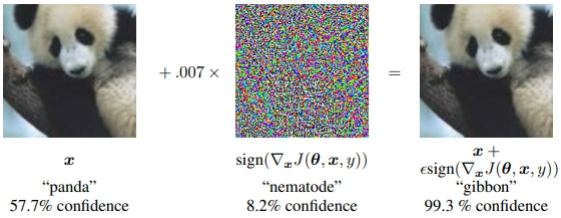

圖 5. FGSM 對抗樣本生成.

Ian Goodfellow 等人於 2017 年提出 (Miyato et al., 2016)，透過梯度的優化取代了 FGSM 中 Sign 的函式 (2) 以取得更好的對抗訓練樣本。

$$r_{adv} = \quad g/gjj_2 \qquad \text{where } g = r_x \log p(yjx : \hat{}): \qquad (2)$$

PGD (Projected Gradient Descent) (Madry et al., 2017) 相較 FGM 透過 *epsilon* 參數進行一次性的擾動可能無法得到最佳解的可能，PGD (3) 透過迭代方式以確保擾動不會過大。

$$x^{t+1} = \prod_{x+S}(x^t + \quad sgn(r\,xL(\ ;x;y))): \qquad (3)$$

本實驗將使用 FGM 及 PGD 作爲對抗訓練的模型用以強化訓練結果。

©uV 混合精度訓練

混合精度訓練 (Mixed Precision Training) (Micikevicius et al., 2017) 用意在於盡可能減少精度損失的前提下利用半浮點數 FP16 替代原 FP32 儲存權重及梯度，同時降低記憶體使用且能達到訓練時間成本降低的作用。雖然透過預訓練模型已大幅降低訓練時間，但對於模型的微調 (fine-tuning) 階段仍可能需要花費一定時間。透過混合精度訓練，在不影響預測精準度前提下有效節省訓練時間及記憶體使用。

混合精度訓練的最大挑戰是如何避免 FP16 半精度導致訊息損失。共有三種方式防止訊息丟失，分別是複製 FP32 的權重 (FP32 Master Copy of Weights)、Loss-scaling 以及高精度計算的改善 (ARITHMETIC PRECISION)。

©uVuo 複製 Sp¨© 的權重

神經網路的正向傳播時，將權重由 FP32 轉成 FP16 並計算 Loss 及梯度，最終再轉回 FP32 進行更新。實際參考如圖6。

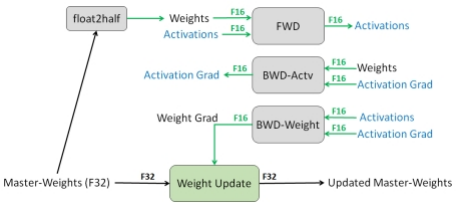

圖 6. 混合精度訓練迭代方式.

©uVu© dn] H9e_jX

訓練過程中部分權重可能因轉換爲 FP16 梯度會變成 0(圖 7)，藉由 Loss-scaling 的方式進行 Loss 的縮放，通過 Loss 的放大在反向傳播時放大梯度，最後再更新 FP32 前再縮放還原。

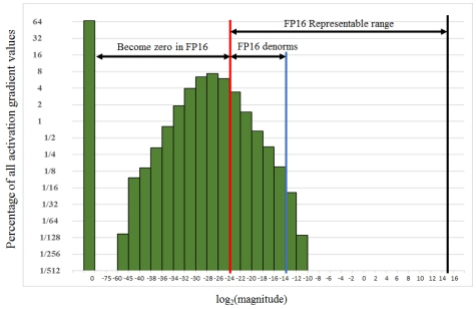

圖 7. Loss Scaling

©uVu¨ 高精度計算的改善

此研究發現重新優化其高精度的計算方式能夠有效減少訊息的損失。其計算方式爲將 FP16 的矩陣相乘後再與 FP32 的矩陣進行加法運算。

3 實驗

本實驗主要分爲幾個步驟，3.1 及 3.2 分別定義實驗的評估指標以及使用的資料集。3.3 嘗試藉由語料擴增提升預測精準度。3.4 說明本實驗所採用的預訓練模型及相關參數設定。3.5 嘗試藉由預測結果後處理提升預測精準度。最後，3.6 說明實驗結果。

¨uo 評估標準

爲同時考慮實驗結果之精確率 (precision) 與召回率 (recall)，本研究中采用 F1-score 作爲實驗模型評估指標，F1-score 計算方式如 (4)，其中 TP、TN、FP、FN 分別代表 True Positive、True Negative、False Positive、False Negative：

$$\begin{aligned} Recall &= TP/(TP + FN) \\ Precision &= TP/(TP + FP) \\ F1 \quad score &= 2 \quad Precision \quad Recall \\ &/(Precision + Recall): \end{aligned} \tag{4}$$

¨u© 資料集

本實驗使用 ROCLING-2022 Shared Task 所提供之資料，訓練集 Chinese Health NER Corpus(Lee and Lu, 2021) 包含 30,692 個句子，總計約 1,500,000 個字元或 91,700 個單詞；共有 68,460 個命名實體，涵蓋 10 種實體類型。測試集包含 3205 句中文句子。由於前期並未給出測試集，我們按照官方所提供之資料，將官方從訓練集切分出的 2,531 筆資料作爲實際測試集，其餘 28,161 筆資料作爲實際訓練集，並從實際訓練集中隨機地切分百分之十作爲訓練時的驗證集，即開發集。

¨u¨ 資料擴增

爲使 NER 的分類能夠獲得更高的精準度，本實驗分別從康健網、醫聯網及 KingNet 國家網路醫藥三個醫療保健相關網站蒐集文章 (共計 2290 篇文章) 並進行人工標註作業。然而在實驗過程中，加入額外的訓練集並沒有在結果上帶來顯著的幫助，反而導致了 F1 降低的情況，推估原因可能在於人工標註仍存在著標註錯誤的可能，另一原因在於標註內容涵蓋範圍超出原訓練集的標註範圍，導致在識別命名時，多出原先無法識別 (標註爲 O) 的實體。文章標註參考如表1。

分類	標籤數量
Body (BODY)	22487
Symptom (SYMP)	17416
Instrument (INST)	706
Examination (EXAM)	1780
Chemical (CHEM)	6423
Disease (DISE)	21386
Drug (DRUG)	3851
Supplement (SUPP)	7037
Treatment (TREAT)	2444
Time (TIME)	952

表 1. 文章標注參考.

¨uV 模型設計

實驗組別部分，我們使用 BiLSTM-CRF 模型作爲實驗的 Baseline，設置了五組實驗，其中前兩組分別用於選擇模型架構、提升訓練速度，後三組用於提升模型性能 (對抗訓練、Bert 選擇、資料增強及後置處理)。實驗程式部分，本實驗主要使用 pytorch、transformers、simple transformers 工具包，以 Bert 爲基礎的模型均來源於 huggingface (Wolf et al., 2020) 中的開源模型，分別用到'hfl/rbt6'、'hfl/chinese-bert-wwm'、'hfl/chinese-electra-base'、'hfl/chinese-macbert-base'。訓練參數部分，模型 learning rate 爲 3e-5，batch size 爲 32，training epoch 爲 50；爲了防止梯度爆炸，採取梯度裁剪 (gradient clipping) 的方式，最大 norm 值爲 5；使用 AdamW 優化器，weight decay 設定爲 0.01；爲了防止 overfitting，採用 early stopping 的方式，設定 patience 值爲 10，min delta 爲 2e-5，每次 F1-score 的值提升值大於 min delta 才算有改善。文字處理設定的部分，句子的最大長度爲當前批次中最長句子的長度，若當前批次所有句子的集合爲 B，則句子最大長度可表示爲 (5)：

$$Lmax = \underset{s \in B}{arg\max}\,(\text{len(s)}): \tag{5}$$

¨uU 資料後處理

在實驗過程中，我們發現部分訓練集內的資料存在歧異性及標註錯誤的可能。嘗試藉由後處理進行模型輸出後的後處理，其處理方式如下：將訓練集的所有詞 (word) 整理成字典檔，並針對所有詞逐一進行結果的替換 (替換方式爲將相同詞進行命名實體的替換，如：維他命的 BIO 爲 [B-SUPP], [I-SUPP], [I-SUPP]；當輸出的句中包含維他命時即將其 BIO 進行替換)。替換後計算 F1-score 的結果；最後再篩選取得大於未進行後處理的測試集 F1 結果製作成字典檔作爲後處理的依據。惟此作法雖能在測試集獲得好的成績，但在最後的結果中並不如預期可有效提高準確率。

¨u£ 實驗結果

第一組實驗的目的是確定基礎的模型架構，我們以 BiLSTM-CRF 模型作爲 Baseline，選用 RoBERTa-wwm-ext 爲 BERT 系列模型代表，設置了 RoBERTa-softmax、RoBERTa-CRF、RoBERTa-BiLSTM-CRF 三種模型架構作爲對照實驗，實驗結果如表2。

model	dev_f1	test_f1
BiLSTM-CRF	0.7301	0.6919
RoBerta-Softmax	0.7541	0.7299
Roberta-CRF	0.7727	0.7453
Roberta-BiLSTM-CRF	0.7613	0.7496

表 2. 模型架構對比實驗結果.

相較 BiLSTM-CRF 模型，RoBERTa-softmax 直接使用具有雙向 Transformers 結構的 RoBERTa，即使未加入更複雜的 layer，亦能有明顯提升；加入 CRF 層的 RoBERTa-CRF 較 RoBERTa-softmax 效果更好；而相較 RoBERTa-CRF，RoBERTa-BiLSTM-CRF 的結構僅在測試集上有少許提升，我們猜測是由於 BERT 系列模型已經具有雙向 Transformers 結構，其效果與 BiLSTM 差不多，故沒有太明顯的提升。考慮到增加 BiLSTM 會增加了模型複雜度，我們將 RoBERTa-CRF 作爲基礎的模型架構，在後續實驗中以其作爲 Baseline。在實驗過程中，爲了提升訓練速度、減少記憶體空間，我們使用混合精度訓練的方式，第二組關於速度提升的實驗數據如表3。

平均一個 epoch 所需時間	平均一個 epoch 所需記憶體	dev_f1	test_f1
439s	10025MiB	0.7712	0.7514
296s	9445MiB	0.7728	0.7559

表 3. 混合精度訓練實驗結果.

使用混合精度訓練並未產生精度損失，並且訓練速度得到明顯提升，每個 epoch 所需時間較原來減少了 32.6%；同時，所需記憶體空間也有少許減少。第三組實驗用於增強模型的魯棒性 (Robustness)，我們通過在 embedding 層增加擾動 (perturbation) 的方式，分別實作了 FGM 和 PGD 兩種攻擊方式，並將其應用在對抗訓練中。對抗訓練具體實作方式分爲四個步驟：首先計算輸入樣本 x 的 loss function 和在 x 處的 gradient；接著使用 FGM 或是 PGD 方法進行攻擊，計算樣本 x 對應的擾動量 r_{adv}；得到對抗樣本 $x_{adv} = x + r_{adv}$ 後，再次輸入模型中，計算 x_{adv} 的 loss，並在正常的 gradient 上累積對抗訓練的 gradient；最後恢復 embedding 參數並進行下一個 batch。實驗中設置 FGM 的 值爲 1.0，PGD 的 值爲 1.0、 值爲 0.3，迭代步數分別設置爲 1、3、5、7。實驗結果如表4。

model	train_f1	dev_f1	test_f1
RoBERTa-CRF	0.9893	0.7727	0.7453
RoBERTa-CRF-FGM	0.9766	0.7727	0.7568
RoBERTa-CRF-PGD,step=1	0.9887	0.7676	0.7487
RoBERTa-CRF-PGD,step=3	0.9767	0.7712	0.7514
RoBERTa-CRF-PGD,step=5	0.9787	0.7707	0.7585
RoBERTa-CRF-PGD,step=7	0.9920	0.7723	0.7461

表 4. 對抗訓練實驗結果.

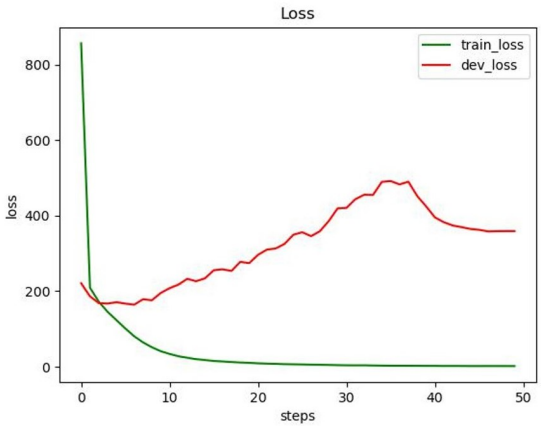

圖 8. The loss of RoBERTa-CRF.

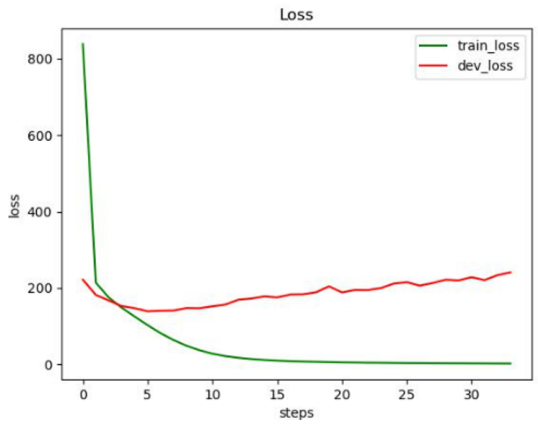

圖 9. The loss of RoBERTa-CRF-FGM.

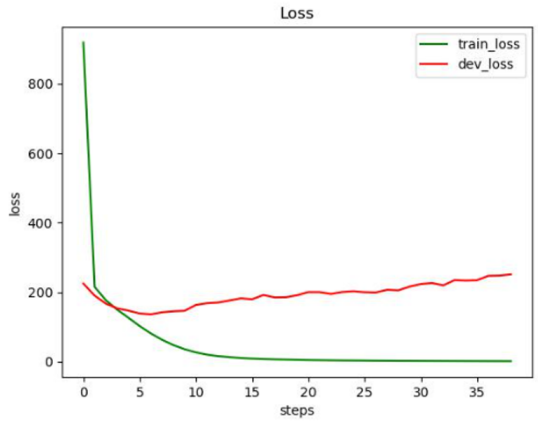

圖 10. The loss of RoBERTa-CRF-PGD,step=3.

使用對抗訓練後，模型在測試集上的 F1-score 分數更高，如圖8、9、10，使用對抗訓練後，能夠一定程度緩解模型過擬合的情況；對於使用 PGD 進行對抗訓練，步數爲 5 時效果最好，步數設定太多或太少模型效果均會變差；使用 FGM 和使用 PGD、步數爲 5 時效果差不多。

第四組實驗用於選擇合適的 BERT 模型，我們分別選用 BERT-wwm、RoBERTa-wwm-ext、ELECTRA、MacBERT-base， 後接 CRF 層的模型架構，比較不同 BERT 系列模型的效果，表5爲實驗結果，我們最終選擇了效果最佳的模型 MacBERT-base。

model	train_f1	dev_f1	test_f1
ELECTRA	0.9923	0.6570	0.6067
BERT-wwm	0.9340	0.7608	0.7448
RoBERTa-wwm-ext	0.9893	0.7727	0.7453
MacBERT-base	0.9769	0.7669	0.7465

表 5. BERT 系列模型訓練結果.

最後一組實驗嘗試通過資料增強與資料後處理的方式提升模型的 F1-score 實驗結果如表6。

4 結論

本實驗最終提交了以 MacBERT-base 爲架構的三個模型作爲 ROCLING-2022 Shared Task 的比賽成績。其中 run1、run2 使用 ROCLING-2022 Shared Task 的所有訓練資料 (3205 筆) 作爲訓練集，run3 則加入了額外標注的 54946 筆資料；run2、run3 加入對資料的後處理；分別在官方測試集上取得了

model	dataset	dev_f1	test_f1
MacBERT-CRF+PGD 對抗訓練 + 混合精度訓練	原始資料集	0.7721	0.7599
MacBERT-CRF+PGD 對抗訓練 + 混合精度訓練	原始資料集 + 拓展資料	0.7338	0.7091
MacBERT-CRF+PGD 對抗訓練 + 混合精度訓練 + 後置處理	原始資料集		0.8004

表 6. 資料增强與後處理實驗結果.

0.7796、0.7512、0.6962 的 F1-score。

賽後我們持續優化模型，將 ROCLING-2022 Shared Task 的 3205 筆訓練資料作爲訓練集，並從訓練集中隨機切分百分之十作爲訓練時的驗證集進行訓練，使用官方提供之測試集作爲實驗測試集，在 MacBERT-base 的基礎上加入 CRF 層，使用混合精度訓練、對抗訓練，得到圖7結果，其中使用 MacBERT-CRF 模型，在使用 PGD 對抗訓練設置步數爲 5 時效果最佳 (表7)，F1-score 爲 0.8076，已趨近於官方公布的最佳成績。

基礎模型	模型改進方式	dev_f1	test_f1
MacBERT	無	0.7659	0.7786
MacBERT	CRF+ 混合精度訓練	0.7719	0.7987
MacBERT	CRF+ 混合精度訓練 +FGM 對抗訓練	0.7701	0.7983
MacBERT	CRF+ 混合精度訓練 +PGD 對抗訓練 (step=3)	0.7682	0.8011
MacBERT	CRF+ 混合精度訓練 +PGD 對抗訓練 (step=5)	0.7725	0.8056
MacBERT	BiLSTM+CRF+ 混合精度訓練 +PGD 對抗訓練 (step=5)	0.7687	0.8076

表 7. 綜合模型訓練結果.

藉由實驗證實混合精度訓練能夠有效提升模型訓練速度。CRF 對預測的精準度有一定程度的提升；PGD 對抗訓練能夠一定程度提升模型魯棒性，並少許提升模型的預測能力；BiLSTM 對模型精準度僅有輕微提升；使用外部資料進行資料擴充則沒有明顯的效果。資料後處理則依據不同的測試集效果不一。

本文的主要貢獻如下：(一) 以 BiLSTM + CRF 作爲實驗 Baseline，在 4 組實驗當中，得知預訓練模型 MacBERT-base + CRF 的成效明顯高於 Baseline 與其它實驗結果。(二) 在 MacBERT-base + CRF 的基礎之下，加入混合精度訓練和對抗訓練，大幅地降低模型訓練的時間，一定程度提升資料標註的效果，與任務評測最佳團隊的系統，有著類似的成效水準。

本研究所需要的運算資源，較過去實驗要多，隨著硬體與機器算力的進步，相信會得到解決。每當實驗次數的增加，便能不斷地提升模型的成效，希望未來能夠建立更完善的 NER 語言模型，以提供更多的應用案例。

參考文獻

Madeleine Bates. 1995. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022. Overview of the rocling 2022 shared task for chinese healthcare named entity recognition. in proceedings of the 34th conference on computational linguistics and speech processing.

Lung-Hao Lee and Yi Lu. 2021. Multiple embeddings enhanced multi-graph neural networks for chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2801–2810.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Lawrence Rabiner and Biinghwang Juang. 1986. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human behavior inspired machine reading comprehension. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 425–434.