# A Novel Named Entity Recognition Model Applied to Specialized Sequence Labeling (創新命名實體識別模型應用於專業化序列標記)

**Ruei-Cyuan Su, Tzu-En Su, Ming-Hsiang Su**
Department of Data Science, Soochow University, Taipei, Taiwan
{70613rex, 70614roy, huntfox.su}@gmail.com

**Matus Pleva, and Daniel Hladek**
Technical University of Kosice, Slovakia
{Matus.Pleva, daniel.hladek}@tuke.sk

## 摘要

近來，序列分割和標記的需求已經劃分了不同的專業領域。在傳統解決方案中，最常用的模型是結合了深度學習和監督學習的群體長短期記憶條件隨機場（Bi-LSTM-CRF），由於無監督學習的重要性已與監督學習並駕齊驅，本研究提出了長短期記憶-無監督監督學習-一般條件隨機場（Bi-LSTM-USL-GRF）模型，將通用隨機條件場（GRF）與無監督監督學習（USL）和 Bi-LSTM 相結合，實現了我們監督學習、無監督學習和深度學習的概念性結合。在本研究中，提供了一種創新的 GRF 架構來取代傳統的 CRF 架構，以及將無監督學習與有監督學習相結合的 USL 原理。我們證明，該模型不僅展示了利用 USL 原理的專業能力，還具有 GRF 的特殊優勢，其性能優於之前的 Bi-LSTM-CRF 架構提高了 1.45%。所提出的 USL 和 GRF 的組合具有更大的靈活性，未來甚至可以在不同的領域得到應用和推廣。

## Abstract

The demand for sequence segmentation and tagging has recently extended to different professional fields. The most commonly used model in conventional solutions is Bidirectional Long Short-Term Memory-Conditional Random Fields (Bi-LSTM-CRF), which combines deep learning and supervised learning. As the importance of unsupervised learning has become equal to that of supervised learning, this study proposes a Bidirectional Long Short-term Memory-Unsupervised Supervised Learning-General Conditional Random Field (Bi-LSTM-USL-GRF) model that combines General Conditional Random Field (GRF) with Unsupervised Supervised Learning (USL) and Bi-LSTM, achieving a conceptual combination of supervised learning, unsupervised learning, and deep learning. In this study, we provide an innovative GRF architecture to replace the traditional CRF architecture, as well as the USL principle, which combines unsupervised learning with supervised learning. We demonstrate that this model not only demonstrates specialized ability in the use of the USL principle but also has the special advantages of GRF, outperforming the previous Bi-LSTM-CRF architecture with a performance improvement of 1.45%. The proposed USL and GRF has more flexibility in its combination and could even be used and promoted in different fields.

關鍵字：深度學習、通用條件隨機場、無監督監督學習，長短期記憶

Keywords: Deep Learning, General Conditional Random Field, Unsupervised Supervised Learning, Bidirectional Long Short-Term Memory

## 1 Introduction

近來，對序列進行分割和標記的需求，不僅出現在 NLP 中部分語意標記(Part-of-speech tagging, POS tagging) (Biemann, 2009) 與命名實體識別(Named Entity recognition, NER) (Li et al., 2020)的兩個經典任務上，也從傳統的情感分析 (Mykhalchuk et al., 2021)，延伸至食品安全領域 (Yuan et al., 2023)、農業文本 (Qian et al., 2023)、機械專利提取技術 (Cui et al., 2023)、以及地震應急中文信息智能識別 (Wang et al., 2023)。使用的方法，也從過往的隱馬爾可夫模型(Hidden Markov model, HMM)與最大的馬爾可夫模型（Maximum-entropy Markov model,

MEMMs）(McCallum et al., 2000)在標記上的應用，像是 HMM 於基因標記 (Lukashin and Borodovsky, 1998)與 MEMMs 於情感內容檢測 (Kang, 2003)，轉變為經常使用的方法，為條件隨機場(Conditional random field, CRF) (Lafferty et al., 2001)，其應用像是在分詞器 (Tseng, 2005)，以及自動韻律預測和檢測的使用 (Qian et al., 2010)。甚至於近期，對於序列進行分割和標記，大部分所使用的是深度學習(Deep learning, DL)與機器學習(Machine learning, ML)的結合，並且以此結合的方向擴展使用，也就是將深度學習結合 CRF 做出改良，從連結 CNN(Convolutional neural network, CNN) (Kamnitsas et al., 2017)與 RNN(Recurrent neural network, RNN) (Wang et al., 2019)，到於現在最常使用的 Bidirectional long-short term memory CRF(Bi-LSTM-CRF) (Thattinaphanich and Prom-on, 2019)，使得 Bi-LSTM-CRF 不僅獲得了機器學習中的監督學習能夠輕易評斷之優點，也獲得深度學習容能夠忍受雜訊高的數據的能力。並且以 Bi-LSTM-CRF 作為基礎，來搭配於深度學習中合適的詞嵌入模型，可以生成最好的標籤序列，像是使用 BERT 來結合對命名實體識別進行標記 (Liu et al., 2023)。

因此在本文的研究中，提出了兩種方法，第一種為不同於傳統的結合方法，也就是它能在兩個監督學習結合中間新增一個隱藏層，實現其在監督學習的概念中有著非監督學習的概念，為非監督化的監督學習(USL)原理，第二種方法，不是傳統 CRF 的前方只能連接一個，而是前面能夠連接兩個的廣條件隨機場(GRF)，也就是其不僅能與 Bi-LSTM 連接，而且也能與從第一種方法中 USL 原理得到的非監督學習連接，且最後得到 Bi-LSTM-USL-GRF 網路。

## 2 General Conditional Random Field

接著，介紹廣條件隨機場原理。設$X$與$Y$與$Z$是隨機變數，三者構成一個無向圖$G=(V,E)$表示的馬爾可夫隨機場。接著，以隨機變數$Z$作為隨機變數$X$與隨機變數$Y$的中心，並且在滿足局部馬爾可夫隨機場下，可得到隨機變數$X$下隨機變數$Z$的條件概率$P(Z|X)$，且以相同的方式也可得到隨機變數$Y$下隨機變數$Z$的條件概率$P(Z|Y)$，接著並且依據局部馬爾可夫性，其與隨機變數$X$與隨機變數$Y$的聯合概率分布$P(Y,X)$可以由隨機變數$Y$下隨機變數$Z$的條件概率$P(Z|Y)$與隨機變數$X$下隨機變數$Z$的條件概率$P(Z|X)$相乘得到，為

$$P(X,Y)=P(X)P(Y) \tag{1}$$

有了隨機變數$X$與隨機變數$Y$的聯合概率分布$P(Y,X)$下的隨機變數$Z$的條件概率分布$P(Z|X,Y)$，若隨機變數$Z$自身再構成另一個無向圖$G=(V,E)$表示的馬爾可夫隨機場，且已知基本條件隨機場的定義(Lafferty et al., 2001 )，可推理得

$$P(Z_v|X,Y,Z_w,w\neq v)=P(Z_v|X,Y,Z_w,w\sim v) \tag{2}$$

與原定義之式等價。對任意結點$v$成立，稱條件概率分佈$P(Z|X,Y)$為條件隨機場，其中$w\sim v$表示圖$G=(V,E)$中與結點$v$有邊連接的所有結點$w$，$w\neq v$表示頂點$v$以外的所有結點，$Y_v$與$Y_W$為結點$v$與$w$對應的隨機變數。此時，假設上述的$X$和$Y$和$Z$，三者是相同的結構，為 $X=(X_1,X_2,\dots X_n), Y=(Y_1,Y_2,\dots Y_n), Z=(Z_1,Z_2,\dots Z_n)$，皆是線性鏈表示的隨機變數序列，此時的隨機變數序列$X$和$Y$兩者的聯合概率分布條件下，隨機變數序列$Z$的條件機率分布$P(Z|X,Y)$，能構成條件隨機場，且滿足馬爾可夫性，為

$$P(Z_i|X,Y,Z_1,\dots,Z_{i-1},Z_{i+1},\dots,Z_n)=P(Z_i|X,Y,Z_{i-1},Z_{i+1}) \tag{3}$$

可稱此條件機率分布$P(Z|X,Y)$是線性條件隨機場，其中$i=1,2,\dots,n$，在$i=1$和$n$時只考慮單邊。若線性鏈表示的隨機變數序列$Z$的無向圖$G=(V,E)$，形狀為樹狀，且與傳統 CRF 不同，它在每節點$E_i(=Z_i)\in E$上，有了兩個分支，分別接著$X_i$與$Y_i$，將此線性鏈表示的隨機變數序列$Z$無向圖。有了其架構後，引入 Hammersley-Clifford 定理 (Hammersley and Clifford, 1971) ，以及增加可變動之參數序列$W$後，可求得在隨機變數序列$X$和$Y$兩者的條件下，隨機變數序列$Z$構成的無向圖之聯合概率分布 $P(Z|X,Y)$為

$$P(Z|X,Y)=\frac{1}{H(X,Y)}\prod_{C\epsilon C_G}\psi_C\big((Z_C|X,Y),W\big),$$
$$H(X,Y)=\sum_{Z}\prod_{C\epsilon C_G}\psi_C((Z_C|X,Y),W) \tag{4}$$

$$= \frac{1}{H(X,Y)} \prod_{C \epsilon C_G} \mathrm{F}_C\left((Z_C|X,Y), W\right),$$
$$\mathrm{F}_C((Z_C|X,Y), W) > 0 \quad (5)$$

其中，$H$是規範化因子，$C_G$為於$G$上全部的最大團集合，$Z_C$是於最大團中之一$C$所對應的隨機變數，$\psi_C$函數稱為勢函數，然而，$\psi_C((Z_C|X,Y), w)$輸出必須為正，因此將$\psi_C$勢函數的符號換成$\mathrm{F}_C$。然而，勢函數於通常情況之下，為指數函數，因此這裡把勢函數正式定義為指數線性的勢函數，則公式可轉換成

$$\mathrm{F}_C\left((Z_C|X,Y), W\right) = \exp[W^T \varphi(\mathrm{X},\mathrm{Y},Z_C)] > 0 \quad (6)$$

其中，$\varphi(\mathrm{X},\mathrm{Y},Z_C)$的定義是由全局輸入$X$以及全局輸入$Y$和局部標籤$Z_C$，所能產生特殊向量結果的一個函數$\varphi$。為了更加明確說明此指數線性的勢函數，我們可知道此線性條件隨機場裡，其中一個完整極大團集合$\mathcal{D}_{(Z_1,Z_2)}$裡，是包含結點、邊的，有$\{Z_{i-1},Z_i\}_2^T$、$\{Z_i,X\}_1^T$、$\{Z_i,Y\}_1^T$、$\{Z_i,X\}_2^T$和$\{Z_i,Y\}_2^T$，其中，$\mathcal{D}_{(Z_1,Z_2)}$表示有包含$z_1$以及$z_2$兩結點的元素的子集們的集合，如此下來，當前集合$\mathcal{D}_{(Z_1,Z_2)}$應與下一個集合$\mathcal{D}_{(Z_2,Z_3)}$會有所交集，但考慮為最大團會有重複計算之問題，因此將$\mathcal{D}_{(Z_i,Z_{i+1})}$集合裡重複的最大團$\{Z_{i+1},X\}_2^T$和$\{Z_{i+1},Y\}_2^T$從集合$\mathcal{D}_{(Z_i,Z_{i+1})}$裡移除，減少重覆計算問題。因此，一個完整極大團集合$\mathcal{D}_{(Z_1,Z_2)}$裡，是有對邊的狀態$\{Z_{i-1},Z_i\}_2^T$、以及兩個對點的狀態$\{Z_i,X\}_1^T, \{Z_i,Y\}_1^T$，將這三狀態分別以函數$\varphi$的定義得出三個不同輸入值的函數，為

$$t(Z_{i-1},Z_i,X,Y) \quad (7.1)$$
$$s(Z_i,X) \quad (7.2)$$
$$q(Z_i,Y) \quad (7.3)$$

之後我們將對邊的狀態函數$t(Z_{i-1},Z_i,X,Y)$稱為移轉狀態函數，第一對點的狀態$s(Z_i,X)$稱為當前狀態函數，第二對點的狀態$q(Z_i,Y)$稱為確認狀態函數。接下來加入可變動之參數序列$W = (\lambda,\mu,\phi)$中的三種可變動參數$\lambda,\mu,\phi$後，融入相應函數可得

$$\exp[W^T \varphi(\mathrm{X},\mathrm{Y},Z_C)] = exp\left[[\lambda,\mu,\phi]\begin{bmatrix} t(Z_{i-1},Z_i,X,Y), \\ s(Z_i,X), \\ q(Z_i,Y) \end{bmatrix}^T\right] > 0 \quad (8.1)$$

$$:\Longrightarrow exp\begin{bmatrix} \lambda t(Z_{i-1},Z_i,X,Y), \\ \mu s(Z_i,X), \\ \phi q(Z_i,Y) \end{bmatrix} > 0 \quad (8.2)$$

最後得

$$P(Z|X,Y) = \frac{1}{H(X,Y)} \prod_{C \epsilon C_G} \exp\begin{bmatrix} \lambda t(Z_{i-1},Z_i,X,Y) \\ ,\mu s(Z_i,X) \\ ,\phi q(Z_i,Y) \end{bmatrix}$$
$$H(X,Y) = \sum_Z \prod_{C \epsilon C_G} \exp[\lambda t(Z_{i-1},Z_i,X,Y) ,\mu s(Z_i,X), \phi q(Z_i,Y)] \quad (9)$$

此為線性鏈表示的隨機變數序列$Z$的無向圖$G = (V,E)$，在隨機變數序列$X$和$Y$兩者的條件下，隨機變數序列$Z$構成的聯合概率分布$P(Z|X,Y)$之證明。

依據上述推導，可得出給出在隨機變數$X$與$Y$的條件下，包含可變動之參數序列$W$的隨機變數$Z$的聯合分布之定裡

$$P(Z|X,Y) = \frac{1}{H(X,Y)} \exp\left(\sum_{v \epsilon V,k} \lambda_k t_k(v,Z_v,X,Y) + \sum_{e \epsilon E,l} \mu_l s_l(e,Z_e,X) + \sum_{e \epsilon E,m} \phi_m q_m(e,Z_e,Y)\right)$$
$$H(X,Y) = \sum_Z \exp\left(\sum_{v \epsilon V,k} \lambda_k t_k(v,Z_v,X,Y) + \sum_{e \epsilon E,l} \mu_l s_l(e,Z_e,X) + \sum_{e \epsilon E,m} \phi_m q_m(e,Z_e,Y)\right) \quad (10)$$

其中，$Z_s$是指在$Z$節點集合中與$s$子圖關聯之子集，$t_i$則是於隨機變數序列$Z$裡，會產生的第$i$的移轉狀態函數，$k$是共有$k$的移轉狀態函數，其餘$s_i$和$q_i$的意義與$t_i$相同。指數函數裡，由一個於邊上之函數$t_k$，以及兩個於點上函數，一個對$X$節點的函數$s_k$，一個對$Y$節點的函數$q_k$，共三個部分組成。$t_k$和$s_k$以及$q_k$函數皆是可被固定的，例如，$t_k$輸出值判斷由在$Z$線性鏈裡，在$Z$上其中有關聯單邊兩節點$Z_i, Z_{i+1}$的標籤，如果與$t_k$要求之標籤相同，則輸出特定值，如果不同則輸出另一特定值。剩下的$\lambda_k$和$\mu_k$以及$\phi_k$則是可變動之參數$W = (\lambda,\mu,\phi)$裡三種可變動參數之內部相應權值，例如，

在$Z$線性鏈裡可表示為$\lambda = (\lambda_1, \lambda_2, \dots \lambda_k), \mu = (\mu_1, \mu_2, \dots \mu_l), \phi = (\phi_1, \phi_2, \dots \phi_m)$。為了使這些值的迭代修正連結至神經網路的權重，以極大似然估計法，極大化對數似然函數來求得其參數$W^*$。若已知數據集$D = \{(X_i, Y_i, Z_i)\}_{i=1}^n$，從之可得經驗概率分布$\tilde{p}(X, Y, Z)$，且$P$使用上式$P(Z|X, Y)$，可得對數似然函數為

$$L(w) = L_{\tilde{p}}(P) = log \prod_{X,Y,Z} P(Z|X,Y)^{\tilde{p}(X,Y,Z)} = \sum_{X,Y} \tilde{p}(X,Y,Z) log\, P(Z|X,Y) \quad (11)$$

$$W^* = \arg\max_W L(W) \quad (12)$$

因配合神經網路對於迭代值為遞減且誤差修正值為正值，則可修改為

$$-L(W) = -\sum_{X,Y} \tilde{p}(X,Y,Z) log\, P(Z|X,Y) \quad (13)$$

$$W^* = \arg\min_W -L(W) \quad (14)$$

得到損失函數之值$-L(W)$後，再從其求得梯度向量，選擇合適的深度學習之優化器，來反向傳播得到最優的權重$W^*$，取得更精確之精度和結果。

接下來剩餘的部分，將說明其如何矩陣化，減輕其計算複雜程度之後作為應用。首先，我們已知隨機變數序列$Z$構成的聯合概率分布$P(Z|X,Y)$之證明結果，接下來，將原本隨機變數序列$Z = (Z_1, Z_2, \dots Z_n)$加入各標籤，再把不同的狀態函數對應的所有函數全部$t = (t_1, t_2, \dots t_k), s = (s_1, s_2, \dots s_l), q = (q_1, q_2, \dots q_m)$寫入，以及把可變動之參數$W$裡三種可變動參數之內部與函數相應權值$\lambda = (\lambda_1, \lambda_2, \dots \lambda_k), \mu = (\mu_1, \mu_2, \dots \mu_l), \phi = (\phi_1, \phi_2, \dots \phi_m)$換入，再將已得出的定理代入後，可得

$$P(Z|X,Y) = \frac{1}{H(X,Y)} \exp\left( \sum_{i,k} \lambda_k t_k(Z_{i-1}, Z_i, X, Y) + \sum_{i,l} \mu_l s_l(Z_i, X) + \sum_{i,m} \phi_m q_m(Z_i, Y) \right)$$

$$H(X,Y) = \sum_Z \exp\left( \sum_{i,k} \lambda_k t_k(Z_{i-1}, Z_i, X, Y) + \sum_{i,l} \mu_l s_l(Z_i, X) + \sum_{i,m} \phi_m q_m(Z_i, Y) \right) \quad (15)$$

其中可以看出，對於同一狀態函數上，在各個位置上皆有定義，且求和於同一狀態函數上，因此將局部狀態函數轉為全局的狀態函數，可恢復成推導前可變動之參數序列$W$與狀態函數的內積。所以先將移轉狀態函數$t(Z_{i-1}, Z_i, X, Y)$與當前狀態函數$s(Z_i, X)$以及確認狀態函數$q(Z_i, Y)$，三者合併，把$k$換成$K_1$，$l$換成$K_2$，$m$換成$K_3$，$K = K_1 + K_2 + K_3$，以$f_k$符號作為全部狀態函數統一表示，得

$$f_k(z_{i-1}, z_i, X, Y) = \begin{cases} t_k(Z_{i-1}, Z_i, X, Y), & k = 1,2, \dots K_1 \\ s_l(Z_i, X), & k = K_1 + l;\ l = 1,2, \dots K_2 \\ q_m(Z_i, Y), & k = K_1 + l + m;\ m = 1,2, \dots K_3 \end{cases} \quad (16)$$

其中，$k$換表示為全部狀態函數的總數。接著，對於移轉狀態函數在各個位置$i$求和，可記成

$$f_k(z_{i-1}, z_i, X, Y) = \sum_{i=1}^n f_k(Z_{i-1}, Z_i, x, i), \quad k = 1,2, \dots, K \quad (17)$$

注意與原本的$W$意義不同，這裡用$W_k$表示為$f_k(z_{i-1}, z_i, X, Y)$的權重，同樣$k$表示為全部權重的總數，為)

$$W_k = \begin{cases} \lambda_k, & k = 1,2, \dots K_1 \\ \mu_l, & k = K_1 + l;\ l = 1,2, \dots K_2 \\ \phi_m, & k = K_1 + l + m;\ m = 1,2, \dots K_3 \end{cases} \quad (18)$$

於是可得內積形式為

$$P(Z|X,Y) = \frac{1}{H(X,Y)} exp \sum_{k=1}^K W_k f_k(z_{i-1}, z_i, X, Y)$$

$$H(X,Y) = \sum_Z exp \sum_{k=1}^K W_k f_k(z_{i-1}, z_i, X, Y) \quad (19)$$

得到內積形式後，簡化成矩陣的變換方式就與 CRF 相同，不同之處為新增了隨機變數序列$Y$。首先，將指數函數裡的加總變換成指數函數外的乘積

$$P(Z|X,Y) = \frac{1}{H(X,Y)}\prod_{i=1}^{n+1}\exp\left(\sum_{k=1}^{K}W_k f_k(Z_{i-1},Z_i,X,Y)\right) \quad (20)$$

已知移轉狀態函數為隨機變數序列$Z$中取$z_{i-1}$和,$z_i$兩個的值，因此假設在$r$種標籤中的取值下，可得到一個$r$階隨機變數矩陣$[M_i(Z_{i-1},Z_i|X,Y)]_{r\times r}$，或稱為標籤矩陣，且再把$W_k$權重以及指數函數皆融入在此矩陣裡後，可得出隨機變數序列$Z$構成的聯合概率分布$P(Z|X,Y)$為

$$P(Z|X,Y) = \frac{1}{H(X,Y)}\prod_{i=1}^{n+1}M_i(Z_{i-1},Z_i|X,Y) \quad (21)$$

其中，$M_i(Z_{i-1},Z_i|X,Y)$的乘積數共$n+1$是因為在標籤序列$Z$前後各新增$Z_0 = 1 = start$以及$Z_{n+1} = 1 = stop$，或稱開始狀態以及結束狀態。規範化因子$H(X,Y)$，是通過狀態的所有路徑$Z_1, Z_2, \dots Z_n$，一一對應非規範化概率$\prod_{i=1}^{n+1}M_i(Z_{i-1},Z_i|X,Y)$的全部加總，所以會考慮到所有路徑，隨機變數矩陣$M_i(Z_{i-1},Z_i|X,Y)$內也不需考慮到移轉狀態$Z_{i-1},Z_i$，可改寫成$M_i(X,Y)$。因此規範化因子$H(X,Y)$，是隨機變數矩陣序列$Z$的隨機變數矩陣$M_1(X,Y)$乘積至$M_{n+1}(X,Y)$，為

$$H(X,Y) = [M_1(X,Y)M_2(X,Y)\cdots M_{n+1}(X,Y)]_{start,stop} \quad (22)$$

最終得到聯合概率分布$P(Z|X,Y)$的矩陣形式

$$P(Z|X,Y) = \frac{\prod_{i=1}^{n+1}M_i(Z_{i-1},Z_i|X,Y)}{\prod_{i=1}^{n+1}M_i(X,Y)_{start,stop}} \quad (23)$$

## 3 Unsupervised Supervised Learning

監督學習，是適用指從標記數據中學習預測模型的學習問題，其本質是學習輸入到輸出的映射的統計規律，而無監督學習，則適用無標記數據中學習預測之模型的學習問題，為學習數據中的統計規律或潛在結構。

無監督化的監督學習(Unsupervised supervised learning, USL)，是將兩個監督學習模型結合於一體，並且將兩個監督學習模型之中的連結處，也就是第一監督學習模型的輸出變數和第二監督學習模型的輸入變數，兩個變數改由一個變數連結兩者，此時，此變數又稱隱藏結構，且此隱藏結構的空間定義是由給定一函數所輸出之變數空間所決定。有了此隱藏結構後，就能以各監督學習模型連結此隱藏結構進而學習，實現無監督學習之原理。因為第一監督學習模型的輸入是源自於數據集，且輸出是映射到未知變數的隱藏結構空間，因此內容上是學習此隱藏結構的空間之潛在結構，接著是與其原理相近但結構不相同的第二監督學習模型，輸出是源自於數據集，輸入是以隱藏結構的空間之未知變數，因此其目的也是學習隱藏結構的空間之潛在結構。由此可知，兩監督學習模型之目的是學習數據中的潛在結構，為無監督學習的方法核心。所以整體而言，是使用了監督學習的手段，實現無監督學習的學習原理的模型，因此又稱之為無監督監督學習模型(Unsupervised Supervised Learning Model, USLM)。

無監督化監督學習模型結構中，由兩個監督學習模型與一個函數和一個作為隱藏結構的隱藏層所構成，此時的隱藏層是與神經網路之定義不一樣，這裡的隱藏層，輸入是連接著第一監督學習的模型，輸出則是連接著第二監督學習的模型，並且在輸入以及輸出之間連接一個函數於此隱藏層。當學習模型時，由函數輸出變數作為隱藏層的變數，與數據集一起提供第一、二監督學習模型的學習使用，且在預測時，函數不輸出變數，則以第一監督學習模的輸出變數作為隱藏層的變數，接著為第二監督學習模型預測所使用，此時隱藏層的變數等同在隱藏結構空間上的變數。其中，裡面所連接之函數有類似於閥門的作用，因為它可以控制於不同時間點對於輸出變數的調控。

以數學方法介紹。首先，對整體的無監督化監督學習的輸入與輸出，以隨機變數$X$與隨機變數$Y$表示，兩者定義在特徵空間與輸出空間上，可以是同類型或不同類型，其中對於隨機變數$X$與$Y$的取值，定義為變數$x$與$y$。接下來看到中間部分，隱藏層則以隨機變數$Z$空間表示，且在隨機變數$Z$取值下可得到變數$z$作為隱藏層之變數。於此同時，隱藏層的空間是定義在函數之輸出的空間上，因此將此連接於隱藏層的外接函數用$g$表示，所以可得$Z = g(R)$，且$R$為任意隨機變數，取值下可得

到變數$r$，是定義在函數$g$的輸入空間上的任意隨機變數。

無監督化監督學習是從訓練集學習模型。無監督化監督學習之訓練集，主要由兩個部分組成，一部分是包含輸入變與輸出變數之訓練集，也就是隨機變數$X$與$Y$取值下作為樣本的訓練集$T$，為

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (24)$$

另一部分是函數之輸入變數，也就是在輸入空間上的任意隨機變數$R$取值下，即將輸入函數$g$的變數$\{r_1, r_2, \dots, r_N\}$，兩者來共同學習模型。其中，樣本的訓練集$T$的$(x_i, y_i)$表示為第$i$的輸入與輸出成對組成的樣本，且$x_N, y_N, r_N$皆表示對其變數的總數為$N$。有了樣本的訓練集$T$與函數之輸入變數後，即可準備模型的學習。

當開始學習模型時，隨機變數$R$的變數$\{r_1, r_2, \dots, r_N\}$會輸入至函數$g$，轉換為隱藏層空間下的隨機變數$Z$的變數$\{z_1, z_2, \dots, z_N\}$，於此同時，隨機變數$Z$之變數也是無標記數據，因此，又稱此隱藏層為在隱藏結構空間下隨機變數$Z$的無標記數據集，為

$$T^U = \{z_1, z_2, \dots, z_N\} \quad (25)$$

其中，$U$代表是無標記的數據集，同樣$z_N$也表示變數$z$的總數為$N$。有了訓練集$T$與無標記數據集$T^U$之後，即可對第一監督學習模型跟第二監督學習模型同時學習。其中，第一監督學習模型的學習，使用的是訓練集$T$的輸入變數與數據集$T^U$，也就是隨機變數$X$的變數$\{x_1, x_2, \dots, x_N\}$與隨機變數$Z$的變數$\{z_1, z_2, \dots, z_N\}$，對第二監督學習模型的學習，則換成訓練集$T$的輸出變數與數據集$T^U$，也就是隨機變數$Y$的變數$\{y_1, y_2, \dots, y_N\}$與隨機變數$Z$的變數$\{z_1, z_2, \dots, z_N\}$。接著，在這同時學習階段裡，要從一、二監督學習模型結合的所有可能的集合中，得到最優的模型，其不僅使兩個監督學習的特點各自發揮最大效果，也實現了無監督學習的學習原理。因此，採取的損失計算方法或學習算法，是採用了各自監督學習的特點，分別選擇出能作為對其揮最大效果的損失計算方法或學習算法，且於此同時，對其所選擇出的學習算法之停止條件，即是收斂時機。例如:支持向量機(SVM) (Platt,1998)使用合頁損失函數(hinge loss function) (Hearst et al., 1998)作為學習的損失函數，且以序列最小優化算法(Sequential minimal optimization, SMO)則作為學習算法。另一個是條件隨機場，是以對數似然損失函數作為其學習的損失函數，且改進迭代尺度算法(Improved iterative scaling, IIS) (Chinneck, 1994)則作為學習算法。本實驗採取的是以對此種的評估的方式進行，於之後實驗介紹中會詳細說明。

接著，當完成學習階段，也就是達到收斂停止學習時，不論採用何種的評估方式，皆可得到兩監督學習合為一體的無監督化監督學習模型 USLM，且在之後預測中，將作為預測所使用到的模型。其中，第一監督學習模型的是概率或非概率模型，可簡單表示為條件概率分布$\dot{P}(Z|X)$或決策函數$Z = \dot{f}(X)$，且第二監督學習模型也是概率或非概率模型，也可簡單表示為條件概率分布$\ddot{P}(Y|Z)$或決策函數$Y = \ddot{f}(Z)$，此時兩者的條件概率分布以及決策函數，主要描述輸入與輸出隨機變數之間的映射關係。

最終，在預測過程時，對於給定預測樣本之隨機變數$X$取值下的$x_{N+1}$，可經由第一監督學習模型之條件概率分布$z_{N+1} = \arg\max_Z \dot{P}(z|x_{N+1})$或決策函數$z_{N+1} = \dot{f}(x_{N+1})$，取得到在隱藏層下之隨機變數$Z$取值下的最大機率之變數$z_{N+1}$，且再經由其中之第二監督學習模型的條件概率分布$y_{N+1} = \arg\max_y \dot{P}(y|z_{N+1})$或決策函數$y_{N+1} = \dot{f}(z_{N+1})$，最後取得到相應輸出隨機變數$Y$取值下最大機率的$y_{N+1}$，完整使用無監督化監督學習模型 USLM 來輸出樣本之預測結果。其中，$x_{N+1}$和$y_{N+1}$變數表示為不屬於訓練集$T$的$N$樣本組內，可以任意從中能得到的$x_i$和$y_i$，$z_{N+1}$變數則表示為第一監督學習模型以$x_{N+1}$作為輸入變數，輸出在隱藏層下的結果。

## 4 Bi-LSTM-USL-GRF Model

在 Bi-LSTM-USL-GRF 網路上，使用了一個 Bi-LSTM 網路和一個 GRF 網路，以及在 USL 的概念下 GRF 的第一監督學習網路結合成的一個三合一之模型，也稱此模型為 USL+BI-LSTM+GRF 模型。其中的一邊，GRF 連結的第一監督學習網路，因為使用了 USL 的概念進而使 GRF 獲得輸入特徵與標籤序列之間的隱藏層來連接，剩下的一邊，GRF 連結的 Bi-LSTM 網路，再進而獲得完整的句

子之輸入特徵，並且因為在訓練時，隱藏層會產生外接函數，因此以$g$表示。接著，因為 Bi-LSTM 網路與第一監督學習網路，以 GRF 為中心構成一個無向圖，因此兩者的條件概率下，對 GRF 產生的機率是相等的，並且得到一個 Y 字型作為連結的模型。因此在 GRF 監督學習網路於輸入上，不僅得到了整體句子的特徵，也獲得了句子特徵和標籤之間的轉換結構，或稱隱藏結構，且於最後標籤輸出上，使用 GRF 的轉移狀態矩陣，可以藉著過去以及未來的標籤預測當前標籤，更加能獲得準確率更高的標籤序列。我們將第一監督學習網路的隱藏層之輸出表示為$\dot{f}([x]^T)$，Bi-LSTM 網路的輸出表示為$f([y]^T)$，以及新增標籤序列表示為$[i]^T$。其中，$[x]^T$為第一監督學習網路的輸入變數，$[y]^T$為 Bi-LSTM 網路的輸入變數，$T$表示為其長度為$T$。接著，對於兩函數輸出後的變數，分別乘上權重$\delta$與$\theta$，以及給予第$i$的標籤種類於第$t$的長度之所有長度為$T$之標籤路徑，因此可將其兩輸出後的變數轉換為$[\dot{f}_\delta([x]^T)]_{i,t}$以及$[f_\theta([y]^T)]_{i,t}$的兩個矩陣。接著，將標籤種類結合了隨機變數矩陣成為標籤矩陣，且其為從第$i$的標籤種類到第$j$的標籤種類之矩陣，因此可以表示成$[M]_{i,j}$，而且不受長度位置變換影響。最後，將上述三個矩陣參數結合成一個新的參數矩陣，此新參數可表達為$\tilde{w} = \delta \cup \theta \cup \{[M]_{i,j} \forall i,j\}$，且再對其長度$T$加總，可以得到作為 GRF 的確認狀態函數之矩陣$[\dot{f}_\delta([x]^T)]_{i,t}$，以及作為 GRF 的當前狀態函數之矩陣$[f_\theta([y]^T)]_{i,t}$，和最後作為 GRF 的移轉狀態函數之矩陣$[M]_{i,j}$的總和，為

$$s([x]^T, [y]^T, [i]^T, \tilde{w}) = \sum_{t=1}^{T} \begin{pmatrix} [M]_{[i]_{t-1},[i]_t} \\ +[\dot{f}_\delta([x]^T)]_{[i]_{t-1},t} \\ +[f_\theta([y]^T)]_{[i]_{t-1},t} \end{pmatrix} \quad (26)$$

其中，$[i]_t$表示為第$i$的標籤種類於第$t$的長度之標記，且對整體 GRF 而言，改良了複雜的計算。接著使用再動態規劃中最有效的維特比算法計算移轉狀態函數之矩陣$[M]_{i,j}$，以求得和最佳的標籤序列。

## 5 Training procedure

我們對於 Bi-LSTM-USL-GRF 模型的完整的訓練算法寫在 Algorithm 1 上，且在我們之後所有的實驗中，所使用到的模型皆是以隨機梯度下降法 SGD 的方式來進行更新模型的參數。其中在每一次的 epoch，將訓練集分成許多的 batch，且每 batch 執行一次模型的過程，而其中的 batch 之數目表示為總共完整句子的個數。在執行一次模型的過程裡，會先分別進行 Bi-LSTM-USL-GRF 的 forward pass 與 backward pass。在 Bi-LSTM-USL-GRF 的 forward pass 中，需要產生兩個輸出作為之後的 GRF 層之輸入，分別是經由 Bi-LSTM 層的輸出$f_\theta([y]^T)$，以及在 USL 的概念下的第一監督學習層與 GRF 層之間的隱藏層$\dot{f}_\delta([x]^T)$。接著，因為在 USL 的概念訓練時，是以外接函數$g$之輸出當作隱藏層，同時訓練與第一、二監督學習模型，因此在訓練第二監督學習模型之前，也就是訓練 GRF 層之前，還要先訓練第一監督學習層。所以，第一監督學習層與外接函數$g$之輸出依照給定 iterations 的次數訓練，更新其參數。之後，將兩個輸出送入 GRF 層進行 forward pass，可得到 GRF 層的輸出以及其損失函數之梯度向量，並且將其損失函數之梯度向量進行 GRF 層的 backward pass 以及 Bi-LSTM-USL-GRF 模型的 backward pass，也就是從輸出至輸入反方向來更新權重，其中包含了標籤矩陣$[M]_{i,j} \forall i,j$和 Bi-LSTM 的參數$\theta$之更新。但其中不包含 GRF 層的更新，因為在先前的 Bi-LSTM-USL-GRF 的 forward pass 裡，早已訓練以及更新。最後，在本文的實驗中，使用了 batch 之數目為 32，且將其輸入句子之長度限定在 100 字作為上限，以進行 Bi-LSTM-USL-GRF 的模型完整之訓練。

**Algorithm 1.** Bi-LSTM- USL-GRF model training procedure

```
for each epoch do
    for each batch do
        1. Bi-LSTM-USL-GRF model forward pass:
            (1) forward pass for Bi-LSTM layer
            (2) USL-Supervised learning layer:
                for each iteration do
                    update USL-Supervised learning layer-parameters
                end for
        2. GRF layer forward and backward pass
        3. USL-Bi-LSTM-GRF model backward
```

pass:
(1) backward pass for Bi-LSTM layer
(2) update Bi-LSTM-parameters
(3) fix USL-Supervised learning layer-parameters
**end for**
**end for**

## 6 Experiment

我們所使用資料集有兩個中文資料集，第一個資料集為 Chinese Healthcare Named Entity Recognition (HealthNER)，是由 NCUEE NLP 研究室人員收集與標記 (Lee et al., 2021)，我們將此資料集做為訓練集以及驗證集使用。第二個資料集是 2022 年台灣計算語言學與語音處理年會(Association for Computational Linguistics and Chinese Language Processing(ROCLING))所提供的資料集 (Lee et al., 2022)，且我們將此資料集做為測試集。訓練集以及驗證集分別有 28,161 個句子以及 2531 個句子，測試集則有 3205 個句子，加上選擇 BIO 標註方式的原因，是因為 BIO 為使用原始標註（Raw labeling）解決聯合標註（Joint segmentation and labeling）之問題的最佳方法，最適合接下來的 NER 任務。其中，關於兩者資料集中使用的 BIO 標註方式，為將單一標籤對應單一中文字進行標記，以醫療相關之中文字進行分類，且對於無醫療相關的中文字則全部歸為同一類別，而對應特殊分類的標籤將作為單詞的語法作用，因此提取一個句子中的一個類別，可能會有兩個中文字，分別對應兩個特殊分類的標籤。

接著，我們的實驗是選擇 Bi-LSTM-USL-GRF 模型與其他不同的網路模型，於 NLP 問題中的序列標註任務上，來做測試與評估。其中，我們的實驗使用的資料集有兩個中文資料集，皆使用 BIO 的方式標註，因此序列會有大量的 O 作為標籤出現。接著，依照 USL 的原理，還需要再選擇一個監督學習作為連接，所以我們選擇將標籤中的 O 與非 O 的兩種標籤作為二分類標準，並且以此來學習其二分類標準的潛在結構，最終選定了最適合二分類的 SVM 以及 Adaboost 兩個監督學習作，作為與 GRF 連接。並且於此同時，在訓練時使用的外接函數$g$之輸出，也以兩不同的值作為此二分類標準出現，因此可以得到 Bi-LSTM-SVM-GRF 和 Bi-LSTM-Adaboost-GRF 兩個網路模型。接著，以原本的 Bi-LSTM 替換成 LSTM，可以得到 LSTM- SVM -GRF 和 LSTM- Adaboost -GRF 兩個網路模型，因為想與原本的 CRF 比較，可以再得到 BiLSTM-CRF 和 LSTM-CRF 兩個網路模型，因此總共有六種網路模型，於 NLP 問題中的序列標註任務上測試。

## 7 Word embedding and Bi-LSTM-USL-GRF Model

在這六種網路模型的測試任務中，我們將這些資料集中的由許多單字構成的句子，全部使用最簡單 Word2Vec (Mikolov et al., 2013) 的方式，把每一個的單字轉成向量，連接後並且以整個句子視作為整體，使用 Word2Vec 的原因為此次實驗為比較傳統的 CRF 和 Bi-LSTM-USL-GRF 的不同，不是比較詞嵌入（Word embedding）的不同。其中的 Bi-LSTM-USL-GRF，是將句子轉換為鏈狀的 Bi-LSTM-USL-GRF 網路之輸入的特徵，且每個 Bi-LSTM-USL-GRF 網路所對應的節點所需要的輸入特徵，分別對應每個單字所屬的向量，也就是輸入按照句子對應的單字順序，並且在輸出時，將每個單字轉化為 BIO 標註方式的標籤。相較於 CRF 結合 Bi-LSTM 之模型 (Huang et al., 2015) 的改進之處，不僅僅新增了隱藏層作為標籤的潛在結構，來得到標籤的潛在規律，也把 CRF 換成了 GRF，使得最後的 GRF 能夠同時得到 Bi-LSTM 的句子特徵以及標籤的潛在規律的特徵作為輸入，來輸出得到最合適的標籤，提升句子對應的 BIO 標註之正確性。

## 8 Parameter Setting

我們使用 Word2Vec 的方式得到訓練集的文字之向量特徵，並且以隨機梯度下降法 SGD 的方式來訓練這七種網路模型的參數。其中，學習率設定為 0.1，而且為了在參數更新的時候於一定程度上保留之前更新的方向，以及防止有過擬合的情形發生，因此新增了 Momentum (Yuan et al., 2018)和權重衰減 (Weight decay) (Zhang et al., 2018)兩個方法，Momentum 方法於設定上保留了 90%的前一刻參數數據來減少震盪，從而加快收斂速度，權重衰減 (Weight decay)方法於設定上加入 0.0001 之值來抑制更新參數的幅度。最後在每

兩 epoch 上所計算出的總更新數值上，再乘以 0.9 來減少最後更新的幅度，更能找到最好的模型之參數。

## 9 Results and Discussions

我們使用訓練集訓練完了這六種網路模型後，再使用驗證集以及測試集來預測標籤，並且以 F1 值(F1 Score)、精確率(Precision)、召回率(Recall)以及準確率(Accuracy)，四項指標來對模型的預測的標籤和正確的標籤做出評估，Table 1 表示出了整體結果。從中可以看出不論在測試集或驗證集下，Bi-LSTM-SVM-GRF 在 F1 Score、Recall 以及 Accuracy 皆達到這六種網路模型的最高值，其中選擇驗證集來觀看可分別得到 F1 Score 為 63.00%、 Recall 為 60.31%以及 Accuracy 為 92.74%。另外，Bi-LSTM-Adaboost -GRF 則是不論在測試集或驗證集下，在 Precision 也達到這六種網路模型的最高值，選擇驗證集來觀看可得到 Precision 為 66.75%。加上，Bi-LSTM-SVM-GRF 和 Bi-LSTM-Adaboost -GRF 為 USL 概念之下連接另一個監督學習的 Bi-LSTM-USL-GRF 網路，而且 Bi-LSTM-USL-GRF 網路的預測結果之指標皆明顯優於傳統的 Bi-LSTM-CRF 的指標，甚至遠優於剩餘網路模型的指標。因此相較於傳統的 Bi-LSTM-CRF，Bi-LSTM-USL-GRF 網路的更有明顯的改進效果，也就是對於 CRF 只取 Bi-LSTM 的有關於句子之輸出特徵，相比之下，GRF 以 USL 的概念多連結另一個不同的監督學習作為隱藏層，更能達到明顯的作用。而且 GRF 不僅能得到句子之輸出特徵，也可以到在隱藏層輸出的有關於標籤的潛在規律。因此在 GRF 相較於 CRF 上，更能使用不同方式改進隱藏層，達成在不同領域對於所需要的獨特之標籤，能有著專業的效果。

Table 1: 分別在驗證集以及測試集之中，對各種模型於 F1 值(F1 Score)、精確率(Precision)、召回率(Recall)以及準確率(Accuracy)，四項指標關於標籤之預測性能之間的比較。

| | F1 | Recall | Precision | Accuracy | Accuracy (non-O) |
|---|---|---|---|---|---|
| **Model using validation set** | | | | | |
| Bi-LSTM-Adaboost-GRF | 62.22 | 58.28 | **66.75** | 92.72 | 64.28 |
| Bi-LSTM-CRF | 61.99 | 58.45 | 65.98 | 92.67 | 64.68 |
| Bi-LSTM-SVM-GRF | **63.00** | **60.31** | 65.94 | **92.74** | **66.36** |
| LSTM-Adaboost-GRF | 60.45 | 55.74 | 66.03 | 92.38 | 62.30 |
| LSTM-CRF | 60.41 | 56.11 | 65.42 | 92.38 | 62.16 |
| LSTM-SVM-GRF | 55.40 | 47.91 | 65.67 | 91.75 | 54.91 |
| **Model using testing set** | | | | | |
| Bi-LSTM-Adaboost-GRF | 65.36 | 60.76 | **70.71** | 86.60 | 64.39 |
| Bi-LSTM-CRF | 65.87 | 61.91 | 70.38 | 86.78 | 65.40 |
| Bi-LSTM-SVM-GRF | **66.23** | **62.54** | 70.39 | **87.10** | **66.85** |
| LSTM-Adaboost-GRF | 62.94 | 57.24 | 69.90 | 85.64 | 61.42 |
| LSTM-CRF | 62.96 | 57.65 | 69.35 | 85.59 | 61.30 |
| LSTM-SVM-GRF | 58.77 | 50.75 | 69.80 | 83.67 | 54.89 |

接著，為了更加比較傳統的 CRF 與 GRF 於訓練過程中的影響，因此選取於與 GRF 連接中指標值最優的 SVM，將 GRF 和 SVM 合為一體與 CRF，也就是 Bi-LSTM-SVM-GRF 以及 Bi-LSTM-CRF 兩者作為 CRF 與 GRF 的比較，其中的訓練過程以每 5 epoch 的倍數來求得各自的模型，再使用驗證集以及測試集來求得各自之指標，Fig. 1 表示了 Accuracy 對於每 5 epoch 所得的值，其中可以看到在 35 epoch 時，不論在測試集或驗證集下，CRF 與 GRF 的指標皆達到最高值，且 CRF 於驗證集所得的 Accuracy 為 92.74%，測試集的 Accuracy 為 87.10%，GRF 於驗證集所得的 Accuracy 為 92.67%，測試集的 Accuracy 為 86.78%，因此可以得知於 35 epoch 時 GRF 的 Accuracy 優於 CRF 的 Accuracy，而且一開始 GRF 的 Accuracy 上升速度快於 CRF，以及後面訓練過程的 epoch 數越多的時候，GRF 的 Accuracy 皆優於 CRF 的 Accuracy，相較之下，CRF 的平緩上升以及平緩下降，而且只於 35 epoch 達到最高值，也就是 GRF 不需要太多時間即可達到顯著效果，以及最後能減少過度擬合發生的狀況，都可以得知 GRF 的隱藏層可能發揮重要之影響，對於訓練過程而言，不但提升訓練速度，也能作為新的方法解決過度擬合狀況的發生。

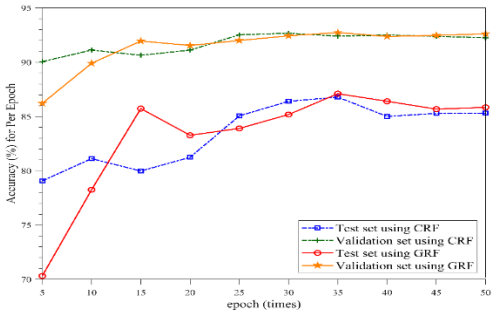


Fig 1: 在驗證集和測試集中 CRF 與 GRF 於不同 epoch 數下的 Accuracy。首先將 CRF 與 GRF 分別皆在驗證集以及測試集下做出測試，並且在不同的 epoch 數之下，求出其在 Accuracy 指標的連續變化情形，有使用驗證集下的 CRF(藍色)、使用測試集下的 CRF(綠色)、使用驗證集下的 GRF(紅色)以及使用測試集下的 GRF(橘色)。

## References

Biemann, C. 2009. Unsupervised part-of-speech tagging in the large. *Research on Language and Computation*, *7*(2-4), 101-135.

J. Li, A. Sun, J. Han and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, 2020. https://doi.org/10.1109/TKDE.2020.2981314.

Mykhalchuk, T., Zatonatska, T., Dluhopolskyi, O., Zhukovska, A., Dluhopolska, T., and Liakhovych, L. 2021. Development of recommendation system in e-commerce using emotional analysis and machine learning methods. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (Vol. 1, pp. 527-535). IEEE.

Yuan, T., Qin, X., and Wei, C. 2023. A Chinese Named Entity Recognition Method Based on ERNIE-BiLSTM-CRF for Food Safety Domain. *Applied Sciences*, *13*(5), 2849.

Qian, Y., Chen, X., Wang, Y., Zhao, J., Ouyang, D., Dong, S., and Huang, L. (2023, March). Agricultural text named entity recognition based on the BiLSTM-CRF model. In *Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022)* (Vol. 12566, pp. 525-530). SPIE.

Cui, R., Deng, N., and Zheng, C. (2023, February). Technology and Efficacy Extraction of Mechanical Patents Based on BiLSTM-CRF. In *International Conference on Emerging Internetworking, Data & Web Technologies* (pp. 223-234). Cham: Springer International Publishing.

Wang, Z., Huang, M., Li, C., Feng, J., Liu, S., and Yang, G. 2023. Intelligent Recognition of Key Earthquake Emergency Chinese Information Based on the Optimized BERT-BiLSTM-CRF Algorithm. *Applied Sciences*, *13*(5), 3024.

McCallum, A., Freitag, D., and Pereira, F. C. 2000,. Maximum entropy Markov models for information extraction and segmentation. In *Icml* (Vol. 17, No. 2000, pp. 591-598).

Lukashin, A. V., and Borodovsky, M. 1998. GeneMark. hmm: new solutions for gene finding. *Nucleic acids research*, *26*(4), 1107-1115.

Lafferty, J., McCallum, A., and Pereira, F. C. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Tseng, H., Chang, P. C., Andrew, G., Jurafsky, D., and Manning, C. D. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Qian, Y., Wu, Z., Ma, X., and Soong, F. 2010, November). Automatic prosody prediction and detection with Conditional Random Field (CRF) models. In *2010 7th International Symposium on Chinese Spoken Language Processing* (pp. 135-138). IEEE.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., ... and Glocker, B. 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis, 36, 61-78.

Wang, S., Yi, L., Chen, Q., Meng, Z., Dong, H., and He, Z. 2019. Edge-aware fully convolutional network with CRF-RNN layer for hippocampus segmentation. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* (pp. 803-806). IEEE.

Thattinaphanich, S., and Prom-on, S. 2019. Thai named entity recognition using Bi-LSTM-CRF with word and character representation. In *2019 4th International Conference on Information Technology (InCIT)* (pp. 149-154). IEEE.

Liu, Z., Wang, H., and Bol, P. K. 2023. Automatic biographical information extraction from local gazetteers with Bi-LSTM-CRF model and BERT. *International Journal of Digital Humanities*, *4*(1-3), 195-212.

Clifford, P., and Hammersley, J. M. 1971. Markov fields on finite graphs and lattices.

Platt, J. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. 1998. Support vector machines. *IEEE*

*Intelligent Systems and their applications*, *13*(4), 18-28.

Chinneck, J. W. 1994. MINOS (IIS): infeasibility analysis using MINOS. *Computers & operations research*, *21*(1), 1-9.

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. H., 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 (ROCLING 2022)* (pp. 363-368).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Huang, Z., Xu, W., and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Yuan, K., Chen, Y., Huang, X., Zhang, Y., Pan, P., Xu, Y., and Yin, W. 2021. DecentLaM: Decentralized momentum SGD for large-batch deep training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3029-3039).

Zhang, G., Wang, C., Xu, B., and Grosse, R. 2018. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*.