Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. 21-36

基於圖神經網路之中文健康照護命名實體辨識

Chinese Healthcare Named Entity Recognition Based on Graph Neural Networks

盧毅*、李龍豪*

Yi Lu and Lung-Hao Lee

摘要

命名實體辨識任務的目標是從非結構化的輸入文本中，抽取出關注的命名實體，例如：人名、地名、組織名、日期、時間等專有名詞，擷取的命名實體，可以做為關係擷取、事件偵測與追蹤、知識圖譜建置、問答系統等應用的基礎。機器學習的方法將其視為序列標註問題，透過大規模語料學習標註模型，對句子的各個字元位置進行標註。我們提出一個門控圖序列神經網路 (Gated Graph Sequence Neural Networks, GGSNN) 模型，用於中文健康照護領域命名實體辨識，我們整合詞嵌入以及部首嵌入的資訊，建構多重嵌入的字嵌入向量，藉由調適門控圖序列神經網路，融入已知字典中的命名實體資訊，然後銜接雙向長短期記憶類神經網路與條件隨機場域，對中文句子中的字元序列標註。我們透過網路爬蟲蒐集健康照護相關內容的網路文章以及醫療問答紀錄，然後隨機抽取中文句子做人工斷詞與命名實體標記，句子總數為 30,692 句 (約 150 萬字/91.7 萬詞)，共有 68,460 命名實體，包含 10 個命名實體種類：人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療與時間。藉由實驗結果與錯誤分析得知，我們提出的模型達到最好的 F1-score 75.69%，比相關研究模型 (BiLSTM-CRF, Lattice, Gazetteers 以及 ME-CNER)表現好，且為效能與效率兼具的中文健康照護命名實體辨識方法。

*國立中央大學電機工程研究所
Department of Electrical Engineering, National Central University
E-mail: ericst91159@gmail.com; lhlee@ee.ncu.edu.tw
The author for correspondence is Lung-Hao Lee

Abstract

Named Entity Recognition (NER) focuses on locating the mentions of name entities and classifying their types, usually referring to proper nouns such as persons, places, organizations, dates, and times. The NER results can be used as the basis for relationship extraction, event detection and tracking, knowledge graph building, and question answering system. NER studies usually regard this research topic as a sequence labeling problem and learns the labeling model through the large-scale corpus. We propose a GGSNN (Gated Graph Sequence Neural Networks) model for Chinese healthcare NER. We derive a character representation based on multiple embeddings in different granularities from the radical, character to word levels. An adapted gated graph sequence neural network is involved to incorporate named entity information in the dictionaries. A standard BiLSTM-CRF is then used to identify named entities and classify their types in the healthcare domain. We firstly crawled articles from websites that provide healthcare information, online health-related news and medical question/answer forums. We then randomly selected partial sentences to retain content diversity. It includes 30,692 sentences with a total of around 1.5 million characters or 91.7 thousand words. After manual annotation, we have 68,460 named entities across 10 entity types: body, symptom, instrument, examination, chemical, disease, drug, supplement, treatment, and time. Based on further experiments and error analysis, our proposed method achieved the best F1-score of 75.69% that outperforms previous models including the BiLSTM-CRF, Lattice, Gazetteers, and ME-CNER. In summary, our GGSNN model is an effective and efficient solution for the Chinese healthcare NER task.

關鍵詞：命名實體辨識、圖神經網路、資訊擷取、健康資訊學
Keywords: Named Entity Recognition, Graph Neural Networks, Information Extraction, Health Informatics

1. 緒論 (Introduction)

命名實體辨識 (Named Entity Recognition, NER) 主要目的為從非結構化的文本中，抽取出所關注的命名實體，主要包括人名、地名、組織名、時間、數量、貨幣、專有名詞等。舉例來說對於「比爾蓋茲創辦了微軟」這個中文句子，如果關注的命名實體包含人名以及組織名，透過 NER 即可抽取出人名「比爾蓋茲」以及組織名「微軟」。命名實體辨識為自然語言處理中的一項基礎任務，其後續的應用包含了關係抽取、事件抽取、知識圖譜以及問答系統等等，像是抽取出人名「比爾蓋茲」以及組織名「微軟」後，我們可以進一步擷取兩者之間關係為「創辦」。

早期的 NER 方法主要是基於字典或是規則，利用字串比對來做辨識，此種方法非常

依賴字典的可靠度以及專業人士所制定出的規則，因此需要耗費大量的人力資源。理論上，並不能夠蒐集到一個涵蓋所有命名實體的字典，或者制定可能的規則找到所有命名實體位置。因此，若是所依賴的字典品質不佳或是規則無法涵蓋所有的情況時，則命名實體辨識的表現會嚴重的下降。

而後，機器學習的方法透過大規模標註語料學習序列標註模型，對句子的各個位置進行標註，但同樣需要透過事先定義好的特徵，因此特徵選取的好壞，對於整個標註的結果有直接的影響，主要的模型有：隱藏式馬可夫模型(Hidden Markov Model, HMM) (Rabiner, 1989)、最大化熵馬可夫模型(Maximum Entropy Markov Model, MEMM) (Toutanova & Manning, 2000) 和條件隨機場域(Conditional Random Field, CRF) (Lafferty, McCallum & Pereira, 2001)。

隨著科技的進步，人類的壽命得以延長，有關健康照護的議題逐漸地浮上檯面，許多的報章雜誌都在談論相關議題，因此本研究所關注的命名實體領域選定為健康照護。有鑑於當前欠缺健康照護的中文命名實體辨識語料庫，我們從網路上蒐集了相關的文章雜誌以及問答紀錄，隨機選取 30,692 句，人工斷詞並標記時後，透過計算 Cohen's Kappa 值以及 Fleiss' Kappa 值確保標記的品質，最後總共有 68,460 個命名實體，橫跨 10 個類別，分別是人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療以及時間。

近年來深度學習技術的興起，神經網路在許多任務皆有著亮眼的表現，在命名實體辨識任務中，BiLSTM-CRF 網路架構為最被廣泛使用的主流模型(Lample, Ballesteros, Subramanian, Kawakami & Dyer, 2016; Ma & Hovy, 2016)。我們以此架構為基礎，並且考量中文的特性，斷詞的精準度會嚴重影響結果，因此以字作為輸入單位，訓練字嵌入、部首嵌入以及詞嵌入語意向量，透過門控圖序列神經網路(Gated Graph Sequence Neural Networks, GGSNN)加入字典資訊，在建置的中文健康照護命名實體辨識語料庫上，達到 F1 分數 75.69%，比當前具代表性相關研究模型(BiLSTM-CRF, Lattice, Gazetteers 以及 ME-CNER)有更好的成效。

本研究一共分為五個章節，第一章節為緒論，介紹命名實體辨識任務以及研究動機與目的。第二章節為探討相關研究，調查目前的中文命名實體辨識語料庫，並且介紹中文命名實體辨識模型。第三章節為模型架構，詳細介紹提出的圖神經網路模型，並對模型的各層做詳盡的說明。第四章節為實驗評估與分析，依序說明語料庫的建置、嵌入向量、實驗設定與效能評估指標，接著討論實驗結果和錯誤分析。第五章為結論和未來研究。

2. 相關研究 (Related Work)

2.1 中文命名實體辨識語料庫 (Chinese NER Corpora)

MSRA 命名實體語料庫(Levow, 2006)總共包含 30 種類別，語料來源為新聞文章，其中較被廣泛使用的類別像是人名(Person)、地名(Location 以及組織名(Organization)，此資料

集的訓練資料總共包含了 46,364 個句子，其中的命名實體總數為 118,643 個，測試資料為 4,365 個句子，其中標記的命名實體總數為 4,362 個。

在社群媒體方面，Weibo 命名實體語料庫(Peng & Dredze, 2015)蒐集了微博此社群媒體從 2013 年 11 月至 2014 年 12 月的訊息並對其標記，隨機挑選訊息的數量總共為 1,890 則，標記的命名實體類別總共有 4 種，分別為地理位置(Geo-political)、地名(Location)、組織名(Organization)以及人名(Person)，其中標記的命名實體總數為 1,981 個。

Resume 資料集(Zhang & Yang, 2018)的來源為個人履歷，履歷的出處為中國上市公司主管，總共隨機挑選了 1,027 份，標註的命名實體種類總共有 8 種，其中包含國家(Country)、人名(Person)以及組織名(Organization)等等，其中標記命名實體的的總數為 16,565 個。

中國知識圖譜與語義計算大會(CCKS: China Conference on Knowledge Graph and Semantic Computing)在 2019 年舉辦的評測任務中，命名實體辨識的資料來源為電子病歷(Electronic Health Record, EHR)，訓練集的文檔數為 1000 筆，而測試集的文檔數為 379 筆，所標註的命名實體包含 6 種，分別為疾病和診斷(Disease and Diagnosis)、檢查(Examination)以及檢驗(Inspection)等等，其中標記命名實體的的總數為 16,565 個。

上述的命名實體語料庫，並沒有關於健康照護領域方面的語料庫，且都為簡體中文，因此本研究建置了一個中文健康照護領域的命名實體語料庫，共有 10 類命名實體，分別為人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療以及時間。

2.2 中文命名實體辨識語模型 (Chinese NER Models)

Dong 等人(2016) 考量了中文字的特性，將中文字拆解成一個個部件，其原因為中文的字是由許多的部件組合而成，而每個部件具有其不同的意義，透過這些部件可以增加字特徵以外的特徵，從該研究可以得知「字」並非中文字具有意義的最小單位。

Xu 等人(2019)加入了除了字特徵以外的部首特徵以及詞特徵，並且將字特徵以及部首特徵利用雙向長短期記憶類神經網路(BiLSTM)以及卷積運算(Convolution)做額外的處理。在此研究中之所以加入中文部首的原因為中文部首具有語意分類，同樣部首的字，可能屬於同樣類別，因此透過部首可以對字做更進一步的分析。

Zhang 和 Yang (2018)提出了一個新的模架構 Lattice LSTM，此模型主要的特點為會將句子中詞彙透過大型自動取得的字典，將所有可能的潛在詞彙找出，利用此種方式可以將考量到可能潛在的詞邊界，此研究結果在命名實體辨識的任務中取得了重大的成果。

Ding 等人(2019)使用到了圖神經網路中的門控圖序列神經網路，並做改良使其能夠將多個字典的資訊加入模型，由於文字訊息常常會有著類似圖結構的訊息，因此透過圖神經網路能夠更充分的表達資訊。

基於上述研究，本研究提出了門控圖序列神經網路(Gated Graph Sequence Neural Networks, GGSNN)模型架構，以 BiLSTM-CRF 做為基礎，以字為單位當作模型的輸入。

除了字的資訊以外，本研究加入了部首以及詞的資訊。在加入字特徵、部首特徵時透過雙向長短期記憶類神經網路以及卷積運算做了有別於 Xu 等人(2019)的處理，使特徵資訊更能夠完整充分。在本研究的模型同樣使用了改良式的 GGSNN 將字典資訊加入，而與 Ding 等人(2019)不同的地方在於透過不同的字典編排方式，使其在相同的硬體設備下，字典的來完能夠更加的龐大且豐富。

3. 模型架構 (Model Architecture)

本研究提出的門控圖序列神經網路(GGSNN)模型架構如下圖 1，此模型使用了目前主流的 BiLSTM-CRF 作為模型的基礎架構，並對其做延伸，模型總共分為四層。

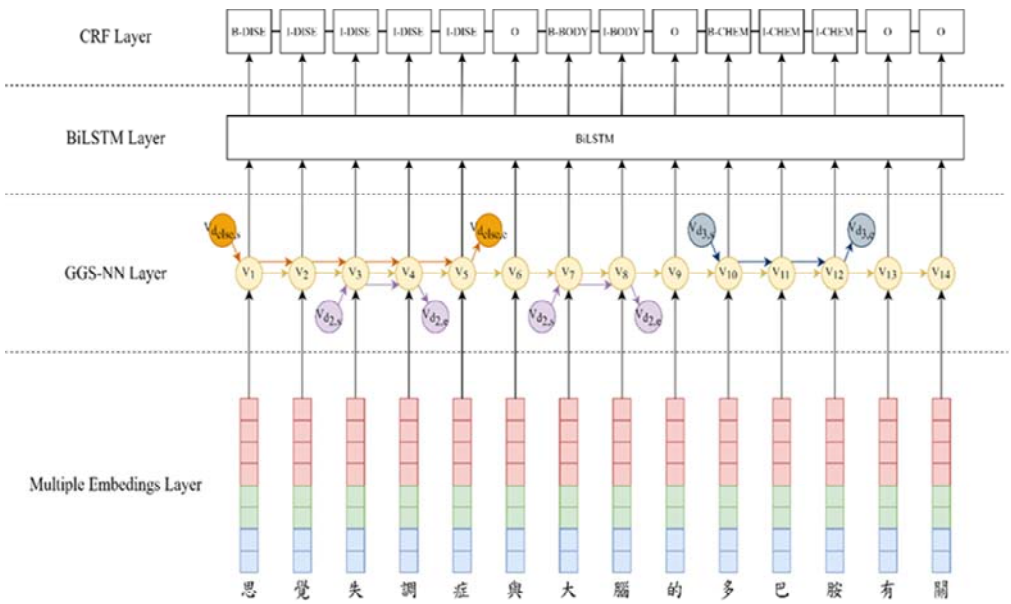

圖 1. GGSNN 模型架構
[Figure 1. GGSNN model architecture]

3.1 多重嵌入層 (Multiple Embeddings Layer)

透過組合字嵌入、詞嵌入以及部首嵌入形成多重嵌入，其中字嵌入、部首嵌入以及詞嵌入的處理分別如下列敘述，假設輸入句子字數長度為 n：

(1)、字嵌入 (Character Embedding)：

輸入字序列 $X = [x_1, x_2, x_3, \dots, x_n]$，分別經過 BiLSTM 以及卷積運算後，將兩者組成新的字嵌入特徵序列，得到序列 $C = [c_1, c_2, c_3, \dots, c_n]$，由於每個字可能與長距離的另一個字或是附近的字有所關聯，因此透過 BiLSTM 可以捕捉到長距離的資訊，而卷積運算可以捕捉到短距離的資訊。

$$[y_1, y_2, y_3, \dots, y_n] = BiLSTM(X) \quad (1)$$

$$[z_1, z_2, z_3, \dots, z_n] = Conv(X) \quad (2)$$

$$c_i = y_i \oplus z_i \quad (3)$$

(2)、部首嵌入 (Radical Embedding)：

輸入部首序列 $X = [x_1, x_2, x_3, \dots, x_n]$，經過卷積運算後，得到新的部首特徵序列 $[r_1, r_2, r_3, \dots, r_n]$，由於每個部首多半與附近的字有關，因此卷積運算可以捕捉到短距離的資訊。

$$[r_1, r_2, r_3, \dots, r_n] = Conv(X) \quad (4)$$

(3)、詞嵌入 (Word Embedding)：

由於模型是以字為基礎作為輸入，而同一個字組成的不同詞語可能有不同的意思，因此相同字的資訊，加入不同詞的資訊，可以解決此種情況，而詞的資訊是屬於較高階的特徵，因此本研究直接將其作合併，不做額外的處理。

$$W = [w_1, w_2, w_3, \dots, w_n] \quad (5)$$

最終將字特徵序列、部首特徵序列以及詞特徵序列，組合成多重嵌入如下：

$$h_i = c_i \oplus r_i \oplus w_i \quad (6)$$

其中c_i代表經過處理後的字嵌入，r_i代表經過處理後的部首嵌入，w_i代表詞嵌入，h_i代表拼接後的多重嵌入。

3.2 門控圖序列神經網路層 (GGSNN Layer)

在本研究中採用改良式 GGSNN 學習句子圖結構化後的訊息，與 Li 等人(2016) 所提出的 GGSNN 不同之處在於改良式的 GGSNN 可以給予邊上標籤不同的權重，透過改良式的 GGSNN 可以將加入多個字典的訊息，並且給予不同的字典不同的權重。但由於硬體的限制，我們無法不受限制的追加多個字典，因此與 Ding 等人(2019) 的字典編排方式不同，本研究將字典裡的詞彙依照字數做分類，總共分成五個字典。

在這層結構中首先會利用字典，透過字串比對產生多維有向圖，建構出的多維有向圖 (Multi-digraph) 範例如圖 2。給定一個多維有向圖 $G := (V, E, L)$，其中V代表節點的集合，E代表邊的集合，L代表邊上標籤的集合。假設輸入的句子為字數為 n 個，字典的使用數量為 m，節點的集合$V = V_c \cup V_s \cup V_e$。其中$V_c$為字序列節點的集合，而當字典比對到詞彙時，會產生除了字序列節的額外兩個節點，分別為$v_{d_i,s}$、$v_{d_i,e}$，其中$v_{d_i,s}$指示出詞彙的起始位置，$v_{d_i,e}$指示出詞彙的結束位置，V_s以及V_e分別代表的為$v_{d_i,s}$以及$v_{d_i,e}$的集合。邊的集合$E = \{e_c\} \cup \{e_{d_i}\}_{i=1}^{m}$，其中$\{e_c\}$為字序列節點連成的邊的集合，$\{e_{d_i}\}_{i=1}^{m}$為所有字典連成的邊的集合。每個邊都帶有標籤，邊上標籤的集合為$L = \{l_c\} \cup \{l_{d_i}\}_{i=1}^{m}$，$l_c$為字序列節點連成的邊上的標籤，$l_{d_i}$為字典連成的邊上的標籤，不同的字典帶有不同的標籤。

以「思覺失調症與大腦的多巴胺有關」當作輸入句子為例，可以得到如圖 2 的多維

有向圖，在此句子中，可以比對到的詞彙有「思覺失調症」、「失調」、「大腦」以及「多巴胺」，其中「思覺失調症」包含在詞彙字數為 5 個字以上 (else) 的字典中，因此「思覺失調症」的開頭「思」，對應到的節點v_1，連結了額外的節點$v_{d_{else},s}$，「思覺失調症」的結尾「症」，對應到的節點v_5，連結了額外的節點$v_{d_{else},e}$，$v_{d_{else},s}$的下標d_{else}以及下標 s 代表的為比對到的字典以及比對到的詞的開頭位置，$v_{d_{else},e}$的下標d_{else}以及下標 e 代表的為比對到的字典以及比對到的詞的結尾位置，其餘依此類推。

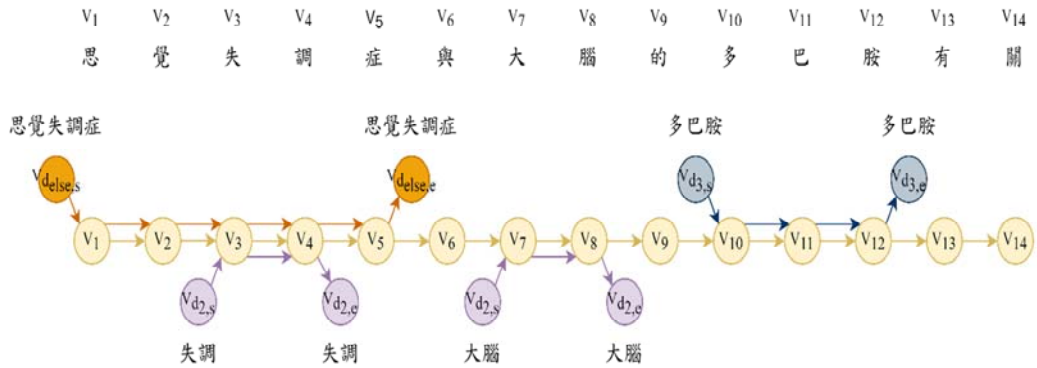

圖 2. 多維有向圖

[Figure 2. A directed multigraphs]

有向圖的結構訊息，可以透過相鄰矩陣 (adjacency matrix) 表達，假設有向圖的結構為圖 3 的左半部，而其對應的相鄰矩陣如圖 3 的右半部，其中A_{in}與A_{out}互為轉置矩陣，而相鄰矩陣由A_{in}以及A_{out}所構成。

以範例句子「思覺失調症與大腦的多巴胺有關」為例，該句子的多維有向圖的拆解成多個有向圖的範例如圖 4，由於詞彙字數為 1 個字的字典以及詞彙字數為 4 個字的字典並沒有比對到詞彙，因此對應的相鄰矩陣為零矩陣。

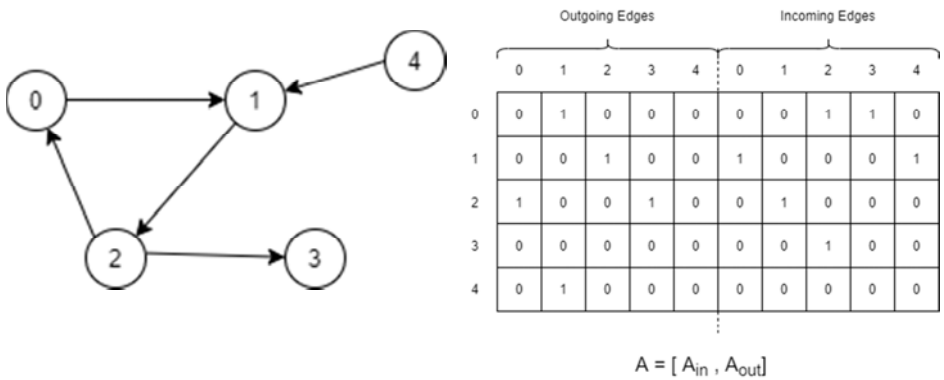

圖 3. 有向圖以及對應的相鄰矩陣

[Figure 3. A directed graph and its corresponding adjacent matrix]

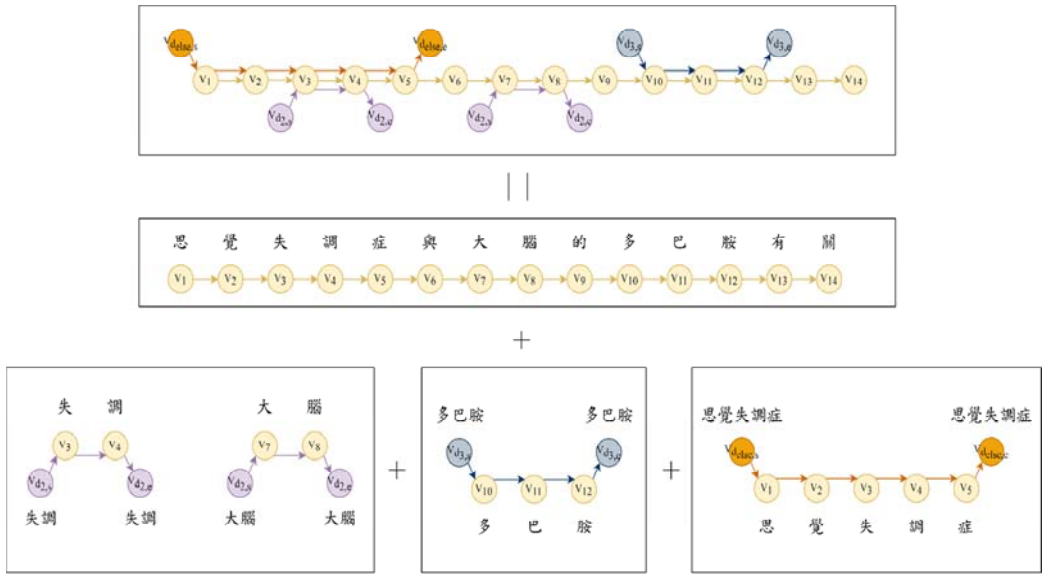

圖 4. 多維有向圖拆解成多個有向圖

[Figure 4. A directed multigraph is composed of multiple directed graphs]

由輸入句子的原始字序列訊息可以得到相鄰矩陣A_c，而由不同的字典可以得到其相對應的相鄰矩陣，依照本研究的字典分類方式可以得到相鄰矩陣A_{d_1}、A_{d_2}、A_{d_3}、A_{d_4}以及$A_{d_{else}}$，其中A_{d_1}代表的為字典詞彙字數長度為 1 的相鄰矩陣，其餘依此類推。

在本研究中，不同字典的相鄰矩陣會分別給定不同的權重，權重由以下的公式決定：

$$[w_c, w_{d_1}, w_{d_2}, w_{d_3}, w_{d_4}, w_{d_{else}}] = \sigma([\alpha_c, \alpha_{d_1}, \alpha_{d_2}, \alpha_{d_3}, \alpha_{d_4}, \alpha_{d_{else}}]) \tag{7}$$

其中$\alpha_c, \alpha_{d_1}, \alpha_{d_2}, \alpha_{d_3}, \alpha_{d_4}, \alpha_{d_{else}}$為可以被訓練的參數，並且透過 sigmod 函數使其轉換成最後的權重$w_c, w_{d_1}, w_{d_2}, w_{d_3}, w_{d_4}, w_{d_{else}}$，將不同的全權分別乘上相對應的相鄰矩陣，即可獲得最後帶有權重的相鄰矩陣。

在本研究的門控圖序列神經網路結構中，節點的初始狀態由以下公式得到：

$$h_v^{(0)} = \begin{cases} h_d(v) & v \in V_s \cup V_e \\ h_i(v) & v \in V_c \end{cases} \tag{8}$$

其中V_c代表的為多重嵌入層最後輸出的字序列特徵中，每個字分別對應到的節點，其值由多重嵌入層最後輸出的字序列特徵的值決定，V_s為命名實體的起始字對應到的節點，V_e為命名實體的最後的字對應到的節點，V_s以及V_e的值為比對到的命名實體的隨機初始狀態決定。

節點的隱藏狀態藉由 GRU 做更新，整個遞迴關係式如下：

$$H = \left[\mathrm{h}_1^{(t-1)}, \mathrm{h}_2^{(t-1)}, \dots\dots, \mathrm{h}_{|v|}^{(t-1)}\right] \quad (9)$$

$$a_v^{(t)} = [(HW_1)^T, \dots\dots, (HW_{|L|})^T]A_v^T + \mathrm{b} \quad (10)$$

$$z_v^{(t)} = \sigma(W^z a_v^{(t)} + U^z h_v^{(t-1)}) \quad (11)$$

$$r_v^{(t)} = \sigma(W^r a_v^{(t)} + U^r h_v^{(t-1)}) \quad (12)$$

$$\hat{h}_v^{(t)} = \tanh(W a_v^{(t)} + U(r_v^{(t)} \odot h_v^{(t-1)})) \quad (13)$$

$$h_v^{(t)} = (1 - z_v^{(t)}) \odot h_v^{(t-1)} + z_v^{(t)} \odot \hat{h}_v^{(t)} \quad (14)$$

其中 $h_v^{(t)}$ 表示的為節點v在時間為t時的隱藏狀態，A_v表示的為節點v對應的相鄰矩陣的行向量，公式(11)-(14)為 GRU 單元，z 與 r 分別代表更新門以及重置門，透過 GRU 單元可以結合來自相鄰節點的信息以及節點的當前隱藏狀態，計算在時間 t 時新的隱藏狀態，經過時間步數 (time step) T 後，可以得到節點的最終狀態。

3.3 雙向長短期記憶神經網路層 (BiLSTM Layer)

在這層結構中，本研究將門控圖序列神經網路層最後的隱藏狀態輸出，當作 BiLSTM 輸入序列，使用的為標準的 BiLSTM。以門控圖序列神經網路層的輸出當做輸入序列，最終可以獲得與原序列長度相同的隱藏層狀態序列。

3.4 條件隨機場域層 (CRF Layer)

命名實體辨識屬於序列標記的多分類問題，傳統上在遇到多分類問題時，會採用 softmax function 作為輸出函數，但在實際情況時，序列標註任務中的當前時刻的狀態，均與當前時刻的前後狀態有所關連，因此條件隨機場域 (Condition Random Fields, CRF) 取代了 softmax function，成為了當前主流的架構，本研究所採用的為標準的 CRF 模型。

4. 實驗評估與分析 (Performance Evaluation and Analysis)

4.1 語料庫建置 (Corpus Construction)

本研究透過爬蟲將網路上的健康照護文章及問答紀錄爬取下來，有三種來源分別為國家網路醫藥[1]、康健雜誌[2]和醫聯網[3]。其中國家網路醫藥以及康健雜誌為醫生或是相關的專業人員所撰寫的文章，而醫聯網則是一般民眾上網提問，醫生回答的問答紀錄，文章內容透過篩選主題選擇健康照護相關。本研究分別在國家網路醫藥以及康健雜誌一共爬取了 425 篇文章以及 799 篇文章，而醫療網一共有 1,818 則問答。

透過計算 Cohen's Kappa (Cohen, 1960) 值以及 Fleiss' Kappa (Fleiss, 1971) 值可以確保標記品質，其主要的功能為評估問題的一致性，其中 Cohen's Kappa 值適用於檢定兩個人意見的一致性，而 Fleiss' Kappa 值則用來檢定三人以上的情況。根據 Landis 以及

[1] 國家網路醫藥：https://www.kingnet.com.tw/knNew/index.html

[2] 康健雜誌：https://www.commonhealth.com.tw/

[3] 醫聯網：https://med-net.com/

Koch 所提出的觀點 (Landis & Koch, 1977)，當 Kappa 值小於 0 時為 Poor agreement，介於 0 到 0.20 為 Slight agreement，介於 0.21 到 0.40 為 Fair agreement，介於 0.41 – 0.60 為 Moderate agreement，介於 0.61 – 0.80 為 Substantial agreement，介於 0.81 – 1.00 為 Almost perfect agreement。

本研究所關注的健康照護命名實體，總共包含 10 類，其定義以及例子如表 1。整個標記資料的流程，我們將其分成兩個階段，參與標記的人員一共有三位師大中文系的大學生，對於每個中文句子做人工斷詞及命名實體標記，第一個階段先取國家網路醫藥 25 篇文章、康健雜 25 篇文章以及醫聯網 100 則問答，先做第一次標記，計算三位標記人員的一致性，得到 Fleiss' Kappa 值為 0.80。對階段一的標記結果做討論，修正標記準則得到一致的標準後，再對另外的 25 篇國家網路醫藥的文章、25 篇康健雜誌的文章以及醫聯網 100 則問答做標記，得到 Fleiss' Kappa 值為 0.89，達到了 Landis 以及 Koch 所認為的 Almost perfect agreement，確認階段二的 Fleiss' Kappa 有明顯上升，且達到可接受的範圍後，剩餘的待標記的中文句子，則由三位標記人員分工各自標記。

表 1. 命名實體類別定義及範例
[Table 1. Named entity definitions and examples]

類別	定義	範例
人體 (Body)	泛指生物體的細胞、組織、器官和系統。	細胞核、神經組織、心、肺、脊髓、呼吸系統等。
症狀 (Symptom)	又稱病徵，由患者描述的主觀感受，而非直接量測得知。	流鼻水、頭昏、發燒、咳嗽、失眠、貧血等。
醫療器材 (Instrument)	包含診斷、治療、減輕與預防人類疾病，使用的儀器、器械、附件、配件與零件。	血壓計、達文西機器手臂、人工髖關節等。
檢驗 (Examination)	利用醫療器材對人體健康狀態及生理功能評估。	聽力檢查、顯微鏡檢查、核磁共振造影等。
化學物質 (Chemical)	人體由不同的化學物質組成，隨著年齡與健康狀況有所增減。	去氧核糖核酸、三酸甘油酯、糖化血色素等。
疾病 (Disease)	指人體在外在因素的損害或內在機能不良情況下，影響部分或全部器官異常，伴隨特定症狀的醫學病症。	小兒麻痺症、帕金森氏症、憂鬱症、青光眼、腦溢血、肺結核等。
藥品 (Drug)	泛指用來做診斷、治療、預防疾病或減輕痛楚的藥物或化學成份。	阿斯匹靈、亞硝酸鈉、亞鐵鹽、抗生素等
營養品 (Supplement)	指從食物中萃取對人體有益的營養素，主要功能是維持健康和預防疾病。	膠原蛋白、益生菌、綜合維他命、葉黃素等。
治療 (Treatment)	讓患者恢復健康的治癒方式。	藥物治療、血漿置換、免疫球蛋白注射等。
時間 (Time)	描述患者患病症狀的持續時間或是某個時刻。	嬰兒期、幼兒時期、青春期、生理期、孕期等。

整個語料庫最後包含 30,692 句，總字數約 150 萬字，接近 92 萬個詞，68,640 個命名實體。訓練資料是三個標記人員各自標記的部分，共有 28,161 句，每個句子平均 49.44 個字 (29.99 個詞)，總共有 61,155 個命名實體，平均每個句子有 2.17 個。測試資料來自三個標記人員共同標記有一致結果的 2,531 句，每個句子平均 47.92 個字 (28.67 個詞)，總共有 7,305 個命名實體，平均每個句子有 2.89 個。10 個類別在訓練和測試資料分佈相似，最多的命名實體類別是人體，約佔 38%，依序是症狀、疾病和化學物質，前四大類佔總數的 82%，其餘 6 類約佔總數的 18%。

4.2 嵌入向量 (Embedding)

本研究所使用的嵌入方式為 Word2vec，訓練的資料來源為維基百科，下載語料庫的日期為 2020 年 2 月 3 日，利用此檔案我們可以訓練出字嵌入、部首嵌入以及詞嵌入，詞頻設定為至少出現 5 次以上，向量的維度的設定皆為 50 維，最終獲得 863,835 個詞嵌入向量，13,581 個字嵌入向量以及 3,209 個部首嵌入向量。

4.3 實驗設定 (Settings)

本研究所使用的字典來源一共分為三個，分別為國家網路醫藥[4]、國家教育研究院[5]以及搜狗網[6]，其中國家網路醫藥的詞彙主要為常見的醫護名詞，國家教育研究院選用的資料為醫學名詞，而搜狗網所包含的內容為 ICD-10、人體穴位名稱、醫學詞彙、醫療檢驗以及醫療器材等等，在使用字典時，將上述字典先合併後分類，依照詞彙字數一共分成五個字典，1 個字的字典有 351 個詞，2 個字的字典有 7,978 個詞，3 個字的字典有 19,282 個詞，4 個字的字典有 31,444 個詞，詞彙字數為 5 個字以上的字典有 95,362 個詞。

在訓練過程中學習率 (learning rate) 以及訓練資料會隨著時期 (epoch) 調整，單數 epoch 的 learning rate 為 0.001，資料為原始整份的訓練資料，雙數 epoch 的 learning rate 為 0.0005，資料為尚未學習好的訓練資料，判斷的依據為命名實體辨識是否有錯誤。其中之所以會針對尚未學習好的資料再學習一遍的原因為理論上在訓練的過程中，會希望模型能夠將所有的訓練資料學習正確，epoch 的設定值為 80， batch size 為 32，LSTM 隱藏層的維度為 200 維，GGSNN 的更新次數 (time step) 設定為 2。

4.4 效能評估 (Evaluation)

目前在命名實體辨識領域的主要評估方法為精確率 (Precision)、召回率 (Recall)、F1-score，在本研究中評估方式採精準比對 (exact match)，意即預測的結果需與正確結果完全相符才算正確。混淆矩陣矩陣範例如表 2，藉此矩陣計算精確率 (Precision) 為「正確被辨識的項目」占「總辨識項目」的比例，召回率 (Recall) 為「正確被辨識的項目」

[4] 國家網路醫藥：https://www.kingnet.com.tw/diagnose

[5] 國家教育研究院：https://terms.naer.edu.tw/

[6] 搜狗網：https://pinyin.sogou.com/dict/

占「應該被辨識的項目」的比例以及 F1-score 此為 Precision 以及 Recall 的調和平均數，計算公式如方程式 (15)-(17)。

表 2. 混淆矩陣
[Table 2. The confusion matrix]

真實值 \ 預測值	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TF)

$$\Pr ecision = \frac{\left|TP\right|}{\left|TP + FP\right|} \tag{15}$$

$$\mathrm{Re}\, call = \frac{\left|TP\right|}{\left|TP + FN\right|} \tag{16}$$

$$F1 - score = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{17}$$

4.5 實驗結果 (Results)

我們比較了以下中文命名實體模型的效能差異

(1)、BiLSTM-CRF (ICCPOL 2016)：

此模型實作了 Dong 等人(2016) 的架構，以字作為基礎當作模型輸入，字嵌入使用透過 4.2 節中所提到的維基百科語料庫當作訓練資料，向量維度為 200 維。

(2)、ME-CNER (CIKM 2019)：

Xu 等人(2019) 提出的模型，本研究實作的模型架構將其稍做更動，本研究認為將字嵌入分別經過 BiLSTM 以及 Convolutions 比起分別經過 BiLSTM-Convolution 以及 Convolutions 後連接，其中前者的 BiLSTM 較能保留原始的訊息，且比原模型效能好。

(3)、Gazetteers (ACL 2019)：

此模型為 Ding 等人所提出 (Ding *et al*., 2019)，在其發表的論文中有提供開源程式碼，因此將資料替換成本研究所使用的資料，參數的設定與原始程式碼相同，由於開源程式碼並未提供模型所會用到的的字嵌入以及二元嵌入，而在其官網的說明為使用維基百科語料庫進行訓練即可，因此使用 4.2 節中所提到的維基百科當作訓練資料，訓練出各 200 維的向量。

(4)、Lattice (ACL 2018)：

此模型為 Zhang and Yang 等人所提出 (Zhang & Yang, 2018)，利用其論文中提到的開源程式碼，將資料替換成本研究所使用的資料，模型設定的參照原始程式碼，而模型會使

用到的字嵌入以及詞嵌入由開源程式碼所提供。

(5)、GGSNN：

此為本研究提出的模型，在第三章有詳細的介紹。其中 - radical 為此模型 GGSNN 去除部首嵌入。- word 為此模型 GGSNN 去除詞嵌入。 - radical - word 則為此模型 GGSNN 同時去除部首嵌入以及詞嵌入。

表 3 為模型效能比較結果。ME-CNER 與 BiLSTM-CRF 兩者的差異為是否有加入部首嵌入以及詞嵌入，從實驗結果得知 ME-CNER 相較於 BiLSTM-CRF 提升了 2.59 的 F1，因此加入部首嵌入以及詞嵌入有助於提升模型的表現。Gazetteers 與 BiLSTM-CRF 兩者主要的差異為是否加入字典的資訊，從實驗結果得知 Gazetteers 相較於 BiLSTM-CRF 提升了 2.7 的 F1，因此透過 GGSNN 將字典的資訊納入考慮，可以有效的提升模型的表現。本研究提出的 GGSNN 模型，同時加入了部首嵌入、詞嵌入以及圖序列神經網路，其表現與 ME-CNER 以及 Gazetteers 比較，分別上升了 1.54 以及 1.43 的 F1。

由 GGSNN 分別去除掉部首嵌入、詞嵌入以及同時去除兩者的實驗比較中，可以更加地確認部首嵌入以及詞嵌入對於模型的表現影響，去除部首嵌入模型的 F1-score 下降了 0.61，去除詞嵌入模型的 F1-score 下降了 1.41，同時去除兩者模型的 F1-score 下降了 1.69，因此我們可以得知詞嵌入對於提升模型的表現的貢獻較大，而不論是詞嵌入或是部首嵌入，皆對模型的表現有幫助。

本研究提出的 GGSNN 模型有最佳的 F1 分數，而 Lattice 次之，兩個模型的差異為略小只有 0.47，然而在訓練的時間方面，相同的硬體設備下，本研究的模型約為 1 天，而 Lattice 約耗時 6.25 天，主要的原因為 Lattice 模型的 batch size 因為模型的特性只能夠設定為 1，當資料量越大時，需要更多時間，無法藉由調整 batch size 加速運算。

表 3. 模型結果比較
[Table 3. Model performance comparisons]

Method	Precision	Recall	F1
BiLSTM-CRF (ICCPOL 2016)	70.38	72.77	71.56
ME-CNER (CIKM 2019)	73.68	74.62	74.15
Gazetteers (ACL 2019)	73.00	75.56	74.26
Lattice (ACL 2018)	74.69	75.76	75.22
GGSNN (ours)	**75.46**	75.76	**75.69**
- radical	73.50	**76.73**	75.08
- word	73.48	75.10	74.28
- radical - word	73.46	74.54	74.00

4.6 錯誤分析 (Error Analysis)

本研究將命名實體的錯誤分成以下 5 種類型，錯誤範例如表 4。

- CONTAIN：正確的命名實體「包含」預測的命名實體。
- CONTAINED：正確的命名實體「被包含於」預測的命名實體。
- SPLIT：正確的命名實體或是預測的命名實體被拆成兩段命名實體。
- CROSS：正確的命名實與預測的命名實體之間「有」重疊的字。
- NO-CROSS：正確的命名實體與預測的命名實體之間「沒有」重疊的字。

5 種類型的錯誤總共有 2,193 個，其中最多的錯誤類型為 NO-CROSS，約佔 72%。我們觀察後得知，有些領域詞彙例如：血清胺基丙酮酸轉化酶 (SGPT)、攝護腺肥大症候群 (BPH) 和胞漿精子注射 (ICSI) 等，沒有在訓練資料中，也不屬於字典中的詞彙，無法被正確辨識。藉由錯誤分析得知，字典詞彙涵蓋程度對模型效能有重要的影響。

表 4. 命名實體預測錯誤類型與範例
[Table 4. NER error types and corresponding examples]

CONTAIN	答案	國際間 德國麻疹$_{DISE}$ 仍有疫情發生，所以有出國計畫要預先做好安排。
	預測	國際間 德國麻疹$_{DISE}$ 仍有疫情發生，所以有出國計畫要預先做好安排。
CONTAINED	答案	肺主脈 指橫膈膜$_{BODY}$ 銜接心臟的部分。
	預測	肺主脈 指橫膈膜$_{BODY}$ 銜接心臟的部分。
SPLIT	答案	喉嚨痛$_{SYMP}$ 主要是我們的扁桃腺發炎。
	預測	喉嚨$_{BODY}$痛$_{SYMP}$ 主要是我們的扁桃腺發炎。
CROSS	答案	對於 痰濁$_{SYMP}$ 瘀阻經絡$_{SYMP}$ 而致的症狀有改善的功能。
	預測	對於 痰濁瘀$_{SYMP}$阻經絡$_{BODY}$ 而致的症狀有改善的功能。
NO-CROSS	答案	鉀離子量若攝取充足，可降低腦血管 阻塞$_{SYMP}$ 風險。
	預測	鉀離子量若攝取充足，可降低腦血管 阻塞 風險。

5. 結論與未來研究 (Conclusions and Future Work)

我們提出門控圖序列神經網路模型，用於中文健康照護命名實體辨識。主要貢獻如下：

一、我們提出一個多重嵌入導向的圖序列網路架構，從部首、字到詞的不同語意資訊被探索，透過圖神經網路調適至健康照護命名實體辨識任務。我們的模型達到 75.69 的 F1 顯著優於先前的命名實體辨識方法。

二、據我們所知，這是第一個健康照護領域的中文命名實體語料庫，包含 30,692 個句子，約莫 150 萬字 (92 萬詞)，共有 68,460 個命名實體，橫跨 10 個類別，包含：人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療以及時間。

利用命名實體辨識的這項技術，我們可以依照各領域不同的需求，從非結構的文章中抽取出該領域所關注的命名實體，透過這些抽取出的命名實體，我們可以充分的掌握文章中的資訊，對文章做更進一步的分析，在未來的應用中，命名實體辨識所標示出的命名實體，可以做為關係擷取、事件偵測與追蹤、知識圖譜建置、智慧問答系統等應用的基礎。

致謝 (Acknowledgements)

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3. We sincerely thank all the annotators for their efforts in the named entity tagging task. We also thank the anonymous reviewers for their insightful comments.

參考文獻 (References)

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: 10.1177/001316446002000104

Ding, R., Xie, P., Zhang, X., Lu, W., Li, L., & Si, L. (2019). A neural multi-diagraph model for Chinese NER with gazetteers. In *Proc. of the ACL'19*, 1462-1467. doi: 10.18653/v1/P19-1141

Dong, C., Zhang, J., Zong, C., Hattori, M., & Di, H. (2016). Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Proc. of the ICCPOL'16*, 239-250. doi: 10.1007/978-3-319-50496-4_20

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. doi: DOI: 10.1037/h0031619

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of the ICML'01*, 282-289.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proc. of the NAACL-HLT'16*, 260-270. doi: 10.18653/v1/N16-1030

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi: 10.2307/2529310

Levow, G. A. (2006). The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In *Proc. of the SIGHAN'06*, 108-117.

Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2016). Gated graph sequence neural networks. In *Proc. of the ICLR'16*. Retrieved from arXiv:1511.05493.

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proc. of the ACL'16*, 1064-1074. doi: 10.18653/v1/P16-1101

Peng, N., & Dredze, M. (2015). Named entity recognition for Chinese social media with jointly trained embeddings. In *Proc. of EMNLP'15*, 548-554. doi: 10.18653/v1/D15-1064

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, *77*(2), 257-286. doi: 10.1109/5.18626

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of the EMNLP/VLC'00*, 63-70. doi: 10.3115/1117794.1117802

Xu, C., Wang, F., Han, J., & Li, C. (2019). Exploiting multiple embeddings for Chinese named entity recognition. In *Proc. of the CIKM'19*, 2269-2272. doi: 10.1145/3357384.3358117

Zhang, Y., & Yang, J. (2018). Chinese NER using lattice LSTM. In *Proc. of the ACL'18*, 1554-564. doi: 10.18653/v1/P18-1144