# 中文醫療文件的命名實體辨識報告
# MIGBaseline at ROCLING 2022 Shared Task: Report on Named Entity Recognition Using Chinese Healthcare Datasets

**馬行遠** Hsing-Yuan Ma　**李韋杰** Wei-Jie Li　**劉昭麟** Chao-Lin Liu
Department of Computer Science
National Chengchi University
{110753132, 110753128, chaolin} @g.nccu.edu.tw

## 摘要

命名實體（Named Entity Recognition , NER）工具發展已久，但少有針對醫療專業領域的 NER 工具，因此建立一個適用於醫療文件的 NER 工具是至關重要的。本研究使用了在中英任務中表現出色的 W2NER 模型 ，藉由更改資料的輸入、選用不同的預訓練語言模型以及運用不同的訓練策略，建立一個適合於中文醫療資料集的 NER 模型。我們的最佳模型在該資料集獲得 81.93%的 F1 分數 ，並在 ROCLING 2022 NER 競賽(Lee et al., 2022)中排名第一。

## Abstract

Named Entity Recognition （NER） tools have been in development for years, yet few have been aimed at medical documents. The increasing needs for analyzing medical data makes it crucial to build a sophisticated NER model for this missing area. In this paper, W2NER, the state-of-the-art NER model, which has excelled in English and Chinese tasks, is run through selected inputs, several pretrained language models, and training strategies. The objective was to build an NER model suitable for healthcare corpora in Chinese. The best model managed to achieve an F1 score at 81.93%, which ranked first in the ROCLING 2022 shared task.

關鍵字: 命名實體辨識、W2NER、醫療、中文
Keywords: NER, W2NER, Healthcare, Chinese

## 1　簡介

命名實體辨識（Named Entity Recognition, NER）在自然語言當中一直是非常重要的一個技術，該技術藉由標記資料來訓練模型，主要處理書籍、字典、新聞等一些非結構化文本，進行專有名詞的抽取與標記，主要針對一些重要的實體，通常包含人名、地名與專有名詞。抽取出來的詞組可以用來分析情意、關係擷取、事件追蹤…等功能。這技術還能讓斷詞（Word Segmentation, WS）的結果更加準確，因此大部分的斷詞工具都會使用這項技術。

現在通用的 NER 技術已經行之有年，技術也一直在進步，而各個專業領域隨著時間發展所創造的詞彙也越來越多，加上艱澀不成用的專業詞彙並不會在通用型 NER 中訓練，導致通用型 NER 在專業領域的標記結果不佳，也顯示基於專業領域資料所開發的 NER 模型的重要性。

至今許多領域的發展越來越離不開資訊與科技的協助，醫療產業也不例外。病人的醫療紀錄與問診都會產生出需要整理的資料，因此能處理醫療文字資料的 NER 已呈迫切的需求，因此我們希望藉由此研究提高相關主題的 NER 準確度滿足相關需求。

## 2　文獻探討

### 2.1　中文 NER 工具發展

NER 技術已經發展多年，實作方式也經過了多次的迭代，相關技術的演進可以分成三個階段(Lee & Lu, 2021)，1.傳統方法：rule-base、大量字典檔 2.機器學習方法：隱馬可夫模型（Hidden Markov Model, HMM）、最大熵馬可夫模型（Maximum Entropy Markov Model, MEMM）、條件隨機場（Conditional Random Field, CRF），University of Stanford 開發的 stanfordNLP 就是運用 CRF 技術完成的 3.深度學習方式：RNN-CRF、CNN-CRF、transformer、attention，例如 CKIP-transformer(Li et al., 2020)

中文 NER 領域的發展也在近期取得了大量的進展，從只有三大套件，Jieba(Sun, 2020)、UnivJersity of Stanford 開發的 StanfordNLP(Manning et al., 2014)、中研院開發

的CKIP(Ma & Chen, 2003)慢慢到現在有更先進與多功能的工具問世，像是最近有名的，中國中央師範大學學所開發的 NLP 工具 HanLP(He & Choi, 2021)以及 University of Stanford 開發的 Stanza(Qi et al., 2020)。

### 2.2 NER 模型與技術介紹

W2NER(Li et al., 2022)有別於常見的 NER 模型將 NER 任務分成四大類的做法，選擇將任務簡化為字與字之間的三種關係分類:

- None：表示兩個字之間沒有關係，且並不屬於同個實體。
- NNW：即 Next-Neighboring-Word，表示這兩個字是在同一個實體中相鄰
- THW-*：即 Tail-Head-Word-*，表示這兩個字是在同一實體中，且分別是開始與結尾。。

使其能夠統一解決扁平實體（flat）、重疊實體（overlapped）以及非連續實體（discontinuous）的 NER 任務。

一個簡單的範例可以參考圖 1 的（a），裡面有兩個症狀實體"aching in legs"和"aching in shoulders"，分別當作 e1 和 e2，該模型會將此資料轉換成關係陣列（如圖 2），並透過陣列釐清關係推導出圖 1 的（b）

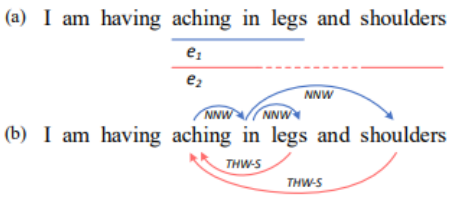


圖 1.NER 任務示意圖 [1]

W2NER 架構主要分成三層（如圖 3），(1). Encoder layer （2）. Convolution layer （3）. Co-predictor layer，在 encoder layer 中，我們將文章經由 BERT 以及 BiLSTM 轉換，得到詞向量，接著輸入 convolution layer，經由 Conditional Layer Normalization 取得 distance、word、region embeddings，接著將這些 embeddings 經由 dilated convolution 處理，輸入至 Co-predictor layer，由 biaffine predictor 以及 multi-layer perceptron predictor 生成字與字的關係矩陣。

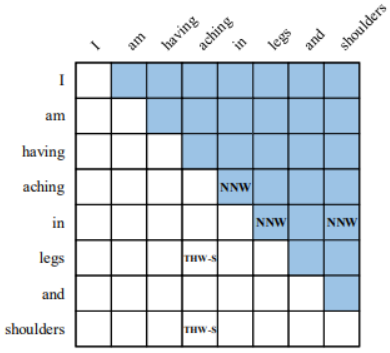


圖 2 .W2NER 生成矩陣 [1]

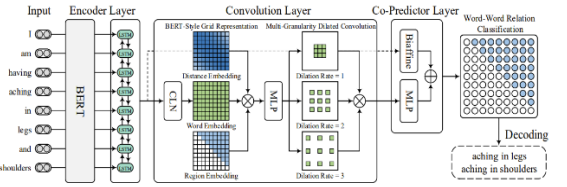


Figure 3: Overall NER architecture. CLN and MLP represent conditional layer normalization and multi-layer perceptron. ⊕ and ⊗ represent element-wise addition and concatenation operations.

圖 3.W2NER 架構圖 [1]

Google 於 2018 年發表了一個預訓練的 Transformer 模型 BERT（ Bidirectional Encoder Representations from Transformers）(Devlin et al., 2018)裡面的主要結構為 Transformer(Vaswani et al., 2017)的 encoding 層，訓練方式為使用英文維基百科與 BookCorpus 資料集配合遮罩預測與下句預測（Next sentence prediction, NSP）的訓練任務。 這個模型如此成功的原因主要是因為其 Context-Based Embedding 的向量轉換方式，他能依上下文的關係給相同的字不同 vector 而不是傳統的 Context-free embedding 方式，像是 word2vec(Mikolov et al., 2013)，因此該模型成為了少數能考量前後文的語言模型，且因為該模型在做下層任務的時候還會改變他的變數，因此可以進行預訓練與微調，而這樣的訓練方法不但獲得比 Feature-based 模型還要多的資訊，還可以針對目標任務進行微調。BERT 有許多變種的模型，本次實驗就選用了哈爾濱工業大學的 PERT(Cui et al., 2022)、

[1]引用來源 Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhanget al.Fei Li. (2022). Unified named entity recognition as word-word relation classification. Proceedings of the AAAI Conference on Artificial Intelligence,

MacBERT(Cui et al., 2020)和 Facebook 的 RoBERTa(Liu et al., 2019)與一同進行比較，這些模型與 BERT 的差別如下：

1. RoBERTa（A Robustly Optimized BERT）：此模型的目的就是要最佳化原本的 BERT 模型，因此該模型在 BERT-large 的基礎上加上了 CC-NEW、OPENWEBTEXT、STORIES 等 160GB 的資料集、更大 bacth size 與動態遮蔽字的訓練方式，動態遮罩方式主要是在資料及輸入的時候才動態產生，這樣就能夠在不同 epoch 相同資料有不同的遮罩，最後比較特別的地方是該模型移除了下一句預測的任務。
2. PERT（PRE-TRAINING BERT WITH PERMUTED LANGUAGE MODEL）：此模型是使用 BERT 原來的模型，僅更改遮罩預測任務的訓練方式，主要的差別在於他不使用遮罩的方式進行訓練（如圖 1），而是利用全詞遮罩（WWM）選定詞組並使用 Ngram 的方式將常見的前後字或是片語打亂掉，並去掉了下一句預測。這樣的好處在於不再使用 MASK 標記，能使訓練集更加接近測試集的樣子，準確度也跟著提高了不少。該中文模型的預訓練集為 EXT 數據集 [2]

| | Input | Output |
|---|---|---|
| **Original Text** | 研究表明这一句话的顺序并不影响阅读。 | - |
| **WordPiece** | 研 究 表 明 这 一 句 话 的 顺 序 并 不 影 响 阅 读 。 | - |
| **BERT** | 研 究 表 明 这 一 句 [M] 的 顺 [M] 并 不 [M] 响 阅 读 。 | $Pos_7 \rightarrow$ 话<br>$Pos_{10} \rightarrow$ 序<br>$Pos_{13} \rightarrow$ 影 |
| **PERT** | 研 究 明 表 这 一 句 话 的 顺 序 并 不 响 影 阅 读 。 | $Pos_2 \rightarrow Pos_3$<br>$Pos_3 \rightarrow Pos_2$<br>$Pos_{13} \rightarrow Pos_{14}$<br>$Pos_{14} \rightarrow Pos_{13}$ |

圖 1.PERT 與 BERT 差異 [3]

3. MacBERT：此模型是在 BERT 原本的基礎上修改了遮罩預測任務的訓練方式的遮罩方式，改用一種偵錯遮罩模型（MLM as correction，Mac）的方式。這種遮罩方式主要的差別在於它會在原有的遮罩基礎上去使用全詞遮罩（WWM）並使用 Ngram 的方式將常見的前後字或是片語直接遮蔽掉，再利用相近詞或是隨機詞去替換掉（如圖 2），這種遮罩方式可以提升詞之間的關聯度，相近詞的採用也使得模型獲得了更多預測的資訊，因此結果比原來有顯著提升。中文模型的預訓練集為 EXT 數據集 [2]

| | Chinese |
|---|---|
| **Original Sentence** | 使用语言模型来预测下一个词的概率。 |
| **+ CWS** | 使用 语言 模型 来 预测 下 一个 词 的 概率 。 |
| **+ BERT Tokenizer** | 使 用 语 言 模 型 来 预 测 下 一 个 词 的 概 率 。 |
| **Original Masking** | 使 用 语 言 [M] 型 来 [M] 测 下 一 个 词 的 概 率 。 |
| **+ WWM** | 使 用 语 言 [M] [M] 来 [M] [M] 下 一 个 词 的 概 率 。 |
| **++ N-gram Masking** | 使 用 [M] [M] [M] [M] 来 [M] [M] 下 一 个 词 的 概 率 。 |
| **+++ Mac Masking** | 使 用 语 法 建 模 来 预 见 下 一 个 词 的 几 率 。 |

圖 2.偵錯遮罩模型範例 [4]

### 2.3 NER 與醫療

NER 技術在醫學用途上一直都有需多應用，最近最大的挑戰就是完成病例分類與建檔，當中比較大的問題是分類項目的特殊性還有過多的專有名詞並不適合用通用性的 NER 來處理。因此華碩公司裡的 AICS 小組就開發 ALFER-BERT 模型來處理這件事情。

中國知識圖譜與語意計算大會（CCKS）也從 2017 年開始到 2020 年每年開放一份電子病歷的資料集給 NER 的開發者使用。以 2020 年的資料集為例，裡面包含了訓練集和測試集，其中訓練集包括 1050 個醫療記錄集，共有六大類項目（包括診斷和診斷、檢查、檢驗、原始數據、藥物、樣品測試）在當時有一組運用 BERT 模型對該份資料集做 NER 預測獲得了 91.54%的準確度(晏阳天 et al., 2020)，因此我們接下來打算去尋找類 BERT 的模型進行訓練。

## 3 實驗方法

### 3.1 實驗環境

本次實驗以 Ubuntu 20.04 的系統下運用 Nvidia GeForce RTX 3090 的 GPU 做為實驗環境，Python 與相關套件的本版如下:

[2] 由哈工大訊飛聯合實驗室中文維基百科，其他百科、新聞、問答等資料，數量多達 5.4B

[3] 引用來源 Yiming Cui, Ziqing Yang, and Ting Liu. (2022). PERT: Pre-training BERT with Permuted Language Model. *arXiv preprint arXiv:2203.06906*.

[4] 引用來源 Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wanget al.Guoping Hu. (2020). Revisiting Pre-Trained Models for Chinese Natural Language Processing.*Findings of the Association for Computational Linguistics: EMNLP 2020* Online.

- Python version: 3.8.10
- Torch version：1.8.0
- Cuda version： 11.1

### 3.2 資料處理

Chinese Healthcare NER Corpus 由中央大學電機系的自然語言實驗室（NCUEE NLP Lab）所製作(Lee & Lu, 2021)，內容為健康相關醫學新聞與醫學問答論壇的文章，裡面包含30,692 個句子、10 種實體類型共 68,460（如表 1）

| 實體類型 | 範例 |
|---|---|
| 身體 | 細胞核、神經組織 |
| 症狀 | 流鼻水、失眠 |
| 醫療器材 | 血壓計、達文西手臂 |
| 檢驗 | 聽力檢查、腦電波圖 |
| 化學物質 | 去氧核糖核酸 |
| 疾病 | 小兒麻痺症、帕金森氏症 |
| 藥物 | 阿斯匹靈、普拿疼 |
| 營養品 | 維他命、膠原蛋白 |
| 治療 | 藥物治療、胃切除術 |
| 時間 | 嬰兒期、幼兒時期 |

表 1.實體類型表格

#### 3.2.1 資料格式

| 參數 | 意義 | 範例 |
|---|---|---|
| Id | 流水號 | 001 |
| genre | 類型 | ‘SM’ |
| sentence | 句子 | 多種維生素 |
| word | 斷詞結果 | [多種", "維生素] |
| word_label | 每個詞的 NER 標記 | ["O", "SUPP",] |
| 字符 | 切字 | ["多", "種", "維", "生", "素"] |
| character_label | 每個字的 NER 標記 | ["O", "O", "B-SUPP", "I-SUPP", "I-SUPP"] |

表 2.資料格式表格

#### 3.2.2 資料前處理

我們首先分析標記內容是否有誤，將出現頻率低於 5 次的標記內容，用人工的方式檢查，並將我們認為明顯有誤的標記內容改正（如表 3）。

| 文字 | 原始標記 | 更正文字 | 更正標記 |
|---|---|---|---|
| “上淋”“巴” | （BODY）（O） | “上”“淋巴” | （O）（BODY） |
| “人參” | （DISE） | “人參” | （DRUG） |
| “放、化療” | （TREAT） | “放化療” | （TREAT） |
| “腫漲” | （SYMP） | “腫脹” | （SYMP） |

表 3.資料修改列表

### 3.3 Encoder 模型選擇

我們的實驗方法主要分成四個方向，第一個方向是嘗試更改 W2NER 輸入層使用的 Encoder 預訓練模型，找出最適合該資料集的 Encoder 預訓練模型。

- $\text{BERT}_{\text{base}}$：110M parameters
- $\text{PERT}_{\text{base}}$：110M parameters
- $\text{RoBERTa}_{\text{large}}$：355M parameters
- $\text{MacBERT}_{\text{large}}$：324M parameters
- $\text{PERT}_{\text{large}}$：330M parameters

### 3.4 統一格式實驗

第二個方向則是統一資料集的格式，鑒於 BERT、PERT、RoBERTa、MacBERT 處理 token 時全形半形會視為不同的 token，因此我們將英文數字、標點符號統一成全形或半形，藉此比較哪種格式會有較好的表現。

### 3.5 斷句實驗

第三個方向則是以句子還是以完整文本輸入的比較，我們考慮在 NER 任務中，標記的內容應該主要以句子為單位，即不需要看完整文本，只看句子也可以標記出實體位置。因此我們比較將文本經由段落標記（逗號、句號、問號與驚嘆號）切割以及保留完整內容的資料型態對於模型的結果是否有影響。

### 3.6 斷詞實驗

第四個方向則是比較有無斷詞資訊是否影響模型效果，基於 W2NER 預設以字元輸入模型，我們參考(Lee & Lu, 2021)中，將斷詞輸入模型，藉此得到更好的結果。我們認為加入斷詞資訊會影響模型效果，因為若 NER 標記皆為一個詞，使用 W2NER 生成字與字的關係矩陣，就可在 Co-predictor layer 把問題簡化，因此我們使用以下三種不同的斷詞法進一步將斷詞資訊輸入至模型中，並比較輸入字元以及輸入詞彙（如表 4）對於模型的結果是否有影響。

- CKIP transformer
- Finetuned CKIP transformer
- 資料集原始的斷詞資訊

| Type | Input | Predict |
|---|---|---|
| Char | ["雞", "蛋", "含", "有", "多", "種", "維", "生", "素", "，", "包", "括", "D", "和", "K"] | [Index:[6,7,8] Type: SUPP] |
| Word | ["雞蛋", "含有", "多種", "維生素", "，", "包括", "D", "和", "K"] | [Index:[3] Type: SUPP] |
| Sentence | ["雞", "蛋", "含", "有", "多", "種", "維", "生", "素"], ["包", "括", "D", "和", "K"] | [Index:[6,7,8] Type: SUPP], [ ] |

表 4. 不同資料型態對應預測標籤之比較

## 4 實驗結果

我們的實驗結果使用 Precision/Recall/F1-score（P/R/F1）評估指標，其中比較的訓練集、測試集以及其他參數除 cross-validation 有切割資料集外其餘皆為固定。

### 4.1 Encoder 模型結果

實驗結果（如表 5）顯示，PERT 的結果略為高於 BERT，而 Large 的模型皆優於 base 的模型，其中 $PERT_{large}$ 有最佳的表現，因此以下的實驗皆會使用 $PERT_{large}$ 作為 encoder 模型。

| | P | R | F1 |
|---|---|---|---|
| $BERT_{base}$ | 77.40 | 75.26 | 76.32 |
| $PERT_{base}$ | 76.19 | 77.10 | 76.64 |
| $RoBERTa_{large}$ | 76.82 | 76.66 | 76.74 |
| $MacBERT_{large}$ | **78.26** | 76.15 | 77.19 |
| $PERT_{large}$ | 76.46 | **78.29** | **77.36** |

表 5.模型實驗結果

### 4.2 統一格式差異

我們發現使用全形資料集訓練的模型，比用半形資料集的模型提高 F1 約 0.5%（如表 6），其原因可能為 PERT 預訓練時的資料集與統一全形的醫療資料集分布較類似，之後的實驗皆使用統一全形資料集。

| | P | R | F1 |
|---|---|---|---|
| 統一半形 | 77.17 | 78.14 | 77.65 |
| 統一全形 | **77.67** | **78.36** | **78.01** |

表 6.統一規格比較實驗結果

### 4.3 斷句結果

實驗結果（如表 7）顯示，斷句不能加強模型的表現，其原因可能在於情境線索對於標記實體是重要的，可以看到缺少上下文訊息的模型雖然 Precision 有約 1.5%的提升，但 Recall 有接近 5%的下降。

| | P | R | F1 |
|---|---|---|---|
| Baseline | 77.67 | **78.36** | **78.01** |
| Sentence | **78.93** | 73.48 | 75.93 |

表 7.斷句結果比較實驗結果

### 4.4 斷詞結果

我們發現在沒有任何預訓練的情況下，使用 CKIP-transformer 斷詞的模型表現比沒有斷詞的 baseline 還差，可見錯誤斷詞會造成模型的結果下降。在使用 finetuned 的 CKIP-transformer 斷詞後，訓練的模型結果有顯著提升，而使用原始斷詞訓練的模型甚至可以比 baseline 模型的 F1 還要高出將近 10%，可見斷詞資訊的好壞顯著影響模型結果（如表 8）。

| | P | R | F1 |
|---|---|---|---|
| Baseline | 77.67 | 78.36 | 78.01 |
| WS-CKIP | 75.29 | 77.26 | 76.27 |
| WS-Finetuned | 79.89 | 82.46 | 81.15 |
| WS-Original | **87.28** | **86.84** | **87.28** |

表 8.斷詞結果比較實驗結果

### 4.5 最終結果

比賽最終的驗證資料集是由中央大學電機系的自然語言實驗室（NCUEE NLP Lab）所收集的醫療資料集(Lee et al., 2022)，但裡面並未提供斷詞資訊，且只能上傳三份預測結果，因此我們最終選擇了 Baseline 模型、WS-finetuned 模型以及 Baseline with 5 fold cross-validation 模型參加比賽(Lee et al., 2022) （如表 9），以 Baseline with 5 fold cross-validation 為最佳。

| | P | R | F1 |
|---|---|---|---|
| Baseline | 78.55 | 79.46 | 79.00 |
| WS-Finetuned | 77.62 | 77.46 | 77.54 |
| Baseline with 5 fold cross-validation | **81.99** | **81.88** | **81.93** |

表 9.三個模型最終比賽結果

## 5 結論

本研究使用中文醫療 NER 資料集，探討從全形半形格式到文本的斷詞、斷句對於 W2NER 模型的影響。我們發現此次任務適合大的 Encoder 模型，其結果普遍比較小的模型要來得好，而資料集的全形與半形對模型的結果也是有影響的，斷句內容雖然提高了 NER 標記的 Precision，但是缺少上下文訊息使得 Recall 大幅下降。雖然斷詞與否在本次的比賽結果與訓練結果相反，但也表示斷詞的結果需要一定的準確度才能使 NER 結果有顯著增加，準確度不夠反而會降低模型的表現，我們可以從使用原始斷詞的模型看到有正確斷詞資料的結果有顯著的提升，但要做到更準確的斷詞是值得探討的難題。

若未來能增加更多資料集像是前面提到的中國知識圖譜與語意計算大會（CCKS）的電子病歷資料集並轉換成繁體，以及使用更多的資料去訓練一個專門處理醫學用的斷詞工具去配合這個 NER 模型，我們認為這樣會有更好的結果，也是未來發展的方向。

## 致謝

本研究承國科會研究計畫 110-2221-E-004-008-MY3 與國立政治大學高教深耕校內補助計畫 111H124D-13 之部分補助，謹此致謝。

## References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wanget al.Guoping Hu. (2020). Revisiting Pre-Trained Models for Chinese Natural Language Processing.*Findings of the Association for Computational Linguistics: EMNLP 2020* Online.

Yiming Cui, Ziqing Yang, and Ting Liu. (2022). PERT: Pre-training BERT with Permuted Language Model. *arXiv preprint arXiv:2203.06906*.

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*.

Han He, and Jinho D. Choi. (2021). The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. *arXiv preprint arXiv:2109.06939*.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. (2022). *Overview of the ROCLING 2022 shared task for Chinese healthcare named entity recognition.* In Proceedings of the 34th Conference on Computational Linguistics and Speech Processing., Taipei.

Lung-Hao Lee, and Yi Lu. (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.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhanget al.Fei Li. (2022). Unified named entity recognition as word-word relation classification. Proceedings of the AAAI Conference on Artificial Intelligence,

Peng-Hsuan Li, Tsu-Jui Fu, and Wei-Yun Ma. (2020). Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER. Proceedings of the AAAI Conference on Artificial Intelligence,

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshiet al.Veselin Stoyanov. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wei-Yun Ma, and Keh-Jiann Chen. (2003). Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. Proceedings of the second SIGHAN workshop on Chinese language processing,

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethardet al.David Mcclosky. (2014). The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations,

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

J Sun. (2020). 'Jieba'(Chinese for'to stutter') Chinese text segmentation: built to be the best Python Chinese word segmentation module.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Joneset al.Illia Polosukhin. (2017). Attention is all you need. Advances in neural information processing systems, 30.

晏阳天, 赵新宇, and 吴贤. (2020). 基于 BERT 与字形字音特征的医疗命名实体识别. Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing,