**國立臺北商業大學管理學院**

書背

**資訊管理系人工智慧與商業應用碩士班**

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

**碩士學位論文**

**結合MAMBA與CRFs在中文醫學命名實體識別**

碩士論文 （12 pt） 論文中文題目（12 pt） 黃鈺傑 撰

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**

[目錄 II](#_Toc169299602)

[圖目錄 III](#_Toc169299603)

[表目錄 IV](#_Toc169299604)

[摘要 5](#_Toc169299605)

[ABSTRACT 7](#_Toc169299606)

[致謝 9](#_Toc169299607)

[第一章 研究背景 10](#_Toc169299608)

[第二章 文獻探討 13](#_Toc169299609)

[第一節 條件隨機場(CRF) 13](#_Toc169299610)

[第二節 卷積神經網路(CNN) 16](#_Toc169299611)

[第三節 循環神經網路(RNN) 17](#_Toc169299612)

[第四節 Mamba 17](#_Toc169299613)

[第四節 長短期記憶網絡（LSTM） 18](#_Toc169299614)

[第四節 雙向長短期記憶網絡（Bi-LSTM） 19](#_Toc169299615)

[第五節 Transformer 20](#_Toc169299616)

[第六節 BERT & BERT-WWM 21](#_Toc169299617)

[第七節 RoBERTa & RoBERTa-WWM 22](#_Toc169299618)

[第八節 MACBERT & MLM as correction 22](#_Toc169299619)

[第九節 評估指標 23](#_Toc169299620)

[第三章 研究方法 25](#_Toc169299621)

[第四章 實驗結果 29](#_Toc169299622)

[第五章 結論與未來展望 37](#_Toc169299623)

[參考文獻 38](#_Toc169299624)

**圖目錄**

[圖1-1　CRFs神經網路 15](#_Toc169299679)

[圖2-1　用於提取單字的字元級表示的CNN神經網路 16](#_Toc169299680)

[圖3-1　LSTM神經網路 19](#_Toc169299681)

[圖3-1　Bi-LSTM神經網路 20](#_Toc169299682)

[圖4-1　模型架構圖 26](#_Toc169299683)

**表目錄**

[表1-1　命名實體類型 24](#_Toc169099237)

[表2-1　Mamba加入CRF在中文醫療命名實體辨識資料庫中混淆矩陣評價分類指標結果 27](#_Toc169099238)

[表3-1　Mamba + CRF與Mamba在中文醫療命名實體辨識資料庫中混淆矩陣評價分類指標結果比較 28](#_Toc169099239)

[表4-1　中文醫療命名實體辨識資料庫與各模型使用微觀平均精確度(Micro F1 Score)比較 31](#_Toc169099240)

[表4-2　中文醫療命名實體辨識資料庫與各模型使用宏觀平均精確度(Macro F1 Score)比較 31](#_Toc169099241)

[表5-1　CoNLL-2003命名實體資料集與各模型比較 33](#_Toc169099242)

**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。

總結來說，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(J. 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)。

第一節 條件隨機場(CRF)

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)。舉例來說，「B-DRUG」表示藥品標記資料的開始，「I-DRUG」表示藥品標記資料的中間或結尾。

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

其中為發射分數(emissions)，表示在給定輸入的情況下的可能性的分數。為轉移分數(transition score)，示後面跟著的可能性的分數。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(Lafferty et al., 2001)：

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，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)：

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，即=中、=樞、=神、=經、=系、=統。輸出為，則、、、、、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。

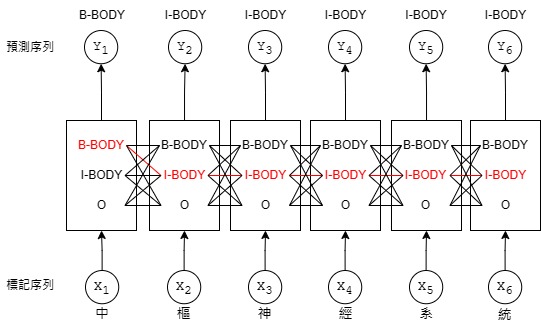


圖1-1　CRFs神經網路

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。

第二節 卷積神經網路(CNN)

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。

圖2-1展示了的CNN神經網路，該CNN與Chiu和Nichols（2015）中的CNN類似，唯一的不同之處在於我們僅使用字符嵌入作為CNN的輸入，而不使用字符類型特徵。虛線箭頭表示在字元嵌入輸入到CNN之前套用的dropout層(Ma & Hovy, 2016)。

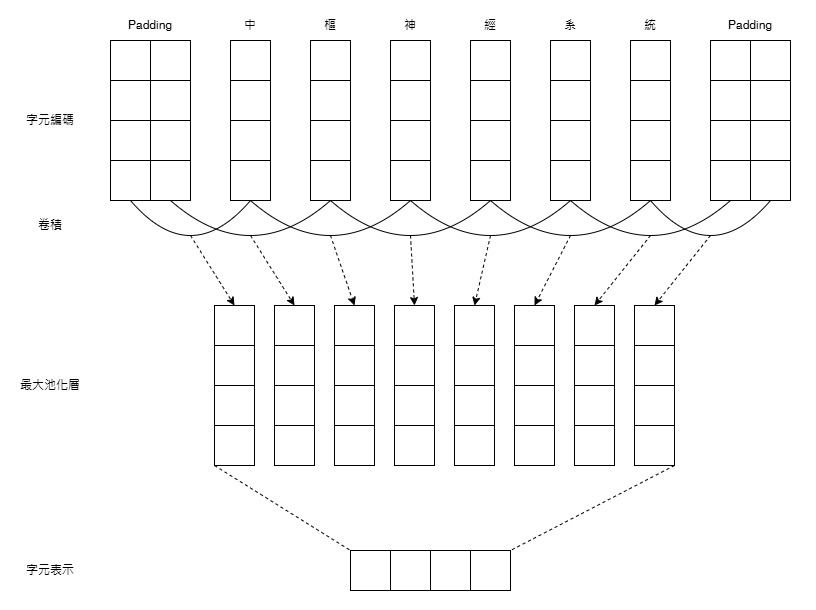


圖2-1　用於提取單字的字元級表示的CNN神經網路

第三節 循環神經網路(RNN)

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。

第四節 Mamba

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：

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：

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，將前一個時間的隱藏狀態與當前輸入相結合，以創建新的隱藏狀態。S4也可以將它用作CNN，之前的離散方程來嘗試計算時：

第四節 長短期記憶網絡（LSTM）

長短期記憶網絡（LSTM, Long Short Term Memory）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：

當為邏輯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)

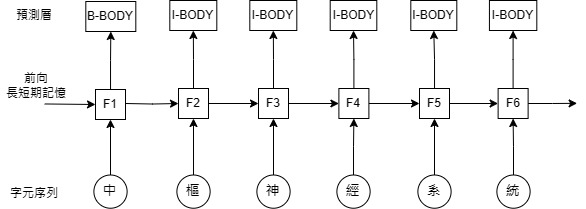


圖3-1　LSTM神經網路

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。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。

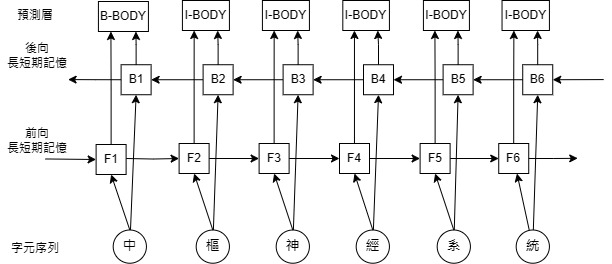


圖3-1　Bi-LSTM神經網路

第五節 Transformer

注意力機制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查詢向量（Query）與一組鍵向量（Key）進行比對，以計算一組相應的值向量（Value）。這種函數的目的是量化查詢與鍵之間的相似度或關聯性。

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)

第六節 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的設計是通過在所有Transformer層中同時條件化左右上下文，來預訓練深度雙向表示(Cui et al., 2019)

。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）。

* 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。
* 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., 2019)。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)。這明確要求模型在MLM預訓練任務中恢復整個單詞，而不僅僅是恢復Word Piece標記，這更具挑戰性。

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., 2019)

第七節 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、批次大小較大、序列長度較長且使用更多數據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NSP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可能是有益的。(Cui et al., 2019)

第八節 MACBERT & MLM as correction

MLM是BERT及其變體中最重要的預訓練任務，它模擬了雙向上下文推理能力。然而，MLM存在預訓練階段中的人工標記（如[MASK]）在真實的下游微調任務中從未出現(Devlin et al., 2018)。

MACBERT（MLM as correction BERT）與BERT具有類似的預訓練任務，但進行了一些修改。MACBERT包含兩個預訓練任務：MLM as correction 和句子順序預測。在MLM as correction（MAC）。這個預訓練任務中，不採用任何預先定義的標記進行遮罩。相反，將原始的MLM轉化為一個文本校正任務，其中模型應將錯誤單詞更正為正確單詞，這會比MLM更自然(Cui et al., 2019)。

第九節 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、、和分別為類別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()、召回率()和F1分數()定義如下：

，，

此外，微觀平均精確度（）、微平均召回率（）、宏觀平均精確度（）和宏平均召回率（）計算如下：

，

，

=，

=

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：

，

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。表格1-1為ROCLING 2023命名實體類型的描述。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：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」。

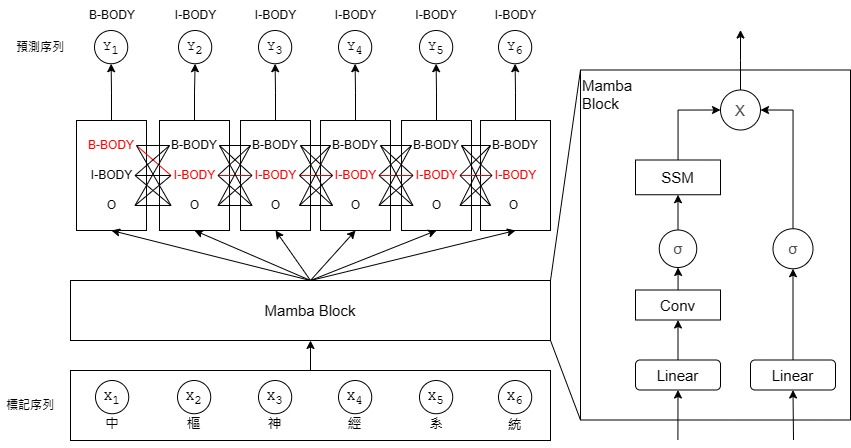


圖4-1　模型架構圖

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。圖4-1為模型架構圖。

表1-1　 ROCLING 2023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)

也有使用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 et al., 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2-1為英文每一個資料集中命名實體數的概述，表2-2為德文每一個資料集中命名實體數的概述。

表2-1　 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 |

表2-2　 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 |

**第四章 實驗結果**

Mamba加入CRF的模型在中文醫療命名實體辨識資料庫的混淆矩陣評價分類指標結果中（表3-1），顯示了在許多標籤下的性能提升。這項改進在身體部位（BODY）、症狀（SYMP）、檢查（EXAM）、化學物質（CHEM）、疾病（DISE）、藥品（DRUG）以及支持（SUPP）等實體識別方面特別明顯，表現出CRF的加入對這些特定類型的實體識別有積極影響。

需要特別注意的部分是，表2-1中的大多數F1-Score主要源於“O”類標籤，即非實體標記資料，這也是該資料庫中最常見的類型。即使如此，Mamba加入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。

然而，值得注意的是，模型在某些特定類型，如治療（TREAT）和時間（TIME），的性能改進有限，甚至可能沒有明顯提升。這可能是由於這些類型的語義模糊性和樣本稀少性對模型性能的影響。因此，未來的研究可以著重於增加這些類型的樣本數量，或者探索更有效的特徵表示方法，以進一步提高模型在這些領域的性能。

除了整體F1-Score，比較準確率和召回率也是評估模型性能的重要指標。準確率反映了模型標記的正確性，而召回率則反映了模型對實體的識別能力。對於召回率較低的情況，特別是對於某些類型的實體，可能需要更多的注意和改進，以確保模型能夠準確識別各種類型的實體。

表2-1　Mamba加入CRF在中文醫療命名實體辨識資料庫(Lee & Chen, 2022)中混淆矩陣評價分類指標結果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 標籤 | 準確率 | 召回率 | 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3-1　Mamba + CRF與Mamba在中文醫療命名實體辨識資料庫(Lee & Chen, 2022)中混淆矩陣評價分類指標結果比較。(ΔF1-score = Mamba with CRF F1-score - Mamba F1-score)

|  |  |  |  |
| --- | --- | --- | --- |
| 標籤 | Mamba + CRF | Mamba | ΔF1-score |
| O | 98.4% | 93.8% | 4.6% |
| B-BODY | 64.0% | 36.1% | 27.9% |
| I-BODY | 71.7% | 32.0% | 39.7% |
| B-SYMP | 57.4% | 20.7% | 36.7% |
| I-SYMP | 61.8% | 29.2% | 32.6% |
| B-INST | 23.3% | 0.0% | 23.3% |
| I-INST | 21.7% | 2.0% | 19.7% |
| B-EXAM | 53.9% | 6.2% | 47.7% |
| I-EXAM | 73.4% | 44.7% | 28.7% |
| B-CHEM | 56.2% | 39.0% | 17.2% |
| I-CHEM | 75.3% | 47.4% | 27.9% |
| B-DISE | 52.4% | 8.5% | 43.9% |
| I-DISE | 71.4% | 31.0% | 40.4% |
| B-DRUG | 37.7% | 0.0% | 37.7% |
| I-DRUG | 62.2% | 3.0% | 59.2% |
| B-SUPP | 53.8% | 37.2% | 16.6% |
| I-SUPP | 76.2% | 27.6% | 48.6% |
| B-TREAT | 23.1% | 0.0% | 23.1% |
| I-TREAT | 34.7% | 29.7% | 5.0% |
| B-TIME | 40.5% | 12.5% | 28.0% |
| I-TIME | 50.0% | 0.0% | 50.0% |
| micro avg | 91.9% | 86.3% | 5.6% |
| macro avg | 55.1% | 23.8% | 31.3% |
| weighted avg | 91.3% | 83.6% | 7.7% |

在中文醫療命名實體辨識資料庫與各模型使用微觀平均精確度比較(表4-1)以及中文醫療命名實體辨識資料庫與各模型使用宏觀平均精確度比較(表4-2)中，呈現了不同模型在兩種平均精確度下的性能指標。精確率(Precision)衡量的是模型預測為正例的樣本中實際正例的比例，召回率(Recall)則衡量的是實際正例中被模型預測為正例的比例。F1-score則是精確率和召回率的調和平均，它給出了Precision和Recall的綜合評價，特別適用於不平衡類別的情況。

在微觀（Micro）水準下，這些指標將所有類別的預測和真實值組合計算，然後計算Precision、Recall和F1-score。在宏觀（Macro）水準下，則是分別計算每一個類別的Precision、Recall和F1-score，然後取平均值。這兩種水準提供了對模型性能的不同視角。

從表格4-1中可以看出，Mamba + CRF 在微觀水準下表現最好，其Precision、Recall和F1-score均達到91.9%，這意味著模型對於所有類別的預測和真實值的匹配都非常準確。然而，在宏觀水準下，這個模型的表現則沒有那麼突出，Precision、Recall和F1-score都較低，分別為66.3%、48.2%和55.1%。相比之下，Bert + Bi-LSTM + CRF在宏觀水準下表現較好，其Precision、Recall和F1-score分別為77.3%、67.3%和68.1%，但在微觀水準下稍微遜色。

這些結果表現出不同模型在預測目標方面的不同優勢和局限性。Mamba + CRF在微觀水準下表現較好，而其他模型在宏觀水準下表現較好。這表現了在選擇模型時，需要根據具體任務的需求和優先考慮的指標來進行選擇。例如，如果微觀水準下的性能對任務非常重要，則可以考慮使用Mamba + CRF；相對的，如果更在意宏觀水準下的性能，則可以考慮其他模型。

表4-1　中文醫療命名實體辨識資料庫(Lee & Chen, 2022)與各模型使用微觀平均精確度(Micro F1 Score)比較

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 準確率 | 召回率 | F1-score |
| Mamba + CRF | 91.9% | 91.9% | 91.9% |
| Mamba | 47.9% | 23.9% | 31.9% |
| Bert + Bi-LSTM + CRF (ROCLING 2023) | 81.1% | 78.2% | 79.6% |
| Bert + CRF | 77.3% | 66.0% | 71.2% |
| ClinicalDistilBERT + Bi-LSTM + CRF | 77.8% | 66.2% | 71.6% |
| MacBERT + Bi-LSTM + CRF | 80.4% | 73.8% | 76.9% |
| BERTWWM + Bi-LSTM + CRF | 78.8% | 75.3% | 77.0% |
| Roberta + WWM+ Bi-LSTM + CRF | 81.3% | 78.5% | 79.9% |

表4-2　中文醫療命名實體辨識資料庫(Lee & Chen, 2022)與各模型使用宏觀平均精確度(Macro F1 Score)比較

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 準確率 | 召回率 | F1-score |
| Mamba + CRF | 66.3% | 48.2% | 55.1% |
| Mamba | 32.0% | 17.0% | 20.3% |
| CRF | 62.2% | 40.3% | 47.9% |
| Bert + Bi-LSTM + CRF (ROCLING 2023) | 77.3% | 67.3% | 68.1% |
| Bert + CRF | 68.2% | 50.9% | 57.2% |
| ClinicalDistilBERT + Bi-LSTM + CRF | 72.7% | 53.5% | 60.1% |
| MacBERT + Bi-LSTM + CRF | 74.5% | 60.1% | 65.6% |
| BERTWWM + Bi-LSTM + CRF | 73.8% | 62.4% | 66.8% |
| Roberta + WWM+ Bi-LSTM + CRF | 76.5% | 67.9% | 71.5% |

在CoNLL-2003命名實體資料集與各模型比較(表5-1)中，將對Mamba + CRF模型與各種先進的命名實體識別模型進行全面的比較和分析，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。

Collobert et al., 2011提出了一種統一的神經網路架構和學習算法，適用於多種NLP任務，包括詞性標記、分塊、命名實體識別和語義角色標記。

Huang et al., 2015引入了一個基於Bi-LSTM和CRF的序列標記模型。

Chiu & Nichols, 2016提出了一種結合了Bi-LSTM和CNN架構，用於自動檢測單詞和字元特徵，減少了特徵工程的需求。

Ma & Hovy, 2016結合Bi-LSTM、CNN和CRF，實現了端到端的序列標記。

Lample et al., 2016開發了一個基於原始LSTM的堆疊式模型，增加模型的深度，有助於更好地捕捉長期依賴性。當加入CRF層時，F1-score有更一步提升。

Hu et al., 2016將邏輯規則轉化為神經網絡可以理解的形式，並通過反覆運算過程融入神經網絡參數中。

Z. Yang et al., 2016提出了深層次RNN用於序列標記，並擴展到多任務和跨語言聯合訓練，並提出了四種不同情況下，模型的結果又會不同，分別是在不使用詞嵌入的情況下、在不使用字符GRU的情況下、在不使用詞典的情況下以及完整模型。

Rei, 2017提出了具有次要訓練目標的序列標記框架。

Strubell et al., 2017提出了反覆運算擴張卷積神經網絡（ID-CNN）作為Bi-LSTM與CRF的快速替代方案。

Peters et al., 2017也提出了三種不同情況下，分別為使用語言模型的上下文嵌入作為輸入、在使用1B詞語數據集時以及使用更大模型時。

Mamba模型專注於通過S4對長序列進行高效建模，尤其是通過三連參數矩陣來關聯狀態，進一步提升對長序列的建模能力。並且結合CRF層，進一步提高F1-score。

雖然Mamba + CRF模型在處理長序列方面具有獨特的優勢，但其在CoNLL-2003資料集上的整體表現仍略低於一些基於深度學習的模型，如Ma & Hovy, 2016及Peters et al., 2017。然而，Mamba模型的創新之處在於其結構化狀態空間模型，這可能在其他應用場景中展現出更強的能力。

表5-1　CoNLL-2003命名實體資料集與各模型比較

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 準確率 | 召回率 | F1-score |
| Mamba + CRF | 85.2% | 85.2% | 85.2% |
| Collobert et al., 2011 | - | - | 88.6% |
| Huang et al., 2015 | - | - | 84.2% |
| Chiu & Nichols, 2016 | 83.4% | 83.2% | 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 (no word embedding) | - | - | 77.2% |
| Z. Yang et al., 2016 (no char GRU) | - | - | 88.0% |
| Z. Yang et al., 2016 (no gazetter) | - | - | 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 結論與未來展望**

本研究不僅在理論上探討了NER技術在中文醫療領域的應用，而且通過實際模型的構建和評估，展示了Mamba模型與CRF結合的優勢，並強調Mamba模型與CRF結合的應用，會進一步提升了在特定類型（如身體部位、症狀、化學物質等）的NER性能。

在未來的研究可以在以下方面進一步擴展和改進：

1. **增強特定類型的識別能力**：對於某些特定類型（如治療和時間）的識別性能仍有限，未來的工作可以集中於增加這些類型的樣本數量，或者探索更有效的特徵表示方法。
2. **跨語言和跨領域的應用**：將Mamba模型和CRF結合，應用於其他語言和不同領域的NER任務。這可以擴展模型的通用性和實用性，使其在全球範圍內更廣泛地應用於醫療信息管理和其他領域。
3. **模型效能的進一步提升**：儘管Mamba模型在處理長序列方面表現出色，但在一些公開的標準數據集（如CoNLL-2003）上的整體性能仍有提升空間。未來可以通過調整模型架構、優化參數和集成更多領域知識來改進模型的性能。
4. **多模態數據的處理**：醫療資訊不僅包括文本，還可能涉及圖像、聲音等多模態數據。未來可以探索如何將文本信息與其他類型的數據進行有效整合和處理，以提升信息提取和理解的全面性和準確性。

未來的工作將繼續致力於解決現有技術的限制，並尋找更好的解決方案來應對不斷變化的醫療保健資訊處理挑戰。

**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

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*. http://arxiv.org/abs/1103.0398

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2019). *Pre-Training with Whole Word Masking for Chinese BERT*. 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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

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

Hu, Z., Ma, X., Liu, Z., Hovy, E., & Xing, E. (2016). *Harnessing Deep Neural Networks with Logic Rules*. https://doi.org/10.18653/V1/P16-1228

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

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). *Neural Architectures for Named Entity Recognition*. 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*. 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

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*. http://arxiv.org/abs/1907.11692

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

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

Peters, M. E., Ammar, W., Bhagavatula, C., & Power, R. (2017). *Semi-supervised sequence tagging with bidirectional language models*. 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*. https://doi.org/10.18653/v1/P17-1194

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

Strubell, E., Verga, P., Belanger, D., & Mccallum, A. (2017). *Fast and Accurate Entity Recognition with Iterated Dilated Convolutions*. https://doi.org/10.18653/v1/d17-1283

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

Tjong, E. F., Sang, K., & De Meulder, F. (2003). *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. http://lcg-www.uia.ac.be/conll2003/ner/

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

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

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.