The 31st

ROCLING 2019

October 3-5, 2019, New Taipei City, Taiwan

Proceedings of the Thirty-first Conference on Computational Linguistics and Speech Processing

Proceedings of the Thirty-first Conference on Computational Linguistics and Speech Processing ROCLING XXXI (2019)

October 3-5, 2019

National Taipei University, New Taipei City, Taiwan

Organizers

National Taipei University
Tamkang University
Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
Research Center for Information Technology Innovation, Academia Sinica

Sponsors

Ministry of Science and Technology
Ministry of Education
Cyberon Corporation
Chunghwa Telecom Laboratories
Institute for Information Industry
eLAND Information Co., Ltd.
Delta Electronics, Inc.

First Published October 2019
By The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

Chen-Yu Chiang, Min-Yuh Day, Jen-Tzung Chien, Ying-Hui Lai, Jenq-Haur Wang, Hung-Yi Lee, Yu Tsao, Kuan-Yu Menphis Chen, Shih-Hung Wu, Chih-Hua Tai (eds.)

Proceedings of the Thirty-first Conference on Computational Linguistics and Speech Processing (ROCLING XXXI)
2019-10-3/2019-10-4
ACLCLP
2019-10

ISBN 978-986-95769-2-5

Welcome Message of the ROCLING 2019

On behalf of the organization committee, it is our pleasure to welcome you to National Taipei University (NTPU) in New Taipei City, Taiwan, for the 31st Conference on Computational Linguistics and Speech Processing (ROCLING), the flagship conference on computational linguistics, natural language processing, and speech processing in Taiwan. ROCLING is the annual conference of the Computational Linguistics and Chinese Language Processing (ACLCLP) which is held in autumn in different cities and universities in Taiwan.

ROCLING 2019 features two distinguished keynote speeches from the renowned speakers in natural language processing as well as speech processing. Prof. Nobuaki Minematsu (Full Professor, Department of Electrical Engineering and Information Systems, Graduate School of Engineering, The University of Tokyo) will give a keynote on "How can speech technologies support learners to improve their skills of speaking, listening, conversation and more?". Prof. Yoshinobu Kano (Associate Professor, Faculty of Informatics, Shizuoka University) will speak on "Beyond end-to-end learning: Dialog system, sentence generation, and conversation analysis for automatic mental disorder diagnosis".

ROCLING 2019 will provide an international forum for researchers and industry practitioners to share their new ideas, original research results and practical development experiences from all NLP areas, including computational linguistics, information understanding, and speech processing. In addition to the regular conference sessions during October 3-4, 2019, the AI Tutorial organized by SIG-AI (Artificial Intelligence Special Interest Group) of ACLCLP (convener: Hung-Yi Lee, Assistant professor of the Department of Electrical Engineering of National Taiwan University) and The Science & Technology Policy Research and Information Center (STPI) will provide Artificial Intelligence Courses that focus on speech processing and NLP applications on October 5, 2019. ROCLING 2019 is sure to be an exciting event for all who attend.

This conference would not have been possible without the tremendous effort of organizing committee of dedicated and motivated research leaders have worked closely together to put together the attractive and intensive scientific program. Their great achievements have contributed much to the visibility of ROCLING 2019. We would like to heartedly thank them all. Special thanks to organizers who have worked hard to produce the proceedings, communicate with participants/authors, and handle the registration, budget, local arrangements and logistics. Thank you to all organizers including Program Chairs: Ying-Hui Lai and Jenq-Haur Wang, Tutorial Chair: Hung-Yi Lee, Industry Track Chair: Yu Tsao, Doctoral Consortium Chair: Kuan-Yu Menphis Chen, Academic Demo Track Chair: Hung-Yi Lee, Publication Chair: Shih-Hung Wu, Web Chair: Chih-Hua Tai, to participants, to authors who submitted papers and to program committee members and the reviewers who invested their valuable time and effort to provide timely and comprehensive reviews. Finally, we thank the generous government, academic and industry sponsors and appreciate your enthusiastic participation and support. Best wishes a successful and fruitful ROCLING 2019 in New Taipei, Taiwan.

General Chairs
Chen-Yu Chiang, Min-Yuh Day, and Jen-Tzung Chien

Organizing Committee

Conference Co-Chairs

Chen-Yu Chiang, National Taipei University

Min-Yuh Day, Tamkang University

Jen-Tzung Chien, National Chiao Tung University

Program Chairs

Ying-Hui Lai, National Yang-Ming University

Jenq-Haur Wang, National Taipei University of Technology

Tutorial Chair

Hung-Yi Lee, National Taiwan University

Industry Track Chair

Yu Tsao, Academia Sinica

Doctoral Consortium Chair

Kuan-Yu Menphis Chen, National Taiwan University of Science and Technology

Academic Demo Track Chair

Hung-Yi Lee, National Taiwan University

Publication Chair

Shih-Hung Wu, Chaoyang University of Technology

Web Chair

Chih-Hua Tai, National Taipei University

Proceedings of the Thirty-first Conference on Computational Linguistics and Speech Processing ROCLING XXXI (2019)

TABLE OF CONTENTS

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 1-2

基於特徵粒度之訓練策略於中文口語問答系統之應用

A Feature-granularity Training Strategy for Chinese Spoken Question Answering

羅上堡 Shang-Bao Luo
國立臺灣科技大學系資訊工程系
Department of computer science and information engineering
National Taiwan University of Science and Technology
M10615012@mail.ntust.edu.tw

陳冠宇 Kuan-Yu Chen
國立臺灣科技大學系資訊工程系
Department of computer science and information engineering
National Taiwan University of Science and Technology
kychen@mail.ntust.edu.tw

摘要

在口語問答系統(Spoken Question Answering, SQA)中，一個簡單且直覺的作法，是先將一段音訊透過自動語音辨識(Automatic Speech Recognition, ASR)轉換成一連串的辨識文字結果，再輸入給現有各式基於文字的問答系統模型來完成任務需求。然而，這樣的做法通常會遭遇自動語音辨識錯誤(ASR Errors)的影響，導致問答系統模型的效果不如預期。為了解決此一問題，本論文提出一種基於輸入特徵粒度的訓練策略，其目標是改善自動語音辨識錯誤所造成的效能損失，而且不需要額外模型的需求即可完成。我們將本論文所提出之訓練策略運用於中文口語機器閱讀理解(Machine Reading Comprehension, MRC)任務之中，驗證此一方法對於自動語音辨識錯誤的影響與改善。

關鍵詞：口語問答系統，語音辨識，特徵粒度，訓練策略。

Abstract

In a spoken question answering (SQA) system, a straightforward strategy is to transcribe given speech utterances into text using an ASR system. After that, classic methods can be readily used to the auto-transcribe text. However, such a strategy usually can not achieve a good performance due to the recognition errors. In order to mitigate the problem, in this paper, we

propose a feature-granularity training strategy for SQA. Specifically, the proposed method is a training strategy, thus we don't need to modify the classic SQA (or QA) methods. In the experiments, we evaluate the proposed feature-granularity training strategy on a Chinese machine reading comprehension task. The results demonstrate that the proposed strategy can overcome the effects caused by the recognition errors on the spoken machine reading comprehension task.

Keywords : Spoken question answering, automatic speech recognition, feature-granularity, training strategy.

致謝

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 108-2636-E-011-005 (Young Scholar Fellowship Program), and by the Project J367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 3-12

新穎的序列生成架構於中文重寫式摘要之研究

Novel Sequence Generation Framework for Chinese Abstractive Summarization

簡靖岳　Chin-Yueh Chien
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
M10615110@mail.ntust.edu.tw

陳冠宇　Kuan-Yu Chen
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
kychen@mail.ntust.edu.tw

摘要

近年網路訊息的爆發式成長，人們每天都能接觸到海量的訊息，但文章中常常包含非必要的資訊、雜訊，降低人們閱讀的效率，若能使用自動文章摘要(Automatic Document Summarization)的技術，將文章中重點萃取出來，便可大幅節省人們閱讀的時間。目前的自動摘要方法，主要分為抽取式(Extractive)摘要與重寫式(Abstractive)摘要，且大部分研究皆驗證在英文的資料集上。本論文提出兩種新穎的序列生成架構於重寫式摘要，包含「以 BERT 為編碼器之指針生成摘要法」與「融合 BERT 與 Transformer 之指針生成摘要法」。此外，目前重寫式摘要研究多半是以英文語料為研究目標，因此在本研究中，我們探討這些模型於中文重寫式摘要的任務成效，以作為後續研究的重要比較基礎。

關鍵詞：自動文章摘要、BERT、Transformer、指針生成網路。

一、緒論

自動文章摘要之研究分為兩大類，抽取式(Extractive)摘要與重寫式(Abstractive)摘要。前者依據特定的摘要比例，從原始文章中選取具代表性的語句來組成摘要。後者則是讓機器閱讀整篇文章，理解文章內容後，重新撰寫摘要代表這篇文章，其使用的詞彙不一定

全來自原始文章，這種摘要方式可說是更貼近人類平常撰寫摘要的形式。

近年來自動摘要的研究中，序列對序列模型應用於重寫式摘要的研究[1-5]，在眾多資料集中驗證其豐碩的成果。特別是近年提出的指針生成網路(Pointer Generator Network, PGN)，其機制可以有效解決文章中存在非字典詞彙(Out-of-vocabulary)的問題，因此被應用在重寫式摘要上。近年由谷歌提出的 Transformer[6]架構，使用注意力(Attention)機制，可以解決遞迴神經網路的長時間序列信息丟失問題並平行處理，加速網運算。谷歌基於 Transformer 架構進一步地提出 BERT，在自然語言處理(Natural Language Processing, NLP)的各項任務中，使用 BERT 效果均獲得顯著的提升。

本論文提出兩個新穎的重寫式摘要模型。第一個模型以指針生成網路為基礎，加上 BERT 作為編碼器，期望 BERT 能生成強健且準確的文章表示特徵，提升重寫式摘要的任務成效，我們稱為「以 BERT 為編碼器之指針生成摘要法」。第二個模型為第一個模型的延伸，除了使用 BERT 作為編碼器外，我們使用 Transformer 架構代替傳統遞迴神經網路，讓注意力機制來獲得時間序列上的關係，用以解決長時間序列訊息丟失問題，產生更加強健且依賴上下文資訊的特徵，期望讓重寫式摘要效能再進一步提升，我們稱為「融合 BERT 與 Transformer 之指針生成摘要法」。此外，我們將探討這些模型在中文重寫式摘要的任務成效，以作為後續研究的重要比較基礎。

二、相關方法

（一）序列對序列模型

序列對序列模型(Sequence-to-sequence)[7]主要包含兩大部分，編碼器(Encoder)與解碼器(Decoder)。編碼器與解碼器大多由遞迴神經網路(Recurrent Neural Networks)構成，例如長短期記憶網路(Long Short-term Memory, LSTM)[8]。

給定一段詞彙序列$\{w_1, w_2 \dots, w_n, \dots w_N\}$依序輸入編碼器中，每個詞彙$w_N$之詞向量$x_n$會與前一個時間點遞迴神經網路的輸出$h_{n-1}$，一起輸入遞迴神經網路，產生此時間點遞迴神經網路的輸出h_n。在解碼器部分，由於輸入文章的每個詞彙對於解碼器產生的每個輸出重要程度並不一樣，有研究提出加入注意力機制[9]，使得解碼器在每個時間點，會對編碼器的所有時間點產生一個注意力權重，表示編碼器中每一個時間點對於此時解碼

器的重要性。在解碼器中，遞迴神經網路會在每個時間點產生一個輸出向量s_t，此一向量將與編碼器的每個時間點輸出h_n計算得到一個相關性權重e_n^t，並透過正規化(Normalize)，獲得t時間點，解碼器對於編碼器的注意力權重$a^t = softmax(e_n^t)$，其中v^T, W_h, W_s, b_{attn}即為注意力機制中的模型參數。接著，將每一個注意力權重a_n^t與所對應的h_n相乘後加總，就可獲得當前時間點之注意力向量s_t^*:

$$e_n^t = v^T tanh(W_h h_n + W_s s_t + b_{attn}) \quad \text{式(1)}$$

$$s_t^* = \sum_{n=1}^{N} a_n^t h_n \quad \text{式(2)}$$

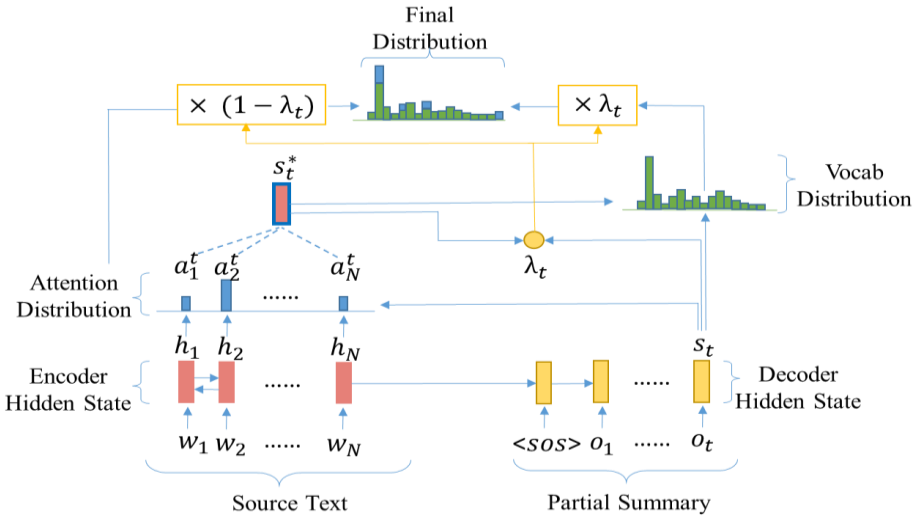
再經過全連接層以及$softmax$激活函數，會輸出一個辭典大小維度的機率分布，每一個維度對應一個字典中的詞。我們將分數最高的詞選取出來，作為此時間點的輸出，同時也是下個時間點的輸入。重複步驟直到解碼器輸出特殊符號 $< EOS >$為止。

（二）指針生成網路(Pointer Generator Network, PGN)

在傳統序列對序列之重寫式摘要的任務中，某些詞彙雖然出現在輸入文字序列中，但若它並非存在字典中(Out-of-Vocabulary, OOV)，這類詞彙是無法被解碼器解碼出來的。為了解決此一問題，近年有研究提出了指針生成網路(Pointer Generator Network, PGN)，其架構如圖一所示。一篇文章之詞彙$\{w_1, w_2 \dots, w_n, \dots w_N\}$，逐一輸入編碼器遞迴神經網路後，產生每個時間點的輸出向量h_n。與序列對序列模型中的解碼器一樣，遞迴神經網路會在每個時間點產生輸出s_t，並與編碼器每個時間點h_n產生e_n^t，藉由正規化得到注意力權重a^t，然後我們將注意力權重$a^t = [a_1^t, \dots, a_n^t, \dots, a_N^t]$乘上編碼器每個時間點之$h_n$，相加後得到當前時間點解碼器之注意力向量$s_t^*$，再透過全連接層以及$softmax$激活函數，可以得到字典(Vocabulary)中所有詞彙的機率分布P_{vocab}：

$$P_{vocab} = softmax(\grave{V}(V[s_t, s_t^*] + b) + \grave{b}) \quad \text{式(3)}$$

其中，$\grave{V}, V, b, \grave{b}$為可學習之模型參數。除了$P_{vocab}$外，指針生成網路會產生一個僅由輸入的文字序列計算而得的語言模型P_{PGN}，這個語言模型的辭典僅由輸入中所有不同的字詞所組成，因此可能包含沒有出現在P_{vocab}中的字詞。P_{PGN}可以快速地由注意力權重a^t計算而得，由於注意力權重a^t已經過正規化，因此P_{PGN}必定滿足機率公設$\sum_w P_{PGN}(w) = 1$。最後，我們利用s_t、s_t^*和解碼器時間點t之輸入x_t計算出P_{vocab}與P_{PGN}的結合係數λ_t，並



圖一、指針生成網路(Pointer Generator Network, PGN)架構圖

透過線性組合，產生解碼器在時間點t的參考機率分布$P(w)$。藉由此方式，不同於序列對序列模型，指針生成網路便可以輸出非字典裡的詞。

$$P_{PGN}(w) = \sum_{n:\ w_n=w} a_n^t \quad \text{式(4)}$$

$$\lambda_t = \sigma(W_{s^*} s_t^* + W_s s_t + w_x x_t + b_{ptr}) \quad \text{式(5)}$$

$$P(w) = (1-\lambda_t)P_{vocab}(w) + \lambda_t P_{PGN}(w) \quad \text{式(6)}$$

（三）Transformer

Transformer 為谷歌近年提出的序列對序列模型[6]，其使用注意力機制搭配捲積神經網路來處理時間序列輸入。在編碼器端，Transformer 透過矩陣運算的方式，將每個輸入詞彙w_n分別乘上Q, K, V三個矩陣，得到q_n, k_n, v_n三個向量，接著詞彙w_n的q_n向量分別會與所有詞彙的k_i向量計算出注意力權重，而後注意力權重再與所對應的v_i相乘後相加，產生詞彙w_n新的向量表示法。於解碼器端，解碼器的詞彙除了與解碼器其他詞彙計算注意力外，也與編碼器每個輸入詞彙計算注意力，藉此來與編碼器產生關係，最後產生當前時間點之輸出。相較於傳統序列對序列模型，Transformer 主要由注意力機制組成，且因為 Transformer 可以進行平行化訓練，相較於遞迴神經網路，可減少相當多的時間成本。

（四）BERT

基於 Transformer，谷歌又進一步地提出了 BERT(Bidirectional Encoder Representations

from Transformers)模型[10]，它使用多層 Transformer 作為主要模型架構，並在兩個任務上進行訓練。第一個任務是遮蔽式語言模型(Masked Language Model)，其作法是隨機將輸入遮蔽，模型訓練的目標為預測被遮蔽的輸入；第二個任務則是下一句預測(Next Sentence Prediction)，做法為同時輸入兩個句子，模型需預測這兩個句子是否為上下文的關係。BERT 的訓練不需要標記資料，因此它可以在非常大量的資料中進行這兩個任務的訓練，而形成一個強健的預訓練模型，而後，其他自然語言任務只需要在這個預訓練模型上進行參數微調，便能取得很好的效果[11]。

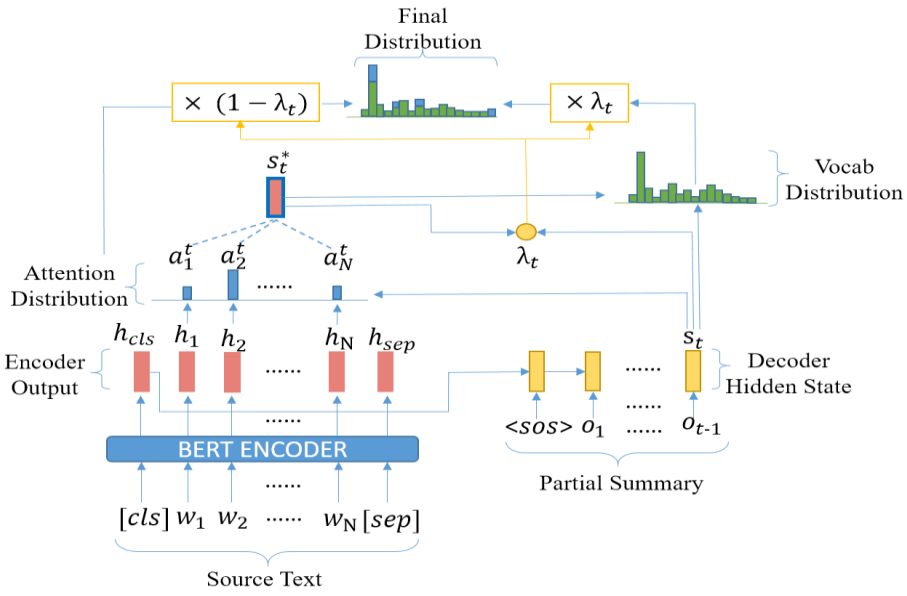
三、方法

本研究嘗試從兩個面向探討現階段重寫式摘要的問題，並藉由指針生成網路、Transformer 和 BERT，提出新穎的重寫式摘要模型，期望增進重寫式摘要的成效。第一個面向是編碼器是否能生成足夠代表文章之表示法：由於解碼器需藉由編碼器生成之文章表示法，來生成出對應此篇文章之摘要，因此如何生成好的文章表示法顯然是一個重要的任務。第二個面向是遞迴神經網路的取代性：當輸入遞迴神經網路的訊息越長，較遠的資訊對於後面時間點的影響越薄弱，但較遠的資訊也可能很重要。此外，遞迴神經網路模型需要依序訓練，無法有效運用平行運算加速處理，常消耗很多時間成本。

有鑑於此，本論文提出兩個新穎的重寫式摘要模型，第一個模型是以指針生成網路為基礎，以 BERT 作為句子表示法的編碼器，期望藉由 BERT 來生成強健且準確的文章表示特徵，使得解碼器能依據此表示法，生成出更好的摘要。第二個模型為第一個模型的延伸，除了使用 BERT 作為編碼器外，更進一步將遞迴神經網路改為使用 Transformer 架構，用注意力機制來獲得時間序列上的關係，藉此解決長時間序列訊息丟失問題，產生更強健且依賴上下文資訊的特徵，期望重寫式摘要的效能再進一步的提升。

（一）以 BERT 為編碼器之指針生成摘要法

為了讓序列對序列模型的編碼器能產生更強健的表示法，我們提出一套以 BERT 為編碼器的指針生成摘要法，其模型架構如圖二所示。我們將文章中文字序列$\{w_1, w_2 \dots, w_n, \dots w_N\}$的前後分別加入$[cls]$與$[sep]$，作為文章開始與結束的符號，輸入到 BERT 編碼器中，BERT 將輸出每個字所對應之向量表示法$\{h_{cls}, h_1, h_2 \dots, h_n, \dots, h_N, h_{sep}\}$，我們將$[cls]$所對應之向量$h_{cls}$視為整篇文章的向量表示法。解碼器端，我們同樣採用傳

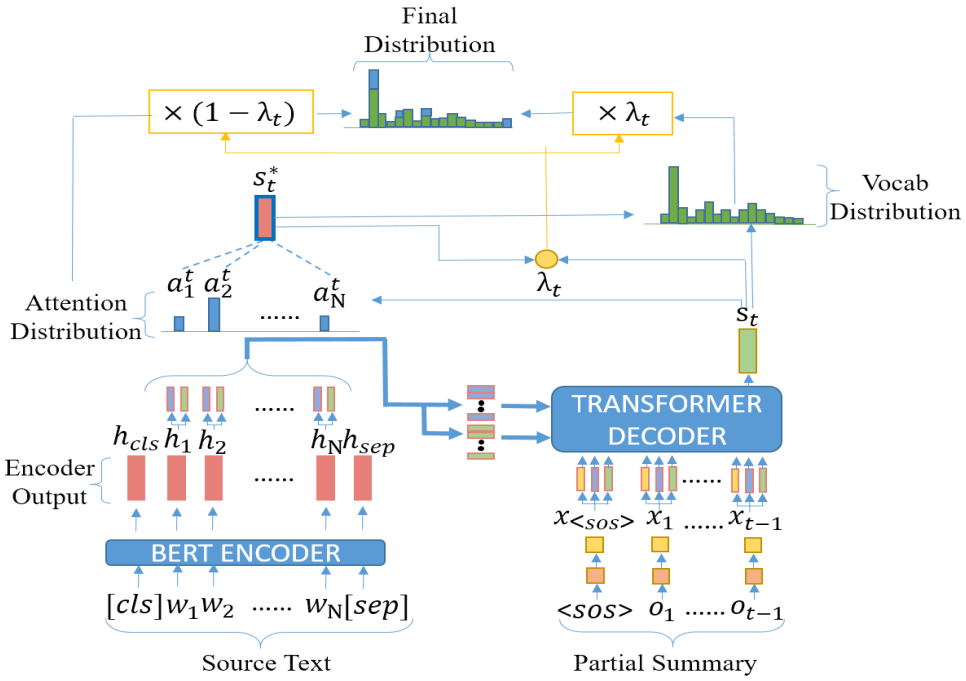


圖二、以 BERT 為編碼器之指針生成摘要法架構圖

統遞迴神經網路為模型基礎。在解碼的過程中，遞迴神經網路會在每個時間點t產生一個輸出s_t，s_t會與編碼器輸出的每個字向量表示法h_n計算出注意力權重$a^t = [a_1^t, \dots, a_n^t, \dots, a_N^t]$，其中$[cls]$與$[sep]$並沒有加入計算。得到注意力權重後，編碼器中每個字向量表示法h_n與時間點t的注意力權重a_n^t相乘並相加後，即得到當前時間點t於解碼器的注意力向量s_t^*。最後，如同指針生成網路，我們使用s_t與s_t^*產生詞彙生成機率分布P_{vocab}；以注意力權重a^t產生P_{PGN}；再利用s_t、s_t^*和解碼器時間點t之輸入x_t計算出P_{vocab}與P_{PGN}的結合係數λ_t，並透過線性組合，產生解碼器在時間點t的參考機率分布$P(w)$（可參閱二-(二)）。與傳統序列對序列、指針生成網路相比，我們改用 BERT 作為文章的特徵抽取器，期待藉由更強健的文章特徵可以產生更佳的重寫式摘要。

（二）融合 BERT 與 Transformer 之指針生成摘要法

此方法為以 BERT 為編碼器之指針生成摘要法之改良模型，方法架構如圖三所示。首先，當我們將文章中文字序列$\{w_1, w_2 \dots, w_n, \dots w_N\}$的前後分別加入$[cls]$與$[sep]$，輸入到 BERT 編碼器之中，獲得對應之向量表示法$\{h_{cls}, h_1, \dots, h_n, \dots, h_N, h_{sep}\}$後，將每個字向量$h_n$各自乘上$K^1, V^1$兩個矩陣，產生兩個對應向量$k_n$與$v_n$，這兩組向量（即$\mathbf{k}_1^N = \{k_1, k_2 \dots, k_n, \dots, k_N\}$與$\mathbf{v}_1^N = \{v_1, v_2 \dots, v_n, \dots, v_N\}$）將用於解碼器之中。在解碼的過程中，當我們解碼第$t$時間點時，會藉由 Transformer 機制，考慮時間點t之前所生成的所有字詞$\{o_1, o_2 \dots, o_{t-1}\}$。更明確地，我們先將$\{o_1, o_2 \dots, o_{t-1}\}$通過詞向量層(Embedding Layer)，



圖三、融合 BERT 與 Transformer 之指針生成摘要法

轉換為詞向量後，$\{h_{o_1}, h_{o_2}, \dots, h_{o_{t-1}}\}$透過$Q^2, K^2, V^2$三個權重矩陣進行轉換，為每一個字分別產生三個向量表示法，即$\mathbf{q}_{o_1}^{o_{t-1}} = \{q_{o_1}, q_{o_2}, \dots, q_{o_{t-1}}\}$、$\mathbf{k}_{o_1}^{o_{t-1}} = \{k_{o_1}, k_{o_2}, \dots, k_{o_{t-1}}\}$與$\mathbf{v}_{o_1}^{o_{t-1}} = \{v_{o_1}, v_{o_2}, \dots, v_{o_{t-1}}\}$。接下來，我們依照 Transformer 自我注意(Self-attention)的機制，計算時間點t時的表示法s_t，最後我們再把編碼器的資訊也考慮進來產生s_t^*:

$$s_t = softmax\left(\frac{q_{o_{t-1}}\left(\boldsymbol{k}_{o_1}^{o_{t-1}}\right)^{\intercal}}{\sqrt{dim}}\right)\boldsymbol{v}_{o_1}^{o_{t-1}} \qquad \text{式(7)}$$

$$s_t^* = softmax\left(\frac{s_t\left(\boldsymbol{k}_1^N\right)^{\intercal}}{\sqrt{dim}}\right)\boldsymbol{v}_1^N \qquad \text{式(8)}$$

dim表示向量的維度，$\intercal$表示矩陣的轉置，並且$softmax\left(\frac{s_t\left(\boldsymbol{k}_1^N\right)^{\intercal}}{\sqrt{dim}}\right)$即為注意力權重$a^t = [a_1^t, \dots, a_n^t, \dots, a_N^t]$。如同指針生成網路，我們使用$s_t$與$s_t^*$產生詞彙生成機率分布$P_{vocab}$；以注意力權重$a^t$產生$P_{PGN}$；再利用$s_t$、$s_t^*$和解碼器時間點$t$之輸入$x_{t-1}$計算出$P_{vocab}$與$P_{PGN}$的結合係數$\lambda_t$，並透過線性組合，產生解碼器在時間點$t$的參考機率分布$P(w)$（可參閱二-(二)）。，此一方法不僅使用 BERT 作為編碼器，更進一步地使用 Transformer 取代傳統的遞迴神經網路，期望這套融合 BERT 與 Transformer 之指針生成摘要法不僅可以平行運算，也可以獲得更好的摘要成效。

表一、基礎系統與本論文所提出方法之實驗結果

		ROUGE-1	ROUGE-2	ROUGE-L
Baseline Systems	seq2seq	0.225	0.153	0.211
	PGN	0.451	0.323	0.368
Our Approaches	Method 1	0.499	**0.346**	0.397
	Method 2	**0.508**	0.340	**0.403**

表二、進階的基礎系統實驗結果

		ROUGE-1	ROUGE-2	ROUGE-L
Advanced Baseline Systems	seq2seq	0.243	0.157	0.212
	PGN	0.469	0.335	0.371

四、實驗設定與結果討論

（一）實驗設定

本論文使用之資料集為 MATBN 中文摘要資料集，共有 205 則文章與對應之摘要，我們依照 80%、10%、10%的比例將資料集隨機切分為訓練集、驗證集以及測試集。實驗結果為三次隨機切分資料集的平均分數。衡量指標為召回率導向的摘要評估(Recall-Oriented Understudy for Gisting Evaluation, ROUGE)[12]，以 ROUGE-1、ROUGE-2 與 ROUGE-L 三種指標為衡量標準。參數設置方面，詞向量設定為 128，隱藏層設定為 256，丟失率設定為 0.3，採用 Adam 優化器，學習率設定為 0.001，批次大小設定為 32。Transformer 參數設置則依照原論文之設置[6]，而層數則設定為 1 層。

（二）實驗結果

首先第一組的實驗中，我們先測試基礎系統（即序列對序列模型(seq2seq)與指針生成網路(PGN)），在中文重寫式摘要的成效，實驗結果如表一所示。實驗結果發現，序列對序列模型在重寫式摘要上已經展現了一定的生成能力。當比較指針生成網路與序列對序列模型時，指針生成網路的 ROUGE 分數有大幅度提升，其中 ROUGE-1 與 ROUGE-2 分數更是兩倍以上。我們認為指針生成網路類似於生成式摘要與抽取式摘要的結合，不僅從文章中抽取重要的詞彙作為摘要，也從詞彙表中選擇合適的詞彙加入摘要，在這種架構下能夠學習語句更通順且更有代表性的摘要，進而提升摘要生成之品質。

接著，我們檢驗本研究所提出的兩個方法，實驗結果如表一所示。首先，我們所提出的第一個方法「以 BERT 為編碼器之指針生成摘要法」(Method 1)，相較於序列對序列模型，成效有大幅度的提升；而當我們進一步比較指針生成網路與我們提出的以 BERT 為編碼器之指針生成摘要法，在 3 種評估指標中，分別有 9.6%、6.8%及 7.1%的相對提升。此結果顯示相較於遞迴神經網路，BERT 能取出更為強健的文章與字詞向量表示法，使得解碼器能透過此表示法生成出更具代表性的文章摘要。而當我們進一步地將遞迴神經網路架構以 Transformer 架構取代，則形成我們所提出的第二個方法「融合 BERT 與 Transformer 之指針生成摘要法」(Method 2)，其實驗結果顯示，這個結合以 BERT 作為特徵抽取、Transformer 作為模型的主幹，再輔以指針生成網路的概念所形成的重寫式摘要法，可以獲得相當優良的實驗成效。

最後，為了分析 Transformer 與傳統遞迴神經網路的差異，我們試著將基礎系統的解碼器皆以 Transformer 取代傳統的長短期記憶模型，其實驗結果如表二所示。比較表二與表一中的基礎系統，實驗結果顯示使用 Transformer 於解碼器的序列對序列模型與指針生成網路，ROUGE 分數皆較使用長短期記憶模型的基礎系統有所提升，說明了 Transformer 確實較傳統遞迴神經網路有較好的效能。

五、結論

本論文透過兩個面向來探討現今重寫式摘要模型，第一個面向是如何在編碼器生成更強健且更代表文章內容之文章表示法，使解碼器端根據更完整的文章訊息，生成出更好的文章摘要。第二個面向是如何解決遞迴神經網路長時間序列下信息丟失問題，使長序列下的文字也能維持著強健的依賴關係。因此，本論文提出兩種新穎的摘要方法，即「以 BERT 為編碼器之指針生成摘要法」與「融合 BERT 與 Transformer 之指針生成摘要法」，並驗證於中文的摘要資料集中。實驗顯示，本研究所提出之改進方法，確實有效地提升摘要任務的成效。在未來，我們會持續深入地探討使用不同 Transformer 層數對於摘要的效能影響外，亦希望能將目前更新穎的 XLNET[13]也應用於重寫式摘要的任務之中。

六、致謝

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under

grant MOST 108-2636-E-011-005 (Young Scholar Fellowship Program), and by the Project J367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

參考文獻

[1] A. M. Rush, S. Harvard, S. Chopra, and J. Weston, "A Neural Attention Model for Sentence Summarization," in *ACLWeb. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2017.

[2] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," *arXiv:1705.04304,* 2017.

[3] R. Nallapati, B. Zhou, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:.06023,* 2016.

[4] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1243-1252: JMLR. org.

[5] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proceedings of the the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 93-98.

[6] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.

[7] I. Sutskever and O. Vinyals, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104-3112.

[8] S. Hochreiter and J. J. N. c. Schmidhuber, "Long short-term memory," vol. 9, no. 8, pp. 1735-1780, 1997.

[9] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv:.00685,* 2015.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805,* 2018.

[11] H. Zheng and M. Lapata, "Sentence Centrality Revisited for Unsupervised Summarization," *arXiv:.03508,* 2019.

[12] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.

[13] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *arXiv:1906.08237,* 2019.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 13-14

EBSUM: 基於 BERT 的強健性抽取式摘要法

EBSUM: An Enhanced BERT-based Extractive Summarization Framework

吳政育　Zheng-Yu Wu
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
M10615079@mail.ntust.edu.tw

陳冠宇　Kuan-Yu Chen
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
kychen@mail.ntust.edu.tw

摘要

目前大部分自動摘要方法，分為抽取式摘要(Extractive)與重寫式摘要(Abstractive)，重寫式摘要雖然能夠改寫文章形成摘要，但這並不是一種有效的方式，困難點在於語意不通順、重複字等。抽取式摘要則是從文章中抽取句子形成摘要，能夠避免掉語意不通順，重複字的缺點。目前基於 BERT(Bidirectional encoder representation from transformers)的抽取式摘要法，多半是利用 BERT 取得句子表示法後，再微調模型進行摘要句子之選取。在本文中，我們提出一套新穎的基於 BERT 之強健性抽取式摘要法(Enhanced BERT-based Extractive Summarization Framework, EBSUM)，它不僅考慮了句子的位置資訊、利用強化學習增強摘要模型與評估標準的關聯性，更直接的將最大邊緣相關性(Maximal Marginal Relevance, MMR)概念融入摘要模型之中，以避免冗餘資訊的選取。在實驗中，EBSUM 在公認的摘要資料集 CNN/DailyMail 中，獲得相當優良的任務成效，與經典的各式基於類神經網路的摘要模型相比，EBSUM 同樣可以獲得最佳的摘要結果。

關鍵詞：自動摘要，抽取式，BERT，強化學習，最大邊緣相關性

Abstract

Automatic summarization methods can be classified into two major spectrums: extractive summarization and abstractive summarization. Although abstractive summarization methods can produce a summary by using different words and phrases that were not in the given document, it usually leads to an influence summary. On the contrary, extractive summarization methods generate a summary by copying and concatenating the most important sentences in the given document, which usually can provide a more continuous and readable summary. Recently, BERT (Bidirectional encoder representation from transformers)-based extractive summarization methods usually leverage BERT to encode each sentence into a fixed low-dimensional vector, and then a fine-tuned BERT model can be used to predict a score for each sentence. On top of the predicted scores, we can rank these sentences and form a summary for a given document. In this paper, we propose an enhanced BERT-based extractive summarization framework (EBSUM), which not only takes sentence position information and RL training into account, but the maximal marginal relevance (MMR) criterion is also be considered. In our experiments, we evaluate the proposed framework on the CNN/DailyMail benchmark corpus, and the results demonstrate that EBSUM can achieve a better result than other classic extractive summarization methods.

Keywords: Extractive, Summarization, BERT, MMR.

致謝

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 108-2636-E-011-005 (Young Scholar Fellowship Program), and by the Project J367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 15-24

GALs: 基於對抗式學習之整列式摘要法

GALs: A GAN-based Listwise Summarizer

郭家銍 Chia-Chih Kuo
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
M10815022@mail.ntust.edu.tw

陳冠宇 Kuan-Yu Chen
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
kychen@mail.ntust.edu.tw

摘要

抽取式摘要（Extractive Summarization）著眼於選擇文本中的幾個句子，使其組成足以代表整篇文本內容的摘要。排序學習（Learning to Rank）[1]最早興起於資料檢索領域，並被應用於各種排序的任務之中。在本研究中，我們將抽取式摘要視為一個整列式（listwise）句子排序問題，提出一套基於對抗式學習之整列式摘要法（GAN-based Listwise Summarizer, GALs）。GALs 以生成對抗網路為架構，將抽取式摘要器作為生成器，並將其生成的摘要與參考答案的表面特徵（Surface Features）輸入給判別器，最後利用強化學習的方式，將判別器的預測做為回饋獎勵用於更新整個模型的參數。因此，本研究所提出之 GALs 融合了對抗式學習、整列式排序的概念、句子與文本的表面特徵以及強化學習，旨於提出一套經典的摘要模型方法。實驗中，我們不僅發現 GALs 在 CNN/Daily Mail 數據集上相較於傳統的最佳模型有明顯的分數提昇，我們亦對 GALs 模型所使用的參數，做了細節上的調查與分析。

關鍵詞：抽取式摘要，整列式排序，生成對抗網路，表面特徵

Abstract

Extractive summarization aims at selecting a set of sentences to form a summary for a given document. Learning-to-rank is first appeared in the field of information retrieval, and it has been employed to solve several ranking-based tasks. In this study, we regard the task of extractive summarization as a listwise sentence ranking problem, and thus a GAN-based

listwise summarizer (GALs) is proposed. On top of the generative adversarial network (GAN), an extractive summarizer is introduced to be the generator, and a discriminator is employed to distinguish the generated summary from the ground truth. Especially, the input to the discriminator is a set of surface features, which are extracted from the generated summary and the ground truth. Finally, GALs can be optimized by leveraging the reinforcement learning (RL) strategy. The experimental results demonstrate the effectiveness of the proposed framework on the CNN/Daily Mail corpus. Moreover, we make detailed investigation and analysis of the parameters used in GALs.

Keywords: Extractive summarization, listwise, GAN , surface features

一、緒論

抽取式摘要（Extractive Summarization）著眼於選擇文本中最顯著且具代表性的句子組成摘要，這些句子不僅須盡可能地保留文本中重要的資訊，也必須限制選取的摘要長度，因此確保摘要結果的低冗餘性就相當重要。為了達成這個目標，許多方法先對文本中的每個句子進行評分，再利用貪婪算法或線性規劃等方法，基於評分與長度等規則挑選理想的句子組合成摘要。在深度神經網路模型盛行的現在，多數模型以各式類神經網路為基礎，搭建句子評分模型，並以句子回歸（Sentence Regression）做為模型訓練的目標[2]。雖然這些方法在許多數據集中已驗證其成效，但實際上它們因為仰賴於手動調整評分、冗餘性與句子長度間的平衡性，而難以獲得全域最佳解。近年來，如指針網路（Pointer Network）[3]等生成模型的使用，也在抽取式摘要的研究中獲得成功[4]，將指針網路運用在抽取式摘要問題上時，模型將能夠自動決定句子選擇的順序與理想的摘要長度，克服了句回歸架構之弊端。然而近年的摘要研究中，許多模型仰賴深度神經網路的各式元件，例如詞嵌入（Word Embedding）[5]與長短期記憶模型（Long Short-Term Memory）[6]等，對文本與句子進行建模以進行各種預測，而疏於使用在文本摘要中富有價值之表面特徵（Surface Features）。

排序學習（Learning-to-Rank）[1]最早興起於資料檢索領域，並被應用於文本排序。在文本排序問題中，給予一串文本與查詢（Query）後，首先須對每筆文本與查詢評定兩者間的相關性，再依相關性之高低排序文本，其中逐點式（Pointwise）排序法會對每個文本做獨立的相關性評分，而整列式（Listwise）排序法則一次考慮所有的文本。我們觀察到逐點式文本排序，與基於句回歸架構的抽取式摘要具有高度的相似

性，因而猜想文本排序的研究設計同樣可適用於抽取式摘要問題，並參考整列式文本排序法，提出一套基於對抗式學習之整列式摘要法(GAN-based Listwise Summarizer, GALs)。我們的方法主要發想於 Jun Wang 所提出之 IRGAN[7]，IRGAN 的架構分為生成器與判別器兩部份，各自代表著資料檢索的兩大派系思想，其中生成器擅於利用向量空間（例：詞向量）對文本與查詢建模，用於預測與查詢較為相關的文本；判別器則擅於學習文本表面特徵與查詢間的潛在相關性分布；利用對抗式學習，IRGAN 能自然地融合生成器與判別器彼此的優勢，使兩者成效皆獲得提昇。因此，在 GALs 的設計上，也分為生成器與判別器,，生成器採用以生成模型為基礎的抽取式摘要器，由生成器產生的每一組摘要與真實的摘要，我們同時抽取其表面特徵作為判別器的輸入，判別器的目標為區分真實的摘要與由生成器產生的摘要。最後，透過基於策略梯度（Policy Gradient）的強化學習，GALs 融合了表面特徵、基於深度神經網路的生成器與判別器，以及整列式學習，希望可以進一步地提昇摘要品質。

二、相關研究

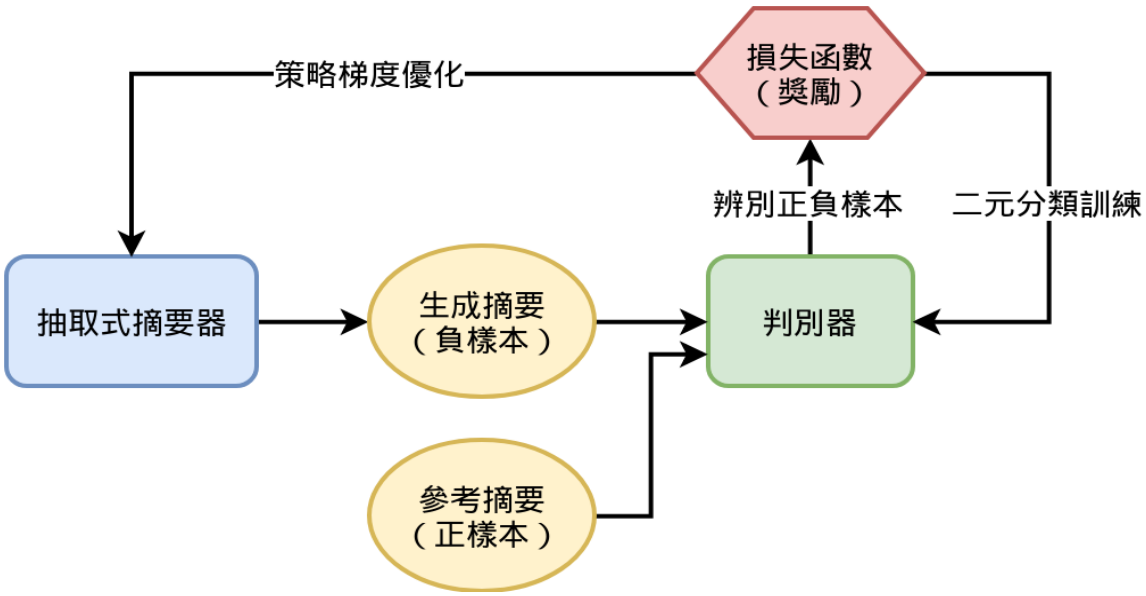
在 Pengjie Ren 提出的基於句回歸架構之抽取式摘要方法中[2]，ROUGE-2 [8]被作為句子重要性評分的參考值，並以類神經網路模型試圖預測之，最後用貪婪算法挑選句子以組成摘要。Pengjie Ren 展示其使用之四項表面特徵與 ROUGE-2 間具有高度的相關性。為了進一步利用文本前後文的特性以提昇摘要品質，Pengjie Ren 使用長短期記憶模型與卷積注意力（Convolutional Attention）機制來評估相鄰句子間的相似度。

Yen-Chun Chen [4] 並未使用表面特徵，而是採用詞嵌入[5]與基於序列到序列[9]與注意力機制[10]的指針網路（Pointer Network）[3]實作抽取式摘要，指針網路之編碼器首先以卷積神經網路對每個句子建立向量，再以雙向長短期記憶模型抽取文本的前後文特徵，而後解碼器將在每個時間步中，抽取最具注意力的句子以組成抽取式摘要。

三、基於對抗式學習之整列式摘要法

（一）總覽

本研究所提出之基於對抗式學習之整列式摘要法(GAN-based Listwise Summarizer, GALs)，是由生成器（抽取式摘要器）與判別器兩大部分所組成。在模型訓練時，當給予一篇文本與參考摘要作為正確答案，首先生成器將對文本生成一份抽取式摘要，



圖一、基於對抗式學習之整列式摘要法（GALs）之模型訓練架構

接著兩份摘要（即正確答案與由生成器產生的摘要）將輸入給判別器，以預測何者為參考摘要（正樣本）、何者為生成摘要（負樣本）。判別器的任務為二元分類問題，其損失函數同時也將作為抽取式摘要器的獎勵，以對其進行基於策略梯度的強化學習，使得抽取式摘要器所生成之摘要能夠更近似於參考摘要，如圖一所示。

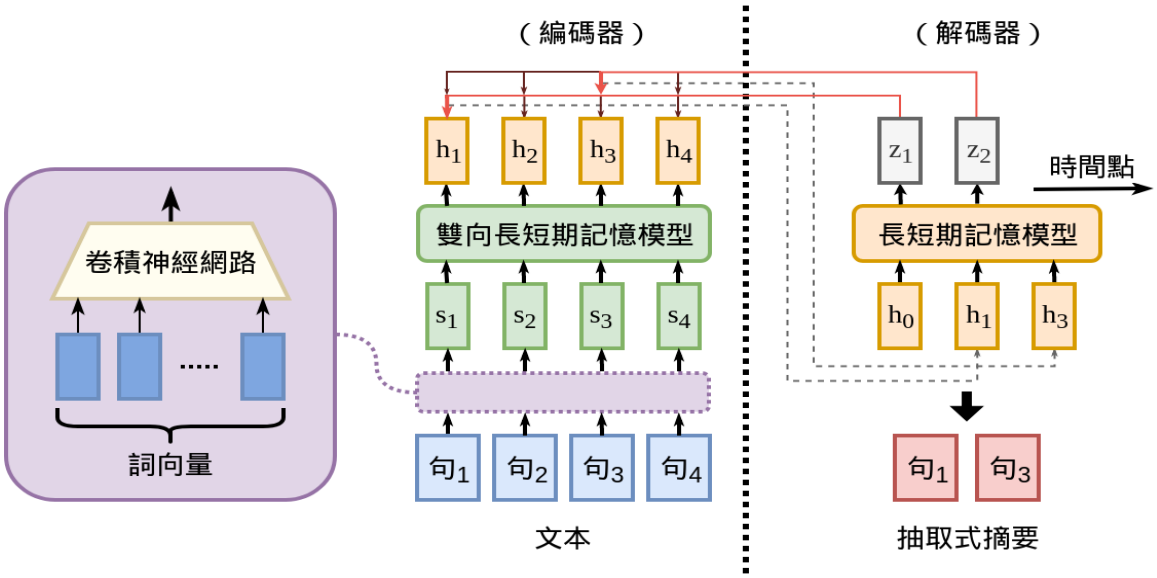
（二）生成器

在本研究中，我們選用 Yen-Chun Chen 於 2018 年所提出的基於指針網路之抽取式摘要器 [4]，如圖二所示。我們選用的抽取式摘要器分為編碼器與解碼器兩部分，編碼器首先會利用文本中每個句子的詞向量序列，以卷積神經網路計算出所有句子的向量表示 s_j。接著卷積神經網路輸出之一序列句子向量 s_j，將再被輸入進雙向長短期記憶模型中，以取得一序列情境導向（Context-aware）之句子向量表示 h_j。解碼器則由單向長短期記憶模型所組成，並基於編碼器輸出之句子向量 h_j以選擇欲抽取之句子，在每個時間點 t，長短期記憶模型會以上一個時間點被抽取之句子的向量 $h_{j_{t-1}}$作為輸入，並輸出一向量 z_t。z_t 首先會與文本中的句子向量 h_j 計算注意力以取得前後文向量 c_t。

$$\alpha^t = \text{softmax}(v_c^\top \tanh(W_{c1}h_j + W_{c2}z_t) \quad (1)$$

$$c_t = \sum_j \alpha_j^t W_{c1}h_j \quad (2)$$

接著將前後文向量 c_t 再次與文本中的句子向量h_j 計算注意力，並排除過去的時間點中已抽取的句子 $j_1, \ldots, j_{t-1}$，最後取得當前時間點抽取句子之機率分佈 $P(j_t|j_1, \ldots, j_{t-1})$。



圖二、基於指針網路之抽取式摘要器

$$e_j^t = \begin{cases} v_p^\top \tanh\left(W_{p1}h_j + W_{p2}c_t\right), & \text{if } j_t \neq j_k \ \forall k < t \\ -\infty, & \text{otherwise} \end{cases} \tag{3}$$

$$P(j_t|j_1, \ldots, j_{t-1}) = \text{softmax}(e^t) \tag{4}$$

基於上述解碼流程，解碼器將遞回抽取句子，直到觸發設定之停止條件。

（三）判別器

GALs 的判別器是一個用於分辨正負樣本的二元分類器，我們單純採用多層感知器（Multilayer Perceptron）[11]組成之類神經網路模型作為判別器。因為一份摘要通常包含了多個句子，因此我們計算句子間各個特徵的平均值與標準差作為判別器的輸入；若摘要中只有一個句子，則設定標準差為零。為了使判別器能提供更富價值之獎勵，我們利用遞歸特徵消除（Recursive Feature Elimination, RFE）[12] 篩除不具效益之特徵，最後挑選出以下八種特徵：

1、句子長度

長度較長的句子含有更多的詞彙，可能涵蓋的資訊量也較多，因此較長的句子通常較為重要，然而抽取過長的句子也可能造成摘要的冗餘性提昇。

2、絕對位置

文本中的前三至前六句話通常有高度的重要性，因此句子在文本中的絕對位置也是判別器的輸入之一。

3、相對位置

撰筆人常常會將多個重要的句子集中在文本中的一處，因此我們將句子的絕對位置除以文本的總句數，計算得相對位置作為特徵。

4、停用詞比例

停用詞（Stop Words）意指語言中極常出現的詞彙（例如英文中的 to, the, and 等等），具有較高停用詞比例之句子很有可能比其他句子來得更不具實際意義。

5、平均詞頻

詞頻（Word Frequency）即詞在文本中出現的頻率，可用於評估詞在文本中的重要性。我們計算句子中所有詞之詞頻的平均值作，平均詞頻較高的句子可能更為重要。

6、平均文本頻

文本頻（Document Frequency）常與詞頻一起使用，若某一詞彙在語料中的大多數文本中都出現過（即文本頻較高），則此詞彙難以突顯出句子間的差異性。我們計算句子中所有詞彙的文本頻平均值作為特徵，平均文本頻較高的句子可能較不重要。

7、句嵌入

我們使用 Sent2Vec [13]，一種旨於探究語意特徵的非監督方法，用以將句子形成分散式表示法（Distributed Representations），也納入考量之中。我們於 CNN/Daily Mail 的訓練集上訓練 100 維的 Sent2Vec 模型，以生成各個句子的句嵌入向量。

8、餘弦相似度

優良的摘要應避免冗餘的語句，意即任兩個句子間都不該擁有過高的相似性。基於句嵌入空間的特性，我們計算摘要中任兩個句子間句嵌入向量的餘弦相似度，以評估任兩個句子間的相似性，最後以平均值與標準差作為特徵使用。

（四）模型細節

在訓練時，GALs 將於生成器與判別器交替訓練，判別器需要一組參考摘要（正樣本）與生成器所生成之摘要（負樣本）作為輸入，以作二元分類的訓練，其中正負樣本的標準答案分別為 1.0 與 0.0。因為輸入之樣本必為一正一負，我們捨棄傳統的二元交叉

熵（Binary Cross-Entropy），改採將負樣本之預測分數減去正樣本之預測分數作為判別器的損失函數，此損失函數同時亦將作為生成器在訓練時使用的獎勵值。我們注意到因為此種計算方式有可能產生負值，這使得生成器與判別器可能傾向於攻擊彼此而非優化自身，因此我們將負值皆歸為 0：

$$D_{loss} = max(0,\ score_{neg} - score_{pos}) \tag{5}$$

而在優化生成器時，我們首先使用生成器產生最多前五個時間點的預測機率分佈 dec_t，分別從中取樣一個句子s，並加總被取樣句子之負對數似然（Negative Log-likelihood）乘上其獎勵值（即D_{loss}）作為生成器的損失函數：

$$G_{loss} = \sum_{t=1}^{5} \mathbb{E}_{s\sim p(s|dec_t)}[-log\ p(s|dec_t)] \cdot D_{loss} \tag{6}$$

（五）訓練流程

我們首先對參考摘要中的每個句子，在文本中提取與其 ROUGE-L 最高的句子組成抽取式摘要的參考答案，以預訓練抽取式摘要器。接著，我們將預訓練過的抽取式摘要器作為 GALs 的生成器，隨機初始化判別器的參數，並交替訓練生成器與判別器，以將作為抽取式摘要器進一步優化，直到在驗證集上獲得最佳的 ROUGE-1 評分。

四、實驗

（一）CNN/Daily Mail 數據集

為了測試本研究所提出之基於對抗式學習之整列式摘要法 GALs 的摘要成效，我們首先依照 Yen-Chun Chen 其論文[4]所附之公開原始碼，重製其在 CNN/Daily Mail 數據集上的實驗，這個結果將是 GALs 的基準系統(Baseline System)。為了克服策略梯度在訓練時的不穩定性與頻繁的梯度爆炸，我們採用 Adam 優化器與 0.00001 的學習率，並裁剪梯度至 2.0。我們在驗證集上根據 ROUGE-1，自動決定早停法（Early Stopping）與學習率衰減的時機。從表一之實驗結果可以看到，經優化後的模型在抽取式摘要的表現上，所有 ROUGE 成績皆有明顯的提昇。

（二）早停法與學習率衰減

在 GALs 優化模型的過程中，我們觀察模型在驗證集上 ROUGE-1、ROUGE-2 與

表一、各式模型於 CNN/Daily Mail 數據集之摘要結果

	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3	40.34	17.70	36.57
Lead-3 (our implementation)	40.27	17.73	36.48
Baseline System [4]	40.17	18.11	36.41
Baseline System (our implementation)	39.69	17.91	36.01
REFRESH [14]	40.00	18.10	36.60
CRSum [2]	40.52	18.08	36.81
EXTRACT [15]	40.62	18.45	37.14
GALs	**40.93**	**18.51**	**37.19**

表二、參考指標與優化結果

		ROUGE-1	ROUGE-2	ROUGE-L
參考指標	ROUGE-1	**40.93**	18.51	**37.19**
	ROUGE-2	40.77	**18.52**	37.07
	ROUGE-L	40.48	18.39	36.80

ROUGE-L 的成績表現，發現三者收斂之時間點不同。為了最大化 GALs 的優化效果，我們嘗試分別以驗證集上的 ROUGE-1、ROUGE-2 與 ROUGE-L 指標，參考其漲跌以自動決定早停法（Early Stopping）與學習率衰減的時機。在實驗於 CNN/Daily Mail 數據集之結果顯示（見表二），參考 ROUGE-1 能帶來最佳的整體 ROUGE 成績

（三）優化的句子數量

在 CNN/Daily Mail 數據集中，大多數文本的參考摘要只包含了三至四句話，對此 Yen-Chun Chen 設定抽取式摘要器固定抽取前三句話，以取得最佳的抽取式摘要成績。上述之設定依然皆適用在 GALs 的抽取式摘要器上，然而我們在實驗過程中發現（見表三），讓抽取式摘要器固定抽取前五個句子以做摘要層次的全局優化，能大幅提昇抽取式摘要器的表現。我們認為原因在於語料中有不少文本，其參考摘要由超過四個句子所組成，當我們將抽取之第四與第五句話也納入優化的對象中，能使抽取式摘要器對這些文本生成更優質之摘要，同時亦不會過度損傷抽取之前三句話的品質。

五、結論

我們將抽取式摘要問題視為整列式文本排序問題，提出 GALs：基於對抗式學習之整列

表三、優化句數與優化結果

優化句數	ROUGE-1	ROUGE-2	ROUGE-L
3	39.97	17.98	36.34
4	40.34	18.26	36.69
5	**40.55**	**18.39**	**36.89**
6	40.48	18.39	36.80
7	40.40	18.33	36.72

式摘要法，是一種對抽取式摘要器做摘要層次優化的方法。相較於曾是世界最佳結果之研究所用的基準系統，我們所提出之 GALs 能在 CNN/Daily Mail 數據集的抽取式與重寫式摘要結果上，所有 ROUGE 成績皆獲得明顯的提昇。最後，我們對 GALs 所使用的參數與特徵，進行細部的調查與分析，藉以觀察發想於資料檢索之研究而設計的 GALs，在優化抽取式摘要時所展現的特性與效能。

致謝

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 108-2636-E-011-005 (Young Scholar Fellowship Program), and by the Project J367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

參考文獻

[1] T.-Y. Liu, “Learning to rank for information retrieval,” Foundations and Trends in Information Retrieval, vol. 3, no. 3, pp. 225-331, 2009.

[2] P. Ren, Z. Chen, Z. Ren, F. Wei, J. Ma, M. de Rijke, “Leveraging contextual sentence relations for extractive summarization using a neural attention model,” Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 95-104, 2017.

[3] O. Vinyals, M. Fortunato, N. Jaitly, “Pointer networks,” Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 2, pp. 2692-2700, 2015.

[4] Y.-C. Chen, M. Bansal, “Fast abstractive summarization with reinforce-selected sentence rewriting,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 675-686, 2018.

[5] T. Mikolov, J. Dean, “Distributed representations of words and phrases and their compositionality,” Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, pp. 3111-3119, 2013.

[6] S. Hochreiter, J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[7] J. Wang, L. Yu, W. Zhang, Y. Gong, Y. Xu, B. Wang, P. Zhang, D. Zhang, “IRGAN: A minimax game for unifying generative and discriminative information retrieval models,” Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 515-524, 2017.

[8] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” Text Summarization Branches Out, pp. 74-81, 2004.

[9] I. Sutskever, O. Vinyals, Q. V. Le, “Sequence to sequence learning with neural networks,” Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 2, pp. 3104-3112, 2014.

[10] D. Bahdanau, K. Cho, Y. Bengio, “Neural machine translation by jointly learning to align and translate,” International Conference on Learning Representations, 2015.

[11] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, B. W. Suter, “The multilayer perceptron as an approximation to a Bayes optimal discriminant function,” IEEE Transactions on Neural Networks, vol. 1, no. 4, pp. 296-298, 1990.

[12] I. Guyon, J. Weston, S. Barnhill, V. N. Vapnik, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46, no. 13, pp. 389-422, 2002.

[13] M. Pagliardini, P. Gupta, M. Jaggi, “Unsupervised learning of sentence embeddings using compositional n-gram features,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 528-540, 2018.

[14] S. Narayan, S. B. Cohen, M. Lapata, “Ranking sentences for extractive summarization with reinforcement learning,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1747-1759, 2018.

[15] X. Zhang, M. Lapata, F. Wei, M. Zhou, “Neural latent extractive document summarization,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 779-784, 2018.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 25-34

結合類神經網路及文件概念圖之文件檢索研究

Document Retrieval based on Neural Network and Document Concept Graph

盧家馨　Chia-Hsin Lu
國立臺北科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taipei University of Technology
t106598005@ntut.edu.tw

王正豪　Jenq-Haur Wang
國立臺北科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taipei University of Technology
jhwang@csie.ntut.edu.tw

摘要

倘若搜尋結果能考慮主題或情境相近的內容，便能搜尋到更符合使用者期待的結果。因此，本研究使用類神經網路及文件概念圖，以探討主題或語意相近之檢索內容，實驗結果顯示，在經由類神經網路所訓練的分類器中，最佳 macro-F1 為 70.0%，而結合文件概念圖的計算後，以查詢內容與結果的相似度而言 nDCG 分數可達 0.959，由此可驗證，基於類神經網路及文件概念圖的結果可以補充和加強資訊檢索的表現。

Abstract

If the search results can consider topics or similar situations, we can find results that are more in line with user's expectations. Therefore, our research uses neural network and document concept graph to explore the topics or semantics similarity. The experimental results show that the best macro-F1 is 70.0% in the classifier trained via the neural network. Combined with the calculation of the concept graph of the document, the nDCG score can reach 0.959 in terms of the similarity between the search content and the results. This proves that the results based on the neural network and the document concept graph can be used to complement and enhance the performance of information retrieval.

關鍵詞：資訊檢索、類神經網路、文件概念圖、語意相似度

Keywords: Information retrieval, Neural network, Document concept graph, Semantic similarity

一、緒論

隨著資訊時代的到來，如何使檢索系統更符合使用者期待，是我們主要的研究主題。為了使檢索系統能搜尋出含有語意或主題相近的結果，我們探討如何取得文件的語意或者主題，本研究旨在增加檢索結果所考慮的因素，例如類別、概念、排序等，期望能提升檢索結果。

我們針對類神經網路預測未知資料之能力進行研究，並利用註釋工具取得概念(Concepts)，建立文件概念圖並探討圖形之間關係，最後我們結合類神經網路及文件概念圖計算，對不同搜尋內容使用不同算法之檢索結果進行討論。透過本研究方法，能經由類神經網路預測文件類別，並藉由文件概念圖的圖形結構關係，找尋具有類別資訊及概念相近之檢索結果，經過實驗驗證，本系統在檢索結果之排名上具有良好的效果，平均 nDCG 分數高於 0.9，此兩種特徵結合確實能輔助我們得到更符合使用者查詢內容的結果。

二、相關研究

（一）. 資訊檢索研究

了解使用者的需求並不容易，如何查詢到適合的結果，是資訊檢索(Information Retrieval, IR)領域持續在研究的其中一個方向。近年來資訊檢索領域逐漸朝著使檢索結果更符合使用者期待的方向發展，本研究嘗試探討檢索結果的相似度，討論何種計算方式能得到與查詢內容更相似的結果。

（二）. 文件分類研究

對於一般的監督式機器學習文件分類而言，訓練集的資料必須包含所有的類別，然而，此種假設在很多應用並不成立，我們無法確保資料都是系統曾經碰過的類型。此種問題被稱作 open world classification 或 open classification[1]，譯為開放式分類問題。

在開放式分類問題中，近年來已經存在一些可以辨別 unseen 類別的研究，例如[2-4]，Shu Lei[5]等人於 2017 年提出一種名為 DOC 的深度學習算法，經實驗發現，簡單的 CNN 模型在此種類型開放式分類問題上具有良好的效果。Hu Xu 等人[6]於 2019 年，參考了 DOC 架構，在電子商務產品分類進行實驗，解決了辨別新產品類別的問題。由上述研究來看，開放式問題已成為近年來探討議題之一。因此，本研究參考 DOC 架構，探討符合我們研究主題之類神經網路模型，詳細方法將在後續進行說明。

（三）. 文件語意研究

近年來實體、概念搜索已成為 Web 研究的一項重要任務，陸續有研究使用此種方式來探討文件語意對於資訊檢索領域之發展。Yuan Ni 等人[7] 於 2016 年提出了利用概念圖之文件表示，測量文件之間的語意相關性。文件使用多個概念節點(node)來表示，其節點為透過工具從文件中提取的概念。節點之間的邊(edge)代表概念之間的語意和結構關係。此概念圖使用 closeness centrality 對概念進行加權，該權重反映了它們與文件的相關性。Zhenghao Liu 等人[8]於 2018 年提出了一種 Entity-Duet Neural Ranking Model（EDRM），它將知識圖(Knowledge graph)引入神經搜索系統，通過其單詞和實體註釋表示查詢和文件，發現此種知識圖語意顯著提高了神經排序模型的泛化能力。而本研究參考上述研究方法，提出一個基於文件之間關係的文件概念圖，並利用此圖形關係計算文件相似程度。不同於過去方法，我們擷取出文件的多個概念來代表一個文件，並用一個節點來表示，而節點之間的邊代表兩節點之間有共同的概念。

三、研究方法

此章節說明研究的方法及架構，如圖一所示，後面章節將針對架構各模組進行說明。

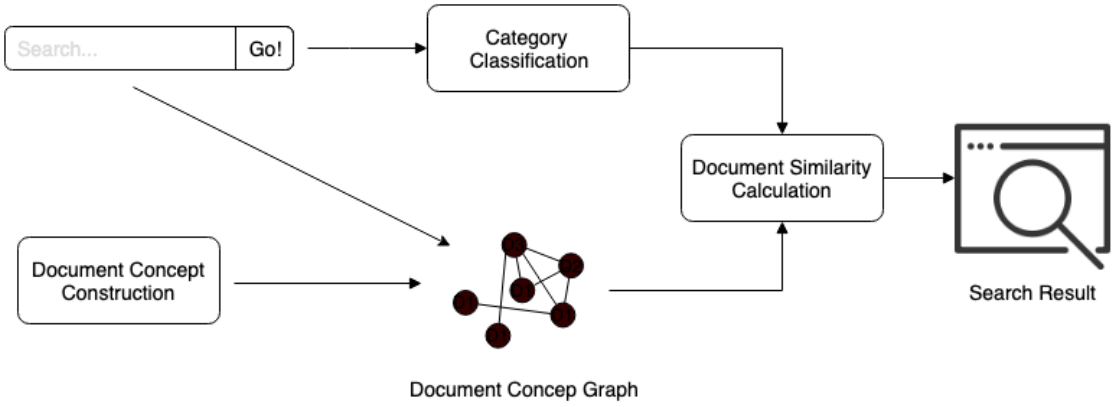


圖 一 、系統架構圖

（一）. Category Classification

本研究參考 DOC[5]利用卷積神經網路(Convolutional neural network, CNN)，此種類神經網路已被實驗證實，對於處理開放式情境問題也具有一定程度的能力，因此，本研究加入此模組，以協助我們探討文件相似度。

1. The Architecture of our CNN model

第一層 Embedding layer 將資料集 D 中的單詞直接 embedding 到密集向量中。近年來在自然語言處理領域上常用的方法會使用 Word2Vec[9]事先訓練一個字詞模型。字詞模型會仰賴不同資料集的特性產生不同向量，所以本研究希望可以不使用字詞模型轉換的方式，僅使用類神經網路的 Embedding layer 計算也能達到良好的效果。

第二層 Convolutional layer，使用不同大小的 filter 分別對密集向量進行卷積，filter 在神經網路中代表對應的過濾器，因此，此計算方式可以得到經由不同過濾條件計算後的結果。

下一步，我們使用 Max-over-time pooling layer 從 Convolutional layer 的結果中挑選最大值以形成 m 維度特徵向量 f，透過兩個 Fully connected layer 和一個中間$ReLU$ activation layer 將 f 降低維度到 n 維向量 x，最後輸出層是應用於 x 的 One-vs-rest 層，此部分在下一小節會做更詳細的描述。

2. One-vs-Rest Layer of CNN

傳統的多分類器使用$softmax$作為最後的輸出層，如此每一個類別的預測機率已經在訓練的時候進行了正規化，便少了彈性調整的能力。因此，我們$sigmoid$函式作為輸

出層，對應的類別採用所有正例的例子，而其餘剩下的例子皆作為反例。

（二）. Document Concept Construction

過去已有研究使用註釋、關鍵字、概念等方式來代表濃縮後的文件[10]。在[7]的研究方法中，提取文件的概念，將概念作為節點，邊代表兩個概念的不同連結關係，並將此概念圖來代表一個文件，計算概念圖的相似來代表文件相似度。而在本研究中，我們參考前述研究方式，將文件參考知識庫取得文件的多個概念，並用多個概念來代表一個文件節點，我們將概念相似的文件建立連結關係，建構文件概念圖，進而利用此圖來計算文件之間的相似度。文件表示為節點d_i，利用式 1 及式 2 計算權重$weight$並建立$edge$。我們使用無向圖(Undirected Graph)來建構文件概念圖，並利用相鄰陣列(Adjacency Matrix)來儲存權重($weight$)。

$$weight_{d_i d_j} = number\ \ of\ \ concepts\ \ shared\ \ between\ \ d_i, d_j \tag{1}$$

$$edge(d_i, d_j) = weight_{d_i d_j} \tag{2}$$

（三）. Document Similarity Calculation

此模組我們將結合分類器及文件概念圖來進行檢索相似度計算。建立圖形時，我們將中心點的概念加入文件概念圖，因為其權重代表兩節點相同概念個數，並且我們要得到該圖形結構中擁有最多概念資訊之節點，因此我們計算節點的 Degree Centrality。

1. Top_{class}計算

在此小節我們詳列Top_{class}的計算方法：

(1)、 將查詢內容輸入分類器，並載入預先訓練好的模型，對查詢內容進行類別預測。

(2)、 利用該類別至文件概念圖找到對應的類別中心點。根據前面對中心點的描述，我們依照鄰居節點多寡進行排序，排序後再挑選出前幾個節點，依據擁有的 weight 多寡再次進行排序，如式 3，找出最後經過兩次排序後最前面的節點，作為類別中心點。

(3)、 取得類別中心點後，我們找出其鄰居節點並依據 weight 高低進行排序，得到與目標節點相似度高的前幾篇文件$Result_C$，如式 4。

$$Center_{class} = (Max(\sum Neighbor(d_i)) \cap (Max(\sum weight(d_i))) \tag{3}$$

$$Result_C = Top_{class}(\ weight(d_i)\ |\ d_i \in Neighbor(Center_{class})) \quad (4)$$

2. Top_{query} 計算

此小節我們詳列Top_{query}的計算方法：

(1)、 為了取得整段查詢內容的向量表示，我們使用類神經網路訓練 Doc2vec[11]向量模型，使用該模型透過 word embedding 轉為向量，將查詢內容與文件概念圖的文件向量化。

(2)、 兩向量計算 similarity 找出相似度最高的目標節點，如式 5。

(3)、 取得目標節點後，我們找出其鄰居節點並依據 weight 高低進行排序，得到與目標節點相似度高的前幾篇文件$Result_Q$，如式 6。

$$Target(q) = argmax_i\ \ similarity(q, d_i) \quad (5)$$

$$Result_Q = Top_{query}(\ weight(d_i)\ |\ d_i \in Neighbor(Target(q))) \quad (6)$$

3. Top_{merge}計算

此小節我們列出Top_{Merge}的計算方法。

(1)、 利用Top_{Class}計算的類別中心點，及Top_{Query}計算的目標節點，找出兩者的所有的權重關係，合併成一個列表，也就是找出所有的鄰居節點之權重。

(2)、 將此具有權重關係之列表由大到小排列。

(3)、 得到權重最高的前幾篇文件，如式 7。

$$Result_M = Top_{merge}(\ weight(d_i)\ |\ d_i \in Result_C \cup Result_Q) \quad (7)$$

四、實驗與討論

（一）. 實驗資料集

本研究使用的資料集為 DOC[11]所使用的 20 Newsgroups。其內容收集了 18846 篇新聞文件，大致均勻分為 20 個類別；以及 Chen 等人[12]所使用的資料集 50-class reviews，其包含 50 種亞馬遜商品的評論，每一種有 1000 則，總共有 50000 則評論。

（二）. 文件分類實驗

為了模擬某一筆資料的類別並未出現在訓練集的類別，透過訓練好的模型之後，能夠被正確分類，我們將資料集類別分成 seen、unseen，訓練集資料為 seen 的類別，而測

試集則是所有的類別。

表一 為以 seen 類別搭配不同的資料及進行實驗，此結果為每一個模型重複執行五次後取平均值。其計算結果經過 paired t-test 檢定後，計算出來的 p-value 皆小於 0.01。從此表我們觀察到，每個資料集的 75%類別效果最好，到了 100%的效果卻降低，可能的推斷是資料量多寡會影響模型的效能，抑或是模型訓練 overfitting 的狀況。

表 一、$macro\ F_1 - score$ for datasets

% of seen classes	25%	50%	75%	100%
20 Newsgroups	0.55367	0.5948	0.62388	0.61624
50-class review	0.59783	0.67283	0.70014	0.64577

（三）. 基於文件概念建圖

我們使用於 2010 年提出的 TAGME[10]作為概念檢測工具，選擇此工具的原因是[13]研究中顯示，TAGME 是各種文檔類型性能最佳的註釋系統，我們列出不同實驗資料集所建構之概念圖的相關屬性，如表二 所示。

表 二、文件概念圖不同資料集屬性表

資料集名稱	資料集屬性
20 Newsgroups[14]	Node: 18,846 Edge: 128,330,600
50-class reviews[12]	Node: 50,000 Edge: 1,343,168,938

（四）. 檢索相似度計算實驗

此小節將分類器及概念圖此兩組模組進行整合計算，我們設計不同的計算方式，並對結果進行討論。後續實驗我們使用的各搜尋編號所對應的查詢內容如下：

1. Semantic Documents Relatedness using Concept Graph Representation.
2. Apple newest product launch
3. how about today's weather
4. convolution neural network
5. machine learning

編號 1 為論文的名稱，較偏學術用語；而編號 2 具有時間、公司名稱、內容資訊等描述；編號 3 為一般使用者日常查詢的內容及用語；編號 4 及編號 5 皆使用了近年來人

工智慧計算之相關用詞。

取得幾組不同算法的結果後，我們將查詢內容與第 i 個檢索文件利用 Doc2vec 向量模型，透過 word embedding 轉為向量，並計算 cosine similarity 作為該位置的相關係數 rel_i，如式 9，其中D_s為查詢內容，D_i為第 i 個檢索文件，最後我們計算正規化折扣累計獲益(Normalized Discounted Cumulative Gain, nDCG)分數。

$$rel_i = cosine\ \ similarity(D_s, D_i) \tag{9}$$

表 三、前 20 篇文件並使用 50%分類器之 nDCG 表

搜尋編號	$Result_C$	$Result_Q$	$Result_M$	Average
1	0.862	0.901	0.918	0.893
2	0.937	0.959	0.902	0.932
3	0.928	0.886	0.922	0.912
4	0.935	0.912	0.913	0.92
5	0.912	0.909	0.921	0.914
Average	0.915	0.913	0.915	

表三 顯示前三種算法取前 20 個檢索結果之 nDCG 分數，可以觀察到普遍的 nDCG 分數皆高於 0.9，可見此系統具有良好的檢索效果。

針對$Result_C$而言，編號 2 的效果最好，或許是此類型的文件得到的檢索結果彼此主題較相近，在取得搜尋結果時可以找到概念較相近的文件。而編號 1 的效果較差，推論是因為敘述較接近學術用語，在新聞資料的分類上效果較不好，另一方面，代表該類型的文件得到的搜尋結果彼此主題內容關係較弱。

下一步，我們觀察$Result_Q$，如同前一種算法，編號 2 的效果最佳，可見此類型的查詢內容，不管是分類器或者是文件概念圖的表現都較良好。而編號 3 在此處表現較差，代表該句子中的四個單字並不能很明確的取得其概念，所以得到的檢索效果較差。

第三種$Result_M$，利用Top_{Class}的類別中心點及Top_{Query}的目標節點，再合併計算，可以理解為以查詢內容直接使用文件概念圖所得到的結果，再藉由分類器所預測的類別輔助，加強檢索效果的可靠度，因此，我們期望$Result_M$的檢索表現大於$Result_Q$，而結果顯示，除了編號 2 之外，$Result_M$在其他編號的分數的確都大於$Result_Q$，而就總體平均值來看$Result_M$ 也比$Result_Q$高了 0.2%，因此，此實驗結果符合我們的期望。

編號 2 在個別$Result_C$及$Result_Q$的效果都比其他查詢內容好。

然而，以五個編號整體平均值來說，編號 2 的效果還是最佳的，可見雖然合併之後效果降低一些，但整體而言其查詢內容仍能檢索到較好的結果，而編號 1 則是效果最差的，代表其語句上的用詞在分類器及文件概念圖上的效果都表現得不是很好。

五、結論

本研究提出結合類神經網路模型及文件概念圖的計算得到檢索結果。在類神經網路的部分，經過實驗驗證，證明此分類器具有一定的分類效果，我們利用得到的類別與文件概念圖整合計算，得到第一組結果；而在文件概念圖的部分，本研究考慮知識庫取得文件的概念，並利用概念來代表文件，計算文件之間的概念相似以建立關係圖，計算時我們直接將查詢內容輸入文件概念圖，得到第二組結果；最後我們合併前面兩個模組在文件概念圖計算的目標點，經由計算得到第三組結果。最後討論三組計算方式對於不同查詢內容之效果，平均 nDCG 分數高於 0.9，證實本研究方法確實具有不錯的效果，得到的檢索結果包含類神經網路預測過的類別資訊，並涵蓋知識庫概念相似。

本研究所提出的方法及計算流程尚有許多改善空間，並且仍有一些限制。

我們使用的卷積神經網路仍有其他多種變形，網路層架構及參數調整更深入的探討或許能提升未知內容的分類效果。在深度學習中，還有很多類型的類神經網路可以進行實驗，或許也能增加文件分類的準確度。

本研究採用的概念檢測工具目前只支援三種語言，英語、德語、義大利語，因此，若是想取得其他語言文件的概念，必須抽換概念檢測工具進行實驗比較。倘若可以加入文件其他特徵，或許能改善文件概念圖的效果。

未來可以實驗更具有評估檢索效果能力之資料集，或者採用資訊檢索公開標準 benchmark 的資料，以多方探討此種結合方式之檢索效果。

參考文獻

[1] L. Shu, H. Xu, and B. Liu, "Unseen class discovery in open-world classification," in

6th International Conference on Learning Representations, 2018.

[2] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563-1572.

[3] G. Fei and B. Liu, "Breaking the closed world assumption in text classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 506-514.

[4] L. P. Jain, W. J. Scheirer, and T. E. Boult, "Multi-class open set recognition using probability of inclusion," in *European Conference on Computer Vision*, 2014: Springer, pp. 393-409.

[5] L. Shu, H. Xu, and B. Liu, "Doc: Deep open classification of text documents," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2911-2916.

[6] H. Xu, B. Liu, L. Shu, and P. Yu, "Open-world Learning and Application to Product Classification," in *The World Wide Web Conference*, 2019: ACM, pp. 3413-3419.

[7] Y. Ni *et al.*, "Semantic documents relatedness using concept graph representation," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016: ACM, pp. 635-644.

[8] Z. Liu, C. Xiong, M. Sun, and Z. Liu, "Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, vol. 1, pp. 2395-2405.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop Papers*, 2013.

[10] P. Ferragina and U. Scaiella, "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010: ACM, pp. 1625-1628.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188-1196.

[12] Z. Chen and B. Liu, "Mining topics in documents: standing on the shoulders of big data," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014: ACM, pp. 1116-1125.

[13] M. Cornolti, P. Ferragina, and M. Ciaramita, "A framework for benchmarking entity-annotation systems," in *Proceedings of the 22nd international conference on World Wide Web*, 2013: ACM, pp. 249-260.

[14] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine Learning Proceedings 1995*: Elsevier, 1995, pp. 331-339.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 35-46

室內遠距離語音辨識實驗

Experiments on In-House Far-Field Speech Recognition

邱炫盛 Hsuan-Sheng Chiu, 楊智合 Jyh-Her Yang

中華電信研究院
Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan

{samhschiu, houseyang0204}@cht.com.tw

摘要

近年來，語音辨識的應用推出各種遠距離操作的系統，例如車用語音助理、智慧音箱等，在這些系統中遠距離語音辨識扮演著關鍵的角色。本文主要提出我們在智慧音箱裝置上的遠距離語音辨識相關實驗及成果，我們利用資料擴充方式、模擬遠距離語音及基於類神經網路聲學模型來降低字元錯誤率。在實驗部分，本文利用智慧音箱錄製了三種距離的平行測試語料，其中 50cm 情境語料可從 13.31% 降至 8.41%，相對改善 36.8%，而 80cm 情境語料從 19.20%降至 10.89%，相對改善 43.2%。

Abstract

In recent years, speech recognition applications have introduced a variety of remote operating systems, such as car voice assistants, smart speakers, etc. In these systems, far-field speech recognition plays a key role. This paper mainly presents our experiments and results on far-field speech recognition on smart speaker devices. We use data augmentation methods, simulated far-field speech and neural network-based acoustic models to reduce the character error rate (CER). In the experimental part, this paper recorded the parallel test corpus of three distances using the smart speaker. The 50cm situation corpus can be reduced from 13.31% to 8.41%, the relative improvement is 36.8%, and the 80cm situation corpus is reduced from 19.20% to 10.89% with relative improvement of 43.2%.

關鍵詞：遠距離語音辨識，智慧音箱，類神經網路聲學模型，資料擴充

Keywords: far-field, smart speaker, neural network-based acoustic models, data augmentation

一、緒論

近年來，在引進深度類神經網絡(Deep Neural Network, DNN)於聲學模型訓練之後，自動語音辨識方面就取得了重大進展[1,2,3]，對於任一語音辨識領域，當提供足夠且具代表性的訓練語料時，DNN 就可以學習其聲學本質上的變異性，如：語者、性別、頻寬、環境等差異。然而，在一個真實室內的空間內，可以想像到遠距離語音辨識能夠讓我們的生活更佳便利，但是遠距離語音辨識仍然是一個具有挑戰性的問題[4]。

目前已經提出許多技術[4,5,6]來處理遠距離語音辨識問題，其中最有效的作法就是資料擴充(Data Augmentation)，它讓 DNN 有機會可以學習到真實環境中可能遭遇到的情況，只要 DNN 能夠學習到真實環境的聲學特徵，在訓練及測試條件匹配之下就會有不錯的效果。而收集大量現實各種樣式的環境語料很耗費大量人力及金錢，因此，利用模擬方法產生訓練語料是一種可行的選擇，藉由模擬各種樣式環境語料進行資料擴充，對於強健聲學模型上能得到非常顯著的效果，可從 IARPA-ASPIRE 遠距離辨識競賽[7,8]中看到使用資料擴充方法得到最大相對 33%的改善效果。

本文主要分享利用前述作法於室內智慧音箱上的遠距離語音辨識實驗，使用模擬的空間脈衝響應(Room Impulse Response, RIR) 捲積於語音信號中來表示一個空間的殘響(Reverberation)，藉由調整不同參數可以改善智慧音箱上的遠距離語音辨識問題。

本文將在接下來的第二節會介紹本實驗所使用的方法，在第三節將會描述本次實驗在真實音箱於不同距離下的結果，最後結論會放在第四節作說明。

二、實驗方法

（一）模擬空間調整

在 Ko[6]的論文中，主要透過鏡像法產生了 3 種 RIR，其產生方式為先依照主要距離限制，例如長寬 1~10 公尺內，高 2~5 公尺內，亂數產生 200 組空間參數，這些參數包含了空間的長寬高、接收端位置與空間的吸收係數，再限制聲音來源端與接收端距離為 0 到 5 公尺，亂數產生 100 組來源端位置。最後每一種 RIR 各有 20000 組模擬結果。另外，Ko 也整理了實際空間錄製的 RIR，共 325 筆來進行相關實驗，所以總共有 4 種類型的 RIR，此資料集稱為 Room Impulse Response and Noise Database（以下稱 RIRND），

表一、RIRND 資料集

參數	設定值
RIR_small	空間長寬 1~10 公尺，高度 2~5 公尺，吸收係數 0.2~0.8，聲音來源與接收位置距離為 5 公尺內，亂數產生共 20000 組
RIR_medium	空間長寬 10~30 公尺，高度 2~5 公尺，吸收係數 0.2~0.8，聲音來源與接收位置距離為 5 公尺內，亂數產生共 20000 組
RIR_large	空間長寬 30~50 公尺，高度 2~5 公尺，吸收係數 0.2~0.8，聲音來源與接收位置距離為 5 公尺內，亂數產生共 20000 組
RIR_real	由三套資料集組成： RWCP sound scene database: 1 個實際空間，長度為 6.66 公尺，寬度為 4.18 公尺，聲音來源與接收距離為 2 公尺，響應時間為 0.3~1.3 秒，共 182 組 REVERB challenge database: 3 個模擬空間，響應時間分別為 0.25 秒、0.5 秒與 0.7 秒，2 種聲音來源接收位置距離分別為 0.5 公尺與 2 公尺；1 個實際空間，響應時間 0.7 秒，2 種聲音來源接收位置距離分別為 1 公尺與 2.5 公尺，共 36 組 Aachen impulse response database: 4 個實際空間，空間分別為 3 x 1.8 x 0.5、5 x 6.4 x 2.9、8 x 5 x 3.1、10.8 x 10.9 x 3.15 公尺，聲音來源接收距離分別為 0.5, 1, 1.5 公尺、1, 2, 3 公尺、1.45, 1.7, 1.9, 2.25, 2.8 公尺、4, 5.56, 7.1, 8.68, 10.2 公尺，共 107 組

詳細資料如表一所示。然而，我們主要的應用是在家裡的客廳內使用智慧音箱，使用 RIRND 資料集未必完全符合我們的需求，所以我們嘗試設計以客廳為空間大小的參數以符合使用情境。我們先依照客廳的可能範圍亂數產生空間大小 200 組，再亂數產生 100 組聲音來源端（即說話者）及接收端（即音箱）位置與殘響時間等參數，同樣總共有 20000 組結果（以下稱 RIR_exp）。說話者與音箱距離的產生方式是先將空間平面劃分成 5 x 5 的區塊，由左上至右下編號為 0 至 24 號，並假設音箱可能放在電視旁邊（編號 1、3）、客廳中間桌子（編號 12）或是空間四邊角落（編號 0、4、20、24），其他編號為可能的說話者位置。至於其他參數如反射次數， RIRND 設定為 10，我們則不限制；收音方式我們與 RIRND 相同假設為全指向性；殘響時間參數部分，RIRND 是使用吸收係數，且假設空間平面皆相同，其殘響時間則可透過 Sabine 公式換算得出，簡單地說，吸收係數越高，殘響時間越短，我們則是參考 REVERB 2014 設定為 0.25~0.7，我們設定為 0.2~0.6。表二是我們的參數詳細設定，我們亦進一步統計，以我們的方法產生出的模擬結果，其說話者與音箱的距離大部分會分佈在 2~3 公尺的距離，其次則是 1~2 公尺及 3~4 公尺，如表三所示。

表二、空間脈衝響應模擬參數設定

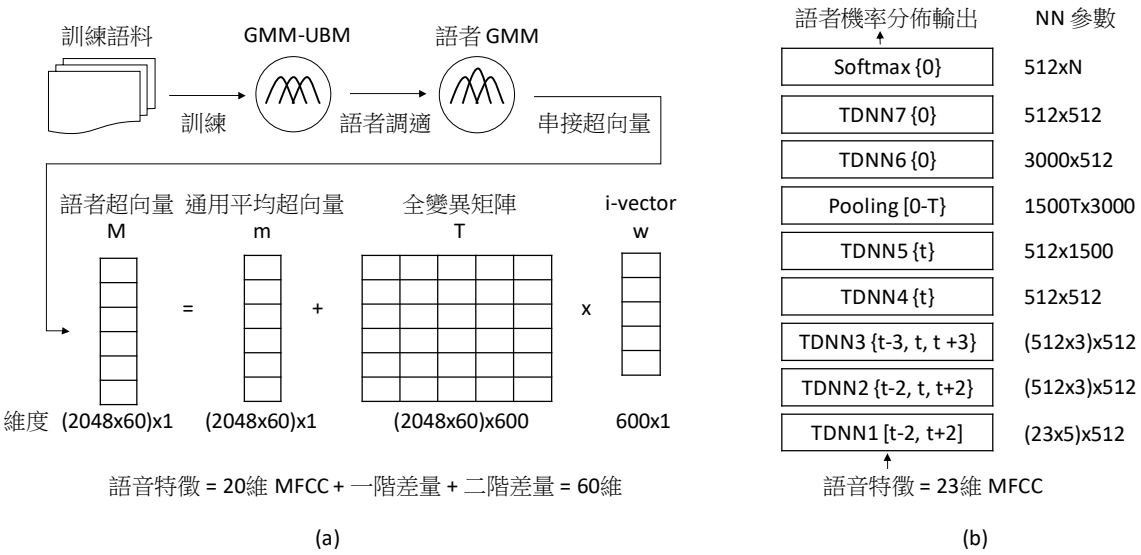
參數	設定值
空間長度 (公尺)	[3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0]
空間寬度 (公尺)	[3.0, 3.4, 3.8, 4.2, 4.6, 5.0]
空間高度 (公尺)	[2.4, 2.6, 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0, 4.2]
說話者高度 (公尺)	[0.9, 1.1, 1.3, 1.5, 1.7]
音箱高度 (公尺)	[0.4, 0.6, 0.8, 1.0, 1.2, 1.4]
音速 (公尺/秒)	340
接收端指向性	全指向
響應時間 (T60)	[0.2, 0.3, 0.4, 0.5,0.6]
反射次數	無限制
說話者位置編號	[10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]
音箱位置編號	[0, 1, 3, 4, 12, 20, 24]

表三、說話者與音箱距離分佈

說話者與音箱距離	筆數	分佈 (%)
0~1 公尺	1871	9.355
1~2 公尺	5869	29.345
2~3 公尺	6774	33.870
3~4 公尺	4347	21.735
4~5 公尺	1139	5.695

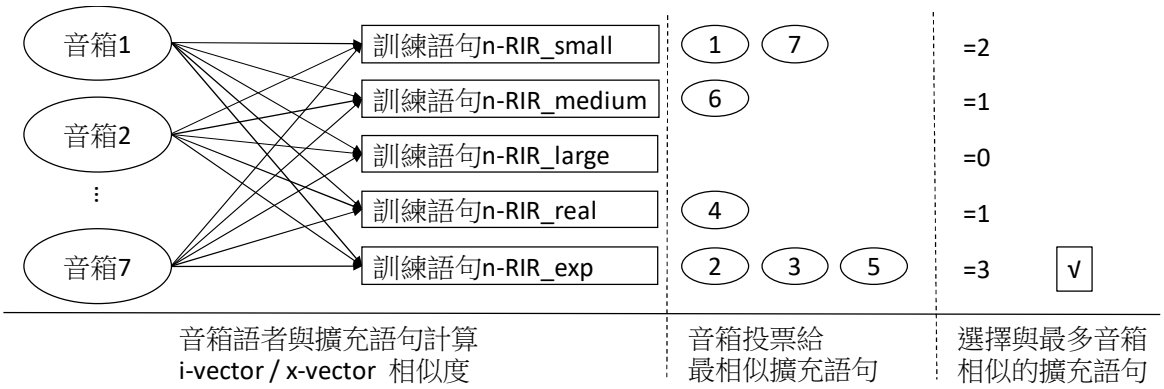
（二）訓練資料選擇

我們已經透過模擬 RIR 方式來產生遠距離的訊號殘響，然而對於我們目標的客廳環境音箱辨識仍不一定一致，尤其原始資料大多來自於手機或麥克風。如我們所知，如果選擇後的訓練資料與測試環境或設備來源一致，與使用全部資料相比，不僅能加快訓練速度，甚至可能得到相同或更好的結果。聲音資料選擇的方法有許多種，大致上可分成依據語言文字相關性，或是依據聲學特性相關性來決定[9]。於本論文中，我們初步嘗試使用 i-vector 及 x-vector 進行資料選擇。I-vector 是近幾年內語者辨識技術上標準的先進方法，其主要是透過以 GMM-UBM (Gaussian Mixture Model-Universal Background Model) 建立訓練語者的超向量(Supervector)，再進一步訓練全變異(Total Variability)矩陣。不同於 JFA (Joint Factor Analysis)分別對語者及通道建模，這個矩陣會同時包含語者與通道的資訊，並使用此矩陣建立目標語者與測試語句的 i-vector，最後再透過 LDA (Linear Discriminant Analysis)降維與 PLDA (Probabilistic Linear Discriminant Analysis)計算相似



圖一、(a) i-vector (b) x-vector 示意圖

度。X-vector 則是最近被提出用在語者辨識上有更佳的效果[10]。X-vector 主要是在 DNN 裡加入時間池化(Temporal Pooling)層，讓音框(Frame)層次資訊轉換成音段(Segment)層次資訊，訓練完成後，輸入語音直接使用池化層的下一層網路輸出當作 x-vector，並可以沿用 LDA 與 PLDA 進行相似度計算；x-vector 也可再透過模型架構調整引用更多資訊，例如音素資訊[11]。此外，i-vector 與 x-vector 皆可以透過資料擴充方式來提升準確度。圖一是 i-vector 與 x-vector 的模型示意圖，i-vector 主要是以 GMM-UBM、特徵空間與 EM 演算法進行訓練，而 x-vector 則是以 NN 模型為訓練架構。除了語者辨識，Siohan [12]也將 i-vector 用來作資料選擇。Siohan 把目標資料集的 i-vector 分佈當作其特性，並透過一句或多句批次的方式，分別計算批次加入前與加入後與目標資料集分佈的 KL 距離，如果距離縮小，則加入，反之則不加入。而我們初步的作法則是以分類與投票方式，先使用原始資料訓練 i-vector 及 x-vector，再準備一部分音箱語料當作發展集，並對發展集與所有擴充資料抽取 i-vector 或 x-vector 特徵後，計算其相似度，再針對每一句訓練語句，選擇與發展集語者有最多且最大相似度的擴充版本加入。訓練資料選擇流程如圖二所示，假設音箱為 7 個，訓練語句 n-RIR_xxx 為第 n 句訓練語句經過 xxx 方式處理的擴充版本。i-vector 與 x-vector 兩者處理語者的聲學特徵方式截然不同，參數量也有很大的差異，雖然主要都是用來區分語者，但是以我們的資料選擇方式來說，對於同一句訓練語句使用不同的擴充方式都是相同語者，主要是希望透過這種方式，嘗試選出與音箱特性最相似的擴充語句。



圖二、訓練語料選擇流程示意圖

（三）實驗設定

我們的訓練語料來自商業採購、專案語料以及網路上各類開源語料的集合，共有 2461 小時（以下稱 STT2461），其中來源設備包含了麥克風、手機與廣播等裝置；錄音樣式包含了腳本錄音與口語對話；情境包含了新聞、各類命令、各類書籍語句念稿與主題式聊天等方式；語言則包含了中文、英文與台語三種語言，其中以中文為主，約佔 76.8%，其餘為 18.8%及 4.4%。我們另外從訓練語料中取出特定資料集，共 507 小時進行實驗（以下稱 K360）。測試語料則是主要來自不同廠商的音箱設備，並包含在辦公室環境錄製的腳本錄音以及用戶實際環境體驗測試。錄製腳本的應用領域包含了各種資訊查詢，如電視頻道、節目、電影、歌曲、廣播、有聲書、股票、店家、天氣、路況等，並搭配多種前綴語或後綴語，如「我要看 OO 台」、「我想聽 OO 的 OO」、「OO 股價多少」、「最近的 OO 在哪裡」、「OO 有沒有下雨」等；或是語音命令，如設定鬧鐘、行事曆、訂車票、設備控制，如「設定 O 點 O 分的鬧鐘」、「打開客廳冷氣」等。其他資訊如表四所示，其中除了 IBPH3 是包含麥克風與不同距離音箱對齊的語料外，其他則都是不同廠商的音箱錄音語料。

我們使用 Kaldi 工具[13]進行相關實驗，採用的聲學模型架構為 TDNN-F[14]，網路層數為 11 層，每一層維度為 1280 維，SVD 分解維度為 256 維，模型架構主要參考 kaldi/egs/swbd/s5c/local/chain/tuning/run_tdnn_7n.sh，語音特徵使用 40 維 MFCC、3 維 PITCH 及 100 維 i-vector 進行訓練。進行實驗時，原始模型的 epoch 設為 4，加入擴充資料時則設為 2。進行訓練資料選擇實驗也是使用 Kaldi 抽取 i-vector 與 x-vector，我們

表四、發展與測試語料集

編號	IBPH3	IBD00	IBQC5K	IBDEV	IB0515
類型	測試	測試	測試	發展	測試
句數	2056	15546	5047	2960	7011
小時數	1.21	11.44	4.56	2.77	6.46
平均秒數	2.13	2.65	3.25	3.38	3.32
收音裝置	麥克風/音箱	音箱	音箱	音箱	音箱
環境	辦公室 20/50/80 公分	辦公室 80 公分	實際環境	實際環境	實際環境
類型	腳本錄音	腳本錄音	腳本錄音	實際用戶	實際用戶
音箱廠商	A	B	B: 1512 句 C: 3535 句	B	B: 1623 句 C: 3226 句 D: 2162 句

參考 kaldi/egs/sre16/v1 與 kaldi/egs/sre16/v2 設定，i-vector 為 600 維，x-vector 為 512 維，但皆不做資料擴充步驟。方法實驗的聲學模型語料使用的是 K360 訓練集與 IBDEV 發展集。語言模型部分，我們使用以中文網頁跟新聞為主的 3-gram 背景模型，再將各種應用情境訓練語料整合訓練出一個應用導向的 5-gram 語言模型，最後透過內插法將兩個模型合併，初步設定背景模型比重為 0.3。此外，詞典大小為 102 萬詞，腳本測試集語言複雜度(Perplexity)約為 1087，實際用戶測試集約為 1388。然而，由於詞彙量大，測試句之間的複雜度差異也很大，從數十到上萬皆有。

三、實驗結果

（一）模擬空間調整結果

我們首先測試加入不同 RIR 進行訓練的效果，其中實驗結果皆以字錯誤率(Character Error Rate, CER)為評估標準，其結果如表五、表六所示。由於商業因素，實際用戶的測試結果會用正規化後的錯誤率表示，即是將錯誤率除以基準錯誤率。首先，顯而易見地，當距離越遠，其辨識效果明顯降低，如 IBPH3 未加入擴充資料的 CER 從 5.69%增加至 19.20%；而如果比較未加入擴充資料與加入後的效果，皆能有所提升，且遠距離的改善更明顯，例如 IBPH3-50cm 可從 13.31% 降至 8.41%，相對改善 36.8%，IBPH3-80cm 可

表五、空間模擬測試 CER 結果 1

K360 加入不同 RIR	IBPH3-mic	IBPH3-20cm	IBPH3-50cm	IBPH3-80cm
NO_RIR	5.69	7.49	13.31	19.20
RIR_small	4.65	6.45	8.50	10.98
RIR_medium	4.78	6.24	8.83	11.79
RIR_large	4.99	6.29	9.64	12.27
RIR_real	5.02	5.99	8.78	11.67
RIR_exp	4.92	5.87	8.41	10.89

表六、空間模擬測試 CER 結果 2

K360 加入不同 RIR	IBD00	IBQC5K	IBDEV	IB0515
NO_RIR	10.53	11.34	1.000	1.000
RIR_small	8.57	8.44	0.863	0.982
RIR_medium	8.53	8.43	0.903	0.974
RIR_large	8.91	8.89	0.909	0.976
RIR_real	8.65	9.05	0.919	0.961
RIR_exp	8.52	8.49	0.928	0.990

從 19.20%降至 10.89%，相對改善 43.2%。而如果我們觀察自行模擬的 RIR_exp，在大部分的測試上都能有改善，但令人意外的是實際環境加上實際用戶的發展與測試集，雖然有改善，但是改善幅度並不大。然而，使用其他類型的 RIR 也是類似情況，觀察結果其主要原因是實際環境除了有遠距離收音問題外，還有背景雜訊干擾如電視聲的問題以及語言模型涵蓋率不夠，如新歌曲、新電影等 OOV(Out-of-Vocabulary)的影響。

（二）訓練資料選擇結果

接著我們呈現訓練資料選擇後的實驗結果，如表七、表八所示。可以觀察到，不管是使用 i-vector 或是 x-vector，效果大部分都會比亂數選擇來得好；而使用 x-vector 與 i-vector 則是互有高低，但差異不大。然而，如果與表五、表六使用單一 RIR 的情況相比，資料選擇後只有在 IBPH3-50cm 的情況下有改善，其他則是沒有改善。我們嘗試統計資料選擇後的 RIR 分佈，如表九所示，亂數是平均分佈，i-vector 以 RIR_exp 與 RIR_real 的比例為最多，x-vector 則是以 RIR_exp 與 RIR_small 為最多。如果 i-vector 與 x-vector 能反應出語音的語者或聲學特性，根據這樣的分佈來看，表示以 RIR_exp 擴充的語料應該是與實際環境最相似；若以 IBDEV 發展集與 IB0515 測試集的單一 RIR 結果來看，的確以 RIR_small 與 RIR_real 擴充的效果最好。資料選擇的效果不如預期的原因或許是因為我們初步資料選擇後僅是讓不同 RIR 之間的混合比例不同，雖然這樣可以確保與單一

表七、訓練資料選擇 CER 結果 1

K360 加入不同 RIR	IBPH3-mic	IBPH3-20cm	IBPH3-50cm	IBPH3-80cm
NO_RIR	5.69	7.49	13.31	19.20
RIR_rand	5.25	6.15	8.67	11.35
RIR_ivector	4.90	6.12	8.41	11.32
RIR_xvector	4.99	6.10	7.95	11.28

表八、訓練資料選擇 CER 結果 2

K360 加入不同 RIR	IBD00	IBQC5K	IBDEV	IB0515
NO_RIR	10.53	11.34	1.000	1.000
RIR_rand	8.54	8.83	0.926	0.967
RIR_ivector	8.55	8.44	0.907	0.983
RIR_xvector	8.60	8.46	0.902	0.978

表九、訓練資料選擇比例分佈

混合 RIR	small	medium	large	real	exp
random	20.01%	19.93%	20.07%	19.92%	20.07%
i-vector	20.16%	12.10%	19.66%	22.43%	25.65%
x-vector	29.87%	10.62%	2.98%	11.51%	45.02%

RIR 的語料數量大小一致，但並沒有使用相關性分數來選擇出最相似或是篩選掉不相似的資料，反而不如使用單一 RIR 的效果，未來可嘗試調整成將所有擴充資料依分數排序後再選擇出不同數量進行實驗。

（三）增加訓練語料結果

最後我們呈現使用全部語料 STT2461 進行訓練的結果，如表十、表十一所示。在這個結果中，訓練的 epoch 設為 6，其他的模型設定與方法實驗相同。除了使用更多語料外，我們也加入了加減速以及加雜訊兩種語料擴充方法，加速與減速分別為 1.1 倍與 0.9 倍，雜訊則使用 MUSAN 資料集[15]，包含一般噪音、音樂及人聲，SNR 分別為 0~15dB、5~15dB 及 13~20dB，同時我們也呈現了 Google 辨識結果。首先，我們以經過加減速與雜訊的結果為基準，累積加入不同 RIR 的訓練，雖然訓練時間增加很多，但在每一個測試集都有改善，讓模型更強健。我們的模型效果在腳本錄音的表現比 Google 好，例如 IBD00 與 IBQC5K 的結果，主要是因為較多特別領域詞彙，例如頻道名稱、節目名稱或歌曲名稱等，透過語言模型的調整，準確率更高。但是在實際環境與實際用戶下，例如 IBDEV 與 IB0515 的結果，Google 表現較好，除了實際用戶比較少說出腳本中的特殊詞彙之外，我們也觀察到主要是我們的模型較無法正確地拒絕背景人聲干擾，導致插入錯

表十、增加語料測試 CER 結果 1

STT2461	IBPH3-mic	IBPH3-20cm	IBPH3-50cm	IBPH3-80cm
Base + sp + noise	4.85	8.46	11.00	17.82
+RIR_small + RIR_medium	4.60	6.79	8.85	13.45
+RIR_large + RIR_real	4.65	6.79	8.41	11.67
+RIR_exp	4.62	6.82	8.11	11.51
GOOGLE	12.62	15.95	19.55	24.24

表十一、增加語料測試 CER 結果 2

STT2461	IBD00	IBQC5K	IBDEV	IB0515
Base + sp + noise	9.04	7.76	0.833	0.884
+RIR_small + RIR_medium	8.02	7.87	0.813	0.871
+RIR_large + RIR_real	7.92	7.65	0.777	0.847
+RIR_exp	7.85	7.26	0.762	0.855
GOOGLE	12.16	8.28	0.667	0.782

表十二、不同廠商音箱測試 CER 結果

IB0515	廠商 B	廠商 C	廠商 D
Base + sp + noise	0.882	0.907	0.835
+RIR_small + RIR_medium	0.824	0.916	0.804
+RIR_large + RIR_real	0.800	0.887	0.801
+RIR_exp	0.832	0.898	0.775
GOOGLE	0.939	0.692	0.877

表十三、測試 CER 錯誤分佈

測試-模型	插入	刪除	取代
IBD00-STT2461+RIR_exp	11.95%	22.08%	65.97%
IBD00-GOOGLE	3.21%	65.01%	31.78%
IB0515-STT2461+RIR_exp	25.67%	25.15%	49.18%
IB0515-GOOGLE	1.70%	86.14%	12.16%

誤較多。另外，也觀察到 Google 在音箱錄音品質稍差的情況下，容易出現沒辨識結果的情況，刪除的錯誤較多。我們也呈現不同廠商音箱在實際用戶環境的結果，如表十二所示，各種廠商的音箱在模型加入擴充語料後，皆有改善；廠商 B 與 D 的效果已經能跟 Google 相比，而廠商 C 的效果較差，觀察主要是某些使用者的設備環境常出現背景聲造成插入錯誤，並非設備問題，這也是整體 CER 比 Google 略差的主要原因。我們也列出 CER 錯誤在辦公室腳本錄音與實際環境的分佈情況，如表十三所示，可以看出 Google 的錯誤大部分是刪除錯誤，而從辦公室環境到實際環境，我們的模型插入錯誤比例也從 11.95%增加到 25.67%，但因為實驗並未加上信心度估測的機制，相信加入後能有所改善。

四、結論

我們實驗了音箱在不同距離及實際情境下的表現，並嘗試使用 i-vector 與 x-vector 來進行資料選擇，我們亦嘗試加入更多的訓練資料及其他資料擴充方式來進行訓練。實驗結果說明，不同音箱由於硬體不同，錄製的聲音品質也有差異，但使用更多語料及模擬方式，在各種音箱上也能有所改善。由於目前訓練語料仍是以手機及麥克風為主，未來將嘗試使用轉移學習(Transfer Learning)相關方法[16]，將音箱錄製語料與現有語料進行結合，或是當音箱語料收集足夠後，可以直接用來訓練更符合音箱特性的聲學模型。

參考文獻

[1] A. Mohamed, G. Hinton, and G. Penn, “Understanding how deep belief networks perform acoustic modelling”, in Proc. ICASSP, 2012, pp. 4273–4276.

[2] D. Yu., M. L. Seltzer, J. Li, J.-T. Huang , F. Seide, “Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks”, in Proceedings of International Conference on Learning Representations, May 2013.

[3] V. Peddinti, D. Povey, S. Khudanpur. “A time delay neural network architecture for efficient modeling of long temporal contexts”, In: 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, Dresden, 205.

[4] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, et al. “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research”, EURASIP Journal on Advances in Signal Processing, 2016.

[5] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno. “Blind separation and dereverberation of speech mixtures by joint optimization”. IEEE Transactions on Audio, Speech, and Language Processing, 19(1):69–84, 2011.

[6] T Ko, V Peddinti, D Povey, ML Seltzer, S Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition”, in Proc. ICASSP, 2017.

[7] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, “Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms,” in Proceedings of 2015 IEEE

Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015, pp. 539–546.

[8] R. Hsiao, J. Ma, W. Hartmann, M. Karafi, I. Sz, J. Honza, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermansk et al., "Robust speech recognition in unknown reverberant and noisy conditions," in Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015, pp. 533–538.

[9] KARAFIÁT Martin, VESELÝ Karel, ŽMOLÍKOVÁ Kateřina, DELCROIX Marc, WATANABE Shinji, BURGET Lukáš, ČERNOCKÝ Jan and SZŐKE Igor. "Training Data Augmentation and Data Selection," in New Era for Robust Speech Recognition: Exploiting Deep Learning. Heidelberg: Springer International Publishing, 2017, pp. 245-260.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in Proc. ICASSP, 2018.

[11] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," in Proc. Interspeech 2018, pp. 2247–2251.

[12] Olivier Siohan and Michiel Bacchiani, "iVector-based acoustic data selection," in Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech), Lyon, France, 2013.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The kaldi speech recognition toolkit," in IEEE 2011 workshop on automatic speech recognition and understanding, no. EPFLCONF-192584. IEEE Signal Processing Society, 2011.

[14] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in Proc. Interspeech, 2018.

[15] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," arXiv, 2015.

[16] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in Proc. ASRU, 2017

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 47-61

基於 BERT 模型之多國語言機器閱讀理解研究

Multilingual Machine Reading Comprehension based on BERT Model

吳承軒 Cheng-Xuan Wu
國立臺北科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taipei University of Technology
t106598068@ntut.edu.tw

王正豪 Jenq-Haur Wang
國立臺北科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taipei University of Technology
jhwang@csie.ntut.edu.tw

摘要

在網路資訊爆炸的現代，人們的生活與網路已密不可分，但受限於檢索技術的瓶頸，雖然能提供多方面的資訊來源，卻不一定是最相關有幫助的資訊。自然語言中的兩個主題：機器問答(Question Answering)與機器理解(Machine Comprehension)，由於對檢索系統，以及服務轉型中重要的聊天機器人，都具有高度相關，因此成為近年熱門的研究議題。本論文使用了 Google BERT 的 pre-trained model 進行詞嵌入向量，以單詞及單字為單位，組織出一個句子的特徵。並且基於問題、答案、與文本間不同組合的答題策略，最終選擇最高餘絃相似度的選項，作為機器作答的依據。本論文分別實驗在英文 TOEFL-QA 資料集，以及中文開放性問答資料集，對比於雙向 GRU 以及 A Strong Alignment IR Baseline 的方法，分別取得 34.87%及 57.5%準確率，實驗結果顯示，雖然不同語言之間具有文法的差異，但本論文所提的方法具有一定程度多國語言的通用性。

關鍵詞：閱讀理解、機器問答、自然語言處理、深度學習

Abstract

In recent years, Internet provides more and more information for people in daily life. Due to the limitation of information retrieval techniques, information retrieved might not be related and helpful for users. Two research topics in natural language processing have attracted much attention due to the important applications of information retrieval and chatbot in the past few years: question answering and machine comprehension. In this paper, we use Google BERT pre-trained model as a word embedding model to form semantic sentence features based on single words and phrases. Based on different strategies for question answering, we use cosine similarity to calculate similarity and choose the option of highest cosine similarity score as machine inferenced answer. In our experiments on TOEFL-QA dataset for English and Formosa Grand Challenge dataset for Chinese, our proposed method was compared with Bi-directional GRU and a strong alignment IR baseline, and obtained an accuracy of 34.87% and 57.5%, respectively. With the grammar difference between difference language, our model is capable of processing multilingual questions with comparable performance to existing methods.

Keywords: Reading Comprehension, Question Answering, Natural Language Processing, Deep Learning

一、緒論

人們每天的生活，與許多的電子產品緊密相連，透過這些便利的工具，有助於改善生活品質。在過去想要探究一個知識，必須查閱大量的書籍，並耗費許多的時間消化了解，才能從中找出想要的資訊；網路的普及與檢索技術的發展，使得人們能透過網路

搜尋服務，快速的從海量的資訊中，得到初步篩選過的結果，節省了閱讀與理解知識的時間。隨著資訊服務不斷的改善人們的生活，不論是社群平台，亦或是購物網站等，皆開始導入聊天機器人，相較於傳統上透過人力行銷及人力客服，除了減少部分人力的成本，更能夠增加與科技之間的互動性，提升黏著度因而增加提升商業產值；科技與人們的生活已經無法分離，進而促使了科技助理的誕生，為了能夠更進一步，滿足人們更多日常生活中的需求，需要讓這些不同類型的機器人，能夠更加了解人們真正所需，理解人們心中所想。傳統機器人與檢索技術的運作方式，大多採用關鍵字為主的方法，若查詢的語句中，與預設設置好的搜索資料相符，則將此資料當成最終的結果；在這種方式下，缺乏考慮所有其他非關鍵字詞的語句，因此可能會遺失真正重要的字詞資訊。為了要能夠更加有效的利用，一個查詢與答案的所有資訊，本論文使用字與詞來組成一句話所代表的真正涵義，並透過計算查詢中所有的句子，與答案句子之間關聯的強弱，來理解不同語句之間真正相似的程度。本論文對比了傳統基於關鍵字的問答方法，以及基於深度學習的方法，不同於大多數問答的研究，通常只專注在某個語言，本論文提出二個架構簡單，且具有一定效果的多國語言答題模型。

二、相關研究

（一）、語言模型

在自然語言處理的研究中，文字處理是一個重要的步驟，要讓機器進行下一步的計算，需要將現實世界中詞語的意義，轉換為代表它們的向量數值，因此這個階段代表了特徵的好壞。現實中使用文字的情境，常常某些詞語會與其他詞一同出現，或者出現在一篇文章的上下文，2013 年 Mikolov 等人[1]基於此特性，設計了一個詞嵌入向量模型，用來表示詞與詞之間的關係。進一步提升語言模型的表現，基於 Vaswani 等人[2]的架構，Devlin 等人[3]建構了一個 12/24 層的 BERT 語言表示模型，BERT 模型在英文問答競賽 SQuAD1.1 中贏過了 BiDAF+ELMO[14][20]、QANet[15]等神經網路的方法，驗證了一個良好的語言表示模型，能夠達到與複雜的神經網路架構相當的水平。

（二）、文字相似性

為了要讓機器更進一步具有理解語意的特性，一個初步需要被探討的問題為：如何找出文字與文字之間，是否具有相似語意，進而使機器能夠以此為基準。一種常見的作法為餘絃相似度(Cosine Similarity)，如 Gomaa 等人[4]的研究中所提到，透過餘絃相似度，使機器判斷兩種具有類似特徵的文章是否相似，更進一步的拓展應用，如 Huang 等人[5]、Karypis 等人[6]的研究中，將餘絃相似度使用在分群不同文章。

（三）、機器理解之選擇題研究

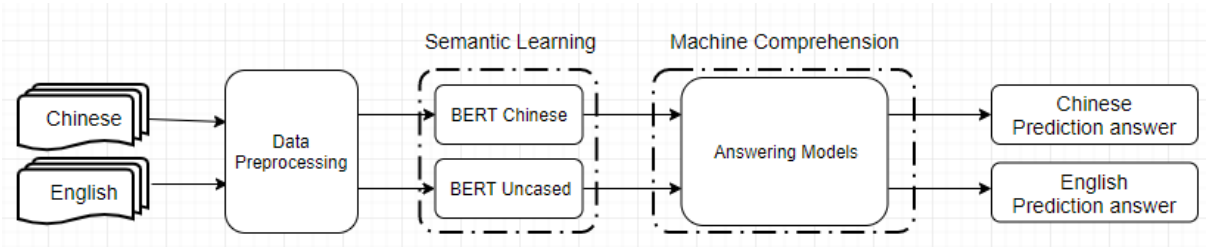
概觀為，給予一段故事、問題、幾個不同的選項，需要透過這些的線索，使機器找到答案，而其中的研究類型大致上可分為兩類，第一類為單選題問答，如中文科技大擂台競賽單選題問答[18]的研究，英文 Tseng 等人[7]的研究，以及第二類，複選題問答如 Richardson 等人[9]的研究。透過雙向 GRU 模型快速的學習故事與問題之間上下文的語意關係，但只在最後階段將雙向 GRU 模型的結果與選項做相似度計算，會喪失將選項與故事及問題一同考慮語意關係的面向，使用 Sliding Window 及 Distance Based 的方法，考慮故事中的詞彙，出現在問題及選項詞彙聯集的數量，但缺乏考慮沒出現詞彙的語意線索。

（四）、機器理解之尋找答案段落研究

概觀為，給予一段故事以及問題，根據問題找出故事中正確答案的段落，如 Yadav 等人[10]基於檢索的角度計算答案與問題的相似度，但依賴於 Embedding 模型的好壞，以及英文 Rajpurkar 等人[11]使用維基百科的文章進行問答研究，和 Tapaswi 等人[12]專注在電影內容交談的問答研究，另外還有 Hermann 等人[13]以新聞文章內容進行問答研究，透過 Word Distance 來計算故事與問題的字詞，但容易有詞彙被忽略，以及 Seo 等人[14]使用 LSTM 搭配 Attention，用來預測答案出現在故事的位置，Yu 等人[15]透過 CNN 以及 Self-Attention，找出答案在故事中最有機會出現的位置，以及中文 Shao 等人[16]使用 BERT 模型，在中文開放性領域問題上的研究，贏過 QANet[15]、R-Net[19]、BiDAF[14]等模型。

三、研究方法

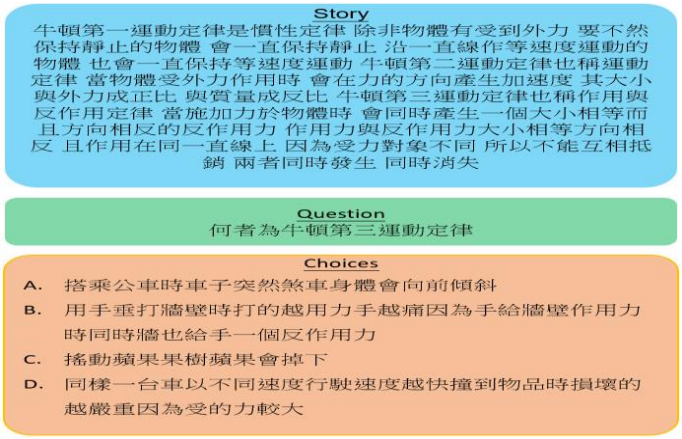
本此章節將說明本論文的研究方法與系統架構，將架構分為四大部分，分別對中英文資料進行前處理，並透過 BERT 將句子轉換為機器能使用的語意資訊，接著進行機器理解答題模型的推論，最終選出預測的答案。以下小節將會針對系統中各模組的細節做進一步說明。



圖一、系統架構圖

（一）、資料前處理

為了要提升資料在模型上的精準度，此階段除掉非文字的標點符號如圖二所示，Story 的部分以逗號及句號為單位來切割內文，而 Question 及 Choices 則去除所有標點符號，當成一個句子來處理。

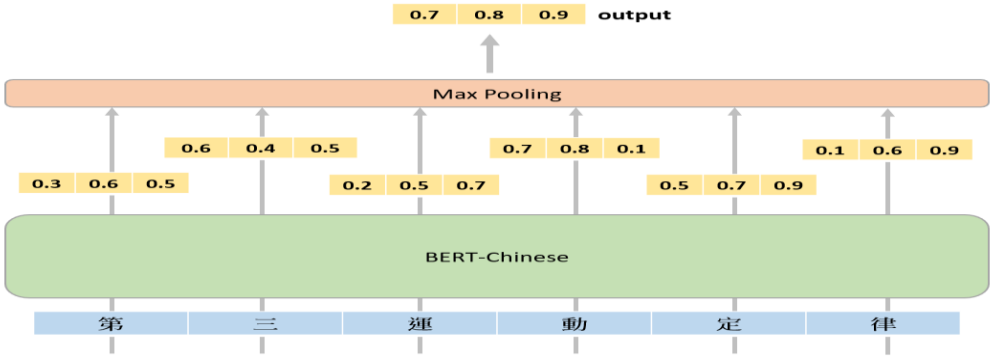


圖二、閱讀測驗題目

（二）、語意學習

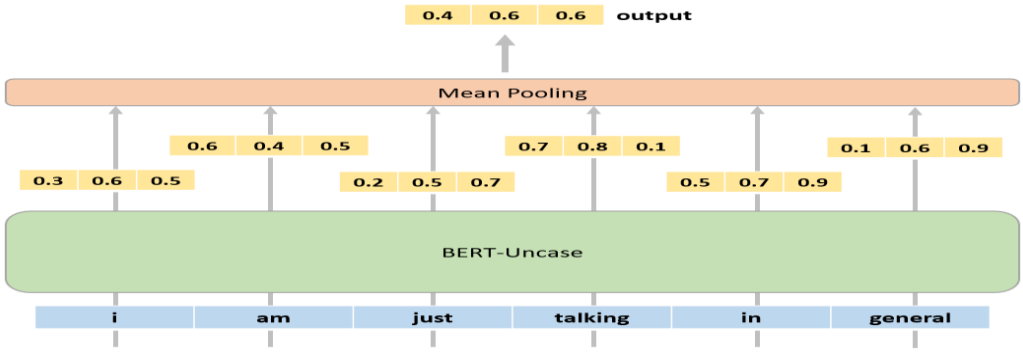
大量詞彙分散在許多上下文中，藉以組織不同文章表達不同主軸的意義，為了要使機器能夠得到這種語意資訊，本論文利用 BERT model 進行詞嵌入向量，對 BERT-Base, Chinese 12-layer 及 BERT-Large, Uncased 24-layer 實驗不同 pooling strategy。Max pooling 為，將每個時間軸上的隱藏層(hidden layer)資訊取最大值，來組織出

該句子所蘊含的語意資訊如圖三。



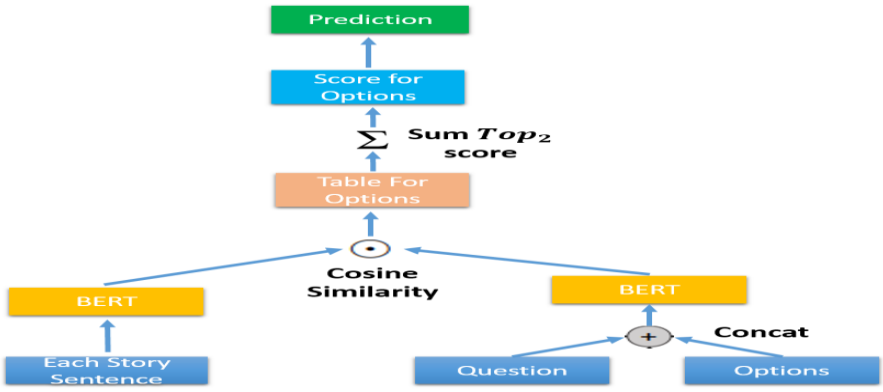
圖三、Max pooling strategy

而 Mean pooling 則是取隱藏層資訊的加總平均值，來組織出該句子所蘊含的語意資訊如圖四，藉此將字詞之間較重要的語意資訊保留，提高推論模型的精準度。



圖四、Mean pooling strategy

（三）、機器理解



圖五、策略一模型圖

本論文基於回答閱讀測驗時，採取不同的答題策略，設計了二種類型的答題模型。

第一種答題策略如圖五，透過將 Story 以句子作切割，能更詳細的思考 Story 中的所

有資訊，將每個句子視為可能保有關鍵資訊來處理如公式(1)，接著思考 Question 以及 Choice 組成的線索如公式(2)，計算兩者之間的相似性如公式(3)，接著計算最相關於選項的故事句子分數，專注於最相關於 Choice 的句子如公式(4)，最終選擇關聯性最高的選項如公式(5)，本模型的公式如下所示：

$$S_n = \boldsymbol{BERT}(Story_n) \quad (1)$$

$$QC_m = \boldsymbol{BERT}(Question + Choice_m) \quad (2)$$

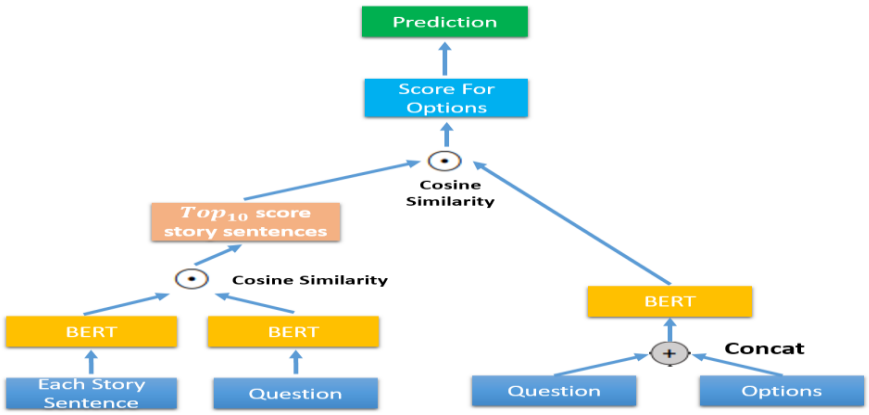
$$Table_m = \{\boldsymbol{Sim}(S_1, QC_m), \ldots, \boldsymbol{Sim}(S_n, QC_m)\} \quad (3)$$

$$Score_m = \sum Top_k(Table_m) \quad (4)$$

$$Prediction\ \ Answer = \boldsymbol{argmax}_m(Score_m) \quad (5)$$

在上述公式中，各個參數及設置的定義如下：

1. n: Story 中句子的數量，m: Choice 的數量，k: 2。
2. BERT: 將句子作為 BERT 的輸入，取得轉換的詞嵌入向量，組織成句子資訊。
3. Sim: 計算 Cosine Similarity 餘絃相似度分數。
4. Table: 一個 Choice 與 Question 合併，對應 Story 所有句子計算的分數表。
5. Score: 將每個 Choice 的分數表，最高的兩個分數相加，為該選項的分數。
6. Prediction Answer: 將四個 Choice 裡最終分數最高的選項作為預測答案。



圖六、策略二模型圖

第二種答題策略如圖六，將每個句子都視為保有關鍵資訊來處理如公式(6)，接著思考 Question 以及 Choice 組成的線索如公式(7)，並且獨立出 Question 的資訊，以免於被選項的資訊所影響如公式(8)，透過這樣的方式思考 Story 中幾個最相關於

Question 的句子，能夠排除不相關的故事內容如公式(9)，接著計算最相關於每個 Choice 的 Story 句子如公式(10)，最終選擇關聯性最高的選項如公式(11)，本模型的公式如下所示：

$$S_n = \boldsymbol{BERT}(Story_n) \quad (6)$$

$$QC_m = \boldsymbol{BERT}(Question + Choice_m) \quad (7)$$

$$Que = \boldsymbol{BERT}(Question) \quad (8)$$

$$Top_k Sentence = Top_k\{\boldsymbol{Sim}(S_1, Que), \ldots, \boldsymbol{Sim}(S_n, Que)\} \quad (9)$$

$$Score_m = \boldsymbol{max}(\boldsymbol{Sim}(\mathrm{Top}_k S, Choice_m) \quad (10)$$

$$Prediction\ \ Answer = \boldsymbol{argmax}_m(Score_m) \quad (11)$$

在上述公式中，各個參數及設置的定義如下：

1. n: Story 中句子的數量，m: Choice 的數量，k: 10。
2. BERT: 將句子作為 BERT 的輸入，取得轉換的詞嵌入向量，組織成句子資訊。
3. Sim: 計算 Cosine Similarity 餘絃相似度分數。
4. $Top_k Sentence$: 取 Story 句子與 Question 計算餘絃相似度的 Top10 句子。
5. Score: 每個選項與 Top10 句子做相似計算，最高相似度為該 Choice 分數。
6. Prediction Answer: 將四個 Choice 裡最終分數最高的選項作為預測答案。

四、實驗與討論

本章節將針對本論文所提出來的研究方法，探討使用 BERT 語言模型作為語意轉換，之後透過答題模型進行機器理解的問答，探討其效果。

（一）、實驗資料集

本論文使用中英文兩種閱讀測驗資料集，英文為 Listening Comprehension Test of TOEFL[7]資料集，合併訓練及測試資料集總筆數為 839 筆。中文為科技大擂台_測試資料集[18]前六次初賽資料，總筆數為 8550 筆，並且中英文資料集皆為單選題。

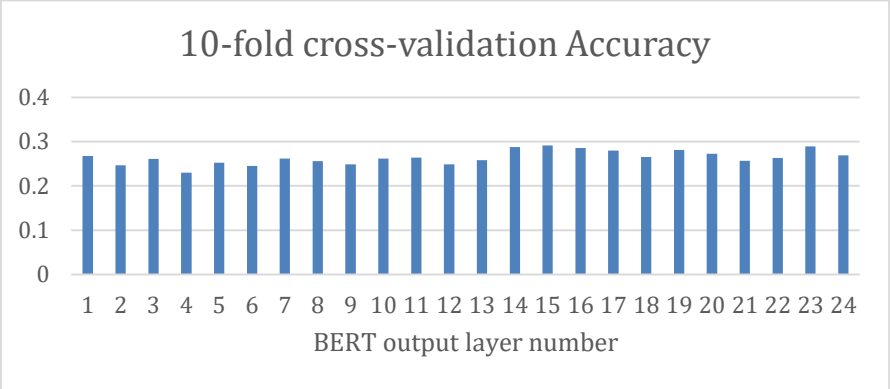
（二）、驗證方法

在中文及英文的資料集之中，本論文採用 10-fold Cross Validation[17]作為進一步評估誤差的指標。

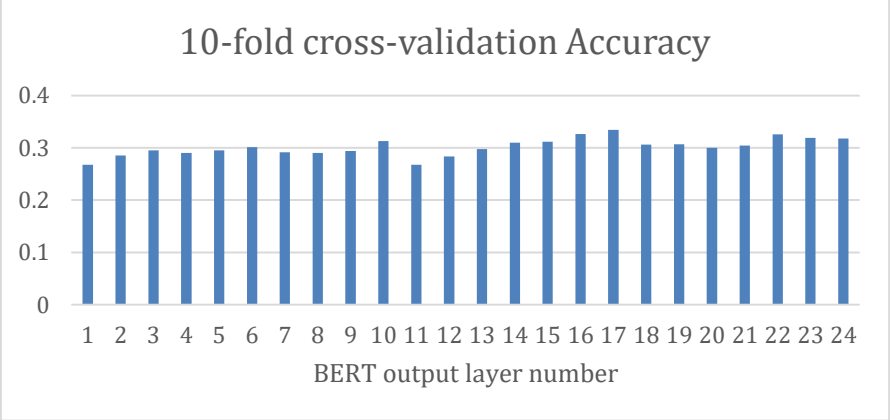
（三）、語言模型詞嵌入比較

我們將比較二種答題模型，中文使用 BERT Chinese，英文使用 BERT Uncased，比較不同 layer 的 output 作為詞嵌入向量，及不同的 pooling strategy 來組織句子後的效果。

1、英文資料集

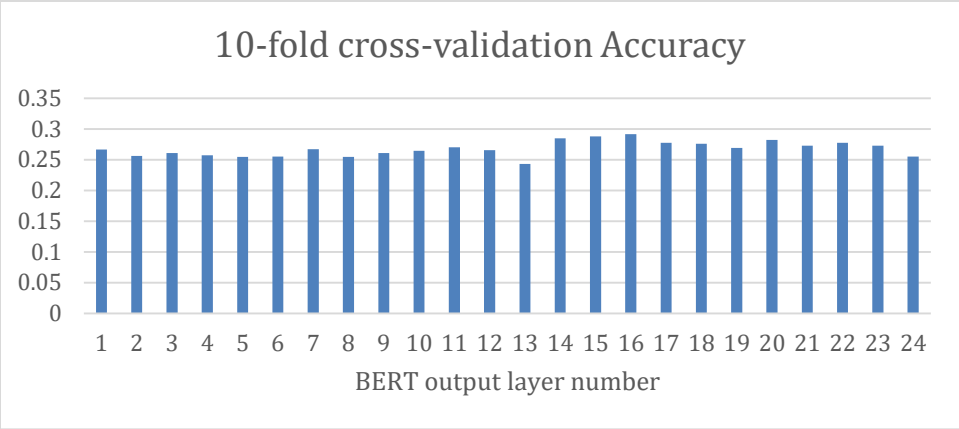


圖八、英文資料集 策略一及 max pooling

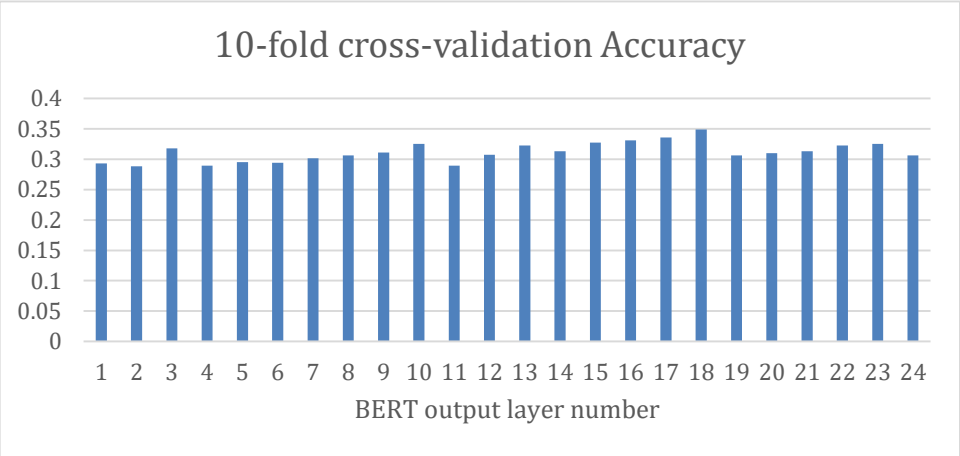


圖九、英文資料集 策略一及 mean pooling

從圖八、圖九中觀察到，英文資料集使用策略一及 mean pooling 時，以 17-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.3345 的準確率。



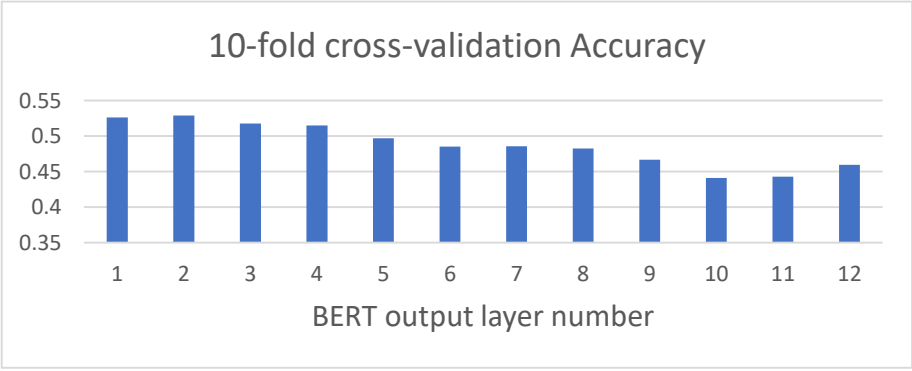
圖十、英文資料集 策略二及 max pooling



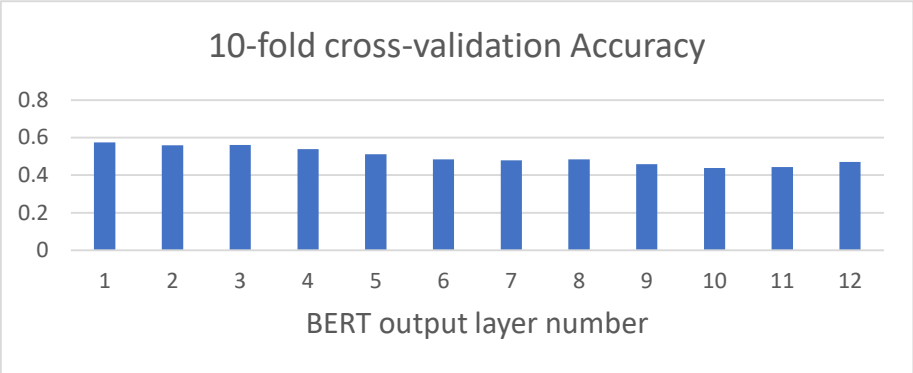
圖十一、英文資料集 策略二及 mean pooling

從圖十、圖十一可以觀察到，英文資料集使用策略二及 mean pooling 時，以 18-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.3487 的準確率。

2、中文資料集

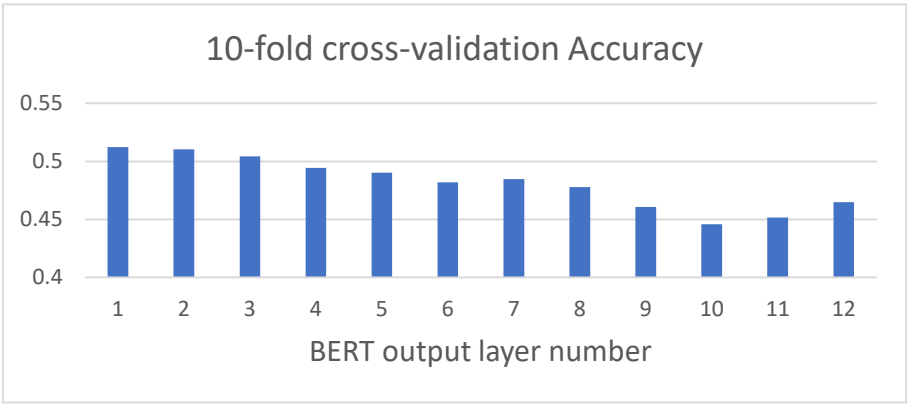


圖十二、中文資料集 策略一及 max pooling

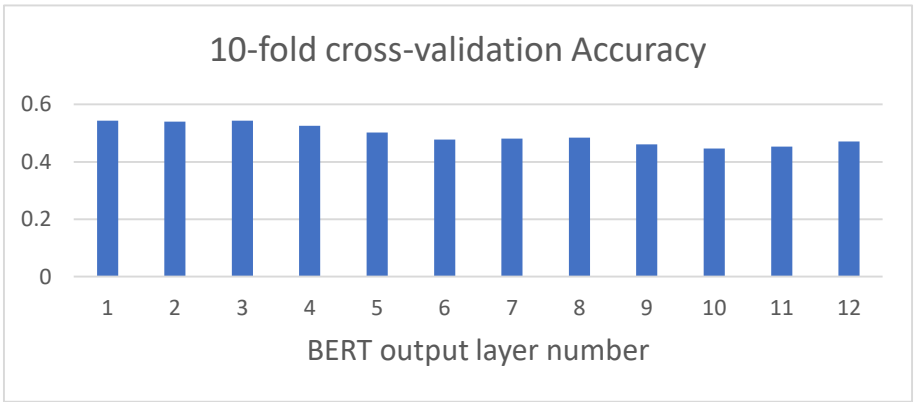


圖十三、中文資料集 策略一及 mean pooling

觀察圖十二、圖十三，中文資料集使用策略一及 mean pooling 時，以 1-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.575 的準確率。



圖十四、中文資料集 策略二及 max pooling



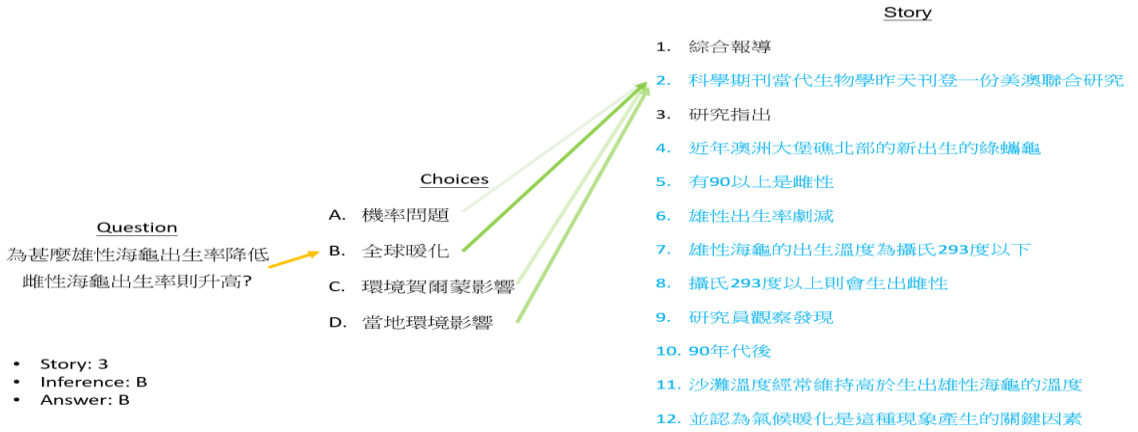
圖十五、中文資料集 策略二及 mean pooling

觀察圖十四、圖十五，中文資料集使用策略二及 max pooling，以 1-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.5439 的準確率。

（四）、策略方法解析

圖十六、策略一解析

策略一結果如圖十六，利用合併 Question 以及 Choice 的資訊(如黃色箭頭)，以及將 Story 切割為句子來檢視故事中的資訊，進一步計算最相關於選項的 2 個句子(綠色越深代表關聯越強，紅色越深代表關聯越差)，透過計算最相關的句子，能更詳細的考慮故事與選項之間的關聯，避免沒有正確資訊的句子，與單一選項關聯過高的情況。

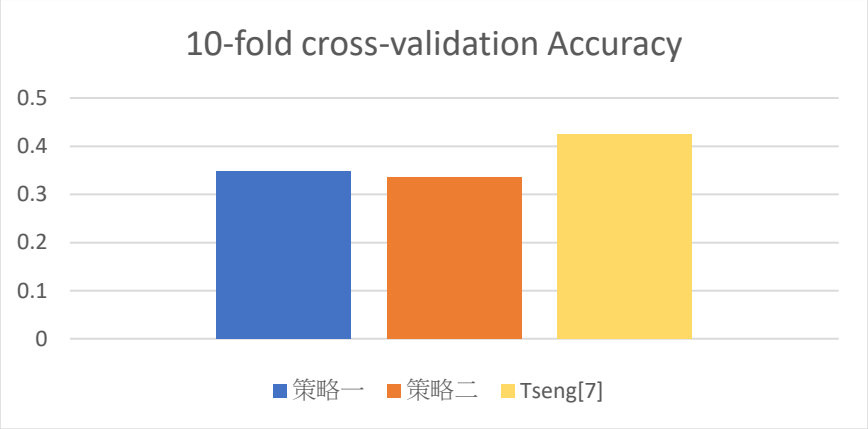


圖十六、策略二解析

策略二結果如圖十六，利用合併 Question 以及 Choice 的資訊(如黃色箭頭)，以及將 Story 切割為句子來檢視故事中所有的資訊，透過保留與問題的關聯度 Top10 的 Story 句子(如圖藍色部分)，能夠幫助篩選掉部分與選項不相關的故事句子，最終只計算與選項有相關的故事句子(綠色越深代表關聯越強，紅色越深代表關聯越差)。

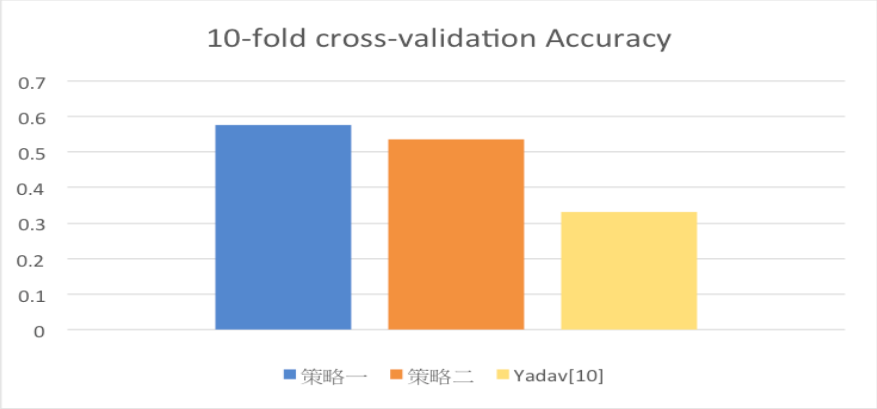
（五）、答題模型結果比較

與提出的二種答題模型比較，在中文上我們實驗了 Yadav 等人[10]的模型，最終得到 33.2%準確率，以及比較 Tseng 等人[7]的結果 42.5%準確率，作為最終結果對照。



圖十七、英文資料集 10-cross validation Accuracy

在英文測試資料集上如圖十七，最佳的答題模型為策略一，最終結果為 34.87%。



圖十八、中文資料集 10-cross validation Accuracy

在中文測試資料集上，最佳的答題模型為策略一，最終結果為 57.5%如圖十八。

五、結論

本論文基於英文與中文兩種語言的閱讀測驗，提出兩種答題模型，在中文資料集上，策略一的模型優於所有結果，此外策略一模型使用在英文時也有一定的效果，經過實驗證實，切割故事的所有句子，有助於保留重要的答案句子資訊，透過合併問題及答案的資訊，可以讓模型更全面的檢視線索，進而讓答題模型找出正確答案。針對答題模型，可以藉由投票機制，組合多個答題模型的結果，以此增加多種不同的答題策略，來彌補不同答題模型之中的缺陷，改善最終的準確率。而資料則可以改善斷句、資料集數量的增加，以及針對不同長度的故事、問題、選項之間優化答提模型的設計方式。

參考文獻

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

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998-6008.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).

[4] Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications, 68*(13), 13-18.

[5] Huang, A. (2008, April). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (Vol. 4, pp. 9-56).

[6] Karypis, M. S. G., Kumar, V., & Steinbach, M. (2000, May). A comparison of document clustering techniques. In *TextMining Workshop at KDD2000 (May 2000)*.

[7] Tseng, B. H., Shen, S. S., Lee, H. Y., & Lee, L. S. (2016). Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine. *Interspeech 2016*, 2731-2735.

[8] Fang, W., Hsu, J. Y., Lee, H. Y., & Lee, L. S. (2016, December). Hierarchical attention model for improved machine comprehension of spoken content. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 232-238). IEEE.

[9] Richardson, M., Burges, C. J., & Renshaw, E. (2013, October). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 193-203).

[10] Yadav, V., Sharp, R., & Surdeanu, M. (2018, June). Sanity check: A strong alignment and information retrieval baseline for question answering. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1217-1220). ACM.

[11] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November).

SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383-2392).

[12] Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4631-4640).

[13] Hermann, K. M., Koč iský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015, December). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1* (pp. 1693-1701). MIT Press.

[14] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In ICLR.

[15] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading compre- hension. In ICLR, 2018a.

[16] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: a chinese machine reading comprehension dataset. CoRR, cs.CL/1806.00920v2.

[17] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*(Vol. 14, No. 2, pp. 1137-1145).

[18] 科技部,科技大擂台_測試資料集：
https://scidm.nchc.org.tw/dataset/grandchallenge

[19] Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017, July). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 189-198).

[20] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT* (pp. 2227-2237).

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 62-71

預訓練詞向量模型應用於客服對話系統意圖偵測之研究

Study on Pre-trained Word Vector Model Applied to Intent Detection of Customer Service Dialogue System

陳冠宇　Guan-Yu Chen
郭敏楓　Min-Feng Kuo
楊宗憲　Tsung-Hsien Yang
陳俊勳　Chun-Hsun Chen
廖宜斌　I-Bin Liao

中華電信研究院 巨量資料研究所
Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan, R.O.C.
robinchen@cht.com.tw
kmf0822@cht.com.tw
yasamyang@cht.com.tw
jeffzpo@cht.com.tw
snet@cht.com.tw

摘要

近年來對話商務的概念在各大科技巨頭間興起，人機互動方式由圖形化介面轉向對話交互介面的方式。因而自然語言成為人機互動介面的關鍵因子。然而教導機器要如何與人類溝通，以完成一項具體任務是相當有挑戰性的。其中一個需要克服的困難是自然語言理解，包含如何辨識使用者在詢問何種問題及如何取得文字間隱藏的資訊。讓機器了解使用者的問題意圖及資訊是相當重要的。本研究主要是針對去識別化後的中文客服對話資料，利用深度學習模型以達到辨識使用者意圖。為了更有效處理中文未知詞以及減少錯誤辨識，本研究比較不同預訓練詞向量模型與深度學習模型來辨識使用者意圖。相較於使用隨機詞嵌入，使用 BERT-WWM-Chinese (BWC)模型的正確率提升近 10%。這表示 BWC 模型產生的向量更能抓住用戶問句字詞間的語意關係。使得語意相近的字詞能產生近似的向量進而提升使用者意圖辨識的準確率。

Abstract

In recent years, the concept of dialogue business has arisen among major technology giants, and the way of human-computer interaction has changed from a graphical interface to a

dialogue interaction interface. Therefore, natural language has become a key factor in the human-computer interaction interface. However, teaching the machine to communicate with humans to accomplish a specific task can be quite challenging. One of the difficulties that needs to overcome is natural language understanding, including how to identify what questions users are asking and how to get information hidden between words. It is important to let the machine know the user's intentions and information.

The dataset of this study is collected from the dialogue of customer service materials. User's intents are recognized by deep learning models. In order to process Chinese unknown words more effectively and reduce false recognition, this study compares different pre-training vector models and deep learning models to understand user's intents. Compared with the use of random word embedding, the correct rate of using BERT-WWM-Chinese (BWC) model is improved by nearly 10%. It shows that the semantic vector generated by BWC model can better represent the relationship between user's words. The recognition rate of user's intent raises because similar vectors can be generated from similar words.

關鍵詞：對話系統，對話行為，深度學習，預訓練詞嵌入模型，注意力機制

Keywords: Dialogue System, Dialogue Act, Deep Learning, Pre-trained Word Embedding, Attention.

一、緒論

在科技不斷進步的今日，電腦、智慧型裝置、網路資訊服務在人類生活中日益扮演著重要的角色，人跟機器之間已有著密不可分的關係。也因為近年來人工智慧發展快速，更多應用深度學習技術來精進傳統機器學習方法的技術。

在許多情景下，對話用戶介面比圖型用戶介面更加自然及高效率，加上智慧型手機之發展，相較於撥打傳統語音客服，年輕人開始轉向使用文字對話介面與客服互動。傳統語音客服的各項任務，皆可經由文字對話描述來實現互動服務。許多公司及個人都嘗試著架構專屬的聊天機器人(Chabot)。然而，聊天機器人的功能不僅僅侷限於聊天，能夠以對話的方式來協助人類完成各式各樣目標才是我們真正想要的人工智慧。

在對話系統中，一般系統的輸入可以為語音或是文字，基於語音系統架構以 Steve Young 提出的架構最為典型[1]，而以文字為輸入的系統架構可以分為三大部分，首先是自然語言理解 (Nature Language Understanding, NLU) ，接著為對話管理 (Dialogue Manager,

DM) ，最後為自然語言產生 (Nature Language Generation, NLG) 。

機器人對話系統大致上可分為 2 種，分別是：

1. 聊天型對話系統：其不需要理解問題，只需要依據模型預測給予答案即可。

2. 任務型對話系統：需要分析問題意圖並給予預先建立之領域知識庫所定義的答案，其實作困難度相較於聊天型對話系統更高，因為回覆必須精準解答問題。

以實用性而言，任務型對話系統遠較聊天型對話系統要高，因為其可以給予用戶較有資訊意涵之回應。其衡量指標在於讓用戶自助服務率提高、並提升用戶滿意度，這包含對話輪次越少就能找到答案越好、可支援的意圖越多越好、回覆的內容越精準確實越好。在傳統文字客服系統上，用戶進線與客服交談以取得對應的解答。但是受限於公司成本考量、客服人力有限且每個值機員依其專業度，回覆內容之品質不一；若是能將大部分較單純、基本的問題，轉交給機器人來回答，那麼就可以節降公司不少值機成本，且可確保回覆內容一致，此外，用戶也可以節省掉等待值機員空閒、打字的時間，快速取得其需要的服務。該如何應用這些客服的文字記錄，研發出針對特定領域能自動問答的對話系統，是本研究的研究動機。

真實的客服對話系統中，使用者不管是表達句子方式或是詢問內容是否為該公司的業務，都不會受到任何限制，想問甚麼就問甚麼。輸入的句子較為口語化，因此經常有省略標點符號，打錯字，表達用詞上、文法上的錯誤，或是詢問非該領域的內容，導致無法使用預先定義的 ontology，在語意理解上也相當困難。因為任務導向型對話系統需要準確的理解語意，如何訓練機器去理解人類口語對話的語意，是相當重要的關鍵部分，如能有效解決此問題，將對自然語言理解能有更大進展。近年來，隨著深度學習技術發展迅速，國內外已經在對話系統有許多相關研究，目標都是在建立一套具有智慧的對話系統[2-5]。針對自然語言理解的研究，使用真實中文對話資料集的研究相當稀少，本文將比較不同深度學習模型是否可以在真實中文客服資料集中正確的辨別對話行為。

二、研究方法

一個說話者在對話中所要傳遞其意圖稱為對話行為(Dialogue act, DA)。最近由於深度學習的一大突破，許多在對話行為研究[6-9]都有比傳統機器學習方法有不錯的提升。對話行為分類在對話系統上佔據關鍵的因素，機器能夠充分了解使用者所表達的句子語意。

(一).系統介紹

1. 資料集簡介：本論文使用之資料集為去識別化後文字及簡訊客服(不含語音客服)的用戶對話資料，內容為行動電話與上網業務之常見客戶提問，擷取了其中較常見的 10 種提問意圖類別，並對語句進行下述文字正規化：(1)除去意圖不明語句(2)除去總字數小於 5 個字元內的語句(3)除去標點符號與未知字元(4)全形轉成半形、英文大寫轉成小寫。

   正規化後，共 15,773 筆訓練語句，1,584 筆測試語句，平均每句字數 12.61 個字，詳細說明如表一：

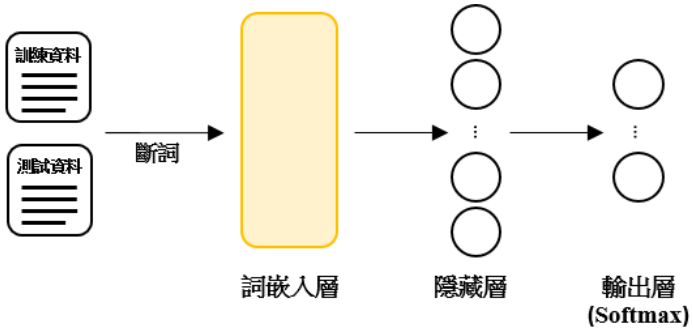
表一、資料集各類意圖筆數

意圖類別名稱	訓練集資料筆數	測試集資料筆數
資費與合約查詢或修改	2,000	253
帳單或繳費問題	2,000	102
加值服務問題	483	24
優惠方案	2,000	668
手機銷售問題	1,239	58
國際漫遊	2,000	113
手機用量問題	1,557	80
障礙申告或收訊問題	2,000	161
APP 使用問題	1,712	81
服務據點問題	782	44

(二).模型設計

近年來，人工智慧發展迅速，其中機器學習分支之一的深度學習在各個領域有許多重大突破。自然語言處理領域中，傳統機器學習方法需要人工設計模型所需的特徵組合，常消耗大量人力與時間。另一方面，深度學習則可以自動找出模型特徵表示，同時深度學習許多架構都能有效處理不同的自然語言處理任務，以下將介紹常見的深度學習模型。本論文討論以下 4 種預訓練詞向量方法對意圖分類器之影響：(1)隨機嵌入(2)Skip-Gram (3)BERT (4)BERT-WWM-Chinese，分類器模型訓練流程如圖

一：



圖一、分類器模型訓練流程圖

(1) 隨機嵌入：訓練分類器時，使用均勻分布(Uniform Distribution)隨機初始化詞嵌入層，每一詞彙索引對應一個值域範圍為[-0.05, 0.05]的 300 維詞向量。

(2) Skip-Gram[10]：Word2vec 是由 Mikolov 等人在 2013 年提出，使用到淺層的神經網路模型，模型分為 Continuous Bag of Words(CBOW)與 Skip-gram，CBOW 利用詞語的前後字建立詞窗當作輸入，來預測出此詞語，而 Skip-gram 則相反過來。由 Word2vec 產生的文字向量，將此向量投射到一個向量空間中，語意相近的詞彙將會在向量空間中非常相近，顯示文字可以在向量空間中有語意近似的關係。近期，許多的 Word2vec 應用在預訓練的詞向量作為訓練深度學習模型的詞語初始向量，對訓練模型時可以更有效的調整向量，加速收斂，同時也會不斷調整(fine-tune)詞向量，更能強化詞語語意關係。使用兩種資料集，分別為 2019 年 6 月繁體中文維基百科[1]，共 340,369 篇文章，斷詞[2]後共 1,042,225 個不重複詞彙；真實中文客服之對話資料集斷詞後共 1,653 個不重複詞彙，各自訓練兩個 300 維的詞向量模型。

(3) BERT(Bidirectional Encoder Representation from Transformers) [11]：從 ELMo[12]之後，根據前後文語意產生句向量之語意表達模型(Contextualized word representations)，成為近年來，自然語言處理領域的熱門研究主題，這類模型的優勢在於：(1) 能夠學習到詞彙在多情境下之不同語意、語法 (2) 能夠學習到詞彙依據上下文變化所帶來之不同語意，ELMo 透過雙向語言模型(Bidirectional language model, biLM)提供了對下游任務效果更佳的詞向量。而 OpenAI GPT[13]則將多層的

[1] https://dumps.wikimedia.org/zhwiki/
[2] https://taku910.github.io/crfpp/

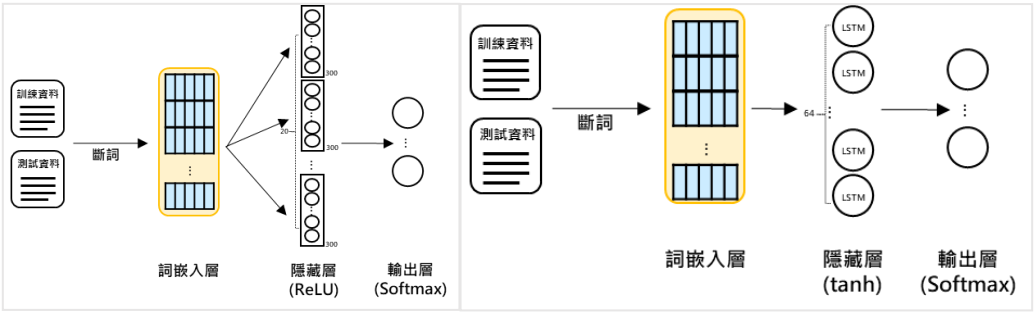
單向 Transformer[14]模組，引入了語意表達模型中。

Google 於 2018 年 10 月所發布的 BERT 模型，也基於 Transformer 模組，不同於 OpenAI GPT 模型的單向 Transformer，BERT 藉由從克漏字(Cloze)任務所帶來的靈感，使用雙向 Transformer 對遮蔽的 Token 進行預測(Masked Language Model, MLM)，以及預測下一句的下游任務，讓模型學習到句子之間的關係，建構一通用的預訓練語言表達模型(Language representation model)，使得自然語言處理的各項任務(如: 意圖分類、問答、翻譯等)，能夠使用較為輕量化的模型，輕易地進行遷移學習(Transfer learning)，在下游任務的訓練過程中，對表達模型進行優化(Fine tuning)即可。使用 Google 於 2018 年 11 月所發布的中文預訓練模型檔案，對每句訓練語句，產生 768 維的雙向語意表示(Bidirectional contextual representation)向量。

(4) BERT-WWM(Whole word masking)-Chinese[15]：中國哈爾濱工業大學於 2019 年 6 月發布，基於 2019 年 5 月 Google 所更新的 BERT 模型，新模型修改了原先訓練過程中，由隨機遮蔽單個字元，改為遮蔽完整詞彙(Whole word masking, WWM)，WWM 將完整詞彙的語意帶入模型中，使得模型較容易學習到字元間常用的組成關係。BERT-WWM-Chinese 採用與(3)相同的 BERT 模型作為基礎，使用中文維基百科及哈爾濱工業大學語言技術平台(Language technology platform, LTP)[3]的斷詞器，斷詞後，加入 WWM 重新訓練，產生 768 維的雙向語意表示向量。

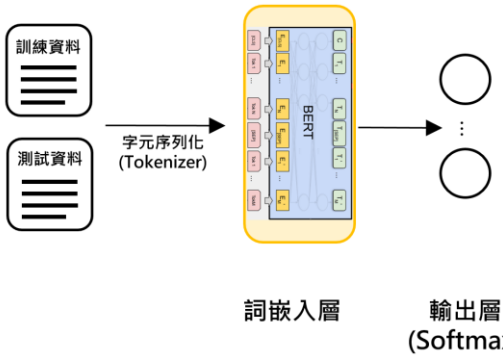
接續在(1)隨機嵌入與(2)Skip-Gram 兩種預訓練詞向量之後，本論文設計了兩種分類器，兩種分類器皆為多層感知機(Multi-Layer Perceptron, MLP)，包含一輸入層、一隱藏層(Hidden Layer)、一輸出層。兩種分類器的差異在於隱藏層之設計：對於句子中的斷詞結果，依每一詞彙出現的順序，產生對應的詞向量，如圖二，(a)將前 20 個詞彙的詞向量拼接起來，形成一 6000 維的句向量，連結一 64 維使用線性整流(Rectified linear unit, ReLU)函數作為激活函數的隱藏層。(b)在限制最多 20 個詞彙輸入的情況下，將詞向量依序放入，傳播至隱藏層共 64 維的長短期記憶(Long short-term memory, LSTM) [16]神經元。輸出為一 10 維隱藏層，使用 Softmax 函數作為激勵函數(Activation function)輸出該語句所預測之意圖分類。

[3] https://ltp-cloud.com/



圖二、詞嵌入層搭配 MLP/LSTM 之分類器

而在(3)BERT 及(4)BERT-WWM-Chinese 的雙向語意向量後，直接連結上述所提相同結構之輸出層，作為輸出該語句所預測之意圖分類，如圖三。



圖三、使用 BERT 及 BERT-WWM-Chinese 之分類器

三、結果與結論

(一). 實驗配置

使用 gensim[4]套件訓練 Skip-Gram 預訓練詞向量模型，過濾 5 個字元以下的語句，循環 10 次(Epochs)。而隨機嵌入與 Skip-Gram 所連接的 MLP、LSTM 分類模型，採 RMSprop 優化器(Optimizer)，學習率(Learning Rate)為 0.001，循環 50 次，批量大小(Batch size)為 64。BERT 與 BERT-WMM 使用 Adam 優化器，採 2e-5 的學習率(Learning rate)，循環 100 次，批量大小為 16。損失函數(Loss function)皆使用分類交叉熵(Categorical cross-entropy)，驗證(Validation)資料集大小皆設定為訓練資料集的 1%，並設定若連續 5 次模型在驗證集上的損失沒有下降，則提早終止(Early stopping)訓練過程。

[4] https://radimrehurek.com/gensim/models/word2vec.html

(二). 實驗結果

由於測試資料中，資料數量最多(668 筆)與最少(24 筆)的類別有不小之差距。為避免因為資料類別不平衡，導致實驗結果陷入正確率悖論(Accuracy paradox)，故採用分類正確率(Accuracy)之外，也使用了 Macro-F1 作為多類別意圖分類模型的評估指標。為了評估的一致性，每個模型皆隨機重覆跑了 5 次結果，取平均值及標準差，實驗結果如表二。

表二、比較不同詞嵌入層之分類器的結果

模型名稱			平均正確率(標準差)		平均 Marco-F1(標準差)	
			不更新詞嵌入層	更新詞嵌入層	不更新詞嵌入層	更新詞嵌入層
隨機嵌入		MLP	0.6830(0.026)	0.7456(0.019)	0.6920(0.026)	0.7290(0.017)
		LSTM	0.6506(0.030)	0.7322(0.046)	0.5956(0.022)	0.6782(0.020)
Skip-Gram	維基百科	MLP	0.6988(0.016)	0.7664(0.008)	0.6718(0.007)	0.7516(0.006)
		LSTM	0.7792(0.016)	0.8106(0.016)	0.7652(0.015)	0.8044(0.014)
	客服資料	MLP	0.6978(0.016)	0.7838(0.010)	0.6656(0.005)	0.7672(0.009)
		LSTM	0.7862(0.020)	0.8110(0.013)	0.7638(0.013)	0.8040(0.013)
BERT				0.8290(0.022)		0.8300(0.015)
BERT-WWM-Chinese				**0.8414(0.006)**		0.8435(0.017)

從表二的實驗數據中，可以得到以下資訊：

1. BERT 架構正確率較高：使用 BERT 架構的模型，與其他模型相比，於分類正確率與 Macro-F1 分數皆取得較好的成績；導入 WWM 之後，與原 BERT 模型相比，兩種指標也有些微提升。

2. 更新詞嵌入層優於不更新：對於每一種預訓練詞向量模型，在訓練分類器的過程持續更新詞嵌入層，分類正確率進步至少 3%左右。

3. MLP 與 LSTM 之比較：值得注意的是，使用隨機嵌入，MLP 會得到比 LSTM 更好的效果，但使用 Skip-Gram 作為預訓練詞向量則反之。

4. 隨機嵌入法與 Skip-Gram 之比較：同是使用客服資料進行訓練的隨機嵌入法與

Skip-Gram，不管使用 MLP 或 LSTM 作為分類器，Skip-Gram 的效果都比隨機嵌入法來的佳。

5. 維基百科及客服資料之比較：兩者用來訓練 Skip-Gram，發現在兩種指標上皆沒有太大的差異，但兩者的語料規模差異很大；顯示若使用與下游任務領域相符的語料，訓練詞向量模型，較可節省計算資源，達到使用一般詞彙語料訓練之水平。

使用維基百科訓練的 Skip-Gram 之 LSTM 分類器，有 70%的錯誤分類來自「優惠方案」，22%的錯誤分類來自「資費與合約查詢或修改」；而使用 BERT 訓練之分類器，有 75%的錯誤分類來自「資費與合約查詢或修改」11%的錯誤分類來自「優惠方案」。這兩個意圖類別也是資料筆數最多的類別，其中可以發現收集到的語句資料變異情況較大，且單詞容易與其他意圖類別之語句重複，故容易分類錯誤。

四、結論

近年來詞嵌入向量(word embedding)在自然語言研究中激起一股研究熱潮，此種表示方式，不僅能以較低維度的向量表示詞彙，還能藉由詞向量間的運算，找出任兩詞彙之間的語意關係。因此，本文實驗探討了新穎的詞向量模型 BERT 與隨機嵌入/Word2Vec 對於特定領域之客服對話意圖辨識效能增進的議題。主要貢獻有兩個部分: 第一部分，本論文將詞向量表示資訊應用於特定領域的客服系統中，透過此種表示方式而能獲取到更多詞彙間的語意資訊，以提升意圖辨識的準確度。第二部分，我們對於近來深度學習強調使用預訓練模型進行微調的訓練模式進行了驗證。我們利用 BERT 模型以及其增益模型 BERT-WWM-Chinese 為預訓練模型，使用少量的訓練資料進行微調，確實幫助意圖辨識準確率的提升。未來，勢必會有更多新穎的預訓練模型不斷的被提出來加強各種 NLP 後端應用的效能。因此，能有效評價各模型於不同後端應用的優劣與整合不同模型達到集成學習的效果也是未來研究重點之一。

參考文獻

[1] S. J. Young, "Probabilistic methods in spoken–dialogue systems," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences,* vol. 358, no. 1769, pp. 1389-1402, 2000.

[2] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," *arXiv preprint arXiv:1605.07683*, 2016.

[3] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models," in *AAAI*, 2016, vol. 16, pp. 3776-3784.

[4] T.-H. Wen *et al.*, "A network-based end-to-end trainable task-oriented dialogue system," *arXiv preprint arXiv:1604.04562*, 2016.

[5] X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, "End-to-end task-completion neural dialogue systems," *arXiv preprint arXiv:1703.01008*, 2017.

[6] L. Meng and M. Huang, "Dialogue Intent Classification with Long Short-Term Memory Networks," Cham, 2018: Springer International Publishing, in Natural Language Processing and Chinese Computing, pp. 42-50.

[7] H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi, "Dialogue act sequence labeling using hierarchical encoder with crf," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] C. Cerisara, P. Kral, and L. Lenc, "On the effects of using word2vec representations in neural networks for dialogue act recognition," *Computer Speech & Language*, vol. 47, pp. 175-193, 2018.

[9] S.-s. Shen and H.-y. Lee, "Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection," *arXiv preprint arXiv:1604.00077*, 2016.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] M. E. Peters *et al.*, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[13] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[14] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.

[15] Y. Cui *et al.*, "Pre-Training with Whole Word Masking for Chinese BERT," *arXiv preprint arXiv:1906.08101*, 2019.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 72-81

A Hybrid Approach of Deep Semantic Matching and Deep Rank for Context Aware Question Answer System

Shu-Yi Xie, Chia-Hao Chang†, Zhi Zhang, Yang Mo,
Lian-Xin Jiang, Yu-Sheng Huang, Jian-Ping Shen
AI Department
Ping An Life Insurance of China, Ltd.
{xieshuyi542,zhangjiahao206†,zhangzhi600,moyang853,jianglianxin769,
huangyusheng112,shenjianping324}@pingan.com.cn
Correspondence†:strategist922@gmail.com

Abstract

Most of the existing Question Answer Systems focused on searching answers from the Knowledge-Base (KB), and ignore context aware information. Many Question Answer models perform well on public data-sets, but too complicated to be efficient in real world cases. Effectiveness, concurrency and system availability are equally important in industry which have large data and requests, we propose a Context Aware Question Answer System based on the Information Retrieval with Deep Semantic Matching and Deep Rank. It has been applied to the online question answer system for insurance Question Answer. By these means, we achieve both high QPS (Query Per Second) and effectiveness. Our approach improves the system's ability to understand the question with context aware coreference resolution, subject completion, and the long sentence compression. After the matching questions are recalled from the ElasticSearch, Siamese CBOW (Continues Bag-Of-Words Model) and KBQA filter some unreasonable ones by entity alignment. After the result is sorted by the deep rank model with co-occurrence words and semantic features, our system does clarification or answer output. Finally, for those questions that we are unable to provide answers, a dialogue mining module as part of our Smart Knowledge-Base Platform is developed. This results in more than 10 times improvement in terms of efficiency for manpower involved in data labeling process.

KEYWORDS

Question Answering, Coreference Resolution, Error Correction, Sentence Compression, Deep Semantic Match, Deep Rank, Knowledge-Base Management, Insurance Domain

All authors contributed equally to this manuscript.

1 INTRODUCTION

The question answer system has been widely used in intelligent customer service, personal assistants, and dialogue robots. In 2018, the pretrain techniques based on a massive corpus pre-training model have made breakthroughs in multiple NLP tasks including Semantic Match. Representative models are Elmo[9], GPT[10], BERT[8]. Higher accuracy, compared with the Siamese CBOW, can be achieved by fine-tuning BERT on downstream tasks, but the model makes inference time much longer, the running efficiency does not meet the requirements of our online products. We propose a high-efficiency contextual referential solution based on syntax analysis to solve the problems of subject missing and pronoun resolution in the question-and-answer scenario in insurance industry that achieved good results. The voice input brings convenience to users but at the same time introduces typos in the results after the text processing. We use the insurance specific noun dictionary with the error correction model of Transformer[7] to improve the input from ASR. For the purpose of increasing the accuracy of matching sentences of the user's input with terms from Knowledge-Base, we use an efficient sentence compression algorithm, which can filter some insignificant content and retain some core content of the insurance industry. We rank all the answers from the retrieval module and do answer output finally. Our contributions are following:

- Propose novel and efficient error correction, sentiment analysis, coreference resolution, sentence compression and other methods to enhance question comprehension ability especially in insurance domain.
- Using ElasticSearch, deep semantic matching and KBQA combined the IR method to quickly recall matching questions. Improve the accuracy of the QA through deep learning rank while ensuring the overall efficiency of the system.
- Proposed a number of new industry test set construction methods and the QA evaluation methods.
- Full-life processing management and optimization for the QA knowledge including question type identification, clustering and annotation dispatch for no answer questions.

2 RELATED WORK

Most of the existing professional domain question answering systems search for the most matching questions (question in KB and user query similarity matching) from the Knowledge-Base through information retrieval. Some existing question-and-answer systems such as the

Ali Xiaomi and the Baidu AnyQ are single-round questions and answers that do not consider the context information. The AliMe from Alibaba, which combines the Knowledge-Base search and Seq2Seq generation, makes achievements in the e-commerce domain[2]. We use the same method as the AliMe and the Baidu AnyQ to match the question and user query similarity and consider context chat history at the same time.

3 SYSTEM OVERVIEW

Our overall system architecture is shown in Figure 1. The user's question (that is query) is used as input. If it is a voice, it will be converted into text first. The context information is passed to the pre-processing module. After error correction and coreference resolution, the processing is passed to the retrieval module. It returns the best matching with the user problem respectively from ElasticSearch based on text retrieval, the semantic retrieval based on the Siamese CBOW and the KBQA based on knowledge graph. The question list is passed into the sorting module, and the multi-way matching list is merged, and some unreasonable matching questions are removed through the entity alignment, and the final related question list will be generated through deep learning sorting. Finally, the answer will be returned to the user according to the matching question with business type.

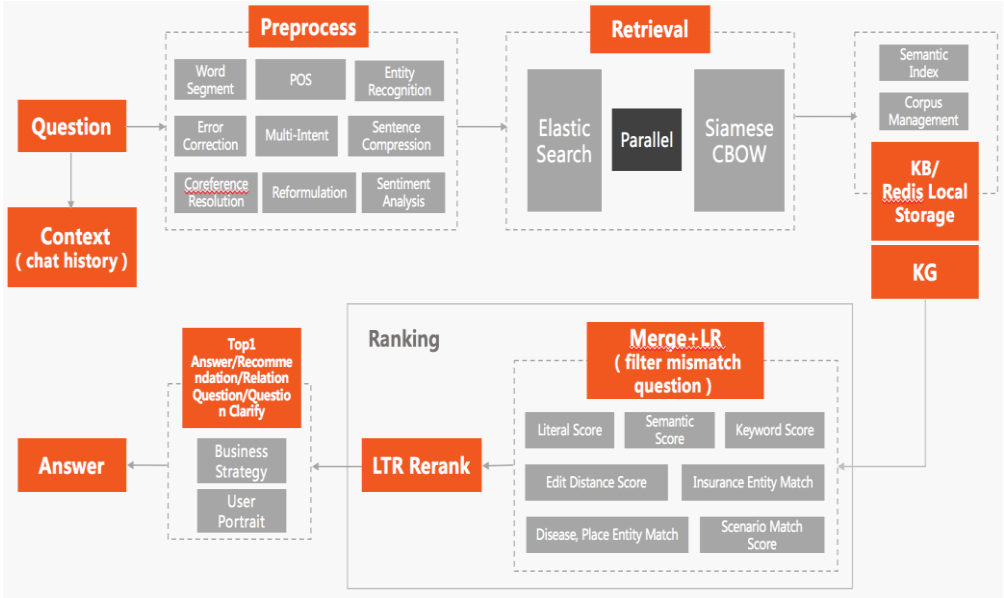


Figure 1: The overall architecture of our QA system

3.1 Pre-processing Module

We use open sourced NLP Tools with the insurance terminology dictionary for word segmentation, part-of-speech tagging and entity recognition. The multi-intention detection uses the method of splitting the sentence by punctuation and then classifying it. The question rewriting is mainly for the insurance product name, and the sentiment analysis is used to judgment the intents of the user's affirmation, negation and double negation. Following we describe more detail implements.

3.1.1 Long Sentence Compression

Step1: Divide the long sentence into several short sentences by punctuation or space, then classify the short sentences and remove the saliva statement

Step2: Based on the sentence compression scheme of probability and syntax analysis, we only retain the core sentence components. Combined with the insurance keyword dictionary to ensure the keywords are retained.

Example: Hello, I bought an insurance for my son in 2006 and I only paid 581 yuan for a year, however I didn’t pay for it after that. Now I want the customer service to refund my money.

Compress result: I bought an insurance in 2006. Now I want to refund my money.

3.1.2 Error Correction

Two solutions are used for business selection. The simple solution is based on the error correction of the insurance noun dictionary. According to the results of the previous word segment and syntactic analysis, the possible nouns are converted into PinYin and compared with the proper nouns in the dictionary for error correction. The general solution is the Transformer model with a special noun dictionary, the training datasets use about 32 million universal corpora from public news and the PinYin dictionary that from insurance domain. The input of encoder in the model is non-dictionary Chinese PinYin and Chinese word characters in the dictionary. The Decoder’s output is a pure Chinese character, where the Chinese characters in the input dictionary do not participate in the prediction, then directly generated.

3.1.3 Coreference Resolution

We use context chat history as Coreference Resolution reference. Our implementation ideas are word segmentation, part-of-speech tagging, dependency syntax analysis, subject-predicate extraction, entity substitution. For example:

(Question) What is the price of life insurance? (Answer) 300 yuan per year.

(Question) How about car insurance? (Coreference resolution result) What is the price of car insurance?

3.2 Retrieval Module

The retrieval module includes keyword search, deep semantic matching and KBQA recall, using the advantages of each of these three methods to increase the number and diversity of recall answers. The keyword search is retrieved using the open source ElasticSearch (ES) engine. As for the deep semantic retrieval, we use the deep semantic model to perform semantic vector representation on the user query and the knowledge in the knowledge-base (standard question and extension question), and use the Annoy algorithm to quickly find and match the semantic vector. The deep semantic model is modeled using the siamese network [5]. For each query, the similarity of the annotations is used as the positive sample, the negative samples are generated by the random sampling method, and random sampling is performed for each iteration, which greatly increases the randomness of the training data and improves the generalization ability of the model. Inspired by the idea from the loss definition of face recognition, we use AM-Softmax as the loss function and achieved the best results.

$$X_i = normalize(\sum_{w\ in\ s_i} embed_w)$$

$$\cos\theta_{y_i} = X_{\mathrm{q}}.X_i^T$$

$$L_{AMS} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s.(\cos\theta_{y_i}-m)}}{e^{s.(\cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c}e^{s.\cos\theta_j}}$$

The sentence vector X_i is obtained by normalizing the summation of word embedding in sentence S_i. The $\cos\theta_{y_i}$ is the similarity of user query vector and question i's vector. Both s and m are hyperparameters where s is the scale factor and m determines the classifier's boundary size.

In terms of feature extraction, for the purpose of extracting local word order relationships and context information better, we use LSTM, CNN, BERT and other networks to extract features. BERT performs best, but it takes a long time for online inference. Due to the limited quality of large-scale industrial corpus annotation, some data noise exists. The more complex models the more noise is fitted so the generalization ability is not as good as the simple model. Therefore, we chose the CBOW[4] model for feature extraction. Considering that Chinese word segmentation has limited effect in specific fields. To reduce the influence of word segmentation errors, we also use multiple dimensions of pre-training vectors to build our model, including: character embedding, word embedding, high-frequency phrase vector, where character embedding can solve literal matching, word embedding can represent the semantics of words, and phrase vectors can capture local-level word order relationships and achieve the best results.

We have done some benchmarks by using insurance domain dataset in different models also, the result show as following:

Method	Siamese LSTM	Siamese CNN	Siamese CBOW	BERT
Recall	80.6%	83.5%	85.2%	88.9%

Table 1: Benchmark results of Deep Semantic models

In KBQA, it receives the pre-processed question information, characterized by the context information, the entity type, and the entity relationship, and predicts the subject entity to be queried through the question recognition model[1], and the neighboring nodes centered on the entity from the KG.

3.3 Ranking Module

The ranking module includes a deep ranking model and a rule sorting. The deep ranking model is mainly used to merge and score the answers of multiple recalls. The rule sorting is mainly used to verify the rules of the sorted answers again to ensure not only the stability but also reasonability of the sorted answers. In the choice of deep ranking model, we use the commonly used pair-wise ranking model. Owing to the model is less difficult for data collection, we define the format of the input sample as the pair of ***<user query, candidate queries>*** when modeling. By constructing the scorer, the scores of the correctly matched samples are as high as possible (normalized to [0,1]) and the scores of the mismatched samples are as low as possible. The deep ranking model uses the interaction model, which not only considers the semantic vectors

of these two parts but also considers the calculation of the interaction information of these two parts so that it could get more accurate matching. In addition to semantic features, our model uses co-occurrence words in ***<user query, candidate queries>*** to model literal features. To better match the user's intention, we construct an intent classifier in the insurance industry, and perform intent feature extraction on the user query and the candidate queries respectively as input to the sorting model. In addition, we have made some attempts on the sentence features and get good results. As the Figure2 shows. We have:

$$\text{score}(\text{q}, \text{d}) = \text{FN}([X_d, X_q, X_{addtion}])$$

$$\text{L}(\text{q}, d^+, d^-; \theta) = \max(0, 1 - \text{score}(\text{q}, d^+) + \text{score}(\text{q}, d^-))$$

$\text{score}(\text{q}, \text{d})$ is the matching score of query and document ,while $X_d, X_q, X_{addtion}$ are the inputs of the neural network; $\text{L}(\text{q}, d^+, d^-; \theta)$ is the hinge loss of a train sample pair. Taking into account the professional requirements of the question-and-answer in the insurance field and the fact that sorting model cannot achieve 100% accuracy, we have added a priori knowledge of the insurance industry in the ordering of rules to ensure the professionalism of question-and-answer. Rule sorting mainly considers the alignment of professional entity information between the user query and the candidate question, and the best matching question should be consistent with the entity described by the user query and avoiding give an irrelevant answer.

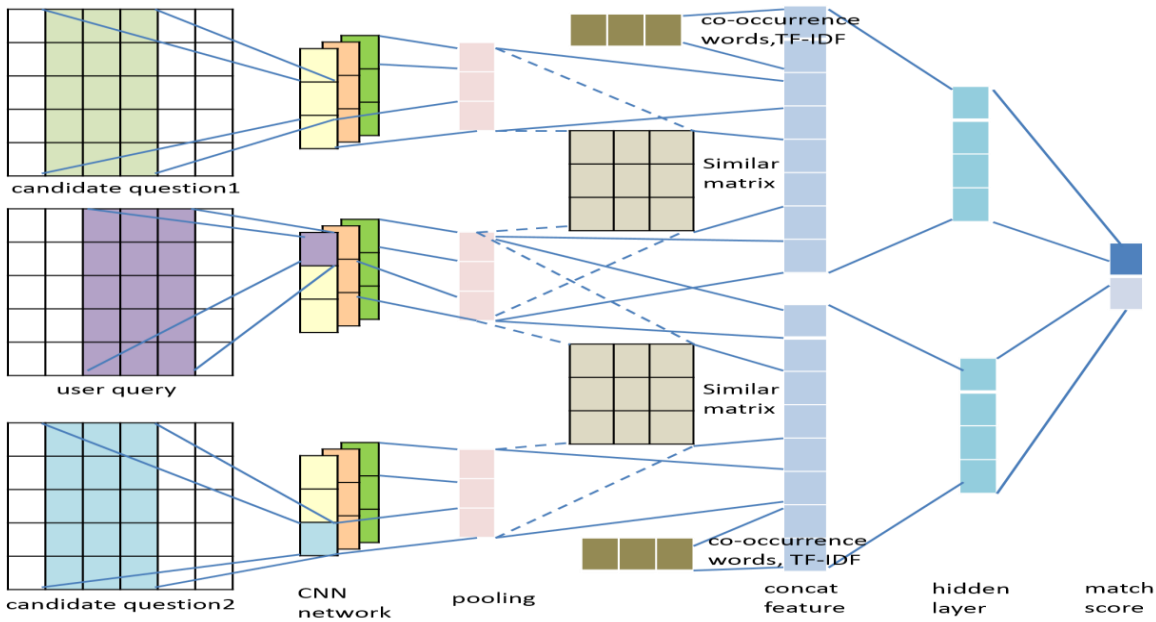


Figure 2: Our Deep Rank Architecture

3.4 Output Module

It gets the matching question list from rank module. If the confidence level is lower than the preset threshold, it will response a question to have user clarification and let the user to confirm the question he wants to ask and make a related question. If the confidence is high, the answer corresponding to the top one matching question or the recommendation question is returned according to the business rule.

3.5 Intelligent Knowledge-Base

The intelligent Knowledge-Base is a behind-the-scenes role in the Q&A system. In addition to providing the FAQ engine with raw materials, it also manages and optimizes the life-cycle of the question-and-answer knowledge. The specific process can be seen in Figure 3.

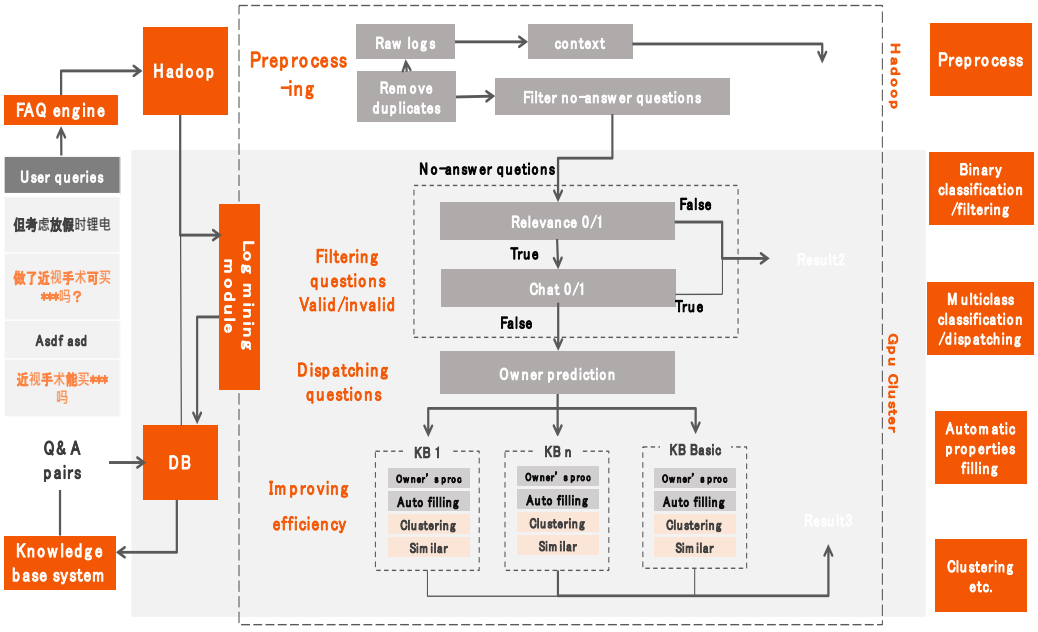


Figure 3: Our Intelligent KB Module Architecture

4 Evaluation Metrics

The Q&A assessment indicators mainly include the number of valid questions, Top1 response accuracy, Top3 accuracy, effective question response accuracy and knowledge coverage. The test set was divided into 5 categories, which are online log sampling used for the evaluation of model, the bad case collection, high frequency question mining used for algorithm regression testing, semantic test sets written according to demands fully cover the requirements, and the

corpus that delete the non-keywords, increase noise, synonym transfer and other methods to generate the literal test set to evaluate the robustness of the model. Our system achieved good results in these insurance business test sets and provided online service for one hundred million customers.

5 Conclusions

This paper proposes a context aware, error correction, coreference resolution, long sentence compression, ElasticSearch and deep semantic matching with the Siamese CBOW and deep learning sorting for the question-and-answer system. Our approaches not only have good performance in engineering but also in model accuracy. Its architecture supports high concurrency requirements in real world use cases and has high availability that fits the standard production environment. We have already applied this system in on-line intelligent customer service bot, AI assistant, AI selling bot and other human-computer interaction AI products. In the future, we hope our question-and-answer system could support multimedia interaction, such as pictures, audios and videos in addition to text and voice so that we could solve more problems for users with more intelligence.

REFERENCES

[1].Yunqi Qiu, Manling Li, Yuanzhuo Wang, Yantao Jia, Xiaolong Jin,2018, Hierarchical Types Constrained Topic Entity Detection for Knowledge Base Question Answering,ACM 2018, April 23–27, 2018 , Lyon, France.

[2] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, Wei Chu, 2017, AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine, ACL 2017 pages 498-503

[3] Kim Y. Convolutional Neural Networks for Sentence Classification[C] Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.

[4] Tom Kenter,Alexey Borisov,Maarten de Rijke, 2016, Siamese CBOW- Optimizing Word Embeddings for Sentence Representations, ACL 2016

[5] Paul Neculoiu,Maarten Versteegh,Mihai Rotaru, 2016, Learning Text Similarity with Siamese Recurrent Networks, ACL 2016 Proceedings of the 1st Workshop on Representation Learning for NLP, pages 148-157

[6] Aliaksei Severyn,Alessandro Moschitti, 2015, Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks, Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval pages 373-382

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, 2017, Attention Is All You Need, NIPS 2017

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Computation and Language 2018

[9] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner,Christopher Clark, Kenton Lee, Luke Zettlemoyer, 2018, Deep contextualized word representations, NAACL 2018

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 82-91

A Real-World Human-Machine Interaction Platform in Insurance Industry

Wei Tan, Chia-Hao Chang†, Yang Mo, Lian-Xin Jiang, Gen Li, Xiao-Long Hou
Chu Chen, Yu-Sheng Huang, Meng-Yuan Huang, Jian-Ping Shen
AI Department
Ping An Life Insurance of China, Ltd.
{tanwei818,zhangjiahao206†,moyang853,jianglianxin769,ligen947,houxiaolong430,
chenchu870,huangyusheng112,huangmengyuan334,shenjianping324}@pingan.com.cn
Correspondence†:strategist922@gmail.com

Abstract

In the insurance industry, lots of effort is putting into helping the customer to solve their problems that occurred during and after purchasing cycle and helping telemarketers to practice selling skills. Chat bots and assistant bots are widely used in these business scenarios, but building a bot application from scratch is expensive. In this paper, a human-machine interaction platform specially designed for intelligent bot applications in insurance industry that combined the technologies of Question Answering (QA), task-oriented dialogue and chit-chat was proposed and we demonstrate the architecture design of this platform, key technologies and the scenario of applications in real-world insurance industry. It has been supporting many intelligent bot applications of insurance industry already, such as Intelligent Coach Bot (ICB) which helps telemarketers to practice their selling skills, Intelligent Customer Service Bot (ICSB) which provides after-sales services and Insurance Advisor Bot (IAB) which helps customer to purchase the most suitable insurance product. Currently, these bot applications serve millions of users per day and are able to solve 80% of the online problems.

KEYWORDS

Human-Machine Interaction Schema, Dialogue Tree, Text Similarity, Text Matching, Chit Chat Engine, Anaphora Resolution, Intention/Slot Driven, Attention Mechanism

1 INTRODUCTION

All authors contributed equally to this manuscript.

With the development of deep learning technology, speech recognition and natural language understanding (NLU) have developed rapidly and the way of human-computer interaction has also changed from GUI to CUI (Conversation User Interface). The traditional GUI depends on the keyboard, mouse, or touch screen. But CUI carries out input and output through dialogue, breaking through the limitation of contact and getting closer to the way of communication between people. CUI is a tremendous change in the way of interaction with a variety of AI techniques. We have implemented a human-computer interaction platform that uses CUI as a starting point to construct a conversational input and output interface and a smart brain. Based on the insurance scenario, our platform has been widely and solidly applied in the fields of salesman training, intelligent customer service, and recommendation of products.

In salesman training scenario, we rely on this platform to build intelligent coach robot that can simulate customers to do human-machine dialogue with the telesales man. It helps them learn sales and handle customer objections, improving their sales skills. In the intelligent customer service scenario, we built a customer service robot under the technical support of the platform which can answer users' questions at anytime and anywhere and guide users to apply the request for business, such as the policy loan. In the scenario of insurance products recommendation, we implemented the insurance advisor bot that could find customers' demands from the human-computer dialogue and recommend insurance products that our customers might need. All of these bots were built on the human-computer interaction technology platform. They share basic technologies provided by the platform, such as intent recognition, semantic understanding, dialogue management, and a unified knowledge center. The main contributions of this paper include: (1) We design a set of human-computer interaction application solutions and based on these to develop the coach bot, the intelligent customer service bot and intelligent insurance advisor bot. (2) We propose the human-machine interaction schema which uses dialogue tree of intentional organization as the core. (3) We propose and implement a calculation method of text similarity that integrated various of technical methods which used for text matching. (4) We propose and implement the intent detection, recognition of chit chat based on the classification technology. (5) We propose and implement a recommendation technical solution based on dialogue.

2 SYSTEM OVERVIEW

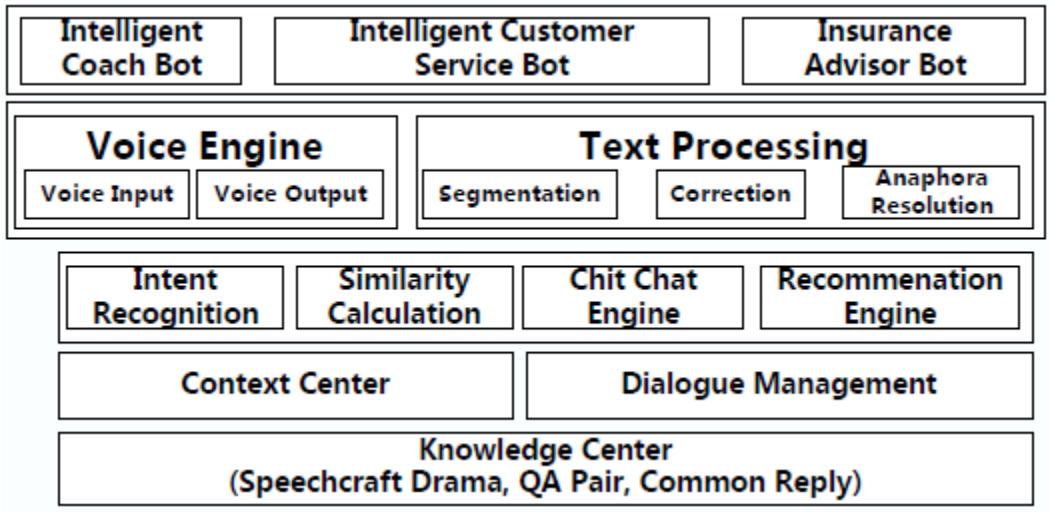


Figure 1: The Overall System Architecture

In Figure 1, it shows the overall architecture of the human-computer interaction technology platform. The first layer is the input and output layer. The system supports the interactive form of speech and text. The input speech needs to be recognized by the ASR engine, and the text needs to be synthesized into the voice output by the TTS engine. Text processing module will pre-process the text input or the text comes from the ASR engine. The pre-processing step includes word segmentation, error correction and anaphora resolution. Using uniform pre-processing module can ensure the consistency of the results. By tuning this module in the system, it will be applied to all downstream modules. The second layer is the basic technical support layer. It includes intent detection, text similarity calculation, the recommendation engine and chit chat engine. The third layer is the context center and dialogue management module used to track the whole human-machine dialogue and manage slot information. The fourth layer is the knowledge center, which includes the speech craft drama, QA pairs, common replies and so forth. The entire human-computer interaction technology platform can support a variety of business scenarios, such as intelligent coaching, intelligent customer service, and insurance products recommendation, which is demonstrated on top of Figure 1. The three bots share basic components of the platform, like the Voice Engine and Text Processing modules. The Intelligent Coach Bot uses Intent Recognition and Similarity Calculation modules for NLU (Natural Language Understanding), and Chit Chat Engine is used to generate reasonable replies when a user has said some task-unrelatable sentences. For Intelligent Customer Service Bot, QA is especially important, so it integrates Similarity

Calculation including semantic similarity and literal similarity as its sub-module. Both Intelligent Coach Bot and Intelligent customer Service Bot use Content Center and Dialogue Management for context understanding and dialogue state tracking [14]. As for Insurance Advisor Bot, Recommendation Engine and Content Center is the core module.

3 SYSTEM FEATURES

3.1 Dialogue Management

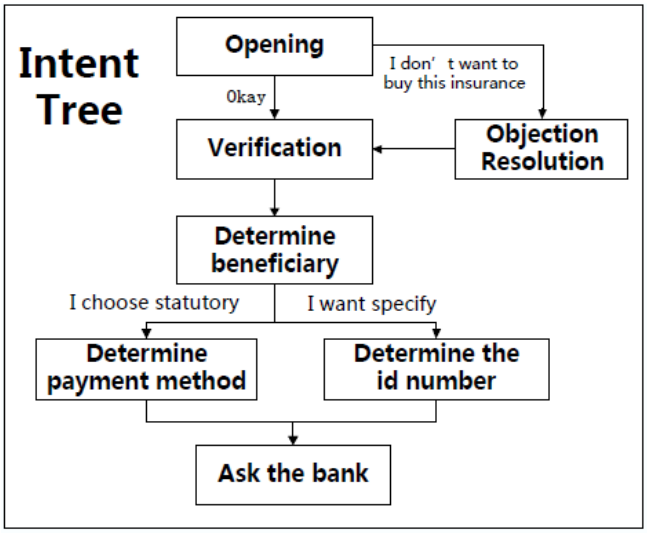


Figure 2: Dialogue Management

In the man-machine dialogue, we need to identify the user's intention or extract the slot value from the conversations. The basic framework for our implementation shows in Figure 2. In salesman training scenario, the system was configured with a lot of dialog dramas in advance. The purpose is to train salesman to improve sales skills by using the system to simulate as a customer and perform human-computer dialogue. The salesman needs to sell the insurance product to the customers according to the pre-defined intention of the script. Each intention has a standard speech and allows the salesman to have a certain degree of free play, but cannot deviate from the main meaning. The customer a machine simulated will give an ordinary response or an objection during the selling process. For example, “I don’t want to buy your insurance” is the customer’s objection, and the salesman needs to resolve the objection before entering the next

round of dialogue. In order to give maximum flexibility for agents, we organize drama by using intent dialog tree, one drama corresponding to an insurance product is organized as one dialog tree. We use the quadruples ***(intent_name, objection, finished_intents, prattle_times)*** to save the information of dialogue status and store in the Redis cache system. In a conversation, ***intent_name*** is used to track the salesman's intent node's position in the last round of dialogue tree, ***objection*** indicates whether the machine raises the objection or not in the last dialogue round, ***finished_intents*** is used to track the intents of the entire conversation. Owing to the limitation on the number of consecutive chats, ***prattle_times*** is used to record the number of chats.

3.2 Intent Detection

In the human-machine dialogue process, the system needs to identify the intention of the agent or the intention in the customer's dialogue. This is a text classification task [1], we use the FastText tool. An important reason why we chose FastText is its fast inference speed so that we can make a robot dialogue system has a good interactive experience. When training, we try to join the n-gram feature, after the addition of this feature the size of the model will reach nearly 1GB, while F1 score of our model is only slightly improved by 0.3%. In our use case, we do training for each dialog drama with an individual model. In order to make memory consumption at a reasonable level, we did not use n-gram feature choice in the final model.

Model	FastText	FastText+2-gram	BERT
Avg. of F1	0.9664	0.9694	0.97245
Avg. Time of Prediction	48.46 μs / sentence	59.38 μs / sentence	213333 μs / sentence
Model Size	8.9M	986M	1.1G

Table 1: Benchmark Results

In Table 1, it shows the performance benchmark results with the FastText model, FastText+2-gram features model and BERT model [13], by using the intent classification dataset. The dataset contains 36 categories, and each category contains 1536 sentences. This dataset corresponds to a

drama, and every category is an intent like “opening”, “verification”. The corpora are collected by many sophisticated salesmen, they divide the insurance sale into several key processes and each of them corresponds to an intent, e.g. the intent “verification” means verify the identity of a customer. We use 80% of the dataset for training and other 20% of the dataset as the test set. We get one F1 score for every category and compute the average F1 over all categories as show in Table 1.

3.3 Semantic Calculation

In the intent-based dialogue, customers will raise objections or questions. The questions need to be matched with the corpus in the knowledge base. The system will match the customer’s objections or questions with the collection set of customer objections in the knowledge base or in our FAQ set one by one. In our system, we used a combination of literal similarity, and Siamese CBOW [2].

3.3.1 Literal Similarity Algorithm

Based on the literal similarity of edit distance, the Levenshtein Distance algorithm is used to calculate the minimum number of steps to transform sentence A into sentence B by adding, deleting, and replacing operations. After finding the edit distance of Levenshtein, the similarity of the two sentences is calculated by the following formula: ***Similarity = (Max(x, y) - Levenshtein) / Max(x, y)***, where ***x***, ***y*** are the length of the sentence ***A*** and ***B***.

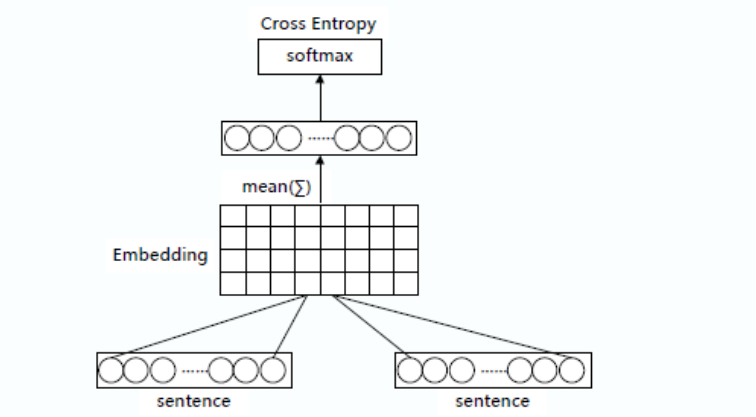


Figure 3: CBOW Optimization

3.3.2 CBOW Optimization

We have optimized the semantic representation of the word vector and adjusted the word vector at the sentence level. The model structure we use was shown in Figure 3. For each sentence, we select a positive case and several negative cases, and calculate the sentence vector according to the formula ***SentenceVec(A)=1/n∑Vec(wi)***. The cosine similarity of the selected sentence and the positive and negative examples is then calculated and normalized using the SoftMax function and uses cross entropy as the loss function. The embedding layer uses a pre-trained Glove word vector and adjusts the word vector weight by adding the sentence similarity as considerations.

3.4 Chit Chat Recognition and Reply

We pre-defined some common Chit Chat response categories and answer templates. The template defines some frequent Chit Chats. For a Chit Chat conversation that cannot be classified as a pre-defined Chit Chat category, we generate a chat response by using the seq2seq model [4] [15] [3].

3.5 Dialog-Based Recommendation

We proposed a dialogue recommendation model show as in Figure 4 to effectively recommend insurance products to users. The dialogue model includes three sub-networks: the graph recommendation sub-network, the semantic recommendation sub-network, the behavior recommendation sub-network and the output layer is a weighted sum of the scores of networks.

The input of the graph recommendation sub-network is the entity embedding in the query. The intelligent insurance advisor bot has established a knowledge graph for insurance approbation and insurance indemnity, with about 7,000 nodes. Based on the knowledge graph, system can learn the entity embedding which including insurance products, diseases, and people. The graph recommendation sub-network employs the attention mechanism. The point multiplication is used to expose the similarity between the insurance product and the entity, which is efficient and performs well. The input of the semantic recommendation sub-network is the word embedding in the query. The word embedding is pre-trained. The insurance, disease and other professional terms are very helpful for the model. Since the conversations are generally short texts, using the CNN

model is more efficient, and the semantic recommendation sub-network uses one-layer convolution-pooling CNN. The input of the behavior recommendation sub-network is the item embedding of the insurance products in the query. Using the user click behavior data in intelligent insurance advisor bot mobile application, item embedding can be learned by item2vector model [5]. The behavior recommendation sub-network employs the attention mechanism and the attention value is calculated by using point multiplication between clicked item and target item. Increasing the behavior recommendation sub-network AUC can increase by 1.2%.

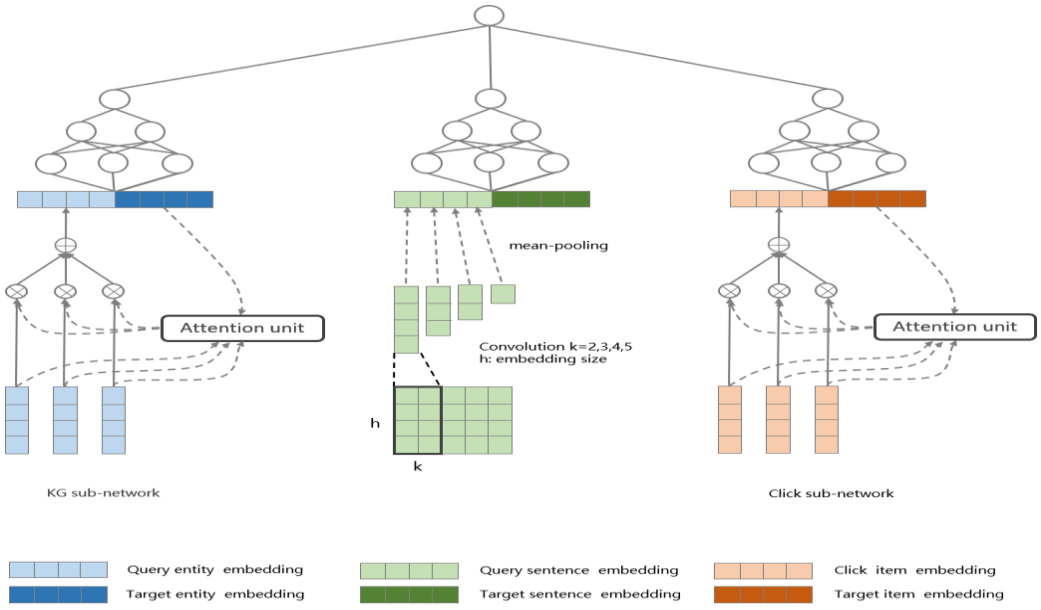


Figure 4: Dialog-Based Recommendation

4 RELATED WORK

Dialog Management. This is a core module of human-computer interaction system, which solving the problem of dialogue state tracking [6] and dialogue logic processing. Here we use the dialog tree to manage user intent and state.

Intent Recognition. In the human-machine dialogue system, intent recognition is crucial. Traditional intent recognition methods typically use a template-based approach. With the development of neural networks, intent recognition based on CNN[7] and RNN[8] methods have emerged. We used the FastText method achieving an accuracy of 86%.

Semantic Computation. It is possible to compute the semantic representation of the query and then retrieve the knowledge base through the semantic matching technology to obtain the best matching question and the corresponding answer. The popular methods of semantic matching include literal matching and semantic matching (semantic matching models include DSSM[9], CLSM[10], Deep Math[11], Siamese). In order to improve the coverage of matching, we use the literal similarity and Siamese CBOW model [2] and set the corresponding weight for each method's results.

Chit chatting. Commonly used methods for open domain chit chatting include IR-based [12] and generation models. We use the IR-based methods on the priority, and use the generative methods when the answer can't be found with IR-based methods, as insurance field is financial related, we need to give a rigorous answer.

Dialogue Based Recommendations. In the human-computer interaction scenario, we recommend products, tasks, and news to users based on user behavior and conversation content to improve user experience and conversion rate in human-machine applications. We constructed a dialogue recommendation model consisting of three sub-networks (semantic recommendation sub-network, map recommendation sub-network, behavior recommendation sub-network).

5 CONCLUSION

The human-machine interaction platform proposed in this paper has been supporting many intelligent bot applications of insurance industry already, such as Intelligent Coach Bot (ICB) which helps telemarketers to practice their selling skills, Intelligent Customer Service Bot (ICSB) which provides after-sales services, and Insurance Advisor Bot (IAB) which helps customer to purchase the most suitable insurance product. Currently, these bot applications serve millions of users per day and are able to solve 80% of the online problems. In the future, the improvements of the platform include intent recognition and semantic computing based on BERT, end-to-end tasks modeling, scene-based sequence labeling and shopping guide based on reinforcement learning.

REFERENCES

[1] Sun, Y., Wong, A. K., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687-719.

[2] Tom.Kenter, et al, 2016, Siamese CBOW: Optimizing Word Embeddings for Sentence Representations

[3] Luong, T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention based Neural Machine Translation. empirical methods in natural language processing,, 1412-1421.

[4] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. neural information processing systems,, 3104-3112.

[5] Barkan, O., & Koenigstein, N. (2016). ITEM2VEC: Neural item embedding for collaborative filtering. international workshop on machine learning for signal processing,, 1-6.

[6] Mahew Henderson. 2015. Machine Learning for Dialog State Tracking: A Review. In MLSLP'15.

[7] Hashemi H B., Asiaee A, Kraft R, Query intent detection using convolutional neural networks, International Conference on Web Search and Data Mining, Workshop on Query Understanding , 2016.

[8] Bhargava A, Celikyilmaz A, Hakkanitur D, et al. Easy Contextual Intent Prediction and Slot Detection, IEEE International Conference on Acoustics. IEEE, 2013: 8337-8341.

[9] Po-Sen Huang, et al., 2013, Learning Deep Structured Semantic Models for Web Search using Clickthrough Data

[10] Yelong Shen, et al, 2014, A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval

[11] Zhengdong Lu & Hang Li, 2013, A Deep Architecture for Matching Short Texts

[12] Zhao Yan, Nan Duan, Jun-Wei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. 2016. DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents. ACL'16

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[14] Julien Perez, Fei Liu. 2013. Dialog state tracking, a machine reading approach using Memory Network. arXiv preprint arXiv:1606.04052

[15] Jiatao Gu, Zhengdong Lu, Hang Li, Victor O.K. Li. 2016**.** Incorporating Copying Mechanism in Sequence-to-Sequence Learning. arXiv preprint arXiv:1603.06393

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 92-101

結合 LDA 與 SVM 之社群使用者立場檢測

Stance Detection of Social Network Users by combining Latent Dirichlet Allocation and Support Vector Machine

翁翊桓　I-Huan Weng
國立臺北科技大學資訊工程學系
Department of Computer Science and Information Engineering
National Taipei University of Technology
t106598025@ntut.org.tw

王正豪　Jenq-Haur Wang
國立臺北科技大學資訊工程學系
Department of Computer Science and Information Engineering
National Taipei University of Technology
jhwang@csie.ntut.edu.tw

摘要

傳統的立場分析常常使用問卷調查、電話訪查等來得知不同的主題下每個人的觀點。但由於傳統的統計方法，採用抽樣的方式，容易因為樣本數的不足，導致效果較差。現有的方法包括以情緒字典、以卷積式類神經網路(CNN)、遞歸式類神經網路(RNN)等，但是因為深度類神經網路需要較多資料集才能提升效果。而文本的特徵則採用 N-Gram 或是 TF-IDF 方法，但這樣無法真正了解文本的語意。本論文提出利用 Word2Vec 字詞表示模型，來取得字詞的向量，並結合 LDA 方法來取得文本的特徵。在立場檢測方面，我們以 SVM 作為分類器，以兩階段方法分辨人們是否中立與否的主觀性問題，並預測使用者的立場。

本論文以 SemEval-2016 的立場偵測任務，作為實驗的資料來源，並使用多種方法(F-Measure, Accuracy, Precision, Recall)來評估效果，相較於 SemEval-2016 的基線或其他隊伍分數，平均而言，本論文所提的方法皆獲得較好的結果 (F-Measure : 83.36%)。

Abstract

In traditional stance analysis, questionnaire survey or telephone survey are often used to know the opinions of each person under different topics. However, due to the traditional statistical methods, the sample size is too small to get good result. Existing methods are usually based on sentiment lexicon, Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN). And the text features are based on N-Gram or TF-IDF, which do not help to understand the semantics of the text. This research proposes to use Word2Vec for word embedding and combine the LDA to obtain the text feature. For stance detection, we use Support Vector Machine (SVM) to train the classifier to detect the subjectivity of texts, and to predict user stances.

In the experiment, we used the data from SemEval-2016 Stance Detection Task, and use a variety of evaluation methods (F-Measure, Accuracy, Precision, Recall) to evaluate performance. Compared with SemEval-2016 official baseline and other teams scores, our proposed method can get better result on average (F-Measure : 83.36%).

關鍵詞：立場檢測、機器學習、社群網路分析、Word2vec、隱含狄利克雷分布

Keywords: Stance Detection, Machine learning , Social network analysis , Word2vec

一、緒論

人們在生活中針對的目標不同時，人們的立場也會變得不一樣，通常人們有所衝突時大多數都是因為立場不同，像是電影評論、產品意見、總統大選等有關的問題。這些問題通常都會被人所收集來進行探討與分析。早期的方法通常是以電訪或紙本問卷來進行抽樣的調查，所以容易因為人力與樣本數量的關係，進而影響到預測的結果。

近年來，網際網路的普及，造就了許多的社群網路平台像是：Twitter、Facebook，而根據 Statista 的數據統計[1]指出：全球社群媒體的使用者在 2019 年估計有 27.7 億人，而光是在台灣，社群網站的使用者數量就佔了總人口 89%，能得知社群網站對於人們來說成為了生活中不可或缺的存在。使用者平常會在這些平台發表自己對不同主題的想法或是意見，因此每天都會有許多訊息存在於社群上。若以傳統的方法來針對這些大量的資訊來對使用者與貼文進行分析，除了需要配置許多人力與成本花費，還需要長時間才能有所結果，而且新的訊息還會隨著時間大量增加，因此透過機器學習的技術與系統自動

化來處理數據的分析將是未來的趨勢。

由於現有文本處理通常都只以提用詞移除(Stopword removal)作前置處理，因此本論文另外再使用詞型還原(Lemmatization)與詞幹提取(Stemming)的方法來評估對於分類器的影響。在立場檢測上，本論文提出以透過 LDA 的方法，來產生主題特徵並與經由 Word2Vec 所轉換的特徵向量作結合，來進行立場的檢測。

最後，根據本論文之實驗，在使用 LDA 結合特徵向量時，確實能使各個目標主題的準確率提升，最高的 F-Measure 為 84.23%，平均 F-Measure 為 76.88%皆高於 SemEval-2016 上的 Baselines 與其他隊伍，因此可以驗證所提出方法是有效的。

二、相關研究

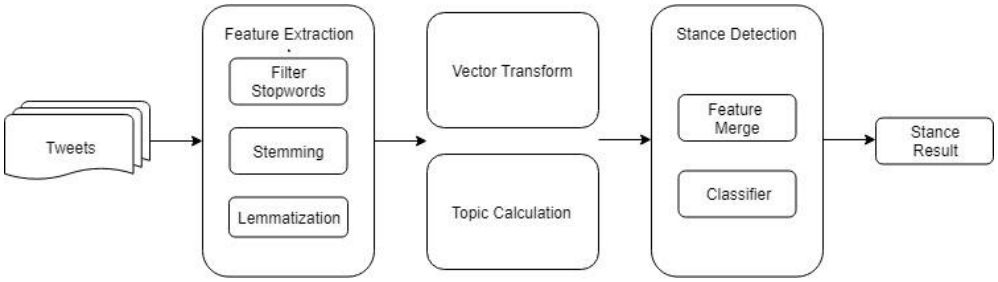
Jannati 等人[2]透過收集部落格的文章來檢測對於政治人物的立場，並以情緒辭典作分析；Tumasjan 等人[3]對 Twitter 上的貼文進行分析並依其結果來預測 2009 年的德國總理大選，來證實社群平台上的貼文確實可以反映出人們的立場；Sasaki[4]等人則是對 Twitter 的貼文添加額外的提示標籤，實驗發現透過添加的額外事件特徵可以提升立場檢測的準確度。 Tutek 等人[5]透過 SVM 對文本作 N-Gram 來進行立場檢測；Bøhler 等人[6]使用 GloVe 詞向量模型，比較了 Naive Bayes, SVM 這兩種分類演算法的效果，發現結合 GloVe 詞向量的特徵能提高檢測效果；[7, 8]使用卷積神經網路(CNN)來進行立場檢測，其中使用 Semeval-2016 的資料集在 F-measure 有 67.33%；Zarrella 等人[9]透過遞歸神經網路(RNN)並使用預訓練的特徵來進行立場檢測，使用 Semeval-2016 的資料集在 F-measure 有 67.8%。根據 Igarashi 等人[10]的觀察發現，深度學習方面效果沒有一般的分類好；而 Mohammad 等人[11]也發現在立場檢測上，SVM 的整體平均高於其他模型。基於資料集的數量與前者的觀察，我們採用 SVM 作為主要的方法進行立場檢測，並在實驗中與 Naive Bayes, LSTM 等分類器作比較。

在相關的論文中，大部分的平台都是以 Facebook 或是 Twitter 平台為主作分析，且主題大多都圍繞在政治上。而目前常用的方法是使用針對文本中目標的情緒極性，這樣方法有些缺點，像在單一目標或跨領域時效果通常不佳，為了提升立場的主題多樣性和準確率，本論文將利用 LDA 主題模型的特性，結合 LDA 立場模型與 Word embedding，期

望能達到更良好的結果。

三、研究方法

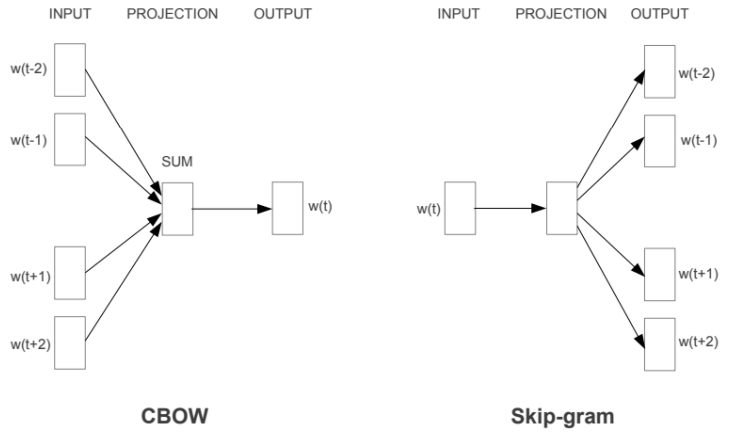
本研究所提出之方法主要可以分為四大部分，一開始會接收文本的輸入，並透過 Feature Extraction 來對輸入的文本進行 Filter Stopwords、Stemming、Lemmatization 來取得文本的特徵。之後分為兩個階段進行，第一階段會透過 Topic Calculation，以 LDA 主題模型來取得各個文本的主題特徵，並對各個文本做主題分佈的標記；第二階段則是透過 WordVec[13]字詞向量空間模型，將各個文本的字詞轉成向量表示。最後合併第一階段與第二階段的結果，透過監督式學習法來訓練分類器並取得立場結果。



圖一、系統架構圖

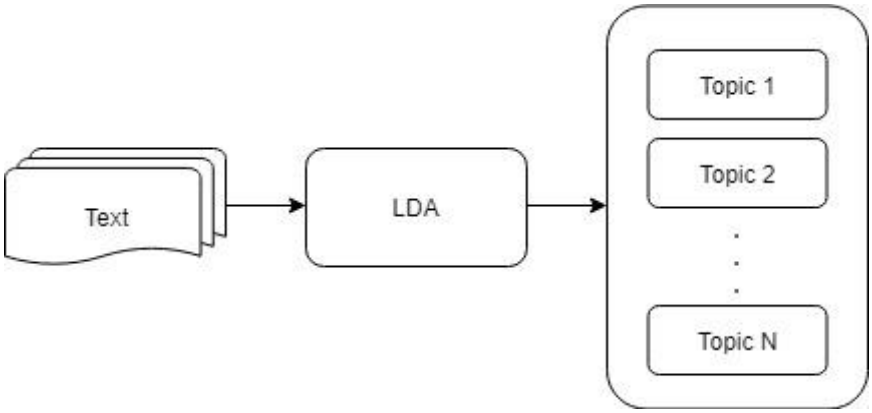
一個文本是由許多的文字所組成的，而文本中有許多字其實是不必要的，從文本中找到有意義或是重要的字，即可獲得較好的數據來進行分析。本研究從文本中移除 Stopwords，接著以詞幹提取(Stemming)、詞性還原(Lemmatization)等方法，來取得文本的完整語義。本研究使用 NLTK 中的斷詞 Stopwords 清單，過濾掉屬於停用詞的文字來提供後續的步驟作使用。詞幹提取是透過抽取字詞的詞綴來獲得詞根的方式，目的是讓變化的字詞簡化，使得文本分析能獲取到較好的字義來使用，本研究使用被廣為接受的 Porter Stemmer 來進行詞幹提取。詞型還原能把一個任何型別的字詞還原為原型，目的是能讓將字詞簡化為最初的字詞原形，使得文本分析能獲得字詞的完整語義，用在更為精確的自然語言處理上的文本分析與表達。將文本的內容透過機器學習進行分類前，必

須先將文本的內容轉換成向量，作為訓練分類器前的輸入。為了表示字詞的語義關係，本研究使用預先訓練好的 Word2Vec 進行文本的向量轉換，以便處理後續的實驗與步驟。Word2Vec[14]是由 Google 的 Tomas Mikolov 等人於 2013 年所提出的一種 Word Embedding model，在 Word2Vec 中，透過神經網路的方法，將文本中的文字與文字的關係轉化成具有語義關係和語法結構的向量形式。Word2Vec 中有兩個模型，如圖二，一個為連續型詞袋模型(CBOW)以及跳躍式模型(Skip-gram)兩種。CBOW 是透過輸入的上下文來預測字詞，而 Skip-gram 是透過輸入的字詞來預測上下文。



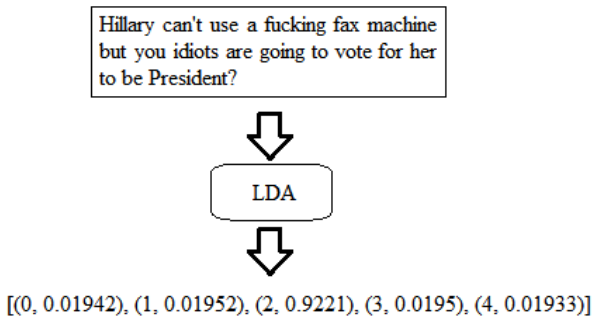
圖二、CBOW 以及 Skip-gram 之 Word Embedding 模型示意圖

在訓練字詞模型的時間上，Skip-gram 相較於 CBOW 的訓練時間會較久，但是在語意分析上，Skip-gram 比 CBOW 還來得好，本研究所使用的 Google 訓練好的模型，是基於 Skip-gram 上作訓練的模型。本研究使用經過前處理的文本，並利用主題模型，對每篇文本作標記，來取得主題特徵。在這一篇章節中，將會說明如何取得，並使之當作本研究特徵。隱含狄利克雷分布(Latent Dirichlet Allocation)簡稱 LDA，是由 Blei[12]等人在 2003 年所提出的主題詞袋模型。利用不同的機率的潛在主題，來描述每篇文章，而每一個主題是由分散的主題字詞所形成的。LDA 主題模型也是一個生成模型，他將每篇文本的主題按照不同的機率來表現每篇文章。LDA 的生成步驟大致上為圖三所示，輸入文本，之後經由 LDA 輸出主題數 N 的各個主題。



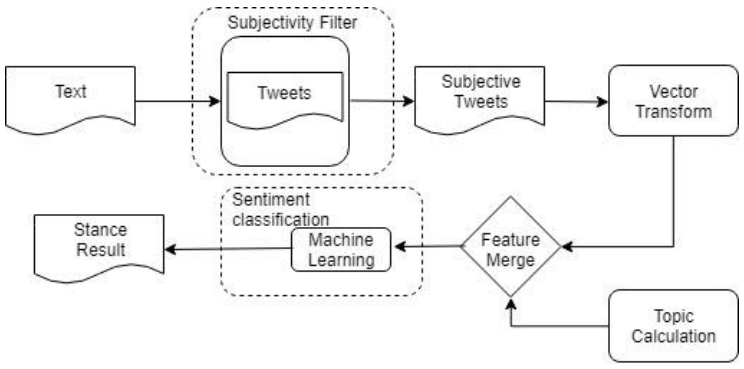
圖三、LDA 生成步驟

主題特徵的擷取步驟為圖四所示，輸入一篇文本，經由 LDA 取得該文本在各個主題中的分布概率。本研究採取文本的主題分布概率來作為主題特徵，並與所產生的字詞向量特徵進行結合，來進行後續的實驗。



圖四、LDA 文本主題分布概率示意圖

為了找出文本作者在貼文中所隱含的立場，我們使用機器學習的方式來針對貼文的內容作立場的分類，在這邊透過兩階段的方法，透過找出貼文者的主觀性立場與情感極性，來預測使用者的立場。



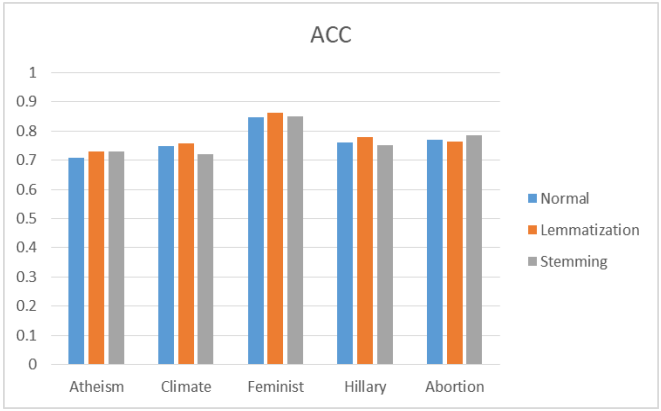
圖五、Stance Detection 架構圖

在立場判斷上，推文的人對於事情的主觀與客觀極為重要，具有中立立場的推文具有非主觀性的看法，而在這邊具有支持與反對的訊息則會有非中立情緒的要素，在這階段如圖五所示，因為有可能會對分類模型的訓練產生影響，所以我們在這對文本作過濾，將文本區分成含中立的文本和不含中立的文本。本研究使用 Scikit-learn 工具來進行 SVM 極感極性分類，在進行分類前，必須將訓練資料和測試資料轉為所需的格式。如圖五，我們所處理完的文本經由向量轉換與主題特徵擷取的結合，透過 SVM 來進行情感極性的判斷。

四、實驗方法

本研究的資料集是使用 SemEval-2016[11]在 Task6 中所提供的測試與訓練資料，共 4063 筆。資料裡共有五個主題分別為：Atheism、Climate Change、Feminist Movement、Hillary Clinton、Legal. Abortion 等，內容為推特使用者針對各主題表達自身的評論或想法。

在機器學習上，文本的處理極為重要，在轉換為向量以前，如果一篇文本的雜訊過多，那麼就會影響到後續的輸入。因此在本部分的實驗，我們將以不同的文字處理方法來處理文本，並比較他們對於 SVM 分類的影響。下圖 4.2 分別有三種處理方法，Normal 為停用詞過濾的基本文本處理；Lemmatization 為詞型還原；Stemming 為詞幹提取。

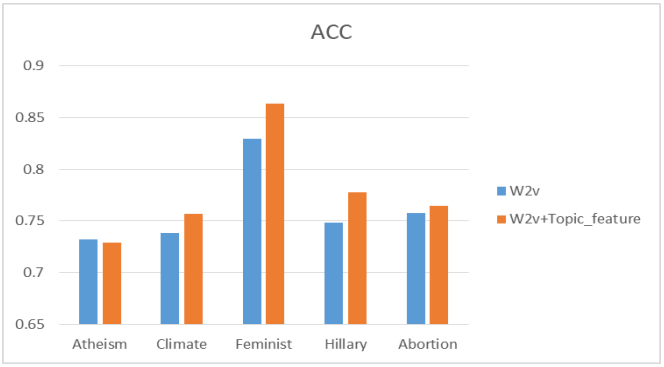


圖六、不同文本處理之分類準確率

由圖六得知，橫軸為各個目標主題，縱軸為準確率，經由詞型還原的資料，在分類的準

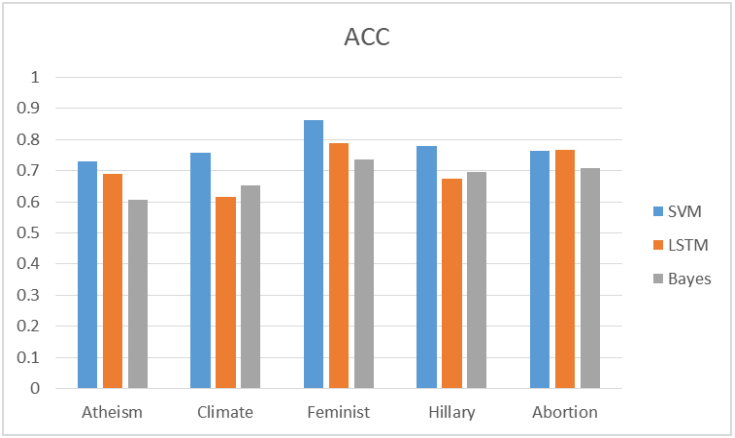
確率上大多有所提升，而一般的處理與詞幹提取的資料多比經由詞型還原的資料的準確率低。

從前面實驗我們知道詞型還原比其他文本處理的方法好，因此我們將使用詞型還原的文本處理方法來做主題特徵的實驗，本實驗將把主題特徵與 Word Embedding 的向量結合，進而與未添加主題特徵的一般字詞向量特徵做比較。



圖七、主題特徵實驗比較之準確率

由圖七得知，橫軸為各個目標主題，縱軸為準確率，我們可以發現 Feminist Movement 與 Hilary Clinton 在添加主題特徵後有明顯提升準確率，在 Legal. Abortion、Climate Change 略為增加；在 Atheism 上則沒明顯增加，原因可能在 Feminist 與 Hillary 的議題通常帶有高主觀性與高情緒極性的用字，而 Atheism 的議題則常出現一些情緒字眼較不明顯的用字。為了驗證分類模型的效果，所以我們以原先的資料集進行測試，並與樸素貝氏分類(Bayes)、長短期記憶神經網路分類模型(LSTM)這幾種方法來進行比較。



圖八、不同分類器之分類準確率

由圖八得知，橫軸為各個目標主題，縱軸為準確率，可以知道三種分類器中，SVM 最為優秀，而 LSTM 與 Bayes 分類器則較差。根據 Igarashi 等人[11]的實驗表示，深度學習的方法不見得會比較好。

五、結論

本論文提出使用 Word2Vec 字詞向量，結合主題特徵來進行立場檢測。在文本處理方面，使用三種文本處理方式，當中的 Lemmatization(詞型還原)於實驗中得證，在文本分析上的準確率較優。在主題特徵的方面，透過隱含狄利克雷分布(LDA)的方法來算出文本的主題分布能有效提升分類效果。在立場檢測方面，本論文透過支援向量機來訓練立場分類器，並且使用兩階段的方法來解決主觀性的問題，經實驗驗證最佳的 F-Measure 為主題目標女權運動的 83.36%。

參考文獻

[1] Statista, Number of social media users worldwide from 2010 to 2020,https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (Viewed on 2019/07/02)

[2] Jannati, R., Mahendra, R., Wardhana, C. W., & Adriani, M. (2018, November). Stance Classification Towards Political Figures on Blog Writing. In 2018 International Conference on Asian Language Processing (IALP) (pp. 96-101).

[3] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010, May). Predicting elections with twitter: What 140 characters reveal about political sentiment. In Fourth international AAAI conference on weblogs and social media.

[4] Sasaki, A., Mizuno, J., Okazaki, N., & Inui, K. (2016, October). Stance classification by recognizing related events about targets. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 582-587). IEEE.

[5] Tutek, M., Sekulic, I., Gombar, P., Paljak, I., Culinovic, F., Boltuzic, F., ... & Šnajder, J.

(2016, June). Takelab at semeval-2016 task 6: stance classification in tweets using a genetic algorithm based ensemble. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)(pp. 464-468).

[6] Bøhler, H., Asla, P., Marsi, E., & Sætre, R. (2016, June). Idi@ ntnu at semeval-2016 task 6: Detecting stance in tweets using shallow features and glove vectors for word representation. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 445-450).

[7] Wei, W., Zhang, X., Liu, X., Chen, W., & Wang, T. (2016, June). pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 384-388).

[8] Vijayaraghavan, P., Sysoev, I., Vosoughi, S., & Roy, D. (2016, June). DeepStance at SemEval-2016 Task 6: Detecting Stance in Tweets Using Character and Word-Level CNNs. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 413-419).

[9] Zarrella, G. and A.J.a.p.a. Marsh, Mitre at semeval-2016 task 6: Transfer learning for stance detection. 2016.

[10] Igarashi, Y., Komatsu, H., Kobayashi, S., Okazaki, N., & Inui, K. (2016, June). Tohoku at SemEval-2016 task 6: feature-based model versus convolutional neural network for stance detection. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 401-407).

[11] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016, June). Semeval-2016 task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 31-41).

[12] Blei, D.M., A.Y. Ng, and M.I.J.J.o.m.L.r. Jordan, Latent dirichlet allocation. 2003. 3(Jan): p. 993-1022.

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

[14] Google-News-dataset,https://code.google.com/archive/p/word2vec/ (Viewed on 2019/07/02)

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 102-111

Sequence to Sequence Convolutional Neural Network for Automatic Spelling Correction

Daniel Hládek, Matúš Pleva, Ján Staš,
Department of Electronics and Multimedia Communications
Technical University of Košice, Slovakia
daniel.hladek@tuke.sk,matus.pleva@tuke.sk,jan.stas@tuke.sk

Yuan-Fu Liao
Department of Electronic Engineering
National Taipei University of Technology
yfliao@mail.ntut.edu.tw

Abstract

The paper proposes a system that compensates most of the noise in a text in natural language caused by technical imperfection of the input device such as keyboard or scanner with optical character recognition, quick typing, or writer incompetence. Correcting the spelling errors in the text improves the performance of the following natural language processing. The incorrect sequence of characters is transcribed into another sequence of correct characters by a neural network with encoder-decoder architecture. Our approach to automatic spelling correction considers characters in an erroneous sentence as words of the source languages. The neural network searches for the best sequence of output characters for the given input. The proposed approach for spelling correction does not require any or minimal amount of training data. Instead, the error model is expressed by a simple component that distorts unannotated data and creates any necessary quantity of training examples for a neural network. The experimental results show that the presented approach significantly improves the distorted data (from 50% WER to 0.09% WER) with distortion lower than 1.5% WER.

Keywords: automatic spelling correction, sequence to sequence, encoder-decoder, deep learning.

1. Introduction

Written or scanned text is often not in the intended form. Writer or the input device often generate deviations that make it less understandable. The errors are usually not a problem in casual communication but make machine processing more complicated.

Removal of spelling errors helps with the following processing of the text in natural language. Automatic spelling correction (ASC) is an essential part of the processing of the documents with noisy data in natural language. ASC helps to recover the intended canonical form of the text and improves the quality of the input data for the following natural language processing (NLP) components. It supports processing of digitized documents, automated proofreaders, or information retrieval systems (e.g TREC-5 confusion track [1]). The main motivation for this work is an improvement of the training data for language model for speech recognition [2].

2. State of the art

The task of automatic error-correction is to generate the most likely correct word-for ms given a misspelled word-form [3].

Previous approaches to ASC, such as correcting spelling errors in the Chinese language [4] use classical statistical methods, such as the hidden Markov model, n-gram language model, log-linear regression, or forward-backward algorithm [5].

The usual form of mathematical formalism is a noisy channel proposed by Shannon [6]. Shannon defines the ASC of a possibly incorrect word s as finding the best correction candidate w_b from a list of possible correction candidates $w_i \in W$ (W is a valid word dictionary) with the best unnormalized probability [7]:

$$w_b = \max_{w_i \in C(s)} P(s|w_i P(w_i),$$

The error model $P(s|w_i)$ estimates the probability of unknown string s instead of real word w_i. The error model characterizes the spelling correction problem.

The context model $P(w_i)$ calculated the probability of the correction candidate according to the surrounding words. A finite-state based system, such as Hunspell[1] proposes a list of correction candidates, and a language model helps to choose the best spelling correction candidate.

The task of spelling correction is similar to machine translation (MT). An ASC trans lates input sentence containing spelling errors into another sentence in a "correct" lang uage. Machine translation converts a sequence of words in the source language into another sequence of words in the target language. Formally, MT is the search for the best target seq uence T given source sequence S using model P [8]:

$$T_b = \max_T P(T|S)$$

There are a couple of approaches that used statistical MT for ASC before, such as machine translation spelling for historical texts [9]. [10] attempts to character-level spelling correction.

3. Sequence to sequence spelling correction

Statistical machine translation uses classical methods, such as hidden Markov models, n-gram language models, and sentence alignment [2]. SMT systems have weaknesses that prevent to reach better results. The statistical approaches can calculate only with relatively short contexts (three items in the input sequence maximum).

Neural networks with encoder/decoder architecture brought significant improvement in the performance of SMT. Current deep neural networks [11] can consider a much bro ader context of words or characters. This ability allows us to use an architecture that is based only on neural networks and considers only characters. A neural model can be used to score any given pair of input and output sequences [12].

[1] http://hunspell.github.io/

[2] Moses toolkit

Sequence to sequence neural networks architecture transforms a sequence of symbols from the source language to another sequence of symbols in the target language. Sequences can have different lengths. One symbol is encoded into an n-dimensional binary vector with one dimension for each possible character. The embedding layer reduces the dimension of the input vector. The transformed input matrix has dimension equal to the embedding dimension and size of the sequence. The neural network that transcribes one sequence of symbols into another consists of the encoder and the decoder.

"The encoder maps a variable-length source sentence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length sentence. The two ne tworks are trained jointly to maximize the conditional probability of the target sequen ce given source sequence.". [12]

Knowing the probability of the next symbol enables the decoder to sample probable sequences of symbols. "Sequence to sequence" systems usually use recurrent neural networks (RNN) or convolutional neural networks (CNN) for encoding and decoding.

"The dominant approach to date encodes the input sequence with a series of bi-directional recurrent neural networks (RNN) and generates a variable-length output with another set of decoder RNNs, both of which interface via a soft-attention mechanism." [13]

Recurrent neural networks perform well with tasks with variable-length input and output. Common types of RNN are gated recurrent units (GRU) [14] and long short-term memory (LSTM) [15].

4. The proposed convolutional network architecture

The RNN always maintain a hidden state and updates it with each new item in the input sequence. Compared to RNN, the current state in the input sequence of a convolutional network does not depend on the previous, which makes the computation easier. The processor can compute convolution for the whole sequence at once.

"Convolutions create representations for fixed-size contexts; however, the effective context size of the network can easily be made larger by stacking several layers on top of each other." [13]

The approach uses a convolutional sequence-to-sequence architecture by [13]. The convolutional architecture uses gated linear units (GLU) [16] with residual connections [17]. "The attention mechanism looks at the input sequence and decides at each step which parts are important." The attention mechanism "writes down" quintessential keywords from the sentence. The attention-mechanism considers several other inputs at the same time and decides which ones are important by attributing different weights to those inputs.

The convolutional architecture was selected because recent results [13] show that they offer superior or comparable performance and higher speed of learning when compared to the more-established recurrent networks.

4. Data preparation

Neural networks are sensitive to the amount of training data. Obtaining reasonable precision requires the sufficient size of the training set. Preparation of the data for training of the neural network is difficult, timely, and expensive. Our approach overcomes the problem of data sparsity by rule-based error model that utilizes any unannotated data in the target language and prepares an artificial training set.

A sequence of edit operations describes a spelling error. Usually, the error model considers insertion, deletion, and substitution of characters. A statistical error model is estimated from training data that contain the original and the erroneous strings. An artificial error function randomly modifies some characters in the dataset and creates a distorted string. Example of the training set is in the Table 1.

The training of the neural network uses the distorted string as input and the original string as the output. Training of the network estimates the reverse function and the network can guess the intended form of a distorted string. The error function can generate any amount of training data from a text in natural language.

Table 1Example of the training data

Distorted input	Correct input
faktom vshak oďtava že stavy zamesnancov	faktom však ostáva že stavy zamestnancov

5. Experiments

The proposed neural network needs a sufficiently large text in a natural language for training. We have composed a set of newspaper articles in the Slovak language.

One training sample for the neural network consists of three words. The "clean" text forms the target part of one sample. Table 2 summarizes the size of the text database.

The error model distorts characters in the sample and creates the source part. The distorted and original sequence form a training pair. The neural network learns a function that is inverse to the one that generated the training data.

Table 2 Experimental dataset size

Set	Samples	Words	Characters
train	12 000 000	36 000 000	206 711 896
test	50 000	150 000	873 358

The error model uses the following rules and probabilities:

- Insertion of arbitrary character 0.02
- Deletion of arbitrary character 0.02
- Replacement of arbitrary character 0.08
- Keeping the character 0.9

A forward-backward algorithm by Ristad and Yanilos [18] can estimate parameters of the error

model for a set of training examples, which is left for the further research. The ASC system uses Fairseq toolkit [19]. The table 3 summarizes the parameters of the neural network (named fconv_iwslt_de_en in Fairseq toolkit).

Table 3 Neural network architecture

Dropout	0.1
Encoder Embedding Dimension	256
Encoder Layers	256,3 * 4
Decoder Embed Dimension	256
Decoder Layers	256,3 * 3
Decoder Out Embed Dimension	256
Decoder Attention	True

Word error rate is a usual form of evaluation of spelling correction models. Its advantage is that the size of the target and source sequence does not have to be the same. The metric first aligns the sequences with the hypothesis and with the golden truth. The WER is defined as a ratio of the counts of the inserted, deleted, and replaced words:

$$WER = \frac{C_i + C_d + C_s}{C}$$

C_i is number of insertions, C_d number of deletions, C_s is number of substitutions.

Character error rate (CER) is a similar metric but considers characters instead words. Sentence error rate (SER) is ratio of incorrect samples to all samples in the testing set.

The first experiment measures CER, WER, SER of correcting randomly distorted testing text omitted from the training. The Table 4 displays performance of the system after selected iterations (1,5,10,15) of the training of the neural network. The first row (0 - no correction) shows measure of preliminary distortion of the testing set by the error model without any processing. Figure 1 displays the complete learning curve in CER for each training iteration.

Performance of the system is improved only slightly after tenth round of training.

Table 4 Performance of the proposed system

Iteration	CER	WER	SER
0 (no correctoin)	0.1096	0.5173	0.8463
1	0.0386	0.1447	0.3396
5	0.0307	0.1108	0.2677
10	0.0279	0.0998	0.2443
15	0.0273	0.0971	0.2387

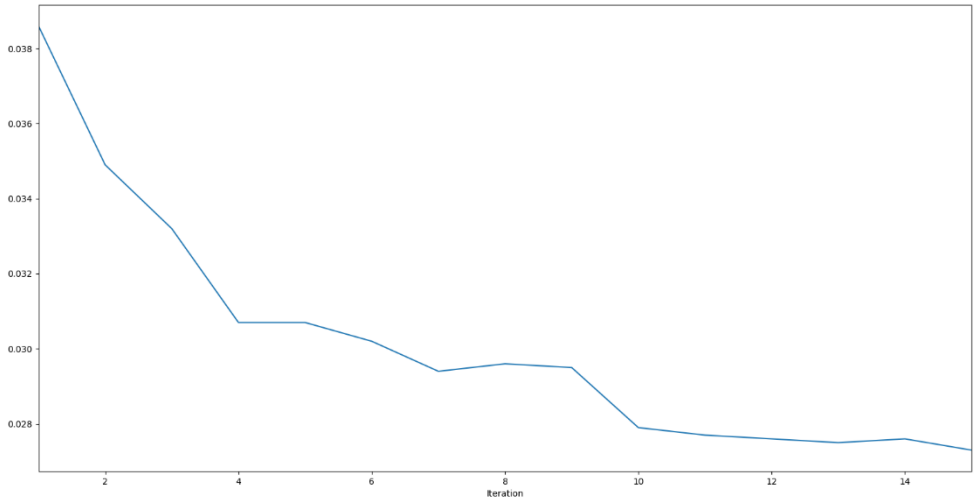


Figure 1 CER Learning Curve

The second experiment measures how the trained neural network damages useful data. Input of the ASC system is a clean text. The Table 4 shows how the neural network distorts the clean data. The distortion of the clean data is very low (0,0022 CER) and decreases with number of training iterations. Distortion CER is marked in the Table 5 for each training round. It shows clear correlation with the learning curve in Figure 1.

Table 5 Distortion on the clean data

Iteration	CER	WER	SER
1	0.00342	0.01594	0.044

5	0.00265	0.01233	0.034
10	0.00254	0.01202	0.033
15	0.00224	0.01029	0.028

6. Conclusion

The experiments confirm that the proposed approach can remove most of the noise from a text in natural language. An expert can design the artificial error model according to the typical error patterns. It is possible to use statistical estimation with relatively small training data, e.g. a letter confusion matrix ([5], [18]). Processing of the clean data has very low distortion and the proposed neural network can be used without damaging the clean data.

Acknowledgements

This work was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under research project VEGA 1/0511/17 of the Slovak Scientific Grant Agency, Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731 and Taiwan Ministry of Science and Technology MOST-SRDA contract no. 108-2911-I-027-501,107-2911-I-027-501,107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067.

References

[1] P. B. Kantor and E. M. Voorhees, “The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text,” *Inf. Retr. Boston.*, vol. 2, no. 2, pp. 165–76, 2000.

[2] M. Rusko *et al.*, *Advances in the Slovak Judicial domain dictation system*, vol. 9561. 2016.

[3] T. A. Pirinen and K. Lindén, “State-of-the-art in weighted finite-state spell-checking,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8404 LNCS, no. PART 2, pp. 519–532.

[4] Y. Zhang, P. He, W. Xiang, and M. Li, “Discriminative Reranking for Spelling Correction,” *Proc. 20th Pacific Asia Conf. Lang. Inf. Comput.*, pp. 64–71, 2007.

[5] D. Hládek, J. Staš, S. Ondáš, J. Juhár, and L. Kovács, “Learning string distance with smoothing for OCR spelling correction,” *Multimed. Tools Appl.*, pp. 1–19, 2016.

[6] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–56, Oct. 1948.

[7] E. Brill and R. C. Moore, “An improved error model for noisy channel spelling correction,” in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics ACL 00*, 2000, pp. 286–93.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” *Adv. Neural Inf. Process. Syst. 27 (NIPS 2014)*, pp. 3104–3112, 2014.

[9] P. Mitankin, S. Gerdjikov, and S. Mihov, “An Approach to Unsupervised Historical Text Normalisation,” in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage - DATeCH ’14*, 2014, pp. 29–34.

[10] R. Mihalcea, “Diacritics restoration: learning from letters versus learning from words,” in *CICLing 2002*, vol. 2276, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2002, pp. 96–113.

[11] A. C. Kinaci, “Spelling Correction Using Recurrent Neural Networks and Character Level N-gram,” in *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, 2019, pp. 1–4.

[12] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[13] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional Sequence to Sequence Learning,” May 2017.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” in *NIPS 2014 Deep Learning and Representation Learning Workshop*, 2014.

[15] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[16] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language Modeling with Gated Convolutional Networks,” Dec. 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[18] E. S. Ristad and P. N. Yianilos, “Learning string-edit distance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 5, pp. 522–32, May 1998.

[19] M. Ott *et al.*, “fairseq: A Fast, Extensible Toolkit for Sequence Modeling,” Apr. 2019.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 112-121

基於深度學習之簡答題問答系統初步探討

A Preliminary Study on Deep Learning-based Short Answer Question Answering System

林鈺宸 Yu-Chen Lin, 廖元甫 Yuan-Fu Liao
國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
t106368024@ntut.edu.tw, yfliao@ntut.edu.tw

Matúš Pleva, Daniel Hládek
Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia
matus.pleva@tuke.sk, daniel.hladek@tuke.sk

摘要

一般問答系統常是以從題目原文中尋找最可能的文字段落的方式來生成答案，但是最近開始有試圖以編碼器(encoder)先抽取題目與問句的隱含意義，再以解碼器(decoder)重新生成回應語句的趨勢。因為它能不受原始文章使用的文字的限制，甚至可以用完全不同的說法來回答。因此在本論文中提出了一個基於 BERT 與 Transformer 的“編碼-解碼”模型，嘗試實現此問答系統。實驗方面使用台達電研究所的公開數據集(Delta Reading Comprehension Dataset, DRCD[4])和『科技大擂臺，與 AI 對話』(Formosa Grand Challenge)的比賽資料，來訓練模型。由實驗結果發現，我們的系統可以由學習產生答案，並達到 ACC=66.22%，F1=41.16%， EM=18.41%， BLEU=3.98%的效能，嘗試實現可自由產生回應語句的問答系統。

關鍵詞：自然語言處理、問題回答、序列到序列模型, NLP, QA, Seq2Seq

一、緒論

在自然語言處理中的問題回答系統領域有許多不同類型的文本理解競賽，如科技部舉辦的『科技大擂臺，與 AI 對話』(Formosa Grand Challenge) 競賽、史丹佛大學發起的挑戰(Stanford Question Answering Dataset, SQuAD[3])以及分別來自於 CMU、Stanford 和 Mila 的學生推出的新型問答挑戰 HotpotQA[5]等等。

因為問題回答系統領域在國外是一個很熱門的挑戰，所以延伸許多不同的閱讀理解數據集，如史丹佛的 SQuAD(表 1)，它不同於一般的閱讀理解數據集，在它的答案不是僅僅一個詞或是物體，而有可能是一段句子，使得答案更難以預測，但是此數據集的問題大多可以使用關鍵字匹配的方式來進行搜索回答；而 HotpotQA(表 2)的問題被設計成

需要以多步推理的方式來回答，所以無法輕易地以關鍵字匹配的方式來解答。另外，此數據集沒有預設任何的知識圖譜，因此問題內容具有多樣性，使得它更有難度。

由於受到了國外的刺激，我國政府機關也開始積極投入這個領域，所以有了 Formosa Grand Challenge 的比賽，此數據庫它需要有對問題的推理能力還有閱讀的理解能力，且不同於 SQuAD 和 HotpotQA 的閱讀理解數據集，它無法輕易以搜索答案在文章中的起始與結束位置得知正確答案，比賽的簡答題的部分需要生成一段完整的句子且需要契合參考答案。

表 1、SQuAD 閱讀理解數據集範例

C	Victoria (abbreviated as Vic) is a state in the south-east of Australia. Victoria is Australia's most densely populated state and its second-most populous state overall. Most of its population is concentrated in the area surrounding Port Phillip Bay, which includes the metropolitan area of its capital and largest city, Melbourne, which is Australia's second-largest city. Geographically the smallest state on the Australian mainland, Victoria is bordered by Bass Strait and Tasmania to the south,[note 1] New South Wales to the north, the Tasman Sea to the east, and South Australia to the west.
Q	Where in Australia is Victoria located?
A1	south-east
A2	the south-east of Australia

表 2、HotpotQA 閱讀理解數據集範例

C1	Gorgeous George (album) Gorgeous George is the third solo studio album by Scottish musician Edwyn Collins. The album was recorded at New River in London, with Collins acting as the producer.
⋮	⋮
C10	Jimmie Ross Jimmie Ross is an American rock guitarist and vocalist who is best known for being a member of Pittsburgh band the Jaggerz, known for their 1970 hit. During the band's initial existence of 1965-1976, the bassist shared the duties of lead vocalist with guitarist Donnie Iris. By the time the Jaggerz regrouped in 1989, Iris was well into his solo career, and Ross became the sole lead vocalist and remained bassist.", " He continues to hold both positions today.
Q	Which musician, Edwyn Collins or Jimmie Ross, played the bass guitar?
A	Jimmie Ross

表 3、Formosa Grand Challenge 閱讀理解數據集範例(簡答題)

C	請聽這段話，然後回答問題： 科學研究結果顯示，人類種植可可樹的起源可以追溯至大約 3600 年前。研究人員抽絲剝繭，分析了 200 棵可可樹的基因組之後認定，最早種植可可樹的是現今厄瓜多境內的古瑪雅人。科學家也指出，早期的可可豆風味相當不同。人工種植可可樹後，人類選擇自己偏好的性狀，然後不斷繁殖，並持續改良植物的大小和風味等特徵。在培育一代又一代可可樹的過程中，可可豆

	的風味逐漸改變，苦味變濃，帶給人興奮感的可可鹼含量也逐漸增加。但在此同時，可可樹對病蟲害的抵抗力也隨之下降，使得它愈來愈珍貴。
Q	為什麼可可樹的價值逐漸增高?
A1	可可樹易得病蟲害，種植成本增加，數量減少。
A2	因為可可豆的風味逐漸改變，帶給人興奮感的可可鹼含量逐漸增加，但同時可可樹對病蟲害的抵抗力也下降了。

為了解決 SQuAD 的類型的題目，目前的方式有 QANet[1]和 BERT[2]與 Transformer[9]組成的 Seq2Seq 模型，它們基於開放性領域的知識庫上建立一個問題回答系統需要透過既有的知識庫中來進行特徵提取，之後將向量將抽取特徵和相同意義的資訊建立在一起進而會答問題。而動態融合圖形網路(Dynamically Fused Graph Network for Multi-hop Reasoning, DFGN[17])是為了解決 HotpotQA 而發展出來的架構，它能從多個片段文字段落動態建構的特徵進行探索，並從特定的文章中逐步找出的相關資訊進行推理來找尋答案。

在 Formosa Grand Challenge 的比賽中我們測試使用 QANet 和 DFGN 的架構配合 DRCD[4]所訓練出來的模型進行問題的解答但其效果都有限，主要是因為這些過去的任務主要還是以搜索答案的位置為主，因此如果一個正確答案不在文章中系統便無法回答。雖然在比賽中還是會遇到單純的分析問題，但在每次比賽過程中會慢慢發現這種分析問題越來越少，且在決賽時會有簡答題的題型，為此我們認為需要一套可以了解文章內容並且可以自然答案生成的系統。

在 Formosa Grand Challenge 的比賽中需要了解文章內容進而回答問題，為了配合後續簡答題的回答策略，所以我們在選擇題的部分先以回答簡答的方式生成答案，再配合提供的選項來進行匹配。我們嘗試以編碼器先抽取題目與問句的隱含意義，再以解碼器重新生成回應語句，而不是從題目原文中尋找最可能的文字段落的方式來生成答案最後在與答案做關聯性的匹配。這不同於過去的知識庫回答系統，他們比起回答一個擁有正確答案且完整句子，更注重於分析問題和搜尋相關答案。我們實驗了(如圖 1)基於注意力機制的自然答案生成模型，在系統裡使用多組閱讀測驗問答式的資料進行訓練，並使用序列到序列(Seq2Seq)的學習框架來完成任務。

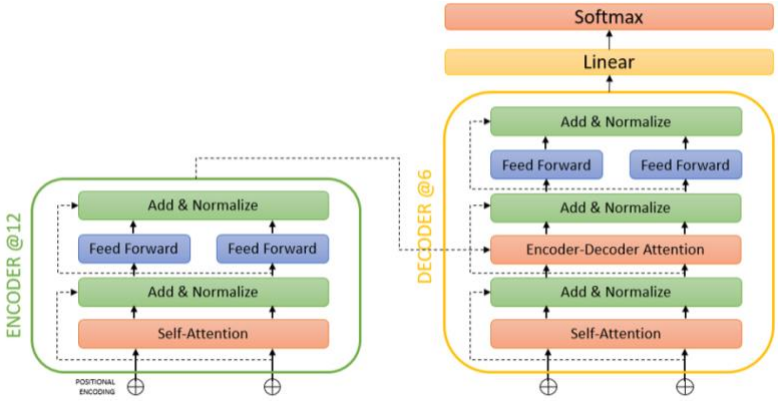


圖 1、模型架構示意圖

二、相關研究

Seq2Seq 模型架構最初提出是為了機器翻譯[12]，它解決 RNN 不定長度的問題並能有效地建立一個基於輸入序列用來預測未知輸出序列，在這之後此架構在機器翻譯、寫作、文本摘要和人機對話等主題有更多的發揮。最初所提出的 Seq2Seq[18]模型想要在一個完整的輸入序列中找出語意來做向量編碼是一件很困難的事情，特別是對於較長的輸入，因此提出了注意力機制[19]的 Seq2Seq 模型架構。由於我們用於訓練所使用的資料經常以較長的輸入為主且需要尋找出關鍵的語意進而回答問題，所以我們傾向使用擁有注意力機制的 Seq2Seq 模型架構來培訓我們的模型。

過去常用於問題回答的方法從 RNN 架構，到近幾年的 QANet 和 DFGN 等等。QANet 它拋棄了 RNN 的作法並設計了一個模型架構，由多個編碼區塊建立而成的，在每個區塊中都包含 Multi-Convolution-Layer、Self-Attention Layer[16]和 Feedforward Layer。其知識庫基本是以詞為單位，使用 200 和 300 不同維度的詞向量(Word2vec[6-8])為基礎，再透過編碼器來學習文章與問題之間的關聯性，由於使用 Self-Attention 因此對全局的資訊做有效的處理來得到較好的結果。

為了改善知識庫的部分而提出新穎的預訓練方法 BERT 並配合 Transformer 組成一個是 seq2seq 模型，BERT 的部分使用 Transformer 的架構來訓練單字級別的知識基礎，由於使用 Transformer 所以它相較於深層遞迴類神經網路(RNN)更能有效捕捉較長句子的資訊，使它在問答解題相較於 QANet 有較好的結果。

三、方法

序列到序列模型架構

過去傳統序列到序列(Seq2Seq)模型架構應用在一連串有相關的連續數據中，如語音數據、影像數據等，因為它具有輸入以及輸出長度不固定的特性以及元素之間的順序關係。在所有的問答題中它們的文章、問題以及答案的長度並不是等長的，而在輸出答案的前後順序也會影響著我們得到結果，因此我們採用此架構為我們的訓練框架。

Seq2Seq 分成兩個部分，編碼器和解碼器。編碼器我們使用 Google 的 BERT 預訓練模型，首先將文章與問題串連後透過 Tokenizer 轉換成輸入序列後再輸入到解碼器得到編碼資訊。在解碼器的部分我們使用 Transformer 的架構，總共有 6 層，它將從編碼器所得到的編碼資訊配合上一個輸出來產生解碼器的輸出資訊。最後，將輸出資訊連接到線性的全連接層透過 softmax 輸出成語字典同樣的維度。

(一) BERT 神經網路模型

BERT 重點在於它能貫穿每層的背景關係來預訓練深度的雙向表示，它與其他模型不同的地方在於它多個一個機制稱為 Masked LM (MLM)，能隨機地屏蔽部分的輸入 token 只計算被遮蓋掉的 token，目的是根據被遮掉標記的背景關係來預測原始對應的 token，另外使用雙向 Transformer，能讓 MLM 標記可以學習到前後的背景關係。

在實驗中我們先將文字以既有的辭典映射成數字的形式來表示，BERT 模型輸入的

地方有兩個輸入，第一個部分是輸入文字序列(Token)，第二個是輸入分割序列(Segment)。在 Token 的部分給出一個中文單字級別的序列$[101, q_0, \cdots, q_L, 102, c_0, \cdots, c_L, 102]$，它包含了問題以及文本的文字編碼，問題前以及文本之間和最後都以特殊字元隔開，問題q和文本c的總長度為L。另外，在 Segment 的部分給入一個由 0 和 1 組的序列，它是根據文章和問題來做分割的，在這邊我們將問題依照 Token 的數量給予相同的 0，而文章給予 1，因此最後會形成一個序列$[0, \cdots, 1, 1, \cdots, 1]$用來區分文章與問題。最後，在將得到的文字序列和分割序列透過 Word Embedding 輸入進 BERT Model，得到該序列的編碼資訊。

(二) Transformer 模型架構

解碼器是使用 Transformer 架構組成的，將訓練答案以跟 BERT 神經網路模型所使用到的辭典映射成數字的形式$[101, a_0, \cdots, a_L, 102]$，之後轉換成訓練對$[101, a_0, \cdots, a_L]$和$[a_0, \cdots, a_L, 102]$，在訓練過程中將從上一個解碼器的輸出值作為輸入，來訓練下一個輸出值的機率分布。應用於解碼器的 Transformer 有一個不同，它多增加一個 Self-Attention 單元，由於序列到序列的模型在訓練時的解碼器需要參考上一個輸出的結果來當作下一個的解碼器的輸入，而編碼器又需要根據 BERT 所分享的資訊來進行運算，因此這邊會先將上一個的輸出會透過 Self-Attention 計算後再跟 BERT 的輸出合併做運算，在來導入線性的全連接層透過 softmax 輸出成語字典同樣的維度，最後將輸出的值再透過字典轉回文字供使用者閱讀。在訓練過程中我們是以字對字的方式去學習，從 input 的每一個字去學習 output 相對應的字的出現機率。

四、實驗結果

(一) 數據集

我們使用兩個數據集來評估我們的方法，第一個是台達電研究院(DRCD)的公開的數據集，另一個是和『科技大擂臺，與 AI 對話』(Formosa Grand Challenge)的比賽數據集(如表 4)。

DRCD 是一個中文閱讀理解數據集，它從 2,108 篇維基百科的文章整理出 30,000 多個問題，人類表現在此數據集的 F1 成績為 93.30%以及 Exact Match 的成績為 80.43%。Formosa Grand Challenge 它是由中華民國政府部門舉辦的一個大型競賽的中文數據集，包含了閱讀理解以及單純的問題回答(問題解答不包含在文本之中)，它從網路新聞、古文以及小說整理出來的 14,000 多個問題。另外，此數據集有 25 題是問答題，由多位專家來出題，且經過討論後給定多個參考解答，並由評審委員修改。它總共分成 6 次初賽、加賽、複賽和決賽，因此我們將初賽和複賽歸納至訓練集並隨機抽出 1,000 題來做驗證並將加賽和決賽的簡答題題目歸類到測試集，在這邊我們捨棄掉大部分決賽的資料是因為決賽中大量的問題需要從答案來判斷答案是否正確。

表 4、數據集統計

資料集	訓練	驗證	測試

DRCD	26,932	3,524	3,485
Formosa Grand Challenge	11,551	1,000	1,025

由於 Formosa Grand Challenge 比賽形式以選擇題為主，因此我們將正確的選項內容當作答案文字內容來進行訓練，另外由於它的資料庫有多餘的換行以及特殊符號，所以我們也將它刪除掉，以保持續訓練文本的統一性。

(二) 評估指標

在進行多類別的標籤分類中採用分類交叉熵(Categorical Cross-Entropy)來幫助我們了解訓練上的差異，其方程式如下所示：

$$loss = -\sum_{i=1}^{n} \hat{y}_{i1} log y_{i1} + \hat{y}_{i2} log y_{i2} + \cdots + \hat{y}_{im} log y_{im}$$

其中，n是樣本數量，m是分類數量，由於此函數是一個多輸出的 loss 函數，所以計算出來的值也是多個的。

另外我們使用了幾個指標來幫我們評估我們的方法：第一個是正確度，主要是為了知道它保存了多少資訊在裡面，其方程式如下所示：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

其中 True Positive(TP)是將正確的預測為正確，True Negative(TN) 是將錯誤的預測為錯誤，False Positive(FP) 是將錯誤的預測為正確最後 False Negative(FN) 是將正確的預測為錯誤的。

第二個是 F1 值，統計學中用來衡量分類模型精確度的一種指標。F1 值就是精確率(precision)和召回率(recall)的調和均值。精確率它計算所有正確被檢出的結果(TP)占實際上被檢索到的(TP+FP)比例，其方程式如下所示：

$$P = \frac{TP}{TP + FP}$$

而召回率它是計算所有正確被檢出的結果(TP)占所有被應該檢索到的(TP+FN)比例。

第三個是 Bilingual Evaluation understudy(BLEU)，評分關鍵在於如何定義生成答案與參考答案之間的相似度，首先須計算詞在句中的匹配數，其方程式如下所示：

$$Count_{clip}(word) = min\{Count(word), MaxRefCount(n-gram)\}$$

其中，$Count(word)$表示詞在生成文中的出現次數，$MaxRefCount(word)$是該參考答案的最大次數。出現詞的精確度P_n定義為：

$$P_n = \frac{\sum_{C \in candidates} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in candidates} \sum_{n-gram \in C} Count(n-gram)}$$

透過上述結果來計算 BLEU，其評分方程式如下：

$$BLEU = BP * exp\left[\left(\sum_{n=1}^{N} W_n log p_n\right)\right]$$

其中BP是懲罰值，其中W_n表示詞出現的權重：

$$\text{BP} = \{_{EXP(1-r/c)}^{1} \; _{if\ c \le r}^{if\ c > r}\}$$

最後是 Exact Match(EM)，它能讓我們觀察生成出的答案與標準答案是否一致。

(三) 實驗設置

訓練過程中，我們將兩個資料庫的訓練集以及驗證集各自合併一起訓練，一開始先設定文本和問題的總長度限制在 384，它可以完全覆蓋數據集中 95%的資料，另外將答案的總長度限制在 30，它能完全覆蓋 98%的答案數據。使用的辭典包含中英文以及各式標籤總共有 21128 個字。

模型架構使用序列到序列模型，在編碼器使用 BERT 中文預訓練模型，它是由 12 層雙向 transformer 組成，隱藏層擁有 768 個節點，Multi-head self-attention 的 heads 為 12；解碼器則是有 6 層雙向 transformer 組成，且每層都共享了編碼器的輸出。

(四) 結果與討論

從損失值(如圖 3)和正確率(如圖 4)可以證明我們的模型架構是可以有效的學習，而正確率越高表示我們的模型在學習過程中透過由 BERT 給予的知識以及上一次的輸出能生成出包含正確的資訊越好。我們所提出的方法用於兩個數據集的測試集，在 DRCD 的正確率擁有 66%，F1 評分有 41.16%，EM 評分有 18%；Formosa Grand Challenge 的正確率有 58%，F1 評分有 37.46%，EM 評分有 16%。如表 5。

表 5、數據集指標

資料庫	Acc	F1	EM	BLEU
DRCD	66.22	41.16	18.41	3.98
Formosa Grand Challenge	58.10	37.46	16.35	3.25

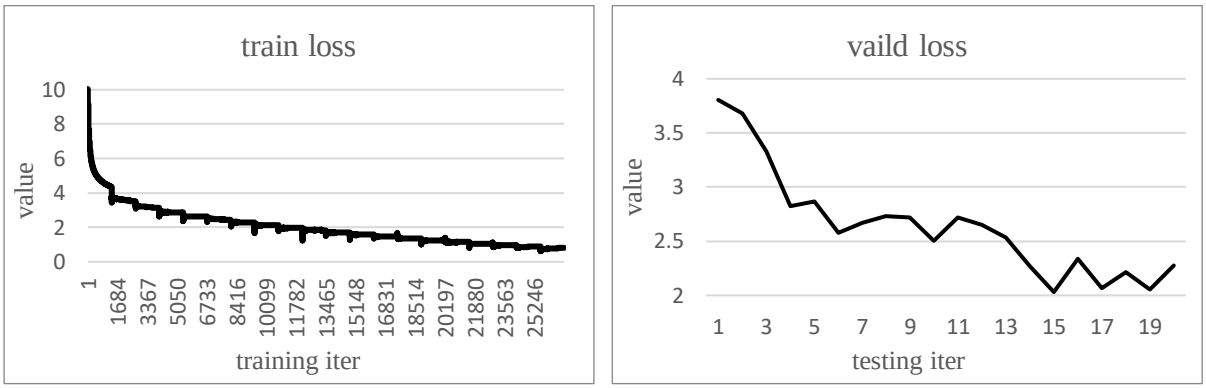


圖 3、訓練/測試的損失值分布

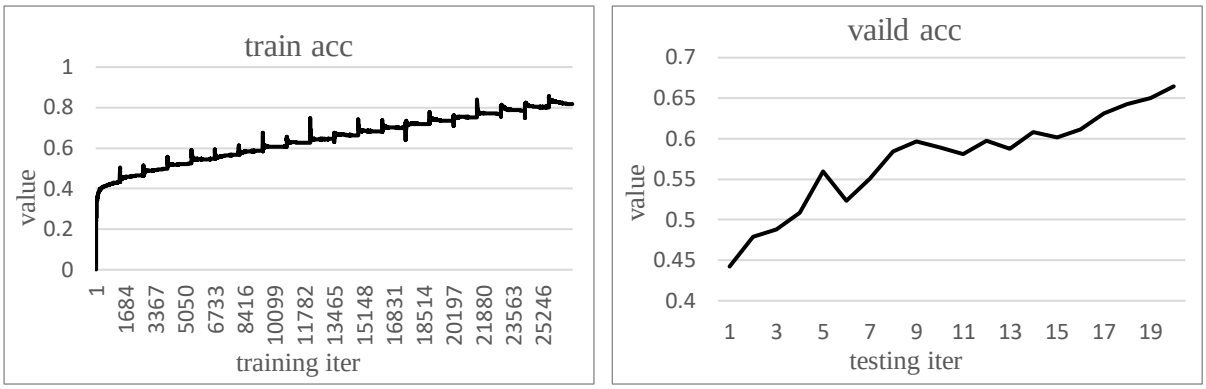


圖 4、訓練/測試的正確率分布

以下我們提出幾個案例來進行討論：從範例觀察我們發現從表 6-8 可以看到我們的模型生成流暢的句子且相似的答案，並具有正確的知識。然而答案還是存在一些問題，例如無法找出問題與文章之間正確的關鍵點，所以回答出的錯誤的答案或產生連續重複的詞彙、回答不完整，以及生成出不流暢的句子。

從上述問題中我們可以發現其實 BERT 是從一個輸入序列(文章)中找尋與問題最具有關聯性的訊息，它還不能算真正擁有解讀文章意義的模型。產生連續重複的詞彙的問題是因為序列到序列模型過度依賴上一次的輸出，因此如果出現一個重複的單字就會出現無限循環，導致產生出的答案出現重複詞彙的現象。最後生成出不流暢的句子，我們認為是訓練答案的資料的類型種類不夠多，由於我們的訓練的答案主要還是以簡短的幾個字詞或專有名詞來輸出，所以在遇到需要回答較長類型的答案時便無法生成出來。

表 6、輸出結果範例一

C1	請聽這段話，然後回答問題： 台灣宇博(Uber)公司遭公路總局罰款 3000 多萬一事，台北高等行政法院今天做出判決，判 Uber 勝訴，無須繳交罰款。根據調查，公路總局發現，Uber 未經核准，於民國 105 年間透過網路招募司機，分別在台北市、新北市等地，藉由 Uber APP 程式平台，指揮調度 233 輛車營業載送客人。載客完成後，收取報酬，違反汽車運輸業管理規則，因此處以總計逾 3000 萬元罰款。台北高等行政法院法官調查，計程車客運業的主管機關並非公路總局，而是市政府。因此撤銷處分，判 Uber 勝訴免罰。全案可上訴。
Q1	台灣宇博(Uber)公司分別在台北市、新北市兩地違法行事，應該由誰開出罰單？
A1	台北市政府、新北市政府
Our	以上皆可

表 7、輸出結果範例二

C2	請聽這段話，然後回答問題： 台東聖母醫院是個小型醫院，除了醫治病人，還肩負起照顧台東弱勢原住民的重責大任。它的廚房每天提供午餐給獨居老人和兒童，讓他們不至於挨餓。 但七月份的強烈颱風重創台東。聖母醫院的倉庫被淹沒，導致白米泡水，連供應午餐的廚房設備也嚴重毀損。醫院估計，重建經費逼近千萬。 聖母醫院表示，他們每天用的食材一向靠志工栽種的有機蔬果供應；此外，醫院向政府承

	租農地，一年兩穫的稻米還可以小包裝透過網路販售。扣除成本之後，販賣的盈餘就是維持農場和廚房營運的資金。 然而風災過後，收成全無，設備毀損，復原之路十分漫長。聖母醫院因而呼籲各界人士慷慨解囊，協助早日重建，讓他們的廚房不至於斷炊。
Q2	聖母醫院平日是靠什麼樣的收入來維持廚房的營運？
A2	靠銷售自己種植的稻米所得盈餘來維持。
Our	自給農業

表 8、輸出結果範例三

C3	A：我們這週末要帶小朋友們去動物園。 B：他們喜歡什麼動物？ A：他們都喜歡大熊貓、無尾熊。哥哥特別想要看老虎和蛇，但妹妹會害怕。妹妹喜歡小兔子和鴨子，可是哥哥沒興趣。兩個吵吵鬧鬧的，真是讓人傷腦筋。
Q3	哪一種動物可以讓兩位小朋友一起看？
A3	無尾熊
Our	大熊貓

五、結論

本論文提出了一個基於 BERT 與 Transformer 的“編碼-解碼”問答模型,嘗試實現可以回答任意簡答題的問答系統,尤其是可以回答文章文中沒有直接提到的答案。其能從文章中搜尋跟問題相關的資訊,並通過自我關注的機制,提高自然答案生成系統的性能。經由實驗結果驗證,我們的系統確實可以由學習產生答案,並達到 ACC=66.22%, F1=41.16%，EM=18.41%, BLEU=3.98%的效能,初步實現了可自由產生回應語句的簡答題問答系統。未來,將往兩個方向進行研究。包括(1)如何讓編碼器能擷取出更豐富的語意特徵,與(2)如何讓解碼器能對問題的理解能有更深一步的能力。

Acknowledgements

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and finally by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 107-2911-I-027-501, 108-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067.

參考文獻

[1]. Yu, Adams Wei, et al. "Qanet: Combining local convolution with global self-attention for reading comprehension." arXiv preprint arXiv:1804.09541 (2018).

[2]. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding."

arXiv preprint arXiv:1810.04805 (2018).

[3]. Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).

[4]. Shao, Chih Chieh, et al. "Drcd: a chinese machine reading comprehension dataset." arXiv preprint arXiv:1806.00920 (2018).

[5]. Yang, Zhilin, et al. "Hotpotqa: A dataset for diverse, explainable multi-hop question answering." arXiv preprint arXiv:1809.09600 (2018).

[6]. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[7]. Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. 2014.

[8]. Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.

[9]. Shao, Taihua, et al. "Transformer-Based Neural Network for Answer Selection in Question Answering." IEEE Access 7 (2019): 26146-26156.

[10]. Lin, Yuhua, and Haiying Shen. "SmartQ: A question and answer system for supplying high-quality and trustworthy answers." IEEE Transactions on Big Data 4.4 (2017): 600-613.

[11]. Karimi, Elaheh, Babak Majidi, and Mohammad Taghi Manzuri. "Relevant Question Answering in Community Based Networks Using Deep LSTM Neural Networks." 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS). IEEE, 2019.

[12]. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[13]. Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[14]. Tu, Zhaopeng, et al. "Modeling coverage for neural machine translation." arXiv preprint arXiv:1601.04811 (2016).

[15]. Tan, Chuanqi, et al. "Context-aware answer sentence selection with hierarchical gated recurrent neural networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.3 (2017): 540-549.

[16]. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[17]. Xiao, Yunxuan, et al. "Dynamically Fused Graph Network for Multi-hop Reasoning." arXiv preprint arXiv:1905.06933 (2019).

[18]. Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).

[19]. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 122-136

基於遞迴類神經網路之麥克風嘯叫抑制系統

Recurrent Neural Network-based Microphone Howling Suppression

林政陽　Cheng-Yang Lin, 廖元甫　Yuan-Fu Liao
國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
chengyang@speech.ntut.edu.tw, yfliao@mail.ntut.edu.tw

潘振銘 Chen-Ming Pan, 郭姿秀 Tzu-Hsiu Kuo
Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan
chenming@cht.com.tw gaga820402@cht.com.tw

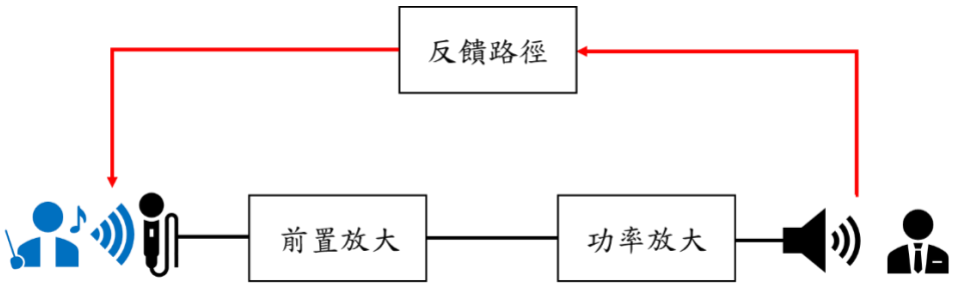
摘要

在使用卡拉 OK 系統唱歌時，常會因麥克風拿離喇叭太近，或是擴大機功率開太大，產生正回授而導致嘯叫，造成歌者跟聽眾都非常不舒服。一般處理麥克風嘯叫，常是利用移頻打斷共振，或是用帶阻濾波器做事後補救，但有可能會造成音質破壞。因此我們想改用適應性回授消除演算法，利用擴大機喇叭的輸入音源當參考訊號，來自動估算在不同空間環境、不同歌曲、不同訊雜比下，麥克風可能錄到的回授訊號，並在做訊號增益前先將其消除，以直接從源頭消除嘯叫發生的可能性。基於以上想法，在本論文中，實現了 normalized least mean square（NLMS）的嘯叫消除演算法，尤其是進一步考慮擴音系統的非線性失真，提出基於 recurrent neural network（RNN）的進階演算法。並在實驗時分別測試在時域或是頻域處理，與使用 NLMS 或是 RNN，對不同曲風、不同環境空間響應情況下，不同演算法的收斂速度、計算量需求與嘯叫抑制效果。由實驗結果發現：(1) 在時域實現收斂比較快，(2) 在頻率可實現計算量小於時域，(3) RNN 在收斂速度及突然變化的頻率消除上優於 NLMS。

關鍵詞： NLMS、遞迴類神經網路 RNN、適應性濾波器、麥克風嘯叫

一、簡介

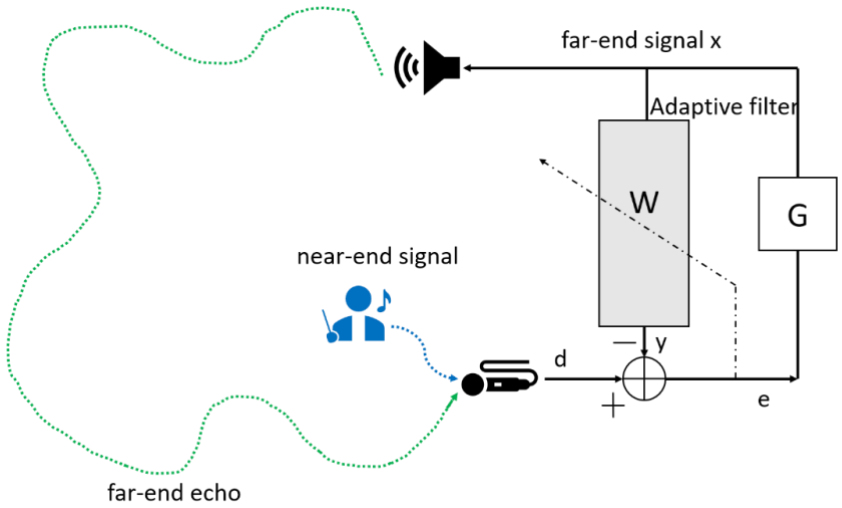
在使用麥克風唱卡拉 OK 的系統中，由於收音的喇叭與麥克風並未隔離在不同區域，當喇叭發出之聲音透過空間傳到麥克風，由於放大電路增益過高而導致正回授反饋如圖一，不斷將放出之聲音重複收入進而發生嘯叫。此種閉鎖迴路的嘯叫主要原因為整個電路與環境對某些共振頻率的增益過大，當提升喇叭通道之增益時，這些增益過大的共振頻率先達到聲學反饋所需的強度條件，若此頻率的反饋類型剛好為正反饋，則必定在此頻率上產生自激震盪現象，便是我們所說的嘯叫。



圖一、麥克風正回授嘯叫

為解決這件事情，可針對硬體上麥克風的指向性收音，將特定方向的聲音收進麥克風，盡量不讓除了近端人聲外的聲音收入，亦或者是增加麥克風之收音敏感度，可減少非必要收入麥克風的聲音;在演算法上有使用移頻打斷共振,將訊號頻率做些許升頻或降頻，破壞了嘯叫的發生條件，進而抑制了嘯叫，或是使用帶阻濾波器將會發生嘯叫的頻段做衰減，如果衰減這些過強的頻率就能抑制住嘯叫，但此兩者雖然只對小範圍的頻率做了調整,但仍會破壞原始訊號聲音,甚至是人耳可聽出的差別,且嘯叫仍是時不時的發生，因而目前在實際控制嘯叫發生的作法仍是治標不治本。

為了改善過去抑制嘯叫的缺點，這裡我們使用適應性濾波器來提前消除回聲以抑制嘯叫，首先我們先介紹基於 NLMS 之適應性濾波器演算法的回聲消除系統[1-2]如下圖二，利用擴大機喇叭的輸入音源作為參考訊號 x，自動估算在不同空間環境、不同歌曲、不同訊雜比下，麥克風可能錄到的回授訊號，再將預測出之期望信號與輸入的麥克風訊號 d 相減，使回聲訊號增益前就將其消除，直接從源頭消除嘯叫發生的可能性。



圖二、傳統聲學回聲消除系統架構圖

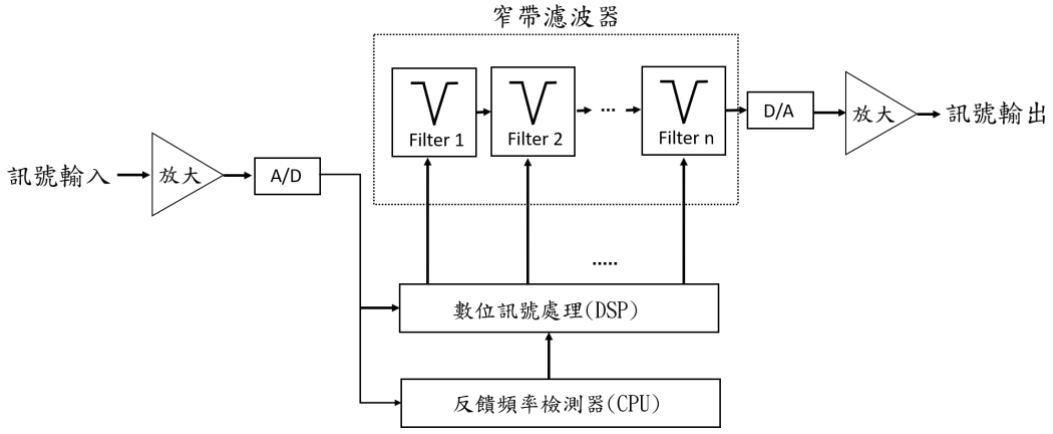
然而在使用卡拉 OK 麥克風時，時常因有能量非常大的擴大機喇叭撥出之聲音，與室內

複雜的空間環境響應，會有嚴重的非線性失真，使得線性演算法的 NLMS 無法有效的解決非線性問題，進而提出了基於 RNN 的進階演算法。由於 RNN 是擁有回授功能的遞迴類神經網路，能將上個過去的時間之輸出值儲存下來，再重新導回到輸入端，使得系統能夠抓取長度較長的時間輸入訊號，讓系統擁有龐大的過去資料來學習環境響應的路徑，改善非線性的部分。

此外嘯叫的發生都是即時且突然多變的，而時域的每一點調變一次，在面對突然變化的音樂或頻率變化，不確定是否仍能有效的收斂與消除，但在頻域做演算法可顧及到不同頻段的環境訊號，因此我們也針對此點做了時域、頻域演算法的比較，觀察時域與頻域在細膩度與收斂速度上是否有著明顯差異。另外也因 RNN 在計算量與 NLMS 有著明顯的差距，我們也觀察其計算量是否有與其效果成正比。因此基於上述考量，於本篇論文中我們將提出遞迴類神經網路麥克風抑制嘯叫系統，比較傳統時域、頻域 NLMS 和時域 RNN 在不同曲風、不同環境空間響應情況下，不同演算法的收斂速度、計算量需求與嘯叫抑制效果，在以之後章節將會介紹其模擬方法與架構。

二、相關研究

常用的回聲消除方法包括過去的帶阻濾波器、移頻，到近年主流的 NLMS 和基於預測誤差方法的適應性濾波器(Prediction Error Method-based Adaptive Feedback Cancellation，PEM AFC)。過去在使用移頻是使用一種可以改變聲音頻率的設備移頻器，其工作原理類似變調器，能夠將聲音訊號增加、減少 5Hz，破壞了嘯叫發生的條件，進而抑制了嘯叫，雖然對聲音的破壞很小，但其在演唱和樂器中就會有明顯差異，光 5Hz 的音調變化對人耳已經有明顯的感覺了；此外帶組濾波器至今在應用上仍是主流抑制嘯叫的工具，在音響系統中出現嘯叫是由於正反饋使音頻信號中的某些頻率點不斷被加強而造成的，因此將這些頻率點切除或進行大幅度衰減，就可以有效抑制聲反饋，其原理如圖三，主要是利用機器快速掃描尋找出發生嘯叫的頻段，並自動生成一組與這些嘯叫頻率相同的窄帶濾波器來切除嘯叫頻率，進而達到消除回聲而抑制嘯叫[3]。



圖三、帶組濾波器架構

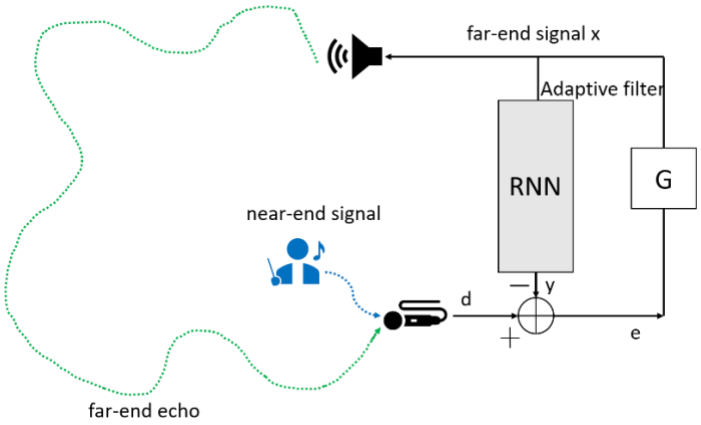
為了有別於過去方法，我們希望可以在嘯叫發生前就將嘯叫清除，因此使用了適應性濾波器，在嘯叫發生前先讓濾波器學習環境響應的路徑而改變其權重，穩定的消除多出來的回聲，其中最小均方演算法(LMS)是最易實現、且穩定及計算量小[4]，時至今日仍受到許多人喜愛且廣泛地運用，同時為了解決系統收斂緩慢的缺點。便將輸入訊號正規化，而演變成 NLMS 演算法，其採用可變步長的方法來穩定收斂過程。

另外由於閉鎖迴路的關係，近端訊號和揚聲器之間有嚴重的相關性問題，造成嘯叫更容易發生，因此後面有關於回聲消除的問題便有部分人著重在去相關性上面，有人在閉鎖迴路中的前向路徑加入全通濾波器來降低相關性[5]，此外也有論文提出了預測誤差方法適應性濾波器(PEM AFC)來減少近端訊號與揚聲器之間相關性[6]。而在 PEM 方法中，是利用反向近端訊號模型對麥克風和喇叭進行預濾波，然後將這些訊號送至自適應濾波演算法中，此便達到了降低相關性的目標。而對於近端語音訊號，通常使用線性預估(Linear Prediction，LP)來估計[7]，語音訊號由於時間相近的訊號點彼此有相關性，每個訊號點可由相近的訊號點藉由線性組合加以逼近或估測。故可藉由 LPC 估計去相關預訓練濾波，將語音訊號中分離成口腔與聲帶訊號，以去除喇叭輸出訊號與人聲訊號各自的相關性。

預測誤差方法次適應性濾波器能夠從期望訊號中去除相關分量，因此該反饋消除器對不相關信號和夾帶(entrainment)能夠使之不反應，但不幸地，當期望訊號是週期性號時，預測誤差方法自適應濾波器將會給出零，倘若是此種情形嘯叫便不能被消除了，因為嘯叫也是有週期性的，因此若發生此種情形或許需要回頭依靠傳統系統，因而也有人對此現象做出了嘯叫偵測器對此情況作出傳統和 PEM 發法的切換[8]。

三、適應性濾波器抑制嘯叫系統

由於麥克風與喇叭的閉鎖迴路會在室內反射造成回聲，而回聲時間會受房間大小影響，但傳統回聲消除系統會被輸入的長度固定而限制住，倘若輸入時間過長，將導致計算量過於龐大而收斂較慢，因此常無法有效地提取使用者長時間的訊號，故本論文改用遞迴類神經網路，其架構圖如圖四，將可以由回授路線看到過去較長時間的資料，以學習較長時間的環境響應路徑，並推測下一時間點的聲音，且僅用單層的 RNN 便足以獲取足夠的長時間資訊，因此多層的 RNN 就能看得更深更廣泛，抓取更多的訊息。下列先說明時域與頻域的 NLMS 做法，後面再介紹遞迴類神經網路的適應性濾波器消除演算法。



圖四、基於遞迴類神經網路之適應性濾波器抑制嘯叫系統架構圖

(一) NLMS

其模擬為利用麥克風所收集到的聲音 d 與透過喇叭輸出的聲音 x，利用 NLMS 預測可能錄到的環境響應，使其相減後將回聲消除，此時系統的誤差訊號 e 為輸出得到的清晰語者聲音。下列為 NLMS 演算法:

$$e(n) = d(n) - \hat{w}^H(n)u(n)$$

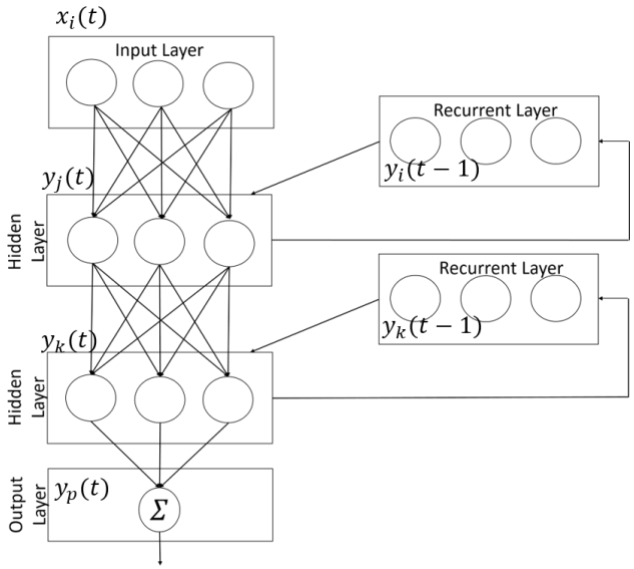
$$\hat{w}(n+1) = \hat{w}(n) + \frac{\tilde{\mu}}{\|u(n)\|^2} u(n)e^*(n)$$

在頻域 NLMS 演算法部分，我們每 512 點取一個音框，將此音框所有點由 FFT 轉為頻域，之後把每個音框的 512 點皆做一次 NLMS，運算完成後再將訊號由 Inverse FFT 轉回時域後做 overlap-add。此外顧慮到環境響應殘響長度，我們將取音框的數量增至每次取 8 個音框，讓演算法有足夠長度的時間序列去做學習。

(二)神經網路 RNN

這裡我們使用的架構為透過近端人聲與喇叭輸出的音檔結合後為麥克風收音 d，利用遞迴類神經網路預測麥克風收進的環境響應 y，相減後將環境響應消除，最後系統輸出誤差訊號 e 即為得到的剩餘近端人聲之聲音。

下圖五為本論文中使用的遞迴類神經網路運作結構。本論文之遞迴類神經網路架構為兩層的神經網路，輸入使用滑動視窗的方式每次抓取 1024 點、兩個隱藏層皆是 100 個節點，最後輸出為當前時間的點，此外擁有兩個的遞迴層。



圖五、深層遞迴網路架構圖

其想法為透過回授的方式，把過去的時間點之隱藏層節點輸出記錄起來，並重新輸入至隱藏層節點之輸入端，將輸入層與輸入值此兩個資訊連結在一起，以之作為下一個時間點的隱藏層輸入，讓當前的節點保有過去的輸出資訊，使得整個網路架構有記憶的特性。此外 RNN 的核心演算法為反向傳遞演算法(Backpropagation)，以下為反向傳遞演算法的推導:

最初輸入$x_i(t)$透過權重v_{ji}傳遞到第一個隱藏層，再加上第一層的隱藏層的偏壓值b_j，經過轉換函數f(·)，產生在第一個隱藏層第一個時間點的神經元輸出$y_j(t)$，如下式(3.1)為第一層隱藏層輸出，n 為輸入個數。

$$y_j(t) = f\left(\sum_i^n v_{ji}x_i(t) + b_j\right) \quad (3.1)$$

此時的第一層隱藏層輸出會輸入到第一層的回饋層，回授連結權重$r_{1,jm}$再到第一層隱藏層，也就是上一個時間點(t-1)神經元的轉換狀態，下式(3.2)為結合輸入$x_i(t)$所產生的新的輸出方程式。

$$y_j(t) = f\left(\sum_i^n v_{ji}x_i(t) + \sum_i^m r_{1,jm}y_i(t-1) + b_j\right) \quad (3.2)$$

由上層隱藏層輸出$y_j(t)$會輸入第二層的隱藏層，權重w_{kj}連接第二層隱藏層，加上第二層隱藏層偏壓值 bk，經轉換函數f(·)轉換，產出在第二個隱藏層這一個時間點的神經元輸出$y_k(t)$，其輸出如下方程式(3.3)。

$$y_k(t) = f\left(\sum_j^n w_{kj}y_j(t) + b_k\right) \quad (3.3)$$

此時的輸出一樣會輸入到第二層回饋層，透過回授連接權重$r_{2,kg}$第二層隱藏層，也是上一個個時間點(t-1)神經元轉換狀態，下式(3.4)為結合輸入$y_j(t)$所產生的新的輸出方程式。

$$y_k(t) = f\left(\sum_j^n w_{kj} y_j(t) + \sum_j^g r_{2,kg} y_j(t-1) + b_k\right) \quad (3.4)$$

最後再由權重 upk 連接第二層隱藏層輸出$y_k(t)$到輸出層$y_p(t)$，此時$y_p(t)$就是深層遞迴式網路最終輸出。

有別於一般 RNN 的反向傳播演算法(Backpropagation Through Time，BPTT)，我們會根據時間前後順序來調整權重值，因此調整權重會經由不同時間點的隱藏層資訊進行，由最後時間點對成本函數(cost function)作偏微分，往前估算出一開始時間點的偏微分值，直到整個權重作出調整完後，再代回網路求新誤差值，使 MSE 接近最小值，讓輸出接近期望的值。

四、實驗結果

本論文使用之音檔為自行錄製真實人聲(A Cappella)與人聲同一曲目的伴唱音檔，比較時域 NLMS、頻域 NLMS、時域 RNN 等三個演算法。

為避免麥克風一打開收音就接收大量能量，以致適應性濾波器尚未收斂便引發不可收拾的嘯叫，因此將測試音檔皆編輯成:第一段(音樂)、第二段(音樂)、第三段(人聲)、第四段(人聲)、第五段(人聲)、第六段(音樂)之長度，而第一段(音樂)為純伴唱音樂且未將濾波器消除結果輸出至喇叭，為單純學習環境響應、調整濾波器權重；第二段(音樂)才將濾波器消除結果加入喇叭輸出，到第三段(人聲)才正式將近端人聲加入系統之中。此模擬類似為 google 之 Google Home 與 apple 之 HomePod，在開機時給予一個提示音，使系統學習環境響應的情況，避免每次開機皆是不同環境的情形。

(一)音檔說明

表一、音檔格式

	位元率	聲道數量	取樣頻率	長度(秒)	曲風
眉飛色舞	16-bit	單聲道	16000 Hz	33.5	快歌
天黑黑	16-bit	單聲道	16000 Hz	33.5	慢歌

(二)環境響應說明(房間大小、回音長度)

以下環境響應為 RIR-Generator 生成之環境響應音檔[9]，我們將此音檔摺積輸出音檔以之來模擬真實環境下麥克風所接收到的環境響應回聲。以下表二為環境響應參數設置，皆為聲速(Sound velocity): 340 (m/s)、取樣頻率(Sample frequency) = 16000 (sample/s)。

表二、環境響應參數設置(單位:公尺)

	大房間(禮堂大小)	中房間(會議室大小)	小房間(車內大小)
Source position	[10.0 4.0 2.0]	[2.5 1.0 1.2]	[1.7 0.2 0.2]
Receiver position	[6.0 4.0 1.5]	[2.5 2.0 1.6]	[1.7 0.8 0.8]
Room dimensions	[15.5 8.5 6.0]	[5.0 4.0 3.0]	[2.3 1.7 1.2]
Beta(殘響長度)(s)	1.8	0.45	0.1

(三)抑制評估-MSE

回聲消除效果除了主觀由耳朵聽聲音之外，我們將使用平均誤差值(Mean Squared Error，MSE)來數值化其回聲消除的效果上，其方程式如下所示:

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(s(n) - e(n))^2$$

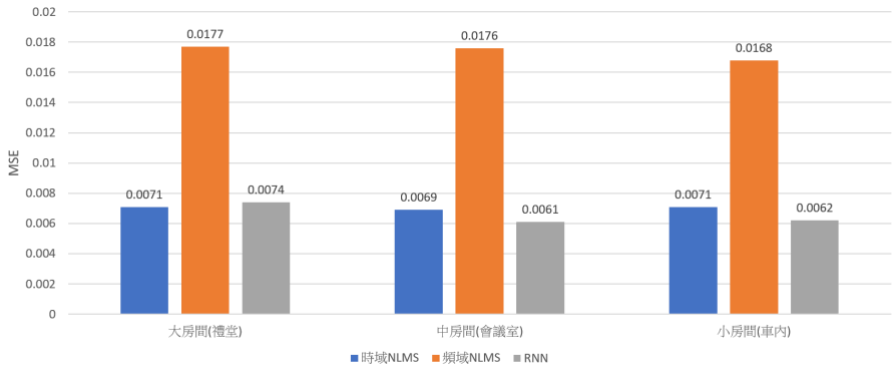
其中，s(n)為近端人聲之原始音檔。e(n)為剩餘訊號:即經回聲消除系統消除環境響應後得到的剩餘人聲。借由原始的近端人聲 s(n)與經回聲消除的剩餘人聲 e(n)，兩者相相減取平方，當 MSE 值愈小，表示消除效果愈好，代表也得到愈乾淨的近端人聲。

(四)實驗結果

每個實驗皆含 (時域 NLMS、頻域 NLMS、RNN)

1.實驗一，同一首歌_不同環境(房間大小、回音程度)

下圖六中，在不同環境下每個演算法的效果皆是穩定的，整理來看是時域會好過頻域，不過以演算法來說 RNN 與 NLMS 彼此的效果卻是相差不多的。

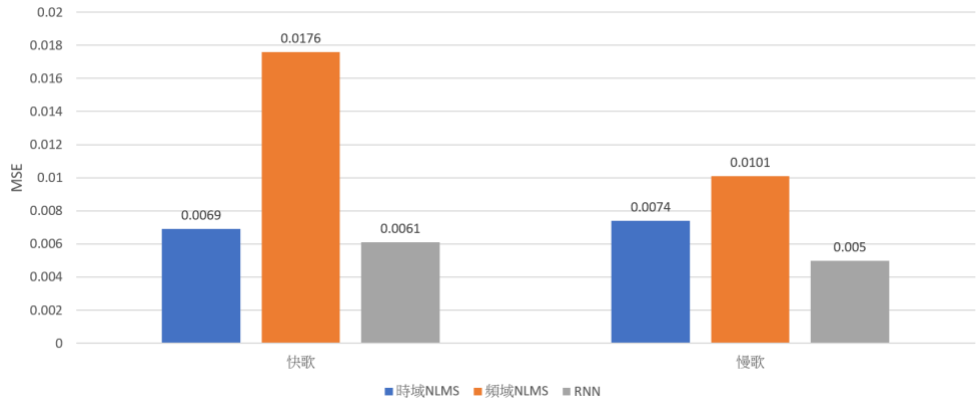


圖六、實驗一比較圖

2.實驗二，同一環境_不同首歌

下圖七中，在不同首歌下慢歌比快歌效果來的更好，這或許跟嘯叫發生頻段有關，在快

歌中的某些頻段剛好與嘯叫發生頻段符合，因此快歌在消除效果上沒慢歌來的好。

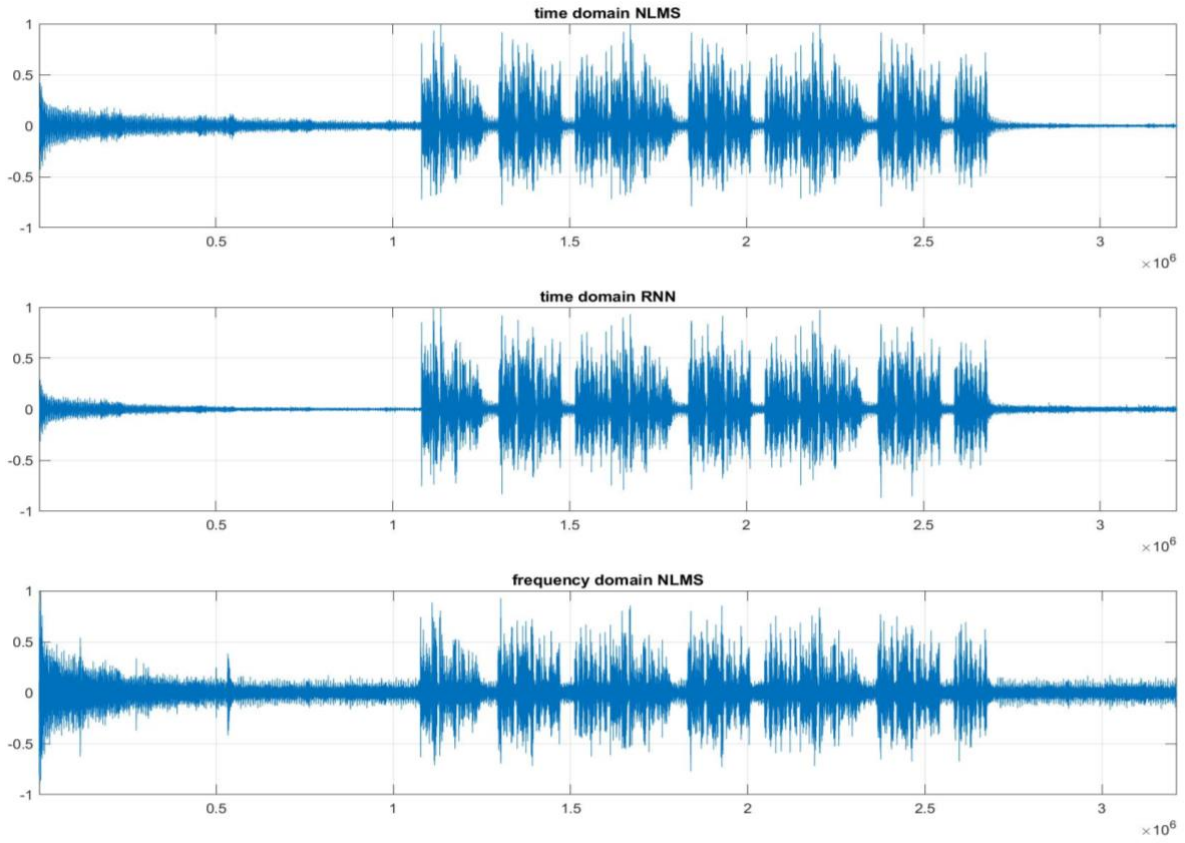


圖七、實驗二比較圖

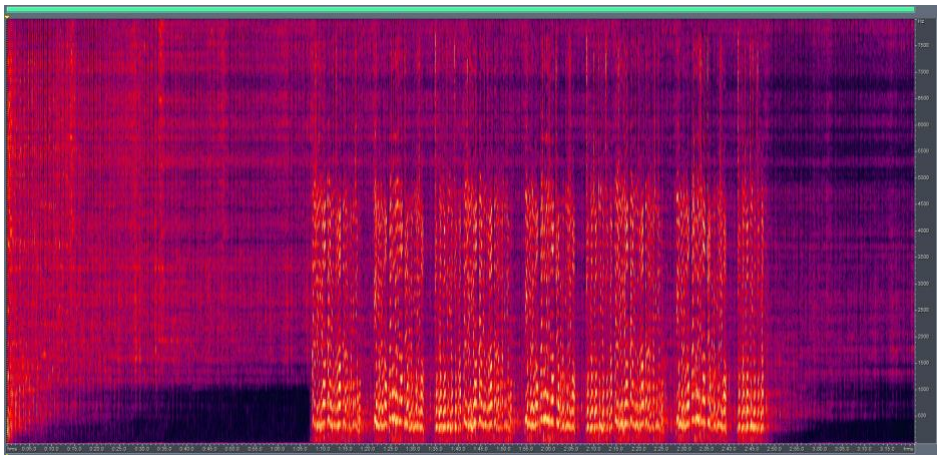
3.時域 NLMS、頻域 NLMS、時域 RNN 比較(同一首歌、同一環境)

下圖八、九、十、十一，為時域 NLMS、時域 RNN、頻域 NLMS 的時域圖、頻譜圖比較，比較下來時域 NLMS 與時域 RNN 效果相差不多，而頻域 NLMS 由於收斂比較慢的關係，所以效果略遜色於其他兩者，接下來會比較第一、二段與最後一段的效果與收斂速度。

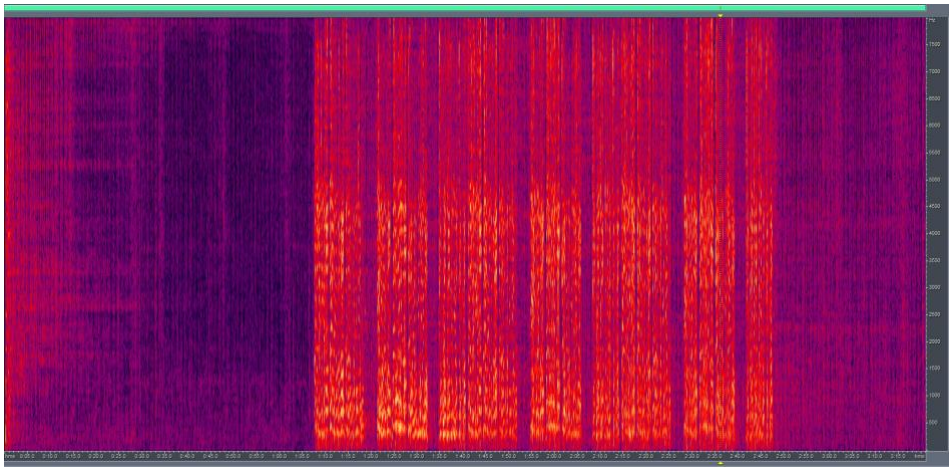
.



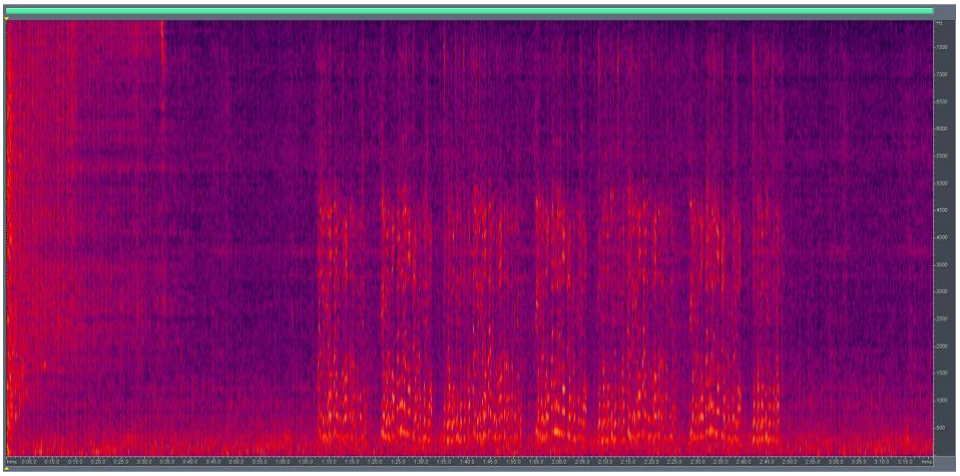
圖八、時域 NLMS、時域 RNN、頻域 NLMS 時域比較圖



圖九、時域 NLMS 剩餘人聲頻域圖



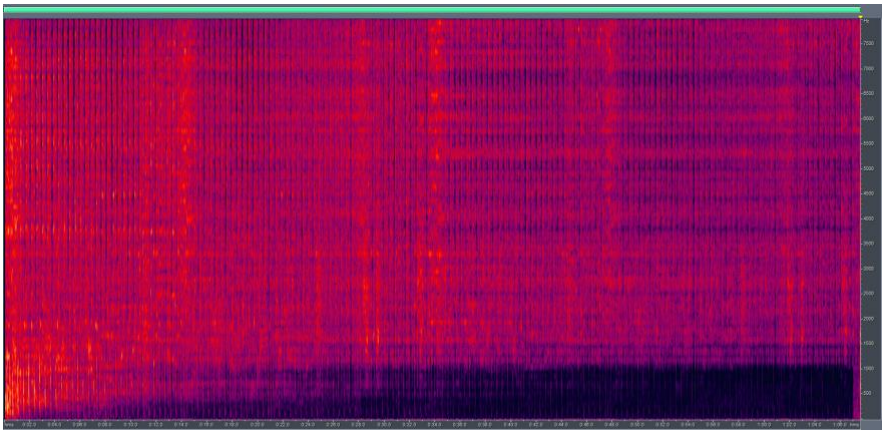
圖十、時域 RNN 剩餘人聲頻域圖



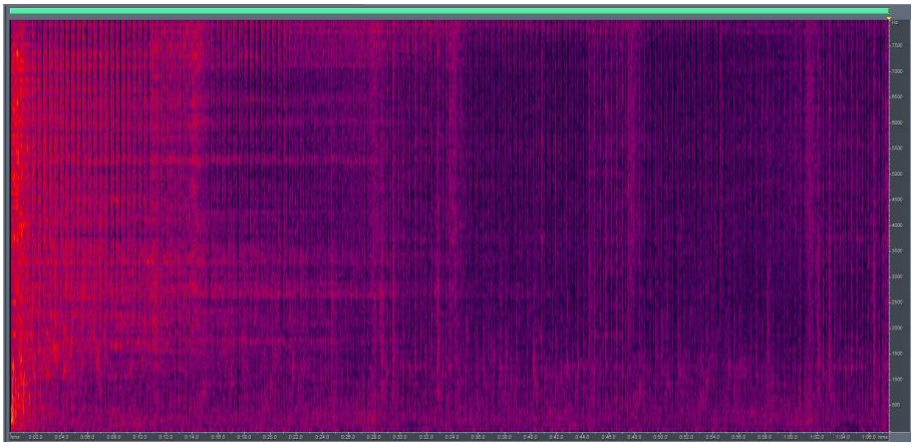
圖十一、頻域 NLMS 剩餘人聲頻域圖

由圖十二、十三、十四可知，第一段、第二段收斂過程中，時域 RNN 的消除效果較好且較快，接著是時域 NLMS，但從演算複雜度來看，時域是每一點都計算一次，也就是說第一、二段 67 秒的歌曲中便演算了 NLMS 1072 千次，每一點都調整一次權重；而頻域由於是每 256 點才取一次 512 點數做運算，512 點每一點都用一樣的調整量，頻域的

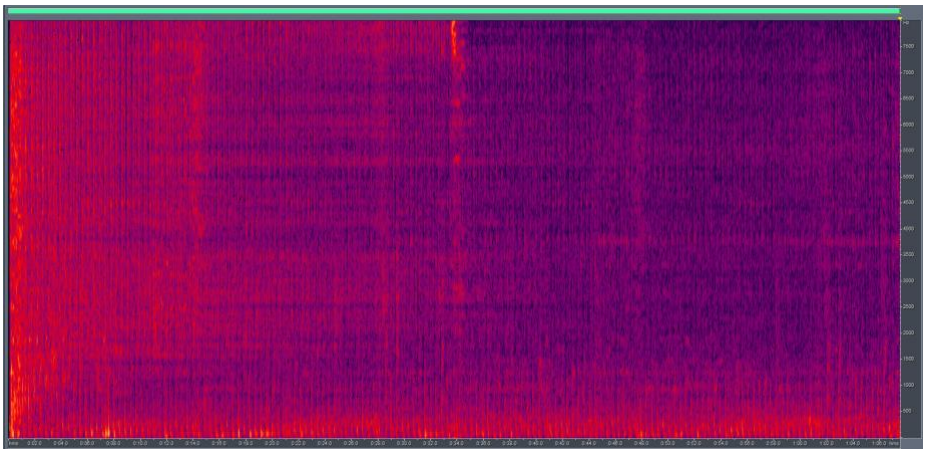
收斂速度會相較慢的。



圖十二、時域 NLMS 第一、二段頻域圖



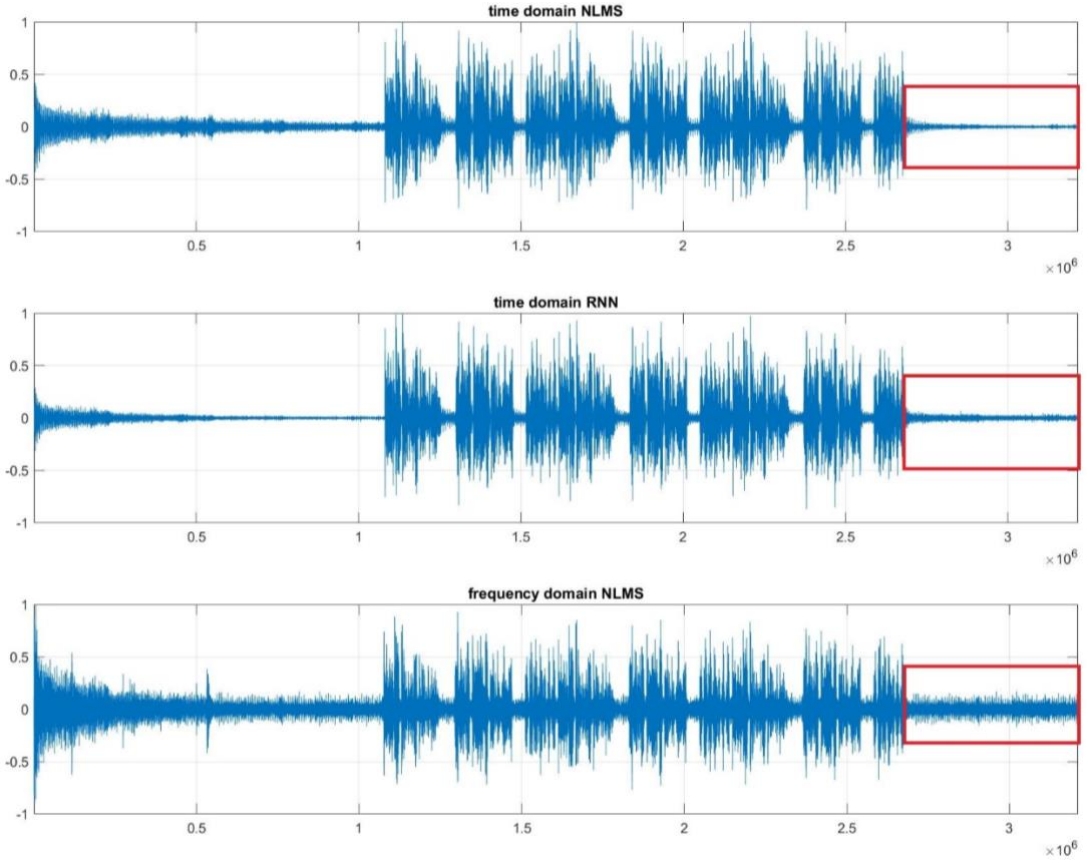
圖十三、時域 RNN 第一、二段頻域圖



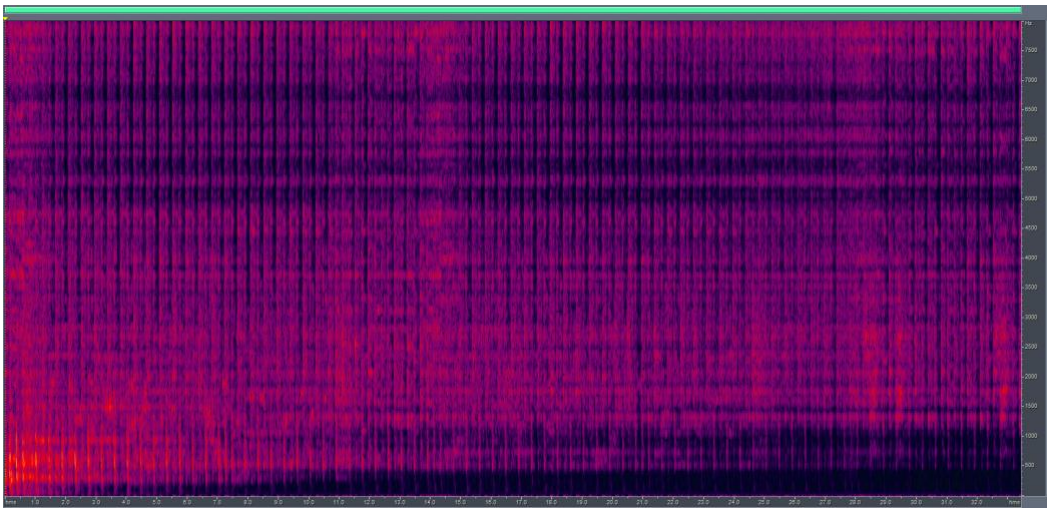
圖十四、頻域 NLMS 第一、二段頻域圖

經過中間第三、四、五段有加入人聲後，回到最後一段僅有音樂的第六段，如圖十五紅框部分所示，此時彼此都已經收斂得差不多了，但時域部分此時已經演算了 3216 千次的運算量了，而頻域演算法才做了 12.56 千次左右的計算與 FFT，彼此在 NLMS 計算量

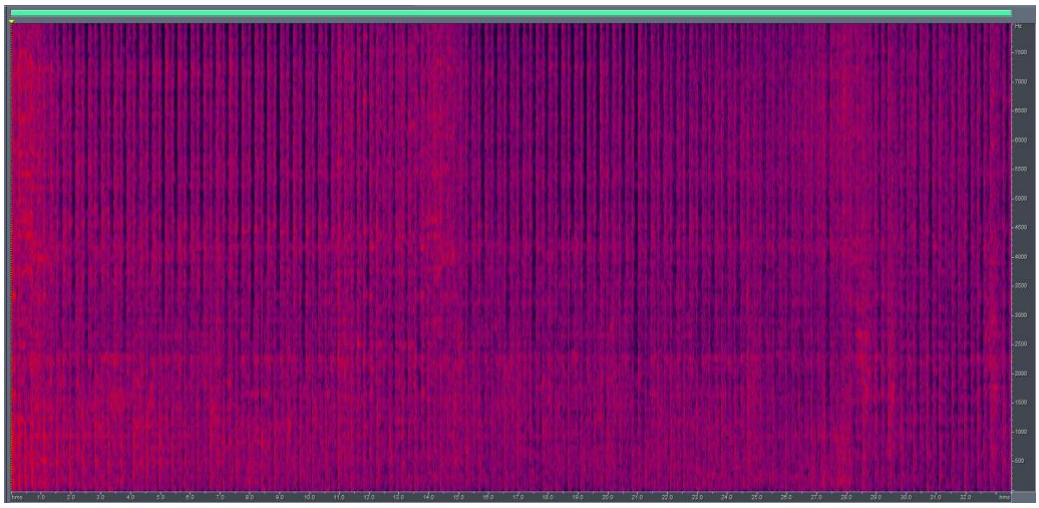
仍相差了 256 倍，而時域 RNN 是其中演算最複雜的。此外從圖十六、十七、十八比較可得知，頻域 NLMS 在 500 Hz 的地方仍有些許沒消乾淨，而時域 NLMS 在音檔突然變換的情形下，就需要一段時間才能重新收斂，但頻域 NLMS 因為有照顧到每個頻段的關係，所以突然的變化，仍可以穩定消除回聲；而 RNN 雖然也為時域運算，但因為其為非線性的演算法，在這種情形下便仍有良好的效果。在做更久的演算下，時域或許仍因每一點都運算一次而比頻域效果來的更好，但由於彼此都收斂的情形下，效果並不會相差太多，但三者的演算量與穩定度在日後的應用上，需要在效果與計算量上做取捨了。



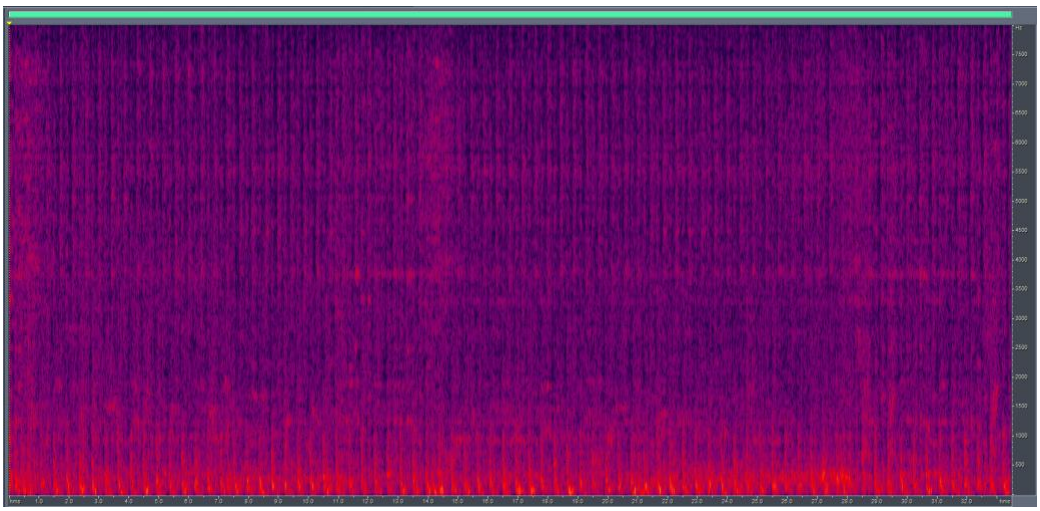
圖十五、時域 NLMS、時域 RNN、頻域 NLMS 第六段時域圖



圖十六、時域 NLMS 第六段頻域圖



圖十七、時域 RNN 第六段頻域圖



圖十八、頻域 NLMS 第六段頻域圖

五、結論

在本實驗中，模擬了真實環境中麥克風在不同空間下的收音與環境響應情形，分別模擬了大房間、中房間、小房間的環境響應，語料部分也分了不同曲風的歌曲來做實驗，接著分別測試了線性的時域 NLMS、頻域 NLMS 以及非線性濾波演算法 RNN，作麥克風嘯叫抑制實驗。在時域 NLMS 與頻域 NLMS 上，在一開始時域的由於是每一點就做一次，其收斂效果會比頻域來的更快，甚至消除效果更好，但在兩者都演算了一段時間後，彼此都已達到了收斂，效果其實是差不多的，但頻域在某些突如的高頻或低頻放面會比時域來的效果更好，計算複雜度上，頻域的摺積比時域來的簡單，且 256 點才做一次運算，但每次都是一次調整 512 點，因此計算量是相差不多的。此外時域 RNN 上由於非線性與有時間的記憶，在某些部分消除的效果其實是最好的，但由於其運算量龐大，日後若有應用，在這三者之間，計算量與效果好壞的取捨便端開使用的情況。

Acknowledgements

This work was partly supported by Taiwan Ministry of Science and Technology MOST contract No. 107-2911-I-027-501, 108-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067 and partly supported by Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan contract No. TL-108-D301.

參考文獻 [References]

[1]. 胡立宁. 自适应回声消除算法的研究与实现. MS thesis. 吉林大学, 2007.

[2]. Stenger, A., L. Trautmann, and R. Rabenstein. "Nonliner Acoustic Echo Cancellstion with 2nd Order Adaptive Volterra Filters, IEEE Int." Conf. on Acoustics, Speech & Signal Procrssing (ICASSP). 1999.

[3]. 杜鷁、吳樂華，電聲技術與音響系統/國防工業出版社/ 2015-05-01

[4]. Tyagi, Ranbeer, Roop Singh, and Rahul Tiwari. "The performance study of NLMS algorithm for acoustic echo cancellation." 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC). IEEE, 2017.

[5]. Boukis, Christos, Danilo P. Mandic, and Anthony G. Constantinides. "Toward bias minimization in acoustic feedback cancellation systems." The Journal of the Acoustical Society of America 121.3 (2007): 1529-1537.

[6]. Ngo, Kim, et al. "Prediction-error-method-based adaptive feedback cancellation in hearing aids using pitch estimation." 2010 18th European Signal Processing Conference. IEEE, 2010.

[7]. Van Waterschoot, Toon, and Marc Moonen. "Adaptive feedback cancellation for audio applications." Signal Processing 89.11 (2009): 2185-2201.

[8]. Kashima, Kakeru, et al. "Adaptive feedback canceller with howling detection for hearing

aids." 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 2015.

[9]. Habets, Emanuel AP. "Room impulse response generator." Technische Universiteit Eindhoven, Tech. Rep 2.2.4 (2006): 1. Available: https://github.com/ehabets/RIR-Generator [Accessed: Jul. 15, 2019]

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 137-151

基於深度類神經網路之多模式情感偵測初步探討

A Preliminary Study on Deep Learning Neural Networks-based Multi-Model Sentiment Detection

陳泰融 Tai-Rong Chen, 廖元甫 Yuan-Fu Liao
國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
t106368030@ntut.edu.tw, yfliao@mail.ntut.edu.tw

潘振銘 Chen-Ming Pan, 郭姿秀 Tzu-Hsiu Kuo
Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan
chenming@cht.com.tw, gaga820402@cht.com.tw

Matúš Pleva, Daniel Hládek
Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia
matus.pleva@tuke.sk, daniel.hladek@tuke.sk

摘要

為慶祝登月計劃五十週年，德州大學達拉斯分校（UTDallas）將登月任務中，所有太空人與任務中心間的通訊對話錄音進行數位化，發行 Fearless Steps Corpus 語料，並舉辦 Fearless Steps Challenge 競賽，希望能增進各種語音處理相關技術發展。本論文即針對其中的語音情緒偵測任務，進行初步探討。主要想法是同時考慮語音訊號中包含的聲學與語意資訊，提出基於深度類神經網路之多模式語音情緒偵測模型，用以偵測語音訊號中傳達的情緒狀態。實際做法包括(1)利用捲積神經網絡(Convolutional Neural Network, CNN)，從聲學頻譜自動求取情緒特徵參數，與（2）以雙向編碼變換器（Bidirectional Encoder Representation from Transformers, BERT），求取語音逐字稿的文字語意特徵參數。再將此兩類特徵參數向量融合，以強化系統的情緒狀態偵測效能。最後由正式比賽結果發現，我們的系統的情緒狀態偵測正確率達到 73.11%，在所有隊伍提交中的 20 個結果中，排第三名，不但超越主辦單位提供的基準參考系統(49.75%)，並只差第一名(74.07)不到 1%。

關鍵詞：情感檢測，CNN，BERT, 多模式情感檢測

一、簡介

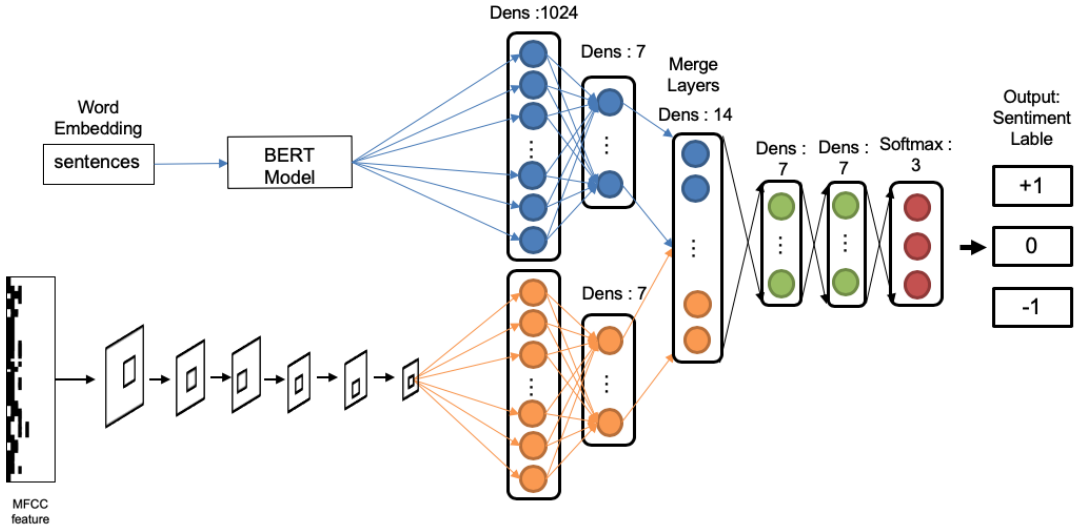
本論文針對 Fearless Steps Challenge 競賽中的 sentiment detection 任務，進行語音情感偵測初步探討。Fearless Steps Challenge 比賽，是為了慶祝登月計劃 50 週年所舉辦的大規模競賽。由德州賽拉達分校將登月任務中所有通訊對話數位化，並發行 Fearless Steps Corpus 語料，支援各個競賽項目，提供大量的訓練資料及測試資料，因為此項比賽主要是希望可以使用自然環境當中所錄製的資料庫進行比賽，所以 Fearless Steps Corpus 的語音資料，是真實太空任務中，太空人與任務中心的通訊對話錄音。

我們會選擇參加此項競賽，主要是因為目前大部分可取得的情緒相關語料庫，大都是由演員表演的，且語料都過於完美或是過於乾淨，導致在這些語料庫上所獲得的結果，不見得可以反應實際運用時的情境。而 Fearless Steps Challenge 的語音資料，是真實太空任務中，太空人與任務中心的通訊對話錄音，因此會有許多自然的雜訊和對話。最重要的是，此 Fearless Steps Corpus 語料庫總共包含 100 小時的語料，而且情緒標籤都是經由人工標註驗證，因此研究獲得的結果會更加有公信力。

傳統上針對語音情緒偵測，通常專注於先提取情緒的低階聲學特徵[3-14]。一些廣泛使用的頻譜特徵是 Mel-Frequency Cepstral Coefficients（MFCC）[1]、線性預測倒譜係數或是音高軌跡。然後再用高斯混合模型、支持向量機或是馬爾可夫模型進行情緒辨認。而若想從語意來求取情緒特徵參數，則需要先有語音辨認器，將語音轉成逐字稿，再以自然語言處理方式，例如以 word-to-vector 求取特徵向量，再以類神經網路進行情緒辨認。然而，人類情感與聲學低階特徵的表現，實際上不見得一致。而若用逐字稿，則通常會有語音辨認錯誤，影響最終判斷的情形。

針對 Fearless Steps Challenge 比賽，我們在進行初步實驗測試時，發現若單獨只用聲音製作模型，或是單獨使用文字訓練模型，所得到效果都有所不足。主要是語音中的情緒特徵，可能同時表現在音色、語氣或是文字用語上。因此在比賽當中，我們除了分別嚐試對於聲音和逐字稿抽取其隱含的情緒相關特徵，並希望以多模式神經網路，將兩者的特徵參數進行結合，同時以聲音中與逐字稿中的情緒特徵來建立模型，以提升情緒偵測的正確率。

因此，本論文提出如圖一的多模式情緒偵測模型。主要想法是同時考慮語音訊號中包含的聲學與語意資訊，提出基於深度類神經網路之多模式語音情緒偵測模型，用以偵測語音訊號中傳達的情緒狀態。實際做法包括（1）利用捲積神經網絡（Convolutional Neural Network, CNN），從聲學頻譜自動求取情緒特徵參數，與（2）以雙向編碼變換器（Bidirectional Encoder Representation from Transformers, BERT），求取語音逐字稿的語意特徵參數。再將此兩類特徵參數向量融合，以強化系統的情緒狀態偵測效能。圖一為我們進行情緒偵測的框架結構：



圖一、多模式神經網路模型架構圖

此系統的運作包含三個模組，包括：（1）我們將原始語音信號轉換為類似圖像的頻譜圖方式，作為 CNN 的輸入[2]。因此，可以使用以大量語音語料預訓練的深度 CNN 模型進行學習，擷取高級聲學情緒特徵。（2）對於逐字稿的多個連續段落，可以用以大規模文字數據集預訓練的 BERT 模型進行訓練，萃練高階的語意情緒特徵。（3）由 2D-CNN 和 BERT 學習的聲學和語意情緒特徵參數，被集成在多模式的融合網絡中。最後，我們採用多模式的最後一個隱藏層的輸出作為分段的情感標籤。

二、Fearless Steps Challenge

(一)、數據集

為了評估本文所提出的模型性能，我們使用 Fearless Steps Challenge 所提供的美國宇航局阿波羅計劃的全程無線電通訊錄音資料庫，共有 100 個小時，包括火箭升空約佔

25 小時，登月約 50 小時，月球漫步約 25 小時。此外由於任務的不同，語料庫的語音活動密度在整個任務中常常變化，且語音數據的質量也常在 0 到 20dB (Signal-to-noise ratio , SNR)之間變化。Fearless Steps Challenge 為了確保能將數據公平地分配到的訓練，評估和開發子集中，根據噪聲水平與活動密度，對數據進行分類。

Fearless Steps Challenge 所提供的訓練子集，皆經人工轉寫逐字稿與標記情緒標籤。評估子集則只提供自動產生的逐字稿與情緒標籤。但測試子集則無提供任何情緒標籤也無逐字稿，因此本論文在測試資料時，需要對音檔先進行文字轉寫處理。由於 Fearless Steps Challenge 所提供得是太空中不同場景的語音記錄，總共提供了五個不同部分的頻道場景，Flight Director (FD)、Mission Operations Control Room (MOCR)、Guidance Navigation and Control (GNC)、Network Controller (NTWK)、Electrical, Environmental and Consumables Manager (EECOM)，表一提供不同事件的五個場景的時間分布表。

表一、Total Speech Durations per Channel and Event

	ECOM	**FD**	**GNC**	**MOCR**	**NTWK**	**Total**
Lift Off	2.1	1.2	1.3	0.8	3.9	9.3
Lunar Landing	3.7	1.3	4.0	0.9	4.4	14.3
Lunar Walking	3.9	1.1	3.0	1.4	2.8	12.2
Total	9.7	3.6	8.3	3.1	11.1	35.8

為了確保 Fearless Steps Challenge 數據公平性，在訓練資料和測試資料中由 Fearless Steps Challenge 來挑選 SNR 較為公平的資料，並根據靜音持續時間和語音持續時間來進行進一步挑選。表二為分別五個不同場景的 SNR 平均值和 SNR 的標準差。表上有分別五種不同的錄音場地，分別不同場景有分別不同的 SNR 標準差，其中 Mission Operations Control Room (MOCR)的標準差最高，但 Mission Operations Control Room (MOCR)在其中 SNR 平均值為最低，在這個錄音場景下的噪音動態範圍較為浮動。

表二、Signal to Noise Ratio Statistics (dB SNR) per channel for Dev Data

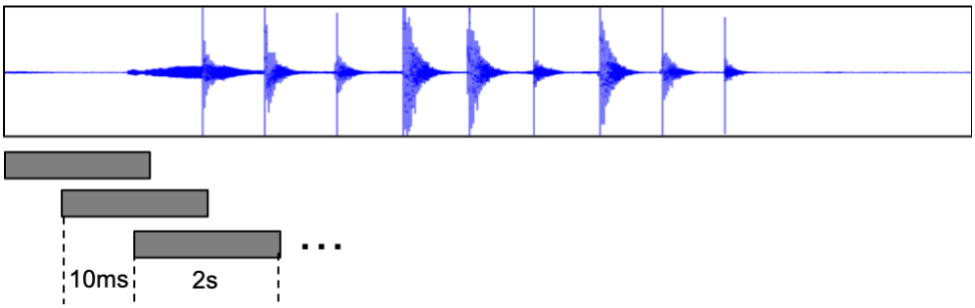
	ECOM	**FD**	**GNC**	**MOCR**	**NTWK**
SNR (Mean)	13.32	14.67	14.91	5.07	10.68
SNR (Std. Dev)	7.40	10.51	11.96	12.60	11.17

在訓練的過程中，由 Fearless Steps Challenge 來做 SNR 資料分群，Fearless Steps Challenge 在依據各個音檔的 SNR(Std Dev)數值去做消除雜訊的動作。

三、基於多模式之情緒檢測系統

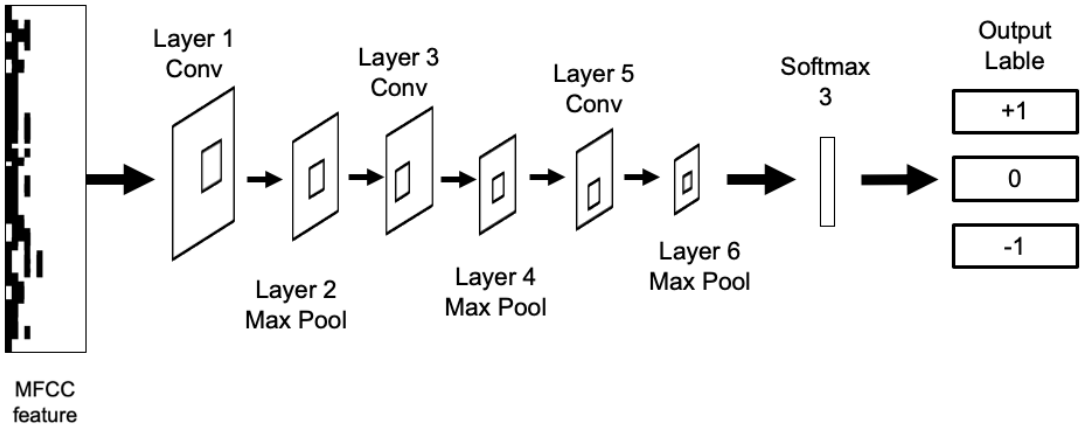
(一)、卷積神經網路聲學情緒模型架構

本篇論文提出的方法的第一階段，先對輸入語音信號執行取音框與求取語音信號的頻譜。其中我們使用 2 秒的窗口大小與 10ms 的音框位移，來獲得足夠可訓練資料。然後將訊號轉換至頻域，在此處以 Mel-frequency 三角形濾波器組過濾頻譜，轉成 Mel-frequency filterbank 參數，再獲得最終的 MFCCs。在本文中，我們還使用 20 個濾波器組和 40-MFCC 進行特徵提取，再將 MFCCs 矩陣輸入 CNN 中。



圖二、Sliding Windows for Sentiment Detection 示意圖

在我們的例子中，CNN 扮演一個從語音訊號頻譜中，提取聲學情緒特徵參數的重要作用。由圖三中可以看出，本論文將 MFCCs 作為 2D 放是作爲輸入，輸入緊接六層 CNN 的基本層數，如圖三所示，CNN 具有[INPUT-CONV-RELU-POOL-CONV-RELUPOOL]的基本架構。CNN 輸入的大小為 40 * 32，為了盡可能保留 Fearless Steps Challenge 提供的信息，我們為每個情緒資料利用 Sliding Windows 窗口採樣訓練數據。最後，將完整的 CNN 架構音頻識別的部分加入混合神經網路模型架構。

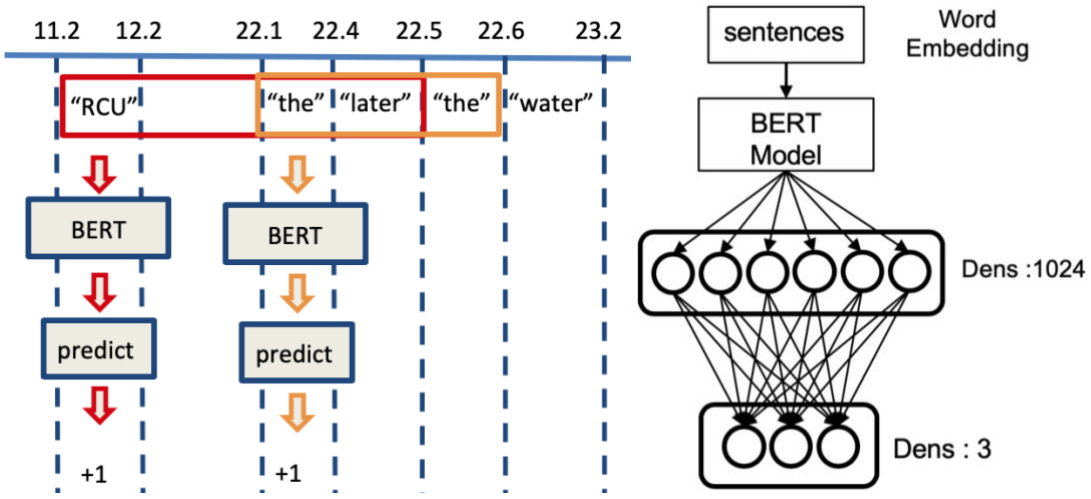


圖三、CNN Architecture for Sentiment Detection

(二)、BERT 神經網路語意情緒模型架構

我們使用 Google 的 BERT 模型。輸入的部分是情緒句子的句向量$[v1, v2, v3 \ldots]$，BERT 與其他模型不同的是，採用了一種簡單的方法，即隨機屏蔽（masking）部分輸入 token，然後只預測那些被屏蔽的 token。將這個過程稱為(masked LM, MLM)，他在訓練雙向語言模型時把少量的詞彙替換成 Mask。

本論文為了輸入較多跟情緒相依性的特徵，一樣採用 Sliding Windows 的方式對文字進行每幀的數據採樣，作法如圖四所示，在當前單字往前取 12 個單詞往後曲 12 個單詞總共 25 個單詞作為 BERT 句子的輸入，對句子做單一個詞的位移來取得下一句，得到句子之後對句子做 Word Embedding 輸入進 BERT Model 如圖四，最後串接 Dens 層連接 Softmax 進行分類。



圖四、BERT 輸入文字採樣示意圖及 BERT 串接 LSTM 示意圖

(三)、混合神經網路情緒模型架構

在長時訓練及辨識的文字和語音由 Fearless Steps Challenge 所提供的美國宇航局阿波羅計劃的全程無線電資料庫裡，我們使用混合模型將語音及文字進行同步訓練。

特徵級融合是最常見和直接的方式，其中所有提取的特徵直接連接成單個高維特徵向量。然後，可以用這種高維特徵向量訓練單個分類器用於情緒識別。大量先前的作品[15-19]證明了情感識別任務中特徵級融合的表現。但是，因為它以直接的方式合併音頻和文字特徵，所以特徵級融合不能模擬複雜的關係。 具體地，每個輸入模態用情緒分類器獨立建模，然後將這些識別結果與某些代數規則組合，例如：”Max”、”Min”、”Sum”等。因此，在情感識別中採用了決策融合。然而，決策層融合無法捕捉不同模態之間的相互關聯，因為這些模態被認為是獨立的。因此，決策級融合不符合人類情緒特徵的特性。

模型級融合作為特徵級融合和決策級融之間的折衷，也被用於情感識別最佳解決方法。該方法旨在獲得音頻和文字模態的聯合特徵表示。其實現主要取決於所使用的融合模型。例如，[4]採用(Hidden Markov Model , MFHMM)來實現模型級融合。 [8]採用誤差半耦合馬爾可夫模型融合以進行情感識別。對於神經網絡，通過首先連接對應於多個輸入模態的神經網絡的不同隱藏層的特徵表示來執行模型級融合。然後，添加額外的隱藏層以從連接的特徵學習聯合特徵表示。現有的模型級融合方法仍然不能有效地模擬音頻和文字模態之間的高度非線性相關性。

四、情緒偵測分類實驗

(一)訓練與測試語料

長時訓練及辨識的文字和語音由 Fearless Steps Challenge 所提供的美國宇航局阿波羅計劃的全程無線電資料庫，包括 100 個小時。選擇的阿波羅 11 號任務主要分為三個階段：（i）升空、（ii）登月、（iii）月球行走。為任務系統開發提供了 80 小時的音頻。在這 80 個小時內，提供了 20 小時的經過人工驗證的答案。對於剩餘的 60 小時音頻，提供 Baseline 系統生成的輸出答案，另外一組 20 小時將發布用於開放測試。

表三、 Fearless Steps Challenge 資料統計與比較

	Fearless Steps Challenge					
	Train			Dev		
	NEUTRAL	POSITIVE	NEGATIVE	NEUTRAL	POSITIVE	NEGATIVE
Avger time	0:00:30	0:00:13	0:00:24	0:00:02	0:00:01	0:00:02
Max time	0:11:10	0:09:22	0:11:10	0:09:03	0:00:16	0:09:03
Min time	0:00:0.4	0:00:0.11	0:00:0.4	0:00:0.17	0:00:0.14	0:00:0.17
#Total time	14:42:44	4:58:34	4:38:38	2:56:00	0:42:30	0:23:03
Count	1724	1327	685	4646	1912	492

(二)評估指標

Fearless Steps Challenge 比賽規則如下，音檔實際判斷正確時間長度單位為 10ms，在參考答案當中只有偵測到和參考答案範圍內一樣才給予得分如圖五，若判斷超出參考得分範圍則不扣分也不予計分只計算真實得分數，每個得分區域將計算每 10ms 幀的真實相同答案的數值（標籤上的最低分辨率）。

圖五、得分範圍參考圖

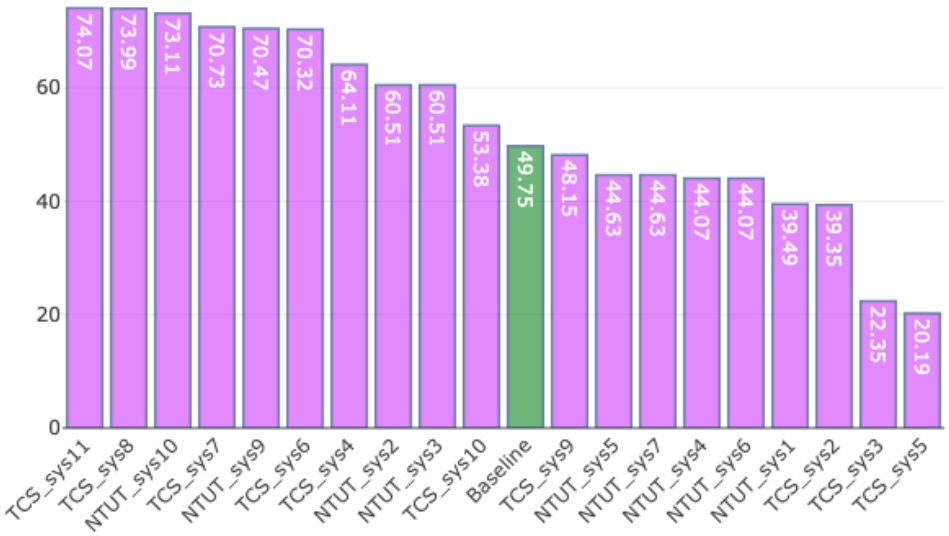
評分公式如下，$tatal\ TP\ time$ 為 System Detected 的真實得分的總時間和，$annotated\ total\ speech\ time$ 為 Reference annotation 的參考答案時間總和，$tatal\ TP\ time$ 除以 $annotated\ total\ speech\ time$再乘以百分比為最後，Fearless Steps Challenge 比賽排名的參考依據。

$$Acc_{sent} = \frac{total\ TP\ time}{annotated\ total\ speech\ time}$$

五、實驗語競賽結果

本實驗使用 Fearless Steps Challenge 資料庫行訓練，Fearless Steps Challenge 資料庫分成了 Train data。本實驗將 Train data 資料庫的每份音檔利用 Sliding Windows 的方式切出訓練資料，窗口大小為 2s 每次位移 10ms 進行 Train data 的資料採樣。以下先單獨對各個部分進行實驗，分為 CNN 架構的音頻部分和 BERT 文字部分別進行討論，再討論多模式情緒偵測模型。

此外，圖六為我們提交至 Fearless Steps Challenge 官方，經過官方評測後的成績排名結果，我們總共提交了 10 個不同設定的系統，在以下實驗中會逐步說明。



圖六、Fearless Steps Challenge 官方排名總表

實驗一，聲學與文字模式情緒偵測

1. 聲學 CNN 模型

多模式模型音頻前及處理部分單獨進行討論， Fearless Steps Challenge 的答案共分為三種 positive、neutral、negative，而在評估指標內還有包含 Non-Sentiment 的部分，因此在訓練同時將測試集 Non-Sentiment 的部分使用 Sliding Windows 進行數據採集。

在音頻測試中可以看到，因資料庫音檔雜訊過多且在大部分音檔當中的對話情緒起伏並不明顯，所以造成 positive、negative 的準確率偏低，但在 neutral、silence 的部分以圖七混淆矩陣來看 silence 的準確率最高，因此在單音頻測試模型下成效較為顯著，但

在 neutral 的判斷還是有部分些許不準確。

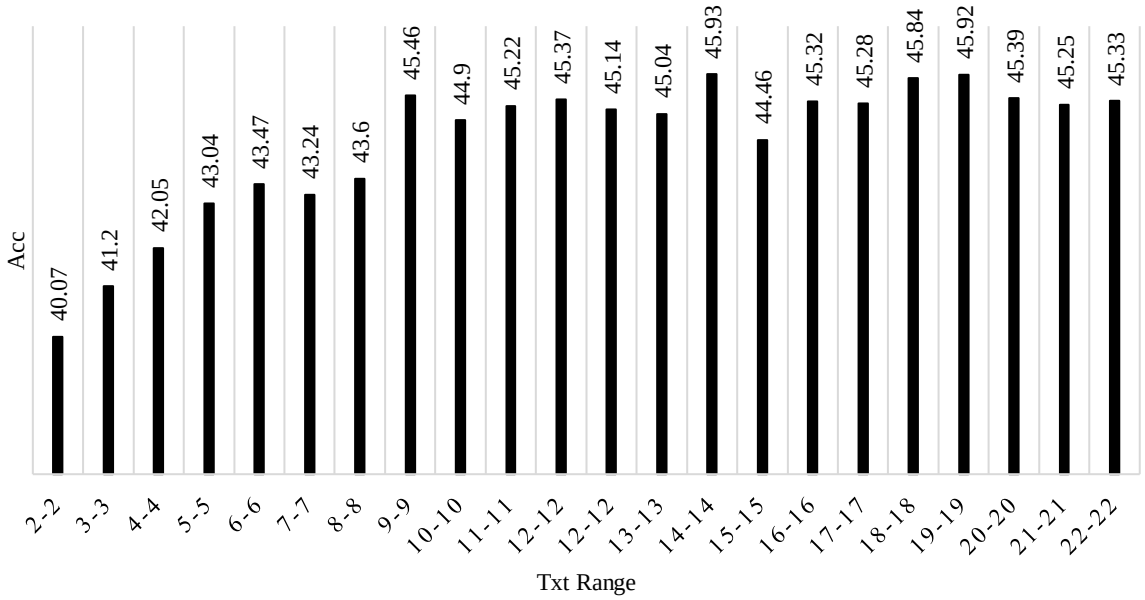
本論文將音頻單獨使用 CNN 神經網路模型進行單獨資料庫訓練，模型如圖三，在 Fearless Steps Challenge 官方網站準確率為 44.07%參考排名如表五，因此單音頻測試對於 silence 和 neutral 偵測有一定的準確度，但離最高準確率還是需要靠文字的輔助下達成。

2. 文字 BERT

由於 Fearless Steps Challenge 測試集並無提供音頻的文字，因此在使用測試集辨識時將音頻使用語音辨識(Automatic Speech Recognition, ASR)進行辨識，但語音辨識只能獲得單詞起始時間和結束時間，所以在文字預處理本論文使用 Sliding Windows 方式，在要辨識單詞時間底下往前往後取一定範圍的單詞量組成句向量，輸入如圖四所示 BERT 模型內進行單文字測試。

在測試文字中發現，文字採樣範圍不同時會有不同準確率，當採樣文字採樣範圍達到往前往後 14 個字時之後準確率趨近於穩定，如表四所示，在採樣範圍從 2 至 8 個字時明顯採樣特徵不足因此造成準確率沒有明顯提升，因此將採樣範圍提升從 9 至 22 個字進行測試，在 8 至 9 個字時準確率有明顯提升，由此實驗可證實當文字採樣範圍會對於情緒識別準確率有一定的成效。

表四、BERT 模型各種文字採樣範圍正確率



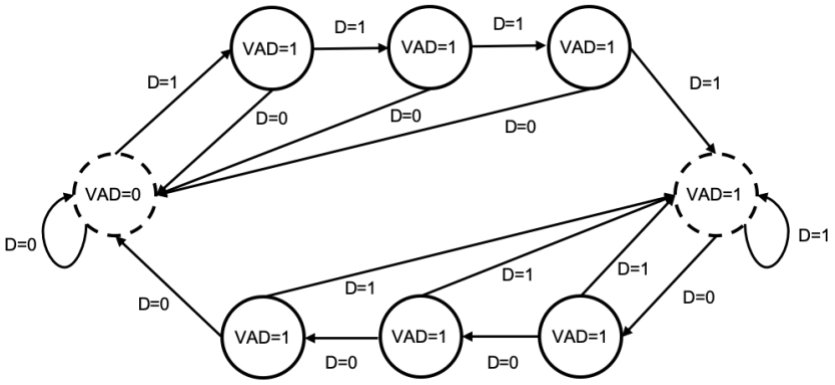
本論文將文字單獨使用 BERT 神經網路模型進行單獨資料庫訓練，模型如圖四，在 Fearless Steps Challenge 官方網站準確率為 44.63%參考排名如表三，因此使用單文字識別情緒時如果有相關情緒字眼出現時則會對準確率照成一定的影響，但某些場景下無法純粹依靠單文字測試，因此本論文使用多模式神經網路模型將兩者模型混合。

實驗二，多模式情緒偵測

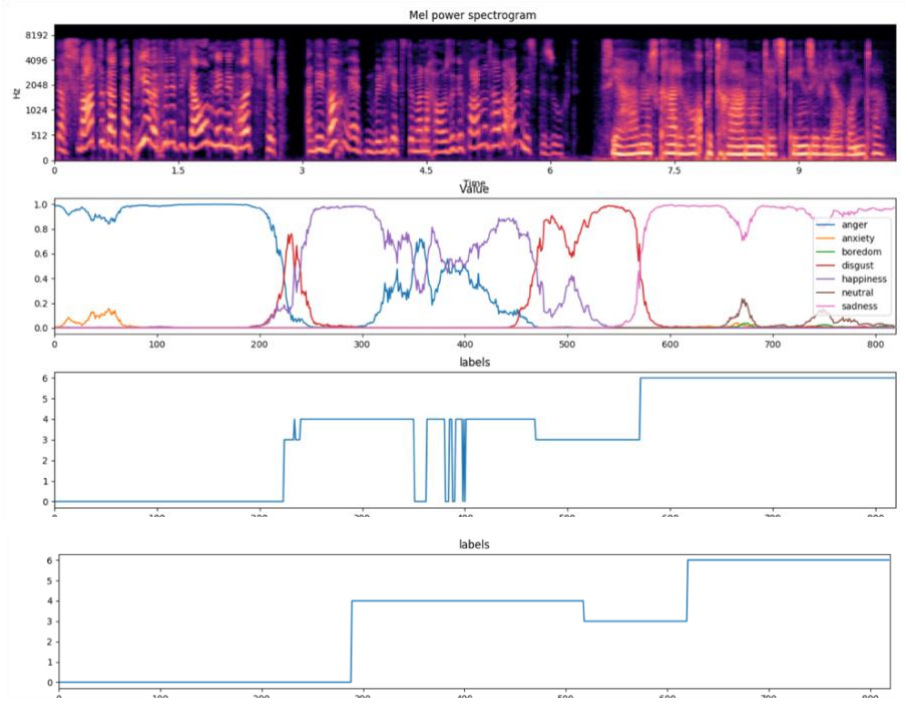
在多模式實驗當中本論文使用音頻模型和文字模型進行情緒識別，交叉測試發現音頻測試當中發現 positive 和 negative 的情緒類別較為不準確，文字測試部分也發現文字採樣範圍對於實驗結果有一定影響，音頻無法識別的 positive 和 negative 利用多模式模型，由文字模型來識別 positive 和 negative 的相關情緒字眼，以利於提升模型準確率，音頻測試當中 silence 和 neutral 的準確度也有一定成效，因此也可輔助多模型識別 Non-Sentiment 的切割位置準確度和 neutral 的正確率，所以本論文使用混合神經網路模型架構來提升模型準確度。

資料庫分為三類 negative，neutral，positive，進行這三類的辨別。因情緒變化動態較慢而我們所使用的 Sliding Windows 的辨識方式讓結果輸出的變化太大，所以在連續辨認時設置了 state machine 的輸出機制，在連續輸出一定數量的答案才會確定輸出否則會繼續輸出前一個答案。例如：圖八在未使用 state machine 時輸出答案不穩定會一直跳動，但在加上 state machine 後可以看到答案輸出趨近於穩定。

state machine 狀態圖如圖七，預設輸出為 0 當 D 連續輸出三次轉態為 1 時 VAD 才會判斷輸出為 1，當D轉態出現中斷或是小於 3 次時回道原始狀態的 VAD 值，反之則將狀態轉為轉態數值。也就是說，當輸出第二次出現不一樣的數值時先放入暫存器，然而繼續輸出相同數值，直到連續得到相同轉態數值，才確定轉態。這可以使本論文模型輸出趨近於穩定。



圖七、state machine 狀態圖



圖八、state machine 前後比較

在 state machine 的幫助下，本實驗使用圖五的多模式神經網路架構，將文字以及聲音使用 Sliding Windows 的方式切出訓練資料，窗口大小為 2s 每次位移 10ms 進行訓練。在 128 訓練次數後，正確率達到 60.51%，

在第二次實驗下，修改 state machine 的暫存器個數來讓情緒浮動的範圍不會變化的太快，將 state machine 暫存器修改為 15 個最本論文最高正確率，正確率來到了 73.11% 為 Fearless Steps Challenge 比賽中 Sentiment Detection 項目的 Rank 3 排名，

實驗三：提交至 Fearless Steps Challenge 官方之系統差異說明

以下分別說明在 Fearless Steps Challenge 官方網站總排名中，不同 NTUT_sys 系統的做法與設定差異：

1. NTUT_sys1 使用 google 語音辨識將音檔切割為 15 秒一個單位音檔不重疊，進行單音頻測試神經網路模型如圖三，Fearless Steps Challenge 官方正確率為 39.49%
2. NTUT_sys2 使用 google 語音辨識將音檔切割為 15 秒一個單位音檔不重疊，進行多模試神經網路模型如圖一，Fearless Steps Challenge 官方正確率為 60.51%
3. NTUT_sys3 為 NTUT_sys2 的重複卻認正確率，因此在回傳一次給 Fearless Steps Challenge 官方卻認正確率為 60.51%
4. NTUT_sys4 使用 Sliding Windows 的方式進行驗證將音框設為 2s 位移時間為 10ms，進行單音頻測試神經網路模型如圖三，Fearless Steps Challenge 官方正確率為 44.07%
5. NTUT_sys5 使用 Sliding Windows 的方式進行驗證將文字採樣範圍調整為前後 14 個字，進行單文字測試神經網路模型如圖四，Fearless Steps Challenge 官方正確率為 44.63%
6. NTUT_sys6 為 NTUT_ sys4 的重複卻認正確率，因此在回傳一次給 Fearless Steps Challenge 官方卻認正確率為 44.07%
7. NTUT_sys7 為 NTUT_ sys5 的重複卻認正確率，因此在回傳一次給 Fearless Steps Challenge 官方卻認正確率為 44.63%
8. NTUT_sys8 使用 Fearless Steps Challenge Train data 所算出的答案進行回傳，因此不列在官方排名中
9. NTUT_sys9 使用 Sliding Windows 的方式進行驗證，多模式神經網路進行識別，state machine 暫存器設為 3 個，Fearless Steps Challenge 官方正確率為 70.47%
10. NTUT_sys10 使用 Sliding Windows 的方式進行驗證，多模式神經網路進行識別，state machine 暫存器設為 15 個，Fearless Steps Challenge 官方正確率為 73.11%

六、結論

在本論文中，我們提出了基於 CNN 與 BERT 的多模式情緒識別神經網路架構，融合聲學與語意情緒特徵參數，用以偵測語音訊號中傳達的情緒狀態，以強化系統的情緒

狀態偵測效能。並以 state machine 減緩輸出跳動的情況，有效的解決輸出時產生的不穩定性，提升準確度。最後，由正式比賽結果發現，我們的系統的情緒狀態偵測正確率達到 73.11%，在所有隊伍提交中的 20 個結果中，排第三名，不但超越主辦單位提供的基準參考系統（49.75%），並只差第一名（74.07）不到 1%。

Acknowledgements

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and partly by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 107-2911-I-027-501, 108-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067 and partly supported by Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan contract No. TL-108-D301.

參考文獻

[1]. Y. Wang, L. Guan, An investigation of speech-based human emotion recognition, pp. 15-18, 2004.

[2]. Z. Huang，M。Dong，Q。Mao，Y。Zhan，Speech Emotion Recognition Using CNN，pp.801-804,2014。

[3]. Y. Wang, L. Guan, "Recognizing human emotional state from audiovisual signals", IEEE Trans. Multimedia, vol. 10, no. 5, pp. 936-946, Aug. 2008.

[4]. Z. Zeng, J. Tu, B. M. Pianfetti, T. S. Huang, "Audio–visual affective expression recognition through multistream fused HMM", IEEE Trans. Multimedia, vol. 10, no. 4, pp. 570-577, Jun. 2008.

[5]. M. Mansoorizadeh, N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech", Multimedia Tools Appl., vol. 49, no. 2, pp. 277-297, 2010.

[6]. M. Glodek et al., "Multiple classifier systems for the classification of audio-visual emotional states" in Affective Computing and Intelligent Interaction, Berlin, Germany:Springer, vol. 6975, pp. 359-368, 2011.

[7]. M. Soleymani, M. Pantic, T. Pun, "Multimodal emotion recognition in response to videos",

IEEE Trans. Affect. Comput., vol. 3, no. 2, pp. 211-223, Apr./Jun. 2012.

[8]. J.-C. Lin, C.-H. Wu, W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition", IEEE Trans. Multimedia, vol. 14, no. 1, pp. 142-156, Feb. 2012.

[9]. J. Wagner, E. Andre, F. Lingenfelser, J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data", IEEE Trans. Affect. Comput., vol. 2, no. 4, pp. 206-218, Oct. 2011.

[10]. A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification", IEEE Trans. Affect. Comput., vol. 3, no. 2, pp. 184-198, Apr./Jun. 2012.

[11]. D. Gharavian, M. Bejani, M. Sheikhan, "Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks", Multimedia Tools Appl., vol. 76, no. 2, pp. 2331-2352, 2017.

[12]. S. Zhalehpour, O. Onder, Z. Akhtar, C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states", IEEE Trans. Affect. Comput..

[13]. R. R. Sarvestani, R. Boostani, "FF-SKPCCA: Kernel probabilistic canonical correlation analysis", Appl. Intell., vol. 46, no. 2, pp. 438-454, 2017.

[14]. M. Bejani, D. Gharavian, N. M. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks", Neural Comput. Appl., vol. 24, no. 2, pp. 399-412, 2014.

[15]. Y. Wang, L. Guan, "Recognizing human emotional state from audiovisual signals", IEEE Trans. Multimedia, vol. 10, no. 5, pp. 936-946, Aug. 2008.

[16]. M. Mansoorizadeh, N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech", Multimedia Tools Appl., vol. 49, no. 2, pp. 277-297, 2010.

[17]. Y. Wang, L. Guan, A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition", IEEE Trans. Multimedia, vol. 14, no. 3, pp. 597-607, Jun. 2012.

[18]. B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations", Proc. 9th Int. Conf. Multimodal Interfaces (ICMI), pp. 30-37, 2007.

[19]. C. Busso et al., "Analysis of emotion recognition using facial expressions speech and multimodal information", Proc. 6th Int. Conf. Multimodal Interfaces (ICMI), pp. 205-211, 2004.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 152-166

適合漸凍人使用之語音轉換系統初步研究

Deep Neural-Network Bandwidth Extension and Denoising Voice Conversion System for ALS Patients

黃百弘 Bai-Hong Huang, 廖元甫 Yuan-Fu Liao

國立臺北科技大學電子工程系

Department of Electronic Engineering, National Taipei University of Technology

tjtkng@gmail.com, yfliao@ntut.edu.tw

Matúš Pleva, Daniel Hládek

Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia

matus.pleva@tuke.sk, daniel.hladek@tuke.sk

摘要

漸凍人症（肌萎縮性脊隨側索硬化症，Amyotrophic lateral sclerosis， ALS）為一種神經退化性疾病，這種疾病目前還沒有治癒的方法，並會讓漸凍人慢慢失去說話能力，最終導致無法利用語音與人溝通，而失去自我認同。因此，我們需要為漸凍人建立適合其使用之語音溝通輔具（voice output communication aids, VOCAs），尤其是讓其能具有個人化的合成語音，即病友發病前的聲音，以保持自我。但大部分在 ALS 後期，已經不能講話的病友，都沒有事先妥善保存好個人的錄音，最多只能找出有少量大約 20 分鐘的低品質語音，例如經過失真壓縮（MP3）、只保留低頻寬（8 kHz），或是具有強烈背景雜訊干擾等等，以致無法建構出適合 ALS 病友使用的個人化語音合成系統。針對以上困難，本論文嘗試使用通用語音合成系統搭配語音轉換演算法，並在前級加上語音雜訊消除（speech denoising），後級輔以超展頻模組（speech super-resolution）。以能容忍有背景雜訊的錄音，並能將低頻寬的合成語音加上高頻成分（16 kHz）。以盡量能從低品質語音，重建出接近 ALS 病友原音的高品質合成聲音。其中，speech denoising 使用 WaveNet，speech super-resolution 則利用 U-Net 架構。並先以 20 小時的高品質（棚內錄音）教育電台語料庫，模擬出成對的高雜訊與乾淨語音語句，或是低頻寬與高頻寬語音，

分別訓練 WaveNet 與 U-Net 模型，再用以處理病友的低品質語音錄音音檔。實驗結果顯示，訓練出來的 WaveNet 與 U-Net 模型，可以相當程度還原具雜訊或是低頻寬的教育電台語音檔。並能用來替 ALS 病友重建出高品質的個人化合成聲音。

關鍵詞：類神經網路、ALS、WaveNet

一、緒論

漸凍症全名，肌萎縮性脊隨側索硬化症(Amyotrophic lateral sclerosis， ALS)，為一種逐漸且要命的神經退化性疾病，ALS 患者因中樞神經系統內控制骨骼肌的神經元退化使大腦逐漸完全喪失控制肌肉運動的能力，病程晚期會影響語言能力、進食、和呼吸系統的運行，但此疾病並不見得會影響 ALS 患者的思考能力；相反，ALS 患者晚期依然維持完整的思緒、具有發病前的人格、智力及記憶，所以病友即便影響到聲帶無法發出聲音，但病友的思緒依然是然全不受影響的，在病友無法自身發出聲音與他人溝通的情況下就需要用到音溝通輔具（voice output communication aids, VOCAs），目前 VOCAs 常用的方法是使用文字轉語音(Text To Speech,TTS)來作為病友語音輸出，病友經過前端裝置輸入文字再由 TTS 轉換成聲音，即便目前 TTS 已能模擬出非常接近人類的聲音，卻無法還原出病友在未發病時的聲音特色，滿足病友想用發病前聲音說話的自我認同需求，且病友家屬也強烈表示希望病友能用具有發病前個性的聲音與他們溝通。

如果要為病友建立個人的 TTS 需要病友在擁有大量的高品質音檔下才能完成，在一般的情況下聲音的資料是非常容易取得的，只要將說過的話錄下來便能成為語料，但這件看似簡單的事情對於晚期的漸凍病友卻不是如此，在病情發展到影響聲帶時，病友將失去語言能力，或在配帶呼吸器的情況下要正常發音是非常困難的事情，所以只能仰賴平日錄製的語音，但漸凍病友的病期難以推敲，且每位病友受病情影響的部位無法預測，所以有預前錄製音檔的病友相當稀少，即便有數量也非常少。

在病友先前自行錄製中我們發現，病友家屬將以手邊最容易取得之錄音系統錄製音檔，如手機或錄音筆，所以錄音語料面臨兩大問題，第一環境音雜音大無法正確分析語音特徵，在實驗中有一位病友家屬提供我們的音檔為測錄音檔，冷氣的低頻聲與環境回音發常嚴重甚至已經超過病友的聲音，聲碼器在提取聲音特徵時大受影響，完全無法有效的提取特徵參數，使得後面的訓練一蹋糊塗，轉換出來的聲音富含嚴重雜音，根本無法辨識為病友聲音，但病友無法再提供其他錄音檔，只有嚴重雜訊音檔為病友轉換音

檔，第二病友語料取樣頻率不足高頻完全消失，情況是病友家屬有事先幫病友保存語音，病友提供我們一小時的語音，但經過處理後只剩下 16 分鐘可用音檔，音檔為 AMR 破壞性壓縮，並且取樣頻率只有 8kHZ，聲音輸出少了高頻聲音顯得不真實，無法完整還原病友發病前聲音，所以本論文將傳統語音轉換系統前級增加 Speech Denosing[1]與在後級再以 Super-Resolution[2]來改善，如圖一所示。

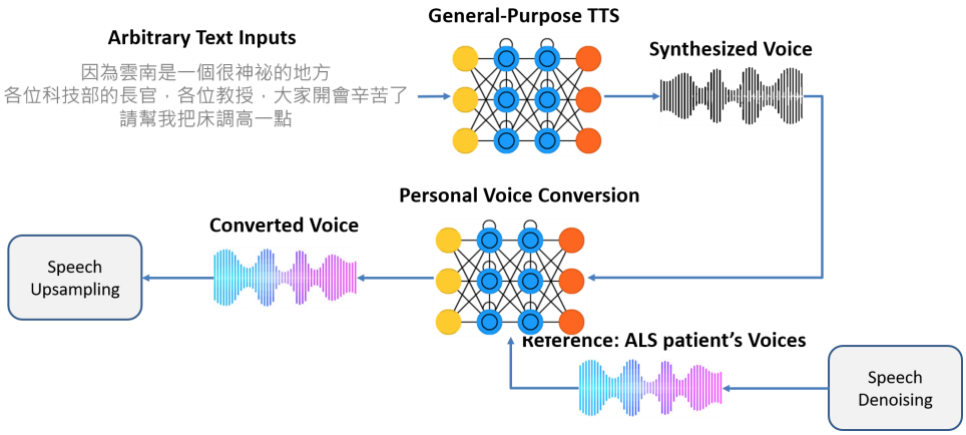


圖 一、ALS 病友語音轉換系統圖

本論文將病友語音輸入前級加上 Speech Denosing 確保病友音檔輸入為無雜訊干擾語音，在轉換後音檔後級加上 Super-Resolution 確保病友轉換音檔輸出為具有高頻的較高品質音檔，期望經過這套系統不僅能確保病友已剩不多的語音，保持輸入訓練音檔品質，也確保輸出為病友具有高頻還原語音。

二、基於類神經網路病友語音轉換系統

本論文提出的 ALS 語音轉換系統如圖 1，輸入語音使用微軟釋出的 Hanhan TTS(Text To Speech)[3]，將病友語音經過前級 Speech Denosing 系統之後作為目標語者(病友)之語音，建立個人語音轉換系統，最後經過後級 Super-Resolution 系統將病友高頻補足。

(一) ALS 語音轉換系統

1 bi-LSTM 語音轉換系統架構

Bi-LSTM 語音轉換系統架構，如圖二，先將 TTS 來源語者經由 Vocoder 取出語者特

徵，在經過 bi-LSTM 模型進行學習，轉換成病友語音特徵，最後經由 Vocder 將特徵組合回來，輸出病友轉換語音。

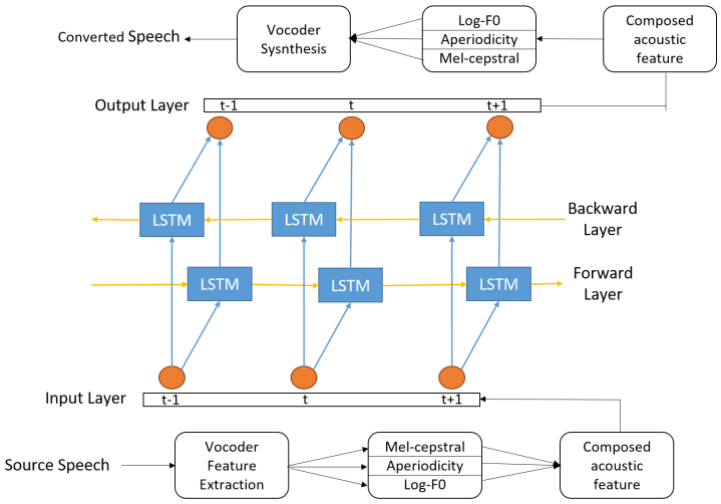


圖 二、BLSTM 語音轉換系統架構

2 實現方法

本論文語音轉換系統使用平行語料架構，系統第一步是要擷取來源與目標語者每一個音框的特徵參數，再利用 Dynamic Time Warping（DTW)[4]進行音框對齊。在語音轉換的過程中，首先透過 Vocoder 來分析音檔裡面的資訊，分別包含了下列三個重要的特徵 Mel-cepstral[5]、Band Aperiodicity、Log-F0，其中 Mel-cepstral 是聲音的頻譜，Band Aperiodicity 是非週期性的特徵用來判斷聲音是否發聲，最後 Log-F0 決定聲音的音高，這些特徵參數將作為模型的輸入進行訓練，在訓練之前因為來源語者與目標與者說話的速度不會相同，Dynamic Time Warping（DTW)解決來源與目標音框長度不同的問題。再以 DTW 的 alignment 結果，計算每一個來源與目標對應音框，應該收縮（shrinking）、拉長（stretching）、還是保持不變（kept）如下圖三所示。

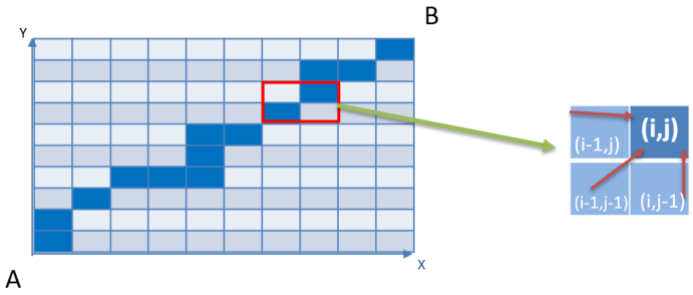


圖 三、 Dynamic Time Warping(DTW)演算法示意圖

接下來將 DTW 對其資料丟入 bi-LSTM 模型中訓練，LSTM 網路是一種特殊的 RNN 結構，可以描述時變的語音訊號，並且把之前的資訊帶到當前的訓練任務中，，LSTM 的結構能夠用來防止長距離依賴問題，也就是可以解決梯度消失的問題，bi-LSTM 是將 LSTM 與 BRNN 結合在一起，這種方法可以在輸入的方向或的長時的上下文信息，效果優於 LSTM，經過系統訓練後會產生病友的個人語音模型，系統後級會將目標語者之 Mel-cepstral、Band Aperiodicity、Log-F0，重新經過 Vocder 組合，將病友轉換語音輸出，最後我們就可以達到使用來源語者說任何語句都能轉換成具有病友語音個性的病友專屬語音轉換系統。

(二) 強化病友轉換語音系統

在摘要我們提到 ALS 病友後期已經無法在錄音，但病友卻都沒有妥善的保存聲音，在處理病友音檔時遇到兩大問題，第一環境音雜音大無法正確分析語音特徵，第二病友語料取樣頻率不足高頻完全消失，在病友已經無法在錄製音檔的情況下，我們希望可以最大限度使用病友已所剩不多的音檔，所以我們將這兩套系統加入傳統的轉換系統中。

1 Speech Denosing(消雜訊系統)

語音消雜(Speech Denosing)的問題上，大多數用於語音消雜的技術使用頻譜圖作為前端 [6,7,8,]。然而，這種做法帶來了潛在丟棄的缺點 有價值的信息（階段）和利用通用特徵提取器（頻譜圖分析）而不是學習給定數據分佈的特定特徵表示，最近，神經網絡已經證明在處理離散化音頻信號的樣本之間的結構化時間依賴性方面是有效的，有趣的是， 大多數這些生成模型都是自回歸模型[9,10]，WaveNet[10]是在自然語音上被廣泛利用得自回歸模型，本論文使用 WaveNet 模型建置 Speech Denosing，目標為最大保留已剩不多的病友語音。

(1)系統架構

圖四為 Speech denoising 系統架構圖，將病友雜訊聲音經過系統消雜後，可以輸出較為乾淨的病友語音，目標為保留病友乾淨語音供後端轉換系統使用。

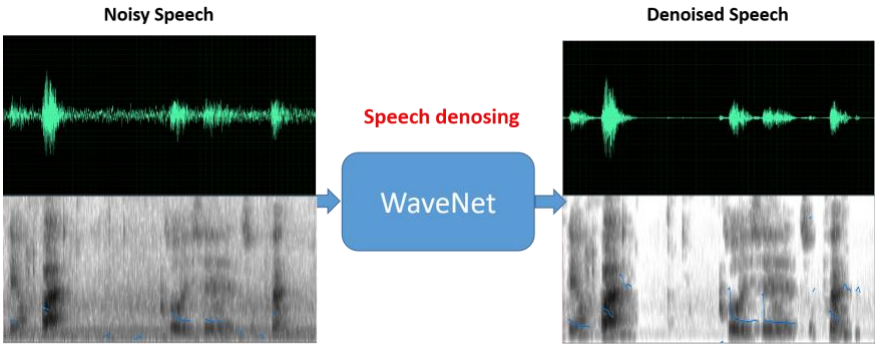
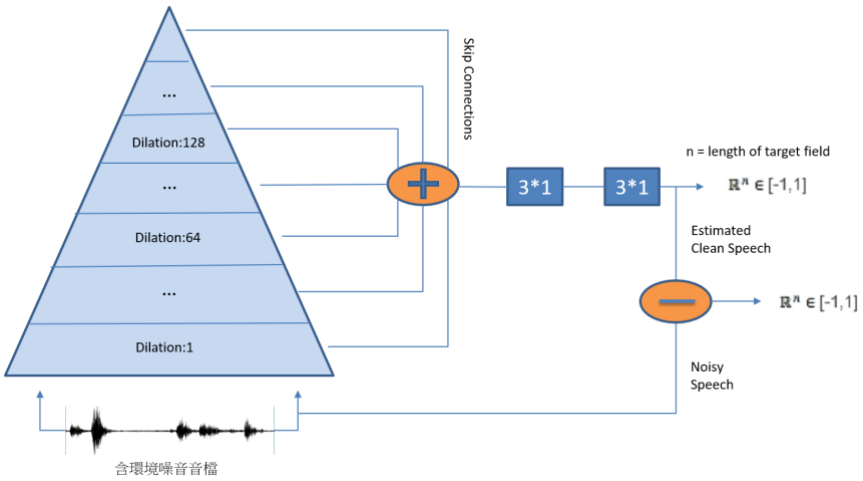


圖 四、Speech denoising 系統架構圖

(2) 實現方法

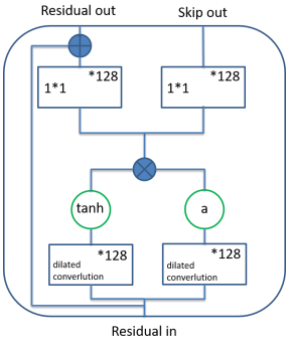
問題表述如下：mt = st + bt，其中：mt＝混合信號，st＝語音信號，bt＝背景噪聲信號。 目標是估計給定的 mt，本論文使用 FaNT[18]將雜音添加入乾淨音檔，所以將混雜噪音的音檔當做 noisy voice，將乾淨的音檔作為 clean voice，系統先經過對齊確定兩個音檔長度是否一致，再以 clean voice 減去 noisy voice 就可以得到我們事先添加的雜訊，將得到資料丟入 Wavenet 架構下訓練，WaveNet 能夠合成自然的發聲語音。 這種自回歸模型的形狀給出先前樣本的一些片段的下一個樣本的概率分佈，本論文的模型的描述在圖五中，



圖五、WaveNet 模型示意圖

本論文所提出的模型具有 30 個殘餘層如圖六，每層中的膨脹因子 以 2 為倍數增加 1，2，...，256,512。該模式重複 3 次（3 個堆疊）。 在第一次擴張卷積之前，1 通道輸入被線性投影到 128 個通道 標準的 3x1 卷積，以符合每個殘留層中的濾波器數量。跳

過 連接是 1x1 卷積，還有 128 個濾波器，在匯總所有後應用 RELU 跳過連接。最後兩個 3x1 卷積層未擴張，包含 2048 和 256 個濾波器，分別由 RELU 分隔。輸出層將要素圖線性投影到 a 使用 1x1 濾波器的單通道時間信號，該參數化導致感受野 共 6,139 個樣本（≈384ms），目標字段由 1601 個樣本（≈100ms）組成。



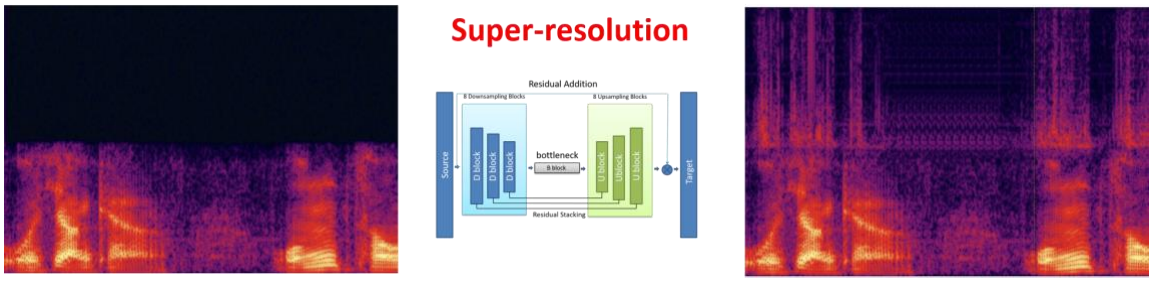
圖六、剩餘層示意圖

2 Super-Resolution(展頻系統)

超分辨率(Super-Resolution)廣泛被使用來解決許多應用中的低分辨率問題[13]，這些技術重建或學習消失的高頻信息以增強成像系統的分辨率。最近，這些技術已被有效地用於提高分辨率，主要是為了提高生物識別系統的識別性能包括 face [14]和 iris [15]，但多半是圖像上的利用，在音頻上一樣可以使用超分辨率(Super-Resolution)來回復消失的高頻音頻，本論文使用 U-Net[16]架構來完成超分辨率(Super-Resolution)系統建置，目標為有效的恢復病友高頻的聲音。

(1) 系統架構

本系統將病友轉換語音輸入為 8kHZ 的低頻音檔，經過 Super-Resolution 系統後可以展頻出 16kHZ 的具有高頻的音檔，圖七為展頻系統架構圖，目標為系統可以有效的還原出病友高頻音檔。



圖七、展頻系統架構圖

(2)實現方法

模型架構是 U-Net，使用虛擬的一維子像素卷積 Sub-Pixel Convolutions[17]，圖八為 U-Net 架構圖，系統經過左側八個下採樣過程分八組捲積來執行，每組捲積後進行 maxpool，將音檔進一步縮小為原本的二分之一，通過八次操作將 39998*8 語音計算成 154*2048，右側的採樣程，使用 8 組反捲積，每次上採樣將語音擴展成 2 倍然後將對應層的語音進行剪輯與複製，然後 concat 到上捲積結果上，完成上採樣後得到 40000*16，最後使用 1*1 捲積合將通道數減為 2 來代替捲積層。

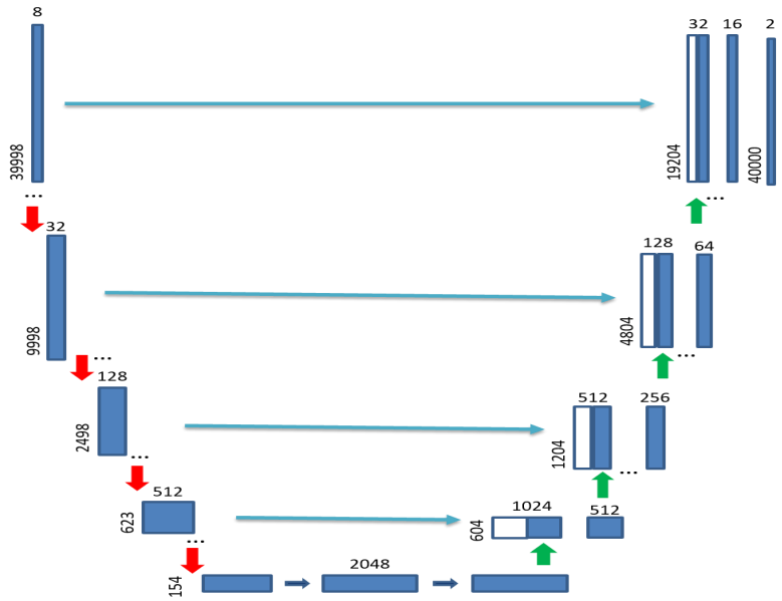


圖 八、 U-net 架構圖

系統將數據經過採樣剪輯並分批丟入模型中以更新其權重。保存驗證分數最低的模型儲存為最佳模型，經過每個下取樣波形(Downsampled waveform)通過八個下取樣塊(Downsampling blocks)發送，經由瓶頸層(Bottleneck)連接到八個上取樣塊(Upsampling blocks)，瓶頸層跟下取樣塊之間有殘差連結(Residual connections)，這些殘差連結彼此

共享特徵資訊。上取樣塊使用子像素卷積，沿一個維度重新排序信息以擴展其他維度，最後由最終一層卷積層將重新排序堆疊後的模塊與殘差加成(Residual addtion)加在原始輸入音檔中產出升頻後得音檔。

三、強化病友合成語音實驗

強化病友合成語音實驗中，目的是為了使病友合成語音更接近病友發病前聲音，本文將比較合成語音經過 Super-Resolution 與 Speech Denosing 系統強化前後是否有效改善語料高頻不足與雜音干擾等問題，在實驗中會比較上述兩個系統的主觀客觀偏好，來評斷結果。

(一) 訓練與測試語料

本文建立展頻系統與消雜訊系統皆使用教育電台廣播節目播出時所錄製的音檔，我們從中挑出其中的 7 個節目共 20 小時，並經過人工剪輯成只有人聲的音檔當作 Data 使用，表 1 為訓練資料統計表，表 2 為測試資料統計表。

表一、訓練資料統計表

類型	節目名稱	集數	音檔總時間(minute)
純人聲	創設市集	6	155
	國際教育心動線	7	185
	技職最前線	5	105
	青農市集 On Air	20	457
	晨間新聞	10	258

表二、 測試資料統計表

類型	節目名稱	集數	音檔總時間(minute)
純人聲	教官不說教	5	92
	兒童新聞	4	60

(二) 實驗設定 - Speech Denosing 消雜訊系統

消雜訊系統實驗中，本文先將音檔消完雜訊完後再輸入 ALS 語音合成系統中重新訓練模型，為滿足大多是環境音條件，本文模擬各種病友在自行錄音時可能會遇到的環境雜訊如表 3。

表三、添加雜訊表

雜訊名稱	音檔時數(hr)
環境有人走動聲響(運動場)	20hr
環境有他人說話雜訊(餐廳)	20hr
音檔總時數	40hr

(三) 實驗設定 - Super-Resolution 展頻系統

展頻系統實驗中，我們將教育電台語料處理成低頻音檔 8kHZ 取樣頻率和高頻音檔 16kHZ 取樣頻率，作為 Data 如圖九，十。

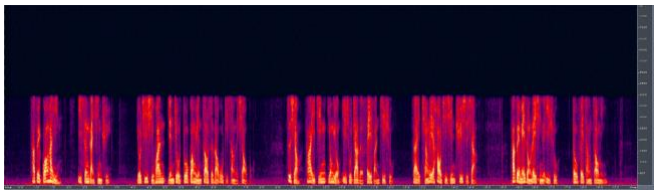


圖 九、 8kHZ 音檔頻譜圖

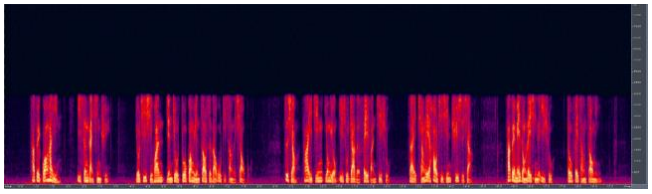


圖 十、 16kHZ 音檔頻譜圖

(四) 評估方法

1 主觀分數評估

本文系統偏好的評估方式，將測試音檔給 5 位包含病友家屬及母語為中文的人員進

行評分，系統偏好度測試是 2 選 1 的方式，為標準的 A/B/X 測試，系統評分採用平均主觀值分數(mean opinion score , MOS)進行評估，分為可理解度評分、相似度評分和自然度評分，評分方式為 1~5 分，分數越高則為越好。

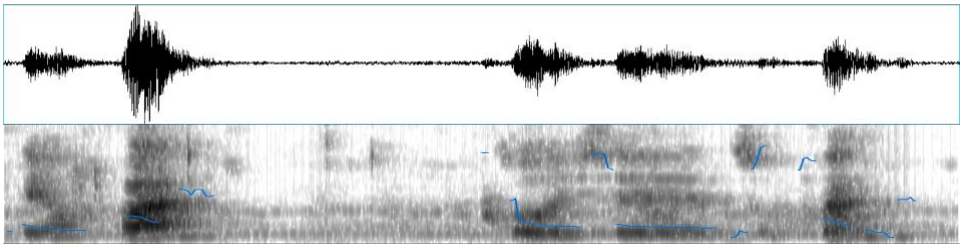
2 客觀分數評估

客觀評估方式比較聲音的參數在訓練後與目標值的差異，客觀評估的分數直接對應到轉換的效果程度好壞評估的內容包含了 Mel-cepstral distortion (MCD): 均方根誤差單位 dB、BAP: 均方根誤差 單位 dB、F0-RMSE: 均方根誤差 單位 Hz 以及 VUV: %(以百分比表示)。

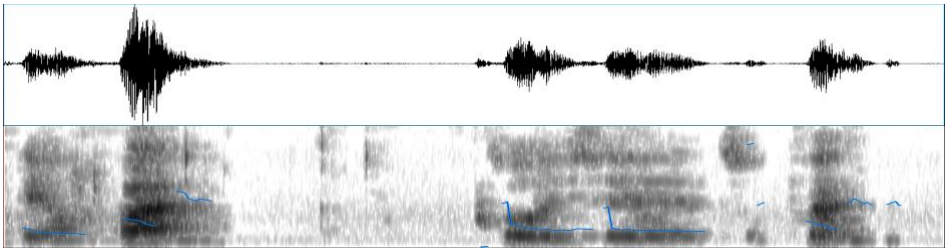
四、實驗結果

Speech Denosing 強化病友合成語音實驗

由圖十一，十二可發現，病友音檔經過消雜訊系統後，低頻雜音已明顯濾除，音檔 F0 部分也可以穩定的求出，表四為未經消雜與消雜與音合成系統客觀評估表，數據顯示 MCD、BAP、F0-RMSE、VUV 的錯誤率明顯降低。經過消雜訊系統後合成聲音在 A/B/X 偏好評估或 MOS 主觀評估中都優越於無消雜續系統，結果如下圖與表



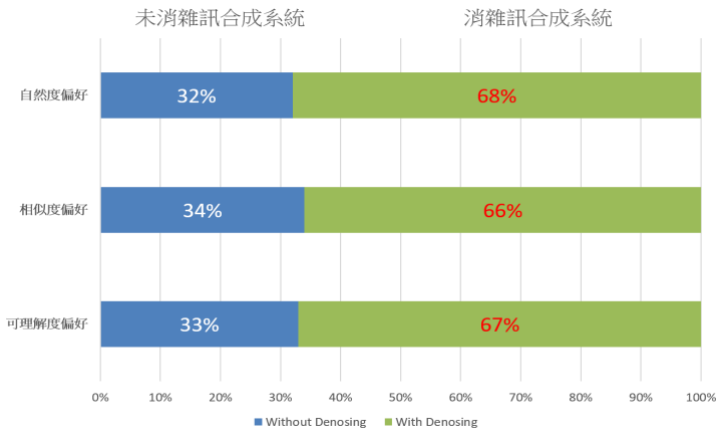
圖十一、未消音音檔圖



圖十二、已消雜訊音檔圖

表四、消雜訊系統客觀評估表

	未消雜與音合成系統	消雜語音合成系統
MCD	9.265 dB	6.869 dB
BAP	0.542 dB	0.208 dB
F0-RMSE	68.597 Hz	49.443 Hz
VUV	62.851%	25.935%



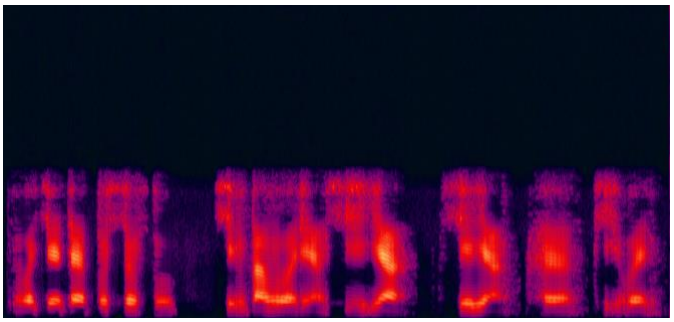
圖十三、消雜尋合成系統 A/B/X 偏好測試圖

表五、消雜訊系統 MOS 主觀評估表

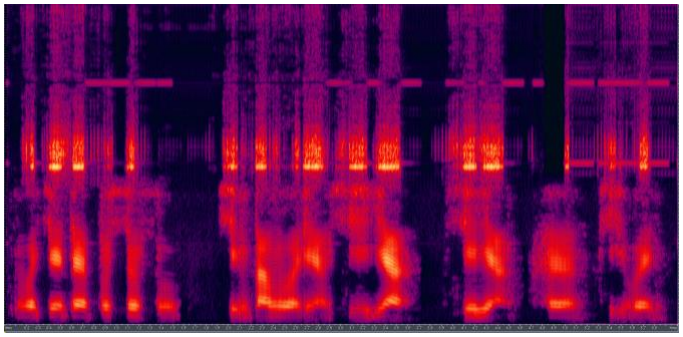
	未消雜訊合成系統	消雜訊合成系統
自然度評分	2	4
相似度評分	2.1	4
可理解度評分	2	4.1

Super-Resolution 強化病友合成語音實驗

經由圖十四，十五比較明顯可以發現，經過展頻系統，系統有相當程度的將病友原先缺少的高頻音域模擬出來。

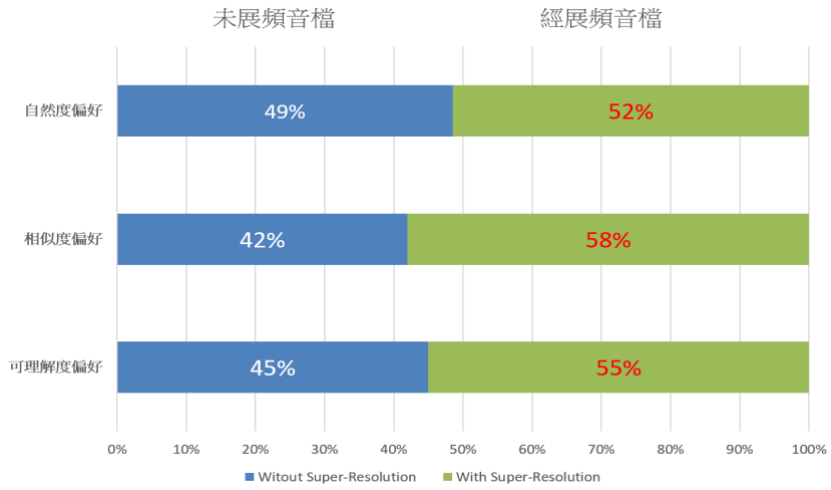


圖十四、未經過展頻系統之病友合成語音頻譜圖



圖十五、經由展頻系統補救之病友合成語音

經由展頻系統之病友合成語音比未展頻合成語音在可理解度、相似度以及自然度偏好效果都比較好，更重要的是無論在 A/B/X 偏好評估或 MOS 主觀評估中，相似度偏好都明顯高過未展頻之音檔。



圖十六、展頻音檔 A/B/X 偏好測試圖

表六、展頻音檔 MOS 主觀評估表

	未展頻音檔	展頻音檔
自然度評分	3.4	3.6
相似度評分	3.1	3.8
可理解度評分	3.3	3.6

五、結論

在實驗中我們成功將病友語音經過 Super-Resolution 系統與 Speech Denosing 系統的強化，不管是相似度或可理解度上都有顯著的提升，病友與病友家屬也給予我們正面的評價。本論文所提出 Super-Resolution 系統與 Speech Denosing 系統僅解決了兩種病友語問題，著重於語料品質上得改善，但事實上在聲音需求上不只漸凍病友需要其他重大傷病病友也有需求，還有許多語音問題需要處理，如:病友還能發聲，但聲音卻不如之前，說話音調與耐力受到影響，是否可以從病友目前語料與發病前語料中找到改善方法，可以從現有資料解決病友需求。

Acknowledgements

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and finally by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 107-2911-I-027-501, 108-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067.

參考文獻

[1]. Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon : AUDIO SUPER-RESOLUTION USING NEURAL NETS ICLR 2017

[2]. Dario Rethage, Jordi Pons, Xavier Serra : A Wavenet for Speech Denoising arXiv:1706.07162v3

[3]. Heiga Zen : Acoustic Modeling for Speech Synthesi Dec. 14th, 2015@ASRU

[4]. Stan Salvador , Philip Chan. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. Dept. of Computer Sciences Florida Institute of Technology Melbourne, FL 32901

[5]. Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition

algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[6]. Anurag Kumar and Dinei Florencio. Speech enhancement in multiple-noise conditions using deep neural networks. arXiv preprint arXiv:1605.02427, 2016.

[7]. Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In Interspeech, pages 436–440, 2013.

[8]. Shahla Parveen and Phil Green. Speech enhancement with missing data techniques using recurrent neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages I–733, 2004.

[9]. Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837, 2016

[10]. Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

[11]. Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759, 2016.

[12]. Kominek, John, Tanja Schultz, and Alan W. Black. "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion." Spoken Languages Technologies for Under-Resourced Languages. 2008

[13]. S.C. Park, M.K. Park, and M.G. Kang, “Super-resolution image reconstruction: a technical overview,” IEEE Signal Processing Magazine, vol. 20, no. 3, pp. 21 – 36, 2003.

[14]. F. Lin, C. Fookes, V. Chandran, and S. Sridharan, “Super-resolved faces for improved face recognition from surveillance video,” in Lecture Notes in Computer Science, Seoul, Korea, Republic of, 2007, vol. 4642 LNCS, pp. 1 – 10.

[15]. Kwang Y.S., Kang R.P., Byung J.K., and Sung J.P., “Super-resolution method based on multiple multi-layer perceptrons for iris recognition,” in Ubiquitous Information Technologies Applications, ICUT ’09. International Conference on, 20-22 2009, pp. 1 –5

[16]. Olaf Ronneberger，Philipp Fischer，Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation arXiv:1505.04597

[17]. Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, Zehan Wang: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network

[18]. Andreas Kitzig: Nieder Rhein University, Room Impulse Response-package: NRU-RIR-package.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 167-181

Bilingual Parallel Sentence Extraction from Comparable Corpora

Chien-Yu Chien, Chin-Hua Chang, Chih-Ping Wei
Department of Information Management
National Taiwan University
r07725026@ntu.edu.tw, r07725031@ntu.edu.tw, cpwei@ntu.edu.tw

Abstract

This research aims to develop a parallel sentence extraction method for automatically extracting parallel sentence pairs from bilingual comparable corpora based on cross-lingual word embeddings. Our task is to effectively identify matched sentence pairs from a Chinese-English corpus with the goal of maximizing F1 score. Our method employs pre-trained, task-specific, and hybrid (a combination of pre-trained and task-specific) monolingual word embeddings to construct a cross-lingual transformation matrix respectively to transform the word embeddings between the two languages, and develops two search strategies (sequential and exhaustive) for parallel sentence extraction. Our empirical evaluation results suggest that task-specific word embeddings (directly trained from a task-relevant corpus, i.e., 25,695 Chinese and English abstracts of theses) outperforms their counterparts. With respect to the two search strategies, our evaluation results suggest that the exhaustive search strategy attains a higher recall rate; the sequential search strategy is more efficient in time. Both strategies achieve a promising performance, with an F1 score up to 60.18%.

1. Introduction

Recently, the tremendous development of neural network techniques for natural language processing has been introduced. Many studies have demonstrated promising results in many important applications, such as neural machine translation [1] and relation extraction [2]. One of the basic requirements for successful neural network model training is a sufficient number of qualified training data examples. In the case of neural machine translation, a bilingual parallel sentence corpus is a required data set. However, many low-resource language pairs (e.g., Chinese- English) or specific domains (e.g., biomedical research) have only limited bilingual parallel sentence corpora, which are not sufficient to support high-performance model construction. Generating parallel sentences by humans is both time consuming and resource intensive. Hence, recent research has focused on how to automatically extract parallel sentences from comparable corpora [3,4,5].

Comparable corpora include non-aligned sentences, phrases or documents that are not an exact translation of each other but share common features such as domain, genre, sampling period, etc. [6]. Compared to parallel corpora, there are more comparable corpora between languages, such as technical documents with bilingual abstracts. Therefore, once the accuracy of the parallel sentence pairs extracted from the comparable corpus can be realized, the problem of lack of a set of parallel sentences as a training corpus for neural machine translation can be effectively relaxed.

Recent research on parallel sentence extraction from comparable corpora has shifted from the feature-based approach [6] to the word-embedding-based approach [3,4], due to the advances on cross-lingual word embedding. Several methods for building cross-lingual word embeddings have been proposed [7,8,9]; among them, a popular method is through transformation matrix [8,9]. Most of existing word-embedding-based parallel sentence extraction methods are conducted on European language corpora (such as English and French). Prior research pays less attention to European-Oriental language pairs (e.g., English and Chinese), which is the focus of our study. Although some studies have investigated Chinese word embeddings [10], these studies are not for bilingual parallel sentence extraction.

Motivated by this research gap, we attempt to propose a word-embedding-based method for extracting parallel sentence pairs from Chinese-English comparable corpora. Our proposed method consists of three stages. First, we train the word embeddings for each language. Specifically, we obtain pre-trained word embeddings from BERT [11] as well as construct, on the basis of a task-relevant corpus, task-specific word embeddings, using the Word2Vec model [12]. Second, we learn a transformation matrix [8,9] to convert word embeddings from one language to another, thus creating cross-lingual word embeddings to align two different embedding spaces. Finally, with the use of the cross-lingual word embeddings, we compare bilingual sentence pairs by calculating their average word-by-word similarity and then extract parallel sentence pairs with a sequential or exhaustive search strategy. Furthermore, observing the phenomenon that an English sentence (segmented by period or question mark) often corresponds to multiple Chinese sentences (segmented by comma, period, or question mark), our proposed method allows many-to-one alignment, mapping multiple Chinese sentences into a single English sentence.

To evaluate the effectiveness of our proposed method, we conduct several experiments. We collect a Chinese-English comparable corpus that consists of 25,695 abstracts of theses. We then randomly selected 100 pairs (the abstracts of theses in both Chinese and English) in the corpus as the testing set. In this parallel sentence corpus, each

comparable document pair contains at least three matched parallel sentence pairs and a number of unmatched sentences. The other 25,595 pairs then serve as the training set for training monolingual word embeddings and constructing cross-lingual transformation matrices for different monolingual word embedding models. Our evaluation results show that our proposed method with task-specific word embeddings and the exhaustive search strategy achieves the highest effectiveness, reaching up to 60.18% in F1 score. The hybrid word embedding model, which combines pre-trained and task-specific word embeddings, is not as effective as the task-specific embedding model. The exhaustive search strategy attains better performance overall, whereas the sequential search strategy achieves a higher precision rate. We also discover the formation (Cbow or Skipgram [12]) of the monolingual word embeddings are sensitive parameters to this extraction task. The remainder of this paper is organized as follows: In Section 2, we describe the design of our proposed parallel sentence extraction method. Subsequently, we detail our evaluation design and discuss important experimental results in Section 3. Finally, Section 4 provides a summary of this study.

2. Our Proposed Method

Our proposed parallel sentence extraction method is to extract bilingual parallel sentence pairs from a comparable corpus. Because aligned documents in a comparable corpus often share similar themes and contents, parallel sentence pairs may exist in these aligned documents [6]. For example, Chinese technical papers or theses typically contain both Chinese and English abstracts, which usually describe the same or highly similar contents in the two languages. A pair of such aligned documents might include sentences that are exact translations or at least share common contents such as subjects, verbs and objectives. These sentence pairs are essential for training a neural machine translation model or for extracting translations for domain-specific terms.

The purpose of this study is to extract all sentence pairs with the same or highly similar content from a set of aligned bilingual documents. The constituent words of a sentence pair are assumed not necessarily consistent with the grammatical order or exact meaning. In order to estimate the similarity of bilingual sentence pairs, we decide to use cross-lingual word embeddings to find out the embedded relations between words and sentences. Accordingly, the research question of this study is formulated as: given a pair of aligned documents written in two different languages, our proposed method is to identify matched sentence pairs in the document pair, with the goal of maximizing the amount of extracted pairs while minimizing the likelihood of extracting wrong pairs.

Our proposed method consists of three stages: monolingual word embedding generation, cross-lingual word embedding generation, and parallel sentence extraction. Figure 1 shows the overall process of the proposed method.

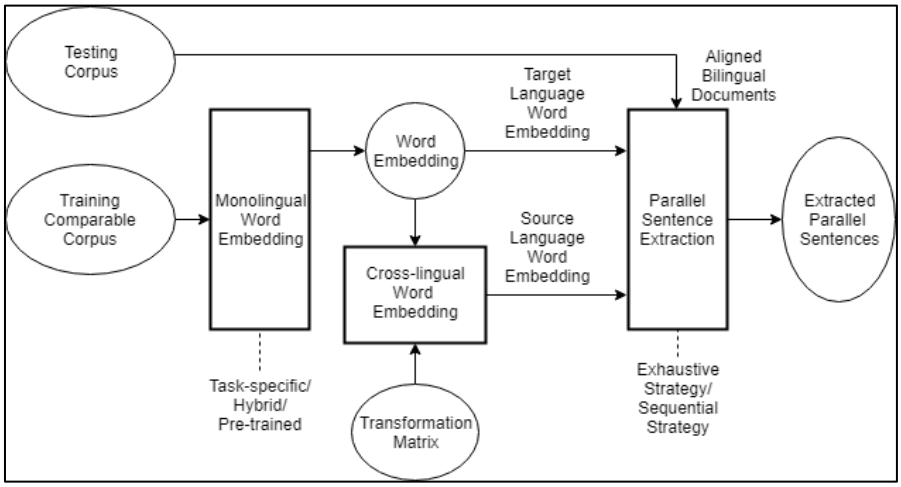


Figure 1. Overall process of our proposed parallel sentence extraction method.

2.1. Monolingual Word Embeddings

Before we link the representations of the two target languages (i.e., Chinese and English in our study), we need to create monolingual word embeddings for the two languages. Specifically, we independently train word embeddings for the source language and the target language, respectively. Several pre-processing steps are involved for the English corpus, including case unification, stemming, and stop word removal. For the Chinese corpus, word segmentation and stop word removal are performed.

Because a domain-specific corpus for word embedding training may not contain a sufficient number of paired documents, the quality of the resultant word embeddings may be compromised. In this study, we will incorporate and empirically evaluate BERT pre-trained models [11] in the proposed method. BERT, a language model developed by Google, uses the bidirectional training of Transformer (attention model) to language modelling and has been applied to many natural language tasks. BERT has released language models in more than 100 languages. In our experiments, we will evaluate the following word embedding models:

1. Pre-trained model: monolingual word embeddings directly from BERT pre-trained models.
2. Task-specific model: monolingual word embeddings directly trained from a task-relevant bilingual training corpus (i.e., 25,595 abstracts of theses).
3. Hybrid model: monolingual word embeddings by concatenating pre-trained and task-specific representations, thus doubling the number of dimensions.

It is noted that the pre-trained Chinese BERT model is based on characters [11], i.e. its vocabulary consists of single Chinese characters rather than words. It may not be optimized for our parallel sentence extraction task.

2.2. Cross-lingual Word Embeddings

Cross-lingual transfer of word embeddings is intended to establish semantic mapping between words in the source and target languages. In this study, we follow the transformation matrix approach to transform the source embedding space (i.e., word embeddings of the source language) to the target embedding one. We use the objective function of cross-lingual word embeddings from [8] to minimize the sum of the loss between Wx and y:

$$\min_{W \in R^{d \times d}} \frac{1}{n} \sum_{i=1}^{n} \ell(W_{x_i}, y_i)$$

ℓ is the loss function, W is transformation matrix with $d \times d$ dimensions, x and y are seed word pairs, x_i is ith x's word embedding, y_j is jth y's word embedding, and n is the number of seed word pairs. In this study, we will construct an optimized transformation matrix for each word embedding model (i.e., pre-trained, task-specific, and hybrid model).

2.3. Parallel Sentence Extraction

Given a pair of aligned bilingual documents, our goal is to extract all possible semantically equivalent or highly similar sentence pairs in the aligned documents as the extracted parallel sentences. Therefore, we need to estimate the similarity of any pair of bilingual sentences based on the word embeddings of their constituent words.

2.3.1 Measuring the Similarity of a Sentence Pair

According to [4], the use of word-by-word similarity can reach a greater effectiveness than the use of sentence embedding similarity when measuring the similarity of two sentences. Thus, in this study, we adopt the word-by-word similarity approach to determine whether two sentences in different languages are similar. Assume that we are estimating the similarity between a source sentence S and a target sentence T, where $S = [s_1\ s_2\ \dots\ s_n]$, s_i is the ith word in S, n is the number of words in S, $T = [t_1\ t_2\ \dots\ t_m]$, and m is the number of words in T. For each word s_i (denoted as source word) in the source sentence S, we calculate the cosine similarity between s_i and every word t_j (denoted as target word) in the target sentence T, according to their word embeddings. That is, for source word s_i and target word t_j,

we obtain CosSim(Vs_i, Vt_j), where Vs_i is the transformed word embedding of s_i (i.e., $W \times s_i$) and Vt_j are the word embeddings of t_j. Among all of the candidates in the target sentence, the word that attains the largest cosine similarity to the source word s_i is identified as the matched word for s_i. After every source word has found a matched word from the target sentence, the average of the similarities of all matched word pairs is calculated and used as the similarity of the source and target sentences.

In [4], when a word pair matches, both the source and the matched target word are removed from the corresponding sentences. This process (word removal during the comparison process of two sentences) may be appropriate for sentences that are written in the same language family (e.g., both the source and target languages are European languages). Because our study deals with Chinese-English language pairs, this process may be inappropriate due to the differences between Chinese and English languages. Specifically, in the Chinese-English scenario, especially after word segmentation, there are many cases in which multiple Chinese words link to the same English word. For example, "新創企業" in Chinese means "startup" in English. However, the Chinese term is typically segmented into "新創" and "企業". Assume that we match "新創" with "startup" and then remove "新創" and "startup" from the subsequent comparison process, we may not be able to match "企業" with any other remaining word, because other similar words such as "company" or "firm" may not appear in the focal English sentence. If we keep the word "startup" in the English sentence, it is likely that the word "企業" in the Chinese sentence can match this English word then. Accordingly, in our study, when a word pair matches, we will not remove both the source word and the matched target word from the subsequent comparison process. The similarity between a source sentence S and a target sentence T is then redefined as follows. We will empirically validate whether the redefined similarity method without word removal can achieve better performance in Section 3.5.

$$Sim(S,T) = \frac{1}{n}\sum_{i=1}^{n} \max_{j \in T} CosSim(V_{s_i}, V_{t_j})$$

2.3.2 Matching Sentence Pairs from an Aligned Document Pair

Previous studies focused on one-to-one sentence matching. When dealing with an aligned Chinese-English document pair, we need to address the difference between sentence segmentation for English documents and that for Chinese documents. For English documents, we generally segment sentences by periods and question marks. However, in Chinese writing, people often concatenate many subsentences by commas into a long sentence (ended with a

period symbol or a question mark). Thus, if we follow the sentence segmentation for English documents to segment a Chinese document into sentences, we may create overly lengthy Chinese sentences, each of which may not be semantically coherent. To avoid this problem, in this study, we segment a Chinese document by commas, periods, and question marks, so that a Chinese sentence may correspond to an English sentence or a part of an English sentence. In other words, in our study, a single English sentence may be aligned to one or multiple Chinese sentences. As shown in the first example in Table 1, one Chinese sentence (C1) is mapped to one English sentence (E1). In contrast, in the second example, two Chinese sentences (C2-1 and C2-2) correspond to one English sentence (E2), where C2-1 is the translation of the subsentence before the comma in E2 and C2-2 is for the subsentence after the comma in E2. To deal with many-to-one (multiple Chinese sentences to one single English sentence) sentence matching, we develop two search strategies (sequential vs. exhaustive search strategy), which will be detailed in the following.

Table 1. Examples of parallel sentence pairs in English and Chinese.

C1: 藥物 開發 成本 高昂 且 費時	E1: drug development is costly and time-consuming
C2-1: 因此 為了 解決 藥物 開發 的 困難 C2-2: 許多 研究 人員 開始 尋求 替代 方法	E2: as a result to overcome the challenges of drug development, researchers start to explore alternative methods for drug development

2.3.2.1 Sequential Search Strategy

Given an aligned bilingual document pair (D_S, D_T) and a similarity threshold α, the sequential search strategy first compares the first source sentence S_1 with the first target sentence T_1. In our study, source sentences in D_S are in Chinese language and target sentences in D_T are in English language. When comparing a source sentence (denote as S_i) and a target sentence (denoted as T_j), the following two cases emerge:

Case 1: If the similarity is lower than a blocking threshold (in this study, we set the blocking threshold = $\alpha/3$, lower than the sentence-similarity threshold α), we skip T_j and move to check the next target sentence T_{j+1}. The search process continues. When all candidate target sentences have examined (the candidate target sentences for S_i include the target sentences in range of $i \pm (\lambda\text{-}1)$, i.e., from $T_{i-(\lambda-1)}$ to $T_{i+(\lambda-1)}$, where λ = the maximum location span to check) and all of the target sentences are dissimilar to the source sentence with respect to the blocking threshold, S_i will be discarded. When this happens, we move to the next source sentence S_{i+1} and start this search process.

Case 2: If the similarity is higher than the blocking threshold, we concatenate with the source sentence (S_i) all possible sequential combinations of the following k-1 source sentences (i.e.,

S_{i+1} to S_{i+k-1}), thus generating k source candidates (including S_i, $S_i + S_{i+1}$, $S_i + S_{i+1} + S_{i+2}$, ..., $S_i + \ldots + S_{i+k-1}$). We then compare each of the source candidates with the target sentence (in this case, T_j). Specifically, we calculate their similarity discounted by length difference. Sentence pairs with greater length differences (measured by the number of words) are unlikely to be parallel sentences. Therefore, the similarity score should be decreased by length difference between the source candidate and the target sentence. The length-difference-penalized similarity score between a source candidate S_x and a target sentence T_y is defined as follows:

$$Score = Sim(S_x, T_y) * (1 - \frac{|len(S_x) - len(T_y)|}{len(S_x) + len(T_y)})$$

Once we complete the calculation of the similarities of these source candidates with the target sentence, we choose the one with the highest similarity and check if it surpasses the predefined similarity threshold α. If the highest similarity does not reach α, we then head to the next target sentence T_{j+1}. However, if the highest similarity exceeds the predefined α, we consider this pair of the specific source candidate and the target sentence as a parallel sentence pair and extract them out of the aligned bilingual document pair (D_S, D_T).

After successfully extracting a parallel sentence pair, we move to the next source sentence and the next target sentence. For example, suppose we find a successful parallel sentence pair (S_1 + S_2, T_1) and we will restart the search process using S_3 as the source sentence and T_2 as the target sentence. If we discard any source sentence, we will anchor the search process from the next source sentence (e.g., S_{i+1}) and the range of the target sentences to check is from $T_{i-\lambda}$ to $T_{i+\lambda}$. For example, suppose S_1 fails and is then discarded. We will start the search process by letting the source sentence as S_2 and the target sentence as T_1 (because $T1$ has not been aligned with any source sentence and is within the range of the target sentences to check for S_2). However, if we discard too many source sentences, we will start discarding target sentences. For example, let λ = 5. Assume that S_1 to S_5 are all failed and discarded. Instead of keeping the target location at T_1, we will start the search process for S_6 with the target sentence T_2 (not T_1, because T_1 is not within the range of the target sentences to check for S_6) as the beginning search point.

2.3.2.2 Exhaustive Search Strategy

The exhaustive search strategy is to compare each source candidate (one or at most k consecutive Chinese sentences in D_S) with every target sentence (an English sentence in D_T) in an aligned bilingual document pair. Then, we select the sentence pair (consisting of a source candidate and a target sentence) with the highest similarity. If the similarity of the

selected sentence pair is equal to or higher than the predefined threshold α, it is extracted as a parallel sentence pair and the corresponding source candidate and target sentence are removed from D_S and D_T, respectively. Subsequently, the sentence pair with the next highest similarity is selected and checks against α to see whether it can be extracted as a parallel sentence pair. The process repeats until the selected sentence pair's similarity is less than α.

The differences between the exhaustive search strategy and the sequential search strategy are twofold. First, the search process of the sequential search strategy is sequential, from the beginning of each document, whereas the search process of the exhaustive search strategy compares all possible sentence pairs. As a result, the sequential search strategy is more efficient than the exhaustive search strategy, especially when source and target documents are large in their length. Second, the sequential search strategy imposes a blocking threshold (i.e., $\alpha/3$ in our study) and a maximum location span (λ) during the search process, while the exhaustive search strategy does not. As a result, the sequential search strategy may result in a suboptimal solution, possibly leading to inferior extraction effectiveness. We will report our evaluation of the two search strategies in Section 3.

3. The Experiments

3.1 Dataset

Our dataset was a corpus containing 25,695 bilingual (Chinese and English) abstracts of theses from the science, engineering, management, and medical colleges in National Taiwan University, Taiwan. We randomly selected 100 bilingual abstracts in this corpus as the testing set. The remaining 25,595 abstracts are the training set for generating monolingual word embeddings and a cross-lingual transformation matrix. Seven coders (graduate students of National Taiwan University) helped manually identify parallel sentence pairs from the testing set as the ground truth for our experiments. Each matched parallel sentence pair contains one English sentence and multiple (one to five) Chinese sentences, and there are at least three matched parallel sentence pairs for each pair of abstracts. 66.7% of the testing English sentences have matched Chinese sentences, and the average number of Chinese sentences in each parallel sentence pair is 1.81. Table 2 lists the statistics of our data set.

Table 2. Statistics of our data set including training and testing sets. 1044 out of 1462 Chinese sentences, and 576 out of 864 English sentences are matched pairs.

	# of Documents	**Word Count**	**Sentence Count**	**# of Sentences per Doc**
Training Set (Chinese, Zh)	25,595	2,388,729	346,591	13.54
Training Set (English, En)	25,595	1,981,510	195,910	7.65
Testing Set (Zh)	100	14,618	1,462	14.62
Testing Set (En)	100	15,000	864	8.64

3.2 Comparative Evaluation Results

As mentioned previously, our proposed parallel sentence extraction method can use one of the following monolingual word embeddings: 1) task-specific word embeddings (denoted as **TS**) directly trained from the training set, using Cbow or Skipgram from [12], 2) pre-trained word embeddings (denoted as **PRE**) extracted from BERT [11], and 3) hybrid word embeddings (denoted as **HB**) that concatenate TS and PRE word vectors. In the following experiment, we first employed Cbow to build task-specific word embeddings. We will compare the performance differential when using Cbow or Skipgram to build task-specific word embeddings in Section 3.4. Furthermore, for the TS model and the PRE model, the number of dimensions for word embedding was set to 200, and for the HB model, it was 400. The number of dimensions of the PRE model was originally 768 and was reduced from 768 to 200 via dimension reduction using principal component analysis.

Before we conduct our experiment, the first test is to decide the transformation direction, i.e., whether the transformation from Chinese (Zh) words to English (En) is better than the opposite direction (the transformation from English words to Chinese. In our test on 576 matched pairs in the testing set and another 576 randomly selected, non-matched pairs, the sentence similarities calculated by the Zh-En transformation attained higher average similarity on the matched pairs, lower average similarity on the random pairs, and greater difference between true and false pairs, as compared to those of the En-Zh transformation, as Table 3 illustrates. As a result, we decided the transformation direction is from Chinese to English, and set the Chinese corpus as source documents and the English corpus as target documents for subsequent experiments.

Table 3. Comparison of sentence similarity conducted by different transformation directions (Zh-En and En-Zh transformation).

Word Embedding	**Source Language**	**Target Language**	**Avg Sim (Matched Pairs)**	**Avg Sim (Random False Pairs)**	**Difference**
TS	Zh	En	0.3142	0.1504	**0.1648**
TS	En	Zh	0.3021	0.1537	0.1484
PRE	Zh	En	0.5556	0.4238	**0.1321**
PRE	En	Zh	0.6726	0.561	0.1116

We built each word embedding model's transformation matrix by linear regression with stochastic gradient descent, using 2,000 most commonly used English words (stop words and words without Chinese translation have been removed) in the training set as seed words. We then evaluated our proposed method using the metrics of precision, recall, and F1.

Table 4 shows the comparative evaluations results, across the three word embedding models and two search strategies, where SEQ denotes the sequential search strategy and EX is the exhaustive search strategy. To determine the similarity threshold α, both strategies took the multiplication of the average sentence similarity of 1,000 prepared parallel sentences and a coefficient (an optimal coefficient was empirically determined). Furthermore, for the sequential search strategy, we set the blocking threshold as one third of α. For each English sentence candidate, we set k = 5 (up to 5 Chinese sentences to be concatenated), and λ (maximum location span) = 5.

Table 4. Performance comparison of our proposed parallel sentence extraction method using different search strategies and word embedding models.

Search Strategy	**Word Embedding**	**Threshold Coefficient**	**Recall**	**Precision**	**F1**
SEQ	TS	0.9	36.54%	**68.55%**	47.67%
SEQ	PRE	0.8	31.17%	49.60%	38.28%
SEQ	HB	0.95	31.53%	59.74%	41.28%
EX	TS	0.7	**56.78%**	64.00%	**60.18%**
EX	PRE	0.7	21.94%	22.77%	22.35%
EX	HB	0.8	54.23%	64.67%	58.99%

As Table 4 shows, the exhaustive search strategy using the task-specific embedding model achieved the best performance in recall and F1 measure, while the sequential search strategy using the task-specific embedding model attained the highest precision rate. With either the sequential search strategy or the exhaustive search strategy, the task-specific embedding model generally outperformed its counterparts, whereas the pre-trained word embedding model performed worst. This finding suggests that the compatibility of the corpus used to generate monolingual word embeddings and the testing corpus (i.e., parallel sentence extraction task) significantly affects the effectiveness of parallel sentence extraction.

We also observed that when using the exhaustive search strategy, the F1 score attained by the pre-trained embedding model was significantly lower than when using the sequential search strategy. The sequential search strategy compares only sentences with similar positions in the two aligned bilingual documents, while the exhaustive search strategy compares all possible sentence pairs. Because of the low compatibility of the pre-trained embedding model with the testing corpus, the exhaustive search strategy identified more false pairs than the sequential search strategy, highlighting the limitation of the pre-trained embedding model.

On the other hand, we expect that the hybrid embedding model could combine the advantages of the pre-trained embedding model and the task-specific embedding model and could achieve a better performance than the other two models. However, according to Table 4, the performance of the hybrid embedding model was in between that of the task-specific embedding model and the pre-trained embedding model. This is because the unacceptable performance attained by the pre-trained embedding model implicates the performance of the hybrid model.

3.3 Performance of Sequential and Exhaustive Search Strategies

Figure 2 shows the performance differences using the sequential or exhaustive search strategy. Since the sequential search strategy does not compare a source sentence with target sentences located far from the corresponding location of the source sentence. This strategy is likely to miss some matched pairs. Thus, the recall rate of the sequential search strategy is expected to be lower than that of the exhaustive search strategy. In contrast, because these relative distant sentence pairs are mainly false positives, the sequential search strategy can reach a higher accuracy than the exhaustive search strategy. Overall, the F1 score attained by the exhaustive search strategy is higher than that by the sequential search strategy, but the exhaustive search strategy is more time consuming. Figure 3 and Figure 4 show that the exhaustive search strategy performed better in recall rate, while the sequential search strategy performed slightly better in precision rate.

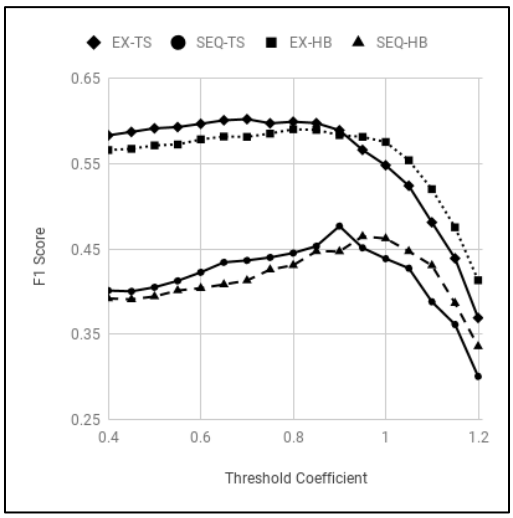


Figure 2. F1 measures obtained by the sequential and exhaustive search strategies. X dimension represents threshold coefficient and Y dimension represents F1 score.

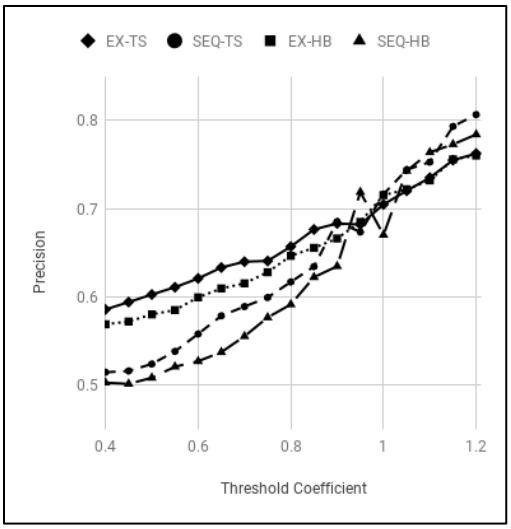


Figure 3. Precision rates obtained by using the sequential and exhaustive search strategies. X dimension represents threshold coefficient and Y dimension represents precision rate.

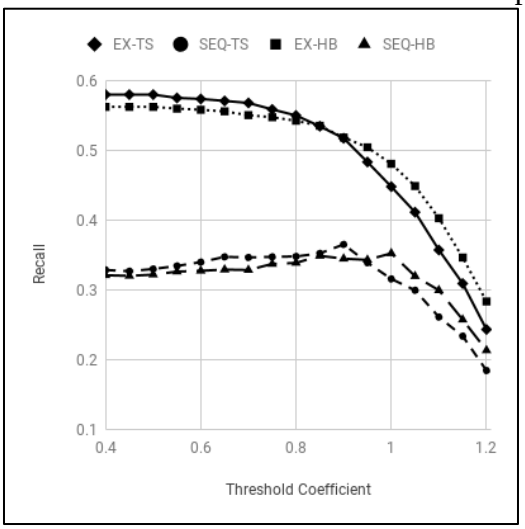


Figure 4. Recall rates obtained by using the sequential and exhaustive search strategies. X dimension represents threshold coefficient and Y dimension represents recall rate.

3.4 Effect of Cbow and Skipgram

We also analyzed the effect of Cbow and Skipgram on the effectiveness of parallel sentence extraction. Table 5 shows that the word embedding models (task-specific and hybrid) constructed by Cbow performed better than the models constructed by Skipgram, similar to the results reported in [13]. According to [12], if Cbow is trained on a large corpus, it would perform better than Skipgram. It seems that our corpus is relatively sufficient for Cbow.

Table 5. Performance comparison across different word embedding structures.

Word Embedding	**Structure**	**Threshold Coefficient**	**F1**
TS	Skipgram	0.7	43.03%
TS	Cbow	0.7	**60.18%**
HB	Skipgram	0.8	30.33%
HB	Cbow	0.8	**58.99%**

3.5 Effect of Word Removal When Measuring the Similarity of Two Sentences

To understand the effect of word removal when measuring the similarity of two sentences, Table 6 shows the performance obtained with or without word removal using the exhaustive search strategy. In general, our proposed parallel sentence extraction method without word removal achieved a lower precision rate, but a higher recall rate, as compared to our proposed method with word removal. With respect to F1 score, our proposed method without word removal outperformed that with word removal, across the two word-embedding models (task-specific and hybrid)

Table 6. Performance comparison with or without word removal during sentence extraction.

Word Embedding	**Threshold Coefficient**	**Word Removal**	**Recall**	**Precision**	**F1**
HB	0.8	Yes	50.44%	68.52%	58.10%
HB	0.8	No	54.23%	64.67%	**58.99%**
TS	0.7	Yes	54.51%	66.25%	59.81%
TS	0.7	No	56.78%	64.00%	**60.18%**

4. Concluding Remarks

In this work we have proposed and implemented an effective method for extracting parallel sentence pairs from bilingual comparable corpora. The effects of differences in word embedding model (task-specific/pre-trained/hybrid), search strategy (sequential/exhaustive), word vector formation (Cbow/Skipgram), and word removal or not have been empirically evaluated. By using the task-specific word embedding with the exhaustive search strategy, our proposed method can achieve the best performance in F1 score.

5. References

[1] M. Artetxe, G. Labaka, E. Agirre, and K. Cho (2018). "Unsupervised Neural Machine Translation." In *Proceedings of the Sixth International Conference on Learning Representations*.

[2] Y. Su, H. Liu, S. Yavuz, I. Gur, H. Sun, and X. Yan (2018). "Global Relation Embedding for Relation Extraction." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 820-830.

[3] H. Bouamor and H. Sajjad (2018). "H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings." In *Proceedings of the*

Eleventh Workshop on Building and Using Comparable Corpora at *International Conference on Language Resources and Evaluation*, pp. 43-47.
[4] V. Hangya , F. Braune, Y. Kalasouskaya, and A. Fraser (2018). "Unsupervised Parallel Sentence Extraction from Comparable Corpora." In *Proceedings of the International Workshop on Spoken Language Translation*, pp. 7-13.
[5] J. Smith, C. Quirk, and K. Toutanova (2010). "Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment." In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 403-411.
[6] D. Wu, and P. Fung (2005). "Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-comparable Corpora." In *Proceedings of the International Conference on Natural Language Processing*, pp. 257-268.
[7] W. Yang, W. Lu, and V. Zheng (2017). "A Simple Regularization-based Algorithm for Learning Cross-Domain Word Embeddings." In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pp. 2898-2904.
[8] A. Joulin, P. Bojanowski, T. Mikolov, H. Jegou, and E. Grave (2018). "Loss in Translation: Learning Bilingual Word Mapping with A Retrieval Criterion." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2979-2984.
[9] M. Artetxe, G. Labaka, and E. Agirre (2017). "Learning Bilingual Word Embeddings with (Almost) No Bilingual Data." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 451-462.
[10] R. Yin, Q. Wang, P. Li, R. Li, and B. Wang (2016). "Multi-granularity Chinese Word Embedding." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 981-986.
[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv preprint arXiv:1810.04805.
[12] T. Mikolov, Q. Le, and I. Sutskever (2013). "Exploiting Similarities among Languages for Machine Translation." arXiv preprint arXiv:1309.4168.
[13] L. Jin and W. Schuler (2015). "A Comparison of Word Similarity Performance Using Explanatory and Non-explanatory Texts." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 990-994.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 182-191

基於卷積神經網路之台語關鍵詞辨識

劉祈宏　Chi-Hung Liu

長庚大學資訊工程學系

Department of Computer Science and Information Engineering
Chang Gung University
m0729015@cgu.edu.tw

呂仁園　Ren-Yuan Lyu

長庚大學資訊工程學系

Department of Computer Science and Information Engineering
Chang Gung University
renyuan.lyu@gmail.com

詹惟中　Wei-Zhong Zhan

b0429016@cgu.edu.tw

吳捷書　Jie-Shu Wu

b0429031@cgu.edu.tw

朱達道　Da-Dao Zhu

b0429052@cgu.edu.tw

施俊良　Jun-Liang Shi

b0429063@cgu.edu.tw

長庚大學資訊工程學系

Department of Computer Science and Information Engineering
Chang Gung University

摘要

本文使用近年興起的卷積神經網路模型來對於台語特定關鍵詞進行訓練，訓練準確度可達 9 成，我們使用 TensorFlow speech command 所使用的 30 單字翻譯成台語，其中包含了十個數字(零到九)及其他日常生活中常用的單詞並提供使用者透過我們建立的平台做單詞音檔即時辨識。

一、 緒論

語言流失已經是全球各地最普遍的危機之一。以台灣為例，目前語言溝通上幾乎以國語為主，幾乎都是老一輩的人在使用台語溝通。根據統計，幾乎大部分會講的年輕人都是七八年級生，也提到家庭環境非常重要。如果從小跟家中長輩經常使用台語，溝通上就不會有太大問題。因此為了保存及推廣台語文化，我們希望開發語音辨識系統，目前網路上並無免費開源台語音檔。首先，我們希望為台語辨識研究者建立一個能夠收集音檔的平台，並針對收集到的音檔進行單詞語音辨識的開發，並提供使用者透過我們建立的平台做單詞音檔即時辨識。

二、 相關研究

由於近年來 GPU 與深度學習與人工智慧興起，我們採用深度學習中的卷積神經網路，主要架構採用此篇論文 [1]，而資料收集方面參考 [2]。在市面上，目前開源的語音資料庫如 LibriSpeech[1], Mozilla Common Voice[2], timit[3] 等皆有國語或英語等語料，但唯獨台語資料在市面上是非常缺乏的，而台語的學習成本也是較高的，臺灣話與其他漢語系語言同為聲調語言，聲調在語句中有辨義作用，亦有不少繁複的變調規則以及文白讀異。

[1] http://www.openslr.org/resources.php 中文語料 SLR18, SLR38, SLR47, SLR62
[2] https://voice.mozilla.org/zh-TW/datasets 華語(台灣) 語料
[3] https://scidm.nchc.org.tw/dataset/darpa-timit DARPA TIMIT 英文語料

三、 台語簡介

台語（英文：Taiwanese Hokkien、Taiwanese，又稱為臺灣閩南語。近代以來常以台語（臺羅：Tâi-gí/gú/gír）稱之。以其為母語的閩南裔臺灣人是臺灣第一大族群。

台灣從明朝末年開始，陸續有部分中國大陸沿海居民遷徙至台灣，特別是”清領後期”渡臺禁令開放後，大量福建南部（閩南）的泉州府和漳州府的居民紛紛遷徙至台灣。由於台灣先後分別歷經了荷蘭及西班牙的統治，後有明鄭與清朝統治，1895 年後更由日本統治長達 50 年，造成閩南語逐漸在台灣各地演變分化，並融入荷蘭語、日語及原住民語言等語言於其中，使得台語與福建的閩南語在詞彙使用及腔調上存在有不少差異，台語也逐漸成為臺灣本島最主要的通行語言之一，而根據 2009 年所發表的《臺灣年鑑》中指出，臺灣民眾約有 73%能夠說台語。

總體上說，台語在北部為偏泉混合腔，中南部平原偏內埔腔，西部沿海偏海口腔。漳州移民主要居住在中部平原地帶、北部沿海地區及蘭陽平原，被稱為內埔腔；泉州移民主要居住在中部沿海地區、臺北盆地，被稱為海口腔，南部則為泉漳混合區。

四、 TensorFlow

TensorFlow 是 Google 基於 DistBelief 進行研發的開源機器學習框架。第一，在這裡會採用 tensorflow 作為開發的原因主要為開源，這意味著世上有志之士皆可以是參與者，提報臭蟲，優化程式碼，使整個社群可以是朝氣蓬勃的。第二為 Google 公司的維護，Google 在 2018 年的市值約 7255 億美元，有如此財力的公司作為後盾，可以確保此框架的延續性。第三，TensorFlow 支援的程式語言種類繁多，目前主流的機器學習框架如 Caffe、PyTorch、CNTK 大多支援 2~3 種語言，或者只有一種，而 TensorFlow 共支援了 python、C++、R、Swift、JavaScript、Go 等… ，可以讓程式設計師更靈活地面對到各種情境，靈活部署。第四、針對行動裝置或是物聯網裝置，TensorFlow Lite 讓模型可在多樣裝置上執行，包括行動裝置、物聯網裝置……等等，意即可以在 Raspberry Pi 或您的手機上，進行機器學習，讓應用場景多樣化。 第五、範例程式碼

與完整的說明文件，我們都知道萬事起頭難，有了官方所提供的範例程式，對於新手而言可以更快地上手，清晰地說明文件可以省下查找程式碼用法時間，提高效率。

五、 台語語音辨識

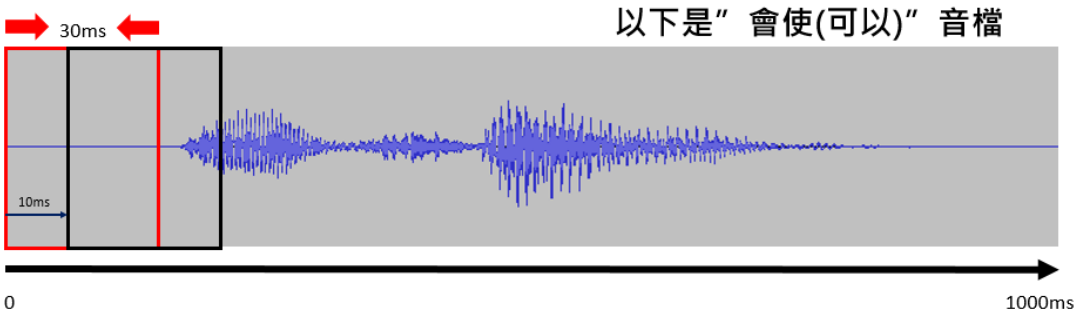
首先，我們為台語辨識研究者建立一個能夠收集音檔的平台，並針對收集到的音檔進行單詞語音辨識的開發，目前我們可以辨識的字共有 30 個單詞如下表所示，而選這三十個詞是因為我們依照 TensorFlow speech command 所使用的 30 單字翻譯成台語，其中包含了十個數字(零到九)及其他日常生活中常用的單詞並提供使用者透過我們建立的平台做單詞音檔即時辨識。

華語	台語	台羅	華語	台語	台羅	華語	台語	台羅
零	零	lîng	上	起去		開	開	khai
一	一	tsi̍t	下	落來	lo̍h-lâi	關	關	kuainn
二	兩	nn̄g	左	倒邊		不可	袂使	b ē -sái/bu ē -sái
三	三	sann	右	正邊		可以	會使	ē -sái
四	四	sì	去	去	khì	志明	志明	
五	五	g ō o	床	眠床	bîn-tshn̂g	春嬌	春嬌	
六	六	la̍k	狗	狗	káu	快樂	快樂	khuài-lo̍k
七	七	tshit	鳥	鳥	tsiáu	房屋	厝	tshù
八	八	peh/pueh	貓	貓	niau	前進	進前	tsìn-tsîng
九	九	káu	樹	樹	tshi ū	後退	退後	thè- ā u

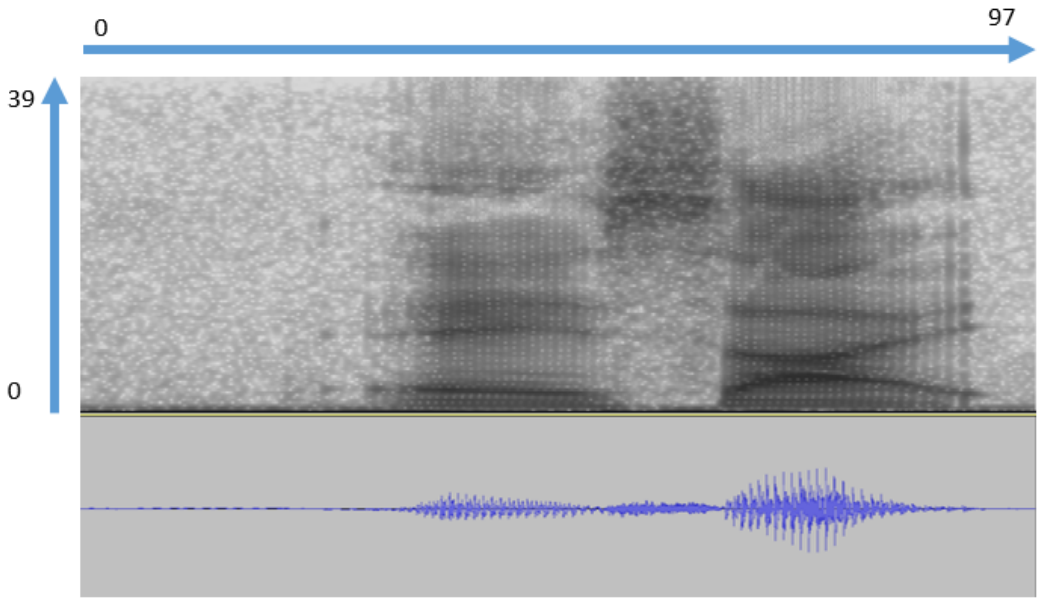
表一、單詞與台語羅馬拼音對照表

(一) 辨識部分

首先在音檔處理部分，我們將這些音檔經過 MFCC 轉換強調人聲的部分後生成頻譜利用 TensorFlow 創建 CNN 模型把頻譜丟入 CNN 進行分類。在音檔格式方面，我們使用 wav 檔 16K 的 sampling rate，即每 1000ms 採 16K 個訊號。設定每次讀取訊號音檔長度為 30ms，而每次訊號讀取移動範圍為 10ms（圖二）。並將讀取到的音檔進行 MFCC 轉換，取 MFCC 特徵點。再來將音檔處理變成頻譜（圖三）。



圖一、音頻圖

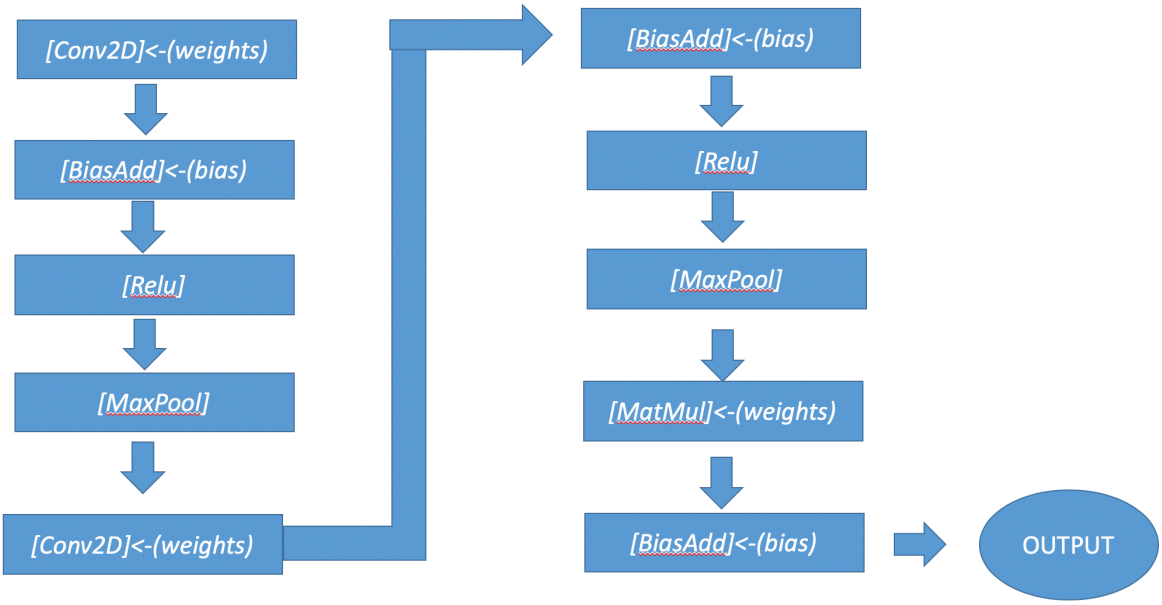


圖二、頻譜圖

(二)CNN 架構

我們選擇使用 CNN 進行訓練，會選擇 CNN 的原因爲通常情況下，語音識別都是基於音頻分析後(MFCC 轉換)的語音頻譜圖完成的，而其中語音頻譜圖是具有結構特點的(各個不同的字音有不同的能量分佈範圍)。要想提高語音識別率，就是需要克服語音信號所面臨各種各樣的多樣性，包括說話人的多樣性(聲音三要素中的音品)，環境的多樣性等(安靜或吵雜的背景音)。而在卷積神經網絡可以達到時域以及頻域的平移不變性(Translation Invariance)，利用此特性將卷積神經網絡應用到語音識別的聲學建模中，則可以此點來克服語音信號本身的多樣性。從這個角度來看，則可以認為是將整個語音信號分析得到的頻譜圖當作一張圖像一樣來處理，採用圖像識別中廣泛應用的

深層卷積神經網絡對其進行識別。而在此我們所採用的架構為 [1]中的 'cnn-trad-fpool3'，因原本架構較為複雜，在此我們將其改為兩層的 Conv2D 跟 MaxPool，詳細結構如下圖所示:



圖三、CNN 結構圖

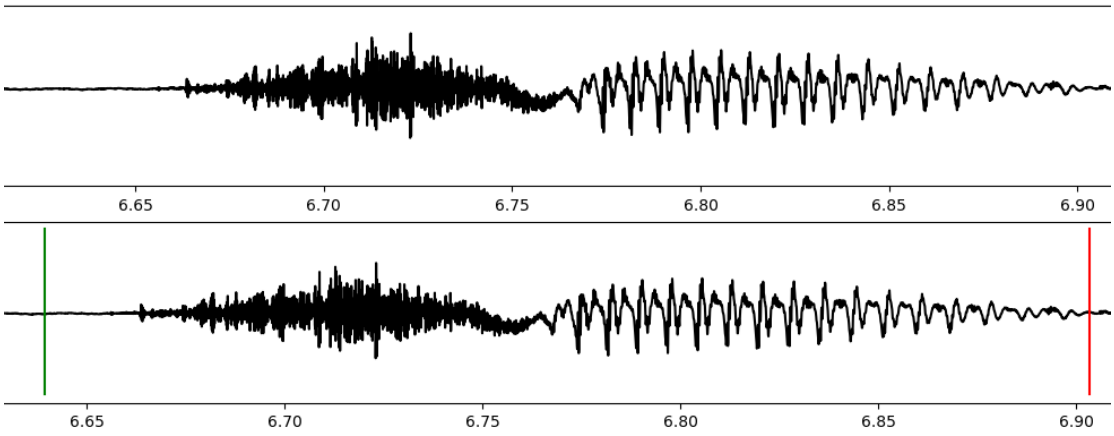
(三)訓練與資料集

市面上並無開源的台語音檔資料，因此我們利用自己架設的網站收集了約 15000 筆台語語音資料，每筆一秒，我們將 15000 筆語音資料分為 3 個子集合 Training data 80%、Validation data 10%、Testing data 10%

(四)端點偵測

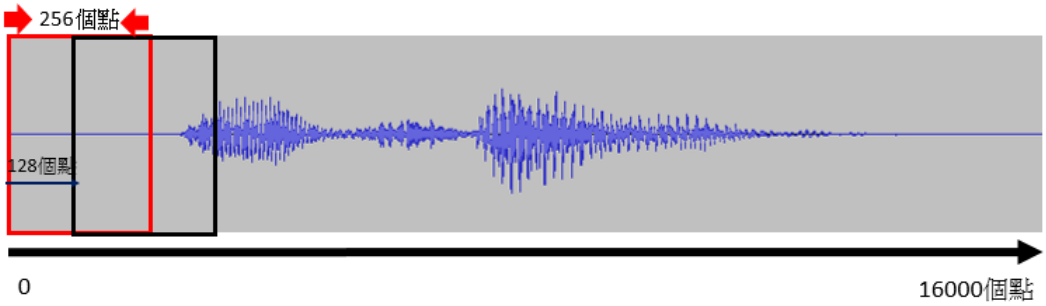
要做到一句話的辨識，會需要做到音檔的分割，從較簡單的連續多個單詞著手，對連續的單詞進行分割，以達成辨識多個單詞。我們使用端點偵測(Endpoint Detection)，定音訊開始和結束的位置，如下圖五所示

我們將端點偵測分為三個部分 1.頻譜正規化:計算分貝數，並給定門檻值，取頻譜中大於門檻值的範圍，認定為有聲音，因此該門檻值能夠決定端點。2.氣音強化: 對於氣音部分，較容易小於門檻值而被摒除在範圍外，所以對頻譜做多次微分，藉此可以將氣音部分凸顯，再經過門檻值的篩選，即可包含較完整的音訊。



圖四 標定選取範圍

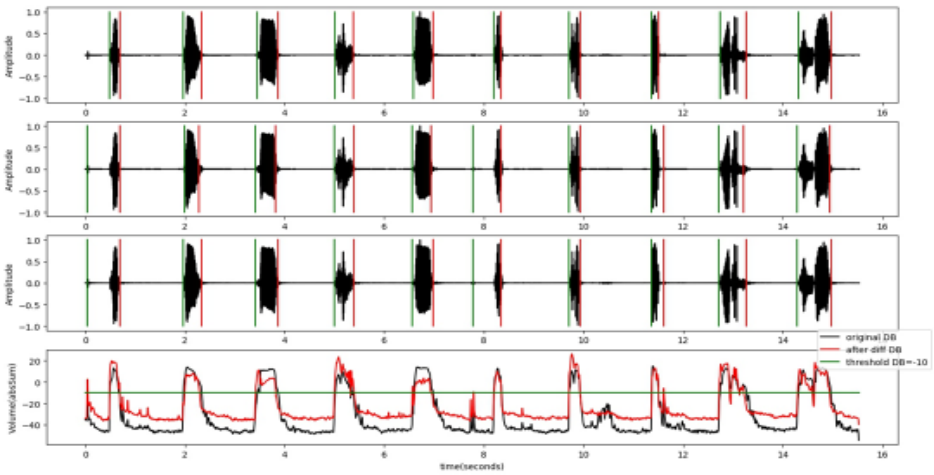
3.後處理: 比較原始頻譜端點及微分後頻譜的端點，並對端點間距做判斷篩選，過濾掉時間較短視為雜音的部分，切割出每個詞彙，並將未滿 1 秒的詞彙補至 1 秒，之後對每個切割出的音檔做辨識。首先我們對音檔進行處理。對頻譜做標準化，將每個點的振幅轉換至[-1,1]每個 frame 取 256 個點，並且每 128 個點作一次平移，計算每個 frame 的 DB 值而我們 DB 值計算方法為對 frame 中每個點做零校正，取平方和後取對數再乘以 10 以 DB=0，DB=-10 等做門檻值進行分割，以下為標準化之示意圖（圖六）



圖五 標準化

圖七是我們切割之結果展示圖，第一列圖為沒有四次微分(將頻譜對時間做四次微分)的圖，而第二列為經過四次微分的圖。第三張是經過後處的圖。第四張是將 DB 門檻加入的展示圖。而綠線為音檔起始線而紅線為音檔結束線。若切割音檔會從綠線到紅

線辨識成一個音檔進行切割，根據實驗結果，經過四次微分後與 DB 值門檻所得結果可以有最大程度的包含進有效頻譜。



圖六 各類切法比較圖

六、 程式環境與訓練結果

我們使用下列規格的電腦配置進行 CNN 模型的訓練

CPU	I7-9800X
MotherBoard	Asus WS X299 Pro
RAM	Kingston 16G*2
SSD	WD Blue 500G
GPU	RTX 2080ti
OS	Ubuntu 16.04
TensorFlow-gpu	1.13.0rc2

表二、環境配置

利用 confusion matrix 顯示訓練結果（圖八），在 CNN 架構中，目前所能辨識的正確率為 90%

```
INFO:tensorflow:Confusion Matrix:
 [[254   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0   0]
 [  0   0  80 ...   0   0   0]
 ...
 [  0   0   0 ...  82   0   0]
 [  0   0   0 ...   0  58   0]
 [  0   0   0 ...   0   0  50]]
INFO:tensorflow:Final test accuracy = 90.1% (N=2789)
```

圖七、訓練結果

在實際辨識上，我們有撰寫出一個網頁可以讓使用者選擇連續辨識(可辨識詞的範圍在30 詞內，圖九)，以及一次一個單詞辨識(錄音一秒，圖十)

圖八、連續辨識

圖九、單詞辨識

從圖九與圖十中右半部皆有上傳、下載與刪除之按鈕，這裡就是緒論中有提到的，我們希望可以製作一個平台，讓使用者可以在使用我們辨識程式時，可以上傳辨識中所錄下的音檔，也可以將錄音下載回自己電腦，如不克上傳，也可以按下刪除鍵。我們希望成為一個收集平台，保存台語文化。

七、　參考資料

[1] Tara N. Sainath, Carolina Parada, “Convolutional Neural Networks for Small-footprint Keyword Spotting,” *INTERSPEECH 2015*, 2015.

[2] P. Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,” *Google Brain Mountain View*, California, April,2018.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 192-201

Extracting Semantic Representations of Sexual Biases from Word Vectors

Ying-Yu Chen
National Taiwan University
r06142009@ntu.edu.tw

Shu-Kai Hsieh
National Taiwan University
shukai@gmail.com

摘要

在台灣的標誌性論壇——PTT 論壇中，帶有性別偏見的線上仇恨言論問題發展已久，詞向量等計算語言學的應用也常帶有性別偏見。本研究利用詞向量表徵（Word2vec），從 PTT「母豬教」鄉民的言談來分析中文厭女言論的詞向量的語意分佈，發現母豬教徒相關言論的確帶有性別偏見與歧視。本研究採取質性加上量化分析，作為第一篇利用自然語言處理技術分析 PTT 論壇上的仇女歧視言論的研究。

Abstract

Sexually biased cyberhate speech has become a fast-growing problem on PTT (a representative online forum in Taiwan). The applications of computational linguistics like word embeddings would also carry similar biases. This paper analyzed the distribution of word representations of netizens from *mu zhu jiao* (a cult that often produces misogynistic cyberhate speech). Word vector representations (word2vec) was utilized for scrutinizing semantic representations of texts found on PTT. The findings from the distributed semantic representation of *mu zhu jiao* implied a sexual bias against them. This paper serves as the first study which investigates the distribution of word representations of the abusive language on PTT forum with an NLP method by taking advantage of both quantitative and qualitative methods.

1 Introduction

During the past few years, the tides of cyberhate speech have surged over online forums and social media. PTT forum, serving as the representative of online Chinese forum, has been to obtain evidence of abusive cyberhate speech on the basis of characterizing gender-biased language, mainly as

disparagingly underestimating the status of the female. Previous research focusing on cyberhate speech of misogyny (Citron, 2011) and negative sentiment of the related word representations (Yu, 2016) made up the big picture of the gender-related cyberhate speech. The emphasis on Distributional Hypotheses (Sahlgren, 2008) also depicts the importance of exploring distributional differences of word representations under the contexts related to a sexually biased speech based on a quantitative computational method. Instead of unsystematic regulations of stopping the propagation of cyberhate speech, there were some methods applying Natural Language Processing (NLP) models for efficiently detecting abusive language online with regard to, but not limited to, cyberbullying and gender-biased speech (Schmidt and Wiegand, 2017;Davidson et al., 2017;Burnap and Williams, 2016). On the ground of previous studies, the relationships and connections between cyberhate speech and gender-biased language on PTT could be reported in both quantitative and qualitative ways. The aim of the present paper is threefold: (1) to analyze the large text sources and compare the distributed representations of words represented by a word vector encoding semantic similarity;(2) to capture linguistic properties from the word vectors;(3) to specify the polarity fluctuation with regard to the keywords from the word vector.

2 Previous Studies

2.1 Misogyny and Cyberhate Speech

Misogyny complex, which Gilmore (2010) suggested best describing as a multi-dimensional phenomenon characterizing the woman hating emotion males direct toward females. This pervasive emotion over the world was specified by Gilmore as a result of the combination of psychogenic in origin and the influence caused by the environment. Specifically, on one hand, the inside conflict between men's abundant need for women and the equally intense fear of that dependence (or fear of losing that dependence) disappointed males themselves, thus leading to their implicit emotional dissatisfaction. Under this circumstance, the female had the possibility to become a convenient and better object to be bullied as a counterbalance to make up the defect. Although society appeared to make a fair amount of progress against gender discrimination, misogyny has, by all means, moved online due to the easy opportunities of speaking on condition of anonymity on the Internet, called "cyberhate attack (Citron, 2011)." Cyberhate attack contained not only speech, but other actions like doctored photographs, privacy invasions, and technical sabotage, while "cyber misogyny" referred to cyberhate attack such as abusive language specifically aiming at women.

As the representative online BBS (Bulletin Board System) forum in the Chinese community, PTT is the place where a wealth of cyberhate speech of misogyny takes place. Yu (2016) claimed that the key concept of misogyny on PTT appeared to follow two patterns: One was emphasizing on the beautiful figure and the virtues that female supposed to have, like performing both emotionally and physically weaker than male in concordance to meet the role expectation from Asian society. The other was restricting the completeness of female as an individual by devaluing female as only an object of sex. This objectification was to a large degree subjective to the critics who tended to take advantage of their ideology to label and define female by using words with specific semantic representations under the context. Such word representations with the aforementioned features were retrieved and visually demonstrated by Yu (2016) as a semantic network, which signaled the distributional word representations based on misogynistic cyberhate speech on PTT. This distribution to an extent specified the words appeared on PTT, such as *gan (fuck), qian (money), chou nu (misogyny), sui bian (easy especially referring to sex), gong zhu (princess syndrome), and po ma (bitch)*. On the basis of the literature review, this present study would be investigating the semantic networks from the abusive speech on PTT forum in the follow-up phases of the study.

2.2 Word Vectors

"You shall know a word by the company it keeps (Firth, 1957)." Distributional hypothesis depicts the idea of the correlation between distributional similarity and meaning similarity so as to make a prediction of semantics by utilizing distributional properties of linguistic entities (Sahlgren, 2008). Exploring the distributional properties of the misogynistic cyberhate speech and the differences between the different distributions are needed. Now, this could be implemented via a neural network inspired vector semantic model called Word2vec (Mikolov et al., 2013;Mikolov et al., 2013) which has been widely used for gathering high-quality vector representations of words from large amounts of unstructured data by using the advantageous Skip-gram model. Previous work (Heuer, 2016) utilized this simple but robust model to provide a view of text source and to compare the words with semantically similar or contrasting relationships, like mapping the country *Finland* to its national sport *hockey*. Schmidt and Wiegand (2017) also gave an overview of hate speech detection, which indicated distributed word representations based on neural networks was an effective way yielding a good classification performance from working the features of word generalization out. Such vector representations were used as classification features as it had the

preference that semantically similar words may end up with similar vectors. With the assumption that distributional models are models of word meaning, this study took advantage of such technique to scrutinize the large text sources for revealing the linguistic properties from the distributional models.

3 Methodology

3.1 Data Collection

Over the decades, PTT has derived various subcultures, among which a new force suddenly arose as *mu zhu jiao*, a cult whose followers believe in the existence of *mu zhu* and worship the popes obov (also known as a000000000) and sumade. On the online forum, everyone can easily share and exchange opinion with each other on more or less condition of anonymity, resulting in the gender-biased speech contributed from the popes of *mu zhu jiao*. With the purpose of getting evidence of the characteristics and nature of the abusive speech on PTT, this study favored making a comparison of the speech between general *xiang min* and the popes of *mu zhu jiao*, trying to draw the differences between them. Hence, the data population of this study consisted of two datasets: Dataset I is made up of the posts from 180 PTT netizens (also known as *xiang min*) randomly chosen from the online PTT user list, while dataset II involved posts from two remarkable popes of *xiang min* of *mu zhu jiao.*

The big picture of the procedure was to collect, organize, and integrate the data. The data were first gathered through crawling the posts from the aforementioned *xiang min* on PTT, after which a popular tool "jieba" was used for segmentalizing text into tokens. In the meantime, a list of 1218 stopwords (e.g., emojis and interjections) was ruled out to avoid getting too much uninformative data. Apart from the list of stopwords, the user dictionary[1] which is comprised of a bunch of frequently used terms and multi-word expressions (MWEs) on PTT was taken into consideration. This user dictionary is consistent with the MWEs and catchphrases employed by pttpedia with the further trimming to 5294 tokens. With the help of the user dictionary, a number of essential terms and fixed constructions were maintained in the data for further analysis. Table 1 provides the overall information of two datasets, including the number of articles, word tokens, word types, and

[1] The user dictionary is retrieved with permission from Liao (https://liao961120.github.io/PTTscrapy/).

vocabulary richness (VR). VR indicated that the words xiang min from dataset II used were slightly richer than that from xiang min from mu zhu jiao.

Table 1: The number of articles, token, types, and vocabulary richness (VR) of two datasets

Population	sumade	obov	a000000000	Dataset I	Dataset II
Articles	1230	573	340	**2143**	**2855**
Tokens	226589	56017	35694	**329127**	**333575**
Types	45339	16508	12501	**63039**	**72648**
VR (%)	0.2001	0.2947	0.3502	**0.1915**	**0.2178**

3.2 Data Analysis

This study employed both quantitative and qualitative approaches, comprising of text analysis and questionnaire of the authentic data from PTT. The quantitative data analysis was performed by using the powerful package word2vec from NLP architecture. Word2vec can be seen as a fashionable and efficient model utilizing large data to compute vector representations of words (Mikolov et al., 2013). With this hypothesis, it was carried out so as to map the representations of words into a vector space to gain word similarity and relatedness. This paper drew a comparison of word representation between the two groups: (1) two well-known *xiang min* of high reputation from *mu zhu jiao*, obov (also known as a000000000) and sumade;(2) 180 randomly chosen *xiang min*. After respectively training the dataset I and dataset II from 329127 and 333575 tokens, two comparing models of representation of vector space were created.

For conducting the analysis, previous researchers typically compared the words with semantically similar or contrasting relationships (Heuer, 2016). Following this, the study was proceeded by generating a bunch of co-occurred words from the model to represent the learned distributional relationships. In the two models, the core word *mu zhu* was queried to predict 20 words with strong context dependency and their corresponding degrees of cosine similarity in the vector space in an effort to detect the distributed difference of word representation between *xiang min* from *mu zhu jiao* and others. Subsequently, the above distributions would be scrutinized to obtain the authentic context and compare the different semantic implication of the keywords in two comparing models. The corresponding concordances of the above 20 words with strong context dependency with *mu zhu* were analyzed as in section 4.1. Furthermore, in the interest of

investigating how the polarity of the related language varied from people who happened to have the related knowledge of misogynistic hate speech on PTT, this study employed a questionnaire to investigate the polarity of the keywords between two groups of people: thirty senior *xiang min* who have logged in PTT more than 3000 times, and thirty people who are not *xiang min*. The keywords were manually selected under human determination in order to filter out the 25 words with the highest correlation with the context of misogynistic subculture on PTT. The result would be discussed in section 4.2.

4 Result and Discussion

4.1 Co-occurred Word and Concordance

The study was proceeded by generating a bunch of co-occurred words from the model to represent the learned distributional relationships. The outcome of the distributed representation from comparing word vectors suggested two preliminary significant differences. First, the words similarly distributed with *mu zhu* shown in the word vector of *mu zhu jiao* indicated a significant semantic consistency with gender-biased language, while the demonstration of the contrasting representation was comparatively neutral without connecting with specific contexts. Second, some famous catchphrases in terms of the abusive language on PTT were presented as similarly appeared word in the word vectors from *mu zhu jiao*. However, the evidence seemed not to imply any connection between the represented words and the implicit implication in word vector from general *xiang min.* That is, the representation of the distributed words from general *xiang min* appeared not to link to any specific context.

Mu zhu, in the configuration of the speech from *mu zhu jiao*, had no doubt serving as the most essential core, assembling its multi-valency on bullying female. In order to compare different semantic implication under the context of two comparing representation, the corresponding concordances of *mu zhu* from two word vectors were extracted. From the concordances, a preliminary difference of semantic implication of *mu zhu* between the two groups of *xiang min* appeared to be detected. In every situation, general *xiang min* talked about *mu zhu* theoretically under the context of breeding pigs, such as the way a boar started to breed a sexually receptive gilt. By contrast, *xiang min* from *mu zhu jiao* strongly criticized *mu zhu* a lot, comparing *mu zhu* to women who have specific characteristics, like double standards, preferring Cross Culture Romance (CCR), worshiping money, and other features. It is noteworthy that some features were judged so hard under this context while

they have no bad essence at all. Take CCR for example, *xiang min* from *mu zhu jiao* considered Taiwanese female having a relationship with a foreigner (especially Caucasian) as a heinous behavior. However, a Taiwanese male would be priding himself if he is dating a Caucasian female. To sum up, the evidence suggested that the speech from *mu zhu jiao* to some extent testified itself as an abusive language on the basis of showing the represented offensive words alluded to gender discrimination. After all, a couple of inferences of *mu zhu* were drawn as a network trying to specify the related semantic implications (Figure 1).

Figure 1: The network of the related semantic implications of *mu zhu* on PTT forum

On the other hand, in contrast to *mu zhu*, this paper also reported the concordances of the counterpart *gong zhu (male pig)* from two word vectors to see if the literally contrasting words *gong zhu (male pig)* are being used as much as *mu zhu (female pig)* in common usage. It turned out that the concordances of *gong zhu* from the discourse from general *xiang min* were exactly the same the concordances of mu zhu, which meant there is little additional use among them. In comparison, *gong zhu* from the discourse from general *xiang min* appeared twice in the overall texts, primarily arguing the issue regarding the existence of "gong zhu jiao (a cult from gong zhu)". In brief, the concordances of the core word *mu zhu* explained its multi-valency from the perspective of language usage on PTT. By contrast, the concordance of *gong zhu* helped to specify that the metaphorically abusive expression of the pig was originated from *mu zhu* and there was not enough evidence to show the gender equality with female in usage. Besides, compared to the word vector of general *xiang min*, the

concordances of both *mu zhu* and *gong zhu* in the word vector of *mu zhu jiao* provided solid evidence of *mu zhu*'s strong context dependency.

4.2 Keywords Polarity and Knowledge-based Features

Abundant semantic implications of the represented words had the tendency to be sentimentally negative. However, the sexually biased hate speech on PTT could be completely understood largely based on related knowledge of misogynistic expressions on PTT because it often employed context-dependent metaphors. In other words, people who didn't engage in PTT would probably have difficulty understanding the implicit meanings from the misogynistic subculture. In order to know how the polarity of the related speech varied from people happened to have the related knowledge, in this subsection, two groups of people were invited to evaluate the polarity of the "keywords" of the abusive speech on PTT: thirty senior *xiang min* who have logged in PTT more than 3000 times and thirty people who are not *xiang min*. 25 keywords were the terms manually selected under human determination to specify their high context-dependency from the speech from m*u zhu jiao*. They might contain words, catchphrases, fixed constructions which often employed metaphoric expressions so as to implicitly identify the intrinsic biased nature. In terms of this questionnaire, the higher the polarity score is, the positive the keyword is, while the lower the polarity score is, the negative the keyword is. The polarity tendency is calculated from 1 (the lowest) to 5 (the highest).

Afterward, the average polarity scores of the 25 keywords from two groups were calculated, rounding to one decimal place. It was not surprising that, 72 % (18 out of 25 terms) of the scores from non *xiang min*, suggested to be lower (from average 0.1 point to 1.2 points) than those from senior *xiang min*. Also, the average polarity score from senior *xiang min* (average 2.16) is 0.24 points lower than which of non *xiang min* (average 2.4). The finding indicated that the perception of particular ideology could be strongly influenced by language use. In this case, it would be more difficult for people without the knowledge of the subculture to perceive the implicit information than people with that knowledge. This situation, in some circumstances, proved the knowledge-based nature of the abusive speech on PTT.

5 Conclusion

This study examined the word representations of biased speech on PTT forum with the application of NLP tools to detect the abusive language online by comparing the distributed representations of

words represented by a word vector encoding semantic similarity and capturing significant linguistic properties from the word vectors. Also, this study took advantage of questionnaires to see the fluctuation of the polarity of the word representation between the party causing injury and others. The findings showed that the distributed semantic representations of *mu zhu jiao* and those of general *xiang min* differed significantly in terms of several perspectives: First, the distributed representations of words represented by a word vector encoding semantic similarity. Second, the keyword concordance and the encoding semantic implications. Third, the related semantic network of the keyword *mu zhu*. Last, the polarity implementation of the related words of the keyword *mu zhu.* The overall results indicated that the distribution of the surrounding words of *mu zhu* tended to display sexually biased semantic implications, making the word vector of *mu zhu jiao* a blatantly sexist.

This study served as the very first paper analyzing the abusive language on PTT with an NLP method by taking advantage of both quantitative and qualitative methods. However, it should be noted that this study has been primarily concerned with posts originating from the popes of *mu zhu jiao* because of the difficulty of defining who really are its followers, and has only addressed some features to detect sexually biased cyberhate speech on PTT instead of developing a well-established computational model to automatically detect the abusive language. Despite its preliminary nature, the study may offer some insight into the gender bias amplification of the distribution of word representations on PTT. To further capture the subtle and highly context-dependent abusive speech in an efficient and scalable computational method, future research may use a text processor to capture the characteristics of biased language while explicitly seen as neutral speech. Hopefully, the embedding model could be further modified to remove gender stereotypes and biases from training data.

References

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349-4357).

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510-526.

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 11.

Citron, D. K. (2011). Misogynistic Cyber Hate Speech.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Gilmore, D. D. (2010). *Misogyny: The male malady*. University of Pennsylvania Press.

Grus, J. (2015). *Data science from scratch: first principles with python*. O'Reilly Media, Inc.

Heuer, H. (2016). Text comparison using word vector representations and dimensionality reduction. *arXiv preprint arXiv:1607.00534*.

Jane, E. A. (2016). Online misogyny and feminist degilantism. Continuum: Journal of Media & Cultural Studies.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London: Pearson.

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

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

Pagano, R. R. (2012). *Understanding statistics in the behavioral sciences*. Cengage Learning.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33-53.

Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10).

余貞誼. (2016). 我說妳是妳就是：從 PTT 母豬教的仇女行動談網路性霸凌的性別階層. *婦研縱橫*, (105), 22-29. [Yu, Zhen-Yi. (2016). Wo shuo ni shi ni jiu shi: Cong PTT muzhujiao de chounu xingdong tan wanglu xing baling de xingbie jieceng. Fu yan zong heng

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 202-211

植基於深度學習假新聞人工智慧偵測：台灣真實資料實作

Deep Learning Based Fake News AI Detection：Evidence From Taiwan News Report

汪志堅　Chih-Chien Wang
國立台北大學資訊管理研究所
Graduate Institute of Information Management
National Taipei University
wangson@mail.ntpu.edu.tw

戴敏育　Min-Yuh Day
淡江大學資訊管理學系
Department of Information Management
Tamkang University
myday@mail.tku.edu.tw

胡林赟　Lin-Lung Hu
國立台北大學資訊管理研究所
Graduate Institute of Information Management
National Taipei University
s710636103@gm.ntpu.edu.tw

摘要

近年來透過網路媒介，使得假新聞得以快速流竄。而許多國家都受到假新聞嚴重影響，使得假新聞偵測變成一個重要課題。本研究蒐集兩個台灣闢謠網站資料，「新聞小幫手」以及「真的假的」。並運用深度學習的三種方法 Gated Recurrent Unit (GRU)、Long Short Term Memory (LSTM) 以及 Bidirectional Long Short Term Memory (BLSTM)，進行假新聞偵測。實驗結果發現，深度學習可以運用在台灣假新聞偵測當中，且 BLSTM 的方法最適合用於假新聞偵測。本研究實驗降低假新聞比例，模擬 25%以及 5%假新聞比例，讓研究樣本可以更趨近於真實情況。最後本研究使用交叉資料集測試，了解實務與學術上的差距。

Abstract

In recent years, because the Internet acts as a medium, fake news can be quickly spread. Many countries have been seriously affected by fake news. making fake news detection become an important issue. This study collects two Taiwan refute rumors sites. And use the three methods of deep learning for fake news detection., Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM), and Bidirectional Long Short Term Memory (BLSTM). The experimental results show that deep learning can be used in Taiwan's fake news detection, and the BLSTM method works best. Research experiments reduce the proportion of fake news, simulating 25%

and 5% fake news ratios. Let the research sample be closer to the real situation. Finally, this study used a cross-data set test to understand the gap between practice and theory.

關鍵詞：假新聞、假新聞偵測、深度學習、人工智慧、雙向長短期記憶

Keywords: Fake News, Fake News Detection, Deep Learning, Artificial Intelligence, Bidirectional Long Short Term Memory

一、緒論

假新聞的定義為有意撰寫且誤導新聞讀者，但驗證是虛假的新聞[1]。近年來各國都受到假新聞嚴重影響，例如：2016 美國總統大選、台灣衛生紙之亂、台灣關西機場假新聞事件等等[2]。近年的假新聞都是透過網路為媒介。而要檢測網路上的內容的真偽，最原始的方式是透過「手動」方式[3]，使得 Facebook 與 Google 兩家網路巨頭公司都想讓假新聞偵測自動化。

過往多位學者曾應用機器學習或者深度學習來進行假新聞偵測，如：Support Vector Machines(SVM) 、 Naive Bayes 、 Convolutional Neural Network(CNN) 、 Logistic Regression(LR)、Neuro Linguistic Programming(NLP)、Recurrent Neural Network(RNN)、Gated Recurrent Unit(GRU)、Long Short-Term Memory(LSTM)。本研究也嘗試使用台灣的假新聞資料，探究以深度學習方式偵測假新聞的可能性。

二、文獻探討

深度神經網路(Deep Neural Network)，是多層的神經網路，可透過電腦找出特徵值，經由深入學習後，讓預測結果更有效，目前主要的神經網路模型有 20 幾種，其中，遞迴神經網路(Recurrent Neural Network, RNN)可用於解決帶有順序性的問題，例如: 自然語言處理、語音辨識、手寫辨識等等[4]。

Hochreiter and Schmidhuber [5]提出長短期記憶(Long Short-Term Memory, LSTM)，是為了解決 RNN 所產生梯度消失及梯度爆炸而產生。LSTM 與 RNN 相比多了 3 個控制器，分別為輸入閘門(Input Gate)、輸出閘門(Output Gate)、忘記閘門(Forget Gate)，當有了控制器閘門的機制 LSTM 就能夠將記憶長期記住。

雙向長短期記憶(Bidirectional- Long Short-Term Memory, BLSTM)在 2005 年由 Graves and Schmidhuber [6]提出來，他是由前向長短期記憶與後向長短期記憶結合而成，是為

了解決後向修飾前向的問題。

閘門循環單元(Gate Recurrent Unit, GRU)與長短期記憶類似。根據 Chung, et al. [7]研究指出閘門循環單元比起長短期記憶會使用更少的記憶、減少中央處理器(CPU)的運算時間以及加快收斂的速度，而閘門循環單元在較小的資料集的表現比長短期記憶來的更加優異。

(二) 假新聞資料庫蒐集

過去假新聞偵測研究中，大多數研究都是根據專家檢核、研究自行蒐集、群眾闢謠網站、社群網站提供以及競賽提供，五種方法所建立起資料庫。例如：BuzzFeed 記者 Silverman, et al. [8]對針對政治相關 Facebook 粉絲專頁，所發的文章做事實檢核、Wang [9]與加拿大維多利亞大學，透過 PolitiFact 事實檢核網站所蒐集的資料庫、The FEVER 1.0 Shared Task 競賽所提供 185,445 筆資料、Pérez-Rosas, et al. [10]透過 GossipCop 八卦檢核網站、Ma, et al. [11].微博(Weibo)管理中心提供假新聞報告資料集、Svärd and Rumman [12]自行蒐集 201 篇美國新聞文章，其中有 120 是假新聞，81 篇是真新聞、Pérez-Rosas, et al. [10]自行蒐集假新聞，並請外包公司改寫。近年來台灣開始有事實檢核網站以及群眾闢謠網站的出現，但仍少有相關的研究成果被提出。

(三) 應用深度學習假新聞偵測研究

過去假新聞偵測使用的特徵大多是使用文字特徵，並使用機器學習的方式進行偵測，例如：使用 Support Vector Machines(SVM)、Unsupervised、Naive Bayes。而因深度學習發展優異，也開始有研究將深度學習應用在假新聞偵測中。不過，在台灣，這方面的研究成果仍較少被提出。本研究整理出過去研究所使用深度學習的假新聞偵測研究：

三、研究方法

(一) 資料集

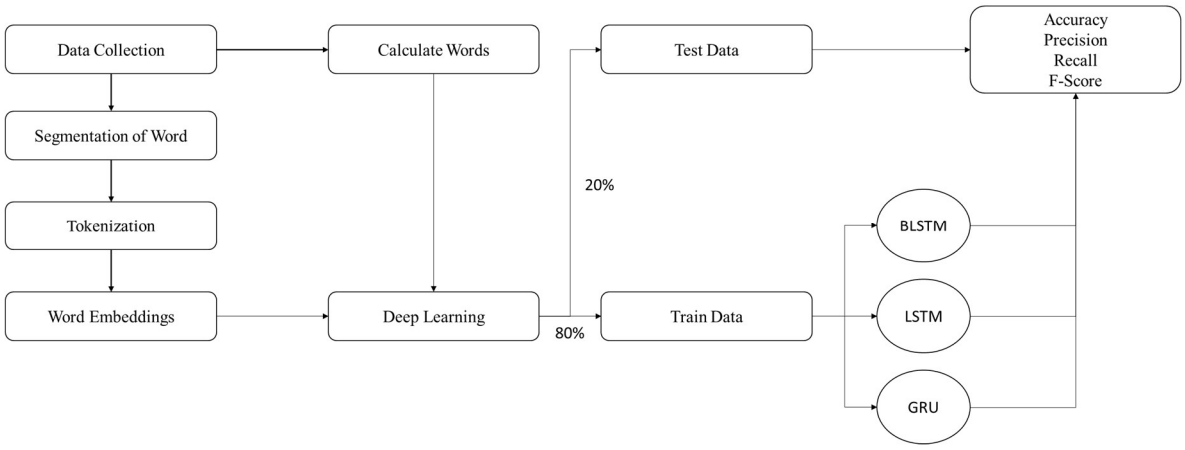
本研究蒐集資料集是台灣組織「g0v.tw 台灣零時政府」，所創立的兩個專案：

「Cofacts 真的假的」以及「新聞小幫手」。「Cofacts 真的假的」透過 Line Chatbot 來蒐集社群上流傳的假新聞，在藉由群眾舉報然後在群眾核對，讓大家一起打擊網路上的謠言以及假新聞。本研究將資料存至 SQL Server，至截稿為止資料庫真新聞有 6,107 筆資料，假新聞有 9,141 筆資料。

「新聞小幫手」是一個闢謠的網站，主要是讓民眾對於可疑新聞來此詢問，由其他民眾蒐集資料後闢謠，但目前於 2018/10/24 停止服務，但資料依舊會保留在網站上面。因「新聞小幫手」沒有真新聞的標記資料，本研究蒐集了真新聞資料，主要來源是蘋果日報，再透過相似度分析將以被舉報假新聞的資料給剔除。至截稿為止真新聞的資料一共蒐集了 11,015 筆，假新聞的資料一共有 5,795 筆，經資料整理後，使用了 2,014 筆。

(二) 研究流程

本研究先將兩個資料集蒐集後，進行文字的預處理。將中文字先進行斷詞後，在將中文字進行 Tokenization 以及文字的詞數計算，以便深度學習處理。然後將預處理資料以 80-20 法則分成訓練集與測試集。本研究實驗三種深度學習的方法，Gated Recurrent Unit (GRU)、Long Short Term Memory (LSTM) 以及 Bidirectional Long Short Term Memory (BLSTM)。然後使用 Accuracy、Precision、Recall 及 F-Score 對模型進行評估。最後將此上述經過繪製成以下流程圖：



圖一、研究流程圖

(三) 深度學習架構

本研究會將三個資料集都以 80-20 法則分成 80%訓練集與 20%的測試集，在讀取資料後建立出 Token 字典，字典依據出現次數的多寡已降冪方式排序，接下來使用 Token 把新聞文字轉成數字陣列。由於新聞字數並不一定，為了讓轉換後的字數相同。本研究將採用截長補短的方式，利用前面所計算的詞數，設定一固定長度為 n，若新聞文字長度大於 n，則截去後面的數字;反之如果新聞文字若小於 n，則會補 0 直到 n 的長度。之後便加入 BLSTM、LSTM 與 GRU 模型，先建立出線性堆疊模型，然後再加入平坦層、隱藏層、輸出層，再以五種不同大小的 Dropout、三種不同的損失函數、三種不同的激活函數以及三種不同的優化器，總共 135 個排列組合，找出最佳組合的模型。在資料經過一連串迭代之後，使 Accuracy、Precision、Recall、F-score 檢測模型。

四、實驗結果

(一) 訓練與測試筆數

本研究將真新聞與假新聞比例成 1:1，假新聞的比例將會佔 50%。接著以 80-20 法則分成訓練與測試筆數，下表為「新聞小幫手」與「真的假的」訓練與測試筆數：

表二、假新聞比例 50%訓練與測試筆數

資料集	訓練筆數 (假：真)	測試筆數 (假：真)
新聞小幫手	1611:1611	403:403
真的假的	4886:4886	1221:1221

(二) 深度學習結果

本研究將實驗中得出的三種方法最佳配置如表二所示。最後以 Accuracy、Precision、Recall 以及 F-Score 對各種模型做評估如表三所示：

表三、50%假新聞比例最佳配置

資料集	方法	激活 函數	損失 函數	優化器	Dropout
新聞 小幫手	GRU	sigmoid	MSLE	Adam	0.6
	LSTM	sigmoid	binary_crossentropy	AdaMax	0.6
	BLSTM	relu	MSE	RMSprop	0.5
真的假的	GRU	sigmoid	binary_crossentropy	Adam	0.4
	LSTM	sigmoid	binary_crossentropy	RMSprop	0.4
	BLSTM	relu	MSLE	AdaMax	0.5

表四、50%假新聞比例模型評估

資料集	方法	Loss	Accuracy	Precision	Recall	F-Score
新聞 小幫手	GRU	0.053	0.881	0.885	0.876	0.88
	LSTM	0.045	0.888	0.881	0.898	0.889
	BLSTM	0.091	0.898	0.898	0.898	0.898
真的假的	GRU	0.104	0.677	0.635	0.836	0.722
	LSTM	0.108	0.685	0.641	0.844	0.728
	BLSTM	0.218	0.683	0.638	0.846	0.728

(三) 不同假新比例結果

本研究認為假新聞佔樣本 50%並不符合真實情況，本研究又實驗了假新聞比例 25%以及 5%，並照一樣的研究方法來實驗。本研究在實驗「新聞小幫手」時發現，在 5%假新聞比例之下，可以再增加資料筆數。所以本研究又實驗了在相同假新聞比例在 5%情況下，會有什麼樣的結果。以下是本研究訓練與測試筆數以及實驗結果：

表五、假新聞比例 25%與 5%訓練與測試筆數

資料集	假新聞比率	訓練筆數(假：真)	測試筆數(假：真)
新聞小幫手	25%	1600:4800	400:1200
	5%	252:4800	63:1200
	5%*	421:8000	105:2000
真的假的	25%	1600:4800	400:1200
	5%	252:4800	63:1200

表六、假新聞比例 25%與 5%最佳配置

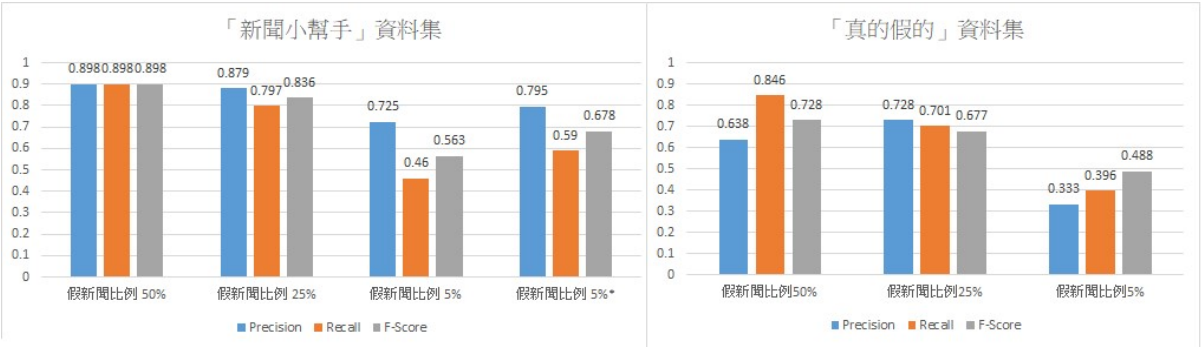
資料集	假新聞比例	最佳方法	激活函數	損失函數	優化器	Dropout
新聞小幫手	25%	BLSTM	sigmoid	binary_crossentropy	Adam	0.3
	5%	BLSTM	tanh	binary_crossentropy	Adam	0.4
	5%*	GRU	Tanh	MSLE	Adam	0.6
真的假的	25%	BLSTM	relu	binary_crossentropy	AdaMax	0.4
	5%	LSTM	tanh	binary_crossentropy	AdaMax	0.6

表七、假新聞比例 25%與 5%實驗結果

資料集	最佳方法	假新聞比例	Loss	Accuracy	Precision	Recall	F-Score
新聞小幫手	BLSTM	25%	0.053	0.922	0.879	0.797	0.836
	BLSTM	5%	0.045	0.964	0.725	0.46	0.563
	GRU	5%*	0.091	0.972	0.795	0.59	0.678
真的假的	BLSTM	25%	0.134	0.845	0.728	0.701	0.677
	LSTM	5%	0.616	0.948	0.333	0.396	0.488

最後以圖二可以發現，雖然降低假新聞比例，可以讓樣本可以更趨近於真實情況，但可以明顯發現 Precision、Recall 以及 F-Score 下降。甚至在「真的假的」在 5%假新聞比例四個指標皆未達到基準線。而在「新聞小幫手」的假新聞 5%實驗當中可以發現，相

同比例下增加訓練與測試筆數，可以增加模型的效率。



圖二、不同假新聞比例效率統計圖

五、結論

本研究實驗兩個中文資料集，發現深度學習確實可以用於偵測假新聞。本研究也模擬了 3 種假新聞比例的實驗，50%、25%以及 50%。在「新聞小幫手」資料集當中，50%的 F-Score 來到 0.898，25%的 F-Score 來到 0.836 以及 5%的 F-Score 來到 0.563。本研究又實驗了將 5%的訓練與測試比數拉高，5%的 F-Score 從原本的 0.563 增長到 0.678 發現訓練與資料測試筆數增加，可以讓深度學習的效率更好。在「真的假的」資料集當中 50%的 F-Score 在來到 0.728，25%的 F-Score 來到 0.701，而 5%的 F-Score 只有 0.396。相較於「新聞小幫手」資料集，「真的假的」資料集效率較差，甚至在 5%假新聞比例完全無法偵測。本研究認為「真的假的」文章長度都短於「新聞小幫手」，資料集筆數不足導致深度學習的效率不佳。

本研究實驗了三種方法，BLSTM、LSTM 以及 GRU，發現大多數的情況下，BLSTM 效果都是最佳。除了在「新聞小幫手」5%假新聞比例的 GRU 效果最好以及「真的假的」5%假新聞比例無法偵測之外，尚餘 5 種皆是 BLSTM 方法效果最佳。

而除此之外，本研究又為了進一步實驗將「新聞小幫手」做交叉資料集測試，發現如要將模型實務所需的發現。

在 50%、25%與 5%模型測試中也發現一樣的結果，Recall 也偏高在 0.93 以及 0.87，這也發現 50%、25%的模型也能夠清楚了解假新聞脈絡，並且成功預測正確，但一樣對多種真新聞來源脈絡很不熟悉，常常真新聞錯當成假新聞。這說明了在訓練的資料

集當中，真新聞的來源並不夠多，導致模型常將真新聞錯當成假新聞。而在 5%的模型當中 Recall 有大幅度下降到 0.4，本研究認為是 5%所放入假新聞減少，導致準確抓出假新聞下降。

綜上所述，如要將模型進一步的實務化，除了假新聞資料蒐集外，必須要蒐集更多來源的真新聞，讓模型在面對各種不同真新聞之時，不會誤將新聞判別成為假新聞。

(三) 研究限制

本研究認為目前假新聞偵測最大限制還是假新聞的資料不足，無法確認此資料集是否能代表整個真實世界的狀況。資料集當中真新聞是否還有假新聞在其中或者假新聞當中有真新聞被誤報為假新聞，以及假新聞比例在真實資料集當中比例為何。本研究認為深度學習已經相當成熟，可以運用在假新聞偵測當中，但最大限制還是語料庫不足的問題，本研究認為假新聞偵測需要有更進一步的分類，以及更多標籤(Label)以及更多特徵(Feature)去標記。本研究認為可以再進一步嘗試蒐集更多真新聞，嘗試更低的假新聞比例，可以讓模型更能運用在實務上。相信有天假新聞偵測有天可以運用在各種裝置上，避免導致訊息接受者被誤導，造成社會的恐慌。

參考文獻

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives,* vol. 31, no. 2, pp. 211-36, 2017.

[2] 汪志堅、陳才, *假新聞：來源、樣態與因應策略.* 台北: 前程文化, 2019.

[3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter,* vol. 19, no. 1, pp. 22-36, 2017.

[4] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013: IEEE, pp. 6645-6649.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[6] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks,* vol. 18, no. 5-6, pp. 602-610, 2005.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated

recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555,* 2014.

[8] C. Silverman, L. Strapagiel, H. Shaba, Ellei Hall, and J. Singer-Vin. "Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate." buzzfeednews. https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis (accessed.

[9] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648,* 2017.

[10] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *arXiv preprint arXiv:1708.07104,* 2017.

[11] J. Ma *et al.*, "Detecting rumors from microblogs with recurrent neural networks," in *Ijcai*, 2016, pp. 3818-3824.

[12] M. Svärd and P. Rumman, "COMBATING DISINFORMATION: Detecting fake news with linguistic models and classification algorithms," ed, 2017.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 212-225

使用生成對抗網路於強健式自動語音辨識的應用

Exploiting Generative Adversarial Network for Robustness Automatic Speech Recognition

楊明璋 Ming-Jhang Yang, 趙福安 Fu-An Chao, 羅天宏 Tien-Hong Lo,
陳柏琳 Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{60547076S, 60747002S, teinhonglo, berlin}@ntnu.edu.tw

摘要

在過去幾年中，深度學習技術的發展在許多領域中大放異彩，應用在語音辨識中也一樣表現優異。儘管語音辨識有了大幅度的改進，然而「雜訊」仍然一定程度的干擾語音辨識之準確度。諸如:背景人聲、火車、公車站牌、汽車噪音、餐館背景雜音…以上皆為易影響語音辨識結果的環境噪音。因此語音辨識的強健性技術研究仍扮演著重要角色。過往於強健性技術的研究主要可區分為以特徵為基礎，以及以模型為基礎兩大面向。以特徵為基礎的強健性技術又可分為特徵正規化以及語音訊號增益。本研究主要採用生成對抗網路(Generative Adversarial Network , GAN)以語音訊號增益方式使用在調變頻譜特徵上。我們的目的在於把受到吵雜環境干擾，或被通道效應破壞的語音特徵轉換成接近乾淨環境下錄製之語音特徵，此方法比起原始梅爾倒頻譜係數特徵可以有效的提升辨識率。

Abstract

In the recent past, deep learning techniques have reached record-breaking performance in a wild variety of applications like automatic speech recognition (ASR). Even though cutting-edge ASR systems evaluated on a few benchmark tasks have already reached human-like

performance, they, in reality, are not robust, in the manner that humans are, to disparate types of environmental noise such as babble, train, bus station, car driving, restaurant, and among others. In view of this, this paper embarks on an effort to develop effective enhancement methods, stemming from the so-called generative adversarial networks (GAN), for use in the modulation domain of speech feature vector sequences. A series of experiments conducted on the Aurora-4 database and task seem to demonstrate the practical merits of our methods.

關鍵詞：生成對抗網路、語音訊號增益、語音強健性技術、強健性語音辨識

Keywords: Generative Adversarial Network, Speech Enhancement, Robustness Techniques, Robust Speech Recognition

一、緒論

在自動語音辨識技術（Automatic Speech Recognition, ASR）的發展中，我們發現環境噪音會大幅度的影響辨識率。因此，為了降低噪聲的影響以及提升辨識率，強健性語音辨識的發展便應運而生。目前強健性語音辨識技術大致可以分為特徵為基礎(Feature-Based)以及模型為基礎之方法(Model-Based)兩種。前者著重於特徵正規化(Feature Normalization)以及特徵增益(Feature Enhancement)兩方面。後者則主要專注在改良聲學模型上，將其「加深」、「加廣」以及其他特殊訓練方法用於提升語音辨識的強健性效果。

時至今日，已有多種新穎之強健性技術可以為語音辨識帶來更好表現，其中在特徵處理方法中有多項採用調變頻譜分析的研究指出在頻率較低之 4Hz 附近存在諸多語意資訊 [1]，而這將有助於提升語音辨識的效果。因此諸多調變頻譜正規化的研究便由此而生 [2] [3] [4]。從以上研究得到啟發，學者發現探索語音特徵的子空間結構可以得到更佳語音辨識效果 [5]，故萌生了子空間學習的概念。依循著這一個脈絡，目前子空間學習已發展出: 字典學習法結合稀疏編碼(Sparse Coding) 、低序表示法(Low Rank Representation, LRR) 等主要方法。

除了語音特徵正規化之外，語音訊號增益法(Speech Enhancement)亦是一種強健性方法。部分採用此作法的研究直接針對語句之波形圖濾波。另一部份則採用深度學習技術生成接近無干擾的語音特徵，例如導入自動編碼器 [6] 用來抑制噪聲干擾。做為新起之秀的生成對抗網路(Generative Adversarial Networks, GAN) [7]也可以作為語音訊號增益的手段，其自動生成與鑑別是否正確的功能在增益特徵強健性上被認為有助益。雖然最初設計是用於影像處理，但是目前已有多種變形應用於語音強健性研究上，著名研究有 SEGAN[8]、Whispered-to-voiced GAN[9]、RSRGAN[10]以及 FSEGAN [11]。這類方法除了可以處理波形圖外亦可處理時域(Time Domain)特徵和頻率域(Frequency Domain)特徵。本研究即是從上述生成對抗網路方法中得到啟發，導入風格轉換概念，並結合調變頻譜相關研究想法，以生成對抗網路處理頻率域特徵，關於詳細方法將於後續章節依序介紹。我們發現使用生成對抗網路處理調變頻譜特徵，可以使語句的特徵分布更接近未受干擾語句，從而提升語音辨識效果。

二、文獻回顧

以處理語音特徵為基礎的強健性技術，目的在於不需要重新設計聲學模型，透過語音訊號增益、特徵向量補償、頻譜補償或正規化等方式還原出乾淨完整的語音特徵。語音訊號增益(Speech Enhancement)技術之目的在於增強語音訊號的可讀性和品質，以便在包含雜訊的情況下可以順利被聽懂亦可用於進行語音辨識 [12]。遮罩與濾波技術是一種很直觀的訊號增益方法，維納濾波就是著名的方法之一。維納濾波器之概念於 1949 正式出版於數學家諾伯特·維納（Norbert Wiener）的著作中 [13]，是一種採用最小化平均誤差(Mean Square Error, MSE)當作最佳化函數的線性濾波器(Linear Filter)，也就是說:在給定約束條件下計算濾波器輸出與期望的輸出之間的平方誤差之最小值，便是維納濾波器的核心概念。簡而言之:目的在於使得經過濾波後的訊號能盡量的接近未受干擾的真實訊號，主要運算可以由方程式(1)表示，其中 ***s(t)***是我們要估計的原始訊號，***n(t)***代表疊加的雜訊，輸入訊號由 ***s(t)***和 ***n(t)***組成和濾波器 ***g(t)***進行摺積運算後得到濾波後的訊號 ***x(t)***:

$$\boldsymbol{x}(\mathrm{t}) = \boldsymbol{g}(\mathrm{t}) * \big(\boldsymbol{s}(\mathrm{t}) + \boldsymbol{n}(\mathrm{t})\big), \tag{1}$$

隨著時光荏苒，在諾伯特·維納之後又有諸多學者以濾波器為題，提出多種可以有效增益語音訊號之遮罩方法 [14] [15]。直接遮罩(Direct Masking) [16]，很直觀的採用$\lambda_{t,f}^{(S)}$作為一個遮罩用來增益嘈雜語音訊號$Y_{t,f}$，經過 [34]改寫，增益後的訊號可以由下列方程式表示:

$$\hat{S}_{t,f} = f_{\theta DM}^{(DM)}(Y,t,f) = \lambda_{t,f}^{(S)} \ dot Y_{t,f}, \tag{2}$$

然而直覺遮罩往往會帶來大量失真，因此便有學者採用由維納濾波器改良成的帶參數維納濾波(Parametric Wiener Filter, PW)來解決這個問題，PW 在抑制噪訊與控制失真這兩者的權衡之間擁有更多靈活性 [17] [18] [19]，PW 具體運算可表示如下:

$$\hat{S}_{t,f} = f_{\theta pw}^{pw}(Y,t,f) = \left|\frac{|Y_{t,f}|^p - l \ .|\widehat{N}_{t,f}|^p}{|Y_{t,f}|^p}\right|^{1/q}.Y_{t,f}, \tag{3}$$

另一方面，除了濾波與遮罩法之外，近年導入深度學習技術，也為語音訊號增益帶來新的想法。我們知道神經網路可以學習與調整一群資料的分布，因此深度學習技術在語音訊號增益中的用途大多以映射為主，將嘈雜訊號映射成乾淨語句的分布便可以達到我們的目的。以下讓我們回顧經幾個經典作法，雖然他們的目的在於消除殘響(Reverberation)，但是仍然對我們想要抑制疊加雜訊干擾以及抑制通道摺積效應有莫大啟發。

以深度遞歸神經網路(Regression Neural Network)做為主要結構 [20]，首先採用乾淨語句提取出的對數功率譜(Log-Power Spectrum, LPS)當作特徵，用以訓練深度學習模型。如此一來模型將學習到乾淨語句的特徵分布，接著輸入嘈雜語句的特徵便可以輸出增益

過後接近原始乾淨語句的語音特徵，再由此重構出波形圖及原始聲音，就可以得到還原後的語句。 [21]之研究發現將語音特徵由 LPS 映射到 MFCC，可以更進一步改善語音辨識結果，由此現象可以得知深度學習技術也可以學習不同特徵之間轉換的對應關係。

上述研究多關注在以深度神經網路生成特徵，但是其估算出的特徵是否夠接近我們的預期呢?這點除了在整體計算完畢後進一步評估語音訊號品質如(Perceptual Evaluation of Speech Quality, PESQ)或是語音辨識結果之詞錯誤率(Word Error Rate, WER%)之外，我們在訓練的同時並無從得知，也無法同時最佳化模型參數，因此強健性效果有可能會大打折扣。此時，導入生成對抗網路中鑑別器(Discriminator)的概念便是一個新的契機，將原本用來生成特徵的模型當作生成器，同時另外訓練一個網路當作鑑別器，由鑑別的結果自動更新生成器的參數，就可以更周全的訓練模型。

三、生成對抗網路應用於語音強健性技術

本研究主要採用生成對抗網路(Generative Adversarial Network, GAN)結合調變頻譜特徵進行語音訊號增益處理，並結合自動語音辨識(Automatic Speech Recognition, ASR)用來達成增進強健性表現之目的。本節將針對本研究使用的生成對抗網路模型與方法進行討論。

生成對抗網路是一種可以減少人類知識介入，而得到更佳學習效果的一種深度學系技術，這項關鍵就在於「生成」與「鑑別」，也有人稱之「新手畫家」 與「鑑賞家」。在訓練過程中，新手畫家不斷臨摹名畫，而鑑賞家持續鑑定畫作，在兩造交手若干次之後，新手畫家有了弄假成真的本事，而鑑賞家漸漸分不出贋品與真品，這便是我們的目的。這樣子對抗學習的過程可以看成是生成器(G)與鑑別器(D)在進行最大與最小值的對局(Minimax-Game)。本研究採用 LSGAN 中的方均根誤差(Mean Square Error, MSE)當作損失函數，在訓練過程中我們需要將 G 與 D 串接起來，因此整體目標函數可以改寫成:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[\log D(\boldsymbol{x})] + \mathbb{E}_{\hat{x}\sim P_{\hat{x}}(\hat{x}),z\sim P_z(z)}\left[log\left(1-D\left(G(\hat{\boldsymbol{x}}+\boldsymbol{z})\right)\right)\right], \quad (4)$$

我們用 $\boldsymbol{x}$ 表示乾淨情境樣本(Clean Condition)，用$\hat{\boldsymbol{x}}$表示噪訊情境樣本(Noisy/Multi Condition)，再這裡 $\boldsymbol{z}$ 為一組 Latent Vector，將其設為介於 0 到 1 之間的隨機雜訊。於訓練時將噪訊樣本加上隨機雜訊輸入 G 使其盡可能有能力把雜亂資料轉換成我們期望的乾淨樣本。

一個完整的生成對抗網路之訓練步驟主要可以分為三個階段: (1)訓練鑑別器認識真實樣本(2)訓練鑑別器認識生成器生成之假樣本(3)固定鑑別器的參數，同時更新生成器參數以達成訓練目標。

圖 1: 生成對抗網路-鑑別器訓練-1

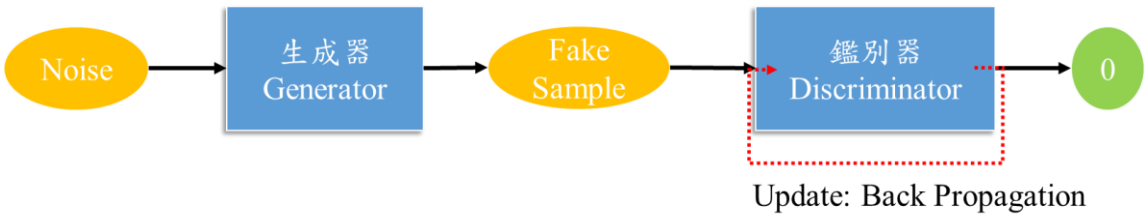


圖 2: 生成對抗網路-鑑別器訓練-2

圖 3: 生成對抗網路-生成器訓練

以上為 GAN 之訓練方法及其示意圖，接著本段將介紹本研究使用之神經網路架構。我們採用兩種不同結構，分別命名為 CAGAN 以及 DNN-LSGAN，前者採用類似於摺積自動編碼器(Convolution Auto Encoder, CAE)作為生成器的主要結構，後者則採用全連接 DNN 作為生成器的主結構，並以此為原則命名。啟發於多項類似於自動編碼器(Auto Encoder, AE)與 GAN 結合的研究[29][22]，加上目前普遍認為 CNN 在學習時間-頻率特徵或圖像的能力比起 DNN 還有更好效果。而我們以調變頻譜特徵作為輸入，在頻率域上進行訊號增益，概念類似電腦視覺領域中於影像降噪的研究，也就是說我們以摺積運算結合自動編碼器當作一項取得強健性特徵的方法。

在消除噪訊干擾效應的深度學習技術中，降噪自動編碼器(Denoise Autoencoder, DAE)與摺積自動編碼器(Convolutional Autoencoder, CAE)是很有效方法。可以輸入被噪訊破壞的原始資料，並還原出未受干擾的資料。而本研究生成對抗網路方法之一就是受到他們啟發，CAGAN 之生成器就是類似於摺積自動編碼器的結構。

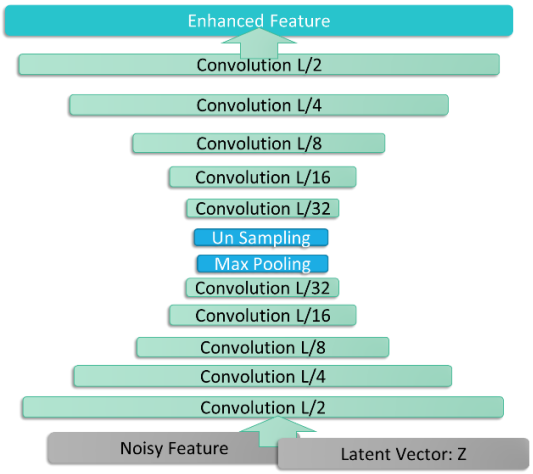


圖 4: CAGAN-生成器結構

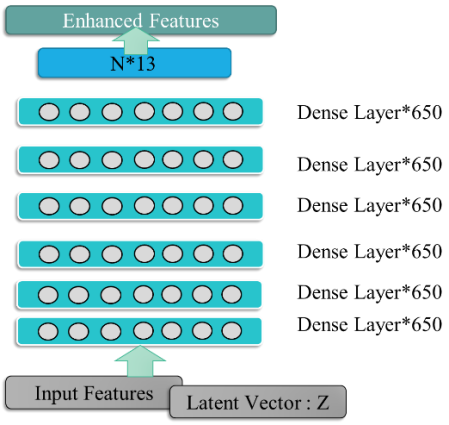


圖 5: DNN-LSGAN-生成器結構

我們知道自動編碼器可以分成編碼階段與解碼階段兩大部分。在編碼階段，隨著深度增加，我們將摺積層的特徵圖(Feature Map)大小減半，並在每一次摺積運算後進行池化(Max Pooling)，目的在於將有效的特徵往下傳遞並且減少不必要的網路參數，使之更方便訓練。在解碼階段，其結構可以視為將編碼階段水平鏡射的對稱關係，唯一不同處在於相對於最大池化法(Max Pooling)，我們在每一次摺積運算之後採用反取樣法(Up-Sampling)，將維度還原成原始大以利進行後續語音辨識步驟。在訓練時我們採用嘈雜環境語料結合隨機雜訊作為輸入資料，其結構圖 4 所示。

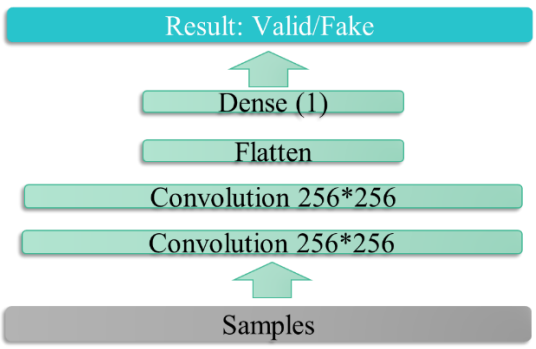


圖 6: CAGAN 之鑑別器

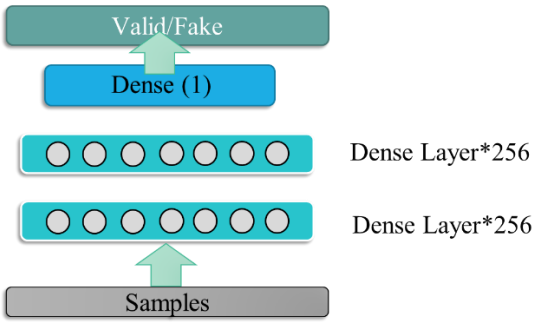


圖 7: DNN-LSGAN 之鑑別器

此外，顧慮到摺積運算比起全連接 DNN 需要更多運算資源，因此也採用以 DNN 結構和 LSGAN 為基礎的 DNN-LSGAN 作為強健性方法之一。為了減輕運算量，我們

捨棄摺積運算，總共使用六層全連接 DNN 網路作為生成對抗網路的生成器(G)，如圖 5。在上述兩個生成對抗網路中，我們分別搭配不同的鑑別器使用，如圖 6 及圖 7 所示。

四、實驗與討論

本研究採用 Aurora-4 作為實驗語料庫，收錄了華爾街日報(Wall Street Journal, WSJ)之朗讀發音，並包含-5dB 至+15dB 的雜訊。簡而言之 Aurora-4 是以華爾街日報為基礎錄音並加上 6 種不同情境下的噪音來組成的，是一個專門設計用來從事語音強健性技術研究的語料庫。其中包含 8KHz 與 16KHz 兩種音頻取樣率並且採用兩種麥克風錄音(Sennheiser, Secondary-Mic)。其訓練資料集可分為無雜訊干擾(Clean-Condition)，和多情境雜訊混合(Multi-Condition)兩種，測試集則包含的 6 種雜訊，分別包含 330 個發音，其種類如下:人聲(Babble)、汽車(Car)、機場(Airport)、火車(Train)、街道(Street)、餐廳(Restaurant)。另一方面，將測試集分為 A、B、C、D 四種子集合，其詳細介紹如表: 1 所示，此外本研究均採用 16Khz 作為取樣率。 語音特徵部分我們採用 13 維 MFCC，聲學模型部分我們採用 5 層 TDNN-F 網路，每一層 650 維，並將瓶頸層設為 128 維，並訓練 8 個 epoch。

表: 1　Aurora4 簡介

取樣率	8kHz/16kHz
語音內容	WSJ 5000 詞
長度	約 15 小時，每一句約 5~12 秒鐘
訓練資料	Clean:7138 個語句 Multi:7137 個語句
測試資料	A 組:330 個無雜訊語句
	B 組:1980 個語句，包含六種環境噪音
	C 組: 受通道效應干擾的 330 個無雜訊的語句
	D 組: 受通道效應干擾的 1980 個包含雜訊的語句

表: 2 實驗結果

Clean Condition Training (WER%)					
	A	B	C	D	AVG
MFCC+TDNN-F	3.61	41.96	33.18	60.01	34.69
WAV+ SEGAN+TDNN-F	4.15	35.20	40.84	55.28	33.86
Modulation Spectrum +LSGAN+TDNN-F	10.29	34.11	23.03	47.48	28.73
Modulation Spectrum +CAGAN+TDNN-F	6.91	30.17	20.08	42.46	**24.95**

本研究比較經典的生成對抗網路方法應用於強健式語音辨識的效果，以及本研究設計的其餘結合生成對抗網路與語音訊號增益法應用於強健式語音辨識的方法。此外，由於 TDNN-F 為新穎的聲學模型，其除了擁有考慮時間資訊的功能之外，也能夠透過因子分解捨去不必要資訊，使整體模型更容易訓練，因此我們以 TDNN-F 做為其餘實驗的 ASR 系統。

從表: 2 中，我們可以看出使用生成對抗網路出直接作用於波形圖上的增益方法，雖然有些微效果，但是對整體語音辨識率的幫助有限。這是由於波形圖中包含太多資訊，並不是所有都有助於我們了解一句話的語意。但是此方法確實可以幫助改善人類聽覺的效果，不過對 ASR 系統幫助有限。所以在 ASR 中，我們如果提取出特徵，再行進一步處理，將會有更好效果。於是本研究之 CAGAN 與 DNN-LSGAN，即採用 MFCC 特徵轉換成調變頻譜特徵，並取出其中之強度頻譜，利用深度學習技術中的映射功能進行訊號增益還原出乾淨語句的特徵，同時也由其他研究得到啟發重新設計 GAN 內部的網路

結構，使其針對我們的任務有更好表現。未來希望能夠再結合傳統強健性技術，使本研究之方法在語音辨識任務下能有更佳效果。

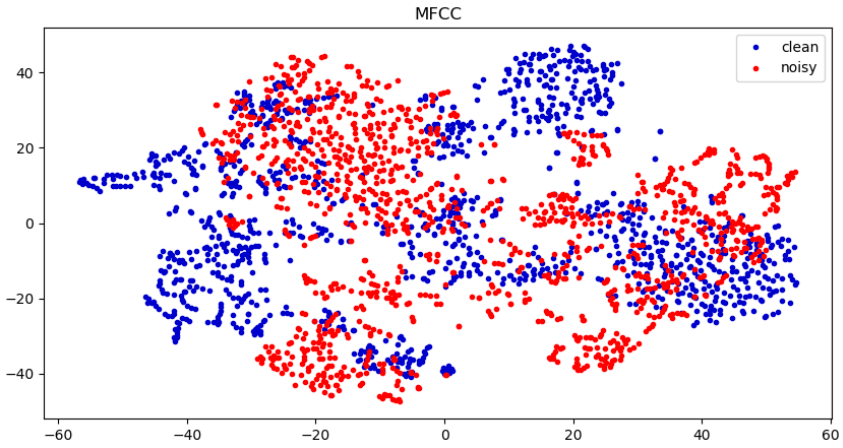


圖 8: MFCC 特徵分布圖

經過 T-SNE 降維後，可將多維度資料投影至某一平面上，方便我們觀察。我們以視覺化方式來討論 MFCC 之特徵分布以及經過強健性技術處理後之差異。乾淨語句與受噪聲干擾語句之 MFCC 分布圖如圖 8 所示，由此可以看出噪聲干擾的確扭曲了語音特徵的分布結構。

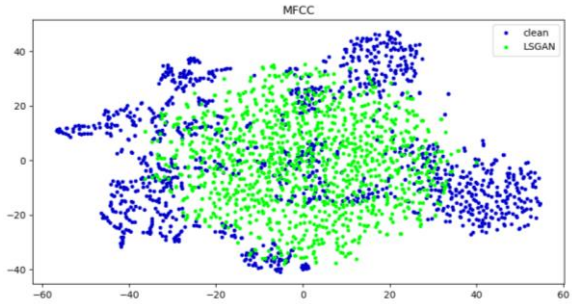


圖 9: DNN-LSGAN 處理後之 MFCC 分布

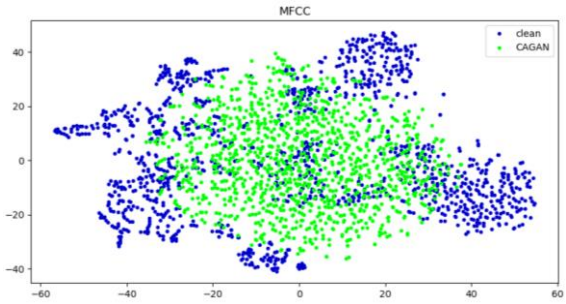


圖 10: CAGAN 處理後之 MFCC 分布

由圖:9 與圖:10 我們可以觀察出一些現象，經過生成對抗網路增益後的語句特徵大幅度的調整了受噪聲干擾的分布，使其與乾淨語句的特徵分布有較多靠近與重疊處。雖

然不尚明顯，但是經過 DNN-LSGAN 與 CAGAN 處理後之特徵分布有較接近乾淨語句分布的趨勢，以上現象也將反映在語音辨識效果上。

五、結論

本研究探討強健式語音辨識的新穎技術，並採用 Aurora-4 語料庫做為實驗基礎，用以比較語音強健技術基於生成對抗網路方法之下對於自動語音辨識的影響。由於調變頻譜可以呈現語音特徵更大尺度變化，所以我們就順著這個脈絡，由調變頻譜特徵著手研究並比較時域特徵與頻率域特徵運用於在語音訊號增益方法上對於提升語音辨識效果的幫助。此外，生成對抗網路的一大特色就是可以自動鑑別生成特徵準確與否，以往多運用在影像處理與電腦視覺領域研究中，用來轉換不同風格圖片，或是用來將含有噪訊之影像映射成特定類型之乾淨影像。本研究由此得到啟發，採用生成對抗網路作為一種語音強健性技術。本研究主要運用生成對抗網路來實現訊號增益方法，從 CAGAN 與 DNN-LSGAN 的實驗中，我們發現在調變頻譜上應用訊號增益方法比起其他媒介更能夠有效提升語音辨識率的效果。與原始 MFCC 比較本研究之方法可分別降低 5.96(WER%)與 9.74(WER%)。

未來希望除了強度頻譜之外，採用相位頻譜也能亦或是結合傳統強健性方法也能成為研究方向之一。此外，由於本研究主要探討以特徵為基礎的強健性方法較少關注以模型為基礎的強健性技術。未來也可採用資料增強法(Data Augmentation)用來增加訓練資料的變異度以及加入更多種模擬雜訊，使用多情境訓練方式來訓練聲學模型，使聲學模型可以學習到更多種情境的資訊，也可以大幅降低訓練與測試的環境不匹配問題，從而大幅提升語音辨識效果。因此，未來希望不只是專注在特徵上，雖然模型方法的缺點在於需要更多計算量，但是隨著硬體運算技術進步，我們可以將精神轉移到資料增強法和模型調適方法上，或許能有更多突破，並且能使研究更具實用價值。

致謝

本論文之研究承蒙行政院科技部研究計畫 (MOST 105-2221-E-003-018-MY3 和 MOST 107-2221-E-003-013-MY2、MOST 108-2221-E-003-005-MY3 和 MOST 108-2634-F-008 -004 -) 之經費支持，謹此致謝。

參考文獻

[1] 汪逸婷, “運用調變頻譜分解技術於強健語音特徵擷取之研究,” *國立臺灣師範大學碩士論文*, 2014.

[2] 朱紋儀, “調變頻譜正規化用於強健式語音辨識之研究,” *國立臺灣師範大學 碩士論文*, 2011.

[3] 張庭豪, “調變頻譜分解之改良於強健性語音辨識,” *國立臺灣師範大學 碩士論文*, 2015.

[4] 顏必成 石敬弘 劉士弘 陳柏琳, “使用字典學習法於強健性語音辨識 The Use of Dictionary Learning Approach for Robustness Speech Recognition,” 於 *ROCLING, ACLCLP*, 2016.

[5] Bi Cheng Yan, Chin Hong Shih, Shih Hung Liu, Berlin Chen, "Exploring Low-Dimensional Structures of Modulation Spectra for Robust Speech Recognition," in *INTERSPEECH*, 2017.

[6] Pierre Baldi, "Autoencoders, Unsupervised Learning, and Deep Architectures," in *JMLR: Workshop and Conference Proceedings*, 2012.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Networks," in *NIPS*, 2014.

[8] Santiago Pascual, Antonio Bonafonte, Joan Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *INTERSPEECH*, 2017.

[9] Santiago Pascual, Antonio Bonafonte, Joan Serrà, Jose A. Gonzalez, "Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks," in *arXiv*, 2018.

[10] Wang, Ke and Zhang, Junbo and Sun, Sining and Wang, Yujun and Xiang, Fei and Xie, Lei, "Investigating Generative Adversarial Networks based Speech Dereverberation for Robust Speech Recognition," in *INTERSPEECH*, 2018.

[11] Chris Donahue, Bo Li, Rohit Prabhavalkar, "Exploring Speech Enhancement With Generative Adversarial Networks," in *ICASSP*, 2018.

[12] P.C.Loizou, Speech Enhancement: theory and Practice, Boca Raton, FL, USA: CRC Press, 2013.

[13] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series, NewYork: WILEY, 1949.

[14] Jahn Heymann, Lukas Drude, Aleksej Chinaev, Reinhold Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *ASRU*, 2015.

[15] Tobias Menne, Ralf Schlüter, Hermann Ney, "Speaker adapted beamforming for multi-channel automatic speech recognition," in *SLT*, 2018.

[16] Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, Shinji Watanabe, "Building state of the art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *INTERSPEECH*, 2018.

[17] Tobias Menne, Ralf Schluter, Hermann Ney, "INVESTIGATION INTO JOINT OPTIMIZATION OF SINGLE CHANNEL SPEECH ENHANCEMENT AND ACOUSTIC MODELING FOR ROBUST ASR," in *ICASSP*, 2019.

[18] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang, Springer Handbook of Speech Processing, Berlin Heidelberg: Springer-Verlag, 2008.

[19] JAE.S. Lim, ALAN.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE,* no. 67,no12, p. 1586–1604.

[20] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters,* vol. 1, no. 21, p. 65–68, 2014.

[21] Kun Han, Yanzhang He, Deblin Bagchi, Eric Fosler-Lussier, DeLiang Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *in Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 226-240

Speech enhancement based on the integration of fully convolutional network, temporal lowpass filtering and spectrogram masking

Kuan-Yi Liu[1], Syu-Siang Wang[2], Yu Tsao[2], Jeih-weih Hung[1]
[1]National Chi Nan University, Taiwan
[2]Academia Sinica, Taiwan
s106323508@mail1.ncnu.edu.tw, sypdbhee@gmail.com, yu.tsao@citi.sinica.edu.tw, jwhung@ncnu.edu.tw

Abstract

In this study, we focus on the issue of noise distortion in speech signals, and develop two novel unsupervised speech enhancement algorithms including temporal lowpass filtering (TLP) and relative-to-maximum masking (RMM). Both of these two algorithms are conducted on the magnitude spectrogram of speech signals. TLP uses a simple moving-average filter to emphasize the low modulation frequencies of speech signals, which are believed to contain richer linguistic information and exhibit higher signal-to-noise ratios (SNR). Comparatively, in RMM we apply a mask that is directly multiplied with the speech spectrogram in a point-wise manner, and the used masking value is directly proportional to the magnitude of each temporal-frequency (T-F) point in the spectrogram. The preliminary experiments conducted on a subset of TIMIT database show that the two novel methods can promote the quality of noise-corrupted speech signals significantly, and both of them can be integrated with a well-known supervised speech enhancement scenario, namely fully convolutional network, to achieve even better perceptual speech quality values.

Keywords: temporal lowpass filtering, relative-to-maximum masking, moving-average filter, speech enhancement

1. Introduction

Nowadays the technologies of communication have been developed quite quickly and they have changed and influenced our life a lot. In particular, speech communication such as speech signal transmission and reception through a wired or a wireless network, has been a widespread use in our daily life [1-2]. Therefore, high speech quality and intelligibility during communication gradually becomes a prerequisite.

However, in the transmission environment of speech signals, there exist lots of distortions, such as additive noise, channel mismatch and reverberation, which inevitably decrease the speech signal quality/intelligibility seriously. To overcome these distortions in speech communication, a lot of researchers in recent decades have been devoted to developing speech enhancement (SE) techniques. These SE algorithms can be classified based on whether a learning/training process is involved. For example, if the noise statistics in spectral-subtractive

SE algorithms [3-6] and the basis of the clean signal subspace in subspace SE algorithms [7-9] are learned via a training set with explicit labels, then the corresponding SE methods are supervised. Comparatively, in the unsupervised methods, such as spectral subtraction (SS) [10], Wiener filtering [11], short-time spectral amplitude (STSA) estimation [12] and short-time log-spectral amplitude estimation (logSTSA) [13], does not employ prior information about speech and/or noise.

In this study, we develop two novel learning-free SE methods. One is called temporal low-pass filtering (TLF) and the other is relative-to-maximum masking (RMM). Briefly speaking, TLF borrows the idea of Mod-WD [14], a learning-free SE method, while it can be implemented significantly more simply than Mod-WD, and the mask used in RMM is totally data-driven, viz. it is determined by the signal being processed and has nothing to do with a training set. We then examine the SE capability of the presented novel methods, and see if they are additive to an advanced SE framework based on a deep learning-based fully convolutional network (FCN) [15] to provide even better speech quality for noise-distorted speech signals.

The remainder of this paper is organized as follows: Section 2 presents the details of two novel SE methods, TLF and RMM. The experimental setup is given in Section 3, and Section 4 exhibits the experimental results together with their discussions. Finally, a concluding remark is provided in Section 5.

2. The presented novel SE methods

In this section, we present two novel speech enhancement methods, which are named temporal lowpass filtering (TLF) and relative-to-maximum masking (RMM), respectively. Both of these two methods modify the input utterances in the spectro-temporal (spectrogramic) domain, and they do not require a learning (training) procedure.

2.1 Temporal lowpass filtering

It has been revealed that the important information helpful for human intelligibility and automatic recognition is mainly dwelled in the relatively low-varying components of a speech temporal stream [16-18]. Thus some well-known speech enhancement and noise-robust feature extraction algorithms are developed via emphasizing/diminishing the low/high modulation frequency components of frame-wise speech feature time series. The ModWD algorithm discussed in the previous section follows this trend and factorizes the spectrogram of a noisy signal and then decrease the resulting detail (high half modulation-frequency) part.

Experimental results have revealed that ModWD can moderately improve the speech quality and it can be also well additive to some well-known SE method.

Partially inspired by the aforementioned concept, in this study we present using a simple moving-average filter to process the time series of the spectrogram of noise-corrupted utterances. The presented scheme is analogous to ModWD in emphasizing the low-varying component of the acoustic spectra along the temporal axis.

The block diagram of TLF is shown in Figure 3.1, which consists of the following three steps:

Step 1: Create the spectrogram $\{X[m,k], 0 \leq m \leq M-1, 0 \leq k \leq K-1\}$ for a given time-domain signal $x[n]$, where m and k are respectively the indices of frame and acoustic frequency, and M and K are the total numbers of frames and acoustic frequency points, respectively.

Step 2: Pass the magnitude spectral sequence $\{|X[m,k]|, 0 \leq m \leq M-1\}$ for each acoustic frequency (with index k) through a length-L moving-average filter. The resulting new magnitude sequence is:

$$\left|\hat{X}[m,k]\right| = \frac{1}{L}\sum_{\ell=0}^{L-1}|X[m-\ell,k]|, \qquad (3.1),$$

where $\left|\hat{X}[m,k]\right|$ is the updated magnitude spectral sequence.

Step 3: Construct the new time-domain signal $\hat{x}[n]$ by applying the inverse STFT to the updated spectrogram, which consists of the new magnitude spectrogram $\left\{\left|\hat{X}[m,k]\right|\right\}$ and the original phase spectrogram $\{\angle X[m,k]\}$.

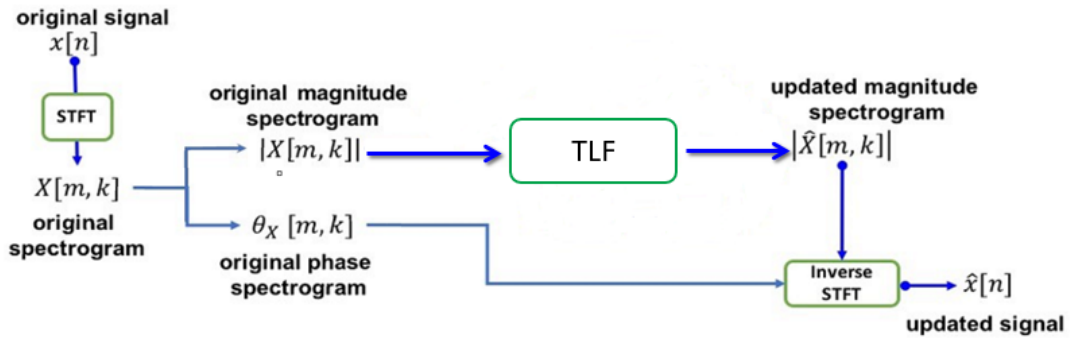


Figure 3.1: The block diagram of TLF.

Since this new method applies a simple lowpass filter (i.e., the moving-average filter) along

the temporal domain of the spectrogram, it is named as temporal lowpass filtering with a short-hand notation "TLF". Some major underlying characteristics of this new method TLF are stated as follows:

1. The used moving-average filter is to emphasize the relatively low modulation frequency portion of the acoustic (magnitude) spectral time series, which is believed to contain helpful linguistic information and be more energy concentrated with a higher signal-to-noise ratio (SNR) compared to the high modulation portion.
2. The greater the length of the employed moving-average filter, the smoother the resulting magnitude spectral curve. However, the filter length needs to be carefully determined in order to diminish the possibly harmful high-frequency part while avoid over-smoothness that ruins the low-varying part.
3. In comparison with the method ModWD that uses DWT and inverse DWT (which consists of at least four filtering processes together with down-sampling and up-sampling), this new approach just applies a filter and is thus simpler in implementation.

2.2 Relative-to-maximum masking

The speech enhancement methods based on time-frequency (T-F) masking have received much attention in the recent decade partially due to its simplicity in computation as well as high capability in segregation speech signals from noise. Among these mask-wise SE methods, a general ideal binary mask (IBM) [19,20] method uses a zero-one masking matrix performing on the spectrogram such that the instantaneous T-F unit for the spectrogram is kept unchanged if it is greater than a threshold that depends on a local SNR criterion (LC), and is set to zero otherwise. By contrast, the method of ideal ratio mask (IRM) [21] applies a soft mask for each instantaneous T-F unit, with the RMM value within the range of zero and one which somewhat reflects the probability of the T-F unit to be speech-wise.

In both methods of IBM and IRM, a key procedure is to estimate the instantaneous signal-to-noise ratio (SNR) of the processed signal. Furthermore, in the recent studies a deep neural work (DNN)-based scenario is used to learn the mask coefficients in IBM and IRM, which inevitably requires a training data set, which contains a great number of noisy-clean signal pairs.

Partially motivated by the ideas of IBM and IRM, in this study we propose a novel RMM scheme which aims to enhance the spectrogram of noise-corrupted utterances. This novel RMM scheme requires no SNR estimation, nor a training stage. The used mask coefficients are totally determined by the utterance being processed. The block diagram of RMM is shown in Figure 3.2, which consists of the following three steps:

Step 1: Create the spectrogram $\{X[m,k], 0 \le m \le M-1, 0 \le k \le K-1\}$ for a given time-domain signal $x[n]$, where m and k are respectively the indices of frame and acoustic

frequency, and M and K are the total numbers of frames and acoustic frequency points, respectively.

Step 2: Compute the RMM coefficients by

$$S_{mk} = \frac{|X[m,k]|}{\max_{m,k}\{|X[m,k]|\}}, \quad 0 \le m \le M-1, 0 \le k \le K-1, \tag{3.2}$$

where S_{mk} is the mask value that will apply to the magnitude spectrogram $|X[m,k]|$ at the m^{th} time frame and k^{th} acoustic frequency bins. From Eq. (3.2), we see that the mask value is simply the ratio of the instantaneous T-F magnitude to the maximum T-F magnitude over the whole spectrogram of the utterance. Thereafter, the new magnitude spectrogram is determined by

$$\left|\hat{X}[m,k]\right| = S_{mk}|X[m,k]|, \qquad 0 \le m \le M-1, 0 \le k \le K-1. \tag{3.3},$$

Step 3: Construct the new time-domain signal $\hat{x}[n]$ by applying the inverse STFT to the updated spectrogram, which consists of the new magnitude spectrogram $\{|\hat{X}[m,k]|\}$ and the original phase spectrogram $\{\angle X[m,k]\}$.

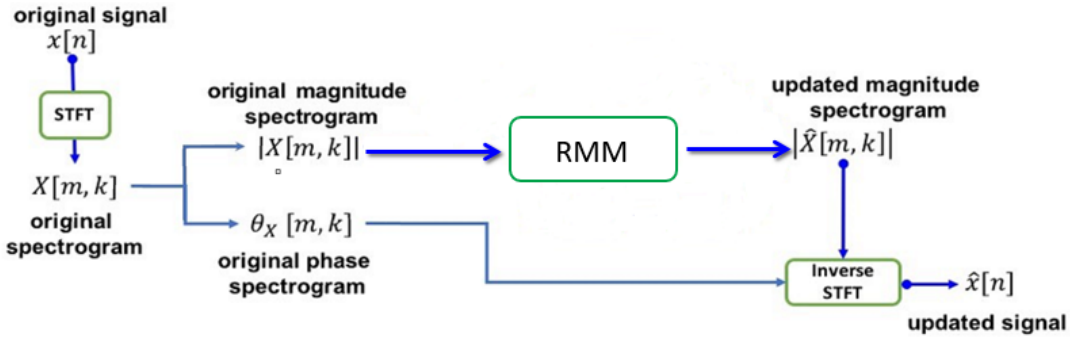


Figure 3.2: The block diagram of RMM.

The new RMM algorithm stated above is termed relative-to-maximum masking, abbreviated by RMM, since the RMM value for each T-F entity is determined by Eq. (3.2). The origin of RMM is a quite simple and naïve idea: a larger magnitude entity in the spectrogram often comes with a high signal-to-noise (SNR) ratio, and it deserves a higher confidence score which is reflected by a larger RMM value. Compared with the other two types of RMM methods, IBM and IRM, RMM does not require noise estimation, nor a supervised/unsupervised learning stage to determine the applied mask. The used mask in RMM is completely data-driven, viz. it totally depends on the test utterance being processed at the present.

3. Experimental setup

As for our evaluation experiments, we use a subset of the TIMIT database [22] to prepare the test set for all SE methods and the training set for the method FCN. TIMIT contains American English utterances produced by 630 speakers. From TIMIT we select 700 utterances pronounced by male speakers and recorded at a sampling rate of 16 Hz. Among these selected 700 utterances, 600 utterances are used to be the training set, while the remaining 100 utterances serve as the test set.
Next, all of the utterances in the training and test sets are manually corrupted by additive noise at various signal-to-noise ratios (SNRs). The numbers of noise types that corrupt the training set and test set are 5 and 3, respectively.

Four speech enhancement (SE) methods, including fully convolutional network (FCN), modulation-domain wavelet denoising (ModWD), temporal lowpass filtering (TLP) and relative-to-maximum masking (RMM), will be evaluated here. As for these four SE methods, FCN is the only one supervised learning method, which requires the training set to learn the associated network parameters. Some important setup factors about the used FCN here are as follows:

1. The FCN model consists of eight convolutional layers with padding, each layer containing filters each with a filter size of 11.
2. The activation function of for each layer output is parametric rectified linear units (PReLUs).
3. The FCN model parameters are trained using Adam optimization algorithm, in which batch normalization is applied so as to minimize the mean square error between the output of the final layer and desired clean time-domain utterance.

Regarding the other three SE methods, ModWD, TLF and RMM, which mainly process the magnitude spectrogram of each test utterance, the general arrangements are listed below:

1. Each test utterance is split into overlapped frames. The frame duration and frame shift are set to be 64 ms and 10 ms, respectively, and thus the frame rate is 100 Hz, which covers the modulation frequency range [0, 50 Hz] for the analyzed speech feature streams.
2. A Hamming window is then applied to each frame signal.
3. The size of the discrete Fourier transform applied to each frame signal is 512, and thus the first 257 frequency bins of the resulting spectrum are used.
4. The biorthogonal 3.7 wavelet basis is used for the DWT and inverse DWT of ModWD.

5. The length of the moving-average filter in TLF is set to 2, with the purpose to cover the modulation frequencies 0-25 Hz approximately, which is highly correlated with linguistic information.
6. Unless otherwise specified, in the RMM method the mask derived with the original or the enhanced test utterance is always applied to the spectrogram of the original (unprocessed) version of the test utterance.

Finally, to evaluate the denoising capability of the aforementioned four SE methods, we employ the well-known objective measure metric, perceptual estimation of speech quality (PESQ) [23], which ranks the level of enhancement for the processed utterances relative to the original noise-free ones. PESQ indicates the quality difference between the enhanced and clean speech signals, and it ranges from -0.5 to 4.5. A higher PESQ score implies that the tested utterance is closer to its clean counterpart.

4. Experimental results and discussions

4.1 Each single SE method

At the outset, we would like to investigate the SE behavior for any individual of the SE methods, which include fully convolutional network (FCN), temporal lowpass filtering (TLP), modulation-domain wavelet denoising (Mod-WD) and relative-to-maximum masking (RMM). Tables 4.1 list the PESQ scores obtained from the baseline and these SE methods with averaging three noisy cases "Engine", "White" and "Crowd". From this tables, we have the following observations:

1. The PESQ score degrades as the signal-to-noise ratio (SNR) of the environment becomes worse, and thus it is believed to be a good metric to reflect the quality of speech utterances.
2. About the cases of the SNR greater than -6 dB for the three noise types, FCN performs the best, closely followed by RMM, and then TLF and ModWD. Notably, as for the two low-pass filtering methods, TLF behaves moderately better than ModWD while it can be implemented in a simpler manner. As mentioned earlier, TLF just uses a moving-average filter, while ModWD requires a DWT-IDWT procedure, which involves both filtering, down-sampling and up-sampling.

To briefly conclude, FCN gives better PESQ scores than the other three methods at moderate noise levels, RMM works quite well for almost all SNR cases, while ModWD and TLF give rise to relatively slight improvement. Since these SE methods are developed along different

directions, it is natural to assume that the combination of two or three of them might cause further improvement relative to each component method. This part will be examined in the subsequent two sub-sections.

Finally, we evaluate different SE methods in the domain of magnitude spectrogram. A speech signal corrupted by engine noise at 0 dB SNR is individually processed by any of these SE methods, and the corresponding magnitude spectrograms are shown in Figure 4.1. From this figure, it is clear that FCN brings an optimal denoising performance compared with the other methods. The novel presented RMM behaves also quite well, but it seems to over-eliminate the portion of high acoustic frequencies partially because these frequencies possess significantly low energy and cause low masking values.

Table 4.1: The PESQ scores obtained from the baseline, and any of four SE methods as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	ModWD	TLF	RMM
-15dB	1.017	0.989*	1.035	1.035	**1.122**
-12dB	1.078	1.076*	1.087	1.087	**1.228**
-6dB	1.283	1.504	1.302	1.310	**1.600**
0dB	1.592	**2.024**	1.611	1.623	1.976
6dB	1.973	**2.494**	1.992	2.005	2.388
12dB	2.391	**2.855**	2.407	2.423	2.737
18dB	2.811	**3.145**	2.819	2.838	2.943

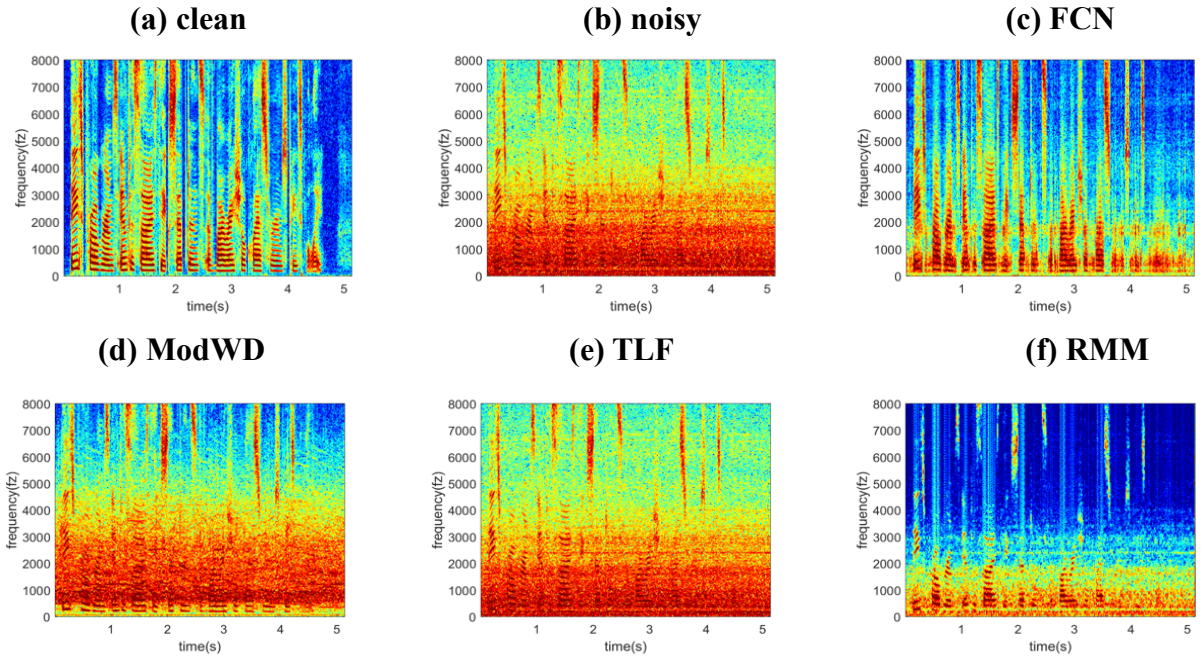


Figure 4.1: The magnitude spectrograms of (a) a clean-noise free signal $\boldsymbol{x}$ (b) the noisy counterpart, $\tilde{\boldsymbol{x}}$, of $\boldsymbol{x}$, which contains 0-dB engine noise, (c) the FCN-enhanced version of $\tilde{\boldsymbol{x}}$, (d) the ModWD-enhanced version of $\tilde{\boldsymbol{x}}$, (e) the TLF-enhanced version of $\tilde{\boldsymbol{x}}$, (f) the RMM-enhanced version of $\tilde{\boldsymbol{x}}$

4.2 The pairing of two SE methods

As mentioned before, almost any of the used four SE methods (FCN, ModWD, TLF and RMM) can enhance the distorted utterances, while the SNR cases for each method to work best are different. In addition, these SE methods might process different parts of the distorted utterances with different goals in speech enhancement. For example, FCN directly minimizes the discrepancy of the noisy speech and its clean counterpart in the training set, ModWD and TLF alleviates the high modulation frequency portions in noisy speech, and TLF emphasizes the high-energy temporal-spectral bins. With this in mind, we would like to investigate whether the cascade of two or three of these SE methods can behave better than each constituent method.

First of all, the cascade of FCN and either of ModWD and TLF is evaluated, in which the test utterances at the three noise conditions are first processed by FCN, and the resulting spectrogram is lowpass filtered by ModWD or TLF. The corresponding PESQ scores are listed in Tables 4.2 and 4.3. From these two tables, we find that both combinations, "FCN plus ModWD" and "FCN plus TLF", give rise to even better results than each component method at almost all SNR cases (except 18 dB for FCN plus ModWD). The amount of PESQ improvement is more significant at lower SNRs. In addition, "FCN plus TLF" outperforms "FCN plus ModWD", which further reveals the advantage of TLF over ModWD, since TLF behaves better with a lower computational cost.

Table 4.2: The PESQ scores obtained from the baseline, FCN, ModWD and the pairing of FCN and ModWD as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	ModWD	FCN+ModWD
-15dB	1.017	0.989*	1.035	**1.126**
-12dB	1.078	1.076*	1.087	**1.216**
-6dB	1.283	1.504	1.302	**1.629**
0dB	1.592	2.024	1.611	**2.116**
6dB	1.973	2.494	1.992	**2.539**
12dB	2.391	2.855	2.407	**2.856**
18dB	2.811	3.145	2.819	3.112

Table 4.3: The PESQ scores obtained from the baseline, FCN, TLF and the pairing of FCN and TLF as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	TLF	FCN+TLF
-15dB	1.017	0.989*	1.035	**1.159**
-12dB	1.078	1.076*	1.087	**1.263**
-6dB	1.283	1.504	1.310	**1.672**
0dB	1.592	2.024	1.623	**2.144**
6dB	1.973	2.494	2.005	**2.559**
12dB	2.391	2.855	2.423	**2.888**
18dB	2.811	3.145	2.838	**3.149**

Next, we examine the integration of RMM with the other three methods. The mask used in RMM is created by any of FCN-, ModWD- and TLF-preprocessed test utterances, which is then applied to the "original" (unprocessed) test utterance counterpart. The respective PESQ results are shown in Tables 4.4, 4.5 and 4.6. From the three tables and figure we have several findings listed below:

1. As for the method "FCN plus RMM", it performs better than FCN and RMM at low SNRs, -15 dB, -6 dB and 0 dB, while it gets worse with the increase of the SNR. One possible reason for the performance degradation at higher SNRs is the phase mismatch in the complex-valued spectrograms of the original and FCN-processed utterances. As we know, FCN updates a test utterance in the time domain, and thus changes both the magnitude and phase parts of the respective spectrogram. However, only the FCN-processed magnitude part is used to create the mask in RMM, which is applied to the original magnitude part. Accordingly, the FCN-processed phase part is discarded in the whole process.

2. Regarding the two combinative methods "ModWD plus RMM" and "TLF plus RMM", the associated PESQ scores are always much higher than the single ModWD and TLF, indicating that for the original noisy spectrogram, the masking operation (with spectrogram masks created by ModWD- and TLF-processed signals) are more effective than the operations of ModWD and TLF. In addition, "ModWD plus RMM" and "TLF plus RMM" outperforms RMM for the SNRs less than 18 dB. At a high SNR as 18 dB, the Mod-WD/TLF-wise masks might over-smooth the spectrogram, and thus are less helpful than the mask created by the nearly clean signal.

Table 4.4: The PESQ scores obtained from the baseline, FCN, RMM and the pairing of FCN and RMM as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	RMM	FCN+RMM
-15dB	1.017	0.989*	1.122	**1.039**
-12dB	1.078	1.076*	**1.228**	1.190
-6dB	1.283	1.504	1.600	**1.718**
0dB	1.592	2.024	1.976	**2.118**
6dB	1.973	**2.494**	2.388	2.390
12dB	2.391	**2.855**	2.737	2.566
18dB	2.811	**3.145**	2.943	2.719*

Table 4.5: The PESQ scores obtained from the baseline, ModWD, RMM and the pairing of ModWD and RMM as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	ModWD	RMM	ModWD+RMM
-15dB	1.017	1.035	1.122	**1.140**
-12dB	1.078	1.087	1.228	**1.252**
-6dB	1.283	1.302	1.600	**1.615**
0dB	1.592	1.611	1.976	**1.998**
6dB	1.973	1.992	2.388	**2.403**
12dB	2.391	2.407	2.737	**2.741**
18dB	2.811	2.819	**2.943**	2.926

Table 4.6: The PESQ scores obtained from the baseline, TLF, RMM and the pairing of TLF and RMM as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	TLF	RMM	TLF+RMM
-15dB	1.017	1.035	1.122	**1.132**
-12dB	1.078	1.087	1.228	**1.257**
-6dB	1.283	1.302	1.600	**1.628**
0dB	1.592	1.611	1.976	**2.012**
6dB	1.973	1.992	2.388	**2.416**
12dB	2.391	2.407	2.737	**2.755**
18dB	2.811	2.819	**2.943**	2.939

Finally, the results for all combinative methods are summarized in Table 4.7 and Figure 4.2. From these results, we find that the method "FCN plus TLF" behaves the best, except for the case of -6 dB-SNR, showing that a simple lowpass filtering is quite additive to FCN to alleviate the noise effect. Comparatively, the two well-behaved methods, FCN and RMM, do not necessarily exhibit the most complementary effect.

Table 4.7: The PESQ scores obtained from several combinative methods as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	FCN+ModWD	FCN+TLF	FCN+RMM	ModWD+RMM	TLF+RMM
-15dB	1.126	**1.159**	1.039	1.140	1.132
-12dB	1.216	**1.263**	1.190	1.252	1.257
-6dB	1.629	1.672	**1.718**	1.615	1.628
0dB	2.116	**2.144**	2.118	1.998	2.012
6dB	2.539	**2.559**	2.390	2.403	2.416
12dB	2.856	**2.888**	2.566	2.741	2.755
18dB	3.112	**3.149**	2.719	2.926	2.939

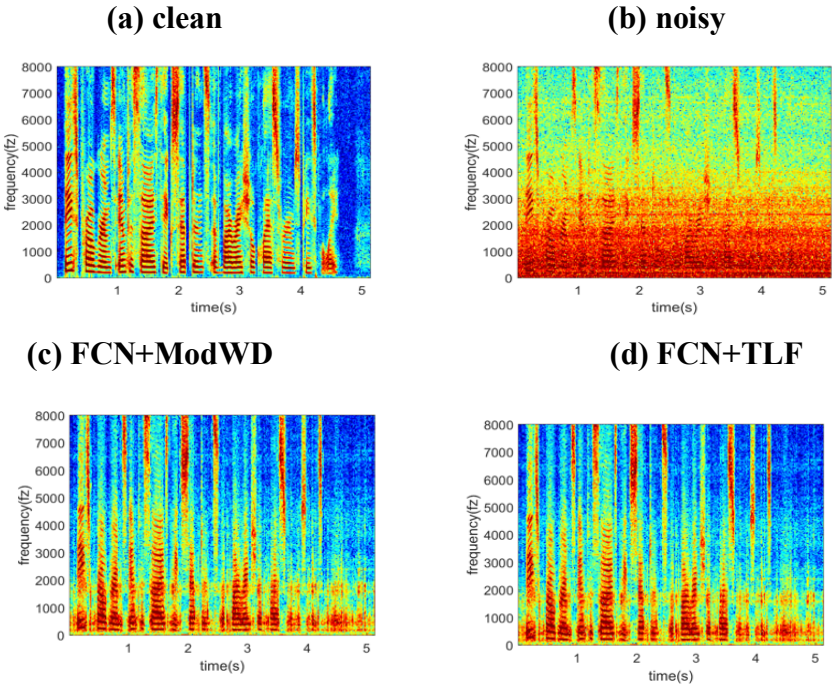


Figure 4.2: The magnitude spectrograms of (a) a clean-noise free signal $\boldsymbol{x}$ (b) the noisy counterpart, $\tilde{\boldsymbol{x}}$, of $\boldsymbol{x}$, which contains 0-dB engine noise, (c) the FCN-plus-ModWD-enhanced version of $\tilde{\boldsymbol{x}}$, (d) the FCN-plus-TLF-enhanced version of $\tilde{\boldsymbol{x}}$

Figure 4.2 shows the magnitude spectrograms for the clean and noisy signals, plus the signals

processed by two combinative methods, "FCN+ModWD" and "FCN+TLF", which have been revealed to promote the PESQ scores apparently. From this figure, we reconfirm that these two combinative methods can reduce the noise effect a lot in the distorted signal and thus bring the recovery of the embedded clean noise-free part.

4.3 The integration of three SE methods

Following the trend in the previous two sub-sections, here we would like to investigate what happens if we use the concatenation of three SE methods to process the test utterances. For simplicity, we use two forms of concatenation: one is FCN followed by ModWD and RMM in turn, denoted by "FCN+ModWD+RMM", and the other is FCN followed by TLF and RMM successively, denoted by "FCN+TLF+RMM". Therefore, these two forms differ in the used lowpass processing method at the median stage. The corresponding PESQ scores are listed in Tables 4.11and 4.12. For the ease of comparison, the results of FCN, FCN plus ModWD/TLF, and ModWD/TLF plus RMM are also listed in these tables. According to the results, we have two findings:

1. The two forms of three-method concatenation outperform the single FCN and the other two-method concatenations at the SNRs of -12 dB, -6 dB and 0 dB. When the SNR becomes higher, adding RMM at the final stage fails to increase the PESQ scores, which is probably due to an effect of over-adjustment.
2. When used in the intermediate or final stage, TLF always behaves superior to ModWD. This again confirms the advantage of TLF over ModWD.

Table 4.11: The PESQ scores obtained from FCN, FCN plus ModWD, ModWD plus RMM, and FCN plus ModWD and RMM, as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	FCN	FCN+ModWD	ModWD+RMM	**FCN+ModWD+RMM**
-15dB	0.989	1.126	**1.140**	1.103
-12dB	1.076	1.216	1.252	**1.262**
-6dB	1.504	1.629	1.615	**1.780**
0dB	2.024	2.116	1.998	**2.157**
6dB	2.494	**2.539**	2.403	2.410
12dB	2.855	**2.856**	2.741	2.567
18dB	**3.145**	3.112	2.926	2.707

Table 4.12: The PESQ scores obtained from FCN, FCN plus TLF, TLF plus RMM, and FCN plus TLF and RMM, as for the utterances in the three noise environments **"Engine",**

"White" and "Crowd".

SNR	FCN	FCN+TLF	TLF+RMM	**FCN+TLF+RMM**
-15dB	0.989	**1.159**	1.140	1.141
-12dB	1.076	1.263	1.252	**1.285**
-6dB	1.504	1.672	1.615	**1.797**
0dB	2.024	2.144	1.998	**2.165**
6dB	2.494	**2.559**	2.403	2.411
12dB	2.855	**2.888**	2.741	2.572
18dB	3.145	**3.149**	2.926	2.714

5 Conclusion

To our knowledge, a fully convolutional network (FCN) applied in an SE framework outperforms conventional neural networks like densely connected network and convolutional neural network (convnet) in promoting the quality of distorted speech signals. Compared with an FCN-based SE framework, the two novel learning-free SE algorithms, temporal lowpass filtering (TLF) and relative-to-maximum masking (RMM) presented in this paper are shown to provide even better denoising performance at some particular signal-to-noise ratio (SNR) cases, despite their simplicity in implementation and their irrelevance with pre-training. Furthermore, our experimental results show that TLF is quite complementary to FCN since the paring of FCN and TLF behaves significantly better than FCN alone. We also show that the two novel methods, TLF and RMM, are quite additive to each other.

In the future avenue, we plan to evaluate FCN, TLF and RMM and the respective possible integrations on the other speech databases, which are recorded in environments that contain various distortions such as additive noise, channel mismatch, and reverberation. In addition, we would like to investigate the theoretical reason why RMM can bring about significant speech quality improvement, and further enhance it by tuning the used mask with a learning scenario.

References

[1] D. O' Shaughnessy, "Speech communications: human and machine," *2nd ed., Hyderabad, India: University Press (India) Pvt. Ltd.*, 2007.

[2] Y. Ephraim, H. L. Ari and W. Roberts, "A brief survey of speech enhancement," *Electrical Engineering Handbook, 3rd ed. Boca Raton, FL: CRC*, 2006.

[3] P. C. Loizou, "Speech enhancement: theory and practice," T*aylor and F. Group, Eds. Boca Raton, FL, USA: CRC Press*, 2013.

[4] R. Martin, "Spectral subtraction based on minimum statistics," *In Proc. European Conference Signal Processing, pp. 1182–1185*, 1994.

[5] P. Krishnamurthy, and S. R. M. Prasanna, "Modified spectral subtraction method for enhancement of noisy speech," *in Proc. International Conference Signal, Image Processing, pp.146-150*, Dec. 2005.
[6] S. Ogata and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal, " *in Proc. International Conference on Electrical and Electronic Technology, vol. 1, pp. 242-245*, 2001.
[7] A. H. Abolhassani, S.-A. Selouani and D. O'Shaughnessy, "Speech enhancement using PCA and variance of the reconstruction error in distributed speech recognition," *in Proc.IEEE ASRU'07,* 2007
[8] B. Nazari, M. Sarkrni and P. Karimi, "A method for noise reduction in speech signal based on singular value decomposition and genetic algorithm," *In Proc. of IEEE EUROCON'09*, 2009
[9] S. J. Rennie, J. R. Hershey and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," *In Proc. of ICASSP'08*, 2008.
[10] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *in Proc. ICASSP*, 2002.
[11] P. Scalart, J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *in Proc. ICASSP*, 1996.
[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoustics, Speech and Signal Process., 1984.
[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. Acoustics., Speech and Signal Process.*, 1985
[14] S.-k. Lee, S.-S. Wang, Y. Tsao and J.-w. Hung, "Speech enhancement based on reducing the detail portion of speech spectrograms in modulation domain via discrete wavelet transform," in ISCSLP, 2018
[15] S. -W. Fu, Y. Tsao, X. Lu and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," *in Proc. APSIPA-ASC*, 2017
[16] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *in Proc. Eurospeech*, 1997.
[17] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech, and Language Processing*, 2006.
[18] X. Xiao, E. S. Chng and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, 2008
[19] G. Kim and P. Loizou, "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints," *IEEE Signal Processing Letters, vol. 17 no. 12 pp. 1010-1013*, 2010.
[20] R. Koning, N. Madhu and J. Wouters, "Ideal time frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," *IEEE Trans. Biomed. Eng. vol. 62 no. 1 pp. 331-341*, 2015.
[21] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *in Proc. ICASSP*, 2013
[22] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database" *NIST Tech Report*, 1988
[23] A. Rix, J. Beerends, M. Hollier and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *Int. Telecommun. Union, T Recommendation, Art. no. 862,* 2001

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 241-245

MONPA: 中文命名實體及斷詞與詞性同步標註系統

葉文照　Wen-Chao Yeh
臺北醫學大學大數據科技及管理研究所
Graduate Institute of Data Science
Taipei Medical University
m946107004@tmu.edu.tw

謝育倫　Yu-Lun Hsieh
中央研究院及國立政治大學
SNHCC, TIGP, Academia Sinica & National Cheng Chi University
morphe@iis.sinica.edu.tw

張詠淳　Yung-Chun Chang
臺北醫學大學大數據科技及管理研究所
Graduate Institute of Data Science
Taipei Medical University
changyc@tmu.edu.tw

許聞廉　Wen-Lian Hsu
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica
hsu@iis.sinica.edu.tw

摘要

有鑑於現今國內外研究繁體中文自然語言處理缺乏合適的斷詞、詞性標註及命名實體辨識的工具，本研究基於 BERT 模型，搭配 CRF 提出以多目標命名實體辨識與詞性標註(Multi-Objective NER POS Annotator, MONPA)系統，並以供學術使用授權條款 CC BY-NC-SA 4.0 License 進行相關安裝套件釋出作業。透過 MONPA 的釋出，嘉惠我國相關學術研究，俾能加快繁體中文自然語言處理之進展。

Abstract

In view of the lack of suitable word segmentation, part-of-speech tagging and named entity recognition tools in the traditional Chinese natural language processing. This study is based on the BERT model with CRF to propose a multi-objective named-entity and part-of-speech annotator, which called MONPA. Our work not only propose a method but also release the relevant python package with the CC BY-NC-SA 4.0 License. We firmly believe that this research project can bridge the technical gap between academia and business applications with our innovation, and enable efficient development of traditional Chinese NLP by all entities in order to enhance our level of competitiveness in the world.

關鍵詞：中文斷詞, 詞性標註, 命名實體辨識, BERT

Keywords: Chinese Word Segmentation, POS tagging, Name Entity Recognition, BERT

一、緒論

綜觀目前繁體中文的斷詞工具主要仰賴 Jieba[1]套件，然而 Jieba 是基於簡體中文語料透過 HMM [1]模型所訓練出來的成果，因此對繁體中文的支援效果不佳，且系統多年未更新。種種的限制讓國內學界或是產業界想要進行繁體中文自然語言處理之研究困難重重。此外，命名實體辨識(named entity recognition)有助於瞭解句子結構進而提升理解能力，但在目前處理繁體中文時尚無可用的工具。繁體中文自然語言處理的基礎設施於此種種的限制之下，勢必使得臺灣的研發能力在這波 AI 浪潮中受阻。有鑑於此，本研究以深度學習方法研發一種能同時完成「命名實體辨識」、「繁體中文斷詞」以及「詞性標註」之系統，並將其完全開源釋出，讓所有想要處理繁體中文的產學界使用者共享此研究成果。

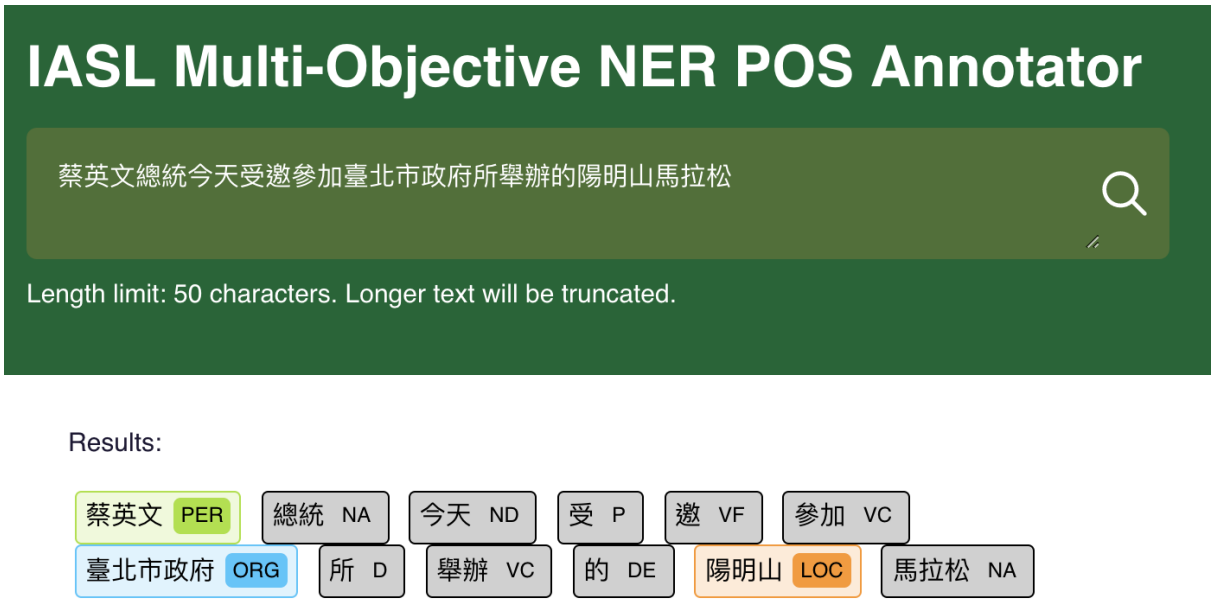
本研究所提出的多目標命名實體辨識與詞性標註(Multi-Objective NER POS Annotator, MONPA)系統，是基於 BERT [2]（應用雙向 Transformer）模型來取得更強健的詞向量（word embeddings）並配合 CRF 同時進行斷詞、詞性標註、及 NER 等多個目標。BERT 模型為現今頂尖的詞向量獲取方法之一，其利用自注意力（self-attention）機制及預訓練（pre-training）等技術以提取更能充分代表整個語句訊息的向量。本研究以授權條款 CC BY-NC-SA 4.0 License 進行相關套件釋出作業。為了使用的便利性，我們

[1]https://github.com/fxsjy/jieba

也將把 MONPA 組成套件發佈到 PyPI，讓使用者能夠透過 pip install 指令安裝。使用者可以透過 Github 獲得 MONPA 相關資訊，進而完成安裝，允許獲得 MONPA 的人依照同一授權條款的情形下再散布。透過 MONPA 的釋出，嘉惠我國相關研究與產業，俾能加快繁體中文自然語言處理之進展。

二、使用 MONPA

MONPA [3]是一個提供繁體中文分詞、詞性標註以及命名實體辨識的多任務模型，初期只有使用原始模型（v0.1）的網站版本[2]（如圖一）。透過本研究的釋出，MONPA 已經包裝成可以 pip install 的 python 套件包，在本次釋出中，我們也透過 BERT 改善 MONPA 的效能(v0.2)，並且發佈在 Github[3]與 PyPI[4]上。使用者能夠在不同的作業平台上透過 pip install 指令完成安裝程序，此外，本研究為了因應 pip 對套件檔案大小的限制，所以在首次引入套件時才會啟動下載最新的 model 檔。



圖一、MONPA v0.1 網頁版示範圖

本研究的釋出包含了三個功能：

- ***斷詞(cut function)***：若只需要中文分詞結果，請使用 cut 功能，回傳值是 list 格式。

[2]http://monpa.iis.sinica.edu.tw:9000/chunk

[3]https://github.com/monpa-team/monpa

[4]https://pypi.org/project/monpa/

程式及輸出如下：

```
1. monpa.cut("蔡英文總統今天受邀參加台北市政府所舉辦的陽明山馬拉松比賽。")
2. ['蔡英文', '總統', '今天', '受', '邀', '參加', '台北市政府', '所', '舉辦', '的', '陽明山', '馬拉松', '比賽', '。']
```

- ***詞性標註(pseg function)***：若需要中文分詞及該詞的 POS 標註，請使用 pseg 功能，回傳值是 list of list 格式，程式及輸出如下：

```
1. monpa.pseg("蔡英文總統今天受邀參加台北市政府所舉辦的陽明山馬拉松比賽。")
2. [['蔡英文', 'PER'], ['總統', 'Na'], ['今天', 'Nd'], ['受', 'P'], ['邀', 'VF'], ['參加', 'VC'], ['台北市政府', 'ORG'], ['所', 'D'], ['舉辦', 'VC'], ['的', 'DE'], ['陽明山', 'LOC'], ['馬拉松', 'Na'], ['比賽', 'Na'], ['。', 'PERIODCATEGORY']]
```

- ***加詞(load_userdict function)***：在 MONPA 元件中，我們提供使用者自訂詞彙的功能，透過 load_userdict function 可以將使用者詞彙檔匯入，請依『詞語 詞頻 詞性』順序製作自訂詞典文字檔。

```
1. 受邀 100 V
```

當要使用自訂詞時，請於執行分詞前先 load_userdict，將自訂詞典載入到 monpa 模組。使用 pseg function 測試，可發現回傳值已依自訂詞典分詞，譬如『受邀』為一個詞而非先前的兩字分列輸出。

```
1. monpa.load_userdict("./userdict.txt")
2. monpa.pseg("蔡英文總統今天受邀參加台北市政府所舉辦的陽明山馬拉松比賽。")
3. [['蔡英文', 'PER'], ['總統', 'Na'], ['今天', 'Nd'], ['受邀', 'V'], ['參加', 'VC'], ['台北市政府', 'ORG'], ['所', 'D'], ['舉辦', 'VC'], ['的', 'DE'], ['陽明山', 'LOC'], ['馬拉松', 'Na'], ['比賽', 'Na'], ['。', 'PERIODCATEGORY']]
```

三、結論

MONPA 提供繁體中文自然語言處理一個全新的分詞、詞性標註暨命名實體辨識模型，從原始的網頁版進化到現今以 Open Source 釋出的套件版，可以看到全然不同的使用效率及應用效益。套件版於釋出前已經近千萬條短文句的處理測試，並於台灣 NLP 研究圈公開後，四天內已逾 6 百多次的安裝數，Github 專案也收到超過 40 多顆星星的鼓勵。相信本研究及釋出的安裝套件必定能嘉惠我國相關研究與產業，加快繁體中文自然語言處理之進展。

致謝

在此感謝中央研究院中文詞知識庫小組的協助。MONPA 在經中央研究院中文詞知識庫小組同意下，使用 CKIP 斷詞元件[4]輔助製作初期訓練資料。

參考文獻

[1] Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 41(1), 164-171.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] Hsieh, Y. L., Chang, Y. C., Huang, Y. J., Yeh, S. H., Chen, C. H., & Hsu, W. L. (2017, November). MONPA: Multi-objective Named-entity and Part-of-speech Annotator for Chinese using Recurrent Neural Network. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 80-85).

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

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 246-255

使用語者轉換技術於語音合成資料庫之音質改進

Speech Enhancement for TTS Speech Corpora by using Voice Conversion Technologies

林衍廷　Yan-ting Lin
國立臺北大學通訊工程學系
Department of Communication Engineering
National Taipei University
s8303232000@gmail.com

江振宇　Chen-yu Chiang
國立臺北大學通訊工程學系
Department of Communication Engineering
National Taipei University
cychiang@mail.ntpu.edu.tw

摘要

本論文將語者轉換技術用於修復語音資料庫的音質，其原由是在建立文字轉語音系統時，所使用的語音資料庫之部分受限於錄音器材與錄音環境而音質不佳，為了讓這些音質不佳的語料能夠被重新不浪費地使用於建立文字轉語音系統，本論文將利用語者轉換技術於語料庫的音質修復，利用同一語者的特性讓語者轉換的問題轉變成音質轉換的研究問題。在轉換技術上，聲學參數用 WORLD 聲碼器來分析語音訊號，轉換模型用傳統高斯混和模型以及深度學習模型，也嘗試了多種輸入及輸出參數組合，最後也探討以不同語速的語料進行音質修復的結果。客觀以及主觀測試結果顯示，轉換(修復)過的音質有明顯提升。另外也嘗試轉換同一位語者錄製的中英夾雜語料，該語料有一樣的問題，即使是跨語言轉換，實驗結果顯示有降低部分回音和雜訊。

一、緒論

（一）研究動機與方向

在語音研究上，好的語音資料庫對於語音相關的研究是很重要的。在建立文字轉語音系統時需要聲學模型來合成聲音，而在訓練聲學模型時，發現訓練用的語料庫有音質上的問題，使得訓練出來的聲學模型不好，進而影響輸出的音質不好。如果能修復這個

語料，就能減少語料產生的問題並繼續使用這些語料。本研究所使用的語音資料庫是由一位專業女性播音員(縮寫成語者 Tao)讀稿錄製之 4 種語速平行語料庫，總計 1,478 個音檔，語速分為：快、正常、中、以及慢，而有雜訊的語料為語速正常，本研究試著用語者轉換的方法，利用同一語者的特性使得語者轉換變成音質轉換，將語速正常的音質轉成其他語速的音質。現有的轉換技術從高斯混合模型(Gaussion mixture model,GMM)到深度類神經網路(deep neuron network,DNN)等等，能多方嘗試找出最好的轉換模型。另外，本研究也嘗試轉換另一個語者 Tao 所錄製的中英夾雜的語料，該語料有一樣的音質問題，嘗試能否也轉換非同一語言的情形。

（二）語者轉換相關研究文獻探討

語者轉換可以分為平行語料的轉換和非平行語料的轉換兩大區塊討論，平行語料的轉換最廣為人知的就是用 Gaussion mixture model(GMM)[1]來做轉換，將來源音框和目標音框用動態時間扭曲(dynamic time warping,DTW)做對齊後，來訓練高斯混合模型描述不同的發音，還有[2]GMM 加上 maximum likelihood parameter generation(MLPG)的作法，用機率來解這一問題到這邊算是有一個不錯的結果。

再來就是近年流行的機器學習，從[3]開始了用類神經網路(artificial neural networks, ANN)來轉換來源語者到目標語者，在[4]中實驗了很多 DNN 的變種，還有可以用一段訊號作為輸入的 DBLSTM [5]。除了用上述的方式做轉換，還有用 spectral differential[6]的作法，去學習來源語者和目標語者的差異；後來出現了用語音辨識加語音合成的作法[7]，先用語音辨識辨別來源語者的內容，再訓練目標語者的聲學模型，用聲學模型將內容合成聲音，這種作法讓非平行語料的語者轉換有更進一步的突破。

（三）語者轉換架構

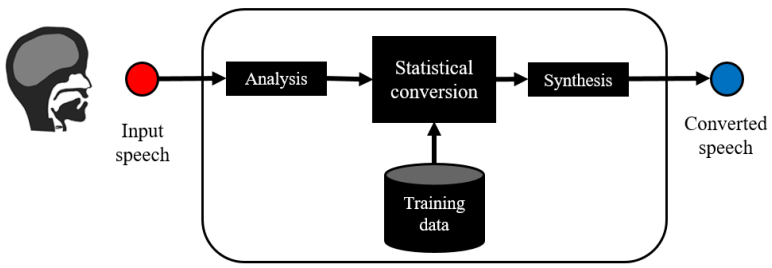


圖 一：語者轉換架構圖

語者轉換作法流程如圖一，輸入訊號經由分析成聲學參數，在訓練資料的部分，訓練的模型都是用來轉換聲學參數，統計式轉換(statistical conversion)所用的模型有 GMM 的機率模型也有神經網路(neuron network, NN)的機器學習模型，都是在將輸入的參數轉換成接近目標的參數，將聲學參數轉換完之後，進入合成器合成出聲音，這就是語者轉換的運作流程。

二、語音資料庫

(i). Treebank-Tao-SR-Corpus：是由一位專業女性播音員(縮寫成語者 Tao)讀稿錄製之 4 種語速平行語料庫，總計 1,478 個音檔，共有 203,746 個音節，語速分為：快、正常、中、以及慢，平均音節長度分別為 0.181 秒、0.198 秒、0.244 秒及 0.264 秒，音檔均為 20kHz 的取樣率及 16-bit 之 PCM 格式，主要內容大多摘錄自新聞、網路文章。

(ii). English-Tao-CEMix-Spell-Corpus：是由語者 Tao 所錄製而成的中英夾雜語料，以中文為主體並穿插英文字母於中文語句中，共 539 個語句，總音節數為 13,540 個音節，包含 11,688 個中文音節與 1,872 個英文字母。音檔為取樣頻率 20,000 赫茲(Hertz)及 16 位元數之 PCM 格式，平均語速為一秒 3.5 個音節。

(iii). English-Tao-CEMix-Word-Corpus：是由語者 Tao 所錄製而成的中英夾雜語料，以中文為主體並穿插英文詞(word)於中文語句中，共 843 個語句，總音節數為 18,103 個音節，包含 15,885 個中文音節與 2,218 個英文音節。音檔為取樣頻率 20,000 赫茲(Hertz)及 16 位元數之 PCM 格式，平均語速為一秒 4.85 個音節。

由於本實驗所需平行的語料，將(i)語料庫中的 4 個語速經文本比對之後，我們用其中 1048 個音檔作為實驗的語料且取樣率為 16kHz。

三、語者轉換技術簡介

（一）聲碼器

聲碼器就是將聲音訊號分解成有意義的參數，像 Mel-cepstral coefficient(MCC)就是將聲音訊號的頻譜取對數後按照人耳對頻率的聽覺敏銳做縮放再做傅立葉逆轉換得到的係數。這個依據人耳特性對頻率軸縮放的就是梅爾刻度(Mel-scale)，每一個刻度都是一維的 MCC。Mel-log spectral approximation filter (MLSA filter)就是將 MCC 轉回頻譜包絡(spectral envelope)，就可以把頻譜包絡和激發訊號卷積來得到聲音訊號。另外，

WORLD [8]聲碼器是近年被認為目前音質最好的聲碼器，它將訊號分成音調(pitch)、頻譜包絡和非週期性(aperiodicity)三個部分，WORLD 將非週期性從頻譜包絡分出來之後，做頻譜包絡的轉換會更好，整體音質會較穩定。

（二）映射函數

1、Gaussian Mixture Model (GMM)

高斯混合模型 GMM，可以很好的近似任意的機率分布，在這邊用來描述聲學參數的機率分布，用於語者轉換。令來源的 MCC 為x_t、目標為y_t，t為音框數，為了考慮音框之間的關聯性，需要一個有前後音框資訊的 delta $= \Delta data = -0.5 * data_{t-1} + 0.5 data_{t+1}$，令$X_t = [x_t \ \ \Delta x_t]$和$Y_t = [y_t \ \ \Delta y_t]$，而$Z_t = [X_t \ \ Y_t]$，假設$Z_t$可以由高斯混合模型表示成$P(Z_t|\lambda^{(Z)})$，$\lambda^{(Z)}$為機率參數。接下來將$P(Z_t|\lambda^{(Z)})$用貝式定理展開，$P(Y_t|X_t, \lambda^{(Z)})$就是將來源轉到目標的映射函數。

$$P(Z_t|\lambda^{(Z)}) = P(X_t, Y_t|\lambda^{(Z)}) = P(Y_t|X_t, \lambda^{(Z)})P(X_t|\lambda^{(Z)}) \tag{3.1}$$

在估計目標$\hat{y}$時，我們會希望 likelihood 最大，也就是$P(Y_t|X_t, \lambda^{(Z)})$最大，寫成

$$\hat{y} = argmax\ logP(Y|X, \lambda^{(Z)}) \tag{3.2}$$

目標函數$Q(Y, \hat{Y})$可寫成式 3.11，為了使目標函數最大，將$Q(Y, \hat{Y})$對y微分算出$\hat{y}$整個過程稱為 maximum likelihood parameter generation(MLPG)。

$$Q(Y, \hat{Y}) = \sum_{all\ m} P(m|X, Y, \lambda^{(Z)}) logP(\hat{Y}, m|X, \lambda^{(Z)}) \tag{3.3}$$

2、Deep Neural Networks (DNN)

類神經網路(Artificial Neural Networks, ANN 亦可以說是 NN)是由多個像神經元的感測器所組成，每個感測器的輸出是由輸入向量與權重向量的內積後，經過一個非線性函數所得的純量，其架構是一層輸入層、一層隱藏層和一層輸出層，而 DNN 就是多層的隱藏層。本研究用的非線性的函數是 sigmoid，訓練整個 DNN 的準則用的是 minimum mean squared error (MMSE)。在語者轉換中，DNN 是音框對音框的轉換，跟 GMM 一樣需要有 delta 作為輔助輸入來考慮前後音框的影響，轉換結果才會比較好。

3、Deep Bidirectional Long Short-Term Memory (DBLSTM)

上述兩種轉換都需要 delta 的資訊來輔助訓練，所以也出現用 Recurrent Neural Networks (RNN)可以考慮整段資料的模型。為了避免 RNN 在資料過長的時候有梯度爆

炸或梯度遺失的問題，而改用 Long Short-Term Memory (LSTM)。本研究使用的是 Deep Bidirectional LSTM (DBLSTM)，DBLSTM 是將多層 BLSTM 疊在一起來加強學習效果。BLSTM 隱藏層分為時序向前和時序向後兩種，這樣的架構可以使下一時刻的輸出考慮前後文的關係，就不需要 delta 作為輔助輸入，$\vec{h}_t$為時序向前，迭代計算$\mathrm{t} = 1$ 到 $\mathrm{t} = \mathrm{T}$；$\overleftarrow{h}_t$為時序向後，迭代計算$\mathrm{t} = \mathrm{T}$ 到 $\mathrm{t} = 1$，$H(*)$表示整個 LSTM 單元過程，BLSTM 單層運算如式(3.4~3.6)，整個 DBLSTM 可以用 back propagation though time(BPTT)算出權重。

$$\vec{h}_t = H(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}) \tag{3.4}$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \tag{3.5}$$

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{3.6}$$

4、Spectral Differential

Spectral differential[6]目的是學x和y的差異，然後作為 MLSA filter 的輸入。令$d_t = y_t - x_t$。在 GMM 中，算出Δd_t，令$D_t = [d_t \;\; \Delta d_t]$，將目標函數寫成下列：

$$\hat{d} = argmax\, P(D|X, \lambda^{(Z)}) \tag{3.7}$$

這個$\hat{d}$就是轉換的x和y距離，最終要合成的 MCC 參數為$y' = x + d'$。在[6]提出的做法中，激發訊號是原始的音訊，然後 MLSA filter 的參數是$\hat{d}$，像是將原本的訊號通過一個 filter，所以音質不會卡在聲碼器的音質上限。

（三）Sprocket toolkit

Sprocket 是一個以 python 為基礎的語者轉換的工具，用的轉換方法是 GMM，而 Sprocket 中的 spectral differential 的作法和[6]中的作法有些出入，Sprocket 並未另外訓練出x對d的轉換模型，而是用原x對y的模型轉出y'後，令$d' = y' - x$來運算，本論文將以 sprocket 的作法來實驗 spectral differential 並和直接轉換的實驗做比較。

四、實驗結果和探討

（一）實驗設計

本節實驗主要為語速正常轉語速中，轉換的方法分為 GMM、DNN 和 DBLSTM，在 GMM 中轉換的目標分為 spectral differential(Diff)和目標語者，另外還有非週期性的轉換。

1、映射函數之比較

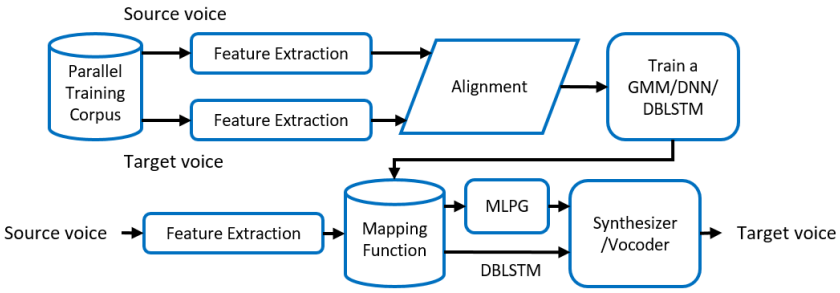


圖 二：GMM/DNN/ DBLSTM 轉換之流程圖

圖二是 GMM/DNN/DBLSTM 語者轉換的流程圖，MCC 的維度都是 40 維，GMM 和 DNN 都需要前後音框的資訊 delta 作為輔助，delta 也是 40 維，所以輸入是 80 維，輸入資料需要做正規化，轉換後的結果經過 MLPG 得到目標的 MCC。DNN 的架構是隱藏層 2 層，每層 2048 個節點。DBLSTM 用雙向訓練來考慮前後音框的影響，一筆資料需要一段的音框最為輸入，這裡用 512 個音框作為一筆資料的長度，整句音檔末端不足 512 的部分會往前湊齊 512 為該音檔的最後一筆資料，而且每個音框的資訊都需要正規化，其架構是隱藏層 2 層，每層 512 個節點，256 個節點向前、256 個節點向後。

2、增加輸入參數之比較

在映射函數比較的實驗中，DBLSTM 來轉換得到較好的結果，但是碰到了在發音邊界會有雜音的問題，為了改善這一問題在模型訓練的輸入端加入了語言參數(language parameter, 以下簡稱 lp)。這語言參數是 64 維 one-hot 的發音標籤(代表音節的聲母和韻母資訊)和 2 維的該音框在該發音中的正規化位置和整段發音的長度，這 66 維的資訊在輸入端時所對應的的目標之語言參數也要作為輸入，因此輸入維度是 172 維，輸出是 40 維的 MCC。另外實驗了加入了兩維的 voice/unvoice(以下簡稱 uv)的資訊，如果第一維是 voice 為 1，是 unvoice 為 0；第二維是 lf0，unvoice 的 lf0 是用前一個 voice 和後一個 voice 的端點做內插取得，也是需要加入對應的目標 uv 資訊，輸入維度是 176 維。

3、加入非週期性轉換之比較

前面都是針對頻譜包絡的轉換，這個實驗是將另一個參數非週期性(aperiodicity, 以下簡稱 ap)也轉換。在來源 MCC 和目標 MCC 動態時間扭曲之後，將其對齊好的索引值拿來對齊 ap(用 MCC 表示)。在轉換 ap 的模型我們也設計兩種，一種是 ap 轉成 ap，

輸入 40 維，輸出 40 維；另一種是加入語言參數的輸入 172 維，輸出 40 維，兩種的訓練除了輸入維度不同，其架構都同 DBLSTM 的實驗。

4、Spectral Differential 與 GMM 轉換之比較

Diff 的做法和 GMM 語者轉換在資料對齊之前都一致，對齊資訊還是依據 MCC 對齊後的索引值，再計算 MCC differential 的部分$d = y - x$，訓練出一個x轉到d的轉換模型，將轉出d'的作為 MLSA filter 的參數，輸入為來源音檔，得出聲音。但 sprocket 中，並未訓練一個x轉到d的轉換模型，而是用x轉到 y 的轉換模型，轉出y'之後，令$d' = y' - x$來得到差異的資訊，本實驗以 sprocket 的作法實踐，並和原 GMM 轉換的聲音做比較。

（二）實驗結果[1]

我們對外部音檔共 25 句來轉到語速中，每一個實驗有客觀評量或主觀評量。客觀評量中，我們並未有正確且長度一致的音檔可以來算 stoi 和 pesq，所以我們提出用 source、target 和 transformed 的 mel-cepstral distortion (MCD)來做相互比較，target 在算 MCD 前須和 source 做 DTW 對齊成 source 的長度。MCD 是描述兩個 MCC 距離的方法， y和x都是 MCC，N 是 MCC 的維度如公式 4.1

$$MCD = \frac{10}{ln\,10}\sqrt{2\sum_{n=1}^{N}(y[n] - x[n])^2} \tag{4.1}$$

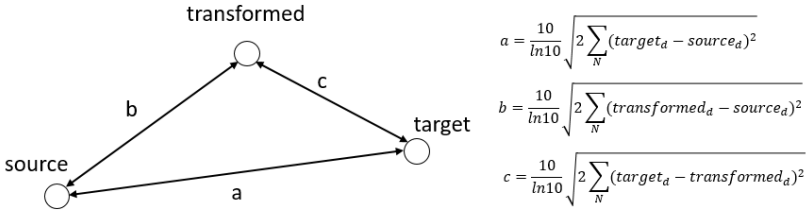


圖 三：MCD 比較示意圖

如圖三將 source、target 和 transformed 的 MCD 算出，只要能符合$a > b > c$，就能保證這個轉換的結果比原來的好，如果只比較$a > c$，那麼以目標為圓心，半徑為 c 的圓都會符合此條件而無法確定圓上的點之間的優劣；另外沒有比較$b > c$的話，也不能確定轉換後的結果是否有比較靠近目標。在做其他變因比較的時候，在符合$a > b > c$情形下比較c也可以看出哪個轉換的效果較優；在主觀測試，我們找了 6 位受測者，每一個實驗從 25 句裡隨機抽 5 句聽測，讓測試者投票哪個音質比較好，再比較票數算比例。

[1] https://drive.google.com/drive/folders/1YsCJNmhw6RFWzfI9DMyiX8ciNBTaAjwu?usp=sharing

1、映射函數之比較結果

表一是 GMM、DNN 和 DBLSTM 的 MCD 比較和票數比較。在客觀評量中，GMM 並未符合$a > b > c$，且b和c還比a長，在聽感上有點像語速中但又有點不像的樣子。DNN 轉換的結果符合$a > b > c$，在距離 c 的差異上 DNN 的表現比較好；在聽覺測試的投票中，73.3%認為 DNN 的音質比 GMM 好。DBLSTM 的轉換結果也符合$a > b > c$。在 DNN 跟 DBLSTM 的比較中，距離 c 的值是 DNN 略小，但在聽覺測試的投票中，僅有 30%認為 DNN 音質較好，70%認為 DBLSTM 音質比 DNN 好，在轉換模型的實驗比較中，DBLSTM 是我們覺得表現較好的。

表 一：映射函數之 MCD 比較和聽覺測試

	a	b	c	聽覺測試	
GMM	7.942	12.815	13.061	26.6%	
DNN	7.942	6.644	5.298	73.3%	30%
DBLSTM	7.942	7.433	5.433		70%

2、增加輸入參數之比較結果

在確認 DBLSTM 是較好的架構之後，我們開始增加輸入參數的實驗，為了使得訓練轉換的結果更接近目標，加入了語言參數- DBLSTM 172 和額外再加入 uv 資訊的- DBLSTM 176，由表二可以看出距離 c 的值有減少了一些；聽覺測試中是 46.6%比 53.3%，聽感上有點接近，必須用耳機聽才能聽出發音邊界雜音的差異。另外增加 voice/unvoice 參數的實驗，由於距離 b 和距離 c 跟 DBLSTM 172 是一樣的所以不做聽力測試。

表 二：增加輸入參數實驗之比較 MCD 和聽覺測試

	a	b	c	聽覺測試
DBLSTM	7.942	7.433	5.433	46.6%
DBLSTM 172	7.942	6.818	4.732	53.3%
DBLSTM 176	7.942	6.818	4.732	

3、加入非週期性轉換之比較結果

在加入 ap 轉換的實驗中，MCC 的轉換是用 DBLSTM 172，所以 MCD 的比較會一致，我們用 ap 加語言參數的實驗來做聽覺測試，聽覺測試的結果 53.3%的票數覺得有轉 ap 的結果聽起來比較好一點，聽感上有轉 ap 的結果比較沒有嗡嗡的聲音。

4、Spectral Differential 與 GMM 轉換之比較結果

因為 GMM Diff 沒有經過聲碼器，所以不做 MCD 比較，只做聽覺測試。聽覺測試的結果 86.6%的票數覺得 Diff 的聲音比較好聽，在沒有經過聲碼器的情況下音質聽起來就像是通過濾波器。

（三）轉到不同語速之音質比較

我們用 DBLSTM172 來實驗語速正常轉到哪種語速的音質比較好，在 MCD 的比較中，由於各語速和語速正常距離並不一致且分布不同，所以只能比較語速正常轉到哪種語速的結果比較像該語速，由表三可以看出語速慢距離語速正常是最遠的，語速快和語速中距離語速正常的長度差不多且語速快的距離 c 比語速中的略小，可以得知語速快和語速中各自的轉換效果差不多好；而在聽覺測試中 42.8%的語速中對上 50%語速快，在聽感上可能語速快會好一些，結論是轉到語速中或是語速快都是不錯的。

表 三：轉到不同語速實驗之 MCD 比較

	a	b	c	聽覺測試
語速中	7.942	6.818	4.732	42.8%
語速快	7.976	6.530	4.710	50%
語速慢	8.632	6.706	5.085	7.1%

（四）中文轉換模型對中英夾雜語料之嘗試

在確認中文對中文的轉換上音質確實有變好之後，我們對 CE_word 和 CE_spell 分別做語者轉換，轉換模型是 DBLSTM，輸入是 40 維。由於是轉成中英夾雜的句子，沒有內容一致且音質較好的音檔可供計算 MCD，只能做聽覺測試。CE_spell 投票結果顯示轉得並不好，聽感上回音的部分少了很多但有點破嗓，整體感覺還是沒有原本好；CE_word 投票結果各 50%，有改善到某些低頻雜音。雖然整體實驗效果沒有特別顯著，但我們發現即使原本中文轉換模型沒看過英文的發音，在轉換的過程還是會轉成類似的發音，使得音檔中英文的部分還是能的聽出來。

五、結論

本論文提出用語者轉換的技術用於修復語音資料庫，藉由同一位語者的特性使得語者轉換變得像是音質轉換，經過研究和分析得出以下結論：(1)在轉換模型的架構上，

DBLSTM 的表現比起 GMM 和 DNN 都來的好；(2)藉由文本的發音標記和發音位置做為輸入來輔助訓練模型，確實能讓結果更近似於目標聲音；(3) 非週期性的資料對齊應跟隨頻譜包絡，且轉換非週期性對於音質有一定的幫助；(4)以本論文所使用的 4 種語速語料庫而言，將語速正常轉到語速快或是語速中是較為恰當的，音質也比較好；(5)跨語言的轉換中，發音相近的會共用相同的轉換對。本研究嘗試用語者轉換技術在同一語者的語料庫進行音質修復，由主觀測試結果可得知，在 DBLSTM、DBLSTM 172 和加入 ap 轉換的實驗中皆有改善音質和減少雜訊。

參考文獻

[1] Y. Stylianou, O. Capp´e, and E. Moulines, “Continuous Probabilistic Transform for Voice Conversion,” in IEEE Trans. on Audio, Speech and Language Processing, 1998

[2] T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” in IEEE Trans. on Audio, Speech and Language Processing, 2007

[3] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectral Mapping Using Artificial Neural Networks for Voice Conversion,” in IEEE Trans. on Audio, Speech and Language Processing, 2010

[4] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, “Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training,” in IEEE Trans. on Audio, Speech and Language Processing, 2014

[5] L. Sun, S. Kang, K. Li and H. Meng, “Voice Conversion Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks,” in ICASSP, Apr. 2015.

[6] K. Kobayashi, T. Toda, S. Nakamura, “F0 Transformation Techniques for Statistical Voice Conversion with Directwaveform Modification with Spectral Differential,” in IEEE Spoken Language Technology Workshop, Dec. 2016.

[7] L. Sun, K. Li, H. Wang, S. Kang and H. Meng, “Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training,” in ICME, July. 2016.

[8] M. MORISE, F. YOKOMORI, K. OZAWA, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” in IEICE, 2016

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 246-255

即時中文語音合成系統

Real-Time Mandarin Speech Synthesis System

鄭安傑 An-Chieh Cheng
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
ajcheng@g-mail.nsysu.edu.tw

陳嘉平 Chia-Ping Chen
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
cpchen@mail.cse.nsysu.edu.tw

摘要

本論文研究與實作即時中文語音合成系統。此一系統採用文字序列到梅爾頻譜序列的轉換模型，再串接一個從梅爾頻譜到合成語音的聲碼器。我們使用 Tacotron2 實作序列到序列轉換模型，配合數種不同的聲碼器，包括 Griffin-Lim，World-Vocoder，與 WaveGlow。其中以實作可逆編碼解碼函數的 WaveGlow 神經網路聲碼器最為突出，無論在合成速度或語音品質方面，皆令人印象深刻。我們使用單人 12 小時的標貝語料實作系統。在語音品質方面，使用 WaveGlow 聲碼器的合成系統語音的 MOS 為 4.08，略低於真實語音的 4.41，而遠勝另兩種聲碼器（平均 2.93）。在處理速度方面，若使用 GeForce RTX 2080 TI GPU，使用 WaveGlow 聲碼器的合成系統產生 10 秒 48 kHz 的語音僅需 1.4 秒，故為即時系統。

Abstract

This thesis studies and implements the real time Chinese speech synthesis system. This system uses a conversion model of the text sequence to the Mel spectrum sequence, and then

concatenates a vocoder from the Mel spectrum to the synthesized speech. We use Tacotron2 to implement a sequence-to-sequence conversion model with several different vocoders, including Griffin-Lim, World-Vocoder, and WaveGlow. The WaveGlow neural network vocoder, which implements the reversible codec function, is the most prominent, and is impressive in terms of synthesis speed or speech quality. We use a single speaker with 12-hour corpus implementation system. In terms of voice quality, the MOS of the synthesized system voice using the WaveGlow vocoder is 4.08, which is slightly lower than the 4.41 of the real voice, and far better than the other two vocoders (average 2.93). In terms of processing speed, if the GeForce RTX 2080 TI GPU is used, the synthesis system using the WaveGlow vocoder produces a voice of 10 seconds and 48 kHz in 1.4 seconds, so it is a real time system.

關鍵詞：文字轉語音，Tacotron2,WaveGlow

Keywords: TTS,Tacotron2,WaveGlow

一、緒論

隨著科技的發展，人機互動的情況也越來越普及，像是 siri 語音助理、智能導航、有聲讀物等都已環繞在我們生活裡。而其中，語音合成的技術就扮演了一個非常重要的腳色。語音合成是透過機械、電子的方式產生人造語音的技術，文字轉語音技術也隸屬於語音合成。而本研究則是致力於開發出一個可合成出更快且更為逼真的文字轉語音合成系統。實現語音合成的方法有多種，其中包含參數式合成以及拼接式合成。基於參數式的語音合成系統主要是透過統計學模型，利用學習出來的語音學特徵和其聲學特徵的對應關係後，預測出相應的參數，接著聲碼器再透過這些參數合成出所期望的音頻。不過這種合成方式最大缺點乃為無法合出接近人類的自然語音，在技術上尚未有明顯的突破。拼接式語音合成系統則是透過同樣的方式去預測出這些聲學特徵，然後再到原始語音庫中找尋近似的音素，最後拼接而成。不過這種合成方法也意味著合成的音質穩定性與語音庫大小成正比。若要能合成出完善的自然語句，就必須要有齊全的資料庫，且同時為了不延遲搜尋上的效率，更必須要有個良好的演算法。而上述這些方法，除了皆有著明顯的

人工痕跡之外，在專業領域上的門檻也都極高。幸運地，還有一種神經網路式的合成技術，可利用神經網路直接學習從文本端到聲學特徵端的對應關係。

二、研究方法

（一）資料集

訓練資料選自標貝資料集，是由「標貝科技有限公司」於 2018 年所開放。由一位女性錄音者錄製而成，全長約略 12 小時，使用 48kHz 16bit 採樣頻率，錄製環境為專業錄音室及錄音軟體，語料涵蓋各類新聞、小說、科技、娛樂等領域，詳細規格如表一。

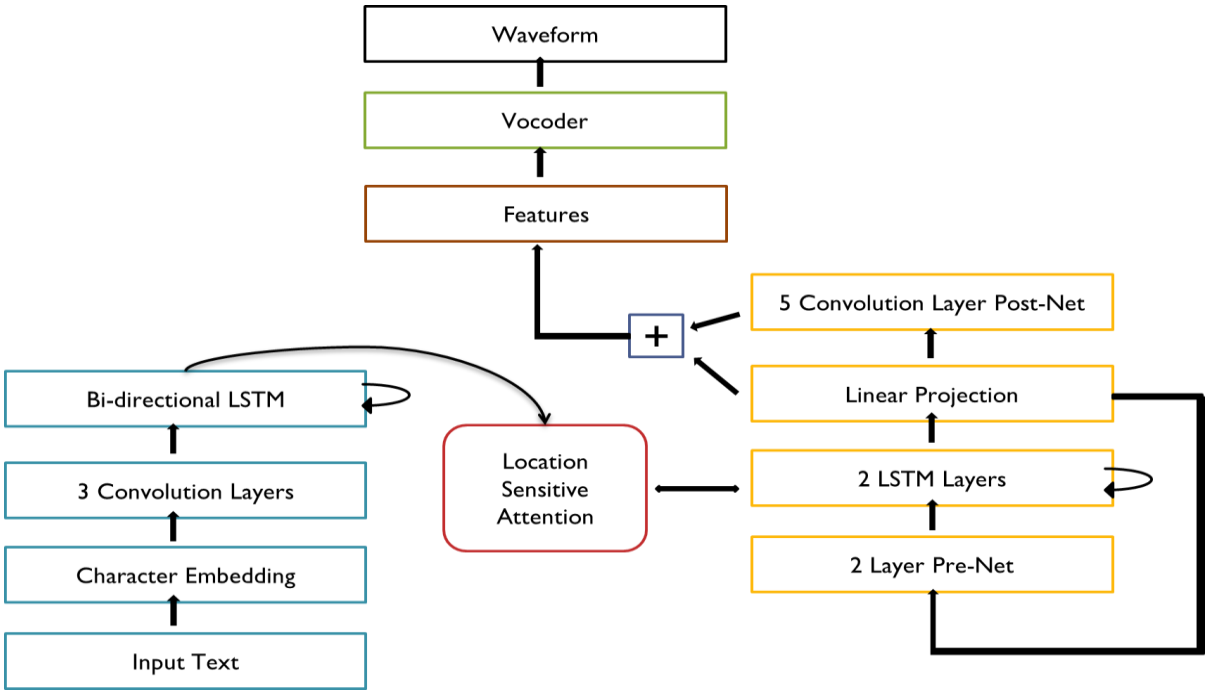
表一、標貝資料集數據規格

數據規格	
數據內容	中文標準女聲語音數據庫
錄音語料	綜合語料樣本量:音節音子的數量、類型、音調、音連以及韻律等進行覆蓋。
有效時長	約 12 小時
平均字數	16 字
語言類型	標準普通話
發音人	女 : 20-30 歲
錄音環境	聲音採集環境為專業錄音室 1. 錄音室符合專業音庫錄製標準 2. 錄音環境和設備自始至終保持不變 3. 錄音環境的信造比不低於 35dB
錄製工具	專業錄音設備及錄音軟體
採樣格式	無壓縮 PCM WAV 格式，採樣率為 48kHz、16bit。

（二）預測網路

在前端預測網路部分我們重現了 Google 的 Tacotron2 [1]，並針對中文語音合成系統做了客製化。在架構上面使用的是一個編碼器-解碼器(Encoder-Decoder)的設置，並加入了位置敏感的注意力機制(Location sensitive attention) [2]，整體架構如圖一。而由於我們使用的是中文資料集，故在文本的內容上先進行了資料的預處理，目的是為了讓神經網

路可以學習到我們中文上的韻律以及抑揚頓挫。由於漢字本身有數萬個相異字，同音異字的情況也不在少數，若以窮舉的方式來對神經網路做訓練顯然不夠明智。我們處理的方式是使用漢語拼音作為字元標註，並採用數字一到四來表示我們的聲調。雖然過程中都是以這樣的形式進行訓練，不過在合成階段時我們可以藉由”pypinyin”的套件直接透過中文輸入合成出所指定的句子。另外，為了提升中文語音的流暢度，我們也透過”jieba”斷詞系統針對文本的內容先進行斷詞，我們利用 TrieTree 的結構去生成句子中所有可能成為詞的情況，並使用動態規劃的方式找出最大機率的路徑。整個前處理的過程可參考表二。



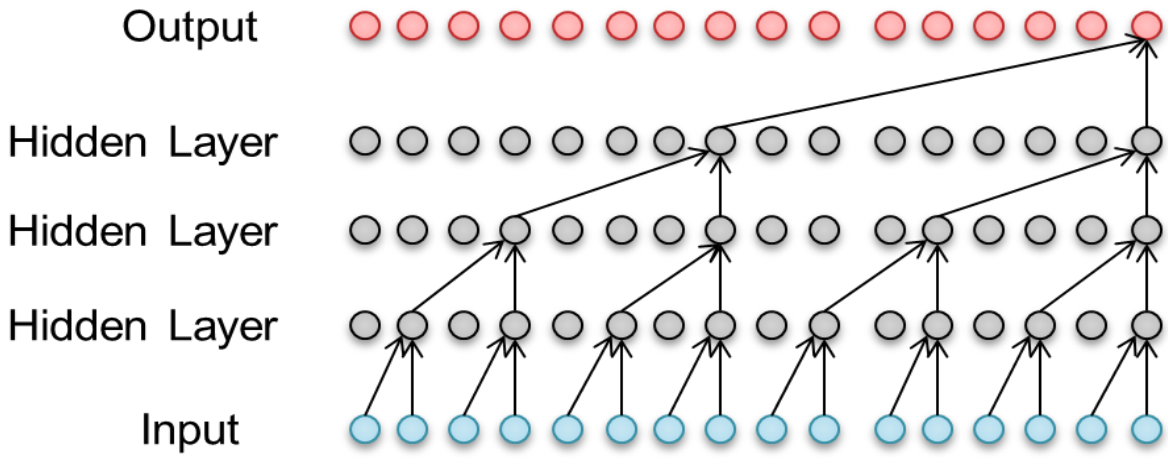
圖一、預測網路之模型架構

表二、資料前處理

原始文本	可想而知，甕中捉鱉顯然比亡羊補牢更可靠更有效。
經過 jieba	可想而知|，|甕中捉鱉|顯然|比|亡羊補牢|更|可靠|更|有效|。
經過 pypinyin	ke3 xiang3 er2 zhi1|，| weng4 zhong1 zhuo1 bie1 | xian3 ran2 | bi3 | wang2 yang2 bu3 lao2 | geng4| ke3 kao4 | geng4| you3 xiao4 |.

（三）聲碼器

Tacotron2 [1]預設的聲碼器為 WaveNet [3]，是使用 Auto-regression 的方式生成音頻，即每在預測當前時刻的值時都是根據前一時刻的輸出結果。模型架構主要是由因果卷積(Causal Convolution)組成，而為了能在時域上獲得更廣的感知能力，模型中加入了擴大卷積(Dilated Causal Convolution) [4]，如圖二，當層數疊加，感知能力就以指數性成長。雖然這樣的模型架構能夠重現極為逼真的人類語音，也在語音合成上達到了很好的效果，但美中不足的卻是其生成速率。根據我們的評估，得花費數十秒的合成時間才能生成一秒鐘的音頻，若要作為實際應用，尚有大幅度的調整空間。而為了使系統能夠即時合成，我們根據了 Tacotron2 [1]的前端預測網路，並針對多種不同的聲碼器進行實測、探討。



圖二、Dilated Causal Convolution

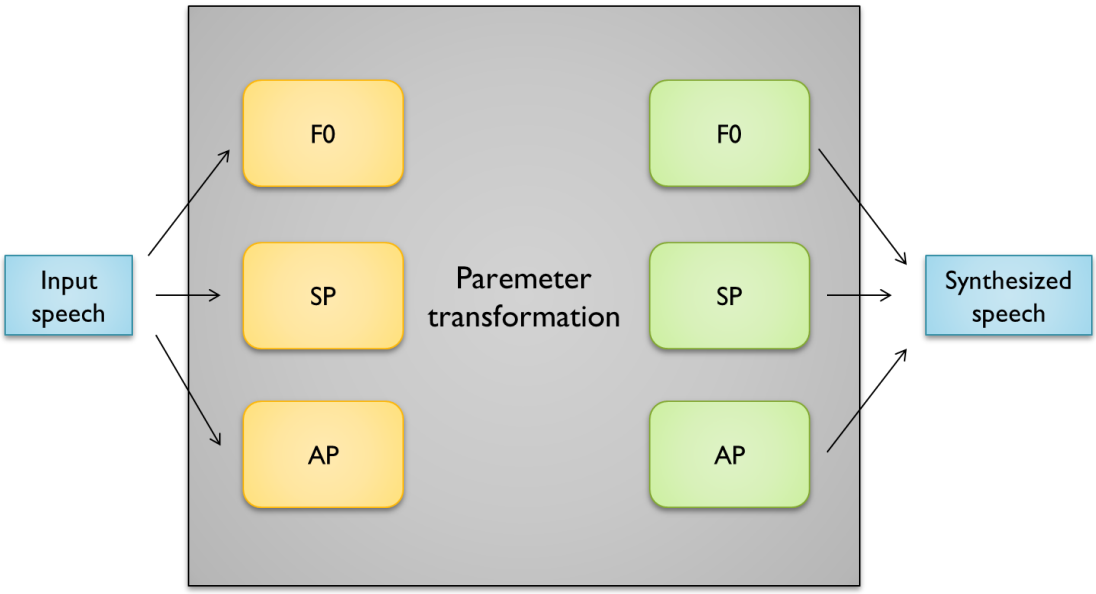
1、Griffin-Lim

Griffin-Lim [5]是一種迭代的演算法，音頻質量雖然不如 WaveNet [3]，但在即時系統中仍保有了競爭力，是許多即時語音合成系統比較的對象。透過迭代的次數來提升合成的音質，我們的實驗中採用了六十次的迭代以確保其穩定性。比起傳統聲碼器需要基頻和倒譜等參數而言，Griffin-Lim [5]可根據文本預測的線性頻譜圖直接重建時域波形。

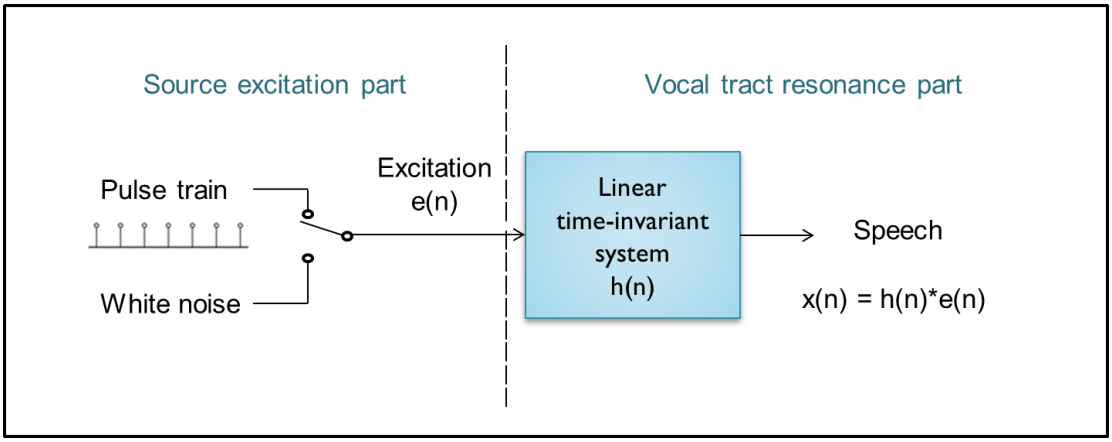
2、World-Vocoder

語音是聲音的一種，是由人的發聲器官發出，具有一定語法和意義的聲音。大腦對發音

器官發出運動神經指令，控制發音器官各種肌肉運動從而振動空氣而形成。整體發聲過程是空氣由肺進入喉部，經過聲帶激勵，進入聲道，最後通過嘴唇輻射形成語音。World-Vocoder [6]參照了此發聲原理，分別將三種聲學特徵:基頻(Fundamental Frequency)、頻譜包絡(Spectral envelope)以及非週期序列(Aperiodic parameter)對應到了人發聲機理的經典源-濾波器(source-filter)模型，最後並利用這些參數重建語音，如圖三、圖四。而在特徵提取的過程，我們先是透過 DIO 演算法提取出基頻，然後利用基頻中的 CheapTrick 提取包絡，最後透過 D4C 將得到後的基頻與包絡計算出非週期信號。不過為了與 Tacotron2 [1]訓練做結合，這種高維度的頻譜包絡以及非週期信號我們必須先將其降維，以緩解神經網路訓練時所帶來的壓力。而透過 Merlin 工具包可幫助實現維度的轉換。以我們的實驗為例，首先將提取到的 MFCC 降至 60 維，接著將非週期序列轉變成 Band 非週期信號，此步驟可有效將維度降至 5 維，基頻部分則保持不變，最後再將上述特徵維度連接起來，輸入 Tacotron2 [1]模型進而實現神經網路之訓練。



圖三、World-Vocoder 透過 F0、SP、AP 重建語音信號



圖四、人發聲機理的經典源-濾波器模型

F0、AP、SP 分別對應到脈衝序列(Pulse train)、非週期序列(White noise)以及聲道諧振部分的 h(n)

3、WaveGlow

隨著神經網路的發展，目前常使用到的生成模型可有:自回歸模型(Autoregressive model)、生成對抗網路(GAN)以及基於流的生成模型(Flow-based generative model) [7]。儘管自回歸模型在許多實驗上得到了很好的效果，但這種一次生成一個樣本的生成方式除了需要龐大的計算資源之外，在可平行性上也受到了限制。相對而言，生成對抗網路則免除了這種煩惱，主要透過生成器與判別器不斷相互學習的迭代從而生得與真實樣本接近的分布。但一直以來生成對抗網路也不免會遇到許多問題，像是生成的多樣性不足以及訓練過程不穩定等等。不過幸運地，基於流的生成模型有效的解決了這些問題，而這樣的生成方式，也被採用在聲碼器-WaveGlow [7]中。

WaveGlow [7]透過分布採樣生成語音，僅需一個網路及一個最大化似然的損失函數即可生成時域波形，並且在高還原度的情況下亦能即時的合成語音。利用多個可逆的變換函數組成序列，將一個簡單的分布透過一系列的可逆函數轉換到一個複雜的分布，並藉此來模擬訓練數據的分布，最後再透過最大似然準則來進行優化。

我們使用的網路架構比照了 [7]中的配置，包含 12 層的對耦映射層、12 個 1*1 的可逆卷積以及 WN 中設有 8 層的 dilated convolutions，同時根據我們的資料集調整了超參數。訓練資料為 48kHz 的音頻，我們將 160 維的梅爾頻譜作為輸入以及 FFT_size、hop_size、window_size 都設定了相符的格式以便訓練，如表三所示。

表三、Biaobei 資料集超參數設定

資料集 ：Biaobei	
sample_rate(Hz)	48k
num_mels	160
FFT_size	4096
hop_size	600
window_size	2400

三、實驗結果

（一）音頻質量

在模型訓練完畢之後，我們從資料集裡面隨機選取了五句與訓練資料不重複的文本進行評估，受測人員一共十位。受測準則如下:每人聽測五種不同句子，每種句子各包含四個音檔，分別來自真實數據以及 World-Vocoder [6]、Griffin-Lim [5]、WaveGlow [7]三種不同聲碼器合成的音檔。在每一句聽完後，都給予一到五分的主觀分數，總計後再平均計算。而整個過程共包含五種句子以及二十個不重複的音檔，測試結果如表四。

表四、Mean Opinion Scores

Model	Mean Opinion Score(MOS)
World-Vocoder	2.71
Griffin-Lim	3.15
WaveGlow	4.08
Ground Truth	4.41

（二）推斷速度

一個高質量的音頻至少需要擁有 16kHz 的採樣點。而我們的實驗從前端預測網路到後端生成語音不僅都符合了標準，甚至都展現了比實時合成還要快的速度。訓練完的模型我們統一放到了 GeForce RTX 2080 TI GPU 上進行推測。以合成一個十秒且 48kHz 的音頻來說，我們分別在 World-Vocoder [6] : 6 秒，Griffin-Lim [5] : 1.2 秒，WaveGlow [7] : 1.4 秒的時間內完成了推斷。另外，我們也整理了 Tacotron2 [1]預測網路搭配不同聲碼器在同一台機器上分別所佔用的資源，雖然 WaveGlow [7]在推測時間上展現了優異的合成

速度，但由表五可看出其所佔據的資源也相當高，說明了我們的模型仍有優化的可能。

表五、Tacotron2 結合不同聲碼器所佔用的計算資源

Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz/64GB/RTX 2080TI	World-Vocoder	Griffin-Lim	WaveGlow
GPU 使用量(GB)	0.93G	1.31G	2.5G
CPU 使用率(%)	6.75%	7.5%	14.3%
MEM 使用率(%)	2.45%	3.25%	5%

四、結論

我們的研究目前在聲碼器上嘗試了多種可能，包含:World-Vocoder [6]、Griffin-Lim [5]以及 WaveGlow [7]，並將這些合成技術都套用在我們的預測網路上。從我們的研究來看，我們發現儘管 World-Vocoder 及 Griffin-Lim 都已開發一段時間，但在音頻的還原度上仍遠不及近期興起的神經網路式合成器，且 WaveGlow [7]不僅在音質的還原度亦或是合成的速度上(在 2080TI 上，一秒約莫可生成 350kHz 以上的採樣點)都給予了我們不錯的展示。但就長期而言，我們的模型還有多種可能的優化方式，像是我們使用的中文資料集規模較小，儘管經過了斷詞系統的調整，仍有部分語句無法良好的呈現人類的自然語音。日後除了收集更完整的語料庫之外，在預測網路部分加入情緒辨識作為條件，使得合成的音頻更生動更有溫度也是我們的任務之一。

參考文獻

[1] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” 於 *arXiv:1712.05884*, 2017.

[2] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio, “Attention-Based Models for Speech Recognition,” 於 *arXiv:1506.07503v1* , 2015.

[3] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio,” 於 *arXiv:1609.03499v2*, 2016.

[4] Fisher Yu, Vladlen Koltun, “Multi-Scale Context Aggregation by Dilated Convolutions,” 於 *arXiv:1511.07122v3*, 2015.

[5] Daniel Griffin and Jae Lim, “Signal estimation from modified short-time fourier transform,” 於 *IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243*, 1984.

[6] Masanori MORISE,Fumiya YOKOMORI,Kenji OZAWA, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” 於 *IEICE Trans. on Information and Systems, vol. 99, no. 7, pp. 1877–1884*, 2016.

[7] Ryan Prenger, Rafael Valle, Bryan Catanzaro, “WaveGlow: A Flow-based Generative Network for Speech Synthesis,” 於 *arXiv:1811.00002v1*, 2018.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 266-280

探究端對端語音辨識於發音檢測與診斷

Investigating on Computer-Assisted Pronunciation Training Leveraging End-to-End Speech Recognition Techniques

張修瑞 Hsiu-Jui Chang, 羅天宏 Tien-Hong Lo, 劉慈恩 Tzu-En Liu, 陳柏琳 Berlin Chen

國立臺灣師範大學資訊工程學系

{ftes90015, teinhonglo, hane0131}@gmail.com
berlin@csie.ntnu.edu.tw

摘要

電腦輔助發音系統(Computer assisted pronunciation techniques, CAPT)，任務可分為錯誤發音檢測(Mispronunciation detection)以及錯誤發音診斷(Mispronunciation diagnosis)。在過往的研究中，這兩種任務主要依賴於傳統語音辨識系統的強制對齊(Forced alignment)方法，並利用強制對齊產生的音素(Phone)段落與觀測到的全部音素或較混淆的音素計算 GOP (Goodness of pronunciation)分數，並以此作為發音好壞的依據。然而傳統語音辨識系統的訓練流程既冗長且複雜。近年來，端對端語音辨識系統不僅大幅簡化此問題，且效能也有追上傳統語音辨識的趨勢。因此，本論文將基於端對端架構下，分別探討(1) 基於辨識產生的信心分數(Confidence score)；(2) 基於語音辨識結果，兩者對於發音檢測任務的影響。實驗結果顯示，使用端對端架構進行發音檢測與診斷，不僅相較於以往基於傳統語音辨識架構有更少的訓練流程，也大幅提升檢測與診斷的效果。

Abstract

One of the primary tasks of a computer-assisted the pronunciation techniques (CAPT) system is mispronunciation detection and diagnosis. Previous research on CAPT mostly relies on a forced-alignment procedure which is usually conducted with the acoustic models adopted from a traditional speech recognition system, in conjunction with a phoneme paragraph, to calculate the goodness of pronunciation (GOP) scores for the phonemes of spoken words with respect to a text prompt. However, the training process of the traditional speech recognition system is complicated. In recent years, the end-to-end speech recognition system has not only greatly simplified this problem, but also has the trend of catching up with

traditional speech recognition. In view of this, this thesis sets out to conduct mispronunciation detection and diagnosis on the strength of end-to-end speech recognition. To this end, we design and develop two mispronunciation detection methods: 1) method leveraging a recognition confidence measure; 2) method simply based speech recognition results; A series of experiments showed that leveraging end-to-end speech recognition architecture on mispronunciation detection and diagnosis not only reduced the training steps originally required for traditional speech recognition but also improve the performance of detection and diagnosis significantly.

關鍵詞：端對端語音辨識、聲學模型、發音檢測、發音診斷

Keywords: end-to-end speech recognition, acoustic model, mispronunciation detection, mispronunciation diagnosis.

一、緒論

在國際化的時代中，學習外語變成了不可或缺的一部份。當今的人們至少需要學習兩種或兩種以上的語言，而語言學習的過程可分為聽、說、讀和寫四個部分，其中以說和寫最為需要專家知識才得以判斷學習的程度。然而大多數的學習平台較注重於聽力以及閱讀練習，口說的練習則通常藉由教學影片讓使用者複誦，缺乏即時的回饋；此機制使學習者較不易發現發音錯誤，對於學習效果有限，因此電腦輔助發音訓練的研究更顯重要。我們希望電腦具備與專業師資相當的聽力，檢測出學習者的發音錯誤，並且給予回饋，使學習者能夠藉由反覆的練習使口說能力更為進步。

電腦輔助發音的研究與語音辨識器息息相關，過去研究中主要依賴傳統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)語音辨識架構。該架構主要由聲學模型(Acoustic model)、語言模型(Language model)、發音詞典(Pronunciation lexicon)所組成，並且在訓練的過程中，必須先由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) [1][2]，取得聲音與文字的強制對齊，才得以訓練 DNN-HMM 聲學模型，而目前常用的類神經網路包含多層感知器(Multiple-layer perceptron, MLP)[1][3]、摺積式類神經網路(Convolutional neural networks, CNN)[4]、遞迴式類神經網路(Recurrent neural

network, RNN)、長短期記憶類神經網路(Long short-term memory, LSTM)[5][6]、時延式類神經網路(Time delay neural network, TDNN)[7]以及這些類神經網路的延伸。

傳統語音辨識具有下列幾點問題：(1)訓練流程與多個模組有關連，無法清楚知道哪一部分影響了語音辨識的效果；(2)需要較多的語言及語音知識將詞彙對到相對應的音素序列來產生發音詞典，以及音素的上下文相關決策樹；(3)聲學模型和語言模型各自使用不同準則分開訓練，導致語音辨識或其應用任務的最後評估標準不一致。近年來端對端語音辨識大幅簡化了傳統語音辨識繁複的訓練流程。其主流為連結時序分類(Connectionist temporal classification, CTC)以及注意力模型(Attention model)兩種方法。前者 CTC 訓練準則使得聲學模型可直接將聲學特徵僅透過類神經網路輸出對應到的標籤序列，通常為字符(Character)或音素[5][8]，並且於解碼時可以不需要使用語言模型。而另一方面，鑑於 CTC 對於端對端語音辨識的成功，且注意力模型已在其他領域被廣泛應用[9][10]，[11]將注意力模型應用於語音辨識的任務上，並得到與 CTC 模型可比較的結果，但在少量語料下仍遜於 DNN-HMM 的效能。[12][13]藉由多任務學習的方法結合 CTC 與注意力模型，希望 CTC-Attention 模型利用 CTC 彌補注意力模型對齊錯誤(Misalignment)及收斂慢的問題。實驗結果顯示，CTC-Attention 模型可在缺乏語料的情況下，更接近甚至低於 DNN-HMM 模型的辨識錯誤率。

本篇論文中，主要探討端對端聲學模型如何應用於電腦輔助發音檢測，其主要任務分為錯誤發音檢測(Mispronunciation detection)以及錯誤發音診斷(Mispronunciation diagnosis)。錯誤發音檢測任務希望藉由學習者朗誦第二外語口說教材，在已知朗誦內容的情況下，由電腦評判學習者的發音是否正確。在過往的發音檢測實驗中，都是利用傳統聲學模型的事後機率(Posterior probability)、對數相似度值(Log-likelihood) [14]，或是 GOP [15][16]作為發音檢測特徵，以此判斷發音的對錯。而發音診斷則是當學習者發音出現錯誤時系統所給予的糾正。假設所希望聽到的是「國語(guo2 yu3)」，但當學習者唸成「狗語(gou3 yu3)」，系統除了能得知學習者發音出錯，也能反饋學習者的「國(guo2) 」唸錯成「狗 gou3」了。過往也有學者將發音檢測與診斷視為語音辨識的任務如[17][18][19]。基於端對端架構的發音檢測方法較為稀少，僅[19]提出以 CTC 進行英

文發音檢測與診斷。因此本篇論文將初步探討端對端架構應用於華語 CAPT 任務。實驗中將分別使用信心分數以及語音辨識結果進行發音檢測。在實驗結果顯示利用信心分數進行錯誤發音檢測，能夠在錯誤檢測上達到與 GOP 相同的效果，然而整體來說較 GOP 的判斷更為嚴格。另外直接視為語音辨識的方法超越了[19]，使得檢測與診斷都達到最好的效果。

二、端對端語音辨識技術

隨著近年來端對端語音辨識技術的發展，端對端聲學模型的辨識率已與傳統聲學模型不相上下，以下將針對近年來主要的端對端語音辨識方法進行說明。

2.1 連結時序分類(CTC)

連結時序分類最早於 2006 年提出[20]，作為取代傳統聲學模型訓練使用交互熵的損失函數希望最小化$-\ln \mathrm{P}_{\mathrm{ctc}}(C^*|X)$，即輸出越接近真實標記越好。常見應用於音素辨識以及手寫辨識。其概念為給定一段長度為 T 的聲學特徵序列 X 及一段長度 L 的標籤序列 C，其中$C = \{c_l \in U | l = 1, \dots, L\}$，U 為存在的標籤集合。並且 CTC 在訓練時引入了額外的空白標籤，作為標籤間的分界，每個音框的標籤序列可表示為$S = \{s_t \epsilon U \cup \{< blank >\} | t = 1, \dots T\}$，其損失函數可表示為：

$$\mathrm{P}_{\mathrm{ctc}}(C|X) \approx \sum_s \prod_{t=1}^{T} \mathrm{P}(s_t|s_{t-1}, C)\mathrm{P}(s_t|X) \tag{1}$$

其中$\mathrm{P}(s_t|s_{t-1}, C)$代表狀態轉移機率，$\mathrm{P}(s_t|X)$則為 Softmax 輸出的結果。

2.2 注意力模型(Attention model)

沿用上一小節中符號設定，注意力模型目標函式可定義為：

$$\mathrm{P}_{\mathrm{att}}(C|X) = \prod_{l=1}^{L} \mathrm{P}(c_l|X, c_{1:l-1}) \tag{2}$$

同樣也希望直接估測聲學特徵對應到標籤的事後機率，然而與 CTC 不同在於注意力模型並無條件獨立的假設，如上式 2 所示，每一當前輸出皆考慮過去的輸出。$\mathrm{P}(\mathrm{c_l}|\mathrm{X}, \mathrm{c_{1:l-1}})$可以由下列式子推得：

$$\boldsymbol{h}_t = \mathrm{Encoder}(X) \tag{3}$$

$$e_{lt} = \mathrm{Attention}(\mathbf{q}_{\mathrm{l}-1}, \mathbf{h}_{\mathrm{t}}, a_{l-1}) \tag{4}$$

$$a_{lt} = \frac{\exp(\gamma e_{lt})}{\sum_l \exp(\gamma e_{lt})} \tag{5}$$

$$\mathbf{r}_l = \sum_{\mathrm{t}=1}^{T} a_{lt}\mathbf{h}_{\mathrm{t}} \tag{6}$$

$$\mathrm{p}(c_l|X, c_{1:l-1}) = \mathrm{Decoder}(\mathbf{r}_l, \mathbf{q}_l, c_{l-1}) \tag{7}$$

其中$\mathbf{h}_{\mathrm{t}}$為 Encoder 的隱藏狀態向量，a_{lt}為注意力權重由e_{lt}經由 Softmax 函數得到，而 γ為 Sharpen Factor，目的在強調權重的分佈，**q** 代表的是前一個 Decoder 隱藏狀態向量。可以想成 **q** 是 Query，**h** 是 Key Value 然後我們透過任意注意力的機制計算注意力權重a_{lt}，同樣地，注意力模型訓練的損失函數也希望最小化$-\ln \mathrm{P}_{\mathrm{att}}(C^*|X)$。

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

由於注意力模型有著非單調的左到右對齊和收斂較慢的缺點，CTC 則是必須使用額外的語言模型才能有較好的效果。因此有學者也將兩者結合[12] [13]，以 CTC 給予注意力模型更強的左到右限制，並在進行光束解碼時(Beam search)同時加入兩種模型的輸出，以達到更佳的解碼效果。訓練時以λ作為兩模型混和參數，其損失函數為：

$$\mathcal{L}_{\mathrm{CTC-ATT}} = -\big(\lambda \ln \mathrm{P}_{\mathrm{ctc}}(C|X) + (1-\lambda)\ln \mathrm{P}_{\mathrm{att}}(C|X)\big) \tag{8}$$

三、端對端語音辨識技術於發音檢測與診斷

基於檢測與診斷可被視為語音辨識任務的想法，一個能夠辨識清楚第一語言者與第二語言的學習者所發出音素差異的辨識器，將對於發音檢測與診斷有很大的突破。利用這樣的方法不僅可以同時進行發音診斷與檢測，也省去以往必須再藉由發音分數評估的兩階段步驟。以下將針對端對端語音辨識發音檢測與診斷方法進行說明。

3.1 基於分數之發音檢測

過去進行檢測的首要步驟在於進行強制對齊，無論發出對與錯的音素皆解碼成目標音素，並標記音素出現於聲音之時間段。在端對端的語音辨識採用光束搜尋算法，輸出時經由 Softmax 函數輸出所有標籤之事後機率，保留每一次輸出前 n 高值直到出現句尾符號 <eos>為止。為了達到在搜尋時能產生我們所想要的目標音素，使用了限制解碼的方法，即在每一次 Softmax 函數輸出時只關注我們所想要的音素集合，如下圖 1 所示。在解碼的過程中，找出音素事後機率總和最高之音素組合，作為最終想要的目標音素序列，並記錄每一音素之事後機率$\mathrm{P}(c^*|x)$。得到音素事後機率可帶入式 9 決策函數$\mathrm{D}(\mathrm{P}(c^*|\mathbf{x}))$，使音素事後機率投影到 0-1 的範圍，並根據門檻值τ決定發音好壞。

$$\mathrm{D}(\mathrm{P}(c^*|\mathbf{x})) = \frac{1}{1+\exp(\mathrm{P}(c^*|\mathbf{x}))} \tag{9}$$

$$\mathbb{I}\big(\mathrm{D}(\mathrm{P}(c^*|\mathbf{x}))\big) = \begin{cases} 1 \text{ if } \mathrm{D}(\mathrm{P}(c^*|\mathbf{x})) \geq \tau \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

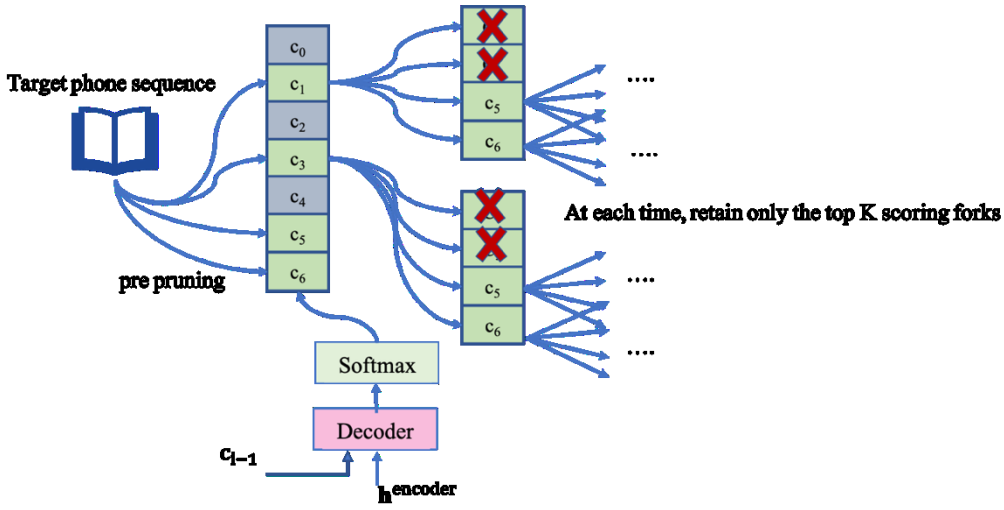


圖1、限制解碼流程

3.2 基於辨識結果之發音檢測與診斷

電腦輔助發音檢測可被視為語音辨識的問題，當語音模型的辨識率為百分之百，發音檢測的問題便可以被解決。儘管當前語音辨識技術仍達不到完美，但已十分進步。在本節中，我們將語音辨識結果與目標語句文字以最短編緝距離演算法(Edit distance)進行對齊

[21]。如下圖 2 所示，紅色箭號代表替換錯誤，藍色箭號為刪除錯誤，綠色箭號為插入錯誤，黑色箭號則為與目標相符。發生替換錯誤與刪除錯誤時則代表發生了發音錯誤，而插入錯誤的發生情況較為特殊，由於中文的一字一音節特性，發生插入錯誤的可能性更低。會發生的情況通常是在學習者發出聲音時意識到自己唸得不夠標準，想要再次發出正確的音所導致，或是環境的噪音被當作語者所說的話。因此，對於插入錯誤的部分我們能夠忽略，僅專注於插入錯誤以外的替換錯誤與刪除錯誤檢測。

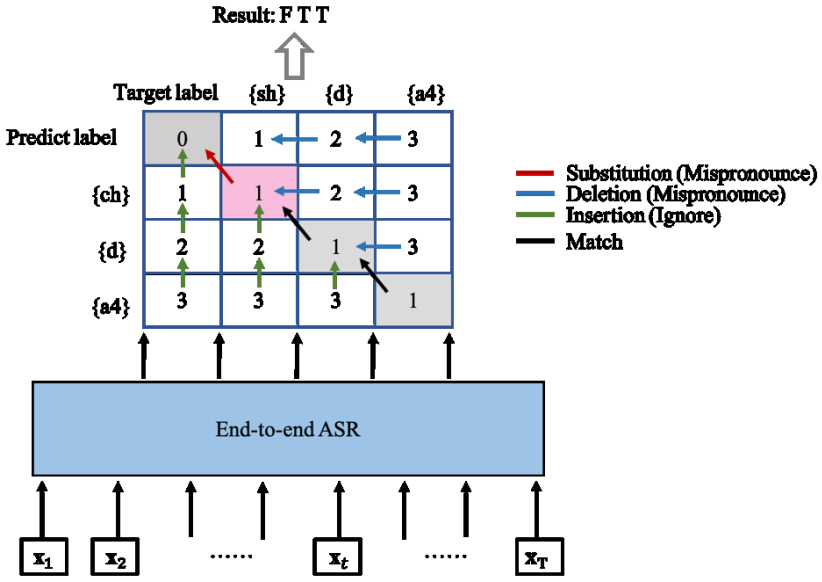


圖2、最短編輯距離發音檢測

四、實驗設定

5.1 語料

本論文使用臺灣師範大學邁向頂尖大學計畫之華語學習者口語語料庫，其中可以分為華語母語者(L1 speaker)以及華語非母語者(L2 speaker)兩部份。語料中的語句包含三種類型分別為單音節(Mono syllable, MS)、雙音節(Double syllable, DS)以及短文(Essay, ES)；詳細統計資訊如表 1 所示。

表1、華語學習者口語語料庫之訓練集、發展集與測試集

		時間(小時)	語者數	音素數量	錯誤發音音素數量
訓練集	L1	6.7	44	72,486	-
	L2	17.4	82	133,102	29,377
發展集	L1	1.4	10	14,186	-
	L2	-	-	-	-
測試集	L1	3.2	25	32,568	-
	L2	7.5	44	55,190	14,247

5.2 聲學模型

本研究分為兩階段，第一階段將比較以 L1 語料訓練傳統聲學模型進行發音檢測與端對端聲學模型進行發音檢測之效果，第二階段則是以端對端聲學模型加入 L2 語料進行發音檢測之效果。傳統聲學模型與端對端聲學模型皆使用美國約翰霍普金斯大學學者發展之大詞彙連續語音辨識工具，分別為“Kaldi”[22]以及“Espnet”[23]。

在傳統聲學模型設定中，初始階段訓練 GMM-HMM 模型輸入特徵為 MFCC 特徵，包含 3 維音調(Pitch)組成共 16 維，並對特徵取一階差量係數(Delta coefficient)，與二階差量係數(Acceleration coefficient)合併為 48 維特徵向量。DNN-HMM 聲學模型輸入特徵則每一音框為 40 維的 Filterbank 特徵加上 3 維音調(Pitch)並且取一階差量係數，與二階差量係數共 129 維。DNN-HMM 模型分別使用了不同層數與神經元數，也嘗試了最新的 DNN-HMM 架構 Factorized TDNN (TDNN-F)以及訓練準則 LF-MMI (Lattice-free maximum mutual information) [24]，也利用了速度擾動進行資料增添分別加快 1.1 倍速與放慢 0.9 倍速，詳細架構如下表 2 所示。

表1、傳統聲學模型架構

	類神經網路層數	每一神經元數
DNN-HMM	6	2048
TDNN-F LFMMI	13	768

端對端聲學模型於第一階段實驗使用的是 CTC-Attention 混合模型，架構主要參考[25][26]，混合參數為 0.5。Encoder 的架構為兩層的 VGG 層加上六層 Long short-term memory projection(LSTMP)一層含 320 個神經元，Decoder 的架構則為一層 LSTM 含 300 個神經元，使用的注意力機制為 Location attention[12]，計算方式如下式 11 所示，為了強化左到右的對齊，除了考慮前一個 Decoder state 以及當前 Encoder state 外，更加入一維摺積層 K 對於過去的 Attention 向量 $\mathbf{a}_{l-1}$抽取的向量。除此之外，也加入了標籤平滑方法(Label smoothing) 參數設為 0.05，目的在於不讓模型過度自信使部分較少出現的標籤也能有點機率分佈使模型更加一般化。由於端對端聲學模型通常需要較大量資料，我們同樣地也使用了速度擾動方法，因此與 TDNN-F LFMMI 之結果較具可比性，詳細架構如下圖 3 所示。

$$e_{lt} = \begin{cases} \mathbf{F_l} = \mathbf{K} * \mathbf{a_{l-1}} \\ \mathbf{g}^{\mathrm{T}} \tanh\left(\mathrm{W_q}\mathbf{q_{l-1}} + \mathrm{W_h}\mathbf{h_t} + \mathrm{W_f}\mathbf{f_{lt}}\right) \end{cases} \tag{11}$$

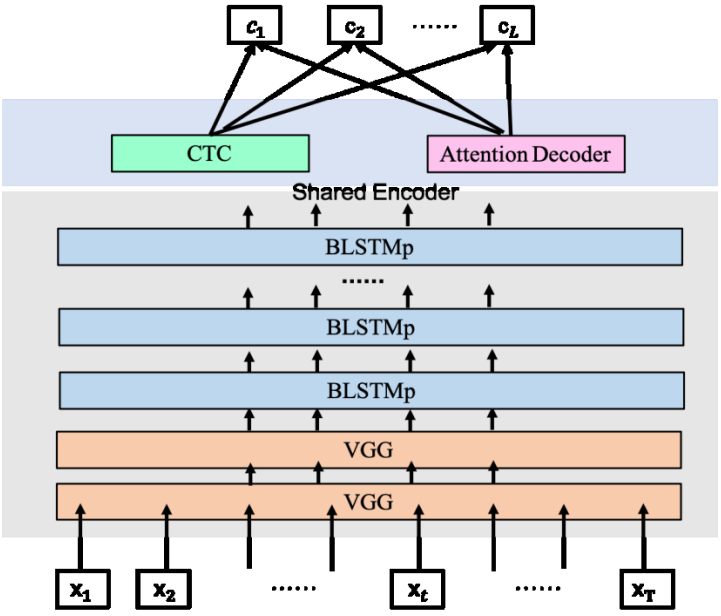


圖3、端對端語音辨識模型架構

第二階段主要探討端對端聲學模型加入 L2 語料對於發音檢測之影響，將基於圖 3 之架構，分別比較只使用 CTC、Attention 模型與 CTC-Attention 混合模型。並且比較加入已知 L2 錯誤模式進行解碼。

五、實驗結果與分析

5.1 L1 辨識結果

為了證實端對端語音辨識於發音檢測的可行性，我們首先不同聲學模型架構於 L1 語料中的表現，下表 3 為不同聲學模型架構於同一語料測試集的音素錯誤率與音節錯誤率，而使用資料量相同的為 TDNN-F LFMMI 與 CTC-Attention 模型。由結果可以得知使用 CTC-Attention 的辨識效果優於任意其他模型，可能原因是端對端聲學模型並不受制於發音詞典，因此相較於傳統 DNN-HMM 模型不會受到未知音素組合的影響，並且 Attention 模型架構設計帶有語言模型的概念，可提升對於已知音素組合的辨識效果。

表3、L1測試集的音素錯誤率與音節錯誤率

	Mono syllable(MS)		Double syllable(DS)	
Model	SER	PER	SER	PER
DNN-HMM	41.8	28.4	28.7	18.0
TDNN-F LFMMI	34.2	26.7	25.3	22.5
CTC-Attention	**32.2**	**18.9**	**6.8**	**5.1**

5.2 基於分數之發音檢測結果

基於門檻值方法，首先我們利用 ROC 曲線觀察門檻值對於發音檢測的評估標準變化，如下圖 4 所示，我們從中發現當門檻值設越小錯誤拒絕率越低，而錯誤接受率會隨之上升，但是增加幅度較小，進而發現當當全域門檻值設為 0.1 時兩者之錯誤接受率與錯誤拒絕率相等。儘管使用門檻值方法在錯誤檢測上與使用 GOP 方法可比較。總體來看分類效果仍然是劣於 GOP 方法。錯誤的拒絕過多導致在判斷錯誤時的精準度偏低。

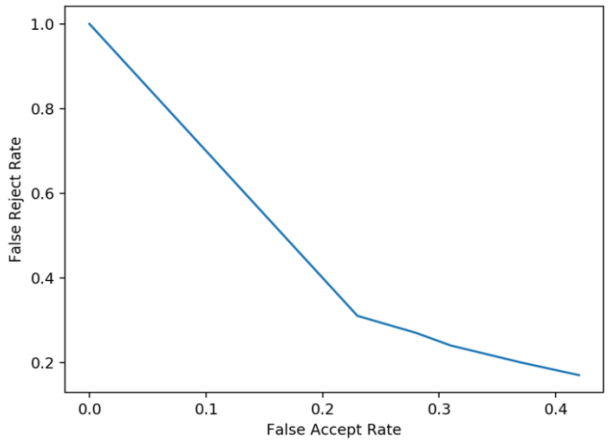


圖4、基於門檻值之ROC曲線圖

表4、第二階段分類效果

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
DNN-HMM (GOP)	0.88	0.85	0.86	0.55	0.61	0.58
End-to-end (Threshold)	0.76	**0.87**	0.81	**0.69**	0.51	**0.59**

5.3 基於辨識之發音檢測與診斷結果

由 5.1 節的實驗結果顯示，CTC-Attention 模型在辨識率上超越其他傳統聲學模型。我們假設此聲學模型所辨識的結果為正確答案，並套用最短編輯距離去對齊目標音素，得到的結果如下表 5 所示：

表5、L1聲學模型發音檢測結果

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
1-best	0.73	**0.87**	0.79	**0.71**	0.49	**0.58**
2-best	0.80	0.84	0.82	0.59	0.52	0.55
3-best	0.84	0.82	0.83	0.52	0.54	0.53
5-best	0.87	0.81	0.84	0.43	0.55	0.49

由表中得知，僅使用 L1 訓練聲學模型判斷 L2 語者發音好壞過於嚴格，也可能因為口音的不同造成模型的誤判。鑑於判斷過於嚴格，我們也逐漸放寬標準，改採以 N-best 的結

果做判斷，然而隨著標準放寬，錯誤接受也隨之上升使得結果缺乏鑑別力。為了使模型能夠判斷發音好壞，接下來的實驗我們將 L2 語料加入訓練加入聲學模型訓練，希望能夠使聲學模型直接學習到 L2 的錯誤模式。

加入含有錯誤模式的L2訓練集後，發音檢測的任務可以直接視為語音辨識的任務，即當對測試集解碼的錯誤率較低，後續的發音檢測與診斷也將有很好的效果。因此我們首先比較音素錯誤率如表 6 所示：

表6、端對端聲學模型音素錯誤率

Model	Weight	Phone error rate
CTC	-	28.3
Attention	-	**24.5**
CTC-ATT	0.2	24.8
CTC-ATT	0.5	**24.6**
CTC-ATT	0.8	24.9

由此實驗得知使用 CTC-Attention 與僅使用 Attention 模型差異不大，然而音素錯誤率的計算有考量到插入錯誤，我們在進行最小編輯距離對齊目標音素時則不考慮插入錯誤因此兩者的效果有待進一步評估，另外也比較[19]基於 CTC 的結果。

發音檢測的效果如下表 7 所示，整體來看使用 CTC-Attention 模型的效果與僅使用 Attention 效果差異不大，但是都比只使用 CTC 做發音檢測效果更好。

表7、端對端聲學模型發音檢測效果

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
CTC	0.831	0.893	0.861	0.706	0.656	0.680
CTC-Att	0.873	0.893	0.883	**0.714**	0.672	**0.693**
Attention	0.875	0.892	0.884	0.710	0.674	0.691

儘管發音檢測效果已經得到良好的結果，對於學習者來說仍然無法得知自己的發音發錯成什麼了，因此繼續探討端對端聲學模型的診斷效果。如下表 8 所示：

表8、發音診斷效果

	Initial	Final	Tone
DNN-HMM	0. 548	0.441	0.752
CTC	0.611	0.582	0.768
CTC-Attention	**0.661**	**0.612**	**0.801**
Attention	0.645	0.609	0.797

由上表診斷結果顯示，儘管在發音檢測 CTC-Attention 與 Attention 的效果差異不大，但是 CTC 與 Attention 的聯合解碼幫助了診斷結果，使診斷更加準確。

六、結論

本論文在端對端語音辨識架構上提出兩種發音檢測方式，分別為使用信心分數以及語音辨識結果，並且比較傳統使用語音辨識器進行發音檢測的方法。在使用信心分數的結果中，仍然遜於 GOP 的方法，希望在未來能夠加入更多發音的特徵，如[27]為了改善單用發音事後機率容易有誤判的情況，額外加入了許多發音特徵，例如發音方式、發音類型、吸氣吐氣等特徵。另一方面，基於語音辨識結果的發音檢測，我們發現加入 L2 語料訓練對於整體的檢測效果影響很大。當僅有母語者語料時，訓練的模型對於 L2 語者來說過於嚴格。而加入 L2 語料不僅能夠使模型進行發音檢測也能夠診斷，並且診斷正確率也超越以往方法。將發音檢測視為語音辨識的問題將使得研究更加簡單，僅需要思考如何讓模型辨識率提升。而未來方向除了改進聲學模型外，也希望能夠處理未知的錯誤。例如在 L2 測試集中有許多不存在於訓練集的錯誤標記，往往是兩個音素或是聲調的組合，而我們的模型診斷結果由於缺少這樣的標記通常只能回饋兩個音素中的其中一種，對於此情況仍然難以解決。期許在未來能夠找到對於未知錯誤的正確回饋方法。另外端對端聲學模型的解碼速度不夠即時，對於實際應用來說還有一段距離，在未來也希望能夠做到即時解碼，如[28][29]，使得我們的架構能被實際應用。

致謝

本論文之研究承蒙行政院科技部研究計畫 (MOST 105-2221-E-003-018-MY3 和 MOST 107-2221-E-003-013-MY2、MOST 108-2221-E-003-005-MY3 和 MOST 108-2634-F-008 -004 -) 之經費支持，謹此致謝。

參考文獻

[1] Lawrence R. Rabiner et al., "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*," Proceedings of the IEEE, 1989.

[2] Mark Gales and Steve Yang, "*The Application of Hidden Markov Models in Speech Recognition*," Foundations and Trends® in Signal Processing, 2008.

[3] Geoffrey Hinton et al., "*Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*," IEEE Signal processing magazine, 2012.

[4] Ossama Abdel-Hamid et al., "*Convolutional neural networks for speech recognition*," IEEE/ACM Transactions on audio, speech, and language processing, 2014.

[5] Alex Graves et al., "*Speech recognition with deep recurrent neural networks*," ICASSP, 2013.

[6] Haşim Sak et al., "*Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition*," arXiv, 2014.

[7] Vijayaditya Peddinti et al., "*A time delay neural network architecture for efficient modeling of long temporal contexts*," Interspeech,2015.

[8] Alex Graves et al., "*Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*," ICML, 2006.

[9] Dzmitry Bahdanau et al., "Neural machine translation by jointly learning to align and translate," ICLR, 2015.

[10] Kelvin Xu et al., "*Show, attend and tell: Neural image caption generation with visual attention*," ICML, 2015.

[11] Jan Chorowski et al., "*Attention-Based Models for Speech Recognition*," NIPS, 2015.

[12] Suyoun Kim et al., "*Joint CTC-Attention based end-to-end speech recognition using multi-task learning*," ICASSP, 2017.

[13] Shinji Watanabe et al., "*Hybrid CTC/attention architecture for end-to-end speech recognition*," IEEE Journal of Selected Topics in Signal Processing 11, 2017.

[14] Yoon Kim et al., "*Automatic pronunciation scoring of specific phone segments for language instruction*," in Proc. Eurospeech-1997. ISCA, pp. 645–648, 1997.

[15] Silke Witt and Steve Young, "*Language Learning Based On Non-Native Speech Recognition*," European Conference on Speech Communication and Technology, 1997.

[16] Silke Witt and Steve Young, "*Phone-level pronunciation scoring and assessment for interactive language learning*," Speech Communiciation, Vol. 30, No. 2-3, pp. 95–108, 2000.

[17] Kun Li et al., "*Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks*," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016.

[18] Shaoguang Mao et al., "*Applying Multitask Learning to Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech*," ICASSP, 2018.

[19] Wai-Kim Leung et al., "*CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis*," ICASSP, 2019.

[20] Alex Graves et al., "*Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*," ICML, 2006.

[21] Robert A. Wagner and Michael J. Fischer. "*The string-to-string correction problem*," Journal of the ACM (JACM) , Vol. 21.1, pp. 168-173, 1974.

[22] Daniel Povey et.al, "*The Kaldi Speech Recognition Toolkit*," ASRU, 2011.

[23] Shinji Watanabe et al., "*ESPnet: End-to-End Speech Processing Toolkit*," Interspeech, 2018.

[24] Povey, Daniel et.al, "*Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks*," Interspeech, 2018.

[25] Takaaki Hori et al., "*Advances in Joint CTC-Attention based End-to-End speech recognition with a Deep CNN Encoder and RNN-LM*," Interspeeh, 2017.

[26] Haşim Sak et al., "*Long short-term memory recurrent neural network architectures for large scale acoustic modeling*," Interspeech, 2014.

[27] Wei Li et al., "*Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models*," Interspeech, 2017

[28] Chung-Cheng Chiu and Colin Raffel. "*Monotonic chunkwise attention*," ICLR, 2018.

[29] Ruchao Fan et al., "*An Online Attention-based Model for Speech Recognition*," arXiv, 2018.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 281-295

基於語境特徵及分群模型之中文多義詞消歧

Using Contextual Information in Clustering Methods for Chinese Word Disambiguation

李右元 周子皓 劉昭麟

Yu-Yuan Lee, Tzu-Hao Chou, Chao-Lin Liu

國立政治大學資訊科學系

Department of Computer Science

National Chengchi University

{107753027, 104753029, chaolin}@nccu.edu.tw

摘要

多義詞是語言中極為常見的現象，在過去，若要查找多義詞的義項及其使用方式，必須翻查傳統辭典，但礙於篇幅問題並非所有義項都會收錄，因此所提供的例句數目也較少。即使隨著科技進步，發展了數位化的辭典與檢索系統，仍有部分問題存在。因此，人文學者必須耗費大量心力以人工判讀方式辨別義項的不同。

本研究以分群模型將大量已向量化之中文語料加以處理，透過 purity 分數比較出最適之模型，並挑選適量的例句供使用者參考。實驗中以人工標記之例句作為評分依據，結果顯示屬於同形異義(homonymy)之多義詞在 macro-average、weighted-average 與 accuracy 皆能達到 0.85 以上之水準。

Abstract

We present preliminary results for searching for useful sentences for learning ambiguous words with clustering methods. First, we search for sentences that contain an ambiguous word (the target word, henceforth). To make the extracted sentences useful for learning the target word, we attempt to guide the clustering methods to separate the sentences that carry different senses of the target word into different clusters. We influence the functioning of a clustering method

by providing example sentences that carry specific senses of the target word. In the terminology of machine learning technology, we label a sentence with the sense of the target word in the sentence. Two sample labeled sentences for the ambiguous word “bank” follow.

1. “financial institution”: Mr. Black deposit the money in the Citi bank.
2. “place”: Along the bank of the Charles river, you may see the MIT campus.

Assume that we can collect a large number of sentences that contain the target word, for which we need sentences that use a specific sense of the target word. Assume that we are willing to label a few of these original sentences as we described above.

A clustering algorithm may employ the labeled sentences to build clusters of sentences for our needs. The algorithm may take advantage of the labeled sentences as informative seeds for initializing the clusters. In addition, when selecting the (unlabeled) sentences from the clustered sentences as the final output, the labeled sentences may also provide guidance for selecting the sentences of “correct” senses. If a cluster has many labeled sentences of a specific sense, the (unlabeled) sentences in this cluster might have the same label of the sample sentences. Furthermore, to select and output the (unlabeled) sentences in this cluster, we may consider the (unlabeled) sentences that are closer to the sample sentences.

Assume that we may find thousands of sentences that use a target word, assume that we provide a certain number of labeled sentences to guide a clustering algorithm, assume that we cluster the thousands of sentences into tens of clusters, and assume that we select just tens of sentences from these tens of clusters. If our clustering methods are good and if we select sentences from a cluster conservatively, we may achieve high precision in the final selection of the unlabeled sentences for the target word.

Empirical evaluations reported in this paper show promising results. Not surprisingly, we found that it was relatively easier to achieve better results for homonym than for polysemy. We hope our methods can be useful for building corpora for learning ambiguous words.

關鍵詞：多義詞，一詞多義，同形異義，分群模型，詞向量，句向量

Keywords: lexical ambiguity, polysemy, homonymy, clustering, word vector, sentence vector

一、 緒論

多義詞存在於大多數的語言當中，Bruce Britton[1]曾保守估計至少 32%的英文單詞存在著不只一個義項，且最常使用的前一百名英文單詞中，高達 93%的單詞是多義詞，可見多義詞與人類的生活是息息相關。此外，多義詞還能細分成一詞多義(polysemy)與同形異義(homonymy)[2]，前者定義為其義項類別彼此有關聯，後者則是彼此無關聯。以 “cup” 為例，茶杯的 “cup” 與獎盃的 “cup” 是一詞多義，而銀行的 “bank” 與河岸的 “bank” 則是同形異義。

以往若要瞭解詞彙的其他義項，必須透過辭典進行查找，但是辭典中提及的義項大多是較有規範化的使用方式，其內容較少包含口語化的義項類別，且礙於篇幅關係，並非所有義項都能有足夠的例句讓使用者瞭解該義項的使用方式。此外，編輯辭典耗時費力，實在難以符合語言發展的速度。

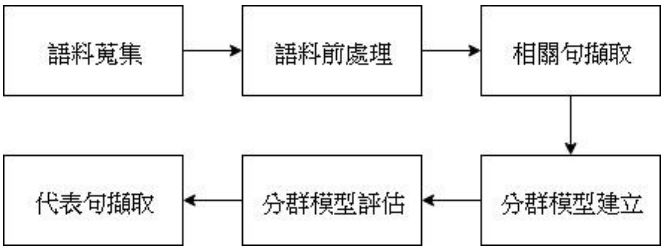
隨著網路與資料庫的發展，有許多的數位檢索系統扮演著新型態的辭典供使用者查詢。以「教育部重編國語辭典修訂本」[3]及「萌典」[4]為例，藉由線上檢索的功能或許能改善部分傳統辭典的問題，但仍觀察到某些較新興的詞彙，如「亞馬遜」不存在於資料庫內，亦或者例句數量較少，難以讓使用者清楚明白適用該義項之句型。再以「中央研究院現代漢語標記語料庫」[5]為例，其搜尋結果沒有顯示出義項類別，因此仍必須透過人工方式逐一閱讀。

過去曾有人文學者對於特定詞彙以人工方式，透過其在特定領域的專業知識，判別出過往辭典中未收錄的義項類別，例如吳美嫻[6]、林香薇[7]、蔡宛玲[8]與許尤芬[9]等人的相關研究。在資訊領域，Navigli[10]曾提出對於消除歧義有監督式、非監督式及以知識為本等三種實驗方式，而 Lesk[11]為首位提出以知識為本的消歧方式，透過計算目標詞彙所在上下文與目標詞彙在知識庫中各義項類別的覆蓋程度，對目標詞彙進行消歧。

因此本研究嘗試透過文字向量化的方式，將大量的中文語料與少數使用者提供之參考句轉換成向量，再以分群模型演算法將相同義項類別之相關句進行區隔，同時搭配 purity[12]分數尋找出最適合該多義詞之分群模型，最後以實驗中設計的方法擷取出一定數量的代表句。透過分群模型，提供例句供使用者閱讀。

二、研究方法

在本研究中，主要分為六個階段，分別如圖一所示。



圖一、研究方法流程圖

（一）、語料搜集

本研究所使用的語料主要是中文維基百科所提供的開放資料[13]、教育部 AI CUP 2019 新聞立場檢索技術第一階段所提供的新聞資料[14]，以及由使用者自行提供的參考句。

維基百科是個開放式的協同寫作平台，其內容是由眾多使用者與專業人士一同撰寫而成，因此不會受限於個人或團體的寫作風格與觀點，其收錄內容不僅包含傳統百科所蒐集的資訊，由於線上協同寫作的關係，其內容更容易與時俱進收錄到較新的詞條訊息。

然而，維基百科雖然是開放式協同合作的資料，但其終究是屬於百科類文體，受限於此，可能導致我們所在意的目標詞彙的義項比例分布不均勻。因此，在實驗中我們加入了約十萬篇的新聞語料藉以平衡這樣的問題。新聞語料內容來自中國時報、自由時報、蘋果日報、聯合報與 TVBS 等媒體，內容橫跨各種議題，包含政治、經濟、教育、…等等，用以增加語料的多元性。

此外，研究中亦會請使用者提供具有人工標記義項類別之句子，不僅為語料增加更多的文字風格外，也能作為後續實驗方法的依據之一。

（二）、語料前處理

不論是維基百科或者是新聞語料，都夾雜了或多或少的額外訊息，因此在進行實驗前必須費工處理以得到乾淨的文本。另外，由於中文的書寫結構與英文不同，因此通常需要額外的方式來加以處理。

以維基百科為例，我們首先使用 WikiExtractor[15]從原本取得之 XML 格式的資料內容中過濾掉大量不需要的資訊，僅留下條目標題、條目超連結、條目編號及條目正文等資訊，接著透過 OpenCC[16]將原本繁簡混雜的內容統一轉換成臺灣繁體，舉例來說，OpenCC 會將原始簡體內容的「光盘」轉換成臺灣使用的「光碟」。最後再去除剩餘不需

要的標籤與資訊，最終得到我們所需要的內容。

將所有語料都處理乾淨後，我們首先必須做斷句的工作，研究中是以逗號、句號、驚嘆號與分號這四種標點符號做為句子之間的斷點，我們將斷開後的句子稱為例句，表一說明實驗中使用的維基百科與新聞語料的數據統計。

表一、語料數據統計

項目	中文維基百科	新聞語料
條目總數	1,092,447 篇	98,250 篇
例句總數	26,815,431 句	5,489,253 句
例句平均長度	15 字	15 字
條目平均例句數	25 句	56 句

在斷句之後必須處理中文的斷詞問題，因為中文不像英文是以空白做為詞與詞之間的分隔，因此必須依賴斷詞器達到詞彙分割的目的。實驗中我們比對了中研院 CKIP 中文斷詞系統[17]與 Jieba 斷詞器繁體版[18]兩者間的差別，從維基百科中隨機抽樣一百句例句並透過人工斷詞作為標準答案，透過萊文斯坦距離(Levenshtein distance)計算兩者與標準答案間的相似度，最終 Jieba 得到較高的相似度，因此我們採用 Jieba 斷詞器繁體版來處理這樣的問題，藉由 Jieba 能夠有效率地將個別詞彙分開，以利後續工作進行。

（三）、相關句擷取

在本研究中，目標詞彙指的是具有多重義項類別的詞彙，例如「亞馬遜」可以代表亞馬遜雨林，也可以代表電商亞馬遜公司。而相關句是指包含目標詞彙的例句。此外，在現實情況下，一句例句所提供的語境可能無法準確地判斷出目標詞彙在此例句中所代表的義項，因此我們將相關句的左右例句也加入相關句中，亦即一筆相關句中包含三句例句，希望藉由增加相關句的長度藉以提供更豐富的語境資訊。然而因為原始語料標點符號的錯用，可能導致即使已經增加相關句長度，其相關句仍然過短無法提供足夠的語境資訊的問題，因此我們根據相關句中例句的長度來進行篩選，將語境相對貧乏的相關句剔除。

（四）、分群模型建立

在建立分群模型之前，必須將處理好的語料轉換成向量的形式，實驗中使用的向量化模型如表二所示，其中包含四種模型，各模型皆有七種向量維度及七種 window size 作為參數調整。

表二、向量化模型表

Embeddings	Models	Vector size	Window size
doc2vec[19]	PV-DBOW	5, 10, 20, 50, 100, 200, 500	2, 5, 10, 15, 20, 30, 50
	PV-DM		
average word2vec[20]	CBOW		
	skip-gram		

而分群模型則使用了 K-means[21]、hierarchical clustering[22]、spectral clustering[23][24][25][26]與 BIRCH clustering[27]等方法，而分群數量則有[2, 10]等 9 種組合。

其中，起始點的選擇方式對於 K-means 的分群結果會有重要的影響，若起始點的選擇與語料數據的分布情形差異過大，會造成效果較差的分群結果。在實驗中，我們採用了三種不同的起始點選擇方式，以下簡略說明：

方法 1： 以 K-means++[28]作為起始點選擇。有別於原始的 K-means 演算法，起始點是隨機選取的，K-means++會選擇離當前起始點最遠的點作為新群集的起始點。

方法 2： 先將參考句分群，再以分群後的群集中心作為相關句分群的起始點。主要是將由維基百科與新聞語料組成的相關句與使用者提供的參考句分別進行分群工作，因為參考句是由使用者提供，能夠掌控目標詞彙不同義項的例句數量，降低因為例句比例差異造成分群錯誤的問題。

方法 3： 依據使用者提供之參考句的義項類別將參考句分群，再以群集中心做為起點。不同於方法 2 是將參考句全部一起進行分群，而是根據人工標記之義項類別各自對參考句進行分群，再將各自分群的群集中心合併，作為相關句的分群起始點。此外，由於是根據義項類別做分群，實驗中的分群數目設置為義項類別的倍數。以目標詞彙「亞馬遜」為例，其包含雨林與電商兩種義項，則分群數目設值為 2 的倍數，即 2、4、6、…等。

此外，hierarchical clustering 的距離計算分別採用了 ward's linkage、complete linkage、average linkage 與 single linkage 等四種方式。

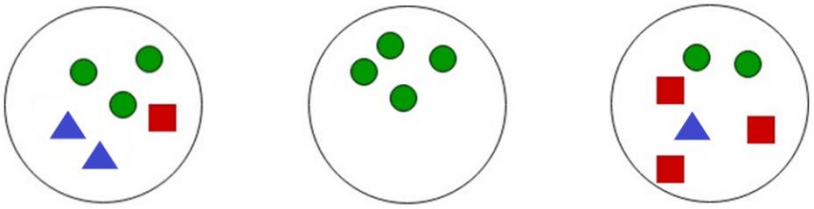
（五）、分群模型評估

如同第四小節所述，本實驗主要採用分群模型判別相關句目標詞彙的義項，並採用了多種分群模型與不同的參數設置模型，目的是要找到一個相對好的模型。實驗中我們採用 purity 作為評估指標，主要是以使用者提供的參考句計算 purity。因為參考句包含人工標記的義項類別，能夠讓我們判斷該群集是以目標詞彙的何種義項為大宗。

若一個群集內的資料其實際類別彼此相同，則 purity 分數高，反之則低，而分數區間是介於 0 到 1 之間。purity 公式如(1)所示，其中 N 代表資料點總數，$\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ 代表 k 個群集，$C = \{c_1, c_2, \dots, c_j\}$ 代表 j 個類別。

$$purity(\Omega, C) = \frac{1}{N}\sum_{k} \max_{j} \left|\omega_k \cap c_j\right| \quad (1)$$

以圖二作為舉例，假設分群模型將實驗語料分為三個群集，圓形資料點為維基百科與新聞語料所組成的目標詞彙相關句，正方形與三角形資料點為不同義項類別的使用者提供參考句。圖二-(a)表示該群集內存在三筆相關句、兩筆屬於三角形義項之參考句及一筆屬於正方形義項之參考句。根據 purity 的定義會加總群集中相對多數的資料點，又因為我們是針對參考句去計算 purity，圓形所代表的相關句不會影響分數的計算，也就是圖二-(b)對於整體 purity 分數不會有影響，僅計算圖二-(a)與圖二-(c)中所佔比例最多之參考句數目。因此範例的 purity 分數為 $\frac{1}{7}(2+3) \approx 0.714$。



(a) (b) (c)

圖二、purity 於本研究中的使用範例

（六）、代表句擷取

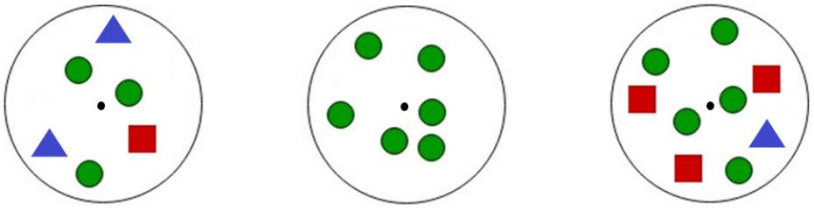
透過分群模型的建立，我們能夠將向量化的相關句分成若干群集，各群集中目標詞彙所代表的義項也不完全相同。為了達到實驗目的，讓使用者瞭解目標詞彙在不同義項上實際使用情形，我們在此定義了代表句為能用以表示目標詞彙使用情形之相關句。並設計了三種代表句擷取方式，再以人工標記之正確答案來評估代表句擷取效果。以下簡略敘

述，其中圓形資料點皆代表相關句，正方形與三角形資料點則為不同義項類別之參考句。

方法 1： 擷取包含參考句在內之群集中的所有相關句做為代表句。以圖三做為舉例，圖三-(a)群集會擷取三筆相關句做為代表句；圖三-(b)群集因為不包含參考句，不做擷取；圖三-(c)群集則擷取五筆相關句做為代表句。

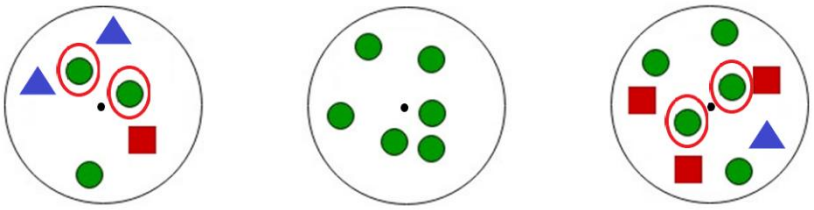
方法 2： 依據相關句與群集中心距離，擷取包含參考句在內之群集中的相關句。以圖四為例，黑點代表群集中心，距離計算採取 cosine similarity 方式，擷取與群集中心距離最近之前兩句相關句做為代表句，其餘距離較遠之相關句則不採納。

方法 3： 依據相關句與參考句之群集中心距離，擷取包含參考句在內之群集中的相關句。以圖五為例，黃色星形代表各群集中參考句之群集中心，同樣採用 cosine similarity 方式計算距離，並擷取與該中心距離最近之前兩句相關句做為代表句。



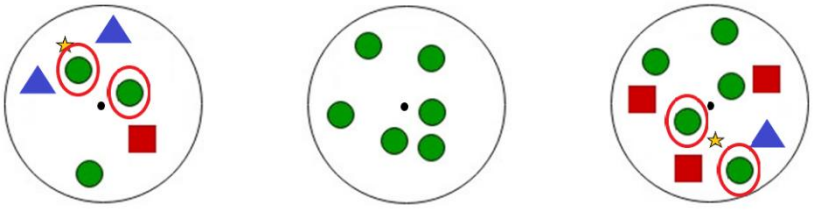
(a) (b) (c)

圖三、代表句擷取方法 1 示意圖



(a) (b) (c)

圖四、代表句擷取方法 2 示意圖



(a) (b) (c)

圖五、代表句擷取方法 3 示意圖

此外，因為向量化具隨機性，為了觀察在不同參數下分群的效果，考量機器效能會以相同參數重複執行十次，透過 hierarchical clustering 將這十次產生的群集視為集合，再度分群。而 hierarchical clustering 的距離計算改為使用 Jaccard Index，而非歐氏距離，其分數越高代表兩群集間相似度越高。最後透過人工標記之答案計算 macro-average[29]、weighted-average 與 accuracy 作為代表句擷取效果評估。

在執行 hierarchical clustering 後可能會有某些相關句同時出現於不同群集中，本研究中設計兩種方式來解決，以下簡略說明。

方法 1： 直接剔除重複出現的相關句。若相關句於不同群集中出現，可能是位於群集邊際所造成，可能與群集內其他相關句的同質性較低，因此參考價值較低，可以直剔除。

方法 2： 計算重複出現之相關句與其群集中心之距離，並將其歸屬於距離較近之群集。為避免直接剔除該相關句可能損失特定資訊，透過距離的比較保留相關句內容。

以下說明代表句合併過程，若資料集內有 {1, 2, 3, 4, 5, 6} 等六筆相關句，將資料集劃分為兩群，重複執行兩次，第一次分群結果為 (1, 2, 3)、(4, 5, 6)，第二次分群結果為 (1, 2, 3, 4)、(5, 6)，再依上述 hierarchical clustering 分成兩群，分別為 [(1, 2, 3), (1, 2, 3, 4)]、[(4, 5, 6), (5, 6)]。此時相關句 4 同時於兩個群集中出現，再依上述方法處理。

三、研究結果分析

（一）、目標詞彙篩選

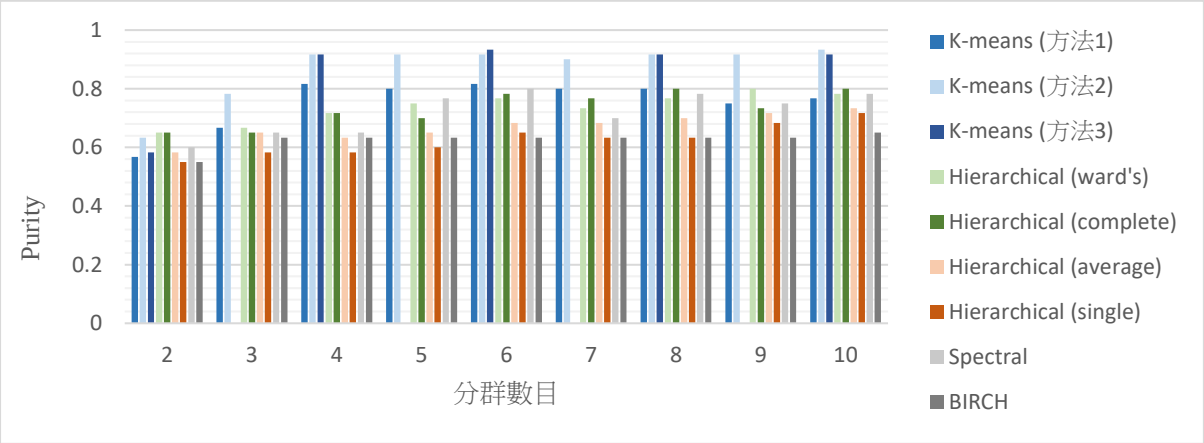
在實驗中，目標詞彙的相關句數量必須足夠，否則機器無法充分學習其語意特徵。此外，因為多義詞有同形異義(homonymy) 與一詞多義(polysemy)的不同，本研究中，同形異義的目標詞彙有「亞馬遜」、「蘋果」、「小米」；一詞多義的目標詞彙則有「出發」、「出入」、「壓力」與「東西」等。表三簡單列出其例句。

表三、目標詞彙義項類別與例句

目標詞彙	義項類別	例句
亞馬遜 (homonymy)	雨林	巴西國家太空研究所氣候學家諾布雷（Carlos Nobre）指出，亞馬遜樹種的死亡和巴西南部
	電商	2015 年黯然退出手機市場。但近日亞馬遜一名高層人員透露，亞馬遜不排除會重返智慧型手機市場
出發 (polysemy)	實際離開	鐵馬進香活動昨清晨 6 點熱鬧展開，300 輛自行車從天后宮出發往北港，挑戰來回 345 公里進香之旅
	從某方面著手	旁觀者到思考者，轉換角色，從整理自己開始，陪考也有重新出發的可能

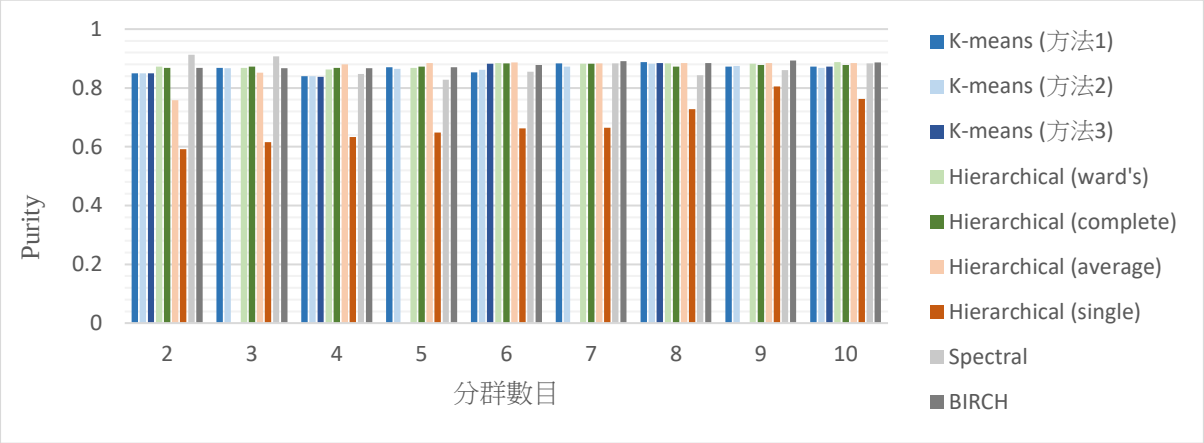
（二）、分群模型評估

在九種分群模型下，分群的數量會對於分群結果造成影響。此外，因為向量化具有隨機性，為了評估分群的效果，實驗中以相同參數重複執行十次並取平均作為該參數下分群的 purity 分數。圖六以「亞馬遜」為例，呈現九種分群模型在分群數目區間為[2, 10]之間，各分群模型效果最優之對應 purity 值。從圖六中可以觀察到，在 K-means 的三種起始點選擇方式下，比起方法 1 單純使用 K-means++的方式，方法 2 與方法 3 能夠獲得較高的 purity 分數。而在 hierarchical clustering 的四種距離計算方法中，ward's linkage 與 complete linkage 都能較另外兩種得到更高的 purity 分數。以實驗結果而言，仍然是 K-means 的方法 2 與方法 3 能夠有效地將相關句進行正確的分群。



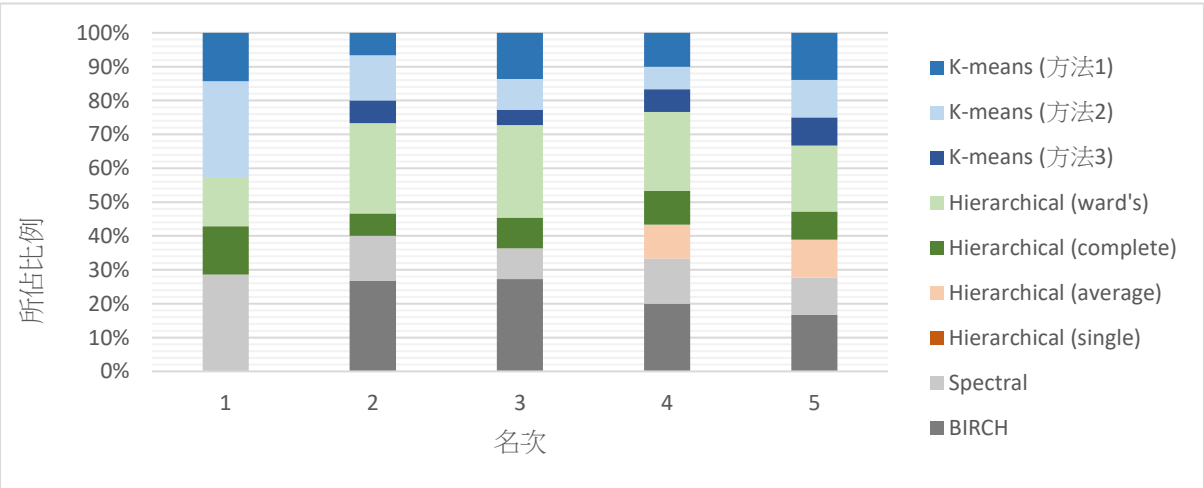
圖六、各分群模型下最優之 purity 分數（亞馬遜）

圖七則是以目標詞彙「出發」為例，以相同的方式重複執行十次，取平均作為 purity 分數。從圖中可以觀察到各分群模型在不同的分群數目下，除了以 single 距離計算方式形成的 hierarchical clustering 外，其餘模型並無明顯變化。根據圖二，可以推測隨著分群數目增加，有許多的群集中可能不包含參考句在內，導致其對 purity 的影響極小。



圖七、各分群模型下最優之 purity 分數（出發）

圖八是依據各目標詞彙在各種分群模型下之最優 purity 分數統計所繪出之圖形，可以觀察到使用 K-means(方法 2) 及 spectral 等分群模型在各目標詞彙最優之 purity 分數出現較多次。而若以各目標詞彙的前 5 名 purity 分數來觀察，可以發現使用 hierarchical (ward's linkage) 方法與 BIRCH 等分群模型所佔比例較高。因此若使用者要以實驗以外之多義詞進行查詢，會以此二種分群模型作為優先推薦。



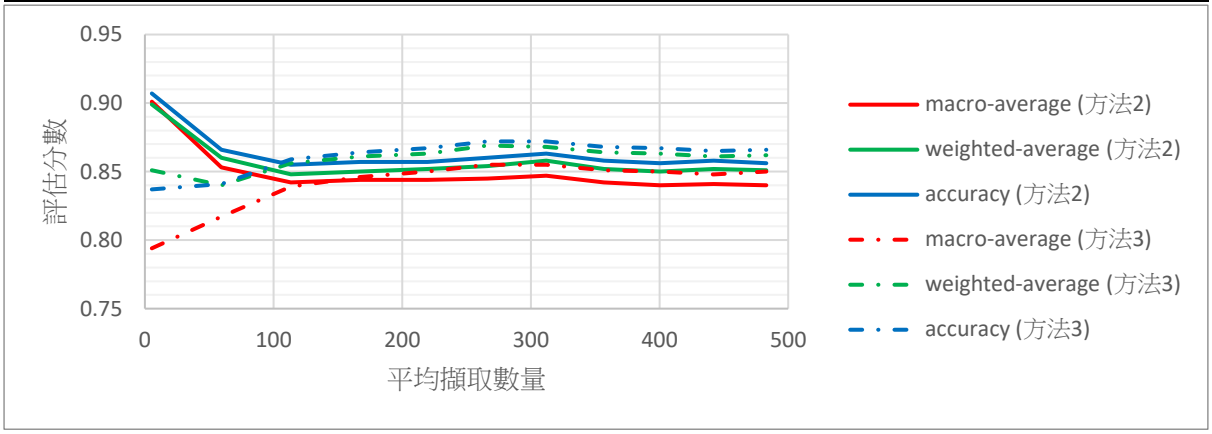
圖八、各分群模型在不同 purity 分數名次所佔之數量比例

（三）、代表句擷取評估

在代表句的擷取，我們採用了三種擷取方式，並透過人工標記的正確答案計算代表句之 macro-average、weighted-average 與 accuracy，作為代表句擷取效果的好壞評估。以下將以「亞馬遜」作為代表，分析在最佳 purity 之參數下，三種擷取方式所得之分數。表四是以擷取代表句方法 1 之參數為實驗，可見三種分數介於[0.840, 0.860]之間。圖九則是以方法 2 與方法 3 之參數為實驗，其中實線為方法 2 之結果，虛線為方法 3 之結果。藉由調整代表句擷取數量，每次重複執行十次取平均，在平均擷取數量下比較其代表句與人工標記答案之分數結果。從圖中可以觀察到方法 2 之分數隨著平均代表句擷取數量增加而降低，但逐漸趨於收斂。方法 3 則在平均擷取數量較低時，與方法 2 些許的不同，隨著擷取數量上升，卻同樣收斂於接近方法 1 的分數。由此可知方法 3 以義項類別之群集中心作為整體分群起始點時，其中心點並無法有效的代表各義項之語境，因為與分群中心較近之代表句所獲得的分數反而較低。方法 2 則符合離中心點越近之代表句，較能表達該群集義項的。

表四、「亞馬遜」代表句擷取分數評估（方法 1）

Macro-average	Macro-average 標準差	Weighted-average	Weighted-average 標準差
0.840	0.004	0.851	0.036
Accuracy	Accuracy 標準差	平均代表句數	平均代表句數標準差
0.856	0.031	483 句	35.364



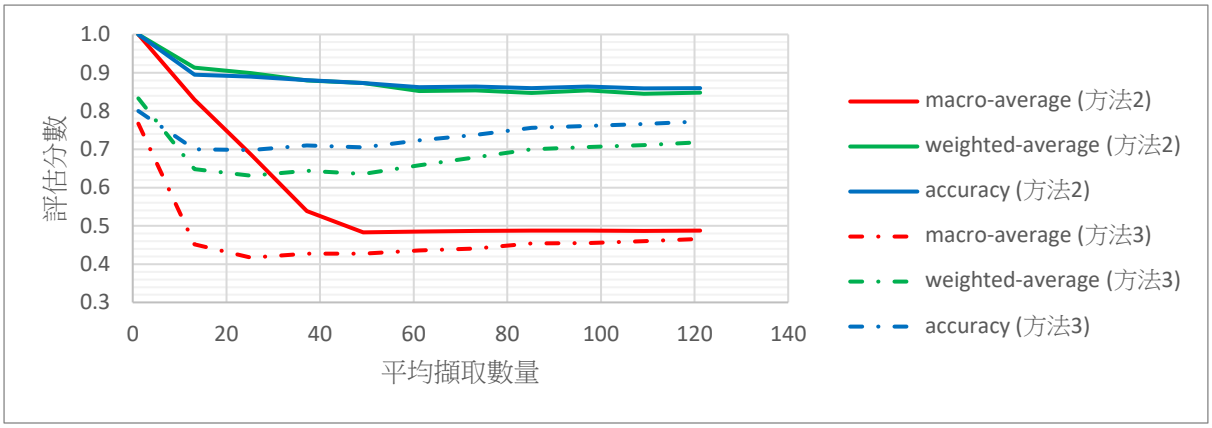
圖九、「亞馬遜」代表句擷取分數評估

接著是以「出發」作為代表，在最佳 purity 參數下的擷取結果，以相同方式進行比較。表五是以代表句擷取方法 1 的實驗結果。圖十則是以實線與虛線分別作為方法 2 及方法 3 之實驗結果。從中可以觀察到方法 2 與方法 3 在代表句擷取數量較低時，其分數皆高於擷取數量較高之分數，亦即距離群集中心越近之代表句越能代表該群集義項之語

境。但是兩者並無在擷取大量代表句時逐漸收斂至接近的分數，且 macro-average 與 weighted-average 相差約 0.236，推測是因為語料數據分布情形差異所造成之影響。

表五、「出發」代表句擷取分數評估（方法 1）

Macro-average	Macro-average 標準差	Weighted-average	Weighted-average 標準差
0.507	0.118	0.743	0.035
Accuracy	Accuracy 標準差	平均代表句數	平均代表句數標準差
0.807	0.007	3571.4 句	0.800



圖十、「出發」代表句擷取分數評估

（四）、代表句擷取之綜合比較

表六展示出各目標詞彙於最佳分群模型時，透過擷取方法 1 所得到的分數進行比較。從中可以發現同形異義之目標詞彙其指標分數皆高於一詞多義之目標詞彙，且三種指標分數皆相去不遠。反之，一詞多義之目標詞彙其 macro-average 與 weighted-average 大多存在一定落差，且 accuracy 分數皆低於同形異義之詞彙。若與表七一同觀察，可以發現多數屬於一詞多義之詞彙其例句分布對於三種指標的影響較同形異義之詞彙來的大。由此可推論若例句分布不均，難以將各義項有關聯，即一詞多義，之詞彙例句區分開來。

表六、各目標詞彙代表句擷取分數比較

目標詞彙	Macro-average	Weighted-average	Accuracy	擷取數量	擷取比例
亞馬遜	0.840	0.851	0.856	483.0	57%
蘋果	0.871	0.900	0.896	4456.0	100%
小米	0.954	0.961	0.961	300.6	40%
出入	0.722	0.777	0.763	935.6	90%
出發	0.507	0.743	0.807	3571.4	99%
壓力	0.394	0.514	0.651	6699.2	99%
東西	0.476	0.763	0.813	4199.8	70%

表七、各目標詞彙之義項比例分布

目標詞彙	義項 1	義項 2	義項 3	總句數
亞馬遜	雨林 51.6%	電商 48.4%		847
蘋果	科技產品 60.0%	水果 21.1%	報紙 18.9%	4367
小米	科技產品 78.4%	農作物 21.6%		667
出入	出外與入內 81.4%	不一致 18.6%		1037
出發	實際離開 81.6%	從某方面著手 18.4%		3574
壓力	緊張不安的狀態 64.9%	單位面積上所受之力 35.1%		6734
東西	物品 53.3%	位置 44.2%	東方與西方 2.5%	5933

四、結論

本研究目的是透過分群機制對於多義詞進行消除歧義。在實驗中能觀察到屬於同形異義之目標詞彙能有效地將多義詞不同的義項類別區隔開來，同時擷取出符合義項之代表句供使用者閱讀。然而屬於一詞多義之目標詞彙則無法得到好的成效，探究其原因，或許是因為一詞多義之義項是具有關連性的，因此難以僅透過現有方式在相關句語境上判斷出明顯的區別。

此外，多義詞的各個義項之例句在語料庫中分布不均，在機器學習上也會是個很大的影響，因此並非給定任意多義詞皆能得到良好的分群結果。如實驗結果表六與表七所示，屬於不同類別之詞彙，對於例句分布的情形存在不一樣的結果。

最後，依據實驗統計，我們仍可以發現特定分群模型在多數的詞彙下有較好的表現，對於本實驗以外的多義詞，可以優先推薦使用者以該分群模型進行分析。不過若要更精確地了解該分群模型是否適合使用者給定的詞彙，以目前實驗方式需要由使用者提供參考句，若能免去人工標記這繁雜的工作，或許未來能有更廣泛的應用。

參考文獻

[1] Bruce K. Britton, “Lexical ambiguity of words used in English text,” *Behavior Research Methods & Instrumentation*, 10(1), 1-7, Jan. 1978.

[2] John Lyons, *Semantics*, Cambridge: Cambridge University Press, Oct. 1977.

[3] 教育部，〈教育部重編國語辭典修訂本〉，網址：http://dict.revised.moe.edu.tw/cbdic/search.htm。

[4] 唐鳳，〈萌典〉，網址：https://www.moedict.tw。

[5] 中央研究院，〈中央研究院現代漢語標記語料庫〉，網址：http://asbc.iis.sinica.edu.tw。

[6] 吳美嫻，《〈長阿含經〉雙音詞研究》，碩士論文，國立東華大學中國語文學系，2010。

[7] 林香薇，〈閩南語歌仔冊中的多義詞「落 loh8」〉，師大學報：語言與文學類，第 61:2 期，頁 1-28，臺灣：國立臺灣師範大學，2016 年 9 月。

[8] 蔡宛玲，《漢語多義詞「跑」之結構及語意分析》，碩士論文，國立政治大學語言學研究所，2017。

[9] 許尤芬，《中文多義詞「發」之語義探討：以語料庫為本》，碩士論文，臺北市立教育大學華語文教學碩士學位學程，2012。

[10] Roberto Navigli, “Word sense disambiguation: a Survey,” *ACM Computing Surveys*, 41(2), 10, Feb. 2009.

[11] Michael Lesk, “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone,” *Proceedings of the 5th Annual Conference on Systems Documentation*, Toronto, Ontario, Canada, 24-26, 1986.

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to information retrieval*, Cambridge: Cambridge University Press, 356-360, 2008.

[13] 中文維基百科，〈維基百科:資料庫下載〉，網址：https://dumps.wikimedia.org/zhwiki。

[14] 教育部，〈新聞立場檢索技術獎金賽〉，
網址： https://aidea-web.tw/topic/b6abbf14-2d60-456c-8cbe-34fdfcd58967。

[15] Giuseppe Attardi, “WikiExtractor,” [Online]. Available: https://github.com/attardi/wikiextractor.。

[16] Carbo Kuo, “OpenCC,” [Online]. Available: https://github.com/BYVoid/OpenCC。

[17] 中央研究院詞庫小組，〈CKIP 中文斷詞系統〉，網址：http://ckipsvr.iis.sinica.edu.tw。

[18] ldkrsi, “jieba-zh_TW,” [Online]. Available: https://github.com/ldkrsi/jieba-zh_TW。

[19] Quoc Le, and Tomas Mikolov, “Distributed representations of sentences and documents,” *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 32, 1188-1196, 2014.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient estimation of word representations in vector space,” *Proceedings of the International Conference on Learning Representations*, Scottsdale, Arizona, United States, 2013.

[21] James MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297, 1967.

[22] Leonard Kaufman, and Peter J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, New Jersey: John Wiley & Sons, Inc., 1990.

[23] Wilm E. Donath, and Alan J. Hoffman, “Lower bounds for the partitioning of graphs,” *IBM Journal of Research and Development*, 17(5), 420-425, Sept. 1973.

[24] Miroslav Fiedler, “Algebraic connectivity of graphs,” *Czechoslovak mathematical journal*, 23(2), 298–305, 1973.

[25] Jianbo Shi, and Jitendra Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888-908, Aug. 2000.

[26] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, “On spectral clustering: analysis and an algorithm” *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 849-856, Dec. 2001.

[27] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, “BIRCH: an efficient data clustering method for very large databases,” *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 25(2), 103-114, Jun. 1996.

[28] David Arthur, and Sergei Vassilvitskii, “K-means++: the advantages of careful seeding,” *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*,” 1027-1035, Jan. 2007.

[29] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to information retrieval*, Cambridge: Cambridge University Press, 279-285, 2008.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 296-309

Influences of Prosodic Feature Replacement on the Perceived Singing Voice Identity

Kuan-Yi Kang
Department of Electrical Engineering
National Tsing Hua University
sjk102061231@gapp.nthu.edu.tw

Yi-Wen Liu
Department of Electrical Engineering
National Tsing Hua University
ywliu@ee.nthu.edu.tw

Hsin-Min Wang
Institute of Information Science
Academia Sinica
whm@iis.sinica.edu.tw

Abstract

Human perception on the singing voice differs with the factors of the singing voice and the subjects. On one hand, the background knowledge influences the understanding of voice for each subject. On the other hand, the difference of the voices presented to the subjects also affects the perception. In this paper, we discuss two factors reflecting on the similarity before and after singing voice conversion: prosodic features and subjects' familiarity to the singers. Three experiments were conducted. The first experiment tested the subjects' ability to identify the singer. The second experiment synthesized the singing voice with different singers' prosodic features, and let the subjects score the similarity. The third experiment presented timbre-converted singing voice with different combinations of prosodic features from two singers to the subjects for them to judge the similarity to the target singer.

The results show that, first, the number of prosodic features contained in the synthesized voice is positively correlated with the scores in identification and similarity. Also, subjects who are more familiar personally with the target singers have better identification scores than target-unfamiliar subjects on the timbre-converted singing voices.

Keywords: Singing voice conversion, Prosody, Human perception, Voice identity.

1. Introduction

In the task of voice conversion, subjective tests examining quality and similarity [1] are used to evaluate the synthesized results from human perception. The quality test asks the subjects to score how great the quality of the converted result is, while the similarity test questions about how similar the result is, comparing to the target speaker.

The acoustic characteristics of voice individuality can be described through timbre, pitch, intensity, duration, etc. [3]. Hence, the synthesis of the voice considers not only the timbre but also other prosodic features such as pitch, duration or intensity [4]. What effects the modification of these features have on the identification of speakers [5] and how these features represent individuality [6,7,8] have also been discussed in previous works.

Studies in voice perception have investigated on the acoustic features affecting voice identity with experiments using the correlation analysis and multidimensional scaling techniques [12]. The modifications of acoustic features [13,16] and vocal identity aftereffects [14] are also used to determine the relative importance factors on humans' ability of speaker identification.

While some studies on the contributing features of voice identity contain timbre and other prosodic features [12,13,16], this paper investigates only on the prosodic features in order to find out the essential prosodic parameters for changing the perceived singer identity. In addition, since that the majority of the voice conversion tasks only focus on the development of the timbre conversion [1], using the source prosodic features for the synthesized voice, we therefore want to examine how changing the prosodic features might potentially help convert the voice more convincingly, and how that would reflect on the subjective similarity test.

The result of a subjective test depends on the background knowledge of the participants. When the subjects know more about the audio presented, the knowledge might influence the test performance; in [2,15], the difference of the human ability on identifying familiar and unfamiliar voices was well discussed. Therefore, we would also like to examine how familiarity to the speaker reflects on the similarity performance in tasks related to the singing voice conversion.

In this paper, we follow a similar experiment design of perception test in [11] by mixing up the features of source, target or converted singing voice to discuss the effects of sound

modification on the perception. Three listening tests for the participants are designed in order to find out the effects due to the subjects' familiarity to the singers and due to the changes of prosodic features. Section 2 describes the experiments designed. Section 3 discusses the listening test results and the effects of different factors. Section 4 then gives the conclusion.

2. Methods

2.1 Recordings

The recordings consist of parallel singing voice data sung by two female singers, F1 and F5. The recording process was conducted in a quiet room with a microphone, an RME interface and Cubase software. Each singer sang 9 pop songs, a total length of 17 minutes. These recordings were further tuned using Cubase and cut into phrases, each phrase ranging from 7 to 15 seconds.

2.2 Participants

We invited 17 participants to the listening test using their own laptops and headphones. 53% had a music training background (have learned musical instruments for more than one year) and 47% did not. 35% of participants were familiar with the voices of the two singers before the listening test, 24% knew one of them, and 41% of subjects were unfamiliar with the voices of the two singers. 47% of subjects were familiar with singer F1, and 47% of subjects were familiar with F5.

2.3 Experiment design

Before starting the test, all the participants were presented with a one-minute singing clip from both singers in order to be acquainted with both singers' singing voices. These audio files could be replayed during the test if the participants wanted to re-learn the singer's voice.

2.3.1 Singer identification

In the first task, we aimed to examine the participants' ability to distinguish signers from their singing voice. The participants were presented with 6 phrases from each singer in random order. The participants would then be asked to distinguish whether the audio presented was sung by F1 or F5.

2.3.2 Identification and similarity task of the timbre-carrying singer

The second task was designed to examine how the prosodic and timbre features would influence human perception, and how the changes of prosodic features would affect the identification task. We fixed the timbre of one singer and selectively replaced the expressive features (pitch, intensity, and duration) from the singer to the other singer. Out of the three expressive features, there were 8 kinds of combinations for feature replacement. For each combination, there were 4 examples (2 examples for a singer). The feature replacement combinations and their nomenclature are summarized in Table 1. Singer A indicates the original timbre-carrying singer, and singer B is the singer whose features are used to substitute singer A's features. In total, 32 audio files were thus prepared for participants to listen to.

The fundamental frequency and spectral envelope of a singing voice were extracted with the WORLD vocoder [10]. We used the fundamental frequency as the pitch feature. The spectral envelope was further compressed into mel-cepstral coefficients, with the first dimension defined as the intensity feature and the other dimensions defined as the timbre feature in this experiment. The duration features were modified through the dynamic time wrapping (DTW) conducted on the timbre feature if the selected identity was singer B; the pitch and intensity of the target would be adjusted to the source length through DTW if the duration identity was singer A. The selected and modified features were then synthesized into the singing voice with the vocoder.

Table 1. Feature Combination of Synthesized Voice

Nomenclature	**Timbre**	**Pitch**	**Intensity**	**Duration**
AAAA	Singer A	Singer A	Singer A	Singer A
AAAB	Singer A	Singer A	Singer A	Singer B
AABA	Singer A	Singer A	Singer B	Singer A
AABB	Singer A	Singer A	Singer B	Singer B
ABAA	Singer A	Singer B	Singer A	Singer A
ABAB	Singer A	Singer B	Singer A	Singer B
ABBA	Singer A	Singer B	Singer B	Singer A
ABBB	Singer A	Singer B	Singer B	Singer B

For each audio file that was listened to, the participants were asked to perform singer identification and score the similarity. The identification task asked which singer sang the audio, in the format of ABX test, while the voice of the two singers were introduced in the

section before the first experiment (2.3.1). The similarity test asked how similar the audio was to the timbre-carrying singer with the minimum opinion score (MOS).

2.3.3 Identification and similarity task of the timbre-converted singer

Existing voice conversion algorithms mostly apply original dynamic changes of source prosodic features on the synthesis results with the timbre converted through models [1]. The third task was to examine how the conversion of the expressive features may play some roles in the similarity test. Based on the timbre converted from the source to the target using a Gaussian mixture model [9], the experiment here tested 8 kinds of feature combinations on pitch, intensity, and duration, shown in Table 2, to test the effects of the prosodic features on human perception. The source singer is F5 and the target singer is F1. Each type contains 3 different singing phrases with timbre converted, so 24 files in total were presented to the participants.

The usage of each expressive feature follows a similar procedure in Sec. 2.3.2. The frame-based spectral features of the source singer were converted with a Gaussian mixture model [9]. The expressive features would be used directly without modification if the features' assigned identity was the source. If the target's duration feature was used, all the other features would be adjusted to match the target's length based on the DTW alignment of the spectral features of the two singers. The target's pitch and intensity could be extracted with the vocoder, and the length would be adjusted through DTW if the duration scale was assigned as the source.

Table 2. Origin of Features of Synthesized Voice

Nomenclature	**Timbre**	**Pitch**	**Intensity**	**Duration**
CSSS	Converted	Source	Source	Source
CSST	Converted	Source	Source	Target
CSTS	Converted	Source	Target	Source
CSTT	Converted	Source	Target	Target
CTSS	Converted	Target	Source	Source
CTST	Converted	Target	Source	Target
CTTS	Converted	Target	Target	Source
CTTT	Converted	Target	Target	Target

For each audio file they listened to, the participants were also asked two questions,

identification and similarity, with the identification task forcing the listener to determine which singer produced the audio, and the similarity task asking the listener to score how similar the audio was to the target singer.

3. Results of experiments

3.1 Singer identification

The performance of the subjects on the identification task of the original singers is shown in Table 3. The overall accuracy of all subjects is 80.39%. When the subjects are more familiar with the singers before the experiment, they will perform better on the identification task. The subjects who knew both singers achieved 95.83% accuracy, while the subjects not knowing any of the singers had 64.29% accuracy. The subjects with a music background were 8% more accurate than the subjects without a music background.

Table 3. Identification accuracy % (mean ± std)

All subjects	
All	80.39±20.61
Familiarity to the singers	
Knowing 0 singer	64.29±20.81
Knowing 1 singer	85.42±14.23
Knowing 2 singers	95.83±6.97
Music background	
Without music background	76.04±23.33
With music background	84.26±18.37

3.2 Identification and similarity task of the timbre-carrying singer

The modification of the expressive features from the original singer could change the perceived singer identity. Using 8 combinations of expressive features, without the conversion of timbre in the singing voice, the scores for each type (4 examples) were first averaged for each subject. Then, for each type, the average scores of the 17 subjects were analyzed to obtain the mean and standard deviation. The identification and similarity scores for each combination of expressive features are shown in Figure 1.

As can be seen from Figure 1, when some prosodic features are replaced, the identification score and the similarity score do decrease, especially AABB, ABAA, and

ABBB have the lowest scores compared with AAAA. The reason why the scores of AAAB and ABAB are comparable to the scores of AAAA could be that the two singers sang in a similar style in the audio files randomly selected for these types, so that after replacing one or two expressive features, the clip could still be recognized as the original singer.

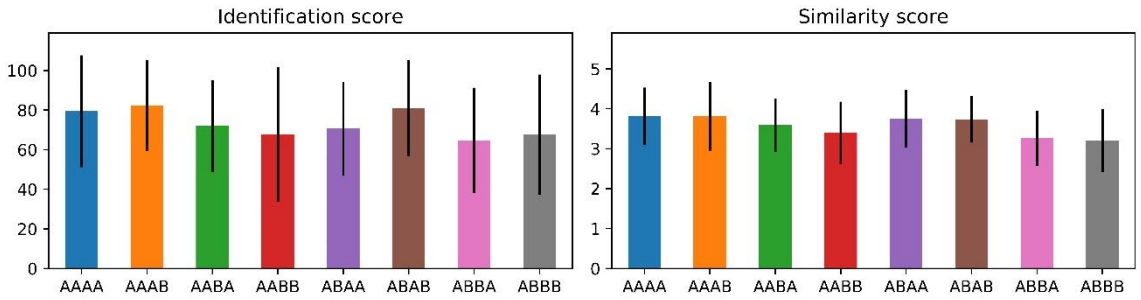


Figure 1. The identification score (%) and similarity score (mos) of the timbre-carrying singer with changed prosodic features.

The dependency of the scores upon the number of replaced expressive features was analyzed and shown in Table 4. Category 0 denotes AAAA while category 3 (all 3 expressive features were replaced) denotes ABBB. Category 1 (1 expressive feature was replaced) consists of AAAB, AABA, and ABAA, while category 2 (2 expressive features were replaced) consists of AABB, ABAB, and ABBA. For categories 2 and 3, the mean and standard deviation of 17 subjects and 3 feature combinations were calculated. It is clear that the identification and similarity scores decrease as more changes were made to the original singer's expressive features. When the number of replaced features increases from 0 to 3, the identification score drops from 79.41% to 67.65% and the similarity score drops from 3.81 to 3.21.

Table 4. Identification and similarity scores (mean ± std) of the timbre-carrying singer with different numbers of replaced expressive features

Number of replaced features	Identification Score %	Similarity Score
0	79.41±28.23	3.81±0.72
1	75.00±23.45	3.72±0.75
2	71.08±28.88	3.47±0.71
3	67.65±30.32	3.21±0.78

Table 5 shows the results with a specific expressive feature replaced. There are 3

expressive features: pitch, intensity and duration. For each feature, the mean and standard deviation of 68 samples (17 subjects and 4 feature combinations) were calculated. For example, for types AABA, AABB, ABBA, and ABBB, the "intensity" was replaced. Compared to duration, changes in pitch and intensity have lower scores and greater reduction from AAAA, indicating a greater impact on human perception of singer individuality.

Table 5. Identification and similarity scores (mean ± std) of the timbre-carrying singer with a specific expressive feature replaced

Features changed	Identification Score %	Similarity Score
Pitch	70.96±26.39	3.49±0.73
Intensity	68.01±28.27	3.36±0.73
Duration	74.63±28.49	3.54±0.78

Since each subject had different levels of prior knowledge about the singing voices that were presented, we divided the subjects into 3 categories: familiar with none, 1, or 2 of the singers before the experiments. The score distributions are depicted in Figure 2. Over all, the scores increase when the number of familiar singers increases. In other words, for the cases where timbre was unchanged but some expressive features were changed, the subjects familiar with more singers had better performance on singer recognition. Even with the changes in expressive features, the subjects who know the two singers have an identification score of higher than 70%, suggesting that the subjects tend to rely on the timbre as the cue to identify the singer.

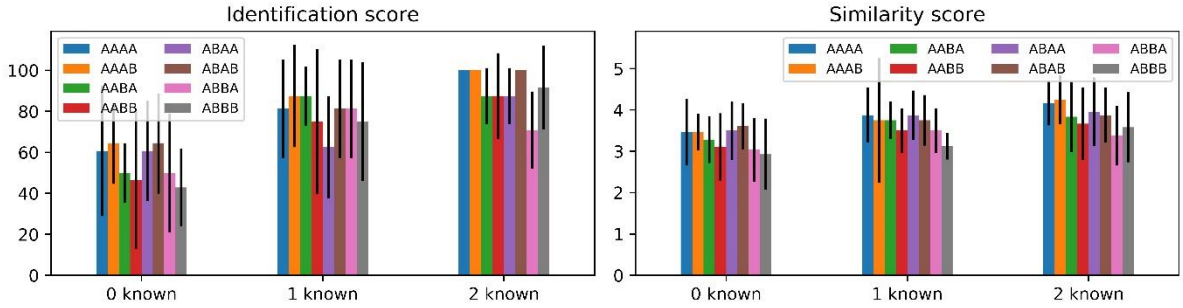


Figure 2. The identification score % and similarity score of the timbre-carrying singer with changed prosodic features and the number of familiar singers.

3.3.3 Identification and similarity task of the timbre-converted singer

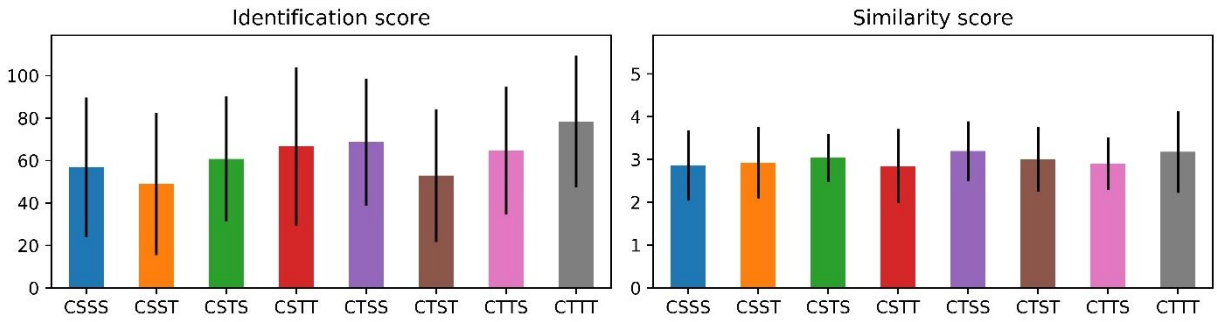


Figure 3. The identification score % and similarity score of the timbre-converted singer with changed prosodic features.

In the conversion task, the expressive features used in the synthesized results could lead to different perceptions in different subjects. For the 8 expressive feature combinations with converted timbre, the scores of each type for each subject were first averaged. We then calculated the statistics of the scores among all 17 subjects for each type. The results are shown in Figure 3.

The similarity score is around 3.0, meaning that the subjects were neutral on the decision of the similarity to the target singer, although the score seemed to slightly increase when more target expressive features were used. For the identification score, the situation CTTT achieved the best performance, which was 78.41% and was 21.55% higher than CSSS as shown in Table 6, suggesting that using all the three kinds of target prosodic features led to much better identification scores than using only the source prosodic features.

Table 6. Identification and similarity scores (mean ± std) of the timbre-converted singer using different numbers of target expressive features

Number of changed features	Identification Score %	Similarity Score
0	56.86±32.84	2.86±0.82
1	59.48±31.49	3.05±0.70
2	61.44±32.91	2.92±0.74
3	78.41±31.05	3.18±0.96

The scores with different numbers of target expressive features used are shown in Table 6. The calculation of the results was the same as in Table 4. It is clear that the identification score increases when more target expressive features are used. The score slightly increases

when using only 1 or 2 target expressive features, and the usage of all the target prosodic features improves the score from 56.86% to 78.41%.

Table 7 shows the results when a specific type of target expressive feature was used in the converted singing voice. The synthesized singing voice including target's pitch and intensity has higher probabilities to be identified as target than the synthesized singing voice including target's duration. The results reconfirm that pitch and intensity have a greater impact of human perception of singer individuality than duration.

Table 7. Identification and similarity scores (mean ± std) of the timbre-converted singer using a specific target expressive feature

Feature changed	Identification Score %	Similarity Score
Pitch	66.18±31.28	3.07±0.76
Intensity	67.65±32.05	2.99±0.76
Duration	61.76±34.68	2.99±0.85

Table 8. The identification score % of the timbre-converted singer with changed prosodic features and subjects' familiarity to the singers. (Subjects: 9 unfamiliar, 8 familiar)

	Target-unfamiliar	Target-familiar
CSSS	55.56±37.27	**58.33**±29.55
CSST	51.85±29.40	45.83±39.59
CSTS	62.96±20.03	58.33±38.83
CSTT	51.85±33.79	**83.33**±35.63
CTSS	70.37±20.03	66.67±39.84
CTST	44.44±28.87	**62.50**±33.03
CTTS	62.96±26.06	**66.67**±35.63
CTTT	77.78±28.87	**79.17**±35.26
All	59.72±29.04	**65.10**±35.85

The effects of subjects' familiarity to the singers on different feature combinations are shown in Table 8. While considering only the factor of target-familiarity (averaging the scores over subjects and feature combinations), the identification score of the target-familiar subjects is 5.38% higher than that of the target-unfamiliar subjects. The target-familiar subjects achieved higher identification scores than the target-unfamiliar subjects in most

combination types. For the cases of CTSS, CTTS and CTTT, the differences between two groups are relatively small and both have the accuracy higher than 60%.

3.3.4 Discussion

In the second experiment (cf. Sec. 3.3.2), the change of the prosodic features in the singing voice of the original timbre-carrying singer degraded the performance of the identification task. In the third experiment (cf. Sec. 3.3.3), the type CTTT achieved the best performance among the 8 types. When more expressive features of the target were used in the converted singing voice, the identification score for the target singer improved. Both experiments show that the modification of pitch and intensity have higher influences on human perception of singer individuality than duration.

Subjects' familiarity to the singing voice also influences the identification task. In the first experiment (cf. Sec. 3.3.1), the subjects who were familiar with the singers personally achieved higher identification scores than the subjects who were not familiar with the singers. In the second experiment, even with the changes in expressive features, the singer-familiar subjects seemed to be able to rely on the timbre for identification, thus maintaining an identification score of greater than 70% (for the subjects who know both singers). The third experiment showed that the target-familiar subjects had a higher identification score than the target-unfamiliar subjects and more subjects in target-familiar groups successfully recognized the synthesized results as produced by the target singer when converted timbre and target expressive features were used.

Each feature combination type in the experiment consisted of 3 or 4 randomly selected files from the data set, and the scores discussed were based on the answer of the 3 or 4 files. For some phrases, the two singers might sing in a similar way with less individuality in the singing voice. If these kinds of files were selected, the identification would depend more on the individuality of timbre itself rather than the expressive features. In the future, more audio files of each types and more subjects could be included in order to cover more situations so we may have stronger conclusions supported by rigorous statistical analysis.

4. Conclusion

In this paper, three perception experiments were designed to find out the influence of the expressive features and the familiarity of the subjects to the singer on the perceived singer

identity in the synthesized singing voice. Identification and similarity tests were conducted.

We found that, first, with the timbre unchanged, the modification of the expressive features to another singer degraded the performance on the identification task. Secondly, during the voice conversion task, the identification scores for the target singer improved when more expressive features of the target were used in the converted singing voice. In addition, in the task of examining the timbre-converted voices, the target-familiar subjects had a higher identification score than the target-unfamiliar subjects when target expressive features were used.

From these experiments, we therefore conclude that not only the timbre but also the expressive features play a role on capturing the singer's identity in voice perception, and the subjects' familiarity to the voice presented also influences the results. The task of voice conversion should also take the conversion of expression features and the subjects' familiarity to the voice into consideration in the future development. In addition, to support our findings by rigorous statistics, more coverage of audio files shall be used and more subjects can be included in future work.

5. Acknowledgements

We would like to thank Prof. Shan-Hung Wu of National Tsing Hua University for the support on the singing voice research. We are also grateful to Hsin-Te Hwang and Yu-Huai Peng who provided suggestions on voice conversion that greatly assisted the research.

References

[1] S. Mohammadi, A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, no. Supplement C, pp. 65–82, 2017.

[2] S. Schweinberger, et al. "Speaker perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.

[3] H. Kuwabara and Y. Sagisak, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, no. 2, pp. 165–173, 1995.

[4] M. Schröder, "Emotional speech synthesis: A review," *Proc. Eurospeech*, pp. 561–564,

2001.

[5] Y. Lavner, I. Gath, and J. Rosenhouse, “The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels,” *Speech Communication*, vol. 30, no. 1, pp. 9–26, 2000.

[6] L. He and V. Dellwo, “Between-speaker variability in temporal organizations of intensity contours,” *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 488–494, 2017.

[7] V. Dellwo, A. Leemann, and M.-J. Kolly, “Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors,” *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513–1528, 2015.

[8] A. Leemann, M.-J. Kolly, and V. Dellwo, “Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison,” *Forensic Science International*, vol. 238, pp. 59–67, 2014.

[9] T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] M. Morise, F. Yokomori, and K. Ozawa, “World: A vocoder based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[11] Z. M. Smith, B. Delgutte, and A. J. Oxenham, “Chimaeric sounds reveal dichotomies in auditory perception,” *Nature*, vol. 416, no. 6876, pp. 87–pp.70, 2002.

[12] O. Baumann and P. Belin, “Perceptual scaling of voice identity: common dimensions for different vowels and speakers,” *Psychological Research PRPF*, vol. 74, no. 1, pp. 110, 2010.

[13] G. Sell, C. Sujed, M. Elhilali, and S. Shamma, “Perceptual susceptibility to acoustic manipulations in speaker discrimination,” *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 911–pp.922, 2015.

[14] M. Latinus and P. Belin, “Perceptual auditory aftereffects on voice identity using brief vowel stimuli,” *PLoS One*, vol. 7, no. 7, pp. e41384, 2012.

[15] S. R. Mathias and K. von Kriegstein, “How do we recognise who is speaking?,” *Front Biosci (Schol Ed)*, vol. 6, pp. 92–109, 2014.

[16] Y. Lavner, I. Gath, and J. Rosenhouse, “The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels,” *Speech Communication*, vol. 30, no. 1, pp. 9–26, 2000.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 310-324

以三元組損失微調時延神經網路語者嵌入函數之語者辨識系統

Time Delay Neural Network-based Speaker Embedding Function Fine-tuned with Triplet Loss for Distance-based Speaker Recognition

葉致廷 Chih-Ting Yehn, 王伯晉 Po-Chin Wang,
張蘇瑜 Su-Yu Zhang, 陳嘉平 Chia-Ping Che
國立中山大學資訊工程學系
Department of Computer Science and Engineering
National Sun Yat-sen University
M063040008@student.nsysu.edu.tw, M063040058@student.nsysu.edu.tw,
M073040069@student.nsysu.edu.tw, cpchen@mail.cse.nsysu.edu.tw,

蕭善文 Shan-Wen Hsiao, 詹博丞 Bo-Cheng Chan, 呂仲理 Chung-li Lu
中華電信研究院
Chunghwa Telecom Laboratories, Taoyuan, Taiwan
swhsiao@cht.com.tw, cbc@cht.com.tw, chungli@cht.com.tw

摘要

本研究工作提出以語者驗證的 x 向量（x-vector）架構為基礎，建立一套語者辨識系統。系統訓練時，我們提出利用三元組損失（triplet loss）來拉開不同語者語句嵌入向量之間的距離。系統辨識時，則是直接使用歐氏距離做為註冊語者與測試音檔之間相似度的量測，並以最小距離的註冊語者為辨識結果。我們以名人聲音（VoxCeleb）語者辨識資料集評估所提出的系統，其中測試資料集包含 1,251 位名人的語音資料。我們所提出的系統單一輸出（top-1）的辨識正確率為 59.57 %，前五個輸出（top-5）的辨識正確率則可以達到 80.32%。

Abstract

In this research work, we build a speaker recognition system based on the x-vector framework for speaker verification. During training, we propose to use the triplet loss to increase the distance between the embedding vectors from different speakers in high-dimensional space. During recognition, we use the European distance between test-utterance embedding vector and enrolled-speaker embedding vector for similarity measure, thus predicting the enrolled speaker with the minimum distance. The proposed system is evaluated with VoxCeleb speaker recognition dataset. The test set consists of utterances from 1,251 test speakers. The proposed

model achieves the top-1 recognition accuracy of 59.57% and the top-5 accuracy of 80.32%.

關鍵詞：時延神經網路、語者辨識、三元組損失

Keywords: TDNN, Speaker Recognition, Triplet Loss

一、緒論

隨著大數據的時代到來，深度學習成為時下的熱門議題之一，並慢慢走入我們的生活之中，而身分驗證與識別就是一個主要的應用範疇，以前只出現在電影裡的生物特徵辨識系統也逐漸可以在我們的日常生活中發現蹤跡，如聲紋這類生物特徵不似傳統的鑰匙或是遙控器，可能會有遺失的風險存在，基於聲紋的語者辨識不需攜帶額外的物品，僅需聲音便可以進行辨識，達到更便利、安全的效果。

其中，語者驗證是當前熱門的研究項目，說話人須先宣稱其身份，再由語者驗證系統進行比對，做出是否通過驗證的決斷。然而在一些實際的應用上，人們開始追求指令簡明與便捷性，若能去除宣稱身份的步驟，便可讓使用者的體驗更佳。如智慧家庭產品許多都採取聲控的方式來下達指令，除了辨識使用者所下達的指令之外，更希望能對下達指令的人進行辨識，來達到一些簡單客製化回應的效果，例如我們若能分辨下達指令的是家中哪位成員，便能提供適合且客製化的回應給使用者，就像是一個簡單的播音樂指令，對家中長者可以播放台語經典歌曲，而年少者則可播放時下偶像團體的歌曲。

然而，在實際使用時需要辨識的語者或是說註冊者，大多時候會與訓練時所使用的訓練資料中的語者不同，所以無法簡單地使用基於 Softmax 的分類器來處理，為了解決這個問題，在本論文中，我們以 x 向量 [1] 架構為基礎，並透過表徵學習 (Representation Learning) [2] 的方式，從時延神經網路 (Time Delay Neural Network, TDNN) [3] 中取得嵌入向量 (Embedding Vector) ，並藉由對註冊語音之嵌入向量進行註冊的動作，為註冊者建立語者模型，在測試語音進入系統時，會對測試語音之嵌入向量與所有註冊者之語者模型進行相似度比對，並選出最相似者做為系統判定測試語音所屬之語者。為了使準確率提升，我們不僅使用交叉熵損失 (Cross Entropy Loss) 來訓練模型，也採用三元組損失 (Triplet Loss) 來對模型進行調適，使不同語者的嵌入向量在高維空間有更好的判別性。

本文主要分為五個部份：第一部份為緒論；第二部份為相關研究的回顧與探討；第三部份為研究方法與流程，介紹使用資料集、資料前處理、模型架構、訓練流程以及註冊與相似度比對等流程；第四部份則為實驗結果與分析，說明實驗環境與設定，並根據實驗結果進行分析；第五部份為結論。

二、相關研究

在語者辨識與驗證技術發展之中，高斯混和模型 (Gaussian Mixture Model) [4] 扮演十分重要的角色，而高斯混和模型-通用背景模型 (Gaussian Mixture Model-Universal Background Model, GMM-UBM) [5] 更是一個重要的應用，通用背景模型是一種大型的高斯混和模型，並非像傳統高斯混和模型針對每個語者訓練一個高斯混和模型來表示語者的特徵分佈，而是先使用所有語者的資料去訓練一個通用的背景模型，表示出語者無關 (Speaker-independent) 的特徵分佈，然後再使用指定語者的資料去調適背景模型產生語者模型。之後，為了解決在不同錄音裝置上，同語者的錄音聽起來會不一樣的問題，聯合因素分析 (Joint Factor Analysis, JFA) [6] 提出將 GMM-UBM 所得出超級向量 (Supervector) 進行因素分析，可分為通道子空間 (Channel Subspace) 與語者子空間 (Speaker Subspace)，JFA 相信若能去除通道因素的影響，那我們就能去除不同錄音條件的影響，使系統更強健。然而在[7] 中卻發現，通道部份仍然包含語者資訊，為了解決這個問題，提出了一種將語者空間與通道空間整合為單一的全局差異空間 (Total Variability Space)，而對應的全局因子則被稱為 i-vector (Identity Vector) [8]。在 2010 年至 2016 年之間的語者驗證系統幾乎都採用 i-vector 或以此為基礎進行改進。

另一方面，由於深度學習 (Deep Learning) 在圖像辨識的成功，人們也嘗試將深度類神經網路應用在語者辨識的任務上。在 2014 年 d-vector [9] 使用四層 256 維隱藏層的多層感知器進行表徵學習，並從最後一層隱藏層擷取出嵌入向量，這也啟發了其他使用卷積神經網路 (Convolutional Neural Network, CNN) [10] 或是遞迴神經網路 (Recurrent Neural Network, RNN) [11] 開發語者驗證系統的想法。到了 2016 年，使用時延神經網路 [3] 的 x-vector [1] 被提出，它最重要的特色在於對訓練音檔進行如：加入噪音、迴響、變速等數據增強 (Data Augmentation) [12]，使訓練資料呈倍數成長並獲得超越當時其他系統的強健性。如今 x-vector 已經是目前最主流的語者識別與驗證系統之一，本論

文的語者辨識系統也以 x-vector 為基礎進行改進。

三、研究方法與流程

（一）、VoxCeleb 資料集

VoxCeleb 資料集可分為 VoxCeleb1 [13] 與 VoxCeleb2 [14]，兩者皆為文本無關 (Text-Independent) 語音資料集，內容源自於 Youtube 中名人的影片，因此內容可能會有背景雜音甚至是其他人說話的聲音，資料集除了提供語者身分之外，也提供該語者國籍以及性別，兩資料集的資料分佈狀況如表一。VoxCeleb1 官方提供兩種資料集分割方式，語者驗證分割 (Verification Split) 與語者識別分割(Identification Split) ，兩種分割包含之音檔相同，差別僅在於依任務目的不同而把資料集進行不同的切割分法，其中語者驗證分割之驗證集與測試集中的語者不重複，而語者識別分割之訓練集、驗證集、測試集中的語者相同。

表一、VoxCeleb 資料分佈表

	VoxCeleb1					VoxCeleb2	
	驗證分割		辨識分割				
	dev	test	train	dev	test	dev	test
語者數	1,211	40	1,251	1,251	1,251	5,994	118
音檔數	148,642	4,874	138,361	6,904	8,251	1,092,009	36,237

在資料使用上，我們使用 VoxCeleb2 驗證集來訓練模型、使用 VoxCeleb1 識別分割的測試集來進行語者辨識的效能評估、使用 VoxCeleb1 驗證分割的測試集來進行語者辨識，更進一步的來研究是否語者驗證的準確率是否與語者辨識正相關。

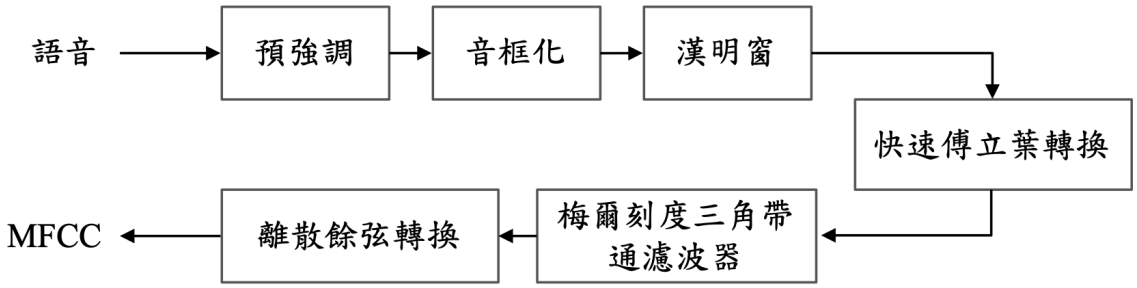
（二）、數據增強

我們採用數據增強 (Data Augmentation) 的技術，對資料加入噪音、或是對音檔加入迴

響，不僅僅增加資料的數量與多樣性，也更能使系統更加強健。我們使用 MUSAN 語料庫 [12] 加入噪音與利用房間脈衝響應 (Room Impulse Response) 加入迴響來進行數據增強。而 MUSAN 語料庫內容包含三個部分，分別為演說 (Speech) 、音樂 (Music) 與噪音 (Noise) ，演說部分的內容為朗讀書本某章節內容的語音或是美國聽證會或辯論會的公開演說；音樂部分則涵蓋古典與現代流行樂；噪音部分包含各類常見噪音，但不包括明顯可辨識說話內容的人聲。在產生數據增強後的音檔後，因為運算資源的考量，我們並不會全部使用，而是隨機取 1,000,000 個數據增強的音檔與原始音檔進行模型的訓練。

（三）、聲學特徵與正規化

我們所使用的聲學特徵為梅爾倒頻譜係數 (Mel-Frequency Cepstral Coefficient, MFCC) ，是一種針對人耳聽覺而設計的一種聲學特徵，由於人耳對不同頻率的聲音有不同的敏銳程度，故我們在頻率座標軸依梅爾刻度 (Mel Scale) 配置在低頻較密集、高頻較稀疏的三角帶通濾波器 (Triangular Bandpass Filters)，表示人耳對低頻聲音感受較為敏銳但面對高頻聲音便相較為遲鈍。梅爾倒頻譜係數聲學特徵處理流程如圖一，語音訊號經過預強調(Pre-emphasis)，來提升高頻的部份，始信號的頻譜變得平坦。之後將多個取樣點集合成一個觀測單位，稱為音框(Frame)，並對每一個音框乘上漢名窗(Hamming window)來增加音框左端與右端的連續性。由於訊號在時域上的變化很難看出訊號的特性，因此會先經由快速傅立葉轉換(Fast Fourier Transfrom, FFT) 將訊號從時域信號轉換到頻域信號上，以能量分佈來觀察。再將得到的頻譜乘上多組三角帶通濾波器，得到每一個濾波器輸出的對數能量(Log energy) 後，將對數能量經離散餘弦轉換(Discrete Cosine Transform, DCT)後即可得梅爾倒頻係數。



圖一、梅爾倒頻譜係數聲學特徵處理流。

由於實際應用時收音容易受環境以及錄音裝置等因素干擾，測試時的環境可能與訓練音檔的錄製環境大不相同，進而影響實際使用時的準確率。除了使用數據增強盡可能的模仿各種不同的環境之外，我們也加入特徵正規化的方法幫助我們增加系統的強健性，其中倒頻譜平均值與變異數正規化法 (Cepstral Mean and Variance Normalization, CMVN) 是常見特徵正規化的方法之一，藉由整段語音特徵的平均與標準差對特徵進行標準化，假設一音檔的音框長度為T，每一音框之特徵維度為N，$1 \leq t \leq T$且$1 \leq i \leq N$，$x_t(i)$表示為第t個音框中第i維的特徵，而將 $x_t(i)$正規化至 $\hat{x}_t(i)$的公式如下：

$$\hat{x}_t(i) = \frac{x_t(i) - \mu(i)}{\sigma(i)} \tag{1}$$

其中

$$\mu(i) = \frac{1}{T}\sum_{t=1}^{T} x_t(i) \tag{2}$$

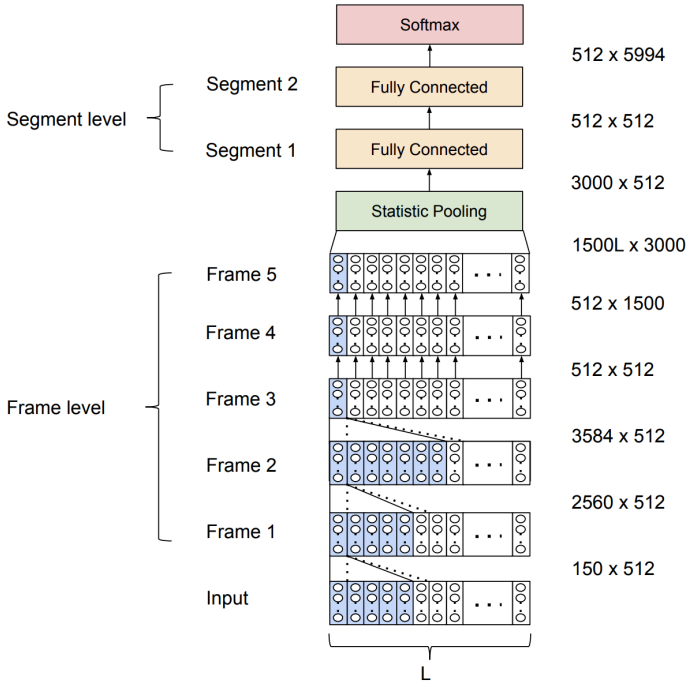
$$\sigma(i) = \frac{1}{T}\sum_{t=1}^{T} (x_t(i) - \mu(i))^2 \tag{3}$$

（四）時延神經網路

我們使用的時延神經網路模型如圖二，時延神經網路可以分為兩個部份，分別為音框層級 (Frame Level) 與音段層級 (Segment Level) ，中間由統計池化層 (Statistic Pooling

Layer) 來將音框層級的資訊轉為音段層級。而模型最後一層輸出層為 5,994 維 softmax 機率。

在我們的模型中音框層級為五層架構，第一層與第二層取相鄰的 5 個音框為輸入進行運算，到第三層則是取相鄰的 7 個音框，第四層與第五層則僅取 1 個。時延神經網路便是透過隱藏層的堆疊來提煉連續音框的特徵，且隨著隱藏層的增加，神經網路可以收集到更大範圍音框的資訊；在音框層級結束時，會在統計池化層對所有音框計算平均與變異數，整合成音段層級的資訊。此外，在這個模型中，每層隱藏層皆經過批量標準化(Batch Normalization) 與 Rectified Linear Unit (ReLU)激活函數。當模型訓練完成後，我們從 Segment 1 輸出取得 512 維的嵌入向量。



圖二、Softmax 訓練階段之時延神經網路架構圖：右側數字各層的輸入與輸出維度；L 為音檔音框數。

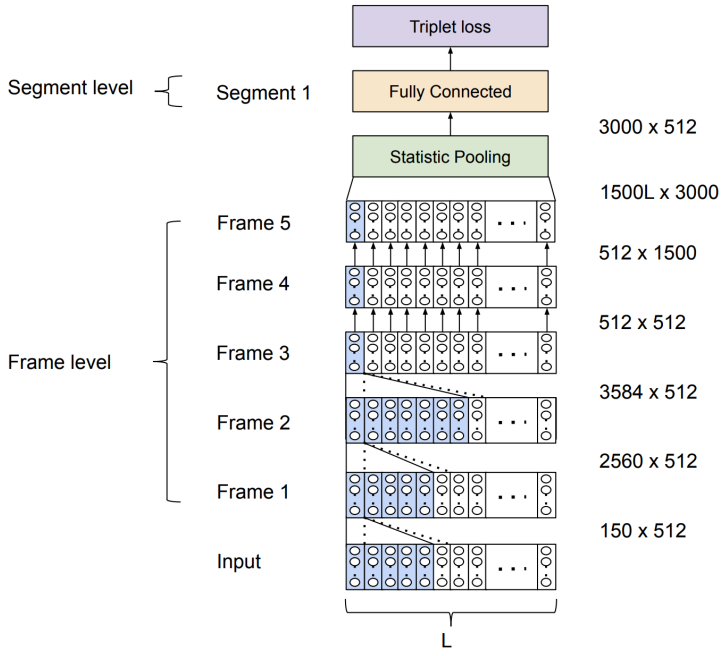
（五）損失函數

在初期訓練時延神經網路時，我們使用交叉熵損失為損失函式來更新模型，單筆資料時

的公式如下：

$$Loss_{CE} = -\sum_{i=1}^{C} y_i \, log(p(i)) \tag{4}$$

C為最後一層分類的類別數；y_i為表示該筆資料的真實類別，僅在該筆資料真實類別屬於第i類時為 1，其餘時候為 0；$p(i)$為神經網路預測第i類的機率。



圖三、三元組訓練階段之時延神經網路架構圖：右側數字各層的輸入與輸出維度；L 為音檔音框數。

在訓練完成後，為了更清楚得分辨出同語者的語音與其他語者的語音之間的差異，我們捨棄了用來提取嵌入向量的那層以後的每一層網路，如圖三所示，並改為使用三元組損失 [15] 來調整模型，三元組的定義如下：

錨點 (Anchor)： 訓練集中一語者 A 的真音檔。

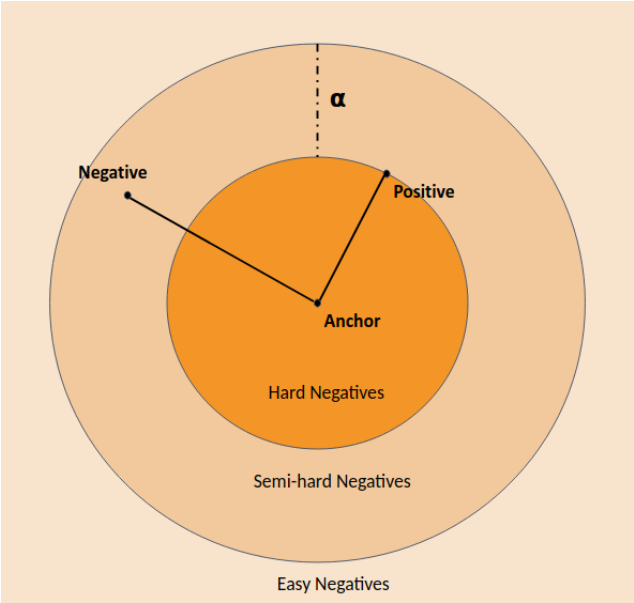
正樣本 (Positive)： 語者 A 與錨點不同的另一音檔。

負樣本 (Negative)：非語者 A 的另一語者之音檔。

在生成三元組的過程中，每筆音檔皆會成為錨點，與所有可能的正樣本去找尋一個符合條件的負樣本組成三元組，其中錨點與正樣本的組合不得重複，也就是說，任一錨點不會在它的正樣本被選為錨點時作為正樣本，而對負樣本的選擇條件為：

$$||E^a - E^p||_2^2 < ||E^a - E^n|| < ||E^a - E^p||_2^2 + \alpha \quad (5)$$

E^a為錨點的嵌入向量、E^p為正樣本的嵌入向量、E^n為負樣本的嵌入向量，α為定義負樣本的邊界值。在選定E^a與E^p的情況下，負樣本與錨點的距離必須小於正樣本與錨點的距離加上α，且大於正樣本與錨點的距離，這樣這筆其他語者的音檔才能被選為負樣本，目的在於選出半難負樣本 (Semi-hard Negatives) 來訓練，如圖四所示，如此一來，可以避免模型收斂在局部最小值 (Local Minima)。



圖四、三元組損失中各類負樣本的定義圖：圖中之負樣本為半難負樣本。

如此一來，三元組損失的損失函數定義如下：

$$Loss_{triplet} = [||E^a - E^p||_2^2 - ||E^a - E^n||_2^2 + \alpha]_+ \quad (6)$$

（六）註冊與相似度比較

不同於傳統語者識別中訓練集語者與測試語者重複，我們預設大多時候訓練集中的語者

會與實際應用時有所不同，所以我們必須對想要註冊的使用者，擷取其語音的嵌入向量，以提供註冊的功能，而我們採用對所有註冊語音得出的嵌入向量取平均做為註冊者的語者模型，當測試語音輸入時，則與所有註冊語者模型比較相似度，回傳相似度最高且高於閾值的註冊者為最終系統做出的辨識結果。

在相似度的比較上，由於三元組損失函式是計算錨點與正樣本、錨點與負樣本歐氏空間下的距離差，所以相似度評估使用歐氏距離來計算，得出的值愈小則表示兩者相似度愈高，歐氏距離的公式如下：

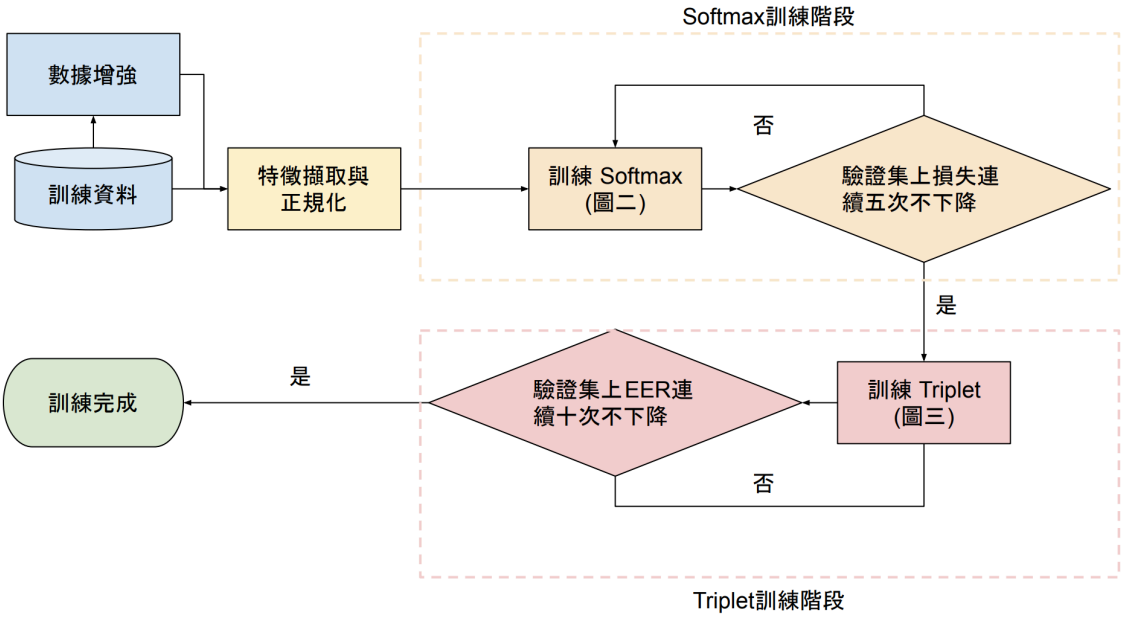
$$||X - Y||_2^2 = \sum_{i=1}^{n} (X(i) - Y(i))^2 \tag{7}$$

X與Y為要計算相似度的兩向量，皆為n維。這個方法與我們在計算三元組損失時計算嵌入向量之間距離的方法相同，也希望能藉此更符合訓練時的訴求，達到更好的效果。

四、實驗結果與分析

（一）實驗設定

在語者辨識的實驗中，我們使用 VoxCeleb2 驗證集共 5,994 位語者來訓練時延神經網路，我們從 VoxCeleb1 識別分割之驗證集每位語者取 5 句音檔進行註冊，再由測試集來測試模型，測試集共包含 1,251 位語者與 8,251 句音檔，模型的訓練流程如圖五。



圖五、模型訓練流程圖。

在此系統中，使用 30 維的 MFCC 特徵為輸入，音框 (Frame) 長度為 25 毫秒、每次移動 10 毫秒。包含特徵提取、數據增強等前處理採用 Kaldi 作為實作工具，模型訓練與相似度比對則是使用 TensorFlow 作為實作工具。

在 softmax 訓練階段過程中，我們取訓練集中每位語者 10 筆音檔做為驗證集，且使用早停法 (Early Stopping) 的機制，每訓練完一輪訓練資料去計算一次在驗證集上的損失，若該損失連續不下降 5 次則停止訓練。

在三元組訓練階段時，使用三元組損失訓練模型，我們並不從所有的訓練資料中組成三元組，而是每次取 90 位語者，每位語者取 20 句音檔，共 1,800 句音檔，我們由這些音檔構成三元組，來訓練與更新模型，這樣對語者進行採樣的方式相對於事先找出所有語者的三元組，能更快速地反應模型的現況，並找出對改善當前模型有幫助的三元組。在三元組訓練階段時，使用 VoxCeleb1 驗證分割的測試資料為驗證集，並進行語者驗證，以語者驗證的 EER (Equal Error Rate) 為判斷是否停止訓練的標準。每次訓練完一次採樣的資料後，便會進行一次語者驗證，由於每次採樣的資料僅 1,800 筆，所以連續 10 次的參數更新 EER 皆沒有下降才停止訓練。

在超參數方面，實驗中學習率皆設為 0.001，優化方法使用 Adam 演算法，三元組損失

的選擇邊界α為 0.2。

（二）實驗結果

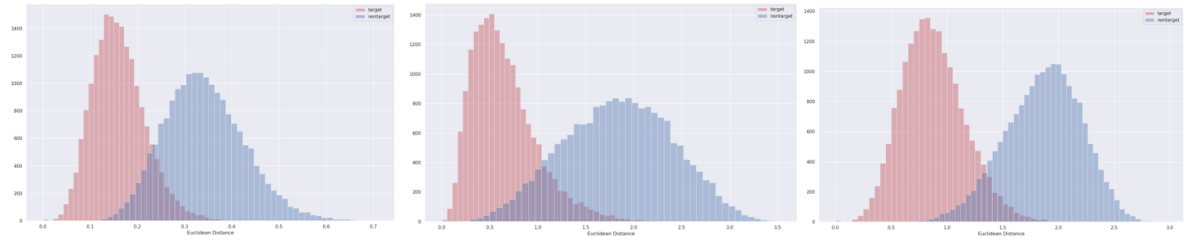
我們在本節比較訓練流程中是否使用三元組損失之間的差別 (是否有圖五中的三元組訓練階段) ，也觀察是否使用 softmax 預訓練模型對三元組損失訓練方法 (是否有圖五中的 softmax 訓練階段) 的影響，探討是否三元組損失在少了 softmax 的條件下，是否使同類資料的嵌入向量在高維空間中聚集。在三元組損失的損失函式中我們可以看出，訓練的核心概念在於拉開不同類別之間的距離，但缺少使同類別內的資料內聚的能力，例如在三元組的選擇中並未對正樣本的選擇做出限制，缺少拉近正樣本與錨點之間的距離的功能，所以在少了 softmax 時，也缺少了使同類資料內聚的能力，僅憑三元組損失把不同類別的距離拉開，是否能對訓練資料未曾出現過的語音得出判別性佳的嵌入向量是值得探討的問題。

首先，表二為在 VoxCeleb1 驗證分割的測試資料中利用官方提供的 trials 進行驗證的實驗結果，也是在使用三元組損失時使用的驗證資料的結果，包含 40 位語者共 37,720 筆 trials，target : nontarget 為 1:1。我們除了 EER 之外，也使用 minDCF (Minimum Decision Cost Function)來評估系統，參數設定比照 [14]，可以發現僅使用 softmax 時 EER 為 9.64%，而使用三元組損失而未使用 softmax 預訓練模型時，EER 則略高來到 10.39%，不過使用三元組損失且加上 softmax 預訓練模型做為起始權重的話，EER 有效降低至 6.84%，minDCF 也是最低，為 0.6278。

表二、在 VoxCeleb1 語者驗證之實驗結果

訓練方法	*EER*	*minDCF*
Softmax	9.64 %	0.7174
Triplet	10.39 %	0.8919
Softmax + Triplet	6.84 %	0.6278

此外，圖六為各訓練方法在語者驗證上分數的分佈圖，橫軸為 trials 中比對目標間的歐氏距離，而縱軸代表預測結果為該距離時資料的數目，紅色為 target trials 的分數分佈情況，藍色為 nontarget trials 的分數分佈情況。我們觀察的重點有二：一是 target 與 nontarget 分佈重疊的部份，代表模型可能分類錯誤的部份，重疊的面積愈小表示愈能將不同語者分類清楚，系統的效能也愈佳，從圖中可以發現 Softmax+Triplet 面積最小，同時也在實驗中有最好的效果；第二是 target 與 nontarget 分佈中心，在 Softmax 我們看到分布狀況與常態分佈相似，但有經三元組調適後，target 與 nontarget 的分佈中心皆有拉開彼此之間距離的情形發生，展現三元組拉開不同語者之間距離的效果。



(a) Softmax　　(b) Triplet　　(c) Softmax+Triplet

圖六、在 VoxCeleb1 語者驗證之分數分佈圖

在語者辨識上，我們以 Top-1 準確率與 Top-5 準確率為評估標準，Top-1 準確率表示僅系統判斷相似度最高者為測試語音所屬的語者才算正確，Top-5 準確率則是系統判定前五相似者中有測試語音所屬語者即算正確。

實驗結果如表三，我們發現有 softmax 預訓練且經三元組損失調適之後，在語者辨識上也有所進步，相較於僅使用 softmax 訓練 Top-1 準確率提升了約 5%，Top-5 準確率更上升了約 6.3%。但從實驗結果中我們也發現，沒有使用 softmax 預訓練模型的話效能會大大的降低，這點在語者辨識時尤其明顯。

表三、在 VoxCeleb1 語者辨識之實驗結果

訓練方法	*Top-1 準確率*	*Top-5 準確率*
Softmax	54.59 %	73.67 %

Triplet	23.68 %	45.58 %
Softmax + Triplet	59.57 %	80.32 %

五、結論

在本論文中，我們以 x 向量架構為基礎，開發語者辨識系統，透過改變損失函式的方法，將原本後端繁瑣的分類流程簡化至計算歐氏距離來比較測試語音與註冊者語音的相似度。此外，不同於常見的語者識別訓練集與測試集中的語者相同，為了使如智慧家庭產品等應用上更加方便，在註冊新增或刪除用戶上不受限制，我們在訓練集語者與測試集語者不同的條件下進行辨識，利用嵌入向量對註冊者建立語者模型，並在測試時找出與測試語音最相似的註冊者做為系統判定的結果。在實驗中，我們比較是否使用三元組損失的差異，發現無論在語者驗證上或是語者辨識上皆有所幫助，在 VoxCeleb1 識別分割測試集單一輸出（top-1）的辨識正確率為 59.57 %，前五個輸出（top-5）的辨識正確率則可以達到 80.32%，但另一方面我們也建議在使用三元組損失時，使用 softmax 預訓練模型，可以使模型更穩定且有更好的效果。

參考文獻

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in Proc. ICASSP, 2018.

[2] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture forefficient modeling of long temporal contexts,” in Proc. Interspeech, 2015.

[4] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” IEEE transactions on speech and audio processing, vol. 3, no. 1, pp. 72–83, 1995.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted

gaussian mixture models," Digital signal processing, vol. 10, no. 1-3, pp. 19–41, 2000.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1435–1447, 2007.

[7] N. Dehak, Discriminative and generative approaches for long-and short-term speaker characteristics modeling: application to speaker verification. PhD thesis, ´ Ecole de technologie sup´erieure, 2009.

[8] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[9] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4052–4056, IEEE, 2014.

[10] Y.-h. Chen, I. Lopez-Moreno, T. N. Sainath, M. Visontai, R. Alvarez, and C. Parada , "Locally-connected and convolutional neural networks for small footprint speaker recognition," in Proc. Interspeech, 2015.

[11] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in Proc. ICASSP, 2016.

[12] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484, 2015.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in Proc. Interspeech, 2017.

[14] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in Proc. Interspeech, 2018.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823, 2015.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 325-333

Building of children speech corpus for improving automatic subtitling services

Matus Pleva, Stanislav Ondas, Daniel Hládek, Jozef Juhar, Ján Staš
Department of Electronics and Multimedia Communications
Technical University of Kosice, Slovakia
matus.pleva@tuke.sk stanislav.ondas@tuke.sk daniel.hladek@tuke.sk
jozef.juhar@tuke.sk jan.stas@tuke.sk

Yuan-Fu Liao
Department of Electronic Engineering
National Taipei University of Technology
yfliao@mail.ntut.edu.tw

Abstract

This paper describes the development and first evaluation of the new Slovak children speech audio corpus for improving the automatic broadcast news subtitling engine developed on the Technical University of Kosice in cooperation with the Slovak Academy of Sciences. The current automatic speech recognition (ASR) systems are reliable for a clean, prepared speech of adults with not very long pause inside the sentences. For speech recognition of children's, it is still a challenge from different reasons. They use much slang, and diminutive words, undeveloped pronunciation, shorter vocal tract (different speech parameters), the sentence syntax is different. The paper presents the results of the children speech automatic recognition from the system built for broadcast news transcription.

Keywords: automatic speech recognition, children speech, audio corpus, annotation, database design.

1. Introduction

The speech technology has significant potential, currently it has growing interest among children and technically enthusiastic people [1]. The International Speech and Communication Association (ISCA) has a Special Interest Group (SIG) for Child Computer Interaction (CHILD) [2] and is organizing a special Workshop on Child Computer Interaction - WOCCI and last years also Language Teaching, Learning and Technology - LTLT. This year a special

session on the prestigious Interspeech conference will be held in September 2019 in Graz called Spoken Language Processing for Children's Speech [3].

The development of children's speech corpora for different languages is in progress [4] (British English, German and Swedish), [5] (non-native English), [6] Chinese, [7] Cantonese, [8] Jamaican English, [9] interactive emotional children speech and many others. For European union also small European languages are essential for electronic communication, so we decided to start the building of Slovak children speech corpus for improvement of the Slovak automatic speech recognition engines already built [10, 11]. Of course, the speech parameters different for children speech because of different vocal tract sizes [12], and they are many algorithms (Vocal-Tract Length Normalization - VTLN) how to handle it [13]. For children's speech, the formant frequencies are higher, the speech rate is slower or higher than in adult speech, and the language contains more home slang, garbled and imaginary words.

The Slovak language belongs to a group of Slavic languages, which are typical of inflection and free word order, which means it is morphologically rich and uses a very large vocabulary [10, 14]. These features make the Slovak automatic speech recognition task very complicated, and a large amount of data is required for automatic large vocabulary spontaneous speech recognition [14].

This article describes the first step, the collection of the first data, manual annotation, and testing of the current ASR system with children and adult speech recordings.

2. Building the database

For children's speech, there are very few freely available recordings on the Internet, especially in the form suitable for speech recognition system acoustic model training. We decided to use the TV series' recordings. There were several problems when using TV series recordings. The vast majority of segments are tinged with music, which would not matter if we were trying to build a model that would recognize where the sound begins and ends. However, when it comes to recognizing children's speech, it can cause distortions that will be undesirable for our purpose, and our results will be affected to some extent [15].

The main problem with a database suitable for acoustic models training is the resources needed for quality data annotation. This task is very time-consuming, and another reason is the lack of publicly available data, and therefore, our database is of a more modest size [15].

The database of children's recordings is made up of segments of children's speech from the television TV series of commercial Slovak broadcasters Markíza and JoJ. Specifically, they are the TV series Daddy (Oteckovia), broadcast on Markíza since early 2018, and Holidays (Prázdniny) from JoJ, the first part aired January 18, 2017.

The recordings were downloaded from premium archives of the TV broadcasters in Full HD. We cut the utterances with children speech out from the .MP4 recording and merged the parts without background music. The audio codec used in original file was AAC LC (Advanced Audio Coding - Low Complexity profile) with 48kHz 257.05 kbps Stereo settings.

Then the WAV file was exported in 48kHz Stereo PCM format and annotated with the Transcriber [16] application (Figure 1.). The collected database statistic is summarized in the Table 1.

Table 1. Database statistics

TV Series_Episode (Date)	Lenght [minutes]	Number of words
Oteckovia_E1(1.1.2018)	2:59	298
Oteckovia_E2(2.1.2018)	5:15	550
Oteckovia_E3(3.1.2018)	4:35	601
Oteckovia_E4(4.1.2018)	2:56	364
Oteckovia_E5(5.1.2018)	4:14	510
Oteckovia_E6(8.1.2018)	3:49	519
Oteckovia_E7(9.1.2018)	4:57	650
Prazdniny_E1(18.1.2017)	4:23	513
Prazdniny_E2(25.1.2017)	7:58	739
Total	**41:01**	**4744**

3. Transcription process

In our database, we have annotated the age, real names and surnames of publicly known children actors, so that we can see how the system performs with different ages of children (Figure 1.).

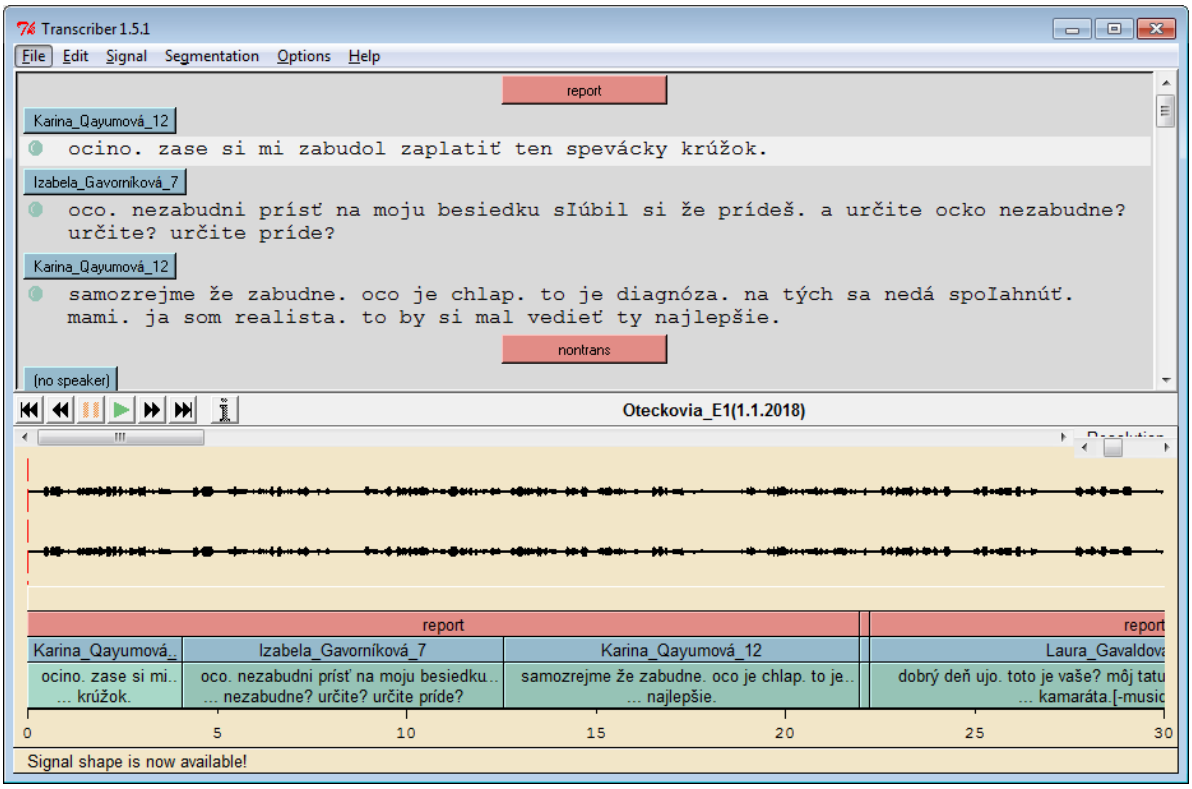


Figure 1. Transcription software used (Transcriber 1.5.1)

The gender, dialect, and the mother tongue (native or non-native speaker) were annotated for each speaker.

The mode field was set for speaker turns. We use the spontaneous option to indicate that the speech is spontaneous, unprepared speech or conversation. Mainly spontaneous speech was annotated for children. The planned speech is commonly used by studio moderators and sport news anchors. We follow the rules from standard broadcast news transcriptions [14] for the fidelity and the channel quality. Similarly, annotations mark the background noise and intermittent noise (see Figure 2.).

Annotations of the speaker turns follow the rule that one speaker turn should be no longer than 5 seconds. The capital letters at the beginning of the sentences were not used for easier named entity recognition.

The transcription process was made manually by the bachelor student and verified by his advisor. The plan is to extend the database following this proposed process in next year using more student annotators and expert verifications.

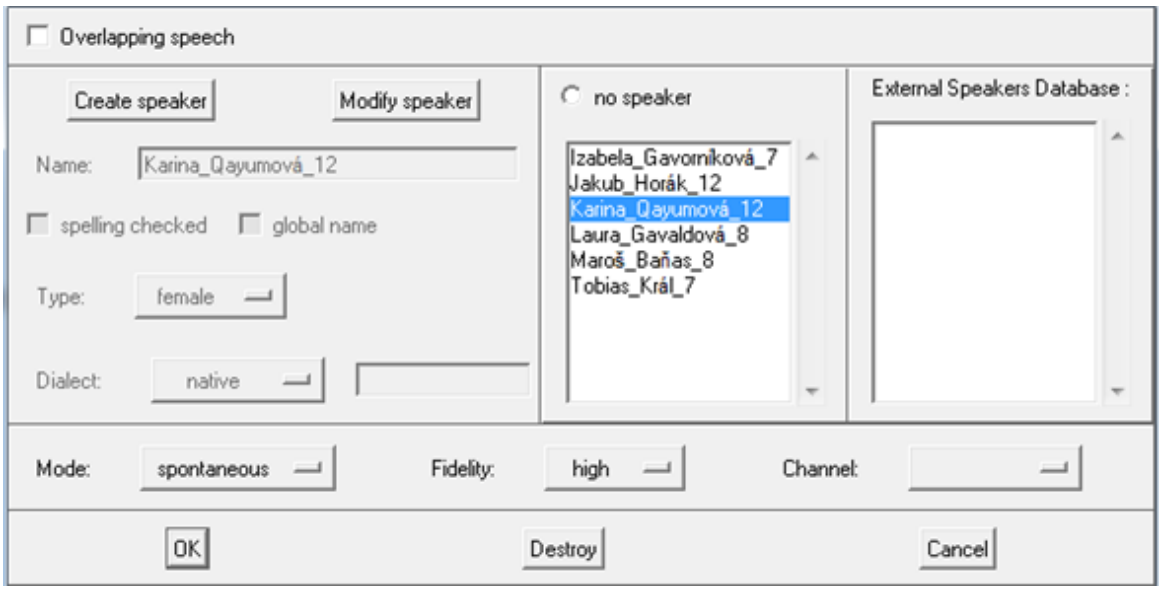


Figure 2. Transcriber window for speaker turn metadata.

4. Evaluation of the current subtitling system with children recordings

The current automatic subtitling system for Slovak TV broadcasters was developed thanks to many years of Slovak automatic speech recognition development of Technical University of Kosice and Slovak Academy of Sciences consortium. The previous system was based on Julius [10] mainly prepared for speech dictation into word processing editor. The next generation was built on Time Delay Deep Neural Network (TDNN) models based on Kaldi [11] for broadcast news transcription [17]. This version was also made online for public testing and evaluation on [18] as seen in Figure 3.

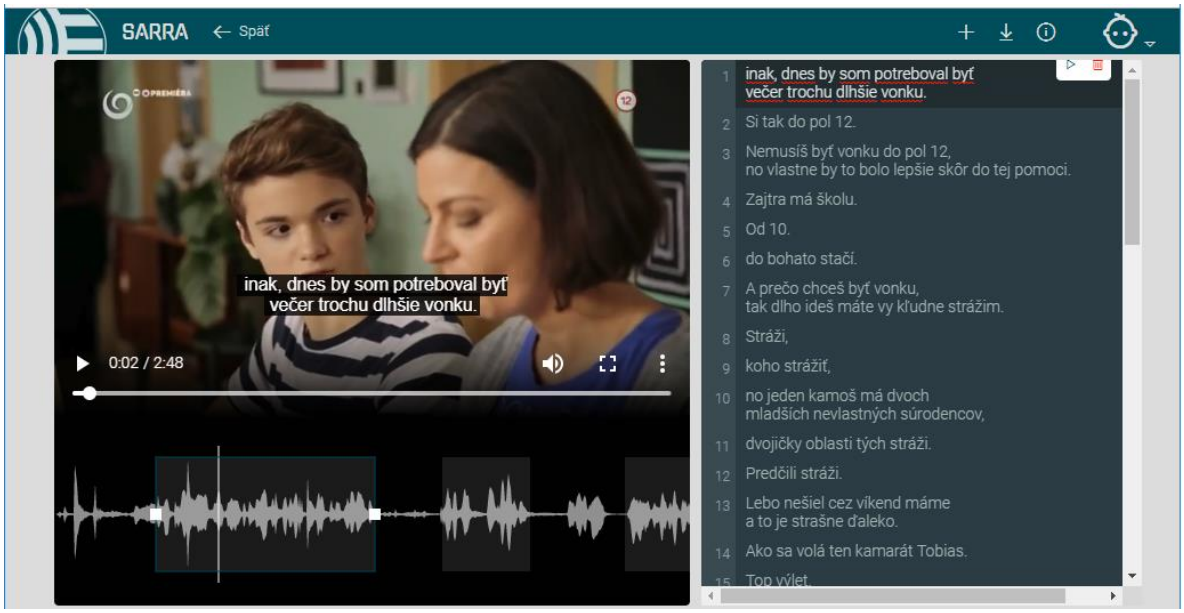


Figure 3. SARRA web user interface of automatically subtitled content

The SARRA system is built to work in multitasking and scaling environment, so the user's task could run on more instances of the recognition toolkit at once. The first part of the process is voice activity detection and speaker diarization for better segmentation of the large audio uploads. The smaller segmented parts of the audio could be scaled better.

The next part is the primary automatic speech recognition process built on models from about 600 hours of Slovak speech from broadcast news and TV discussions recordings. The acoustic model is using 40MFCC coefficients with online Cepstral mean normalization (CMN, in first training phase) and 100 dim i-vector. The language model was built from 1.89 billion token corpus with 500 thousand unique words vocabulary smoothed by the Witten-Bell algorithm [17].

The last part is the post-processing of the recognized text to convert it to the TV subtitle suitable form. The requirements were that the amount of text is limited on one subtitle caption and also the time for showing the caption should be long enough to read them by the viewers. Finally, it is expected to have the subtitles adapted to speaker changes where the diarization engine results are used.

The current engine could achieve 14.6% WER (Word Error Rate – the number of errors divided by the number of words in ground trough data) for broadcast news transcriptions where the variety of speakers and speech styles is wide [17]. For comparison, the dictation engine could achieve 3.93% WER for prepared Slovak dictation [10].

After the uploading and evaluation of the results from presented children Slovak speech database, we achieve only 47.81% WER, mainly because of 9.18% OOV (Out of vocabulary words) rate and very spontaneous speech segments.

5. Conclusions

The resources of children speech are scarce, even for major languages [7]. The presented database of children speech is the first one for the Slovak language and provides essential experiences about acoustic and mainly linguistic features of Slovak children speech. The development of adapted acoustic and language models for the Slovak automatic children speech recognition is in progress. There are several goals ahead, but mainly the extension of the presented dataset is planned for next year using more undergraduate students and expert verifications of the transcriptions. The goal is to present a special version of the SARRA models [17, 18] for children speech and evaluation by real users also for dictation and Human-Robot interaction purposes [19] based on the running international collaboration and projects.

Acknowledgments

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and finally by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 108-2911-I-027-501, 107-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067.

References

[1] Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A.: A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, ACM, p. 7, 2009.

[2] ISCA Special Interest Group: Child Computer Interaction (CHILD). [Online]. Available: https://www.isca-speech.org/iscaweb/index.php/sigs?id=129 [Accessed: July 30, 2019].

[3] Spoken Language Processing for Children's Speech, Interspeech 2019 Special session proposal. [Online]. Available: https://sites.google.com/view/wocci/home/interspeech-2019-special-session. [Accessed: July 30, 2019].

[4] Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., & Wong, M.: The PF_STAR children's speech corpus. In *Ninth European Conference on Speech Communication and Technology – INTERSPEECH 2005*. pp. 3761-3764, 2005.

[5] Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., Price, P., Anderson, E., Narayanan, S., & Alwan, A.: TBALL data collection: the making of a young children's speech corpus. In *Ninth European Conference on Speech Communication and Technology – INTERSPEECH 2005*. pp. 1581-1584, 2005.

[6] Xiangjun, D., & Yip, V.: A multimedia corpus of child Mandarin: The Tong corpus. *Journal of Chinese Linguistics*, 46(1), pp. 69-92, 2018.

[7] Wang, J., Ng, S. I., Tao, D., Ng, W. Y., & Lee, T.: A study on acoustic modeling for child speech based on multi-task learning. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP).* Taipei, IEEE, pp. 389-393, 2018.

[8] Watson, S., & Coy, A.: JAMLIT: A Corpus of Jamaican Standard English for Automatic Speech Recognition of Children's Speech. In *SLTU*. pp. 243-247, 2018.

[9] Pérez-Espinosa, H., Martínez-Miranda, J., Espinosa-Curiel, I., Rodríguez-Jacobo, J., Villaseñor-Pineda, L., & Avila-George, H.: IESC-Child: An Interactive Emotional Children’s Speech Corpus. *Computer Speech & Language*, *59*, pp. 55-74, 2020.

[10] Rusko, M. et al.: Advances in the Slovak Judicial Domain Dictation System. In: Vetulani

Z., Uszkoreit H., Kubis M. (eds) *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2013 - Revised selected papers*. Lecture Notes in Computer Science, vol 9561. Springer, Cham, pp. 55-67, 2016.

[11] Staš, J. et al.: Automatic subtitling system for transcription, archiving and indexing of Slovak audiovisual recordings. In *Proceedings of the 7th Language & Technology Conference, LTC 2015.* pp. 186-191, 2015.

[12] Lee, S., Potamianos, A., and Narayanan, S.: Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp.1455–1468, 1999.

[13] Shivakumar, P. G., Potamianos, A., Lee, S., and Narayanan, S.: Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In: Proc. *Workshop on Child, Computer and Interaction (WOCCI)*, pp. 15-19, 2014.

[14] Pleva, M. and Juhár, J.: TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation. In: *LREC*, Reykjavik, pp. 1709-1713, 2014.

[15] Fehér, M.: Automatic speech recognition for children. Bachelor thesis, Technical University of Kosice, p. 39, 2019.

[16] Transcriber - a tool for segmenting, labeling and transcribing speech. [Online]. Available: http://trans.sourceforge.net/en/presentation.php [Accessed: July 30, 2019].

[17] Lojka M., Viszlay P., Staš J., Hládek D., Juhár J.: Slovak Broadcast News Speech Recognition and Transcription System. In: Barolli L., Kryvinska N., Enokido T., Takizawa M. (eds) *Advances in Network-Based Information Systems*. NBiS 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 22. Springer, Cham, pp 385-394, 2019.

[18] SARRA - the automatic subtitling system for transcription, archiving, and indexing of Slovak audiovisual recordings. [Online]. Available: https://marhula.fei.tuke.sk/sarra/ [Accessed: July 30, 2019].

[19] Pleva, M., Juhar, J., Ondas, S., Hudson, C. R., Bethel, C. L., & Carruth, D. W.: Novice User Experiences with a Voice-Enabled Human-Robot Interaction Tool. In *2019 29th International Conference Radioelektronika*, Pardubice. IEEE, pp. 1-5, 2019.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 334-342

基於階層式編碼架構之文本可讀性預測

A Hierarchical Encoding Framework for Text Readability Prediction

翁詩諺　Shi-Yan Weng
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
40547041S@ntnu.edu.tw

曾厚強　Hou-Chiang Tseng
國立臺灣師範大學資訊工程學系 / 心理與教育測驗研究發展中心
Department of Computer Science and Information Engineering/ Research Center for Psychological and Educational Testing
National Taiwan Normal University
ouartz99@gmail.com

宋曜廷　Yao-Ting Sung
國立臺灣師範大學教育心理與輔導學系
Department of Educational Psychology and Counseling
National Taiwan Normal University
sungtc@ntnu.edu.tw

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

摘要

以教育的角度來看，為了幫助學生獲得更好的學習效果，對每個年級安排適當的難度文本是非常重要的。因此，長久以來，陸續有許多學術機構致力於可讀性模型或特徵的研究。為了解決這個問題，在先前的研究常使用一些人為定義的特徵，例如難詞頻率或字數等特徵來對文本進行可讀性難易程度預測。然而，這些特徵可能太淺層而不能表示文本的語法、語意或更深層的內涵。近期，由於深度學習或表示學習技術的蓬勃發展，使得更有代表性的語言特徵能從文本中被萃取出來增進可讀性難易程度的準確性。延伸此技術發展趨勢，在本論文中我們設計並實作出具有階層式編碼的類神經網路來做為可讀性預測模型，以擷取在文本中的詞彙到語句、語句到文本的語意和結構表示資訊。此外，

我們並嘗試在此模型中額外加入傳統的人為定義特徵作為輔助資訊。從實驗結果可以發現，我們提出的可讀性預測模型具有良好的效能表現，而加入傳統的人為定義特徵，亦可以進一步增進其預測的準確性。

Abstract

From an educational perspective, it is important to provide students of different grades with reading material of appropriate difficulty for better learning retention. To deal with this problem, it is common practice to use a set of handcrafted features, for example, hard word rate or word count, to distinguish articles into different readability levels. However, these traditional readability features are often too shallow to represent deeper semantic or syntactic structures of the articles. In view of this, we present a modeling approach that leverages a recurrent neural network to hierarchically encode both the semantic or syntactic structures of a given article for better readability classification. Furthermore, we also seek to make extra use of traditional handcrafted feature as side information to further boost the performance.

關鍵詞：可讀性、語言特徵、表示學習法、卷積式類神經網路、遞迴式類神經網路

Keywords: Readability、Language Feature、Representation Learning、Convolutional Neural Network、Recurrent Neural Network

一、緒論

可讀性(Readability)是指閱讀材料能夠被讀者所理解的程度[1],[2],[3],[4]；當讀者閱讀較高可讀性的文本時，會產生較好的理解及學後保留效果[2],[3]。西方的可讀性公式發展的非常早[5],[6]，據 Chall 與 Dale[7]在 1995 年的統計，到 1980 年為止相關的可讀性公式就已經超過 200 多種。這些傳統的可讀性研究大多使用較淺層的語言特徵來發展線性的可讀性公式。例如著名的 Flesch Reading Ease 公式以詞彙音節數做為語意的指標、以語句的長度作為語法的指標，透過計算詞彙的平均音節數與文本的平均語句長度來評估文本的可讀性難易程度：當文本的詞彙音節數愈多、語句愈長時，則該文本在閱讀理解時愈為困難。然而，傳統可讀性公式所採用的淺層語言特徵，並不足以反映文本閱讀理解的難度。Graesser、Singer 和 Trabasso[8]便指出，傳統可讀性公式無法反映閱讀的真實歷程，並沒有考量文本的深層特性，例如語句的前後順序。Collins-Thompson[9]亦指出傳統可讀性公式僅著重在使用文本的表淺資訊，而忽略文本其它的重要特徵，例如文

本結構資訊或 *N*-連語言樣式(*N*-gram Patterns)。這也讓傳統可讀性公式在預測文本可讀性的結果常遭受到質疑。甚至因為可讀性公式所採用的可讀性特徵過少，導致容易受到有心人士為了達到特定的可讀性數值，而刻意針對可讀性特徵的特性來修改文本，使得文本呈現許多簡短而破碎的句子，反而降低了文本的流暢度與連貫性，因此增加閱讀難度[10]。直至今日，可讀性的研究仍持續不斷。研究人員為了克服傳統可讀性公式的缺點，嘗試利用較複雜的機器學習演算法來發展出更細緻、非線性的可讀性模型；同時，亦納入更多元的可讀性指標來共同評量文本的可讀性，除了可以提升可讀性模型的效能，亦可防止有心人士去操弄文本的可讀性[11],[12],[13]。

而在近年，由於深度學習(或表示學習)在自然語言處理領域的蓬勃發展，讓我們能夠藉由相關技術從文本中萃取出更深層的語意或結構特徵。例如，在自動文件摘要的研究上，有所謂的序列對序列(Sequence-to-Sequence, Seq2Seq)模型架構提出，透過在不同階段使用卷積式類神經網路(Convolutional Neural Network, CNN)與遞迴式類神經網路(Recurrent Neural Network, RNN)來達成語句或文本(文本)的固定長度語意向量表示[14]，並藉此產生出一份較為簡短扼要的摘要來代表原始文件。而在[15]中，研究人員採用了更複雜的序列對序列模型並加入了強化學習(Reinforcement Learning)使得自動摘要的表現更好。

基於上述的研究與相關技術發展，在本論文中我們設計並實作出具有階層式編碼的類神經網路來做為可讀性預測模型。首先，透過以卷積式類神經網路(CNN)以擷取每一句語句內部的局部詞彙使用特徵；接著，經由遞迴式類神經網路(RNN)依照語句向前或向後讀取卷積式類神經網路所產生之語句的語意向量表示，來產生同時含有文本語句結構資訊之文本的語意向量表示，最後用於於文本可讀性難易程度預測。同時，我們並嘗試在此模型中額外加入傳統的人為定義特徵作為輔助資訊以期能更進一步提升可讀性預測模型的效能表現。本論文接下來的安排如下：第二節描述我們提出的文本階層式編碼模型架構；第三節將呈現實驗材料及相關的實驗設定及結果；最後第四節是總結及未來研究的方向。

二、階層式編碼架構

基於階層式編碼模型架構之文本可讀性預測模型是如圖一所意示，它主要可分為兩個部分：首先為結合卷積式類神經網路(CNN)和遞迴式類神經網路(RNN)所組成的文本編碼

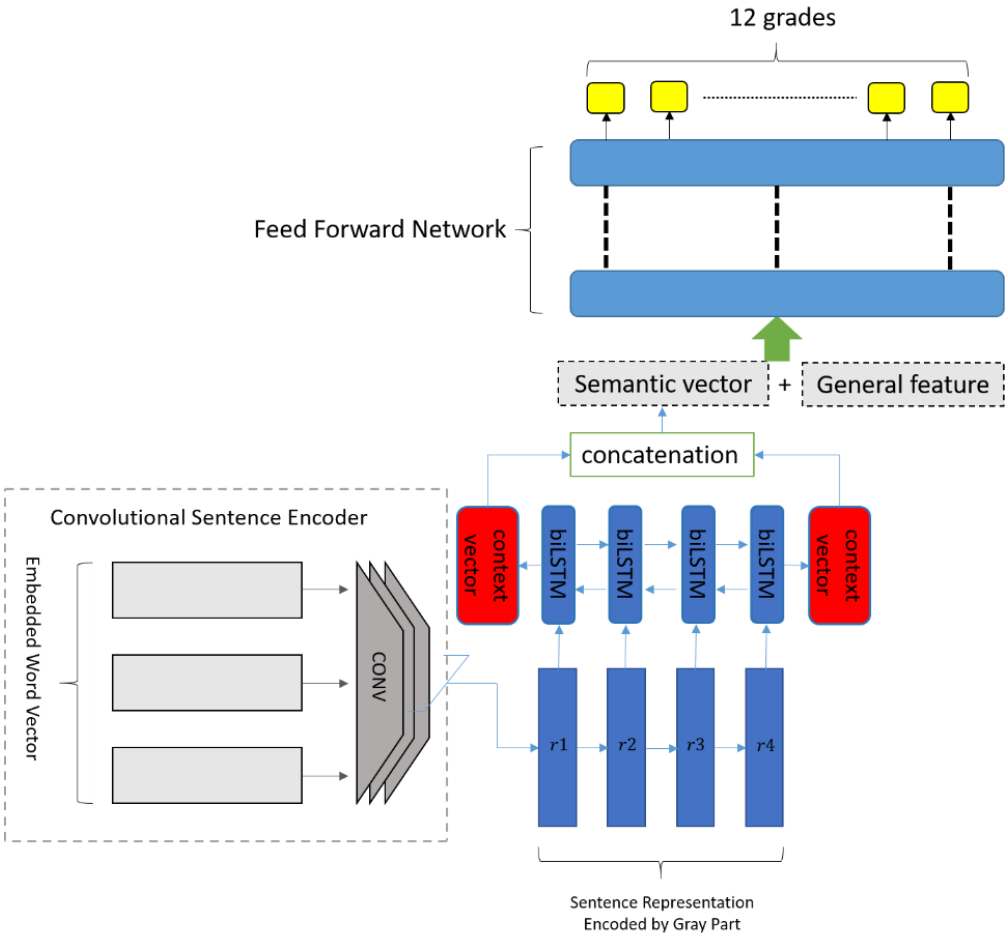
(Encoding)結構，此部分主要的目標在於得到對於每一篇文本的高階特徵；此結構的優點在於融合卷積式類神經網路與遞迴式類神經網路的長處: 卷積神經網路能夠捕捉語句內局部的 *N*-連語言樣式(*N*-gram Patterns)，而遞迴式類神經網路則可以捕捉長距離的語句與語句之間的語意和結構關係。透過上述依序結合使用的方式，可以彌補單純將文本視為一個詞彙序列(忽略語句間結構資訊)並僅單獨使用卷積式類神經網路或遞迴式類神經網路來產生文本固定長度向量表示的不足處。更詳細地說，在實作時輸入文本中的每一句語所含的詞彙會先經過詞嵌入(Word Embedding)轉換方法，例如 Skip-Gram 和 CBOW，轉換成固定長度的詞彙語意向量[16]。再者，每一句語句中詞彙的語意向量會經過含三層卷積式類神經網路，每一層含有不同數量和大小(由大到小)的核函數(Kernel Function)以萃取不同階段的語句內部的語意向量表示。

語句會經過三層核函數由大到小的卷積式類神經網路以取得對每個句子的向量表達(Sentence Representation)，再將其送入雙向長短期記憶網路(Bi-directional LSTM)，雙向長短期記憶網路能夠達成捕捉每一句對其他句之間的關係的目標，最後得出兩個方向相反的內文向量，將其串接之後，送入最後的前饋網路(Feed Forward Network)做出最後的預測。

在文本編碼及前饋網路之間，會考慮是否加入文本的一般特徵(表一)，此些一般特徵由考慮文本的不同面向統計而來。一般特徵會在得到對整個文本的特徵表達之後與之串接，得到一個包含同時具有文本語句結構資訊及一般統計特徵之文本的語意向量，以期作出最後的文本可讀性難度預測。

表一、文本之一般特徵共 15 個

	特徵名稱
詞相關特徵	詞數、動詞數、領域詞頻對數平均、負向連接詞數、中筆劃字元數、副詞數、實詞數、低筆劃字元數、二字詞數、複雜語意類別數、正向連接詞數、難詞數
句相關特徵	單句數比率、複雜結構句數、複雜語意類別句子數



圖一、階層式編碼之文本可讀性預測模型架構圖

三、實驗設定與結果

(一)、實驗材料

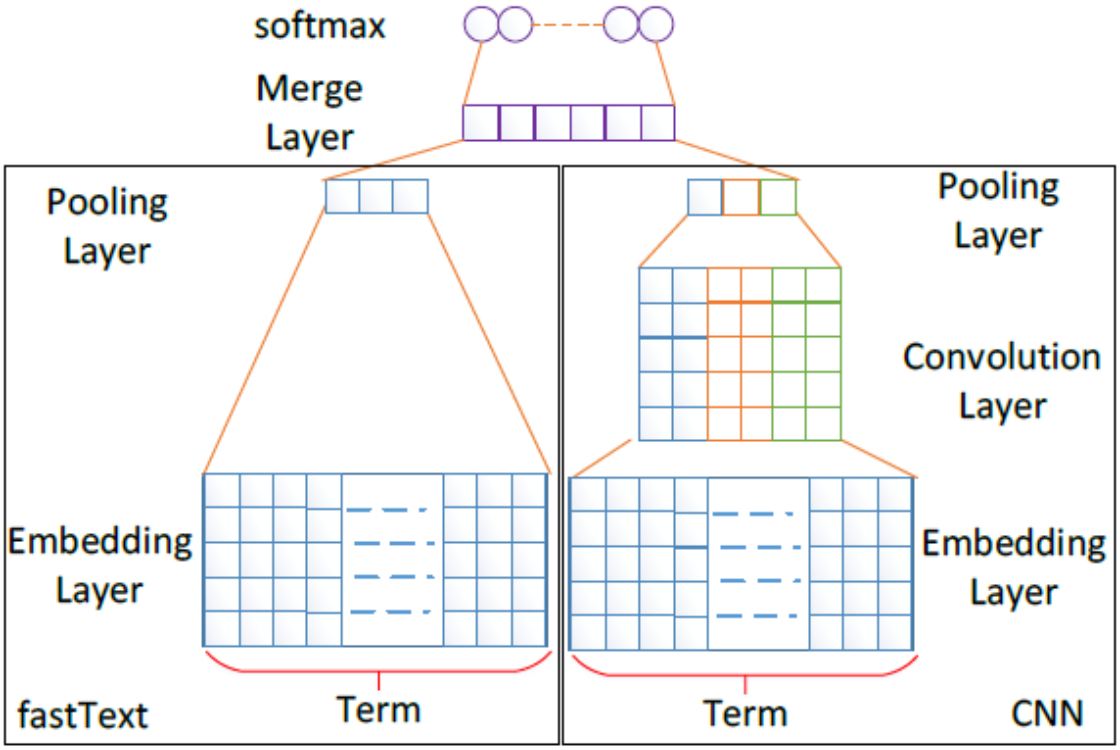
本研究材料選自 98 年度臺灣 H、K、N 三大出版社所出版的 1-12 年級審定版的國語科、社會科和自然科等三個領域的教科書全部共計 4,648 篇[17]，各版本教科書均經由專家根據課程綱要編制而成，其實驗材料的年級分佈如表一所示，模型的目標則是預測輸入屬於 12 個年級中何者。

表一、實驗材料在各年級的數量分佈

年級	1	2	3	4	5	6	7	8	9	10	11	12
國文科	24	67	61	71	69	70	37	34	28	84	41	47
自然科	0	0	72	67	67	62	172	175	157	211	355	295
社會科	0	0	80	74	85	81	389	407	325	340	331	270

(二)、實驗結果

我們的結果顯示在表二中。在表二中，我們將結果與[17]進行比較。在[17]中提出了一種結合了快速文本(fastText)和 CNN 的混合架構。[17]中提到，若在訓練可讀性模型的過程中可以同時融合快速文本(fastText)和 CNN 這兩種演算法，其訓練可讀性模型的過程中就可以相互分享資訊，為可讀性模型帶來更豐富的資訊去評估文本的可讀性。基於這個想法，快速文本(fastText) [18]用作學習可用於對語義進行分類的功能的角色。利用類神經網路來融合卷積神經網路及快速文本兩種不同的表示學習演算法，其示意圖如圖二所示，在[17]的可讀性模型的訓練過程中，卷積式類神經網路和快速文本所產生的特徵會以向量的形式在融合層進行相加、相乘、平均或串聯等不同的運算方式進行融合，而融合後的特徵便可以讓可讀性模型在訓練的過程中享有不同表徵學習法所帶來資訊。



圖二、融合卷積神網路及快速文本的可讀性模型架構

在表二中，可以觀察到我們的模型表現在沒有加上一般特徵的情況下表現略輸於[17]約 0.5 個百分點，但在加入一般特徵輔助判斷後我們的模型表現在準確率進步了 1.91%，相鄰準確率進步了 2.71%，表現都略優於[17]，可見的一般特徵還是有提供資訊

增進模型判斷的能力，相鄰準確率則是考慮了正確答案及前後一個年級:如果正確答案為 5 年級，則可以容忍模型的預測為 4、5、6 年級。

表二、實驗結果

Readability model	準確率	相鄰準確率
Tseng et al.,(2018)	79.42%	91.59%
Hierarchical Encoding w/o general feature	78.81%	89.64%
Hierarchical Encoding w/ general feature	**80.72%**	**92.35%**

我們還在前饋網路(feed-forward network)中比較不同數量的層數(6,7,8,9)對模型的影響。在表 3 中，我們可以看到，在前饋網路中如果層數更多，我們可以略微提高準確性。但太多層沒有對模型表現有更多的增進，可能還會有過度擬合的疑慮。

表三、前饋網路不同層數之結果

層數	準確率	相鄰準確率
6	79.12%	91.65%
7	79.82%	91.24%
8	**80.72%**	**92.35%**
9	80.04%	91.86%

四、結論

本研究結合卷積神經網路及遞歸神經網路並兼取兩者之長處，訓練出一個能夠抽取更深層特徵的可讀性模型。實驗結果顯示，此模型不考慮文本的一般特徵就已經可以達到不錯的模型效能，但若在模型中考慮一般特徵則能夠在得到些微的提升。在未來，我們希望能夠不只用遞歸神經網路來取得句與句之間的關係，而是更精細的計算句子間的相關性，並且探討相關性是否能可讀性研究有所幫助，並考慮使用更細緻的方法來結合文本的各種特徵，使的模型能夠更全面的評估文件的可讀性難易程度。

致謝

本論文之研究承蒙行政院科技部研究計畫 (MOST 105-2221-E-003-018-MY3 和 MOST 107-2221-E-003-013-MY2、MOST 108-2221-E-003-005-MY3 和 MOST 108-2634-F-008 -004 -) 之經費支持，謹此致謝。

參考文獻

[1] E. Dale and J. S. Chall, “The concept of readability,” Elementary English, vol. 26, pp. 19–26, 1949.

[2] G. R. Klare, “Measurement of Readability,” 1963.

[3] G. R. Klare, “The measurement of readability: useful information for communicators,” ACM Journal of Computer Documentation (JCD), vol. 24, pp. 107-121, 2000.

[4] G. H. McLaughlin, “SMOG grading: A new readability formula,” Journal of reading, vol. 12, pp. 639–646, 1969.

[5] B. A. Lively and S. L. Pressey, “A method for measuring the vocabulary burden of textbooks,” Educational administration and supervision, vol. 9, pp. 389–398, 1923.

[6] M. Vogel and C. Washburne, “An objective method of determining grade placement of children's reading material,” The Elementary School Journal, pp. 373–381, 1928.

[7] J. S. Chall and E. Dale, Readability Revisited: The new Dale-Chall Readability Formula, Brookline Books, 1995.

[8] A. C. Graesser, M. Singer, and T. Trabasso, “Constructing inferences during narrative text comprehension,” Psychological Review, vol. 101, pp. 371, 1994.

[9] K. Collins-Thompson, “Computational assessment of text readability: A survey of current and future research,” International Journal of Applied Linguistics, vol. 165, pp. 97–135, 2014.

[10] B. Bruce, A. Rubin, and K. Starr, “Why readability formulas fail,” IEEE Transactions on Professional Communication, pp. 50-52, 1981.

[11] S. E. Petersen and M. Ostendorf, “A machine learning approach to reading level assessment,” Computer Speech & Language, vol. 23, pp. 89–106, 2009.

[12] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, “A comparison of features for automatic readability assessment,” in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 276–284.

[13] Y. T. Sung, J. L. Chen, J. H. Cha, H. C. Tseng, T. H. Chang, and K. E. Chang, “Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning,” Behavior research methods, vol. 47, pp. 340 – 354, 2014.

[14] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, “Sequence to Sequence Learning with Neural Networks”, in Proc. NIPS, Montreal, CA , 2014

[15] Yen-Chun Chen, Mohit Bansal, “Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting”, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(ACL’2018), Pages 675-686

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[17] Hou-Chiang Tseng, Berlin Chen, Yao-Ting Sun, “Exploring Combination of FastText and Convolutional Neural Networks for Building Readability Models”, in Proceedings of the the 2018 Conference on Computational Linguistics and Speech Processing, Pages 116-125

[18] A.Joulin, E.Grave, P.Bojanowski, and T.Mikolov, “Bag of tricks for efficient text classification,”arXiv preprint arXiv:1607.01759. 2016

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 343-357

國語語音辨識系統中之人名語言模型

The Personal Name Modeling in Mandarin ASR System

梁鴻彬　Hong-Bin Liang
國立交通大學電機工程學系
Department of Electrical Engineering
National Chiao Tung University
hbliang@speech.cm.nctu.edu.tw

王逸如　Yih-Ru Wang
國立交通大學電機工程學系
Department of Electrical Engineering
National Chiao Tung University
yrwang@speech.cm.nctu.edu.tw

摘要

本論文主要有兩個目的：一是訓練一個高效能的中文語音辨識系統；二是改善因人名而造成的 OOV(Out-Of-Vocabulary)問題，並將其辨認出來，以便日後自動轉寫不同類型的語音訊息並產生逐字稿。而人名之辨識對於將來自然語言處理也是一重要的訓練資料。

本論文使用 Kaldi speech recognition toolkit 的環境為基礎，在聲學模型的方面，本實驗使用類神經網路 TDNN 以達到聲音資訊轉成音素序列(phone sequence)的目的；在語言模型方面，本論文透過加入中文特有的語言資訊如形音義詞的合併、專有名詞的拆解，並使用 n-gram 語言模型的訓練，以達到音素序列轉成詞序列(word sequence)的目的，並於解碼過程中調整參數與權重，找出最佳操作點，以得到即時性與辨識率兼顧的語音辨識系統，此外，針對以往人名無法辨認出來的問題，本論文建立特別的人名語言模型以類似 class-based model 的方式置換原 word-based model 中的人名，以達到辨識人名的目的。

Abstract

There are two purposes in the paper, one is training an efficient ASR system, the other is improving the OOV problem caused by the personal name, and we want to recognize it for the purpose of making transcription of different kind of speech data. Name recognition data is also an important training data for the NLP.

The paper base on the environment of Kaldi speech recognition toolkit. In the acoustic model part, we use many different kind of neural network such as TDNN to transform the speech information into phone sequence. In the language part, we add Chinese special

language information such as variant word combination and name entity decomposition, using n-gram language model and lattice rescoring to transform the phone sequence into word sequence. We also tune the parameters and weights during the decoding process to get the best operation point to obtain a ASR system which is not only good at recognition rate but also efficient at recognition time. Moreover, we focus on the problem of difficulty in personal name recognition. We build a class-based like model to replace the original word-based model of personal name to reach the goal of personal name recognition.

關鍵詞：聲學模型、語言模型、中文大辭彙連續語音辨識、時延神經網路(Time Delay Neural Network, TDNN)、專有名詞辨識(Named Entity Recognition, NER)。

Keywords: acoustic model, language model, Mandarin large vocabulary continuous speech recognition, TDNNs, named entity recognition.

一、 緒論

(一)研究動機

隨著訓練語料的增加，電腦運算速度的大幅提升，再加上越來越多製作語音辨識器的工具庫的發明，如 HTK Speech Recognition Toolkit[1], Kaldi ASR[2], et al. 訓練一可用甚至是商用級別的語音辨識已經是唾手可得。

但是，辨識器普遍存在著對於專有名詞辨識率不佳的問題，由台北科技大學廖元甫老師提供之比賽[3]結果，比較國內多所學校與企業所訓練之辨識器，整體辨識率之正確率普遍可高達 90%做右(字錯誤率 10%左右)，但是仔細評估辨認錯誤之詞彙後發現，各辨識器對於專有名詞之辨認率相對較差，尤其在專有名詞中最為重要且為數眾多的人名辨識之正確率僅不到 50%(如圖 一)，也就是說，只要辨識器一遇到人名，辨識率通常不會太高。然而，人名對語言來說又是一個非常重要的資訊，例如在日常生活中，常用的句型裡不外乎由人事時地物所組成，而這裡的「人」，指的就是人名，再者，若辨識器將人名全部當成 OOV 處理，不僅人名這個詞彙辨識不出來以外，也會影響包含人名的句子的辨識能力，因此，本實驗將致力於提升語音辨識器對人名之辨識率，藉由對語言模型之訓練語料加入人名資訊，以期望提高辨識器對於人名之辨識率。

[1] http://htk.eng.cam.ac.uk/
[2] http://kaldi-asr.org/
[3] 指 ISCSLP 2018 Formosa Speech Recognition 會外賽。

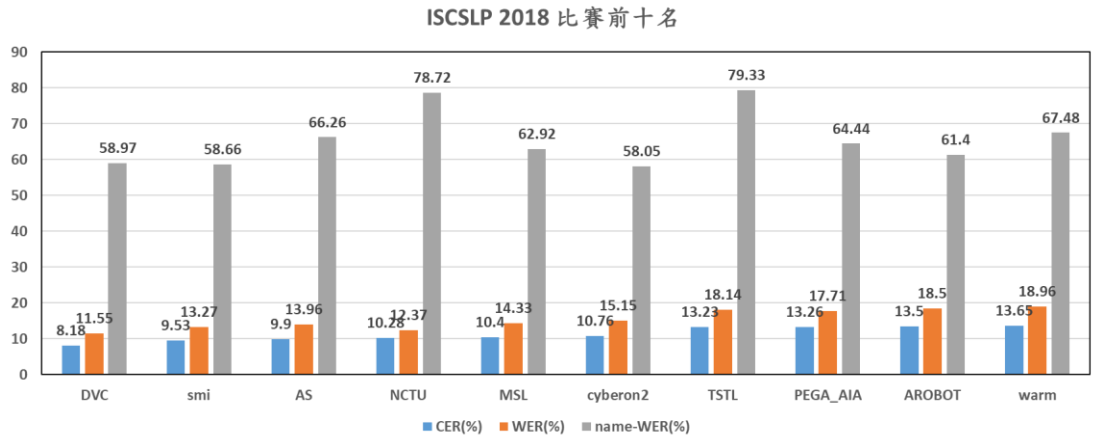


圖 一、比賽前十名之成績（圖例由左至右分別為字錯誤率(CER)、詞錯誤率(WER)與人名詞彙錯誤率(name-WER)）

(二)研究方向

本研究將致力於提升語音辨識器之辨識率與辨識速度，並藉由訓練資料的改善，使之得以辨識以往所不能辨識之人名。

本篇論文首先透過調整聲學模型架構、辭典大小、語言模型參數，使得辨識器在整個語音辨認過程上取得辨識率與辨識速度的平衡。在得到一個辨識率與辨識速度兼具的辨識器之後，將針對辭彙不在辭典裡之問題(Out of Vocabulary, OOV)做改善。

由於人名為 OOV 組成中較為重要、且出現頻率最高的辭彙，本實驗將針對 OOV 裡的人名做特別處理，首先，先建立一 word-based Language Model，並找出所有人名(Word)，將人名拆解成二到三個小單元(Sub word unit)，因本實驗僅要求辨識出之人名的發音正確，部分未被選入辭典之人名將會被轉成音節(Syllable)，以音節的方式加入進辭典裡，之後再對文字語料後處理，將人名資訊加進訓練語料，利用統計特性計算各種不同可能人名互相連接的機率，以產生不同人名之組合，此方法可以讓辨識器對人名的辨識不再侷限於辭典裡出現過的人名，使得語言模型獲得辨識人名的能力。

二、 語料庫介紹

本節將介紹用於本實驗中之所有語料庫。其中用來當作訓練聲學模型之語料的有 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乾淨語料(Clean)以及背景有人為雜訊或音樂參雜其中之其他語料(Other)。

(一)TCC300 語料庫

本次實驗中所使用的 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。其中訓練語料約為 24.4 小時，304780 個音節數；測試語料約為 2.4 小時，26357 個音節數。

(二)NER 語料庫

NER 語料庫 [2]，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，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。其中訓練語料約為 111.5 小時，1715091 個音節數；測試語料可分為乾淨語料(Clean)的 1.9 小時，33660 個音節數與其他語料(Other)的 9.0 個小時，133746 個音節數。

(三)AIShell

AIShell 語料庫 [3]，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，此語料庫文本經人工校正過，正確率為 95%以上。其中訓練語料約為 162.4 小時，1862171 個音節數；測試語料：約為 16.6 小時，178041 個音節數。

三、 深層類神經網路模型配置

由於 CLDNN 網路結構中之 LSTM 層的原因，導致模型之解碼速度大幅的下降，有鑑於此，有學者提出了一種 Resnet-like TDNN-F 模型 [4]，以加深網路層數的方式增強模型之學習能力，且在層與層之間加上一層 Bottleneck layer 降低參數量，以解決參數隨著層數增加而暴增之問題，並加上 Skip connection，跨接上一層所訓練之參數。

此模型亦使用 TCC300、NER、AISHELL 作為訓練語料，特徵參數的抽取為 40 維之 Fbank，並前後串接 2 個音框(2-1-2) ，形成 200 維特徵向量，後經過一個 LDA 矩陣，做為 TDNN-F 之輸入特徵向量，每一層隱藏層神經元個數為 1536，Bottleneck layer 之神經元個數為 160，Bypass-scale 為 0.66，輸出的觀測狀態數目為 2672，總共為 15 層。架構如圖 二。

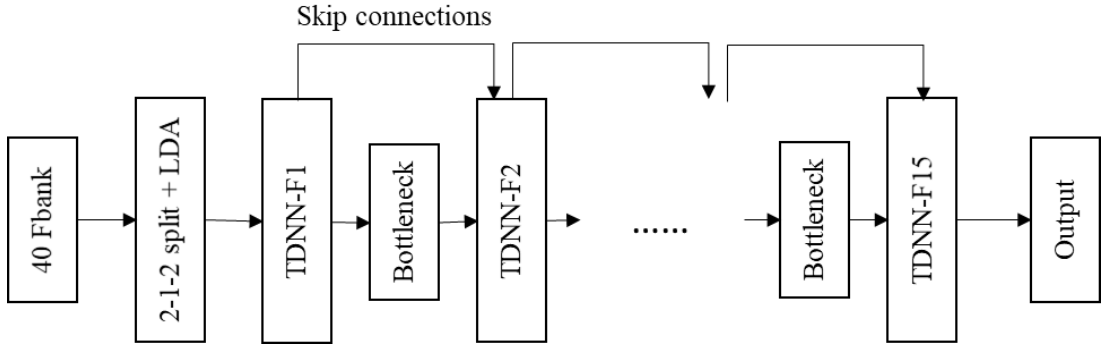


圖 二、TDNN-F 模型架構圖

四、 加入人名語言模型之語音辨識器

在解碼的過程中，若遇到辭典裡沒有出現的詞(Out of Vocabulary, OOV)，則此一辭彙將永遠不可能辨認正確，此時，辨識器會以一相似讀音的詞取代之，使得此一 OOV 附近的辭彙將受到影響，也就是形成所謂的搶詞，進而導致錯誤率的提高[4]。因此，OOV 對於語音辨識來說，一直是個非常重要的問題。在 OOV 組成中，尤以專有名詞(Nb)的出現為最大宗，又專有名詞中，人名為較具意義且出現頻率最高之種類，因此，本實驗將針對人名之辨識率做改善。

為了解決人名被當作 OOV 的問題，本論文將人名從字轉音成音節(G2P)，使原訓練語料中人名的位置是以音節的方式存在，之後再以製造人名隨機填入的方式，使語言模型在原人名位置能看到許多不同種類之人名。詳細內容將在接下來的小節一一解說。

(一)文字語料簡介

本研究用於訓練與研模型之文字語料庫共約 25 億個詞彙，包含以下：

- 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。
- 其他由交大語音處理實驗室蒐集之語料

(二)文字語料後處理中之形音義分合詞(variant word)處理

由於訓練語料多來自於新聞語料，為了使訓練資料更接近一般人說話，本實驗將斷詞後的文字做一些特別的處理，使其能更接近口語。例如將文章中出現的科學符號、度量衡等等轉為中文，賦予其統一的文字表示方式，使其被選入辭典後，被語言模型訓練到。後處理中，最重要之步驟如同義詞(variant word)之代換合併。中文之同義詞百百種，能表達同樣意思之詞語的選擇見仁見智，例如同義異音之周一、星期一；同義同音之周一、週一，當人們在使用這些同義詞構句時，這些同義詞之前後連接詞彙幾乎一模一樣，因此，若未合併這些詞彙，統計語言模型將會視這些詞彙為個別不同的詞彙，這將影響辭典收納詞彙之能力，且分散詞句在統計語言模型所統計之機率。依照發音之異同，大致上可以分為兩類，即發音相同與發音相異之詞類，置換原則是建立在字義相同之上，置換的目的一是為了讓文章中同義詞正規化，以利選詞時容納更多詞彙，二是使得合併訓練後的詞可以獲得更多的訓練資料，但是在語言模型建立後，辨識端會產生一個狀況：同義異音之詞彙在合併後將會導致某些發音在解碼圖路徑上消失，導致解碼時無法被搜尋到，如範例中的「禮拜一」被置換成「週一」，故在語言模型中無法找到「禮拜

[4] 這裡指的 OOV 所造成之辨識錯誤率計算方式為 OOV rate 乘以平均詞長。

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」。如表 一。

表 一、同義詞修改範例

形音義分合詞類型	置換前文字	置換後文字
發音相同	周一	週一
發音相異	禮拜一、星期一	週一

本實驗使用實驗室長期累積蒐集之 3160 個同義詞代換，在 7 個不同之語料集上做混淆度(Perplexity, ppl)測試，此實驗使用之語言模型為 4-gram 語言模型，平滑化方法使用 Witten-Bell smoothing[5]，訓練語料為 25 億詞，辭典大小為 120k，差別僅在於一個是使用置換過 variant word 的辭典，另一個是使用未置換 variant word 的辭典。

比較有合併同義詞、與沒有合併同義詞之辭典所建立之語言模型可以發現，有合併同義詞之語言模型的 ppl 較小(見圖 三)，也就是說，平均每搜尋一個詞，所需要使用的辭典數大小較小，表示新的語言模型在詞組搜尋上之效能較佳。由此可見當合併同義詞之後，可以使得語言模型在搜尋詞組上會更有效率。

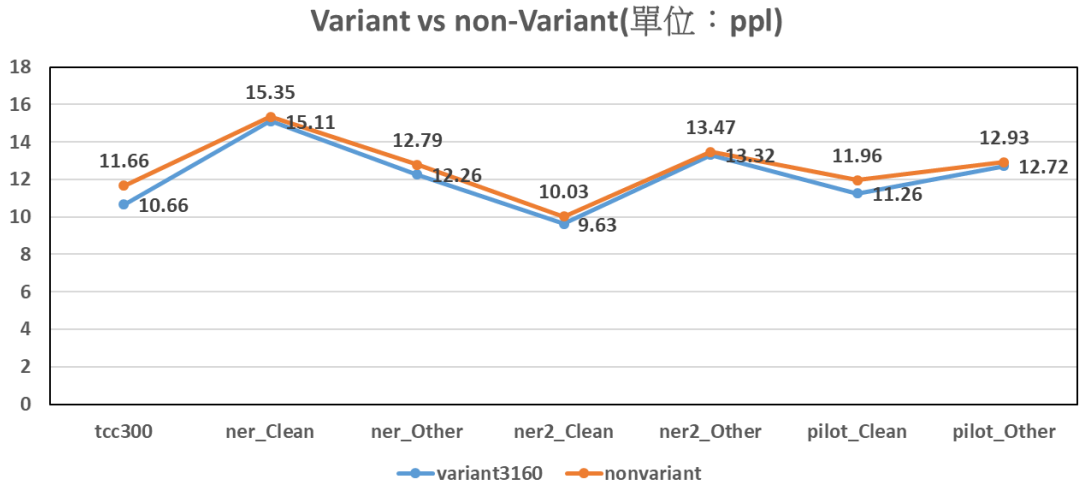


圖 三、辭典使用 120k 且置換 3160 個 variant word 前(non-variant)與置換後(variant3160)之 ppl 比較

[5] 依據經驗，Witten-Bell smoothing 在大詞彙語音辨識系統中的表現會較常見的 kneser-ney 佳。

(三)人名語言模型之建立

語音辨識器之本質是達成聲音轉文字之目的，對於日常對話之句子，將字詞辨認完全正確是非常合理且容易達成的，但是當聽到一人名時，即使是人工轉寫，對於字詞之選擇，仍是無一個百分之百正確之選擇，僅能依日常生活所聽到之常見人名選字選詞，加上中文之同音異字詞實在是多如牛毛，因此，針對辨識器對於人名之辨識結果，僅要求其音節正確，例如「王小明」這個人名將以"wang xiao ming"的形式被辨識器解碼出來。另外，本研究將人名分成兩類：常出現之人名(Somebody)與不常出現之人名(Nobody)，根據本實驗之 25 億詞語料庫統計，人名多為政治人物或新聞上常報導之人物，針對此常出現之人名，取統計數前 5000 名放入辭典中，使其能被以詞的形式解碼出來，剩下之人名，將拆成多個音節放入辭典中，例如「王小明」這一個三字詞將被拆解成"wang xiao ming"三個音節放入辭典中，也就是說，辭典中有關人名的部分將由 5000 個中文人名加上 411 個音節[6]所組成。接下來將介紹本實驗如何建立人名語言模型，使原本被辨識器當成 OOV 之人名能以音節的形式辨識出來。

1. 找出語料中所有人名位置

為了找出文本中出現之人名並置換為#Name，需先將訓練語料中所有人名位置找出來，本研究訂定了一套搜尋規則，首先，利用斷詞器所標記出之詞性(Part of Speech, POS)，挑出所有專有名詞(Nb)，接下來，查詢姓氏排名前三百名[7]作為判斷依據，假如一專有名詞的開頭出現在前三百姓氏表裡，且後續連接字數為一至二字，則判斷為人名，並用#Name 取代之(如圖 四)，以標記此處為一人名，供後續展開人名之處理。

依此規則搜尋在本 25 億訓練語料中，符合人名的個數，總共約為 4100 萬個，約占總訓練語料之 1.6%，字詞之數量僅次於「的」(如下圖 五)。由此可再次看出人名在中文語句中之重要性，不容忽視，若不對人名字詞做專門的處理，僅以一般的辨識器辨認，且假設人名之辨識率為 0%，則所有人名將被當成 OOV 處理，並造成錯一個人名導致前後字詞選詞錯誤之連鎖反應，如表 二，崔蓉芝為人名，此處辨識器因為未在語言模型找到「崔蓉芝」之前後連接詞之機率，而將這一個三字詞解碼為一個一字詞「推」加上一個二字詞「融資」，導致統計模型將計算「推」與「融資」之前後字詞連接機率，造成字詞錯誤接二連三的傳遞下去，也就是形成所謂的搶詞問題。以未經處理之 12 萬詞所建立之語言模型中，統計分析其辨識答案可知，一個二至三字人名平均會造成左右附近 2.03 個詞的錯誤，也就是說，訓練語料中之 1.6%人名最高可造成約 3.25%之詞錯

[6] 中文之所有音節數即為 411 個。
[7] https://news.cnyes.com/news/id/3660142

誤率，這對辨識器來說，實在是一不容小覷之詞錯誤率。

表 二、人名解碼錯誤導致搶詞之問題

正確答案	…	此	案	***	***	崔蓉芝	已經	勝訴	在望	…
辨識答案	…	此	案	推	融資	以及	因	勝訴	在望	…

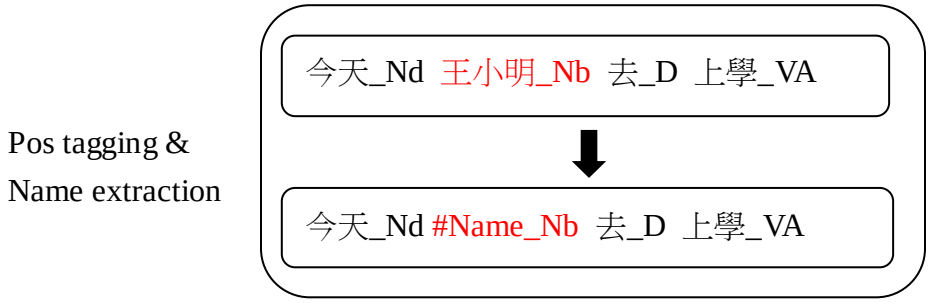


圖 四、找出人名之位置並代換為#Name

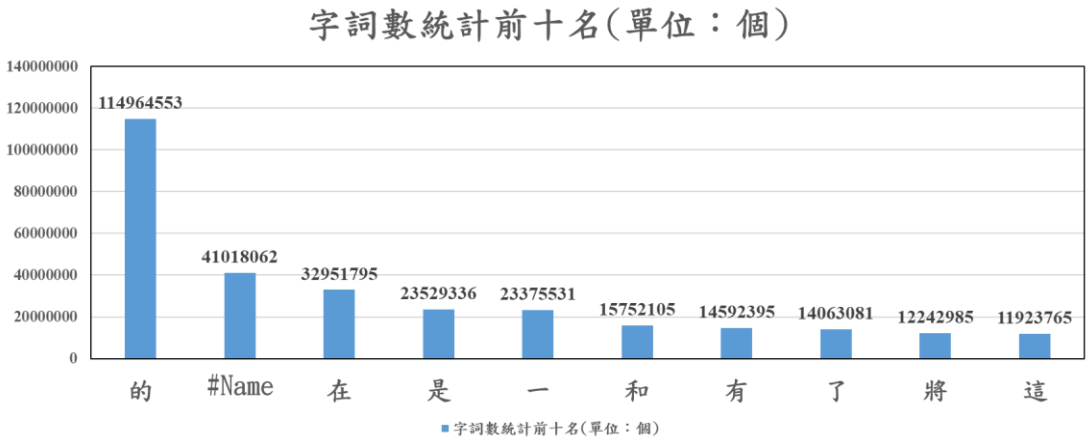


圖 五、字詞數統計前十名

2. class-based 人名語言模型

為了使訓練語料原句中#Name token 處之人名種類更多元，本實驗複製好幾份訓練語料，並根據在前一小節(五之三之 1)所找到的所有#Name token 位置，將所有找到的人名(約 4100 萬個)以隨機亂數的方式填入，若填入之字詞為語料中常出現人名(Somebody)，以字詞(word)填入；若填入之字詞為語料中不常出現之人名(Nobody)，則以音節填入，之後再將這些語料拿去做 LM training。如圖 六。

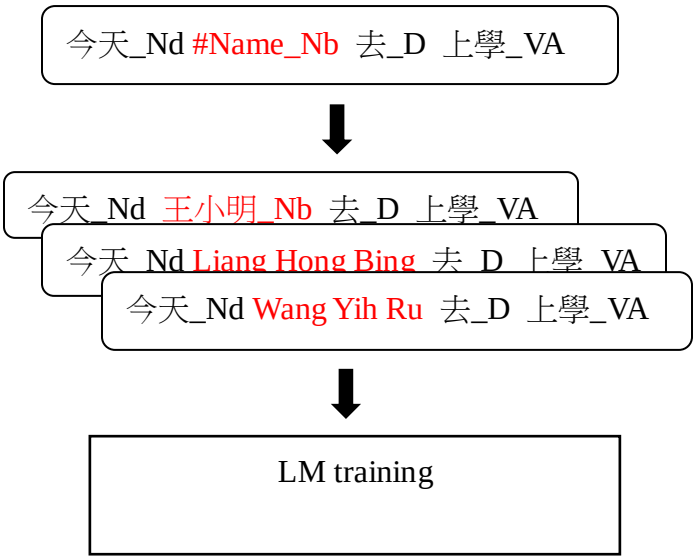


圖 六、製造人名填入#Name 之位置

經由不斷的以亂數方式將人名填入訓練語料中，如此一來，原語言模型中的人名(#Name)位置將出現許多不同的人名，也就是說，語言模型中的人名位置將不再僅能看到原訓練語料中所出現之人名機率，而是類似將#Name 這個標記(Token)展開成一個擁有許多人名機率的 class(如圖 七)，使我們的語言模型組合成 word-based(一般字詞)加 class-based(人名字詞)的語言模型。之後當辨識器遇到一個未曾出現在辭典裡之人名時，將有一定機率被解碼成人名字詞(若為 Somebody)或人名音節(若為 Nobody)。

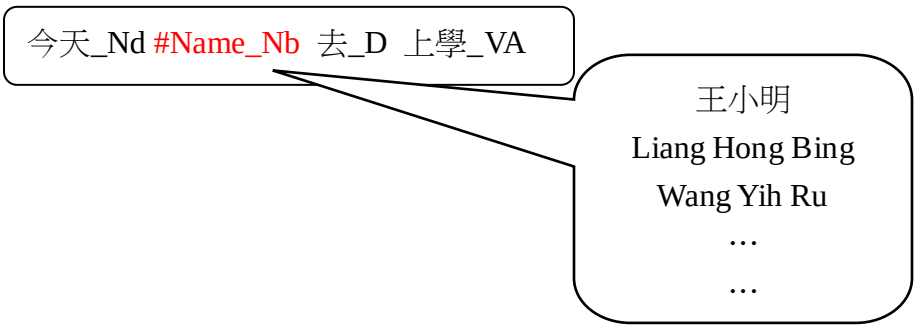


圖 七、人名 word 展開成人名 class

(四)調整一般名詞被當成人名之機率

在之前曾經提到過人名出現之機率非常高(如圖 五)，僅次於「的」，也就是說，辨識器有很高的機率會將一般名詞當成人名來解碼，若剛好欲解碼之詞彙不是人名，且此人名容易與一般名詞混淆時，此將不只影響該位置辨識錯誤，更將影響前後詞彙相連接之機率，例如：有一人名「何平」在訓練語料中出現的機率非常高，當測試句子提到「我喜歡和平」時，此句之意思應是指喜歡和平這種狀態，然而，辨識器因人名之出現機率

過高而將一般名詞「和平」解碼為人名「何平」，進而解碼成「我喜歡何平」，表示喜歡一個叫做何平的人，此將導致不僅「和平」這個字詞辨識錯誤，該位置之前後詞彙連接機率也會大不相同，將有可能會發生搶詞的情況。

因此，為了降低此種錯誤發生，本實驗額外訓練一完全不包含人名之語言模型，並定義一權重 α，以內插的方式(interpolation)與人名語言模型合併，試著降低詞彙被當人名之機率，如圖 八，詳細之實驗結果將在第六章做更詳細的說明。

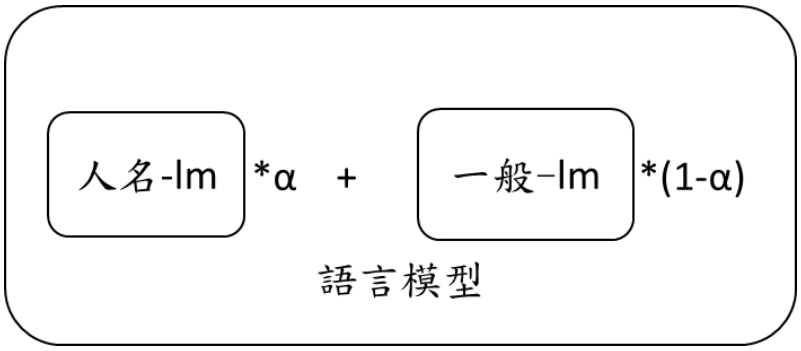


圖 八、降低人名語言模型之權重後之語言模型。

五、 實驗結果分析與討論

此實驗之測試語料除了 tcc300、ner_Clean、ner_Other 外，另加入同樣為自發性語音、新聞廣播語料之 ner2_Clean(音節數 12171 個，約 52 分鐘)、ner2_Other(音節數 223385，約 12.6 小時)、pilot_Clean(音節數 63695，約 4.3 小時)、pilot_Other(音節數 107326，約 6.8 小時)。

在評估辨識率之前，本章節將先說明因人名而導致之搶詞情形的嚴重性，並介紹錯誤率之計算方式，最後，再以調整語言模型之內插權重的方式，控制各種錯誤的發生機率。

在前面的章節曾經提到過，若放任人名在辨識器中解碼，則人名將很有可能被拆解成由許多小字數詞組(Sub-word unit)所組成，而這些 sub-word unit 又將影響前後字詞之連接機率，導致錯誤傳遞下去，例如先前提到的例子「崔蓉芝」，但也有可能很幸運地，辨識器以存在於辭典之發音相似之人名取代之，在此情況下，錯誤僅發生在人名本身，並不會發生搶詞的情況，綜合上述兩種情形，定義人名之錯誤傳遞率(Error Propagation Rate, EPR)，計算方式如式(4.1)。由於本實驗之人名語言模型會將人名解碼為音節或文字，在計算 EPR 時仍會先將辨識正確之音節記為錯誤，直到後面再做校正。本測試語料之總詞數為 387510 個，總人名為 1812 個，在加入人名語言模型後，比較各測試語料之人名錯誤傳遞率，即每單位人名所造成左右詞彙之錯誤量，如表 三。在加入人名語言模型後，僅僅在錯誤傳遞率上，就由原先的 2.03(未加入人名語言模型前)降至 1.37，也就是說，原先人名 OOV 為 1%將造成 2.03%WER 的結果，將降為 1.37%。

$$EPR = \frac{\text{人名所導致之錯誤個數}}{\text{人名錯誤個數}} \tag{4.1}$$

表 三、加入人名語言模型前後之 EPR 比較

	加入人名語言模型前	加入人名語言模型後
tcc300	1.84	1.13
ner_Clean	2.25	1.52
ner_Other	2.03	1.37
ner2_Clean	1.78	1.23
ner2_Other	2.41	1.66
pilot_Clean	1.73	1.17
pilot_Other	2.38	1.71
Overall	2.03	1.37

在評估人名語言模型之優劣時，與一般詞彙不同，且在評估時，本實驗僅針對人名解碼成音節之正確率做觀察，換句話說，那些本來就存在於詞典裡的有名人物(Somebody)將自然而然的解碼成字詞，不是此實驗之重點。在加入人名語言模型後，雖然理想上可增加原本被當成 OOV 的人名被辨識出來的機率，但同時，也可能造成幾種因人名語言模型的加入所導致的錯誤，因此，本實驗定義了兩種類型的錯誤，與一種類型的正確方式，說明如下：

1. False Alarm(type I error):

一般字詞被當成人名解碼。這裡指的一般字詞也包含 OOV，即所有非人名字詞。此類錯誤是由於語言模型中人名詞彙之出現機率過高，導致辨識器選用人名來替換一般詞彙。例：「長話短說」被解碼成「張華頓 說」。

2. Wrong Detection(type II error):

人名被當成一般字詞解碼。此類錯誤是由於人名詞彙與一般常用詞彙之發音相似，導致辨識器解碼錯誤。例：「何平」被解碼成「和平」;「鍾國仁」被解碼成「中國人」。

3. Hit:

人名被當成人名解碼。不論解碼出的人名音節是否完全正確，只要該詞彙在辨識器中是被當成人名來看待，就算 Hit。

為了評估辨識器在解碼人名詞彙之能力，定義下列參數：

定義辨識器抓到之所有人名且為 Hit 的比率 Precision 為：

$$\text{Precision} = \frac{\text{Hit}}{\text{Hit+False Alarm}} \tag{4.2}$$

定義測試語料中所有人名且為 Hit 的比率 Recall 為：

$$\text{Recall} = \frac{\text{Hit}}{\text{Hit+Wrong Detection}} \tag{4.3}$$

以調和平均數綜合以上兩個參數，計算其 F1-score：

$$F1score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{4.4}$$

當 Precision 越低，表示 False Alarm 越多，非人名被當成人名的情況越嚴重，反之亦然；當 Recall 越低，表示 Wrong Detection 越多，人名沒有被當成人名的情況越嚴重，反之亦然。

各測試語料上之結果如下圖 九所示：

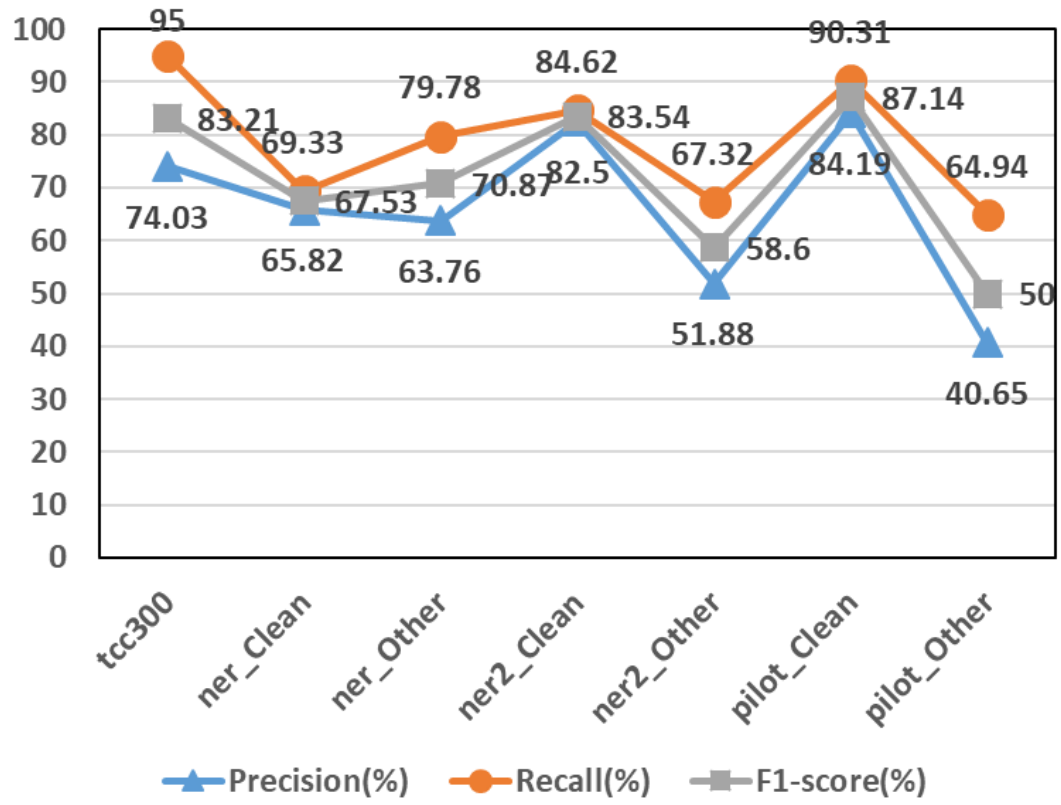


圖 九、人名辨識能力之評估

由實驗結果得知，各測試語料之 Precision 普遍偏低，這是因為辨識器容易將 OOV 當成人名來辨識，而此類錯誤又被計入 Precision 之緣故，表 四為各測試語料，OOV 佔 False Alarm 的比例。

表 四、False Alarm 中為 OOV 的比例

Tcc300	Ner_Clean	Ner_Other	Ner2_Clean	Ner2_Other	Pilot_Clean	Pilot_Other
76.67%	29.63%	53.01%	78.57%	67.19%	79.22%	54.34%

欲控制辨識器將詞彙當成人名字詞辨識的機率，以調整 Flase Alarm 或 Wrong Detection 的高低，本實驗將訓練語料中不含人名之句子特別分離出來製做一個完全不含人名之語言模型，並給定一個權重將此語言模型內插進原本的人名語言模型中，以稀釋辨識器整體人名詞彙出現之機率如式(4.5)。 分別調整 α 為 1.0(人名語言模型完全未

調整前，即圖 九使用之人名語言模型)、0.5、0.1，結果如圖 十所示。當人名語言模型權重越來越低時，Precision 因 OOV 越來越不容易被當成人名辨識而提高；Recall 將因人名越來越不容易被當成人名辨識而降低。

$$\mathrm{LM} = \alpha \mathrm{LM}_{\text{人名}} + (1-\alpha)\mathrm{LM}_{\text{不含人名}} \tag{4.5}$$

α=1.0

	Tcc300	Ner_Clean	Ner_Other	Ner2_Clean	Ner2_Other	Pilot_Clean	Pilot_Other
Precision(%)	74.03	65.82	63.76	82.50	51.88	84.19	40.65
Recall(%)	95.00	69.33	79.78	84.62	67.32	90.31	64.94
F1-score(%)	83.21	67.53	70.87	83.54	58.60	87.14	50.00

α=0.5

	Tcc300	Ner_Clean	Ner_Other	Ner2_Clean	Ner2_Other	Pilot_Clean	Pilot_Other
Precision(%)	75.57	72.86	68.25	81.82	51.33	83.99	47.54
Recall(%)	92.78	68.00	78.26	80.77	65.85	88.99	63.04
F1-score(%)	83.29	70.34	72.91	81.29	57.69	86.42	54.21

α=0.1

	Tcc300	Ner_Clean	Ner_Other	Ner2_Clean	Ner2_Other	Pilot_Clean	Pilot_Other
Precision(%)	86.13	85.45	81.58	85.29	61.62	89.14	66.32
Recall(%)	82.78	62.67	66.67	74.36	55.61	79.52	55.17
F1-score(%)	84.42	72.31	73.37	79.45	58.46	84.05	60.24

圖 十、調整詞彙被當成人名之機率

前述實驗在計算 Hit 的時候，僅要求解碼出的答案為人名，而 Hit 裡又可再細分為三種類型：1.音節完全辨識正確，例如「梁鴻彬」解碼為「liang_hong_bin」。2.音節相似但不完全正確，例如「郭振興」原本應解碼為「guo_zheng_xing」但辨識器解碼為「guo_zheng_xin」，一個是「ㄒㄧㄥ」的拼音，一個是「ㄒㄧㄣ」的拼音，由於人們在平常講話時「ㄥ」跟「ㄣ」本來就分辨不出來，故此種音節相似之解碼錯誤實在是無可厚非。3.由一個發音相似，且存在於詞典裡之人名代替之，例如「黃朝興」被辨識器解碼為「黃昭星」，其中，「黃昭星」為有被選入詞典中之常出現人名，也就是前面章節所定義的 Somebody。

在所有測試語料之總詞彙數 387510 個、總人名詞數為 1812 個中[8]，本實驗之辨識器能辨識人名的正確率(hit/總人名數)約為 81%；辨識出人名且音節完全正確[9]之正確率(hit 且音節完全正確/總人名數)約為 73%，也就是說，即使測試語料出現一個從未出現在詞典中之人名，本辨識器仍有 73%之機率將其音節完全正確的辨識出來。

最後，為了觀察人名語言模型對於整體辨識率的影響，我們重新計算加入人名語言模型後辨識器之詞錯誤率，這裡僅將 Hit 中音節完全解碼正確之人名當作解碼正確，並與未加入人名語言模型前之辨識器做比較，實驗結果如下圖 十一所示。此實驗使用之

[8] 這裡的 1812 個人名包含存在於詞典之人名與未存在於詞典之 OOV 人名。
[9] 這裡指僅計入 Hit 中音節完全正確之類型。

聲學模型為上述(第四章)的 TDNN-F 架構，語言模型為 4-gram 語言模型，平滑化方法使用 Witten-Bell smoothing，訓練語料為 25 億詞，辭典大小為 120k，唯一差別僅在於語言模型有或沒有加入人名語言模型做調整。

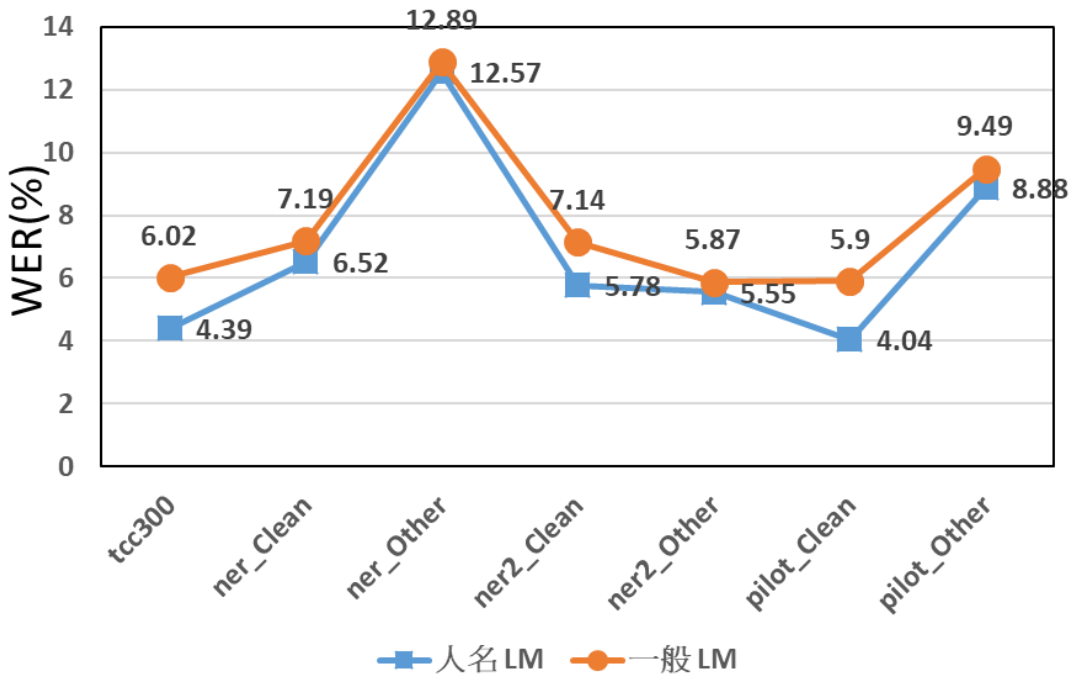


圖 十一、加入與未加入人名語言模型之詞錯誤率比較

六、 結論

參考 Kaldi 之作者 Daniel Povey 在 2018 發表之論文，使用 TDNN-F 之架構下訓練之聲學模型，辨識速度的確比傳統之 LSTM 快很多，即使音節辨識率有些差距，那些差距在後級語言模型解碼後也顯得微乎其微。

在人名辨識率方面，加入人名語言模型可以使以前被當成 OOV 處理之人名被解碼出來，不但救回了以前不在詞典裡之人名以外，也減少了左右搶詞的情形。

本實驗建構之大詞彙中文語音辨識器統對於乾淨語料(Clean)有絕佳的辨識能力，但是對於噪音環境(Other)語料之辨識率仍然有改進之空間，尤其在聲學模型方面，本實驗並沒有對噪音做特別的處理，導致在噪音環境下之音節錯誤率仍然偏高，這會使得即使後級之語言模型再強，某些字詞還是無法挽救回來之情形，而語言模型如果能克服記憶體限制之問題往 5-gram 語言模型發展，將會使辨識率又再更為提升，尤其在加入人名語言模型後，訓練語料裡之人名被展開至 2 到 3 個音節數，使用 5-gram 語言模型將可以看得更遠，以增加人名詞彙之 Hit 數。此外，目前實驗室辨識器對於未出現在詞典裡之人名是以音節的形式解碼出來，也許未來能以機率或其他搜尋的方式，選擇大家普遍能接受的字詞呈現。

參考文獻

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

[2] "Formosa Speech Recognition Challenge 2018," [Online]. Available: https://sites.google.com/speech.ntut.edu.tw/fsw/home/challenge#h.p_I_b8URx26NXZ.

[3] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline,” Proc. Oriental COCOSDA, 2017.

[4] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohamadi, Sanjeev Khudanpur, "Semi-OrthogonalLow-RankMatrixFactorizationforDeepNeuralNetworks," in *interspeech*, 2018.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 358-367

基於 Seq2Seq 模型的中文文法錯誤診斷系統

A Chinese Grammatical Error Diagnosis System Based on Seq2Seq Model

王鈞威　Jun-Wei Wang, 簡聖倫 Sheng-Lun Chien
陳義昆 Yi-Kun Chen, *吳世弘 Shih-Hung Wu
朝陽科技大學資訊工程系
Department of Computer Science and Information Engineering
Chaoyang University of Technology
s10427098@gm.cyut.edu.tw
s10727614@cyut.edu.tw
kun26712930@gmail.com
*shwu@cyut.edu.tw (contact author)

摘要

本文將以中文句子的錯誤診斷為實例，說明如何利用深度學習演算法序列對序列(Seq2Seq)模型[1]，使用其中的編碼器與解碼器架構，實作出能夠從學習者的句子當中生成出修改過後的句子，並且識別錯誤的類型。一個句子是由許多詞所組成，我們透過修正前與修正後的兩個句子配成一對讓演算法進行學習，盡可能的使模型識別原始與正確之間的關係，並將有錯誤或是不通順的句子加以修正與改正。此研究利用 Pytorch 所提供的範例更改為我們所想要的功能，以此理論作為基礎的中文文法錯誤診斷系統；此研究分為兩部分：首先利用 NLP-TEA2 至 NLP-TEA5 的 Shared Task 所提供的資料訓練模型。其次因應資料集數量不夠讓機器充分學習，所以我們利用 Ge 等人[2]所提出的方式來擴大訓練的資料集。過去 Chen [3]在 NLP-TEA3 的 Shared Task 使用條件隨機域[4](Conditional Random Field，CRF)得到當時最佳的準確度與精確度。所以我們主要針對 NLP-TEA3 當時所完成的任務結果來做比較，另外為了確保我們所使用的序列對序列的可行性與公平性，在此我們重新訓練 CRF 不做任何的調整與現在的序列對序列一樣做比較。

關鍵詞：文法錯誤診斷系統，深度學習，序列對序列模型，條件隨機域

一、緒論

根據網路調查中文是全世界第二多人學習的語言，現在使用中文的人口已經超過十三億，有越來越多的外國人學習中文。可是中文包含很多漢字符號五音加上沒有固定的文法規則等等因素，所以它被認為是全世界最難學習的語言之一，導致外國人要學習中文是非常困難的。

為開發新的文法錯誤診斷系統，需要使用大量的語料庫讓深度學習演算法調整出合適的模型，透過使用 NLP-TEA2 至 NLP-TEA5 Shared Task 所提供的所有訓練集，將訓練集裡面的原始的句子與修正後的句子倆倆進行配對便會產生 22,656 個句對，利用這些句對建立一個新的文法錯誤診斷系統與之前的 CRF 診斷系統進行比較，但這樣並不能讓模型充分的發揮出它的效果，而這邊我們會再使用 Ge 等人[2]提出的擴大訓練集的方式，因為一個句子裡面存在著不只一種的錯誤，然而模型也未必能夠一次性的完全修改正確，當然模型可有可能將原本正確的句子修改成錯誤，因此利用這幾種特性使得尚未修改完全的句子與修改錯誤的句子都成為了不同的錯誤，就能讓模型以多對一的方式訓練讓模型遇到不同種狀況的時候，能夠靈活的辨識出句子的錯誤，而繁體中文的 TOCFL 的資料集與簡體中文 HSK 的，是根據官方蒐集外國人寫作中文的句子，進行分析得出來的經常性錯誤，大致可以分為四類：冗字(Redundant word，簡稱 R)、缺字(Missing word，簡稱 M)、用字不當(word Selection error，簡稱 S)與詞序錯誤(Word ordering error，簡稱 W)。

二、方法

(一) 條件隨機域(Conditional Random Fields, CRF)

CRF 是條件機率分布模型P($Y|X$)，給定一個 X 序列的標籤，CRF 可利用這標籤經由訓練產出另一組序列輸出 Y，而 Y 是因任務的不同而也會有不同的標籤集合，由圖一所示，若是給定一個資料 X 則會計算 Y 所持有的標籤集合裡所有機率，最後返回一個機率值最大的標籤 Y 作為輸出，而根據模板的設定 X 也可以配合前幾個 X 所輸入的資料或是其他內外部特徵，都會是有助於模型是別出更為精確的輸出 Y。

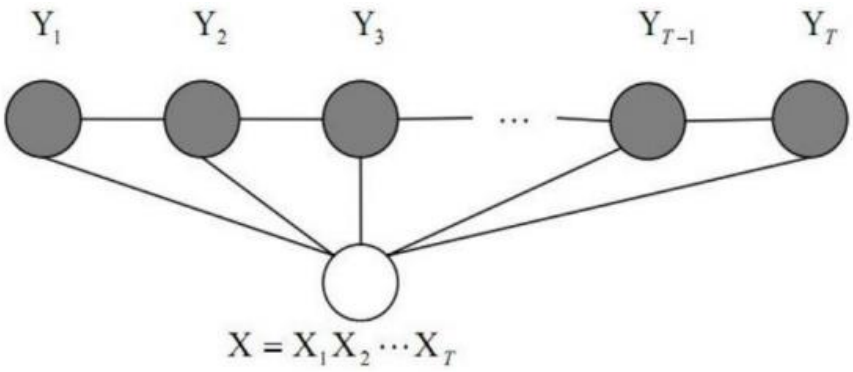


圖 一、條件隨機域

這系統 X 所代表輸入的句子與詞性，而 Y 就是跟 X 相對應的錯誤類型標籤，我們把標籤集定義為：{ O, R, M, S, W }，分別是以下幾種類型：沒有錯誤、冗詞、缺字、用字不當與詞序錯誤。若是將輸入 X 的序列放入到 CRF，此時 CRF 就會利用事先建立好的模板抓取所需要的特徵，而輸入的序列不只是單單放入詞這一種特徵加入 POS 等等的特徵對於 CRF 可以有更好的效果，而在測試時將輸入 X 的序列給 CRF 這時會產生多組可能的標籤組合而每一組組合都會產生符合 X 的機率，最後將機率最大的那一組作為 X 序列的輸出。

(二) 序列對序列(Sequence to Sequence, Seq2Seq)模型

在本次的中文文法錯誤診斷系統，本團隊使用的技術核心為 Seq2Seq 加入 Bahdanau[11]等人所提出來 Attention 專注機制以及雙向 GRU 架構的模型，而在 Seq2Seq 裡面 Encoder 就是負責將輸入序列消化、吸收成一個向量，我們通常把這個向量稱為 context vector，顧名思義，這個向量會把原序列的重要訊息包含起來送至 Decoder 當中，而 Decoder 則是根據 context vector 來生成文字，如圖二所示。

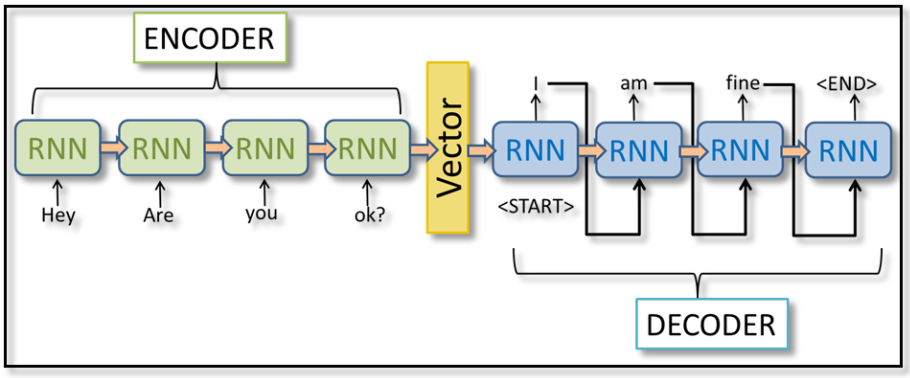


圖 二、Encoder 與 Decoder 示意圖

1. 編碼器(Encoder)

在此模型的編碼器，最主要的工作在於將每一個接收到的文字轉換成一個文字向量與隱藏狀態，並且將每一次的文字向量以及隱藏狀態存取起來，最後將整個句子的向量

及隱藏狀態串起並傳向給解碼器，解碼器將會使用這些向量和隱藏狀態來生成有對於先前的輸入有意義的文字輸出。

在此模型中的編碼器之核心是由 Cho 等人在 2014 年所發明的 multi-layered Gated Recurrent Unit—GRU[7]，此模型使用的是雙向變通的 GRU，這意味著此模型基本上有兩個獨立的 RNN：(1)一個正常方向的序列輸入(2)一個反向的序列輸入 在這兩個 RNN 的輸出端都會計算每個時間的向量和隱藏狀態以便抓取最佳解，使用雙向變通 GRU 將會使模型在編碼過去和未來上下文時有明顯的優勢。雙向變通 GRU 如圖三所示。

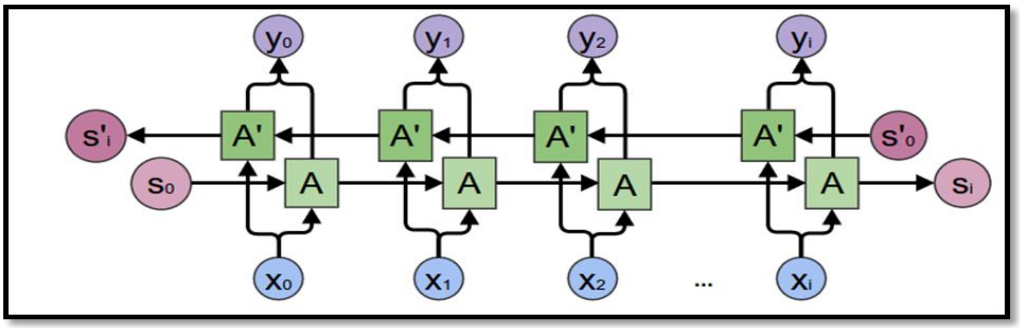


圖 三、雙向變通 GRU 示意圖

2 解碼器(Decoder)

在此模型的解碼器部分，解碼器會使用編碼器傳過來的文字向量以及隱藏狀態來做文字生成的工作，解碼器會依照收到的向量和隱藏狀態去計算並且逐字的生成單字直到解碼器最後生成 EOS_taken 表示句子的結尾，此時解碼器便會停止成文字，但是在一般 Sequence to Sequence 解碼器當中的一個常見問題是，如果我們只依賴於上下文的向量來編碼整個輸入文字的含義，那麼我們很可能會遺漏訊息或是完全讀取錯誤訊息，尤其在處理一段長的輸入序列時更是如此，這極大地限制了解碼器的能力。

為了解決這個問題，Bahdanau[6]等人，創建了一種“Attention 機制”，允許解碼器關注於輸入文字的某些部分，而不是整個輸入的句子，進而提升解碼器再生成文字時能更加的正確率，而在進行錯誤句子與正確句子的分析時，Attention 會將兩者的句子互相差異的地方視為一個需要修改的重點，讓這些有差異的錯誤再生成時不會再出現。在 PyTorch 所提供的 Chatbot 中將 Attention 機制加入此模型中，提高了解碼器的效能。

(三) 內部擴增訓練資料

訓練資料過於少，這是一個在實作深度學習時經常會遇到的棘手問題，特別在處理自然語言處理上的深度學習，更會因為訓練資料量非常的小而使機器訓練的效果不太理想，在經過許多實驗後，決定採用生成的方法擴大訓練資料，Ge 等人提出將第一次 Seq2Seq 所產生的錯誤句子當作訓練資料，然後重新訓練機器並且持續循環到產生出最好的正確句子。利用此方法來擴大機器自己的訓練資料進而提升機器的文字生成正確率，由此方法解決訓練資料過少的問題，本團隊將此概念運用在中文文法錯誤診斷系統上，

將中文訓練資料擴大。

(四) 編輯距離(Edit Distance)

在此次的語法改錯中，我們基於語法錯誤類型中使用了 Edit_distance[10]，由於官網所提供的答案有四種錯誤類型的形式，而我們現階段只能計算三種錯誤類型，所以此次不考慮其中的一種錯誤是否詞序錯誤(W)，因本身詞序錯誤在我們所獲得的資料集裡並不是很多，也因本身想要找出兩句之間字詞調換的位置是困難的，利用此方法將 Edit distance 中的刪除視為冗字、插入為缺字與替換並非詞序錯誤是將原本字詞替換成正確的字詞，使用套件的方式是將官網所提供的句子跟模型產生的句子做字串比對，透過套件的 opcode 會回傳句子的錯誤類型及位置。

三、實驗與結果

(一) 實驗設定

此次的實驗分為四個部分，實驗一是使用 NLP-TEA2 與 NLP-TEA3 的訓練集，實驗二是利用 Ge 的方法擴大訓練集使原本的訓練集擴大兩倍來增加模型對於不同錯誤的修改。實驗一與實驗二分別是模擬當時的 NLP-TEA2 的 Shared Task 利用有限的資料並完成當時的任務。實驗三將 NLP-TEA2 至 NLP-TEA5 所有的資料集，而實驗四同樣是將實驗二的擴大資料集的方式擴大 NLP-TEA2 至 NLP-TEA5 的所有的資料集並放大兩倍，實驗三與四是利用能夠得到的所有資料都套用到模型裡面，盡可能讓模型多看到一些文字。

表 一、訓練集大小

	NLP-TEA2	NLP-TEA3	NLP-TEA4	NLP-TEA5
Redundant	434	10,010	5,852	208
Missing	622	15,701	7,010	298
Disorder (word ordering)	306	3,071	1,995	87
Selection	849	20,846	11,591	474

而評估方式上主要利用模型所生成出來的句子與原始句子作比對，透過比對將所需要的四個類型(冗詞、缺字、詞序錯誤與用字不當)辨識出來並使用 FPR、Accuracy、precision、recall 與 F1-score 再配合表二混淆矩陣來評估。

- False Positive Rate (FPR) = FP / (FP+TN)

- Accuracy = (TP+TN) / (TP+FP+TN+FN)
- Precision = TP / (TP+FP)
- Recall = TP / (TP+FN)
- F1 = 2*Precision*Recall / (Precision+Recall)

表 二、混淆矩陣

混淆矩陣		系統結果	
		Positive	Negative
模型生成	Positive	TP	FN
	Negative	FP	TN

除了使用上述四項評估標準還會再分為 Detection Level、Identification Level 兩個等級再繼續細分下去，Detection Level 會是一個二分類的問題，看此句子正確或是不正確，Identification Level 檢測是否為該錯誤類別，模型預測出來的必須與原本給定的錯誤類型相同。

(二)實驗一

在使用 Pytorch 所提供的 Chatbot 模型裡我們將隱藏層更改為 500 層與 750 層訓練都訓練 50000 回讓兩個模型來比較看哪個能夠有比較好的效果，而測試集的部分是使用 NLP-TEA3 所提共 TOCFL 與 HSK 兩種的測試資料，並且使用官方所釋出的測試工具以達到一個公平的測試結果。

而我們可以根據表三所示，從 RUN1 至 RUN3 是 Chen 參加 NLP-TEA3 所得到的分數。可以看出在各種的分數上我們還是無法與當時最好的成績來比較，但是雖然成果不好但在 FPR(False Positive Rate)的這個部分還是有不錯的成績，因 FPR 是必須越小越好也就表示模型的誤判程度是比原先來的低的。

表 三、實驗一 NLP-TEA3 TOCFL 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
RUN1	0.347	0.595	0.625	0.541	0.580	0.515	0.460	0.302	0.364
RUN2	0.355	0.595	0.623	0.550	0.584	0.513	0.456	0.306	0.366
RUN3	0.363	0.594	0.620	0.554	0.585	0.508	0.447	0.300	0.359
500H	0.294	0.496	0.523	0.301	0.383	0.373	0.255	0.204	0.227
750H	0.297	0.497	0.523	0.305	0.386	0.375	0.264	0.219	0.239

另外從表四可以看到如果怎模型在 HSK 的效果會原比預測 TOCFL 來的好，是因為簡體字是將很多原本繁體字的簡化而成的，所以模型所需要認識的字進而減少也就導致預測出來的句子會比繁體字來的好，可以從 Detection Level 來看除了 Recall 與 F1 其他兩個都比原本的分數來的高，而 FPR 也比當初的分數來的低這就可以看出如果好好的加以調整這個模型是可以比原本來的更好。

表 四、實驗一 NLP-TEA3 HSK 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
RUN1	0.401	0.614	0.600	0.630	0.615	0.571	0.530	0.437	0.479
RUN2	0.419	0.611	0.595	0.644	0.618	0.566	0.523	0.450	0.484
RUN3	0.401	0.614	0.600	0.630	0.615	0.572	0.530	0.435	0.478
500H	0.250	0.619	0.648	0.484	0.554	0.528	0.455	0.362	0.403
750H	0.258	0.617	0.643	0.487	0.554	0.521	0.444	0.357	0.396

(三)實驗二

將原先實驗一的資料加入後兩屆所提供的訓練資料，讓模型可以識別更多的文字，但後來加入的資料多數為簡體中文，對於 TOCFL 的結果從表五可以看到在效果上實驗二的結果會比實驗一來的差，識別過多的簡體中文反而會對繁體中文造成一定的影響，在進行修正的過程裡模型會將原本 TOCFL 答案修正為簡體中文，而在最後將預測結果轉換為比賽格式會與最初的測試資料進行比對，過程中因識別出簡體中文就會導致錯誤的出現。

而在 HSK 的表現可以由表六得知，在兩個等級的 Recall 有了成長，因加入前兩屆的資料讓模型識別更多的句子與不同的錯誤方式，在面對尚未識別過的句子能夠抓出更多的錯誤。

表 五、實驗二 NLP-TEA3 TOCFL 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
500H-1	0.294	0.496	0.523	0.301	0.383	0.373	0.255	0.204	0.227
750H-1	0.297	0.497	0.523	0.305	0.386	0.375	0.264	0.219	0.239
500H-3	0.300	0.491	0.515	0.297	0.377	0.374	0.260	0.205	0.230
750H-3	0.304	0.494	0.518	0.305	0.384	0.369	0.255	0.211	0.231

表 六、實驗二 NLP-TEA3 HSK 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
500H-1	0.250	0.619	0.648	0.484	0.554	0.528	0.455	0.362	0.403
750H-1	0.258	0.617	0.643	0.487	0.554	0.521	0.444	0.357	0.396
500H-3	0.285	0.609	0.626	0.499	0.555	0.509	0.431	0.365	0.395
750H-3	0.283	0.613	0.630	0.504	0.560	0.510	0.431	0.375	0.401

(四)實驗三

此次實驗是由將重新訓練 CRF 與 Seq2Seq 進行比對，此次的 CRF 與實驗一和二最大的不同是不使用 Chen 所使用的搭配詞僅使用原始的詞與詞性所得到的結果，而 Seq2Seq 將實驗一與二透過預測訓練資料產生不同的句子來擴大因訓練資料，但與實驗一相同在這次實驗中只採用 NLP-TEA2 與 NLP-TEA3 的資料集以此做為限制完成當時的任務。

如表七、八所示，CRF 相較於 Seq2Seq 注重於 Precision 反而忽略 Recall 才會使得得到較好的 FPR，然而 Seq2Seq 傾向於句子全面性的修改，雖然 Precision 下降但 Recall 的升高反而在三等級的 F1 都拿到比 CRF 還要好的成果。

表 七、實驗三 NLP-TEA3 TOCFL 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
CRF	0.042	0.521	0.772	0.129	0.221	0.484	0.610	0.058	0.106
500H	0.308	0.494	0.519	0.307	0.388	0.378	0.271	0.220	0.243
750H	0.319	0.494	0.517	0.318	0.394	0.369	0.263	0.227	0.244

表 八、實驗三 NLP-TEA3 HSK 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
CRF	0.071	0.573	0.818	0.273	0.409	0.530	0.755	0.187	0.300
500H	0.317	0.608	0.615	0.530	0.570	0.510	0.439	0.398	0.416
750H	0.333	0.600	0.603	0.529	0.564	0.500	0.428	0.393	0.410

(五)實驗四

重複實驗三之實驗，將 CRF 與 Seq2Seq 訓練集放大到使用 NLP-TEA2 至 NLP-TEA5 所有的資料集，同樣的 Seq2Seq 使用與實驗三同樣的方式將資料放大一倍。

如表九、十所示，在 Recall 的方面都有所成長相對的會導致誤判的上升，但這樣使得 F1 來得比實驗三都來的好。

表 九、實驗四 NLP-TEA3 TOCFL 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
CRF	0.062	0.528	0.743	0.159	0.262	0.481	0.560	0.070	0.125
500H	0.327	0.492	0.515	0.323	0.397	0.363	0.261	0.243	0.252
750H	0.328	0.494	0.517	0.327	0.400	0.360	0.255	0.239	0.247

表 十、實驗四 NLP-TEA3 HSK 測試結果

Runs	FPR	Detection Level				Identification Level			
		Acc.	Pre	Rec	F1	Acc.	Pre	Rec	F1
CRF	0.096	0.614	0.832	0.380	0.522	0.550	0.775	0.265	0.400
500H	0.335	0.606	0.608	0.544	0.574	0.499	0.423	0.409	0.416
750H	0.350	0.604	0.603	0.555	0.578	0.487	0.409	0.411	0.410

四、結論

從實驗一至實驗四看下來，雖然無法比當時 CRF 所得到的分數來得高，但從這幾次的實驗裡我們看到 Seq2Seq 的可能性，而從實驗一 Chen 的表現我們可以知道以現在的 Seq2Seq 還是很難與當時 CRF 做抗衡在各方面的分數都是低於他們的表現，實驗二是不局限於 2016 年 NLP-TEA2 當時資料集，透過將 NLP-TEA2 至 NLP-TEA5 所有的資料集全部套入到模型裡做訓練，可以得知在增加資料的同時也會讓 Recall 隨之地增加利用這一項原理再加上 Tao Ge 等人的方法進行了實驗三與實驗四，將訓練資料集利用訓練好的模型產生不同的錯誤模式並加入到原先的訓練集增加到兩倍的訓練量，讓平時一對一的模型看到各種不同的錯誤方式，來達成二對一甚至是三對一的方式來使得自己的模型更加的靈活而若是再搭配 PreTrain 的技巧，例如：Word2Vec[8]或是 BERT[9]的等等技巧，將文字轉成向量能夠讓詞與詞有良好的連接。

參考文獻

[1] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, 2014, “Sequence to Sequence Learning with

Neural Networks”, In Advances in neural information process-ing systems, pages 3104–3112.

[2] Tao Ge, Furu Wei, Ming Zhou, 2018, “Fluency Boost Learning and Inference for Neural Grammatical Error Correction”, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1–11.

[3] Po-Lin Chen, 2017, “Chinese Grammatical Error Diagnosis System”, Retrieved from http://ir.lib.cyut.edu.tw:8080/handle/310901800/34228

[4] Taku Kudo, 2007, “CRF++：Yet Another CRF toolkit”, https://taku910.github.io/crfpp/.

[5] NLP-TEA3 CGED Shared Task, 2016, https://www.aclweb.org/anthology/W16-4906.

[6] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio, 2014, “Neural Machine Translation by Jointly Learning to Align and Translate”, arXiv preprint arXiv:1409.0473.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio, 2014, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling”, arXiv preprint arXiv:1412.3555v1.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (NIPS'13), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 2. Curran Associates Inc., USA, 3111-3119.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ,(Submitted on 11 Oct 2018 (v1), last revised 24 May 2019 (this version, v2))

[10] Eric Sven Ristad, Peter N. Yianilos, 1998, IEEE,” Learning String-Edit Distance”, IEEE Transctions On Pattern Analysis And Machine Intelligence, Vol. 20, No. 5.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 368-377

應用文脈分析於中英夾雜語音合成系統

Linguistic Analysis for English/Mandarin Speech Synthesis System

洪翌翔　Yi-Hsiang Hung
國立屏東大學資訊科學系
Department of Computer Science
National Pingtung University
gbaian10@gmail.com

黃奕欽　Yi-Chin Huang
國立屏東大學資訊科學系
Department of Computer Science
National Pingtung University
ychuangnptu@mail.nptu.edu.tw

鄧廣豐 Guang-Feng Deng
資訊工業策進會
Institute for Information Industry, Taipei, Taiwan
raymalddeng@iii.org.tw

摘要

本論文將藉由文脈分析的處理，實作出一套中英夾雜的語音系統。在語音模型的建模上，採取統計式模型中的隱藏式馬可夫模型(Hidden Markov Model)做為基礎針對中文以及英文進行處理。在系統的實作中，首先在合成語音前先將文字做前語言處理切割成中文和英文的部分，接著將中文與英文分別已預先訓練好的的中文/英文之語音模型分別進行合成，最終將各自合成的部份進行語音段的串接。其中，由於中文以及英文為不同的語言，為了維持整段話的連貫性，若整個句子以中文句當作主體，並且將此中英夾雜句中的英文字的部份，透過其詞性分析(POS Analysis)找出其詞性後，將此英文字置換成與其詞性相同的中文字(Substitute Word, 縮寫為 SW)，使其與原英文字的詞性相同，在中文主體句中，則透過置換過後的中文句來進行文脈分析，挑選合適的中文語音模型，並用來為合成整段中文句子，並且將合成好的英文部分替換回該句中完成中英文夾雜的句子。透過實驗分析顯示，透過文脈的分析，能夠幫助合成的句子的語流較為順暢，因而提升中英夾雜句的何成語音更為自然。

關鍵字 ： 中英夾雜句、隱藏式馬可夫模型、文脈分析、語音串接、語音合成

Abstract

In this study, we analysis the effect of the linguistic information for the English/Mandarin speech synthesis system. In order to construct the acoustic models for both languages, we

adopted the Hidden Markov Model. For the system implementation, we firstly detected the language segments for each language of the input bilingual sentence, and then independently generate the feature sequences for each language. However, for generating fluent synthesized speech, the linguistic information should be taken into account. Here, if the bilingual sentence is mainly written in Mandarin with a few English words, we firstly analyze the Part-Of-Speech information for the English words. Then, we adopted some substitute words (SW) to translate the English parts into Mandarin which have the same POS tags as their corresponding English words. Finally, The entire sentence consists of only one language and could be analyzed linguistically and keep its context information. Finally, the synthesized speech should be more fluent since the contextual linguistic information is used for choosing the suitable acoustic model sequence. In order to construct the original bilingual speech utterance, the English segment is substituted back to the synthesized speech. Experimental results showed that adding the contextual linguistic information is indeed helpful for generating fluent speech for the bilingual sentences.

Keywords: English/Mandarin bilingual sentence, Hidden Markov Model, Linguistic analysis, Speech concatenation, Speech synthesis

一、緒論

(一)、研究動機

隨著世界朝向國際化的發展，不同語言之間的交流越來越盛行，不管是在學界、業界，都免不了會接觸到不同語言間的各式各樣問題，而某些特定的專有名詞翻成當地語言時，時常會有無法充實表達其原意的困境，亦或者可能發生類似繁中與簡中的翻譯完全不同甚至到了相反的意義(如:'行'與'列')。此時為了避免問題，我們常常會在語言中直接使用該詞的原本發音來溝通，這在人與人之間或許並沒有什麼太大的問題，但實際上因為中文跟英文不管在發音還是整個語言以即文字的結構上都有著非常大的不同，這使的如果要合成一句多語言的語音合成容易發生語調不順暢的問題，為此我們必須找一個方法來解決此問題。本論文提出一種依照前後文分析文脈關係以及根據其詞性來確認中文的發音方式，使得合成中英夾雜語句時能保持中文的整體脈絡，進而提升合成語音的流暢性與自然度。

(二)、相關研究

近年來語音合成系統廣泛被使用的主要有兩種，一是基於大型語料庫樣本做串接，如:單元選擇(Unit selection approach) [1]，另一種則是基於統計方法。如:隱藏式馬可夫模型(HMM-based approach) [2]。單元選擇合成雖然有著極佳的合成音質，但卻需要非常大量的語料庫做支持，所以在製作語料成本上有著極大的代價。而隱藏式馬可夫模型對語料庫的需求則不像前者需求這麼大。

語音相關的應用也非常多，例如在不同語速下的應用 [3]、合成歌唱合成系統 [4]、情緒轉換 [5]、多語言語音合成 [6]、基於深度神經網路在多語言間的應用 [7]等，其中關於 [7]DNN 部份所合成的多語言系統為使用單一合成器，並且在合成出來的語音上有著不錯的結果，但其缺點是必須選擇語系相似的語言，且通常要找到精通多語言的相同語者是困難的，對此問題也有 [8]這些使用類似音素來進行不同語言間資料的補足的方法，且中英合成中也有使用混合兩語言決策樹的方法。

(三)、隱藏式馬可夫模型系統概述

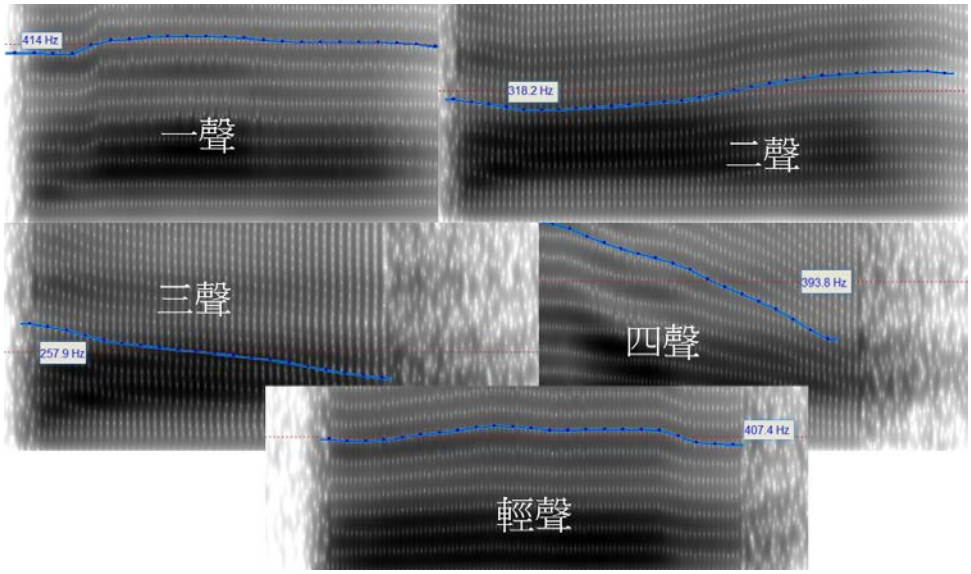
近年來隱藏式馬可夫模型在語音合成的領域中已經有著舉足輕重的地位，因為其所合出的語音流暢度已經不亞於傳統串接合成的結果，且其所訓練出來的模型所需要的儲存空間也相對較小，有較高的攜帶性。在本文中所使用的基於隱藏式馬可夫模型的語音合成系統(HMM-based Speech Synthesis System, HTS)，是由 HTS 工作團隊 [9]所研發，該技術由 Hidden Markov Model Toolkit(HTK) [10]研發修改而來，HTS 團隊提供了一個便於開發的研究平台，並有效幫助 HMM 的訓練。

本研究目標為建立一套基於 HMM 的語音模型，主要可分為訓練與合成兩個階段。訓練階段時由聲音語料藉由 SPTK [11]估算頻譜及音韻參數，聲音語料所相對應的文字經由文字分析器(Text analysis)產生相應的文脈訊息，在經由問題集(Question set)分類樹產生與文脈訊息相關的 HMM 模型。合成階段將欲合成的語音文字當作輸入丟入文字分析器中產生相應的文脈訊息，再經由問題集分類樹找出該段文字對應的 HMM 模型序列，再經由 HMM 模型產生對應的頻譜及音韻參數，最後合成出所需的語音訊號當作輸出。

(四)、章節概述

本論文主要敘述中英夾雜語音合成系統為了確保連貫性，而使用 SW 替換英文字來合成整句中文字以保持整體中文脈絡，本論文主要分成五節:第一章 ： 緒論，主要說明研究動機、相關研究與討論、以及系統概述。第二章 ： 中英文語音模型，定義分別介紹中

文、英文語音模型的定義。第三章 ： 中英夾雜語音合成系統實作，詳述中英切割的方法、中文音素模型的建立、前後文相關之問題集決策樹。第四章 ： 實驗結果及討論，說明實驗目的、語料庫來源以及內容、分析及討論結果。第五章 ： 結論，總結整篇論文的結論。



圖一、中文五聲變化

二、中英文語音模型定義

(一)、中文模型定義

中文約有 420 個不含聲調的基本單元，加入五聲變化後，則有超過 1200 個含聲調的單元。若直接對這約 1200 個模型進行訓練或許可以達到不錯的結果，但是這需要非常非常大量的語料庫來訓練才有可能，若資料量不足可能導致某些音訓練不足難以發出正確的音調，甚至可能發生某些音連一個資料都沒有的情況。

為了解決模型過多導致訓練不足的問題我們必須盡可能壓低模型的數量，本論文中，我們以“Segmental Tonal Phone Model” (STPM) [12]做為定義我們中文模型的方法，對此我們必須將模型考慮至聲母、韻母以及五聲來進行設計。然而五聲的變化差別在音高上，所以要區別五聲就必須得觀察五聲音高的的變化，圖一分別為五聲的頻譜圖以及其音高(圖中聲音分別為:巴、拔、把、爸、吧)。如圖所示，音頻的範圍大概可分為高音頻範圍(H)、中音頻範圍(M)、低音頻範圍(L)，五聲的變化趨勢分別為:

一聲:H→H 、 二聲:L→H 、 三聲:L→L 、 四聲:H→L 、 輕聲:M→M

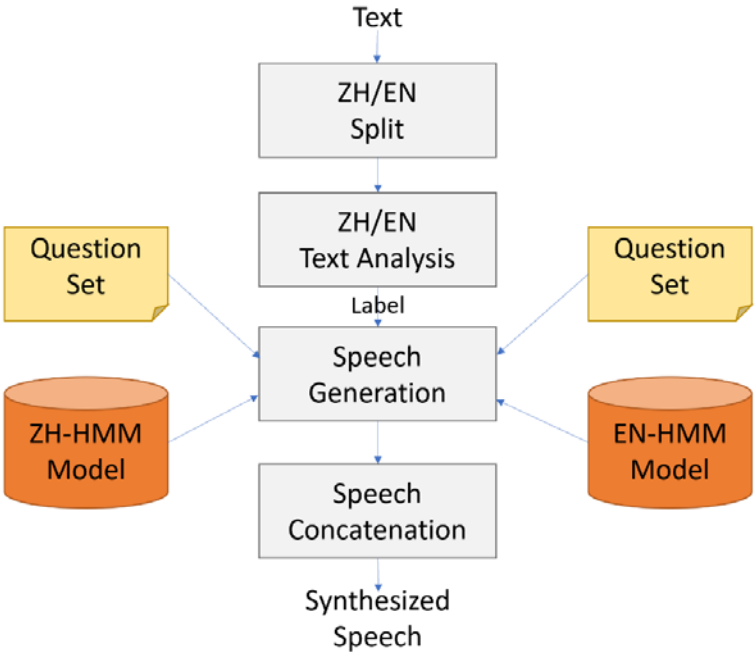
為了能表達五聲的變化趨勢，我們將中文的基本單元分割成三個音素模型:

C+V1+V2

其中 C 為第一個音素模型用來表示聲母的音，V1+V2 則同時用來表示韻母以及五聲變化趨勢的前(V1)後(V2)，以此方法最後我們可能如下圖 107 個音素模型(含一個 pause 模型)。相對於原本將所有中文約 1200 多個音的模型數量，此方法大大減少了模型的數量

表一、音素模型範例

	Syllable	C	V1	V2
ㄏㄨㄟˋ	huei4	hu	eiH	erL
ㄕˊ	shr2	shr	shrL	shrH
ㄧㄡˇ	yiou3	yi	ouL	ouL0



圖二、中英夾雜合成系統流程圖

(二)、英文模型定義

英文模型方面則直接基於 ARPAbet 標示並以國際音標(IPA)為準來做為我們的音素模型，其中包含 13 個元音、3 個雙元音、27 個輔音，共 43 個音。元音又因發音位置可細分至前、中、後元音，輔音也可依發音位置分成塞音、擦音、半元音、鼻音、塞擦音等。其中值得注意的是音標中的[l]、[m]、[n]雖然在音標內是同個樣子，但實際上根據是否在母音前，看似相同的音標會有不同的發音，這在 ARPAbet 標示法中需以不同的記號

做標記。

三、中英夾雜語音合成系統實作

本系統在合成文字前須先將原始文字做中英文字切割的前處理，在得到中英文後，先將整個句子中的英文以一個可識別的 SW 做替換並將整段文字丟入中文合成器直接合成出整段中文句子以保留整段中文的文脈資訊。英文單字則丟入英文合成器做合成，最後在將合好的英文單字取代前面的 SW 做語音串接，整個架構的流程圖如圖二所示。

(一)、中英夾雜文字切割

使用者輸入文字後(圖二. ZH/EN Split 的步驟)，本系統主要以 Unicode 來識別中英文字的區別。在 Unicode 中中文的範圍為 19968~40869，其餘部分包含半形空白當作英文部分(非中文)。

當輸入一段文字後，逐一由前至後針對每一個文字做中英文字判別，並使用一個變數來儲存當前文字屬於中文還是英文的狀態做紀錄，同時有兩個陣列分別來儲存由上一次狀態轉換前至現在的中/英文字收集器，當遇到狀態轉換時候就代表中英文字做了切換，此時就把所有文字收集器的中/英文丟入結果集內，當所有文字都判別完畢後，在把文字收集器內的結果放入結果集就將原始句子切割完成。

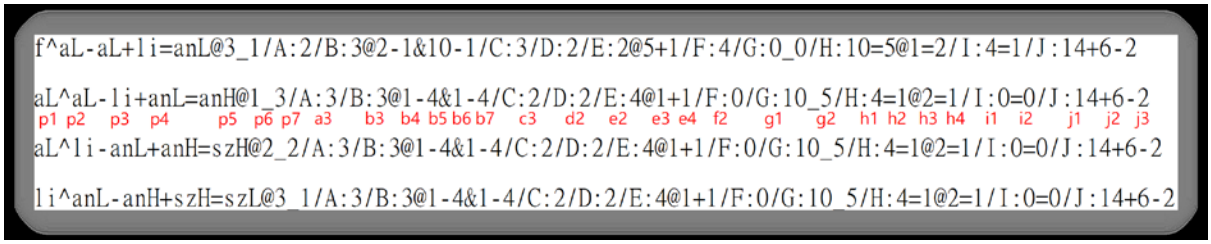
(二)、語音串接

1、文脈分析

為了使中英文連接處有著更自然的發音與韻律，我們必須保有整句中文字中的前後關係並記錄下來，因此在合成語音前必須先對文字做前處理(圖二. ZH/EN Text Analysis)，為此我們必須定義以下中文文脈資訊，並加以記錄成 Label 檔。細節如下所列: **音素(Extended Final):**在中文中即表示一個聲母/韻母/介音，如:ㄅ、ㄆ、ㄧ、ㄨ、ㄢ、ㄥ ；**音節(Syllable):**在中文中即表示一個字，如:我、剛、回、家；**音韻詞(Prosodic Word)** 多個音節組合成的一個詞，如: 螞蟻、無法、我們；**音韻短語(Prosodic Phrase):** 多個音韻詞組成的一小段話，通常兩個短語間會有停頓，所以短語很容易發生在連接詞上；**句子(Utterance):** 一個以上的短語組成一個句子，即我們要合成的目標語句。以上文脈資訊會依照其前前一個單元、前一個單元、當前單元、下一個單元、下下一個單元以及當前單元層級由前至後和由後至前數來的位置都納入考慮，並將以上的韻律階層可由分類回歸決策樹(Classification and Regression Tree, CART) [13]得到。

藉由以上文脈資訊記錄，我們可由 Label 檔中看出整句話的斷詞是如何斷的，在圖

三中，原句子為”家裡網路出了問題無法連伺服器”當中的’連’字，由 e2 可知當前音韻詞由四個組成，由 b4 可知他是由前數來第一個音節，即代表’連’在斷詞中被斷成”連伺服器”；英文方面則是在文脈分析上與記錄音節及重音(stress)方面[14]，與中文有著類似的記錄方法。



圖三、Label 檔範例

2、問題集

根據中文模型定義以及上述所記錄的文脈資訊，便可開始設計問題集之決策樹以讓模型達到最佳狀態，對此我們將對音素模型考慮以下五大類問題 [15]

(1) 音素相關(Phoneme related): 若為韻母 : 其音高範圍(H/M/L)；若為聲母 : 單一或由兩個音素組成(即是否含介音)；聲母發音類別 : 塞音、塞擦音、鼻音、擦音、邊音、唇音、舌尖音、舌根音、舌面音、翹舌音、齒舌音；韻母發音類別 : 單韻、複韻、聲隨韻、捲舌韻；音素在音節中的位置 : 由前/後數來位置。

(2) 音節相關(Syllable related): 音節中音素的數量 : 考慮前/當前/下一個音節；在音韻詞、音韻片語中的位置 : 由由前/後數來位置。

(3) 音韻詞相關(Prosodic Word related): 音韻詞中音節的數量 : 考慮前/當前/下一個音韻詞；音韻詞在音韻片語中的位置 : 由前/後數來位置。

(4) 音韻短語相關(Prosodic Phrase related): 音韻片語中音節、音韻詞的數量 : 考慮前/當前/下一個音韻片語；音韻片語在整段話的位置 : 由前/後數來位置。

(5) 句子相關(Utterance related): 句子中音節、音韻詞、音韻短語的數量。

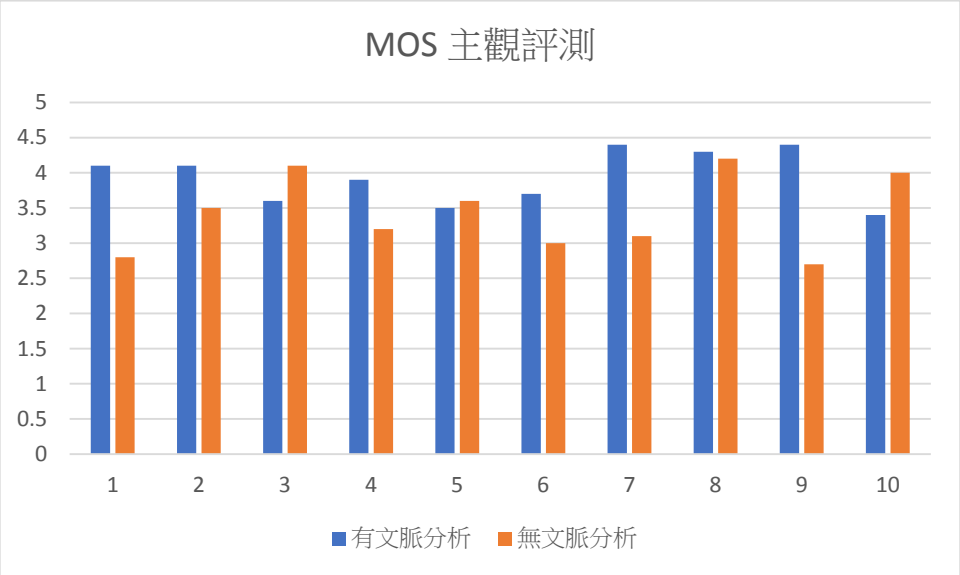
當其在訓練階段透過問題集分類並訓練完成模型後，在合成語音接段時就可再次從此問題集(圖 3. Question Set)中從所需合成文字的 Label 中找到其對應的模型並合成該語音，最後在將中英文部分合併，由英文字取代 SW 並得到最中的語音輸出結果。

四、實驗結果及討論

(一)、實驗目的與語料庫介紹

實驗將請受試者分別對每個句子來評論有分析並考慮前後文且記錄文脈訊息來一次合成整個句子後再替換掉 SW 所合出的語音，以及與沒有分析前後文並分開合成每段文字後直接合併的語音，比較哪一個句子較為順暢並評分，但由於實驗中英夾雜語句時容易使受試者混淆，所以本實驗僅以純中文並無將 SW 換回英文的方法來讓受試者接受本次測試。在實驗中所使用的語料庫分為中英兩部分。中文語料庫使用資策會所錄製的女性語者，主要內容為新聞語料，總共有 5102 個句子，92388 個字，英文語料庫方面使用 CMU ARCTIC 語料庫[16]。共有 1132 個句子，其中包含 10045 個單字(2974 個不重複的單字)，39153 個音素。

本次實驗目的在於確認有分析前後文文脈是否會影響句子的流暢性，評分方式採平均主觀值分數(Mean Opinion Score, MOS)，由 10 位受試者對 10 組隨機順序的不同的語句做評分，受試者成員包含同班同學、指導老師、社交平台上的朋友等，語音的自然度越高則給越高分(最高 5 分)，越不自然則給越低分(最低 1 分)，最後得到的 10 組句子的平均分數長條圖如圖四。其中，經由計算平均的結果後發現，有文脈分析較無文脈分析平均來得高(3.94 vs. 3.42)



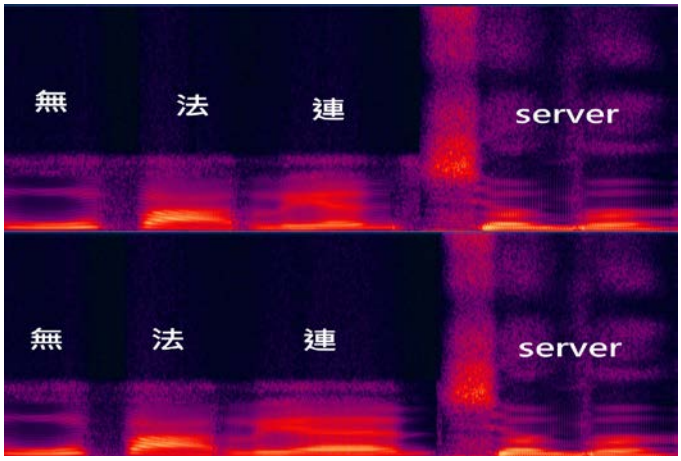
圖五、MOS 主觀評估

(二)分析與討論

從實驗平均和長條圖來看，多數句子都顯示有分析文脈後的結果較好。以下將用一個正式的中英夾雜句子作為範例來觀察頻譜圖和文脈中如何知道為什麼不分析前後文會導

致句子不自然。

在此使用圖四中文脈的句子”家裡網路出了問題無法連伺服器”，但實際上其實我們想合的是 Server 而不是伺服器，所以當使用者輸入”家裡網路出了問題無法連 Server”時候，我們的文字前處理器會先判斷 Server 是個名詞並換成一個作為名詞用的 SW 後再合成整句中文，此時文脈中的’連’被斷成”連+某個名詞 SW”，其被當為動詞後面有個受詞。但如果我們沒有使用文脈分析，則合成句子時候會先合成”家裡網路出了問題無法連”，再另外合成一個 Server 後直接做串接，此時因為’連’為最後一個字，後面並無受詞也沒有其他句子使其成連接詞，最後斷詞就將他斷成”無法連”，後面原本該有的名詞被落單成一個單字了，使的整句頓時失去了連貫，其兩者頻譜圖比較如圖六，圖中上半不為有分析文脈，下半部則無分析，可看出無分析文脈時，連與 sever 上出現了斷層。最後使的整句話念起來不通順，由此可見文脈分析對於一句話的重要性。



圖六、文脈分析頻譜圖比較

五、結論

本論文為中英夾雜語音合成系統，在一句中英混雜的語句以 SW 替換英文字並藉此合出整句考慮前後文關係的中文以保證句子的流暢性，最後在以合成的英文換掉替換用的 SW 來還原原本的中英夾雜句子。實作方法以隱藏式馬卡夫模型並在音素模型以及句子的文脈資訊使用問題集之決策樹來進行模型優化。

在實驗數據上顯示，本論文所提出的將句子整句合成在將字替換回原本的字，確實比每段字分開合成後在合併來的好，因為其保有了整句話的流暢性而非像是各個單字拼接而成。

參考文獻 [References]

[1] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, in *Acoustics, Speech and Signal Processing (ICASSP)*, vol.1, pp. 373-376, 1996.

[2] Yoshimura, Takayoshi, et al. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis." *Sixth European Conference on Speech Communication and Technology*. 1999.

[3] 江振宇, 黃啟全, 王逸如, 余秀敏, & 陳信宏. (2012). 可變速中文文字轉語音系統. *中文計算語言學期刊*, *17*(1), 27-41.

[4] Ju-Yun Cheng, Yi-Chin Huang, and Chung-Hsien Wu. "合成單元與問題集之定義於隱藏式馬可夫模型中文歌聲合成系統之建立 (Synthesis Unit and Question Set Definition for Mandarin HMM-based Singing Voice Synthesis)." *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*. 2013.

[5] 吳尚鴻. "基於隱藏式馬可夫模型之中文語音合成與吼叫情緒轉換." 清華大學電機工程學系所學位論文 (2010): 1-66.

[6] Chia-Ping Chen, Yi-Chin Huang, Chung-Hsien Wu, & Kuan-De Lee. (2014). Polyglot speech synthesis based on cross-lingual frame selection using auditory and articulatory features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1558-1570.

[7] Reddy, M. Kiran, and K. Sreenivasa Rao. "DNN-based Bilingual (Telugu-Hindi) Polyglot Speech Synthesis." *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018.

[8] Qian, Y., Cao, H., & Soong, F. K. (2008, December). HMM-based mixed-language (Mandarin-English) speech synthesis. In *2008 6th International Symposium on Chinese Spoken Language Processing* (pp. 1-4). IEEE.

[9] Zen, H., Nose, T., Yamagishi, J., Sako, S. and Tokuda, K., The HMM-based Speech Synthesis System (HTS) Version 2.0, 2007. http://hts.sp.nitech.ac.jp/

[10] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.Y., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., The Hidden Markov Model Toolkit (HTK) Version 3.4, 2006. http://htk.eng.cam.ac.uk/

[11] SPTK Working Group, “Reference Manual for Speech Signal Processing Toolkit Ver 3.3,”. http://sp-tk.sourceforge.net/

[12] T. Lin, and L.-J. Wang, “Phonetic Tutorials”, Beijing University Press, pp. 103-121, 1992.

[13] Hsia, C. C., Wu, C. H., & Wu, J. Y. (2010). Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 1994-2003.

[14] Zen, H. (2006). An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*, *133*.

[15] 謝雲飛, “語音學大綱”, 臺灣學生書局, 民國 63 年.

[16] Kominek, John, Alan W. Black, and Ver Ver. "CMU ARCTIC databases for speech synthesis." (2003).

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 378-387

基於有向圖與爭論導向摘要的網路辯論之爭論元素辨識

Identifying Argument Components in Online Debates through Directed Graph and Argument-oriented Summarization

魏奇安　Chi-An Wei
國立成功大學資訊工程研究所
Institute of Computer Science and Information Engineering
National Cheng Kung University
kevinwei3@gmail.com

高宏宇　Hung-Yu Kao
國立成功大學資訊工程研究所
Institute of Computer Science and Information Engineering
National Cheng Kung University
hykao@mail.ncku.edu.tw

摘要

辨識觀點在「辯論語意挖掘」的領域中扮演著重要的角色。在任意文章中辨識出論點，不僅可以用在被立場分類做為特徵來源增加辨識的準確性，同時也可以做為一篇文章表明其支持或反對一個目標的理由。在這項研究中，已經有許多方法嘗試解決這一項任務，較廣泛提起的是文字分類方法和自動化摘要的技巧。然而，將辨識觀點轉化成文字分類的問題不僅著重在特徵的選擇和使用，同時也因為將文章中一個一個句子拆出視為獨立的句子，所以也失去了文章整體的關係；自動化摘要的技巧雖然將文章中的所有句子一起處理評估出其中一句適合做為文章代表的句子，然而現有的一些方法多使用詞袋模型以及通用的自動化摘要方法，不僅沒有表示的特徵過於稀疏，同時亦沒有考量及引入辯論文章的特性。

本研究主要在可以保留文章為一體的自動化摘要方法上進行深刻的觀察及調整，使其不僅可以以辯論強度為導向摘要文章，並建構一有向圖來表示這些文章。在這些增強和調整下，不僅讓原本無辯論導向的自動化摘要方法可以更好的排序出辯論的句子。最終，本研究提出的方法可以在辯論文章中識別論點句子的任務上提升 8%的辨識準確率。

Abstract

Identifying argument components has become an important area of research in argument mining. When argument components are identified, they can not only be used for stance classification but also can provide reasons for determining an article is supporting or opposing about a specific target.

Previous research mainly used text classification and summarization techniques to solve this task. However, by transforming the task to a classification problem, not only rely heavily on choosing and using bag-of-words features, but also lose the article entity information due to extract the sentences out of the article and treat as an individual training instance. In the other hand, although summarization techniques handle on entire article and try to figure out which sentence can best represent the core concept of the article, in identifying argument components still heavily relies on bag-of-words feature representation and lack of argument-oriented features to concern about argument components characteristics.

In our study, we dive down to the core of the summarization method, not only makes it based on argument strength to summarize articles and identify argument components, but also proposed a directed graph construction approach. Experiments show that our proposed method outperforms 8% better than those without argument-oriented methods.

關鍵詞：辯論語意挖掘、辨識論點、自動化摘要、有向圖

Keywords: Argument Mining, Argument Components, Summarization, Directed Graph.

1. Introduction

In “Argument Components Extraction (ACE),” researchers try to get which sentence can be treated as arguing points, as known as “Argument Components (AC).” According to [1], which summarized tons of state-of-art knowledge about AM, summarized that how can a sentence become an argument component. The argument components in argumentative articles are usually formed by five types of sentences: “Claim” are the sentences that represent the statement being argued, “Data” are the facts or evidence used to prove the claim, “Warrant” are the sentences that make a connection between data and claim, “Backing” and “Rebuttal” are the sentences that support and against the warrant, respectively.

ACE is what our study is targeting on, to find out which are the key sentences that make

people explain why they support or against something. Furthermore, the results can feedback to SC task, since we know the key sentences that lead an article to support or against.

We are using the dataset released by [2], which collected posts under the domain “Abortion,” “Gay Rights,” “Obama” and “Marijuana” from an online debate forum. They manually labeled argument components in each post under each domain.

To accomplish the goal of ACE, previous methods commonly use bag-of-words or feature-based approach to represent the sentence in a feature vector form. These methods will cause the problems that lead to sparse feature, and it needs to treat sentences individually rather than processing with other sentences in the post.

We based on existed summarization method, TextRank to be one of our baselines. In [3], they aim to extract argument components by applying TextRank, which first using TF-IDF to represent each sentence then create an undirected graph, then applies PageRank on the undirected graph to acquire ranked sentences; In the end, the top-ranked sentence will be treated as argument components. However, we can say that the summarization algorithm was proved to perform well in extracting keywords or key sentences from an article, but cannot confidently say they are argument components.

In [4], which their work motivated us to integrate argument-oriented information in the graph, shows that changing the edge construction method can improve TextRank performance. We proposed an argument-oriented TextRank, ArguRank to address previously issued problems: With integrating subjectivity score to change the calculation within TextRank, we can confidently say that the result will be argument-ranked. Moreover, we proposed a directed graph construction approach to retrieve the nodes relation and direction, which aim to gain more performance by concern about how the score will be propagated.

To summarize, we make the following contributions:

- A model to retrieve subjectivity score of words through manual compiled argument-oriented corpus.
- An argument-oriented TextRank to identify the argument component.
- A directed graph construction approach to pursuing better ranking performance.

2. Related Works

2.1 Lexicon Expansion

The lexicon expansion approach aims to enlarge an existing lexicon using possibility, so it can not only be used by itself but can also lead a classifier or a model to learn its core concepts.

Previously, a lexicon-based method uses the attributes provided in the lexicon to do a specific task. The attribute in the lexicon usually becomes a feature that whether the target article contains the word or not, if contains, then the feature will switch to a state, otherwise will have another state. But more chances words in the target article will not appear in the chosen lexicon, make the feature is not as efficient as it should be. Researchers like [5] proposed a concept that using an existing lexicon as seed, and choose a model to learn the concept of the lexicon and then the model can try to predict attributes of words that not contain in the lexicon to maximize the feature that we want to retrieve from the lexicon. The lexicon expansion is accomplished in the following step, which proposed in [5]: (1) Choose a lexicon as seed, (2) Transfer words from text to its representation in vector space, finally (3) Construct and train a classifier or model to predict unknown words to retrieve the information the seed lexicon can give. The main workflow is shown in our research; we will use such this approach to enlarge an existing subjectivity lexicon to acquire words subjectivity strength probability for further procedure.

2.2 Graph-based Summarization

Another field and category to summarize documents are the graph-based methods. One of this kind is LexRank, proposed by [6]. It adopts TF-IDF to represent each words' importance in the sentences, then uses a modified cosine similarity equation to construct the edges between sentences.

The other approaches are TextRank and TextRank-based variation. [3] modified TextRank (shorten as PsTK) to rank sentences to determine which sentences are highly possible to be an argument component. The work constructs the graph for PageRank to iterate with sentences as nodes and similarity between each sentence as edges. In the original TextRank [7], it uses the following Equation 1 to calculate the similarity between sentences.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \quad (1)$$

After the edges are calculated, TextRank will use PageRank, proposed by [8] to iterate the graph and get the scores of each node, here we treat nodes as sentences in an article. In PsTK, they use a different way to construct the edges. However, the method cannot convince that summarization is a suitable method to solve the task, since none of the operation related to argument, stance or reasoning. In our work, we proposed an argument-oriented TextRank, ArguRank, which considered argument specific characteristics to identify argument component.

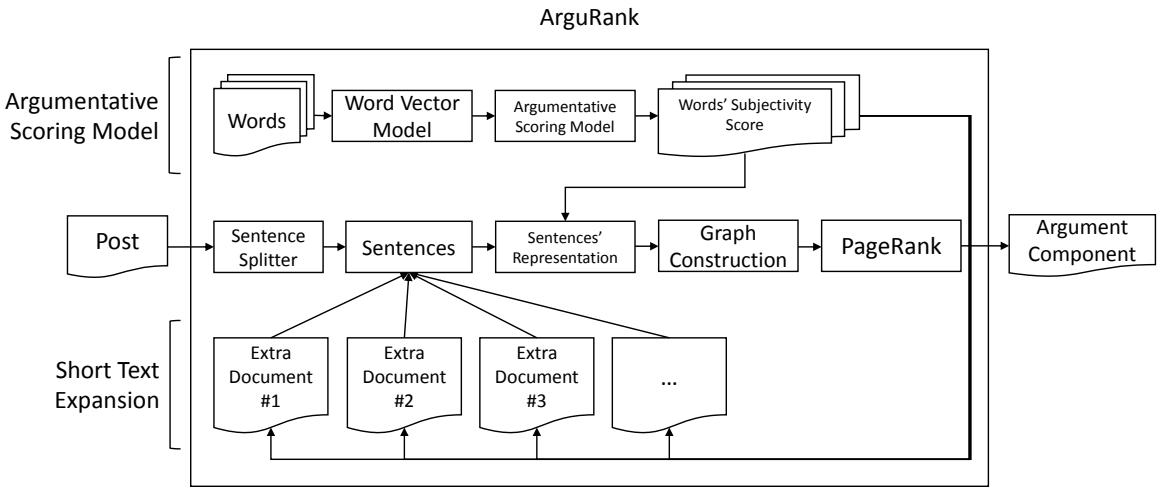


Figure 1. System framework and flowchart of our proposed method.

3. Method

We based on PsTK, the state-of-the-arts method on our dataset to develop our method, ArguRank. In the previous section, we address the issue of TextRank, motivated us to proposed argument-oriented TextRank, ArguRank, which aim to solve these issues and performs better in identifying argument components. Our proposed system will (1) read online debate posts, (2) preprocess the texts into sentences, (3) create sentence representation for calculating edge, (4) build a directed graph then (5) apply PageRank to scoring sentences. During this pipe which shows in Figure 1, we will apply our method to certain steps to make become argument-oriented TextRank, ArguRank.

3.1 Argumentative Scoring Model

To make TextRank argument-oriented, we develop a model to help enhance extract hidden argumentative information. The corpus we use to build the model is "Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon," was compiled by [9]. It contains 8222 words; each word was labeled as "strong subjectivity" or "weak subjectivity." To get the score of words that are not contained in the lexicon, we use the word vector model and binary classifier to build an argumentative scoring model to predict the argumentative score of the word. To construct the argumentative scoring model, we (1) use a word vector model to get its vector representation, then (2) feed to a binary classifier to make it learn how to separate the vector into strong or weak subjectivity.

3.2 Argument-oriented Summarization

After training the classifier, we use it to get subjectivity score of each word in each sentence. First, we normalized the subjectivity score of words in each sentence by a softmax function. The sentence representation will then be calculated via Equation 2.

$$\vec{S} = \sum_{i}^{K} \vec{V_{W_i}} a_i \tag{2}$$

Where $\vec{V}$ represents the word vector of W_i and $\vec{S}$ denotes the sentence representation before constructing the graph for PageRank.

3.3 Graph Construction

After we acquire various sentence representations, the next step is to construct the graph that represents the article. We choose cosine similarity to retrieve sentence relationship as edges, which shows in Equation 3.

$$\hat{E}(\vec{S_i}, \vec{S_j}) = \frac{\vec{S_i} \cdot \vec{S_j}}{||\vec{S_i}|| \cdot ||\vec{S_j}||} = \frac{\sum_{k=1}^{n} \vec{S_{i_k}} \cdot \vec{S_{y_k}}}{\sqrt{\sum_{k=1}^{n} \vec{S_{i_k}}^2} \sqrt{\sum_{k=1}^{n} \vec{S_{j_k}}^2}} \tag{3}$$

where $\vec{S_i}, \vec{S_j}$ are representations of two sentences.

After the edges been calculated, the graph G will be constructed. S_1, S_2, S_3 denotes the sentences respectively, and E_{12}, E_{23}, E_{13} indicates the similarity between sentences.

Based on the undirected graph, we then further apply two conditions to make it becomes directional. First is the condition to determine what edges are going to be discarded due to low similarity, which determined by $Threshold_{Dis}$ and shows in Equation 4.

$$E_{ij} = \begin{cases} \hat{E}(\vec{S_i}, \vec{S_j}), & \text{if } \hat{E}(\vec{S_i}, \vec{S_j}) \geq Threshold_{Dis} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Next, the construction of directional graph representation will be completed in two sub-steps. First, every sentence will have a direction that points to its next sentence, which shows in Equation 5.

$$E_{ij} = \begin{cases} E(\vec{S_i}, \vec{S_j}), & \text{if } j - i = 1 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Last, the sentences which have high similarity (determined by $Threshold_{Sim}$) will be used to applied direction that makes two nodes points together, which similar to Equation 4 and shows in Equation 6.

$$E_{ij}E_{ji} = \begin{cases} \hat{E}(\vec{S_i}, \vec{S_j}), & \text{if } \hat{E}(\vec{S_i}, \vec{S_j}) \geq Threshold_{Sim} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where the difference is the threshold will be interpreted as how similar of the sentences will have a bi-directional connection. The directed graph can be visualized in Figure 2.

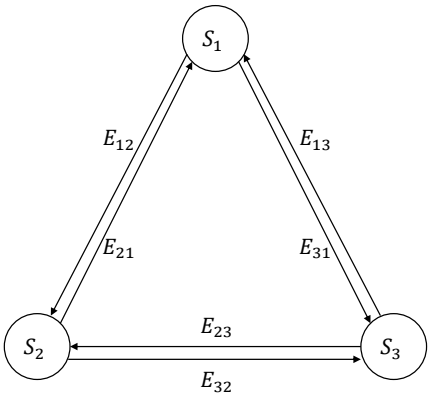


Figure 2. A directed graph that represents the article.

In the end, PageRank will applied on the directed graph to output the rank of these sentences. The top-1 ranked sentences will then be considered as argument components of the online debate article.

4. Experiments and Discussion

4.1 Evaluation Metrics

The methods we proposed required different evaluation metrics. For assessing the argumentative scoring model, first, we will use accuracy to evaluate how well the model can correctly predict whether the word in the lexicon that is strong or weak subjectivity.

Moreover, to prevent the illusion that the accuracy gives we adopt sensitivity (also called the true positive rate, or recall) and specificity (also called the false positive rate) to observe how the model performs on the answer in the strong and wrong subjectivity, respectively.

To evaluate ArguRank, we use accuracy again to assess how many articles in the dataset are correctly found the argument components. If the top-1 ranked sentence match one of the annotated argument components within an online debate article then the accuracy will raise.

4.2 Result Discussion

There are two main parts we are going to discuss in this section: (1) Building Argumentative Scoring Model and (2) Graph Construction.

4.2.1 Building Argumentative Scoring Model

We select three word vector models: Word2vec, GloVe, and fastText for comparison. The result shows in Table 1, which using fastText and debate posts plus Wikipedia articles as training corpus plus the original lexicon words can get acceptable performance on predicting subjectivity score of words than other methods. The model than further being used in applying the score of each word in sentence representation.

Table 1. Final argumentative scoring model where boldfaced scores show that better than others.

	Word2vec	**GloVe**	**fastText**
Accuracy	0.776	0.767	**0.777**
Sensitivity	**0.798**	0.789	0.767
Specificity	0.729	0.719	**0.798**
AUC	0.834	0.829	**0.86**

4.2.2 Graph Construction

In constructing a directed graph, due to the sentence splitting operation, some articles in our dataset exist only one sentence, which cannot construct the graph since no enough nodes. After filtering these kinds of articles, the filtered datasize is shown in Table 2, also with the performance of various directed graph construction approaches. By only drop connections on the undirected graph, the accuracy will get slightly improvement, and the performance will get better after direction construction with different granularity.

Table 2. The results of different graph constructions using cosine similarity. The different number of size about valid dataset in use to construct graph is also shown in here.

	ABO	**GAY**	**OBA**	**MAR**	**avg**
PsTK	0.469	0.528	0.582	0.641	0.555
ArguRank	**0.490**	**0.538**	**0.628**	**0.693**	**0.587**
Improvement	0.021	0.01	0.046	0.052	0.032

4.2.3 Positions of Argument Components

We apply our method to the datasize that removes the posts which one of the argument components is in the first sentence and PsTK predicted. The reason to remove these posts is in PsTK or other methods based on undirected graph, it will face a problem if the ranked scores are the same, the algorithm will predict the first sentence of the post to be the argument components. To show how our directed graph construction approach can deal with such this problem, we experimented on the datasize after the removal. In Table 3, we run the directed graph construction methods on the datasize and filtered graph size; the results show that by representing sentence by specific weighting mechanism and the directed graph construction, our method can identify argument components better than PsTK.

Table 3. After removing the posts that one of its argument components is located at the first place of sentences.

	ABO	**GAY**	**OBA**	**MAR**	**avg**
PsTK	0.430	0.475	0.512	0.568	0.496
ArguRank	0.503	0.553	0.609	0.634	0.575
TextRank + ArguRank	**0.515**	**0.548**	**0.606**	**0.637**	**0.577**
Improvement	0.085	0.073	0.094	0.069	0.081

5. Conclusions

We based on TextRank to develop an argument-oriented and directed ranking method called "ArguRank," which makes TextRank argumentative and directed. Also, we show how we build our research environment to expand a lexicon for identifying argumentative words and construct an argument representation.

The experiments show the proof that using argument-oriented graph-based summarization method by applying the subjectivity lexicon to construct the sentence representation can get better result on extracting argument components. Moreover, the approach of directed graph construction significantly improves the performance of identifying argument components via graph-based summarization.

References

[1] da Rocha, G.F. *ArgMine: Argumentation Mining from Text*. 2016; Available from: http://hdl.handle.net/10216/89719.

[2] Hasan, K.S. and V. Ng. *Why are you taking this stance? Identifying and classifying reasons in ideological debates*. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[3] Petasis, G. and V. Karkaletsis. *Identifying argument components through TextRank*. in *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. 2016.

[4] Barrios, F., et al., *Variations of the similarity function of textrank for automated summarization.* arXiv preprint arXiv:1602.03606, 2016.

[5] Bar-Haim, R., et al. *Improving claim stance classification with lexical knowledge expansion and context utilization*. in *Proceedings of the 4th Workshop on Argument Mining*. 2017.

[6] Erkan, G.u., nes and D.R. Radev, *Lexrank: Graph-based lexical centrality as salience in text summarization.* journal of artificial intelligence research, 2004. **22**: p. 457-479.

[7] Mihalcea, T., P.T. Textrank, and others. *Bringing order into texts*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2004.

[8] Page, L., et al., *The PageRank citation ranking: Bringing order to the web*. 1999, Stanford InfoLab.

[9] Wilson, T., J. Wiebe, and P. Hoffmann. *Recognizing contextual polarity in phrase-level sentiment analysis*. in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. 2005.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 388-397

利用 Attentive 來改善端對端中文語篇剖析遞迴類神經網路系統

Using Attentive to improve Recursive LSTM End-to-End Chinese Discourse Parsing

王育任　Yu-Jen Wang
國立中央大學資訊工程學系
Department of Computer Science & Information
National Central University
leowang@g.ncu.edu.tw

張嘉惠　Chia-Hui Chen
國立中央大學資訊工程學系
Department of Computer Science & Information
National Central University
chiahui@g.ncu.edu.tw

摘要

篇章剖析，可以幫助我們以不同角度來理解文句之間的關係與連結，但篇章剖析資料結構目前仰賴人工標記，使這項技術無法直接利用在任意篇章中。因此至目前為止，有許多研究著手於讓電腦能夠自動對篇章進行篇章剖析，並建構出一個完整的剖析樹。以中文語料庫 CDTB 來說，欲建立完整的篇章剖析程式，其問題主要可以被分成四項，分別是子句分割、剖析樹建立、子句關係辨識與中心關係辨識。由於深度學習近幾年發展快速，因此針對篇章剖析的建構方法也從傳統的 SVM, CRF 等方法，進展到目前以遞迴類神經的方式來建構剖析篇章程式。在本篇論文中，我們使用了許多目前最新的深度學習技術，例如 Attentive RvNN、self-attentive、BERT 等方法，來提高模型的準確度。最後，我們成功將每一項任務的 F1 都提高了近 10% 左右，達到目前我們所知研究中最好的效能。

Abstract

Discourse parser helps us understand the relationship and connection between sentences and sentences from different angles, but the tree structure data still need to rely on manual marking, which makes this technology unable to be directly used in daily life. So far, there have been many research and studies on how to automatically construct the complete tree structure on the computer. Since deep learning has progressed rapidly in recent years, the construction method for discourse parser has also changed from the traditional SVM, CRF

method to the current recursive neural network. In the Chinese corpus tree library CDTB, the parsing analysis problem can be divided into four main problems, including elementary discourse unit (EDU) segmentation, tree structure construction, center labeling, and sense labeling. In this paper, we use many state-of-the-art deep learning techniques, such as attentive recursive neural networks, self-attentive, and BERT to improve the performance. In the end, we succeeded in increasing the accuracy by more than 10% of F1 in each task, reaching the best performance we know so far.

關鍵詞：深度學習, 篇章剖析, 注意力機制, 遞迴類神經網路

Keywords: Deep Learning, Discourse Parsing, Attention, RvNN

一、緒論

篇章剖析主旨在分析文本之間的關係和結構信息，最終能建構成完整的剖析樹，至目前為止，篇章剖析語料庫有許多不同的體系，例如對目前研究影響最深的 Rhetorical Structure Theory (RST)體系[1]，擁有更多標記資料且結構更為自由的 Penn Discourse Treebank (PDTB) 體系[2]，或是以中文篇章結構為主的 Chinese Discourse Treebank (CDT) 體系[3]，每一種體系在標記的定義與見解皆不同，所建立起來的關係結構也截然不同，因此相同的篇章在不同體系之間，資料可以視為一種互補的關係。由於篇章剖析中豐富的關係標記，使得篇章剖析資料運用在許多不同的研究上，皆獲得更好的準確率，但篇章剖析語料庫仰賴人工標記，使得篇章剖析的許多研究無法直接使用在未經結構化的資料上，因此我們的研究以 CDT 體系為主，目的是要建立一個中文的剖析程式，將串列結構的篇章轉換成剖析樹結構的篇章資料，並標註出結構關係。

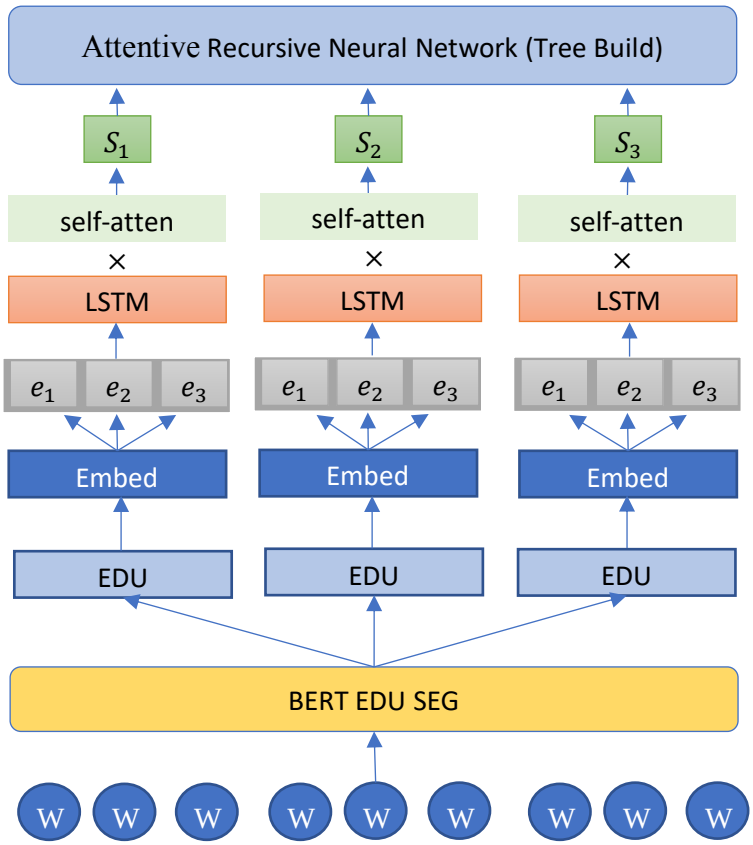
目前已經有許多針對篇章剖析進行的研究，例如 SVM Base Parser [4]、DCRF Base Parser [5]、Recursive Deep Learning [6]、Transition-Based Dependency Parser [7]，在這些研究中，將篇章剖析主要分成兩種問題，分別是子句分割與剖析樹結構建立，並將這兩項問題視為分類問題，但這些研究有以下問題無法直接應用在我們的研究上，其一是這些研究皆是針對英文進行剖析程式的建立，研究中設計了許多英語的文字特徵難以複製到中文剖析研究上，其二是大部分的研究不做子句分割的問題，而是以正確的子句當作輸入來進行訓練，而剖析模型中子句分割的好壞很重要，這會直接影響後續所有問題的效能表現，因此忽略子句分割的剖析程式，不但無法實際運用在現實生活中，也不符合我們的要求。

2018 年 Lin 提出了以中文為主的 RvNN Base Parser [8]，此研究使用中文語料庫 CDTB 來進行建構，不需要擷取額外的文字特徵，且包含完整端對端剖析結構，此研究是據我們所知，第一個針對中文的端對端篇章剖析開源程式，且達到目前我們所知中文剖析程式中最高的準確率，但研究中還有許多不足的地方，因此受到他們的啟發，我們希望以此研究架構為主軸，在效能與程式結構上進行優化，使之準確率更貼近人類的標準。Lin 於這份研究中，將篇章剖析分為四個步驟，分別是子句分割、樹結構構建、主次關係標示與篇關係辨識，在研究中使用標點符號對子句進行分割，接著輸入 Recursive neural network (RvNN) 對節點子樹資訊進行擷取，並對不同問題進行預測，最後使用 CYK 演算法做剖析樹的建構。我們針對 Lin 的架構做出以下四種優化：(1) 在訓練階段 Lin 並未使用任何預先訓練的 word embedding，但好的 word embedding 對模型的訓練極其重要，因此我們加入了 FastText embedding 來幫助訓練。(2) 一個子句中，每個字詞之間的重要程度皆不同，如果能找出字詞之間的權重資訊將能有效幫助模型的訓練，在程式中我們加入了 self-attentive[9]，讓模型能在不同順序點上學習文字或句子的重要程度。(3) 在 2016 年 Zhou 等人於 COLING 發表一篇論文[10]，將 Attentive 機制加入 RvNN 裡，幫助 RvNN 學習節點下每個子樹的重要程度，並且實驗在四個不同問題中，相較於加入 Tree Attention 前，獲得更好的效能。因此我們相信，一個樹的向量推導並不需要依賴於所有的子樹，樹狀結構中一定會有某個部份的子樹相對於其它子樹影響更大，我們並不需要依賴全部子樹來得到節點的訊息，因此我們將 attentive 機制也加入 RvNN 中，讓 RvNN 在訓練過程中找出相對重要的子樹，幫助模型進行訓練。(4) Lin 在模型中以標點符號作為子句分割的依據，但此方法的效果不理想，子句分割的 F1 比 Lin 實驗中的 baseline 還要低，因此我們將此問題看作是 sequence labeling problem，以目前做此類問題效能最好的 BERT [14] 進行子句分割實驗，並達到目前子句分割最好的效能。

二、模型設計

在這章節，我們會先針對模型的主要架構作解釋，隨後會再針對模型中個別使用到的深度學習技術做介紹。本論文架構主體以 RvNN 來進行訓練，目的是建構出一個端對端的中文剖析程式，我們程式的架構和資料準備方式參照了 Lin 所提出的方法，並在此架構上進行更進，針對我們更進的部分，我們會在次章節中進行詳細介紹。

Lin 對輸入的句子使用標點符號做子句分割，而後輸入 LSTM 擷取子句訊息，最後輸入 RvNN 內進行節點資訊提取，並建立剖析樹。我們與之不同，如圖一，將輸入的句子使用預先訓練好的 BERT 模型做子句分割，子句長度為 n，使用 FastText 做為文字的 embedding，每個字會映射到一個大小為 300 的陣列，每個子句會以字為單位轉換成大小為$300 \times n$的矩陣，則每個子句則可以表示為 $e^i = (e_1^i, \dots, e_n^i)$。之後將子句輸入 LSTM 學習，同時使用 self-attentive layer 算出每段子句文字資訊的重要程度，對文字依據不同的權重做加權，得到計算過後的子句資訊S_i，此資訊作為我們 Attentive RvNN 的輸入，並且以下到上的方式，使用 CYK 演算法針對子句做結合，最後組建出一顆完整的剖析樹。



圖一、篇章剖析訓練結構

我們採用與 Lin 相同的方式計算 RvNN，針對剖析樹的建構、主次關係與篇關係辨識等問題，會將 RvNN 對目標節點所輸出的隱藏層 $\vec{h}$ 與狀態層 $\vec{c}$，使用 SoftMax 分類器來算出不同問題選項的機率，並選取機率最高的選項做為答案，機率公式如下

$$\vec{p} = softmax\left(W_s\left[\begin{matrix}\vec{h}\\ \vec{c}\end{matrix}\right] + \overrightarrow{b_s}\right) \quad (1)$$

使用了三個 SoftMax 分類器來針對三種不同的問題做輸出，並同樣使用 CYK 演算法對剖析樹進行建構。

（一）子句分割

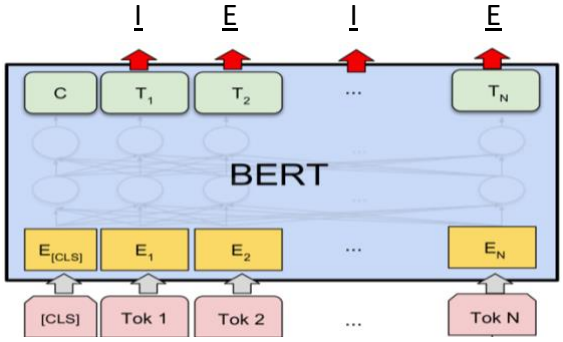
建構 CDTB 中，子句分割極為重要，子句分割點一定是位於標點符號上，但並非只要出現標點符號，句子就必須做分割。如下圖表一的句子，第一個逗號為子句分割點，但第二個逗號並不需要做子句分割。因此我們將子句分割獨立出來，視為一個序列標記問題，並且分別使用 BERT Fine-Tuning 此種在序列標記上非常熱門的方法來做訓練。

訓練中我們以 BERT 語言模型作為基礎，訓練資料字為單位做分割，使用 I 代表的 inside 與 E 代表的 End 兩種標籤來做標記，標記方式如表一，子句分割段落以 E 做標記，其餘皆以 I 來做標記。

表一、序列資料標記範例

句子	他	非	常	認	真	，	所	以	，	把	重	要	任	務	交	給	他	。
標籤	I	I	I	I	I	E	I	I	I	I	I	I	I	I	I	I	I	E

在 BERT Fine-Tuning 模型當中，我們使用 Google 所提供 12-layer, 768-hidden, 12-heads, 110M parameters 預先訓練好的中文模型作為我們的初始模型，並在此之上學習序列標記問題，我們會將輸入的文字轉換成 BERT 所定義的 Token，再將此 Token 輸入模型中進行訓練，模型架構如圖二



圖二、BERT 模型架構

（二）Self-Attentive

在 2017 年 Lin, Feng 等人於 ICLR 提出了 Self-Attentive 的機制，Self-Attentive 機制不需依靠外界的額外資訊，即可在句子內部進行 Attentive 的學習，尋找序列內部資訊的聯繫。因此，在我們的模型中，也加入了此機制，使用二维矩陣來表示句子，矩陣的每一行則表示句子的不同關係。公式如下：

$$s = (w_1, w_2 \ldots . w_n) \tag{2}$$

$$h_t = LSTM(w_t,\ h_{t-1}) \tag{3}$$

$$H = (h_1, h_2\ \ldots h_t) \tag{5}$$

$$a = softmax(W_2 \tanh(W_1 H^T)) \tag{6}$$

模型的輸入資料 s 含有 n 個資訊的序列，其中 w 代表序列的 word embedding，我們會將此資訊 s 輸入一個 LSTM，得到一個可用於計算的 h_t，LSTM 的隱藏單元為 u，則 $h_t \in R^u$，而 H $\in R^{n*u}$ 則表示為所有隱藏狀態h_t的集合，之後將 H 帶入 self-attentive 層中，最終使用 SoftMax 對每一段文字做歸一化，$a \in R^{1*n}$，此時 a 的每一維度可以認為是對應位置文字的 attentive，至此 Self-Attentive 完成。

（三）Attentive RvNN

在樹狀結構中，子樹之間彼此的資訊是有關聯的，因此將 Attentive 機制加入 RvNN 中，能讓模型更好的學習不同子樹之間的重要性。

對於我們 RvNN 輸入的隱藏層$\overrightarrow{h_s^1}, \overrightarrow{h_s^2}$，我們在計算新的隱藏層之前，會針對 $\overrightarrow{h_s^1}, \overrightarrow{h_s^2}$ 做 Attentive ，公式如下：

$$m_{k1} = \tanh\left(W^{(m_{k1})} h_k + U^{(m_{k1})} S\right) \tag{7}$$

$$m_{k2} = \text{relu}\left(W^{(m_{k2})} m_{k1}\right) \tag{8}$$

$$m_{k3} = \text{relu}\left(W^{(m_{k3})} m_{k3}\right) \tag{9}$$

$$a_k = \frac{\exp(w^t m_{k3})}{\sum_{j=1}^{n} \exp(w^t m_{k3})} \tag{10}$$

$$g = \sum_{1 \le k \le n} a_k h_k \tag{11}$$

h_k為我們要學習的任一子樹的隱藏層，S 為一個額外資訊，定義為子句經過 LSTM 後得到的資訊，我們將子樹的資訊h_k與 S 做訓練得到子樹之間的權重 a，最後將 a 乘回相

對應的子樹完成 Attentive。而此公式與 Zhou 等人在原始論文提出的公式略有不同，由於在原始方法中只使用了一個 m_{k1} 做學習，但我們發現只使用一個 m_{k1} 效能不盡理想，為了增強注意力機制的學習成果，我們增加了多層神經層 m_{k2}, m_{k3} 到模型中幫助訓練，且在這個階段，我們的資訊 S 就不參與學習。

三、實驗

在實驗中，由於我們模型的原架構來至 Lin，為了更好的與 Lin 的成果做比較，我們使用了與 Lin 相同的實驗方式，使用 standard evaluation tool PARSEVAL [11] 來作為樹狀結構 F1 的計算方法。在 Lin 的實驗中，將 Kang 等人於 2016 年提出來的實驗結果作為 Baseline，我們則將 Lin 與 Kang 的結果一起當作 Baseline 與我們的模型做比較。

實驗一中，我們針對完整的端對端篇章剖析程式做測試，並對樹狀結構建構、主次關係標示與篇章關係辨識做測試，實驗中我們的模型皆加入 FastText Embedding 和 Self-Attentive Layer，並針對是否使用 Attention RvNN 與 BERT 子句分割做比較。

實驗結果如表二，我們的模型使用模型在使用 Attentive-RvNN 後，準確率反而相較於未使用 Attentive-RvNN 來得低，而模型在加入 BERT 做子句分割後，在子句分割問題(EDU)上獲得最高的準確率，受惠於此，各項問題在後續的表現上也相繼獲得準確率的提升，得到所有研究中最好的準確率。

表二、端對端篇章剖析效能

Mod	EDU	Structure	Sense	Center	all
Kang	93.8%	46.4%	28.8%	23.1%	20.0%
Lin	87.6%	50.7%	27.8%	25.7%	22.2%
RvNN	88.6%	54.3%	35.3%	34.2%	30.8%
Atten RvNN	87.8%	55.1%	34.4%	33.1%	30.1%
BERT RvNN	**94.6**%	**57.7**%	**37.2**%	**36.0**%	**31.9**%

在前實驗一中，模型在使用 Attentive-RvNN 後，準確率反而相較於未使用 Attentive-RvNN 來得低，我們可以從下列表三來理解。Lin 的程式架構是以二元樹為建構目標，但 CDTB 本身為多元樹的結構，使用原始的多元樹來當作測試資料，並無法真實體現 Attentive-RvNN 所帶來的學習效能，因此我們將測試資料轉成二元樹後可發現，使用 Attentive-RvNN 的模型在二元樹的測試資料中，的確能幫助模型提高準確率，但模型的

效果並不能顯現在未經訓練過的多元樹結構上，因此效能在第一個實驗中被低估。

表三、標準子句二元樹與多元樹篇章剖析實驗

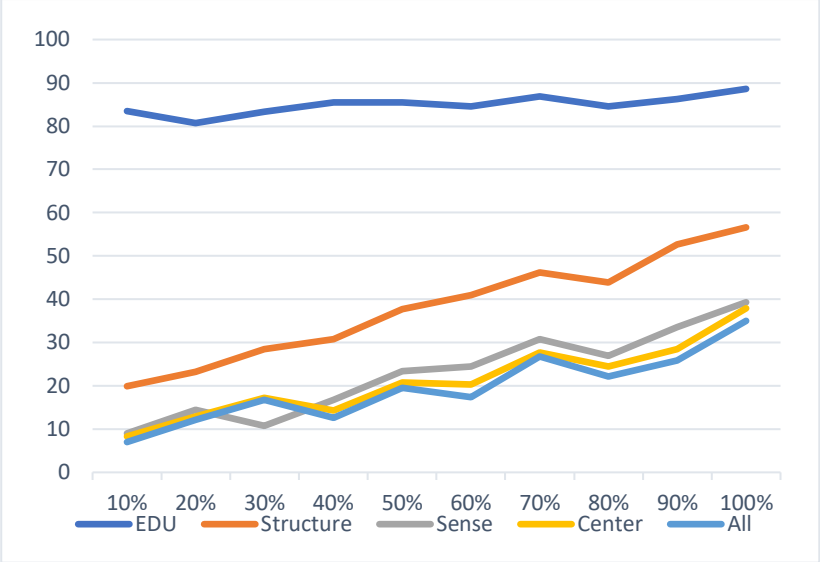
Tree Type	Model	Structure	Sense	Center	Overall
多元樹	RvNN	**63.4%**	**40.7%**	**40.5%**	**35.1%**
	Attentive-RvNN	62.4%	39.9%	38.1%	33.7%
二元樹	RvNN	69.2%	48.8%	**48.5%**	**44.2%**
	Attentive-RvNN	**69.9%**	**49.3%**	47.6%	**44.2%**

在下表四中，我們模型針對四種子句關係的準確率做比較，並將 Lin 的準確率列在最右側做參考，可以看到在我們的模型中，並列類與解說類獲得最好的準確率，而表中轉折類擁有所有類別中占比最高顯性連接詞，但卻因為訓練資料少於其他類別，所以效能表現明顯不足。

表四、子句關係分析表

Sense	Node (Gold)	Node (Pred)	True Positive	F1	Lin_F1
並列類	414	369	192	49.0%	67.0%
因果類	119	44	14	17.1%	16.5%
轉折類	151	146	54	36.3%	29.7%
解說類	11	11	8	72.7%	0.0%

最後，我們將訓練資料平分成十等份，每次訓練皆增加 10%的訓練資料，並在二元樹的測試資料下進行測試。從圖三中可以看出，子句分割在不使用 BERT 的情況下，準確率已經達到極限無法再增長，但樹狀結構建置、關係標記與中心標記還是成持續增長的狀態，因此我們認為模型目前尚未達到效能的極限，若能標記更多訓練資料，此模型的準確率會有更好的成果。



圖三、學習曲線

四、結論

本篇論文中，我們針對 RvNN Base 的中文端對端篇章剖析程式進行改進，並使我們的剖析程式達到目前最好的效能。我們加入的 word embedding 能有效幫助模型針對語言詞向量做學習，且可以避免使用過多人為產生的特徵來影響模型的訓練。而我們使用了 self-Attentive layer 與 Attentive-RvNN，在實驗中也證實，此方法能有效地幫助程式對於文句或子樹不同部分的重要程度做學習，以提高模型準確率。最後我們使用 BERT 的方法改善子句的分割，使模型的整體準確率有顯著的提升。以本實驗學習曲線的結果來看，此模型的效能受限於訓練資料，尚未達到最好的準確率，倘若能夠繼續增加訓練資料，模型將會有更好的表現。

參考文獻

[1] University of Chicago, *The Chicago Manual of Style Online*. Chicago, IL: University of Chicago Press, 2006. [Online].

[2] Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation.

[3] Li Yancui, Feng Wenhe, Sung Jing, Kong Fang, Zhou Guodong. Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure[C]. In Proceedings of the 2014 conference on Emporical Methods in Natural Language Processing, pages 2105–2114.

[4] David A. duVerle , Helmut Prendinger .A Novel Discourse Parser Based on Support Vector Machine Classification. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 665–673

[5] Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pages 486–496

[6] iwei Li, Rumeng Li and Eduard Hovy. Recursive Deep Models for Discourse Parsing. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2061–2069

[7] Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with LSTMs. In EMNLP'15, Lisbon, Portugal. 349–359.

[8] Chuan-An Lin. A Unified RvNN Framework for End-to-End Chinese Discourse Parsing. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 73–77 Santa Fe, New Mexico, USA, August 20-26, 2018.

[9] Zhouhan Lin , Minwei Feng, Cicero Nogueira dos Santos, Mo Yu,Bing Xiang, Bowen Zhou& Yoshua Bengio. A S TRUCTURED S ELF-ATTENTIVE S ENTENCE EMBEDDING. Published as a conference paper at ICLR 2017.

[10] Yao Zhou, Cong Liu, Yan Pan. Modelling Sentence Pairs with Tree-structured Attentive Encoder. 10 pages, 3 figures, COLING2016.

[11] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in theframe work of rhetorical structure theory. In Proceedings of the Second SIGdial Workshop on Discourse and Dialogue (SIGDIAL'01), pages 1–10.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 398-407

Four-word Idioms Containing Opposites in Mandarin[1]

Siaw-Fong Chung, National Chengchi University

sfchung@nccu.edu.tw

Abstract

This study aims to investigate the occurrences of opposites in Mandarin four-word idioms. We used the Revised Ministry of Education Online Dictionary to search for a total of 12 combinations of opposites in four-word idioms based on 32 opposites. These idioms were analyzed in terms of their dictionary definitions, combinations, and internal structure. Based on the definitions and the POS information on the components in the online dictionary, as well as the combinations of opposites, we collected 910 types of idioms. Our analysis used quantitative method to analyze not only idioms but idioms containing opposites, and our study is one of the few studies that have utilized this approach.

1. Introduction

There have been multiple studies on Mandarin idioms, but those that have inspected contrasting words in idioms have seldom been carried out in the past. Wang, Wu, Li, Huang, and Hui (2010: 501) [1], one of the few studies that conducted a corpus-based analysis of the co-occurrence of Chinese opposites, proved that little research on the issue of Chinese antonymy has been performed in the past[2]: "We searched 'antonymy' and 'corpus|corpora' as keywords in the CNKI, an e-resource containing all Chinese journals, but no matches were retrieved. Therefore there is a need for further analysis of Chinese antonymy in this way." Our goal was to examine opposites in Mandarin four-character constructions and the same

[1] This paper was supported by the Ministry of Science and Technology grants 104-2420-H-004 -003-MY2 and 108-2410-H-004-095-. Assistance from Wei-yu Chen and Yin-Wen Li was appreciated. The research outcome of this work was discussed at the 成果發表會of the Joint Research project 臺灣語言詞彙、構式及語意浮現機制：跨界的探討. Comments from the research team and Prof. Jung-hsing Chang were highly appreciated.
[2] We used "opposites" as a more general term instead of "antonyms" in this paper, but we kept Wang et al.'s (2010) [1] term in the literature review.

observation was found—little research has been conducted on this topic. Our search in several major databases (e.g., EBSCOHost, ProQuest, CNKI, Google Scholar, etc.) using the keywords 'corpus' AND 'four-character idiom'/'idiom'/成語 (*chengyu*) AND 'antonym'/ 'opposites' returned hardly any satisfactory results. This indicated that either the research of Mandarin opposites has evoked less interest or opposites have not been looked at because they are too straightforward.

In this paper, we will show how a study on Mandarin opposites can be interesting and is worth running. The aim of the study was to investigate idiomatic four-character lexical constructions containing opposites in Mandarin (hereafter, four-word idioms). The present study intended to answer the following questions presented in (1) below:

(1) (a) What are the patterns of the opposites in Mandarin four-word idioms?

(b) What kinds of meanings are carried by the opposites in the four-word idioms?

The occurrence of paired opposites in the structure of four-character idioms is an issue worth looking into in depth. The meaning changes and effects due to the construction of the paired opposites are also worthy of closer observation. In what follows, we will review past literature on antonyms/opposites and Chinese four-word idioms.

2. Literature Review

There has been much discussion on opposites in past literature, but mostly regarding their constructions. It is also important to know that opposites are not always in opposition. Pioneered by Jones (2002) [2], Jones, Murphy, Paradis, and Willners (2007 [3], 2012 [4]), and Murphy, Paradis, Willners, and Jones (2009) [5], a new view of antonyms was postulated—antonyms can be identified by a certain construction in the corpus (e.g., *big and small; rich and poor*) as well as comparatively (*he was more feminine than masculine*) (Jones [2002] [2] used antonyms instead of opposites).

As for Chinese idioms, Wu (1992:10-12 [6]; 1995:65 [7]) defined Chinese idioms (成語)

as “an old expression that has prevailed in society for a long period of time” and “[t]he meanings of most idioms can be deducible from their constituents. With some of the idioms, integrated meanings are unlikely to be directly inferred from their constituents” (p. 10-11).

謝健雄 (2006) [8] analyzed the number of syntactic combinations in《超活用成語大辭典》by 遲嘯川 and 許易人 (2002) [9] and found that four-word combinations constituted 98.6% in that dictionary, while the remaining four-word combinations were longer expressions. In Wang et al. (2010) [1], they investigated the formations of Chinese antonym pairs and phrases using several fixed frames and found six significant patterns in Chinese, which are shown in Table 1 below. We added the examples from 陳曉燕 (2004) [10] and 韓漢雄 (1993) [11], which Wang et al. (2010) [1] cited in their paper but with no examples.

Table 1: Patterns of Chinese antonyms

Features of Patterns	**Patterns**	**Examples**
Interleaved with a second antonym pair	X+Y+!X+!Y	送舊迎新 陽奉陰違
Two juxtaposed antonym pairs	X+!X+Y+!Y	悲歡離合 浮沉起落
Interleaved with synonymous words	X+Y1+!X+Y2	生離死別 冷嘲熱諷 南腔北調
Followed by a fixed phrase	X+!X+m+n	悲喜交集 賞罰分明
Interleaved with grammatical particles	半/忽/亦+X+半/忽/亦+!X	半信半疑 忽冷忽熱 亦喜亦憂
Repetition of antonym pairs	X+X+!X+!X	吞吞吐吐 進進出出 斷斷續續

Note: An exclamation mark marks the opposite meaning.

Wang et al. (2010) [1] also examined 371 antonyms proposed by 譚達人 (1989) [12], which were extracted “partly from dictionaries and partly intuitively” (p. 502). Examples of these antonyms include 褒貶, 東西, 多少, 教學, 來往, 冷熱, 利害, 兄弟, etc. (p. 502). Wang et al. (2010) [1] first checked whether these antonyms appeared in reverse order (e.g., *貶褒) in the Corpus of the Center of Chinese Linguistics (CCCL, Peking University), but none of them were reversed. Nonetheless, examples such as 怯勇 and 瘦胖 could be

reversed (p. 503). They then delimited the X and Y elements in Table 1 to be 東西, 南北, 上下, 左右, 前後, 遠近, 裡外, and 內外. Next, they used the frame "a+X+b+!X+c, in which, a, b and c denotes one Chinese character and !X is the opposite of X" (p. 503). This frame contains two pairs, namely, "a+X+b+!X and X+b+!X+c" (p. 503), in which the "a+X+b+!X" pair is called the B1 Pair and the "X+b+!X+c pair" is called the B2 Pair (cf. Jones, 2002) [2]. This part of the results returned examples such as 宮裡宮外 and 對裡對外 (as few examples were provided, these two were the ones Wang et al. [2010] [1]used to discuss incorrectly tagged examples). This pattern of antonyms, as shown in a lexical construction, was close to the ones targeted by the current study. Nonetheless, our analysis was more refined, which will be elaborated in the methodology section.

Apart from the research above, studies that have focused on Chinese four-word idioms only are far too many to be cited individually. Those that dealt especially with opposites are sparse, while some were conducted in a less systematic way, but many provided random mentions of examples. Qiu (2015: 84) [13] examined the existence of numbers in Chinese four-character idioms, such as (a) 同語素類疊 (e.g., 七早八早 and 賊頭賊腦); (b) 異語素鑲嵌－ (e.g., 一問一答); and (c) 數字式套語 [~A~B], [A~B~], [A~A~] (e.g., 挑三揀四, 一時半刻, 一心一意). Although the study was not corpus-based, it provided a useful categorization of idioms with numbers. Another research on numbers in Mandarin was by Nall (2009) [14]. Both Nall (2009) [14] and Wu (1992 [6]; 1995 [7]) summarized the semantic relations of Mandarin idioms provided by Ma (1978) [15]. These relations are (a) Synonymic Relation (百伶百俐、七拉八扯, Nall, 2009: 93) [14]; (b) Contrast Relation (一出一入、一死一生, Nall, 2009: 95) [14]; (c) Sequential Relation (三思而行、重九登高, Nall, 2009: 95) [14]; (d) Purposive Relation (懲一警百、五斗解酲, Nall, 2009: 96) [14]; and (e) Causative Relation (一謙四益、一言喪邦, Nall, 2009: 96).

When opposites serve as an element in four-word idioms, they can appear in various orders, and their order might affect the internal structure, the part-of-speech (POS), or even

the meanings. In our study, we intended to find all possible combinations of opposites in four-word idioms. We also looked at their sequence, syntactic role, morphological structure, and whether they carried the same meanings in different positions.

3. Methodology

The current version of 教育部重編國語辭典修訂本[3] is an officially edited national reference dictionary and it allows searches for different combinations. By entering a dot (.) in this online reference dictionary, we retrieved different four-word idiom combinations containing opposites, such as "真假..", "假真..", ".真假.", ".假真.", "..真假", "..假真", "真.假.", "假.真.", ".真.假", ".假.真", "真..假", "假..真", etc., which are instances of four-word idioms that contained opposites. This dictionary also provides the definitions, word origins, and example sentences for some of the words.[4] As for the Sinica Corpus, which contains six different categories (literature, living, social, science, philosophy, and art) from 1981 to 2007, with a total of 19,247 articles,[5] we only used POS-tagging from the Sinica Corpus to determine the POS of each "word" in the four-word idioms. For instance, when we tagged 忽冷忽熱, we used the Sinica Corpus to check the POS of 忽, and the adverb tag "D" was found (thus, "DADA").

Jones (2002: 29) [2] suggested that "the best one can do is to investigate a wide range of pairs which a majority of speakers might recognise as being 'good opposites'." Following this, we looked for "good opposites" by generating a list of opposites with a group of trained language researchers. The list was then verified by dictionaries and online lexical resources.

[3] http://dict.revised.moe.edu.tw/cbdic/search.htm 中華民國 104 年 11 月臺灣學術網路第五版試用版

[4] The online dictionary stated that there are 3,017 entries of *chengyu* (成語) in the dictionary. This number of *chengyu* might have slightly different definitions than our "four-word idioms" as *chengyu* is traditionally defined as expressions that come with a history. Therefore, when we computed the overall number of four-word idioms in the whole dictionary, we were aware of the slight differences it might cause. Still, when we examined most of the four-word idioms with opposites that we retrieved, most of them fell into the category of *chengyu*, too. Therefore, the discrepancy might be minimal, if it exists.

[5] http://asbc.iis.sinica.edu.tw/

As Jones (2002: 26-27) [2] indicated that “no single definition of antonymy has been universally agreed upon”, from the list of opposites, we selected 32 to be observed.

Table 2: Thirty-two selected opposites

天地	有無	始終	明暗	尊卑	前後	是非	內外	強弱	粗細	表裡
生死	上下	高低	多少	繁簡	長短	男女	陰陽	盛衰	動靜	左右
異同	冷熱	輕重	真假	逆順	軟硬	虛實	裡外	大小	增減	

The actual procedures comprised two stages. The first stage was to search for four-word idioms that contained opposites, record the dictionary definitions (辭典定義), and then observe whether there were idioms that appeared with different word orders (字序調換) and POS-tagging (詞類標記). At this stage, this study also examined the correspondence between word meanings and whether positive words usually appeared before negative words.

4. Results and Discussion

The number and types of idioms retrieved from the dictionary are summarized in Table 3 below. From a total of 3,017 four-word idioms, 910 constituted 30.16% of the total instances.

Table 3: Thirty-two selected opposites and the number and types of idioms found

Opposites	**Freq.**	**%**	**Opposites**	**Freq.**	**%**	**Opposites**	**Freq.**	**%**
天地	**182**	**20.00**	異同	24	2.64	表裡	8	0.88
生死	**71**	**7.80**	始終	23	2.53	真假	7	0.77
前後	**67**	**7.36**	高低	22	2.42	冷熱	6	0.66
長短	**66**	**7.25**	內外	19	2.09	尊卑	4	0.44
大小	**58**	**6.37**	陰陽	17	1.87	軟硬	4	0.44
有無	**48**	**5.27**	裡外	16	1.76	繁簡	3	0.33
上下	**46**	**5.05**	明暗	16	1.76	粗細	3	0.33
是非	**44**	**4.84**	強弱	15	1.65	逆順	3	0.33
男女	38	4.18	多少	15	1.65	增減	2	0.22
輕重	33	3.63	盛衰	12	1.32	動靜	2	0.22
左右	25	2.75	虛實	11	1.21			
Total							**910**	**100**

In Table 3, 天地 was the most frequently found opposite collected from the dictionary, followed by 生死, 前後, and 長短. We found that the first eight pairs of opposites formed 63.96% of the total 910 idioms, indicating that 天地, 生死, 前後, 長短, and 大小 constituted a majority of the four-word idioms among all types.

We also calculated the proportions of four-word idioms that showed a changed order of AB and whether there was a semantic change. Only 15.6% had a word order change and the remaining 84.2% had the same word order that we expected them to appear in (i.e., AB). Among the word order change, only a small number appeared in the form *上*情*下*達 and *下*情*上*達; a majority was in 花*明*柳*暗* and 柳*暗*花*明*, although we preferred to see the former type. In Table 4 below, because of the latter type of word changes, it was not surprising that only a small number of idioms (2.2%) had meaning changes, while 97.8% kept their original idiom meanings. Apart from those that changed with their components, as in 花*明*柳*暗* and 柳*暗*花*明*, those that had repeated words did not change in meaning, such as 不*生*不*死*.

Table 4: Analysis of word orders of opposites and semantic changes

有無字序調換	**成語數量**	**有無語意改變**	**成語數量**
有字序調換	144 (15.8%)	有語意改變	20 (2.2%)
無字序調換	766 (84.2%)	無語意改變	890 (97.8%)
總計	**910 (100%)**	**總計**	**910 (100%)**

More than 84.2% of the four-word idioms were more fixed; and even if the remaining 15.8% had word order changes, their meanings were often not affected because they changed with their neighboring components, such as 花*明*柳*暗* and 柳*暗*花*明*. Only a small number of opposites changed and had contrastive meanings, as in 上情下達 and 下情上達. We then analyzed the AB patterns (成語格式) of each opposite pair (13 patterns in total):

Table 5: AB patterns of the four-word idioms containing opposites

成語格式	例子	成語數量	成語格式	例子	成語數量
?A?B	*大*題*小*作	337 (37.0%)	??BA	飛流*短長*	10 (1.1%)
A?B?	*明*查*暗*訪	206 (22.6%)	AABB	*裡裡外外*	10 (1.1%)

?B?A	視*死*如*生*	111 (12.2%)	B??A	*無*奇不*有*	6 (0.7%)
AB??	*左右*逢源	94 (10.3%)	A??B	*盛*極*必衰*	5 (0.5%)
B?A?	*弱*肉*強*食	68 (7.5%)	?AB?	半大小子	4 (0.4%)
??AB	別有*天地*	45 (4.9%)	?BA?	絕*地*天通	1 (0.1%)
BA??	*終始*如一	13 (1.4%)	總筆數		**910** (100%)

As can be seen in Table 5 above, when we calculated all the AB word orders, we found a total of 701 (77%) instances among the 910 idioms. This shows that our AB word order instinct was matched a majority of the time in the database, indicating that an AB order was expected. In other words, we mentally stored the opposites in an AB word order (e.g., 天地 instead of 地天; 明暗 instead of 暗明) and they appeared almost 77% in a similar word order in the idioms. The BA word order accounted for about 23% of the remaining idioms. In Table 5, the "?A?B" pattern appears to be the most frequent pattern, followed by the "A?B?" pattern, as a result of the POS of the idioms. There were 71 POS patterns in total (POS patterns with frequency fewer than seven are not shown):

Table 6: Distributions of four-word idiom types in terms of POS patterns (Type Freq. ≥ 7)

POS	Example	Freq.	%	POS	Example	Freq.	%
VNVN	想*前*顧*後*	301	33.08	AAAA	*長長短短*	11	1.21
NANA	*天差地*遠	112	12.31	VAVA	知*高*識*低*	10	1.10
ANAN	千*生*萬*死*	86	9.45	NNVN	*左右*開弓	9	0.99
NVNV	*天*造*地*設	74	8.13	NNDA	*表裡*相合	8	0.88
NNNN	*天*涯*地*角	50	5.49	VVVV	*有*屈*無*伸[6]	8	0.88
DVDV	*後*擁*前*推	28	3.08	ANDV	*小*題*大*作	7	0.77
NNVV	*生死*存亡	25	2.75	DVNN	不顧*前後*	7	0.77
VVNN	撥弄*是非*	17	1.87	NNAN	*天地*萬物	7	0.77
DADA	半*大*不*小*	16	1.76	AVAV	*冷*譏*熱*嘲	7	0.77
NNDV	*是非*不分	13	1.43				

This analysis helped us anticipate the kind of structure that would appear under each pattern and the kind of POS patterns denoted by each.

[6] 有 has a tag of "V_2" (a verb) in the Sinica Corpus.

5. Conclusion and Future Research

Mandarin four-word idioms have a long history, and many studies have focused on them. However, idioms with opposites have seldom been studied. The few cases that were studied were not analyzed in the way they were in our study. In this paper, we addressed the following: (a) the internal structure of four-word idioms that contain opposites; (b) the order of opposites in the four-word idioms; and (c) the most frequently found patterns of opposites in four-word idioms. Moreover, we analyzed the possibility of word order changes and how the word order changes affected the meanings, if at all. We also analyzed the POS of each component in the four-word idioms. All this information is useful in teaching idioms and when considering whether idioms should be collected in dictionaries. We also found which pairs of opposites are more prominent in Mandarin as well as all about these opposites.

As noted in this paper, our methodology has limitations. We overcame the limitations by decision-making, which may have affected some of the results, but they were weighted throughout by minimizing the problems these decisions caused regarding the complete results. Thus, more studies are needed. The frequency of idioms in a corpus was the original focus of this study, but the frequency of the idioms was lower than expected and not all combinations could be found in the corpus. In the future, this will need to be overcome by finding a suitable corpus for the study of idiom frequency.

References

[1] Wang, X.-F., Wu, Z.-F., Li, Y., Huang, Q., & Hui, J. (2010). Corpus-based analysis of the co-occurrence of Chinese antonym pairs. In *Advanced data mining and applications* (pp. 500-507). Heidelberg: Springer.

[2] Jones, S. (2002). *Antonymy: A corpus-based perspective*. London and New York: Routledge.

[3] Jones, S., Murphy, M. L., Paradis, C., & Willners, C. (2007). Googling for opposites—A web-based study of antonym canonicity, *Corpora, 2.2*, 129-155.

[4] Jones, S., Murphy, M. L, Paradis, C., & Willners, C. (2012). *Antonyms in English: Construals, constructions and canonicity*. Cambridge: Cambridge University Press.

[5] Murphy, M. L., Paradis, C., Willners, C., & Jones S. (2009). Discourse functions of antonymy: A cross-linguistic investigation of Swedish and English. *Journal of Pragmatics, 41*, 2159-2184.

[6] Wu, C.-H. (1992). *Semantic-based synthesis of Chinese idioms (Chéngyǔ)*. Unpublished doctoral dissertation. Georgetown University, Washington D.C.

[7] Wu, C.-H. (1995). On the cultural traits of Chinese idioms. *Intercultural Communication Studies, 5.1*, 61-84.

[8] 謝健雄（2006）。漢語成語情感譬喻之概念模式與語言結構。*UST Working Papers in Linguistics,* 2, 43-65。

[9] 遲嘯川、許易人（2002）。超活用成語大辭典。中和市：華文網。

[10] 陳曉燕（2004）。 現代漢語詞彙中反義語素並行構詞現象說略。 鹽城工學院學報(社會科學版)，1，56-60。

[11] 韓漢雄（1993）。漢英反義詞的成對使用比較。杭州師範學院學報。1，134-140。

[12] 譚達人（1989）。 略論反義相成詞。語文研究，1，27- 33。

[13] Qiu, X. (2015). Four-character set phrases in Taiwanese Mandarin: A cognitive approach to studying phrases with numbers. In *Recent developments of Chinese teaching and learning in higher education: Applied Chinese language studies* (pp. 79-90). London: Sinolingua London Ltd.

[14] Nall, T. M. (2009). *An analysis of Chinese four-character idioms containing numbers: Structural patterns and cultural significance*. Unpublished doctoral dissertation. Ball State University, Indiana.

[15] Ma, K.-f. (1978). *Cheng Yu*. Inner Mongolia: People.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 408-422

漢語及物化的大數據研究

A Data Scientific Study of Transitivization in Chinese

蔡維天　Wei-Tien Dylan Tsai
國立清華大學語言學研究所
Linguistics Institute
National Tsing Hua University
wttsai@mx.nthu.edu.tw

楊謦瑜　Ching-Yu Helen Yang
國立清華大學資工系
Department of Computer Science
National Tsing Hua University
chingyu@nlplab.cc

陳映竹　Chen Ying-Zhu
國立清華大學資工系
Department of Computer Science
National Tsing Hua University
jocelyn@nlplab.cc

陳志杰　Jhih-Jie Chen
國立清華大學資工系
Department of Computer Science
National Tsing Hua University
jjc@nlplab.cc

張俊盛　Jason S. Chang
國立清華大學資工系
Department of Computer Science
National Tsing Hua University
jason@nlplab.cc

摘要

本文從資料科學的角度來考察漢語中一個新興的現象「及物化」：亦即原本以動前介詞組引介域內論元的謂語(如「為人民服務」)轉化為直接引介動後賓語的謂語(如「服務人民」)。我們認為這個現象其實是一種復古的趨勢，如古漢語的為動式「壯士死知己」即「壯士為知己而死」之意，是一種隱性的輕動詞用法。值得一提的是，這種趨勢仍處於變動之中：它可能逐步消亡，也可能引發爆炸性的發展(如同「語言癌」一般)；這便

是為甚麼我們需要從資料科學的角度切入為其把脈，不但能總結之前的演化歷程，更能預測未來的發展趨勢。此外，我們發現及物化現階段只在幾種文體中有高度的能產性，這也讓我們有機會一窺其使用上的限制及其背後的制約因素(如語用和韻律上的考量等)，並印證於文法搜尋引擎如 Linggle 的數據統計之上。

關鍵詞：及物化，漢語句法，輕動詞，資料科學

一、緒論：何謂及物化？

在歷史的長河中，語言的變遷無異是學者們考察關注的重點之一：其原因即在於漢語有大量的文本典籍和歷史記錄(如災荒、遷徙、戰爭、駐軍及地緣政治等)可供查對，同時還有從唐宋以下各朝各代建立起的韻書撰寫傳統，反映出多層次、多方言的動態音韻體系，豐富了我們探究漢語歷史演化的工具箱，讓專家學者得以一窺中古漢語和上古漢語的風貌，也使重建祖語(proto-language)成為可能。

而現代語言學的發展也正處在一個關鍵期：語言定義「人之所以為人」的生物本能是如何跟種種外緣因素互動(如使用、接觸、混合等)，進而驅動了演化的歷程。這些困難複雜的議題都需要有嚴謹的田野、實驗和大數據研究做為工具，才能抽絲剝繭，澄清問題的本質，並提出明確的解決之道。

根據以上理念，我們將探索重點聚焦在一個近年來興起的語法現象，可姑且稱之為「及物化」(transitivization)。漢語文獻中其實已經有了相當深入的觀察和後續討論：齊滬揚(2000)即指出「漢語述賓結構是一種優勢結構，許多原非述賓結構有向述賓結構靠攏的趨勢這樣使得一些原先不能帶賓語的動詞也逐漸可以帶賓語了，及物動詞的數量呈擴大趨勢。」他舉了以下幾個淺顯易懂的例子，可作為參考：

(1) a. 他常[為人民]服務。
    b. 他常服務[人民]。
(2) a. 他常[用毛筆]寫字。
    b. 他常寫[毛筆]字。

(3) a. 我們[在北京]相見。

b. 我們相見[北京]。

如(1a)中動前的由「為」引介的受惠者「人民」在(1b)中變成了動後的直接賓語；又如(2a)中動前的由「用」引介的工具「毛筆」在(2b)中變成了動後的直接賓語；最後(3a)中動前的由「在」引介的地點「北京」在(3b)中變成了動後的直接賓語。

蔡維天(2017)則注意到上述及物化現象在某些文體或地區特別顯著：比如說臺灣媒體下簡短標題時就常採用此種及物句構：表對象關係的例子如(4a-c)，可分別轉寫為(5a-c)，原本動後的直接賓語放到了動前介詞組的位置：

(4) a. 馬林魚仍甜蜜復仇[紅雀]。

b. 張艾亞生日告白[許孟哲]。

c. 法院嗆聲[王寶強]。

(5) a. 馬林魚仍[對紅雀]甜蜜復仇。

b. 張艾亞生日[對許孟哲]告白。

c. 法院[對王寶強]嗆聲。

表示與同(comitative)關係的例子如(6a-c)，可分別轉寫為(7a-c)，亦即動前不及物用法的例子：

(6) a. 伊朗斷交沙國。

b. 蔡康永牽手小 S，…

c. 蕭亞軒分手百億男友。

d. 秦凱求婚何姿。

(7) a. 伊朗跟沙國斷交。

b. 蔡康永跟小 S 牽手，淚灑錄影現場。

c. 蕭亞軒跟百億男友分手。

d. 秦凱跟何姿求婚。

最後是表達蒙受關係的(8a,b)，可分別用動前的「給」、「把」轉寫為(9a,b)：

(8) a. 法陸空罷工添亂歐國杯。
b. 美國聯邦法官打臉川普。
(9) a. 法陸空罷工給歐國杯添亂。
b. 美國聯邦法官把川普打臉。

相對而言，這種新興的及物化在敍事文本中則較少見。更有趣的是，這類用法在口語中反而能產性頗高，而且愈來愈普遍：在動後賓語位置引介對象論元的有(10a-d)，可分別改寫為(11a-d)：

(10) a. 有沒有人發問[你]？
b. 他常輕挑[女生]。
c. 你別刻薄[人家]。
d. 但我絕對一定要來靠北[她]。
(11) a. 有沒有人[對你]發問？
b. 他常[對女生]輕挑。
c. 你別[對人家]刻薄。
d. 但我絕對一定要來[對她]靠北。

此外，引介受惠者或原因論元當直接賓語的有(12a-d)，可分別改寫為(13a-d)中的不及物用法(也有人稱為半及物)：

(12) a. 報紙一直吹牛[他]。
b. 薇如緊張[我]，…
c. 他很傷心[這件事]。
d. 他很高興[這件事]。
(13) a. 報紙一直[為他]吹牛。
b. 薇如[為我]緊張，就一直說你怎麼了？

c. 他[為這件事]很傷心。

d. 他[為這件事]很高興。

最後，引介蒙受論元當直接賓語的則是(14a,b) ，可分別改寫為 (15a-d)：

(14) a. 他還想要索賠[對方]。

b. 你怎麼可以放鳥[人家]？

(15) a. 他還想要[跟對方]索賠。

b. 你怎麼可以[把人家]放鳥？

二、研究方法

我們更進一步使用語料庫驗證上述漢語及物化演變的趨勢（參閱蔡維天 2017），考量及物化可能因年代與語體而有不同的變化，本研究選用不同年代的新聞資料作為研究標的，並區分標題與內文以對比語體。我們分別將這三份資料建構成以依存關係（Universal dependency）（Nivre et al., 2015, 2016）為框架的語料庫，並抽取所需資料做統計，方法步驟如下：

1. 重新斷句、為標題補上標點符號
2. 使用中研院詞庫小組 CKIP CoreNLP 系統斷詞與做詞性剖析[1]
3. 使用 Chinese Stanford Parser 剖析依存關係[2]

本節將介紹使用的語料庫、進行的預處理與運用的剖析器（Parser）。

[1] http://ckip.iis.sinica.edu.tw/service/corenlp/

[2] https://github.com/UniversalDependencies/UD_Chinese-GSD

(一)資料介紹

我們將 LDC Tagged Chinese Gigaword[3] 的標題與內文分開儲存為兩份資料庫，涵蓋的新聞來源有：中央通訊社、新華通訊社與聯合早報，資料年份為 1991 年到 2004 年，新聞約 181 萬 6 千多篇。第三份資料為 2004 年到 2017 年間的新聞內文，主要新聞來源為：聯合報、聯合晚報、經濟日報，約有 177 萬多篇新聞。

1. 新聞資料__標題：1991 年至 2004 年，176 萬 8 千多則
2. 新聞資料__內文一：1991 年至 2004 年，176 萬 8 千多則
3. 新聞資料__內文二：2004 年至 2017 年，177 萬多則

（二）資料處理與剖析

為提升 Chinese Stanford Parser 的依存關係剖析效果，我們先進行以下預處理：將標題（資料一）結尾補上句點符號，解決 Stanford Parser 傾向將最後一個詞剖析成標點符號的問題（如表一）。內文部分（資料二、資料三），以句點、問號、分號、驚嘆號斷句，以避免句子不完整而造成依存關係剖析錯誤（如表二）。將處理好後的資料，以中研院詞庫小組 CKIP CoreNLP 系統斷詞與做詞性剖析。最後，我們使用 Chinese Stanford Parser（UD_Chinese-GSD）做依存關係剖析。處理後，語料庫大小，條列如下：

1. 新聞資料__標題(1991 年至 2004 年)：176 萬 8 千多則，181 萬 6 千多句，1650 萬 2 千多詞。
2. 新聞資料__內文一(1991 年至 2004 年)：176 萬 8 千多則，1699 萬 8 千多句，4 億 6293 萬多個詞。
3. 新聞資料__內文二(2004 年至 2017 年)：177 萬多則，1665 萬 5 千多句，5 億 2 千多萬詞。

[3] https://catalog.ldc.upenn.edu/LDC2007T03

表一、添加標點符號

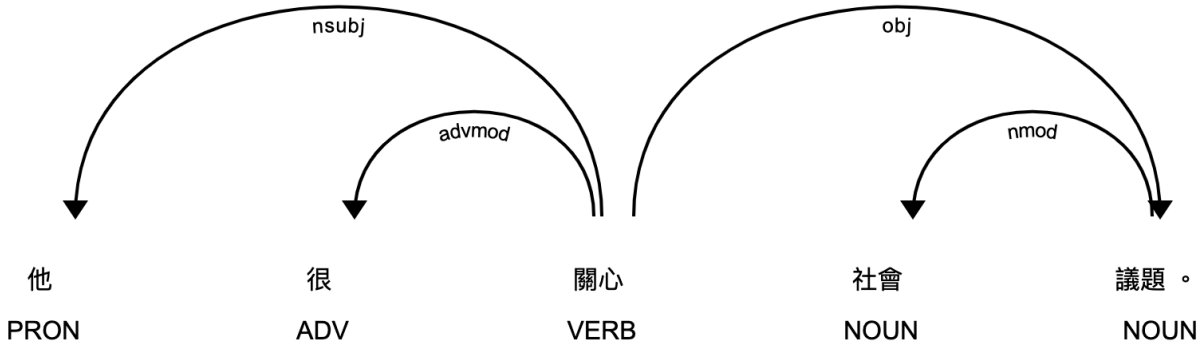
處理前	處理後
環保署推出環保建設六年計畫 法國不在意日本對其核試的杯葛 豐原醫院貼心手推車服務解決住出院行李煩惱	環保署推出環保建設六年計畫。 法國不在意日本對其核試的杯葛。 豐原醫院貼心手推車服務解決住出院行李煩惱。

表二、將 Tagged Chinese Gigaword 重新斷句

處理前	處理後
戈巴契夫曾表示， 他相信國家安全委員會前首腦克留契柯夫是政變主謀。 愛滋病患者對他社交圈內別人對他的看法， 與對待他的方式， 非常在意。	戈巴契夫曾表示，他相信國家安全委員會前首腦克留契柯夫是政變主謀。 愛滋病患者對他社交圈內別人對他的看法，與對待他的方式，非常在意。

三、 抽取依存關係的規則

近年來，很多自然處理的研究使用依存句法分析樹（Universal Dependencies Treebank）來訓練句法剖析器。Universal Dependencies（Nivre et al., 2015, 2016）是一個跨語言語法標註的框架，涵蓋的語言超過 70 種，中文也包含其中，相關研究學者訂定了跨語言標註分析的準則（Nivre et al., 2015, 2016），超過 70 個語言。在依存句法分析的架構下，直接標註詞與詞之間的依存關係，結構較為扁平，如圖一所示。



圖一、依存句法分析

一個句子只有一個獨立成分，不必依附於（depend on）其他詞，圖一的「關心」即為此例，標示的依存關係為“root”，其餘詞彙則須依附在其他的詞彙上，以箭頭表示，如「他」、「很」與「議題」皆依附於「關心」，依存關係分別為主詞（nsubj）、副詞修飾語（advmod）以及受詞（obj），而「社會」則依附於「議題」，依存關係為名詞修飾語（nmod）。依存句法樹直接標註詞與詞的關係，而非詞與詞組之間的關係，如圖一中依附在「關心」的詞為「議題」，而非「社會 議題」，如此一來，可直接得知與考察目標詞彙最直接相關的詞有哪些，以及依存關係為何，方便語法的抽取與搭配詞的計算。

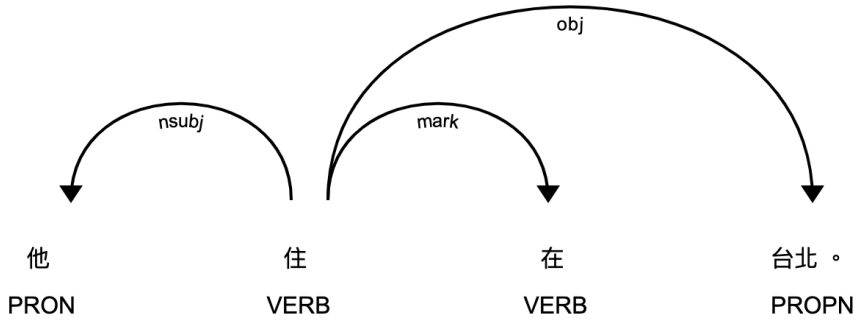
本研究目標為考察動詞及物與不及物用法在不同語料庫的對比，需抽取以下兩個結構：(1) 動詞 受格論元; (2) 介系詞 旁格論元 動詞。我們依照表三、表四的規則抽取依存關係，組成標的結構，留下例句並統計次數。

表三、「動詞 受格論元」抽取規則

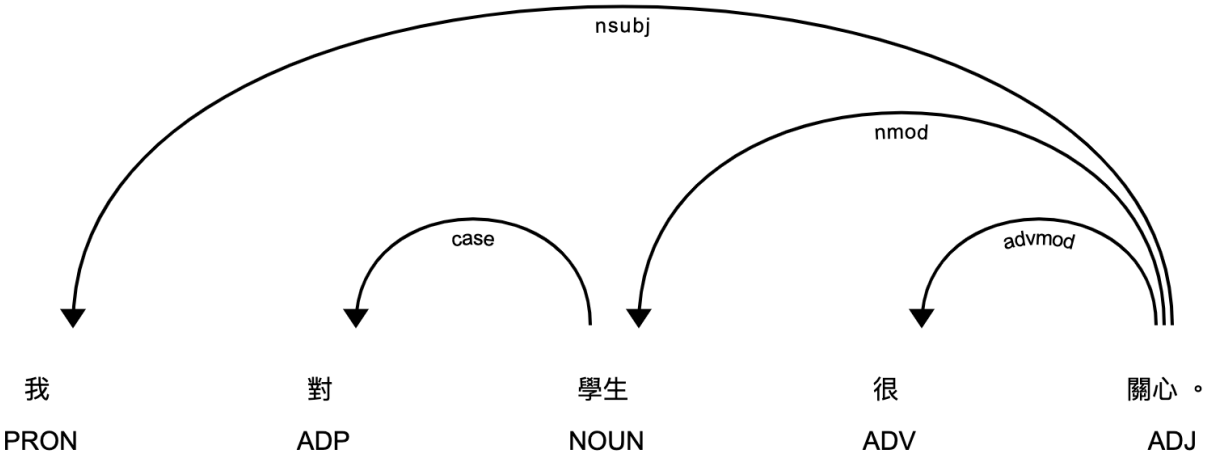
抽取規則	1. 目標動詞為 root，受格論元依存於 root，兩者依存關係為 obj。 2. root 與其他詞的依存關係不包含依存關係 mark。
抽取範例	(1) 他 很 關心 社會 議題 。 (符合抽取規則，參見圖一) 抽取詞彙：關心(root) 議題(obj)
濾除範例	(2) 他 住 在 台北 。(不符合抽取規則 2，參見圖二) 住(root) 在(mark) 台北(obj)

表四、「介系詞 旁格論元 動詞」抽取規則

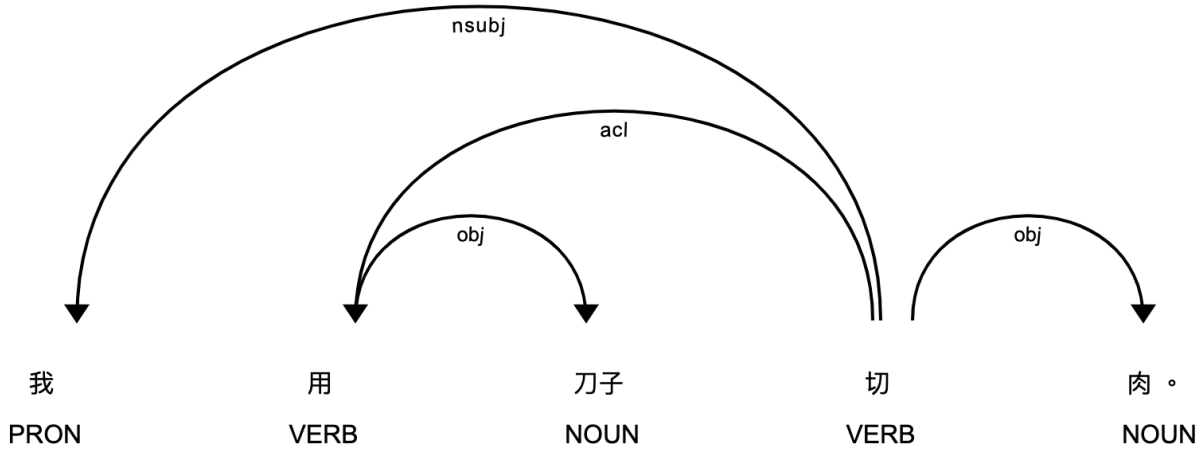
抽取規則	1. 介系詞 A. 句子包含下欄的介系詞，且中研院 Core NLP 詞性分析為介系詞（P）。 B. 依附於動詞(root)或是動詞的依附詞（見抽取範例(2)）。 C. 依附關係為 case、mark 或是 acl。 2. 旁格論元 A. 依附於動詞(root)或是介系詞 B. 依附關係為 obj、obl 或是 nmod
介系詞列表	在, 以, 與, 對, 從, 爲, 由, 到, 向, 於, 和, 用, 跟, 給, 至, 經, 往, 靠, 替, 藉, 朝, 按, 憑, 沿, 把, 對於
抽取範例	(1) 他 住 在 台北 。 (符合抽取規則，參見圖二) 抽取詞彙：住(root) 在(mark) 台北(obj) (2) 我 對 學生 很 關心 。(符合抽取規則，參見圖三) 抽取詞彙：對(case) 學生(nmod) 關心(root) (3) 我 用 刀子 切 肉。(符合抽取規則，參見圖四) 抽取詞彙：對(case) 學生(nmod) 關心(root)



圖二、依存句法分析



圖三、依存句法分析



圖四、依存句法分析

四、統計結果與分析

我們使用大數據考察漢語及物化的趨勢，聚焦在以下兩個項目：

1. 標題比其他書面文體更傾向使用及物句構。
2. 年代較近的資料比較久遠的資料更傾向使用及物句構。

我們先呈現及物化在標題與內文的差異，再探究年代對及物化的影響。我們以蔡維天(2017)一文中提及的動詞作為起點，來觀察漢語及物化的趨勢：在意、關心、服務、相見、擔心、復仇、告白、嗆聲、斷交、牽手、分手、求婚、添亂、打臉、背信、偷情、發問、輕挑、刻薄、靠北、吹牛、緊張、傷心、高興、索賠、放鳥、混、劈腿、求救、前進、褒獎、怪罪、加害、看好、眼紅、頭痛。

（一）標題與內文的差異與分析

資料一與資料二皆為 1991 年至 2004 年的新聞資料，前者為標題，後者為內文，可用來探究趨勢一指出的及物化情形。我們根據第三節的抽取規則，統計標的動詞的及物結構與不及物結構的次數，結果請見表五。

表五、資料一與資料二及物性統計結果

資料一（標題 1991-2004）				資料二（內文 1991-2004）				比較
	及物	不及物	及物性		及物	不及物	及物性	1 > 2
在意	10	0	1	在意	29	37	0.44	✓
復仇	1	0	1	復仇	1	1	0.5	✓
怪罪	7	0	1	怪罪	22	17	0.56	✓
加害	2	0	1	加害	2	4	0.33	✓
關心	489	19	0.96	關心	481	230	0.68	✓
擔心	176	9	0.95	擔心	417	185	0.69	✓
看好	383	22	0.95	看好	325	310	0.51	✓
前進	56	12	0.82	前進	101	182	0.36	✓
服務	168	41	0.8	服務	555	563	0.5	✓
混	2	1	0.67	混	15	17	0.47	✓
告白	5	3	0.63	告白	7	6	0.54	✓
斷交	14	11	0.56	斷交	7	112	0.06	✓
分手	3	4	0.43	分手	2	35	0.05	✓
相見	3	5	0.38	相見	3	29	0.09	✓
求婚	5	10	0.33	求婚	4	19	0.17	✓
索賠	16	45	0.26	索賠	31	64	0.33	×
緊張	3	15	0.17	緊張	2	81	0.02	✓
求救	3	16	0.16	求救	11	129	0.08	✓
嗆聲	0	2	0	嗆聲	4	10	0.29	×
發問	0	2	0	發問	8	31	0.21	×
傷心	0	1	0	傷心	1	5	0.17	×
高興	0	26	0	高興	7	211	0.03	×
頭痛	0	2	0	頭痛	0	18	0	--

我們先排除資料一或資料二中及物結構「動詞 受格論元」與不及物結構「介系詞 旁格論元 動詞」筆數為零的動詞 13 筆（如：「添亂」、「打臉」），第二欄與第六欄為及物結構的筆數，第三欄與第七欄為不及物結構的筆數，第四欄與第八欄為標的動詞的及物性，計算方式為：及物結構除以總次數（不及物結構的次數加上及物結構的次數），分數越高，及物性越強，最後一欄比較標的動詞在資料一的及物性是否比在資料二高。對比 23 個動詞中，有 17 個動詞在標題資料中的及物性比在內文資料中來得高（如：「在意」、「關心」），有 6 個動詞在內文資料的及物性較高（如：「發問」、「嗆聲」）。

（二）及物性於跨年代的演變

趨勢二的及物化情形可透過對比資料二與資料三來一虧端倪，資料二與資料三皆為新聞內文，前者整年代較久(1991 年至 2004 年)，後者年代較新。及物化在不同年代對比結果請見表六。

表六、資料二與資料三及物性統計結果

資料三（內文 2004-2017）				資料二（內文 1991-2004）				比較
	及物	不及物	及物性		及物	不及物	及物性	3 > 2
擔心	923	265	0.78	擔心	417	185	0.69	✓
關心	919	287	0.76	關心	481	230	0.68	✓
吹牛	5	1	0.83	吹牛	3	0	1	×
怪罪	28	11	0.72	怪罪	22	17	0.56	✓
在意	188	84	0.69	在意	29	37	0.44	✓
看好	674	309	0.69	看好	325	310	0.51	✓
服務	1201	743	0.62	服務	555	563	0.5	✓
加害	3	2	0.6	加害	2	4	0.33	✓
背信	3	2	0.6	背信	1	0	1	×
索賠	64	60	0.52	索賠	31	64	0.33	✓
混	115	109	0.51	混	15	17	0.47	✓
傷心	12	12	0.5	傷心	1	5	0.17	✓
前進	339	426	0.44	前進	101	182	0.36	✓
偷情	3	5	0.38	偷情	1	0	1	×
刻薄	1	2	0.33	刻薄	0	1	0	×
復仇	4	10	0.29	復仇	1	1	0.5	×
發問	8	31	0.21	發問	8	31	0.21	--
緊張	19	108	0.15	緊張	2	81	0.02	✓
高興	25	147	0.15	高興	7	211	0.03	✓
頭痛	5	29	0.15	頭痛	0	18	0	✓
牽手	2	13	0.13	牽手	1	3	0.25	×
求婚	27	198	0.12	求婚	4	19	0.17	×
分手	26	204	0.11	分手	2	35	0.05	✓
告白	8	81	0.09	告白	7	6	0.54	×
斷交	3	29	0.09	斷交	7	112	0.06	✓
相見	10	97	0.09	相見	3	29	0.09	--
嗆聲	29	279	0.09	嗆聲	4	10	0.29	×

求救	40	647	0.06	求救	11	129	0.08	×

資料二或資料三中及物結構「動詞 受格論元」與不及物結構「介系詞 旁格論元 動詞」筆數為零的動詞有 8 筆（如：「添亂」、「打臉」），暫不列入考量，表六表格安排與及物性按表五方式計算，分數越高，及物性越強，對比結果為，28 個動詞中，有 16 個動詞的及物性有隨年代而遞增的趨勢（如：「擔心」、「關心」），有 2 個動詞的及物性（如：「發問」、「相見」）未應年代而產生變化，而有 10 個動詞的及物性不增反減（如：「牽手」、「刻薄」）。

（三）結果分析

資料一與資料二的對比反應出標題比內文文體更傾向使用及物句構（趨勢一），23 個動詞中有 16 個動詞支持這個趨勢。資料二與資料三的差異呈現，在年代較近的語料有較多及物結構，28 個動詞中有 16 個動詞有及物化的趨勢，些微反應出年代對及物化的影響（趨勢一）。從數據上看來，年代對及物化的影響沒有文體差異顯著，可能句法上的新興演變會優先反應在口語上，需要一段時間才會反應在書面語上，因此動詞有隨時間演進有及物化的趨勢，但仍不顯著。而媒體下標題需簡短、吸引讀者注意力，不及物動詞使用及物化結構，可達到媒體因此可因語用因素而違反句法規則。

動詞有 16 組未支持及物化趨勢，其中一個可能的原因是動詞的詞頻較低（如：「偷情」、「刻薄」、「復仇」、「嗆聲」等），使用度較低的動詞，對語法變化會較保守，未來研究可先以詞頻較高的動詞為研究標的，以排除詞頻的干擾因素。另一個可能為，資料的差異性不夠高，資料一、資料二的文體不同，一為標題，一為內文，但仍同屬新聞文體，未來關於文體差異的研究，可對比書面語與口語資料，反應出的及物化趨勢會更加明顯。關於年代的對比，統計方式可改為統計所有標的動詞的逐年及物化變化，或許更適合考察及物化歷年演變的趨勢。

五、結論與未來展望

前述觀察與研究顯示語法研究應該和資料科學充份結合，取得實證面上的數據支持，並在語言教學和人工智能等面向上發揮其潛在應用價值。事實上，若從中英文平行語料庫做初步觀察，及物化的用法很難找到跨語言的對應，但卻凸顯了中英文在語言類型上的差異：以(16)中的對譯為例，首先是語序上英語的介詞結構在動後而非動前，這與其「中心在前」(head-initial)的特色一致。其次是英語名動互轉的機制非常發達（亦即 quality ⇒ qualify），而漢語只能靠引介輕動詞(light verb)如「取得」來翻譯。

(16) Andrew **qualified as** a teacher in 1995.
安德魯 於 1995 年 **取得** 教師 **資格** 。

這點在(17)、(18)的對譯中也得到印證：相當於英語中一個簡單的介詞 *for*，中文翻譯卻需要添加「喝」、「找」等動詞，才能形成動後的及物用法：其實此處 *for* 的性質非常接近漢語中語意泛化的輕動詞；一旦翻成中文便需要照顧到賓語的選擇限制，必須用更明確的動詞才行：

(17) It was freezing outside and Marcia **longed for** a hot drink
外面 很 冷 ， 瑪西雅 **很 想 喝** 一 杯 熱飲 。

(18) She **groped for** her glasses on the bedside table
她 在 床頭櫃 上 **摸索 著 找 眼鏡** 。

最後是 *qualify* 其實是一種使動動詞，相當於 *cause someone to have the quality/certificate of a teacher*。由於漢語沒有類似用法，也不太能說「**取得資格成**教師」，因此「教師」便從補語轉化為名前定語，相當英語中名後的 *of a teacher* 或是動後的 *as a teacher*。這些現象若有足夠多的語料可供大數據研究，相信一定能在華語教學、文法寫作、機器翻譯等面向上開啟更具突破性的發展，讓文法理論和應用做更完美的結合。

此外，我們可以開始發展「客製化」的中文文法搜尋引擎，用大數據理念來研究各種不同文體的語法差異和通則，進而應用到修辭學、文體學、高級寫作、辭典編纂以至符號學、社會語言學等做具有前瞻性的跨界研究。一旦我們有了夠大的資料庫，便可將觸角延伸至認知結構、本體論(ontology)、歷史語法、語言類型學及普遍語法的研究，讓不同領域的學者既能各取所需又可相互支援。

參考文獻

[1] 蔡維天. 2017.〈及物化、施用結構與輕動詞分析〉，《現代中國語研究》，第19期，1-13頁。

[2] 齊滬揚. 2000.《現代漢語短語》，張斌主編，上海：華東師範大學出版社。

[3] Chang, P.-C., Tseng, H., Jurafsky, D., and Manning, C. D. Discriminative reordering with Chinese grammatical relations features. In *Third Workshop on Syntax and Structure in Statistical Translation*, pages 51–59, 2009.

[4] de Marneffe, M.-C., Connor, M., Silveira, N., Bowman, S. R., Dozat, T., and Manning, C. D. More constructions, more genres: Extending Stanford dependencies. In *DepLing*, 2013.

[5] Elming, J., Johannsen, A., Klerke, S., Lapponi, E., Martinez, H., and Søgaard, A. Down-stream effects of tree-to-dependency conversions. In *NAACL HLT*, 2013.

[6] De Marneffe, Marie-Catherine, and Christopher D. Manning. *Stanford typed dependencies manual*. Technical report, Stanford University, 2008.

[7] de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC), 2014.

[8] McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Ta¨ckstro¨m, O., Bedini, C., Bertomeu Castello´, N., and Lee, J. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL), 2013.

[9] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajicˇ, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. Universal dependencies v1: A multilingual tree- bank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC), 2016.

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 423

基於訊息配對相似度估計的聊天記錄解構

劉至咸　ZhiXian Liu
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
zhixian989@gmail.com

張嘉惠　Chia-Hui Chang
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
chia@csie.ncu.edu.tw

摘要

一般而言，為建立 Retrieval-based 聊天機器人，我們可以從聊天紀錄中來產所需的問答配對（Question-Answer Pair），然而問答配對並非完全依序地呈現在聊天紀錄中，不同內容的問答配對可能互相穿插，而從互相穿插的訊息中分離出內容不同的會話的任務稱為對話解構（conversation disentanglement）。

現有的對話解構研究大多透過計算兩個訊息的相似度來解決問題，在此論文中，我們發透過計算訊息相似度判斷訊息是否屬於相同會話是非常困難的，但若我們透過計算相似度來預測訊息的回覆關係則可以解決此問題。此外我們指出過去研究中的模型無法處理未經訓練的訊息，並無法在實務上運用的缺陷。

此論文中，我們使用 IRC 與 Reddit 資料集進行實驗，並使用聊天記錄進行對話解構。其中人工合成的 Reddit 資料集提供額外的大量訓練資料，且 BERT 模型在此資料集上的回覆關係預測獲得良好的效能。

關鍵詞：對話解構，回覆關係預測，BERT 模型應用

The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 424-433

標註英中同步樣式文法之研究

Annotating Synchronous Grammar Patterns across English and Chinese

楊謦瑜　Ching-Yu Helen Yang, 陳映竹　Chen Ying-Zhu, 張俊盛　Jason S. Chang
國立清華大學資工系
Department of Computer Science
National Tsing Hua University
chingyu@nlplab.cc, jocelyn@nlplab.cc, jason@nlplab.cc

林依蒨　Yi-Chien Lin
國立清華大學外語系
Department of Foreign Languages
National Tsing Hua University
nicayclin@gmail.com

蔡維天　Wei-Tien Dylan Tsai
國立清華大學語言學研究所
Linguistics Institute
National Tsing Hua University
wttsai@mx.nthu.edu.tw

摘要

本文從語料庫與機器學習的角度，來進行英語和華語的同步文法研究。我們認為這種對比性研究，可以借助既有的英語樣式文法的研究成果，提供華語辭典學、華語教學新的研發方向。在我們的研究路線上，運用了辭典中的雙語例句，來發掘英語、華語的動詞句法規則。我們的方法涉及自動辨識例句中的英語文法規則、運用詞彙對應的技巧產生華語規則的建議，最後，透過人為分析產生正確英華語法規則的資料集。我們把這個研究方法，運用在劍橋大學出版社的線上英漢辭典的例句，初步完成英語「動介賓」規則的華語對應規則的分析。我們就初步的研究結果，說明標註華語文法規則的指導原則，觀察分析所得到的華語文法規則的統計分布。最後完成的資料集，可望有助於提供華語文法規則自動擷取的機器學習研究。

關鍵詞: 樣式文法 pattern grammar，同步文法 synchronous grammar，自動文法推導 grammar induction

一、緒論

實證性、語料庫為本的句法研究，有幾個不同的作法。最常見的方式，透過抽樣取少部分句子樣本，人為分析這些句子的句法結構，以建構剖析樹庫的方式，來得到一個文法剖析的代表性樣本（Marcus, Santorini, and Marcinkiewicz 1993）。另外一條詞彙式的研究路線（Sinclair 2000），是分析語料庫中個別詞彙的樣本，分析其常見文法規則，以羅列詞彙化文法規則。在華語句法研究上，已經有中研院的樹庫，Chinese Treebank 8.0 等樹庫資料集的研究發展（Huang, et al. 2000; Xue and Palmer 2003, Xue, et al. 2005）。然而，華語詞彙化文法（如樣式文法 pattern grammar），卻相當缺乏相關資料與研究。

本文描述一項計畫，透過英語既有的詞彙化文法規則，以及相關的平行雙語句子，標示華語句子中出現的對應規則。其目的在於產生對應的華語文法規則的小量訓練資料，以利後續可以饋入機器學習系統，在單語語料庫中，擷取更全面的華語文法規則。

樣式文法（Pattern Grammar ）是一種描述個別詞彙的句法環境的語言模型。PG 源自在大型語料庫中，觀察個別詞彙的實例。在知名的考林斯出版社與伯明罕大學的 COBUILD 計畫中，Hunston, Francis, and Manning (1996, 1997) 發展出的文法理論，和一般習知的片語結構文法（phrase structure context-free grammar），有很大的區隔[1]。

PG 主張，語言學家、辭典學家可以針對每個實詞（動詞、名詞、形容詞），觀察語料庫來賦與一組文法規則（patterns），用以描述該詞彙的主要用法。PG 更進一步主張，通常規則有類似語意。PG 的符號形式，和 CFG 有些不同：

- 文法規則由一串符號構成，其中有個大寫符號（V, N, 或 ADJ ）代表中心語。其餘的符號為小寫的文法結構（'v', ‘n’, ‘adj’, ‘adv’, ‘prep’ 分別代表動詞片語、名詞片語、形容詞片語、副詞片語、介詞），但是也可以是特定的介詞。這些小寫的元素（個別，或整體而言）可以視為中心詞的句法搭配（grammatical collocations 或 colligations）。

[1] https://www.collinsdictionary.com/word-lovers-blog/new/what-are-grammar-patterns,524,HCB.html

- 小寫元素除了上述的幾種之外，還可以是句法結構（that 子句）、動詞型態（如 to-inf, inf, v-ing, v-ed, passive）、虛詞（wh, ord）語意搭配（如 amount, color, number, name）或者表面詞彙（如 get, be, way）。

所以，Pattern Grammar 是一種線性編碼，用以把通常的多層 CFG 規則的結構壓扁成為一層，並強調實詞（如動詞）與虛詞、文法結構（如副詞、子句）搭配的現象，可以直覺地溝通字詞的用法，明確地說明文法現象。因此，很適合作為辭典、教學之用。PG 最早應用在 Collins COBUILD English Dictionary (1995) 來標示詞條下每個詞意的文法編碼（Grammar coding）提供一種簡單容易理解的符號形式。然而，雖然簡明易解，PG 又是非常有彈性，有豐富的表達能力。而華語詞彙知識庫(如中研院 eHowNet)有句法訊息，但是並沒有透過嚴密的語料庫語言學分析，所以其文法規則的涵蓋不完備。有鑑於此，我們認為有必要透過英語文法規則的投射(projection) 來加速華語詞彙文法的研究，以提升未來華語辭典的文法方面的教學效果。

二、相關研究

用辭典、文法書、分級讀本來學習語言，有悠久的歷史，也是直覺合理的作法。然而，Sinclair (1991) 指出辭典傳統上過於重視詞意的解釋，而忽略了文法、語用的說明，也缺乏文體、領域的資訊。所以，Sinclair 主張運語料庫語言學，計算辭典學，詞彙索引典，來改善語言學習，和參考工具書的編輯。在知名的 Collins 出版社和 Birmingham 大學的 COBUILD 合作計畫下，Hunston, Francis, and Manning (1996, 1998) 出版了兩冊以詞彙為中心（lexical approach or lexical syllabus）的文法規則彙編——Collins COBUILD Grammar Patterns: Volumes 1 and 2。

在語料庫為本的研究上，Weber (2001) 描述如何運用語料庫和學生習作的交互作用，來幫助法學院學生的法律寫作。Sun (2007) 分析並標示語料庫，開發 *Scholarly Writing Template* (SWT) 系統，來幫助研究生寫作。我們聚焦在動詞的文法規則，不論是否和常見的寫作樣板有關。

在自動寫作評分的研究上，美國的 Education Test Service (ETS) 用機器學習與統計方法開發了 *Criterion* 系統 (Burstein, Chodorow and Leacock, 2003) 可以提供包含文法錯誤的寫作回饋。該系統已經用於對 4 to 12 年級的學生寫作，以及 TOFEL 和 GRE 的作文測驗部分。我們的研究集中在英語、華語的動詞文法規則，可以用來檢驗語言學習者最容易觸犯的動詞錯誤。

Chang and Chang (2015) 提出一個輕度督導式方法，自動推導英文的文法規則，並透夠提示來輔助學生寫作。這個系統的構想，延伸搜尋（如 *Google Suggest*）、翻譯（如 *TransType*）中的自動提示與完成的功能(Langlais, Foster and Lapalme, 2002)。

有別於先前的研究，我們提出一套開發文法標註資料集的作法，可以輔助語言分析師，標註資料集，以提供英語、華語文法，甚至雙語同步文法，自動推導的訓練資料，以期有助於華語，以及雙語同步文法的語言學，語言工程研究。

三、建構英華同步文法標註資料集

為了推導出華語文法規則，我們打算利用英華雙語例句，參照既有的英語文法規則，斯尋文法規則出現的實例，選取相關例句，並在例句的華語部分標註對應的華語文法規則。這項工作，並不如想像中那麼簡單。平行語料庫中的句子，常常結構過於複雜，有時也不容易找到特定既有英語文法規則的實例。所以，我們採取雙語辭典中的代表性例句，逕行此項工作。我們的方法，有三個步驟：產生雙語文法規則標示的草稿、人工標註、分析標註的結果。以下分別敘述之。

3.1 產生雙語標示詞料的草稿

我們將介紹一個將英語的動詞文法規則（grammar pattern） 找出對應的華語文法規則的方法。文法規則的資料來源為 Collins COBUILD Grammar Patterns: Verbs（arts-ccr-002.bham.ac.uk/ccr/patgram/）書中第二章節所列的規則，共 34 條。我們擷取符合 34 條規則的動詞。接著，我們蒐集劍橋大學線上英漢字典的例句（dictionary.cambridge.org/us/dictionary/english-chinese-traditional/）。我們先透過預處理辭例句，選出含相關動詞（如 talk）以及相關文法規則（我們透過相依關係分析，確認有 **V about n** 的句法結構，並擷取英文介詞、賓語）。接著，我們運用詞彙對應（word

alignment）的技巧，產生動詞、介詞、賓語的華語對應詞。憑藉著這些資料，我們產生如下格式的資料。標記的資料格式如下：資料點標號、四項資訊標號，四項資訊（包括英語句、華語句、英語實例＋文法規則、華語實例＋文法規則。以下顯現一筆 TALK: V about n 的例子：

1.1 ||| We were just talking about Gareth's new girlfriend

1.2 ||| 我們 剛才 在 談論 葛瑞 的 新 女友 。

1.3 ||| talk about girlfriend ||| talk : V about n

1.4 ||| 談論 女友 ||| 談論 : V n

第三項的動詞以及規則也為已知的資料，但仍須人工再確認是否和句子結構有相符。

第三項的英語詞組抽取方式為，先將 英語句子使用 spaCy 標記詞性，再抓取動詞與介系詞後面的名詞。其餘的項目為需要人工標記的部分，皆先使用 heuristic 的方式得到一個暫時的結果，再由人工校正。

第四項的華語動詞透過事先訓練好的中英雙語 word2vec model，計算英語動詞與華語句中詞語的相似度，找出相似度最高的華語，作為英語動詞的翻譯。華語詞組抽取即是透過找出的華語動詞，經由 Stanford Parser 得到華語句子的相依關係（dependency），找到與動詞相依的介系詞和名詞。

3.2 資料集標註指南

人工標記校正的準則，英文、中文分別列舉如下：

3.2.1 英語標註原則

應正確地標示出規則的實例，確認「動詞、介詞、賓語」是否正確，並確認由介詞與賓語組成的介賓詞組為動詞的必備成分。以下按賓語的詞類分項說明標注原則：

(a) 賓語若為名詞組：確認抽取實例是否為名詞組的中心語，而非修飾語或是所有格。如例子 (1)。

(1) We were just **talking about** Gareth's new **girlfriend**.

(b) 賓語若為動詞組，實例應擷取動詞組中的動詞原型。如例子 (2)。

(2) I'm **thinking about buying** a new car.

(c) 賓語若為疑問子句，應該標示 *wh*-詞以及 “to” (若句子含有 “to”)。如例子(3)。

(3) We couldn't **agree on what to buy**.

(d) 賓語若多於 1 個字, 則用底線連接，如例句 (4)。

(4) Who is going to **speak for** (= represent in a court of law) **the accused**.

此外，依照文法規則實例的情況，標示文法規則的元素，包含中心詞、V、介詞、賓語類型，如例句 (5) – (8)。

(5) We were just talking about Gareth's new girlfriend. （talk : V about n）

(6) I'm thinking about buying a new car. (think : V about ving)

(7) She was dithering about what to wear. (dither:V about wh_to-inf)

(8) Nick was enthusing about how well things worked out. (enthuse:V about wh)

3.2.2 華語標註原則

在華語詞組實例抽取的部分，與英文一樣，主要任務為確認「介系詞」、「賓語」以及「動詞」，並確認由介詞與賓語組成的介賓詞組為動詞的必備成分，標記資料中部分資料的斷詞不理想，但考慮後續訓練模型方便，標記時盡量維持原始資料的斷詞，不做修改。中文與英文的詞彙語法不同，有中英文能直接對應的語料，但有許多語料無法直接對應，因此華語標示原則在力求保持中英語料的平行性，以求能區分華語不同結構。以下分賓語、動詞進行討論。

3.2.3 華語文法規則標註原則

依照文法規則實例的情況，標示文法規則的元素，包含中心詞、V、介詞、賓語類型，中心詞以動詞實例標注，介系詞以介系詞實例標注，如(27-29)。

(27) 下 一 位 講 者 將 **就** 瀕危 **昆蟲 作 報告** 。（作_報告:就 n V）

(28) Andrew **qualified as** a **teacher** in 1995.

安德魯 於 1995 年 **取得 教師 資格** 。 (取得^資格 : V n)

(29) 餐館 裏 所有 的 人 都 **擁 到 他們 周圍** 唱 了 起來 。(擁:V 到^周圍 n)

若英文介系詞對應到華語動詞，則標註為 “v n”，如(30)呈現。

(30) Janet is **speaking for** the **motion**.
珍妮 特 **發言 支持** 這 項 **動議**。 (發言 : V v n)

賓語則按詞類標註：

(a) 賓語為名詞，標註為 n，如(27-30)。

(b) 賓語為動詞組，若英語有相對應的動詞，標示為 vp 如(31)，否則標註為 v_n 如(32)。

(31) I'm **thinking about buying** a new car.
我 在 **考慮 買** 輛 新 車 。 (考慮 : V vp)

(32) She **groped for** her **glasses** on the bedside table.
她 在 床頭櫃 上 **摸索** 著 **找 眼鏡** 。(摸索 : V v_n)

(a) 賓語為疑問子句時，若疑問詞為子句的謂語，直接標註「疑問詞」，如(33)，若不為子句的謂語，則標註為「疑問詞_v」，如(34)。

(33) 你 **覺得** 這 個 改善 地鐵 系統 的 最 新 **計劃 怎麼樣** ？(覺得:V 怎麼樣)

(34) 你 就 **假裝** 好像 **什麼** 事 都 沒 **發生** 過 。 (假裝 : V 什麼_v)

(d) 英語賓語對應到華語賓語的修飾語，如(35)，應標註為 “n_的_n”。

(35) Did you **ask about** the **money**?
你 **打聽 錢 的 事** 了 嗎 ？ （打聽 : V n_的_n）

四、初步標示結果的摘要分析

英語動介賓的文法規則（包含各種介詞）對應到的華語文法規則，有以下幾個現象：

1. 視介詞的不同（如 about 和 against），其對應的華語規則，有很大差異。所以，文法規則凸顯個別介詞，是很有必要的。對機器學習、語言學習都會有比較好的效果。

2. 一般而言，不論介詞為何，英語不及物動詞的 V p n 的文法規則，對應到華語動詞

文法，集中在不及物轉為及物（V n），介詞片語往前移動（p n V），或規則不變（V p n）三大類。而英語介詞（p）所對應的華語介詞，變化範圍也不大。

3. 少部分的介詞（如 against, for）有很強的傾向，對應到華語動詞（如「反對」、「支持」、「贊成」等）。如此一來，對應的華語規則，就變成 V v n。

4. 對應到的華語介詞有有一些現象和英語介詞不同。具體而言，英語介詞片語，對應到華語常常傾向於「介詞＋名詞＋方位詞」的型態。例如 CHASE: V after n 對應到「追：在 n 後面 V」。為了表示「在」和「後面」的成雙成對的關係，我們改寫為「在^後面 n V」。其中的「^」符號，表示其後的 「n」插入「在」和「後面」兩者之間。類似辭典常用的 「在～後面」或「在 … 後面」的表達方式。

5. 歸納起來，有近 6 成，對應的華語文法規則，是 p n V 或 V p n。另外，有三成多為 V v n。介詞以「在」或「在^方位詞」為多，其餘的介詞包括「向、為、冲、朝、對、就、與、跟、和、到鬧、於、給、替」等等。

五、結論

未來有許多方向可以繼續探索，並改進目前的作法和結果。例如，目前抽取英文規則的方法還可以考慮用搭配的統計分析，篩選教具代表性的例子，同時規避剖析錯誤，抽取了不正確的英文規則。我們也可以透過更大量的雙語資料分析，得到比較正確的華語文法規則的建議，減低語言分析師的工作負荷。另外一個有趣的研究方向，是擴大英語、華語文法規則的範圍，以包括更複雜的句法現象。更進一步的研究，應該從動詞，延伸到名詞、形容詞。還有一個更重要的研究方向，是用完成的資料集，訓練一個自動產生雙語同步文法規則的系統。該系統的結果，可以作為編撰華語辭典，開發機器翻譯系統的基礎。

總結起來，我們呈現一個方法，可以採電腦輔助的方式，開發同步文法規則資料集。我們的研究路線涉及運用了辭典中的雙語例句, 來發掘英語、華語的動詞的對應句法規則。我們的方法有三個步驟：自動辨識例句中的英語文法規則、運用詞彙對應的技巧產生華語規則的建議、人為分析產生正確英華同步文法規則的資料集。我們把這個研究方法, 運用在劍橋大學出版社的線上英漢辭典的例句, 初步完成英語「動介賓」規則的華

語對應規則的分析。我們就初步的研究結果，說明標註華語文法規則的指導原則，觀察分析所得到的華語文法規則的統計分布。最後完成的資料集，可望有助於提供華語文法規則自動擷取的機器學習研究。

參考文獻

[1] Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." (1993).

[2] Sinclair, John. "Lexical grammar." *Naujoji Metodologija* 24 (2000): 191-203.

[3] Hunston, Susan, and Gill Francis. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Vol. 4. John Benjamins Publishing, 2000.

[4] Huang, Chu-Ren, et al. "Sinica Treebank: design criteria, annotation guidelines, and on-line interface." *Second Chinese Language Processing Workshop*. 2000.

[5] Xue, Naiwen, et al. "The Penn Chinese TreeBank: Phrase structure annotation of a large corpus." *Natural language engineering* 11.2 (2005): 207-238.

[6] Xue, Nianwen, and Martha Palmer. "Annotating the propositions in the Penn Chinese Treebank." *Proc of the 2nd SIGHAN workshop on Chinese language processing*, 2003.

[7] Johnson, Christopher (2000). "Review of Pattern grammar: A corpus-driven approach to the lexical grammar of English". Computational Linguistics. **27** (4): 318–320.

[8] Francis, Gill; Hunston, Susan; Manning, Elizabeth, Collins COBUILD Grammar Patterns 1: Verbs, HarperCollins, 1996.

[9] Francis, Gill; Hunston, Susan; Manning, Elizabeth, Collins COBUILD Grammar Patterns 2: Nouns and Adjectives, HarperCollins, 1997.

[10] Hunston, Susan; Francis, Gill, Pattern Grammar: A corpus-driven approach to the lexical grammar of English, John Benjamins, 2000.

[11] Francis, G. (1993). A corpus-driven approach to grammar – principles, methods and examples. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds). *Text and Technology: in Honour of John Sinclair*. Amsterdam: Benjamins, pp. 137–156.

Francis, G., Hunston, S. & Manning, E. (1996). *Collins COBUILD Grammar Patterns 1:*

Verbs. London: HarperCollins. Francis, G., Hunston, S. & Manning, E. (1998). *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

[12] Hunston, S. & Francis, G. (2000). Pattern Grammar: A corpus-driven approach to the lexical grammar of English. Amsterdam: Benjamins.