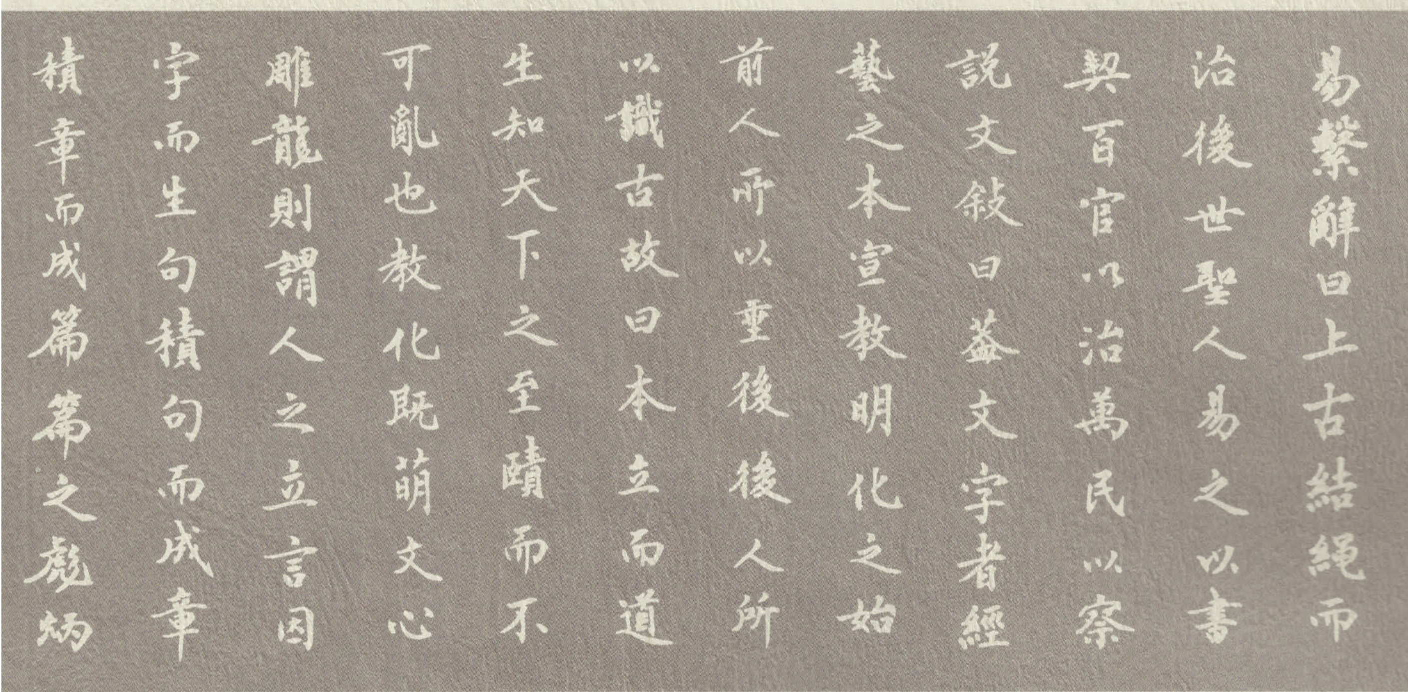
International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.



Vol.24 No.1 June 2019 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 24, no. 1, June 2019

Contents

Computational Linguistics and Chinese Language Processing
Vol. 24, No. 1, June 2019, pp. 1-14

應用記憶增強條件隨機場域與之深度學習及自動化詞彙特徵於中文命名實體辨識之研究

Leveraging Memory Enhanced Conditional Random Fields with Gated CNN and Automatic BAPS Features for Chinese Named Entity Recognition

簡國峻*、張嘉惠*

Kuo-Chun Chien and Chia-Hui Chang

摘要

命名實體辨識是在自然語言處理當中一個重要的任務。現今基礎深度學習模型應用於資料品質較為優良的命名實體擷取，雖有不錯的效果，但在社群媒體資料集中卻未能達到傳統條件隨機場域之基準值。由於一個命名實體有能可多次在文中提及，因此藉由上下文資訊來改進命名實體的擷取也是近年來的研究方向。在本研究中，我們延伸記憶增強條件隨機場域 MECRF 於中文的命名實體擷取，利用門控卷積網路及雙向 GRU 網路來增強記憶條件隨機場域，以利模型抓取長距離的文章資訊。此外，也藉由特徵探勘擷取命名實體前後詞彙以及前綴後綴詞彙特徵（簡稱為 BAPS），並使用模型可自動訓練的參數，自動調整詞向量及 BAPS 詞彙特徵。最後我們同時採用字元及詞彙向量來增進模型的效能。本研究所提出之模型，在網路社群媒體的人名辨識資料中可以達到的 91.67％準確率，在 SIGHAN-MSRA 中也得到最高的 92.45％地名實體辨識效果及 90.95％整體召回率。

關鍵詞：機器學習、命名實體辨識、神經網路、特徵探勘

* 國立中央大學資訊工程學系
Department of Computer Science & Information Engineering, National Central University
E-mail: qk0614@gmail.com; chia@csie.ncu.edu.tw

Abstract

Named Entity Recognition (NER) is an essential task in Natural Language Processing. Memory Enhanced CRF (MECRF) integrates external memory to extend Conditional Random Field (CRF) to capture long-range dependencies with attention mechanism. However, the performance of pure MECRF for Chinese NER is not good. In this paper, we enhance MECRF with Stacked CNNs and gated mechanism to capture better word and sentence representation for Chinese NER. Meanwhile, we combine both character and word information to improve the performance. We further improve the performance by importing common before and common after vocabularies of named entities as well as entity prefix and suffix via feature mining. The BAPS features are then combined with character embedding features to automatically adjust the weight. The model proposed in this research achieve 91.67% tagging accuracy on the online social media data for Chinese person name recognition, and reach the highest F1-score 92.45% for location name recognition and 90.95% overall recall rate in SIGHAN-MSRA dataset.

Keyword: Machine Learning, Named Entity Recognition, Memory Network, Feature Mining

1. 緒論 (Introduction)

命名實體辨識(Named Entity Recognition, NER)是自然語言處理中訊息理解的第一步，其目標是提取當中的命名實體並歸類到預先定義的分類當中，如：人名、地名、組織等。傳統的機器學習於命名實體的辨識任務中，大多使用統計式條件隨機場域進行序列標記，因此受限於小範圍的特徵擷取。如何在中文的資料集當中擷取參考長距離上下文資訊，判斷當前字詞正確的語意，進而正確的辨識命名實體，是機器理解訊息根本的任務。

近年來深度學習被運用在序列標記的模型建立，得到不錯的進展。例如 Huang 在序列標記的任務上使用長短期記憶(Huang, Xu & Yu, 2015)，應用於英文的資料集當中獲得了非常好的效能。Liu 等人於 IJCNLP 2017 將記憶網路的概念加入條件隨機場域當中(Liu, Baldwin & Cohn, 2017)，提出 MECRF 架構，透過整合上下文額外的記憶，使模型能夠獲取較長範圍以外的文章特徵，同樣在英文資料集上獲得了出色的表現。然而這些基礎深度學習模型應用於資料品質較為優良的資料集上，雖均有不錯的效果，但在社群媒體資料集中卻未能達到傳統機器學習方式之基準值，因此如何有效地擷取文字中所隱含的資訊，使模型有較好的濾除雜訊之能力，也是在應用上非常重要的一環。

為改善上述的限制，本研究延伸記憶增強條件隨機場域 MECRF 於中文命名實體辨識任務；MECRF 的概念是基於上下文可能不只一次提及實體名稱，以及 Attention 機制的應用，藉以更正確找出命名實體。我們首先透過訓練詞向量模型，將字元轉換為數值

化之資料；再藉由卷積層、雙向 GRU 層提供模型更多的特徵，及整合長距離文章資訊的記憶層，使命名實體任務不同於往常僅能夠擷取小範圍的資訊，能夠獲取豐富完整的文章訊息。此外，也藉由特徵的探勘(Chou & Chang, 2017)，並使用深度學習模型可自動訓練的參數，自動調整詞向量及詞彙特徵，除長距離的文章資訊外，更能充分獲得文章所隱藏的訊息。

本研究所使用的資料為 Chou 及 Chang 所使用的 PerNews 測試資料集，但其資料集是以句子為單位進行標記，並無上下文，因此我們自製爬蟲程式，蒐集原始資料的網路新聞及社群媒體做為訓練及測試資料。經實驗結果比較，在網路社群媒體的資料中可以達到的 91.67%的標記準確率，與尚未加入記憶的模型相比大幅提升 2.9%，再加入詞向量及詞彙特徵，與基礎的記憶模型相比更是提升了 6.04%。本研究所提出之模型也在 SIGHAN-MSRA 中得到最高的 92.45%地名實體辨識效果及 90.95%召回率。

2. 相關文獻回顧 (Related Work)

序列標記已經發展許久，常見的模型有隱藏式馬可夫模型(Hidden Markov Model, HMM)、最大化熵馬可夫模型(Maximum Entropy Markov Model, MEMM)以及條件隨機場域(Conditional Random Field, CRF)。Lafferty 等人(2001)所提出的條件隨機場域在自然語言處理序列標記(Sequence Labeling)的任務中，是多數人的選擇且被廣泛的應用，但是條件隨機場域僅能夠抓取小範圍的文章資訊(Finkel, Grenager & Manning, 2005)，對於獲取整篇文章中的資訊則是條件隨機場域關鍵的限制。

2.1 卷積神經網路(Convolutional Neural Networks)

卷積神經網路是一種前饋神經網路，通常由卷積層(Convolutional)、池化層(Pooling)、全連接層(Fully-Connected)組成，相較於其他的網路，卷積神經網路所需要使用的參數較少，因而成為一種頗具吸引力的深度學習模型。卷積神經網路擁有能夠自動抓取相鄰特徵的優點，Collobert 等人(2011)首先將卷積神經網路自動抓取相鄰特徵的優點應用在自然語言處理的序列標記任務中，讓自然語言處理不再相依於專業知識特製而成的特徵模板。近期，Wang 等人(2017)透過堆疊式的卷積神經網路更有結構、多階層地萃取中文語意特徵，同時結合 Dauphin 等人(2016)提出的閘門線性單元(Gated linear unit, GLU)，應用於中文斷詞任務中。

2.2 遞歸神經網路(Recurrent Neural Networks)

RNN 則是另一種處理序列型輸入的神經架構，但是單純的 RNN 模型無法擷取長距離的文章資訊，為了不受局部限制的影響，因此有常短期記憶(Long Short Term Memory)的提出。Huang 等人(2015)在序列標記的任務上使用長短期記憶，導入雙向(Bidirectional)的概念來擷取正向及反向的資訊，應用於英文的資料集當中獲得了非常好的效能。

但是遞歸神經網路隨著輸入句子的長度增加(Cho, van Merrienboer, Bahdanau &

Bengio, 2014)，會帶來效能的惡化。在相關的研究(Lai, Xu, Liu & Zhao, 2015; Linzen, Dupoux & Goldberg, 2016)更顯示，遞歸神經網路包括其變化之類型，儘管已經加入時間序列的標記，但仍偏向於相鄰的字元資訊，在涉及遠程上下文依賴性的判斷中表現不佳。

2.3 記憶網路(Memory Networks)

傳統的條件隨機場域沒有能力去抓取較長範圍以外的文章特徵，而遞歸神經網路在長距離的文章資訊擷取上效能也並不出色，因此，Weston 等人提出記憶網路(Memory Network)來增強擷取長範圍文章特徵的表現，並應用於問答(QA)的任務當中(Weston, Chopra & Bordes, 2014)，證明記憶的增加對於執行需要常距離文章資訊的推理至關重要。

近期，Liu 將記憶網路的概念加入條件隨機場域當中(Liu *et al.*, 2017)，透過整合額外的記憶(Memory)，使模型能夠獲取較長範圍以外的文章特徵，並且在英文資料集上獲得了出色的表現。

3. 模型架構及方法(Model Architecture and Method)

在命名實體辨識標記任務中，每一個句子 S 是由 T 字元(character)組合而成的序列 $S = \{w_1, \cdots, w_T\}$，其對應的標籤序列可表示為$Y = \{y_1, \cdots, y_T\}$。不同於傳統的條件隨機場域僅需要輸入句子，MECRF 的特點是另有上下文資訊或稱之為記憶體M_s。假設每篇文章是由$|D|$句子組成$D = \{S_1, \cdots, S_{|D|}\}$，與其對應的序列標籤集合$L = \{Y_1, \ \ldots, Y_{|D|}\}$。為避免輸入整篇文章造成記憶體消耗過大，我們僅抓取當前句子S_t的前後 B 句共抓取 2B+1 個句子M_s =$\{S_{t-B}, \cdots, S_t, \cdots, S_{t+B}\}$做為短期記憶(short context)，其長度可記為 N=$\sum_{i=t-B}^{t+B} T_i$（其中T_i表句子S_i的長度）。每個輸入字元w_j可以透過 word2vec 或 GloVe 對文字進行編碼, 以EMB(w_j)來表示。假設 D 為 Embedding 的維度，則短期記憶增強隨機場域的輸入序列為大小 TxD 的句子E^{S}、和 LxD 的短期記憶E^{M}。

3.1 Stacked CNNs with Gated Mechanism

在本篇論文中，我們應用多層次卷積（Convolution Layer）來萃取文字特徵，並參考Dauphin 等人做法在層與層間加入門控機制來泛化萃取的特徵。門控機制廣泛地應用於循環神經網路架構中，用來控制長期神經網路中資訊的流動；在卷積神經網路中雖沒有長期依賴的問題，不需要輸入閥門以及遺忘閥門，但是 Dauphin 等人認為在多層次的卷積神經網路中，層與層之間可以透過類似輸出閥門的門控機制來決定神經元的流通與否，並有效率地擷取有效的特徵。假設前面嵌入層輸出為E^{S}（$\epsilon R^{T \times D}$)，則此處卷積運算可表示為：

$$A = \mathrm{E}^{\mathrm{S}} \oplus W_K \ + b \tag{1}$$

其中W_K為大小為$K \times D$的卷積運算過濾器 Kernel Filter，K 若過小導致不能含括有效資訊；若過大導致含括冗餘資訊對系統產生不必要的干擾，本研究中將 K 設定為 3，再透過多層卷積層擴及字元前後資訊；我們將滑動視窗移動的格數(strides)設為 1，並將補零方式

(padding)設為 SAME，意即輸出長度等於輸入長度。此處卷積運算後不採用池化層(Pooling Layer)，其原因為中文語意中每個特徵都有其意義，不像影像可能會經過放大、縮小或者位移，因此本研究直接將 L 個卷積 Filters 輸出的 feature maps 做連接(Concatenation)。

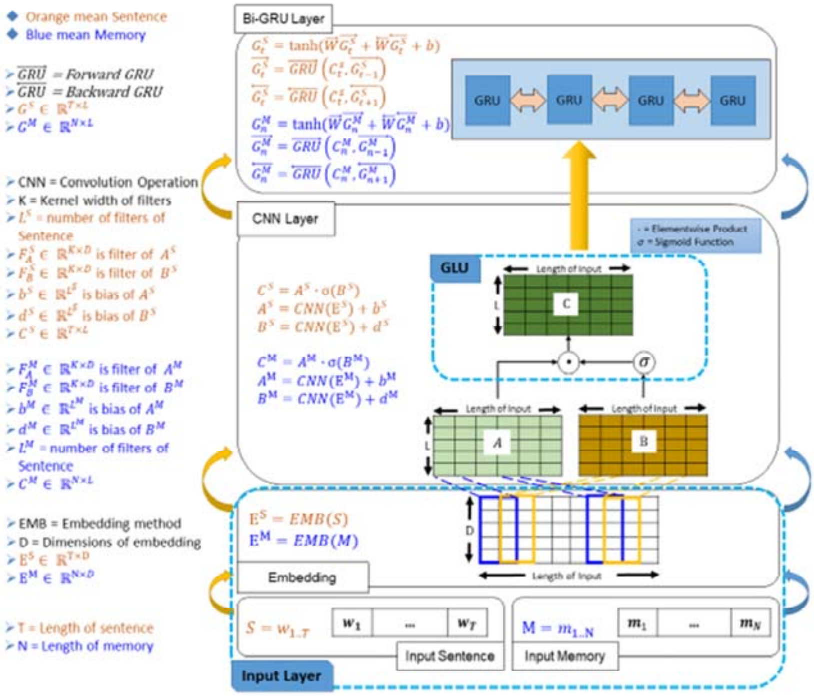


圖 1. Input sentence and memory representation

令 A 及 B 分別為經由兩組 CNN 卷積運算之後所產生的矩陣。前者不經過任何啟動函數，後者將通過一非線性轉換(sigmoid function)用來決定神經元的取捨，再將兩輸出做矩陣逐元素乘法(element-wise multiplication)，如式 (1)所示。

$$C = A \cdot \sigma(B) \tag{2}$$

應用多層卷積（Stacked Convolution）及門控機制（Gated-CNNs）擷取相鄰字詞特徵後，我們參考 MECRF 採用遞歸神經網路(RNN)的變化體 GRU，且透過雙向的技術來擷取當下位置處的文字C_t正向及反向的資訊，並且於輸出時，將正向及反向的資訊套用一個非線性單元 tanh，做為位置 t 的輸出資訊G_t，如式(2)。

$$G_t = tanh(\overrightarrow{W}\vec{G}_t + \overleftarrow{W}\overleftarrow{G}_t + b) \tag{3}$$

如圖 1 所示，我們以C^S及C^M分別代表句子 S 及記憶 M 經過兩組卷積層後的輸出，G^S及G^M分別代表句子 S 及記憶 M 經過雙向 GRU 層後的輸出。

3.2 記憶層(Memory Layer)

我們參考 MECRF 做法，使用二組雙向長短期記憶(LSTM)分別對記憶 GM 進行編碼，將時間序列訊號加入模型當中，產生輸入記憶(Input Memory)以及輸出記憶(Output Memory)， 如式(4)、(5)。

$$I_j = tanh(\overrightarrow{LSTM}(G_j^M) + \overleftarrow{LSTM}(G_j^M)) \tag{4}$$

$$O_j = tanh(\overrightarrow{LSTM}(G_j^M) + \overleftarrow{LSTM}(G_j^M)) \tag{5}$$

假設當前輸入是句子的第 t 個字元G_t^S，為了計算G_t^S與記憶當中每個元素I_j的注意力值$A_{t,\,j}$，我們將當前輸入G_t^S與輸入記憶I_j做內積運算，但是不同於 MECRF 採用 Softmax，此處我們採用 tanh 函數強化重要的記憶位置，如式(6)，其中 j ∈ [1, N]。

$$A_{t,\,j} = tanh((G_t^S)^\top I_j) \tag{6}$$

最後使用加權和來計算當前的輸出p_t，如式(7)，並結合當前輸入G_t^S，做為最後的輸出，如式(8)。

$$p_t = \sum_{j=1}^{N} A_{t,j} O_j \tag{7}$$

$$U_t = G_t^S + p_t \tag{8}$$

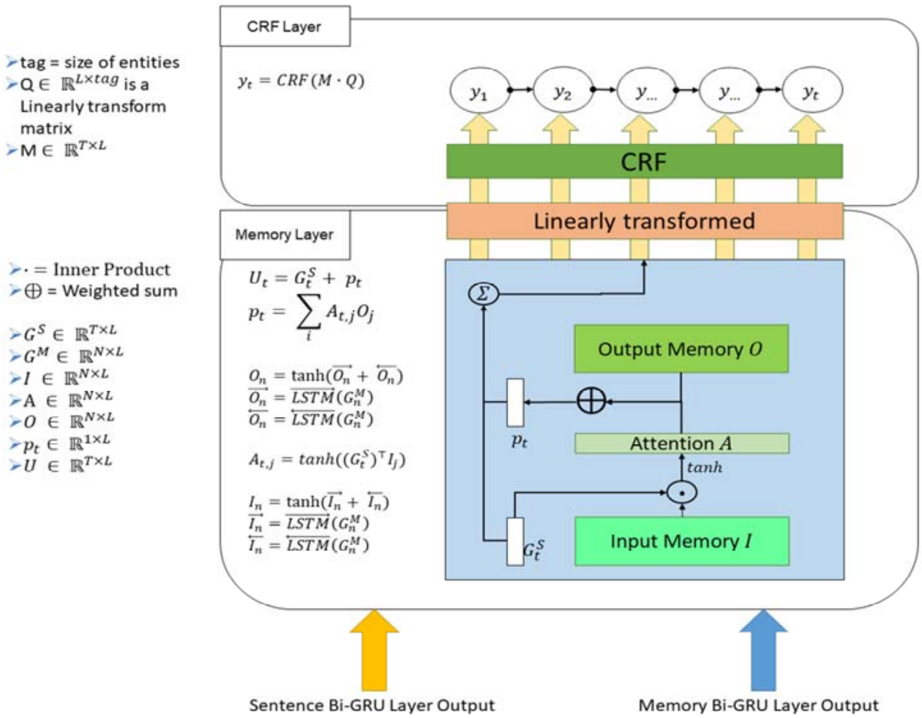


圖 2. Memory Enhanced Model with CRF output layer

Attention 的機制允許模型可以不受限制的訪問文章中短期記憶涵蓋的位置，讓我們的模型可以獲取較豐富的文章資訊。最後我們採用條件隨機場域，經由轉移矩陣考慮標記之間的依賴關係，用以增加準確率。完整架構可參考圖 2。

4. 實驗與系統效能(Experiments and System Performance)

在本章節中，我們將針對本研究所提出的各層模組效能及可調整的變數進行比較。我們使用 PerNews 及 SIGHAN-MSRA 兩組資料評估模型之效能。其中 PerNews 資料集(Chou & Chang, 2017)係以辨識社群媒體上的人名實體為主要研究目標，藉由 7053 個人名清單，自動標記出現於資料集中的人名。不同於 Chou 與 Chang 的做法僅留包含人名的句子，本研究因需要參考上下文資訊，因此在標記時不會過濾掉未含有任何實體的句子。SIGHAN-MSRA 則是學術界普遍用來評估中文斷詞與命名實體辨識工具效能的標準數據集(Levow, 2006)，本研究主要針對人名、地名、組織名進行命名實體辨識。資料集之基本統計資料如 Table 1 所示。

Table 1. Two Datasets

Dataset	Sentences	Average Characters/ per Sentence	Entity
PerNews Train	335,056	13.13	PERSON:54,338
PerNews Test	363,572	13.14	PERSON:54,546
SIGHAN-MSRA Train	141,546	14.94	PERSON:17,615 LOCATION:36,861 ORGANIZATION:20,584
SIGHAN-MSRA Test	11,679	14.45	PERSON:1,973 LOCATION:2,886 ORGANIZATION:1,331

本研究採用的標記法為 BIESO 標記法，評估方式為精準比對完整命名實體後，以常用的指標，即精確率、召回率以及 F1-Score 來進行效能的評估。模型所採用的參數如 Table 2 所示：中文字元嵌入維度 250、三層卷積層、每層 50 個 Kernel Filters、短期記憶體為 200 字元，學習率與 dropout rate 分別為 0.0005 及 0.2。

Table 2. Model Hyper-Parameters

Hyper-parameters	value
Character Embedding	250
Conv layer # filters	50
Kernel width of filters	3
Learning rate	0.0005
Dropout rate	0.2
Memory size	200

4.1 PerNews Dataset

首先我們針對本篇提出的門控多層卷積雙向 GRU 資訊表示方式，與 DS4NER 工具(Chou & Chang, 2017)所提供的基於前後字詞及首尾字詞特徵的 CRF++方法進行比較。如 Table 3 所示，DS4NER 搭配 CRF++工具的效能僅有 0.8603，而單純使用字元嵌入的 CE-MECRF 效能也僅有 0.8572 左右，顯示並非只採用記憶架構就能達到好的字詞及句子的表達方式有其重要性。我們發現多層卷積雙向 GRU 架構優於單純 CE-MECRF 效能 2.1% ，加入短期記憶的模型更有效提升 2.9%F1。

Table 3. Performance on PerNews Dataset

Models	Min/epoch	Precision	Recall	F1
DS4NER-CRF++	25*	0.9347	0.7968	0.8603
CE-MECRF	15	0.8881	0.8284	0.8572
CE-CNNs-BIGRU-CRF	25	0.9345	0.8289	0.8785
CE-CNNs-BIGRU-MECRF	65	0.9067	0.9084	0.9075

* DS4NER-CRF++為全部訓練時間

4.1.1 卷積層過濾器數量 (Fiters Number of Convolution Layer)

在此實驗中，我們調整卷積(CNN)層的過濾器的數量，比較不同過濾器數量對於效能的影響。如圖 3 所示，將過濾器數量設定為 50 的時候，效能表現最佳，過濾器數量逐漸增加的情形下，並無法顯著提升效能，且值得注意的是，越多的卷積過濾器雖因產生更多的特徵，可以得到較好的精準率，但是召回率的表現上則是逐步下滑。

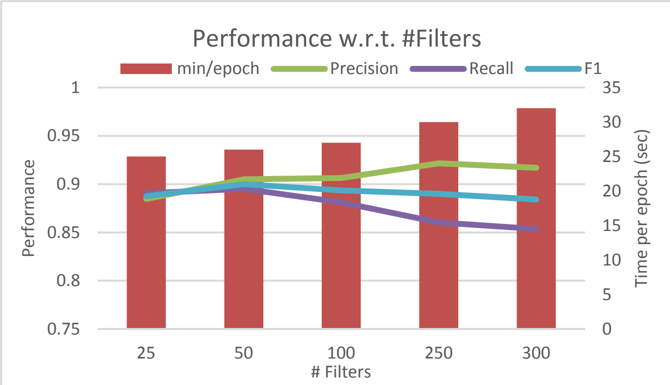


圖 3.不同過濾器數量對於效能的影響
[Figure 3. Effects of # of Filters for CNN Layer]

4.1.2 記憶體大小 (Memory Preparation Method and Memory Size)

在這個實驗中，我們不僅比較記憶體大小對效能的影響，同時也比較從文章起始做為參考上下文對於效能的影響。如圖 4所示，採用前後 B 句相較從文章起始的短期記憶方法，效能較佳。我們以 3 句、 7 句，分別對應 100、200 字元以及全文 300 字元進行實驗。由於 PerNews 為網路上之資料，當中擁有許多雜訊，因此在記憶過大的情況下，模型參考到較多的雜訊資料，效能反而有所減損，而在 200 字元記憶體時效能最佳。

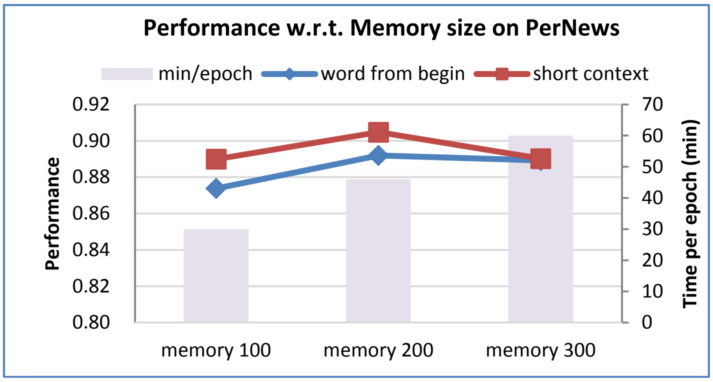


圖 4.記憶體與效能的影響
[Figure 4. Effects of Memory Size]

4.1.3 加入詞向量 (Word Embedding)

雖然堆疊卷積可以找出相鄰字元之間的關係，但是其實無法像中文詞彙詞向量那麼有意義，因此我們在基於字元的標記當中加入以當前字w_i為中心，與前後字元結合，加入適當的詞向量。如式(8)所示，我們主動加入$w_{i-1}w_i$、w_iw_{i+1}、$w_{i-2}w_{i-1}w_i$、$w_{i-1}w_iw_{i+1}$、$w_iw_{i+1}w_{i+2}$等五個詞的詞向量，若查無此詞彙，則補零向量。

$$X_i = [V_{char}(w_i),\ V_{word}(w_{i-1}w_i),\ V_{word}(w_i w_{i+1}),\ V_{word}(w_{i-2}w_{i-1}w_i),\ V_{word}(w_{i-1}w_i w_{i+1}),\ V_{word}(w_i w_{i+1} w_{i+2})] \quad (9)$$

Position	Char
Pre_2	外
Pre_1	交
Cur	部
Suf_1	昨
Suf_2	天

Kind	Combination	Words
Bichar_Pre	Pre_1 + Cur	交部
Bichar_Suf	Cur + Suf_1	部昨
Trichar_Pre	Pre_2 + Pre_1 + Cur	外交部
Trichar_Mid	Pre_1 + Cur + Suf_1	交部昨
Trichar_Suf	Cur + Suf_1 + Suf_2	部昨天

圖 5. 詞彙詞向量產生方法
[Figure 5. Illustration of Adding Word Embedding]

如圖 5，以「部」為範例，在新增的五個詞彙向量，僅有「外交部」這個詞有對應向量，其於四個字詞均以零向量取代。

在詞向量前處理中，我們採用結巴斷詞系統的精確斷詞模式，再使用 CBOW 建立 Word2vec 模型(Mikolov, Chen, Corrado & Dean 2013)，設定詞頻為至少出現 5 次，訓練 50 維的詞彙(word)的詞向量模型。

4.1.4 自動前後字詞典特徵(Automatic BAPS Dictionary-Based Features)

由於 PerNews 在原始資料中，是藉由 Chou and Chang (2017)所提出的方式進行特徵的探勘，找出 Common Before、Common After、Entity Prefix、Entity Suffix 等特徵(Support 閾值設定為 0.5)，做為 Dictionary-based Features（簡稱 BAPS），得到不錯的效能。因此我們試圖將此四類特徵各三種長度(1-gram, 2-gram, 3-gram)共 12 個特徵，經過 CNN-BiGRU-MECRF 與上述模型結合，再使用一個可由模型自動訓練的變數α (a∈[0,1])來調整嵌入向量(EMB)與 BAPS 特徵所佔的比重，經過式(10)的計算後，最後再使用條件隨機場域進行序列標記。

$$output = \alpha \cdot EMB + (1-\alpha) \cdot BAPS \quad (10)$$

換言之，BAPS 特徵也同樣經過多層卷積雙向 GRU 以及記憶網路的計算，學習得新的特徵（如圖 6 所示），最後與字元向量與前面加入的字元及詞向量表示（EMB）統整為條件隨機場域的輸入。

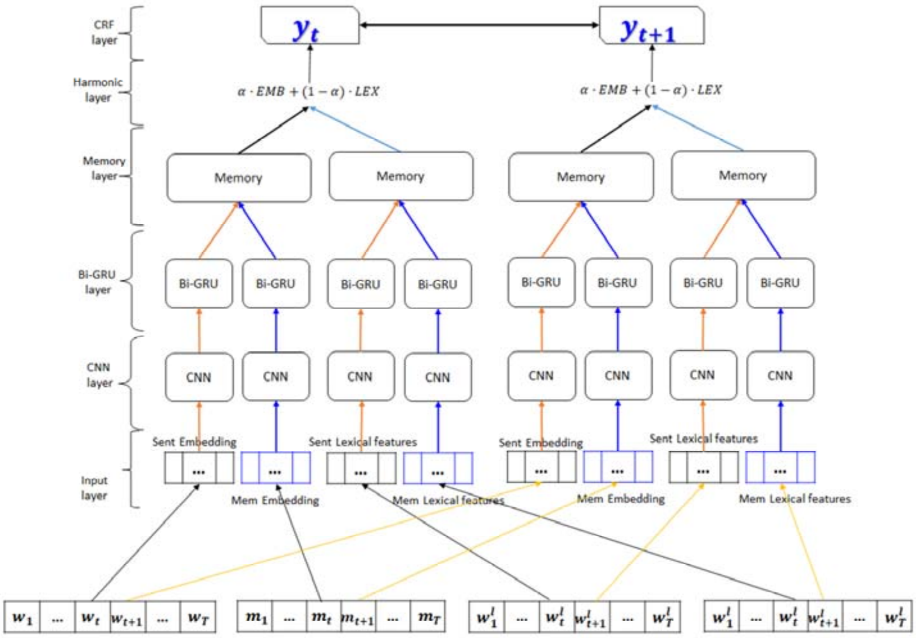


圖 6. Embedding and BAPS Hybrid Model

加入詞向量與 BAPS 特徵的效能如 Table 4 所示，兩者各別改進的幅度不大，綜合來看有 1%的進步，讓 F1 效能達到 0.9176。

Table 4. Effects of adding word embedding and dictionary-based features

Models	Min/epoch	Precision	Recall	F1
CE-CNNs-BIGRU-MECRF	65	0.9067	0.9084	0.9075
CWE-CNNs-BIGRU-MECRF	101	0.9467	0.8779	0.9110
BAPS-CNNs-BIGRU-MECRF	31	0.9134	0.7951	0.8502
Harmonic (CWE+BAPS) -CNNs-BIGRU-MECRF	137	0.9307	0.9048	0.9176

4.2 SIGHAN-MSRA 資料集

在 SIGHAN-MSRA 資料集上，實驗結果如 Table 5 所示。本研究所提出的調和(Harmonic)模型，引入詞向量以及自動前後字詞典特徵的資訊，在效能上均有不錯的表現。此外，在地名的評估效能高達 92.45%，在整體召回率也達到了 90.95%的出色表現。

Table 5. Performance on SIGHAN-MSRA

Model	PER-F	LOC-F	ORG-F	Precision	Recall	F1
Zhou-CRF (2006)	90.09	85.45	83.10	88.94	84.20	86.51
Chen-CRF (2006)	82.57	90.53	81.96	91.22	81.71	86.20
Zhou (2013)	90.69	91.90	86.19	91.86	88.75	90.28
Zhang-MEMM (2006)	**96.04**	90.34	85.90	**92.20**	90.18	**91.18**
Dong-BiLSTM-CRF (2016)	91.77	92.10	**87.30**	91.28	90.62	90.95
Liu-MECRF (2017)	91.09	91.87	83.81	89.16	90.47	89.81
Lex-CNNs-BiGRU-MECRF	81.70	75.00	67.22	85.30	67.33	75.26
CWE-CNNs-BiGRU-MECRF	91.92	90.84	84.76	89.44	90.16	89.80
Harmonic(CWE+BAPS)-CNNs-BiGRU-MECRF	92.70	**92.45**	86.31	91.34	**90.95**	91.14

5. 結論與未來展望(Conclusion and Future Work)

本研究所提出的模型，除了使用門控式多層卷積層來自動編碼鄰近字詞外，再使用 Bi-GRU 增加對上下文序列的資訊擷取功能，達到較佳的語意表示，最後使用記憶增強來加強長距離的文意擷取效能，充分獲得文章中所隱含的資訊，可找出有效的特徵做為序列標記的判斷依據。相較於 Liu 等人所提的基本 MECRF 記憶模型而言，本研究所提出的模型在社群媒體資料集中更具有穩定性及效能，而將模型應用於資料品質較好的官方資料上，同樣也有優良的效能展現。

加入 BAPS 特徵探勘所獲得的資訊雖然可能增加部份效能，是否對不同語言有同樣的效能，是後續我們想要探討的地方；另外文字在句子中的位置能否做為一種資訊，或許可以透過特別的編碼來達成。最後，由於近年來深度學習於語言領域應用日廣，對於文字的理解能否有通用的解法，也是未來努力的方向。

參考文獻(References)

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of Eighth*

Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 103-111.

Chou, C.-L. & Chang. C.-H. (2017). Mining features for web ner model construction based on distant learning. In *2017 International Conference on Asian Language Processing (IALP)*, 322-325. doi: 10.1109/IALP.2017.8300608

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, *12*, 2493-2537.

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2016). Language Modeling with Gated Convolutional Networks. CoRR abs/1612.08083

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 *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing (NLPCC 2016), and 24th International Conference on Computer Processing of Oriental Languages (ICCPOL 2016)*, 239-250.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating Nonlocal Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*, 363-370. doi: 10.3115/1219840.1219885

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. CoRR abs/1508.01991

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, 282-289.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*, 2267-2273.

Levow, G.-A. (2006). The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In *Proceedings the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *TACL, 4* (2016), 521-535. doi: 10.1162/tacl_a_00115

Liu, F., Baldwin, T., & Cohn, T. (2017). Capturing Long-range Contextual Dependencies with Memory-enhanced Conditional Random Fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP 2017)*, 555-565.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781

Sun, J. (2012). "Jieba" (Chinese for "to stutter") Chinese text segmentation: built to be the best Python Chinese word segmentation module.

Wang, C., Chen, W., & Xu, B. (2017) Named Entity Recognition with Gated Convolutional Neural Networks. In Proceedings of Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, (NLP-NABD 2017, CCL 2017), 110-121. doi: 10.1007/978-3-319-69005-6_10

Wang, C. & Xu, B. (2017). Convolutional Neural Network with Word Embeddings for Chinese Word Segmentation. CoRR abs/1711.04411

Weston, J., Chopra, S., & Bordes, A. (2014). Memory Networks. CoRR abs/1410.3916

Computational Linguistics and Chinese Language Processing
Vol. 24, No. 1, June 2019, pp. 15-38

Discovering the Latent Writing Style from Articles: A Contextualized Feature Extraction Approach

Yen-Hao Huang*, **Ting-Wei Liu***, **Ssu-Rui Lee***, **Ya-Wen Yu***,

Wan-Hsuan Lee*, **Fernando Henrique Calderon Alvarado*** and

Yi-Shin Chen*

Abstract

With the growth of the Internet, the ready accessibility and generation of online information has created the issue of determining how accurate or truthful that information is. The rapid speed of information generation makes the manual filter approach impossible; hence, there is a desire for mechanisms to automatically recognize and filter unreliable data. This research aimed to create a method for distinguishing vendor-sponsored reviews from customer product reviews using real-world online forum datasets. However, the lack of labelled sponsored reviews makes end-to-end training difficult; many existing approaches rely on lexicon-based features that may be easily manipulated by replacing word usages. To avoid this word manipulation, we derived a graph-based method for extracting latent writing style patterns. Thus, this work proposes a Contextualized Affect Representation for Implicit Style Recognition framework, namely CARISR. Transfer learning architecture was also adapted to improve the model's learning process with weakly labeled data. The proposed approach demonstrated the ability to recognize sponsored reviews through comprehensive experiments using the limited available data with 70% accuracy.

Keywords: Reliability, Transfer Learning, Writing Style, Text Classification, Natural Language Processing.

* Department of Computer Science, National Tsing Hua University
E-mail: yenhao0218@gmail.com

1. Introduction

With the popularization of the Internet and communication devices, information can be sent more quickly and widely than ever before. However, technological advances have also made it difficult to avoid incorrect information. Sponsored reviews, which have recently become a popular marketing strategy in online forums, can provide incorrect information. The intention of these articles is to give their consumers a positive impression of the product. Some advertisement companies have even begun to use sponsored reviews as a new method of promoting their commodities. Such sponsored reviews usually only provide positive information about a product. Thus, these reviews may hide the disadvantages of a product and potentially mislead consumers into making an unbeneficial purchase.

As unreliable data may contain incomplete or incorrect information, it is important to avoid them. Most of the filtering approaches on online social platforms rely on mutual reviewing from users or human-designed rules. However, no matter which approach is used, automatic filtering is still limited due to the various methods of writing sponsored reviews and how quickly information is generated. Consequently, a system to automatically identify these kinds of information has become an important issue in the information reliability research field.

In this work, we focus on recognizing the information reliability of review articles on online web platforms. Review articles are widely consumed by readers in order for them to purchase the best products. General filtering methods fail to address two main difficulties. First, current filters are easily fooled if the method only considers word-based characteristics; writers can simply avoid specific words/phrases to pass the filtering check. Second, there is a lack of defined and labeled sponsored review article data for testing reliability problems. It is difficult enough to manually collect these articles, let alone to create rules for automatically gathering them, because these articles are written by experienced writers.

To address the first issue of keywords bias, this research focused on extracting the latent writing style of review articles to avoid specific word biases found in word-level methods. The presented research proposes a Contextualized Affect Representation for Implicit Style Recognition (CARISR) method to recognize the writing styles of various reviews. The proposed CARISR consists of an unsupervised approach for generating stylistic word patterns, which condenses patterns into distributed matrix representations, and a learning-based model. Sections 4 and 5 describe the details of the stylistic patterns and model, respectively.

The biggest difference between the general methods and CARISR is that the latter defines two specific word groups, stylistic skeleton words (CW) and stylistic content words (SW), to capture the writing style information. A set of stylistic word patterns are extracted based on the constructive relationship of different stylistic skeleton words and content words in the sentence. By adopting stylistic word patterns, the experiment results show that CARISR is more robust compared to the word-based approaches, including neural network methods. In other words, the contextualized effect representation model is less susceptible to changes to specific words. Consequently, CARISR has a better ability to deal with the first challenge, that is, to detect the implicit word usages of advertisement writers.

For the second difficulty, the lack of labeled data, we defined our recognized targets as sponsored reviews (業配文), trial product reviews (產品試用文), and self-purchased product reviews (自購心得文). Since it is rare for sponsored reviews to actually be labelled as such, we introduced a similar class that is more easily obtained, called official advertisements (廣告), as the weak label concept for model pre-training. The transfer learning approach can then be applied to the target label of sponsored review.

This work proposes that the purpose of the sponsored review is more similar to official advertisements than self-purchased product reviews. This similarity allows for transfer learning to be adopted in our work. After preliminary training leveraging a large number of advertisements, the model should have the ability to classify the implicit writing style of advertisements. Further, we manually collect small amounts of sponsored review for transfer learning and fine-tune. The proposed model achieves around 70 percent accuracy and shows better robustness than the compared models, which demonstrates that our framework works successfully, even with the scarce sponsored review resources.

To shortly summarize this research, we highlight the following contributions:

- To quantify the problem of review articles' reliability, we defined different levels of reviews and collect the corresponding dataset for the training model.
- To prevent our model from being defrauded by intentional word selection, our model recognizes reliability based on the implicit writing style instead of word-level features.
- To capture the implicit writing style, we first applied graph-based pattern extraction to the review articles. Then, we designed the embedding strategy of contextual stylistic patterns for the convolutional neural network model.
- To overcome the insufficient quantity problem, we combined the weak label concept and the transfer learning approach to stabilize the learning process and improve the performance and robustness of our model.

2. Related Work

2.1 Information Reliability

Information reliability research aims to distinguish whether the given information is reliable or not. Most of the information reliability research could be consider as credibility analysis on news. The main difficulty of credibility analysis is how to find the effective features to identifying the news is reliable or not. To address the problems, the researchers attempt to extract different features, which could be categorized as the propagation-based, knowledge-based and content-based approaches.

For propagation-based approach, social media could be one major domain for news sharing, the analysis within social media relies heavily on social context features like author profiles, retweets, likes, etc. Social media rumor detection (Derczynski *et al.*, 2017) utilized conversation on Twitter to determine the veracity as RumorEval tasks. By modeling the sequence posts and behaviors on social media, researchers (Kochkina, Liakata, & Zubiaga, 2018; Ruchansky, Seo, & Liu, 2017; Volkova, Shaffer, Jang, & Hodas, 2017) proposed supervised method to detect the rumors and fake content. These approaches assume that the footprint and network of fake news are different from real news. Moreover, it has been shown that the spread speed of fake news is faster than real news (Vosoughi, Roy, & Aral, 2018). The propagation-based methods rely on social context feature; therefore, it is difficult to capture enough information for fake news detection right after the newly emerged news. Also, they are limited to social network for social context features. In contrast, this work studied reliability only on textual information, therefore, it can recognition the unreliable information in real time.

Knowledge-based method includes the tradition manual fact-checked by expert and automatic factchecking (Shi & Weninger, 2016; Shiralkar, Flammini, Menczer, & Ciampaglia, 2017; Wu, Agarwal, Li, Yang, & Yu, 2014). Several organizations, such as PolitiFact and Snopes, investigate the news and related document to report the credibility of the claim. The manual fact-checking method is time-consuming and expert oriented, which is difficult to handle the huge amount of false claim in online news media. Thus, the automated knowledge-based fact-checking system has been developed. The system will extract the claims in news content and try to match the claim to relevant data on the external knowledge base. In our work, we do not count on the external knowledge bases or web evidences; instead, we extract the stylistic features from articles to automatically capture the implicit style of unreliable article information.

Content-based methods aim to capture the keywords or writing style of malicious fabrication news from its content. The advantage of content-based methods is that it can immediately alarm the reader only from its content no matter the news is newly emerged or not. Previous works on content-based methods can be categorized into two groups by their method. One focused on the "textual content classification" (Al-Anzi & AbuZeina, 2017; Pavlinek & Podgorelec, 2017; Qu *et al.*, 2018; Wang, Luo, Li, & Wang, 2017). It classified content by "Content words", which were meaningful and different depended on the content. The other interested in "writing style recognition" (Gomez Adorno, Rios, Posadas Durán, Sidorov, & Sierra, 2018; Rexha, Kröll, Ziak, & Kern, 2018; Stamatatos, 2009) which aimed to find out the articles that have the same style but different content. These word-based methods concerned more about the "Function words" and the structure of sentence, which were often regarded as less important part before. Several research Karimi and Tang (2019); Khan, Khondaker, Iqbal, and Afroz (2019); Wang *et al*. (2018) has shown the promising result by taking advantage of machine learning technique. However, Janicka, Pszona, and Wawer (2019) address the issue that the failure of cross-domain detection, which can be interpreted as a type of overfilling on the training domain. The work conducts the experiment on four types of domain including short-text claim, full-text content. generated fake new via Amazon Mechanical Turk (AMT), and fake news on Facebook. The experiment shows that the model can fit well in the same domain, but the accuracy drops sharply when testing on the other domain.

2.2 Text Representation

To represent unique characteristics of different text documents, several features extraction methods have been proposed. Before the widespread use of the deep learning models, there are many methods relied on the hand-crafted, lexicon-based and syntactic approaches.

The hand-craft approaches are based on predefined dictionaries or linguistic resources such as the linguistic inquiry and word count (LIWC) affect lexicon (Pennebaker, Booth, & Francis, 2007). One of the advantages of using predefined dictionaries is that they are usually of high quality due to the rigorous process of labeling. However, this also presents a scalability problem as these features may not be representative of the dynamically evolving language used.

The lexicon-based approaches automatically extract the representative tokens from corpus, such as bag of word (BOW) or term frequency-inverse document frequency (TF-IDF). BOW learns the distribution of word usages to present the corpus. By integrating the n-grams consideration, the token units of BOW could be extended to n words as phrases rather than a single word to extract more high-level features. TF-IDF further introduces the statistical concept to reduce the importance of common tokens, such as "the" and "or". One of the benefits of the lexicon-based approach is that are robust to misspellings and the out of vocabulary (OOV) problems. However, it result in a extreme large size of vocabularies in memories and the curse of the dimensionality from the sparsity of vocabularies.

The syntactic approaches including part of speech (POS) parsing tree and graph-based word pattern, which considering the relation among the words. The POS parsing tree converts words by the POS tags and models the syntactic structure of sentence. The syntactic POS tree benefits the understanding for sentence, however, the POS tagging process relies on predefined dictionaries and may encountered OOV and not perform stably for specific terminologies or among different languages. The graph-based word pattern approaches (Argueta, Saravia, & Chen, 2015; Saravia, Liu, Huang, Wu, & Chen, 2018) analyze the hidden word relation by learning a word relation graph dynamically from the corpus. By adopting the graph analysis techniques, words that is important in the connection of graph structure could be extracted and used to construct the n-grams word patterns. As the word graph could present a longer connection of words than n-gram approaches, the hidden relations among words could be better preserved. The word pattern derived from graph structure learns the syntactic features of the corpus rather than n-grams key tokens; the syntactic word pattern is thus considered as a representation of the writing style. Although the method could learn the syntactic writing styles from word relation graph, however, the current approaches only focused on the English corpus. This work aims to leverage the benefits of word relation graph and propose the modification to extract syntactic writing style features from Mandarin corpus.

In the deep learning approaches, words are embedded as the vector representations by different contextual learning techniques, such as word2vec (Mikolov, Chen, Corrado, & Dean, 2013) and GloVE (Pennington, Socher, & Manning, 2014). The word vectors preserve the semantic reasoning capabilities of the word and are treated as the input feature representations to the deep learning models, such as the sequence-modeling recurrent neural network (RNN) and the convolution neural network (CNN) which focus on the local pattern extraction.

By integrating the traditional methods and the modern neural network approaches, this study proposes an approach that leverages the graph pattern features and a convolutional neural network model to identify the unreliable text information. The proposed model not only captures the textual and stylistic feature from articles but also has the adaptability for different writing styles.

3. Contextualized Affect Representation for Implicit Style Recognition

To prevent keyword bias, we studied various writing styles with a focus on frequent word usages and corresponding co-located words for each writing style. In this work, we adapted the concept of graph-based pattern extraction approaches to dynamically learn the writing style of Mandarin product review datasets. This approach has been applied in related works on emotion analysis by extracting the word patterns for each emotion. In the following sections, we highlight the adaptation of the graph-based emotion pattern approach to extract stylistic word patterns as the writing style.

The overall framework, which can be separated into stylistic pattern feature extraction (titles highlighted in orange) and model architecture (title highlighted in yellow), is shown in Figure 1. By constructing the word relation graph, the hidden word relations are preserved to enrich the stylistic words patterns in comparison to traditional lexicon-based approaches. A weighting mechanism was proposed to learn the significance of each pattern for each style.

Articles were first transformed into stylistic patterns by encoding each matched pattern and determining the corresponding score vector, which represents the article's stylistic pattern. In this work, the pattern representations were treated as the input of a neural network model for document classification based on writing style features. The details of the stylistic pattern feature extraction and model architecture are summarized in the following subsections.

4. Stylistic Pattern Features Extraction

4.1 Stylistic Graph Construction

Given a set of corpuses $\boldsymbol{C} = \{c\}$ and the sentences $\boldsymbol{S}_c$ in corpus c, the sequences of word are denoted as V_{s_c} in sentence s_c. The word graph G_c then represents the graph structure for the corpus set $\boldsymbol{C}$, such that G_c= (V_{C}, E_{C}, W_{C}). Vertices $\boldsymbol{V_C}$ is a set of nodes which represent all the word tokens v in corpus $\boldsymbol{C}$, and $A_{\boldsymbol{C}}$ is a set of arcs that represents a bi-gram relationship between each two adjacent tokens. For example, the tokenized sentence "用 _ 起來 _ 還有 _ 飾色 _ 效果 _ ， _ 給 _ 你 _ 無可取代 _ 的 _ 透亮 _ 蘋果光 _ 唷 _ ！！" could construct the following bi-gram relations: "用 → 起來", "起來 → 還有", "還有 → 飾色", ..., "蘋果光 → 唷", "唷 → ！！". Note that the under-dash "_" shows how the sentence is tokenized and the arrow "→" denotes the link relation in the word graph.

For the edge weights W, instead of initialized with binary representation, which is align with the adjacency matrix, the edge weight w_{v_i,v_j} are defined as the bi-gram probability between two word tokens v_{i} and v_{j} in order to capture the significance of link relation. The bi-gram probability is designed with a denominator of global bi-gram frequency, the frequency of all the bi-grams, rather than the degree of word node v_i or the frequency of out nodes v_{j} from node v_i. By comparing to all the bi-gram tokens, the word graph could better capture and compare the global significance for each node. Consistent to the setting of edge weight, the weighted adjacency matrix $\boldsymbol{M}$ is designed as the matrix representation of the edge weights $\boldsymbol{W}$ and defined in Definition 1.

By having the weighted mechanism, the word graph G_c could have a better ability to preserve the syntactic structure of words by a graph representation.

Definition 1 ***(Weighted Adjacency Matrix)*** *Let* $\boldsymbol{M}$ *be the weighted adjacency matrix that each entry* $M_{i,j}$ *represents the relation of word pair in the word graph* G:

$$M_{\mathrm{i,j}} = \frac{\mathrm{freq}(v_i,v_j)}{\sum_{v_k,v_l \in V, k \neq l} freq(v_k,v_l)} \tag{1}$$

where the freq() *denotes the frequency of two bi-gram words* v_i, v_j *or* v_k, v_l.

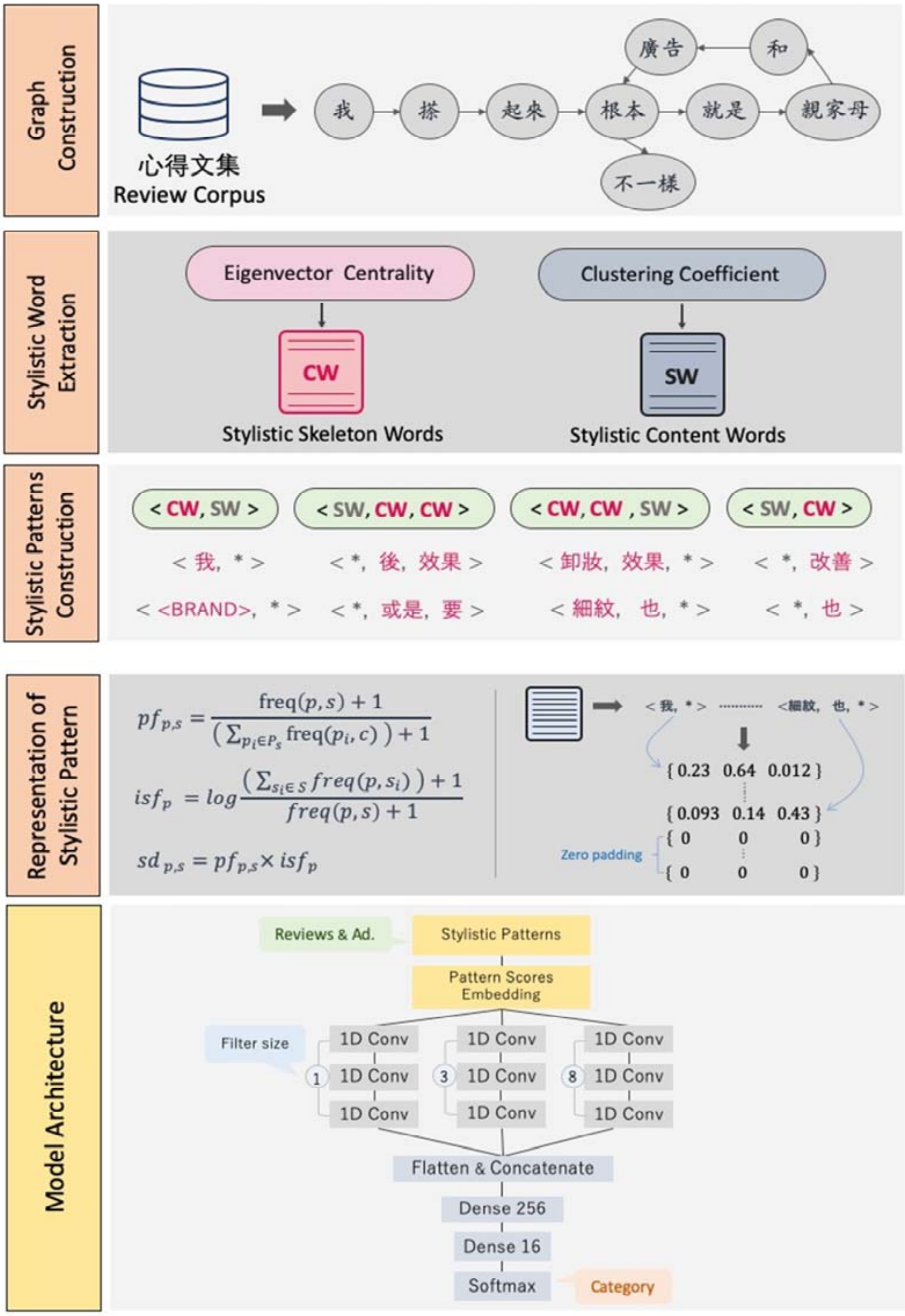


Figure 1. The framework of CARISR.

4.2 Stylistic Word Extraction

Writing styles vary from individual to individual. The idea that people utilize different distributions of words for different topics has widely been accepted in several topical methods, such as latent dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003). This work also uses this concept to extract and decompose the writing style into two elements: the stylistic skeleton and the stylistic contents. This work assumes that sentence and corpus are constructed by choosing the words of selected style to form skeleton and deciding the contents words to complete the sentence structure.

To extract the stylistic elements, two types of graph analyses—centrality and clustering—were applied to the word graph G_c. Each analysis method helps to generate a set of words: stylistic skeleton words *(CW)* (i.e., stylistic stop words) and stylistic content words *(SW)*.

4.2.1 Stylistic Skeleton

The stylistic skeleton represents the fundamental elements of word usages in a style, where such words should be widely used in all the corpuses of a given style. That is, all of the words included in the stylistic skeleton of a style should consistently appear in all of the corpuses of that style. In the structure of the graphical representation, skeleton words that represent a strong connection to other words are considered suitable candidates for stylistic skeleton words, as those words act as the fundamental nodes in the word relation graph $G_{\boldsymbol{C}}$. Inspired by Google's PageRank (Page, Brin, Motwani, & Winograd, 1999), in which nodes with high connection word nodes contribute more importance than low connection word nodes, the eigenvector centrality was selected to measure the influence of each node in $G_{\boldsymbol{C}}$.

Definition 2 ***(Eigenvector Centrality)*** *The eigenvector centrality is calculated as*:

$$e_i = \frac{1}{\lambda} \sum_{j \in \boldsymbol{V_C}} M_{i,j}\, e_j \tag{2}$$

where λ is a proportionality factor and e_i is the centrality score of word node v_i. Let λ be the corresponding eigenvalue, the equation could be rewritten into vector form Me = λe, where e is the eigenvector of M.

A word is selected as a connecter word if its eigenvector centrality e_i is higher than the empirically defined threshold θ_{eig} to ensure the quality of the high connectivity word. The higher the centrality e_i of a word v_i, the more important the word is in the graph $G_{\boldsymbol{C}}$. By the centrality measurement, a set of connector words with both high frequency and connectivity to other high-rank nodes are extracted from the word relation graph G_C and considered stylistic skeleton words $\boldsymbol{CW}$, such that $\boldsymbol{CW} = \{cw\,|e_{cw} > \theta_{eig}\}, cw \in V_C$. The examples of the stylistic skeleton words in this task (the makeup advertisement dataset) were as follows: “我,” “的,” “因為,” “肌膚,” and “特別.” The extracted stylistic skeleton words not only contained numerous traditional stopwords but also style-specific words, which are known as stylistic stopwords.

4.2.2 Stylistic Content

The stylistic contents represent frequently appearing topics within a style, where topics could be formed by several separated words (i.e., LDA) or continuous word sequences. Apart from the skeleton, a topic could be presented by using the words in different ways; however, to represent the similar semantics of the topic, the topic words are generally interchangeable. For example, in the makeup advertisement dataset, there are several ways to describe the product’s effect on skin care, such as “能 _ 有效 _ 保養 _ 肌膚,”“保護 _ 嫩白 _ 肌膚,” or “擁有 _ 水嫩 _ 臉頰.” In the above example, some word tokens can be changed while keeping the meaning the same, such as “保養” to “保護” or “嫩白” to “水嫩” and so on.

To capture the stylistic content cues, this work focuses on interchangeable word usages. By converting the style corpus in the word relation graph, the cross connections between these interchangeable word nodes are discovered. Such stylistic content word nodes tend to cluster with other nodes that share this or similar concepts. The clustering behavior in the graph can be measured by a graph analysis factor, namely the *clustering coefficient*, which determines how a node interconnects with its neighbor nodes. This work therefore applied the *clustering coefficient* to dynamically extract the stylistic contents, as shown below.

Definition 3 ***(Clustering Coefficient)*** *The clustering coefficient is defined by clustering coefficient as:*

$$cl_i = \frac{\sum_{j\neq i,k\neq j,k\neq i,} M_{i,j}\times M_{i,k}\times M_{j,k}}{\sum_{j\neq i,k\neq j,k\neq i,} M_{i,j}\times M_{i,k}} \times \frac{1}{|\boldsymbol{V_C}|} \tag{3}$$

where cl_i *denotes the average clustering coefficient of node* v_i.

Similarly, the word nodes v_i were also filtered by a predefined threshold θ_{tri} for clustering coefficient cl_i to ensure the clustering quality. During the computing process of clustering coefficient cl_i for each node v_i, we discovered that there were many nodes with high coefficients. However, many of them belonged to local mini-clusters in which the degree of node was too small, resulting in too many specific words for stylistic contents. A post-filtering step was then applied to remove the local mini-cluster and small cluster words based on the number of triangles tri_i of the word nodes v_i, where less node triangles indicated a smaller cluster. With the post-filtering step, a set of qualified stylistic content words *SW* were retrieved, such that $\boldsymbol{SW}$ = $\{sw \mid cl_{sw} > \theta_{cl},\ tri_{sw} > \theta_{tri}\}$, $sw \in \boldsymbol{V_C}$, where θ_{tri} denotes the empirical threshold for the number of triangles for the word node. Some examples of stylistic content words in this task were “森林系,” “世界級,” “黏稠度,” and “可愛感.”

4.3 Stylistic Pattern Construction

With the extracted stylistic skeleton and stylistic content words, this step aimed to construct the stylistic word pattern template. The stylistic word pattern is designed to capture hidden word usages in a writing style. For a word pattern, the length *l* of the pattern can be dynamic; that is, there may exist a longer stylistic word pattern (i.e., slogans) or a shorter one (i.e., topic tokens). In this work, a short length was adapted, as a longer word pattern may be difficult to match in a real-world case.

To construct the word pattern templates $\boldsymbol{P} = \{p\}$, the permutation of stylistic skeleton and content words, *CW* and *SW*, were adopted in our work using the rules below:

- The stylistic skeleton words are required to exist in the pattern at any position as such words have the top connectivity in the corpus.
- A word pattern could contain more than one skeleton words.

For example, in pattern length $l = 3$, each pattern feature is composed of an arbitrary permutation, such as “cw sw cw” or “cw sw sw,” from the set of *CW* and *SW*. The word patterns are then used to search the corpus set $\boldsymbol{C}$ to retrieve the pattern frequency. The word patterns that belongs to last 20% infrequent patterns are dropped, as they are not general enough.

Instead of utilizing the word pattern by exact matching (bag-of-word matching) as n-gram does, this work adopts a flexible representation to increase the versatility of the pattern template due to the issue of easily overfitting for n-grams and pattern size consumption. Compared to the stylistic skeleton words, the stylistic content words are relatively easier to update or replace (i.e., develop new terms) as these are determined by the clustering coefficient, which captures interchangeable words. With respect to the stylistic content characteristics, various words that may be beyond the knowledge coverage of the training dataset could be used to describe a topic. Therefore, flexible representation was designed and performed by replacing the *SW* in the word pattern with a placeholder <*>, which means any token could be considered in the stylistic patterns during the matching process (i.e., “我 <*> 肌膚”, “特別 <*> 的”).

The flexibility of the pattern (the wildcard representation <*>) enables our model to possess robust generalization ability, which increases pattern coverage for dealing with out-of-vocabulary words and slang or coded words used in specific domains when extracting features during testing. The complete steps for stylistic word extraction and stylistic pattern construction are formally summarized in Algorithm 1.

Algorithm 1 Stylistic Pattern Features Extraction Algorithm

Calculate eigenvector centrality (e) and clustering coefficient (cl) for topic graph.

Set θ_{eig}, θ_{cl}, θ_{tri} thresholds of centrality, clustering coefficient and number of triangles.
CW← a set of stylistic skeleton words
TW← a set of stylistic content words
for all node v in **V do**
  tri_v= number of triangles for v
  if $e_v > \theta_{eig}$ **then**
    CW← v **end if if** $cl_v > \theta_{cl}$ **and** $tri_v > \theta_{tri}$ **then**
    SW← v
  end if
end for
Construct patterns **P** with the permutation of stylistic skeleton words and content words.
for all pattern p in **P do**
  p = Replace the sw with wildcard (<*>) from p
end for

4.4 Representation of Stylistic Pattern

With the stylistic word pattern, it is critical that how to transform a set of patterns to features for the classification. One of the traditional ways is to present the word pattern as a set of bag-of-patterns with the frequency or normalized frequency (probability of occurrence) as the numerical features. However, such bag-of-pattern representations limited in the current state-of-the-art deep neural network (DNN) models, which applied several word embedding techniques to present the hidden information for a word. Such embedding features are very flexible which could be utilized not only in traditional classifiers (i.e. support vector machine (SVM) or random forest), but also the DNN models.

Inspired from it, this work aims to proposed a flexible numerical vector representation for the extracted word patterns in a pre-training manner which could perform as the initialized parameters for the classification models. The numerical representation is designed to leverage the uniqueness of each word pattern for each label, which is the style in this work. The uniqueness of the pattern for different labels is calculated by a weighting schema, namely *identical stylistic degree*. Formally, given a set of corpuses $\boldsymbol{C} = \{c\}$ and a set of possible style $\boldsymbol{S} = \{s\}$, where each corpus *c* belongs to a style *s*, the identical stylistic degree is defined by three components, which are *pattern frequency*, *inverse style frequency*.

Definition 4 ***(Pattern Frequency)*** *The pattern frequency pf is defined as:*

$$pf_{p,s} = log\,\frac{\mathrm{freq}(p,s)+1}{1+\sum_{p_i \in \boldsymbol{P_s}} \mathrm{freq}(p_i,s)} \tag{4}$$

where $\mathrm{freq}(p,s)$ *represents the frequency of the pattern p in the style s, and* $pf_{p,s}$ *is the logarithmic scaled frequency of p in all the articles of the style s.*

Pattern frequency is designed to capture the frequently appeared word pattern under the assumption that the more a pattern exists in the corpus of a style, the more important the pattern is. As the frequency is dramatically different from pattern to pattern, the scale of the $\mathrm{freq}(p,s)$ score may encounter biased due to the large frequency gap. A logarithm function is thus applied to avoid the identical stylistic degree dominated by pattern frequency.

Definition 5 ***(Inverse Style Frequency)*** *The inverse style frequency* isf *is computed as:*

$$isf_p = log\,\frac{1+\sum_{s_i \in \boldsymbol{S}} \mathrm{freq}(p,s_i)}{\mathrm{freq}(p,s)+1} \tag{5}$$

where isf_p *is the measurement of the rareness of the pattern p in all articles.*

The inverse style frequency aims to decrease the importance for the commonly appeared pattern among many styles. The traditional inverse document frequency in TF-IDF is designed to examine whether the pattern exist in how many styles. However, the pattern frequency in a style is able to be treated as the intensity of the pattern existence. This work then refines the inverse style frequency by introducing the pattern frequency as indicator to calculate the cross styles uniqueness.

Finally, the uniqueness of each stylistic pattern could be presented by the identical stylistic degree as below.

Definition 6 ***(Identical Stylistic Degree)*** *The identical stylistic degree sd is calculated as:*

$$sd_{p,s} = pf_{p,s} \times isf_p \tag{6}$$

where $sd_{p,s}$ *is the identical stylistic degree that represents the importance of the pattern* p *to the style s.*

With the identical stylistic degree $sd_{p,s}$, it is able to quantify the uniqueness of each stylistic word pattern p for a style s. The stylistic pattern p is then able to present in a vectorized form $X_p = | sd_{p,s} |$, $X_p \in R^{|\boldsymbol{S}|}$, namely stylistic pattern embeddings, where each component represents the identical stylistic degree $sd_{p,s}$ of pattern p for a style s. The flexibility of the proposed identical stylistic degree also allows the weighting schema to be extended when the number of styles $|\boldsymbol{S}|$ is increased.

5. Model Training

In this section, we describe the classification model and the transfer learning procedure.

5.1 Model Architecture

Due to the well performance of Convolutional Neural Network architecture on several text classification tasks in the past, CARISR was based on Multi-layer ConvNet (Kim, 2014) architecture, as shown in the bottom of Figure 1. Consider a set of corpuses $C = \{c_1, c_2, \dots, c_n, \dots c_N\}$, where $n \in [1, N]$. Each article c_n was transformed into pattern degree matrix X_n based on the stylistic pattern embedding described in previous section.

$$X_n = PatternEmbedding(c_n), \text{where } X_n \in R^{L \times |\boldsymbol{C}|} \tag{7}$$

where L denotes the parameter as the threshold for the maximum number of patterns for an article, and $|\boldsymbol{C}|$ denotes the number of categories, respectively. If the number pattern for an article is less than L, it will be filled with zero as pattern scores. For the sake of brevity, we used X to present single instance X_n. Each entry $X_{i,j}$ in the pattern degree matrix X represented identical stylistic degree for pattern i in category j, where $i \in [1, |\boldsymbol{C}|], j \in [1, L]$.

X is following fed into three paths which are composed by 1-D convolutional layer with different filter size of 1, 3, and 8. The output is passed through a ReLU activation function (Nair & Hinton, 2010) that produces a feature map. A 1-D max pooling layer of size 3 is then applied to each feature map.

$$a_i = ReLU(conv(X, filter\_size = i)) \tag{8}$$

$$\widehat{a_i} = MaxPooling(a_i) \tag{9}$$

the above two steps are simplified as following equation:

$$\widehat{a_i} = conv\_block(X, i) \tag{10}$$

where i denotes filter size. Stacked with three $conv\_block$, the results were concatenated together and passed through two fully connected layers of dimensions 256 and 16 in order.

$$a = \widehat{a_1} \oplus \widehat{a_3} \oplus \widehat{a_8} \tag{11}$$

$$d_1 = ReLU(W_a a + b_a) \tag{12}$$

$$\text{Classification: } s = softmax(W_d d_1 + b_d) \tag{13}$$

where $\oplus$ denotes the concatenate operation, $\widehat{a_i}$ is the output of stacked block which kernel size is i. We used softmax to get the probability of each category and used cross entropy as loss function. In order to prevent overfitting to training data, Dropout was applied to convolution layers and fully connected layers. The corresponding dropout rate is 0.5 and 0.7. The L2 regularization is also applied in the loss function, and the coefficient is 0.05. We chose a batch size of 64 and trained for 12 epochs using Adam optimizer (Kingma & Ba, 2014). We used Keras (Chollet *et al.*, 2015) to implement the CARISR architecture.

5.2 Transfer Learning

Due to the difficulty of collecting labelled sponsored reviews and self-purchased product reviews, a limited dataset was available to train the classifier to distinguish sponsored reviews from self-purchased product reviews. Inspired by the idea of transfer learning, we predicted that the flexibility of the proposed stylistic patterns could enable the proposed model to be transferable. This research thus proposes a two-stage training process to recognize sponsored reviews.

In the first stage, a large amount of advertisement and product review data were collected as weak label data to pre-train the CARISR model. In terms of writing styles, advertisements are designed to highlight the features of sale products, while sponsored reviews are written in a manner similar to trial reviews. However, sponsored reviews are considered a special kind of advertisement, as they aim to both introduce the product and spotlight it. More specifically, both advertisements and sponsored reviews have the same objective, which is to advertise the product in a positive manner. In other words, the model could learn the diverse writing styles of advertisements in the early stages (learning from advertisement) through the weak label pre-trained procedure.

In the second stage, the transfer learning concept was applied to fine-tune the pre-trained model with what little sponsored review data were available. Having the prior knowledge of the advertisement writing style, the model could more easily learn to distinguish sponsored reviews. To fine-tune it, the parameters of CNN blocks were fixed, and the first fully connected layer in CARISR was taken as the feature vector of articles. The feature vector was fed into another fully connected layer to examine the transformation from feature vector to classification result. This approach allows CARISR to distinguish sponsored reviews from true product reviews.

In this two-stage transfer learning process, the model's feature representation improved thanks to pre-training with a large amount of weak label data. It learned to distinguish the writing style of sponsored reviews and product reviews through fine-tuning with the small amount of true label data available. Based on the training process, we predict that even with the lack of true labeled data, the model could still perform well and avoid overfitting.

6. Experiments

6.1 Data

To distinguish the sponsored and product review, this research utilized the transfer learning concept which leveraged user reviews and advertisement articles as pre-training corpus and fine-tune the model with sponsored and self-purchased product reviews. For the entire training process, two datasets are collected and introduced below.

The first dataset was collected from UrCosme, a famous makeup product review website in Taiwan, with three classes *Self-purchased product review*, *Trial product review*, and *Advertisement*, where the three classes are tagged and verified by UrCosme. It has total 194,099 makeup reviews from 17,006 users from 2015 to 2018 June and includes 22,094 products and 4,594 articles from 498 brands.

The second dataset was from PIXNET, an online social blog in Taiwan, makeup product-related articles are collected with three classes Self-purchased product review, Trial product review, and the target Sponsored review. Since there are no article tags provided from PIXNET, several rules are defined for identifying the three classes. Firstly, the Sponsored review are the articles which contain the URL links with specific blogger's identification tokens. To trace the web reference from which bloggers to the product web page, this kind of URLs are widely been used to record the number of clicks and make profits to the bloggers. The text content from articles with specific URLs are collected with the Sponsored review label. Second, based on matching the keywords, "邀稿" and "試用", to label the Trial product review and other normal product reviews are labeled as Self-purchased product review. After categorizing the articles, we manually pick 125 articles from each category as the PIXNET dataset and cross valid the dataset with 5 experts. To prevent our model learned from the specific contents, all the clues (including URLs and keywords, tokens that have used to create labels) are removed in advance.

Due to the lack of the sponsored review, the UrCosme dataset is considered as the weak label dataset for the main task, the classification of sponsored and product review. The PIXNET dataset is treated as the ground truth dataset as it is labeled by manual efforts. The detail data distribution of two datasets are shown in Table 1 and Table 2. The experiment 6.3 takes the training part of the UrCosme dataset for model pre-training but evaluates on the testing part of PIXNET dataset. In experiment 6.4, the completed PIXNET dataset is involved for evaluating the pre-training model from UrCosme dataset. For experiment 6.5, the PIXNET dataset is down sampled following the ratio 4:1 for fine-tuning and evaluating.

Table 1. The data distribution of UrCosme dataset.

	Total	**Training**	**Testing**
Advertisement	9,681	9,681	2,423
Trial product review	87,508	10,000	2,423
Self-purchased product review	106,591	10,000	2,423

Table 2. The data distribution of manual labeled PIXNET dataset.

	Total
Sponsored review	125
Trial product review	125
Self-purchased product review	125

6.2 Baseline Methods

To represent a text corpus, the term frequency-inverse document frequency (TF-IDF) has been widely used in several text classification tasks. It could automatically learn the important n-grams from the corpus and present the corpus based on the extracted important n-grams. Represented by the TF-IDF features, all the articles were transformed into TF-IDF feature vector with 2500 dimensions for the extraction of the important n-grams.

In deep neural network (DNN) approaches, a text corpus is frequently represented by a sequence of the word vectors, namely *word embeddings*. The word embeddings could be either provided by a pre-trained word vectors or derived by the DNN models during the training procedure. In this work, a pretrained 400 dimensions word vector from *YZU NLP Lab*[1], trained from traditional Mandarin Wikipedia, were applied as initialized representation to present the words. The word embeddings were set as trainable to be fine-tuned in the learning procedure.

For the classification model, both traditional model and DNN model were applied in our

[1] http://nlp.innobic.yzu.edu.tw/demo/word-embedding.html

work, which were the Logistic Regression (LR) model and the Long Short-term Memory (LSTM) model. The LR model learned a specific weight for each dimension of the features, which could provide a more interpretable explanation for analysis. For DNN models, the text-CNN and LSTM were applied in the experiments. The text-CNN (Kim, 2014) considers local word features by *n*-gram windows. By adopting multiple convolutional layer, model could summarize the local word features and representation the corpus. This work set the filter size of convolution layer as 3, stacked 3 convolution layers and following with 512,128 dense layers for feature summary. The LSTM model takes the input word sequence in a word by word manner and models the words relation step by step. In this work, the bi-directional LSTM with attention mechanism was applied which achieved several state-of-the-art performance for many NLP tasks. The LSTM model was connected with a 128-dimension fully connected layer for feature summary. For two DNN models, the categorical predictions were done by the Softmax activation function for feature summaries.

6.3 Weak Label Classification Training

In the first training stage, all of the models were trained to distinguish the three different classes with the UrCosme dataset as weak label pre-training for the main task, which was the classification of sponsored and product reviews. After the model pre-training, the testing data from UrCosme was applied to evaluate the pre-training performance, the results of which are shown in Table 3. Overall, the proposed CARISR did not have the best performance in the first stage of the training process compared to the TF-IDF baseline method and LSTM-based models. However, after analyzing the weight of the model, we observed that the baseline method result was easily influenced by specific keywords. An example from a real article is discussed below:

感謝 UrCosme 與 SK-II，讓我參與「超肌因鑽光淨白精華」新品活動！

超肌因鑽光淨白精華 0.7ml x 28 包使用方式

‧於清潔肌膚後，先使用 SK-II 青春露調理肌質，有效提升細滑度、緊緻度、抗皺度、白皙度、光澤度等五大美肌度。

‧ 接著 ...... 乳白色精華無特別香氣，它使肌膚好吸收無黏膩，說實在的，當每晚保養擦上精華後，我都覺得肌膚看起來變得平滑、有光澤、膚質超好的，總覺得它有美肌般的效果！連續使用幾天，肌膚的黯沉、泛黃有改善，轉為明亮、光澤度大大提升，真心滿意，會想買正貨！

Thanks for UrCosme and SK-II for inviting me to join this campaign!

*How to use SK-II Facial Treatment Essence 0.7ml * 28*

After cleaning the face, apply SK-II Facial Treatment Essence can keep your face moisturized, brighten and firming.

then… …it makes my skin without stinging, literally, once applied the essence, it spreads easily and gets absorb quickly into the skin, besides, my skin felt moisturized without any greasy feeling. Continuing using for 2 weeks, my skin feels more brighten and firmer. I am really satisfied with this product and will order again once I run out!

The example articled was a trial product review, which it was correctly classified as by the baseline models but was incorrectly classified as an advertisement by the CARISR model. Although this article was misclassified as an advertisement, the writing style of the article showed more similarity to an advertisement than a real review by human judgement. By analyzing the weight of each term in the LR model, the result showed that the model relied on some specific terms, such as activity (活動), satisfy (滿意), and invite (邀請). In this example, the model would be easily misled by malicious writers due to these specific terms.

Based on this example, although the accuracy of the CARISR model result was lower, it gave greater consideration to the relation between word structures in the article as a whole. The following experiment shows that the CARISR model was better able to resist the influence of specific terms.

Table 3. The classification result of four methods on UrCosme dataset.

Method	**Avg.F1-Score**	**Ad.F1-Score**
TF-IDF	0.79	0.97
text-CNN	0.79	0.98
bi-LSTM-attention	0.82	0.98
CARISR	0.70	0.97

6.4 Sponsored Review Testing

The pre-trained models were evaluated with the testing data from the weak labeled UrCosme dataset discussed in the previous section. The pre-trained models were evaluated with the human-labeled dataset; that is, the reviews from PIXNET were used as testing data with the advertisement label in UrCosme replaced by sponsored review label. As shown in Figure 2, although the baseline models had better performance using the pre-trained settings, they performed worse than CARISR using the PIXNET dataset. More importantly, in the classification of sponsored reviews, baseline methods could not successfully differentiate sponsored reviews. This indicates that the baseline models had a good ability to learn but were hampered by the overfitting issue when using the training dataset. The main reason for this was that the baseline methods relied heavily on specific terms as clues, which resulted in the models not being general enough to apply to different testing data, even data from the same domain dataset (in this task, both were sponsored makeup reviews). Instead, CARISR leveraged the stylistic patterns to keep the features of sentence structure and writing style rather than only specific keywords or n-grams. Therefore, even if the testing dataset was slightly changed, the model was still able to determine the advertisement writing style.

In real-world sponsored reviews, malicious writers usually pretend that the advertisement is a self-purchased product review. Many words used in commercial reviews usually appear in self-purchased product reviews; therefore, it is easy for them to avoid detection if the model relies heavily on specific terms or baseline methods. The proposed model, CARISR, was better able to avoid this problem, making it more suitable to real-world situations.

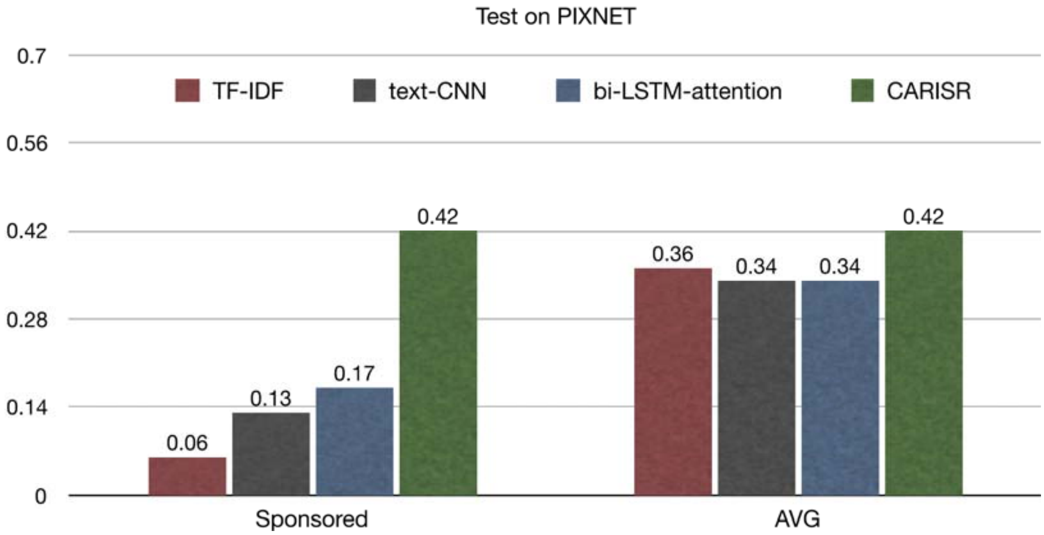


Figure 2. Comparison of TF-IDF, text-CNN, bi-LSTM-attention, and CARISR when applied to the PIXNET dataset. AVG is the average F1-score for all three categories, and Sponsored is the F1-score for sponsored reviews.

6.5 Transfer Learning with Sponsored Reviews

According to the classification results presented in the previous section, CARISR demonstrated the ability to recognize the latent writing styles of sponsored articles. Transfer learning was applied to fine-tune the DNN models to boost its performance based on a small number of manually collected sponsored reviews on PIXNET. One-fifth of the PIXNET dataset (25 samples for each class) was kept for the final testing, and the rest of the data were utilized for fine-tuning (100 samples for each class). Note that the TF-IDF model was excluded from this section, as it is not able to perform standard transfer learning based on the

TF-IDF and LR algorithms. The experimental result, labelled Transfer-3, is shown in Figure 3.

All three of the tested models manifested better performance after adjusting the parameters using transfer learning. For three-label classification, the text-CNN, bi-LSTM-attention and CARISR had F1-scores of 0.21, 0.47 and 0.51, respectively. Furthermore, our analysis found that a large percentage of collected sponsored reviews were very similar to advertisements. This may be the reason why the CARISR-Trans3 did not perform as well as expected.

Therefore, we conducted another experiment that only used sponsored reviews and self-purchased product reviews, as checked by humans, to build a binary classification model. As shown in Figure 3, with the application of two-category transfer learning (Transfer-2), the CARISR F1-score was improved to 0.70 and outperformed the bi-LSTM-attention by 0.07 points.

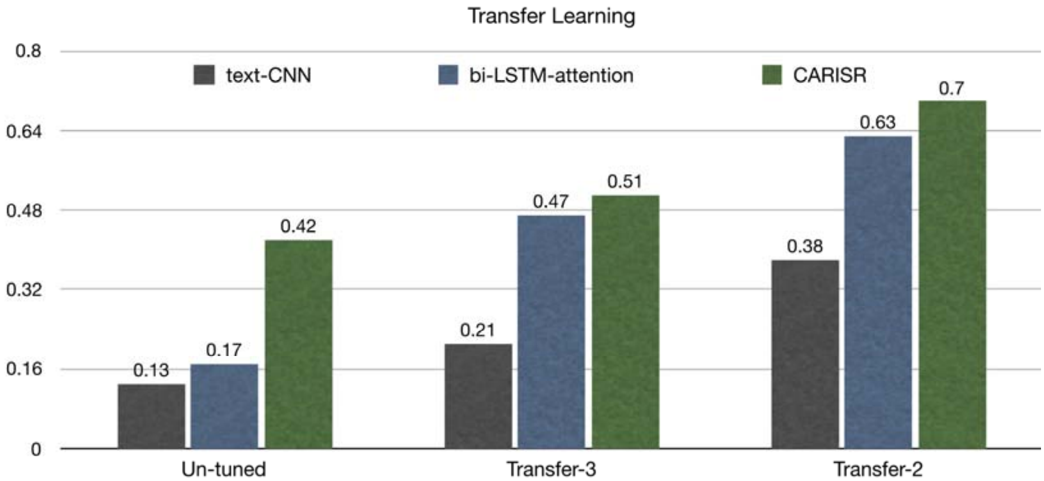


Figure 3. Comparison between original method and transfer learning. Transfer-3 indicates the result of the models after fine-tuning using three categories: sponsored, trial product, and self-purchased product review. Transfer-2 shows the results of the models after fine-tuning with only sponsored and self-purchased product reviews.

7. Conclusion

This research mainly focused on quantifying the reliability problem that results from sponsored articles on popular Mandarin forums or websites. To address the problem with limited labeled data, we first proposed a framework, CARISR, that combines weak label and transfer learning methods. CARISR can learned implicit writing styles from weak label data, and it can be further improved by transfer learning with minimal amounts of manually labelled data. Thanks to its graph-based feature, CARISR is not only more robust, but it also has better generalization compared to the traditional token-based features. Experimental results showed that our model can correctly recognize around 70% of sponsored articles from the human-labeled dataset.

Our work provides a new perspective on and further improvement to reliability tasks. In the future, we plan to merge graph-based and semantic features to capture more underlying meaning in context. Meanwhile, the enrichment of stylistic word patterns could also improve model comprehension.

References

Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced arabic text classification using cosine similarity and latent semantic indexing. *Journal of King Saud University-Computer and Information Sciences*, *29*(2), 189-195. doi: 10.1016/j.jksuci.2016.04.001

Argueta, C., Saravia, E., & Chen, Y.-S. (2015). Unsupervised graph-based patterns extraction for emotion classification. In *Proceedings of the 2015 ieee/acm international conference on advances in social networks analysis and mining 2015*, 336-341. doi: 10.1145/2808797.2809419

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Chollet, F. et al. (2015). *Keras*. https://keras.io.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In arXiv preprint arXiv:1704.05972.

Gomez Adorno, H. M., Rios, G., Posadas Durán, J. P., Sidorov, G., & Sierra, G. (2018). Stylometrybased approach for detecting writing style changes in literary texts. *Computación y Sistemas*, *22*(1), 47-53. doi: 10.13053/CyS-22-1-2882

Janicka, M., Pszona, M., & Wawer, A. (2019). Cross-domain failures of fake news detection. *Computación y Sistemas*, *23*(3), 1089-1097. doi: 10.13053/CyS-23-3-3281

Karimi, H., & Tang, J. (2019). Learning hierarchical discourse-level structure for fake news detection. In arXiv preprint arXiv:1903.07389.

Khan, J. Y., Khondaker, M. T. I., Iqbal, A., & Afroz, S. (2019). A benchmark study on machine learning methods for fake news detection. In arXiv preprint arXiv:1905.04749.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In arXiv preprint arXiv:1408.5882.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In arXiv preprint arXiv:1412.6980.

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). All-in-one: Multi-task learning for rumour verification. In arXiv preprint arXiv:1806.03713.

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

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (icml-10)*, 807-814.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. (Technical Report No. 1999-66). Stanford InfoLab. Previous number = SIDLWP-1999-0120. Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/422/

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and lda topic models. *Expert Systems with Applications*, *80*, 83-93. doi: 10.1016/j.eswa.2017.03.020

Pennebaker, J., Booth, R., & Francis, M. (2007). Linguistic inquiry and word count: Liwc [computer software]. Austin, TX: liwc. net.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 1532-1543. doi: 10.3115/v1/D14-1162

Qu, Z., Song, X., Zheng, S., Wang, X., Song, X., & Li, Z. (2018). Improved bayes method based on TF-IDF feature and grade factor feature for chinese information classification. In *Proceedings of 2018 ieee international conference on Big data and smart computing (bigcomp)*, 677-680. doi: 10.1109/BigComp.2018.00124

Rexha, A., Kröll, M., Ziak, H., & Kern, R. (2018). Authorship identification of documents with high content similarity. *Scientometrics*, *115*(1), 223–237. doi: 10.1007/s11192-018-2661-6

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 acm on conference on information and knowledge management*, 797-806. doi: 10.1145/3132847.3132877

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018). Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 3687-3697. doi: 10.18653/v1/d18-1404

Shi, B., & Weninger, T. (2016). Fact checking in heterogeneous information networks. In *Proceedings of the 25th international conference companion on world wide web*, 101-102. doi: 10.1145/2872518.2889354

Shiralkar, P., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2017). Finding streams in knowledge graphs to support fact checking. In *Proceedings of 2017 ieee international conference on data mining (icdm)*, 859-864. doi: 10.1109/ICDM.2017.105

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556. doi: 10.1002/asi.21001

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th annual meeting of the association for computational linguistics*,volume 2: Short papers, 647-653. doi: 10.18653/v1/P17-2102

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. doi : 10.1126/science.aap9559

Wang, T., Luo, T., Li, J., & Wang, C. (2017). Reasearch on feature mapping based on labels information in multi-label text classification. In *Proceedings of 2017 7th ieee international conference on Electronics information and emergency communication (iceiec)*, 452-456. doi: 10.1109/ICEIEC.2017.8076603

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., … Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849-857. doi :10.1145/3219819.3219903

Wu, Y., Agarwal, P. K., Li, C., Yang, J., & Yu, C. (2014). Toward computational fact-checking. *Proceedings of the VLDB Endowment*, *7*(7), 589-600. doi: 10.14778/2732286.2732295

UrCosme. Retrieved July 18, 2018, from https://www.urcosme.com/

PIXNET. Retrieved July 18, 2018, from https://www.pixnet.net/

Computational Linguistics and Chinese Language Processing
Vol. 24, No. 1, June 2019, pp. 39-50

探究端對端混合模型架構於華語語音辨識

An Investigation of Hybrid CTC-Attention Modeling in Mandarin Speech Recognition

張修瑞*、趙偉成*、羅天宏*、陳柏琳*

Hsiu-Jui Chang, Wei-Cheng Chao, Tien-Hong Lo, Berlin Chen

摘要

近年來端對端(End-to-End)語音辨識的出現，簡化了許多傳統語音辨識的繁複流程。端對端語音辨識中，最主要的模型架構分別為連結時序分類(Connectionist Temporal Classification, CTC)與注意力模型(Attention Model)。本論文嘗試結合上述兩種模型架構(即 CTC-Attention 混合模型)於華語會議語音辨識之使用，以期能進一步提升語音辨識的效能。為此，我們分析模型結合時混合權重調整的影響，並進一步探究 CTC-Attention 混合模型對於短句的辨識效果。在中文會議語料的實驗結果顯示，相較於傳統語音辨識的 TDNN-LFMMI 模型，CTC-Attention 混合模型在語句較短時，可具有較好的一般化能力(Generalization)。

Abstract

The recent emergence of end-to-end automatic speech recognition (ASR) frameworks has streamlined the complicated modeling procedures of ASR systems in contrast to the conventional deep neural network-hidden Markov (DNN-HMM) ASR systems. Among the most popular end-to-end ASR approaches are the connectionist temporal classification (CTC) and the attention-based encoder-decoder model (Attention Model). In this paper, we explore the utility of combining CTC and the attention model in an attempt to yield better ASR performance. we also analyze the impact of the combination weight and the

* 國立台灣師範大學資訊工程研究所
Department of Computer Science and Information Engineering, National Taiwan Normal University
E-mail: {60647061S , 60647028S , teinhonglo , berlin}@ntnu.edu.tw

performance of the resulting CTC-Attention hybrid system on recognizing short utterances. Experiments on a Mandarin Chinese meeting corpus demonstrate that the CTC-Attention hybrid system delivers better performance on short utterance recognition in comparison to one of the state-of-the-art DNN-HMM settings, namely, the so-called TDNN-LFMMI system.

關鍵詞：CTC、Attention、端對端中文語音辨識、短句辨識

Keywords: CTC, Attention-based Encoder-Decoder, End-to-End Mandarin Chinese Speech Recognition, Short Utterance Recognition

1. 緒論 (Introduction)

隨著近幾年來深度學習技術的長足發展，在語音辨識任務上，深度類神經網路結合隱藏式馬可夫模型(Deep Neural Network-Hidden Markov Model, DNN-HMM) (Hinton *et al*., 2012)與傳統的高斯混合模型結合隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM) (Rabiner, 1989) (Gales & Yang, 2008)相比，在字錯誤率(Character Error Rate, CER)和詞錯誤率(Word Error Rate, WER)有了大幅度的下降。然而，儘管 DNN-HMM 已取得不錯的成果，但 DNN 聲學模型仍無法充分利用語音信號之時間依賴性的缺點，為了更好地捕捉該性質，過往學者們引入了遞歸類神經網路(Recurrent Neural Network, RNN) (Hochreiter & Schmidhuber, 1997) (Gers, Schmidhuber & Cummins, 1999)及長短期記憶模型(Long Short-Term Memory, LSTM) (Graves, Mohamed & Hinton, 2013) (Graves, Jaitly & Mohamed, 2013) (Sak, Senior & Beaufays, 2014) (Sak, Vinyals & Heigold, 2014) (Li & Wu, 2015)組成聲學模型。這類的聲學模型與 DNN 相同，在訓練時仍是使用最小交互熵(Cross Entropy, CE)的準則，並且也能夠再進一步結合序列式鑑別式訓練(Kingsbury, Sainath & Soltau, 2012) (Veselý, Ghoshal, Burget & Povey, 2013)得到更好的辨識效果。

語音辨識可以視為一種序列對序列的任務，將輸入的語音訊號對應輸出的文字序列。在傳統語音辨識器的訓練中，分別由聲學模型、語言模型及發音詞典構成，並且在訓練 DNN 前，還得透過預先訓練的 GMM-HMM 將聲音與文字強制對齊，因此需要額外的冗餘步驟。有別於傳統的語音辨識訓練，CTC 訓練準則使得聲學模型可直接將聲學特徵透過類神經網路輸出對應到的字符(Character)或音素(Phone) (Graves *et al*., 2013) (Graves, Fernández, Gomez & Schmidhuber, 2006)，甚至在資料量夠大(通常大於 3000 小時)時能夠直接對應到單詞(Soltau, Liao & Sak, 2016) (Li, Ye, Das, Zhao & Gong, 2018)，並且在解碼時可以不需要語言模型，這樣的做法稱之為端對端的訓練方式。另一方面，有鑑於 CTC 端對端模型的成功，且基於 Attention 的遞歸類神經網路已被廣泛應用於各個研究領域(Bahdanau, Cho & Bengio, 2015) (Xu *et al*., 2015)，(Chorowski, Bahdanau, Serdyuk, Cho & Bengio, 2015)也將此模型應用於語音辨識的任務上，得到接近 CTC 的 WER。在後續其他學者研究中，在大量語料的情況下，Attention 模型的 WER 甚至能逼近辨識效果很好的 CLDNN-HMM 模型(Convolutional Long Short-Term Memory, Fully Connected Deep

Neural Networks, CLDNN) (Chan, Jaitly, Le & Vinyals, 2016)。

雖然端對端的訓練方式相較於傳統的 DNN-HMM 訓練更加簡單，但在少量語料下，其效能仍與傳統的 DNN-HMM 模型有一段差距。為此，(Kim, Hori & Watanabe, 2017) (Watanabe, Hori, Kim, Hershey & Hayash, 2017)，使用 CTC-Attention 模型 (Hybrid CTC-Attention Model)。該方法為結合 CTC 與 Attention 模型的多任務學習架構，目的是希望利用 CTC 彌補 Attention 模型對齊錯誤(Misalignment)及收斂慢的問題。在(Kim *et al*., 2017) (Watanabe *et al*., 2017)的實驗結果顯示，CTC-Attention 模型可在少量語料下，能夠更接近甚至低於 DNN-HMM 模型的辨識率。因此，本篇論文希望基於此模型對於中文會議語料的辨識做研究探討，我們的貢獻可分為：

1. 不同 Attention 機制的辨識結果：在長句測驗集實驗結果中發現使用 Coverage Location 效果比 Location 機制好，而在短句實驗則反之。
2. CTC 的權重對於辨識結果之影響：一般來說情況下，多任務架構訓練之聲學模型可優於傳統 CTC 或 Attention 模型。
3. CTC-Attention 混合模型於短語句測試之影響：短句辨識任務上，當使用較大的 CTC 權重作為解碼參數，可以得到最好的效果。

2. 方法 (Method)

2.1 CTC (Connectionist Temporal Classification)

給定一段長度為 T 的聲學特徵序列 X 及一段長度 L 的標籤序列 C，其中$\mathrm{C} = \{c_l \in \mathrm{U} | \mathrm{l} = 1, \dots, L\}$，U 為存在的標籤集合。並且 CTC 引入了額外的空白標籤，作為標籤間的分界，每個音框的標籤序列可表示為$S = \{s_t \epsilon \mathrm{U} \cup \{< blank >\} | t = 1, \dots T\}$。 X 對應 C 的後驗機率可表示為：

$$P(C|X) = \sum_S P(C|S,X)P(S|X)$$

$$\approx \sum_S P(C|S)P(S|X) \tag{1}$$

由於 CTC 假設每一時間下的聲音輸入對應字符為條件獨立，因此$P(C|S,X) \approx P(C|S)$，其中$P(C|S)$可以視為 CTC 標籤模型，可以分別由貝氏定理(Bayes' Rule)、鏈式法則(Chain Rule)展開。最後帶入條件獨立的假設可推導為：

$$P(C|S) = \frac{P(S|C)P(C)}{P(S)}$$
$$= \prod_{t=1}^{T} P(s_t|C, s_{1:t-1}) \frac{P(C)}{P(S)}$$
$$\approx \prod_{t=1}^{T} P\left(s_t \middle| s_{t-1,} C\right) \frac{P(C)}{P(S)} \quad (2)$$

其中，$P(C)$為字符級別的語言模型，$P(S)$為每一狀態的先驗機率，$P\left(s_t \middle| s_{t-1,} C\right)$為狀態轉移機率，為了使輸出有空白標籤，CTC 將上述長度 L 的標籤序列 C 調整為：

$$c' = \{< blank >, c_1, < blank >, c_2, < blank >, \dots c_L\}$$
$$= \{c'_l \in \mathrm{U} \cup \{< blank >\} | l = 1, \dots 2L + 1\} \quad (3)$$

狀態轉移機率$p\left(s_t \middle| s_{t-1,} C\right)$可以表示為：

$$P\left(s_t \middle| s_{t-1,} C\right) \begin{cases} 1 & s_t = c'_l \; and \; s_{t-1} = c'_l \; for \; all \; possible \; l \\ 1 & s_t = c'_l \; and \; s_{t-1} = c'_{l-1} \; for \; all \; possible \; l \\ 1 & s_t = c'_l \; and \; s_{t-1} = c'_{l-2} \; for \; all \; possible \; even \; l \\ 0 & otherwise \end{cases} \quad (4)$$

其依序分別為相似於 HMM 的自我轉移(Self-loop)，轉移至下一狀態，而第三個則是在l為偶數時且c'_l及c'_{l-2}皆屬於標籤序列 S 時跳過 blank 狀態，如同下圖的拓樸結構：

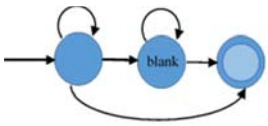


圖 1. CTC 拓墣結構
[Figure 1. CTC's topology]

另一方面，P(S|X)為 CTC 聲學模型，由鏈式法則展開後，再帶入條件獨立的假設可以表示為：

$$P(S|X) = \prod_{\mathrm{t}=1}^{T} P(s_t|s_1, \dots, s_{t-1}, X)$$
$$\approx \prod_{t=1}^{T} P(s_t|X) \quad (5)$$

其中$P(s_t|X)$為 softmax 輸出的結果，綜合上述式(2)、式(5)，可以得到：

$$P(C|X) \approx \sum_s \prod_{t=1}^{T} P(s_t|s_{t-1}, C) P(s_t|X) \frac{P(C)}{P(S)} \quad (6)$$

而 CTC 的目標函數通常不包含$\frac{P(C)}{P(S)}$，因此可定義為：

$$P_{ctc}(C|X) \approx \sum_s \prod_{t=1}^{T} P(s_t|s_{t-1}, C) P(s_t|X) \quad (7)$$

上式為 CTC 目標函數，而訓練時希望最小化損失函數(Loss Function)便是$-\ln P_{ctc}(C^*|X)$，C^*為訓練語料的正確字符序列的標籤，損失函數越小等同於輸出正確標籤的機率越大。

2.2 Attention 模型 (Attention-based Encoder-Decoder Network)

有別於 CTC 對於聲音對應字符的條件獨立假設，Attention 模型直接估測聲學特徵對應到字符的後驗機率，其目標函式可定義為：

$$P_{att}(C|X) = \prod_{l=1}^{L} P(c_l|X, c_{1:l-1}) \quad (8)$$

$P(c_l|X, c_{1:l-1})$可以由下列式子推得：

$$h_t = Encoder(X) \quad (9)$$

$$e_{l\langle t} = \begin{cases} LocationAttention: & \\ \boldsymbol{F_l} = \boldsymbol{K} * \ \boldsymbol{a_{l-1}} & (10) \\ \mathbf{g}^T \tanh\left(W_q \boldsymbol{q_{l-1}} + W_h \boldsymbol{h_t} + W_f \boldsymbol{f_{lt}}\right) & (11) \\ CoverageLocationAttention: & \\ \boldsymbol{F_l} = \boldsymbol{K} * \ \boldsymbol{a_{l-1}} & (12) \\ \boldsymbol{v_l} = \sum_{l'=1}^{l-1} \boldsymbol{al'} & (13) \\ \mathbf{g}^T \tanh\left(W_q \boldsymbol{q_{l-1}} + W_h \boldsymbol{h_t} + W_f \boldsymbol{f_{lt}} + W_v \boldsymbol{v_{lt}}\right) & (14) \end{cases}$$

$$a_{lt} = \frac{\exp(\gamma e_{lt})}{\sum_l \exp(\gamma e_{lt})} \quad (15)$$

$$\boldsymbol{r_l} = \sum_{t=1}^{T} a_{lt} \boldsymbol{h_t} \quad (16)$$

$$p(c_l|X, c_{1:l-1}) = \mathrm{Decoder}(\boldsymbol{r_l}, \boldsymbol{q_l}, c_{l-1}) \quad (17)$$

其中$\mathrm{h_t}$為 Encoder 的隱藏狀態向量，$\mathrm{a_{lt}}$為 Attention 的權重由$\mathrm{e_{lt}}$作 Softmax 得到，而 γ 為強調權重的 Sharpen Factor，而我們可藉由 Decoder 的隱藏狀態向量$\mathrm{q_{l-1}}$ 為 Query 去查找做為 Key-Value 的$\mathrm{h_t}$得到$\mathrm{e_{lt}}$，g、$\mathrm{W_q}$、$\mathrm{W_h}$、$\mathrm{W_f}$、$\mathrm{W_v}$為可訓練的矩陣參數。F_l為 Location Attention 機制(Chorowski *et al.*, 2015)中由一維摺積層 K 對於過去的 Attention 向

量 $\{a_1, a_2, \dots a_{l-1}\}$抽取的向量集合，Fl={fl1, fl2, …, flT}。v_l為 Coverage Attention 機制(Watanabe *et al*., 2017)中負責紀錄所有 Decoder 過去的 Attention 權重分佈，加入該機制的目的是希望能夠減少插入錯誤(Insertion)與刪除錯誤(Deletion)的出現，以達到更低的WER或CER。Attention模型訓練時損失函數也同樣希望最小化$-\ln P_{att}(C^*|X)$。Attention模型與 CTC 損失函數差異在於前者計算時必須考慮過去輸出的字符。

2.3 CTC-Attention模型 (Hybrid CTC-Attention model)

由於語音的每個音框間彼此相關，所以 CTC 中對於每個音框對應文字輸出的獨立性假設是飽受批評。另一方面，Attention 模型有著非單調的左到右對齊和收斂較慢的缺點。(Kim et al., 2017) (Watanabe et al., 2017)通過使用 CTC 目標函數作為輔助函數，將 Attention 模型與 CTC 結合作多任務學習。這種訓練方式可保留 Attention 模型的優勢，並能有效改善 Attention 模型的收斂速度與對齊錯誤的問題。綜合式(7)及式(8)，CTC-Attention 混合模型透過線性組合兩種模型的目標函數，其訓練的損失函數可以表示成：

$$\mathcal{L}_{MOL} = -\left(\lambda lnP_{ctc}(C|X) + (1-\lambda)lnP_{att}(C|X)\right) \tag{18}$$

其中 λ 的範圍為 0 ≤ λ ≤ 1，而在解碼時，，我們可同時使用 CTC 及 Attention 模型的輸出，可表示為：

$$\begin{aligned} &logp(c_n|c_{1:n-1}, h_{1:T'}) \\ &= \alpha logp_{ctc}(c_n|c_{1:n-1}, h_{1:T'}) + (1-\alpha)logp_{att}(c_n|c_{1:n-1}, h_{1:T'}) \end{aligned} \tag{19}$$

2.4 聲學模型 (Acoustic model)

本篇論文在聲學模型的 Encoder 部分使用的是兩層的 VGG 層加上八層 Long Short-Term Memory Projection(LSTMP)，LSTMP (Sak *et al*., 2014)是 LSTM 的變形，通過添加投影層來進一步優化 LSTM 的速度和效能。而 VGG 與(Chan *et al*., 2016)的金字塔型的 LSTM 結構作為 Encoder 相比，使用 VGG 的效果在(Watanabe *et al*., 2018)說明了在大多數情況會優於金字塔型的 LSTM，因此我們採用 VGG-LSTMP 作為 Encoder，完整模型架構如圖 2，其中 X 代表輸入特徵，C 代表輸出的字符序列。解碼算法採用光束搜尋，搜尋時的分數結合可參考 2.3 節式 19。

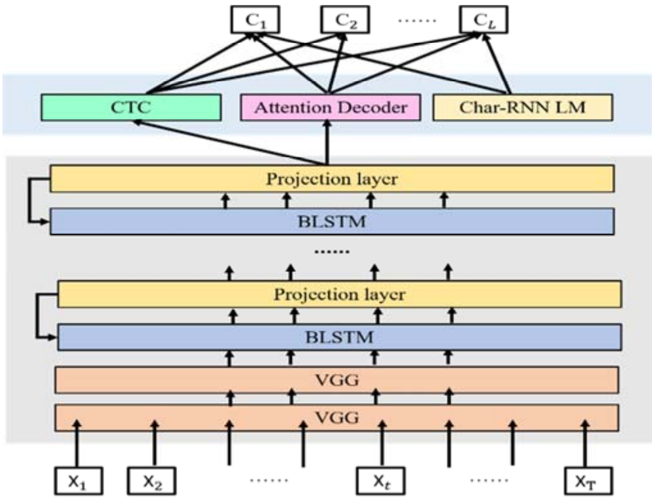


圖 2. CTC-Attention 混合模型架構
[Figure 2. Hybrid CTC-Attention model architecture]

3. 實驗結果與分析 (Experiments and Results)

3.1 實驗語料與設定 (Corpus and Setup)

本論文實驗使用的語料為華語會議語料，該語料為國內企業所收集整理的語料庫。其中談話內容沒有經過設計，而是一般公司在實際開會中討論面臨的問題與技術，而說話方式屬於正常交談，所以會有不少停頓、口吃、中英文轉換等情形，相較於新聞語料，較具有挑戰性。其訓練集為 230 小時，而測試集則為 2.6 小時兩場會議的內容，另外還有一額外 3 小時短句測試集，其內容為多為在訓練語料中未曾出現的專有名詞，在辨識上更有難度。

表 1.語料庫訓練集、測試集小時數與句數
[Table 1. hours of training set and test set]

	總小時數	句數
訓練集	230	367434
測試集	2.6	2306
短句測試集	3	2809

特徵部份，我們使用 80 維的 Filterbank 加 Pitch 特徵；聲學模型部分，我們使用兩層 VGG 層及八層 LSTMP 作為 Encoder，每層 LSTMP 各有 320 個單元，Decoder 部分則使用單層 300 個單元的 LSTM，如圖 2 所示。Attention 機制分別為 Location 及 Coverage Location。語言模型部分我們用訓練集的轉寫作為語料訓練字符級別的 RNN 語言模型，訓練時 CTC 權重設為 0.5，在解碼時使用(Watanabe *et al*., 2017)的解碼算法並利用

Shallow Fusion (Gulcehre *et al*., 2015)的方式，插入額外的語言模型分數以提升整體辨識效能，實作上使用 Espnet (Watanabe *et al*., 2018)工具，另外為我們也使用了 Kaldi (Povey *et al*., 2011)工具實作時延式類神經網路(Time-delay Neural Network, TDNN)結合 Lattice-free Maximum Mutual Information (LF-MMI) (Povey *et al*., 2016)訓練的聲學模型與端對端混和模型做比較。

3.2 實驗結果 (Experiment result)

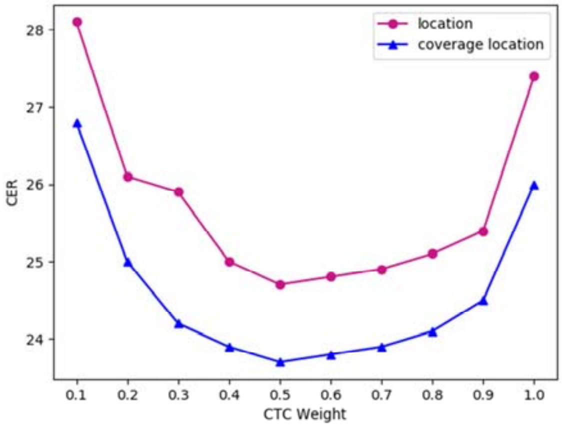


圖 3.不同的 CTC 權重對於測試集 CER 的影響
[Figure 3. Character error rate when using different CTC weight in test set]

圖 3 橫軸代表 CTC 的權重，而縱軸代表 CER。由於 CTC 的權重在解碼時是可以變動的，我們利用窮舉的方式嘗試不同的權重組合。由實驗結果得知，我們發現 Location 及 Coverage Location 皆發現權重設為 0.5 在測試集上表現最好，而權重偏向 CTC 或是 Attention 都使 CER 有上升趨勢。當 CTC 權重為 1.0 時可視為傳統 CTC 模型，反之當權重為 0.0 時為傳統 Attention 模型。另一方面，Coverage Location 在任一權重下其 CER 皆比 Location Attention 模型低，因此我們進一步去分析其解碼結果。

表 2.不同 Attention 機制的表現
[Table 2. Different attention mechanism performance in test set]

Attention	CER	#Deletion	#Insertion
location	24.7	3637	1474
Coverage location	**23.7**	3378	1467

由圖 3 已知道 CTC 的權重設為 0.5 時其 CER 為最低，因此表 1 為該權重下的辨識率，CER 分別為 24.7 及 23.7。在實驗的結果中，我們發現由 Coverage 機制的模型解碼後，插入錯誤與刪除錯誤數有些微但一致的進步，其結果也反映在 CER 上。其中可能的原因是 Coverage 機制，該機制避免了模型的注意力過度集中在同個音框的語音特徵上。

另外 TDNN-LFMMI 於此測試集的 CER 為 17%，相較之下我們的方法仍有進步空間。

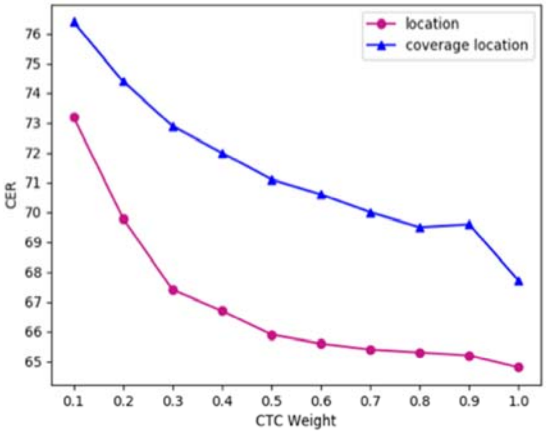


圖 4. 不同的 CTC 權重對於短句測試集 CER 的影響
[Figure 4. Character error rate when using different CTC weight in external short utterance test set]

在這次的實驗中，我們額外比較 CTC-Attention 混合模型於短句辨識任務上的表現，由圖 4 可以得知在任一權重下的 CER，與前一個測試集的實驗相反，Location 機制的模型反而較 Coverage Location 好。推測其原因可能在於語句過短，使得 Coverage Location 模型無法發揮 Coverage 機制的作用，因而表現較差。而 CTC 權重為 1.0 時，即僅使用 CTC 解碼，兩種模型皆為最佳表現，其原因可能在於 CTC 模型是為了解決輸出的文字序列長度小於輸入的聲音長度的情況而設計，而 Attention 模型，也出現了如同(Chan *et al*., 2016) 的實驗結果，當測試語句與訓練語句長度差異太大時，解碼出來的 CER 變差許多，然而因為 CTC 權重的可變動性，可以看到 CTC-Attention 混合模型具有因應不同語句長度的彈性。

表 3.不同 Attention 機制於短句測試集表現
[Table 3. Different attention mechanism performance in in external short utterance test set]

Model	CER
location	**64.8**
Coverage location	**67.7**
TDNN-LFMMI	85.5

4. 結論與未來展望 (Conclusion and Future works)

本篇論文探討了兩種端對端語音辨識的主流方法，以及 CTC-Attention 模型權重對於語句長短的辨識效果，我們發現在短語句辨識上 CTC-Attention 模型相不僅相較於 TDNN-LFMMI 的表現更加出色，同時具有能夠依據語句長短改變權重解碼的彈性。另一方面，並且由於使用字符級別的預測目標及語言模型，更能有效處理未知詞的問題。

近年來在序列對序列模型上有學者提出許多優化訓練的方法如(Pereyra, Tucker, Chorowski, Kaiser & Hinton, 2017)，能夠避免 Overconfidence，以及 Cold Fusion (Sriram, Jun, Satheesh & Coates, 2018) 在訓練聲學模型時加入預先訓練語言模型，以上方法都能夠有更好的泛化效果與收斂速度，我們在未來也將在訓練中嘗試加入該方法。其次，在語言模型則將加入目前訓練集外以外的語料，並希望能針對語種切換做額外研究；最後，聲學模型方面也希望能夠再多嘗試不同的 Attention 機制，以及不同的類神經網路架構對於華語語音辨識的效果，以期待未來能夠得到更低的字錯誤率。

參考文獻 (References)

Bahdanau, D., Cho, K.H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of ICASSP 2016*. doi: 10.1109/ICASSP.2016.7472621

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In *Proceedings of NIPS 2015*, 577-585.

Gales, M. & Yang, S. (2008). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends® in Signal Processing*, *1*(3), 195-304. doi: 10.1561/2000000004

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. In *Proceedings of ICANN 1999*. doi: 10.1049/cp:19991218

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML 2006*, 369-376. doi: 10.1145/1143844.1143891

Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *Proceedings of ASRU 2013*. doi: 10.1109/ASRU.2013.6707742

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of ICASSP 2013*.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., …Bengio, Y. (2015). On Using Monolingual Corpora in Neural Machine Translation. In arXiv preprint arXiv: 1503.03535

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., …Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82-97. doi: 10.1109/MSP.2012.2205597

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Kim, S., Hori, T., & Watanabe, S. (2017). Joint CTC-Attention based end-to-end speech recognition using multi-task learning. In *Proceedings of ICASSP 2017*. doi: 10.1109/ICASSP.2017.7953075

Kingsbury, B., Sainath, T. N., & Soltau, H. (2012). Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization. In *Proceedings of Interspeech 2012*, 10-13.

Li, X. & Wu, X. (2015). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *Proceedings of ICASSP 2015*. doi: 10.1109/ICASSP.2015.7178826

Li, J., Ye, G., Das, A., Zhao, R., & Gong, Y. (2018). Advancing Acoustic-to-word CTC model. In *Proceedings of ICASSP 2018*. doi: 10.1109/ICASSP.2018.8462017

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., & Hinton, G. (2017). Regularizing neural networks by penalizing confident output distributions. In *Proceedings of ICLR 2017*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., …Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU 2011*.

Povey, D., Peddinti, V., Galvez, D., Ghahrmani, P., Manohar, V., Na, X., …Khudanpur, S. (2016). Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proceedings of Interspeech 2016*. doi: 10.21437/Interspeech.2016-595

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257 - 286. doi: 10.1109/5.18626

Sak, H., Senior, A. & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proceedings of INTERSPEECH-2014*, 338-342.

Sak, H., Vinyals, O., & Heigold, G. (2014). Sequence discriminative distributed training of long short-term memory recurrent neural networks. In *Proceedings of Interspeech 2014*, 1209-1213.

Soltau, H., Liao, H., & Sak, H. (2016). Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. In arXiv preprint arXiv: 1610.09975

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2018). Cold Fusion: Training Seq2Seq Models Together with Language Models. In *Proceedings of ICLR 2018*.

Veselý, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence discriminative training of deep neural networks. In *Proceedings of Interspeech 2013*, 2345-2349..

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., …Ochiai , T. (2018). ESPnet: End-to-End Speech Processing Toolkit. In *Proceedings of Interspeech 2018*, 2207-2211. doi: 10.21437/Interspeech.2018-1456

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayash, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1240-1253. doi: 10.1109/JSTSP.2017.2763455

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., …Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML 2015*, 2048-2057.

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member： NT$ 1,000 (US$50.-)
group membership： NT$20,000 (US$1,000.-)
life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： ________________

Name： __________________________________ Date of Birth： _______________

Country of Residence： ___________________ Province/State： ___________________

Passport No.： ___________________________ Sex: ________________________

Education(highest degree obtained)： ____________________________________

Work Experience： ____

____

Present Occupation： ____

Address： ____

____

Email Add： ____

Tel. No： ____________________________ Fax No： ___________________________

Membership Category：☐ Regular Member ☐ Life Member

Date： _____/____/____ （Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）

Life Member ： US$500.- （NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）
（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目
（三）收集國內外有關計算語言學知識之圖書及最新發展之資料
（四）發行有關之學術刊物，論文集及通訊
（五）研定有關計算語言學專用名稱術語及符號
（六）與國際計算語言學學術機構聯繫交流
（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中 華 民 國 計 算 語 言 學 學 會
個 人 會 員 入 會 申 請 書

會員類別	☐終身 ☐個人 ☐學生	會員編號		（由本會填寫）	
姓　　名		性別		出生日期	年　月　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
申請人:　　　　　　（簽章） 中　華　民　國　　　年　　　月　　　日					

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
PAYMENT FORM

Name: ____________________________________(Please print) Date: _______________

Please debit my credit card as follows: US$ _____________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:___________________

Card No.: ________ -_________-_________ -_________ Exp. Date:____________(M/Y)

3-digit code: _________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: ____________________________________

Phone No.: __________________________E-mail: _____________________________________

Address: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: ______

US$ _____ ❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: ______

US$ _____ ❑ Publications:_________________________________

US$ _____ ❑ Text Corpora: _______________________________

US$ _____ ❑ Speech Corpora:______________________________

US$ _____ ❑ Others: _____________________________________

US$ _________ ❑ Membership Fees ❑ Life Membership ❑ New Membership ❑Renew

US$ _________ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP
% IIS, Academia Sinica
Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: ____________________________(請以正楷書寫)　　日期:：____________________

卡別：❑ VISA CARD　　❑ MASTER CARD ❑ JCB CARD　　發卡銀行：______________

信用卡號：__________-__________-__________-__________　　有效日期：__________(m/y)

卡片後三碼：______________（卡片背面簽名欄上數字後三碼）

持卡人簽名：_______________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：____________________________E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________ ❑ 中文計算語言學期刊(IJCLCLP) _______________

NT$__________ ❑ Journal of Information Science and Engineering (JISE)

NT$__________ ❑ 中研院詞庫小組技術報告_____________________

NT$__________ ❑ 文字語料庫 ____________________________________

NT$__________ ❑ 語音資料庫 ____________________________________

NT$__________ ❑ 光華雜誌語料庫1976~2010

NT$__________ ❑ 中文資訊檢索標竿測試集/文件集

NT$__________ ❑ 會員年費：❑續會　　　　❑新會員　　　❑終身會員

NT$__________ ❑ 其他: _____________________________________

NT$_______________ = 合計

填妥後請傳真至 02-27881638 或郵寄至:
11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： ☐ Credit Card (Preferred)
  ☐ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會
相關出版品價格表及訂購單

編號	書目	會員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本) ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05 中文詞類分析	185	205	_____	_____
7.	no.93-06 現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02 古漢語字頻表	180	200	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02 論語詞頻表	90	100	_____	_____
16	no.98-01 詞頻詞典	395	440	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義（中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期） 年份：______（過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合 計	_____	_____

※ 此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：______________________ 收據抬頭：______________________

地　　址：____

電　　話：______________________ E-mail:______________________

Information for Authors

nternational Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites ubmission of original research papers in the area of computational linguistics and speech/text processing of atural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously npublished and cannot be under consideration elsewhere. Submissions should report significant new research esults in computational linguistics, speech and language processing or new system implementation involving gnificant theoretical and/or technological innovation. The submitted papers are divided into the categories of egular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full etails. Short papers can focus on a smaller research issue. And survey papers should cover emerging research ends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no rict length limitation on the regular and survey papers. But it is suggested that the manuscript should not xceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced 4 pages. All contributions will be anonymously reviewed by at least two reviewers.

opyright: It is the author's responsibility to obtain written permission from both author and publisher to eproduce material which has appeared in another publication. Copies of this permission must also be enclosed ith the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to cilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright rm, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP ciety, will be required before the manuscript can be accepted for publication. The papers published by CLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database rvices.

yle for Manuscripts: The paper should conform to the following instructions.

Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size 11 points or larger.

Title and Author: The first page of the manuscript should consist of the title, the authors' names and stitutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and nail address. The title of the paper should use normal capitalization. Capitalize only the first words and such ner words as the orthography of the language requires beginning with a capital letter. The author's name ould appear below the title.

Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords required. The abstract should not only indicate the scope of the paper but should also summarize the author's nclusions.

Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-nd margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with perscript numbers. Footnotes may appear at the end of manuscript

Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or itten clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in rentheses.

References: All the citations and references should follow the APA format. The basic form for a reference ks like

`Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article.` *`Title of Periodical, volume number`*`(issue number), pages.`

re shows an example.

`Scruton, R. (1996). The eclipse of listening.` *`The New Criterion, 15`*`(30), 5-13.`

e basic form for a citation looks like `(Authora, Authorb, and Authorc, Year)`. Here shows an example. ruton, 1996).

ase visit the following websites for details.

APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

APA Stytle (http://www.apastyle.org/)

page charges are levied on authors or their institutions.

al Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply l manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

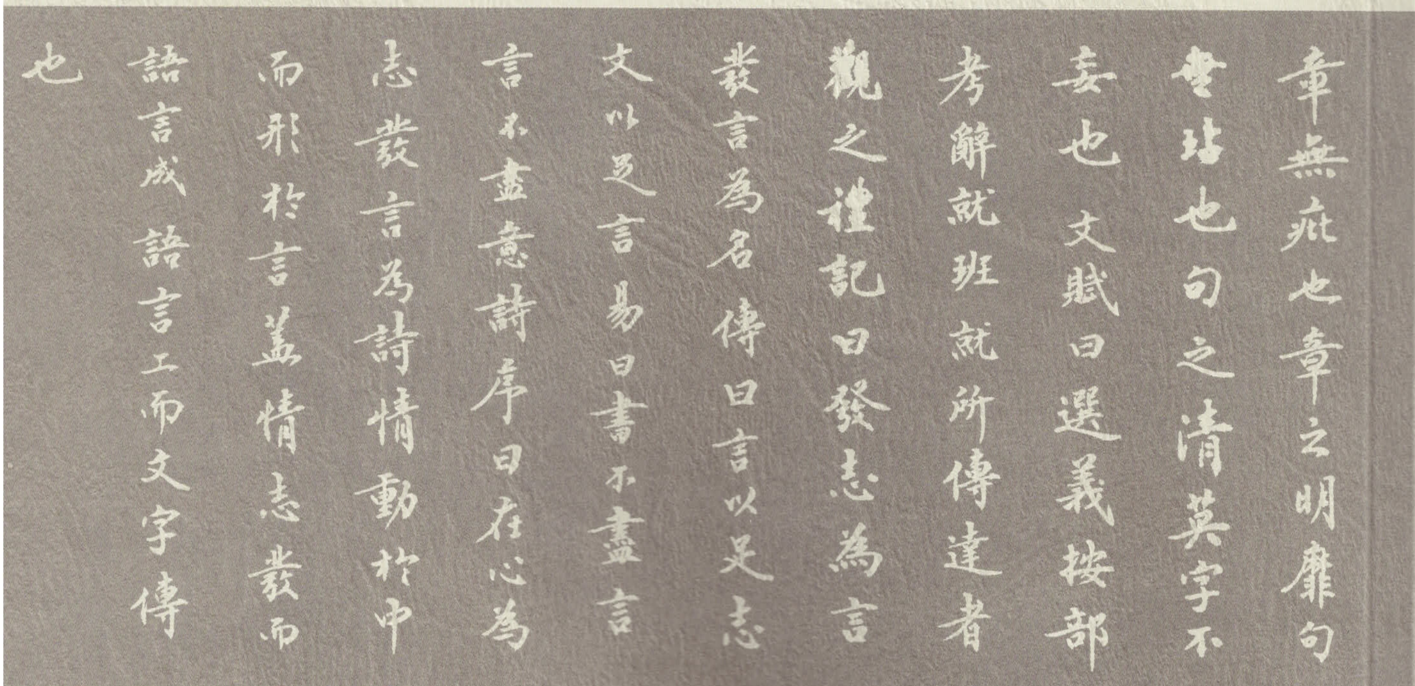
line Submission: http://www.aclclp.org.tw/journal/submit.php

ase visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.24 no.1 June 2019

Contents

Papers



ISSN: 1027-376X
The Association for Computational Linguistics and Chinese Language Processing