# 基於 RoBERTa 的中藥命名實體識別模型
# ( RoBERTa-based Traditional Chinese Medicine Named Entity Recognition Model)

**Ming-Hsiang Su, Chin-Wei Lee, Chi-Lun Hsu and Ruei-Cyuan Su**
Department of Data Science, Soochow University, Taipei, Taiwan
{huntfox.su,vivibank888, laurenhsu31,70613rex}@gmail.com

## 摘要

本研究構建了一個命名實體識別，並將其應用於中藥名稱和疾病名稱的識別，其結果可進一步用於人機對話系統，為人們提供正確的中藥用藥提醒。首先，本研究利用網路爬蟲整理網路資源，成為中藥命名實體語料庫，收集了 1097 篇文章，1412 個疾病名稱和 38714 個中藥名稱。然後，我們使用中藥名稱和 BIO 標籤方法對每篇文章進行標註。最後，本研究用 BiLSTM 和 CRF 對 BERT、ALBERT、RoBERTa、GPT2 進行訓練和評估。實驗結果表明，RoBERTa 結合 BiLSTM 和 CRF 的 NER 系統取得了最好的系統性能，其中精準率為 0.96，召回率為 0.96，F1-score 為 0.96。

## Abstract

In this study, a named entity recognition was constructed and applied to the identification of Chinese medicine names and disease names. The results can be further used in a human-machine dialogue system to provide people with correct Chinese medicine medication reminders. First, this study uses web crawlers to sort out web resources into a Chinese medicine named entity corpus, collecting 1097 articles, 1412 disease names and 38714 Chinese medicine names. Then, we annotated each article using TCM name and BIO tagging method. Finally, this study trains and evaluates BERT, ALBERT, RoBERTa, GPT2 with BiLSTM and CRF. The experimental results show that RoBERTa's NER system combining BiLSTM and CRF achieves the best system performance, with a precision rate of 0.96, a recall rate of 0.96, and an F1-score of 0.96.

關鍵字：中藥、疾病、命名實體識別模型
Keywords: Traditional Chinese Medicine, Disease, Named Entity Recognition Model

## 1 Introduction

中藥(Traditional Chinese Medicine, TCM) 是中華民族傳統藥物的總稱，根據中醫理論指導下應用的藥物 (Wiki, 2022)。台灣擁有一個融合了傳統中藥和西藥的多元化醫療環境，中藥對保健和慢性病的作用近來逐漸受到公眾的重視，中藥的使用也得到了廣泛的普及 (顏秋蘭等人, 2020; 葉明功, 2020)。民眾普遍認為中藥溫和，沒有副作用。因此，人們經常在沒有醫生處方的情況下到中藥店購買中藥，或者聽地下電台誇大療效，購買相關中藥產品，而忽視了中藥的安全性 (楊榮季, 2012)。根據統計，約有 88.2%的大眾在過去一年中有過購買和服用中藥的經歷 (顏秋蘭等人, 2020)。然而，市場上存在著各種非法的藥品廣告和不合格的產品。國人的用藥習慣不正確，往往會加重肝腎負擔，甚至誤用、混用、中西藥相互作用，導致藥物的療效或毒性發生變化，對健康的影響是不可低估的。王海征等人 (王海征等人, 2015) 研究指出銀髮族患者病情複雜，慢性病居多，加上認為中藥其中以銀髮族用藥錯誤所造成的健康影響最為嚴重。在 520 例用藥錯誤中，以用法、用量錯誤為最高，佔了 48.5%。在聊天過程中，如何提供人們正確的中藥使用方式及用藥資訊，以避免因錯誤用藥造成健康惡化，是對話系統值得研究的議題。

深度學習是目前機器學習領域最具前瞻性的方法，如對話理解 (Su et al., 2018)、對話回應生成 (Su et al., 2019)、圖像分類 (Chhillar et al., 2020) 或語音翻譯 (McCarthy et al., 2020) 等成功案例。具深度學習模型的機器，可以從

大量的數據中自行摸索出潛在的抽象規則，而不需要他人的指導。鑒於近年來深度學習出色的識別能力，本研究將應用深度學習技術於中藥命名實體識別(Named Entity Recognition, NER)，從而提供對話系統與人的互動，並提醒避免用藥錯誤。目前，中藥領域的語料庫還沒有像中文或英文語料庫那樣多樣化和豐富，這大大增加了中藥命名實體識別的難度。如果能夠克服中藥命名實體識別的問題，就有可能正確識別人機對話中人們提到的藥名，這對後續用藥提醒回應的生成有很大的幫助。對於中藥領域，目前還沒有適合使用的中藥公共知識庫。如果能通過網路爬蟲將網路資源整理成中藥知識庫，包括各種類型的中藥和中藥處方及療效，將對對話系統和問答系統有很大幫助。

現今，在命名實體識別任務中，主要採用基於循環神經網路(Recurrent Neural Network, RNN)的方法 (Chiu & Nichols, 2016; Lample et al., 2016) 作為序列標註的模型，並輔以字元級的詞向量或有其他文本特徵。Chiu 和 Nichols (Chiu & Nichols, 2016) 在序列標註模型中使用了雙向長短期記憶循環神經網路(Bidirectional Long Short-Term Memory, Bi-LSTM)。在循環神經網路輸出後，使用人工網路來確定哪些電流應該被標記。命名實體，作者使用字元級模型使用了具有一些字元特徵的卷積神經網路 (Convolutional Neural Network, CNN)，例如首字母是否大寫。

而 Lample et al. (Lample et al., 2016) 使用一層條件隨機域 (Conditional Random Fields, CRF)來判斷當前的標註結果。他們利用 CRF 的特點，讓前一個時間點的標註結果影響當前的標註，從而提高準確率。另一方面，字元層面的編碼也發生了變化，他們在字元層面使用了預先訓練好的詞向量語另一個雙向的長短期循環神經網路。最後，他們將原始單詞向量語字元級向量連接起來，作為單詞的代表向量。除了使用 RNN 來判斷命名實體外，Strubell et al. (Strubell et al., 2017) 使用疊代擴張卷積神經網路 (Iterated Dilated Convolutional Neural Networks, ID-CNN)來處理命名實體以達到加速的目的。他們通過卷積網路的特殊結構，解決了卷積網路不適合於序列標註的缺點。

NER 系統通常通過將其輸出與人類注釋進行比較來評估，這可以通過精確匹配來量化。NER 涉及到實體邊界和實體類型，通過精確匹配評估，只有當時體邊界和實體類型都與基礎事實相匹配時，命名實體才被認為是正確的 (Tjong et al., 2003; Pradhan et al., 2012; Li et al., 2020)。由於大多數NER系統涉及多種實體類型，因此通常需要評估所有實體類型的性能，這方面通常使用兩種方法，宏觀平均的 F-score 和微觀平均的 F-score，宏觀平均 F-score 是針對每個實體類型獨立計算的，然後取其平均值。微觀平均 F-score 將所有類別的實體貢獻彙總，計算出一個平均值。後者可能會受到語料庫中大類別識別實體質量的嚴重影響。

## 2 Dataset Collection

在台灣，雖然人們經常使用中藥來保養身體，但很少有人整理中藥數據集來訓練 NER 模型。在本研究中，採用爬蟲從 KingNet 網站 (KingNet, 2022) 和 CloudTCM 網站 (CloudTCM, 2022) 檢索中藥數據。對於中藥和處方的種類，根據其名稱、功效、用法、禁忌等，自動整理成適合NER模型的中藥數據集。在 KingNet 網站上，共收集了 730 篇文章，包含 678,846 個單詞;而在 CloudTCM 網站上，共收集了 367 篇文章，包含 1,219,168 單詞。中藥名稱和處方名稱的例子見 Table1。然後本研究採用內-外-開始 (IOB) 進行標註，我們將中藥名稱標註為 ”B-TMC” 和 “I-TMC”，症狀標註為 “B-SYMP” 和 “I-SYMP”，其他標註為 “O”。例如，”人蔘的安神益智功效主要表現在促進學習記憶方面。適量人蔘可安神舒眠，緩解壓力” 被標為 “人 B-TMC” 和 “蔘 I-TMC”，其他的則被標為 ”O”。

| | |
|---|---|
| **中藥名稱** | 白芷、薄荷、人蔘等。 |
| **處方名稱** | 桂枝湯、溫脾湯、清脾飲等。 |

Table 1: 中藥名稱和處方名稱的例子。

## 3 Proposed Methods

### 3.1 Word Embedding

語言模型預訓練已被證明在改善許多自然語言處理任務方面是有效的 (Devlin, 2018)。這些任務包括句子級任務和標記級任務，如自然語言推理和 NER。在本研究中，我們使用 Bidirectional Encoder Representations from

Transformers (BERT) 作為詞嵌入模型，其中 BERT 是 Google AI 團隊近年來發布的自然語言預訓練模型。

BERT 是一種自然語言預訓練模型，與其他預訓練語言模型相比，它的訓練方法更加新穎。它採用 Transformer 的編碼器雙向連接，在訓練雙向語言模型時採用兩個無監督的預測任務，即遮罩語言模型 (Masked Language Model, MLM) 和下一句預測 (Next Sentence Prediction, NSP)。雙向和單向的區別主要是由於單詞語言表現的訓練方向不同。單向語言模型會根據每個詞的左邊或右邊的詞來訓練每個詞的語言表示。例如，假設你想得到”我訪問了銀行帳戶”句子中”銀行”這個詞的語言表示。單向語言模型根據”銀行”左邊的”我訪問了”而不是”銀行”右邊的”帳戶”來訓練”銀行”這個詞。對於雙向語言模型，通過考慮”我訪問了”和”帳戶”來訓練該詞的詞嵌入。一個好的詞語語言表現對於 Natural language processing (NLP) 任務非常重要，而雙向語言模型可以讀取雙向訊息，所以詞語的語言表現會比單向的好。

但 BERT 的作者解釋說，一般語言模型的雙向發展可能是因為它可以間接地”看到自己”，所以預測單詞時只需要直接考慮它的已知上下文訊息，為了解決雙向語言模型面臨的問題，BERT 提出了在預測單詞的任務中加入遮罩的訓練技術，將輸入句子中 15%的單詞遮罩，並預測這些遮罩的單詞。該任務實例如 Table 2 所示。如何選擇遮擋的比例是一個問題。首先，如果所選單詞 100% 被遮擋，會導致模型只通過遮擋來學習上下文語言表現。對於未被遮擋的詞，很難學習到好的語言表現。其次，由於遮擋本身並不出現在實際的預測階段，為了迫使模型關注所有的詞，一定比例的被選中的詞並設有被遮罩，而是被替換成其他的詞或者保持不變。通過這種訓練機制，模型可以學習到更好的上下文語言表現。

| **Input:** |
| --- |
| The man [MASK1] to [MASK2] store |
| **Label:** |
| [MASK1] = went ; [MASK2] = store |

Table 2: 遮罩語言模型預測人物的例子。

BERT 訓練策略中加入的另一項創新是其他語言模型沒有考慮的兩個句子之間的關係，這也是許多自然語言任務的一個重要特徵。因此，為了讓模型學習句子之間的關係，將給出兩個句子 A 和 B，模型將判斷 B 是否是真實語料庫中 A 的下一句。任務實例見 Table 3，預測兩個句子間的關聯學習深度雙向的上下文表示。此外，BERT 還有實驗表明，在模型的輸出中加入線性層，通過微調可以在各種自然語言處理任務中表現良好，適用於自然語言任務，如情感分類或問題回答。BERT 的輸入是標記嵌入、分段嵌入、位置嵌入，如 Figure 1 所示，還有兩個特殊符號[CLS]、[SEP]。[CLS]可用於後續的自然語言分類任務，而[SEP]則用於區分兩個句子。

有許多基於 BERT 的改進模型，包括 ALBERT (Lan et al., 2019)，RoBERTa (Liu et al., 2019) 和 GPT2 (Lagler, 2013)。本研究將評估這些不同的模型，以獲得 NER 系統的最佳性能。

| **Input:** |
| --- |
| The man went to the store [SEP] he bought a gallon of milk |
| **Label:** |
| IsNext |
| **Input:** |
| The man went to the store [SEP] penguins are flightless birds |
| **Label:** |
| NotNext |

Table 3: 下一句話預測任務的例子。

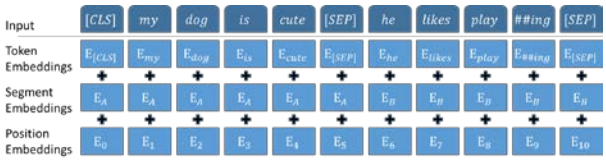


Figure 1: BERT 輸入表示的示意圖。

### 3.2 Bi-LSTM

長短期記憶 (Long Short-Term Memory, LSTM) 是一種特殊的 RNN。與傳統的 RNN 不同，LSTM 使用三個不同的閥來控制單元的狀態。這些閥是輸入閥、遺忘閥和輸出閥。這三個閥在 Figure 2 中分別三個綠框表示。

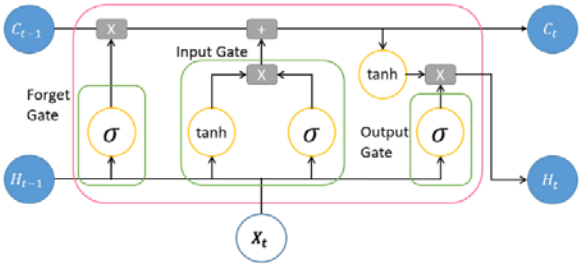


Figure 2: LSTM 模型的示意圖。

遺忘閥通過 (1) 來控制遺忘，其中$W_f$和$U_i$代表要與前一個時間點的輸出和當前輸入相乘的權重矩陣，$h_{t-1}$代表前一個時間點的輸出，$X_t$代表當前輸入，$b_f$代表偏移量向量，所得的$f_t$可以決定哪些訊息應該被遺忘。輸入閥分為兩小部分，一部分稱為候選狀態向量$\tilde{\boldsymbol{c}}_t$和輸入閥向量$\boldsymbol{i_t}$，操作方法為 (2) 和 (3)，其中$W_c, W_i$, $U_c$ 和 $U_i$代表權重矩陣，$\boldsymbol{b_c}$和 $\boldsymbol{b_i}$代表偏移量向量。

用這兩個向量$\tilde{\boldsymbol{c}}_t$和 $\boldsymbol{i_t}$來控制多少個單元狀態受到當前輸入的影響，新的單元狀態$\boldsymbol{c_t}$將由$f_t$, $\boldsymbol{c_{t-1}}$, $\boldsymbol{i_t}$和 $\tilde{\boldsymbol{c}}_t$決定，如 (4) 所示。輸出閥是為了控制將輸出多少個單元狀態，如(5)所示，這也是由當前的輸入$\boldsymbol{X_t}$和前一輪的輸出$h_{t-1}$決定的。最後，本輪的輸出向量$h_t$取決於本輪的單元狀態$\boldsymbol{c_t}$和輸出閥的向量$\boldsymbol{o_t}$，如(6)所示。由於這些閥的機制，LSTM 可以記住長期的依賴關係。

$$f_t = \sigma(W_f h_{t-1} + U_i X_t + \boldsymbol{b_f}) \quad (1)$$
$$\tilde{\boldsymbol{c}}_t = tanh(W_c h_{t-1} + U_c X_t + \boldsymbol{b_c}) \quad (2)$$
$$\boldsymbol{i_t} = \sigma(W_i h_{t-1} + U_i X_t + \boldsymbol{b_i}) \quad (3)$$
$$\boldsymbol{c_t} = f_t \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \tilde{\boldsymbol{c}}_t \quad (4)$$
$$\boldsymbol{o_t} = \sigma(W_o \boldsymbol{h_{t-1}} + U_o \boldsymbol{X_t} + \boldsymbol{b_o}) \quad (5)$$
$$\boldsymbol{h_t} = \boldsymbol{o_t} * \tanh(\boldsymbol{c_t}) \quad (6)$$

大多數LSTM的輸出會是一個或多個向量，與地面實況相比，得到兩者之間的誤差，然後通過隨機梯度下降或其他優化算法矩陣更新網路中的權重。由於網路中存在多個閥，大大降低了部分分化過程中梯度消失或爆炸的可能性，這是 LSTM 比一般 RNN 的優勢。Bi-LSTM 是一種雙向 LSTM，其結構圖如 Figure 3 所示。Bi-LSTM 用於學習時間序列的依賴關係，通過訓練輸入閥、遺忘閥和輸出閥的權重來學習序列輸入中應該注意的關鍵點。

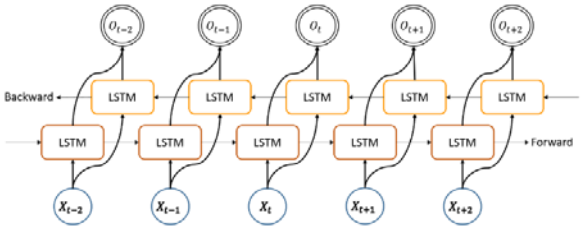


Figure 3: Bi-LSTM 模型的示意圖。

### 3.3 CRF

輸入序列中的每個向量進入 Bi-LSTM，並與前一個時間點的隱藏向量相匹配，判斷當前時間點的輸出。這個輸出將作為特徵進入 CRF 層，並讓 CRF 學習每個權重的特徵函數，如 Figure 4 所示。CRF 的訓練方法主要有兩個步驟。第一步是由訓練數據集生成特徵函數，並初始化每個特徵函數對應的權重。第二步是使用最大似然估計、梯度下降等方法來更新每個特徵函數的權重，直到權重變化收斂。以 Bi-LSTM 和 CRF 作為 NER 模型，對於 CRF 來說，Bi-LSTM 在每個時間點的輸出序列就是 CRF 的觀察向量，可以將命名實體的標註序列與 CRF 預測的序列進行比較，計算出誤差函數的梯度。通過反向傳播算法，這個誤差梯度被反饋給 Bi-LSTM 和 CRF，權重可以被更新以最小化誤差。

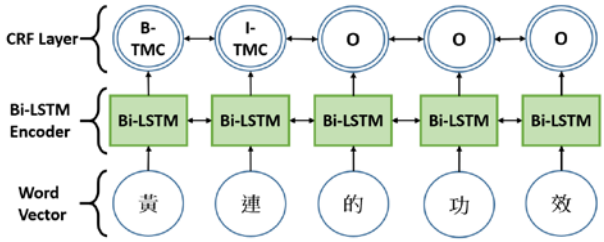


Figure 4: NER 模型的示意圖。

## 4 Experimental Results and Discussion

本研究針對命名實體識別任務提出了 BERT, Bi-LSTM 和 CRF 串聯模型，用 BERT 表示單句的訊息關係，然後通過 Bi- LSTM 和 CRF 模型判斷命名實體的標註位置。本研究比較了 BERT、ALBERT、RoBERTA 和 GPT2 結合 Bi-LSTM 和 CRF 模型的性能，以確定最終 NER 系統的架構。測試集的實驗結果表明，BERT 的精確度為 0.86，召回率為 0.91，F1-score 為 0.89; ALBERT 的精確度為 0.93，召回率為 0.94，

F1-score 為 0.93; RoBERTA 的精確度為 0.96，召回率為 0.96，F1-score 為 0.96; GPT2 的精確度為 0.93，召回率為 0.92，F1-score 為 0.92。

Table 4 和 Table 5 分別顯示了不同模型在 TMC、SYMP 標籤、微觀平均值、宏觀平均值和加權平均值上的實驗結果。實驗結果顯示，RoBERTa 模型優於 BERT、ALBERT 和 GPT2 模型。我們認為，在我們收集的 TCM 語料庫中，RoBERTa 模型更能夠提取語料相關的訊息，這導致了最佳的整體性能。

| | BERT | | | ALBERT | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| SYMP | 0.66 | 0.83 | 0.74 | 0.89 | 0.88 | 0.88 |
| TMC | 0.5 | 0.56 | 0.53 | 0.68 | 0.75 | 0.71 |
| micro avg | 0.86 | 0.91 | 0.89 | 0.93 | 0.94 | 0.93 |
| macro avg | 0.79 | 0.85 | 0.82 | 0.89 | 0.9 | 0.9 |
| weighted avg | 0.88 | 0.91 | 0.89 | 0.93 | 0.94 | 0.94 |

**P**: 精確度; **R**: 召回率; **F1**: F1-score

Table 4: 對 BERT 和 ALBERT 的評估。

| | RoBERTa | | | GPT2 | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| SYMP | **0.92** | **0.91** | **0.92** | 0.82 | 0.79 | 0.81 |
| TMC | **0.8** | **0.79** | **0.79** | 0.66 | 0.62 | 0.64 |
| micro avg | **0.96** | **0.96** | **0.96** | 0.93 | 0.92 | 0.92 |
| macro avg | **0.93** | **0.92** | **0.93** | 0.87 | 0.85 | 0.86 |
| weighted avg | **0.96** | **0.96** | **0.96** | 0.93 | 0.92 | 0.92 |

**P**: 精確度; **R**: 召回率; **F1**: F1-score

Table 5: 對 RoBERTa 和 GPT2 的評估。

## 5 Conclusion and future work

在這項研究中，構建了命名實體識別模型，並將其應用於中藥名稱和疾病名稱的識別。其結果可進一步用於人機對話系統，為人們提供正確的中藥用藥提醒。此外，本研究利用網路爬蟲將網路資源整理成中藥命名實體語料庫，共包括 1097 篇文章、1412 個疾病名稱和 38714 個中藥名稱。然後我們用中藥名稱和 BIO 標籤方法對每篇文章進行標籤。最後，實驗結果表明，RoBERTa 組合 Bi-LSTM 和 CRF 的 NER 系統取得了最好的系統性能，其中精準度為 0.96，召回率為 0.96，F1-score 為 0.96。

在未來的工作中，我們希望能獲得更多的中藥對話數據集，以便我們能訓練出更適合對話系統的 NER 系統。此外，我們還希望在 NER 系統中加入自我注意力的機制，以提高系統性能。最後，我們希望擴大 NER 的標籤，使 NER 能夠識別更多語中藥有關的命名實體。

## References

王海征, 林曉蘭, 張鵬, 王雅葳, and 陳文強. 2015. 老年患者中藥用藥錯誤報告 520 例分析. *藥物不良反應雜誌*. 17(5): 353.

楊榮季. 2012. 「老人」及「婦女」醫學保健之用藥安全調查與知能研究. *行政院衛生署中醫藥年報*, 1(6): 1-120.

顏秋蘭, 黃林煌和葉明功. 2020. 藥師介入提升民眾中醫藥就醫用藥安全. *藥學雜誌*, 29(3).

葉明功. 2020. 中醫藥就醫用藥之停、看、聽、選、用專業. 南投醫院. Retrieved October 12, 2020, from https://www.nant.mohw.gov.tw/public/ufile/c909c79547bd1528856c92f9a08e9361.pdf.

Ankit Chhillar, Sanjeev Thakur, and Ajay Rana. 2020. Survey of Plant Disease Detection Using Image Classification Techniques. In *Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, pp. 1339-1344.

Arya D. McCarthy, Puzon Liezl, and Pino Juan. 2020. SkinAugment: auto-encoding speaker conversions for automatic speech translation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7924-7928.

Chiu, Jason PC, and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4: 357-370.

CloudTCM website, Retrieved October 11, 2022, from https://cloudtcm.com/

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pp. 2670-2680.

Erik F. Tjong Kim Sang, and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 142-147.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the Conference of the North American Chapter of the Association for*

*Computational Linguistics: Human Language Technologies*, pp. 260-270, 2016.

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

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1): 50-70.

K. Lagler, Schindelegger, Michael, Böhm, Johannes, Krásná, Hana, and T. Nilsson. 2013. GPT2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6): 1069-1073.

KingNet website, Retrieved October 11, 2022, from https://www.kingnet.com.tw/tcm/

Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, and Chu-Kwang Chen. 2018. Attention-based dialog state tracking for conversational interview coaching. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6144-6148.

Ming-Hsiang Su, Chung-Hsien Wu, and Liang-Yu Chen. 2019. Attention-Based Response Generation Using Parallel Double Q-Learning for Dialog Policy Decision in a Conversational System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 131-143.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task*, pp. 1-40.

Wiki. Traditional Chinese medicine. Retrieved October 11, 2022, from https://en.wikipedia.org/wiki/Traditional_Chinese_medicine.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite Bert for self-supervised learning of language representations. *arXiv preprint* arXiv:1909.11942.