基於語言模型與詞典方法的三種命名實體辨識模型架構之比較

NERVE at ROCLING 2022 Shared Task: A Comparison of Three Named Entity Recognition Frameworks Based on Language Model and Lexicon Approach

林柏劭 Bo-Shau Lin
國立高雄科技大學
資訊工程系
Department of Computer Science and Information Engineering National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C

陳建和 Jian-He Chen
國立高雄科技大學
資訊工程系
Department of Computer Science and Information Engineering National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C

張道行 Tao-Hsing Chang
國立高雄科技大學
資訊工程系
Department of Computer Science and Information Engineering National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C

{c108151121, c107151129, changhth}@nkust.edu.tw

摘要

此次任務的目的是設計一個方法標記在句子中的醫療實體詞以及它們的類別。本研究提出三種模型。第一種是以 BERT 模型結合線性分類器；第二種是一個兩階段模型，兩階段都是 BERT 模型結合分類器的次模型，但一階段只判斷句子中是否有醫療實體詞、二階段才專注於實體類別分類。第三種是結合前兩種模型以及一個基於詞典的模型，整合三個模型的結果後預測。實驗顯示這些模型在驗證與測試集的表現差異不大，最佳的模型 Run 1 在 F1 的值為 0.7569。

Abstract

ROCLING 2022 shared task is to design a method that can tag medical entities in sentences and then classify them into categories through an algorithm. This paper proposes three models to deal with NER issues. The first is a BERT model combined with a classifier. The second is a two-stage model, where the first stage is to use a BERT model combined with a classifier for detecting whether medical entities exist in a sentence, and the second stage focuses on classifying the entities into categories. The third approach is to combine the first two models and a model based on the lexicon approach, integrating the outputs of the three models and making predictions. The prediction results of the three models for the validation and testing datasets show little difference in the performance of the three models, with the best performance on the F1 indicator being 0.7569 for the first model.

關鍵字: 中文命名實體辨別, BERT, 集成式學習
Keyword: Chinese NER, BERT, Ensemble Learning

1 緒論

命名實體(Named Entity, NE)是指一種真實存在的事物，例如人、地點、組織以及產品等等，通常以專有名稱命名，例如梅克爾、柏林、基督教民主聯盟等等。由於命名實體通常是文件中的重要訊息，因此如何辨識命名實體成為自然語言處理領域重要且持續研究的問題，也稱為命名實體辨識(Named Entity Recognition, NER)。在中文文本上這個問題又更加困難，因為中文句子中的詞彙間並無空白加以區隔，因此對中文句而言，不僅要判斷句子中是否有 NE、也要判斷 NE 在句中的起始位置與結束位置。此外，NE 的類別辨識也相當重要，因為在

如	何	治	療	胃	食	道	逆	流	症
O	O	O	O	B-DISE	I-DISE	I-DISE	I-DISE	I-DISE	I-DISE

圖 1. 句子被標記後的標籤樣式

實際應用中，不同類別的NE有著不同的性質、功能或用途，若能正確分類對於實際應用上有很大的幫助。而由於不同專業領域的 NE 特徵也有不同，因此為了提高 NER 的正確率，會針對特定領域探討 NER 如何解決。

ROCLING 2022 Shard Task(以下簡稱此次任務)由 Lee et al.(2022)提出，是一項針對中文醫療 NER 的任務，其難度除了包含中文可以同時以單字詞與多字詞表達語意的複雜性，還包含中文醫療命名實體的詞典資源稀少、而新產生的 NE 會不斷產生。因此，此次任務目標為：研究者須找出一個句子中是否有 NE，並分辨該 NE 屬於 10 種實體類別(如表 1 所列)中的何者。

此次任務具體要求如下。研究者需要設計一個模型，針對一個句中每一個字元給予標籤。該標籤由兩部分資訊組成：該字元是否是 NE 的一部分，以及若是 NE、其所屬類別。第一部分有三個標記：B 表示該字元為一個 NE 的起始字元、I 表示該字元為一個 NE 的非起始字元、O 表示該字元不是任一 NE 的一部份。當一個字元被標記為 B 或 I 時，在第二部分需標記其所屬的類別，標籤種類如表 1 中所列。上方圖 1 為一個句子被標記後標籤樣式的範例。

實體類別	標籤	範例
人體	BODY	脊髓
症狀	SYMP	咳嗽
醫療器材	INST	達文西手臂
檢驗	EXAM	腦電波圖
化學物質	CHEM	糖化血色素
疾病	DISE	帕金森氏症
藥品	DRUG	青黴素
營養品	SUPP	益生菌
治療	TREAT	胃切除術
時間	TIME	青春期

表 1. 此次任務要辨識的 10 個實體類別

由於「如何治療」不是 NE，所以四個字元都標記為「O」;「胃」是命名實體「胃食道逆流症」的第一個字，「食道逆流症」是非起始字，而胃食道逆流症是一種疾病，因此「胃」標記為「B-DISE」、其他字標記為「I-DISE」。綜上所述，此次任務是要將句子中的每個字元標記 21 個標籤之一。

此次任務訓練與驗證資料集是由 Chinese HealthNER(Lee and Lu, 2021)所提供給研究者作為建立模型之用，資料集中各類別數量與比例如表 2 所示。本文針對此次任務設計了三種方法，在以下小節說明本文所提方法。本文第二節會回顧 NER 任務相關的研究；第三節介紹本文提出的三種模型；第四節會介紹本文所使用的實驗資料集以及評估指標。最後我們會從實驗結果探討本文所提方法的特性與限制，並提出未來工作的可能方向。

實體類別	訓練集(比例)	驗證集(比例)
人體	23,240(38.01%)	3,171(43.41%)
症狀	11,423(18.67%)	1,481(20.27%)
醫療器材	1,047 (1.71%)	42 (0.58%)
檢驗	2,218 (3.63%)	404 (5.53%)
化學物質	6,090 (9.96%)	744 (10.18%)
疾病	9,074 (14.84%)	1,005(13.76%)
藥品	2,146 (3.51%)	79 (1.08%)
營養品	1,403 (2.29%)	122 (1.67%)
治療	2,905 (4.75%)	203 (2.78%)
時間	1,609 (2.63%)	54 (0.74%)

表 2. 此次任務訓練與驗證資料集各類別數量及比例

2 相關工作

近年來 NER 研究多聚焦在深度學習神經網路模型。Luo et al. (2018)指出，在化學領域的 NER 任務上，普遍都是以傳統的機器學習的方法來解決，但這些傳統方法的效能取決於特徵工程。該研究提出了基於注意力機制的 BiLSTM-

修	齊	指	甲	O	K	,	有	一	位	3	0	多	歲	男	性
O	O	B-BODY	I-BODY	O	O		O	O	O	O	O	O	O	O	O

圖 2. 特殊句標記範例

CRF 模型，透過注意力機制去學習一個標籤在不同前後文中都能被標記為同一個標籤。該研究指出此設計在 CHEMDNER 以及 CDR 的兩個資料集中，其 F1 分別達到 91.14 以及 92.57。

Liu et al. (2021)認為先前 NER 模型在使用由巨量的訓練語料庫預訓練後的模型直接進行 NER 通常表現不佳，可能原因為數據集中包含專業領域的資料量佔比極低，所以建立的語意空間與專業領域的語意空間有所差別，導致模型的性能受限。為此 Liu et al. 進行了 BERT 的訓練語料庫改良，此研究創建了一個質量較高的大量 NER 語料庫，用此語料庫進行 BERT 的訓練，得到的預訓練模型稱之為 NER-BERT。此研究指出此模型的效能比原來的 BERT 在 NER 的表現更佳。

Lu & Lee (2020)提出一個門控圖序列神經網路的模型架構，用於中文健康照護領域命名實體辨識。該架構整合了輸入句的詞嵌入資訊與字典中已知的詞彙，之後輸入給一個 BiLSTM-CRF 模型對句子進行序列標註。該研究以網路爬蟲的方式擷取資料後以人工標記，再以人工標記結果測試該模型。實驗結果顯示該模型架構能比單純的 BiLSTM-CRF 或是 ME-CNER 模型的表現來得更好。

也有研究專注在辨識效率的問題。Gui et al. (2019)則指出，雖然 BiLSTM 在 NER 任務上有相當好的效果，但運算相當耗時。該研究利用基於 CNN 的模型架構，此方法的特點在於可以平行處理構建句中每個字的語意向量與找出句中的 NE，此研究指出此架構的效率為原本 3.21 倍。

3 實驗方法

本文對此次任務主要是以 Bidirectional Encoder Representations from Transformers (BERT)為基礎設計模型。BERT (Devlin et al., 2019)是知名的語意向量模型，其主要運用注意力機制(self-attention)為基礎，可以輸出句子的句意向量以及句中每個字的語意向量。BERT 主要運作原理是透過注意力機制(self-attention)使得模型可以因為單字的前後文來產生語意向量，這意味著儘管是同樣的一個字，也會因為出現的位置和前後文不同，因而產生出不同的輸出。

本文設計了三個基於 BERT 的模型，分別稱為 Run1、Run2 與 Run3。這三個模型雖然都使用 BERT，但在策略與架構上有所不同，本文將在下列各小節分別介紹。近幾年，有越來越多的語意向量模型被提出，如 RoBERTa (Liu et al., 2019)、ELECTRA (Clark et al., 2020)等等。由於此次任務只能送出三個結果，而本文希望比較不同的策略和架構的效果，因此只使用 BERT 為產生語意向量的核心。

3.1 Run 1 模型

Run1 是以非常直觀的方法來使用 BERT 解決 NER 任務，也就是用 BERT 輸出每個字的語意向量直接進行分類。此方法的想法是經過微調訓練後的 BERT，同一標籤的不同字其語意向量應該彼此接近、不同標籤的字其語意向量應距離較遠。只要在使用一個分類器就能學習將彼此相近的字輸出同一個類別，進而完成任務。此方法實際設計是在 BERT 輸出一個字的 768 維度向量後輸入給一層線性層(linear layer)，線性層輸出一個 21 維度的向量，每個維度代表一個標籤，維度值代表對應標籤的機率值，機率值最高者為此方法標定該字元的標籤。此模型有採用 dropout 與 fine-tune 策略優化模型。

經過上述程序，句子中的每一個字都會得到一個分類。但由於 BERT 會將數字、英文單字及多字符符號(例如刪節號)合併視為一個字，與此次任務的標記規則不同，因此需要進行前處理。舉例來說，此次任務資料集中句子「修齊指甲 OK」與「有一位 30 多歲男性」應該被標記為如圖 2 所示。

由於 BERT 會將「OK」和「30」視為一個字，不符合此次任務的輸出規格，因此本文會先將 10 個阿拉伯數字替換成在訓練資料中未曾出現過的 10 個特殊符號、例如@、#等等。另外，本文將會發生上述問題之英文字母與多字符符號的 57 句從資料集中移除，移除後訓練資料集中有 28,106 個句子，驗證資料集中有

2,529 個句子。

此外，訓練資料集有資料不均衡的問題，其中數量最少的標籤「B-INST」僅有 1,040 個，數量最多的標籤「O」則有多達 1,229,263 個，是前者的 1,000 多倍。這樣的情況可能會導致模型傾向將所有的字元都預測為標籤「O」就能有不錯的效能。本文因此設計了 loss 函數如下：

$$loss = \sum_{i=1}^{n} w_i \times y_i \times \log \hat{y}_i \qquad (1)$$

其中n為標籤類別數，y為正確答案經過 one-hot encoding 後的結果，$\hat{y}$為模型輸出經過 softmax 轉為機率的結果，w為每項特徵的權重。此公式改良自 cross entropy 函數，差別在w參數。而類別i的w計算公式如下：

$$w_i = {a_o}/{a_i} \qquad (2)$$

其中a_i表示類別i在訓練集中出現的次數；a_O表示類別 O 在訓練集中出現的次數。此公式會使得出現次數越少的類別得到愈大的權重，使得模型會重視資料量少的類別的損失。

一個句子經本文所提方法標記後，可能會出現標記結果不合理的情形。例如連續的標籤「O」之後出現一個標籤「I-xxxx」(xxxx 表示 10 種實體類別之一)，由於任何一個 NE 的第一個字一定是「B-xxxx」而非「I-xxxx」，出現上述情況是明顯地不合理。本文所提方法會對模型輸出結果進行後處理程序，以修正輸出結果邏輯上不合理之處。後處理程序由以下兩條規則所組成：

(1) 若句中單獨出現標籤「I-xxxx」，且其前後字元為標籤「O」，則將其替換為標籤「O」。

(2) 若是連續出現標籤「I-xxxx」，但這些標籤之前沒有標籤「B-xxxx」，則將第一個標籤「I-xxxx」替換為標籤「B-xxxx」。

3.2 Run 2 模型

Run 2 模型是基於以下構想：BERT 分類器模型一開始不需要將字元直接分類成 21 種不同的標籤，而是僅需要分辨與醫療 NE 無關或有關的字。若判斷一串連續字元與醫療 NE 有關，則再經由一個模型判斷這個字串屬於哪個分類。圖 3 為此構想所設計出的 Run 2 模型架構。

圖 3 中第一階段模型與 Run 1 模型相似，差別在於線性層的輸出為 2 維度，分別代表該字元與 NE 有關或無關。若無關，則該字元被標記標籤「O」；若有關，則進入第二階段模型。在第一階段中被視為有關的連續字元會被視作一個句子，輸入給第二階段模型。第二階段模型也與 Run 1 模型相似，差別在於 BERT 輸出給線性層的向量不是單一字元的語意向量，而是整個連續字串的句向量。線性層將句意向量分類為 10 種實體類別。最後，此連續字串的第一個字元被標記為「B-xxxx」，其餘標記為「I-xxxx」。此方法不會發生 Run 1 模型標記不合邏輯的情況，因此不需要對輸出加上後處理。

此模型所需的訓練與驗證資料需要進行前處理。對於第一階段模型，其訓練集資料中標籤「O」以外的各種類別標記重新標記為類別 1，而標籤「O」則標記為類別 0。對於第二階段模型，其訓練集是從原訓練集中抽出非標籤「O」之詞彙，並照原標籤「B-xxxx」或「I-xxxx」都替換成標籤「xxxx」。因此第二階段的訓練集由 61,155 個詞組成，而驗證資料集由 7,305 個詞組成。由於這兩階段模型的訓練集都沒有太嚴重的資料不均衡問題，因此在損失函數上只使用一般的 Cross Entropy 函數運算。

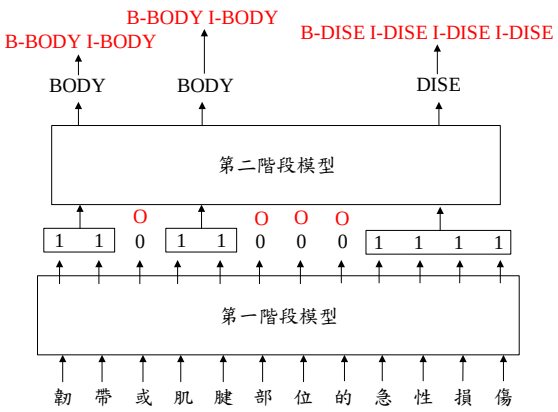

圖 3. Run 2 模型架構

3.3 Run 3 模型

Run 3 模型則是嘗試一種集成式 (ensemble) 架構。此架構由三個次模型組成，包括前兩小節提到的 Run 1 與 Run 2 兩個模型，以及一個詞典模型。詞典模型的核心是一個由訓練集中抽取曾出現過的 NE 所形成的詞典。例如在第一節提到的例子「如何治療胃食道逆流症」，由於訓練集中「胃食道逆流症」已被標記為 DISE，因此會被收錄至詞典並被記錄為 DISE。本文

所提方法的詞典在收錄詞時會排除同時屬於多個類別的 NE。此外，由於單字詞 NE 很容易造成誤判，因此也被排除在詞典收錄之外。得到詞典後，本文所提方法會利用詞典對驗證集進行初步標記。標記方式為一個句子中如果有出現在詞典中的詞，則將該詞標記為該詞在詞典中紀錄的類別。由於只用此方式標記結果不完全準確，因此我們會蒐集驗證集中每個類別的預測精確率。以 DISE 類別為例，若是預測驗證集中有 1,000 個字元被預測為 DISE，其中人工標記 DISE 有 850 個、CHEM 有 100 個和 SYMP 有 50 個，本文所提方法就建立 DISE 的機率分布為 DISE 是 0.85、CHEM 是 0.10 和 SYMP 是 0.05。對於沒有出現的類別，會給予一個極小值避免機率為 0 的情形發生。

圖 4 中的詞典模型對於每一個字元都會視其所屬類別輸出一個 11 維的向量，這個向量就是前述的機率分布。例如當一個句子輸入詞典模型後，會先檢視句中是否有存在於詞典的 NE 並標記每個詞的類別。對於每個字元就輸出每個字元所屬類別在各類別的機率分布，也就是一個 11 維的向量。

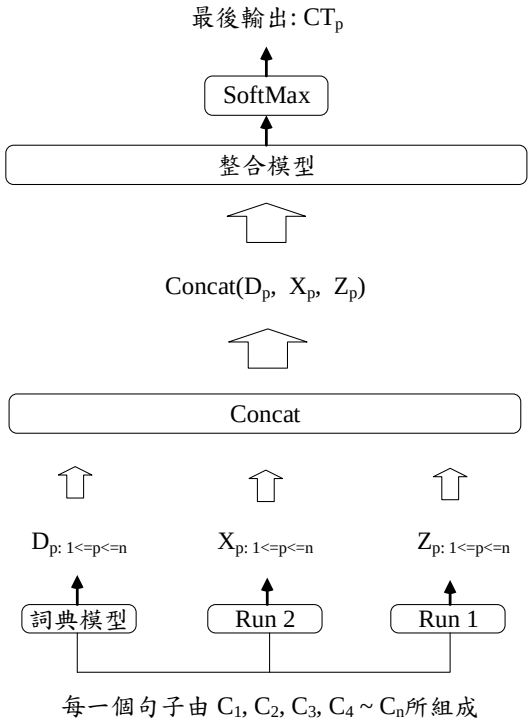

圖 4. Run 3 模型架構

由於 Run 3 模型包含三個次模型，需要一個整合模型將三個次模型的輸出加以整合，輸出最後的預測結果。圖 4 為 Run 3 模型的架構圖。Run 3 模型輸入一個句子時，詞典模型和 Run 2 會分別對每個字輸出一個 11 維向量(也就是圖 4 的 D_p 與 X_p，p 為 1 到 n 的值)、Run 1 模型會對每個字輸出一個 21 維向量(也就是圖 4 的 Zp)。這三個向量會被串接成一個 43 維的向量輸入給整合模型。整合模型是一個三層全連接模型，各層神經元分別為 43、30 以及 21。輸出層的 21 個神經元分別輸出該字元屬於神經元對應類別的機率值。

最後經由 softmax 程序判定該字元類別為機率值最大的類別(也就是圖 4 的 CT_p)。此整合模型的參數設計都比照 Run 1 的分類器。

4 實驗

除了此次任務提供之資料集，以及由 Huggingface (Wolf et al., 2019)提供的已預訓練 BERT 模型外，本文各項模型沒有使用其他的外部資料。此次任務是以精確率(Precision)、召回率(Recall)以及 F1-score 作為評估指標，4.1 節將說明評估指標的算法。本文所提模型的評估結果於 4.2 節討論。

4.1 評估指標

此次任務會針對 21 個類別的每個類別分別計算其混淆矩陣的四個值。以表 3 的 I-BODY 標籤的預測結果為例，四個值為預測為 I-BODY 且真實值為 I-BODY 的真陽性(true positive, TP)、預測為 I-BODY 但實際值為其餘類別的偽陽性(false positive, FP) 、預測為其餘特徵但實際值為 I-BODY 的偽陰性(false negative, FN)以及預測為其餘特徵且實際值為其餘特徵的真陰性(true negative, TN)，如表 3 所示。

預測 / 人工	I-BODY	Others
I-BODY	TP	FN
Others	FP	TN

表 3. 單一類別之混淆矩陣示例

透過混淆矩陣得到的四個值，即可計算每個類別之精確率(Precision)以及召回率(Recall)，計算公式如下：

模型	NERVE Run 1		NERVE Run 2		NERVE Run 3	
	驗證集	測試集	驗證集	測試集	驗證集	測試集
Precision	0.7873	0.7959	0.6800	0.7165	0.7871	0.7573
Recall	0.6812	0.7309	0.7353	0.7895	0.7294	0.7358
F1-score	0.7304	0.7620	0.7056	0.7512	0.7571	0.7464

表 4. 本文所提模型之效能

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1-score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (5)$$

在此次任務中，對於每一個模型，都會先計算該模型預測的每一個類別的精確率、召回率以及 F1，再將 21 個類別精確率加總平均後求得模型的預測精確率。模型的召回率以及 F1-score 亦復如是。

4.2 實驗結果

表 4 為 Run 1、Run 2 以及 Run 3 三個模型分別對此次任務所提供的驗證與測試資料集的預測結果。從表 4 可以發現，Run 1 和 Run 2 分別有較佳在精確率和召回率，而 Run 3 整合模型表現較預期為差。Run 3 對驗證集的結果是三種模型中最好的，不過對測試集的評估結果卻是最差的。我們猜測是由於加入了詞典模型導致的，因為訓練集和驗證集中有很多重複的 NE，因此詞典模型能夠正確指出詞的類別，使得評估數據較高；但測試資料中的 NE 卻是沒出現過的，使得詞典沒有收錄，就無法發揮其效能。

我們觀察資料發現 Run 1 與 Run 2 兩個模型對連續字元形成的較長字串會有不同的處理結果。Run 1 傾向將字串分割成多個不同類別的 NE，而 Run 2 則是傾向將字串視為單一類別的 NE。以圖 5 的句子為例，對於「中耳積水」這個詞，Run 1 會判斷「中耳」為 BODY 與「積水」為 SYMP，Run 2 則會判斷整個詞為 DISE。這是因為 Run 2 是將可能的 NE 交由第二階段判斷所屬類別，因此會將字串全部歸類於單一類別。

5 未來工作

在此次任務中，本文設計了三種不同的架構。其中最直觀的 Run 1 模型在整體表現是最好的，而另外兩個模型雖然有較細緻的設計，但測試資料集與驗證資料集的差異使得兩個模型雖然在部分指標或驗證資料集有較佳表現，但對測試資料集的整體表現不如直接而簡單的 Run 1 模型。由於另外兩個模型的設計理論上應該至少不遜於第一個模型，因此如何提高這兩個模型的強健性會是後續重要的研究方向。

此外，Run 3 模型表現較預期差的可能原因之一是詞典來源過於依賴訓練集。由於此次任務本文並未使用外部資源，因此未來可考慮使用外部資源、例如 wikipedia 或醫學書籍等，讓詞典模型能發揮應有功能。最後，已經有許多研究提出不同的語言模型，也有研究提出對特定領域修正後的預訓練語言模型，這些語言模型是否能提高本文所提方法的效能也值得嘗試。

誌謝

本研究由國家科學與技術委員會計畫編號 MOST 110-2511-H-992-003-MY3 提供部分補助，特此致謝。

文字	中	耳	積	水	的	定	義
NERVE Run1	B-BODY	I-BODY	B-SYMP	I-SYMP	O	O	O
NERVE Run2	B-DISE	I-DISE	I-DISE	I-DISE	O	O	O

圖 5. Run 1 以及 Run 2 輸出結果差異之說明範例

參考文獻

Lee, L. H., and 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.

Lee, L. L., Chen, C. Y., Yu, L. C., and Tseng, Y. H., 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.

Liu, Z., Jiang, F., Hu, Y., Shi, C., & Fung, P. 2021. *NER-BERT: a pre-trained model for low-resource entity tagging.* arXiv preprint arXiv:2112.00405.

Lu, Yi., & Lee, L. L. 2020. *Chinese Healthcare Named Entity Recognition Based on Graph Neural Networks*. Computational Linguistics and Chinese Language Processing Vol. 25, No.2, December 2020, pp. 21-36

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. 2018. *An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics*, *34*(8), 1381-1388.

Gui, T., Ma, R., Zhang, Q., Zhao, L., Jiang, Y. G., & Huang, X. 2019, August. *CNN-Based Chinese NER with Lexicon Rethinking*. In *IJCAI* (pp. 4982-4988).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

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*. arXiv preprint arXiv:1907.11692.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. 2020. *Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shlerifer, S.,von Platen, P., Ma, C., Jernite, Y., Plu, Julien., Xu, Canwen., Scao, L. T., Gugger, S., Drame, M., Lhoest, Q., Rush, M. A., Hugging Face & Brew, J. 2019. *HuggingFace's Transformers: State-of-the-art Natural Language Processing.* ArXiv, arXiv-1910.