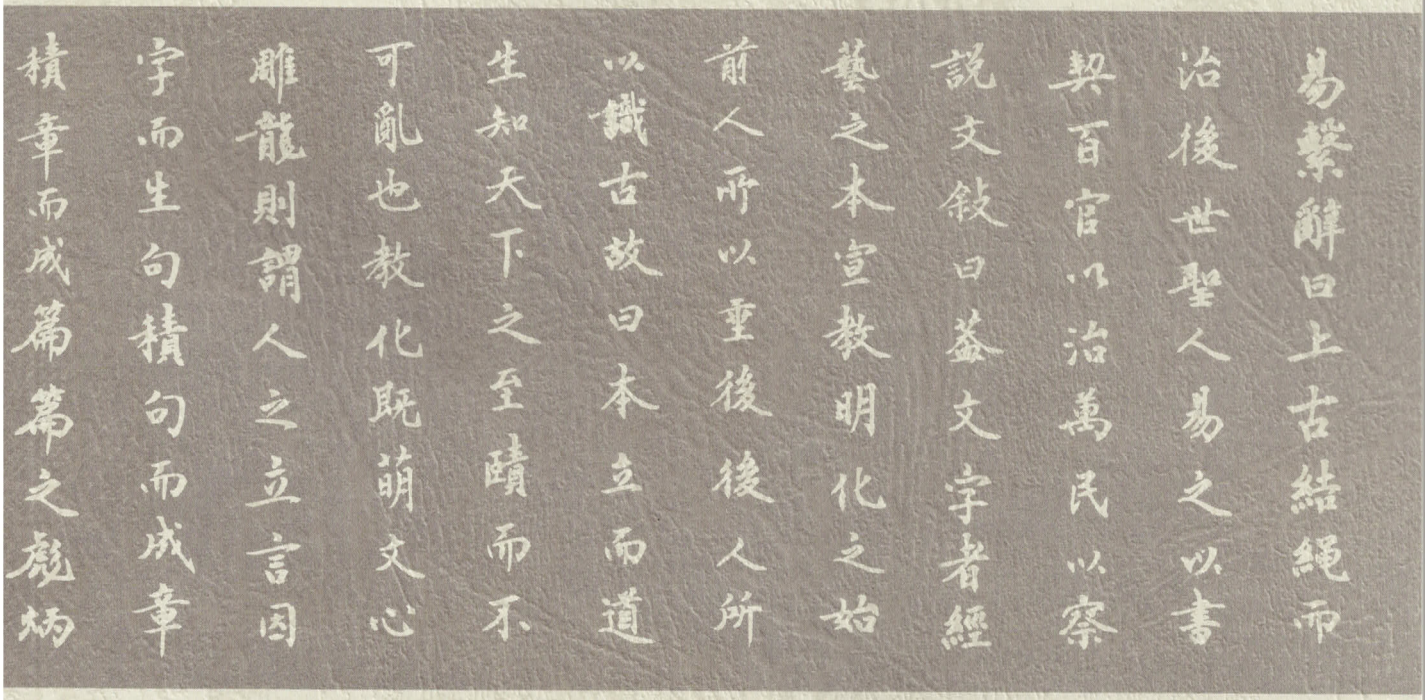
International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

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

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



Vol.25 No.2 December 2020 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

The Association for Computational Linguistics and Chinese Language Processing, Taipei

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

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

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

Membership & Subscriptions

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

Copyright

Cover

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

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 25, no. 2, December 2020

Contents

Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. i-ii

Forewords

The 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020) was held at GIS Convention Center of National Taipei University of Technology (NTUT) in Taipei, Taiwan during September 24-26, 2020. ROCLING 2020 is the leading and most comprehensive conference on computational linguistics and speech processing in Taiwan, which was hosted by the Association for Computational Linguistics and Chinese Language Processing. The conference brings together researchers, scientists, and industry participants from fields of natural language processing and speech processing to present their works and to discuss recent trends in the related subjects. This special issue presents the extended and reviewed versions of six papers meticulously selected from ROCLING 2020, including three natural language processing papers and three speech processing papers.

The first paper, which comes from National Taiwan Ocean University and National Taiwan University, focuses on analyzing the morphological structures in Seediq words. A set of morphological rules is created and involved in the developed system to detect the existence of infixes and suffixes. Besides, the structure of suffixes is predicted by probabilistic models. The second paper from National Central University presents a gated graph sequence neural network (GGSNN) model for Chinese healthcare named entity recognition. An extended character representation is derived based on multiple embeddings at 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. This paper is awarded as one of the two best papers of ROCLING 2020. The third paper, which comes from National Tsing Hua University and National Chung Hsing University, proposes a statistical method to extract translations of nouns and prepositions from Chinese-English bilingual parallel corpora, which is then used to improve phrase translation based on sentence alignment. In their developed system, a user inputs an English phrase with a noun and a preposition, and the system retrieves translations with example sentences for showing to the user. The fourth paper, which comes from National Sun Yat-sen University and Chunghwa Telecom Laboratories, describes a speaker verification system for the 2020 Far-field Speaker Verification Challenge (FFSVC 2020). A TDResNet, which combines the advantages of both TDNN and CNN, is proposed to extract informative acoustic representations. Next, two methods are evaluated to encapsulate a set of acoustic features. Finally, a PLDA-based classifier is used to predict the results. The fifth paper, which is proposed by National Taipei University of Technology and Chunghwa Telecom Laboratories, concentrates on proposing a Chinese Text-to-Taiwanese speech synthesis system. The system mainly composes of a machine translation module, a text-to-spectrogram component, and a

spectrogram-to-waveform synthesis system. The MOS score of the system is 4.30, which confirms the effectiveness of the speech synthesis system. The last paper from National Chi Nan University focuses on developing a novel speech feature extraction technique to achieve noise-robust speech recognition. Instead of retraining and adapting the complicated acoustic models, a neural network-based model is proposed to learn the front-end acoustic speech feature representation that can achieve the maximum state accuracy obtained from the original acoustic models.

The Guest Editors of this special issue would like to thank all the authors and reviewers for sharing their knowledge and experience at the conference. We hope this issue can provide useful insights for directing and inspiring new pathways of natural language processing and speech processing studies within the research field.

Guest Editors

Lung-Hao Lee

Department of Electrical Engineering, National Central University, Taiwan

Kuan-Yu Chen

Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taiwan

Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. 1-20

Analyzing the Morphological Structures in Seediq Words

Chuan-Jie Lin*, Li-May Sung+, Jing-Sheng You*,

Wei Wang*, Cheng-Hsun Lee*, and Zih-Cyuan Liao*

Abstract

NLP techniques are efficient to build large datasets for low-resource languages. It is helpful for preservation and revitalization of the indigenous languages. This paper proposes approaches to analyze morphological structures in Seediq words automatically as the first step to develop NLP applications such as machine translation. Word inflections in Seediq are plentiful. Sets of morphological rules have been created according to the linguisitic features provided in the Seediq syntax book (Sung, 2018) and based on regular morpho-phonological processing in Seediq, a new idea of "deep root" is also suggested. The rule-based system proposed in this paper can successfully detect the existence of infixes and suffixes in Seediq with a precision of 98.88% and a recall of 89.59%. The structure of a prefix string is predicted by probabilistic models. We conclude that the best system is bigram model with back-off approach and Lidstone smoothing with an accuracy of 82.86%.

Keywords: Seediq, Automatic Analysis of Morphological Structures, Deep Root, Natural Language Processing for Indigenous Languages in Taiwan, Formosan Languages

* Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan
E-mail: {cjlin, 10857039, 00657120, 00657140, 00672042}@email.ntou.edu.tw
+ Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan
E-mail: limay@ntu.edu.tw
The authors for correspondence are Chuan-Jie Lin and Li-May Sung.

1. Introduction

1.1 Motivation

Machine learning and deep learning have been the most popular techniques in recent days. Systems built by machine learning or deep learning often achieve good performance, but the scale of the training sets in general should be large enough. Comparing to English, the amount of resources in Mandarin is far more small, not to mention the resources in the Southern Min, Hakka, even the indigenous languages in Taiwan. The United Nations has declared the Year of 2019 as the International Year of Indigenous Languages[1] in order to highlight the preservation issues of these endangered languages and gain more attention from the world. Following the same spirit, in January 2019 Taiwan has also promulgated the National Languages Development Act[2] (國家語言發展法) to speed up the preservation and revitalization of the indigenous languages in Taiwan.

The indigenous languages in Taiwan, well-known as Formosan languages[3] (台灣南島語言) in the Austronesian languages (南島語系) family, include 16 languages of 42 dialects in total. All are endangered to some degree according to the investigation by UNESCO in 2009. So far we have not found many researches on the natural language processing of the Formosan languages. Collaborating with a linguist and an expert in Seediq, one of the authors, this paper aims to provide an innovative first step on Seediq. In addition, we expect that the research results can be applied to linguistically related languages, Atayal (泰雅語) or Truku (太魯閣語) (Li, 1981), without much effort, or even to Amis (阿美語) which has the largest population of speakers and a similar writing system to Seediq.

The morphology in Seediq is quite complicated, including many word inflections to represent verbal focus, aspect and causation etc. For example, the morphological structure of the word "*psetuq*" (break, 斷) is "*p-setuq*", and the structure of the word "*qnyutan*" (bite, 咬) is "*q<n>yuc-an*", where "*p-*" (CAU, causative, 使動), "*<n>*" (PRFTV, perfective aspect, 完成貌), and "*-an*" (LV, locative voice, 處所焦點) are prefix, infix, and suffix, respectively. As we will discuss later, it is not easy to decompose the affixes and the stem in a Seediq word, but they carry important information for NLP tasks such as machine translation. This paper proposes the automatic approaches to analyze the morphological structures in Seediq as the first step of machine translation or other NLP tasks.

There is no large corpus in Seediq available so far. The experimental data in this paper came from the book "賽德克語語法概論" (A Sketch Grammar of Seediq) (Sung, 2018)

[1] https://en.iyil2019.org/

[2] https://law.moj.gov.tw/LawClass/LawAll.aspx?pcode=H0170143

[3] https://zh.wikipedia.org/wiki/台灣南島語言

(referred to as ***the Seediq syntax book*** hereafter). This book provides many sentences with morphological information as illustrations. We used these data to construct the training set. Dr. Li-May Sung, the author of the Seediq syntax book and one of the authors of this paper, provided another batch of sentences tagged with morphological information as well. We used them to construct the test set. There are 394 and 322 affixed Seediq words in these two datasets, far less than the necessary amount to train a classifier by machine learning or deep learning. One additional Seediq resource is an online Seediq dictionary "賽德克語德固達雅方言" (Tgdaya Seediq, referred to as ***the CIP Seediq dictionary*** hereafter) (Sung, 2011), compiled by Dr. Sung for the Council of Indigenous Peoples (原住民族委員會). There are about 5,600 words in this dictionary but with no morphological analysis. In the future we will apply the techniques developed in this paper to analyze these dictionary words in order to build up a larger dataset.

1.2 Related Work

To our best knowledge, there are not many researches on natural language processing of the Formosan languages. The most related studies are the ones done by Dr. Meng-Chien Yang in Tao (達悟語, aka. Yami 雅美語), including the construction of a wordnet and a lexicon in Yami (Yang & Rau, 2011; Yang *et al*., 2011; Rau *et al*., 2015), and machine translation between Yami and Mandarin under a small bilingual corpus (Yang & Rau, 2015).

There are many NLP studies for other local languages in Taiwan though, including machine translation for Taiwanese (Lin & Chen, 1999), speech recognition and synthesis in Taiwanese (Iunn *et al.*, 2007; Yu & Lin, 2012), and prosodic models in Hakka (Gu *et al*., 2007; Chiang, 2018). As we know that Taiwanese Southern Min, Hakka, and Mandarin belong to the Sinitic languages (漢語群), and they do not share similar language structure with the Formosan languages. Thus the research results cannot be applied directly to the Formosan languages.

In addition, there are limited electronic resources in Seediq available in the Internet. The CIP Seediq dictionary contains 5,595 words and 6,019 sentences with Mandarin translations. It is the largest dataset we can find so far. There are also textbooks for the elementary, junior-high and high schools available in "原住民族電子書城[4]" (Taiwanese Indigenous ebooks) and "族語 E 樂園[5]" (Formosan Languages E-Land), but their amounts are still comparatively small with no morphological analysis. Only sentences in the Seediq syntax book are tagged with morphological information.

A Seediq ontology was built by Dr. Shu-Kai Hsieh and Dr. Chu-Ren Huang (Hsieh *et al.*,

[4] https://alilin.apc.gov.tw/tw/ebooks

[5] http://web.klokah.tw/

2007). It contains 270 Seediq words mapping to the senses of WordNet in English in order to study the hyponymy relationships between Seediq words. As the ontology only covers a small set of Seediq words and provides mainly semantic information, we will not use it in this paper.

The development of a machine translation system usually requires a large size bilingual corpus in order to train a good-quality MT system by machine learning or deep learning (Bahdanau *et al*., 2015; Luong *et al*., 2015). It is important to create a large corpus efficiently by the help of NLP techniques, and this is the main goal we plan to do on Seediq in this paper.

2. Introduction to Seediq

2.1 Seediq Writing System

Seediq as one of the Formosan languages, the Seediq people mainly live in Nantou County and Hualien County. Linguistically belonging to the Atayalic subgroup (Li, 1981), Seediq is closely related to Atayal (泰雅語) and Truku (太魯閣語).

It has three dialects, including Tgdaya (德固達雅), Toda (都達), and Truku (德路固). Our experimental data came from the Seediq syntax book "賽德克語語法概論" (Sung, 2018) focusing on the Tgdaya dialect. Most morphological information about Seediq provided in this section also came from this syntax book.

The Seediq writing system follows the definition of "原住民族語言書寫系統" (writing systems of Formosan languages) published by the Ministry of Education and the Council of Indigenous Peoples on December 15th, 2005. It is a Romanization system. There are 18 consonants (including 2 half-vowels) and 5 vowels in Seediq. An example of a Seediq sentence is as follows.

[Seediq] Teta su kmkelun psetuq qnyutan su!
[Chinese] 看你咬得斷咬不斷！
(English: See if you can bite this off!)

2.2 Morphology in Seediq

The Seediq syntax book (Sung, 2018) provides detailed morphological information in each exemplar sentence to help the reader understand Seediq more efficiently. Words, especially the verbs, in the sentence are affixed to indicate actor voice (AV), patient voice (PV), locative voice (LV), beneficiary/instrumental/referential voices (BV/IV/RV), *etc*., and aspects such as perfective aspect (PFV). Affixation is overwhelmingly prevailing in Seediq. Such information is very useful in our study. One example of the morphological information is as follows.

[Morpho Info] teta=su kmekul-un p-setuq q<n>iyuc-an=su
[Explanation] 看看=你.屬格 能夠-受事焦點 使動-斷 <完成貌>咬-處所焦點=你.屬格
(English: See=you.GEN able-PV CAUS-break <PFV>bite-LV = you.GEN)

In this example, the root of the word "*qnyutan*" (bitten by, 被..咬) is "*qiyuc*". This word is affixed with a suffix "*-an*" (LV, locative voice, 處所焦點) and an infix "*<n>*" (PFV, perfective aspect, 完成貌), and becomes "*q<n>iyuc-an*". Similarly, the root of the word "*psetuq*" (broken, 使斷) is "*setuq*". This word is affixed with a prefix "*p-*" (CAUS, causative, 使動) and becomes "*p-setuq*". (GEN means genitive case. The symbol '=' represents the attachment of pronouns and other cases. It will not be discussed in this paper.) Several examples of word inflections are provided in Table 1.

Table 1. Examples of Seediq Word Inflections

Seediq	**Root**	**Morphological Structure**	**Meaning (Seediq / Root)**
mpkbeyax	beyax	m-p-k-beyax	hard-working, 努力 / do with force, 用力
cmnebu	cebu	c<m><n>ebu	shot successfully, 打中了 / shoot, 擊射
qyaanun	qeya	qeya-an-un	hang, 掛 / hang, 掛
pndsanan	adis	p<n>adis-an-an	bring back, 帶回 / bring, 帶

Notes: "m-": AV, agent voice 主事焦點; "p-": FUT, future 未來 or CAUS, causative 使動; "k-": STAT, stative 靜態; "<m>": AV, actor voice 主事焦點; "<n>": PFV, perfective aspect 完成貌; "-an": LV, locative voice 處所焦點; "-un": PV, patient voice 受事焦點

Another type of prefixes is reduplication (RED, 重疊) which repeats some part of the word. It is used for plurality, intensification, and *etc*. For example, the word "*sseediq*" ("*s-seediq*") (RED-person, 重疊-人) means "many people", and the word "*mkrkere*" ("*m-kr-kere*") (AV-RED-strong, 主事焦點-重疊-強壯) means something is very strong. Even prefixes can be repeated, such as in the word "*pposa*" ("*p-p-osa*") (RED-CAUS-go, 重疊-使動-去) which means "forced to go to somewhere". The reduplication usually does not change the meaning of a word but its amount or intensity, which could also be an issue in machine translation.

When a Seediq word is affixed, the final writing form can be different from its original combination, as we can see in the examples in Table 1. This is the reason why the morphological structure of a Seediq word cannot be generated directly from its surface form.

We discuss only three variation cases here (Sung, 2018; Yang, 1976; Li, 1977; Li, 1991).

The first case is related to vowel neutralization (元音中性化) and vowel reduction (元音脫落). In Seediq, vowels other than the last two syllables are weakened (neutralized) and omitted when writing. It usually happens in the suffixation process in Seediq. Take examples from Table 1. In the word "*qyaanun*" ("*qeya-an-un*"), the first vowel "*e*" of its root "*qeya*" is omitted when affixed. And in the word "*pndsanan*" ("*p<n>adis-an-an*"), both vowels of its root "*adis*" are omitted.

Consider another example. The word "*dngei*" ("*dengu-i*") consists of a root "*dengu*" (sun-dry; 曬乾) and a suffix "*-i*" (IMP, imperative, 祈使). We suggest that the root word "*dengu*" may originally be "*denge*": that is, the second vowel '*e*' is neutralized as '*u*' when it appears at the end of a word. When "*denge*" is suffixed with "*-i*", the accent falls on the second vowel '*e*' (hence not neutralized any more) and makes it remain as "*e*"; meanwhile, the first vowel "*e*" of "*denge*" is neutralized and omitted, resulting in "*dngei*". That is, the word "*dngei*" comes from the original structure of "*denge-i*". We refer to such original form of a root as its "deep root" and will discuss it in details in Section 3.1.

The second case is about vowel harmony (元音和諧變化). When a root word starts with a vowel, the preceding prefix usually ends with the same vowel. For example, if the prefix "*s-*" (RV, referential voice, 參考焦點) attaches to the root "*osa*" (go, 去), the prefix becomes "*so-*" and the final writing form is "*soosa*" ("*so-osa*").

The third case is about word-final consonant mutation (詞尾輔音變化). Some word-final consonants will be changed if there is no suffix attached. When such a word is suffixed, its final consonant changes back to the original one. Take the word "*qnyutan*" (bite, 咬) as an example. Its root "*qiyuc*" is in fact the result of word-final consonant mutation from its original form (deep root) "*qiyut*". When "*qiyuc*" is attached with a suffix "*-an*" (LV, locative voice, 處所焦點), the final consonant '*c*' changes back to '*t*' and the affixed word is in fact "*q<n>iyut-an*" and the final writing form is "*qnyutan*" (note that the first vowel '*i*' of the root is omitted).

Word inflections in Seediq are overwhelmingly plentiful. In the CIP Seediq dictionary, for example, there are 39 words relating to the same root "*adis*" (bring, 帶走): *desan, dese, desi, deso, desun, dnsanan, dsanan, dsane, dsani, dsanun, dsdesan, dsdesi, dsdesun, knddesi*, *maadis, madis, mdaadis, mkdesun, mkmadis, mnadis, nadis, paadis, pdaadis, pdesan, pdese, pdesi, pdeso, pdesun, pdsanan, pdsane, pdsani, pdsanun, pnaadis, pnadis, pndesan, pndsanan*, *ppaadis, saadis*, and *spaadis*.

3. Seediq Morphological Structure Analysis

The main issue focused in this paper is: when we have a Seediq word and its root word, we want to know its morphological structure, i.e. the combination of prefixes, infixes, and suffixes in that word. In the CIP Seediq dictionary, words and their roots are available. By the techniques developed in this paper, we can generate those words' morphological structures automatically and efficiently.

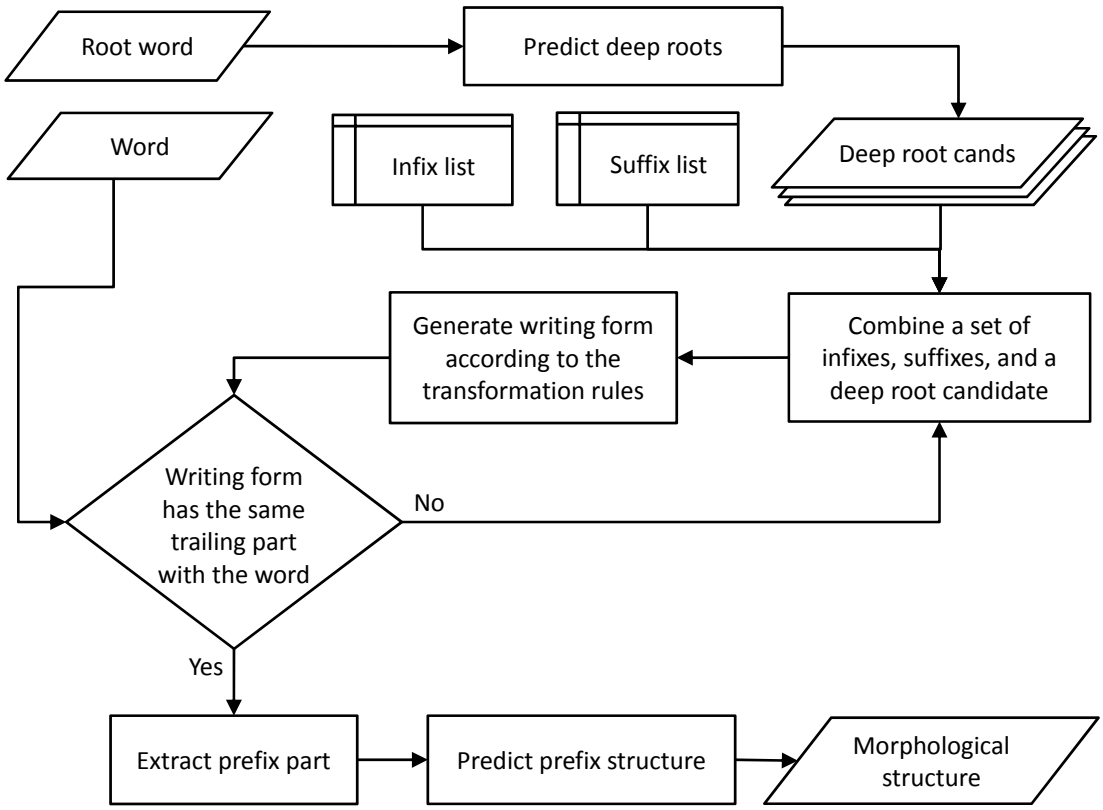


Figure 1. Flowchart of Automatic Seediq Morphological Structure Analysis

Figure 1 demonstrates our proposed flowchart to analyze Seediq morphological structure automatically. Take the word "*pnsltudan*" (whose root word is "*lutuc*") as an example to explain the flowchart. First, a list of deep root candidates {"*lutud*", "*lutuc*"…} of the root word "*lutuc*" is prepared by the method introduced in Section 3.1. (The definition of *deep root* is also given in Section 3.1.) Each deep root candidate is combined with a set (or none) of known infixes and suffixes to form a partial morphological structure (cf. Section 3.2). For example, by selecting a deep root "*lutud*", no infix, and a suffix "*-an*", we will have a partial morphological structure "*lutud-an*". Transformation rules (described in Section 3.3) are then

applied to the partial structure and its writing form "*ltudan*" will be generated (note that the first vowel '*u*' and all the structural symbols are omitted). Since that "*ltudan*" is exactly the trailing substring of the given word "*pnsltudan*", the leading substring "*pns*" is extracted as the prefix part, and its structure "*p<n>s*" is decided by prefix structure analysis methods (as discussed in Section 3.4). Finally, the overall predicted morphological structure of the word "*pnsltudan*" is "*p<n>s-lutuc-an*". Note that we still use root words in the morphological structures, not the deep roots.

3.1 Deep Root Prediction

As discussed briefly above, some root words (原形詞) when suffixed in Seediq will change back to their original forms before vowel neutralization or word-final consonant mutation (cf. Section 2.2). We refer to such original form of a root word as its ***deep root*** (深層原形). For example, when generating writing form of the word "*p-adis-o*", the root word "*adis*" should be replaced with its deep root "*ades*", so that, by omitting neutralized vowels, "*p-ades-o*" becomes the correct writing form "*pdeso*" (bring, 帶).

Table 2 provides more examples of deep roots. All four root words in Table 2 have the same trailing substring "*uk*". However, when they are attached with the suffix "*-i*", the result words (in the third column) do not all end with "*uki*" but change into different trailing substrings. That is because the deep root of "*aduk*" is "*adup*", the deep root of "*ciyuk*" is the same as the root word, the deep root of "*dehuk*" is "*dehek*", and the deep root of "*eluk*" is "*eleb*".

Table 2. Examples of Deep Roots

Root Word	Suffixed Structure	Word	Struct w. Deep Root
aduk (repel, 趕走)	aduk-i	dupi	adup-i
ciyuk (reply, 回覆)	ciyuk-i	ciyuki	ciyuk-i
dehuk (arrive, 到達)	dehuk-i	dheki	dehek-i
eluk (close door or window, 關門窗)	eluk-i	lebi	eleb-i

Predicting deep roots is not an easy task. Neither dictionaries nor syntax books provide information of deep roots. Vowel neutralization or word-final consonant mutation could also be many-to-one mapping. In the following we discuss how we gain a list of deep roots.

Method 1. Inductive Deep Root Prediction

Our first proposed method is based on inductive method. From the CIP Seediq dictionary, we can collect a set of suffixed words referencing to the same root word. The most frequent common trailing substrings among the suffixed words is extracted as its deep root. Note that it

may be identical to the root word, but we are only interested in the transformed deep roots. Details of steps to predict deep roots as well as some examples are given as follows.

Step 1. Collect words referencing to the same root word. In the CIP Seediq dictionary, "參考條目" (cross reference) often provides the root word information. For example, as shown in Section 1.2, 39 words including *desan*, *dese*, *desi*, *etc*., all refer to the same root word "*adis*". Words referencing to the same root word should have the same deep root.

Step 2. Select words with suffixes. Only suffixed words will reveal its deep root, so we need to decide if a word is suffixed or not. Note that the CIP Seediq dictionary does not provide detailed morphological structural information.

If a word ends with its root word, it is not suffixed. For example, both "*ppaadis*" and "*maadis*" end with their root word "*adis*", so there is no suffix in these two words.

If a word ends with its root word after removing possible infixes, it is not suffixed. For example, the word "*cmnebu*" does not ends with its root word "*cebu*". By removing infixes "<*m*><*n*>" after the first consonant '*c*', this word appears exactly the same as its root word and hence not suffixed.

Words other than the two cases above and ending with known suffixes are considered as suffixed words.

Step 3. Predict deep roots by induction. When there is only one vowel in a suffix, the last vowel of the suffixed deep root will not be omitted. We can check if these words end with the same trailing substring and decide the deep root. For example, the structures of the words "*desan*", "*dese*", and "*desi*" are "*ades-an*", "*ades-e*", and "*ades-i*", respectively. After removing suffix parts, they all end with "*es*". Moreover, the preceding consonant '*d*' appears in the root word "*adis*". By replacing the trailing substring of the root word with the most common substring induced from these suffixed words, we can obtain the deep root "*ades*".

Unfortunately, some root words do not have enough related words to induce their deep roots. Moreover, in some rare cases, we found two different deep roots related to the same root word. In order to increase the coverage of deep root prediction and morphological analysis, below we further propose a mapping table for the deep root prediction.

Method 2. Deep Root Mapping Table

The deep root mapping table lists the mapping of trailing substrings between root words and their deep roots. This table is constructed from the <root_word, deep_root> pairs collected by Method 1. For example, the pairs in Table 2 of Section 3.1 tell us that a root word ending with

"*uk*" may have a deep root ending with "*up*", "*ek*", or "*eb*". These mappings are saved in the deep root mapping table. The real data show that "*uk*" maps to "*ek*" for 4 times, "*up*" for twice, and "*eb*" for once.

Since deep roots are closely related to the processes of vowel neutralization and word-final consonant mutation, we only need to consider trailing substrings consist of the last vowel and the word-final consonant. For example, the extracted trailing substring of the word "*aduk*" is "*uk*" and trailing substring of the word "*beebu*" is "*u*".

Figure 2 illustrates the steps of building the deep root mapping table. First, by applying Method 1 to the words in the CIP Seediq dictionary, a set of predicted <root_word, deep_root> pairs are collected. Words in the pairs are then replaced with their trailing substrings. Finally, by counting the mapping pairs, a mapping table is constructed where the mappings are sorted by their frequencies.

When we do not know the deep root of a root word, we can still propose deep root candidates by replacing the trailing substrings according to the deep root mapping table. For example, we know that the root word of "*hligan*" is "*haluy*" but we do not know its deep root. According to the mapping table, the trailing substring "*uy*" often maps to "*ig*". By replacing the trailing substring, we guess that its deep root is "*halig*". The result structure of "*halig-an*" matches indeed with the target word "*hligan*".

To recap, deep root candidates in Figure 1 are generated according to the following order:

(1) The deep root induced from the CIP Seediq by Method 1 (if any)

(2) The original root word

(3) Trail-replacement results according to the deep root mapping table built by Method 2

3.2 Affixation with Infixes and Suffixes

With the list of deep root candidates of the root words, we then move to the next step. Each deep root candidate will be combined with known infixes and suffixes for string matching in the next steps. Infixes and suffixes are first considered because their sets are rather fixed; we only see 3 kinds of infixes {"<*m*><*n*>", "<*m*>", "<*n*>"} and 10 kinds of suffixes {"*-an-an*", "*-an-un*", "*-ane*", "*-ani*", "*-ano*", "*-an*", "*-un*", "*-e*", "*-i*", "*-o*"} in the training set. Note that an infix appears after the first consonant. For example, when the word "*quyux*" is infixed with "<*m*>", it becomes "*qmuyux*" ("*q*<*m*>*uyux*") (raining, 下雨). But if a root word starts with a vowel, the infix appears at the beginning of the word. For example, when the word "*apa*" is infixed with "<*n*>", it becomes "*napa*" ("<*n*>*apa*") (carry, 揹). For convenience, we leave the infixes in such cases together with the prefix part (extracted in Section 3.3) to be processed in Section 3.4.

3.3 Transformation Rules to Generate Writing Form

For the case when a root word is affixed with only prefixes and infixes, its writing form can be derived directly from the combination by removing structural symbols. For instance, "*m-p-k-beyax*" becomes "*mpkbeyax*" (work hard, 努力) and "*h<m>aduc*" becomes "*hmaduc*" (send, 送).

But when a root word is suffixed, two cases should be considered. The first case is vowel reduction (元音脫落) where vowels other than the last two are neutralized and omitted. For example, "*hetur-ani*" (block out, 擋) becomes "*htrani*" where the first two vowels '*e*' and '*u*' are omitted. The rule of vowel reduction can be applied by programs easily.

One exception of vowel reduction happens when only one of the two adjacent identical vowels is about to be omitted. In such case, both vowels will not be omitted. Take "*osa-an-un*" (go, 去) as an example. According to the general rule of vowel reduction, both vowels '*o*' and '*a*' in the root word "*osa*" should be omitted. However, the second vowel '*a*' of the root word "*osa*" is followed by the suffix "*-an*" which starts with the same vowel. Therefore, the second vowel '*a*' is not omitted, and the final writing form becomes "saanun" where only the first vowel '*o*' is omitted.

The second case is the addition of '*y*' or '*w*'. We find some cases that a '*y*' or '*w*' is added between the root word and the suffix. For example, the final writing form of "*chungi-an*" is "*chngiyan*" (forget, 忘記) and the final writing form of "*cebu-an*" is "*cbuwan*" (to be shot, 被擊中). We have not figured out the rules for such cases. Currently we simply insert a '*y*' or '*w*' to see if the transformation result matches the final writing form.

The complete transformation rules to generate the final writing form are defined as follows. Given a morphological structure represented as *pfx-root<ifx>str-sfx*, *rootstr* is the root part (root word or deep root), *pfx* is the prefix part, *ifx* is the infix part, and *sfx* is the suffix part. Any affix part may be empty. The writing form of the morphological structure is generated by the following steps:

Step 1. When the suffix part is not empty, the last two vowels in the structure remain unchanged. Vowels other than the last third one are all omitted. As for the last third vowel,

a) If it is the same as the last second vowel and they are adjacent to each other, the last third vowel remains unchanged

b) Otherwise the last third vowel is omitted

Step 2. If the suffix *sfx* starts with a vowel but is different from the word-final vowel of rootstr, one '*y*' or '*w*' may be inserted between them to generate a correct writing form.

Step 3. Remove all morphological structural symbols (including -, <, and >).

An interim summary for Section 3.1 to Section 3.3: Given a Seediq word and its root word, a list of deep root candidates is generated by methods proposed in Section 3.1. Each deep root candidate is then combined with every infix and suffix (including empty strings) as described in Section 3.2. Each combination is then transformed into the writing form by rules explained in this Section 3.3. If this writing form matches the trailing substring of the target Seediq word, this combination of deep root, infix and suffix is proposed as the predicted morphological structure, and the unmatched part is extracted as the prefix part for further analysis by methods proposed next in Section 3.4.

3.4 Prefix Structure Analysis

The prefix structure analysis also encounters ambiguity problem. One prefix string can be segmented into several different prefix combinations. For example, the prefix string "*kn-*" can be either "*kn-*" (NMLZ, nominalization, 名物化) or "*k<n>-*" (STAT<PFV>, 靜態<完成貌>), and the prefix string "*sk-*" can be either "*sk-*" (deceased, 已故) or "*s-k-*" (existential-STAT, 有-靜態).

To our best effort, we so far cannot find much information about prefix combinations. To solve the prefix problem in Seediq, we here propose several approaches similar to the classical solutions for Chinese word segmentation, including probability models and machine learning, which will be discussed in details below. Our goal is to find the best system in which we can predict the morphological structures of words in the CIP Seediq dictionary with high accuracy in order to reduce the effort of human checking in the future.

First of all, we need to prepare a list of atomic prefixes. There are 29 atomic prefixes found in the Seediq syntax book, including {"*k-*", "*n-*", "*kn-*", "*m-*"…}. We further found 10 different atomic prefixes in the test data, including {"*de-*", "*gn-*", "*km-*"⋯}. The following experiments are based on these atomic prefixes. We do not know whether there will be more new atomic prefixes in the CIP Seediq dictionary or not.

Reduplication (introduced in Section 2.2) also appears in the prefix part. It is used to emphasize the amount of something or the intensity of an action. It can be attached to a root word or an atomic prefix. It repeats either the first consonant (e.g. "*s-*" in "*s-seediq*" and the first "*p-*" in "*p-p-heyu*"), or the first consonant with the word-initial vowel (e.g. "*le-*" in "*k-le-eluw*"), or the first two consonants (e.g. "*kr-*" in "*m-kr-kere*").

During training, all reduplication prefixes are replaced with a special symbol and treated as one type of the atomic prefixes. Therefore, there are totally 40 types of atomic prefixes in the experiments in Section 4. When segmenting a prefix string, a segment matching any of the 3 reduplication cases shown in the previous paragraph is considered to be a reduplication prefix.

Probability Models

One common approach for Chinese word segmentation is to build probability models. In a similar way, we propose unigram and bigram models for prefix structure analysis in Seediq. Given a prefix string *pfx* and one of its segmentation $x_1x_2\ldots x_m$ where x_i is an atomic prefix, the probability of this segmentation is defined as follows, where \$ denotes the beginning of the prefix string.

$$\text{Unigram model: } P(pfx) = \prod_{i=1}^{m} P(x_i) \quad (1)$$

$$\text{Bigram model: } P(pfx) = P(x_1|\$) \prod_{i=2}^{m} P(x_i|x_{i-1}) \quad (2)$$

Because the amount of training data is not large enough, we still need to apply smoothing methods to avoid zero probabilities. But some well-known smoothing methods such as Witten-Bell or Good-Turing are good for large training data. We did not choose them in this paper. Instead, we use Lidstone smoothing to build our unigram model. That is, the frequency of each atomic prefix (seen or unseen) is added with a value λ before building the probability model. Let *N* be the original sum of the frequencies of all atomic prefixes and *B* be the number of types of atomic prefixes. Lidstone smoothing will assign a probability of λ / (*N*+*B*λ) to each unseen atomic prefix.

We use back-off approach to deal with zero probabilities in the bigram model. That is, we consider the unigram probability (weighted by an α value) of the second prefix in an unseen bigram. When $P(x|y)=0$, we use $P(x|y)=\alpha P(x)$ instead.

The unigram model provides the probabilities of 40 atomic prefixes. The bigram model provides the probabilities of bigram of these 40 prefixes and the starting sign \$ (thus 41×40 types of bigrams). An unknown prefix x_i or a bigram containing such an unknown prefix has no probability. Smoothing is designed for known but unseen atomic prefixes in our work.

The steps of prefix structure analysis are as follows. Given a prefix string *pfx*, all segmentations $x_1x_2\ldots x_m$ are enumerated by inserting one or zero '-' between any two adjacent letters. For example, the prefix string "*mss-*" can be segmented into {"*mss-*", "*m-ss-*", "*ms-s-*", "*m-s-s-*"}. The segmentation having the best probability is selected as the final answer. Note that the strings "*mss-*" and "*ss-*" do not appear in the list of atomic prefixes and thus have no probability; so the probability of "*mss-*" and "*m-ss-*" is also 0.

Machine Learning and Deep Learning Methods

Machine learning methods are also tried to guess the prefix structure. However, we have too little features so far, and the only features we know are contextual information and the list of atomic prefixes. More useful features need to be discovered in the future. The following example illustrates the features of each letter in the prefix string "*psq-*" where its correct structure is "*ps-q-*".

c	c_{-2}	c_{-1}	c_1	c_2	[B	E	S]	Class
p	$	$	s	q	1	0	1	B
s	$	p	q	$	1	1	1	E
q	p	s	$	$	1	0	1	S

The feature c_k denotes the letter in the context. The Boolean features [B E S] denotes the position where this letter appear in the atomic prefixes. For example, the [B E S] values of the letter '*p*' are [1 0 1] because it appears in the beginning (B) of some atomic prefixes {"*pn*-", "*ps*-"…}, and it can be a single-letter prefix "*p*-" itself (S). E means appearing the end of an atomic prefix. Note that no atomic prefix is longer than 2 letters in our datasets. The final classification is also one of the BES labels.

In addition, deep learning methods such as the encoder-decoder model are explored. The input is the prefix string where letters and the symbol '-' are denoted by one-hot encoding. The output is the prediction of morphological structure.

4. Experiments

4.1 Experimental Datasets

The first dataset comes from the Seediq syntax book "賽德克語語法概論". There are 509 sentences provided as illustrations in this book. The morphological structures of words in the sentences are also provided. There are 817 distinct Seediq words appearing in the sentences and 394 of them contain affixes. We took these 394 affixed words as the training data.

The second dataset comes from 515 new sentences provided by Dr. Li-May Sung, the author of the Seediq syntax book and one of the authors of this paper. These sentences are also tagged with morphological structures. 322 new Seediq words with affixes are extracted from these sentences as the test data.

4.2 Infix and Suffix Detection Experiments

Sections 3.1 ~ 3.3 propose approaches to detect deep root, prefix, infix, and suffix parts in a given Seediq word (in which the structure inside the prefix part has not been predicted). Table 3 lists the performance of these approaches, where precision is the percentage of system-detected units (words or affixes) being correct, and recall is the percentage of gold-standard units being detected by the system.

Table 3. Performance of Infix and Suffix Detection

	Training Data					**Test Data**				
Unit	**Gold**	**System**	**Correct**	**P (%)**	**R (%)**	**Gold**	**System**	**Correct**	**P (%)**	**R (%)**
Word	394	357	353	98.88	89.59	322	286	278	97.20	86.34
Infix	79	77	77	100.0	97.47	55	47	47	100.0	85.45
Suffix	169	135	135	100.0	79.88	127	98	98	100.0	77.17
Prefix	221	207	203	98.07	91.86	194	186	180	96.77	92.78

For more details, the third row of Table 3 shows that 79 of the 394 words in the training set contain infixes, where 77 of them can be detected by the system (recall 77 / 79 = 97.47%) and all of them are correct (precision 77 / 77 = 100%). All precision scores of prefix, infix, and suffix detections are around 98% to 100%. Recall scores are a little lower, because 37 of the 394 affixed words are exceptions of morphological rules.

4.3 Prefix Structure Analysis Experiments

In the training set, only 221 words are prefixed as shown in Table 3. 116 of them are prefixed by one single-letter prefix and hence no further analysis is needed. Therefore, the training set of prefix structure analysis contain only 105 words whose prefix parts are longer than one letter. When evaluating on the training set, we adopt leave-one-out cross-validation method due to the small amount of data. Each word is predicted by the classifier trained with the other 104 words.

Table 4. Performance of Prefix Analysis by Unigram Models

	Training Data			**Test Data**		
λ	**Word**	**Correct**	**A (%)**	**Word**	**Correct**	**A (%)**
0	**105**	**86**	**81.905**	103	61	59.223
0.1	**105**	**86**	**81.905**	103	64	62.136
0.3	105	85	80.952	103	65	63.107
0.5	**105**	**86**	**81.905**	103	69	66.990
1	105	85	80.952	103	71	68.932
2	105	81	77.143	103	83	80.583
3	105	79	75.238	103	86	83.495
4	105	81	77.143	103	86	83.495
5	105	79	75.238	**103**	**87**	**84.466**

As for the testing set, there are 194 prefixed words in the test set and 103 of them have prefixes longer than one letter. The metric of evaluation is accuracy. The prefix structure prediction has to be exactly the same as the gold standard to be counted as "correct".

The experimental results of unigram models with different λ values are shown in Table 4. The λ value does not affect much performance on the training set. It means that most unseen prefixes only appear in the test set. Interestingly, when λ value is set to be 3 or larger, the performance on the test set is improved in a great degree. It seems to hint that we need a training set where each atomic prefix should appear at least 3 times.

The experimental results of bigram models with different λ values are listed in Table 5. Again, the λ value does not affect much performance on the training set, but improves the performance on the test set a lot when it is set to be 2 or larger.

Table 5. Performance of Prefix Analysis by Bigram Models with Different λ

		Training Data			Test Data		
λ	**α**	**Word**	**Correct**	**A (%)**	**Word**	**Correct**	**A (%)**
0	0.7	105	82	78.095	103	64	62.136
0.01	0.7	105	81	77.143	103	67	65.049
0.1	0.7	105	81	77.143	103	67	65.049
0.2	0.7	105	81	77.143	103	68	66.019
0.3	0.7	105	82	78.095	103	68	66.019
0.4	0.7	105	82	78.095	103	68	66.019
0.5	0.7	105	83	79.048	103	68	66.019
0.6	0.7	105	83	79.048	103	68	66.019
1	0.7	105	82	78.095	103	70	67.961
2	0.7	105	84	80.000	103	87	84.466
3	0.7	**105**	**87**	**82.857**	103	88	85.437
4	0.7	105	82	78.095	**103**	**89**	**86.408**
5	0.7	105	80	76.191	103	87	84.466
6	0.7	105	81	77.143	103	87	84.466
7	0.7	105	82	78.095	103	87	84.466
8	0.7	105	80	76.191	103	87	84.466

The experimental results of bigram models with different α values are shown in Table 6. Comparing the first system (where α = 0) with the others, we can see that back-off method does improve the performance. However, the α value does not affect the performance much.

Parameters in the best system are λ = 3 and α = 0.7, which achieves an accuracy of 82.86% on the training set and 85.44% on the test set.

Table 6. Performance of Prefix Analysis by Bigram Models with Different α

		Training Data			Test Data		
λ	**α**	**Word**	**Correct**	**A (%)**	**Word**	**Correct**	**A (%)**
3	0	105	80	76.191	103	83	80.583
3	0.1	105	86	81.905	103	87	84.466
3	0.4	105	86	81.905	**103**	**88**	**85.437**
3	0.7	**105**	**87**	**82.857**	**103**	**88**	**85.437**
3	1	105	86	81.905	**103**	**88**	**85.437**

Machine learning and deep learning methods described in Section 3.4 are also tested in this paper. Many well-known classifiers including Naïve Bayes, SVM, and decision tree are tried, and an encoder-decoder system by LSTM is also constructed. But unfortunately, the best accuracy is only 52.06%. The training set is too small for machine learning and deep learning at this stage.

4.4 Final Remarks

In general, our infix and suffix detection system can successfully predict structures for nearly 90% of words. It will greatly reduce the human effort needed to construct a larger dataset from the CIP Seediq dictionary. In our preliminary observation, only 335 of the 5,600 words in the CIP Seediq dictionary cannot be predicted.

Error analysis indicates that some words are inflected in an exceptional way. For example, the word "*kesa-un*" is "*kesun*" (do this way, 這樣做), but our system incorrectly predicts it as "*ksaun*"; and the word "*p-uqi-un*" is "*puqun*" (eat, 吃), but our system incorrectly predicts as "*puqiun*" or "*puqiyun*". A list of exceptional words should be constructed in the future.

As for our prefix analysis system, it can successfully analyze structures for around 83% of prefixed words. Again, it will greatly reduce the human effort in the future. However, it is not easy to improve the performance of the prefix structure analysis system. To solve the ambiguity problem (such as "*kn-*" vs. "*k<n>-*"), we might need the semantic information of the prefixed word or even the information about its functionality in that sentence. This will also be explored in the near future.

5. Conclusions

This paper proposes approaches to analyze morphological structures of Seediq words automatically. The experimental datasets contain 716 affixed Seediq words with their morphological structures.

Morphological analysis starts from the infix and suffix detection. Deep root candidates generated by our proposed methods are combined with known infixes and suffixes. The writing form of the combination is then generated by the transformation rules. If the writing form matches the trailing substring of the target word, this combination is selected as the result of infix and suffix detection. This approach achieves a precision of 98.88% and a recall of 89.59%.

Prefix structure analysis is treated similar to the word segmentation problem and predicted by probabilistic models. Zero probability problem in the bigram model is solved by the back-off approach, i.e. using the unigram probability weighted by α instead. Zero probability problem in the unigram model is solved by the Lidstone Smoothing. i.e. adding λ to frequencies of unigrams. We conclude that the best system is based on bigram model where $\lambda = 3$ and $\alpha = 0.7$, with an accuracy of 82.86%.

In the future, we would like to apply the techniques developed in this paper to analyze the 5,595 words in the CIP Seediq dictionary to create a larger dataset and build a more reliable probabilistic model. Moreover, if the morphological structures of all words appearing in the 6,019 exemplar sentences in the CIP Seediq dictionary are available, it will be possible to build a large bilingual corpus for machine translation then.

Acknowledgement

This research was funded by the Ministry of Science and Technology in Taiwan (Grant: MOST 109-2221-E-019 -053 -).

References

Bahdanau, D., Cho, K.H., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Chiang, C.-Y. (2018). Cross-Dialect Adaptation Framework for Constructing Prosodic Models for Chinese Dialect Text-to-Speech Systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 108-121. doi: 10.1109/TASLP.2017.2762432

Gu, H.-Y., Zhou, Y.-Z., & Liau, H.-L. (2007). A System Framework for Integrated Synthesis of Mandarin, Min-Nan, and Hakka Speech. *International Journal of Computational Linguistics & Chinese Language Processing*, *12*(4), 371-390.

Hsieh, S.-K., Su, I-L., Hsiao, P.-Y., Huang, C.-R., Kuo, T.-Y., & Prévot, L. (2007). Basic Lexicon and Shared Ontology for Multilingual Resources: A SUMO+MILO Hybrid Approach. In *Proceedings of OntoLex Workshop in the 6th International Semantic Web Conference*.

Iunn, U.-G., Lau, K.-G., Tan-Tenn, H.-G., Lee, S.-A., & Kao, C.-Y. (2007). Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods. *International Journal of Computational Linguistics & Chinese Language Processing*, *12*(4), 349-370.

Li, P. J.-K. (1977). Morphophonemic Alternations in Formosan Languages. *Bulletin of the Institute of History and Philology (中央研究院歷史語言研究所集刊)*, *48*(3), 375-413. doi: 10.6355/BIHPAS.197709.0375

Li, P. J.-K. (1981). Reconstruction of Proto-Atayalic Phonology. *Bulletin of the Institute of History and Philology (中央研究院歷史語言研究所集刊)*, *52*(2), 235-301. doi: 10.6355/BIHPAS.198106.0235

Li, P. J.-K. (1991). *Vowel Deletion and Vowel Assimilation in Sediq*. In Papers on Austronesian languages and ethnolinguistics in honour of George W. Grace, Pacific Linguistics C-117, 163-169.

Lin, C.-J. & Chen., H.-H. (1999). A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan. *International Journal of Computational Linguistics & Chinese Language Processing*, *4*(1), 59-84.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1412-1421. doi: 10.18653/v1/D15-1166

Rau, D. V., Wu, Y.-H., & Yang., M.-C. (2015). A Corpus-Based Approach to the Classification of Yami Emotion. *New Advances in Formosan Linguistics, Asia-Pacific Linguistics*, 533-554.

宋麗梅(2011)。「**原住民族語言字詞典編纂四年計畫—第 3 階段計畫**」(**賽德克語**)。新北市:原住民族委員會。[Sung, L.-M. (2011). *Revitalization of Formosan Languages: Compilation of Seediq Dictionary*. New Taipei City, Taiwan: Council of Indigenous Peoples.] 2009/8/3-2011/8/2.

宋麗梅(2018)。**臺灣南島語言叢書 5: 賽德克語語法概論**(2 版)。新北市:原住民族委員會。[Sung, L.-M. (2018). *A Sketch Grammar of Seediq, Formosan Series #5, 2018* (2nd Edition). New Taipei City, Taiwan: Council of Indigenous Peoples.]

楊秀芳(1976)。賽德語霧社方言的音韻結構。**中央研究院歷史語言研究所集刊**，**47**(4)，611-706。[Yang, H.-F. (1976). The Phonological Structure of the Paran Dialect of Sedeq. *Bulletin of the Institute of History and Philology*, *47*(4), 611-706.]

Yang, M.-C. & Rau, D. V. (2011). Constructing a Yami Language Lexicon Database from Yami Archives. In *Proceeding of the 2011 TELDAP (Taiwan e-Learning and Digital Archives Program) International Conference*.

Yang, M.-C., Rau, D. V., & Chang, A. H.-H. (2011). A Proposed Model for Constructing a Yami Wordnet. *International Journal of Asian Language Processing*, *21*(1), 1-14.

楊孟蒨、何德華(2015)。**建構台灣原住民語自然語言處理技術探討與研究**。科技部計畫期末報告(編號: MOST 103-2221-E-126-008-) [Yang, M.-C. & Rau, D.V. (2015). *Exploring the NLP Techniques for Formosa Indigenous Languages*. (MOST 103-2221-E-126-008-), 2014/8~2015/7.

Yu, M.-S. & Lin, Y.-J. (2012). The Polysemy Problem, an Important Issue in a Chinese to Taiwanese TTS System. *International Journal of Computational Linguistics & Chinese Language Processing*, *17*(1), 43-64.

Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. 21-36

基於圖神經網路之中文健康照護命名實體辨識

Chinese Healthcare Named Entity Recognition Based on Graph Neural Networks

盧毅*、李龍豪*

Yi Lu and Lung-Hao Lee

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 Networks, 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, 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。

*國立中央大學電機工程研究所
Department of Electrical Engineering, National Central University
E-mail: ericst91159@gmail.com; lhlee@ee.ncu.edu.tw
The author for correspondence is Lung-Hao Lee

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, Lattice, Gazetteers, and ME-CNER. In summary, our GGSNN model is an effective and efficient solution for the Chinese healthcare NER task.

關鍵詞：命名實體辨識、圖神經網路、資訊擷取、健康資訊學
Keywords: Named Entity Recognition, Graph Neural Networks, Information Extraction, Health Informatics

1. 緒論 (Introduction)

命名實體辨識 (Named Entity Recognition, NER) 主要目的為從非結構化的文本中，抽取出所關注的命名實體，主要包括人名、地名、組織名、時間、數量、貨幣、專有名詞等。舉例來說對於「比爾蓋茲創辦了微軟」這個中文句子，如果關注的命名實體包含人名以及組織名，透過 NER 即可抽取出人名「比爾蓋茲」以及組織名「微軟」。命名實體辨識為自然語言處理中的一項基礎任務，其後續的應用包含了關係抽取、事件抽取、知識圖譜以及問答系統等等，像是抽取出人名「比爾蓋茲」以及組織名「微軟」後，我們可以進一步擷取兩者之間關係為「創辦」。

早期的 NER 方法主要是基於字典或是規則，利用字串比對來做辨識，此種方法非常

依賴字典的可靠度以及專業人士所制定出的規則，因此需要耗費大量的人力資源。理論上，並不能夠蒐集到一個涵蓋所有命名實體的字典，或者制定可能的規則找到所有命名實體位置。因此，若是所依賴的字典品質不佳或是規則無法涵蓋所有的情況時，則命名實體辨識的表現會嚴重的下降。

而後，機器學習的方法透過大規模標註語料學習序列標註模型，對句子的各個位置進行標註，但同樣需要透過事先定義好的特徵，因此特徵選取的好壞，對於整個標註的結果有直接的影響，主要的模型有：隱藏式馬可夫模型(Hidden Markov Model, HMM) (Rabiner, 1989)、最大化熵馬可夫模型(Maximum Entropy Markov Model, MEMM) (Toutanova & Manning, 2000) 和條件隨機場域(Conditional Random Field, CRF) (Lafferty, McCallum & Pereira, 2001)。

隨著科技的進步，人類的壽命得以延長，有關健康照護的議題逐漸地浮上檯面，許多的報章雜誌都在談論相關議題，因此本研究所關注的命名實體領域選定為健康照護。有鑑於當前欠缺健康照護的中文命名實體辨識語料庫，我們從網路上蒐集了相關的文章雜誌以及問答紀錄，隨機選取 30,692 句，人工斷詞並標記時後，透過計算 Cohen's Kappa 值以及 Fleiss' Kappa 值確保標記的品質，最後總共有 68,460 個命名實體，橫跨 10 個類別，分別是人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療以及時間。

近年來深度學習技術的興起，神經網路在許多任務皆有著亮眼的表現，在命名實體辨識任務中，BiLSTM-CRF 網路架構為最被廣泛使用的主流模型(Lample, Ballesteros, Subramanian, Kawakami & Dyer, 2016; Ma & Hovy, 2016)。我們以此架構為基礎，並且考量中文的特性，斷詞的精準度會嚴重影響結果，因此以字作為輸入單位，訓練字嵌入、部首嵌入以及詞嵌入語意向量，透過門控圖序列神經網路(Gated Graph Sequence Neural Networks, GGSNN)加入字典資訊，在建置的中文健康照護命名實體辨識語料庫上，達到 F1 分數 75.69%，比當前具代表性相關研究模型(BiLSTM-CRF, Lattice, Gazetteers 以及 ME-CNER)有更好的成效。

本研究一共分為五個章節，第一章節為緒論，介紹命名實體辨識任務以及研究動機與目的。第二章節為探討相關研究，調查目前的中文命名實體辨識語料庫，並且介紹中文命名實體辨識模型。第三章節為模型架構，詳細介紹提出的圖神經網路模型，並對模型的各層做詳盡的說明。第四章節為實驗評估與分析，依序說明語料庫的建置、嵌入向量、實驗設定與效能評估指標，接著討論實驗結果和錯誤分析。第五章為結論和未來研究。

2. 相關研究 (Related Work)

2.1 中文命名實體辨識語料庫 (Chinese NER Corpora)

MSRA 命名實體語料庫(Levow, 2006)總共包含 30 種類別，語料來源為新聞文章，其中較被廣泛使用的類別像是人名(Person)、地名(Location 以及組織名(Organization)，此資料

集的訓練資料總共包含了 46,364 個句子，其中的命名實體總數為 118,643 個，測試資料為 4,365 個句子，其中標記的命名實體總數為 4,362 個。

在社群媒體方面，Weibo 命名實體語料庫(Peng & Dredze, 2015)蒐集了微博此社群媒體從 2013 年 11 月至 2014 年 12 月的訊息並對其標記，隨機挑選訊息的數量總共為 1,890 則，標記的命名實體類別總共有 4 種，分別為地理位置(Geo-political)、地名(Location)、組織名(Organization)以及人名(Person)，其中標記的命名實體總數為 1,981 個。

Resume 資料集(Zhang & Yang, 2018)的來源為個人履歷，履歷的出處為中國上市公司主管，總共隨機挑選了 1,027 份，標註的命名實體種類總共有 8 種，其中包含國家(Country)、人名(Person)以及組織名(Organization)等等，其中標記命名實體的的總數為 16,565 個。

中國知識圖譜與語義計算大會(CCKS: China Conference on Knowledge Graph and Semantic Computing)在 2019 年舉辦的評測任務中，命名實體辨識的資料來源為電子病歷(Electronic Health Record, EHR)，訓練集的文檔數為 1000 筆，而測試集的文檔數為 379 筆，所標註的命名實體包含 6 種，分別為疾病和診斷(Disease and Diagnosis)、檢查(Examination)以及檢驗(Inspection)等等，其中標記命名實體的的總數為 16,565 個。

上述的命名實體語料庫，並沒有關於健康照護領域方面的語料庫，且都為簡體中文，因此本研究建置了一個中文健康照護領域的命名實體語料庫，共有 10 類命名實體，分別為人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療以及時間。

2.2 中文命名實體辨識語模型 (Chinese NER Models)

Dong 等人(2016) 考量了中文字的特性，將中文字拆解成一個個部件，其原因為中文的字是由許多的部件組合而成，而每個部件具有其不同的意義，透過這些部件可以增加字特徵以外的特徵，從該研究可以得知「字」並非中文字具有意義的最小單位。

Xu 等人(2019)加入了除了字特徵以外的部首特徵以及詞特徵，並且將字特徵以及部首特徵利用雙向長短期記憶類神經網路(BiLSTM)以及卷積運算(Convolution)做額外的處理。在此研究中之所以加入中文部首的原因為中文部首具有語意分類，同樣部首的字，可能屬於同樣類別，因此透過部首可以對字做更進一步的分析。

Zhang 和 Yang (2018)提出了一個新的模架構 Lattice LSTM，此模型主要的特點為會將句子中詞彙透過大型自動取得的字典，將所有可能的潛在詞彙找出，利用此種方式可以將考量到可能潛在的詞邊界，此研究結果在命名實體辨識的任務中取得了重大的成果。

Ding 等人(2019)使用到了圖神經網路中的門控圖序列神經網路，並做改良使其能夠將多個字典的資訊加入模型，由於文字訊息常常會有著類似圖結構的訊息，因此透過圖神經網路能夠更充分的表達資訊。

基於上述研究，本研究提出了門控圖序列神經網路(Gated Graph Sequence Neural Networks, GGSNN)模型架構，以 BiLSTM-CRF 做為基礎，以字為單位當作模型的輸入。

除了字的資訊以外，本研究加入了部首以及詞的資訊。在加入字特徵、部首特徵時透過雙向長短期記憶類神經網路以及卷積運算做了有別於 Xu 等人(2019)的處理，使特徵資訊更能夠完整充分。在本研究的模型同樣使用了改良式的 GGSNN 將字典資訊加入，而與 Ding 等人(2019)不同的地方在於透過不同的字典編排方式，使其在相同的硬體設備下，字典的來完能夠更加的龐大且豐富。

3. 模型架構 (Model Architecture)

本研究提出的門控圖序列神經網路(GGSNN)模型架構如下圖 1，此模型使用了目前主流的 BiLSTM-CRF 作為模型的基礎架構，並對其做延伸，模型總共分為四層。

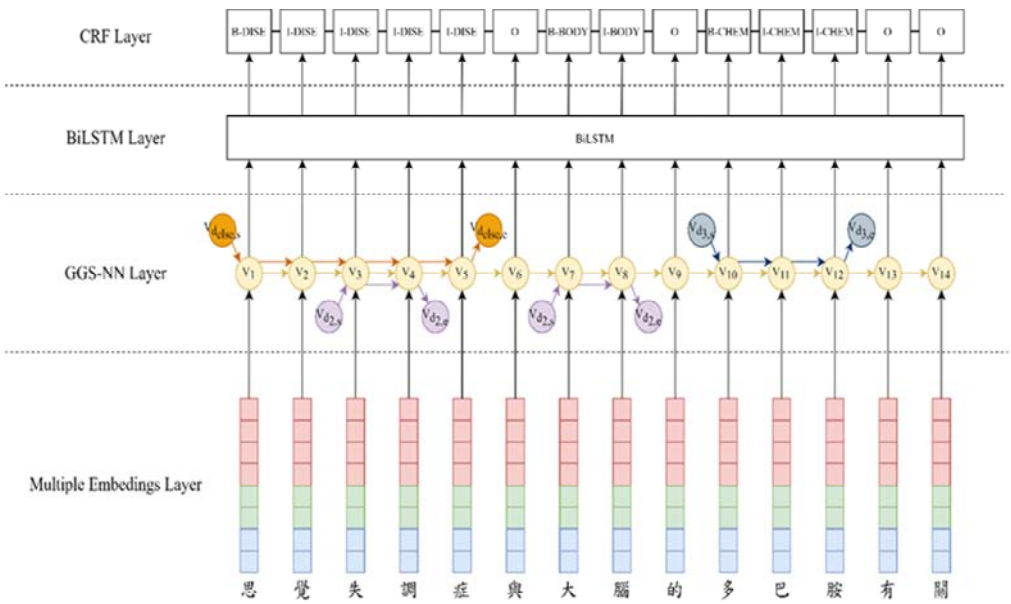


圖 1. GGSNN 模型架構
[Figure 1. GGSNN model architecture]

3.1 多重嵌入層 (Multiple Embeddings Layer)

透過組合字嵌入、詞嵌入以及部首嵌入形成多重嵌入，其中字嵌入、部首嵌入以及詞嵌入的處理分別如下列敘述，假設輸入句子字數長度為 n：

(1)、字嵌入 (Character Embedding)：

輸入字序列 $X = [x_1, x_2, x_3, \dots, x_n]$，分別經過 BiLSTM 以及卷積運算後，將兩者組成新的字嵌入特徵序列，得到序列 $C = [c_1, c_2, c_3, \dots, c_n]$，由於每個字可能與長距離的另一個字或是附近的字有所關聯，因此透過 BiLSTM 可以捕捉到長距離的資訊，而卷積運算可以捕捉到短距離的資訊。

$$[y_1, y_2, y_3, \dots, y_n] = BiLSTM(X) \quad (1)$$

$$[z_1, z_2, z_3, \dots, z_n] = Conv(X) \quad (2)$$

$$c_i = y_i \oplus z_i \quad (3)$$

(2)、部首嵌入 (Radical Embedding)：

輸入部首序列 $X = [x_1, x_2, x_3, \dots, x_n]$，經過卷積運算後，得到新的部首特徵序列 $[r_1, r_2, r_3, \dots, r_n]$，由於每個部首多半與附近的字有關，因此卷積運算可以捕捉到短距離的資訊。

$$[r_1, r_2, r_3, \dots, r_n] = Conv(X) \quad (4)$$

(3)、詞嵌入 (Word Embedding)：

由於模型是以字為基礎作為輸入，而同一個字組成的不同詞語可能有不同的意思，因此相同字的資訊，加入不同詞的資訊，可以解決此種情況，而詞的資訊是屬於較高階的特徵，因此本研究直接將其作合併，不做額外的處理。

$$W = [w_1, w_2, w_3, \dots, w_n] \quad (5)$$

最終將字特徵序列、部首特徵序列以及詞特徵序列，組合成多重嵌入如下：

$$h_i = c_i \oplus r_i \oplus w_i \quad (6)$$

其中c_i代表經過處理後的字嵌入，r_i代表經過處理後的部首嵌入，w_i代表詞嵌入，h_i代表拼接後的多重嵌入。

3.2 門控圖序列神經網路層 (GGSNN Layer)

在本研究中採用改良式 GGSNN 學習句子圖結構化後的訊息，與 Li 等人(2016) 所提出的 GGSNN 不同之處在於改良式的 GGSNN 可以給予邊上標籤不同的權重，透過改良式的 GGSNN 可以將加入多個字典的訊息，並且給予不同的字典不同的權重。但由於硬體的限制，我們無法不受限制的追加多個字典，因此與 Ding 等人(2019) 的字典編排方式不同，本研究將字典裡的詞彙依照字數做分類，總共分成五個字典。

在這層結構中首先會利用字典，透過字串比對產生多維有向圖，建構出的多維有向圖 (Multi-digraph) 範例如圖 2。給定一個多維有向圖 $G := (V, E, L)$，其中V代表節點的集合，E代表邊的集合，L代表邊上標籤的集合。假設輸入的句子為字數為 n 個，字典的使用數量為 m，節點的集合$V = V_c \cup V_s \cup V_e$。其中$V_c$為字序列節點的集合，而當字典比對到詞彙時，會產生除了字序列節的額外兩個節點，分別為$v_{d_i,s}$、$v_{d_i,e}$，其中$v_{d_i,s}$指示出詞彙的起始位置，$v_{d_i,e}$指示出詞彙的結束位置，V_s以及V_e分別代表的為$v_{d_i,s}$以及$v_{d_i,e}$的集合。邊的集合$E = \{e_c\} \cup \{e_{d_i}\}_{i=1}^{m}$，其中$\{e_c\}$為字序列節點連成的邊的集合，$\{e_{d_i}\}_{i=1}^{m}$為所有字典連成的邊的集合。每個邊都帶有標籤，邊上標籤的集合為$L = \{l_c\} \cup \{l_{d_i}\}_{i=1}^{m}$，$l_c$為字序列節點連成的邊上的標籤，$l_{d_i}$為字典連成的邊上的標籤，不同的字典帶有不同的標籤。

以「思覺失調症與大腦的多巴胺有關」當作輸入句子為例，可以得到如圖 2 的多維

有向圖，在此句子中，可以比對到的詞彙有「思覺失調症」、「失調」、「大腦」以及「多巴胺」，其中「思覺失調症」包含在詞彙字數為 5 個字以上 (else) 的字典中，因此「思覺失調症」的開頭「思」，對應到的節點v_1，連結了額外的節點$v_{d_{else},s}$，「思覺失調症」的結尾「症」，對應到的節點v_5，連結了額外的節點$v_{d_{else},e}$，$v_{d_{else},s}$的下標d_{else}以及下標 s 代表的為比對到的字典以及比對到的詞的開頭位置，$v_{d_{else},e}$的下標d_{else}以及下標 e 代表的為比對到的字典以及比對到的詞的結尾位置，其餘依此類推。

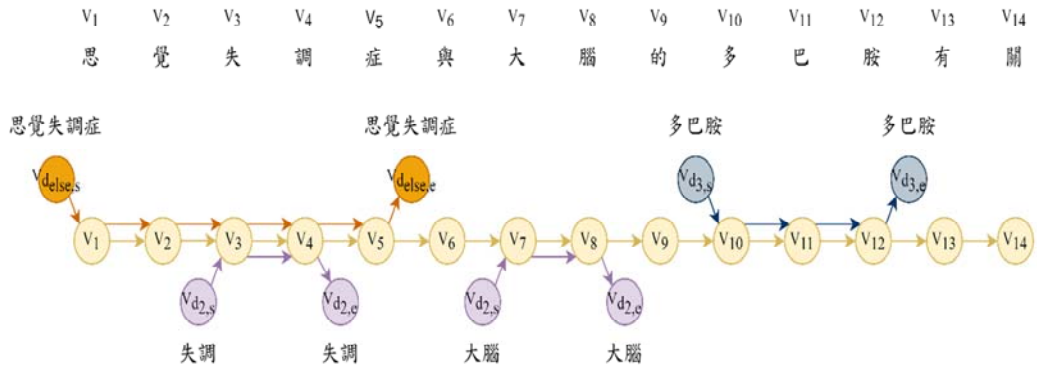


圖 2. 多維有向圖

[Figure 2. A directed multigraphs]

有向圖的結構訊息，可以透過相鄰矩陣 (adjacency matrix) 表達，假設有向圖的結構為圖 3 的左半部，而其對應的相鄰矩陣如圖 3 的右半部，其中A_{in}與A_{out}互為轉置矩陣，而相鄰矩陣由A_{in}以及A_{out}所構成。

以範例句子「思覺失調症與大腦的多巴胺有關」為例，該句子的多維有向圖的拆解成多個有向圖的範例如圖 4，由於詞彙字數為 1 個字的字典以及詞彙字數為 4 個字的字典並沒有比對到詞彙，因此對應的相鄰矩陣為零矩陣。

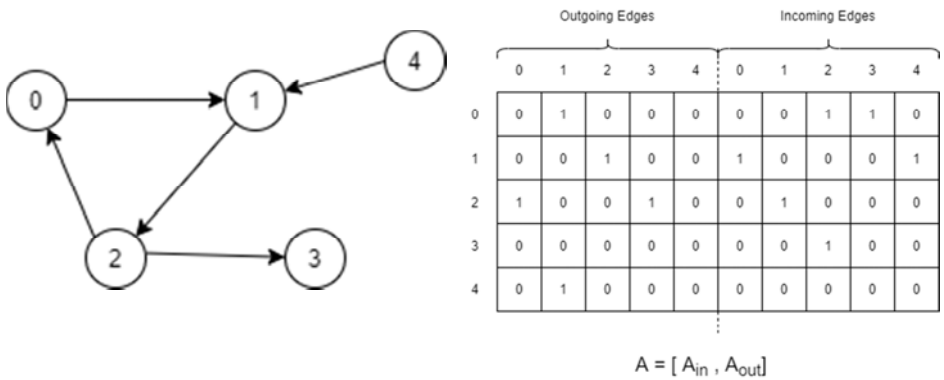


圖 3. 有向圖以及對應的相鄰矩陣

[Figure 3. A directed graph and its corresponding adjacent matrix]

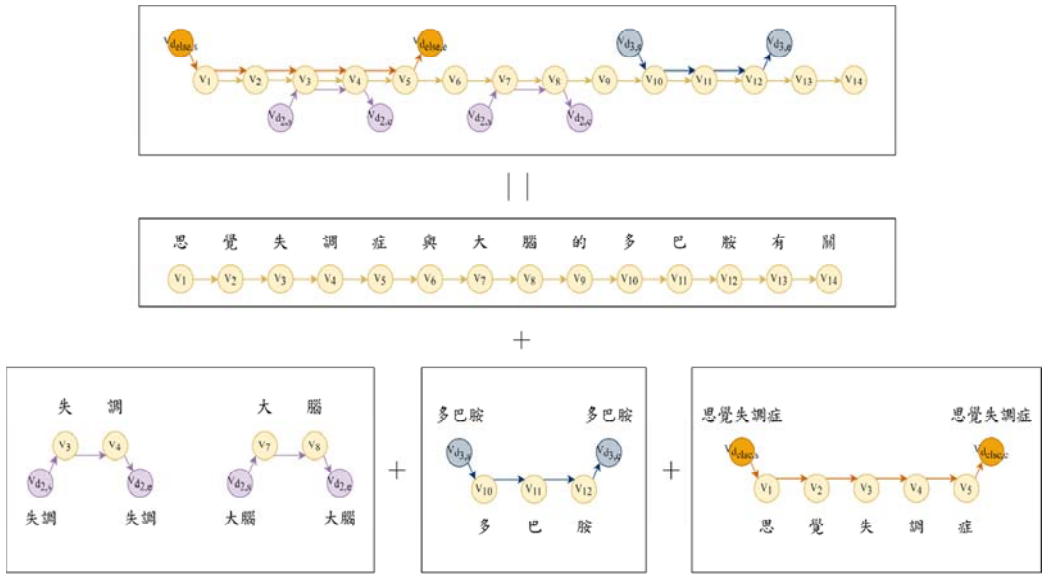


圖 4. 多維有向圖拆解成多個有向圖

[Figure 4. A directed multigraph is composed of multiple directed graphs]

由輸入句子的原始字序列訊息可以得到相鄰矩陣A_c，而由不同的字典可以得到其相對應的相鄰矩陣，依照本研究的字典分類方式可以得到相鄰矩陣A_{d_1}、A_{d_2}、A_{d_3}、A_{d_4}以及$A_{d_{else}}$，其中A_{d_1}代表的為字典詞彙字數長度為 1 的相鄰矩陣，其餘依此類推。

在本研究中，不同字典的相鄰矩陣會分別給定不同的權重，權重由以下的公式決定：

$$[w_c, w_{d_1}, w_{d_2}, w_{d_3}, w_{d_4}, w_{d_{else}}] = \sigma([\alpha_c, \alpha_{d_1}, \alpha_{d_2}, \alpha_{d_3}, \alpha_{d_4}, \alpha_{d_{else}}]) \tag{7}$$

其中$\alpha_c, \alpha_{d_1}, \alpha_{d_2}, \alpha_{d_3}, \alpha_{d_4}, \alpha_{d_{else}}$為可以被訓練的參數，並且透過 sigmod 函數使其轉換成最後的權重$w_c, w_{d_1}, w_{d_2}, w_{d_3}, w_{d_4}, w_{d_{else}}$，將不同的全權分別乘上相對應的相鄰矩陣，即可獲得最後帶有權重的相鄰矩陣。

在本研究的門控圖序列神經網路結構中，節點的初始狀態由以下公式得到：

$$h_v^{(0)} = \begin{cases} h_d(v) & v \in V_s \cup V_e \\ h_i(v) & v \in V_c \end{cases} \tag{8}$$

其中V_c代表的為多重嵌入層最後輸出的字序列特徵中，每個字分別對應到的節點，其值由多重嵌入層最後輸出的字序列特徵的值決定，V_s為命名實體的起始字對應到的節點，V_e為命名實體的最後的字對應到的節點，V_s以及V_e的值為比對到的命名實體的隨機初始狀態決定。

節點的隱藏狀態藉由 GRU 做更新，整個遞迴關係式如下：

$$H = \left[\mathrm{h}_1^{(t-1)}, \mathrm{h}_2^{(t-1)}, \dots\dots, \mathrm{h}_{|v|}^{(t-1)}\right] \quad (9)$$

$$a_v^{(t)} = [(HW_1)^T, \dots\dots, (HW_{|L|})^T]A_v^T + \mathrm{b} \quad (10)$$

$$z_v^{(t)} = \sigma(W^z a_v^{(t)} + U^z h_v^{(t-1)}) \quad (11)$$

$$r_v^{(t)} = \sigma(W^r a_v^{(t)} + U^r h_v^{(t-1)}) \quad (12)$$

$$\hat{h}_v^{(t)} = \tanh(W a_v^{(t)} + U(r_v^{(t)} \odot h_v^{(t-1)})) \quad (13)$$

$$h_v^{(t)} = (1 - z_v^{(t)}) \odot h_v^{(t-1)} + z_v^{(t)} \odot \hat{h}_v^{(t)} \quad (14)$$

其中 $h_v^{(t)}$ 表示的為節點v在時間為t時的隱藏狀態，A_v表示的為節點v對應的相鄰矩陣的行向量，公式(11)-(14)為 GRU 單元，z 與 r 分別代表更新門以及重置門，透過 GRU 單元可以結合來自相鄰節點的信息以及節點的當前隱藏狀態，計算在時間 t 時新的隱藏狀態，經過時間步數 (time step) T 後，可以得到節點的最終狀態。

3.3 雙向長短期記憶神經網路層 (BiLSTM Layer)

在這層結構中，本研究將門控圖序列神經網路層最後的隱藏狀態輸出，當作 BiLSTM 輸入序列，使用的為標準的 BiLSTM。以門控圖序列神經網路層的輸出當做輸入序列，最終可以獲得與原序列長度相同的隱藏層狀態序列。

3.4 條件隨機場域層 (CRF Layer)

命名實體辨識屬於序列標記的多分類問題，傳統上在遇到多分類問題時，會採用 softmax function 作為輸出函數，但在實際情況時，序列標註任務中的當前時刻的狀態，均與當前時刻的前後狀態有所關連，因此條件隨機場域 (Condition Random Fields, CRF) 取代了 softmax function，成為了當前主流的架構，本研究所採用的為標準的 CRF 模型。

4. 實驗評估與分析 (Performance Evaluation and Analysis)

4.1 語料庫建置 (Corpus Construction)

本研究透過爬蟲將網路上的健康照護文章及問答紀錄爬取下來，有三種來源分別為國家網路醫藥[1]、康健雜誌[2]和醫聯網[3]。其中國家網路醫藥以及康健雜誌為醫生或是相關的專業人員所撰寫的文章，而醫聯網則是一般民眾上網提問，醫生回答的問答紀錄，文章內容透過篩選主題選擇健康照護相關。本研究分別在國家網路醫藥以及康健雜誌一共爬取了 425 篇文章以及 799 篇文章，而醫療網一共有 1,818 則問答。

透過計算 Cohen's Kappa (Cohen, 1960) 值以及 Fleiss' Kappa (Fleiss, 1971) 值可以確保標記品質，其主要的功能為評估問題的一致性，其中 Cohen's Kappa 值適用於檢定兩個人意見的一致性，而 Fleiss' Kappa 值則用來檢定三人以上的情況。根據 Landis 以及

[1] 國家網路醫藥：https://www.kingnet.com.tw/knNew/index.html

[2] 康健雜誌：https://www.commonhealth.com.tw/

[3] 醫聯網：https://med-net.com/

Koch 所提出的觀點 (Landis & Koch, 1977)，當 Kappa 值小於 0 時為 Poor agreement，介於 0 到 0.20 為 Slight agreement，介於 0.21 到 0.40 為 Fair agreement，介於 0.41 – 0.60 為 Moderate agreement，介於 0.61 – 0.80 為 Substantial agreement，介於 0.81 – 1.00 為 Almost perfect agreement。

本研究所關注的健康照護命名實體，總共包含 10 類，其定義以及例子如表 1。整個標記資料的流程，我們將其分成兩個階段，參與標記的人員一共有三位師大中文系的大學生，對於每個中文句子做人工斷詞及命名實體標記，第一個階段先取國家網路醫藥 25 篇文章、康健雜 25 篇文章以及醫聯網 100 則問答，先做第一次標記，計算三位標記人員的一致性，得到 Fleiss' Kappa 值為 0.80。對階段一的標記結果做討論，修正標記準則得到一致的標準後，再對另外的 25 篇國家網路醫藥的文章、25 篇康健雜誌的文章以及醫聯網 100 則問答做標記，得到 Fleiss' Kappa 值為 0.89，達到了 Landis 以及 Koch 所認為的 Almost perfect agreement，確認階段二的 Fleiss' Kappa 有明顯上升，且達到可接受的範圍後，剩餘的待標記的中文句子，則由三位標記人員分工各自標記。

表 1. 命名實體類別定義及範例
[Table 1. Named entity definitions and examples]

類別	定義	範例
人體 (Body)	泛指生物體的細胞、組織、器官和系統。	細胞核、神經組織、心、肺、脊髓、呼吸系統等。
症狀 (Symptom)	又稱病徵，由患者描述的主觀感受，而非直接量測得知。	流鼻水、頭昏、發燒、咳嗽、失眠、貧血等。
醫療器材 (Instrument)	包含診斷、治療、減輕與預防人類疾病，使用的儀器、器械、附件、配件與零件。	血壓計、達文西機器手臂、人工髖關節等。
檢驗 (Examination)	利用醫療器材對人體健康狀態及生理功能評估。	聽力檢查、顯微鏡檢查、核磁共振造影等。
化學物質 (Chemical)	人體由不同的化學物質組成，隨著年齡與健康狀況有所增減。	去氧核糖核酸、三酸甘油酯、糖化血色素等。
疾病 (Disease)	指人體在外在因素的損害或內在機能不良情況下，影響部分或全部器官異常，伴隨特定症狀的醫學病症。	小兒麻痺症、帕金森氏症、憂鬱症、青光眼、腦溢血、肺結核等。
藥品 (Drug)	泛指用來做診斷、治療、預防疾病或減輕痛楚的藥物或化學成份。	阿斯匹靈、亞硝酸鈉、亞鐵鹽、抗生素等
營養品 (Supplement)	指從食物中萃取對人體有益的營養素，主要功能是維持健康和預防疾病。	膠原蛋白、益生菌、綜合維他命、葉黃素等。
治療 (Treatment)	讓患者恢復健康的治癒方式。	藥物治療、血漿置換、免疫球蛋白注射等。
時間 (Time)	描述患者患病症狀的持續時間或是某個時刻。	嬰兒期、幼兒時期、青春期、生理期、孕期等。

整個語料庫最後包含 30,692 句，總字數約 150 萬字，接近 92 萬個詞，68,640 個命名實體。訓練資料是三個標記人員各自標記的部分，共有 28,161 句，每個句子平均 49.44 個字 (29.99 個詞)，總共有 61,155 個命名實體，平均每個句子有 2.17 個。測試資料來自三個標記人員共同標記有一致結果的 2,531 句，每個句子平均 47.92 個字 (28.67 個詞)，總共有 7,305 個命名實體，平均每個句子有 2.89 個。10 個類別在訓練和測試資料分佈相似，最多的命名實體類別是人體，約佔 38%，依序是症狀、疾病和化學物質，前四大類佔總數的 82%，其餘 6 類約佔總數的 18%。

4.2 嵌入向量 (Embedding)

本研究所使用的嵌入方式為 Word2vec，訓練的資料來源為維基百科，下載語料庫的日期為 2020 年 2 月 3 日，利用此檔案我們可以訓練出字嵌入、部首嵌入以及詞嵌入，詞頻設定為至少出現 5 次以上，向量的維度的設定皆為 50 維，最終獲得 863,835 個詞嵌入向量，13,581 個字嵌入向量以及 3,209 個部首嵌入向量。

4.3 實驗設定 (Settings)

本研究所使用的字典來源一共分為三個，分別為國家網路醫藥[4]、國家教育研究院[5]以及搜狗網[6]，其中國家網路醫藥的詞彙主要為常見的醫護名詞，國家教育研究院選用的資料為醫學名詞，而搜狗網所包含的內容為 ICD-10、人體穴位名稱、醫學詞彙、醫療檢驗以及醫療器材等等，在使用字典時，將上述字典先合併後分類，依照詞彙字數一共分成五個字典，1 個字的字典有 351 個詞，2 個字的字典有 7,978 個詞，3 個字的字典有 19,282 個詞，4 個字的字典有 31,444 個詞，詞彙字數為 5 個字以上的字典有 95,362 個詞。

在訓練過程中學習率 (learning rate) 以及訓練資料會隨著時期 (epoch) 調整，單數 epoch 的 learning rate 為 0.001，資料為原始整份的訓練資料，雙數 epoch 的 learning rate 為 0.0005，資料為尚未學習好的訓練資料，判斷的依據為命名實體辨識是否有錯誤。其中之所以會針對尚未學習好的資料再學習一遍的原因為理論上在訓練的過程中，會希望模型能夠將所有的訓練資料學習正確，epoch 的設定值為 80， batch size 為 32，LSTM 隱藏層的維度為 200 維，GGSNN 的更新次數 (time step) 設定為 2。

4.4 效能評估 (Evaluation)

目前在命名實體辨識領域的主要評估方法為精確率 (Precision)、召回率 (Recall)、F1-score，在本研究中評估方式採精準比對 (exact match)，意即預測的結果需與正確結果完全相符才算正確。混淆矩陣矩陣範例如表 2，藉此矩陣計算精確率 (Precision) 為「正確被辨識的項目」占「總辨識項目」的比例，召回率 (Recall) 為「正確被辨識的項目」

[4] 國家網路醫藥：https://www.kingnet.com.tw/diagnose

[5] 國家教育研究院：https://terms.naer.edu.tw/

[6] 搜狗網：https://pinyin.sogou.com/dict/

占「應該被辨識的項目」的比例以及 F1-score 此為 Precision 以及 Recall 的調和平均數，計算公式如方程式 (15)-(17)。

表 2. 混淆矩陣
[Table 2. The confusion matrix]

真實值 \ 預測值	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TF)

$$\Pr ecision = \frac{\left|TP\right|}{\left|TP + FP\right|} \tag{15}$$

$$\mathrm{Re}\, call = \frac{\left|TP\right|}{\left|TP + FN\right|} \tag{16}$$

$$F1 - score = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{17}$$

4.5 實驗結果 (Results)

我們比較了以下中文命名實體模型的效能差異

(1)、BiLSTM-CRF (ICCPOL 2016)：

此模型實作了 Dong 等人(2016) 的架構，以字作為基礎當作模型輸入，字嵌入使用透過 4.2 節中所提到的維基百科語料庫當作訓練資料，向量維度為 200 維。

(2)、ME-CNER (CIKM 2019)：

Xu 等人(2019) 提出的模型，本研究實作的模型架構將其稍做更動，本研究認為將字嵌入分別經過 BiLSTM 以及 Convolutions 比起分別經過 BiLSTM-Convolution 以及 Convolutions 後連接，其中前者的 BiLSTM 較能保留原始的訊息，且比原模型效能好。

(3)、Gazetteers (ACL 2019)：

此模型為 Ding 等人所提出 (Ding *et al*., 2019)，在其發表的論文中有提供開源程式碼，因此將資料替換成本研究所使用的資料，參數的設定與原始程式碼相同，由於開源程式碼並未提供模型所會用到的的字嵌入以及二元嵌入，而在其官網的說明為使用維基百科語料庫進行訓練即可，因此使用 4.2 節中所提到的維基百科當作訓練資料，訓練出各 200 維的向量。

(4)、Lattice (ACL 2018)：

此模型為 Zhang and Yang 等人所提出 (Zhang & Yang, 2018)，利用其論文中提到的開源程式碼，將資料替換成本研究所使用的資料，模型設定的參照原始程式碼，而模型會使

用到的字嵌入以及詞嵌入由開源程式碼所提供。

(5)、GGSNN：

此為本研究提出的模型，在第三章有詳細的介紹。其中 - radical 為此模型 GGSNN 去除部首嵌入。- word 為此模型 GGSNN 去除詞嵌入。 - radical - word 則為此模型 GGSNN 同時去除部首嵌入以及詞嵌入。

表 3 為模型效能比較結果。ME-CNER 與 BiLSTM-CRF 兩者的差異為是否有加入部首嵌入以及詞嵌入，從實驗結果得知 ME-CNER 相較於 BiLSTM-CRF 提升了 2.59 的 F1，因此加入部首嵌入以及詞嵌入有助於提升模型的表現。Gazetteers 與 BiLSTM-CRF 兩者主要的差異為是否加入字典的資訊，從實驗結果得知 Gazetteers 相較於 BiLSTM-CRF 提升了 2.7 的 F1，因此透過 GGSNN 將字典的資訊納入考慮，可以有效的提升模型的表現。本研究提出的 GGSNN 模型，同時加入了部首嵌入、詞嵌入以及圖序列神經網路，其表現與 ME-CNER 以及 Gazetteers 比較，分別上升了 1.54 以及 1.43 的 F1。

由 GGSNN 分別去除掉部首嵌入、詞嵌入以及同時去除兩者的實驗比較中，可以更加地確認部首嵌入以及詞嵌入對於模型的表現影響，去除部首嵌入模型的 F1-score 下降了 0.61，去除詞嵌入模型的 F1-score 下降了 1.41，同時去除兩者模型的 F1-score 下降了 1.69，因此我們可以得知詞嵌入對於提升模型的表現的貢獻較大，而不論是詞嵌入或是部首嵌入，皆對模型的表現有幫助。

本研究提出的 GGSNN 模型有最佳的 F1 分數，而 Lattice 次之，兩個模型的差異為略小只有 0.47，然而在訓練的時間方面，相同的硬體設備下，本研究的模型約為 1 天，而 Lattice 約耗時 6.25 天，主要的原因為 Lattice 模型的 batch size 因為模型的特性只能夠設定為 1，當資料量越大時，需要更多時間，無法藉由調整 batch size 加速運算。

表 3. 模型結果比較
[Table 3. Model performance comparisons]

Method	Precision	Recall	F1
BiLSTM-CRF (ICCPOL 2016)	70.38	72.77	71.56
ME-CNER (CIKM 2019)	73.68	74.62	74.15
Gazetteers (ACL 2019)	73.00	75.56	74.26
Lattice (ACL 2018)	74.69	75.76	75.22
GGSNN (ours)	**75.46**	75.76	**75.69**
- radical	73.50	**76.73**	75.08
- word	73.48	75.10	74.28
- radical - word	73.46	74.54	74.00

4.6 錯誤分析 (Error Analysis)

本研究將命名實體的錯誤分成以下 5 種類型，錯誤範例如表 4。

- CONTAIN：正確的命名實體「包含」預測的命名實體。
- CONTAINED：正確的命名實體「被包含於」預測的命名實體。
- SPLIT：正確的命名實體或是預測的命名實體被拆成兩段命名實體。
- CROSS：正確的命名實與預測的命名實體之間「有」重疊的字。
- NO-CROSS：正確的命名實體與預測的命名實體之間「沒有」重疊的字。

5 種類型的錯誤總共有 2,193 個，其中最多的錯誤類型為 NO-CROSS，約佔 72%。我們觀察後得知，有些領域詞彙例如：血清胺基丙酮酸轉化酶 (SGPT)、攝護腺肥大症候群 (BPH) 和胞漿精子注射 (ICSI) 等，沒有在訓練資料中，也不屬於字典中的詞彙，無法被正確辨識。藉由錯誤分析得知，字典詞彙涵蓋程度對模型效能有重要的影響。

表 4. 命名實體預測錯誤類型與範例
[Table 4. NER error types and corresponding examples]

CONTAIN	答案	國際間 德國麻疹$_{DISE}$ 仍有疫情發生，所以有出國計畫要預先做好安排。
	預測	國際間 德國麻疹$_{DISE}$ 仍有疫情發生，所以有出國計畫要預先做好安排。
CONTAINED	答案	肺主脈 指橫膈膜$_{BODY}$ 銜接心臟的部分。
	預測	肺主脈 指橫膈膜$_{BODY}$ 銜接心臟的部分。
SPLIT	答案	喉嚨痛$_{SYMP}$ 主要是我們的扁桃腺發炎。
	預測	喉嚨$_{BODY}$痛$_{SYMP}$ 主要是我們的扁桃腺發炎。
CROSS	答案	對於 痰濁$_{SYMP}$ 瘀阻經絡$_{SYMP}$ 而致的症狀有改善的功能。
	預測	對於 痰濁瘀$_{SYMP}$阻經絡$_{BODY}$ 而致的症狀有改善的功能。
NO-CROSS	答案	鉀離子量若攝取充足，可降低腦血管 阻塞$_{SYMP}$ 風險。
	預測	鉀離子量若攝取充足，可降低腦血管 阻塞 風險。

5. 結論與未來研究 (Conclusions and Future Work)

我們提出門控圖序列神經網路模型，用於中文健康照護命名實體辨識。主要貢獻如下：

一、我們提出一個多重嵌入導向的圖序列網路架構，從部首、字到詞的不同語意資訊被探索，透過圖神經網路調適至健康照護命名實體辨識任務。我們的模型達到 75.69 的 F1 顯著優於先前的命名實體辨識方法。

二、據我們所知，這是第一個健康照護領域的中文命名實體語料庫，包含 30,692 個句子，約莫 150 萬字 (92 萬詞)，共有 68,460 個命名實體，橫跨 10 個類別，包含：人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療以及時間。

利用命名實體辨識的這項技術，我們可以依照各領域不同的需求，從非結構的文章中抽取出該領域所關注的命名實體，透過這些抽取出的命名實體，我們可以充分的掌握文章中的資訊，對文章做更進一步的分析，在未來的應用中，命名實體辨識所標示出的命名實體，可以做為關係擷取、事件偵測與追蹤、知識圖譜建置、智慧問答系統等應用的基礎。

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. We sincerely thank all the annotators for their efforts in the named entity tagging task. We also thank the anonymous reviewers for their insightful comments.

參考文獻 (References)

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: 10.1177/001316446002000104

Ding, R., Xie, P., Zhang, X., Lu, W., Li, L., & Si, L. (2019). A neural multi-diagraph model for Chinese NER with gazetteers. In *Proc. of the ACL'19*, 1462-1467. doi: 10.18653/v1/P19-1141

Dong, C., Zhang, J., Zong, C., Hattori, M., & Di, H. (2016). Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Proc. of the ICCPOL'16*, 239-250. doi: 10.1007/978-3-319-50496-4_20

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. doi: DOI: 10.1037/h0031619

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of the ICML'01*, 282-289.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proc. of the NAACL-HLT'16*, 260-270. doi: 10.18653/v1/N16-1030

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi: 10.2307/2529310

Levow, G. A. (2006). The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In *Proc. of the SIGHAN'06*, 108-117.

Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2016). Gated graph sequence neural networks. In *Proc. of the ICLR'16*. Retrieved from arXiv:1511.05493.

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proc. of the ACL'16*, 1064-1074. doi: 10.18653/v1/P16-1101

Peng, N., & Dredze, M. (2015). Named entity recognition for Chinese social media with jointly trained embeddings. In *Proc. of EMNLP'15*, 548-554. doi: 10.18653/v1/D15-1064

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, *77*(2), 257-286. doi: 10.1109/5.18626

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of the EMNLP/VLC'00*, 63-70. doi: 10.3115/1117794.1117802

Xu, C., Wang, F., Han, J., & Li, C. (2019). Exploiting multiple embeddings for Chinese named entity recognition. In *Proc. of the CIKM'19*, 2269-2272. doi: 10.1145/3357384.3358117

Zhang, Y., & Yang, J. (2018). Chinese NER using lattice LSTM. In *Proc. of the ACL'18*, 1554-564. doi: 10.18653/v1/P18-1144

Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. 37-54

改善詞彙對齊以擷取片語翻譯之方法

Improving Word Alignment for Extraction Phrasal Translation

陳怡君*、楊謦瑜+、張俊盛*

Yi-Jyun Chen, Ching-Yu Helen Yang and Jason S. Chang

摘要

本研究專注於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，其結果可用於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，再根據中文搭配詞，將名詞及介系詞的翻譯做適當調整，並產生例句。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。

Abstract

This thesis presents a method for extracting translations of noun-preposition collocations from bilingual parallel corpora. The results provide researchers a reference tool for generating grammar rules. In this paper, 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. The evaluation is done using randomly 30 selected phrases. We used human judge to assess the translations.

*國立清華大學資工系
Department of Computer Science, National Tsing Hua University
E-mail: {yijyun; chingyu; jason}@nlplab.cc
+國立中興大學外語系
Department of Foreign Languages and Literatures, National Chung Hsing University

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

關鍵詞：雙語詞彙對齊、文法規則、搭配詞、片語翻譯

1. 簡介 (Introduction)

介系詞在英文使用頻率很高，因爲中英文性質不同，翻譯的變化非常多元，所以當我們將英文實詞與介系詞的搭配翻譯至中文時，會有很複雜的情況，以下列句子為例：

(1) a. In her **speech on** the motion of thanks , the hon margaret ng touched upon ...

b.吳靄儀 議員 **就** 致謝 議案 **發言** 時 ， 曾 談及 …

(2) a. ... the extent of business via ec was still relatively limited , so was its **impact on** the statistical systems .

b. …涉及 電子 貿易 的 業務 範圍 仍 相對 有限 ， **對** 統計 系統 的 **衝擊** 也 有限 。

(3) a. **impact on** the financial market

b.**衝擊** 金融 市場

(4) a. Mr downer believed the close **relationship between** hong kong and australia would continue to strengthen .

b.唐納 相信 ， 澳洲 與 香港 **之間** 的 密切 **關係** 是 會 繼續 加強 的 。

(5) a. up to now , the change in **relationship between** china and hong kong can be divided into three stages .

b.直至 目前 ， 中港 **關係** 的 變化 ， 可 分為 三 個 階段 。

相同的介系詞在搭配不同名詞時，可能會有不同的翻譯，例如在例句(1a)中的「speech on」，翻譯至例句(1b)中的「就 … 發言」，此時「on」翻譯為「就」，而在例句(2a)中的「impact on」，翻譯至例句(2b)中的「對 … 衝擊」，此時「on」翻譯為「對」。有時介系詞也可能在翻譯時被省略，例如例句(4a)中的「relationship between」翻譯為例句(4b)中的「之間 … 關係」，但例句(5a)中的「relationship between」，在例句(5b)中則是僅翻譯成「關係」。另外，在某些情況下，介系詞也可能和名詞一起翻譯為一個動詞，例如例句(3a)中的「impact on」翻譯為例句(3b)中的動詞「衝擊」。這些不同翻譯變化對於語言學習者或是機器翻譯演算法來說，是一個難解的議題。

本研究的目標，是由英文名詞和介系詞搭配（例如「speech on」），產生其中文翻譯（例如「就...發言」）。現有的自動對齊演算法雖可自動產生雙語平行語料的詞彙對應(word alignment)，但部分對應不準確，若直接以自動對齊的結果，來提取英文搭配的中文翻譯，將會產生許多錯誤的結果，我們認為若同時考量英文對中文的對應與反向的對應，對應的準確度將會更高，另外，在自動對齊的結果中，有時英文搭配所對應到的中文翻譯，若演算法能同時考量其是否符合中文搭配的性質，就能做出更正確的英文搭配翻譯。因此在本研究中，我們除了使用雙語自動對齊的平行語料庫，也使用中文單語語料庫，在平行語料庫中，我們不僅統計英文對中文的對應，也統計中文對英文的對應，同時，我們也在中文單語語料庫中，統計並篩選中文的高頻搭配，藉由結合中英對應的結果與中文搭配的性質，取得更精確的翻譯。

我們的訓練階段主要有四個步驟，依序是改善文本斷詞與對齊、產生實詞翻譯、產生介系詞翻譯、產生實詞介系詞搭配的翻譯。在改善文本斷詞與對齊這個步驟，我們統計雙語詞彙對應機率並計算各種不同斷詞及對應方式的機率，修正一些斷詞及對齊的錯誤；在產生實詞翻譯這個步驟，我們統計英文實詞所對應到的中文詞彙及次數，以及這些中文詞彙的反向對應，篩選中文詞彙做為實詞的翻譯；在產生介系詞翻譯這個步驟，我們計算當介系詞出現在實詞後面時對應至各中文詞性及詞彙的機率，並從中篩選中文詞彙做為介系詞的翻譯；在產生實詞介系詞搭配的翻譯這個步驟，我們由中文語料庫統計中文高頻搭配詞，用中文高頻搭配詞組合實詞的翻譯和介系詞的翻譯，產生比較符合搭配性質的翻譯，並再回到雙語語料庫中檢驗所產生的翻譯及篩選例句。

本研究所產生的翻譯，可用於協助語言學習者學習中文或英文，或為語言研究者提供撰寫文法規則的參考，也可輔助改進機器翻譯系統。

2. 相關研究 (Related Work)

機器翻譯一直是自然語言處理領域中的活躍研究領域，過去幾十年中，大量的雙語語料庫資源，使得統計式機器翻譯越來越可行，在 1990 年代，雙語句子對齊技術快速發展(Gale & Church, 1991a, 1991b, 1993; Brown, Lai & Mercer, 1991; Simard, Foster & Isabelle, 1992; Chen, 1993)。

除了找出相對應的雙語句子(Debili & Sammouda, 1992; Kay & Roscheisen, 1993)，有些研究使用統計模型以改善自動對齊所產生的詞彙對應，如隱藏式馬可夫模型(Hidden Markov Model, HMM) (Brown, Lai & Mercer, 1991) 、對數似然比(log-likelihood ratio) (Gale & Church, 1991a, 1993)，以及 K-Vec algorithm (Fung & Church, 1994)。奠基在前人基礎之上，Melamed (1999) 提出 the Smooth Injective Map Recognizer (SIMR)，將雙語片語對齊視為 x 軸與 y 軸在二維空間的最佳分佈，與前人研究不同之處在於，SIMR 採貪婪演算法，以最小可能運算出最佳分佈的二維空間為演算單位。SIMR 分為兩個階段，生成階段產生所有可能雙語對應（x 軸與 y 軸對應），以及辨識階段已選出最佳對應。然而，句子對齊技術仍有其限制，難免有許多錯誤的對應。

本研究是專注於使用統計方法擷取名詞介系詞搭配的翻譯，目的是獲得精確的翻譯。過去與本研究較相關的研究，主要集中於學習單詞翻譯並從現有語料庫中提取雙語單詞翻譯對(Catizone, Russell & Warwick, 1989; Brown *et al.*, 1990; Gale & Church, 1991a; Wu & Xia, 1994; Fung, 1995; Melamed, 1995; Moore, 2001)，計算平行句子中的單詞對之間的相互關係程度，從而推導出翻譯。我們提出了一種結合了雙語對齊統計和中文搭配詞統計的方法。

在 2006 年柯明憲(柯明憲，2006)提出的「雙語語料庫之多字詞語對應」論文，與本研究有類似目標，皆為擷取片語翻譯。該論文是針對特定動詞片段（例如：make a report to police），使用自動對齊技術計算雙語之間的搭配關係（例如：當「make」和「report」一起出現時，「report」常對應至「報案」），再於雙語句子中擷取對應（例如：make a report to police 的對應「向警察報案」），並統計翻譯。

本研究與上述論文同樣使用到自動對齊技術，與上述論文不同的是，本研究是針對名詞和介系詞的搭配，並且在計算完詞彙機率後，並非直接由句子擷取對應，而是計算出中文高頻搭配詞組合名詞與介系詞的翻譯後，再回到雙語語料庫中以句子篩選最合適的翻譯。本研究提出的新方法，成功達到較高的翻譯正確率，大量篩選掉錯誤的翻譯，減少查詢翻譯資訊的困難。

3. 方法 (Method)

在本章中，我們會說明如何進行資料前處理，以改善其斷詞和雙語詞彙對應（第（一）小節），接著分別詳細描述如何從已自動對齊的雙語平行語料庫中統計並篩選實詞和介系詞兩者個別的翻譯（第（二）、（三）小節），最後說明如何統計中文搭配詞以及使用中文搭配詞的統計結果來計算兩者精確翻譯（第（四）小節）。

3.1 改善斷詞及雙語詞對應 (Improving Word Segmentation and Word Alignment)

本階段的目標是改善資料的斷詞和詞彙對應，提升詞彙對應準確率。本階段的輸入為已標註中文斷詞及雙語自動對齊的雙語平行資料。此標註斷詞和詞彙對應的資料仍有不少錯誤，表 1 為句子範例，其中「中英詞對齊」代表中英詞彙的對應位置（由 0 起算），例如「4-0」代表中文句的第 4 個詞彙「反應熱」對應至英文句的第 0 個詞彙「enthusiastic」，在此句中，「反應熱烈」斷詞為「反應熱｜烈」，然而正確斷詞應為「反應｜熱烈」，也因為斷詞錯誤，導致詞彙對應的錯誤：

表 1. 英中詞對齊錯誤範例
[Table 1. An example of wrong word alignments of English and Chinese sentence]

中文例句	社區 投資 共享 基金 <u>反應熱 烈</u>
英文例句	**<u>enthusiastic response</u>** to community investment and inclusion fund
中英詞對齊	**<u>4-0 5-0 5-1</u>** 0-3 1-4 2-5 2-6 3-7

我們從資料中統計每個中文詞彙對應至每個英文詞彙的機率 P_{CtoE}，以及每個英文詞彙對應至每個中文詞彙的機率 P_{EtoC}，接著針對每一組中英詞彙的二對二對應，產生各種可能的斷詞及對應方式，並計算機率，取出機率最高者。以表 1 的句子為例，中英詞對齊中的「4-0 5-0 5-1」表示「反應熱」對應至「enthusiastic」、「烈」對應至「enthusiastic」和「response」，故「反應熱｜烈」共對應至「enthusiastic」和「 response」兩個英文詞彙，為一組二對二對應，表 2 即為產生之各種可能的斷詞及對應方式，經過計算機率後，取出機率最高的斷詞及對應方式，即斷為「反應｜熱烈」，且反應對應至「response」、「熱烈」對應至「enthusiastic」。

本階段的輸出為改善之後的資料，我們以上述方法將整份資料中的這些詞彙進行修正，有效減少部分詞彙的錯誤對應。

表 2. 斷詞及對應方式機率計算範例
[Table 2. An example of probability calculation for word segmentation and word alignment]

斷詞方式	對應方式	機率計算
反｜應熱烈	反 → enthusiastic 應熱烈 → response	P_{CtoE}(反,enthusiastic) × P_{CtoE}(應熱烈,response) × P_{EtoC}(enthusiastic,反) × P_{EtoC}(response,應熱烈)
	反 → response 應熱烈 → enthusiastic	P_{CtoE}(應熱烈,enthusiastic) × P_{CtoE}(反,response) × P_{EtoC}(enthusiastic,應熱烈) × P_{EtoC}(response,反)
反應｜熱烈	反應 → enthusiastic 熱烈 → response	P_{CtoE}(反應,enthusiastic) × P_{CtoE}(熱烈,response) × P_{EtoC}(enthusiastic,反應) × P_{EtoC}(response,熱烈)
	反應 → response **熱烈 → enthusiastic**	P_{CtoE}(熱烈,enthusiastic) × P_{CtoE}(反應,response) × P_{EtoC}(enthusiastic,熱烈) × P_{EtoC}(response,反應)
反應熱｜烈	反應熱 → enthusiastic 烈 → response	P_{CtoE}(反應熱,enthusiastic) × P_{CtoE}(烈,response) × P_{EtoC}(enthusiastic,反應熱) × P_{EtoC}(response,烈)
	反應熱 → response 烈 → enthusiastic	P_{CtoE}(烈,enthusiastic) × P_{CtoE}(反應熱,response) × P_{EtoC}(enthusiastic,烈) × P_{EtoC}(response,反應熱)

3.2 篩選實詞翻譯 (Extracting Translations of Content Words)

本階段的目標是產生實詞的翻譯。本階段的輸入為我們要查找的英文實詞（例如：「speech on」中的「speech」）和經前一階段（第三章第（一）節）改善後的資料。由於某些英文搭配經常被包含在更長的片語中（例如「connection with」經常被包含在「in connection with」中，且兩者中的實詞翻譯不同），因此在開始統計英文實詞所對應的中文詞彙之前，我們會先針對輸入的英文搭配，統計出現在其前面的詞是否高機率集中在某些詞，藉此排除英文搭配被包含在更長片語中的狀況。接著，我們統計資料中英文實詞對應至

的中文詞彙及次數，並從中選出次數較高者。由於翻譯後詞性經常出現變化，因此在統計時我們不限制中文詞彙的詞性。以「speech」為例，我們選出以下這些詞彙：

演辭	中	發言	致辭	全文	時
言	講話	發表	預算案	施政	報告
辭	演	演講	講	篇	演說

此步驟所選出的中文詞彙中，仍有許多並非原英文實詞正確的翻譯（如上例中的「預算案」、「施政」等），因此我們會再針對這些中文詞彙統計反向對應，也就是其對應至的英文詞彙及次數，並從中選出次數較高者，做為計算中文詞彙分數之用。以上例中的「演辭」為例，我們會選出以下這些詞彙：

speech	by	speeches
my	his	's

正確翻譯的反向對應時常也會對應到一些和原英文實詞的衍生詞，如複數、動詞變化形等，像是上例的「演辭」除了對應至原來的實詞 「speech」之外，也對應到「speech」的複數「speeches」。我們希望在以反向對應結果計算中文詞彙分數時，將這些情況也考慮進去，因此我們建立並結合了複數表、動詞時態表、動詞名詞變化型態表以及相似詞表，如表 3：

表 3.「discussion」衍生詞表
[Table 3. Derivatives of "discussion"]

與原名詞關係	詞彙
複數	discussions
動詞變化	discuss, discussed, discusses, discussed, discussing
近義詞	conference, argument, consideration, talk, consultation, session...

在一些情況下，中文詞彙對應至某英文詞彙的次數雖然很高，但當該中文詞彙對應至該英文詞彙的時候，大多同時對應到不只一個詞彙。當這樣的狀況發生時，此對應很可能不是正確的翻譯，例如「重視」一詞對應至「importance」的次數相當高，但實際上「重視」並不適合做為「importance」的單詞翻譯，因為當「重視」對應至「importance」的時候，其完整對應多為「attaches great importance to」，而不是單獨對應至「importance」。因此，在統計中文詞的反向對應時，我們會計算當中文詞對應至某個英文詞時，同時對應至多個英文詞彙的機率，並排除此機率過高的對應。

本階段的輸出為英文實詞的翻譯篩選結果。我們以反向對應的統計結果，計算中文詞彙的分數，決定該中文詞彙是否做為原英文實詞的翻譯。若中文詞彙 u 對應至英文詞彙$v_1, v_2, v_3, \dots, v$ ，$Pro(u, v_i)$表示 u 對應至v_i的機率，則 u的分數為

$$Score(u) = \sum_{i=1}^{n} (Pro(u\ \ , v_i) \times f(v_i))$$

若v_i為原輸入實詞則$f(v_i) = 1$；若v_i為原輸入實詞的複數或動詞變化形，由於英文的複數變化及動詞名詞變形翻譯至中文時，經常翻譯成相同的中文詞彙，因此我們仍將$f(v_i)$設為 1 ；若v_i為原輸入實詞的相似詞，因相似詞在翻譯至中文時，可能會有些許差異，故我們將$f(v_i)$設為 0.5；另外由於實詞的正確翻譯也時常對應到與實詞搭配的介系詞，故若v_i為原輸入實詞所搭配的介系詞，我們也將$f(v_i)$設為 0.5 ；其他狀況則$f(v_i) = 0$。經過測試與觀察，我們訂定分數標準為 0.15，即若計算結果大於 0.15，則此中文詞彙入選為原英文實詞的翻譯。

以「speech」的翻譯「發言」為例，下表為「發言」所對應到的與原英文實詞「speech」相關的英文詞彙及機率，則發言的分數為 0.074×1 + (0.044 + 0.037 + 0.305 + ⋯) ×1 + (0.021 + 0.001 + 0.003 + ⋯)×0.5 = 0.531，大於 0.15，故入選為「speech」的翻譯。

表 4.「發言」所對應的英文詞彙及機率
[Table 4. Corresponding English words of "發言" fayan "speech" and the translation probability]

	詞彙	機率
原實詞	speech	0.074
原實詞的複數或動詞變化形	speeches	0.044
	speaking	0.037
	speak	0.305
	spoke	0.018
	speakers	0.003
	spoken	0.033
	speaks	0.002
	speaker	0.001

	詞彙	機率
原實詞的相似詞	address	0.021
	addressing	0.001
	addresses	0.003
	talked	0.0001
	addressed	0.0003
	articulate	4e-05
	voice	0.001
	talk	0.001
	voices	4e-05

3.3 篩選介系詞翻譯 (Extracting Translations of Prepositions)

本階段的目標是產生介系詞的翻譯。本階段的輸入為我們要查找的英文介系詞（例如：「speech on」中的「on」）和上一階段（第三章第（二）節）所輸出的實詞翻譯，以及上上階段（第三章第（一）節）改善後的資料。

我們首先統計英文介系詞出現在該英文實詞後面的時候，對應至每一種中文詞性的次數及機率，以及對應至該詞性的各種中文詞彙的次數。由於介系詞在翻譯時也時常被省略，因此我們除了統計介系詞對應至各詞性的機率，也計算介系詞「沒有對應」（記

為 NULL）的機率。由於在某些介系詞省略翻譯的狀況中，介系詞不會沒有對應，而是會對應至其所搭配的實詞所對應的中文實詞，例如「problem of」中，「of」可能和「problem」一起對應至「問題」，因此我們會將此類狀況也列入介系詞沒有對應的機率。以「problem of」為例，下表為「problem of」中的「of」的統計結果：

表 5.「problem of」的「of」翻譯結果
[Table 5. Chinese translations results of "of" in "problem of"]

詞性	機率	詞彙及次數
DE	0.69	的(3098) 之(14)
NULL	0.28	NULL(1247)
T	0.01	的(58)
Na	0.01	工作(7) 人數(6) 程度(5) 精神(5) 的(5) 過程(3) 成員(2) 問題(2) ...

介系詞的正確翻譯大多為特定詞性，因此許多錯誤的對應源自對應到錯誤的詞性，因此在統計之後，我們首先以詞性做篩選，我們認為較合理的詞性有「DE」（如「的」、「之」）、「P」（如「在」、「對」）「Ng」（如「上」、「之間」）、「Caa」（如「與」、「和」）。以上表的「problem of」為例，經過詞性篩選後，只會留下詞性「DE」。

然而在某些情況下，介系詞所對應的詞性會較為特殊，例如「action against」可翻譯為「反對...的行動」，其中介系詞「against」翻譯為「反對」，但「反對」是動詞，不屬於上述我們認為合理的詞性，因此若僅以上述的方法篩選，「反對」將不會被列入可能的翻譯。因此，我們人工整理了雙語辭典中的資料，並加入做為例外條件。

經過篩選後，刪除了許多不合理的對應，因此我們將將篩選後的對應的機率值做正規化，將機率值等比例放大至總和為 1。最後，我們以正規化後的機率，篩選詞性，再從選出的詞性中，以詞彙次數篩選詞彙。以「discussion on」中的「on」為例，表 6 為「on」的對應經過篩選並將機率值正規化後的結果，我們從詞性機率篩選出「NULL」及詞性「P」，並從詞彙次數篩選出「在」、「對」、「就」、「於」等詞彙。

表 6.「discussion on」中的「on」的對應機率
[Table 6. The translation probability of "on" and Chinese correspondences in "discussion on"]

對應中文詞性	機率	詞彙及次數
NULL	**0.724**	
P	**0.226**	**在|35 對|21 就|10 於|9** 關於|3 從|2 對於|2 以|1 自|1 針對|1 至於|1
Ng	0.039	上|7 時|7 後|1
DE	0.01	的|4

3.4 產生實詞與介系詞搭配後的翻譯 (Translating Content Words and Preposition Collocations)

本階段的目標是產生實詞和介系詞搭配後的翻譯。本階段的輸入是在上上階段（第三章第（二）節）所輸出的實詞翻譯和上階段（第三章第（三）節）所輸出的介系詞翻譯，我們也會使用在第一階段（第三章第（一）節）改善後的雙語對齊資料，以及中文語料庫。

首先我們要從中文語料庫中擷取中文高頻搭配。我們用中文斷詞與詞性標註系統，處理中文單語語料庫，將句子斷詞並標註詞性，然後使用 Smadja 於 1993 年(Smadja, 1993)提出的搭配詞提取方法「Retrieving Collocation from Text: Xtract」，在純中文語料庫中擷取每個詞彙的高頻搭配，建立高頻搭配表。由於英文介系詞所對應到的中文翻譯多為特定詞性，因此在計算高頻搭配詞時，我們將各種詞性分開計算，以「發言」為例，其搭配詞提取的部分結果如表 7：

表 7.「發言」的搭配詞
[Table 7. The collocation of "發言" fayan "speech"]

基本詞	搭配詞性	搭配詞	位置
發言	P	在	-3
發言	P	在	-2
發言	D	就	-2
發言	D	就	-3

接著，我們針對在前面階段中所擷取的實詞翻譯和介系詞翻譯，嘗試各種組合，檢查是否為中文高頻搭配，若是，則做為搭配後的翻譯。以上述提到的「speech on」為例，「speech」的翻譯有「發言」，而「on」的翻譯有「在」和「就」，則由上表我們可以找到「在 _ _ 發言 」、「在 _ 發言」、「就 _ _ 發言 」、「就 _ 發言」這幾組翻譯（此處以底線代表空格）。若介系詞翻譯包含「NULL」，代表此介系詞在翻譯時經常被省略，故我們會將實詞皆做為搭配後的翻譯。為了得到更精確的翻譯，產生翻譯後，我們會在回到平行語料庫中，檢查這些翻譯在平行語料庫中出現的次數，並篩除次數太少者。下表以「speech on」為例，展示得出搭配翻譯的過程。

表 8.「speech on」搭配翻譯的產生過程
[Table 8. The process of generating translation of "speech on"]

英文搭配	實詞翻譯	介系詞翻譯	搭配翻譯
speech on	演說、講、發言、演講、言論、辭、演辭、演、演詞	P:在 於 就 關於 NULL: NULL	在 _ _ 發言、在 _ 發言、就 _ _ 發言、就 _ 發言、演說、講、發言、演講、言論、辭、演辭、演、演詞

然而，有時某個介系詞翻譯雖然為某個實詞翻譯的高頻搭配詞，但當兩者搭配時，該介系詞翻譯卻經常不是對應到原輸入中與英文名詞搭配的英文介系詞，以「problem of」為例，在前述的篩選方法中，我們擷取了「problem」的翻譯「問題」及「of」的翻譯「的」，而在中文搭配的統計中，「問題 的」為一個相當高頻的搭配，因此在前一個步驟中，「問題 的」會被列為「problem of」的翻譯，但此時「的」通常並不會對應到「problem of」中的「of」，「問題 的」並不適合做為「problem of」的翻譯。為了改善這樣的狀況，我們針對每一組由中文搭配組合而成的搭配翻譯，計算當此組合出現時，介系詞翻譯對應正確（即對應至原輸入中的英文介系詞）的比例，並以此比例值篩選出更為精確的搭配翻譯。

產生搭配翻譯後，我們針對翻譯選取適合的例句。首先我們在平行語料庫中爲每一組英文搭配詞的每組中文搭配翻譯，抽取含有此搭配的句子，因為中文搭配在句子中時常跨越超過 1~3 個詞彙，因此我們在選取例句時我們放寬距離的限制，允許中間的空格填入較多詞彙。為了減少選取錯誤句子的機會，我們將句子原來的自動對齊納入考量，在此句子原來的自動對齊中，此中文搭配翻譯確實對應至此英文搭配，我們才會選取這個句子作為這個翻譯的例句。以「speech on」翻譯至「就 ... 發言」為例，表 9 呈現抽取例句中詞彙自動對齊：

表 9.「speech on」翻譯至「就 ... 發言」的例句
[Table 9. An example pair of sentences including translating "speech on" to "就 … 發言"]

中文例句	我(0) **在(1)** 二讀(2) **發言(3)** 時(4) ，(5) 已經(6) 頗為(7) 詳盡(8) 地(9) 講述(10) 這(11) 項(12) 動議(13) 。(14)
英文例句	i(0) have(1) dealt(2) with(3) this(4) at(5) some(6) length(7) in(8) my(9) **speech(10)** **on(11)** the(12) second(13) reading(14) .(15)
中英詞對齊	0-0 5-1 6-1 10-2 10-3 11-4 4-5 7-6 7-7 8-7 9-7 10-7 9-8 0-9 **3-10** 4-10 **1-11** 11-12 2-13 12-13 2-14 13-14 14-15

部分句子可能同時含有兩組以上可能的翻譯，以表 10 的句子為例，在前述的方法中我們擷取了「speech on」的翻譯「就 … 發言」及「在 … 發言」，此句同時含有「就 … 發言」及「在 … 發言」，但僅適合做為「就 … 發言」的例句，不適合做為「在 … 發言」的例句。因此，在以自動對齊進行初步的例句篩選後，我們再針對同時含有兩組以上翻譯的例句，考量搭配翻譯在句子中所跨越的距離等因素，進行篩選，得到更精確的例句。最後，我們以篩選後例句數量，對搭配翻譯做最後一次篩選，得出本研究的翻譯結果。

表 10. 同時含有「speech on」兩組翻譯的例句
[Table 10. An example sentence containing two Chinese translations of "speech on"]

中文例句	… **在** 我 **就** 動議 議案 **發言** 後 …
英文例句	… after making the **speech on** my motion.

4. 實驗與評估 (Experiment and Evaluation)

4.1 資料集與工具 (Datasets and Tools)

4.1.1 香港立法局會議資料 (Minutes of Legislative Council of the Hong Kong Special Administrative)

我們採用香港立法局會議資料作為統計詞彙對應及擷取例句時使用的語料，此資料為中英雙語平行語料庫，共約 222 萬句，本研究實際使用約 164 萬句。（資料來源：catalog.ldc.upenn.edu）

4.1.2 聯合報 (United Daily News)

我們採用聯合報資料作為研究中文高頻搭時使用的語料，此資料為中文單語語料庫，涵蓋約 230 萬篇中文新聞，共約 7,118 萬句。（資料來源：udn.com）

4.1.3 CKIP中文斷詞系統 (Chinese Knowledge and Information Processing System)

我們使用 CKIP 中文斷詞系統處理中文句子，產生詞性標註。此系統是由中研院詞庫小組開發，提供中文的斷詞與詞性標註。（資料來源：ckipsvr.iis.sinica.edu.tw）

4.2 實驗設定 (Experimental Settings)

4.2.1 名詞介系詞搭配之片語翻譯 (Translations of Praises Including Nouns and Prepositions)

本實驗的輸入為 30 組英文名詞和介系詞的搭配。我們從香港立法局會議資料中，隨機抽取 30 組搭配，作為本實驗的輸入，抽取結果如下。

period in	scheme to	environment for	place at	emphasis on
development by	market for	scope of	time by	agreement between
extension of	help to	potential for	agreement with	pressure on
help from	return to	power in	industry in	communication with
view on	system at	success of	gap between	pressure from
link between	scheme by	satisfaction with	increase from	information about

我們使用第三章所介紹之方法，改善香港立法局會議資料的中文斷詞及中英對應，然後從資料中統計並擷取實詞翻譯及介系詞翻譯，最後將兩者結合產生翻譯及例句，即為本實驗的輸出。

本實驗評估分為翻譯正確率、翻譯召回率及例句正確率三個部份。翻譯正確率即本

實驗產生之翻譯的正確比例，由人工逐一評判得出。翻譯召回率則是指本實驗產生的正確翻譯在香港立法局會議資料全文所有正確翻譯中所佔的比例，由程式計算得出例如若原輸入為「disparity between」，產生之翻譯為「之間 … 差距」、「差距」、「懸殊」，則我們計算所有英文句包含「disparity between」的句子中，其中文句出現「之間 … 差距」、「差距」、「懸殊」的比例，作為翻譯召回率。例句正確率則是針對經人工評判為正確的翻譯，分別抽取 5 句例句，由人工逐一檢視，得出例句正確比例。

4.2.2 單詞翻譯 (Translations of Single Content Word)

因實詞翻譯為本研究中相當重要的一部分，因此我們另外設計了一組實驗評估實詞翻譯的效果。本實驗的輸入為 30 個英文名詞。我們從香港立法局會議資料中隨機抽取了 30 個名詞，作為本實驗的輸入，抽取結果如下。

super	seriousness	nurse	placing	inland
abuse	inject	final	designation	urgent
death	send	city	charge	pain
outlook	divorce	degree	adding	providing
signal	enhancement	cut	wisdom	auditor
position	hotel	identification	confirmation	administrator

我們使用第三章的二之一小節所介紹之方法，從香港立法局會議資料中統計並擷取翻譯，所得之翻譯即為本實驗的輸出。最後使用與評估名介搭配翻譯相同的方法，計算翻譯正確率、翻譯召回率及例句正確率，做為本實驗的評估結果。

4.3 實驗結果與討論 (Evaluation and Discussion)

名詞介系詞搭配翻譯的評估結果如下表：

表 11. 名詞介系詞搭配翻譯評估結果
[Table 11. Evaluation of translations of nouns and prepositions]

英文搭配	翻譯精確率	翻譯召回率	例句精確率	英文搭配	翻譯精確率	翻譯召回率	例句精確率
help from	1.0	0.42	1.0	gap between	1.0	0.08	1.0
power in	1.0	0.35	0.67	success of	0.86	0.65	1.0
help to	0.86	0.61	1.0	scope of	1.0	0.73	1.0
agreement with	0.75	0.51	1.0	period in	0.8	0.84	0.75
environment for	0.5	0.8	0.6	pressure from	1.0	0.07	1.0
agreement between	1.0	0.24	1.0	emphasis on	1.0	0.75	1.0

return to	1.0	0.38	0.6	communication with	1.0	0.71	0.8
extension of	0.78	0.63	1.0	view on	1.0	0.3	1.0
information about	1.0	0.77	1.0	scheme to	1.0	0.82	1.0
potential for	1.0	0.52	1.0	link between	1.0	0.03	1.0
market for	1.0	0.75	1.0	increase from	None	0	None
pressure on	1.0	0.44	0.4	place at	None	0	None

單詞翻譯的評估結果如下表：

表 12. 單詞翻譯評估結果
[Table 12. Evaluation of translations of single content words]

英文搭配	翻譯精確率	翻譯召回率	例句精確率	英文搭配	翻譯精確率	翻譯召回率	例句精確率
seriousness	1.0	0.42	1.0	degree	1.0	0.45	1.0
nurse	1.0	0.89	1.0	signal	1.0	0.49	1.0
inland	1.0	0.03	1.0	enhancement	1.0	0.44	1.0
abuse	1.0	0.69	1.0	cut	1.0	0.48	1.0
final	1.0	0.58	1.0	wisdom	1.0	0.54	1.0
designation	1.0	0.33	1.0	auditor	1.0	0.85	1.0
death	1.0	0.41	1.0	position	1.0	0.59	1.0
city	1.0	0.32	1.0	hotel	1.0	0.69	1.0
charge	1.0	0.39	1.0	identification	1.0	0.37	1.0
pain	1.0	0.52	1.0	confirmation	1.0	0.45	1.0
outlook	1.0	0.35	1.0	administrator	1.0	0.21	1.0

統計以上評估結果，得到兩組實驗的平均翻譯精確率、翻譯召回率及例句精確率，如下表所示：

表 13. 綜合評估結果
[Table 13. Evaluation]

	翻譯精確率	翻譯召回率	例句精確率
名介搭配翻譯	93%	47%	91%
名詞單詞翻譯	100%	49%	100%

我們嘗試直接擷取自動對齊結果做為翻譯，並以相同方式評估翻譯精確率及翻譯召回率，並與本論文方法比較，結果如下表：

表 14. 評估結果比較
[Table 14. Results and discussion]

	自動對齊 翻譯精確率	本論文方法 翻譯精確率	自動對齊 翻譯召回率	本論文方法 翻譯召回率
名介搭配翻譯	60%	93%	52%	47%
名詞單詞翻譯	73%	100%	54%	49%

本論文方法在翻譯精確率方面有大幅提升：於名介搭配翻譯較自動對齊方法精確率提升 33%，於名詞單詞翻譯方面，本論文方法提升 27%，達到 100%正確率。然而，本論文於翻譯召回率方面表現較差，分別為 47%與 49%。

我們深入探討未能擷取翻譯導致召回率較差的原因，原因大致可分為四類，分別為「無對應中文翻譯」、「翻譯為非獨立詞彙」、「斷詞錯誤」、「本方法未成功擷取翻譯」，說明如下表：

表 15. 錯誤類型說明
[Table 15. Descriptions of error types]

錯誤類型	說明
無對應中文翻譯	因雙語句子對齊錯誤或句法改寫，中文句不存在該英文詞彙的翻譯，或該英文詞彙翻譯至中文時被省略。
對應中文翻譯為非獨立詞彙	中文句中確實存在該英文詞彙的翻譯，但並不是一個獨立的詞彙，而是包含於某個中文詞彙中，導致找不到翻譯。例如「high degree of autonomy」譯為「高度 自治」，此時名詞「degree」的翻譯為「高度」中的「度」，而非一個獨立的詞彙。
斷詞錯誤	中文句中確實存在該英文詞彙的翻譯，但因斷詞錯誤導致無法找到翻譯。
本方法未成功擷取翻譯	中文句中確實存在該英文名詞的翻譯，亦沒有發生斷詞錯誤等問題，但本研究的方法無法成功擷取翻譯。

我們共抽取 50 個句子，人工觀察後，得到各類型錯誤的數量及比例，如下表：

表 16. 錯誤類型比例
[Table 16. Ratios of error types]

錯誤類型	數量	百分比
無對應中文翻譯	25	50%
對應中文翻譯為非獨立詞彙	19	38%
斷詞錯誤	2	4%
本方法未成功擷取翻譯	4	8%

由此可知未找到翻譯的句子中，約有 50%在原始資料即無相對應的中文翻譯，因此我們不將錯誤類型一（無對應中文翻譯）的資料納入統計，以此比例估計本方法真實召回率，以及直接擷取自動對齊結果做為翻譯的召回率，結果如下表：

表 17. 評估結果比較（修正召回率後）
[Table 17. Results and discussions with revised recall rates]

	自動對齊 翻譯精確率	本論文方法 翻譯精確率	自動對齊 翻譯召回率 （修正後）	本論文方法 翻譯召回率 （修正後）
名介搭配翻譯	60%	93%	68%	64%
名詞單詞翻譯	73%	100%	70%	66%

實驗結果顯示，相較於直接擷取自動對齊結果做為翻譯，本方法可在翻譯召回率僅小幅下降（兩組實驗分別下降 4%）的情況下，精確率大幅提升（兩組實驗分別提升 33%及 27%），表示本方法能在僅犧牲少數正確翻譯的情況下，篩選掉大量的錯誤翻譯。本方法所擷取的搭配翻譯及單詞翻譯對使用者來說是相當可信的，但召回率仍有改善的空間，且有少數搭配未能找到翻譯（例如「increase from」），顯示仍有不少正確的翻譯本方法尚無法成功擷取。

我們觀察實驗結果較不佳的搭配，發現兩個效果不佳的可能原因，其一，和實詞及介系詞的中文翻譯在句子中的位置距離有關，實詞與介系詞距離太遠、位置不定，導致搭配組合抽取困難，因而未能篩選出可做為翻譯的組合，例如「increase from」這組搭配，在擷取實詞翻譯時，成功擷取「增加」、「提高」、「增長」等翻譯，在擷取介系詞翻譯時，也成功擷取「由」、「從」等翻譯，但在最後根據中文高頻搭配組合的階段，卻未能篩選出可做為翻譯的組合，例如聯合報資料中的句子「**從** 去年 底 的 十一 人 **增加** 到 十四 人」雖包含「從...增加」，但「從」和「增加」的位置距離較遠，導致在計算中文高頻搭配時，未能擷取這樣的搭配。另一個召回率較低的原因為：單詞翻譯中英文的中文翻譯不是一個詞而是語素（例如「death penalty」翻譯至「死刑」，單詞「death」翻譯至「死」，單詞「penalty」翻譯至「刑」），然而本方法無法處理詞彙非一對一翻譯的狀況，因此無法擷取這些翻譯，這可能是導致召回率較低的重要原因。

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

從我們的實驗結果能觀察到，我們的方法所擷取出的翻譯，已能達到不錯的精確率，但在召回率的部分仍有改善的空間，顯示仍有部分翻譯無法由我們的方法找到。

目前有許多方向可以繼續研究。計算高頻搭配時可嘗試考慮更大的距離，以應付介系詞與實詞的對應距離較遠的狀況（例如「increase from」對應至「從...增加」），提升召回率。目前的方法無法處理一詞翻譯至多詞的情況（例如「partnership」對應至「夥伴關係」、「死刑」對應至「death penalty」）若能將這些情況加入考慮，就能更精準擷取翻譯。目前僅限制在擷取名詞及名詞介系詞搭配的翻譯，未來可以嘗試用類似的方法，

來擷取其他詞性（如動詞及動詞介系詞搭配）的翻譯，或擴充至更長片語的翻譯。

綜上所述，我們的研究提出了一套方法，從已做雙語自動對齊的雙語語料庫中，擷取名詞單詞及名詞和介系詞搭配時的翻譯，使用的方法包含統計雙語語料庫中的正反向對應，以及統計單語語料庫中搭配詞，並結合以上兩者。經過評估後，證實我們的方法找到的翻譯已能達到較佳的精確率，並大多能找到正確的例句。

參考文獻 (References)

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D. … Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*(2), 79-85.

Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of 29th Annual Meeting of the ACL*, 169-176. doi: 10.3115/981344.981366

Catizone, R., Russell, G., & Warwick, S. (1989). Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Lexical Acquisition Workshop*, 15-21.

Chen, S. F. (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings of 31st Annual Meeting of the ACL*, 9-16. doi: 10.3115/981574.981576

Debili, F., & Sammouda, E. (1992). Aligning sentences in bilingual texts french-english and french-arabic. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992)*, *2*, 5178524. doi: 10.3115/992133.992151

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of ACL-1995*, 236-243

Fung, P., & Church, K. (1994). K-vec: A new approach for aligning parallel texts. In arXiv preprint arXiv: cmp-lg/9407021

Gale, W. A., & Church, K. W. (1991a). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the workshop on Speech and Natural Language*, 152-157. doi: 10.3115/112405.112428

Gale, W. A., & Church, K. W. (1991b). A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of 29th Annual Meeting of the ACL*, 177-184. doi: 10.3115/981344.981367.

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75-102. doi: 10.5555/972450.972455

Kay, M., & Roscheisen, M. (1993). Text-translation alignment. *Computational linguistics*, *19*(1), 121-142. doi: 10.5555/972450.972457

Melamed, I. D. (1995). Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, 184-198.

Melamed, I. D. (1999). Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics*, *25*(1), 107-130.

Moore, R. C. (2001). Towards a Simple and Accurate Statistical Approach to Learning Translational Relationships Among Words. In *Proceedings of ACL-2001 Workshop on Data-Driven Methods in Machine Translation*, 79-86.

Simard, M., Foster, G. F., & Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, 67-81.

Smadja, F. (1993). Retrieving Collocation from Text: Xtract. *Computational Linguistics*, *19*(1), 143-177.

Wu, D. & Xia, X. (1994). Learning an English-Chinese Lexicon from a Paarallel Corpus. In *Proceedings of AMTA-94*, 206-213.

柯明憲(2006)。**雙語語料庫之多字詞語對應**(碩士論文)。[Ko, M. H. (2006). *Alignment of Multi-word Expressions in Parallel Corpora* (Master's thesis).

Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. 55-68

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*、陳嘉平*、蕭善文+、詹博丞+、呂仲理+

Yu-Jia Zhang, Chia-Ping Chen, Shan-Wen Hsiao,

Bo-Cheng Chan, and Chung-li Lu

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

*國立中山大學資訊工程學系
Department of Computer Science and Engineering, National Sun Yat-sen University
E-mail: M083040025@student.nsysu.edu.tw; cpchen@mail.cse.nsysu.edu.tw
+中華電信研究院
Chunghwa Telecom Laboratories, Taoyuan, Taiwan
E-mail: {whsiao; cbc; chungli@cht.com.tw

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

1. 緒論 (Introduction)

自動語者驗證(Automatic Speaker Verification, ASV)系統隨著深度學習技術的發展，有著顯著的提昇，每一年舉行的相關競賽更是不勝枚舉，不管是 NIST Speaker Recognition Evalution (SRE) (NIST, 2019)，抑或是防止欺騙語音攻擊的 ASVspoof (Todisco *et al.*, 2019)，這些競賽都促使自動語者驗證系統日趨成熟。目前最被廣泛採用的自動語者驗證系統是基於嵌入(Embedding)的架構，該架構由前端的特徵提取器(Feature Extractor)，以及後端的評分器(Scorer)組合而成。前端在音框層(Frame level layer)將原始輸入提取成高階的表徵，並經由池化層整合音框層資訊成為音段層(Segment level layer)，而後透過全連接層提取嵌入並 softmax 計算機率以分類語者。前端架構從傳統的深度神經網路(Deep Neural Network, DNN) / i 向量(i-vector) (McLaren, Lei, & Ferrer, 2015)，一直到近年在語者驗證比賽中，大放異彩的時延神經網路(Time Delay Neural Network, TDNN) / x 向量 (x-vector) (Snyder, Garcia-Romero, Sell, Povey & Khudanpur, 2018) 與以其為延伸的架構：擴展時延神經網路 (Extend-TDNN) (Snyder *et al.*, 2019)；而原先基於影像辨識所建立的卷積神經網路，被認為在語者辨識任務中也能取得不錯的表現，像是殘差神經網路(Residual Neural Network, ResNet) (Nagrani, Chung, Xie & Zisserman, 2020; Xie, Nagrani, Chung & Zisserman, 2019; Qin, Bu & Li, 2019)；另一方面，也有許多研究是針對池化層與損失函數來作改進，池化層除了在時延神經網路中最常使用的統計池化層(Statistic Pooling)之外，自注意池化層(Self-attentive Pooling) (Okabe, Koshinaka & Shinoda, 2018; Zhu, Ko, Snyder, Mak & Povey, 2018)，NetVLAD (Chen *et al.*, 2018)都能使系統在整合音框層資訊的效果更好；損失函數則是從人臉辨識領域借鑒而來，嘗試了許多 softmax 不同的變體：L-softmax (W. Liu, Wen, Yu & Yang, 2016)、A-softmax (W. Liu *et al.*, 2017)、AM-softmax (Wang, Cheng, Liu & Liu, 2018)、AAM-softmax (Deng, Guo, Xue & Zafeiriou, 2019)。而後端的評分器，除了餘弦相似度(Cosine Similarity)，機率線性判別分析(Probabilistic Linear

Discriminant Analysis, PLDA) (Kenny, 2010)成為了最常使用的方法之一。

近年來隨著物聯網設備與智慧家居產品的普及，短語音指令的處理，以及在遠場噪音的真實使用場景下，成為了自動語者驗證系統的新挑戰，而錄音設備的不匹配，更再加深了識別的難度，為了推動該情景下的自動語者驗證系統研究，FFSVC 2020 (Qin *et al.*, 2020)因應而生。

因此，本論文旨在參加 FFSVC 2020，並針對任務一：單一麥克風陣列的遠場文本相關語者驗證(Far-Field Text-Dependent Speaker Verification from single microphone array)與任務二：單一麥克風陣列的遠場文本無關語者驗證(Far-Field Text-Independent Speaker Verification from single microphone array)採取不同的解決方法。我們以基於時延神經網路的擴展時延神經網路，與基於卷積神經網路的殘差神經網路，建立了前端的特徵提取器。聲學特徵採用 FBank (Filter Bank)配上音調(Pitch)。而後也針對殘差神經網路架構進行修改，將其與擴展時延神經網路結合，成為一個新的網路架構稱為時延殘差神經網路(Time Delay Residual Neural Network, TDResNet)，並使用機率線性判別分析作為後端評分器，分別實驗上述模型架構在各任務上的表現。此外，我們也針對池化層做改變，將原先的統計池化層替換成 NetVLAD 的改進：GhostVLAD (Zhong, Arandjelovic & Zisserman, 2018)。更多的實作細節將會在後續的章節詳細說明。

2. 網路架構 (Network Architecture)

2.1 音框層 (Frame Level Layers)

在此章節，我們總共實作了三種不同的架構，一種是擴展時延神經網路，另一種是殘差神經網路，而最後一種則是我們發現前兩者在網路架構上有互補之處，因此我們將其結合，成為一個全新的架構，稱為時延殘差神經網路。

2.1.1 基於時延神經網路 (TDNN-based)

我們參照(Snyder *et al.*, 2019)建立了擴展時延神經網路作為我們的基準(Baseline)，其架構使用了十層來提取音框層的特徵，並且在第 3、5、7 層中使用到了擴張(dilation)的概念，這也是時延神經網路的精髓所在，以擴張來擴大音框層資訊的感知範圍，擴張數分別為 2、3、4，當擴張數為 2 時，我們會取相鄰的 5 個音框進行運算，擴張數為 3 時，則取 7 個音框，以此類推，透過層層堆疊，因此最終可以感知 23 個音框的資訊。接著將音框層輸出經過統計池化層，將音框層資訊整合成音段層資訊，而後由兩層全連接層組成音段層，最後輸出經 softmax 計算機率進行分類；在推論階段，我們從音段層的第一層取出 512 維代表語者的嵌入。每一層皆經過批量標準化(Batch Normalization)與 Rectified Linear Unit (ReLU)激活函數。

2.1.2 基於卷積神經網路 (CNN-based)

根據 (Nagrani *et al.*, 2020; Xie *et al.*, 2019; Qin, Bu & Li, 2019) 的研究，都表明殘差神經網路架構，在有噪音與迴響 (reverberate) 的遠場環境下，對於特徵的擷取是相當出色的，因此採用殘差神經網路作為我們基於卷積神經網路的音框層架構，並參考 (Nagrani *et al.*, 2020) 所提到的 thin-ResNet 架構，實作了參數量較少的殘差神經網路，接著同樣將音框層的輸出經過統計池化層整合，而後的音段層與擴展時延神經網路不同，只採用一層全連接層，並將 softmax 替換成 AM-softmax 計算機率進行分類。每一層皆經過批量標準化與 ReLU 激活函數。每一個殘差區塊 (Residual Block) 皆使用殘差連結 (Residual connect) 連接，最終架構圖如表 1。

表 1. 殘差神經網路架構
[Table 1. Network architecture of ResNet]

#	Module	Structure	Size
0	-	Input 43 Fbank-pitch($43 \times T$)	43
1	Conv	Conv1d, $1 \times 43, 64$	64
2	Conv	$\begin{bmatrix} 1,48 \\ 3,48 \\ 1,96 \end{bmatrix} \times 2$	96
3	Conv	$\begin{bmatrix} 1,64 \\ 3,64 \\ 1,128 \end{bmatrix} \times 3$	128
4	Conv	$\begin{bmatrix} 1,128 \\ 3,128 \\ 1,256 \end{bmatrix} \times 3$	256
5	Conv	$\begin{bmatrix} 1,256 \\ 3,256 \\ 1,512 \end{bmatrix} \times 3$	512
6	Statistic Pooling	Full-seq	2×512
7	Segment	FC	512
8	AM-Softmax		# of speakers

2.1.3 結合時延神經網路與卷積神經網路 (Combination of TDNN and CNN)

擴展時延神經網路與殘差神經網路的差別在於，擴展時延神經網路的前幾層是由擁有擴張的卷積層所組合而成，透過擴張來獲取較大範圍的音框層資訊，隨後緊接著數層無擴張且相同卷積核大小的卷積層來提取並整合音框層資訊，而殘差神經網路的層數雖然較擴展時延神經網路深，但其最終感知的音框層範圍卻不如擴展時延神經網路來得廣闊，所以我們認為兩者有互補之處，擴展時延神經網路後幾層無擴張的部份，如果採用殘差神經網路的機制，就能讓後續的層數越深越大，使得特徵萃取能力更好，因此按照該想法，設計了一個混合的架構，架構如表 2，前五層採用了原先擴展時延神經網路的擴張設計，擴張數分別為 2、3、4、5，後幾層則使用了不同通道大小的殘差區塊各 3 個，該架構保留了擴展時延神經網路獲取較大範圍音框層資訊的能力，同時也擁有殘差神經網

路良好萃取特徵的能力。

表 2. 時延殘差神經網路架構
[Table 2. Network architecture of TDResNet]

#	Module	Structure	Size
0	-	Input 43 Fbank-pitch($43 \times T$)	43
1	TDNN	$[t-2, t+2]$	512
2	TDNN	$\{t-2, t, t+2\}$	512
3	TDNN	$\{t-3, t, t+3\}$	512
4	TDNN	$\{t-4, t, t+4\}$	512
5	TDNN	$\{t-5, t, t+5\}$	512
6	ResNet	$\begin{bmatrix} 1,512 \\ 3,512 \\ 1,1024 \end{bmatrix} \times 3$	1024
7	ResNet	$\begin{bmatrix} 1,1024 \\ 3,1024 \\ 1,2048 \end{bmatrix} \times 3$	2048
8	Statistic Pooling	Full-seq	2×2048
9	Segment	FC	512
10	AM-Softmax		# of speakers

而在 (Li *et al.*, 2018) 論文中，作者提出了與我們想法相近的時延殘差區塊(Time Delay Residual Block, TDResBlock)架構，但他們選擇將時延神經網路模組加入到殘差區塊中。這樣的不同之處在於，我們透過前幾層的擴張得到了固定的感知範圍，才接著使用殘差區塊來萃取與整合特徵，但他們的作法則是讓每個殘差區塊的感知範圍皆不相同，因此每個殘差區塊整合著不同感知範圍的資訊，同時，他們的擴張數最終可達 11，但我們資料的音框數並不足以應付這麼大的擴張，從而導致最終結果可能會受到零填充的影響而變差，因此我們在設計時延殘差神經網路架構時，擴張數只到 5。

2.2 池化層 (Pooling)

除了原先透過計算平均值跟標準差的統計池化層之外，為了解決遠場噪音對於分類的影響，我們嘗試使用 GhostVLAD (Xie *et al.*, 2019)方法，該方法是由 NetVLAD 為基礎改進而來的，NetVLAD 是一種可訓練的分群法，主要做法是將每個音框層的特徵分配到不同的群，接著計算該特徵到群中心的殘差並編碼成最後的輸出，產生 $K \times D$ 大小的矩陣 V。以下為 NetVLAD 計算公式：

$$V(k,j) = \sum_{t=1}^{T} \frac{e^{a_k x_t + b_k}}{\sum_{k'=1}^{K} e^{a_{k'} x_t + b_{k'}}} (x_t(j) - c_k(j)) \quad (1)$$

其中 K 表示群總數，是一個自訂的超參數，D 表示每一個群的維度，與音框層的輸出

通道數相同，a_k， b_k， c_k 是由網路訓練得到的參數，該公式的前半部為 softmax，表輸入 x_t 屬於群 k 的機率，後半部為計算 x_t 與群中心的距離，並以前半部計算出來的 softmax 值作為該距離的權重，再將所有結果相加，最終把所有的群串連成最後的輸出向量 V，而 GhostVLAD 的改進在於向後傳遞時，有些群並不會被包含在最後的輸出當中，如此一來，再訓練網路時，能讓網路自主學習哪些特徵作用較低，應該被分類到需要被排除的群中，而因為被排除的群不會參與到整個網路權重的更新，因此在訓練中似有非有，所以又被稱為 Ghost 群，這也是這個方法的由來，而 Ghost 群也是事先設定好的超參數，它將原先的 K 個群額外增加 G 個 Ghost 群，最後再將 $(K+G)\times D$ 的輸出，只採用 $K\times D$，將代表噪音的 Ghost 群排除掉。我們按照原始論文中的設定，$K=8$，$G=2$。

2.3 損失函數 (Loss Function)

近年來，基於 AM-Softmax 損失函數訓練的語者驗證系統，比起傳統的 softmax 效果有著很大的提昇(Y. Liu, He & Liu, 2019)，因此比起原先的 softmax，我們更偏向採用 AM-softmax，該損失函數將角度間隔的概念引入 softmax 。AM-softmax 損失函數公式如下：

$$L=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot(cos\theta_{y_i}-m)}}{e^{s\cdot(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c}e^{s\cdot(cos\theta_j)}} \tag{2}$$

$cos\theta_{y_i}$ 代表第 i 個輸入的特徵向量與權重向量的角度餘弦值，m 則代表角度邊界，s是尺度係數，用於調整角度餘弦值的大小，m 和 s 皆是超參數，這個損失函數的目標是要最大化$cos\theta_{y_i}-m$ 來讓特徵向量與權重向量的夾角最小。我們參考原論文設定$s=30$，$m=0.2$。

2.4 後端評分 (Back-end Scoring)

2.4.1 高斯機率線性判別分析 (Gaussian PLDA)

後端評分器是基於高斯機率線性判別分析，我們先針對擷取出來的語者嵌入作平均正規化，來降低語者嵌入數值的變異性，接著經由線性判別分析 (Linear discriminant analysis, LDA) 來將嵌入的維度降維到 250 維，並用降維過後的嵌入訓練機率線性判別分析，以及用於機率線性判別分析調適 (Adaptation) 的調整，最後以訓練好的機率線性判別分析模型，計算經轉換過後的語者嵌入間的分數。

2.4.2 分數融合 (Score Fusion)

每個系統都有其不同的嵌入提取器的架構，以及機率線性判別分析和機率線性判別分析調適的評分器，而為了能得到最佳的系統表現，我們結合了多個系統所計算出來的分數，

結合方式如圖 1，我們依照機率線性判別分析與機率線性判別分析調適評分器，將一個模型拆分成兩個子系統，並使用 BOSARIS toolkit (Brummer & De Villiers, 2013) 來校正我們系統分數之間的權重，校正資料集採用 FFSVC 2020 開發集。

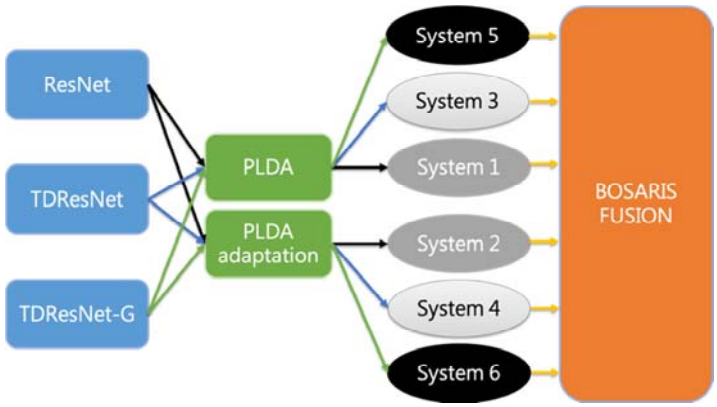


圖 1. 融合策略的示意圖
[Figure 1. Schematic illustration of fusion strategy]

2.5 模型微調 (Model Fine-tuning)

我們採用大量文本無關資料來訓練我們的模型，用以吻合任務二的條件，但如果直接套用在任務一的情境下，所得到的效果會很差，而最直接的解決方法是使用文本相關資料重新訓練嵌入提取器，或者以文本無關資料訓練好的模型作為預訓練(Pre-train)模型，採遷移學習(Transfer Learning)的方式，使用文本相關資料調適模型，但這兩種方法勢必得花費大量的時間，因此我們的作法是選擇將訓練機率線性判別分析與機率線性判別分析調適的資料更換成文本相關資料，能夠節省時間且能達到一定的效果。

3. 實驗設置 (Experimental Setup)

3.1 訓練資料 (Training Data)

競賽方提供的 FFSVC 2020 訓練集，採用麥克風陣列及手機，在不同空間距離、不同雜訊以及不同語速下錄製而成，錄音內容與智慧家居產品的使用情境有關，而除了使用該資料之外，依據競賽要求，我們也從 OpenSLR (OpenSLR, 2020)上挑選開放資料集，其中包含競賽評估計畫中有提及的 SLR-85 (HI-MIA)，而為了要符合本競賽的測試情境，我們也選擇了中文且錄音內容與智慧家居有關的資料集，分別是 SLR-33 (AISHELL)，SLR-38 (FreeST Chinese)，以及 SLR-68 (MAGICDATA)，同時為了增加訓練資料的語者多樣性，我們也加入了在語音任務上經常使用 SLR-49 (VoxCeleb)與 SLR-12 (LibriSpeech)資料集，因此最後總共採用了 7 個不同的資料集用以訓練模型。至於訓練的超參數，我們設定批量大小(Batch Size)為 32，起始學習率為 0.001，並隨著訓練迭代數遞減至 0.0001，模型使用 Nvidia GeForce GTX 1080 Ti GPU 訓練 6 個 epoch。

3.2 資料增強 (Data Augmentation)

訓練資料採用資料增強，一直以來都是被使用於增強語者嵌入模型的強健性(Robustness)，而該篇論文(Qin, Cai & Li, 2019)中有提及，在遠場的環境下，訓練資料與測試資料存在著不匹配的現象，因此為了要模擬遠場的環境，我們針對幾個較大的資料集，使用KALDI toolkit (Povey *et al.*, 2011)以迴響的方式增強我們的訓練資料，最終採用經增強過後的資料來訓練模型，表 3 為我們訓練過程的資料數量與使用方式。

表 3. 訓練過程的資料數量與使用方式
[Table 3. Data usage in the training process]

資料集	語者數	音檔數	語言	資料增強	嵌入提取器訓練	PLDA/PLDA Adaptation
FFSVC 2020 訓練集	120	1,403,383	中	√	√	√
HI-MIA	296	1,157,723	中		√	√
AISHELL	2,331	1,129,626	中	√	√	
FreeST	443	102,600	中		√	
MAGICDATA	1,080	609,550	中		√	
VoxCeleb	7,363	1,281,762	英	√	√	
LibriSpeech	5,831	292,367	英	√	√	

3.3 聲學特徵 (Acoustic Feature)

我們的聲學特徵，採用 KALDI 40 維的 FBank 配 3 維的音調，並且統一取樣頻率為16kHz，音框長度為 25-ms，音框偏移為 10-ms，而特徵擷取完後，使用基於能量的語音活性偵測(Energy-based Voice Activation Detection)來除去沒有聲音的語音片段，許多實驗表明，有無採用基於能量的語音活性偵測對於結果的影響是很大的，接著針對特徵做倒頻譜平均值與變異數正規化(Cepstral Mean And Variance Normalization, CMVN)，降低離群特徵的影響，使模型的訓練效能提昇。

3.4 開發集與驗證集 (Development and Evaluation Data)

使用競賽方所提供的 FFSVC 2020 開發集與驗證集，錄製方式與內容同 FFSVC 2020 訓練集，但彼此語者不重疊。依照競賽要求，測試方式需以手機錄製的音檔作為註冊，麥克風陣列錄製的音檔作為測試。所有實驗皆經由開發集來測試結果，並以其結果來預估在驗證集上的表現，因此開發集並無參與到任何形式的訓練中，僅用來評估模型訓練結果的好壞。

4. 結果 (Result)

我們總共實驗了五種不同的模型在開發集上的表現，逐一對應表 4 的五個系統，分別是擴展時延神經網路、殘差神經網路、時延殘差神經網路、時延殘差神經網路-G（採用GhostVLAD 的時延殘差神經網路）和融合模型，並且僅有時延殘差神經網路、時延殘差

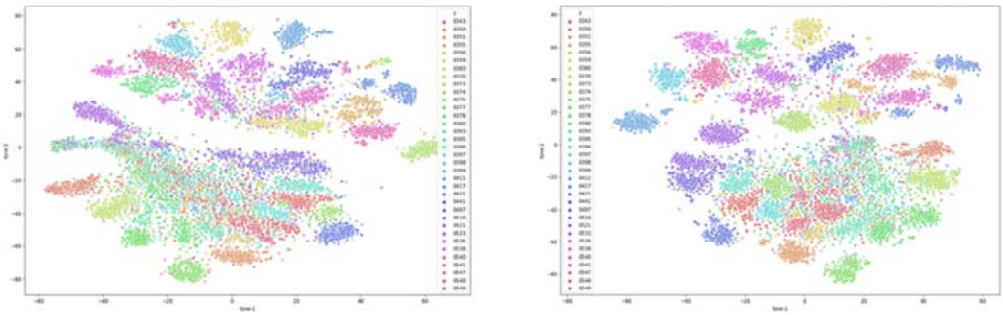
神經網路-G 和融合模型在驗證集上測試並上傳成績，並以有無經過分數融合來區分上傳的系統，經過分數融合的融合模型作為我們競賽的 Primary System 1，而這也是我們在該競賽的最佳系統，另外沒有經過分數融合的時延殘差神經網路與時延殘差神經網路-G 則作為 Single System 1、2， 其中又以 Single System 2 表現較佳。所有結果如表 4 所示。

表 4. FFSVC 2020 開發集與驗證集的最小 DCF 和 EER
[Table 4. Minimum DCF and EER of the FFSVC 2020 development data and evaluation data]

		Development Set								Evaluation Set			
		Task 1				Task 2				Task 1		Task 2	
		PLDA		PLDA Adapt		PLDA		PLDA Adapt					
ID	Model	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER
1	E-TDNN	0.9286	11.94%	0.9231	11.82%	0.9503	12.86%	0.9479	12.43%	-	-	-	-
2	ResNet	0.8755	11.21%	0.8851	10.41%	0.9131	12.47%	0.9191	12.02%	-	-	-	-
3	TDResNet	0.8694	11.02%	0.8740	10.23%	0.9220	11.6%	0.93	10.88%	0.8566	11.45%	0.9132	11.36%
4	TDResNet-G	0.8374	11.59%	0.8295	10.33%	0.8650	12.11%	0.87	11.39%	0.8197	12.19%	0.8994	12.11%
5	Fusion(2+3+4)	-	-	-	-	-	-	-	-	0.7703	9.94%	0.8762	10.31%

4.1 單一系統 (Single System)

Single System 1 的架構如時延殘差神經網路章節中所描述，由 5 層的時延神經網路與 6 個殘差區塊組合而成，池化層採用統計池化層，損失函數是 AM-softmax，嵌入提取器的訓練資料使用訓練資料章節提到的 7 種不同資料集。而因為任務一與任務二的差別在於，任務一為文本相關，註冊與測試音檔的內容皆為“你好，米雅”。任務二則為文本無關，內容與智慧家居設備指令與日常用語，因此機率線性判別分析與機率線性判別分析調適依照不同任務使用不同資料，在任務一我們使用 FFSVC 2020 訓練集(只使用編號 1-30) 加上 SLR-85，其錄音內容皆是“你好，米雅”，這樣做的目的是為了與文本相關的任務一測試情境相同，而文本無關的任務二則使用全部的 FFSVC 2020 訓練集。



(a) TDResNet ***(b) TDResNet-G***

圖 2. FFSVC2020 開發集在不同模型所擷取出來的嵌入經 t-SNE 視覺化
[Figure 2. The t-SNE visualization of the embeddings extracted from the different model embedding layer on FFSVC 2020 development data]

Single System 2 則是採取與 Single System 1 相同的模型架構，唯一的不同在於池化層從原來的統計池化層替換成GhostVLAD，訓練資料與超參數並無做任何的更動。此外，我們也使用 t-distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008) 來分別對 Single System 1、2 的高維度嵌入視覺化，以此評估不同池化層對於最終嵌入學習的影響，結果展示於圖 2，我們可以發現，採用 GhostVLAD 所擷取出來的嵌入經 t-SNE，分群表現較統計池化層佳，尤其是在圖片上半部鮮有重疊者。而從 minDCF 的評估標準來看，分群結果較佳的 Single System 2 也確實表現較佳，這也就表示在 false alarm 與 miss 相同權重的條件下，Single System 2 的驗證效果比 Single System 1 好。

4.2 主系統 (Primary System)

使用經 BOSARIS toolkit 融合過後的系統作為 Primary System 1，選用 ID 2、3、4 作為前端的模型，並且每一個模型分別對應後端的機率線性判別分析和機率線性判別分析調適評分器，因此最終的融合結果由 6 個不同的子系統參與融合後產生，而這個融合系統是我們所有系統中最佳的，於任務一上 EER 9.94%，minDCF 0.7703，在 22 隊參賽隊伍中排名 14 名；於任務二上 EER 10.31%，minDCF 0.8762，在 19 隊參賽隊伍中排名 11 名。

4.3 開發集分析 (Development Data Analysis)

針對 FFSVC 2020 開發集，我們以任務一的時延殘差神經網路測試了空間、雜訊與語速對於結果的影響，為了公平性，我們直接採用競賽方提供的開發集 trials 共 53,996 筆進行測試，並依據我們的測試情形，從 53,996 筆測試中挑選指定的配對，因此每個測試情形的 trials 數量會有些微的差異。

首先，我們實驗註冊與測試在不同空間距離下對於結果的影響，如表 5 所示，0.25m、1m、-1.5m、3m 及 5m 分別表示錄音裝置不同的收音距離，0.25m 為錄音裝置面對說話人 0.25 公尺遠，以此類推，而負號則表示收音距離相反，即是背對說話人來收音。我們發現在 1m 的距離下效果最好，而 -1.5m 的距離效果最差，因為其收音方向與其他距離相反，因此推測效果差與收音方向有關。

表 5. 不同空間距離影響的結果
[Table 5. Effects of different spatial distances]

註冊 / 測試	1m	-1.5m	3m	5m
0.25m	8.519	10.99	10.75	9.312

接著我們實驗雜訊的影響，表 6 為各雜訊表示情境與影響的結果，從結果可以看出，註冊與測試在相同的噪音環境下，結果都表現的較好，即對角線的部分，而當採用無噪音的音檔來註冊時，效果比起其他有噪音的來得好。

表 6. 各雜訊表示情境與影響的結果
[Table 6. Each noise condition and result]

F - 電視 / 辦公室 + 電風扇
T - 電風扇
S - 無噪音

註冊 / 測試	F	T	S
F	3.922	13.06	12.15
T	10.65	6.25	7.543
S	9.735	8.561	7.143

最後測試語速的影響，表 7 顯示各語速平均秒速與影響的結果，結果顯示慢語速的效果是最好的，其次是快語速，最差的是正常語速，我們也發現，當註冊與測試秒速差距越大，效果也越差。

表 7. 各語速平均秒速與影響的結果
[Table 7. The average speed of each speech and the result]

註冊 / 測試	Slow	Normal	Fast
Slow	8.318	9.457	11.69
Normal	9.683	9.584	11.46
Fast	11.75	11.5	9.324
平均秒數(s)	2.39	1.96	1.76

5. 結論 (Conclusions)

在這篇論文中，我們參加了 FFSVC 2020， 並基於時延神經網路與卷積神經網路實作前端的系統，同時也將兩個不同基礎的系統結合，設計出一個新的被稱為時延殘差神經網路的系統，後端實作機率線性判別分析與機率線性判別分析調適並用來融合系統。各系統分別在 FFSVC 2020 的開發集與驗證集上評估，從結果看出時延殘差神經網路勝過原先的兩個系統， 此外，我們也實驗了 GhostVLAD 並與原先的統計池化層做比較。最終，我們的最佳融合系統能在任務一上達到 minDCF 0.7703，EER 9.94%，在任務二上則是 minDCF 0.8762， EER 10.31%。

參考文獻 (References)

Brümmer, N., & De Villiers, E. (2013). The bosaris toolkit: Theory, algorithms and code for surviving the new dcf. arXiv preprint arXiv:1304.2865.

Chen, J., Cai, W., Cai, D., Cai, Z., Zhong, H., & Li, M. (2018). End-to-end language identification using netfv and netvlad. In *Proceedings of 11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018)*, 319-323. doi: 10.1109/ISCSLP.2018.8706687

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690-4699. doi: 10.1109/CVPR.2019.00482

Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Proceedings of Odyssey 2010*, 14.

Li, S., Lu, X., Takashima, R., Shen, P., Kawahara, T., & Kawai, H. (2018). Improving very deep time-delay neural network with vertical-attention for effectively training ctc-based asr systems. In *Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT)*, 77-83. doi: 10.1109/SLT.2018.8639675

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 212-220. doi: 10.1109/CVPR.2017.713

Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. In *Proceedings of ICML 2016*, *48*, 507-516.

Liu, Y., He, L., & Liu, J. (2019). Large margin softmax loss for speaker verification. In arXiv preprint arXiv:1904.03479.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*, 2579-2605.

McLaren, M., Lei, Y., & Ferrer, L. (2015). Advances in deep neural network approaches to speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 4814-4818. doi: 10.1109/ICASSP.2015.7178885

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, *60*, 101027. doi: 10.1016/j.csl.2019.101027

NIST. (2019). NIST speaker recognition evaluation. Retrieved from https://www.nist.gov/itl/iad/mig/nist-2019-speaker-recognition-evaluation

Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. In arXiv preprint arXiv:1803.10963.

OpenSLR. (2020). Open Speech and Language Resources. Retrieved from https://openslr.org/

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Vesely, K. (2011). The kaldi speech recognition toolkit. In *Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Qin, X., Bu, H., & Li, M. (2019). Hi-mia: A far-field text-dependent speaker verification database and the baselines. In arXiv preprint arXiv:1912.01231.

Qin, X., Cai, D., & Li, M. (2019). Far-field end-to-end text-dependent speaker verification based on mixed training data with transfer learning and enrollment data augmentation. In *Proceedings of Interspeech 2019*, 4045-4049. doi: 10.21437/Interspeech.2019-1542

Qin, X., Li, M., Bu, H., Das, R. K., Rao, W., Narayanan, S., & Li, H. (2020). The ffsvc 2020 evaluation plan. In arXiv preprint arXiv:2002.00387.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 5329-5333. doi: 10.1109/ICASSP.2018.8461375

Snyder, D., Villalba, J., Chen, N., Povey, D., Sell, G., Dehak, N., & Khudanpur, S. (2019). The jhu speaker recognition system for the voices 2019 challenge. In *Proeedings of Interspeech 2019*, 2468-2472. doi: 10.21437/Interspeech.2019-2979

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado,H., Nautsch, A., . . . Lee, K. A. (2019). Asvspoof2019: Future horizons in spoofed and fake audio detection. In arXiv preprint arXiv:1904.05441.

Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, *25*(7), 926-930. doi: 10.1109/LSP.2018.2822810

Xie, W., Nagrani, A., Chung, J. S., & Zisserman, A. (2019). Utterance-level aggregation for speaker recognition in the wild. In *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, 5791-5795. doi: 10.1109/ICASSP.2019.8683120

Zhong, Y., Arandjelović, R., & Zisserman, A. (2018). Ghostvlad for set-based face recognition. In *Proceedings of Asian Conference on Computer Vision 2018*, 35-50. doi: 10.1007/978-3-030-20890-5_3

Zhu, Y., Ko, T., Snyder, D., Mak, B., & Povey, D. (2018). Self-attentive speaker embeddings for text-independent speaker verification. In *Proceedings of Interspeech 2018*, 3573-3577. doi: 10.21437/Interspeech.2018-1158

Computational Linguistics and Chinese Language Processing
Vol. 25, No. 2, December 2020, pp. 69-84

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*、曾證融*、廖元甫*、王文俊+、潘振銘+

Wen-Han Hsu, Cheng-Jung Tseng, Yuan-Fu Liao,

Wern-Jun Wang and Chen-Ming Pan

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。

*國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
E-mail: jeff3136169@gmail.com; {t107368030, yfliao}@ntut.edu.tw
+中華電信實驗室
Chunghwa Telecom Laboratories
E-mail: {wernjun, chenming}@cht.com.tw

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 Tacrtron2 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。

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, Tacrtron2 and WaveGlow models. In addition, the real-time factor of the whole system achieved 1/3.5.

關鍵詞：機器翻譯、臺灣閩南語羅馬字拼音、台語語音合成

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

1. 緒論 (Introduction)

造成台語使用人口式微的原因很多，最早可追溯至民國 34 年，國民政府接管臺灣後極力推行的「國語運動」，使得當時的學校禁止各省方言及原住民語，並嚴格推行「國語教育」(李玟逸、李祐萱、周楚，2017)。時至今日，人民生活中大多說國語為主，導致現代人熟悉台語的人數越來越少，尤其是年輕人，大部分懂的台語詞彙不多，講得也不甚

流利。此外，台語語言的演進也慢慢地與生活脫節，常有一些新的時、事、物，例如“磐石艦、討拍、滑鼠”等等，都不知道如何用台語來說，造成大家在用台語講話時，只好常常夾雜國語。

針對台語現階段的困境，若能做出一套中文文字轉台語語音合成的機器翻譯系統，讓使用者輸入中文文字後，能自動合成台語語音，就可以教大家如何講台語。讓使用者對台語提起興趣，並加強台語在日常生活中的應用，進而活化台語。尤其若能同時顯示教育部官方推薦的台羅拼音書寫系統，就能讓學生對台羅拼音有初步的認識及瞭解，進而能直接書寫台語。建立起一套中文文字轉台語語音合成的機器翻譯系統，通常需要三個模組，包括(1)將中文文字轉成以台羅（Tâi-lô）拼音表示的台語講法，(2)將台羅拼音轉為台語合成語音參數，最後(3)將合成語音參數轉成實際台語合成音檔。其中，其中以將中文文字轉成台羅拼音的機器翻譯模組最為重要，因為若翻譯的正確率不高，合成端有再好的音色品質和合成速度都是徒勞。

較早期的機器翻譯方法，有基於規則的字對字機器翻譯(RBMT)，基於範例的句對句機器翻譯(EBMT)，以及統計機器翻譯(SMT) (“機器翻譯，” 2020)。詞對詞的規則法即為將一個中文詞，依據規則與台華平行辭典，對照到一個台語拼音的翻譯法，適用於只注重單詞的非完整句子之翻譯，但翻譯出來的台語文法可能不正確。句對句的範例法為一整串中文句子對照到一整串台語的台羅拼音，可適用於句子的翻譯，也較能考慮文法差異。但此法常需依賴語料庫中台華平行句子的多樣化和數量，如要翻譯從未出現於語料庫中的句子，通常會較為困難。統計法目前為非限定領域機器翻譯中性能較佳的一種方法，通過對大量的台華平行語料進行統計分析，構建統計翻譯模型並進行翻譯，已經可以融合文句中語法等信息進一步提高翻譯的精確性。

例如，交大陳信宏(Kuo, Wang & Chen, 2004) (趙良基，2012)，中興余明興(潘能煌、余明興、許書豪，2011)與台大陳信希(Lin & Chen, 1999)等老師與都曾進行過中文文字轉成台羅拼音機器翻譯的相關研究。其中，陳信希老師曾在 1999 年，就發展出一個基於辭典翻譯，具有語音合成功能的 Mandarin to Taiwanese Min Nan Machine Translation System(目前已終止維護)。而且意傳科技也採用統計方法訓練出一套網頁版本的中文轉台語機器翻譯[1]。不過，此種機器翻譯模組，還需要先有一個華文斷詞與 POS 剖析器(自然語言剖析器，NLP parser)，才能順利進行後續的機器翻譯程序。但 NLP parser 本身就已經是一個難解的問題，而且通常會有大約 5%的分析錯誤(包括斷詞與 POS 標記)。若還是使用此傳統兩階段架構，就會讓前級產生的錯誤，連帶導致後面的翻譯與語音合成錯誤，而且後級只能接受，無法再加以挽救。

而近年來主流的機器翻譯方法為類神經網路機器翻譯(NMT)，顧名思義使用類神經網路(Neural Network)來做機器翻譯，其通常是基於 sequence-to-sequence 模型，使用 encoder-decoder 架構來學習輸入來源語言與輸出目標語言間的對應關係。NMT 尤其常使用 CNN 或是 RNN，來學習自然語言這種具有時間順序的序列數據(Sequence Data)的關

[1] 鬥拍字，https://suisiann.ithuan.tw/

係。例如給 encoder 端的 RNN 輸入一個來源語言的句子後，先利用 RNN 分析來源文字的語意，編碼成一個能代表原語句的語意向量序列。再讓 decoder 端的 RNN，以目標語言的語言模型知識，重新解譯該語意，輸出合乎目標語言架構的語句(Lee, 2019)。這樣就可以讓翻譯結果同時符合詞彙、文法與語意。

另一方面，目前的主流語音合成，也幾乎都是基於類神經網路技術，尤其以 Google 提出的 Tacotron2+WaveNet Vocoder 較為出名。Tacotron2 可直接以類神經網路，進行文脈訊息處理，建立一「文字」轉「Mel-Spectrogram」的 end-to-end 架構。WaveNet Vocoder 接著將「Mel-Spectrogram」轉成「Speech Waveform」。此 Vocoder 出現以後，語音合成的音質就幾乎接近人聲。Tacotron2+WaveNet Vocoder 兩者的組合基本上就是目前的 State-of-the-Art 語音合成技術。但此處的 WaveNet Vocoder，是一個以 sample 為單位做計算的序列式遞迴網路架構，sample 需要一個接著一個照前後順序產生。除計算量相當大外，也不易平行化，導致語音生成速度非常慢，幾乎無法用一般的 GPU 設備達到 real-time 的效能要求。

因此，目前語音合成研究主要是要解決合成速度問題。例如 Wave-RNN 與 WaveGlow。其中，NVIDIA 提出的 WaveGlow 跟 WaveNet 相比，可以避開遞迴網路架構計算量大，且不易平行化的問題，合成所需時間比大幅減少，約為 1:400，若是合成約 10 秒以下的語音，大幅減少的合成時間已經幾乎接近體感的即時合成，且其公開的平均意見得分(MOS)測試也表明，WaveGlow 的音質也不遜於 WaveNet。

因此，基於以上討論，我們將使用 sequence-to-sequence + Tacotron2 + WaveGlow 等模型來實現高品質且即時之台語語音合成。其架構為使用者輸入的中文文本透過 C2T (Ott, Edunov, Grangier & Auli, 2018)轉為台羅拼音，再透過 Tacotron2 (Shen *et al.*, 2018) 將台羅拼音轉為頻譜，最後透過 WaveGlow (Prenger, Valle & Catanzaro, 2018)將頻譜合成出台語語音，如圖 1 所示。

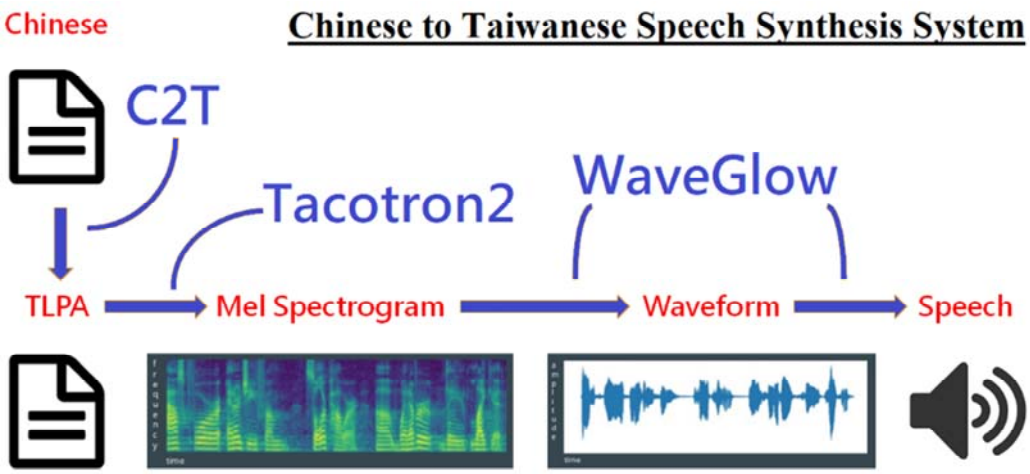


圖 1. 中文轉台語語音合成系統流程圖

[Figure 1. Chinese text to Taiwanese speech synthesis system flow chart]

為訓練此系統中的 C2T 模組，我們將利用中研院的 iCorpus 台華平行語料庫，與從網路上收集的多本台華平行辭典（包含教育部閩南語常用詞辭典），並採用 sequence-to-sequence 深度類神經網路架構，讓模型去學習如何將中文文字，轉換成台羅拼音。並利用中華電信錄製的單一語者台語語音合成語料庫，訓練 Tacotron2 與 Waveglow。希望能盡可能地達到中文文字翻譯成台羅拼音的正確性，與合成台語語音的高度自然度。

2. 中文文字轉台語語音合成系統 (Chinese text to Taiwanese speech synthesis system)

此系統基於深度學習之語音合成技術，以實現高品質且即時之台語語音合成。以下將進一步敘述 C2T，Tacotron2 和 WaveGlow 三個模型的實際作法。

2.1 Chinese to Taiwanese (C2T)

本文中的 C2T 採用 Facebook AI 研究院發表的 fairseq (Ott *et al*., 2018)為架構進行訓練，選用此法的原因為 fairseq 使用了比 RNN 效率和成果表現都更為優秀的 CNN 架構作為基礎。RNN 相比 CNN 有以下幾項缺點，(1)RNN 的模型是時序的，在處理序列的信息時只能逐項處理，不能並行操作，導致運行速度慢。(2)RNN 在處理較長的語句時，間格較遠的詞很難去學到詞與詞之間的依賴關係，並不能很好地處理句子中的結構化等更複雜的信息。(3) RNN 輸入多個單詞時，第一個單詞會經過 n 次單元的計算和非線性，但是最後一個單詞只會經過 1 次。

相比之下，CNN 改善了以上問題，除了能夠並行處理數據，Position Embedding 時輸入除了詞向量還加入位置向量，且 CNN 為層級結構，可顧慮到整段文句的每一個單詞，較底層 CNN 捕捉間隔較近的詞之間的依賴關係，較高層 CNN 則捕捉間隔較遠的詞之間的依賴關係。CNN 在 encoder 端，以 GLU 作為非線性單元，輸入與輸出相加後，才輸入到下一層網絡中，在 decoder 端，有 multi-hop attention 機制，encoder 端的輸出進行加權時，會考慮原始的輸入向量，在每一個卷積層都會進行 attention 的操作，使得模型在得到下一個 attention 時，能夠考慮到之前的已經 attention 過的詞。從 IBM Research 發表的研究論文”Comparative Study of CNN and RNN for Natural Language Processing“ (Yin, Kann, Yu & Schütze, 2017)，也可以看出在處理句子配對的任務上，比起 RNN 的 GRU，LSTM 等模型，CNN 擁有一定的優勢。

2.1.1 語料及辭典 (Corpus and Lexicon)

要訓練一個 C2T 模型，必須先準備好台華平行語料以及台華平行辭典。本模型使用的語料，為中研院資訊所陳孟彰老師計畫內的 iCorpus，此語料庫收集 3266 篇新聞，共 83544 句。算標點符號，台語 504037 詞、1030671 字，華語 501202 詞、1028218 字。以下為 iCorpus 的部份文章內容，如圖 2 所示。

	在地 力量 企業 贊助 萬國 戲院 重生		chai7-te7 lek8-liang7 khi3-giap8 chan3-chou7 Ban7-kok hi3-hng5 tiong5-seng
012	民視新聞報導	012	Bin5-si7-sin-bun5-po3-to7
	位於 大林鎮 的 萬國 戲院 曾經 風光一時 ，		ui7-ti7 Toa7-na5-tin3 e5 Ban7-kok hi3-hng5 chan-keng chhiann-iann7-chit8-si5 ，
	隨著 產業 沒落 經濟 蕭條 人潮 散去 ，		toe3-tioh8 san2-giap8 pang-pai7 keng-che3 chhin3-chhi7 jin5-tiau5 soann3-khi3 ，
	最後 發生 火災 停業 二十多年 。		siang7-boe2 hoat-seng hoe2-chai theng5-giap8 ji7-chap8-goa7-ni5 。
	現在 在 當地人 的 努力 規劃 ，		chit-ma2 ti7 chai7-te7-lang5 e5 lou2-lek8 kui-oe7 ，
	企業 的 贊助 加上 電視台 戲劇 演出 。		khi3-giap8 e5 chan3-chou7 ka-siang7 tian7-si7-tai5 hi3-kiok8 ian2-chhut 。
	配合 當時 的 時空 背景 場景 重建 ，		phoe3-hap8 tong-si5 e5 si5-khong poe3-keng2 tiunn5-keng2 tiong5-kian3 ，
	讓 萬國 戲院 彷彿 回到 風華 時代 。		hou7 Ban7-kok hi3-hng5 na2-chhiunn7 tng2-kau3 hong-hoa5 si5-tai7 。

圖 2. iCorpus 華台平行語料庫
[Figure 2. iCorpus Chinese-TLPA parallel corpus]

辭典方面，則是透過”ChhoeTaigi 找台語 ”網站之台語字詞資料庫，蒐集到的 9 本不同台語辭典合併成的台華平行辭典，各辭典統計的華語詞數，如圖 3 所示。

台語辭典	詞數
1. 台文華文線頂辭典	87670
2. 台日大辭典（台文譯本）	69552
3. Maryknoll台英辭典	55903
4. Embree台語辭典	36820
5. 教育部台語辭典	27487
6. 甘字典	24367
7. iTaigi華台辭典	8713
8. 台灣白話基礎語句	5301
9. 台灣植物名彙	1722
總共	317526

圖 3. 台語辭典華語詞數統計
[Figure 3. statistics of number of lexicon words]

因各個辭典由不同作者所撰寫，格式並非一致，為能在合成系統中使用，需要經過多次校正，校正目的主要是將台語詞翻譯成華語詞，華語詞即為平常生活中口語的慣用文字，檢查華語詞的意思與格式是否正確，並且與之對應的台羅拼音是否為一對一。台羅拼音也需檢查，剔除多餘之意思或符號。最終可使用的台華詞條數 225965，華語詞條數 88881，台語詞條數 153132。

2.1.2 模型訓練 (Model training)

中文轉台羅拼音為一種機器翻譯，做法為將中文文字序列轉台羅拼音序列，利用基於 sequence-to-sequence 深度類神經網路架構，學習如何進行轉換。此 C2T 採用網路上開源的 fairseq 架構(Ott *et al.*, 2018)進行訓練，其包括一 encoder 前端與一 decoder 後端。前端 encoder 負責接收輸入中文文字序列，分析其語意並擷取出文脈資訊向量。後端 decoder 在文脈資訊向量之間加入 attention 之機制與 Convolutional Neural Network 之訓練模型下每個 encoder 權重，利用一中文對應台語拼音平行語料庫(iCorpus)，再加上台華平行辭典進行訓練，以此得到最佳的轉譯台羅拼音序列，如圖 4 所示。

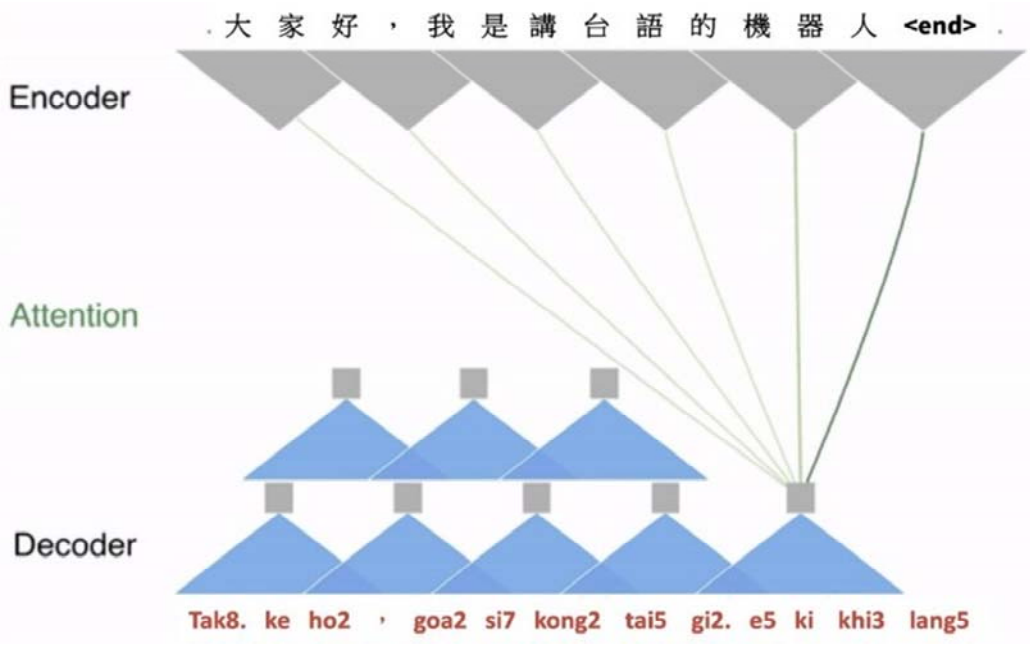


圖 4. 中文轉台語拼音模組
[Figure 4. C2T seq-to-seq model]

2.2 Tacotron2

此處模型訓練用之語料庫來源，為臺北科技大學和李江確台語文教基金會以及意傳科技合作產製，為一有大學教育程度之 34 歲男性錄製，台語腔調偏漳州腔，音檔筆數 9625 筆，長度約 10.4 小時。Tacotron2 為一 end-to-end 方式做訓練與推論之模型，使用架構為 encoder-decoder + Location Sensitive Attention。做法為將翻譯完成的台羅拼音輸入後，類神經網路進行文本分析，把語法與語意轉成語言特徵參數，讓系統知道文本中哪些是詞，哪些是句子，發什麼音，怎麼發音，發音時到哪應該停頓，停頓多長等等。語言特徵參數接著送入韻律產生器來產生文本裡每個音節的對應韻律訊息，包含基頻軌跡，音量，音長等，然後把說話的聲調，語氣，停頓方式，發音長短轉換成韻律參數(朱孝國，2005)。

最後輸出梅爾頻譜圖，再經由對齊達成台羅拼音與頻譜一對一的對應，如圖 5 所示。

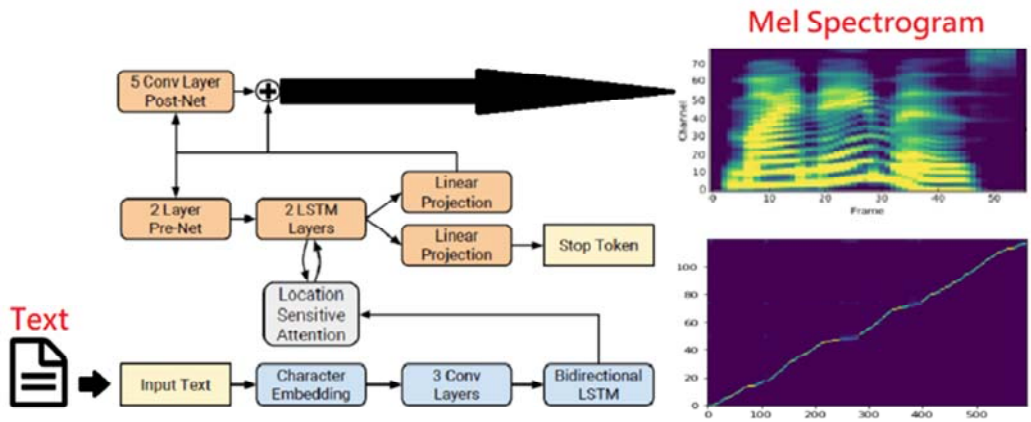


圖 5. Tacotron2 流程圖
[Figure 5. Tacotron2 flow chart]

2.3 WaveGlow

此處模型訓練用之語料庫來源同 Tacotron2，WaveGlow 是一種 flow-based generative networks，結合 Glow 和 WaveNet 的原理，透過輸入音檔與其生成之頻譜，僅使用單個網路與單個損失函式進行訓練，生成一高斯分布 z，合成時只需透過 z 與頻譜就可即時合成高品質語音，如圖 6 所示。

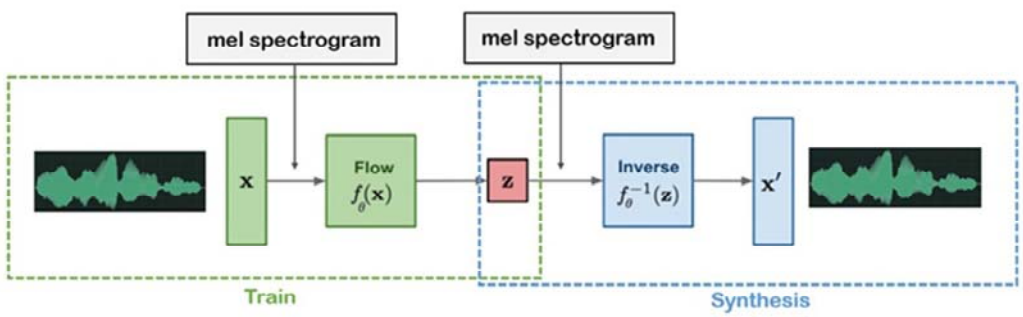


圖 6. WaveGlow 訓練及合成過程
[Figure 6. WaveGlow process of synthesis and training]

訓練時依據基於機率之 cost function 導引，多次利用函式轉換，逐步學習如何將真實語音波形訊號 x 投射到一具高斯分佈之隱藏變數 z 的空間。並在訓練時限制 mapping 函式為可逆函式，WaveGlow 在生成波型圖時即可依據隱藏變數 z 空間取樣的結果，經多次函式轉換，逐步轉換成真實語音波形訊號 x。最後根據需要發出的聲音從資料庫中選擇出合適的聲學參數，然後根據在韻律模型中得到的韻律參數，透過語音合成演算法

產生語音(朱孝國，2005)。如圖 7 所示。

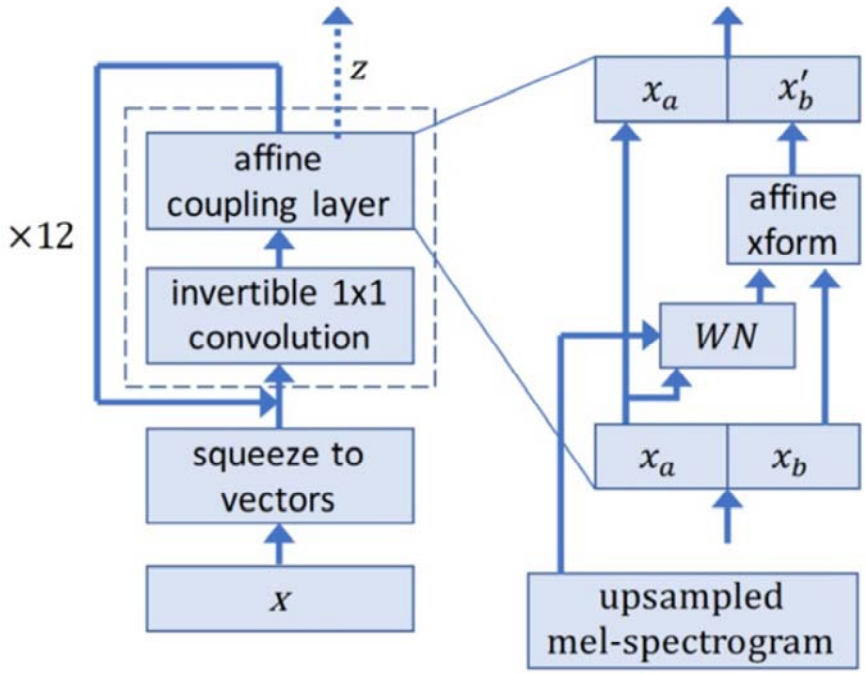


圖 7. WaveGlow 網路結構圖
[Figure 7. WaveGlow network structure diagram]

3. 中文文字轉台語語音合成雛形展示系統 (Website of Chinese text to Taiwanese speech synthesis system)

此基於深度學習之中文轉台語語音合成系統，已架設網頁版本以供使用，連結網址為 http://140.115.54.90:31810/。使用者輸入中文文字，按下合成按鈕就能撥放對應的台語語音，並能一併顯示出翻譯過後的台羅拼音供使用者查詢，且設計了可輸入台羅拼音的欄位讓擁有相關台羅知識的使用者可以鍵入不同的發音方式並合成語音。圖 8 展示為本文之使用者介面。網頁之紅色分隔線上方使用本文之 C2T 機器翻譯，下方為作為比較用之意傳科技統計法機器翻譯，合成端一律使用本文之 Tacotron2+WaveGlow。

其中，初步測試統計式機器翻譯後可以發現，統計式翻譯的結果較為接近中文文字本身念法的音譯，亦即如果要翻譯出中文轉台語後正確的台羅拼音，應輸入台文為佳，所謂台文即為中文轉為閩南語的另一種書寫表示形式。如中文的「現在是晚上八點」，統計式機器翻譯結果為「hian7 tsai7 si7 mng2 siong7 peh4 tiam2」，不是正確的台語發音，而台文的「這馬是暗時八點」翻譯後的「tsit4 ma1 si7 am3 si5 peh4 tiam2」，才是正確的台語發音。因此意傳科技此機器翻譯的「中文」轉台語翻譯，還是有所牽強。

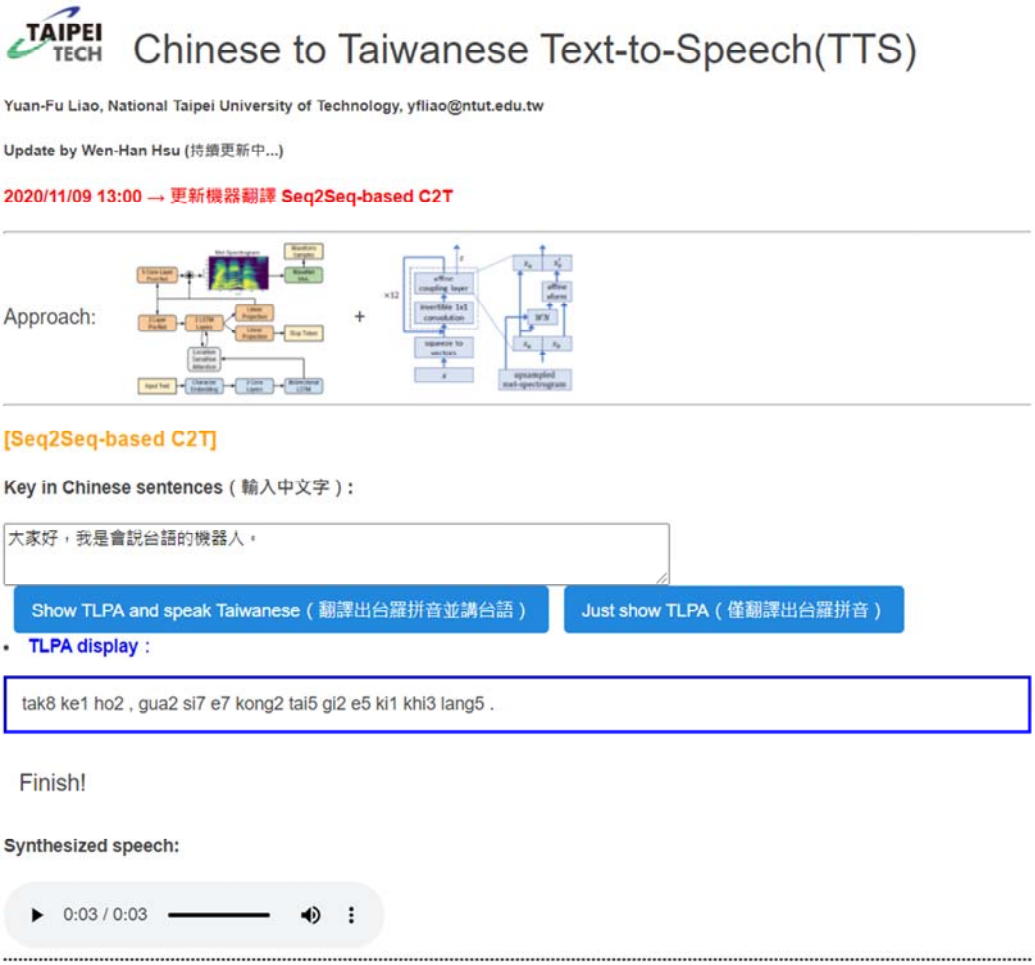


圖 8. 中文轉台語語音合成系統網頁
[Figure 8. Website of Chinese text to Taiwanese speech synthesis system]

4. 實驗 (Experiment)

4.1 C2T效能實驗 (C2T efficacy experiment)

本文之 C2T 模型分別以三種運行模式進行效果測試。模式一為中文句子透過斷詞後，判斷各個詞之詞性，再將各詞轉為台羅拼音，因此可以得到斷詞後含詞性之台羅拼音，如表 1 所示；模式二為將整句中文句子直接轉換為台羅拼音，如表 2 所示；模式三為將整句中文句子直接轉換為台羅拼音，同時學習台語之斷句規則，如表 3 所示。

表 1. 中文轉台羅拼音^斷詞/詞性
[Table 1. Chinese to TLPA^Hyphenation/ Part of speech]

中文句子	傅達仁今將執行安樂死，卻突然爆出自己 20 年前遭緯來體育台封殺，他不懂自己哪裡得罪到電視台。
斷詞	傅達仁 今 將 執行 安樂死 ， 卻 突然 爆出 自己 20 年 前 遭 緯來 體育台 封殺 ， 他 不 懂 自己 哪裡 得罪到 電視台。
詞性	Nb Nd D VC Na COMMACATEGORY D D VJ Nh Neu Nf Ng P Nb Na VC COMMACATEGORY Nh D VK Nh Ncd VJ Nc PERIODCATEGORY
台羅拼音^斷詞/詞性	poo3^B/Nb tat8^I/Nb jin5^E/Nb kim1^S/Nd tsiong3^S/D tsip4^B/VC hing5^E/VC an^B/Na lok8^I/Na si2^E/Na , khiok^S/D tut8^B/D jian5^E/D pok8^B/VJ chhut^E/VJ ka^B/Nh ki7^E/Nh ji7^B/Neu tsap8^E/Neu ni5^S/Nf tsing5^S/Ng cho^S/P hu7i^B/Nb la5i^E/Nb the2^B/Na iok8^I/Na tai5^E/Na hong^B/VC sat^E/VC , i^S/Nh bo5^S/D tong2^S/VK ka^B/Nh ki7^E/Nh to2^B/Ncd ui7^E/Ncd tioh8^B/VJ choe7^I/VJ kau3^E/VJ tian7^B/Nc si7^I/Nc tai5^E/Nc .

表 2. 中文轉台羅拼音
[Table 2. Chinese to TLPA]

中文句子	中央流行疫情指揮中心，今日表示，國內無新增確診個案。
台羅拼音	Tiong iang liu5 heng5 ek8 cheng5 chi2 hui tiong sim , kin a2 jit8 piau2 si7 , kok lai7 bo5 sin cheng7 chin2 ko3 an3 .

表 3. 中文轉台語詞
[Table 3. Chinese to words of TLPA]

中文句子	里長的言論在 PTT 引發熱議許多網友紛紛留言。
台語詞	li2-tiunn2-e5 gian5-lun7 ti7 PTT in2-huat4 jiat8-gi7 tsiann5-tse7 bang7-iu2 hun1-ue7 .

為測試以上三種 C2T 模式的效能，我們以 iCorpus 台華平行語料與辭典進行測試。實驗資料庫包括 iCorpus(78821 句)與台華辭典合集(225965 詞條)，並切分成三個子集，包括 Train 90%，Valid 5%，Test 5%。系統效能則以 Perplexity，及 Word error rate(WER)來衡量結果。考慮聲調的情況下，模式一到模式三的 WER 分別為 25.265%，7.102%以及 9.211%。不考慮聲調的情況下，模式一到模式三的 WER 分別為 18.660%，6.530%以及 8.699%。綜上數據得知，模式二由中文句子直接轉換為台羅拼音的效果最佳，如圖 9 所示。因此本文之 C2T 最終採用將中文直接轉換為台羅拼音的轉換法，以利接下來的台語語音合成工作。

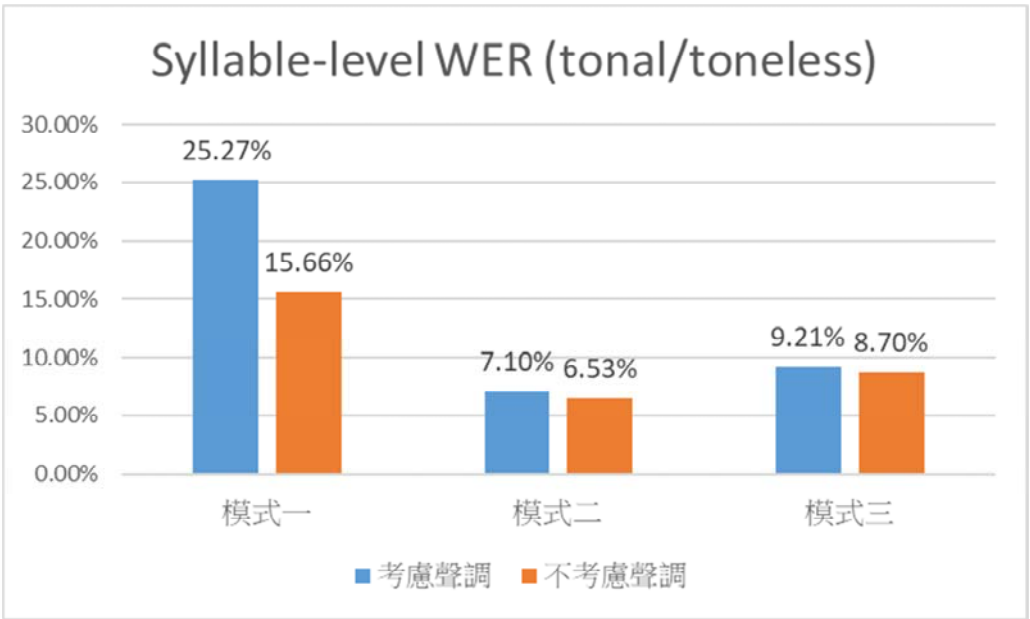


圖 9. C2T 錯誤率比較
[Figure 9. C2T Syllable-level WER]

4.2 Tacotron2+WaveGlow 合成語音品質實驗 (Tacotron2+WaveGlow synthesized speech quality experiment)

將事先準備好的 15 句合成音檔放上 Google 表單，並開放一般人對每個句子單獨進行評分。忽略語音內容翻譯錯誤或是語意不順等因素，僅根據聽到的「品質」評分 1.0 到 5.0 分。最低分為 1.0 分 為最接近機器人講話的聲音；最高分為 5.0 分 為最接近真人講話的聲音。評分到小數點後一位。開放評分時間約兩天，截止時有 20 位聽者評分。表 4 之 S1 到 S15 代表 15 句中文語句內容，為了不讓翻譯錯誤或是語意不通順等因素影響聽者對音檔品質的評分，所有句子皆有透過人工校正台羅拼音，且適當的加入斷詞標記使句子整體語氣通順，讓聽者可專心針對音檔「品質」評分。實驗最終結果之盒鬚圖如圖 10 所示。

總共 300 筆評分資料，最終平均意見得分(mean opinion score，MOS)約為 4.30 分。此實驗所使用之 15 個音檔皆由系統雛型展示網頁合成，由此高得分可知本文所用的 Tacrtron2 與 WaveGlow 模型確實可以正確合成出接近真人聲音之台語語音。

表 4. 測試中文語句內容
[Table 4. Chinese sentence content for experiment]

S1	大家好，我是會說台語的機器人
S2	請根據聽到聲音的音質打分數
S3	今天一早起來，天氣就非常炎熱
S4	一千兩百三十四萬五千六百七十九點零一美元
S5	韓國瑜是台灣歷史上，第一位被罷免的縣市首長
S6	武漢肺炎的出現，讓全世界的人都開始戴口罩
S7	現在是晚上八點，有一些老人應該想睡了
S8	歐美國家如:美國、加拿大、德國、法國、英國、西班牙、瑞典、瑞士、挪威、芬蘭等等
S9	這個猴死囝仔，竟然偷恁爸的錢去買那個垃圾
S10	吃予肥肥，裝予錘錘，裝予水水，等領薪水
S11	昨天地震時，我們家的花瓶掉下來摔破了
S12	龜笑鱉無尾，鱉笑龜粗皮
S13	歡迎光臨，請問有幾位
S14	有颱風從太平洋來的時候，中央山脈常常幫台灣的西部擋去很多災情
S15	謝謝你付出寶貴的時間，參加這次的調查

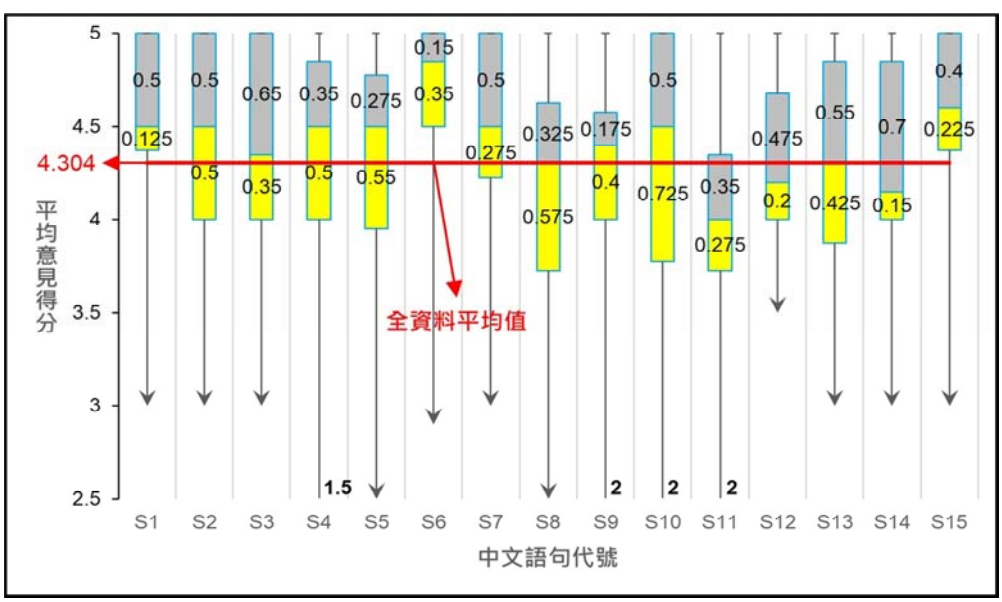


圖 10. 合成語音品質實驗結果盒鬚圖
[Figure 10. Box-plot of experimental results]

4.3 WaveGlow語音合成速度實驗 (WaveGlow speech synthesis speed experiment)

選用 WaveGlow 當作合成端的好處在於即時的語音合成速度，表 5 為一個簡單的 WaveGlow 合成音檔的速度實驗，時間的單位為秒。

表 5. WaveGlow 合成速度實驗
[Table 5. WaveGlow synthesis speed experiment]

音檔長度	5.83	4.25	7.48	3.96	9.16
合成花費時間	1.90	1.22	1.83	1.08	2.74

由表 5 可以得知，本文中的 WaveGlow 一秒約可合成 3.5 秒的音檔，相比原始合成速度非常緩慢的 WaveNet，已經可以達到即時合成的效益。

5. 結論(Conclusion)

我們提出的 C2T 機器翻譯，最好的 CER 值為 6.53%。表示 sequence-to-sequence 模型確實可以將輸入的中文文本，翻譯成正確的台羅拼音串。台語語音合成的 MOS 得分為 4.30 分，表示 Tacrtron2+WaveGlow 模型確實可以正確將台羅拼音串轉成台語語音。且系統的語音合成速度為一秒可合成約 3.5 秒之音檔，達到即時語音合成的要求。由以上實驗結果可以驗證我們中文文字轉台語語音的初步作法，確實有得到一定成效。未來將繼續增加台華平行辭典以及台語語料，進一步改善 C2T 的正確率，與合成語音的自然度。

致謝(Acknowledgements)

感謝中華電信研究院對於本論文提供的資源與協助。This work is supported partially by 中華電信 under the project “基於深度學習之台語語音合成『文字－聲學參數模型』”，Taiwan’s Ministry of Education under the project ”教育部閩南語語音語料庫建置計劃” and partially by Ministry of Science and Technology under the contract number 107-2221-E-027-102, 107-2911-I- 027-501, 107-3011-F-027-003, 108-2221-E-027-067 and 109- 2221-E-027-108.

參考文獻 (References)

Kuo, W.-C., Wang, Y.-R., & Chen, S.-H. (2004). A MODEL-BASED TONE LABELING METHOD FOR MIN-NAN/TAIWANESE SPEECH. In *Proceddings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2004*, 505-508. doi: 10.1109/ICASSP.2004.1326033

Lin, C.-J. & Chen, H.-H. (1999). A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan. *Int. J. Comput. Linguist. Chinese Lang. Process.*, *4*(1), 59-84.

Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling Neural Machine Translation. In arXiv preprint arxiv: 1806.00187

Prenger, R., Valle, R., & Catanzaro, B. (2018). WaveGlow: A Flow-based Generative Network for Speech Synthesis. In arXiv preprint arxiv: 1811.00002

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., …Wu, Y. (2018). Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions. In arXiv preprint arxiv: 1712.05884

Yin, W., Kann, K., Yu, M., & Schutze, H. (2017). Comparative Study of CNN and RNN for Natural Language Processing. In arXiv preprint arXiv:1702.01923.

Lee, M. (2019 年 6 月 17 日)。淺談神經機器翻譯 & 用 Transformer 與 TensorFlow 2 英翻中 [部落格文字資料]. 取自 https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html [Lee, M. (2019, June 17). Talk about neural machine translation & Using Transformer and TensorFlow 2 translates English into Chinese. [Web blog message]. Retrieved from https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html

朱孝國(2005 年 4 月 28 日)。語音合成 Speech Synthesize Note [部落格文字資料]。取自 https://irw.ncut.edu.tw/peterju/speech.html#categories [Chu, H.-K. (2005, April 28). 語音合成 Speech Synthesize Note [Web blog message]. Retrieved from https://irw.ncut.edu.tw/peterju/speech.html#categories]

李玟逸、李祐萱、周楚(2017 年 3 月 31 日)。年輕人不諳台語 語言傳承不能等【部落格文字資料】。取自 http://shuj.shu.edu.tw/blog/2017/03/31/年輕人不諳台語-語言傳承不能等/。[Wen-Yi Li, You-Shiuan Li, Chu Jou. (2017, March 31). Young people do not know Taiwanese. Language inheritance cannot wait. [Web blog message]. Retrieved from http://shuj.shu.edu.tw/blog/2017/03/31/年輕人不諳台語-語言傳承不能等/

趙良基 (2012)。**台語語音合成技術之研究**(碩士論文)。取自 http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22GT070060254%22.&searchmode=basic [Chao,L.-J. (2020). *The Research of Speech Synthesis Technology for Taiwanese (Master ' s thesis*). Retrieved from http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22GT070060254%22.&searchmode=basic]

潘能煌、余明興、許書豪(2011)。中文文句轉台語語音系統之連音變調預估模組。**資訊科技國際期刊**，**5**(1)，118-128。[Pan, N.-H., Yu, M.-S., & Shiu, S.-H. (20110. A Tone Sandhi Prediction Module for Chinese to Taiwanese Text-to-Speech Systems. *Int. J. Adv. Inf. Technol*., *5*(1), 118-128.

機器翻譯(2020 年 9 月 17 日)。In Wikipedia, the free encyclopedia. Retrieved November 10, 2020, from https://zh.wikipedia.org/wiki/机器翻译 [Machine Translation (2020, September 17). In Wikipedia, the free encyclopedia. Retrieved November 10, 2020, from https://zh.wikipedia.org/wiki/机器翻译]

Computational Linguistics and Chinese Language Processing
Vol. 25, No. 2, December 2020, pp. 85-98

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*、洪志偉*

Li-Chia Chang and Jeih-weih Hung

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的語音特徵，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。

Abstract

In this study, we focus on developing a novel 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

*國立暨南國際大學電機工程學系
Department of Electrical Engineering, National Chi Nan University
E-mail: s108323518@mail1.ncnu.edu.tw; jwhung@ncnu.edu.tw

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

關鍵詞：雜訊強健性之語音特徵、語音辨識、深度學習

1. 緒論 (Introduction)

在現今的時代，電子產品和相關服務，例如手機、助聽器、耳機、電話會議系統等，在我們的生活中逐漸變成日常且不可或缺的需求，在這些設備和服務中，語音的功能和應用（語音互動，語音通話，語音辨識等）是一個相當重要的環節。然而，現實通訊環境中中存在各種干擾源，而這些干擾源會干擾語音訊號，因此減損了上述語音功能和應用的性能。這些干擾源包括加成性雜訊、通道干擾和混響等，近幾十年來，各方已經研究開發出多種技術來降低這些干擾效應、以改進語音相關的功能。沿此方向，在本研究中，我們著眼於語音訊號中加成性雜訊的干擾的問題，提出了一種專門用於提高語音辨識準確度的新型降噪方法。

語音通話和辨識中，環境雜訊的存在很可能反映在語音訊號的品質和能辨度以及辨識的準確度上。為了降低雜訊的影響，研究者從語音處理系統的不同角度提出了多種方法，例如前端訊號處理(front-end signal processing)，聲學特徵擷取(acoustic feature representation)和後端聲學模型(back-end acoustic model)等。

在聲學特徵擷取方法中，例如，相對頻譜法(relative spectral analysis, RASTA) (Hermansky & Morgan, 1994)設計一個基於聲學知識的無限脈衝響應濾波器，其應用在語音訊號的對數頻譜中可以有效的抑制訊號中非語音的成分，著名的 RASTA-PLP (Hermansky, Morgan, Bayya & Kohn, 1991)語音特徵表示就是將感知線性估計(perceptual linear prediction, PLP) (Hermansky, 1990)的語音特徵經由 RASTA 處理。此外，對語音特徵序列進行不同階級的正規化可有效地減輕訓練與測試資料的統計不匹配，而研究中也指出藉此可以同時降低雜訊的影響，相關的方法包括平均正規化法(mean normalization, MN) (Liu, Stern, Huang & Acero, 1993)、正規化法(mean and variance normalization, MVN) (Viikki & Laurila, 1998)和統計圖等化法(histogram equalization, HEQ) (Torre *et al.*, 2005)，上述方法分別對語音特徵正規化了平均、平均與變異數、機率密度函數。

在後端聲學模型中，聲學模型之調適法旨在調整聲學模型去適應嘈雜環境下的輸入語音特徵，一些知名的方法例如最大後驗機率自適應模型(maximum a posteriori

adaptation, MAP) (Su, Tsao, Wu & Jean, 2013)、最大似然線性回歸(maximum likelihood linear regression, MLLR) (Stolcke, Ferrer, Kajarekar, Shriberg & Venkataraman, 2005)與最大似然線性轉換(maximum likelihood linear transformation, MLLT) (Gales, 1998)應用於聲學模型的參數（例如高斯混和模型的平均與共變異數）並進行映射轉換。此外，鑑別式聲學模型同樣有非常好的效果，其透過設計訓練時的目標函式，以達到直接改善辨識時的準確率，例如，在單語句辨識中，最小化分類錯誤(minimum classification error, MCE) (Juang, Hou & Lee, 1997)的目標函式是最佳化其辨識的分類結果，而非只是模擬輸入的語句；在大辭彙連續語音辨識中，最小化音素錯誤(minimum phone error, MPE) (Povey, 2003)和最小化詞錯誤(minimum word error, MWE) (Kuo & Chen, 2005)所得的聲學模型訓練目標是對訓練資料能最小化對辨識錯誤的平滑估計。

由於近年來深層神經網絡(deep neural network, DNN)技術的蓬勃發展，語音處理的前後端方法都得到了顯著改善和進步，從而獲得更好的效能，例如，在語音強化領域中，可以使用 DNN 透過大量成對的雜訊語音與乾淨語音，來學習二者之間的映照(mapping)關係；在聲學特徵擷取與聲學模型方面，DNN 同樣也部份甚至全面地取代了傳統的方法，例如，ANN-HMM(artificial neural network-hidden Markov model) (Bourlard & Morgan, 1994)的雙向系統藉由 ANN 更精確地估計出語音特徵的似然分數，此外，TANDEM 系統(Hermansky, Ellis & Sharma, 2000)訓練 DNN 產生語音特徵的後驗機率，並將其作為額外的資訊去訓練傳統的聲學模型，研究中也指出此方法可使模型具有更好的強健性，同樣地，瓶頸特徵(bottleneck feature)技術(Grezl, Karafiat, Kontar & Cernocky, 2007)擷取了 ANN 的中間特徵作為傳統聲學模型的輸入，也可以有效地提高其辨識準確度。

特別一提的是，由於語音強化或是嘈雜語音特徵的映射(Han, He, Bagchi, Fosler-Lussier & Wang, 2015)旨在將雜訊語音訊號或是其特徵轉換回受雜訊干擾前的原始值，這是機器/深度學習中典型的回歸(regression)問題，因此，在其方法中 DNN 的訓練經常使用均方誤差(mean squared error, MSE)作為損失函數、藉由其最小化來學習 DNN 模型參數。然而，在評估方法的性能時，通常會使用其他一些客觀的指標，例如語音品質的感知評估(perceptual evaluation of speech quality, PESQ) (Rix, Beerends, Hollier & Hekstra, 2001)、短時客觀能辨度(short-time objective intelligibility, STOI) (Taal, Hendriks, Heusdens & Jensen, 2010)或詞錯誤率(word error rate, WER)。這些評估分數不一定與還原後的語音和原始語音之間的均方誤差(MSE)有直接的相關，亦即 DNN 訓練目標與評估指標並不一致，因此降低 MSE 未必可直接提升這些評估分數。有鑑於此，在一些近年開發的基於深度學習的語音強化法中(Zhang, Zhang & Gao, 2018)，直接將 PESQ 和 STOI 作為 DNN 模型訓練的目標函數、加以最佳化，而獲得更好的效能。

受上述觀察和其他文獻的啟發(Fu, Liao, Tsao & Lin, 2019; Xia & Bao, 2014)，本研究提出一種基於深度學習模型之強健性特徵擷取的新方法，其利用了與 MSE 無關的目標函數來訓練其中的深度網路。在此新方法中，我們使用的目標函數是給定語音特徵序列下，對應的聲學模型其中狀態序列(state sequence)與真實狀態序列相較之下的精確度，與語音識別的精確度有直接的相關。簡而言之，我們訓練一個深度神經網絡來進行語音特徵映

射，用以最大化地提高語音辨識系統中後端聲學模型狀態的後驗機率(posterior probability)。我們預期得到的新語音特徵在辨識準確度方面將優於原始特徵，並且具有對雜訊的強健性。

在以下章節中，我們介紹新提出的語音特徵擷取方法，並探討它的特點與可能的優勢。然後進行實驗與分析結果。而後以結論作終。

2. 基於最大化狀態精確度的語音特徵擷取法 (Proposed Method: Feature Extraction based on Maximizing State Accuracy)

在本研究中，我們提出了一種使用深度學習之強健語音特徵擷取法，這種方法目的在於增強原始語音辨識系統中，聲學特徵的抗噪能力，但毋需更改其後所使用的聲學模型。該方法在訓練其中的深度模型時，所使用的目標函數與辨識準確度有直接的相關，簡而言之，我們提出的方法的主要想法是找到在雜訊干擾環境中的聲學語音特徵序列$\bar{O}$，從而使後端的隱藏式馬可夫聲學模型 (hidden Markov model, HMM)對應的狀態序列 $\bar{q}'$ 相較於真實序列 $\bar{q}$（其由乾淨語音特徵序列對應之 HMM 的狀態序列）的相似度（精確度）最大化。該方法的流程圖如圖 1 所示，它包括以下步驟：

步驟 1:

對乾淨狀態(clean-condition)語音與訓練集與多條件狀態(multi-condition)的訓練集中的每個句子計算其 MFCC 與 FBANK 特徵序列，之後對於這些特徵序列使用平均值與變異數正規化法(Viikki & Laurila, 1998)加以處理，這些特徵以$\{\bar{o}_t\}$表示，其中t是音框的索引。

步驟 2:

利用訓練集中的 MFCC 特徵，透過 Kaldi (Povey *et al*., 2011)所提供的標準程序訓練單音素(monophone)與三連音素(triphone)的高斯混合(Gaussian mixture model) —隱藏式馬可夫的聲學模型(GMM-HMM)。值得一提的是，在訓練過程中，線性鑑別分析(linear discriminant analysis, LDA) (Haeb-Umbach & Ney, 1992)、最大相似度線性變換(Gales, 1998)和語者自適應訓練(speaker adaptive training, SAT) (Anastasakos, McDonough, Schwartz & Makhoul, 1996)幾種演算法在聲學模型訓練期間也應用於語音特徵上。在訓練後，我們得到訓練集中每個特徵序列根據 GMM-HMM 所對齊(aligned)之狀態序列 $\bar{q}$，此狀態序列包括單音標籤和三音標籤。

步驟 3:

藉由步驟 1 所得之訓練集的 FBANK 特徵以及步驟 2 所得之特徵對應的聲學模型狀態序列標籤，我們訓練相應的 DNN-HMM 聲學模型(Hinton *et al*., 2012)，其目標是找到語音特徵$\bar{o}_t$及其對應的狀態標籤q_t之間的轉換，因此這是一個多元分類的問題，在 DNN 的最後一層可以產生每個特徵$\bar{o}_t$的狀態觀察機率。值得注意的是，我們是使用多條件狀態(multi-condition)的訓練集來訓練聲學模型，其語音訊號摻雜了不同種類與訊雜比(signal-to-noise ratio, SNR)的雜訊，因此，預期產生的 DNN-HMM 會比使用乾淨狀態的訓練集對應的聲學模型具有更好的抗噪能力。

此外，我們使用乾淨狀態的訓練集特徵（與多狀態訓練集有相同的原始乾淨語句內容），另外訓練一套 DNN-HMM，藉此 DNN-HMM，我們為每句語音特徵求其對應的狀態序列$\bar{q}$，我們把它視為真實狀態序列(ground-truth state sequence)，因為它是由乾淨語音求得，沒有雜訊干擾。

步驟 4:

此步驟是我們方法的核心。我們訓練一個去噪深度神經網路(denoising network)，用來將原始語音特徵$\bar{o}_t$ 轉換為$\bar{o}_t'$，如下所示：

$$\bar{o}_t' = f_{DN}(\bar{o}_t) \tag{1}$$

其中$f_{DN}(.)$表示欲訓練之去噪網路函數，其訓練目標是使新的特徵$\bar{o}_t'$在步驟 3 中創建的 DNN-HMM 裡，可以預測出更接近真實狀態的聲學模型狀態序列，其數學式如下：

$$f_{DN} = \text{argmax}_f(\text{Acc}(\bar{q}, \bar{q}' = g(f(\bar{o}_t) \mid \lambda)) \tag{2}$$

其中λ是 DNN-HMM 聲學模型、g 是一個給定模型λ的一個函數，用來產生新特徵$f(\bar{o}_t)$對應的最高相似度(maximum likelihood)狀態序列 $\bar{q}'$ 、$\bar{q}$ 是由前一步驟所得之真實狀態序列(ground-truth state sequence)、Acc是對數相似度(log-likelihood)函數，用於評估 $\bar{q}'$ 相對於$\bar{q}$ 的精確度。

當訓練好去噪深度神經網路f_{DN}後，在辨識過程中，我們將其用於雜訊干擾的語音特徵$\bar{o}_t$映射到新特徵$\bar{o}_t'$，然後將$\bar{o}_t'$輸入至原本 (無須重新訓練) DNN-HMM 聲學模型與語言模型、分別生成最高相似度狀態序列與詞序列。與原始特徵 $\bar{o}_t$ 相比，新特徵$\bar{o}_t'$預期有更強的抗噪能力，因為它是在多條件訓練的 DNN-HMM 幫助下創建的，並有乾淨訓練的 DNN-HMM 所提供的真實狀態，相當於整合了雜訊環境對映至乾淨狀態的資訊；同時，由於新特徵 $\bar{o}_t'$所對應的狀態序列，相較於原始 $\bar{o}_t$ 而言應會具有較高的狀態精確率，因此它們在辨識中理應產生較低的詞錯誤率。

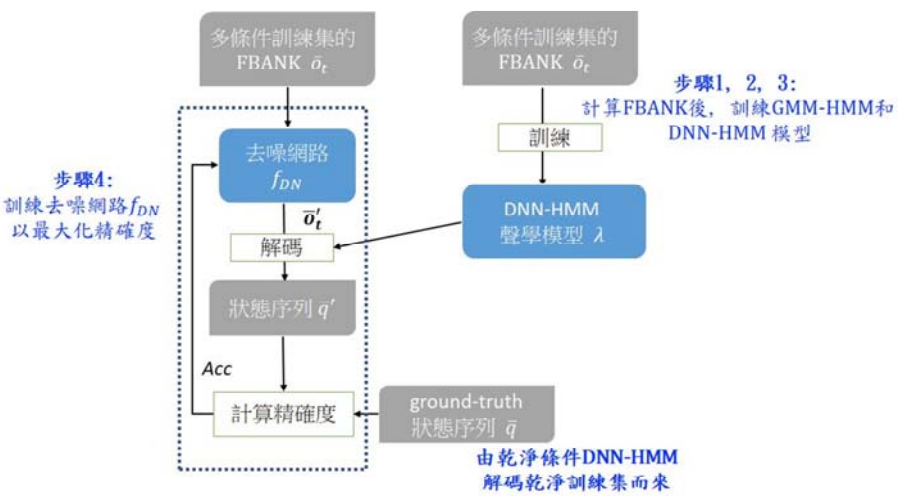


圖 1. 所提方法之流程圖

[Figure 1. The flowchart of the presented method]

與使用平均平方誤差(mean squared error, MSE)作為損失函數的之 DNN 求取抗噪之語音特徵方法相比(Garofolo, Lamel, Fisher, Fiscus & Pallett, 1993)，我們提出的方法具有以下潛在優勢：

1. 我們的方法其中的抗噪網路其訓練目標是在雜訊環境下，最大化聲學模型的狀態精確率，進而直接降低語音辨識的詞錯誤率(word error rate, WER)。相對而言，直接最小化訓練集內雜訊語音特徵與原始乾淨語音特徵之間的平方誤差(MSE)的 DNN 抗噪網路，可能存在如前一章節討論的目標不匹配問題，並不能保證在測試的雜訊環境下降低辨識錯誤率。
2. 我們的方法採用多條件訓練集依序獲得 GMM-HMM 和 DNN-HMM 聲學模型，藉此訓練我們的降噪神經網路。但是，我們預期當使用*乾淨無雜訊的訓練集所訓練的聲學模型*時，此降噪神經網路仍然可以使輸入雜訊語音特徵之輸出，對應到較高辨識率。根本的原因是整個降噪架構要改善原始語音特徵來擬合後端的聲學模型。在後面的章節中，我們將會在評估實驗中觀察並討論這方面的結果。

3. 實驗設置 (Experimental Setup)

我們使用著名的 TIMIT 語料庫(Garofolo *et al*., 1993)進行實驗，TIMIT 包含來自不同性別和方言的美國英語使用者的語句，其標註語句對應的音素和詞彙序列，TIMIT 中的語句在我們的評估實驗中用作訓練和測試集。在訓練語句上，除非特別提及之例外，我們是使用多條件狀態(multi-condition)的訓練集，先任意挑選 1000 句乾淨語音、再個別摻入不同種類及不同訊雜比(signal-to-noise ratio, SNR)的雜訊干擾，雜訊有三種，分別為：Babble、Car、Street，而訊噪比有五個等級，分別為 -5 dB、0 dB、5 dB、10 dB 和 15 dB，因此訓練集共有 15,000 句語音。在測試語句上，選擇了與訓練集不同的 400 句乾淨語音，再分別摻入三類別(White, Engine 和 Jackhammer)及六種訊雜比(-6 dB, -3 dB, 0 dB, 3 dB, 6 dB 和 12 dB)的雜訊干擾，共有 7,200 句語音。

訓練集中的語音特徵用於訓練上下文相關(context-dependent)的三連音素(tri-phone)聲學模型，它們被訓練成兩種不同的結構，分別是 GMM-HMM 和 DNN-HMM，具體來說，GMM-HMM 和 DNN-HMM 分別是使用 GMM 和 DNN 表示 HMM 的各種狀態。對於 GMM-HMM，每個單音素(monophone)的語音訊號和靜音分別由具有 3 個狀態的 HMM(總共 1000 個 Gaussian)來表示，而每個三連音素由具有 3 個狀態的 HMM 來表示，總共 2500 個 leaves。總共有 15000 個 Gaussian。此外，在三連音素模型訓練期間，將 LDA、MLLT 和 SAT 應用於語音特徵。另一方面，對於 DNN 的結構，使用了 5 層隱藏層，每個隱藏層包含 1024 個節點，並且分別連接到 DNN-HMM 中用於三連音和單音的兩個獨立輸出層。此 DNN 的訓練使用 Dropout 法，比例為 15%，進行 24 個 epochs 和使用 SGD 優化器，並且使用對數相似度(log-likelihood)作為目標函數。在模型訓練中，會將三連音素和單音素的誤差相加，並將其最小化。我們使用 Kaldi 工具包(Povey *et al*., 2011)來創建 GMM-HMM，而 Pytorch-Kaldi (Ravanelli, Parcollet & Bengio, 2019)工具包則用於創建 DNN-HMM。

此外，透過 Kaldi 的標準程序，我們建構了一組用於訓練語音的三元語法的語言模型(tri-gram)。

對於訓練和測試集中的每個語句，我們使用 69 維的 FBANK 特徵（每個音框 23 維的 FBANK 以及其 delta 和 delta-delta，音框長度為 20 毫秒，每次位移 10 毫秒）來作為基礎特徵(baseline feature)。我們提出的降噪 DNN 框架將 FBANK 作為輸入，按照上個章節中的步驟產生新的特徵，以進行後續辨識。去噪 DNN 模型是一個卷積神經網絡(convolutional network)，具有 4 個相同尺寸的一維卷積層，每層 kernel 數為 30，kernel 大小為 5，padding 數為 2。此外，這四個卷積層後面接著兩個相同的全連接層，各層具有 759 個節點。每層輸出的激活函數是線性整流函數(ReLU)。該降噪框架的訓練過程使用 Adam 優化器進行了 30 個 epochs，並使用了對數相似度(log-likelihood)作為目標函數。

4. 實驗結果與討論 (Experimental Results and Discussions)

在本章節中，我們呈現實驗結果並加以討論，為了方便討論，我們將所提出的方法命名為"最大化聲學模型狀態精確率法"，英文為"maximum state accuracy"，以縮寫"MSA"表示，同時，我們使用了兩種語音強化法進行比較，分別為最小均方誤差短時頻譜強度估測(minimum mean-square error short-time spectral amplitude estimation, 縮寫為 MMSE-STSA) (Ephraim & Malah, 1984)，及理想比例遮罩法(ideal ratio masking, 縮寫為 IRM) (Wang, 2005)。此外，我們採用一種基於 DNN 求取聲學特徵轉換的方法(Han *et al*., 2015)進行比較，該方法主要使用深度神經網絡(DNN)來轉換輸入的 FBANK 特徵，透過直接最小化多條件訓練集中雜訊—乾淨配對(noisy-clean pair)之語音的 FBANK 特徵之間的均方誤差(mean squared error, MSE)來學習其 DNN，這種方法稱為 feature-based MSE，縮寫為 FMSE。

在這裡，我們的實驗結果分為兩部分，分別為多條件訓練模式(multi-condition training mode)和乾淨狀態訓練模式(clean-condition training mode)。值得注意的是，我們所提出的方法 MSA 中的降噪模型在兩個模式下皆是藉由多條件訓練集所學習而得，但我們想測試由此產生的強化後語音特徵在兩種模式下可否都能表現良好。

- 多條件訓練模式(multi-condition training mode)之結果與討論

當利用多條件訓練集所訓練而得的聲學模型時，表 1、表 2 與表 3 列出了我們的方法 MSA 及三種比較法 FMSE、MMSE-STSA 與 IRM 在三種雜訊測試集所得之詞錯誤率(word error rate, WER)，值得注意的是，MMSE-STSA 與 IRM 二種語音強化法只使用於測試集的語句，並未作用於訓練集之語句，原因是我們之前實驗發現，若它們同時作用於訓練集，將使測試集之詞錯誤率明顯增加。從這三個表中，我們有以下觀察結果：

1. 平均而言，各種方法在 Jackhammer 雜訊環境中得到的詞錯誤率明顯低於 White 和 Engine 雜訊環境中的 WER，這表明與 White 和 Engine 雜訊相比，Jackhammer 雜訊對語音訊號的失真較小。但是，我們發現所有的方法均無法改善 Jackhammer 雜訊下的基礎實驗結果(baseline)，這表明語音強化或強健性特徵方法可能會對干擾較少的語句引

入更多可觀察到的失真。

2. 對於 MMSE-STSA 與 IRM 兩種語音強化法而言，MMSE-STSA 效果明顯比 IRM 差，且比基礎實驗得到較高的詞錯誤率，而 IRM 法相較於基礎實驗結果而言，只能在部分 SNR 環境小幅較低詞錯誤率，此結果部分驗證了語音強化方法雖可改善語音品質，但未必能有效提升語音辨識精確度。
3. 在 White 和 Engine 雜訊環境中，新提出的 MSA 法在大多數 SNR 環境下可以得到較低的詞錯誤率，並且更勝過其他方法，這驗證了 MSA 藉由提高語音特徵之狀態精確度，可改善特徵對雜訊的強健性並增加辨識準確率。特別的是，新提出的 MSA 中的降噪網路，其訓練所使用之語音包含的雜訊種類並非是測試集之 White 雜訊與 Engine 雜訊。因此，MSA 在某種程度上顯示出其一般化(generalization)的能力，在未知雜訊(unseen noise)環境下，仍可提升語音特徵的強健性。
4. FMSE 方法是為了最小化雜訊語音和乾淨語音其 FBANK 特徵之間的均方誤差(MSE)，然而在幾乎所有雜訊情況下，其效果都比基礎實驗結果差，如同之前討論，其可能原因是其評估與優化指標的不匹配，造成 FMSE 方法轉換後的語音特徵反而產生較差的辨識準確率，另一個原因是，在 FMSE 中學習到的 DNN 過度擬合訓練資料，因此無法很好地改善測試資料的失真問題。

表 1. 多條件訓練模式在“White”雜訊環境下測試集的基礎實驗、MSA、FMSE、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 1. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the White noise-corrupted test set under the multi-condition-training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	66.1	62.1	57.0	49.8	44.8	34.4
MSA	65.5	61.0	54.9	48.9	43.2	34.7*
FMSE	69.2*	63.3*	57.2*	50.4*	44.8	35.3*
MMSE-STSA	70.3*	66.8*	60.4*	55.0*	49.7*	40.1*
IRM	65.8	61.9	56.5	50.4*	44.4	34.3

表2. 多條件訓練模式在 "Engine" 雜訊環境下測試集的基礎實驗、MSA、FMSE 、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 2. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Engine noise-corrupted test set under the multi-condition-training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	65.3	61.7	55.1	48.2	41.5	31.1
MSA	65.5*	60.2	54.6	47.9	41.7*	32.3*
FMSE	70.4*	64.5*	56.2*	48.2	41.0	31.2*
MMSE-STSA	68.3*	63.6*	57.3*	51.4*	44.9*	34.2*
IRM	65.9*	61.4	54.8	48.2	41.2	31.1

表 3. 多條件訓練模式在"Jackhammer"雜訊環境下測試集的基礎實驗、MSA 、FMSE 、 MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 3. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Jackhammer noise-corrupted test set under the multi-condition-training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	31.4	27.8	25.9	23.8	23.0	21.9
MSA	32.6*	29.4*	27.5*	25.7*	25.1*	23.9*
FMSE	34.4*	31.3*	29.1*	27.6*	27.0*	26.1*
MMSE-STSA	32.9*	29.0*	27.2*	25.2*	24.4*	23.2*
IRM	32.4*	29.8*	26.6*	25.2*	23.8*	22.8*

- 乾淨訓練模式(clean-condition training mode)之結果與討論

利用乾淨訓練集所訓練而得的聲學模型，表 4、表 5 與表 6 列出了我們的方法 MSA 及三種比較法 FMSE、MMSE-STSA 與 IRM 在三種雜訊測試集所得之詞錯誤率(word error rate, WER)，值得注意的是，由於訓練集是乾淨語音，我們並不使用任何方法對其特徵作進一步強化，意即我們使用乾淨訓練集中的原始 FBANK 特徵來訓練聲學模型，而各個方法只用在測試集上。從這三個表，我們有以下觀察：

1. 若與前三個表(表 1、2、3)相比，乾淨訓練模式得到的基礎實驗(baseline)結果比多條件訓練模式所得到的基礎實驗結果較差(前者得到較高的詞錯誤率)，這很可能是因為與多條件訓練語句相比，乾淨訓練語句與測試集的雜訊語句之間之不匹配更為明顯。
2. 類似於之前的觀察，相較於基礎實驗結果，兩種語音強化法 (MMSE-STSA 和 IRM) 只能得到相近或更高的詞錯誤率，這再次顯示了直接改善語音的品質並不一定能提高其

辨識準確率。

3. 對於大多數雜訊之狀態(除了 SNR 高於-3 dB 的 Jackhammer 雜訊環境)，我們所提出的 MSA 法相較於基礎實驗結果，獲得明顯較低的詞錯誤率，這些結果表明，即使訓練集特徵未經 MSA 處理，若測試語句特徵經過 MSA 法處理後，仍可改善其語音辨識精確度。我們認為，這再次證實了我們先前的陳述，即 MSA 具有一般化的能力，可克服未見雜訊之問題。
4. 作用在特徵上的 FMSE 法，在部分雜訊環境下能比基礎實驗結果表現較佳(即得到較低的詞錯誤率)，但其效果仍不及我們所新提出的 MSA 法。

表 4. 乾淨訓練模式在"White"雜訊環境下測試集的基礎實驗、MSA 、FMSE 、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 4. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the White noise-corrupted test set under the clean-condition training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	67.6	64.6	61.0	55.6	50.9	41.2
MSA	64.3	60.6	56.3	50.8	45.4	36.6
FMSE	68.6*	64.3	58.9	53.0	47.7	38.9
MMSE-STSA	69.9*	67.1*	63.7*	59.1*	54.4*	45.6*
IRM	67.4	64.3	60.6	55.0	50.5	40.7

表 5. 乾淨訓練模式在"Engine"雜訊環境下測試集的基礎實驗、MSA 、FMSE 、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 5. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Engine noise-corrupted test set under the clean-condition training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	67.3	64.7	61.1	56.3	50.4	39.4
MSA	64.4	60.9	55.2	49.8	43.8	34.7
FMSE	69.9*	65.5*	59.7	52.3	46.8	36.6
MMSE-STSA	69.1*	65.9*	62.5*	57.3*	52.1*	40.7*
IRM	66.8	65.1*	60.6	55.7	49.7	39.4

表6. 乾淨訓練模式在"Jackhammer"雜訊環境下測試集的基礎實驗、MSA、FMSE、MMSE-STSA 和 IRM 所得的詞錯誤率(WER, %)
[Table 6. Word error rates (WER, %) achieved by different methods (baseline, MSA, FMSE, MMSE-STSA and IRM) for the Jackhammer noise-corrupted test set under the clean-condition training mode]

	Signal-to-noise ratio (SNR)					
	-6 dB	-3 dB	0 dB	3 dB	6 dB	12 dB
baseline	35.3	31.5	28.5	26.7	24.9	23.1
MSA	33.8	30.6	28.7*	27.0*	26.3*	24.9*
FMSE	35.0	32.2*	29.7*	28.1*	27.0*	25.4*
MMSE-STSA	36.5*	32.8*	29.6*	27.7*	25.3*	23.7*
IRM	34.8	31.8*	28.8*	26.6	24.9	23.3*

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

在本研究中，我們主要關注在自動語音辨識中的雜訊問題，提出一種基於深度學習的新方法來建立抗噪語音特徵，該方法利用深度神經網路來最大化語音特徵所對應的聲學模型的狀態精確度。初步實驗表明，新提出的方法可以提高 FBANK 特徵的辨識準確率，特別是在中度和重度雜訊干擾狀態中；且無論聲學模型的訓練集是在多條件環境下或是乾淨環境，它都能表現良好。關於未來的改良方向，我們將透過採用更多種類的訓練數據或增加訓練資料量來進一步增強此降噪神經網路，然後將其與其他基於特徵或基於模型的雜訊強健性演算法組合，以實現更好的性能。

參考文獻 (References)

Anastasakos, T., McDonough, J., Schwartz, R., & Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proceedings of Fourth International Conference on Spoken Language Processing (ICSLP) 1996*. doi : 10.1109/ICSLP.1996.607807

Bourlard, H. & Morgan, N. (1994). *Connectionist Speech Recognition: A Hybrid Approach*. New York, NY: Springer.

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans on Acoustics, Speech, and Signal Processing*, *32*(6), 1109-1121. doi: 10.1109/TASSP.1984.1164453

Fu, S.-W., Liao, C.-F., Tsao, Y., & Lin, S.-D. (2019). MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *Proceedings of the 36th International Conference on Machine Learning 2019*, 2031-2041.

Gales, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, *12*(2), 75-98. doi: 10.1006/csla.1998.0043

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., & Pallett, D. (1993). *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1* (NASA STI/Recon Technical Report N, vol. 93, p. 27403).Retrieved from https://ui.adsabs.harvard.edu/abs/1993STIN...9327403G

Grezl, F., Karafiat, M., Kontar, S., & Cernocky, J. (2007). Probabilistic and bottleneck features for lvcsr of meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2007*. doi: 10.1109/ICASSP.2007.367023

Haeb-Umbach, R. & Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 1992*. doi : 10.1109/ICASSP.1992.225984

Han, K., He, Y., Bagchi, D., Fosler-Lussier, E., & Wang, D. (2015). Deep neural network based spectral feature mapping for robust speech recognition. In *Proceedings of INTERSPEECH 2015*, 2484-2488

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738-1752. doi: 10.1121/1.399423

Hermansky, H., Ellis, D. P. W., & Sharma, S. (2000). TANDEM connectionist feature extraction for conventional hmm systems. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2000*. doi: 10.1109/ICASSP.2000.862024

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 578-589. doi: 10.1109/89.326616

Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1991). Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proceedings of EUROSPEECH 1991*, 1367-1370.

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

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

Kuo, J.-W. & Chen, B. (2005). Minimum word error based discriminative training of language models. In P*roceedings of Interspeech'2005 - Eurospeech*, 1277-1280.

Liu, F. H., Stern, R. M., Huang, X., & Acero, A. (1993). Efficient cepstral normalization for robust speech recognition. In *Proceedings of the workshop on Human Language Technology (HLT '93)*, 69-74. doi: 10.3115/1075671.1075688

Povey, D. (2003). *Discriminative training for large vocabulary speech recognition* (Doctoral dissertation). University of Cambridge, UK.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. … Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding 2011*.

Ravanelli, M., Parcollet, T., & Bengio, Y. (2019). The PyTorch-Kaldi speech recognition toolkit. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. doi: 10.1109/ICASSP.2019.8683713

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*. doi: 10.1109/ICASSP.2001.941023

Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., & Venkataraman, A. (2005). MLLR transforms as features in speaker recognition. In *Proceedings of Eurospeech 2005*, 2425-2428.

Su, Y.-C., Tsao, Y., Wu, J.-E., & Jean, F.-R. (2013). Speech enhancement using generalized maximum a posteriori spectral amplitude estimator. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2013*. doi: 10.1109/ICASSP.2013.6639114

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2010*. doi: 10.1109/ICASSP.2010.5495701

Torre, A., Peinado, A., Segura, J., Pérez-Córdoba, J., Benitez, C., & Rubio, A. (2005). Histogram equalization of the speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *13*(3), 355-366. doi: 10.1109/TSA.2005.845805

Viikki, O. & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, *25*(1-3), 133-147. doi: 10.1016/S0167-6393(98)00033-8

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi P. (eds) *Speech Separation by Humans and Machines* (pp. 181-197), Springer, Boston, MA. https://doi.org/10.1007/0-387-22794-6_12

Xia, B. & Bao, C.-c. (2014). Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Communication*, *60*, 13-29. doi: 10.1016/j.specom.2014.02.001

Zhang, H., Zhang, X., & Gao, G. (2018). Training supervised speech separation system to improve STOI and PESQ directly. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2018*. doi: 10.1109/ICASSP.2018.8461965

The individuals listed below are reviewers of this journal during the year of 2020. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Guo-Wei Bian
Jing-Shin Chang
Yung-Chun Chang
Tao-Hsing Chang
Yu-Yun Chang
Ru-Yng Chang
Fei Chen
Alvin C.-H. Chen
Chung-Chi Chen
Kuan-Yu Chen
Mei-Hua Chen
Yun-Nung (Vivian) Chen
Tai-Shih Chi
Jia-Fei Hong
Shu-Kai Hsieh
Chun-Hsien Hsu
Hen-Hsen Huang
Yi-Chin Huang
Jeih-Weih Hung
Hung-Yu Kao
Wen-Hsing Lai
Ying-Hui Lai
Lung-Hao Lee
Hong-Yi Lee
Chuan-Jie Lin
Shu-Yen Lin
Yi-Fen Liu
Shih-Hung Liu
Ming-Hsiang Su
Wei-Ho Tsai
Jenq-Haur Wang
Syu-Siang Wang
Hsin-Min Wang
Jiun-Shiung Wu
Jheng-Long Wu
Cheng-Zen Yang
Yi-Hsuan Yang
Jui-Feng Yeh

2020 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 25

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2020.

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

AUTHOR INDEX

SUBJECT INDEX

S

T

W

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

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

Activities：

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

To Register：

Please send application to:

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

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

Annual Fees：

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

Contact：

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

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

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

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

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： _____

Name： _____ Date of Birth： _____

Country of Residence： _____ Province/State： _____

Passport No.： _____ Sex: _____

Education(highest degree obtained)： _____

Work Experience： _____

_____

Present Occupation： _____

Address： _____

_____

Email Add： _____

Tel. No： _____ Fax No： _____

Membership Category： ☐ Regular Member ☐ Life Member

Date： ____/____/____（Y-M-D）

Applicant's Signature：

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

Regular Member ： US$ 50.- （NT$ 1,000）

Life Member ： US$500.-（NT$10,000）

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

Committee Assessment：

中華民國計算語言學學會

宗旨：

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

活動項目：

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

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會
2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

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

連絡處：

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

中華民國計算語言學學會
個人會員入會申請書

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

中　華　民　國　　年　　月　　日</td></tr>
</table>

審查結果:

1. 年費：

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

2. 連絡處：

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

3. 本表可自行影印

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

Name: _____(Please print) Date: _____

Please debit my credit card as follows: US$ _____

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

Card No.: _____ -_____-_____ -_____ Exp. Date:_____(M/Y)

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

CARD HOLDER SIGNATURE: _____

Phone No.: _____E-mail: _____

Address: _____

PAYMENT FOR

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

Quantity Wanted: _____

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

Quantity Wanted: _____

US$ ____ ❑ Publications:_____

US$ ____ ❑ Text Corpora: _____

US$ ____ ❑ Speech Corpora:_____

US$ ____ ❑ Others: _____

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

US$ _____ = Total

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

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

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: _____(請以正楷書寫)　　日期:：_____

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

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

卡片後三碼：_____（卡片背面簽名欄上數字後三碼）

持卡人簽名：_____(簽名方式請與信用卡背面相同)

通訊地址：_____

聯絡電話：_____E-mail：_____

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$_____ ❑ 中文計算語言學期刊(IJCLCLP) _____

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

NT$_____ ❑ 中研院詞庫小組技術報告_____

NT$_____ ❑ 文字語料庫 _____

NT$_____ ❑ 語音資料庫 _____

NT$_____ ❑ 光華雜誌語料庫1976~2010

NT$_____ ❑ 中文資訊檢索標竿測試集/文件集

NT$_____ ❑ 會員年費：❑續會　❑新會員　❑終身會員

NT$_____ ❑ 其他: _____

NT$_____ = 合計

填妥後請傳真至 02-27881638 或郵寄至:

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

Publications of the Association for Computational Linguistics and Chinese Language Processing

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

10% member discount: _____Total Due:_____

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

Name (please print): _____ Signature: _____

Fax: _____ E-mail: _____

Address：_____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表 (甲)	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義（中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊 (一年兩期)　年份：_____（過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內（台灣地區）使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人：　黃琪 小姐、何婉如 小姐　　E-mail:aclclp@aclclp.org.tw

訂購者：_____ 收據抬頭：_____

地　　址：_____

電　　話：_____ E-mail: _____

Information for Authors

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

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

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

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

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

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

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

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

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

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

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

Here shows an example.

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

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

Please visit the following websites for details.

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

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

No page charges are levied on authors or their institutions.

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

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

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

International Journal of
Computational Linguistics & Chinese Language Processing
vol.25 no.2 December 2020

Contents

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing