SCU-MESCLab at ROCLING 2023 "MultiNER-Health" Task：Named Entity Recognition Using Multiple Classifier Model

Tzu-En Su, Ruei-Cyuan Su, Ming-Hsiang Su, and Tsung-Hsien Yang Department of Data Science, Soochow University, Taipei, Taiwan
{70614roy, 70613rex, huntfox.su, yasamyang}@gmail.com

摘要

本研究旨在命名實體識別模型的任務上，設計出多重分類模型，並運用在醫學領域。其中訓練資料以 BIO 格式進行標記，我們採用不同組合的模型進行選擇，以六種模型進行評估及挑選，最後篩選出最佳三種模型分別為 BERT-based NER Model、RoBERTa (base)-based NER Model，以及 RoBERTa (large) + BiLSTM + CRF Model，並應用於多重分類模型，使 RUN1 獲得最佳預測，其平均精確度為 68.69%、平均召回率為 67.64%，平均 F1-score 為 68.13。

Abstract

This study aims to design a multi-class classification Model for the task of named entity recognition and apply it in the medical field. The training data is labeled in the BIO format. We employed various combinations of Models for selection, evaluating, and choosing among six Models. Finally, we identified the top three Models: a BERT-based NER Model, a RoBERTa (base)-based NER Model, and a RoBERTa (large) + BiLSTM + CRF Model. These Models were applied to a multi-class classification setup, with RUN1 achieving the best predictions. The average precision for RUN1 is 68.69%, the average recall is 67.64%, and the average F1 score is 68.13%.

關鍵字：BERT, RoBERTa, NER
Keywords: BERT, RoBERTa, NER

1 Introduction

自然語言處理 NLP (Natural Language Processing) 的技術發展使醫療領域應用逐漸取得優異成果，從臨床決策 (Dina Demner-Fushman, 2021)到醫學研究 (Weiner, 2012)，都得到了有效支持。但同時此應用必須面對一系列挑戰。例如，醫學領域專業術語的多樣性以及在海量信息中準確辨識所需信息至關重要。只有正視這些問題，才能在醫療領域的方面，實現更有效的醫療信息傳遞 (Konam and Rao，2021)。因此，命名實體識別 (Named Entity Recognition, NER) 在醫療領域上具有明顯的應用價值，也可以幫助解決這一問題。NER 是一種常見的自然語言處理任務，能夠自動辨識文本中的實體，如人名、地點、組織機構、日期、藥物、疾病等，並從中提取有價值的信息。

而隨著技術的不斷演進，命名實體識別從傳統機器學習的隱藏式馬可夫模型(hidden Markov Model, HMM)，到深度學習中的雙向長短時記憶網路 Bidirectional Long Short-Term Memory (Bidirectional LSTM) 結合條件隨機場 (Conditional Random Field, CRF) (Huang et al., 2015)，甚至結合 Bidirectional Encoder Representations from Transformer (BERT)、Robustly Optimized BERT Pretraining Approach (RoBERTa) 等預訓練模型，使得模型可以更有效提升其效率以及精確度。

但不同於英文，中文的命名實體識別的方法上面臨許多問題，如分詞或是歧義性，使得在中文的訓練集上，需要更多資訊以確保模型訓練上的穩定，為此在任務選擇上，我們採用中文的預訓練模型，並選擇不同模型的組合進行微調及改進，並分別訓練出各個

模型。最後我們將各個預測資料採用我們的多重分類模型，並評估此方法的可行性。

2 Dataset

本次的研究實驗中，我們採用 Chinese Healthcare Named Entity Recognition (HealthNER) 數據集，該數據集是由 NCUEE NLP 研究室的團隊成員進行收集與標註 (Lee and Lu, 2021; Lee et al., 2022a)。這些中文數據是通過社交媒體網路爬取而來，並涵蓋了醫療保健資訊、健康相關新聞以及醫療問答論壇網站上的文章。經過人工標註後，這個數據集包括了 30,692 個句子，總計 150 萬個字符或 91,700 個單詞。

在標註過程中，共識別出了 68,460 個命名實體。這些從網路爬取的文章使用 BIO 格式進行標記。例如，"鈣質"被標記為"B-CHEM"和"I-CHEM"，而"骨骼"會被標記為"B-BODY"和"I-BODY"，以此類推。並且依據不同的類型，數據集分別包含 10 種不同的實體類型，它們的名稱分別為人體(BODY)，症狀(SYMP)，醫療器材(INST)，檢驗(EXAM)，化學物質(CHEM)，疾病(DISE)，藥品(DRUG)，營養品(SUPP)，治療(TREAT)，時間(TIME)。而類別以外的字全標為"O"。訓練資料最終包含 28,161 個句子，而測試資料作為驗證集包含 2,531 個句子，其中訓練資料有 61,155 個命名實體，而測試資料有 7305 命名實體。最後大會提供 MultiNER-Health_truth 當作最終的測試集，而分別為"FT"和"SM"以及"WA"共 6,626 句，因此我們可以將 HealthNER 中的測試資料當作我們模型的驗證集使用。

3 Experimental Model

3.1 BERT

BERT 模型，即 Transformer 的 Encoder。是谷歌以無監督方式利用大量無標記文本的方法訓練而成 (Devlin, 2018)。其訓練資料來自於 Wikipedia 2.5B 語料集加上 BookCorpus 800M 的語料集。批量大小為 1,024 * 128 長度或 256 * 512 長度。BERT 分為 BERT-Base (12-layer, 768-hidden, 12-head) 和 BERT-Large (24-layer, 1,024hidden, 16-head) 兩種形式。其結構基於多層的編碼器，但不包含解碼器，因此無法用於生成需要預測的信息。然而，BERT 的主要創新在於其先前訓練的方法，而不是其模型架構本身。這種先前訓練方法使模型能夠在預訓練階段學習大量文本數據的上下文信息，從而使其在各種下游任務上表現出色。

3.2 RoBERTa

RoBERTa 是基於 BERT 模型的優化版本，由 Facebook AI 於 2019 年發表 (Liu, 2019)。RoBERTa 在訓練策略上進行了一系列的優化，包括使用更大的批次大小、更長的訓練時間和更多的訓練數據。這些調整有助於讓模型更好學習語言表示。具體而言，英文的 RoBERTa 主要是在維基百科及書籍語料庫上進行訓練，而中文的 RoBERTa 主要使用了哈工大訊飛聯合實驗室發布的 RoBERTa 模型 (Cui et al., 2020)，該模型分為 RoBERTa-Base (768-hidden) 和 RoBERTa-Large (1024hidden) 兩種形式 (Xu et al., 2020)。這個模型經過了第三方中文基準測試 CLUE 的驗證。CLUE 的基準測試包含了 6 個中文文本分類數據集和 3 個閱讀理解數據集，其中包括哈工大訊飛聯合實驗室發布的 CMRC 2018 閱讀理解數據集，在中文訓練中，我們選擇了他們的模型作為基準。

3.3 LSTM

LSTM (Hochreiter, 1997) 是用於處理序列數據的循環神經網路 (Recurrent Neural Network, RNN) 的特殊變種。LSTM 解決長序列數據上的梯度消失及梯度爆炸等問題。LSTM 主要由四個閘 (gate) 及一個元所組成，輸入閘 (Input Gate)，遺忘閘 (Forget Gate)，輸出閘 (Output Gate) 及記憶元 (Memory Cell)。其中(1) 至 (3) 分別為 Input Gate, Forget Gate 和 Output Gate 計算公式。Input Gate 用於決定何時將數據輸入單元，Output Gate 負責從單元中輸出結果，Forget gate 管理單元內容的重置。其中H_{t-1}為其一個時間的隱藏層，W以及b分別為權重和偏差。其中 (4) 為候選記憶元 (candidate memory cell) 計算和上述相似。而 (5) 為控制多少輸入和遺忘資料，輸入閘I_t控制採用多少來自$\tilde{C}_t$的新數據，而遺忘閘F_t控制保留多少過去的記憶元C_{t-1}的內容。最後隱藏層(6)經由輸出閘 O_t和新的記憶元C_t的計算確保H_t的取值始終在區間(-1,1)之間內。

$$I_t = \sigma(X_t W_{Xi} + H_{t-1} W_{hi} + b_i) \qquad (1)$$

$$F_t = \sigma(X_t W_{Xf} + H_{t-1} W_{hf} + b_f) \quad (2)$$
$$O_t = \sigma(X_t W_{Xo} + H_{t-1} W_{ho} + b_o) \quad (3)$$
$$\tilde{C}_t = tanh(X_t W_{Xc} + H_{t-1} W_{hc} + b_c) \quad (4)$$
$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \quad (5)$$
$$H_t = O_t \odot tanh(C_t) \quad (6)$$

因此這種機制幫助LSTM模型有效地捕捉序列中的長期依賴關係。而BiLSTM被用於學習時間序列之間的相互關係，使得模型能夠類似於隱馬爾可夫模型一樣具備建模能力(Schuster & Paliwal, 1997)。 BiLSTM通過訓練輸入閘、遺忘閘、輸出閘等雙向傳遞的信息更新 (Graves & Schmidhuber, 2005)。換句話說，我們在預測時使用了過去和未來的文字等信息來做出預測。在我們研究中並不是預測下一個字，而是藉由整個句子的分析並且各個字之間帶有時間前後輸出信息向量，因此我們最佳選擇是使用 BiLSTM 完成此任務。

3.4 CRF

CRF 是一種統計模型，主要使用於各種序列標籤的任務，它是屬於監督式學習模型，廣泛應用於自然語言處理等領域。它透過相鄰元素之間的關係，來預測序列中間的關係，與最大熵分類器 (Ratnaparkhi, 1996) 和最大熵馬爾可夫模型 (MEMMs) (McCallum et al., 2000) 相比 CRF 更具靈活性和強大性，而 CRF 模型包含兩個部分：特徵函數以及參數。特徵函數將輸入序列的元素映射為一組特徵，這些特徵可以包括詞彙、詞性、上下文等信息。通過從訓練數據中學習參數，這些參數控制不同特徵對標籤的影響。模型的訓練過程旨在最大化在給定輸入條件下的標籤序列可能性，並找到最佳的參數組合。由於 CRF 能夠全面考慮序列的上下文信息之間的關係，因此，它在詞性標註、命名實體識別和詞分割等任務中表現更佳的出色。

4 Experimental Result

此次競賽中，大會允許提交每種測試集各三個最佳的預測結果。我們在以下各小節分別先說明每種模型的訓練成果，最後在說明三次提交(RUNS)採用的方法與相關參數設置。

4.1 BERT-based NER Model

在獲得訓練結果之前，我們採取了以下步驟。首先，我們分別使用不同的預訓練模型進行訓練，以獲得各自的輸出向量。接著，我們將這些輸出向量輸入至一個線性分類器中，以進行最終的分類。預訓練模型方面分別採用中研院中文計算語言研究 (Chinese Knowledge and Information Processing, CKIP) 所發佈的 BERT 繁體中文預訓練模型 (Yang and Ma, 2021) 、哈工大訊飛聯合實驗室發布的 RoBERTa-wwm-ext (base) 模型以及 RoBERTa-wwm-ext-large 模型。這些模型分別生成 768 維、1024 維的輸出向量。我們將訓練集中的最大句子長度設定為 441 與 batch size 為 16 並以 adamw 為優化器進行訓練。首先，我們以大會提供之訓練集進行模型訓練，並以驗證集進行模型測試。在三個模型訓練過程中，經過 3 個 epoch 後，綜合考慮 F1 值和準確率的平均表現，整體性能相較其他 epoch 表現更優越。以此為基準，進行模型訓練，並以大會最終提供的測試集進行衡量，整體來說 BERT 模型表現較佳，平均 F1 score 為 66.94%，如表 1 所示。RoBERTa (large) 模型 Accuracy 為 91.74%、Precision 為 66.05%、Recall 為 67.38% 與 F1 score 為 66.68。雖然 RoBERTa (large) 在 Accuracy 及 Recall 較佳，但主要以 F1 score 為基準，因此選擇所有模型之外的最佳 F1 score 當作我們的 RUN 2。

4.2 BERT+BiLSTM+CRF NER Mdeol

同樣，我們分別使用不同的預訓練模型進行訓練，以獲得各自的輸出詞向量。接著，我們將這些輸出詞向量輸入至 BiLSTM 和 CRF 中，以進行我們的研究。各個預訓練模型分別為 CKIP BERT 繁體中文預訓練模型、哈工大訊飛聯合實驗室發布的 RoBERTa-wwm-ext (base) 模型以及 RoBERTa-wwm-ext-large 模型，並且訓練使用 SGD 隨機梯度下降，學習率為 0.012，weight decay 為 1e-5，且設定 scheduler 每兩次 epoch 時學習率減少 0.9。實驗結果採用大會最終提供的測試集進行衡量，整體來說 RoBERTa (large) + BiLSTM + CRF 模型表現較佳， Accuracy 為 89.75%、Precision 為 67.18%、Recall 為 62.76% 與 F1 score 為 65.24%，而 BERT + BiLSTM + CRF 效能表現較差，其 F1 score 只有 63.91%。

4.3 Multiple NER Model

得 F1 Score 為 70.57、Marco-Averaging 中取得

表 1：各個模型訓練數據

Model	MultiNER_F1			Average (FT, SM, WA)			
	FT	SM	WA	Accuracy	Precision	Recall	F1
BERT-based NER	61.55	70.21	69.07	91.57	67.11	66.75	**66.94**
RoBERTa (base)-based NER	60.76	70.36	69.41	91.57	67.1	66.65	66.84
RoBERTa (large)-based NER	60.81	69.88	69.37	**91.74**	66.05	**67.38**	66.68
BERT + BiLSTM + CRF	61.69	58.49	71.55	89.72	**68.02**	60.45	63.91
RoBERTa (base) + BiLSTM + CRF	60.82	60.46	71.51	88.94	65.84	62.81	64.26
RoBERTa (large) + BiLSTM + CRF	62.09	61.34	72.3	89.75	67.18	62.76	**65.24**

多重 NER 模型的操作流程如下。首先，從三個不同的模型中獲取各自輸出的標籤結果。接著，基於這些標籤進行多數決，即選擇在多數模型中獲得的標籤作為最終的預測標籤。若所有模型的輸出標籤不盡相同，則會選擇其中一個具有較高 F1 score 的模型所產生的標籤作為最終的預測標籤。我們採取一個選擇步驟，排除了 F1 score 最差的 BERT + BiLSTM + CRF 模型，以及在任何分數上都表現不突出的 RoBERTa (base) + BiLSTM + CRF 模型。最後在 RUN1 ，我們定義 Model_1，包含三個模型，分別為 BERT-based NER Model、RoERTa (large)-based NER Model 以及 RoBERTa (large) + BiLSTM + CRF NER 模型。同樣，在 RUN3 中，我們定義 Model_2，包含 RoBERTa (base)-based NER、RoERTa (large)-based NER Model 以及 RoBERTa (large) + BiLSTM + CRF NER 模型作為實驗對象。在 RUN 1 ，我們得出 Model_1 Accuracy 為 91.81%、Precision 為 68.69%、Recall 為 67.64%、F1 score 為 68.13%。而 Model_2 的 Accuracy 為 91.87%、Precision 為 68.26%、Recall 為 67.84%、F1 score 為 68.01%。最後，Model_1 在測試集預測表現明顯突出。

4.4 Competition Results

本研究將 Model_1 做為比賽使用之模型，並上傳比賽數據所預測之結果。其中在 Formal Texts 中取的 F1 Score 為 62.51、Social Media 中取得 F1 Score 為 71.33、Wikipedia Articles 中取 F1 Score 為 68.14，最終取得第四名的佳績。

5 Conclusion and Future Work

在這項研究中，我們運用了三種不同的命名實體識別模型，各自分別為 BERT-based NER Model、RoBERTa (base)-based NER Model 及 RoBERTa (large) + BiLSTM + CRF Model，從而獲得了最佳的結果。並將其應用於醫療領域。根據其分類名稱分別為人體(BODY)，症狀(SYMP)，醫療器材(INST)，檢驗(EXAM)，化學物質(CHEM)，疾病(DISE)，藥品(DRUG)，營養品(SUPP)，治療(TREAT)，時間(TIME)。資料是以 BIO 格式去標記。例如，"鈣質" 會被標記為"B-CHEM"和"I-CHEM"，而 " 骨骼 " 會被標記為"B-BODY"和"I-BODY"，以此類推。而類別以外的字全標為"O"。最終的實驗中，我們使用 HealthNER 的全部資料，共有 30,692 句子作為訓練與驗證資料集，同時使用大會提供的 6,626 個句子為測試資料集，來進行三種不同模型的驗證。根據實驗結果顯示，RUN3 中的 Model_2 取得了良好的實驗結果。而 RUN2 使用的是 BERT-based NER Model 效能最差。然而，最優的結果出現在 RUN1 中。在 RUN1 中，Model_1 獲得了最佳的系統效能，其 Accuracy 為 91.81%、Precision 為 68.69%、Recall 為 67.64%，而 F1 score 則達到 68.13%。Model_1 在指標上均優於其他模型。由此可知，這項研究顯示，透過多重分類模型的方法在這個特定任務上的效果優於使用單一模型進行訓練。未來，我們有潛力進一步拓展這個

方法，引入更多不同種類的模型，以探討是否能夠進一步提升成果。

References

Dina Demner-Fushman, Noémie Elhadad and Carol Friedman. 2021. *Natural Language Processing for Health-Related Texts*. In: Edward H. Shortliffe, James J. Cimino (eds) *Biomedical Informatics*. Springer, Cham. https://doi.org/10.1007/978-3-030-58721-5_8

Weiner, Jonathan P. 2012. Doctor-patient communication in the e-health era. *Israel journal of health policy research*. 1(33):1-7. https://doi.org/10.1186/2045-4015-1-33.

Konam, Sandeep, and Shivdev Rao. 2021. Abridge: A Mission Driven Approach to Machine Learning for Healthcare Conversation. *Journal of Commercial Biotechnology*. 26(2):62-66.

Huang, Zhiheng, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for sequence tagging. *arXiv preprint* arXiv:1508.01991. https://doi.org/10.48550/arXiv.1508.01991

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Lung-Hao Lee, & 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. https://doi.org/10.1109/JBHI.2020.3048700.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022a. 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*, pages 363-368.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint.* arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. *arXiv preprint.* arXiv:2004.05986. https://doi.org/10.48550/arXiv.2004.05986.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735-1780.

Schuster, Mike, and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*. 45(11):2673-2681. https://doi.org/10.1109/78.650093.

Graves, Alex, and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*. 18(5-6):602-610. https://doi.org/10.1016/j.neunet.2005.06.042.

Ratnaparkhi, Adwait. 1996. A maximum entropy Model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*.

McCallum, Andrew, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy Markov Models for information extraction and segmentation. In *Proceedings of the 17th International Conference on Machine Learning,* pages 591-598.

Yang, Mu, and Ma, W.-Y. 2021. ckiplab/ckip-transformers. https://github.com/ckiplab/ckip-transformers