一個結合文本內容與標籤語意模型的三階段 NER 模型
ISLab at ROCLING 2023 MultiNER-Health Task: A Three-Stage NER Model Combining Textual Content and Label Semantics

吳俊傑 Jun-Jie Wu　　張道行 Tao-Hsing Chang
國立高雄科技大學
資訊工程系
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
{c109151150, changth}@nkust.edu.tw

許福元 Fu-Yuan Hsu
國立臺灣師範大學
學習科學跨國頂尖研究中心
心理與教育測驗研究發展中心
Institute for Research Excellence in Learning Sciences,
Research Center for Psychological and Educational Testing
National Taiwan Normal University
Taipei, Taiwan, R.O.C
kevinhsu@ntnu.edu.tw

摘要

本次任務為醫療保健領域的命名實體識別。這些被辨識出來的命名實體需要進一步被歸類在十種類別之一。本研究提出一個同時結合文字內容與標籤語意的三階段模型。模型的第一階段是使用 Label Semantics Model 辨識與分類命名實體。第二階段是使用 Label Correction Model 對異常標籤進行更正。第三階段是以一個規則式方法將第二階段無法處理的異常標籤進行更正。本研究所提方法以 ROCLING 2023 MultiNER-Health Task 所提供的三個測試集進行評估，平均 F_1score 為 0.6838，並且在所有參與任務模型中是對 SM 測試集表現最佳的模型。

Abstract

ROCLING 2023 MultiNER-Health Task is to identify named entities in the healthcare domain. These recognized named entities need to be further classified into one of ten categories. This study proposes a three-stage framework. The first stage of the framework is to recognize and categorize named entities using Label Semantics Model. This model employs both of textual content and label semantics to identify name entities and their categories. In the second stage, the Label Correction Model is used to correct abnormal labels. In the third stage, a rule-based approach is employed to revise the abnormal labels that cannot be corrected in the second stage. The proposed method is evaluated on three test sets provided by the task and its average F_1score is 0.6838. It is the best performing model for SM test set among all participating task models.

關鍵字：命名實體識別、標籤語意、語意模型、校正模型

Keywords: NER, Label semantics, Semantic model, Correction model

1 緒論

在自然語言處理中，命名實體(Named Entity, NE)指的是文本中的基本資訊片段或者是專有名詞，常見的類別有人名、地名、日期、組織等等。而命名實體辨識(NER)指的是對文本

中的命名實體進行辨識，例如「孫中山是國父」即可辨識出「孫中山」為人名。辨識結果可以應用至問答系統、文本分類、自動摘要等等其他任務中，所以 NER 的正確率將會影響後續應用的正確性。

本研究參與的 ROCLING 2023 Shared Task I(以下簡稱本次任務)是延伸 Lee, Chen et al. (2022)提出 ROCLING 2022 Shared Task(以下簡稱 ROCLING 2022)的中文 NER 任務。在本次任務中，參與者需要設計一個模型判斷文本中是否含有醫療保健領域的 NE。若含有 NE 則需將該 NE 的每個字元標記實體類別。

表 1 列出本次任務所標記的 10 種類別。此外，模型還需再將 NE 的第一個字的標記前面加上「B-」型態，表示此為 NE 的開始(begin)，後續字則加上「I-」型態，表示此為 NE 的內容(inside)。最後，不屬於 NE 的字元都標記為 O(other)型態。例如「耳鳴很吵」中的「耳鳴」屬於 DISE 類別，則需將「耳」標記為「B-DISE」，將「鳴」標記為「I-DISE」，而「很」與「吵」標記為「O」。一個字元的型態與類別形成的組合(例如 B-DISE)，本研究稱為標籤。由於型態 O 沒有類別，因此不屬於 NE 的字元其標籤即為「O」。

類別	標記	範例
Body	BODY	細胞核
Symptom	SYMP	流鼻水
Instrument	INST	血壓計
Examination	EXAM	腦電波圖
Chemical	CHEM	膽固醇
Disease	DISE	肺結核
Drug	DRUG	普拿疼
Supplement	SUPP	維他命
Treatment	TREAT	標靶治療
Time	TIME	生理期

表 1. 十種 NE 類別的範例

本次任務使用的資料集以 ROCLING 2022 提供的為基礎，另外新增了一個資料集，並且將所有資料集依來源分成以下三種類別。第一種是 formal texts(FT)，此類別包括專業編輯或記者撰寫的文章；第二種是 social media(SM)，此類別包括論壇上民眾的問答；第三種是 Wikipedia articles(WA)，此類別為維基百科的文章。

本次任務有以下兩個資料集(Lee et al., 2023)，第一個資料集 Chinese HealthNER Corpus (Lee & Lu, 2021)分為 FT 以及 SM 兩種類別，FT 含有 23,008 個句子、1,109,918 個字元以及 42,070 個 NE；SM 含有 7,684 個句子、403,570 個字元以及 26,390 個 NE。第二個資料集 CHNER (Lee, Chen et al., 2022)為 WA 類別，含有 3,205 個句子、118,116 個字元以及 13,369 個 NE。最後會使用三種資料各至少兩千句的文本對模型進行評估，評估以三種資料集的平均 F1 分數為最終數值。

表 2 列出前述兩個資料集在各類別的實體數量總和。表 2 的數據顯示數量最高與最低的類別在整體資料中的所佔比例落差達 36.69%，且在資料量少的類別中，某些 NE 的數量也寥寥無幾。此現象容易造成模型對特定類別的標記效果不佳。因此，如何充分應用資料集的各種訊息是提升模型標記正確率的方法之一。

類別	數量	比例
Body	31,719	38.78%
Symptom	14,848	18.15%
Instrument	1,339	1.46%
Examination	2,829	3.55%
Chemical	8,552	10.46%
Disease	12,688	15.52%
Drug	2,706	3.32%
Supplement	1,708	2.09%
Treatment	3,574	4.37%
Time	1,857	2.30%

表 2. 各類別的 NE 總和數量與比例

我們認為目前許多 NER 方法只利用文本內容的語意進行標記，而標籤僅被用來分類的符號，沒有特別的意義。然而，我們認為標籤、特別是類別的語意也能為模型帶來更多的訊息。因此，本研究所提方法以 Ma et al. (2022)提出的一種同時使用文本內容以及標籤語意的模型作為基礎模型。另外，Lin et al. (2022)指出模型預測的標籤有形式上不正確的問題(以下簡稱此類標籤為異常標籤)。我們認為異常標籤可視為自然語言中的別字問題，因此本研究所提模型利用別字校正的概念，在本研究所提方法中加入了校正模型來處理異常標籤的問題。

2 相關研究

NER 最早期的方法為基於人工撰寫規則式的線性模型。之後因為有了可以相同條件比較不同方法與模型的標記資料集，促進了監督式學習模型的發展，例如條件隨機場(CRF)(Lafferty et al., 2001)曾經是效果最好的模型之一。之後因為標記資料數量的限制，使模型有時會難以學習到文本語意，此時能夠使用大量未標記資料來對上下文進行學習、生成豐富特徵的無監督式學習模型也相繼問世(Roy, 2021)。

近年來，NER 的研究持續發展，各項研究所提出的方法推陳出新，效能越來越好。例如由 Huang et al. (2015)提出的 LSTM-CRF 架構至今仍然常被用於處理 NER 問題，像是由 Lu & Lee. (2020)提出的門控圖序列神經網路(GGSNN)模型就基於該架構進行研究。該研究針對中文 NER 會因為斷詞而大幅影響結果的特性，組合了字、詞、部首作為多重嵌入層，並且對原始的 GGSNN (Li et al., 2015)加入字典訊息進行改良。最後將 GGSNN 的結果再輸入至 BiLSTM-CRF 進行序列標記，來取得最終的標記結果。該研究利用網路爬蟲的方式收集了醫療保健領域的句子並進行人工標記，最後此架構的性能表現比傳統的 BiLSTM-CRF 或是 MECNER 都來得更好。

ROCLING 2022 的 Shared Task (Lee, Chen et al., 2022) 是一項對醫療保健領域的中文 NER 問題任務，有許多研究提出各種方法處理此任務。例如 Lin et al. (2022)利用 BERT (Devlin et al., 2019)進行以下三種方法：第一種方法為利用 BERT 輸出語意向量，各類別的實體在向量空間會彼此接近，藉此學習到各類別會在語意空間哪些範圍，之後只要計算出輸入的語意向量，如果較接近某個類別，就可以判斷該詞屬於此類別；第二種方法為一個兩階段模型，第一階段會先判斷文本是否可能為 NE，第二階段會對可能為 NE 的文本使用上述第一種方法進一步判斷出該詞屬於哪個類別；第三種方法結合了前兩種方法提出的模型以及詞典模型，該詞典模型會收錄訓練集中曾被標記為 NE 的字元，並計算此字元以不同實體類別出現過幾次，建立起每個 NE 對於不同類別的機率分布，最後將三個模型的結果輸入至一個全連接模型，並輸出字元對每個類別的機率值，再經過 softmax 取得最終標籤結果。

另外，ROCLING 2022 的有些研究(Lin et al., 2022)有發現模型產生的結果會出現形式錯誤的標記結果，並且使用了後處理進行修正，基於此發現，本研究所提方法也採用校正模型讓模型自動學習如何修正此類錯誤。

3 本研究所提方法

本研究設計了一個新的模型來處理本次任務。圖 1 為該模型的架構圖，由三個部分串接組成。第一部分是 Ma et al. (2022)所提出的模型，以下簡稱為 Label Semantics Model (LSM)。這個模型的目的為初步產生預測標籤，也就是將文本輸入至 LSM 後，該模型將會產生文本中每個字相對應的標籤。第二部分為校正模型，以下簡稱為 Label Correction Model (LCM)，這個模型的目的是對 LSM 產出的標籤中不合理的結果進行校正。該模型首先從 LSM 的結果中過濾出含有異常標籤之句子，再重新輸出更為合理的標籤，以達到校正的目的。由於校正結果仍可能有漏網之魚，因此最後一個部分還會以一個規則式模組再次檢查結果是否出現異常標籤，並且以規則式方法重新標記成邏輯上合理之標籤。各部分模型具體內容在以下各小節進行說明。

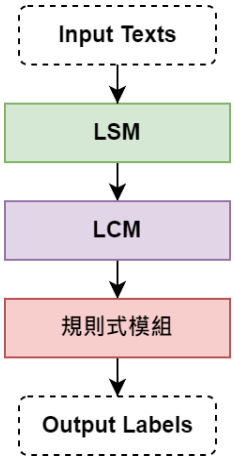

圖 1. 本研究所提模型架構圖

3.1 Label Semantics Model

Ma et al. (2022)認為在 NER 任務中，標籤的語意訊息也能夠為訓練提供更多資訊，以達到在使用同樣的訓練資料量時，模型能夠更精確地標記詞彙。圖 2 為 Ma et al. (2022)所提模型的架構圖，本研究也採用這個模型作為第一階段的 NER 模型。

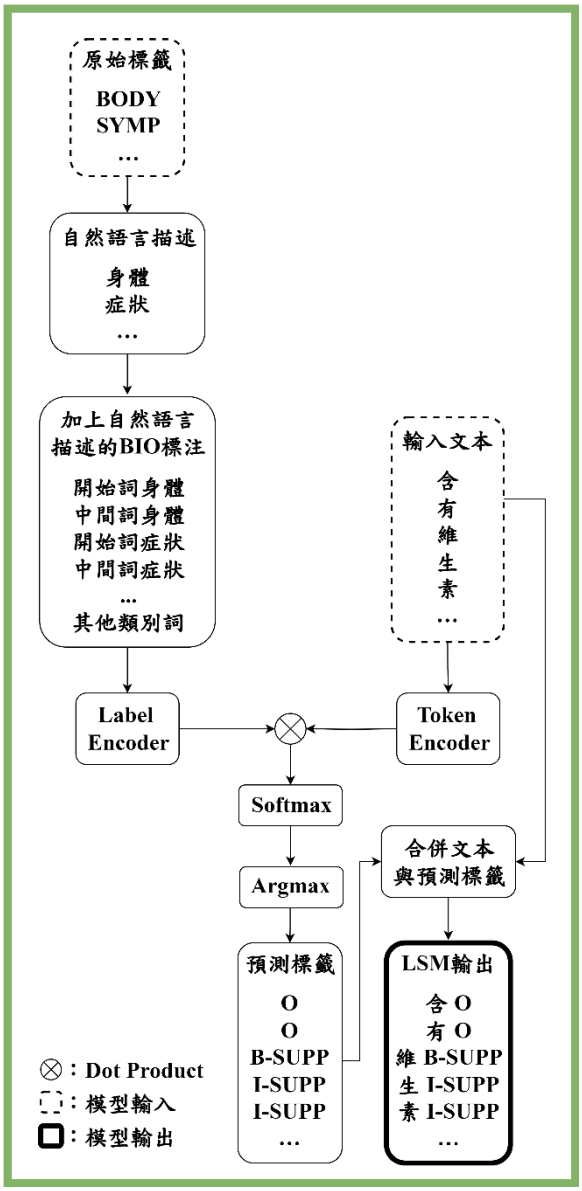

圖 2. LSM 架構圖

LSM 使用了兩個用以進行詞嵌入(word embedding)工作的編碼器(encoder)，一個用以處理原始文本中的字彙語意，一個用以處理標籤語意。第一個編碼器將原始文本的每個字彙由 Token Encoder 產生其表徵(representation)，也就是語意向量。第二個編碼器是將每一種標籤轉換為自然語言描述的文本，接著這個文本輸入至 Label Encoder 取得這個文本對每一個標籤的表徵，再將所有標籤表徵組合成一個矩陣。舉例來說，若是有 21 種不同的標籤，而每個表徵為 768 維的向量，就會有一個 21X768 維的標籤表徵矩陣。

接著 LSM 會以內積計算一個字彙的文本表徵和標籤表徵矩陣，可以得到一個表示這個字彙與每種標籤間關聯的向量。最後透過 softmax 以及 argmax 預測該字彙最可能的標籤。在訓練過程中，預測結果與真實結果的差異會被用來微調兩個編碼器。在這個設計中，編碼器有許多模型可以選擇。基於 4.2 節的實驗，我們提交的系統是採用 RoBERTa 模型(Liu et al., 2019)。

在 LSM 中，如何將標籤轉換為表徵是重要的關鍵，而標籤如何轉換為自然語言描述的文本會影響轉換的結果。Ma et al. (2022)的作法是將標籤類別直接轉換為單詞後，在單詞前再加上自然語言描述的 BI 型態。對於不屬於任何類別時則描述為 other。例如「SYMP」會直接轉換為詞彙「symptom」，並且在前面加上 begin 或 inside，最後形成「begin symptom」和「inside symptom」兩種標籤。但我們認為若將類別轉換為對類別更具體且深入的描述，應該能使模型更精確地生成標籤的語意向量。

被描述類別	**BODY**
描述方法一	身體
描述方法二	構成人或動物的整個物理結構，包括生物細胞、組織、器官和系統。
描述方法三	指人或動物的整個物理結構，包括生物細胞、組織、器官和系統。例如，細胞核、神經組織、左心房、脊髓以及呼吸系統都屬於身體的組成部分。
描述方法四	指人或動物的整個生理組織，有時特指軀幹和四肢。人或動物各生理組織構成的整體、健康狀況。

表 3. 產生類別描述的四種方法之範例

因此我們嘗試了產生類別的自然語言描述文本的四種方法：描述方法一是同樣是輸入一個詞彙；描述方法二是輸入本次任務網站中對於該類別的描述；描述方法三是以第二種方法為基礎，再加上此次網站中對該類別的舉例；描述方法四為查詢外部資料(例如詞典)對該類別的描述。表 3 是以類別「BODY」為例說明產生類別描述文本的四個範例。此外，我們將 O 類別一律描述為「其他類別詞」、B 型態描述為「開始詞」、I 型態描述為「中間詞」。在經過 4.3 節的實驗後，在本次任務本研究採用描述方法二。

3.2 Label Correction Model 與規則式模組

圖 3 是 LCM 的模型架構圖。LCM 的目的是對標記異常的結果進行再標記，希望藉由建立專門辨識異常標籤並修正的模型來提高標記正確率。因此，LCM 的設計是採用一個已經預訓練完成的 LSM 作為核心模型，然後以第一階段有異常標籤的句子作為 LCM 的訓練資料，微調已經預訓練的核心模型。

LCM 處理後可能仍有些輸出仍維持原先的異常標籤、或是重新輸出後仍是異常標籤。因此我們在第三部分以規則式方式修正異常標籤。異常標籤的判斷規則與相對應的修正方式說明如下：

規則一：若一個 I 標籤字元前面是一個 O 標籤字元，則該字元改為 O 標籤。

規則二：若一個實體內標記的標籤正常但類別不一致，則將該命名實體所有字元的類別改標記為與 B 標籤字元相同的類別。

4 實驗

本研究的訓練與驗證集 Chinese HealthNER Corpus 與 CHNER、以及測試資料集均由本次任務所提供。除此之外本研究無使用任何其他公開或非公開數據。

本次任務中對模型效能以精確率(precision)、召回率(recall)以及 F1 分數(F_1 score)三個指標進行：精確率是正確預測實體數量除以預測的實體總數量；召回率正確預測實體數量除以測試集中實體總數量；而 F1 分數定義如下：

$$F_1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1)$$

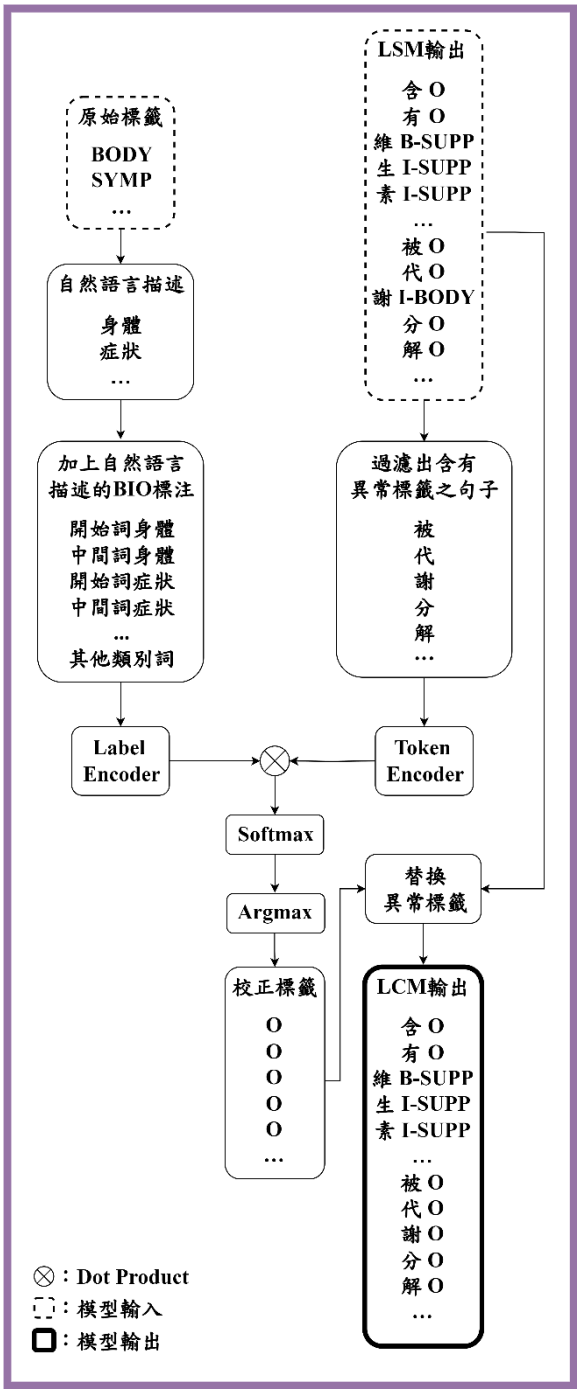

圖 3. LCM 架構圖

另外，本研究以 Lee, Lu, & Lin (2022)提出的 Word2vec-BiLSTM-CRF 與 BERT-BiLSTM-CRF 模型作為比較，以評估本研究提出的方法在測試集上的表現。

4.1 在測試集的表現

本研究所提方法與基準模型對本次任務提供的測試集 (Lee et al., 2023)之評估結果如表 4 所示，其中 Macro-Averaging 是指 FT、SM、WA 三個測試集所得之 F1 分數的平均值。在本次

任務中，本研究所提方法在 FT 與 SM 測試集以及 Macro-Averaging 的效能皆高於基準模型。另外，在所有參與本次任務的模型中，本研究所提方法在 SM 測試集上表現最佳。

測試集	基準模型		本研究所提方法
	Word2vec	BERT	
FT	0.6099	0.6108	**0.6252**
SM	0.6716	0.7077	**0.7142**
WA	0.6791	**0.7254**	0.7119
Macro-Averaging	0.6535	0.6813	**0.6838**

表 4. 兩種模型對測試集取得的 F1 分數

我們從三個測試集中產生的錯誤標籤隨機抽樣 45 句子，並分為以下五種錯誤類別。第一種為多類別標記錯誤。這種錯誤是指模型通常能正確預測實體部分字元的標籤與類別，但部分字元會被判斷成其他類別。例如類別為 SYMP 的實體「胎盤異常」中，「胎盤」被視為 BODY，而「異常」被判斷為 SYMP。

第二種為未能由關鍵詞識別。在訓練資料中無此實體存在，但依實體中的字詞應能判斷出類別，而模型卻標記為 O。例如，「新冠肺炎」的標籤為 DISE，「炎」是標籤 DISE 的常見關鍵詞，但標記結果卻為 O。第三種為未能由上下文識別。被錯誤標記的實體在訓練資料中無相似實體存在，實體中也沒有關鍵詞能判別出特定類別，但真人能依句子中的前後文來判斷出類別，但模型無法正確標記。例如，「截瘤達」的標籤為 DRUG，模型應該可由句子「醫師現在開截瘤達治療我」判斷該詞為 DRUG，但模型標記為 O。

第四種為未能正確標記多類別 NE。有些 NE 在訓練資料中存在兩種以上的類別，並且這些類別意義相近、該 NE 以各類別出現次數又相近，造成系統只能依前後文而不幸將 NE 誤判成另一類別。例如「貧血」有些會被標記成 DISE、有些標記成 SYMP。在標記待處理 NE 時容易誤判成另一種類別。第五種為其他無法歸類以及無法判斷錯誤原因的標記錯誤。

表 5 列出五種錯誤類別佔總體錯誤量的比例。其中第二與第三種類別錯誤我們認為都有進一步減少錯誤量的機會。因為這兩類都有可輔助辨識的資訊，但目前的模型可能因訓練資料量不足未能正確標記。而這兩類錯誤佔了超過四成的比例，因此我們認為此標記模型仍有許多可提升正確性的機會。

錯誤類型	比例
多類型標記錯誤	35.56%
未能由關鍵詞識別	22.22%
未能由上下文識別	20.00%
類別混淆	13.30%
其他	8.89%

表 5. 各種標記錯誤類型之發生比例

4.2 語意模型的選擇

本研究測試了多種可做為 LSM 模型中所使用之編碼器的語意模型。這個測試使用的資料集為 Chinese HealthNER Corpus，並且將其切割成 27,622 個句子作為訓練集，3,070 個句子作為驗證集。我們分別使用 BERT，RoBERTa (Liu et al., 2019)和 MacBERT (Cui et al., 2020)作為 LSM 的編碼器。實驗結果如表 6 所示，可以看到在雖然 RoBERTa 在驗證集中的 precision 最低，但 recall 與 F1 明顯優於 BERT，也略勝於 MacBERT，考慮到本次任務採納的評估指標，故此本研究將使用 RoBERTa 為基礎模型，模型的參數設定如下：epochs 為 5、batch size 為 32、embedding 大小為 128、optimizer 使用 Adam 方法、learning rate 為 10^{-5}。

模型	Precision	Recall	F1
BERT	**0.6997**	0.7126	0.7061
RoBERTa	0.6907	**0.7562**	**0.7220**
MacBERT	0.6942	0.7534	0.7189

表 6. 使用不同語意模型之效能

4.3 標籤描述對效能之差異

我們也分析了四種類別描述方法的差異。NER 模型部分只使用以 RoBERTa 為預訓練語意模型的 LSM；使用的訓練集與驗證集同 4.2 節。實驗結果如表 7 所示，比起使用單詞描述標籤，使用 ROCLING 描述的描述規則二表現較佳，故本研究選擇採用描述規則二作為本研究所提模型的標籤轉換方式。雖然如此，實際上前三種描述方法差異相當有限，反而是最詳細的描述規則四表現較差。我們推測

原因可能是描述資訊太過雜亂，導致模型學習到過多與類別無關的資訊。

描述方法	Precision	Recall	F1
一	0.7115	0.7796	0.7440
二	**0.7162**	0.7814	**0.7473**
三	0.7156	**0.7816**	0.7471
四	0.6967	0.7582	0.7262

表 7. 使用不同自然語言描述的結果

4.4 各模組效能分析

表 8 列出 LSM、LCM 以及規則式模組對整體效能的影響。由表 10 可得兩項結論：第一、只有使用 LSM 時最低、三者同時使用時最高，這表示 LCM 與規則式模組都有發揮作用。第二、使用 LSM 搭配 LCM 與規則式模組之一都能有效提高效能，但規則式模組更為顯著。但我們認為這不代表規則式模組優於 LCM 的設計，會有此項數據主要是異常標籤的數量不夠多，使得 LCM 的訓練成效有限；而規則式是採用經驗法則所擬定，未必符合真實情境。但是這項數據可以得知 LCM 的確學習了一部分規則式模型沒有涵蓋的規則並成功校正，因此有相當大的發展潛力。

模組組合	Precision	Recall	F1
LSM	0.7382	0.7831	0.7600
LSM+LCM	0.7578	0.8043	0.7806
LSM+規則式	0.7933	0.8081	0.8012
LSM+LCM +規則式	**0.8108**	**0.8140**	**0.8123**

表 8. 各模組組合的效能

5 結論與未來工作

本研究提出了一個由 LSM、LCM 以及規則式模組組合而成的 NER 模型。實驗結果顯示個子模型都有發揮作用，對本次任務的 SM 測試集在所有參與者中有最好的表現，是一個有效的設計。

我們認為未來在本研究的基礎上可以針對 LSM、異常標籤判斷規則與校正訓練再改良。首先是在 LSM 加入 Bi-LSTM (Bi-directional Long Short-Term Memory)模型(Zhou et al., 2016)至原先的編碼器之後，利用該模型雙向編碼的特性來增強文本上下文關係。此外，LSM 與 LCM 中的 softmax 與 argmax 程序，可以嘗試 Lafferty et al. (2001)提出的 CRF (Conditional Random Fields)取代，因為許多研究指出 CRF 是個對序列式標籤標記問題有相當好的效果。

此外，本研究所提方法中對於異常標籤的判斷規則僅有兩項，無法排除有更多異常類別的可能性，所以也可以嘗試挖掘更多的異常標籤類別。最後，目前的 LCM 是基於 LSM 再做微調的方式，且輸入的文本是經過異常標籤判斷規則式篩選的結果。可以考慮修改成在 LSM 中加入專門修正對於異常標籤權重的子模型，並在訓練模型時一併進行微調，這樣一來就可以免去規則式篩選遺漏問題，也可以由模型自動尋找最佳的修正方式。

致謝

本研究由國科會計畫編號 110-2511-H-992-003-MY3 以及教育部補助國立臺灣師範大學高等教育深耕計畫「學習科學跨國頂尖研究中心」支持，特此致謝。

References

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting Pre-trained Models for Chinese Natural Language Processing. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 657-668).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). 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 (pp. 4171-4186).

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991.

Lin, B. S., Chen, J. H., & Chang, T. H. (2022). NERVE at ROCLING 2022 Shared Task: A Comparison of Three Named Entity Recognition Frameworks Based on Language Model and Lexicon Approach. In Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022) (pp. 343-349).

Lee, L. H., Chen, C. Y., Yu, L. C., & 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).

Lu, Y., & Lee, L. H. (2020). Chinese Healthcare Named Entity Recognition Based on Graph Neural Networks. International Journal of Computational Linguistics and Chinese Language Processing, 25(2), 21-36.

Lee, L. H., & 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., Lin, T. M., & Chen, C. Y. (2023). Overview of the ROCLING 2023 shared task for Chinese multi-genre named entity recognition in the healthcare domain. In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing.

Lee, L. H., Lu, C. H., & Lin, T. M. (2022). NCUEE-NLP at SemEval-2022 Task 11: Chinese Named Entity Recognition Using the BERT-BiLSTM-CRF Model. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 1597-1602).

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (pp. 282-289).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2015). Gated Graph Sequence Neural Networks. arXiv preprint arXiv:1511.05493.

Ma, J., Ballesteros, M., Doss, S., Anubhai, R., Mallya, S., Al-Onaizan, Y., & Roth, D. (2022). Label Semantics for Few Shot Named Entity Recognition. In Findings of the Association for Computational Linguistics: ACL 2022 (pp. 1956-1971).

Roy, A. (2021). Recent Trends in Named Entity Recognition (NER). arXiv preprint arXiv:2101.11420.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 207-212).