雙向神經網路應用於臺灣客語詞性標記之成效比較

葉秋杏[1]，吳明宗[2]，劉吉軒[3]，賴惠玲[4]

[1] 國立政治大學英國語文學系
E-mail: csyeh.corpus@gmail.com
[2] 國立政治大學資訊科學系
E-mail: 105971011@nccu.edu.tw
[3] 國立政治大學資訊科學系
E-mail: jyishane.liu@gmail.com
[4] 國立政治大學英國語文學系
E-mail: hllai.nccu@gmail.com

摘要

在現代自然語言處理領域中，深度學習模型已然成為處理複雜語言任務的核心技術之一，尤其在詞性標注的任務中展現出顯著優勢，亦即根據預定義的標籤集，自動識別和標記文本中每個詞彙的詞性。本研究選用兩種雙向神經網路模型：BiLSTM 與 BERT 來進行臺灣客語詞性標記之成效比較，資料來源取自臺灣客語語料庫之客語四縣腔與海陸腔語料，並分別在兩個模型中使用 K-Folds 交叉驗證法進行驗證，計算出臺灣客語語料之驗證集與測試集的平均損失與平均準確率。實驗表明，此二種深度學習模型在詞性標注上均可達到約 93%的準確率，展現了優異的標記能力，並證明了深度學習技術在臺灣客語語言處理上的潛在價值。

關鍵字(5 個字)：自然語言處理、深度學習、雙向神經網路、詞性標注、臺灣客語語料庫

Abstract

Deep learning models have become one of the core technologies for handling complex language tasks in the field of modern natural language processing, particularly excelling in part-of-speech tagging, where they automatically identify and label the part of speech for each word in a text according to a predefined set of tags. This study employs two bidirectional neural network models, BiLSTM and BERT, to compare their effectiveness in POS tagging for the Sixian and Hailu Hakka data sourced from Taiwan Hakka Corpus. K-Folds cross-validation is applied in both models to evaluate performance, calculating the average loss and accuracy for the validation and test sets of Hakka data. Experimental results indicate that both models achieve an accuracy rate of approximately 93% in POS tagging, demonstrating excellent tagging performance and highlighting the potential value of deep learning technology in processing Taiwan Hakka language.

Keywords：Natural Language Processing, Deep Learning, Bidirectional Neural Networks, Part-of-speech Tagging, Taiwan Hakka Corpus

1. 前言

自然語言處理是人工智慧領域中的一個重要研究方向，旨在實現語言間的有效轉換，以促進資訊的相互交流。當前世界上有數百種語言，各自擁有獨特的結構和表達方式，即使傳達相同的語意，不同語言間的表達形式往往存在顯著差異。為了讓機器能夠學習和理解人類的語言結構，詞性標注是一個重要且不可或缺的步驟，其可將連續的文字資料結構化，從而提高語言分析和應用的準確性和效率。針對華語的詞性標注已有大量研究[2, 8, 11]，標注系統也相對成熟，並廣泛應用於華語語料庫中（如中央研究院的現代漢語平衡語料庫以及國家教育研究院的國教院語料庫索引典系統）。然而，針對臺灣客語的詞性標注仍面臨許多挑戰，這些挑戰來自於客語自身的語言特性以及其在現代科技

中的資源相對匱乏。儘管客語和華語同屬漢語系，然兩者之間的語法結構存在顯著差異，特別是在詞序和句法方面。舉例來說，客語中的雙賓語或主謂結構，在華語中可能以不同的語法形式呈現，此會導致詞性標記上的不一致性，進而難以精確對應。客語詞彙在詞性上的轉換也相當靈活，許多詞彙可以在不同語境下充當不同的詞性，例如某些詞彙在華語中可能固定為名詞或動詞，但在客語中，這些詞可能會根據上下文變為副詞甚至是詞組形式。在客語中亦常見多功能詞，即同一詞彙在不同語境中可擔任不同的詞性角色。這類詞彙的多義性或是詞性轉換的靈活性，使得單純套用華語詞性標記變得困難，因為一個詞在不同語言中的語法功能可能截然不同。而深度學習模型能夠從大量的語料中學習語言的上下文訊息，並藉此捕捉語言的多層次結構與語意特徵。因此，在面對多義詞或結構模糊的情況時，深度學習模型可更準確地進行詞語辨識以及詞性標記。

本研究目標係採用深度學習模型中的雙向神經網路：BiLSTM 與 BERT 模型，個別訓練跟預測臺灣客語詞性之自動序列標記，並進行相關的深度學習模型測試。本文架構分為五節，第一節為研究動機與目的，並於第二節介紹研究模型與資料來源。第三節說明臺灣客語詞性自動標注系統架構，有關於實驗設計與結果討論則列於第四節，最後是結論。

2. 研究模型與資料來源

詞性標記可為深度學習模型提供更多的語言層次資訊，有助於模型更精確地判斷詞語的邊界及其在句子中的功能，對於如臺灣客語這類具有複雜詞形變化和語法結構的語言特別重要。以下將逐一說明雙向神經網路模型的運作原理，以及資料集來源和臺灣客語詞性標記。

2.1 深度學習模型：雙向神經網路

深度學習旨在模擬人腦的神經元結構並用以處理訊息，其利用多層神經網路（亦即深度神經網路）來自動學習和提取數據中的特徵，這些多層結構使模型能夠從複雜的數據中提取抽象特徵，並且具有較強的表示學習能力。而深度學習模型中，單向式模型指的是僅考慮序列中前一部分的訊息來預測或處理後一部分（如「遞歸神經網路（Recurrent neural network，RNN）或「長短期記憶網路」（Long Short-Term Memory，LSTM）」，其結構較為簡單，對於計算和存儲需求較低，參數調教也相對容易。然而因其僅考慮前向上下文，無法利用序列中後向的訊息，因此也無法全面捕捉全局資訊，尤其遇到複雜語境或長序列文本時，可能無法有效捕捉詞語邊界，並可能遇到梯度消失或梯度爆炸問題，導致詞性標注效果不佳。而雙向式深度學習模型在處理序列數據時，將序列分為兩個方向進行處理：一個是從左到右（前向）的網路，另一個是從右到左（後向）的網路。這樣的結構允許模型同時獲取和利用序列中的過去和未來訊息，在處理需要考慮前後上下文的詞性標注任務中，雙向模型能夠更好地理解詞語的語境，並可處理長期依賴的問題，從而提供更高的準確性。因此，本文選用「雙向長短記憶網路」（Bi-directional Long Short-Term Memory，BiLSTM）[4, 5]與「基於變換器的雙向編碼器表示技術」（Bidirectional Encoder Representations from Transformers，BERT）[3, 7, 9]兩個雙向式深度學習模型來處理臺灣客語詞性標注，並比較兩者之成效。

2.2 資料集來源與詞性標記

本實驗所使用的語料為臺灣客語語料庫（Taiwan Hakka Corpus，以下稱 THC）所收錄之書面及口語語料，涵括四縣、海陸、大埔、饒平、詔安、南四縣六種腔調，其中少數腔（大埔、饒平、詔安）受限於人口數量以及使用比例，收錄目標為總字數約 3% [10, 12]。THC 係由客家委員會於 2017 年 12 月底委託政治大學所建構，目前（截至 2024 年 11 月）已收錄書面 751 萬字，口語 127 萬字。語料係依照教育部及客委會之規範進行用字校訂，並透過長詞優先法與詞庫查找進行斷詞與詞性標注，以提高語料庫分析及應用的效率與準確性。THC 之詞性標記參考中研院平衡語料庫詞類標記集和賓州中文樹庫詞性標記集，並依照客語真實語言表現進行調整後，共計為 24 類：AD（副詞）、AS（時態）、B (附著語素)、BUN (「分」)、C（連詞）、DED (「得」)、DET（限定詞）、DO (「到」)、

GE(「个」,結構助詞)、HE(「係」,繫動詞)、IJ(感嘆詞)、LAU(「摎」)、M(量詞)、N(名詞)、NEG(否定詞)、P(介詞)、PN(代名詞)、PRT(助詞)、PU(標點)、RN(專有名詞)、SYM(符號)、TUNG(「同」)、VA(行動動詞)、VS(狀態動詞)。24 標記中有 7 個係以客語發音來標記,其中包含 BUN、DO、LAU、TUNG 4 個為客語中的多功能詞,較不易於區別,因此以客語發音標記。繫動詞 HE 則是參考中研院現代漢語平衡語料庫之作法(華語繫動詞「是」以其華語發音 SHI 標記),同樣以客語發音標記。結構助詞亦依循中研院的作法(華語結構助詞「的」、「得」、「地」標記為 DE),依照客語結構助詞之發音,標記為 GE(「个」)與 DED(「得」,此字也是多功能詞,除了可擔任結構助詞外,亦可作為行動動詞、表情態之狀態動詞以及動詞名物化之名詞)。

3. 臺灣客語詞性自動標注系統架構

臺灣客語詞性之自動標注系統架構流程主要包括客語語料初步處理、資料前處理、模型建立與模型準確率評估。首先是客語語料初步處理階段,自資料庫(THC)中提取客語語料後,經過語料庫工作人員將每個詞語的詞性標記出來,做為標準答案。接著進行資料前處理,將整個資料集分為訓練集、測試集和驗證集。於此基礎上,建立臺灣客語詞性標記模型,並使用訓練集、驗證集進行模型訓練。在模型訓練完成後,使用測試集來評估模型的準確率。各流程細節說明如下。

3.1 客語語料初步處理

本文訓練資料來源取自 THC 之四縣腔與海陸腔語料各 2 萬句,語料文本中的換行符號用以標示出不同語句之間的分隔點,具有結束當前語句的作用。另由於標點符號在自然語言處理領域中被視為一種標記符號,主要作用在於劃分句子中的語法結構,提供了語言單元之間的邊界資訊,可幫助模型識別句子結構與斷句邊界,從而提高斷詞的準確度,因此,所有文本先以換行符號斷句後,再將所有標記為 PU(Punctuation,標點符號)的符號做為斷句標記。在語料清洗過程中,句子結尾不是 PU 者均從資料集中剔除,並排除長度僅為 1 與 2 的句子,因句子過短,詞與詞之間的詞性缺乏足夠參照,容易引發混淆並可能導致錯誤率較高。長度大於 12 的句子也須排除,因過長的句子通常複雜度較高,容易有雜訊產生。語料中若有重複的句子也予以剔除。完成語料初步清理後,留下長度 3 至長度 12 的句子,每個長度句子數量各 2,000 句(長度統計不記入標點符號(PU))。所有詞彙均帶詞性標記,係透過 THC 之斷詞及詞性標注系統所完成。此斷詞工具採用長詞優先法搭配詞庫查找比對功能,連線至 THC 詞庫以獲取各詞條的詞性資料,並透過運算匹配合適的斷詞及標記組合來進行斷詞與標注,故斷詞與詞性標記仍存在一定的錯誤率。因此,所有資料均需再由受過語言學訓練的工作人員針對斷詞與詞性錯誤進行人工校正。完成後,語料的句子長度、斷詞方式以及詞性均有程度不一的修正,各長度句子數量也因而有所增減,部分句子在修正過後,長度變成不到 3 或是超過 12,因此必須再做進一步剔除。經再度篩選後,每個長度的句子數量各取 1,500 句,作為後續的模型訓練使用。

3.2 資料前處理

完成句子清理後,還需進行資料前處理,以確保資料格式適合模型讀取。主要步驟可分為三階段:第一階段先整理各語料長度的集合,第二階段將語料分割為「詞的序列集合」及「詞性類別的序列組合」,第三階段則是將所有文本資料轉成數值,以提供模型輸入使用。

在第一階段,經清洗過的語料已分別組成長度 3 到 12 的句子集合,為使資料均勻分布,在各長度集合中隨機抽取固定數量且足夠的句子合併成資料集,再分割成訓練集、驗證集及測試集。第二階段則是將資料集每個句子中的詞與詞性類別分割開來,分別組成詞的序列資料集合及詞性類別的序列資料集合。到了第三階段,即是須將文本資料轉成電腦可以處理的數值。首先,針對詞的序列資料集合建立一個相異詞性的字典,並將句子中每一個詞根據字典編碼,做為「詞編碼」。另一方面,詞性類別的序列資料集合也建立一個相異詞性類別的字典,將所有詞性類別也進行編碼,做為「詞

性類別編碼」。因模型期望接收固定長度的輸入，為確保輸入資料的統一長度，接下來將編碼過後的詞組序列資料及詞性類別序列資料，都補滿長度 12 的詞組。不足長度 12 之處，則以 0 顯示。最後，使用熱編碼[1]消除特徵之間的數學大小關係，讓相異詞性的每個詞性編碼擴增為句子最大長度來標示文本中的詞性，如圖 1。

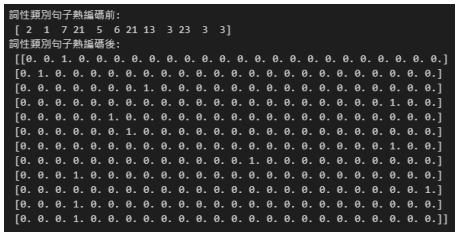

圖 1. 詞性類別句子熱編碼後

3.3 模型建立與準確率評估

本研究選擇使用的 BiLSTM 深度學習模型實作是採用 Google 開源的 TensorFlow 框架，並利用其所內建之 Keras 提供的順序模型結構(Sequential Model)實作。此模型結合嵌入層(Embedding Layer)並加入時間分布層（TimeDistributed Layer)，其中包含一個全連接層（Dense Layer)。這樣做的效果是將全連接層應用於序列數據的每個時間步，以及將輸出維度設置為與類別數量相同，並使用 softmax 激活函數進行分類概率的計算。此設置通常用於多類別分類問題，其中模型目標是預測每個輸入序列的詞性。

BERT 訓練模型則是採用 Pytorch 的架構所建立，此模型使用了一個 BERT 模型(預訓練模型)，再設定嵌入層的維度，並由一個線性層組成，用於進行分類。該模型的 forward 方法接受文本輸入，將其通過 BERT 模型，然後通過線性層，以獲得詞類標記的預測結果。預訓練模型內容包含了嵌入層和編碼器（Encoder)。嵌入層負責將輸入的文字轉換成向量表示，這一層包含了詞嵌入、位置嵌入和片段嵌入，而這些嵌入向量將文本的語義和位置訊息編碼成固定維度的向量。編碼器是 BertModel 的核心組件，由多個 BertLayer 組成，每個 BertLayer 都包含了自注意力機制和前饋神經網路。自注意力機制用於在文本中建立詞與詞之間的關聯性，將每個詞與其他詞計算相似度並獲得加權表示。前饋神經網路則對自注意力機制的輸出進行非線性變換，以增強表示能力。編碼器中的每個 BertLayer 都有一個殘差連接（Residual Connection）和層正規化（Layer Normalization)操作，以幫助梯度流動和模型的穩定訓練。編碼器可以由多個 BertLayer 堆疊而成，通常使用 12 或 24 個 BertLayer。通過多層的疊加，模型可以捕捉到更多不同層次的語義訊息，從而提升模型的性能。線性輸出層是詞類標記的輸出維度，代表了有多少不同的詞性標記，最後使用了 Dropout 層，透過隨機關閉一部分神經元，可以阻止它們過度依賴其他神經元的特定特徵，從而減少神經元之間的相互依賴性，用於減少過擬合。

模型訓練完後，再利用交叉驗證過程中，透過驗證集把 loss 最小的模型參數留存下來，以利後續提供給測試集用以預測每個輸入序列的詞性。

所有資料處理完畢後，即是進行訓練集之訓練，並使用 K-Folds 交叉驗證法（K-Folds Cross Validation)[6]進行驗證，計算出驗證集與測試集的平均損失與平均準確率。所有資料被分割為 k 個不同的子集，使用 k-1 個子集來訓練資料，最後一個子集則留下作為測試資料。接下來，對每個子集計算損失（Loss)和準確率（Accuracy)，並在測試集上進行測試。訓練結束後，將所有測試結果的詞性預測準確率加總並取平均值，即可用以評估模型之詞性標注能力。

4. 實驗設計與結果討論

本研究之客語詞性標注模型於Windows 10 使用 Python 3.9.16 進行開發 BiLSTM 模型，深度學習框架使用 Tensorflow 版本 2.13.0。另於 Ubuntu 20.04.5 LTS 使用 Python 3.9.5 進行開發 BERT 模型，深度學習框架使用 Pytorch 版本 1.8.0，GPU 使用 NVIDIA GeForce RTX 3090/24G。

4.1 實驗資料集

實驗語料為臺灣客語四縣腔與海陸腔語料，經人工校正並篩選淘汰後，得到長度 3 至 12 的句子各約 1,500 句，總數為兩腔調各約 1 萬 5 千句供模型使用。語料集以亂數平均取樣後，再分割成訓練集 11,700 句、驗證集 1,300 句及測試集 2,000 句。

4.2 實驗結果分析

4.2.1 BiLSTM 實驗結果

BiLSTM 透過嵌入層和單層的 RNN 隱藏層進行處理，最後經由全連接層輸出類別概率分佈。本研究使用了 K-Folds 交叉驗證迴圈，對於每個 Fold，模型進行多個訓練迭代和驗證迭代。基於驗證損失的最佳模型被保存下來。在交叉驗證迴圈結束後，使用驗證階段的最佳模型對測試集進行評估。四縣腔語料的實驗結果顯示，測試集之單一詞彙正確率取得的準確率為 93%，測試集之單一句子正確率（指句中每個詞彙的詞性標記均正確）的準確率為 62%，另外從統計表中可發現長度越長單一句子的準確率就越低，低於 60%有 4 組，60%以上的則有 6 組。(圖 2)。

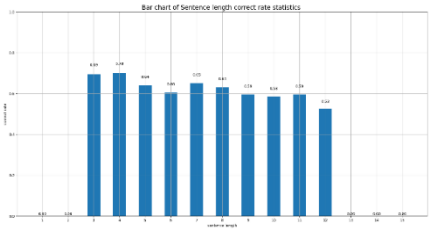

圖 2. BiLSTM 之四縣腔測試集實驗結果
（單一句各長度的正確率）

而在海陸腔語料的測試集中，單一詞彙正確率取得的準確率為 94%，測試集單一句子正確率的準確率為 65%，而統計表（圖 3）顯示，海陸腔各長度讀單一句子的準確率比較平均，除了長度 11 的組為 54%以外，其他 9 組的準確率均為 60%以上。

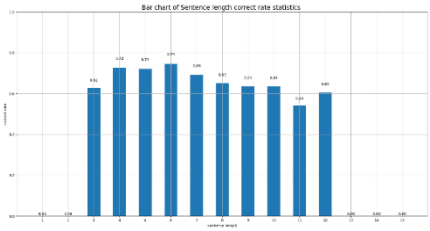

圖 3. BiLSTM 之海陸腔測試集實驗結果
（單一句各長度的正確率）

4.2.2 BERT 實驗結果

在 BERT 模型的訓練中，本文使用中研院預先訓練好的模型（ckiplab/Bert-base-chinese-pos），再對下游任務做微調（Fine-tuning），使用單一句子標註任務的情境。在微調之前，由於預訓練模型為繁體中文且為單一字的模型，而臺灣客語語料集中的詞彙包含兩個字以上，導致模型無法正確辨別因而將二字以上的詞均判別為 UNK（unknown）。因此，客語語料也進行前處理，將被判別為 UNK 的詞都加入到模型字彙中（如表 1），以符合 BERT 的相容模式。

THC 句子	'湯', '老先生', '再', '講', '另', '一', '隻', '有關', '大', '茄苳樹', '个', '故事'
字彙新增前	'湯', '[UNK]', '再', '講', '另', '一', '隻', '[UNK]', '大', '[UNK]', '个', '[UNK]'
字彙新增後	'湯', '老先生', '再', '講', '另', '一', '隻', '有關', '大', '茄苳樹', '个', '故事'

表 1. 新增字彙前後的模型辨識結果

於此模型同樣地也使用 K-folds 交叉驗證迴圈，並於完成後使用驗證階段的最佳模型對測試集進行評估。在 BERT 的測試集中，四縣腔單一詞彙正確率取得的準確率為 93%，單一句子正確率取得的準確率為 62%，並且隨著長度越長，單一句子的準確率也出現降低的趨勢，2 組低於 60%，8 組超過 60%以上準確率（圖 4）。

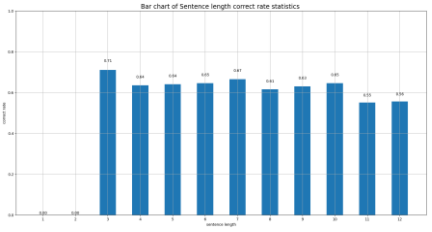

圖 4. BERT 之四縣腔測試集實驗結果
（單一句各長度的正確率）

而在海陸腔的結果中，測試集單一詞彙正確率取得的準確率為 93.7%，測試集單一句子正確率取得的準確率為 63.5%，並且也呈現出當句子長度越長，準確率就越低的現象，其中 3 組低於 60%，7 組超過 60%以上準確率（圖 5）。

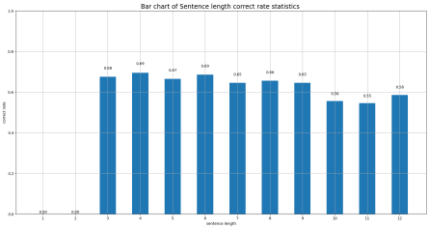

圖 5. BERT 之海陸腔測試集實驗結果
（單一句各長度的正確率）

4.3 模型結果綜合比較分析

在本研究中，針對客語的不同腔調（四縣、海

陸)使用兩種不同的雙向神經網路模型 BiLSTM 與 BERT 進行訓練和測試，並針對模型訓練後使用驗證集取得 loss 較低的模型去對測試集預測，記錄了相關結果呈現如表 2。

腔調	四縣		海陸	
模型	BiLSTM	BERT	BiLSTM	BERT
複雜度	複雜	更複雜	複雜	更複雜
參數量	2,640,024	112,225,560	2,511,128	111,503,640
執行時間	80 分鐘	67 分鐘	80 分鐘	68 分鐘
Loss	0.21	0.37	0.19	0.393
詞彙準確率	0.931	0.926	0.936	0.93
句子準確率	0.62	0.62	0.65	0.63

表 2. 腔調模型綜合比較表

BiLSTM 在隱藏層中用了 12 個神經元，因多了輸入閘、輸出閘、遺忘閘及雙向的 LSTM，1 個神經元等同於配置 8 個參數。BERT 是更大且更為複雜的模型，僅 BertEncode 層就用了 12 層，每一層的神經元都使用 768 個參數，整體的參數量遠大於 BiLSTM 模型。而在執行時間方面，由於 BiLSTM 是在 CPU 上運行，相對於 GPU 而言，執行時間較長，但屬可接受範圍。在參數設置為批次大小（batch size）為 100、訓練輪數（epoch）為 40 的情況下，BiLSTM 的執行時間需要 80 分鐘。BERT 模型因其需要大量的矩陣運算和並行計算，在 CPU 上所需的執行時間較長，加上 CPU 的核心數量有限，面對大規模的並行計算時，效率相對較低，因此改為在 GPU 上運行。參數設置的批次大小調整為 1,024，訓練輪數調整為 100，執行時間約為 67 分鐘。在詞彙準確率方面，BiLSTM 和 BERT 表現相近，均可達約 93%；而句子準確率因為係指句中每個詞彙的詞性標記都必須正確，因此準確率較低，略高於 60%。

5. 結論

本研究應用深度學習技術於客語詞性自動序列標記任務，並比較 BiLSTM 和 BERT 模型的表現，以期提升客語語料自動標注的準確性。在客語四縣腔與海陸腔語料上，兩個模型均展現出良好的性能。由於本次實驗使用的語料量較小，相較於參數數量與模型複雜度均較高的 BERT，BiLSTM 於此任務中已足以滿足客語詞性標注之需求。值得一提的是，隨著臺灣客語語料庫持續收錄的書面及口語語料不斷增加，語料規模日益龐大，語言結構也會愈加複雜，特別是臺灣客語的六種腔調在語音和語法結構上存在著一定的差異，對於語言處理模型而言將會是一大挑戰，而 BERT 模型之結構特性讓其在處理更大規模且更為複雜的語料時，可以發揮出更強的處理能力與泛化潛力。本次實驗成果可作為後續研究之基礎，在未來針對語料多樣性更高或語境依賴性更強的任務中，可進一步綜合比較不同的深度學習模型，探索如何針對各個腔調進行優化，提升在不同語境下的詞性標注準確性，以期促進客語語言處理技術之發展與應用。

6. 參考文獻

[1] É. Arnaud, M. Elbattah, M. Gignon, and G. Dequen, “NLP-based prediction of medical specialties at hospital admission using triage notes,” in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pp. 548-553, 2021.

[2] S.W.K. Chan, M.W.C. Chong, and T.B.Y. Lai, “Unknown Chinese composite words tagging using selective back-off smoothing,” in *Journal of Chinese Linguistics Monograph Series*, Vol. 25, pp. 139-159, 2015.

[3] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL2019)*, pp. 4171-4186, 2019.

[4] A. Graves, and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” in *Neural Networks*, Vol. 18, No. 5,pp. 602-610, 2005.

[5] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” in *Neural Computation*, Vol. 9, No. 8, pp.1735-1780, 1997.

[6] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1137-1143, CA: Morgan Kaufmann Publishers Inc, San Francisco, 1995.

[7] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 2227-2237, Association for Computational Linguistics, New Orleans, Louisiana, 2018.

[8] Y.F. Tsai, and K.J. Chen, “Reliable and cost-effective pos-tagging,” in *International Journal of Computational Linguistics & Chinese Language Processing*, Vol. 9, No. 1, pp. 83-96, 2004.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

[10] 客家委員會，**臺灣客語語料庫**，網址：https://corpus.hakka.gov.tw/

[11] 張詠淳、葉文照、謝育倫、許聞廉，「MONPA: 多目標中文命名實體辨識與詞性標註系統」，**第 24 屆人工智慧與應用研討會**（TAAI 2019），臺北：中華民國人工智慧學會，臺灣，R.O.C，2019。

[12] 葉秋杏、賴惠玲、劉吉軒，「臺灣客語語料庫建置與客語詞彙使用初探」，**數位典藏與數位人文**，8，pp. 75-131，2021。