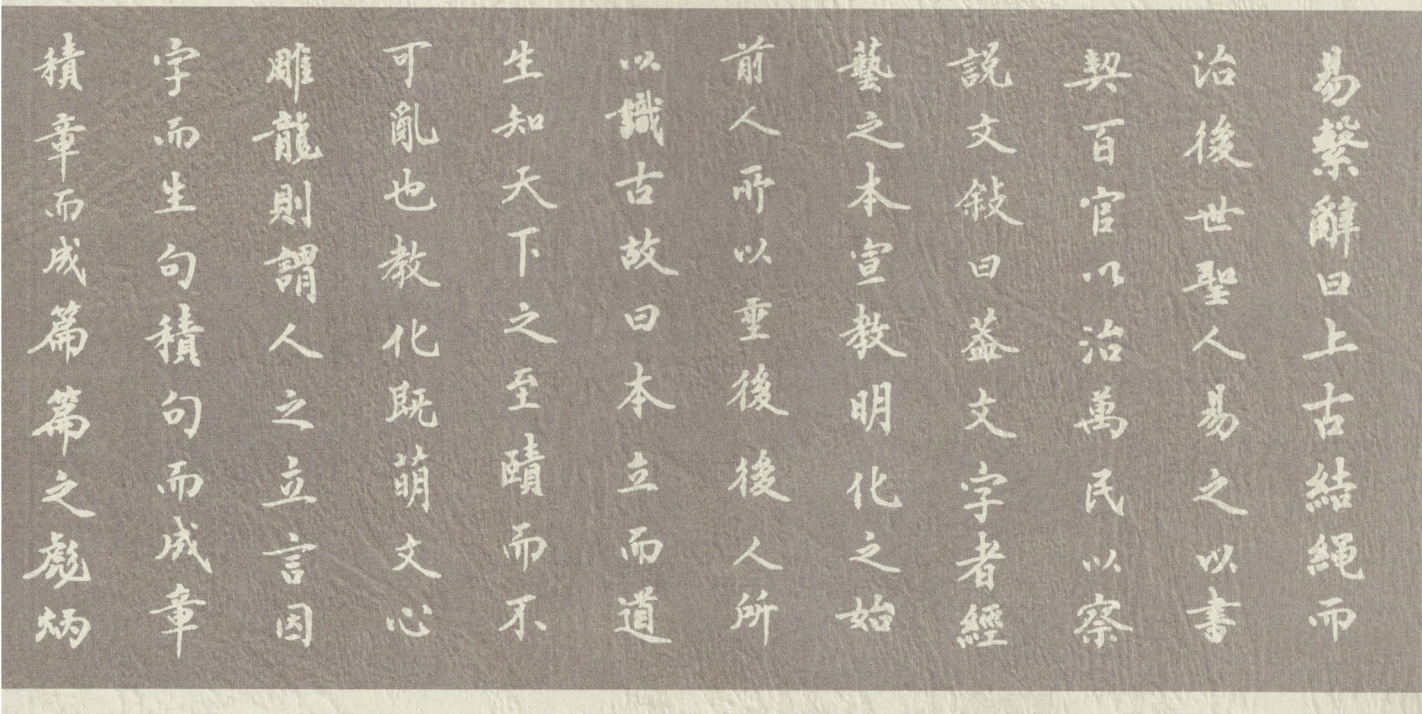
International Journal of

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.23 No.2 December 2018 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

The Association for Computational Linguistics and Chinese Language Processing, Taipei

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. 23, no. 2, December 2018

Contents

Forewords…….....…….....… i
Chen-Yu Chiang and Min-Yuh Day

Papers

使用長短期記憶類神經網路建構中文語音辨識器之研究 [A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network]…….……... 1
賴建宏(Chien-hung Lai),王逸如(Yih-Ru Wang)

結合鑑別式訓練與模型合併於半監督式語音辨識之研究 [Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition]…….…..... 19
羅天宏(Tien-Hong Lo),陳柏琳(Berlin Chen)

結合鑑別式訓練聲學模型之類神經網路架構及優化方法的改進 [Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method]……... 35
趙偉成(Wei-Cheng Chao),張修瑞(Hsiu-Jui Chang),
羅天宏(Tien-Hong Lo),陳柏琳(Berlin Chen)

Supporting Evidence Retrieval for Answering Yes/No Questions..... 47
Meng-Tse Wu, Yi-Chung Lin and Keh-Yih Su

未登錄詞之向量表示法模型於中文機器閱讀理解之應用 [An OOV Word Embedding Framework for Chinese Machine Reading Comprehension]……...…….. 67
羅上堡(Shang-Bao Luo),李青憲(Ching-Hsien Lee),
涂家章(Jia-Jang Tu),陳冠宇(Kuan-Yu Chen)

以深層類神經網路標記中文階層式多標籤語意概念 [Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network]……...……. 85
周瑋傑(Wei-Chieh Chou),王逸如(Yih-Ru Wang)

Reviewers List & 2018 Index…… 99

Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. i-ii

Forewords

The 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018) was held at National Tsing Hua University (NTHU), Hsinchu, Taiwan, during October 4-5, 2018. ROCLING, which sponsored by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), is the leading and most comprehensive conference on computational linguistics and speech processing in Taiwan, bringing together researchers, scientists and industry participants from fields of computational linguistics, information understanding, and speech processing, to present their work and discuss recent trends in the field. This special issue presents the extended and reviewed versions of six papers meticulously selected from ROCLING 2018, including 3 natural language processing papers and 3 speech processing papers.

The first paper from National Chiao Tung University presents a neural network acoustic model with the CLDNN architecture for Mandarin speech recognition system. The CLDNN architecture cascades CNNs, LSTMs, and DNNs as one unified framework to achieve significant word error rate reduction. This paper is awarded as one of the two best papers of ROCLING 2018. The second paper from National Taiwan Normal University investigates acoustic model combination and semi-supervised discriminative training for meeting speech recognition. A promising reduction of word error rate is achieved when speech corpora is small in size under a semi-supervised training condition. The third paper from National Taiwan Normal University (the same research group as the second paper) discusses the discriminative training of acoustic models for speech recognition with improved neural network architecture and optimization method. A significant reduction of character error rate is reported with the Backstitch optimization method for the TDNN-LF-MMI acoustic model.

The fourth paper from Academia Sinica proposes a new n-gram matching approach for retrieving the supporting evidence for answering Yes/No questions. The proposed approach is evaluated on a task of answering Yes/No questions of Taiwan elementary school Social Studies lessons. The experiment results outperform Lucene Apache search engine substantially. This paper is awarded as one of the two best papers of ROCLING 2018. The fifth paper from National Taiwan University of Science and Technology and Industrial Technology Research Institute introduces an OOV word embedding framework for generating reasonable low-dimensional dense vectors. A series of experiments and comparisons demonstrate the efficacy of the proposed framework in Chinese machine reading comprehension. The last paper from National Chiao Tung University aims to classify the concept of word in E-HowNet and proposes a deep neural network training method with hierarchical relationship in E-HowNet taxonomy. The experimental results indicate the proposed order-award 2-Bag Word2Vec with

hierarchical classification achieved higher accuracy than flatten classification.

The Guest Editors of this special issue would like to thank all of the authors and reviewers for contributing their knowledge and experience at the ROCLING 2018. We hope this special issue provide for directing and inspiring new pathways of natural language processing and spoken language research within the research field.

Guest Editors

Chen-Yu Chiang

Department of Communication Engineering, National Taipei University, Taiwan

Min-Yuh Day

Department of Information Management, Tamkang University, Taiwan

Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 1-18

使用長短期記憶類神經網路建構中文語音辨識器之研究

A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network

賴建宏*、王逸如+

Chien-hung Lai and Yih-Ru Wang

摘要

近年來類神經網路(Neural network)被廣泛運用於語音辨識領域中，本論文使用遞迴式類神經網路(Recurrent Neural Network)訓練聲學模型，並且建立中文大辭彙語音辨識系統。由於遞迴式類神經網路為循環式連接(Cyclic connections)，應用於時間序列訊號的模型化(Modeling)，較於傳統全連接(Full connection)的深層類神經網路而言更有益處。

然而一般單純遞迴式類神經網路在訓練上隨著時間的遞迴在反向傳播(Backpropagation)更新權重時有著梯度消失(Gradient vanishing)以及梯度爆炸(Gradient exploding)的問題，導致訓練被迫中止，以及無法有效的捕捉到長期的記憶關聯，因此長短期記憶(Long Short-Term Memory, LSTM)為被提出用來解決此問題之模型，本研究基於此模型架構結合了卷積神經網路(Convolutional Neural Network)及深層類神經網路(Deep Neural Network)建構出 CLDNN 模型。

訓練語料部分，本研究使用了 TCC300(24 小時)、AIShell(162 小時)、NER(111 小時)，並加入語言模型建立大辭彙語音辨識系統，為了檢測系統強健度(Robustness)，使用三種不同環境之測試語料，分別為 TCC300(2.4 小時，朗讀

* 國立交通大學電信工程研究所
Institute of Communications Engineering, National Chiao Tung University
E-mail: lsr950082@speech.cm.nctu.edu.tw
+ 國立交通大學電機工程學系
Department of Electronic Engineering, National Chiao Tung University
E-mail: yrwang@cc.nctu.edu.tw

語速)、NER-clean(1.9 小時，快語速，無雜訊)、NER-other(9 小時，快語速，有雜訊)。

關鍵詞：遞迴式類神經網路、長短期記憶、梯度消失(爆炸)、聲學模型、中文、大辭彙語音辨識、卷積類神經網路、深層類神經網路

Abstract

In recent years, neural networks have been widely used in the field of speech recognition. This paper uses the Recurrent Neural Network to train acoustic models and establish a Mandarin speech recognition system. Since the recursive neural networks are cyclic connections, the modeling of temporal signals is more beneficial than the full connected deep neural networks.

However, the recursive neural networks have the problem of gradient vanishing and gradient exploding in the backpropagation, which leads to the training being suspended. And the inability to effectively capture long-term memory associations, so Long Short-Term Memory (LSTM) is a model proposed to solve this problem. This study is based on this model architecture and combines convolutional neural networks and deep neural networks to construct the CLDNN models.

Keywords: RNNs, LSTMs, Gradient Vanishing (Exploding), Acoustic Model, Mandarin, LVCSR, CNNs, DNNs

1. 緒論 (Introduction)

近年來，人工智慧(Artificial intelligence, AI)儼然已成為隨處可聽見的關鍵詞，綜觀歷史，AI 浪潮共出現過三次，而每一次浪潮的興起，都和語音辨識技術的發展脫離不了關係。早期的語音辨識技術是由語言學學者透過研究聲學以及語言學之間的關聯，統整歸納出一套規則法(Ruled-based)的語音辨識系統；但是由於聲學和語言學二者間的變化，無法單單使用規則法完成描述，而後發展出像機器學習(Machine Learning)這樣透過資料驅動(Data-driven)的方法，讓機器從輸入帶有標籤(Label)的資料中，自動剖析並從中獲取規則，並對於未知的資料進行預測。

在近期的大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)系統中，聲學模型部分有別於傳統的高斯混合模型(Gaussian Mixture Model, GMM) (Reynolds, 2009)，使用了深層類神經網路(Deep Neural Network, DNN) (Zhang, Trmal, Povey & Khudanpur, 2014) (Mohamed, 2014) 取代之。而以 DNN 建構的聲學模型，在訓練的過程中，必須使用大量的語料，對於不同的發聲才能有較佳的辨識結果。由於語音為時序相關訊號，因此本研究加入了遞迴式類神經網路(Recurrent neural network)訓練聲學模型，並探討其辨識結果

在語音辨識系統中，語言模型(Language model)亦扮演相當重要的角色，本研究基於

本實驗室擁有的文字語料，選擇八萬詞、十萬詞、十二萬詞的詞典(Lexicon)分別建構出三種 Tri-gram 語言模型，並且對於 TCC300、NER-clean、NER-other 三種不同環境的測試語料進行分析及探討，其中 TCC300 屬於朗讀語速且無雜訊之一般語料，NER-clean 為快語速且無雜訊之自發性語料、NER-other 則為快語速且有背景雜訊之自發性語料。

2. 實驗流程與實驗環境介紹 (Experimental Process and Experimental Environment)

由於訓練深層類神經網路非常耗時，本研究使用繪圖處理器(Graphics Processing Unit, GPU)來訓練含有 DNN、CNN (Abdel-Hamid *et al*., 2014) (Abdel-Hamid, Mohamed, Jiang & Penn, 2012) 及 LSTM (Sak, Senior & Beaufays, 2014a) (Sak, Senior & Beaufays, 2014b) 之聲學模型，並使用 Kaldi speech recognition toolkit (Povey *et al*., 2011) 中 nnet3 所提供的深層類神經網路訓練流程，進行聲學模型訓練。表 1 為實驗所使用之硬體規格；表 2 則為 GPU 規格表。另外為了使矩陣運算效能加速，本研究使用的 Kaldi 經由 Intel 開發之數學運算核心函式庫(Math Kernel Library, MKL)進行編譯。

表 1. 硬體規格描述
[Table 1. Hardware specification:]

CPU	Intel® Core™ i7-8700K @ 3.70GHz
RAM	64 GB DDR4-3000
HDD	4 TB SATA-III 7200RPM
GPU	NVIDIA GeForce GTX 1080TI
OS	Arch Linux 4.17.5-1 64bit

表 2. GPU 規格描述
[Table 2. GPU specification]

型號	NVIDIA GeForce GTX 1080TI
CUDA 核心數	3584
基礎時脈	1480 MHz
加速時脈	1582 MHz
記憶體時脈	11 Gbps
記憶體容量配置	11264 MB
記憶體介面型號	GDDR5X
記憶體介面頻寬	484 GB/s

3. 語料庫介紹 (Databases)

本節將分別介紹用於本實驗中之所有語料庫，其中用來當作訓練語料的有 TCC300、NER 及 AIShell 語料庫，而為了測試本實驗之辨識系統對於不同環境的辨識能力，因此在測試語料的選擇上，使用 TCC300 及 NER 語料庫，其中 NER 為廣播語料，又可細分為背景乾淨無雜訊之 NER-clean 以及背景有人為雜訊或音樂參雜其中之 NER-other。

3.1 TCC300語料庫 (TCC300 Corpus)

實驗中所使用的 TCC300 麥克風語音資料庫[1]是由國立交通大學(National Chiao Tung University, NCTU)、國立成功大學(National Cheng Kung University, NCKU)、國立台灣大學(National Taiwan University, NTU)共同錄製而成，並且由中華民國計算語言學學會(The Association for Computational Linguistics and Chinese Language Processing, ACLCLP)發行，此語料庫屬於麥克風朗讀語音，主要目的為提供台灣腔之中文語音辨認研究使用。

詳細資訊如表 3 所示，台灣大學部分主要包含詞以及短句，文本經過設計，考慮音節與其相連出現之機率，共由 100 人錄製而成；交通大學與成功大學部分則為長文語料，其語句內容透過中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百字，再切割分成 3 至 4 段，每段至多包含 231 字，兩校分別各錄製 100 人而成，且每人朗讀的文章皆不相同。每個學校之錄音取樣頻率皆為 16000 Hz，取樣位元數為 16 位元。

本實驗進一步將整個 TCC300 語料庫分為訓練語料與測試語料，訓練與測試比例約為 9 : 1，分別資訊如下：

- 訓練語料：約為 24.4 小時，共 284 位語者，8633 句發音，304780 個音節數。
- 測試語料：約為 2.4 小時，共 19 位語者，225 句長句發音，26357 個音節數。

表 3. TCC300 語料庫資訊
[Table 3. TCC300 corpus information]

學校名稱	文章屬性	語者總數		音節總數		檔案總數	
台灣大學	短句	男	50	男	27541	男	3425
		女	50	女	24677	女	3084
		總數	100	總數	52218	總數	6590
交通大學	長文	男	50	男	75059	男	622
		女	50	女	73555	女	616
		總數	100	總數	148614	總數	1238
成功大學	長文	男	50	男	63127	男	588
		女	50	女	68749	女	582
		總數	100	總數	131876	總數	1170

[1] Mandarin Microphone Speech Corpus-TCC300. http://www.aclclp.org.tw/use_mat_c.php#tcc300edu

3.2 NER語料庫 (NER Corpus)

NER 語料庫，全名為 NER Manual Transcription Vol1，為國立臺北科技大學和國家教育廣播電台合作錄製之語料庫，主要目的為大量轉寫教育電台之節目，產生節目逐字稿，以建置大規模台灣腔之語料庫，詳細內容如表 4 所示，內容大部份為談話性節目，多為自發性(Spontaneous)語音，僅少部分為新聞報導之朗讀式(Reading)語音。

此語料庫依照 1. 為錄音室內或為錄音室以外之場所錄製，2. 有無任何背景襯樂或非人聲之噪音兩項條件分為兩個部分：乾淨語料(Clean，約 19.4 小時，共 5106 個檔案)及其他語料(Other，約 107.4 小時，共 15983 個檔案)合計共約 126.8 小時，21089 個檔案，取樣頻率為 16000 Hz，取樣位元數為 16 位元，聲道數為 1(mono)。

語料庫中逐字稿來源由國立臺北科技大學之雙語語音辨識器進行初步轉寫逐字稿，後經由人工校正以及切割，並移除有版權疑慮之音樂段落後產生。

本實驗亦進一步將此語料庫分為訓練語料及測試語料，詳細資訊如下：

- 訓練語料：約為 111.5 小時，共 18710 句發音，1715091 個音節數。
- 測試語料：
  - Clean：約為 1.9 小時，共 549 句發音，33660 個音節數。
  - Other：約為 9.0 小時，共 1322 句發音，133746 個音節數。

表 4. NER 語料庫資訊
[Table 4. NER corpus information]

環境類型	節目名稱	代碼	總時數	音節總數	檔案總數
Clean	創設市集	CS	14.4	235052	4028
	技職最前線	JZ	1.8	34352	438
	國際教育心動線	GJ	3.2	55057	640
Other	多愛自己一點點	DA	13.6	212821	2347
	科學 SoEasy	KX	1.8	23415	208
	青年故事館	QG	17.3	260116	3202
	不太乖學堂	BG	9.5	143138	1586
	星期講座	WK	8.4	113202	1102
	遇見幸福幼兒園	YX	5.6	90419	826
	收藏人生	SR	16.5	280074	2670
	雙語新聞	SY	34.5	434851	4015

3.3 AIShell語料庫 (AIShell Corpus)

AIShell 語料庫(Bu, Du, Na, Wu & Zheng, 2017)，是由北京希爾貝殼科技有限公司釋放之開源語音資料庫，錄製內容如表 5，涉及智能居家、無人駕駛等 11 項領域，錄製過程皆在安靜的室內環境。

使用高效能麥克風錄製而成，取樣頻率為 44100 Hz，後降低取樣頻率至 16000 Hz，取樣位元數為 16 位元，由 400 名來自中國不同口音地區的參與者錄製而成，語者資訊如表 6、表 7 所示，此語料庫文本經人工校正過，正確率為 95%以上。

本實驗進一步將此語料庫分為訓練語料及測試語料：

- 訓練語料：約為 162.4 小時，共 129341 句發音，1862171 個音節數。
- 測試語料：約為 16.6 小時，共 12259 句發音，178041 個音節數。

表 5. AIShell 語料庫文本內容
[Table 5. AIShell corpus text contents]

主題	語句數
智能居家	5
地理訊息	30
音樂播放指令	46
數字串	29
電視與電影播放指令	10
金融	132
科學與科技	85
體育	66
娛樂	27
新聞	66
英文拼寫	4

表 6. AIShell 語料庫語者資訊
[Table 6. AIShell corpus speaker information]

年齡範圍	語者數	地區	語者數
16 - 25	316	北方	333
26 – 40	71	南方	56
> 40	13	其他	11
合計	400	合計	400

表 7. AIShell 語料庫資訊
[Table 7. AIShell corpus information]

語者總數		音節總數		檔案總數	
男	186	男	939132	男	65205
女	214	女	1101080	女	76395
合計	400	合計	2040212	合計	141600

4. 深層類神經網路模型配置 (Deep Neural Network Model Configuration)

CLDNN 為近年來被提出(Sainath, Vinyals, Senior & Sak, 2015)適合用來建立聲學模型的一種架構，其名稱來源為卷積類神經網路(CNN)加上長短期記憶(LSTM)後再接上深層類神經網路(DNN)，普遍認為，CNN 能夠學習特徵參數在頻域上的變化程度，LSTM 則擅長時域上的模型建立，最後 DNN 適合將特徵映射至更可分離的空間上。

此主要模型亦使用 TCC300 作為訓練語料，特徵參數的抽取也是 40 維之 Fbank，本研究使用的 LSTM 帶有映射層及窺視孔，詳細架構參見圖 1，虛線連結部分即為窺視孔作用之途徑，目的在於讓閘門做決定時能同時考慮短期記憶與長期記憶，而映射層之目的在於降低 LSTM 輸出或遞迴的神經元數量，降低模型總參數量，幫助網路訓練更為快速，和 CNN 後連接的降維全連接層有異曲同工之妙。

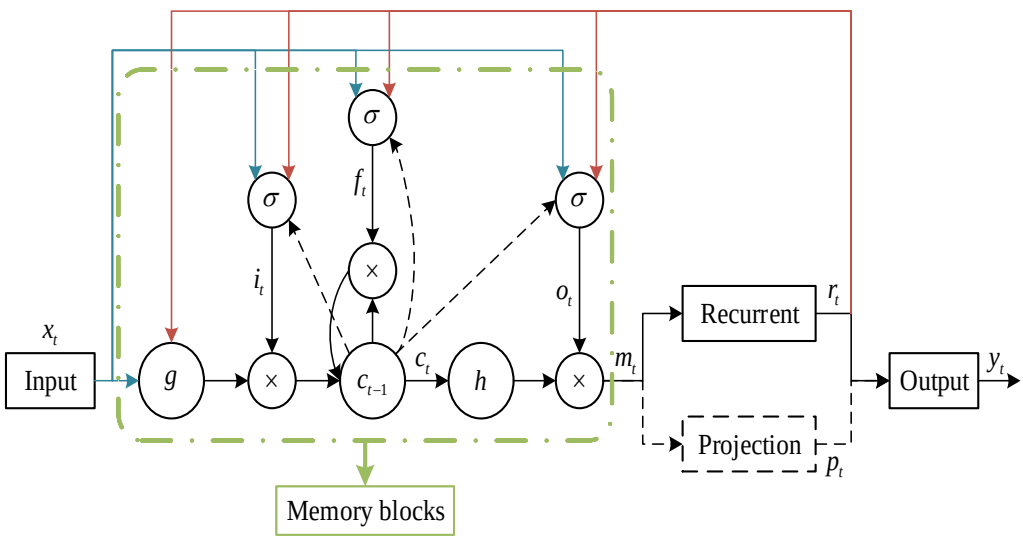


圖 1. 長短期記憶內部結構圖
[Figure 1. Long Short-Term Memory internal structure]

每一層 LSTM 中心細胞數目皆為 512，映射(Projection)層及遞迴(Recurrent)層數目皆為 256，而在訓練過程中，為了避免梯度產生爆炸，在遞迴的過程中會設定限幅閥值

(Clipping-threshold)為 30，即當梯度大於此閥值時，將梯度設定為 30，如此一來便解決 梯度在反向傳播的時候數值過大的問題。

另外，在 DNN 部分，為了解決層數過多導致學習困難的問題，有研究提出批次正規化(Batch normalization, BN)方法(Ioffe, & Szegedy, 2015)，將每一層 DNN 之輸出依照小型批次數(mini-batch)進行正規化，如此一來就可以大幅增加訓練學習率 讓模型訓練加速以及避免層數過深而造成的過度擬合(Over-fitting)的問題。

5. 語言模型之建立 (Language Model Establishment)

本研究目的於建立一中文大詞彙辨識系統，因此需要建立語言模型，並加入至系統中，辨識出中文詞彙序列。如圖 2 所示，建立流程為：將文字語料經國立交通大學語音處理實驗室王逸如老師撰寫之繁體中文斷詞器進行斷詞，後將文字進行正規化、移除冗餘贅字、取代同義異字詞(Variant Word, VW)等前處理，接著依照詞頻(Term Frequency, TF)及檔案頻率(Document Frequency, DF)進行選詞，一般來說，語音辨識系統之語言模型需要 TF 高及 DF 亦高之詞彙，本研究選擇了八萬詞、十萬詞及十二萬詞分別建立三個 3-gram 語言模型，而最後須將前處理置換的同義異字詞置換回來，詳細內容在五之(二)章節解說，最後以有限狀態轉換機表示此語言模型。

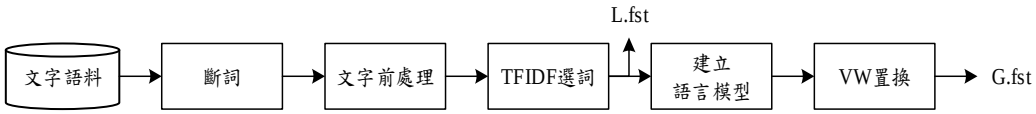


圖 2. 中文語言模型建立流程圖
[Figure 2. Chinese language model establishment flow chart]

5.1 文字語料庫簡介 (Introduction to the Text Corpus)

本研究用於訓練與研模型之文字語料庫共約 4.4 億個詞彙，包含以下：

- 光華雜誌(Sinorama)：內容為一般雜誌之文章，資料年份介於 1976 至 2000 年。
- NTCIR：為一個建立資訊檢索系統的標竿測試集，其內容由數種不同學科領域文章構成。
- 中研院平衡語料庫(Sinica)：由中研院收集，內容包含多種主題，以語言分析研究為目的之資料庫。
- Chinese Gigaword：由 Linguistic Data Consortium (LDC)整合發行，內容包括台灣中央社、北京新華社等國際新聞。
- 中文維基百科語料(Wiki)：中文維基百科內容廣泛，且資訊較新，能使語言模型更為多元，且增加資料庫。
- TCC300：包含詞、短句、長句，內容由中研院 500 萬詞標示語料庫中選取。

5.2 形音義分合詞前處理 (Preprocess of Variant Words)

漢字具有三大要素：「形、音、義」，其中字義為語文之核心，字形、字音皆因字義而存在。而在文字的演進當中，有些字形變的不一致，或者因為沒有創立而借用，甚至可能是錯用，各種複雜的因素導致漢字形成了「多形、歧音、異義」的狀況，因此目前漢字在使用上呈現字形不一、字音分歧、且字義寬廣的特性。

中文語音辨識因形音義之不同，有些詞類是可以合併的，如表 8 所示，大致上可以分為三類，即同形異音、異形同音及異形異音，置換原則是建立在字義相同之上，目的是為了讓文章中同義詞正規化，以利選詞時容納更多詞彙，但是在語言模型建立後，辨識端會產生一個狀況：異音類的同義字無法被搜尋到，如範例中的「禮拜一」被置換成「週一」，因此在語言模型中無法找到「禮拜一」這個詞彙，因此我們在語言模型建置的最後一步，需要處理不同發音之同義異字詞的置換，將「週一」展開成「週一」、「星期一」及「禮拜一」。本研究使用之同義異字詞表(variant word table)為 4261 詞。

表 8. 形音義分合詞範例
[Table 8. Example of variant words]

形音義分合詞類型	置換前文字	置換後文字
同形異音	爸	爸爸
	媽	媽媽
異形同音	手表	手錶
	瓩	千瓦
異形異音	禮拜一	週一
	星期一	週一

6. 實驗結果分析與討論 (Analysis and Discussion of Experimental Results)

本章節將進行實驗結果的分析與探討，其中包含使用無文法(Free-grammar)之語言模型測試音節錯誤率(Syllable Error Rate, SER)，以及加入不同詞典大小之語言模型測試詞錯誤率(Word Error Rate, WER) 與其即時係數(Real-Time Factor, RTF)。

即時係數如式(1)所示，表示平均一個音框需要解碼(decode)之時間，又因為本研究設定音框之間隔為 10ms，因此可以解釋為每秒辨識系統所需之解碼時間，若建立之系統為即時系統(Real-time System)，則 RTF 須小於 1.0；辨識錯誤的分析則可以分為以下三種錯誤：取代型錯誤(Substitution)、插入型錯誤(Insertion)及刪除型錯誤(Deletion)；而辨識錯誤率計算方式如式(2)所示。

解碼過程我們使用維特比演算法(Viterbi algorithm)，透過神經網路輸出狀態序列，搜尋計算找出最佳路徑，但是一般維特比算法過於耗時，因此加入光束搜尋演算法(Beam searching algorithm)，設定最大存活狀態數(Max-active states)及光束值(Beam)，找出當下

音框所有可能路徑(Hypotheses)，並刪除分數之光束臨界值，即當下路徑與最高分差大於光束值，則刪除該路徑，最後將狀態數控制於最大存活狀態數下，如此雖然會犧牲些許辨識率，但能大幅提升辨識速度，本研究設定最大存活狀態數為 7000、光束值為 15.0。

$$RTF = \frac{Seconds}{Frames} \times 100 \tag{1}$$

$$ER = \frac{S + I + D}{N} \times 100\% \tag{2}$$

6.1 各式類神經網路聲學模型辨識結果 (Various Types of Neural Network Acoustic Model Recognition Results)

為了探討遞迴式類神經網路對於聲學模型之影響，本實驗設計四組模型，使用的訓練語料皆為 TCC300，CDNN 為一般卷積類神經網路(CNN)結合深層類神經網路(DNN)，輸入之間彼此獨立，並沒有記憶特性；LDNN 為長短期記憶(LSTM)結合深層類神經網路，多了時間軸之資訊，過去的隱藏層狀態被保留，且透過閘門篩選控制，避免發生過擬現象；而 CLDNN 則結合以上三種類神經網路，詳細架構如圖 3 所示。

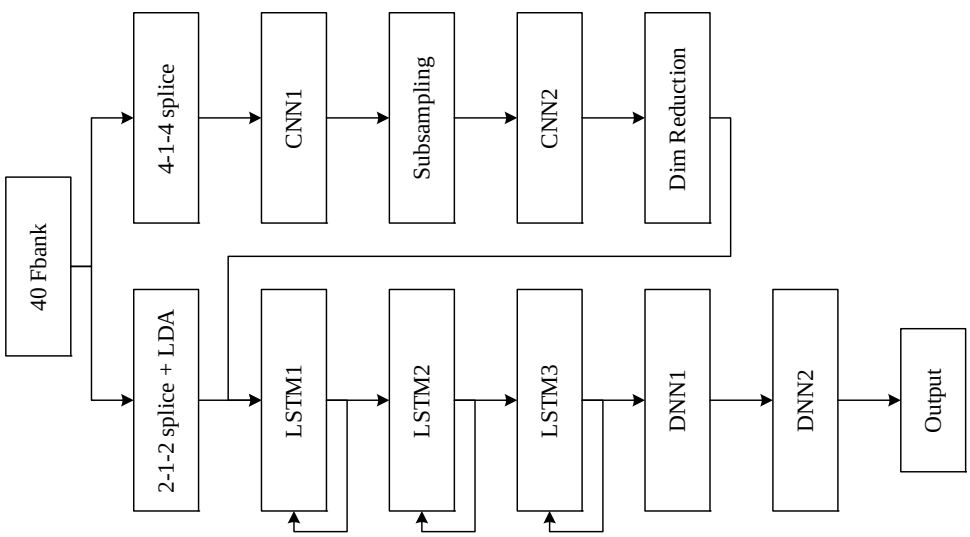


圖 3. CLDNN 模型架構圖
[Figure 3. CLDNN model architecture diagram]

測試語料部分亦選擇使用 TCC300，音節數為 26357，辨識結果如表 9 所示，卷積類神經網路對比傳統類神經網路而言，對於特徵的學習是有幫助的，相對改善率約 5%，而長短期記憶模型在聲學模型的建立上則是非常有益處的，相對改善率高達約 25%，但是使用遞迴式類神經網路在解碼時相對耗時，對照 CDNN 及 CLDNN 之 RTF，將近多出兩倍的時間。

表 9. 各式類神經網路模型之音節錯誤率
[Table 9. Syllable error rate of various neural network models]

Model	SER (%)	RTF
DNN	21.17	0.04
CDNN	19.52	0.05
LDNN	15.72	0.10
CLDNN	15.23	0.15

另外，本實驗亦使用鏈式模型(Chain model) (Povey *et al*., 2016)建構聲學模型，一般聲學模型的訓練使用的是最大化相似度(Maximum Likelihood, ML)，用於最大化模型及其特徵參數之相似度；鏈式模型則是使用最大交互資訊法則(Maximum Mutual Information, MMI)進行訓練如式(3)，其中 $P(W)$ 表示給定逐字文本(Transcription)中序列 W 之語言模型機率，而交互資訊可以拆成兩項相減，M^{num} 表示參考文本序列，M^{den} 表示所有可能之文本序列，最大化 F_{MMI} 表示讓參考文本的路徑機率 $P(O_r|W_r)$ 在所有路徑中最為突出，但是一般語言模型皆建立在詞彙(word)上，這會導致訓練過程效率不彰，因此鏈式模型在訓練上，會先以音素(phone)為單位，建立一個 4-gram 之語言模型，作為訓練時參考用。另外參考圖 4 及圖 5，鏈式模型使用降低 3 倍之音框速率 (Sak, Senior, Rao & Beaufays, 2015)，即一次觀察 30ms 之音框，以及更為簡單的 HMM 拓樸圖，一個音素(phone)僅用一個 HMM 描述，因此鏈式模型在解碼時比一般類神經網路模型加速三倍左右，實驗結果如表 10 所示，Chain-CLDNN 在音節錯誤率以及 RTF 都表現較 CLDNN 模型佳。

$$F_{MMI} = \sum_{r=1}^{R} \log \frac{P(O_r|W_r)P(W_r)}{\sum_W P(O_r|W)P(W)} = \log P(O|M^{num}) - \log P(O|M^{den}) \tag{3}$$

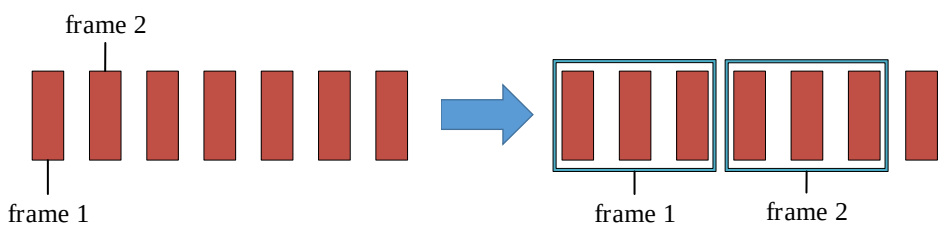


圖 4. 鏈式模型音框速率示意圖
[Figure 4. Chain model frame rate diagram]

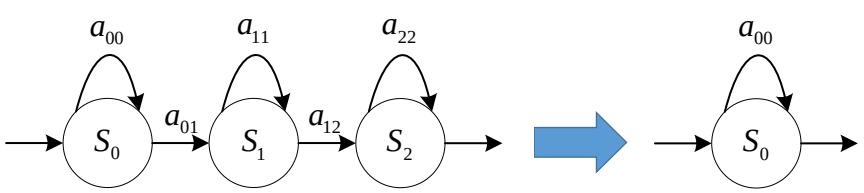


圖 5. 鏈式模型使用之 HMM 拓樸
[Figure 5. HMM topology used by chain models]

表 10. Chain/Non-chain 模型音節錯誤率
[Table 10. Chain/Non-chain model syllable error rate]

Model	SER (%)	RTF
CLDNN	15.23	0.15
Chain-CLDNN	13.66	0.05

6.2 加大訓練語料對辨識率的影響 (Impact of Increasing Training Corpus on Recognition Rates)

在深度學習(Deep learning)或是機器學習中，常常會遭遇模型過度擬合(Over-fitting)之問題，若能順利解決，則能使得模型訓練得更為深層，方法除了本研究所使用的批次正規化之外，另一個方法就是直接增加訓練語料，然而在資料有限的情形下，可以透過資料轉換技術來增加訓練資料，此概念在影像處理(Image Processing)領域已被實現(Krizhevsky, Sutskever & Hinton, 2012)，圖片可以透過旋轉(Rotation)、翻轉(Flip)、縮放(Zoom)、平移(Shift)、尺度轉換(Rescale)等方法產生新的圖片。

語音辨識方面，亦能使用類似的方法，比如改變音檔之音高(Pitch)、節奏(Tempo)、語速(Speed)等產生出假造之資料，擴充語料庫，本研究除了使用 TCC300、NER 及 AIShell 語料庫，亦利用上述方法產生語速 1.1 及 0.9 之擾動語料(Speed perturbation data)，並加入訓練語料。

實驗結果如表 11 所示，首先針對一般 CLDNN 模型，加入 AIShell 語料庫，訓練語料由原本的 24 小時增加到 186.4 小時，雖然 AIShell 語料庫來自中國各地口音，但是對於音節辨識率之相對改善率仍高達約 15.5%，若以語者個別分析，如圖 6 所示，則可以發現到，主要降低音節錯誤率之貢獻來自原本音節錯誤率高之語者，對於錯誤率低之語者無太多改善，換句話說，辨識系統更具強健性(Robustness)。

接著加入屬於台灣腔調之 NER 自發性語料，訓練語料增加至 297.9 小時，並使用上述之擾動語速方法，增加至 900.7 小時，逐一訓練出 CLDNN 鏈式模型，實驗結果如表 12、圖 7 所示。

表 11. 使用不同訓練語料之 CLDNN 模型比較
[Table 11. Comparison of CLDNN models using different training corpora]

Model	Training data	SER (%)
CLDNN	TCC300	15.23
	TCC300+AIShell	12.87

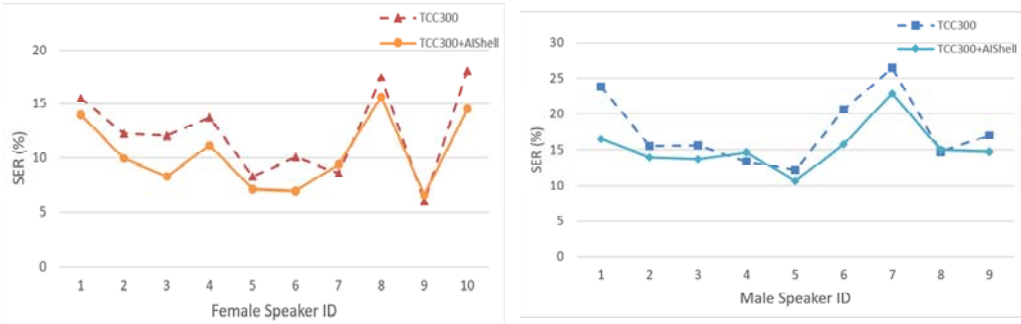


圖 6. TCC300 女性/男性測試語者之音節辨識率
[Figure 6. Syllable error rate of TCC300 female/male testers]

表 12. 使用不同訓練語料之 Chain-CLDNN 模型比較
[Table 12. Comparison of Chain-CLDNN models using different training corpora]

Model	Training data	SER (%)
Chain-CLDNN	TCC300	13.66
	TCC300+AIShell	11.97
	TCC300+AIShell_sp	11.49
	TCC300+AIShell+NER_sp	8.92

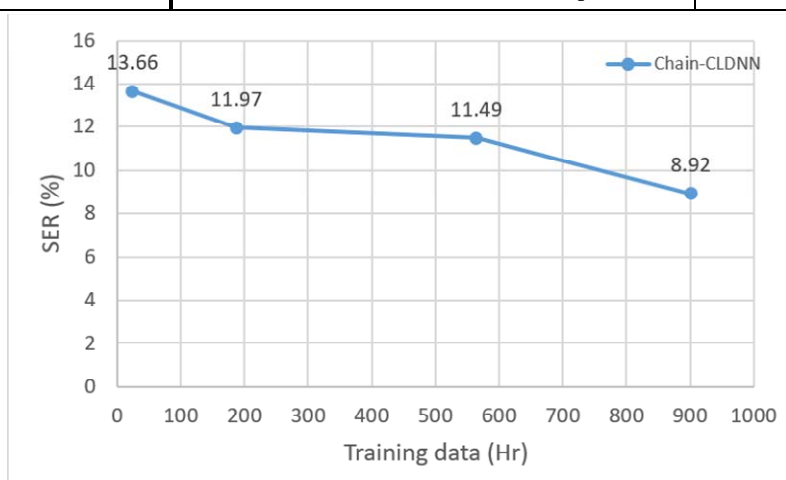


圖 7. 訓練語料量對音節辨識率之影響
[Figure 7. The effect of the amount of training corpus on the syllable recognition rate]

6.3 加入語言模型並探討不同環境對辨識率之影響 (Add Language Models and Explore the Impact of Different Environments on Recognition Rates)

本節我們選擇使用透過 900.7 個小時訓練語料建立之 Chain-CLDNN 模型作為聲學模型，測試語料部分選擇 1. 朗讀語速之 TCC300、2. 自發性語音且背景無噪音之 NER-clean、3. 自發性語音且具雜訊之 NER-other，音節辨識率如表 13，後加入三組 Tri-gram 語言模型，分別將詞典大小設定為：八萬詞、十萬詞及十二萬詞，分別測試其最佳辨識結果(Oracle)，即當聲學模型輸出之音素序列皆完全正確情況下，語言模型辨識出之詞錯誤率，以及 EDO(Error Due to OOVs)，即 OOV 造成之錯誤率，平均一個 OOV 影響 2.103 個詞，最後結合聲學模型解碼計算出詞錯誤率(WER)，如表 14 至表 16 所示。RTF 部分對於三個測試語料分別為 0.27、0.48 及 0.59。然而本實驗室之文字語料庫大多取自新聞文章，domain 相對偏向 TCC300 測試集，而 NER 測試集則多為談話性節目，因此本實驗利用 NER 之訓練語料逐字稿進行語言模型之調適，如式(4)所示，實驗結果如表 17 所示，WER 獲得大幅度的改善。

$$LM_{adapt} = 0.3LM_{ori} + 0.7LM_{ner} \tag{4}$$

表 13. Chain-CLDNN 模型對於各測試集之音節錯誤率
[Table 13. The syllable error rate of the Chain-CLDNN model for each test set]

Model	Test data	SER (%)
Chain-CLDNN [TCCAINER-sp]	TCC300	8.92
	NER-clean	16.89
	NER-other	22.14

表 14. 八萬詞語言模型辨識結果
[Table 14. 80K word LM recognition results]

Model	Test data	80K-LM		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	7.73	6.74	5.85
	NER-clean	24.95	9.39	2.48
	NER-other	31.92	11.92	3.91

表 15. 十萬詞語言模型辨識結果
[Table 15. 100K word LM recognition results]

Model	Test data	100K-LM		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	7.12	6.06	5.19
	NER-clean	24.80	9.27	2.25
	NER-other	31.69	11.57	3.26

表 16. 十二萬詞語言模型辨識結果
[Table 16. 120K word LM recognition results]

Model	Test data	120K-LM		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	6.56	5.34	4.52
	NER-clean	24.72	9.05	2.02
	NER-other	31.61	11.42	2.92

表 17. 十二萬詞調適語言模型辨識結果
[Table 17. 120K word adaptation LM recognition results]

Model	Test data	120K-LM-Adapt		
		WER (%)	Oracle (%)	EDO(%)
Chain-CLDNN [TCCAINER-sp]	TCC300	7.79	5.87	4.52
	NER-clean	15.12	4.00	2.02
	NER-other	21.66	4.74	2.92

7. 結論與未來展望 (Conclusion and Future Prospects)

本論文使用 Kaldi speech recognition toolkit 來實現結合卷積類神經網路、長短期記憶及深層類神經網路的聲學模型(CLDNN)，經過各式類神經網路模型的比較後，確定長短期記憶對於聲學模型的建構上助益良多，也確定卷積類神經網路對於特徵的學習對整體模型有幫助，且加入大量不同來源之訓練語料(NER、AIShell)，並使用資料增強轉換技術，能使模型之強健度提升，最後再以實驗室 4.4 億詞彙量文本訓練 Tri-gram 語言模型，以建構中文大詞彙語音辨識系統，從實驗結果顯示，系統之詞辨識率與語言模型有非常密切的關聯，也就是說，測試語料之領域依存性(domain dependence)相當高。

本實驗建構之中文辨識系統雖然在朗讀語速及自發性且無噪音環境下的辨識率(6.56%、15.12%)有不錯的表現，但是在自發性且環境雜訊高的環境下，詞錯誤率仍高達 21.66%，因此在聲學模型方面如何抗噪，亦是一個研究課題，另外許多研究加入 I-vector

作為特徵參數進行聲學模型訓練 (Madikeri, Dey, Motlicek & Ferras, 2016)，目的為了學習語者特性，增加模型強健性，訓練語料不足的方面，可以使用半監督式學習 (semi-supervised learning) (Manohar, Hadian, Povey & Khudanpur, 2018)，蒐集無轉寫文本之語料，透過辨識結果之信心分數決策是否加入為訓練語料；至於語言模型部分，解決人名造成 OOV 之問題，且將文本進行分類，以建構出不同 domain 之語言模型，以及快速進行調適語言模型之建立與轉換。

參考文獻 (References)

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533-1545. doi: 10.1109/TASLP.2014.2339736

Abdel-Hamid, O., Mohamed, A., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *Proceedings of ICASSP 2012*, 4277-4280. doi: 10.1109/ICASSP.2012.6288864

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *Proceedings of 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. doi: 10.1109/ICSDA.2017.8384449

Ioffe, S., & Szegedy. C. (2015). Batch normalization:Accelerating deep network training by reducing internal covariate shift. In *Proceedings of ICML' 15*, *37*, 448-456.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, *1*, 1097-1105.

Madikeri, S., Dey, S., Motlicek, P., & Ferras, M. (2016). *Implementation of the standard i-vector system for the kaldi speech recognition toolkit* (Idiap- RR Idiap-RR-26-2016). Retrieved from IDIAP Research Institute website: http://publications.idiap.ch/downloads/reports/2016/Madikeri_Idiap-RR-26-2016.pdf

Manohar, V., Hadian, H., Povey, D., & Khudanpur, S. (2018). Semi-supervised training of acoustic models using lattice-free MMI. In *Proceedings of ICASSP 2018*. doi: 10.1109/ICASSP.2018.8462331

Mohamed, A. (2014). *Deep Neural Network Acoustic Models for ASR* (Doctoral dissertation). Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/44123/1/Mohamed_Abdel-rahman_201406_PhD_thesis.pdf

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., …Vesely´, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of IEEE 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*, 2751-2755. doi: 10.21437/Interspeech.2016-595

Reynolds, D. A. (2009). Gaussian mixture models. In S. Z. Li (Eds.), *Encyclopedia of Biometrics* (pp. 659-663) 2009. doi: 10.1007/978-0-387-73003-5_196

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional Long Short-Term Memory Fully Connected Deep Neural Networks. In *Proceedings of 2015 IEEE International Conference on Acoustics Speech and Signal Processing*. doi: 10.1109/ICASSP.2015.7178838

Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. Retrieved from arXiv:1402.1128

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., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. In *Proceedings of Sixteenth Annual Conference of the International Speech Communication Association*.

Zhang, X., Trmal, J., Povey, D., & Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *Proceedings of ICASSP 2014*. doi: 10.1109/ICASSP.2014.6853589

Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 19-34

結合鑑別式訓練與模型合併於半監督式語音辨識之研究

Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition

羅天宏*、陳柏琳*

Tien-Hong Lo and Berlin Chen

摘要

近年來鑑別式訓練(Discriminative training)的目標函數 Lattice-free Maximum Mutual Information (LF-MMI)在自動語音辨識(Automatic speech recognition, ASR)上取得了重大的突破。儘管 LF-MMI 在監督式環境下斬獲最好的成果，然而在半監督式設定下，由於種子模型(Seed model)常因為語料有限而效果不佳。且由於 LF-MMI 屬於鑑別式訓練之故，易受到轉寫正確與否的影響。本論文利用兩種思路於半監督式訓練。其一，引入負條件熵(Negative conditional entropy, NCE)權重與詞圖(Lattice)，前者是最小化詞圖路徑的條件熵(Conditional entropy)，等同對 MMI 的參考轉寫(Reference transcript)做權重平均，權重的改變能自然地加入 MMI 訓練中，並同時對不確定性建模。其目的希望無信心過濾器(Confidence-based filter)也可訓練模型。後者加入詞圖，比起過往的只使用最佳辨識結果，可保留更多假說空間，進而提升找到參考轉寫(Reference transcript)的可能性；其二，我們借鑒整體學習(Ensemble learning)的概念，使用弱學習器(Weak learner)修正彼此的錯誤，分為假說層級合併(Hypothesis-level combination)和音框層級合併(Frame-level combination)。實驗結果顯示，加入 NCE 與詞圖皆能降低詞錯誤率(Word error rate, WER)，而模型合併(Model combination)則能在各個階段顯著提升效能，且兩者結合可使詞修復率(WER recovery rate, WRR)達到 60.8%。

關鍵詞：自動語音辨識、鑑別式訓練、半監督式訓練、模型合併、LF-MMI

* 國立臺灣師範大學資訊工程研究所
Institute of Linguistics, National Taiwan Normal University
E-mail: {teinhonglo, berlin}@ntnu.edu.tw

Abstract

In recent years, the so-called Lattice-free Maximum Mutual Information (LF-MMI) criterion has been proposed with good success for supervised training of state-of-the-art acoustic models in various automatic speech recognition (ASR) applications. However, when moving to the scenario of semi-supervised acoustic model training, the seed models of LF-MMI are often show inadequate competence due to limited available manually labeled training data. This is because LF-MMI shares a common deficiency of discriminative training criteria, being sensitive to the accuracy of the corresponding transcripts of training utterances. This paper sets out to explore two novel extensions of semi-supervised training in conjunction with LF-MMI. First, we capitalize more fully on negative conditional entropy (NCE) weighting and utilize word lattices for supervision in the semi-supervised setting. The former aims to minimize the conditional entropy of a lattice, which is equivalent to a weighted average of all possible reference transcripts. The minimization of the lattice entropy is a natural extension of the MMI objective for modeling uncertainty. The latter one, utilizing word lattices for supervision, manages to preserve more cues in the hypothesis space, by using word lattices instead of one-best results, to increase the possibility of finding reference transcripts of training utterances. Second, we draw on the notion stemming from ensemble learning to develop two disparate combination methods, namely hypothesis-level combination and frame-level combination. In doing so, the error-correcting capability of the acoustic models can be enhanced. The experimental results on a meeting transcription task show that the addition of NCE weighting, as well as the utilization of word lattices for supervision, can significantly reduce the word error rate (WER) of the ASR system, while the model combination approaches can also considerably improve the performance at various stages. Finally, fusion of the aforementioned two kinds of extensions can achieve a WER recovery rate (WRR) of 60.8%.

Keywords: Automatic Speech Recognition, Discriminative Training, Semi-supervised Training, Model Combination, LF-MMI

1. 緒論 (INTRODUCTION)

近年來基於類神經網路的聲學模型(Deep neural network-hidden Markov model, DNN-HMM)取得重大的突破(Seide, Li & Yu, 2011) (Dahl, Yu, Deng & Acero, 2012)。傳統的 DNN-HMM 透過交互熵訓練(Cross-Entropy training, CE)和鑑別式訓練(Discriminative training) (Valtchev, Odell, Woodland & Young, 1996) (Valtchev, Odell, Woodland & Young, 1997) (Woodland & Povey, 2002) ，兩階段的訓練提升聲學模型的辨識效果。尤其是第二

階段的鑑別式訓練，由於提升效果顯著，吸引了許多研究者的目光。過往於鑑別式訓練的研究主題種類繁多，如 MMI (Bahl, Brown, de Souza & Mercer, 1986), MCE (Juang, Hou & Lee, 1997), MPE (Povey & Woodland, 2002), sMBR (Kaiser, Horvat & Kacic, 2000) (Gibson & Hain, 2006)和 bMMI (Povey *et al,* 2008)等。最近，隨著語料的增長，不透過第一階段的 CE 訓練，將鑑別式訓練做一階段訓練的端對端訓練(End-to-End)也越來越流行。目前兩種主流的端對端架構的目標函數為 CTC (Graves, Fernández, Gomez & Schmidhuber, 2006)和 Lattice-free MMI (LF-MMI) (Povey *et al*., 2016)。前者在語料非常充足(通常大於 500 小時)的情況下，表現可以媲美甚至超越傳統的二階段方法。而後者證明了在語料較為缺乏的情況下，儘管效能會下降，但仍可以勝過前者，因此成為了目前最具魅力的研究主題。在(Povey *et al*., 2016)的實驗中展示，基於 LF-MMI 的目標函數下，可從亂數初始化參數後，以鑑別式準則訓練類神經網路。實驗結果顯示 LF-MMI 效果更勝兩階段訓練的 sMBR 一籌，且還可結合 sMBR 進一步提升辨識結果。然而，在這樣的訓練準則下，仍受限於需大量訓練語料的問題(Data hungry)。進一步來說，便是在小語料庫上的表現(通常小於 100 小時)仍無法勝過在大語料庫的優異結果(Pundak & Sainath, 2016)。

在現實生活中，相對於高成本的人工轉寫語料，未轉寫語料十分容易取得。當我們沒辦法取得大量的轉寫語料時，就必須更有效地利用大量的未轉寫語料訓練模型。換句話說，探索存在於未轉寫語料的線索，並加入半監督式訓練的聲學模型就更顯重要。另一方面，半監督式訓練用途多元，不僅可用在自動語音辨識的訓練，也同樣適用於自動轉寫(Automatic labeling)及遷移學習(Transfer learning)、語者調適(Speaker adaptation)。過往的研究於半監督式聲學模型(Zavaliagkos, Siu, Colthurst & Billa, 1998)，最常見的訓練方法是自我訓練(Self-training) (Vesely, Hannemann & Burget, 2013) (Grezl & Karafiát, 2013) (Zhang, Liu & Hain, 2014)。自我訓練的架構主要分成兩階段，第一階段為利用轉寫語料訓練種子模型直到穩定，第二階段則是利用種子模型辨識未轉寫語料，並以此為答案重新訓練模型。在第二階段的辨識結果與真實答案難免會有誤差，因此會再加入信心過濾器(Confidence-based filter) (Lamel, Gauvain & Adda, 2002) (Chan & Woodland, 2004) (Liu, Chu, Lin & Chen, 2007) 挑選訓練語料，該動作可在不同層級上進行，分為音框層級(Vesely, Hannemann & Burget, 2013)、詞層級(Thomas, Seltzer, Church & Hermansky, 2013)以及語句層級(Grezl & Karafiát, 2013) (Vesely *et al*., 2013) (Zhang *et al*., 2014)。

最近 LF-MMI 訓練方法在 ASR 取得了重大的突破。有別於傳統的二階段訓練，LF-MMI 提供更快的訓練與解碼，同時在模型準度上取得目前最優異的表現。儘管 LF-MMI 在監督式環境下獲得最好的成果，但在半監督式環境下的研究成果仍然有限。在過往的研究中，鑑別式訓練的好壞很大層度地仰賴於訓練語句的正確性(Mathias, Yegnanarayanan & Fritsch, 2005) (Yu, Gales, Wang & Woodland, 2010) (Cui, Huang & Chien, 2011)，而屬於鑑別式訓練的 LF-MMI 也同樣對於正確性十分敏感。然而，在半監督式訓練過程中，由於第二階段訓練時無法保證語句的正確性，因此在過往研究常著重於二階段鑑別式訓練前的信心過濾器，如(Liu *et al*., 2007) (Mathias *et al*., 2005)將音框層

級的信心過濾器加入鑑別式訓練。而在(Walker, Pedersen, Orife & Flaks, 2017) 加入語句層級的信心過濾器以及後處理最佳辨識結果(One-best result)。本論文與(Manohar, Hadian, Povey & Khudanpur, 2018) (Manohar, Povey & Khudanpur, 2015)相同，是將詞圖的不確定性以條件熵(Conditional entropy)的形式加入，保留整個詞圖來做二階段的訓練。本論文與其不同的是，我們將這樣的方法做在更口語化的會議語料，以及基於這個方法之上，利用整體學習的觀念，進一步地探討模型合併帶來的成效。

整體的模型合併在自動語音辨識上能取得優於單一模型的成果(Fiscus, 1997) (Evermann & Woodland, 2000) (Deng & Platt, 2014) (Xu, Povey, Mangu & Zhu, 2011)，這樣效能的進步歸功於下列幾點，各別模型可以修正彼此的錯誤；減少選擇到較差模型的可能性；增加整體模型搜尋時的假說空間(Dietterich, 2000)，用以修正訓練時的問題。如語料選擇(Data selection)、目標函數(Objective function)、模型(Model)。這裡我們期待利用整體學習增加的假說空間，解決在半監督式訓練時有限語料造成效能降低的問題。訓練的過程為各別訓練每個模型，接著在訓練結束後的階段加入合併模型的技術，讓模型修正彼此的錯誤，進一步提升效能。這裡我們採用兩種不同層級的合併方法，音框層級的合併(Frame-level combination or score fusion) (Deng & Platt, 2014)，以及假說層級的合併(Hypothesis-level combination) (Fiscus, 1997) (Xu *et al*., 2011)。在(Senior, Sak, Quitry, Sainath & Rao, 2015)的研究中音框層級合併無助於 CTC 的表現，而 LF-MMI 被視為 CTC 的延伸，因此探討半監督式 LF-MMI 的合併結果是具有價值的事情。

本論文的實作目的便是在語料缺乏的半監督式環境下，使用負條件熵與詞圖輔助 LF-MMI 的訓練，並利用模型合併技術，進一步提升模型的辨識結果。我們希望即使在語料不足的情況下，仍能達到不錯的辨識效果，甚至媲美原先轉寫語料的訓練結果。

2. 相關文獻回顧 (RELATED WORK)

本篇論文於半監督式的環境研究 LF-MMI 的表現，並進一步利用模型合併的技術提升效能。其中相關研究可分為三大方向：半監督式聲學模型；MMI 與 LF-MMI 的改變；最後則是模型合併技術。

2.1 半監督式訓練於聲學模型 (Semi-supervised Acoustic Modeling)

半監督式聲學模型目的是解決下列問題：低資源的語料庫、大量的未轉寫語料、測試語料與訓練語料的不匹配。首先，充足的語料庫是讓目前最新穎的 ASR 系統可以表現優異的原因之一，但我們擁有的轉寫語料通常不大；其次，儘管取得足夠的轉寫語料十分困難，但取得未轉寫語料卻容易得多，要如何利用好大量的未轉寫語料便成了重要的問題；最後，也是最廣泛的問題，訓練與測試環境的不匹配。相關研究裡最常見的方法為自我訓練(Self-training) (Zavaliagkos *et al*., 1998) (Vesely *et al*., 2013) (Grezl & Karafiát, 2013) (Zhang *et al*., 2014)。自我訓練的步驟分為兩階段，首先使用轉寫語料訓練種子模型直到穩定(通常為 CE 訓練，但也可加入鑑別式訓練)，第二階段則利用種子模型辨識未轉寫語料，加入信心過濾器(Lamel *et al*., 2002) (Chan & Woodland, 2004) (Liu *et al*., 2007) 篩選

訓練語料，過濾可能會影響訓練的語料，再重新訓練模型。而信心過濾器(Confidence filter)可在音框層級(Vesely *et al*., 2013)、詞層級(Thomas *et al*., 2013)、語句層級(Grezl & Karafiát, 2013) (Vesely *et al*., 2013) (Thomas *et al*., 2013)多種層級上進行。由於鑑別式訓練對於訓練語句的正確性十分敏感(Mathias *et al*., 2005) (Yu *et al*., 2010) (Cui *et al*., 2011)，因此過往的研究著重於信心過濾器的選擇。在(Liu *et al*., 2007) (Mathias *et al*., 2005)中將音框層級的信心過濾器加入鑑別式訓練。而在(Walker *et al*., 2017)在鑑別式訓練中加入語句層級的信心過濾器以及後處理最佳辨識結果。(Manohar *et al*., 2018) 則將詞圖加入在半監督式 LF-MMI 的訓練。

2.2 Lattice-free Maximum Mutual Information

2.2.1 MMI

條件最大化可能性(Conditional maximum likelihood, CML) (Nadas, 1983)的目標函數是在給予聲學特徵 O 和模型參數下，估測轉寫(Transcript)的對數可能性。分子為正確轉寫(Reference transcript)的機率，而分母為所有可能答案的機率。因為一些歷史的原因，CML 成為了我們目前熟知的 MMI (Bahl *et al*., 1986)，式子如下：

$$\mathcal{F}^{\text{MMI}} = \sum_u \log P(S_u|O_u, \lambda) \tag{1}$$

式(1)的 u 為語句，S_u為語句 u 的正確狀態序列(Reference state sequences)，O_u為語句 u 的聲學特徵，λ為模型參數。MMI 的目標便是最大化上述的式子。詳細的計算可透過貝式定理(Bayes' theorem)拆解成式(2)：

$$\mathcal{F}^{\text{MMI}} = \sum_u \log \frac{P(O_u|S_u,\lambda)P(S_u)}{\sum_{S'} P(O_u|S',\lambda)P(S')} \tag{2}$$

在式(2)中，S'_u為語句 u 的競爭狀態序列(Competing state sequence)。可透過鑑別式訓練，將模型目標函數定義成接近正確狀態序列和遠離競爭狀態序列。若要計算語句 u 在時間 t，而輸出層為$\mathbf{y}(u,t)$，則可偏微分式(1)：

$$\frac{\partial \mathcal{F}^{MMI}}{\partial y(u,t)} = \delta_{S_u:\mathbf{y}(u,t)} - \gamma^{DEN}_{\mathbf{y}(u,t)} \tag{3}$$

式(3)中的 $\delta_{S_u:\mathbf{y}(u,t)}$為指示函數(Indicator function)，當$\mathbf{y}(u,t)$的輸出屬於正確狀態序列S_u時為 1，反之則為 0。$\gamma^{DEN}_{\mathbf{y}(u,t)}$則表示$\mathbf{y}(u,t)$為正確狀態序列的事後機率(Posterior)，可表示如下：

$$\gamma^{DEN}_{\mathbf{y}(u,t)} = \sum_S \delta_{S:\mathbf{y}(u,t)} P(S|O_u, \lambda)$$

$$= \frac{\sum_S \delta_{S:\mathbf{y}(u,t)} P(O_u|S,\lambda)P(S)}{\sum_{S'} P(O_u|S')P(S')} \tag{4}$$

式(3)裡最繁雜的問題便呈現在式(4)，式(4)為計算所有可能存在於假說的競爭序列。在較早期的研究裡，學者們利用 CE 作預先訓練限制假說空間的大小，使得 MMI 的競爭序列可由有限的詞圖中產生。這樣是二階段的訓練取得了不錯的成果，但為了產生可能序列的詞圖，不僅需要多餘的 CE 訓練，且受限於 CE 的訓練，第二階段的 MMI 訓練僅能找到一階段 CE 訓練結果的局部最佳解。LF-MMI 主要解決的是式(4)的計算，使得不用一階段 CE 預先訓練產生詞圖，即可直接計算所有可能的競爭訓練。

2.2.2 LF-MMI

近年來，(Povey *et al*., 2016)中提出 LF-MMI，避開需要 CE 訓練產生詞圖的冗餘步驟，可視為 CTC (Graves *et al.*, 2006)的延伸架構。主要改變有四種，利用 4 連音素語言模型(Four-gram phone LM) 且不會退化小於 3 連音素語言模型(Tri-gram phone LM)，取代傳統鑑別式訓練時的詞圖，使得搜尋的假說空間減少；提出多種避免過度擬合(Overfitting)的訓練技巧，如多任務架構的 CE 正則項(CE-based regularization)，讓訓練能同時最佳化 LF-MMI 和 CE；採用類似 CTC 的兩個左到右狀態 HMM (2-state left-to-right HMM)的拓樸架構，且第一個狀態沒有 self-loop，相似於 CTC 的空白輸出(Blank)；最後的假設則是類神經網路的輸出沒有軟式最大化(Softmax)，因此不是狀態的事後機率，而是偽對數可能性(Pseudo log likelihood)。前兩者的改變使得 MMI 的訓練可在一階段的聲學模型便加入訓練，且式(4)也不是計算在候選詞圖上，而是完整搜尋(Full search)所有的可能序列，最終效果可媲美甚至超越兩階段的聲學模型訓練。後兩者的改變則是模仿 CTC 的架構，因此 LF-MMI 也可視為 CTC 的延伸架構。

2.3 模型合併技術 (Model Combination)

整體模型可藉由多個模型互補的假說空間，用以修正單一模型難以解決的問題。如語料選擇(Data selection)、目標函數(Objective function)、模型(Model)。為了實現最大的組合增益，在整體系統裡的模型必須單獨且準確(Dietterich, 2000)。在 DNN-HMM 的模型中，可引入五種多樣性。特徵多樣性，如隨機特徵投影(Random feature projection)；架構多樣性，如 DNN、LSTM；模型參數多樣性，如隨機初始化(Random Initialization)；輸出目標多樣性，如隨機森林(Random forest) (Dietterich, 2000)；轉換模型(Transition model)和語言模型(Language model)的多樣性。過往在語音辨識的模型合併可分為兩種，假說層級合併(Hypothesis-level combination) (Evermann & Woodland, 2000) (Xu *et al*., 2011)和音框層級合併(Frame-level combination or score fusion) (Deng & Platt, 2014)。ROVER(Fiscus, 1997).利用多個 ASR 產生的可能轉寫(n-best)結果的聯集，透過詞頻(Word frequency)或信心分數(Confidence score)合併成單詞轉換網路(Word translation network)，自動重新計搜索生成的網路，選擇得分最高的輸出序列；而(Xu *et al*., 2011)則是將多個模型的解碼結果的詞圖取聯集，得到一個新的詞圖。結果證明可以在最小化貝式決策風險解碼(Minimum Bayes-risk decoding)中，改進貝式風險的界限；在(Deng & Platt, 2014)中結合聲學模型的網路輸出並進行解碼；(Evermann & Woodland, 2000)利用維特比(Viterbi)產生

的詞圖與混淆網路(Confusion network)，其提供最可能的單詞假設及其相關單詞的事後機率，實驗結果優於 ROVER 的性能。雖然上述的幾種模型合併皆證明可優於單一模型的表現。然而，在過往的研究中，由於 CTC 的輸出為高峰分佈(Peaky distribution)，音框層級合併不僅無助於 CTC 模型，甚至會惡化原先的表現(Senioret *et al*., 2015)。而 LF-MMI 被視為 CTC 的延伸，且音框層級合併比假說層級合併來得更高效，因此探討半監督式 LF-MMI 的音框合併是具有價值的事情。

3. 基於 LF-MMI 的半監督式訓練 (SEMI-SUPERVISED TRAINING USING LF-MMI)

3.1 半監督式LF-MMI (Semi-supervised LF-MMI)

在有參考轉寫的情況下，傳統 MMI 估測方式為 CML，計算的式子為式(2)。然而在半監督式的環境下，未轉寫語料的自動轉寫(分子項)未必正確。因此在半監督式環境下，我們可將原先的式(2)改寫如下：

$$\mathcal{F}^{\text{SemiMMI}} = \sum_{s_u \in \mathcal{H}} \log \frac{P(O_u|S_u,\lambda)P(S_u)}{\sum_{S'} P(O_u|S',\lambda)P(S')} \tag{5}$$

上式的 u 為語句，S_u為語句 u 的正確狀態序列，但在半監督式環境下的S_u來自於種子模型產生的假說$\mathcal{H}$，因此不能保證其正確性。O_u為語句 u 的聲學特徵。S' 為語句 u 的競爭狀態序列，早期的聲學模型透過 CE 第一階段的訓練限制產生競爭序列的假說空間，使得競爭的序列只能從 CE 訓練後的詞圖中產生。而 LF-MMI 透過一些實作上的機制避開上述冗餘的步驟，可以在訓練時直接計算所有的競爭序列。

當我們計算式(5)分子項的正確序列，與過往只取最佳辨識結果的計算方式不同，而是將整個詞圖加入計算，透過設定光束(Beam)保留搜尋時的數量。保留越多就越可能搜尋到最佳答案，但同時會增長計算複雜度。其餘實驗設定與(Povey *et al*., 2016)中一致。

3.2 條件熵 (Conditional Entropy)

前一段中提到正確序列的S_u來自於種子模型產生的假說$\mathcal{H}$，我們不能保證其分子項的正確性。因此直接加入第二階段訓練是危險的行為，甚至會惡化原先模型的表現。在過往的研究中為了解決此問題，最常見的便是在第一階段和第二階段中間，加入信心過濾器排除分數過低的語句，用以確保訓練語句的「品質」，但挑選過濾器的門檻值並不容易且非常浪費訓練時間。有別於以往的排除訓練語句，我們希望在訓練時仍保留分數較低的語句，並與分數高的語句一起訓練。這裡我們在原先的向前向後算法(Forward-backward algorithm)加入了權重機制，並將原先的式(1)改寫如下：

$$\mathcal{F}^{\text{NCE}} = \sum_{u \in \mathcal{H}} \sum_{s} P(S_u|O_u,\lambda) \log P(S_u|O_u,\lambda) \tag{6}$$

上式為未轉寫語料的估測方式。式(6)與式(1)相似，但在計算可能的正確序列S_u時，加入

了$P(S_u|O_u,\lambda)$的權重於詞圖中，用以改變詞圖中的分數矩陣。式(6)進一步化簡成下式：

$$\mathcal{F}^{\mathrm{NCE}} = -\sum_u H(S_u|O_u,\lambda) \tag{7}$$

式(7)便是 NCE(Grandvalet & Bengio, 2005) (Huang & Hasegawa-Johnson, 2010)。我們可以稱式(7)為給予模型參數λ和聲學特徵O_u條件下，參考轉寫序列S_u的條件熵$H(S_u|O_u,\lambda)$。式(7)的改變可利用資訊量對轉寫的「品質」建模，並且自然地加入 LF-MMI 目標函數，在不用信心過濾器的情況下也能提升訓練結果。

4. 模型合併技術應用於聲學模型 (MODEL COMBINATION OF ACOUSITIC MODELING)

模型合併的成果可透過修正各別模型的錯誤、減少較差選擇的可能性、增加模型搜尋時的假說空間來達到更好的模型效能。在聲學模型的合併可分為音框層級和假說層級。兩者的比較紀錄於表 1，前者因為是在聲學模型的輸出直接合併，因此具有較快的即時性。後者則是在模型產生詞圖後合併，與解碼標準更相關，有較好的辨識結果。

表 1. 合併方式比較
[Table 1. Frame combination vs. hypothesis combination]

面向	音框層級合併	假說層級合併
解碼詞圖	• 強制共享詞圖中的時間同步的狀態 • 僅需處理整體的單個詞圖	• 不需要時間同步的狀態 • 需先各別處理整體模型數的詞圖再合併
事後機率	• 旨在產生更好的音框事後機率或觀察可能性，從而產生更好的詞圖	• 旨在產生更好的假說事後機率，其與解碼標準更密切相關

4.1 音框層級合併 (Frame-level Combination)

音框層級合併是根據某個時間點中音框輸出的對數可能性(Log likelihood)，給予不同的權重後合併。因為合併的是音框，所以必須保持輸出時間的同步。聲學模型的對數可能性是類神經網路的輸出。式子如下：

$$P(S_{ut}|O_{ut},\lambda) = \sum_{m=1}^{M} \alpha_m P(S_{ut}|O_{ut},\lambda_m) \tag{8}$$

式(8)中S_{ut}為類神經網路的輸出，代表語句 u 在時間點 t 的狀態 S 的機率。M 為合併的模型總數，α_m為各別模型混和權重，且$\alpha_m \geq 0$, $\sum_{\mathrm{m=1}}^{\mathrm{M}} \alpha_m = 1$。α相似於在(Dietterich, 2000)中的利用對角線矩陣(Diagonal matrices)對各別模型的線性合併(Linear ensemble)。合併後的音框事後機率(Frame posterior)會當成隱藏式馬可夫模型(Hidden Markov model, HMM)的聲學特徵 O 的對數可能性，並進行標準的解碼程序。

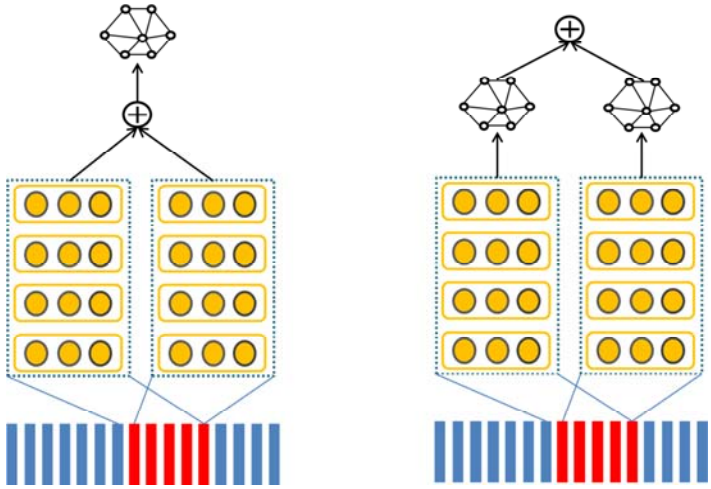


圖 1. 兩種層級的聲學分數合併。左方為音框層級，右方為假說層級。
[Figure 1. Frame-level vs. Hypothesis-level combination]

4.2 假說層級合併 (Hypothesis-level Combination)

假說層級合併則是利用 ASR 系統經過一般的解碼機制產生的詞圖，給予不同權重和損失函數進行合併。相較於音框層級合併，假說層級合併可允許非同步時間輸出，但因為需要合併各別 ASR 的輸出結果，因此較為費時。

$$h_u^* = \text{argmin}_{h'_u} \sum_{h_u} \mathcal{L}(h_u, h'_u) \sum_{m=1}^{M} \beta_m P(h_u|O_u, \lambda_m) \quad (9)$$

其中$\mathrm{h_u}$為各別 ASR 系統解碼時產生的詞序列。M為合併的模型總數。β_m為各別模型混和權重，且 $\beta_m \geq 0$,$\sum_{\mathrm{m=1}}^{\mathrm{M}} \beta_m = 1$。$L$為詞層級的損失函數，這裡使用編輯距離(Edit distance)。式(9)可理解為各別 ASR 產生詞圖的聯集，並透過最小化貝式決策風險對合併的詞圖解碼。音框層級和假說層級的示意圖可參考圖 1。

5. 實驗 (EXPERIMENTS)

5.1 實驗設定 (Experimental Setup)

實驗使用 Kaldi (Povey *et al.*, 2011)語音識別工具包。語料庫為 AMI (Augmented Multi-party Interaction) (McCowan *et al.*, 2005)。AMI 語料庫是來自歐盟發起的會議瀏覽(Meeting browser)計畫，其中包含情境會議 (Scenario meetings)和非情境的會議，情境會議是指明確的會議目標、會議間彼此有關連，如其中一個會議的主題為討論電視遙控器的設計；另一方面的非情境會議 (Non-scenario meetings)則反之，較沒有明確的主題，主要為英國愛丁堡大學、瑞士 Idiap 研究中心、荷蘭 TNO 人為因素研究所的學生或研究者組成討論的小型會議，如線性代數、微積分等。AMI 的語料庫也包含了影像、文字、語音，影像紀錄的是會議視角、投影機畫面和白板書寫記錄；文字有語音轉寫、對話特性，

表 2. AMI 會議之訓練、發展與測試集
[Table 2. The table shows that some basic statistics of the AMI corpus]

語料單位	訓練集	發展集	測試集 1	測試集 2	總計
小時數	70.09	7.81	8.71	8.97	95.79
語句數	97,222	10,882	13,059	12,612	133,775

可用於摘要、情緒與對話；最後是語音的部分，可分為耳掛式近距離麥克風、固定式遠距離麥克風。本實驗只用到了語音語料。表 2 為 AMI 的基本統計數據，由於原先 AMI 的訓練中並沒有用到發展集，因此實際訓練集為訓練集加發展集。我們用詞錯誤率(Word error rate, WER)和詞修復率(WER recovery rate, WRR)作為評估。WRR 如下：

$$\text{WRR} = \frac{BaslineWER - SemisupWER}{BaselineWER - OracleWER} \tag{10}$$

5.1.1 半監督式實驗流程與設定 (Semi-supervised Setup)

本實驗將 AMI 原先的訓練集切割成 16 小時的監督(轉寫)語料和 62 小時的非監督(未轉寫)語料，發展集和測試集。整體實驗的訓練為兩階段，第一階段為利用 16 小時的監督語料訓練種子模型，以及再使用 62 小時的非監督語料提升模型效能。整體實驗的詳細架構可參考圖 2。LF-MMI 的設定與(Povey *et al.*, 2016)一樣，特徵是 40 維 MFCC 和 100 維的 i-vector，類神經網路是使用時間延遲網路(Time-delay neural network, TDNN) (Peddinti, Povey & Khudanpur, 2015)。實驗分為訓練準則的有效性，以及後處理的模型合併。這裡需要注意的是合併時的權重皆為模型數量的倒數(e.g. M 個模型，權重為 1/M) 。

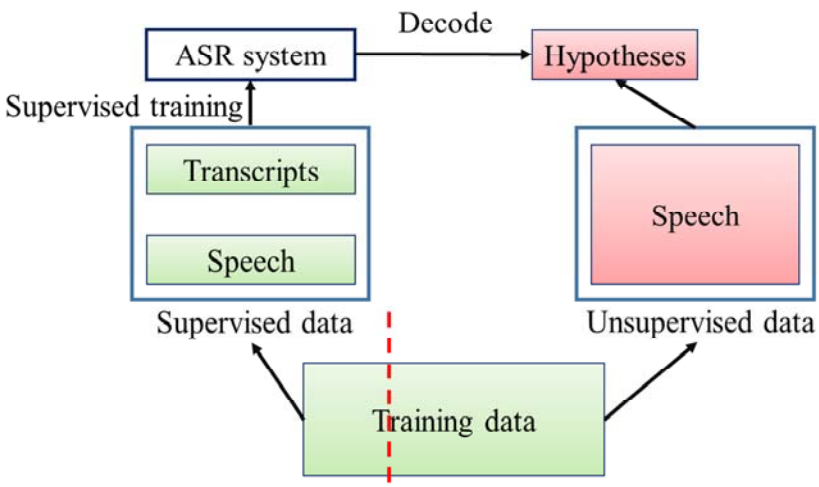


圖 2. 整體實驗架構
[Figure 2. A Flow chart of the experimental design.]

5.2 實驗結果與分析 (Results and Discussion)

5.2.1 加入NCE與詞圖的影響 (NCE and Lattice for supervision)

表 3 中呈現的是是否加入 NCE 權重和詞圖於訓練中的結果。第二欄中的 lm-scale 為第二階段的語音模型重新評分的縮放常數、Beam 為搜尋詞圖保留的種子個數、Tol 為在訓練時用到的詞圖的允許音框位移(Frame shift)，1 代表 30ms，這裡基於經驗上的設置，並無特地調動。第一欄由上到下，Baseline 是只用 16 小時訓練的聲學模型；No weight (NW)則是直接加入 62 小時未轉寫的語料；Best Phone Path(BPP)基於 NW 之上，加入 NCE 的權重的最佳辨識結果；Lattice for supervision (LS)則是基於 BPP 並加入詞圖。為了計算方便，LS 將原先的詞圖切成 1.5 秒的塊(Chunks)，再用向前向後算法計算；Oracle 則是將 78 小時的語料直接加入訓練。Dev 和 Eval 分別是測試集 1 和測試集 2。從實驗中的結果可以看出，直接進行傳統二階段訓練時，便可稍微提升 WRR 為 24%，這要歸功於良好的 LF-MMI 種子模型，只用 16 小時的訓練語料便達到尚可的辨識率，但 WRR 提升不夠可能是未轉寫語料相對使用太少導致修復率不佳。使用 NCE 可進一步地提升 WRR 至 33%，這也證明了加入音框層級的條件熵能有效輔助半監督式訓練，因此後續實驗探討皆會以 NCE 為主；最後則是加入整個詞圖後可將 WRR 提升至 45%，可看出多保留幾個搜尋的可能性後，增加的計算空間能輔助模型的訓練，進一步提升辨識結果。

表 3. 加入 NCE 與詞圖的影響
[Table 3. Negative conditional entropy and lattice for supervision]

Supervision	lm-scale	Beam	Tol	Dev	Eval	WRR
Baseline	-	-	-	27.2	27.8	-
NW	0	0	1	26.2	26.8	24%
BPP	0	0	1	26.0	26.2	33%
LS	0.5	4	1	**25.5**	**25.7**	**45%**
Oracle	-	-	-	23.5	23.1	-

5.2.2 模型合併於半監督式訓練 (Model combination in semi-supervised training)

這裡探討不同的訓練機制下，模型合併的成效。可分為音框層級的合併及假說層級的合併。雖然合併的增益主要來自模型的各別準確與多樣性，但這裡主要是探討整體學習合併的成效，因此採用簡單地調整訓練方式達成多樣性。實驗記錄於表 5，從第一列由左至右分別為不同方式訓練下的 LF-MMI 聲學模型，比較紀錄於表 4；FCOMB 和 HCOMB 則分別是音框層級和假說層級的合併；Test 則是 Dev 和 Eval 之 WER 的相加取平均。從上述實驗中可看出加入 Proportional shrink 和 L2-regularization 可比原先的訓練再降低詞錯誤率 1%，而調整初始化也會些微地影響 WER，也再一次證實了 LF-MMI 容易過度擬合的問題。另一方面，雖然在 WER 上看到成效，不過在 WRR 上則沒有顯著差異。這可

以看出這兩種方法的泛用性，不會受到不同的訓練準則影響進步成效。另一方面，從合併的觀點來看，音框層級合併在各個階段，比起單一系統的準度皆能提升 0.5 至 1.5 的 WER，證明了合併模型的技術應用於半監督式環境的有效性。另一方面，假說層級的合併效果更勝於音框層級的合併，這樣的結果歸功於假說合併是做在各個 ASR 的詞圖上，而音框合併則是類神經網路的輸出。比起音框合併，假說層級的合併更接近於辨識目標的詞，因此效果較好。但另一方面，音框合併雖然在效果上略輸假說合併，但由於可直接在音框階段就合併，不需要各別 ASR 產生詞圖，因此有更好的即時性。

表 4. 不同網路的設定差異
[Table 4. The table shows that different training criteria in the experiment. We combine four TDNN model generated by different random seed at both frame level and hypothesis level.]

	TDNN0	TDNN1	TDNN2	TDNN3
設定差異 (與 TDNN0)	基於表 2 的設定	+Proportional shrink	+L2-regularization	與 TDNN1 初始化不同

表 5. 音框和假說層級的聲學分數結合
[Table 5. Results on model combinations including frame-level and hypothesis-level combination.]

	TDNN1	TDNN1	TDNN2	TDNN3	FCOMB	HCOMB
	Test	Test	Test	Test	Test	Test
Baseline	27.5	26.7	26.5	26.5	25.6	**25.5**
BPP	26.1	25.2	25.5	25.1	24.5	**24.4**
LS	25.6	25.1	25.5	24.9	24.4	**24.2**
Oracle	23.3	22.5	22.8	22.5	21.5	**21.3**

5.2.3 不同半監督式準則的模型合併 (Model Combination and Semi-supervised Training)

表 6 中的第一列分別為音框層級合併與假說層級合併。第二欄由上至下為 NW、BPP 和 LS。而這裡合併的模型是表二的訓練結果。合併的方式為種子模型與 NW；種子模型、NW 和 BPP；種子模型、NW、BPP 和 LS。從實驗的結果中可看出基於音框層級與假說層級的合併十分有效，且假說層級的合併在大部分的情況下，仍勝過音框層級的合併，少部分是兩者持平。這裡我們可分析 NW、BPP 和 LS 這三種方法彼此的互補，最好的 WRR 為 60.8%。這些不同準則的合併雖有助於 WER 與 WRR 的提升，但可從實驗結果中觀察到 WER 進步的幅度約為 0.5，沒有比各別訓練多個模型並在同個半監督準則合併(表 5)來得更好。因此我們可得知，半監督準則在各別準確與多樣性上，相較使用 Proportional shrink 和 L2-regularization 來得小。儘管如此，不論是那種方式，在半監督式

環境下，我們皆可透過簡單地改變超參數，再以音框層級與假說層級的合併達到更好的辨識結果。

表 6. 不同半監督準則的模型合併
[Table 6. Results on model combination in conjunction with different semi-supervised criteria]

		F-COMB			H-COMB		
		Dev	Eval	WRR	Dev	Eval	WRR
TDNN-0	+NW	25.9	26.1	35%	25.7	26.0	39%
	+BPP	25.4	25.5	48%	25.3	25.5	50%
	+LS	25.1	25.1	57%	**24.9**	**25.0**	**60%**

6. 結論 (CONCLUSION AND FUTURE WORK)

本論文探討兩種思路於半監督式 LF-MMI。其一，利用 NCE 權重與詞圖模擬未轉寫語料的不確定性；其二，探討不同層級的合併，較快的音框層級合併和較準的假說層級合併。實驗結果得知，在無需信心過濾器的語料挑選下，這兩種思路可直接應用於半監督式 LF-MMI，並能有效地降低 WER 與提升 WRR 且相輔相成，最終 WRR 為 60.8%。未來的研究方向會針對有效性與即時性兩個方向繼續研究。根據這次的實驗結果，我們得知可透過模擬不確定性或更改參數的模型合併提升準度，未來會繼續朝如何利用未轉寫語料與互補多樣性的合併繼續研究，如 1) 利用不同模型種類，以產生更好的互補性。另一方面，2) 轉寫語料與未轉寫語料的比例，要到多少才能達成最好的 WRR；再者，儘管這次透過模型合併得到了不錯的結果，但同時也付出相較於單一 ASR 系統更高昂的運算資源，即便是相較於假說合併較為輕量的音框合併也是如此。因此 3) 未來會加入模型壓縮(Model combination)的技術，期許有一天能夠以少量的轉寫語料便達到有效且即時的辨識結果。

參考文獻 (REFERENCES)

Bahl, L., Brown, P., de Souza, P., & Mercer, R. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proceedings of ICASSP 1986*. doi: 10.1109/ICASSP.1986.1169179

Chan, H. Y., & Woodland, P. (2004). Improving broadcast news transcription by lightly supervised discriminative training. In *Proceedings of ICASSP 2004*. doi: 10.1109/ICASSP.2004.1326091

Cui, X., Huang, J., & Chien, J.-T. (2011). Multi-view and multiobjective semi-supervised learning for large vocabulary continuous speech recognition. In *Proceedings of ICASSP 2011*. doi: 10.1109/ICASSP.2011.5947396

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, *20*(1), 30-42. doi: 10.1109/TASL.2011.2134090

Deng, L., & Platt, J. C. (2014). Ensemble deep learning for speech recognition. In *Proceedings of Interspeech 2014*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of International workshop on MCS 2000*, 1-15.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139-157. doi: 10.1023/A:1007607513941

Evermann, G., & Woodland, P. C. (2000). Posterior probability decoding, confidence estimation and system combination. In *Proceedings of STW 2000*.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In *Proceedings of ASRU 1997*, 347-352. doi: 10.1109/ASRU.1997.659110

Gibson, M., & Hain, T. (2006). Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. In *Proceedings of Interspeech 2006*.

Grandvalet, Y., & Bengio , Y. (2005). Semi-supervised learning by entropy minimization. In *Proceedings of NIPS 2005*, 529-536.

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*. doi: 10.1145/1143844.1143891

Grezl, F., & Karafiát, M. (2013). Semi-supervised bootstrapping approach for neural network feature extractor training. In *Proceedingd of ASRU 2013*. doi: 10.1109/ASRU.2013.6707775

Huang, J.-T., & Hasegawa-Johnson, M. (2010). Semi-supervised training of gaussian mixture models by conditional entropy minimization. In *Proceedings of INTERSPEECH 2010*, 1353-1356.

Juang, B.-H., Hou, W., & Lee, C.-H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, *5*(3), 257-265. doi: 10.1109/89.568732

Kaiser, J., Horvat, B., & Kacic, Z. (2000). A novel loss function for the overall risk criterion based discriminative training of HMM models. In *Proceedings of ICSLP 2000*, 887-890.

Lamel, L., Gauvain, J.-L., & Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech & Language*, *16*(1), 115-129. doi: 10.1006/csla.2001.0186

Liu, S.-H., Chu, F.-H., Lin, S.-H., & Chen, B. (2007). Investigating data selection for minimum phone error training of acoustic models. In *Proceedings of ICME 2007*. doi: 10.1109/ICME.2007.4284658

Manohar, V., Hadian, H., Povey, D., & Khudanpur, S. (2018). Semi-supervised training of acoustic models using lattice-free MMI. In *Proceedings of ICASSP 2018*. doi: 10.1109/ICASSP.2018.8462331

Manohar, V., Povey, D., & Khudanpur, S. (2015). Semi-supervised maximum mutual information training of deep neural network acoustic models. In *Proceedings of Interspeech 2015*.

Mathias, L., Yegnanarayanan, G., & Fritsch, J. (2005). Discriminative training of acoustic models applied to domains with unreliable transcripts. In *Proceedings of ICASSP 2005*. doi: 10.1109/ICASSP.2005.1415062

McCowan, I., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., …Wellner, P. (2005). The ami meeting corpus. In *Proceedings of ICMTBR 2005*.

Nadas, A. (1983). A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *31*(4), 814-817. doi: 10.1109/TASSP.1983.1164173

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of Interspeech 2015*, 3214-3218.

Povey, D., & Woodland, P. C. (2002). Minimum phone error and i-smoothing for improved discriminative training. In *Proceedings of ICASSP 2002*. doi: 10.1109/ICASSP.2002.5743665

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for feature and model space discriminative training. In *Proceedings of ICASSP 2008*. doi: 10.1109/ICASSP.2008.4518545

Povey, D., Ghoshal, A., Boulianne, G., Goel, N., Hannemann, M., Qian, Y., …Stemmer, G. (2011). The Kaldi speech recognition toolkit. In *Proceedings of ASRU 2011*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, 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

Pundak, G., & Sainath, T. N. (2016). Lower frame rate neural network acoustic models. In *Proceedings of Interspeech 2016*. doi: 10.21437/Interspeech.2016-275

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Proceedings of Interspeech 2011*.

Senior, A., Sak, H., Quitry, F. C., Sainath, T., & Rao, K. (2015). Acoustic modelling with CD CTC-SMBR LSTM RNNs. In *Proceedings of ASRU 2015*. doi: 10.1109/ASRU.2015.7404851

Thomas, S., Seltzer, M. L., Church, K., & Hermansky, H. (2013). Deep neural network features and semisupervised training for low resource speech recognition. In *Proceedings of ICASSP 2013*. doi: 10.1109/ICASSP.2013.6638959

Valtchev, V., Odell, J. J., Woodland, P. C., & Young, S. J. (1996). Lattice-based discriminative training for large vocabulary speech recognition. In *Proceedings of ICASSP 1996*. doi: 10.1109/ICASSP.1996.543193

Valtchev, V., Odell, J. J., Woodland, P. C., & Young, S. J. (1997). MMIE training of large vocabulary recognition systems. *Speech Communication*, *22*(4), 303-314. doi: 10.1016/S0167-6393(97)00029-0

Vesely, K., Hannemann, M., & Burget, L. (2013). Semi-supervised training of deep neural networks. In *Proceedings of ASRU 2013*. doi: 10.1109/ASRU.2013.6707741

Walker, S., Pedersen, M., Orife, I., & Flaks, J. (2017). Semi-supervised model training for unbounded conversational speech recognition. Retrieved from https://arxiv.org/abs/1705.09724

Woodland, P. C., & Povey, D. (2002). Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech & Language*, *16*(1), 25-47. doi: 10.1006/csla.2001.0182

Xu, H., Povey, D., Mangu, L., & Zhu, J. (2011). Minimum bayes risk decoding and system combination based on a recursion for edit distance. *Computer Speech and Language*, *25*(4), 802-828. doi: 10.1016/j.csl.2011.03.001

Yu, K., Gales, M., Wang, L., & Woodland, P. C. (2010). Unsupervised training and directed manual transcription for LVCSR. *Speech Communication*, *52*(7-8), 652-663. doi: 10.1016/j.specom.2010.02.014

Zavaliagkos, G., Siu, M., Colthurst, T., & Billa, J. (1998). Using untranscribed training data to improve performance. In *Proceedings of ICSLP 1998*.

Zhang, P., Liu, Y., & Hain, T. (2014). Semi-supervised DNN training in meeting recognition. In *Proceedings of SLT 2014*. doi: 10.1109/SLT.2014.7078564

Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 35-46

結合鑑別式訓練聲學模型之類神經網路架構及優化方法的改進

Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method

趙偉成*、張修瑞*、羅天宏*、陳柏琳*

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

摘要

本論文探討聲學模型上的改進對於大詞彙連續中文語音辨識的影響。在基礎聲學模型的訓練上，有別於以往語音辨識通常使用交互熵(Cross Entropy)作為深度類神經網路目標函數，我們使用 Lattice-free Maximum Mutual Information (LF-MMI)做為序列式鑑別訓練的目標函數。LF-MMI 使得能夠藉由圖形處理器(Graphical Processing Unit, GPU)上快速地進行前向後向運算，並且找出所有可能路徑的後驗機率，省去傳統鑑別式訓練前需要提前生成詞圖(Word Lattices)的步驟。針對這樣的訓練方式，類神經網路的部分通常使用所謂的時間延遲類神經網路(Time-Delay Neural Network, TDNN)做為聲學模型可達到不錯的辨識效果。因此，本篇論文將基於 TDNN 模型加深類神經網路層數，並藉由半正交低秩矩陣分解使得深層類神經網路訓練過程更加穩定。另一方面，為了增加模型的一般化能力(Generalization Ability)，我們使用來回針法(Backstitch)的優化算法。在中文廣播新聞的辨識任務顯示，上述兩種改進方法的結合能讓 TDNN-LF-MMI 的模型在字錯誤率(Character Error Rate, CER)有相當顯著的降低。

關鍵詞：中文大詞彙連續語音辨識、聲學模型、鑑別式訓練、矩陣分解、來回針法

* 國立臺灣師範大學資訊工程研究所
Institute of Linguistics, National Taiwan Normal University
E-mail: {60647028S, 60647061S, teinhonglo, berlin}@ntnu.edu.tw

Abstract

This paper sets out to investigate the effect of acoustic modeling on Mandarin large vocabulary continuous speech recognition (LVCSR). In order to obtain more discriminative baseline acoustic models, we adopt the recently proposed lattice-free maximum mutual information (LF-MMI) criterion as the objective for sequential training of component neural networks in replace of the conventional cross entropy criterion. LF-MMI brings the benefit of efficient forward-backward statistics accumulation on top of the graphical processing unit (GPU) for all hypothesized word sequences without the need of an explicit word lattice generation process. Paired with LF-MMI, the component neural networks of acoustic models implemented with the so-called time-delay neural network (TDNN) often lead to impressive performance. In view of the above, we explore an integration of two novel extensions of acoustic modeling. One is to conduct semi-orthogonal low-rank matrix factorization on the TDNN-based acoustic models with deeper network layers to increase their robustness. The other is to integrate the backstitch mechanism into the update process of acoustic models for promoting the level of generalization. Extensive experiments carried out on a Mandarin broadcast news transcription task reveal that the integration of these two novel extensions of acoustic modeling can yield considerably improvements over the baseline LF-MMI in terms of character error rate (CER) reduction.

Keywords: Mandarin Large Vocabulary Continuous Speech Recognition, Acoustic Model, Discriminative Training, Matrix Factorization, Backstitch

1. 緒論 (INTRODUCTION)

近幾年來，語音辨識技術已有了長足的進步。其中，隨著深度學習技術以及電腦運算能力的突破性發展，聲學模型化技術已從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)，轉變成以使用交互熵(Cross Entropy)作為損失函數的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)。DNN-HMM 將以往用 GMM 計算的生成機率透過 DNN 的輸出層所代表的事後機率來近似，輸入特徵使用當前幀還有相鄰的幀，輸出則和 GMM-HMM 常用的 Triphone 共享狀態相同，以得到更低的詞錯誤率(Word Error Rate, WER)或字錯誤率(Character Error Rate, CER)。另一方面，進一步透過鑑別式訓練估測的聲學模型在語音辨識的表現上往往比僅以交互熵做為深度類神經網路損失函數的訓練方式來的好。但由於傳統上進行鑑別式訓練需要使用先進行交互熵訓練的聲學模型來產生詞圖(Word Lattices)，才能再進行下一步聲學模型鑑別式訓練(Bahl, Brown, de Souza & Mercer, 1986) (Veselý, Ghoshal, Burget & Poveyet, 2013)。近年來為了減少時間及空間複雜度，有學者對於 Maximum

Mutual Information (MMI)訓練，提出了所謂的 Lattice-free 的方式，使產生詞圖的步驟能夠在 GPU 上完成(Povey *et al.*, 2016)，因而讓鑑別式訓練得以做到端對端的訓練方式(Hadian, Sameti, Povey & Khudanpur, 2018)，因而大幅縮減了聲學模型訓練所需時間。

傳統 DNN-HMM 模型用於語音辨識的缺點在於無法充分利用語音信號之時間依賴性；而如同在(Graves, Mohamed & Hinton, 2013)所提到，基於遞迴類神經(Recurrent Neural Network, RNN)能對於序列性資料能夠有好的建模效果的想法所發展的 RNN-HMM 聲學模型其辨識效果卻是不如 DNN-HMM 模型來的好，因此以長短期記憶(Long Short-Term Memory, LSTM)取代簡單 RNN 所形成的聲學模型(LSTM-HMM) (Sak, Senior & Beaufays, 2014)，解決了 RNN-HMM 梯度消失的問題，在語音辨識上能夠達到比 DNN-HMM 好的效果。但在實務上，這樣的聲學模型很難像 DNN-HMM 一樣平行化訓練(Pascanu, Mikolov & Bengio, 2013)，以致於模型訓練時間的增加。另一方面，也由於其模型架構較為複雜使得運算量較大，較不適合需即時反應的語音辨識任務；相對來說時間延遲類神經網路(Time-Delay Neural Network, TDNN) (Waibel, Hanazawa, Hinton, Shikano & Lang, 1989)可以包含歷史和未來輸出、對長時間依賴性的語音訊號建模，使 TDNN-HMM 與傳統 DNN-HMM 訓練效率也相仿，因此在使用 LF-MMI 進行鑑別式訓練時，聲學模型的類神經網路部分通常是使用 TDNN。

從經驗上來看，類神經網路的深度對模型的性能非常重要(Ba & Rich, 2014)，增加層數之後能有更加複雜的特徵擷取能力。對於 TDNN 而言，增加層數可以說是提取更長時間的特徵；我們希望加深 TDNN 的網路層數來達到更好的結果，但以往的實驗發現深度的網路常有退化問題，類神經網路的深度之增加準確率反而會下降。因此本篇論文將比較並結合當前先進的聲學模型訓練方法，例如(Povey *et al.*, 2018)對網路的矩陣分解訓練可以使網路訓練更穩定，以期達到更佳的語音辨識表現。另一方面，梯度下降是執行優化的最流行的算法之一，也是迄今為止優化類神經網絡的最常用方法。而常見的優化算法有隨機梯度下降法(Stochastic Gradient Descent, SGD)、RMSprop、Adam、Adagrad、Adadelta (Ruder, 2016)等演算法；其中，SGD 算法在語音辨識任務上最被廣為使用。而本論文則採用來回針法(Backstitch) (Wang *et al.*, 2017)做為模型優化的演算法；它是一種基於 SGD 上的改進，希望能夠藉由兩步驟的更新 Minibatch，以達到更好的效果。

總合以上所述，我們認為加入對網路的矩陣分解來可順利訓練更深層的類神經網路模型；同時，使用 Backstitch 亦可提升模型泛化性，最終能使辨識結果更加進步。因此，本論文將分別比較使用 TDNN-LF-MMI，TDNN-LF-MMI 加入半正交低秩矩陣分解，TDNN-LF-MMI 加入半正交低秩矩陣分解及來回針法優化算法的辨識效果，最終在 TDNN-LF-MMI 加入半正交低秩矩陣分解及來回針法優化算法達到較佳的中文廣播新聞語音辨識的 CER 表現。

2. 聲學模型 (ACOUSTIC-MOLEL)

2.1 基本聲學模型-時間延類神經網路 (Time-Delay Neural Network, TDNN)

TDNN 在 1989 年被提出(Waibel *et al.*, 1989)，最初用於音素辨識；基於 TDNN 所產生的模型架構，能適用於處理語音所擁有特徵向量序列之時間長度不一致的特性。TDNN 對每一個隱藏層的輸出在時間上進行擴展，即每個隱藏層收到的輸入會有前一層在不同時刻的輸出。語音在考慮上下文長時間相關性很重要，TDNN 的優點在可以比傳統 DNN 看更長的時間，而且速度不會比 DNN 在訓練和辨識(解碼)時來的慢。

2.2 半正交低秩矩陣分解 (Semi-Orthogonal Low-Rank Matrix Factorization)

減少類神經網路參數的方法之一是透過 SVD 分解已經估測好的權重矩陣；近期有學者(Povey *et al.*, 2018)提出基於一個隨機的初始參數，用同樣的分解架構開始訓練類神經網路聲學模型，但是要讓其中一個分解的矩陣保持正交，避免有不穩定的問題。

在實作上，我們可以在進行若干次 SGD 後強迫參數矩陣變成半正交的更新，假設M是參數矩陣，定義$\mathrm{P} \equiv \mathrm{MM}^{\mathrm{T}}$目標是要讓 P 變成單位矩陣，學習率(Learning Rate)決定了權值更新速率的快慢，愈大的學習率會更快達到半正交的結果，但是設置太大會變得很不穩定，在接近半正交矩陣的時候 0.125 的設置是最好的，數學上可以達到平方收斂。令X是每次M更新的值，所以我們做一次更新$\mathrm{M} \leftarrow \mathrm{M} + \mathrm{X}$，我們希望$\mathrm{tr}(\mathrm{MX}^{\mathrm{T}}) = 0$以達到正交效果，下式為更新公式:

$$\mathbf{M} \leftarrow \mathbf{M} - \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\alpha}^{\mathbf{2}}}(\mathbf{MM}^{\mathbf{T}} - \boldsymbol{\alpha}^{\mathbf{2}}\mathbf{I})\mathbf{M} \tag{1}$$

α是一個縮放的參數，I是單位矩陣不考慮常數項，我們要使$\mathrm{tr}\left(\mathrm{MM}^{\mathrm{T}}(\mathrm{P} - \alpha^2\mathrm{I})\right) = 0$，因為$\mathrm{MM}^{\mathrm{T}} = \mathrm{P}$，所以$\mathrm{tr}(\mathrm{P}^2 - \alpha^2\mathrm{P}) = 0$移項之後$\alpha = \sqrt{\frac{tr(\mathrm{P}^2)}{tr(P)}}$，因為 P 是對稱矩陣所以$\mathrm{P}^2 = \mathrm{PP}^{\mathrm{T}}$，為了計算上較快會使用$\alpha = \sqrt{\frac{tr(\mathrm{PP}^{\mathrm{T}})}{tr(\mathrm{P})}}$。圖 1 是 TDNN+NF(Networks Factorized)內部架構，1536 維的隱藏層經矩陣分解後變成 1536*160*1536，SMAT 是要做正交限制的矩陣，後面再接上線性整流函數(ReLU)和批次標準化(Batch Normalization)。

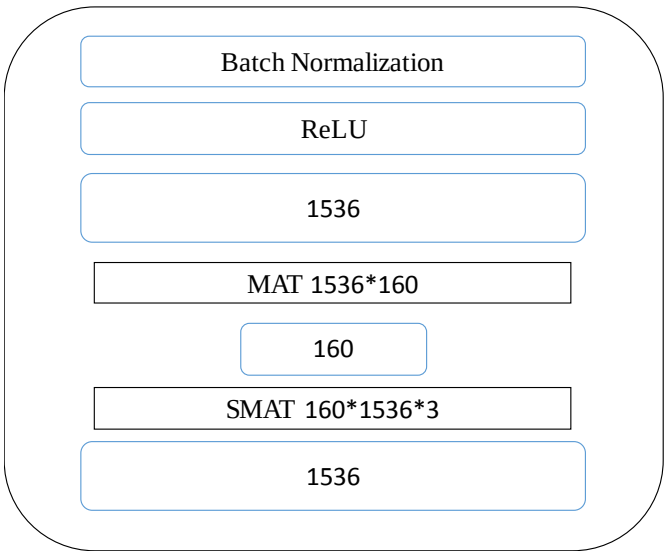


圖 1. 矩陣分解
[Figure 1. Matrix Factorization]

2.3 來回針法 (Backstitch)

Backstitch 是修改 SGD 增進對沒有看過資料的效能(Wang *et al*., 2017)；這個方法分成兩個步驟，先用一個較小的負學習率，再跟著一個較大的學習率，和對抗訓練的概念很像，可以消除有限數據集的偏見，用同樣的 Minibatch 但重新計算梯度，傳統的 SGD 單一的迭代如下式:

$$\boldsymbol{\theta_{n+1} \leftarrow \theta_n - \nu g(x_n, \theta_n)} \tag{2}$$

θ 是更新的參數，x_n 是第 n 個迭代的樣本 g(x_n, θ _n)是 f(x, θ)關於 θ 的導函數

$$\boldsymbol{\theta'_{n+1} \leftarrow \theta_n + \alpha \nu g(x_n, \theta_n)} \tag{3}$$

$$\boldsymbol{\theta_{n+1} \leftarrow \theta'_n - (1+\alpha) \nu g(x_n, \theta'_{n+1})} \tag{4}$$

其中，式(3)為 Backstitch 第一步驟退回更新，而式(4)為 Backstitch 第二步驟前進更新；α 這個常數決定要做多大步伐的更新，我們可以調整要幾個 Minibatch 做一次這種更新，根據原始論文比較有效率的設定是 α =1 和 n=4。

3. 模型架構和訓練方法 (METHODS)

3.1 聲學模型之類神經網路架構(Structure)

每層 TDNN 使用 1,536 維，加上兩層分解過後的矩陣 160 維，加上 ReLU 和 Batch

Normalization 合起來稱為 TDNNF 層，層數可比以前的 TDNN 網路(9 層以內)都還深。如圖 2 所示，最下層為隨著時間輸入之特徵，最上層為多任務的輸出，LF-MMI 的目標函數(Povey *et al.*, 2016)對應由決策樹定義的 Senone 函數，Cross Entropy Regularization 為輔助正規化訓練輸出，中間的 TDNNF 層有捷徑連結(Skip Connections) 。

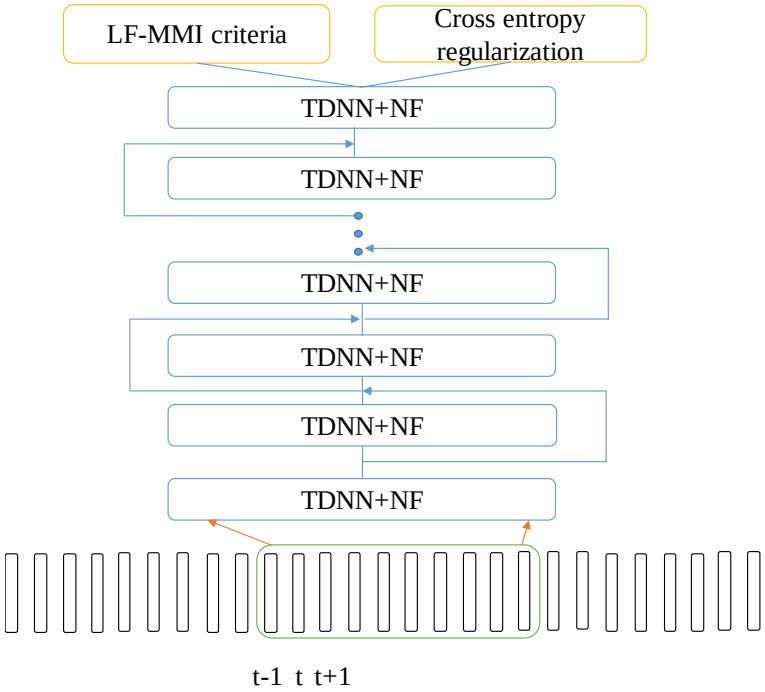


圖 2. 聲學模型中基於 TDNN 之類神經網路架構
[Figure 2. TDNN artitecture]

3.2 LF-MMI

近年來，透過可視為鑑別式模型(Discriminant Model)的類神經網路能夠有效提升系統效能。訓練開始之前先要有所有單詞序列組合，透過交叉熵訓練一個模型，配合語言模型生成一個詞圖，詞圖要有正確結果的路徑和足夠靠近的其他路徑，鑑別式模型用目標是要提高走正確路徑之機率，降低走相似路徑之機率。

MMI（Maximum Mutual Information）準則是要加大正確路徑和其他路徑的機率差。另一方面，LF-MMI 透過在類神經網路輸出計算所有可能的序列，根據這些序列計算 MMI 的梯度來訓練，因為 LF-MMI 在訓練中計算所有路徑的後驗機率(Posterior Probability)，所以不用事先生成訓練語句之詞圖。由於 TDNN 相鄰節點的變化通常不大，而且訊息重複機率很高，所以可以跳過一些幀的計算。LFMMI 的實驗設定在(Povey *et al.*, 2016)是降低幀率和將傳統 3-state 的 HMM 拓撲改為 2-state 的 HMM，端對端的鑑別式訓練也借鑒 CTC 上特殊的空白標籤(Graves, Fernández, Gomez & Schmidhuber, 2006)，在少量語料下

也達到很好的效果，在辨識率和解碼速度都有很大提升。

3.3 結合半正交限制和來回針法訓練 (Combine Training)

我們嘗試兩種訓練參數的方法用在聲學模型的效果，在 Natural Gradient (NG) (Povey, Zhang & Khudanpur, 2014)上面做改變，做 SGD 更新時每若干次在 backstitch 第一步驟後做半正交限制的更新，第二步驟維持原本的做法，在退回和前進步驟間做正交限制。

3.4 Dropout

在機器學習中，模型的參數太多會發生過度擬合現象，在每個批次訓練中忽略一些特徵檢測器，在 TDNN 中是橫跨時間的，(Povey *et al*., 2018)不是用二值的零一丟棄遮罩，而是用一個連續型均勻分布[1-2α, 1+2α]。我們使用一個丟棄排程表，在訓練一開始設定α=0.2，訓練到一半提升到 0.5，最後又下降到 0，這個設定在普通未分解的 TDNN 看起來沒有效果，但在分解後的網路架構中有明顯改善。

3.5 捷徑連結 (Skip Connections)

根據影像辨識裡 VGG (Simonyan & Zisserman, 2014)的發展，經驗上增加層數可以增加準確度，但是會增加訓練上的難度。ResNet (He, Zhang, Ren & Sun, 2016)在其上進行修改在網路上增加捷徑，可以防止類神經網路太深而無法訓練，我們在 TDNNF 裡也做上類似的機制，每層加上輸入更前面一層的三分之二和前一層相加當成新的輸入。

4. 實驗 (EXPERIMENTS)

4.1 實驗設定 (Experiment Setups)

表 1. 中文廣播新聞的實驗語料
[Table 1. MATBN]

	長度(小時)	句數
訓練集	114.7	38,556
發展集	3.7	2,001
測試集一	3.6	1,957
測試集二	1.4	307

本論文實驗語料來自公視新聞(Wang, Chen, Kuo & Cheng, 2005) (Mandarin Across Taiwan-Broadcast News, MATBN)。公視新聞語料是 2001 年至 2003 年間由中研院資訊所口語小組與公共電視台合作錄製，共計 197 個小時，取其中部分用於實驗，表 1 為實際用於實驗的語料長度和句數，包含內場新聞與外場新聞，其中內場新聞為新聞主播語料，外場包含採訪記者語受訪者語料。背景語言模型使用 5-gram 語言模型，訓練語料來自

2001 年至 2002 年中央通訊社(Central News Agency, CNA)的文字新聞語料，內含一億五千萬個中文字，經斷詞後約有八千萬個詞(本論文使詞典約七萬二千詞)。本論文是使用 SRI Language Modeling Toolkit(SRILM) (Stolcke, 2002) 來訓練語言模型。語言模型的訓練集由 2001 及 2002 年的新聞語料所篩選出來的。測試集一包含五場錄音在 2003/01/28，2003/01/29，2003/02/11，2003/03/07 和 2003/04/03，測試集二為只選擇了採訪記者語料並濾掉了含有語助詞之語句。另一方面，聲學模型都是在開源的語音辨識工具 Kaldi (Povey *et al*., 2011)上訓練，首先在語音語料庫上訓練具語者調適性(Speaker-adaptive)高斯混和隱馬可夫模型(GMM-HMM)，並利用該模型來獲得所有訓練語句的詞圖來準備後續聲學模型之類神經網路訓練。然後使用 TDNN-LF-MMI 準則來訓練出一類神經網路；其中，最佳化的部分則使用 NG 和 Backstich，遵循(Povey *et al*., 2016)中描述的方法來創建聲學模型，即對於 5,600 個依賴於上下文的語音中的每一個具有一個狀態的 HMM 拓撲，以原始幀速率的三分之一操作。有用變速擾動的資料擴充，特徵使用 40 維的 MFCC 和 3 維的聲調特徵加上 100 維的 i-vectors 做調適。

4.2 實驗結果 (Experiment Results)

表 2 比較了基本的 TDNN 和其他論文的 Attention 模型(Povey, Hadian, Ghahremani, Li & Khudanpur, 2018)包含 TDNN 和 3 層 LSTM 訓練而成，從實驗可以看出交互熵(CE)會遜於 LF-MMI 訓練結果，所以之後的改進模型皆使用 LF-MMI 訓練。基礎 TDNN 模型有 9 層隱藏層，每一個隱藏層有 625 維，前後文音窗各 15 幀，TDNN+NF 模型使用 1536 維的隱藏層，矩陣分解瓶頸 160 維，前後文音窗各 33 幀。TDNN+NF 在隱藏層維度變大時效果較好(Waibel *et al*., 1989)，而基本的 TDNN 沒有這種變化。

表 2. 基礎語音辨識實驗
[Table 2. Baseline experiment results]

	WER	CER	Parameters	RTF
Attention(LF-MMI)	26.76	18.96	50M	1.66
TDNN(LF-MMI)	**26.22**	**18.34**	**15M**	**0.42**
TDNN(CE)	27.84	19.17	**15M**	0.47

改進模型實驗結果如表 3，分別實做了 TDNN 和 TDNN+NF 及各自加上 ackstitch 的實驗，分解後的網路在參數上只多了一些，解碼速度相當但是字錯誤率降低很多，在不同深度的 TDNN+NF 比較實驗中 15 層表現最好。表 4 可見在不同難度測試集解碼後最終模型的字錯誤率都有顯著改善。

表 3. 改進模型在測試集一的實驗結果
[Table 3. Experiment results for test sets]

	WER	CER	Parameters	RTF
TDNN+Backstitch(9 層)	25.14	17.45	15M	0.42
TDNN+NF(15 層)	23.98	16.27	18M	0.47
TDNN+NF+Backstitch(10 層)	23.30	15.64	**13M**	**0.37**
TDNN+NF+Backstitch(15 層)	**22.56**	**15.15**	18M	0.47
TDNN+NF+Backstitch(20 層)	22.75	15.26	23M	0.58

表 4. 改進模型在測試集二與發展集的實驗結果
[Table 4. Experiment results for other test sets]

	測試集二 (CER)	發展集 (CER)
TDNN	5.05	33.39
TDNN+Backstitch	4.88	33.15
TDNN+NF	3.69	23.56
TDNN+NF+Backstitch	**3.67**	**22.73**

RTF(Real Time Factor)是一個常用於度量自動語音辨識系統解碼速度的值，如果處理一段長度為 a 的音訊信號需要花費時間 b，則 RTF 為 b/a，圖 3 是不同模型解碼速率比較，可以看出使用 LSTM 會提升大量時間，而其他模型隨著參數提昇會稍微增加時間。

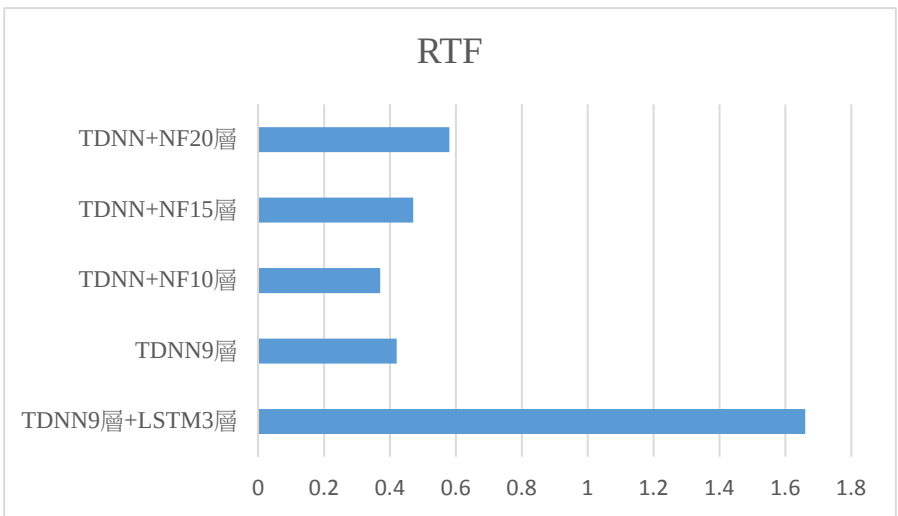


圖 3. 不同模型解碼速率比較
[Figure 3. Decoding speed comparison]

從訓練過程的準確率來看，有沒有使用 Backstitch 看不出太大差異，但是反應在字的錯誤率上有進步。有做矩陣分解的模型因為有正交限制的更新，在迭代 160 次後準確率會超越基本的 TDNN。關於 Backstitch 方面，圖 4 中虛線為訓練集，實線為驗證集，藍色 α =1 相較紅色 α =0.3 需要多兩倍的迭代才會收斂到相同準確率，可能是第二步驟沒有做正交化延遲了收斂的速度。

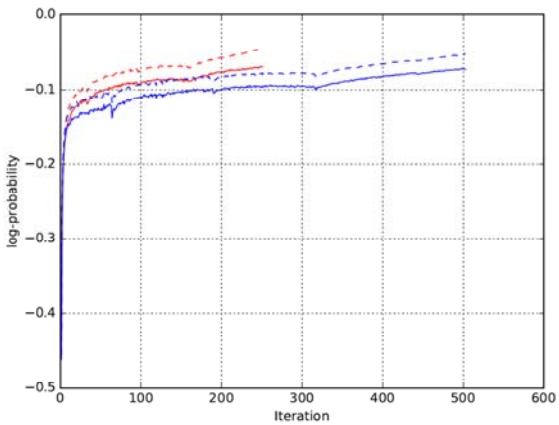


圖 4. 不同 Backstitch 的更新步伐之比較
[Figure 4. Different Backstitch steps comparison]

5. 結論 (CONCLUSION AND FUTURE WORK)

本論文探討聲學模型中的類神經網路權重參數更新並優化對於語音轉成文字錯誤率的影響。且應用了幾個其他改進，像是捷徑連接和隨著時間改變網路節點丟失的方法，從實驗結果發現，加上矩陣分解的時延遲網路用結合半正交限制和交叉針法的訓練效果最佳，在不同的測試集的字錯誤率上都有顯著的進步，訓練時間和解碼速度也不比基本的模型差，未來希望繼續探究不同結合方式在自動語音辨識的表現，並且更詳細與廣泛地探討各種聲學模型之優缺點。

參考文獻 (REFERENCES)

Ba, J., & Rich, C. (2014). Do deep nets really need to be deep? In *Proceedings of NIPS 2014*, *2*, 2654-2662.

Bahl, L., Brown, P., de Souza, P., & Mercer, R. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proceedings of ICASSP 1986*. doi: 10.1109/ICASSP.1986.1169179

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

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML '06*, 369-376. doi: 10.1145/1143844.1143891

Graves, A., Mohamed, A.-r., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of ICASSP 2013*. doi: 10.1109/ICASSP.2013.6638947

Hadian, H., Sameti, H., Povey, D., & Khudanpur, S. (2018). End-to-end speech recognition using lattice-free MMI. In *Proceedings of Interspeech 2018*. doi: 10.21437/Interspeech.2018-1423

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. (2016). In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2016)*. doi: 10.1109/CVPR.2016.90

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

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of ICML 2013*, *28*, 1310-1318.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohamadi, M., …Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proceedings of Interspeech 2018*. doi: 10.21437/Interspeech.2018-1417

Povey, D., Ghoshal, A., Boulianne, G., Goel, N., Hannemann, M., Qian, Y., …Stemmer, G. (2011). The Kaldi speech recognition toolkit. In *Proceedings of ASRU 2011*.

Povey, D., Hadian, H., Ghahremani, P., Li, K., & Khudanpur, S. (2018) A time-restricted self-attention layer for ASR. In *Proceedings of ICASSP 2018*. doi: 10.1109/ICASSP.2018.8462497

Povey, D., Peddinti, V., Galvez, D., Ghahremani, 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

Povey, D., Zhang, X., & Khudanpur, S. (2014). Parallel training of DNNs with natural gradient and parameter averaging. Retrieved from https://arxiv.org/abs/1410.7455

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

Ruder, S. (2016). An overview of gradient descent optimization algorithms. Retrieved from https://arxiv.org/abs/1609.04747

Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. Retrieved from arXiv:1402.1128

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Retrieved from https://arxiv.org/abs/1409.1556

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, 901-904.

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

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(3), 328-339. doi: 10.1109/29.21701

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese broadcast news corpus. *International journal of computational linguistics & Chinese language processing, Special Issue on Annotated Speech Corpora*, *10*(2), 219-236.

Wang, Y., Peddinti, V., Xu, H., Zhang, X., Povey, D., & Khudanpur, S. (2017). Backstitch: counteracting finite-sample bias via negative steps. In *Proceedings of Interspeech 2017*. doi: 10.21437/Interspeech.2017-1323

Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 47-66

Supporting Evidence Retrieval for Answering Yes/No Questions

Meng-Tse Wu*, Yi-Chung Lin* and Keh-Yih Su*

Abstract

This paper proposes a new n-gram matching approach for retrieving the supporting evidence, which is a question related text passage in the given document, for answering Yes/No questions. It locates the desired passage according to the question text with an efficient and simple n-gram matching algorithm. In comparison with those previous approaches, this model is more efficient and easy to implement. The proposed approach was tested on a task of answering Yes/No questions of Taiwan elementary school Social Studies lessons. Experimental results showed that the performance of our proposed approach is 5% higher than the well-known Apache Lucene search engine.

Keywords: Supporting Evidence Retrieval, Q&A for Yes/No Questions.

1. Introduction

Supporting evidence retrieval is a key step in the question-answering task. It locates the related text passage from the given documents according to the question content so that the system can efficiently answer the question only based on the retrieved passage. The goal of supporting evidence retrieval is to merely keep necessary information (but filter out the irrelevant content as much as possible) to reduce the associated inference time.

Previous supporting evidence retrieval approaches can be classified into three categories: (1) Term matching approaches (Chen, Fisch, Weston & Bordes, 2017), (2) Syntactic/Semantic scoring approaches (Murdock, Fan, Lally, Shima & Boguraev, 2012; Jansen, Sharp, Surdeanu & Clark, 2017), and (3) Translation model based approaches (Berger, Caruana, Cohn, Freitag & Mittal, 2000; Jeon, Croft & Lee, 2005; Xue, Jeon & Croft, 2008; Zhou, Cai, Zhao & Liu, 2011). Term matching approaches, such as Lucene search[1], used the vector space model and some language models adopted in *Information Retrieval* (Manning, Raghavan & Schütze,

* Institute of Information Science, Academia Sinica
E-mail: {moju, lyc, kysu}@iis.sinica.edu.tw

[1] http://lucene.apache.org/

2008). On the other hand, syntactic/semantic scoring approaches (Murdock *et al*., 2012; Jansen *et al*., 2017) retrieved the supporting evidence by conducting the syntactic/semantic analysis of each document sentence. They detected certain terms or structures in the question and then weighted the candidates differently by the appearance of those terms or structures. Finally, approaches that utilize a translation model were widely adopted in the *Community QA* systems (Berger *et al*., 2000; Jeon *et al*., 2005; Xue *et al*., 2008; Zhou *et al*., 2011). They used phrase-based or word-based translation models to find the similar historical questions from the new queried question. In the task of supporting evidence retrieval, we could let the question play the role of new queried question and the supporting evidence play the role of historical questions, and then adopt the translation model to find the supporting evidence.

Term matching approaches are widely adopted in the search engine due to its efficiency. However, they do not consider the local context of each term, not even mentioning the associated syntactic/semantic information. Therefore, they usually result in low accuracy. On the other hand, syntactic/semantic scoring approaches utilize syntactic/semantic meaning of each document sentence. They can understand the questions more in the syntactic/semantic level. However, those approaches are not only time consuming but also task orientated. Finally, translation model based approaches are widely adopted in the Community QA systems. However, they need large training data to train the translation models, and are thus not suitable for the tasks with only small amount of training data.

To overcome the problems mentioned above, we aimed at the approach that is efficient, general and accurate enough. Therefore, the approach of term (most of them are unigrams) matching is still adopted in this paper for computation efficiency and generalization. However, to further consider the phrase and local context, it is extended into n-gram for considering the local dependency. It thus avoids the drawbacks of previous approaches.

Given a question, our goal is to find a related passage, from the given corpus, that contains minimum but sufficient information to answer the question. In other words, good supporting evidence should include sufficient related information and less irrelevant and redundant information for the given question. On the other hand, supporting evidence can be extracted in different granularity. For instance, they are specified as top 5 articles in (Chen *et al*., 2017). The smaller the granularity is, the harder the approach is to find the appropriate supporting evidence (since we need to locate it more accurately). In our task, we define the supporting evidence as a text passage with consecutive sentences in the same paragraph, which will be explained in Section 4.3. We propose two scoring functions for finding the supporting evidence: QE-BLUE and modified F-measure. QE-BLUE is converted from the CR-BLEU score (Papineni, Roukos, Ward & Zhu, 2002) which only considers n-gram precision and is used in evaluating the performance of a machine translation system. In contrast, the modified F-measure takes both recall and precision of n-grams into consideration.

Therefore, the modified F-measure is able to evaluate the portion of the matched terms in the question. In comparison with those term matching approaches, the proposed method provided better performance. On the other hand, in comparison with those semantic scoring approaches, the proposed method is more efficient, easy to implement and task independent. In summary, we make the following contributions in this paper:

- We studied the desired characteristics of extracted supporting evidence.
- We proposed a novel scoring function for retrieving the supporting evidence by jointly considering precision and recall of n-grams.
- We adopted and tested several techniques for improving the supporting evidence retrieval.
- We conducted the experiments to show the superiority of the proposed approach.

The remainder of this paper is organized as follows. Section 2 illustrates the desired characteristics that an effective supporting evidence retrieval algorithm should possess. The proposed approach is introduced in Section 3. Section 4 shows the experimental result. The error analysis of the proposed approach is then given in Section 5. The related work is introduced in Section 6. Finally, Section 7 concludes this paper.

2. Desired Characteristics

Question:我們|應該|完全|聽從|父母|的|建議|，|選擇|加入|學校|的|團隊|。

"**We** should fully follow the **advice** of **parents** for **choosing** which school **group to join**."

Evidence:我們|可以|依照|自己|的|興趣|，|參考|老師|和|父母|的|建議|，|選擇|加入|不同|的|團隊|學習|。

"**We** can consider our own interest and refer to the **advice** from the teachers and **parents** for **choosing** which learning **group to join**."

Figure 1. A question and its corresponding supporting evidence

From the question and its supporting evidence shown in Figure 1, we can see that they share many words (which are marked in bold and underlined). This is because the questions usually use the same words or sentences to describe the same thing.

Let s_i stand for the i-th matched word, w_j stand for the j-th unmatched word, w_j^* stand for the j-th string which purely consists of an arbitrary number of unmatched words, and $|w_j^*|$ denote the number of words contained in w_j^*. The desired characteristics of an effective supporting evidence retrieval algorithm are listed as follows.

Characteristic-1: Prefer more matching occurrences

Candidate1: $\underline{\mathbf{s_1}}\ w_1^*$

Candidate2: $\underline{\mathbf{s_1}}\ w_2^*\ \underline{\mathbf{s_1}}\ w_3^*$

In the above pattern, we prefer Candidate-2 as the supporting evidence since the same matched term appears more times. Consider the following Example-1:

Example 1

Question: 安平古堡|是|位於|台灣|南部|的|古蹟|。

"**Fort Zeelandia** is a monument located in south Taiwan."

Candidate-1: 安平古堡|位於|台灣|南部|，

"**Fort Zeelandia** is located in south Taiwan."

Candidate-2: 安平古堡|位於|台灣|南部|，|*安平古堡*|*有約*|*400*|*年*|*歷史*|。

"**Fort Zeelandia** is located in south Taiwan. ***Fort Zeelandia*** *has a history of about 400 years.*"

This preference is illustrated with the above Example-1. We prefer Candidate-2 here since it additionally mentions that 安平古堡 ("Fort Zeelandia") has a long history which entails that it is a monument. As a result, we prefer more occurrences of a matching term because it may contain more information we need.

Characteristic-2: Prefer less unmatched terms

Candidate1: $\underline{\mathbf{s_1}}\ w_1^*$

Candidate2: $\underline{\mathbf{s_1}}\ w_2^*$

Suppose $|w_1^*|$ is larger than $|w_2^*|$, then we prefer Candidate-2 in this case because it contains less number of unmatched terms whic are assumed to be the irrelevant information. Consider the following Example-2:

Example 2

Question: 小丑魚|是|一|種|熱帶|海水魚|。

"**Clownfish** is a tropical sea fish."

Candidate-1:小丑魚|原|生於|印度洋|和|太平洋|較|溫暖|的|水中|，|*包括*|*大堡礁*|*和*|*紅海*|。

"**Clownfish** are native to the warmer waters of the Indian Ocean and Pacific Ocean, *including the Great Barrier Reef and the Red Sea.*"

Candidate-2:**小丑魚**|原|生於|印度洋|和|太平洋|較|溫暖|的|水中|，
"**Clownfish** are native to the warmer waters of the Indian Ocean and Pacific Ocean."

Candidate-1 in Example-2 contains the extra information "*包括大堡礁和紅海*" ("including the Great Barrier Reef and the Red Sea") which is irrelevant to our question. Therefore, we prefer Candidate-2 which contains less unmatched terms.

Characteristic-3: Prefer more different term-types

Candidate1: $\mathbf{s_1}$ w_1^* $\mathbf{s_1}$ w_2^*

Candidate2: $\mathbf{s_1}$ w_3^* $\mathbf{s_2}$ w_4^*

Suppose $|w_1^*| = |w_3^*|$ and $|w_2^*| = |w_4^*|$, and both Candidate-1 and Candidate-2 match two terms in the above pattern. However, Candidate-1 has the same two terms s_1 but Candidate-2 has two different terms s_1 and s_2. In this case we prefer Candidate-2 as the supporting evidence because it recalls more terms from the question. Consider the following Example-3:

Example 3

Question: **電腦**|和|**手機**|已|成為|現代人|生活|的|必需品|。
"**Computers** and **mobile phones** have become necessities for modern life."

Candidate-1: **電腦**|在|許多|工作|中|廣泛|使用|，|***電腦***|也|讓|我們|生活|更加|進步|。
"**Computers** are widely used in many jobs. **Computers** also make our lives more advanced."

Candidate-2: **電腦**|在|許多|工作|中|廣泛|使用|，|***手機***|也|讓|我們|生活|更加|進步|。
"**Computers** are widely used in many jobs and **mobile phones** make our lives more advanced."

Candidate-1 only mentions the information about *computer* twice; however, Candidate-2 contains the information about both *computer* and *mobile phone* (which provide more question-related information). As the result, we prefer the candidate-2 that matches more term-types.

According to the desired characteristics of the supporting evidence mentioned above, "*Prefer more matching occurrences*" and "*Prefer less unmatched terms*" could be reflected through the *precision-rate*; and "*Prefer more different term-types*" could be reflected through the *recall-rate*. Following two cases (Table 1 and Table 2) illustrate the effect of precision and recall in retrieving the supporting evidence candidates.

Table 1. Precision and Recall for question-case-1: Question: w_1 w_2 w_3 $\underline{s_1}$ w_4

Candidate	Terms	Precision	Recall
1	w_5 $\underline{s_1}$ $\underline{s_1}$	2/3	1/5
2	w_6 w_7 w_8 $\underline{s_1}$ $\underline{s_1}$ $\underline{s_1}$	3/6	1/5

Table 1 shows that the precision-rate could truly reflect the desired Characteristic-1 and Characteristic-2. Therefore, with the precision-rate, we can successfully select the human desired Candidate-1 as the supporting evidence.

Table 2. Precision and Recall for question-case-2: Question: w_1 w_2 w_3 $\underline{s_1}$ w_4

Candidate	Terms	Precision	Recall
1	w_5 w_6 w_7 $\underline{s_1}$ $\underline{s_1}$	2/5	1/5
2	w_8 w_9 w_{10} $\underline{s_1}$ $\underline{s_2}$	2/5	2/5

However, the precision-rate alone is not enough to meet the desired Characteristic-3. For example, the precision-rate cannot tell the difference between two candidates in Case-2, since both the candidates match two terms. However, by measuring the recall-rate we can choose the better candidate that matches more terms of the question.

According to the above two cases, it clearly shows that both precision-rate and recall-rate should be involved in the scoring function for obtaining the best supporting evidence

3. Proposed Method

3.1 QE-BLEU Scoring Function

Intuitively, BLEU score (Papineni *et al*., 2002), which is a widely used metric in evaluating machine translation quality via comparing the machine-translation output with human-translation references, could be adopted for this task as it can check the similarity between the question content and the passage of the supporting evidence. BLEU score (also called CR-BLEU score, where C stands for candidate and R stands for reference) is originally defined as:

$$BLEU = BP * \prod_{n=1}^{4} p_n^{w_n} \tag{1}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \tag{2}$$

where p_n is the modified n-gram precision between machine translation candidate and a set of human translation references, w_n is the n-gram weight, r and c are the reference and candidate

lengths, respectively. BP is the brevity penalty which penalizes the candidate that is shorter than the reference. BLEU score combines each n-gram precision by multiplication.

As shown in Equation (1), CR-BLEU score only cares about the precision-rate of a candidate. However, we actually more care about the recall-rate in retrieving supporting evidence. We thus adapt the original CR-BLEU metric by letting the given question plays the role of translation candidate and each possible supporting evidence as the translation reference. Therefore, we propose an alternative QE-BLEU score which is defined as follows:

$$QE\text{-}BLEU = BP * \prod_{n=1}^{4} p_n^{w_n} \tag{3}$$

$$BP = \begin{cases} 1 & if\ \ len_E \leq len_Q \\ exp\left(1 - len_E/len_Q\right) & if\ \ len_E > len_Q \end{cases} \tag{4}$$

where Q and E denote the question and the evidence passage, respectively. Question and evidence thus correspond to the candidate and the reference, respectively, in the original function of CR-BLEU.

3.2 Modified F-measure Scoring Function

On the other hand, F-measure is a widely used evaluation metric in information retrieval which considers both precision and recall of the information retrieved (Chinchor, 1992; Sasaki, 2007). We thus prefer the F-measure, instead of BLEU score, for this task as both precision and recall are required to meet the desired characteristics listed in Section 2.

Inspired by BLEU score metric, we also apply n-gram model to consider the word order information. Therefore, we proposed a new Modified F-measure:

$$Modified\ F\text{-}measure = \sum_{n=1}^{4} \left(\frac{1}{\frac{\alpha}{p_n} + \frac{1-\alpha}{r_n}} \right)^{w_n} \tag{5}$$

where p_n and r_n denote the n-gram precision and recall of the question passage, respectively; and w_n is the corresponding n-gram weight as that in BLEU score; α is an adjustable parameter ranging from 0 to 1. If α is close to 0, Modified F-measure becomes more recall-oriented; on contrary, it becomes more precision-oriented when α is close to 1. The adopted precision and recall are defined as follows:

$$Precision = \frac{\#matched\ words\ in\ evidence}{\#words\ in\ evidence} \tag{6}$$

$$Recall = \frac{\#matched\ words\ in\ question}{\#words\ in\ question} \tag{7}$$

4. Experiments

4.1 Data Sets Adopted

We evaluate various approaches on a Taiwan elementary school social studies Yes/No question supporting evidence benchmark data set, which was created by two part-time workers and decided by the third person when there is a conflict. The original corpus consists of 178 lessons, and each lesson is composed of several paragraphs and then followed with its associated questions. We randomly divide those lessons into a development-set (124 lessons) and a test-set (54 lessons). Afterwards, we arbitrarily selected 202 and 414 questions from the development-set and the test-set, respectively. Afterwards, each question is annotated with its supporting evidence benchmark. The statistics of the benchmark is showed in Table 3.

Table 3. The statistics of the benchmark data set.

Data-Set	Development-Set	Test-Set
#Lesson	124	54
#Question	202	414
Averaged #paragraphs per lesson	26.8	30.6
Averaged #sentences per paragraph	3.7	3.6
Averaged #words per sentence	5.0	5.0
Averaged #characters per sentence	9.1	9.0

4.2 Procedure

Step 0: Preprocessing:

The raw texts of lessons and questions are segmented into words via HanLP[2] package. The punctuations are then eliminated after the segmentation (as the punctuations are only used for segmenting sentences). We had tested the case of eliminating *stop words*, but the result seems not much different. Therefore, we keep all the words in the following experiments.

After the preprocessing process, we retrieve the supporting evidence via following four steps:

Step 1: Paragraph-based search

Given a question and its corresponding lesson, we first locate the top-1 paragraph with *Apache Lucene* search engine. This step is used to cut down the search space of locating the supporting evidence.

[2] https://github.com/hankcs/HanLP

Step 2: Sentence-level candidate generation

After the above paragraph-based search, we generate various supporting evidence candidates by increasingly concatenating the consecutive sentences (up to the whole paragraph). For example, if we have a paragraph with three consecutive sentences A, B and C in order, then we will generate the following six different candidates: A, B, C, AB, BC, and ABC.

Step 3: Candidate scoring

This step is the focus of our approach. We use either QE-BLEU or Modified F-measure to score each candidate according to the given question passage.

Step 4: Select the top-1 candidate

After scoring the candidates with a specific scoring function, we then choose the candidate with the highest score as the supporting evidence.

4.3 Experiments Settings

Smoothing: We adopt the package jbleu[3], which uses the smoothing method-3[4] adopted in (Chen & Cherry, 2014) to smooth both QE-BLEU and Modified F-measure. After smoothing, they will get a small non-zero value (instead of zero) when there is no match for a given n-gram.

Weight optimization: Last, there are four n-gram weights in QE-BLEU; however, there are four n-gram weights and one additional parameter α in Modified F-measure. These parameters affect the performance of the proposed scoring functions significantly. We adopt *Particle Swarm Optimization*[5], which is known for being able to escape from the local maximum points, to automatically search for their optimal values on the development-set. We then use the obtained optimal parameters to evaluate the performance on the test-set. There are two α values tested in the Modified F-measure approach. α=0.5 is the situation to weight precision and recall equally; α=0.13 is obtained by optimizing the Modified F-measure with equal n-gram weights. And finally, α=0.12 (without smoothing) and α=0.21 (with smoothing) are the optimal values obtained by jointly optimizing the n-gram weight and α value.

4.4 Experiment Results

For various reasons, there are some benchmarks that cannot be generated by our candidate generation procedure (Step-2). Table 4 briefly lists different reasons and their associated percentages. As shown in Table 4, 16.2% of the questions are originally marked as the case

[3] GitHub repository, https://github.com/jhclark/multeval/tree/master/src/jbleu

[4] It basically assigns a geometric sequence to the n-gram that has 0 matches.

[5] https://pythonhosted.org/pyswarm/

that no appropriate evidence can be found in the text. 12.8% of the selected top-1 paragraph is different with the desired paragraph. 13.8% of the benchmarks are not a consecutive passage within a paragraph. In order to focus on comparing the effectiveness of various scoring functions, we eliminate those types of questions that the desired benchmark cannot be included in the candidate-set, and only evaluate the performance on the remaining questions (total 237 questions remained) in the following tests.

The performances of various approaches are shown in Table 5. Apache Lucene Core 5.5.0 is regarded as our baseline which uses the vector space model and a pre-specified scoring function for ranking. We adopted two widely used scoring functions, TF-IDF and BM25, as our baselines. The performances of equally weighting the n-gram are listed in the table "Equal N-gram Weight". The "+Smoothing" column shows the experiments that involve smoothing technique. The table "Optimal Weight" shows the experiments that adopt the optimized parameters which include various n-gram weights and the α value (for Modified F-measure). Again, the columns labeled with "+Smoothing" are the experiments that adopt smoothing technique with optimal weights. Table 5 shows that the overall performance of both QE-BLEU and Modified F-measure with optimal weight and smoothing technique outperform the baseline Apache Lucene (TF-IDF) about 5%.

Table 4. The statistics of the benchmark evidences that are not covered by the generated candidate-set (measured on the test set).

No evidence in the text	16.2% (67/414)
Non-Top-1 paragraph	12.8% (53/414)
Non-consecutive passage	13.8% (57/414)
Total	42.8% (177/414)

Table 5. The performances of various approaches.

Baseline:

Apache Lucene(TF-IDF)	54.43%
Apache Lucene (BM25)	46.84%

Equal N-gram Weight:

	Equal N-gram Weight	+Smoothing
QE-BLEU	37.13%	52.32%
Modified F-measure (α=0.5)	37.55%	42.19%
Modified F-measure(α=0.13)	58.23%	50.63%

Optimal Weight: (α=0.12, 0.21)

	Optimal N-gram Weight	+Smoothing
QE-BLEU	40.93%	**59.49%**
Modified F-measure	**59.92%**	**59.49%**

5. Error Analysis and Discussion

Apache Lucene: We find that Apache Lucene makes errors (for selecting the desired candidate) in the cases that the top-1 paragraph contains more sentences. This is mainly due to that IDF weight is adopted in both BM25 and TF-IDF, and IDF weight is based on the diversification of the documents to give the term weights. The term which appears in many documents is thus given a lower weight. However, various supporting evidence candidates are actually from the same paragraph (due to the way that they are created). Therefore, the term which appears in many candidates may actually be the key word (in the question) that we should pay attenuation to. As a result, Apache Lucene is not a preferable method for supporting evidence retrieval because it is related to the term distribution in the supporting evidence candidates. As shown in Table 5, the performance of BM25 is lower than that of TF-IDF. The reason is that the IDF matrix in BM25 is more sensitive, which deteriorates the performance in this task.

QE-BLEU: We find that most errors resulted from the QE-BLEU approach is due to the brevity penalty factor, as it penalizes the length of evidence candidates when the length of a candidate is longer than that of the question. In principle, the brevity penalty factor is mainly introduced to avoid involving unnecessary sentences in the evidence. However, as we mentioned in Section 1, the supporting evidence selection is only affected by the relevant and irrelevant information but not the question length. If we punish the evidence of which the length is larger than the question length, we tend to get the supporting evidence that is shorter, and might lose some relevant information.

Modified F-measure: As shown in Table 5, we test two α values: 0.5 (i.e., equally weighting precision and recall) and 0.13 (which is the optimal value obtained from the developoment-set). The performances are found about 8%~20% better when we adopt the optimal α value. However, both QE-BLEU and Modified F-measure get the same performance in the "+Smoothing" column in "Optimal Weight" (Modified F-measure improves 14 cases against QE-BLEU, but it also deteriorates the same number of cases). Furthermore, the optimal α value (α = 0.13) shows that recall is more important than precision since α= 0.13 < 0.5.

However, this model is found that it tends to find the evidence which is the longest among the candidates if we only consider recall. To avoid involving unnecessary sentences in

the evidence, the proposed approaches actually adopt two different stategies: QE-BLEU relies on *Brevity Penalty* (which penalizes the longer passage regardless of its content) and Modified F-measure relies on *Precision* (which penalizes the passage with more irrelevant content). However, utilizing *Precision* is better than adopting *Brevity Penalty* since *Brevity Penalty* only penalizes the passage with the length being longer than that of the quesiton without considering its content. To show the effect of this issue, we further extend the experiments to test Top-N (intead of Top-1 only) accuracy-rates to demonstrate the superiority of Modified F-measure. Table 6 shows that the performances of Modified F-measure are better under the columns of Top-2 and Top-3. Where "+Both" means that the experiments are under the setting of the optimal N-gram weight and smoothing technique.

Table 6. Top-N accuracy rates of QE-BLEU (+Both) and Modified F-measure (+Both)

	Top-1	Top-2	Top-3
QE-BLEU (+Both)	59.49%	72.15%	79.32%
Modified F-measure (+Both)	59.49%	73.84%	79.75%

Last, we further check 30 wrong cases from the Modified F-measure with optimal parameters along with smoothing technique. It is observed that those associated errors are mainly due to six different types as shown in Figure 2. They will be further illustrated as follows.

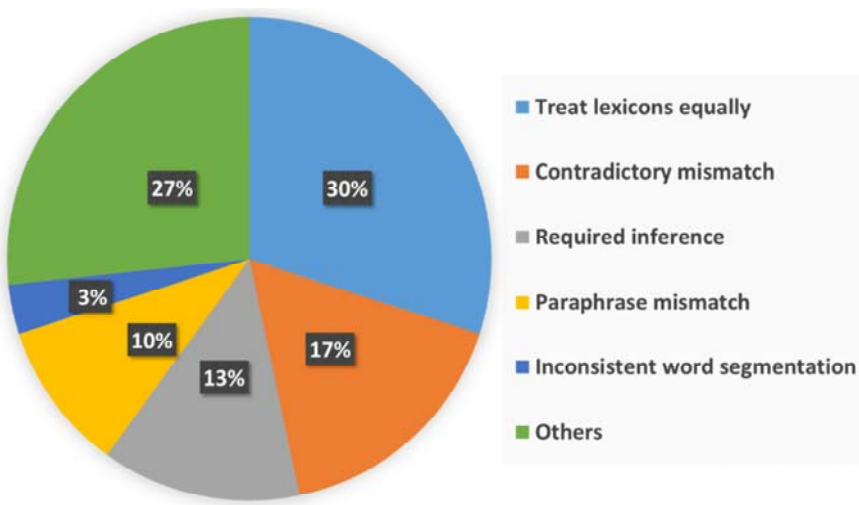


Figure 2. Error types of Modified F-measure

(1) Treat lexicons equally (30%): Since we match the terms without considering which terms are more important in the sentence, some error occurs due to the *focusing-words* are not weighted more. For example:

Question: 班級|幹部|要以|公平|，|公正|的|態度|，|引|導|同學|遵守|團體|秩序|。

"**Class leaders** should guide the classmates to abide by group order with a fair and just attitude."

Top-1 candidate: 並以|公平|，|公正|的|態度|，|引|導|同學|遵守|團體|秩序|，

"and guide them to abide by group order with a fair and just attitude."

Benchmark: 擔任|班級|幹部|，|要能|作為|同學|的|榜樣|，|並以|公平|，|公正|的|態度|，|引|導|同學|遵守|團體|秩序|，

"As a **class leader**, you should be able to serve as a role model for classmates and guide them to abide by group order with a fair and just attitude.

The terms "班級" and "幹部" ("Class leaders") are the important topic words in this Yes/No question. However, they are interleaved with other unmatched words in the first half of the benchmark. The Top-1 candidate, instead of the benchmark, is thus selected because it possesses a higher precision-rate. This kind of error need a specific technique to find the focusing-words in the sentence and give different term weights according to the degree of importance of the terms in the sentence.

(2) Contradictory mismatch (17%): Some Yes/No questions are designed to describe the wrong fact. Therefore, the sentence which describes the wrong fact would not match the evidence sentence in the lesson, but this unmatched evidence sentence still should be regarded as a part of the supporting evidence. For example:

Question: 在|從前|，|農民|參與|民俗|藝|陣|的|目的|，|是|為了|反抗|政府|而|集結|組成|的|。

"In the past, **the purpose** that farmers participate in folk art array was to assemble against the government."

Top-1 candidate: 臺灣|的|民俗|藝|陣|從前|多|是|業|餘|的|組織|，|村民|利用|農閒|時|參與|藝|陣|，

"Taiwanese folk art array used to be an amateur organization, and villagers used leisure time to participate in the folk art array."

Benchmark: 臺灣|的|民俗|藝陣|從前|多|是|業|餘|的|組織|，|村民|利用|農閒|時|參與|藝|陣|，|**既**|**可**|休閒|娛樂|、|練武|強身|，|也|間接|連絡|情誼|，|凝聚|地方|的|向心力|。

"Taiwanese folk art array used to be an amateur organization, and villagers used leisure time to participate in the folk art array. **The purpose** of it is for leisure, martial arts and also connect with friendship, condense the centripetal force of the place."

The sentence "是為了反抗政府而集結組成的" ("was assembled against the government") is the wrong fact in the question which describe the incorrect purpose of forming "民俗藝陣" ("folk art array"). Although the sentences "既可休閒娛樂、練武強身，也間接連絡情誼，凝聚地方的向心力" are not matched, they in fact provide the supporting evidence to conclude that the associated statement in the given question is incorrect. Therefore, they should be included in the supporting evidence. This kind of error also need to identify the focusing-words in the sentence, and emphasize them with larger weights.

(3) Require real-world knowledge (13%): This kind of errors is caused by the shortage of real-world knowledge. For example:

Question: 開|漳|聖王|陳元光|因|開發|漳州|有|功|而|被|當地|人們|所|**信仰**|。

"Chen Yuanguang, the Kaizhang Shengwang, was **believed** by the local people for his contribution in developing Zhangzhou."

Top-1 candidate: 宜蘭縣|壯圍鄉|開|漳|聖王|廟|祭祀|開|漳|聖王|。|因|唐朝|武進士|陳元光|開發|漳州|有|功|，

"In the Zhuangwei Township of Yilan County, the Kaizhang Shengwang Temple worship the Kaizhang Shengwang. Because Chen Yuanguang had contributed in developing Zhangzhou,"

Benchmark: 宜蘭縣|壯圍鄉|開|漳|聖王|廟|祭祀|開|漳|聖王|。|因|唐朝|武進士|陳元光|開發|漳州|有|功|，|**當地人**|**建**|**廟**|**祭祀**|，|**是**|**漳州人**|**的**|**保護神**|。

"In the Zhuangwei Township of Yilan County, the Kaizhang Shengwang Temple worship the Kaizhang Shengwang. Because Chen Yuanguang had contributed in developing Zhangzhou, **the local people built temples to honor him, he is the protecting god of the people of Zhangzhou.**"

To deal with the errors of this category, we need to know that the sentences "當地人建廟祭祀，是漳州人的保護神" ("the local people built temples to honor him, he is the protecting god of the people of Zhangzhou") implies "信仰" ("believe").

(4) Paraphrase mismatch (10%): Since we only count those "exactly matched" words, we

cannot match two terms that describe similar concepts but use different word-types. For example:

Question: **參觀**|名勝古蹟|時|要|維護|環境|的|整潔|。
"We should maintain a clean environment when **visiting** famous places and monuments."

Top-1 candidate: 古蹟|時|，|應|遵守|規定|並|維護|環境|整潔|；
"monuments, you should abide by the regulations and maintain a clean environment."

Benchmark: **拜訪**|名勝|，|古蹟|時|，|應|遵守|規定|並|維護|環境|整潔|；
"When **traveling to** famous places and monuments, you should abide by the regulations and maintain a clean environment."

The terms "參觀" ("visiting") and "拜訪" ("traveling to") have similar meaning but are not matched in string. Therefore, the capability of detecting paraphrasing is needed to deal with this kind of problems.

(5) Inconsistent word segmentation (3%): This type of errors is caused by the inconsistent word segmentation between the word in questions and lessons. For example:

Question: **名勝古蹟**|的|環境|維護|是|政府|的|責任|，|與|參訪|民眾|無關|。
"The environmental maintenance of **historical sites** is the responsibility of the government and has nothing to do with the visitors."

Top-1 candidate: 維護|家鄉|的|**名勝**|，|**古蹟**|，|需要|政府|機關|與|民間|機構|積極合作|，
"The maintenance of **historical sites** in the hometown requires active cooperation between government agencies and private institutions."

Benchmark: 維護|家鄉|的|**名勝**|，|**古蹟**|需要|政府|機關|與|民間|機構|積極合作|，|加強|對|**名勝**|，|**古蹟**|的|管理|與|修復|，|也|需要|居民|共同|關心|與|愛護|。
"The maintenance of **historical sites** in the hometown requires active cooperation between government agencies and private institutions in order to strengthen the management and restoration of **historical sites**. It also requires the resident's care and protection."

Because the same string "名勝古蹟" ("historical sites") is segmented differently in the question (as one word: "名勝古蹟") and in the candidates (as two words: "名勝" and "古蹟"), the system thus regards the second sentence in the benchmark as a purely irrelevant string.

(6) Others (27%): The errors in this category are either the cases that are caused by multiple error types mentioned above or the errors that only occupy a small portion. In the following example:

Question: 位|在|山地|丘陵|的|地方|適合|發展|林業|，|畜牧業|。

"It is suitable for the development of forestry and animal husbandry in the hilly areas."

Top-1 candidate: 山地|丘陵|等|地方|，|發展|出|林業|，|畜牧業|等|活動|；|而|居住|在|平原|地區|的|居民|，

"In hilly areas where forestry, animal husbandry and other activities are developed; for those residents living in the plains,"

Benchmark: 山地|丘陵|等|地方|，|發展|出|林業|，|畜牧業|等|活動|；

"In hilly areas where forestry, animal husbandry and other activities are developed;"

The error is caused by multiple reasons. First, because we treat lexicons equally, the last sentence in the Top-1 candidate matches the stop words which are not important. Second, the last sentence in the Top-1 candidate cannot express a meaning completely by its own. We need to detect the coherent of the sentence to deal with this kind of problem. An another example:

Question: 近年來|各縣市|親水|步道|，|河濱公園|的|設立|都|是|河川|整治|的|成果|，|不但|改善|了|河流|的|水質|，|也|提高|了|居民|的|生活品質|。

"In recent years, **some city's** hydrophilic trails and the establishment of the riverside park are the result of river remediation, which not only improves the water quality of the river, but also improves the quality of life of the residents."

Top-1 candidate: 在|整治|過程|後|，|改善|了|河流|的|水質|，|也|提高|居民|的|生活品質|。

"After the remediation process, the water quality of the river has been improved and the quality of life of the residents has also been improved."

Benchmark: **高雄市|的|愛河|曾|遭受|嚴重|污染|**，|在|整治|過程|後|，|改善|了|河流|的|水質|，|也|提高|居民|的|生活品質|。

"**The love river in Kaohsiung has been seriously polluted.** After the remediation process, the water quality of the river has been improved and the quality of life of the residents has also been improved."

In this case, we need an extra module to link "高雄市" (Kaohsiung) to "各縣市" (some city) because "Kaohsiung" is an instance of "some city".

6. Related Work

As mentioned in Section 1, the previous studies of retrieving supporting evidence can be grouped into three categories: matching terms, conducting syntactic/semantic analysis, and scoring with a translation model. Term matching approaches focus on retrieving the related query from a large scale of documents by using similarity functions and word weight functions. For example, DrQA system (Chen *et al*., 2017) was developed for large scale applications such as retrieving the relevant documents from Wikipedia. In their document retriever model, they evaluated the similarity of the articles and questions by the score of TF-IDF weighted bag-of-word vectors. They also improved the model by taking bi-gram counts. However, those approaches usually do not consider word order and local context.

Syntactic/semantic scoring approaches are specially developed to deal with certain QA datasets. The DeepQA pipeline in IBM Watson system (Murdock *et al*., 2012), which is used in the task Jeopardy![6], presented four passage-scoring algorithms to retrieve the supporting evidence by scoring the passages. The scoring algorithms operate on the syntactic-semantic graphs constructed from analyzing the syntactic and semantic information of the documents. The QA system in (Jansen *et al*., 2017) was developed for standardized science exams. They extracted the focus words according to their scores of the concrete concepts. The words are scored by the psycholinguistic concreteness and rated from 1 (highly abstract) to 5 (highly concrete) by human. Nonetheless, this kind of approaches is more complex and their operations are usually more time-consuming.

Translation model based approaches are widely adopted in community Q&A tasks. They mainly check the similarity between the queried question and those historical questions kept in the archive with a translation model (in which a higher translation score implies that they are more similar). In our case, this approach translates the given question into the specified supporting evidence candidate via a translation model, and then assigns the obtained translation score as the associated score of that candidate. These approaches can be further categorized into word-based approaches and phrase-based approaches. Word-based approaches (Berger *et al*., 2000; Jeon *et al*., 2005; Xue *et al*., 2008) adopt word translation probabilities in a language model to rank the similarity. Zhou *et al*. (2011) further extended this model into a phrase-based one and obtained better performances. This kind of approaches clearly needs large benchmark data which is expensive to construct in our task.

In comparison with previous term matching approaches, our proposed n-gram matching

[6] Jeopardy! is an American television game show.

approaches further consider word order and local context, and thus improve the retrieval accuracy. On the other hand, for those syntactic/semantic scoring approaches, the proposed approaches can operate more efficiently due to the use of simple string matching. Last, comparing with those translation model based approaches, our approaches do not need large training data.

7. Conclusion

Two different models are proposed in this paper to retrieve supporting evidence for the given Yes/No question: *QE-BLEU* and *Modified F-measure*. In comparison with previous approaches, the proposed approaches provide better accuracy and efficiency. Both of them adopt n-gram to incorporate phrases and local context; however, the Modified F-measure takes care of both precision and recall, while QE-BLEU only handles recall of the question. Experiment results have shown that both of them outperform Lucene Apache search engine by 5%.

Our main contributions mainly are: (1) We proposed and tested two novel approaches to retrieve the supporting evidence, and have obtained better performances. (2) We list the desired characteristics of the supporting evidence retrieved. (3) We implement and compare various refinement techniques, including smoothing and optimization, for the proposed approaches.

References

Berger, A., Caruana, R., Cohn, D., Freitag, D. & Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 192-199. doi: 10.1145/345508.345576

Chen, B. & Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. doi: 10.3115/v1/W14-3346

Chen, D., Fisch, A., Weston, J. & Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. Retrived from *arXiv preprint arXiv:1704.00051*

Chinchor, N. (1992). The statistical significance of the MUC-4 results. In *Proceedings of the 4th conference on Message understanding*, 30-50. doi: 10.3115/1072064.1072068

Jansen, P., Sharp, R., Surdeanu, M. & Clark, P. (2017). Framing QA as Building and Ranking Intersentence Answer Justifications. *Computational Linguistics*, *43*(2), 407-449. doi: 10.1162/COLI_a_00287

Jeon, J., Croft, W. B. & Lee, J. H. (2005). Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 84-90. doi: 10.1145/1099554.1099572

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge university press.

Murdock, J. W., Fan, J., Lally, A., Shima, H. & Boguraev, B. K. (2012). Textual evidence gathering and analysis. *IBM Journal of Research and Development*, *56*(3-4), 8:1-8:14. doi: 10.1147/JRD.2012.2187249

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, 311-318. doi: 10.3115/1073083.1073135

Sasaki, Y. (2007). The truth of the F-measure. Teach Tutor mater. Retrived from *http://www.flowdx.com/F-measure-YS-26Oct07.pdf*

Xue, X., Jeon, J. & Croft, W. B. (2008). Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 475-482. doi: 10.1145/1390334.1390416

Zhou, G., Cai, L., Zhao, J. & Liu, K. (2011). Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 653-662.

Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 67-84

未登錄詞之向量表示法模型於中文機器閱讀理解之應用[1]

An OOV Word Embedding Framework for Chinese Machine Reading Comprehension

羅上堡*、李青憲+、涂家章+、陳冠宇*

Shang-Bao Luo, Ching-Hsien Lee, Jia-Jang Tu and Kuan-Yu Chen

摘要

在使用深度學習(Deep Learning)方法於自然語言處理的問題時，我們通常會先將每一個詞以一個相對應的詞向量(Word Embedding)表示，再輸入至各式神經網路模型。當遭遇未登錄詞(Out-of-Vocabulary, OOV)的問題時，最常見的處理方式是略去該未登錄詞、以一個零向量表示或是用一個隨機產生的向量表示這個未登錄詞。就我們所知，在目前的研究裡，似乎仍未有一套合理且快速的做法，用於產生未登錄詞的詞向量表示法，並進一步地探索未登錄詞的詞向量對於任務成效的影響性。因此，本論文提出一套新穎的詞向量表示法學習技術，其目標是為未登錄詞產生一個較為合理且可靠的低維度向量表示法；除此之外，我們將進一步地把此一技術運用於中文機器閱讀理解任務之中，探究未登錄詞對於中文機器閱讀理解任務之影響，並驗證本論文所提出的詞向量表示法學習技術之成效。

關鍵詞：自然語言處理、詞向量表示法、未登錄詞、機器閱讀理解

---

[1] This research was partially supported by the Project H367B83300 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

* 國立臺灣科技大學系資訊工程系
Department of Computer Science & Information Engineering, National Taiwan University of Science and Technology
E-mail: {M10615012, kychen}@mail.ntust.edu.tw

+ 工業技術研究院巨資中心
Computational Intelligence Technology Center, Industrial Technology Research Institute
E-mail: {C.H.Lee, santu}@itri.org.tw

Abstract

When using Deep Learning methods in NLP-related tasks, we usually represent a word by using a low-dimensional dense vector, which is named the word embedding, and these word embeddings can then be treated as feature vectors for various neural network-based models. However, a major challenge facing such a mechanism is how to represent OOV words. There are two common strategies in practiced: one is to remove these words directly; the other is to represent OOV words by using zero or random vectors. To mitigate the flaw, we introduce an OOV embedding framework, which aims at generating reasonable low-dimensional dense vectors for OOV words. Furthermore, in order to evaluate the impact of the OOV representations, we plug the proposed framework into the Chinese machine reading comprehension task, and a series of experiments and comparisons demonstrate the good efficacy of the proposed framework.

Keywords: Natural Language Processing, Word Embedding, Out-of-vocabulary, Machine Reading Comprehension

1. 緒論 (Introduction)

機器閱讀理解(Machine Reading Comprehension, MRC)是一個自然語言處理(Natural Language Processing, NLP)領域中相當重要的任務，其目標是希望讓機器像人類一樣進行文本閱讀，並根據對該文本之理解，進而回答相關的問題。讓電腦幫助人類在大量文本中找到想要的答案，可以減輕資訊獲取的成本、加速資訊處理的速度、以及提升資訊的利用率。進一步地，如果電腦能具備相當高水準的閱讀理解能力，許多應用將會有更進一步的發展，例如問答(Question Answering)、對話系統(Dialogue System)以及搜尋引擎(Search Engine)等。因此機器閱讀理解不論在學術界或產業界都有極高的研究價值。

目前機器閱讀理解的研究主要有完型填空(Cloze Style)與文本段(Text Span)預測等兩種型式。完型填空是去掉文本中的某個詞語，讓系統進行填空，但答案往往是單一的字詞，並不需要對於整段文本進行理解，因此這類型的回答形式較難以延伸應用於實際生活中。有鑑於完型填空之不足，2016 年時，一個大規模的文本段類型數據集 SQuAD (The Stanford Question Answering Dataset) (Rajpurkar, Zhang, Lopyrev & Liang, 2016)應運而生。SQuAD 資料集包含十萬多個問題答案組，文本來至維基百科，因此答案就是維基百科文本中的一個小段落；更明確地，文本段的預測方式為給定文本與問題後，機器需以文本中一個連續的小片段來回答給定的問題。

由於深度學習的蓬勃發展並且在許多領域中取得空前的好成績，目前多數的機器閱讀理解模型皆是建構於深度學習的方法上。基於深度學習之機器閱讀理解模型，主要可區分成五個模塊，嵌入層(Embedding Layer)、嵌入編碼層(Embedding Encoder Layer)、段落問題注意機制(Passage-question Attention)、注意力編碼層(Attention Encoder Layer)以及

輸出層(Output Layer)。嵌入層主要是將每一個詞轉換成一個相對應的詞向量表示法，有些模型會額外加入各式語言特徵(Linguistic Features)來豐富每一個詞的語意或語言資訊；嵌入編碼層目標於探究詞與詞之間的上下文關係，並基於這個資訊，為每一個詞產生一個新的低維度向量表示法；段落問題注意機制是藉由注意力機制(Attention Mechanism)將每一段落（或文本）表示為含有問題意識之段落表達(Question-aware Passage Representation)；注意力編碼層(Attention Encoder Layer)則是堆疊在段落問題注意機制之後，利用段落問題注意機制所產生的詞向量與雙向循環神經網路並搭配各式注意力機制，產生更進階的詞向量表示法；在文本段預測的機器閱讀理解模型裡，輸出層通常採用指針網路(Pointer Network)生成兩個分別代表開始與結束的機率分布，藉由簡單的搜尋演算法，將文本中被預測為答案開始與結束的位置標示出來。值得一提的是，段落問題注意機制可視為機器閱讀理解模型中最重要且關鍵的元件，各式機器閱讀理解模型多半著墨於此層中，並提出各式不同的模型架構來改善機器閱讀理解模型之成效。另外，傳統的機器閱讀理解模型多是以循環神經網路(Recurrent Neural Network, RNN)為主要架構，但由於循環神經網路較難以實現並行運算，因此多數模型皆易遭受訓練時間過久的問題。為解決此一問題，近年來有研究提出一套新穎的模型方法 QANet (Yu *et al.*, 2018)，將機器閱讀理解模型中常見的循環神經網路架構全部捨棄，只採用卷積神經網路(Convolution Neural Network, CNN)與可並行的注意力機制來建立機器閱讀理解模型，不僅大幅降低訓練所耗費的時間，其成效依然相當傑出。

在各式自然語言處理的任務中，未登錄詞是一個基礎且重要的問題。在機器閱讀理解任務裡，各式模型為了減緩此一問題造成的影響，多半在嵌入層裡除了每一個詞的詞向量外，會再額外利用字向量(Character Embedding)來豐富每一個詞的語彙資訊。QANet、BiDAF (Seo, Kembhavi, Farhadi & Hajishirzi, 2016)、jNet (Zhang *et al.*, 2017)、MEMEN (Pan *et al.*, 2017)與 ReasoNet (Shen, Huang, Gao & Chen, 2017)等，皆是利用卷積神經網路對一連串的字向量提取字與字之間相連的結構資訊(Kim, Jernite, Sontag & Rush, 2016)，再轉換成固定維度的向量表示法；R-NET (Wang, Yang, Wei, Chang & Zhou, 2017)與 S-NET (Tan *et al.*, 2017)則是利用雙向循環神經網路提取字向量的前後規則資訊；RMR (Reinforced Mnemonic Reader) (Hu, Wei, Mao & Chikina, 2017)、Conductor-net (Liu *et al.*, 2017)、V-Net (Wang *et al.*, 2018)、FastQA (Weissenborn, Wiese & Seiffe, 2017)與 Smartnet (Chen *et al.*, 2017)是將字向量直接與詞向量進行串接，形成新的向量表示法。綜觀上述各式模型方法，皆是以英文為處理目標，並且以機器閱讀理解任務為導向，在嵌入層中，訓練一組字向量表示法模型，用以彌補未登錄詞在機器閱讀理解任務中所造成的影響。然而，這樣的方式並非實際地為每一個未登錄詞產一個合理且適當的向量表示法，且使其可以應用於各式任務之中。有鑑於此，本論文旨於提出一套新穎的詞向量表示法學習技術，為每一個未登錄詞產生一個較為可靠的詞向量表示法，並且，本論文進一步地將此一技術運用於中文機器閱讀理解任務之中，實驗結果顯示，結合本論文提出之詞向量表示法學習技術，可以有效提升中文機器閱讀理解任務之成效。

2. 相關方法 (Related Methods)

2.1 詞向量表示法 (Word Embedding)

詞向量表示法(Mikolov, Chen, Corrado & Dean, 2013)是深度學習於自然語言處理中相當成功的應用之一，其目的是將每一個詞以一個低維度的向量表示之。常見的詞向量表示法模型有連續型詞袋模型(Continuous Bag-of-Words, CBOW)、略詞模型(Skip-gram)與全局向量模型(GloVe) (Pennington, Socher & Manning, 2014)。

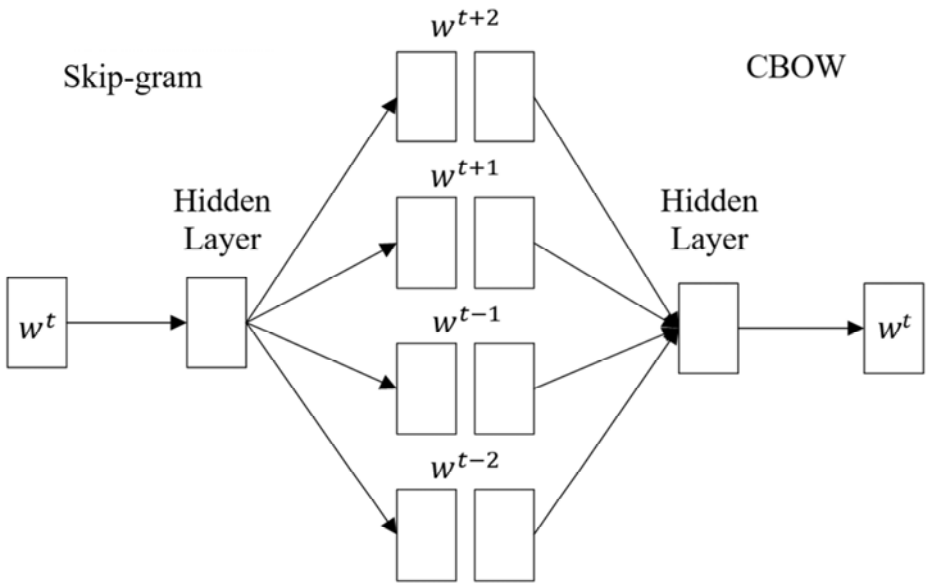


圖 1. 略詞模型(Skip-gram)與連續型詞袋模型(CBOW)示意圖
[Figure 1. Illustrations of Skip-gram and CBOW models.]

詞向量表示法學習的起源可以追溯至 2003 年被提出的神經網路語言模型(Bengio, Ducharme, Vincent & Jauvin, 2003)，其使用前饋式神經網路(Feed-forward Neural Network)來建立 N 連語言模型，在這個模型架構下，自然地獲得了每一個詞的向量表示法，通常稱之為詞向量。衍生至今，目前多數的研究著眼於將每一個字、詞用一個連續數值的向量(Distributed Vector)來表示。不同於神經網路語言模型以建立一個 N 連語言模型為目標，而每一個詞的向量表示法僅是模型建立過程的副產物，連續型詞袋模型的目標即是為每一個詞建立專屬的詞向量。連續型詞袋模型的架構類似於前饋式神經網路，但為了節省運算時間與參數量，省略了前饋式神經網路中的隱藏層，不僅打破了過去各式神經網路模型訓練耗時的缺點，並大幅提升此表示法學習的實用性，其架構如圖 1 所示。當給定一連串的文字序列：$w^1, w^2, \dots, w^T$，連續性詞袋模型的目標函數(Objective Function)是最大化每一個詞出現的可能性：

$$\sum_{t=1}^{T} logP(w^t | w^{t-c}, \dots, w^{t-1}, w^{t+1}, \dots w^{t+c}) \quad (1)$$

其中，c 表示相對於詞w^t的左右窗函數(Window Function)，T 表示文字序列的總長度，而條件機率$P(w^t|w^{t-c},\dots,w^{t-1},w^{t+1},\dots w^{t+c})$則為：

$$P(w^t|w^{t-c},\dots,w^{t-1},w^{t+1},\dots w^{t+c}) = \frac{\exp(v_{\bar{w}^t}\cdot v_{w^t})}{\sum_{i=1}^{V}\exp(v_{\bar{w}^t}\cdot v_{w_i})} \quad (2)$$

其中，V 為辭典的總詞數，v_{w^t}是詞w^t的詞向量，$v_{\bar{w}^t}$則是出現在詞w^t左右相鄰詞的詞向量之加權平均向量：

$$v_{\bar{w}^t} = \sum_{\substack{j=-c \\ j\neq 0}}^{c} \alpha_j v_{w^{t+j}} \quad (3)$$

其中，α_j為與距離有關的權重值。連續型詞袋模型的概念源起於分散式假說(Distributional Hypothesis)，也就是語意相近的詞通常其左右相鄰的詞不會差異很大，因此，建立詞向量時左右相鄰詞的詞向量是很重要的參考。

略詞模型的訓練目標函數則與連續型詞袋模型相反。當給定一連串的文字序列：$w^1, w^2, \dots, w^T$，略詞模型的目標函數為：

$$\sum_{t=1}^{T}\sum_{\substack{j=-c \\ j\neq 0}}^{c} log P(w^{t+j}|w^t) \quad (4)$$

其中，條件機率$P(w^{t+j}|w^t)$可表示為：

$$P(w^{t+j}|w^t) = \frac{\exp\left(v_{w^{t+j}}\cdot v_{w^t}\right)}{\sum_{i=1}^{V}\exp\left(v_{w_i}\cdot v_{w^t}\right)} \quad (5)$$

V 為辭典的總詞數，v_{w^t}與$v_{w^{t+j}}$分別表示詞w^t與w^{t+j}的詞向量，其模型架構如圖 1 所示。雖然連續型詞袋模型以及略詞模型在模型架構已相當簡化，在實作的過程中，前人的研究更提出了階層式軟性最大化演算法(Hierarchical Soft-max Algorithm) (Mnih & Hinton, 2009)以及負取樣演算法(Negative Sampling Algorithm) (Mikolov, Sutskever, Chen, Corrado & Dean, 2013)來加速模型參數（即詞向量）的估算過程。

由於連續型詞袋模型以及略詞模型在訓練的過程中，僅考慮短距離的詞彙規則關係，全局向量模型認為在求取詞向量時，應當考慮的是整個訓練語料中，詞與詞之間的相互關係，並且詞與詞之間的關係不應該以最大化預測機率來描述，應該考慮詞對(Word Pair)與詞對之間的比例關係，綜合以上考量，全局向量模型的目標函數為：

$$\sum_{i=1}^{V}\sum_{j=1}^{V} f\left(X_{w_iw_j}\right)(v_{w_i}\cdot v_{w_j} + b_{w_i} + b_{w_j} - log X_{w_iw_j})^2 \quad (6)$$

同樣地，V 為辭典的總詞數，v_{w_i}與v_{w_j}分別表示詞w_i與w_j的詞向量，$X_{w_iw_j}$代表詞w_i與w_j在訓練語料中共同出現的次數，$f(\cdot)$為一個單調的平滑函數，用來調整每一個詞對在訓練過程中的影響（重要）性，而b_{w_i}則為詞w_i的基數(Bias)。

除了經典的詞向量表示法模型外，上下文向量表示法(Context Vectors, CoVe) (McCann, Bradbury, Xiong & Socher, 2017)借鑑於遷移學習(Transfer Learning)的概念，將

各式詞向量表示法模型所求得的詞向量做為預訓練(Pre-trained)的模型參數，接著利用序列至序列(Sequence-to-Sequence) (Sutskever, Vinyals & Le, 2014)的方式進行機器翻譯(Machine Translation)模型的訓練，藉由最佳化機器翻譯為目標，調適出一組新的詞向量表示法。最後，上下文向量表示法是將新的詞向量與原始的詞向量表示法串接，作為一個新的詞向量表示法。在許多任務上已證實，上下文向量表示法確實可以獲得更好的任務成效。快文向量模型(FastText) (Bojanowski, Grave, Joulin & Mikolov, 2016)則為略詞模型之延伸，不同之處在於，略詞模型是以詞為單位，通過目標單詞來預測其上下文中之其他詞彙的出現機率，而快文向量模型則是以字符為單位，因此，每一個詞彙的向量表示法是詞彙中所有字符向量的平均。

2.2 問答模型 (Question Answering)

有鑑於傳統的機器閱讀理解模型多半採用循環神經網路為基礎，容易遭受訓練時間過長的問題，QANet 改以卷積神經網路為基礎，提出一套嶄新的機器閱讀理解模型，不僅有效地縮短訓練所需的時間，也在許多實驗中，被驗證可獲得相當不錯的任務成效。

QANet 包含五個主要元件：嵌入層(Embedding Layer)、嵌入編碼層(Embedding Encoder Layer)、語境查詢注意力層(Context-query Attention Layer)、模型編碼層(Model Encoder Layer)以及輸出層(Output Layer)。除了在嵌入編碼層與模型編碼層捨棄傳統主流的循環神經網路，改採卷積神經網路外，QANet 亦引入自我注意力機制(Self-attention Mechanism) (Vaswani *et al*., 2017)，用以擷取每一個詞與整個文本中每一個詞之間的關係，並且可以採用平行化的訓練方式，使得訓練速度與推論(Reasoning)的速度更快。QANet 的模型架構如圖 2 所示，值得注意的是，模型中有許多堆疊架構，架構裡的神經網路結構皆是相同的，並且每一層之間皆使用層正規化(Layer Normalization) (Ba, Kiros & Hinton, 2016)與殘差(Residual Network) (He, Zhang, Ren & Sun, 2016)技術來穩定訓練過程。除此之外，為了增加泛化能力，QANet 的文本與問題編碼器是共享權重的。

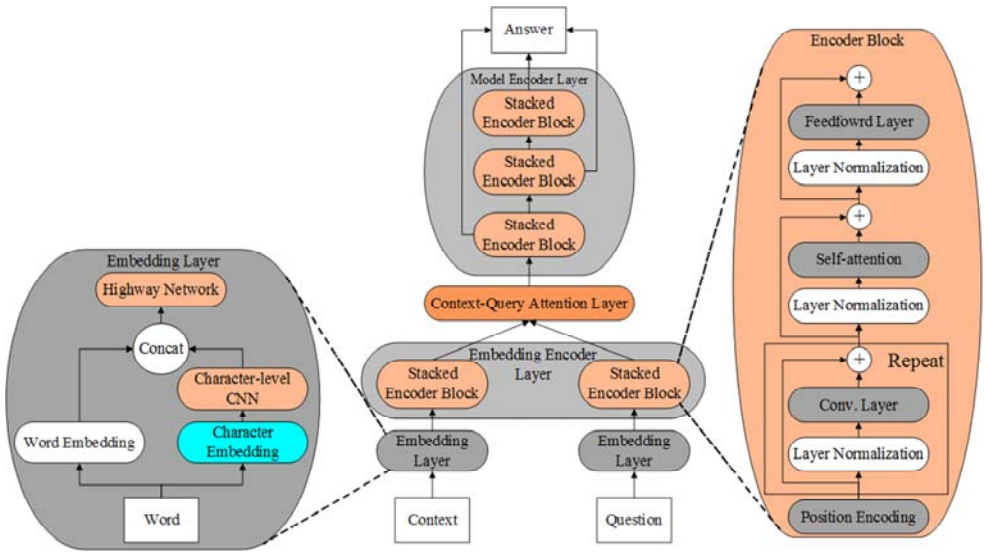


圖 2. QANet 模型示意圖
[Figure 2. Illustration of QANet Model]

3. 新穎的詞向量表示法模型技術 (Novel Word Embedding Model Technique)

在使用深度學習方法於自然語言處理的問題時，我們通常會先將每一個詞以一個相對應的詞向量進行表示，作為各式神經網路模型之輸入。當遭遇未登錄詞的問題，最常見的處理方式是略去未登錄詞、以一個零向量表示之或是用一個隨機產生的向量表示這個未登錄詞。為此，本論文提出一套新穎的詞向量表示法學習技術，為未登錄詞產生一個較為可靠的詞向量表示法，並進一步地將此一技術運用於中文機器閱讀理解任務之中，探究未登錄詞對於中文機器閱讀理解任務之影響，並驗證本論文提出的詞向量表示法學習技術之成效。

中文是字符語言(Character-based Language)，通常每一個字都有其獨特的意義，或是字符的形狀與其所對應的事物有關聯。詞(Word)通常是由兩個以上的字(Character)所組成，用以表達某一種事、物或現象，而只用一個字所成的詞通常稱之為單字詞。由於新字的生成是複雜且繁瑣的，所以我們假設所有的中文字符是固定且已知的$V_C = \{c_1, c_2, \dots, c_{|V_C|}\}$。藉由一份文字語料，傳統的各式詞向量表示法模型可用來為字典 V 中的每一個詞 $\{w_1, w_2, \cdots, w_{|V|}\}$產生一個相對應的低維度向量表示法$\{\{v_{w_1}, v_{w_2}, \cdots, v_{w_{|V|}}\}$。由於每一個詞所欲表達的事、物或現象常與詞中字的意義或字與字之間的排列順序有關，因此我們可以利用每一個詞的詞向量為學習目標，為每一個字求取一個相對應的字向量表示法(Chen, Wang & Chen, 2015)。之後，當遇到未登錄詞時，就可以透過字向量表示法取得此一未登錄詞的詞向量表示法。

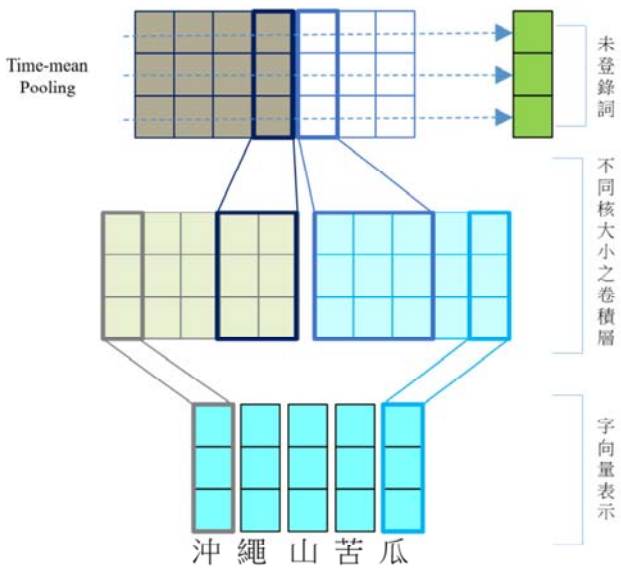


圖 3. 基於卷積神經網路的未登錄詞之向量表示法模型
[Figure 3. Convolutional Neural Network-based Out-of-Vocabulary Embedding Model, COEM]

為了達成此一目的，我們嘗試提出兩種模型架構：基於卷積神經網路的未登錄詞之向量表示法模型(Convolutional Neural Network-based Out-of-Vocabulary Embedding Model, COEM)，如圖 3 所示；基於循環神經網路的未登錄詞之向量表示法模型(Recurrent Neural Network-based Out-of-Vocabulary Embedding Model, ROEM)，如圖 4 所示。在基於卷積神經網路的未登錄詞之向量表示法模型中，我們首先將每一個詞以數個相對應的字向量表示之，這些字向量先經由全連接層進行線性轉換後，接著透過不同核大小(Kernel Size)的卷積神經網路，探究字符間特定距離的排列順序資訊，最後利用池化層(Pooling Layer)，將各式特徵取平均後作為輸出。在基於循環神經網路的未登錄詞之向量表示法模型裡，每一個詞同樣以一連串的字向量表示之，接著藉由雙向循環神經網路抽取字與字之間長距離的意義關係，再將雙向循環神經網路最後一個時間點所產生的隱藏層輸出進行串接，最後經過全連接層產生輸出結果。

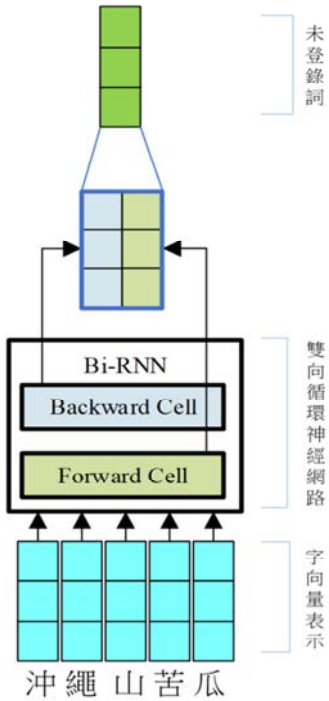


圖 4. 基於循環神經網路的未登錄詞之向量表示法模型
[Figure 4. Recurrent Neural Network-based Out-of-Vocabulary Embedding Model, ROEM]

總言來說，本論文提出的未登錄詞之向量表示法模型訓練可分為兩階段：第一階段為利用經典的各式詞向量表示法模型求取一組詞向量；在第二階段中，我們以第一階段所獲得的詞向量為目標，訓練本論文所提出的未登錄詞之向量表示法模型，其中包含神經網路的模型參數以及一組字向量表示法。更明確地，我們將目標詞向量與模型之輸出向量進行目標函示定義為：

$$Loss = \sum_{i=1}^{V}(v_{w_i} - \varphi(v_{c^{i,1}}, \cdots, v_{c^{i,|w_i|}}))^2 \tag{7}$$

其中，v_{w_i}是詞w_i的在第一階段所獲得的詞向量表示法，$|w_i|$表示詞w_i內所含字的個數，$v_{c^{i,j}}$表示詞w_i中第j個字的字向量，$\varphi(\cdot)$為本論文所提出的未登錄詞之向量表示法模型。本論文藉由此目標函數來訓練未登錄詞之模型，並且訓練流程架構如圖 5 所示，其中輸入為該詞之字向量表達，輸出為未登錄詞之詞向量。

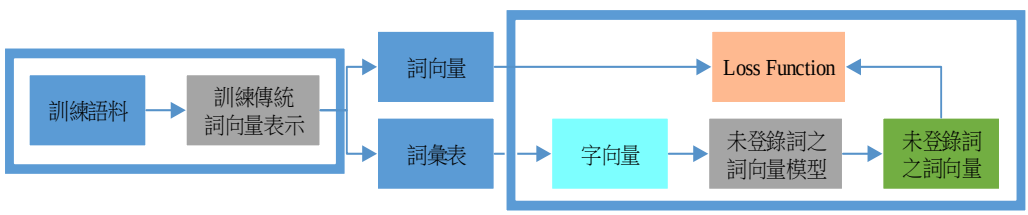


圖 5. 未登錄詞之詞向量模型整體訓練示意圖
[Figure 5. Training Flowchart for the Proposed Framework]

4. 實驗設定與結果 (Experimental Settings and Results)

4.1 實驗設定 (Experimental Settings)

本論文蒐集批踢踢實業坊(PTT)與中央社新聞(CNA)作為文本資料，用於訓練各式詞向量表示法模型(即連續型詞袋模型(CBOW)、略詞模型(Skip-gram) 與全局向量模型(GloVe))，詞向量的維度大小設定為 300 維，並將詞頻小於 5 的詞彙捨去。在本論文所提出之新穎的詞向量表示法模型（即 COEM 與 ROEM）裡，同樣將字向量表示法的維度設定為 300 維，因此，每個詞都會重新以 16 個 300 維的字向量表示之，若某一個詞長度超過 16 個字，則將過長的字捨去，相反地，長度不足 16 個字的詞，則進行貼補(Padding)。在基於 COEM 中，全連接層輸入與輸出皆為 300 維，卷積層共包含四組分別核大小為 2、4、6 與 8，跨步(Stride)大小為 1，過濾器個數(Filter Size)皆設定為 300。在 ROEM 中，雙向循環神經網路的隱藏層大小設定為 300，全連接層輸出大小為 300。此外，除了循環神經網路激活函數採用雙曲函數(Tanh)外，其它激活函數皆採用 TanhShrink。中文文本斷詞是採用結巴(Jieba) (Sun, 2012)作為斷詞器。在訓練新穎的詞向量表示法模型時，我們採用 Adam (Kingma & Ba, 2014)演算法做為求取參數的優化器。所有實驗皆以 python 3.5.2 與 Tensorflow1.6.0 (Abadi *et al*., 2016)套件實現。

為了比較傳統詞向量表示法與本論文提出的未登錄詞詞向量表示法模型於中文機器閱讀理解任務之成效，我們使用台達電閱讀理解資料集(Delta Reading Comprehension Dataset, DRCD) (Shao, Liu, Lai, Tseng & Tsai, 2018)進行各式實驗與觀察，並將此資料集切分為訓練集、發展集與測試集，其統計資訊如表 1 所示，其中 DRCD 在我們蒐集之文本資料上，未登錄詞在訓練集、發展集與測試集各佔 65.50%、46.19%與 46.20%。由於本論文著眼於不同詞向量表示法在中文機器閱讀理解任務的成效，因此我們將基於前人在 SQuAD 上所提出之機器閱讀理解模型架構 QANet，把各式詞向量或其結合作為輸入，驗證其成效。在 QANet 中，隱藏層大小設定為 96，多重注意力機制(Multi-attention head number)數量為 2，訓練資料批次大小為 32，總期次數為 75000。值得一提的是，各式詞向量（即連續型詞袋模型(CBOW)、略詞模型(Skip-gram)與全局向量模型(GloVe）與本論文所提出之新穎的詞向量表示法模型皆不會在訓練閱讀理解模型時再被調整、更新，但 QANet 架構中的字向量表示法，則會隨著模型訓練而改變。

表 1. 台達電閱讀理解資料集之統計資訊
[Table 1. Statistics on Delta Reading Comprehension Dataset]

訓練集 Training Set			發展集 Development Set			測試集 Test Set		
文本數	問題數	OOV Ratio	文本數	問題數	OOV Ratio	文本數	問題數	OOV Ratio
1,960	26,936	65.50%	383	3,524	46.19%	383	3,524	46.20%

在中文機器閱讀理解的實驗中，我們採用標準的 F1 與 EM(Exact Match)分數作為評估指標。EM 指標是當模型預測之答案與正確答案完成一致，才會獲得分數；F1 指標是

資訊檢索領域常用的評分方法，是基於精確度(Precision)與召回率(Recall)計算而得。

4.2 實驗結果 (Experimental Results)

在第一組實驗中，我們首先比較各式基於 QANet 與傳統詞向量表示法模型之基準系統(Baseline Systems)，發展集與測試集實驗結果分別詳列於表 2 與表 3。在基準系統中，我們將傳統詞向量作為 QANet 的輸入，而未登錄詞可分別以零向量或隨機向量來表示，分別標示為 Baseline(Zero)與 Baseline (Rand)。除了以詞向量表示法作為輸入外，我們進一步地將字向量亦加入模型之中，其實驗結果則標註為 Baseline(Zero)＋Char 與 Baseline (Rand)＋Char。除了標準的基準系統外，我們亦嘗試將訓練文本資料切割成一個個的字，接著分別利用連續型詞帶模型、略詞模型以及全局向量模型訓練一組字向量表示模型，因此未登錄詞的詞向量表示法則為該詞中所有字向量表示法的平均。此一系統可以視為一個強健性基礎系統，我們標示為 Strong Baseline。當然字向量亦可加入模型之中，其實驗結果則標註為 Strong Baseline＋Char。首先，結合傳統詞向量表示法模型與 QANet（即 Baseline(Zero)與 Baseline(Rand)），不論是使用連續型詞袋模型、略詞模型或全局向量模型，都可獲得超過 5%與 7%的 F1 與 EM 分數。當我們進一步地將字向量表示法也加入模型中後，不論使用何種 F1 與 EM 分數作為評估指標，都可以進一步地獲得效能的提升。另外，我們發現在 F1 上，略詞模型比連續型詞袋模型和全局模型有更好的表現；然而在 EM 評估標準上，則是連續型詞袋模型與略詞向量有更好之結果。我們還發現在於為每個未登錄詞產生隨機向量的做法，在不考慮字向量表示的時候，是有實質提升的作用，但在加入字向量表示與零向量的作法可以發現，結果都會略低於零向量的效能。

表 2. 運用各式詞向量表示法模型於基礎系統中之機器閱讀理解任務成效(發展集)
[Table 2. Experimental Results on Development Set with Respect to Various Word Embedding Methods and Baseline Systems]

Development Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
Baseline(Zero)	71.34%	55.34%	70.82%	55.59%	70.44%	53.43%
Baseline(Zero)+Char	**80.24%**	**68.00%**	**80.73%**	**68.00%**	**80.58%**	**67.25%**
Baseline(Rand)	76.92%	63.49%	76.97%	62.88%	75.86%	61.83%
Baseline(Rand)+Char	79.46%	66.62%	79.71%	66.20%	78.44%	64.56%
Strong Baseline	78.72%	64.65%	79.20%	65.43%	79.44%	65.37%
Strong Baseline+Char	**79.84%**	**67.25%**	**79.77%**	**66.03%**	**80.60%**	**66.09%**

表 3. 運用各式詞向量表示法模型於基礎系統中之機器閱讀理解任務成效(測試集)
[Table 3. Experimental Results on Test Set with Respect to Various Word Embedding Methods and Baseline Systems]

Test Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
Baseline(Zero)	70.72%	54.85%	70.26%	54.94%	69.42%	53.21%
Baseline(Zero)+Char	**79.53%**	**67.22%**	**80.10%**	**67.54%**	**80.07%**	**66.69%**
Baseline(Rand)	76.14%	62.60%	77.29%	63.12%	77.79%	63.73%
Baseline(Rand)+Char	79.32%	66.46%	79.84%	66.41%	79.32%	66.37%
Strong Baseline	78.26%	64.36%	79.76%	66.29%	78.92%	64.59%
Strong Baseline+Char	**79.88%**	**66.40%**	**80.79%**	**67.85%**	**79.94%**	**66.92%**

接著，我們探討本論文所提出之未登錄詞詞向量表示法模型之成效。這組實驗共分為三種輸入方式來進行實驗，如圖 6 所示。首先，我們使用本論文提出之 COEM 與 ROEM 為每個詞（包括登錄詞與未登錄詞）產生對應的詞向量，做為機器閱讀理解模型的輸入（實驗結果標註為 COEM 與 ROEM）；其二，基於本論文提出之 COEM 與 ROEM 為每個詞(包括登錄詞與未登錄詞)產生的對應詞向量外，再加上字向量，一起做為機器閱讀理解模型的輸入（實驗結果標註為 COEM+Char 與 ROEM+Char）；接著，我們將傳統詞向量表示法與本論文提出之未登錄詞向量表示法結合，其作法是將未登錄詞使用 COEM 與 ROEM 來產生詞向量，而登錄詞則使用原本詞向量表示法所得的向量（實驗結果標註為 WE+COEM 與 WE+ROEM）；最後，我們進一步地將傳統詞向量表示法與本論文提出之未登錄詞向量表示法結合，並且加上字向量來進行實驗（實驗結果標註為 WE+COEM+Char 與 WE+ROEM+Char）。發展集與測試集的實驗結果如表 4 與表 5 所示。

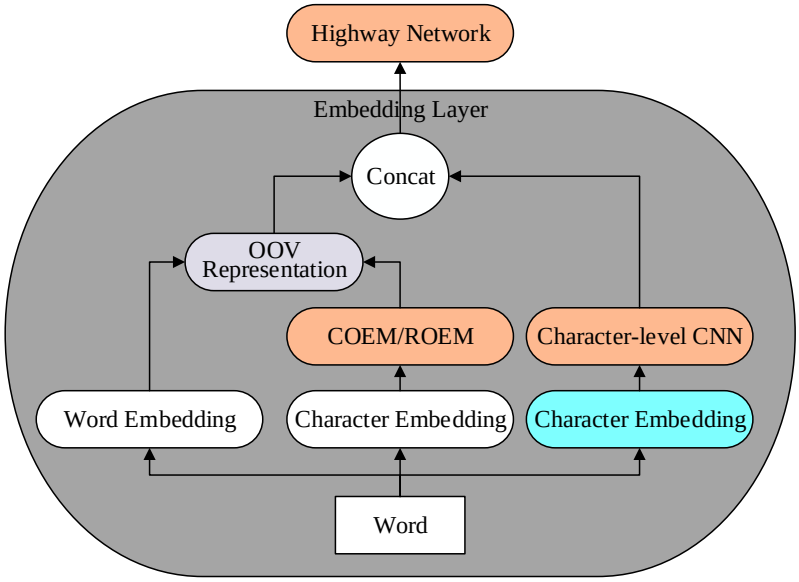


圖 6. QANet 中詞向量、字向量與未登錄詞向量之關係圖
[Figure 6. Relationship Among Word, Character and Out-of-Vocabulary Embeddings in the QANet]

表 4. 結合各式詞向量表示法模型於本論文提出之未登錄詞向量表示法模型在機器閱讀理解任務的成效(發展集)
[Table 4. Experimental Results on Development Set with Respect to Various Word Embedding Methods and the Proposed Framework]

Development Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
COEM	80.29%	67.19%	78.37%	64.74%	77.89%	63.71%
COEM+Char	80.60%	67.16%	79.91%	66.72%	**80.90%**	67.75%
WE+COEM	**80.82%**	67.94%	80.50%	66.94%	80.31%	67.22%
WE+COEM+Char	80.69%	**68.32%**	**81.28%**	**68.25%**	80.86%	**67.88%**
ROEM	75.44%	60.76%	73.48%	63.59%	74.25%	59.10%
ROEM+Char	79.90%	66.97%	79.89%	67.00%	79.49%	66.31%
WE+ROEM	79.00%	65.72%	78.70%	65.12%	77.78%	64.02%
WE+ROEM+Char	**80.16%**	**67.35%**	**80.82%**	**68.29%**	**81.06%**	**68.00%**

表 5. 結合各式詞向量表示法模型於本論文提出之未登錄詞向量表示法模型在機器閱讀理解任務的成效(測試集)
[Table 5. Experimental Results on Test Set with Respect to Various Word Embedding Methods and the Proposed Framework]

Development Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
COEM	79.86%	66.79%	77.83%	64.25%	77.36%	63.59%
COEM+Char	80.23%	66.80%	79.33%	66.17%	80.25%	67.13%
WE+COEM	80.12%	67.20%	80.01%	66.32%	79.84%	66.69%
WE+COEM+Char	**80.28%**	**67.93%**	**80.38%**	**67.34%**	**80.45%**	**67.40%**
ROEM	74.65%	59.96%	73.07%	58.46%	73.74%	58.83%
ROEM+Char	79.44%	66.52%	79.26%	66.43%	78.72%	65.66%
WE+ROEM	78.35%	65.07%	78.07%	64.47%	77.00%	63.14%
WE+ROEM+Char	**79.79%**	**66.91%**	**80.19%**	**67.57%**	**80.57%**	**67.76%**

首先，當基於卷積神經網路的未登錄詞模型(COEM)來取代所有詞向量與未登錄詞來進行訓練時，在每個詞向量表示模型上，都可以比基礎系統單用詞向量提升 7%至 8%之效果（請參考表 2 與表 3）；當我們採用基於循環神經網路的未登錄詞模型(ROEM)時，相較於基礎系統只能獲得 3%至 4%之效能提升（請參考表 2 與表 3），但其他實驗結果成效

皆不顯著，探究可能的原因是循環神經網路並不像卷積神經網路的未登錄詞模型，利用多種不同核大小，抽取字與字之間相鄰的特徵，而是直接一次性的在字與字之間進行雙向掃描，細緻的相鄰資訊較不容易保留。再來，我們進一步地討論將字向量表示法加入模型中（即 COEM+Char 與 ROEM+Char），在各種實驗結果上，皆可獲得一定程度之效能提升。至此，我們可以歸納出，不論是傳統的詞向量表示法模型或本論文提出之未登錄詞詞向量表示法模型，皆屬於學習全域性資訊的特徵，而 QANet 中的字向量是利用訓練閱讀理解模型時一併獲得，可看作是任務導向之特徵向量，因此當我們將這兩種資訊結合，通常可以進一步的提升任務成效。並且可以發現除了基於卷積神經網路的未登錄詞向量模型在連續型詞袋模型可以獲得相差不遠之效果，其餘甚至循環神經網路皆低於基礎模型詞向量加字向量之結果。我們認為這是因為我們將已有的詞向量也以未登錄詞來進行取代，並且模型的學習分布還不夠強健，導致雖然有字向量的幫助下，還不能贏過基礎系統詞向量加字向量之結果。接著，我們探討結合傳統詞向量表示法模型以及未登錄詞詞向量表示法模型（即 WE+COEM 與 WE+ROEM）相較於基礎系統之成效。實驗結果顯示，結合傳統詞向量表示法可獲得更進一步的效能提升，並且，基於卷積神經網路的未登錄詞模型在連續型詞袋模型、略詞模型與全局向量模型，都可以達到甚至超越基礎系統加上字向量之成果。值得一提的是，綜觀實驗結果，我們可歸納出，在多數情況下，基於卷積神經網路的未登錄詞模型會較基於循環神經網路的未登錄詞模型獲得較好的成效，推測其原因，應是由於在中文裡，每一個詞所欲表達的事、物或現象，經常與詞彙中字與字之間的排列順序有關，因此基於卷積神經網路的未登錄詞模型，長於擷取短距離的字與字的排列關係，並將此資訊用於生成未登錄詞的向量表示法，是較合理且有效的。當與強基礎系統（即 Strong Baseline 與 Strong Baseline+Char）進行比較，透過額外訓練字表示的文本資料為每一個未登錄詞產生詞向量之結果，基本上都與基礎系統詞向量加上字向量的效能差不多；甚至 F1 評估上，連續型詞袋模型反而得到更低之分數。最後，我們結合字向量表示法、傳統詞向量表示法模型以及未登錄詞詞向量表示法模型（即 WE+COEM+Char 與 WE+ROEM+Char）相較於基礎系統之結果，多數情況可再進一步獲得效能提升，並且超越基礎系統之最好結果，其中又以略詞模型最為突出。因此，我們可以歸納出，基於循環神經網路的未登錄詞詞向量模型，當使用全局向量模型為訓練目標時，可以獲得比基於卷積神經網路更好的任務成效；而略詞向量模型則是搭配基於卷積神經網路的未登錄詞向量模型可以獲得較好的任務成效。值得一提的是，本論文所採用的台達電閱讀理解資料集，未登錄詞在訓練集中的比例高達 65.50%，在發展集與測試集中則分別為 46.19%與 46.20%，而綜觀上述實驗結果，我們可以發現，未登錄詞的詞向量表示法確實會影響機器閱讀理解任務的成效，因此當利用本論文提出之方法時，可以大幅的改善任務的成效，甚至只需使用 COEM 或 ROEM 產生的詞向量表示法為輸入，就可以與傳統結合詞向量與自向量表示法為輸入的基礎系統擁有相近（甚至更好）的任務成效。

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

有鑑於自然語言處理中，未登錄詞一直是一個亟待解決的研究議題，本論文提出一套簡單又有效的未登錄詞詞量表示法模型，並藉由中文閱讀理解任務來驗證未登錄詞詞向量表示法模型之成效。實驗結果顯示，未登錄詞的詞向量表示法確實會影響任務的成效，因此，當使用本論文所提出的方法時，在未登錄詞的表達上更加合理與可靠，相較於基礎系統，可以獲得更好的結果。在未來，我們將繼續延伸本論文提出的未登錄詞詞向量表示法模型，並進一步地運用於英文的實驗語料中，也將探討再不同比例之未登錄詞比率之成效。再者，我們亦希望將未登錄詞詞向量表示法模型運用於其他任務之中。

參考文獻 (References)

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. In *Proceeding of OSDI'16 Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, 265-283.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. Retrived from arXiv:1607.06450.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*, 1137-1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. Retrieved from arXiv:1607.04606.

Chen, K.-Y., Wang, H.-M., & Chen, H.-H. (2015). A Probabilistic Framework for Chinese Spelling Check. *Transactions on Asian and Low-Resource Language Information Processing (Special Issue on Chinese Spell Checking)*, *14*(4), 15. doi: 10.1145/2826234

Chen, Z., Yang, R., Cao, B., Zhao, Z., Cai, D., & He, X. (2017). Smarnet: Teaching Machines to Read and Comprehend Like Human. Retrived from arXiv:1710.02772.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 770-778. doi: 10.1109/CVPR.2016.90

Hu, M., Peng, Y., Huang, Z., Qiu, X., Wei, F., & Zhou, M. (2017). Reinforced Mnemonic Reader for Machine Reading Comprehension. Retrived from arXiv:1705.02798

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-Aware Neural Language Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2741-2749.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, Retrieved from arXiv:1412.6980.

Liu, R., Wei, W., Mao, W., & Chikina, M. (2017). Phase conductor on multi-layered attentions for machine comprehension. Retrived from arXiv:1710.10504.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Proceedings of Advances in Neural Information Processing Systems*, 6294-6305.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in neural information processing systems*, 3111-3119.

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. In *Proceedings of Advances in neural information processing systems*, 1081-1088.

Pan, B., Li, H., Zhao, Z., Cao, B., Cai, D., & He, X. (2017). MEMEN: multi-layer embedding with memory networks for machine comprehension. Retrived from arXiv:1707.09098.

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.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of empirical methods in natural language processing (EMNLP)*, 2383-2392. doi: 10.18653/v1/D16-1264

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. Retrived from arXiv:1611.01603.

Shao, C. C., Liu, T., Lai, Y., Tseng, Y., & Tsai, S. (2018). DRCD: a Chinese Machine Reading Comprehension Dataset. Retrieved from arXiv:1806.00920.

Shen, Y., Huang, P. S., Gao, J., & Chen, W. (2017). Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1047-1055. doi: 10.1145/3097983.3098177

Sun, J. (2012). 'Jieba'Chinese word segmentation tool.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of Advances in neural information processing systems*, 3104-3112.

Tan, C., Wei, F., Yang, N., Du, B., Lv, W., & Zhou, M. (2017). S-net: From answer extraction to answer generation for machine reading comprehension. Retrived from arXiv:1706.04815.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, 5998-6008.

Wang, Y., Liu, K., Liu, J., He, W., Lyu, Y., Wu, H., ... & Wang, H. (2018). Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. Retrived from arXiv:1805.02220.

Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, *1*, 189-198. doi: 10.18653/v1/P17-1018

Weissenborn, D., Wiese, G., & Seiffe, L. (2017). Making neural QA as simple as possible but not simpler. Retrived from arXiv:1703.04816.

Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. Retrived from arXiv:1804.09541.

Zhang, J., Zhu, X., Chen, Q., Ling, Z., Dai, L., Wei, S., & Jiang, H. (2017). Exploring question representation and adaptation with neural networks. In *Proceedings of 2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 1975-1984. doi: 10.1109/CompComm.2017.8322883

Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 85-98

以深層類神經網路標記中文階層式多標籤語意概念

Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network

周瑋傑*、王逸如*

Wei-Chieh Chou and Yih-Ru Wang

摘要

傳統上對超過 100 個階層式標籤分類可以使用扁平 (flatten) 標籤做分類，但如此會喪失架構樹 (taxonomy) 的階層資訊。本研究旨在對廣義知網中文詞彙做概念分類與標記，提出考慮廣義知網架構樹階層關係之深層類神經網路訓練方法，其輸入為詞彙樣本點的詞向量，詞向量方面本研究亦提出考慮上下文前後關係之 2-Bag Word2Vec，而各階層的訓練結果有不同的重要性，所以在模型的最後使用最小分類誤差法，賦予各階層在測試階段時不同的權重。實驗結果顯示階層式 (hierarchical) 分類預測正確率會比扁平分類還高。

關鍵詞：詞向量、類神經網路、最小分類誤差、廣義知網、階層式分類、多標籤分類

Abstract

Traditionally, classifying over 100 hierarchical multi-labels could use flatten classification, but it will lose the taxonomy structure information. This paper aimed to classify the concept of word in E-HowNet and proposed a deep neural network training method with hierarchical relationship in E-HowNet taxonomy. The input of neural network is word embedding. About word embedding, this paper proposed order-aware 2-Bag Word2Vec. Experiment results shown hierarchical classification will achieved higher accuracy than flatten classification.

* 國立交通大學電機工程學系
Department of Electrical Engineering, National Chiao Tung University
E-mail: m0450743.eed04g@g2.nctu.edu.tw; yrwang@mail.nctu.edu.tw

Keywords: Word2Vec, Neural Network, Minimum Classification Error, E-HowNet, Hierarchical Classification, Multi-label Classification.

1. 緒論 (Introduction)

文字是資訊傳遞的重要媒介，其使用上豐富多變，中文語言中有所謂同義詞(synonym)，例如: 星期一與禮拜一，其為同義並且可相互替換字型的詞。另外也有部分詞彙為相同概念(concept)但其義不同的詞，例如: 戰鬥機與轟炸機，而這兩者都有飛行器以及戰鬥的概念，但其兩者意旨不同實體。我們希望可以在搜尋引擎中鍵入一個詞彙就可以搜尋到相似概念詞彙之搜尋結果，因此如何對中文詞彙進行語意概念標記將是本研究之重點。

在概念的範疇分析(ontology of concept)中，本研究使用廣義知網(Extended-HowNet, E-HowNet[1])的詞彙概念做為樣本點，廣義知網是 2003 年中央研究院資訊所詞庫小組將詞庫小組詞典(CKIP Chinese Lexical Knowledge Base)的詞條與董振東先生創建之知網(HowNet)的語意定義機制做連結、擴充以及修改(Huang, Chung & Chen, 2008)，以新的語義義原(sememe)通過義原的組合來標記各種單純或複雜的概念，以及各個概念與概念之間，概念的屬性與屬性之間的關係(Su, Li & Li, 2002) (Liu & Li, 2002)。

當一個樣本點對應到多個類別時可稱為多標籤(multi-label)，而一個樣本點對應到的類別數僅有一個時是為多類別(multi-class)。廣義知網中每一個詞彙樣本點皆會在架構樹(taxonomy)中對應至一個階層式標籤，其標籤為人工標記，圖 1 為某詞彙之概念分類為{mental|精神}在廣義知網上截至該節點之階層式關係圖，該詞彙之階層式多標籤資訊為物體 -> 萬物 -> 抽象物 -> 精神 -> null (null 為停止節點，原先架構並無該資訊，停止節點使用方式將在後續章節介紹)。

扁平(flatten)分類是不考慮架構樹階層式資訊，其在訓練上簡單容易，但資料本身的結構化標籤關係就未被考慮，故本研究以階層式(hierarchical)分類為基礎，提出考慮廣義知網架構樹上下位階層式標籤資訊並以類神經網路建構之階層式多標籤分類模型，網路之輸入為詞彙的詞向量，本研究另將 Mikolov 提出之 Word2Vec[2] (Mikolov, Chen, Corrado & Dean, 2013) (Mikolov, Sutskever, Chen, Corrado & Dean, 2013)模型做修改，提出考慮目標詞上下文前後關係之 2-Bag Word2Vec 架構，希望以不同的詞向量模型評測詞向量間的效能。若詞向量可以將語意及語法隱含在其中，那麼詞向量也可以運用在詞彙的概念分類上。

[1] http://ehownet.iis.sinica.edu.tw/index.php

[2] https://code.google.com/archive/p/word2vec/

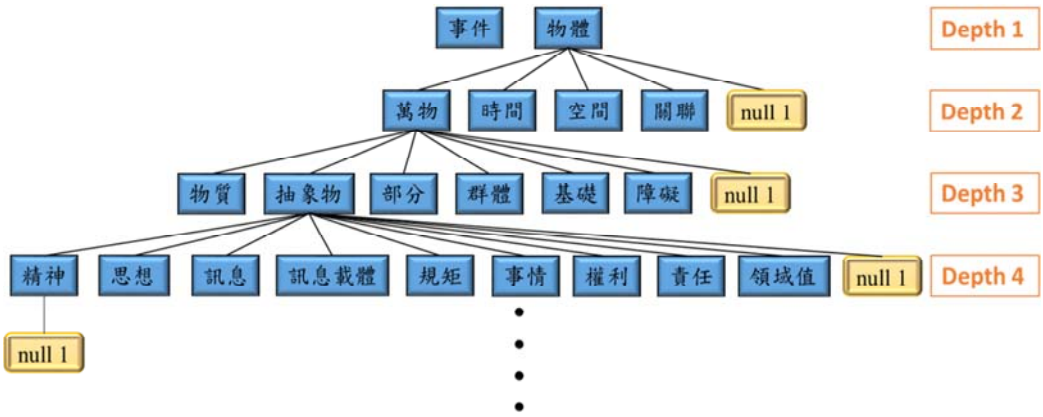


圖 1. 廣義知網樹狀階層式架構 — 截至{mental\|精神}之節點為
[Figure 1. Taxonomy of E-Hownet --- take {mental\|精神} as example]

2. 詞向量模型 (Word Vector Model)

本章節我們將會介紹本研究使用的詞向量模型，此處將介紹 Mikolov 所提出的 Word2Vec 模型，其中包含 Continuous Bag-Of-Words (CBOW)和 Skip-gram，本研究亦提出 2-Bag Word2Vec，後者與前者不同的是後者為考慮目標詞前後文關係之詞向量模型。

在 CBOW (圖 2 左側) 和 Skip-gram (圖 2 右側) 的架構中，投影層至輸出層的預測矩陣(prediction matrix) 皆相同為 $M' \in \mathbb{R}^{(|V|) \times d}$，在不同位置的詞使用共同的預測矩陣，也就是說這兩種模型在學習上並沒有加入詞在句子中的位置資訊，其會導致詞向量訓練結果不包含相鄰詞的順序關係。

本研究將對 Mikolov 所提出的 Word2Vec 模型修改，提出 2-Bag Word2Vec，該模型下包含 2-Bag CBOW 以及 2-Bag skip-gram。以圖 3 左側之 2-Bag CBOW 為例，2-Bag CBOW 中輸入為 (2 × d) 維的向量，其中投影層 (projection layer) 加大，目標詞之前與之後的投影層分開成兩個輸入相鄰詞[$e(W_{-c},\ldots, W_{-1})$, $e(W_1, , \ldots, W_c)$]，投影矩陣變為 $M' \in \mathbb{R}^{(|V|) \times 2d}$，意即加大投影矩陣分開保留字詞前後的關係。圖 3 右側之 2-Bag skip-gram 則使用兩個預測矩陣M'_a、M'_b，大小各為$M' \in \mathbb{R}^{(|V|) \times d}$，也是考慮了相鄰詞的前後順序。

僅考慮目標詞的前後關係而非完整考慮目標詞之順序的原因在於，系統參數量(weight) 增加，M′的大小變為兩倍，訓練語料要夠大才能得到好處，故沒有考慮各個輸入詞彙的順序，以觀察在系統參數量較少的情況下，詞向量訓練的成效如何。

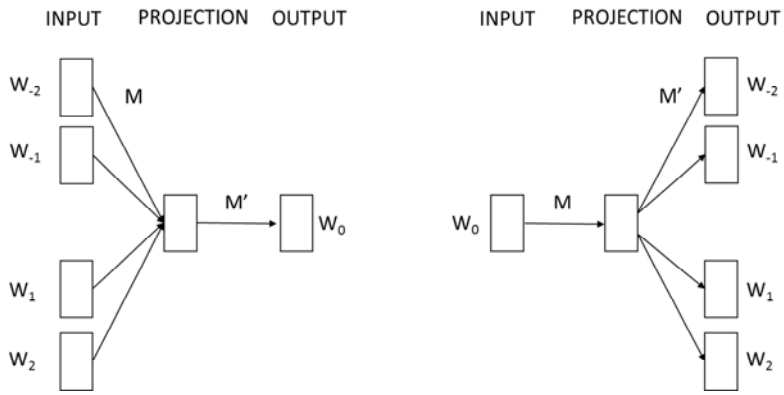


圖 2. Word2Vec 模型，左側為 CBOW，右側為 Skip-gram
[Figure 2. Word2Vec model. The left side of the figure is CBOW. The right side of the figure is Skip-gram.]

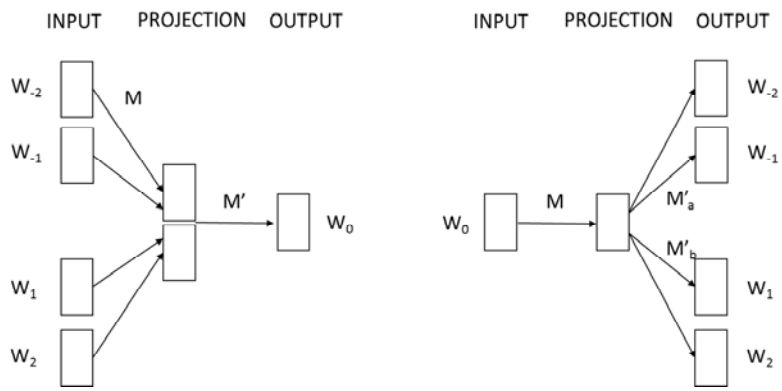


圖 3.本研究提出之 2-Bag 模型，左側為 2-Bag CBOW，右側為 2-Bag Skip-gram
[Figure 3. 2-Bag model proposed by this research. The left side of the figure is 2-Bag CBOW. The right side of the figure is 2-Bag Skip-gram

3. 以類神經網路建構之階層式多標籤分類 (Hierarchical Multi-label Classification using Neural Network

3.1 模型 (Model)

相比扁平分類，階層式分類為考慮資料標籤之階層式架構。階層式多標籤分類模型以架構樹之階層式標籤使用多個類神經網路建立，使用類神經網路可讓模型建構上變得較有彈性，圖 4 為本研究所提出之階層式模型架構，在不同深度 (depth) 之各階層分類上皆使用一個類神經網路，而各階層皆有一個輸出，各輸出分別對應至各階層的標籤，其為一個多輸出 (Multi-Output) 模型。階層式標籤可能有長有短，在各階層之類神經網路輸出層中加入一個停止節點 (圖 1 的 null 標籤) ，所以一個長度較短的標籤從模型由上而

下 (top-down) 遇到停止節點後，其後在輸出層皆為停止標籤 (null) 直到模型的最底。

訓練階段一開始，一個神經網路負責架構樹資料的第一層資訊 (depth 1，較靠近 root 節點的階層)，此網路有一個隱藏層以及一個輸出層 (Depth 1 的類別)，網路權重更新的方式為倒傳遞演算法 (Back-propagation)，每一個神經網路僅負責預測一個架構樹階層中的類別。當第一階層的神經網路訓練完畢後 (圖 4 上方對應到架構樹第一階層)，第二階層會另外有一個類神經網路負責訓練，差別在於第一階層神經網路的輸出與整個網路一開始的輸入特徵相接 (concatenate) 後做為第二階層神經網路的輸入 (圖 4 下方第二階層之輸入)，此訓練程序會不斷重複直到最後一階層之神經網路被訓練，如此將各階層神經網路相接形成一個深層神經網路。

測試階段將測試資料餵入第一個神經網路 (第一階層之神經網路)，而後第一階層的輸出做為第二階層的輸入，此過程將不斷重複直到抵達最後一階層，在各階層神經網路的輸出後使用 softmax，最後執行下一節要介紹的修正矛盾現象。

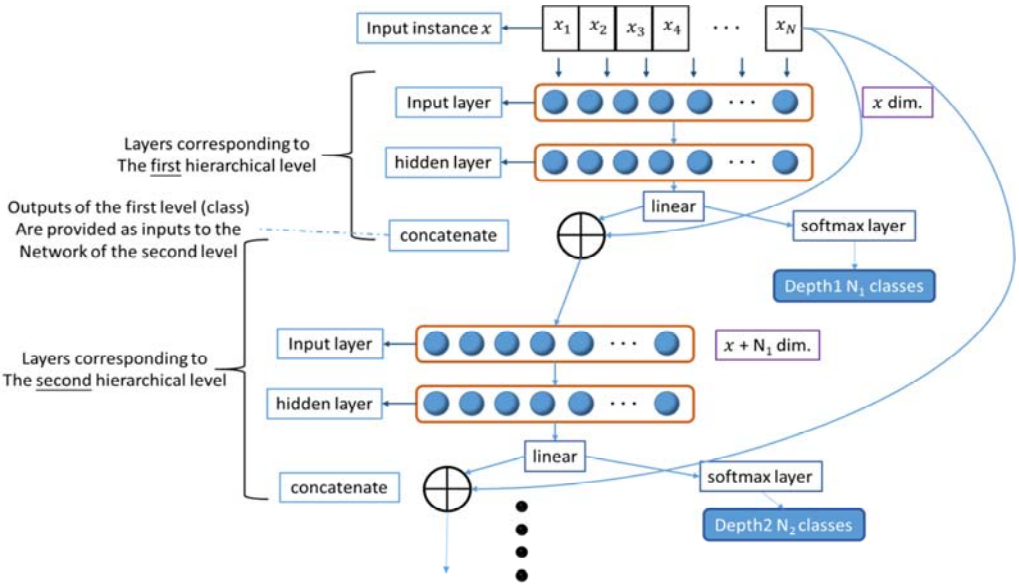


圖 4.以深層類神經網路架構之階層式多標籤分類模型
[Figure 4. The structure of Hierarchical Multi-Label Classification using neural network]

3.2 階層式標籤矛盾現象 (Correcting Inconsistencies)

當階層式多標籤類神經網路完成其訓練之後，在測試階段，還需做一個後處理來校正上下位標籤關係矛盾情況，例如: 一個下位階層的節點 (subclass) 被預測，但其所屬上位節點 (superclass) 卻沒有。這些矛盾情況的發生原因在於每個階層的神經網路都是各別輸出結果的，也就是說在測試階段，一個詞彙的預測結果是與其他階層的結果沒有相依性的，所以會造成上下位階層前後矛盾 (inconsistent)，意即出現架構樹上不存在的階層式標籤資訊，使用後處理可保證在最後階段不會存在矛盾情況。

$$p(X) = \sum_{L=1}^{N} \log\left(\sigma(X_{Lk})\right) \quad (1)$$

針對各階層內的多類別分類上本研究在輸出層後使用 Softmax，式(1)為考慮架構樹路徑資訊後對單筆路徑計算 softmax 機率和的方法，其中 L 為該階層架構下的某一層，k 為該階層中某一分類的機率，在此依照資料集之架構樹的資訊對架構樹的多標籤路徑資訊做加總機率和後的結果稱為分數，最後總共有 N 個分 數 (架構樹共有 N 條路徑)，再依照分數的大小作排序，求其正確率。

3.3 效能評估方式 (Measurement)

實驗所使用的評估方式為正確率 (accuracy)，其計算為標記正確的詞彙數與總詞數的比值，而在考慮架構樹後的多標籤深層類神經網路中我們將探討考慮架構樹階層時的正確率，故我們將正確率再細分為總體正確率與各層正確率，分別如下:

總體正確率: 該層類別是否正確需考慮上層所有測試路徑是否正確

各層正確率: 僅考慮該層類別是否正確

表 1. word α 正確率計算示意
[Table 1. word α test path]

	type	Depth 1	Depth 2	Depth 3	Depth 4	Depth 5
word α	TEST	classA	classF	classC	null	null
	TRUTH	classA	classB	classC	null	null

以表 1 為例，如 word α 的 TRUTH 所示，word α 落在架構樹的節點 C，在測試階段 word α 在架構樹的預測結果也落在節點 C，但其預測之多標籤路徑 A -> F -> C (圖 5 橘色路徑，正確路徑為綠色) 為一個不存在的路徑，計算正確率時以 Depth 3 為例，因為 word α 在 Depth 2 就錯誤，所以在計算總體正確率時僅有 Depth 1 為正確，而 Depth 2 和 Depth 3 皆為錯誤 ; 而在各層正確率方面 Depth 1、3 都計為正確，而 Depth 2 計為錯誤。

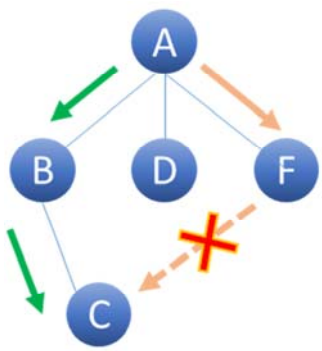


圖 5. word α之預測與實際路徑
[Figure 5. word α test and truth path]

4. 實驗與分析 (Experiment and Analysis)

4.1 實驗資料與實驗設定 (Experiment Setup)

在建立詞向量方面使用的語料包含 (1) LDC Chinese Gigaword Second Edition[3]、(2) Sinica Balanced Corpus ver. 4.0、(3) CIRB0303[4] (Chinese Information Retrieval Benchmark, version 3.03)、(4) Taiwan Panorama Magazine[5]、(5) TCC300[6] 和 (6) Wikipedia (ZH_TW version)，以上各語料庫經交通大學王逸如老師開發之斷詞器(Wang, Wu, Liao & Chang, 2013)斷詞後共約 4.4 億詞。為了使詞向量更精確，在建立詞向量之前會先經過文字正規化、同義詞替換、少數人名和數字合併。訓練詞向量時，目標詞前後窗口為 7 (c = 7)，維度為 200 維，詞頻篩選剔除小於 25 的詞彙，最後共建立出 19 萬個詞向量。

廣義知網詞彙概念樣本點方面，廣義知網未對所有的詞彙標上概念展開式，例如「新歌」在詞向量有被建立，但是廣義知網卻未標上其概念展開式，扣除未標上概念展開式以及詞向量未被建立的詞彙，最後約使用 5 萬筆資料當作概念樣本點，其中 90%為訓練資料集，其餘 10%為測試資料集。

某些概念的詞彙樣本點過少，故本研究也將廣義知網的少數概念類別向上一層合併，合併完成後為 403 個概念類別。

在未進行概念類別合併之前，從廣義知網的詞彙統計結果發現，{人}的類別所佔的數目最多佔了 3404 個，而第二名的{事物}為 1511 個 ；如果看到最少的資料，其中數量小於 10 的類別有 521 個，小於 5 的有 321 個，小於 1 的有 107 個。因此針對不平衡資料做處理，此處使用隨機資料增生，將所有個別數量小於 30 的類別增生至 30 類，如此原先資料從 50924 筆增加至 52906 (+1982) (+3.9%)

4.2 扁平分類 (Flatten Classification)

扁平分類為不考慮廣義知網架構樹上下位階層關係之分類方法，此處將使用最近鄰居法 (K-Nearest Neighbors Algorithm, KNN) 與類神經網路(Neural Network)。

在進行最近鄰居法時，必須先決定 K 值，然而最佳 K 值取決於樣本點等等因素，隨著樣本點的改變其值亦會改變，雖然我們可以透過一些最佳化演算法去求得最佳的 K 值，但我們未採用任何演算法求最佳 K 值，而是希望透過實驗結果決定。如果在投票時平票 (tied-vote)，則考慮相似度最大者為第一名 (Top 1)，如不依照相似度排序則為第一名最大值 (Top 1 MAX)。此處最近鄰居法使用餘弦距離 (cosine distance)。

---

[3] https://catalog.ldc.upenn.edu/LDC2005T14

[4] http://www.aclclp.org.tw/use_cir.php

[5] https://www.taiwan-panorama.com/en

[6] http://www.aclclp.org.tw/use_mat.php#tcc300edu

表 2. 最近鄰居法標記正確率
[Table 2. Accuracy of KNN]

	CBOW200	Skip-gram200	2Bag-CBOW200	2Bag-Skip-gram200
Top 1	**55.8%**	53%	55.2%	52.7%
Top 1 Max	63.9%	62%	62.6%	61.1%
Top 3	74.4%	73.4%	74.3%	72.5%
Top 3 Max	80.2%	79.4%	80.1%	78.7%
K value	8	10	8	9

由表 2 發現 Mikolov 所提出的 Word2Vec 模型之效能比本研究所提出之 2-Bag 模型要來得好，而採用以鄰近詞預測目標詞的 CBOW 和 2-Bag CBOW 正確率皆較高。

在最近鄰居法方面的實驗結果驗證在詞向量模型方面 Word2Vec 的 CBOW 和 Skip-gram 有較好的效能，故在類神經網路方面選用 Word2Vec 的詞向量模型架構做後續詞彙標記模型的輸入。

表 3. 類神經網路法標記正確率
[Table 3. Accuracy of Neural Network]

word vector	Top1 Acc.	Top1 Acc. + POS	Top3 Acc. + POS
CBOW 200	54.1%	58.4%	75.3%
Skip-gram 200	53.7%	56.7%	74.6%
CBOW 200 + Skip-gram 200	**55.4%**	**59.5%**	**76.4%**

此處探討加入 POS (part of speech) 後對正確率的影響，POS 經 one-hot encoding 為 11 維向量，該 POS 向量與 CBOW 或 Skip-gram 向量相接後各為 211 維向量，而 CBOW 200 + Skip-gram 200 + POS 為 411 維向量，由表 3 得知，加入 POS 資訊後在各詞向量模型上可以提高正確率，而其中詞向量又以 CBOW 200 + Skip-gram 200 + POS 的正確率最佳，儘管 CBOW 200 + Skip-gram 200 中兩個詞向量相接上並未做任何優化處理，但也得到了較好的效果，使用 CBOW 200 + Skip-gram 200 + POS 的特徵達到了 59.5%的正確率。

4.3 階層式多標籤分類 (Hierarchical Multi-Label Classification)

從扁平式分類章節發現 CBOW + Skip-gram + POS 之 411 維特徵之正確率表現較好，故階層式多標籤分類模型的輸入也使用相同設定，其中 CBOW 和 Skip-gram 各 200 維，另外加入 POS one-hot 之 11 維資訊。

在訓練階段，將各詞彙抽取 head concept 後，依照架構樹的資訊從樹葉節點 (leaf node) 往 root 節點尋找每個詞彙的路徑資料，各標籤的路徑長短不一，資料未達到 N 層的情況下將在下位階層的節點補上 null。

表 4. 階層式多標籤分類模型之測試階段正確率 (更正標籤矛盾情況)
[Table 4. Accuracy of Hierarchical Multi-Label Classification (Correcting inconsistencies)

	Overall Accuracy		Layer Accuracy		
Depth N	**Acc. Top1**	**Acc. Top3**	**Acc. Top1**	**Acc. Top3**	**類別數**
Depth 1	98.3%	99.1%	98.3%	99.1%	3
Depth 2	90.7%	95.0%	90.7%	95.0%	7
Depth 3	81.9%	89.8%	81.9%	89.8%	18
Depth 4	77.1%	87.0%	77.3%	87.2%	83
Depth 5	70.9%	83.1%	75.1%	86.3%	85
Depth 6	65.8%	79.4%	72.4%	85.2%	127
Depth 7	63.4%	77.7%	79.8%	88.9%	78
Depth 8	62.2%	76.9%	87.1%	92.9%	23
Depth 9	61.7%	76.5%	91.8%	95.6%	31
Depth 10	61.3%	76.2%	95.6%	97.6%	38
Depth 11	**61.0%**	76.1%	98.2%	99.1%	10

在測試階段採用 3.2 節更正標籤矛盾情況方法，測試階段正確率顯示於表 4 中，觀察正確率發現在 Depth 1 到 Depth 2 時僅僅是增加 4 個類別 (3 類->7 類)，總體正確率就下降了將近 8%。

4.3.1 賦予各階層不同權重 (Given Different Layer Different Weight)

在階層式多標籤類神經網路測試結果中發現某些階層的正確率偏低，由表 5 中觀察到 Depth 4 到 Depth 7 之正確率都不足 80%，低於其他階層的正確率。可能原因為該層類別數較多，導致該層分類困難。此處可以給予每一層不同的 weight 來較相信或較不相信才自某幾層的資訊，幫助辨認結果。

在此可以將各階層的權重視為一個參數集來最大化總體正確率，在此應用最小分類誤差法，以分類的方式決定每一個階層的權重。最小分類誤差之決策式如下:

$$d_k(X) = -g_i(X;w) + \log\left[\frac{1}{M-1}\sum_{j,j\neq c} exp\big(g_j(X;w)\eta\big)\right]^{1/\eta} \quad (2)$$

其中η為一個正整數，$d_k(X) > 0$時為分類錯誤，而$d_k(X) \leq 0$為分類正確， logarithm 與 exponential 互為反函數，其目的為避免方程式$g_j(X;w)$進行η次方時產生計算機 underflow 問題。

此處可以看成是正確類別和所有錯誤類別的競爭學習過程，當 η 趨近∞ 時，(2)中

括弧內的方程式會變為$\max_{j,j\neq i} g_j(X;w)$，意即僅找所有錯誤類別中錯誤分數最大的結果來訓練以加快訓練速度，如此(2)會變為:

$$d_k(X) = -g_i(X;w) + g_j(X;w) \quad (3)$$

(2)中$g_j(X;w)$為類別條件似然函數(class conditional likelihood functions)，可將 softmax 後結果 X 視為已知，而欲找到一組最佳權重 w 來最大化似然函數，(2)帶入各階層 softmax 後結果以及 weight，可將方程式寫為:

$$d_k(X) = -\sum_{L=1}^{11} X_{LC} W_L + \log\left[\frac{1}{M-1}\sum_{j,j\neq c} exp\left(\left(\sum_{L=1}^{11} X_{Lj} W_L\right)\eta\right)\right]^{1/\eta} \quad (4)$$

其中C ∈廣義知網架構樹的正確路徑，該方法將所有錯誤以及正確路徑做競爭學習。

而在尋找最佳的參數叢集 X 時，也可將個別的錯誤資料是其重要程度，改變 η 和 M 將所有錯誤競爭類別加入考慮，而此處把(3)嵌入 zero-one function，定義損失函數 (loss function)，此處以 sigmoid function (5)當作考量。

$$\ell(d) = \frac{1}{1+exp(-\gamma d+\theta)} \quad (5)$$

損失函數的 θ 通常是 0，$\gamma \geq 0$，改變 θ 和 γ 可改變 loss 被調整的範圍。

表 5. 各別階層正確率中，部分階層 (框選處) 正確率較低
[Table 5. Accuracy of each layer. Accuracy of some layer (circled) are lower than the others.]

	Accuracy of each layer	
Depth N	**Accuracy**	**Classes per layer**
Depth 1	98.3%	3
Depth 2	90.7%	7
Depth 3	81.9%	18
Depth 4	77.3%	83
Depth 5	75.1%	85
Depth 6	72.4%	127
Depth 7	79.8%	78
Depth 8	87.1%	23
Depth 9	91.8%	31
Depth 10	95.6%	38
Depth 11	98.2%	10

在 Minimum Classification Error 訓練階段以梯度下降法做參數的更新:

$$W_L(t+1) = W_L(t) - \varepsilon \frac{\partial \ell(X;W)}{\partial W_L} \tag{6}$$

其中 ε 為學習率，而偏微分項次可以透過鏈鎖律(chain rule)求得

$$\frac{\partial \ell(X;W)}{\partial W_L} = \frac{\partial \ell(X;W)}{\partial d}\frac{\partial d}{\partial W_L} \tag{7}$$

其中 $\frac{\partial \ell(X;W)}{\partial \mathrm{d}}$ 為對 sigmoid 微分的結果，可寫為:

$$\frac{\partial \ell(X;W)}{\partial \mathrm{d}} = \gamma \ell(d)\big(1-\ell(d)\big) \tag{8}$$

其中$\frac{\partial d}{\partial W_L}$

$$= -X_{LC} + \frac{1}{\eta}\Big[\frac{1}{M-1}\sum_{j,j\neq c} exp\big[\big(\sum_{L=1}^{11} X_{Lj}W_L\big)\eta\big]\Big]^{-1}\frac{\partial d}{\partial W_L}\Big[\frac{1}{M-1}\sum_{j,j\neq c} exp\big[\big(\sum_{L=1}^{11} X_{Lj}W_L\big)\eta\big]\Big]$$

$$= -X_{LC} + \frac{1}{\eta}\Big[\frac{1}{M-1}\sum_{j,j\neq c} exp\big[\big(\sum_{L=1}^{11} X_{Lj}W_L\big)\eta\big]\Big]^{-1}\Big[\frac{1}{M-1}\sum_{j,j\neq c} \eta X_{Lj}\, exp\big[\big(\sum_{L=1}^{11} X_{Lj}W_L\big)\eta\big]\Big]$$

$$= -X_{LC} + \frac{\sum_{j,j\neq c}\big[exp\big[\big(\sum_{L=1}^{11} X_{Lj}W_L\big)\eta\big]X_j\big]}{\sum_{j,j\neq c} exp\big[\big(\sum_{L=1}^{11} X_{Lj}W_L\big)\eta\big]}$$

訓練完成的各階層權重如圖 6 所示，觀察圖表發現最小分類誤差法訓練出的權重在第八層時為最小，而在階層中正確率最低者為第六層，與預期上權重最低應該在第六層有所不符，但權重分布與階層正確率一樣為山谷型走勢。

原先未使用最小分類誤差法時各階層的權重皆為1。在最小分類誤差訓練法完成後，在測試階段計算各路徑分數時可以考慮各階層的權重的不同再進行 3.2 更正矛盾情況時計入不同的 weight，此處將(1)稍作修改如下所示:

$$\mathrm{p}(X) = \sum_{L=1}^{N} W_L \log\big(\sigma(X_{Lk})\big) \tag{9}$$

W_L為第 L 層之權重，此處在計算各階層分數並加總時考慮權重W_L。

計入權重後的測試階段預測結果如表 6 所示，在加入 weight 後其總體正確率之第 11 層較未修正 weight 的情況多了 0.3%正確率，為 61.3%，而在 Top 3 情況其總體正確率上升 0.8%。

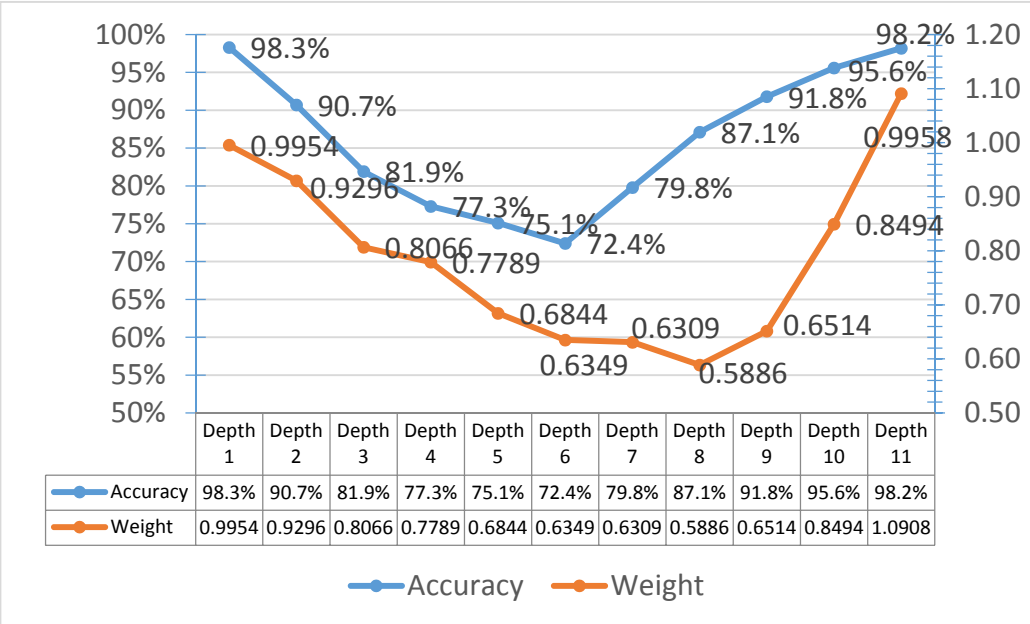


圖 6. 各階層正確率和 weight 走勢
[Figure 6. Weight and accuracy of each layer]

表 6. 調適各階層權重後的正確率
[Table 6. Accuracy of each layer after adjusting the weight]

	Accuracy Top1		Accuracy Top3		
Depth N	**original**	**add weight**	**original**	**add weight**	**classes**
Depth 1	98.3%	98.4%	99.1%	99.2%	3
Depth 2	90.7%	90.8%	95.0%	95.6%	7
Depth 3	81.9%	81.9%	89.8%	90.5%	18
Depth 4	77.1%	77.1%	87.0%	87.6%	83
Depth 5	70.9%	70.9%	83.1%	84.1%	85
Depth 6	65.8%	65.8%	79.4%	80.3%	127
Depth 7	63.4%	63.5%	77.7%	78.4%	78
Depth 8	62.2%	62.3%	76.9%	77.7%	23
Depth 9	61.7%	61.8%	76.5%	77.2%	31
Depth 10	61.3%	61.5%	76.2%	76.9%	38
Depth 11	61.0%	61.3%	76.1%	76.9%	10

4.4 標記結果討論 (Discussion)

由於詞向量是根據前後文相鄰的關係所訓練的，所以兩個同義的詞彙在用法上類似其餘弦相似度就會高，但也可能會有兩詞彙互為反義但其兩者因前後文類似所以其餘弦相似度高造成標記錯誤的情況發生，例如「瘦」與「胖」。

自動標記之第一名類別和廣義知網所標記的類別有所不同，例如: 戰車_N 在廣義知網中的類別為{車}，而自動標記第一名為{武器}，但在常識上似乎也不能直接算是錯誤類別，此處參考教育部線上辭典或維基百科對於戰車的解釋: 戰車 1. 作戰用的車輛 2. 全裝甲結構之全履帶車輛，有砲塔、自動武器、通信等裝置。

但有時廣義知網內的類別也並非完全客觀，看到下列例子: 魚丸_N 在廣義知網中的類別為{身體部件}，而自動標記第一名為{食品}，而要判斷標記之類別是否正確，不妨先了解其真實的語意再做判斷，以下列出依照教育部線上辭典或維基百科對於此詞彙的解釋：魚丸，魚肉調蛋清製成的丸子。此例子中自動標記結果比起廣義知網更貼近真實詞義，故自動標記之類別有時反而比廣義知網所標記的類別更加接近真正的語意，但這些標記錯誤的詞彙最後仍須人工檢查，且其為錯誤或正確之標記常常是見仁見智。

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

本研究提出考慮架構樹 (taxonomy) 之階層式多標籤資訊後以類神經網路建立的階層式多標籤分類模型，其應用於廣義知網的詞彙語意概念標記，神經網路的輸入為詞向量，在詞向量方面本研究亦提出 2-Bag model，其為將詞向量之投影層至輸出層的 weight 數量增加且考慮目標詞前後關係之模型，唯因系統參數量增加的情況下，訓練語料 4.4 億詞過少(Wang, Dyer, Black & Trancoso, 2015)，因而無法有效地訓練 2-Bag model。

實驗階段比較了階層式 (hierarchical) 與扁平式 (flatten) 分類，其兩者同樣以類神經網路建立分類模型，不同的是階層式架構之類神經網路是深層且階層的，扁平式分類的輸出層節點數與資料集中的類別數量相同。從實驗結果來看，階層式分類之正確率會比扁平分類還高。而在測試階段也採用最小誤差分類法 (minimum classification error)，讓機器自行學習各階層的重要性，賦予不同層權重，改善最後測試階段之正確率。

本研究輸入詞彙有消歧義情況，未來可以在本研究中加入中文消歧模型，增加模型辨認率。另外階層式多標籤分類法也可應用在其他同樣具有階層式多標籤資訊的資料庫，例如檔案目錄系統、生物資訊系統。

致謝 (Acknowledgements)

This work was supported by the Ministry of Science and Technology, Taiwan with contract MOST-105-2221-E-009-142-MY2.

參考文獻 (References)

Huang, S.-L., Chung, Y.-S., & Chen, K.-J. (2008). E-HowNet: the expansion of HowNet. In *Proceedings of the First National HowNet Workshop*, 10-22.

Liu, Q. & Li, S.-j. (2002). Word Similarity Computing Based on How-net. *International Journal of Computational Linguistics and Chinese Language Processing*, *7*(2), 59-76. [In Chinese]

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*, Retrived from arXiv:1301.3781v1

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS 2013*, 3111-3119.

Su, W.-F., Li, S.-Z., & Li, T.-Q. (2002). A Module of Automatic Chinese Documents Classification Based on Concept. *Computer Engineering and Applications*, *2002*(6).

Wang, L., Dyer, C., Black, A. W., & Trancoso, I. (2015). Two/Too Simple Adaptations of word2vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299-1304. doi: 10.3115/v1/N15-1142

Wang, Y.-R., Wu, Y.-K., Liao, Y.-F., & Chang, L.-C. (2013). Conditional random field-based parser and language model for traditional Chinese spelling checker. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, 69-73.

The individuals listed below are reviewers of this journal during the year of 2018. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Guo-Wei Bian
Yung-Chun Chang
Tao-Hsing Chang
Yu-Yun Chang
Fei Chen
Alvin C.-H. Chen
Chien Chin Chen
Kuan-Yu Chen
Yun-Nung (Vivian) Chen
Tai-Shih Chi
Chen-Yu CHIANG
Wen-Lih Chuang
Hung-Yan Gu
Wei-Tyng Hong
Shu-Kai Hsieh
Hen-Hsen Huang
Jen-Wei Huang
Yi-Chin Huang
Jeih-Weih Hung
Hsin-Te Hwang
Wen-Hsing Lai
Ying-Hui Lai
Chi-Chun Lee
Hong-Yi Lee
I-Bin Liao
Bor-Shen Lin
Shu-Yen Lin
Chao-Hong Liu
Chih-Hua Tai
Wei-Ho Tsai
Chin-Chin Tseng
Jenq-Haur Wang
Jiun-Shiung Wu
Cheng-Zen Yang
Jui-Feng Yeh
Ming-Shing Yu

2018 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 23

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2017

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of paper or other item, and its location, specified by the publication volume, number, and inclusive pages. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication volume, number, and inclusive pages.

AUTHOR INDEX

C

Chang, Hsiu-Jui
see Chao, Wei-Cheng, 23(2): 35-46

Chao, Wei-Cheng
Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method; 23(2): 35-46

Chen, Berlin
see Lo, Tien-Hong, 23(2): 19-34
see Chao, Wei-Cheng, 23(2): 35-46

Chen, Chia-Ping
Sentiment Analysis on Social Network: Using Emoticon Characteristic for Twitter Polarity Classification; 23(1): 1-18

Chen, Chin-Po
see Liu, Yu-Shuo, 23(1): 19-34

Chen, Kuan-Yu
see Luo, Shang-Bao, 23(2): 67-84

Chou, Wei-Chieh
Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network; 23(2): 85-98

G

Gau, Susan Shur-Fen
see Liu, Yu-Shuo, 23(1): 19-34

H

Hsieh, Wan-Ting
Joint Modeling of Individual Neural Responses using a Deep Voting Fusion Network for Automatic Emotion Perception Decoding; 23(1): 35-48

L

Lai, Chien-hung
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; 23(2): 1-18

Lee, Chi-Chun
see Liu, Yu-Shuo, 23(1): 19-34
see Hsieh, Wan-Ting, 23(1): 35-48

Lee, Ching-Hsien
see Luo, Shang-Bao, 23(2): 67-84

Lin, Yi-Chung
see Wu, Meng-Tse, 23(2): 47-66

Liu, Yu-Shuo
A Lexical Coherence Representation Computational Framework using LSTM Forget Gate For Autism Recognition; 23(1): 19-34

Lo, Tien-Hong
Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition; 23(2): 19-34
see Chao, Wei-Cheng, 23(2): 35-46

Luo, Shang-Bao
An OOV Word Embedding Framework for Chinese Machine Reading Comprehension; 23(2): 67-84

S

Su, Keh-Yih
see Wu, Meng-Tse, 23(2): 47-66

T

Tseng, Tzu-Hsuan
see Chen, Chia-Ping, 23(1): 1-18

Tu, Jia-Jang
see Luo, Shang-Bao, 23(2): 67-84

W

Wang, Yih-Ru
see Lai, Chien-hung, 23(2): 1-18
see Chou, Wei-Chieh, 23(2): 85-98

Wu, Meng-Tse
Supporting Evidence Retrieval for Answering Yes/No Questions; 23(2): 47-66

Y

Yang, Tzu-Hsuan
see Chen, Chia-Ping, 23(1): 1-18

SUBJECT INDEX

A

Acoustic Model
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18
Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method; Chao, W.-C., 23(2): 35-46

Autism Spectrum Disorder
A Lexical Coherence Representation Computational Framework using LSTM Forget Gate For Autism Recognition; Liu, Y.-S., 23(1): 19-34

Automatic Speech Recognition
Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition; Lo, T.-H., 23(2): 19-34

B

Backstitch
Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method; Chao, W.-C., 23(2): 35-46

Behavioral Signal Processing
A Lexical Coherence Representation Computational Framework using LSTM Forget Gate For Autism Recognition; Liu, Y.-S., 23(1): 19-34

C

CNNs
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18

D

Deep Voting Fusion Neural Net
Joint Modeling of Individual Neural Responses using a Deep Voting Fusion Network for Automatic Emotion Perception Decoding; Hsieh, W.-T., 23(1): 35-48

Discriminative Training
Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition; Lo, T.-H., 23(2): 19-34

Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method; Chao, W.-C., 23(2): 35-46

DNNs
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18

E

E-HowNet
Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network; Chou, W.-C., 23(2): 85-98

F

fMRI
Joint Modeling of Individual Neural Responses using a Deep Voting Fusion Network for Automatic Emotion Perception Decoding; Hsieh, W.-T., 23(1): 35-48

G

Gradient Vanishing (Exploding)
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18

H

Hierarchical Classification
Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network; Chou, W.-C., 23(2): 85-98

I

Individual Difference
Joint Modeling of Individual Neural Responses using a Deep Voting Fusion Network for Automatic Emotion Perception Decoding; Hsieh, W.-T., 23(1): 35-48

L

Lexical Coherence Representation
A Lexical Coherence Representation Computational Framework using LSTM Forget Gate For Autism Recognition; Liu, Y.-S., 23(1): 19-34

LF-MMI
Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition; Lo, T.-H., 23(2): 19-34

LSTM
A Lexical Coherence Representation Computational Framework using LSTM Forget Gate For Autism Recognition; Liu, Y.-S., 23(1): 19-34

LSTMs
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18

LVCSR
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18

M

Machine Learning
Sentiment Analysis on Social Network: Using Emoticon Characteristic for Twitter Polarity Classification; Chen, C.-P., 23(1): 1-18

Machine Reading Comprehension
An OOV Word Embedding Framework for Chinese Machine Reading Comprehension; Luo, S.-B., 23(2): 67-84

Mandarin
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18

Mandarin Large Vocabulary Continuous Speech Recognition
Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method; Chao, W.-C., 23(2): 35-46

Matrix Factorization
Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method; Chao, W.-C., 23(2): 35-46

Minimum Classification Error
Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network; Chou, W.-C., 23(2): 85-98

Model Combination
Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition; Lo, T.-H., 23(2): 19-34

Multi-label Classification
Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network; Chou, W.-C., 23(2): 85-98

N

Natural Language Processing
An OOV Word Embedding Framework for Chinese Machine Reading Comprehension; Luo, S.-B., 23(2): 67-84

Neural Network
Sentiment Analysis on Social Network: Using Emoticon Characteristic for Twitter Polarity Classification; Chen, C.-P., 23(1): 1-18

Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network; Chou, W.-C., 23(2): 85-98

O

Out-of-vocabulary
An OOV Word Embedding Framework for Chinese Machine Reading Comprehension; Luo, S.-B., 23(2): 67-84

P

Polarity Classification
Sentiment Analysis on Social Network: Using Emoticon Characteristic for Twitter Polarity Classification; Chen, C.-P., 23(1): 1-18

Q

Q&A for Yes/No Questions
Supporting Evidence Retrieval for Answering Yes/No Questions; Wu, M.-T., 23(2): 47-66

R

RNNs
A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network; Lai, C.-h., 23(2): 1-18

S

Semi-supervised Training
Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition; Lo, T.-H., 23(2): 19-34

Sentiment Analysis
Sentiment Analysis on Social Network: Using Emoticon Characteristic for Twitter Polarity Classification; Chen, C.-P., 23(1): 1-18

Story-telling
A Lexical Coherence Representation Computational Framework using LSTM Forget Gate For Autism Recognition; Liu, Y.-S., 23(1): 19-34

Supporting Evidence Retrieval
Supporting Evidence Retrieval for Answering Yes/No Questions; Wu, M.-T., 23(2): 47-66

V

Vocal Emotion Perception
Joint Modeling of Individual Neural Responses using a Deep Voting Fusion Network for Automatic Emotion Perception Decoding; Hsieh, W.-T., 23(1): 35-48

W

Word Embedding
Sentiment Analysis on Social Network: Using Emoticon Characteristic for Twitter Polarity Classification; Chen, C.-P., 23(1): 1-18

An OOV Word Embedding Framework for Chinese Machine Reading Comprehension; Luo, S.-B., 23(2): 67-84

Word Vector
Exploring the Use of Neural Newtork based Features for Text Readability Classification; Tseng, H.-C., 22(2): 31-46

Word2Vec
Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network; Chou, W.-C., 23(2): 85-98

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

<table>
<tr><td>會員類別</td><td colspan="2">□終身 □個人 □學生</td><td>會員編號</td><td colspan="2">（由本會填寫）</td></tr>
<tr><td rowspan="2">姓　　名</td><td rowspan="2"></td><td rowspan="2">性別</td><td rowspan="2"></td><td>出生日期</td><td>年　月　日</td></tr>
<tr><td>身分證號碼</td><td></td></tr>
<tr><td>現　　職</td><td></td><td>學　　歷</td><td colspan="2"></td></tr>
<tr><td>通訊地址</td><td colspan="5">□□□</td></tr>
<tr><td>戶籍地址</td><td colspan="5">□□□</td></tr>
<tr><td>電　　話</td><td colspan="2"></td><td>E-Mail</td><td colspan="2"></td></tr>
<tr><td colspan="6">申請人:　　（簽章）

中　華　民　國　　年　　月　　日</td></tr>
</table>

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

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

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

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

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

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

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

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

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

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

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

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

Here shows an example.

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

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

Please visit the following websites for details.

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

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

No page charges are levied on authors or their institutions.

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

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

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

International Journal of
Computational Linguistics & Chinese Language Processing
vol.23 no.2 December 2018

Contents

Forewords..... i
Chen-Yu Chiang and Min-Yuh Day

Papers

使用長短期記憶類神經網路建構中文語音辨識器之研究 [A Study on Mandarin Speech Recognition using Long Short- Term Memory Neural Network]..... 1
賴建宏(Chien-hung Lai), 王逸如(Yih-Ru Wang)

結合鑑別式訓練與模型合併於半監督式語音辨識之研究 [Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition]..... 19
羅天宏(Tien-Hong Lo), 陳柏琳(Berlin Chen)

結合鑑別式訓練聲學模型之類神經網路架構及優化方法的改進 [Leveraging Discriminative Training and Improved Neural Network Architecture and Optimization Method]..... 35
趙偉成(Wei-Cheng Chao), 張修瑞(Hsiu-Jui Chang), 羅天宏(Tien-Hong Lo), 陳柏琳(Berlin Chen)

Supporting Evidence Retrieval for Answering Yes/No Questions..... 47
Meng-Tse Wu, Yi-Chung Lin and Keh-Yih Su

未登錄詞之向量表示法模型於中文機器閱讀理解之應用 [An OOV Word Embedding Framework for Chinese Machine Reading Comprehension]..... 67
羅上堡(Shang-Bao Luo), 李青憲(Ching-Hsien Lee), 涂家章(Jia-Jang Tu), 陳冠宇(Kuan-Yu Chen)

以深層類神經網路標記中文階層式多標籤語意概念 [Hierarchical Multi-Label Chinese Word Semantic Labeling using Deep Neural Network]..... 85
周瑋傑(Wei-Chieh Chou), 王逸如(Yih-Ru Wang)

Reviewers List & 2018 Index..... 99

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing