

**ROCLING 2020: The 32nd Conference on
Computational Linguistics and Speech Processing**

第三十二屆自然語言與語音處理研討會

Sep. 24-26, 2020

National Taipei University of Technology, Taipei, Taiwan, ROC

主辦單位：

國立臺北科技大學、國立陽明大學、中華民國計算語言學學會

協辦單位：

中央研究院資訊科學研究所、中央研究院資訊科技創新研究中心

贊助單位：

科技部、教育部、賽微科技股份有限公司、中華電信研究院、台達電子工業股

份有限公司、科技部補助人工智慧生技醫療創新研究中心、財團法人國家實驗

研究院科技政策研究與資訊中心

First Published in September 2020

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

Copyright©2020 the Association for Computational Linguistics and Chinese Language
Processing (ACLCLP), Authors of Papers

Each of the authors grants a non-exclusive license to the ACLCLP to publish the paper
in printed form. Any other usage is prohibited without the express permission of the
author who may also retain the on-line version at a location to be selected by him/her.

Jenq-Haur Wang, Ying-Hui Lai, Lung-Hao Lee, Kuan-Yu Chen, Hung-Yi Lee, Chi-
Chun Lee, Syu-Siang Wang, Hen-Hsen Huang, and Chuan-Ming Liu (eds.)

Proceedings of the 32nd Conference on Computational Linguistics and Speech
Processing (ROCLING XXXII)

2020-09-24—2020-09-26

ACLCLP

2020-09

ISBN: 978-986-95769-3-2

Organizing Committee

Conference Co-Chairs

Jenq-Haur Wang, National Taipei University of Technology

Ying-Hui Lai, National Yang-Ming University

Program Chairs

Lung-Hao Lee, National Central University

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

Tutorial Chair

Hung-Yi Lee, National Taiwan University

Industry Chair

Chi-Chun Lee, National Tsing Hua University

Demo Chair

Syu-Siang Wang, Academia Sinica

Publication Chair

Hen-Hsen Huang, National Chengchi University

Web Chair

Chuan-Ming Liu, National Taipei University of Technology

Program Committee

Guo-Wei Bian (邊國維), Huafan University
Chia-Hui Chang (張嘉惠), National Central University
Ru-Yng Chang (張如瑩), AI Clerk International Co., LTD.
Yu-Yun Chang (張瑜芸), National Chengchi University
Yung-Chun Chang (張詠淳), Taipei Medical University
Cheng-Hsien Alvin Chen (陳正賢), National Taiwan Normal University
Chung-Chi Chen (陳重吉), National Taiwan University
Fei Chen (陳霏), Southern University of Science and Technology
Mei-Hua Chen (陳玫樺), Tunghai University
Yun-Nung (Vivian) Chen (陳縕儂), National Taiwan University
Tai-Shih Chi (冀泰石), National Chiao Tung University
Jia-Fei Hong (洪嘉馥), National Taiwan University
Shu-kai Hsieh (謝舒凱), National Taiwan University
Chun-Hsien Hsu (徐峻賢), National Central University
Yi-Chin Huang (黃奕欽), National Pingtung University
Hen-Hsen Huang (黃瀚萱), National Chengchi University
Jeih-weih Hung (洪志偉), National Chi Nan University
Wen-Hsing Lai (賴玟杏), National Kaohsiung First university of Science and Technology
Ying-Hui Lai (賴穎暉), National Yang Ming University
Hong-Yi Lee (李宏毅), National Taiwan University
Lung-Hao Lee (李龍豪), National Central University
Yuan-Fu Liao (廖元甫), National Taipei University of Technology
Chuan-Jie Lin (林川傑), National Taiwan Ocean University
Shu-Yen Lin (林淑晏), National Taiwan Normal University
Chao-Lin Liu (劉昭麟), National Chengchi University
Yi-Fen Liu (劉怡芬), Feng Chia University
Shih-Hung Liu (劉士弘), Delta Electronics, Inc.
Wen-Hsiang Lu (盧文祥), National Cheng Kung University
Ming-Hsiang Su (蘇明祥), Soochow University
Richard Tzong-Han Tsai (蔡宗翰), National Central University
Wei-Ho Tsai (蔡偉和), National Taipei University of Technology
Yuen-Hsien Tseng (曾元顯), National Taiwan Normal University
Jenq-Haur Wang (王正豪), National Taipei University of Technology

Syu-Siang Wang (王緒翔), National Taiwan University
Hsin-Min Wang (王新民), Academia Sinica
Jiun-Shiung Wu (吳俊雄), National Chung Cheng University
Shih-Hung Wu (吳世弘), Chaoyang University of Technology
Jheng-Long Wu (吳政隆), Soochow University
Cheng-Zen Yang (楊正仁), Yuan Ze University
Yi-Hsuan Yang (楊奕軒), Academia Sinica
Jui-Feng Yeh (葉瑞峰), National Chia-Yi University
Liang-Chih Yu (禹良治), Yuan Ze University

Welcome Message from ROCLING 2020

On behalf of the organizing committee, it is our pleasure to welcome you to National Taipei University of Technology (NTUT), Taipei, Taiwan, for the 32nd 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 Association for Computational Linguistics and Chinese Language Processing (ACLCLP) which is regularly held by different universities in different cities of Taiwan.

ROCLING 2020 features two distinguished keynote speeches from the renowned researchers in natural language processing as well as speech processing. Prof. Tomoki Toda (Professor, Information Technology Center, Nagoya University, Japan) will give a keynote on the “Recent Trend of Voice Conversion Research and Its Possible Future Direction”. Prof. Hiroyuki Shinnou (Professor, Department of Computer and Information Sciences, Ibaraki University, Japan) will talk about the “Use of BERT for NLP tasks by HuggingFace's transformers”.

ROCLING 2020 is going to 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. To facilitate more cross-domain communication and collaboration, we organize a special session on Natural Language Processing for Digital Humanities with Taiwanese Association for Digital Humanities (TADH). In addition to the regular sessions during the first two days, the AI Tutorial organized by SIG-AI (Artificial Intelligence Special Interest Group) of ACLCLP and the Science & Technology Policy Research and Information Center (STPI) will provide Artificial Intelligence Courses that focus on speech processing and NLP applications on the last day. It's sure to be an exciting event for all participants.

This conference would not have been possible without the tremendous effort of organizing committee and program committee who have worked closely to put together the attractive and intensive scientific program. Their great achievements have contributed much to the visibility of ROCLING 2020. We would like to express our sincere thank and gratitude to all of them. 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. Thanks to all organizers including Program Chairs: Lung-Hao Lee and Kuan-Yu Chen, Tutorial Chair: Hung-Yi Lee, Industry Chair: Chi-Chun Lee, Demo Chair: Syu-Siang Wang, Publication Chair: Hen-Hsen Huang, Web Chair: Chuan-Ming Liu. Thanks to special session organizer: Chao-Lin Liu, and the invited speakers: Jen-Jou Hung, Su-bing Chang, and Wu, wan-yi. Thanks to all participants, authors, and program committee members and reviewers who contributed 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. With the best for a successful and fruitful ROCLING 2020 in Taipei, Taiwan.

General Chairs

Jenq-Haur Wang and Ying-Hui Lai

Keynote Speaker I



Tomoki Toda

Professor, Information Technology Center, Nagoya University, Japan

Biography

Tomoki Toda was born in Aichi, Japan on January 18, 1977. He earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively.

He is a Professor at the Information Technology Center, Nagoya University. He has also been a Visiting Researcher at the NICT, Kyoto, Japan, since 2006. He was a Research Fellow of JSPS in the Graduate School of Engineering, Nagoya Institute of Technology, Aichi, Japan, from 2003 to 2005. He was then an Assistant Professor (2005-2011) and an Associate Professor (2011-2015) at the Graduate School of Information Science, NAIST. From 2001 to 2003, he was an Intern Researcher at the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan, and then he was a Visiting Researcher at the ATR until 2006. He was also a Visiting Researcher at the Language Technologies Institute, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the Department of Engineering, University of Cambridge, Cambridge, UK, from March to August 2008. His research interests include statistical approaches to speech, music, and sound information processing.

He received more than 10 paper awards including the 18th TELECOM System Technology Award for Students and the 23rd TELECOM System Technology Award

from the TAF, the 2007 ISS Best Paper Award from the IEICE, the 2009 Young Author Best Paper Award from the IEEE SPS, and the 2013 Best Paper Award (Speech Communication Journal) from EURASIP-ISCA. He also received the 10th Ericsson Young Scientist Award from Nippon Ericsson K.K., the 4th Itakura Prize Innovative Young Researcher Award from the ASJ, the 2012 Kiyasu Special Industrial Achievement Award from the IPSJ, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, the Young Scientists' Prize in 2015. He served as a member of the Speech and Language Technical Committee of the IEEE SPS from 2007 to 2009 and 2014 to 2016. He has served as an Associate Editor of IEEE Signal Processing Letters since Nov. 2016. He is a member of IEEE, ISCA, IEICE, IPSJ, and ASJ.

Keynote Speech A

Recent Trend of Voice Conversion Research and Its Possible Future Direction

September 24, 2020 (Thursday) 9:30-10:30

Venue: The Lecture Hall, GIS Convention Center

Abstract

Voice conversion is a technique for modifying speech waveforms to convert non-/paralinguistic information into any form we want while preserving linguistic content. It has been dramatically improved thanks to significant progress in machine learning techniques, such as deep learning, as well as significant efforts to develop freely available resources. In this talk, I will review recent progress of voice conversion techniques, overviewing recent research activities including Voice Conversion Challenges, and then, I will also discuss possible future directions of voice conversion research.

Keynote Speaker II



Hiroyuki Shinnou

Professor, Department of Computer and Information Sciences, Ibaraki University,
Japan

Biography

Prof. Hiroyuki Shinnou worked as a researcher in Fuji Xerox Co., Ltd. and Panasonic Corporation during 1987 and 1993. He joined the Faculty of Engineering, Ibaraki University in 1993, as a research assistant. After receiving his Ph.D. degree in Tokyo Institute of Technology in 1997, he worked as a lecturer, and an associate professor in Ibaraki University, respectively. He is currently a professor at the Department of Computer and Information Sciences in Ibaraki University.

Prof. Shinnou has long been active in the academic associations related to natural language processing, including ACL (Association of Computational Linguistics), JSAI, and IPSJ. Now, he serves as the director of the Association for Natural Language Processing (ANLP), and serves as the conference chairman in the annual conference NLP 2020 this year, which is the most important conference on Natural Language Processing in Japan.

Prof. Shinnou has published many academic papers in international journals such as ACM TALIP, and Natural Language Processing (in Japanese), and international conferences including ACL, PACLIC, LREC. His research interests include Bayes statistics, machine learning, natural language processing and image processing. Since he integrates theory with practice, he also published many books (mostly in Japanese),

which have great impact in related fields. Recently, he is actively researching deep learning technology, especially transfer learning.

Keynote Speech B

Use of BERT for NLP tasks by HuggingFace's transformers

September 25, 2020 (Friday) 9:30-10:30

Venue: The Lecture Hall, GIS Convention Center

Abstract

The pre-trained BERT model has been improving states of many NLP tasks. I believe that the use of BERT is essential when we build some kind of NLP system in the future. Initially, it was hard to use BERT because the concept of the pre-trained model was unfamiliar, and BERT was available only by using TensorFlow which is cumbersome for beginners. However, today, there is the HuggingFace's transformers library. Thanks to this library, everyone can utilize BERT easily.

In this talk, first I will explain what BERT is and what we can do by BERT, and then I show some examples of the use of BERT by HuggingFace's transformers. As an application, I will do fine-tuning of BERT for a document classification task. Additionally, I will show the technique to learn just some of the layers in BERT. As one of the improvements of BERT, the study on smaller BERT model have been active, for example, Q8BERT, ALBERT, DistilBERT, TinyBERT and so on. Even simple pruning of BERT is effective. I will introduce these studies and show that some of these models are available through HuggingFace's transformers.

Contents

Oral Papers

Analyzing the Morphological Structures in Seediq Words	1
Gated Graph Sequence Neural Networks for Chinese Healthcare Named Entity Recognition	4
Improving Phrase Translation Based on Sentence Alignment of Chinese-English Parallel Corpus	6
Mitigating Impacts of Word Segmentation Errors on Collocation Extraction in Chinese	8
Japanese Word Readability Assessment using Word Embeddings	21
A Hierarchical Decomposable Attention Model for News Stance Detection	35
Combining Dependency Parser and GNN models for Text Classification	50
A Preliminary Study on Using Meta-learning Technique for Information Retrieval	59
NLLP for the Understanding and Prediction of Construction Litigation Based on Multiple BERT Model	72
Real-Time Single-Speaker Taiwanese-Accented Mandarin Speech Synthesis System	87
Taiwanese Speech Recognition Based on Hybrid Deep Neural Network Architecture	102
NSYSU+CHT Speaker Verification System for Far-Field Speaker Verification Challenge 2020	114

A Preliminary Study on Deep Learning-based Chinese Text to Taiwanese Speech Synthesis System	116
The preliminary study of robust speech feature extraction based on maximizing the accuracy of states in deep acoustic models	118
Multi-view Attention-based Speech Enhancement Model for Noise-robust Automatic Speech Recognition	120
A Preliminary Study on Leveraging Meta Learning Technique for Code-switching Speech Recognition	136
Innovative Pretrained-based Reranking Language Models for N-best Speech Recognition Lists	148
Lectal Variation of the Two Chinese Causative Auxiliaries	163
The Semantic Features and Cognitive Concepts of Mang2 ‘Busy’: A Corpus-Based Study	178
An Analysis of Multimodal Document Intent in Instagram Posts	193

Posters and System Demonstrations

A Chinese Math Word Problem Solving System Based on Linguistic Theory and Non-statistical Approach	208
An Adaptive Method for Building a Chinese Dimensional Sentiment Lexicon	223
Nepali Speech Recognition Using CNN, GRU and CTC	238
A Study on Contextualized Language Modeling for FAQ Retrieval	247
French and Russian students’ production of Mandarin tones	260
Sentiment Analysis for Investment Atmosphere Scoring	275
Exploiting Text Prompts for the Development of an End-to-End Computer-Assisted Pronunciation Training System	290
Combining Hybrid Attention Networks and LSTM for Stock Trend Prediction	304

Low False Alarm Rate Chinese Misspelling Detection Model Based on BERT Task Model	319
The Analysis and Annotation of Propaganda Techniques in Chinese News Texts	331
Exploring Disparate Language Model Combination Strategies for Mandarin-English Code-Switching ASR	346
Scientific Writing Evaluation Using Ensemble Multi-channel Neural Networks	359
Building A Multi-Label Detection Model for Question classification of Auction Website	372
Email Writing Assistant System	387
Aspect-Based Sentiment Analysis Based on BERT-DAOA	398

Special Session: NLP for Digital Humanities

Natural Language Processing for Digital Humanities	413
The Opportunities and Challenges of Natural Language Processing Technology in the Field of Digital Humanities—Taking the Study of Buddhist Scriptures as an Example	415
The Taiwan Biographical Database (TBDB): An Introduction	418
How to Analyze the Related Materials of Traditional Chinese Drama in the Early 20th Century (1900–1937) from the Perspective of Digital Humanities—Focusing on Newspaper Databases, Record Databases, and Script Collections	421
Optical Character Recognition, Word Segmentation, Sentence Segmentation, and Information Extraction for Historical and Literature Texts in Classical Chinese	423

賽德克語構詞結構之自動解析

Analyzing the Morphological Structures in Seediq Words

林川傑 Chuan-Jie Lin[†]

國立臺灣海洋大學資訊工程學系

Department of Computer Science and Engineering

National Taiwan Ocean University

cjlin@email.ntou.edu.tw

宋麗梅 Li-May Sung[†]

國立臺灣大學語言學研究所

Graduate Institute of Linguistics

National Taiwan University

limay@ntu.edu.tw

游景勝 Jing-Sheng You, 王瑋 Wei Wang, 李政勳 Cheng-Hsun Lee,

廖子權 Zih-Cyuan Liao

國立臺灣海洋大學資訊工程學系

Department of Computer Science and Engineering

National Taiwan Ocean University

{10857039, 00657120, 00657140, 00672042}@email.ntou.edu.tw

摘要

原住民族語言保存及振興的問題已日益受到重視。如果能開發出原住民族語相關自然語言處理技術，有助於原住民語資料保存及族語推廣等工作。賽德克語的詞形變化相當多樣，其中一大部分主要是為了標示動詞焦點或時貌，包括完成貌、主事焦點、受事焦點、處所焦點等等。因為這種焦點系統為南島語系所特有，若要研究台灣原住民語和中文之間的自動翻譯系統，辨別這類構詞資訊非常重要。

[†] 通訊作者 corresponding authors

一個賽德克詞的構詞結構會以它所參考的原形詞加上前、中、後綴的組合來呈現，然而構詞結構資訊無法由詞面直接獲得，詞典中也僅能查到各賽德克詞參考的原形詞。更特別的是，賽德克語構詞律有元音脫落規則，因此賽德克詞並非直接由原形詞接上詞綴而得。因此本論文的主要目標是自動解析賽德克語的構詞結構，在給定一個賽德克詞及其參考原形詞時，能夠解析出該賽德克詞裡所出現的前綴、中綴及後綴組合。

此外，賽德克語構詞律有元音中性化和詞尾輔音變化等等規則。在研究構詞情形的過程中我們發現，加上後綴時原形詞部份會恢復回變化之前的樣子，我們將之定義為「深層原形」。因為字典中並無此項資訊，本論文也會探討如何猜測一個賽德克詞的深層原形。實驗資料主要來自宋麗梅教授著作「賽德克語語法概論」及協助原住民族委員會開發的「賽德克語德固達雅方言」線上詞典。

首先由語法書中整理出的構詞相關知識來撰寫規則，用以偵測詞綴的出現。深層原形則是利用詞典中參考同一原形詞的不同賽德克詞來統計猜測，這些猜測結果又可再歸納出常見的變化規則用來推測新詞彙的深層原形。詞綴及深層原形解析工作在測試資料的精確率是 98.66%，而召回率是 88.29%。

至於前綴的部份，因為同一個前綴字串可能可以拆解出多種不同的前綴組合而產生歧異情形，因此改以機器學習方式進行。在測試各種方法後，效果最好的是以基本前綴為單位的二元機率模型。解決零機率的方法是降階至一元機率模型（權重設為 α ），而一元機率模型解決零機率的方法又以 Lidstone smoothing 效能最好（頻率增加值設為 λ ）。前綴組合最佳解析正確率為 76.92%。

Abstract

The issue of preservation and revitalization of the indigenous languages is gaining attention from the public in recent days. Developing NLP techniques related to the indigenous languages will help to preserve and promote these languages. Word inflection or morphological forms in Seediq are plentiful. Major categories of the inflections are mainly for representing the focus or aspect, such as perfective aspect, active voice, patient voice, locative voice, etc. The focus system of the Austronesian languages is quite different from Chinese. It is important to identify the information of focus or aspect in words if we want to study machine translation among Taiwanese indigenous languages and other languages.

The morphological structure of a Seediq word consists of its word root, prefixes, infixes, and suffixes. This kind of information cannot be obtained directly from the surface of a Seediq

word. Dictionaries only offer the information of word roots. Furthermore, due to the rule of vowel reduction in Seediq, the surface of a Seediq word is not the same as the concatenation of affixes and word root. This paper focuses on automatically analyzing the morphological structure of a Seediq word given its word root.

Moreover, there are also rules of vowel neutralization and final consonant variation. During the research, we found that a word root would return to its original form when combining with the suffixes. We define the original form of a root word as a “deep root”. Since there is no information about deep roots in the dictionary, this paper also proposes methods to predict deep roots of Seediq words. The experimental data come from the works of Prof. Li-May Sung: the grammar book “賽德克語語法概論” (An Introduction to Seediq Grammar) and the online dictionary “賽德克語德固達雅方言” (Tgdaya Seediq) from the Council of Indigenous Peoples.

First, several morphological analyzing rules were created from the knowledge provided in the grammar book. These rules were used to detect the occurrences of affixes. Deep roots were learned from the set of different words referencing to the same root words. The mapping of root words with their deep roots could be further used to derive deep-root-prediction rules for unknown words. The rule-based system successfully detected the deep root and the existence of affixes with a precision of 98.66% and a recall of 88.29% on the test data.

Because one prefix string can be divided into several different structures, we used machine learning methods to solve the ambiguity. The best system was developed by bigram model where grams were atomic prefixes. Zero probability in the bigram model was replaced by the unigram probability (weighted by α), where the unigram model was also smoothed by the Lidstone smoothing method (with an addition of λ to the frequencies). The best prefix analysis system achieved an accuracy of 76.92% on the test data.

關鍵詞：賽德克語，構詞結構自動解析，深層原形，臺灣原住民族語之自然語言處理

Keywords: Seediq, automatic analysis of morphological structures, deep root, natural language processing for Taiwanese indigenous languages

致謝

本論文之研究承蒙行政院科技部研究計畫 (MOST 109-2221-E-019 -053 -) 之經費支持，謹此致謝。

門控圖序列神經網路之中文健康照護命名實體辨識

Gated Graph Sequence Neural Networks for Chinese Healthcare Named Entity Recognition

盧毅 Yi Lu, 李龍豪 Lung-Hao Lee
國立中央大學電機工程學系
Department of Electrical Engineering
National Central University
ericst91159@gmail.com、lhlee@ee.ncu.edu.tw

摘要

命名實體辨識任務的目標是從非結構化的輸入文本中，抽取出關注的命名實體，例如：人名、地名、組織名、日期、時間等專有名詞，擷取的命名實體，可以做為關係擷取、事件偵測與追蹤、知識圖譜建置、問答系統等應用的基礎。機器學習的方法將其視為序列標註問題，透過大規模語料學習標註模型，對句子的各個字元位置進行標註。我們提出一個門控圖序列神經網路 (Gated Graph Sequence Neural Network, GGSNN) 模型，用於中文健康照護領域命名實體辨識，我們整合詞嵌入以及部首嵌入的資訊，建構多重嵌入的字嵌入向量，藉由調適門控圖序列神經網路，融入已知字典中的命名實體資訊，然後銜接雙向長短期記憶類神經網路與條件隨機場域，對中文句子中的字元序列標註。我們透過網路爬蟲蒐集健康照護相關內容的網路文章以及醫療問答紀錄，然後隨機抽取中文句子做人工斷詞與命名實體標記，句子總數為 30,692 句 (約 150 萬字/91.7 萬詞)，共有 68,460 命名實體，包含 10 個命名實體種類：人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療與時間。藉由實驗結果與錯誤分析得知，我們提出的模型達到最好的 F1-score 75.69%，比相關研究模型 (BiLSTM-CRF, BERT, Lattice, Gazetteers 以及 ME-CNER) 表現好，且為效能與效率兼具的中文健康照護命名實體辨識方法。

關鍵詞：命名實體辨識、圖神經網路、資訊擷取、健康資訊學

Abstract

Named Entity Recognition (NER) focuses on locating the mentions of name entities and classifying their types, usually referring to proper nouns such as persons, places, organizations, dates, and times. The NER results can be used as the basis for relationship extraction, event detection and tracking, knowledge graph building, and question answering system. NER studies usually regard this research topic as a sequence labeling problem and learns the labeling model through the large-scale corpus. We propose a GGSNN (Gated Graph Sequence Neural Networks) model for Chinese healthcare NER. We derive a character representation based on multiple embeddings in different granularities from the radical, character to word levels. An adapted gated graph sequence neural network is involved to incorporate named entity information in the dictionaries. A standard BiLSTM-CRF is then used to identify named entities and classify their types in the healthcare domain. We firstly crawled articles from websites that provide healthcare information, online health-related news and medical question/answer forums. We then randomly selected partial sentences to retain content diversity. It includes 30,692 sentences with a total of around 1.5 million characters or 91.7 thousand words. After manual annotation, we have 68,460 named entities across 10 entity types: body, symptom, instrument, examination, chemical, disease, drug, supplement, treatment, and time. Based on further experiments and error analysis, our proposed method achieved the best F1-score of 75.69% that outperforms previous models including the BiLSTM-CRF, BERT, Lattice, Gazetteers, and ME-CNER. In summary, our GGSNN model is an effective and efficient solution for the Chinese healthcare NER task.

Keywords: Named Entity Recognition, Graph Neural Networks, Information Extraction, Health Informatics

致謝 Acknowledgements

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3.

奠基於雙語自動對齊之
動介片語翻譯改進

Improving Phrase Translation Based on Sentence Alignment
of Chinese-English Parallel Corpus

陳怡君 Yi-Jyun Chen
國立清華大學資工系
Department of Computer Science
National Tsing Hua University
yijyun@nplab.cc

楊馨瑜 Ching-Yu Helen Yang
國立中興大學外文系
Department of Foreign Languages and Literatures
National Chung Hsing University
chingyu@nplab.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@nplab.cc

摘要

本研究呈現片語查詢的雛形系統 PrecisePhraseBook，能從雙語語料庫中，自動擷取英文名詞與介系詞搭配的中文翻譯及例句，可輔助語言學習者學習語言，亦可改善機器翻譯或提供語言研究者撰寫文法規則之參考。本方法利用雙語語料庫擷取英文片語的中文翻譯。其方法為使用統計方法由雙語語料庫中的詞彙自動對齊，分別擷取名詞及介系詞的翻譯，再根據由中文語料庫統計而來的中文高頻搭配詞，將名詞及介系詞的翻譯做適當調整，並產生例句。系統執行時，使用者輸入一組英文名詞與介系詞的搭配，系統會呈現資料庫中此搭配的翻譯及例句。本研究的評估方式是隨機抽取三十組名詞及介系詞的搭配，人工評估本研究方法產生的翻譯。

關鍵詞：雙語句子對齊、文法規則、搭配詞、動介片語翻譯

Abstract

This thesis presents a phrases searching system, PrecisePhraseBook, which provides Chinese translations and example sentences of English phrases with a noun and preposition to assist learners in learning English or Chinese. PrecisePhraseBook provides researchers a reference tool for generating grammar rules. We propose a method for extracting Chinese translations of English phrases from bilingual parallel corpora. We use statistical methods to extract translations of nouns and prepositions from bilingual parallel corpora with sentence alignment, and then adjust the translations according to the Chinese collocations extracted from a Chinese corpus. Finally, we generate example sentences for the translations. At run-time, the user enters an English phrase with a noun and a preposition, and the system retrieves translations and example sentences from the database and presents the results to the user. The evaluation is done using randomly 30 selected phrases. We used human judge in assess the translations.

Keywords: Sentence Alignment, Grammar Patterns, Collocations, Phrase Translation

Mitigating Impacts of Word Segmentation Errors on Collocation Extraction in Chinese

廖永賦 Yongfu Liao
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
liao961120@gmail.com

謝舒凱 Shu-Kai Hsieh
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

摘要

隨著網路的盛行，自動斷詞與標記的大規模語料庫逐漸普及。自動化不可避免地引入一些斷詞與標記的錯誤，並可能對下游任務產生負面影響。搭配詞的自動抽取是一項受斷詞品質影響的任務。本文探討一些方法試圖減輕斷詞錯誤對漢語搭配詞抽取之影響。我們嘗試了一個結合多個共現訊息的簡單線性模型，試圖減少抽取出之搭配詞含有的斷詞錯誤。實驗結果顯示，此模型無法區分搭配詞是否為斷詞錯誤所導致。因此，我們使用了 FastText 詞向量的訊息進行了另一個案例研究。結果顯示，由斷詞錯誤所產生的假搭配詞與真正的搭配詞，其之間的語義相似性具有不同的特徵。未來研究可嘗試在搭配詞抽取中加入詞向量的訊息。

Abstract

The prevalence of the web has brought about the construction of many large-scale, automatically segmented and tagged corpora, which inevitably introduces errors due to automation and are likely to have negative impacts on downstream tasks. Collocation extraction from Chinese corpora is one such task that is profoundly influenced by the quality of word segmentation. This paper explores methods to mitigate the negative impacts of word segmentation errors on collocation extraction in Chinese. In particular, we experimented with a simple model that aims to combine several association measures linearly to avoid retrieving false collocations resulting from word segmentation errors. The results of the experiment show that this simple model could not differentiate between true collocations and false collocations

resulting from word segmentation errors. An ad hoc case study incorporating information from FastText word vectors is also conducted. The results show that collocates resulting from correct and erroneous word segmentation have different profiles in terms of the semantic similarities between the collocates. The incorporation of word vector information to differentiate between true and false collocations is suggested for future work.

關鍵詞：搭配詞抽取、中文斷詞、詞向量

Keywords: Collocation Extraction, Chinese Word Segmentation, Word Vector

1 Introduction

A collocation, in Firthian sense, is a combination of words that tend to occur near each other in natural language [1]. To measure the tendency for words to co-occur, various statistical measures are proposed to quantify the association strengths of word pairs. These association measures are often used to rank and extract collocations from corpora. As the concept of collocation was developed in the western world, which has a writing system that clearly delimits word boundaries, the adoption of the concept of collocation in languages where the notion of *wordhood* is not clear necessitates a preprocessing step that segments the text into sequences of “words”. Computing association measures based on the segmented text to extract collocations thus requires an additional assumption—the word segmentation must return correct results. Otherwise, the collocations extracted might be nonsense—instead of being recurrent “word” combinations, the “collocations” may in fact be “character” combinations that have a tendency to co-occur.

With the prevalence of the internet, large corpora constructed from texts collected from the web has become common. At the same time, manual checking of the automatic segmentation and tagging of the corpora to ensure the quality has become nearly impossible, as the amount of data collected is enormous. In addition, out-of-vocabulary words such as named entities, new terms, and special usage of particular subcultures frequently appear in web texts [2], further casting doubt on the performance of automatic segmentation of the constructed corpora.

Since manual checking and corrections are not practical solutions to counter automatic preprocessing errors in large corpora, it is crucial to be aware of the negative impacts that such errors could have on downstream tasks. For instance, collocations extracted from Chinese social media texts may contain several instances of false collocations that resulted from word

segmentation errors. Table 1 lists the top 16 collocates of the node word 三 ‘three’ retrieved from the social media PTT (see section 2.1 for the data). Strikingly, the top 10 ranked collocations in Table 1 all resulted from word segmentation errors. In other words, there are only 37.5% of the word pairs in this list that count as “true” collocations. The others are not even “word” pairs.

In this paper, we explore the potential of leveraging existing association measures, originally designed to quantify the association strengths between *words*, to detect or filter out false collocations resulting from word segmentation errors in collocation extraction. The assumption is that false collocations may behave differently from true collocates in the patterns of association measures. In particular, we explore the possibility of constructing a new association measure from existing ones that is robust against retrieving false collocations resulting from word segmentation errors.

Table 1. Top 16 collocates that have the highest tendency to co-occur (as measured by MI) with the node word 三 ‘three’ in PTT corpus.

Word 1	Word 2	Frequency	MI	Rank	Word 1	Word 2	Frequency	MI	Rank
三	池崇史	17	10.062	1	三	丁	219	9.154	9
	浦友	5	10.062	2		日月	18	9.022	10
	浦春馬	7	10.062	3		次元	24	7.559	11
	角頭	20	9.683	4		餐	261	7.457	12
	人房	8	9.603	5		小	2367	6.585	13
	人行	52	9.496	6		秒	82	5.655	14
	倍速	8	9.477	7		年前	95	5.563	15
	班制	12	9.325	8		年	716	5.377	16

2 Combining Association Measures

The purpose of this research is to explore the possibility of constructing a robust association measure by combining several association measures, with the aim of mitigating the impact of retrieving false collocations resulting from word segmentation errors in Chinese. Below, we describe the data, the model for combining several association measures, and the training of the model.

2.1 Data

As a preliminary study, we focus here only on association strengths of word pairs occurring in a running window of two (i.e., bigrams). The corpus used to calculate various association measures was constructed from 36,000 texts from PTT forum¹, which is one of the largest online forums in Taiwan. The texts were collected from 12 categories (*BabyMother*, *Boy-Girl*, *gay*, *Gossiping*, *Hate*, *HatePolitics*, *Horror*, *JapanMovie*, *joke*, *LGBT_SEX*, *NTU*, *sex*)², with 3000 texts sampled from each category. The corpus was segmented with Jseg³. Word pair frequencies were then calculated from the corpus.

Eight association measures—MI, MI³, MI.log-f, t, Dice, logDice, $\Delta P_{1|2}$, $\Delta P_{2|1}$ —were calculated from the corpus. The first six measures follow the statistics used in the Sketch Engine [3], and the last two measures, $\Delta P_{1|2}$ and $\Delta P_{2|1}$, are directional association measures proposed in [4]. The MI measure measures the ratio between the observed frequency ($O = f_{AB}$) and the expected frequency ($E = f_A \cdot f_B / N$) of a word pair (w_A, w_B) on a logarithmic scale. Since MI tends to assign low-frequency word pairs (having low value of E) high scores, variants of the MI measure are proposed to counter this effect. MI³ achieve this by taking the cube of the observed frequency to strengthen its influence relative to the expected frequency, and MI.log-f counters MI’s low-frequency bias by multiplying the MI score with $\ln(O + 1)$. T measures the discrepancy between the observed and expected frequency against the square root of the observed frequency. The Dice coefficient compares the cooccurrence frequency of the word pair against the summed frequencies of the words in the pair. As proposed in [4], $\Delta P_{1|2}$ and $\Delta P_{2|1}$ are different from the other measures in that they are “directional” while others are “symmetric”. That is, instead of assigning a single score that indicates the strength of “mutual” attraction between a word pair (w_1, w_2), $\Delta P_{1|2}$ and $\Delta P_{2|1}$ assign two separate scores to a single word pair— $\Delta P_{1|2}$ indicates how predictable w_1 is given w_2 , and $\Delta P_{2|1}$ indicates how predictable w_2 is given w_1 .

$$\begin{array}{lll}
 \text{MI} = \log_2(O/E) & \text{MI}^3 = \log_2(O^3/E) & \text{MI.log-f} = \text{MI} \cdot \ln(O + 1) \\
 t = (O - E)/O^{0.5} & \text{Dice} = 2 \cdot f_{AB} / (f_A + f_B) & \text{logDice} = 14 + \log_2(\text{Dice}) \\
 \Delta P_{2|1} = p(w_2 | w_1) - p(w_2 | \text{not } w_1) & & \Delta P_{1|2} = p(w_1 | w_2) - p(w_1 | \text{not } w_2)
 \end{array}$$

Figure 1. Formula of the association measures used in this study. f_A is the frequency of a word

¹ <https://www.ptt.cc/bbs>

² <https://www.ptt.cc/bbs/{category}>

³ A modified version of Jieba trained with Sinica Corpus. <https://github.com/amigacamel/Jseg>

w_A in the corpus; O (or f_{AB}) is the observed frequency of a word pair (w_A, w_B) ; E (equals $f_A \cdot f_B / N$, where N is the corpus size) is the expected frequency of a word pair (w_A, w_B) ; $p(w_A | w_B)$ is the probability that w_A occurs before w_B , and $p(w_A | \text{not } w_B)$ is the probability that w_A occurs before words other than w_B .

Due to the limitation of computing power, only word pairs with one of the words being a single-character word occurring in the Chinese Lexical Database [5] were calculated for the association measures. In addition, word pairs with frequencies below or equal to 3 were excluded from the calculation. This resulted in a dataset of 334,686 word pairs with their corresponding eight association measures.

2.2 Model

As a preliminary investigation, the model used in this study was intended to be simple and transparent. The model M_{comb} is a simple linear combination of several association measures, as shown in equation (1).

$$M_{\text{comb}} = \alpha_1 \cdot \mathbf{M}_1 + \alpha_2 \cdot \mathbf{M}_2 + \alpha_3 \cdot \mathbf{M}_3 + \dots + \alpha_n \cdot \mathbf{M}_n \quad (1)$$

In equation (1), M_i is the percentile rank of one of the eight association measures mentioned in the previous section, and α_i is the weight of M_i on the model M_{comb} . The weights α_i are determined by a grid search [9] that finds the best configuration of $(\alpha_1, \alpha_2, \dots, \alpha_n)$.

The goal of the model is to retrieve a list of collocations that has a low portion of false collocations resulting from word segmentation errors. To achieve this, we score the model during the grid search as *the portion of “correct” collocations in a list of top n collocations ranked according to M_{comb}* . “Correct” collocations are defined as collocations that (1) do not result from word segmentation errors, and (2) have ranks below 100 in at least m association measures (the parameter “low rank num” in Table 2). Word segmentation errors are defined using a dictionary constructed from tokens in ASBC [6], lexical entries in the Chinese dictionary compiled by the Ministry of Education⁴, and Chinese Wikipedia page titles⁵. A word pair is defined as a false collocation if it results from a single lexical entry in the dictionary that is split apart due to a word segmentation error. Note that this definition is limited in that only a certain kind of word segmentation errors (e.g. “蔡 | 英文”) is captured.

⁴ <https://github.com/g0v/moedict-data>

⁵ <https://dumps.wikimedia.org/zhwiki/20200620>

Other kinds of segmentation errors such as “軍官軍 | 銜”, which segments a string of two words (“軍官” and “軍銜”) into two words in a wrong way (“軍官軍” and “銜”), would not be captured by this dictionary checking approach. This restricted definition is a compromise since a more precise definition would require costly human annotation of the word pairs.

2.3 Training

The dataset described in section 2.1 was split into 80% for training and 20% for testing. The training set was used to perform the grid search to find the optimal weight configurations for the component association measures. For each iteration (a set of weights α_i), an M_{comb} score can be calculated for each word pair in the training set. The top n collocations were then retrieved according to the M_{comb} scores, from which a score (the proportion of “correct” collocations) could then be assigned to this weight configuration. After the grid search, weight configurations with the highest score were then used to retrieve the top n collocations from the testing set, from which the model was evaluated.

3 Evaluation

To see whether combining several association measures in a linear fashion could improve the results of collocation extraction, we evaluated the top n collocations retrieved by the model against the top n collocations retrieved by each association measure constituting the model. Ninety percent of the testing data were sampled and used for the retrieval of the top n collocations. For each set of top n collocations retrieved by the model and its component association measures, the proportion of “correct” word segmentation was calculated. This process was repeated 100 times, and the distribution of the proportion of the “correct” collocations for the model and its component association measures were compared.

Several configurations of the model were tested, most of which show qualitatively similar results. In the following sections, we describe two versions of the model—one consisting of three component association measures and the other consisting of eight. Table 2 summarizes the models and their performance.

3.1 Model 1: Linear Combination of Three Measures

The first model is a linear combination of three association measures—MI, logDice, and $\Delta P_{1|2}$:

$$M_{\text{comb}} = \alpha_1 \cdot \text{Percentile}(\text{MI}) + \alpha_2 \cdot \text{Percentile}(\text{logDice}) + \alpha_3 \cdot \text{Percentile}(\Delta P_{1|2}) \quad (2)$$

3.1.1 Model Parameters

During training, the weight configurations $(\alpha_1, \alpha_2, \alpha_3)$ were searched over the space:

$$\{(i, j, k) \mid \forall i, j, k \in S\}, \text{ where } S = \{-1, -0.95, -0.9, \dots, 0.9, 0.95, 1\}$$

For each weight configuration, 20 collocations with the highest M_{comb} scores were retrieved. The score of a weight configuration is the proportion of “correct” word segmentation in this list of top 20 collocations.

3.1.2 Comparing with Single Association Measures

Training with these parameters resulted in eight weight configurations that reached an optimal score of 0.7 in the training set. For each of them, the distribution of the proportion of the “correct” collocations in the testing set is shown in Figure 2. The optimal M_{comb} model on the testing set has a mean of 46.7% “correct” collocations.

Retrieving the top 20 collocations with the component measures of the M_{comb} model, on the other hand, yields better results—two of the three measures performed better than 46.7%, and even the least performant measure (MI) has an average score of 45.9% (Figure 3).

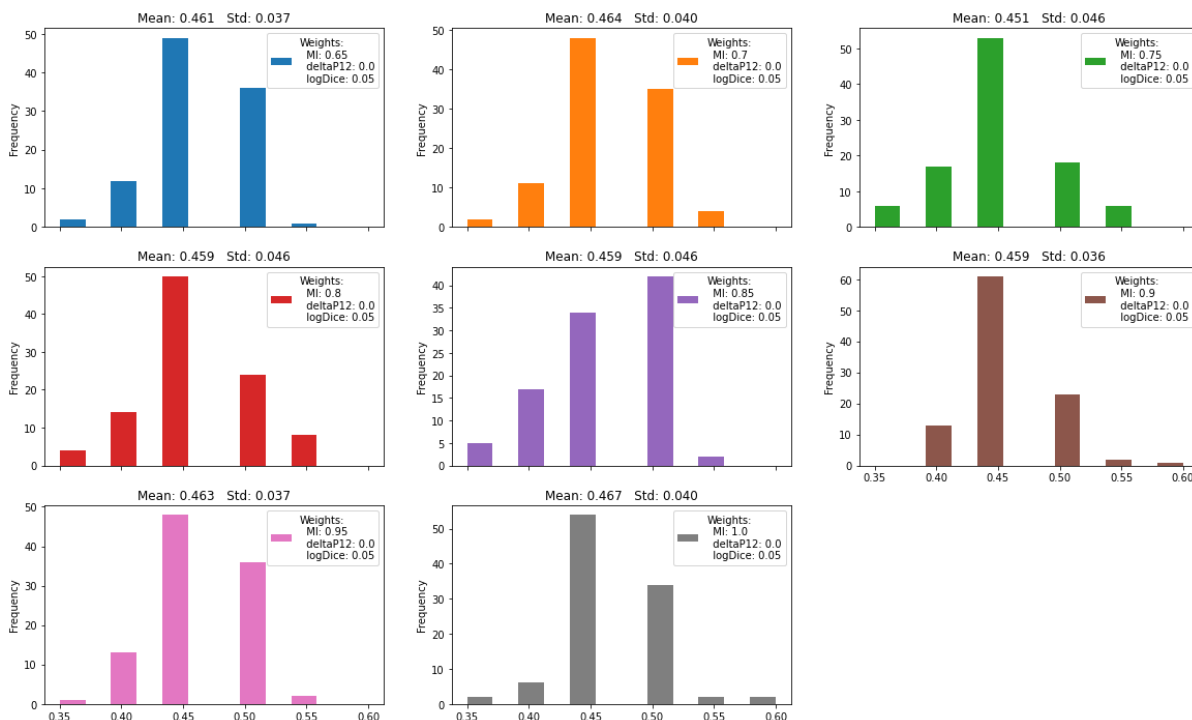


Figure 2. The distribution of the proportion of the “correct” collocations retrieved by each of the eight optimal M_{comb} scores in the testing set. Among these eight optimal weight configurations (on training set), the configuration $1.0 \cdot \text{Percentile}(\text{MI}) +$

$0.05 \cdot \text{Percentile}(\log\text{Dice}) + 0.0 \cdot \text{Percentile}(\Delta P_{1/2})$ (the subplot in the 3rd row and the 2nd column) achieved the best performance (46.7% “correct”) on the testing set.

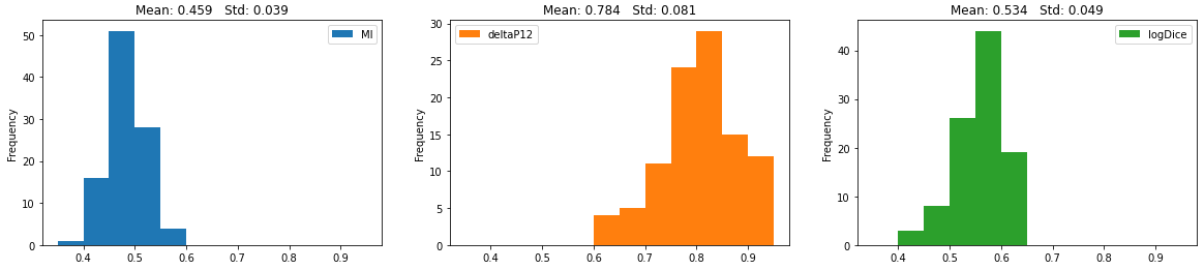


Figure 3. The distribution of the proportion of the “correct” collocations retrieved by each of the component association measures of the M_{comb} model—MI, $\Delta P_{1/2}$, and logDice. The component measures, at least for $\Delta P_{1/2}$ (the center subplot, 78.4% “correct”) and logDice (the rightmost subplot, 53.4% “correct”), performed better individually than combining together into M_{comb} on the testing set.

3.2 Model 2: Linear Combination of Eight Measures

The setup of Model 2 is identical to Model 1 except that there are now 8 components in the model:

$$\begin{aligned}
 M_{\text{comb}} = & \alpha_1 \cdot \text{Percentile}(\text{MI}) + \alpha_2 \cdot \text{Percentile}(\log\text{Dice}) + \alpha_3 \cdot \text{Percentile}(\Delta P_{1/2}) + \\
 & \alpha_4 \cdot \text{Percentile}(\Delta P_{2/1}) + \alpha_5 \cdot \text{Percentile}(\text{MI}^3) + \alpha_6 \cdot \text{Percentile}(\text{MI} \cdot \log\text{-f}) + \\
 & \alpha_7 \cdot \text{Percentile}(t) + \alpha_8 \cdot \text{Percentile}(\text{Dice})
 \end{aligned} \quad (3)$$

Due to the huge search space resulting from the eight weight configurations, instead of a full grid search, 1/10,000 of the search space was sampled and searched on. In addition, the space of the possible values for α_i was set smaller to $S = \{-1, -0.875, -0.75, \dots, 0.75, 0.875, 1\}$.

Evaluated using the procedure identical to Model 1, Model 2 showed no qualitatively different results. Several configurations of the parameters of Model 2 all resulted in models that do not perform better than their component association measures, again, showing that combining several association measures in a linear fashion cannot protect the model from retrieving false collocations resulting from word segmentation errors. Table 2 summarises several parameter settings of Model 1 and Model 2 and the results of the evaluation.

Table 2. Parameter settings and performance of the models in the experiment. For most

models (except 1-3), the performance is worse than at least one of their component measures.

Model ID	Component Measures	Parameters			Training	Testing (max correct %)	
		Top n	Low rank number	Search space	Number of optimal weight configs	Model	Component measures
1-1	MI, $\Delta P_{2 1}$, logDice	20	2	{1, -0.95, ..., 0.95, 1}	101	0.557	0.568
1-2	MI, $\Delta P_{1 2}$, logDice	20	2	{1, -0.95, ..., 0.95, 1}	8	0.462	0.794
1-3	MI, t, logDice	20	2	{1, -0.95, ..., 0.95, 1}	261	0.632	0.555
1-4	MI, MI3, t	20	2	{1, -0.95, ..., 0.95, 1}	436	0.504	0.554
2-1	All	10	1	{-1, -0.875, ..., 0.875, 1}	10	0.613	0.796
2-2	All	20	3	{-1, -0.875, ..., 0.875, 1}	3	0.503	0.793
2-3	All	20	2	{-1, -0.875, ..., 0.875, 1}	1	0.552	0.787
2-4	All	10	2	{-1, -0.875, ..., 0.875, 1}	13	0.632	0.797

4 Discussion

As seen in Table 2, generally, the model performs worse than its component measures. The only exception is Model 1-3, which by combining MI, t, and logDice, attained better performance than its component measures in 3 of 261 weight configurations. Hence, at least in the case of MI, t, and logDice, the combination of association measures may lead to better collocation extraction.

The general failure of the model suggests that if combining association measures could indeed capture patterns of word segmentation errors in collocation extraction, combining the measures in a linear fashion is too simple to capture these patterns. One direction for future research then is to use more complicated models, such as adding interaction terms to the model and see whether these more complicated models could capture word segmentation errors in the collocations. This approach, however, suffers from the exponential growth of the search space, making it computationally expensive or even impossible to find the optimal configurations.

Another direction of future work is to incorporate information additional to association measures into the model. Word vectors are promising candidates for this direction of work, as word segmentation errors might result in nonsense “words”, and these nonsense words might reveal themselves from the pattern of semantic similarities between normal and nonsense words, and between the words in each of these categories. To confirm our intuition, we carried out a pilot case study to inspect the semantic similarities among the words in two lists of collocations. The word 林 ‘Lin (family name)’ and 三 ‘three’ were used in the two lists

respectively as the node word, and their right collocates were extracted. For each of the two lists, collocations were extracted using seven measures (the eight measures except Dice mentioned in section 2.1)—20 collocations ranked as highest were retrieved for each measure, resulting in a list of 140 collocations (with duplications). Then, collocations that appeared less than 3 times were removed from the list (i.e., a collocation needs a rank of at least 20 in at least 3 measures to retain in the list). We then calculated the semantic similarities (cosine similarity of word vectors) between all words with FastText pre-trained word vectors [7]. The results are represented as network plots shown in Figure 4 and 5. The node in the network represents a word (either a node word or its collocates) in a list of collocations. The thickness of the edge between a pair of words indicates the degree of similarity between them, with higher similarity represented by a thicker edge.

One feature that instantly pops out from the figures is that correctly segmented collocates (blue nodes) form clusters. That is, these collocates are similar to each other in terms of semantic similarities as measured by the cosine similarity of their word vectors. On the other hand, collocates resulting from word segmentation errors are much more spread out throughout the network. This contrast between correctly and erroneously segmented collocates makes sense since word vectors are known to capture the extent to which words are replaceable (i.e., second-order, or paradigmatic, similarity) [8]. Thus, the collocates appearing within the same paradigm, such as 林{同學/老師/醫生} or 三{次/位/名/天/年/秒/小時}, are expected to have high pairwise similarities. Word segmentation errors, on the other hand, distort the well-formedness of the words, which may result in noisy patterns in similarities between these anomalous words, and the patterns are likely to vary case to case for collocations retrieved with different node words.

This simple ad hoc study, which shows that erroneously word segmented collocates may have a different profile to correctly segmented collocates, thus hints at a potential direction for future research by incorporating word vector information to improve the quality of collocation extraction. In addition, this pattern of similarities between collocates, which is observed in collocations that are defined with word pairs occurring in a window size of two (bigrams), is expected to generalize to collocations defined with word pairs occurring in larger window sizes. This is because the pattern observed seems to result from the well-formedness of the collocates. As long as word segmentation errors produce nonsense collocates, this approach is likely to capture the pattern of semantic similarities between true and false collocates.

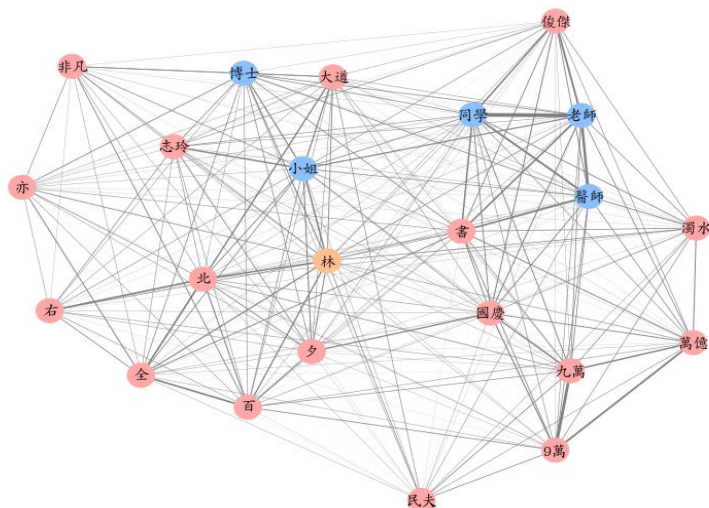


Figure 4. Semantic similarities between right collocates of 林 ‘Lin (family name)’. The thickness of the edge indicates the degree of similarity between a pair of nodes (edges with higher values of similarity are thicker). The orange node indicates the node word. Blue nodes indicate words of collocations that are correctly segmented, and red nodes are words resulting from word segmentation errors.

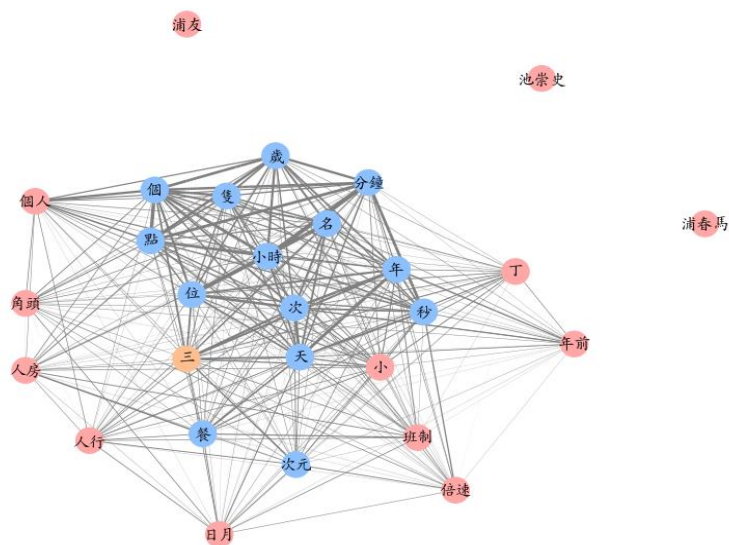


Figure 5. Semantic similarities between right collocates of 三 ‘three’. See the caption of Figure 4 for the meaning of the edges and the node colors. The nodes 浦友, 池崇史, 浦春馬 are isolated (i.e., similarities with other words cannot be measured) because word vectors of these words cannot be constructed due to the absence of necessary subword information in FastText pre-trained model.

5 Conclusion

Word segmentation is an important step in the NLP pipeline for Chinese, as the result of word segmentation largely influences the downstream tasks in the pipeline, such as PoS tagging, NER, and collocation extraction. In addition, with the prevalence of the internet, large corpora are constructed from texts collected from the web. With automatic word segmentation and PoS tagging performed on such large corpora, it is nearly impossible for manual checking on the correctness of such results. Thus, it is crucial to explore ways to mitigate the impacts of erroneous results stemming from such automatic tasks on downstream tasks.

In this paper, we investigated methods for improving the results of collocation extraction in automatically word segmented Chinese corpora, which suffers from retrieving false collocations resulting from word segmentation errors. A simple model, which combines several association measures in a linear fashion, are explored. Experiments with the simple linear model show that this model could not capture the necessary patterns to distinguish correctly word segmented collocations from erroneously segmented ones, as in most cases, the model performed worse than the association measures constituting the model. Facing this null result, we explore the potential for word vectors to capture the patterns of word segmentation errors in a list of collocations. An ad hoc case study of two lists of collocations shows that, in a list of collocations retrieved with a node word, correctly word segmented collocates are much more similar to each other in terms of semantic similarities computed from word vectors compared to erroneously segmented collocates. In future work, a study that investigates formal methods of incorporating word vector information to mitigate impacts of word segmentation errors on collocation extraction is suggested.

References

- [1] S. Evert, “Corpora and collocations,” in *Corpus linguistics. An international handbook*, vol. 2, A. Lüdeling and M. Kytö, Eds. Berlin: Walter de Gruyter, 2008, pp. 1212–1248.
- [2] S.-K. Hsieh, “Why chinese web-as-corpus is wacky? Or: How big data is killing chinese corpus linguistics,” in *Proceedings of the ninth international conference on language resources and evaluation (LREC’14)*, 2014, pp. 2386–2389.
- [3] A. Kilgarriff *et al.*, “The Sketch Engine: ten years on,” *Lexicography*, vol. 1, no. 1, pp. 7–36, Jul. 2014.
- [4] S. Th. Gries, “50-something years of work on collocations: What is or should be next ...,” *International Journal of Corpus Linguistics*, vol. 18, no. 1. John Benjamins, pp.

137–166, 2013.

- [5] C. C. Sun, P. Hendrix, J. Ma, and R. H. Baayen, “Chinese lexical database (CLD),” *Behavior Research Methods*, vol. 50, no. 6, pp. 2606–2629, Dec. 2018.
- [6] C. Huang and K.-J. Chen, *Academia Sinica Balanced Corpus*. 1998.
- [7] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, “Learning word vectors for 157 languages,” 2018.
- [8] O. Levy, Y. Goldberg, and I. Dagan, “Improving distributional similarity with lessons learned from word embeddings,” *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015.
- [9] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, no. null, pp. 281–305, Feb. 2012.

應用詞向量模型於日文單詞可讀性評估之研究

Japanese Word Readability Assessment using Word Embeddings

楊正仁¹, 林淑璋², 薛芸如², 鄭靖潔¹, 王縉蓁¹, 黃依賢¹

Cheng-Zen Yang¹, Shu-Chang Lin², Yun-Ju Hsueh², Ching-Chieh Cheng¹, Yih-Chen Wang¹,
I-Shyan Hwang¹

¹元智大學資訊工程學系

²元智大學應用外語學系

¹Department of Computer Science & Engineering

²Department of Foreign Languages and Applied Linguistics

Yuan Ze University

Email: {czyang,shuchang,connie,ishwang}@saturn.yzu.edu.tw

{s1051530, s1051531}@mail.yzu.edu.tw

摘要

在文本分析上，文本可讀性多年來都是一項重要的研究議題。然而過往研究大多偏向討論文章的可讀性，鮮少有研究討論單詞的可讀性。在單詞可讀性自動化評估問題上，最大的挑戰來自於單詞沒有豐富的可讀性特徵，不像一般文本有語句長度、音節多寡等特徵來辨別文本可讀性，因此難以利用機器學習技術來協助人工評估。然而由於詞向量技術的進步，在本論文中，我們提出一個新的單詞可讀性評估方式，使用詞向量技術將單詞取出對應的詞向量特徵，然後利用詞向量的語意相似關係，以最近鄰居法建構一個單詞可讀性評估分類模型WR-kNN。由於單詞可讀性語料的貧瘠，本研究以目前可獲得的日文單詞可讀性來進行探討。實驗結果顯示，WR-kNN的單詞可讀性分類準確率最多可以達到42.50%，比基於單詞上下文周邊詞的計算方法可以得到2.51倍的改善。

Abstract

In text analysis, text readability has been an important research topic for many years. However, most studies focus on the document readability rather than the word readability. To decide the readability levels of words, linguists need to spend a large amount of human effort and assessment time. The most challenging problem faced in the task of automatic word readability assessment is the lack of research on readability for each word. In this study, we propose a novel assessment model for word readability called WR-kNN based on the word embedding technology by calculating the word vectors

and classifying them with a k-nearest neighbor model. Since the resource of word readability corpus is rare, this study uses a Japanese corpus to discuss. The experimental results show that WR-kNN can predict the word readability with 42.50% accuracy. Compared with a co-occurrence-based approach, WR-kNN can achieve an improvement by a factor of 2.51.

關鍵字：單詞可讀性，詞向量，自然語言處理，效能評估

Keyword: Word Readability, Word Embedding, Natural Language Processing, performance evaluation

一、緒論

在文本分析上，可讀性（Readability）一直是許多研究所探討的對象[1,2]。Geroge Klare [1] 指出，可讀性的重要是因為它能夠用來 (1)表示文件手寫或排版的易讀性（Legibility）；(2)表示由於閱讀興趣或寫作生動而易於閱讀的程度；(3)表示由於寫作風格而易於理解或理解的程度。可讀性高的文章就容易被讀者理解，因此文本可讀性的評估對於出版或教育等皆具有重要的意義。自 20 世紀初開始，就有許多研究對文本可讀性進行探討，並且紛紛提出不同的評估模型，用來分析文章的可讀性，例如 [3,4,5,6,14,15,16,17,18,19]。

在這些文本可讀性的研究中，考慮了許多文本以及語言特性，例如語句長度、英文音節（Syllable）個數、難詞的詞數，中文不同筆劃的詞數、日文和語詞數、外來語等特徵。雖然有些研究會對單詞的難易度加以考慮，然而過往對於單詞可讀性的研究卻相當缺乏。一些文本可讀性研究曾考慮單詞可讀性的難易度，但是它們只有單純考慮單詞在語料庫當中出現的頻率(The frequency of words)或是考慮音節個數等單字特徵。例如在詞頻上，Lively與Pressey是以Thorndike編著的常用詞彙表來進行研究[3]，劉憶年等人也考慮文章中難詞（不在常用詞表的詞）出現次數[16]。以上這些研究都沒有再深入探討單詞可讀性。從這些研究中也顯示單詞可讀性對於文本可讀性的判斷會有影響。如果單詞可讀性語料具有一定規模，這些資訊將可以有效被運用來進行文本可讀性的評估。

雖然單詞可讀性對於文本可讀性評估有其重要性，但目前缺乏相關研究，所能獲得單詞可讀性語料規模也有限。就目前所搜尋的相關文獻所及，我們只看到Sunakawa等人發表的「日本語學習詞書」(Japanese Language Learners' Dictionary)[7]曾標註日文單詞的

テキスト詳細

結果保存(CSV:Shift-JIS) 結果保存(CSV:UTF-8)

総文数:4 文の平均語数:43.25

色の付いた語をクリックすると辞書引きを行います。

内容語レベル: 初級前半 初級後半 中級前半 中級後半 上級前半 上級後半

機能語レベル: 初級 初中級 中級

- 1 岐阜県は15日、県農業大 学校(可児市)で飼育している実習用のメスの親豚から豚(とん)コレラの陽性反応が出たと発表した。
- 2 岐阜県内の飼育施設での豚コレラ確認は5例目、県立施設での確認は今年5日の県畜産研究所(美濃加茂市)に次いで2例目。
- 3 県によると、感染した豚から検出されたウイルスを農業・食品産業技術総合研究機構で精密検査したところ、これまでに県内で感染が確認されたウイルスと同一のものと確認された。
- 4 畜産研究所の感染が判明したため7日に検査したが、この時は同大 学校で飼育していた13頭全てが...

圖 1、日本毎日新聞社有關豬瘟的新聞文章，經由 Jreadability 分析。

可讀性。他們將日文單詞可讀性分成6個級別：初級前半，初級後半，中級前半，中級後半，上級前半，上級後半。Jreadability (<http://jreadability.net/>) [13]日文文本可讀性分析網站便使用這些語料來標註日文單詞的可讀性。

雖然「日本語學習詞書」提供了可讀性語料，但是其中的單詞數量仍然有限。如圖1所顯示，一篇有關豬瘟的新聞「豚コレラ、岐阜で5例目確認 県農業大 学校」(<https://mainichi.jp/articles/20181216/k00/00m/040/012000c>)經過Jreadability分析，可以發現許多字詞依然沒有可讀性級別。在圖1的例子中，總共173個單詞的84個相異詞型(Word Type)當中，有26個詞型沒有可讀性級別。我們在分析其他文章的時候，發現甚至可以高達35%以上的單詞沒有可讀性級別。

在單詞可讀性自動化評估問題上，最大的挑戰來自於過往缺乏單詞可讀性特徵的研究，以至於缺乏可讀性特徵。不像一般文本有語句長度、音節多寡等特徵來辨別文本可讀性，因此過往是以人工研讀文本的方式來決定單詞可讀性，沒有機器學習的人工智慧方式來判斷。由於單詞數量龐大，人工研讀文本會耗費大量人力與時間。更因為目前網路時代造成資訊爆炸，每年都有不少新的詞彙產生。這些數量可觀的新詞彙將使單詞可

讀性評估更具挑戰。

最近在單詞特徵技術上，Mikolov等人所提出的詞向量（亦稱之為詞嵌入，Word Embedding）技術[8]，為單詞的特徵表示方式開啟新的里程碑。由於詞向量技術的發展，在本研究中我們提出一個新的單詞可讀性評估方式，使用詞向量技術從所蒐集的語料當中計算出單詞的詞向量特徵。利用詞向量特徵表現出單詞在文本當中的語意特性，在語意空間中找出語意關係相似的詞向量，以最近鄰居法(k-Nearest Neighbor, kNN) [9]建構一個單詞可讀性評估分類模型稱之為WR-kNN (Word Readability kNN)，從這些相似的詞向量的可讀性等級來預測單詞的可讀性。

由於過往缺乏單詞可讀性的研究，本研究的主要貢獻在於首先針對單詞可讀性，以詞向量技術來建構機器學習模型，進行單詞可讀性預測。這個機器學習模型將可以協助語言學專家判讀單詞的可讀性級別，同時可以進而運用在文本可讀性判讀的相關技術中，提昇文本可讀性的評估。

然而由於單詞可讀性語料稀少，因此本研究以Jreadability與「日本語學習詞書」標註的日文單詞可讀性來探討。我們蒐集了福娘童話集(<http://hukumusume.com/douwa/>)、青空文庫(<https://www.aozora.gr.jp/>)、及IEICE日文期刊論文摘要(<https://www.ieice.org/>)進行詞向量計算。實驗結果顯示，WR-kNN的單詞可讀性分類準確率最多可以達到42.50%，比基於單詞上下文周邊詞的計算方法可以得到2.51倍的改善。

本論文其餘內容安排如下：第二節將介紹最近的可讀性相關研究。第三節將說明基於詞向量的最近鄰居法單詞可讀性評估模型WR-kNN的設計。第四節將說明實驗環境設定、蒐集的文本語料、以及實驗的結果。第五節是本論文的結論。

二、 相關研究

目前關於可讀性的研究，幾乎都集中在文本可讀性的探討，例如在經驗法則的設計上，1921年時Kitson就使用單詞的音節數與句子長度來討論英文新聞文章的可讀性[1]。Lively與Pressey在1923年對英文文章提出一個可讀性評估公式來計算詞彙負擔

(Vocabulary Burden) [3]。1948年Flesch 發表著名的Flesch Reading Ease公式[4]。在中文文章的可讀性評估中，楊孝滌[14]與陳世敏[15]分別提出不同的迴歸方程式來計算。在日文文章的可讀性評估研究中，Tateisi等人在1988年使用Principal Component Analysis (PCA)分析文章的10種特性，提出一個可讀性評估公式[5]。2008年時Sato等人以教科書為語料庫，提出一個機率式語言模型來評估文章的可讀性[6]。然而這些研究並沒有涉及單詞可讀性的討論。

隨著機器學習技術以及詞向量技術的發展，劉憶年等人對於中小學國語文教材，利用詞向量技術結合逐步迴歸以及支援向量機，建構中文文本可讀性分析模型[16]。曾厚強等人並提出快速文本(fastText)與卷積神經網路(Convolutional Neural Network, CNN)來建構中文文章的可讀性模型[17,18]。曾厚強等人並於2019年提出以BERT (Bidirectional Encoder Representation from Transformers)技術[10]，利用WECA[11]來進行中文文本可讀性評估的研究[12]。即便如此，單詞可讀性也沒有機器學習模型的研發。

在單詞可讀性的研究上，Sunakawa等人在「日本語學習詞書」(Japanese Language Learners' Dictionary)[7]曾對日文單詞進行探討。他們邀請五位日語專業教師以人工方式

表 1、「日本語學習詞書」的單詞可讀性級別屬性範例。單詞的可讀性級別（語彙難易度）分為 6 種等級。取自[7]。

語彙 ID	標準的表記	読み	語彙難易度	品詞	語種	旧試験語彙級	意味分類	アクセント情報
10	アート	アート	中級前半	名詞-普通 名詞-一般	外来語		体-活動-芸術・美術	1
40	アイスコーヒー	アイスコーヒー	初級前半	名詞-普通 名詞-一般 and 名詞-普通 名詞-一般	外来語		体-生産物-食料-飲料・たばこ	6
109	明かり	アカリ	中級前半	名詞-普通 名詞-一般	和語	2級	体-生産物-機械-灯火 体-自然-自然-光	0
222	足掛かり	アシガカリ	上級後半	名詞-普通 名詞-一般	和語		体-関係-空間-点	3
294	厚かましい	アツカマシイ	中級後半	形容詞-一般	和語	2級	相-活動-心-自信・誇り・恥・反省	5
262	温まる	アタタマル	中級前半	動詞-一般	和語	2級	用-自然-物質-熱	4

檢查18010個單詞，他們邀請五位日語專業教師以人工方式檢查18010個單詞，將它們分成6個可讀性級別：初級前半，初級後半，中級前半，中級後半，上級前半，上級後半，如表1所示。基於「日本語學習詞書」的開發，這些資料後來彙整為日本語教育語彙表 (<http://jisho.jpn.org/>)，運用在由Hasebe等人所研發的日語可讀性評估工具Jreadability [13]文本可讀性分析網站之中。就目前文獻回顧所及，目前還沒有國內外研究針對單詞可讀性的機器學習模型進行討論。因此在本研究中，我們將探討單詞可讀性評估分類模型的建構。

三、 研究方法

在本論文中我們提出一個基於詞向量技術與kNN的單詞可讀性分類模型WR-kNN。這一節將說明WR-kNN的設計。首先將說明WR-kNN的處理流程以及各個處理部份的方法。接下來說明實驗中用來作為效能比較baseline的基於單詞上下文分類的作法。

(一) 單詞可讀性分類模型

針對一個單詞 t_i ，以及一個具有 m 個可讀性級別的分類集合 $L = \{l_1, l_2, \dots, l_m\}$ ，本研究定義單詞可讀性評估為一個分類問題，要找到 t_i 的可讀性級別 l_i 如下：

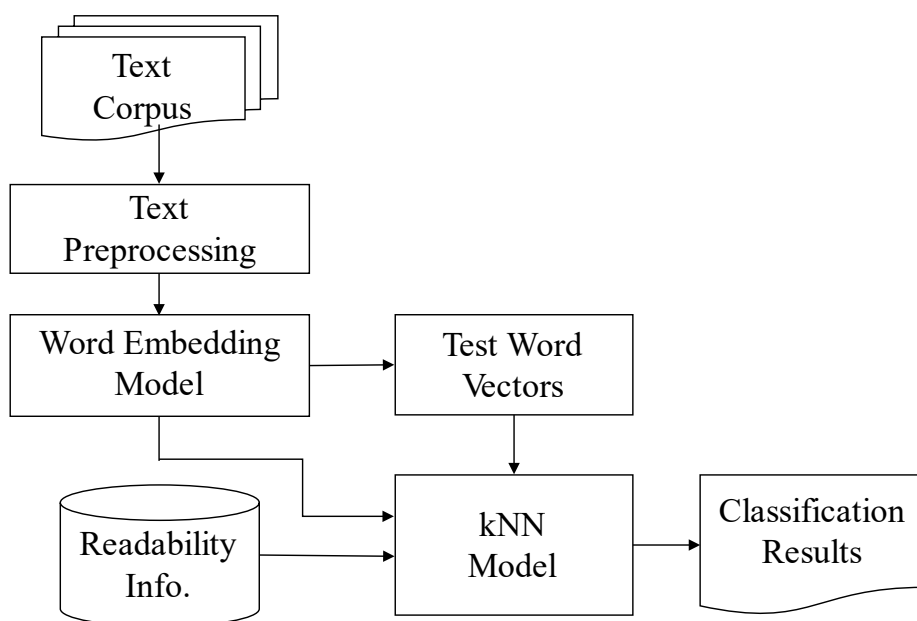


圖2、單詞可讀性分類模型。

$$l_i = \underset{L}{\operatorname{argmax}} P(l_x | t_i)$$

其中 $P(l_x | t_i)$ 是 t_i 的可讀性級別為 l_x 的機率。

圖2是本研究所建構的單詞分類模型架構。以下將說明文本前處理方式、詞向量生成方式、以及kNN分類器處理方式。

(二) 文本前處理

對於日文文本，首先將文本中的一些標點符號去除，並將一些文本中的註解消去，以免影響文本內容的品質。接著進行斷詞，我們使用 MeCab (<http://taku910.github.io/mecab/>)與Unidic字典對日語文本進行斷詞以及相關詞性分析。我們取得MeCab斷詞結果以及相關單詞資訊，包括：表層形、詞性、詞性細分類1、讀音等資訊。圖3是一個文本前處理的範例，顯示句子「居留地女の間では」經過斷詞後的結果。

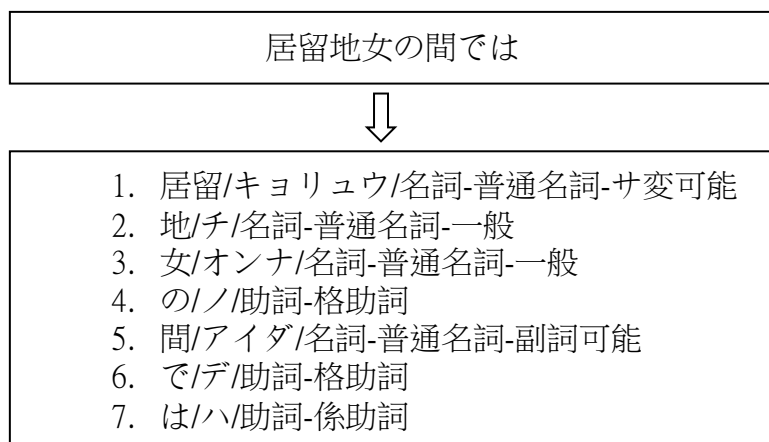


圖3、日文斷詞範例。

(三) 詞向量計算

透過文本中的上下文關係，單詞可以經由N-gram語言模型(Language Model)計算出單詞的詞向量。在本研究中，我們使用Mikolov等人提出的Word2vec模型[8]來計算出單詞的詞向量。

Word2vec詞向量模型有兩種不同的架構，分別是Continuous Bag-of-Words model (CBOW)與Skip-Gram model (SG)。在CBOW模型中，對於目標字詞 $w(t)$ 的詞向量，會將

表2、和「居留」語意最相似的5個單詞。

單詞	相似度	可讀性級別
浴室/ヨクシツ/名詞-普通名詞-一般	0.7632547616958618	中級後半
いかり/イカリ/名詞-普通名詞-一般	0.7366321682929993	上級後半
桜/サクラ/名詞-普通名詞-一般	0.7297971248626709	初級後半
停車/テイシャ/名詞-普通名詞-サ変可能	0.7289541363716125	中級後半
倚つ/ヨル/動詞-一般	0.7197038531303406	中級後半

上下文其他字詞的詞向量經過淺層神經網路加以計算，得出 $w(t)$ 的詞向量。在SG模式中，則剛好相反，利用 $w(t)$ 詞向量，來計算 $w(t)$ 上下文字詞的機率分佈而得出詞向量。經由詞向量計算，每個詞都映射到 D 維度的語意空間。在這個語意空間中，呈現了單詞的語意特性。

在本研究中，單詞都經過Word2vec的計算得出它的詞向量。利用詞向量的特性，可以找出語意相似的單詞。表2是一個利用詞向量來找出與「居留」語意相似單詞的例子。在這個例子中，我們透過Cosine similarity的計算，從目前實驗的文本中找出與語意「居留」最相似的五個單詞。表中也列出這些單詞的可讀性級別。

(四) kNN分類

在文本的詞向量空間當中，單詞之間的語意相似度(Semantic Similarity)會嵌入在詞向量資訊中。對於語意相似度高的單詞，其可讀性級別可能大部分相近。因此基於這個

Algorithm 1 kNN classification

Input: t_i : target word; W : vocabulary with readability ; k : # of nearest neighbor;

Output: l_i : readability level of t_i ;

1: **for** $(t_j, l_j) \in W$ **do**

2: calculate the similarity $Sim(t_i, t_j)$;

3: **end for**

4: Sort W according to the similarity;

5: Count the number of occurrences of each class l_j among the k nearest neighbors;

6: Assign to t_i the level l_i which is the most frequent class;

圖4、kNN分類演算法。

構想，當Word2vec計算出單詞的詞向量後，我們利用kNN方式來進行單詞可讀性級別的分類。圖4是kNN演算法的演算法。

在本研究中，為了與kNN預測模型進行效能比較，我們基於單詞的共現關係（Co-occurrence），利用所有在 t_i 上下文跨距為 k 的視窗內的有可讀性級別周邊詞，以 Majority Voting的方式來決定 t_i 的可讀性級別。例如圖5中，如果 $k=5$ ，我們要預測「夫婦は貧乏で食べる物がないので、他の家が捨てたイモの尻尾ばかりを食べています。」中「家」的可讀性級別，就以「家」為目標詞，用前後 k 個有可讀性級別的周邊詞來進行投票。如果「家」在文本中總共出現 m 次，這些 $2km$ 個有可讀性級別的周邊詞都一起進行投票。

單詞	夫婦	は	貧乏	で	食べる	物	が	ない	の
級別	N/A	N/A	4	N/A	1	3	N/A	1	N/A
單詞	で	他	の	家	が	捨て	た	イモ	...
級別	N/A	3	N/A	目標詞	N/A	N/A	N/A	3	...

圖5、考慮共現關係來決定目標詞的可讀性級別。

四、實驗結果及分析

（一）實驗環境設定

在本研究中，我們所探討的日語的單詞可讀性級共分為6級：「初級前半」、「初級後半」、「中級前半」、「中級後半」、「上級前半」及「上級後半」。日文文章的可

表3、實驗的日文文章數量。

文本來源	數量
福娘童話集-日本童話	368
福娘童話集-世界童話	372
青空文庫	273
IEICE論文摘要	214
總計	1227

表4、日文文章可讀性級別的分佈。

文本來源	初級 前半	初級 後半	中級 前半	中級 後半	上級 前半	上級 後半
福娘童話集	23	504	208	5	0	0
青空文庫	1	14	128	74	47	9
IEICE論文摘要	0	0	2	25	96	91
總計	24	518	338	104	143	100

讀性級別也是分成相同的6級。為了進行效能評估，我們從福娘童話集、青空文庫以及IEICE的論文摘要分別蒐集日文文章，希望能涵蓋不同可讀性級別的文章。表3是本研究所蒐集的日文文章的數量，表4是文章可讀性級別分佈。

(二) 實驗結果

在本研究中，我們提出來基於詞向量的kNN模型稱之為WR-kNN。在實驗中，由於Word2vec的CBOW模型與Skip-gram模型，因此實驗分別量測WR-kNN-C (使用CBOW模型)與WR-kNN-SG (使用Skip-gram模型)的效能。由於詞向量在語意空間維度 D 的多寡會影響詞向量蘊含語意的豐富度，也會影響整體的效能表現，因此在實驗中，我們分別討論了50、100、200、300這4種維度，來觀察不同維度的影響。所進行比較的共現周邊詞的計算方式稱為kCO。在kCO中，我們考慮周邊詞所選的上下文跨距範圍是否受到語句邊界的限制。kCO-S表示受到語句邊界限制。kCO-D則表示不受到語句邊界限制，只受到文章本身邊界的限制。

在測試的時候，我們按照文本的可讀性級別，在各個級別中隨機選取100分文本，希望能保持各個級別文本數量的平衡。但是因為「初級前半」的文章數量很少，因此我們總共取出524篇文章。我們再從這524篇文章中按照單詞的可讀性級別，在各個級別中取

表5、測試資料集中的文章與單詞數量。

	初級 前半	初級 後半	中級 前半	中級 後半	上級 前半	上級 後半
文章數量	24	100	100	100	100	100
單詞數量	72	110	252	556	494	95

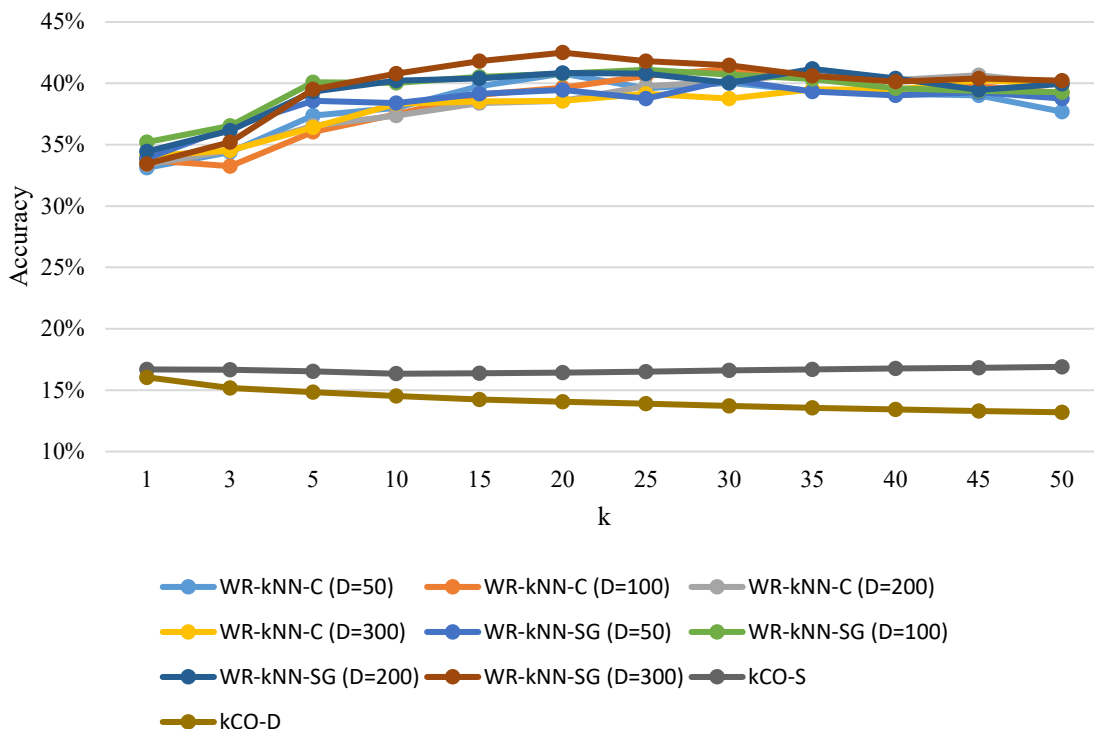


圖6、單詞可讀性準確率預測效能。

出前10%具有可讀性級別標註的高頻詞進行測試。從中總共取了1579個高頻單詞進行測試。表5是這些用來測試的文本與單詞在不同級別中的數量。

圖6是進行測試的準確率結果。從圖中可以看到WR-kNN的表現都比kCO為佳。WR-kNN的準確率表現在33.12%~42.50%，而kCO的準確率表現在13.20%~16.90%。其中WR-kNN-SG在詞向量維度 $D=300$ 、 $k=20$ 的時候有最佳準確率42.50%的表現。整體來說，使用Skip-gram模型的詞向量要比使用CBOW模型的詞向量有好的準確率表現。

我們另外計算了可讀性級別誤差一個級別也在可容許範圍內的準確率(± 1 Accuracy)。這是由於在一些實際情況中，即使預測的結果有一個級別的誤差，也還是能夠對專家的辨識有所幫助。圖7是 ± 1 準確率的效能表現。從圖中可以看到，kCO的表現隨著 k 值的增加而逐步下降，但是WR-kNN卻有逐步上升的趨勢。WR-kNN的 ± 1 準確率表現在78.21%~87.78%，而kCO的 ± 1 準確率表現在30.41%~40.82%。其中WR-kNN-C在詞向量維度 $D=200$ 、 $k=45$ 的時候有最佳準確率42.50%的表現。整體來說，使用CBOW模型的詞向量要比使用Skip-gram模型的詞向量有好的 ± 1 準確率表現。

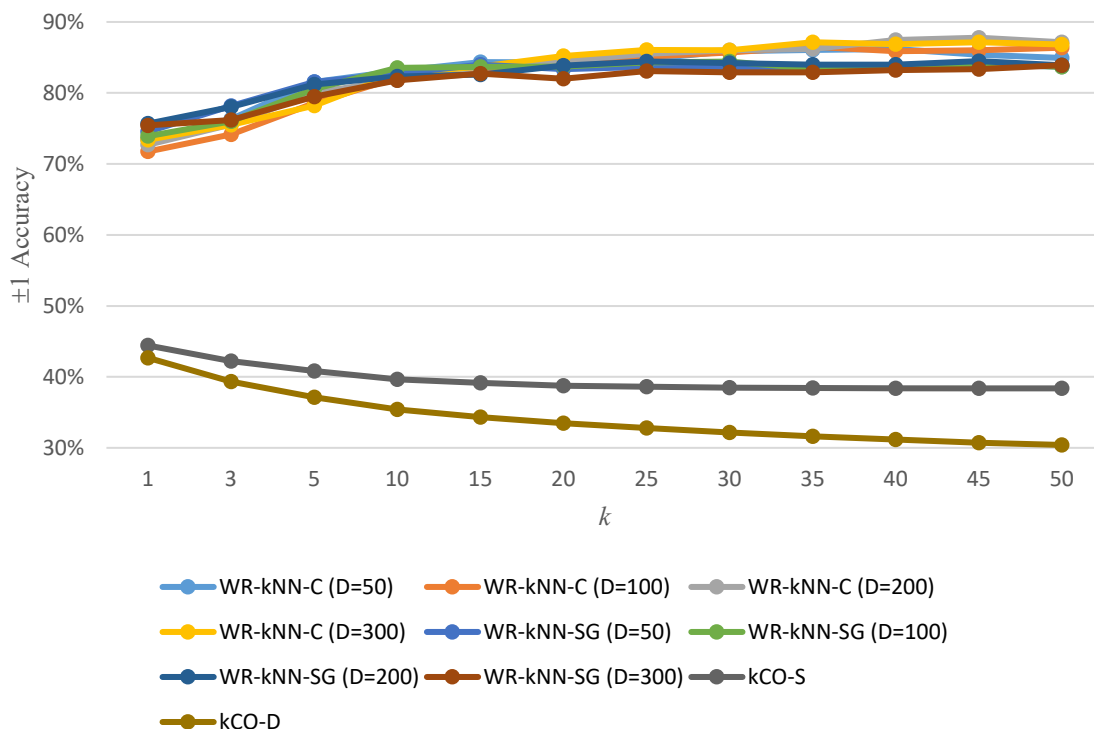


圖7、單詞可讀性±1準確率預測效能。

五、 結論

文本可讀性多年來都是一項重要的研究議題。然而過往研究大多偏向討論文章的可讀性，鮮少有研究討論單詞的可讀性。在本研究中，我們利用詞向量的特性，使用kNN分類器提出一個新的單詞可讀性級別評估的機器學習模型WR-kNN。經由與共現周邊詞的計算方式進行比較，實驗結果顯示，WR-kNN的單詞可讀性分類準確率最多可以達到42.50%，比基於單詞上下文周邊詞的計算方法可以得到2.51倍的改善。

在未來研究規劃中，我們預計將探討其他分類器的設計，希望能提昇評估的準確率效能。我們同時繼續蒐集更多文本來進行訓練與測試，希望能夠透過大量文本的訓練，能更準去表現出單詞之間的語意關係。

六、 致謝

本研究感謝科技部計畫編號 MOST 109-2221-E-155-028 部份支持。

七、 參考文獻

- [1] G. R. Klare, *The Measurement of Readability*: Iowa State University Press, 1963.
- [2] G. R. Klare, “The measurement of readability: useful information for communicators,” *ACM Journal of Computer Documentation*, vol. 24, no. 3, pp. 107-121, 2000.
- [3] 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.
- [4] R. F. Flesch, “A New Readability Yardstick,” *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221-233, 1948.
- [5] Y. Tateisi, Y. Ono, and H. Yamada, “A Computer Readability Formula of Japanese Texts for Machine,” in Proceedings of the 12th International Conference on Computational Linguistics (COLING '88), 1988, pp. 649-654.
- [6] S. Sato, S. Matsuyoshi, and Y. Kondoh, “Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus,” in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08), Marrakech, Morocco, 2008.
- [7] Y. Sunakawa, J. Lee, and M. Takahara, “The Construction of a Database to Support the Compilation of Japanese Learners’ Dictionaries,” *Acta Linguistica Asiatica*, vol. 2, no. 2, 10/23, 2012.
- [8] T. Mikolov, W.-T. Yih, and G. Zweig, “Linguistic Regularities in Continuous Space Word Representations,” in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), 2013, pp. 746-751.
- [9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer -Verlag, 2nd edition, 2009.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *CoRR*, vol. arXiv:1810.04805, 2018.
- [11] Y.-T. Sung, T.-H. Chang, W.-C. Lin, K.-S. Hsieh, and K.-E. Chang, “CRIE: An automated analyzer for Chinese texts,” *Behavior Research Methods*, vol. 48, no. 4, pp. 1238-1251, 2016/12/01, 2016.
- [12] H.-C. Tseng, H.-C. Chen, K.-E. Chang, Y.-T. Sung, and B. Chen, “An Innovative BERT-Based Readability Model,” in Proceedings of the 2nd International Conference on Innovative Technologies and Learning (ICITL 2019), pp. 301-308, 2019.
- [13] Y. Hasebe, and J.-H. Lee, “Introducing a Readability Evaluation System for Japanese Language Education,” in Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J 2015), pp. 19-22, 2015.

- [14] 楊孝滌, “中文可讀性公式,” *新聞學研究*, 8 卷, 77-101, 1971.
- [15] 陳世敏, “中文可讀性公式試擬,” *新聞學研究*, 8 卷, 181-226, 1971.
- [16] 劉憶年、陳冠宇、曾厚強、陳柏琳, “可讀性預測於中小學國語文教科書及優良課外讀物之研究,” in *Proceedings of the 27th Conference on Computational Linguistics and Speech Processing (ROCLING 2015)*, 2015, 71-86.
- [17] 曾厚強、宋曜廷、陳柏琳, “探究不同領域文件之可讀性分析,” in *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017)*, 2017, 116-118.
- [18] 曾厚強、陳柏琳、宋曜廷, “探索結合快速文本及卷積神經網路於可讀性模型之建立,” in *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018)*, 2018, 116-125.
- [19] 荊溪昱, “中文國文教材的適讀性研究：適讀年級值的推估,” *教育研究資訊*, 3 卷, 3 期, 113-127, 1995.

應用階層可解構式注意力模型於新聞立場辨識任務

A Hierarchical Decomposable Attention Model for News Stance

Detection

黃晨郁 Chen-Yu Hunag

國立中央大學資訊工程系

Department of Computer Science & Information Engineering

National Central University

lensixtwo@gmail.com

張嘉惠 Chia-Hui Chang

國立中央大學資訊工程系

Department of Computer Science & Information Engineering

National Central University

chia@csie.ncu.edu.tw

摘要

新聞立場辨識任務的目的為判斷一篇新聞對於某個議題的立場是中立、贊成或反對。此項任務與自然語言推理 (Natural Language Inference, NLI) 任務類似，目標在給定兩個句子，判斷兩者之間是否無關或存在蘊涵、矛盾關係。本論文以新聞立場檢索競賽提供的資料作為參考，但其大部分的新聞文章都屬於支持特定議題的立場，造成資料在不同類別的分佈不平衡。本篇論文提出 Hierarchical Decomposable Attention Model 來解決新聞立場辨識任務，我們以句子為單位分割新聞文章，並基於 Decomposable Attention 的原理找出文章中的每個句子與特定議題的關係。針對資料不平衡的問題，我們建立反義的議題，並手動標記新聞對反義議題的立場，以改善模型效能。實驗結果顯示，我們提出的模型效能優於其他模型。

Abstract

The goal of News Stance Detection task is to detect whether the stance of a news article is neutral, approval or opposition with respect to a given query. The task is similar to Natural Language Inference (NLI) task, which aims to determine if one given statement (a premise)

semantically entails another given statement (a hypothesis). Since most news articles hold neutral stances with respect to the given query, the training data is often unbalanced. In this paper, we proposed a Hierarchical Model based on the Decomposable Attention Model for NLI tasks to compare individual sentences with the given query and jointly predict the stance of the complete article. For the data imbalance problem, we heuristically create opposite queries and label supporting news articles from unrelated ones of the original query to identify unrelated news articles. The experiment result showed that the performance of our architecture is better than other models.

關鍵詞：新聞立場辨識，自然語言推理，篇章分析，注意力機制

Keywords: News Stance Detection, Natural Language Inference, Discourse Analysis, Attention Mechanism.

一、緒論

在新聞立場辨識 (News Stance Detection) 中，我們必須根據給定議題及新聞文章，去判斷此篇新聞立場為中立或是偏向贊成或是對立。[AI CUP 2019](#) 的新聞立場檢索競賽，其目的是開發一搜尋引擎，使用者能根據其輸入的議題，從大量新聞文章中找出與議題相關且立場一致的新聞文章，幫助閱聽人釐清新聞文章所代表的立場，從不同角度去了解各種爭議性的議題。競賽主辦單位提供國內新聞的網頁連結、包含立場的爭議性議題作為查詢題目以及標記好的訓練資料。表一為主辦單位提供的訓練資料範例，每筆資料包含議題、新聞文章以及兩者的相關程度。在評估的部分，主辦單位採用 MAP@300 (Mean Average Precision at 300) 指標來評估系統效能。MAP@300 的值介於 0 到 1 之間，值愈高表示搜尋結果愈好，其計算方式可參考[競賽網站](#)。

目前資訊檢索領域已發展出許多現成的工具，如：[Solr](#) 和 [Elasticsearch](#)，方便我們根據議題從大量文件中找尋相關的文件。然而，在相關的文件中，我們會發現內容與議題相關，但立場卻相反的新聞文章。如表一的 D1，該內容雖然與「陳前總統保外就醫」的議題相關，但包含了「反對陳前總統保外就醫」的論述，因此該篇文章與「支持陳前總統保外就醫」的議題存在著相反的立場，並且標記為 Irrelevant。

表一、新聞立場辨識資料範例

Query	支持陳前總統保外就醫	
D1	前總統陳水扁申請出席國慶大典昨遭到台中監獄駁回...中監應該考慮撤銷他保外就醫資格，讓他回去服刑。...	Irrelevant
D2	前總統陳水扁在獄中罹患重度憂鬱症、重度阻塞型睡眠呼吸暫止症...扁真的病了，早就應該儘快「保外就醫」...	Relevant

新聞立場辨識與自然語言推理任務的目標相似，藉由輸入兩個訊息後，辨識一個給定的陳述（**premise**）在語義上是否蘊涵另一陳述（**hypothesis**）。然而，由於新聞文章屬於由多個句子組成的結構（篇章），因此，在僅知道整篇文章與議題之間的關係，且文章中的每個句子與議題間的關係都是未知的情况下，如何去結合每個句子和議題的蘊涵資訊，以及運用句子間的篇章關係（**Discourse Relations**）成為本項任務困難之處。另一方面，本論文以新聞立場檢索競賽提供的資料作為訓練模型之參考，但我們發現該資料中，新聞內容與議題的立場不同的資料比例很少，此現象造成訓練模型不易，也是本項任務的困難之處。

根據上述的問題，我們參考了 **Decomposable Attention Model**[1]，藉由該模型針對輸入的兩個句子編碼生成句子表示（**Sentence Representation**），並結合 Durmus 等人[2]在立場辨識任務中使用了階層式架構的想法，提出 **Hierarchical Decomposable Attention Model**，階層式分析文章每個句子與議題的關係。針對資料不平衡的問題，我們試圖加入新聞內容與議題立場不相關的資料來平衡資料，藉此提升模型在不相關資料的效能。

二、相關研究

（一） Natural Language Inference (NLI)

在 NLI 相關的研究中，Parikh 等人提出的 **Decomposable Attention Model**[1]是基於對齊的概念[5]的神經網路架構。該架構包含：**Attend**、**Compare** 以及 **Aggregate** 三個部分，輸入為兩個句子（以 a 和 b 表示），輸出為兩者關係預測（以 y 表示）。首先，在 **Attend** 的部分對齊（**soft-align**） a 和 b 中每個詞。接著，在 **Compare** 的部分會比較原有的句子和上個部分所計算的對齊子句段（**aligned subphrase**）。最後，在 **Aggregate** 的部分會結合 **Compare** 的結果並輸出最後的標記。相較於其他現有的架構，由於 **Decomposable Attention Model** 所需訓練的參數量少，訓練的時間也較短，但效能卻能超越其他更為複雜的模型。然而，

由於沒有考慮詞的順序，因此在某些情形下會有準確率不高的現象。

（二） Stance Detection

立場辨識（**Stance Detection**）是根據一段帶有立場的文本，將作者對目標（如：議題或事件）的立場進行分類（如：贊成或反對）的問題。目前與立場辨識相關的任務主要可以分成三種類別[6]，分別是：**Generic Stance Detection**、**Rumour Stance Classification** 以及 **Fake News Stance Detection**。

Generic Stance Detection 的例子為 **SemEval 2016 stance dataset**[7]，該資料集的每筆資料由一段文字以及目標組成，此任務的目的即是根據以上資訊辨識出作者對於目標的立場是 *favor*、*against* 或是 *neither*。與 **Rumour Stance Classification** 相關的問題出現在 **RumourEval task**[8]，該任務必須根據包含謠言（*rumour*）的來源推文和同一對話線程中的推文，將對話線程中的推文對於謠言的立場分類成 *supporting*、*denying*、*querying* 或是 *commenting*。[Fake News Challenge](#) 的立場辨識任務被視為辨識假新聞的首要步驟，其目的是建立一個可評估新聞來源對特定主張所含有的立場的自動化系統。立場分為四種：*agree* 表示文章內容與標題立場一致；*disagree* 表示文章內容與標題立場不一致；*discuss* 表示文章內容討論到標題，但不含有立場；*unrelated* 表示文章內容沒有討論到標題。

Durmus 等人[2]定義了另一種立場辨識的任務：**Claim Stance Detection**。在該任務中，必須根據 **Argument Path** 上的前後兩個主張，判斷後者的立場是支持或反對前者。實驗結果顯示，透過階層式（*hierarchical*）而非平坦式（*flat*）來表示 **Argument Path** 上的內容可以達到較好的結果。

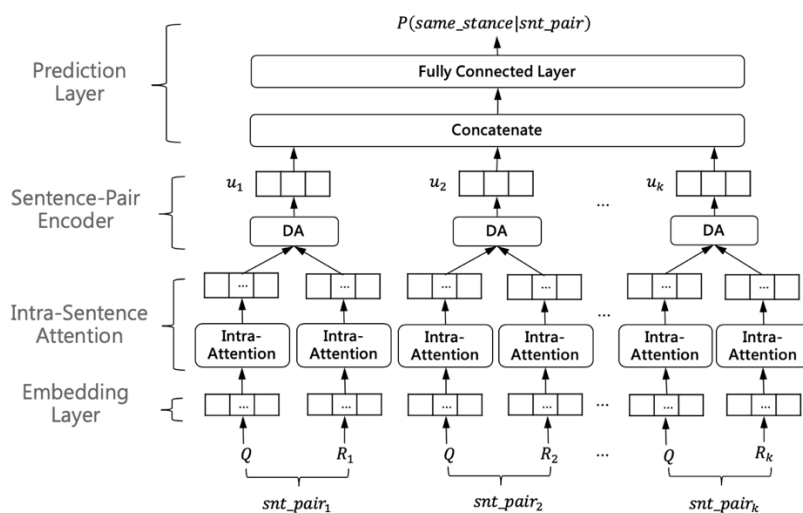
（三） Discourse Parsing

篇章分析是分析篇章中子句之間的層次結構和語義關係，進而了解篇章所要表達的意義。篇章剖析（**Discourse Parsing**）模型以標記好的篇章剖析語料庫作為訓練資料，如英文的 **Rhetorical Structure Theory (RST)**[9]和 **Penn Discourse Tree-bank Project (PDTB)**[10]，前者以樹狀結構（*tree structure*）來表示篇章中句子之間的關聯，後者以平坦結構（*flat structure*）表示含有顯式連接詞或隱式連接詞的兩個句子之間的關係；中文的 **Chinese Discourse Tree Bank (CDTB)**[11]語料庫則與 **RST** 相似，以樹狀結構來表示篇章關係，並將關係分成：因果（*Causality*）、轉折（*Transition*）、解說（*Explanation*）以及並列（*Coordination*）。

篇章剖析的過程包含四個步驟：基本篇章單位的切割、剖析樹結構建立、子句關係標記以及中心關係標記。基本篇章單位 (Elementary Discourse Unit, EDU)，又稱為子句，是篇章剖析樹上的子節點。在建立剖析樹之後，所有子句會以階層式樹狀架構呈現。最後，標記剖析樹中相連子樹之間的關係 (Sense Label) (如：並列、因果、轉折或解說) 以及中心關係標記 (Center Label) (如：核心結構在前、在後或相等)。

三、模型架構

新聞立場辨識的目的為根據給定的文章與議題，辨識兩者的立場是否相同。根據此任務，我們提出 Hierarchical Decomposable Attention Model。首先，基於 Decomposable Attention Model[1]，對輸入的兩個句子進行編碼生成句對表示 (Sentence-Pair Representation)。接著，受到 Durmus 等人[2]在 Claim Stance Detection 任務中使用了階層式架構的啟發，我們以階層式的方式處理新聞文章，找出新聞文章中的句子彼此間的關係，進而判斷文章與議題之間的立場是否相同。模型架構如圖一所示，包含了以下部分：**Embedding Layer**、**Intra-Sentence Attention**、**Sentence-Pair Encoder** 以及 **Prediction Layer**。模型的輸入由新聞文章與議題配對而成，輸出為對應的預測標記以表示兩者的立場是否相同。



圖一、Hierarchical Decomposable Attention Model 架構圖

(一) Input Representation

模型的輸入為議題與新聞文章，以 (Q, D) 表示，文章 D 包含了 n 個句子，以 $D = \{S_1, \dots, S_n\}$ 表示。輸出為兩者的立場是否一致，以 y 表示。訓練資料以集合 $\{Q^{(n)}, D^{(n)}, y^{(n)}\}_{n=1}^N$ 表示，

其中 $y \in \{0, 1\}$ 。由於文章平均大約為 53 個句子，加上文章當中可能包含與議題較為無關的部分，因此我們嘗試使用以下設定，從中選取 k 個句子，以 $R = \{S_{f_1}, \dots, S_{f_k}\}$ 表示， f_k 為被選取到的句子索引。為了簡化符號，我們令 $R_i = S_{f_i}$ ，並以 $R = \{R_1, \dots, R_k\}$ 表示被選取到的 k 個句子。

1. **Original order:** 根據句子在文章中原有的順序，選取前 k 個句子。
2. **Cosine similarity:** 根據每個句子和議題的詞嵌入 (Word Embedding)，計算兩者之間的餘弦相似度，並根據相似度排序，選取前 k 個與議題相似的句子。
3. **Discourse parsing:** 使用篇章剖析模型 RvNN-CYK2 Seq-EDU + self-attentive Model[12] 處理新聞文章，產生對應的子句、關係標記 (Sense Label) 和中心關係標記 (Center Label)。由於我們認為含有“Causality (因果)”和“Transition (轉折)”關係的句子在文章中含有較為重要的資訊，因此我們從中選取前 k 個被標為“Causality (因果)”和“Transition (轉折)”關係的中心子句。

有了選取好的句子後，我們以兩種設定形成句子配對 snt_pair_i ，作為模型的輸入：

1. **Parallel:** 議題 Q 與每個被選取到的句子 $R = \{R_1, \dots, R_k\}$ 配對，形成 $\{(Q, R_1), (Q, R_2), \dots, (Q, R_k)\}$ 。
2. **Series:** 議題 Q 僅與 R_1 配對，被選取到的句子 R' 則依照順序配對，形成 $\{(Q, R_1), (R_1, R_2), \dots, (R_{k-1}, R_k)\}$

(二) Embedding Layer

針對長度為 m 議題 Q 和 l 的句子 R_s ，以訓練好的 Word2Vec 模型表示每個詞的向量，分別為 $Q = (q_1, \dots, q_m)$ 和 $R_s = (w_1, \dots, w_l)$ ， $\forall s \in [1, \dots, k]$ ， $q_i, w_j \in \mathbb{R}^{d_w}$ ， d_w 為 Word2Vec 模型的維度。

(三) Intra-Sentence Attention

透過計算每個句子中每個詞的權重，找出句子本身重要的詞。如公式(1)(2)所示， c_{ij} 為同一個句子的詞向量經過全連階層 F_{intra} 後彼此內積的結果； c_{ij} 透過 softmax 運算，並與原本的詞相量進行加權總合後，可得 q'_i 和 w'_i 。在後續的運算中，我們將原有的詞向量與 q'_i 和 w'_i 連接，作為新的詞向量，以 $q_i := [q_i, q'_i]$ 和 $w_i := [w_i, w'_i]$ 表示。其中， $W_1 \in \mathbb{R}$ 和 $b_1 \in \mathbb{R}$ ，

皆為模型中可訓練的參數。

$$\begin{aligned} c_{ij} &= F_{intra}(q_i)^T F_{intra}(q_j) \\ F_{intra}(q_i) &= \text{ReLU}(q_i^T W_1 + b_1) \end{aligned} \quad (1)$$

$$q'_i = \sum_{j=1}^m \frac{\exp(c_{ij})}{\sum_{k=1}^m \exp(c_{ik})} q_j \quad (2)$$

(四) Sentence-Pair Encoder

我們使用 Decomposable Attention model[1] (以 DA 表示) 作為 Sentence-Pair Encoder，針對 snt_pair 進行編碼以取得對應的句對表示 $u_s \in \mathbb{R}^{d_u}$ 。最後，我們將 u_s 向量連接取得向量 $u \in \mathbb{R}^{kd_u}$ ，如以下公式：

$$u_s = DA(snt_pair_s), \quad \forall s \in [1, \dots, k] \quad (3)$$

$$u = [u_1, u_2, \dots, u_k] \quad (4)$$

(五) Prediction Layer

向量 u 經過 Prediction Layer 計算文章與議題為相同立場的機率 P ，如下列公式所示， $W_{p_1} \in \mathbb{R}^{kd_u \times hidden}$ 、 $W_{p_2} \in \mathbb{R}^{hidden}$ 、 $b_{p_1} \in \mathbb{R}^{hidden}$ 以及 $b_{p_2} \in \mathbb{R}$ 皆為模型中可訓練的參數， σ 為激勵函數。

$$P(\text{same_stance} | snt_pair) = \sigma(\text{ReLU}(u^T W_{p_1} + b_{p_1}) W_{p_2} + b_{p_2}) \quad (5)$$

模型在訓練的過程中的損失函數為交叉熵(cross-entropy)，並使用 L2 正規化 (權重為 λ)， θ 為模型中可訓練的參數集合， N 為訓練資料的大小。計算方式如下：

$$L(\theta) = \sum_{n=1}^N [z^{(n)} \cdot \log P^{(n)} + (1 - z^{(n)}) \cdot \log(1 - P^{(n)})] + \lambda \|\theta\| \quad (6)$$

四、實驗與分析

在本章節中，將詳細分析本論文所使用的資料，並比較其他架構與本論文提出之模型在新聞立場辨識的效能。最後，我們針對不同的資料輸入形式進行比較，找出合適的方式處理新聞資料。

(一) 資料集

本研究所使用的資料來自 AI CUP 2019 的[新聞立場檢索競賽](#)所提供的新聞語料庫 (NC) 以及訓練語料 (TD)。新聞語料庫包含了 80 萬篇新聞編號以及連結，我們根據這些連結爬取了對應的文章。訓練語料包含了 4,662 筆資料，資料範例如表二所示，包含 Query、News_Index、Relevance 三個欄位，Query 為訓練查詢題目，News_Index 為編號，Relevance 為相關程度，相關程度 0、1、2、3 分別代表不相關 (0)、部分相關 (1)、相關 (2)、非常相關 (3)。Query 必定含有立場，若某一文件之內容與 Query 內的議題有關，但立場與 Query 不一致，仍視為不相關 (0)。

表二、訓練語料範例

Query	News_Index	Relevance
贊成流浪動物零撲殺	news_000109	3
核四應該啟用	news_000156	1
遠雄大巨蛋工程應停工或拆除	news_000684	0
拒絕公投通過門檻下修	news_000091	2

表三為在訓練語料所包含文章的相關數據統計。我們使用[中研院詞庫小組提供的工具](#)進行中文斷詞，並以“，。、！？”等符號來分隔句子。

表三、訓練語料中的文章的相關數據統計

	Mean
# Sentence / Article	53
# Word / Article	383
# Char / Article	704
# Word / Sentence	7
# Char / Sentence	13

訓練語料中各相關等級所含有的資料個數如表四所示。在實際觀察訓練語料後，我們發現相關程度與一篇新聞中對於議題立場所佔的比例有正向關係，且大部分文章僅含有支持或反對其中一方的觀點。另一方面，該競賽在評估 MAP@300 時認定相關程度在 1 以上即屬於立場相同。因此，本研究參考主辦單位的評估標準，根據立場相同與否來分類新聞文章，而不考慮細部的相關程度。若依照此分類方式，其資料比例如表五的 Original 欄位所示，由於不相關立場的資料比例大約只佔整體資料的 12%，當分成兩類時，即出現了資料不平衡的問題 (Data Imbalance Problem)。

表四、訓練語料各類資料的比例

Relevance	#Articles	Pic %
0 (不相關)	546	12%
1 (部分相關)	2071	44%
2 (相關)	1537	33%
3 (非常相關)	508	11%
Total	4662	100%

表五、根據立場相關與否將訓練語料分成兩類後的資料比例

Relevance	Original	Opposite	Original+Opposite
0 (不相關)	546 (12%)	4559 (98%)	4615 (55%)
1 (相關)	4116 (88%)	103 (2%)	4219 (45%)
Total	4662	4662	9324

針對資料不平衡的問題，我們試圖增加標記為「不相關」的資料個數以平衡資料。我們以手動的方式產生與原有議題立場相反的議題（以下稱 **Opposite Query**），將原有訓練語料的 **Query**（以下稱 **Original Query**）中的一些詞替換成相反意義的詞。例如：「支持」替換成「反對」、「應該」替換成「不應該」…依此類推。所有立場與 **Original Query** 「相關」的文章，對 **Opposite Query** 的立場變為「不相關」；考慮到對 **Original Query** 「不相關」的文章可能包含立場中立的情形，因此我們利用人工標記的方法，去標記其對 **Opposite Query** 的立場。如表五的 **Original+Opposite** 欄位所示，兩種類別的資料比例變得較為平衡。

（二）實作細節

在資料前處理方面，我們將文章中所有出現的網址以“_url_”符號表示，並移除掉所有中文、英文以及數字以外的符號。在詞嵌入方面，用前處理好的新聞語料庫（**NC**）作為訓練資料，設定 **window size** 為 5，並以 **Skip-gram** 建立維度為 100 的 **Word2Vec** 模型[12]。

在超參數的部分，輸入的批次量為 32、**Epoch** 最大為 300、模型在驗證資料的效能若超過 60 個 **Epoch** 沒有提升，將停止訓練。**Hidden Size** 為 100、**Dropout** 為 0.3、學習率為 $1e-3$ 、最佳化器使用 **Adam**、**L2** 正規化權重 λ 設定為 $1e-4$ 。

競賽單位提供的訓練語料中包含了 20 種不同的 **Query**，為避免相同的 **Query** 同時出現在訓練與測試資料中，因此我們根據 **Query** 的不同去分割資料，並重複以下切割方式十次來評估模型：如表六所示，我們隨機選取 **Original Query** 中 80%的資料以及其 **query-article pair** 作為訓練資料；測試資料的部分，我們分成兩種測試資料，一種是

Original Query 中訓練資料以外的 20% (以下稱為 **Original Test**)，一種則是 Original Test 相對應 Opposite Query (以下稱為 **Opposite Test**)。

表六、訓練與測試資料的分割

Split	# Original queries	# Opposite queries
Train	16 (80%)	16
Test	4 (20%)	4

考慮到資料在兩類別的數量不平衡，我們同時計算了 macro-average 和 micro-average F1-score，並以兩者的平均作為後續實驗的評估方法：

$$Avg_F = \frac{1}{2}(Marco_F + Micro_F) \quad (7)$$

(三) 模型效能比較

本節將針對其他三種架構與本論文提出的模型進行比較，同時比較在加入訓練資料之後對效能的影響。在本節的實驗中，每篇文章被選取的句子個數 (以 k 表示) 皆為 5。我們比較的三種架構如下所述。

Decomposable Attention model[1] (以 **DA (Flat)** 表示)：原始輸入的句子配對 (a, b) 改為議題與文章配對 (Q, R') ， Q 為議題， R' 由文章 R 中前 k 個句子 $R' = \{R_1, \dots, R_k\}$ 串接而成，以 Flat 的形式輸入，議題與文章最大長度皆設定為 100 個詞。

Fine-tuned BERT[14] (以 **BERT (Flat)** 表示)：輸入序列為 (Q, R') 配對之間插入分隔符號 [SEP]，並以分類符號 [CLS] 作為序列開頭，並在 [CLS] 的輸出位置接上一個線性分類器。議題與文章最大長度皆為 100 個字元。使用預訓練好的 BERT 中文模型，其參數設定 Transformer 層數為 12、Hidden Size 為 768、Multi-Head 數量為 12。

Fine-tuned BERT (hierarchical)[2] (以 **Hi-BERT** 表示)：每對 sentence pair 經過 BERT 後編碼取得 sentence pair representation R_{pair} ，接著，透過雙向 GRU 找出每對 R_{pair} 之間的關係，最後，接上線性分類器後輸出預測結果。Bi-GRU 的 Hidden Size 為 100。兩句子合計最大長度設定為 30 個字元 (含 [CLS] 符號與 [SEP] 符號)。

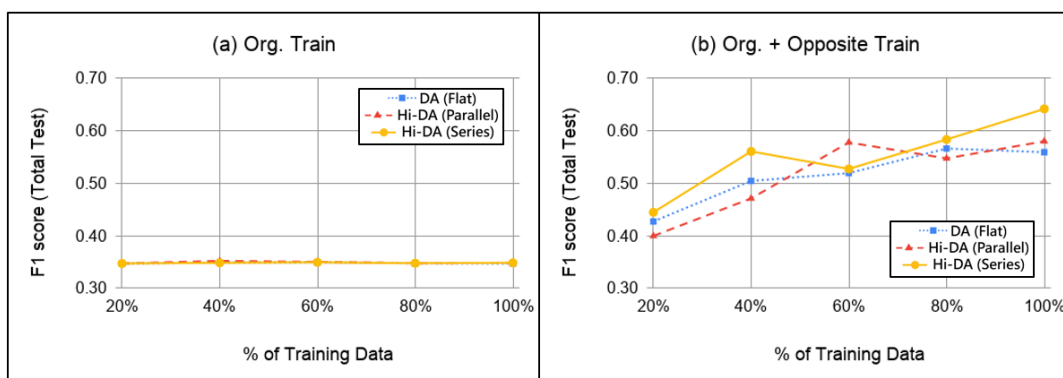
表七為在 Original Train 和加入 Opposite Train 兩種情形下訓練模型後，在 Original Test 和 Opposite Test 兩種資料集測試的結果。首先，比較各模型後，我們發現比起以 BERT 為基礎的模型 (BERT (Flat) 和 Hi-BERT)，以 Decomposable Attention 為基礎的模型

可以達到較好的效能 (DA (Flat)和 Hi-DA)。接著，在我們將 DA (Flat)改為階層式的架構之後，模型在兩種測試資料集的效能皆有所提升。最後，可以發現對所有模型而言，若僅以 Original Train 訓練，模型在 Opposite Test 上的效能明顯不佳，我們推論這是由於 Original Train 當中僅有少部分「不相關」的資料，因此模型難以學習如何辨識「不相關」的資料；當我們加入 Opposite Train 訓練後，可有助於模型辨識出「不相關」的資料，因此在 Opposite Test 的結果有明顯提升。

表七、不同模型在加入 Opposite Train 後的效能比較

Model	Avg. F1 score					
	Trained with Org. Datasets			Trained with Org.+ opposite Train		
	Original Test	Opposite Test	Total	Original Test	Opposite Test	Total
BERT (Flat)	0.6556	0.0385	0.3471	0.4319	0.5345	0.4832
Hi-BERT (Parallel)	0.6567	0.0389	0.3478	0.3326	0.5671	0.4498
Hi-BERT (Series)	0.6554	0.0386	0.3470	0.3407	0.5599	0.4503
DA (Flat)	0.6522	0.0498	0.3510	0.6046	0.4973	0.5509
Hi-DA (Parallel)	0.6605	0.0387	0.3496	0.6571	0.5040	0.5806
Hi-DA (Series)	0.6559	0.0380	0.3470	0.6592	0.6245	0.6419

為了進一步驗證加入 Opposite Train 能夠提升效能，我們比較模型在使用 Original Train 以及 Original + Opposite Train 訓練的學習曲線，如圖二所示。在圖二左半部中，所有模型的效能皆沒有隨著資料的增加而上升；在圖二右半部中，各模型的效能整體都比左半部好，同時，相較於其他兩種模型，Hi-DA (Series)的效能大致上隨著資料的增加而上升，並達到比其他兩種架構高的效能。



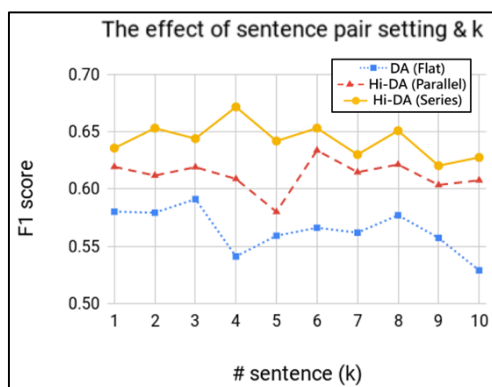
圖二、訓練資料為 Original Train 以及 Original + Opposite Train 的學習曲線

(四) 資料形式的影響

此節將分析我們所提出的從文章中選擇句子的三種方法（以下稱 Sentence Filter）和配

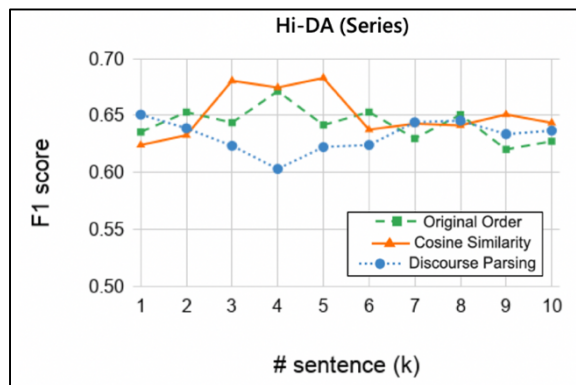
對句子的兩種設定（以下稱 **Sentence Pair Setting**）對模型效能的影響，並皆以 **Original Train + Opposite Train** 作為訓練資料。

在 **Sentence Pair Setting** 比較的實驗中，比較當句子以 **Flat** 和 **Hierarchical** 的形式作為輸入後，以 **Decomposable Attention** 為基礎的模型在測試資料的表現，如圖三所示。可以觀察到當句子以 **Series** 的方式配對時，結果普遍比其他兩種形式好。顯示出 **Series** 的方式相較於其他兩種方法，更有助於模型判別文章中的每個句子之間的關聯，並進一步推論文章與議題的立場相同與否。



圖三、不同 **Sentence Pair Setting** 對模型效能的影響

在 **Sentence Filter** 比較的實驗中，我們以三種 **Sentence Filter**：依照句子在文章中原有的順序（**Original Order**）、議題與句子之間的相似度（**Cosine Similarity**）以及透過篇章剖析模型處理（**Discourse Parsing**）等方法從每篇文章中選取 k 個句子後，比較 **Hi-Decomposable Attention (Series)** 模型在測試資料的表現，如圖四所示。可以發現若以議題與句子之間的相似度（**Cosine Similarity**）作為選擇句子的依據，當 $3 \leq k \leq 5$ 時，相較於其他兩者方法可以取得較好的結果。因此，我們認為使用相似度（**Cosine Similarity**）找出和議題較相關的句子，對模型效能的提升是有幫助的。另一方面，我們也發現使用篇章剖析模型處理文章（**Discourse Parsing**）後無法提升效能的現象。



圖四、Hi-DA (Series)在使用不同 Sentence Filter 後的結果

五、結論與未來工作

本篇論文提出 Hierarchical Decomposable Attention Model，來解決新聞立場辨識任務，同時藉由實驗比較不同的方式來配對和選擇新聞文章中的句子。根據實驗結果，我們提出的模型效能優於其他架構之外，各種模型在加入訓練資料後，效能都有所改善。

本論文提出的模型可與 Solr 或 Elasticsearch 等搜尋平台結合：先透過搜尋平台找出與議題相關的文章後，再以訓練好的立場辨識模型去判別該文章與議題的立場是否一致。但由於我們目前無法取得競賽單位所提供的測試題目對應的標記，因此未來若取得相關標記資料，我們將驗證該模型是否能幫助搜尋平台在立場判別的辨識。另一方面，針對使用篇章剖析模型處理文章後卻無法提升效能的現象，我們提出以下方法改善：第一，透過人工的方式對篇章模型產生的標記進行驗證，再以此資料作為模型的輸入；第二，使用相關的情緒詞典（如：增廣中文意見詞詞典（AUTUSD）[15]）來辨識文章中的句子與特定議題所包含的字詞情感是否一致，再進一步辨識文章與議題的立場。

參考文獻

- [1] A. P. Parikh, O. Tackström, D. Das, and J. Uszkoreit, “A decomposable attention model for natural language inference”, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP2016, Austin, Texas, USA, November 1-4, 2016*, The Association for Computational Linguistics, 2016, pp. 2249–2255.

- [2] E. Durmus, F. Ladhak, and C. Cardie, “Determining relative argument specificity and stance for complex argumentative structures”, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul.2019, pp. 4630–4641.
- [3] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference”, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642.
- [4] A. Williams, N. Nangia, and S. R. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference”, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Association for Computational Linguistics, 2018, pp. 1112–1122.
- [5] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- [6] D. K'uk and F. Can, “Stance detection: A survey”, *ACM Comput. Surv.*, vol. 53, no. 1, Feb. 2020, issn: 0360-0300.
- [7] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, “Semeval-2016 task 6: Detecting stance in tweets”, in *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, The Association for Computer Linguistics, 2016, pp. 31–41.
- [8] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, “Semeval-2017 task 8: Rumoureal: Determining rumour veracity and support for rumours”, in *Proceedings of the 11th International Work-shop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, Association for Computational Linguistics, 2017, pp. 69–76.
- [9] W. C. Mann and S. A. Thompson, “Rhetorical structure theory: Toward afunctional

- theory of text organization”, *Text & Talk*, vol. 8, no. 3, pp. 243–281, 1988.
- [10] R. Prasad, B. Webber, and A. Joshi, “Reflections on the Penn discourse Tree Bank, comparable corpora, and complementary annotation”, *Computational Linguistics*, vol. 40, no. 4, pp. 921–950, Dec. 2014.
- [11] Y. Li, W. Feng, J. Sun, F. Kong, and G. Zhou, “Building chinese discourse corpus with connective-driven dependency tree structure”, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014*, pp. 2105–2114.
- [12] Y.-J. Wang and C.-H. Chen, “Using attentive to improve recursive lstm end-to-end chinese discourse parsing”, in *The 2019 Conference on Computational Linguistics and Speech Processing ROCLING 2019, The Association for Computational Linguistics and Chinese Language Processing, 2019*, pp. 388–397.
- [13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space”, in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013*.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [15] S.-M. Wang and L.-W. Ku, “ANTUSD: A large Chinese sentiment dictionary”, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2697–2702.

結合依存句法分析及圖神經網路的文本分類方法

Combining Dependency Parser and GNN models for Text

Classification

周冠勳 Kuan-Hsun Chou

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

t107598035@ntut.org.tw

吳炎蒼 Yen-Tsang 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

buddyswu@gmail.com

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

隨著數據量的擴增，人為地對文本進行分類是相當耗成本的，因此自動化的文本分類變得十分重要，如垃圾郵件偵測、新聞分類、情緒分析等。目前自然語言方面的深度學習模型大致上分為兩類：sequential 和 graph based，sequential 模型通常都是使用 RNN 和 CNN，以及近年來在各方面都很突出的 BERT 模型及其變種；近年來有許多的研究，開始將 graph based 的深度模型應用在 NLP 上，利用文字之間的 co-occurrence 關係，從而學習到文字和文本的特徵，以進行分類。本論文首先使用 RNN 計算出文本中的文字特徵，將所有文字當作 node，並用文字之間的修飾關係建 graph，使用 graph model 重新得到文字特徵，並預測文本類別。

在實驗中，我們使用多種資料集，MR、R8、R52 和 Ohsumed 作為驗證。與多種模型，TF-IDF+LR、CNN、LSTM、PV-DBOW、PV-DM、PTE、fastText、SWEM、LEAM 和 Text GCN 進行比較。在 MR 上獲得較好的結果 (Accuracy: 79.42%)。

關鍵詞：依存句法分析，圖神經網路，文本分類

ABSTRACT

As the amount of data increases, manually classifying texts is expensive. Therefore, automated text classification has become important, such as spam detection, news classification, and sentiment analysis. Recently, deep learning models in natural language are roughly divided into two categories: sequential and graph based. The sequential models usually use RNN and CNN, as well as the BERT model and its variants; In recent years, researchers started to apply the graph based deep learning model to NLP, using word co-occurrence and TF-IDF weights to build graphs in order to learn the features of words and documents for classification.

In the experiment, we use different datasets, MR, R8, R52 and Ohsumed for verification. Comparing with sequential and graph-based models, the accuracy of our proposed method on MR can achieve 0.79.

Keywords: dependency parser, graph neural network, text classification

一、緒論

隨著網路的發展，資訊傳播的速度與數量快速的增加，若想搜尋某種資料，沒有特別分類方法，使用者便需要仔細地閱讀整筆資料，從而將不必要的資訊過濾掉。如何減少使用者的負擔，同時輔助查詢的準確度，文本分類可以說是一個基本的技術。

文本分類是各種應用軟體的核心，Email 系統裡面使用分類器來區分是不是垃圾郵件；在電商裡面，可以用來區分商品底下的評論是否為垃圾評論，幫助消費者與賣家更好地從評論中得到回饋，或是從文字中，了解文本所述說的主題。

早期的文本分類方法都是使用 **bag-of-word** 來代表文本的意思，即計算文本中出現哪些字詞，來作為文本的表示，而這種方法並未考慮字詞的上下文，隨著類神經網路模型的興起，開始出現了 RNN 和 CNN 應用在文本分類上，近年來開始有新的研究，將類神經網路應用在 **graph** 的資料結構上，常見使用在 **knowledge graph**、社群網路、文章引用、分子結構.....。在文本分類上，由於句子或文章本身沒有這種結構，會需要其他方式進行處理。其中一種方法是將文字和文本作為 **graph** 的 **node**，以他們的 **co-occurrence**

作為邊，以這種方式學習文本特徵，然而隨著文字和文本的增加，graph 也會變大，需要更多的記憶體來載入整個 graph；另一種方法則使用 dependency parser 對句子做處理，使其變成 tree 的結構，再進行後續的任務。

根據本論文實驗，使用 dependency parser 的結果結合圖神經網路 (Graph Neural Network, GNN) 在情緒分類文本 MR 上，Accuracy 可以到達 79.42%。

二、相關研究

Mikolov 等人[1]提出 Word2vec 將文字轉成向量的模型，用大量的文本輸入到模型中，來捕獲文字之間的相關性，同時也避免了 bag-of-word 的缺點；Omer 等人[2]將 dependency parser 與 word2vec 結合，學習到句法 (Syntax) 特徵；Melamud 等人[3]提出 context2vec，使得文字特徵包含了前後文資訊，隨後 Peters 等人[4]基於 bidirectional language model (biLM)的想法學習文字特徵，在 NLP 各種 task 上得到高分；Kim 等人[5]將 CNN 模型應用在文本分類。

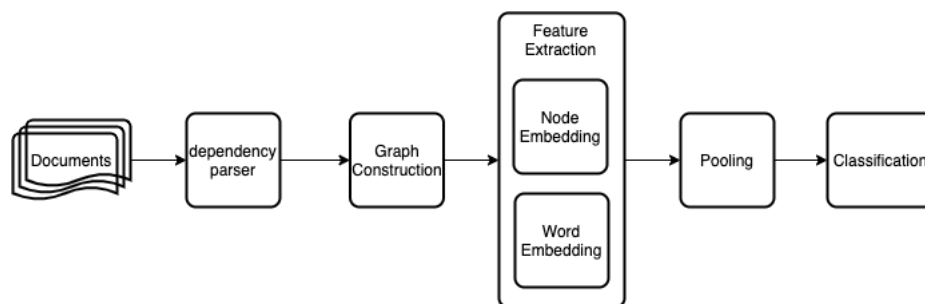
近年來有許多的 GNN 模型應用在 NLP 的任務上。Kipf 等人[6]提出 GCN 模型，GCN 模型為一種多層的架構，藉由 node 與鄰居 node 之間的 aggregate，來更新 node 的特徵，在 knowledge graph 和引文網路上取得了很好的分類效果。Veličković 等人[7]提出 GAT 模型，將 attention 和 multi-head 機制[8]加到 GCN 中，在 aggregate 階段，給予每個 node 不同的重要度，效果超出 GCN。隨後便有人將 GNN 和 dependency parser 結合，應用在 aspect level sentiment classification[9, 10]和 word embedding[11]。Yao 等人[12]提出 textGCN 則將 GCN 應用於一般的文本分類。

本論文嘗試將 dependency parser 與 GAT 結合，應用在一般的文本分類，並且與 TextGCN 進行比較，期望得到較好的分類效果。

三、研究方法

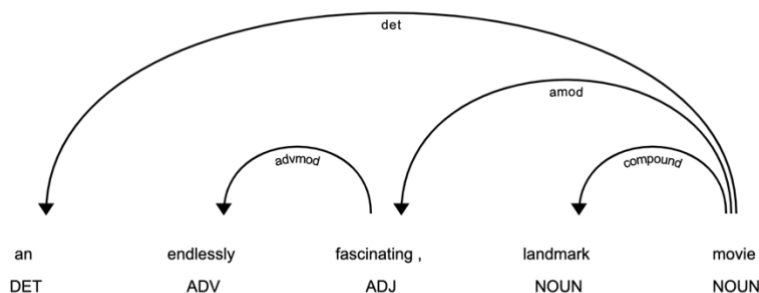
本方法分為五大步驟，如下圖一，首先，使用 dependency parser 取出文本中每個文字的

修飾關係，利用這種關係建 graph 的 edge，而文本中的文字則作為 graph 的 node。在 Feature Extraction 階段，會先使用 RNN 模型得到 word embedding，隨後將此作為初始的 node embedding，與 graph 一同輸入到 GAT 模型中，得到新的 node(word) embedding；在 Pooling 階段，我們對 node embedding 進行 pooling 作為整個 document 的特徵，輸入到分類當中。



圖一、系統架構圖

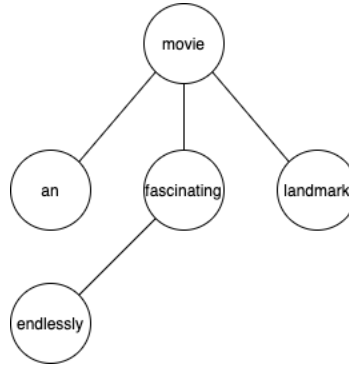
dependency parser 的結果如圖二所示，每個句子都有個 root 文字，不修飾其他文字，如同圖二中的 movie，而其他的文字則會去修飾其他的文字，如 fascinating、a 和 landmark 會去修飾 movie，endlessly 會修飾 fascinating，我們將會在後續利用這種關係建 graph。



圖二、dependency parser 示意圖

建出的 graph 如圖三所示，原本 dependency parser 的修飾關係是有方向的，此處我們將其去掉，改成了無向邊，邊的方向藉由後面的模型學習，若 node 沒有通到另一個 node，則邊的權重視為 0。在文字特徵的結合上，movie 會佔有比較多 an、fascinating 和 landmark 的特徵，而 endlessly 是 movie 的 2-hop neighborhood，兩個文字之間並無直接的關係，得到的特徵會相對的少；對 fascinating 而言，由於無向的關係，可以結合 movie 和

endlessly 的特徵，由此來學習到 node embedding。



圖三、dependency graph 示意圖

整個 node embedding 的更新使用 GAT 來計算，如公式(1)和(2)。公式中 x 為 node embedding， l 為更新的次數， Θ 為線性轉換， α 為 node 之間的權重， $N(i)$ 為 node i 的鄰居 node，從整個公式來看，所有 node embedding 經過一次線性轉換，然後計算 node i 和他的所有鄰居 node j 之間的權重，計算完權重之後，乘上 node embedding 做加總。以圖三為例，movie 會計算 an、movie、fascinating 和 landmark 之間的權重，若 movie、fascinating 和 landmark 所佔的權重較高，則新的 movie 特徵則包含了 movie、fascinating 和 landmark 的特徵。

權重的計算方法如公式(2)將線性轉換過後的 node 特徵 x_i, x_j 作串接，隨後乘上一個 weight vector \vec{a} ，再經過一個 LeakyReLU activation，接下來使用 softmax 求權重。

$$x_i^{(l+1)} = \alpha_{i,i} \Theta x_i^{(l)} + \sum_{j \in N(i)} \alpha_{i,j} \Theta x_j^{(l)} \quad (1)$$

$$\alpha_{i,j} = \frac{\exp(\text{LeakyReLU}(\vec{a}^T [\Theta x_i^{(l)} \parallel \Theta x_j^{(l)}]))}{\sum_{k \in N_i} \exp(\text{LeakyReLU}(\vec{a}^T [\Theta x_i^{(l)} \parallel \Theta x_k^{(l)}]))} \quad (2)$$

而最初的 node embedding， $x^{(0)}$ ，我們使用 biRNN[13]的架構來取得各個文字包含上下文的特徵，RNN 的部分則使用 GRU[14]，如公式(3)， w_0, w_1, \dots, w_n 為文本中的文字經過 word2vec 轉換過的向量， n 為該文本的長度。

$$x^0 = [h_0, h_1, \dots, h_n] = \text{biRNN}(w_0, w_1, \dots, w_n) \quad (3)$$

在 pooling 階段我們測試兩種做法，1. 以 root 文字作為整個句子特徵如公式(4)，如同前面所述，dependency parser 會有一個 root 文字，其會被其他文字修飾或間接的修飾，在

經過 node embedding 的計算過後，root 文字本身就包含了修飾詞的特徵。2.我們將所有的 node embedding 相加取平均，作為整句話的特徵如公式(5)。隨後將句子特徵輸入到一層 NN 做分類，如公式(6)。

$$\vec{doc} = x_{root}^l \quad (4)$$

$$\vec{doc} = \frac{1}{n} \sum_{i=0}^n x_i^l \quad (5)$$

$$class = softmax(\theta \vec{doc}) \quad (6)$$

四、實驗結果

本論文參照 TextGCN[12]所使用的資料集，包括：MR、R8、R52 和 Ohsumed。MR 資料集為電影評論，主要是作為情緒分類使用，包括正面情緒 5331 筆，負面情緒 5331 筆。R8 和 R52 為 Reuters-21578 部分資料，分別取其中的 8 種和 52 種類別作為資料集。Ohsumed 為醫學摘要，經過過濾[12]，每篇文章都只描述一種心血管疾病，總共有 23 種類別。

表一顯示本論文提出的模型和其他模型的分類結果比較[12]，我們對每個資料集跑 10 次，來計算平均值和標準差。

表一、各個模型與對應資料集的 accuracy

Model	R8	R52	Ohsumed	MR
TF-IDF+LR	0.9347	0.8695	0.5466	0.7459
CNN-non-static	0.9571±0.0052	0.8759±0.0048	0.5844±0.0106	0.7775±0.0072
Bi-LSTM	0.9631±0.0033	0.9054±0.0091	0.4927±0.0107	0.7768±0.0086
PV-DBOW	0.8587±0.0010	0.7829±0.0011	0.4665±0.0019	0.6109±0.0010
PV-DM	0.5207±0.0004	0.4492±0.0005	0.2950±0.0007	0.5947±0.0038
PTE	0.9669±0.0013	0.9071±0.0014	0.5358±0.0029	0.7023±0.0036
fastText	0.9613±0.0021	0.9281±0.0009	0.5770±0.0049	0.7514±0.0020
SWEM	0.9532±0.0026	0.9294±0.0024	0.6312±0.0055	0.7665±0.0063
LEAM	0.9331±0.0024	0.9184±0.0023	0.5858±0.0079	0.7695±0.0045
Text GCN	0.9707±0.0010	0.9356±0.0018	0.6836±0.0056	0.7674±0.0020
Dep-GAT-root	0.9654±0.0025	0.9263±0.0062	0.6194±0.0118	0.7942±0.0059
Dep-GAT-avg	0.9611±0.0075	0.9229±0.0066	0.5630±0.0146	0.7839±0.0029

TF-IDF+LR 為傳統模型，使用 TF-IDF 作為文本特徵，使用 Logistic Regression 分類，LDA+LR 使用文本的主題分佈作為特徵，輸入到 Logistic Regression 分類；sequential deep learning 的模型有 CNN[5]和 bi-LSTM[15]，PV-DBOW 和 PV-DM 為 doc2vec[16]，將文本作為一組向量表示，使用 Logistic Regression 進行分類；PTE[17]、fastText[18]、SWEM[19]和 LEAM[20]使用不同方式學習 word embedding，並對文本所有文字的 word embedding 相加取平均或用其他 pooling 的方式，來算出文本的向量；graph deep learning 的模型為 TextGCN[12]。Dep-GAT-root 和 Dep-GAT-avg 為本論文所提出的方法。

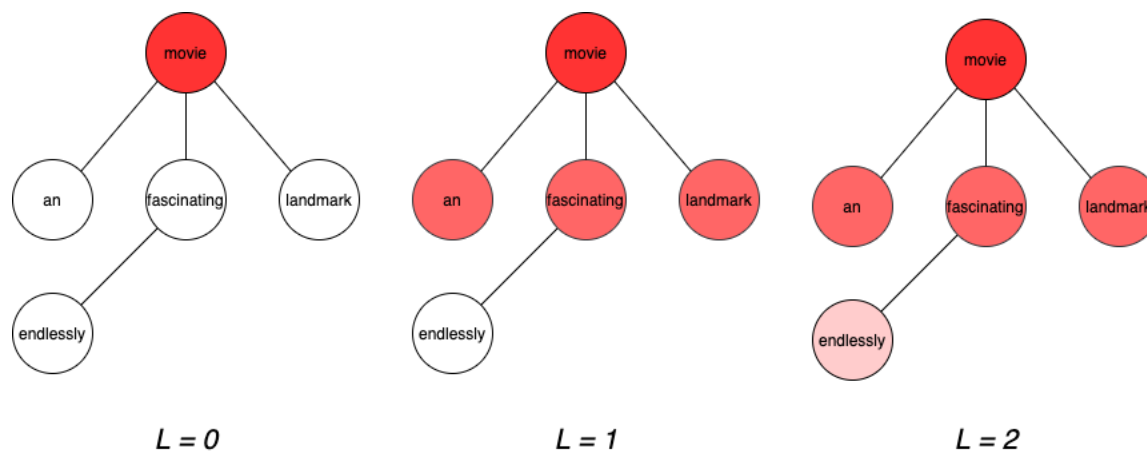
結果顯示在 MR 資料集上取得了較好的效果，accuracy 為 0.7942，BiRNN 可以在較短的文本上得到好的結果，同時 graph 學出來的 embedding 也能很好的輔助分類；然而在 R8、R52 和 Ohsumed 的分數反而不如 TextGCN，其原因可能在於單純 LSTM 和 GRU 在長文本摘要上面效果不是很好，且 dependency parser 的結構對於普通分類文本不太有幫助，在特徵擷取上，dependency parser 找出 root 文字，可將其視為句子的主要文字 (被其他詞修飾)，使得分類效果雖然不是在所有模型中表現最好，但也不會太差。

表二、各個資料集和卷積次數(L)對應的 accuracy

Dataset \ L	1	2	3	4	5
MR	0.784±0.006	0.786±0.009	0.790±0.078	0.794±0.005	0.788±0.006
R8	0.965±0.002	0.965±0.004	0.964±0.003	0.960±0.004	-
R52	0.926±0.008	0.926±0.006	0.922±0.012	0.911±0.127	-
Ohsumed	0.619±0.011	0.611±0.015	0.591±0.012	0.572±0.014	-

表二為 Dep-GAT-root 的結果，不同的卷積次數代表著每個 node 的更新次數，同時也代表著每個 node 結合了多少 hop neighborhood 的特徵，如下圖四所示，當 L=0 時，movie node 尚未得到其他 node 特徵，當 L=1 時，則得到了 1-hop neighborhood 的特徵，以此類推，而過多的卷積會帶來更多 noise。在 R8、R52 和 Ohsumed，取得 root 和修飾 root

的文字特徵，在分類上就已達極限；在 MR 上或許會需要更多的修飾詞特徵，而這些修飾詞也許跟情緒很有相關，所以需要更多的卷積次數。



圖四、node embedding 結合示意圖

五、結論

本論文嘗試將 dependency parser 與 GAT 的結合應用在文本分類中，在情緒分類的資料集上，dependency parser 能提升分類效果達到 0.794 accuracy，但是在一般文本分類上，這種修飾的關係沒有太大的幫助。在建 graph 過程中我們並無利用 dependency relation，而是單純使用 attention 來結合 node embedding，未來或許可用別種方式將 relation 的特徵加入計算中，並且將其用在情緒分類的相關任務中。

參考文獻

- [1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [2] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 302-308.
- [3] O. Melamud, J. Goldberger, and I. Dagan, "context2vec: Learning generic context embedding with bidirectional lstm," in *Proceedings of the 20th SIGNLL conference on computational natural language learning*, 2016, pp. 51-61.
- [4] M. E. Peters *et al.*, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.
- [5] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.
- [6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional

- networks," *arXiv preprint arXiv:1609.02907*, 2016.
- [7] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.
- [8] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.
- [9] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," *arXiv preprint arXiv:1909.03477*, 2019.
- [10] B. Huang and K. M. Carley, "Syntax-aware aspect level sentiment classification with graph attention networks," *arXiv preprint arXiv:1909.02606*, 2019.
- [11] S. Vashishth, M. Bhandari, P. Yadav, P. Rai, C. Bhattacharyya, and P. Talukdar, "Incorporating syntactic and semantic information in word embeddings using graph convolutional networks," *arXiv preprint arXiv:1809.04283*, 2018.
- [12] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 7370-7377.
- [13] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.
- [14] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [15] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *arXiv preprint arXiv:1605.05101*, 2016.
- [16] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188-1196.
- [17] J. Tang, M. Qu, and Q. Mei, "Pte: Predictive text embedding through large-scale heterogeneous text networks," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1165-1174.
- [18] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.
- [19] D. Shen *et al.*, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," *arXiv preprint arXiv:1805.09843*, 2018.
- [20] G. Wang *et al.*, "Joint embedding of words and labels for text classification," *arXiv preprint arXiv:1805.04174*, 2018.

使用元學習技術於資訊檢索任務之初步研究

A Preliminary Study on Using Meta-learning Technique for Information Retrieval

林崇恩 Chong-En Lin

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

M10815018@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

資訊檢索(Information Retrieval)是從海量的文件中，依據使用者提供的查詢(Query)關鍵字，從而回傳和關鍵字相關的文件，這個任務是自然語言處理(Natural Language Processing, NLP)中相當重要的研究方向。在資訊檢索中，使用少量訓練資料進行模型訓練，往往得到的成績不是很理想。所以本研究提出運用元學習(Meta Learning)中與模型無關的元學習演算法(Model-Agnostic Meta-Learning, MAML)於資訊檢索任務，透過與模型無關的元學習演算法從不同主題的文件集學習模型初始化參數，實現在新主題資料集中快速適應文件的能力。在實驗中，我們的方法使用深度類神經網路(Deep Neural Networks)架構的檢索模型（深層結構語義模型 Deep Structured Semantic Model, DSSM）搭配與模型無關的元學習演算法進行訓練，並在 TREC Robust04 資料集上進行評估，實驗結果初步表明使用與模型無關的元學習演算法訓練方式相較於傳統訓練方式有明顯的分數提升。就我們所知，本研究是首先使用元學習技術應用在資訊檢索領域，開啟了後續研究的新方向。

關鍵詞：資訊檢索、元學習、與模型無關的元學習、深層結構語義模型

Abstract

Information retrieval aims at searching documents to satisfy a user's query from a large collection of documents. The research subject is a fundamental challenge in the context of natural language processing. We all agree upon that obtaining a neural network-based retrieval model usually needs a large amount of training examples. To modulate the problem, in this study, we make a step forward to leverage the meta learning strategy for training a neural network-based retrieval model. More formally, the model-agnostic meta-learning algorithm (MAML) is employed to learn a set of model parameters from a set of training queries. After that, the model is easily to be adapted for each test query. By doing so, query dependent retrieval model can be simply obtained and achieves better results than conventional retrieval models. In the experiments, a simple neural network-based retrieval model (i.e., deep structured semantic model, DSSM) is adopted, and the resulting model is evaluated on the TREC Robust04 dataset. Preliminary results show that the use of model-agnostic meta-learning algorithm training method has a significant improvement compared with traditional training methods. As far as we know, this research is the first one, which introduces the meta-learning technique to information retrieval task.

Keywords: Information Retrieval, Meta-learning, Model-agnostic Meta-learning, Deep Structured Semantic Models

一、緒論

伴隨著互聯網蓬勃的發展和網路文件資訊的增長，各種不同瀏覽器的搜尋引擎功能，已經是使用者挑選瀏覽器的必要條件之一，而在網路的世界中，存在的文件檔案就有成千上萬筆資料，如何能有效率搜尋到和使用者相關的文件，這就是檢索系統中資訊檢索 (Information Retrieval) 的重要研究課題。資訊檢索的技術大致可以分為三大類：多媒體資訊檢索、資料庫檢索與文件資訊檢索，本研究主要探討和文件檔案相關的文件資訊檢索技術。和文件檔案相關的文件資訊檢索技術以在瀏覽器中的搜尋引擎為主，運作方式為使用者在搜尋引擎中的查詢介面提供查詢 (Query) 的關鍵文字，然後搜尋引擎會協助使用者快速檢索到和關鍵文字相關 (Relevance) 的網路文件資訊。

近年來在機器學習領域中有一個新興的研究方向：元學習 (Meta Learning)，這項技術近幾年在機器學習的各大領域引起了廣泛的討論。元學習的概念是讓機器學習到「如何學習 (Learning to Learn)」的能力，與一般的機器學習最大的差別在於元學習會先學習到一個先驗知識，這會讓機器在面對新的任務時，能更快更好得適應新的數據。在資源匱乏的訓練資料環境設置中，元學習的技術已經在很多領域上取得了不錯的成績，例如

影像視覺方面[1]、機器翻譯方面[2]和語音處理方面[3-6]。就我們所知，到目前為止，尚無研究探討使用元學習技術於資訊檢索任務之中，為了探究此一研究方向，本研究提出使用元學習技術於資訊檢索的任務，期望可以獲得任務成效的提升。

二、相關研究

(一)傳統資訊檢索模型

1、向量空間模型(Vector Space Model, VSM)

在向量空間模型中，我們會將每個查詢和文件以向量表示，其中每一個分量對應於字典中的每個單字，在空間中靠近的向量在遣詞用字上較為相近，所以我們會依照查詢文件向量與不同檢索文件向量之間的餘弦相似度(Cosine Similarity)進行文件排序。更明確地，向量中的每一個分量，通常使用詞頻－倒文件頻(Term Frequency-Inverse Document Frequency, TF-IDF)的方式計算權重：

$$TF - IDF_{i,j} = tf_{i,j} \times idf_i \quad (1)$$

$$tf_{i,j} = 1 + \log(count(w_i, d_j)) \quad (2)$$

$$idf_i = \log \left[\frac{1 + N}{1 + df(w_i)} \right] + 1 \quad (3)$$

其中， $TF - IDF_{i,j}$ 代表單詞 w_i 在文件 d_j 向量中的權重，這個權重由局部權重($tf_{i,j}$)和全局權重(idf_i)的乘積所組成。局部權重是以單詞 w_i 在文件 d_j 中出現的頻率 $count(w_i, d_j)$ 經過簡單的計算而得；全局權重則是以單詞 w_i 出現在整個文件集中的文章數 $df(w_i)$ 和整個文件集的文章數 N ，透過簡單的計算而得。因此，如果單詞 w_i 在文件 d_j 中出現的次數很高，並且它僅在少數的文件中出現，則此一單詞是相當重要的資訊；反之，若單詞 w_i 在很多篇文件中都有出現，或是在文件 d_j 中出現的次數很低，甚至是 0，則表示此一單詞對文件 d_j 而言並不重要。

2、最佳匹配模型 25(Best Match 25, BM25)

最佳匹配模型 25(Best Match 25, BM25)是資訊檢索領域用來計算查詢文件與檢索文件相似度分數的經典模型。最佳匹配模型是在對機率模型的變化進行一系列實驗後創建的，包括最佳匹配模型 1、包括最佳匹配模型 11、包括最佳匹配模型 15 以及最佳匹配模型 25 等模型。最佳匹配模型 25 是一個融合語句的倒文件頻率資訊、詞頻資訊以及文件長

度標準化的排序計算公式：

$$BM25(d_j, q) = \sum_{w_i \in \{d_j \cap q\}} tf_{i,j}^1 \times tf_{i,q}^2 \times idf_i \quad (4)$$

$$tf_{i,j}^1 = \frac{(K_1 + 1) \times count(w_i, d_j)}{K_1 \times \left[(1 - b) + b \times \frac{len(d_j)}{avg_{doclen}} \right] + count(w_i, d_j)} \quad (5)$$

$$tf_{i,q}^2 = \frac{(K_3 + 1) \times count(w_i, q)}{K_3 + count(w_i, q)} \quad (6)$$

$$idf_i = \log \left(\frac{N - df(w_i) + 0.5}{df(w_i) + 0.5} \right) \quad (7)$$

其中， d_j 為檢索文件， q 為查詢文件， N 為所有檢索文件的數目， $count(w_i, d_j)$ 和 $count(w_i, q)$ 分別為單詞 w_i 在文件 d_j 和查詢 q 中出現的次數， $df(w_i)$ 為 w_i 出現在整個文件集中的文章數， $len(d_j)$ 為文件 d_j 的長度（字數）， avg_{doclen} 為所有文件的平均長度（字數）， K_1 、 K_3 和 b 為可調整的模型參數。由最佳匹配模型 25 的計算方式可知，字詞出現在查詢文件 q 的頻率資訊會藉由權重函數 $tf_{i,q}^2$ 進行適當的調整，其中參數 K_3 會考慮到字詞是否出現於查詢文件當中，並且更進一步地將字詞在查詢文件中出現的頻率資訊做加權運算；權重函數 $tf_{i,j}^1$ 則用於計算當前文件與查詢共同出現的字詞在檢索文件中的重要性，並且對文件的語句長度進行標準化，最後用可調參數 K_1 負責進行適當加權調整；倒文件頻率資訊 idf_i 是用來決定一個字詞普遍重要性的度量，也就是會加強比較稀少出現字詞的權重，而削弱比較常出現字詞的重要性。

(三)元學習相關研究

通常在進行元學習或小樣本學習時，會有幾個重要的專有名稱，分別是： N 類別 K 樣本 (N -way K -shot)、支撐集(Support Set)、查詢集(Query Set)、元學習訓練集($D_{meta-train}$)以及元學習測試集($D_{meta-test}$)。 N 類別 K 樣本是在進行小樣本學習中常見的實驗設置，是指模型訓練一次任務時所會用到的資料數量，例如在分類任務中，表示每個訓練任務包含 N 個類別，並且每個類別分別有 K 筆訓練資料，因此，每個訓練任務共會使用 $N \times K$ 筆資料進行訓練。元學習訓練集則是由多組訓練任務資料組合而成，每組訓練任務皆包含支撐集和查詢集。這些資料通常是利用具有豐富資源的資料來進行模型預訓練。元學

習測試集同樣是由多組支撐集和查詢集組合而成。這些資料通常是用在新的任務上，模型會先經過元學習訓練集的資料訓練後得到一個預訓練的模型或者是一組模型參數，繼續使用元學習測試集中的支撐集進行模型微調，最後使用元學習測試集中的查詢集進行模型測試。

元學習方法大致分為三類：基於度量(Metric-based)、基於模型(Model-based)和基於參數優化(Optimization-based)。基於度量的元學習方法是學習度量空間，然後將其比較低資源的測試樣本和豐富的訓練樣本之間距離分佈進行建模，使得同類樣本靠近，異類樣本遠離，此類別的代表方法包括孿生網路(Siamese Neural Networks)[7]、原型網路(Prototype Network)[8]、匹配網路(Matching Networks)[9]和關係網路(Relation Network)[10]。基於模型的元學習方法是使用一個新的元學習器模型，通過一些訓練示例學習參數，讓原始模型使用其參數後有更好結果，此類別的代表方法包括元學習長短期記憶(Meta Learning Long Short-term Memory)[11]。基於優化參數的元學習方法通常設計成有利於快速適應新任務的演算法，此類別的代表方法為可擴展的元學習演算法(Reptile)[12]。

孿生網路是屬於基於度量的元學習方法。孿生網路具體的運作原理是透過監督式(Supervised)的方式來訓練具有兩個輸入的類神經網路，訓練時通過不同的成對輸入樣本進行訓練，兩個輸入樣本分別對應到元學習訓練(Meta-train)的支撐集(Support Set)和查詢集(Query Set)，在最後一層對經過訓練所得到的兩個嵌入向量進行相似度計算，判斷它們之間是否屬於同一個類別，並且產生對應的機率分布。原型網路是屬於基於度量的元學習方法，它的運作原理和孿生網路很類似，兩者之間的差別在於孿生網路每次輸入到類神經網路的資料只能兩筆，而原型網路輸入到類神經網路則可以輸入兩筆以上資料。原型網路在訓練時，每個任務中的每個類別裡，樣本經過學習後得到相對應的度量空間，每個類別的嵌入會在內部類別樣本中進行均值計算，然後會得到各自代表相對應類別的一個原型來表示這個類別的嵌入，最後就是拿測試樣本經過學習後得到的度量空間，將測試的度量空間對每個類別的原型表達進行相似度計算，最後進行交叉熵計算損失來更新網路。匹配網路是屬於基於度量的元學習方法，它的運作原理和原型網路很相似，兩者之間的差別在於原型網路輸入到類神經網路的資料都是分開處理，而匹配網路則是輸入到類神經網路的資料是一起處理，因為匹配網路覺得輸入的資料彼此間互相關係。另外一個差別是匹配網路在計算出分數以後，會經過多層次(Multiple Hop)的處理，

最後再經過軟性最大化函數(Softmax Function)得到每個輸入相對應的類別機率。關係網路是屬於基於度量的元學習方法，它的運作原理和原型網路很類似，最主要的差別是關係網路先把訓練資料和測試資料抽取其各自的嵌入，接著將測試資料的嵌入接在每個訓練資料的嵌入後面，最後經過另一個新的類神經網路訓練算出其測試資料和其對應訓練資料的相似度為多少。

元學習長短期記憶(Meta Learning Long Short-term Memory)是屬於基於模型的元學習方法，它的構想是把梯度下降法用長短期記憶(Long Short-term Memory, LSTM)來取代，透過長短期記憶進行模型參數更新。這個方法目的是藉由吸取所有任務的基礎知識，以及透過少量的資料訓練，來提升模型泛化的能力和提供對模型有更好的初始參數。可擴展的元學習演算法(Reptile)是屬於基於梯度優化的元學習演算法，它的運作原理是學習類神經網路初始化參數的方法，讓類神經網路在新的目標任務上，僅使用少量新任務訓練就可以得到好的結果。可擴展的元學習演算法的運作方式是在每個任務中執行大於一次的隨機梯度下降法(Stochastic Gradient Descent, SGD)或適應性矩估計演算法(Adam)，每個任務結束後再更新參數。

(四)與模型無關的元學習演算法(Model-Agnostic Meta-Learning, MAML)

與模型無關的元學習演算法[13]是屬於基於優化參數的元學習方法，這種類型的方法通常應用在類神經網路架構的模型上，目的是希望藉由使用與模型無關的元學習演算法直接優化梯度，讓模型進行參數優化，最終學習到一組表現很好的初始參數，並且將訓練過後的模型參數當作新任務的初始參數，接續在新任務上對模型進行微調(Fine-tuning)動作，也就是將新的任務進行傳統訓練，最後得到比直接使用傳統預訓練方式有更好的結果。通常元學習的方法使用背景是在資源匱乏(Low-resource)的任務上進行訓練，然而在低資源的任務上進行模型訓練，最常遇見的問題可能會發生模型過度擬合(Overfitting)的現象，更詳細地說模型在面對訓練集有很好的成績，但面對測試集時模型就會表現得很差。所以近年來有許多新興的技術在研究如何在資源匱乏的任務上得到好成績，而元學習的優點剛好是它能在新的任務中，僅使用較低的資料進行訓練，並且經由少次的梯度下降法(Gradient Descent)進行參數更新，模型就會取得良好的泛化效果。演算法一是與模型無關的元學習演算法虛擬碼。

演算法中的 $p(\mathcal{T})$ 代表的是元學習訓練集(即 $D_{meta-train}$)中任務的分布， α 、 β 指的

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks

Require: α, β : step size hyperparameters

```
1: Randomly initialize  $\theta$ 
2: while not done do
3:   Sample batch of tasks  $\mathcal{T}_i \sim p(\mathcal{T})$ 
4:   for all  $\mathcal{T}_i$  do
5:     Evaluate  $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$  with respect to  $K$  examples
6:     Compute adapted parameters with gradient descent:  $\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$ 
7:   end for
8:   Update  $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i})$ 
9: end while
```

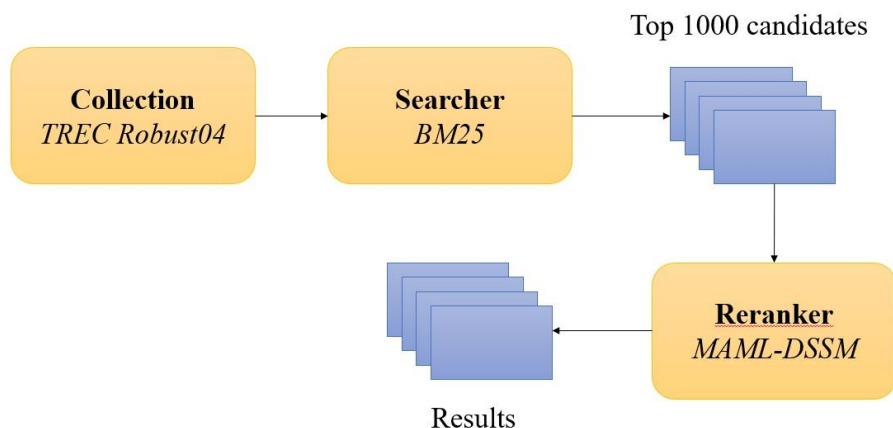
演算法一：與模型無關的元學習演算法虛擬碼。

是進行梯度下降法的學習率超參數。與模型無關的元學習演算法會先隨機初始化一組參數當作模型的初始參數，接著在每一輪迭代的過程中，先從 $p(\mathcal{T})$ 中取樣一堆任務，然後對每個任務中的支撐集計算參數梯度，並且更新當前任務中模型參數，重要的是每個任務都是從參數為 θ 的模型開始更新，做完全部任務後，會將每個任務的查詢集拿來和剛剛更新過參數的相對應模型來計算損失，然後使用得到的梯度來更新參數為 θ 的模型，最後迭代過程做完後會得到一組強而有力的參數，我們通常會把這組參數當作新任務的模型初始參數，然後進行模型微調動作。

三、使用元學習演算法於資訊檢索任務

在本研究中，我們提出的方法是把不同的查詢想像成不同的類別，而在資訊檢索中，每個查詢文件剛好會對應到幾篇和自身相關的檢索文件，這些對應到的文件可以視為這個類別下的資料，這樣的行為和機器學習中的分類問題很相似，所以我們就想把資訊檢索中的分類行為結合元學習中的與模型無關的元學習演算法；換句話說，我們使用與模型無關的元學習演算法來訓練資訊檢索模型，讓模型的參數在進行檢索任務時可以獲得更好的成效。

圖一為我們所提出方法的系統架構圖，整個流程為先使用最佳匹配模型 25 來為每個查詢進行檢索，檢索完後再利用本研究所提出的方法對前一千篇文件進行重新排序，重新排序過後的檢索結果即作為最終的輸出。



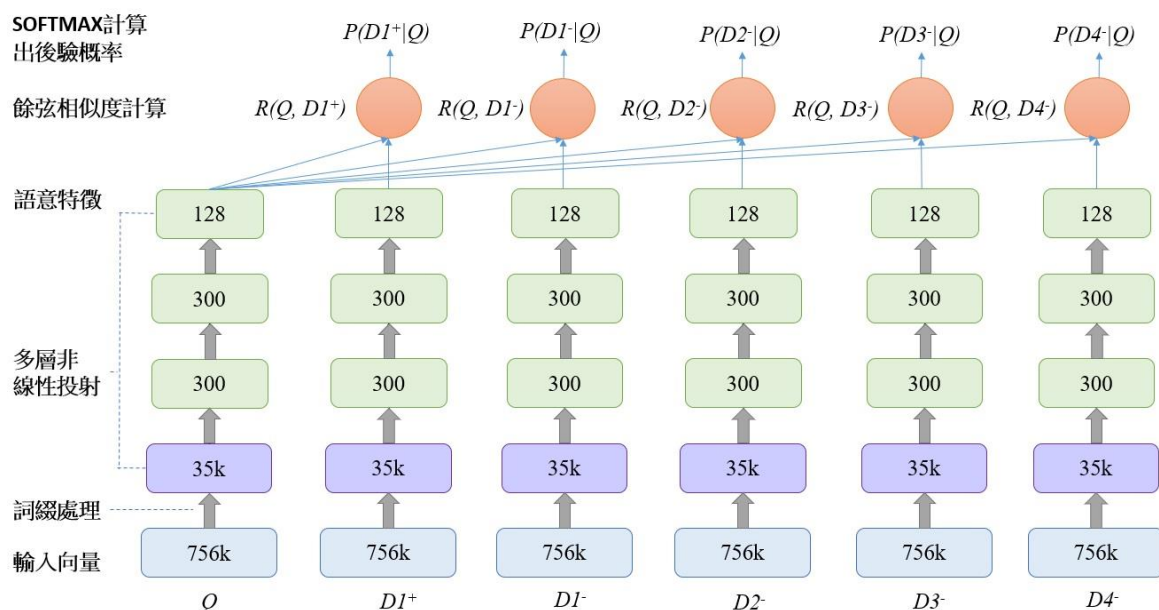
圖一、MAML-DSSM 系統架構圖。

(一)深層結構語義模型(Deep Structured Semantic Models, DSSM)

深層結構語義模型(Deep Structured Semantic Models, DSSM)[14]是基於深度神經網路的著名資訊檢索模型，它將查詢和文件投影至低維度的向量表示，再通過餘弦相似度計算查詢和檢索文件的相似度，最終訓練出可預測兩個句子之間語義相似度的低維度語義向量表達模型。更明確地，深層結構語義模型利用詞綴處理(Word Hashing)後，將查詢或文件轉換成詞頻－倒文件頻的向量表示，然後將這個向量經過三層全連接神經網路(Fully-connected Neural Network)進行降維，就可以獲得低維度的語義特徵向量。接下來，在訓練時，我們將一篇查詢和一篇正相關檢索文件以及四篇負相關檢索文件經過模型輸出後，分別獲得低維度的向量表示法，接著透過餘弦相似度計算，可以獲得這五篇文件對於這個查詢的相關分數，然後將這五個數值通過軟性最大函數(Softmax Function)，最後輸出一組機率分佈，而模型的訓練目標，即是希望正相關的檢索文件可以獲得非常接近 1 的分數，其餘四篇附相關文件則是獲得相當接近 0 的分數。圖二為深層結構語義模型的模型架構圖。

(二)結合與模型無關的元學習演算法與深層結構語義模型

本研究提出使用與模型無關的元學習演算法於訓練深層結構語義模型應用在資訊檢索的任務之中，此後我們簡稱為此一方法為 MAML-DSSM。在 MAML-DSSM 中，我們可以將訓練資料中不同的查詢想像成不同的類別，這樣的構想就和用無關的元學習演算法做分類問題的方法類似，所以每個類別的每一筆資料是由一篇查詢、一篇和查詢相關的



圖二、深層結構語義模型之模型架構圖。

檢索文件以及四篇和查詢不相關的文件組合而成。在我們的設定中，每一次訓練迭代過程會訓練 32 個任務，每個任務分別會獲取 10 篇查詢，也就是會提取 10 個類別，每個類別分別會拿 5 筆資料進行訓練和 5 筆資料進行測試，所以每訓練一個任務會一次訓練到 50 筆資料（即支撐集），然後這些資料在與模型無關的元學習演算法中的內部迴圈會進行 10 次隨機梯度下降法來更新模型參數，當完成內部迴圈後會使用另外 50 筆資料（即查詢集）計算損失。有一點要注意的是在每一次迭代過程中，做每個任務的模型初始參數都是同一組數值，所以在本實驗中，每一次迭代會藉由 32 個任務得到不同的損失值，當做完內部迴圈運算後，會用加總過的內部迴圈損失值經由適應性矩估計演算法來更新外部迴圈的模型參數，最後做完迭代次數後，我們將模型參數存下來以利在新任務上進行評估。在測試過程中，本研究的測試模型和訓練模型是相同架構，所以將經過與模型無關的元學習演算法得到的模型參數直接應用在測試模型上，查詢文件和檢索文件分別藉由測試模型得到 128 維語意特徵向量，然後去計算兩者間的餘弦相似度，最後就可以得到彼此的相似度分數。

四、實驗

(一) TREC Robust04 資料集

TREC Robust04 是由金融時報(Financial Times)、聯邦公報 94(Federal Register 94)、洛杉

磯時報(LA Times)和外國廣播資訊處(FBIS)的文件組合成一個大型資料集。該資料集包含 250 個主題文件：文件編號 301 到 450 為 TREC6 到 TREC8 資料集的主題文件，文件編號 601 到 650 為在 2003 年熱門發展的新主題文件，文件編號 651 到 700 為 2004 年新發展的主題文件，這些主題文件會取用文件內的主題(Title)語句當成實驗中所配置的查詢關鍵字。另外檢索文件則是由上述來源所提及的資料集彙整而成，共有 528,155 篇文章，並且使用文件中的文字區段(TEXT)當作實驗配置所對應的檢索文件。其中，TREC 特別設計此系列資料集，目的是用來評估資訊檢索系統面對困難查詢關鍵字的處理能力。

(二)實驗設置

我們對 TREC Robust04 資料集進行實驗，在 528,155 篇檢索文件中，先使用最佳匹配模型 25 進行初步檢索，然後對初步檢索後排名前 1,000 篇的候選文件進行重新排序。在本實驗中，我們採用 5-fold 交叉驗證，其中 4-fold 用於訓練，剩下的 1-fold 用於評估結果。並且使用平均精確度(Mean Average Precision, MAP)來進行效能評估。在資料前處理方面，我們會先將查詢文件和檢索文件做 Porter 詞幹提取處理，詞幹處理是去除詞綴得到單詞最一般寫法的過程，目的是讓整個詞庫不會很龐大。在參數設置方面，最佳匹配模型 25 的超參數我們設定 K_1 為0.8、 b 為0.75和 K_3 為1000；深層結構語義模型的批次大小設定為 4，採用適應性矩估計演算法優化器，學習率設定為 0.00001，訓練 100 個世代(Epoch)。在本研究提出的方法 MAML-DSSM 中，我們對適應性矩估計演算法使用單個梯度步長訓練模型，步長大小設定為0.00001，元批次大小為 32 個任務，並使用隨機梯度下降優化器進行十個梯度步長($\alpha = 0.001$)來評估內部模型，迭代次數設定為 1000。對於 N 類別 K 樣本的設置，我們採用每個任務會有 10 個類別，每個類別使用 5 筆資料，所以每個梯度都是由 50 筆資料來計算，並且每個類別也會有 5 筆資料來評估更新後的元梯度。以上實驗參數為多次實驗後調整所得到。

(三)實驗結果

首先在第一組實驗中，我們使用三組基礎方法來進行比較，分別是向量空間模型(VSM)和最佳匹配模型 25(BM25)，以及使用類神經網路架構的深層結構語義模型(DSSM)，實驗結果如表一所示，我們可以發現深層結構語義模型可以獲得最好的成績。接著，我們亦比較這些基礎系統與本研究所提出之 MAML-DSSM，實驗結果同樣如表一所示。經過實驗證實，我們提出的方法相對於基礎方法來說有明顯的改進，並且在評估指標中，

Model		MAP
(a)	VSM	0.1297
(b)	BM25	0.2409
(c)	DSSM	0.2502
(d)	MAML-DSSM	0.2584

表一、基礎方法與本研究提出方法之實驗結果。

Model		MAP
(a)	MAML-DSSM	0.2584
(b)	(a)+ Finetune	0.2595

表二、本研究提出方法加上微調之實驗結果。

Model	MAP	
	1-shot	5-shot
MAML-DSSM	0.2601	0.2584

表三、不同設定在 MAML-DSSM 中的實驗結果。

相對於深層結構語義模型有 3.1%的效能提升。由此可知，在資訊檢索中使用元學習方法確實能讓模型訓練的更好。接著，我們探討對模型進行微調的成效，所以我們使用經過最佳匹配模型 25 檢索後的前 10 篇文件對 MAML-DSSM 再進行微調，實驗結果如表二所示，進行微調動作有助於提升資訊檢索的表現。最後，為了分析訓練樣本多寡所帶來的差異，我們分別在 10 類別 1 樣本(10-way 1-shot)和 10 類別 5 樣本(10-way 5-shot)的設置下進行訓練，其實驗結果如表三所示。我們可以發現通常使用較多的樣本進行訓練，所帶來的結果會比較好，但是在本實驗中的結果卻是少樣本的效能比較好，這一部分值得進行更深一步探討。

五、結論

本研究提出使用元學習方法應用在資訊檢索領域上。經過實驗結果顯示，我們所提出之結合與模型無關的元學習演算法與深層結構語義模型在資訊檢索任務上，可以獲得一定的任務成效提升。值得一提的是，在本研究中，我們只使用深層結構語義模型架構的模型來進行訓練，但是基於與模型無關的元學習演算法與模型無關的特性，這種方法可以應用在更多不同類型的網路架構，所以在未來的研究中，我們會將此方法應用在資訊檢索中不同的模型上面，並使用不同的資料集來評估其有效性。

致謝

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 109-2636-E-011-007 (Young Scholar Fellowship Program), and by the Project K367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

參考文獻

- [1] A. A. Rusu *et al.*, "Meta-Learning with Latent Embedding Optimization," in *International Conference on Learning Representations*, 2018.
- [2] J. Gu, Y. Wang, Y. Chen, V. O. Li, and K. Cho, "Meta-Learning for Low-Resource Neural Machine Translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622-3631.
- [3] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844-7848: IEEE.
- [4] G. I. Winata *et al.*, "Learning fast adaptation on cross-accented speech recognition," 2020.
- [5] F. Mi, M. Huang, J. Zhang, and B. Faltings, "Meta-learning for low-resource natural language generation in task-oriented dialogue systems," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 3151-3157: AAAI Press.
- [6] O. Klejch, J. Fainberg, and P. J. a. p. a. Bell, "Learning to adapt: a meta-learning approach for speaker adaptation," 2018.
- [7] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, 2015, vol. 2: Lille.
- [8] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077-4087.
- [9] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630-3638.
- [10] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199-1208.
- [11] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.
- [12] A. Nichol, J. Achiam, and J. J. a. p. a. Schulman, "On first-order meta-learning algorithms," 2018.
- [13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of

- deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1126-1135.
- [14] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2333-2338.

基於多 BERT 模型之 NLLP 應用於建築工程訴訟之理解與預測

NLLP for the Understanding and Prediction of Construction Litigation Based on Multiple BERT Model

鍾文傑 Wen-Chieh Chung, 陳哲文 Che-Wen Chen, 王駿發 Jhing-Fa Wang

國立成功大學電機工程學系

Department of Electrical Engineering, National Cheng Kung University
n26074883@gs.ncku.edu.tw, kfcmax300@gmail.com, wangjf@mail.ncku.edu.tw

曾世邦 Shih-Pang Tseng

常州信息職業技術學院軟件與大數據學院

Software Department, Changzhou College of Information Technology, Changzhou
tsp@tajen.edu.tw

王宗松 Tzong-Song Wang

大仁科技大學數位多媒體設計系

Department of Digital Multimedia Design, Tajen University
tswang@tajen.edu.tw

摘要

本研究以深度學習之 BERT 技術提出一個工程訴訟案件篩選與歷審統計表建立及案件預測系統，並分為三個部分。第一部份是工程訴訟案件篩選，由中華民國司法院提供之判決書資料中經由基於 BERT 的模型架構篩選出屬於建築工程訴訟之案件，其準確率達到 93.55%。第二部分是案件歷審統計表建立，將案件的歷審判決書利用正則表達式進行資訊擷取並彙整成個案之歷審統計表，準確率達到 86.75%。第三部分是案件預測，利用基於多 BERT 的模型架構預測法院判決之結果，並找出相似的案例及同案件類型之統計表格，而判決預測在金額上及時間上準確率分別達到 82.22% 及 88.89%。

關鍵詞：案件篩選，資訊擷取，文本相似度，判決預測，BERT

Abstract

This research uses the multiple BERT model to propose an construction litigation case screening and summary table creation and case prediction system, which is divided into three parts. The first part is the screening of construction litigation cases. From the judgment data provided by the Judicial Court of the Republic of China, the cases belonging to the construction litigation are selected through the BERT based model structure, and the accuracy rate reaches 93.55%. The second part is summary table creation, which uses regular expressions to extract information from the judgments and integrate them into a case summary table, with an accuracy rate of 86.75%. The third part is the case prediction. The multiple BERT model framework is used to predict the outcome of court judgments, and to find similar cases and statistical tables of the same case type. The accuracy rates of judgment prediction in terms of amount and time are respectively 82.22% and 88.89%.

Keywords: Case Screening, Information Extraction, Text Similarity, Judgment Prediction, BERT

一、緒論

隨著人工智慧技術日新月異，各個領域也逐漸加入深度學習的技術以加速產業之發展，而在法律的領域也有相關應用，近年 NLP 在法律上的應用被稱為 Natural Legal Language Processing (NLLP)[1]，利用 NLP 技術有效且精確解決法律上的相關問題，解決以往需要大量基層人員花費多數工時在進行法律資訊檢索等工作。NLLP 相關應用像是基於法律文本的問答系統[2]，預測法院的投票與判決預測[3][4]，判斷案例的描述是否違反人權條款[5]和以單獨親權酌定裁判的預測模型為例[6]等。

在中華民國法院訴訟中，工程糾紛訴訟具有高度複雜性，且往往涉及之金額都非常龐大，訴訟所需時間也時常曠日廢時，而具有建築工程相關知識的法律專家更是少之又少，因此不論是法官、律師或是從事建築工程之人員都須參考過往之訴訟案件來解決紛爭。

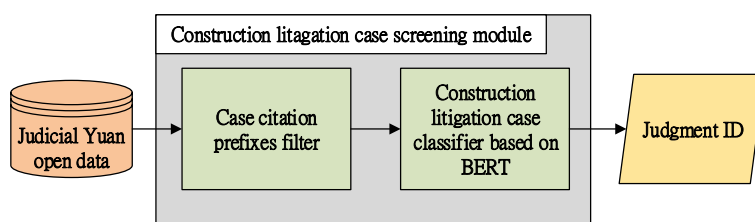
以現今中華民國司法院裁判書查詢系統提供之裁判書查詢方式僅以裁判書全文是

否出現某些關鍵字做為搜尋方法，雖能大致篩選出想查詢的相關裁判書資料，但其中可能混雜了一些與建築工程訴訟不相關的案例，且查詢到的資料對於工程法律缺乏全面性的整理與統計分析，使得在查找相關案例時必須花費大量的時間。

為了有效解決以往花費大量人力及時間在法律資訊檢索上，以及在訴訟前可以有效的評估訴訟效率，本系統將分為三個部分。第一部分是從中華民國司法院公開之裁判書資料中篩選出屬於建築工程訴訟之案件。第二部分是將屬於建築工程訴訟之案件利用正則表達式對判決書進行資料擷取並整理成案件歷審統計表格。第三部分是對新的案件利用 BERT 模型進行判決結果的預測，並找出相似之案例及同類型案例的統計表。

二、工程訴訟案件之案件篩選

案件篩選將從中華民國司法院提供之裁判書中篩選出屬於建築工程訴訟的案件，此模組會先利用案件的貫字進行第一次篩選，再透過基於 BERT 架構的網路模型進行第二次的篩選，取出其中屬於建築工程訴訟的案件。



圖一、工程訴訟案件篩選架構圖

(一) 案件貫字篩選

在利用神經網路篩選工程訴訟案例前，我們希望用簡單且快速的方法過濾掉一些案件，進而節省神經網路分類案件的時間，而從案件的貫字可以大致知道案件的類型，我們從中刪除確定不屬於工程訴訟的案件，確定不屬於工程訴訟案件的貫字如表一，此為案件貫字的連結。

表一、刪除之案件貫字

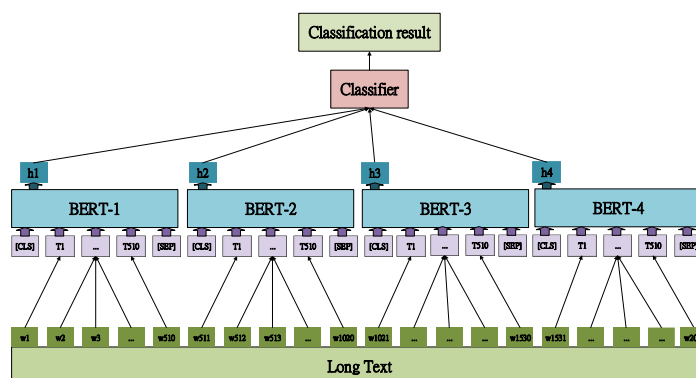
案件貫字	描述
婚	婚姻事件
家	家事訴訟事件
繼	繼承事件
醫	醫療糾紛事件

海	海商事件
國貿	國際貿易事件
金	證券、金融事件
選	選舉訴訟事件
親	親子關係事件
拍	拍賣事件
除	除權事件
智	智慧財產事件
聲	聲請、聲明事件
簡	簡易事件(金額 50 萬以下)

(二) 基於 BERT 模型建構之工程訴訟案件分類器

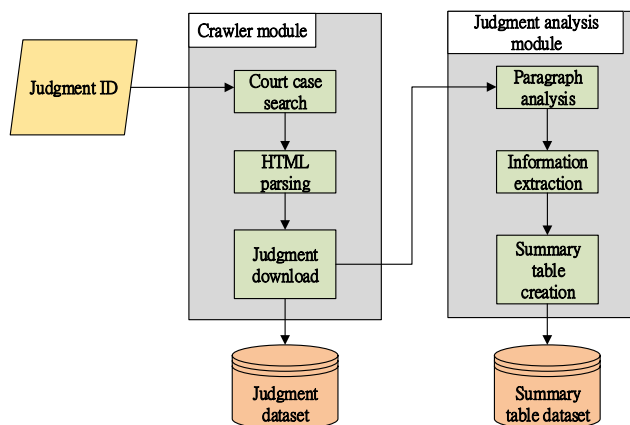
在刪除確定不屬於工程訴訟的案件後，剩餘的案件會透過由 BERT 模型[7]構建的分類器來進一步篩選，由於 BERT 模型有輸入文字長度的限制，故本系統中會使用 BERT 模型對切分成多段的輸入文本進行文本嵌入，再將多個 BERT 輸出的向量串接後進行後續的分類任務。

BERT 模型的輸入序列長度限制為 512 個字，扣除輸入序列最前面的[CLS]及最後的[SEP]後，最多可以輸入 510 個中文字，而在法院判決書中往往是數千字的文本，若由只保留文本前 512 字或文本後 512 字會損失大量的訊息，無法達到準確分類的效果。在本系統中將取出判決書中前 2040 個字，並拆分成四段後分別送入 BERT 模型中進行分類任務的模型微調，而四個 BERT 模型的參數為共享的，在模型訓練時，我們微調 BERT 最後四層的參數，其輸出的 4 個 768 維的向量串接後再進行下游的分類任務，此分類器的目的在於分類出屬於建築工程訴訟的案件與非建築工程訴訟的案件。



圖二、基於 BERT 架構之分類器

三、工程訴訟案件之歷審統計表建立



圖三、工程訴訟案件歷審統計表建立之架構圖

此部分包含兩個模組:(1)網路爬蟲模組用於自動化至中華民國司法院的裁判書查詢系統將指定的判決書下載，並儲存至判決書資料庫中。(2)判決書分析模組用於分析下載下來的判決書，其中會對判決書進行段落分析再提取各段落中重要的資訊，最後生成該案件的歷審統計表。

(一) 爬蟲模組

1. 案件判決書搜尋

案件判決書搜尋是使用 python 的 selenium 套件至中華民國司法院的裁判書查詢系統(<https://law.judicial.gov.tw/FJUD/default.aspx>)自動化鍵入判決書 ID，並透過比對法院之區域與裁判書年份及編號，以確定搜索到正確的裁判書。

2. HTML 解析

HTML 解析是利用 python 的 beautiful soup 套件將搜索到的判決書網頁進行 HTML 解析，轉換為結構化的 beautiful soup 物件，可藉由指定 HTML 標籤的方式來快速找到判決書內文。

3. 歷審判決書下載

案件的歷審裁判在個案判決書的網頁中，可以透過 HTML 解析取得歷審裁判書的連結網址再透過解析歷審判決書網頁找到判決書內文並將歷審判決書內文下載儲存至判決書資料庫。

(二) 判決書分析模組

1. 段落分析

在本系統中，建築工程訴訟的判決書內文可分為 5 個段落:(1)原、被告資訊(2)主文(3)原告主張(4)被告抗辯(5)法官判決。

- (1) 原、被告資訊描述了原告及被告雙方當事人及雙方法定代理人及訴訟代理人，主要位於「主文」二字以上的段落。
- (2) 主文部分簡要描述了法院的判決結果，包含賠償金額與訴訟費用的分配，主要位於「主文」二字以下至「事實及理由」以上的段落。
- (3) 原告主張描述了原告方在訴訟中提出的聲明、主張，主要位於「原告主張」以下至「被告抗辯」以上的段落，依據書記官的風格可能使用不同的詞彙「原告起訴主張、原告之主張、原告方面、原告之聲明及陳述」等。
- (4) 被告抗辯描述了被告方在訴訟中對於原告論述的辯駁，主要位於「被告抗辯」以下至「心證之理由」以上的段落，同樣詞彙像是「被告則以、被告之答辯、被告方面、被告之聲明及陳述」等。
- (5) 法官判決部分描述了法官的想法與做出決斷的理由，主要位於「心證之理由」以下的段落，相似詞彙如「心證理由、法院之判斷、本院之判斷、兩造之爭執、兩造不爭執」等。

2. 資訊擷取

我們將從上個段落分出的判決書的 5 個段落裡以及司法院網站中提取 12 項資訊:(1)原告當事人(2)原告訴訟代理人(3)被告當事人(4)被告訴訟代理人(5)工程標的(6)契約價金(7)原告索賠金額(8)原告訴求項目(9)歷審判決金額(10)歷審訴訟費用(11)歷審字號(12)訴訟期程。其中第 1 項到第 10 項可以從判決書中提取，第 11 項及第 12 項則要從司法院網站中提取。

本系統中用於資訊提取的方法為正則表達式，正則表達式是電腦科學的一個概念，使用單個字串來描述或匹配一系列符合制定的句法規則之字串，在很多文字編輯器中，常用正則表達式來檢索、替換符合規則的文字。

下面將逐項說明 12 項資訊提取的方法:

- (1) 原告當事人位於原、被告資訊段落，找到「原告」二字後的公司或機關，即為原告當事人。
- (2) 原告訴訟代理人位於原、被告資訊段落，找到「被告」二字前且在「訴訟代理人」後的姓名即為原告訴訟代理人。
- (3) 被告當事人位於原、被告資訊段落，找到「被告」二字後的公司或機關，即為被告當事人。
- (4) 被告訴訟代理人位於原、被告資訊段落，找到「被告」二字後且在「訴訟代理人」後的姓名即為被告訴訟代理人。
- (5) 工程標的是當事人雙方發生糾紛的工程案，通常位於原告主張段落，「系爭工程」前的工程名稱即為工程標的。
- (6) 契約價金為當事人雙方所簽訂之合約金額，通常位於原告主張段落，「契約金額、契約總價、契約價金」後的金額即為契約價金。
- (7) 原告索賠金額為原告向法院請求被告給付之金額，位於原告主張段落，「聲明」後的「被告應給付原告○○○元」即為原告索賠金額。
- (8) 原告訴求項目為原告向法院請求被告給付之各個項目，位於原告主張段落，由於各個訴求項目之金額加總為原告索賠金額，故此部分會將原告主張部分的所有金額以正則表達式找出，並在刪除大於原告索賠金額的項目後以窮舉法排列所有組合，並找出加總等於原告索賠金額之組合，而含有這些金額的句子極有可能包含原告訴求項目。
- (9) 歷審判決金額為各審法院判決之金額，位於主文段落，「被告應給付原告○○○元」即為法院判決之金額。
- (10) 歷審訴訟費用為各審原告與被告要負擔之訴訟費用，位於主文段落，「訴訟費用由○○負擔」或「訴訟費用由○○負擔○分之○，其餘由○○負擔」，訴訟費用除了要判斷由哪一方負擔外，還需要計算負擔的比例。
- (11) 歷審字號為案例的各審裁判字號，在司法院網站裁判書查詢系統中個案裁判書頁面的歷審裁判欄位，其中分為裁定與判決，但由於裁定主要用於與訴訟程序事項

相關的程序裁判，與案件事實較無關係，故只擷取判決部分。

(12) 訴訟期程為案例起訴至結案所花費的天數，在司法院網站裁判書查詢系統中個案裁判書頁面的歷審裁判欄位，當中記錄了歷審判決的日期，而起訴日期並未標明，故以第一審裁判日期往前算 180 天做為預估的起訴日期。

3. 歷審統計表建立

歷審統計表建立是將上節擷取到的資訊統整成表格，並對一些歷審的資訊數值的計算，如各審裁判費用的總和及判決比例的計算。生成的歷審統計表如圖四。

臺北地方法院 99 年建字第 號

1. 當事人

關係人	一審	二審	三審
原告	原告		
<input type="text"/> 營造股份有限公司			
律師	陳 <input type="text"/> 芳律師	陳 <input type="text"/> 芳律師	陳 <input type="text"/> 芳律師
被告	被告		
<input type="text"/> 市政府工務局衛生下水道工程處			
律師	陳 <input type="text"/> 玲律師	陳 <input type="text"/> 玲律師 陳 <input type="text"/> 安律師 洪 <input type="text"/> 彬律師	陳 <input type="text"/> 玲律師

2. 工程標的: 第七期分管網工程第十標 (區天和公園附近地區)

3. 契約價金: 1 億 1,080 萬元

4. 索賠金額: 19125239

5. 歷審字號:

(1) 第一審: 臺灣臺北地方法院 99 年度建字第 號判決

(2) 第二審: 臺灣高等法院 102 年度建上字第 號判決

(3) 第三審: 最高法院 104 年度台上字第 號判決

(a)

6. 訴訟效率

原告訴求	一審判決	二審判決	三審判決	確定判決	
				金額	比例
索賠金額合計	19125239	15736113	1271939	1271939	6.7%
原告訴訟費用	訴訟費 -168321 律師費 -80000 計 -248321	-252481 -80000 -332481	-270516 -80000 -350516	-691318 -240000 -931318	
原告訴訟所得				340621	1.8%
期程	起(上)訴日期 判決日期 合計天數	101.12.30 103.04.16 101.12.10	103.5.6 104.03.19 103.04.16		

原告訴求:

[14382911, '處原告逾期罰款 14,382,911 元', '又逾期罰款 14,382,911 元']

[4317526, '被告應給付展延工期之工程管理費 4,317,526 元']

[90000, '支出保險費 9 萬元']

[334802, '依法定利率計算為 334,802 元']

一審判決細項:

[14382911, '處原告逾期罰款 14,382,911 元', '又逾期罰款 14,382,911 元', '工期逾期 138 日應罰款 14,382,911 元', '原告依系爭契約第 9 條第 1 項第 6 款、第 12 條第 3 項約定及民法第 179 條規定請求被告返還 14,382,911 元', '原告得依系爭契約第 9 條第 1 項第 6 款、第 12 條第 3 項約定及民法第 179 條規定請求被告返還 14,382,911 元']

[1353202, '原告依系爭契約第 11 條第 5 項約定得請求展延工期之工程管理費 1,353,202 元', '另得依系爭契約第 11 條第 5 項約定請求被告給付展延工期之工程管理費 1,353,202 元']

(b)

圖四、系統生成的個案歷審統計表

原告訴求部分除了將金額找出外，同時也會擷取出包含金額的句子，減少人工找查判決書中原告訴求項目判決書的時間，如圖四(b)。

四、工程訴訟案件之案件預測

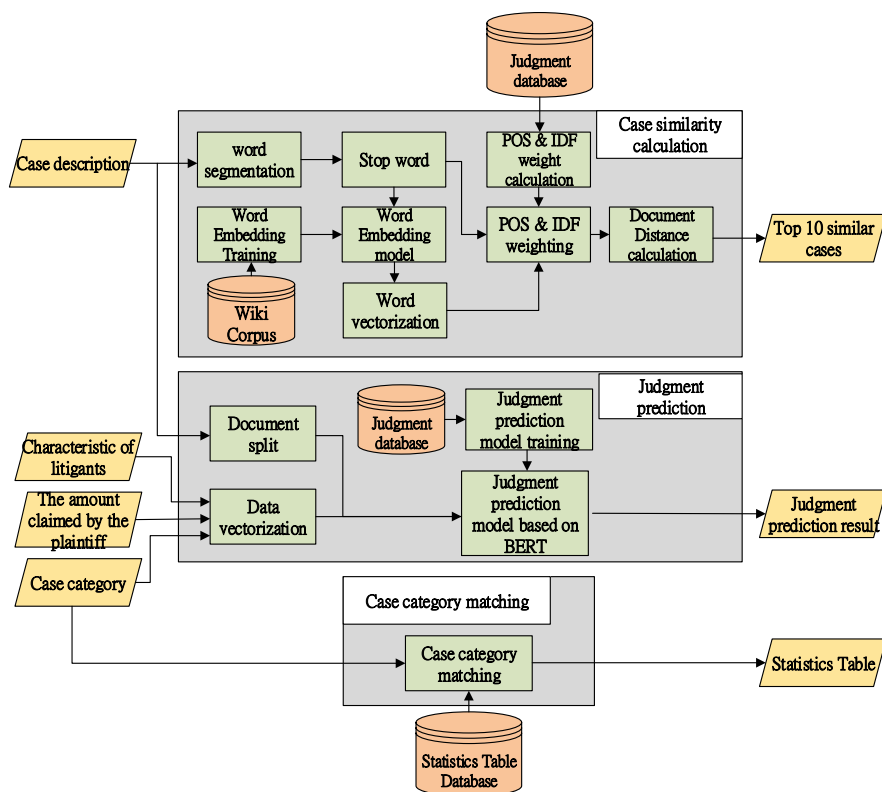
本系統預測部分分為三個功能，案例相似度計算、法院判決預測及案件類別匹配，本章節將依序介紹這三個功能，其架構圖如圖五所示。

(一) 案例相似度計算

1. 中文斷詞及刪除停用詞

中文與英文在自然語言處理中最大的差異就是英文的每個詞在句子中都以空格分開，而在處理中文文本時，往往都需要先進行斷詞的處理，將句子拆分成更小的單位，以利於在後續處理中保留句子的完整意義。

在本系統中使用的斷詞方式是 CKIP Lab 所開發的斷詞系統[8]，相較於 jieba 斷詞在中文斷詞的表現上更準確。在做完中文斷詞後，接著，我們會進行刪除停用詞，停用詞主要是頻繁會出現的詞彙且即使刪除也可以表達句子意義的詞。



圖五、工程訴訟案件預測架構圖

2. 詞嵌入

經過前面的斷詞即刪除停用詞處理後，系統仍然無法理解每個字的涵義，所以我們必須將文本轉成向量的形式。本系統在案件相似度計算部分的向量化方式使用 word2vec[9]，其方法採用 Skip-gram 和 CBOW 兩種方法來訓練模型，將詞語映射到同一座標空間，其目的是讓相似上下文的詞會產生相似的詞嵌入結果。

3. 詞性(POS)與逆向文件頻率(IDF)加權

在將文本單詞向量化後，我們使用 POS 和 IDF 對向量進行加權。詞性標註的結果分為名詞、動詞、形容詞、副詞及其他。詞性的權重列於表二，我們起初設定名詞、動詞的權重為 5，其他為 1，再經過實驗調整後，得到形容詞、副詞的權重為 4 時會有最好的結果。

表二、詞性權重

詞性	權重
名詞、動詞	5
形容詞、副詞	4
其他	1

IDF 權重用於衡量單詞在文本普遍重要性的度量，其計算式如式(1)， idf_i 為詞語 t_i 的 IDF 權重， $|D|$ 為語料庫中的檔案總數， j 為包含詞語 t_i 的檔案數目，其中分母加一為避免除以零的情況。

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}| + 1} \quad (1)$$

結合 POS 和 IDF 權重的文本向量化表示式如式(2)。其中 w_i 表示字詞經 Word2vec 向量化結果， Pos_weight 為根據每個詞的詞性所賦予的權重， Pos_{w_i} 則是字詞經由 POS tagging 轉換的詞性，將文本內單詞分別經過 POS 和 IDF 的加權後相加，使每個文本都可以用 300 維的向量表示。

$$V = \sum_i w_i \times \log \frac{|D|}{|\{j: t_i \in d_j\}| + 1} \times Pos_weight(Pos_{w_i}) \quad (2)$$

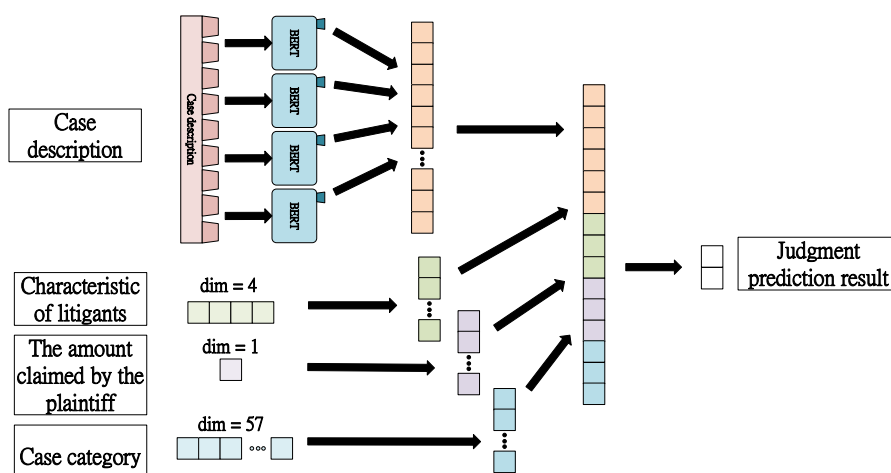
4. 相似度計算

在本系統中我們使用餘弦相似度來衡量文本之間的相似程度。餘弦相似度式通過測量兩向量夾角的餘弦值來度量他們之間的相似度。

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (3)$$

(二) 法院判決預測

判決預測的架構圖如圖六。判決預測模型的輸入包括(1)案件描述，即判決書中原告主張部分。(2)雙方單位性質，即公部門或私部門。(3)索賠金額。(4)案件類別，本系統中案件類別共 57 類。



圖六、判決預測模型架構圖

案件描述將取前 2040 個字做為輸入，將 2040 個字拆分成 4 個段落，每段最多 510 個字，並利用 BERT 模型進行嵌入，將 BERT 模型得到的 4 個 768 維向量串接。雙方部門性質因為雙方都有可能是公部門或是私部門，因此共有 4 種組合，並以 one-hot 的形式表示，故其維度為 4。索賠金額由於數值過大，我們會將輸入的索賠金額除以資料庫案件中最大的索賠金額。案件類別共有 57 類，以 one-hot 表示為 57 維向量做為輸入之一。

除了案件描述的其他三個輸入會經由全連接層提取 64 維特徵向量，並與 BERT 模型得到的向量串接再進行分類，兩個輸出維度為訴訟期間是否超過三年及金額上的勝敗訴。

(三) 案件類型匹配

案件類型匹配是將輸入系統的案件類型與 57 個案件類型做匹配，並輸出由建築工程訴訟方面的法律專家依照案件類型進行統計的統計表格。

五、實驗結果

(一) 案件篩選實驗

1. 案件篩選之資料庫

用於案件篩選的資料是從司法院裁判書開放資料網站(<http://data.judicial.gov.tw/>)所下載之 2000 年至 2018 年的裁判書，其資料庫內容如表 1，我們在 6,398,460 件民事案件中挑選了 1,231 件案件進行工程訴訟案件的標記，其中有 828 件為工程訴訟案件，403

件不是工程訴訟案件，而訓練集與測試集的分切比例為 9:1。

表三、案件篩選資料庫內容範例

項目	內容
裁判 ID	TPDV,88,訴,384,20000222
裁判年分	88
裁判貫字	訴
裁判編號	384
裁判日期	20000222
裁判案由	給付工程款
判決書內文	臺灣臺北地方法院民事判決 八十八度訴字第三八四號…… 事實 甲、原告方面：壹、聲明：除假執行擔保金額外，如主文所示。……

2. 案件篩選之實驗結果

我們使用第二章提到的模型進行工程訴訟案件的分類，並與其他方法進行比較，其中 keyword 是利用文本中出現特定關鍵字即將案件列為工程訴訟之案件，Word2vec+CNN 則是利用 Word2vec 將斷詞後的裁判書文本向量化並利用 CNN 作為分類模型，而 BERT 與 BERT-fine tune 則是未進行微調與進行微調的模型。結果如表四。

表四、案件篩選實驗結果比較表

方法	準確率
Keyword	62.67%
Word2vec+CNN	84.55%
BERT	91.86%
BERT-fine tune	93.55%

(二) 資訊擷取實驗

在資訊擷取實驗中，我們將系統產生出的 100 個案例的 word 檔與人工整理的 word 內容做比較，比較的內容包含(1)原告當事人(2)原告訴訟代理人(3)被告當事人(4)被告訴訟代理人(5)工程標的(6)契約價金(7)原告索賠金額(8)原告訴求項目(9)歷審判決金額(10)歷審訴訟費用(11)歷審字號(12)訴訟期程等 12 個項目。，正確率計算公式如式(4)。計算其每個案件的準確率後，取 100 件之平均準確率為 86.75%。

$$Accuracy = \frac{Wrong\ number\ of\ items}{Number\ of\ items} \times 100\% \quad (4)$$

(三) 案件相似度計算實驗

1. 案件相似度計算之資料庫

案例相似度計算的資料庫是由建築工程背景之法律專家所收集並整理 899 件工程訴訟案件，並針對個案分析其類別、雙方當事人、工程標的、索賠金額等資訊及判決書內文，本系統之案件相似度計算模組主要使用個案判決書中原告主張部分來比較案件之間的相似程度，並針對個案的案件類別進行各個方法的評比。

2. 案件相似度計算之實驗結果

將 899 件案件依次作為輸入案件計算案件相似度，並將每次的前 10 個最相似的案件之案件類別與輸入的案件類別比對，若輸出之案件之案件類別至少與輸入之案件之案件類別有一項相同，表示兩案之間有一定相關性，在前 10 個最相似的案件計算相關的案件數，並平均 899 件案件的相關案件數，其平均相關案件數如表五所示。

表五、案件相似度計算實驗結果比較表

方法	Average precision in TOP 10(AP@10)
TF-IDF	0.761
Word2vec	0.835
POS tagging + Word2vec	0.902
IDF + Word2vec	0.894
TF-IDF + Word2vec	0.866
POS tagging + IDF + Word2vec	0.911

(四) 法院判決預測

1. 判決預測之資料庫

判決預測使用的資料庫與案例相似度計算的資料庫相同，而在判決預測時要用到的部分是原告主張、雙方部門性質、原告索賠金額、案件類別，而訓練集與測試集的比例為 9:1。

2. 判決預測實驗結果

判決預測分為時間上勝敗訴及金額上勝敗訴，時間上勝敗訴以訴訟期程 3 年做區分，訴訟期程在 3 年以下為時間上勝訴，3 年以上為時間上敗訴，金額上勝敗訴則以訴訟所得佔據原告索賠金額之百分比為勝敗訴之依據，0%~25%為慘敗、25%~50%為小敗、

50%~75%為小勝、75%~100%為大勝，又可依 50%為分界，50%以上為勝訴，50%以下為敗訴，實驗結果如表六。

表六、判決預測實驗結果比較表

	方法	準確率
時間上	Word2vec+CNN	84.1%
	BERT	86.67%
	BERT-fine tune	88.89%
金額上(4 分類)	Word2vec+CNN	58.88%
	BERT	60.67%
	BERT-fine tune	64.44%
金額上(2 分類)	Word2vec+CNN	73.33%
	BERT	74.45%
	BERT-fine tune	82.22%

六、結論

本系統利用案件篩選及案件統計表建立兩個部分來輔助法律專家蒐集資料，解決了以往在法律資訊檢索上需花費大量人力及時間的問題。判決預測部分讓當事人或律師在訴訟前可以有效的評估訴訟效率，在評估是否進行法律訴訟。並可以針對相似的案例進行研究也有助於訴訟效率的提升。

在我們的系統中，先用貫字進行第一次的案件篩選，接著使用 BERT 模型來準確的分類屬於建築訴訟的案件，以避免神經網路所帶來的巨大運算時間以及傳統方法的低準確率問題，我們的分類器達到了 93.55%的準確率。在統計表建立部分使用正則表達式快速地對判決書進行分析，找出其重點部分並整理成表格，相較於以往人工整理表格大約可減少 6 倍以上的時間且準確率達到 86.75%，與本研究合作的建築師事務所之負責人表示，此系統預估可減少 30%的人力於此法律資訊檢索工作，節省花費的時間成本約每年 50 萬元。判決預測部分使用 BERT 模型對新的案件進行判決的預測，針對金額與時間進行分析，當事人可自行評估是否有訴訟的價值，而我們的模型在金額上及時間上分別得到 82.22%與 88.89%的準確率。此外，透過 POS 及 IDF 加權的詞向量進行相似案例的計算，則可讓律師對相似的案例深入研究以提高訴訟的勝率。

參考文獻

- [1] Aletras, Nikolaos, et al. "Proceedings of the Natural Legal Language Processing Workshop 2019." *Proceedings of the Natural Legal Language Processing Workshop 2019*. 2019.
- [2] Do, Phong-Khac, et al. "Legal question answering using ranking SVM and deep convolutional neural network." *arXiv preprint arXiv:1703.05320* (2017).
- [3] Katz, Daniel Martin, Michael J. Bommarito, and Josh Blackman II. "A general approach for predicting the behavior of the Supreme Court of the United States." *PloS one* 12.4 (2017).
- [4] Virtucio, Michael Benedict L., et al. "Predicting decisions of the philippine supreme court using natural language processing and machine learning." 2018 *IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 2. IEEE, 2018.
- [5] Aletras, Nikolaos, et al. "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective." *PeerJ Computer Science* 2 (2016): e93.
- [6] 黃詩淳, and 邵軒磊. "人工智慧與法律資料分析之方法與應用: 以單獨親權酌定裁判的預測模型為例." *臺大法學論叢* 48.4 (2019): 2023-2073.
- [7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [8] Ma, Wei-Yun, and Keh-Jiann Chen. "Design of CKIP Chinese word segmentation system." *Chinese and Oriental Languages Information Processing Society* 14.3 (2005): 235-249.
- [9] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

單語者台灣腔中文即時語音合成系統

Real-Time Single-Speaker Taiwanese-Accented Mandarin Speech Synthesis System

王奕雯 Yih-Wen Wang

國立中山大學資訊工程學系

Department of Computer Science and Information Engineering

National Sun Yat-sen University

M083040011@student.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

本論文研究單語者台灣腔中文即時語音合成系統，架構上採用文字序列端到梅爾頻譜圖序列端的合成器，再串接一個從梅爾頻譜圖到語音訊號的聲碼器。首先，我們嘗試使用 GST Tacotron-2 合成器串接 Griffin-Lim 聲碼器，搭配不同的資料集，包括北京腔中文語料與台灣腔中文語料等等，以及不同的訓練方式，包括遷移式學習(Transfer Learning)與集成式學習(Ensemble Learning)等等，進行了三種系統設定實驗。接著我們使用 Tacotron-2 串接 Griffin-Lim 架構與中文語料，實驗是否使用預訓練模型(Pretrained Model)，再進行了兩種系統設定實驗。最後，我們從上述五種系統設定中挑選出 MOS 最高者，再將其聲碼器從 Griffin-Lim 替換成 WaveGlow，評估兩種聲碼器對 MOS 的影響。我們使用的資料集包含單人中文 12 小時的標貝語料、單人中文 4.5 小時的個人錄製語料、單人中文 2.2 小時的教育廣播電台語料，以及單人英文 24 小時的 LJSpeech 語料。最終 MOS 最高的單語者台灣腔中文即時語音合成系統為，使用標貝語料預訓練、再使用教育廣播電台語料接續訓練的 Tacotron-2 模型，並串接使用 LJSpeech 語料預訓

練、再使用標貝語料接續訓練的 WaveGlow 模型，MOS 評分可達 4.32，且該語音合成系統產生 10 秒 48kHz 的語音只須 1.3 秒，因此為即時語音合成系統。

Abstract

In this paper, we study a real-time single-speaker Taiwanese-accented Mandarin speech synthesis system. This system uses an end-to-end sequence-to-sequence model from the text sequence to the Mel spectrogram sequence, and a vocoder to map the Mel spectrogram sequence to synthesized speech waveform. We first use the GST Tacotron-2 sequence-to-sequence model and the Griffin-Lim vocoder. The system is trained with several datasets, such as Mainland-accented Mandarin corpus and Taiwanese-accented Mandarin corpus, and with different training methods including transfer learning and ensemble learning. In this stage, three experiments were carried out. In addition, we use Tacotron-2 and Griffin-Lim with the same data sets and experimented with using model pretraining. In this stage, two experiments were carried out. Finally, the system setting with the highest MOS in the experiments is selected, and the Griffin-Lim vocoder is replaced by WaveGlow vocoder. The datasets we use include 12-hour Biaobei Mandarin corpus, 4.5-hour personal recording Mandarin corpus, 2.2-hour National Education Radio Mandarin corpus, and 24-hour LJSpeech English. At the end of day, the Real-Time Single-Speaker Taiwanese-Accented Mandarin Speech Synthesis System with the highest MOS we achieved is the system as follows: Tacotron-2 is pretrained with the Biaobei corpus, and then trained with the National Education Radio corpus, and the WaveGlow vocoder is pretrained with the LJSpeech corpus, and then trained with the Biaobei corpus. This system achieves the MOS score of 4.32 and generates 10 seconds of 48kHz speech in 1.3 seconds.

關鍵詞：Tacotron-2、GST Tacotron-2、Griffin-Lim、WaveGlow、Transfer Learning、Ensemble Learning、Pretrained Model

一、緒論

大數據時代的到來，深度學習成為熱門議題之一，人機互動的情況也早已普及，像是數位助理、智能導航、以及有聲書等等。在這些廣泛的應用當中，語音合成的技術就扮演了相當重要的角色。雖然語音合成的產品眾多，且能產生中文語音的技術也相當成熟，但合成的中文語音其實大多數為「北京腔調的中文語音」，會形成此結果的主要原因是因為可大量取得的中文語料，多為北京腔調的語者錄製而成。因此該研究希望透過神經網路的語音合成技術，利用端到端直接學習從文本到聲學特徵的對應關係，並且使用有限的中文語料，搭配不同的訓練方式，達到「台灣腔調的中文語音合成」。

一個完整的語音合成系統需要合成器與聲碼器，在神經網路的訓練過程中，合成器使用可訓練的神經網路 Tacotron-2[3]、GST Tacotron-2[4]，其輸入為成對的文字與音檔，透過端到端的神經網路，一次輸出一幀的梅爾頻譜圖(Mel-Spectrogram)；聲碼器則使用了演算法 Griffin-Lim[1]以及可訓練的神經網路 WaveGlow[2]，在演算法 Griffin-Lim 中，輸入是梅爾頻譜圖中幅度譜的資訊，透過六十次迭代演算輸出時域波形；而在神經網路 WaveGlow 中，輸入為成對的音檔與梅爾頻譜圖，透過多個可逆的變換函數組成序列，最後輸出時域波形。訓練完成後，於推斷時只要輸入一段欲合成的文字，透過合成器輸出梅爾頻譜圖，最後由聲碼器輸出語音訊號。本文分為四個部分：第一部分為緒論；第二部分為研究方法，介紹資料集的使用、資料前處理、合成器與聲碼器的模型架構、不同模型架構與不同資料搭配不同的訓練方式；第三部分為實驗結果的分析與評估；第四部分為結論。

二、研究方法

(一) 資料集

1. 標貝資料集

標貝資料集是由「標貝(北京)科技有限公司」於 2018 年所開放。由一位中年女性錄音者錄製而成，時長共 12 小時，採樣頻率為 48KHz、16-bit，錄製環境為專業錄音棚環境，語料內容包含各類新聞、小說、科技等領域，詳細的資料規格如表一。後續內容提及時，將該資料集簡稱為「Biaobei 資料集」。

2. 客製化資料集

該資料集為我個人製作。會製作此資料集的主要原因是希望合成的中文語音能具備台

表一、Biaobei 資料集數據規格

數據內容	中文標準女聲語音庫數據
語音類型	標準普通話(北京腔調)
錄音者	單一語者，中年女性
錄音環境	專業錄音棚環境，無背噪
錄音工具	專業錄音設備
有效時長	共約 12 小時，10000 個 wav，3~5s/wav
採樣格式	無壓縮 wav 格式，採樣率為 48KHz、16-bit
標註格式	文本標註為.txt 格式；音節音素標註為.interval

表二、YW 資料集數據規格

數據內容	客製化中文女聲語音庫數據
語音類型	標準國語(台灣腔調)
錄音者	單一語者，青年女性
錄音環境	研究室，略有背噪
錄音工具	個人筆電麥克風
有效時長	共約 4.5 小時，3800 個 wav，3~5s/wav
採樣格式	無壓縮 wav 格式，採樣率為 48KHz、16-bit
標註格式	文本標註為.txt 格式

灣腔調，而目前開放免費使用的單一語者中文語料大多數為北京腔調。整份資料集的製作由我進行錄音，錄製的文本內容為 **Biaobei** 資料集的文本，錄製的環境為安靜無人的研究室，詳細的資料規格如表二。後續內容提及時，將該資料集簡稱為「YW 資料集」。

3. **NER-Trs-Voll** 資料集

此資料集全名為「北科大教育電台廣播節目語音語料庫」，主要是大量轉寫教育電台節目，產生節目音檔逐字稿，並作人工校正與切割，形成長度約 30 秒的音檔，以建置大規模台灣腔語料庫。但該資料集的錄製者為多語者，將會導致合成的語音其語者具有隨機性、非單一性，並不符合這次的單語者語音合成目標，此外，該資料集的每個音檔時長過長，使得訓練難度提升。因此，決定從多語者中，提取出單一語者的音檔，並將每個 30 秒的音檔人工切割成約 10 秒的音檔，詳細的資料規格如表三。後續內容提及時，將該資料集簡稱為「NER-2hr 資料集」。

表三、NER-2hr 資料集數據規格

數據內容	NER 中文女聲語音庫數據
語音類型	標準國語(台灣腔調)
錄音者	單一語者，中年女性
錄音環境	錄音室內或錄音室以外之場所，略有背噪
錄音工具	教育電台節目錄製
有效時長	共約 2.2 小時，1495 個 wav，5~8s/wav
採樣格式	無壓縮 wav 格式，採樣率為 16KHz、16-bit
標註格式	文本標註為.txt 格式

表四、LJSpeech 資料集數據規格

數據內容	LJSpeech 英文女聲語音庫數據
語音類型	美式英文
錄音者	單一語者，中年女性
有效時長	共約 24 小時，13100 個 wav，1~10s/wav
採樣格式	無壓縮 wav 格式，採樣率為 22050Hz、16-bit
標註格式	文本標註為.csv 格式

4. LJSpeech 資料集

這是一個公開的英文語音數據集，文本在 1884~1964 年之間出版，音檔由「LibriVox」於 2016 年至 2017 年錄製。總共 13,100 個音頻，每個音頻平均長度 1~10 秒不等，總時長約 24 小時，錄製者為同一女性，內容來自七部非小說類書籍，詳細的資料規格如表四。後續內容提及時，將該資料集簡稱為「LJSpeech 資料集」。

(二) 資料前處理

1. 文字

由於中文字本身有數萬個相異字，加上有許多同音異字的情況，這導致無法以窮舉的方式對神經網路進行訓練。為了解決此問題，我們使用漢語拼音作為文本的輸入，並且加上數字 1~5 來表示聲調。此外，我們也對文本進行斷詞，進而提升合成中文語音的流暢度。在訓練完成後進行合成時，首先先對輸入的文本經過 jieba 套件進行斷詞，再透過 pypinyin 套件形成漢語拼音，進行後續的語音合成，整體流程可參考表五。

2. 語音訊號

在進到神經網路訓練前，會將語音訊號進行前處理，生成「梅爾頻譜圖」作為輸入。前處理的部分是使用幀大小為 50 毫秒、幀移為 12.5 毫秒，以及漢明窗(Hanning Window)進行計算，然後通過短時傅立葉轉換(STFT)得到線性頻譜。接著使用頻率範圍在 125Hz~7.6kHz、通道數為 80 的梅爾濾波器組，對 STFT 的線性頻率進行過濾，再對函數進行壓縮，從而把 STFT 幅度轉換到梅爾刻度上。

表五、文本資料前處理流程

原始文本	不好意思，我找不到我想要的書。
經過 jieba 進行斷詞	不好意思 ， 我 找 不到 我 想要 的 書 。
經過 pypinyin 進行漢語拼音	bu4 hao3 yi4 si1 ， wo3 zhao3 bu2 dao4 wo3 xiang3 yao4 de shu1 。

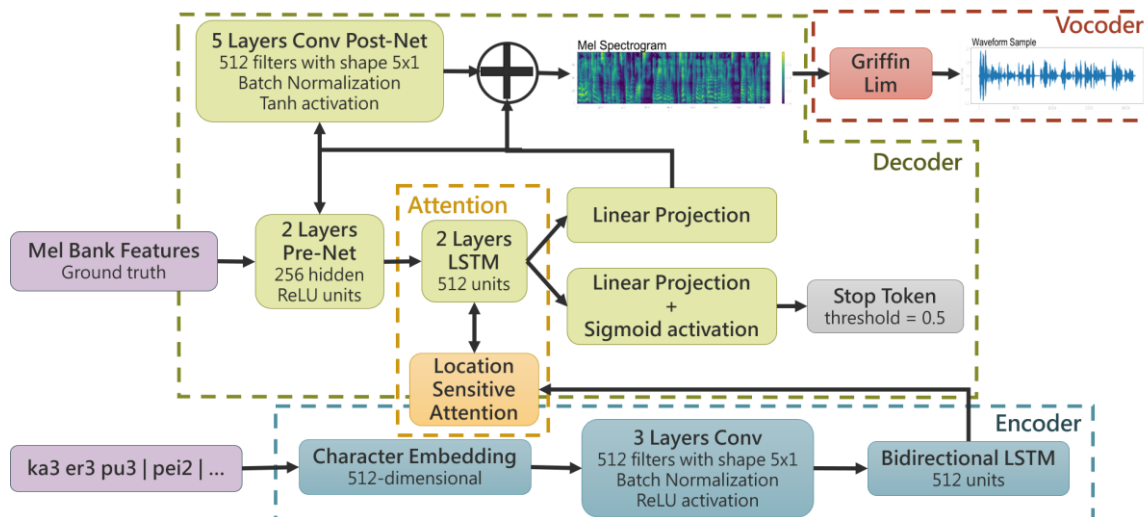
(三) 合成器

1. Tacotron-2

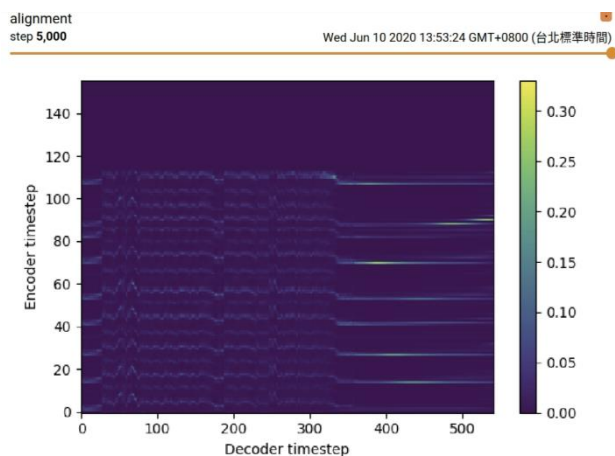
Tacotron-2 是由 Google Brain 於 2018 年提出來的一個語音合成框架[3]，模型架構如圖一，主要由三個部分組成，分別為編碼器(Encoder)、位置敏感的注意力機制(Location Sensitive Attention)與解碼器(Decoder)。編碼器將輸入的字元編碼成 512 維的字元向量(Character Embedding)，透過三層卷積(Convolution Neural Network)獲取序列中的上下文訊息、雙向長短期記憶(Bi-LSTM)將文本編碼成一個固定向量。接著透過位置敏感的注意力機制，給予解碼器在不同時間步有不同的權重。梅爾頻譜圖作為解碼器中兩層全連接預處理網路(Pre-Net)的輸入，預處理網路的輸出會與上一個時間步的上下文向量拼接送入兩層單向的長短期記憶(LSTM)，長短期記憶的輸出被用作計算本時間步的上下文向量，並且經過線性投映(Linear Projection)後，分別用來預測線性頻譜圖，每一次預測一幀，也用來計算停止符機率(Stop Token)。為了提取更高維的特徵，線性頻譜圖會經過五層的卷積後處理網路(Post-Net)來預測一個殘差，疊加到未經後處理網路的線性頻譜圖，形成梅爾頻譜圖。而停止符機制的運作原理，是將經過線性投影的結果由 Sigmoid Activation 去預測輸出的頻譜序列是否完成的一個機率，當機率大於 0.5 時，頻譜圖的生成即停止。

在經過 Post-Net 之前，會將經過線性投影預測出的線性頻譜圖與真實頻譜圖計算一個損失；在經過 Post-Net 之後，會將經過殘差疊加後產生的梅爾頻譜圖與真實頻譜圖也計算一個損失。這兩項計算原先皆是使用 MSE Loss Function，而我們將其改成 Huber Loss Function 並比較兩者的結果。Huber Loss Function 主要是結合 MSE 與 MAE 的優點。MSE 的優點是收斂較快，因為它的梯度是隨著損失值在改變，但缺點是遇到離群值時，經過平方後計算的損失值會較大，對模型造成不好的影響。而 MAE 的優點則是對離群值較有魯棒性，損失值較低，但缺點是收斂速度慢，因為其梯度始終為 1，也因此容易錯過損失值最低的點。整個公式如(1)、(2)，使用一個超參數 δ 來控制要側重 MSE 或是 MAE，當誤差值小於 δ 時，使用 MSE，使其收斂快速；當誤差值大於 δ 時，使用 MAE，避免離群值造成較大的損失值。圖二、三分別為使用 MSE 與 Huber Loss 在訓練過程中編碼與解碼的對齊圖，可以發現同樣在步數(Step)為 5K 時，圖三已有較明顯的對齊圖，圖二則尚未。修改成 Huber Loss 後可以使訓練收斂更加快速，並且損失值越低，代表預測出的頻譜圖與真實的頻譜圖越接近，語音合成的效能也進而提升。而後續提及 Tacotron-2 的模型架構皆為修改成此損失函數的模型架構。

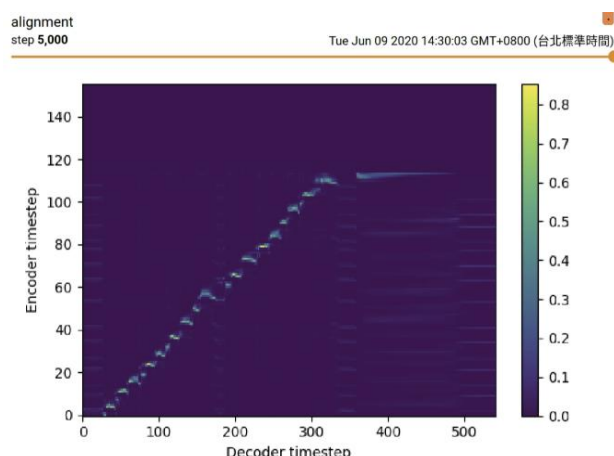
$$\text{Huber Loss}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & |y - \hat{y}| \leq \delta \quad (1) \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2, & |y - \hat{y}| > \delta \quad (2) \end{cases}$$



圖一、Tacotron-2 模型架構



圖二、Mean Square Error Loss Function



圖三、Huber Loss Function

2. GST Tacotron-2

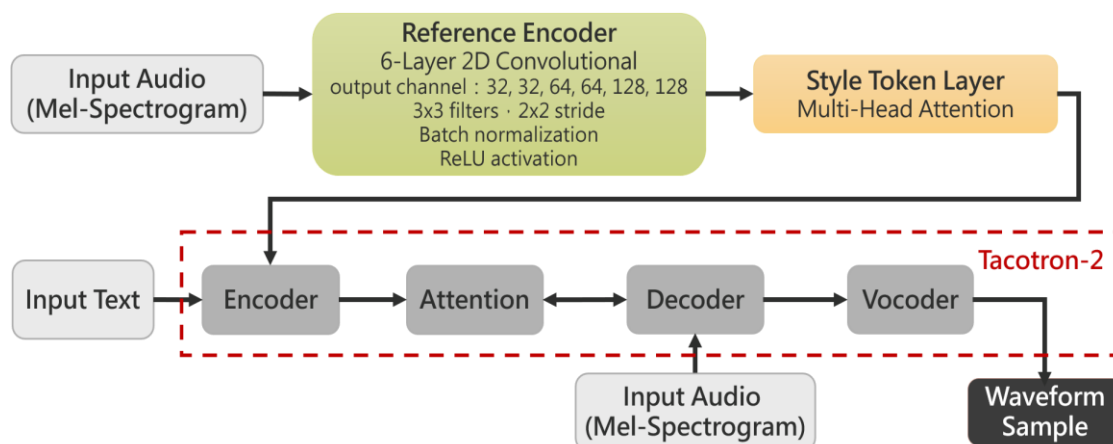
為了合成高自然度的人聲，語音合成系統必須學會對韻律建模，韻律包含語音的所有表現力因素，例如語調、節奏和重音。透過向 Tacotron-2 多增加一個注意力機制，它將語音片段中的韻律嵌入，並表達成一組基礎嵌入固定集合的線性組合，這種嵌入方式稱為，並且這些韻律無須事先標記，因此是屬於無監督式學習。基於 Tacotron-2 往上增加的模塊主要為參考編碼器(Reference Encoder)，與風格標記層(Style Token Layer)，如圖四。參考編碼器的輸入為梅爾頻譜圖，先通過六層二維的卷積 (Convolution Neural Network)，再接一層單向門控循環單元(GRU)，將聲音編碼成一個

512 維的固定向量，稱為參考特徵(Reference Embedding)。風格標記層主要是由多頭注意力機制(Multi-Head Attention)構成，也就是作多次的自注意力機制(Self-Attention)，這裡設定作 8 次。模型隨機初始化一組風格特徵(Style Token Embedding)集合，而我們使用的訓練資料其風格多樣性並未很高，因此設定風格數為 4。在自注意力機制中，Q(query)為 512 維的參考特徵，K(key)與 V(value)皆為 64 維的風格特徵。運作流程是先將 512 維的參考特徵切割成 8 個 64 維的參考特徵，每一個參考特徵皆與 4 個風格特徵作點積(Dot)，進行相似度計算得到權重，將權重透過 Softmax Activation 使得權重落在 0~1 之間，接著將權重與 4 個風格特徵進行加權，得到一個 64 維的風格特徵，如公式(3)。總共進行 8 次上述的自注意力機制，最後將 8 個 64 維的風格特徵作拼接(Concat)，形成一個 512 維的風格特徵，如公式(4)。經過 GST 後得到的風格特徵，會與 Tacotron-2 中編碼器的輸出向量作拼接，進行後續解碼的動作。

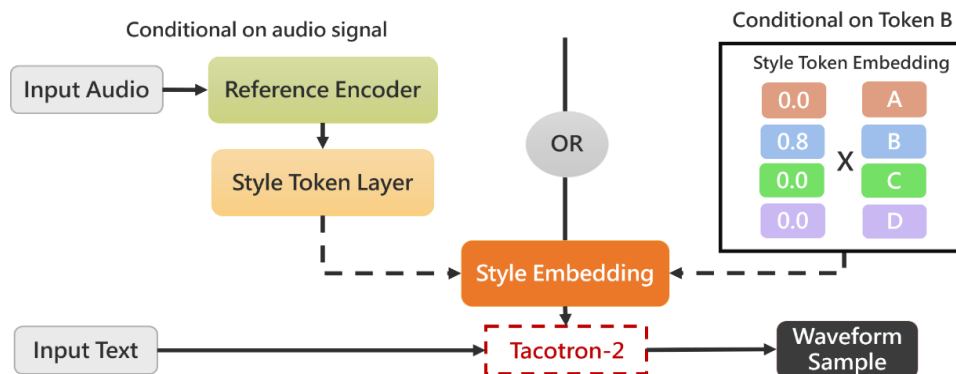
訓練完成後，推斷方法有兩種，如圖五。第一種是輸入欲合成的文本、欲合成此風格的音檔，透過參考編碼器與風格標記層得到風格特徵；第二種方法是輸入欲合成的文本，並給予一組指定風格特徵進行加權的權重，直接得到需要的風格特徵，此方法還能得知每種風格的資訊。因為在訓練過程中，模型隨機初始化一組風格特徵，我們只知道風格數量，但不會得知每種風格可能代表的是語速快慢，或是音調高低等等，那麼可以透過只給特定風格大於 0 的權重，其餘皆為 0，藉此可知該風格的資訊為何。得知每種風格的資訊後，可以更自由的進行線性組合，計算出想要的風格特徵，得到更多樣風格的語音合成。而在接續的實驗中，兩種推斷方式我們皆會嘗試，並從中比較出哪一種推斷方式以及風格參數設定，可以得到最好的合成效果。

$$Head_i = \text{Self-Attention}(QW_i^Q, KW_i^K, VW_i^V) = \sum \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

$$\text{Multi-Head Attention} = \text{Concat}(Head_1, Head_2, \dots, Head_i)W^{output} \quad (4)$$



圖四、GST Tacotron-2 模型架構



圖五、GST Tacotron-2 推斷流程

(四) 聲碼器

1. Griffin-Lim

這是一種由演算法來合成語音的聲碼器[1]，這種聲碼器不需要事先訓練，它的輸入為合成器輸出的梅爾頻譜圖。Griffin-Lim 重建語音訊號時，需要使用幅度譜與相位譜，但在梅爾頻譜圖中是不包含相位訊息的。於是演算法第一步驟是用噪聲隨機初始化一個相位譜，第二步驟是將已知的幅度譜與初始化的相位譜經過逆傅立葉轉換(ISTFT)，得到一個初步的時域訊號。接著第三步驟，是將上一步得到的時域訊號經過傅立葉轉換(STFT)，得到新的幅度譜與新的相位譜，此時是從一個不準確的時域訊號得到幅度譜與相位譜，於是第四步驟用原先已知的幅度譜取代新的幅度譜，接著與新的相位譜再透過逆傅立葉轉換，得到一個更準確的時域訊號。如此重複上述步驟二到四，直至迭代出一個穩定的時域訊號，這裡設定的迭代次數為六十次。此演算法的整個流程可以參考表六。

2. WaveGlow

WaveGlow 是由 NVIDIA 研究小組於 2018 年提出，它是一個基於流的生成模型，透過分佈採樣生成語音，只需一個神經網路與一個最小化負對數似然(negative log likelihood)的損失函數，即可生成時域波形。主要是由多個可逆的轉換函數組成序列，

表六、Griffin-Lim 演算步驟

Step1：隨機初始化一個相位譜
Step2：此相位譜與已知的幅度譜經過 ISTFT，合成新的語音
Step3：對合成的語音作 STFT，得到新的相位譜與新的幅度譜
Step4：丟棄新的幅度譜，用已知的幅度譜與新的相位譜，再次合成新的語音
Step5：重複 Step2、3、4，直到迭代次數達到六十次，即得最終的合成語音

將一個簡單的分佈轉換到一個複雜的分佈，藉此模擬訓練數據的分佈，最後再透過最小化負對數似然值，進行優化。圖六為整個網路架構，首先將 8 個聲音採樣值拼接成一個向量，此動作稱為 squeeze，接著通過 12 層的 1*1 可逆卷積(Invertible Convolution)與仿射耦合層(Affine Coupling Layer)。在仿射耦合層中，只會將 x 一半的通道數 x_a 作為輸入，並與梅爾頻譜圖進入 WN function。WN 是由 8 塊多種卷積層與殘差模組組成，包含單位卷積層(Pointwise Convolution)、空洞卷積層(Dilated Convolution)，以及殘差跳躍連接(Residual and Skip Connection)，此架構類似於 WaveNet[8]，計算後輸出 s 和 t ，接著將剩下的另一半通道數 x_b ，由 s 和 t 的轉換公式(9)得到 x_b' ，並將 x_a 與 x_b' 作拼接(Concat)。此外，在仿射耦合層中，同一半部中的通道不會直接被修改，但若不跨通道的混合訊息，那會有部分的參數不會被調整，因此在每層仿射耦合層前，會添加 1*1 的可逆卷積，使得通道間的訊息可被混合。

$$z \sim \mathbf{N}(z; \mathbf{0}, \mathbf{I}) \quad (5)$$

$$x = f_0 \circ f_1 \circ \dots \circ f_k(z) \quad (6)$$

$$x_a, x_b = \text{split}(x) \quad (7)$$

$$(\log s, t) = \text{WN}(x_a, \text{Mel-spectrogram}) \quad (8)$$

$$x_b' = s \odot x_b + t \quad (9)$$

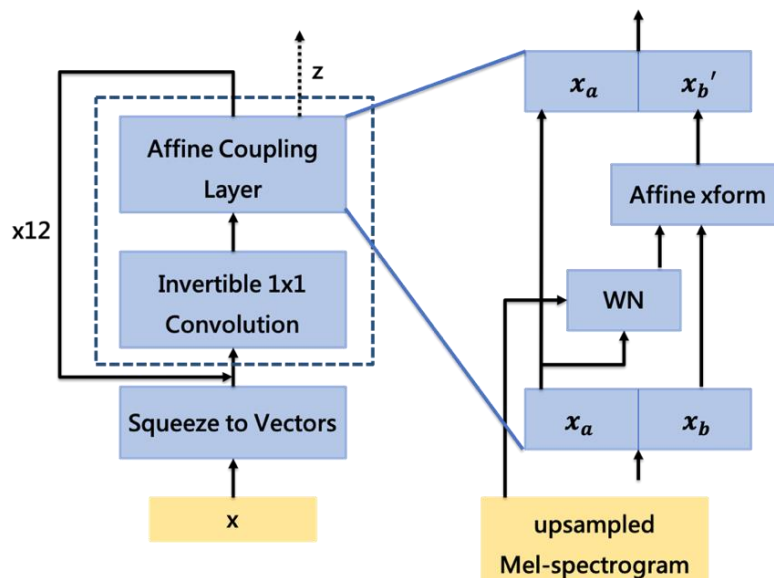
$$f_{\text{coupling}}^{-1}(x) = \text{concat}(x_a, x_b') \quad (10)$$

$$\log|\det(J(f_{\text{coupling}}^{-1}(x)))| = \log|s| \quad (11)$$

$$f_{\text{conv}}^{-1} = Wx \quad (12)$$

$$\log|\det(J(f_{\text{conv}}^{-1}(x)))| = \log|W| \quad (13)$$

$$\log p_{\theta}(x) = -\frac{z(x)^T z(x)}{2\sigma^2} + \sum_{j=0}^{\#\text{coupling}} \log s_j + \sum_{k=0}^{\#\text{conv}} \log W_k \quad (14)$$



圖六、WaveGlow 模型架構

(五) 合成器－訓練方式

這部分使用五種不同的訓練過程，對合成器進行訓練，目的都是希望能使合成的中文語音具備台灣腔調。我們利用不同的資料集、凍結某部分參數的權重，或是只取出部分參數權重與另一模型結合，詳細過程會在以下說明，而以下五項實驗於聲碼器統一使用 Griffin-Lim 合成語音。

1. Tacotron-2(YW 資料集)

使用個人錄製的 YW 資料集對 Tacotron-2 進行訓練。合成的效果的确是台灣腔的中文語音，但因為個人錄製的環境及設備並非專業，使得資料集的音檔略帶雜音，且個人的發音並非完全正確清晰，語調較平淡，語速較緩慢，導致最後合成的語音品質尚須許多改善。

2. GST Tacotron-2(Biaobei 資料集)，freeze Tacotron-2，GST Tacotron-2(YW 資料集)

為了改善方法一的合成品質，我們嘗試了方法二。首先使用 Biaobei 資料集訓練 GST Tacotron-2，當此模型的損失值已經收斂時，將模型所有參數的權重保存，接著凍結 Tacotron-2 的參數權重，目的是希望先保有 Biaobei 資料集的語音品質，最後使用 YW 資料集繼續訓練 GST Tacotron-2，去微調 GST 的參數權重，希望可因此學習到 YW 資料集的風格與腔調。訓練完畢後並進行語音合成時，於 GST 中兩種推斷方式皆嘗試：由於我們希望合成語音能具有 YW 資料集的腔調，因此首先使用 YW 資料集中未經過訓練的音檔作為參考音檔，透過參考編碼器與風格標記層得到其風格特徵；接著我們嘗試指定 4 種風格參數的權重為 $\text{style}(A, B, C, D)=\text{style}(0.0, 0.4, 0.0, 1.2)$ 。而兩者的合成效果相似，其語音的確具有 Biaobei 資料集的語音品質，但 YW 資料集的風格只學習到了語速較慢的特色。

3. GST Tacotron-2(YW 資料集)，freeze GST，GST Tacotron-2(Biaobei 資料集)

與方法二同時進行的是方法三。首先使用 YW 資料集訓練 GST Tacotron-2，接著將所有參數權重保存，凍結 GST 的參數權重，這樣的流程是希望先保有 YW 資料集的風格與腔調，最後使用 Biaobei 資料集繼續訓練 GST Tacotron-2，去微調 Tacotron-2 的參數權重，希望具有 Biaobei 資料集的語音品質。訓練完成後並進行語音合成時，同樣先使用 YW 資料集中未經過訓練的音檔作為參考音檔，接著也指定 4 種風格參數的權重為 $\text{style}(A, B, C, D)=\text{style}(0.0, 0.4, 0.0, 1.2)$ ，進行兩種推斷方式。兩者的合成結果也近似，其合成語音的確有 Biaobei 資料集乾淨清晰的品質，且有 YW 資料集的平淡語調、語速

較慢的風格，但台灣腔的成分卻仍不夠明顯，但此方法的合成效果是目前最接近實驗的目標。

4. GST Tacotron-2(Biaobei 資料集)與 GST Tacotron-2(YW 資料集)，各取部分之參數 權重進行組合

嘗試過凍結參數權重並微調的方法後，我們設想是否能透過模型融合的方式達到想要的結果。首先使用 Biaobei 資料集與 YW 資料集，分別訓練 GST Tacotron-2，使得目前有兩組相同模型架構、經由不同資料集訓練後的參數權重。接著，於使用 Biaobei 資料集訓練的模型中，取出 Tacotron-2 的參數權重；使用 YW 資料集訓練的模型中，取出 GST 的參數權重，進行合併，得到一組新的 GST Tacotron-2 的參數權重。訓練完成後，使用兩種推斷方式：取 YW 資料集中未經過訓練的音檔作為參考音檔、指定 4 種風格參數的權重為 $\text{style}(A, B, C, D) = \text{style}(0.0, 0.4, 0.0, 1.2)$ 。在不作微調而是直接給定參數權重的情況下，兩者的合成效果卻仍與方法三大同小異，由此可知，方法三的參數權重已微調至與方法四的參數權重相近了。

5. Tacotron-2(Biaobei 資料集)，作為 pretrain model，Tacotron-2(NER-2hr 資料集)

方法一到四中，我們希望透過 GST Tacotron-2 這個神經網路搭配不同的訓練方式，能夠獲得具有音質好、台灣腔調的語音合成，但合成的結果總是無法完全學習到腔調的部分。因此想要透過整理出來的 NER-2hr 資料集，單純訓練 Tacotron-2，來達成目標。但因為 NER-2hr 資料集的資料量偏少，且人工切割音檔容易導致音檔的平均長度範圍偏大，致使訓練的難度提升，於是利用 pretrain model 的想法來解決此問題。首先使用 Biaobei 資料集訓練 Tacotron-2，且將 Biaobei 資料集的採樣率調至與 NER-2hr 資料集相同，當模型的損失值已收斂時，在不凍結任何參數權重的情况下，繼續訓練 NER-2hr 資料集。此方法的合成結果，順利達成我們想要的音質好、台灣腔調的中文語音合成。

(六) 聲碼器－訓練方式

透過上述五種實驗，可藉由方法五順利達成這次的目標。但上述的方法中聲碼器的部分，皆是以演算法 Griffin-Lim 來進行梅爾頻譜圖轉語音的動作，而聲碼器卻也是影響語音品質的因素之一。因此，我們將方法五搭配不同的聲碼器，希望達到更好的合成品質。

1. Griffin-Lim

Griffin-Lim 為一種迭代的演算法，並非需要透過資料集訓練的神經網路，因此不須討論資料集以及其訓練方式。而在實驗中，我們設定演算法的迭代次數為六十次，以確保穩定性。

2. WaveGlow

針對此神經網路，我們首先使用 LJSpeech 資料集進行訓練，作為一個 pretrain model，接著再使用 Biaobei 資料集接續訓練。會選擇這樣的訓練方式，是因為我們直接使用 Biaobei 資料集訓練時，其合成的語音略帶雜音，而改成使用 NER-2hr 資料集時，因為資料量偏少，即導致欠擬合的情況發生。

(七) 小結

1. YW 資料集與 NER-2hr 資料集之差別

YW 資料集為個人自行錄製的語料集，錄製環境與設備皆非專業等級，該資料集只適用於一般的個人研究上，當要建一個完整並供大多數人使用的語音合成系統時，此資料集的質量會導致語音合成的品質低落；反之，NER-2hr 資料集是透過專業設備、專業人士製作而成，其品質能有一定的保障。因此，即便起初已有客製化的台灣腔中文語料集，但仍想嘗試後續實驗，希望使用有限且質量好的資料集，藉由不同的訓練方式，來達成台灣腔中文語音合成的目標。

2. 微調(Fine-Tune)不同模型參數對合成結果之影響

在合成器的訓練方式中，方法二、三皆嘗試凍結模型中部份不須再更新的參數，並微調模型中其餘須再繼續更新的參數，而從實驗結果可以發現，方法三比方法二更接近實驗目標。我們的目標是希望藉由 GST 模組學習到台灣腔，因此直接先使用 YW 資料集對 GST 模組做訓練再凍結該參數，比起使用 Biaobei 資料集訓練 GST 模組後再微調成 YW 資料集的參數權重，台灣腔的效果能夠更加明顯。

3. 預訓練模型(Pretrained Model)對合成結果之影響

當我們使用較小資料量的語料集，重新訓練一個複雜度高的模型時，容易導致過擬合(overfitting)的情形發生，即在訓練集上能合成完整的語音，但在測試集上卻有漏字、靜音、雜音等等的問題，因此在合成器的訓練方式其方法五、聲碼器的訓練方式其方法二中，皆使用 pretrained model 的方式進行後續訓練，改善合成不佳的問題。

三、實驗結果

在合成器與聲碼器的各項訓練方式皆訓練完畢後，我們將生成五句與訓練資料不重複的中文文本，進行語音合成，由八位受測人員進行評估，評分標準採用 MOS(Mean Opinion Score)，在每一句子聽完後給予主觀分數，分數範圍為 1~5 分。首先針對合成器的五項實驗進行評分，越高分表示合成的語音其台灣腔調越明顯、合成音質越乾淨無噪，測試結果如表六。接著針對合成器五項實驗中，分數最高的實驗串接兩種不同的聲碼器進行評分，越高分表示合成品質越好，測試結果如表七。

四、結論

我們目前的研究在 Tacotron-2[3]、GST Tacotron-2[4]上嘗試了許多訓練方式，包括凍結參數的權重、組合兩個模型參數的權重，以及使用預訓練模型進行訓練，並將這些訓練結果與 Griffin-Lim[1]、WaveGlow[2]進行結合，分別評估其 MOS 的高低。透過八名受測人員評分的結果，可以發現方法五，使用 Biaobei 資料集訓練 Tacotron-2 作為預訓練模型，接著使用 NER-2hr 資料集接續訓練，並串接聲碼器 Waveglow，此實驗 MOS 為最高者。雖然最後合成台灣腔的中文語音，並非藉由 GST Tacotron-2 提取聲音韻律特徵的模型架構徹底達成，但可以發現透過不同的訓練方式，該模型能夠合成出風格多樣性的中文語音。在語音合成的領域，除了透過學習韻律使得聲音更接近人聲之外，也期望能合成更接近日常用語的語音。越來越多人使用的文字內容不再只有單一語言，經常中英夾雜的使用，因此我們未來的方向將會朝混合語言的語音合成進行研究，了解並應用代碼轉換(Code Switching)[5]、語音克隆(Voice Cloning)[6]等技術，期望能夠合成具有韻律且混合語言的語音。

表六、合成器五項實驗之 MOS

方法	實驗方法簡述	於推斷時，相關參數設定		
		MOS (無須相關設定)	MOS (設定參考音檔)	MOS (設定風格權重)
一	Tacotron-2(YW)	3.91	-	-
二	GST Tacotron-2(Biaobei)， freeze Tacotron-2， GST Tacotron-2(YW)	-	2.30	2.83

三	GST Tacotron-2(YW) , freeze GST , GST Tacotron-2(Biaobei)	-	3.03	3.37
四	GST Tacotron-2(Biaobei) 與 GST Tacotron-2(YW) , 各取部分之參數權重進行組合	-	3.10	3.21
五	Tacotron-2(Biaobei) , 作為 pretrain model , Tacotron-2(NER-2hr)	4.25	-	-

表七、聲碼器兩項實驗之 MOS

方法	實驗方法簡述	MOS
一	表六、方法五 + Griffin-Lim	4.25
二	表六、方法五 + WaveGlow	4.32

參考文獻

- [1]. Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. Deep griffin-lim iteration. *CoRR*, abs/1903.03971, 2019.
- [2]. Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. *CoRR*, abs/1811.00002, 2018.
- [3]. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017.
- [4]. Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *CoRR*, abs/1803.09017, 2018.
- [5]. Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng. End-to-end code-switched tts with mix of monolingual recordings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6935–6939, 2019.
- [6]. Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R. J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *CoRR*, abs/1907.04448, 2019.
- [7]. Bohan Zhai, Tianren Gao, Flora Xue, Daniel Rothchild, Bichen Wu, Joseph E. Gonzalez, and Kurt Keutzer. Squeezewave: Extremely lightweight vocoders for on-device speech synthesis. *CoRR*, abs/2001.05685, 2020.
- [8]. Aˆaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

Taiwanese Speech Recognition Based on Hybrid Deep Neural Network Architecture

Yu-Fu Yeh , Bo-Hao Su , Yang-Yen Ou , Jhing-Fa Wang

Department of Electrical Engineering

National Cheng Kung University Tainan, Taiwan

n26070106@gs.ncku.edu.tw , xtlettle99360017@gmail.com , ouyang0916@gmail.com ,
wangjf@mail.ncku.edu.tw

An-Chao Tsai

Department of Computer Science and Entertainment Technology

Tajen University, Pingtung, Taiwan

actsai@tajen.edu.tw

Abstract

In this research, we developed the Taiwanese speech recognition system which used the Kaldi toolkit to implement. The Taiwanese corpus was collected by Taiwan Taiwanese National Reading Competition and Classmate Recording, and a total of about 11 hours of audio files were collected. Because the training data is small dataset, two audio augmentation methods are used to increase the training data, so that the acoustic model can be more robust and more effective training. One method is speed perturbation, which speeds up the original data by 1.1 times and slows it down by 0.9 times. Another method is to use multi-condition training data to simulate reverberation of the original speech and add background noise. The background noise includes music, speech, and noise. The acoustic model is trained for different hybrid deep neural network architectures which can use the advantages of each neural network by hybrid different neural networks, including TDNN, CNN-TDNN and CNN-LSTM-TDNN. In the experimental results, the CER in the domain of language modeling reaches 3.95%, and the CER of online decoding test is 3.06%. Compared with other researches on Taiwanese speech recognition of similar dataset size, the recognition results are better than other studies.

Keywords: Speech Recognition, Taiwanese, Data Augmentation, Deep Neural Network Acoustic Model.

1. Introduction

In the past few years, more and more products using speech recognition technology. Because these speech recognition applications make people's lives more and more convenient, no longer need to type to allow the machine to receive our message input. Taiwanese language is one of the commonly used languages of Taiwanese. From [1], we can know that in 2013, the social change survey results showed that 31.4% of the people in the family spoke Mandarin Chinese most often, and 44.2% of them spoke Taiwanese most often. 19.5% of the people advocate using both Mandarin and Taiwanese, but the proportion of the older generation is much larger than that of the younger generation, which shows that Taiwanese is still the main language for the elderly. Most of them are learning by word of mouth, which leads to relatively scarce resources in Taiwanese. It makes the research of Taiwanese-related technologies much more difficult, and also causes people who speak Taiwanese to not enjoy these conveniences. Therefore, we have established a Taiwanese dataset and a Taiwanese speech recognition system for this problem.

2. Related Work

Establishes a deep neural network architecture in kaldia[2], the input features used in addition to the Mel frequency cepstral coefficients[3] will also concatenate the ivector feature [4], which is a feature vector that can represent the speaker. First, a general background model is trained on the data of all speakers. The universal background model(UBM) is a Gaussian mixture model containing many components, and then the UBM is modified with the speech features of different speakers to achieve the speaker adaptation model, and the expected values of each Gaussian component are concatenated to form a GMM super-vectors. A section of GMM super-vectors can be used to represent the feature vector of a speaker. Finally, the GMM super-Vectors of the general background model are related to the speaker. GMM super-Vectors calculates ivectors.

In recent years, more and more DNNs have been used to replace GMM to increase the modeling capabilities of acoustic models, indicating that DNN-HMM is better than traditional GMM-HMM, and Kaldi continues to update the DNN architecture to build acoustic models, such as Time Delay Neural Network (TDNN) [5], CNN-TDNN or LSTM-TDNN [6], etc. TDNN is a deep neural network structure. It can include historical and future outputs and model long-term dependent speech signals. It was first proposed to classify phonemes in speech signals. Used

for speech recognition [7]. For TDNN, increasing the number of layers allows the network to capture features for a longer period of time; usually it is desirable to deepen the number of network layers of TDNN to achieve better results. However, previous experiments have found that the deeper the network is, the more often the problem of degradation is, so that the increase in the depth of the neural network will result in a decrease in accuracy. Therefore, another TDNN network architecture [8] is proposed. The Matrix Factorization training of the network can make the network training more stable, in order to achieve better speech recognition performance.

Traditional Discriminative Training requires cross-entropy training to obtain a lattice, which must take extra time. Therefore, the extended framework of CTC is proposed, Lattice-free maximum mutual information [9]. The principle is the same as the method of MMI, the formula is as formula (1), and the following changes are made: (a) The denominator FST uses training text to generate a 4-gram phone language model, and does not use backoff less than 3-gram, instead of lattice.(b) Use different training techniques to avoid Overfitting, like: L2 regularization on the network output, Cross-entropy regularization and Leaky HMM

$$F^{MMI} = \sum_u \log P(S_u | O_u, \lambda) = \sum_u \log \frac{P(O_u | S_u, \lambda) P(S_u)}{\sum_{S'} P(O_u | S', \lambda) P(S')} \quad (1)$$

3. Proposed system

The overall architecture of this system is shown in Figure 1, which include Pre-processing, Deep Neural Networks Acoustic model, Decoding Graph and Recognition.

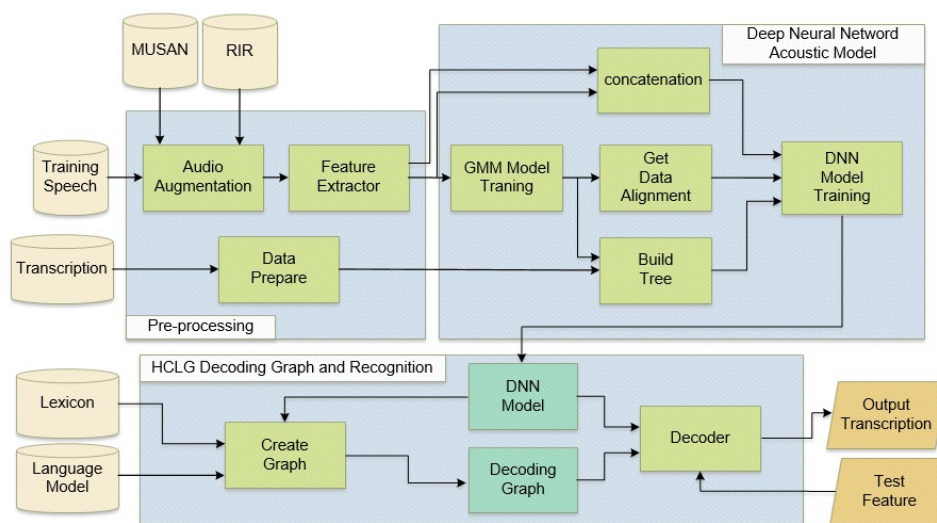


Figure 1. System Flow diagram

3.1 Pre-processing

In the field of speech recognition, the data augmentation is commonly used to increase the quantity of training data, avoid overfitting and improve robustness of the models. The system uses two types of data augmentation, including speed perturbation [10] and using multi-condition training data [11]. In this system, the speed perturbation first generates 3 times the amount of original data, and then this data generates 15 times the amount of original data by adding multi-conditional background noise. Increase the original 10 hours of training data to 150 hours.

The 39-dimensional MFCC feature is used in the GMM-HMM system, and with the addition of Cepstral Mean and Variance Normalization, the standard features of mean 0 and Variance 1 are obtained to solve the effects of different microphones and audio channels. The DNN-HMM system uses high resolution MFCC and ivector. The ivector extraction process is: (1) use 40-dimensional features and 512 Gaussian training diagonal universal background model to obtain final.dubm (2) use the obtained UBM to train ivector extractors (3) Use ivector extractors to extract the ivector of each training data. Feature parameter of TDNN architecture is shown in Figure 2. In order to obtain more context information when inputting deep neural network training, the input will be $(t-1, t, t+1)$ three times 40-dimensional high resolution MFCC feature stitching, followed by 100-dimensional ivector features.

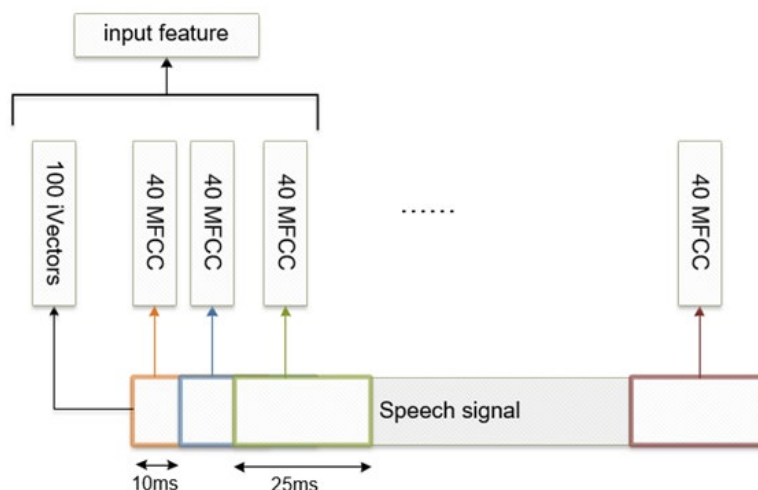


Figure 2. TDNN input feature

The feature parameters of CNN-TDN and CNN-LSTM-TDNN architecture will first convert 40-dimensional high resolution MFCC into Mel-FilterBanks features through Inverse discrete cosine transform layer. Linear transform a 100-dimensional ivector into a 200-dimensional ivector and concatenate with Mel-FilterBanks features. Finally, convert 240-dimensional input features into 40×6 input feature map, such as Figure 3.

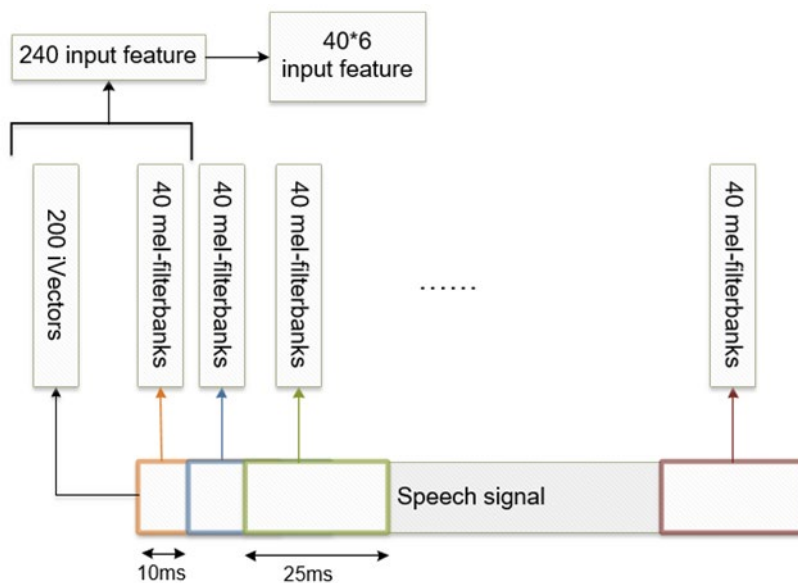


Figure 3. CNN-TDNN and CNN-LSTM-TDNN input feature

3.2 Deep Neural Networks Acoustic model

In this chapter, before training the deep neural network acoustic model, the GMM-HMM system must be pre-trained, and the alignment result obtained by the GMM-HMM system should be used as the training target of the deep neural network acoustic model. This system models HMM at the phone level. Each phone HMM model has 3 states. In the Taiwanese Pinyin system, there are 85 phones including initials and finals. If the tone is considered, the system has 299 HMM models and GMM-HMM model training steps are mono, tri1, tri2, tri3[2].

This research establishes three DNN architectures, including (a) TDNNF, (b) CNN-TDNNF, (c) CNN-LSTM-TDNN.

3.2.1 TDNNF Architecture

The data alignment obtained by the GMM-HMM system is used to establish a decision tree, and the number of leaves corresponds to the output dimension of the deep neural network. Therefore, the architecture output dimension of this chapter is 2776. The TDNNF architecture is shown in Figure 4. This architecture refers to the WSJ recipe and uses the TDNNF architecture proposed by Povey, Daniel, et al. [9], which uses a total of 13 layers of TDNNF layer. The first layer will be 100-dimensional ivector Features and three consecutive 40-dimensional MFCC features make up a total of 220-dimensional input features. Layers 2-4 are $(t-1, t, t+1)$ three-time input vectors, and layers 6-13 are $(t-3, t, t+3)$ three time input vectors, so each frame output can get the information of the first 28 frames and the last 28 frames. The dimension of each layer is 1024, and the SVD decomposition dimension is 128. The internal

architecture of each TDNNF block is shown in Figure 3-10, and the output is divided into chain output and Cross-Entropy output as shown in Figure 4.

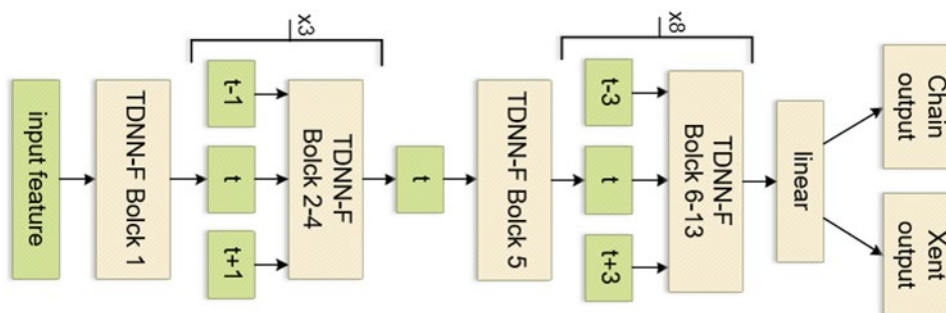


Figure 4. TDNN Architecture

3.2.2 CNN-TDNNF Architecture

In this section, the TDNN architecture used in section 3.3.1 is added to the CNN architecture. The CNN operation method is shown in Figure 5, which is characterized by a 40×6 matrix, and consists of 3 consecutive times to form $3 \times 40 \times 6$ three-dimensional input matrix, before doing convolution, first zero-padding the height to become a $3 \times 42 \times 6$ matrix, using 48 $3 \times 3 \times 6$ size filters for convolution, the output is the first layer. The output of the convolutional layer, if there is subsampling, will only reduce the dimension of the height.

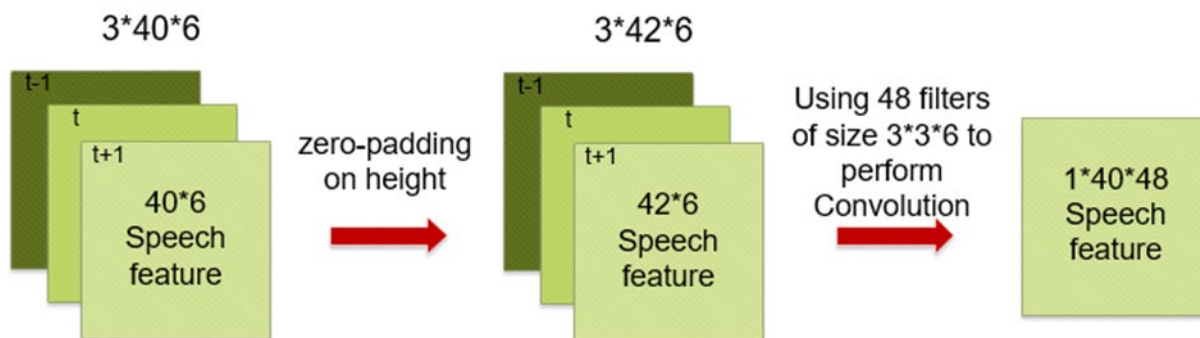


Figure 5. Convolution Neural Network Operation

The output dimension of the CNN-TDNNF architecture in this chapter corresponds to the number of decision tree leaves is 2776. The CNN-TDNNF architecture is shown in Figure 6. This architecture refers to the mini_librispeech recipe, uses 6 layers of convolution layer, and the first layer receives three consecutive times 40×6 . The dimension of the speech feature matrix is $3 \times 40 \times 6$. After the first layer of convolution layer operation, the output is $1 \times 40 \times 48$. After that, each layer uses three consecutive input times, and at the 3rd, 5th and 6th layers, the height will be subsampled, and finally the output dimension will be $1 \times 5 \times 128$. After the CNN, 9-layer

TDNNF layer is used, where each layer has a dimension of 1024, and the SVD decomposition dimension is 128 dimensions, and each layer of TDNNF layer uses $(t-3, t, t+3)$ three Time is used as the input vector, so each output can get the information of the first 30 frames and the last 30 frames.

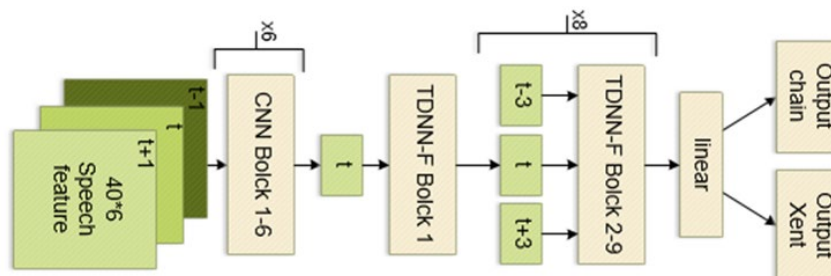


Figure 6. CNN-TDNNF Architecture

3.2.3 CNN-LSTM-TDNN Architecture

CNN-LSTM-TDNN is a deep neural network architecture designed by us. Using CNN can effectively extract feature parameters from a small corpus, and LSTM can model the advantages of long-term sequences to find out this deep neural network architecture. The output dimension of the CNN-LSTM-TDNN architecture in this chapter corresponds to the number of decision tree leaves is 2776. The CNN-LSTM-TDNN architecture is shown in Figure 7, which uses 6 layers of convolution layer, 8 layers of TDNN layer and 2 layers of LSTM layer, among which the convolution layer The architecture parameters are the same as in CNN-TDNNF. The TDNN layer is a general TDNN non-matrix decomposition, and the LSTM cell dimension is 1024. The dimensions of the recurrent and non-recurrent projection layer are all 256, so the input gate, forget gate and output gate input are $1024+256=1280$, and then the 1024-dimensional vector is output to the Cell state and Hidden state through the Nonlinear activation function, and the final output is r_t and p_t concatenation. Each TDNN layer has a dimension of 1024 and receives $(t-3, t, t+3)$ three time inputs, so each output can get the information of the first 33 frames and the last 33 frames, and finally output to the chain output and Cross-Entropy output.

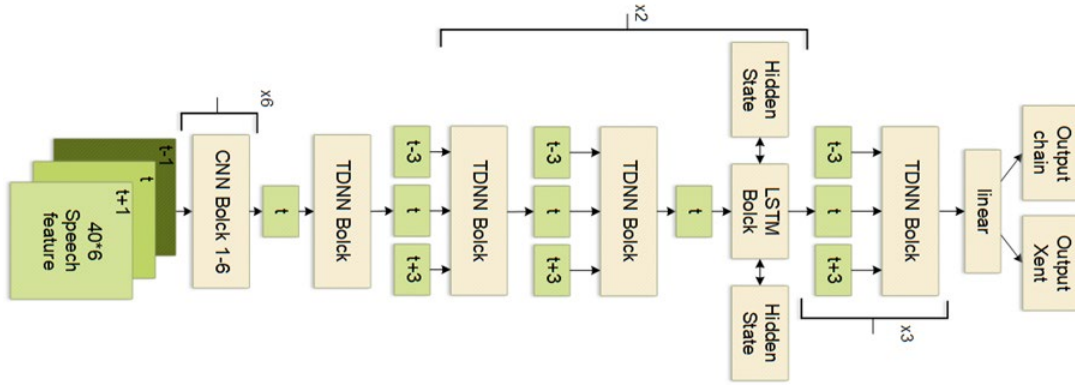


Figure 7. CNN-LSTM-TDNN Architecture

4. Experimental Results

4.1 Taiwanese Dataset

The audio files and the corresponding transcriptions of Taiwanese characters were collected by Taiwan Taiwanese National Reading Competition [12] and recorded by classmates in the laboratory. The sampling frequency is 16kHz, the sampling accuracy is 16bit, and the number of channels is 1 (mono). 11.23 hours, 10439 utterances, 101596 syllables. The experimental part divides the corpus into 10.22 hours of training data and 1.01 hours of testing data.

Lexicon grabs each word and corresponding phone from Taiwanese transcription, and adds the Taiwanese vocabulary of the text of the language model to lexicon, a total of 31331 words are obtained.

The language model text dataset of this system has about 540,000 words, including: 29724 uni-grams, 22123 bi-grams, and 66118 tri-grams.

4.2 Evaluation Method

We use Character Error Rate metrics to evaluate model accuracy in Taiwanese. Character Error Rate (CER), is a common metric of the performance of a speech recognition or machine translation system. The formulas to calculated accuracy as follows:

$$CER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (2)$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, N is the number of words in the reference (N=S+D+C).

4.3 Experimental Results

The Taiwanese corpus of this system is a small dataset, and the amount of data is far from the size of other language corpus, so consider whether to use tone to label phones. If tones are considered in the part of the dataset marked with phones, the number of phones will increase. This will cause more HMM models, and vice versa, reduce the number of HMM models. And this experiment is to explore whether tones are added to the dataset as the Taiwanese corpus used by this system. The number of HMM models without adding tones is 86, which is much smaller than the HMM with adding tones, which is 299. This experiment uses 1.01 hours of testing data to test the performance of the model. The testing data has a total of 9692 syllables. Table 1 shows the experimental results of whether the corpus adds tones. It can be seen from the experimental results that in the case of the same language model, the traditional GMM-HMM system does not add tones better than tones, but the acoustic model of the deep neural network is the opposite. It can be seen that the architecture of the deep neural network has a better ability to model speech signals, so all subsequent experiments will consider the tone as the pinyin label of the Taiwanese corpus.

Table 1. Compare whether the dataset adds tones

Model	With tone(CER%)	Without tone(CER%)
Mono	36.32	29.94
Tri1	27.24	26.06
Tri2	24.39	23.68
Tri3	19.13	17.33
TDNNF	10.21	12.12

However, in training speech recognition systems, overfitting problems are often encountered. In order to solve this problem, the easiest way is to add training data. But this is a thorny problem in the case of limited data and manpower, so training data is often obtained through data augmentation. This experiment compares two data augmentation methods, including speed perturbation[10] and using multi-condition training data[11]. The increase in training data can make the model training deeper and more efficient. Table 2 shows the comparison results of adding different data augmentation methods. The first column is a different system. Take the TDNNF architecture as the acoustic model and add SP and multi respectively, where SP stands for speed perturbation and MULTI stands for using multi-condition training data,

The second column shows the amount of data after data augmentation. The TDNNF + SP + MULTI system adds SP first and then MULTI, which increases the total amount of data by 15 times. The third and fourth columns represent the character error rate of the testing data. It can be seen that for the same acoustic model system, the character error rate decreases as the data increases.

Table 2. Comparison of data augmentation methods

System	Duration of data (hours)	CER (%)	Error/Total
TDNNF	10	10.21	991 / 9692
TDNNF + SP	30	8.91	864 / 9692
TDNNF + MULTI	50	8.14	789 / 9692
TDNNF + SP + MULTI	150	7.94	770 / 9692

In order to explore the impact of different deep neural network models on acoustic models, four sets of models were set up in this experiment, including TDNNF, CNN-TDNNF, LSTM-TDNN, and CNN-LSTM-TDNN. The recognition results of each deep neural network acoustic model are shown in Table 3. It can be seen that the effect of LSTM-TDNN is worse than the other three, and the best model is the CNN-LSTM-TDNN mixed three deep neural network architecture acoustic models. It can be seen that the effect of the LSTM-TDNN architecture is very poor, possibly because the number of training data in the corpus is too small. Although LSTM is an algorithm for time series training, it requires too many parameters, so a larger amount of training data is needed for training. The disadvantage of CNN is that there is no concept of time series, and the use of too many parameters leads to an increase in training time. But shows that for acoustic models, convolutional neural networks can effectively help feature extraction in small dataset and overall deep neural network learning.

Table 3. Comparison of different model recognition results

Model	CER (%)	Error/Total	ins	del	Sub
TDNNF	7.94	770 / 9692	79	91	600
CNN-TDNNF	7.68	744 / 9692	80	56	608
LSTM-TDNN	10.20	989 / 9692	91	103	796

CNN-LSTM-TDNN	7.61	738 / 9692	88	49	601
----------------------	-------------	------------	----	----	-----

Finally, a total of 10 people in the laboratory are asked to do an online decoding test. Each person tests 15 sentences. The test text is a daily language in Taiwanese, and the training text of the language model has been added. There are 150 sentences and 1078 syllables in total. The recognition results are shown in Table 4. The online decoding test text example is shown in Table 5.

Table 4. Online decoding test results

	CER (%)	Error/Total	ins	del	Sub
Online decoding test	3.06	33 / 1078	6	9	18

Table 5. Online decoding test text

Testing data number	Text
1	gua2 beh4 khi3 tai5 pak4
2	gua2 siunn7 beh4 tshut4 khi3 tshit4 tho5
3	gua2 siunn7 beh4 tsiah8 mih8 kiann7
4	kin1 a2 lit8 thinn1 khi3 be7 bai2
5	u7 siann2 mih4 ho2 tsiah8 e5
...	...
150	gua2 beh4 khi3 siong2 kho3 ah4

5. Conclusions

In this paper, we collected a Taiwanese corpus and use the architecture of the deep learning HMM model to build a Taiwanese speech recognition system. Finally, the CNN-LSTM-TDNN architecture is the best. The language model can be changed according to the domain used to greatly improve the accuracy. In our experiment, the character error rate of the testing data of inside domain is 3.95%. The experimental results show that if the text is inside domain, the Taiwanese speech recognition system does get good results. Finally, in actual application, the laboratory classmates were asked to test the online decoding character error rate of 3.06%.

References

- [1] 葉高華. "臺灣歷次語言普查回顧." 臺灣語文研究 13.2 (2018): 247-273.

- [2] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF. IEEE Signal Processing Society, 2011.
- [3] Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE transactions on acoustics, speech, and signal processing 28.4 (1980): 357-366.
- [4] Dehak, Najim, et al. "Front-end factor analysis for speaker verification." IEEE Transactions on Audio, Speech, and Language Processing 19.4 (2010): 788-798.
- [5] Peddinti, Vijayaditya, Daniel Povey, and Sanjeev Khudanpur. "A time delay neural network architecture for efficient modeling of long temporal contexts." Sixteenth Annual Conference of the International Speech Communication Association. 2015.
- [6] Peddinti, Vijayaditya, et al. "Low latency acoustic modeling using temporal convolution and LSTMs." IEEE Signal Processing Letters 25.3 (2017): 373-377.
- [7] Waibel, Alex, et al. "Phoneme recognition using time-delay neural networks." IEEE transactions on acoustics, speech, and signal processing 37.3 (1989): 328-339.
- [8] Povey, Daniel, et al. "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks." Interspeech. 2018.
- [9] Povey, Daniel, et al. "Purely sequence-trained neural networks for ASR based on lattice-free MMI." Interspeech. 2016.
- [10] Ko, Tom, et al. "Audio augmentation for speech recognition." Sixteenth Annual Conference of the International Speech Communication Association. 2015.
- [11] Ko, Tom, et al. "A study on data augmentation of reverberant speech for robust speech recognition." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.
- [12] "台灣國賽台語(閩南語)朗讀篇目整理,"[Online]. Available:
<http://ip194097.ntcu.edu.tw/longthok/longthok.asp>.

NSYSU+CHT 團隊於 2020 遠場語者驗證比賽之語者驗證系統

NSYSU+CHT Speaker Verification System for Far-Field Speaker Verification Challenge 2020

張育嘉 Yu-Jia Zhang, 陳嘉平 Chia-Ping Chen

國立中山大學資訊工程學系

Department of Computer Science and Engineering

National Sun Yat-sen University

M083040025@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

在本論文中，我們描述了 NSYSU+CHT 團隊在 2020 遠場語者驗證比賽 (2020 Far-field Speaker Verification Challenge, FFSVC 2020) 中所實作的系統。單一系統採用基於嵌入的語者識別系統。該系統的前端特徵提取器是結合了時延神經網路，與卷積神經網路模組兩者的優點，稱為時延殘差神經網路的架構。在池化層，我們實驗了不同方式：統計池化層和 GhostVLAD。而後端的評分器則採用機率線性判別分析，我們訓練跟調適機率線性判別分析用以不同系統的融合。我們分別參加了 FFSVC 2020 採單一麥克風陣列資料的文本相關(任務一)與文本無關(任務二)的語者驗證任務。我們提出的系統在任務一上取得 minDCF 0.7703，EER 9.94%，在任務二上則是 minDCF 0.8762，EER 10.31%。

Abstract

In this paper, we describe the system Team NSYSU+CHT has implemented for the 2020 Far-field Speaker Verification Challenge (FFSVC 2020). The single systems are embedding-based neural speaker recognition systems. The front-end feature extractor is a neural network architecture based on TDNN and CNN modules, called TDResNet, which combines the advantages of both TDNN and CNN. In the pooling layer, we experimented with different methods such as statistics pooling and GhostVLAD. The

back-end is a PLDA scorer. Here we evaluate PLDA training/adaptation and use it for system fusion. We participate in the text-dependent(Task 1) and text-independent(Task 2) speaker verification tasks on single microphone array data of FFSVC 2020. The best performance we have achieved with the proposed methods are minDCF 0.7703, EER 9.94% on Task 1, and minDCF 0.8762, EER 10.31% on Task 2.

關鍵詞：遠場語者驗證、時延神經網路、卷積神經網路、時延殘差神經網路、GhostVLAD

Keywords : Speaker Verification, TDNN, CNN, TDResNet, GhostVLAD

基於深度學習之中文文字轉台語語音合成系統初步探討

A Preliminary Study on Deep Learning-based Chinese Text to Taiwanese Speech Synthesis System

許文漢 Wen-Han Hsu, 曾證融 Cheng-Jung Tseng, 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

jeff3136169@gmail.com, t107368030@ntut.edu.tw, yfliao@ntut.edu.tw

王文俊 Wern-Jun Wang, 潘振銘 Chen-Ming Pan

中華電信實驗室

Chunghwa Telecom Laboratories

wernjun@cht.com.tw, chenming@cht.com.tw

摘要

台語在台灣歷史悠久，使用的族群眾多，有著很重要的存在價值。語音合成在追求跟人類一樣的聲音以及語調的同時，語言的多樣性也是一個需要深入探討的領域。本論文針對目前較少有的台語語音合成系統來作探討，利用翻譯模型 Chinese to Taiwanese(C2T) 將輸入的中文文字轉成台羅拼音數字調(TLPA)，再將拼音輸入 Tacotron2 模型(Text to Spectrogram)後輸出頻譜，最後由 WaveGlow 模型(Spectrogram to Waveform)來實現語音合成。同時有架設網頁可供使用者一同來測試成效。

本文 C2T 機器翻譯的實驗方面採取三種模式，包括(1)輸入中文字詞，先進行斷詞，再輸出每個中文詞的台語台羅 (Tâi-lô) 拼音。(2)輸入中文字元串，直接輸出台羅拼音串。(3)輸入中文字元串，輸出台語的台羅拼音串與台語詞的斷詞關係。若不考慮聲調，方法(1)的 syllable error rate(SER)為 15.66%。而方法(2)的 SER 更可達 6.53%。這表示我們所用的 sequence-to-sequence 模型確實可以正確地將輸入的中文字元串，直接輸出台羅拼音串。

在台語語音合成品質實驗方面，我們找了 20 位聽者，各聽取 15 句不同內容的合成音檔後，以平均主觀意見進行評分(mean opinion score, MOS, 完全不像人講話的聲音為 1 分，完全像真人講話聲音為 5 分)。總計收集到 300 個評分，最後得到我們系統的 MOS

得分為 4.30 分。這表示我們所用的 Tacotron2 與 WaveGlow 模型確實可以正確將台羅拼音串轉成台語語音。此外此系統的語音合成速度為一秒可合成約 3.5 秒之音檔，的確可以達到即時語音合成的要求。

關鍵詞：機器翻譯、臺灣閩南語羅馬字拼音、台語語音合成

Abstract

This paper focuses on the development and implementation of a Chinese Text-to-Taiwanese speech synthesis system. The proposed system combines three deep neural network-based modules including (1) a sequence-to-sequence-based Chinese characters to Taiwan Minnanyu Luomazi Pinyin (shortened to as Tâi-lô) machine translation (called C2T from now on), (2) a Tacotron2-based Tâi-lô pinyin to spectrogram and (3) a WaveGlow-based spectrogram to speech waveform synthesis subsystems.

Among them, the C2T module was trained using a Chinese-Taiwanese parallel corpus (iCorpus) and 9 dictionaries released by Academia Sinica and collected from internet, respectively. The Tacotron2 and Waveglow was tuned using a Taiwanese speech synthesis corpus (a female speaker, about 10 hours speech) recorded by Chunghwa Telecom Laboratories. At the same time, a demonstration Chinese Text-to-Taiwanese speech synthesis web page has also been implemented.

From the experimental results, it was found that (1) the best syllable error rate (SER) of 6.53% was achieved by the C2T module, (2) and the average MOS score of the whole speech synthesis system evaluated by 20 listeners gains 4.30. These results confirm that the effectiveness of integration of C2T, Tacotron2 and WaveGlow models. In addition, the real-time factor of the whole system achieved 1/3.5.

Keywords: Machine Translation, Taiwanese Speech Synthesis, Tacotron2, Waveglow

基於深度聲學模型其狀態精確度最大化之強健語音特徵擷取的初步研究

The preliminary study of robust speech feature extraction based on maximizing the accuracy of states in deep acoustic models

張立家 Li-chia Chang
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
s108323518@mail1.ncnu.edu.tw

洪志偉 Jieh-weih Hung
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
jwhung@ncnu.edu.tw

摘要

在本研究中，我們提出一種新穎的強健性語音特徵擷取技術，以增進雜訊干擾環境下的語音辨識效能。此新技術，利用語音辨識系統中後端的原聲學模型所提供的資訊，在不重新訓練聲學模型的前提下，藉由深度類神經網路架構，學習得到最大化聲學模型狀態之精確度對應之語音特徵，進而使此語音特徵擁有對雜訊的強健性，相較於其他改善聲學模型以達到雜訊強健性的技術，本研究所提出的新技術具有計算量小且訓練快的優點。

在初步實驗中，我們使用了 TIMIT 此中型語料庫來加以評估，實驗結果顯示所提之新語音特徵擷取法，相對於基礎實驗，能有效地降低各種雜訊種類與雜訊程度之環境下語音的音素錯誤率，凸顯此方法的效能及發展價值。

關鍵詞：雜訊強健性之語音特徵、語音辨識、深度學習

Abstract

In this study, we focus on developing a novel noise-robust speech feature extraction technique to achieve noise-robust speech recognition, which employs the information from the backend acoustic models. Without further retraining and adapting the backend acoustic models, we use deep neural networks to learn the front-end acoustic speech feature representation that can achieve the maximum state accuracy obtained from the original acoustic models. Compared with the robustness methods that retrain or adapt acoustic models, the presented method exhibits the advantages of lower computational complexity and faster training.

In the preliminary evaluation experiments conducted with the median-vocabulary TIMIT database and task, we show that the newly presented method achieves lower word error rates in recognition under various noise types and levels compared with the baseline results. Therefore, this method is quite promising and worth developing further.

Keywords: noise-robust speech feature, speech recognition, deep learning

基於多視角注意力機制語音增強模型於強健性自動語音辨識

Multi-view Attention-based Speech Enhancement Model for Noise-robust Automatic Speech Recognition

趙福安 Fu-An Chao,

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

fuann@ntnu.edu.tw

洪志偉 Jieh-weih Hung

國立暨南國際大學電機工程學系

Department of Electrical Engineering

National Chi Nan University

jwhung@ncnu.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

仰賴深度學習(Deep Learning)的發展，近年來許多研究發現相位(Phase)資訊在語音增強(Speech Enhancement, SE)中至關重要。亦有學者發現，透過時域單通道語音增強技術，可以有效地去除雜訊，進而顯著提升語音辨識的精確度。啟發於此，本研究從時域及頻域面分別探討兩種考慮相位資訊的語音增強技術，並提出多視角注意力機制語音增強模型、融合時域及頻域兩者特徵運用於語音辨識中。我們藉由 Aishell-1 中文語料庫評估這些語音增強技術，透過使用各種雜訊源，模擬不同的雜訊狀態作為訓練及測試，進而驗證所提出的新方法皆優於基於其他時域及頻域的方法。具體而言，當測試於訊噪比為-5dB、5dB、15dB 的三種環境下，使用新提出之方法中重新訓練(Retraining)之聲學模型(Acoustic Model, AM)，與基於時域的方法相比較，在已知雜訊的測試集，分別使相對字錯誤率下降 3.4%、2.5%及 1.6%；而在未知雜訊的測試集，則使相對字錯誤率分別

下降了 3.8%、4.8%及 2.2%。

Abstract

Recently, many studies have found that phase information is crucial in Speech Enhancement (SE), and time-domain single-channel speech enhancement techniques have been proved effective on noise suppression and robust Automatic Speech Recognition (ASR). Inspired by this, this research investigates two recently proposed SE methods that consider phase information in time domain and frequency domain of speech signals, respectively. Going one step further, we propose a novel multi-view attention-based speech enhancement model, which can harness the synergistic power of the aforementioned time-domain and frequency-domain SE methods and can be applied equally well to robust ASR. To evaluate the effectiveness of our proposed method, we use various noise datasets to create some synthetic test data and conduct extensive experiments on the Aishell-1 Mandarin speech corpus. The evaluation results show that our proposed method is superior to some current state-of-the-art time-domain and frequency-domain SE methods. Specifically, compared with the time-domain method, our method achieves 3.4%, 2.5% and 1.6% in relative character error rate (CER) reduction at three signal-to-noise ratios (SNRs), -5 dB, 5 dB and 15 dB, respectively, for the test set of pre-known noise scenarios, while the corresponding CER reductions for the test set of unknown noise scenarios are 3.8%, 4.8% and 2.2%, respectively.

關鍵詞：語音強化、自動語音辨識、深度學習、單通道語音增強、重新訓練、聲學模型

Keywords: Speech Enhancement, Automatic Speech Recognition, Deep Learning, Single-Channel Speech Enhancement, Re-training, Acoustic Models

一、緒論

近年來，隨著深度學習的蓬勃發展，現今使用之基於深度學習架構的自動化語音辨識 (ASR)系統在無雜訊干擾的情況下，已可達到近乎人類感知水平的辨識水準。但是，在真實環境中往往存在背景雜訊等聲學干擾，若在此環境下，使用事先訓練的 ASR 系統，其辨識性能可能會嚴重下降。為了降低這類干擾效應，長期以來已經發展了許多相關研究及技術，而在這些研究相關技術中，語音增強(Speech Enhancement, SE) 是一個主要的類別，其作為語音在訓練聲學模型(Acoustic Models, AM)前消除雜訊干擾的預處理 (Preprocessor)，是一項備受重視的研究方向。

然而，在大多數缺少多麥克風陣列的真實情況，單通道 SE 技術(Single-channel SE) 通常效果較差，只能提高語音品質和理解度指標，例如語音品質的感知評估(Perceptual Evaluation of Speech Quality, PESQ)和短時客觀理解度(Short-time Objective Intelligibility,

STOI)，而在這些前端 SE 的指標有所提升，並不能有效地反映在後端 AM 的辨識結果。其原因大致分為兩種：一、在使用前端 SE 模型時產生額外的失真(Distortion)以及雜訊殘留(Artifacts)，進而影響了後端 AM 的辨識效果；二、在訓練 SE 模型時，目標函數(Objective)常設計為優化或近似原乾淨訊號的品質，但 AM 的訓練目標則為最小化分類錯誤、即降低辨識之錯誤率，兩模型訓練目標不一致，造成前後端不匹配的問題。為了解決這樣的現象，儘管許多學者提出使用聯合訓練(Joint Training)的方法，如[8]，針對如何訓練一個強健(Robust)的單通道 SE 模型以減少輸出的失真，還是一項值得研究的方向[9]。

近年來，從時域分析的角度出發，來處理語音訊號的做法，引起了越來越多的關注[6][7]，其中，全摺積時域音訊分離網絡(Fully Convolutional Time-domain Audio Separation Network, Conv-TasNet)[7]在語音分離(Speech Separation, SS)任務上取得了顯著的成果，且超越了頻域上的相關方法。這個想法也被用於單通道 SE 的相關研究，並同時驗證在 ASR 系統上[10]，其對應之使用多情境訓練(Multi-condition Training, MCT) 的 AM，使辨識率獲得顯著進步。然而，此架構若以時域的訊號作為輸入，其特徵抽取(Feature Extraction)的步驟須仰賴摺積神經網路(Convolution Neural Network, CNN)的處理。因此，當沒有足夠的資料數據來找到適當的特徵分佈時，此方法不利於進行小規模資料集的訓練，且增強效果較頻域方法差[11]。

有鑑於此，在本論文中，我們研究了兩種語音特徵，一種是藉由一維摺積神經網路(Convolution Neural Network, CNN)的濾波器組得到的時域特徵，另一種為傳統 STFT 得到的頻域特徵，後者為人工特徵(Hand-crafted Feature)，因此這兩種特徵分別為可訓練的以及固定的。接著我們提出了多視角注意力機制編碼器(Multi-view Attention-based Encoder)，動態地選擇逐音框(Frame-wise)之特徵，合併為單一強健的語音表示作為最終特徵，以輸入至 SE 模型。

為了評估我們提出的方法其效能，我們使用開源中文語料庫 AISHELL-1[12]，並收集多種雜訊資料庫，合成各種帶噪的語音作為訓練以及測試語料。根據評估實驗結果得知，相較於時域上的對應方法，使用新方法之重新訓練之聲學模型(AM)，在已知雜訊的測試集，於三種訊噪比： -5dB、5dB、15dB 之情境中，分別得到相對字錯誤率 3.4%、2.5%及 1.6%的下降；而在未知雜訊的測試集，則使相對字錯誤率分別下降了 3.8%、4.8%及 2.2%。

二、文獻回顧

考慮一段在單通道麥克風中，受加成性雜訊干擾的帶噪離散語音訊號 $y(t)$ ，其公式如下：

$$y(t) = x(t) + n(t) \quad (1)$$

其中 $x(t)$ 是目標之乾淨語音訊號， $n(t)$ 是加成性雜訊(Additive Noise)， t 是時間索引。我們旨在消除帶噪語音訊號 $y(t)$ 中的加成性雜訊 $n(t)$ ，以恢復乾淨語音訊號 $x(t)$ 。

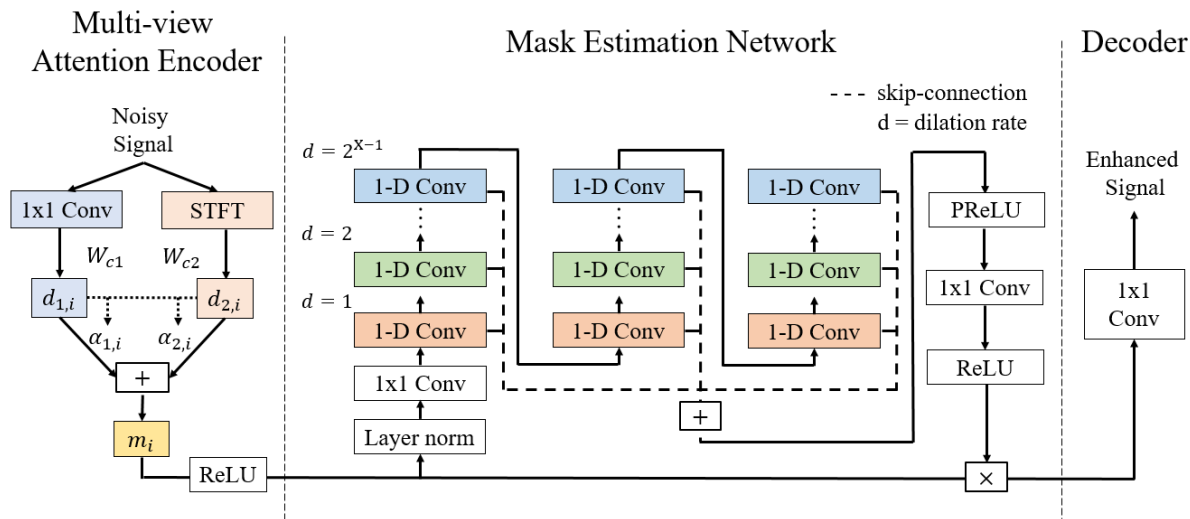
在大多數的研究中，選擇透過短時傅立葉轉換(short-time Fourier transform, STFT)，以在頻域上利用時頻分析(Time-Frequency Analysis)分析帶噪語音訊號，其可視為一特徵抽取的步驟，其公式如下：

$$X(t, f) = \int_{-\infty}^{\infty} w(t - \tau)x(\tau)e^{-j2\pi f\tau} d\tau \quad (2)$$

其中 $w(\cdot)$ 為窗函數，用於將訊號截短， $x(\tau)$ 為待轉換的訊號，窗函數隨著時間在時間軸上位移，因此訊號將只留下窗函數截取部分做傅立葉轉換。式(2)是將一維的時域訊號 $x(t)$ 透過短時傅立葉轉換求得 $X(t, f) \in \mathbb{C}^{t,f}$ ，稱為複數時頻圖 (complex-valued spectrogram)，當以極座標表示，可得 $X(t, f) = |X(t, f)|e^{j\theta_X(t, f)}$ ，即其可拆解成訊號隨著時間與頻率變化的幅度(Magnitude) $|X(t, f)|$ 及相位(Phase) $\theta_X(t, f)$ 。一般來說，相位在估測上比較困難，因此頻域上的語音增強技術通常只針對頻譜幅度做調整、保留原始相位不加以更動。

傳統的 SE 技術主要針對輸入訊號作統計分析，試圖還原乾淨語音訊號，如頻譜消去法(Spectral Subtraction)、維納濾波器(Wiener Filter)等[1]，相較於這些傳統方法所使用的統計分析技術，著名的深度類神經網路(Deep Neural Network, DNN)有更好的非線性轉換能力，已被廣泛應用於 SE 並取得了顯著成果。[2]首先引入了深層去噪自動編碼器(Deep Denoising Auto Encoder, DDAE)，[3]提出了時頻遮罩法(Time-frequency Masking, T-F Masking)。隨後便沿著這個脈絡發展了許多後續研究，這些方法都透過各式 DNN 模型得到優異的去噪效果，大致可以依據其目標分成映射(Mapping)方法或時頻遮罩(T-F Masking)方法[4][5][6]。

近期許多研究發現，相位資訊對語音強化的效能至關重要[4][5][13]。為考慮相位資



圖一、多視角注意力機制語音增強模型

訊，我們需在短時訊號的頻域中同時考慮幅度(Magnitude)和相位(Phase)，或是選擇直接在時域中分析訊號，兩種方法均被證明是有效的，並且優於先前僅考慮頻譜幅度的語音強化技術。在頻域語音強化法中，[13]提出了使用兩個分支結構分別預測幅度遮罩和相位的方法，同時重建幅度和相位，[5]使用遮罩方法，預測複數比率遮罩(Complex Ratio Mask, CRM)。[10]提出了 Denoising-TasNet，其採用[7]中摺積編碼器、解碼器的架構，在時域上對語音訊號做強化，與頻域上的方法相比，其透過控制摺積核(Kernel)以及步長(Strides)大小做摺積運算(Convolution)，類似於短時傅立葉運算(Short-time Fourier Transform, STFT)，隱性地考慮了訊號的相位(Phase)資訊、將其編碼成潛在表示式(Latent Representation)。其去噪效果不僅在訊號失真比(Signal-to-distortion Ratio, SDR)指標有所提升，更驗證在自動語音辨識(ASR)上有效地降低了詞錯誤率(Word Error Rate, WER)。

三、所提出的新方法：多視角注意力機制語音增強模型

在本論文中，我們提出多視角注意力機制語音增強模型，其架構可見圖一，可以拆解成多視角注意力機制編碼器(Multi-view Attention-based Encoder)、遮罩估計網路(Mask Estimation Network)以及解碼器(Decoder)如[7]，藉此對輸入訊號進行建模。首先，我們將輸入之帶噪語音在時域上切割為許多長度為 L 的相互重疊片段，各片段以列向量 $x_m \in \mathbb{R}^{1 \times L}$ 表示，其中 m 表示每個片段的索引， $m \in \{1, \dots, M\}$ ， M 為總片段數。我們將所有片

段縱向串連在一起，表示為一矩陣 $X \in \mathbb{R}^{M \times L}$ ，其每一列即為 x_m 。

(一) 多視角注意力機制編碼器

多視角注意力機制編碼器考慮了時域以及頻域兩種不同的語音特徵，如圖一，它們分別以 C_1 與 C_2 表示，兩種特徵抽取的步驟可視為一線性轉換，首先，時域特徵是藉由矩陣相乘將 X 轉換至 N 列的特徵矩陣 $C_1 \in \mathbb{R}^{N \times L}$ 如下：

$$C_1 = \mathcal{F}(UX) \quad (3)$$

其中 $U \in \mathbb{R}^{N \times M}$ 為轉換之矩陣，包含 N 個基函數(Basis Function)，它們對應到一系列可訓練的一維摺積層(1-D Convolution Layer)係數，而 $\mathcal{F}(\cdot)$ 為可選擇的激活函數(Activation Function)，這裡通常選擇線性整流單元(Rectified Linear Unit, ReLU)[7]為激活函數 $\mathcal{F}(\cdot)$ ，以確保特徵為非負值。此外，頻率特徵矩陣 C_2 是透過傳統的 STFT 對時域訊號矩陣 X 轉換而得，相較於可訓練的時域特徵矩陣 C_1 ，頻譜特徵矩陣 C_2 是固定式的，因其求取所用的傅立葉轉換矩陣為一常數矩陣。

根據近期研究，注意力(Attention)模型廣泛被應用於序列至序列(Sequence-to-sequence)的任務中，其中在語音處理領域也獲得了優異的成果。有鑑於此，我們提出多視角注意力機制編碼器，其利用注意力機制將語音特徵，透過注意力加權後，得到單一特徵表示作為輸入之特徵。希望透過注意力機制，使 SE 模型可以更加關注輸入特徵之變異(Variants)並提高 SE 的性能。為了計算注意力權重，我們首先將個別音框特徵透過投影層(Projection Layer)投影至相同維度的向量，如下式：

$$d_{k,i} = W_{ck} c_{k,i} + b_{ck} \quad (4)$$

其中， $c_{k,i}$ 為語音特徵矩陣 C_k 之個別音框的特徵向量、 i 表示音框索引(Frame index)、 W_{ck} ， b_{ck} 分別是投影層的權重矩陣(Weight Matrix)和偏差(Bias)。藉由投影得到之向量 $d_{k,i}$ ，我們可以透過時序注意力機制(Temporal Attention Mechanism)計算注意力權重 $\alpha_{k,i}$ ：

$$\alpha_{k,i} = \frac{\exp(v_{k,i})}{\sum_{k=0}^{K-1} \exp(v_{k,i})} \quad (5)$$

其中 $v_{k,i}$ 是 $d_{k,i}$ 和其他 $d_{k',i} (k' \neq k)$ 間的相似度(Similarity)， K 是特徵種類的個數，在考慮時域及頻域兩種特徵情況下， $K=2$ 。在我們的實驗中，我們考慮了三種不同相似度分數計算公式，如表一所示：

表一、不同注意力機制相似度分數之計算 (對兩種特徵而言, $k = 0, 1$)

注意力機制	公式
加成性(Additive)	$v_{k,i} = \omega_A^T \tanh(W_A d_{k,i} + B_A d_{1-k,i} + b_A)$
串接性(Concatenate)	$v_{k,i} = \omega_C^T \tanh(W_C [d_{k,i}; d_{1-k,i}] + b_C)$
縮放點積(Scaled dot-product)	$v_{k,i} = \frac{d_{k,i}^T d_{1-k,i}}{\sqrt{n_d}}$

其中 W_A 、 W_C 、 B_A 為權重矩陣， b_A 、 b_C 為偏差， ω_A 、 ω_C 為向量， n_d 為投影向量之維度， $\frac{1}{\sqrt{n_d}}$ 為控制兩向量點積(Dot-product)後值的縮放係數。

最後我們利用計算完的注意力權重 $\alpha_{k,i}$ 將特徵之投影向量 $d_{k,i}$ 加權後相加，得到融合後的特徵表示式 z_i ：

$$z_i = \sum_{k=0}^1 \alpha_{k,i} d_{k,i} \quad (6)$$

(二) 遮罩估計網路

在基於遮罩的 SE 模型中，通常採用 DNN 模型來估計輸入帶噪語音特徵的遮罩，藉此來分離雜訊與語音。為了有效地捕獲時間資訊並考慮音框之間的長期依賴性(Long-term Dependency)以預估遮罩，大部分的研究透過堆疊雙向長短期記憶層(Bidirectional Long-short Term Memory, BLSTM)或擴張摺積層(Dilated Convolution Layer)來實現。其輸出可以直接是目標語音的單個遮罩 M_x ，也可以選擇輸出語音和雜訊兩個各別遮罩 M_x, M_n ：

$$[M_x, M_n] = \mathcal{M}_\theta(Z), \quad (7)$$

其中 $\mathcal{M}_\theta(\cdot)$ 為遮罩估計網路， Z 為輸入語音特徵矩陣，可為式(6)所得之個別音框特徵的排列： $Z = [z_1, z_2, \dots, z_L]$ 。在得到估計之遮罩後，我們可以將語音遮罩與輸入特徵 Z 逐項相乘得到強化後的語音特徵矩陣 $D_x \in \mathbb{R}^{N \times L}$ ：

$$D_x = M_x \odot Z \quad (8)$$

其中， \odot 表示逐元素相乘運算(Element-wise Product)。

(三) 解碼器

得到強化後的語音特徵 D_x 後，我們透過解碼器(Decoder)將特徵表示轉換並重建時域之波形，此步驟可以視為矩陣相乘運算，如下式：

$$\hat{S}_x = VD_x \quad (9)$$

其中 $\hat{S}_x \in \mathbb{R}^{M \times L}$ 為重建的語音片段組成之矩陣，而解碼器矩陣 $V \in \mathbb{R}^{M \times N}$ 是由長度為 M 的 N 個基函數排列而成。在頻域中，矩陣 V 可以是短時逆傅立葉轉換(Inverse Short-time Fourier Transform, iSTFT)，在時域中，矩陣 V 則對應至一維轉置摺積運算(1-D Transpose Convolution)。最後我們使用重疊相加(Overlap-add)的方法從片段矩陣 \hat{S}_x 重構語音訊號。

四、實驗設置

我們使用 AISHELL-1[12]語料集來執行評估實驗，其為北京希爾貝殼科技有限公司提供的開源中文 ASR 語料庫，包含 400 位語者和 170 個小時的中文語音。為了評估我們提出的方法，我們從原始的訓練集中生成了一些模擬雜訊的訓練資料，並使用各種雜訊資料集在不同的訊雜比(SNR)及不同雜訊種類之狀態下設計了四種類型的測試集。

為了進行訓練，我們採用的雜訊來自 MUSAN[14]、DEMAND[15]、QUT-NOISE[16]和環境背景雜訊資料集[17][18]，總共 2553 種雜訊，乾淨語音和雜訊音檔皆重新採樣至 16kHz。我們將這些雜訊以 5dB 之 SNR 值混入原始訓練資料中的每個語句，合成與原資料相同(即 SNR 5 dB 的 120098 個帶噪語句)的多條件訓練(Multi-condition Training, MCT)資料(以下稱為”MCT 資料”)。

在測試資料的準備上，我們將原始測試集語音及上述之雜訊，根據三種 SNR 值設定: -5 dB、5 dB 與 15 dB 加以混合，分別建立了三個額外的測試集。另外，為模擬測試於未知雜訊的情形，我們使用了不同的雜訊資料集: Nonspeech 雜訊資料集[19]，以 SNR 5dB 來合成第四種測試集。

(一) 語音增強模型設置

針對語音增強系統，我們在原始 AISHELL-1 訓練集和發展集中，每個語者隨機挑出 20 則語句(分別為 6800 及 800 則語句)來訓練我們的語音增強系統，所有資料均使用前述

MCT 雜訊以 5dB 之 SNR 加以混合。訓練的所有語音增強模型只輸出乾淨語音，訓練皆不採用語音分離問題[7]中的置換不變訓練(Permutation invariant Training, PIT)，且損失函數皆為負比例無關訊噪比(Scale-invariant Signal-to-noise ratio, SISNR)，相關公式如下：

$$\mathbf{s}_{target} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \quad (10)$$

$$\mathbf{e}_{noise} = \hat{\mathbf{s}} - \mathbf{s}_{target} \quad (11)$$

$$SISNR = 10 \log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2} \quad (12)$$

$$\mathcal{L}_{SISNR} = -SISNR(\mathbf{s}, \hat{\mathbf{s}}) \quad (13)$$

其中 $\hat{\mathbf{s}} \in \mathbb{R}^{1 \times T}$ 、 $\mathbf{s} \in \mathbb{R}^{1 \times T}$ ，分別為預測語音訊號向量及乾淨語音訊號向量， T 為訊號之長度， \mathbf{s}_{target} 為 $\hat{\mathbf{s}}$ 於 \mathbf{s} 之投影向量。

我們考慮了三種不同特徵的語音增強模型，所有模型遮罩估計網路皆使用時間摺積網路(Temporal Convolution Network, TCN)[7]，包含殘差連結(Residual-connection)以及跨層連結(Skip-connection)，可見圖一。相關之超參數 (hyperparameter) 設置為 $X = 8$ 、 $R = 3$ 、 $B = 128$ 、 $H = 512$ 、 $S = 128$ 、 $P = 3$ ，為[7]中最佳的模型設置，不同之處在於編碼器與解碼器之結構：

1. STFT-TCN

在頻域 SE 模型設定上，我們使用 STFT 以及 iSTFT 作為編碼器與解碼器，藉由 STFT 我們得到複數頻譜(Complex-valued Spectrogram)作為頻域特徵，特徵的前半部 $(1, \dots, N/2)$ 表示實數(Real Value)部分，後半部 $((N/2 + 1, \dots, N)$ 表示虛數(Imaginary Value)部分，與[11]相同。我們設置傅立葉轉換點數為 512 點，窗函數為漢寧窗(Hanning Window)，取窗長為 64 (個樣本數)，窗移為 32 (個樣本數)擷取特徵，得到 512 維之頻域特徵。

2. Denoising-TasNet(M)

在時域 SE 模型設定上，我們參考[10]提出 Denoising-TasNet 之架構，使用一維摺積層及轉置摺積層作為編碼器與解碼器。且根據[7]，在提取時域特徵時使用較小的窗長可以有較佳的效果，因此我們設置窗長為 16 (個樣本數)，窗移為 8 (個樣本數)擷取特徵，濾波器數為 512，最後得到 512 維之時域特徵。由於這裡使用的模型較原[16]提出的模型，多了跨層連結且有更好的效果，我們將其稱之為 Denoising-

TasNet(M)。

3. Multi-view-TCN

在我們新提出之多視角注意力機制語音增強模型中，我們使用了多視角注意力機制編碼器及 1-D 轉置摺積層作為解碼器，取 256 維的頻域特徵與 256 維的時域特徵進行融合，為了固定特徵之總音框數，我們將窗長及窗移皆設置為 16 (個樣本數)以及 8 (個樣本數)，映射層之維度設置為 128，最後得到 128 維的融合特徵。

(二) 語音辨識模型設置

於後端語音辨識系統，我們使用 Kaldi 建構 Hybrid DNN-HMM 聲學模型，遵循[18]提供的 GMM-HMM 訓練流程。在 DNN 模型的訓練，我們堆疊分解式時延神經網絡(Factorized Time Delay Neural Network, TDNN-F)，採取詞圖無關最大交互資訊(Lattice-free Maximum Mutual Information, LF-MMI)目標函數進行訓練，其可在原官方提供之測試集獲得更好的結果(7.46%及 6.51%之字錯誤率, CER)，我們將此系統作為我們實驗的基線(Baseline)。另外，我們訓練了兩種聲學模型(AM)進行比較，一種訓練資料包含原始訓練資料和 MCT 資料，為多情境訓練之 AM，以 MCT-AM 表示；另一種訓練資料包含原始訓練資料，MCT 資料和使用對應 SE 模型增強 MCT 資料得到的 ENH 資料，以彌除增強後的特徵與模型不匹配的問題，為重新訓練(Retraining)之 AM，以 ENH-AM 表示。

五、實驗結果及分析

我們使用 SISNR 作為 SE 之評估方式，見式(15)，單位為 dB 其值越高越好；於 ASR，我們則採用字錯誤率(Character Error Rate, CER)作為評估以百分比表示，值為越低越好。

(一) 不同注意力機制之比較

表二、不同注意力機制之比較

AM Model	SE Model	Attention Mechanism	Test (5dB)	
			CER	SISNR
Baseline	—	—	43.81	5.02
MCT-AM	—	—	18.75	5.02
MCT-AM	Multi-view-TCN	Additive	17.05	14.81
		Concatenate	16.92	14.88
		Scaled dot-product	16.49	14.91

我們首先針對提出的 Multi-view-TCN 模型，比較不同注意力機制的效果，於此實驗我們測試在已知雜訊為 5dB SNR 之測試集，SE 法只作用於測試集，實驗數據如表二所示。由此表之第一、二列可以發現，相較於基線結果，使用 MCT-AM，在帶噪的環境中可以大幅地降低 CER。此外，根據第二至四列之數據可知，在不重新訓練聲學模型的前提下，所新提出之 Multi-view-TCN 法在三種不同注意力機制下，不僅提升語音增強指標 SISNR，同時也增加了語音辨識率(即 CER 降低)。其中，以伸縮點積(Scaled dot-product)注意力機制效果最為顯著，因此，我們在此後的實驗，於 Multi-view-TCN 模型上皆使用伸縮點積注意力機制。

(二) 已知雜訊環境之語音辨識結果

在第二部分的實驗中，我們測試各模型在已知雜訊且不同訊噪比的環境，分別針對語音增強及語音辨識的表現進行探討，實驗結果請見表三。

表三、已知雜訊環境之語音辨識結果

AM Model	SE Model	Test (-5dB)		Test (5dB)		Test (15dB)	
		CER	SISNR	CER	SISNR	CER	SISNR
Baseline	—	81.41	-4.94	43.81	5.02	15.14	15.02
MCT-AM	—	58.19	-4.94	18.75	5.02	8.66	15.02
	STFT-TCN	49.78	5.43	18.92	13.71	9.42	18.73
	Denoising-TasNet(M)	47.47	4.52	16.99	14.69	9.14	19.52
	Multi-view-TCN	46.62	5.01	16.49	14.91	8.90	19.83

根據表三之數據，在語音增強的效果上，所提出的 Multi-view-TCN 模型相較於其他兩方法幾乎都可得到更佳的 SISNR 值，唯在 -5dB 之 SNR 的測試集較基於頻域的方法 (STFT-TCN) 差；在語音辨識效能上，Multi-view-TCN 相較其他兩方法皆得到更高的辨識精確率，由此可知，SISNR 指標的提升雖不能完全反映在 CER 的下降，但大致有一致的現象。然而，所有模型在 15dB 之 SNR 環境下，語音辨識表現皆比原始 MCT-AM 差，我們猜測在雜訊干擾較少的環境，語音透過語音增強模型會產生較多額外的失真，導致前後端較顯著之不匹配現象。因此我們採取重新訓練的方法，使用各模型對原 MCT 資料增強後加入訓練資料，重新訓練聲學模型，進行後續的實驗。

(三) 重新訓練聲學模型之結果

表四、重新訓練聲學模型之結果

AM Model	SE Model	Test (-5dB)	Test (5dB)	Test (15dB)
		CER		
MCT-AM	—	58.19	18.75	8.66
MCT-AM	STFT-TCN	49.78	18.92	9.42
ENH-AM		40.20	12.69	7.81
MCT-AM	Denoising-TasNet(M)	47.47	16.99	9.14
ENH-AM		40.41	12.02	7.79
MCT-AM	Multi-view-TCN	46.62	16.49	8.90
ENH-AM		39.03	11.71	7.66

根據表四可發現，在重新訓練聲學模型後，在不同訊噪比之測試環境下皆有顯著進步的語音辨識率，且在高訊噪比的環境(SNR 15dB)，所有模型皆比 MCT-AM 佳，因此驗證了重新訓練聲學模型的方法，可以彌除前後端不匹配的現象。而所新提出的 Multi-view-TCN 在各測試集皆表現最佳，相較於 Denoising-TasNet(M)法，在-5dB、5dB 及 15dB 三種 SNR 測試環境下可獲得 3.4%、2.5%及 1.6%相對字錯誤率下降。

(四) 未知雜訊環境之語音辨識結果

表五、未知雜訊環境之語音辨識結果

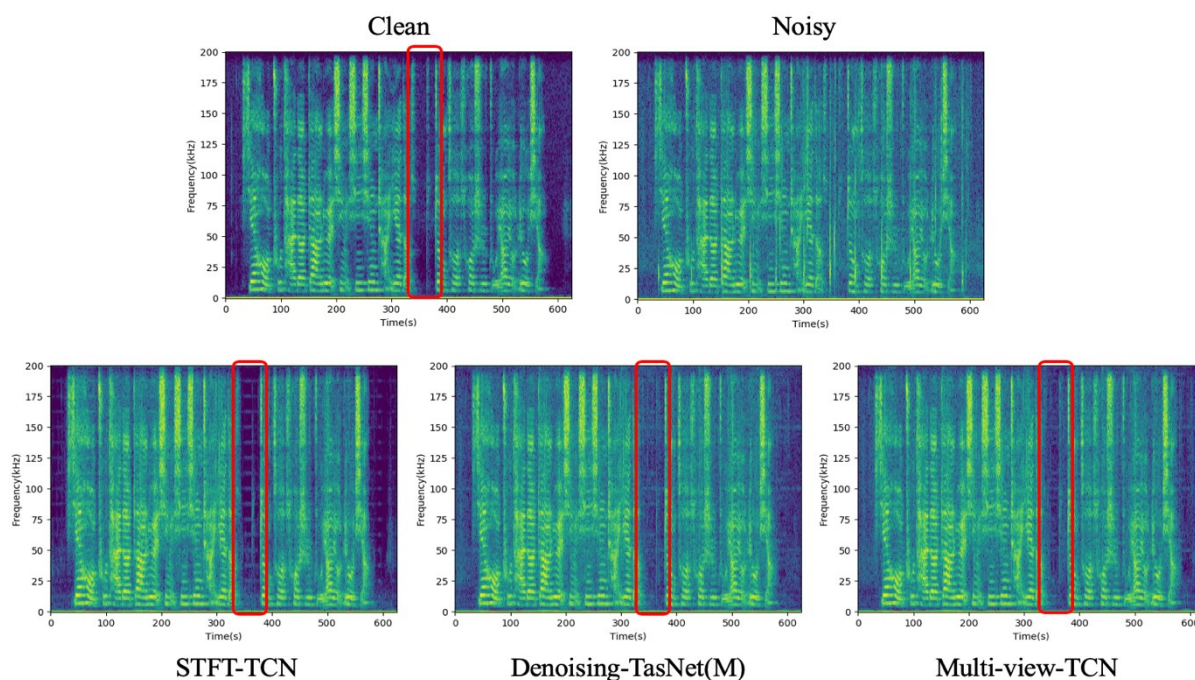
AM Model	SE Model	Test (unseen) (-5dB)		Test (unseen) (5dB)		Test (unseen) (15dB)	
		CER	SISNR	CER	SISNR	CER	SISNR
Baseline	—	88.42	-4.97	50.58	5.01	16.92	15.01
MCT-AM	—	70.90	-4.97	25.37	5.01	10.45	15.01
MCT-AM	STFT-TCN	60.56	5.98	25.86	13.79	11.80	19.09
ENH-AM		47.40	5.98	15.51	13.79	8.62	19.09
MCT-AM	Denoising-TasNet(M)	56.88	5.73	22.82	14.54	11.44	19.96
ENH-AM		47.31	5.73	14.88	14.54	8.63	19.96
MCT-AM	Multi-view-TCN	55.38	6.10	22.08	14.71	10.92	20.15
ENH-AM		45.49	6.10	14.16	14.71	8.44	20.15

表五為未知雜訊環境的實驗結果，從此表我們可以觀察到，在雜訊未知的環境下，各方法於語音辨識的表現與前述實驗有一致的趨勢，且在低訊噪比環境(-5dB SNR)時，新方

法其表現皆優於頻域的 STFT-TCN 法及時域的 Denoising-TasNet(M)法，因此，所新提出的 Multi-view-TCN，展現了在未知雜訊的環境下之泛化(Generalization)的能力，在 SISNR 及 CER 的評估上皆得到最好的結果，相較於 Denoising-TasNet(M)，在 SNR 為 -5dB、5dB 與 15dB 之雜訊環境下，可得到 3.8%、4.8%及 2.2%的相對字錯誤率下降率。

(五) 語音增強效果比較

最後，我們比較不同系統在語音增強上的效果，其音檔範例可見¹。另外，我們比較不同系統於同一音檔增強後的時頻圖(Spectrogram)，如圖二所示。可以觀察紅色粗體方框的範圍中，頻域之 STFT-TCN 法會殘留許多雜訊頻譜，並將某些原音框中乾淨語音的頻譜一併消除；而基於時域的模型 Denoising-TasNet(M)，雖保留了較多乾淨頻譜成分，但同時也包含了許多雜訊頻譜成分；Multi-view-TCN，則似乎是在兩者之中權衡，因此獲得了比較好的增強效果。



圖二、語音增強系統之時頻圖比較 (BAC009S0764W0193, 腳步聲, 5dB SNR)

六、結論

在本研究中，我們提出了多視角注意力機制語音增強模型於強健語音辨識，同時考慮了頻域特徵(複數時頻圖)以及時域的特徵，透過注意力機制將兩者特徵融合為單一特徵，

¹ https://smildemo.csie.ntnu.edu.tw/rocling_demo/index.html

藉由 Aishell-1 開源中文語音語料庫的評估實驗，透過使用各種不同的雜訊源，模擬不同的雜訊情形作為訓練及測試，充分顯示此新方法皆優於基於時域的語音增強方法。而在已知雜訊的測試集中，於 -5dB、5dB、15dB 三種 SNR 環境，使用重新訓練之聲學模型，相對於時域上的方法，新方法可分別下降相對字錯誤率 3.4%、2.5%及 1.6%；而在未知雜訊的測試集，則獲得相對字錯誤率 3.8%、4.8%及 2.2%的下降。

從實驗結果而論，於時域處理的語音增強模型，不僅對於語音增強效果優異，同時也較適用於後端語音辨識模型。於未來，我們將專注在時域特徵之研究，並考慮更多語音特徵應用在所提出之基於注意力機制之特徵融合方法，並將問題進一步延伸至處理摺積型雜訊(Convolution Noise)，如混響(Reverberation)之干擾等。

參考文獻

- [1] N. Wiener, “Extrapolation, Interpolation, and Smoothing of Stationary Time Series”, *New York: WILEY*, 1949.
- [2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. INTERSPEECH*, pp. 436–440, 2013.
- [3] Y. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381–1390, 2013.
- [4] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. Hershey, and B. Schuller, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *Proc. LVA/ICA*, pp. 91–99, 2015.
- [5] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.
- [6] D. Rethage, J. Pons, and X. Serra, “A Wavenet for speech denoising,” in *Proc. ICASSP*, pp. 5069–5073, 2018

- [7] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [8] T. Menne, R. Schlüter and H. Ney, “Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR,” *arXiv preprint arXiv:1904.09049*, 2019.
- [9] K. Tan and D. Wang, “Improving robustness of deep learning based monaural speech enhancement against processing artifacts,” in *Proc. ICASSP*, 2020.
- [10] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, “Improving noise robust automatic speech recognition with single-channel time-domain enhancement network,” in *Proc. ICASSP*, pp. 7009–7013, 2020
- [11] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, “Exploring the best loss function for dnn-based low-latency speech enhancement with temporal convolutional networks,” *arXiv preprint arXiv:2005.11611*, 2020.
- [12] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AISHELL1: An open-source Mandarin speech corpus and a speech recognition baseline,” in *Proc. O-COCOSDA*, pp. 1–5, 2017
- [13] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “Phasen: A phase-and harmonics-aware speech enhancement network,” *arXiv:1911.04697*, 2019.
- [14] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.
- [15] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.
- [16] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, “The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms,” in *Proc. INTERSPEECH*, pp. 3110–3113, 2010

- [17] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, “Smart phone-based real-time classification of noise signals using subband features and random forest classifier,” in *Proc. ICASSP*, pp. 2204–2208, 2016
- [18] F. Saki and N. Kehtarnavaz, “Automatic switching between noise classification and speech enhancement for hearing aid devices,” in *Proc. EMBC*, pp. 736–739, 2016
- [19] G. Hu and D. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 2067–2079, 2010.

使用元學習技術於語碼轉換語音辨識之初步研究

A Preliminary Study on Leveraging Meta Learning Technique for Code-switching Speech Recognition

余福浩 Fu-Hao Yu

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

M10815004@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

語碼轉換(Code-switching)的語音辨識近年來愈來愈受到研究學者的重視，雖然人們在日常生活中使用語碼轉換的情形逐漸增加，但是可以用來訓練語碼轉換之語音辨識器的語音資料，相較於主流的單語言（如：英語或中文）語料，更是少之又少，這是因為語音語料的標註(Labeling)相當耗時費工。為了提升語碼轉換之語音辨識器的效能，在本研究中，我們提出利用近期逐漸興起的元學習(Meta Learning)方法，希望可以利用資料量較多的單語言語料，提升語碼轉換的語音辨識任務之成效。實驗中，我們採用 SEAME (South East Asia Mandarin-English)資料集，透過元學習中的與模型無關之元學習法(Model-Agnostic Meta-Learning, MAML)進行語音辨識器訓練，實驗結果證明，模型能夠快速適應於最終的語碼轉換語音辨識任務。

關鍵詞：語碼轉換，語音辨識，元學習

Abstract

In recent years, code-switching speech recognition has become an important research topic. Code-switching in conversation speech is gradually increasing in our daily lives. However, compared with monolingual languages (e.g., English or Chinese), only a few resources can be obtained for training a code-switch speech recognizer. To mitigate the deficiency, in this paper, we propose a meta-learning approach for code-switching speech recognition. In other words, following the model-agnostic meta-learning (MAML) procedure, we first train the speech recognizer by using monolingual corpora, and then a fine-tune stage is performed to obtain the final code-switching speech recognizer by using code-switching data. We evaluate the proposed method on the SEAME (South East Asia Mandarin-English) dataset. A series of experiments show that the meta-learning method can improve the performance of the low-resource code-switching speech recognition task.

Keywords: Code-switching, Speech Recognition, Meta-learning

一、緒論

語音辨識中語碼轉換(Code-switching)的議題目前愈來愈受到重視。在全球化的基礎之下，人類學習多種語言的情形開始蔚為風潮，如非英語系國家通常會學習英語或學習使用頻率較高的主流語言作為第二外語來使用，在這樣的時空背景下造成人們在說話時產生語碼轉換的可能性日益增加，以亞洲地區為例，於許多不同的場景中，人們在對話或溝通時開始產生容易在中文間穿插英語等外語的情形，從大專院校校園甚至到工作職場中，這樣的狀況逐漸普及，因此含有語碼轉換的語音辨識在自然語言領域開始成為一個重要的問題。

近年來，由於深度神經網路的興起，深度學習在影像辨識任務上獲得了非常優秀的成績，辨識準確率甚至能夠超越人類，因此研究學者們開始傾向使用深度學習來進行機器翻譯和語音辨識等自然語言方面的任務[1, 2, 3, 4, 5]，研究成果顯示，基於深度學習的模型可以獲得比傳統模型還要傑出的成績，因而此類方法逐漸變成了近年自然語言研究的主流與趨勢。

在使用深度學習方法來訓練模型時，通常必須藉由大量的訓練資料進行訓練，才能夠成功訓練出良好的模型，若模型的訓練資料過於稀少，則容易產生擬合不足

(Underfitting)的問題，使模型在訓練集和測試集的表現都不夠良好。在語音辨識的任務中，含有語碼轉換的語音資料相當稀少，這使得要訓練一個成功的語碼轉換語音辨識器變得不太容易，因此有許多研究開始將含有語碼轉換的語音辨識視為低資源(Low-resource)的任務，嘗試用不同的訓練方式或模型方法來解決這個問題。有鑑於此，本論文提出使用元學習(Meta Learning)的方式來進行含有語碼轉換的語音辨識，並嘗試使用目前主流的元學習方法中與模型無關之元學習法(Model-Agnostic Meta-Learning, MAML)[8]於語碼轉換之語音辨識，期望可以提升語音辨識的準確率。

二、相關研究

(一) 元學習(Meta Learning)

元學習(Meta Learning)是近年來逐漸興起的一種深度學習方式，元學習的理論背景建立在希望所訓練出的深度學習模型，能夠有如同人類的學習行為，人類對於學習一件新事物的能力很強，通常不用像深度學習模型一樣要看過大量學習資料才能做得很好，例如可以只看過一幅特定畫家的畫作就能大致判斷大多數的畫作是否出自於此畫家，或是只須看過一個器物的樣貌，就能夠成功辨認出相同器物的不同樣式或形狀。元學習理論認為人類之所以能夠學習得這麼快速，是因為已經累積了很多先前學習的經驗，所以才能夠達到快速學習(Rapid Learning)的能力。因此，在元學習理論中，為了使模型能過獲得快速學習的能力，將透過額外的訓練資料產生不同的訓練任務(Task)，讓模型「學習如何學習(Learning-to-learn)」，使模型在面對目標任務時並不是從頭開始學起，而是有了過往學習的經驗或知識，也就是具備了一定的先驗知識，並學會了「如何學習」的經驗與技巧，成為了更厲害的學習者，藉由過往的學習經驗，深度學習模型在未來遇到目標任務時，便可以利用少量的資料達到快速學習的成果。

近期由於元學習逐漸受到重視，在研究方面發展出三大不同的主流方法，分別為：以含有記憶性功能的神經網路，如使用具有長短期記憶遞迴神經網路[18]、神經網路圖靈機(Neural Turing Machines, NTM)[19]來實踐的黑箱適應方法(Black-box Adaptation)[6, 7]；以及基於最佳化(Optimization-based)的元學習方法[8, 9]，像是與模型無關之元學習(Model-Agnostic Meta-Learning, MAML)[8]和可擴展的元學習演算法(Reptile)[9]等方法；最後則是基於測度(Metric-based)理論的非參數化(Non-parametric)方法，例如著名的孿生網路(Siamese Network)[10]等。其中，與模型無關之元學習方法更成為目前最主流的方法。

法之一，與模型無關之元學習方法屬於一種基於最佳化方式來達成元學習概念的方法。元學習旨在希望能夠訓練出一個模型成為好的學習者（即元學習者(Meta-learner)），在各種學習任務下只需要利用少量的訓練資料，就可以快速地解決或適應一個新的學習任務，也就是希望我們訓練出的元學習者能夠像是人類一樣，能夠在少量的訓練下達成快速學習的目標；而與模型無關之元學習演算法便發展出了一種與模型無關的元學習方式，由於他在演算法的設計上是完全與模型無關，因此我們便能夠使用梯度下降的訓練方式，直接地應用在任何一種學習問題和模型上，例如：分類(Classification)、迴歸(Regression)和強化學習(Reinforcement Learning)等常見的不同任務中，和以往發展的元學習方法之不同點在於與模型無關之元學習方法不會增加模型的參數量，也不需要限制模型的架構，因此模型不受限於各種遞迴神經網路，也可以和全連接(Fully-connected)神經網路與卷積神經網路(Convolutional Neural Network, CNN)等不同神經網路進行組合。透過與模型無關之元學習演算法的訓練，可以使訓練後的模型在給定一個新的學習任務時，能夠快速地適應於新任務且在新任務上能有較好的結果。

（二）語碼轉換語音辨識模型之現況

傳統上常用來進行語碼轉換語音辨識任務的模型有像是基於高斯混合模型結合隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)[20]或是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)的語音辨識器 [15, 16]，而近年來由於端對端(End-to-end)的語音辨識器成為研究主流，因此近期亦有基於端到端的語碼轉換語音辨識器模型，例如基於 CTC 以及注意力機制的混合模型(Hybrid CTC-Attention based Models)[12, 17]，並嘗試結合加入語言辨識(Language Identification, LID)進行多任務學習(Multitask Learning)的訓練策略，來改善語碼轉換語音辨識的準確度。

三、方法

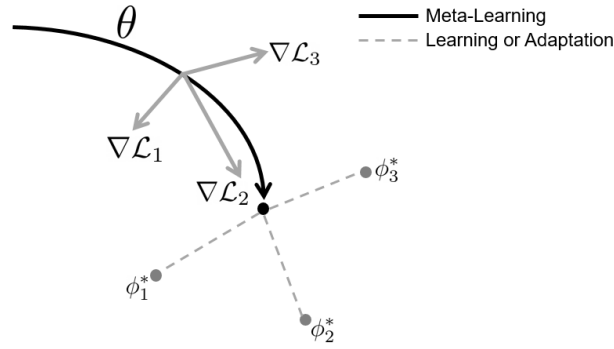
由於目前並不容易取得擁有大量訓練資料的語碼轉換語音資料集，這使得基於深度學習的語音辨識模型難以在語碼轉換的任務上有良好的表現。所幸的是，近期開始興盛起的元學習方法，目標便是使得訓練出的模型能如同人類一般，在少量的資料下能夠做到快速學習的學習行為，也正因為如此元學習方法十分適合使用在訓練資料不多、低資源(Low-resource)的學習任務[13, 14]。在語碼轉換的語音辨識任務上，我們認為在宏觀的

概念上，可以將含有語碼轉換的語音視為一種新的語言，若我們將它當成一種新的語言來看待，當然也就能夠將語碼轉換這個語言和單語言當作不同的訓練任務，基於這個想法，我們就可以利用元學習的概念，希望藉由單語言語料訓練出的模型，能夠快速地適應於目標任務，也就是含有語碼轉換的語音辨識任務上。有鑑於此，本論文提出使用元學習中，與模型無關之元學習來進行語音辨識器的訓練，期望在單語言語料的訓練下，可以獲得一個更好的語碼轉換語音辨識器，進一步地提升語碼轉換語音辨識的準確率。

(一) 與模型無關之元學習(Model-Agnostic Meta-Learning, MAML)

在與模型無關的元學習方法中[8]，我們首先定義模型為 f ，負責進行預測任務，在給定輸入資料 x 即可映射到輸出資料 y ，與模型無關的元學習演算法和傳統元學習的概念相同，訓練過程可分為元訓練(Meta-training)與元測試(Meta-testing)兩大部分，在進行元訓練時，可將原本的訓練資料集 \mathcal{D} 經過隨機抽樣產生許多不同的子資料集 $\mathcal{D}_i = \{(x_1, y_1), \dots, (x_n, y_n)\}$ ，此時每一個子資料集都可以切分成訓練資料(或稱為支撐集(Support Set)) $\mathcal{D}_i^{tr} = \{(x_1, y_1), \dots, (x_k, y_k)\}$ 以及測試資料(或稱為查詢集(Query Set)) $\mathcal{D}_i^{ts} = \{(x_1, y_1), \dots, (x_l, y_l)\}$ ，此時便可以將每一個子資料集視為一個不同的任務 \mathcal{T}_i ，也就是說 $\mathcal{D}_{meta-train} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\} = \{(\mathcal{D}_1^{tr}, \mathcal{D}_1^{ts}), \dots, (\mathcal{D}_n^{tr}, \mathcal{D}_n^{ts})\} = \{\mathcal{T}_1, \dots, \mathcal{T}_n\}$ ，此外若是每一個任務的訓練資料集 \mathcal{D}_i^{tr} 都有 k 筆訓練資料，我們也可稱之為 K 樣本學習(K -shot Learning)。而在元測試時也可定義出 $\mathcal{D}_{meta-test} = (\mathcal{D}^{tr}, \mathcal{D}^{ts})$ ，也就是我們想適應於目的領域或目的任務的訓練與測試資料，我們會利用其中的訓練資料 \mathcal{D}^{tr} 訓練模型，使模型能夠快速適應在新的目標任務中，最後便可以將模型應用於新任務的測試資料中進行預測，以達到比較好的預測結果。

更明確地，我們希望能夠在元訓練資料集 $\mathcal{D}_{meta-train}$ 上找到一個較好的元參數 θ ，利用此參數在新的任務上得到一個好的模型參數 ϕ ，在與模型無關的元學習演算法中，我們將模型的初始化參數視為元參數 θ ，將模型記為 f_θ ，當要訓練一個新任務 \mathcal{T}_i 時，透過訓練資料與誤差函數 \mathcal{L} 計算出的損失 $\mathcal{L}_{\mathcal{T}_i}$ ，可以利用梯度下降的方式來更新模型的參數(通常只進行一次，但也可直觀地擴展為多次更新)，在與模型無關的元學習方法中，經過參數更新適應於新任務的模型參數即為 ϕ_i ，也就是說 $\phi_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ ，其中 α 為梯度下降的可調整參數。與模型無關的元學習方法希望模型的初始化參數 θ 更新變為



圖一、與模型無關的元學習方法參數優化路徑示意圖。

ϕ_i 後，能夠在新任務 \mathcal{T}_i 上有良好的表現，因此我們可以訂定出元學習的元目標函數(Meta Objective Function)為：

$$\min_{\theta} \sum_i \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i}) = \sum_i \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})}) \quad \text{式 (1)}$$

意即希望能找到一個最好的初始化參數，使得此參數在新任務上只需透過一次梯度下降進行參數更新，就能在新任務上有最好的表現，也就是使損失最小化。此外，要優化元學習目標函數同樣是一個最佳化問題，我們可以再次利用梯度下降的方式對元學習的目標函數進行最佳化（又稱為元最佳化(Meta Optimization)），代表元參數 θ 可依照公式 $\theta \leftarrow \theta - \beta \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$ 進行更新，其中 β 亦為梯度下降的可調整參數。與模型無關的元學習方法的優化過程可以參考圖一之視覺化示意圖。

完整的與模型無關的元學習方法之演算法可參考以下演算法 1 之虛擬碼，值得一提的是，與模型無關的元學習方法如同上述介紹會有兩種最佳化，第一種最佳化於演算法 1 中第 6 行，是負責進行讓模型適應於新任務之中的最佳化，又稱為內層迴圈最佳化 (Inner Loop Optimization)；第二種最佳化於演算法 1 中第 8 行，則是負責讓模型能夠快速適應於新任務，使模型能成為一個更好的學習者的最佳化，又稱為外層迴圈最佳化 (Outer Loop Optimization)。

雖然與模型無關的元學習演算法不會增加模型的參數量，但是卻有著在進行外層迴圈最佳化時需要計算 $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$ 的缺點，而 $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$ 的計算則是牽涉到需要對損失函數進行二階微分(Second-order Derivatives)，要在高維度的向量空間中進行二階微分則必須用 損失建立起其完整的黑塞矩陣(Hessian Matrix)，代表在進行損失函數優化時計算量

演算法 1 : Model-Agnostic Meta-Learning (MAML)

輸入： α, β

- 1: 隨機初始化模型參數 θ
 - 2: **while** 訓練尚未完成 **do**
 - 3: 產生一個（或多個）新任務 \mathcal{T}_i
 - 4: **for all** \mathcal{T}_i **do**
 - 5: 利用 \mathcal{D}_i^{tr} 計算 $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$
 - 6: 計算參數 $\phi_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$
 - 7: **end for**
 - 8: 更新模型初始化參數 $\theta \leftarrow \theta - \beta \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$
 - 9: **end while**
-

演算法 1、與模型無關的元學習方法之演算法。

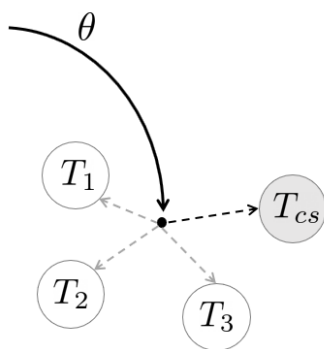
將會大幅提升，並消耗更多記憶體空間，使得模型的訓練速度大幅下降。為了避免計算出完整的黑塞矩陣，我們也可考慮利用一階近似(First-order Approximation)的方式來計算出 $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$ 的近似值，以對演算法加速。在原本的梯度計算中，由於 $\phi = \theta - \alpha \nabla_{\theta} \mathcal{L}(f_{\theta})$ ，我們可先利用連鎖律對其展開，接著則會出現一項 $\nabla_{\theta}(\theta - \alpha \nabla_{\theta} \mathcal{L}(f_{\theta}))$ 的梯度計算，由於 $0 < \alpha < 1$ 使 $\alpha \nabla_{\theta} \mathcal{L}(f_{\theta})$ 很接近 0，此項會十分接近 $\nabla_{\theta}(\theta)$ ，因此可以選擇忽略計算此項內部的梯度，直接以 $\nabla_{\theta}(\theta)$ 進行近似，而又因 $\nabla_{\theta}(\theta) = I$ 所以經過一階近似後可得 $\nabla_{\theta} \mathcal{L}(f_{\phi}) \approx \nabla_{\phi} \mathcal{L}(f_{\phi})$ ：

$$\begin{aligned} \nabla_{\theta} \mathcal{L}(f_{\phi}) &= \left(\nabla_{\phi} \mathcal{L}(f_{\phi}) \right) (\nabla_{\theta} \phi) \\ &= \left(\nabla_{\phi} \mathcal{L}(f_{\phi}) \right) \left(\nabla_{\theta} (\theta - \alpha \nabla_{\theta} \mathcal{L}(f_{\theta})) \right) && \text{式 (2)} \\ &\approx \left(\nabla_{\phi} \mathcal{L}(f_{\phi}) \right) (\nabla_{\theta}(\theta)) = \nabla_{\phi} \mathcal{L}(f_{\phi}) \end{aligned}$$

利用一階近似方法實作的與模型無關的元學習演算法又特稱為一階與模型無關的元學習演算法(First-order MAML, FOMAML)，即在計算上以 $\theta \leftarrow \theta - \beta \nabla_{\phi_i} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$ 取代原本於演算法 1 中第 8 行的更新式，即可達成一階近似[8]。

(二) 基於元學習的語碼轉換語音辨識系統

由於希望能夠藉由使用元學習技術改善語碼轉換任務的語音辨識準確度，我們將語碼轉換的語音資料作為目標任務，即用來進行元測試階段的訓練，而訓練集中僅含有中英文之單語言資料則用於元訓練階段的訓練和測試中。在進行元訓練階段時，每一次都將從



圖二、使用元學習的概念於語碼轉換語音辨識的任務 T_{cs} 。

單語言資料中隨機抽取出部分資料，於元學習中可將其視為一個新的語音辨識任務，接著使用與模型無關之元學習方法來訓練、更新模型，則可使模型學習到要如何快速適應於一個新的語音辨識任務之中。最後在元測試階段，則使用含有語碼轉換情形的語料對模型進行訓練，使其能夠適應於我們想解決之目標任務，由於模型已經學會如何快速適應於不同的語音辨識任務，因此就算語碼轉換的訓練資料不如單語言語音資料多，模型依然能夠快速適應於語碼轉換的語音辨識任務，且能夠有較好的表現。圖二為使用元學習的概念於語碼轉換語音辨識系統的示意圖。

四、實驗

為了瞭解使用元學習方法是否能在語碼轉換的資料集上帶來改善語音辨識的效果，我們採用 LAS(Listen, Attend and Spell)[4]架構作為語音辨識器，並使用與模型無關之元學習演算法的方式進行元學習實驗，並與傳統語音辨識系統的結果進行比較。

(一) 資料集

我們使用 SEAME (South East Asia Mandarin-English)[11, 12]作為實驗的資料集，由於想了解元學習對語碼轉換語音辨識是否帶來改進，我們將資料集中的訓練資料進行分類，根據語音資料文字中是否為語碼轉換資料來將訓練資料分成兩大部分，分別為只有中文或只有英文的單語言資料以及含有中英文夾雜的語碼轉換資料，測試集則使用 SEAME 中的 dev_man 與 dev_sge 資料合併進行測試，相關統計資訊如表一所示。最後，我們使用字錯誤率(Character Error Rate, CER)與詞錯誤率(Word Error Rate, WER)作為評估標準。

	小時數	小時(比例)		
		純中文	純英文	語碼轉換
<i>train</i>	101.13	16.18 (16%)	16.18 (16%)	68.76 (68%)
<i>dev_{man}</i>	7.49	1.04 (14%)	0.52 (7%)	5.91 (79%)
<i>dev_{sgc}</i>	3.93	0.23 (6%)	1.61 (41%)	2.08 (53%)

表一、SEAME 資料集統計資訊。

Model	CER	WER
LAS	55.0%	63.4%
LAS+MAML	49.7%	58.9%

表二、實驗結果。

(二) 語音辨識模型

實驗時我們採用 Google 於 2015 年提出之 LAS(Listen, Attend and Spell, LAS)模型[4]作為語音辨識器使用，LAS 模型由金字塔式堆疊的雙向長短期記憶(Long Short-term Memory)作為編碼器(Encoder)，以及含有注意力(Attention)機制的解碼器(Decoder)組合而成，結合以上兩種特殊機制在語音辨識上取得了非常好的結果。實驗中 LAS 模型依照以下設定進行設置，在編碼器中使用三層雙向各 180 維的長短期記憶，在解碼器中使用兩層 360 維的長短期記憶，詞嵌入(Word Embedding)的維度大小為 180 維，以 Uniform Distribution [-0.1,0.1]進行模型初始化，其餘設置參考原模型設定。

(三) 實驗結果

實驗中，基礎系統使用 LAS 模型訓練於僅有中文或英文之單語言資料 5 個世代(Epoch)後，再更新(Fine-tune)於語碼轉換的資料 5 個世代，最後的字錯誤率約為 55.0%、詞錯誤率為 63.4%。當我們將元學習方法運用於語碼轉換之語音辨識系統時，首先是使用單語言資料作為元訓練資料集 $\mathcal{D}_{meta-train}$ ，每一次都將隨機從元訓練集中隨機取出 8 筆資料作為訓練資料以及 8 筆資料作為測試資料，外層迴圈使用學習率 0.001 的適應性矩估計演算法(Adam)進行優化，內層迴圈使用學習率 0.1 之隨機梯度下降法(Stochastic Gradient Descent, SGD)進行優化，之後則將語碼轉換的資料作為元測試 $\mathcal{D}_{meta-test}$ 中的訓練資料並訓練 5 個世代，讓模型快速適應於語碼轉換的目標任務，最後在測試集上進行測試，最後的字錯誤率約為 49.7%、詞錯誤率為 58.9%。最終實驗結果如表二所示。

四、結論

本論文提出以元學習方式改善語碼轉換語音辨識的方法，並使用近期元學習中的主流方法與模型無關之元學習方法進行語音辨識器的訓練，由於目前語碼轉換的訓練資料稀少，以元學習的方式進行學習，可以在相同的訓練資料集之下，訓練出更好的語音辨識器，於實驗中我們所提出的方法，可以在字錯誤率與詞錯誤率上獲得改善，在這樣的研究結果下，我們為語碼轉換語音辨識提供了一種新的解決方法，未來，我們將繼續這個研究方向，期望可以結合其他單語言資料集或是對資料進行資料擴增(Data Augmentation)，以及發展出一套專屬於語碼轉換或語音辨識器的元學習演算法，為語音辨識任務提供一個新的方向與效能的提升！

致謝

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 109-2636-E-011-007 (Young Scholar Fellowship Program), and by the Project K367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

參考文獻

- [1] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.
- [2] A. Graves, “Sequence transduction with recurrent neural networks,” in *ICML Representation Learning Worksop*, 2012.
- [3] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-Based Models for Speech Recognition,” in *Neural Information Processing Systems*, 2015.
- [4] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 4960-4964, doi: 10.1109/ICASSP.2016.7472621.

- [5] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4835–4839.
- [6] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1842–1850.
- [7] T. Munkhdalai and H. Yu, “Meta networks,” in *Proc. ICML*, 2017, pp. 2554–2563.
- [8] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” *arXiv preprint arXiv:1703.03400*, 2017.
- [9] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” *arXiv preprint arXiv:1803.02999*, 2018.
- [10] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in *ICML deep learning workshop*, 2015, vol. 2: Lille.
- [11] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, “Seame: a mandarin-english code-switching speech corpus in south-east asia,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [12] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, “On the end-to-end solution to mandarin-english code-switching speech recognition,” *arXiv preprint arXiv:1811.00241*, 2018.
- [13] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li, “Meta-learning for low-resource neural machine translation,” *arXiv preprint arXiv:1808.08437*, 2018.
- [14] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, “Meta learning for end-to-end low-resource speech recognition,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 7844-7848.
- [15] E. Yilmaz, H. v. d. Heuvel, and D. A. van Leeuwen, “Acoustic and textual data augmentation for improved ASR of code-switching speech,” *arXiv preprint arXiv:1807.10945*, 2018.
- [16] P. Guo, H. Xu, L. Xie, and E. S. Chng, “Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition,” *arXiv preprint arXiv:1806.06200*,

2018.

- [17]N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards end-to-end code-switching speech recognition," *arXiv preprint arXiv:1810.13091*, 2018.
- [18]S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [19]A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.
- [20]L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

新穎基於預訓練語言表示模型於語音辨識重新排序之研究

Innovative Pretrained-based Reranking Language Models for N -best Speech Recognition Lists

邱世弦 Shih-Hsuan Chiu, 陳柏林 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
{shchiu, berlin}@ntnu.edu.tw

摘要

本論文提出兩種基於 BERT 的語言排序模型，用於準確地重新排序自動語音辨識之候選結果(通常以 N -best 列表的形式來表示)。在過往的研究中，已經證明對於從聲學模型解碼出的 N -best 列表進行重新排序可以明顯改善兩階段式的語音辨識系統。然而在另一方面，隨著眾多從巨量文本預訓練的上下文語言模型的興起，它們在自然語言處理的領域中都達到了最先進的效能，像是問答系統或是機器翻譯，但是卻缺少有學者探討此種預訓練語言模型對於語音辨識的有效性。因此在本論文中，我們採用了 BERT 以開發簡單而有效的方法，對 N -best 列表進行重新排序。具體而言，我們可以將 N -best 重新排序問題視為 BERT 模型的下游任務，並提出了兩種基於 BERT 的語言排序模型，分別稱為(1) uniBERT: 給定一個 N -best 列表，輸出最理想的一連詞(Ideal Unigram)，(2) classBERT: 給定一個 N -best 列表，視為一道選擇題，輸出最好的候選排名(Oracle 當前名次)。這些模型試圖撼動 BERT 之強大之力僅僅透過一層附加輸出層，來重新排序第一階段語音辨識產生的 N -best 列表。我們評估模型於 AMI 會議語料庫，並實驗出比廣泛使用且堅強的基準 LSTMLM 改進了多達 3.14% 的 WER 相對下降率。

關鍵詞：自動語音辨識，語言模型，BERT， N -best 列表重新排序

Abstract

This paper proposes two BERT-based models for accurately rescore (reranking) N -best speech recognition hypothesis lists. Reranking the N -best hypothesis lists decoded from the acoustic model has been proven to improve the performance in a two-stage automatic speech

recognition (ASR) systems. However, with the rise of pre-trained contextualized language models, they have achieved state-of-the-art performance in many NLP applications, but there is a dearth of work on investigating its effectiveness in ASR. In this paper, we develop simple yet effective methods for improving ASR by reranking the N -best hypothesis lists leveraging BERT (bidirectional encoder representations from Transformers). Specifically, we treat reranking N -best hypotheses as a downstream task by simply fine-tuning the pre-trained BERT. We proposed two BERT-based reranking language models: (1) uniBERT: ideal unigram elicited from a given N -best list taking advantage of BERT to assist a LSTMLM, (2) classBERT: treating the N -best lists reranking as a multi-class classification problem. These models attempt to harness the power of BERT to reranking the N -best hypothesis lists generated in the ASR initial pass. Experiments on the benchmark AMI dataset show that the proposed reranking methods outperform the baseline LSTMLM which is a strong and widely-used competitor with 3.14% improvement in word error rate (WER).

Keywords: Automatic Speech Recognition, Language Models, BERT, N -best Lists Reranking

一、緒論

近年來，在眾多新穎精緻的神經網路引入下，自動語音辨識(Automatic Speech Recognition, ASR, 亦簡稱語音辨識)得到了快速而活躍的進展，基於語音辨識的各種應用(包括語音搜尋或是口語對話系統)因而取得了巨大的進步[1, 2]。儘管他們取得了很大的進步，但是在某些情境下，例如在嘈雜的環境中或在隨性風格(Casual-style)口語中執行語音辨識，普遍 ASR 系統的準確性仍然無法令人滿意[3, 4]。

在某些任務或應用中需要高準確率的 ASR，就採用了多個語音辨識候選假設(Hypotheses) (詞序列)，這些後選詞序列會以某種形式表示，例如詞圖(Lattice or Word Graph)、 N 最佳列表(N -best List)或是詞混淆網路(Word Confusion Network, WCN) [5]來顯現。進行候選假設是因為第一階段的語音辨識結果(1-best)可能會包含許多錯誤在上述的嚴重情境中，但是如果經過重新排序(Reranking or Rescoring)，則從多個候選句中，都可以找到詞錯誤率(Word Error Rates, WERs)明顯低於 1-best 的其他候選句。舉例來說，在噪音環境的語音辨識任務 CHiME-4 [6]，在 ASR 的最後階段，就使用了遞迴神經網路語言模型(Recurrent Neural Network Language Models, RNNLMs)，執行 N -best 或是詞圖的重新排序。 N -best 列表也在口語對話系統被採用[7, 8]。

在本論文，我們專注於語音辨識候選 N -best 之重新排序。目前，最廣為使用進行

N -best 重新排序的模型為 RNNLMs [9, 10] (之後都內涵了 LSTM cell [11], 亦可稱為 LSTMMLs) [12], 此模型在近幾年達到了最先進的效能, 比起稱霸多年的基於頻率計數的傳統回退 n 連詞模型(Back-off n -gram) [13, 14, 15]有更大的改善, 這是因為 RNNLMs 能夠考慮到更長的上下文資訊(Long-term Context)。隨後, 許多研究專注於探索 LSTMMLs 的調適(Adaptation)方法, 以進行更準確的 N -best 重新排序。但是要注意的是, 即使 LSTMMLs 在 N -best 重新排序表現傑出, 但它最初是為了預測下一個單詞而開發的, 而不是為了 N -best 重新排序任務而開發的。

在另一方面, 隨著自然語言處理(Natural Language Processing, NLP)的技術大量發展, 當前 ASR 系統用於評估 N -best 候選句的語言與語意合法性的資訊還是相當有限。在自然語言處理的領域中, 許多膾炙人口的預訓練語言表示法模型 (Pre-trained Language Representation Models), 在近幾年如雨後春筍般的湧出, 像是 ELMO (Embeddings from Language Models) [16], GPT (Generative Pre-Training Transformer) [17], BERT (Bidirectional Encoder Representations from Transformers) [18]...等等預訓練模型, 來提取上下文相關(Context-dependent or Contextualized)的詞嵌入(Word Embedding), 此種 Contextualized 詞嵌入已經被證實在眾多下游 NLP 的任務下達到了最先進的效能, 像是口語語言理解[19]、文本分類[20]和問答任務[21]...等等。然而, 據我們所知, 鮮少有相關研究探討將上述的預訓練語言模型, 應用於 ASR 系統中並探討其有效性。因此在本論文, 我們嘗試利用 Google 近來提出的 BERT 來對從 ASR 第一階段產生的 N 最佳候選列表(N -best List), 執行重新排序, 希望提高 ASR 的效能。

我們提出了兩種基於 BERT 的語言排序模型, 都是將 N -best 重新排序視為 BERT 的下游任務, 都是基於在 BERT 之上, 僅僅疊加一層全聯接層(Fully Connected Layer, FC), 分別稱為(1)uniBERT: 給定一個 N -best 列表, 輸出理想的一連詞(Ideal Unigram)和(2)classBERT: 給定一個 N -best 列表, 輸出最好的候選排名(Oracle 的排名, Oracle 代表的是與該正確文句做計算, WER 最小的那條候選句), 這兩種模型將會在第四章做詳細的介紹。在 BERT 的預訓練階段中, 主要對模型進行訓練以從上下文, 來預測被遮蔽的單詞, 以使模型能夠“融合”左和右的表示, 與以前的 bi-LMs (包括 bi-RNNLMs) [22, 23] 不同, 後者使用各方向的獨立編碼表示來淺層連接(Shallow Concatenation), 因此可能會限制 bi-RNNLMs 的潛力。有鑑於此, 我們認為 BERT 對於 N -best 列表重新排序是有前途的, 因而提出了兩種基於 BERT 的語言排序模型, 這些模型試圖借助 BERT 之力僅通過微調(Fine-tuning)一層附加的輸出層。我們在基準語料庫 AMI 上評估我們的模型, 並

表明所提出的模型比強大且廣泛使用的 LSTMLM 獲得了更好的性能。

二、文獻回顧

在本節中，我們將簡要回顧有關 ASR 系統中 N -best 重新排序方法的先前研究。隨著近年來深度神經網絡的興起，RNNLM (LSTMLM) [12] 直接稱霸語言模型界成為流行且廣泛使用於 N -best 重新排序，遠勝過傳統的統計式 n -gram 模型[13, 14, 15]，因為前者能考慮更長距離的資訊。因此有許多研究都集中在探索 LSTMLM 的調適方法，以進行更準確的 N -best 重新排序，像是有一些研究利用**歷史資訊**(History Information)對 RNNLM 作語言模型調適[24, 25]。而有更多研究專注於對**主題資訊**(Topic Information)作語言模型調適，例如 Mikolov [26] 使用上下文感知向量(Context-aware Vectors)作為 RNNLM 的額外輸入，以適應大範圍的主題資訊。同樣地，Chen [27] 探究主題建模方法，以提取主題特徵作為 RNNLM 的附加輸入，用於多類型廣播轉錄任務中的類型和主題的調適。Ma [28] 探索了基於 LSTMLM 的三種微調策略。Lam [29] 對 LSTMLM 的激活函數(Activation Function)作高斯分佈處理(Gaussian Process)，得到了些微的進步。Irie [30] 提出了一種基於 LSTM 的動態加權之混合器(Mixer)，各個主題模型在特定領域(Specific Domain)上分別進行訓練，並擁有動態的權重，可以勝過簡單的線性插值。之後，Li [31] 使用上述前者的方法，但他改成使用基於 Transformer 的 LM 與加權混合器。

但要注意的是，即使 LSTMLM 在 N -best 重新排序方面表現出色，但它最初的設計是為了預測下一個單詞而開發的，而不是為進行 N -best 重新排序任務而開發的。所以有研究者直接提出專為 N -best 重新排序任務而設計的模型。像是鑑別式語言模型(Discriminative Language Models, DLM) [32-35] 最初就是為 N -best 重新排序而開發的，它利用 ASR 的錯誤資訊來訓練鑑別式語言模型。Ogawa [36] 受 DLM 啟發，但認為其損失函數(Loss Function)的設計很複雜，因此他們開發了一個簡單的編碼器-分類器模型(Encoder-Classifer Model)，該模型訓練一個分類器進行一對一的候選句比較(氣泡排序(Bubble Sort))來執行 N -best 重新排序。Tanaka [37] 提出了一種將端到端(End-to-End) ASR 系統視為一個神經語音到文本語言模型(Neural Speech-to-Text LMs, NS2TLM)的想法，該模型以輸入的聲學特徵為條件，並將其用於對 DNN-HMM hybrid ASR 系統中生成的 N -best 進行重新排序。Song [38] 受資訊檢索(Information Retrieval, IR)中的核心問題，即排名學習(Learning-to-Rank, L2R)的啟發，提出了重新計分學習(Learning-to-

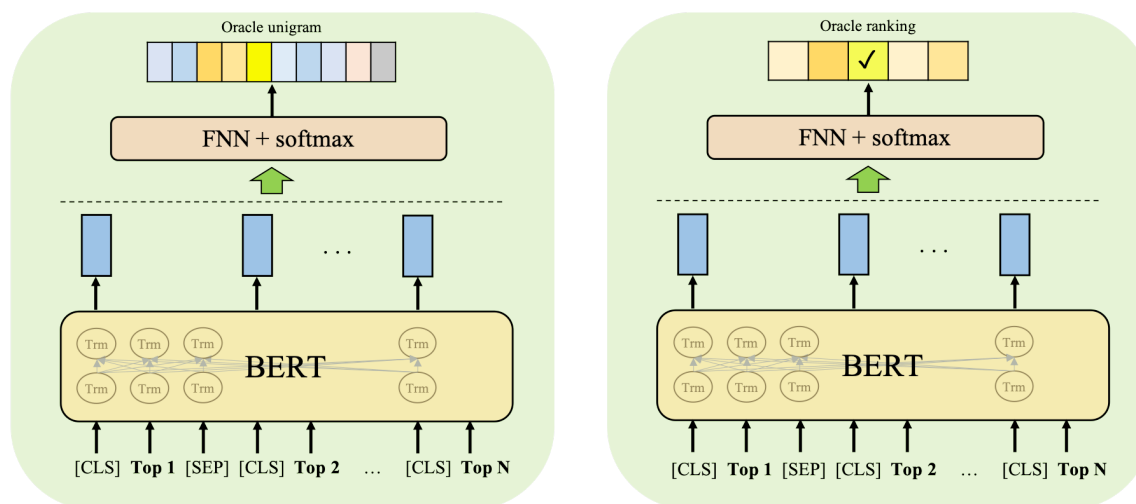
Rescore, L2RS) 機制，這是第一個研究將 N -best 重新排序視為一種學習問題(Learning Problem)。亞馬遜 Gandhe [39]在 2020 年提出一種基於注意力機制(Attention-based)的鑑別式語言模型，該模型在訓練鑑別式 LSTM 時會同時考慮到詞與聲音的特徵，後者是借助端對端系統，把當前詞與聲音片段藉由注意力機制所獲取，該方法也得到了顯著的效果。

近年來，隨著預訓練語言表示法模型的興起且蔚為風潮，像是 BERT [18]，在 2018 年一問世直接打破 12 項最先進效能的 NLP 任務。因此在 2019 年，Wang [40]首度提及可以將 BERT 拿來作句子評分，但是他們並沒有以實驗證明。不久後，首爾大學 Shin 等人[41]，以預測每個位置遮蔽處([MASK])的機率加總作為句子的分數，雖然相較傳統的 LM 它不是真正的句子機率，但仍然可以作為 N -best 重新排序。隨後，亞馬遜 Salazar 等人[42]，也作[MASK]處的機率加總，把句子分數定義為 PLL (Pseudo-log-likelihood)，並以數學形式證明該方法的有效性，並提出為該方法執行加速，藉由知識蒸餾(Knowledge Distillation)訓練一個不用預先遮蔽([MASK])某處單詞的加速版句子評分模型，雖然效果變差了，但把計算複雜度從 $O(|W| \cdot V)$ (其中 $|W|$ 是句子長度， V 是詞典大小)降到 $O(1)$ ，且比較了多種預訓練模型的效能。首爾大學 Shin 等人[43]，受到 Salazar 的啟發，不同於原始 BERT 要重複性的遮蔽再預測(Mask-and-Predict Repetition)，在 BERT 內部 self-attention 處利用對角遮蔽(Diagonal Masking)就能達到原先功能且加快速度，而實驗證明，該方法不僅為 N -best 重新排序加快了 6 倍時間甚至連效能也得到提升。有鑑於此，我們也提出了兩種基於 BERT 的語言排序模型，這些模型希望撼動 BERT 強大之力透過僅僅一層附加輸出層，都將會在第四章節作詳細介紹。下一章節將會詳細介紹 BERT。

三、BERT

BERT (Bidirectional Encoder Representations from Transformers)是一個近期發表且廣為使用的語言表示模型，內部由多層雙向的 Transformer [44]的編碼器(Encoder)所組合起來。而 Transformer 是一個基於注意力機制(Attention-based)的結構，且能夠考慮全域的輸入與輸出的相依性。訓練 BERT 分成兩個階段，分別是預訓練(Pre-training)和微調(Fine-tuning)。在預訓練階段，有兩個訓練準則且同時訓練於大量且廣泛主題的無標注文本資料，一個是遮蔽語言建模(Masked Language Modeling, MLM)，另一個則是下句預測 (Next Sentence Prediction, NSP)，前者是一種填空任務，學習如何從過去和未來的上下

文(Context)預測出被遮蔽的詞是什麼詞彙，後者學習兩兩之間的句子是否有連貫性(Contiguous)。BERT 特別受益於 MLM 的訓練方式，因為它能夠“融合”(Fuse)歷史與未來詞的資訊，不像傳統的 LM 只考慮先前詞，或是 bi-RNNLM 只能淺層連接(Shallow Concatenation)兩個方向的上下文資訊。在微調(Fine-tuning)階段，預訓練完成的 BERT 僅僅只要附加一層輸出層(Output Layer)針對特定任務視為下游任務，就能撼動強大的 BERT。輸出層從頭訓練(Training from Scratch)，而 BERT 本體的參數會被“微調”。BERT 在多項 NLP 的領域得到了最新進的效能，包括問答系統(Question Answering, QA)，自然語言推論(Natural Language Inference, NLI)，神經機器翻譯(Neural Machine translation, NMT) 和語音文件檢索(Spoken Document Retrieval, SDR)...等等。因此我們也將 N -best 重新排序視為一種 NLP 任務，認為 BERT 是一個有潛力的雙向語言模型(bi-LMs)，將 N -best 重新排序作為 BERT 的下游任務。



圖一：uniBERT 與 classBERT 的架構

四、基於 BERT 之語言排序模型(BERT-based Reranking Language Models)

本論文著重於嘗試借助 BERT (本論文採用“bert-base-uncased”版本)之力應用於第一階段之語音辨識結果 N -best 重新排序。在本節中，我們將提出兩種基於 BERT 的 N -best 重新排序模型，分別稱為 uniBERT 和 classBERT。簡單來說，首先使用預訓練的 BERT 參數初始化該模型，然後使用標記好的訓練資料，僅附加一層額外的輸出層即可對預訓練的 BERT 進行微調(Fine-tuning or Adaptation)。具體而言，輸出層將從頭訓練(Training From Scratch)，而預訓練的 BERT 將進行微調。

4.1 uniBERT

我們在原始(Vanilla)預訓練的 BERT 之上疊加了一層前饋式全連接層(Feed-forward Neural Network, FNN)，直接輸出 V 維(詞典大小)最理想的一連詞(Ideal Unigram)。具體來說，給定一組 N -best 列表，模型能夠在 N -best 列表中輸出“最好”(Oracle)的候選句的 unigram。我們希望利用 BERT 來萃取出 N -best 列表中，多個候選句中的詞與詞甚至是句與句之間的關係，輸出到一個理想的 unigram，我們稱此模型為 uniBERT。模型架構如圖一之左圖所示。在微調(Fine-tuning)階段，一次輸入 N 句(在本論文實驗設定為 $N = 10$)候選句，並在每個候選句的頭跟尾分別加入特殊符號[CLS]和[SEP])，uniBERT 要學習如何輸出最理想的 unigram。而模型輸入 N -best 後的流程如下所述，每一候選句會先藉由[CLS] token，BERT 自動編碼成句子表示法 h_k (如圖 1 藍色長方形所示)，而全部的表示法 h_k 會互相逐項(Element-wise)的取平均，或是相連接起來(在實驗中會比較該兩種方法)，再經過一層線性分類層(FC layer)和 softmax 讓此 unigram 正規化(滿足機率公設，總和為 1)，就能得到理想的 unigram。uniBERT 的演算過程以數學式表示如下：

$$\begin{aligned}
 [h_1, h_2, \dots, h_{10}] &= \text{BERT}([hyp_{[CLS]}^1, hyp_{[CLS]}^2, \dots, hyp_{[CLS]}^{10}]) \\
 h^{nb} &= \text{Average}([h_1, h_2, \dots, h_{10}]) \text{ or } \text{Concat}([h_1, h_2, \dots, h_{10}]) \\
 z^{nb} &= \text{linear}(h^{nb}) \\
 P_{bert_{uni}}(\cdot | h^{nb}) &= \text{Softmax}(z^{nb})
 \end{aligned} \tag{1}$$

其中 $P_{bert_{uni}} = uniBERT(nb) \in R^V$ 是模型輸出最理想的 unigram， nb 為一組 N -best 列表。而訓練模型的資料收集於每個訓練文本(語句人工轉錄)解碼出的 N -best 列表中 WER 最低的那條候選句(與正確文本做計算)，並且創造它的 unigram 表示法，舉例來說，例如 Oracle 候選句是：“我 愛 你 你 愛 我 媽”，unigram 表示法為 $P_{ora_{uni}} = [\dots, 0, \frac{2}{7}, \frac{2}{7}, 0, 0, \frac{2}{7}, 0, \frac{1}{7}, 0, \dots]$ 。訓練準則(Training Criterion)使用 Kullback-Leibler (KL)散度：

$$L = D_{KL}(P_{ora_{uni}} | P_{bert_{uni}}) = \sum_{w \in Vocab} P_{ora_{uni}}(w) \log \left(\frac{P_{ora_{uni}}(w)}{P_{bert_{uni}}(w)} \right) \tag{2}$$

此模型對於每筆訓練資料去做最小化 KL 散度的訓練，找到模型最佳化參數。在測試階段時，我們可以使用 $P_{bert_{uni}}$ 去替代或是插值於原本的語言模型分數，去為第一階段的語

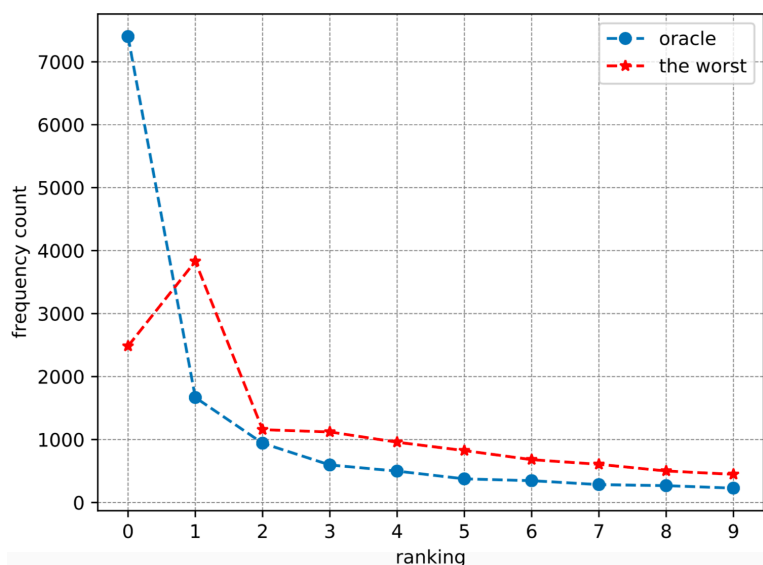
音辨識產生的 N -best 列表之每一候選句進行重新打分(Rescoring)，希望改善語音辨識效能。

4.2 classBERT

classBERT 模型結構與 uniBERT 非常相似，如圖一之右圖所示，唯一的區別是，給定一組 N -best 列表，它會直接輸出 Oracle 候選的目前所在位置(也就是名次)。在微調階段，classBERT 要學習如何輸出 Oracle 候選的排名。形式上，我們創建訓練資料 $\mathcal{T} = \{(nb_1, y_1), \dots, (nb_t, y_t), \dots, (nb_T, y_T)\}$ 來訓練 classBERT，其中 nb_t 是某句訓練語句 (Utterance) U_t 的 N -best 列表，而 y_t 是在 nb_t 中 Oracle 候選句的當前排名位置，其中 y_t 會以 one-hot 的形式 $P_{y_t} \in R^N$ 表示， N 是 N -best 列表的大小(本論文採用 $N = 10$)。模型嘗試學習如何輸出 Oracle 候選的排名，訓練準則是交叉熵(Cross Entropy)損失函數，廣泛應用於許多多類別分類(Multi-class Classification)問題中：

$$L = H(P_y, P_{bert_{cls}}) = - \sum_i^N P_y(i) \log P_{bert_{cls}}(i) = - \log P_{bert_{cls}}(k) \quad (3)$$

其中 $P_{bert_{cls}} = classBERT(nb) \in R^N$ 是模型輸出， k 是預期的 Oracle 排名位置。classBERT 學習在每筆訓練資料去最小化 Cross Entropy。在測試階段，給定一組 N -best 列表，我們



圖二：表示 Oracle 排名的頻率分佈和 Worst 排名的頻率分佈，分別由藍線和紅線顯示

就可以直接輸出哪一句為“最佳”的候選句，希望藉此改善第一階段語音辨識效能。

表一：AMI 的評估集於第一階段的語音辨識結果(1-pass)、Oracle、隨機選擇、最差的 WERs，
第二行顯示 3-Gram 困惑度(PPL)

	1-pass	Oracle	Random	Worst
WER	22.79	14.39	29.28	40.65
3G PPL	154.62			

五、語音辨識實驗

我們評估模型於 AMI 會議語料庫[45]上，這是一個眾所周知的基準(Benchmark)語音辨識語料庫，內含 100 個小時的會議對話記錄。100 小時的音檔用來訓練 DNN-HMM 結構的聲學模型 $p(X|W)$ ，以及相應的轉錄文本(總共 108221 條語句)以訓練基於 Kneser-Ney (KN) [13]平滑技巧的 3-Gram 語言模型 $P(W)$ 。這兩個分開訓練的模型構成了我們基礎的第一階段 ASR 系統。在本論文中，基礎 ASR 系統是使用語音辨識實驗的著名工具包 Kaldi [46]搭建的。本論文中，我們致力於將我們提出的方法（基於 BERT 的兩種新穎的重排模型）應用於 N -best 重排任務來改善第二階段的 ASR。首先，我們會使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search)對第一階段 ASR 系統在評估集(Evaluation Set, 12612 條待測語句(音檔))建立的詞圖(Lattice or Word Graph)進行解碼，從而獲得每個音檔的前 N 個最佳候選列表 (N -best list，本論文採用 $N = 10$)。因此，我們能夠利用更高階的語言模型，例如: NNLMs (在本論文是使用提出的兩種模型)來替換或內插語言模型分數 $P(W^{Hyp})$ ，並與相應的聲學模型分數 $p(X|W^{Hyp})$ 結合以重新排列 N -best 列表：

$$W^* = \underset{W^{Hyp} \in N\text{-best}}{\operatorname{argmax}} p(X|W^{Hyp})P(W^{Hyp}) \quad (4)$$

期望獲得更好的辨識結果(詞序列) W^* 。表一表示 ASR 系統在評估集第一階段的（使用 3-Gram LM）WER，而 Oracle WER 是 10-best 的理論上限(Ceiling Performance)，表示每個測試語句都選擇 WER 最低的候選句，Random 表示每個測試語句都隨機選擇一條候選句，Worst 表示每個測試語句都選擇 WER 最高的候選句，第二行顯示 3-Gram 在第一

階段語音辨識的困惑度(Perplexity, PPL)。圖二表示 Oracle 排名的頻率分佈和 Worst 排名的頻率分佈，分別由藍線和紅線顯示。

表二：uniBERT 應用於 AMI 語音辨識的結果(WERs)

AM = 27.55	LM	AM + 10 * LM
1-pass	26.80	22.79
LSTM	25.00	21.33
uniBERT	26.84	22.86
LSTM + uniBERT	25.12	21.24

5.1 uniBERT 之 N -best 重新排序實驗

在第一個提出的語言重排模型 uniBERT 中，當我們向模型輸入一組 N -best 列表，它會輸出理想的 unigram 語言模型 $P_{bert_{uni}}(W)$ 。此 unigram LM 可用於重新計分語言模型 $P(W)$ 分數：

$$P(W^{Hyp}) = \alpha P_{rnn}(W^{Hyp}) + (1 - \alpha) \left(\beta P_{bert_{uni}}(W^{Hyp}) + (1 - \beta) P_{tri}(W^{Hyp}) \right) \quad (5)$$

其中 $P_{tri}(W)$ 是第一階段語音辨識的 3-Gram 語言模型，然後與模型輸出 $P_{bert_{uni}}(W)$ 用係數 β 進行線性插值，分配兩者模型的相對貢獻，在本研究中我們在發展集(Developing set) 中調配出最好的效能為 $\beta = 0.2$ 或是 0.1 。此外，我們還使用 RNNLM (LSTMLM) 的分數 $P_{rnn}(W)$ 與上述組合後的分數做線性插值，用自由超參數 α 來分配彼此的貢獻，並根據經驗法則將其設置為 $\alpha = 0.7$ 或是 0.8 。在這部分的實驗，我們主要是期望利用 BERT 萃取出理想的 unigram 來輔助最先進的基準 LSTMLM，並提供額外的資訊，例如詞頻。如

表三：classBERT 應用於 AMI 語音辨識的結果(WERs)

AM = 27.55	LM	Consider AM and LM	
1-pass	26.80	22.79	
LSTM	25.00	21.33	
classBERT	23.18	+2-dim	+1-dim
classBERT+3G	-	21.69	21.27
classBERT+LSTM	-	20.66	21.61

表二所示，雖然單獨使用 uniBERT 輸出的 unigram 不會直接改善 ASR 性能，但可以輔助 LSTMLM 並使其(LSTMLM)改善 0.2% 的 WER 相對下降率。

5.2 classBERT 之 N -best 重新排序實驗

在第二種提出的語言重排模型 classBERT 中，給定一組 N -best 列表，模型能直接選擇出哪一條是最佳(Oracle)的候選句。如表三所示，有三種實驗設定，第一種是 classBERT 僅考慮文本(候選句)，並且勝過基準 LSTMLM 相對減少了 7.28% 的 WER。第二種是我們考慮了 ASR 的兩種分數(聲學分數和語言模型(3G 或是 LSTMLM)分數)，在 BERT 編碼出的候選句嵌入的頂端連接(Concatenating)該二種分數(即[CLS]的 768-dim + 2-dim)作為特徵，該方法在加入聲學和 LSTMLM 分數時，比基準的 LSTMLM 進步了 3.14% 的 WER 相對下降率。第三種方法是將 AM 和 LM 得分利用我們的先備知識(即 $AM + 10 * LM$) 事先結合起來，成為了單個分數，然後我們如同前者的方法，把該分數連接在候選句嵌入的頂端(也就是[CLS]的 768-dim + 1-dim)，該方法在加入 3G 分數時比原先沒有先結合 ($AM + 10 * LM$)獲得了改善，也比 LSTMLM 進步了 0.3% 的 WER 相對下降率。

六、結論與未來展望

在本文中，我們提出了兩種基於 BERT 之 N -best 重新排序模型，分別是 uniBERT 與 classBERT。uniBERT 給定一組 N -best 列表，輸出最理想的 unigram，而 classBERT 將 N -best 重新排序視為一道選擇題。我們已經通過實驗證實了兩者優異的 N -best 重新排序效能。這些方法都是通用框架，可以應用於使用 N -best 列表形式作為候選假設的其他研究領域，例如：機器翻譯(Machine Translation, MT)和資訊檢索(Information Retrieval, IR)。

在未來的研究中，我們計畫像以往的研究[32-35、47-48]一樣，通過使用鑑別式訓練(Discriminative Training)，主要是利用 ASR 的錯誤當作特徵，來提高語言排序模型的效能。我們還希望考慮語者之前所說過的內容(歷史資訊)，來幫助預測當前的話語。

參考文獻

- [1] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling

- in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
- [2] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*, Springer-Verlag London, 2015.
- [3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr. 2014.
- [4] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, “Low-latency realtime meeting recognition and understanding using distant microphones and omnidirectional camera,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 499–513, Feb. 2012.
- [5] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: Word error minimization and other applications of confusion networks,” *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, Oct. 2000.
- [6] E. Vincent, S. Watanabe, A.A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Computer Speech and Language*, vol. 46, pp. 535–557, Nov. 2016.
- [7] J.D. Williams, “Exploiting the ASR N-Best by tracking multiple dialog state hypotheses,” in *Proc. Interspeech*, 2008, pp. 191–194.
- [8] S. Young, M. Gašić, B. Thomson, and J.D. Williams, “POMDP-based statistical spoken dialogue systems: A review,” *Proc. IEEE*, vol. 101, no. 5, pp. 1160–1179, Nov. 2016.
- [9] T. Mikolov, M. Karafiat, L. Burget, F. Černocký, and S. Khudanpur, “Recurrent neural network based language model,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2010.
- [10] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur, “Extensions of recurrent neural network language model,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5528–5531.
- [11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [12] M. Sundermeyer, R. Schluter, and H. Ney, “Lstm neural networks for language modeling,” in *Thirteenth annual conference of the international speech communication association*, 2012.

- [13] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in ICASSP, 1995, vol. 1, p. 181e4.
- [14] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.
- [15] J. T. Goodman, “A bit of progress in language modeling,” *Computer Speech & Language*, vol. 15, no. 4, pp. 403–434, 2001
- [16] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. NAACL, 2018.
- [17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” Technical Report, OpenAI, 2018.
- [18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.
- [19] Chao-Wei Huang and Yun-Nung Chen, “Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding,” In Proceedings of The 45th IEEE ICASSP, 2020.
- [20] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.
- [21] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016.
- [22] Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen, “Bidirectional recurrent neural network language models for automatic speech recognition,” in Proc. ICASSP, 2015, pp. 5421–5425.
- [23] Xie Chen, Anton Ragni, Xunying Liu, and Mark Gales, “Investigating bidirectional recurrent neural network language models for speech recognition.,” in Proc. ICSA INTERSPEECH, 2017.
- [24] Mittul Singh, Youssef Oualil, and Dietrich Klakow, “Approximated and domain-adapted lstm language models for first-pass decoding in speech recognition.,” in Proc. Interspeech, 2017.
- [25] Ke Li, Hainan Xu, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, “Recurrent neural network language model adaptation for conversational speech recognition.,” in Proc. Interspeech, 2018.

- [26] Tomas Mikolov and Geoffrey Zweig, “Context dependent recurrent neural network language model,” in Proc. SLT, 2012.
- [27] Xie Chen, Tian Tan, Xunying Liu, Pierre Lanchantin, Moquan Wan, Mark JF Gales, and Philip C Woodland, “Recurrent neural network language model adaptation for multi-genre broadcast speech recognition,” in Proc. Interspeech, 2015.
- [28] Min Ma, Michael Nirschl, Fadi Biadsy, and Shankar Kumar, “Approaches for neural-network language model adaptation.,” in Proc. Interspeech, 2017.
- [29] M. W. Y. Lam, X. Chen, S. Hu, J. Yu, X. Liu, and H. Meng, “Gaussian process lstm recurrent neural network language models for speech recognition,” in ICASSP 2019, pp. 7235–7239, May 2019.
- [30] Kazuki Irie, Shankar Kumar, Michael Nirschl, and Hank Liao, “Radmm: recurrent adaptive mixture model with applications to domain robust language modeling,” in Proc. ICASSP, 2018
- [31] Ke Li, Zhe Liu, Tianxing He, Hongzhao Huang, Fuchun Peng, Daniel Povey, Sanjeev Khudanpur, “An Empirical Study of Transformer-Based Neural Language Model Adaptation” in Proc. ICASSP, 2020
- [32] B. Roark, M. Saraclar, and M. Collins, “Discriminative n-gram language modeling,” *Computer Speech and Language*, vol. 21, no. 2, pp. 373–392, Apr. 2007.
- [33] F.J. Och, “Minimum error rate training in statistical machine translation,” in Proc. ACL, 2003, pp. 160–167.
- [34] M. Collins and T. Koo, “Discriminative reranking for natural language parsing,” *Computational Linguistics*, vol. 31, no. 1, pp. 25–70, Mar. 2005.
- [35] T. Oba, T. Hori, A. Nakamura, and A. Ito, “Round-robin duel discriminative language models,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1244– 1255, May 2012.
- [36] Atsunori Ogawa, Marc Delcroix, Shigeki Karita, and Tomohiro Nakatani, “Rescoring n-best speech recognition list based on one-on-one hypothesis comparison using encoder-classifier model,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 6099–6103.
- [37] Tomohiro Tanaka, Ryo Masumura, Takafumi Moriya, and Yushi Aono, “Neural speech-to-text language models for rescoring hypotheses of dnn-hmm hybrid automatic speech recognition systems,” in 2018 AsiaPacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018, pp. 196–200.

- [38] Yuanfeng Song, Di Jiang, Xuefang Zhao, Qian Xu, Raymond Chi-Wing Wong, Lixin Fan, Qiang Yang, “L2RS: A Learning-to-Rescore Mechanism for Automatic Speech Recognition,” arXiv preprint arXiv:1910.11496, 2019.
- [39] Ankur Gandhe, Ariya Rastrow, “Audio-attention discriminative language model for ASR rescoring,” In Proceedings of The 45th IEEE ICASSP, 2020.
- [40] Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In NeuralGen.
- [41] Joongbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using BERT for speech recognition. In ACML.
- [42] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In ACL.
- [43] Joongbo Shin, Yoonhyung Lee, Seunghyun Yoon, Kyomin Jung. 2020. Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning. In ACL.
- [44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.
- [45] J. Carletta et al., “The AMI meeting corpus: A pre-announcement,” The International Workshop on Machine Learning for Multimodal Interaction, 2005.
- [46] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi Speech Recognition Toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.
- [47] Y. Tachioka and S. Watanabe, “A discriminative method for recurrent neural network language models,” in Proc. ICASSP, 2015, pp. 5386–5389.
- [48] T. Hori, C. Hori, S. Watanabe, and J.R. Hershey, “Minimum word error training of long short-term memory recurrent neural network language models for speech recognition,” in Proc. ICASSP, 2016, pp. 5990–5994.

Lectal Variation of the Two Chinese Causative Auxiliaries

Cing-Fang Shih, Mao-Chang Ku, Shu-Kai Hsieh

Graduate Institute of Linguistics, National Taiwan University

r08142004@ntu.edu.tw, d08142002@ntu.edu.tw, shukaihsieh@ntu.edu.tw

摘要

本文旨在從語料庫的觀點研究中文兩個使役助動詞使 ‘cause’ 和讓 ‘let’ 之間的差異。我們對從兩個語料庫中提取的中文語料進行邏輯迴歸分析，認為此兩個助動詞之間的差異可視為 Verhagen and Kemmer (1997) 所提出的直接/間接使役區分。回歸模型得到的結果表明，直接/間接使役的理論為動詞的特徵和詞義提供了合理的解釋。我們指出，動詞使與「直接使役」相關，因為它通常使用於涉及無生命參與者的使役事件中，在這種情況下，起因論旨角色必然且直接地導致受使役者的結果狀態。另一方面，讓應該被歸類為「間接使役」，因為它通常用於涉及有生命參與者的場景，並且除了使動者之外，亦有其他一些驅動來源也導致使役事件的發生。

Abstract

This paper aims to investigate the variation between two Chinese causative auxiliaries *shi* ‘使’ and *rang* ‘讓’ from a corpus-based perspective. We conduct a logistic regression analysis to the Chinese data extracted from two corpora and propose a direct/indirect distinction (Verhagen and Kemmer 1997) between the two auxiliary verbs. The results retrieved by the regression model show that the theory of direct/indirect causation provides a reasonable account for the characteristics and lexical meanings of the verbs. We indicate that the verb *shi* is correlated with “direct causation” because it is typically used when inanimate participants are involved in the causing event, in which the force initiated by the cause inevitably and directly leads to the resulted stage of the causee. On the other hand, the verb *rang* should be classified as “indirect causation” because it is typically used in scenarios where animate participants are both involved, and some extra force besides the causer also plays a role in the effected event.

關鍵詞：語言變異，使役結構，邏輯迴歸，R 語言統計

Keywords: language variation, causation, logistic regression, R statistics.

1. Introduction

The causative construction has been a debatable subject in linguistic studies. It is widely accepted that there are two participants encoded in a causative construction, which are the causer and the causee. The causing event led by the causer, and the caused event formed by the cause, are two components of a causative construction [1]. Verhagen and Kemmer [2] described the causative verb as a ‘causal predicate’, and the infinitive in the construction is called ‘effected predicate’, which includes two varieties: intransitive and transitive. In Mandarin Chinese, causative verbs *shi* ‘使’ and *rang* ‘讓’ can form causative constructions, see (1).

- (1) a. 你又說了幾句讓我印象深刻的話

nǐ yòu shuō-le jǐ-jù ràng wǒ
you again say-PERF several-CL make me
yìxiàng shēnkè de huà
impression deep MOD words

‘You again say something that has deeply impressed me.’

- b. 現代通訊科技使我們可以天天通話

xiàndài tōngxùn kējì shǐ wǒmen
modern communications technology make us
kěyǐ tiāntiān tōnghuà
able every.day call

‘Modern communications technology enables us to call every day.’

In (1), the subject before *shi* or *rang* is the causer, and the object after the predicate is the causee. Constructions with causal predicates *shi* and *rang* are categorized as direct and indirect causation, respectively. Most of the time, direct causation is more likely to indicate non-human interaction than the indirect one is. To clarify the usages of the two causal predicates, this study is going to demonstrate a corpus-based regression analysis to explore the word choice between *shi* and *rang*. Furthermore, the regression analysis explains how the property of the causal

predicates influences the tendency of choosing direct or indirect causation.

This paper is organized as follows. Section 1 is the introduction. Section 2 briefly reviews related literature. Section 3 describes our research methods. Section 4 presents our results. A direct/indirect dichotomy is argued for and a comparison between Chinese and Dutch causative predicates is made. Section 5 concludes this paper.

2. Literature Review

The structure of causative construction reflects human's real-world experience of the relationship between the cause and the result. It is widely discussed from the typological aspect and the cognitive aspect. From the typological point of view, causatives are widely classified into three different types: (i) lexical causatives, (ii) morphological causatives, and (iii) analytic causatives [3]. From the cognitive point of view, Croft [4] explained the Idealized Cognitive Model (ICM) based on Lakoff [5]. Croft [4] views the causative construction as a single event, and it falls into three categories: (i) causative, (ii) inchoative, and (iii) stative. Both Comrie's [3] and Croft's [4] classifications of causative construction are defined as a continuum, which expresses that a linguistic expression does not always neatly fall into one of the three types. Instead, it can fit in between the two adjacent types.

Croft [6] schematized the causation types proposed by Talmy [7, 8], as shown in Figure 1. Two dimensions distinguish the four causation types. The first dimension makes distinctions between the initiator and endpoint in a causative construction. The other dimension shows differences between the animate and inanimate. Animates are seen as the mental dimension, and inanimates are physical. As demonstrated by Figure 1, the two arrows starting from the physical entity, which are affective and physical, are rather straight and direct. It shows that physical entities can act on other entities directly. On the other hand, the two arrows starting from the mental entity are not straightforward. The arrow of mental-on-mental causation, which is inductive, is rather bent. Also, the arrow of mental-on-physical causation, which is volitional, is slightly bent. It shows that mental entities cannot act on others as directly as physical entities.

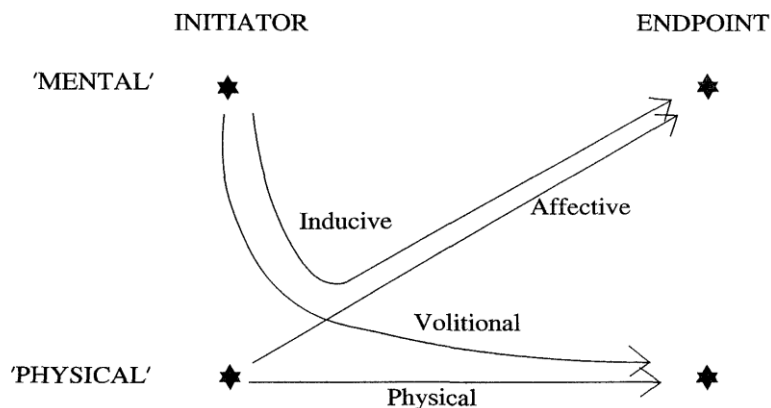


Figure 1. A Model of Causation Types (Croft 1991: 167; based on Talmy 1976)

The model of causation types ([6, p. 167]; based on [7]) is used in the study of Verhagen and Kemmer [2] to analyze the causative constructions in modern standard Dutch formed by *doen* and *laten*. According to the estimate of Verhagen and Kemmer [2], *laten* should indicate inducive (mental-on-mental) causation and should have more animate causers than inanimate ones, for it forms indirect causation. By contrast, *doen* should have more inanimate causers, for it is thought to be the component of direct causation.

3. Research Methods

To understand the usages of the two predicates under different circumstances, the data that contains *shi* and *rang* were extracted. The traditional Chinese data is collected from Academia Sinica Balanced Corpus of Modern Chinese, and the simplified Chinese data is from The Chinese Web Corpus (zhTenTen). The data contains four categories, which are traditional *shi*, traditional *rang*, simplified *shi*, and simplified *rang*. Two hundred items of each category are selected randomly for further analysis. After removing the data in which *shi* and *rang* are not used as causative verbs, the remaining 606 data were annotated.

First of all, whether the subject and the object of the data are mental or non-mental will be decided. If the subject, usually a human being or an institute that is operated by humans, can conduct the causing event of their own will, it is a mental subject. By contrast, if the subject is inanimate, it is marked as non-mental. A mental object is decided when it spontaneously executes the caused event. Otherwise, it will be considered as non-mental. The properties of

the causal predicates are decided by subjects and objects in causative constructions. If the causative construction contains a mental subject and a mental object, its property is annotated as inductive, as exemplified in (2), repeated from (1a).

(2) 你又說了幾句讓我印象深刻的話

nǐ yòu shuō-le jǐ-jù ràng wǒ
you again say-PERF several-CL make me
yìnxàng shēnkè de huà
impression deep MOD words
'You again say something that has deeply impressed me.'

In (2), the causer is *nǐ* 'you', which is a mental subject, and the causee is *wǒ* 'me', which is a mental object. The causer can directly influence the causer, while the causer can decide to perform the influence of his or her own will.

As shown in (3), volitional causation is defined when the construction contains a mental subject and a non-mental object.

(3) 他讓事件的終點等於起點

tā ràng shìjiàn de zhōngdiǎn děngyú qǐdiǎn
he make event MOD end.point equal.to starting.point
'He makes the endpoint of the event equal to its starting point.'

In (3), the causer is *tā* 'he', and the causee is 'the endpoint of the event'. This causative construction contains a mental subject the performs influence on the non-mental object.

If the construction features a non-mental subject and a mental subject, it is considered affective, as demonstrated in (4), reproduced from (1b).

(4) 現代通訊科技使我們可以天天通話

xiàndài	tōngxùn	kējì	shǐ	wǒmen
modern	communications	technology	make	us
kěyǐ	tiāntiān	tōnghuà		
able	every.day	call		

‘Modern communications technology enables us to call every day.’

The example given in (4) contains a non-mental subject *xiàndài tōngxùn* ‘modern communications’ and a mental object *wǒmen* ‘us’. The non-mental subject does not voluntarily act on the object; however, the object has influenced.

Finally, physical causation is found when both the subject and the object are non-mental, as shown in (5).

(5) 長壽能使文化承繼較完整

chángshòu	néng	shǐ	wénhuà	chéngjì	jiào	wánzhěng
longevity	can	make	culture	inheritance	more	complete

‘Longevity can make cultural inheritance more complete.’

Both the subject and object in (5) are non-mental. It presents an indirect act which is done by the inanimate subject to the object that is also inanimate.

After the properties are recorded, the transitivity variable is annotated by the transitivity of verbs after the causal predicates. The verb expresses the function of an ‘effected predicate’ [2], and it can be transitive or intransitive. If the verb requires an object, it is marked transitive (TR). Otherwise, it will be considered intransitive (INTR).

Finally, the varieties of the data are being marked for further analysis. There are two varieties, Chinese Traditional (CHT) and Chinese Simplified (CHS), based on their sources.

The annotated data is then being fitted to a logistic regression model (cf. Levshina [9], Geeraerts [10]). We choose to adopt the model because logistic regression is suitable for modeling a set of binary dependent variables. In this study, the statistics returned by the logistic regression model will be examined for the analysis of the word choice between two auxiliaries.

4. Results

4.1. Evaluation

The output retrieved by the regression model is given below in Table 1, which contains several columns with different statistics.

Table 1. A Logistic Regression Analysis to the Two Chinese Causative Auxiliaries

		Model Likelihood		Discrimination		Rank Discrim.	
		Ratio Test		Indexes		Indexes	
Obs	606	LR chi2	152.46	R2	0.297	C	0.777
rang	278	d.f.	5	g	1.324	Dxy	0.555
shi	328	Pr(> chi2)	<0.0001	gr	3.759	gamma	0.600
max deriv	3e-07			gp	0.274	tau-a	0.276
				Brier	0.190		
		Coef	S.E.	Wald Z	Pr(> Z)		
Intercept		0.4308	0.2172	1.98	0.0473		
Property=inducive		-1.3660	0.3091	-4.42	<0.0001		
Property=physical		1.5052	0.2146	7.01	<0.0001		
Property=volitional		0.1508	0.3704	0.41	0.6839		
Transitivity=TR		-0.2606	0.1913	-1.36	0.1733		
Varieties=CHT		-0.8006	0.1971	-4.06	<0.0001		

As illustrated in Table 1, the column on the upper left reports the total number of observations and the frequency of each verb in our dataset.

The “Model Likelihood Ratio Test” column in the middle of the upper part of Table 1 provides an overall picture of whether the model is significant in general. In this column, one can find the Likelihood Ratio test statistic, the number of degrees of freedom, and the p -value. Since the p -value is smaller than 0.05 (< 0.0001), our model is significant, i.e. at least one predictor is significant in our model.

The rightmost column of the upper part of Table 1 contains the concordance index C , which is the proportion of the times when the model predicts a higher probability of *shi* for the sentence with *shi*, and a higher probability of *rang* for the sentence with *rang*. The statistic C in our model is 0.777. This means that for 77.7% of the pairs of *shi* and *rang* examples, the

predicted probability of *shi* is higher for the sentence where the speaker actually used *shi* than for the example where *rang* occurred. According to the scale proposed by Hosmer and Lemeshow [11, p. 162] given in Table 2 below, the discrimination in our result is acceptable.

Table 2. A Scale for the Index C (Hosmer and Lemeshow 2000: 162)

$C = 0.5$	no discrimination
$0.7 \leq C < 0.8$	acceptable discrimination
$0.8 \leq C < 0.9$	excellent discrimination
$C \geq 0.9$	outstanding discrimination

Finally, the lower part of Table 1 contains the figures of coefficients. These values represent the estimated log odds of the outcome when all predictors are at their reference levels, which correspond to affective causation, intransitive effected predicates, and CHS materials.

If the coefficient is positive, the level specified in the table boosts the chances of *shi* and decreases the odds of *rang*. If the coefficient is negative, the specified level decreases the odds of *shi* and boosts the chances of *rang*. For the predictor of Causation Property, the reference level is ‘affective’. We can see that only inductive causation has negative coefficients. This means that inductive causation decreases the odds of *shi*, and, conversely, boosts the chances of *rang*, in comparison with affective causation. Physical causation has the biggest positive estimate, so it seems to significantly boost the chances of *shi*, i.e. has a strong preference for choosing *shi* instead of *rang*, in comparison with the reference level. Transitive effected predicates seem to disfavor *shi* when compared with intransitives, though the difference is merely subtle. The odds of *shi* in the CHS variety are much higher than those in the CHT variety.

These findings can be nicely accounted for if we adopt a direct/indirect distinction [2] between *shi* and *rang*. As the verb *shi* is correlated with “direct causation”, it is typically used when inanimate participants are involved in the causing event, in which the force initiated by

the cause inevitably and directly leads to the resulted stage of the causee. Therefore, the fact that physical causation, characterized as having both a non-mental causer and a non-mental causee, particularly favors the use *shi* but not *rang* is not difficult to imagine.

In contrast, since the verb *rang* should be regarded as “indirect causation”, it is typically used in scenarios where animate participants are both involved, and some other force besides the causer becomes the most immediate source of energy in the effected event. This explains why inductive causation, which features both a mental causer and a mental causee, has a strong tendency for choosing *rang* rather than *shi*.

A plot for the outliers and discrepancy values in our dataset is provided in Figure 3 below. One can see that there are a few observations with large discrepancies and large Cook’s distance values distributed around the borders of the plot. The outliers are extracted in Table 3.

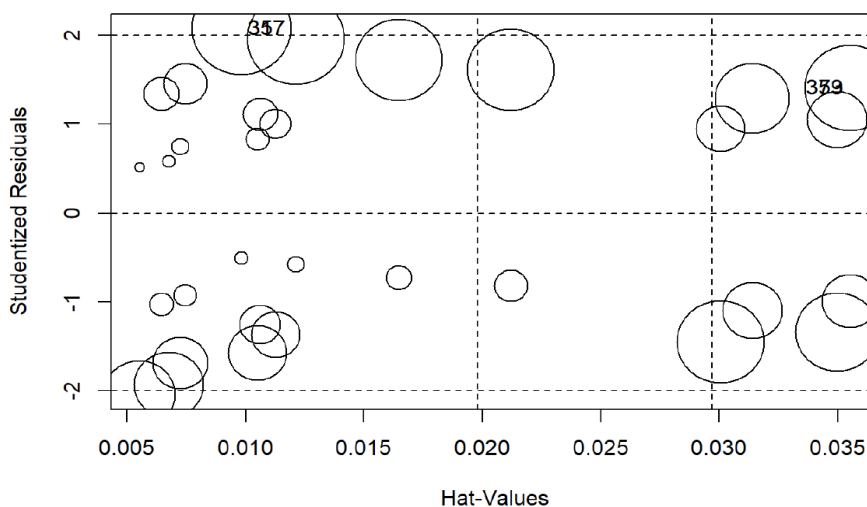


Figure 3. Plot with Outliers and Discrepancy Values

Table 3. Outliers in the Dataset

ResuPred	Property	Transitivity	Varieties
shi	inducive	TR	CHT
shi	inducive	TR	CHT

ResuPred	Property	Transitivity	Varieties
shi	volitional	TR	CHT
shi	volitional	TR	CHT

Table 3 presents contexts that are not typical of *shi*. As mentioned previously, *shi* is typically used with physical causation but not inductive or volitional causation. This is an indicator that the dataset we collected may be too coarse-grained for subtle conceptual differences, a common problem for corpus-based semantic studies.

To avoid undermining the value of the logistic regression model, overfitting is also tested, and its performance on new data is checked. The methods used in this study is to validate the model with bootstrapping (cf. Levshina [9]). The function refits the model 200 times, and the results are shown in Table 4.

Table 4. The Results of Testing for Overfitting

##	index.orig	training	test	optimism	index.corrected	n
## Dxy	0.5451	0.5462	0.5392	0.0070	0.5381	200
## R2	0.2893	0.2954	0.2822	0.0132	0.2761	200
## Intercept	0.0000	0.0000	0.0041	-0.0041	0.0041	200
## Slope	1.0000	1.0000	0.9691	0.0309	0.9691	200
## Emax	0.0000	0.0000	0.0077	0.0077	0.0077	200
## D	0.2423	0.2486	0.2355	0.0131	0.2292	200
## U	-0.0033	-0.0033	0.0002	-0.0035	0.0002	200
## Q	0.2456	0.2519	0.2353	0.0166	0.2290	200
## B	0.1922	0.1905	0.1942	-0.0037	0.1959	200
## g	1.3011	1.3168	1.2700	0.0467	1.2543	200
## gp	0.2693	0.2695	0.2642	0.0053	0.2640	200

The model is more likely to be overfitted if the ‘optimism’ of the estimates is high. As shown in Table 4, the optimism value is 0.0386 in the line with ‘Slope’, which is relatively small. It indicates that the estimates of the regression coefficients should be trustworthy.

4.2. A Comparison between Chinese and Dutch

This subsection showcases a comparison between our results and Levshina’s [9] work on the

two causative verbs *doen* and *laten* in modern Dutch. First of all, Chinese and Dutch behave very differently with respect to the transitivity of the matrix verb. Generally speaking, in Chinese, the CHT variety seems to favor the use of *rang*, i.e. indirect causation, while the CHS variety prefers to use *shi*, i.e. direct causation. However, CHT speakers tend to use *rang* when the matrix verb is transitive, whereas CHS speakers are more inclined to use *shi* when the matrix verb is transitive.

In Dutch, the indirect variant *laten* is more frequently used than the direct variant *doen* in both dialects, the reason why Geeraerts [10] regarded *laten* as the default form in causative constructions. Besides, the two dialects behave the same with respect to transitivity as both dialects are particularly more likely to choose *laten* when the main verb is transitive. The interactions between the predictors of Varieties and Transitivity in Chinese and Dutch are schematized below in Figure 2 and Figure 3, respectively.

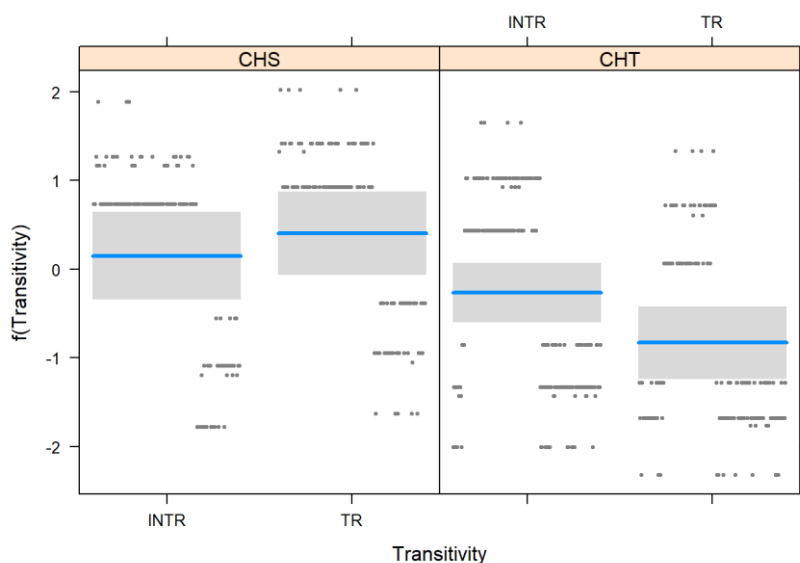


Figure 2. Interaction between Varieties and Transitivity in Chinese

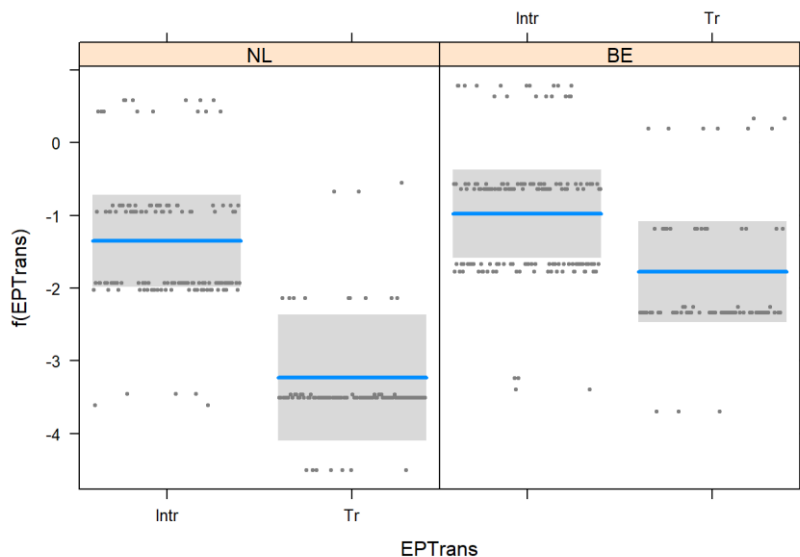


Figure 3. Interaction between Varieties and Transitivity in Dutch (Levshina 2015: 269)

The second difference between Chinese and Dutch has to do with causation types. In Chinese, physical causation tends to use *shi*, while inducive causation will opt for *rang*. Affective and volitional causation, however, have no obvious preference. No obvious difference between the two dialects is observed either. On the other hand, in Dutch, affective and physical causation are more likely to choose *doen*, while inducive and volitional causation favor *laten*. Again, no clear dialectal difference can be observed. The interactions between the predictors of Varieties and Causation Types in Chinese and Dutch are schematized below in Figure 4 and Figure 5, respectively.

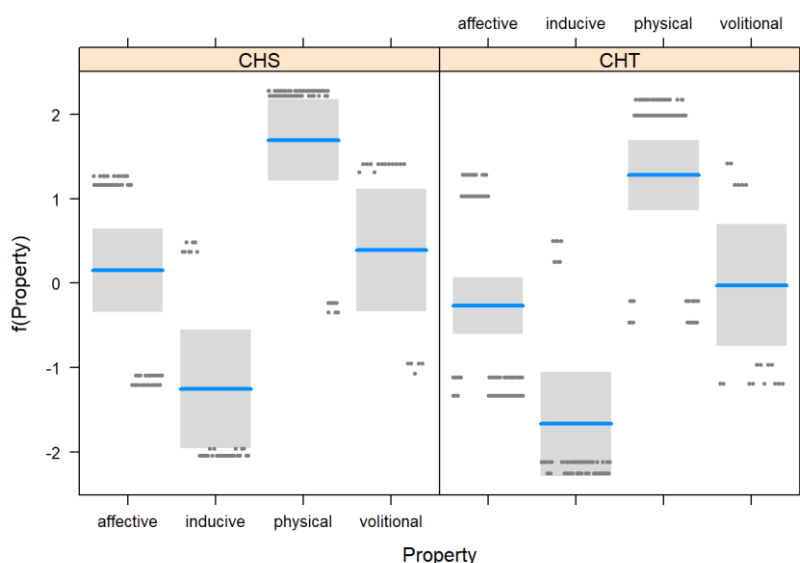


Figure 4. Interaction between Varieties and Causation Types in Chinese

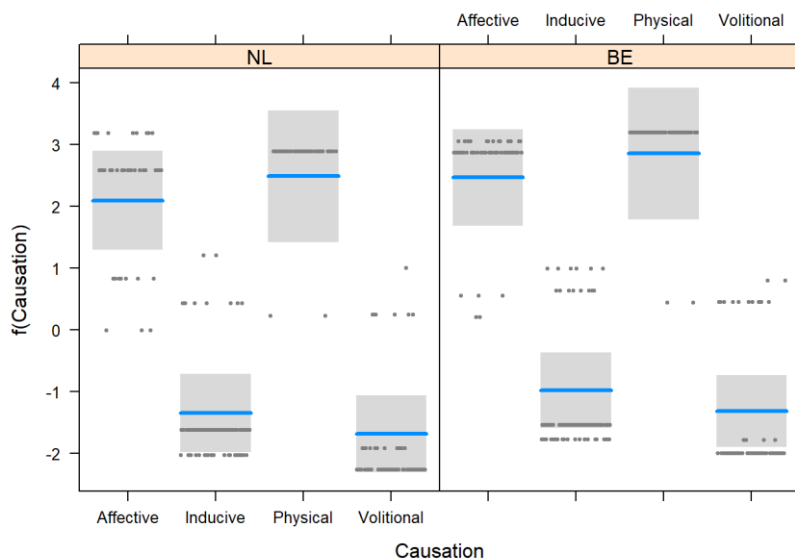


Figure 5. Interaction between Varieties and Causation Types in Dutch

5. Conclusion

Although the causative construction has been a frequently discussed subject in linguistic studies and has been widely studied in the literature, it remains unclear when it comes to the difference between its variants. This paper managed to fill the gap by providing a direct/indirect dichotomy in differentiating causation types.

We conducted a logistic regression analysis of the two Chinese causative auxiliary verbs *shi* and *rang*. The results retrieved by the regression model showed that the theory of direct/indirect causation provides a reasonable account for the characteristics and lexical meanings of the verbs. We propose that the verb *shi* is correlated with “direct causation” because it is typically used when inanimate participants are involved in the causing event, in which the force initiated by the cause directly gives rise to the resulted stage of the causee. On the other hand, the verb *rang* should be considered “indirect causation” because it is typically used in situations when animate participants are involved, and some extra force besides the causer also participates in the causal event.

In natural language processing tasks, it can be difficult to recognize how causers interact with causees in causative constructions. The results of this study explain the different usages

of direct *shi* and indirect *rang*. It is hoped that this research as well as the annotated data can make improvements to the performance in other related tasks.

To conclude, the findings help shed light on the nature of the two Chinese causative predicates. We demonstrate that the word choice between the two verbs in different contexts is influenced by the intimate relation between cognitive factors, pragmatic contextual effects, and even lexical semantics as well. A cross-linguistic survey on causation in more other languages is also necessary for future work with a view to verifying our proposal.

References

- [1] M. Shibatani, “The grammar of causative constructions: A conspectus,” in *The Grammar of Causative Constructions*, M. Shibatani, Ed. New York: Academic Press, 1976, pp. 1-42.
- [2] A. Verhagen and S. Kemmer, “Interaction and causation: Causative constructions in modern standard Dutch,” *Journal of Pragmatics*, vol. 27, no. 1, Jan., pp. 61-82, 1997.
- [3] B. Comrie, *Language Universals and Linguistic Typology: Syntax and Morphology*, Chicago: University of Chicago Press, 1989.
- [4] W. Croft, “Possible verbs and the structure of events,” in *Meanings and Prototypes: Studies in Linguistic Categorisation*, S. Tsohatsidis, Ed. London & New York: Routledge, 1990, pp. 58-83.
- [5] G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago: University of Chicago Press, 2008.
- [6] W. Croft, *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*, Chicago & London: University of Chicago Press, 1991.
- [7] L. Talmy, “Semantic causative types,” in *The Grammar of Causative Constructions*, M. Shibatani, Ed. New York: Academic Press, 1976, pp. 43-116.

- [8] L. Talmy, “Force dynamics in language and cognition,” *Cognitive Science*, vol. 12, no. 1, Jan., pp. 49-100, 1988.
- [9] N. Levshina, *How to Do Linguistics with R: Data Exploration and Statistical Analysis*, Amsterdam & Philadelphia: John Benjamins Publishing Company, 2015.
- [10] D. Geeraerts, *Ten Lectures on Cognitive Sociolinguistics*, Leiden & Boston: Brill, 2017.
- [11] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, New York: Wiley, 2000.

「忙」的語意及認知概念分析——以語料庫為本

The Semantic Features and Cognitive Concepts of *Mang2* ‘Busy’:

A Corpus-Based Study

林昕柔 Hsin-Rou Lin

國立政治大學華語文教學碩士學位學程

Master’s Program in Teaching Chinese as a Second Language

National Chengchi University

107161009@nccu.edu.tw

鍾曉芳 Siaw-Fong Chung

國立政治大學英國語文學系

Department of English

National Chengchi University

sfchung@nccu.edu.tw

摘要

本研究¹旨在探析「忙」的語意特徵以及認知概念。本文使用「COCT 書面語語料庫 2017」採不重複隨機抽樣的方式蒐得 1,314 筆語料，在排除 220 筆名詞性的「忙」、二字詞、四字詞等無效語料後，共得 1,094 筆有效語料。分析「忙」的語意特徵、歸納其語義類別，及透過分析其後方之補語結構，探討「忙」的認知概念。本研究結果顯示「忙」的語意特徵可分為「投注大量心力和時間做某事」、「在有限的時間內抓緊時間做事」、「表示密集繁多的狀態」三類。人類對於「忙」的認知概念有事件狀態、生理表現、情緒三個類別。「忙」的語意特徵和認知概念相互作用、相輔相成，展現「忙」的急迫性、消耗性、密集性。筆者期望本研究成果能為「忙」訂定更加完整的定義，補足漢語心理動詞之研究缺口，並提供研究方法予學界參考。

Abstract

This study utilized the written data from the ‘2017 Corpus of Contemporary Taiwanese Mandarin’ to analyze the semantic features and cognitive concepts of *mang2* ‘busy’. Through

¹本研究由第二作者之科技部計畫 MOST 109-2410-H-004-163 的支持與協助，特此誌謝。

random sampling, 1,314 instances were obtained, and among which, 220 instances, including the ones of *bang1mang2* ‘help’, four-character idioms, etc., were removed. Finally, there were a total of 1,094 instances analyzed. For analysis, the study firstly analyzed the semantic features and categorized the meanings of *mang2*. Next, the complements following *mang2* were analyzed. The results of the study showed that the semantic features of *mang2* can be sorted into three types: (1) to describe someone taking a great effort and time to do something; (2) to describe someone seizing the time to finish something urgently; and (3) a dense condition given a specific period of time. There are three categories of cognitive concepts of *mang2*: the state of events, physical performances, and emotions. The semantic features and cognitive concepts of *mang2* display its urgency, consumptiveness, and density. It is expected that this paper should provide a more complete interpretation of *mang2* and a corpus-based method for lexical analysis.

關鍵詞：「忙」，語意特徵，認知概念，語料庫

Keywords: 'busy', semantic features, cognitive concepts, corpus

一、緒論

臺灣《華語八千詞》將「忙」歸類為準備級一級的詞彙；在中國 2012 年公布之《新 HSK 5000 詞》中，「忙」被列入二級詞彙，是故「忙」在華語教學中為相當基礎的詞。然而字詞典對「忙」的定義皆有所出入，以臺灣的《教育部重編國語辭典》與《中文詞彙網路》的搜尋結果為例，於前者中「忙」可為形容詞、動詞、副詞和名詞，有五種義項；後者則將「忙」視為不及物動詞、及物動詞與名詞，有四種意思。由此可見，字詞典對於「忙」缺乏具系統性、詳盡的說明和定義。而漢語研究中針對表達情感、情緒的心理類詞彙亦未有深入探討「忙」意義之相關研究。

基於「忙」在字詞典與漢語研究方面的缺口，本文目的在於進行語料庫為本之研究，分析「忙」的語意特徵，進一步延伸至「忙」認知概念的剖析，並針對這兩個部分提供新的分類。本文以三個研究問題為核心進行，如下所示：

1. 漢語「忙」的語意特徵與意義為何？
2. 漢語「忙」在人類的認知概念中為何？
3. 「忙」的義項以及認知概念的分類為何？

為達到本文之研究目的並釐清三個研究問題，本研究分為語意分析、認知概念分析兩部分進行，以「忙」和「忙得」作為分析單位。第一部分於「COCT 書面語語料庫 2017」輸入關鍵詞「忙」，僅討論單字詞的「忙」，故在不重覆隨機抽樣、排除無效語料後，得到 1,094 筆有效語料，並分析「忙」的語意特徵和語義類別。第二部分同樣使用「COCT 書面語語料庫 2017」，經由分析「忙得」後方之補語研究「忙」的認知概念。

本研究結果顯示「忙」的語意特徵可分為「投注大量心力和時間做某事」、「在有限的時間內抓緊時間做事」、「表示密集繁多的狀態」三類。人類對於「忙」的認知概念有事件狀態、生理表現、情緒三個類別。「忙」的語意特徵和認知概念相互作用、相輔相成，展現「忙」的急迫性、消耗性、密集性。筆者期望本研究成果能為「忙」訂定更加完整的定義，補足漢語心理動詞之研究缺口，並提供研究方法予學界參考。

二、文獻探討

(一)「忙」的字詞典義

在進入正式研究前，本文先使用《教育部重編國語辭典》與《中文詞彙網路》搜尋「忙」字，並列出其詞類、定義和例子（表一、表二）。基於字詞典義所進行的初步觀察結果顯示，「忙」可作動詞和名詞，可能具有副詞、形容詞的特性。「忙」作名詞為姓氏，而作其他詞類時具有「事情繁多」、「緊急」、「時間不足或被事情佔據」的意義。

表一、《教育部重編國語辭典》中「忙」的字詞典義

詞類	定義	例句
形容詞	事情繁多，沒有空閒。	X
	急迫、慌張。	X
副詞	趕緊。	胡屠戶〈忙〉躲進女兒房裡，不敢出來。
動詞	做，急迫不停的做。	最近你都〈忙〉些什麼？
名詞	姓。	明代有〈忙〉義。

註：該辭典針對形容詞的「忙」僅提供複合詞，無例句。

表二、《中文詞彙網路》中「忙」的意義

詞類	定義	例句
及物動詞 (VC)	將大部分時間或心力用在後述對象上。	白天忙開會，晚上也要〈忙〉讀書。
不及物動詞 (VH)	形容事情繁多，佔據所有的時間。	我不〈忙〉，但是卻能將該做的事都做完了。
	形容心裡著急，行動快速。	陳達海〈忙〉用力關上了門。
名詞 (Nb)	專有名詞。姓。	明代有〈忙〉義。

註：本表整理自《中文詞彙網路》，請見[1]。

(二) 漢語心理情緒類詞語或狀態動詞語義之相關研究

Ng、Lin 和 Yao (2010) 以問卷的方式要求受測者評斷 372 個情感詞的情感極性，極性依極為正面至極為負面為 3~-3，0 表示中性。根據統計出來的極性強度可將研究中的情感詞彙分為偏負面極性詞（如舒坦、高興，佔 38.4%）、偏中性極性詞（如急忙、惦念，佔 33.6%）、偏正面極性詞（如暴怒、悲哀，佔 28.0%）[2]。

除了透過問卷、數據統計的方式測出詞語的情感強弱程度，也能藉由不同的方法客觀分析情感詞彙在人類認知中的概念為何。為證明「歡樂總是向上的嗎？」的說法是否成立，王海丹 (2016) 安排漢語和英語的母語者接受問卷調查，分別分析英、漢兩種語言中的 28 個動詞在例句中的位移方向以及動詞所引起的情感。前者包含「無位移或方向」、「向前」、「向後」、「向上」、「向下」、「向左或右」。後者包含「無聯繫到任何情感」、「難過／悲哀」、「寧靜／平和」、「歡樂／喜悅」、「憤怒／氣憤」、「愛戀／熱愛」、「憂愁／沮喪」、「思念／懷戀」。研究結果顯示漢語組動詞的「向上」與「歡樂」特徵之相關性顯著，漢語中向上位移常連結至歡樂的概念 [3]。

若要進一步探討心理詞語的意義和概念，可使用字詞典、語料庫等工具分析詞彙之義項，或從近義詞或多義詞方面切入，系統性地歸納出詞語的意義。謝佩璇 (2020) 對「恐懼」、「恐慌」進行語料庫為本的研究，透過句法分析和語義角色（承蒙者、客體等）探討此組近義詞的句法結構、語義及搭配詞，基於框架語義理論 (Frame Theory) 與自然後設語義理論 (Natural Semantic Metalanguage) 探析兩者的認知概念。研究結果指出

「恐懼」是「內向」的，多與表心理活動的動詞搭配，作定語時其後方接人和人類內心相關的名詞。「恐慌」主要與引起外在現象的動詞搭配，因具有「外向」的特質，作定語時表外在狀況的名詞多接於其後。在認知概念上，兩者亦有不同。須有原因才能引起「恐懼」的感覺；「恐慌」則指對於某事件背後的影響所突然產生的感覺 [4]。

關芳芳 (2016) 的博士論文分析「愛」的賓語範疇、詞素，並進行情感分析，藉此探究其語義特徵、義項和情感強弱。主要使用「中研院平衡語料庫」以「愛」為關鍵字收集 1990 年至 1997 年間之語料(共 11,796 筆)。基於「同義詞詞林檢索軟體」中以「愛」為核心語義的近義詞以及該軟體對「愛」近義詞的分類，分析語料中「愛」的句法結構、共現詞和語義。研究結果顯示「愛」能分為「親近愛」、「情侶愛」、「社會愛」、「喜好愛」、「惜護愛」、「傾向愛」六個義項 [5]。

深入探析詞彙的概念及意義，也能基於理論或模組分析詞語的各義項，並詳細說明、定義，而詞彙語義的研究可採用 MARVS 理論 (The Module-Attribute Representation of Verbal Semantics)。動詞表達事件，而事件訊息結構包含「事件類型 (Event Types)」、「特質屬性 (Inherent Attributes)」、「參與角色 (Roles)」、「角色內部屬性 (Role-internal Attributes)」。

事件類型由「過程」、「狀態」、「階段」、「端點」、「瞬時」組成，如「下雨」需要一個起因 (端點) 才能持續進行 (過程)，故以「•/////」表示事件類型。特質屬性是基於動詞的核心概念判斷事件的性質，如「哭」是意志可操控的動作，故該動詞具控制的屬性。參與角色為動詞的論元角色 (「主事者」、「受事者」、「經驗者」等)，角色內部屬性則指參與角色本身的特性，如主事者能根據自身想法、意願行事，故該角色具意志 (volition) 屬性 [6]。基於 MARVS 理論分析動詞，須透過語料庫觀察或提取語料進行句法分析，找出動詞句法功能 (調語功能、補語功能等)、論元數量與類型，觀察動詞傾向和哪些時態標記搭配，並會出現於何種句子結構中 [7]。

以狀態動詞為例，鄭語箴 (2018) 的研究主要利用《教育部重編國語辭典》及《中文詞彙網路》歸納出「大」的主要字詞典義，依此分類分析在「中研院平衡語料庫」中「大」作為狀態動詞的語料，共歸納出十三個類別：「度量衡」、「數量」、「懷孕」、「時間」、「排序」、「尊稱」、「重點」、「力量」、「程度 (深度)」、「程度 (廣度)」、「煩惱」、「搬

弄是非」、「其他」。接著進一步以 MARVS 理論分析各類「大」的義項。將中文「大」與由歸納法語字詞典而得到的法文“grand（大）”之義項進行對比，並要求法語母語者接受測驗與問卷調查，理解母語者對“grand”的認知情形。藉由兩者研究法語母語者學習中文「大」的難點並提供教學建議 [8]。

心理情緒類詞彙的意義也能以 MARVS 理論進行分析，張麗麗、陳克健與黃居仁（2000）以心理情緒類的狀態動詞「快樂」、「高興」為例，說明以 MARVS 理論分析漢語詞彙的方法，並敘述分析過程。首先初步觀察兩個詞彙的意義、詞類，使用語料庫蒐集語料並分析語法功能（名物化、謂語功能、補語功能、狀語功能、定語功能）、論元結構及語意角色（例如：「快樂」只有一個論元，而此論元為經驗者）。在事態方面的觀察，則偏向分析詞彙作不同語法功能時所搭配之成分：(1)謂語時所搭配的狀語、補語和句型、(2)作狀語或補語所配的動詞中心語、(3)作定語時所搭配的名詞中心語。為區分該組近義詞，須對比兩者之各分析項目和規律，針對兩個詞彙提供 MARVS 理論事件結構及具系統性之詳細解釋 [6]。

（三）語意分析及語意特徵

王智儀（2012）以語料庫觀察「因為」、「由於」在各文體的分布、句法功能與結構、詞彙搭配辨析該組近義詞。透過分析兩者與副詞/副詞修飾語的搭配情形，發現「因為」具「突顯事件中的肇因」的語意特徵，「由於」的語意特徵則「強調事件的過程」[9]。

賴淑芬（2008）之研究旨在辨析「懂」、「明白」、「知道」、「曉得」、「清楚」、「了解」、「理解」七個認知動詞，透過中研院平衡語料庫分析認知動詞的詞彙結構、句法功能以及與共現詞彙的搭配情形，分析這七個認知動詞的語意及句法特徵。研究結果顯示這七個動詞能依是否具有「狀態改變」、「意圖控制」或「程度比較」的語意特徵分類 [10]。

（四）認知概念分析

戴浩一（2007）認為語法形式並不全然是任意、獨立的，語言與人類認知有緊密的關係。不同個體或族群經由自身的感知、經驗等對存於客觀世界的事物、事實發展出不同的概念、語言形式以及言談策略 [11]。許多如心理空間、隱喻、語用功能等概念結

構的「映照 (mappings)」深深影響著語言的創造性 (creativity)。為達到溝通之目的，人類透過句法結構表達各種存在於大腦中的概念 [12]。

基於以上觀點，一些學者透過分析詞彙，研究特定族群語言中所蘊含的認知概念及文化。張付海、楊曉峰、方燕紅和張積家 (2016) 調查城市、農村、牧區的蒙古族大學生對紅、白、藍、綠等基本顏色詞進行分類。主要研究結果顯示城市和農村受試者將顏色分類的第一維度皆為亮色和暗色，前者的第二維度為實物色和背景色，後者則以天空色、大地色作為第二維度。牧區受試者以實物色、背景色作為第一維度，第二維度則分為亮色、暗色。對顏色詞的分類能夠反映出不同族群對於顏色的認知概念，藉此能進一步探析人類不同的生活環境與宗教文化 [13]。

王娟、張積家、劉翔、肖二平、和秀梅和盧大克 (2012) 要求白族、彝族的受試者將 62 個親屬詞分類，並以語義空間圖、分類樹狀圖呈現實驗結果，從親屬關係的性質、性別等面向觀察不同族群對於親屬的認知概念。彝族為男尊女卑的社會，男性、父系親屬詞主要分布於圖表上方，親密度也較高。白族在過去為母系社會，男性、女性在圖表的分布較對稱。從受試者分類結果也能發現白族、彝族對舅舅的重視。該研究透過親屬詞分析不同族群對親屬關係的認知概念以及社會文化特色 [14]。

基於上述狀態動詞、情感類詞彙、語意分析、認知概念的相關研究，筆者決定進行語料庫為本的研究，從語法結構、語意特徵分析的面向切入，研究漢語「忙」的意義和語意特徵，並進一步探討「忙」所表達情感、感受以及人類腦中對於「忙」的認知概念。

三、研究方法

本研究使用「國教院語料庫索引典系統」的「COCT 書面語語料庫 2017」蒐集語料，CQP syntax 指令為[word="忙" & pos="VH"]。搜尋結果共得 19,711 筆語料，點選“Thin”的欄位，以隨機不重複抽樣的設定縮減語料筆數，取 19,711 筆語料中的 1,314 筆（每十五筆語料中取一筆），並以單字詞「忙」為單位進行分析。分析過程中共排除 220 筆語料，其中包含名詞化的「忙」（如「忙裡偷閒」）、雙字詞（如「幫忙」、「忙線」）和四字詞（如「忙進忙出」）、無法辨識句義之語料，故有效語料為 1,094 筆。

利用 Excel 檔案整理 1,094 筆有效語料並分析，首先根據語法結構將語料分類，分成「忙 (V.)」、「忙+N.」、「忙+V.」、「忙於+V.」、「忙 (Adj.)」、「忙 (Adv.) +V.」、「忙得」、「忙到」及無效語料九類。之後再觀察各類語料將「忙」意義相同的併為同一類，進一步歸納出語意特徵的分類。

根據「忙」的語意、句法分析結果，抽樣語料中 10.32% (113 筆) 的「忙」與表結果和狀態的「到」或「得」連用，當中有 105 筆為「得」。因此，筆者進一步以「忙得」為單位分析其後方的補語結構。此階段使用「COCT 書面語語料庫 2017」，輸入[word="忙" & pos="VH"][word="得"] (CQP Syntax) 蒐集「忙得」的語料，共顯示 1,715 筆語料。接著點選語料庫介面“collocations”一欄，搜尋「忙得」左側第一個詞至右側第四個詞範圍中 MI 值高的詞彙。提取 MI 值前 40 高的詞並使用《教育部重編國語辭典修訂本》查詢各詞彙之意義，再依意義將詞彙分類，探究「忙」的認知概念。

四、 研究結果

(一)「忙」之語意分析與句法特徵

本研究針對單字詞「忙」進行語意分析，歸納出三類「忙」的語意特徵 (表三)。第一類表示「某生命體在某一期間內集中大量心力及時間從事某事」，偏動詞性。於此類中 33.85% 的「忙」的後方會點出生命體 (多為人) 所從事的事情，如「〈忙〉殺雞」、「〈忙〉於事業」等，而從事事務可為名詞性或動詞性之語法成分，並與「忙」連用。另外 66.14% (340 筆) 的部分則是「將生命體做的事情省略」，於這 340 筆語料中有 34.11% 的語料顯示「忙」會以「忙到」、「忙得」之形式描述生命體做事的狀態或結果，提供附加資訊，如「〈忙〉到沒時間回家」、「〈忙〉得身心俱疲」。第二類用以描述「某生命體在有限的時間內，企圖抓緊時間來進行某事」，後方多接動詞性成分，「忙」用以表示該動作的急迫性強，偏向副詞的用法。「忙」的左側時常出現如「愕然」、「驚」等表達驚慌之情緒性詞彙；「忙」右側以「說」、「道」、「問」等與口部動作相關的動詞居多。第三類表達「繁多密集」的概念，以類形容詞的功能描述修飾對象的狀態或本身的特質。依修飾對象的生命性，此類別又能再細分為兩類。當修飾對象為生命體時，「忙」用以

表示該對象在一段期間內處於事務繁多，無法分心處理其他事情的狀態，例如「醫師非常〈忙〉，不能去打擾他」。另一類以「這個星期公司特別〈忙〉」為例，當「忙」修飾無生命體時，則表示該事物本身在某一期間內具有「繁多密集」的性質。前述例句中的「公司」為無生命體，並受「忙」之修飾，表「公司」的某一性質或內容在「這個星期」（特定期間）是密集繁多的。

表三、「忙」的語意特徵

類別	百分比	語意特徵	例句
第一類 (514 筆)	46.98%	某生命體在某一期間內集中大量心力及時間從事某事。	<ul style="list-style-type: none"> • 每天一早來就〈忙〉個不停。 • 拉拉從早晨起便一直〈忙〉家務。 • 媽媽〈忙〉殺雞，弟弟、妹妹躲在房後。 • 她〈忙〉於自己看書以及做家事、給孩子做飯。 • 她〈忙〉於事業，幾乎沒有私生活。
第二類 (305 筆)	27.87%	某生命體在有限的時間內，企圖抓緊時間來進行某事。	<ul style="list-style-type: none"> • 光秀不禁愕然，〈忙〉追問原因。 • 眾人都是一驚，〈忙〉問：「怎麼打傷了張提督？」
第三類 (275 筆)	25.13%	「忙」的修飾對象為生命體：某生命體在某一期間所從事的事情繁多，無從分心於其他的事物。	<ul style="list-style-type: none"> • 醫師非常〈忙〉，不能去打擾他。
		「忙」的修飾對象為無生命體：描述無生命體自身的特質是繁多密集的。	<ul style="list-style-type: none"> • 生活實在太〈忙〉，很難撥出時間靜坐。 • 這個星期公司特別〈忙〉。

(二)「忙」之認知概念分析

根據第一節的分析結果，抽樣語料中 10.32% (113 筆) 的「忙」與表結果和狀態的「到」或「得」連用，當中有 105 筆為「得」。因此，為大量蒐集「忙得」接子句的語料，筆者於「COCT 書面語語料庫 2017」輸入[word="忙" & pos="VH"][word="得"] (CQP Syntax)，共顯示 1,715 筆語料。選取語料庫介面中“collocations”一欄，範圍設定在「忙得」左側第一個詞到右側第四個詞之間，搜尋高 MI 值的詞彙 (表四)。

表四、「忙得」後方之高 MI 值詞彙

共現詞彙	MI 值	共現詞彙	MI 值
不可開交	14.137	一塌糊塗	8.138
沾地	14.025	起勁	8.093
不亦樂乎	13.121	手忙腳亂	7.848
團團轉	12.812	團團	7.521
焦頭爛額	12.659	一天到晚	7.307
昏天暗地	11.785	大汗	7.307
分身乏術	11.666	樂趣	7.093
昏頭轉向	11.458	整天	7.075
人仰馬翻	11.237	喘	7.048
滿頭大汗	10.641	蒼蠅	6.72
暈頭轉向	10.459	效率	6.661
雞飛狗跳	10.288	沾	6.308
值不值得	10.025	天天	5.969
沒空	9.902	抽	5.709
天昏地暗	9.764	觀察到	5.68
頭昏腦脹	9.086	忘	5.02
要命	8.479	平常	5.011
陀螺	8.45	累	5.003
不得了	8.315	脫	4.948
無暇	8.193	忙	4.68

註：筆者自「COCT 書面語語料庫 2017」的搜尋結果取前 40 個 MI 值高的詞彙，並繪製此表。

本研究進一步依意義分類上述詞彙，由於無法清楚判斷「沾地」、「值不值得」、「蒼蠅」、「沾」、「抽」、「觀察到」、「脫」、「團團」與「忙」的連結性，而「忙」本身已是索引關鍵詞，故分類時排除這 8 個詞彙及「忙」一詞。經篩選過後，以《教育部重編國語辭典修訂本》逐一查詢詞彙之意義，並依此分為事件狀態、生理表現、情緒三類(表五)。事件狀態類描述事件本身或生命體(多為人)於該事件中所呈現的狀態，但該狀態強調的是事件與人的緊密關聯。從該組詞彙能發現人將「忙」視為混亂、難以脫身的狀態，事情繁多使得人無法自由使用時間去做其他事情，只能專注於手邊所進行的事務。

生理表現表生命體（多為人）的生理機制受事件刺激所致之狀態。該組詞彙反映出「忙」刺激肉體做出流汗、喘氣、頭暈等不適反應，消耗身體能量，導致疲累的感受。

情緒類則為人類對於事件的主觀感受，「不亦樂乎」、「起勁」、「樂趣」這三個詞彙皆表達說話者對於「忙」的正面感受。

表五、「忙得」後方高 MI 值詞彙之分類

類型	詞彙	《教育部重編國語辭典修訂本》的定義
事件狀態	不可開交	形容無法擺脫或結束。
	團團轉	繞來繞去。形容人著急或忙碌的樣子。
	昏天暗地	形容忙亂的情況。
	分身乏術	比喻非常繁忙，無法再兼顧他事。
	人仰馬翻	形容非常混亂騷動的樣子。
	雞飛狗跳	比喻因驚擾引來的混亂。
	沒空	沒有空閒。
	天昏地暗	形容極度的行為。
	要命	形容程度非常嚴重。
	不得了	形容程度很深。
	無暇	沒有空閒的時間。
	一塌糊塗	形容紊亂糊塗，以致不可收拾。
	手忙腳亂	形容做事慌亂，失了條理。
	一天到晚	成天、整天。
	整天	從早到晚。
	天天	日日、每天。
	平常	平時、往常。
效率	所付出之能力與所獲得之功效的比率。	
陀螺	一種木頭製的圓錐形玩具。下端有鐵尖，繞上繩子，急甩出去，落地後就能在地上直立旋轉。	
生理表現	焦頭爛額	比喻做事困苦疲勞的樣子。
	昏頭轉向	形容頭腦不清，無法冷靜思考。
	滿頭大汗	頭上流滿了汗。
	暈頭轉向	神志昏眩的樣子。
	頭昏腦脹	頭部昏暈，心思不清。
	忘	不記得。
	喘	急促呼吸。
	累	操勞、使疲勞。
	大汗	形容汗流不停的樣子。

情緒	不亦樂乎	本指喜悅、快樂。
	起勁	情緒熱烈，興致高昂。
	樂趣	趣味、情趣。

五、 結果討論

本研究分為兩部分，第一部分以「忙」為單位進行語意分析，第二部分以「忙得」為單位探討「忙」的認知概念。本文語意分析的結果顯示「忙」的意義可分為三類。第一類表示某生命體在某一期間內投注大量心力及時間做某事；第二類描述某生命體在有限的時間內，企圖抓緊時間來進行某事；第三類表示生命體所從事的事務性質或無生命體本身具「繁多密集」的特性。綜合這三個類別，可歸納出「忙」的意義具有三種特性：

(一) 急迫性：時間短，行事速度快。(二) 消耗性：大量消耗力量和時間。(三) 密集性：事情呈現多而密的狀態。

各語意特徵的類別可進一步整理出特定的語法結構(表六)。第一類的「忙」可作不及物動詞表示主語的動作；「忙」後方接動詞或名詞時，可在「忙」與其後方成分之間插入「於」，此處的「忙」作及物動詞。第二類的「忙」可於動詞前方進行修飾，表示其緊急、急迫，為副詞。第三類則類似形容詞，用來修飾主語的狀態或性質。

表六、語意特徵分類各類之語意特徵

語意特徵類型	句式與例句
第一類(動詞性成分)	忙(作不及物動詞) • 沒看到我正在〈忙〉嗎?
	忙(於)+動詞/名詞 • 這些人卻〈忙〉於 <u>養雞(動詞)</u> 。 • 丞相〈忙〉於 <u>戰事(名詞)</u> 。
第二類(副詞性成分)	忙+動詞 • 店主一見韓山進來，〈忙〉 <u>帶(動詞)</u> 著他走到小店最裡面的桌子旁。
第三類(形容詞性成分)	忙 • 有會議的話，那一定是在俱樂部最〈忙〉的時候舉行。 • 我的爸爸、媽媽好〈忙〉好〈忙〉，每天都很晚才回家。

完成「忙」的語意分析後，本研究從描述「忙」狀態的子句及詞彙探析「忙」的認知概念，了解中文母語者對於「忙」的感受及看法為何。分析「忙得」後方所接的子句

和詞組中的高 MI 值詞彙後，歸納出三類「忙」的認知概念：事件狀態類、生理表現類、情緒類。

事件狀態類詞彙顯示「忙」的狀態是混亂的，事情繁多且密集，人須投注大量時間與精力於某事件上，「無暇」、「沒空」兩個詞彙深刻體現此概念。「人仰馬翻」、「雞飛狗跳」、「一塌糊塗」等描述「忙」的詞彙除了表示狀況混亂，也傳達出說話者對「忙」的負面觀感。「陀螺」以及「團團轉」則具隱喻性功能，將人忙碌、混亂的情況連結至陀螺旋轉不停的樣子。藉由生理表現類詞彙，能發現人類對人體生理機制於「忙」之狀態的觀察和既定印象。人類認為「忙」會刺激身體做出不適的反應，並消耗精力。「暈頭轉向」及「昏頭轉向」除了顯示因「忙」所致的生理現象外，其「轉」的意象與「陀螺」相關。人類對於「忙」的主觀評價則能透過情緒類詞彙反映，儘管前兩類的認知概念有不少對於「忙」的負面觀感，但情緒類詞彙也顯示人類對於「忙」依然有正面的評價，而「不亦樂乎」的高 MI 值顯示華人文化中對於「忙」可能是看重或是讚許的，此假設需要日後進一步的探究。

六、 結論

本研究旨在探析「忙」的語意特徵以及認知概念。本研究結果顯示「忙」具有三類語意特徵，能表達大量投注心力和時間做事情，或在有限的時間內急於做某事。「忙」也能形容事物的狀態繁多密集，或是生命體置身於事情多且密的情境之中。

在「忙」的認知概念分析結果中能夠發現人類對於「忙」的認知與三種層面相關，即事件狀態、生理表現以及情緒。人們透過對三者的觀察和認識建構出「忙」的概念，而「忙」的認知概念又透過語言體現。本研究所歸納出來的語意特徵類別與認知概念分類實際上具有緊密的關聯性，兩者在急迫性、消耗性、密集性相互呼應。從認知概念²出發，事件狀態類表示「忙」是事務繁多密集且大量消耗時間、力氣的，正好對應至語意

² 於審查人予本文的寶貴建議中，提及了喚醒度 (arousal)，經查詢後，筆者找到一篇相關論文：“Analysis of Affiliation-Related Traits in Terms of the PAD Temperament Model”，Mehrabian (1997) 於此文提出情緒面向模組——“Pleasure, Arousal, Dominance” (簡稱 PAD) [15]。審查人建議「忙得」之喚醒度高於「忙」，因此「忙得」後方常出現表達生動身心狀態的詞彙。筆者對此深表認同，期望未來能以 PAD 模組深入分析「忙」的認知概念及情緒強弱。

特徵分類的第一類(表示某生命體在某一期間內投注大量心力及時間做某事),突顯「忙」的急迫性、消耗性與密集性,這三種特性的影響亦透過生理表現類詞彙體現,並與語意特徵的三種類別環環相扣。

至於認知概念分類中的情緒類則需進一步的深入研究,探討人類認知概念中的「忙」是好是壞,而在什麼情況下人類會將「忙」視為正面或負面。筆者認為這方面的研究與華人社會價值或個人經驗相關,日後需透過心理學和社會語言學的證據來深入探討。此外,漢語中「陀螺」與「忙」的連結也是值得討論的方向,能夠進一步發展與「忙」概念相關的隱喻研究。

參考文獻

- [1] 黃居仁、謝舒凱、洪嘉馥、陳韻竹、蘇依莉、陳永祥、黃勝偉,〈中文詞彙網路：跨語言知識處理基礎架構的設計理念與實踐〉,中國語文,第24卷,第二期。
- [2] Chin Loong Ng, Jingxia Lin, and Yao Yao, “Polarity of Chinese Emotion Words: The Construction of a Polarity Database Based on Singapore Chinese Speakers,” in *Chinese Lexical Semantics*, Minghui Dong, Jingxia Lin, and Xuri Tang, Eds. Cham: Springer, 2016, pp. 110-119.
- [3] 王海丹,〈歡樂總是向上的嗎?——漢語的位移動詞與「歡樂」情感的表達〉,華語文教學研究,第13卷,第4期,頁55-76,2016年12月。
- [4] 謝佩璇,《近義詞「恐懼」與「恐慌」之辨析——以語料庫為本》,碩士論文,國立政治大學語言學研究所,2020年。
- [5] 關芳芳,《漢語心理動詞「愛」字近義詞群區辨架構》,博士論文,國立台灣師範大學華語文教學系暨研究所,2016年。
- [6] 張麗麗、陳克健、黃居仁,〈漢語動詞詞彙語意分析：表達模式與研究方法〉,中文計算語言學期刊,第5卷,第1期,頁1-18,2000年2月。
- [7] Chu-Ren Huang, Kathleen Ahrens, Li-Li Chang, Keh-Jiann Chen, Mei-Chun Liu, and Mei-Chi Tsai, “The Module-Attribute Representation of Verbal Semantics: From

Semantics to Argument Structure,” *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 5, no. 1, Feb., pp. 19-46, 2000.

- [8] 鄭語箴,《法籍學習者之狀態動詞語意分析研究——以「大」之多義性為例》,碩士論文,國立台灣師範大學華語文教學系暨研究所,2018年。
- [9] 王智儀,〈關係詞「因為、由於」的語意特徵與句法功能探析——以語料庫為本的方法〉,應華學報,第10期,頁203-236,2012年5月。
- [10] 賴淑芬,〈華語「認知動詞」的語意辨析〉,漢語學報,第2008卷,第2期,頁40-45,2008年5月。
- [11] 戴浩一,〈中文構詞與句法的概念結構〉,華語文教學研究,第4卷,第1期,頁1-30,2007年6月。
- [12] James H-Y. Tai, “Conceptual Structure and Conceptualization in Chinese,” *Language and Linguistics*, vol. 6, no. 4, Oct., pp. 539-574, 2005.
- [13] 張付海、楊曉峰、方燕紅、張積家,〈生活環境和宗教文化對蒙古族基本顏色詞概念結構的影響〉,華南師範大學學報:社會科學版,第1期,頁112-118,2016年2月。
- [14] 王娟、張積家、劉翔、尚二平、和秀梅、盧大克,〈彝族人、白族人的親屬詞概念結構——兼與摩梭人的親屬詞概念結構比較〉,華南師範大學學報:社會科學版,第1期,頁45-54,2012年2月。
- [15] Albert Mehrabian, “Analysis of Affiliation-Related Traits in Terms of the PAD Temperament Model,” *The Journal of Psychology*, vol. 131, no. 1, Jan., pp. 101-117, 1997.

文本意圖的多模態分析：以 Instagram 為例

An Analysis of Multimodal Document Intent in Instagram Posts

陳盈瑜 Ying-Yu Chen
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
r06142009@ntu.edu.tw

謝舒凱 Shu-Kai Hsieh
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

摘要

時至今日，社群媒體(如 Instagram)趨向結合圖片以及文字表徵，建構出一種新的「多模態」溝通方式。利用計算方法分析多模態關係已成為一個熱門的主題，然而，尚未有研究針對台灣的百大網紅發文中的多模態圖文配對(Image-caption Pair)來分析文本意圖和圖文關係。利用文字和圖片的多模態表徵，本研究沿用 Kruk et al. (2019)的圖文關係分類方法(contextual relationship/semiotic relationship/author's intent)，對此三種分類提出新的圖文表徵方式(Sentence-BERT 及 image embedding)，並利用計算模型(Random Forest, Decision Tree Classifier)精準分類三種圖文關係，研究結果顯示正確率高達 86.23%。

Abstract

Present-day, a majority of representation style on social media (i.e., Instagram) tends to combine visual and textual content in the same message as a consequence of building up a modern way of communication. Message in multimodality is essential in almost any type of social interaction especially in the context of social multimedia content online. Hence, effective computational approaches for understanding documents with multiple modalities are needed to identify the relationship between them. This study extends recent advances in authors intent classification by putting forward an approach using Image-caption Pairs

(ICPs). Several Machine Learning algorithm like Decision Tree Classifier (DTC's), Random Forest (RF) and encoders like Sentence-BERT and picture embedding are undertaken in the tasks in order to classify the relationships between multiple modalities, which are 1) contextual relationship 2) semiotic relationship and 3) authors intent. This study points to two possible results. First, despite the prior studies consider incorporating the two synergistic modalities in a combined model will improve the accuracy in the relationship classification task, this study found out the simple fusion strategy that linearly projects encoded vectors from both modalities in the same embedding space may not strongly enhance the performance of that in a single modality. The results suggest that the incorporating of text and image needs more effort to complement each other. Second, we show that these text-image relationships can be classified with high accuracy (86.23%) by using only text modality. In sum, this study may be essential in demonstrating a computational approach to access multimodal documents as well as providing a better understanding of classifying the relationships between modalities.

關鍵詞：多模態文本分析，自然語言處理，決策樹，隨機森林

Keywords: multimodal documents understanding, contextual relationship, semiotic relationship, authors intent, Natural Language Processing, Decision Tree Classifier, Random Forest, Sentence-BERT, image embedding

1. Introduction

Up to date, a majority of representation style on social media tends to combine visual and textual content in the same message, the growing of multimodal documents is thus at a staggering rate. The developing multimodal document builds up a modern way of communication. The fact is that social multimedia, such as Instagram, inevitably hosting the way for information conveying. When reading a post on a multimodal social media, a simple question is raised: how do people analyze the relationship between image and text? Regarding the meaning of a word, we found out that it would help to know the difference between denotation and connotation, which can be something suggested or implied by a word or constructions of words. As in semiotics, the terms denotation and connotation may be seen as different ways comprising a particular semantic domain, making up of the two obligatory relata of the sign function - (1) the expression and content, and (2) of a portion of the world in correspondence of the content, that is, the referent (Wason & Jones, 1963).

To access multimodal data in a computational way, an essential inclination of current studies utilized an amount of data with annotation for solving different scales of computer science problems in data mining and multimedia (Jin et al., 2010). Besides, a substantial body of related research documents the tendency of assuming images accompanied by basic text labels or captions such as interpreting author intent (Kruk et al., 2019). With the growing volume of multimodal social media, the detection of the relationships between text and image has become more critical. Research engaging fully with the multimodal aspects of the texts and images on social media is still in its infancy. Some researchers (i.e., Castro et al. (2019), Kruk et al. (2019), O'Halloran et al. (2019), Phan et al. (2019)) considered image and text as both the primary content, viewing that incorporating multimodal features can result in Meaning Multiplication, hence improving the automatic classification. Unfortunately, there is little general consensus on how does Meaning Multiplication (Lemke, 1998) comes with online multimodal documents, and few empirical studies have been done on this issue. In addition, more people consider that ignoring images means ignoring a large portion of potential meaning (Summaries & Panel, n.d.). Castro et al. (2019) further indicated that multimodal information can reduce the relative error rate compared to the use of individual modalities, thus motivating the purposes in this study - to clarify whether there is Meaning Multiplication in the multimodal documents and to distinguish the alignment between the image and text modalities.

As a consequence, two major sets of research questions are addressed in this study. The first purpose is to describe whether the combined modality “stronger” - means which has better performance in the multimodal classification task - than individual modalities or not, and in what way and how? This study uses the text-only feature, image-only feature, and combined model from both modalities of Image-caption Pairs (abbreviated as ICPs) to develop an automatic classification system, aiming to classify the presence of the *contextual relationship* between the literal meanings (referred to what is said) of the image and caption, the *semiotic relationship* between what is signified by the image modality and text modality, and the *author's intent* hiding behind the ICPs. The second purpose is to access the relationship between text and image, and their combination given Meaning Multiplication of multimodal documents. Given the complex nature of multimodal documents, many researchers are motivated to develop a framework to classify different relationships in the multimodal documents (Chancellor et al., 2017; Illendula & Sheth, 2019; Kruk et al., 2019; Zeppelzauer & Schopfhauser, 2016). Among these recent advances, Kruk et al. (2019) proposed a

framework for a sounder theoretical bases to solve this problem. In this study, we adapt the taxonomy designed by Kruk by modifying existing taxonomies (Bateman, 2014; Marsh & White, 2003) to explore the relationship of the posts of Key Opinion Leaders (KOLs) in Instagram, and establish a new dataset with 936 posts from a variety of KOLs. We further put forward a multimodal approach that can classify the relationships between multiple modalities.

2. Literature Review

2.1 Taxonomies

Since this study focuses on exploring the ways that images and text interact, we adapt three taxonomies extracted by Kruk et al. (2019), which are possible to identify their relationships applicable to all subject areas and document types according to the closeness of the conceptual relationship. In the three proposed taxonomies introduced by Kruk et al. (2019), two (contextual and semiotic) are taken advantages to capture different aspects of the relationship between image and caption, and one to capture speaker intent (Kruk et al., 2019) while the three taxonomies address the underlying concept of Marsh and White (2003) – framing the image only as subordinate to the text. A survey was conducted to identify those three relationships in the Instagram contexts.

First, it is investigated that the **contextual taxonomy** claimed to report the relationship between the literal meaning of the image and text (Kruk et al., 2019; Marsh & White, 2003). Kruk et al. (2019) considered three categories of Marsh and White (2003) taxonomy which generally captured the closeness of the relationship between image and text essential; hence, they further generalized them to three top-level categories to make them symmetric for the Instagram domain: *minimal*, *close*, and *transcendent*. Second, to answer questions concerning the more complex forms of meaning multiplication, capturing the relationship between what is signified by the respective modalities turned out to be the priority of the **semiotic relationship**. Kruk et al. (2019) categorized the semiotic relationship between ICPs as *divergent*, *parallel*, and *additive* by taking advantage of the earlier 3-way distinction (Bateman, 2014; Kloepfer, 1976) and the two-way (parallel vs. non-parallel) classification (Zhang et al., 2018). Third, with the advantages that prior work has drawn on **author’s intent**, like Goffman’s proposal of self-presentation (Goffman et al., 1978; Mahoney et al.,

2016), eight illocutionary intents had been developed (Kruk et al., 2019). Details of the eight intents are quoted below:

Advocative: advocate for a figure, idea, movement, etc.

Promotive: promote events, products, organizations, etc.

Exhibitionist: create a self-image for the user using selfies, pictures of belongings, etc.

Expressive: express emotion, attachment, or admiration at an external entity or group.

Informative: relay information regarding a subject or event using factual language.

Entertainment: entertain using art, humor, memes, etc.

Provocative/Discrimination: directly attack an individual or group.

Provocative/Controversial: be shocking.

2.2 Multimodal Document Understanding

The literature is full of discussions surrounding the definitions of modality, and scholars have debated its nature for decades. “A modality is a communication channel, for instance, related to the human senses or the form of expression (Bongers & van der Veer, 2007)”. Modality, which most people associate sensory modalities with, is to represent our primary channels of communication and sensation, such as vision or touch (Baltrušaitis et al., 2018). The objects we see, the sounds we hear, the odors we smell are different modalities that we received in the surrounding world. For example, text and images are sometimes considered from a different modality, that is, different “modes” of communication. Text-image relations and their related work hence fall within the general area of multimodality - the investigation of diverse modes of expressions and their combinations. Additionally, the research regarding multiple modalities of document understanding is hence called multimodality document understanding.

As discussed in the last paragraph, such a combination of diverse modalities sometimes results in Meaning Multiplication, a metaphor first promoted by the socio-functional semiotician Jay Lemke (Lemke, 1998). Prior work (Bateman, 2014) states that “under the right condition, the value of a combination of different modes of meaning can be worth more than the information (whatever that might be) that we get from the modes when used alone. In other words, text ‘multiplied by’ images is more than text simply occurring with or alongside images. (...) Somehow the meanings of one and the meanings of the other resonate so as to produce more than the sum of the parts”. As for NLP perspective, Morency and

Baltrušaitis (2017) provided the view that, with the goals of recognizing language and vision projects such as image and video captioning, Multimodal Machine Learning is, inevitably, a vibrant multi-disciplinary research field for Artificial Intelligence, for example, integrating and modeling multiple communicative modalities, like linguistics, acoustic, and visual messages. In this study, we focus primarily on two modalities: the natural language that can be written, and visual signal which is represented with images. So far, seminal work on defining modality was carried out.

In order to interpret and reason about multimodal messages, it is necessary to develop a computational model that can not only deal with the heterogeneity of the data and the contingency often found between modalities but understand the dependencies across modalities. Additionally, in this study, it also requires the knowledge of the multimodal language, and thus, a multimodal framework was created and briefly introduced below. Most researchers working on exploring the relationship between text and image and extracting meaning often assigning a subordinate role to either text or images, i.e., image captioning, visual question answering, which claimed that text is a subordinate modality to image. Thus, our work builds on the framework of Marsh and White (2003) who offers a taxonomy of the relationship between image and text which Kruk et al. (2019) draw on to create a new one. With respect to this, Baltrušaitis et al. (2018) brought out five unique challenges regarding the research field of Multimodal Machine Learning as follows.

Representation: The first challenge state how to represent and summarize multimodal data to highlight the complementarity and synchrony between modalities. To construct multimodal data representations, the researcher may face several difficulties like combining the data from heterogeneous origins and dealing with different levels of data noise and missing data. In place of author intent on Instagram, almost all researchers explore classification tasks on Instagram by taking advantage of word embedding by Word2vec (Le & Mikolov, 2014). With reference to image representation, some Singla et al. (2018) exploit ResNet50 (He et al., 2016) to convert the images into vectors based on the different research purpose and essence. However, word-embedding seemed to be out of state now. With the target of modeling intra-modality dynamics, Yu and Jiang (2019) first apply Bidirectional Encoder Representations from Trans- formers (BERT) (Devlin et al., 2018) to get target-sensitive textual representations. In the context of multimodal language understanding, a

majority of multimodal research (Rahman et al., 2019) find this model outperforms several highly competitive approaches.

Translation: The second difficulty announces how to translate, or say, map, data from two or more modalities. On one hand, the data is disparate due to the way of representations between modalities. On the other hand, the relationship between modalities is often open-ended or subjective. For instance, although there are several ways of translation to explain an image, there may not be one. In addition, the evaluation and characterization of the multimodal translation may be subjective.

Alignment: The third obstacle indicates how to confirm the direct relationships between (sub)elements of instances from one modality to another. For example, given an image and a caption, the mission is to align which part of an image could be corresponded to the caption's representation. With reference to combining information from visual image and text, incorporating two synergistic modalities in a combined model is a high-efficient way adapted by most researchers, no matter in ICPs (Kruk et al., 2019), emoji-text pair (Barbieri et al., 2018), or feature-extraction for multimodal sentiment analysis (Soleymani et al., 2017). These studies, in fact, most studies, employ a competitive architecture in many image classification tasks, Convolutional Neural Networks (CNN) (Baltrušaitis et al., 2018; Lin et al., 2014; Russakovsky et al., 2015) or Random Forest (Breiman, 2001). What's more, Residual Networks (abbreviated as ResNets afterward) (He et al., 2016) is involved with CNN showed to be one of the best CNN models for image recognition. However, existing approaches to this task primarily rely on the textual content, but ignoring the other increasingly vibrant multimodal data sources, like images. This kind of ignoring will somehow enhance the robustness of these text-based models. Inspired by the recently proposed BERT architecture, a multimodal BERT architecture is applied, firstly by (Yu & Jiang, 2019), to obtain target-sensitive textual representations in order to model intra-modality dynamics.

Fusion: The fourth face-off remarks on how to combine information from two or more different modalities to perform a prediction, discrete or continuous. Take an image for example, the visual description of an image is fused with a caption to predict authors intent. The varying predictive power and noise typology may be the consequence of information coming from different modalities. Baltrušaitis et al. (2018) claim that two types of multimodal representations - joint and coordinated. Joint representation often projects

multimodal input into a common space while coordinated representation project each modality into a separate but coordinated space where only one modality is present at test time.

Co-learning: The last challenge suggests transferring knowledge between different modalities, their representation, and their predictive models. It will be unexpectedly dominant when one of the modalities has limited resources such as a lack of annotated data, noisy data, and unreliable labels.

3. Methodology

3.1 Datasets

The study comprises data labeling and text analysis of a corpus published posts from the discourse community Instagram. The primary criterion for selecting objects was that they are 100 famous Key Opinion Leader (KOLs) in Taiwan in 2019. Ten posts are crawled employing each KOL's Instagram official web page using a python package beautifulSoup. Corresponding posts from April 20 in 2020 are collected with the goal of developing a rich and diverse set of posts. Since the posts contain not only texts and images, but hashtags, emojis, and name-taggings, those features would be directly analyzing by being integrated. Although the posts include a variety of texts, images, hashtags, emojis, and name-taggings to convey information, only under two circumstances are the posts collected. First, to ensure some homogeneity of meta-data, this study only includes images of photos, rather than any short video or long video in a post. The second circumstance is that we only recruit the first photo of multiple photos in a post if have ones. Currently, an amount of 906 posts are extracted from Instagram.

3.2 Annotation

Data were pre-processed, converting all albums to single ICPs. A simple annotation toolkit built on an online google sheet was developed and displayed with the form of ICPs. The annotators, who are non-expert in linguistics, are asked to confirm whether the data was acceptable and if so, to identify the post's intent, the contextual relationship, and the semiotic relationship. Every image was labeled by at least two independent human annotators. We retained only those images on which all annotators agreed. From the collected datasets,

exploratory analysis can be conducted by an amount of analysis and visualizations. We take advantage of Cohen’s Kappa to measure the agreement between different annotators who classify 936 items into three taxonomies. The scores of Cohen’s Kappa calculated on the three taxonomies are 0.5803 (the contextual relationship), 0.5834 (the semiotic relationship), and 0.6223 (the author’s intent). The scores for three taxonomies are all complied to the moderate agreement. The scores are used to ensure that annotators have high enough reliability in giving the same degree of annotating.

3.3 Model

After obtaining and annotating the caption and image on Instagram, it is necessary to compute embeddings when working with both contextual and image data in the machine learning pipeline. For text embedding, we utilize Sentence-BERT (Reimers & Gurevych, 2019) which is pretrained character-based contextual embeddings. For image embedding, we use ResNet50 (He et al., 2016) which has a model pretrained on ImageNet as the image encoder by implementing a Python package `pic2vec` to convert the image modality to embedding. After, based on the collected datasets, two Machine Learning models are applied to train the classifiers: Decision Tree Classifier (DTC’s) (Swain & Hauska, 1977) and Random Forest (RF) (Breiman, 2001). In this study, both Machine Learning models were trained on both multimodal features. Our model takes input image (Img), text (Txt), or both (Img+Txt), plus modality-specific encoders, a fusion layer, and a class prediction layer. Consistent with our purpose that caption is seen as an integration, BERT sentence embedding fits the most because it considers caption as a whole sentence. For the combined encoding model, we take advantage of a simple fusion strategy that linearly projects encoded vectors from both modalities in the same embedding space and then adds two vectors (Kruk et al., 2019). According to Kruk et al. (2019), despite naive, this simple strategy has demonstrated high effectiveness at different related tasks. At last, we use the fused vector to predict scores with a fully connected layer.

4. Result and Discussion

We use a 906-sample dataset and only use a corresponding image and text information which is aligned manually for each post. Due to the small dataset, 10-fold cross-validation is conducted in our implementation. In order to report the result, the classification accuracy (ACC) and area under the ROC curve (AUC) are reported, using micro-average across all

classes (Jeni et al., 2013; Stager et al., 2006). Additionally, for image, we use 2048 dimensional vectors trained from scratch. For character-based embeddings, we use a pretrained model with layers resulting in a 512-dimensional vector. The results will be shown after training in DTC and RF models. Due to the small dataset, we conducted a 10-fold cross-validation. To report the result, I'll present the classification accuracy (ACC) and area under the ROC curve (AUC) with micro-average across all classes. There will be two tables summarizing the main results as follows.

Table 1. Results with DTC's models – image only (Img), text-only (Text-BERT) and combined model (Img+Txt-BERT)

Method	Contextual		Semiotic		Intent	
	ACC	AUC	ACC	AUC	ACC	AUC
Img	50.54	50.70	51.37	49.86	67.23	50.00
Txt-BERT	72.67	69.08	73.63	67.67	82.20	65.14
Combined	70.97	67.88	71.94	64.49	81.82	64.89

First, the result with Decision Tree Classifier is shown in Table 1. Overall, the result shows a striking effect of text embedding on performance. For all the taxonomies, text embedding was significantly superior to the other models. It outperforms consistently than just using images embedding and the combined model. Following Kruk's work, we hypothetically assume that the combined model would help across the board. However, it has been disproved from several aspects of the result. For the contextual and semiotic relationship, text embedding and combined model both performs much better than image model for more than 20%. Toward this result, this might because Sentence-BERT is originally designed for text classification, so it outperforms in the text classification task. As for author's intent, the performance of text embedding (82.20%) is just slightly higher than image embedding (81.02%), and reaches almost 15% difference from the combined model (67.23%). The reason might be that the data of author's intent is more homogeneous than the others. Digging into the annotation details, when annotating the author's intent, annotators tend to label the target with high identical consistency, hence making the high performance in the author's intent.

Table 2. Results with RF models – image only (Img), text-only (Text-BERT) and combined model (Img+Txt-BERT)

Method	Contextual	Semiotic	Intent
--------	------------	----------	--------

	ACC	AUC	ACC	AUC	ACC	AUC
Img	62.45	84	60.37	83	81.02	92
Txt-BERT	84.43	94	83.99	91	86.23	95
Combined	78.31	93	80.48	90	82.36	95

Next, the results with the RF models are given in Table 2. Overall, like in Decision Tree Classifier, the result also shows a powerful performance of the text embedding. For the three taxonomies, text embedding (average 84.88) was significantly superior to those other models. It outperforms than the image embedding for 16.93% and than the combined model for 4.5%. For the contextual relationship, text embedding (84.83%) and the combined model (78.31%) both perform much better than Image model (62.45%) for at least 15.86%. As for semiotic relationship, text embedding (83.99%) and the combined model (80.48%) both perform also much better than image model (60.37%) for more than 20%. Likewise, sentence-BERT again performs its advantage in the text classification experiment. As for author's intent, the performance of text embedding (86.23%) is just slightly higher than image embedding (82.36%) and the combined model (81.02%). The result of the Random Forest model tends to show its advantage in training the features, especially in the intent relationship (average 83.20%). Concerning the RF model, the author intent again displays its advantage of this task.

On one hand, compare the result of DTC's and RF, the accuracy of the text embedding of RF is even 8.71% higher than that of DTC's. The image embedding is 11.57% higher, and the combined model is 5.7% higher, in that RF is constructed on the foundation of multiple decision trees. Resulted from the growing ensemble of Decision trees, in this study, RF combines 100 tree predictors and makes them vote for the most popular class. By selecting features on randomly training and creating an amount of decision trees, it has significantly improved the classification accuracy. On the other hand, in comparison to the result of Kruk, this study performs better than Kruk's in both single modalities and combined modality. For text embedding, Kruk uses ELMO model, while this study uses Sentence-BERT architecture, showing an absolute advantage by improving the performance for 24.95%. For image embedding, this study utilizes ResNet50 instead of ResNet18 and makes progress on the performance by 15.65%. For the combined model, RF in this study outperforms Kruk's DCNN for 17.04%. Kruk has set the baseline classifier models as a preliminary effort, while this study examines dataset from a different domain and adapts more sounder classifiers, making outstanding performance in the multimodal document classification.

5. Conclusion

This study utilizes a computational model to capture the complex relationship between text and image modality, and how they cue authors intent in Instagram posts. In response to the first research question that if the combined modality stronger than individual modalities, text did assign a subordinate role to image. Toward this result, this might because Sentence-BERT is originally designed for text classification, and it performs very strong in the text classification task. Up to this point, these results are consistent with the prior studies which indicate that either text or image should be assigned a subordinate role. Although these two modalities encode different information on the use of classifying the relationship, the result suggests that the incorporating of text and image needs more effort to complement each other. The second main finding is that we present the results of the relationships of author's intent, the contextual relationship and the semiotic relationship between the ICPs. Furthermore, among the all studied modalities, the captions are no doubt the strongest feature for classifying relationships, in that the caption encoder (Sentence-BERT) shows its powerful advantages in text classification. In addition, we make two comparisons: the results between the two Machine Learning models and the result between this study and that of Kruk et al (2019).

Even though there are a variety of tasks being carried in the study, the design of the present study is not without limitations. The first limitation concerns the data size used in this current study. 906 posts might be too small to make a classification with high accuracy. The deep learning model used in these tasks needs more data to get better training. This may probably be one of the reasons that we cannot reach better performance in classifying the relationship with the image embedding model. The second limitation is rooted in the labor for annotation. The Kappa score in this study is to the average of 0.5953 (moderate agreement). If it may reach a substantial agreement (0.61-0.80) or even almost perfect agreement (0.81- 1.00), the result should appear to be better. In order to raise agreement between annotators, we should have completed more annotating norms with an abundance of details besides the two norms applied in this study. Third, a more stable architecture to build the image describer is needed. To characterize the images by representing an n-dimensional feature vector, it is necessary to explore the image describer to better generalize the results. Last but not least, the synergistic model in our task did not outperform the model of single modality. This issue might because the linearly project vectors did not highlight the significance of image embedding. Having

acknowledged the limitations above, future studies should be alerted to the disadvantages of this study.

For future studies, new possibilities can be listed. Since sentence-BERT is good at processing the text input, the literal meaning of the image automatically generated from the computer vision technique should be involved as a part of the image encoder to reach a possibly better performance. Additionally, it is suggested that future research should explore more linguistic features of multimodal documents to improve the working. More solid visual features and other meta-data features are needed. Further, considering a large amount of literature in predicting sentiments of multimodal documents, the key challenge is to collect a sufficient amount of training labels to train a discriminative model for multimodal prediction. Although preliminary research in the area is already being undertaken by researchers, more extensive research would be necessary to make any definite claims along these lines.

References

- [1] Wason, P. C., & Jones, S. (1963). Negatives: denotation and connotation. *British Journal of Psychology*, 54(4), 299–307.
- [2] Jin, X., Gallagher, A., Cao, L., Luo, J., & Han, J. (2010). The wisdom of social multimedia: using flickr for prediction and forecast, In *Proceedings of the 18th ACM international conference on Multimedia*.
- [3] Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., & Divakaran, A. (2019). Integrating text and image: determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*.
- [4] Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (an *_obviously_* perfect paper). *arXiv preprint arXiv:1906.01815*.
- [5] O'Halloran, K. L., Tan, S., Wignell, P., Bateman, J. A., Pham, D.-S., Grossman, M., & Moere, A. V. (2019). Interpreting text and image relations in violent extremist discourse: a mixed methods approach for big data analytics. *Terrorism and Political Violence*, 31(3), 454–474.
- [6] Phan, T.-T., Muralidhar, S., & Gatica-Perez, D. (2019). # drink or# drunk: multimodal signals and drinking practices on instagram, In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*.
- [7] Lemke, J. (1998). Multiplying meaning. *Reading science: Critical and functional perspectives on discourses of science*, 87–113.

- [8] Summaries, P. E., & Panel, C. (n.d.). Esrc centre for corpus approaches to social science (cass).
- [9] Chancellor, S., Kalantidis, Y., Pater, J. A., De Choudhury, M., & Shamma, D. A. (2017). Multimodal classification of moderated online pro-eating disorder content, In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.
- [10] Illendula, A., & Sheth, A. (2019). Multimodal emotion classification, In Companion Proceedings of The 2019 World Wide Web Conference.
- [11] Zeppelzauer, M., & Schopfhauser, D. (2016). Multimodal classification of events in social media. *Image and Vision Computing*, 53, 45–56.
- [12] Bateman, J. (2014). *Text and image: a critical introduction to the visual/verbal divide*. Routledge.
- [13] Marsh, E. E., & White, M. D. (2003). A taxonomy of relationships between images and text. *Journal of Documentation*.
- [14] Kloepfer, R. (1976). Komplementarität von sprache und bild am beispiel von comic, karikatur und reklame.(la complémentarité de la langue et de l'image. l'exemple des bandes dessinées, des caricatures et des réclames). *Sprache in Technischen Zeitalter Stuttgart*, (57), 42–56.
- [15] Zhang, M., Hwa, R., & Kovashka, A. (2018). Equal but not the same: understanding the implicit relationship between persuasive images and text. *arXiv preprint arXiv:1807.08205*.
- [16] Goffman, E. et al. (1978). *The presentation of self in everyday life*. Harmondsworth London.
- [17] Mahoney, J., Feltwell, T., Ajuruchi, O., & Lawson, S. (2016). Constructing the visual online political self: an analysis of instagram use by the scottish electorate, In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.
- [18] Bongers, B., & van der Veer, G. C. (2007). Towards a multimodal interaction space: categorisation and applications. *Personal and Ubiquitous Computing*, 11(8), 609–619.
- [19] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 423–443.
- [20] Morency, L.-P., & Baltrušaitis, T. (2017). Multimodal machine learning: integrating language, vision and speech, In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts.
- [21] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents, In International conference on machine learning.

- [22] Singla, K., Mukherjee, N., Koduvely, H. M., & Bose, J. (2018). Evaluating usage of images for app classificatio, In 2018 15th IEEE India Council International Conference (INDICON). IEEE.
- [23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition, In Proceedings of the IEEE conference on computer vision and pattern recognition.
- [24] Yu, J., & Jiang, J. (2019). Adapting bert for target-oriented multimodal sentiment classification.
- [25] 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.
- [26] Rahman, W., Hasan, M. K., Zadeh, A., Morency, L.-P., & Hoque, M. E. (2019). M-bert: injecting multimodal information in the bert structure. arXiv preprint arXiv:1908.05787.
- [27] Barbieri, F., Ballesteros, M., Ronzano, F., & Saggion, H. (2018). Multimodal emoji prediction. arXiv preprint arXiv:1803.02392.
- [28] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65, 3–14.
- [29] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: common objects in context, In European conference on computer vision. Springer.
- [30] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.
- [31] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- [32] Reimers, N., & Gurevych, I. (2019). Sentence-bert: sentence embeddings using Siamese bert-networks. arXiv preprint arXiv:1908.10084.
- [33] Swain, P. H., & Hauska, H. (1977). The decision tree classifier: design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3), 142–147.
- [34] Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data—recommendations for the use of performance metrics, In 2013 Humaine association conference on affective computing and intelligent interaction. IEEE.
- [35] Stager, M., Lukowicz, P., & Troster, G. (2006). Dealing with class skew in context recognition, In 26th IEEE International Conference on Distributed Computing Systems Workshops (ICDCSW'06). IEEE.

以「語言學理論」為基礎用「非機率模型」建立的數學應用問題 作答系統

A Chinese Math Word Problem Solving System Based on Linguistic Theory and Non-statistical Approach

王文傑 Wen-jet Wang
卓騰語言科技

Droidtown Linguistic Tech. Co. Ltd.
peter.w@droidtown.co

陳加容 Chia-Jung Chen
卓騰語言科技

Droidtown Linguistic Tech. Co. Ltd.
éclair.c@droidtown.co

賴建豫 Chien-yu Lai
卓騰語言科技

Droidtown Linguistic Tech. Co. Ltd.
keith.l@droidtown.co

李家名 Chia-ming Lee
卓騰語言科技

Droidtown Linguistic Tech. Co. Ltd.
trueming.l@droidtown.co

林信宏 Hsin-hung Lin
卓騰語言科技

Droidtown Linguistic Tech. Co. Ltd.
oliver.l@droidtown.co

摘要

本論文介紹一個基於語言學知識而非機率模型或類神經網路機器學習方法來實作的「中文數學應用問題解析」系統。一方面，該系統採納功能語言學學派 (Functional Linguistics) 的方法論，保存了「句式和意義 (form and meaning)」之間的連結，將句型分類做為意

圖偵測之用；另一方面，在逐步求解過程中，則採用形式語意學 (Formal Semantics) 的真值計算方式抽取事件結構中的論元 (argument) 詞彙。最後把與計算目標相關的論元取出後，傳給相應事件的函式進行數值計算。

和統計或機器學習方法相比，本論文提出的「功能-形式混合系統」不只具有極高的應用靈活性，亦能以相當少的資料進行訓練即可取得優良的意圖分類結果。使用本研究建立的模型，對國小一年級加減法及比較類的數學應用問題答題正確率達到 99.29%。

本論文提出的中文數學應用問題解析系統除提供線上操作網頁外，亦將該系統的程式原始碼公開於 Github 專案頁面中。本論文主要貢獻如下：(1) 提出一基於語言學知識解析自然語言中數學運算概念的方法與實作的系統；(2) 以系統實作證明透過適當的中文斷詞處理及 POS/NER 標記後，句型和意義之間的發散度可迅速收斂成為人類可閱讀並理解的邏輯表徵方式；(3) 提供基於台灣小學數學課本中的數學應用問題而編寫的繁體中文測試題庫，並以 MIT 授權釋出。

Abstract

Aside from statistics based and NN machine learning based approaches, this paper presents a Chinese math word problem (CMWP) solving system that is implemented with linguistic reasons. On one hand, the system adopts the functional approach to keep the relation between form and meaning for intent detection. On the other hand, its argument extraction design follows how formal semantics calculate meanings of languages.

The proposed system shows great flexibility with minimal training data requirement. When applying the model to 1st-year elementary level CMWP, the correct rate is between 98.57% and 99.29%. This paper also presents an adjustment procedure to reveal the potentials of the system to improve edging problems.

The proposed hybrid system provides an operational webpage, its source codes are also accessible on Github.com. The main contributions of this paper are listed: (1) It implements a working system that is based on linguistic knowledge to solve CMWP. (2) The system proves that with proper Chinese word segmentation and POS/NER tagging, the divergence between form and meaning can converge to a set of human-readable regular expressions. (3) The CMWP based on Taiwan elementary math textbooks are released under MIT license on Github.com.

關鍵詞：Loki，意圖偵測，數學應用文題，語言學理論，非機率模型方法。

Keywords: Loki, Intent Detection, Math Word Problem, Linguistic Theory, Non-Statistical Approach

一、緒論

自從大數據 (Big Data) 帶動的數據驅動方式興起，許多以數據驅動的統計機率方法以及機器學習方法中，都不再以「語言學方法」做為自然語言處理方法論上的首選。然而，1950 年的語言學至 2020 年之間，語言學已有七十年的發展。即便部份議題仍在研究中，但許多在 1950 年代無法描述的語言內在規則和人類語言能力的普遍表現，在今日的語言學研究中都已得到了廣泛接受的結論。

本研究採用發展自 Chomsky 提出的句子內部結構 X-bar 的觀點 [1]，並以輕動詞 (light verb) 的內部結構 [2] 和中文動詞內部事件結構的結果貌成份 [7] 實作了中文斷詞 (Chinese Word Segmentation, CWS) 及詞性標記 (Part-of-Speech, POS tagging) 的規則與流程，以此建立斷詞與詞性標記工具 - Articut [16]。以 Articut 為基礎，再依 Sinclair & Coulthard 的語言互動模式觀察 [3]，進一步打造建立意圖模型的工具 - Loki (Linguistic Oriented Keyword Interface) [17, 18]。本論文即介紹使用「意圖模型工具 Loki」來處理中文數學應用問題的流程與方法。

傳統的數學應用問題 (Math Word Problem, MWP) 研究都建立在語意的理解上。透過各種自然語言處理 (Natural Language Processing, NLP) 技術進行 MWP 的語意分析及資訊抽取，之後進行邏輯解析，將語意資訊輸出成可操作的數學邏輯結構或訓練成模型。最後再對應到數學公式並計算結果 [14]。由於近年來深度學習的方法和工具發展迅速，騰訊 2017 究嘗試將 End-to-end 模型 [11] 應用在 MWP 的研究中 [12]，走了不同於傳統語意理解的另一條路。本論文以解析數學題目「意圖」的方法來處理 MWP，試圖解析數學題目文字的語意後，再進行數學運算。是屬於語意理解方法的 MWP 研究。

此外，現有的 MWP 成果，不論用語意理解的方式或 End-to-end 模型，都運用了機率模型來訓練資料並預測結果。而本研究從底層的中文斷詞、詞性標記到建立意圖模型，均不使用任何機率模型來預測結果。綜合前述，1. 依循語言學의各種理論(包括 X-bar、語音、構詞、句法、等等規則)來實作 NLP 解析工具，且 2. 不使用機率模型，這是本研究的兩大特色。

以語言規則為基礎來處理 NLP 問題，不需要大量的訓練資料集。可以解決缺乏訓練資

料的問題；不以機率模型預測結果，讓 NLP 的處理結果有絕對的一致性。可解決機率模型預測結果無法解釋、不可追溯，難以除錯等等的問題。

和資訊技術相較之下，MWP 對人類而言是非常容易的問題。但 MWP 對 NLP 和 NLU 技術卻是極大挑戰的一個研究領域。目前所有的 MWP 研究與題庫都僅以國小數學題目為範圍 [12, 15] 都採用 Math23k 的資料集 [13] 做為評估和訓練的樣本。考量到 Math23k 資料集為簡體中文，且許多數學問題的描述文字和台灣的慣用句型有所出入。本研究另外參考國內國小一年級課本與習作內容，編寫了符合台灣的中文數學應用問題描述習慣和句型的題目。本研究已完成國小一年級程度的「加減法」目標函式做為評估和訓練用的樣本，並將模型及原始碼公開於 [ArticutAPI Github](#) 專案頁面中 [16]。本研究對此題庫的答題正確率達到 99.29% (commit code: 0a4057a)。

二、Articut 斷詞工具與 Loki 意圖模型工具

(一) Articut 斷詞與詞性標記工具

Articut 是一個商用中文斷詞及詞性、命名實體標記工具 [16]。Articut 依 Chomsky 對句子內部結構的觀點，定義了 X-bar 語言樹狀結構框架 [1]，再將輸入的中文句子由下而上地，透過多組中文詞組構詞原則 [2] 決定詞的邊界。以能最接近句法樹的最高點為輸出結果。

依句法樹的運作原則，一個詞彙被定位在句法樹上的某個節點時，其詞性亦固定下來。因此節點之間的分界，就是詞組的分界；而節點的位置，就標示了詞性的推算結果。此外，Articut 在設計上另外收集了一組用以表示台灣地址、台灣道路名稱、法條索引、網址以及金錢的字串模式 (string pattern) 以及可供使用者動態選用的 WikiData 詞條和政府公開資料中的景點名稱資訊做為外部字典以完成命名實體辨識的工作。

綜合以上流程，Articut 的斷詞、詞性標記以及命名實體辨識是同步完成而無法分割的。其輸出結果除了詞彙邊緣外，亦已隱含句法、句型資訊在內。

(二) Loki 設計原理

Loki [17, 18] 為 Linguistic Oriented Keyword Interface 之字首組合詞。其設計之目的在於透過句型比對，以及挑選語意計算時所需的詞彙做為參數，以便在保有「句型-語意」

的語言表現關係之餘，也能擷取出關鍵詞彙做為計算介面之所需。

Loki 意圖分析工具的架構，是依 Sinclair, J. & Coulthard, R.M. 1975 年在 *Toward an Analysis of Discourse*. [3] 書中所提，注意到的課堂上教師與學生之間的言語互動模式而設計。在該研究中提出的三個層次分別為：

效果 (Act)：例如「教學中與學生互動」、「考試時令學生安靜」...等效果。

功能 (Function)：例如「問與答的互動」、「點名與答有的互動」

實例 (Example)：例如「教師：『ㄎㄨㄞ 四聲？』學生：『快』」

這三個層次在 David Nunan (1993). *Introducing Discourse Analysis*. Penguin Group [4]一書中，被擴充為：

場景 (Discourse) 對應效果 (Act)

對話的語境 (Context) 對應功能 (Function)

對話的實際內容 (Utterance) 對應實體 (Example)

本工具在設計時，使用 NLP 領域較熟悉的詞彙，將對話分成三層，分別是：

專案名稱 (Project)：某組意圖適用的場景。例如在便利商店的場景，具備繳費意圖、購票意圖。但不具備住宿意圖。對應 Nunan 的「場景」。

意圖名稱 (Intent)：某一種意圖。例如在便利商店場中的繳費意圖。對應 Nunan 的語境。

語言表達 (Utterance)：一組可以用來表達某一場景下，某一意圖的語言表達。可以是完整的句子或是不完整的句子。對應 Nunan 的對話的實際內容。

首先，本研究利用 Loki 建立「數學應用問題」做為專案名稱 (Project name)，說明這一組意圖將適用於數學應用問題的語言場景 (Discourse)。接著建立「加減法」的意圖名稱 (Intent name)。在這意圖下，所有的句子都是為了描述「加減法」的語言表達 (Utterance)。例如「爸爸吃掉兩顆蘋果」、「姐姐弄破三張」或「哥哥又給他兩枝筆」...等。

透過 Loki 呼叫 Articut 進行斷詞、詞性標記與命名實體辨識處理後，所有的句子都將轉化為只保留標記做為句式辨識用的正規表式示。以「爸爸吃掉兩顆蘋果」為例：完整的模型產生流程為：

原句	爸爸吃掉兩顆蘋果
----	----------



Articut 處理結果	<ENTITY_pronoun> 爸爸 </ENTITY_pronoun>	<ACTION_verb> 吃掉 </ACTION_verb>	<ENTITY_classifier> 兩顆 </ENTITY_classifier>	<ENTITY_nouny> 蘋果 </ENTITY_nouny>
-----------------	---	---------------------------------------	---	---



Loki	<ENTITY_UserDefined>	<ACTION_verb>	<ENTITY_classifier>	<ENTITY_UserDefined>
產生模型	[^<]*?	[^<不]*?[吃掉][^<不]*?	[^<]*?	[^<]*?
	</ENTITY_UserDefined>	</ACTION_verb>	</ENTITY_classifier>	</ENTITY_UserDefined>

在 Loki 將訓練句的「爸爸吃掉兩顆蘋果」轉寫為：

```
<ENTITY_UserDefined>[^<]*?</ENTITY_UserDefined>((<ACTION_verb>[^<不]*?[吃掉][^<不]*?</ACTION_verb>)(<VerbP>[^<不]*?[吃掉][^<不]*?</VerbP>))<ENTITY_classifier>[^<]*?</ENTITY_classifier><ENTITY_UserDefined>[^<]*?</ENTITY_UserDefined>
```

的正規表示式同時，亦將動詞轉寫為以方括號標記的 [吃掉]，即可擴充兼容任何以「掉」為結尾的動詞，或是包含「吃」的動詞與動詞組來表示「減少」的意圖，同時把否定表述的「不」會造成的反向語意也予以排除。如此設計，便能用極少的資料，透過保留語言表達 (Utterance) 句型的方式來區分語意意圖，獲得最大的兼容性。

三、研究方法與流程

本研究將 Loki 意圖分類工具操作流程分成四部份來達成解數學應用問題的目標：

1. MWP 文字的斷詞及詞性標記預處理 (Articut)
2. 完成預處理後，建立 Loki 意圖模型 Loki_Math.atm
3. 設計數學運算的函式 (本研究以「加法事件」和「減法事件」來處理加減法)
4. 使用 Loki_Math.atm 進行中文數學應用問題的解題

其中流程 1 的預處理，在 Loki 意圖分析工具中會自行於後台處理。

(一) 建立 Loki 意圖模型 Loki_Math.atm

Loki 的架構分成專案名稱 (Project)、意圖名稱 (Intent) 和語言表達 (Utterance) 三層。

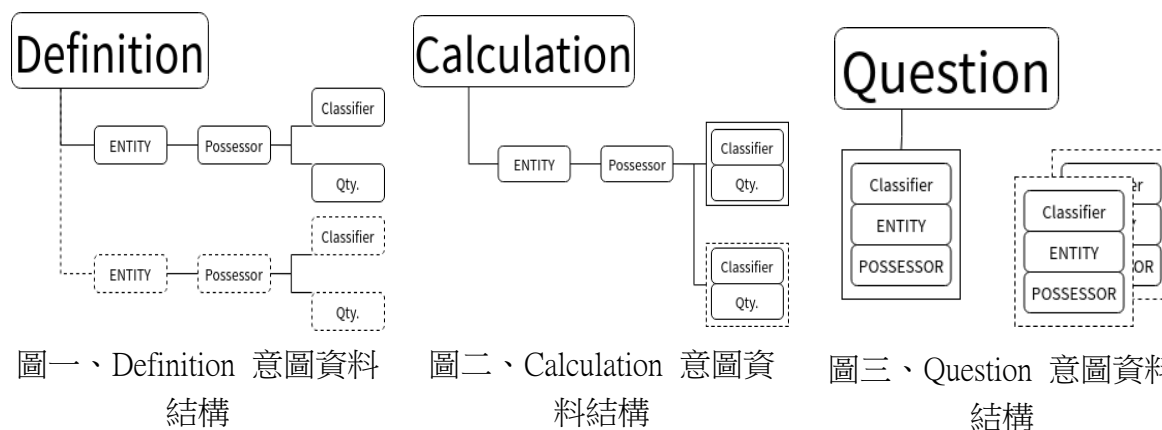
本研究建立一專案名為「國小數學」，接著透過將數學題目中的各段文字定義為三個主要意圖，分別為：

- a. 用來表示「定義語境場景 (Definition)」的意圖
- b. 用來描述「計算過程 (Calculation)」的意圖
- c. 用來說明「求解目標 (Question)」的意圖

以例題：「桌上有三顆蘋果，小明吃掉一顆，現在還有幾顆蘋果？」來說明，語境場景是「桌上有三顆蘋果」，計算過程是「小明吃掉一顆」，求解目標是「還有幾顆蘋果？」

本研究利用參考國小一年級數學課本及習作題目，另行編寫的應用問題題目 [16] 做為建立意圖模型使用。將資料集裡，所有用以表示定義語境場景的句子，例如「小明有兩枝鉛筆」或「桌上有一顆蘋果」，輸入定義語境場景 (Definition) 中建立相關意圖 (Intent) 模型；再將資料集裡用來表示計算過程的句子，例如「爸爸再給她兩瓶」或「姐姐折斷了一支」...等句子，輸入計算過程 (Calculation) 的意圖中；最後將說明求解目標的句子，例如「請問池子裡有幾隻烏龜」或「妹妹剩下多少塊蛋糕」輸入求解目標 (Question) 中建立 Loki_Math.atm 模型。有了 Loki 意圖模型，符合描述場景句型的文字，就會被 Loki 分類為 Definition 意圖。以此類推，「媽媽再給他三個」就會被分類為 Calculation 意圖，而「有幾個蘋果」則是 Question 意圖。

題目文字轉成 Definition, Calculation, Question 三種意圖的資料結構如下圖一、二、三：



在定義語境場景 (Definition) 的意圖中，依句型取出其實體名稱 (Entity)、物體持有人 (Possessor)、分類/量詞 (Classifier) 以及其數量 (Quantity)。以若無，則儲存空字串。例題中的「蘋果」是 Entity，「桌上有」是 Possessor，「顆」是 Classifier，而「一」是 Quantity。若在同一題目中有多組定義，例如「姐姐有三張黑紙，妹妹有五張白紙」，則儲存多組定義。

在計算過程 (Calculation) 的意圖中，依句型取出其實體名稱 (Entity)、物體持有人 (Possessor)、分類/量詞 (Classifier) 以及其數量 (Quantity)。若在同一題目中有多組計算過程，例如「姐姐早上搞丟三枝筆，下午去買了五枝」，則儲存多組過程以記錄發生順序。

在求解目標 (Question) 的意圖中，依句型取出其實體名稱 (Entity)、物體持有人 (Possessor)、分類/量詞 (Classifier) 以及其數量 (Quantity)。若在同一題目中有多組計算

過程，例如「姐姐總共吃了幾塊蛋糕，妹妹剩下多少蛋糕」，則儲存多組求解目標以待稍後依序求解時使用。

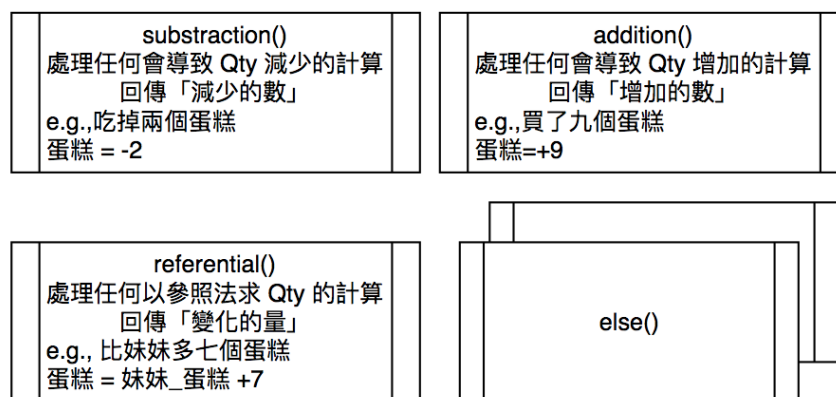
模型建立後，Loki 系統會將產生的 Loki_Math.atm 儲存於雲端，同時產生用以表示每一個「句型 - 意義」連結關係以及自動取出該句型中可供計算的「詞彙元素 (e.g., 兩顆, 蘋果)」的程式碼。

(二) 設計「加法事件」和「減法事件」的函式

基於意圖模型，我們有了可運算的詞彙元素 (e.g., 蘋果)、計量單位 (e.g., 顆) 和數字 (e.g., 二)。但欲處理「加法」和「減法」兩種計算方式，則需要定義「加法」和「減法」兩種事件的函式，以便處理當句子裡出現諸如「弟弟有 X 顆橘子，再給他 Y 顆」的 X+Y 以及「哥哥有 X 枝鉛筆，借給弟弟 Y 支」時的 X-Y 的「加法事件」和「減法事件」的需求。

不同的數學計算，例如加、減、乘、除、集合運算等等，需要定義不同的事件計算函式。不同的 MWP 研究對數學題目 (事件函式) 的分類略有不同。本論文以加法和減法為例進行說明。

在定義事件函式這個步驟中，本研究將運算的事件函式定義於事件池 (Event Pool) 中，參見圖四。各函式內有「動詞、句型組合」和「數學運算子及運算公式」的對照規則。計算意圖的文字，經過配合事件的比對，即產生該計算意圖文字所應進行的數學運算。



圖四、事件池 (Event Poll)

從「文字」到「數學邏輯」的轉換工作，本論文和現有研究 [19] 最大的不同有兩點。一是本研究不以「資料對齊算式」以求其數學意義的方式進行訓練，而是以「資料比對

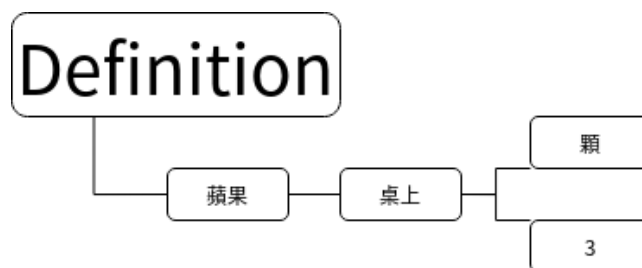
句型」以解析其數學意義的方式進行訓練；二是本研究不需依每一個動詞都重新設計一次事件框架。例如「給予」事件的框架是加法、「贈予」事件的框架也是加法、「購買」事件的框架也是加法...等。本研究直接將「加法事件」設計為一個獨立的函式，透過 Loki 將事件轉譯為正規表示式時，即能表現「在某些句型下，正規表示式可透過 (贈予|給予|購買) 的標記方式來說明這些句子的語意都是表示加法事件」的功能。如此一來，便能依不同應用的需求，只需設計最小需求的函式，置於事件池 (event pool) 中，供後續使用即可。而不需要一開始就依不同的動詞事件設計不同的框架。

(三) 使用 Loki_Math.atm 的中文數學應用問題解題系統

本論文是少數提供系統展示的 MWP 研究 [18]。本研究的系統流程，在中文數學應用問題載入後，系統先將題目中的句子一次一句送出至雲端的 Loki_Math 專案，Loki_Math 專案將句子經 Articut 處理後含有 POS/NER 標記的結果字串，再轉譯為可表示意圖的句型字串。經比對過該專案下所有的意圖內包含的句型字串後，將比對成功的意圖和句型一次回傳。

取得回傳的意圖和句型後，即能依本機程式中的句型正規表示式取出該句型中可做為語意計算論元的詞彙元素，依不同的句型表達的語意，區分事件是「加法事件」或「減法事件」，分別呼叫事件池中的函式，並將取出的論元輸入進行計算，即能求解。

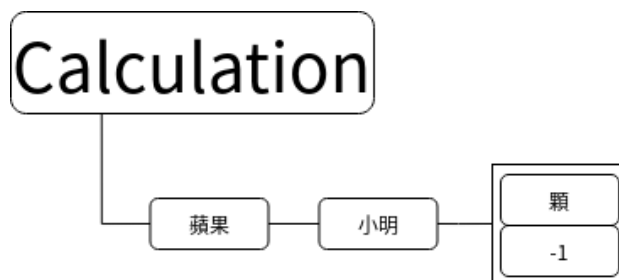
以「桌上有三顆蘋果，小明吃掉一顆，還剩下幾顆蘋果」的題目為例。依前述流程，系統將依次送出「桌上有三顆蘋果」、「小明吃掉一顆」和「還剩下幾顆蘋果」三個句子。在雲端經比對 Loki_Math.atm 專案後，「桌上有三顆蘋果」將回傳「定義語境場景 (Definition)」的意圖，以及句型中可供計算的詞彙單位為「桌上 (Possessor)」、「蘋果 (Entity)」和「三 (Quantity) 顆 (Classifier)」等三個論元。如下圖五所示：



圖五、Definition 架構儲存實際資料示意圖

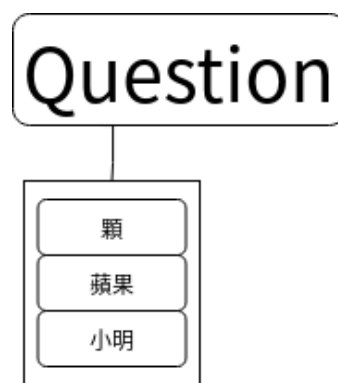
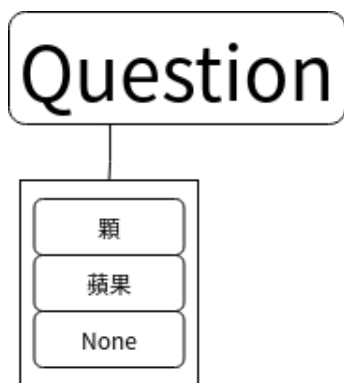
第二句「小明吃掉一顆」將回傳「計算過程 (Calculation)」的意圖，故句型中可供計算

的詞彙單位為「小明 (Possessor)」、「蘋果 (Entity)」、「一 (Quantity) 顆 (Classifier)」等論元，因「吃掉」為「減法事件」，故「一」加上負號成為“-1”，儲存如下圖六所示：



圖六、Calculation 架構儲存實際資料示意圖

最後一句「還剩下幾顆蘋果」將從雲端的 Loki_Math.atm 專案得到「求解目標 (Question)」的意圖回傳。詢問的目標是「(幾) 顆」，詢問的實體則是「蘋果」。因句子中沒有提到持有者，故將該欄位留空。如下圖七所示：



圖七、Question 架構儲存實際資料示意圖 圖八、Question 架構儲存實際資料示意圖

依求解指示，要詢問的實體是「蘋果」，故從「定義語境場景」開始進行計算，得到「蘋果 = 3 顆」的初始定義。接著進入計算過程，得到「蘋果 -1」的計算過程。因沒有其它的計算過程，故得到 $3-1=2$ ，還剩下 2 顆蘋果的最終答案。

若題目不是問「還剩下幾顆蘋果」，而是問「小明總共吃掉幾顆蘋果」，則「求解目標 (Question)」會是如圖八的結果：

則計算過程中，因為在「定義語境場景 (Definition)」的資料中不存在「小明持有蘋果」的記錄，因此會略過定義語境場景，直接進入「計算過程 (Calculation)」的資料中，取得「小明持有蘋果數為 -1」的記錄。再因為「吃掉」是一「減法事件」，因此取得的數值要再加上負號，得到“-(-1)”為“1”。故小明總共吃掉「1 顆蘋果」的最終答案。

四、資料及與實驗結果

基於三項考量，本研究沒有採用常見的 Math23k 資料集：

1. 範圍差異：Math23k 的程度包含國小全年級，超過本研究設定的國小一年級目標。
2. 解答標準：Math23k 的題目是中國普通話的譯文，內容有部份詞彙和台灣使用的國語不同。
3. 打字錯誤：Math23k 的題目中有部份句子內容繕打錯誤。

為避免前述因素干擾，本研究參考台灣的國小一年級數學課本和習作題目，另行編輯 140 題摻雜以「加減法」計算數量和透過「比大小」推估數量兩種意圖的題目。

在實驗過程中，本研究的事件池中只定義了「加法事件」(例如「媽媽又買給他兩枝」)、「減法事件」(例如「早上吃掉一個」) 以及「比大小事件」(例如「蘋果比水梨多兩顆」) 共 3 種事件。

在 140 題個題目中，共有 464 個句子，從中取出 99 句做為訓練 Loki_math.atm 模型之用。解答正確的標準有三項：

1. 數值部份需與正確答案一致，例如，若正確解答是「50 箱蘋果」，則計算結果的「50」必須與正確答案一致。
2. 單位部份需與正確答案一致。例如，若正確解答是「50 箱蘋果」，則計算結果的「箱」必須與正確答案一致。
3. 實體部份需正正確答案一致，例如，若正確解答是「50 箱蘋果」，則計算結果的「蘋果」必須與正確答案一致。

只要三個標準中未達到其中一個，就當做沒有正確理解題目語意，視為錯誤。本研究在 140 題中文數學應用問題所使用的 464 個句子中，使用了其中的 99 個句子，即能答對其中的 138 題。

透過逐步反查，可得知模型在遇到「小美有 33 張貼紙，媽媽又給她 5 張，小美現在有幾張？」的題目時，因缺乏「小美現在有幾張」的句型訓練，而只取用了「現在有幾張」的句型，由於題目文字中出現兩個人物 (小美和媽媽)，而「現在有幾張」的句型只能處理題目文字中只有一個人物的題型。因此回覆的答案就錯了。

表一、實驗數據表及實驗程式 Github commit code

	第一次實驗	第二次實驗
--	-------	-------

訓練資料佔全部資料比例 (訓練句數/全部句數)	99 / 464 = 21.33%	100 / 464 = 21.55%
模型正確率 (答對題數/全部題數)	138 / 140 = 98.57%	139 / 140 = 99.29%
ArticutAPI [16] commit code	46edb4f	0a4057a

在新增了「小美現在有幾張」的句型做為訓練材料後，訓練句型的數量從 99 提升到 100 句。而模型的答對題數則從 138 提升到 140 題。兩次實驗也凸顯出透過訓練「句型 - 意義」再取出計算論元的方式，可以輕易溯及錯誤發生原因，並加以調整模型以便容納邊緣問題的特性。實驗結果數據如表一所示。

五、相關研究

Mandal 在 2017 的回顧研究中 [14]，將 MWP 的 NLP 研究分成初期階段 (Early Stage, 1964–2006) 和新世代 (New Era, 2007–2017) 兩個階段。2007 之後的研究因為各種 NLP 和機器學習技術在語意理解上有長足的進步，使得 MWP 可處理的題型範圍擴大，方法上也更多，因此稱為新世代。在 Mandal 的回顧中，所有研究都包括四個共同的步驟：

1. 原文 NLP 前處理：對題目原文進行句法結構及語意的 NLP 處理，抽取所需的語意資訊。
2. 數學邏輯解析：將前處理過的語意資訊，輸出成用數學邏輯形式描述（或理解）的資料結構或建立類似的模型。
3. 產生對應計算方程式：將數學邏輯解析的結果轉成數學等號方程式。
4. 依答案計算結果：有了可計算結果的方程式，經過分析所問的答案後（原文前處理階段），計算結果。

然而，自 2017 至今，因為深度學習演算法蓬勃發展，有研究將端對端 (End-to-end) 模型 [11] 應用於數學應用問題。端對端 (End-to-end) 模型跳脫了上述的四個步驟，直接設計一個序列對序列 (Sequence-to-sequence/End-to-end) 模型 [12]，準備好所有「題目文字」和「對應的數學算式解答」的訓練資料，使用深度學習訓練出解題模型。這樣的方法完全不處理語意的問題，任何語言只要準備好訓練資料，即可直接訓練。騰訊在 2017 就使用 LSTM 架構訓練了中文的數學題庫 Math23k [13] 的 MWP 模型，此模型可達到 58% 的解題正確率 [12]。

由於使用端對端 (End-to-end)模型跟之前的語意理解的方法不同，因此可以算是初期階段 (Early Stage) 和新世代 (New Era) 之後的第三個階段。

除了這三個發展階段外，從 MWP 難易度來看，目前所有研究都是以國小 (elementary school) 應用數學題目為目標。從語言資料的角度來看，跟其他 NLP 研究一樣，所有 MWP 的研究都是從英文開始。中文的 MWP 則是參考英文的研究成果後，再進行調整而來。Mandal 的研究回顧 [14] 中所整理的研究，都是解英文的 MWP 研究。在針對中文 MWP 的成果上，近期的成果有許聞廉教授的研究 [19]，中研院 CKIP 團隊的研究 [10]、台大 MiuLab 的研究 [15]，以及上述的騰訊研究團隊的成果 [12]。

許聞廉教授進行過一個「小學數學深度理解解題系統 Demo」計畫 [19] 是以各種「動詞語意事件框架 (Frames)」為語意解析的核心。流程上從 MWP「問題句」開始，拉起每個陳述句的關係。再透過各個陳述句對應的框架來建立起整個數學的解題邏輯。本研究則是直接以自然語言本身做為後設語言 (meta language)，因此不是以框架做為模型目標。在實作解析時，成本較低。

表二是在 Mandal 整理的 MWP 四個步驟下，本論文的研究方法，和中研院 CKIP 團隊在 2015 年提出的研究 [10] 的比較。

表二、研究步驟比較

MWP 研究步驟 [14]	本論文研究流程	CKIP 團隊研究流程 [10]
1. 原文 NLP 前處理	Articut 斷詞、詞性標記及命名實體標記	Language Analysis module: Sinica Treebank E-HowNet
2. 數學邏輯解析	Loki 意圖模型 (Definition, Calculation, Question Structures)	Problem Resolution module: Logic Form Structure (LF)
3. 產生對應計算方程式	事件池 (Event Pool)	Problem Resolution module: Inference Engine (IE)
4. 依答案計算結果	依 Question Structure 分析所需結果並計算	Explanation Generation module (Question & Answer Mapping)

CKIP 團隊在 NLP 前處理階段，使用了中研院進行多年的中文句結構樹 (Sinica Treebank) 以及廣義知網 (E-HowNet) 來進行 MWP 題目的語意解析。中文句結構樹包含了 61,087 個中文樹圖及 361,834 個詞 [5, 6]。而廣義知網包含了九萬多個知識條目 [8, 9]。相對於本研究使用的 Articut 斷詞及詞性標記及 Loki 意圖工具，兩者皆以語言學規則建立，僅使用一個 Chomsky 提出的句法樹圖 X-bar，且不需維護龐大的詞典資源。

在「數學邏輯解析」和「產生對應計算方程式」兩個階段，CKIP 的研究均使用機率來

預測最可能的對應結果。本研究在所有階段都不使用任何機率模型，其優點是一致性高且可明確知道不能處理的邊緣問題在哪裡，亦可以持續針對未能處理的 MWP 邊緣問題逐步調整。缺點則和其它基於統計機率或機器學習方法一樣，不易在未知題目中，估計尚需處理的範圍和所需時間。

六、結論

本論文提出了一個基於「語言學知識」而「非機率模型」方法來實作的「中文數學應用問題解析」系統。方法上本研究將數學題目的意圖分成「環境背景 / Definition」、「數學計算 / Calculation」以及「問題 / Question」三種意圖。透過 Loki 建立意圖偵測模型。再搭配事件池，對計算意圖的句子進行數學運算邏輯解析。最後透過問題句子的分析，進行所需的計算並取得答案。

本研究使用台灣小學一年級的題目語料庫，並以加法和減法為範圍實作了本方法的理論，本方法對國小一年級加減法及比較類的數學應用問題答題正確率達到 99.29%。

本論文提出的中文數學應用問題解析系統除提供線上操作網頁外，亦將該系統的程式原始碼公開於 ArticutAPI Github (commit code: 0a4057a) 專案頁面中 [16]。

參考文獻

- [1] Chomsky, N. (1968). Remarks on Nominalization: Linguistics Club, Indiana University.
- [2] Chomsky, N. (1995). The Minimalist Program: Cambridge: MIT.
- [3] John McHardy Sinclair, M. C. (1975). Towards an Analysis of Discourse: Oxford Univ Pr.
- [4] Nunan, D. (1993). Introducing Discourse Analysis: Penguin Group.
- [5] Feng-Yi Chen, P.-F. T., Keh-Jiann Chen, Chu-Ren Hunag. (1999). 中文句結構樹資料庫 (Sinica Treebank) 的構建. IJCLCLP.
- [6] Chen Keh-Jiann, Y.-M. H. (2004). Chinese Treebanks and Grammar Extraction. IJCNLP.
- [7] 王文傑. (2008). 結局與結果：中文的兩種動詞後結果貌成份研究. (碩士). 國立交通大學,

- [8] Wei-Te Chen, S.-C. L., Shu-Ling Huang, You-Shan Chung, Keh-Jiann Chen. (2010). E-HowNet and Automatic Construction of a Lexical Ontology. COLING.
- [9] Shu-Ling Huang, K.-J. C. (2013). Semantic Analysis and Contextual Harmony of Durations. *Journal of Chinese Linguistics*, 41.
- [10] Yi-Chung Lin, C.-C. L., Kuang-Yi Hsu, Chien-Tsung Huang, Shen-Yun Miao, Wei-Yun Ma, Lun-Wei Ku, Churn-Jung Liao, Keh-Yih Su. (2015). Designing a Tag-Based Statistical Math Word Problem Solver with Reasoning and Explanation. Paper presented at the Computational Linguistics and Chinese language Processing.
- [11] Ian Goodfellow, Y. B., Aaron Courville. (2016). *Deep Learning*: MIT Press.
- [12] Yan Wang, X. L., Shuming Shi. (2017). Deep Neural Solver for Math Word Problems. Paper presented at the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark.
- [13] Math 23k Dataset. (2019). Retrieved from https://github.com/ShichaoSun/math_seq2tree
- [14] Sourav Mandal, S. N. (2019). *Solving Arithmetic Mathematical Word Problems: A Review and Recent Advancements*: Springer Nature Singapore Pte Ltd.
- [15] Ting-Rui Chiang, Y.-N. C. (2019). Semantically-Aligned Equation Generation for Solving and Reasoning Math Word Problems. Paper presented at the Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2019), Minneapolis.
- [16] 卓騰語言科技. (2020). ArticutAPI. Retrieved from <https://github.com/Droidtown/ArticutAPI>
- [17] 卓騰語言科技. (2020). LokiAPI. Retrieved from <https://api.droidtown.co/document/#loki>
- [18] 卓騰語言科技. (2020). Loki 算數學. Retrieved from <https://api.droidtown.co/#lokiMath>
- [19] 許聞廉. (2020). 小學數學深度理解解題系統 Demo. Retrieved from <http://iasl.iis.sinica.edu.tw/hsu/zh/小學數學深度理解解題系統 demo/>

自適應中文維度型情感詞典之建立

An Adaptive Method for Building a Chinese Dimensional Sentiment Lexicon

林應龍 Ying-Lung Lin

元智大學資訊管理學系

Department of Information Management, Yuan Ze University
fxm900206216@gmail.com

禹良治 Liang-Chih Yu

元智大學資訊管理學系

Department of Information Management, Yuan Ze University
lcyu@staur.yzu.edu.tw

摘要

在文本的情感分析(Sentiment Analysis)的任務中，基於詞典的方法因具有高可解釋性且容易使用，中文維度型情感詞典(Chinese Valence-Arousal Words, CVAW)已是重要的基礎工具，本研究的主要目的則是發展一種自適應方法(Adaptive Method)擴充該情感詞典，使其可擴充並適應到不同領域，故本研究利用深度學習的嵌入(Embedding)技術，從健保領域專家標記結果取得新詞的維度型情感(Dimensional Sentiment)，擴充中文維度型情感詞典為自適應中文維度型情感詞典。為驗證該方法之有效性，我們以中文維度型情感詞典作為基線(Baseline)，並加入支援向量機(Support Vector Machine, SVM)及極限梯度提升(Extreme Gradient Boosting, XGBoost)等熱門演算法進行比較，實驗結果顯示，自適應中文維度型情感詞典在交叉驗證實驗中之均方誤差(Mean Square error, MSE)為 0.95、皮爾森相關係數(Pearson's Correlation coefficient)為 0.71，效能略優於基線及其他機器學習演算法。未來亦將結合標記推薦系統，完善具方向性的標記及學習循環，使自適應中文維度型情感詞典能更有效的持續發展。

關鍵詞：自適應，維度型情感分析，情感詞典，情感嵌入，健保政策

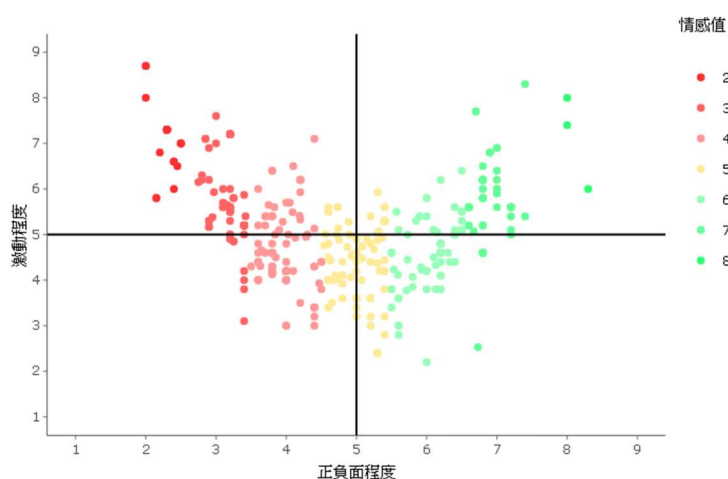
一、實驗目的

因中文維度型情感詞典建立在比較通用的領域中，可能缺少某些公共事務領域的情感詞，導致情感分析的誤差，故為擴充該詞典至不同領域，我們提出一種自適應方法，以專家標記作為基礎事實，透過深度學習的嵌入能力學習新詞的維度型情感，使不同領域可有效持續發展維度型情感詞典，強化情感分析的效能與正確性。

二、文獻探討

在文本的情感分析任務中，大致可分為基於詞典(Lexicon-based)[1]及基於學習(Learning-based)[2、3、4]等 2 種方法，基於詞典的方法因具有高可解釋性且容易使用[5]，亦可融入基於學習的方法[6]，並結合注意機制(Attention mechanism)強化模型的效能[7、8]，因此情感詞典在情感分析領域有著十分重要的地位，已是不可或缺的基礎工具。

故我們在進行中文的情感分析時，中文維度型情感詞典[9]將是有用的基礎工具，而維度型情感如圖一所示，係透過兩個維度之情感，輔助同時判斷情感之正負面及其強度，X 軸為正、負面程度(Valence, 1-9 分)、Y 軸為激動程度(Arousal, 1-9 分)，X 軸愈高分愈正面，愈低分愈負面，Y 軸愈高分愈激動，愈低分愈平靜。應用該詞典於情感分析時，相較常見的單維度情感正、負面模型，可進一步區分輿情是否激動，並利用兩維度呈正相關的特性，輔助判斷情感分析的正確性，即情感值趨正、負面極值時，通常激動值將較高，反之則較低。



圖一、維度型情感模型

中文維度型情感詞典是從較通用的領域所建立的，在特定領域中則可能因缺乏領域情感詞導致情感分析誤差，雖人工擴充詞典[10]可有效解決此問題，但標記領域情感詞成本較高，因此本研究發展一種可使維度型情感詞典自動擴充及適應到不同領域的方法。

在演算法部份，我們加入支援向量機[11]、極限梯度提升 [12]等常見的機器學習 (Machine Learning)演算法進行比較，並使用了深度學習方法[13]進行情感嵌入。支援向量機係為透過核函數(Kernel Function)嘗試將資料從低維度映射到更高維度的空間，使超平面(Hyper Plane)可在映射後的空間最小化誤差，在本研究使用徑向基函數(radial basis function, RBF)作為核函數；極限梯度提升則是決策樹(Decision Tree)集成(Ensemble)及提升(Boosting)的一種方法，透過生成弱學習器(Weak Learner)使決策樹學習資料的某個部分，並透過對誤差的監督決定新的弱學習器是否生成，當弱學習器達指定數量後再透過投票(Voting)機制集成，此演算法對特徵及學習樣本同時進行自動化調整，可使學習更為有效穩定。

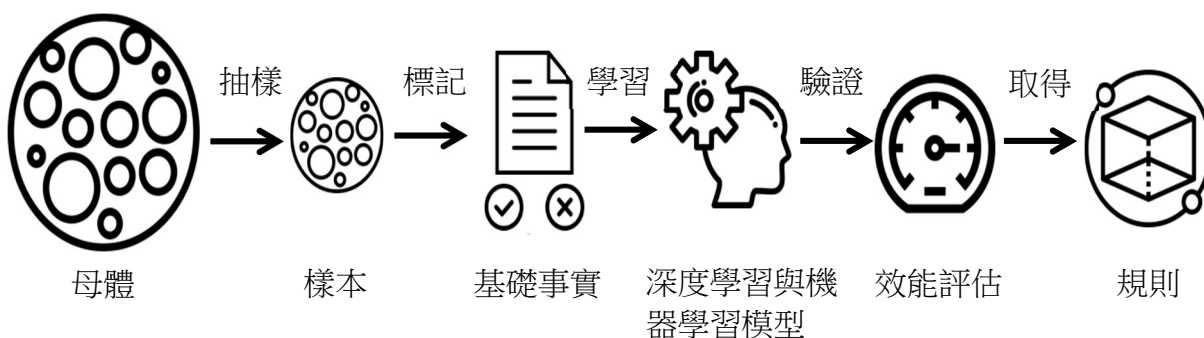
深度學習為一種神經網路架構，其目的為透過不同的架構設計及微調(Fine-tune)[14、15、16、17、18]，使端到端(End-to-End)的倒傳遞(Back-propagation)過程中，自動調整神經元的參數達成最小化誤差，而在架構上則大致可分為編碼器(Encoder)及解碼器(Decoder)等 2 個部分，編碼器部分負責從原始資料中萃取特徵，解碼器則負責從萃取完成的特徵解碼為目標值。因深度學習架構具有編碼器，其透過映射(Mapping)可保留表徵(Representation)，因此擁有優異的表徵學習能力[19、20、21]。

如詞嵌入(Word Embedding)[22、23、24]即為自然語言處理(Natural Language Processing, NLP)領域中應用深度學習取得表徵的重要方法，其透過預測前、後詞的任務調整投影層(Projection Layer)，在訓練完成後該層取出即為詞向量(Word Vector)，若某些詞在訓練文本中的前、後詞相似，則詞向量的相似度將較高，因此保留了前、後詞資訊，相較 One-hot 編碼有更多的資訊量，且因映射到固定長度的向量空間，相較 One-hot 編碼可減少運算量。

而本研究中將藉由深度學習的嵌入能力，應用在自適應中文維度型情感詞典，透過自動編碼從訓練資料中嵌入新詞的情感值，使中文維度型情感詞典可自動擴充詞典並適應到不同應用領域，強化應用範圍及效能。而情感詞典的自適應，已有研究透過基因演算法[25]或深度學習嵌入[26、27、28]，而本研究使用自動編碼[29、30、31]實現情感自適應。

三、實驗設計

為實作自適應中文維度型情感詞典，其流程如圖二所示，可分為建立基礎事實(Ground Truth)、學習規則、驗證規則及應用規則等 4 個階段，第 1 階段將使用抽樣技術取得具代表性之樣本，並進行人工標記，第 2 階段使用機器學習及深度學習演算法學習情感正負面程度與激動程度之規則，第 3 階段則以評估指標驗證規則的效能，最後的第 4 階段則是取得新詞維度型情感並應用在情感分析，透過這 4 個階段的流程，實現可持續自適應，改進情感分析效能與正確性。



圖二、自適應中文維度型情感詞典實驗流程

在資料部分我們透過網路問卷方式取得真實健保政策輿情共 6,920 則輿情，並經過縮減樣本為 1,200 則後，由國立陽明大學醫務管理研究所熟悉健保政策領域的 2 名研究人員標記資料，將以此作為基礎事實，評估中文維度型情感詞典的效能並作為基線。接著我們評估了深度學習與機器學習模型，並從深度學習的嵌入隱藏層(Latent layer)取得新詞，新詞即中文維度型情感詞典中尚未被收錄的詞，透過增加新詞將中文維度型情感詞典擴充為自適應中文維度型情感詞典。

在標記前我們考量人工標記成本高昂，為了使標記工作更有效率，採用了 4 個步驟的樣本縮減方法，第 1 個步驟為去除句子長度低於或高於 2 倍標準差的句子，使句子的長度適中，較適合用於後續標記及學習情感。第 2 個步驟為去除相似句，此步驟係為減少重複標記，以 One-hot 編碼為詞向量後計算歐幾里得距離(Euclidean Distance)，並將小於兩倍標準差的結果視為相似樣本，只保留其中 1 個。第 3 個步驟為符合情感分布，避免隨機抽樣有過度集中在負面或正面情感的狀況，我們先使用中文維度型情感詞典預先計算每一個句子的情感正負程度及激動程度，再分別計算各組別的平均值及標準差等統計值，每組抽樣 100 筆，並以抽樣後的統計值與原統計值計算誤差，作為目標函數，經 1000 次隨機抽樣後取目標函數之最佳結果，此方法確保抽樣後情感正負程度與激動程

度的統計值與原統計值較相近。第 4 個步驟將每組樣本的句子隨機排序後合併，如表一所示，6 個組別在 106 年有 600 筆資料、107 年有 600 筆資料，共 1,200 筆資料。

在經過熟悉健保政策領域的 2 名研究人員標記完成後(結果如表二)，我們使用平均值、標準差及皮爾森相關係數觀察標記結果。皮爾森相關係數可用以觀察兩組標記的相關程度，該係數之區間為 0 至 1，越接近 1 代表越相關。其中情感正負程度的相關程度較高，標記較一致；而激動程度的標記則相關程度較低，從原始標記資料可看出 2 名研究人員的標記較容易出現相反結果，且標記均集中在 4.5 到 5.5 區間，故標準差較低。

表一、健保政策問卷樣本縮減統計

問卷來源	年度	蒐集數	去除句子長度 離群值	去除相似句	抽樣
中醫	106	645	611	556	100
	107	547	534	490	100
牙醫	106	698	660	626	100
	107	572	547	511	100
全民健保	106	673	645	610	100
	107	624	591	523	100
西醫基層	106	628	593	562	100
	107	269	256	237	100
門診透析	106	532	502	497	100
	107	527	500	479	100
醫院	106	666	640	614	100
	107	539	515	473	100
總計		6,920	6,594	6,178	1,200

表二、健保政策問卷樣本標記統計

標記種類	樣本數	平均值	標準差	相關係數
Valence	1,200	5.3613	1.4243	0.8672
Arousal	1,200	5.3327	1.0248	0.3313

在斷詞部分我們使用結巴(Jieba)斷詞，並自建含 8 萬多個詞的自定義詞典，其權重以長詞優先，並加入領域相關的特定用語，以提升情感分析的正確性，減少因無法正確斷詞而造成的情感值誤差。在效能的實驗設計部分，我們使用預訓練的詞嵌入進行編碼後，評估極限梯度提升、支援向量機及深度學習等 3 種演算法，亦評估新詞及自適應中文維度型情感詞典等 2 種字典法，並以中文維度型情感詞典作為基線。

評估指標部分，連續型指標為均方誤差及皮爾森相關係數，均方誤差可評估實際值與預測值的誤差大小，誤差越小代表模型學習效果越佳，皮爾森相關係數可評估實際值與預測值之間的相關程度，避免演算法僅以特定值縮小誤差的未正確學習狀況。

我們接著將情感正負程度之連續值依據標記結果之平均值(5.3)離散化為偏向正面(5.3~9)及偏向負面(1~5.3)2 個類別，以離散型指標準確值(Accuracy)、精確值(Precision)、召回值(Recall)及 F 值進行評估，透過離散化指標我們可進一步觀察模型在不同類別的效能。準確值為全類別的效能評估，準確值的區間為 0 至 1，越高代表越準確；精確值、召回值及 F 值則為各類別的效能評估，區間亦為 0 至 1，精確值高代表預測類別與實際類別相符的比例高，召回值高則代表實際類別可被預測出來的比例高，而 F 值則是精確值與召回值的調和分數，F 值越高代表精確值與召回值的平均效能越高。

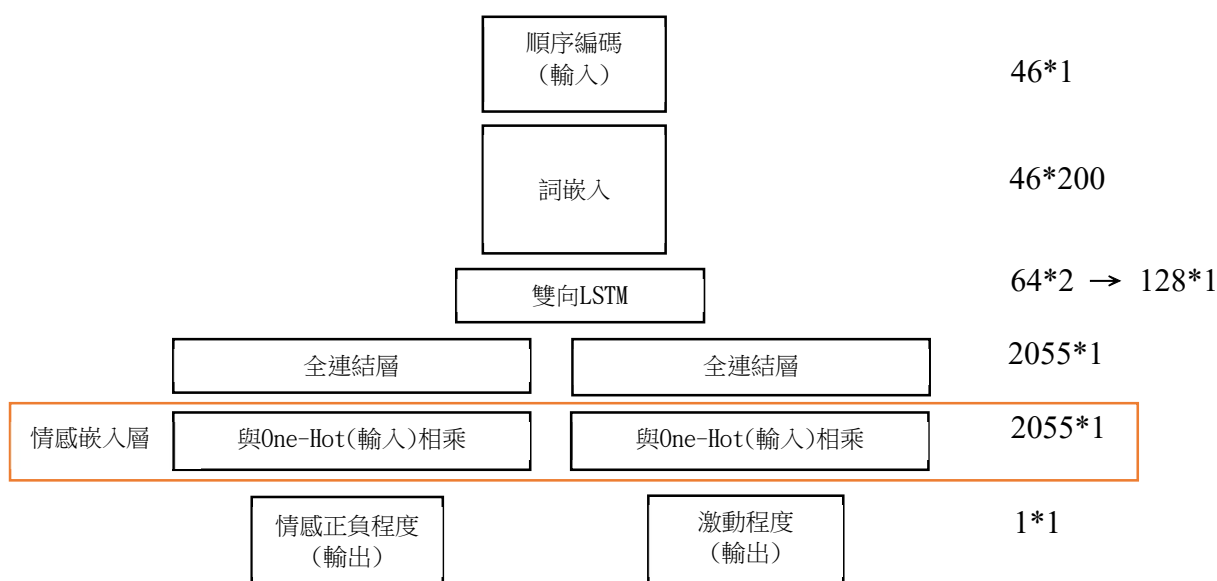
為了確認各模型是否能有效學習，我們第 1 組實驗先將全部的資料作為訓練集及測試集，各評估指標越高則代表學習效果越佳。而使用已知資料學習出來的模型預測已知資料，可能會有過度擬合的狀況發生，即模型過度合適特定資料導致無法用於其他資料，為了避免該狀況，第 2 組實驗進行 5 折交叉驗證，模擬測試集未被學習過的狀況，與第 1 組實驗相比，各模型效能通常會下降，若效能下降幅度過大即可能為過度擬合，導致模型應用在新資料的狀況較不理想。

在確定各模型的學習效能後，我們會希望用在新資料的預測上，但考量機器學習與深度學習皆為啟發式演算法，目的是針對因變數自動找出合適的自變數函數組合，因此處於一種難以被觀察的黑箱模式，不如字典法來的直觀及容易解釋。因此我們透過深度學習的嵌入技術，在學習過程中調整情感嵌入的隱藏層，定位每個詞對短句情感值的貢獻，再將隱藏層的權重取出後正規化(Normalization)，即為詞的維度型情感。透過前述嵌入技術取得新詞後，我們就可以評估新詞及自適應中文維度型情感詞典之效能，並與機器學習演算法效能進行比較，分析自適應方法是否有效。

本實驗當中所使用之深度學習架構如圖三所示，係將句子經過順序編碼後，使用預訓練的 200 維詞嵌入取得詞向量(46*200)，其中 46 代表句子的最大長度，經過長度為 64 的雙向長短期記憶(Long Short-Term Memory, LSTN)編碼後取得長度為 64*2 的向量，再將該向量連接成為 128*1 的向量作為情感程度及激動程度的共用編碼，解碼部分則使用情感正負程度及激動程度等 2 組全連結層，長度均為詞總數 2055*1，最後以長度為 1*1 的

全連結層作為輸出。另為定位詞的權重，取出新詞的模型會加上與句子輸入 One Hot 編碼相乘之隱藏層，長度同樣為 $2055*1$ ，使權重以詞存在與否進行調整。

前述深度學習模型的設計，在詞嵌入部分使用預訓練的詞嵌入，是因為本研究樣本數較少，在大量文本預訓練的詞嵌入可保留較好的詞位置資訊，而雙向長短期記憶則是透過由前往後及由後往前的編碼取得順序資訊，綜合兩者則為輸入的語意資訊編碼，此編碼器之設計在自然語言處理中已廣泛被使用。而解碼器部份我們假設輸入的情感正負程度及激動程度係每個詞貢獻而成的，因此使用全連結層乘上 One-Hot，定位每個詞的位置及貢獻程度，最後將貢獻程度正規化即為每個詞的維度型情感，此假設單純而易解釋。



圖三、自適應中文維度型情感詞典深度學習架構

四、實驗結果

在第 1 組實驗中，我們先以全部資料進行訓練及測試，確認資料是否可被有效學習及演算法是否正確學習，並以連續型指標及離散型指標評估學習結果。測試結果如表三所示，以極限梯度提升最佳，其次為支援向量機及深度學習，效能皆優於中文維度型情感詞典。

接著我們取出隱藏層，並將值正規化至原情感的尺度(Scale)，將其作為新詞的情感，透過字典法評估，新詞效能亦優於中文維度型情感詞典，因新詞係透過學習標記資料後所擷取出來的，而中文維度型情感詞典則是在通用領域取得，尚未適應至健保政策領域。

而考量新詞僅在本資料中訓練後擷取，為增加通用性，我們結合新詞與中文維度型情感

詞典為自適應情感詞典，僅加入尚未存在於中文維度型情感詞典的新詞，並去除單字詞，該詞典之評估結果顯示情感正負程度效能些許提升，說明中文維度型情感詞典在情感正負程度可能有補充效果。

而為了進一步觀察不同情感正負程度的效能，我們將值離散化至正面及負面，如表四所示，可觀察到新詞在負面效能高於正面，因此我們假設在本份資料負面較容易學習，並在後續進一步分析為何負面較容易學習。

在第 2 組實驗中，我們進行 5 折交叉驗證，確認演算法在學習時沒有過度擬合。結果如表五所示，以支援向量機最佳，極限梯度提升可能因過度擬合而使效能下降，而深度學習效能下降幅度最低，可能具有較佳的穩定性，3 種演算法雖效能下降但同樣優於中文維度型情感詞典。

表三、訓練測試連續型指標評估

演算法	均方誤差 (Valence)	相關係數 (Valence)	均方誤差 (Arousal)	相關係數 (Arousal)
極限梯度提升	0.0006	0.9998	0.0006	0.9996
支援向量機	0.3984	0.8953	0.2032	0.8513
深度學習	0.4650	0.8731	0.2518	0.8000
新詞	0.9593	0.7099	0.5180	0.4924
自適應中文維度型情感詞典	0.9503	0.7091	0.6209	0.3903
中文維度型情感詞典	1.4837	0.5628	1.4472	0.1454

表四、訓練測試離散型指標評估

演算法	準確值	情感正負	精確值	召回值	F值
極限梯度提升	0.9992	正面	1.0000	0.9984	0.9992
		負面	0.9983	1.0000	0.9991
支援向量機	0.9200	正面	0.9111	0.9322	0.9216
		負面	0.9294	0.9076	0.9184
深度學習	0.8817	正面	0.8546	0.9105	0.8817
		負面	0.9105	0.8546	0.8817
新詞	0.8400	正面	0.7900	0.8875	0.8359
		負面	0.8933	0.7997	0.8439
自適應中文維度型情感詞典	0.8083	正面	0.7351	0.8733	0.7982
		負面	0.8864	0.7585	0.8175
中文維度型情感詞典	0.7008	正面	0.6947	0.7167	0.7055
		負面	0.7074	0.6850	0.6960

表五、交叉驗證連續型指標評估

演算法	均方誤差 (Valence)	相關係數 (Valence)	均方誤差 (Arousal)	相關係數 (Arousal)
極限梯度提升	1.1807	0.6107	0.4990	0.4931
支援向量機	0.9656	0.6962	0.4546	0.5526
深度學習	0.9730	0.7017	0.5229	0.4987
新詞	1.4498	0.4825	0.6278	0.2429
自適應中文維度型情感詞典	0.9503	0.7103	0.6209	0.3895
中文維度型情感詞典	1.4837	0.5647	1.4472	0.1483

在離散化後，如表六所示，我們在本組實驗亦重複觀察到新詞負面效能高於正面效能，再次說明本實驗資料可能在負面較容易學習。另值得注意的是中文維度型情感詞典在正面的效能略優於負面，與演算法學習後的結果恰好相反，可能係因中文維度型情感詞典在健保政策領域中缺乏有效的負面詞。

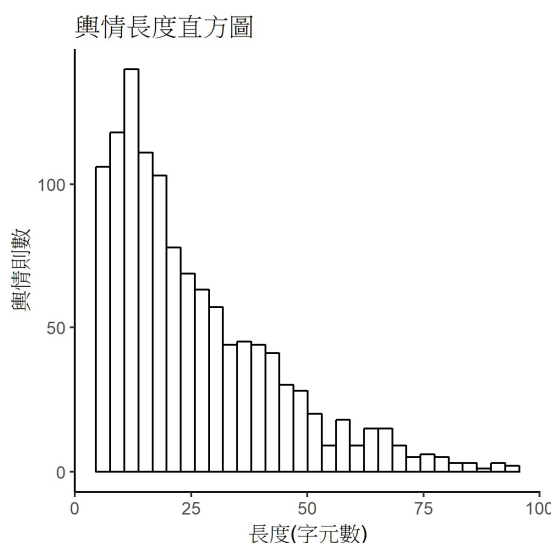
表六、交叉驗證離散型指標評估

演算法	準確值	情感正負	精確值	召回值	F值
極限梯度提升	0.7458	正面	0.7779	0.7422	0.7592
		負面	0.7119	0.7501	0.7300
支援向量機	0.7950	正面	0.7643	0.8238	0.7926
		負面	0.8263	0.7685	0.7961
深度學習	0.7992	正面	0.7701	0.8290	0.7966
		負面	0.8261	0.7755	0.7988
新詞	0.6875	正面	0.5613	0.7732	0.6444
		負面	0.8173	0.6399	0.7154
自適應中文維度型情感詞典	0.8083	正面	0.7353	0.8747	0.7979
		負面	0.8862	0.7579	0.8162
中文維度型情感詞典	0.7008	正面	0.6949	0.7170	0.7052
		負面	0.7082	0.6852	0.6959

字典法在效能一般難以超越機器學習及深度學習，因字典法每個詞只能對應一個值，即其表徵複雜程度較低，而機器學習及深度學習則是映射到高維空間的複雜函數，可處理的複雜程度並不相同。舉例來說，反諷詞在字典法仍為正面，即字典法無法分辨同一個詞在不同句子中的情況，但機器學習與深度學習則可透過句子中不同詞組成複雜的規則，區分詞在不同句子中的差異，即軟表徵(Soft Representation)可保留較多資訊量，使用詞嵌入取代詞袋模型即為其中一種應用。而經本實驗微調及適應後，自適應中文維度型情感詞典已貼近健保政策領域，在效能表現上略優於機器學習中效能最佳之支援向量機。

為了進一步釐清為何在負面學習效果較佳、正面學習效果較差，我們針對 1,200 則健保政策輿情進行統計及視覺化，分析資料的差異並檢視對學習結果的影響。首先是字元數的分布狀況，1,200 則健保政策輿情字元數之平均值為 25.73 個字元、標準差為 17.75 個字元，可從圖四中看出呈右偏態，資料集中在左側。

為了觀察字元少及字元多的差異，我們定義及篩選短輿情及長輿情，短輿情為平均值減標準差，即 7.98 字元以下，長輿情為平均值加標準差，即 43.48 字元以上，並觀察該 2 組輿情在正負面程度與激動程度的分布情況。從圖五中可得知短輿情在正面多於負面，而長輿情則相反，分析結果與常識相符：「滿意的人少回覆，不滿意的人多抱怨」。



圖四、健保政策輿情字元數分布

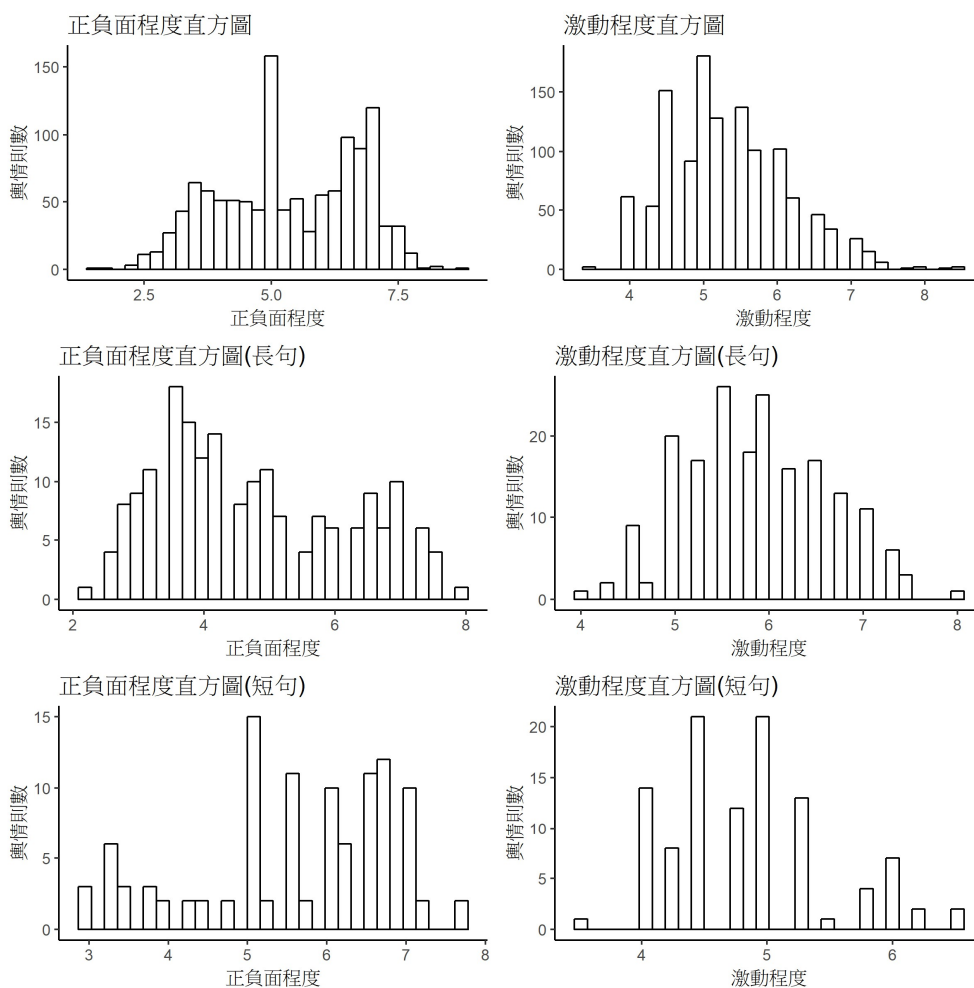
而短輿情偏向正面這樣的分布情形，即導致特徵不足，可能使學習的效果較差；反之長輿情多負向這樣的分布情形使特徵較多，則可能較適合學習。為解釋特徵數量的問題，我們取出短輿情最正面的 5 則及長輿情最負面的 5 則，可觀察出短輿情較易出現重複詞，長輿情之詞彙則較多樣。結合前述分析結果，當長輿情正面比例低但短輿情正面比例高的狀況同時發生時，可能是正面輿情不易被學習的原因。

表七、健保政策短輿情最正面 3 則

情感正負面程度	激動程度	字元數	輿情
7.75	6	5	很滿意健保
7.75	6.5	6	就很滿意啦
7.25	5.75	6	非常滿意健保

表八、健保政策長輿情最負面 3 則

情感正負面程度	激動程度	字元數	輿情內容
2.25	6.75	73	很怕健保會倒掉，有些人在 A 健保很過分，醫療系統不好，明明醫不好八九十歲中風還要氣切插管，浪費醫療資源，這樣很缺德，要規定八十歲以上不要有侵入性治療
2.5	6.75	72	健保藥物給付的條件越來越嚴苛，所以很多人會覺得一般診所的藥越來越差，寧願多花錢去醫院。鼓勵民眾先去診所治療，但又讓人感覺藥比較差，這樣怎麼會願意
2.5	6.75	60	醫療資源時常被濫用浪費，很多人明明不需要就診，還是跑來看醫生，還有很多中醫說要義診，叫老人家拿卡去刷，這都是浪費健保資源



圖五、健保政策長、短輿情情感正負面程度及激動程度分布

我們已知新詞在負面的學習效果優於正面，因此挑選 10 個較特別的負面詞如表九，可看出福利國、勞工、大陸人等特別的負面詞，在一般領域中則偏向中性，但在健保政策領域出現時可能較為負面，這些健保政策領域中特殊的負面詞可說明深度學習有學習到新詞的維度型情感，補充了中文維度型情感詞典在健保政策領域不足的部分。

相較傳統針對特殊詞個別標記之方法，從句子中自動提取新詞之方法已有效率的提升，惟本方法仍受資料數量及標記品質影響，當資料數量不足時或標記錯誤較多時，詞的學習效果就會不理想，因此所學習出之新詞並非完全正確，僅係根據基礎事實自動學習出來的結果，因此仍有正面之新詞夾雜負面詞或負面之新詞夾雜正面詞等錯誤結果。在自適應中文維度型情感詞典的方法下，前揭錯誤可透過增加資料數量或改善標記品質改善，若須確保新詞的正確性，則可透過人工篩選保留正確的新詞。

表九、經挑選後之負面新詞

新詞	情感正負面程度	激動程度
福利國	1.6754	6.9010
病房	1.6796	5.1799
倒掉	1.7034	5.0382
專利	1.7133	5.2003
費率	1.7259	5.0074
大陸人	1.7406	5.0936
勞工	1.7577	5.2355
掛號費	1.8044	3.1129
健保費	1.9870	5.0051
漲價	2.0198	4.9987

五、結論

本節實驗在正確性部分，我們蒐集了實際的健保輿情，並經過健保領域的研究人員進行情感標記後，比較中文維度型情感詞典與真實輿情標記結果，說明中文維度型情感詞典在健保政策領域效能較差的情形。

為了改進效能，我們透過機器學習及深度學習等演算法重新學習，並以連續型及離散型指標評估效能，在機器學習演算法中以支援向量機表現最佳，交叉驗證實驗之均方誤差為 0.97、皮爾森相關係數為 0.7，若不考慮黑箱問題，該模型可有效應用在情感預測。

而考量健保政策領域需要易理解的規則，因此我們提出透過深度學習隱藏層提取新詞的方法，即自適應中文維度型情感詞典，在經過效能評估後，驗證中文維度型情感詞典可透過增加新詞改善在健保政策領域的效能，其交叉驗證實驗之均方誤差為 0.95、皮爾森相關係數為 0.71，效能略優於支援向量機，且可直觀的觀察到每個詞的情感。

在資料分析的過程中，我們觀察到了正面輿情長度較短，而負面輿情長度較長的狀況，

符合「滿意的人少回覆，不滿意的人多抱怨」的常識，因此並非資料蒐集過度集中所導致。而正面輿情缺少特徵，可能造成學習效果略低於負面輿情，故在新詞提取中，負面詞有較佳的學習效果。

透過實驗我們初步驗證自適應方法，在正確性部分可以透過增加真實健保政策輿情資料量及標記人員數量等方式強化，在效能部分則可透過深度學習提取新詞進行自適應，而在解釋性部分則可透過人工篩選改進。綜上，本節研究之自適應中文維度型情感詞典，已可將中文維度型情感詞典適應至特定領域，強化情感分析的正確性，且效能在本實驗資料中優於機器學習或深度學習模型。

五、未來展望

本研究使用之深度學習模型仍有許多改進空間，例如：對於情感正負程度的貢獻未必每個詞都是同等重要，透過注意力機制可能會有更佳的效能；每個詞的情感嵌入到單一神經元僅取得詞的整體情感貢獻，但詞未必只有一種情感，在不同位置可能會代表不同情感，例如反諷時就常利用相同的詞表示負面情感，因此改為嵌入多個神經元或許可以取得詞在不同語境下的情感編碼，但更複雜的模型往往需要更大量且優質的標記資料。

因此考量標記任務為自適應中文維度型情感詞典重要的一環，當標記的數量或品質較差時，可能導致無法有效學習，因此為完善自適應中文維度型情感詞典的有效循環，應搭配有效的標記推薦方法。我們將在未來設計標記推薦系統，其透過隨機的刪減小樣本，監控深度學習模型的學習誤差，並結合刪減樣本數的懲罰係數設計目標函數，透過多次迭代使目標函數收斂，並將刪減的樣本即視為雜訊樣本，而雜訊樣本則假設可能為新詞數量較多或斷詞錯誤造成學習效果不佳，因此可透過修正結巴自定義詞典修正斷詞，再將雜訊樣本跟未被刪除的樣本進行詞彙差集，取得雜訊樣本中的差集詞彙，最後再透過搜尋引擎查詢差集詞彙取得較相似的新樣本，推薦領域專家進行標記，完善具方向性的標記及學習循環，降低標記成本，使自適應中文維度型情感詞典能更有效的持續發展。

六、致謝

本研究特別感謝陽明大學醫務管理所林寬佳教授、碩士生江婉琪及衛生福利部中央健康保險署相關人員，協助資料蒐集與標記並提供專業知識指導。本研究承蒙科技部 MOST 107-2628-E-155-002-MY3 經費補助特此致謝。

參考文獻

- [1] F. Z. Xing, F. Pallucchini, and E. Cambria, "Cognitive-inspired domain adaptation of sentiment lexicons," *Information Processing & Management*, vol. 56, no. 3, pp. 554-564, 2019.
- [2] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, 2018.
- [3] J. Wang, L. C. Yu, K. R. Lai and X. Zhang, "Community-based Weighted Graph Model for Valence-Arousal Prediction of Affective Words," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 1957-1968, 2016.
- [4] L. C. Yu, J. Wang, K. R. Lai and X. Zhang, "Pipelined Neural Networks for Phrase-level Sentiment Intensity Prediction," *IEEE Transactions on Affective Computing*, 2018.
- [5] K. Z. Aung and N. N. Myo, "Sentiment analysis of students' comment using lexicon based approach," *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, Wuhan, 2017, pp. 149-154.
- [6] X. Fu, J. Yang, J. Li, M. Fang and H. Wang, "Lexicon-Enhanced LSTM With Attention for General Sentiment Analysis," *IEEE Access*, vol. 6, pp. 71884-71891, 2018.
- [7] B. Shin, T. Lee, and J. D. Choi, "Lexicon Integrated CNN Models with Attention for Sentiment Analysis," *arXiv preprint arXiv:1610.06272*, 2016.
- [8] Y. Zou, T. Gui, Q. Zhang and X. Huang, "A Lexicon-Based Supervised Attention Model for Neural Sentiment Analysis," *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 868-877, 2018.
- [9] L. C. Yu, L. H. Lee, S. Hao, J. Wang, Y. He, J. Hu, K. R. Lai and X. Zhang, "Building Chinese affective resources in valence-arousal dimensions," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 540-545, 2016.
- [10] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng and X. Wu, "Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary," *IEEE Access*, vol. 7, pp. 43749-43762, 2019.
- [11] P. T. Noi and M. Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 imagery," *Sensors*, vol. 18, no. 1, pp. 18, 2018.
- [12] H. Dong, X. Hu, L. Wang and F. Pu, "Gaofen-3 PolSAR Image Classification via XGBoost and Polarimetric Spatial Information," *Sensors*, vol. 18, no. 2, pp. 611, 2018.
- [13] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85-117, 2015.
- [14] M. Lin, Q. Chen and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
- [15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.
- [16] S. Ioffe and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [17] J. Wang, L. C. Yu, K. R. Lai and X. Zhang, "Tree-Structured Regional CNN-LSTM Model

- for Dimensional Sentiment Analysis," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 28, no. 1, pp. 581-591, 2020.
- [18] J. L. Wu, Y. He, L. C. Yu and K. R. Lai, "Identifying Emotion Labels from Psychiatric Social Texts Using a Bi-directional LSTM-CNN Model," *IEEE Access*, vol. 8, pp. 66638-66646, 2020.
- [19] 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.
- [20] Y. Wang and S. Hu, "Exploiting high level feature for dynamic textures recognition," *Neurocomputing*, vol. 154, pp. 217-224, 2015.
- [21] C. Yang, Z. Liu, D. Zhao, M. Sun and E. Chang, "Network representation learning with rich text information," *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [22] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, 2013.
- [24] L. C. Yu, J. Wang, K. R. Lai and X. Zhang, "Refining Word Embeddings Using Intensity Scores for Sentiment Analysis," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 26, no. 3, pp. 671-681, 2018.
- [25] H. Keshavarz and M. S. Abadeh, "ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs," *Knowledge-Based Systems*, vol. 122, pp. 1-16, 2017.
- [26] B. Shi, Z. Fu, L. Bing, W. Lam, "Learning domain-sensitive and sentiment-aware word embeddings," *arXiv preprint arXiv:1805.03801*, 2018.
- [27] J. Barnes, R. Klinger and S. S. I. Walde, "Projecting embeddings for domain adaptation: Joint modeling of sentiment analysis in diverse domains," *arXiv preprint arXiv:1806.04381*, 2018.
- [28] L. C. Yu, J. Wang, K. R. Lai and X. Zhang, "Refining Word Embeddings Using Intensity Scores for Sentiment Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 671-681, 2018.
- [29] X. Fu, Y. Wei, F. Xu, T. Wang, Y. Lu, J. Li and J. Z. Huang, "Semi-supervised aspect-level sentiment classification model based on variational autoencoder," *Knowledge-Based Systems*, vol. 171, pp. 81-92, 2019.
- [30] S. Zhai and Z. Zhang, "Semisupervised autoencoder for sentiment analysis," *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [31] C. Wu, F. Wu, S. Wu, Z. Yuan, J. Liu, Y. Huang, "Semi-supervised dimensional sentiment analysis with variational autoencoder," *Knowledge-Based Systems*, vol. 165, pp. 30-39, 2019.

Nepali Speech Recognition Using CNN, GRU and CTC

Bharat Bhatta

brb4344@gmail.com

Basanta Joshi

basanta@ioe.edu.np

Ram Krishna Maharjhan

rkmahajn@ioe.edu.np

Department of Electronics and Computer Engineering
Pulchowk Campus, Institute of Engineering
Tribhuvan University
Nepal

Abstract

Communication is an important part of life. To use communication technology efficiently we need to know how to use them or how to instruct these devices to perform tasks. Automatic speech recognition plays an important role in interaction with the technology. Nepali speech recognition involves in conversion of Nepali speech to its correct Nepali transcriptions. The purposed model consists of CNN, GRU and CTC network. The feature in the raw audio is extracted by using MFCC algorithm. CNN is for learning high level features. GRU is responsible for constructing the acoustic model. CTC is responsible for decoding. The dataset consists of 18 female speakers. It is provided by Open Speech and Language Resources. The build model can predict the with the WER of 11%.

Keywords: Nepali Speech Recognition, Automatic Speech Recognition, Gated Recurrent Unit (GRU), Convolution Neural Network (CNN)

1 Introduction

Speaking and writing are the two important things that help us to communicate among us. Deficient of either writing or speaking affects our daily activities. Most of the people in rural area are able to speak properly but not able to write properly. Most of communication technology (gadgets, mobiles, computers etc) needs text as an input for their operation. To make familiar with the technology Automatic Speech Recognition (ASR) can play significant role. The Nepali ASR converts the spoken Nepali voice to its textual representation.

The ASR can be built by two different approaches. The first approach is traditional based that implement Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). The input feature vector is efficiently processed by the GMM and calculated the emission probabilities for each HMM states[1]. The second is to deep learning approach to build the acoustic model. The use of deep learning approach significantly increases the performance of the ASR system[2]. Traditional ASR system needs separate block of phonetic/linguistic constructs, acoustic model and the language model. The requirement of separate phonetic/linguistic constructs is eliminated by the deep learning approach[3].

Previously, a work is carried out to recognise the ten Nepali unique words[4] and another work is carried out taking the dataset of sports news. It consists of 1320 words among which 617 are unique[5]. The second model could not recognise the characters 'दु', 'वि', 'के', 'लि' that occurs very close together suggesting that the network learned them as single sounds although they are multiple characters. The CNN is used to learn the features that can easily distinguish those close words.

This paper presents an idea to build the Nepali ASR system that can convert spoken Nepali language to its textual representation. The model used MFCC as input feature vector. These MFCC features are used by CNN to generate more spatial features. CNNs are used because they are exceptionally good at capturing high level features in spatial domain[6]. GRU is used to develop an acoustic model. Training duration for GRU is less compared to LSTM network. CTC is used because it is alignment free[3]. The loss function used is CTC loss and decoding is carried out through the CTC network. The use of CTC eliminates the use of frame-wise labeling.

2 Review

Starting from single speaker based digit recognizer the modern Automatic Speech Recognition (ASR) reaches to speaker independent Hidden Markov Model based ASR[7]. With the evolution of the deep learning the accuracy of the ASR system further increases[8]. Deep Neural

Network (DNN) domination in ASR started, which showed that feed-forward DNN outperforms (Gaussian Mixture Model) GMM in the task of estimation of context-dependent HMM state emitting probabilities[9].

The development of ASR for speech recognition passes through series of steps. Development of ASR starts from digit recognizer for single user , passing through HMM, GMM based and reaches to deep learning[10, 9]. Some research work has been carried on Nepali speech recognition and Nepali speech synthesis. The initial work on Nepali ASR is carried out by using HMM based approach. This system is trained with 10 different words. Four female and four male speakers record the data. This models is able to predict limited words only[4]. Since limited word based ASR system is not able to generalised unseen words. Increasing the size of vocabulary and building own dataset a model is built. The model is not able to predict the some words 'दु', 'वि', 'के', 'लि' that occurs very close together suggesting that the network learned them as single sounds although they are multiple characters[5]. To eliminate the wrong prediction on close character a CNN layer is added. The accuracy of the Nepali ASR model can be further increased by using n-gram language model. The n-gram model, which defined the probability of occurrence of an ordered sequence of n words[11].

3 Method

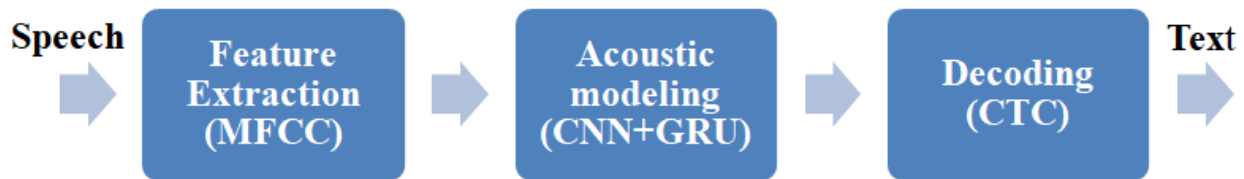


Figure 1: Architecture of proposed ASR system

The proposed ASR system involves in the conversion of speech input feature vector $X = [x_1, x_2, x_3, \dots, x_n]$ into its textual representation (label) $Y = [y_1, y_2, y_3, \dots, y_m]$ as shown in in Figure 1. The label is group of Nepali characters. The MFCC feature vector x_t of dimension D at time frame corresponds to label sequences $L = [l_1, l_2, l_3, \dots, l_N]$. If vocabulary of labels is V , $l_u \in V$ is the label at position u in L , then V^* represents the collection of all label sequences formed by labels in V . From these information the ASR need to find most likely label sequence \hat{L} for given X This can be represented by the equation 1:

$$\hat{L} = \underset{L \in V^*}{\operatorname{argmax}} p(L/X) \quad (1)$$

The the main objective of an ASR is to established a model that can accurately calculate the posterior probability $p(L/X)$.

3.1 Feature Extraction

Features are the elements that represents the phonemes in the speech. The features presents in the raw audio speech is is extracted by using the MFCC feature extraction algorithm. MFCC features are sequence of Acoustic feature vectors where each vector representing information in a small time window of signal[12]. The first step is pre-emphasis of the signal. The emphasis is carried out by choosing the value of $\alpha = 0.97$. Pre-emphasis boost the amount of energy present in high frequency signal. A signal is said to be stationary if its frequency or spectral contents are not changing with respect to time. Speech signal are non-stationary in nature because its spectral is constantly changing. So to simplify things it is assumed that on short time scales the audio signal doesn't change much, i.e. statistically stationary, obviously the samples are constantly changing on even short time scales[13]. Windows is taken to make them stationary signal[14]. The size of the window is 25 ms. 512 point FFT is taken from windowed signal. Thus obtained signal now model with human auditory system.26 filter-bank is implemented as the set of triangular shaped band-pass filters arranged in non-uniform frequency scale. This MFCC filter-bank is responsible to make them similar to human auditory system[15] and is converted to Log scale because human are less sensitive to higher energy change.

3.2 Feature Learning

CNN is used to capture high level spatial features from the image. The plot of MFCC can be view as a transformed intensity of frequencies over time which resembles to images[6], hence CNN can be used to capture high level feature in spatial domain. 1-dimensional CNN of filter size 200, kernel size (K) 11 and dilation 1 is used. Compared with conventional speech features, CNN can use local filtering and maximum pooling techniques to get more robust features[16].

3.3 Acoustic model generation

Gated Recurrent Unit (GRU) is used to learn the sequential data[17]. GRU manage the input weights to solve the vanishing gradient that present in RNN and consists of two gates: reset gate and update gate[18] which can be represented by the Egn. 2 and 3 respectively.

$$z_t = \sigma(W_z[h_t - 1, x_t]) \quad (2)$$

$$r_t = \sigma(W_r[h_t - 1, x_t]) \quad (3)$$

The reset gate is responsible for determining how much amount of past information to be forgotten. The update gate is responsible for determining how much amount of past information should be forwarded to for future[19]. The current memory and the final memory at current time step is given by the Egn. 4 and 5 respectively.

$$\hat{h} = \tanh(r_t.W_r[h_t - 1, x_t]) \quad (4)$$

$$h = (1 - z_t) * h_t - 1 + z_t * \hat{h} \quad (5)$$

3.4 CTC layer

The decoding is carried out is using CTC network. The CTC is based on the Bayes' on decision theory[2]. It receives output from softmax function. For each output layer, posterior probability (that represents 89 symbols) is computed. The character (symbol) having highest probability is given as the output. The decoding is carried out by the CTC network[20]. The output is generated in the form of numeric form. This output is mapped to the character. This is predicated output. CTC loss is computed by using true label and predicated output.

3.5 Evaluation

The Word Error Rate (WER) and Character Error Rate (CER) indicate the amount of text that the applied model did not read correctly. WER is the common evaluation metric for speech recognition system[21] which lies in the range between 0% and 100%.

4 Experiment

The experimental setup is carried out on the GPU MX150. For the pre-processing, feature extraction, training and testing, python and its library has been used.

4.1 Dataset

Speech recognition need the label data. It consist of audio corpus and its label file[8]. Every voice clip consists of phoneme transcription aligned with the sentence transcription label. The dataset consists of high quality female spoken corpus which is provided by Open Speech and Language Resources[22]. The dataset contains Eighteen unique female speaker records. The dataset is divided as 80% train and 20%. test set.

4.2 Result and Analysis

At first MFCC feature is taken from the raw audio. These features are passed to CNN to extract high level features. These features are used to generate the acoustic model. Several experiments has been conducted by varying several hyperparameter such as batchsize, learning rate, number of epochs. The size of CNN filter is 200, kernal size is 11 and in GRU units is 200.

The first experiment is carried out keeping 16412 as training utterance out of 26000 utterances. The learning rate is 0.03, momentum is 0.9, batch size is 100 and total number of epochs is 400. Total training duration is 1.5 days. The training is carried out in CPU Intel Core i7-8550U. The model get overfitted. It can predict well for train data but doesn't predict well for unseen data.

The second experiment is carried out keeping 16412 as training utterance out of 2064 utterances. The learning rate is 0.03, momentum is 0.9, batch size is 300 and total number of epochs is 100. Total training duration is 1.5 days. The other conditions remain unchanged.

The third experiment is carried out keeping 16412 as training utterance out of 2064 utterances. The learning rate is 0.015, momentum is 0.9, batch size is 50 and total number of epochs is 100. Total training duration is 1.5 hrs. The other conditions remain unchanged. The Batch size is changes to observe the effect. The result can be summarised by the Table 1. The parameters from this experiment is considered and some sample outputs are tabulated as shown in Table 2.

Table 1: Summary of experiments with the results

Experiment	learning rate	batch size	total epochs	WER
1	0.03	100	44	90
2	0.03	300	100	80
3	0.015	50	100	11

4.3 Model validation

The performance of the model is validate with the RNN-CTC model [5]. This research work is carried out to enhance the performance of the existing model. The Character Error Rate (CER) of that model is 52% and our CNN-GRU-CTC model gives the CER of 1.836%. Based on these results, it can be concluded that our model provides the better generalization capabilities.

Table 2: Sample output of the System

Sample	Ground Truth	Model prediction	WER
1	सानैदेखि सङ्गीतमा रुचि राख्ने अधिकारीले सुरुमा हार्मोनियममा शिव शङ्करबाट तालिम लिएका थिए	सानैदेखि सङ्गीतमा रुचि राख्ने अधिकारीले सुरुमा हार्मोनियममा शिव शङ्करबाट तालिम लिएका थिए	0.00
2	इन्डोनेसियाली पपुवा प्रान्तमा रहेको राष्ट्रिय निकुञ्ज	इन्डोनेसियाली पपुवा प्रान्तमा रहेको राष्ट्रि निकुञ्ज	0.167
3	चलचित्रमा केमियो रोलमा नायक राजबल्लभ कोइरालालाई पनि देख्न पाइनेछ	चलचित्रमा केमियो रोलमा नायक राजबल्लभ कोइराँलालाई पनि देख्न पाइनेछ	0.111
4	उनले दुई हजार दसमा जर्जियामा सुरु हुने स्टोर्स टुर्नामिन्टमा भाग लिन थाले	उनले दुई हजार दसमा जर्जियामा सुरुहुने स्टोर्स टुर्नामिन्टमा भागलिन थाले	0.25

5 Conclusion

On performing several experiments it seem the performance of model depends upon several factors such as learning rate, number of epochs, momentum, batch size, training duration etc. The system predict the unseen data with the WER of around 11% which is quite satisfactory. It can be conclude that CNN-RNN architecture can be used for speech to teach conversion. The quality of the model depends upon the quality of the data. So before the training phase the data must kept clean by preprocessing on data. The model depends upon several factor. From the experiment it seem the batch size and learning rate greatly effect on model development. Several parameters must be hypertuned to obtain the best model. Finally it is concluded that CNN-GRU model can be implemented to develop Nepelai ASR system.

References

- [1] Dong Liu, Antoine Honore, Saikat Chatterjee, and Lars K Rasmussen. Powering hidden markov model by neural network based generative models. *arXiv preprint arXiv:1910.05744*, 2019.
- [2] Vishal Passricha and Rajesh Kumar Aggarwal. Convolutional neural networks for raw speech recognition. In *From Natural to Artificial Intelligence-Algorithms and Applications*. IntechOpen, 2018.
- [3] Dong Wang, Xiaodong Wang, and Shaohe Lv. End-to-end mandarin speech recognition combining cnn and blstm. *Symmetry*, 11(5):644, 2019.

- [4] Manish K Ssarma, Avaas Gajurel, Anup Pokhrel, and Basanta Joshi. Hmm based isolated word nepali speech recognition. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 71–76. IEEE, 2017.
- [5] Paribesh Regmi, Arjun Dahal, and Basanta Joshi. Nepali speech recognition using rnn-ctc model. *International Journal of Computer Applications*, 178(31):1–6, Jul 2019.
- [6] William Song and Jim Cai. End-to-end deep neural network for automatic speech recognition. *Stanford CS224D Reports*, 2015.
- [7] Hani S Matloub, David L Larson, Joan C Kuhn, N John Yousif, and James R Sanger. Lateral arm free flap in oral cavity reconstruction: a functional evaluation. *Head & neck*, 11(3):205–211, 1989.
- [8] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.
- [9] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.
- [10] Sadaoki Furui. 50 years of progress in speech and speaker recognition research. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 1(2):64–74, 2005.
- [11] Frederick Jelinek. The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, 73(11):1616–1624, 1985.
- [12] R Gupta and G Sivakumar. Speech recognition for hindi language. *IIT BOMBAY*, 2006.
- [13] Adrien Meynard and Bruno Torr sani. Spectral analysis for nonstationary audio. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2371–2380, 2018.
- [14] Sunil Kumar Kopparapu and M Laxminarayana. Choice of mel filter bank in computing mfcc of a resampled speech. In *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, pages 121–124. IEEE, 2010.

- [15] Deividas Eringis and Gintautas Tamulevičius. Improving speech recognition rate through analysis parameters. *Electrical, Control and Communication Engineering*, 5(1):61–66, 2014.
- [16] Chongchong Yu, Yunbing Chen, Yueqiao Li, Meng Kang, Shixuan Xu, and Xueer Liu. Cross-language end-to-end speech recognition research based on transfer learning for the low-resource tujia language. *Symmetry*, 11(2):179, 2019.
- [17] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *arXiv preprint arXiv:1905.01997*, 2019.
- [18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [19] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):92–102, 2018.
- [20] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE, 2017.
- [21] Xiang Kong, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel. Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5810–5814. IEEE, 2017.
- [22] Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De Silva, and Supheakmungkol Sarin. A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 66–70, Gurugram, India, August 2018.

上下文語言模型化技術於常見問答檢索之研究

A Study on Contextualized Language Modeling for FAQ Retrieval

曾琬婷 Tseng, Wen-Ting

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

d0409857@gmail.com

許永昌 Hsu, Yung-Chang

易晨智能股份有限公司

mic@ez-ai.com.tw

陳柏林 Chen, Berlin

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

近年來，深度學習技術有突破性的發展，並在很多自然語言處理的相關應用領域上也有相當亮眼的效能表現，例如 FAQ (Frequently Asked Question) 檢索任務。FAQ 檢索無論在電子商務服務或是線上論壇等許多領域都有廣泛的應用；其目的在於依照使用者的查詢(問題)來提供相對應最適合的答案。至今，已有出數種 FAQ 檢索的策略被提出，像是透過比較使用者查詢和標準問句的相似度、使用者查詢與標準問句對應的答案之間相關性，或是將使用者查詢做分類。因此，也有許多新穎的基於上下文的深層類神經網路語言模型被用於以實現上述策略；例如，BERT(Bidirectional Encoder Representations from Transformers)，以及它的延伸像是 K-BERT 或是 Sentence-BERT 等。儘管 BERT 以及它的延伸在 FAQ 檢索任務上已獲得不錯的效果，但是對於需要一般領域知識的 FAQ 任務仍有改進空間。因此，本論文中探討如何透過使用知識圖譜等的額外資訊來強化 BERT 在 FAQ 檢索任務上之效能，並同時比較不同策略和方法的結合在 FAQ 檢索任務之表現。

Abstract

Recent years have witnessed significant progress in the development of deep learning techniques, which also has achieved state-of-the-art performance for a wide variety of natural language processing (NLP) applications like the frequently asked question (FAQ) retrieval task. FAQ retrieval, which manages to provide relevant information in response to frequent questions or concerns, has far-reaching applications such as e-commerce services and online forums, among many other applications. In the common setting of the FAQ retrieval task, a collection of question-answer (Q-A) pairs compiled in advance can be capitalized to retrieve an appropriate answer in response to a user's query that is likely to reoccur frequently. To date, there have many strategies proposed to approach FAQ retrieval, ranging from comparing the similarity between the query and a question, to scoring the relevance between the query and the associated answer of a question, and performing classification on user queries. As such, a variety of contextualized language models have been extended and developed to operationalize the aforementioned strategies, like BERT (Bidirectional Encoder Representations from Transformers), K-BERT and Sentence-BERT. Although BERT and its variants has demonstrated reasonably good results on various FAQ retrieval tasks, they still would fall short for some tasks that may resort to generic knowledge. In view of this, in this paper, we set out to explore the utility of injecting an extra knowledge base into BERT for FAQ retrieval, meanwhile comparing among synergistic effects of different strategies and methods.

關鍵詞：常見問答集檢索，知識圖譜，自然語言處理，資訊檢索，深度學習

Keywords: Frequently Asked Question, Knowledge Graph, Natural Language Processing, Information Retrieval, Deep Learning

一、緒論

隨著網際網路上文本資料或是多媒體資訊的蓬勃發展，FAQ(Frequently Asked Questions)檢索技術的發展已經成為各種應用領域(如電子商務服務、線上論壇等)中極為重要的需求[1],[2]。過去許多網站通常會針對經常被問到的問題與其對應答案經由人工方式整理成問答配對可直接提供給使用者進入網站時瀏覽。但是資料量隨著時間增加，使用者要直接找到需要的答案也越來越困難[3]。時至今日，已經有許多網站提供 FAQ 查詢的服務讓使用者能更快速找到自己的需求，像是 FAQ Finder[4]系統以及 Ask Jeeves[5]服務網頁。

FAQ檢索基本上可視為以自然語言查詢為主的資訊檢索(Information Retrieval)之應用。目前實作於FAQ檢索技術大致可以分為非監督式方法以及監督式方法。前者利用計算使用者查詢和標準問句的相似度方式找到最適合的答案，像是Okapi BM25[6]模型或向量空間模型[7] (Vector Space Model, VSM)。後者主要利用使用者查詢與標準問句的相關答案之間相關性做判斷，BERT (Bidirectional Encoder Representations from Transformers)[8]或K-BERT (Knowledge-enabled Language Representation Model)[9]。其中BERT是能夠理解上下文的語言模型，主要使用Transformers中的注意力機制(Attention Mechanism)[10]讓模型可以學習到上下文的關係。但是BERT較缺乏特定領域知識，因此K-BERT基於知識圖譜並注入至模型中，使得BERT能夠如專家般針對相關知識進行推理。

儘管BERT在FAQ檢索研究上已經有亮眼的成績，但其效能需要有一般或特定領域知識的FAQ檢索任務上仍然有改進的空間。因此，本論文嘗試比較BERT以及它的兩種延伸方法，即K-BERT和Sentence-BERT[11]於兩種不同特性的FAQ資料集上的表現。同時，我們也比較使用者查詢和標準問句的相似度、使用者查詢與標準問句的相關答案之間相關性以及將使用者查詢做分類三種FAQ檢索策略實作在這兩種不同特性的資料集上之效果。從實驗中發現在含有多個標準問句對應單個答案(類別少)的任務情境適合利用將使用者查詢做分類方式，而單個標準問句對應單個答案(類別多)的任務情境適合利用比較使用者查詢與標準問句的相關答案之間相關性方式。另外，在模型中加入額外知識圖譜的資訊有助於提升FAQ檢索的效能。關於詳細方法與實驗討論將於後續章節依序介紹。

二、相關研究

(一)、語言模型

語言模型在自然語言處理的研究中佔有極重要的地位。近年來，預訓練的深層類神經網路語言模型，像是ELMO (Embeddings from Language Models)[12]、OpenAI GPT (Generative Pre-Training)[13]、BERT等在各種自然語言處理任務上皆提升整體效能。其中以BERT模型效果最為亮眼，它利用Transformers[10]學習文本中單詞之間上下文關

係的注意力機制。基於此特性應用於現實使用文字的情境中理解到文本的上下文資訊。BERT 模型在多種 FAQ 檢索任務中皆被發現其效能贏過 BiLSTM 和 ELMO 等神經網路的方法，也驗證了 BERT 是一個優質的語言表示模型，能夠達到甚至超越傳統的類神經網路模型。

(二)、知識圖譜

知識圖譜(Knowledge Graph)[14]曾於 2012 年由 Google 所提出，其本質上是基於圖的數據結構主要由節點(實體)和邊(關係)組成三元組(Triplets)。過去知識圖譜經常被應用在金融或醫療專業領域中[9]。近年來，隨著人工智慧的蓬勃發展，知識圖譜又被廣泛得應用在問答系統和聊天機器人中，協助系統更深地理解自然語言並做推理來提升整體問答的效果。目前有許多高品質且大規模的開放式知識圖譜，以英文來說包含 WordNet[15]、Wikidata 和 Yago 等[16]。中文的知識圖譜則包含知網(HowNet)[17]、CnDbpedia¹[9]和 Zhishi.me²等。

其中 HowNet 為大陸學者於 1998 年所創建，主要為電腦所設計的大型雙語知識庫。提供了設計人工智慧軟體所需的外部知識。HowNet 總共收錄了 50,220 筆漢語詞語，所涵蓋的概念總量達 62,174 筆，目前仍在持續擴充中。詞彙是最小的語言使用單位，但卻不是最小的語意單位。HowNet 使得自然語言理解上更深入地了解詞彙背後豐富的語意。對於一個詞在不同的情境之下可能會有不同的概念。在 HowNet 中 W_C 為一個概念，而 G_C 表示概念所屬於的詞性，DEF 則表示其定義。表一列舉幾個 HowNet 中的例子，其中每個『|』代表一個義原，左邊為英文右邊為中文。

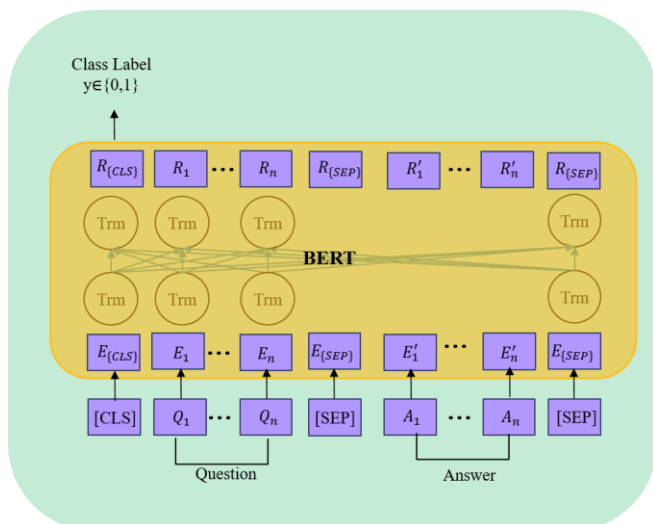
表一、HowNet 定義舉例

W C	G C	DEF
頂點	N	location 位置:belong={angular 角}, modifier={dot 點}
大學生	N	person 人, education 教育, highrank 高等
鮮花	N	FlowerGrass 花草

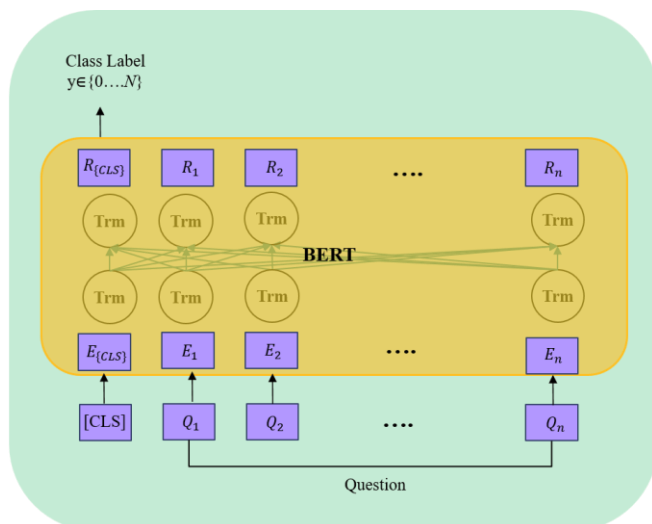
(三)、FAQ 相關研究

¹ <http://kw.fudan.edu.cn/cndbpedia/search/>

² <http://zhishi.me/>



圖一、BERT 架構預測問句與答案相關性



圖二、BERT 架構預測問句所屬答案

FAQ 檢索任務主要是利用使用者的查詢來找出一句相對應最適合之答案。而在過去的研究中可透過非監督式學習方式，例如 Okapi BM25 模型或向量空間模型，比較使用者的查詢以及 FAQ 樣本中的標準問句，若兩者的語意相近，則該 FAQ 樣本的答案就可能包含使用者需求的資訊。另外，FAQ 樣本的答案中也可能包含其他額外的資訊。因此，也可以透過監督式學習方式，例如 BERT 或 XLNet[18]，比較使用者查詢與標準問句的相關答案之間相關性。

除此之外，也可以藉由將非監督式學習的排名結果以及監督式學習排名的結果做線性組合並重新排序，以達到同時考慮上述兩種面向的效果。如 Wataru[1]等人使用 Okapi BM25 模型查找使用者查詢和標準問句相似度結合 BERT 模型查找標準問句和答案相關性結果來達到考慮兩個不同面向以提升整體效果，贏過 BiLSTM 和 CNN 等模型。

三、研究方法

本研究主要針對兩個層面做探討。第一個層面，針對不同特性的 FAQ 資料集所適用之策略作探討，分為使用者查詢與標準問句的相關答案之間相關性(Pair)、將使用者查詢做分類(Multiclass)以及比較使用者查詢和標準問句的相似度(Similarity)。第二個層面，我們再針對上述策略透過不同模型方法做實驗測試，使用的模型分為 BERT、K-BERT 以及 Sentence-BERT。

(一)、BERT 神經網路模型

BERT 為雙向 Transformers 的解碼器(Encoder)，主要的特色在於預訓練的方法上使用了 Masked LM (MLM)和 Next Sentence Prediction(NSP)。其中以 MLM 機制更為重要，能隨機性屏蔽掉部分輸入的 Tokens 並預測這些被遮蔽掉的 Tokens，目的在於讓模型依照上下文的資訊學習填補被遮蔽掉的 Tokens。

在實驗中我們使用兩種 Fine-Tuning 的方式來訓練 BERT，分別應用在使用者查詢與標準問句的相關答案之間相關性(Pair)如圖一和將使用者查詢做分類(Multiclass)如圖二，兩種 FAQ 檢索實作策略。前者的輸入分為兩個部分問題和答案中間以[SEP]特殊符號來分隔兩句，問題前面加上[CLS]用於分類，最後輸出{0,1}表示問題和答案是否相關。後者的輸入為單句，以問題作為輸入並在句首加上[CLS]用於分類，最後輸出{0,...,N}表示此問句所屬於的答案類別，N 表示為答案總共有 N 個。

(二)、K-BERT 神經網路模型

BERT 模型透過預訓練方式可以從大型語料庫中獲取通用的語言表示，但是缺少特定領域知識。所以在處理需要領域知識的任務上表現不佳，例如，醫學問答任務或是法律知識問答。在理解領域知識資料時，普通人只能根據其上下文理解單詞，而專家則可以利用相關領域知識進行推斷。為了達到此目的，引入知識圖譜至 BERT 模型中使得模型成為領域專家是一個較佳的方式。

在實驗中我們採用開放式一般領域知識圖譜，例如:HowNet 或是 CnDbpedia，作為知識三元組注入至語句之中。但是過多的知識注入可能會使得語句偏離其真正含義，因而產生知識噪聲 (Knowledge Noise) 問題。為了克服此問題，K-BERT 架構如圖三引入了軟位置(Soft-position)和可見矩陣(Visible Matrix)來限制不適當知識注入所產生的負面影響。在 K-BERT 中，首先透過知識層(Knowledge Layer)給定輸入句子 $s = \{w_0, w_1, \dots, w_n\}$ 和知識圖譜 k ，它會將知識圖譜注入到語句中並且轉換成語句樹 (Sentence tree) $Tr = \{w_0, w_1, \dots, w_i\{(r_{i0}, w_{i0}), \dots, (r_{ik}, w_{ik})\}, \dots, w_n\}$ ，語句樹允許每一個詞最多有兩個分支但其深度只能為一。接著在輸入 K-BERT 前會將資訊轉換成三層

的編碼層(Embedding layer)，包含 Token Embedding、Soft-position Embedding 以及 Segment Embedding。在 K-BERT 中，首先會將語句樹平鋪，平鋪以後的句子是雜亂不易閱讀的。K-BERT 通過 Soft-position 編碼方式來恢復語句樹的信息順序。其中在 Seeing Layer 利用 Masked-Transformer 的概念引入 Visible Matrix，讓詞的詞嵌入只源自於同一句語句樹的枝幹上，不同枝幹的詞之間相互不影響。以矩陣方式表示為：

$$M_{ij} = \begin{cases} 0 & , \text{visible} \\ -\infty & , \text{invisible} \end{cases} \quad (1)$$

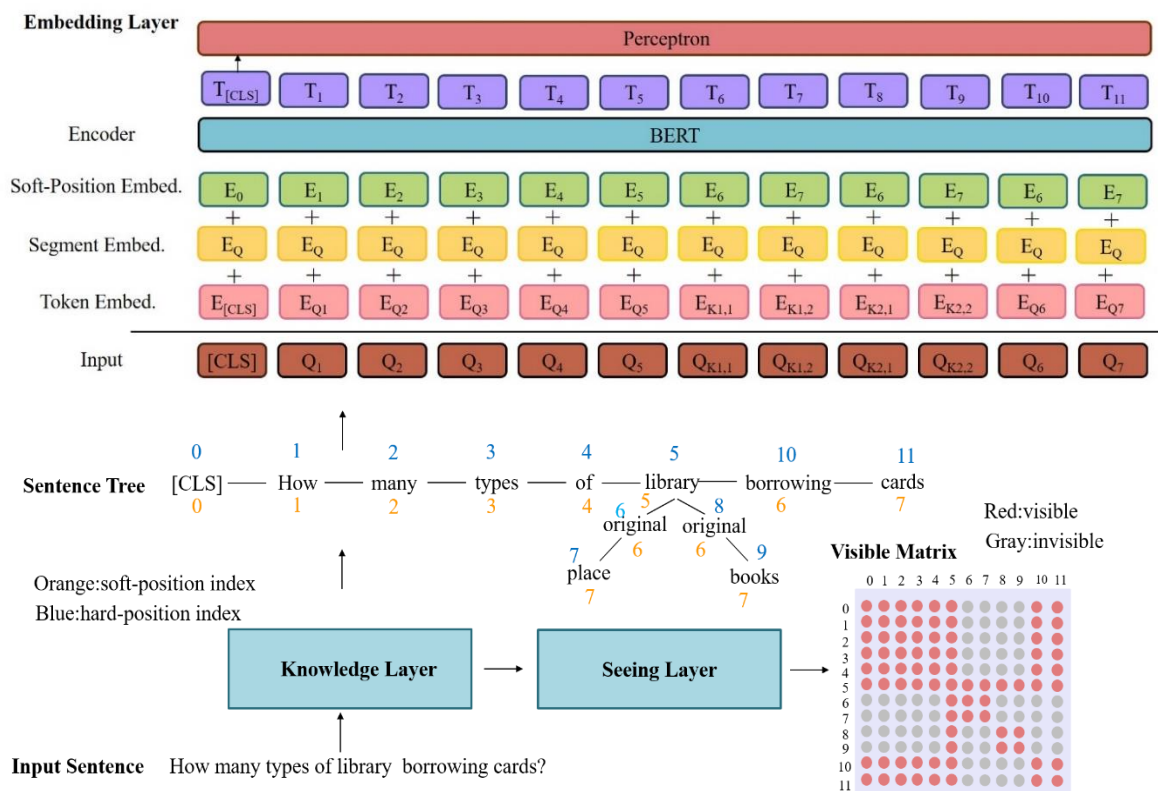
其中 0 表示詞彙之間於語句樹中同一條枝幹上是相互可見的， $-\infty$ 則表示詞彙之間於語句樹中不同條枝幹上是相互不可見的。將 Visible Matrix 加入至 Self-attention 中表示為：

$$Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q, h^i W_k, h^i W_v \quad (2)$$

$$S^{i+1} = \text{softmax}\left(\frac{Q^{i+1}K^{i+1T} + M}{\sqrt{d_k}}\right) \quad (3)$$

$$h^{i+1} = S^{i+1}V^{i+1} \quad (4)$$

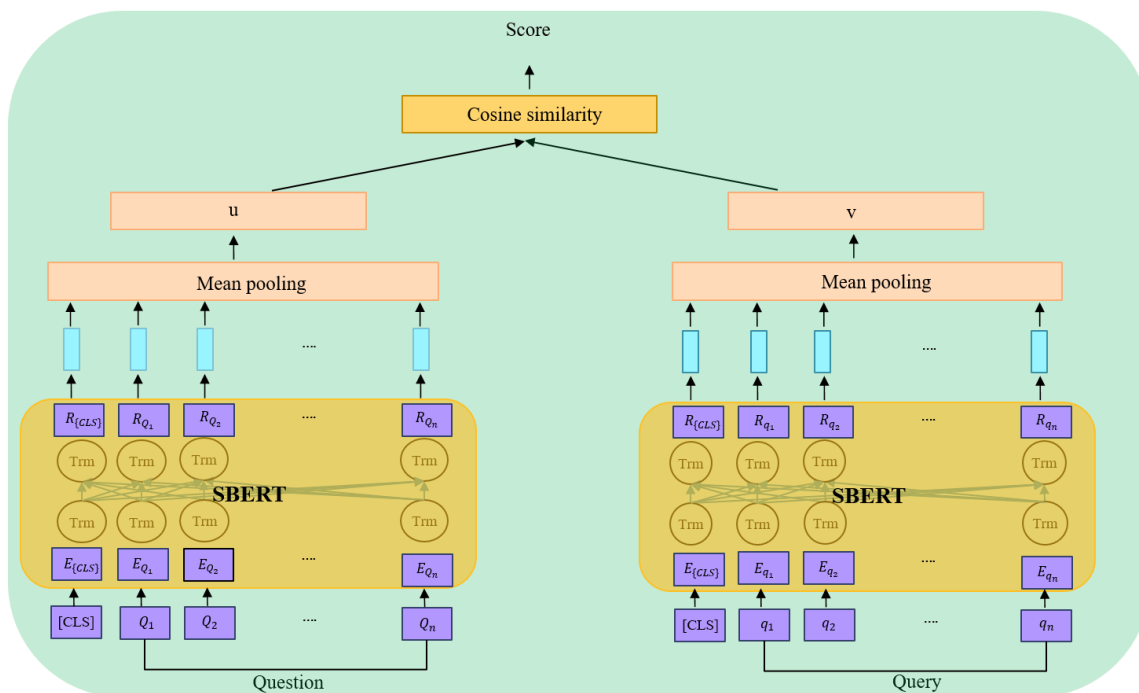
其中式(3)在原始 Self-attention 中多加上遮罩矩陣 M(Mask matrix)，讓語句樹中不同枝幹上的詞彙之間互不影響。



圖三、K-BERT 架構[9]

(三)、Sentence-BERT 神經網路模型

BERT 在語意相似度計算(Semantic Similarity)任務上已經有不錯的表現，但是它必須將兩句語句都輸入至模型中，因而導致大量的計算成本。Sentence-BERT (SBERT)網路結構可以解決 BERT 模型的不足，將兩句不同語句輸入到兩個 BERT 中，其中兩個 BERT 模型的參數是共享的，得到每一句語句的表徵向量。直接利用兩句話的表徵向量計算餘弦相似度，如圖四。



圖四、Sentence-BERT 架構[11]

四、實驗結果與討論

(一)、實驗材料

本研究材料採用兩個公開的中文數據集 TaipeiQA 以及 LawQA[9]。TaipeiQA 是從台北市政府官方網站爬蟲取得的 FAQ 問答數據集，多種問題可能對應至一種答案。LawQA 為法律相關的 FAQ 問答數據集，一句問題只會對應至一句答案。其中將上述兩個數據集整理成三種格式的資料如表二。並將資料切分成訓練集、驗證集和測試集，其中測試集為使用者查詢(Query)。第一種格式資料(Pair)包含多個問題與答案之配對，問題與答案若相關就標記為 1，相反地非相關就標記為 0。第二種格式資料(Multiclass)將每個答案當作一個類別，每個問題皆有相對應的答案。第三種格式資料(Similarity)包含多個問題與答案之配對以及使用者的查詢問題(Query)當作測試資料。

另外在實驗中我們有加入開放式一般領域知識圖譜的資料，分別使用到 HowNet 總

表二、三種 FAQ 檢索策略資料格式

	Question	Answer	Relevance
Pair	我在網路貸款上借了 1500 百塊不還會怎麼樣？	若無力償還會面臨法院後果的，建議與債權人積極協商，爭取延遲還款。債權人也會起訴你，然後申請執行你的財產的。	1
		我可以幫你	0

	Question	Class
Multiclass	我在網路貸款上借了 1500 百塊不還會怎麼樣？	20

	Question	Query
Similarity	我在網路貸款上借了 1500 百塊不還會怎麼樣？	我貸款借 1500 百元不還會怎樣

表三、資料集統計

資料集	策略	訓練	驗證	測試
TaipeiQA	Pair	147,998	172	2070
	Multiclass	5,821	1,665	1,035
	Similarity	7,485	-	1,035
LawQA	Pair	29,003	3,708	3,631
	Multiclass	14,656	2,657	14,467
	Similarity	14,656	-	1,750

共 52,576 筆以及 CN-Dbpedia 總共 7526,076 筆，並且限制掛載數量最多為 2 深度最多為 1。HowNet 是一個大型的語言知識庫表示每個詞彙背後更深層的語意。CN-DBpedia 是由復旦大學研發並維護的大規模結構化百科知識圖譜資料。百科領域延伸至法律、工商、金融、文娛、科技、軍事、教育和醫療等十多個領域，已經成為業界與學界開放中文知識圖譜的首選。

(二)、評估指標

在評估方法上我們採用 F1 值，主要是為了觀察各個模型分別在三種 FAQ 檢索策略上精確度的一種指標表示為：

$$F1 = \frac{2 * precision * recall}{Precision + recall} \quad (5)$$

F1 值就是精確率 (precision)和召回率(recall)的調和平均值。其中精確率計算所有正確被檢出的結果(TP)占實際上被檢索到的(TP+FP)比例表示為：

$$precision = \frac{TP}{TP + FP} \quad (6)$$

召回率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，表示為：

$$recall = \frac{TP}{TP + FN} \quad (7)$$

(三)、實驗結果

我們的實驗結果顯示於表四中。首先，從三種 FAQ 檢索策略中可以觀察到，TaipeiQA 數據集中多種問題可能對應至一種答案且答案變化性少，所以適合使用 Multiclass 分類的策略直接將每個答案當作一個類別，而使用者查詢透過模型預測相對應最適合的答案。另外，在 LawQA 數據集中一句問題只會對應至一句答案，所以直接使用 Pair 策略計算使用者查詢與標準問句的相關答案之間相關性上效果比較好。

表四、實驗結果

Strategies\Models\Datasets		TaipeiQA	LawQA
Pair	BERT	0.518	0.864[9]
	K-BERT(HowNet)	0.521	0.873[9]
	K-BERT(CnDbpedia)	0.507	0.875[9]
Multiclass	BERT	0.697	0.168
	K-BERT(HowNet)	0.706	0.178
	K-BERT(CnDbpedia)	0.704	0.139
Similarity	SBERT	0.261	0.827

接著，在模型方法的比較上可以觀察到 BERT 模型中加上 HowNet 知識圖譜後優於原始 BERT 模型。但是在加上 CnDbpedia 知識圖譜後可能會造成過多知識噪音使得準

確率下降。可見加入適當量且高品質的知識圖譜才會增進模型的判斷能力。此外，在 TaipeiQA 數據集中使用者查詢與標準問題問法的差異性較大所以在 Sentence-BERT(SBERT)相似度比對上效果較差。在 LawQA 數據集使用者查詢與標準問題問法較為相似，所以在 Sentence-BERT(SBERT)相似度比對上效果較佳。

最後，我們還嘗試在 LawQA 資料集中的語句樹(Sentence tree)掛載不同數量的知識圖譜三元組(0,1,2,3,4)並比較對實驗結果的影響。在表五中，我們可以觀察到掛載量增加可以提高準確率，但是掛載過多的知識圖譜三元組反而會造成噪音因而造成準確率下降。

表五、掛載不同量知識圖譜之結果

掛載量	準確率
0	0.864
1	0.868
2	0.873
3	0.872
4	0.871

五、結論

本研究使用三種基於 BERT 的模型方法，應用到三種自然語言 FAQ 檢索策略上。經實驗驗證 K-BERT 模型中加上知識圖譜效果優於單獨使用 BERT 模型。另外，在 FAQ 檢索任務上資料集屬於多種問題對應至一種答案的情況下適合使用 Multiclass 分類的策略。若資料屬於一句問題只會對應至一句答案，則適合使用 Pair 策略計算使用者查詢與標準問句和相關答案對之間相關性。使用者的查詢與標準問句較為相似，則適合使用 Similarity 策略計算使用者查詢與標準問句相似度並找出對應的答案。從中可以觀察到不同特性的資料集皆有各自合適的方法和策略。在未來，我們希望能夠考慮到資料集中不同面向的資訊；例如，問題與答案相對應的主題，並加入到模型之中與目前的方法做比較。

參考文獻

- [1] Wataru Sakata et al., “FAQ retrieval using query-question similarity and BERT-based query-answer relevance,” In *Proceedings of the International ACM SIGIR Conference on*

- Research and Development in Information Retrieval*, pages 1113–1116, 2019.
- [2] Mladen Karan and Jan Šnajder, “Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. *Expert Systems with Applications*,” 91: 418-433, 2018.
 - [3] Yu-Sheng Lai et al., “Intention Extraction and Semantic Matching for Internet FAQ Retrieval Using Spoken Language Query,” In *Proceedings of Research on Computational Linguistics Conference XIII*. 2000.
 - [4] Kristian Hammond et al., “FAQ finder: a case-based approach to knowledge navigation,” In *Proceedings the 11th Conference on Artificial Intelligence for Applications*, IEEE, 1995.
 - [5] Ask Jeeves, [Online]. Available: <http://www.ask.com>
 - [6] Stephen Robertson and Hugo Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” *Foundations and Trends in Information Retrieval*, 3(4): 333–389, 2009.
 - [7] Gerard Salton et al., “A vector space model for automatic indexing,” *Communications of the ACM*, 18(11), pages 613–620, 1975.
 - [8] Jacob Devlin et al., “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2019.
 - [9] Weijie Liu et al., “K-BERT: enabling language representation with knowledge graph,” In *Proceedings of the AAAI Conference on Artificial Intelligence AAAI*, pages 2901–2908, 2020.
 - [10] Ashish Vaswani et al., “Attention is all you need,” In *Proceedings of Conference on Neural Information Processing Systems*, pages 5998–6008, 2017.
 - [11] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese BERT-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
 - [12] Matthew Peters et al., “Deep contextualized word representations,” In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, 2018.
 - [13] Alec Radford et al., “Improving language understanding by generative pre-training,” 2018.
 - [14] Shaoxiong Ji et al., “A survey on knowledge graphs: Representation, acquisition and applications,” *arXiv preprint arXiv:2002.00388*, 2020.
 - [15] George A. Miller. “WordNet: a lexical database for English,” *Communications of the ACM*, 38(11): 39–41, 1995.
 - [16] Fabian M. Suchanek et al., “YAGO: a core of semantic knowledge,” In *Proceedings of the international conference on World Wide Web*, pages 697–706, 2007.
 - [17] Zhendong Dong et al., “HowNet and Its Computation of Meaning,” In *Proceedings of the International Conference on Computational Linguistics*, pages 53–56, 2010.
 - [18] Zhilin Yang et al., “XLNet: Generalized autoregressive pretraining for language understanding,” In *Proceedings of Conference on Neural Information Processing Systems*, pages 5753–5763, 2019.

French and Russian students' production of Mandarin tones

Dr Felicia Zhang

Chinese Culture University

Email: feliciazhang8381@gmail.com

Abstract

This paper discusses the tone acquisition of Chinese by students in the context of Indo-European languages. This paper conducted two experiments: (1) Using Zhang (2006)'s 'Somatically-Enhanced Approach'(SEA) to conduct small-scale teaching experiments to the effectiveness of SEA on error correction of intermediate French and Russian students. "Somatically-Enhanced Approach" is centered on the body, teaching through humming, clapping, rhythm and movement to increase learners' sensitivity to tone and rhythm through language rhythm. The data in this thesis comes from the output of a Chinese class oral test of six French and Russian exchange students in a private university in Taiwan. (2) In the second experiment, all the spoken language corpus of French and Russian students were provided to ten native speaking Chinese teachers for analysis. After a one-semester study of the "Somatically-Enhanced Approach" in this research, Russian students and French students demonstrated that they could correctly pronounce the correct tones when speaking Chinese, with enhanced fluency in natural speech. The results of this study will be presented through quantitative (statistical data) and visualization and Praat was used to analyze the collected classroom spoken data and explore the sources of the errors.

Keywords: tone analysis, French and Russian students, interlanguage, Somatically-Enhanced-Approach.

1. Introduction

Tone is an important feature for distinguishing meaning in Chinese, and it is also one of the evaluation criteria for the accuracy of Chinese pronunciation. Chao (1930) also pointed out that Chinese tones have a distinguishing function, and the accurate mastery of Chinese tones can eliminate ambiguity and communicate effectively. Tones are supra-segmental components attached to syllables which are realized in pitch. To depict tones graphically to facilitate understanding, in Chinese language teaching the "Tone letter system" designed by Chao (1930) is usually adopted.

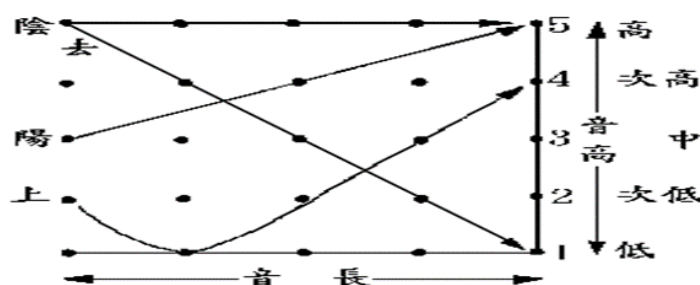


Figure 1: A system of tone letters

In Figure 1, the vertical axis represents a person's voice range which can be divided into five equal levels represented by 1, 2, 3, 4, and 5. Among them, 1 means the lowest, 2 means the second lowest, 3 means the 3rd highest, 4 means the second highest, and 5 means the highest level of pitch. 葉德明 (2005) pointed out that tones are realized by the pitch of speech which is determined by the frequency of the sound wave. Therefore, the higher the number for frequency, the higher the pitch for the sound, and vice versa. It is related to the number of vocal cord vibrations. The pitch ranges of a person also vary by gender. For instance, for males, it is roughly between 100 and 200 Hz, while for females, it is roughly between 150 and 300 Hz.

There are four tones in Mandarin Chinese, five if neutral tone is included.

Using Figure 1 as an illustration, Tone 1 is represented as 55, which means a word with a 55 tone tends to be high and flat like 'ma1: mother'. Tone 2 is represented as 35 which means that it starts at level 3 and increases in pitch until the voice reaches level 5 (ma2: hemp). Tone 2 is not of a vertical upward shape. Its shape goes from level 3 to level 5 across the entire horizontal axis. Tone 3 has a zigzag shape which goes from level 2 down to level 1 and then back up to level 4 like ma3: horse. Finally, Tone 4 is represented by 51 which means it starts at level five and then go straight down to 1 across the horizontal axis.

These tonal representations represent tones of citation form of characters consisted of single syllables. However, in real running speech, tones are affected by words surrounding that tone. This is most evident in the realization of tone 3. Even for native speakers, Tone 3 is often mixed up with Tone 2 and is the last tone to be acquired (Guo, 1991). For foreigners learning Chinese as a second language (TCSL), the mastery of Tone 3 is notoriously difficult with Tone 2 and Tone 3 equally unstable and error prone (葉德明, 2005).

There is also strong evidence that the beginnings and endings of sentences were a problem area for learners whose mother tongue was English and especially when Tone 4 occurs in the final positions of a sentence or a question (Zhang, 2006). This finding agrees with Wang's finding that large discrepancies exist between the tonal patterns of Tone 2, Tone 3 and Tone 4 before internal phrase boundaries and those at the end of the utterances exist (Wang, Jongman, & Sereno, 2001; Wang, Sereno, Jongman, & Hirsch, 2001). Such tonal problems also occur with students from other nationalities such as Russian and French learners of Mandarin (娜斯佳,

2014; 娜達莉婭, 2014; 林宥榛, 民 101; 隋偉靜, 2012). Regardless of which tone posed most difficulty for which students, the majority of the foreign students investigated in previous research demonstrated that especially students from non-tonal languages, i.e. those from alphabetic languages, their voice ranges for their native languages are consistently narrower than what is required to speak native level Mandarin Chinese. Zhang (2006) demonstrated that Mandarin native speakers' voice range is about 35% wider than those of the English speakers'. This suggests that the widening foreign learners' voice range is necessary in order for foreign learners to successfully acquire tones in Mandarin Chinese. This was also true with Russian, French, Italian and Polish students (周汪融, 2018; 娜斯佳, 2014; 娜達莉婭, 2014; 辛亞寧, 2007; 隋偉靜, 2012). Given this universality, Zhang (2006) used a movement oriented active approach known as the Somatically-enhanced Approach (SEA) in her study. The results of her study showed that using the body to enhance students' perception and production of Mandarin tones was largely successful except for Tone 3. The steps in SEA are intended as tools for implicit learning without help from pinyin, tone diacritics or English explanation. To solve the problems regarding Tone 3, this study employed an explicit learning paradigm in which the different realizations of Tone 3 are explained, with the aid of Pratt. to the student and then intensely practices immediately following the explanation.

2. Research methods

The research utilized a pre-posttest research design. The pre-test was held during the mid-semester test period and the post-test was conducted in the semester final period.

2.1 Setting and participants

Participants were 6 Russian and French exchange students who came to a private university in Taiwan to study Chinese. They usually only stay in Taiwan for about 6 months. Therefore, opportunities to collect oral data were limited which limits the effectiveness of this research. The participating students in the course had already acquired a certain level of Chinese with the Russian students having taken HSK level 3 exam in China. The teaching material was created by this researcher. The purpose of the course was to correct their language errors. Oral data analyzed came from mid-semester and final semester oral exams. Table 1 shows the basic information of the participating students.

Table 1: Basic information of the participating students

Student	age	gender	Length of studying Chinese	Certification
1	20	Female	2 years	HSK level 3
2	19	Female	2 years	HSK level 3
3	19	Female	3 years	HSK level 3
4	19	male	2 years	None
5	21	Male	2 years	None
6	20	Male	2years	None

2.2 Data collection instrument

Data was collected through mid-semester and end of semester oral tests in a 36-hour Mandarin course. The mid-semester test consisted of three parts. The first task was to read a passage, which only consisted of characters. The second task was for students to use natural speech to describe a house and the third task was a prepared monologue narrating ten things one must not do in French or Russian cultures. In the end of semester test, only two tasks were chosen. Task 1, the same passage used in Task 1 of mid semester test was used. With the content of the passage under control, it was possible to more accurately compare tonal errors between the two tasks. The second task was a pre-prepared monologue on taboos in French and Russian cultures. The rationale for the design of the controlled and natural tasks was to see whether errors made under controlled environment were transferred to free natural speech tasks

2.2.3 Auditory analysis

In order to explore in more detail, the nature of L2 learner errors in Mandarin, an auditory analysis was also conducted. This analysis was done by two native speakers of Mandarin who are also experienced teacher of Mandarin. Markers marked according to her judgment, the deviant tonal production of each syllable using Chao's tone letter system (Chao, 1930). They were allowed to listen to each utterance as often as she wanted. From the detailed auditory analysis, it was possible to observe the patterns of errors made by the student. An inter-rater coefficient is calculated to ensure the consistency between the markings of the two markers. ICC was used to calculate the interrater correlation coefficient for the two markers.

2.2.4 The perceptual experiment

In the perceptual experiment, speech productions produced by the student were given to 10 native speakers of Mandarin in Taiwan in random order. The researcher was not one of the ten judges. The native speaking judges were TCSL trainee teachers of Mandarin from Taiwan who were asked to listen to the speech production with no script and mark the "naturalness" of the speech from each recording according to a scale of 1 to 9. They were not specifically asked to mark tones or intonation or prosody as any judgment of these aspects of the language required specialized knowledge of the Mandarin phonological system. "Naturalness" is defined as how close the utterances are to native speaker speech in terms of rhythm, tones, intonation, stress and discourse features. The results obtained from the perceptual experiment represent subjective judgment of TCSL student's performances from native speakers of Mandarin. The judges did not know which oral performances were produced at which time.

2.2.5 Teaching procedure

The course consisted of two-hour per week for 18 weeks, a total of 36 hours of Chinese language instruction. In the first few weeks, the teacher found that the 6 students had some problems with tones, especially in Tone 3 of Chinese through oral recording assignments.

Before launching into phonetic correction, the teacher decided to find out whether the large number of errors with Tone 3 was caused by students not being able to hear the different realizations of Tone 3 or from the fact they have been taught to annotate the Tone 3 always with the diacritic [ˇ].

First, the teacher verbally explained the different realization of tone 3 based on Miracle’s research (Miracle, 1989) based on native speaker samples, i.e.:

- when in the word initial position, the NSs produced the low level contour (22);
- while in the word final position, the low falling contour (21) was more prevalent.
- A consistently falling contour was found preceding the neutral tone (21) and Tone 1 (21).
- In all other combination environments there was variation among the NSs between a falling and level contour.

Secondly, she illustrated these phenomena by using Pratt so that students can confirm with their eyes how these Tone 3 combinations are realized in reality. Then the teacher asked students to notate what they hear of the Tone 3 using [22] when tone 3 appears in the initial position; [21] when tone 3 in the final position; [21] before a neutral tone or tone 1. Of course, she also reminded them of the tone sandhi phenomena that when two or a series of third tones together, the preceding tone 3 change to a tone 2 with the last tone 3 preserving its 214 shape. Thus when two tone 3s are together such as in [ni3hao3], therefore its tones are represented as 35 for ni3 and 214 for hao3.

Thirdly, the teacher then read out 87 tone 3 combinations and asked them to note on the sheet what tone 3 combinations they hear and annotate them according to the new rules and notational system they established.

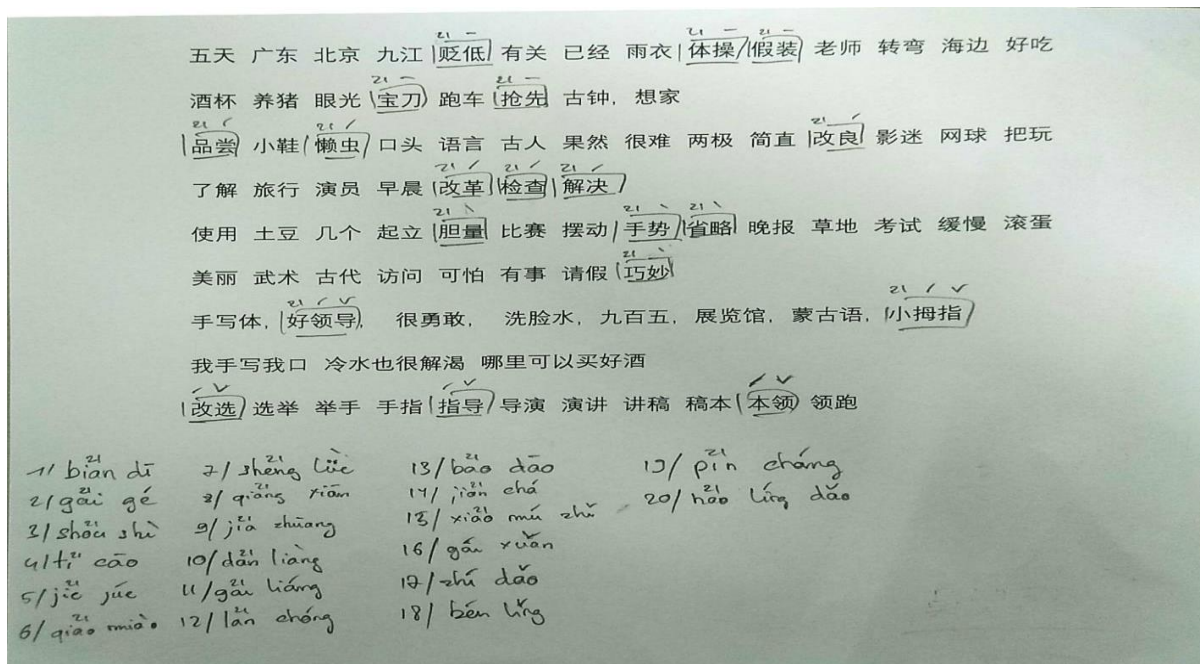


Figure 2: 87 tone 3 combinations used in the pre-testing stage. (曹文, 2002, p. 130)

This perception exercise established that most of the errors that these students made, with regards to Tone 3, came from the fact during their course of learning Chinese, the different realizations of Tone 3 were not explained to them adequately. This confusion with Tone 3 is further compounded by the convention of notating Tone 3, no matter what realization, with the tonal diacritic [ˇ].

Having found the source of Tone 3, the teacher then introduced them to the famous poem by Li qing zhao (李清照), sheng1sheng1 man4, xun2xun2mi4mi4 (聲聲慢，尋尋覓覓)(曹文, 2002, p. 132). Again, they were asked to listen and note the tones of the poem using the new system, then they used movements in the “Somatically-Enhanced Approach” to correct their tones. They were not allowed to read the text. In the ‘walking in circle’ procedure, they hummed, clapped and used movements to perceive tones. When the teacher finds that there is a problem with the tone, the teacher will use the movements to correct errors. For example, after ‘^{xiàbān}下班: get off work’, the Russian students are always unable to pronounce Tone 4 xia4 properly. So the teacher asked them to use the "stomping" method of physical teaching to correct it.

2.2.5.1 A brief sketch of the Somatically-enhanced Approach to Mandarin prosodic correction

This study was about how L2 learners learn to ‘structure’ L2 language input in the L2 learning process. Trubetzkoy defined the tasks of phonology as how the infinite variety of physical sounds fit in with the finite nature of language structures that govern a particular language (Trubetzkoy, 1939). As far as perception is concerned, structuring happens when a learner is called upon to select different acoustic stimuli through the senses in order to integrate the message. In other words, in terms of perception, the structuring activity implies re-ordering, usually in an unconscious way, by selective filtering out of redundant data that are usually perceived globally. For instance, in production, the learner must structure the non-linguistic experiences that he/she wants to talk about so that the available extra-linguistic and linguistic means can be applied to it.

Before we describe the structuring activity through SEA in detail, it is essential to understand what makes up a speech sound. Take the sound [i] in French. If it is recorded and listened to successively through octave filters: at different frequency ranges, different sounds can be heard. This is to illustrate how one’s ears are trained to hear sounds at a particular frequency which makes sense to the person by his/her mother tongue.

between 150 and 300 cps, we hear [u]
between 300 and 600 cps, we hear an intermediate sound
between [u] and [o]
between 400 and 800 cps, we hear [o]
between 600 and 1,200 cps, we hear [ɔ]
between 800 and 1,600 cps, we hear [ə]
between 1,200 and 2,400 cps, we hear [ɛ]
between 1,600 and 3,200 cps, we hear [e]
between 2,400 and 4,800 cps, we hear [i] lax
between 3,200 and 6,400 cps, we hear [i] tense — Cf. R. RENARD,

Figure 3: “L’appareil Suvaglingua, instrument de recherché et de correction phonétique”, R.P.A., 4, 1967, notes 13 and 14, pp. 62 and 63. Taken from (Renard, 1985)

According to the above figure, the French [i] is made up of all those sounds, each sound occupying a different frequency. Yet when French speakers were asked to identify these sounds, all the sounds were identified by French speakers as productions of the phoneme /i/. This manipulation of the French [i] sound showed very clearly the superabundance of the acoustic sound in reality. Through perception a distinction between what is necessary for [i] to be recognized ([i] is recognized between 3200 Hz to 6400 Hz, previously in cps: cycles per second) and French NSs would filter out the other superfluous frequencies.

However, when it comes to a L2 learner learning French, the learner would perceive a sound through hearing all the frequencies that are contained in a sound. When it comes to perceiving a sound in L2, he/she is likely *not* to recognize the sound /i/ at the frequency recognized by a French person but is likely to recognize the sound /i/ at a frequency dictated by his/her mother tongue such as between 300-600 Hz because his/her perception is likely to be mediated through his/her mother tongue. Thus, he/she is in danger of confusing /u/ (between 300-600 Hz) with /i/ (between 3200 and 6400 Hz), or with /o/ (between 400 and 800 Hz) as these sounds also occupies part of the spectrum that contains /i/ at different frequencies.

In order for L2 learners of French to perceive [i], the sounds at frequencies above 300 Hz could be eliminated and only leave frequencies between 3200 and 6400 Hz through a process of filtering so that L2 learners can be exposed to the French [i] at the correct frequency. This was a strategy similar to the strategy adopted by McCandliss et al in the perceptive training Japanese speakers to discriminate between the difference between [r] and [l] in English ((McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002) in which exaggeration was used to highlight the difference. A filtering process was also used in the Verbo-tonal method (VTM) of phonetic correction, developed at the Institute of Phonetics of the University of Zagreb by the late Professor Petar Guberina. Filtering in VTM (Renard, 1975) is a load lightening measure through which only the “relevant frequencies” for a particular sound, in this case, that of /i/ in French, is allowed to remain. In the teaching of Mandarin Chinese, the researcher in this study

did not have the equipment to supply students with a filtering mechanism. However, within the sensitization process of SEA, steps have been taken to prioritize the tonal aspects of Mandarin Chinese perceptually through the employment of humming, clapping and gestures. With the specific aim of defeating the ‘phonological sieve’ postulated by Trubetzkoy (1939), the following procedure was employed so as to develop a ‘feel’ for Mandarin prosody in TCSL students. The following descriptions has been simplified from Zhang (2006).

Step 1: Relaxation

The first step in the learning process is a relaxation procedure adapted from the success of Suggestopaedia in the 1980s. This relaxation step also is designed to reduce the language shock experienced by many learners especially when they are required to speak in the target language. In this first step, students are asked to lie on their backs on the floor and if possible, with the classroom darkened, then carry out mind-calming exercises for some five to ten minutes. This constitutes the relaxation phase of the classroom procedure.

Step 2: Humming

“Now, get up and stand in a circle.” The teacher joins the circle. The teacher says “I will hum to the sentence and please hum with me while walking slowly in a circle”. This is done for 5 times. (Step 2)

Step 2 involves humming to the intonation of the sentences without the vowels and consonants (5 times). This is used to highlight the intonation and tones of Mandarin. At this stage, students take an explicitly active part in the proceedings through humming along to the model. They are asked to repeat by “humming along” to the intonation. This is a way for them to produce an uncluttered sound string free from interfering vowel and consonant sounds.

The removal of vowel sounds is particularly important for learners of Mandarin from alphabetic language backgrounds because it forces them to prioritize the tones and prosodic aspects of Mandarin. As the input and output of the language uttered mutually reinforce each other, the language structure to be hummed should be a maximum of 5 to 7 syllables.

Step 3: Clapping to the rhythm of the sentences

“Now, I will clap to the rhythm of the sentence and then you can clap after me while walking in a circle.” (Step 3) This is done for 5 times again. The intonation of the sentence is again hummed in this fashion while the clapping is taking place.

The students, while listening to and “feeling” the intonation patterns, begin to move in harmony with the rhythm and intonation of the sentences modelled by the teacher. The teacher provides the beat and the rhythm of the sentences according to the stress and discourse features of the sentences. For instance, in teaching the sentence “*nǐ jiào shénme míngzi?* What is your name?” in spoken speech “*shénme:what*” always go together. If a learner only learns this through reading then it is highly likely that he/she will always introduce a pause in between “*shén*” and “*me*” and another pause between “*míng*” and “*zi*”. However, in prioritising the spoken over the written language in this course, the teacher demonstrated the beat of this sentence by providing

a beat for that group of words in the following manner:

[nǐ] [jiào] [shénme] [míngzi]?

1beat 1 beat 2 beats 2 beats

The clapping to the intonation patterns create a rhythm that students could follow while walking in a circle. This also allow students to observe and experience how stress, realized by length and loudness in Mandarin, is tied to meaning. This also allow them to observe the key words in a sentence and realize that not all words were of equal value and that in making oneself understood, some words are more important than others. This training was also essential in equipping them with the strategies of prediction and advanced planning in listening comprehension.

Step 4: Incorporation of movement and gesture

“Self-synchrony” refers to a process whereby the body of a speaker moves closely in time with speech (Condon & Ogston, 1966). For instance, spoken English is produced in groups of words, typically averaging about five in length, where there is only one primary vocal stress, conveyed principally through changes in pitch, also through changes in loudness or rhythm (Bull & Connelly, 1985). English, French and Russian are stress timed languages (Nespor, Shukla, & Mehler, 2011) To learn Mandarin Chinese successfully, they need to change their vocal and bodily behavior from speaking a life time of a stress-timed language to that of a syllable time language. Russian and French both are perceived as stress-timed languages and Mandarin a syllable time language.

Zhao (1987) described the four tones of Mandarin in terms of varying degree of tenseness of the vocal cords. For instance, in order to produce Tone 1, the vocal cords should be kept tense; to produce Tone 2, the vocal cords at first neither tense nor lax, then tense rapidly; to produce Tone 3, the vocal cords become lax immediately after being tensed, and then tense again; to produce Tone 4, the vocal cords suddenly tense, and then lax gradually.

Thus the corresponding gestures have been developed to produce the various tensions of the four tones in a sentence environment are as follows:

Tone 1: requires the vocal cords to be tensed and to be kept tensed. In order for students to experience the tensing of the body tension when pronouncing the Tone 1, students and the teacher need to tense up their hands, with the fingers spread out and the palms facing upwards. The elbows should be vertical and held close to the body. Students then push upward as though trying to touch the ceiling and keep to this posture, when pronouncing Tone 1.

Furthermore, as Tone 1 starts at a higher frequency with the tensing of the body greater than what most foreign speakers are used to, extra physical efforts need to be made to remind one that one must start high. To stretch one’s muscular system to express these Mandarin tones, one must not slouch in seats. By asking students to stand up straight and walk in a circle with various gestures would enable them to experience the coordination and synchronization of various muscles with the sounds uttered.

Tone 2: the vocal cords are at first neither tense nor lax, then become tense rapidly.

In order for students to experience the gradual tensing of the vocal cords, students are advised to adopt a forward slumping of the shoulders or a forward motion of the head initially, using very tense hands with the fingers spread out and the palms facing downwards, then tense up their arms and the whole of the upper body with the elbows held close to their body, then gradually push their hands up directly over their heads while pronouncing the Tone 2. This movement was used a great deal when practicing counting from one to ten in Mandarin because ten is pronounced in Tone 2 as “shí”. The movement thus reproduces the tonal contour of this tone. The key is to start the movement from the waist level and go up gradually until tenseness in the muscles are experienced.

Tone 3: the vocal cords become lax immediately after tense, and then tense up again. However, this description is only accurate on a lexical level. In running speech, Tone 3 is either realized as a lower level tone before a Tone 1, Tone 2, Tone 4 or a neutral tone or a Tone 2 before another Tone 3 according to the tone sandhi rules. Therefore, it is more accurate to describe Tone 3 before Tone 1, Tone 2, Tone 4 or a neutral tone as a lower level tone which requires the body to relax; and as a Tone 2 before another Tone 3 thus requiring the same movement as those described for Tone 2.

In instructing students to produce the low level Tone 3, students were advised to adopt a Relaxed posture with the shoulders accompanied by a forward motion of the head while producing the sound. In instructing students to produce the Tone 3 before another Tone 3, the movement adopted in producing Tone 2 was recommended with the first Tone 3 syllable.

Tone 4: the vocal cords suddenly tense, and then lax gradually. When it is necessary to go from tense muscle to lax muscle very quickly such as producing Tone 4, they were instructed to first raise their hands up high like what they were doing in Tone 1, then stamp their foot and if this movement fails for students to pronounce the Tone 4, throwing their arms downwards quickly while stamping their foot proved to be effective.

Step 5: Mouthing the words

In this step, the teacher instructs students by saying “Continuing with the movements, now mouth the sentences while I say them out loud.” (Step 5)

For the first time in the learning sequence, so far, students were hearing an intelligible sentence. They were asked not to say anything but merely to mouth the words.

Mouthing the words gives students the opportunity to practice the articulation of the sounds of the words without, in fact, placing them on an intonational background actually produced themselves.

Step 2-5 isolates each element of articulation e.g. filtered intonation, humming and mouthing before restoring them to a normal context has the further advantage of eliminating as many difficulties as possible in terms of comprehension of the sentence. Consequently, by the time students are actually asked to repeat a full sentence, they will have practiced each of its

constituent elements many times. They will look forward to achieving success in the next step of the process which will follow naturally and which should present little additional difficulty.

Step 6-7: Adding words to the intonation patterns

The teacher then says “Now repeat after me, and then add words to the intonation.” This again is done for five times. (Step 6)

The teacher then instructs each individual to repeat the sentence by themselves; checking that each student is reproducing the sentence correctly (Step 7).

The prosodic patterns are hummed again by the teacher for a further 5 times, and students are asked to say the sentences at the same time as they hear the prosodic patterns. This provides a transition between the kinds of exercises performed so far and the production of normal speech.

3.Results

Due to lack of space in this paper, the results of this research will provide most of the results in a simplified form but will comment in detail on the following two issues: (1) What are the causes of tonal errors in Russian and French students? (2) Are their errors a result of first language interference?

3.1 Result of perceptual experiment by ten native speakers of Mandarin Chinese:

Did the ten native speakers of Mandarin Chinese also perceive the students’ improvement as a result of conducting experiment two?

Table 2 Assessment of 6 students’ Mandarin Chinese production by 10 native speaking Chinese teachers

	No. of students	Average	Standard deviation
Pretest	6	6.1	1.02
Posttest	6	7.3	1.01

The T-test shows that there was a significant difference between the pre-test and the post-test, and there was a significant growth in the post-test of students’ production. The scoring criteria was to score on “naturalness” ranging from 1=totally unnatural and 9=natural like a native speaker. The mean value of the difference in the perceptual rating scores given by the teachers has a significant level ($p = 0.011$) where $p < 0.05$, which proves that the use of the "Somatically-enhanced approach can successfully improve foreign students’ Mandarin tone acquisition.

3.2 Causes of tonal errors:

To answer the question concerning the causes of student errors, the analysis of the production of the sentence

「zài měiguó měitiān zǎoshàng jiǔ diǎn 在美国每天早上九点上班，xiàwǔ wǔ diǎn xiàbān 下午五点下班」 was used. According to the mid-semester oral data (pre-test), all students made errors with the word 美國. All pronounced ‘mei3’ with 35 rather than the 21.

Table 3 Frequency of tonal errors in the sentence 「在美国每天早上九点上班，下午五点下班」(pretest)

	Errors: French students (n3) (time)	Errors: Russian students (n3) (time)
美 (21) 國 (35)	1 (s3)	1 (s4)
每 (21) 天 (55)	1 (s3)	0
早 (214) 上 (51)	2 (s1、s2)	2 (s4、s6)
九 (35) 點 (214)	1 (s3)	1 (s4)
上 (51) 班 (55)	1 (s3)	3 (s4、s5、s6)
下 (51) 午 (214)	0	2 (s4、s5)
五 (35) 點 (35)	2 (s1、s3)	0
下 (51) 班 (55)	0	0
total	8	9

key : s1, s2, s3=French students ; s4, s5, s6=Russian students

Examples of errors made are as follow:

(1) Comparison of [美國] between a male native speaker teacher and student 3.

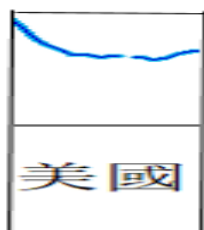


Figure 4; Male native speaker

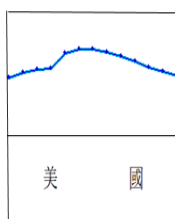


Figure 5: student 3 (pretest)

(2) Comparison of 「每天」 between a male native speaker teacher and student 3.

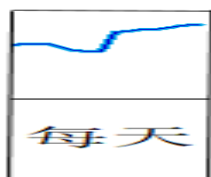


Figure 6 Male native speaker

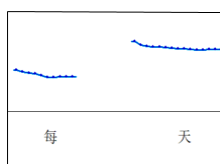


Figure 7: student 3 (pretest)

(3) Comparison of 「上班」 between a male native speaker teacher and Russian students

4 and 5.

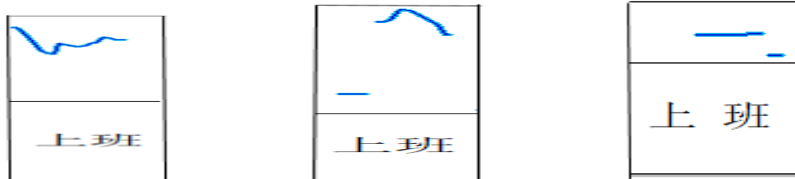


Figure 8: NS female teacher Figure 9: student 4 (pretest) Figure 10: student 4 (posttest)

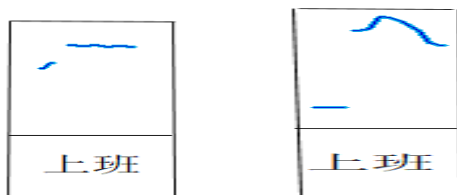
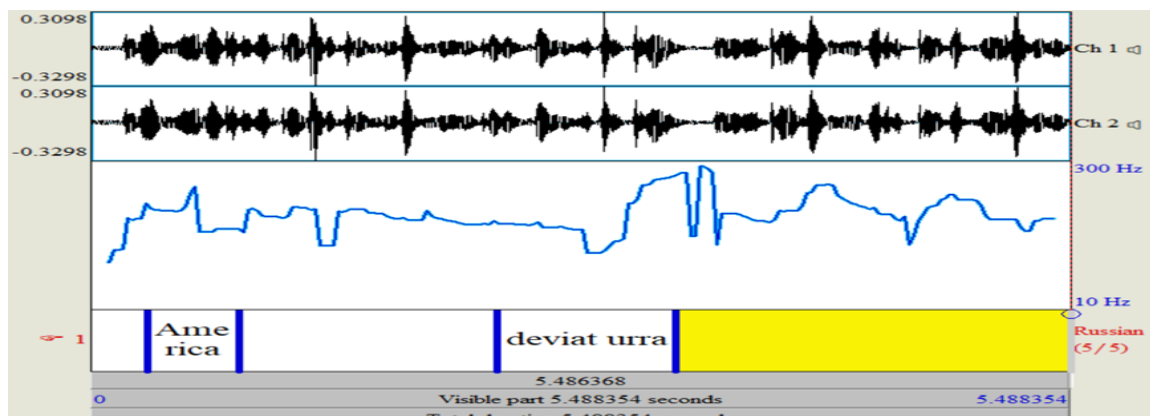


Figure 11: Student 5 (pretest) Figure 12: student 6 (posttest)

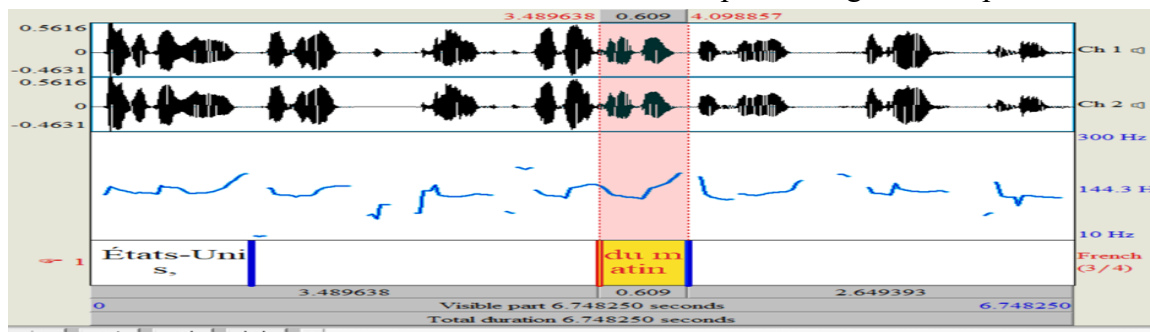
The above results clearly demonstrate that errors with 上班(shang4ban1) persisted. In order to investigate what might be the cause of this persistent error, the researcher asked a native speaker Russian to say 「在美国每天早上九点上班，下午五点下班」 in Russian.



Russian : V Amerike rabochi den nachinaetsa v deviat urra , a zakanchivaetsa v piat vecher a .

Figure 13 : 俄國人用俄文說「在美國每天早上九點上班，下午五點下班。」

The Praat figure 13 clearly shows that for the word ‘America’, the pitch goes up and similarly for the words ‘deviat urra’ the pitch goes up as well.



French : Aux États-Unis, le boulot commence à 9 heures du matin et finit à 5 heures du soir.

Figure 14 : 法國人用法文說「在美國每天早上九點上班，下午五點下班。」

Similarly, when the same sentence is spoken in French, figure 14 “Aux États-Unis” means

America, the intonation also goes up. For “du matin” it also goes up. In French ‘9 ‘clock’ goes before ‘in the morning’.

5. Conclusions

Tone 3 errors produced by the French and Russians may be due to poor or none existent explanation of Tone 3 realizations in real running speech. However, the Russians’ fourth tone errors is most likely affected by their mother tongue’s intonation. Like the Russian, when French people want to change sentences, their intonation usually goes up, which causes tonal problems in Chinese. Zhang (2006) pointed out that when students are saying a series of things, the end of each phrase, regardless of the tone present, always rises, while the trend of the last syllable at the end of the same phrase of native speakers is downward. In this study, the fourth-tone errors of Russian students confirmed the conclusion of Zhang (2006). To further confirm this conclusion, it will be advisable to conduct a comprehensive experiment involving a larger number of students be recording them both in Chinese and Russian.

This study also proves that the same problem occurs in the language of Russian-speaking Chinese students. When saying a series of things or to pick up the next sentence, they would raise the tone of the ending sound, just like the phrase "下班". The French students in this study are also affected in the same way. Therefore, it may be said that if a student’s mother tongue is an alphabetic language, most of the tonal errors that occur when they speak Chinese may be related to the intonation of their mother tongue. Since intonation and stress are more important than tones in their mother tongues, these features of their mother tongues are a likely to persist even in advanced speakers of Chinese’s Mandarin speech. The job facing teachers and phoneticians is still how to suppress the ‘phonological sieve’ (Trubetzkoy, 1939) of the first language in the second language process of Mandarin. In terms of research method, this study has demonstrated that the use of running speech to analyze students’ oral production is the way to allow first language interference errors at phrase junctures to appear.

6. Suggestions for teaching

The data contained in this paper suggests that Tone 3 is teachable and information about the realization of Tone 3 in sentences has always been available from Miracle’s research (Miracle, 1989). To avoid confusing students during the instructional process, the researcher suggests that the following strategies to be included in teaching practice:

- (1) Given that it is likely that many topics related to one’s family and friends will be the content of a zero beginner’s course, the different realizations of Tone 3 need to be explicitly shown to students and discussed with them through Praat. At the initial stage, avoid reliance on pinyin and tone diacritics in the printed textbook or the use of a language laboratory.
- (2) In the teaching process, it might be advisable to introduce Chao’s system of tones after the common 20 questions about oneself have been introduced. Once the tone letter system is introduced, use ‘21’ to represent the lower third tone; ‘24/14 as the change to a second tone

when the next character is another third tone. Then conduct perception exercises with third tone words or phrases to sensitize students' perception of third tone realizations. Then always ask students to look for rules about what third tone combinations cause changes in the realization of third tones. Always correct using gestures in student production.

These measures serve to expand the support system to include body, movement, gesture and other forms of technical support for students. More importantly, through the use of these support systems, learners will learn new strategies for learning Chinese on their own.

References

Chinese references:

- 周汪融. (2018). 對外漢語中漢語聲調的實驗語音研究. *文教資料*(10).
- 娜斯佳. (2014). 俄羅斯留學生上聲變調習得研究. (碩士), 遼寧師範大學.
- 娜達莉婭. (2014). 母語為俄語的留學生初級華語聲調偏誤及教學對策. (碩士), 北京外國語大學, 北京.
- 林宥榛. (民 101). 法文版零起點華語發音 紙本教材之設計原則初探. (碩士), 臺灣國立師範大學, 臺北, 臺灣.
- 葉德明. (2005). *華語語音學(上篇)*. 台北, 台灣: 台北:師大書苑有限公司.
- 辛亞寧. (2007). 意大利學生習得漢語聲調的實驗研究.
- 隋偉靜. (2012). 法國初級漢語水平學生聲調習得實驗分析.

English references:

- Bull, P. E., & Connelly, G. (1985). Body movement and emphasis in speech. *Journal of Nonverbal Behaviour*, 9, 169-187.
- Condon, W. S., & Ogston, W. D. (1966). Sound Film Analysis of Normal and pathological Behavior Patterns. *Journal of Nervous and Mental Diseases*, 143(4), 338-347.
- McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioural Neuroscience*, 2(2), 89-108.
- Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-timed vs. syllable-timed languages. *The Blackwell companion to phonology*, 1-13.
- Renard, R. (1985). Structuro-Global and Autonomy. *R. P. A.*, 73-74-75, 233.
- Trubetzkoy, N. S. (1939). *Principles of Phonology (Grundzuge de Phonologie, Travaux du cercle linguistique de Prague)* (C. Baltaxe, Trans. 1969 ed.): University of California Press.
- Zhao, J.-M. (1987). *ji1chu3han4yu3yu3yin1jiao4xue2deruo4gan1wen4ti2*. Paper presented at the Paper presented at the 1st International Conference on Teaching Chinese, Beijing.
- 曹文. (2002). *漢語語言教程*. 北京: 北京語言大學出版社.

情感分析於投資溫度評分之應用

Sentiment Analysis for Investment Atmosphere Scoring

彭志翔 Chih-Hsiang Peng

元智大學資訊管理學系

Department of Information Management College of Informatics

Yuan Ze University

metgnr89@apb.gov.tw

禹良治 Liang-Chih Yu

元智大學資訊管理學系

Department of Information Management College of Informatics

Yuan Ze University

lcyu@saturn.yzu.edu.tw

摘要

網路財經文本和金融商品投資收益的相關性在近年的研究議題上十分熱門。在此類型研究中，以使用自然語言處理領域中的情感分析技術量化投資大眾心理最受研究者青睞。在情感分析中，有類別型及維度型的分析方法，而維度型的分析方法又可分為單維度型與多維度型。本研究使用類別型、單維度型及多維度型三種情感分析方法，將蒐集的網路財經文本量化為投資溫度評分。實驗結果顯示，以多維度型的投資溫度評分與市場漲跌趨勢最具相關性。

Abstract

The correlation between Internet financial texts and financial commodity investment income has been an emerging research topic in recent years. Sentiment analysis is a useful technique to accomplish the investment psychology. In sentiment analysis, there are two main approaches to emotion state representation: categorical and dimensional. The dimensional approach can be further divided into single dimension and multiple dimensions. This study investigates the use of categorical, single-dimensional and multi-dimensional approaches to quantifying the Internet financial texts as investment atmosphere scoring. Experimental results show that the multi-dimensional method achieves the best correlation between the

investment atmosphere scoring and stock price trends.

關鍵詞：自然語言處理，情感分析，相關係數

Keywords: Natural language processing、Sentiment analysis、Correlation coefficient

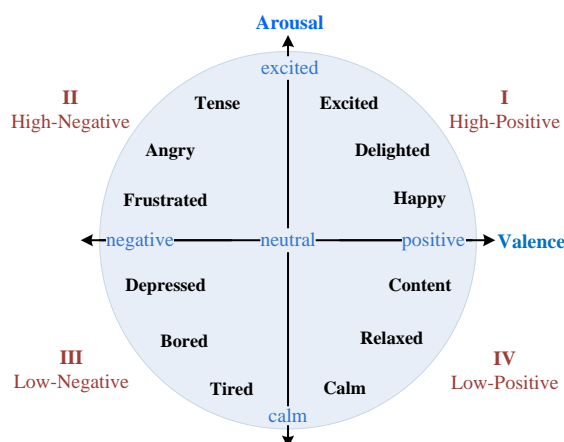
一、緒論

現今民眾從事工商經濟活動後，常會將剩餘可支配所得用來進行投資理財。目前常見的投資理財工具包括了股票、基金及其他衍生性金融商品，在這個高度專業化的社會環境中，民眾無法花太多時間研究、分析產業或區域的發展，也難以掌握精深的金融專業知識。據研究指出，我國投資人透過「網路」獲取投資相關資訊的比例最高，占 80.5%，其次是「報紙雜誌」，占 46.3%[1]。

在投資實務界中，有德國證券界教父之稱的安德烈·科斯托蘭尼（André Kostolany）強調「行情=資金+心理」[2]，認為單靠資金並不能影響股市，還要看投資大眾的心態，也就是「心理因素」，此一論點目前仍被許多投資人奉為圭臬。而現在最能影響投資大眾心態的資訊來源，就是網路。網路時代的投資市場資訊流動快速，但一般人通常沒有能力及時間，即時且完整地掌握網路上大量投資訊息，因此使用新聞報導預測股價走勢一直引起許多研究人員的興趣，此類研究大多係透過文字探勘技術(Text Mining)以及自然語言處理(Natural Language Processing) 中的情感分析(sentiment analysis) [3], [4], [5] 技術達成。

情感分析可分為類別型及維度型 2 種，類別型是將情感區分為正向(positive)、負向(negative)2 類，或是加上中立(neutral)分為 3 類。而維度型情感分析，可分為單維度與多維度 2 種不同的分析方法，單維度指的是採取一個情感程度值來分析，以連續型類值來表示情感的正負向。而多維度是用多個情感特徵來分析，Russell[6]認為情感間存在相關，相關性能夠透過一個 Valence-Arousal(VA)二維空間模型來表示，如圖一。Valence 代表正面或負面的程度，Arousal 則代表激動及冷靜的程度。情感經分析後，可在該二維空間模型上表示為一點。多維度的情感分析運用在財經文本分析上面，例如「美股上漲」與「美股狂漲」都是正向形容美股，但前者的強度並不及後者，影響程度應會有差距。相較於單維度情感分析，多維度情感分析應該更能精確的表達出文本中高度正面或高度負面的情緒。

本研究的動機就是希望比較上述三種情感分析方法將網路財經文本量化的結果，目的是運用情感分析技術將網路財經文本情感量化成投資溫度評分，並以實驗方式驗證投資溫度評分與選定市場趨勢間的相關性，因此本研究將針對以下 3 項進行實驗與探討：



圖一、Valence-Arousal 二維情感模型示意圖[7]

- 1、在網路財經文本的來源上，本研究係使用網路財經新聞，時間區間為 2017 年 1 月至 2020 年 5 月。
- 2、在與投資溫度評分進行相關性檢定的市場選擇上，本研究選擇台股、美股、大中華、亞太市場、新興市場、歐洲等 6 個市場區域。不同於以往多數研究進行個股與網路財經文本情感分析的實驗，本研究認為個股或其新聞較容易受到少數關鍵人士的操控或影響，而國家及區域指數較難被少數人操控，故選擇較大區域的市場趨勢做為檢定目標。
- 3、在投資溫度評分計算上，本研究將計算方式分為類別型、單維度及多維度 3 種計算方式，以實驗驗證哪種計算方式將財經文本情感分析結果量化為投資溫度評分後，與目標市場趨勢最具相關性。

二、相關研究

如依金融學中的效率市場理論，財經新聞出現之際，股價早已反應完畢，因此該理論認為，看新聞做股票係沒有辦法獲利和預測股價，但有許多研究證實投資人的心理因素在大大程度上影響了投資決策，因此投資大眾心理應該是影響金融市場波動不可或缺的一部分。網路財經文本和金融商品投資收益的相關性在近年的研究議題上十分熱門，近年的研究中，在財經文本的來源上，有網路新聞如蔡宇祥[8]的研究係使用微博財經貼文、

Qu et al.的研究係使用 Yahoo Financial News[9]，網路論壇如盧奕叡的研究係使用批踢踢實業坊的文本[10]，社群媒體貼文、搜尋首頁的關鍵字如 Zhao et al.的研究[11]係使用百度搜尋熱門關鍵字。在探討相關性或預測的對象的選擇上，大部分文獻是以個股為主如 Hwang et al.[12]的研究係使用特斯拉股價，部分文獻係採單一市場指數，如，Bourezk et al.[13]的研究係使用摩洛哥指數。文本的情感分析(Sentiment Analysis)，又稱意見挖掘或意見探勘(Opinion Mining)，是運用自然語言處理(Natural Language Processing)、文字探勘(Text Mining)及計算機語言學等技術，進而提取並辨識原文本中作者的情感及主觀意見。常見的情感分析所適用的範圍可分為句子層次、段落層次和文章層次的情感分析方法。類別型的情感分析方法相關研究，Hwang et al.[12]，通過將文本中的正向詞和負向詞計數，以「正詞數」-「負詞數」來計算情感得分，如果情感分數大於 0，則文本標記為「正」，如果小於 0，則文本標記為「負」。維度型情感分析在財經領域之應用並不多見，但已有維度型詞典，如：Affective Norms for English Words (ANEW)[14]、Extended ANEW[15]及語料庫，如：Affective Norms for English Texts (ANET) [16]、EmoBank[17]，相關情感詞向量[18], [19]及迴歸模型[20], [21], [22], [23], [24], [25]也有相關研究發表，均可做為財經領域維度型情感分析之研究材料。

三、研究方法

(一) 中文情感字典

本研究使用中文情感字典 CVAW (The Chinese Valence-Arousal Words) 進行情感分析[16]，CVAW 是一個包含 5,512 個情感多維度型字典，每個詞彙皆包含一組二維的 Valence-Arousal 實數值，維度 Valence 的範圍為 1 至 9，值 1 和 9 分別表示最負面和最正面的情緒表現，值 5 表示沒有特定傾向的中性情緒表現。維度 Arousal 的範圍為 1 至 9，值 1 和 9 分別表示平靜或激動。使用 CVAW 進行預測的實驗結果顯示，與使用英文文本情感分析獲得的結果相當[26]。

(二) 投資溫度

本研究將「投資溫度」定義為網路財經文本的情感，可作為投資大眾心態的參考。

1、財經文本資料蒐集

本研究先蒐集財經新聞做為情感分析之資料，資料來自蘋果日報、蘋果即時、工商時報、

工商即時、中國時報、中時即時、自由即時、中央社、經濟日報、聯合即時，再透過市場區域關鍵字，如：台股、美股、那斯達克、道瓊、標普、陸股、港股、歐股等，將財經文本歸類至特定市場。

依市場區域分類的財經文本，再利用情感字典計算每日的 Valence 及 Arousal 數值，計算結果如圖二所示，其中欄位「V」代表每日所有財經文本 Valence 的平均值，欄位「A」代表每日所有財經文本 Arousal 的平均值，欄位「Pos_num」代表每日所有正向財經文本數量(Valence 值大於或等於 5.5 之文章分類為正向)，欄位「Neg_num」代表每日所有負向財經文本數量(Valence 值小於或等於 4.5 之文章分類為負向)，中性的文章在本研究中不列入計算，欄位「Pos_V」代表每日所有正向財經文本 Valence 的平均值，欄位「Neg_V」代表每日所有負向財經文本 Valence 的平均值，欄位「Pos_A」代表每日所有正向財經文本 Arousal 的平均值，欄位「Neg_A」代表每日所有負向財經文本 Arousal 的平均值，欄位「Pos_VA」是每日所有正向財經文本 Valence 值乘以 Arousal 值之平均值，欄位「Neg_VA」是每日所有負向財經文本 Valence 值乘以 Arousal 值之平均值。

Date	V	A	Pos_num	Neg_num	Pos_V	Pos_A	Neg_V	Neg_A	Pos_VA	Neg_VA
20170101	5.320	5.820	3	2	6.600	5.833	3.400	5.800	27.200	-18.640
20170102	6.200	4.300	2	0	6.200	4.300	0.000	0.000	10.440	0.000
20170103	5.626	4.485	20	4	6.145	4.600	3.600	4.775	106.475	-27.020
20170104	5.721	4.629	11	2	6.255	5.055	3.575	3.550	71.610	-10.905
20170105	5.806	4.727	17	2	6.232	5.059	3.600	4.100	110.023	-11.880
20170106	5.863	4.315	20	0	6.002	4.583	0.000	0.000	92.458	0.000
20170107	None	None	0	0	0.000	0.000	0.000	0.000	0.000	0.000

圖二、財經文本每日 Valence 及 Arousal 數值示例

使用 CVAW 的情感分析結果，以「台股狂瀉近 700 點」為例，Valence 值為 1.5，Arousal 值為 8.5，呈現激動負面的情緒。而「台股重新衝牛市」Valence 值為 9，Arousal 值為 8.3，呈現激動正面的情緒。

2、投資溫度評分計算

本研究將投資溫度評分計算分為類別型、單維度型及多維度型三種。分數計算的時間皆以「週」為單位，較不受短期間過度反應的情緒影響，亦可減少網路財經文本日資料缺失值對實驗結果的影響。

(1)類別型：本研究定義之類別型投資溫度，只將情感區分為正向及負向，不考慮正向及負向的程度問題。計算方式係將一週的正向文章數總和除以一週所有文章數總和，計算公式如下：

$$\sum_{i=1}^n \frac{Pos_num_i}{Pos_num_i + Neg_num_i} \quad (1)$$

(2)單維度型：本研究定義之單維度型投資溫度，是以 Valence 值做為情感區分的依據，Valence 值係介於 1 至 9 之間的連續型數值，數值越接近 1 代表情感越負向，數值越接近 9 代表情感越正向。計算方式係一週內所有正向文章 Valence 值的總和加上一週內所有負向文章 Valence 值的總和，除以一週內所有文章數，再將結果標準化為 0 至 1 的數值，計算公式如下：

$$\frac{\left(\sum_{i=1}^n \frac{Pos_num_i \times Pos_V_i + Neg_num_i \times Neg_V_i}{Pos_num_i + Neg_num_i} \right) - 1}{8} \quad (2)$$

(3)多維度型：本研究定義之多維度型投資溫度，除了考量代表情感正負向的 Valence 值外，還將代表情感激動程度 Arousal 值加入，Arousal 值係介於 1 至 9 之間的連續型數值，數值越接近 1 代表情感越平靜，數值越接近 9 代表情感越激動。計算方式係一週內每日所有正向文章的 Valence 值乘以 Arousal 值相加後，加上一週內每日所有負向文章的 Valence 值乘以 Arousal 值，再將結果標準化為 0 至 1 的數值。計算公式如下：

$$\chi = \sum_{i=1}^n Pos_VA_i + Neg_VA_i \quad (3)$$

$$\chi_{norm} = \frac{\chi - \chi_{min}}{\chi_{max} - \chi_{min}} \in [0,1] \quad (4)$$

3、目標市場參考指數

本研究選定 6 個市場區域各 2 種參考指數做為實驗資料，分別為臺灣(台股加權指數、MSCI Taiwan Index)、美國(S&P500、Nasdaq)、大中華區域(FTSE Greater Chinese Index、MSCI Gold Dragon Index)、亞洲太平洋區域(MSCI Asia Pacific Index、FTSE Asia Pacific Index)、新興市場(MSCI Emerging Markets Index、FTSE Emerging Index)及歐洲(MSCI Europe Index、FTSE Europe Index)，參考指數主要以 MSCI 及 FTSE 所發行的指數為主。參考指數的歷史資料係由 Yahoo Finance 及 Investing.com 取得。

四、實驗結果

(一)實驗流程

在投資溫度評分計算方式部分，本研究使用 3 種方式，分別為類別型、單維度型、多維度型。在市場區域部分本研究選擇 6 個區域，在各區域中分別選定 2 種參考指數，總計 12 種參考指數，在參考指數漲跌趨勢計算部分為本週收盤價減上週收盤價。

1、實驗資料

本研究使用之網路財經文本及參考指數，蒐集時間為 2017 年 1 月 1 日至 2020 年 5 月 31 日。每篇文本經計算完 Valence 值及 Arousal 值後，先將 Valence 值大於或等於 5.5 之文章分類為正向，小於或等於 4.5 之文章分類為負向，介於 4.5 至 5.5 之文章分類為中性文章，本研究將中性文章排除後，共計收集 13 萬 2,564 篇網路財經文本。

2、評估標準

本研究採用皮爾森相關係數(r)做為相關性評估標準，其值在 0.1 以下屬於微弱或無相關，在 0.1 至 0.39 間屬於低度相關，在 0.4 至 0.69 間屬中度相關，0.7 至 0.99 間屬高度相關。

(二)實驗結果與分析

1、實驗結果

投資溫度評分與市場區域參考指數漲跌趨勢的相關性檢定結果如表一所示。

(1)維度型投資溫度評分計算方式結果比類別型為好。

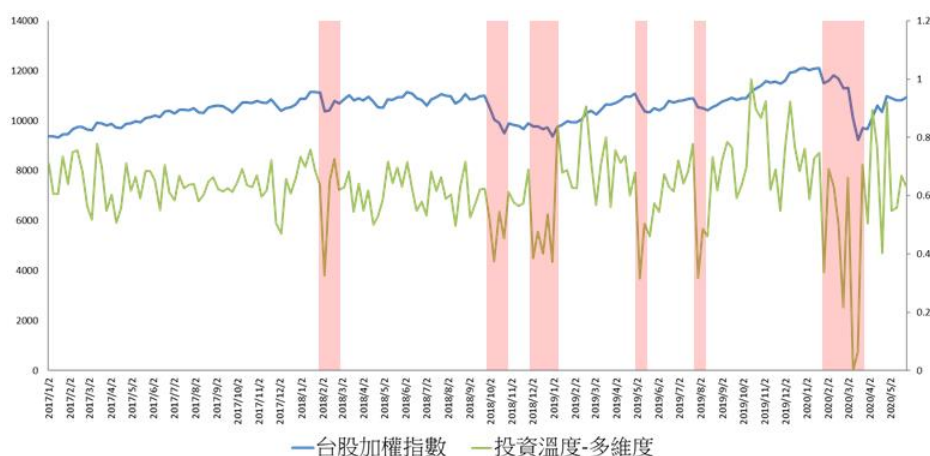
(2)在維度型的投資溫度評分中，多維度型投資溫度評分計算方式絕大多數情形下比單維度型投資溫度評分好。在 12 個指數實驗結果中有 11 個是多維度型計算方式較好，只有在大中華區域的 MSCI Gold Dragon Index 相關性檢定實驗中係以單維度型投資溫度評分較好，但與多維度型投資溫度評分相關係數僅相差 0.002。

表一、各投資溫度評分與參考指數漲跌相關性檢定結果表

市場區域	參考指數	類別型	單維度	多維度
臺灣	台股加權指數	0.687	0.730	<u>0.839</u>
	MSCI Taiwan Index	0.705	0.741	<u>0.830</u>
美國	S&P500	0.564	0.615	<u>0.810</u>
	Nasdaq	0.574	0.615	<u>0.751</u>
大中華區域	FTSE Greater Chinese Index	0.725	0.752	<u>0.754</u>
	MSCI Gold Dragon Index	0.725	<u>0.751</u>	0.749
亞洲太平洋區域	MSCI Asia Pacific Index	0.688	0.721	<u>0.800</u>
	FTSE Asia Pacific Index	0.682	0.716	<u>0.798</u>
新興市場	MSCI Emerging Markets Index	0.659	0.684	<u>0.717</u>
	FTSE Emerging Index	0.649	0.673	<u>0.713</u>
歐洲	FTSE Europe Index	0.443	0.501	<u>0.717</u>
	MSCI Europe Index	0.391	0.459	<u>0.666</u>

2、臺灣區域實驗結果分析

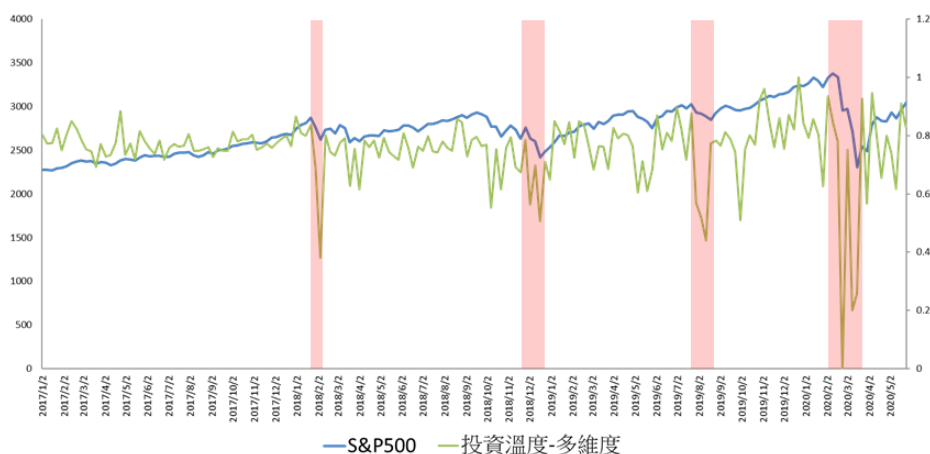
臺灣區域最佳相關係數為 0.839 為高度相關。相關性以折線圖方式表示如圖六，在下列時間點有很高的相關性：2018 年 2 月，股市漲幅過高、通膨預期升溫、公債殖利率上漲，引發程式交易避險，2 月 6 日台股創史上第 6 跌幅。2018 年 10 月，美國升息、預告 2019 年起對中國加徵關稅，10 月 11 日台股創史上最大跌點。2018 年 12 月，華為財務長遭捕，引發貿易戰，12 月跌幅 2.16%，為 2001 年來最差。2019 年 5 月，美中貿易戰升溫，5 月 14 日台股跌破年線。2019 年 8 月，川普下令美企撤出中國，8 月台股重挫逾 380 點。2020 年 2 月，新型冠狀肺炎造成全球經濟停滯。



圖三、台股加權指數與多維度型投資溫度相關圖

3、美國區域實驗結果分析

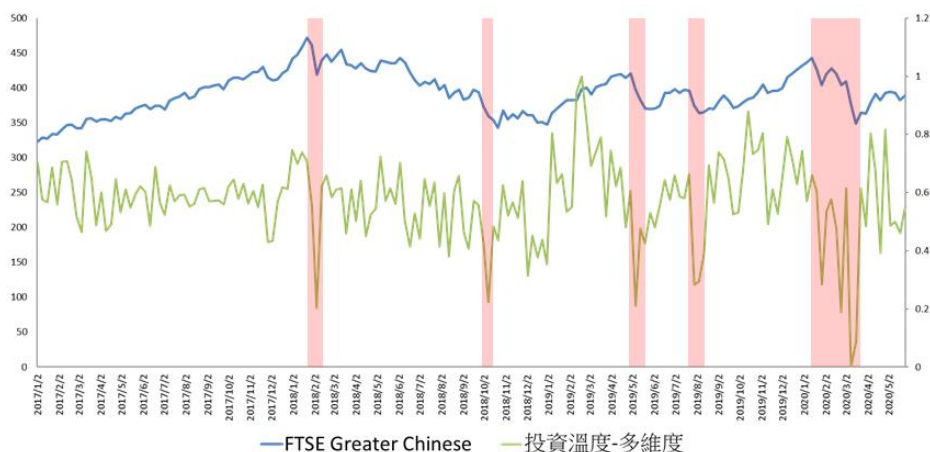
美國區域最佳相關係數為 0.810 為高度相關，相關性以折線圖方式表示如圖七，在下列時間點有很高的相關性：2018 年 2 月，2 月 4 日 5 日美股大跌 1,800 多點。2018 年 12 月，華為財務長孟晚舟遭捕，引發貿易戰戰火，12 月創 10 年來最糟單週表現。2019 年 8 月，川普下令美企撤出中國，美股在 8 月三度重挫至少 2.5% 以上。2020 年 2 月，新型冠狀肺炎造成全球經濟停滯。



圖四、S&P500 與多維度型投資溫度相關圖

4、大中華區域實驗結果分析

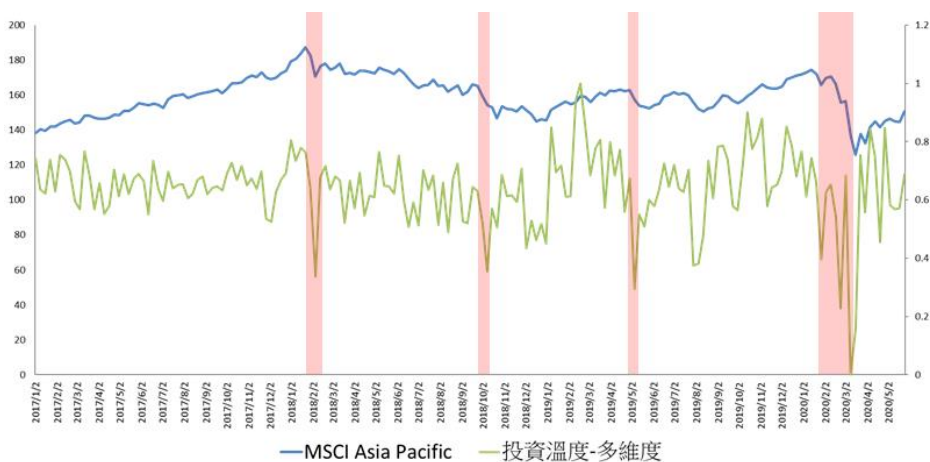
大中華區域最佳相關係數為 0.757 為高度相關。相關性以折線圖方式表示如圖八，在下列時間點有很高的相關性：2018 年 2 月，股市漲幅過高、通膨預期升溫、引發美股大跌，2 月 8 日港股開盤跌 800 點、上證 50 跌破半年線。2018 年 10 月，美國預告 2019 年起開始對中國加徵關稅，10 月上海 A 股跌幅 9.6%，港股跌幅 7.97%，台灣店頭市場指數跌幅 12.26%。2019 年 5 月，美中貿易戰對峙氛圍升溫，上證綜指 5 月創 2011 年以來最差的 5 月表現，港股 5 月下跌 2584 點或 8.7%。2019 年 8 月，川普下令美企撤出中國，港股 8 月中旬前下跌 2000 點，跌幅 7.6%。2020 年 1 月，新型冠狀肺炎造成中國大陸全面封城。



圖五、FTSE Greater China Index 與多維度型投資溫度相關圖

5、亞洲太平洋區域實驗結果分析

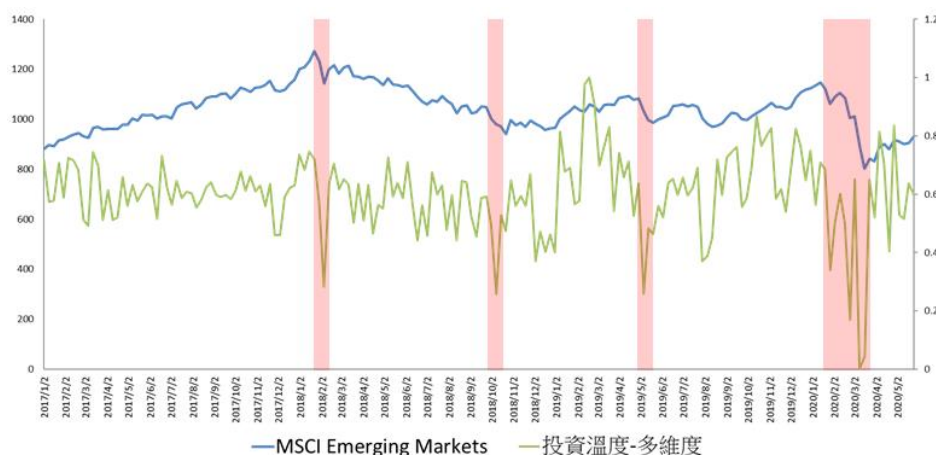
亞洲太平洋區域最佳相關係數為 0.800 為高度相關。相關性以折線圖方式表示如圖九，在下列時間點有很高的相關性：2018 年 2 月，股市漲幅過高、通膨預期升溫、引發美股大跌，2 月 6 日亞股趴成一片，上證指數收跌 4.05%，香港恆生指數跌逾 3%，日股跌逾 2%，韓股跌 1.8%，台股收跌 156 點。2018 年 10 月，美國預告 2019 年起開始對中國加徵關稅，10 月 11 日亞股哀鴻遍野，日經 225 指數收跌 3.89%；南韓股市收跌 4.44%，為 2011 年 11 月以來最大跌幅 2019 年 5 月，美國宣布對 2,000 億美元中國商品加徵懲罰性關稅，亞股 5 月 6 日重挫。2020 年 2 月，新型冠狀肺炎造成全球性經濟停滯。



圖六、MSCI Asia Pacific Index 與多維度型投資溫度相關圖

6、新興市場區域實驗結果分析

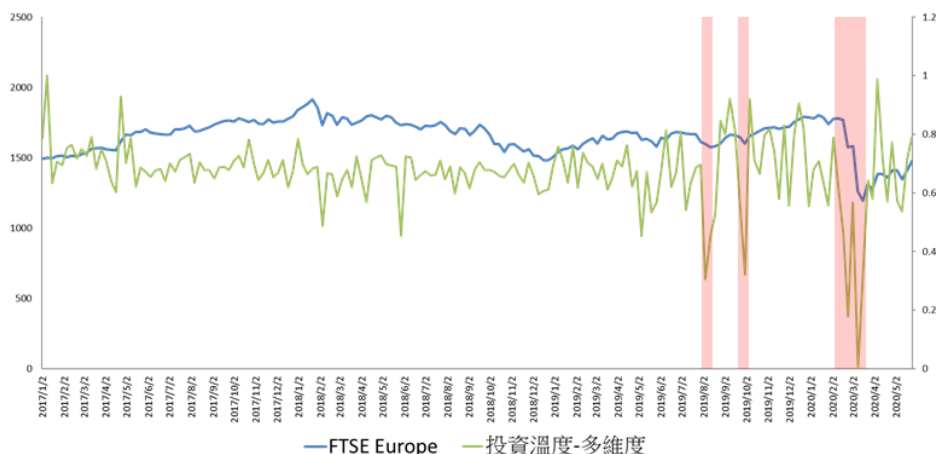
新興市場區域最佳相關係數為 0.724 為高度相關。相關性以折線圖方式表示如圖十，在下列時間點有很高的相關性：2018 年 2 月，股市漲幅過高、通膨預期升溫、引發美股大跌，恐慌氣氛導致新興市場資金顯著外流。2018 年 10 月，美國預告 2019 年起開始對中國加徵關稅，10 月新興市場股匯市表現慘烈。2019 年 5 月，美中貿易戰對峙氛圍升溫，新興市場 5 月中旬將 2019 年漲幅全部回吐。2020 年 2 月，新型冠狀肺炎造成全球性經濟停滯。



圖七、MSCI Emerging Markets Index 與多維度型投資溫度相關圖

7、歐洲區域實驗結果分析

歐洲區域最佳相關係數為 0.717 為高度相關。相關性以折線圖方式表示如圖十一，在下列時間點有很高的相關性：2019 年 8 月，川普下令美企撤出中國，美中貿易衝突再起，引發全球股市震盪，歐股也難逃修正壓力。2019 年 10 月，英國正式提出脫歐新方案，歐股全面大跌超過 3%。2020 年 2 月，新型冠狀肺炎造成全球性經濟停滯。



圖八、FTSE Europe Index 與多維度型投資溫度相關圖

五、結論及未來展望

本研究所做的實驗僅著重於驗證投資溫度評分計算方式與目標市場漲跌趨勢的相關性檢定，在投資溫度評分計算的方式上，本研究係採用類別型、單維度型及多維度型 3 種方式，實驗結果顯示，維度型投資溫度評分計算方式比類別型為好，而多維度型投資溫度評分計算方式在大多數的情形下比單維度型為好，在各組實驗結果相關係數平均值為 0.760，屬高度相關。未來展望及方向主要有幾個重點：

- (一)持續增加網路財經文本蒐集來源。本研究使用的文本來源為網路論壇及網路新聞，如能從其他財經網站或社群媒體中蒐集更多文本，對於將網路投資訊息量化為能夠代表投資大資心理的投資溫度評分更具客觀性。
- (二)本研究的目標是設定在國家及區域等大範圍市場，依據「行情=資金+心理」的投資法則，可以嘗試使用央行的貨幣供給總量、指數交易月均量、全球共同基金資金流向等資料做為「資金」面的變數，使用本研究的投資溫度評分做為「心理」面的變數，進而建構大範圍市場區域的金融商品薦購系統。
- (三)在網路財經文本情感分析的研究中，用來預測股價走勢是大多數研究者的目的，可以嘗試使用本研究投資溫度評分的計算方式，結合基本分析指標(總體經濟資料等)及技術分析指標(移動平均線、KD 值等)，用於建構市場趨勢預測模型。

參考文獻 [References]

- [1] 財團法人中華民國證券暨期貨市場發展基金會，*基金投資人投資行為與偏好問卷調查分析*(資產管理人才培育與產業發展基金委託專題研究)，P.31，民國 107 年。
- [2] 安德烈·科斯托蘭尼，*一個投機者的告白之金錢遊戲* (增修版)，民國 107 年。
- [3] B. Pang and L. Lee, *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval, 2008. 2(1–2): p. 1-135, 2008.
- [4] R. A. Calvo and Sidney. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Computing*, vol. 1, no. 1, pp. 18-37, 2010.
- [5] R. Feldman. "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82-89, 2013.
- [6] Russell, J.A., *A circumplex model of affect*. Journal of personality and social psychology, 1980. 39(6): p. 1161, 1980.
- [7] Jin Wang, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang, "Community-based Weighted Graph Model for Valence-Arousal Prediction of Affective Words," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 1957-1968, 2016.
- [8] 蔡宇祥，*股市趨勢預測之研究－財經評論文本情感分析*，民國 104 年。
- [9] Qu et al.，*Quantifying Correlation between Financial News and Stocks*，IEEE Symposium Series on Computational Intelligence (SSCI)，2016.
- [10] 盧奕叡，*深度學習與情感分析應用於股價預測*，民國 107 年。
- [11] Zhao et al.，*Inferring private information from online news and searches: Correlation and prediction in Chinese stock market*，Physica A，2019.
- [12] Hwang et al.，*Interdependency between the Stock Market and Financial News*，IEEE International Conference on Big Data (Big Data)，2019.
- [13] Bourezk Hind et al.，*Analyzing Moroccan Stock Market using Machine Learning and Sentiment Analysis*，2020 1st International Conference on Innovative Research in

Applied Science, Engineering and Technology (IRASET) Innovative Research in Applied Science, Engineering and Technology (IRASET), 2020 1st International Conference on. :1-5 Apr, 2020.

- [14] M. M. Bradley and P. J. Lang, “Affective norms for English words (ANEW): Instruction manual and affective ratings,” Technical Report C-1, University of Florida, 1999.
- [15] A. B. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 English lemmas,” *Behavior research methods*, vol. 45, no. 4, pp. 1191-1207, 2013.
- [16] M. M. Bradley and P. J. Lang, “Affective norms for English text (ANET): Affective ratings of text and instruction manual. Technical Report. D-1, University of Florida, 2007.
- [17] S. Buechel and U. Hahn. “EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis,” in *Proc. of EACL*, pp. 578-585, 2017.
- [18] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, “Sentiment embeddings with applications to sentiment analysis,” *IEEE Trans. Knowledge and Data Engineering*, vol. 28, no. 2, pp. 496-509, 2016.
- [19] L. C. Yu, J. Wang, K. R. Lai, and X. J. Zhang, “Refining word embeddings using intensity scores for sentiment analysis,” *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 26, no. 3, pp. 671-681, 2018.
- [20] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, “Distributional semantic models for affective text analysis,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2379-2392, 2013.
- [21] G. Paltoglou and M. Thelwall, “Seeing stars of valence and arousal in blog posts,” *IEEE Trans. Affective Computing*, vol. 4, no. 1, pp. 116-123, 2013.
- [22] S. Du and X. Zhang, “Aicyber's system for IALP 2016 shared task: Character-enhanced word vectors and boosted neural networks,” in *Proc. of IALP*, pp. 161-163, 2016.
- [23] P. Goel, D. Kulshreshtha, P. Jain, and K. K. Shukla, “Prayas at EmoInt 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets,” in *Proc.*

of *WASSA*, pp. 58-65, 2017.

- [24] J. Wang, L. C. Yu, K. R. Lai and X. Zhang, "Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 28, pp. 581-591, 2020.
- [25] L. C. Yu, J. Wang, K. R. Lai and X. Zhang, "Pipelined Neural Networks for Phrase-level Sentiment Intensity Prediction," *IEEE Trans. Affective Computing*, to appear.
- [26] L. C. Yu, L. H. Lee, S. Hao, J. Wang, Y. He, J. Hu, K. R. Lai and X. Zhang, "Building Chinese Affective Resources in Valence-Arousal Dimensions," in *Proc. of NAACL-HLT-16*, pp. 540-545, 2016.

探究文本提示於端對端發音訓練系統之應用

Exploiting Text Prompts for the Development of an End-to-End Computer-Assisted Pronunciation Training System

鄭宇森 Yu-Sen Cheng, 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

sam841009@yahoo.com.tw

teinhonglo@gmail.com

berlin@csie.ntnu.edu.tw

摘要

近年來，電腦輔助發音訓練(Computer assisted pronunciation training, CAPT)系統的需求日益上升。然而，現階段基於端對端(End-to-End)類神經網路架構之系統在錯誤發音檢測(Mispronunciation detection)的效能仍未臻完美，其原因是此類系統的內部模型本質上仍是屬於自動語音辨識(Automatic speech recognition, ASR)模型。ASR 目的是儘量正確地辨識出語者所說內容，縱使其發音是有偏誤的；而 CAPT 目的恰巧相反，是要能儘量正確地偵測出語者的錯誤發音。有鑒於此，本論文基於 CAPT 任務通常會有文本提示的特殊性，嘗試將文本提示資訊融入於端對端模型架構。我們研究使用兩個編碼器(Encoders)分別處理發音特徵以及文本特徵，並以分層式注意力機制(Hierarchical attention mechanism, HAN)來動態地結合不同編碼器產生特徵表示。本論文在一套華語學習者語料庫進行一系列實驗；透過不同評估準則所獲得結果顯示，我們所提出的方法較現有方法有較佳的錯誤發音檢測效能。

Abstract

More recently, there is a growing demand for the development of computer assisted pronunciation training (CAPT) systems, which can be capitalized to automatically assess the pronunciation quality of L2 learners. However, current CAPT systems that build on end-to-end (E2E) neural network architectures still fall short of expectation for the detection of

mispronunciations. This is partly because most of their model components are simply designed and optimized for automatic speech recognition (ASR), but are not specifically tailored for CAPT. Unlike ASR that aims to recognize the utterance of a given speaker (even when poorly pronounced) as correctly as possible, CAPT manages to detect pronunciation errors as subtlety as possible. In view of this, we seek to develop an E2E neural CAPT method that makes use of two disparate encoders to generate embedding of an L2 speaker's test utterance and the corresponding canonical pronunciations in the given text prompt, respectively. The outputs of the two encoders are fed into a decoder through a hierarchical attention mechanism (HAM), with the purpose to enable the decoder to focus more on detecting mispronunciations. A series of experiments conducted on an L2 Mandarin Chinese speech corpus have demonstrated the effectiveness of our method in terms of different evaluation metrics, when compared with some state-of-the-art E2E neural CAPT methods.

關鍵詞：端對端語音辨識、電腦輔助發音訓練、分層式注意力機制、發音檢測、發音診斷

Keywords: end-to-end speech recognition, Computer assisted pronunciation training, hierarchical attention mechanism, mispronunciation detection, mispronunciation diagnosis.

一、緒論

近年來，不少語言學習者透過智慧型裝置在網際網路(Internet)上學習，主要的原因在於資訊科技的日漸普及讓許多不容易得到學習資源的人，譬如偏遠地區的學子、生活貧困的人與不易抽出時間上課的人，都可以透過網際網路簡單的取得學習資源。線上學習存在著諸多優點，譬如學習者可以自行選擇合適的教材編排自己的進度、不管通勤或者在家都可以隨時隨地的進行學習，達到 24 小時沉浸式學習(Immersive learning)的效果，並且可以避免因為群聚而得到傳染病的風險。然而目前線上學習技術在語言學習上，仍舊存在著不足。語言學習可以分為聽、說、讀和寫四大部分，而其中口說尤為困難，由於學習者難以自行察覺發音錯誤，一般需要藉由具備專業知識的專家進行評估，才能判斷學習的成果。因此在口說的部份，本論文希望可以藉由電腦輔助人類專家，檢測出學習者的錯誤發音，以利於學習者進行修正。而這樣的技術就被稱為電腦輔助發音訓練(Computer assisted pronunciation training, CAPT) [1][2]。

在 CAPT 過去的研究中[3-5]，多採用語音辨識(Automatic speech recognition, ASR)和似然機率比例(Likelihood ratio)。兩者皆可使用高斯混合模型結合隱藏式馬可夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM) [3][4]或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) [5][6]。在語音辨識的方法中，透過計算最短編輯距離(Edit distance)將模型預測結果與標準答案強制對齊(Forced alignment)，使用對齊後的結果判斷使用者是否存在發音錯誤[2]。另一方面，似然機率比例則是以與人類專家成高度相關的 GOP (Goodness of pronunciation) [7]最為知名。近年來，由於 DNN-HMM 模型架構過於繁雜訓練不易，研究者[2][8]多採用端對端自動化語音辨識(End-to-end ASR) [9]簡化傳統模型繁複的訓練流程。

然而，上述做法都忽視了 CAPT 與 ASR 的目標相異性，CAPT 的目標是要能儘量正確地偵測出語者的錯誤發音，但 ASR 在使用上希望儘量正確地辨識出語者所說內容，縱使其發音是有偏誤的，因此 ASR 系統會傾向將錯誤發音辨識為正確發音。因為上述原因 ASR 在錯誤發音檢測(Mispronunciation detection)以及錯誤發音診斷(Mispronunciation diagnosis)上無法達到最佳化的效果。另一方面，當採用傳統的 ASR 方法執行 CAPT 任務時，忽視了 CAPT 任務相較於傳統 ASR 擁有發音詞彙相應的文本提示。因此也有相關學者[10][11]陸續開始研究文本提示在端對端 CAPT 任務的重要性。在[10]的研究中，將文本提示的資訊加入注意力模型(Attention model)的權重計算，並將具有文字加權影響的聲學隱藏向量(Audio hidden vector)與原始聲學隱藏向量串接，然後作為解碼器的輸入進行預測。而[11]採用多視角(Multi-view)架構，將文本提示視為額外的提示資訊，擁有獨立的編碼器(Encoder)且共享解碼器(Decoder)的參數，結合輸出端的聲學損失函數(包含 CTC 和 Attention)，用以輔助模型判斷聲音的正確與否。

本論文主要探討如何在端對端架構中使用 CAPT 的文本提示增強錯誤發音檢測。模型採用多編碼器(Encoder)架構平行擷取文本提示的音素(Phone)文本特徵與發音的聲學特徵，並透過分層式注意力機制(Hierarchical attention mechanism, HAN)[12][13]結合兩種特徵。過去 HAN 在[13]中被應用於多麥克風陣列的語音辨識，透過注意力機制(Attention)[14][15][16]動態分配權重給來自不同麥克風的資訊並且合併為單一陣列。本

論文希望藉由 HAN 的合併資訊的特性，動態結合來自不同編碼器的文本特徵或聲學特徵中的資訊。解碼器則針對聲學資訊參考[9]所提出的架構，使用連接時序分類(Connectionist temporal classification, CTC)[17]模型對注意力機制的結果進行限制來取得更高的效能。實驗顯示採用文本特徵後的模型在各個評估標準(F-measure, accuracy, precision, and recall)下，我們所提出的方法較現有方法有較佳的錯誤發音檢測效能。

二、對於錯誤發音檢測的端對端語音辨識技術

本論文實驗模型主要採用與[2][8][9][11]相同的端對端架構，並在本章節中介紹與此架構相關的主要技術。

2.1 連結時序分類(CTC) 模型

連結時序分類最早於 2006 年由[17]提出，類似 DNN-HMM 架構，基於條件獨立假設使用貝氏決策法則找出最大事後機率。其過程尋求輸出符號(字母、單字或音素)序列 $C = c_1, c_2, \dots, c_L$ ，在輸入的聲學特徵序列 $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ ，發生時出現的最大機率。並且公式可以分解下方的定義：

$$P_{\text{ctc}}(C|\mathbf{X}) \approx \sum_S P(C|S)P(S|\mathbf{X}) \quad (1)$$

其中 S 代表每個音框的標籤序列，而 CTC 為了避免重複出現的字造成辨識錯誤，在訓練時引入了額外的區塊(blank)標籤 $\langle b \rangle$ 作為標籤間的區隔，因此 S 可表示為 $S = \{s_t \in U \cup \{\langle b \rangle\} | t = 1, \dots, T\}$ 。

2.2 注意力模型(Attention model)

注意力機制過去在機器翻譯(Machine translation)上取得了卓越的成果，近年來也在語音領域上得到優秀的成果[14][15]。其特點是可以在不需要條件獨立假設的情況下直接計算輸出符號對應輸入聲學向量序列的事後機率，注意力模型目標函式可定義為：

$$P_{\text{att}}(C|\mathbf{X}) = \prod_{l=1}^L P(c_l|\mathbf{X}, c_{1:l-1}) \quad (2)$$

上式 2 中的 $P(c_l|\mathbf{X}, c_{1:l-1})$ 藉由編碼器與解碼器的交互作用取得。可以由下列式子推導：

$$\mathbf{h}_t = \text{Encoder}(X) \quad (3)$$

$$e_{lt} = \text{Attention}(\mathbf{q}_{l-1}, \mathbf{h}_t, a_{l-1}) \quad (4)$$

$$a_{lt} = \frac{\exp(\gamma e_{lt})}{\sum_l \exp(\gamma e_{lt})} \quad (5)$$

$$\mathbf{r}_l = \sum_{t=1}^T a_{lt} \mathbf{h}_t \quad (6)$$

$$p(c_l|X, c_{1:l-1}) = \text{Decoder}(\mathbf{r}_l, \mathbf{q}_l, c_{l-1}) \quad (7)$$

其中 \mathbf{h}_t 為編碼器的隱藏向量， a_{lt} 是由 e_{lt} 經由 Softmax 函數轉換為機率分佈後得到的注意力機制權重，而 γ 為 Sharpen Factor，用於在強調權重的分佈， \mathbf{q}_l 是 Decoder 每一層的隱藏向量。

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

注意力模型允許非序列化的對齊，這在機器翻譯或者其他不強調順序的任務中沒有問題。然而，語音辨識是種序列化的任務，因此非單調的對齊會讓其訓練時收斂較慢，但注意力模型因為不需要條件獨立假設，更加貼近真實環境因此在辨識時可以取得優越的表現。與之相對的是，CTC 具有由左至右嚴格單調的對齊。然而，傳統上 CTC 必須搭配其他額外的語言模型才能達到最佳效果。其原因是 CTC 的條件獨立假設會使它與真實環境偏離對效能造成負面影響。因此就有學者[9]提出了結合兩者的優點彌補彼此缺點的 CTC-Attention 混合模型。藉由注意力模型可以得到非條件獨立的前後資訊，並且藉由 CTC 的嚴格單調特性限制注意力模型的計算範圍，在[9]的實驗中指出，這樣的模型能夠比注意力模型更快收斂得到更高的效能。模型訓練時以 λ 作為兩種損失函數的線性相加參數，新的損失函數定義如下：

$$\mathcal{L}_{\text{CTC-ATT}}(C|X) = \lambda \ln P_{\text{ctc}}(C|X) + (1 - \lambda) \ln P_{\text{att}}(C|X) \quad (8)$$

需要注意的是因為文字資訊不需要 CTC 的對齊，所以本論文的雙編碼器架構中的文字資訊注意力模型，並沒有使用 CTC 輔助。

三、文本提示在端對端發音檢測與診斷的使用

本論文將在本章節介紹受到[13]的多編碼器架構與 HAN 啟發的採用文本提示的多編碼器端對端架構。模型採用兩個平行獨立的編碼器，分別為發音編碼器與文本提示編碼器，透過這兩個編碼器分別抽取聲學和文本的特徵。接著透過 HAN 技術動態整合兩種不同維度的特徵。以下部分將針對本架構細節部分進行說明。

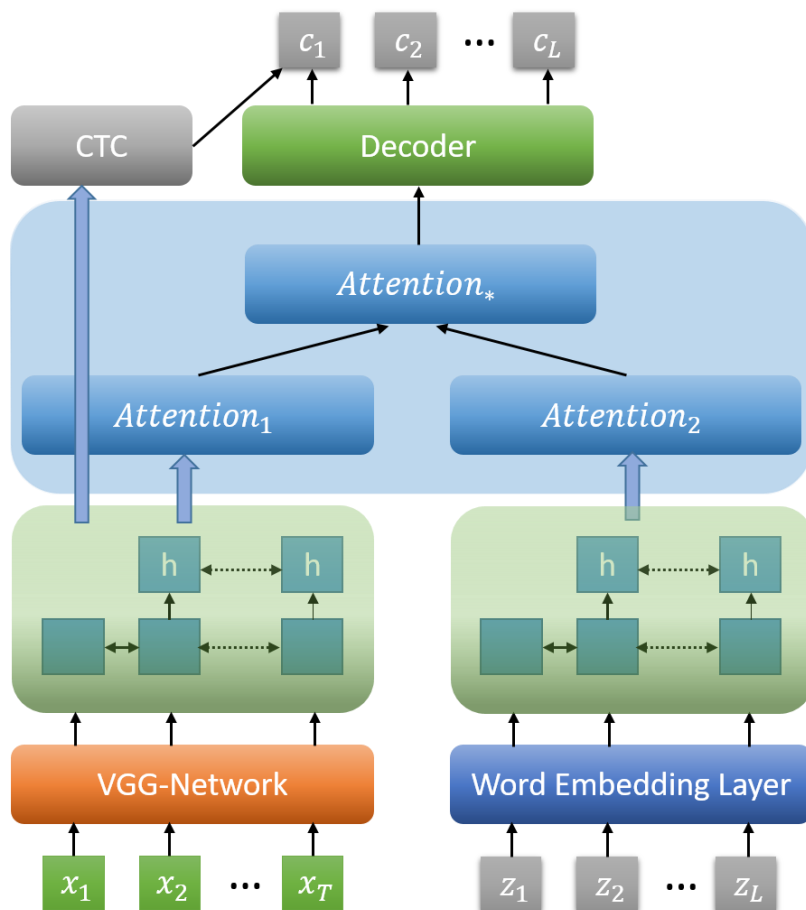


圖 1、多編碼器端對端錯誤發音檢測系統

3.1 文本提示的添加

如圖 1 所示，文本提示以音素層級的符號序列 $Z = z_1, z_2, \dots, z_L$ 被輸入到音素向量層轉換為對應的向量，並輸入到長短期記憶類神經網路(Long short-term memory, LSTM)[18][19] 中抽取文本特徵，然後將文本特徵輸入到 HAN 層中動態分配與聲學特徵合併的權重。

3.2 聲學特徵處理

針對聲學特徵的處理，本論文參考[8][9]的 CTC-Attention 混合模型架構，採用 VGG-LSTM 的架構，透過 VGG 層提取聲學特徵，然後用 LSTM 學習聲學特徵中的時序資訊，之後將編碼器輸出的具聲學資訊的隱藏向量輸入到 HAN 層與文本特徵合併。

3.3 分層式注意力機制(Hierarchical attention mechanism, HAN)

由於聲學與文本兩種資訊的空間維度不同，無法直接合併，本論文採用 HAN 技術整合兩種來自不同維度的資訊。公式定義如下：

$$h_t^1 = \text{Encoder}_1(X), \quad h_t^2 = \text{Encoder}_2(Z) \quad (9)$$

$$r_i^i = \sum_{t=1}^{T^1, L^2} \alpha_{it}^i h_t^i, i \in \{1,2\} \quad (10)$$

$$r_i^* = \beta_{i1} r_i^1 + \beta_{i2} r_i^2 \quad (11)$$

$$\beta_{ii} = \text{Attention}(q_{i-1}, r_i^i), i \in \{1,2\} \quad (12)$$

公式中(9)的兩個Encoder_{1,2}分別發音與文本提示的編碼器，這裡定義 $i \in \{1,2\}$ 作為編碼器與對應隱藏向量的索引。第(10)式中的 α 是兩個編碼器各自的隱藏向量輸入注意力模型所得到的輸出，具體計算過程請參照公式(4)與(5)。透過公式(12)對公式(10)的兩個加權向量 r_i^i 計算權重，並於公式(11)相加得出最後的加權向量 r_i^* 作為公式(7)解碼器的輸入。

四、實驗設定

在這個章節會介紹本實驗所使用的語料集，以及實驗相關參數設定還有開發架構，以利於其他研究者覆現本實驗之結果。

4.1 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[20]，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，我們將 L1 語料均視為正確發音，而 L2 的部分則標記有文本提示以及學習者的真正發音。為了貼近訓練與測

試的條件，我們將 L1 訓練集與 L2 訓練集合併作為我們的訓練集，而測試集本實驗採用的是 L2 測試集，詳統計資訊如表 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

4.2 模型設定

本論文的實驗皆採用開源端對端語音辨識工具“Espnet”[21]完成，在參數上聲學部分參考 [2]的模型設定，聲學部分採用 VGG 連接雙向 LSTM，並且設置了投影層，透過大於[2]的節點數量得到更好的效果，文本部分考量文字與聲音的相異性將 VGG 層換成音素向量層對文本部分編碼，具體設定如表 2 所示，名詞表示參考“Espnet”設定所需使用之名詞。

表2、實驗參數設定

	聲學模型($Encoder_1$)	文本模型($Encoder_2$)
特徵	80-dim fbank + 3-dim pitch	pytorch word2vec
編碼器種類	VGGBLSTMP	BLSTMP
編碼器層數	3	3
編碼器節點數	1024 (BLSTMP)	1024 (BLSTMP)
解碼器種類	LSTM	LSTM
解碼器層數	2	
解碼器節點數	1024	
CTC/Attention 混合比	0.5/0.5	0/1

五、實驗結果與分析

在本章節會將展示使用華語學習者語料庫進行的一系列實驗的數據結果，並且對數據進行 nbest 以及華語在語言學上的分析。

5.1 辨識結果

為了驗證採用文本提示後，系統在 CAPT 任務的有效性，本實驗對前述(4.1)測試集進行多種評估標準(F-measure, accuracy, precision, and recall)研究，比較基線為未採用文本提示的傳統端對端 CTC-Attention 混合模型[8][9]，其參數設定與多編碼器模型中聲學部分相同。

從評估結果可以發現本論文提出的模型在各個評估標準下皆優於原始沒有採用文本提示的基線。此外本論文比較了沒有使用 CTC 輔助聲學部分對齊的結果，發現在 precision 的部分得到了提升，然而其他部分是下降的，可以判斷當沒有 CTC 參與時，聲學部分的辨識結果會往文本提示過度擬合。具體數據參見表 3。

另一方面，本論文將相近研究[11]納入比較後，發現本實驗的自動分配權重的方法能夠與[11]在解碼器人工調適文本提示與聲學損失函數結合權重的結果旗鼓相當，且因為少了人工調適的過程，在開發上成本上相對較少。另一方面，從[11]的數據可以表較使用 DNN-HMM 的模型(GOP+MFC)的 F1 65.2%相較低於端對端基線的 F1 69.2%，所以目前端對端方法在錯誤發音檢測上，明顯優於過去傳統的 DNN-HMM 方法。另外一份相近研究[10]研究英語學習者，但本論文主要在探討華語學習者的 CAPT，因此不列入比較。

表3、L2測試集的音素錯誤率與音節錯誤率

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
GOP [11]	-	-	-	51.8%	63.5%	57.0%
GOP+MFC [11]	-	-	-	69.5%	61.3%	65.2%
CTC-ATT(SR)	-	-	-	70.8%	67.9%	69.2%

CTC-ATT(SR) +PS [11]	-	-	-	71.8%	68.4%	70.2%
Baseline	87.7%	89.1%	88.4%	70.7%	67.7%	69.1%
Multi-encoder (without CTC)	86.9%	89.2%	88.0%	66.4%	71.2%	68.7%
Propose (with CTC)	88.3%	89.4%	88.9%	71.3%	69%	70.2%

5.2 N-best 結果

參考[2][11]的實驗，本論文將表3中具有 CTC 的本實驗模型，採用了 N-best 作為比較，可以發現隨著條件的放寬對於 CAPT 的影響是負面的，對於錯誤發音檢測而言，雖然 Precision 提高了，但 Recall 會嚴重下降，具體數據見於表5。

表5、N-best 對於 CAPT 任務的效能影響

	Correct pronunciation			Mispronunciation		
	Recall	Precision	F1	Recall	Precision	F1
1-best	88.3%	89.4%	88.9%	71.3%	69%	70.2%
2-best	61.3%	85.8%	71.5%	40.5%	72.3%	51.9%
3-best	55.0%	87.3%	67.5%	38.8%	78.1%	51.8%
4-best	53.7%	88.6%	66.9%	39.0%	81.1%	52.7%
5-best	53.3%	90.0%	66.9%	39.5%	83.7%	53.7%

5.3 語言學分析

本論文以語言學的角度分析系統的辨識能力。在表 7 中探討聲母(Initial)與韻母(Final)在四種量測象限的分布情況，量測象限定義如下：

1. 正確接受(True accept, TA): 系統辨識此發音為正確發音，且確實為正確發音。
2. 錯誤接受(False accept, FA): 系統辨識此發音為正確發音，但實際為錯誤發音。
3. 正確拒絕(True rejection, TR): 系統辨識此發音為錯誤發音，且確實為正確發音。
4. 錯誤拒絕(False rejection, FR): 系統辨識此發音為錯誤發音，但實際為正確發音。

量測象限對應關係可以參考表 6，其中 CP 代表正確發音，MP 代表錯誤發音，Ground Truth 為文本提示。

表6、錯誤發音量測象限

		Ground Truth	
		CP	MP
Model Prediction	CP	True accept (TA)	False accept (FA)
	MP	True rejection (TR)	False rejection (FR)

表7、聲母韻母量測象限分布

	TA	FA	TR	FR
Initial	19595 (76%)	1318(05%)	3347(13%)	1615(06%)
Final	14364 (54%)	2701(10%)	6654(25%)	2874(11%)
Total	33959(65%)	4019(08%)	10001(19%)	4489(09%)

數據上可以發現不管是系統的辨識或者是使用者本身的發音，在聲母上的正確率較高。本論文推測這是因為在韻母的發音上具有聲調(Tone)，所以不利於使用者發音，且也不利於系統辨識。因此在表 8 中本論文嘗試觀察中文的 5 種聲調對於辨識的影響，可以發現在一聲與四聲的 TA 比例較高，而三聲與輕聲的比例較低，其中輕聲因為資料所佔比例較少，所以容易判斷錯誤。本論文進一步在表 9 調查當韻母聲調錯誤時，L2 學習者在面對各個時，會發成何種聲調，藉此了解學習者在聲調上發音錯誤的狀況，其中 None 代表這種 L2 學習者的聲調無法被歸類在五種聲調中，可能是介於兩種聲調之間，或者為受到其 L1 語言影響的聲調。從表 9 可以觀察到，三聲最常與二聲混淆，本論文判斷主要的原因是因為相較於其他聲調，三聲的 F0 輪廓類似於 V 字型[22]，後段上揚部分容易被視為二聲。其中四聲發音混淆的比例最低，本論文判斷是因為相較於其他聲調，四聲相對低頻，因此可以讓人耳以及模型獲得更多資訊以利判斷。

表8、發音聲調量測象限分布

	TA	FA	TR	FR
--	----	----	----	----

Tone 1	4413(66%)	536(08%)	1142(17%)	592(09%)
Tone 2	2544(47%)	603(11%)	1498(28%)	726(14%)
Tone 3	2386(35%)	885(13%)	2677(40%)	789(12%)
Tone 4	4780(67%)	604(08%)	1047(15%)	700(10%)
Tone 5	241(36%)	73(11%)	290(43%)	67(10%)
Total	14364(54%)	2701(10%)	6654(25%)	2874(11%)

表9、辨識結果聲調混淆矩陣

		Non-native Pronunciation					
		Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	None
Canonical	Tone 1	6071(91%)	183(03%)	95(01%)	234(04%)	9(00%)	91(01%)
	Tone 2	335(06%)	3840(71%)	1010(19%)	27(01%)	8(00%)	151(03%)
	Tone 3	86(01%)	2239(33%)	4088(61%)	122(02%)	7(00%)	195(03%)
	Tone 4	101(01%)	16(00%)	110(01%)	6757(95%)	82(01%)	65(01%)
	Tone 5	178(27%)	20(03%)	24(04%)	101(15%)	344(51%)	4(01%)

六、結論

本論文實驗了在端對端錯誤發音檢測系統上使用多編碼器結構處理文本提示的特徵，並且以 HAN 動態的合併不同來源的資訊。實驗部分相較於端對端基線在多種評估標準下都取得了良好的進步，證明了這個方法的有效性。另一方面，使用了混淆矩陣進行分析，具體地顯示出此方法是如何影響評估結果。在未來的部分，希望對於文本特徵與發音特徵可以找到更有效的結合法。此外考量模型可能會過於依賴文本提示而造成偏差，希望尋找一個有效的方法，針對文本特徵與發音特徵嘗試計算其是否映射到相同結果，如果是則加強文本特徵的影響，否則降低文本特徵的影響或反向拉遠希望預測結果偏離文本特徵的映射目標。最後，本論文目前只有使用一個資料集進行實驗，未來希望可以測試在更多大型資料集上是否會有不一樣的表現。

參考文獻

- [1] Eskenazi, Maxine, “An overview of spoken language technology for education,” Speech

- Communication, vol. 51, no. 10, pp. 832–844, 2009.
- [2] Chang, Hsiu-Jui et al., “*Investigating on Computer-Assisted Pronunciation Training Leveraging End-to-End Speech Recognition Techniques*,” ROCLING, 2019.
 - [3] Lawrence R. Rabiner et al., “*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*,” Proceedings of the IEEE, 1989.
 - [4] Mark Gales and Steve Yang, “*The Application of Hidden Markov Models in Speech Recognition*,” Foundations and Trends® in Signal Processing, 2008.
 - [5] 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.
 - [6] Metallinou, Angeliki, and Jian Cheng, “*Using deep neural networks to improve proficiency assessment for children English language learners*,” Interspeech, 2014.
 - [7] Witt, Silke M., and Steve J. Young, “*Phone-level pronunciation scoring and assessment for interactive language learning*,” Speech communication vol. 30.2-3, pp. 95-108, 2000.
 - [8] Leung, Wai-Kim et al., “*CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis*,” ICASSP, 2019.
 - [9] Watanabe, Shinji et al., “*Hybrid CTC/attention architecture for end-to-end speech recognition*,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240-1253, 2017.
 - [10] Feng, Yiqing et al., “*SED-MDD: Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis*,” ICASSP, 2020.
 - [11] Lo, Tien-Hong et al., “*An Effective End-to-End Modeling Approach for Mispronunciation Detection*,” arXiv, 2020.
 - [12] Yang, Zichao et al., “*Hierarchical attention networks for document classification*,” Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016.
 - [13] Wang, Xiaofei et al., “*Stream attention-based multi-array end-to-end speech recognition*,” ICASSP, 2019.
 - [14] Chorowski, Jan et al., “*End-to-end continuous speech recognition using attention-based recurrent NN: First results*,” arXiv, 2014.
 - [15] Chorowski, Jan, et al., “*Attention-based models for speech recognition*,” Advances in neural information processing systems, 2015.
 - [16] Vaswani, Ashish et al., “*Attention is all you need*,” Advances in neural information processing systems, 2017.
 - [17] Graves, Alex et al., “*Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*,” ICML, 2006.
 - [18] Graves, Alex et al., “*Speech recognition with deep recurrent neural networks*,” ICASSP, 2013.

- [19] Sak, Haşim et al., “*Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition,*” arXiv, 2014.
- [20] Hsiung, Y. et al., “*Development of Mandarin annotated spoken corpus (MAS Corpus) and the learner corpus analysis,*” WoALF, 2014.
- [21] Watanabe, Shinji et al., “*ESPnet: End-to-End Speech Processing Toolkit,*” Interspeech, 2018.
- [22] Lin, Ju et al., “*Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features using extended recognition networks,*” Journal of Signal Processing Systems 90.7, 2018.

基於混合注意力機制與長短期記憶之股票趨勢預測

Combining Hybrid Attention Networks and LSTM for Stock Trend Prediction

劉馨文 Hsin-Wen Liu

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

t107598027@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@ntut.edu.tw

摘要

本研究結合長短期記憶(LSTM)中股價時間序列的特徵，以及混合注意力模型(HAN)中模擬人判讀新聞影響股票重要性，有效地學習股價時間序列和新聞報導中的訊息順序，藉此訓練新聞文章與歷史股票交易資料之間的關聯，建構股票漲跌趨勢之模型。根據實驗結果，相較於僅單獨使用新聞或是股價資訊的模型，加入時間序列的新聞文章訊息使模型更能精準的預測股價市場趨勢，在兩種資訊的結合中，與 HAN、LSTM 模型相比，最佳準確度為 80%，整體最高可提升 40%的準確度。

Abstract

Our research merge two different models of the Hybrid Attention Networks (HAN) and the Long Short-Term Memory (LSTM) to improve the stock trend prediction. The combination of the two algorithms helps leverage the advantages of both models to learn sequential information in time series and news articles. The experimental results show that the best accuracy score, combined with news and stock prices, is 80 %. The performance of the proposed model compared to HAN and LSTM model increased by up to 40%.

關鍵詞：深度神經網路、注意力機制、股票預測

Keywords: deep attention networks, attention mechanism, stock prediction

一、緒論

股票市場一直是非常熱門且具有挑戰性，而一般投資人往往參考財經相關新聞或是技術指標來評估投資標的，而隨著科技的進步，網路時代的發達，電子媒體則是網路族群獲得資訊的主要來源，當重大的新聞消息發佈時，往往衝擊投資人的交易行為，造成股價一定程度的波動。因此，面對各式各樣的消息，投資者在投資時的規劃與預測分析是相當重要。

目前對股票趨勢預測的方法主要有：

- 利用時間序列中一段股票歷史數據
- 利用自然語言處理技術，金融新聞能夠有效地影響股票價格

但是股票歷史數據與新聞文本之間的結合與研究相對較少，大部分考慮單一新聞或是股價資訊影響，因此本研究將探討如何融合兩種不同的資料來源，並且使用深度學習的方式，藉此訓練兩者之間的關係，提高股票預測的準確度。

二、相關研究

(一) . 基於股價之股票預測

股票市場每天都會產生大量的交易數據，提供大量資料，有利於深度神經網路訓練以及提高其預測能力。Hiransha M 等人[1]提出使用不同的深度學習網路，像是多層感知器(MLP)模型、卷積神經網路(CNN)、遞歸神經網路(RNN)和長短期記憶神經網路(LSTM)來比較各個模型之間的表現差異，並且觀察到神經網路模型優於時間序列的線性模型。Kai 等人[2]根據股票的歷史資訊結合 LSTM 神經網路，來預測後三天的股票趨勢是屬於哪一個區間，與隨機預測方法相比，LSTM 模型將股票收益預測的準確性從 14.3% 提高到 27.2%。

許多研究發現，深度學習神經網路具有儲存能力，因此它比其他機器學習方法

有更好時間序列數據的學習能力。

(二) . 基於新聞之股票預測

隨著自然語言處理(Natural Language Processing) 的發展，使電腦把輸入的語言變成有意思的符號和關係，根據其目的再處理。Heeyoung Lee 等人[3]實驗有包含文字或財經訊息的模型，發現文字，是影響股票預測的重要性。自然語言處理(NLP)中單詞嵌入的發展，深度神經網路可以通過學習單詞向量來有效地掌握文本中的訊息。Xiao Ding 等人[4]使用 Open Information Extraction techniques(Open IE)提取事件 Event Extraction 並且用 WordNet 和 VerbNet 概括結構化事件特徵，以減少其稀疏性，Xiao Ding 等人[5]提出該方法的延伸，使用神經張量網路(Neural Tensor Network)來訓練新聞標題的事件嵌入，並使用卷積神經網絡(CNN)來預測 S&P 500 及其成分股的波動率。Hu Liu 等人[6]用新聞向量和價格數據訓練雙向 GRU 模型來預測股票的日波動率。他們的結果發現，市場新聞和股票價格的綜合訊息可以提高在日內交易上下文中對股票期貨價格回報預測的準確性。

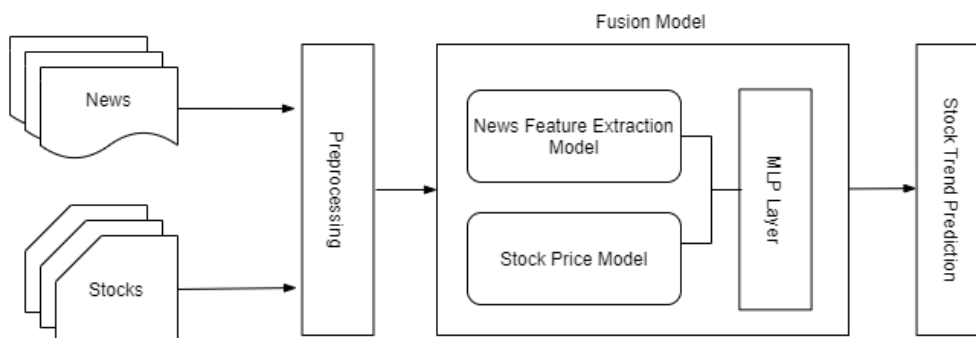
(三) . 基於結合新聞和股價之股票預測

Yuzheng Zhai 等人[7]使用支援向量機(Support Vector Machine)的方式，提出結合相關的新聞和技術指標可增強股票趨勢的可預測性，實現更高的精準度。Xiaodong Li 等人[8]使用深度學習的方式，結合新聞文章的情緒、技術指標和股價，輸入於 LSTM 進行訓練和預測，此篇論文證實，包含新聞與股價的模型優於僅使用單一新聞或是股價資訊的模型，並發現於四個情緒字典中，特定於金融領域的情緒字典(Loughran - McDonald Financial Dictionary)對新聞情緒進行了更好的建模，與其他三個詞典相比，可以更好地提高預測性能。

總結先前的方法，我們知道用來預測股票的趨勢可以分為三大類：股價、新聞、以及結合新聞與股價。本研究中，我們將嘗試結合新聞文本與股價以進行股票之預測，透過結合兩種資料的序列特徵，加強模型對股票趨勢預測之精準度。

三、研究方法

我們在此章節說明研究的方法及架構，本研究主要分為四大部分，分別是系統架構、資料前處理、融合模型之選擇、模型融合之方法。



圖一、系統架構圖

(一) 系統架構

Preprocessing 負責處理新聞文章資料集、歷年股價資料集、融合模型資料之產生。Fusion Model 負責結合兩種模型，分別為針對新聞文章進行訓練與針對股價進行訓練，並且把 Fusion Model 訓練出來之特徵矩陣輸入於多層感知器

(MLP) 訓練。Stock Trend Prediction 負責把多層感知器(MLP)輸出之向量轉為漲、跌或是持平。

(二) 資料前處理

1. 新聞資料前處理

標準普爾 500(S&P 500)為在美國股市的兩大股票交易市場。我們在處理新聞資料時，會以此 500 間公司當成目標，從新聞資料集中找出提到這幾間公司的文章。

首先我們會過濾掉標題和內容沒有出現於 S&P 500 公司的新聞，並且根據新聞發布的日期時間序列排序。根據 Jey Han Lau 等人[9]證實長文本中 Doc2Vec 的效果優於 Word2Vec，因此本論文中採用 Doc2Vec 的 PV-DM 模型來訓練文本向量。PV-DM 模型保留文檔中的單詞順序，經有轉換後新聞的格式為每一筆

[5,40,200]的三維矩陣，在三維矩陣中 5 代表天數，40 代表一天最多有 40 篇文章，200 代表文章向量維度。

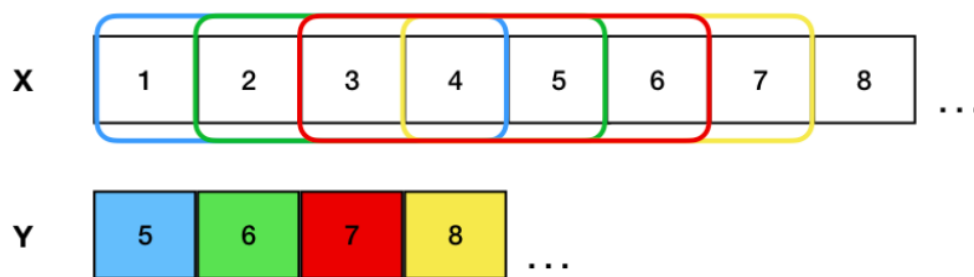
2. 股價資料前處理

我們將股價按照時間序列排序，並經由 Z 分數(Z-Score)標準化值，公式如式

(1)，資料將符合標準常態分佈(Standard Normal Distribution)，可透過 Z 分數標準化來降低離群值對整個模型的影響。資料經過標準化後，它能帶給模型兩個優點，提升模型的收斂程度、提高模型的精準度。

$$Z = \frac{X-\mu}{\sigma}, \sigma \neq 0 \quad (1)$$

為了輸入於長短期記憶神經網路(LSTM)模型，我們必須經過以下轉換格式，假設給定時間序列 N 天的長度，預測第 N+i 天的股票趨勢 Yi 為漲、跌或是持平，時間從 i 到 i + N - 1 天的股價資料表示為 Xi: [Xi, ..., Xi+N-1]，預測的資料為 Yi 而 (Xi, Yi) 被用作 LSTM 神經網路的輸入，如圖二[12]所示，window size 長度為 4 天的滑動窗口示意圖。



Sliding window algorithm of sequence length 4, for data (X) and corresponding labels (Y).

圖二、模型輸入示意圖

3. Fusion Model 資料前處理

在上面小節我們有介紹到，針對新聞文章為了降低 Model 訓練時的記憶體使用量，我們使用 Doc2Vec 將複雜文章轉換為 200 維度之矩陣，針對股價為了避免各家公司股價高低不同造成模型預測失準，我們使用 Z 分數標準化(Z-Score

Standardization)來正規化股價資訊。

Fusion Model 由於是結合兩種不一樣之模型，因此輸入也必須做處理才能使 Fusion Model 進行訓練，我們必須串接文章向量與股價資訊，而我們為了要讓模型能夠同時拿到新聞文章 200 維度之特徵和股價向量開盤、收盤之特徵，我們使用串接(Concatenate)的方式結合兩種資料，讓產生出來的資料集每一筆都包含新聞文章與股價之特徵。

4. 股價標籤分類

股價趨勢預測通常有上漲、持平、下跌三個級距，以下介紹歸類之方法。對於給定日期 t 和股價 s ，參考 Ziniu Hu 等人[6]對於股票標籤分類的評估方法，通過以下公式計算其上漲百分比，公式如式(2)。把當日開盤價格與前一天的開盤價格相減，除以前一天開盤價格，可以計算出其上漲的百分比幅度，並依照其上漲百分比歸為不同的類別。

$$Rise_Percent_{(t)} = \frac{Open_Price_{(t+1)} - Open_Price_{(t)}}{Open_Price_{(t)}} \quad (2)$$

為了使類別的判斷能夠平均分布，因此定義閾值來幫助類別判斷之篩選，經過實驗後，發現 0.45% 閾值，能使三個類別(漲、跌、持平)的資料集數量平均分布。

- 漲(Up):
 - $Rise_Percent(t) > 0.45\%$
- 跌(Down):
 - $Rise_Percent(t) < -0.45\%$
- 持平(Preserve):
 - $-0.45\% \leq Rise_Percent(t) \leq 0.45\%$

公式(3)為 S&P 500 所有公司在 2016~2017 年期間，股價趨勢為漲的資料集總數、公式(4)為 S&P 500 所有公司在 2016~2017 年期間，股價趨勢為跌的資料

集總數、公式(5)為 S&P 500 所有公司在 2016~2017 年期間，股價趨勢為持平的資料集總數， n 為 S&P 500 公司總數量、 $threshold$ 為不同 Rise_Percent 參數。

$$total_{Up} = \sum_{i=1}^n Up_i^{threshold} \quad (3)$$

$$total_{Down} = \sum_{i=1}^n Down_i^{threshold} \quad (4)$$

$$total_{Preserve} = \sum_{i=1}^n Preserve_i^{threshold} \quad (5)$$

公式(6)為 S&P 500 所有公司在 2016~2017 年期間，股價趨勢資料集總數量、公式(7)是股價趨勢為漲的資料集在全部資料集中所佔的百分比、公式(8)是股價趨勢為跌的資料集在全部資料集中所佔的百分比、公式(9)是股價趨勢為持平的資料集在全部資料集中所佔的百分比。如表 3.1 顯示，當 $threshold$ (Rise_Percent) 為 0.45%時，能使股價趨勢為漲、跌、持平的資料集在全部資料集中所佔的百分比趨近於相等。

$$total = total_{Up} + total_{Down} + total_{Preserve} \quad (6)$$

$$Distrubition_Percent_{Up} = \frac{total_{Up}}{total} \quad (7)$$

$$Distrubition_Percent_{Down} = \frac{total_{Down}}{total} \quad (8)$$

$$Distrubition_Percent_{Preserve} = \frac{total_{Preserve}}{total} \quad (9)$$

(三) 模型融合之選擇

本章節將介紹新聞與股價模型之選擇，以及其選擇的原因。

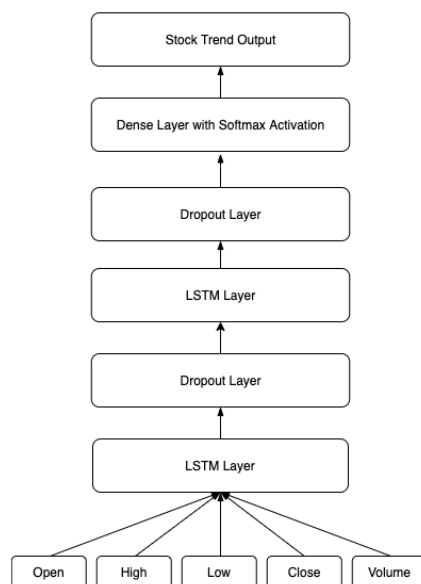
1. 混合注意力機制(Hybrid Attention Networks)

由於 Zichao Yang 等人[10]提出了 Hierarchical Attention 方式去進行文本分類，相較於先前研究於文本分類模型中是最好的，Ziniu Hu 等人[6]提出的混合注意力機制(Hybrid Attention Networks)模型，是根據 Zichao Yang 等人[10]提出的 Hierarchical Attention 架構應用於股票趨勢預測，賦予文章不同的權重，找出影

響股票的重要資訊，經由實驗證實效果為最佳。

在新聞文章訓練時，我們會需要各個文章彼此之間的關聯性而不是每篇文章獨立去進行訓練，因此我們需要選擇一種模型來幫助我們達到訓練文章彼此之間的關聯性，針對影響股價的重要性，賦予新聞文章不同的權重，使模型有效地學習人在閱讀新聞文章中的順序訊息，能夠幫助我們更精準的達到股票趨勢預測之效果。

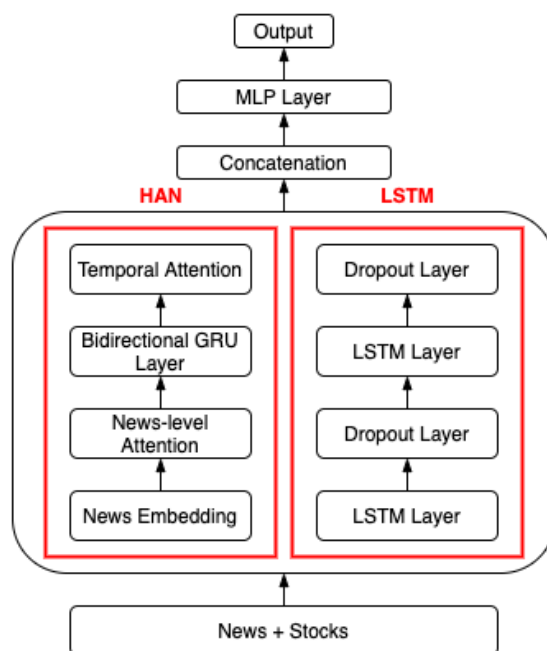
2. 階層式長短期記憶神經網路(Hierarchical LSTM)



圖三、Hierarchical LSTM 架構圖

本篇論文架構採用長短期記憶網路(LSTM)，參考 Kai 等人[2]的設計，為了在順序訊息中找出時間序列的特徵，加上 LSTM 具有存儲能力，可以幫助改善預測效能，因此本研究選擇 LSTM 模型來訓練股價。本研究所提出的階層式長短期記憶神經網路(Hierarchical LSTM)架構中，經過實驗，發現兩層的 LSTM 效果最佳。為了避免過度擬合，每層 LSTM 連接一個 Dropout Layer 於每一個序列時間長度中，Dropout 將會根據訓練中的給定機率分佈將某些向量設置為零，對於遞迴神經網路，Dropout 可以迫使循環層更穩健地執行其中間運算，而不會移除單元狀態中的所有訊息。

(四) 模型融合之方法



圖四、Fusion Model 架構圖

Fusion Model 結合方式如圖四。使用兩個模型完整的向量輸出，結合 (concatenate) 時間序列注意機制 (Temporal Attention) 與 LSTM 模型架構中第二層 LSTM 輸出最後一個隱藏狀態作為輸出能達到以下效果，Temporal Attention 能夠在時間序列的新聞文章當中找出影響力最大的新聞文章子集合、透過使用兩層 LSTM 和 Dropout layer 的結合能夠使股票預測之效果更為精確方便找出股價的重要特徵，結合後再輸入於 MLP 進行分類來達到使用新聞文章與股價進行股票趨勢預測之效果。時間序列注意機制 (Temporal Attention) 設計的位置也會對模型造成一定的影響，在 Fusion Model 中只針對新聞文章進行注意力權重訓練，從新聞文章中找出影響力較大的文章跟 LSTM 輸出的股票預測向量結合，以此達到透過影響力大之新聞文章來影響股價預測之效果。

透過混合注意力機制 (HAN) 模型強化判讀新聞影響股票重要性、長短期記憶神經網路 (LSTM) 學習時間序列的特徵，使模型更有效的結合兩種不同類型的資

訊，提高預測的精準度。

四、實驗與討論

(一) . 資料集

本研究所使用的資料集為 Kaggle 上所提供的路透(Reuters)財經新聞文章、標準普爾 500(S&P 500)的股價歷史資訊。如表一所示，分別為新聞文章數量以及股價公司資料筆數。資料的處理方式為 80%訓練集，20%測試集。

表 一、新聞文章與股價的資料數量

	Reuters	S&P 500
數量	10686 篇	500 間
時間	2016/1/1~2017/7/5	2016/1/1~2017/7/5

(二) . 環境設置

研究使用 Adam[11]作為優化器，並設定其初始學習率為 5×10^{-5} ，每一次訓練都使用 Early Stopping，而 Early Stopping 的設定為 monitor = 'val_loss', min_delta = 0.0001, mode = 'min', verbose = 2, patience = 200，避免過度擬合、學習率過大導致不收斂。Batch size 為 8，最大 epoch 數為 500。使用 Three-fold-cross-validation 來評估模型之精準度。

(三) . HAN 模型實驗

我們針對混合注意力機制(HAN)模型進行實驗，根據之前的相關研究以 Window Size 為 5 天預測第 6 天和 window size 為 10 天來預測第 11 天的股票趨勢，來探討是否影響股票趨勢預測。預測新聞文章中出現過的標準普爾 500(S&P 500)公司之股票趨勢準確率，並根據全部準確率加總平均，由表二可以發現，長(Window Size 為 10)和短(Window Size 為 5)，準確度皆為 0.4，對於混合注意力機制(HAN)模型沒有太大的差異和影響。

表 二、HAN 模型在不同 Window size 之比較圖

	Window Size = 5	Window Size = 10
Average Accuracy	0.4	0.4
Average F1 Score	0.33	0.31

(四) . Hierarchical LSTM 模型實驗

本小節將探討，Window Size 長短對於階層式長短期記憶神經網路(Hierarchical LSTM)模型之影響以及股票特徵選取之比較。針對標準普爾 500(S&P 500)每一間公司算出各自的準確率並且加總平均。由表三可以發現對於時間長短之影響，其中又以短期(Window Size 為 5)的時間序列效果較好。針對特徵之選取，相較於其他特徵的選取，單獨只取收盤(Close)特徵的股票趨勢預測較佳。根據實驗結果，可以發現 LSTM 適合處理和預測時間序列的問題，對於短期的預測效果較佳，選取單一特徵，將不受其他資訊的干擾，可以有更好的表現。

表 三、Hierarchical LSTM Model 不同 Window Size 比較圖

Window Size = 5	Close	Close/Volume	Close/Open/High/Low /Volume
Average Accuracy	0.72	0.46	0.65
Average F1 Score	0.7	0.41	0.63
Window Size = 10	Close	Close/Volume	Close/Open/High/Low /Volume
Average Accuracy	0.68	0.42	0.55
Average F1 Score	0.56	0.39	0.51

(五) . Fusion Model 實驗

為了探討 Fusion Model 的有效性，我們討論時間序列對於模型的影響以及特徵之選取的比較，針對標準普爾 500(S&P 500)每一間公司算出各自的準確率並且加總平均。由表四可以發現，本篇論文所提出的方法架構，結合新聞和股價，

對於整體股票預測趨勢有顯著的提升，其中，又以短期(Window Size 為 5)預測效果較長期(Window Size 為 10)為佳，特徵選去的部分，則是拿單一收盤(Close)的準確率最佳。

表 四、Fusion Model 不同 Window Size 比較圖

Window Size = 5	Close	Close/Volume	Close/Open/High/Low /Volume
Average Accuracy	0.8	0.61	0.71
Average F1 Score	0.79	0.57	0.69
Window Size = 10	Close	Close/Volume	Close/Open/High/Low /Volume
Average Accuracy	0.77	0.47	0.62
Average F1 Score	0.71	0.45	0.57

(六) . 模型加入股價資訊的效果

本小節針對 Ziniu Hu 等人[6]提出混合注意力機制(HAN)模型相比，此論文的 Window Size 設定為 10，因此我們也把 Window Size 設定為 10，輸入我們的資料集以及比較本研究所提出的混合深度模型(Fusion Model)，綜合比較不同模型之間的效果。如表五。可以觀察到本篇論文所提出的架構，加入股價時間資訊，對於整體有非常大的提升。原本只有新聞的混合注意力機制(HAN)模型從 0.4，加入股價後，準確率可以達到 0.77。

表 五、加入股價資訊的效果比較

Model	HAN With Reuters	Fusion Model With Reuters + S&P 500
Average Accuracy	0.4	0.77
Average F1 Score	0.33	0.71

(七) . 新聞與股價資訊結合方法的效果

表 六、各個模型綜合比較表

	HAN	Hierarchical LSTM	Fusion Model
Average Accuracy	0.4	0.72	0.8
Average F1 Score	0.33	0.7	0.79

由表六中，我們可以看到本論文提出的 Fusion Model 準確度(80%)在所有模型當中表現最佳，Fusion Model 於最後結合新聞與股價有助於提升模型整體的效果，因為不同模型能夠針對不同資料分別擷取更有效的特徵。

五、結論

本研究提出的 Fusion Model，此模型結合新聞與股價資料集，新聞採用混合注意力機制(HAN)模型，模擬人判讀新聞影響股票重要性，股價方面則是採用長短期記憶神經網路(LSTM)模型，擷取股價時間序列的特徵，有效地學習新聞報導中的訊息順序和股價時間序列，藉此訓練新聞文章與歷史股票交易資料之間的關聯，建構股票漲跌趨勢之模型。由實驗結果可以證實，結合新聞與股價的混合深度模型優於單僅使用新聞或是股價訊息的模型。兩種資訊的結合中，與 HAN、LSTM 模型相比，最佳準確度為 80%，整體平均最高可提升 40%的準確度。相較於 Ziniu Hu[6]等人提出只有新聞資料的 HAN 模型相比，我們提出的 Fusion Model，整體平均高於 37%的準確率。本研究所提出的方法還有能改善的方向，針對股票趨勢預測中，除了本論文結合新聞文本與股價，還可以使用技術指標、情緒，增加深度學習神經網路在股票趨勢預測上的準確度。在處理新聞文章上，不是輸入整篇新聞文章，而是從新聞文章中提取對股票有影響的事件，並且把這些事件輸入到模型中進行訓練，降低多餘新聞文章中不重要的資訊，增加模型預測之精準度。

參考文獻

- [1] Indu Kumar, Kiran Dogra, Chetna Utreja, and Premlata Yadav. A comparative study of supervised machine learning algorithms for stock market trend prediction. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pages 1003–1007. IEEE, 2018.
- [2] Kai Chen, Yi Zhou, and Fangyan Dai. A lstm-based method for stock returns prediction: A case study of china stock market. In 2015 IEEE international conference on big data (big data), pages 2823–2824. IEEE, 2015.
- [3] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In LREC, volume 2014, pages 1170–1175, 2014.
- [4] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Using structured events to predict stock price movement: An empirical investigation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1415–1425, 2014.
- [5] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In Twenty-fourth international joint conference on artificial intelligence, 2015.
- [6] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In Proceedings of the eleventh ACM international conference on web search and data mining, pages 261–269, 2018.
- [7] Yuzheng Zhai, Arthur Hsu, and Saman K Halgamuge. Combining news and technical indicators in daily stock price trends prediction. In International symposium on neural networks, pages 1087–1096. Springer, 2007.
- [8] Xiaodong Li, Pangjing Wu, and Wenpeng Wang. Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong. *Information Processing & Management*, page 102212, 2020.
- [9] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196, 2014.
- [10] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 1480–1489, 2016.
- [11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

arXiv preprint arXiv:1412.6980, 2014.

- [12] 參考網址來源: <https://towardsdatascience.com/stock-prediction-using-recurrent-neural-networks-c03637437578>

基於 BERT 任務模型之低誤報率中文別字偵測模型

Low False Alarm Rate Chinese Misspelling Detection Model Based on BERT Task Model

沈峻毅 Jyun-Yi Shen

張道行 Tao-Hsing Chang

國立高雄科技大學資訊工程系

Department of Computer Science and Information Engineering

National Kaohsiung University of Science and Technology

1106108125@nkust.edu.tw

changth@nkust.edu.tw

摘要

中文別字偵測技術可以應用在教育及出版等許多實務領域。雖然近期許多研究提出了一些能提高效能的模型，但這些模型卻有著誤報率偏高的缺點。在真實的應用中，減少誤報情況的發生是很重要的，因為使用者在操作時，一直出現誤報的情況，會讓使用者體驗不佳，所以如何生成低誤報率且高效率的模型，成為一個要處理的問題。本文採用 BERT 在 Single Sentence Tagging 任務模型來解決中文別字偵測的問題，並配合此模型設計了訓練資料的大量生成方法。實驗顯示本文所提方法對 SIGHAN 2015 測試資料集的誤報率(False Alarm Rate)達到 0.0297。與先前其他低誤報率方法相比，此方法有最低的誤報率以及最高的召回率。

Abstract

Chinese misspelling detection technology can be applied in fields such as education and publishing. This research topic has garnered considerable attention. Recently, although many studies have proposed models that are based on deep learning and that are capable of improving detection accuracy, these models have the disadvantage of high false alarm rates. In real application scenarios, it is important to reduce the occurrence of false alarms because false alarms, while using the system, lead to poor user experience. Therefore, it is important to create a model with low false alarm rate and high efficiency. In this paper, BERT Single Sentence Tagging task model is used to solve the Chinese misspelling detection problem. To work with this model, mass training data generation methods were designed. Experiments showed that the method employed in this study has a false alarm rate of 0.0297 for the

SIGHAN 2015 test set. Compared to other previous methods with low false alarm rates, this method has the lowest false alarm rate and the highest recall rate.

關鍵字：BERT，中文別字偵測，低誤報率

Keywords: BERT, Chinese misspelling detection, low false alarm rate

一、緒論

錯別字問題是一個持續受到討論的重要議題，特別在語言教育上。一般認知的錯別字應分成兩類，一是字形本身是不存在字的錯字，二則是字本身是存在字但被誤用的別字。但現代人寫文件或報告大多都使用電腦和手機的輸入法，所以本文討論聚焦在別字偵測。目前已經有許多研究與方法被提出，然而，雖然近年來的方法在實驗報告中看起來有不錯的效果，例如 Wang, Song, Li, Han, & Zhang (2018)述及其方法的精確率 (precision)、召回率(recall)與 F1 等三項評估指標分別為 0.57、0.70、0.62；FASpell (Hong, Yu, He, Liu, & Liu, 2019) 則提到其三項指標分別為 0.68、0.60、0.64。然而在實際應用中，這些模型仍然遇到很大的挑戰。其中最關鍵的因素是在實際應用時必須將誤報率 (False Alarm Rate)儘量壓低。這是因為在大多數情境下別字只是文件中相當少量的發生，若假設某方法將正確字誤判成別字的誤報率達 0.2，那麼即使每 10 句就有一個別字且百分之百被找出，但該方法還是會同時指出 2 個沒有問題的句子有別字，這將造成不大理想的使用者體驗。因此低誤報率是別字自動偵測能否在實務場域被採用的重要因素。

我們認為 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019) 提出的 BERT for Single Sentence Tagging Task (以下簡稱 BSST)架構相當適合用於別字偵測並克服這個問題。以 BERT 模型為基礎的方法近年在許多自然語言處理(NLP)問題上成為好的解決方案，BSST 是其中之一。BSST 原始用途之一是命名實體識別(Name Entity Recognition, NER)。舉例來說，我們的目標是將句子中的字分成 5 類，分別是名詞詞首、名詞詞幹、形容詞詞首、形容詞詞幹和其他字。假設給 BSST 一個句子「今天天氣真好」，那 BSST 的目標就是對照原句順序依序輸出下列結果：

今	天	天	氣	真	好
名詞詞首	名詞詞幹	名詞詞首	名詞詞幹	其他字	形容詞首

由此模式，我們可以套用在別字偵測。當我們輸入一個句子，BSST 的輸出為是別字以及不是別字兩個類別。如此一來 BSST 便可用來偵測別字。

因此，本文的目的是提出一個以 BSST 模型為基礎的別字偵測方法，並針對 BSST 所需要的大量訓練資料提出一個模擬資料生成方法。以下各節的組織如下：在第 2 節我們說明別字偵測常見的一些方法及其原理。第 3 節將略述 BERT 與 BSST 的基本架構。由於訓練 BSST 需要龐大的資料量，但別字的語料庫規模多有限，因此第 4 節說明我們產生訓練 BSST 資料的方法，並於第 5 節說明實驗結果。最後對本文提出方法進行討論。

二、文獻回顧

中文別字偵測一直是受到矚目的研究議題。近年來有許多方法被提出，也有許多歸納與分析這些方法的評論性論文 (Wu, Liu, & Lee, 2013; Yu, Lee, Tseng, & Chen, 2014; Tseng, Lee, Chang, & Chen, 2015)。Chang (1995)提出了早期的中文別字自動偵測的架構，雖然有著誤判率太大、偵測時間長等問題，但也為這領域開啟了研究起點。早期的研究利用混淆字表偵測別字，也就是事先蒐集常見別字，之後為每個別字依照字形、字音、涵義及輸入法等規則建立混淆字表，系統會將句子中的每個字替換為字表中的字，最後再利用模型計算所有修改句的機率。

之後一些研究發現採用字音、字形與語法特徵偵測別字有不錯的效果。Liu, Lai, Tien, Chuang, Wu, & Lee (2011)除了把相同音的字製作成混淆字集，也使用倉頡輸入法來編碼，目的是可以更快速的去計算字形相似度。Chang, Su, & Chen (2012)基於一個假設：有別字的詞在斷詞後被切割成多個單字詞。基於這個假設，此方法偵測到句子有連續的單字詞出現，便進一步以各單字與候選正確字的字音、字形相似度以及字詞頻和詞性組合機率，判斷是否有別字。Wang, Liao, Wu, & Chang (2013)先檢查句子裡出是否有出現混淆字集裡的高頻別字，之後使用 CRF 把句子斷詞，最後再使用 tri-gram 模型進行判斷，判斷完後會把少於 3 個字元的詞都當成是別字的可疑字，並依照字形或字音

去替換這些可疑字，再重複輸入進 tri-gram 模型內計算分數，直到找出最高分數的句子。

後續許多研究多基於前面兩個研究的基本架構進行改良或修正。Chang, Chen, & Zheng (2014) 進一步修正了 Chang et al. (2012)的方法，採用筆畫結構取代部件結構評估字形相似度、並加入了常見單字詞別字的規則式方法。Xiong, Zhang, Zhang, Hou, & Cheng (2015)發表的 HANSpeller 採用先前一些方法的基本原理，提出一個兩階段的架構。第一階段此方法會將會句子送入一個簡單的分類器，把過於明顯不是錯別字的選項給篩選掉。第二階段則會把句子翻譯成英文，接著把翻譯後的英文句輸入Microsoft Web n-gram Service，去計算英文翻譯的 n-gram 分數，依照這個分數去篩選剩下的可疑字。類似做法還有 Chu, & Lin (2015)與 Xie, Huang, Zhang, Hong, Huang, Chen, &Huang (2015)提出的方法。Chu, & Lin 是先將句子斷詞，斷詞後將句子內的字或詞用混淆字集做替換，再使用 Google n-gram Viewer 去計算最有可能有別字的選項。Xie et al. (2015)則是採用 Chang et al. (2012) 的假設，先把句子斷詞，再去判斷是否有連續的單字詞，如果有錯字的話，會有很大的機率被斷成連續的單字詞。之後再透過混淆字集與語料庫做修正，最後使用 bi-gram 和 tri-gram 計算並挑選正確的句子。

Wang, Song, Li, Han, &Zhang (2018)也是採用以字音字形為特徵輸入一個分類器預測別字的類似架構，但由於其採用雙向 LSTM 模型(Bi-LSTM)作為預測器，因此需要大量訓練資料。為此，此方法使用光學字符辨識(Optical Character Recognition, OCR)和自動語音辨識(Automatic Speech Recognition, ASR)在辨識時會將相似字形和字音的字列為候選詞的特性，產生每個字的混淆字集，並透過大量數據集與混淆字集去訓練 Bi-LSTM。其實驗結果的精確率(precision)和召回率(recall)分別達到了 0.54 和0.69。這也顯示需要有足夠貼近真實錯別字的訓練資料是很重要的。

近年來由於深度語意網路的發展，有研究開始採用這類模型辨識別字。例如 FASpell (Hong, Yu, He, Liu, & Liu, 2019) 使用 BERT 的 masked language model，將所有的字都當成可疑字，並逐步把每個可疑字遮蔽，再來預測被遮蔽的字是什麼字。如果預測的結果內沒有原先被遮蔽字的話，那就認定被遮蔽字是別字。SpellGCN (Cheng, Xu, Chen, Jiang, Wang, Chu, & Qi, 2020)也是採用 BERT 的 masked language model，比較不同的是，他們將字音與字形相似度做成graphs，再使用graph convolutional network (GCN)去計算最佳解。Zhang, Huang, Liu, & Li (2020)則將模型分成偵測層與校正層，偵測層使用雙向

GRU 模型 (Bi-GRU)，校正層使用BERT模型；偵測層的輸入是句子的embedding，輸出是代表這個字是否為別字的機率。將這些機率做 soft masking 的計算，計算結果再輸入校正層。校正層會去預測被遮蔽的字，再將預測出來的結果和偵測層輸入的embedding相加，作為每個字的最終特徵。將這些特徵輸入到 softmax 分類器，再由分類器去篩選哪個候選字是最佳解，最後輸出校正句子。

這類模型的優勢在於能使用前後文語意特徵作為判別依據，因為先前方法所採用的別字出現在連續單字詞的假設無法處理有別字的詞仍是一個成詞的問題、也無法處理字音、字形相似度與 n-gram 模型以外的例子，雖然這些模型的 F1值都有不錯的成績，但還是沒有適合的方式去解決誤判率過高的問題。

三、BERT for Single Sentence Tagging Task

BSST 主要是由 BERT 模型的輸出加上一層簡單的線性分類器所組成，如圖 1 所示。該模型一次接受一個句子輸入，句首輸入符號 CLS 以便於 BERT 模型分析。由於每個句子長度不一，故須設定一個最大值 n，如果句長小於 n 時，則會使用 zero padding 方式補滿。BERT 模型輸出的結果會進一步輸入給 n 個線性分類器(Linear classifier)，這些分類器會再依據 BERT 輸出的語意特徵作出對每個字做出是否為別字的決定，以圖 1 來說，輸出 0 為正確字，1 為別字。

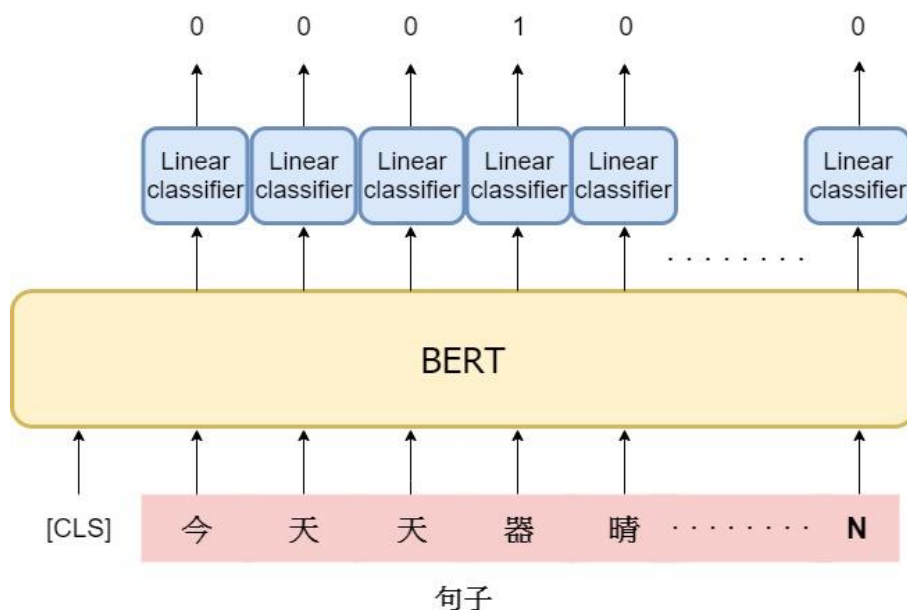


圖 1、BERT for Single Sentence Tagging Task 之架構圖

這個模型的核心 BERT 是由多個 Transformer 的 Encoder (以下簡稱 TE)所組成的，如圖 2 所示，圖中 E1 到 En 為輸入字的 embedding，輸出的 T1 至 Tn 為 E1 至En 在此句語境中的語意向量，n 為輸入句子長度的最大值。輸入與輸出間的中間層使用 12 層的 TE，內部的計算機制則使用 Self-Attention (Vaswani et al., 2017)。該機制有助於整合句子前後詞的語意，對於模型理解整個句子有相當大的幫助。每個字在每一層的 Transformer 要做 12 個 Head 的 Self-Attention，並把 12 個 Head 的 Self- Attention 結果做運算再輸入到下一層 Transformer 中，最後輸出該字的語意向量。

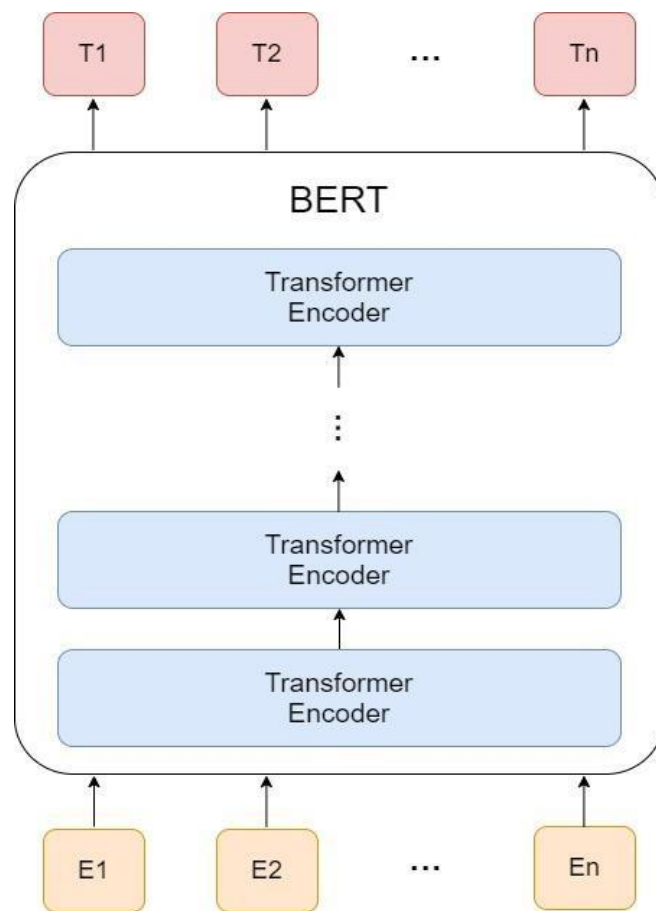


圖 2、BERT 架構圖

四、訓練資料的生成

在第二節中曾提到對於深度學習模型如何產生訓練資料是相當關鍵的問題。訓練資料必須盡可能滿足兩個條件：資料量大且為真實資料。然而，就目前已知的資料集而言，資料量仍有所不足。此外，這些資料集雖然都是真實資料，但由於資料量不足，對真

實環境中別字發生樣態的涵蓋率不足。而訓練資料需要同時有無別字以及有別字的句子集(以下分別稱為正常集與別字集)，因此除了偏誤語料，也需要同性質的正常語料。根據上述需求，本文提出一個訓練資料生成方法，藉由現有大量語料生成模擬真實資料的訓練語料。

首先，我們使用聯合報 2002~2009 年新聞報導為語料來源(以下稱為源語料)。這個來源不但有足夠的語句，而且當時報導內容幾乎完全沒有別字，所以適合作為當作正常集來源。我們從源語料中隨機抽取 40 萬句組成正常集。

對於別字集則利用下列步驟產生。第一、我們建立每個中文字元的候選字集(以下簡稱選字集)，建立的方法是將所有中文字元兩兩計算在字音與字形相似性，計算方法是使用 Chang et al. (2014)的方法。此方法在字音部分利用兩個字的聲母、介母、韻母和聲調的相同與相異計算出一個相似機率。在字形部分利用筆畫結構以最長共同子串列(LCS)為基礎的方法計算相似度。計算出所有的字音和字形分數後，我們以一個加權線性方程式算出相似值，最後針對每個字蒐集與該字最相似的前 k 個字形成該字的選字集。第二、從源語料另外隨機抽取與正常集不重複的 40 萬句。第三，對於每個別字集的句子，都根據中研院平衡語料庫統計的字頻表作為機率計算依據挑選句中將被變更的字。但每個字被挑中的次數有限制，若某字的別字句數量已達上限，則不在被挑選的範圍內。這個限制是確保別字集不會被少數高頻字的別字所涵蓋。第四、從被挑選字的選字集中隨機挑選一個字替換原先的字模擬成錯別字。

五、實驗

本文將以 SIGHAN 2015 (Tseng, Lee, Chang, & Chen, 2015)資料集作為評估本文所提方法的測試集。此測試集共有 1100 句，其中正常句與別字句各有 550 句。由於我們的模型暫不處理長度超過 60 字的句子，因此經排除過長句子後共有正常句有 538 句，別字句有 534 句。

實作 BSST 部分我們則是依據 Wolf et al. (2019)的 HuggingFace's Transformers 所設計的 BERT for Token Classification 開源碼。測試各模型與模式的評估指標與 SIGHAN 2015 相同，各指標計算方式及使用符號說明如下。

誤報率(False-Alarm Rate)： $FP / (FP+TN)$

正確率(Accuracy)： $(TP+TN) / (TP+FP+TN+FN)$

精確率(Precision)： $TP / (TP+FP)$

召回率(Recall)： $TP / (TP+FN)$

F1： $2 * Precision * Recall / (Precision + Recall)$ 。

其中

TP：所有被辨識為有別字的別字句數量。

TN：所有被辨識為正常的正常句數量。

FP：沒有別字卻被辨識為有別字的正常句數量。

FN：有別字卻被辨識為正常句的別字句數量。

表 1 比較本文所提方法與其他誤報率(False-Alarm Rate)低於 0.1 的先前方法的效能。先前方法包括 NTOU (Chu, & Lin, 2015)以及 NCTU+NTUT (Wang, & Liao, 2015)的兩個 Run，這些方法的數據引自 SIGHAN 15 對測試集的實驗結果(Tseng, Lee, Chang, & Chen, 2015)，由於高誤報率的方法在應用上的限制，在此不列入比較。由表 1 可知，本文所提方法比先前誤報率最低方法的誤報率還低 42%，但召回率與 F1 比先前方法最佳者還高 12%。這顯示本文所提方法在降低誤報率同時也能提升別字的偵測率，更接近真實應用需求。

表 1、誤報率低於 0.1 之模型的評估比較

Models	誤報率	正確率	精確率	召回率	F1
本文所提方法	0.0297	0.6446	0.9135	0.3165	0.4701
NTOU	0.0909	0.5445	0.6644	0.1800	0.2833
NCTU+NTUT-Run1	0.0509	0.6055	0.8372	0.2618	0.3989
NCTU+NTUT-Run2	0.0655	0.6091	0.8125	0.2836	0.4205

表 2 說明本文所提方法在不同訓練資料產生模式下造成的效能差異。一個訓練資料只有別字集而沒有正常集(以下簡稱無正常集)；另一個是別字集的句子在選定別字發生位置而挑選替換字時，不是從選字集中挑選，而是採用隨機選擇任一中文字元(以下簡稱無選字集)。由表 2 可知，無正常集模式的誤報率大幅提高。我們認為是因為該模型並未有學習正常句子的經驗，以至於完全沒有錯誤的句子中若有些微語意不一致，

就會被判斷成別字句。而無選字集的模式表現更差，我們認為原因是因為隨機挑選中文字元的方式和真實的情況是有相當大的落差，無選字集模式替換的別字有很高的可能性並不會出現在句子中的那個位置，對模型而言即使有很多的訓練資料也很難擷取語意不一致的正確模式。而選字集有效限縮別字選擇範圍，可以讓模型比較容易捕捉發生語意不一致的模式。

表 2: 訓練資料產生方法差異之效能比較

訓練集產生模式	誤報率	正確率	精確率	召回率	F1
本文所提方法	0.0297	0.6446	0.9135	0.3165	0.4701
無正常集	0.6022	0.4534	0.4564	0.5094	0.4814
無選字集	0.5428	0.3246	0.2589	0.1910	0.2198

六、結論與未來工作

本文提出一個以 BERT 為基礎的別字偵測方法，以及一個產生能訓練此模型的訓練資料的模擬資料生成方法。初步的實驗結果顯示本文所提方法能有效降低誤報率，也能維持整體效能。我們認為能有這樣的效果，除了採用語意面向的辨識方法外，足夠的訓練資料以及選字集的產生是模擬實際別字發生的過程是重要的因素。

雖然誤報率有效的降低，但若召回率可以再提升，則應用的範圍將可以更擴展。我們認為有幾個值得努力的方向。第一，這次的實驗因為受限於系統的設計，所以暫不處理六十字以上的句子，未來也希望可以將此缺點加以改良，往後也會嘗試各種的測試資料，去更進一步的證明此系統的可行性。第二，本次實驗可以初步證明不同的訓練資料產生方式能有效的降低誤報率，但對於BERT是否有助於降低誤報率，我們還需要更完整及更多的實驗才能證明。第三，別字對句子各面向造成的影響進行更多地探索。近期採用深度學習模型的方法多基於別字造成語意不協調性的假設，不過這些方法仍然注意到字音與字形相似性仍是一個重要的特徵，因此在選字集階段採用不同的方法使用這些特徵。我們認為可以嘗試在決策階段融合不同面向的特徵。第四、我們可以進一步檢視目前無法辨識的別字具有的特性。在本文中我們專注在降低誤報率的設計，我們可以嘗試進一步分析未被正確辨識的別字句，探索本文所提方法未知的侷限，進而找出提升召回率的方法。

七、致謝

本研究部分經費由科技部計畫(MOST 107-2511-H-992-001-MY3)支持。

參考文獻

- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Chang, C. H. (1995). A new approach for automatic Chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium* (Vol. 95, pp. 278-283).
- Chang, T. H., Chen, H. C., & Zheng, J. L. (2014, October). Using Chinese Orthography Database to Correct Chinese Misspelling Words With Graphemic Similarity. In *Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014)* (pp. 153-162).
- Chang, T. H., Su, S. Y., & Chen, H. C. (2012, December). Automatic correction for graphemic Chinese misspelled words. In *24th Conference on Computational Linguistics and Speech Processing, ROCLING 2012* (pp. 125-139).
- Chu, W. C., & Lin, C. J. (2015, July). NTOU Chinese spelling check system in SIGHAN-8 bake-off. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing* (pp. 137-143).
- Cheng, X., Xu, W., Chen, K., Jiang, S., Wang, F., Wang, T., Chu, W., & Qi, Y. (2020, July). SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 871-881).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.
- Hong, Y., Yu, X., He, N., Liu, N., & Liu, J. (2019, November). FASpell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)* (pp. 160-169).
- Liu, C. L., Lai, M. H., Tien, K. W., Chuang, Y. H., Wu, S. H., & Lee, C. Y. (2011). Visually and phonologically similar characters in incorrect Chinese words: Analyses,

- identification, and applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(2), 1-39.
- Tseng, Y. H., Lee, L. H., Chang, L. P., & Chen, H. H. (2015, July). Introduction to sighthan 2015 bake-off for chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing* (pp. 32-37).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).
- Wang, Y. R., & Liao, Y. F. (2015, July). Word vector/conditional random field-based Chinese spelling error detection for SIGHAN-2015 evaluation. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing* (pp. 46-49).
- Wang, D., Song, Y., Li, J., Han, J., & Zhang, H. (2018). A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2517-2527).
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistav, P., Rault, T., Louf, L., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., Rush, A. M., & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, arXiv-1910.
- Wu, S. H., Liu, C. L., & Lee, L. H. (2013, October). Chinese spelling check evaluation at SIGHAN Bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing* (pp. 35-42).
- Xie, W., Huang, P., Zhang, X., Hong, K., Huang, Q., Chen, B., & Huang, L. (2015, July). Chinese spelling check system based on n-gram model. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing* (pp. 128-136).
- Xiong, J., Zhang, Q., Zhang, S., Hou, J., & Cheng, X. (2015, June). HANSpeller: a unified framework for Chinese spelling correction. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 1, June 2015- Special Issue on Chinese as a Foreign Language*.
- Yu, L. C., Lee, L. H., Tseng, Y. H., & Chen, H. H. (2014, October). Overview of SIGHAN 2014 bake-off for Chinese spelling check. In *Proceedings of The Third CIPS-SIGHAN*

Joint Conference on Chinese Language Processing (pp. 126-132).

Zhang, S., Huang, H., Liu, J., & Li, H. (2020, July). Spelling Error Correction with Soft-Masked BERT. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 882-890).

中文新聞文本之宣傳手法標記與分析¹

The Analysis and Annotation of Propaganda Techniques in Chinese News Texts

施孟賢 Meng-Hsien Shih
國立中正大學通識教育中心
Center for General Education
National Chung Cheng University
simon.xian@gmail.com

段人鳳 Ren-feng Duann
國立臺東大學通識教育中心
Center for General Education
National Taitung University
rdumann@nttu.edu.tw

鍾曉芳 Siaw-Fong Chung
國立政治大學英國語文學系
Department of English
National Chengchi University
sfchung@nccu.edu.tw

摘要

新聞媒體常在政治新聞文本中運用宣傳手法（propaganda techniques）表達媒體本身之政治立場，企圖影響讀者之立場。目前尚無具宣傳手法標記之中文語料供立場分析，本文以考量可解釋性的方式，人工細部標記中文新聞文本所使用之宣傳手法、並以 Bootstrap 方式擴展標記規模的資料集，最後再人工校正確保標記之品質。本宣傳手法之人工標記資料已公開釋出，可應用於未來機器訓練與學習預測新文本之立場。

Abstract

In political news media, propaganda techniques are often employed to express one's political view, or to influence the audience's stance. Chinese corpora with the annotation of propaganda

¹ 本研究經費由科技部計畫 MOST 109-2811-H-004-503 及 109-2410-H-004-163- 補助，特此誌謝。

techniques are yet to be developed. In this paper, with an explainable approach we annotated the use of propaganda techniques in Chinese political news texts, and enlarged the dataset by bootstrapping using a small set of manually annotated data. We also manually corrected the bootstrapped dataset to increase the data quality. The data manually annotated with propaganda techniques is available online for the application of machine training and learning to predict the stance of new texts.

關鍵詞：情感（立場）分析、語言資源、宣傳手法

Keywords: Sentiment (Stance) Analysis, Language Resource, Propaganda Techniques.

一、緒論

宣傳（propaganda）一詞²，原為宗教用語，於 19 世紀中葉該詞漸漸用於政治領域，且開始帶有貶意（Diggs-Brown, 2011）。Lasswell（1927, p. 9）將宣傳定義為「使用重要的符號來控制意見」。大英百科全書也將之定義為「訊息（事實、論點、謠言、半事實、或謊言）的傳播，以期影響輿論」（Smith, 2020）。報紙做為宣傳的平台由來已久，Riegel（1935, p. 206）即指出，單就新聞選取而言，報紙即遵循一套本質為政治的準則。而美國人普遍認為報紙充斥商業與政治宣傳，因而對報紙內容始終存疑。Riegel 確立了報紙作為一個宣傳場域的事實。另一方面，該文（p. 202）質疑當代對於商業與政治宣傳的檢視方法，大多基於主觀、短暫並且不確定的參考方式。過往研究試圖提出一套準則，俾使研究者檢視宣傳的技巧（如 Lee & Lee, 1939; Weston, 2018）。近年來，隨著科技的發展及語料蒐集技術的演進，計算學者提出新方法，以客觀、一致的準則來偵測報紙或新聞媒體中的宣傳手法（如 Barrón-Cedeño et al., 2019; Da San Martino et al., 2019）。然而，這些研究都是以英文新聞媒體為探討對象，針對中文新聞媒體的宣傳手法辨識的方法則付之闕如。本研究以 Da San Martino et al.（2019）所提出的 18 種宣傳手法為出發點，分析台灣兩個政治立場相反的報紙對同一事件撰寫的

² Diggs-Brown（2011）指出，Propaganda 始於 1622 年由天主教會所創的 Congregatio de propaganda fide，意指由樞機主教組成的行政單位，其任務為在天主教徒國家執行教會事務。1790 年代，該詞的英文擴及非宗教層面，19 世紀中葉該詞漸漸用於政治領域，且開始帶有貶意。本文將 propaganda 翻譯為「宣傳」。

社論，提出在中文報紙中使用的宣傳手法，為第一篇以語料庫探究、提取中文新聞媒體宣傳手法的研究。本研究辨識的方法與詞彙，可做為日後自動偵測的訓練語料。

研究發現，由於報紙社論的特質，Da San Martino et al. (2019) 的諸多方法並未出現在本研究語料中。再者，我們也發現在 Da San Martino et al. 提出的手法之外，中文報紙尚採用了兩個手法：(1) 引用歷史：報紙引用歷史事件或人物，讓讀者連結類比當前事件與歷史事件，意圖影響讀者立場；(2) 預設立場：報紙透過修辭性問句和評價標記 (evaluative marker) 來表明其預設立場，藉以暗示並影響讀者對該事件的看法。就部分中文報紙使用的宣傳手法，本研究也給予更明確的語言上的定義。同時，某些手法出現次類，我們也針對這些次類舉例說明。

本文主要回答下列三個研究問題：

1. 中文政治新聞文本中使用的宣傳手法為何？
2. 中、英文政治新聞採用的宣傳手法有何異同？
3. 以 Bootstrap 方式擴增中文政治新聞宣傳標記之成效如何？

二、文獻回顧

立場偵測 (stance detection) 與情感分析 (sentiment analysis) 是相關但不同的工作，情感分析通常著重於決定文本是正面、負面或中性，立場偵測則在於決定文本傾向於特定對象，亦即「立場」。而相同的立場可能經由正面的語言表達，也可能以負面的語言表達 (Mohammad, Sobhani & Kiritchenko, 2017)，並無法直接藉由文本的情感分析獲知其立場，因此立場偵測有其獨特重要性，且廣泛運用於資訊擷取 (information retrieval) 與文本摘要 (text summarization) 領域。

新聞媒體的立場對大眾有一定的影響力，且常藉由宣傳手法 (propaganda techniques) 表達其立場，進而影響讀者對新聞事件之看法。早期以英文為研究主體的文獻中，大多從文件的層次 (document level) 偵測宣傳文章，甚至僅標記文章出處是否為宣傳來源，而將該來源所有文章視為宣傳。但 Horne et al. (2018) 指出，宣傳的文章來源也會定期發佈客觀的非宣傳文章以提高該來源之可信度。此外，以文件層次標記訓練出來之計算模型較缺乏可解釋性 (explainability)。另一方面，Barrón-Cedeño et al. (2019) 提出一套以新聞寫作風格、可讀性和詞彙豐富性為基礎的模型，

試圖偵測新聞報導是否隱含宣傳目的，並計算其宣傳分數（propaganda score）。該研究結果顯示，這套系統優於過往的偵測系統，但該文也建議相關學者未來可研究以語句的片段層次（fragment level）來辨識新聞文本宣傳手法。有鑑於此，Da San Martino et al. (2019, pp. 5637–5639) 聚焦於語句的片段層次，針對來自宣傳性新聞媒體網站的文章中，可從內部判定、無須外在訊息支援之宣傳文章，列出表一的 18 種宣傳手法，並據此檢視 372 篇英文新聞（約 35 萬詞）。

表一、英文新聞的 18 種宣傳手法（Da San Martino et al., 2019, pp. 5637–5639）

1. Loaded language³	7. Flag-waving	13. Whataboutism
2. Name calling	8. Causal oversimplification	14. Reductio ad Hitlerum
3. Repetition	9. Slogans	15. Red herring
4. Exaggeration	10. Appeal to authority	16. Bandwagon
5. Doubt	11. Black-and-white fallacy	17. Obfuscation
6. Appeal to fear	12. Thought-terminating <i>cliche</i>	18. Straw man

本研究以這 18 種手法為出發點，參考 Da San Martino, Barrón-Cedeño, Wachsmuth, Petrov, & Nakov (to appear) 新標記之 550 篇新聞文本資料⁴，以分析中文報紙社論的宣傳手法。我們發現，Da San Martino et al. (2019) 提出的用於標記英文新聞的 18 種宣傳手法並未完全出現於中文的新聞文本上。由於本研究採用的語料為報紙社論，而社論乃「報社對新聞議題深思熟慮後所形成之見解，此見解以肅穆、說理的筆觸，常態地呈現在每期報刊上，目的在於聲明報社對該議題之立場，並欲勸服讀者採取一致的立場。」（朱, 2003, pp. 18–19）；社論可用來倡導或反對某主張，內容多半採取「理性客觀的口吻、莊重持平的語調」（p. 358），因此，諸如過度簡化因果（causal over simplification）⁵、「那又怎麼說」主義（Whataboutism）⁶、刻意模糊與混淆

³ 表一中粗體字者為在本研究的中文新聞也出現之宣傳手法，另有兩新類別未在此英文列表中，完整定義和實例請參考「三、研究方法」。

⁴ 標記新聞資料取得方法參考此網頁 <http://propaganda.qcri.org/semEval2020-task11>。

⁵ 當一項議題或事件有諸多成因，然而撰稿者卻僅咎責其一，無視其他原因，即為過度簡化因果。

⁶ 參考維基百科，本文將 whataboutism 譯為「『那又怎麼說』主義」，意指撰稿者不直接反駁對手論點，轉而攻擊對手言行不一而削弱對手立場的可信度。

(obfuscation, intentional vagueness, confusion) ⁷、轉移焦點 (red herrings) ⁸...等缺乏理性論證與完整討論，或不合邏輯的手法便不會出現在本研究的語料中。

以中文為主體的相關研究多半聚焦中國的社群媒體檢查制度。Knockel et al. (2015) 探討中國如何執行資訊管控，該研究針對四個中國人常用的社群影音平台，透過逆向工程 (reverse engineering)，辨識客戶端的關鍵詞檢查制度，並歸納出分屬六大主題 (政治、社會、人物、事件、科技、雜項) 共計 17,547 關鍵詞詞表，該研究發現，不同平台對執行審查仍留有各自的彈性，且與「集體行動」和「批評政府」相關的貼文常為審查標的。Arefi et al. (2019) 沿用了 Knockel et al. (2015) 的詞表，採用深度學習、卷積神經網路定位 (CNN localization) 及自然語言處理技術 (NLP techniques)，針對 14 個受審查的類別⁹，探討中國新浪微博受審查和未受審查的貼文與圖片。該研究發現，情感 (sentiment) 為不同主題的唯一共通審查指標，不同主題也受到不同層次的審查，而受審查的貼文僅有少於三小時的網路存活時間。上述研究所使用的關鍵詞，皆為以社群平台貼文主題為分類基準，以宣傳手法作為詞彙/語句片段層次為分類基準的研究仍有待發展。

在標記品質的評估測量上，有別於傳統的分類標記問題，宣傳手法之標記屬於序列標籤 (sequence labeling) 問題，需同時考慮標記不定長度之文字，以及該文字之標籤，另需考量不同標記者之標記文字部份重疊之評測。有鑑於此，Mathet, Widlocher & Métivier (2015, pp. 441-447) 提出 γ 評測量以評估此類序列標籤的標記者間一致性。惟本文做為首篇標記中文宣傳手法之研究，目前僅初步評測標記者間針對同一句的標記是否一致的百分比 (Scott, 1995, p. 323; Artstein & Poesio, 2008, p. 558)，詳細的評測基準請見本文第四章之結果與討論。

三、研究方法

本文選擇的中文報紙為台灣兩份政治立場相反的主流報紙：自由時報與聯合報。自由時報傾向由民主進步黨領導的陣營，在國家認同上主張台灣獨立，因而對中國及

⁷ 撰稿者刻意使用不清楚的或有多重語意的文字，讓讀者有自行解讀與詮釋的空間。

⁸ 原文為紅鯡魚 (red herring)，即撰稿者引入與討論議題無關的題材，以轉移讀者焦點。

⁹ 這 14 個類別為：(1) 薄熙來、(2) 鄧小平、(3) 火災、(4) 死亡/傷害、(5) 劉曉波、(6) 毛澤東、(7) 人民大會、(8) 軍/警、(9) 抗爭、(10) 色情/裸露、(11) 暴雨、(12) 小熊維尼、(13) 習近平、(14) 周克華。

其相關事務採取保留的態度；聯合報則傾向由中國國民黨領導的陣營，認同 1912 年建立的中華民國，對中國及相關議題較為友善。

本研究選出上述兩報有關太陽花學運的社論。選擇社論的原因，乃根據 Smith (2020) 說明，社論可視為隱蔽型 (covert) 的宣傳，即閱聽人對宣傳者身份或來源一無所知。Riegel (1935) 不諱言社論具有宣傳的特質，以往亦有社論作為宣傳的研究 (例如，賴，1965；游，2001)，我們因而將報紙社論視為宣傳的一類。

太陽花學運為一場由大學生主導的社會運動，肇因於當時的執政黨 (中國國民黨) 在立法院企圖迅速通過海峽兩岸服務貿易協定 (簡稱「服貿」)，這項協定被抗議學生與公民團體認為有損台灣現狀並會危及台灣中小企業，他們因而進入立法院，佔領立法院長達 24 天之久 (2014 年 3 月 18 日至 4 月 10 日，Lin & Hsieh, 2017)，創下台灣歷史上立法院被佔領的首例。

表二、本研究從自由時報和聯合報取得太陽花學運相關社論文章的統計

	自由時報	聯合報	總和
文章數	78 (34%)	150 (66%)	228 (100%)
詞數	147,562 (41%)	213,437 (59%)	360,999 (100%)

本研究新聞資料來自台灣的自由時報與聯合報網站，搜尋日期介於 2014 年 3 月 18 日至 2016 年 12 月 31 日之間，含「服貿」或「太陽花」關鍵詞之社論。如表二所示，自由時報有 78 篇 (佔本語料庫文章數的 34%) 共 147,562 詞 (佔本語料庫詞數的 41%)，聯合報有 150 篇 (佔本語料庫文章數的 66%) 共 213,437 詞 (佔本語料庫詞數的 59%)。經過初步標記自由時報與聯合報日期最早的前 10% 社論 (自由時報自 2014 年 3 月 18 日至 2015 年 3 月 20 日共 8 篇 1,073 句，聯合報自 2014 年 3 月 18 日至 2014 年 4 月 5 日共 15 篇 1,825 句) 的宣傳手法之後，我們提出中文新聞的 11 種宣傳手法，定義和舉例如下，並參考此定義和範例進行宣傳手法之標記工作：

1. **Loaded language (LL, 情緒語言)**：以帶有情緒的語言描述某人、某政治群體或事件，可能是單詞、短語或子句。因社論主要以評論時事為主，且具有監督

執政黨的責任，通常負面語言較多。負面語言的例子：「以盧淺的理由宰殺了社會正義」（聯合 2014-03-22）；正面語言的例子：「激勵愈來愈多人民抗爭」（自由 2014-03-26）。

2. **Name calling or labeling (NCL, 貼標籤)**：將宣傳陣營的目標貼上讀者所恐懼、厭惡、不歡迎，或者受讀者喜愛或讚揚的標籤。可正面或負面，以名詞片語呈現，大多是修飾結構 (**modifier-modified**)。可以是人物的標籤，也可以是行為、政策或事件的標籤。負面標籤的例子：「馬卡茸總統」（自由 2014-03-26）；「憲政荒謬劇」（聯合 2015-08-02）；正面標籤的例子：「司法鐵漢」（聯合 2014-03-22）、「太陽花學子」（自由 2014-04-04）。
3. **Appeal to authority (ATA, 訴諸權威)**：引用有名的／有影響力的媒體或人物等專有名詞。例如：「《時代》駐北京特派員指出」（自由 2014-03-26）。
4. **Doubt (DT, 質疑)**：通常以問號結尾，為媒體質疑政治人物或政黨可信度 (**credibility**)，進而影響讀者立場的手法。例如：「哪來『深自反省』？」（自由 2015-03-20）；「難道連自己面臨什麼危機都不明白？」（聯合 2014-04-17）。
5. **Thought-terminating cliché (TTC, 格言論證)**：使用某些訴諸讀者常識的詞彙，讓讀者認為報社論點屬於常識，進而影響其立場。例如：「眾所皆知」（自由 2014-08-15）；「可謂已是在兩岸議題上的常識」（聯合 2014-03-27）。
6. **Flag-waving (FW, 高舉大旗)**：訴諸國民或某群體（種族、性別或政治群體）的價值偏好，以便對讀者宣揚某些想法，文字上彰顯某些（普世）價值與願景。例如：「民意向背」（自由 2015-04-15）、「台灣的出路是由 E C F A 到 T P P」（聯合 2014-03-23）。
7. **Historical allusion (HA, 引用歷史)**：使用歷史事件或人物，引發讀者對當下事件與歷史事件的類比與聯想。語言特徵為「『台（灣）版』+ 歷史事件名稱」或「歷史事件名稱+ 『翻版』」，或使用類比或比較文字「如」、「像」、「相較」，例如「台灣版天安門事件」（自由 2014-03-26）、「仍如三十多年前美麗島事件的翻版」（自由 2014-04-04）、「有人將這場太陽花學運與一九九〇年的野百合學運相較」（聯合 2014-04-05）。
8. **Presupposition (PS, 預設立場)**：利用修辭性問句 (**rhetorical question, RQ**) 和評價標記 (**evaluative marker**) 來傳達報紙預設立場。（1）修辭性問句：湯

(1981) 主張國語的疑問句可分為要求回答的「徵訊問句」和不要求回答的「非徵訊問句」，修辭性問句為「非徵訊問句」的一種，即形式上雖是問句，但說話者實則表示個人的觀點或看法，語言特色為否定疑問句或者是附加問句，例如：「豈不反諷之至？」（聯合 2014-04-29）；（2）評價標記：Bednarek (2006, p. 67) 提出六種評估英文媒體的「核心評價參數」，其中的預期性 (Expectedness) 的評價標記如 *strikingly* 和 *unexpectedly*，即對應到中文新聞文本中「竟(然)」、「居然」這類的評價標記，我們認為，中文新聞使用這些標記來凸顯報社的預設立場，例如：「竟開放了一些不該開放的項目給中國」（自由 2014-03-18）。

9. **Black-and-white fallacy, Dictatorship (BW, 非黑即白)**：引導讀者認為僅存在兩種可能，但其實尚存在其他可能。有以下兩個次類別：（1）非 A 即 B (Black-and-white fallacy, BWF)；（2）A 或非 A (Dictatorship, DS)。例如（1）「學生的要求是畫蛇添足還是無理找碴？」（聯合 2014-03-25）；（2）「這不是私有化是什麼？」（聯合 2014-03-19）。
10. **Appeal to fear (ATF, 訴諸恐懼)**：使用引發讀者恐懼的文字，進而影響讀者對該事件的看法：「統戰」（自由 2014-03-18）；「最厲的鬼魅」（聯合 2014-03-22）
11. **Exaggeration or minimization (誇大或淡化)**：用過度的方法再現事件，使其看起來更好或更壞，進而影響讀者立場。例如：「一夕之間」（聯合 2014-06-09）。

在進行標記的工作之前，我們先以中央研究院 CKIP 的斷詞系統將兩報關於太陽花學運的社論進行斷詞，並以其分行結果為原則，以單行做為標記的單位。以下列句子為例：

在國會全武行的混亂中，這紙影響重大的政治與經濟開門條款，是否如國民黨團宣稱的「視為已審查」？（自由 2014-03-18）

經中研院 CKIP 斷詞及分行的結果如下：

第一行：在國會全武行的混亂中，

第二行：這紙影響重大的政治與經濟開門條款，

第三行：是否如國民黨團宣稱的「視為已審查」？

基於未來計算應用上的考量，我們在判斷標記時，並不考慮前後文，亦不標記跨行之宣傳手法。也就是說，這三行各自形成各自的單位，例如，第一行「在國會全武行的混亂中」，單從該行可判斷「混亂」為情緒語言（Loaded Language），因此我們僅就該行中的「混亂」一詞進行標記，第二、三行亦如此。

標記程序分成兩階段，在第一階段時，兩位作者根據前述表二之宣傳手法定義對同一份文件分別獨立進行標記。在第二階段時，對於兩份標記結果之間的分歧，請第三位作者判斷，並做最終版本的標記。

為了加快標記速度與增加標記規模，初步人工標記完前 10% 的文章後，以 Bootstrap 方式將人工標記做為種子，比對其他 90% 的文章內文是否含有與人工標記相同之內容，進而擴展資料規模，最後再以人工檢核方式提高標記的品質。在此過程中，我們也發現有特定類別的標記 Bootstrap 的成效較佳（或者單字詞可能因斷詞錯誤而不適合做為 Bootstrap 的種子），可列為未來機器學習訓練特徵的考量之一。

四、結果與討論

本研究總共標記了 2,312 句自由時報太陽花學運社論（斷句標準乃根據前述之中研院 CKIP 斷詞系統），以及 2,413 句聯合報相關社論之宣傳手法。由於本標記工作較傳統分類問題複雜，第一階段在自由時報前 8 篇共 1,073 句中，第一位標記者共標記了 449 次宣傳手法，第二位標記者則有 86 個宣傳手法標記；聯合報前 15 篇共 1,825 句中，第一位標記者有 577 個宣傳手法標記，第二位標記者有 104 個宣傳標記。第一階段兩位標記者間的一致性（inter-annotator agreement）僅達 63.3%¹⁰。

第二階段請第三位標記者針對不一致處做最終版本之標記，標記者間一致性以最嚴謹的三位標記者的宣傳手法標記皆一致的百分比（unanimous percentage）進行評估，則分別為自由時報的 14.9%（共有 67 個三位標記者皆一致的宣傳手法標記）以及聯合

¹⁰ 以第一標記者為基準，與第二標記者對同一句標記結果是否一致的百分比。新聞文本大部分的句子並未使用宣傳手法，本研究僅針對含有宣傳手法的語句進行統計分析。

報的 9.9%（57 個完全一致的宣傳手法標記）。最後再進行第三階段以 Bootstrap 方式擴增資料標記規模，並以人工校正。

人工校正的筆數如表三所示，以 Bootstrap 方式擴展但遺漏處，再以人工新增筆數分別為 266 筆（自由時報）以及 328 筆（聯合報）；有 77 筆（自由時報）與 105 筆（聯合報）Bootstrap 不盡正確處再以人工更正，其中含 Bootstrap 誤標之 59 筆（自由時報）與 66 筆（聯合報）。第一階段的標記最後成為最終版本之百分比達 85.2%（自由時報）以及 82.1%（聯合報）。

表三、人工校正兩報社論之宣傳手法標記之統計

校正方式	自由時報	聯合報
新增 (A)	266 (11.5%)	328 (13.2%)
更正 (Y)	77 (3.3%)	105 (4.4%)
正確	1,969 (85.2%)	1,980 (82.1%)
總計	2,312 (100.0%)	2,413 (100.0%)

表四、表五分別呈現兩報各項宣傳手法之 Bootstrap 正確率（與人工檢核比較），其中更正筆數表示 Bootstrap 標記錯誤、經人工更正成正確的標記筆數；新增筆數為 Bootstrap 未比對出之標記，經人工新增之標記筆數；其他未經變動的筆數皆為正確筆數，總筆數則代表最終標記版本中，該宣傳手法出現的筆數，正確率之算法則為正確筆數除以總筆數。表格中第 12 類宣傳手法標記為其他 (X) 者，表示以 Bootstrap 方式標記錯誤，該句並未含有任何宣傳手法。

以表四自由時報社論中情緒語言宣傳手法的標記為例，有 15 筆為 Bootstrap 方式錯標為其他手法，後來經人工更正為情緒語言標記；另 Bootstrap 遺漏 199 筆情緒語言標記，而在最後人工檢核階段，經人工新增為情緒語言標記；剩下 1,116 筆為以 Bootstrap 方式比對後完全正確，未再經人工變動的正確筆數；最後自由時報社論中，總共標記了 1,330 筆情緒語言，正確率為 $1,116 / 1,330 = 83.9\%$ 。

表四、自由時報社論各宣傳手法標記之 Bootstrap 正確率統計（依照總筆數排序）

宣傳手法	更正筆數	新增筆數	正確筆數	總筆數	正確率
1 情緒語言	15	199	1,116	1,330	83.9%
2 高舉大旗	0	16	416	432	96.3%
3 貼標籤	2	39	90	131	68.7%
4 非黑即白	0	3	127	130	97.7%
5 訴諸恐懼	0	4	123	127	96.9%
6 預設立場	0	4	42	46	91.3%
7 訴諸權威	1	0	34	35	97.1%
8 引用歷史	0	0	12	12	100.0%
9 格言論證	0	0	6	6	100.0%
10 質疑	0	0	3	3	100.0%
11 誇大或淡化	0	1	0	1	0.0%
12 其他 (X)	59	0	0	59	-
總和	77	266	1,969	2,312	85.2%

表五、聯合報社論各宣傳手法標記之 Bootstrap 正確率統計（依照總筆數排序）

宣傳手法	更正筆數	新增筆數	正確筆數	總筆數	正確率
1 情緒語言	13	278	1,449	1,740	83.3%
2 非黑即白	15	0	257	272	94.5%
3 預設立場	0	9	146	155	94.2%
4 貼標籤	10	23	26	59	44.1%
5 訴諸恐懼	0	4	31	35	88.6%
6 高舉大旗	1	3	27	31	87.1%
7 誇大或淡化	0	0	17	17	100.0%
8 格言論證	0	4	9	13	69.2%
9 引用歷史	0	1	10	11	90.9%
10 質疑	0	3	7	10	70.0%
11 訴諸權威	0	3	1	4	25.0%
12 其他 (X)	66	0	0	66	-
總和	105	328	1,980	2,413	82.1%

從表四和表五中可見，兩報社論以 **Bootstrap** 比對方式標記的平均正確率皆達八成以上（自由時報為 85.2%，聯合報為 82.1%），惟其他未經 **Bootstrap** 的句子尚有可能含有遺漏之標記。

我們發現常見的宣傳手法之比例如表六所述，相較於 **Da San Martino et al. (2019, p. 5641)** 在英文新聞的標記結果，中文報紙和英文報紙的「情緒語言」的手法出現頻率都是最高，然而其餘手法卻呈現不同的面貌：排除「其他」這個類別，英文報紙第二、第三和第四高比例的手法分別為「貼標籤」、「誇大或淡化」和「質疑」，而中文報紙第二、第三和第四高比例的手法則是「高舉大旗」和「非黑即白」，以及未出現在英文新聞中的「預設立場」。

表六、兩中文新聞媒體社論之常見宣傳手法與比較（依照中文的百分比排序）

宣傳手法	自由時報	聯合報	總標記數	百分比	Da San Martino et al.	百分比
1. 情緒語言	1,330	1,740	3,070	65.0%	2,547	34.1%
2. 高舉大旗	432	31	463	9.8%	330	4.4%
3. 非黑即白	130	272	402	8.5%	134	1.8%
4. 預設立場	46	155	201	4.3%	-	-
5. 貼標籤	131	59	190	4.0%	1,294	17.3%
6. 訴諸恐懼	127	35	162	3.4%	367	4.9%
7. 訴諸權威	35	4	39	0.8%	169	2.3%
8. 引用歷史	12	11	23	0.5%	-	-
9. 格言論證	6	13	19	0.4%	95	1.3%
10. 誇大或淡化	1	17	18	0.4%	571	7.6%
11. 質疑	3	10	13	0.3%	562	7.5%
12. 其他 (X)	59	66	125	2.6%	1,411	18.9%
總數	2,312	2,413	4,725	100.0%	7,480	100.0%

五、結論

本文已回答了前述三個研究問題，首先我們針對中文的政治新聞提出 11 種宣傳手法，其中「引用歷史」和「預設立場」是有別於英文、在中文首見試圖影響讀者立場之宣傳手法。而根據標記自由時報與聯合報共 228 篇（360,999 詞）太陽花學運相關社論的結果統計，中英文新聞除了皆最常使用「情緒語言」手法之外，其他宣傳手法呈現完全不同的分佈。惟本研究之中文宣傳型新聞取材僅限於社論之文類，與 **Da San Martino et al. (2019)** 所採用宣傳型新聞網站之新聞文類不盡相同，因而此處中、英統計數據之比較會有所限制，未來在文類選擇上，我們將選取更高可比性的文類。最後，

以 Bootstrap 比對方式擴增標記之正確率兩報皆達八成左右，但尚有其他未經 Bootstrap 比對出之句子，也可能使用宣傳手法。我們未來將繼續檢視本宣傳手法標記資料中，是否尚有其他未比對出之標記，並進行評估。

由於宣傳手法標記之複雜度，本文除了基本的標記一致性之百分比，僅將人工標記之宣傳手法資料用 Bootstrap 方式擴展資料集，最後進行人工檢核。關於運用機器計算模型於此類型標記之可能性，將列為未來繼續研究之議題。

本文人工標記之宣傳手法語料，已開放用於偵測宣傳手法以及相關研究¹¹，亦可用於媒體中立程度之評估。本研究未來將延續目前的宣傳資料標記，並擷取其中常見之語言特徵（如關鍵詞頻或句法結構），自動判讀其他未標記新聞，並偵測其所採用之宣傳手法。

致謝

本文特別感謝國立政治大學資訊科學系江玥慧老師幫助分析英文新聞之宣傳手法。

參考文獻

- [1] Diggs-Brown, B. (2011). *Strategic public relations: An audience-focused approach*. Boston: Wadsworth Cengage.
- [2] Lasswell, H. D. (1927). *Propaganda Technique in the World War*. Michigan: Peter Smith.
- [3] Smith, B. L. (2020). Propaganda. In *Britannica*. Retrieved from www.britannica.com/topic/propaganda
- [4] Riegel, O. W. (1935). Propaganda and the Press. *The Annals of the American Academy of Political and Social Science*, 179, 201–210.
- [5] Lee, A. M., & Lee, E. B. (1939). *The Fine Art of Propaganda*. New York: The Institute for Propaganda Analysis.
- [6] Weston, A. (2018). *A Rulebook for Arguments* (5th ed.). Cambridge: Hackett Publishing.
- [7] Barrón-Cedeño, A., Jaradat, I., Da San Martino, G. & Nakov, P. (2019). Propy: Organizing the News Based on Their Propagandistic Content. *Information Processing and Management* 56, 1849-1864.

¹¹本標記資料集可於此網址下載 <http://simonshih.tw/propaganda/{LT,UDN}-bootstrap-checked.tsv>。

- [8] Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-Grained Analysis of Propaganda in News Articles. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5636–5646.
- [9] Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology*, 17(3), 26:1-23.
- [10] Horne, B. D., Dron, W., Khedr, S., & Adali, S. (2018). Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, 518–527.
- [11] Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., & Nakov, P. (to appear). SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. *Proceedings of the 14th International Workshop on Semantic Evaluation*.
- [12] 朱灼文 (2003) 。《社論的論證結構分析》。政治大學新聞研究所碩士論文。
- [13] Knockel, J., Crete-Nishihata, M., Ng, J. Q., Senft, A., & Crandall, J. R. (2015). Every Rose Has Its Thorn: Censorship and Surveillance on Social Video Platforms in China. In *Proceedings of the 5th USENIX Workshop on Free and Open Communications on the Internet*, 1–10.
- [14] Arefi, M. N., Pandi, R., Crandall, J. R., Tschantz, M. C., Fu, K.-W., Shi, D. Q., & Sha, M. (2019). Assessing Post Deletion in Sina Weibo: Multi-modal Classification of Hot Topics. *Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 1–9.
- [15] Mathet, Y., Widlocher, A., & Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3), 437–479.
- [16] Scott, W. (1955). Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3), 321–325.
- [17] Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4), 555–596.
- [18] 賴賢 (1965) 。《人民日報社論的反美宣傳分析》。政治大學新聞研究所碩士論文。
- [19] 游承季 (2001) 。《大陸經濟改革的內容和宣傳策略研究(1992-1996)--以「人民日報」社論內容為例》。政治大學新聞研究所碩士論文。

- [20] Lin, Y., & Hsieh, J. F. (2017). Change and Continuity in Taiwan's Public Opinion on the Cross-Strait Economic Interactions. *Journal of Asian and African Studies*, 52(8), 1103–1116.
- [21] 湯廷池 (1981) 。國語疑問句的研究 。《師大學報》 26:1-59 。
- [22] Bednarek, M. (2006). *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. New York: Continuum.

探究語言模型合併策略應用於中英文語碼轉換語音辨識

Exploring Disparate Language Model Combination Strategies for Mandarin-English Code-Switching ASR

林韋廷 Wei-Ting Lin, 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

{60347014S, berlin}@ntnu.edu.tw

摘要

語碼轉換 (Code-Switching, CS) 在多語言社會中是一種常見的現象；例如，在台灣的官方語言是中文，但居民們日常對話時而會夾雜一些英文詞彙、片語或語句。語碼轉換語音的轉寫，在自動語言辨識 (Automatic Speech Recognition, ASR) 上仍被視為一個重要且具有挑戰性的任務。而為了提升 CS ASR 效能，改進其語言模型是最直接且有效的方法之一。有鑒於此，我們提出多種不同階段的語言模型合併策略以用於中英文語碼轉換自動語言辨識。在本篇論文的實驗設定中，會有兩種中英文 CS 語言模型和一種中文的單語言模型，其中 CS 語言模型使用的訓練資料與測試集同一領域 (Domain)，而單語言模型是用大量一般中文語料訓練而成。我們透過多種不同階段的語言模型合併策略以探究 ASR 是否能結合不同的語言模型其各自的優勢以在不同任務上都有好的表現。在本篇論文中三種語言模型合併策略，分別為 *N*-gram 語言模型合併、解碼圖 (Decoding Graph) 合併和詞圖 (Word Lattice) 合併。經由一系列在企業應用領域的多種語料之實驗結證實，透過語言模型的合併的確能讓 CS ASR 對不同的測試集都有好的表現。

關鍵詞：語碼轉換、語言模型、語音辨識、解碼圖、詞圖

Abstract

Code-switching (CS) speech is a common language phenomenon in multilingual societies. For example, the official language in Taiwan is Mandarin Chinese, but the daily conversations of the ordinary populace are often mingled with English words, phrases or sentences. It is

generally agreed that transcription of CS speech remains an important challenge for the current development of automatic speech recognition (ASR). One of the straightforward and feasible ways to promote the efficacy of CS ASR is to improve the language model (LM) involved in ASR. Given these observations, we put forward disparate strategies that conduct combination of various language models at different stages of the ASR process. Our experimental configuration consists of two CS (i.e., mixing of Mandarin Chinese and English) language models and one monolingual (i.e. Mandarin Chinese) language models, where the two CS language models are domain-specific and the monolingual language model is trained on a general text collection. Through the language model combination at different stages of the ASR process, we purport to know if the ASR system could integrate the strengths of various language models to achieve improved performance across different tasks. More specifically, three strategies for combining language models are investigated, namely simple N -gram language model combination, decoding graph combination and word lattice combination. A series of ASR experiments conduct on CS speech corpora complied from different industrial application scenarios have confirm the utility of the aforementioned LM combination strategies.

Keywords: code-switching, language model, automatic speech recognition, decoding graph, word lattice

一、緒論

當在對話中使用兩種以上的語言時，這種現象稱為語碼轉換，根據語言切換情形又可細分為兩種類型：句子間的語碼轉換（inter-sentential CS）和句子內的語碼轉換（intra-sentential CS）。其中又以 intra-sentential CS 的語音辨識任務較為困難，因為其語言切換的情形更多種，更難訓練出一個好的 CS ASR。

語碼轉換有幾種方法，可以簡單地分為四個方面：語言識別（Language identification, LID）、資料增強（data augmentation）、模型調適（model adaptation）和模型改進（model improvement）。

首先，最常見的方法是用 LID 標記每個句子或每個單詞，然後分別使用每個單語言 ASR 系統進行語音識別。在 [1-7] 中提出了使用 LID 的相關任務。但使用這種方法的缺點是若是前端的語言辨識錯誤，後端的語音辨識就會錯誤，產生錯誤傳導的問題。

CS 訓練資料的缺乏也是訓練 ASR 模型的瓶頸之一。因此，資料增強在 CS ASR 中

也起到重要作用。資料增強包括聲學 (acoustic) 和文本 (textual) 資料增強，透過資料增強能訓練更加可靠的聲學和語言模型。在聲學資料增強方面，音檔可另外混和噪音 (noise) 或對其進行速度擾動 (speech perturbation) [8]，以及針對大量未標記資料的自動標記方法[9, 10] 或文本轉語音 (text-to-speech, TTS) 技術[11-13]產生更多的訓練資料被應用於聲學模型訓練。在文本資料增強方面，CS 語句可透過句子生成技術[9, 10, 14-17]產生而被用於語言模型訓練。

與資源較少的 CS 資料相比，單語言訓練資料更多。要如何利用大量的單語訓練數據來改善 CS ASR，常用的方法為使用 CS 資料來調適預訓練模型 (pre-trained model)，讓模型保有原本預訓練模型的效能外，也能有好的 CS ASR 效果。因此，以前有一些任務採用遷移式學習[5, 18-23]以利用大量的單語言資料來彌補 CS ASR 的資料稀疏性問題。

除了模型調適之外，還有一些任務是改進模型結構[4, 7, 24, 25]，讓模型可以學習處理多語言 ASR。此外，還引入了多任務學習[3-6]，使模型能同時學習 LID 和 CS ASR。期望 LID 資訊可以幫助 CS ASR 的提升。

CS ASR 還有其他較新穎的方法。在[26]中，他們提出了一種只使用單語言資料來訓練端到端 (End-to-end, E2E) CS ASR 的方法。對不同語言的輸出向量加上限制，讓每種語言的輸出向量的分布相近以達到語言轉換的效果。在[27]中也使用相同方法來訓練 CS 語言模型。在[28, 29]中，他們提出了一種用多解碼圖 (Multi-graph)進行解碼 (decode) 的方法應用於 CS ASR。其中，多解碼圖為多語言解碼圖和單語言解碼圖結合而成。此種解碼方法可讓每個單語言或多語言語音辨識任務有平行且獨立的搜索空間，以更有效地使用每種語言的文本資源。

在本篇論文也將用到論文[28]的方法，和其他種語言模型的合併方法作比較。採用語言模型的合併技術其目的為結合 CS 語言模型和單語言模型各自的優勢以在不同任務上都能有好的表現。在接下來的章節二將介紹聲學模型的模型架構及方法，章節三將介紹三種不同階段的語言模型的合併方法，章節四則會簡述實驗設定及對實驗結果進行探討與分析。

二、聲學模型

(一) TDNN-F

時延神經網路 (Time Delay Neural Network, TDNN) 最早用於音素辨識[30]。因為難

以對語音訊號的時間定位有精確的標記，所以 TDNN 有時移不變性的特性，在語音辨識上會與時間位置無關。另外，TDNN 在建模時也會考慮上下文關係，每層隱藏層會接收到前層不同時間的輸出，舉例來說，若時間延遲 (time delay) 為 2，那就會考慮連續 3 個音框 (frame) 的特徵。藉由這種特性，讓 TDNN 可以表現出語音在時間上的連續關係，也可以考慮特徵序列的長時間相關性。

在論文[31]將 TDNN 應用於語音辨識，和原始的 TDNN 不同，多引進了子採樣 (subsampling)，只保留神經網路的部分連接 (connection)，因此降低計算量，加快訓練速度，也沒影響到原始的模型效能。在[31]實驗中也證明其訓練速度比深度神經網路(Deep Neural Network, DNN)、遞歸神經網路(Recurrent Neural Network, RNN)快，效能也相對提升。

TDNN-F[32]之後也被提出，和 TDNN 的差別主要有兩點，第一點就是將原來的權重矩陣分解成兩個矩陣，而且第二個矩陣需為半正交矩陣(semi-orthogonal matrix)。如此一來，不僅降低參數量，加快計算速度，也能保持相同的建模能力。第二點則是多了跳層連接(skip connections)，將前一層的輸出和當前層的輸出相加當作下一層的輸入，讓模型架構可以架得更深，且避免有梯度消失的問題。

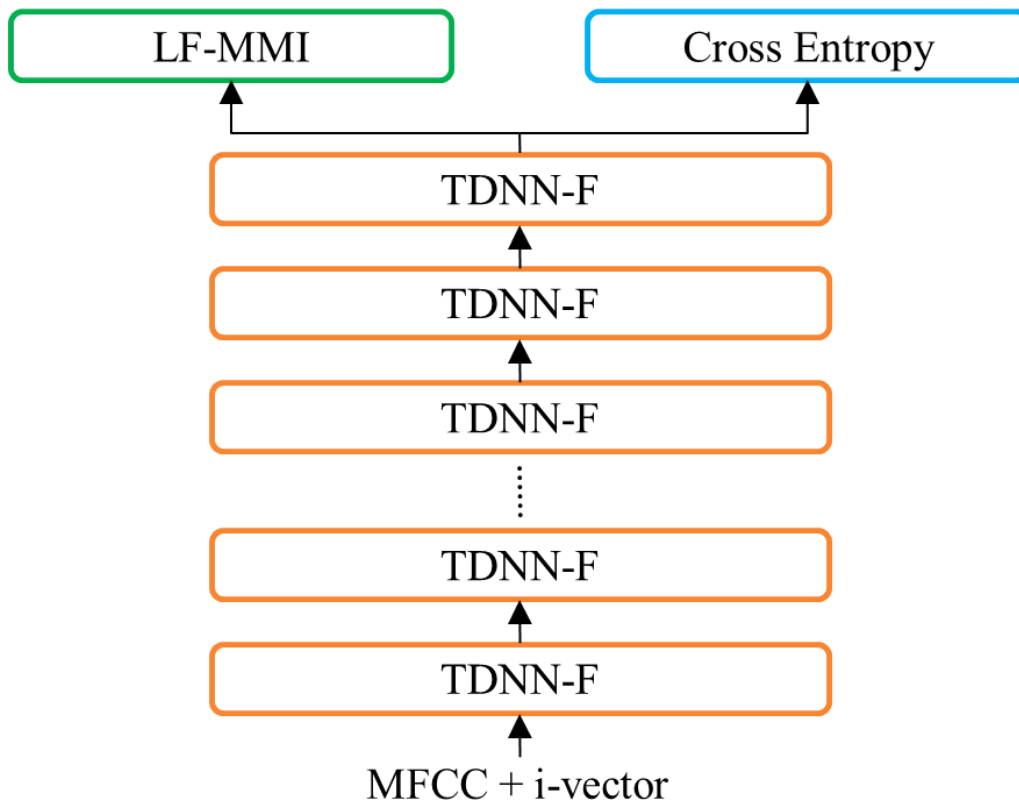
(二) LF-MMI

由於進行鑑別式訓練會提升聲學模型的效能，所以在語音辨識上除了以交叉熵 (Cross entropy)作為損失函數訓練模型外，也會加上鑑別式訓練。進行鑑別式訓練前需要先進行交叉熵訓練，配合語言模型以產生 lattice，然而產生 lattice 是一個解碼的過程，會耗費不小的時間和空間的複雜度。

之後就有論文[33]提出了 Lattice-free Maximum Mutual Information (LF-MMI) 的方法，以音素或狀態 (state) 取代詞 (word) 作為語言模型的單元，使產生 lattice 的計算可以在 GPU 上進行，除了降低空間複雜度，也加快了 MMI 的訓練速度。

(三) Chain model

本篇論文實作的語音辨識工具為 Kaldi[34]，在 Kaldi 中的 chain model 使用了 LF-MMI 的鑑別式訓練，且另外加入了一些技巧使模型訓練更穩定、快速，例如：將隱藏式馬可夫模型 (Hidden Markov Model, HMM) 從三狀態改為單狀態的 HMM；使用音素作為語言模型的單元；加入交叉熵正規化 (cross entropy normalization) 進行多任務學習



圖一、TDNN-F 聲學模型架構

(multi-task learning)。其模型架構圖如圖一。

三、語言模型合併

(一) N -gram 語言模型合併

N -gram 是一種統計語言模型。假設一段 M 個詞組成的句子其機率為 $P(w_1, w_2, \dots, w_M)$ ，根據連鎖律 (chain rule) 可展開成 $\prod_{i=1}^M P(w_i | w_{i-1}, \dots, w_1)$ ，因為第 i 個字需考慮到前 $i - 1$ 個字，若遇到較長的句子時，計算量會變大，所以會根據 $n - 1$ 階馬可夫假設 ($n - 1$ order Markov assumption) 簡化，只需考慮前 n 個字，公式如下：

$$P(w_1, w_2, \dots, w_M) \approx \prod_{i=1}^M P(w_i | w_{i-n+1}, \dots, w_{i-1}) \quad (1)$$

不同的 n -gram 語言模型進行插值合併：

$$LM_{fusion} = \lambda LM_1 + (1 - \lambda) LM_2 \quad (2)$$

其中 LM_* 為 n -gram 的機率， $n = 3$ ，即 trigram 語言模型。

(二) Graph 合併 (WFST 合併)

在 Kaldi 工具會使用 WFST 表示詞序列 (word sequences) 對應的 HMM state。在解碼時即可通過聲學模型計算出的 HMM 發射機率，搜尋出最佳路徑得出解碼結果。在 Kaldi 工具中以 HCLG[35]當作搜尋空間，其可分為四個部分：(1) G：從語言模型抽取的詞序列資訊；(2) L：存有每個詞對應的所有音素的發音辭典；(3) C：音素的上下文關係；(4) H：HMM 的拓撲結構。融合成 HCLG 的過程中都會進行確定化 (determinization) 和最小化 (minimization)：

$$HCLG = \min(\det(H \circ \min(\det(C \circ \min(\det(L \circ G)))))) \quad (3)$$

其中 \circ 為 composition， \det 為 determinization， \min 為 minimization。

如論文[28]，提出在 decode 端利用多個 graph 結合的 multi-graph 進行解碼。將不同的 graph 用聯集 (union) 結合成一個 graph，讓各個 graph 都有各自的搜索空間 (searching space)，而不互相影響。

(三) Lattice 合併

Lattice 合併[36-38] 是將不同系統解碼後產生的 lattice，用不同的權重和損失函數進行合併：

$$W^* = \operatorname{argmin}_W \frac{1}{N} \sum_{n=1}^N \sum_{W'} \lambda_n P_n(W'|\mathcal{O}) L(W, W') \quad (3)$$

其中 $P_n(W'|\mathcal{O})$ 為第 n 個 lattice 的後驗機率， λ_n 為第 n 個 lattice 的結合權重， $L(W, W')$ 為詞序列 W 和 W' 間的 Levenshtein 編輯距離 (Levenshtein edit distance)。 W^* 為對各個系統的 $\sum_{W'} P(W'|\mathcal{O}) L(W, W')$ 的平均透過最小化貝式決策風險 (Minimum Bayes Risk) 估計其最小值。

四、實驗設定與結果

(一) 資料集

本篇論文使用的語料為中英文混合會議語音資料，共有 230 小時，是國內某企業會議時錄製的語料庫。由於錄製內容都是實際對話狀況，所以會出現不一樣的說話腔調、頓點、速度等等。對話中除了一般對話外，也會出現一些專有名詞，甚至會突然出現英文專有名詞或日常用語。這語料庫因為沒經過特別設計，而且出現 CS 的問題 (包含 inter-sentential CS 和 intra-sentential CS)，所以挑戰性較大。

除了原本的訓練資料，另外加上了台灣的中文語料 Formosa 資料集和 YouTube 上收集的語料，除了擴增訓練集，也增加其豐富性。其中 Formosa 語料庫為從廣播、電視、開放課程等收集而來的真實的台灣自發性語音 (spontaneous speech) 中文語料，我們選了其中的 NER-Trs-Vol1、NER-Trs-Vol2 和 NER-Trs-Vol3 資料集和原有訓練集合併。另外也從 YouTube 上的開放課程或演講收集語料，其語料使用的語言為以中文為主，英文為輔。因為我們會把訓練資料中過長或過短句刪除，所以整合出的訓練集共為 635 小時。

而實驗中測試集有三種：(1) 會議的某一場錄音，是以中文為主的中英文 intra-sentential CS，共有 1 小時，資料集被命名為會議錄音測試集。(2) 短句測試集，內容為一些較少見的專有名詞，包含中文、英文和 CS 的專有名詞，共有 3 小時。(3) 長句測試集，由 33 位語者利用手機或平板錄製而成。100 句皆為經過設計過的語句，語句為

以中文為主的中英文 intra-sentential CS，共有 6 小時。

另外實驗中會用三種不同資料集來訓練 N-gram 語言模型，包括：(1) Meeting LM，用包含於 230 小時訓練集中的一部分會議錄音的文本資料訓練而成的語言模型。(2) General LM，用 Chinese Gigaword 資料集中的大部分繁體語料訓練而成的語言模型。(3) Keyword LM，用短句測試集的文本資料訓練而成的語言模型。其中，Meeting LM 和 Keyword LM 為 CS 語言模型，General LM 為中文單語言模型。其訓練資料細節如表一。

表一、語言模型的訓練資料細節

	詞彙數	字數
Meeting LM	551,141	1,605,545
General LM	627,819,651	1,534,226,867
Keyword LM	3,043	16,184

選擇用短句測試集的文本資料訓練一個語言模型，是為了實驗若結合和測試資料集同領域 (domain) 的語言模型，是否能在提升短句測試集辨識正確率時，也不會降低在其他測試集的辨識正確率。

(二) 實驗設定

在聲學特徵方面，會對音檔抽取 40 維的梅爾頻率倒譜系數 (Mel-Frequency Cepstral Coefficients, MFCC) 和 3 維的音調 (pitch)，並另外加上 100 維的 i-vector。

關於 TDNN-F 的模型訓練，參照了 Kaldi 的腳本 (script) 進行訓練。模型包含了 17 層 1536 維的 TDNN-F，每層的矩陣分解瓶頸 (bottleneck) 皆為 160 維。

實驗結果的評估方法採用混和錯誤率 (Mixed error rate, MER)，即英文採用詞錯誤率 (Word error rate, WER)，中文採用字錯誤率 (Character error rate, CER)。

(三) 實驗結果與分析

實驗結果如表二，方法(1)–(3)為只用單個語言模型進行解碼得出的結果，從數據可發現 Meeting LM、General LM 和 Keyword LM 分別在會議錄音測試集、長句測試集和短句測試集和其他語言模型相比都有較好的表現。

方法(4)–(6)為 Meeting LM 和 Keyword LM 在三種不同層次的合(結合比例為 1：

表二、不同層次的語言模型合併於測試集的 MER

方法 \ 測試集	會議錄音 測試集	短句 測試集	長句 測試集	平均
Meeting LM (M) – (1)	27.85	40.48	18.75	29.03
General LM (G) – (2)	41.72	39.83	18.54	33.36
Keyword LM (K) – (3)	94.31	2.46	80.05	58.94
N-gram LM 合併(M+K) – (4)	28.61	2.97	18.61	16.73
Graph 合併(M+K) – (5)	28.06	3.02	19.29	16.79
Lattice 合併(M+K) – (6)	70.07	3.25	49.28	40.87
N-gram LM 合併(G+K) – (7)	42.46	2.92	19.47	21.62
Graph 合併(G+K) – (8)	42.04	2.81	19.25	21.37
Lattice 合併(G+K) – (9)	84.66	3.28	46.98	44.97
N-gram LM 合併(M+G+K) – (10)	29.68	3.19	15.82	16.23
Graph 合併(M+G+K) – (11)	28.18	3.00	18.80	16.66
Lattice 合併(M+G+K) – (12)	68.06	3.68	40.44	37.39
Mixed LM (M+G+K) – (13)	35.57	6.55	16.70	19.61

1)；方法(7)–(9)為 General LM 和 Keyword LM 在三種不同層次的合併（結合比例為 1：1）。由方法(1)可發現 Meeting LM 其實在長句測試集的 MER 和 General LM 只有些微差距，所以方法(4)–(5)能在三個測試集都有好的效果；反之，General LM 因為在會議錄音測試集的表現不好，所以方法(7)–(8)的平均 MER 比方法(4)–(5)差。另外透過方法(6)和(9)可發現 Lattice 合併只在短句測試集有好效果，其他測試集的 MER 都很高，我們推測因為 Keyword LM 是用特定領域的資料訓練而成，透過 Lattice 合併的方法相比其他方法較容易影響到其他語言模型的效能。

方法(10)–(12)為 Meeting LM、General LM 和 Keyword LM 在三種不同層次的合併（結合比例為 1：1：1）。方法(12)和方法(6)、(9)一樣只在短句測試集有好效果，而方法(10)和(11)的平均 MER 比只用兩個語言模型合併還要低，證實了合併三個語言模型能讓 ASR 系統在三個測試集都有好的表現。

另外，方法(13)為將三個語言模型的訓練資料合併在一起再訓練成一個語言模型進行解碼得出的結果，從平均 MER 的比較結果可證明透過方法(10)和(11)的合併方法會比

方法(13)好。而在長句測試集上，方法(10)和(13)比(11)好的原因，我們推測在語言模型比較早的階段開始合併，會對長句測試集比較有幫助，因為方法(10)和(13)都是在建成解碼圖的階段之前合併，在其產生的解碼圖上，各個語言模型的搜索空間並非各自獨立的，彼此間會互相影響，所以會較易解碼出混合各個語言模型詞語的句子。但是在其他測試集上反而因為這種性質而導致 MER 較高。

五、結論

本篇論文採用多種不同層次的語言模型合併於 CS ASR 上，目的為結合不同語言模型的優勢以在不同的任務上都能有好的表現。透過實驗數據可發現 *N-gram* 語言模型的合併和 *Graph* 的合併被證實能有效地結合不同語言模型的優勢，並能在各個測試集上都有好的表現，最後以結合三種不同的語言模型的效果最好，也證實了比直接混合三種訓練資料訓練的語言模型表現得還要好。另外 *Lattice* 合併於實驗中因為較易受到一個特定領域資料訓練的語言模型所影響，而沒有預期的好表現於不同的測試集上。未來我們也將對這部分進行深入探討。

參考文獻

- [1] Shinji Watanabe, Takaaki Hori, and John R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *Proc. ASRU*, 2017.
- [2] Hiroshi Seki, Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey, “An end-to-end language-tracking speech recognizer for mixed-language speech,” in *Proc. ICASSP*, 2018.
- [3] Ne Luo, Dongwei Jiang, Shuaijiang Zhao, Caixia Gong, Wei Zou, et al., “Towards end-to-end code-switching speech recognition,” in *Proc. ICASSP*, 2019.
- [4] Ke Li, Jinyu Li, Guoli Ye, Rui Zhao, and Yifan Gong, “Towards code-switching asr for end-to-end ctc models,” in *Proc. ICASSP*, 2019.
- [5] Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, et al., “Investigating end-to-end speech recognition for mandarin-english code-switching,” in *Proc. ICASSP*, 2019.

- [6] Zhiping Zeng, Yerbolat Khassanov, Van Tung Pham, Haihua Xu, Eng Siong Chng, et al., “On the End-to-End Solution to Mandarin-English Code-switching Speech Recognition,” in *Proc. INTERSPEECH*, 2019.
- [7] Metilda Sagaya Mary N J, Vishwas M. Shetty, and S. Umesh, “Investigation of Methods to Improve the Recognition Performance of Tamil-English Code-Switched Data in Transformer Framework,” in *Proc. ICASSP*, 2020.
- [8] Duo Ma, Guanyu Li, Haihua Xu, and Eng Siong Chng, “Improving code-switching speech recognition with data augmentation and system combination,” in *Proc. APSIPA* 2019.
- [9] Emre Yilmaz, Henk van den Heuvel, and David A. van Leeuwen, “Acoustic and Textual Data Augmentation for Improved ASR of Code-Switching Speech,” in *Proc. INTERSPEECH*, 2018.
- [10] Emre Yilmaz, Henk van den Heuvel, and David A. van Leeuwen, “Code-Switching Detection with Data-Augmented Acoustic and Language Models,” in *Proc. SLTU*, 2018.
- [11] Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, “Zero-Shot Code-Switching ASR and TTS with Multilingual Machine Speech Chain,” in *Proc. ASRU*, 2019.
- [12] Yuewen Cao, Xixin Wu, Songxiang Liu, Jianwei Yu, Xu Li, et al., “End-to-end Code-switched TTS with Mix of Monolingual Recordings,” in *Proc. ICASSP*, 2019.
- [13] Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li, “End-to-End Code-Switching TTS with Cross-Lingual Language Model,” in *Proc. ICASSP*, 2020.
- [14] Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung, “Learn to Code-Switch: Data Augmentation using Copy Mechanism on Language Modeling” in *Proc. ICASSP*, 2019.
- [15] Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee, “Code-switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation,” in *Proc. INTERSPEECH*, 2019.
- [16] Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee, “Code-switching Sentence

Generation by Generative Adversarial Networks and its Application to Data Augmentation,” in *Proc. INTERSPEECH*, 2019.

- [17] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che, “CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP,” in *Proc. IJCAI*, 2020.
- [18] Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung, “Towards End-to-end Automatic Code-Switching Speech Recognition,” in *Proc. ICASSP*, 2019.
- [19] Min Ma, Bhuvana Ramabhadran, Jesse Emond, Andrew Rosenberg, and Fadi Biadsy, “Comparison of Data Augmentation and Adaptation Strategies for Code-switched Automatic Speech Recognition,” in *Proc. ICASSP*, 2019.
- [20] Gustavo Aguilar and Tamar Solorio, “From English to Code-Switching: Transfer Learning with Strong Morphological Clues,” in *Proc. ACL*, 2020.
- [21] Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, et al., “Meta-Transfer Learning for Code-Switched Speech Recognition,” in *Proc. ACL*, 2020.
- [22] Sanket Shah, Basil Abraham, Gurunath Reddy M, Sunayana Sitaram, and Vikas Joshi, “Learning to Recognize Code-switched Speech Without Forgetting Monolingual Speech Recognition,” *arXiv:2006.00782*, 2020.
- [23] Gurunath Reddy Madhumani, Sanket Shah, Basil Abraham, Vikas Joshi, and Sunayana Sitaram, “Learning not to Discriminate: Task Agnostic Learning for Improving Monolingual and Code-switched Speech Recognition,” *arXiv:2006.05257*, 2020.
- [24] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W. Black, “Sequence-based Multi-lingual Low Resource Speech Recognition,” in *Proc. ICASSP*, 2018.
- [25] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, “Meta Learning for End-to-End Low-Resource Speech Recognition,” in *Proc. ICASSP*, 2020.
- [26] Yerbolat Khassanov, Haihua Xu, Van Tung Pham, Zhiping Zeng, Eng Siong Chng, et al., “Constrained Output Embeddings for End-to-End Code-Switching Speech Recognition with Only Monolingual Data,” in *Proc. INTERSPEECH*, 2019.
- [27] Shun-Po Chuang, Tzu-Wei Sung, and Hung-Yi Lee, “Training a code-switching language model with monolingual data,” in *Proc. ICASSP*, 2020.

- [28] Emre Yilmaz, Samuel Cohen, Xianghu Yue, David van Leeuwen, and Haizhou Li, “Multi-Graph Decoding for Code-Switching ASR,” in *Proc. INTERSPEECH*, 2019.
- [29] Xianghu Yue, Grandee Lee, Emre Yilmaz, Fang Deng, and Haizhou Li, “End-to-End Code-Switching ASR for Low-Resourced Language Pairs,” in *Proc. ASRU*, 2019.
- [30] Alexander H. Waibel, Toshiyuki Hanazawa, Geoffrey E. Hinton, Kiyohiro Shikano, and Kevin J. Lang, “Phoneme recognition using time-delay neural networks,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, Mar. 1989.
- [31] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. INTERSPEECH*, 2015.
- [32] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, et.al., “Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks,” in *Proc. INTERSPEECH*, 2018.
- [33] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, et al., “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in *Proc. INTERSPEECH*, 2016.
- [34] Daniel Povey, Arnab Ghoshal, Ghoshal Boulianne, Lukas Burget, Ondrej Glembek, et al., “The kaldı speech recognition toolkit,” in *Proc. ASRU*, 2011.
- [35] Daniel Povey, Mirko Hannemann, Gilles Boulianne, Lukas Burget, Arnab Ghoshal, et al., “Generating exact lattices in the WFST framework,” in *Proc. ICASSP*, 2012.
- [36] Haihua Xu , Daniel Povey , Lidia Mangu , and Jie Zhu, “An improved consensus-like method for Minimum Bayes Risk decoding and lattice combination,” in *Proc. ICASSP*, 2010.
- [37] Haihua Xua , Daniel Poveyb , Lidia Manguc , and Jie Zhua, “Minimum Bayes Risk Decoding and System Combination Based on a Recursion for Edit Distance,” in *Proc. CSL*, 2011.
- [38] Tien-Hong Lo and Berlin Chen, “Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition,” in *IJCLCLP*, 2018.

運用集成式多通道類神經網路於科技英文寫作評估

Scientific Writing Evaluation

Using Ensemble Multi-channel Neural Networks

王昱翔 Yuh-Shyang Wang, 李龍豪 Lung-Hao Lee
國立中央大學電機工程學系
Department of Electrical Engineering
National Central University
yswang135@gmail.com, lhlee@ee.ncu.edu.tw

林柏霖 Bo-Lin Lin, 禹良治 Liang-Chih Yu
元智大學資訊管理學系
Department of Information Management
Yuan Ze University
ss27713084@gmail.com, lcyu@saturn.yzu.edu.tw

摘要

現在有許多母語非英語人士撰寫的科學論文，幫助作者撰寫科學論文的自動化工具產生了巨大的需求。國際科技英文寫作評估評測任務藉由評估一個論文中的英文句子，是否需要語言編輯為任務目標，幫助開發自然語言處理工具，用以改善科技英文寫作的品質。本研究透過實驗設計比較通道數、模型架構和集成數，提出一個集成式多通道類神經網路架構，在該評測資料集下獲得 F1 分數 63.28，比當時參與評測的系統有更好的效能。

Abstract

A huge number of scientific papers have been authored by non-native English speakers. There is a large demand for effective computer-based writing tools to help writers composing scientific articles. The Automated Evaluation of Scientific Writing (AESW) shared task seeks to promote the use of NLP tools for improving the quality of scientific writing in English by predicting whether a given sentence needs language editing or not. In this study, we propose an ensemble multi-channel BiLSTM-CNN model based on a series of experiments in comparing the number of channels, network architectures, and ensemble size. Our model achieved an F1 score of 63.28 outperforms participating systems in the AESW 2016 task.

關鍵詞：集成學習、多通道神經網路、自動寫作評估, 科技英文

Keywords: Ensemble Learning, Multi-channel Neural Networks, Automated Writing Evaluation, Scientific English

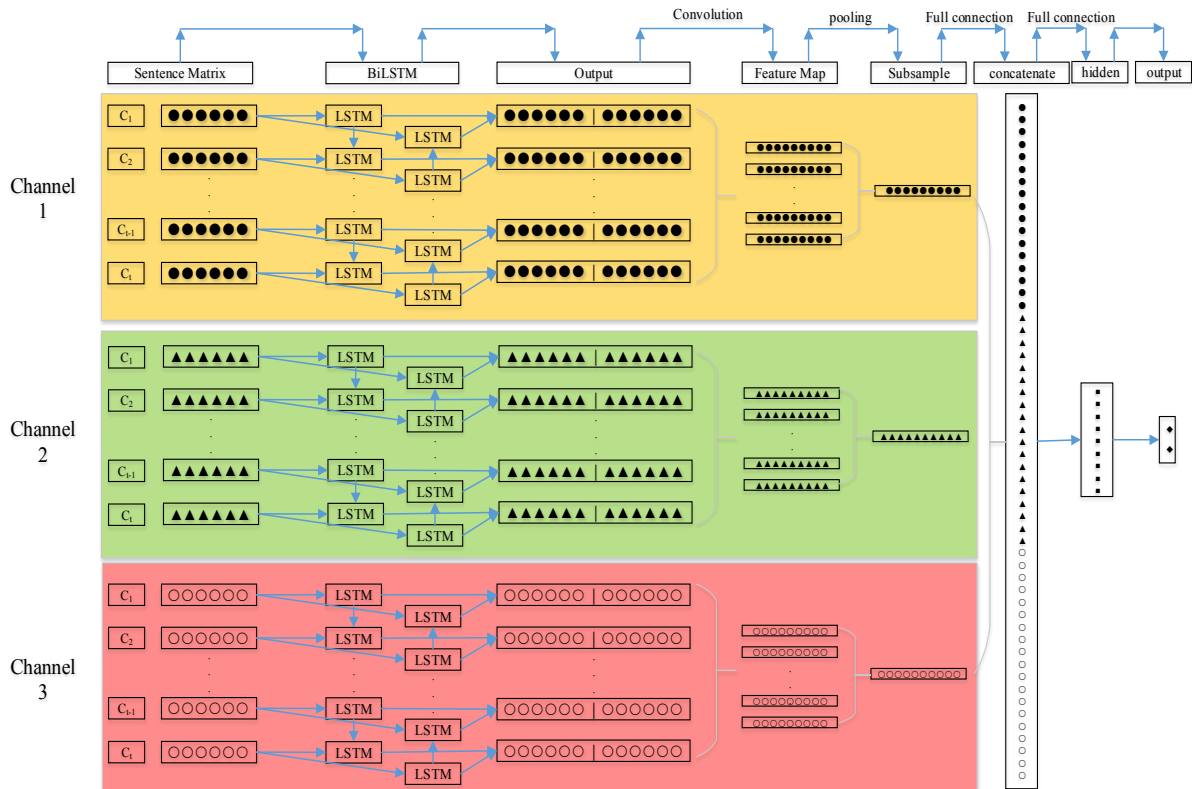
一、緒論

英文是世界上最常被使用的語言之一，有 67 個國家將英文列為他們的官方語言，非英語系的國家通常都將英文視為第一外語或第二外語學習。語言學習有四個重點「聽、說、讀、寫」，在正式的場合如信件、商業合約、論文等等，對於寫作者需要額外的人力去校稿。近年有越來越多非英文母語者撰寫的科學論文，因而對科技英文寫作的自動化評測工具的需求與日俱增。一個有效的自動化寫作工具，可以幫助寫作者減少語意表達上的錯誤，提升寫作品質，進而減少校稿的時間以及人力成本。為了建立自動化工具，語法錯誤檢測和更正是必要的一環，也有許多的任務競賽：Helping Our Own (HOO)是一系列用於寫作者文法錯誤校正任務 [1][2]。CoNLL 2013/2014 的評測任務則是英文為外語學習者的文法錯誤更正[3][4]。

本研究使用的資料來源為科技英文寫作評估競賽(Automated Evaluation of Scientific Writing, AESW)。AESW 2016 評測任務的目的是分析科學寫作的語言特徵，促進科技論文的自動化寫作評估工具的發展[5]。任務內容分為兩個子任務：一是二元分類預測，輸入的句子判定是否需要語言校正，如果是則預測為 True，反之為 False；二是機率估算，系統需要估計句子需要被校正的機率值。有鑑於不同的詞嵌入以及神經網路模型在語法偵錯任務上各有優缺點，本研究透過完整的實驗流程，結合各種詞嵌入(word embedding)與卷積神經網路 (Convolutional Neural Network, CNN) 以及長短期記憶神經網路 (Long Short-Term Neural Network, LSTM) 兩個深度學習的基礎模型，藉由實驗比較通道數、模型架構、集成數對子任務一分類效能的差異，建構出集成式多通道類神經網路，在測試集上達到 63.28 的 F1 分數，與競賽時的方法相比較，有更好的分類成效。

二、模型架構

我們提出的方法為集成式多通道類神經網路 (Ensemble Multi-Channel BiLSTM-CNN) 模型，其架構圖如圖一，由詞向量輸入、多通道輸入、雙向長短期記憶網路、卷積神經網路所組成。



圖一、集成式多通道類神經網路架構

(一)、詞嵌入向量 (Word Embedding)

詞向量是在自然語言處理中常用的方法，要將語言輸入給機器運算前，需要將其數值化，詞向量就是將文句中的詞數值化的方式，將每一個單詞以一個向量表示。最簡單的方式維 One-hot Encoding 是用一個維度等同文本中詞彙數的向量來表示，向量中只包含一個 1 與多個 0，字典中第一個詞表示為 $[1,0,0,\dots,0]$ 、第二個詞為 $[0,1,0,0,\dots,0]$ 以此類推。這種表示的缺點為當文本、字典較大時，代表每個詞的維度就會變得極為巨大，引發維度災難，造成運算上的困難，而且這種表示方式對於詞與詞之間的關係沒有代表性。

為了解決該上述問題，Hinton 於 1986 年提出 Distributed Representations，將所有詞向量組合成一個詞向量空間，每個向量則為空間上的一點，點與點之間的距離即為詞之間的相似性[6]。常用的傳統靜態詞向量工具有 Word2vec, GloVe, fastText。Word2Vec 為 Mikolov 於 2013 年提出，使用連續詞袋模型(continuous bag of words, CBOW)和 skip-gram，以非監督式學習的算法學習單詞的含義[7]。GloVe 為史丹佛大學於 2014 年提出，基於文本內詞彙的共現矩陣(Co-occurrence Matrix)，對共現矩陣進行訓練，計算出詞向量[8]。fastText 為 Facebook 於 2016 提出，相較於 Word2vec，fastText 引入了 N-gram 考

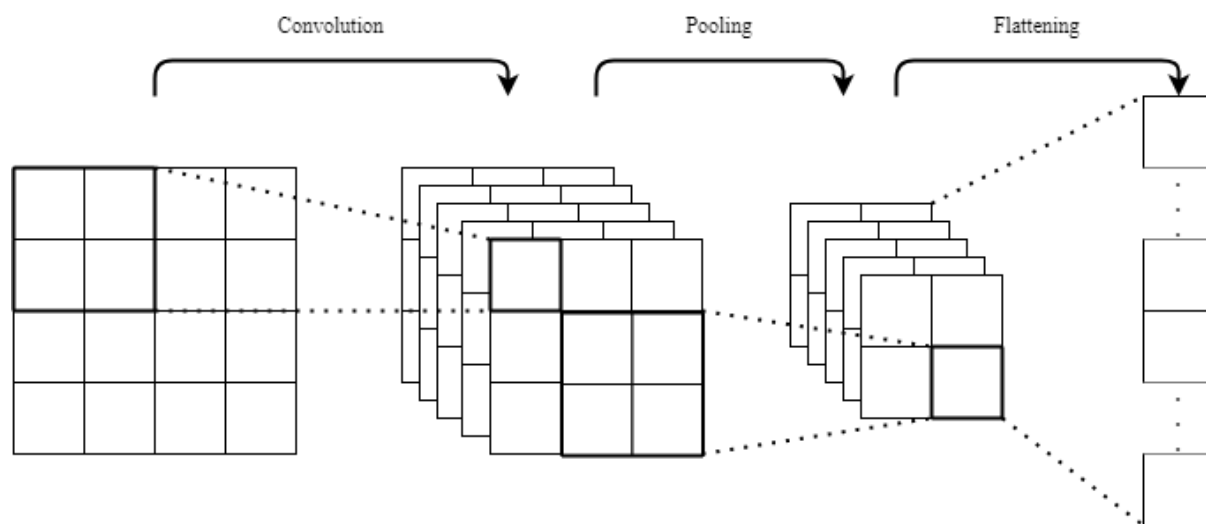
慮詞序特徵，並使用 subword 來處理未登錄詞[9]。

本研究採用三個詞向量模型分別為：Word2vec 官方利用 GoogleNews dataset 所訓練的詞向量模型，內含各為 300 維向量的 300 萬個字詞。GloVe 官方利用 Common Crawl 資料集所訓練的 glove.840.300d 詞向量模型，包含 300 維的 220 萬個字詞。fastText 官方提供使用 Wikipedia 所訓練的 294 個不同語言的詞向量中的英文詞向量，其維度也是 300 維。

(二)、卷積神經網路(Convolutional Neural Network, CNN)

卷積神經網路(CNN)示意圖如圖二，在圖像處理上有出色的表現，也能有效處理自然語言中的語意分析、分類、預測等任務。在 2016 年 DeepMind 透過結合蒙地卡羅搜尋法 (MCTS) 與深度卷積神經網路(DCNN)提出的演算法開發出 AlphaGo，並與韓國職業九段棋士世界冠軍李世乜對弈以四勝一敗獲勝，引起世界大量關注。

CNN 有兩個主要部分：卷積層(Convolution Layer)和池化層(Pooling Layer)，卷積層透過在輸入圖像上的數個卷積核滑動計算並提取資料的特徵，各個卷積核可以分別得出一個特徵圖(Feature Maps)。而池化層將前面獲得的特徵圖做次採樣(Subsampling)，最常見的方法為最大池化(Max Pooling)，將特徵圖切為多個矩形，輸出各區的最大值，透過池化可以保留顯著的特徵並降低特徵的數量。

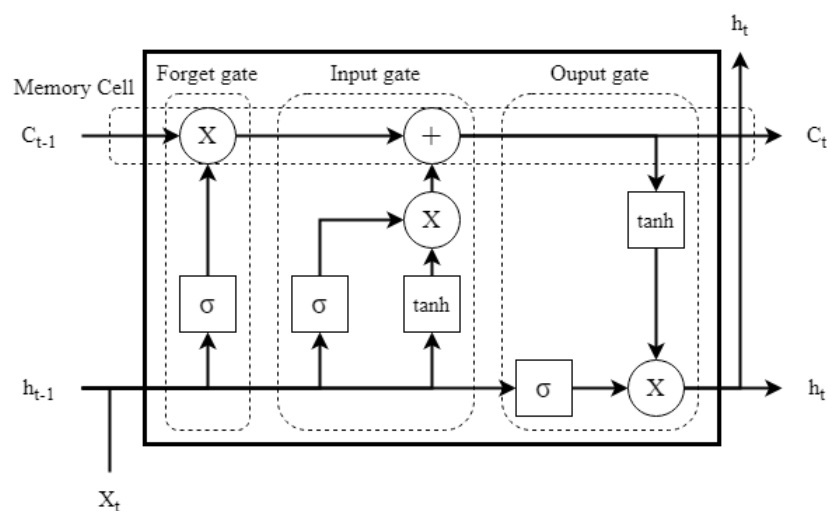


圖二、卷積神經網路

(三)、長短期記憶網路 (Long Short-Term Memory, LSTM)

1、單向長短期記憶網路 (LSTM)

長短期記憶網路(LSTM) 是一種時間循環神經網路 (Recurrent Neural Network, RNN)。一般的 RNN 各節點的輸入為輸入的資料以及前一個節點的輸出，當序列較長時前面的資訊便無法完整傳遞到後面，因此 RNN 在處理短文句時會有很好的表現，但是在較長的句子時就會無法順利預測。

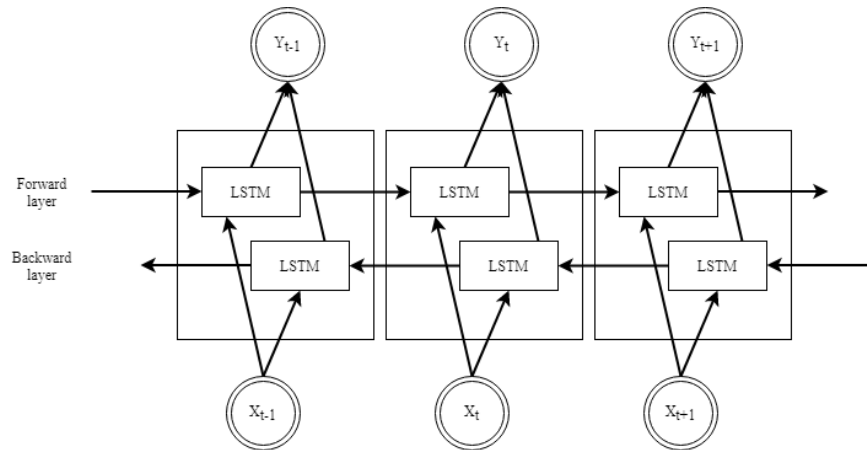


圖三、LSTM Cell

因此 1997 年 Hochreiter 和 Schmidhuber 提出了 LSTM，基於 RNN 架構引入了記憶單元來解決這個問題。LSTM Cell (如圖三)由四個元件組成輸入門(Input Gate)、輸出門(Output Gate)、遺忘門(Forget Gate)、記憶單元(Memory Cell)，輸入門決定要輸入到記憶單元的特徵、遺忘門決定要刪除的特徵訊息、而輸出門則是決定記憶單元內的特徵是否能輸出。

2、雙向長短期記憶 (Bi-directional Long Short-Term Memory, BiLSTM)

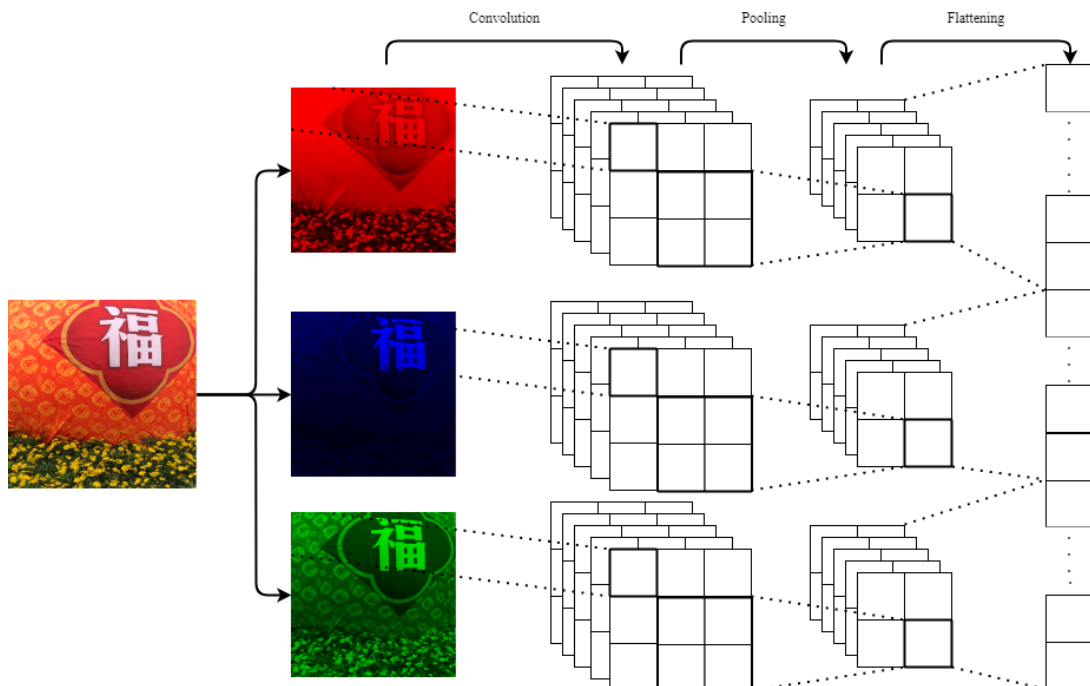
單向長短期記憶在學習到時間點 t 時，只能獲得該時間點之前的訊息，而雙向長短期記憶網路(架構圖如圖四)則能在每個時間點都獲得前後文的序列狀態。在英文文法偵錯時，需要同時注意前後文以確認時態、助動詞等是否正確使用，因此 BiLSTM 應較 LSTM 更適合本實驗。



圖四、雙向長短期記憶網路架構

(四)、多通道輸入 (Multi-Channel)

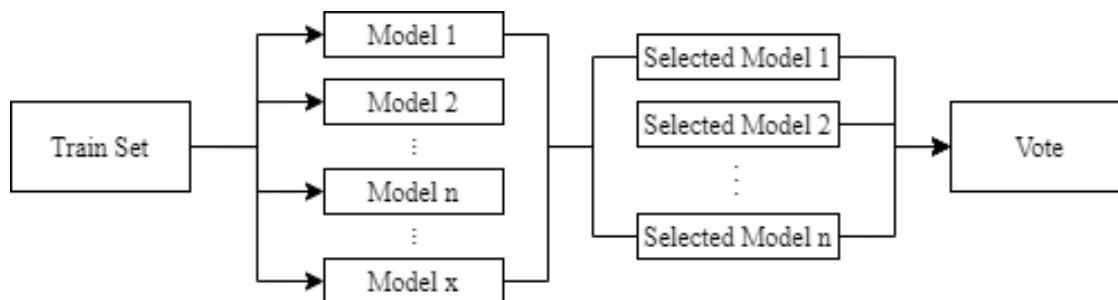
單通道僅能用將資料以一種表示方式輸入網路，若使用多通道則可將資料以多種方式輸入類神經網路之中。以彩色圖片使用多通道卷積神經網路進行分類如圖五為例：將圖片以 RGB 三元素表示時，將 R、G、B 數值分為三通道輸入，分別提取特徵，拓展卷積神經網路的視野。



圖五、多通道卷積神經網路

(五)、集成學習 (Ensemble Learning)

集成學習為使用多種學習算法來獲得比單獨使用一種學習算法更好的預測性能，如俗話說「三個臭皮匠勝過一個諸葛亮」。集成學習的常見方法有 Bagging、Boosting、Stacking 等，我們採用類似 Bagging(如圖六)的作法，將同一份訓練集，訓練多個模型並選擇表現較好的模型進行投票。



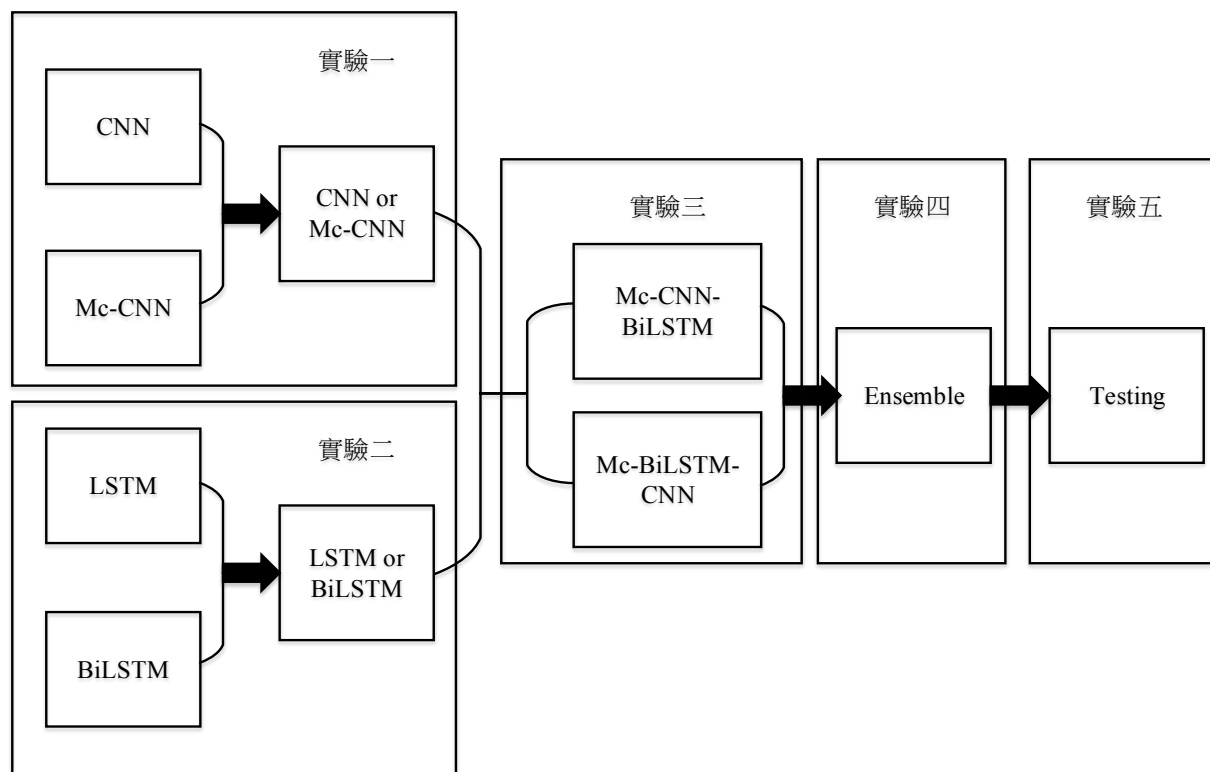
圖六、集成學習投票機制

三、實驗與結果

(一)、實驗設計

實驗使用的資料來源為 AESW 2016 科技英文自動評測國際競賽提供的資料，它是第一個大規模公開的科技寫作資料集，資料中包含需要語言編輯以符合科技論文寫作的體裁的文句，以及不須編輯便已符合體裁的文句，需要經過編輯的文句會附上如何編輯以符合體裁。標籤中為需要刪除的字詞，而<ins>中則是需要加入的字詞。訓練集共有 1,196,940 筆，其中 466,672 筆為經語言編輯(True 標記)，發展集有 148,478 筆，其中 57,340 筆為經語言編輯，而測試集有 143,784 筆。三組資料中經語言編輯比例均約佔四成、未經修改的則為六成。該競賽目標為輸入一個未經編輯的文句，預測其是否需要語言編輯以符合寫作體裁，評估方式是將模型預測結果與實際答案做比對，以 F1 為主要評分標準。

本實驗分為五個階段(如圖七)，透過逐一比較效果並決定參數，實驗一比較單通道與多通道的表現，實驗二做單向與雙向 LSTM 的比較，實驗三再依照前兩個比較出較好的模型做組合，實驗四是集成多模型的結果，決定出最佳模型及參數。前四個實驗均是在發展集上調整參數，以三次實驗平均作為比較標準，最後實驗五再對測試集做預測。



圖七、實驗流程

(二)、實驗一結果

第一個實驗比較單通道、雙通道、三通道在 CNN 最佳參數時的平均表現，以決定後續實驗所使用的通道數，以及詞向量組合順序。不同通道數的表現結果分見於表一、表二和表三，單通道是以使用 GloVe 為詞向量時表現最佳，平均 F1 為 0.6392，而多通道包含雙通道及三通道中表現最佳的是三通道的 Word2vec+ FastText+GloVe 組合，平均 F1 為 0.6422，此處實驗可得知三通道在此實驗上表現較好，結合訓練方式各異的詞向量特徵，更能有效獲得語句中的訊息。

表一、單通道 CNN 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
GloVe	0.5484	0.7661	0.6392	0.0004
fastText	0.5991	0.6729	0.6365	0.0012
Word2vec	0.5763	0.7065	0.6347	0.0016

表二、雙通道 CNN 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
Word2vec + GloVe	0.5636	0.7354	0.6375	0.0014
GloVe + Word2vec	0.5753	0.7129	0.6367	0.0011
GloVe + fastText	0.5683	0.7266	0.6375	0.0031
fastText + GloVe	0.5507	0.7617	0.6390	0.0024
fastText + Word2vec	0.5647	0.7241	0.6341	0.0029
Word2vec + fastText	0.5746	0.7121	0.6360	0.0010

表三、三通道 CNN 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
Word2vec + GloVe + fastText	0.5617	0.7473	0.6404	0.0028
GloVe + Word2vec + fastText	0.5638	0.7383	0.6391	0.0003
GloVe + fastText + Word2vec	0.5654	0.7376	0.6400	0.0023
fastText + GloVe + Word2vec	0.5423	0.7838	0.6403	0.0030
fastText + Word2vec + GloVe	0.5722	0.7264	0.6396	0.0036
Word2vec + fastText + GloVe	0.5524	0.7671	0.6422	0.0017

(三)、實驗二

第二個實驗比較單雙向長短期記憶網路(LSTM vs. BiLSTM)在此任務上的表現，作為後續實驗模型選擇的依據。單雙向模型的表現分別在表四及表五，LSTM 以 GloVe 為詞向量時表現最好，平均 F1 為 0.6419；BiLSTM 以 GloVe 為詞向量時表現最好，平均 F1 為 0.6473，勝過 LSTM。與前面推論相同，BiLSTM 能獲取前後文訊息，較適合語法語意偵錯的任務。

表四、LSTM 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
GloVe	0.5201	0.8383	0.6419	0.0007
fastText	0.5151	0.8359	0.6373	0.0020
Word2vec	0.5221	0.8264	0.6399	0.0013

表五、BiLSTM 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
GloVe	0.5529	0.7803	0.6473	0.0002
fastText	0.5458	0.7888	0.6452	0.0004
Word2vec	0.5443	0.7889	0.6441	0.0006

(四) 實驗三

第三個實驗根據前兩個實驗的結果，選用多通道、BiLSTM 與 CNN 組合，建立更複雜的網路，使用多通道表現最好的詞向量組合 Word2vec+fastText+GloVe，實驗結果記錄於表六。兩模型中前者為將詞向量經由 CNN 取得 feature map 再進入 BiLSTM，後者則反之。Mc-BiLSTM-CNN 的平均 F1 為 0.6536 較 Mc-CNN-BiLSTM 的 0.6491 來得好。

表六、Mc-BiLSTM-CNN 與 Mc-CNN-BiLSTM 於發展集的結果

Model	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
Mc-BiLSTM-CNN	0.5529	0.7803	0.6473	0.0002
Mc-CNN-BiLSTM	0.5458	0.7888	0.6452	0.0004

(五)、實驗四

第四個實驗為了驗證集成(ensemble)學習的效果，利用實驗三所決定的模型架構、參數，訓練了數十個模型，取出表現最佳的前 N 名，若超過 $(N/2) + 1$ 個模型的預測結果認為該句有誤，便認為該句文法有誤，反之亦然。實驗結果如表七，可以看出 N 為 3 和 5 時表現提升不少，增加到 7 和 9 時幾乎沒有提升，11 時反而表現變差，因此決定採用最佳的前 9 名為最終參數。

表七、Ensemble Mc-BiLSTM-CNN 於發展集的結果

Ensemble (N)	Precision	Recall	F1
3	0.5606	0.7980	0.65860
5	0.5599	0.8010	0.65911
7	0.5586	0.8040	0.65928
9	0.5676	0.7866	0.65994
11	0.5652	0.7907	0.65924

(六)、實驗五

將以上實驗中的最佳模型，在 AESW2016 國際評測的測試集，驗證效能結果如下表八。與實驗預期相同，多通道比單通道好，深層網路架構，以及集成學習都是有效提升模型表現的方法。而表九則是與該競賽參賽隊伍的成績做比較，我們的最佳模型 Ensemble(N=9) Mc-BiLSTM-CNN 達到 F1 分數 0.6328，超越了表現最好的哈佛大學團隊的 0.6278。

表八、實驗中各模型測試結果

Method	Precision	Recall	F1
CNN	0.5274	0.7153	0.6071
Mc-CNN	0.5256	0.7405	0.6148
LSTM	0.4786	0.8316	0.6076
BiLSTM	0.5157	0.7735	0.6188
Mc-CNN-BiLSTM	0.5257	0.7581	0.6209
Mc-BiLSTM-CNN	0.5144	0.7988	0.6258
Ensemble (N=9) Mc-BiLSTM-CNN	0.5359	0.7724	0.6328

表九、測試結果與參賽隊伍的比較

Team	Method	Precision	Recall	F1
Hu	CNN, RNN, LSTM	0.5444	0.7413	0.6278
HITS	HMM, Logistic Regression	0.3765	0.9480	0.5389
ISWD	SVM, SubSet Tree kernel	0.4482	0.7279	0.5548
Knowlet	MaxEnt	0.6241	0.3685	0.4634
NTNU-YZU	CNN	0.5025	0.7785	0.6108
UW-SU	MaxEnt	0.4145	0.8201	0.5507
Ours	Ensemble (N=9) Mc-BiLSTM-CNN	0.5359	0.7724	0.6328

四、結論

本研究將科技英文寫作的句子是否需要語言編修，視為一個標準的二元分類問題，使用 AESW2016 國際競賽的資料集驗證方法成效。在文句分類時，最重要的就是文句的表示方式以及找到合適的模型，透過實驗由不同的詞向量(GloVe, Word2vec, fastText)、通道選擇(單通道、雙通道、三通道)、模型的挑選(CNN vs. Mc-CNN, LSTM vs. BiLSTM, Mc-CNN-BiLSTM vs. Mc-BiLSTM-CNN)，到 Ensemble 個數，最後提出了 Ensemble (N=9) Mc-BiLSTM-CNN 集成式多通道類神經網路模型，這是一種利用多種靜態詞向量，透過多個神經網路集成預測結果的分類模型，此方法在測試集上得到最好的 F1 分數為 0.6328，較當時參賽隊伍中表現最好的表現高。未來希望除了靜態詞向量之外，可以採用動態詞向量如：BERT 或者是 ELMO，或是使用 Transformer 進行分類。

致謝

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3 and MOST 107-2628-E-155-002-MY3.

參考文獻

[1] R. Dale and A. Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In

Proceedings of the 13th European Workshop on Natural Language Generation, pages 242–249.

- [2] R Dale, I Anisimoff, and G Narroway. 2012. A report on the preposition and determiner error correction shared task. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.
- [3] H. T. Ng, S. M. Wu, C. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.
- [4] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.
- [5] V. Daudaravicius, R. E. Banchs, E. Volodina and C. Napoles. 2016. A report on the automated evaluation of scientific writing shared task. In *Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 53–62.
- [6] G. E. Hinton. Learning distributed representations of concepts. 1986. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 1–12.
- [7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Neural Information Processing Systems 2013*, pages 1–10.
- [8] J. Pennington, R. Socher and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Empirical Methods on Natural Language Processing*, pages 1532–1543.
- [9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

建構拍賣網提問類型之多重標籤辨識模型

Building A Multi-Label Detection Model for Question classification of Auction Website

林怡如 I-Ju, Lin

東吳大學巨量資料管理學院
School of Big Data Management

Soochow University

06770024@gm.edu.tw

吳政隆 Jheng-Long Wu

東吳大學巨量資料管理學院
School of Big Data Management

Soochow University

jlwu@gm.scu.edu.tw

摘要

網路購物已成為現今非常重要的消費型態，每日都會產生成千上萬的銷售問題。此時，客服中心是介於客戶與公司間的第一線服務單位，為了解決客戶服務提問相關問題的需求，本研究運用自然語言處理與機器學習技術，嘗試建構客戶之發問類型辨識模型。其本研究目的為建構一個多重標籤之分類任務資料集，其主是露天採用拍賣之問題類型部分作為主要研究資料來源，以及將資料分為粗細分類的兩組資料集。本研究嘗試以 TF-IDF 和 Word Embeddings 等兩種詞向量演算法進行特徵轉換，以及採用 Extra Trees、Logistic Regression、Random Forest、SVM 等四種機器學習分類演算法進行分類模型建立。本研究實驗結果顯示透 TF-IDF 的詞特徵轉換，以及搭配 Extra Trees 分類模型，其分類效果為最佳，高達 0.8285 F1 分數。顯示露天拍賣網之多重問題類型標籤辨識可有效被機器識別正確。

Abstract

Online shopping (e-commerce) has become an indispensable type of consumption today. The problems faced by the customer service center in e-commerce back-offices are more complicated. This research's core content is to explore how the customer service center to improve the workflow to provide customer's needs solutions by using machine learning

technology. In this study, the detailed and rough types of multi-label detection predicted model was trained by the customers' questions texts about the products of auction website "RUTEN". The TF-IDF and Word Embeddings methods were used to extract the text features, and we experiment with Extra Trees、Logistic Regression、Random Forest、SVM classification models to build a multi-label detection model. The overall result of experiments, the features extracted by the TF-IDF method, and detected by the Extra Trees classification model have performed a better F1 score with 0.82846.

關鍵詞：問題類型分類、多重標籤辨識，文字探勘，詞向量，機器學習。

Keywords: Question Classification, Multi-Label Detection, Text Mining, Word Embeddings and Machine Learning

一、緒論

根據金管會銀行局公開之信用卡消費資料統計，信用卡消費中用來支付網路購物之金額佔比，逐年成長至民國一百零八年來到百分之十八點一，而其EC(Electronic Commerce)消費金額在 108 年來到了新台幣 3,700 億元，與前一年相比成長比達 34%。客戶在網路購物過程中所產生的服務需求問題，無論是網站流量速度（平台系統面）、行銷活動、商品面、物流收送貨、退款速度（金流面）等所有問題，都將全數導向客服中心處理，而客戶需求通常不會是單一存在反而是多樣複雜且難以區分辨識的。在大量進線與留言湧進的同時，單純透過傳統客服人員依循著標準作業流程以及過去的服務經驗值，辨識及處理客戶提出的問題，其處理效率確實非常不佳。在葉靜縈[1]針對壽險業客服對話文本之多標籤主題預測研究中，以客服人員發聲之文本資料的分析結果相對具一定之準確率。由於客服人員具備專業服務訓練，其對話應對被要求必須符合標準作業規範；相較於本研究針對客戶提問內容進行多標籤辨識，口語化文字留言與錯別字的可能性，都增加分析預測的困難度。且因產業屬性的不同，網路購物不論在商品種類、行銷活動、銷售／服務供應鏈等影響因素，都遠比壽險業來得多樣多變。由表 1 客戶提問內容範例可得知，客戶需求通常不會是單一存在反而是多樣複雜且難以區分辨識的。

表 1 客戶提問內容範例

編號	提問內容	類別
1	請問可以貨到付款嗎？可以指定時段嗎？	配送方式、付款方式
2	請問如果要下訂 450 個:1.數量夠嗎？可否挑色？2.希望 7/15 前到貨，是否來的及？	庫存、商品相關、到貨出貨日期

基於上述情況，本研究針對電商客服中心最常見的問與答日常任務，進行自動化辨識客戶需求問題之多重分類模型建置。本研究著採用自然語言處理（Natural Language Processing, NLP）與機器學習技術（Machine Learning）技術，針對電商客服中心收到的客戶需求問題內容，透過詞向量化技術將文本轉換特徵向量，再應用機器學習分類方法建立分類模型，以期望達到對於留言內容進行提問類型的多標籤分類預測，以達成自動分類效果，進而加速回覆客戶的效率與品質為目的。本研究主要貢獻在於產生一組電商平台之提問問題多重標籤分類資料集，能夠讓電商產業業者以此資料集作為雛型進行開發。本研究也嘗試採用不同的文本向量化技術來證實其多重標間分類效果為何，促使學術研究專家學者能夠理解電商平台之多重提問標籤任務的困難度。

二、文獻探討

（一）自然語言處理技術於電子商務問題

客服問題是電子商務領域中相當長常見的議題，多數以建立分類模型等相關研究為主，如在國營事業、金融業、資訊服務業等不同領域的研究中都提到，將客戶提問內容進行建模分類後，客服系統就能依照客戶問題歸屬類別自動分派給權責單位處理，將問題分流指派給專職擅長處理的部門或客服，能快速準確回覆、縮短處理時間；問題被分類系統化進而建立作業標準程序後，使得客服人員能縮短查找該類別問題解答的時間、依循標準作業流程為客戶解答降低錯誤發生，進而分析找出各類別問題產生之原因，並使得後續報表分析能正確地分類。參考過去類似案件的處理經驗回覆眼前客戶的問題，立即且快速、能以好的服務品質提升企業形象以及顧客對商家的滿意度與忠誠度[2-5]。非結構化之文本資料已經成為許多研究議題的主要資料來源之一，文本資料需要轉換為可以被計算的特徵後，才得以應用於各式各樣的求解任務。最常見的特徵轉換方法為 TF-IDF (Term-Frequency Inverse Document Frequency)，是一種統計方法用來評估單詞在許多文章中的重要程度，越重要的單詞越有可能成為關鍵字。而 Word Embeddings 是將文字轉換為詞向量，供後續機器學習分類模型分析使用[6-9]。無論為何種特徵轉換方法，其目的都是為了表達字詞與特定事物的關係。

（二）多分類與多標籤辨識

有關非結構化文本資料多分類預測的相關研究為數不少，多以系統軟體使用者問題、新

聞社群文章、醫療病症、公文文件、客訴問題等單一分類內容居多。相較於本研究處理的是電商客戶提問文本多標籤辨識問題，商品及問題內容方向相對複雜、預測準確困難度較高。對於問題分類的相關研究中，Desai 等人將 Quora 問答平台對語意相似的問題進行分類，其中以 Random Forest 和 XGBoost 等的集成模型提供了最佳性能[10]；Sulaiman 等人則提出有關教育評估的技能和知識，基於認知水平對考試問題進行分類，實驗結果表明 SVC (Support Vector Classification) 有最好的分類效果[11]。而在相關多標籤分類預測研究中，Rokham 等人使用 SVM (Support Vector Machine) 從情緒和精神病性疾病的結構磁共振成像數據中檢測標籤噪聲[12]；Zhou 等人提出了一種多標籤學習方法來聯合學習零件檢測器以捕獲部分遮擋模式，通過共享一組決策樹以增強利用零件相關性[13]。如 PTT 電影版文章[14]與 Google Map POI[15]等多標籤分類的研究，他們嘗試將多標籤分類轉換為單標籤分類問題，分別對各個類別訓練二元分類模型後，測試時將全部分類器的結果組合作為輸出結果，來進行多類別預測分析；相較於本研究以支援 Multi-label Detection 的分類演算法，來預測每一個樣本多個屬性的方式略有不同。它們的研究也將多標籤分類問題進一步針對大小（粗細）類別，分別進行不同難易程度的辨識，其研究結果也是以大（粗）類別的多標籤辨識預測效果較佳。

（三）機器學習分類模型

Logistic Regression 邏輯式迴歸和線性迴歸很類似，都是在確認自變數和因變數之間的關聯。何欣儒[16]在網購低溫食品快速送達與超商取貨服務對購買行為影響之研究中，運用邏輯式迴歸模型，計算分析不同產品類別，使用何種配送方式較符合成本效益。而葉丞峻[17]則以邏輯式迴歸演算法，在解決二元分類不平衡問題有較好的成效。呂育如[18]的實驗在決策變數超過兩種類別時，以邏輯式迴歸模型有較佳的分類效果。隨機森林 (Random Forest) 的名稱是由 Random Decision Forests 而來，隨機森林可以看作一個多棵決策樹的集合，即是一種整合演算法。高詠富[19]在建立財務詐欺檢測模型、陳希聖[20]訓練超音波多特徵脂肪肝疾病分類模型、邱惠君[21]針對社群發言網路霸凌預測，其運用多項機器學習分類演算法進行預測實驗，均以隨機森林演算法能獲得較好的分類效果。Extra Trees 或 Extremely Randomized Trees 演算法由 Geurts 等人[22]提出，此演算法與 Random Forest 十分相似，都是由許多決策樹所組成。與 Random Forest 主要的差異在，Random Forest 是以 Bootstrap 隨機對特徵採樣，作為決策的分裂點，而不是計算最優的相關性的節點進行分裂，之後再基於 entropy 信息增益、gini 基尼係數之類的原

則，選擇一個最佳的特徵值分支屬性。范琪[23]在智能椅活動監控成效評估實驗中，將壓力傳感器採集到的原始數據，使用機器學習分類演算法進行活動分類分析，結果以極限隨機樹的分類精準度高達 98%以上為最佳。在國外相關的研究中，以疾病預測分類於心血管疾病預測系統[24]、在腦腫瘤分類研究中，都使用 **Extremely Randomized Trees** 演算法得到顯著的預測結果，而對於腦腫瘤分類的成效更勝於 **Random Decision Forest** 過去在這個領域的表現[25]。**Support Vector Machine (SVM)** 模型是將訓練資料集中不同類別的各個資料投射為高維特徵空間中的點，透過找到一個超平面將不同屬性類別的點能被盡可能的間隔分開，並使得不同類別之間的邊界間隔距離最大化。接著將測試資料對映到同一個空間中，依照它落在分隔開的哪一側來預測其類別。陳昱瑾[9]針對電商 APP 的使用者／商品／時間特徵來做消費預測，藉以提升 APP 推薦系統的系統效能與推薦效果。實驗結果使用 **RBF** 核函數 **SVM** 預測效果較好且相當，但因 **SVM** 模型預測運算費時，相較之下結構學習法較可以滿足推薦的即時性。**SVM** 分類模型訓練與測試所需時間較長，為了解決這個費時問題，楊鎧謙[15]對 **Google Map POI** 標籤的研究中，嘗試採用 **KDE(Kernel Density Estimation)+SVM** 的混合模型進行多標籤預測，實測時間較單純的 **SVM** 分類模型幾乎快一倍，而預測效果以 **Micro-F1 score** 評估，在大小分類都只略低單純的 **SVM** 模型約 3~7 個百分比。以上分類器是目前對常見的模型，多重標籤分類模型通常採用單一類別模型進行建模和訓練，之後再透過整合每個單一類別模型的機率，就可以達成多重標籤分類任務，因此本研究將嘗試採用多重分類器進行實驗。

三、研究方法

建立提問問題多重標籤分類模型，主要是資料取得與彙整、多重標籤標記、問題內容中文斷詞、問題內容萃取特徵值，以及建模預測四個部分進行說明。其整體概念如圖所示。

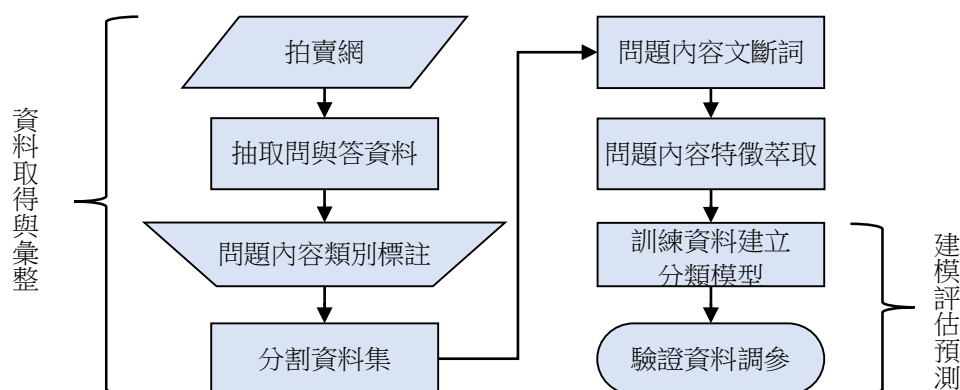


圖 1 本研究方法流程示意圖

(一) 資料取得與彙整

(1) 取得資料與標記

本研究以露天拍賣網客戶對商品的提問內容做為數據來源，主要針對網站首頁分類目錄中，排除成人專區共計 22 類商品類別熱銷商品。針對客戶提問內容進行人工檢閱分類／歸類。本研究參考相關網路購物問題種類，以及實務分析經驗，將客戶對商品提問內容，劃分為十類細分類，並將其歸納為五類粗分類，詳如表 2 所示。逐筆檢閱提問內容後至少須歸屬但不限於一類細分類，再以細分類對應出歸屬之大分類。本研究所設計的粗細分類是為了驗證模型對分類數量的學習效果。

表 2 粗細分類表

五類粗分類	十類細分類	相關項目
金流	發票收據	發票、收據、統一編號、發票抬頭
	價格	確認金額、運費、折扣
	付款方式	付款方式、變更付款方式、退款進度
商品	商品相關	商品規格、使用範圍、保固、包裝、材質、內容物、使用方式、贈品、維修、類似替代品、相關配件、客製化、訂做
	到貨出貨日期	出貨相關、售後查詢配送進度、何時貨到客端
物流	庫存	有沒有貨、何時進貨入庫
	配送方式	面交、取貨方式、自取、變更配送地址、回收
訂單	確認訂單	已下單、確認、修改、取消、退換貨、評價評分、露露通（私訊工具）
	下標相關	下單、加購、合購、預購、平台操作問題、其他賣場、單買
非相關	非相關	問題不完整、非提問

(2) 分割資料集

考量不同商品類別造成客戶提問內容差異化的可能性，為避免各資料集因不同商品類別分配不均影響建模預測結果，本研究將 22 類商品客戶提問資料，依照客戶提問時間先後順序，各自分割為 70%訓練用、10%驗證用、20%測試用，其中 10%的驗證及作為模型調參之使用。

1. 顧客提問之內容中文斷詞

本研究使用 Jieba 演算法對中文資料進行詞語的斷詞。在實驗初期發現，同時建立裝載使用者自訂辭典，以提升斷詞準確效果；並排除標點符號、英文數字字元，以避免因斷詞後字詞數量過多影響分析效率。

2. 問題文本特徵轉換

針對客戶提問內容的特徵值轉換，本研究採用 TF-IDF 與 embeddings 方法，說明如下：

(1) TF-IDF

TF-IDF 是基於頻率的方式將文本轉換，在電商提問的句子，特性用詞有極高機率是特定分類。以訓練資料集斷詞後的字詞作為單一特徵，即計算 TF-IDF 特徵矩陣，除了使用最大詞彙數量，並依據不同詞彙數設定進行多組特徵值實驗效果比較。根據訓練資料集獲得的字詞特徵，分別針對訓練集、驗證集和測試集，轉換為 TF-IDF 特徵矩陣，供後續模型訓練使用。

(2) Word Embeddings

Word embeddings 可以估算其上下文特性，而非只有詞頻，所以在電商提問句子中，常常會出現特性的商品名稱等，所以採用 word embeddings 方法可以有效的辨識出專有名詞的句型。仍是以訓練資料集作為語料庫進行訓練，本計畫分別採用 Word2Vec、FastText、Doc2Vec 詞向量模型進行詞向量訓練。主要的做法是將一整篇的所有字詞的詞向量作加總，因此，每一筆提問的向量就可以產生，即將內容的每個字詞對應的 word embeddings 進行總和計算。

(二) 建模評估預測

在多標籤分類模型部分，本研究運用 Extra Trees、Logistic Regression、Random Forest、SVM 四種機器學習分類演算法進行模型建立。所使用的機器學習演算法將分別對 5 分類與 10 分類任務進行模型訓練。以驗證集進行分類預測後，其分類效果作為最佳參數設定選擇，並以最佳參數設定對測試集進行預測，並評估分類效果。

(三) 評估指標

為了驗證模型在所設計的問題類型辨識效果，以下將針對每個實驗的預測結果，進行評估成效優劣，其評估指標說明如下：

各項評估指標計算方法，統一假設以變數 Q 代表預測筆數、變數 T 代表問題類別數（十類細分類則 $T=10$ 、五類大分類則 $T=5$ ），第 q 筆預測資料的第 t 項問題類別之預測結果計數分別以 $TP_{q,t}$ 、 $FP_{q,t}$ 、 $FN_{q,t}$ 、 $TN_{q,t}$ 表示，其中： TP : True Positive；預測正確，實際提問類別和預測提問類別都是 Yes； FP : False Positive；預測錯誤，實際提問類別是 No、預測提問類別是 Yes； FN : False Negative；預測錯誤，實際提問類別是 Yes、預測提問類別是 No； TN : True Negative；預測正確，實際提問類別和預測提問類別都是 No。

1. 準確率 (Accuracy)

本研究為多重標籤辨識，一筆預測資料中的每一項問題類別必須預測正確為 $TP_{q,t}$ 或 $TN_{q,t}$ 才視為該筆資料預測正確，則第 t 項問題類別的準確率及整體準確率計算如下：

$$Accuracy_t = \frac{\sum_{q=1}^Q TP_{q,t} + TN_{q,t}}{Q}$$

$$Accuracy = \frac{\sum_{q=1}^Q \begin{cases} 1 & \text{if } (\sum_{t=1}^T TP_{q,t} + TN_{q,t}) = T \\ 0 & \text{otherwise} \end{cases}}{Q}$$

2. 精確度 (Precision)

精確度是模型所有的預測值為 Yes 的部分裡，預測正確的效果。但 Precision 並無法反映有多少實際值為 Yes 的樣本被預測錯誤，所以通常不會單獨使用 Precision 指標進行預測效能評估。多重標籤辨識預測第 t 項問題類別的精確度為 $Precision_t = \frac{\sum_{q=1}^Q TP_{q,t}}{\sum_{q=1}^Q TP_{q,t} + FP_{q,t}}$ ，整體精確度計算公式如下。

$$Precision = \frac{\sum_{q=1}^Q \sum_{t=1}^T TP_{q,t}}{\sum_{q=1}^Q \sum_{t=1}^T TP_{q,t} + FP_{q,t}}$$

3. 召回率 (Recall)

又稱 True Positive Rate 或 Sensitivity。召回率是模型所有的實際值為 Yes 的部分裡，被預測正確的程度。和 Precision 相反，Recall 無法反映有多少實際值為 No 的樣本被錯誤預測成 Yes，所以通常也會和其他指標並行評估預測效能。多重標籤辨識預測第 t 項問題類別的召回率為 $Recall_t = \frac{\sum_{q=1}^Q TP_{q,t}}{\sum_{q=1}^Q TP_{q,t} + FN_{q,t}}$ ，整體召回率計算公式如下。

$$Recall = \frac{\sum_{q=1}^Q \sum_{t=1}^T TP_{q,t}}{\sum_{q=1}^Q \sum_{t=1}^T TP_{q,t} + FN_{q,t}}$$

4. F1-Score

F1-Score 是將精準度與召回率合併評估的指標，即精準度與召回率的調和平均值，所以理論上如果 F1-Score 比較高的時候，代表 Precision 和 Recall 也會比較高。本研究以 F1 做為特徵值萃取方法、分類模型、參數設定各種組合後效能評估指標。多重標籤辨識預測第 t 項問題類別的 F1 為 $F1_t = 2 \times \frac{(Precision_t \times Recall_t)}{(Precision_t + Recall_t)}$ ，整體 F1 計算公式如下。

$$F1 = 2 \times \frac{(Micro - Precision \times Micro - Recall)}{(Micro - Precision + Micro - Recall)}$$

四、實驗數據與結果

(一) 實驗資料

本研究針對露天拍賣網 22 類熱銷商品，共爬取 2,108 項商品、84,630 筆發問內容，接續對提問內容進行人工檢閱分類標記，自最新發問內容往回標記，總計共標記 5,760 筆提問資料。人工標記後進行資料集分割，各資料集之問題分類佔比分佈如表 3 和表 4 所示，各資料集客戶提問歸屬類別數分佈如表 5 所示，屬於多標籤問題筆數佔比約 20%。Jieba 斷詞後各資料集字詞數如表 6 所示。

表 3 十分類各資料集之問題分類佔比分佈表

問題資料集	發票收據	商品相關	到貨出貨日期	庫存	非相關	價格	確認訂單	配送方式	下標相關	付款方式	合計
訓練	65	1,474	266	1,457	55	725	198	304	450	137	5,131
驗證	18	202	37	238	-	97	19	43	58	12	724
測試	20	390	81	470	10	186	32	89	112	26	1,416
合計	103	2,066	384	2,165	65	1,008	249	436	620	175	7,271

表 4 五分類各資料集之問題分類佔比分佈表

問題資料集	非相關	物流	金流	商品	訂單	合計
訓練	55	1,889	888	1,474	641	4,947
驗證	-	302	124	202	77	705
測試	10	589	218	390	143	1,350
合計	65	2,780	1,230	2,066	861	7,002

表 5 各資料集客戶提問歸屬類別數分佈表

歸屬類別數	十分類				五分類			
	訓練	驗證	測試	合計	訓練	驗證	測試	合計
1	3,107	454	929	4,490	3,197	465	964	4,626
2	783	101	181	1,065	763	96	171	1,030
3	120	20	31	171	72	16	12	100
4	22	2	8	32	2	-	2	4
5	2	-	-	2				
合計	4,034	577	1,149	5,760	4,034	577	1,149	5,760

表 6 各資料集之字詞數統計分佈表

資料集	字詞總數	平均字數	中間值	眾數	標準差	變異數	標準誤
訓練	4,754	10.52	8	5	9.47	89.60	0.149
驗證	1,206	9.93	8	4	8.28	68.56	0.345
測試	2,029	9.59	7	4	7.81	60.92	0.23

(二) 實驗設計

在實驗設計部分，如表 7 所示，本研究嘗試採用不同大小之 TF-IDF 詞彙數，根據訓練集的資料，最多為 4,754 個單詞，最少使用 50 個單詞。而 Word Embeddings 部分，設計 20 組向量維度設定，以 50 為級距，最大為 1,000。而 4 種機器學習分類演算法之超參數設定值，可以根據表 8 所示。

表 7 詞彙數和詞向量維度設定

向量化方法	設定值
TF-IDF	50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 4754
Word Embedding	50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000

表 8 分類演算法參數設定

分類演算法	參數名稱	參數設定
Extra Trees	決策樹數量 (n_estimators)	100, 200, 300, 400, 500, 750, 1000
	正規化(penalty)	l2
Logistic Regression	求解方式(solver)	lbfgs、newton-cg
	正規強度的倒數(C)	1, 10, 100, 1000, 10000, 100000, 1000000
Random Forest	子樹數量(n_estimators)	100, 200, 300, 400, 500, 750, 1000
SVM	核函數(kernel)	rbf、linear、poly、sigmoid
	scale'(gamma)	0.01, 0.1, 1, 10, auto

※ 其餘未列參數均使用預設值。

(三) 多標籤分類之實驗結果

實驗最後以驗證資料集預測結果進行最佳參數設定選擇，並以最佳參數設定對測試資料集預測結果進行分類預測成效評估。測試集十類標籤分類預測成效評估，各項指標彙整如表 9 所示。整體以 TF-IDF 搭配 Extra Trees 預測效果最佳，F1 分數達 0.82846；Word Embeddings 則以 SVM 演算法有較好的預測效果。

表 9 測試集之十類標籤辨識分成效表

特徵法	分類模型	最佳參數	Accuracy	Precision	Recall	F1
Doc2Vec	Extra Trees	特徵=1000, estimators=400	0.6214	0.9146	0.63560	0.7500
	Logistic Regression	特徵=800, solver=newton-cg, C=1	0.6736	0.8519	0.7797	0.8142
	Random Forest	特徵=750, estimators=400	0.6275	0.9162	0.6483	0.7593
	SVM	特徵=550, kernel=linear, gamma=auto	0.6832	0.8560	0.7811	0.8168
FastText	Extra Trees	特徵=750, estimators=300	0.6171	0.9166	0.6285	0.7457
	Logistic Regression	特徵=600, solver=newton-cg, C=100	0.6658	0.8266	0.7910	0.8084
	Random Forest	特徵=750, estimators=200	0.6240	0.9042	0.64620	0.7537
	SVM	特徵=850, kernel=linear, gamma=auto	0.6710	0.8558	0.7670	0.8089
TF-IDF	Extra Trees	特徵=400, n_estimators=750	0.7172	0.8820	0.7811	0.8285
	Logistic Regression	特徵=3500, solver=newton-cg, C=10	0.6954	0.8720	0.7698	0.8177
	Random Forest	特徵=450, estimators=1000	0.7215	0.8901	0.7719	0.8268
	SVM	特徵=3500, kernel=linear, gamma=auto	0.7102	0.8911	0.7691	0.8256
Word2Vec	Extra Trees	特徵=850, n_estimators=500	0.6310	0.9138	0.6511	0.7604
	Logistic Regression	特徵=650, solver=newton-cg, C=10	0.6641	0.8432	0.7712	0.8056
	Random Forest	特徵=850, estimators=400	0.6414	0.9098	0.6695	0.7714
	SVM	特徵=550, kernel=linear, gamma=auto	0.6728	0.8551	0.7754	0.8133

如表 10 數據所示，測試資料集五類標籤分類預測成效整體仍以 TF-IDF 搭配 Extra Trees 的預測效果最佳，F1 分數達 0.8734；而 Word Embeddings 則以搭配 Logistic Regression 的預測效果較好。

表 10 測試集之五類標籤辨識分成效表

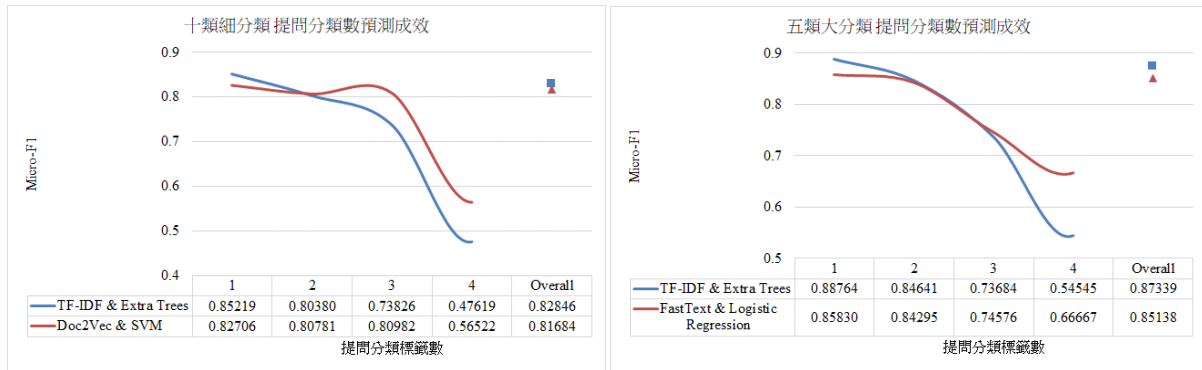
特徵法	分類模型	最佳參數	Accuracy	Precision	Recall	F1
Doc2Vec	Extra Trees	特徵=900, estimators=1000	0.7041	0.9029	0.7437	0.8156
	Logistic Regression	特徵=850, solver=newton-cg, C=1000	0.7285	0.8668	0.8244	0.8451
	Random Forest	特徵=1000, estimators=750	0.7032	0.8987	0.7489	0.8170
	SVM	特徵=750, kernel=linear, gamma=auto	0.7206	0.8728	0.8133	0.8420
FastText	Extra Trees	特徵=900, estimators=1000	0.6928	0.8943	0.7333	0.8059
	Logistic Regression	特徵=850, solver=newton-cg, C=1000	0.7372	0.8584	0.8444	0.8514
	Random Forest	特徵=1000, estimators=750	0.7084	0.8919	0.7578	0.8194
	SVM	特徵=650, kernel=linear, gamma=auto	0.7163	0.8689	0.8148	0.8410
TF-IDF	Extra Trees	特徵=4500, n_estimators=300	0.7807	0.8944	0.8533	0.8734
	Logistic Regression	特徵=4754, solver=newton-cg, C=10	0.7502	0.8867	0.8289	0.8568
	Random Forest	特徵=350, estimators=300	0.7641	0.8764	0.8348	0.8551
	SVM	特徵=4754, kernel=linear, gamma=auto	0.7572	0.8895	0.8289	0.8581
Word2Vec	Extra Trees	特徵=450, n_estimators=100	0.6928	0.9044	0.7356	0.8113
	Logistic Regression	特徵=750, solver=newton-cg, C=10000	0.7102	0.8393	0.8356	0.8374
	Random Forest	特徵=550, estimators=300	0.7041	0.9041	0.7541	0.8223
	SVM	特徵=500, kernel=linear, gamma=auto	0.7111	0.8653	0.8089	0.8361

(四) 結果分析

1. 從客戶提問分類標籤數面向看預測成效

十類細分類、五類大分類分別以 TF-IDF 和 Word Embeddings 演算法預測效果最佳的分類模型預測結果進行比較。如圖 2 所示，單一類別問題被成功預測的效果較高，F1 分

數隨著問題歸屬類別數增加而遞減；而比起 TF-IDF 演算法，對於多重標籤問題辨識，以 Word Embeddings 萃取特徵值的分類模型反而得到了較好的預測效果。

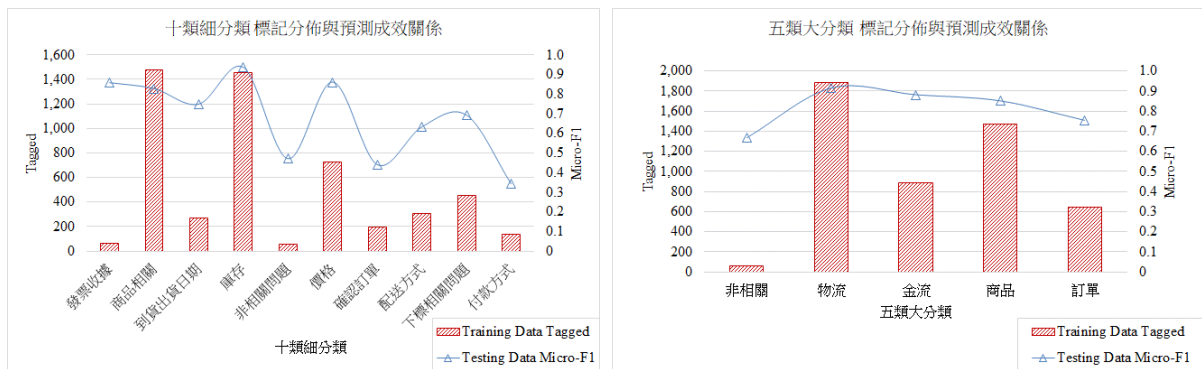


a. 十類細分類不同特徵值萃取法最佳組合比較 b. 五類大分類不同特徵值萃取法最佳組合比較

圖 2 各提問分類標籤數預測成效

2. 從各分類標籤面相看預測成效

以整體預測效果最好的 TF-IDF 搭配 Extra Trees 的預測結果，從分類標籤各自的預測效果來看，對照訓練資料集各分類標籤標記數量，由圖 3 可看出，預測效果較差的分類標籤，對應到訓練資料集中該分類標籤標記數量也相對的較少。



a. 十類細分類 b. 五類大分類

圖 3 標記分佈與預測成效關係

3. 預測結果差異性比較

十類細分類以 TF-IDF、Word Embeddings 特徵值最佳預測效果進行比較。各分類標籤預測優劣比較如表 12 所示，其中 TF-IDF 特徵值預測效果較差的問題類別，Word Embeddings 特徵值反而有較好的預測效果。

表 12 十類細分類各分類標籤預測效果差異比較表

	發票收據	商品相關	到貨出貨日期	庫存	非相關問題	價格	確認訂單	配送方式	下標相關問題	付款方式
TF-IDF	0.8571	0.8269	0.7482	0.9349	0.4706	0.8603	0.4400	0.6323	0.6919	0.3429
Word Embeddings	0.7692	0.8361	0.7733	0.9280	0.5600	0.8362	0.5455	0.6282	0.5864	0.4167

（五）討論

從模型導入實際運用層面來看，首先需要檢視的是：現有分類模型問題類別是否足夠、是否適用，因為每個購物網站的營運模式、作業流程都不盡相同，對於客戶提問分類層級多寡也因其分析需求而有所不同，本研究建議，在模型預測架構實際導入運用時，以本研究提出之十類細分類作為問題類別基礎，依照各平台實務狀況增減調整分類項目，或是擴增另一分類更細微的微分類層級進行運用。從組織層級來區分，十類細分類預測結果，適合提供給客服人員、執行層級，經過分類模型自動預測出客戶問題類型，快速理解客戶需求、縮短人工分派案件時間，進而透過平台系統的整合串接，提供客服人員該問題類別回覆話術參考範本，在客服人員確認調整回覆內容的同時，亦可將正確的提問分類回饋給預測模型，達到校正效果。客服單位透果此模型可以立即發現有疑慮的商品或營運異常狀況，警示相關單位介入處理，有助於在第一時間使得問題獲得解決。

五、結論與建議

本研究整體以 TF-IDF 搭配極限樹得到較好的分類預測效果。對於粗略的五類大分類預測，*F1* 分數可達到 0.8734；而分類較精準、較難辨識準確的十類細分類預測結果也有 0.8285。單一標籤問題預測仍以 TF-IDF 特徵值有較高的 *F1* 分數，而多標籤問題部分，則以 Word Embeddings 特徵值的預測成效較好。由於無法取得現成資料進行實驗，本研究自行撰寫爬蟲程式取得客戶提問文本資料，依實務經驗定義客戶問題分類並人工檢閱標記歸屬類別。本研究嘗試以既有的特徵值萃取演算法及分類模型進行模型建立，實驗結果也顯示分類效果可達八成以上。對此，本研究也創造了電商客戶提問多重標籤辨識資料集，爾後能夠做為相關電商問題之多重標籤研究的參考基礎。

本研究同時使用 TF-IDF 以及 Word Embeddings 兩種特徵值演算法進行實驗比較，雖然整體預測效果以 TF-IDF 較好，但對於多重標籤問題，則以 Word Embeddings 能帶來較高的預測成效，而 TF-IDF 則能對單一標籤案例獲得較佳成效。本研究的成果是在特定拍賣網站下所得分類效果，因此無法直接採用本研究所出的標記原則，但其電商提問多重標籤之標記概念相似，針對特定電商設計特定類別將可以有效建立適合的分類模型。由於缺乏大量電商相關語料庫可供 Word Embeddings 訓練，導致其預測效果無法顯著提升。在未來研究方面，可以採用預訓練的 Word Embeddings 模型，進行詞向量微調，期望有效提升預測效果。例如使用 BERT 的預訓練模型，來強化文本特徵可用性。

參考文獻

- [1] 葉靜縈,「客服對話式文本資料之多主題標籤辨識研究—以某壽險公司為例」,輔仁大學統計資訊學系應用統計碩士在職專班,碩士論文,2019。
- [2] 鄭哲明,「應用資料探勘於顧客問題自動分類之研究-以自來水公司民眾意見信箱為例」,國立交通大學理學院科技與數位學習學程,碩士論文,2015。
- [3] 施瑋蘋,「運用資訊科技輔助客服系統應答最佳化之研究」,國立臺北教育大學資訊科學系碩士班,碩士論文,2019。
- [4] 蔡佩珊,「建置自動化客服回覆機制之聊天機器人」,國立臺北科技大學工業工程與管理系,碩士論文,2019。
- [5] 林益陞,「電子商務網站服務品質、顧客滿意與顧客忠誠之研究 - 以 PChome 為例」,朝陽科技大學應用英語系,碩士論文,2018。
- [6] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, Sentiment analysis based on improved pre-trained word embeddings, *Expert Systems with Applications*, 117, 2019, pages 139-147.
- [7] B. Guo, C. Zhang, J. Liu, and X. Ma, Improving text classification with weighted word embeddings via a multi-channel TextCNN model, *Neurocomputing*, 363, 2019, pages 366-374.
- [8] M. Aydogan, and A. Kaci, Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification, *Physica A: Statistical Mechanics and its Applications*, 541, 2020.
- [9] J. A. Gonzalez, L.-F. Hurtado, and F. Pla, Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter, *Information Processing & Management*, 57, 2020.
- [10] N. Desai, and A. Mahendran, classification of semantically similar question pairs using machine learning. *International Journal of Computing and Digital Systems*, 9, 2020.
- [11] S. Sulaiman, R. A. Wahid, A. Ariffin, and C. Z. Zulkifli, question classification based on cognitive levels using linear SVC. *Test Eng. Manag*, 83, 2020, pages 6463-6470.
- [12] H. Rokham, G. Pearlson, A. Abrol, H. Falakshahi, S. Plis, and V. D. Calhoun, Addressing inaccurate nosology in mental health: A multi label data cleansing approach for detecting label noise from structural magnetic resonance imaging data in mood and psychosis disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2020.
- [13] C. Zhou, and J. Yuan, Multi-label learning of part detectors for heavily occluded pedestrian

- detection. *In Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- [14] 黃冠傑, 「多標籤分類方法應用於 PTT 資料」, 淡江大學統計學系應用統計學碩士班, 碩士論文, 2019。
- [15] 楊鎧謙, 「On Large-Scale Multi-Label Classification for POI Tagging」, 國立中央大學資訊工程學系, 碩士論文, 2017。
- [16] 何欣儒, 「網購低溫食品快速送達與超商取貨服務對購買行為影響之研究:二元 logistic Regression 模式之建立與應用」, 銘傳大學企業管理學系, 碩士論文, 2018。
- [17] 葉丞峻, 「適用於分類變數資料的二元不平衡資料自動分類系統」, 淡江大學統計學系碩士班, 碩士論文, 2017。
- [18] 呂育如, 「應用資料探勘分類超音波病人檢查時間」, 中原大學工業與系統工程研究所, 碩士論文, 2019。
- [19] 高詠富, 「法則庫方法萃取財務報表詐欺規則」, 國立雲林科技大學資訊管理系, 碩士論文, 2019。
- [20] 陳希聖, 「使用隨機森林實現超音波多特徵脂肪肝疾病分類」, 國立臺灣大學應用力學研究所, 碩士論文, 2019。
- [21] 邱惠君, 「應用文字探勘技術進行網路霸凌預測」, 國立交通大學管理學院資訊管理學程, 碩士論文, 2019。
- [22] P. Geurts, D. Ernst, and L. Wehenkel, Extremely randomized trees, *Machine Learning*, 63, 3-42, 2006.
- [23] 范琪, 「具有動作辨識之智慧椅研發」, 元智大學通訊工程學系, 碩士論文, 2018。
- [24] R. Shafique, A. Mehmood, S. Ullah, and G. S. Choi, Cardiovascular Disease Prediction System Using Extra Trees Classifier. *Research Square*, 4, 2016.
- [25] M. Goetz, C. Weber, J. Bloecher, B. Stieltjes, H. P. Meinzer, and K. Maier-Hein, Extremely randomized trees based brain tumor segmentation, *Proceeding of BRATS challenge-MICCAI, 006-011*, 2014.

電子郵件輔助寫作

Email Writing Assistant System

張俊盛 Jason S. Chang

楊馨瑜 Ching-Yu Yang

彭冠復 Guan-Fu Peng

國立清華大學資訊工程學系 Department of Computer Science, National Tsing Hua
University

elmon@nplab.cc, Jason@nplab.cc, chingyu@nplab.cc

摘要

在本論文中，我們介紹一種在書寫特定目的之郵件時提供精準寫作建議的方法。在我們方法中，電子郵件先依照意圖分門別類，然後從每個類別的電子郵件中，擷取常用片語，並且講類似的片語，聚集成為一組一組的聚叢。執行時，系統檢視使用者輸入，比對信件類別的片語聚叢，以提供寫作建議。我們將此一方法，加以實際製作，成為一套電子郵件寫作建議系統 *EmailDr*。

Abstract

We introduce a method for learning to provide writing suggestions for writing an email for a specific purpose. In our approach, emails are divided into categories aimed at finding common phrases for making suggestions. The method involves automatically extracting common phrases for every email category, and automatically clustering phrases for more effective suggestions. At run-time, the system accepts user's input and matches the input with common phrases for the specific purpose, to offer suggestions for what to write next. We have implemented the propose method and present a prototype email suggestion system, *EmailDr*.

關鍵詞：自然語言處理, 智慧寫作, 電子郵件

Keywords: NLP, Smart Compose, Email

一、緒論

隨著科技的日新月異，人們愈加依賴電子郵件做為通訊的方式，因此支援電子郵件寫作的工具也愈來愈多，如 Grammarly (<https://www.grammarly.com/>)、Gmail (<https://www.gmail.com>)、Whitesmoke (<http://www.whitesmoke.com/>) 等等。目前既存的電子郵件寫作工具，多半協助使用者，改正拼字與文法錯誤，並且提供改善寫作風格建議，這些功能都與自然語言處理的技術息息相關，可以做到相當好的效果，然而如何寫作某一特定性質的書信，更是使用者迫切需要的功能，目前這方面還有相當改進的空間。

近年，Gmail 已經開始在使用者書寫郵件時，建議使用者往下可以使用的片語，然而這些建議並未考慮使用者書信的類別。對於使用者寫作的特定類別的書信，未能直接提供對應的寫作建議。在網路上，有許多網站提供不同目的的書信範本，給使用者參考。然而使用者不容易一面寫作，一面查考書信範本的資訊。此外，在使用者尚未開始寫作時，Gmail 無法提供寫作建議。

因此本論文提出方法，利用既有的書信類型與範本，統計分析常用詞語，在使用者寫作時，比對使用者已經輸入的用詞，建議適當的常用片語，讓使用者可以寫出適當的書信。

舉例來說，寫作一封邀請信時，當使用者輸入 “*I would like to*” 時，最好的建議，可能是 “*I would like to invite you to . . .*”。直覺上，為了提供適當的寫作建議，我們可以分析出現在「邀請類」的多篇書信範本的高頻的片語。如此一來，我們可以提供比較正確的寫作預測，顯示比較適合的片語。

我們提出一套新的電子郵件寫作輔助系統 *EmailDr*，可以自動地學習如何接受使用者寫作的未完成句（例如 *I would like to*），來預測特定類別書信的接續片語（例如 *I would like to invite you to*），如圖一所示。*EmailDr* 藉由分析 Enron Email Dataset 中的資料，並計算個別類別書信的高頻片語，已得到適合提供給使用者參考的片語。此外，此一系統也可以透過互動的方式，協助學習者，檢視不同類別之書信常用詞彙與片語，提升學習者的寫作能力。

二、相關研究

近年來為了資訊交流愈加方便，拼字文法校正、建議用詞等等在寫作輔助系統上的功能，是自然語言處理中熱門的研究領域。在拼字文法校正，Grammarly 以及 Linggle 在這方面做了許多研究，而建議用詞的功能則在 EmailPro 以及 Gmail 之中有被提供，EmailPro 是在以不考慮類別下，提供使用者建議用詞，而 Gmail 的功能較偏向精準用詞，所以當建議用詞為高成功率的搭配詞，才會被顯示給使用者作為利用，因此在本論文中，會針對電子郵件的建議用詞上，提出將類別納入考量的方法。

學者針對建議詞彙的生成，也提出不同的方法在這過程之中。[1]先將文法模式取出後，利用 Macmillan English Dictionary 的資料，去做建議詞彙的優先順序。[2] 則是利用機器學習，針對來信的內容去做分析，藉此提供回信的建議用詞。

本文我們是參考 Smadja 演算法去生成搭配詞，設計如何過濾出較佳的詞彙束方式。搭配詞的篩選通常是利用統計的方式從語料庫中選取候選的詞彙，再判斷這些候選的搭配詞何者是合理的。計算兩個單詞相隔的距離在資料中出現的次數，可有效的協助篩選良好的詞彙束

與我們最相關的研究，[3] 也是利用 Smadja 演算法，這個演算法有兩個假設：搭配詞的出現頻率遠高於非搭配詞，以及搭配詞出現的次數在詞與詞之間的距離上的分布有峰值。會先計算任意兩個單字在特定距離的時候出現的次數此為 特定距離時的 skip bigram，進而選出次數最高的 skip bigram 作為最終的搭配詞，而與[3]在過程中的差別是預先篩選搭配詞，利用被預期的搭配詞排序先後，以及實際的搭配詞排序先後，相除得到比例去得到它認為較好的搭配詞，再進行詞彙束的篩選。但我們的文章是直接針對詞彙束做篩選，以 NLTK 中的停用詞作為功能詞，先留下有功能詞的資料，再利用我們生成的搭配詞進行查詢以提取我們要的詞彙束。此外，我們也利用 Linggle 的相似詞功能，去擴增資料，藉以提升輸出的效果。

相對於以往的郵件寫作輔助系統，在本文中會提出一套寫作輔助系統，利用已經標記好的片語整理成 N-連詞(N-gram)，將搭配詞與現有的樣板資訊相互比對，篩選出效果較佳的詞彙束，套用至系統上，幫助使用者書寫郵件。

三、研究方法

本論文的目標為開發一個電子郵件輔助寫作系統，提供不同類別的建議用語。研究方法我們依下列階段進行，分別為（一）資料前處理，（二）生成搭配詞，（三）篩選良好的詞語束，（四）建立預測模型。

（一）資料前處理:

利用 spaCy 套件解析 WriteExpress (<https://www.writeexpress.com/>)

資料的句子，得到對應的詞性 (part-of-speech tag)、名詞片語 (noun chunk) 等資訊，再以名詞片語為 token 單位來切 Ngram (n 從 3~5)。之後，再根據 Ngram 中的每個 token 詞性來找出相對應的 grammar pattern (這邊的 grammar pattern 會做 lemmatization)，詳細步驟如下：

1. 辨別句子之詞性(Part-of-Speech Tag)

句子 (e.g., "My husband and I will be delighted to be part of the celebration.") 經由 spaCy 產生資訊，例如詞性為 [DET NOUN CCONJ PRON VERB AUX ADJ PART AUX NOUN ADP DET NOUN PUNCT]，名詞片語為 [My husband, I, part, the celebration]

2. 產生 Ngram 資料

將句子，以三個到五個詞為單位，產生 Ngram，以上述句子為例，切出來的 Ngram 有[My husband and, My husband and I, My husband and I will, ...]等等組合。

3. 將 Ngram 轉為針對每組 Ngram 搭配名詞片語的資訊，再轉換成 grammar pattern 形式 (e.g., "My husband and I will "轉換為 "My n. and SOMEONE v." 以及 "My husband and SOMEONE v.")

我們共有 63 種信件類別，其中有 2 類資料短缺，實作時只能排除。這個步驟後產生 4,132,282 種 Ngram，共有 3069634 個 grammar pattern 分布在 61 類中，共有 818,169 種不同的 grammar pattern。

（二）生成搭配詞

研究搭配詞時通常只研究 base word 和 collocate 兩個單詞，可以記錄成(w,w1)，比如 listen 和 music 的 collocation 會表示成 (listen, music)，省略中間的單詞。這種類似 bigram 的形式，稱為 skip bigram。藉由統計 Enron Email Dataset 的 skip-

gram 等資訊，並利用 Smadja 演算法篩選出合理的搭配詞。結果可以用來過濾掉不適合或不完整的 Ngram (使用 詞彙束 Lexical Bundle 的概念，因此在第三步我們將過濾完的 Ngram 稱為 Lexical Bundle)。詳細的步驟如下 (見表一)：

1. 找出 window size = 5 之內的所有 skip-bigram。
2. 計算每個候選搭配詞的頻率、分佈等資訊。
3. 保留符合 Smadja's Algorithm 篩選標準的搭配詞。

left token 左字元	right token 右字元	distance	count
invite	you	1	2087
invite	to	2	2081
invite	attend	3	210
thank	you	1	5861
thank	for	2	3655
thank	your	3	1208
total collocations		969750	

表一、Smadja 演算法生成的搭配詞

(三) 生成詞彙束

利用第二步所產生的搭配詞來過濾第一部產生的 Ngram。因為使用到 lexical bundle 的概念，因此我們稱過濾出的乾淨資料為 lexical bundle。lexical bundle 的概念提到，一組 lexical bundle 通常都擁有 function word (這裡我們使用 nltk 的 stopword 作為 function word)，詳細步驟如下 (可見表二)：

1. 先留下有 function word 的 Ngram，接著找出每個 Ngram 的 skip-bigram。
2. 用第二步產生的搭配詞表查詢，只留下「至少有一個 skip-bigram 為搭配詞表中出現過」的 Ngram。

未過濾的 Ngram	過濾後的 Lexical Bundle
Mr. and Mrs. John Doe	accept your kind invitation to
accept your kind invitation to	to brunch at

to brunch at	the twentieth of
of October at	
the twentieth of	

表二、詞彙束過濾前後的結果

(四) 分析同義詞

由於雖然有些 **pattern** 有不同的用字，用法實際上卻是很相近的（如 “I am glad to” 及 “I am happy to” ），我們利用查詢 Linggle (Linggle.com) 得到相近詞來擴展 **pattern**，讓輸入更容易被對應到 **pattern**。

1. 計算出信件樣板之 關鍵詞

利用 Chi-square test，從 WriteExpress data 得到關鍵詞，在這裡我們取分數最高的前 10 名。

2. 藉由 Linggle 查詢相似詞

利用 Linggle API 查詢，得到各個配對出現在例句的次數，從而得知哪些配對較常一起出現、可能有相近的意思。例如我們從上一步得到三個關鍵詞

“pleased”、 “honored” 及 “sorry”，透過 Linggle API 查詢

“pleased/honored/sorry and pleased/honored/sorry”，便能知道 “pleased” 及

“honored” 較常一起被提及，因此較有可能具有相近的詞義。在這裡，我們取出次數前三名的配對，以及第四名後次數超過 10000 的配對列為相似詞。

3. 擴展片語

將片語中的單字，以其同義詞取代，產生另一組片語，藉此可以擴增片語，形成片語的聚叢。

(五) 統計式預測模型

我們以 **lexical bundle** 為 **knowledge base** 來跟使用者輸入做比對，最主要是比對輸入的最後兩個 **token** 和 **knowledge base** 的每個 **pattern**，最後產生相對應的建議。詳細步驟如下：

1. 輸入句(通常是不完整句)先使用 spaCy 斷詞，再取出最後兩個 **token**。
2. 首先比較 **pattern** 的前兩個 **token** 和輸入句的最後兩個 **token**，相同則 **hit**。
3. 若沒有，則比較輸入的最後一個 **token** 和 **pattern** 的第一個 **token**，相同則 **hit**。

4. 若沒有，則比較輸入的倒數第二個 token 和 pattern 的第一個 token，相同則 hit。
5. 從 hit 中按照頻率取出前五名成為給使用者的建議。

四、實驗

(一)、資料集

透過網路爬蟲收集 WriteExpress(<https://www.writeexpress.com/>) 所提供的各項類別的樣本，並依照信件樣本、例句、搭配用語作為分類，另一個則是 Enron Email Dataset 共五十萬筆的信件內容，這兩項資料皆以 3-grams、4-grams，以及 5-grams 的方式將文字做切割，再利用這些資料進行分類的預測模型訓練。

(二) 實驗模型與結果

1. 利用範例信件進行評估準確率

我們使用《How To Say It》一書中的範例信件來評估系統，書中共有 50 類書信，其中有 29 類與我們的系統相同，因此我們取這些類別中的每篇範例信件來進行評估。評估方式如下：

$$\sum_{i=1}^w \frac{((N_i - n_i + 1)/N_i)}{\text{總字數 } w \text{ (去除標點符號)}}$$

輸入為 token[:i]，系統輸出共輸出 N_i 筆 pattern，其中第 n_i 名 pattern 的 Ngram 與 token[i-2:i+1] 或 token[i-1:i+1] 相同，i 從 1 ~ w，w 為範例信件長度。我們共做了 156 篇測資，去除標點符號及不在 Training corpus 內的字 (out of vocabulary) 後共 12950 字。由於測資的部分並沒有完全對應到 WriteExpress 上所提供的類別，所以單以類別作為評估標準，而測試結果準確率最高的是 response 類別，最低的是 application，下表（見表三）顯示前三高的類別與最後三名的類別。我們認為由於訓練的資料不夠豐富，以及分析測資內容時，有特別針對某些詞語做歸類，會影響由 WriteExpress 資料所統計出的常用搭配詞，進而導致資料不平均，推測藉由其他資料來源，以及擴增特定用語的數量，可以將準確率不佳的類別做程度上的改善。

類別	準確率
response	0.611

love	0.607
appreciation	0.566
refusal	0.422
introduction	0.416
application	0.387

表三、信件類別與準確率

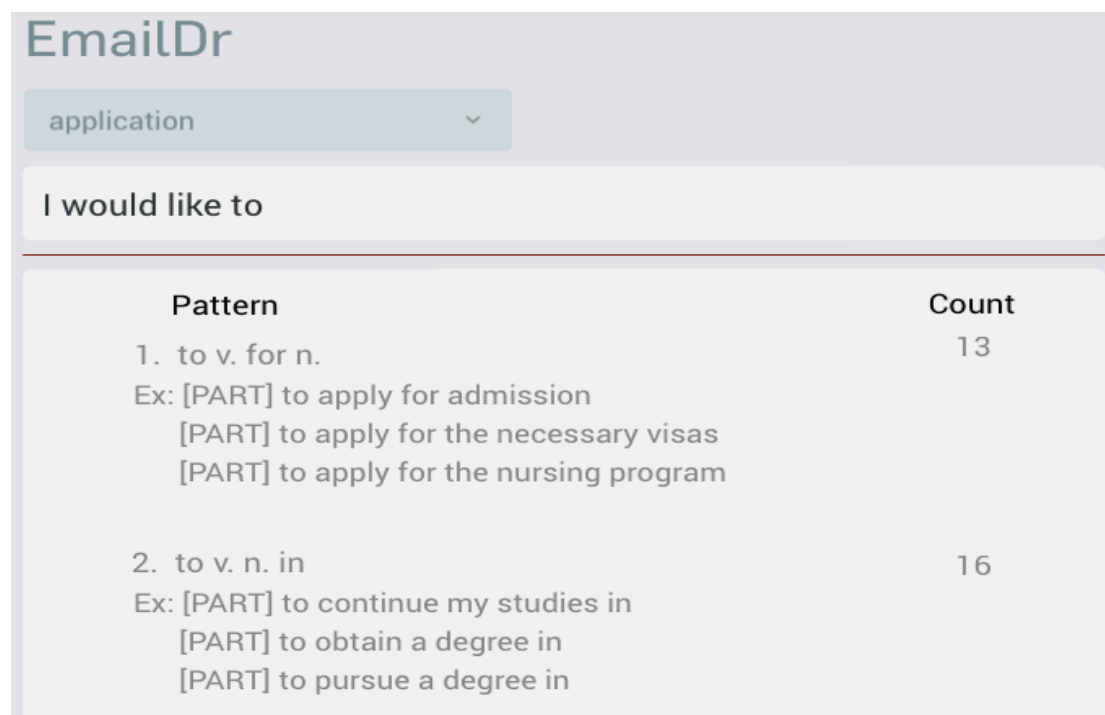
2. EmailDr 與其他系統的比較

(1)EmailDr 與 EmailPro

我們利用 EmailDr 選出類別後所提供的資訊（可參見圖二、圖三），以及圖四，能夠發覺在有類別的選擇之下，可以比較精準而快速的提供建議用語，相反的 EmailPro 呈現的資訊較為繁雜（可見圖四），容易造成查詢者使用上的困擾。

(2)EmailDr 與 Gmail

以一封申請信的 “*I would like to...*” 為例，在我們的系統中，選擇類別後，能夠在開頭輸入少量資訊（可參見圖二），便提供該類別的相關建議用語，而 Gmail 可能無法立即地提供資訊（可參見圖五）。除此之外，我們所提供的資訊可能稍多，較適合作為使用者教學以及學習使用，Gmail 則是比較偏向高成功率搭配詞出現時，才給予使用者建議，使其能降低其錯誤率。



The screenshot shows the EmailDr interface. At the top, the title "EmailDr" is displayed. Below it, a dropdown menu is set to "application". The input field contains the text "I would like to". The results are presented in a table with two columns: "Pattern" and "Count".

Pattern	Count
1. to v. for n. Ex: [PART] to apply for admission [PART] to apply for the necessary visas [PART] to apply for the nursing program	13
2. to v. n. in Ex: [PART] to continue my studies in [PART] to obtain a degree in [PART] to pursue a degree in	16

圖二、EmailDr 選取 Application 類別所提供的建議用語


EmailDr

apology

I would like to

Pattern	Count
1. to v. n. Ex: [PART] to make amends [PART] to repaint your garage door [PART] to help those affected secure new jobs	236
2. like to v. Ex: [VERB] like to apologize [VERB] like to express [VERB] like to meet	19

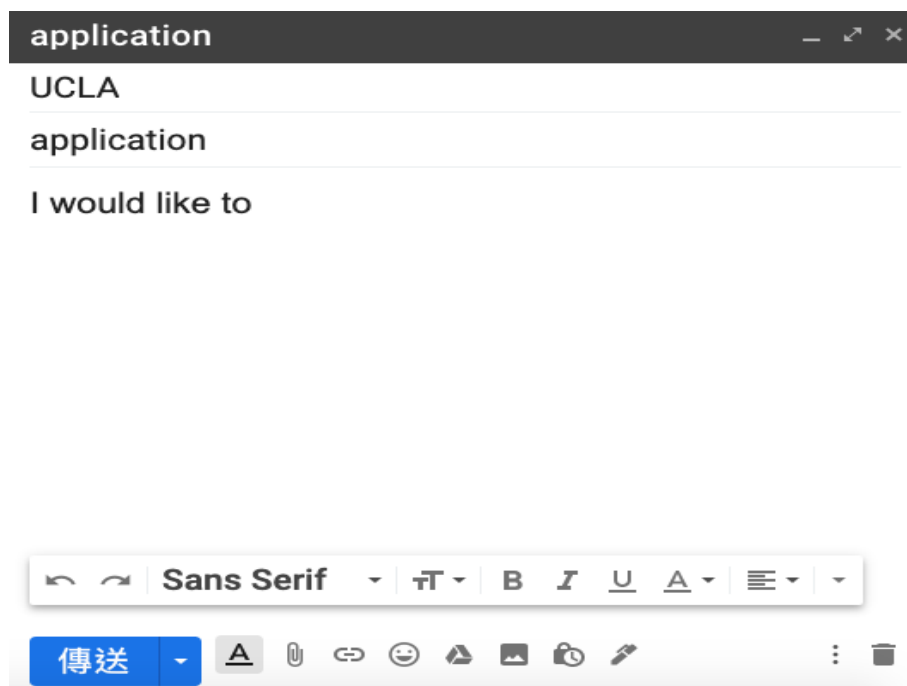
圖三、EmailDr 選取 Apology 的類別所提供的建議用語



I would like to

Pattern	Percentage	Count
I would like to v. I would like to get I would like to have I would like to see	95%	3203
I would like to v. n. I would like to invite you I would like to thank you I would like to ask you	15%	516
I would like to v. the	10%	350

圖四、EmailPro 呈現的建議用語



圖五、在 Gmail 輸入未完成句所呈現的情形

五、結論

本篇論文建立一個電子郵件輔助寫作系統，從資料收集、分析處理、統計頻率，至產生供使用者利用的建議用語。實驗結果顯示，我們的系統在參考類別後所提供的建議用語具有不錯的效果。此外也在少量詞彙資訊輸入時，即時提供建議，即使有些建議可能無法貼近使用者的需求，但還是盡可能達到輔助寫作的作用。

未來，我們將會持續收集與標記信件의 樣板資料，以及利用機器學習的方式使大量的郵件內容可以被自動分類成各個類別。此外，加強信件開頭語的建議用詞，讓使用者能夠更便捷完成一封信件的書寫。

六、參考文獻

- [1] Jim Chang, JS Chang, “WriteAhead2: Mining Lexical Grammar Patterns for Assisted Writing”, Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 106-110. 2015
- [2] Hsu, C. L., Ju, H. H., Wu, Y. H., Peng, H. C., Chen, J. J., Chang, J., & Chang, J.. Computer Assisted English Email Writing System. Proceedings of the

International Conference on Computational Linguistics and Intelligent Text
Processing, CICLing, 2017

- [3] Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, Yonghui Wu, "Gmail Smart Compose: Real-Time Assisted Writing." Knowledge Discovery and Data Mining, KDD, 2019

- [4] Email Pattern dataset source web site: <https://www.writeexpress.com/>

基於 BERT-DAOA 的意見目標情感分析

Aspect-Based Sentiment Analysis Based on BERT-DAOA

陳震瑜 Chen-Yu Chen

國立中央大學資訊工程學系

Department of Computer Science and Information Engineering

National Central University

107522103@cc.ncu.edu.tw

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

社群媒體網站包含了豐富且多樣性的資訊，使輿情分析變成市場調查的方法之一。情感分析是輿情分析的重要一環，目的在取得意見目標網路聲量之外更精準的消費者愛好。然而，即使是句子中包含談論目標，句子的情感類別與文中被評論的目標的情感類別也不盡然一致。因此，本研究應用 Aspect-based sentiment analysis (ABSA)，分析社群網站對意見目標(Opinion Target)的情感類別。在模型的設計上，我們採用 Google 的 BERT[1]作為字詞的嵌入層方法，並參考了 Huan 等人[2]的意見目標情感分類方法，以及 Parikh 等人[3]的自然語言推論的方法，結合兩者來分析意見目標情感分類。實驗結果顯示，在 BERT 之上搭配 Attention-over-Attention (AOA)注意力模型，效能優於 BERT CLS 效能。

Abstract

Social media networks provide rich and diverse information, making opinion analysis and network volume analysis a new method to investigate and understand the market. Sentiment analysis aims to determine the emotional category in a given text. Since there might be several targets being commented on in the text, Aspect-base sentiment analysis (ABSA) has been proposed to explore the sentiment categories of a target in different aspects. In this paper, we explore the idea of ABSA for sentiment analysis of singers on social networks. We utilize

BERT as the embedding layer method for characters and words to explore the relationship between a given sentence and a mentioned target. We consider two attention mechanisms to enhance the performance. The experimental results show that adding the attention layer on top of BERT outperforms the basic BERT-CLS model.

關鍵詞：深度學習，情感分析，意見目標情感分析

Keywords: Deep Learning, Sentiment Analysis, Aspect-based Sentiment Analysis.

一、緒論

現今許多網路論壇、串流影音平台等熱門的網路服務，都設置有留言區或是收藏功能等的反饋機制，讓這些網路平台中累積了多樣且豐富的資料，使得網路輿情分析 (Opinion Analysis) 成為了調查與理解市場的方法之一。在商業市場上，透過輿情分析能夠消費者對於企業的服務、產品等的正負面評價，同時也能發掘消費者所在意的優缺點，讓產業能藉由這些資訊進而調整行銷策略、改進產品生產方向與服務品質。

情感分析是輿情分析的重要一環。傳統的情感分析是針對文章(document-level)或整句話(sentence-level) 進行情感分類，當文本中包含多個討論主題時，如此所得的情感結果不見得是我們所關注角色的情感，而是對其他意見目標的情感結論。為了更精準地評估個別意見目標的情感，近年的情緒分析研究即有意見目標情感分析 (Aspect-level sentiment analysis, ABSA)，對於給定的句子不只要預測句中目標 (Target) 的情感類別，還須預測該句屬於何種面向 (Aspect)。舉例而言，對於同一家旅館或餐廳的評論中，使用者可以針對其設備、服務、清潔、地點、菜色等不同層面提出個人觀點，以及對於不同的面向或目標給予情緒上的評論。

本研究的主要目標是針對社群網路上與歌手相關的評論文章進行情感分析，不論評論內容對於評論者而言情緒為何，評論內容對於主要被評論歌手的情感評價才是我們關注的重點。因此，我們以歌手名作為評論目標 (Opinion Target)，透過 ABSA 中的子任務——意見目標的情緒分類 (Aspect-level Sentiment Classification)，運用此方法來更準確的判斷文本中的評論目標 (Opinion Target) 的情感類別，並可做為歌曲點播率及市場反應度預測，進而做為歌手行銷決策的考量。

本研究在模型的設計上，運用透過 2018 年 Google AI 團隊所提出的預訓練 BERT

中文模型[1]來進行參數微調，在 BERT 的使用上與以往的任務不同，本研究不直接採用 BERT 的分類向量 CLS 直接進行情感分類，而是將 BERT 作為來對輸入的語句進行字詞向量的轉換，並在 BERT 模型之上，運用 Huan 等人[2]的意見目標情感分類方法，以及 Parikh 等人[3]的自然語言推論 (Natural Language Inference) 的方法，進一步探討句子與目標之間的相關性，並對意見目標進行情感分類。

本篇論文貢獻主要包含兩個方向：(1)應用意見目標情感分析於社群網路評論，預測主要評論對象之情感分類結果，提供廠商掌握市場動向。(2)運用 BERT 作為嵌入層，並在 BERT 後運用多種不同的架構模型，來解決意見目標情感分類，提供更多樣化的 BERT 使用方法。

二、 相關研究

意見目標情感分析(Asspect Based Sentiment Analysis, ABSA)屬於情感分析中的一項任務，意見目標情感分析可以看成多項子任務問題，意見目標類別分類(Asspect Category Extraction)、意見目標擷取(Asspect Target Extraction)與意見目標情感分類(Asspect-level Sentiment Classification)等，而 Schouten 等人[4]與 Zhou 等人[5]也針對上述的 ABSA 相關知識與任務進行統整與介紹。先前的研究大都是單獨解決部分的子問題，但近年來有不少研究將多項子問題結合起來，即先做目標擷取再做目標的情感分類，為了讓模型可以借鏡個別標記所帶來的資訊，透過 multi-task 架構同時處理目標擷取與情感分類的聯合訓練(Joint Learning)方法變成近年來主要的研究趨勢[6][7]。

本研究專注於意見目標的情感分類(Asspect-level Sentiment Classification, ASC)。以往的情感分類(Sentiment Classification)問題主要是判斷整體文章(Document-level)或整句話(Sentence-level)的情感，而 ASC 的任務中則需要考慮到文本中對於目標(Target)的情感類別。在研究上，目標可以是指一個實體(Entity)或實體的某個層面(Aspect)，為了簡單起見，這裡通稱目標 (Target)；在 ASC 的任務中需要給定一個文本與出現在文本中的一個目標，用來判斷句子對目標的情感類別。

過去在 ASC 任務中，已有不少深度學習方法的應用，如：卷積神經網(CNN)與門控機制(Gating Mechanism)[8]、注意力機制 (Attention mechanism)[2]、Transformation Networks [9]等方法。這些方法都是使用 Word Embedding 的方式來做為輸入句子與目標的詞向量轉換，近年來 BERT 也逐漸被運用於 ABSA 的任務。例如 Sun 等人[10]將句子

與目標以 Sentence Pair 的形式作為 BERT 的輸入，透過製造輔助句子(Auxiliary Sentence)，來重新構建輸入的 Sentence Pair 的內容，並使用 BERT 的用於分類任務的向量 CLS 進行分類，獲得比原本 BERT 更好的效能。Xu 等人[11]將 BERT 應用於機器閱讀理解和 ABSA 任務中，同樣是透過目標與句子所組成的 Sentence pair 做為輸入，並透過作者提出的 Post-training 方法，運用與任務和數據領域相關的大型語料庫，來增強預訓練好的 BERT，最後使用 Post-training 後的 BERT 分類任務向量 CLS 進行目標的情感類別預測，並在實驗結果上獲得比原本 BERT 更加好的效能。

上述使用 BERT 的方法主要是以 Sentence Pair 的形式作為模型的輸入，並用額外數據資料、改善模型的輸入，來改進模型的預測效果，顯示 BERT 分類任務的 CLS 向量難以超越之處。本研究則希望能超越 BERT 分類任務的 CLS 向量的效能，應用 BERT 作為字詞向量的轉換，對於輸入句子與目標所組成的 Sentence pair 的輸出結果，搭配不同的 ASC 相關的模型，進一步探討句子與目標之間的相關性。

三、 BERT 模型嵌入層方法

本研究使用過大量中文資料預訓練過的 BERT 模型，參數設定 transformer blocks 層數有 12 層、嵌入大小為 768、transformer blocks 的 Self-Attention Head 數量為 12，以及使用參數微調 (Fine-Tuning) 的方法，來使用預訓練好的 BERT 模型將字詞轉換成嵌入向量，在中文模型的嵌入單位是以「字」為單位。並 BERT 之上運用 Huang 等人[2]所提出的 Attention-over-Attention (AOA)的注意力模型與 Parikh 等人[3]在自然語言推論任務提出的 Decomposable Attention 方法（簡稱 DA）的注意力模型，以及比較 AOA 與 DA 模型架構，所提出結合兩者架構的 Decomposable- Attention-over-Attention (DAOA)模型。如圖 1 所示，在模型的設計上總共分成 4 層：

1. 輸入層 (Input Layer) :

以句子(Sentence)與句子中的意見目標(Opinion Target)做為輸入，並將兩者進行以字元(character-based)進行分割，表示符號為分別為 $S = [s_1, \dots, s_n]$, $T = [t_1, \dots, t_m]$ ，代表句子中有 n 個字元，而目標則 m 字元，且目標詞是句子中的某

一小段連續字元所組成。並將句子與目標處理成 BERT 的 Sentence Pair 輸入格式，符號表示為 $BERT_{input} = [[CLS], s_1, \dots, s_n, [SEP], t_1, \dots, t_m, [SEP]]$ ，輸出的結果為形成一個長度為 $n+m+3$ 的序列。

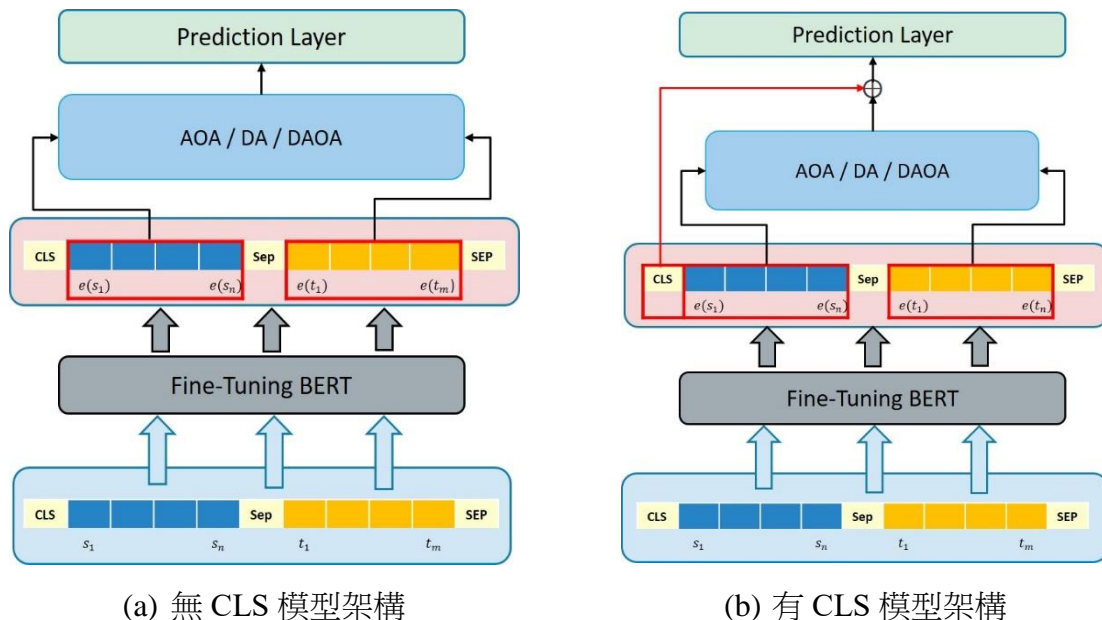


圖 1、BERT 模型架構

2. BERT 層(BERT Layer):

本研究使用預訓練好的 BERT 模型將 $BERT_{input}$ 轉換成嵌入向量，而嵌入大小為 768。而在 BERT 模型的輸出有分為兩種：pooler output 和 sequence output。前者的輸出為整個 sentence pair 序列的 representation 的向量 $[CLS] \in R^{768}$ ，是專門用於分類任務上的特殊向量；後者則是輸出句中每個字元經過 BERT 計算後的都會獲得一個對應的向量。

本研究使用 sequence output，輸出結果為 $BERT_{output} \in R^{(n+m+3)*768}$ ，並從矩陣中取出句子與目標經由 BERT 轉換後的字向量。取出結果表示 $\bar{S} = [\bar{s}_1, \dots, \bar{s}_n] \in R^{n*768}$ ， $\bar{T} = [\bar{t}_1, \dots, \bar{t}_{i+m-1}] \in R^{m*768}$ 。

3. 特徵擷取層:

根據 BERT 層的輸出做為輸入。運用 AOA、DA 以及比較 AOA 與 DA 模型架構，並結合兩者架構提出的 DAOA 注意力模型(詳細模型如 3.1 節所述)，來計算句

子(Sentence)與句子中的意見目標(Opinion Target)的嵌入向量之間的關係。

4. 預測層(Prediction Layer) :

預測層我們設計了兩種方法，第一種是如圖 1 (a) 的架構，運用特徵擷取層的輸出向量 $M_{output} \in R^{d_M}$ ，透過前饋式神經網路分類預測， d_M 表示輸出向量的維度。第二種則是如圖 1 (b)的架構，將 BERT 的 Pooler Output 的分類任務的向量 CLS，符號表示為 $BERT_{CLS} \in R^{768}$ ，與特徵擷取層的輸出結果結合 (公式 1)，再透過前饋式神經網路進行分類預測。

$$C = [BERT_{CLS}, M_{output}] \quad \text{公式 1}$$

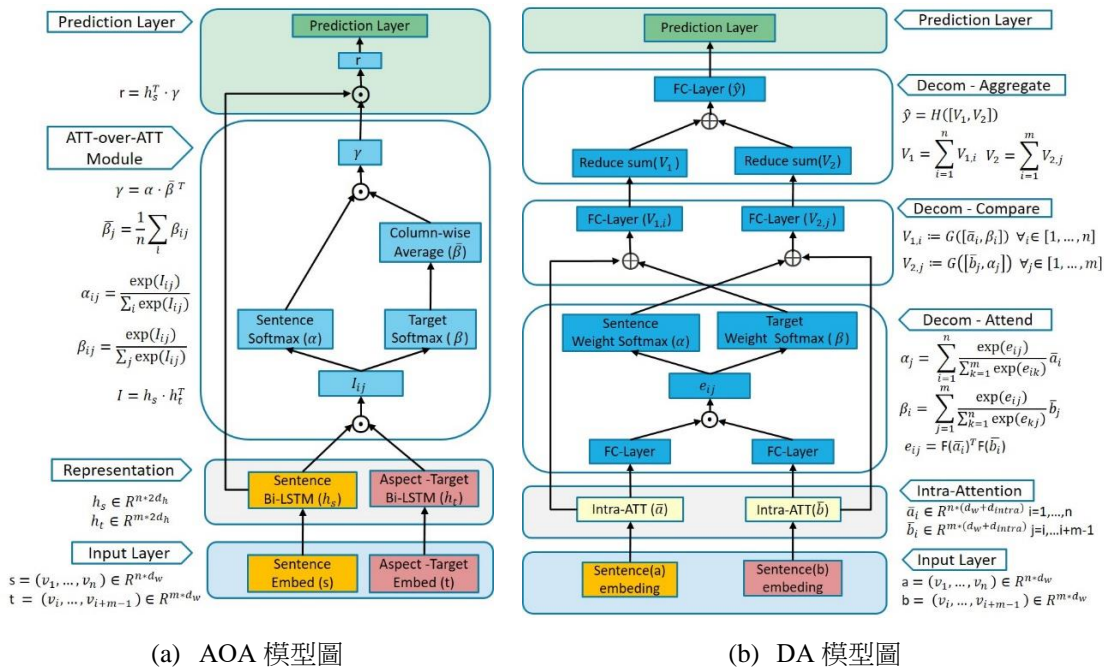


圖 2、AOA 與 DA 模型比較圖

3.1 DA 與 AOA 架構比較

再介紹 DAOA 的模型架構前，我們畫出 AOA 與 DA 兩篇論文模型從輸入到輸出的實作圖(圖 2)，並詳細比較兩者的注意力方法。根據圖 2 (a)的 AOA 的注意力計算部分(ATT-over-ATT)是一種非對稱的注意力模型方法，透過計算兩句話之間的注意力權重，並將權重加權於其中一句話。根據圖 2 (b)，DA 的注意力計算分成三個部分

Attend、Compare、Aggregate，且是屬於對稱式的計算，透過前饋神經網路 F 計算兩句話之間的注意力權重，運用 columns-wise 與 row-wise 的 Softmax 對兩句原始句子進行加權與正規化，並在將加權後的句子，與另一方的原始句子進行訊息的交互，運用前饋神經網路 G 進行比較，來探討兩句話詞與詞的相關性，最後在將比較結果透過前饋神經網路 H 進行整合。在注意力計算部分 DA 較 AOA 來的更複雜。

3.2 DAOA 架構

本研究的注意力層機制的設計想法上，運用根據圖 2 (b)中 DA 對於句子與意見目標句子之間的注意力權重計算方法(Decom-Attend)與交互比較注意力權重的方法(Decom- Compare)，然而 DA 最後的計算 Decom- Aggregate 只是單純的將比較結果進行總和。因此，我們想透過 AOA 對於資訊彙整的方式來將比較結果進行整合，來加強其中一方的權重，並進行情感預測，其架構如圖 3 所示。

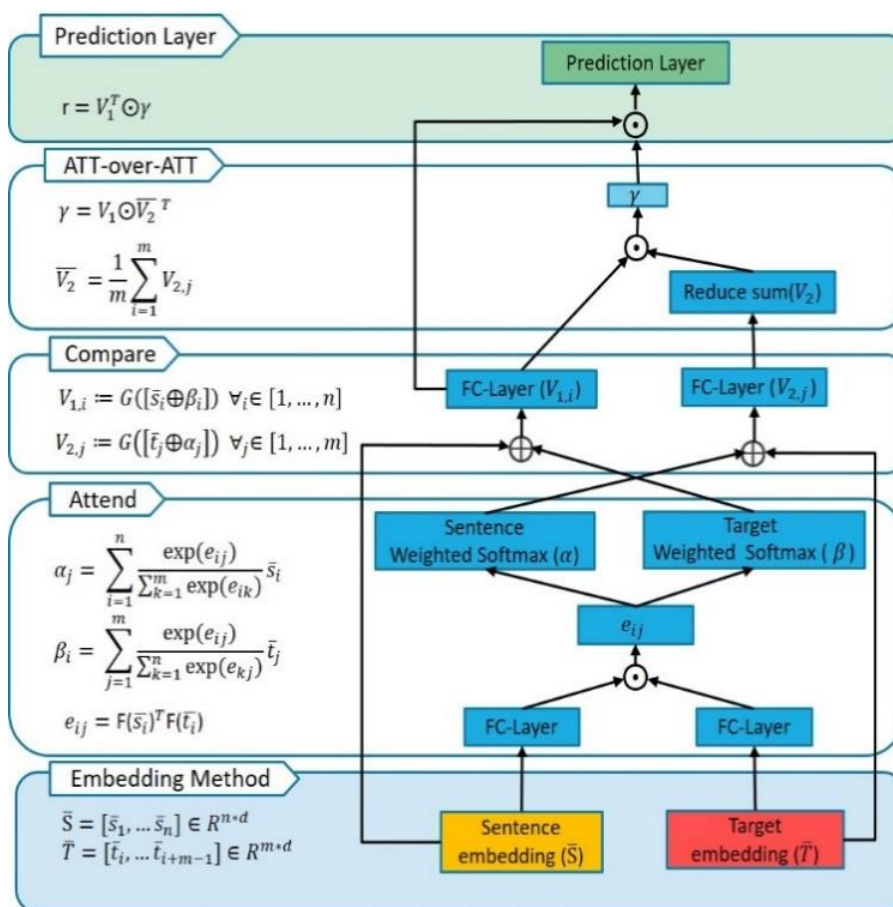


圖 3、Decom-AOA 架構

A. 注意力層(Attend Layer) :

我們首先透過激勵函數為 ReLU 的前饋式神經網路 F 來，來計算兩句話中每個詞的注意力權重 e_{ij} (公式 2)， β_i 表示 b 句中每個詞對 a 句中第 i 個詞的注意力權重經過加權與正規(normalization) 的總和結果(公式 3)， α_j 可用相同方式獲得(公式 4)。

$$e_{ij} = F(\bar{s}_i)^T F(\bar{t}_j) \quad \text{公式 2}$$

$$\beta_i = \sum_{j=1}^m \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})} \bar{t}_j \quad \text{公式 3}$$

$$\alpha_j = \sum_{i=1}^n \frac{\exp(e_{ij})}{\sum_{k=1}^m \exp(e_{ik})} \bar{s}_i \quad \text{公式 4}$$

B. 比較層 (Compare Layer) :

經過 Attend 加權後的兩個句子則分別其原始句子透過前饋神經網路 G 來進行特徵擷取(公式 5)，這部分的用意是希望能夠透過此種方式交換相互加權後的訊息，進一步的比較兩句中詞與詞的相關性。

$$\begin{aligned} V_{1,i} &:= G([\bar{s}_i \oplus \beta_i]) \quad \forall_i \in [1, \dots, n] \\ V_{2,j} &:= G([\bar{t}_j \oplus \alpha_j]) \quad \forall_j \in [1, \dots, m] \end{aligned} \quad \text{公式 5}$$

C. Att-over-ATT 層 (Att-over-ATT Layer) :

接著透過 Attention-over-Attention (AOA)的方法，將前一層所得的目標權重平均(公式 6)，做為新的權重，並與前一層句子進行比較，透過內積加強句子與目標每個詞之間的相關性(公式 7)。

$$\bar{V}_2 = \frac{1}{m} \sum_{i=1}^m V_{2,i} \quad \text{公式 6}$$

$$\gamma = V_1 \odot \bar{V}_2^T \quad \text{公式 7}$$

D. 輸出層(Output Layer) :

最後 DAOA 輸出層的部份，則是經由 Att-over-ATT 所獲得的比較權重，與比較層所獲得的 V_1 ，進行矩陣相乘 (公式 8)。其用意是希望，原始的句子已經與目標的注意力權重進行比較，並在運用 Att-over-ATT 所獲得的比較權重，進一步的加強句子與目標每個詞之間的相關性。

$$r = V_1^T \odot \gamma \quad \text{公式 8}$$

四、 資料集準備與分析

本研究所使用之資料集，是來自台灣的「批踢踢 PTT 實業坊」，由於我們關注的主題是歌手的網路聲量，因此我們僅抓取與音樂相關討論版的文章。資料準備方式如下，首先對 PTT 的貼文進行預處理，解決 PTT 貼文特定格式、特殊符號、URL 等，排除會影響文本分析的訊息，並使用歌手字典做為種子，搜尋含有歌手的段落，並透過人工的方式判斷該段落主要被討論的實體，以及該段落所描述的內容對於被標記實體的情感。

為了確保測試資料的標記品質，測試資料經由三組人員以進行標記，最後採用多數決的方式來決定情感類別。如圖 4 範例所示，紅色代表該段落主要被討論的對象，而標記者會根據整段文字的敘述，來判斷該段落對於討論對象情感類別，而情感類別包含「正向、負向、中立」三類的情感。

Target	Target Sentiment	Corpus
張藝興	中立	張藝興回母校設立獎學金未來10年贊助100萬人民幣今日(5月13日)，有網友透過微博稱男團EXO成員張藝興(LAY)向母校湖南師大附中捐款100萬元人民幣和一架鋼琴。據悉，張藝興為湖南師大附中設立藝術獎學金，每年捐款10萬元人民幣幫助有夢想的學弟...
五月天	正評	謝謝你、你、你、你、你。謝謝你們願意成為五月天，謝謝你們願意當五月天。謝謝你們讓我當五月天的家人。想和你們說的感謝不知道還能如何表達。我在五月天音樂中療癒，傷口變成勇氣，把眼淚擦乾，又能勇敢邁進...
徐懷鈺	負評	吳宗憲安慰徐懷鈺訊息曝光！要她問自己錄影時快不快樂2016年04月10日1750記者黃子瑋/台北報導「國民天后」徐懷鈺日前上「綜藝玩很大」被批評為求勝負使出奧步，耍賴硬盧，試圖讓吳宗憲心軟...

圖 4、標記資料範例

表 1、Aspect-Based 資料集統計

Dataset	# of Sent.	Class			# Distinct Targets
		Negative	Positive	Neutral	
Training	1,238	30	426	782	247
Testing	705	17	313	375	189

我們透過 Kappa 值來檢視 3 組人員的情感標記與透過多數決所得情感類別的一致性。各組 Kappa 值分別為 0.66、0.66、0.70，都達到 0.6 以上，根據 Kappa 值的標準，測試資料的標記一致性尚可接受。

五、 實驗與模型效能評估

(一) Bert 模型效能

本階段實驗採用 BERT 的 pooler output 所輸出的分類任務的向量 CLS 之預測結果做為 Baseline，在此簡稱 BERT CLS。首先比較運用 BERT 嵌入層方法，在其架構上搭配不同的注意力模型 AOA、DA 以及 DAOA，來計算轉換後句子與目標之間相關性模型的效能。第二部份，則是在注意力模組輸出後，加入 CLS 向量來探討是否能夠增加效能。各項模型如表 2 所示，表中的每個模型的效能，都是模型執行 10 次後平均的結果。

表 2、BERT 模型之效能

Model	F1-Score				Accuracy (%)
	Negative	Positive	Neutral	W. Avg	
BERT CLS	0.428	0.705	0.758	0.726	72.81
BERT-AOA	0.334	0.715	0.759	0.728	73.16
BERT-DA	0.289	0.689	0.746	0.710	71.49
BERT-DAOA	0.107	0.703	0.758	0.719	72.40
BERT-CLS-AOA	0.321	0.711	0.756	0.726	72.65
BERT-CLS-DA	0.250	0.698	0.753	0.718	72.10
BERT-CLS-DAOA	0.208	0.700	0.760	0.721	72.70

首先，我們先比較表 2，前四項的效能，在模型整體效能的表現上 BERT-AOA 的模型表現最好(Weighted avg. F1 = 72.8)，高於 BaseLine 效能(72.8 v.s. 72.6) 以及其餘的模型效能，且 BERT-AOA 在正向與中立類別的效能分數也是所有模型中表現出色的。不過在負向情感類別的效能分數上 BERT 展現最好的效能 (Negative F1 = 0.428)。

接著，我們加入了 CLS 的向量，來觀察是否能夠改善效能，加入 CLS 向量後對於 BERT-DA (71.0 -> 71.8) 與 DAOA (71.9 -> 72.1) 都可獲得了效能上的改善，但對於的 BERT-AOA 模型在效能的表現上是下降的。

透過實驗，我們發現在 BERT 的 sequence output 上運用太複雜的注意力模型架構可能會導致效能的下降。舉例而言，AOA 的運算比 DA 來的相對簡單許多，而本研究提出的架構 DAOA，雖然在效能上比 DA 來得好(71.9 v.s. 71.0)，但在計算的複雜性上，還是比 AOA 來的更加複雜，因此在 BERT 上應用較為簡單的運算架構，會比複雜的運算架構效能表現上來得更好。而加入 CLS 向量，從實驗結果顯示是有幫助於改善原先效能表現較差的模型。

由於表 2 表現較好的前三個模型在整體效能差異較小，本研究透過盒鬚圖 (box plot) 進一步分析模型執行 10 次的過程中的模型效能差異。圖 5(a)是三個情感類別的 F1-Score 效能，在中立與正向的部分三個模型的最大值以及效能分數的浮動差異不大，但在負向類別，能看出 BERT 的效能明顯高於另外兩者，且效能分數的浮動較其餘模型穩定。圖 5(b)為整體模型的 Weighted avg F1 與 Accuracy，雖然三個模型的 Weighted avg F1 最大值近乎相同，但 BERT-AOA 是三者模型中平均 Weighted avg F1 分數最高且分數浮動最穩定的模型。而在 Accuracy 的部份 BERT-AOA 在模型不管在最大值還是平均分數，都明顯好於其它模型。根據盒鬚圖的分析，得知 BERT-AOA 模型是所有模型中效能最好且最為穩定的模型。

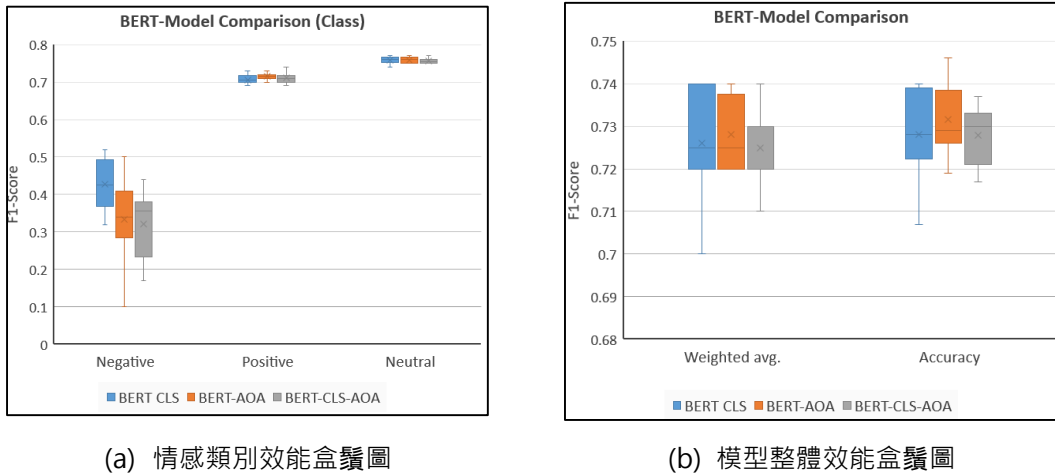


圖 5、BERT 模型盒鬚圖

(二) 預測錯誤分析

本階段根據 Weighted avg F1 前三名的模型 BERT-AOA, BERT CLS 以及 BERT-CLS-AOA 的共同預測錯誤的結果進行錯誤分析，並列出模型在預測上常見的錯誤範例。在第一、二個範例中，"疑似劈腿"、"孝順"並不是情感詞彙，但根據人的認知判斷，這些詞彙對於目標來說是帶有正負評價的詞彙。第三個範例中，雖然有明顯的情感詞彙，但這些詞彙並不是指向於目標的詞彙，而是描述留言者本人。第四個範例中，出現了強烈的負面詞彙，但是這些詞彙並不指向於目標，而是一種情境上的敘述方式。

Target	Target Sentiment	Corpus	BERT	BERT-AOA	BERT-CLS-AOA
唐從聖	負評	從哥大方認了一認了未婚生女！唐從聖PO全家福照「今天自由時報即時新聞」藝人唐從聖先是被爆未婚生女，又被媒體揭露正與一名護士交往，疑似劈腿，但唐從聖昨晚在臉書上貼出一家四口的照片，等於認了未婚生女	中立	中立	中立
羅志祥	正評	羅志祥講小時候的故事，感覺很辛苦他會這麼孝順是知道要回頭！媽媽突然出現，小豬沒猜到很驚訝！...	中立	中立	中立
蘇打綠	中立	在這誠心誠意地徵求蘇打綠謝天謝地演唱會門票2張非常喜歡春日光專輯，最喜歡日光，交響夢，早點回家三首歌，如果能夠如願前往聆聽，一定會遵守蘇打綠演唱會的所有規定，	正評	正評	正評
吳宗憲	中立	媒體來源蘋果吳宗憲為節目哭窮被酸放屁諷馮光遠值8千元再談金鐘50唇槍舌箭「蔡維歡、顏馨宜／台北報導」吳宗憲（憲哥）在金鐘獎典禮上怒嗆評審，引起網友共鳴，並創下當天典禮最高收視417，吸92萬人收看。但他引言...	負評	負評	負評

圖 6、錯誤分析範例

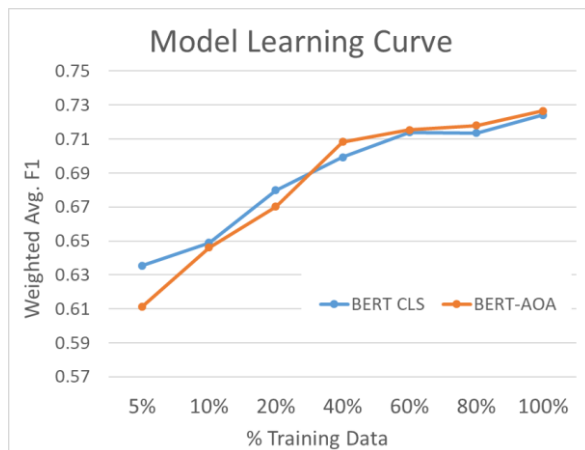


圖 7、模型學習曲線

從上述錯誤分析來看，現有模型對於具有情感涵義的詞彙，以及對於情感詞彙與意見目標的關係，並未能完全理解，難以判斷。

(三) 模型學習曲線

根據 BERT 模型中 Weighted Avg F1 表現最好的 BERT-AOA 模型與 Baseline 模型 BERT CLS，來畫出模型學習曲線(Learning curve)，並探討模型與資料之間的關係。根據圖 7 中可以觀察到，兩個模型在訓練資料量只有 5%的情況下模型的 F1-score 就能夠到達 60 以上，而 BERT CLS 在數據量少(< 40%)的時候效能高於 BERT-AOA 模型，但在模型隨著資料量的增加 BERT-AOA 效能比 BERT CLS 來的更好，且隨著資料量的增加，BERT-AOA 模型的 F1-score 都可以隨之提升，若是增加更多訓練資料，也許能獲得更好的效能。

(四) 新增資料效能影響

由於本研究的負向情感訓練資料偏少，使模型在負向類別學習效能表現上不佳，因此我們希望透過增加負向標記訓練資料，是否能否提升模型對於負向類別的效能，以及資料量的增加數是否會再次提升效能。

我們以模型中表現最好的 BERT-AOA 來進行這項實驗。如圖 8 所示，BERT-AOA 隨著負向資料的增加，負向的 F1-score 獲得了提升。不過在負向 F1-score 提

升過程中，並沒有改善其餘類別的 F1-score 與整體模型的 Accuracy。當負向訓練資料增加超過 100 筆時，整體效能略為下滑，顯示負向資料僅有助於改善模型對於負向類別的預測效能。

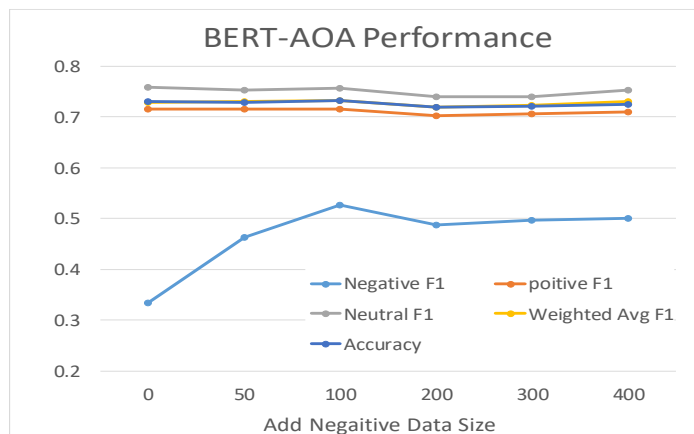


圖 8、BERT-AOA 效能曲線圖

六、 結論

本研究中透過 BERT 模型作為嵌入層來對句子與目標的字進行轉換，並搭配不同 NLP 任務的注意力模型來找出句子與目標之間的相關性。實驗結果顯示，BERT 模型之上搭配注意力模型以及加入 BERT 的 CLS 分類向量是有助於 BERT 提升模型的效能，但在 BERT 輸出後搭配過於複雜的注意力模型可能導致模型效能的下降。在本研究的模型中由於訓練資料負向資料過少，本研究嘗增加負向的訓練資料，根據實驗結果，增加負向資料有助於改善模型負向資料的效能，且對於其餘類別以及整體效能的提升有限。

參考文獻

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [2] B. Huang, Y. Ou, and K. M. Carley, “Aspect level sentiment classification with attention-over-attention neural networks”, vol. abs/1804.06536, 2018.
- [3] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, “A decomposable attention model

- for natural language inference”, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Nov. 2016, pp. 2249–2255.
- [4] K. Schouten and F. Frasincar, “Survey on aspect-level sentiment analysis”, IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 813–830, 2016.
- [5] J. Zhou, J. X. Huang, Q. Chen, Q. V. Hu, T. Wang, and L. He, “Deep learning for aspect-level sentiment classification: Survey, vision, and challenges”, IEEE Access, vol. 7, pp. 78 454–78 483, 2019.
- [6] D. Yin, X. Liu, and X. Wan, “Interactive multi-grained joint model for targeted sentiment analysis”, in CIKM ’19, 2019.
- [7] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, “An interactive multi-task learning network for end-to-end aspect-based sentiment analysis”, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 504–515.
- [8] W. Xue and T. Li, “Aspect based sentiment analysis with gated convolutional networks”, in Proceedings of the 56th Annual Meeting of the association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, Jul. 2018, pp. 2514–2523.
- [9] X. Li, L. Bing, W. Lam, and B. Shi, “Transformation networks for target-oriented sentiment classification”, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 946–956.
- [10] C. Sun, L. Huang, and X. Qiu, “Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence”, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 380–385.
- [11] H. Xu, B. Liu, L. Shu, and P. Yu, “BERT post-training for review reading comprehension and aspect-based sentiment analysis”, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2324–2335.

自然語言處理與數位人文

Natural Language Processing for Digital Humanities

劉昭麟 Chao-Lin Liu

國立政治大學資訊科學系

Department of Computer Science

National Chengchi University, Taiwan

chaolin@g.nccu.edu.tw

洪振洲 Jen-Joe Hung

法鼓文理學院佛教學系

Department of Buddhist Studies

Dharma Drum Institute of Liberal Arts, Taiwan

jenjou.hung@dila.edu.tw

張素玢 Su-bing Chang

國立臺灣師範大學臺灣史研究所

Graduate Institute of Taiwan History

National Taiwan Normal University, Taiwan

109682@ntnu.edu.tw

吳宛怡 Wu, wan-yi

香港理工大學中國文化學系

Department of Chinese Culture

The Hong Kong Polytechnic University, Hong Kong

wan.yi.wu@polyu.edu.hk

摘要

數位人文近十幾年以來在國際學術界已然蓬勃發展。相對之下，我國計算語言學界參與數位人文研究的程度並不如一般所預期。這一個座談會藉著簡短介紹四個具備數位人文屬性的研究工作，希望能夠吸引更多計算語言學界的學者參與數位人文的研究。中華電子佛典學會的大正藏網路資料庫 (CBETA) 的內容與附屬的服務都是免費的。洪振洲將分享建構這一個資料庫的過程中，語文分析科技包含人工智慧與自然語言處理技術等，的相關應用。歷史人物的傳記資料是歷史研究的重要基石，張素玢會介紹「臺灣傳記人物資料庫」 (TBDB) 的研究團隊從多個臺灣的地方志中抽取個人傳記資料來建構 TBDB 的經驗和展望。戲劇是中國傳統藝術的重要部分，相關資料以許多不同形式留存至今。吳宛怡將分享關於中國戲劇的研究經驗，並且討論語文分析技術對於各種戲劇研究的潛在貢獻機會。如果時間允許，劉昭麟將介紹漢文古文書數位化的文字辨識工作，

文言歷史文本的分句工作，諸如《全唐詩》、《全宋詩》和《全臺詩》等格律詩的斷詞工作，還有從文言史學資料中抽取有用資訊，例如個人傳記資料，的相關經驗。

Abstract

The research of digital humanities has flourished in the past decade internationally. In contrast, the participation of researchers of computational linguistics in domestic research projects remains less common than one may have anticipated. The goal of this panel is to introduce sample research projects of digital humanities to the community of computational linguistics, hoping to promote further cooperation between the two communities. The panel consists of four parts. Hung introduces the online repository of the Taishō Tripiṭaka that the Chinese Buddhist Electronic Text Association (often called CBETA) offers. Applications of language technology, including artificial intelligence and natural language processing, for the construction of CBETA will be discussed. Chang and her colleagues aim to build the Taiwan Biographical Database (TBDB), which, in the long term, will serve as part of the bedrock for historical studies about Taiwan. Experience about how the research team extracted and integrated the information from some collections of local gazetteers to build the TBDB will be discussed. Dramas are important part of Chinese arts. Relevant materials about dramas are available in some different databases and in different forms. Wu will share with us her study on Chinese dramas, and elaborates on the potential contributions of language technology to the studies of Chinese dramas. If time allows, Liu plans to outline his work on optical character recognition (OCR) for ancient Chinese documents, sentence segmentation for classical Chinese, word segmentation for classical Chinese poems, including the Tang and Song poems, and information extraction from historical documents in classical Chinese.

自然語言處理技術於數位人文領域的機會與挑戰 --以佛教經典研究為例

The Opportunities and Challenges of Natural Language Processing Technology in the Field of Digital Humanities – Taking the Study of Buddhist Scriptures as an Example

洪振洲 Jen-Jou Hung
法鼓文理學院佛教學系
Department of Buddhist Studies
Dharma Drum Institute of Liberal Arts, Taiwan
jenjou.hung@dila.edu.tw

摘要

佛教自東漢傳入中國以來，歷經長時間的發展，成為人民主要的信仰之一。佛教在中國發展的千年時間中，不僅發展出獨特的漢傳佛教風格，同時間，佛教也融合成為了中華文化的一部分。而佛教傳入中國後，另一個重要的產出，便是引發了歷經千年的佛經翻譯活動，因此產生了大量的漢文佛典文獻。這些佛教典籍，經由歷代僧人精心創作、發展、甄別、校對、編排與整理，形成我們今日所見的漢文大藏經。漢文大藏經中內收錄的大量佛典文獻衍生出多元並存的部類與思想體系，表現於外在書目有著一定的規範結構組織，其編輯的內在理路，包括編藏目的、典籍選編入藏的標準、編纂藏經方法，通常也反映著一個時代的知識構成與跨時代的知識演變。

隨著國家推動數位典藏發展，漢文大藏經及相關之大量佛教典籍已陸續數位化，並製造出多項佛學資料庫。此項成果，讓佛學、人文學科以及各領域的研究者可以前所未見的便利方式，取得佛經內容與相關參考資料。然而，人文學者過去以傳統方式運用網路環境蒐集資料、進行研究工作，如果研究主題涉及寬廣的背景知識，以往只能以關鍵詞搜尋到單一而零碎的訊息，需要將許多碎片化的資料來源，逐項解讀、判斷與比對，才能整合組成具有關連意義的背景資料。隨著大量資料在網際網路湧現，以關鍵詞搜尋資料的傳統方式，不僅耗費大量人工時間，也不免出現毫無意義或重覆訊息，乃至天壤地別的錯誤結果。

人工智慧領域中之自然語言處理之研究不斷推進，對於文字內容裡的重要資料擷取、語法及語意分析、文本生成與自動問答等技術持續發展，並且隨著深度學習方法的導入，在許多應用上得到令人欣喜的突破。然此類自然語言處理之技術發展，由於語料來源與商業利益之考量，多半仍集中於現代語言之處理，對於古典語言文獻處理上仍未得到廣泛的關注，也缺乏合適之自然語言處理工具與訓練語料集。如能將現前已深度發展之自然語言處理技術，帶入漢文佛典文獻之處理，前述之人文學者進行研究時所遭遇之問題將可望得到有效之解決。目前應用人工智慧技術於佛典數位資料研究，雖有獲取少數成功案例，但仍屬萌芽階段。本演講將描述佛教研究領域的數位資料建置現況、簡單說明

傳統佛教研究學者關心之研究議題，以及自然語言處理技術應用於此研究領域的發展與成果。藉此希望引發更多共鳴，以吸引更多計算語言學界的研究者參與數位人文的研究。

Abstract

Since Buddhism was introduced into China in the Eastern Han Dynasty, it has undergone long-term development and has become one of the main beliefs of the people. During the millennia of development of Buddhism in China, not only did it develop a unique style of Chinese Buddhism, but at the same time, Buddhism also became part of Chinese culture. After Buddhism was introduced to China, another important effect was to trigger a millennium-long process of Buddhist scripture translation, which resulted in a large number of Chinese Buddhist texts. These Buddhist scriptures, through the careful creation, development, screening, proofreading, arrangement and collation by monks of the past dynasties, form the Chinese canon that we see today. The large number of Buddhist texts included in the Chinese canon derives from the coexistence of diverse divisions and ideological systems, which is reflected in the external bibliography having a certain standardized structure and organization, and the internal rationale of its editing, including the purpose of the collection, the standards of the selection and inclusion of scriptures, and the method of compiling canonical scriptures usually reflects the knowledge structure of one era and the evolution of knowledge across the ages.

As the country promotes the development of digital collections, the Chinese canon and a large number of related Buddhist scriptures have been digitized one after another, and a number of Buddhist studies databases have been created. This achievement allows researchers in Buddhism, the humanities, and various other fields to access the content of Buddhist scriptures and related reference materials in a convenient way that has never been seen before. However, humanities scholars use traditional methods to collect data and conduct research in the Internet environment. If the research topic involves broad background knowledge, in the past, only single and fragmented pieces of information could be searched for by using keywords, and many fragmented pieces of data were needed. Sources, item-by-item interpretation, assessment, and comparison can be integrated to form relevant background information. With the emergence of a large amount of information on the Internet, the traditional way of searching for information with keywords not only consumes a lot of time, but also inevitably produces meaningless or repeated messages, and even widely erroneous results.

Natural language processing research in the field of artificial intelligence is constantly advancing. The technology of capturing important data in text content, grammar and semantic analysis, text generation, and automatic question answering continues to develop, and with the introduction of deep learning methods, it has been applied in many ways to provide gratifying breakthroughs. However, the technological development of this type of natural language processing, due to the source of the corpus and the consideration of commercial interests, is mostly still focused on the processing of modern languages. The processing of classical language documents has not yet received widespread attention, and there is a lack of suitable natural language processing tools and training corpus. If the previously developed natural language processing technology can be brought into the processing of Chinese Buddhist texts,

the foregoing problems encountered by humanities scholars in their research will be expected to be effectively solved. At present, the application of artificial intelligence technology in the research of digital data of Buddhist scriptures has obtained a few successful cases, but it is still in its infancy. This lecture will describe the current situation of digital data construction in the field of Buddhist research, and briefly explain the research topics that traditional Buddhist scholars are concerned with, as well as the development and results of natural language processing technology in this research field. It hopes to arouse greater interest and attract more researchers in computational linguistics to participate in the research of digital humanities.

致謝 (Acknowledgments)

This research was supported in part by the contracts MOST-106-2420-H-655-001-MY3 and MOST-106-2420-H-655 -002 -MY3 of the Ministry of Science and Technology of Taiwan.

人物資料庫的建置：以「臺灣歷史人物傳記資料庫」為例

The Taiwan Biographical Database (TBDB): An Introduction

張素玢 Su-bing Chang
國立臺灣師範大學臺灣史研究所
Graduate Institute of Taiwan History
National Taiwan Normal University, Taiwan
109682@ntnu.edu.tw

摘要

人物是歷史學研究的重要基礎，舉凡人物的個性、家庭背景、經歷、社會階層，甚至於整個社會的階層流動、婚姻與政治網絡等都是歷史研究的議題。「中國歷史人物傳記資料庫 (The China Biographical Database, CBDB)」可謂目前與歷史人物研究相關最具規模的資料庫之一。目前臺灣已有的人物傳記資料多以數位文件的形式存在，而須用人力從大量人物傳記中爬梳、彙整資料。此外，儘管臺灣過去已建置眾多資料庫，也有各種人物傳和可資應用的資料文獻，卻較少進行歷史人物資料庫勘考、分析工具的開發。有鑑於此，研究者乃組成研究團隊，以《新修彰化縣志·人物志》為文本來源，發展資料庫檢索、全文檢索、文本探勘與社會網絡等分析工具，並運用前述工具協助歷史人文學者進行研究。

臺灣歷史人物傳記資料庫 (Taiwan Biographical Database, TBDB) 的建置計畫，源自中國歷代人物傳記資料庫 (China Biographical Database, CBDB) 的啟發，期望能提供臺灣歷史人物傳記資料的文本探勘、社會網絡分析工具及相關軟體服務，梳理錯綜複雜的人際網絡，開拓臺灣史的研究視野。基於建置者對資料的掌握度及著作權的取得，建置團隊首先選擇四位人文學者成員共同編纂的《新修彰化縣志·人物志》做為文本來源，其體例統一、考訂嚴謹，含括各時代、各類型人物，是 TBDB 良好的試金石。除了借鏡 CBDB，建置 TBDB 的重點在於依照臺灣的時空環境，設計符合臺灣歷史的人物屬性類別。目前初步完成檢索服務、人物地域分布及社群關係分析等功能，開放學界利用以除錯優化。目前資料內容已從彰化縣人物擴及臺北市、澎湖縣、臺中市、南投縣等地區。

本演講主旨在於描述「臺灣歷史人物資料庫」現階段所收錄之人物特性，闡述系統架構以及初步成果。此外，也說明建置「臺灣歷史人物資料庫」過程中的經驗和未來發展方向。

Abstract

Personage is an important kind of entities in the study of history. Comprehensive understanding of personage biographies is beneficial for researching into historical events. In the digital era, many personage biographies are available in digital formats; as a result, it is time-consuming and labor-intensive for researchers to explore invaluable findings from massive personage biographies. Facing this situation, researchers may be helped to utilize the information efficiently with information technologies.

The idea of developing Taiwan Biographical Database (TBDB) was inspired by China Biographical Database (CBDB) project, and aims to provide digital tools, such as text mining, social network analysis, and other related tools, for analyzing complex social networks and broaden research visions in the study of Taiwanese history. Considering the participants' knowledge of texts and the ability to acquire copyright, the team chooses the "Treatise of Historical Figures" of the New Edition of Changhua Local Gazetteer, compiled by four of the team members, as the base text. There are several advantages of this choice. First, the biographies in the Treatise are written according to the same format; second, the biographies are rigorously verified; third, they include people from all periods and all walks of life. Therefore, they should be a good starting point for the TBDB project. Besides learning from CBDB's model, the emphasis in developing TBDB is to build attributes for historical figures that are suited for Taiwan's historical contexts. Currently, we have built initial versions of basic search functions, a map which shows the geographical distribution of historical figures, and social network analysis tools. They are now open to the public in order to optimize and debug. In the future, we will continue to increase both the quality and quantity of the database and also develop new analysis tools.

This speech introduces the development of a text retrieval and mining system for Taiwanese historical people -- Taiwan Biographical Database (TBDB). It describes the characteristics of personages in TBDB, highlights the system architecture and preliminary achievement of TBDB. Finally, this talk elaborates on the lessons learned through the creation of TBDB, and the future plans.

致謝

本研究為科技部研究計畫的部分成果，計畫編號為：MOST 106-2420-H-003-010 和 MOST 105-2420-H-003 -016，上述整合型計畫的參與學者為柯浩仁、謝順宏、李宗翰、李毓嵐、李昭容、顧雅文等，共同建置了「臺灣歷史人物傳記資料庫」(TBDB)的基礎，特此致謝。

參考文獻

- 張素玢, 2018.04, 〈以《新修彰化縣志·人物志》建置臺灣歷史人物傳記資料庫(TBDB)的構想和做法〉, 《彰化文獻》22, 彰化:彰化文化局, 頁 15-31。
- 張素玢、李宗翰、李毓嵐、李昭容、顧雅文、柯皓仁、謝順宏, 2018.10, 〈從 CBDB 到 TBDB: 以《新修彰化縣志·人物志》為試金石〉, 《數位典藏與數位人文》2, 頁 91-115。
- 謝順宏、張素玢, 2018.06, 〈臺灣歷史人物文本檢索與探勘系統之建置〉, 《圖資與檔案學刊》92, 頁 67-87。
- Bol, P. K., Hsiang, J., & Fong, G. (2012). Prosopographical databases, text-mining, GIS and system interoperability for Chinese history and literature. In *Proceedings of the 2012 International Conference on Digital Humanities*.
- Digital Humanities Network (n.d.) *About*. Retrieved from <https://www.digitalhumanities.cam.ac.uk/about/about-page>.
- Fuller, M. A. (2015). *The China Biographical Database User's Guide*, Revised Version 2.0. Retrieved from https://projects.iq.harvard.edu/files/cbdb/files/cbdb_users_guide.pdf.
- Harvard University (n.d.). *Home. China Biographical Database Project (CBDB)*. Retrieved from <https://projects.iq.harvard.edu/cbdb/home>.
- Liu, C. L., Huang, C. K., Wang, H., & Bol, P. K. (2015). Toward Algorithmic Discovery of Biographical Information in Local Gazetteers of Ancient China. In *29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)*, Shanghai, China, October 30 – November 1, 2015.
- Martin, L. (2016). The university library and digital scholarship: A review of the literature. In Mackenzie, A. & Martin, L. (ed.) (2016). *Developing Digital Scholarship: Emerging Practices in Academic Libraries*. Facet Publishing.
- Sie, S. H., Ke, H. R., & Chang, S. B. (2017, November). Development of a text retrieval and mining system for Taiwanese historical people. In *Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, 2017 (pp. 56-62). IEEE.

如何從數位人文的角度分析二十世紀初期(1900-1937)傳統戲劇相關資料—試以報刊數據庫、唱片資料庫以及劇本合集為中心

How to Analyze the Related Materials of Traditional Chinese Drama in the Early 20th Century (1900-1937) from the Perspective of Digital Humanities—Focusing on Newspaper Databases, Record Databases, and Script Collections

吳宛怡 Wu, wan-yi

香港理工大學中國文化學系

Department of Chinese Culture

The Hong Kong Polytechnic University, Hong Kong

wan.yi.wu@polyu.edu.hk

摘要

二十世紀初期中國傳統戲劇是個蓬勃發展的領域，京劇，梆子劇等逐漸從成熟到鼎盛時期，新劇種更是層出不窮、百花齊放。然而，學界對於本時期的研究多偏向於京劇領域，梆子劇為首等其他地方劇種不在其視野之內，甚為遺憾。導致如此狀況的主要因為資料方面的缺乏。

近年來，《申報》、《順天時報》等近現代報刊數據庫的出現，使得研究者擁有更多的新興資料來源。多種報刊上所刊登的劇評、演員介紹、劇本及演出節目表等項目，均是理解本時期戲劇發展的重要資訊；但由於資訊過於龐大，人文學者往往僅能侷限於單一報刊或劇種，進行收集與解讀。又，二十世紀初期唱片技術普及，唱片工業日趨茁壯，唱片成為時人另外一種新興的聽劇體驗。現已有「中華老唱片數字資料庫」、「中華傳統音樂資源系列數據庫」等資料庫收錄眾多的聲音資料；藉由提取本時期的各劇種唱片目錄，演唱者，聲音資料，或能重新建構各類劇種表演流派的形成歷史。最後尚有豐富的戲劇作品合集有待勘查。民國初年出版眾多的劇本合集，例如由王大錯編纂，出版於1912-25年的《戲考》，收錄當時流行的作品約有五百多齣，裡面有詳細的演唱內容，人物行當等資訊。同樣地，也因數量龐大，無法以單一查找的方式精確判讀所有文本內容。本回座談，試想討論若能經由系統性地數據整理，或可具體化各類劇種的發展歷程，理解表演形式的改變，劇目的演進狀況，藉此重構多向度的戲劇研究之可能性。

Abstract

In the early 20th century, traditional Chinese drama was a vigorous development field, Peking Opera, Bangzi Opera and so on gradually developed from maturity to heyday, and new types of opera emerged in an endless stream. However, the academic research on this period is mostly focused on the field of Peking Opera. It is a pity that other local operas such as Bangzi Opera are not in its field of vision. The main reason for this situation is the lack of information.

In recent years, With the emergence of modern newspaper databases such as Shen Bao (申報), Shuntian Shibao (順天時報) researchers have more new sources of information. The dramatic criticisms, actor introductions, scripts, and performance schedules published in various newspapers are all important information for understanding the development of theater in this period; however, due to the huge amount of information, humanities scholars can only collect and interpret them in a single newspaper or drama. In addition, at the beginning of the 20th century, with the popularization of recording technology and the growing prosperity of the recording industry, records became another new experience of listening to drama. At present, there are a lot of sound data in databases such as "Chinese Old Records Digital Database", "Chinese Traditional Music Resource Series Database"; by extracting the album catalogs, singers, and sound data of various dramas of this period, it may be reconstructed the formation history of various performance genres. Finally, there are still plenty of drama works to be explored. In the early years of the Republic of China, a large number of play scripts were published, such as Xi Kao (戲考), published in 1912-25. It contains more than 500 popular dramas at that time, including detailed singing content, characters profession and other information. Similarly, due to the large number of texts, it is impossible to accurately interpret all the text content with a single search method. In this session, I would like to discuss the possibility of reconstructing multi-dimensional drama research by systematically sorting out data, or by specifying the development process of various types of drama, and understanding the changes in performance forms and the evolution of drama repertoires.

致謝 (Acknowledgments)

Wu is supported by General Research Fund, The Research Grants Council, Hong Kong. Project number:15610219.

漢文文言文史文本的數位化、斷詞、分句與資訊擷取

Optical Character Recognition, Word Segmentation, Sentence Segmentation, and Information Extraction for Historical and Literature Texts in Classical Chinese

劉昭麟 Chao-Lin Liu

國立政治大學資訊科學系

Department of Computer Science

National Chengchi University, Taiwan

chaolin@g.nccu.edu.tw

摘要

文言文文本是研究二十世紀之前中國歷史、社會與文學議題的主要素材。近年以來語言分析技術的進展大都集中於白話文本。在這一簡短的報告中，我們介紹一些應用自然語言處理、機器學習(包含深度學習)技術以從文言文資料源抽取有用資訊的數位人文研究議題[4]。

自動化的軟體服務，讓我們得以以較低的人力代價，從大量的歷史文獻中擷取傳記資料[8]，進而建構諸如 CBDB[2]、TBDB[1][14]、CBETA[3][6] 和 DocuSky[5] 等專業服務。例如，人物和地名是史學研究的重要根基[11][15]；在獲得大量的人物和地名等資訊之後，我們可以嘗試推測文言文的句法，進一步增進資訊擷取的能力[11]。諸如條件隨機場和深度學習等技術讓研究者比較容易發覺大量文獻中這一些命名實體的資訊；進階的關係推演技術，使得推論文本中人際關係和社會網絡的難度大幅降低[13]。

數位人文分析技術同時也讓我們可以透過分析大量的傳統漢詩來探索詩人內心的語言和文學世界[12]。從統計面來說，我們發現了歷代漢詩的用字分佈呈現了齊夫分佈。借助量化分析的取徑，我們也可以觀察詩人如何巧妙地安排和組合優美字詞，來營造讀者所感受到的美學意境。

以上提到的這一些研究，雖然都已經有不錯的進展，但是計算工具所傳回的結果，都還需要專業人員相當精度的精度來驗證和詮釋。原始資料源中的資料，不僅僅詞彙之間沒有明顯邊界，就連句子之間也沒有現代人所熟知的標點標記。要精確了解這一些連續漢字的古文並不容易。為了提高人類專家分析這一些語料的效能，為文言資料進行斷詞和分句是非常基礎的研究工作[7][9][10]。事實上，尚有許多文言語料目前只有紙本資料，還沒有完全的數位化；需要以人工繕打或者以計算方法進行文字辨識。由於文言文書印刷與書寫的方式、排版的方式有諸多變化，文本數位化的自動化仍然有許多工作仍待解決。

Abstract

Texts written in classical Chinese are the major sources for studying the Chinese history, society, and literature, particularly for the periods before the twentieth century. Recent advances in language technology focus mostly on the analysis of modern mandarin Chinese, or vernacular Chinese. In this brief presentation, we shall go through some digital-humanities topics of applying techniques of natural language processing and machine-learning methods, including deep learning, for extracting useful information from sources of classical Chinese texts [4].

The abilities to recognize and extract named entities (NEs) such as person and place names [15] and to infer about personal relationships and their social networks [13] are fundamental for assisting historical research. We may apply conditional-random-field models or deep learning methods to extract the NEs, and attempt to infer the grammar of classical Chinese when sufficient samples are available [11]. By reasonably automating the extraction of biographical information from historical documents [8], domain experts can build such research-oriented databases and platforms as CBDB [2], CBETA [3][6], DocuSky [5], and TBDB [1][14] at affordable costs.

With the availability of ample text collections of classical Chinese poems, we can explore the linguistic and literature worlds of poets with the help of computational tools [7][12]. We were not surprised to find that the distributions over Chinese characters in classical Chinese poems follow the Zipf's law. We may even study how poets organized their words to induce aesthetic imagery in the minds of their readers, from a certain analytical perspectives.

To verify the raw findings reported by algorithmic procedures still require non-negligible amount of close reading by human experts. We have not achieved the results mentioned completely automatically yet. The original classical Chinese texts do not have delimiters between consecutive words [7][10]. Unless added by human experts, most of the original classical texts do not use modern punctuation marks either, so we need to split sentence segments as well [9][10]. Furthermore, there are still a myriad of printed books in classical Chinese that need to be manually typed or algorithmically digitized, and optical character recognition for many ancient books remains a practical challenge due to the wide variety of page layouts and writing/printing styles.

致謝 (Acknowledgments)

Liu was supported in part by the contracts MOST-104-2221-E-004-005-MY3 and MOST-107-2200-E-004-009-MY3 of the Ministry of Science and Technology of Taiwan and in part by the mini-project 109H124D-09 of the National Chengchi University in Taiwan. The Chinese Biographical Database Project of Harvard University and the Harvard-Yenching library provide the data for some of the reported research, and we have discussed our work in OCR with Professor Donald Sturgeon of the Durham University and with a research team at the Academia Sinica Center for Digital Cultures (中研院數位文化中心) before.

References

- [1] S.-b. Chang, T.-h. Lee, Y.-l. Lee, C.-j. Li, Y.-w. Ku, H.-R. Ke, and S.-H. Sie. From CBDB to TBDB: The “Treatise of Historical Figures” of the new edition of Changhua Local Gazetteer as a starting point, *J. of Digital Archives and Digital Humanities*, 2:91–115, 2018.
- [2] China Biographical Database Project (CBDB): <https://projects.iq.harvard.edu/cbdb/home>
- [3] Chinese Buddhist Electrical Text Association (CBETA): <https://www.cbeta.org/>
- [4] J. Hsiang. Editorial for the inaugural issue: From digital archiving to digital humanities, *J. of Digital Archives and Digital Humanities*, 1:i–iv, 2018.
- [5] I M. Hung, C. Hu, and J. Hsiang. Exploring Guangxu-era missionary activities in Taiwan from Chinese Recorder, Dan-Hsin Archives and Ming-Qing Taiwan Administrative Archives through DocuSky, *Proc. of the 2020 Int’l Conf. on Digital Humanities*, 2020.
- [6] J.-J. Hung. CBETA research platform: A digital tool for studying Chinese Buddhist texts in the new era, *J. of Digital Archives and Digital Humanities*, 1:149–174, 2018.
- [7] C.-L. Liu and W.-T. Chang. Onto word segmentation of the Complete Tang Poems, *Proc. of the 2019 Intl Conf. on Digital Humanities*, 2019.
- [8] C.-L. Liu, W.-T. Chang, T.-Y. Zheng, and P.-S. Chiu. Toward building chronicles from biographies in local gazetteers: An application of syntactic and dependency parsing, *Proc. of the 2019 Int’l Conf. on Digital Humanities*, 2019.
- [9] C.-L. Liu and Y. Chang. Classical Chinese sentence segmentation for tomb biographies of Tang dynasty, *Proc. of the 2018 Int’l Conf. on Digital Humanities*, 231–235, 2018.
- [10] C.-L. Liu, C.-T. Chu, W.-T. Chang, T.-Y. Zheng. When classical Chinese meets machine learning: Explaining the relative performances of word and sentence segmentation tasks, *Proc. of the 2020 Int’l Conf. on Digital Humanities*, 2020.
- [11] C.-L. Liu, C.-K. Huang, H. Wang, and P. K. Bol. Mining local gazetteers of literary Chinese with CRF and pattern-based methods for biographical information in Chinese history, *Proc. of the 3rd Workshop on Big Humanities Data*, 2015 IEEE Int’l Conf. on Big Data, 1629–1638, 2015.
- [12] C.-L. Liu, T. J. Mazanec, and J. R. Tharsen. Exploring Chinese poetry with digital assistance: Examples from linguistic, literary, and historical viewpoints, *J. of Chinese Literature and Culture* (a special issue on Digital Methods and Traditional Chinese Literary Studies), 5(2):276–321, 2018.
- [13] C.-L. Liu and H. Wang. Matrix and graph operations for relationship inference: An illustration with the kinship inference in the China biographical database, *Proc. of the 2017 Annual Meeting of the Japanese Assoc. for Digital Humanities*, 94–96, 2017.
- [14] Taiwan Biography Database (TBDB): <http://tbdb.ntnu.edu.tw/>
- [15] T.-H. R. Tsai, C.-H. Wu, P.-L. Pai, and I.-C. Fan. Automatic labeled data generation for person named entity disambiguation on the Ming Shilu, *Proc. of the 2020 Int’l Conf. on Digital Humanities*, 2020.