國 立 中 央 大 學

電 機 工 程 學 系

碩 士 論 文

多重嵌入增強式門控圖序列神經網路之中文健康照護命名實體辨識

Multiple Embeddings Enhanced Gated Graph Sequence Neural Networks for Chinese Healthcare Named Entity Recognition

研 究 生：盧毅

指導教授：李龍豪 博士

中 華 民 國　　一百零九　　年　　六　　月

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。我們提出一個多重嵌入增強式門控圖序列神經網路 (Multiple Embeddings Enhanced Gated Graph Sequence Neural Network, ME-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, BERT, 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。

關鍵詞：嵌入向量、圖神經網路、命名實體辨識、資訊擷取、健康資訊學

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 ME-GGSNN (Multiple Embeddings enhanced 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, BERT, Lattice, Gazetteers, and ME-CNER. In summary, our ME-GGSNN model is an effective and efficient solution for the Chinese healthcare NER task.

Keywords: embedding representation, graph neural networks, named entity recognition, information extraction, health informatics

致謝

時光飛逝，研究所的兩年時光已接近尾聲，包含大學的四年，我總共在國立中央大學待了整整六年，在此由衷的感謝這一路上曾經幫助過我的人。

在研究所的生涯當中，首先最要感謝的是我的指導教授李龍豪老師，老師每個禮拜都會將自己的寶貴時間撥出給所有研究生，與我們仔細地討論研究，給予我研究上的寶貴意見，每當有研究上的問題或是缺乏的資源，只要跟老師說一聲，老師便會馬上的想辦法解決，對於研究的態度以及方法，是我良好的學習典範。

謝謝口試委員曾元顯教授以及禹良治教授對於論文提出的寶貴建議，使得我的論文內容能夠更加的完整充分，並且在百忙之中抽空來中央大學，讓我可以順利的完成口試。

感謝我研究室的同學鼎鈞以及昱翔，每當有研究上的問題彼此都可以互相討論以及交流，待在實驗室一起為研究努力，還有要感謝實驗室學弟們昌浩、少鈞、浩銓以及柏翰的加入，讓原本三個人的實驗室變得加歡樂，讓我的研究所生活更加的充實，我會懷念與所有實驗室夥伴的相處時光。

最後我要感謝我的父母以及凱琦，支持著我完成研究所學業，讓我可以無後顧之憂的專心研究，給予我穩定溫暖的力量，往人生的下一道關卡邁進。

盧毅 謹致於國立中央大學電機所
中華民國 109 年 7 月

目錄

圖目錄

表目錄

第一章　緒論

1-1　研究背景

人類之所以與其他動物有所不同的主要原因之一為我們擁有文字，透過文字我們可以將所學的知識傳遞下去，因此閱讀文字一直以來便是人類認識世界獲取知識的方式之一，透過報章雜誌、書籍以及文獻等等，可以滿足我們對於知識的渴求。然而在現在資訊化時代，上述的報章雜誌、書籍以及文獻等等許多皆已電子化，因此，如何透過電腦幫助我們處理這些龐大的資訊量，便是人們近年來所關注的課題。

「自然語言處理」 (Natural Language Processing, NLP) 即為能夠幫助我們達成上述目標的技術，其主要研究目的為讓電腦能夠有處理、理解以及運用人類語言的能力，屬於計算機科學與語言學的交叉學科，因此又被稱作為計算語言學。而所謂的自然語言，即為人們溝通時自然地發展出來的語言，與之相對應的則是程式語言，程式語言為人類人工設計的語言。

目前在中文領域的自然語言處理發展遇到的主要困難點之一為單詞的邊界判定，即所謂的斷詞，在英文領域中，可以透過空白字元當作斷詞的分割依據，然而中文並沒有空白字元可以判定詞與詞的邊界，因此中文領域的自然語言處理較英文領域更為複雜困難，斷詞的精準度往往會對後續的處理以及應用有重大的影響。

本研究所關注的主題為「命名實體辨識」 (Named Entity Recognition, NER)，此任務的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、專有名詞等等。舉例來說「比爾蓋茲創辦了微軟」，假設所關注的命名實體為人名以及組織名，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，像是透過抽取出的人名「比爾蓋茲」以及組織名「微軟」，我們可以透過後續的處理來推斷兩者之間關係為「創辦」。

早期的 NER 方法為主要是基於字典或是規則，利用比對來做辨識，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以及制定規則得知命名實體位置的所有可能情況，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過的語料來學習出標註模型，從而對句子的各個位置進行標註，但同樣需要事先透過人工定義好特徵，因此特徵的選取也仰賴專業人士的制定，特徵的好壞對於整個標註的結果有直接的影響，主要使用的模型有：隱藏式馬可夫模型 (Hidden Markov Model, HMM) [1] 、最大化熵馬可夫模型 (Maximum Entropy Markov Model, MEMM) [2] 和條件隨機場域 (Conditional Random Field, CRF) [3]。

近年來由於電腦計算能力的進步，帶動深度學習的興起，因此近期的主流為深度學習模型，與前兩類方法不同，此方法的好處為不需要透過人工制定複雜的特徵，深度學習模型會自動學習出重要的特徵，因此只需要標註過的大規模語料庫即可訓練出標註模型。目前較常見的深度學習模型主要有：卷積網路 (Convolutional Neural Network, CNN) [4]、遞歸神經網 (Recurrent Neural Network, RNN) [5]，其中 CNN 較常使用在影像辨識領域，而 RNN 則是較常使用在自然語言處理領域，其原因 RNN 可以處理時間序列的問題，但是單純的 RNN 模型無法擷取長距離的文章資訊，因此目前較常被使用的為其改良後的長短期記憶模型 (Long Short Term Memory, LSTM) [6]，LSTM 與 RNN 不同的地方在於 LSTM 在神經單元中加入了遺忘、更新以及輸出三個步驟，進而大幅提高了其在長期記憶的表現。

1-2 研究動機與目的

隨著科技的進步,人類的壽命得以延長,因此有關健康照護的議題逐漸地浮上檯面,在數位化的時代,在找醫生進行診斷前,人們往往會先在網路上搜尋相關的文章、雜誌以及問答紀錄以獲取相關的訊息,事前所搜尋到的資訊,往往會決定人們對於所面臨的健康照護問題採取的態度以及措施,然而有時候文字內容涉及一些專業的名詞,可能會造成閱讀理解上的障礙,導致對於相關資訊的理解並不全面,這時如果能夠將內容中某些艱澀的專有名詞,做簡單的名詞解釋,可以幫助閱讀者更容易了解文章的內容,對於所面臨到的健康照護問題採取更妥當的解決方式。

透過本研究所要探討的主題中文健康照護命名實體辨識,即可完成上述的功能,利用相關的語料庫進行模型的訓練後,將訓練好的命名實體辨識模型,對文章進行序列標註,將其中所關注的命名實體找出,例如:疾病、症狀、化學物質以及治療等等,並透過後續像是維基百科、相關字典等等的應用,即可讓文章讀者對於所標註出的命名實體有一定程度的了解。

有鑑於此,本研究的主要目的為以下兩點:

一、建立一個中文健康照護領域的命名實體語料庫

有鑑於目前缺乏健康照護的中文命名實體辨識語料庫,因此本研究從網路上蒐集了有關健康照護相關的文章雜誌以及問答紀錄,總共分成三個來源,分別為國家網路醫藥、康健雜誌和醫聯網,將蒐集完的文章雜誌以及問答紀錄給標記人員標記,並且在標記時透過計算 Cohen's Kappa 值以及 Fleiss Kappa 值以確保標記資料的品質

二、提出一個中文康照護命名實體辨識模型

利用標註好的資料訓練適合的健康照護領域模型,根據健康照護領域的資料特殊性,本研究針對其領域挑選適合加入的資訊,透過此訓練好的標註模型,即可對於所關注的命名實體,例如:疾病、症狀、化學物質以及治療等命名實體類型進行標註。

1-3 章節概要

本論文一共分為五個章節：

- 第一章節為緒論，內容包含研究背景、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，說明建立中文健康照護語料庫的原因，以及介紹命名實體辨識的模型演變過程，並整理當前已知的中中文命名實體辨識模型。
- 第三章節為模型架構，詳細介紹本研究所提出的神經網路模型，並對模型的各層做詳盡的說明。其中多重嵌入層的功能為將字嵌入、部首嵌入以及詞嵌入組合成多重嵌入，門控圖序列神經網路層的功能為將所準備的字典，利用字串比對產生圖後，透過門控圖序列神經網路，把每個字融入字典的資訊，雙向長短期記憶神經網路層的功能為抽取含有字、部首、詞以及字典資訊的序列特徵。條件隨機場域層的功能為對序列進行標記，輸出機率值最高的序列。
- 第四章節為實驗結果，此章節首先說明語料庫的建置，並且對資料進行統計，實驗設定的部分包含了使用的字典、模型訓練流程以及模型參數，接著描述如何使用字嵌入、詞嵌入以及部首嵌入，而模型評分的指標採用 Precision、Recall 和 F1-score，最後對模型進行比較與效能分析，並對模型預測做錯誤分析。
- 第五章為結論與未來工作。

第二章　相關研究

在做命名實體辨識的任務時，首先會遇到的問題便是語料庫來源，而當有了資料以後便能根據資料找出適合的方式標註命名實體，因此相關研究的章節中涵蓋了兩個主題，分別為語料庫和辨識模型，在接下來的兩個小節中分別將會統整較為人所知的中文命名實體語料庫以及近年來較受人關注的中文命名實體辨識模型。

2-1　中文命名實體辨識語料庫

目前在中文領域較為常見的命名實體語料庫有 SIGHAN 2006 MSRA [7]、Weibo [8]、Resume [9] 以及 CCKS-2019 [10]，這四個語料庫的來源不一且所關注的命名實體也不盡相同，但每個語料庫皆具一定的規模並且都有相關研究人員使用，因此本研究將對這四個語料庫進行介紹。

MSRA 其所包含的命名實體包含總共 30 種類別，語料來源為新聞文章，其中較被廣泛使用的類別為像是人名 (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，人名、地名以及組織名分別的個數為 17,615 個、36,860 個以及 20,584 個，平均每個句子含有 2.56 個命名實體，而測試資料為 4,365 個句子，其中的命名實體總數為 4,362 個，人名、地名以及組織名分別的個數為 1,973 個、2,886 個以及 1,331 個，平均每個句子含有 1 個命名實體。

在社群媒體方面，Weibo 語料庫蒐集了微博此社群媒體從 2013 年 11 月置 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)，其中個命名實體的數量分別為 243 個、126 個、255 個以及 1,357 個，而每則訊息所包含的命名實體至少為 3 個以上才會被加入語料庫。

Resume 的來源為個人履歷，履歷的出處為中國上市公司的主管，總共隨機挑選了 1,027 份，這當中所標註的 8 類命名實體種類分別為國家 (Country)、人名 (Person)、地

名 (Location)、組織名 (Organization)、專業 (Profession)、教育程度 (Educational Institution)、種族背景 (Ethnicity Background) 以及工作職稱 (Job Title)，其中命名實體的數量分別為 321 個、1,174 個、55 個、5,687 個、338 個、1,076 個、144 個和 7,770 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 年舉辦的比賽中的其中一個項目為命名實體辨識，而該組織也曾在 2017 以及 2018 舉辦過類似的命名實體辨識比賽，比賽的資料來源為電子病例 (Electronic Health Record, EHR)，訓練集的文檔數總共為 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)、檢驗 (Inspection)、手術 (Operation)、藥物 (Drug) 以及解剖部位 (Anatomy)，其中各命名實體的數量分別為 682 個、91 個、193 個、140 個、263 個以及 447 個。

上述的命名實體語料庫，並沒有關於醫療照護領域方面的語料庫，且都為簡體中文，而在繁體中文領域的公開命名實體語料庫更是不常見，因此本研究建了一個有關醫療照護的命名實體語料庫，其中所關注的命名實體總共有 10 類，分別為人體、疾病、症狀 、化學物質、藥品、 營養品、醫療器材、 檢驗、治療以及時間。

表 1、中文命名實體辨識語料庫列表

資料集	年份	命名實體	資料來源
MSRA [7]	2006	人名、地名以及組織名	新聞文章
Weibo [8]	2015	地理位置人名、地名以及組織名	社群媒體
Resume [9]	2018	國家、人名、地名、組織名、專業、教育程度、種族背景以及工作職稱	履歷
EHR [10]	2019	疾病和診斷、檢查、檢驗、手術、藥物以及解剖部位	電子病例

2-2 中文命名實體辨識模型

命名實體辨識在許多相關研究中，被視為序列標註的問題，在中文領域的命名實體辨識，早期所使用的模型包含隱藏式馬可夫模型 (Hidden Markov Model, HMM) [11]、最大化熵馬可夫模型 (Maximum Entropy Markov Model, MEMM) [12]以及條件隨機場域 (Conditional Random Field, CRF) [13] [14]，目前最被廣為使用的為 Lafferty [3] 所提出的條件隨機場域，大部分的模型皆使用為其當作最後的輸出。

由於中文的獨特性，斷詞的好壞對於整個模型的表現有直接的影響，然而在 2006 年 SIGHAN 的 Closed Track 比賽中，以詞為單位的模型與以字為單位的模型，兩者之間的表現並無明顯的差異，因此這意味著在中文領域的命名實體辨識能以字為單位做輸入，而模型仍舊可以找到命名實體的邊界，這樣的好處在於即使斷詞不正確，模型依然可以將斷詞錯誤的命名實體辨識正確，特別是在某些斷詞特別容易錯誤、文字內容易牽涉專有名詞以及包含許多未知新詞 (Out-Of-Vocabulary, OOV) 的領域，以字為單位作為輸入，是一個比較好的方法。

近年來深度學習的興起，神經網路在許多地方皆有著亮眼的表現，眾多神經網路中的長短期記憶模型 (Long Short-term Memory , LSTM)主要功能為處理序列問題，因此非常適合使用在自然語言處理領域，該模型的效果已在 2015 年 Huang 等人 [15] 的研究中被證實。在 Huang 的研究中除了 LSTM 外，同時也使用了雙向長短期記憶神經網路 (Bidirectional Long Short-Term Memory, BiLSTM)，所謂的 BiLSTM 即為將前向 LSTM 以及後向 LSTM 組合而成的雙向 LSTM，在此研究中比較了 CRF、LSTM-CRF 以及 BiLSTM-CRF，其中以 BiLSTM-CRF 架構表現最好。透過此研究結果可以得知在命名實體辨識領域中，深度學習模型正式的成為往後大家普遍所採用的模型，而其中的 BiLSTM-CRF 在後續其他人的研究中最被廣為使用，為目前在命名實體辨識領域中的主流模型，圖 1 為 BiLSTM-CRF 的模型架構範例，左邊為以字作為基礎當作輸入，右邊為以詞為基礎當作輸入。

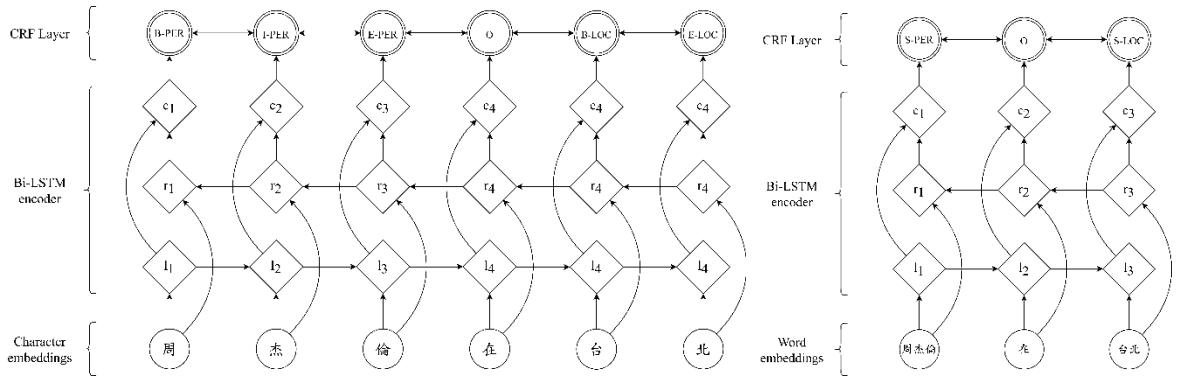


圖 1、BiLSTM-CRF 架構以字或詞作為序列輸入單位

有鑑於在英文領域中，Lample 等人 [16] 的研究考量了英文的特性，將英文單字中的字首已及字尾的特徵納入考慮，因此在中文的領域中，Dong 等人 [17] 也同樣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。如圖 2 以「朝」字為例，該字所代表的意思為「早晨」，而「朝」字可以被拆解成 4 個部件，分別為「十」、「日」、「十」以及「月」，兩個「十」代表的意思為「草」，而「日」與「月」分別帶別「太陽」以及「月亮」，所有部件所構成的意思為「太陽剛從草叢升起，月亮剛要消失的時候」，即為「早晨」的意思，因此「字」並非中文字具有意義的最小單位。

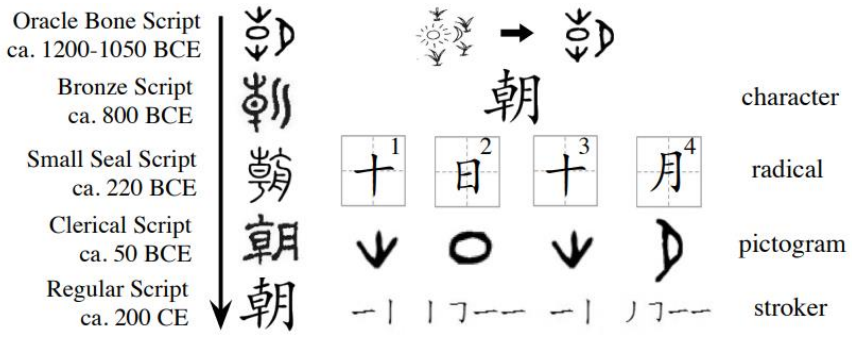


圖 2、「朝」字的部件拆解
資料來源：Dong et al. [17]

在 Xu 等人 [18] 的研究中，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 BiLSTM 以及 Convolutions 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。舉例來說，由於中國人的文化傾向在取名字時會避開帶有不好意義的字，像是部首為「疒」的字，因為「疒」代表著疾病的意思，而在像是「金」、「木」、「水」以及「火」這類部首則常常出現在名字當中。

在中文領域的自然語言處理，其中一項最大的挑戰即為斷詞，許多自然語言處理中的子任務在一開始都需要先解決斷詞，其主要原因為不好的斷詞會降低模型的表現，反之，能夠有精確的斷詞，對於模型表現的提升有重大的幫助，因此如何獲得良好的詞邊界也是命名實體辨識的一個重要課題，透過像是語料庫以及字典等等，可以幫助我們更好的判定詞的邊界。

Zhang 等人 [9] 提出了一個新的模型 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，其研究結果在命名實體辨識的任務中取得了重大的成果。該研究找出所有可能潛在詞彙的範例如圖 3，以句子「南京市長江大橋」為例，當中所包含的潛在詞會有「南京」、「市長」、「長江」、「大橋」、「南京市」以及「長江大橋」，透過該研究的 Lattice LSTM 架構，即可考慮到所有潛在詞彙的資訊。

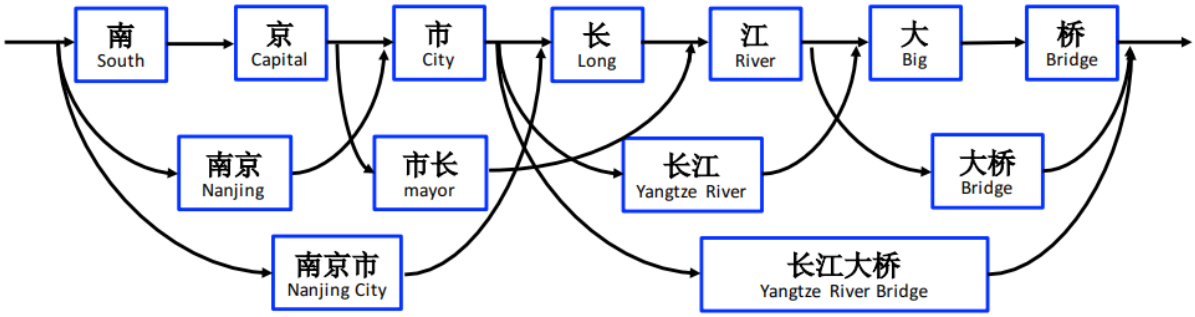


圖 3、句子中的潛在詞彙範例
資料來源：Zhang et al. [9]

在命名實體辨識的領域中，常常能夠蒐集到相關的字典，而如何將這些字典加入模型中使用，便是相關研究探討的重點之一。因此在 Ding 等人 [19]所提出的模型當中，使用到了圖神經網路中的門控圖序列神經網路 (Gated Graph Sequence Neural Network, GGSNN) [20]，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。字典引入的方入式為將句子與字典中的詞彙做比對，將所有比對到的詞彙利用圖結構表達資訊後，透過 GGSNN 學習圖結構的資訊。

基於上述研究，本研究提出了多重嵌入增強式門控圖序列神經網路 (ME-GGSNN)，由於目前當前的主流架構為 BiLSTM-CRF，因此本研究所提出的模型採用了同樣架構。而本研究所關注的領域為健康照護，在此領域中常常牽涉專有名詞以及可能包含許多的 OOV (Out-Of-Vocabulary)的詞彙，容易造成斷詞錯誤而降低模型的表現，所以選擇了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，由於在此研究的健康照護領域中所關注的命名實體如「人體」以及「疾病」等等，常常會帶有「肉」或是「疒」等部首，因此選擇將部首資訊加入，在詞的資訊方面，由於同樣的字可以構成不同的詞，而不同的詞有不同的意思，因此加詞的資訊加入可以更精準的表達文字的意思。在加入字特徵、部首特徵時透過 BiLSTM 以及 Convolution 做了有別於 Xu 等人的處理，使特徵資訊更能夠完整充分。在本研究的模型與 Ding 等人同樣透使用了改良式的 GGSNN 將字典資訊加入，而與 Ding 等人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。

表 2、中文命名實體辨識模型列表

作者	年份	主要特點
Chuanhai Dong et al. [17]	2016	加入了中文字的部件資訊
Canwen Xu et al. [18]	2019	加入了部首以及詞的資訊
Yue Zhang & Jie Yang [9]	2018	利用 Lattice LSTM 將句子的潛在詞納入考量
Ruixue Ding et al. [19]	2019	透過 GGSNN 將字典的資訊加入

第三章　模型架構

在研究中提出的多重嵌入增強式門控圖序列神經網路(ME-GGSNN)模型架構如下圖 4，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。

(1) 多重嵌入層 (Multiple Embeddings Layer)：

將輸入的字序列做前處理後，找出字序列中每個字對應的部首以及詞，得到部首序列以及詞序列後做為輸入，接著將輸入的字、部首以及詞透過預先訓練好的 Word2vec [21]，找出其對應的向量後，分別針對字嵌入、部首嵌入以及詞嵌入經過 BiLSTM 以及 Convolutions 組合成多重嵌入。

(2) 門控圖序列神經網路層 (GGSNN Layer)：

在此層結構中會使用門控圖序列神經網路將字典的資訊加入，將所蒐集的字典，透過字串比對產生多維有向圖後，利用相鄰矩陣表達多維有向圖的資訊，最後透過門控圖序列神經網路，學習多維有向圖的資訊，將字典的資訊融入模型當中。

(3) 雙向長短期記憶神經網路層 (BiLSTM Layer)：

在此層結構中，將會使用雙向長短期記憶神經網路來將帶有字特徵、部首特徵、詞特徵以及字典特徵的序列做序列特徵的抽取。

(4) 條件隨機場域層 (CRF Layer)：

由於命名實體辨識為序列標記的問題，因此在最後本研究利用 CRF 進行序列標記，輸出機率值最高的序列。

此章節接下來的內容將會對這四層結構做更仔細的描述。

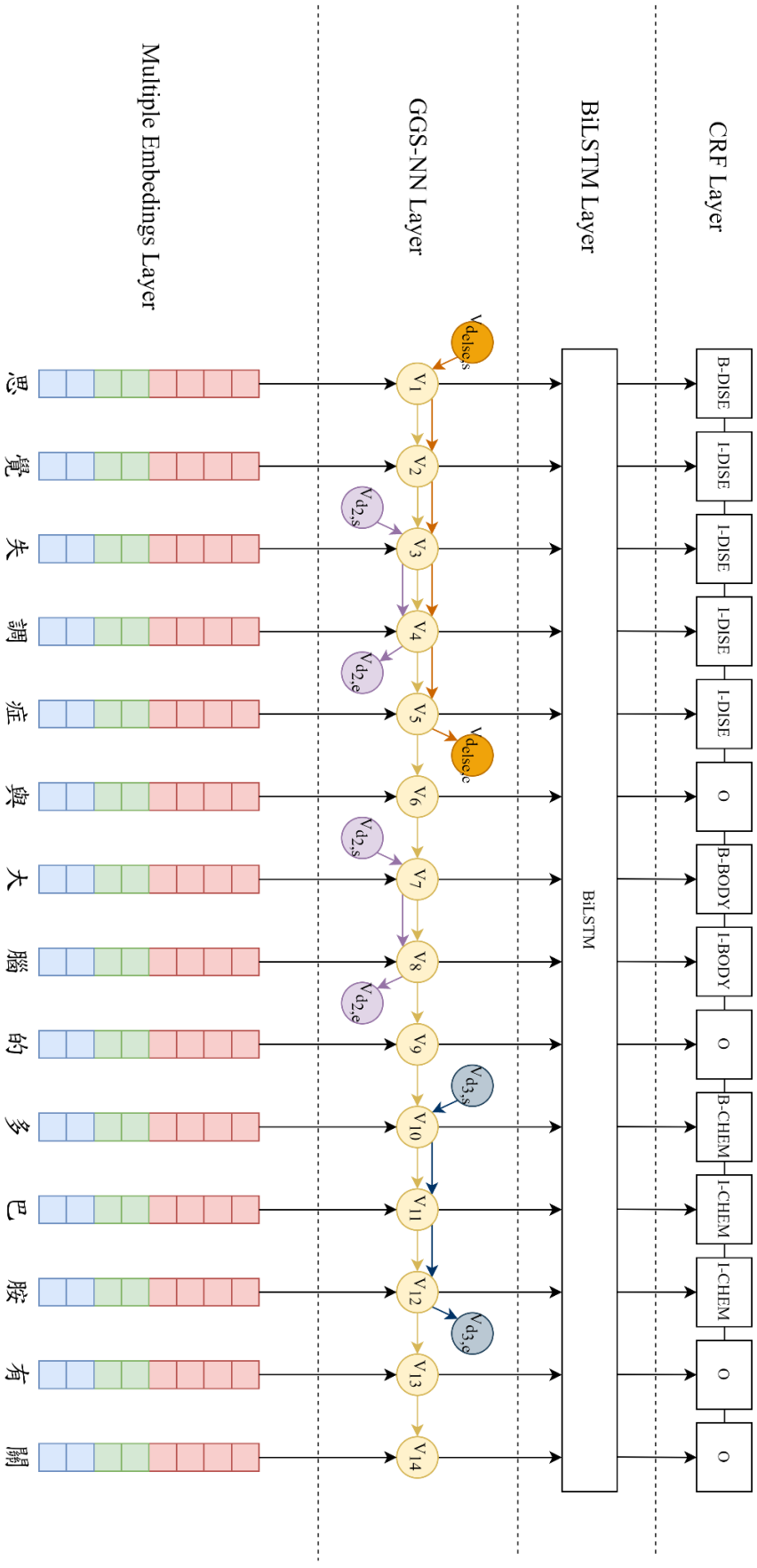


圖 4、ME-GGSNN 模型整體架構圖

3-1　多重嵌入層：

透過組合字嵌入、詞嵌入以及部首嵌入形成多重嵌入，多重嵌入層的整體架構如圖5。資料經過整理，輸入資料以句子為單位，取得句子的字序列後，找出字序列中每個字分別對應的部首以及詞，即可得到與字序列相同長度的部首序列以及詞序列。

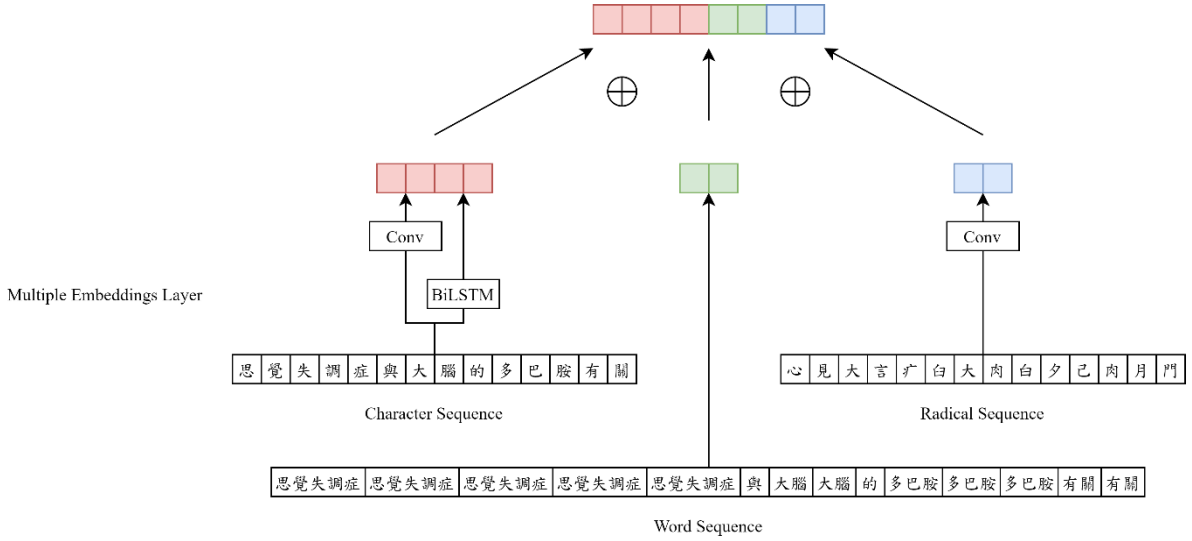


圖 5、多重嵌入向量組成示意圖

在將文字輸入深度學習模型前，需要將文字做數值化，否則電腦是無法分析的，因此在本研究中使用了 Word2vec 預訓練字嵌入、部首嵌入以及詞嵌入，並且利用 BiLSTM 以及 Convolutions 做特徵抽取，組合成多重嵌入，當作最後以字為基礎的字序列特徵。其中字嵌入、部首嵌入以及詞嵌入的處理分別如下，假設輸入句子字數長度為 n：

(1) 字嵌入 (Character Embeddings)：

輸入字序列 $X = [x_1, x_2, x_3, \ldots, x_n]$，分別經過 BiLSTM 以及 Convolutions 後，得到與原嵌入維度相同的特徵序列，接著將兩者組成新的字嵌入特徵序列，最後得到與原長度相同的序列 $C = [c_1, c_2, c_3, \ldots, c_n]$，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 Convolutions 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 Embeddings)：

輸入部首序列 $X = [x_1, x_2, x_3, \dots, x_n]$，經過 Convolutions 後，得到與原長度以及嵌入維度相同的特徵序列$[r_1, r_2, r_3, \dots, r_n]$，由於每個部首多半與附近的字有關，因此 Convolutions 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 Embeddings)：

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。

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

最終將字特徵序列、部首特徵序列和詞特徵序列，組合成多重嵌入，組合方式如下：

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

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

3-2　門控圖序列神經網路層：

相較於鏈狀結構數據或者樹狀結構數據，圖結構數據往往更加靈活，而文字的訊息常常會有類似圖結構的訊息，因此本研究透過門控圖序列神經網路，學習句子圖結構化後的訊息。

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 [20] 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，而之所以選擇此結構的原因在於命名實體辨識的任務中，往往可以蒐集到相關命名實體的字典，而蒐集到的字典常常不只一個，因此透過改良式的 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，分類的規則如 4-2 節中所提到的方式，總共分成五個字典，分別為 (1) 詞彙長度為 1 個字 (2) 詞彙長度為 2 個字 (3) 詞彙長度為 3 個字 (4) 詞彙長度為 4 個字 (5) 詞彙長度為 5 個字以上。

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) 範例如下圖 6：

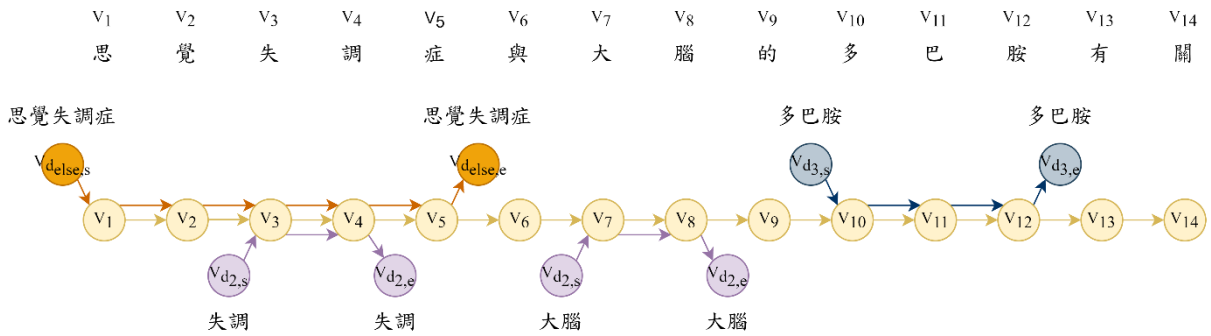
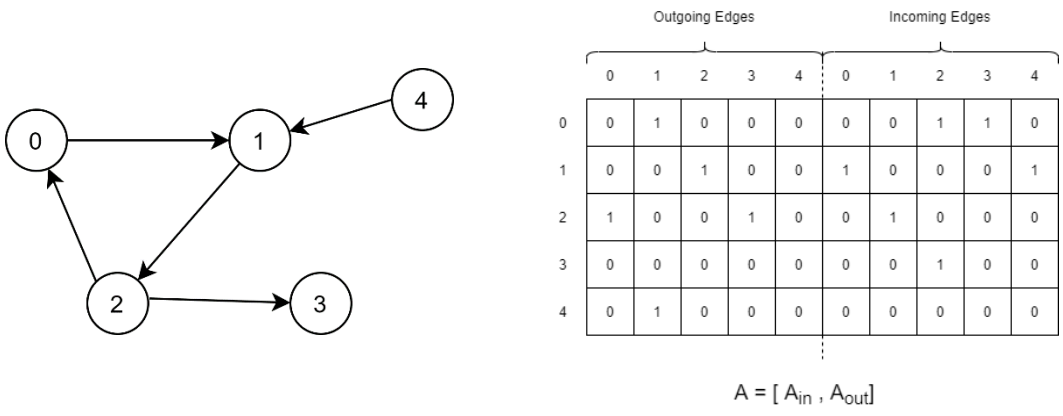


圖 6、多維有向圖範例

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$E = \{e_c\} \cup \{e_{d_i}\}_{i=1}^{\mathrm{m}}$，其中$\{e_c\}$為字序列節點連成的邊的集合，$\{e_{d_i}\}_{i=1}^{\mathrm{m}}$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$L = \{l_c\} \cup \{l_{d_i}\}_{i=1}^{\mathrm{m}}$，$l_c$為字序列節點連成的邊上的標籤，$l_{d_i}$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，可以得到圖 6，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」包含在詞彙長度為 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。

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 7 中的左半部，而其對應的相鄰矩陣如圖 7 的右半部，其中A_{in}與A_{out}互為轉至矩陣，而相鄰矩陣由A_{in}以及A_{out}所構成。



	0	1	2	3	4	0	1	2	3	4
0	0	1	0	0	0	0	0	1	1	0
1	0	0	1	0	0	1	0	0	0	1
2	1	0	0	1	0	0	1	0	0	0
3	0	0	0	0	0	0	0	1	0	0
4	0	1	0	0	0	0	0	0	0	0

圖 7、有向圖以及對應的相鄰矩陣範例

一個多維有向圖是由多個有向圖所構成，而一個相鄰矩陣可達一個有向圖的資訊，因此如果要表達一個多維有向圖，需要多個相鄰矩陣。由於本研究一共使用了 5 個字典，因此句子的多維有向圖一共包含 6 個有向圖，分別為 (1) 原始字序列的有向圖 (2) 詞彙長度為 1 個字的有向圖 (3) 詞彙長度為 2 個字的有向圖 (4) 詞彙長度為 3 個字的有向圖 (5) 詞彙長度為 4 個字的有向圖 (6) 詞彙長度為 5 個字以上的有向圖，而由於要表達的有向圖有 6 個，因此總共會產生 6 個相鄰矩陣。

以剛剛的輸入句子「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下圖 8，圖 9-12 代表原始字序列的有向圖、詞彙長度為 2 個字的有向圖、詞彙長度為 3 個字的有向圖以及詞彙長度為 5 個字以上的有向圖對應的 A_{in}矩陣，由A_{in}轉至可得A_{out}，構成最後的相鄰矩陣，由於詞彙長度為 1 個字的字典以及詞彙長度為 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。

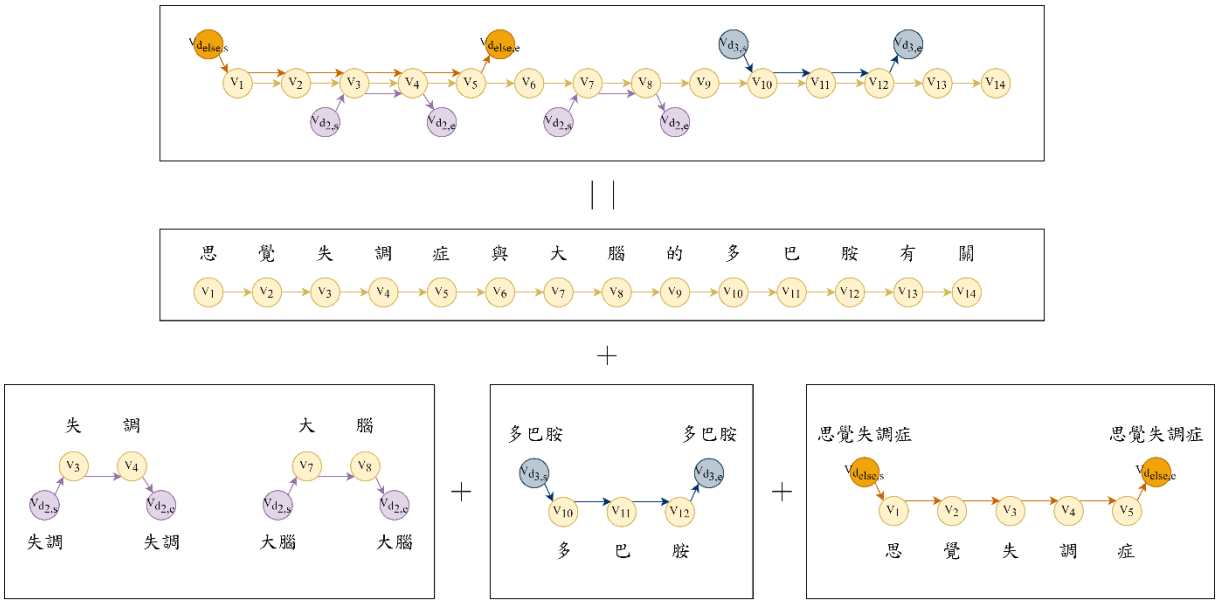


圖 8、多維有向圖拆解成多個有向圖範例

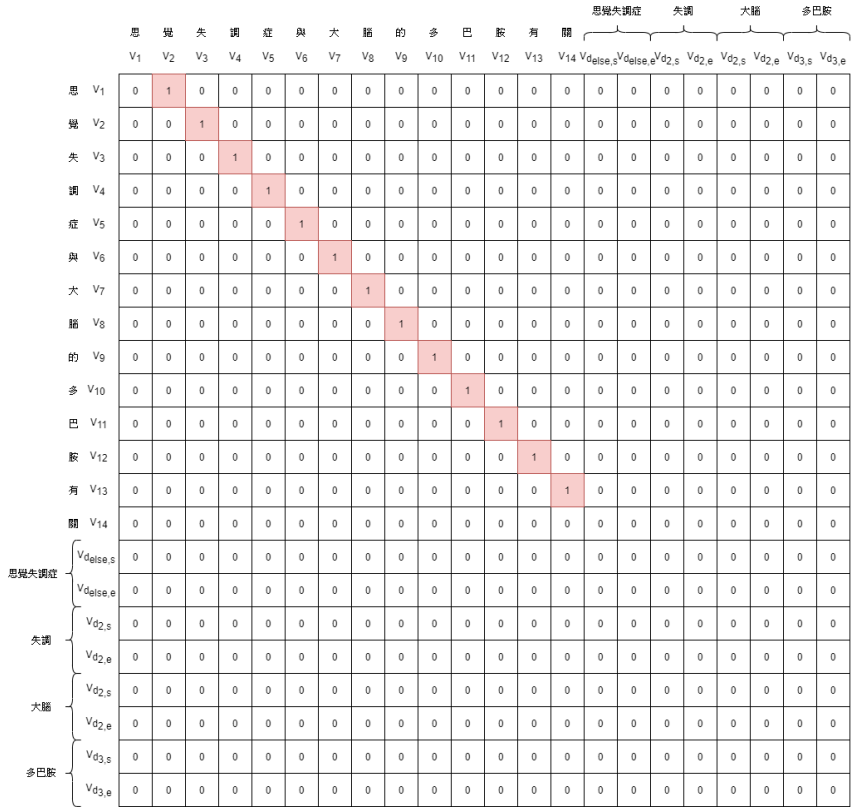


圖 9、原始字序列的有向圖對應的A_{in}矩陣

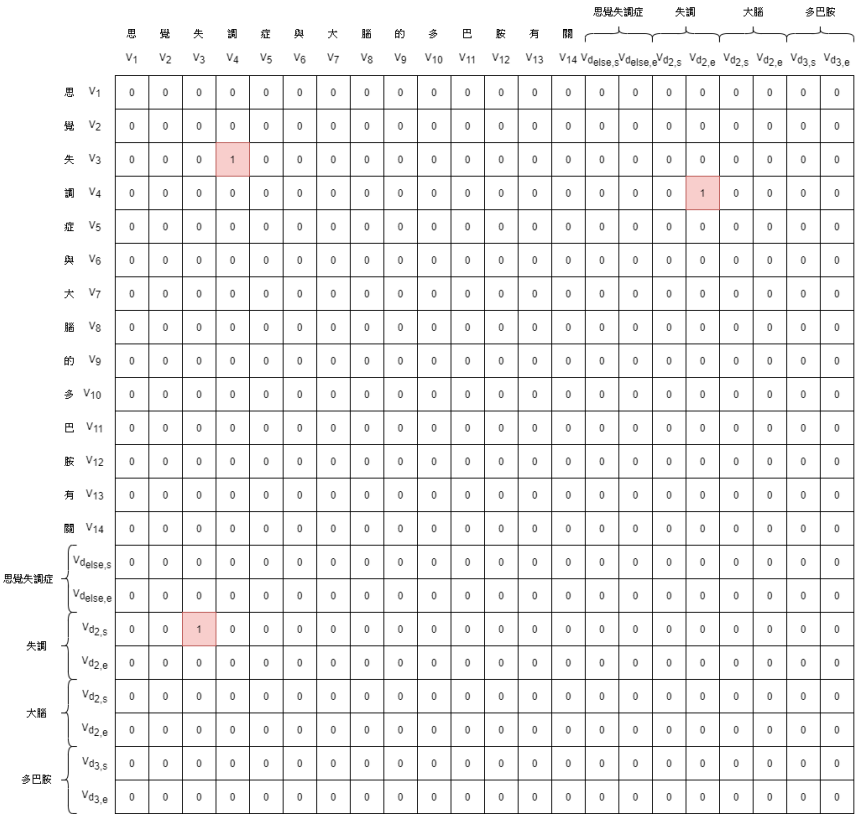


圖 10、詞彙長度為 2 個字的有向圖對應的A_{in}矩陣

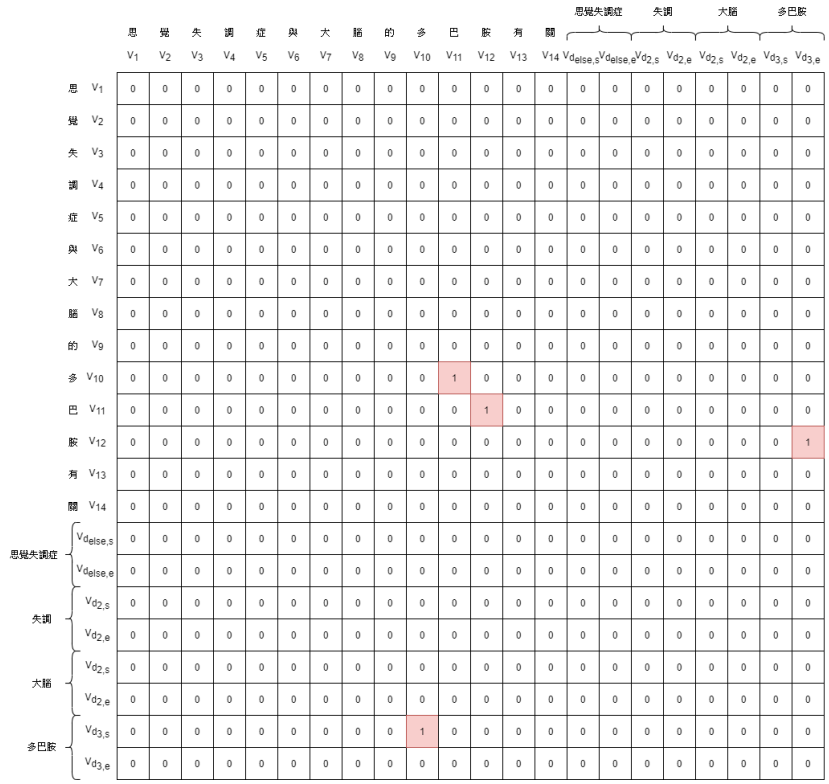


圖 11、詞彙長度為 3 個字的有向圖對應的A_{in}矩陣

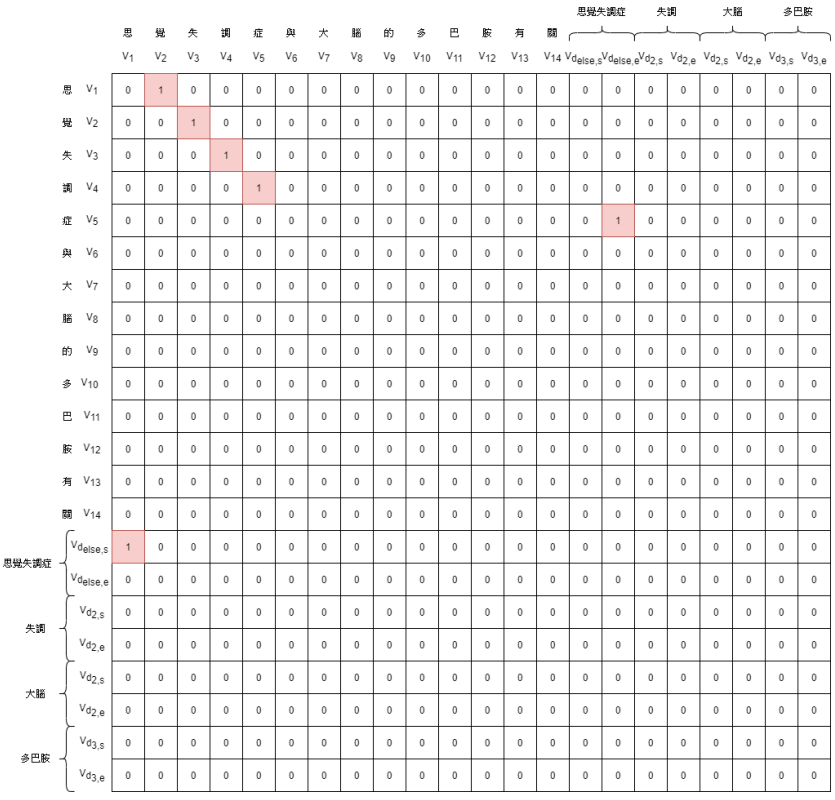


圖 12、詞彙長度為 5 個字以上的有向圖對應的A_{in}矩陣

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，不同字典得到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：

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

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：

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

其中V_c代表的為多重嵌入層 (Multiple Embeddings Layer)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 (Multiple Embeddings Layer)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：

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

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

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

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

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

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

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對應的相鄰矩陣的列向量，公式(11)-(14)為 GRU 單元 [22]，如下圖 13，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。

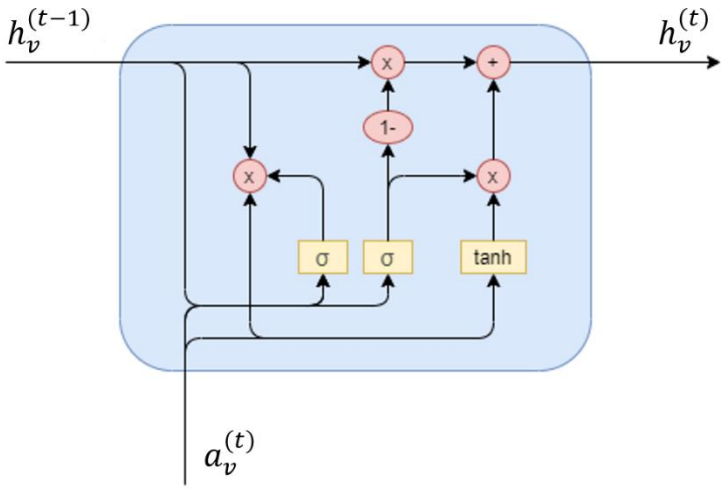


圖 13、門控循環單元(GRU)

3-3 雙向長短期記憶神經網路層：

雙向長短期記憶神經網路 (Bidirectional Long Short-Term Memory, BiLSTM)是由前向 LSTM 與後向 LSTM 組合而成，適合做上下有關係的序列標註任務，因此在 NLP 中常被用來建模上下文資訊。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，整個 BiLSTM 計算過程及架構總共包含以下 6 個公式，以門控圖序列神經網路 (Gated Graph Sequence Neural Networks)層的輸出當做輸入序列。

$$f_t = \sigma(W_f \cdot \left[\acute{h}_{t-1}, x_t\right] + b_f) \tag{15}$$

$$i_t = \sigma\left(W_i \cdot \left[\acute{h}_{t-1}, x_t\right] + b_i\right) \tag{16}$$

$$\tilde{C}_t = \tanh(W_C \cdot \left[\acute{h}_{t-1}, x_t\right] + b_C) \tag{17}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{18}$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{19}$$

$$h_t = o_t * \tanh(C_t) \tag{20}$$

其中x_t為門控圖序列神經網路時刻 t 的輸出，並以其當作 BiLSTM 時刻 t 的輸入，$\acute{h}_{t-1}$為 BiLSTM 前一時刻隱藏層狀態輸出，C_{t-1}為 BiLSTM 前一時刻的細胞狀態，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：

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，如下圖 14，因此條件隨機場域 (Condition Random Fields)取代了 softmax function，成為了當前主流的架構。而目前較為常見的標記格式包含 BIO 格式以及 BIOES 格式，圖 15 為 BIO 標記格式的一範例，在進行實體辨識時，正確的標記序列中標記 O 後面是不會接連著標記 I，因此在本研究的輸出層中採用條件隨機場域 (Condition Random Fields, CRF)做為輸出層，以確保預測的標記是合理的。

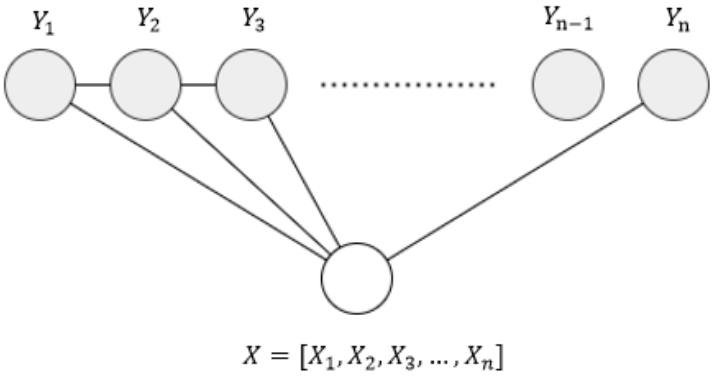


圖 14、CRF 模型示意圖

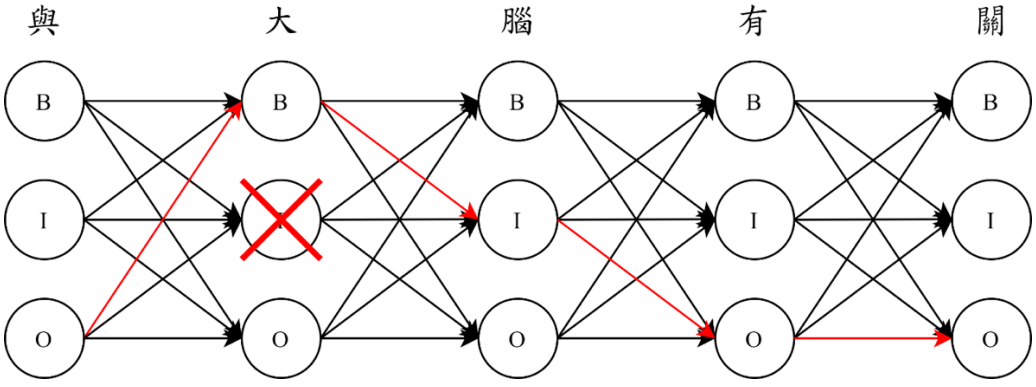


圖 15、BIO 標記格式範例

輸入觀察序列為 X=$(x_1, x_2, \ldots\ldots, x_n)$，輸出標記序列 Y=$(y_1, y_2, \ldots\ldots, y_n)$，透過下列公式(21)可以獲得觀察序列 X 對應的輸出標記序列 Y 的分數：

$$\mathrm{s(x,y)} = \sum_{t=1}^{n+1}(A_{yt-1,yt} + P_{t,yt}) \tag{21}$$

給定觀察序列 X 得到的標記序列 Y 的條件機率如公式(22)：

$$\mathrm{p(y|x)} = \frac{\prod^n e^{s(x,y)}}{\sum_{\tilde{y}\in Y_x}\prod^n e^{s(x,\tilde{y})}} \tag{22}$$

訓練時，我們使用最大似然估計來最大化正確標籤序列的對數概率：

$$\log\big(\mathrm{p(y|x)}\big) = \mathrm{s(x,y)} - \log(\sum_{\tilde{y}=Y_x} e^{s(x,\tilde{y})}) \tag{23}$$

在測試時，模型預測標籤係使用最大後驗概率 (Maximum posteriori probability)，在解碼時採用維特比 (Viterbi)演算法，尋找最大機率的標記序列。

$$y^* = arg\max_{\tilde{y}\in Y_x} p(\tilde{y}|x) \tag{24}$$

第四章　實驗結果

在此章節中首先會說明如何建置健康照護領域的中文命名實體辨識語料庫，實驗設定的小節內容包含所使用的字典、模型的訓練流程以及模型的相關參數，接下來小節會介紹本研究所使用的字嵌入、部首嵌入以及詞嵌入，內容包含訓練的資料來源以及相關設定，在模型的效能評估小節中會介紹目前被大眾所採用的評估方法，並依照此評估方式在下一個節中，將本研究所提出的模型與其他當前的模型進行比較、分析與結果討論，並且探討調整本研究模型細部構造所造成的影響，最後對於模型預測錯誤進行分析。

4-1　語料庫建置

由於目前健康照護領域命名實體語料庫的缺乏，所以本研究透過上網爬蟲的方式來蒐集符合主題的相關文章，並將整理後的資料將給標記人員標註，透過計算 Cohen's Kappa 值以及 Fleiss Kappa 值來確保標記品質，在標註完後統計整個語料庫的資訊，將其切分成訓練集以及測試集。

Cohen's Kappa 值 [23]以及 Fleiss Kappa 值 [24]可以評估問題的一致性，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 Koch [25] 所提出的觀點，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。

為了能夠獲得有關健康照護的命名實體語料庫，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，所爬取的來源一共分為三種，分別為：國家網路醫藥、康健雜誌以及醫聯網。各文章以及問答紀錄的範例如圖 16-18，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健康照護文章，而醫聯網則是一般民眾上網提問，醫生回答的問答紀錄。在文章內容方面，國家網路醫藥以及康健雜誌文章所包含的內容主要為醫療保健，國家網路醫藥的主題涵蓋健康新知、中醫保健、婦幼保健、運

動保健、疾病保健以及銀髮族等等，而康健雜誌的主題涵蓋醫療、養生保健、食物營養、高齡等等，醫聯網的問答不侷限於特定的主題，任何病人的提問都被涵蓋。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 1818 則問答。

圖 16、康健雜誌文章範例

資料來源：https://www.commonhealth.com.tw/

圖 17、國家網路醫藥文章範例

資料來源：https://www.kingnet.com.tw/knNew/index.html

圖 18、醫聯網問答紀錄範例

資料來源：https://med-net.com/

在將資料給標記人員標記以前，已先將所有文章及問答利用 CKIP 斷詞 [26] 系統做斷詞，整個標記資料的流程，我們將其分成階段一、階段二，參與標記的人員一共有三位，都是國立臺灣師範大學的中文系大學生，各階段的內容如下：

(1) 階段一：

在此階段分別取國家網路醫藥 25 篇文章、康健雜 25 篇文章以及醫聯網 100 則問答，三位標記人員分別對這些資料做標記，並對標記結果計算 Cohen's Kappa 值以及 Fleiss Kappa 值。

(2) 階段二：

三位標記人員對階段一的標記結果做討論並且得到一致的標準後，再對另外的國家網路醫藥 25 篇文章、康健雜 25 篇文章以及醫聯網 100 則問答做標記，並對標記結果計算 Cohen's Kappa 值以及 Fleiss Kappa 值確認階段二的 Cohen's Kappa 值以及 Fleiss Kappa 值有上升，且達到可接受的範圍，剩餘的文章以及問答各自請三位標記人員標記。

階段一以及階段二所得到的 Cohen's Kappa 值以及 Fleiss Kappa 分別如表 3，我們可以看到說在經過階段一標註完後的討論，在階段二的一致性有明顯的上升，並且達到 Landis 以及 Koch 所認為的 Almost perfect agreement。

表 3、標記結果一致性 Cohen's Kappa 與 Fleiss Kappa 值

標記人員	階段一	階段二
A vs. B	0.93	0.93
B vs. C	0.72	0.86
A vs. C	0.74	0.88
A vs. B vs. C	0.80	0.89

在健康照護領域中本研究想所關注的命名實體，總共有包含十種，其定義以及例子如下：

(1) 人體 (Body)：

泛指生物體的細胞、組織、器官和系統。例如：細胞核、 神經組織、心、肺、腦、脊髓、呼吸系統、消化系統、泌尿系統等。

(2) 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。例如：流鼻水、頭昏、發燒、咳嗽、失眠、倦怠感、貧血、心悸、耳鳴、胸痛等。

(3) 醫療器材 (Instrument)：

包含診斷、治療、減輕與預防人類疾病，或足以影響人體結構及機能之儀器、器械、附件、配件與零件。例如：血壓計、耳溫槍、達文西機器手臂、內視鏡裝置、人工髖關節、心律調整器、輪椅等。

(4) 檢驗 (Examination)：

利用醫療器材對人體健康狀態及生理功能評估。例如：聽力檢查、心電圖、顯微鏡檢查、核磁共振造影、X 光攝影、電腦斷層掃描等。

(5) 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。例如：去氧核糖核酸、三酸甘油酯、糖化血色素、低密度膽固醇、尿酸、甲狀腺刺激素等。

(6) 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。例如：小兒麻痺症、帕金森氏症、憂鬱症、青光眼、腦溢血、肺結核、胃食道逆流等。

(7) 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。例如：阿斯匹靈、嗎啡、亞硝酸鈉、硫酸鎂、青黴素、亞鐵鹽、流感疫苗、抗生素等。

(8) 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。例如：膠原蛋白、益生菌、綜合維他命、安素、葡勝納、完膳、葡萄糖胺、葉黃素等。

(9) 治療 (Treatment)：

讓患者恢復健康的治癒方式。例如：藥物治療、血漿置換、免疫球蛋白注射、標靶治療、放射線治療、外科手術等。

(10) 時間 (Time)：

描述患者患病症狀的持續時間或是某個時刻。例如：嬰兒期、幼兒時期、青春期、生理期、孕期等。

最後整個語料庫總共為訓練資料 28,161 句以及測試資料 2,531 句，整個資料統計如表 4 和 5 以及圖 19 和 20，從這些統計完的資訊，可以看出訓練及以及測試集的每句平均字數、平均詞數以及平均命名實體個數差異不大，並且在各命名實體佔總命名實體的個數比例也十分相近。在測試資料的選擇上，本研究使用的測試資料為三個標註人員分別標註後，經過討論的標記檔案，其中包括國家網路醫藥文章 50 篇、康健雜誌文章 50 篇以及醫療網問答 100 則。

表 4、訓練資料集統計

訓練資料總句數 28,161		
總字數	總詞數	總命名實體個數
1,392,204	844,517	61,155
平均字數	平均詞數	平均命名實體個數
49.44	29.99	2.17

表 5、測試資料集統計

測試資料總句數 2,531		
總字數	總詞數	總命名實體個數
121,284	72,574	7,305
平均字數	平均詞數	平均命名實體個數
47.92	28.67	2.89

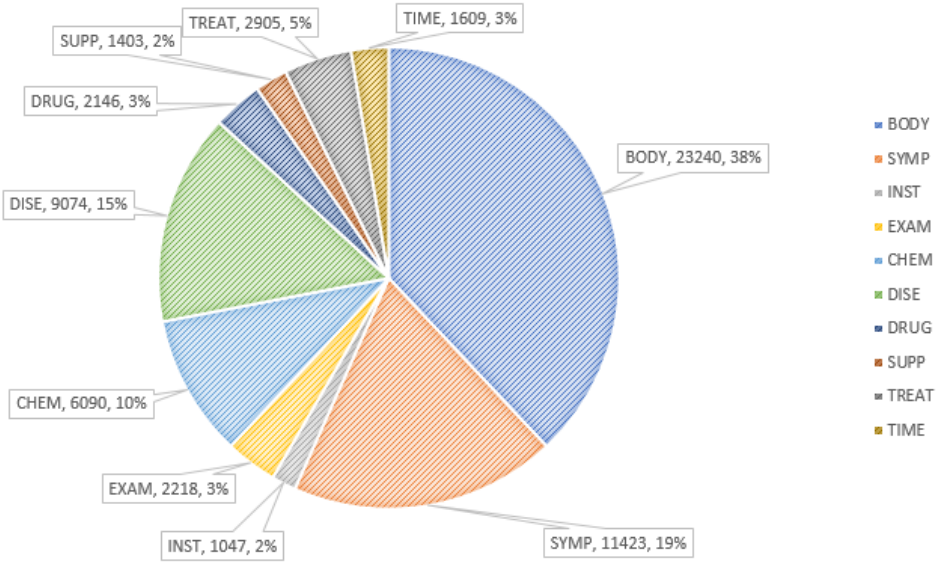


圖 19、訓練資料命名實體類型分佈

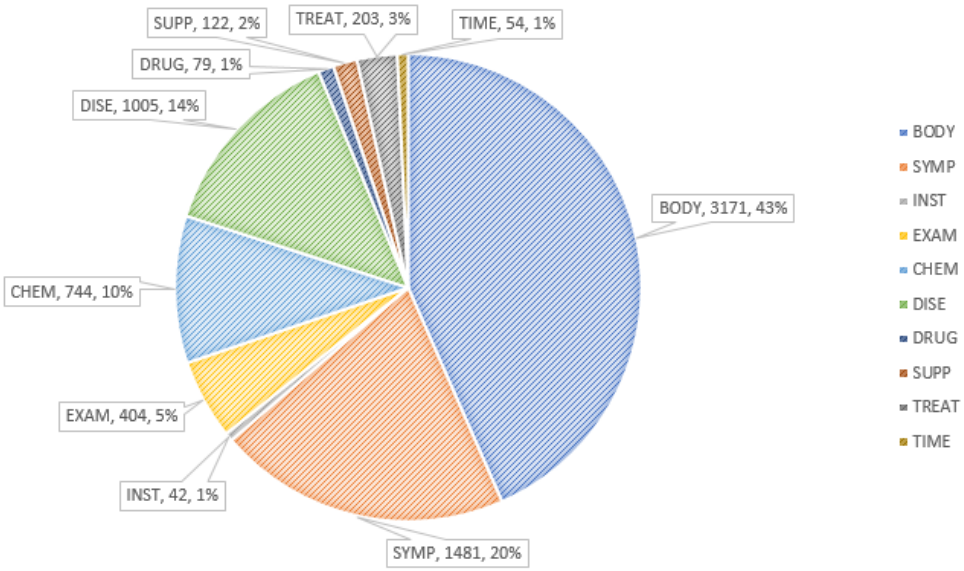


圖 20、測試資料命名實體類型分佈

4-2 實驗設定

在命名實體辨識的任務時，往往可以蒐集到相關的字典，因此如何將字典的資訊融入到模型中，便是會面臨到的問題之一。在本研究中從網路上蒐集了醫療照護相關的字典，並透過 GGSNN 將字典的資訊融入在模型中。

本研究中所使用的字典來源一共分為三個，分別為國家網路醫藥、國家教育研究院和搜狗網，其中國家網路醫藥的詞彙內容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，由於搜狗網的詞彙為簡體字，因此本實驗在使用時透過 OpenCC 將簡體字轉換成繁體字。

在使用字典時，本研究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，各字典的詞彙個數的統計如表 6，其中字典詞彙的數量以詞彙長度為 5 個字以上的最多，詞彙長度為 1 個字的最少。

表 6、字典詞彙數量統計

字典	詞彙數量	範例
詞彙長度為 1 個字	351	耳、鈣、吐
詞彙長度為 2 個字	7,978	紅腫、紫斑、肝癌
詞彙長度為 3 個字	19,282	多汗症、尿蛋白
詞彙長度為 4 個字	31,444	下肢無力、老人痴呆
詞彙長度為 5 個字以上	95,362	子宮鏡檢查、扁桃腺切除

本研究整個流程總共分成以下步驟 (1)資料的蒐集、(2)進入模型前的準備工作、(3)將資料輸入模型、(4) 輸出預測的結果，在資料的蒐集的部分會透過網路爬蟲蒐集且完成標註，在進入模型前會需要先將字嵌入、部首嵌入以及詞嵌入預先訓練完成，並且將後續會使用到的字典蒐集完成，資料以句子為單位進入模型訓練，訓練的參數如下列表

7，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，調整的範例如表 8，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，Dropout rate 的設定為 0.5，批次大小 (Batch size) 的大小設定為 32，GGSNN 中的 time step 設定為 2，time step 的作用為決定 GGSNN 利用 GRU 更新的次數，當設定為 2 時，最後計算出新的節點資訊包含了鄰近的節點以及鄰近節點的鄰近節點，在命名實體辨識的任務中，time step 較合適的大小為 2，如果過大會導致訓練時間拉長以及考慮了太遠節點的資訊，LSTM 的隱藏層大小為 200 維，總共訓練的 epoch 次數為 80。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，因此透過此方法將錯誤的資料在學習一次，以達到希望模型能夠將所有的訓練資料學習正確的目標。

表 7、ME-GGSNN 模型參數值列表

模型參數	**值**
epoch	80
Learning rate	0.001 或 0.0005
Dropout rate	0.5
Batch size	32
time step	2
LSTM hidden	200

表 8、調整 learning rate 以及訓練資料的範例

epoch	**Learning rate**	**訓練資料**
1	0.001	原始整份
2	0.0005	錯誤句子
3	0.001	原始整份
…	…	…

4-3 嵌入向量

在以往要將文字數值化的方式主要是透過 one hot encoding，此種編碼方式為將資料裡所有的字，利用 0 1 做編碼，其中的缺點為並不能表達字原始的意義。而為了能夠將文字做有意義的編碼，因此延伸出許多將文字數值化的技術有，其中較有名的有 Word2vec [21]、Glove [27] 以及 fastText [28] 等等。

本研究所使用的嵌入方式為 Word2vec，透過 Word2vec 可以將文字所隱含的資訊映射至多維空間，並藉由此多維空間中的位置向量代表文字數值化後的數值，獲得向量方法主要分為兩種，分別為 CBOW (Continuous Bag-of-Words) 模型以及 Skip-gram 模型，其中 CBOW 是從上下文字推測當前文字，而 Skip-Gram 正好相反，是從當前文字推測出上下文字。透過 Word2vec 可以訓練出含有語意的空間向量，在向量空間中語意越相近距離會越相近。

在本研究中 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，其中各資料的詳細輸入方式以及訓練參數將會在此小節的接下來有詳細的描述。

首先在字嵌入的部分，本研究會將句子中的每個字分別拆開 (方式如表 9)，將拆開的句子當作輸入，利用 Word2vec 訓練，其中詞頻的設定為至少出現 5 次以上，向量的維度的設定為 50 維，最後訓練出的字嵌入數量為 13,581 個字。

表 9、字嵌入的前處理範例

拆開前	思覺失調症與大腦的多巴胺有關
拆開後	思 覺 失 調 症 與 大 腦 的 多 巴 胺 有 關

由於中文並沒有詞的邊界，因此在詞嵌入的部分本研究使用了 CKIP 斷詞系統，將句子做完斷詞 (斷詞的結果如表 10)，切割出句子中包含的詞後，將這些詞當作輸入，利用 Word2vec 訓練，其中詞頻的設定為至少出現 5 次以上，向量的維度的設定為 50 維，最後訓練出的詞嵌入數量為 863,835 個詞。

表 10、詞嵌入的前處理範例

斷詞前	思覺失調症與大腦的多巴胺有關
斷詞後	思覺失調症 與 大腦 的 多巴胺 有關

如同前面字嵌入，會將句子中的每個字分別拆開，但拆開後的字，在部首嵌入的部分會將其每個拆開後字對應到該字的部首 (對應後的結果如表 11)，將其當作輸入資料，利用 Word2vec 訓練，其中詞頻的設定為至少出現 5 次以上，向量的維度的設定為 50 維，最後訓練出的部首嵌入數量為 3,209 個部件。

表 11、部首嵌入的前處理範例

拆開前	思覺失調症與大腦的多巴胺有關
拆開後	思 覺 失 調 症 與 大 腦 的 多 巴 胺 有 關
對應部首	心 見 大 言 疒 臼 大 肉 白 夕 巴 肉 月 門

4-4 效能評估

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 12，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」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(25)-(27)。

表 12、混淆矩陣

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

$$\text{Precision} = \frac{|TP|}{|TP + FP|} \tag{25}$$

$$\text{Recall} = \frac{|TP|}{|TP + FN|} \tag{26}$$

$$\text{F1-score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{27}$$

4-5 模型比較

在此小節會利用 4-1 節所描述的資料驗證模型效果，將本研究所提出的模型與其他模型透過 4-4 節的評估方法進行比較，下表 13 模型實驗結果的比較。

表 13、命名實體辨識模型實驗結果

Method	Precision	Recall	F1
BiLSTM-CRF [17] (ICCPOL 2016)	70.38	72.77	71.56
BERT Fine-tuning [29] (arXiv 2018)	71.45	76.36	73.82
Lattice [9] (ACL 2018)	74.69	75.76	75.22
Gazetteers [19] (ACL 2019)	73.00	75.56	74.26
ME-CNER [18] (CIKM 2019)	73.68	74.62	74.15
ME-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
- radical - word, Conv⊕BiLSTM → Conv	72.75	72.35	72.55
- radical - word, Conv⊕BiLSTM → BiLSTM	73.74	72.68	73.20

- BiLSTM-CRF：此模型實作了 Dong 等人的架構 [17]，以字作為基礎當作模型輸入，在字嵌入的方面使用的為透過 4-3 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。
- BERT Fine-tuning：此模型為 Devlin 等人所提出 [29]，所使用的預訓練為該官網所下載的 BERT-Base, Chinese，並使用官網的開源程式碼，參照其論文中的描述時做出命名實體辨識的模型。
- Lattice：此模型為 Zhang and Yang 等人所提出 [9]，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。
- Gazetteers：此模型為 Ding 等人所提出 [19]，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，因此本研究使用 4-3 節中所提到的維基百科語料庫當作訓練資料，訓練出各 200 維的向量。

- ME-CNER：此模型為 Xu 等人所提出 [18]，本研究實作的模型架構將其稍做更動，原始的架構如下圖 21，更動後的架構如下圖 22。之所以要更動的原因為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 BiLSTM 的訊息。

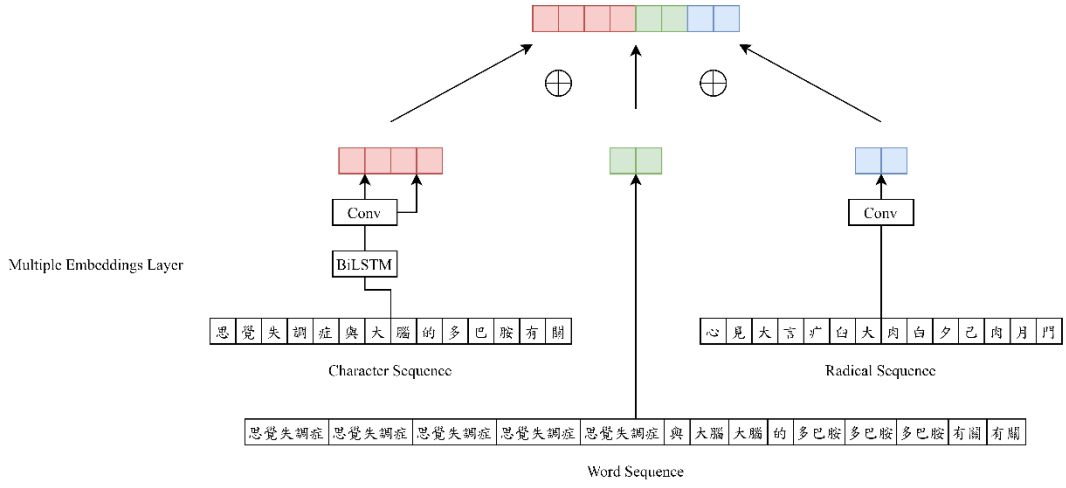


圖 21、原本 ME-CNER 模型的多重嵌入向量架構

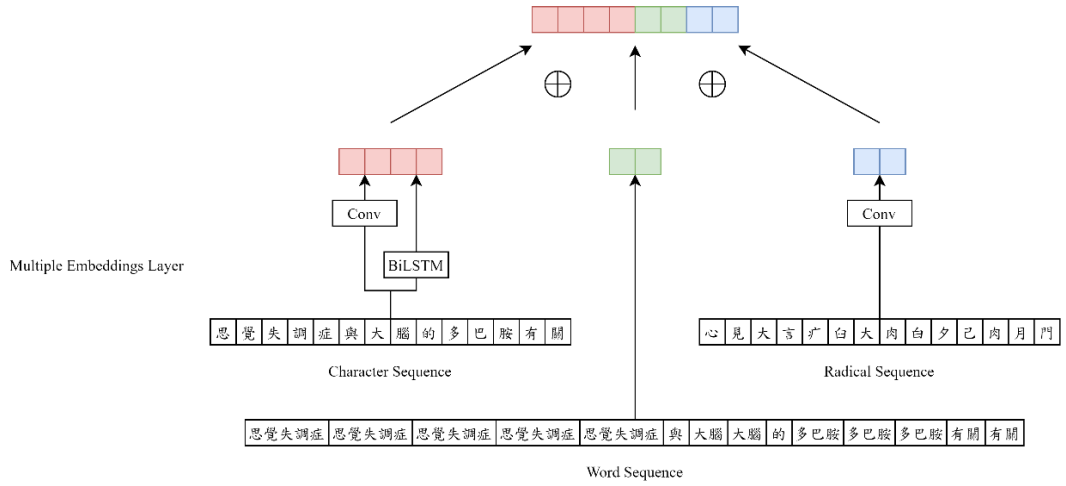


圖 22、修改後 ME-CNER 模型的多重嵌入向量架構

- ME-GGSNN：為本研究提出的模型，在第三章有詳細的介紹。
  (1) - radical：此模型為 ME-GGSNN (ours)去除部首嵌入。
  (2) - word：此模型為 ME-GGSNN (ours)去除詞嵌入。
  (3) - radical - word：此模型為 ME-GGSNN (ours)去除部首嵌入以及詞嵌入。
  (4) - radical - word, Conv⊕BiLSTM → Conv：此模型為 ME-GGSNN (ours)去除部首嵌入以及詞嵌入，在字嵌入的部分只經過 Convolutions。
  (5) - radical - word, Conv⊕BiLSTM → BiLSTM：此模型為 ME-GGSNN (ours)去除部首嵌入以及詞嵌入，在字嵌入的部分只經過 BiLSTM。

由最基礎只包含字嵌入的 BiLSTM-CRF 與 ME-CNER 的比較，可以得知增加除了字嵌入以外特徵是否有幫助，兩者的差異為是否有加入部首嵌入以及詞嵌入，從實驗結果的數據我們可以看出 ME-CNER 相較於 BiLSTM-CRF 提升了 2.59%，因此加入部首嵌入以及詞嵌入有助於提升模型的表現，其可能原因為單純的字嵌入所包含的資訊量並不足夠應付健康照護領域的命名實體辨識，像是本研究所關注的命名實體種類「身體」以及「疾病」，常常會帶有「肉」或是「疒」等部首，透過加入部首資訊可以幫助模型辨識。

由 BiLSTM-CRF 與 Gazetteers 進行比較，可以探討加入字典的資訊是否可以提升模型的表現，兩者主要的差異為是否有加入字典的資訊，從實驗結果的數據得知 Gazetteers 相較於 BiLSTM-CRF 提升了 2.7%，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，其可能的原因為字典包含了需要辨識的命名實體，而透過將字典的資訊納入可以使模型更能將命名實體辨識出。

以前述兩個比較的結論為基礎下，可以得知加入除了字嵌入以外的特徵如部首嵌入以及詞嵌入，以及透過 GGSNN 將字典的資訊納入考慮，皆可以提升模型的表現。因此本研究提出的 ME-GGSNN，同時加入了部首嵌入、詞嵌入以及 GGSNN。

由本研究所提出的 ME-GGSNN 與 Gazetteers 進行比較，兩者的差異為是否有使用多重嵌入，在字典方面使用同為依詞彙長度分類過後的字典，在 Gazetteer 的模型當中，使用的嵌入為字嵌入以及二元嵌入各 200 維，總共 400 維，而本研究的字嵌入、部首嵌入以及詞嵌入組合而成的多重嵌入，其維度總共為 200 維，而本研究的 ME-GGSNN 相較於 Gazetteers 的表現，F1-score 提升了 1.43%。理論上越高維度的嵌入，所包含的資訊

量會越豐富，然而本研究的多重嵌入維度為 200 維，相較於 Gazetteer 所使用的嵌入 400 維少了 200 維，但模型的表現本研究的 ME-GGSNN 卻比 Gazetteer 優秀，因此可以推斷出多重嵌入會是一種較好的嵌入方式，並且較低維度的嵌入在做模型模型訓練時，在硬體方面的要求也相對較低。

由本研究所提出的 ME-GGSNN 與 ME-CNER 進行比較，ME-GGSNN 透過了 GGSNN 將字典的資訊加入，而 ME-CNER 沒有，而兩者皆使用了相同的多重嵌入，因此兩者的處要差異為是否有字典的資訊，由實驗的結果得知，本研究所提出的 ME-GGSNN 較於 Gazetteers 的表現，F1-score 提升了 1.54%，其原因可能為透過 GGSNN 加入字典資訊，能夠有效的幫助模型辨識命名實體。

BERT 為當前非常火紅的架構，其模型特點為使用了 transformer 做特徵抽取，BERT 的整個訓練流程分成兩個階段，分別為 Pre-training 和 Fine-tuning，在 Pre-training 階段時，Google 使用大量文本資料，以非監督式學習的方式訓練模型，本研究使用其官網所提供的 Pre-training，而在 Fine- tuning 階段則是針對不同的任務，使用有標籤的資料訓練並對模型微調。由本研究所提出的 ME-GGSNN 與 BERT Fine-tuning 進行比較， ME-GGSNN 的 F1-score 相較於 BERT 高出了 1.87%，其可能的原因為 BERT 官網的 Pre-training，可能不適合使用在健康照護領域，並且模型只單獨使用了字嵌入。在硬體設備方面，BERT 的所需的要求相較於本研究所提出的 ME-GGSNN 的需較高，在訓練時所佔的 GPU 資源較多。

由本研究所提出的 ME-GGSNN 與 Lattice 進行比較，ME-GGSNN 的 F1-score 相較於 Lattice 高出了 0.47%，兩者的差異包含嵌入的使用、用來比對的字典以及學習字典資訊的結構，在 Lattice 中使用的為字嵌入以及詞嵌入，兩者皆為 50 維，總共 100 維，而本研究所使用的有字嵌入、部首嵌入以及詞嵌入，三者皆為 50 維，組合成多重嵌入後總共 200 維，在用來比對句子中詞彙的字典，Lattice 是透過大量自動取得的字典，將句子中的潛在詞彙找出，而本研究的為使用健康照護領域的相關字典來做句子中的詞彙比對，在學習字典資訊的結構，Lattice 使用的為 Lattice LSTM，而本研究使用的為改良式的 GGSNN。因此本研究的模型 ME-GGSNN 相較於 Lattice LSTM 表現較好的原因可能

為,在嵌入方面,由前述幾個模型的比較中可以得知組合成多重嵌入為較好的嵌入方式。而在使用字典比對時,本研究使用的是健康照護相關的字典,而 Lattice 使用的為不分領域的字典,其包含任何中文可能的詞彙,但健康照護相關領域相關的詞彙並非常見的詞彙,因此在 Lattice 所使用的字典可能沒有包含。在學習字典資訊的結構方面,比對完字典的資訊較類似於圖結構的訊息,因此以圖神經網路學習圖結構的資訊可能為較好的方法。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,在此種情況下,當資料量越大時,其餘平常的模型能夠隨之調整 Batch size 以便加快模型訓練速度,但 Lattice 卻無法達到此效果。

接下來為不同組合方式的多重嵌入比較,由 ME-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。

除了比較有無部首嵌入以及詞嵌入之外,在字嵌入的部分也做了不同的實驗比較,在都只有字嵌入的基礎下,將字嵌入只經過 BiLSTM 抽特徵、只經過 Convolutions 抽特徵以及同時經過 BiLSTM 抽特徵以及 Convolutions 抽特徵,其中以同時經過 BiLSTM 抽特徵以及 Convolutions 抽特徵的表現最好。只經過 BiLSTM 的 F1-score 為 73.20%,而只經過 Convolutions 的 F1-score 為 72.55%,相差了 0.65%,其可能的原因為 BiLSTM 會捕捉長距離的資訊,而 Convolutions 會捕捉短距離的資訊,但由於中文字的特性,長距離的資訊較為重要,因此只經過 BiLSTM 的表現較佳,而同時經過 BiLSTM 以及 Convolutions,可以同時捕捉長距離以及短距離的資訊,比起只單獨經過其中一個的資訊量更為豐富,模型的 F1-score 達到了 74.00%,相較於只經過 BiLSTM 抽特徵、只經過 Convolutions 抽特徵,模型的 F1-score 分別上升了 0.8%以及 1.45%。

表 14 為 ME-GGSNN 模型預測各類命名實體的 Precision、Recall 以及 F1-score,其中以「疾病」的 F1-score 為最高,其次依序是「人體」、「化學物質」以及「檢驗」,此 4

種的 F1-score 高於整體的 F1-score，剩下的 6 種命名實體類別低於整體的 F1-score，分別為「藥品」、「醫療器材」、「營養品」、「症狀」、「時間」以及「治療」，其中以「醫療器材」的 F1-score 最低，該種類的命名實體在訓練資料所佔的數量也最少，因此之所以該種類的 F1-score 不高的可能原因為在訓練資料數量不多，其餘低於整體 F1-score 的命名實體種類除了「症狀」以外，在訓練資料的數量也普遍偏低，而高於整體 F1-score 的命名實體種類「檢驗」與「症狀」相反，雖然在訓練資料的數量不多，但辨識效果相比整體的 F1-score 卻比較好。

表 14、ME-GGSNN 模型各類命名實體辨識結果

	ME-GGSNN		
	Pr.	**Re.**	**F1**
人體	75.41	78.81	77.07
化學物質	72.79	80.91	76.64
疾病	81.16	84.88	82.98
藥品	78.69	60.76	68.57
檢驗	69.73	84.41	76.37
醫療器材	62.50	35.71	45.45
營養品	67.67	73.77	70.59
症狀	74.79	67.52	70.97
時間	64.81	64.81	64.81
治療	53.23	68.97	60.09
Total	74.45	76.97	75.69

4-6　效能分析

接著對於本研究所提出的 ME-GGSNN 模型進行更進一步的分析與討論，分別從以句子為單位，和以命名實體為單位這兩個角度出發，討論加入字典資訊以及各字典對於模型的影響。

(1) 以句子為單位：

表 15 的實驗結果為是否有透過 GGSNN 加入字典資訊的比較，而不是單純的只使用 GGSNN，其中不使用字典 (without dict)為只使用相鄰矩陣A_c，而使用字典 (with dict)使用的相鄰矩陣為A_c、A_{d_1}、A_{d_2}、A_{d_3}、A_{d_4}以及$A_{d_{else}}$，透過相鄰矩陣A_{d_1}、A_{d_2}、A_{d_3}、A_{d_4}以及$A_{d_{else}}$可以將字典的資訊納入考慮，以句子「思覺失調症與大腦的多巴胺有關」不使用字典以及使用字典的多維有向圖範例如圖 23。其中在表格上方左邊的 Train 代表的意思為命名實體是否有出現在訓練資料，其中 All、Some 以及 None 代表的意思分別為，該句子中的命名實體「全」在訓練資料裡、該句子中的命名實體「部分」在訓練資料裡以及句子中的命名實體「不」在訓練資料裡，。

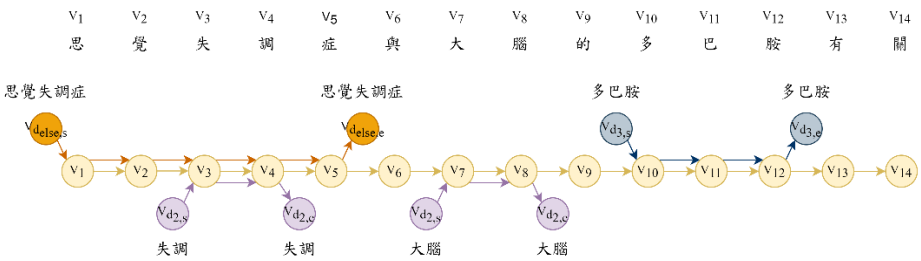
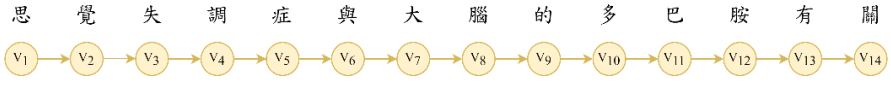
A_c、A_{d_1}、A_{d_2} A_{d_3}、A_{d_4}、$A_{d_{else}}$ (with dict)	
A_c (without dict)	

圖 23、不使用字典和使用字典的多維有向圖

由表 16 可以看出不論是命名實體全在訓練資料 (All)、命名實體部分在訓練資料 (Some) 或是命名實體不在訓練資料 (None) 皆為使用字典 (with dict) 的表現較佳，特別是當命名實體不在訓練資料時的效果特別顯著 F1-score 上升 4.06%，命名實體部分在

訓練資料 (Some) 次之為 2.18%，命名實體全在訓練資料 (All) 上升最少為 0.94%。因此可以得知字典對於不在訓練資料的命名實體有重大幫助。

由於句子中的命名實體「不」在訓練資料裡的 F1-score 上升最顯著，因此表 11 為對該情況做更細部討論的實驗結果，當命名實體不在訓練資料時，理論上模型較難將其判斷正確，但由表 14 的實驗結果，可以看出當命名實體不在訓練資料但全在或是部分在字典時，相較於不使用字典，F1-score 有明顯的上升。

表 15、由訓練資料涵蓋程度探討字典的影響

Entities appear in the training data	With dictionaries			Without dictionaries		
	Pr	Re	F1	Pr	Re	F1
All	79.56	82.00	**80.76**	79.18	80.47	79.82
Some	74.43	72.67	**73.54**	72.72	70.04	71.36
None	72.04	70.53	**70.73**	69.32	64.21	66.67

表 16、由字典詞彙涵蓋程度探討字典的影響

Entities appear in the dictionaries	With dictionaries			Without dictionaries		
	Pr	Re	F1	Pr	Re	F1
All	76.74	82.50	**79.52**	71.11	80.00	75.29
Some	76.47	56.52	**65.00**	60.00	39.13	47.37
None	63.64	65.62	64.62	71.43	62.50	**66.67**

(2) 以命名實體為單位：

在使用字典時，本研究將字典分成 (1)詞彙長度為 1 個字 (2)詞彙長度為 2 個字 (3)詞彙長度為 3 個字 (4)詞彙長度為 4 個字 (5)詞彙長度為 5 個字以上，因此下表 17 主要探討的目標為是否上述 5 個字典皆對於模型有正面的幫助。

實驗結果如表 17，第 1 列 All 為所有字典全用，即相鄰矩陣A_c、A_{d_1}、A_{d_2}、A_{d_3}、A_{d_4}以及$A_{d_{else}}$皆考慮，第 2-6 列為加入 5 種字典中的其中一個，即為相鄰矩陣A_c加上A_{d_1}、A_{d_2}、A_{d_3}、A_{d_4}以及$A_{d_{else}}$中的其中一個，而第 7 列為不使用字典，單獨使用相鄰矩陣A_c，以句子「思覺失調症與大腦的多巴胺有關」所有字典全用 (All)、使用A_{d_1}字典以及不使

用字典的有向圖範例如圖 24。在此實驗結果的表格中，主要關注的項目為粗黑體字的部分，我們可以看到說，除了第 6 列的實驗結果以外，當加入 5 種字典中的其中一個，相較於第 7 列不使用字典的 F1-score 皆有上升，而第 1 列使用所有字典的 F1-score，相較於使用 5 種字典中的其中一個，單獨比較加入某個長度的命實體時，雖未皆比較高，但整體模型的效果 (overall)，以第 1 列的表現為最佳。

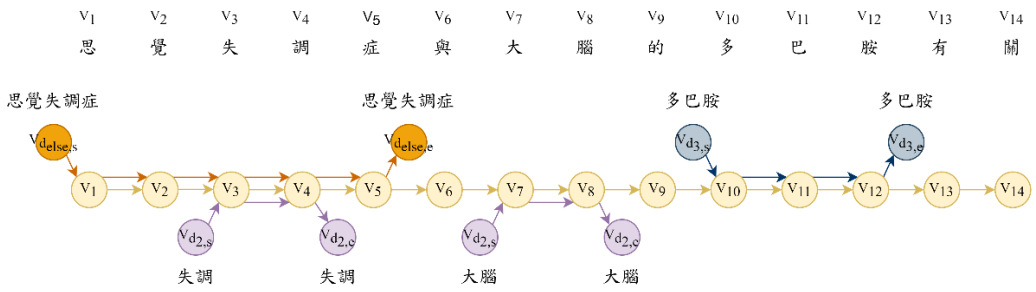
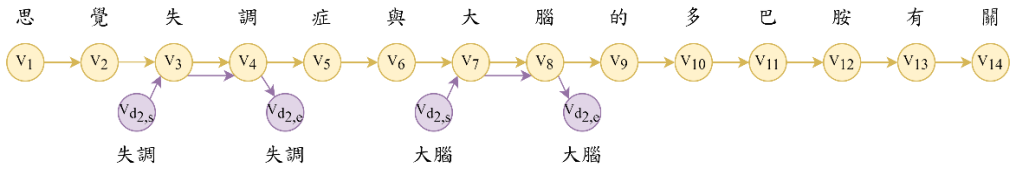
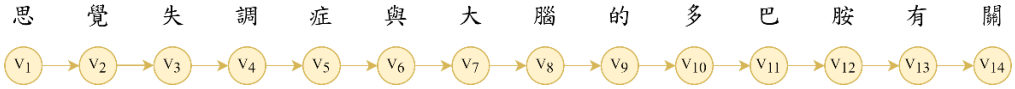
All	
A_c+A_{d_2}	
A_c	

圖 24、所有字典全用、使用A_{d_1}字典以及不使用字典的有向圖

表 17、字典組合對不同詞彙長度的命名實體辨識結果

Dict.	one character			two character			three character			four character			else character			Overall		
	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1
All	69.92	49.70	**58.10**	74.10	83.74	**78.62**	78.17	80.87	**79.49**	71.58	60.50	**65.58**	73.31	68.04	**70.58**	**75.46**	75.76	**75.69**
A_c+A_{d_1}	71.82	52.48	**60.64**	73.67	81.54	77.40	78.86	79.45	79.15	72.19	58.92	64.88	74.52	65.29	69.60	74.57	75.29	74.93
A_c+A_{d_2}	70.87	50.10	58.70	73.61	82.92	**77.99**	78.25	79.31	78.78	73.49	60.76	66.52	74.83	66.82	70.60	74.49	76.17	75.32
A_c+A_{d_3}	74.07	51.49	60.75	74.07	82.07	77.87	79.24	80.80	**80.01**	69.21	59.58	64.03	70.83	64.98	67.78	74.43	75.82	75.12
A_c+A_{d_4}	72.89	49.50	58.96	74.63	80.92	77.65	79.37	77.76	78.55	72.12	59.06	**64.94**	73.74	66.97	70.19	75.21	74.58	74.89
A_c+$A_{d_{else}}$	65.00	51.49	57.46	71.57	84.64	77.55	77.18	80.73	78.92	72.33	60.37	65.81	70.08	63.76	**66.77**	72.29	**77.15**	74.64
A_c	65.49	47.72	**55.21**	74.03	81.18	**77.44**	78.23	77.01	**77.61**	70.02	57.61	**63.21**	71.74	65.60	**68.53**	73.91	74.17	74.04

4-7　錯誤分析

本研究將命名實體的錯誤分成以下 5 種，並且統計各錯誤的數量，數量的分布如圖 25，命名實體預測錯誤的範例如表 18：

- 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。

表 18、命名實體辨識預測錯誤範例

CONTAIN	答案	國際間 德國麻疹$_{DISE}$ 仍有疫情發生。
	預測	國際間 德國麻疹$_{DISE}$ 仍有疫情發生。
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。

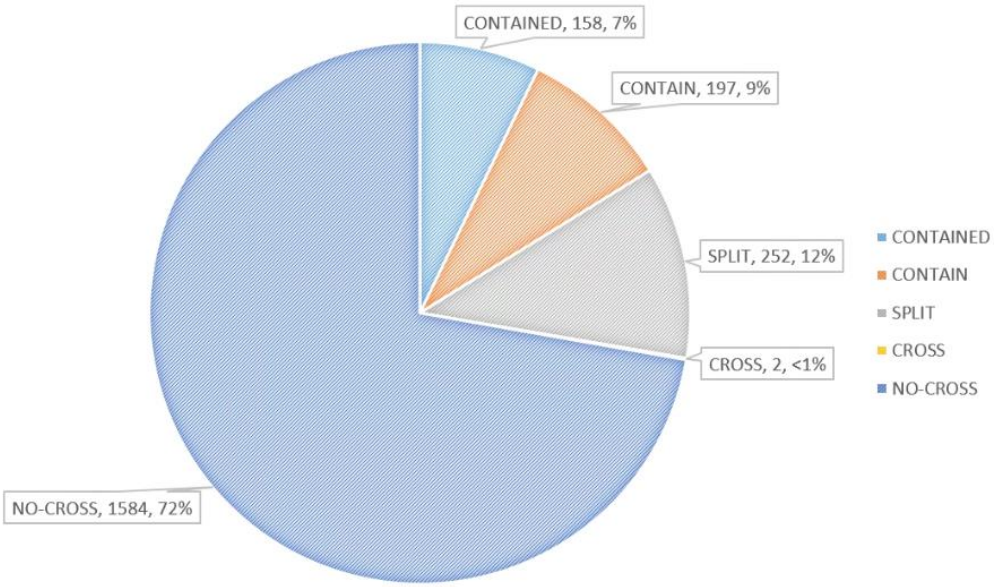


圖 25、命名實體辨識錯誤類型分佈

從圖 25 可以看出，最多種類錯誤的為 NO-CROSS，即為正確的命名實體與預測的命名實體之間「沒有」重疊的字，而最少的為正確的命名實與預測的命名實體之間「有」重疊的字 (CROSS)，該種類的錯誤所佔的比例非常的低，錯誤種類中的 CONTAIN、CONTAINED 以及 SPLIT 三者相加所佔的比例約為 28%左右，此三種錯誤命名實體預測結果與正確答案的差異最小，例如像是 CONTAIN 例子中，正確答案為「德國麻疹」，而模型辨識出的結果為「麻疹」，雖然並非完全正確，但與原本的正確答案相距不遠，因此某方面來說並非完全的辨識錯誤。

第五章　結論與未來工作

本研究的貢獻分別有以下兩點：

一、建立了一個健康照護相關的中文命名實體辨識語料庫，總共的句子數為 30,692 句，總共的詞數為 917,091 個，總共的字數為 1,513,488 個，總命名實體的個數為 68,460 個，其中所關注的命名實體包含：(1) 身體 (2) 疾病 (3) 症狀 (4) 化學物質 (5) 藥物 (6) 營養品 (7) 醫療設備 (8) 檢驗 (9) 治療方式 (10) 時間。

二、本研究的模型在此資料集上的表現，相較於其他當前的模型更為優秀。我們提出的 ME-GGSNN 模型達到 F1-score 75.69%，高於目前較為人所知的模型 (BERT、Lattice、Gazetteers、ME-CNER)，此模型以字基礎的序列作為輸入，透過加入詞嵌入以及部首嵌入的特徵組合成多重嵌入，並且透過 GGSNN 同時引入字典的資訊。藉由多重嵌入可以使的原本單獨的字嵌入，更能夠強化其所代表的字特徵，而字典資訊的引入，可以將原本已知命名實體訊息納入模型考慮。

透過本研究模型 ME-GGSNN 的表現以及分析，可以得知將字嵌入融入詞嵌入以及部首嵌入的資訊，經過 BiLSTM 以及 Convolutions 處理組合成多重嵌入，均能夠提升模型的表現。而透過整理過的字典，從以句子為單位的角度或是以命名實體的角度分析，皆能夠顯示字典對於模型的效能有所幫助。並且透過將字典依字數分類，使得較能夠不受硬體限制的加入字典來源。

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，我們可以更中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、問答系統等應用的基礎。

參考文獻

[1] Lawrence R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77 (2), p. 257 – 286, February 1989.

[2] Toutanova, Kristina; Manning, Christopher D., Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. Proc. J. SIGDAT Conf. on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-2000). pp. 63 – 70.

[3] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of 18th International Conference on Machine Learning, ICML 2001, pp. 282 – 289 (2001).

[4] Krizhevsky, A., Sutskever, I., & Hinton, G., (2012). ImageNet classification with deep convolutional neural networks. In NIPS.

[5] Williams, Ronald J.; Hinton, Geoffrey E.; Rumelhart, David E., (October 1986). "Learning representations by back-propagating errors". Nature. 323 (6088): 533 – 536.

[6] Hochreiter, S., Schmidhuber, J., Long short-term memory. Neural Comput. 9, 1735 – 1780 (1997).

[7] Levow, G.A., The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In: Computational Linguistics, pp.

[8] Nanyun Peng and Mark Dredze, 015. Named entity recognition for Chinese social media with jointly trained embeddings. In EMNLP. pages 548 – 554.

[9] Zhang, Y. and Yang, J., (2018). Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(ACL’ 18),Long Papers, pages 1554-1564.

[10] Xianpei Han, Overview of the CCKS 2019 Knowledge Graph Evaluation Track: Entity, Relation, Event and QA (2019). arXiv.

[11] Fu, G., Luke, K.K., Chinese named entity recognition using lexicalized HMMs. ACM SIGKDD Explor. Newsl. 7, 19 – 25 (2005).

[12] Gideon S. Mann and Andrew McCallum., 2010. Generalized Expectation Criteria for SemiSupervised Learning with Weakly Labeled Data. J. Mach. Learn. Res. 11 (March 2010), 955 – 984.

[13] Duan, H., Zheng, Y., A study on features of the CRFs-based Chinese. Int. J. Adv. Intell. 3, 287 – 294 (2011).

[14] Han, A.L.-F., Wong, D.F., Chao, L.S., Chinese named entity recognition with conditional random fields in the light of Chinese characteristics. In: Kłopotek, M.A., Koronacki, J., Marciniak, M., Mykowiecka, A., Wierzchoń, S.T. (eds.) IIS 2013. LNCS, vol. 7912, pp. 57 – 68. Springer, Heidelberg (2013).

[15] Huang, Z., Xu, W., Yu, K., Bidirectional LSTM-CRF models for sequence tagging (2015). arXiv.

[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer (2016).Neural architectures for named entity recognition.In Proceedings of the NAACL'16, pp. 108-117

[17] Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di., 2016. Character based LSTM-CRF with radical-level features for Chinese named entity recognition. In International Conference on Computer Processing of Oriental Languages. Springer, pages 239 – 250.

[18] Canwen Xu, Feiyang Wang, Jialong Han, and Chenliang Li, Exploiting multiple embeddings for chinese named entity recognition. In CIKM, pages 2269–2272. ACM, 2019.

[19] Ruixue Ding, Pengjun Xie, Xiaoyan Zhang, Wei Lu, Linlin Li, and Luo Si., 2019. A neural multidigraph model for chinese ner with gazetteers. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1462–1467.

[20] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel, 2016. Gated graph sequence neural networks. In Proc. of ICLR.

[21] Mikolov, T., Chen, K., Corrado, G., & Dean, J., (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[22] Cho, K. et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proc. Conference on Empirical Methods in Natural Language Processing 1724–1734 (2014).

[23] Cohen, Jacob, (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement. 20 (1): 37–46.

[24] Fleiss, J. L., (1971) "Measuring nominal scale agreement among many raters." Psychological Bulletin, Vol. 76, No. 5 pp. 378–382.

[25] Landis, J. R. and Koch, G. G., "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, pp. 159–174.

[26] Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532 – 1543.

[28] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017. Enriching word vectors with subword information. TACL 5:135 – 146.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova., BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short apers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.