**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**

**使用BERT在中文醫學命名實體識別分析之研究**

碩士論文 （12 pt） 論文中文題目（12 pt） 黃鈺傑 撰

Research of using and BERT in Chinese medical NER analysis

**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](#_Toc147483598)

[圖目錄 III](#_Toc147483599)

[摘要 4](#_Toc147483600)

[ABSTRACT 5](#_Toc147483601)

[第一章 緒論 6](#_Toc147483602)

[第二章 文獻探討 8](#_Toc147483603)

[第一節 詞向量表示 8](#_Toc147483604)

[第二節 卷積神經網路（CNN） 8](#_Toc147483605)

[第三節 長短時記憶網路(LSTM) 8](#_Toc147483606)

[第四節 雙向循環神經網絡(Bi-LSTM) 10](#_Toc147483607)

[第五節 條件隨機場(CRF) 11](#_Toc147483608)

[第六節 注意力機制(Attention) 12](#_Toc147483609)

[第七節 BERT 12](#_Toc147483610)

[第八節 評估矩陣(Evaluation matrix) 13](#_Toc147483611)

[參考文獻 14](#_Toc147483612)

# **圖目錄**

[圖 1 從N元語法生成多序列 9](#_Toc147488778)

[圖 2 LSTM 神經網路 10](#_Toc147488779)

[圖 3 Bi-LSTM神經網路 11](#_Toc147488780)

[圖 4 CRF 神經網路 11](#_Toc147488781)

**摘要**

本研究聚焦在自然語言處理（NLP）領域，特別是在中文NER（命名實體識別）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。

特別強調了Transformer模型，這種模型通過自注意機制實現對文本的深度理解，改進了文本處理效率，並有助於理解文本的全部語義和上下文。命名實體識別（NER）也因深度學習和Transformer模型的應用而受益。

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，本研究旨在開發一種能夠預測句子中命名實體邊界和分類的方法，並使用BIO格式進行命名實體識別（NER）。將採用多種評估指標，如精確度、召回率和F1分數等，來評估模型性能。此外，本研究將深入探討不同的NER方法，包括基於規則的、統計機器學習和深度學習方法，以確定最適合特定任務的方法。我們還將考察模型優化策略，包括超參數調整和特徵工程，以提高預測準確性。本研究將應用於實際文本數據，跨足不同領域和語言，並進行多層次的實驗和比較，以為自然語言處理領域的命名實體識別研究提供寶貴貢獻。

關鍵詞：自然語言處理、命名實體識別、雙向循環神經網絡、注意力機制、條件隨機場、BERT

**ABSTRACT**

This paper focuses on the field of Natural Language Processing (NLP), particularly in the research and application of Chinese Named Entity Recognition (NER). NLP is a crucial area in artificial intelligence that enables computers to understand, analyze, and generate natural language text. In recent years, the rise of deep learning and Transformer models, along with the availability of large datasets and powerful computing capabilities, has driven rapid advancements in NLP. NLP has made significant breakthroughs not only in text classification, machine translation, and automatic question-answering but also in areas such as sentiment analysis, speech recognition, and conversational system development.

It is worth highlighting the significance of Transformer models, which achieve deep understanding of text through self-attention mechanisms, improving text processing efficiency and aiding in comprehending the full semantics and context of the text. Named Entity Recognition (NER) has also benefited from the application of deep learning and Transformer models.

Chinese NER presents more challenges compared to English NER due to the Chinese language's character-based nature and the close relationship between word segmentation and NER. Incorrect word segmentation can lead to NER errors. This research will focus on the application of NLP and NER techniques in the healthcare domain to more effectively classify and label a large volume of medical texts, ultimately contributing to the establishment of more comprehensive and accurate medical databases, enhancing the quality of healthcare decision-making and research.

In summary, this research aims to develop a method for predicting named entity boundaries and classification within sentences, using the BIO format for Named Entity Recognition (NER). Various evaluation metrics, such as precision, recall, and F1 score, will be employed to assess the model's performance. Additionally, this study will delve into different NER methods, including rule-based, statistical machine learning, and deep learning approaches, to determine the most suitable method for specific tasks. We will also investigate model optimization strategies, including hyperparameter tuning and feature engineering, to enhance prediction accuracy. This research will be applied to real-text data across different domains and languages, conducting multi-layered experiments and comparisons, aiming to provide valuable contributions to the field of Named Entity Recognition in natural language processing.

Keyword：NLP、NER、Bi-LSTM、Attention、CRF、BERT

**第一章 緒論**

自然語言處理（Natural Language Processing，NLP）作為人工智慧領域的一個關鍵的分支，增加了對於自然語言的理解、分析以及生成的能力。最近幾年以來，越來越多人注重對於NLP領域的發展，主要原因是深度學習技術的創新、可用的大量資料集以及日益強大的計算能力。這使得NLP不僅僅在文本分類、機器翻譯和自動問答等應用中突破，還在情感分析、語音識別和對話系統的建構上取得了重要進展，為改進人機互動、資訊檢索和知識管理等領域帶來了新的可能性。

深度學習在NLP領域中迅速發展，特別是隨著Transformer模型的崛起。 Transformer模型透過自注意機制（self-attention）成功實現對文本資料有更深入的理解，並在處理語言中的語句和詞語時取得了極高的成就。相較於循環神經網路（Recurrent Neural Network，RNN）和長短時記憶網路（Long Short-Term Memory，LSTM）等模型，Transformer模型大幅提升了處理效率，同時能更好捕捉長時間任務的依賴關係，助於理解文本的全部語義和上下文。命名實體識別（Named Entity Recognition，NER）在NLP中也受益於深度學習和Transformer模型的應用。

NER是NLP領域的一個關鍵基礎任務，其主要目標是在非結構化文本中定位和分類命名實體的提及，例如人名、組織機構和地點等。NER不僅是NLP中的一個重要任務，還為關係抽取、事件提取、知識圖譜、機器翻譯、問答系統等多個NLP任務提供基礎支援。

傳統上，NER被視為一種序列標記問題，其中我們需要聯合預測實體的邊界和其對應的類別標籤。相對於英文NER，中文NER會更具挑戰性。中文以符號為基本單位，不像英文一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詞與分詞有著密切相關，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”可能被錯誤地分割為三個詞：“上皮組織”、“惡性”和“腫瘤”。

在數位時代，尋求醫療保健相關資訊的使用者通常會在網路上搜索和瀏覽網頁內容，以獲取有關健康議題的資訊，然後再預約醫生進行診斷和治療。網路文本是提供這些醫療保健資訊的重要來源，包括健康相關的新聞、數位健康雜誌和醫學問答社群。這些資訊涵蓋了多個專業術語和具體名詞，主要是關於醫學實體的命名，例如：中樞神經系統(central nervous system)和固有結締組織（Connective tissue proper）。

綜上所述，中文NER是NLP中一項至關重要且不可或缺的任務，旨在自動識別醫療領域中的各種實體，包括醫療設備（Instruments）、化學物質（Chemicals）、疾病（Diseases）和治療方法（Treatments）。這有助於使機器能夠讀取和理解醫學文本。

本研究預測每一個給定句子的命名實體邊界和類別。我們使用通用的BIO（開始、內部和外部）格式進行NER任務。標記前的B前置字元是命名實體的開頭，標記前的I前置字元位於命名實體內。O標記表示令牌不屬於任何命名實體。通過檢查機器預測標籤和人工解釋標籤之間的差異來評估性能。採用標準精度、召回率和F1分數，這是NER系統在字元級別上最典型的評估指標。如果以BIO格式預測的字元標籤與黃金標準（即定義的BIO標籤之一）完全相同，則測試實例中的字元被視為正確識別。精度定義為NER系統找到的正確命名實體的百分比。召回率是NER系統找到的測試集中存在的命名實體的百分比。不同的類型將獨立評估。

**第二章 文獻探討**

**第一節 詞向量表示**

在醫學文本中應用命名實體識別（NER）具有廣泛的應用，其中之一是提取重要的醫學資訊，如疾病、症狀和藥物。首先，我們在模型的第一層中使用詞嵌入技術。這些詞嵌入是通過Skip-gram方法訓練得到的向量空間模型。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組，提高了自然語言處理任務的性能表現。

這一方法的關鍵優勢在於它的能力，不僅可以識別和分類醫學文本中的命名實體，還可以説明理解它們之間的關聯和語義資訊。這對於從大量的醫學文本中提取關鍵資訊，例如患者的病史、疾病的傳播趨勢或藥物的副作用等內容，都非常有説明。

**第二節 卷積神經網路（CNN）**

卷積神經網路（CNN）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。在本研究中，我們將CNN層使用在從給予的句子中提取形態資訊。這些CNN層通常緊隨去除層（dropout layer）的應用，並且完全依賴於資料本身，而不需要手工製作的特徵工程。圖1-1展示了通過N-gram生成的多序列，這些序列對於本研究的形態資訊提取過程非常關鍵。

為了更好地理解CNN在文本處理中的應用，我們可以補充一些關於CNN工作原理的資訊。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。

**第三節 長短時記憶網路(LSTM)**

長短期記憶網絡（LSTM）是一種循環神經網路（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。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。圖2-1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(forger 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。(Huang et al., 2015)

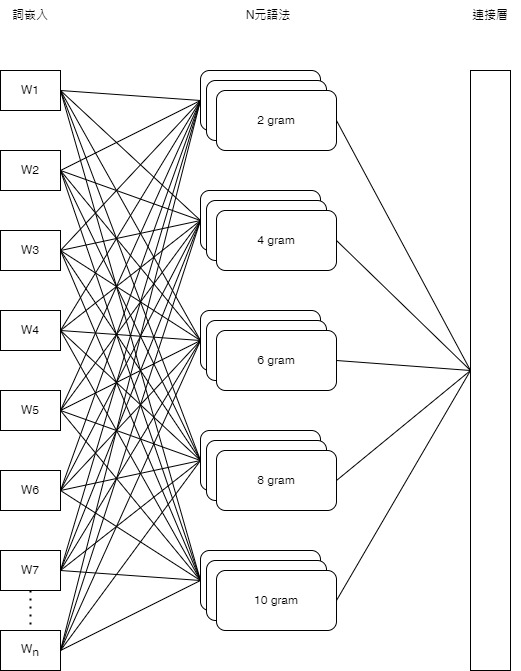


圖 1 從N元語法生成多序列

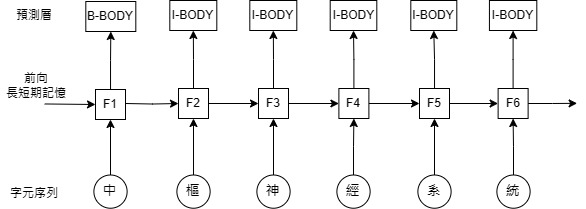


圖 2 LSTM 神經網路

**第四節 雙向循環神經網絡(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3-1)。

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。

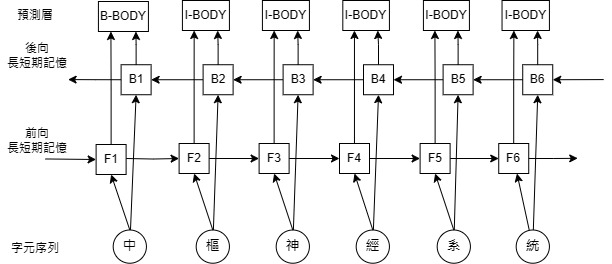


圖 3 Bi-LSTM神經網路

**第五節 條件隨機場(CRF)**

條件隨機場（CRF）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，CRF模型(圖4-1)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。(Wei et al., 2021)

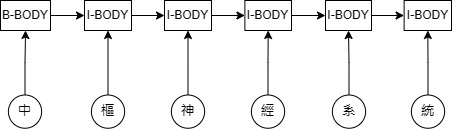


圖 4 CRF 神經網路

**第六節 注意力機制(Attention)**

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。

注意函數可以描述為將查詢和一組鍵值對映射到輸出，其中查詢（query）、鍵（keys）、值（values）和輸出都是向量。輸出計算值的加權總和，其中分配給每一個值的權重由查詢與相應鍵的相容性函數計算。(Vaswani et al., 2017)，將輸出矩陣計算為

**第七節 BERT**

BERT是一種新提出的上下文語言模型，可以改變預訓練生成的詞向量與下游特定NLP任務之間的關係。BERT框架是一訓練過程包含兩個主要步驟：預訓練和微調。在預訓練階段，BERT模型通過在龐大的未標記文本數據上進行訓練，透過這個過程涉及多個預訓練任務，如遮蔽語言模型（Masked Language Model，MLM）和下一句預測（Next Sentence Prediction，NSP）等。使模型學會了理解單詞、詞組和句子之間的關係，並捕捉語言的語義結構。

微調是BERT的第二步關鍵階段，它允許模型應用在特定下游任務上所學到的知識。在微調過程中，使用從預訓練過程中獲得的權重來初始化BERT模型。然後，我們使用來自特定下游任務的標記數據對模型進行微調。這些下游任務可以是各種自然語言處理任務，例如情感分類、文本分類、問答系統等。每一個下游任務都有一個獨立的微調模型，即使它們最初是使用相同的預訓練參數初始化的。(Devlin et al., 2018)

**第八節 評估矩陣(Evaluation matrix)**

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

，

，

= ，

=

多類別的分類任務整體通常由Micro-F1和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。(Yuan et al., 2021)

**參考文獻**

1. Chiu, J. P. C., & Nichols, E. (2015). *Named Entity Recognition with Bidirectional LSTM-CNNs*. http://arxiv.org/abs/1511.08308
2. 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
3. Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. http://arxiv.org/abs/1508.01991
4. 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
5. Wei, T., Qi, J., He, S., & Sun, S. (2021). *Masked Conditional Random Fields for Sequence Labeling*. http://arxiv.org/abs/2103.10682
6. 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
7. Lee, L. H., & Lu, Y. (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. Article 9312396. https://doi.org/10.1109/JBHI.2020.3048700
8. Lee, L. H., Chen, C. Y., Yu, L. C., & Tseng, Y. H. (2022). *Overview of the ROCLING 2022 Shared Task for Chinese Healthcare Named Entity Recognition.* 於 Y-C. Chang, Y-C. Huang, J-L. Wu, M-H. Su, H-H. Huang, Y-F. Liu, L-H. Lee, C-H. Chou, & Y-F. Liao (編輯), ROCLING 2022 - Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (頁 363-368). (ROCLING 2022 - Proceedings of the 34th Conference on Computational Linguistics and Speech Processing). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).