書背

111 國立臺北商業大學資訊管理系 碩士論文（12 pt） 論文中文題目（12 pt） 黃鈺傑 撰

國立臺北商業大學管理學院
資訊管理系人工智慧與商業應用碩士班
碩士學位論文

結合 MAMBA 與 CRF 在中文醫學命名實體識別

Apply MAMBA model and CRFs in Named Entity Recognition: using Chinese medical dataset

研究生：黃鈺傑

指導教授：林俊杰 博士

中華民國一一三年六月

目錄

圖目錄

表目錄

摘要

自然語言處理（NLP）領域是人工智慧(AI)中一個關鍵領域，它使機器能夠理解、分析和生成自然語言文本。近年來，深度學習和 Transformer 模型的崛起，以及大量可用的資料和強大的計算能力，推動了 NLP 的快速發展。NLP 不僅在文本分類、機器翻譯和自動問答等方面取得了重要突破，還在情感分析、語音識別和對話系統建構等領域實現了重要進展。但隨著處理序列長度和模型規模的增加，Transformer 也面臨著計算效率下降的問題，基於上述原因，本研究透過 Mamba 模型，在命名實體辨識(NER)的問題上驗證，透過 Mamba 結構能夠更有效地處理長序列，並且能夠在計算上實現線性擴展，突破傳統 Transformer 在長序列上的計算瓶頸，並結合 CRF 加強序列中的依賴關係。透過本研究的方法所生成的文字序列在實驗中得到很好的結果，從實驗結果來看模型帶給中文醫療命名實體辨識資料集 91.9%的 F1 值(F1-score)。

關鍵詞：自然語言處理、命名實體識別、結構化狀態空間、條件隨機場、Mamba

ABSTRACT

Natural language processing (NLP) is a key area in Artificial Intelligence(AI) that enables machines to understand, analyze, and generate natural language text. In recent years, the rise of deep learning and Transformer models, as well as the large amount of available data and powerful computing power, have promoted the rapid development of NLP. NLP has not only made important breakthroughs in text classification, machine translation, and automatic question answering, but also made important progress in the fields of sentiment analysis, speech recognition, and dialogue system construction. Based on the above reasons, this study uses the Mamba model to verify that the Mamba structure can process long sequences more efficiently, and can achieve linear expansion in computing, breaking through the computational bottleneck of traditional Transformer on long sequences, and combining with CRF to strengthen the dependencies in the sequence. The text sequences generated by the method in this study obtained good results in the experiment, and from the experimental results, the model brought 91.9% of the F1-score by using Chinese medical data set.

Keyword：NLP、NER、SSM、CRF、Mamba

第一章 研究背景

第一節 研究背景與動機

近年來，自然語言處理（Natural Language Processing, NLP）作為人工智慧的一個重要分支，不斷提升對自然語言的理解、分析和生成能力(Dai et al., 2019)。這一領域的快速發展主要得益於深度學習（Deep Learning）技術的創新、豐富的資料集以及計算能力的增強。這些因素不僅使 NLP 在文本分類、機器翻譯和自動問答等方面取得了突破，同時也在情感分析、語音識別和對話系統構建方面取得了顯著進展。這些進展為改進人機互動、資訊檢索和知識管理等領域帶來了全新的機遇(Praful Bharadiya, 2023)。

中文命名實體識別（Named Entity Recognition，NER）是 NLP 領域中至關重要的基礎任務，主要目標是在非結構化的文本中識別和分類命名實體。除了在 NLP 中扮演關鍵角色外，NER 還為多項 NLP 任務如關係抽取、事件提取、知識圖譜、機器翻譯以及問答系統等提供基礎支援(Lee & Chen, 2022)。

最早期的 NER 研究採用基於規則的線性模型，這些模型往往過度擬合於特定的結構化文本資料集(Jehangir et al., 2023)。隨著在大規模標記資料集上進行的發展，深度學習技術已被廣泛使用，一些新的方法不斷湧現。如 Bi-LSTM+CRF(Huang et al., 2015)，BERT(Devlin et al., 2019)，RoBERTa (Liu et al., 2019)等，並取得了不錯的結果，Transformer 模型的崛起無疑是一大亮點。它利用了自注意機制（self-attention），成功對理解文本取得了傑出的成就(Vaswani et al., 2017)。深度學習被證明是直接從文本資料中提取特徵表示的有效策略，這在 NER 領域取得了突破性進展(J. Yang et al., 2024)。

現今 NER 還適用於特殊領域，如生物學，它可以發現蛋白質和基因等實體；在製造業中，它可以識別產品和品牌(Pakhale, 2023)；在醫療保健中，保健資訊包括許多專有名稱，主要是作為命名實體，這些資訊涵蓋了許多專業術語和具體名詞，主要涉及醫學實體的命名，例如中樞神經系統（central nervous system）和固有結締組織（Connective tissue proper）(Eickhoff et al., 2020)。

隨著處理序列（Sequence）長度和模型規模的增加，Transformer 也面臨著一些限制。其中一個主要問題是，隨著上下文長度的增加，self-attention 的計算量呈指數級增長，導致計算效率下降。雖然有一些高效的變化模型被提出來，但會以降低模型效能作為代價。名為「Mamba」的架構模型似乎改變了這個情況。

Schiff et al. (2024)提出以 Mamba 架構作為基礎的雙向 Mamba 模型，一次處理正向序列，一次處理反向序列，並將反向序列輸出添加到正向的維度中，為了避免參數增加一倍，所以共用兩次序列的權重，證明雙向 Mamba 模型模型優於基於 Transformer 模型的能力(Schiff et al., 2024)。Ren et al. (2024)提出 Mamba 結合 Sliding Window Attention (SWA)+多層感知器（Multi-Layer Perceptron, MLP）的模型 (Samba)，證明 Samba 比基於注意力(Attention)的模型和基於 State Spaces Model(SSM)的模型還要良好。本研究將建立基於 Mamba 架構(Gu & Dao, 2023a)並且結合 CRF 的模型，在解決 Transformer 的運算效率低下的問題，並利用 CRF 加強對序列的依賴關係。

第二節 研究目的

本研究將於解決醫療領域的文本通常涵蓋各種專業術語、縮寫以及不同風格的描述，這使得特定醫療資訊的查找變得相對複雜。因此，結合自然語言處理技術和醫療保健資訊的深度分析，能夠為醫療專業人員和普通使用者提供更簡便、快速和精確的資訊檢索途徑。

在本研究中應用 Mamba 模型將序列標記轉化為實體提取。在特定文本的訓練下，Mamba 模型從文字中提取相關的實體類別並分配其對應實體類別標籤。除此之外，本研究還在模型中加入 CRF 模型，使本研究模型能夠捕捉序列中的依賴關係，理解資料之間的關係和語義資訊，提高模型的評估性能。本論文之貢獻為以下部分：

1. 提出將 Mamba 模型架構結合 CRF 的新模型。
2. 與 LSTM、Transformer 相關的模型進行比較，並提出 Mamba 結合 CRF 的優點

第三節 論文架構

本研究的論文架構主要包含五個章節，依序為研究背景、文獻探討、研究方法、研究結果、結論，如圖 1-1 所示。

下列為各章節介紹：

第一章　研究背景：此章節闡述本研究的研究背景與動機以及研究目的。

第二章　文獻探討：此章節回顧與本研究相關的文獻和先前學者的研究。

第三章　研究方法：此章節提供本研究所使用的資料集與模型架構。

第四章　研究結果：此章節呈現研究的具體成果。

第五章　結論：此章節是對於整個研究的總結，並提出未來能繼續研究的方向。

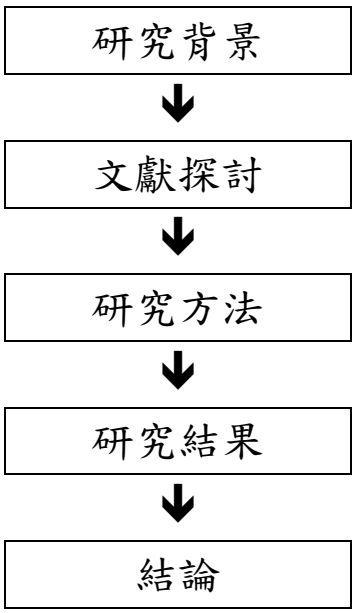


圖 1-1　論文架構

第二章 文獻探討

第一節 命名實體識別（NER）

在傳統上，NER 一直被視為序列標記（Part-of-Speech Tagging, POS Tagging）問題的一種，其中我們需要同時預測實體的邊界和其對應的類別標籤(Lee & Chen, 2022)。相較於英文 NER，中文 NER 更加具有挑戰性，中文以字為基本單位，不像英文那樣有明顯的規則性特徵可供參考。由於中文字與字之間沒有明確的分隔符號，因此中文 NER 中的字與分詞密切相關，這也意味著命名實體的邊界通常也會是分詞的邊界。然而，錯誤的實體分詞決策可能會造成在 NLP 中學習錯誤的資訊。例如，在特定情況下，正確的身體類別實體為“上皮組織惡性腫瘤”，但可能會被分割成三個錯誤的類別實體：“上皮組織”、“惡性”和“腫瘤”(Lee & Chen, 2022)。

現在的 NER 多用深度學習技術的模型，包括循環神經網路（Recurrent Neural Network, RNN）、卷積神經網路（Convolutional Neural Network）, CNN）、長短期記憶（Long Short Term Memory, LSTM）以及 BERT 和 GPT 等。這些模型具有文本資料中捕捉上下文關係與語義特徵的能力(Pakhale, 2023)。RNNs (Schmidt, 2019)和 LSTMs(Lample et al., 2016)在序列建立模型方面表現出色，可以捕捉單詞之間的依賴關係。另一方面，CNN(Chiu & Nichols, 2016)可以有效地捕捉局部模型，對於字元表示特別有用。

以 BERT (Devlin et al., 2019) 和 GPT (Brown et al., 2020) 為例的 Transformers 通過其注意力機制徹底改變了 NER，使模型能夠同時考慮句子中的所有單詞。例如，BERT (Devlin et al., 2019)豐富了在更廣泛的語言語境中對單詞的理解。這些技術的意義在於它們能夠通過預訓練利用大量未標記的資料，從而增強了它們在具有有限標記資料的 NER 任務中的表現(Han et al., 2023)，從而有助於提高最先進的 NER 準確性並推動自然語言處理領域的發展。

本研究運用 BIO（即開始、內部和外部）表示法來執行命名實體識別（NER）任務。在標記中，以實體標記為「B-」開頭的表示命名實體的開始，實體標記為「I-」開頭的表示命名實體的中間或結尾，實體標記為「O」則表示該輸入不屬於任何命名實體(Lee et al., 2023)。根據這些約束，不應該出現"O"標記後面立刻是「I-」標記，因為這代表著不連續的實體標記。CRF 模型有時會在生成預測標記時忽視這些約束，因此可能生成非法的序列(Wei et al., 2021)。舉例來說，完整命名實體句子為「需要定期做子宮頸抹片檢查。」,其標籤為「O, O, O, O, O, B-EXAM, I-EXAM, I-EXAM, I-EXAM, I-EXAM, I-EXAM, I-EXAM, O」。可以發現這句子中的「B- EXAM」表示測試實體標記資料開始，「I- EXAM」表示測試實體標記資料中間或結尾，「O」標記則表示該輸入不屬於測試實體標記。

第二節 循環神經網路（RNN）

RNN 是基於順序資訊的深度學習演算法。與前向神經網路（feed forward neural networks）不同，RNN 保留了一種狀態，該狀態可以表示來自任意長度上下文窗口的資訊。儘管 RNN 傳統上很難訓練，並且通常包含數百萬個參數，但網路架構、優化技術和並行計算的最新進展已經使大規模學習成為可能(Baviskar et al., 2023)。

Auli 等人,(2013)的 RNN 模型使用 Mikolov 的詞嵌入（Word Embedding）的表示,以方便在可能的翻譯空間中進行搜索。詞嵌入是一種用於文本分析的表示,它允許具有相似含義的單詞以實值向量的形式具有相似的表示(Mikolov et al., 2013)。可以使用一組語言建模技術獲得單詞嵌入，其中單詞被映射到實數的低維向量空間。通過上下文單詞表示，幾乎每一個 NLP 都得到了顯著改進（Ethayarajh, 2019）。

第三節 長短期記憶網絡（LSTM）

LSTM 是一種 RNN 的延伸變化,與標準 RNN 不同,LSTM 具有內部記憶單元,可以更有效地捕捉長期依賴性,這使其能夠更好地處理長序列,同時降低梯度消失的問題(Sak et al., 2014)。LSTM 具有選擇性記憶和遺忘機制,使其能夠有效地捕捉重要資訊,並長期保存有用的資訊。長期短期記憶網路與 RNN 相同,只是隱藏層更新被專用存儲單元所取代(Sak et al., 2014)。圖 2-1 顯示了採用上述 LSTM 模型。LSTM 儲存單元實現如下：

$$I_t = \sigma\left(W_{h_i} h_{t-1} + W_{x_i} \chi_t + b_i\right) \tag{1}$$

$$F_t = \sigma\left(W_{h_f} h_{t-1} + W_{x_f} \chi_t + b_f\right) \tag{2}$$

$$O_t = \sigma\left(W_{h_o} h_{t-1} + W_{x_o} \chi_t + b_o\right) \tag{3}$$

$$C_t = F_t C_{t-1} + I_t \, tanh\left(W_{h_c} h_{t-1} + W_{x_c} \chi_t + b_c\right) \tag{4}$$

$$H_t = O_t \, tanh(C_t) \tag{5}$$

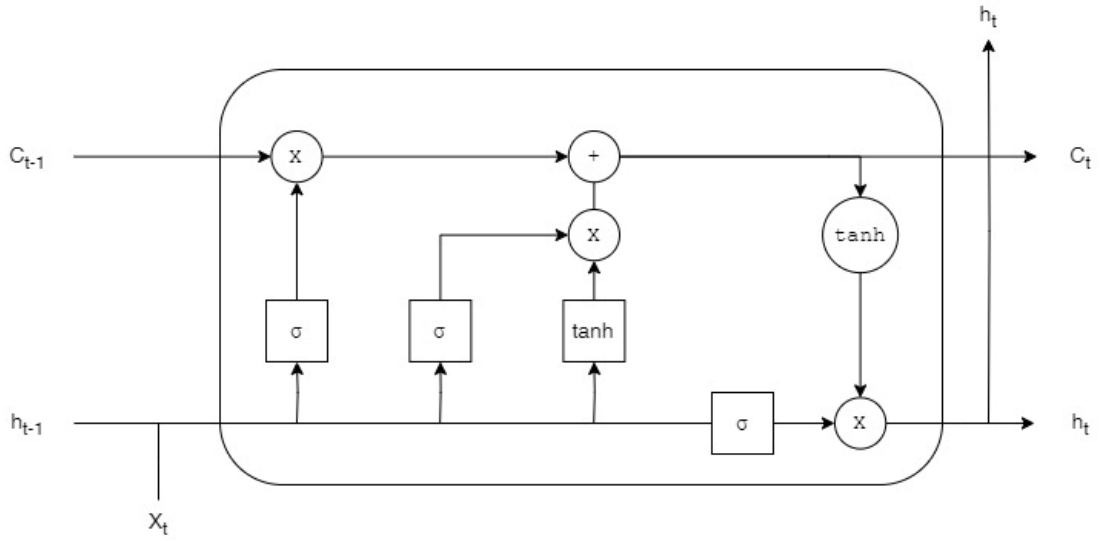


圖 2- 1　LSTM 模型

圖 2-2 為 LSTM 神經網路的應用，以身體類別的「中樞神經系統」為舉例，第一個字輸入「中」經由模型判預測這個字為「B-BODY」並產生一個初始的隱藏層，下一個字「樞」會與模型剛剛產生的隱藏層參數一起進行預測出「I-BODY」並更新隱藏層，第三個字「神」就會與第二個字更新的隱藏層一起預測，後面的字也依照同樣的方式進行預測。

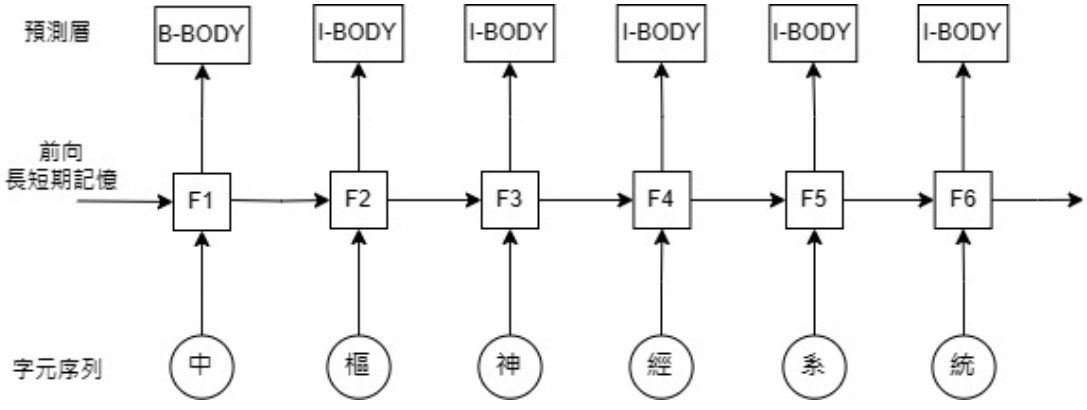


圖 2- 2　LSTM 神經網路的應用

第四節 雙向長短期記憶網絡（Bi-LSTM）

這種堆疊的雙向遞歸神經網路在自然語言處理領域中具有廣泛的應用。每一個單詞的特徵首先被送入一個前向 LSTM 和一個後向 LSTM，每一個時間軸，前向和後向 LSTM 網絡都會生成一個特徵向量，它們分別表示了單詞在文本序列中的上下文關係（圖 2-3）。這樣的設計可以在特定範圍內利用過去的特徵（前向狀態）和未來的特徵（後向狀態）。

這些特徵向量隨後通過一個線性層和一個 SoftMax 層的解碼過程，以計算每一個標記類別的對數概率（log-probabilities）。這一步驟使得模型能夠對每一個時間軸的每一個單詞預測其可能的實體類別標記。(Chiu & Nichols, 2015)

最後，為了生成最終的輸出，前向和後向 LSTM 生成的特徵向量簡單地組合起來。這樣的組合可以捕捉到更豐富的上下文關係，並有助於提高對命名實體的識別。

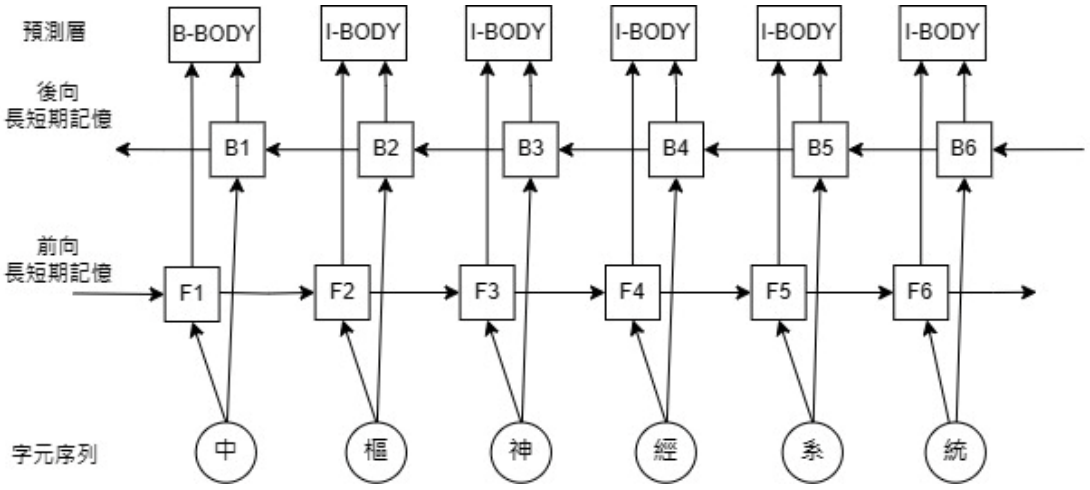


圖 2-3　Bi-LSTM 神經網路

第五節 注意力機制（Attention）

注意力機制（Attention）在深度學習模型中被廣泛運用，被視為一種將資訊從輸入到輸出傳遞的機制。它在處理序列資料、自然語言處理和計算機視覺等領域發揮著關鍵作用。該機制的核心思想是將向量（Query）與一組鍵向量（Key）進行比對，以計算一組相應的值向量（Value）。這種函式的目的是量化查詢與鍵之間的相似度或關聯性。

$$\text{Attention}(\text{Q},\text{K},\text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V} \tag{6}$$

注意力機制的優勢在於它的靈活性，使得模型能夠有針對性地關注輸入中的不同部分，並在不同上下文中取得更好的性能。這在處理長度較長的序列、擁有多重語義的詞語或解決聚焦問題時非常有幫助。因此，注意力機制函式已經成為許多深度學習架構的核心組成部分，使它們能夠更好地處理複雜的任務和大量的資料(Vaswani et al., 2017)。

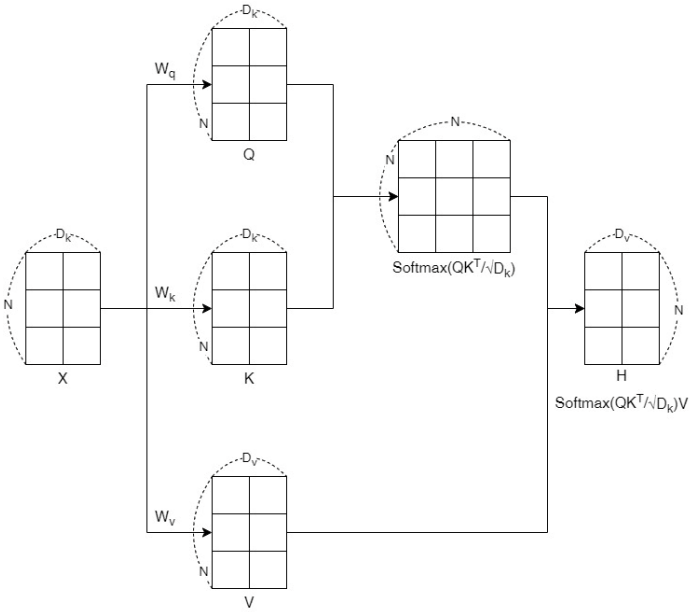


圖 2-4　Attention

第六節 BERT

壹、 BERT & BERT-WWM

BERT（Bidirectional Encoder Representations from Transformers）已在各種自然語言處理任務中展示了其有效性。BERT 透過同時考慮上下文來生成，因此只需要添加一個輸出層(output layer)，就可以對 BERT 模型進行調整，從而建立是用於各任務的模型(Pakhale, 2023)。BERT 主要由兩個預訓練任務組成：遮罩語言模型(Masked Language Model, MLM)和預測下一句任務(Next Sentence Prediction, NSP）(Cui et al., 2021)。

遮罩語言模型（MLM）：隨機對輸入的一些標記進行遮罩，目標是僅基於其上下文來預測原始單詞。

預測下一句任務（NSP）：預測句子 B 是否是句子 A 的下一句。

之後，進一步提出了一種技術，稱為全詞掩碼技術（Whole Word Masking, WWM），用於優化 MLM 任務中的原始遮罩(Cui et al., 2021)。在這種設置中不是隨機選擇 Word Piece 標記進行遮罩，而是一次性遮罩與整個單詞相對應的所有標記(Wu et al., 2016)。

在中文環境中，由於中文字符不是由類似字母的符號組成，因此不再使用 Word Piece 分詞器，而是使用傳統的中文分詞工具（CWS）將文本分成幾個詞。這樣可以在中文中採用 WWM，以單詞作為遮罩的單位(Cui et al., 2021)。

貳、 RoBERTa & RoBERTa-WWM

RoBERTa 是基於原始 BERT 架構進行了微調以加強 BERT 潛力的方法(Liu et al., 2019)。這種方法對 BERT 的各個組成部分進行了仔細比較，包括遮罩策略、輸入格式、訓練步驟等。就會發現訓練時間較長、批次（batch）大小較大、序列長度較長且使用更多資料的訓練方式能夠提高 BERT 的性能。

WWM 也可以應用於 RoBERTa 模型，雖然不再使用預測下一句任務，但仍然使用成對的輸入進行預訓練，這對於文字分類和閱讀理解任務是有益的。(Cui et al., 2021)

第七節 條件隨機場（CRF）

壹、 馬可夫隨機場到條件隨機場

條件隨機場（CRF）是馬爾可夫隨機場（Markov Random Field）的變化，CRF 是假設馬爾可夫隨機場中只有 X 與 Y 兩種變數，X 為固定的，Y 是再給定 X 的條件下的輸出。從此得知，假設 X 與 Y 為隨機變數，$P(Y|X)$是給定 X 時 Y 的條件機率分佈，若隨機變數 Y 是一個馬爾可夫隨機場，則稱$P(Y|X)$為條件隨機場(Sutton et al., 2004)。

貳、 線性條件隨機場

在 CRF 的定義中，並沒有要求 X 與 Y 有相同的維度，但通常都假設輸入 X 和相對應得輸出標籤 Y 為相同維度（n），如公式(7)(X. Yang et al., 2018)：

$$X = (x_1, x_2, \dots, x_n), Y = (y_1, y_2, \dots, y_n) \tag{7}$$

在給定輸入 X 的情況下，相對應的輸出標籤 Y 的條件機率分佈$P(Y|X)$構成條件隨機場，即滿足馬爾可夫性，則稱$P(Y|X)$為線性條件隨機場。當輸入$X = (x_1, x_2, \dots, x_n)$、相對應的輸出標籤$Y = (y_1, y_2, \dots, y_n)$與輸入特徵條件為$\emptyset_c$，公式如下(X. Yang et al., 2018)：

$$P(Y|X) = \frac{1}{Z(X)} \prod_c \emptyset_c(Y, X) \tag{8}$$

$Z(X)$為標準化公式，如下(X. Yang et al., 2018)：

$$Z(X) = \sum_Y \prod_c \emptyset_c(Y_c, X) \tag{9}$$

參、 CRF 的損失函式（Loss Function）定義

在監督分類問題中，目標是最小化訓練期間的預期誤差，可以透過定義一個損失函數 L 來做到這一點，該函數將預測和真實標籤作為輸入，如果它們相等則傳回零分，如果不同則傳回正分，表示錯誤。Sutton et al. (2004)描述了一種反覆運算縮放演算法，能使對數似然目標函式(log-likelihood objective function)最大化(Sutton et al., 2004)：

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{N} \log P_\theta\left(y^{(i)} \mid x^{(i)}\right) \\ &\propto \textstyle\sum_{x,y} \tilde{p}(x, y) \log P_\theta(y \mid x) \end{aligned} \tag{10}$$

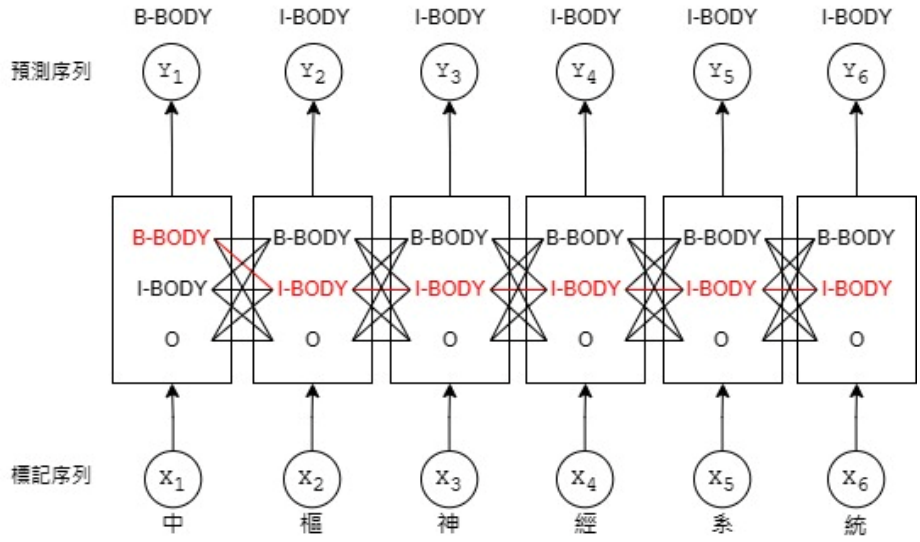


圖 2-5　CRF 神經網路

第八節　卷積神經網路（CNN）

CNN 是一種高效的方法，用於從圖片中提取形態資訊，也可以利用於句子，例如 Zhang et al. (2016)表明可以在不了解單字、片語、句子和任何其他與人類語言有關的句法或語義結構的情況下進行訓練，也能理解文本，並且不僅適用於英語，也適用於中文(Zhang et al., 2015)。CNN 在深度學習模型的多個任務中發揮著關鍵作用，能夠有效地捕捉每一個詞序列的資訊(Kim, 2014)。CNN 可以自動學習不同級別的特徵，這使其在文本分類、命名實體識別、情感分析等任務中表現出色。

圖 2-6 展示了的 CNN 神經網路在句子上的應用，Chiu & Nichols (2016)從每一個單字中提取字元特徵，然後再連接並傳遞到 CNN(Chiu & Nichols, 2016)。Ma & Hovy (2016)與 Chiu & Nichols (2016)中的 CNN 類似，但唯一的不同之處在於使用字元嵌入作為 CNN 的輸入，而不使用字元特徵(Ma & Hovy, 2016)。

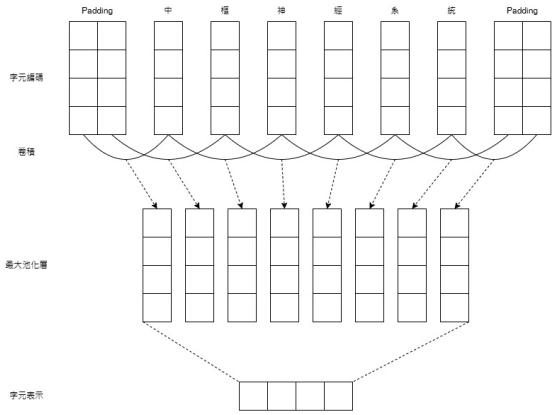


圖 2-6　用於提取單字的字元級表示的 CNN 神經網路

第九節 Mamba

結構化狀態空間序列模型（即 S4, Structured State Spaces for Sequence Modeling）是用於深度學習的一類最新序列模型，與 RNN、CNN 和經典狀態空間模型廣泛相關(Dao & Gu, 2024)。Mamba 模型架構主要依賴 S4，模型的核心是其線性時不變性（LTI），其核心是利用遞迴或選擇性掃描（Selective Scan）有效地將中心遞歸映射到並行 GPU 硬體。模型的重複性使它們能夠在沒有注意力機制的 Q、K、V 的情況下有效地用於生成結果，還使 Mamba 可以隨序列長度線性擴展(Anthony et al., 2024)。

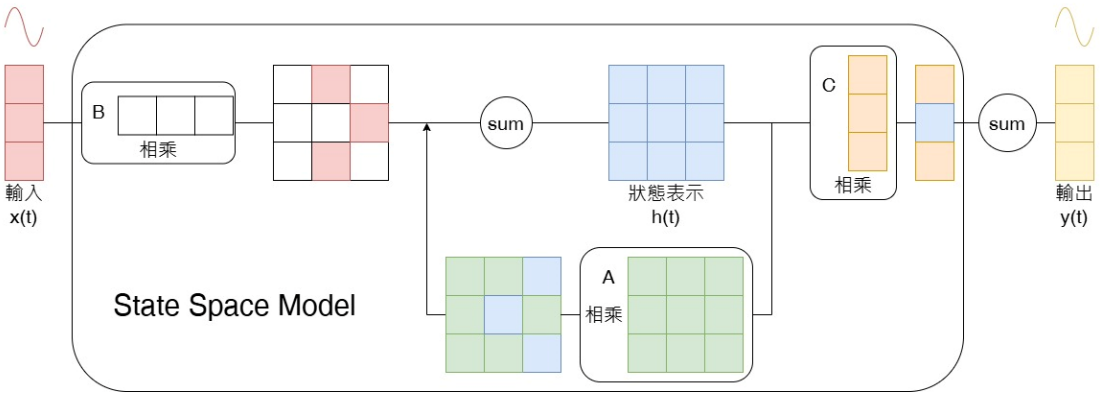


圖 2-7　結構化狀態空間序列模型

壹、 SSM 的狀態方程式與輸出方程式

在 S4 中，假設輸入為 x，狀態空間模型 SSM（A, B, C）和 t 時刻的狀態$h(t)$，A 表示控制 h（t）隨時間變化的矩陣，B 表示控制 x（t）如何與模型交互的動態矩陣，C 的觀察矩陣將狀態轉換為「觀察值」，表示為 y。三個連續參數矩陣 A、B 和 C 將與$h(t)$相互關聯(Gu & Dao, 2023b)，方程式如下：

$$h'(t) = Ah(t) + Bx(t) \tag{11}$$

$$y(t) = Ch(t) \tag{12}$$

貳、 離散化 SSM

由於現實中，一般不會利用連續型資料進行處理，都是離散資料，所以需要對 SSM 進行離散化。離散化是透過固定公式將連續參數轉換為離散參數的關鍵過程，使 S4 模型能夠與連續時間系統保持聯繫，從而增強模型的穩定性和性能(Gu & Dao, 2023b)。方程式如下：

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t \tag{13}$$

$$y_t = Ch_t \tag{14}$$

SSM 在離散化資料上訓練，是仍能學習連續資訊，因為對於 SSM，句子是連續信號的抽樣，或是說信號模型是離散的序列模型的，利用零階保持技術（Zero-order hold technique）就可以處理離散化(Gu & Dao, 2023b)。

每次收到離散信號時，都會保留值，直到收到新的離散信號，如此就可以創建 SSM 能使用的連續信號。保持值可以利用步長（Δ）代表輸入的保持。有了連續的信號後，便可以產生成連續的輸出，並根據輸入的時間步長對值進行抽樣訓練。而這些就是離散輸出，並且可以針對 A、B 利用雙線性方法(Bilinear Method)推導(Gu et al., 2021)：

$$\bar{A} = \left(\mathrm{I} - \Delta/2 \cdot \mathrm{A}\right)^{-1} \cdot ((\mathrm{I} + \Delta/2 \cdot \mathrm{A})) \tag{15}$$

$$\bar{B} = \left(\mathrm{I} - \Delta/2 \cdot \mathrm{A}\right)^{-1} \cdot \Delta \mathrm{B} \tag{16}$$

參、 循環結構

向後傳遞前一個隱藏狀態並重新計算新的隱藏狀態，類似於在 RNN 一樣。在每一個步驟 t 中，將前一個時間h_{t-1}的隱藏狀態與當前輸入x_t相結合，以創建新的隱藏狀態h_t (Gu et al., 2021)。利用之前的離散方程來嘗試計算h_2時：

$$h_0 = \bar{B}x_0 \tag{17}$$

$$h_1 = \bar{A}h_0 + \bar{B}x_1 \tag{18}$$

$$h_2 = \bar{A}\left(\bar{A}h_0 + \bar{B}x_1\right) + \bar{B}x_2 \tag{19}$$

由此可知，我們可以推導出y_2：

$$\begin{aligned} y_2 &= Ch_2 \\ &= C\left(\bar{A}h_1 + \bar{B}x_2\right) \\ &= C\left(\bar{A}\left(\bar{A}h_0 + \bar{B}x_1\right) + \bar{B}x_2\right) \\ &= C\left(\overline{AAB}x_0 + \overline{AB}x_1 + \bar{B}x_2\right) \\ &= C\overline{AAB}x_0 + C\overline{AB}x_1 + C\bar{B}x_2\right) \end{aligned} \tag{20}$$

肆、 卷積結構

在經典的圖像識別中，會使用卷積核（kernels）來產生出其特徵，SSM 也可以以卷積的方式表示。但由於 NER 要處理的是文本，而不是圖像，因此我們需要的是一維卷積(Gu et al., 2021)，而這個卷積核源自 SSM 的公式：

$$\overline{K} = （C\overline{B}, C\overline{AB}, \dots, C\overline{A^k B}, \dots）\tag{21}$$

$$y = x \cdot \overline{K}\tag{22}$$

換個形式，可以發現利用點積的方式也能達成y_2的計算，向量為輸入 x：

$$y_2 = \left(C\overline{AAB} \quad C\overline{AB} \quad C\overline{B}\right)\begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}\tag{23}$$

SSM 表式用卷積的一個主要好處可以像 CNN 一樣進行並行訓練，但因為內核大小式固定的，所以速度不如 RNN。

一個 Mamba 塊可以在兩種模式下運行，第一種模式是遞迴方法，它直接遵循此處描述的步驟。這種方法在單一步驟的記憶體和計算成本上都是線性的，因為它只利用迴圈狀態來預測下一個權杖。第二種方法是引入的“選擇性掃描”操作和內核，一次在整個序列中運行 SSM (Anthony et al., 2024)。

伍、 選擇性掃描（Selective Scan）

Gu & Dao (2023a)建立了一種「選擇性掃描」演算法，目標獲得更好的性能(Gu & Dao, 2023a)。從圖形處理單元（GPU）的角度來看高頻寬記憶體（HBM）有大量空間來保存資料，但速度很慢。靜態隨機存取記憶體（SRAM）速度很快，但無法容納大量資料。

與選擇性掃描不同，標準的掃描操作是將維度是批次大小(batch size)、序列長度(sequence length)、維度大小(dimension size)以及後續張量大小 N 的整個資料放進高頻寬記憶體並進行，因為靜態隨機存取記憶體沒有足夠大的空間處理整個資料，然後就會在運算中花費大量時間。

相比之下，選擇性掃描不需要操作整個資料，而是將整個資料分成維度是批次大小、後續張量大小 N 的張量與維度是批次大小、序列長度、後續張量大小 N 的張量，使用這 2 個張量更新矩陣 A、矩陣 B、矩陣 C 以及 Δ，這些運算會在靜態隨機存取記憶體進行由於處理的資料要小得多，所以縮短了計算時間，本質上是用計算換取記憶體(Gu & Dao, 2023a)。

陸、 長距離依賴問題

在循環結構中發現矩陣 A 捕捉先前狀態資訊建立新狀態（$h_k = \bar{A}h_{k-1} + \bar{B}x_k$，當 k=2 時，$h_2 = \bar{A}h_1 + \bar{B}x_2$）。由於矩陣 A 只記住幾個 token 和先前狀態的每一個 token 之間的差異，特別是循環表示的上下文中。

為了保留比較長的記憶先前資訊的方式建立矩陣 A，可以使用 Hippo（High-order Polynomial Projection Operator）解決在有限的儲存空間有效處理序列模型的長距離依賴問題。通過函數逼近產生矩陣 A 的最優解，公式如下(Gu et al., 2020)：

$$A_{nk}\begin{cases}(2n+1)^{\frac{1}{2}}(2k+1)^{\frac{1}{2}} & if\ n>k \quad (24)\\ n+1 & if\ n=k \quad (25)\\ 0 & if\ n<k \quad (26)\end{cases}$$

由於 Hippo 矩陣可以產生一個隱藏狀態來記住以前的資訊，使得在被應用循環結構和卷積結構中，可以處理長距離依賴性。

柒、 多層感知器（Multi-Layer Perceptron, MLP）

多層感知器（MLP）在神經網路架構中極為常見。它們是前向神經網路，其中一層的每一個神經元都連接到前一層的每一個神經元。MLP 可以被看作是一個有向圖(Directed graph)，由多個的節點層所組成，每一層都全連接到下一層。

捌、 GRU（Gated Recurrent Unit）

GRU 是一種 LSTM 的變化，在每一次輸入時會透過重置門（reset gate）和更新門（update gate）為資料輸入時添加控制。重置門確定需要重置來自先前隱藏狀態的資訊量，而更新門控制更新其隱藏狀態的程度(Z. Yang et al., 2016)。

玖、 Mamba 區塊（Mamba Block）

Mamba 區塊是結合 Hippos 與門控多層感知器（Gated MLPs）在一起。Mamba 從門控多層感知器取得門控功能，然後將 Mamba 與卷積與選擇性 SSM 變換結合。現在有了一個新的區塊結構，它可以透過門控機制傳遞輸入的某些部分，然後也可以透過選擇性 SSM 關注輸入的某些部分(Gu & Dao, 2023a)。

由於Mamba沒有注意力機制，所以不需要擔心更大的輸入大小帶來的影響。無論輸入的序列長度如何，傳遞的狀態都是相同的大小，雖然更大的輸入序列長度需要更多的計算，但它只會以線性速率增加。

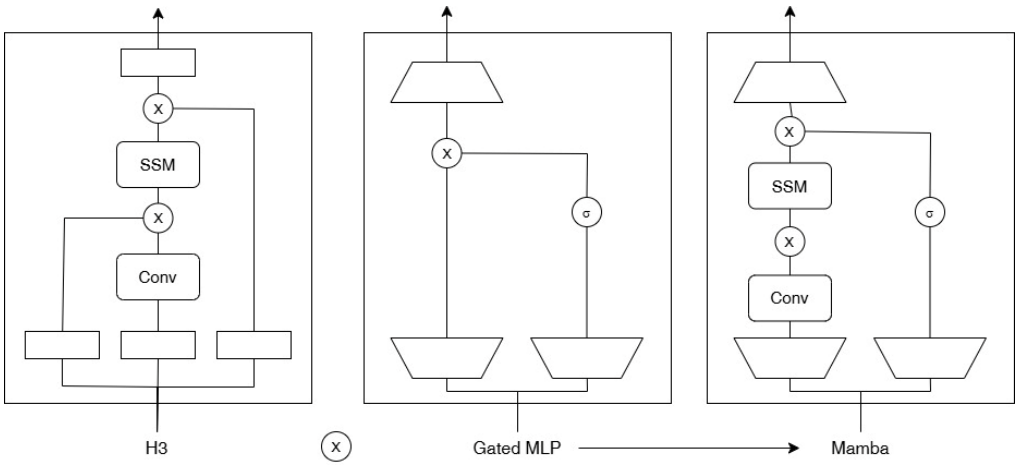


圖 2-8 Mamba 區塊

第十節 評估指標

本研究使用 Micro-f1 和 Macro-f1 指標來評估模型。假設得到的TP_i、TN_i、FP_i和FN_i分別為類別 i={1,...,z}的真陽性、真陰性、假陽性和假陰性計數，z 是類別的數量，類別 i 的精確度（P_i）、召回率（R_i）和 F1 分數（$F1_i$）定義如下：

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{27}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{28}$$

$$F1_i = \frac{2P_i * R_i}{P_i + R_i} \tag{29}$$

此外，微觀平均精確度（P_{Micro}）、微觀平均召回率（R_{Micro}）、宏觀平均精確度（P_{Macro}）和宏觀平均召回率（R_{Macro}）計算如下：

$$P_{Micro} = \frac{\sum_1^z TP_i}{\sum_1^z (TP_i + FP_i)} \tag{30}$$

$$R_{Micro} = \frac{\sum_1^z TP_i}{\sum_1^z (TP_i + FN_i)} \tag{31}$$

$$P_{Macro} = \frac{\sum_1^z \{\frac{TP_i}{TP_i + FP_i}\}}{z} = \frac{\sum_1^z P_i}{z} \tag{32}$$

$$P_{Macro} = \frac{\sum_1^z \{\frac{TP_i}{TP_i + FN_i}\}}{z} = \frac{\sum_1^z R_i}{z} \tag{33}$$

多類別的分類任務整體通常由微觀平均 F1 值（Micro-F1）和宏觀平均 F1 值（Macro-F1）來評估，可規定如下：

$$Micro - F1 = \frac{2P_{Micro} * R_{Micro}}{P_{Micro} + R_{Micro}} \tag{34}$$

$$Macro - F1 = \frac{2P_{Macro} * R_{Macro}}{P_{Macro} + R_{Macro}} \tag{35}$$

這些指標提供了評估模型在不同類別以及整體性能方面的資訊。Micro-F1 關注在所有類別的整體性能，而 Macro-F1 關注了所有類別的平均性能，對於不平衡的資料集將會特別有用。(Yuan et al., 2021)

第三章 研究方法

第一節 使用資料集

本研究將使用 2 種類別的任務，分別是中文醫療命名實體辨識與 CoNLL-2003。

壹、 中文醫療命名實體辨識

針對醫療領域 NER 的任務。表格 3-1 和表格 3-2 為中文醫療命名實體辨識資料集命名實體類型的描述以及數量。有三種類型：

- 正式文本：這包括健康新聞和由專業編輯或記者撰寫的文章。
- 社交媒體：這包含來自醫療問答論壇中擁擠使用者的文本。
- 維基百科文章:這本免費的在線百科全書包括由全球志願者創建和編輯的文章對於這個中文醫療保健命名實體辨識任務。

舉例來說，輸入為「抑酸劑，又稱抗酸劑，抑制胃酸分泌，緩解燒心。」，「抑酸劑」、「抗酸劑」以及「胃酸」屬於在人體中發現的基本化學元素，故為化學（CHEM）類別，「燒心」是“胃食道逆流症”的口語，屬於由感染或健康失敗而不是事故引起的人或動物疾病，故為疾病（DISE）類別，則輸出為「B-CHEM, I-CHEM, I-CHEM, O, O, O, B-CHEM, I-CHEM, I-CHEM, O, O, O, B-CHEM, I-CHEM, O, O, O, O, O, B-DISE, I-DISE, O」。

表 3-1　中文醫療命名實體辨識資料集命名實體類型

實體類型	描述
身體（BODY）	形成人或動物的整個物理結構，包括生物細胞、組織、器官和系統。
症狀（SYMP）	由特定疾病引起的任何疾病或身體或精神變化的感覺。
設備（INST）	用於執行特定醫療任務（如診斷和治療）的工具或其他設備。
測試（EXAM）	仔細觀察或檢查某物以發現可能的疾病的行為。
化學（CHEM）	通常在人體中發現的任何基本化學元素。
疾病（DISE）	由感染或健康失敗而不是事故引起的人或動物疾病。
藥品（DRUG）	任何用作藥物的天然或人工製造的化學品。
補充物（SUPP）	添加到其他東西中以改善人類健康。
治療（TREAT）	一種用於治療疾病的行為方法。
時間（TIME）	以分鐘、天、年為單位的存在元素。

資料來源：(Lee et al., 2023)

表 3-2　中文醫療命名實體辨識資料集數量

	資料集		
類型	正式文本	社交媒體	維基百科文章
句子	23,008	7,684	3,205
字元	1,109,918	403,570	118,116
實體	42,070	26,390	13,369

資料來源：(Lee et al., 2023)

貳、 CoNLL-2003

CoNLL-2003 提供有關英文與德文資料集，對參與任務的模型進行了總體概述，並討論了它們的性能(Tjong Kim Sang & De Meulder, 2003)。專注於四種類型的命名實體：人名（persons）、地名（locations）、組織（organizations）以及其他實體（miscellaneous names）。表 3-3 為英文每一個資料集中命名實體數的概述，表 3-4 為德文每一個資料集中命名實體數的概述。

表 3-3　CoNLL-2003 英文每一個資料集中命名實體數

	人名（PER）	地名（LOC）	組織（ORG）	其他實體（MISC）
訓練集	6600	7140	6321	3438
驗證集	1842	1837	1341	922
測試集	1617	1668	1661	702

資料來源：(Tjong Kim Sang & De Meulder, 2003)

表 3-4　CoNLL-2003 德文每一個資料集中命名實體數

	人名（PER）	地名（LOC）	組織（ORG）	其他實體（MISC）
訓練集	2773	4363	2427	2288
驗證集	1401	1181	1241	1010
測試集	1195	1035	773	670

資料來源：(Tjong Kim Sang & De Meulder, 2003)

第二節 模型架構

本研究旨在探討如何利用 Mamba 與 CRF 結合來提高模型的性能。之所以利用 Mamba 模型架構是因為 Gu & Dao （2023）展現出處理序列資料集時有更好的準確率。在訓練過程中，隨著序列長度的增加，計算量和記憶體需求也會相應增加，但這種增長是線性的，而不是呈指數級增長。這使得 Mamba 模型架構能夠有效地處理大型資料集和長序列，而無需過多擔心性能下降。

另外，在推理過程中，Mamba 模型架構的另一優勢在於它的高效率。因為在推理時不需要儲存以前的元素，每一步的計算時間是固定的，不會隨著序列長度的增加而增加。這意味著即使處理大型輸入資料，Mamba 模型架構也能夠提供快速而穩定的推理性能。這些條件可以在訓練資料時被 CRF 自動學習得到。圖 3-1 為模型架構圖。

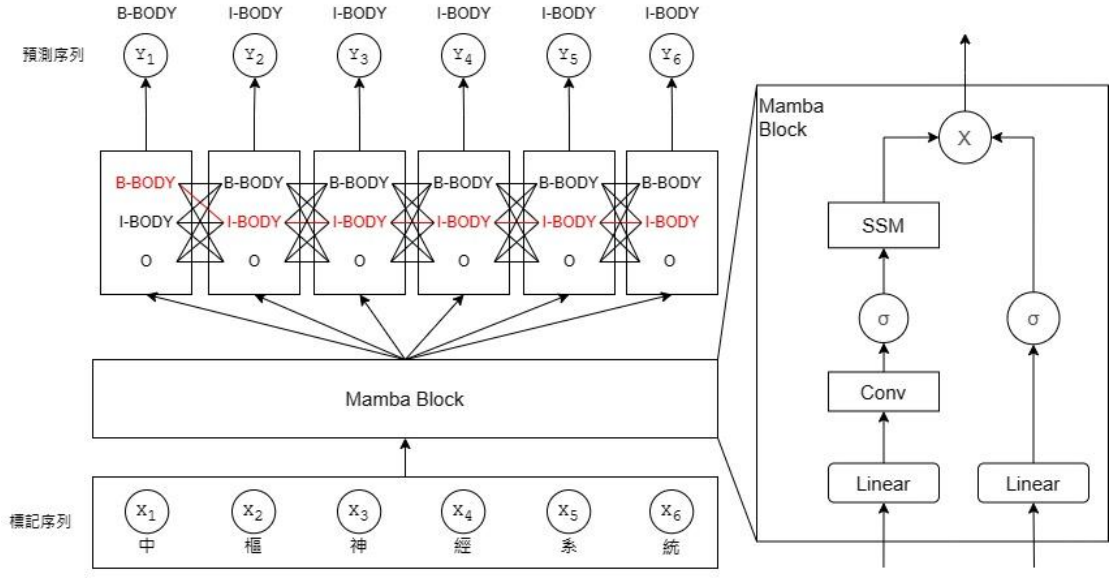


圖 3-1　模型架構圖

第三節 研究流程

本研究的研究流程主要包含五個章節，依序為環境建置、資料預處理、模型訓練、模型測試、分析結果，如圖 3-2 所示。

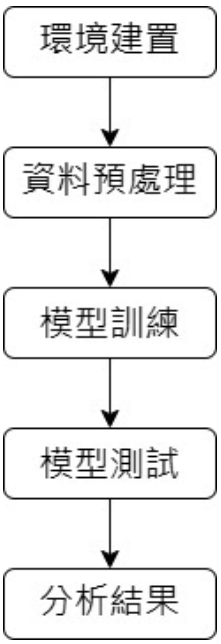


圖 3-2　研究流程

壹、 環境建置

Mamba 對於環境建置有以下需求，使用的電腦為 Linux 環境並需要有 NVIDIA GPU。需要安裝 PyTorch 版本 1.12 以上、CUDA 版本 11.6 以上、causal-conv1d 版本 1.4.0 以及 mamba-ssm。

貳、 資料預處理

將從中文醫療命名實體辨識資料集獲取的資料，將要訓練的輸入與標籤整理。舉例來說：整段句子為「如何治療胃食道逆流症？」，本研究將訓練用輸入變成

「如, 何, 治, 療, 胃, 食, 道, 逆, 流, 症, ？」與其相對應標籤「O, O, O, O, B-DISE, I-DISE, I-DISE, I-DISE, I-DISE, I-DISE, O」，並將所有資料集單獨分成訓練集、驗證集以及測試集。

參、 模型訓練

Mamba 模型可以對 Mamba 參數設定（Mamba Config）進行一些調整，本研究沒有對參數設定過多的調整，以下為本研究對 Mamba 模型的參數設定：隱藏層維度（d _ model）為 2560；模型層數（n _ layer）為 64；詞彙表大小（vocab _ size）為 50277；狀態空間維度（d _ state）為 16；卷積核維度（d _ conv）為 4，迭代（Epoch）為 5，訓練批次大小(train _ batch _ size)為 64，資料類型皆為整數(int)型態。學習率(learning _ rate)為 5e-5，資料類型皆為浮點(float)型態。並使用 AdamW 作為學習優化器(optimizer)。

肆、 模型測試

本研究中文醫療命名實體辨識分成 3 個部分：訓練、驗證以及測試。訓練與測試是中文醫療命名實體辨識資料集(Lee & Chen, 2022)的訓練集與測試資料集，而驗證會使用 ROCLING 2022 中文醫療命名實體辨識資料集(Lee & Chen, 2022)的驗證集。

CoNLL-2003 訓練、驗證以及測試皆是使用 CoNLL-2003 的訓練集。

伍、 分析結果

對於模型的所有的實驗結果進行分析，所有實驗結果會放在本論文第四章研究結果的實驗結果裡。

第四章 研究結果

第一節 模型訓練結果

在第一次 Mamba 訓練結果（表 4-1），本研究選則迭代次數(Epoch)10 次作為訓練，可以發現在訓練中第一次迭代就有超過 89%的準確率，後續也只在 90%的準確率，沒有明顯的提高了，在第三次迭代訓練損失就已經達到最低的效果，從第四迭代訓練損失到第十迭代訓練損失都沒有第三迭代訓練損失還要低，故而可以在第三次迭代停止訓練。

為了驗證本研究能停止迭代在第三次的想法，所有參數設定只有修改迭代次，其餘沒有進行任何更動，進行第二次的 Mamba 訓練，完成訓練後依照第二次 Mamba 訓練結果（表 4-2）來看，訓練損失的最低值仍然是第三次迭代，準確率也沒有隨著每一次的迭代有明顯的提升，而每次的訓練集資料以及驗證集資料是相同的，因為在訓練前已經區分出來的，成功驗證本研究能停止迭代在第三次的想法。

在表 4-3　各模型在迭代次數為五次的驗證準確率（Validation Accuracy）中發現除了 Mamba 結合 CRF 的模型與 Bert + Bi-LSTM + CRF(ROCLING 2023)模型的驗證準確率沒有再提高，相較於其他比較模型有在緩慢的提升，但因為研究為了相同的基準，基準是以五次迭代，故而停止了訓練，持續增加迭代次數是有增加模型性能的可能性。

在 Mamba 結合 CRF 的模型與各模型在中文醫療命名實體辨識資料集的訓練損失（Training Loss）與驗證損失（Validation Loss）（圖 4-1）中，x 軸為迭代

次數，y 軸為損失值。可以發現 Mamba 結合 CRF 的模型驗證損失在第三次迭代轉從原本的下降轉變成上升，發生過擬合（overfitting）的情況(Ying, 2019)。表 4-1 中也可以發現在第 7 次迭代以及表 4-2 中在第 4 次迭代也發生訓練損失上升的情況。相反地，圖 4-1 中發現除了 Mamba 結合 CRF 的模型，Bert + Bi-LSTM + CRF(ROCLING 2023)模型在第 5 次迭代時也有往上的趨勢，但限制於實驗只有跑五次，不能確定是否發生同樣的情況，其餘模型就沒有發生任何上漲趨勢。

為了解決過擬合的情形，本研究針對 Mamba 結合 CRF 的模型進行參數調整，(調整內容)(調整結果)。

表 4-1　第一次 Mamba 訓練結果

迭代次數 （Epoch）	訓練損失 (Training Loss)	驗證損失 (Validation Loss)	準確率 (Accuracy)
1	33.01%	34.42%	89.93%
2	27.86%	33.52%	90.15%
3	21.89%	33.44%	90.10%
4	27.81%	33.13%	90.16%
5	28.28%	32.94%	90.19%
6	26.66%	32.80%	90.14%
7	37.67%	32.63%	90.26%
8	53.77%	32.58%	90.27%
9	35.02%	32.56%	90.24%
10	30.78%	32.56%	90.27%

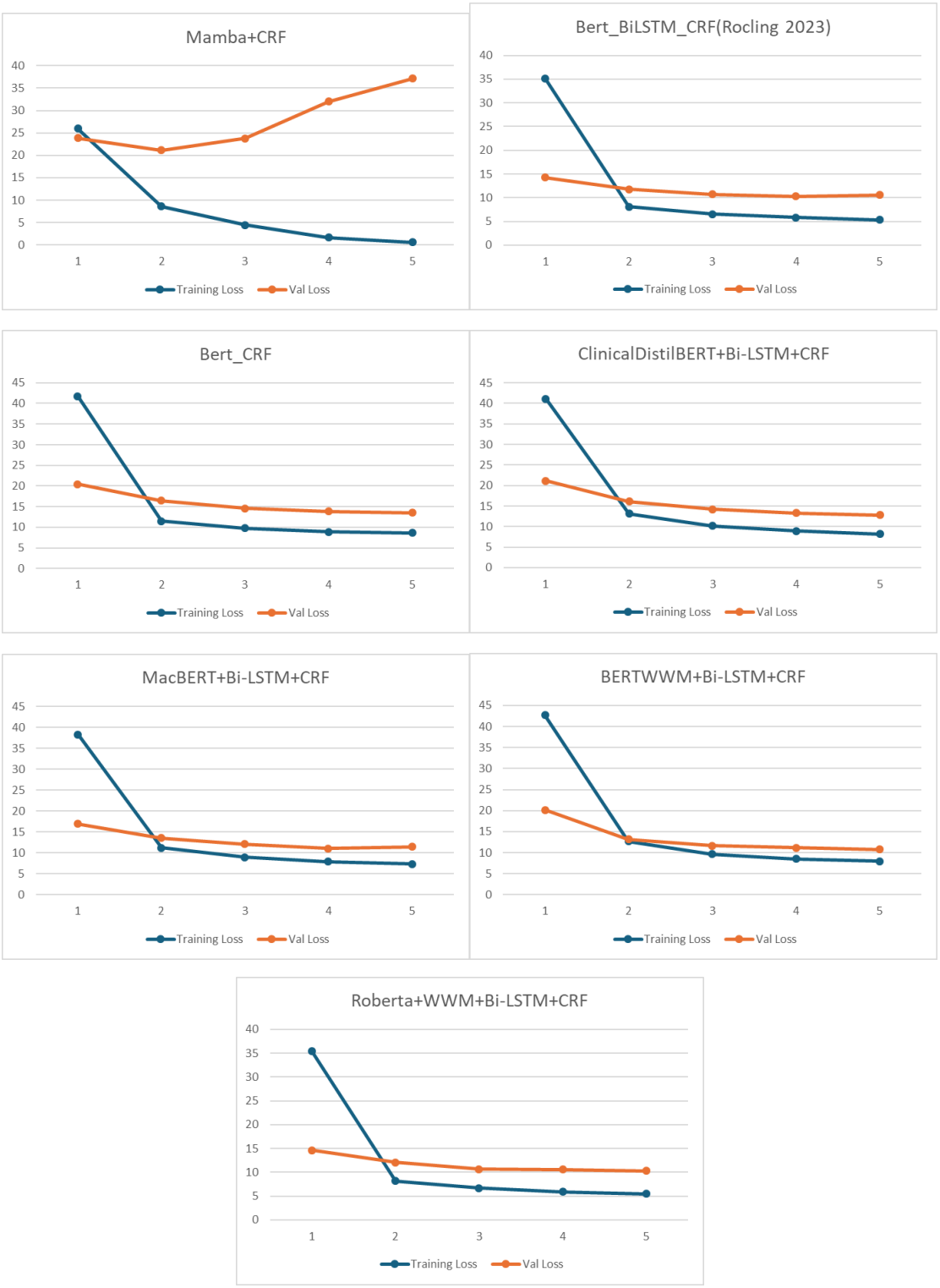
表 4-2　第二次 Mamba 訓練結果

迭代次數 （Epoch）	訓練損失 (Training Loss)	驗證損失 (Validation Loss)	準確率 (Accuracy)
1	40.36%	36.98%	89.61%
2	27.07%	35.17%	89.92%
3	24.31%	34.55%	90.01%
4	48.50%	34.06%	89.98%
5	41.35%	33.63%	90.16%

表 4-3　各模型在迭代次數為五的驗證準確率（Validation Accuracy）

模型 (Model)	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Mamba + CRF	81.92%	83.18%	82.93%	82.93%	81.80%
Bert + Bi-LSTM + CRF （ROCLING 2023）	88.18%	89.20%	89.65%	89.77%	89.50%
Bert + CRF	82.49%	84.64%	85.43%	85.79%	85.90%
ClinicalDistil BERT + Bi-LSTM + CRF	82.38%	84.96%	86.05%	86.33%	86.73%
Mac BERT + Bi-LSTM + CRF	85.87%	87.84%	87.87%	88.48%	88.05%
BERTWWM + Bi-LSTM + CRF	84.22%	87.81%	88.62%	88.61%	88.83%
Roberta + WWM+ Bi-LSTM + CRF	87.50%	88.80%	89.72%	89.38%	89.66%

圖 4-1　Mamba 結合 CRF 的模型與各模型在中文醫療命名實體辨識資料集的訓練損失（Training Loss）與驗證損失（Validation Loss）



第二節 中文醫療命名實體辨識資料集實驗結果

在 Mamba 模型分類指標結果中（表 4-4），微觀平均 F1 值有 86.3%，但是仔細去細看每一個實體類型的時候，需要特別注意的部分是大多數 F1-Score 主要源於「O」類標籤，即非實體標記資料，這也是該資料集中最常見的類型。甚至是「B-INST」實體類別標籤、「B-DRUG」實體類別標籤、「B-TREAT」實體類別標籤以及「I-TIME」實體類別標籤不管是準確率、召回率還是 F1 值都是 0。

在加入 CRF 的 Mamba 模型中（表 4-5），顯示了在每一個標籤都有性能提升。這向改變在藥品（DRUG）、疾病（DISE）、時間(TIME)等實體類別標籤有特別明顯，原本在 Mamba 模型中 F1 值為 0 的實體類別標籤也都有很好的提升。舉例來說，「I-DRUG」實體類別標籤從未加入 CRF 的模型（3%）到加入 CRF 的模型（62.2%），將近 60%的漲幅，而「I-SUPP」實體類別標籤也從 27.6%漲幅到 76.2%。微觀平均 F1 值也從 86.3%增長突破 90%，來到了 91.9%，然而宏觀平均 F1 值從 23.8%增長到 55.1%，比想像中的還要差，但也是有增長。Mamba 結合 CRF 的模型在其他類型的性能提升也非常顯著，這表明了 CRF 對於多類型實體識別的有效性。

相對的在某些其他特定實體類別，如治療（TREAT）和設備（INST）的性能改進非常有限，幾乎沒有可能沒有明顯提升。舉例來說，「B-INST」實體類別標籤從未加入 CRF 的模型（0%）到加入 CRF 的模型（23.3%），雖然成長 23.3%，但因為 F1 值還是太低，不能作為良好的訓練結果。這可能是由於這些實體類別的樣本稀少性對模型在訓練時造成語義模糊性影響和性能的優劣，像是剛剛舉例的「B-INST」實體類別標籤在所有字超過 10000 個字裡只有少少的 42 個字。

在中文醫療命名實體辨識資料集與各模型使用微觀平均精確度比較(表 4-6)以及中文醫療命名實體辨識資料集與各模型使用宏觀平均精確度比較(表 4-7)中，呈現了不同模型在兩種平均精確度下的性能指標。

可以看出，Mamba 結合 CRF 的模型在微觀平均精確度下表現最好，其準確率、召回率與 F1 值均達到 91.9%，其次就是 Mamba 模型（86.3%），Bert + Bi-LSTM + CRF（ROCLING 2023）模型作為本次研究的基準模型也有 79.6%的 F1 值型，其餘模型只有 Roberta + WWM+ Bi-LSTM + CRF 模型（79.9%）高於基準模型，其他都略低於基礎模型。

然而，在宏觀平均精確度下，Mamba 結合 CRF 的模型的表現則沒有那麼突出，準確率、召回率與 F1 值都較低，分別為 66.3%、48.2%和 55.1%，而 Mamba 模型更加慘烈，準確率、召回率與 F1 值分別為 34.8%、20.9%和 23.8%。相比之下，Roberta + WWM+ Bi-LSTM + CRF 模型是比較中的有最佳的結果，其準確率、召回率與 F1 值分別為 76.5%、67.9%和 71.5%，其次就是作為基準的 Bert + Bi-LSTM + CRF(ROCLING 2023)模型在宏觀平均精確度下準確率、召回率與 F1 值的表現分別為 77.3%、67.3%和 68.1%。

表 4-4　Mamba 模型分類指標結果

標籤 (Label)	準確率 (Accuracy)	召回率 (Recall)	F1 值 (F1-score)	數量
O	89.7%	98.4%	93.8%	99473
B-BODY	37.2%	35.1%	36.1%	3171
I-BODY	54.1%	22.7%	32.0%	4077
B-SYMP	64.8%	12.3%	20.7%	1481
I-SYMP	57.5%	19.6%	29.2%	2096
B-INST	0.0%	0.0%	0.0%	42
I-INST	12.5%	1.1%	2.0%	93
B-EXAM	30.4%	3.5%	6.2%	404
I-EXAM	68.9%	33.1%	44.7%	1074
B-CHEM	43.5%	35.3%	39.0%	744
I-CHEM	59.4%	39.4%	47.4%	1634
B-DISE	23.2%	5.2%	8.5%	1005
I-DISE	60.8%	20.8%	31.0%	2438
B-DRUG	0.0%	0.0%	0.0%	79
I-DRUG	17.6%	1.7%	3.0%	180
B-SUPP	33.6%	41.8%	37.2%	122
I-SUPP	32.0%	24.3%	27.6%	301
B-TREAT	0.0%	0.0%	0.0%	203
I-TREAT	35.5%	25.5%	29.7%	337
B-TIME	9.4%	18.5%	12.5%	54
I-TIME	0.0%	0.0%	0.0%	89
micro avg	-	-	86.3%	19624
macro avg	34.8%	20.9%	23.8%	19624
weighted avg	83.2%	86.3%	83.6%	19624

表 4-5　Mamba 結合 CRF 的模型分類指標結果

標籤 (Label)	準確率 (Accuracy)	召回率 (Recall)	F1 值 (F1-score)	數量
O	94.6%	98.4%	98.4%	99473
B-BODY	66.4%	61.8%	64.0%	3164
I-BODY	76.0%	67.9%	71.7%	4063
B-SYMP	70.3%	48.5%	57.4%	1481
I-SYMP	80.9%	50.0%	61.8%	2096
B-INST	38.9%	16.7%	23.3%	42
I-INST	48.1%	14.0%	21.7%	93
B-EXAM	64.5%	46.3%	53.9%	404
I-EXAM	85.7%	64.2%	73.4%	1074
B-CHEM	62.8%	50.8%	56.2%	744
I-CHEM	80.0%	71.2%	75.3%	1634
B-DISE	63.3%	44.8%	52.4%	1005
I-DISE	82.1%	63.1%	71.4%	2438
B-DRUG	53.5%	29.1%	37.7%	79
I-DRUG	79.3%	51.1%	62.2%	180
B-SUPP	63.3%	46.7%	53.8%	122
I-SUPP	84.0%	69.8%	76.2%	301
B-TREAT	31.6%	18.2%	23.1%	203
I-TREAT	44.2%	28.5%	34.7%	337
B-TIME	56.7%	31.5%	40.5%	54
I-TIME	65.5%	40.4%	50.0%	89
micro avg	91.9%	91.9%	91.9%	119076
macro avg	66.3%	48.2%	55.1%	119076
weighted avg	91.1%	91.9%	91.3%	119076

表 4-6　各模型在中文醫療命名實體辨識資料集使用微觀平均精確度的比較

模型 (Model)	準確率 (Accuracy)	召回率 (Recall)	F1 值 (F1-score)
Mamba + CRF	91.9%	91.9%	91.9%
Mamba	86.3%	86.3%	86.3%
Bert + Bi-LSTM + CRF（ROCLING 2023）	81.1%	78.2%	79.6%
Bert + CRF	77.3%	66.0%	71.2%
Roberta + WWM+ Bi-LSTM + CRF	81.3%	78.5%	79.9%

表 4-7　各模型在中文醫療命名實體辨識資料集使用宏觀平均精確度的比較

模型 (Model)	準確率 (Accuracy)	召回率 (Recall)	F1 值 (F1-score)
Mamba + CRF	66.3%	48.2%	55.1%
Mamba	34.8%	20.9%	23.8%
Bert + Bi-LSTM + CRF（ROCLING 2023）	77.3%	67.3%	68.1%
Bert + CRF	68.2%	50.9%	57.2%
Roberta + WWM+ Bi-LSTM + CRF	76.5%	67.9%	71.5%

第三節 CoNLL-2003 資料集實驗結果

在 CoNLL-2003 命名實體資料集與各模型比較（表 4-8）中，將對 Mamba 結合 CRF 的模型與各種先進的命名實體識別模型進行分析和比較，這些模型的性能通過在 CoNLL-2003 命名實體識別資料集上的評估結果來衡量其性能。

從第二節中文醫療命名實體辨識資料集實驗結果可以看出在中文醫療命名實體辨識資料集中 Mamba 結合 CRF 的模型有相對好的性能，但是將資料集轉換成 CoNLL-2003 資料集時，性能卻沒有了獨特的優勢，F1 值為 85.2%。

從表 4-5 中可以發現 Chiu & Nichols (2016) 的 Bi-LSTM-CNNs 模型與 Ma & Hovy (2016) 的 Bi-LSTM-CNNs-CRF 模型的比較與 Lample et al. (2016)的 Bi-LSTM 模型與 Lample et al. (2016)的 Bi-LSTM-CRF 模型的比較，應證了 CRF 的結合可以增加模型的性能。令人有趣的地方是，排除多個神經網路(Neural Network, NN)結合 CRF 的模型，在所有單純某一個神經網路結合 CRF 的模型裡，是 Z. Yang et al. (2017) 的門控迴圈單元(Gated Recurrent Unit, GRU)結合 CRF 模型是最高的 F1 值為 91.2%。

Peters et al., (2017)建立四個不同模型，分別是使用預訓練模型的詞嵌入、語言模型的詞嵌入 (LM embedding) 、基於訓練十億文字資料集的模型 (1B word dataset) 以及同樣使用基於訓練十億文字資料集的模型，但將前向語言模型輸入字元、LSTM 隱藏單元以及投影層 (projection layer) 做了變化，改成 4096 個輸入字元、8192 個隱藏單元、1024 層投影層，並添加後向語言模型。可以發現改變詞嵌入不會造成太大的影響，而添加後向語言模型可以提高性能。

表 4-8 CoNLL-2003 命名實體資料集與各模型 F1-Score 比較

模型	F1-score
Mamba + CRF	85.2%
Collobert et al., (2011)	88.6%
Huang et al., (2015)	84.2%
Chiu & Nichols, (2016)	83.2%
Ma & Hovy, (2016)	91.3%
Lample et al., (2016) （Bi-LSTM）	89.1%
Lample et al., (2016) （Bi-LSTM + CRF）	90.9%
Hu et al., (2016)	91.1%
Z. Yang et al., (2016) （GRU+CRF no word embedding）	77.2%
Z. Yang et al., (2016) （GRU+CRF no char GRU）	88.0%
Z. Yang et al., (2016) （GRU+CRF no gazetteer）	90.9%
Z. Yang et al., (2016)	91.2%
Rei, (2017)	87.3%
Strubell et al., (2017)	90.5%
Z. Yang et al., (2017)	91.2%
Peters et al., (2017) （word embedding）	90.8%
Peters et al., (2017) （LM embedding）	90.7%
Peters et al., (2017) （1B word dataset）	91.6%
Peters et al., (2017) （1B word dataset+4096-8192-1024）	91.9%

第五章 結論

近年來，因為深度學習技術的迅速發展，在許多的領域中有顯著的進展，而NER 隨著語言模型與 Transformer 序列長度和模型增加或變化，計算量也跟著成指數成長。Mamba 模型的到來，改變 Transformer 的架構，Mamba 繞過 self-attention 所造成的二次縮放，在未來一定有類似於 Mamba 模型這樣的 SSM 架構，會比現在的更新穎更強大，也可以利用這些模型增加 NER 的性能。

從中文醫療命名實體辨識資料集來看，會發現資料集之間實體類別數量差距太大，某些實體類別太稀少，舉例來說「設備」的實體類別數量總共也不超過 200 個，而有的實體類別的比例就佔多數，在未來的研究中可以添加資料集中所缺乏的實體類別進行強化訓練，期望可以增加 Mamba 模型對於此領域的性能。

從研究成果可以得知在中文醫療命名實體辨識資料集中 Mamba 模型有比其他模型的性能效果更加顯著，但只是改變資料集，會發現並沒有當初看到 Mamba 模型的驚豔，而是與其他模型來的弱小。NLP 目前的研究中大多都是利用 LSTM、CNN 等等神經網路結合 Transformer 架構的方法，本研究利用 Mamba 模型結合 CRF 模型，也成功提高模型與其他模型的抗衡能力。在未來中也希望有相似方向的研究者可以利用此模型在一些公開的標準資料集（如 CoNLL-2003）上的提升整體性能。

本論文之貢獻為以下部分，第一項貢獻是提出將 Mamba 模型架構結合 CRF 的新模型，第二項貢獻是與 LSTM、Transformer 相關的模型進行比較，並提出 Mamba 結合 CRF 的優點，本研究使用 Mamba 模型架構結合 CRF 的模型，在解決 Transformer 的運算效率低下的問題，並利用 CRF 加強對序列的依賴關係，對

於不同的模型在相同的資料集有不一樣的結果，並與 LSTM、Transformer 相關的模型進行比較後取得更好的結果(91.9%)。

參考文獻

Anthony, Q., Tokpanov, Y., Glorioso, P., & Millidge, B. (2024). *BlackMamba: Mixture of Experts for State-Space Models*. http://arxiv.org/abs/2402.01771

Baviskar, V., Verma, M., Chatterjee, P., & Singal, G. (2023). Efficient Heart Disease Prediction Using Hybrid Deep Learning Classification Models. *IRBM*, *44*(5), 100786. https://doi.org/10.1016/j.irbm.2023.100786

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. http://arxiv.org/abs/2005.14165

Chiu, J. P. C., & Nichols, E. (2015). *Named Entity Recognition with Bidirectional LSTM-CNNs*. http://arxiv.org/abs/1511.08308

Chiu, J. P. C., & Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, *4*, 357–370. https://doi.org/10.1162/tacl_a_00104

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, *12*, 2493–2537. http://arxiv.org/abs/1103.0398

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., & Bai, X. (2019). Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–5. https://doi.org/10.1109/CISP-BMEI48845.2019.8965823

Dao, T., & Gu, A. (2024). *Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality*. http://arxiv.org/abs/2405.21060

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Eickhoff, C., Kim, Y., & White, R. W. (2020). Overview of the Health Search and Data Mining (HSDM 2020) Workshop. *Proceedings of the 13th International Conference on Web Search and Data Mining*, *2551*, 901–902. https://doi.org/10.1145/3336191.3371879

Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65. https://doi.org/10.18653/v1/D19-1006

Gu, A., & Dao, T. (2023a). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. https://arxiv.org/abs/2312.00752

Gu, A., & Dao, T. (2023b). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. https://github.com/state-spaces/mamba.

Gu, A., Dao, T., Ermon, S., Rudra, A., & Re, C. (2020). *HiPPO: Recurrent Memory with Optimal Polynomial Projections*. https://api.semanticscholar.org/CorpusID:221150566

Gu, A., Goel, K., & Ré, C. (2021). Efficiently Modeling Long Sequences with Structured State Spaces. *Processing and Chinese Computing, NLPCC 2016*. http://arxiv.org/abs/2111.00396

Han, R., Peng, T., Yang, C., Wang, B., Liu, L., & Wan, X. (2023). *Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors*. http://arxiv.org/abs/2305.14450

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. http://arxiv.org/abs/1508.01991

Hu, Z., Ma, X., Liu, Z., Hovy, E., & Xing, E. (2016). Harnessing Deep Neural Networks with Logic Rules. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2410–2420. https://doi.org/10.18653/v1/P16-1228

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on Named Entity Recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, *3*, 100017. https://doi.org/10.1016/j.nlp.2023.100017

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270.

https://doi.org/10.18653/v1/N16-1030

Lee, L.-H., & Chen, C.-Y. (2022). Overview of the ROCLING 2022 Shared Task for Chinese Healthcare Named Entity Recognition. *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)* . https://aclanthology.org/2022.rocling-1.46

Lee, L.-H., Lin, T.-M., & Chen, C.-Y. (2023). Overview of the ROCLING 2023 Shared Task for Chinese Multi-genre Named Entity Recognition in the Healthcare Domain. *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*. https://aclanthology.org/2023.rocling-1.42

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://api.semanticscholar.org/CorpusID:198953378

Ma, X., & Hovy, E. (2016). *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*. http://arxiv.org/abs/1603.01354

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems* . https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Pakhale, K. (2023). *Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges*. https://api.semanticscholar.org/CorpusID:262465984

Peters, M., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1756–1765. https://doi.org/10.18653/v1/P17-1161

Praful Bharadiya, J. (2023). A Comprehensive Survey of Deep Learning Techniques Natural Language Processing. *European Journal of Technology*, *7*(1), 58–66. https://doi.org/10.47672/ejt.1473

Rei, M. (2017). Semi-supervised Multitask Learning for Sequence Labeling. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2121–2130. https://doi.org/10.18653/v1/P17-1194

Ren, L., Liu, Y., Lu, Y., Shen, Y., Liang, C., & Chen, W. (2024). *Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling*. http://arxiv.org/abs/2406.07522

Sak, H., Senior, A., & Beaufays, F. (2014). *Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition*. http://arxiv.org/abs/1402.1128

Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., & Kuleshov, V. (2024). *Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling*. http://arxiv.org/abs/2403.03234

Schmidt, R. M. (2019). *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. http://arxiv.org/abs/1912.05911

Strubell, E., Verga, P., Belanger, D., & McCallum, A. (2017). Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2670–2680. https://doi.org/10.18653/v1/D17-1283

Sutton, C., Rohanimanesh, K., & McCallum, A. (2004). Dynamic conditional random

fields. *Twenty-First International Conference on Machine Learning - ICML '04*, 99. https://doi.org/10.1145/1015330.1015422

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 -*, *4*, 142–147. https://doi.org/10.3115/1119176.1119195

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/1706.03762

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. http://arxiv.org/abs/1609.08144

Yang, J., Zhang, T., Tsai, C.-Y., Lu, Y., & Yao, L. (2024). Evolution and emerging trends of named entity recognition: Bibliometric analysis from 2000 to 2023. *Heliyon*, *10*(9), e30053. https://doi.org/10.1016/j.heliyon.2024.e30053

Yang, X., Gao, Z., Li, Y., Pan, C., Yang, R., Gong, L., & Yang, G. (2018). Bidirectional LSTM-CRF for biomedical named entity recognition. *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 239–242. https://doi.org/10.1109/FSKD.2018.8687117

Yang, Z., Salakhutdinov, R., & Cohen, W. (2016). *Multi-Task Cross-Lingual Sequence Tagging from Scratch*.

Yang, Z., Salakhutdinov, R., & Cohen, W. W. (2017). *Transfer Learning for Sequence*

Tagging with Hierarchical Recurrent Networks. http://arxiv.org/abs/1703.06345

Yuan, H., Zheng, J., Ye, Q., Qian, Y., & Zhang, Y. (2021). Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, *151*. https://doi.org/10.1016/j.dss.2021.113633

Zhang, X., Zhao, J., & Lecun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, *2015-January*, 649–657.

Zhang, X., Zhao, J., & LeCun, Y. (2016). Character-level Convolutional Networks for Text Classification. *Advances in Neural Information Processing Systems*, *2015-January*, 649–657. http://arxiv.org/abs/1509.01626