結合MAMBA與CRFs在中文醫學命名實體識別

作者為 鈺傑 黃

提交日期: 2024年04月29日 01:16下午 (UTC+0800)
作業提交代碼: 2365185082
文檔名稱: 結合MAMBA與CRFs在中文醫學命名實體識別.pdf (671.27K)
文字總數: 4564
字符總數: 15187

書背

111 國立臺北商業大學資訊管理系人工智慧與商業應用碩士班 碩士論文（12pt） 論文中文題目（12pt） 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 CRFs 在中文醫學命名實體識別

Apply MAMBA model and CRFs in Chinese medical Named Entity Recognition

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

本研究聚焦在自然語言處理（NLP）領域，特別是在中文命名實體識別（Named Entity Recognition，NER）的研究和應用。NLP 是人工智慧中一個關鍵領域，它使計算機能夠理解、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。

特別強調 Mamba 架構來提升模型性能的可能性。Mamba 架構通過引入一種高效的序列建模方法，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。

中文 NER 比英文 NER 更具挑戰性，因中文以符號為基本單位，並且分詞和 NER 之間有緊密的關係。錯誤的分詞可能導致 NER 的錯誤。本研究將在醫療保健領域，NLP 和 NER 技術的應用，以更有效地分類和標記大量的醫療文本，有助於建立更完整和準確的醫療資料庫，提升醫療決策和研究的質量。

總的來說，本研究旨在解決醫療保健領域資訊的專業性和多樣性，並提供更有效的資訊搜索方法，同時強調 NLP 和 NER 技術在醫療保健領域的應用，為醫學專業人員和普通使用者提供更好的資訊檢索方式，並促進醫療資訊的可及性以及可理解性。

關鍵詞：自然語言處理、命名實體識別、循環神經網路、條件隨機場、Mamba

ABSTRACT

This paper focuses on the field of natural language processing (NLP), especially the research and application of Chinese Named Entity Recognition (NER). NLP is a key area of artificial intelligence that enables computers 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 and answer, but has also achieved important progress in the fields of sentiment analysis, speech recognition, and dialogue system construction.

Special emphasis is placed on the possibility of Mamba architecture to improve model performance. By introducing an efficient sequence modeling method, the Mamba architecture can handle long sequences more effectively and achieve linear expansion in calculations, breaking through the computational bottleneck of traditional Transformers on long sequences.

Chinese NER is more challenging than English NER because Chinese uses symbols as the basic unit and there is a close relationship between word segmentation and NER. Wrong word segmentation may lead to NER errors. This study will apply NLP and NER technology in the field of healthcare to more effectively classify and label large amounts of medical text, help build a more complete and accurate medical database, and improve the quality of medical decision-making and research.

In summary, this study aims to address the professionalism and diversity of information in the healthcare field and provide a more effective information search method, while emphasizing the application of NLP and NER technology in the healthcare field for both medical professionals and general use. Provide researchers with better ways to retrieve information and promote the accessibility and understandability of medical information.

Keyword：NLP、NER、RNN、CRFs、Mamba

致謝

第一章 研究背景

近年來，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)。

當談到 NLP 的發展時，Transformer 模型的崛起無疑是一大亮點。它利用了自注意機制（self-attention），成功地對文本實現更深層次的理解，取得了巨大的成就(Vaswanietal., 2017)。但是，隨著處理序列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體被提出來，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。Mamba 能夠在處理語言時隨著上下文長度的增加實現線性擴展，這使得它在處理長達百萬個 token 的序列時性能依然出色，同時提升了推理速度(Gu&Dao,2023)。

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，如人名、組織機構和地點等。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(Leeetal.,2022)。

在傳統上，NER 一直被視為序列標記問題的一種，其中我們需要同時預測實體的邊界和其對應的類別標籤。相較於英文 NER，中文 NER 更加具有挑戰

性。中文以符號為基本單位，不像英文那樣有明顯的大小寫區別等規則性特徵可供參考。由於中文字符之間沒有明確的分隔符號，因此中文 NER 中的詞與分詞密切相關，這也意味著命名實體的邊界通常也會是分詞的邊界。然而，錯誤的分詞決策可能會導致 NER 的錯誤傳播。例如，在特定情況下，身體類別的實體“上皮組織惡性腫瘤”可能會被錯誤地分割為三個詞：“上皮組織”、“惡性”和“腫瘤”(Lee&Chen,2022)。

在數位時代，人們通常會透過網路搜索和瀏覽各種網頁來獲取與健康相關的資訊，再預約醫生進行診斷和治療。網路上的文字內容是提供這些醫療保健資訊的主要來源，包括健康新聞、數位健康雜誌和醫學問答社群。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）(Tarcar et al., 2020)。

綜上所述，中文 NER 在 NLP 中具有重要且擁有關鍵性的任務，其核心目標是自動識別醫學領域中的各種實體，包括症狀(Symptom)、醫療設備（Instruments）、化學物質（Chemicals）、營養品(Supplement)。進而有助於機器閱讀與理解醫學文本。

本研究將致力於解決醫療保健領域信息處理中的專業性和多樣性挑戰。醫療領域的文本通常涵蓋各種專業術語、縮寫以及不同語言風格的描述，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。我們運用通用的 BIO（即開始、內部和外部）格式來執行命名實體識別（NER）任務。在標記中，以"B"開頭的表示命名實體的開始，而以"I"開頭的表示命名實體的內部。而 "O" 標記則表示該令牌不屬於任何命名實

體(Lee et al., 2023)。並且通過比對機器預測的標籤和人工標記之間的差異來評估性能。標準的精確度、召回率和 F1 分數是評估 NER 系統在字元級別上的常見指標。

第二章文獻探討

深度學習被證明是直接從文本數據中提取特徵表示的有效策略，這在命名實體識別（NER）領域取得了突破性進展(Yang et al., 2024)。NER 是文本處理的一項任務，用於在文本中發現不同類型的命名實體，例如人名、地名、日期，甚至是網路連結或電話號碼等。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)。

最早期的 NER 研究採用手工設計的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庫，如軍事情報集、海軍作戰報告等。隨著在大規模標記資料庫上進行監督學習技術的發展，NER 取得了最先進的結果。尤其是條件隨機場（CRFs）是最有效的 NER 演算法。在 NER 中，需要利用許多前後相連的非局部序列來訓練輸出標籤的機率，這使得 CRFs 模型比其他生成模型更適用於 NER (Roy, 2021)。

CRFs 是一種被廣泛用於序列標記問題的統計模型，它能夠捕捉序列中的依賴關係，這使得它在多個領域中都獲得了重要的成功。然而，儘管其應用廣泛，CRFs 模型也存在一個不可忽視的缺點，即偶爾會生成非法的標記序列。這個問題通常體現在遵守標記約束方面，特別是在使用基本的 BIO 標記方案時。在這種方案中，"B-"標記表示一個實體的開始，"I-"標記表示實體的中間部分，而"O"則表示非實體。根據這些約束，不應該出現"I-"標記後面立刻是"O"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(Weietal.,2021)。

假定ℓ為訓練集序列長度，X_i為長度為ℓ的輸入向量序列，y_i為長度為ℓ的標籤向量序列，在常規分類問題透過乘以第 k 位置的每一個機率來計算$P(y|X)$，方程式如下(Lafferty et al., 2001)：

$$P(y|X) = \frac{\exp\left(\sum_{k=1}^{\ell} U(x_k, y_k) + \sum_{k=1}^{\ell-1} T(x_k, y_{k+1})\right)}{\mathrm{Z}(X)}$$

其中$U(x_k, y_k)$為發射分數(emissions)，表示在給定輸入x_k的情況下y_k的可能性的分數。$T(x_k, y_k)$為轉移分數(transition score)，示y_k後面跟著y_{k+1}的可能性的分數。$Z(x)$為標準化函式，為了得到序列上的機率分佈，但須要正確定義標準化函式$Z(x)$(Lafferty et al., 2001)：

$$Z(x) = \sum_{y'_1} \sum_{y'_2} \cdots \sum_{y'_k} \cdots \sum_{y'_\ell} \left(\sum_{k=1}^{\ell} U(x_k, y_k) + \sum_{k=1}^{\ell-1} T(x_k, y_{k+1}) \right)$$

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。我們正在計算$P(y|X)$，想要最大化的值。為了將其視為最小化問題，我們取該機率的負對數，得(Lafferty et al., 2001)：

$$L = -\log\big(P(y|x)\big) = \mathrm{Z}_{log}(X) - \left(\sum_{k=1}^{\ell} U(x_k, y_k) + \sum_{k=1}^{\ell} T(x_k, y_{k+1}) \right)$$

舉例來說，已知中文的文本「中樞神經系統」(圖 3-1)，採用 BIO 標記方式，輸入標記序列$\mathrm{X} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$，即$x_1$=中、$x_2$=樞、$x_3$=神、$x_4$=經、$x_5$=系、$x_6$=統。輸出為$\mathrm{Y} = (y_1, y_2, y_3, y_4, y_5, y_6)$，則$y_1$、$y_2$、$y_3$、$y_4$、$y_5$、$y_6$取得{B-BODY, I-BODY, I-BODY , I-BODY , I-BODY , I-BODY }。

在醫學文本中也有廣泛的應用，提取重要的醫學資訊是其中一種，例如疾病、癥狀和藥物。首先，我們在模型的初始階段採用詞嵌入技術。這些詞嵌入是通過 Skip-gram 方法進行訓練得到的向量空間模型(Melamud et al., 2016)。這種方法有助於將單詞的分散式表示嵌入到向量空間中，從而可以將語義相似的詞彙進行分類，提高了自然語言處理任務的性能表現。

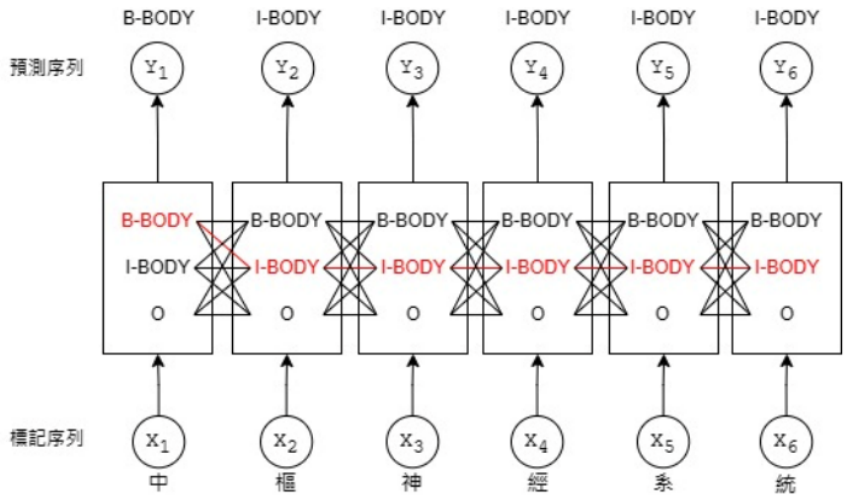

圖 1-1　CRFs 神經網路

這個方法的主要優勢在於它不僅可以識別和分類醫學文本中的命名實體，還可以理解它們之間的關係和語義資訊。這對於從大量的醫學文本中提取重要資訊非常有幫助，例如患者的病史、疾病的傳播趨勢或藥物的副作用等。

卷積神經網路(CNN, Convolutional Neural Network)是一種高效的方法，用於從句子或單詞中提取形態資訊。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。(Zhang et al., 2015)表明可以在不瞭解單字、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。

循環神經網路(RNN, Recurrent Neural Network)是基於順序資訊的有前途的深度學習演算法。與前饋神經網路(feedforward 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)認為成人疾病患者的心臟病診斷是一個關鍵問題，所以使用多個 RNN 從患者的診斷資料序列中學習，以預測高危疾病的發生。

Mamba 架構主要依賴結構化狀態空間對長序列進行高效建模（即 S4, Structured State Spaces for Sequence Modeling）架構。在 S4 中，透過三個連續參數矩陣 A、B 和 C 將這些狀態相互關聯。(Gu & Dao, 2023)方程式如下：

$$h'(t) = Ah(t) + Bx(t)$$
$$y(t) = Ch(t)$$

由於現實中，一般不會利用連續型數據進行處理，都是離散數據比如文本，所以需要對 SSM 進行離散化。方程式如下：

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t$$
$$y_t = Ch_t$$

方程會形成遞迴，情況類似於在 RNN 中一樣。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。S4 也可以將它用作 CNN，之前的離散方程來嘗試計算h_2時：

$$h_0 = \bar{B}x_0$$
$$h_1 = \bar{A}h_0 + \bar{B}x_1$$
$$h_2 = \bar{A}(\bar{A}h_0 + \bar{B}x_1) + \bar{B}x_2$$
$$y_2 = C(\bar{A}(\bar{A}h_0 + \bar{B}x_1) + \bar{B}x_2)$$

長短期記憶網絡（LSTM）是一種 RNN 的延伸變化，特別適用於處理序列資料，如自然語言處理和時間序列預測(Huang et al., 2015a)。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。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 1-1 顯示了採用上述 LSTM 記憶單元的 LSTM 序列標記模型。

LSTM 儲存單元實現如下：

$$I_t = \sigma\left(W_{h_i} h_{t-1} + W_{x_i} \chi_t + b_i\right)$$
$$F_t = \sigma\left(W_{h_f} h_{t-1} + W_{x_f} \chi_t + b_f\right)$$
$$O_t = \sigma\left(W_{h_o} h_{t-1} + W_{x_o} \chi_t + b_o\right)$$
$$C_t = F_t C_{t-1} + I_t \, tanh\left(W_{h_c} h_{t-1} + W_{x_c} \chi_t + b_c\right)$$
$$H_t = O_t \, tanh(C_t)$$

當σ為邏輯 sigmoid 函數，而且 I、F、O 與 C 是輸入門(input gate)、忘記門(forge gate)、輸出門(output gate)和單元向量(cell vectors)，它們都與隱藏向量 H 的大小相同。權重矩陣下標具有顧名思義的含義。(Huangetal.,2015)

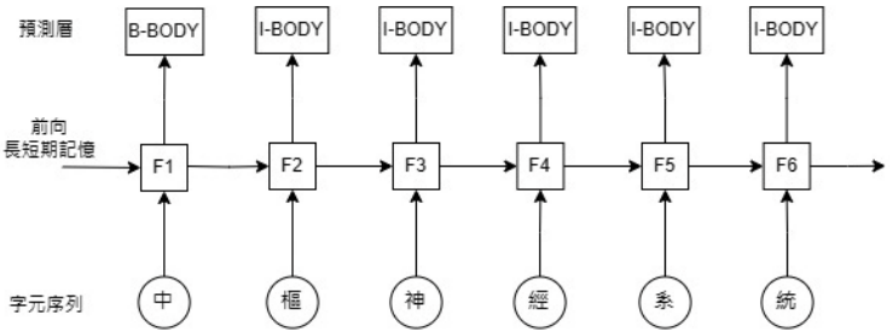

圖 2-1　LSTM 神經網路

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。該方法利用 LSTM 來處理文本中的單詞特徵，將它們轉化為對應的命名實體標記分數。為了實現這個目標，每一個單詞的特徵首先被送入一個前向 LSTM 和一個後向 LSTM。這樣的設計允許同時考慮到單詞的上下文關係，這對於準確地識別命名實體非常重要。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 2-1)。

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。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標記類別。(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相加。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性能。

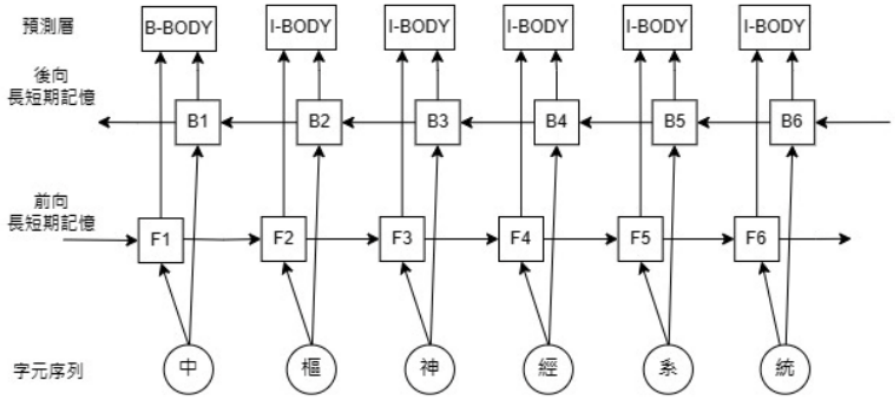

圖 3-1　Bi-LSTM 神經網路

一個注意力函數，通常在深度學習模型中廣泛應用，可以被視為一種將資訊從輸入到輸出的機制。它在處理序列資料，自然語言處理，計算機視覺等領

域發揮著關鍵作用。這種函數的核心思想是將查詢向量與一組鍵值進行比對，以計算輸出向量。該函數可以量化查詢和唯一值之間的相似度或關聯性。

舉例來說，當我們使用注意機制來翻譯一個句子，查詢可以是正在翻譯的目標單詞，而鍵值則可能是源語言句子中的單詞和它們的嵌入表示。通過計算查詢與每一個鍵值之間的相似性，我們可以為每一個鍵值分配一個權重，而權重是用於加權總和計算，以生成最終的輸出單詞的表示。

這種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，並在不同上下文中達到更好的性能。這在處理長序列、擁有多重語義的詞語或解決聚焦問題時非常有幫助。因此，注意力函數已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數據。(Vaswanietal.,2017)

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：

$$P_i = \frac{TP_i}{TP_i+FP_i}，R_i = \frac{TP_i}{TP_i+FN_i}，F1_i = \frac{2P_i*R_i}{P_i+R_i}$$

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

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

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

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

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

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

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

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

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。(Yuanetal.,2021)

第三章 研究方法

本研究資料集使用中文醫療命名實體辨識資料庫(Lee & Chen, 2022)以及 ROCLING 2022 中文醫療保健命名實體識別資料集(Lee & Chen, 2022)。針對醫療領域多類型 NER 的任務。有三種類型：表格 1-1 為命名實體類型的描述：

- 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：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」。

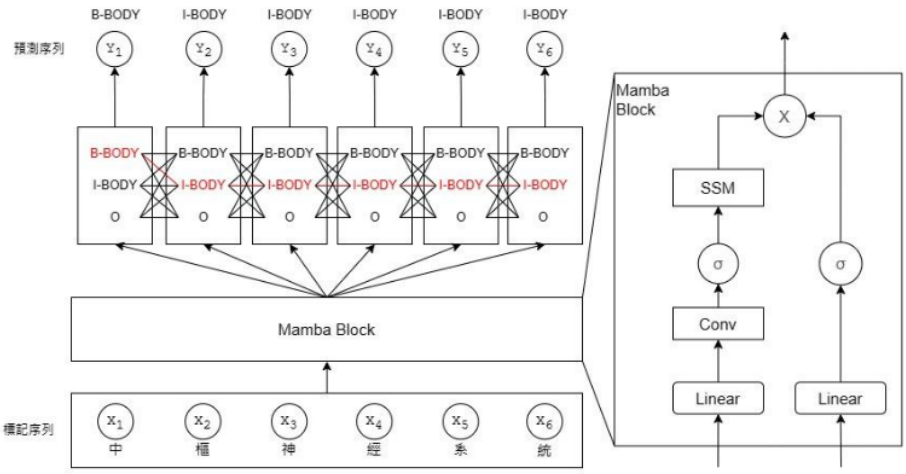

圖 4-1 模型架構圖

本研究旨在探討如何利用 Mamba 與 CRFs 結合來提高模型的性能。之所以利用 Mamba 架構是因為它在處理序列數據時展現出了卓越的性能。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 架構能夠有效地處理大型數據集和長序列，而無需過多擔心性能下降。

另外，在推理過程中，Mamba 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數據，Mamba 架構也能夠提供快速而穩定的推理性能。

CRFs 一種被廣泛用於序列標記問題的統計模型，它能夠捕捉序列中的依賴關係(Wei et al., 2021)，這使得它在多個領域中都獲得了重要的成功。或是增加訓練數據量以提升模型的泛化能力，並改進訓練策略以更好地優化模型參數。 綜合採取這些措施有望提高模型在實體標記任務上的性能表現。圖 4-1 為模型架構圖。

表 1-1　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)

參考文獻

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

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

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

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

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

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

Lafferty, J., Mccallum, A., & Pereira, F. (2001). *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. https://api.semanticscholar.org/CorpusID:219683473

Lee, L.-H., & Chen, C.-Y. (2022). *Overview of the ROCLING 2022 Shared Task for Chinese Healthcare Named Entity Recognition*. 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*. https://aclanthology.org/2023.rocling-1.42

Lee, L.-H., Lu, C.-H., & Lin, T.-M. (2022). NCUEE-NLP at SemEval-2022 Task 11: Chinese Named Entity Recognition Using the BERT-BiLSTM-CRF Model. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 1597–1602. https://doi.org/10.18653/v1/2022.semeval-1.220

Melamud, O., McClosky, D., Patwardhan, S., & Bansal, M. (2016). The Role of Context Types and Dimensionality in Learning Word Embeddings. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1030–1040. https://doi.org/10.18653/v1/N16-1118

Pakhale, K. (2023). *Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges*. http://arxiv.org/abs/2309.14084

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

Roy, A. (2021). *Recent Trends in Named Entity Recognition (NER)*. http://arxiv.org/abs/2101.11420

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

Tarcar, A. K., Tiwari, A., Rao, D., Dhaimodker, V. N., Rebelo, P., & Desai, R. (2020). Healthcare NER models using language model pretraining. *CEUR Workshop Proceedings*, *2551*, 12–18. https://doi.org/10.1145/3336191.3371879

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

Wei, T., Qi, J., He, S., & Sun, S. (2021). *Masked Conditional Random Fields for Sequence Labeling*. http://arxiv.org/abs/2103.10682

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*, e30053. https://doi.org/10.1016/j.heliyon.2024.e30053

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.

結合MAMBA與CRFs在中文醫學命名實體識別

原創性報告

7%	7%	2%	0%
相似度指數	網際網絡來源	出版物	學生文稿

主要來源

1	codeqingyun.com 網際網絡來源	1%
2	nccur.lib.nccu.edu.tw 網際網絡來源	1%
3	www.cnblogs.com 網際網絡來源	<1%
4	www.ygxb.ac.cn 網際網絡來源	<1%
5	patents.google.com 網際網絡來源	<1%
6	aclanthology.org 網際網絡來源	<1%
7	discourse.matplotlib.org 網際網絡來源	<1%
8	www.coursehero.com 網際網絡來源	<1%
9	"Natural Language Understanding and Intelligent Applications", Springer Nature, 2016	<1%

出版物

10	apps.who.int 網際網絡來源	<1%
11	huggingface.co 網際網絡來源	<1%
12	www.woshipm.com 網際網絡來源	<1%
13	Kumar, M.. "A second order finite difference method and its convergence for a class of singular two-point boundary value problems", Applied Mathematics and Computation, 20031231 出版物	<1%
14	arxiv.org 網際網絡來源	<1%
15	stats.moe.gov.tw 網際網絡來源	<1%
16	Di Lena, G.. "An extension of Chu-Moyer's theorem to two variable functions", Nonlinear Analysis, 20081115 出版物	<1%
17	mailman.uib.no 網際網絡來源	<1%
18	pkucaters.github.io 網際網絡來源	<1%

19	static.dergipark.org.tr 網際網絡來源	<1 %
20	www.cjig.cn 網際網絡來源	<1 %
21	www.xjishu.com 網際網絡來源	<1 %
22	xa8.net 網際網絡來源	<1 %
23	zh-v2.d2l.ai 網際網絡來源	<1 %
24	"Text, Speech, and Dialogue", Springer Science and Business Media LLC, 2021 出版物	<1 %
25	Asep Ridwan Lubis. "Balancing the Equation: Investigating AI Advantages, Challenges, and Ethical Considerations in the Context of GPT-3, Natural Language Processing, and Researcher Roles", SAR Journal - Science and Research, 2023 出版物	<1 %

排除引述	開	排除相符處	關閉
排除參考書目	開		