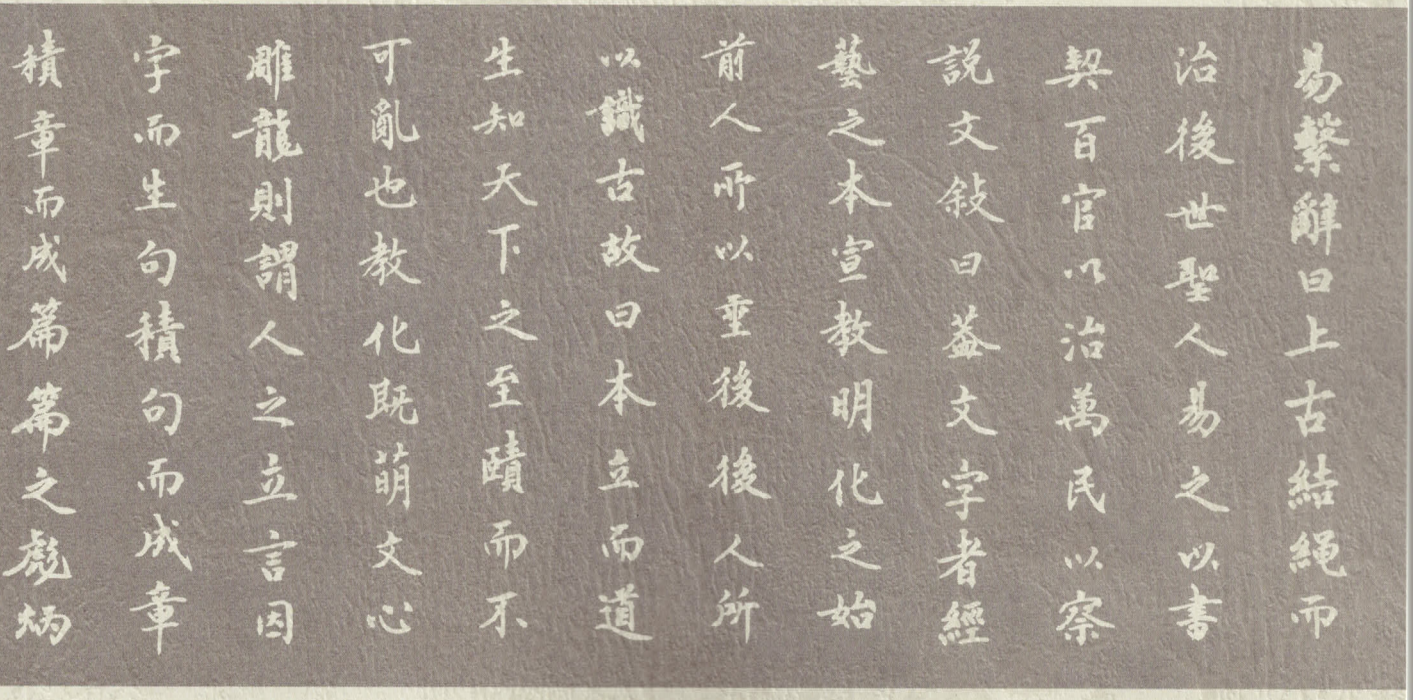
International Journal of

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.1 June 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. 1, June 2020

Contents

Papers

Computational Linguistics and Chinese Language Processing
Vol. 25, No. 1, June 2020, pp. 1-28

Chinese Spelling Check based on Neural Machine Translation

Jhih-Jie Chen*, Hai-Lun Tu+, Ching-Yu Yang*,

Chiao-Wen Li# and Jason S. Chang*

Abstract

We present a method for Chinese spelling check that automatically learns to correct a sentence with potential spelling errors. In our approach, a character-based neural machine translation (NMT) model is trained to translate the potentially misspelled sentence into correct one, using right-and-wrong sentence pairs from newspaper edit logs and artificially generated data. The method involves extracting sentences contain edit of spelling correction from edit logs, using commonly confused right-and-wrong word pairs to generate artificial right-and-wrong sentence pairs in order to expand our training data , and training the NMT model. The evaluation on the United Daily News (UDN) Edit Logs and SIGHAN-7 Shared Task shows that adding artificial error data can significantly improve the performance of Chinese spelling check system.

Keywords: Chinese Spelling Check, Artificial Error Generation, Neural Machine Translation, Edit Log

1. Introduction

Spelling check is a common yet important task in natural language processing. It plays an important role in a wide range of applications such as word processors, assisted writing systems, and search engines. For example, search engine without spelling check is not user-friendly, while assisted writing system must perform spelling check as the minimal requirement. Web search engines such as Google (www.google.com) and Bing

* Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan
+Department of Library and Information Science, Research and Development Center for Physical Education, Health, and Information Technology, Fu Jen Catholic University, New Taipei, Taiwan
Department of Information System and Application, National Tsing-Hua University, Hsinchu, Taiwan
E-mail: {jjc, helentu, chingyu, chiaowen, jason}@nlplab.cc

(www.bing.com) typically perform spelling check on queries, in order to retrieve documents better meeting the user information need. The users' queries would be corrected first by the spelling check component in order to avoid irrelevant or low-quality search results. In contrast to Web search engines, while Microsoft Word has a very effective spelling checker for English, there is still considerable room to improve the one for Chinese.

Consider a sentence "他在文學方面有很高的造旨。" ('He is highly accomplished in literature.'). In the context of this sentence, the character "旨" (pronounced 'zhi') is a typo. For another sentence "他在文學方面有很高的造藝。", the character "藝" (pronounced 'yi') is also a typo. For these two typos, the correct character is "詣" (pronounced 'yi'). Chinese spelling errors are due to two main reasons: one is similar sound (e.g., *藝 and 詣) and the other is similar shape (e.g., *旨 and 詣), as pointed by Liu *et al*. (2011).

Unfortunately, such spelling error is probably uneasy to correct due to limited training data. In fact, there is a lack of training data for the Chinese spelling check task. Compared to western languages (e.g., English and German), relatively little work has been done on Chinese spelling check and few datasets are available. More spelling errors can be corrected with a machine learning model trained on more data. It could be that there are some fundamental problems such as no word boundaries, too many characters, and inconsistent use along time. Chinese spelling check could be more practical if more training data is available.

One solution to the lack of training data is to create artificial one for training. Researches on artificial error generation for English have shown great potential in improving underlying models for writing error correction (Felice & Yuan, 2014; Rei, Felice, Yuan, & Briscoe, 2017). In other words, by generating artificial errors to increase data, we might have a chance to make spelling check models better and stronger. However, very few works have focused on generating artificial errors for Chinese.

In this paper, we present *AccuSpell*, a system that automatically learns to generate the corrected sentence for a potentially misspelled sentence using neural machine translation (NMT) model. The system is built on a new dataset consisting of edit logs of journalists from the United Daily News (UDN). Moreover, we collect a number of confusion set for generating artificial errors to augment the data for training. The evaluation on the UDN Edit Logs and SIGHAN-7 Shared Task shows that adding artificial error data can significantly improve the performance of Chinese spelling check system. The model is deployed on Web and an example *AccuSpell* searches for the sentence "今晚月色很美，我想小灼一杯。" ('The moon is so beautiful tonight, and I want a drink.') is shown in Figure 1. *AccuSpell* has determined that "今晚月色很美，我想小酌一杯。" is the most probably corrected sentence. *AccuSpell* learns how to effectively correct a given sentence during training by using more data, including real edit logs and artificially generated data. We will describe how to

create artificial data and training process in detail in Section 3.

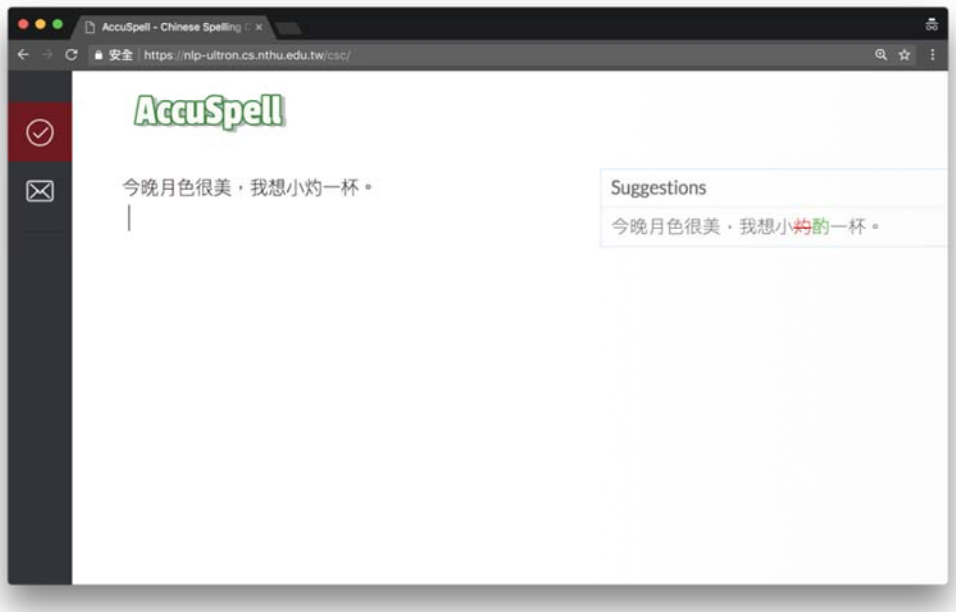


Figure 1. An example the Web version of AccuSpell searches for input "今晚月色很美，我想小灼一杯。" ('The moon is so beautiful tonight, and I want a drink.')

At run-time, *AccuSpell* starts with a sentence or paragraph submitted by the user (e.g., "今晚月色很美，我想小灼一杯。"), which was first divided into clauses. Each clause then is splitted into Chinese characters before being fed to the NMT model. Finally, the model outputs an n-best list of sentences. In our prototype, *AccuSpell* returns the best sentence to the user directly (see Figure 1); alternatively, the best sentence returned by *AccuSpell* can be passed on to other applications such as automatic essay rater and assisted writing systems.

The rest of the article is organized as follows. We review the related work in the next section. Then we describe how to extract the misspelled sentences from newspaper edit logs and how to generate artificial sentences with typos in Section 3. We also present our method for automatically learning to correct typos in a given sentence. Section 4 describes the resources and datasets we used in the experiment. In our evaluation, over two set of test data, we compare the performance of several models trained on both real and artificial data with the model trained on only real data in Section 5. Finally, we summarize and point out the future work in Section 6.

2. Related Work

Error Correction has been an area of active research, which involves Grammatical Error Correction (GEC) and Spelling Error Correction (SEC). Recently, researchers have begun

applying neural machine translation models to both GEC and SEC, and gained significant improvement (e.g., Yuan & Briscoe, 2016; Xie, Avati, Arivazhagan, Jurafsky, & Ng, 2016). However, compared to English, relatively little work has been done on Chinese error correction. In our work, we address the spelling error correction task, that focuses on generating corrections related to typos in Chinese text written by native speakers.

Early work on Chinese spelling check typically uses rule-based and statistical approaches. Rule-based approaches usually use dictionary to identify typos and confusion set to find possible corrections, while statistical methods use the noisy channel model to find candidates of correction for a typo and language model to calculate the likelihood of the corrected sentences. Chang (1995) proposed an approach that combines rule-based method and statistical method to automatically correct Chinese spelling errors. The approach involves confusing character substitution mechanism and bigram language model. They used a confusion set to replace each character in the given sentence with its corresponding confusing characters one by one, and use a bigram language model built from a newspaper corpus to score all modified sentences in an attempt to find the best corrected sentence. Zhang, Huang, Zhou, and Pan (2000) pointed out that Chang (1995)'s method can only address character substitution errors, other kinds of errors such as character deletion and insertion cannot be handled. They proposed an approach using confusing word substitution and trigram language model to extend the method proposed by Chang (1995).

In recent years, Statistical Machine Translation (SMT) has been applied to Chinese spelling check. Wu, Chen, Yang, Ku and Liu (2010) presented a system using a new error model and a common error template generation method to detect and correct Chinese character errors that can reduce false alarm rate significantly. The idea of error model is adopted from the noisy channel model, a framework of SMT, which is used in many NLP tasks such as spelling check and machine translation. Chiu, Wu and Chang (2013) proposed a data-driven method that detect and correct Chinese errors based on phrasal statistical machine translation framework. They used word segmentation and dictionary to detect possible spelling errors, and correct the errors by using SMT model built from a large corpus.

More recently, Neural Machine Translation (NMT) has been adopted in error correction task and has achieved state-of-the-art performance. Yuan and Briscoe (2016) presented the very first NMT model for grammatical error correction of English sentences and proposed a two-step approach to handle the rare word problem in NMT. The word-based NMT models usually suffer from rare word problem. Thus, a neural network-based approach using character-based model for language correction was proposed by Xie *et al*. (2016) to avoid the problem of out-of-vocabulary words. Chollampatt and Ng (2018) proposed a multilayer convolutional encoder-decoder neural network to correct grammatical, orthographic, and collocation errors. Until now, most work on error correction done by using NMT model aimed

at grammatical errors for English text. In contrast, we focus on correcting Chinese spelling errors.

Building an error correction system using machine learning techniques typically require a considerable amount of error-annotated data. Unfortunately, limited availability of error-annotated data is holding back progress in the area of automatic error correction. Felice and Yuan (2014) presented a method that generates artificial errors for correcting grammatical mistakes made by learners of English as a second language. They are the first to use linguistic information such as part-of-speech to refine the contexts of occurring errors and replicate them in native error-free text, but also restricting the method to five error types. Rei *et al*. (2017) investigated two alternative approaches for artificially generating all types of writing errors. They extracted error patterns from an annotated corpus and transplanting them into error-free text. In addition, they built a phrase-based SMT error generator to translate the grammatically correct text into incorrect one.

In a study closer to our work, Gu and Lang (2017) applied sequence-to-sequence (seq2seq) model to construct a word-based Chinese spelling error corrector. They established their own error corpus for training and evaluation by transplanting errors into an error-free news corpus. Comparing with traditional methods, their model can correct errors more effectively.

In contrast to the previous research in Chinese spelling check, we present a system that uses newspaper edit logs to train an NMT model for correcting typos in Chinese text. We also propose a method to generate artificial error data to enhance the NMT model. Additionally, to avoid rare word problem, our NMT model is trained at character level. The experiment results show that our model achieves significantly better performance, especially at an extremely low false alarm rate.

3. Methodology

Submitting a misspelled sentence (e.g., “今晚月色很美，我想小灼一杯。”) to a spelling check system with limited training data often does not work very well. Spelling check systems typically are trained on data of limited size and scope. Unfortunately, it is difficult to obtain a sufficiently large training set that cover most common errors, corrections, and contexts. When encountering new and unseen errors and contexts, these systems might not be able to correct such errors. To develop a more effective spelling check system, a promising approach is to automatically generate artificial errors in presumably correct sentences for expanding the training data, leading the system to cope with a wider variety of errors and contexts.

3.1 Problem Statement

We focus on correcting spelling errors in a given sentence by formulating the Chinese spelling check as a machine translation problem. A sentence with typos is treated as the source sentence, which is translated into a target sentence with errors corrected. The plausible target sentence predicted by a neural machine translation model is then returned as the output of the system. The returned sentence can be viewed by the users directly as suggestion for correcting a misspelled sentence, or passed on to other applications such as automatic essay rater and assisted writing systems. Thus, it is important that the misspelled characters in a given sentence be corrected as many as possible. At the same time, the system should avoid making false corrections. Therefore, our purpose is to return a sentence with most spelling errors corrected, while keeping false alarms reasonably low. We now formally state the problem that we are addressing.

***Problem Statement*:** We are given a possibly misspelled sentence X with n characters $x_1,x_2,...,x_n$. Our goal is to return the correctly spelled sentence Y with m characters $y_1,y_2,...,y_m$. For this, we prepare a dataset of right-and-wrong sentence pairs in order to train a neural machine translation (NMT) model. The sentences come from real edit logs and artificially-generated data.

In the rest of this section, we describe our solution to this problem. First, we describe the process of automatically learning to correct misspelled sentences in Section 3.2. More specifically, we describe the preprocessing of edit logs in Section 3.2.1, and how to artificially generate similar sentences with edits in Section 3.2.2. We then describe the process of training NMT model in Section 3.2.3. Finally, we show how *AccuSpell* corrects a given sentence at run-time by applying NMT model in Section 3.3.

3.2 Learning to Correct Misspelled Sentence

We attempt to train a neural machine translation (NMT) model using right-and-wrong sentence pairs from edit logs and artificial data, which to translate a misspelled sentence into a correct one. In this training process, we first extract the sentences with spelling errors from edit logs (Section 3.2.1) and generate artificial misspelled sentences from a set of error-free sentences (Section 3.2.2). We then use these data to train the NMT model (Section 3.2.3).

3.2.1 Extracting Misspelled Sentences from Edit Logs

In the first stage of training process, we extract a set of sentences with spelling errors annotated by simple edit tags (i.e., <[-, -]> for deletion and <{+, +} > for insertion). For example, the sentence “希望未來主要島嶼都有完善的[-馬-]{+碼+}頭，” (Hope that the main islands will have perfect docks in the future.) contains the edit tags “[-馬-]{+碼+}” that means the original character “馬” (pronounced ’ma’) was replaced with “碼”

(pronounced 'ma').

【記者葉子菁／台北報導】12月台指期合約將於明日結算，台指期今日開高後震盪走低，並回測9<FONT class=1 title=李定強新增, color=#265e8a>,</FONT>200點位置。永豐期貨副總廖祿民表示，台股目前屬於盤跌、慢慢走弱的盤勢，從外資在期貨淨多單的留倉來看，仍未有企圖撐在高點結算的意味，且適逢耶誕假期，外資也不急著<FONT style="TEXT-DECORATION: line-through" class=3 title=李定強刪除, color=#555588>佈</FONT><FONT class=1 title=李定強新增, color=#265e8a>布</FONT>局明年，惟從選擇權的Put/Call Ratio來看，1.4仍屬於多方架構，後續9<FONT class=1 title=李定強新增, color=#265e8a>,</FONT>200點為觀察支撐點位的基礎，而預期在12/5日的低<FONT class=1 title=李定強新增, color=#265e8a>點</FONT>9<FONT class=1 title=李定強新增, color=#265e8a>,</FONT>138點具有較強勁的支撐力道。</P>

Figure 2. An example of edit logs in HTML format

1. 食藥署**[-今-]{+昨+}**宣布將開放食鹽添加微量氟化物，
2. 現在「**[-b-]{+B+}**lue Monday變**[-g-]{+G+}**reen Monday了，
3. 記得拍照帶走美景就好**{+。+}[-；-]**如果時間許可，
4. **[-她-]{+他+}**昨日接受專訪時說明，
5. 參加世界盃拔河賽獲4金2銀的大笨牛夢想拔河隊教練陳**[-鵬-]{+建+}**文，
6. 最後再撒上適量起**{+司+}[-士-]**絲，
7. 這項計畫將持續募款到今年**[-聖-]{+耶+}**誕節，
8. 使得泰山今年**[-上-]{+下+}**半年獲利成長樂觀。
9. 饗蔬職人**[-除了提供全素（非奶蛋素）的新鮮食材外，-]**還用心烹調3種湯頭、7種醬料，
10. 價值上百萬的好禮**[-通通-]{+統統+}**帶回家。
11. 希望未來主要島嶼都有完善的**[-馬-]{+碼+}**頭，

Figure 3. Examples of different edit types in edit logs

The input to this stage are a set of edit logs in HTML format, containing the name of editor, the action of edit (1 is insertion and 3 is deletion), the target content and some CSS attributes, as shown in Figure 2. We first convert HTML files to simple text files by removing HTML tags and using simple edit tags "{+ +}" and "[- -]" to represent the edit actions of insertion and deletion respectively. For example, the sentence in HTML format

“外資也不急著<FONT style= ” TEXT-DECORATION: line-through” class=3 title=XXX 刪除, color=#555588>佈</FONT><FONT class=1 title=XXX 新增, color=#265e8a>布</FONT>局明年，”

is converted to “外資也不急著[-佈-]{+布+}局明年，” (”Foreign investment is not in a hurry to layout next year,”).

After that, we attempt to extract the sentences that contain at least one typo. As shown in Figure 3, the edit logs could contain many kinds of edits, including spelling correction, content changes, and style modification (such as synonyms replacement). Among these edits, we are only concerned with spelling correction. However, lack of edit type annotation makes it difficult to directly identify spelling errors. Thus, we consider consecutive single-character edit pairs of deletion and insertion (e.g., “[-佈-]{+布+}” or “{+布+}[-佈-]”) as spelling correction, and extract the sentences containing such edit pairs. Furthermore, we use a set of rules to filter out some kinds of edits such as time-related and digital-related. Figure 3 shows some edited sentences, the fifth, sixth, seventh, eighth and eleventh sentences are regarded as sentences with spelling errors according these simple rules. The output of this stage is a set of sentences with spelling errors annotated using simple edit tags, as shown in Figure 4.

- 一些較落後地區（如孟加拉）因表面水體受到[-污-]{+汙+}染，
- 到大陸創業條件首要是膽[-試-]{+識+}，
- [-盡-]{+敬+}請鎖定相關報導。
- 現場{+儼+}[-嚴-]然成為超跑展示中心，
- 十六支球隊要{+爭+}[-整-]取十二張高雄複賽門票。
- 也連續兩年創下歷史新高{+記+}[-紀-]錄。
- 一口氣追回昨天創下英國脫歐以來最大[-鵚-]{+頹+}勢，
- 把施工圍籬變成美麗的彩繪或塗[-鴨-]{+鴉+}，
- 也通報捕狗隊來協助；對受害學童[-己-]{+已+}派員慰問。
- 不論在市[-佔-]{+占+}率、獲利表現、品牌及服務等各方面，

Figure 4. Example outputs for the step of extracting misspelled sentences

Although this approach for extracting the edited sentences involving spelling correction can obtain quite a few results, there is still a room for improvement. For example, the edited sentence “價值上百萬的好禮[-通通-]{+統統+}帶回家。” (’Bring millions of good gifts home’) contains a consecutive two-character edit pair “[-通通-]{+統統+}” (both pronounced ’tong tong’), which is also spelling error correction. However, it is not extracted because we only consider consecutive single-character edit pairs. In some cases, an edited sentence might be wrongly regarded as misspelled sentence. For example, the sentence “這

項計畫將持續募款到今年[-聖-]{+耶+}誕節，” (’This project will continue to raise funds until this Christmas,’) contains an edit pair “[-聖-]{+耶+}” about style modification. Consider the context of the edited character, the word “聖誕節” (pronounced ’sheng dan jie’, it means the birthday of the holy child Jesus) and “耶誕節” (pronounced ’ye dan jie’, it means the birthday of Jesus) are both correct, and they almost mean the same thing. For such case, using word segmentation and meaning similarity measure of two words may be helpful.

3.2.2 Generating Artificially Misspelled Sentences

In the second stage of training process, we create a set of artificial misspelled sentences for expanding our training data. These generated data are expected to make the Chinese spelling checker more effective.

```
procedure GenerateErrorSentence_Map(CorrectSentences)

    for each Sentence in CorrectSentences
       for each Word_i in Sentence
(1)       WrongWords = getConfusionSet(Word_i)
          for each WrongWord_j in WrongWords
(2a)         WrongSentence = Sentence
(2b)         replace WrongSentence_i with WrongWord_j
(3a)         WordPair = Word_i + "|||" + WrongWord_j
(3b)         SentencePair = Sentence + "|||" + WrongSentence
(4)          output <WordPair, SentencePair>


procedure GenerateErrorSentence_Reduce(WordPairs, SentencePairs)

(1) N = n
    for each WordPairs, SentencePairs
(2)    shuffle SentencePairs
(3)    output top N SentencePairs
```

Figure 5. Generating artificial misspelled sentence

Table 1. Examples of confusion set

Correct Word	Wrong Words
部署(’arrange’, pronounced ’bu shu’)	布署, 部處, 佈署, 步署
賠罪(’apologize’, pronounced ’pei zui’)	培罪, 陪罪

The input to this stage is a set of presumably error-free sentences from published texts with word segmentation done using a word segmentation tool provided by the CKIP Project (Ma & Chen, 2003). Artificially misspelled sentences are generated by injecting errors into these error-free sentences. Although a correct word could be misspelled as any other Chinese word, some right-and-wrong word pairs are more likely to happen than others. In order to generate realistic spelling errors, we use a confusion set consisting of commonly confused right-and-wrong word pairs (see Table 1). The wrong words in confusion set are used to replace counterpart correct words in the sentences. For example, we use error-free sentence "也跟患者賠罪了十分鐘" ('also apologized to the patient for ten minutes') to generate three misspelled sentences, as shown in Table 2. Figure 5 shows the procedure for generating artificial misspelled sentences using the MapReduce framework to speed up the process.

Table 2. Artificial misspelled sentences for '也跟患者賠罪了十分鐘'

Artificial Misspelled Sentence	Replaced Word	Wrong Word
也跟患者培罪了十分鐘	賠罪	培罪
也跟患者陪罪了十分鐘	賠罪	陪罪
也跟患者賠罪了十分鍾	分鐘	分鍾

- **Map procedure**: In Step (1), for each word in the given (presumably) error-free sentence with length not longer than 20 words, we obtain the corresponding confused words. For example, the confusion set of word "賠罪" contains two confused wrong words: "培罪" and "陪罪". The original word is then replaced with its corresponding confused words in Steps (2a) and (2b). To work with *MapReduce* framework, we then format the output data to key-value pair in Step (3a) and (3b). In order to group the generated misspelled sentences according to replacement (e.g., "賠罪" is replaced with "培罪"), we use a right-and-wrong word pair (e.g., "賠罪|||培罪") to be the key, and a right-and-wrong sentence pair (e.g., "也跟患者賠罪了十分鐘|||也跟患者培罪了十分鐘") to be the value. Finally, the key-value pair is outputted in Step (4).
- **Reduce procedure**: In this procedure, the inputs are the key-value pairs outputted by Mapper. For each word pair, there might be too many sentence pairs. Thus, in Step (1), we set a threshold N to limit the number of sentences generated. In order to randomly sample a set of sentences, we make these sentence pairs redistributed by shuffling in Step (2), and output the first N of sentence pairs in Step (3).

The output of this stage is a set of right-and-wrong sentence pairs, as shown in Table 3.

The confusion set plays an important role in this stage, so it is critical to decide what kinds of confusion set to use. There are several available word-level and character-level confusion sets. However, compare to word-level, a Chinese character could be confused with

more other characters based on shape and sound similarity. For example, the character "賠" is confused with 23 characters with similar shape and 21 characters with similar sound in a character-level confusion set, while the word "賠罪" is confused with only two words in a word-level confusion set. Moreover, an occurring typo might involve not only the character itself but also the context. If we use the character-level confusion set, an error-free sentence would produce numerous and probably unrealistic artificial misspelled sentences. Therefore, we decide to use word-level confusion sets.

Table 3. Example outputs for the step of generating artificial misspelled sentences

Right Sentence	**Wrong Sentence**
可見酒精會讓白老鼠上癮，	可見酒精會讓白老鼠上蔭，
導致水圳混濁不堪，	導致水圳混濁不勘，
媒體何嘗沒有一點責任？	媒體何賞沒有一點責任？
地處偏僻且巷弄狹窄，	地處編僻且巷弄狹窄，
希望他的覺醒為時不晚。	希望他的覺省為時不晚。

3.2.3 Neural Machine Translation Model

In the third and final stage of training process, we train a character-based neural machine translation (NMT) model for developing a Chinese spelling checker, which translates a potentially misspelled sentence into a correct one.

The architecture of NMT model typically consists of an encoder and a decoder. The encoder consumes the source sentence $X = [x_1,x_2,...,x_I]$ and the decoder generates translated target sentence $Y = [y_1,y_2,...,y_J]$. For the task of correcting spelling errors, a potentially misspelled sentence is treated as the source sentence X, which is translated into the target sentence Y with errors corrected. To train the NMT model, we use a set of right-and-wrong sentence pairs from edit logs (Section 3.2.1) and artificially-generated data (Section 3.2.2) as target-and-source training sentence pairs.

In the training phase, the model is given (X, Y) pairs. At encoding time, the encoder reads and transforms a source sentence X, which is projected to a sequence of embedding vectors $\mathbf{e} = [e_1,e_2,...,e_I]$, into a context vector $\mathbf{c}$:

$$c = q(h_1,h_2,...,h_I) \tag{1}$$

where q is some nonlinear function.

We use a bidirectional recurrent neural network (RNN) encoder to compute a sequence of hidden state vectors $\mathbf{h} = [h_1,h_2,...,h_I]$. The bidirectional RNN encoder consists of two independent encoders: a forward and a backward RNN. The forward RNN encodes the normal

sequence, and the backward RNN encodes the reversed sequence. A hidden state vector h_i at time i is defined as:

$$fh_i = \text{ForwardRNN}(h_{i-1}, e_i) \tag{2}$$

$$bh_i = \text{BackwardRNN}(h_{i+1}, e_i) \tag{3}$$

$$h_i = [fh_i || bh_i] \tag{4}$$

where || denotes the vector concatenation operator.

At decoding time, the decoder is trained to output a target sentence Y by predicting the next character y_j based on the context vector c and all the previously predicted characters $\{y_1, y_2, \ldots, y_{j-1}\}$:

$$p(\mathrm{Y} \mid \mathrm{X}) = \prod_{j=1}^{J} p(y_j \mid y_1, y_2, \ldots, y_{j-1}; c) \tag{5}$$

The conditional probability is modeled as:

$$p(y_j \mid y_1, y_2, \ldots, y_{j-1}; c) = g(y_{j-1}, h'_j, c) \tag{6}$$

where g is a nonlinear function, and h'_j is the hidden state vector of the RNN decoder at time j.

We use an attention-based RNN decoder that focuses on the most relevant information in the source sentence rather than the entire source sentence. Thus, the conditional probability in Equation 5 is redefined as:

$$p(y_j \mid y_1, y_2, \ldots, y_{j-1}; \mathbf{e}) = g(y_{j-1}, h'_j, \mathbf{c_j}) \tag{7}$$

where the hidden state vector h'_j is computed as follow:

$$h'_j = f(y_{j-1}, h'_{j-1}, \mathbf{c_j}) \tag{8}$$

$$c_j = \sum_{i=1}^{I} a_{ji} h_i \tag{9}$$

$$a_{ji} = \frac{\exp(\text{score}(h'_j, h_i))}{\sum_{i'=1}^{I} \exp(\text{score}(h'_j, h_{i'}))} \tag{10}$$

Unlike Equation 6, here the probability is conditioned on a different context vector c_j for each target character y_j. The context vector c_j follows the same computation as in Bahdanau, Cho, and Bengio (2014). We use the global attention approach (Luong, Pham & Manning, 2015) with general score function to compute the attention weight a_{ji}:

$$\text{score}(h_j^{'}, h_i) = h_{j\mathrm{T}}^{'} W_a h_i \tag{11}$$

Instead of implementing an NMT model from scratch, we use *OpenNMT* (Klein, Kim, Deng, Senellart, & Rush, 2017), an open source toolkit for neural machine translation and sequence modeling, to train the model. The training details and hyper-parameters of our model will be described in Section 4.2.

3.3 Run-time Error Correction

Once the NMT model is automatically trained for correcting spelling errors, we apply the model at run time. *AccuSpell* then corrects a given potentially misspelled sentence with the character-based NMT model using the procedure in Figure 6.

```
procedure CorrectSpellingError(Sentence)

(1)  sourceSentence = tokenize(Sentence)
(2)  targetSentence = NMTModel(sourceSentence)

(3a) copy sourceSentence to Result
     for each sourceChar_i in sourceSentence:
        if sourceChar_i not equals to targetChar_i
(3b)       replace Result_i with "[-sourceChar_i-]{+targetChar_i+}"

(4)  return Result
```

Figure 6. Correcting spelling errors in a sentence

With a character-based NMT model, the input sentence is expected to follow the format that tokens are space-separated. Thus, in Step (1), the characters in the given sentence are separated with space. For example, “今晚月色很美，我想小灼一杯。” is transformed into “今晚月色很美，我想小灼一杯。”. In Step (2), the source sentence is fed to our NMT model. During processing, the encoder first transforms the source sentence into a sequence of vectors. The decoder then computes the probabilities of predicted target sentences given the vectors of source sentence. Finally, a beam search is used to find a target sentence that approximately maximizes the conditional probability. Table 4 shows the top three target sentences predicted by our NMT model for the source sentence “今晚月色很美，我想小灼一杯。”, and the highest-score one “今晚月色很美，我想小酌一杯。” is returned as the correction.

Table 4. Top three target sentences of the source sentence "今晚月色很美，我想小灼一杯。" predicted by NMT model

Target Sentence	Predicted Score	Rank
今晚月色很美，我想小酌一杯。	-0.0047	1
今晚月色也美，我想小酌一杯。	-6.93	2
今晚月色很美，我想小灼一耶。	-7.36	3

To give useful and clear feedback, we convert the correction result into a informative expression instead present users with the output of NMT model directly. Therefore, in Steps (3a) and (3b), we compare the source sentence with the target sentence to find out the differences between them, and use simple edit tags to mark these differences. Finally in Step (4), the converted result (e.g., "今晚月色很美，我想小[-灼-]{+酌+}一杯。") is returned by *AccuSpell*. As shown in Figure 1, the characters to be deleted (e.g., "[-灼-]") are colored in red, while the inserted characters (e.g., "{+酌+}") are colored in green.

4. Experimental Setting

AccuSpell was designed to correct spelling errors in Chinese texts written by native speakers. As such, *AccuSpell* will be trained and evaluated using mainly real edit logs and a newspaper corpus. In this section, we first give a brief description of the datasets used in the experiments in Section 4.1, and describe the hyper-parameters for the NMT model in Section 4.2. Then several NMT models with different experimental setting for comparing performance are described in Section 4.3. Finally in Section 4.4, we introduce the evaluation metrics for evaluating the performance of these models.

4.1 Dataset

United Daily News (UDN) Edit Logs: UDN Edit Logs was provided to us by UDN Digital. This dataset records the editing actions of daily UDN news from June 2016 to January 2017. There are 1.07 million HTML files with more than 30 million edits of various types, with approximately 11 million insertions and 20 million deletions. However, lack of edit type annotation makes it difficult to directly identify spelling errors. Thus, we extracted a set of annotated sentences involving spelling error correction from this edit logs using the approach described in Section 3.2.1. To train on NMT model, we transformed every annotated sentence into a source-and-target parallel sentence. For example, "外資也不急著[-佈-]{+布+}局明年，" is transformed into a source sentence "外資也不急著佈局明年，" and a target sentence "外資也不急著布局明年，". In total, there are 238,585 sentences extracted from UDN Edit Logs, and each sentence contains only edits related to spelling errors. We divided these extracted sentences into two parts: one (226,913 sentences) for training NMT models,

and the other (11,943 sentences) for evaluation in our experiments.

United Daily News (UDN): The UDN news dataset was also provided by UDN Digital. The dataset consists of published newswire data from 2004 to 2017, which contains approximately 1.8 million news articles with over 530 million words. Unlike UDN Edit Logs, UDN are composed of news articles which had been edited and published. We used the presumably error-free sentences in this dataset to generate artificially misspelled sentences, as described in Section 3.2.2.

Table 5. Examples of 聯合報統一用字(Uniform Words List of UDN)

	Recommended word	**Unrecommended word**
巴吧 (pronounced 'ba')	啞巴('dumb')	啞吧
背揹 (pronounced 'bei')	背著('carrying') 背黑鍋('take the blame')	揹著 揹黑鍋
刨鉋 (pronounced 'bao')	刨冰('shaved ice')	鉋冰
杯盃 (pronounced 'bei')	市長杯('mayor cup')	市長盃
澹淡 (pronounced 'dan')	慘澹('miserable') 淡泊 ('indifferent')	慘淡 澹泊
闆板 (pronounced 'ban')	老闆('boss')	老板

Confusion Set: We used five distinct confusion sets collected from different sources:

- **聯合報統一用字(Uniform Words List of UDN):** The dataset of 聯合報統一用字 provided by UDN Digital contains 1,056 easily confused word pairs. As shown in Table 5, the confused word pairs indicate that which words are recommended and which ones should not be used for UDN news articles. However, not all the unrecommended words are wrong because the suggestions are just preference rules for writing news articles for the UDN journalists. For example, a confused word pair [“市長杯”, “市長盃”]('Mayor CUP') in Table 5, the former is recommended and the latter is not recommended, but they are both correct and in common use. In our work, we collect all the word pairs, and consider them as right-and-wrong word pairs
- **東東錯別字(Kwuntung Typos Dictionary):** This dataset was collected from the Web (www.kwuntung.net/check/), which contains a set of commonly confused right-and-wrong word pairs. For each word pair, there is one distinct character with similar pronunciation or shape between right and wrong word. We obtain 38,125 different right-and-wrong word pairs in total, which constitutes the main part of our confusion set.
- **新編常用錯別字門診(New Common Typos Diagnosis):** This dataset comes from the print publication: 新編錯別字門診 (蔡有秩, 2003) and contains 492 right-and-wrong

word pairs.

- **常見錯別字辨正辭典(Dictionary of Common Typos):** This dataset is from a print publication: 常見錯別字辨正辭典 (蔡榮圳, 2012). There are 601 right-andwrong word pairs in total.
- **國中錯字表(The Typos List for Middle School):** This dataset contains a set of commonly misused right-and-wrong word pairs for middle school students. There are 1,720 word pairs in original. However, some pairs are composed of phrases (e.g., “觀念不佳” and “為自己的未來鋪路”) instead of words. To ensure that all pairs are at word level, we used some rules to transform the phrase pairs into word pairs. For example, the right-and-wrong phrase pair [“為自己的未來鋪路”, “為自己的未來捕路”] (’Pave the way for your own future’) is transformed to the word pair [“鋪路”, “捕路”] (pronounced ’pu lu’ and ’bu lu’). Moreover, we discarded the pairs cannot be transformed such as [“十來枝的掃具”, “十來隻的掃具”] (’A dozen brooms.’). After that, 1,551 word pairs remained.

The confused word pairs of five confusion sets are combined into a collection with over 40,000 word pairs. However, for a given confused word pair, the judgments in different confusion sets might be inconsistent. Consider a confused word pair [“鐘錶”,“鐘表”](’Clock’, pronounced ’zhong biao’). “鐘錶” is right and “鐘表” is wrong in Kwuntung Typos Dictionary, while “鐘表” is adopted and “鐘錶” is not recommended in Uniform Words List of UDN. Furthermore, the confusion sets are not guaranteed to be absolutely correct. To resolve these problems, we used the Chinese dictionary published by Ministry of Education of Taiwan as the gold standard. After filtering out the invalid word pairs, the new confusion set **CFset** with 33,551 distinct commonly confused word pairs were obtained. Table 6 shows the number of word pairs of all confusion sets.

Table 6. Number of word pairs of five confusion sets

Confusion Set	Number of confused word pairs
Uniform Words List of UDN	1,056
Kwuntung Typos Dictionary	38,125
New Common Typos Diagnosis	492
Dictionary of Common Typos	601
The Typos List for Middle School	1,460
CFset	33,551

Table 7. The statistics of test sets

	UDN Edit Logs	SIGHAN-7
# of sentences	1,175	6,101
# of sentences with errors	919	1,222
# of sentences without errors	256	4,879
# of error characters	919	1,266
Average # of errors in sentences with errors	1	1.04
Average length of sentences	17.47	12.16

Test Data: We used two test sets for evaluation, and Table 7 shows the statistical analysis of them in detail:

- **UDN Edit Logs:** As mentioned earlier, UDN Edit Logs were partitioned into two independent parts, for training and testing respectively. The test part contains 11,943 sentences and we only used 1,175 sentences for evaluation, 919 out of which contain at least one error.
- **SIGHAN-7:** We also used the dataset provided by SIGHAN 7 Bake-off 2013 (Wu, Liu & Lee, 2013). This dataset contains two subtasks: Subtask 1 is for error detection and Subtask 2 is for error correction. In our work, we focus on evaluating error correction, so we used Subtask 2 as an additional test set. There are 1,000 sentences with spelling errors in Subtask 2, and the average length of sentences is approximately 70 characters. To be consistent with UDN Edit Logs, we segmented these sentences into 6,101 clauses, and 1,222 of which contain at least one error.

4.2 Hyper-parameters of NMT Model

We trained several models using the same hyper-parameters in our experiments. For all models, the source and target vocabulary sizes are limited to 10K since the models are trained at character level. For source and target characters, the character embedding vector size is set to 500. We trained the models with sequences length up to 50 characters for both source and target sentences.

The encoder is a 2-layer bidirectional long-short term memory (LSTM) networks, which consists of a forward LSTM and a backward LSTM, and the decoder is also a 2layer LSTM. Both the encoder and the decoder have 500 hidden units. We use the Adam Algorithm (Kingma & Ba, 2014) as the optimization method to train our models with learning rate 0.001, and the maximum gradient norm is set to 5. Once a model is trained, beam search with beam size set to 5 is used to find a translation that approximately maximizes the probability.

4.3 Models Compared

Our experimental evaluation focuses on writing of native speakers. Therefore, we used UDN Edit Logs and the artificially generated misspelled sentences as the training data. To investigate whether adding artificially generated data improves the performance of our Chinese spelling check system, we compared the results produced by several models trained on different combination of datasets.

In addition, we use some additional features on source and target words in the form of discrete labels to train the NMT model[1]. As Liu *et al*. (2011) stated, around 75% of typos were related to the phonological similarity between the correct and the incorrect characters, and about 45% were due to visual similarity. Thus, we use the pronunciation and shape of a character from the Unihan Database[2] as the additional feature of the source and target characters. As an example, for the character “詣”, the pronunciation feature is “ㄧ” (without considering the tone) and the shape features are “言” and “旨”. On the other hand, a spelling error might involve not only the character itself but also the context, so we use the context (with window size 1) of a character as additional features to train another model.

Table 8. Features for the sentence “我想小酌一杯。”

Feature	我	想	小	酌	一	杯	。
Sound	ㄨㄛ (wo)	ㄒㄧㄤ (xiang)	ㄒㄧㄠ (xiao)	ㄓㄨㄛ (zhuo)	ㄧ (yi)	ㄅㄟ (bei)	N
Shape	(戈,我)	(心,相)	(小,小)	(酉,勺)	(一,一)	(木,不)	(N,N)
Context	(BEG,想)	(我,小)	(想,酌)	(小,一)	(酌,杯)	(一,。)	(杯,END)

Table 8 gives an example to illustrate the pronunciation, shape, and context features.

There are totally eight models trained for comparing, and only last two were trained with features. The eight models evaluated and compared are as follows:

- **UDN-only:** The model was trained on 226,913 sentence pairs from the training part of UDN Edit Logs.
- **UDN + Artificial (1:1):** The model was trained on 226,913 sentence pairs from the training part of UDN Edit Logs plus 225,985 artificially generated sentence pairs (452,871 in total).
- **UDN + Artificial (1:2):** The model was trained on 226,913 sentence pairs from the training part of UDN Edit Logs plus 440,143 artificially generated sentence pairs (667,056

[1] https://opennmt.net/OpenNMT/data/word_features/
[2] http://www.unicode.org/charts/unihan.html

in total).

- **UDN + Artificial (1:3):** The model was trained on 226,913 sentence pairs from the training part of UDN Edit Logs plus 673,006 artificially generated sentence pairs (899,919 in total).
- **UDN + Artificial (1:4):** The model was trained on 226,913 sentence pairs from the training part of UDN Edit Logs plus 899,385 artificially generated sentence pairs (1,126,298 in total).
- **Artificial-only:** The model was trained on 899,385 artificially generated sentence pairs.
- **FEAT-Sound & Shape:** The model was trained on the same data in *UDN +Artificial (1:3)* model with pronunciation and shape of character features.
- **FEAT-Context:** The model was trained on the same data in *UDN + Artificial (1:3)* model with context features.

4.4 Evaluation Metrics

Chinese spelling check systems are usually compared based on two main metrics, precision and recall. We use the metrics provided by SIGHAN-8 Bake-off 2015 for Chinese spelling check shared task (Tseng, Lee, Chang, & Chen, 2015), which include False Positive Rate, Accuracy, Precision, Recall, and F1, to evaluate our systems.

The confusion matrix is used for calculating these evaluation metrics. In the matrix, TP (True Positive) is the number of sentences with spelling errors that are correctly identified by the developed system; FP (False Positive) is the number of sentences in which non-existent errors are identified; TN (True Negative) is the number of sentences without spelling errors which are correctly identified as such; FN (False Negative) is the number of sentences with spelling errors that are not correctly identified. The following metrics are calculated using TP, FP, TN and FN:

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN} \tag{12}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{16}$$

Table 9. The given test sentences with gold standards

Sentence ID	Sentence	Gold Standard
S1	希望藉此鼓勵自己和他人要積極樂觀實現夢想。	0
S2	PM2.5 對人體健康為害大，	11, 危
S3	因為難以達到連數門檻，	8, 署
S4	他仍記得自已當年還是學校棒球隊員，	6, 己
S5	剛推動的社會住密也要設一定比例的大陽光電。	8, 宅, 17, 太
S6	美麗的勇士山頭將被掏空了嗎？	10, 淘
S7	未來發展需要新的能力、新的動能，	0
S8	學生因宗教、重族、國籍而遭羞辱者大幅增加。	7, 種

Table 10. The results outputted by the system

Sentence ID	Output Sentence	Correction
S1	希望藉此鼓勵自己和他人要積極樂觀實現夢想。	0
S2	PM2.5 對人體健康危害大，	11, 危
S3	因為難以達到連數門檻，	8, 署
S4	他還記得自己當年還是學校棒球隊員，	2, 還, 6, 己
S5	剛推動的社會住宅也要設一定比例的大陽光電。	8, 宅
S6	美麗的勇士山頭將被掏空了嗎？	0
S7	未來發展需要新的能力、新的動能，	15, 力
S8	學生因宗教、種族、國籍而遭羞辱者大幅增加。	7, 種

For example, given 8 test sentences with gold standards shown in Table 9. Assume that our system outputs the results as shown in Table 10, the evaluation metrics will be measured as follows:

- FPR = 0.5 (= 1/2)

  Notes: {S7}/{S1, S7}

- Accuracy = 0.5 (= 4/8)

  Notes: {S1, S2, S3, S8}/{S1, S2, S3, S4, S5, S6, S7, S8}
- Precision = 0.5 (= 3/6)

  Notes: {S2, S3, S8}/{S2, S3, S4, S5, S7, S8}
- Recall = 0.75 (= 3/4)

  Notes: {S2, S3, S8}/{S2, S3, S6, S8}
- F1 = 0.6 (= 2*0.5*0.75/(0.5+0.75))

5. Results and Discussion

In this section, we report the results of experimental evaluation using the resources and metrics described in previous chapter. Specifically, we report the results of our evaluation, which contains two test sets evaluated by false positive rate (FPR), accuracy, precision, recall, and F1 score. First, we present the results of several models evaluated on two test sets in Section 5.1. We then give some analysis and discussion of the errors in the two test sets in Section 5.2.

5.1 Evaluation Results

Table 11 shows the evaluation results of UDN Edit Logs. As we can see, all models trained on edit logs and artificially generated data perform better than the one trained on only edit logs. Moreover, the model trained on only edit logs performs slightly worse, while the model trained on only artificially generated data performs the very worst on all metrics. Even though the model trained with sound and shape features performs relatively poorly on FPR, it has the best performance on accuracy, precision, recall, and F1 score.

Table 11. Evaluation results of UDN Edit Logs

Model	FPR	Accuracy	Precision	Recall	F1
UDN-only	.066	.64	.80	.64	.71
UDN + Artificial (1:1)	.090	.69	.84	.69	.76
UDN + Artificial (1:2)	.063	.71	.86	**.72**	.78
UDN + Artificial (1:3)	.066	.70	.86	.69	.76
UDN + Artificial (1:4)	**.059**	.71	.87	.71	.78
Artificial-only	.137	.35	.43	.26	.33
FEAT-Sound & Shape	.098	**.72**	**.88**	**.72**	**.79**
FEAT-Context	.059	.71	.87	.70	.78

Table 12. Evaluation results of SIGHAN-7

Model	FPR	Accuracy	Precision	Recall	F1
UDN-only	.109	.74	.19	.17	.18
UDN + Artificial (1:1)	.089	.83	.50	.59	.54
UDN + Artificial (1:2)	.081	.84	.54	.61	.57
UDN + Artificial (1:3)	.078	**.85**	.56	.62	.58
UDN + Artificial (1:4)	**.073**	**.85**	**.58**	.63	**.61**
Artificial-only	.079	.84	.53	.58	.56
FEAT-Sound & Shape	.097	.83	.51	**.64**	.57
FEAT-Context	.080	.84	.56	.61	.58

For the other test set, SIGHAN-7, the evaluation results are shown in Table 12. UDN + Artificial (1:4) performs substantially better than the other models, noticeably improving on all metrics. Interestingly, in contrast to the results of UDN Edit Logs, the model trained on only edit logs has significantly worse performance than others, while the model trained on only artificially generated data performs reasonably well. We note that there is no obvious improvement in the performance of the model trained with additional features of either sound and shape or context.

In general, we obtain extremely low average FPR evaluated on the two test sets. There are three obvious differences between the results of two test sets. First, the model trained on only edit logs (**UDN-only**) and the model trained on only artificially generated data (**Artificial-only**) have the opposite results on UDN Edit Logs and SIGHAN-7. As we can see, **UDN-only** performs well on UDN Edit Logs but very poorly on SIGHAN-7. In contrast, **Artificial-only** has worst performance on UDN Edit Logs but acceptable performance on SIGHAN-7. Second, we obtain relatively high precision compared with recall on UDN Edit Logs, while higher recall than precision on SIGHAN-7. Third, in Table 13, it is worth noting that the model trained with sound and shape features has significantly better accuracy, recall, and F1 score on UDN Edit Logs. However, on SIGHAN-7, only the recall is a little better than the model trained without using features.

Table 13. Evaluation results related to the models trained with features

Test Set	Model	FPR	Accuracy	Precision	Recall	F1
UDN Edit Logs	UDN + Artificial (1:3)	.066	.70	.86	.69	.76
	FEAT-Sound & Shape	.098	**.72**	**.88**	**.72**	**.79**
	FEAT-Context	**.059**	.71	.87	.70	.78
SIGHAN-7	UDN + Artificial (1:3)	**.078**	**.85**	**.56**	.62	**.58**
	FEAT-Sound & Shape	.097	.83	.51	**.64**	.57
	FEAT-Context	.080	.84	**.56**	.61	**.58**

Table 14. Distribution of the relations between typos and corrections in test sets

	UDN Edit Logs	SIGHAN-7
# of error characters	919	1,266
Similar Sound	70%	84%
Similar Shape	36%	40%
Similar Sound and Shape	30%	30%

5.2 Error Analysis

The nature of our two test sets are different, UDN Edit Logs are produced by newspaper editors, while SIGHAN-7 are collected from essays written by junior high students. Therefore, we analyze and discuss the details of the two test sets in this section.

We use the confusion sets provided by SIGHAN 7 Bake-off 2013 (Wu *et al.*, 2013), which contains a set of characters with similar pronunciation and shape, to analyze the relations between typos and the corresponding corrections in our test data. There are 919 typos in UDN Edit Logs and 1,266 typos in SIGHAN-7. As shown in Table 14, the analysis results of UDN Edit Logs and SIGHAN-7 are similar. Most of typos are related to similar pronunciation, and over 35% of typos are due to similar shape. Moreover, around 30% of typos are associated with similar pronunciation as well as shape.

Table 15 and 16 show some analysis of evaluation results of UDN Edit Logs and SIGHAN-7 respectively. As we can see, according to the analysis of the errors which were not corrected by models, there is no significant difference among these different models. In both UDN Edit Logs and SIGHAN-7, around half of the spelling errors not corrected are related to similar pronunciation no matter which model we used.

Table 15. Distribution of the relations between not corrected typos and corrections of the evaluation results using UDN Edit Logs

Model	# of errors not corrected	Similar Sound	Similar Shape	Similar Sound and Shape
UDN-only	404	52%	7%	27%
UDN+Artificial (1:3)	340	54%	8%	26%
Artificial-only	733	43%	6%	26%
FEAT-Sound&Shape	299	57%	8%	25%

Table 16. Distribution of the relations between not corrected typos and corrections of the evaluation results using SIGHAN-7

Model	# of errors not corrected	Similar Sound	Similar Shape	Similar Sound and Shape
UDN-only	1,092	57%	9%	27%
UDN+Artificial (1:3)	596	60%	8%	22%
Artificial-only	641	58%	8%	24%
FEAT-Sound&Shape	597	58%	8%	24%

It is worth discussing that there are some special cases in the test sets. For example, an error character “怖” (pronounced ’bu’) occurring in some words such as “怖告欄” (pronounced ’bu gao lan’) and “怖置” (pronounced ’bu zhi’) should be corrected to “佈” (pronounced ’bu’) in SIGHAN-7. However, the correction predicted by our models is “布” since we used the Chinese dictionary published by Ministry of Education of Taiwan as the gold standards of our training data. According to the dictionary, “佈置” and “佈告欄” are invalid, while “布置” (’decorate’) and “布告欄” (’bulletin board’) are legal. Another case is related to grammatical errors. Our models aim to correct spelling errors, but there are some sentences with grammatical errors in SIGHAN-7 such as “要如何在站起來呢？” (’How to stand up again?’) and “哪激的起美麗的浪花？” (How can it stir up the beautiful spray?), where “在” (pronounced ’zai’) and “的” (pronounced ’de’) should be “再” (pronounced ’zai’) and “得” (pronounced ’de’) respectively. These kinds of errors are involved the dependency structure of sentences. In the predicted results of our models, we found that the model trained on only artificially generated data cannot correct such errors. Other models using edit logs have slightly better performance on correcting these kinds of errors, but there isn’t too much of a difference.

Besides the test data, we also found that the model trained with additional features could correct some new and unseen errors. For example, the sentence “他在文學方面有很高的造

酯。” with a typo “酯” (pronounced ’zhi’), which is not corrected by a model trained without features because our training data does not cover this typo. However, the sentence is correctly translated into “他在文學方面有很高的造詣。” by the model trained with sound and shape features.

6. Conclusion and Future Work

Many avenues exist for future research and improvement of our system. For example, the method for extracting misspelled sentences from newspaper edit logs could be improved. When extracting, we only consider the sentences contain consecutive single-character edit pairs. However, two-character edit pairs could also involve spelling correction. Moreover, we could investigate how to use character-level confusion sets to expand the scale of confused word pairs. If we have more possibly confused word pairs, we could generate more comprehensive artificial error data. Additionally, an interesting direction to explore is expanding the scope of error correction to include grammatical errors. Yet another direction of research would be to consider focusing on implementing the neural machine translation model for Chinese spelling check.

In our work, we pay more attention to the aspect of data and methods of augmenting data for CSC. We collect a series of confusion set from the Web, including 東東錯別字 (Kwuntung Typos Dictionary), 新編常用錯別字門診(New Common Typos Diagnosis), 常用錯別字(Dictionary of Common Typos), 國中錯字表(The Typos List for Middle School). To augment more data for training an NMT model, we develop a way of injecting artificial errors into error-free sentences with the confusion sets. In addition, we compare the different ratio of mixture of real and artificial data and more artificial data improves the performance. Finally, we conduct experiments on models with additional features (e.g., pronunciation, shape components, and context words) to show that phonological, visual, and context information can improve the recall and reveal the ability to generalize common typos.

In summary, we have proposed a novel method for learning to correct typos in Chinese text. The method involves combining real edit logs and artificially generated errors to train a neural machine translation model that translates a potentially erroneous sentence into correct one. The results prove that adding artificially generated data successfully improves the overall performance of error correction.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In arXiv preprint arXiv:1409.0473.

Chang, C.-H. (1995). A new approach for automatic chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, *95*, 278-283.

Chiu, H.-w., Wu, J.-c., & Chang, J. S. (2013). Chinese spelling checker based on statistical machine translation. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, 49-53.

Chollampatt, S. & Ng, H. T. (2018). A multilayer convolutional encoder-decoder neural network for grammatical error correction. In arXiv preprint arXiv:1801.08831.

Felice, M. & Yuan, Z. (2014). Generating artificial errors for grammatical error correction. In Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 116-126. doi: 10.3115/v1/E14-3013

Gu, S. & Lang, F. (2017). A chinese text corrector based on seq2seq model. In *Proceedings of 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 322-325. doi: 10.1109/CyberC.2017.82

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

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Opensource toolkit for neural machine translation. In arXiv preprint arXiv:1701.02810.

Liu, C.-L., Lai, M.-H., Tien, K.-W., Chuang, Y.-H., Wu, S.-H., & Lee, C.-Y. (2011). Visually and phonologically similar characters in incorrect chinese words: Analyses, identification, and applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, *10*(2),10. doi: 10.1145/1967293.1967297

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attentionbased neural machine translation. In arXiv preprint arXiv:1508.04025.

Ma, W.-Y. & Chen, K.-J. (2003). Introduction to ckip chinese word segmentation system for the first international chinese word segmentation bakeoff. In *Proceedings of the 2nd SIGHAN on CLP*, 168-171. doi: 10.3115/1119250.1119276

Rei, M., Felice, M., Yuan, Z., and Briscoe, T. (2017). Artificial error generation with machine translation and syntactic patterns. In arXiv preprint arXiv:1707.05236.

Tseng, Y.-H., Lee, L.-H., Chang, L.-P., & Chen, H.-H. (2015). Introduction to sighan 2015 bake-off for chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, 32-37. doi: 10.18653/v1/W15-3106

Wu, S.-H., Chen, Y.-Z., Yang, P.-C., Ku, T., & Liu, C.-L. (2010). Reducing the false alarm rate of chinese character error detection and correction. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese spelling check evaluation at sighan bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, 35-42.

Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., & Ng, A. Y. (2016). Neural language correction with character-based attention. In arXiv preprint arXiv:1603.09727.

Yuan, Z. & Briscoe, T. (2016). Grammatical error correction using neural machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies, 380-386. doi: 10.18653/v1/N16-1042

Zhang, L., Huang, C., Zhou, M., & Pan, H. (2000). Automatic detecting/correcting errors in chinese text by an approximate word-matching algorithm. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 248-254. doi: 10.3115/1075218.1075250

蔡有秩 (2003)。新編錯別字門診。 語文訓練叢書，螢火蟲。[Tsai, Y.-J. (2003). *New Common Typos Diagnosis*, Fireflybooks.]

蔡榮圳 (2012)。常見錯別字辨正辭典。中文可以更好，商周出版。 [Tsai, R.-J. (2012). *Dictionary of Common Typos*, Business Weekly.]

Computational Linguistics and Chinese Language Processing
Vol. 25, No. 1, June 2020, pp. 29-56

基於端對端模型化技術之語音文件摘要

Spoken Document Summarization Using End-to-End Modeling Techniques

劉慈恩*+、劉士弘#、張國韋*、陳柏琳+

Tzu-En Liu, Shih-Hung Liu, Kuo-Wei Chang, and Berlin Chen

摘要

本論文主要探討端對端(End-to-End)的節錄式摘要方法於語音文件摘要任務上的應用，並深入研究如何改善語音文件摘要之成效。因此，我們提出以類神經網路為基礎之摘要摘要模型，運用階層式的架構及注意力機制深層次地理解文件蘊含的主旨，並以強化學習輔助訓練模型根據文件主旨選取並排序具代表性的語句組成摘要。同時，我們為了避免語音辨識的錯誤影響摘要結果，也將語音文件中相關的聲學特徵加入模型訓練以及使用次詞向量作為輸入。最後我們在中文廣播新聞語料(MATBN)上進行一系列的實驗與分析，從實驗結果中可驗證本論文提出之假設且在摘要成效上有顯著的提升。

Abstract

This thesis set to explore novel and effective end-to-end extractive methods for spoken document summarization. To this end, we propose a neural summarization approach leveraging a hierarchical modeling structure with an attention mechanism to understand a document deeply, and in turn to select representative sentences as

*中華電信研究院巨量資料研究所
Big Data Laboratory, Telecommunication Laboratories, Chunghwa Telecom Co., Ltd
E-mail: hane0131@gmail.com; muslim@cht.com.tw
+國立臺灣師範大學資訊工程研究所
Department of Computer Science and Information Engineering, National Taiwan Normal University
E-mail: hane0131@gmail.com; berlin@csie.ntnu.edu.tw
#台達知識管理部
Delta Management System
E-mail: journey.liu@deltaww.com

its summary. Meanwhile, for alleviating the negative effect of speech recognition errors, we make use of acoustic features and subword-level input representations for the proposed approach. Finally, we conduct a series of experiments on the Mandarin Broadcast News (MATBN) Corpus. The experimental results confirm the utility of our approach which improves the performance of state-of-the-art ones.

關鍵詞：語音文件、節錄式摘要、類神經網路、階層式語意表示、聲學特徵
Keywords: Spoken Documents, Extractive Summarization, Deep Neural Networks, Hierarchical Semantic Representations, Acoustic Features

1. 緒論 (Introduction)

隨著大數據時代的來臨，巨量且多元的資訊透過網際網路快速地在全球各地傳播，資料內容的呈現方式已不侷限於傳統的紙本形式，包含語音及影像的多媒體資訊逐漸取代靜態的文字資訊，如何有效率地閱讀多樣化形式的多媒體資訊，已成為一個刻不容緩的研究課題。此外，在社會逐步行動化的情況下，人手一機已是常態，且伴隨著科技不斷地創新，行動設備不再只能通話和傳遞文本訊息，多媒體訊息如語音及影像等亦能完好地傳遞，更甚於我們能透過聲音及手勢等指令操作設備。

在眾多的研究方法中，自動摘要 (Automatic Summarization) 被視為是一項關鍵的技術，其在自然語言處理 (Natural Language Processing, NLP) 領域中一直都是熱門的研究議題，因其具有能擷取文件重要資訊的特性，在許多應用上更是不可或缺的一項技術，如問答系統 (Question Answering)、資訊檢索 (Information Retrieval) 等。另一方面，語音是多媒體文件中最具語意的主要成份之一，如何透過語音(文件)摘要技術有效率地處理時序資料，更是顯得非常重要。其關鍵在於影音文件往往長達數分鐘或數小時，使用者不易於瀏覽與查詢，而必須耗費許多時間閱讀或聆聽整份文件，才能理解其內容，不符合人們想要快速地獲取資訊之目的。

對於含有語音訊號的多媒體資訊，我們可先經由自動語音辨識 (Automatic Speech Recognition, ASR) 技術將文件轉成易於瀏覽的文字內容，再透過文字文件摘要的技術作處理，以達到摘要語音文件之目的。但因現階段的語音辨識技術仍存在辨識錯誤的問題，也缺乏章節與標點符號，使得語句邊界定義模糊而失去文件的結構資訊；此外，語音文件通常含有一些口語助詞、遲疑、重複等內容，進而使得語音摘要技術的發展更為艱鉅。

本論文主要探討端對端的節錄式語音文件摘要任務常見的自動摘要技術大致上可分為兩種，節錄式 (Extractive) 摘要與重寫式 (Abstractive) 摘要。節錄式摘要方法是本論文的研究重點，其主要會辨別文章中的語句是否具代表性，並依照特定的摘要比例從其中選取作為摘要；重寫式摘要方法則需理解文章後，依文章的主旨重新撰寫摘要，其所使用的詞彙與文法不全然從原文中複製，與人們日常撰寫的摘要較為相似。

常見的語音文件摘要任務主要是分為兩階段，自動語音辨識（Automatic speech recognition, ASR）和自動文件摘要（Automatic document summarization）。當我們得到

一語音文件，自動語音辨識系統會先對語音訊號進行特徵抽取，進而透過預先訓練完成之聲學模型（Acoustic model）和語言模型（Language model）進行語音辨識得到其轉寫文件 （Transcription）。本論文中所使用的語音辨識系統，是採用國立臺灣師範大學資訊工程學系研究所語音暨機器智能實驗室所發展之大詞彙語音辨識器（Large vocabulary continuous speech recognition system, LVCSR）(Chen, Kuo & Tsai, 2004; 2005) 進行自動語音辨識。常見的節錄式文件摘要方法大多是以資料驅動 (Data-driven) 方法為主。其中，又以深度學習 (Deep Learning) 方法發展出的序列對序列 (Sequence-to-Sequence) 架構 (Bahdanau, Cho & Bengio, 2015; Sutskever, Vinyals & Le, 2014)在摘要任務上獲得較多學者的青睞。尤其重寫式摘要被認為是一種序列對序列的問題(Sutskever *et al.*, 2014)，更以此發展出許多方法(Chen, Zhu, Ling, Wei & Jiang, 2016; Chopra, Auli & Rush, 2016; Nallapati, Zhou, dos Santos, Gùlçehre & Xiang, 2016; Paulus, Xiong & Socher, 2017; Rush, Chopra & Weston, 2015; See, Liu & Manning, 2017; Tan, Wan & Xiao, 2017)；而節錄式摘要一般則被視為一種序列標記 (Sequence Labeling) 的問題，對文章中每個語句作標記，標示出其是否為摘要(Cheng & Lapata, 2016; Nallapati, Zhai & Zhou, 2017)。

雖然語音辨識的錯誤對於語音文件摘要任務上會有一定的影響，其主要的影響在於自動轉寫文件中的內文會與人工轉寫結果有差異，進而導致文件摘要系統無法完全準確地理解文件含義，因此使得摘要成效不佳；此外，摘要的呈現亦是一項重要的課題，如何呈現出易於閱讀的摘要，是文件摘要系統中必須學會的重點。而一個良好的摘要表達應該著重於以下四個要素：

- **資訊性（Informativity）**：摘要結果所包含原文件中的資訊程度，應盡可能涵蓋所有重要資訊。
- **文法性（Grammaticality）**：摘要中的語句應符合語言的文法，所得之摘要才易於閱讀；若不符合文法，則會較常被視為關鍵詞擷取（Keyword Extraction）。此要素於重寫式摘要任務上較受關注。
- **連貫性（Coherency）**：此要素所指的是摘要中上下文間的連貫程度，若前後句不存在連貫性，則會類似於畫重點的方式條列出重點，而非根據文件主旨所生成之摘要。此要素於節錄式摘要任務上常被提及。
- **非重複性（Non-Redundancy）**：為了能簡化描述，應避免出現過多重複的詞句或相似的資訊，若重複的資訊太多會影響使用者閱讀。

因此本論文主要會針對上述之資訊性及連貫性兩項要素討論，並嘗試以不同方法避免受到語音辨識錯誤的影響。首先於摘要資訊性部分，本論文發展並改進一個端對端的階層式類神經網路架構，其受益於摺積式類神經網路（Convolutional neural networks, CNNs）之語言模型應用以及遞迴式類神經網路（Recurrent neural networks, RNNs）於自然語言處理領域的優秀表現，使得我們能夠階段式（先語句後全文）地閱讀文件並快速地理解語意；另外我們亦嘗試應用注意力機制（Attention mechanism）更進一步提升模型對於文章的理解度，進而提升摘要資訊性。其次對於摘要連貫性，由於節錄式摘要往

往是挑選較符合摘要語句的結果，因此其通常沒有根據語意進行排序，因此本論文亦嘗試將摘要語句的排序及摘要評估指標應用於強化學習（Reinforcement learning, RL）輔助模型訓練。最後為了避免語音辨識錯誤，我們在模型預測摘要的過程中參考語句的聲學特徵（Acoustic features）及次詞資訊（Subword information），其中前者包含原語音文件中的語音特性，可改善兩階段語音文件摘要系統上，進行摘要時無法參考之原語音特性；而後者則是為了改善前述之詞彙辨識錯誤，因辨識錯誤可能發生在詞彙中的部分區塊，而導致斷詞時無法辨別正確的詞彙，若使用次詞資訊則可以使用周邊資訊推測錯誤的部分其正確的語意。

2. 文獻回顧 (Related Work)

圖 1. 自動文件摘要的分類
[Figure 1. Category of Automatic document summarization]

自動文件摘要方法主要可依照四個面向分類（如圖 1），可依照來源、目的、功能及方法等細分為不同類型：

- **來源**：主要分為單文件與多文件，前者指針對單一文件擷取摘要，後者則是統整歸納多篇主題相近的文件重點產生摘要。多文件摘要通常會與查詢共同進行為以查詢為主之多文件摘要，同時進行檢索與摘要。
- **目的**：可分為一般性和查詢導向，一般性的摘要主要專注在文件中的主要重點；而查詢導向則會根據查詢字串決定其摘要內容，而查詢導向的摘要通常會與多文件摘要同時出現。
- **功能**：大多數摘要是資訊性的，主要專注在產生原文件的簡短版本，能保留其重要資訊；而較少數為指示性和批判性，此二者給予的摘要皆不包含原文的重要內容，前者會指出文件的題目或領域等詮釋資料（Metadata）；而後者則是會判斷整份文件是正面的還是負面的。

- **方法**：此分類方式最為常見，可概分為三種：

  ■節錄式摘要（Summarization by extraction）

  ■重寫式摘要（Summarization by abstraction）

  ■語句壓縮式摘要（Summarization by sentence compression）

  節錄式摘要與重寫式摘要之差異在於其產生摘要的原理不同。節錄式摘要是依據固定之摘要比例(Summarization ratio)，從原文件中選出重要性高的語句、段落或章節簡單組合成摘要。摘要比例是指摘要長度與原文件長度的比例，一般我們通常選用10%的摘要比例，也就是摘要長度為原文件長度的 10%。而重寫式摘要主要會依原文件中的完整概念，重新撰寫出摘要，因此摘要內容中可能還有非原文件中所使用但不影響其語意的詞語。綜上所述，我們可以(Torres-Moreno, 2014)之示例簡單描述節錄式摘要與重寫式摘要的優缺，以學習者為例，一個好的學習者在撰寫摘要時會先閱讀過整篇文章，再以自己的方式撰寫，而得之摘要內容能前後通順且符合文章旨意；而不好的學習者在撰寫摘要時，只會大略看過文章，並且挑選出「可能」重要的語句，組合在一起作為摘要。但此方法得到之摘要可能包含某些不相關的內容，且語句間的銜接可能會有內容不連貫或不通順的情況發生。除了較常見的節錄式摘要及重寫式摘要外，語句壓縮式摘要比較特別一點，主要用於將語句長度縮減，此方法可與節錄式摘要共同使用，而目前通常會將此方法歸類為重寫式摘要的一部分。

本論文主要專注於一般性單文件節錄式摘要的研究。此外摘要亦可針對文件形式分類，如常見的文字文件(Text documents)及包含語音資訊的語音文件(Spoken documents)，針對不同文件形式，所使用的摘要模型細節也應有所變化。文字文件摘要係指一般以文字內容為主的文件產生之摘要，大部分的摘要研究都屬於文字文件摘要；而語音文件摘要則是使用含有語音資訊的文件，通常是透過語音辨識後得到的轉寫文件，其中可能會含有一些語音辨識產生之錯誤，以及口語上無意義的資訊。因此，語音文件摘要會比文字文件摘要更為困難，反之，語音文件包含語音資訊，可以提供摘要方法更多有意義的資訊，能有效地抵銷其辨識錯誤。

此外，有鑒於深層學習的蓬勃發展，現今的技術大多是以端對端的深層類神經網路架構為主。深層學習主要是模擬人類之學習模式，將深層類神經網路架構視為人類大腦神經系統，並輔以大量資料進行訓練，使其能夠學習如何解決該研究問題。其架構中主要學習的是輸入與輸出之間的關係，藉由將不同的輸入樣本投影至相同的空間中，我們即可在該空間中將每個輸入樣本對應至正確的輸出，進而得到正確的結果。因此後續之文獻探討將以端對端之深層學習方法為主。

2.1 節錄式摘要 (Extractive Summarization)

在節錄式文件摘要任務中，我們通常可以將其視為分類問題，因為我們要判斷文件中的語句「是否」為摘要。而分類問題在深層學習技術中是最基本的問題，但是節錄式摘要

任務還是有相當的難度，因為除了簡單的分類外，我們還需理解並解析出文件的重要資訊，才能知道哪些語句有機會成為摘要。

(Cheng & Lapata, 2016) 將節錄式摘要任務視為一種序列標記及排序問題，其方法主要的特色在於使用一階層式編碼器和含有注意力機制(Attention Mechanism)的解碼器。階層式的編碼器有兩層，第一層為摺積式類神經網路(Convolutional Neural Networks, CNNs)，是參考(Kim, 2014)的方法，使用 CNN 計算語句的向量表示；第二層為遞迴式類神經網路(Recurrent Neural Networks, RNNs)，將語句向量做為每個時間點的輸入，而將最後一個時間點的輸出視為文件的向量表示。此作法對於較長的文章而言是相當有效的，因為文章過長時，若單使用一個 RNN，則有可能會遺失掉許多重要的資訊。最後透過另一個 RNN 對每個語句進行標記，並使用預測出的分數進行排序，進而得到最後的摘要成果。此外，(Cheng & Lapata, 2016)還嘗試用節錄式的方法模擬出重寫式摘要，與前述標記語句的不同，主要是從原文件中挑選單詞後組合成摘要句，而生成之摘要相當不符合文法性也不通順，不過關鍵詞彙基本上都能涵蓋。以此得知，(Cheng & Lapata)的方法在語言理解(Language Understanding)及資訊擷取(Information Extraction)有不錯的成效。

除了(Cheng & Lapata, 2016)同時進行節錄式摘要與重寫式摘要的研究外，(Nallapati *et al*., 2017)提出的 SummaRuNNer 亦嘗試生成重寫式摘要。與(Cheng & Lapata, 2016)不同之處在於 SummaRuNNer 在節錄式摘要任務上，並非使用編碼-解碼器架構，僅是單純地建立兩層雙向 RNN 後便判斷語句標記為何。相似之處在於其 RNN 也是階層式的架構，第一層輸入為詞彙向量，第二層則是第一層輸出所得之語句向量。此種作法中使用的參數量較少，因此收斂速度也較為快速。除了節錄式摘要任務外，(Nallapati *et al*., 2017)也嘗試將最後一層預測標記，改為一個簡易解碼器用於重寫式摘要任務。此外，由於摘要任務使用之資料集一般是沒有摘要標記的，(Nallapati *et al*., 2017)提出一種貪婪法對每個語句標記摘要，這個方法能夠找到較好的摘要組合而非只是找單獨比對每句的重要性，亦有許多學者嘗試將此方法用於自身的任務上。

隨著近幾年強化學習(Reinforcement Learning)的熱潮，亦有學者將強化學習應用於節錄式摘要任務上，(Narayan, Cohen & Lapata, 2018a)為了解決前述之節錄式摘要沒有正確摘要標記的情況，因此加入強化學習。其主要架構是改良自(Cheng & Lapata, 2016)，不同之處在於其在第二層編碼器的語句輸入是以倒序方式輸入，因為大多數文件通常會將主旨置於較前面的段落，再加上 RNN 比較容易記得後面時間點資訊的特性，此方式能夠將重要資訊更清楚記得。(Narayan *et al*., 2018a)所使用的強化學習方法，是最基礎的策略梯度(Policy Gradient)，也就是透過計算得之獎勵(Reward)分數與模型訓練梯度加成，使其能夠往我們期待的方向進行訓練。(Narayan *et al*., 2018a)所使用的獎勵分數是使用預測摘要與標準摘要的評估分數，而此方法讓模型收斂速度增加，同時也提升準確度，是一項跳躍性地成長。

然而，對於節錄式摘要任務來說，模型對文件的理解應該要能達到支撐後續分類摘要語句的程度，意即模型所得之文件向量表示應完整涵蓋文件主旨。根據不同的撰寫方式，文件主旨可能分散於文件的不同部分，除去文件主旨的段落，文件的其他部分應為

支持主旨的相關論述。如何讓模型可以準確地理解文件主題呢？(Ren *et al*., 2017)針對此議題提出一個有效的方法，其在產生語句向量表示時，亦將前面的語句以及後面的語句與該句的相關性串接，同時放入一些與該句相關的人工特徵（語句長度、位置等），使得分類時能使用更具語意的語句向量。此方法之架構相當大，但得到之摘要效果也相當不錯。不過從實驗分析可以發現對於摘要結果有較多貢獻的部分大多在於人工特徵上，以此我們可以推論，類神經網路的學習仍需人工特徵輔助方可更加提升成效。

單單只讓類神經網路架構自動學習語句或文件向量表示的效果仍有限，若能加入一些相關的額外資訊輔助訓練，可以讓我們的方法更深入地學習到文件重要資訊。(Narayan *et al*., 2018b)提出在摘要方法中參考文件的標題資訊，可以讓我們的方法更快速地找到文件的主旨，而以此得到的文件向量表示也較能涵蓋文件主旨，因而能提升摘要的成效。而(Narayan *et al*., 2018b)主要用的基本架構是由(Narayan *et al*., 2018a)變化而成，差異在於其將額外資訊向量與語句向量共同用於判斷是否為摘要。此方法更是驗證類神經網路架構有額外資訊輔助能學習更好。

2.2 重寫式摘要 (Abstractive Summarization)

(Rush *et al*., 2015) 是最早將類神經網路架構應用於重寫式摘要的研究，其主要的架構是改良至 (Bahdanau *et al*., 2014) 提出的編碼解碼器 (Encoder-Decoder) 與注意力機制，亦稱之為序列對序列模型，並應用於重寫式摘要任務。注意力機制能讓輸入文件內容與輸出摘要中的文字作一個對應，能找到文件與摘要中詞彙間的關係。(Rush *et al*., 2015) 的架構與 (Bahdanau *et al*., 2014) 不同之處在於其並非使用遞迴式類神經網路作為編碼器與解碼器，而是使用最基本的前向式類神經網路 (Feed-forward Neural Networks) 結合注意力機制作為其編碼器，而解碼器則是基於(Bengio, Ducharme, Vincent & Jauvin, 2003) 提出的 NNLM 變化。此方法在語句摘要 (Sentence Summarization) 任務上得到相當優異的成效，因此也證實類神經網路能夠適用於重寫式摘要任務上。

隨著深層學習的快速發展，遞迴式類神經網路在序列相關任務上的成功亦漸漸廣為人知，因此(Chopra *et al*., 2016) 則提出一個遞迴式類神經網路的編碼解碼器架構，應用於語句摘要任務上。此方法主要是 (Rush *et al*., 2015) 的延伸，其編碼器使用摺積式類神經網路，而解碼器則使用長短期記憶 (Long Short-Term Memory, LSTM) (Hochreiter & Schmidhuber, 1997) 單元作為遞迴式類神經網路的基本單元。LSTM 是 遞迴式類神經網路演變的架構，因其具有三個閘門: 輸入閘 (input gate)、遺忘閘 (forget gate) 及輸出閘 (output gate)，以及一個記憶單元 (memory cell)，所以可以改善消失的梯度(Vanishing Gradient)問題，同時透過不斷更新記憶單元，能保留更多重要資訊，不會隨著時間太長而遺忘以前的資訊。

與此同時，(Nallapati *et al*., 2016) 從 (Rush *et al*., 2015) 和 (Chopra *et al*., 2016) 發想出許多架構，同時也解決許多重寫式摘要潛在的問題。基本的架構是跟(Bahdanau *et al*., 2014) 提出的序列對序列模型相似，同時也加入注意力機制，而與 (Chopra *et al*., 2016) 不同之處則是在於其編碼器與解碼器皆使用遞迴式類神經網路，且使用 (Cho *et al*., 2014)

提出的 Gated Recurrent Unit (GRU) 而非 LSTM，GRU 同樣具有閘門，但是僅有兩個，且沒有額外的記憶單元，但是整體的記憶效果是一樣的，訓練參數量減少很多，可以比 LSTM 更快速地建構和訓練。(Nallapati *et al*., 2016) 中提到在語言生成時會遇到未知詞 (Out-of-vocabulary, OOV) 問題，為了解決此問題，加入 Large Vocabulary Trick (LVT) (Jean, Cho, Memisevic & Bengio, 2014)，此技術是對每小批 (mini-batch) 訓練資料建立單獨的解碼用詞典，因此能夠讓詞典不會太大，同時又能在訓練的時候減少發生未知詞問題。除了基本架構外，還提出三種改良的版本，第一種是在輸入時加入一些額外的特徵，如：詞性、詞頻等；第二種則是在解碼器生成詞彙之前，加入一個控制器，控制解碼器是否要生成新詞或從輸入文件複製，此一機制是參考(Vinyals, Fortunato & Jaitly, 2015)提出的 Pointer Network 架構，當文件中有專有名詞出現時，但解碼器的詞典中可能沒有該詞彙，就需要從輸入資料中複製使用；最後則是將編碼器改成階層式的編碼器，一般的編碼器輸入都是整篇文章的每個詞彙，不考慮語句的分界，而階層式編碼器第一層的輸入一樣是整篇文章的每個詞彙，當遇到每個語句的結尾詞時，就會將此時的輸出向量視為語句的向量表示，並作為第二層的輸入，也就是說，第二層的輸入是文章中的語句，這種方法能夠得到更細部的文件資訊，也使得產生之摘要內容較符合文章主旨。雖然在 (Nallapati *et al*., 2016) 已經有嘗試將 Pointer Network 的想法結合進模型中，但是此種方法過於強硬，因為此控制器得到的結果僅能二選一。

因此 (See *et al*., 2017) 提出的架構能有效的解決此狀況，此篇研究提出的方法是以同時進行產生新詞與選取原有詞彙的動作，最後利用一機率值簡單線性結合兩者所得到的機率分佈，以此得到最終的詞典機率分佈，詞典中包含解碼詞典與輸入文件的詞彙。此外，(See *et al*., 2017)亦提出一種 Coverage 機制，此機制主要是為了解決在語言生成任務上容易出現 OOV 和重複詞的問題，其在每個時間點會將以前時間點得到的注意力分佈加總後作為一 coverage 向量，維度大小為編碼器的時間點數量，而後在當前時間點會參考此向量計算注意力分佈，同時也會將此向量和注意力分佈進行比較，找出每個維度最小值後加總便得到一 coverage 損失，之後會做為訓練時使用的懲罰值，讓模型可以將重複詞的機率降低。此研究所得到的摘要效果比以往的重寫式摘要優異許多，而實驗結果亦顯示摘要成果比較偏向於節錄式摘要，因為複製的比例比生成的比例高出許多，與此同時我們也發現節錄式摘要的成效仍比重寫式摘要更為顯著。

3. 階層式類神經摘要模型 (Hierarchical Neural Summarization Model)

我們將語音文件摘要問題視為一語句分類暨排序問題，以期能依文件主旨選出可能為摘要的語句，且同時能學習到摘要語句間有意義的排序，使得摘要內容能更流暢地表達文件主題及概念。因此，我們提出一基本架構，其中包含一階層式編碼器及一解碼器，亦稱之為語句選取器。階層式編碼器中主要有兩個階層，我們會先針對文件中的語句找到對應的語句表示，再從語句表示中學習到文件中的重要概念，亦可稱為文件表示；最後會將語句表示及文件表示皆放置於語句選取器中，使其能夠根據文件表示及語句表示，辨別及排序摘要句。

此外，為了避免摘要結果受到過多語音辨識錯誤的影響，我們嘗試加入聲學特徵和次詞向量輔助訓練；同時我們亦加入注意力機制和強化學習機制於模型訓練中，以期能增加摘要的資訊性。

3.1 問題定義及假設 (Problem Formulation)

首先我們將語音文件摘要任務定義為一序列標記問題，主要是針對文件中的語句進行摘要的標註。其中摘要類別可分為摘要和非摘要，分別以 1 和 0 表示，因此我們將任務目標定義為最大化類別機率，亦為最大化似然性，並可將目標函式定義為下式：

$$log\, p(\boldsymbol{y}|D,\, \theta) = \sum_{i=1}^{N} p(y_i|s_i,\, D,\, \theta) \tag{1}$$

當給定一文件 D 時，其為一語句序列 $(s_1, \dots, s_n)$，我們的方法會從 D 中選取 M 個語句經由排序後作為其摘要。對於每個語句 $s_i \in D$，我們會預測一分數 $p(y_i|s_i, D, \theta)$ ，作為判定是否為摘要的依據 $y_i \in (0, 1)$。之後會依照語句被視為摘要的分數 $p(y_i = 1|s_i, D, \theta)$ 對所有語句進行排序，取前 M 個語句作為此文件摘要。

對於每個語音文件，我們定義以下幾點假設：

- 語音資訊可透過額外的聲學特徵參考進模型訓練
- 使用字向量可有效改善語句表示的成效並抵銷語音辨識錯誤
- 摘要句可被其他非摘要句解釋
- 強化學習技術可訓練摘要之排序

後續我們會針對上述之假設對模型架構進行不同的改進，且會詳細闡述其動機。

3.2 基本架構 (Basic Architecture)

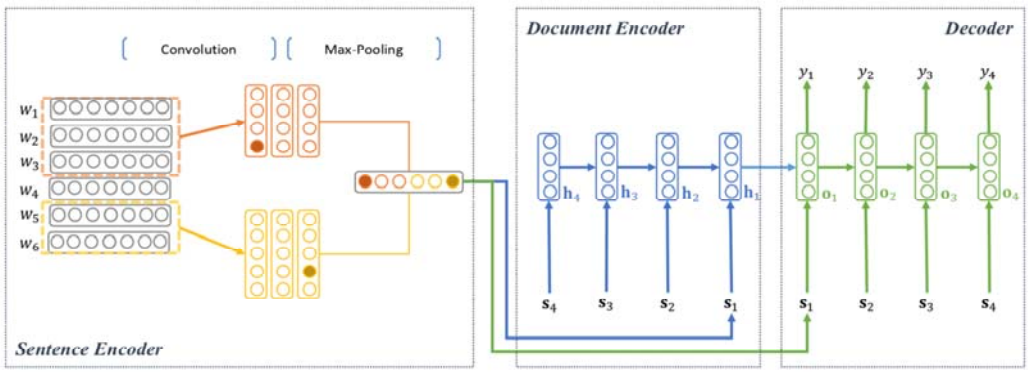


圖 2. 階層式類神經摘要模型 - 基本架構
[Figure 2. Basic architecture]

基本架構中包含一階層式編碼器及一解碼器，亦稱之為語句選取器。階層式編碼器中主要有兩個階層，我們會先針對文件中的語句找到對應的語句表示，再從語句表示中學習到文件中的重要概念，亦可稱為文件表示；最後會將語句表示及文件表示皆放置於語句選取器中，使其能夠根據文件表示及語句表示，辨別及排序摘要句。

3.2.1 語句編碼器 (Sentence Encoder)

我們利用摺積式類神經網路 (Convolutional Neural Networks, CNNs) 將每個不同長度的語句投影至向量空間，能夠得到固定長度的向量表示 (Representation)。在過去的研究中顯示，CNNs 在 NLP 領域的任務中有相當不錯的成效(Cheng & Lapata, 2016; Collobort *et al*., 2011; Kalchbrenner, Grefenstette & Blunsom, 2014; Kim, Jernite, Sontag & Rush, 2016; Lei, Barzilay & Jaakkola, 2015; Zhang, Zhao & LeCun, 2015)。我們使用 1-D 摺積 (Convolution) 並給定寬度 h 的摺積核 (Kernel) K，其定義為每次看 h 個詞彙，類似於 N 元模型 (N-gram) 的概念，可得到特徵圖 (Feature map) f。之後，對每個特徵圖沿著時序使用最大池化 (Max Pooling)，將特徵圖中的最大值視為語句特徵。為了能找到更好的特徵，我們使用多種寬度的摺積核，且每種寬度有多個不同的摺積核，最後將所得到的特徵串接在一起，即為語句的向量表示。

3.2.2 文件編碼器 (Document Encoder)

在文件編碼器中，我們使用遞迴式類神經網路 (Recurrent Neural Networks, RNNs)，將每個文件的語句序列轉換成一固定長度之向量表示，其能夠擷取到文件中的重要資訊。其中為了避免產生消失的梯度 (Vanishing Gradient) 問題，我們選擇使用 GRU (Gated Recurrent Unit) (Cho *et al*., 2014) 作為 RNN 的基本單元。此外，我們參考相關實作，將文件以倒序的方式作為輸入(Narayan, Papasarantopoulos, Cohen & Lapata, 2017; Narayan *et al*., 2018a; Narayan *et al*., 2018b; Sutskever *et al*., 2014)。由於我們使用的訓練語料是以新聞為主，而大多數新聞的主旨通常座落於開頭幾句，因此以倒序方式輸入文章，能使得 RNN 對重要資訊記憶更深。，因此可定義下列算式：

$$\mathbf{h}_i = f^e(\mathbf{h}_{i+1}, \mathbf{s}_i) \tag{2}$$

$$\mathbf{d} = \mathbf{h}_1 \tag{3}$$

其中 $f^e(\cdot)$ 為 RNN，$\mathbf{h}_i$ 是序列中每個時間點經過 RNN 運算後得到的隱藏層輸出，而 $\mathbf{s}_i$ 為語句向量。因輸入方式為倒序，所以每個時間點 $\mathbf{h}_i$ 都會參考後一時間點的輸出 $\mathbf{h}_{i+1}$ 及當前時間點的語句向量 $\mathbf{s}_i$ 。最後為了能得到整篇文章的隱含資訊，我們將最後一個時間點的輸出 $\mathbf{h}_1$ 視為文件向量 $\mathbf{d}$ ，並供之後摘要擷取時使用。

3.2.3 摘要選取器 (Summary Extractor)

我們的摘要選取器主要會將文件中每個語句標示為 1 (摘要) 或 0 (非摘要)。在此部分，我們將會使用另外一個 RNN，其中輸入一樣以語句向量為主，而語句向量同樣是經由語句編碼器所產生。此處與文件編碼器不同之處在於，摘要選取時是以文件的正序輸入，因此可定義成下列方程式：

$$\mathbf{o}_i = f^d(\mathbf{o}_{i-1}, \mathbf{s}_i) \tag{4}$$

$$\mathbf{o}_0 = \mathbf{d} \tag{5}$$

$$\mathbf{y}_i = \text{softmax}(\text{MLP}(\mathbf{o}_i)) \tag{6}$$

其中 $\mathbf{o}_i$ 為隱藏層輸出，$f^d(\cdot)$ 為一 RNN 架構，其輸入包含前一時間點的隱藏層輸出 $\mathbf{o}_{i-1}$ 和當前時間點的語句輸入 $\mathbf{s}_i$。為了在選取摘要時能參考到整篇文章的主旨，我們將初始的隱藏層 $\mathbf{o}_0$ 設定為文件向量 $\mathbf{d}$。此舉可以同時參考局部 (單一語句) 及整體 (文件) 的資訊，因此能更好的辨別語句。最後我們會透過 (6) 計算每個語句的類別 $\mathbf{y}_i$，其中 MLP(·) 為一簡單的前向式類神經網路(Feed-forward Neural Networks) 之後經由一個 softmax 函式得到語句類別的機率 $p(\mathbf{y}_i|\mathbf{s}_i, D, \theta)$ ，並依據 $p(\mathbf{y}_i = 1|\mathbf{s}_i, D, \theta)$ 將每個語句進行排序，依照固定的摘要比例選取排名高的語句作為完整的摘要結果。

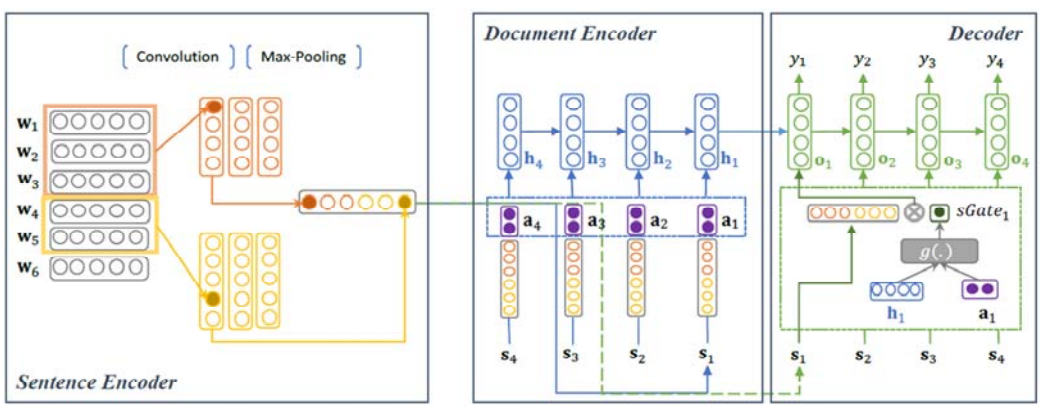


圖 3. 階層式類神經摘要模型 - 結合聲學特徵
[Figure 3. Basic architecture with acoustic features]

3.2.4 聲學特徵 (Acoustic Features)

為了能夠避免摘要結果受到辨識錯誤的影響，我們認為聲學特徵能夠保留每個文件的語音資訊且不受辨識錯誤之影響，因此提出三種方式將聲學特徵與上述架構結合，使得在判斷摘要的時候能夠參考，以得到更好的摘要成果。聲學特徵是以語句為單位，每個語

句會有對應的聲學特徵，因此令聲學特徵向量為 $\mathbf{a}$，我們的方法可定義下列方程式：

$$\mathbf{h}_i = f^{e\prime}(\mathbf{h}_{i+1}, [\mathbf{s}_i; \mathbf{a}_i]) \tag{7}$$

$$\mathbf{o}_i = f^{d\prime}(\mathbf{o}_{i-1}, [\mathbf{s}_i; \mathbf{a}_i]) \tag{8}$$

$$\mathbf{sGate}_i = g(\mathrm{W}_g[\mathbf{h}_i; a_i] + \mathbf{b}_g) \tag{9}$$

$$\mathbf{s}'_i = \mathbf{s}_i \odot \mathbf{sGate}_i \tag{10}$$

$$\mathbf{o}_i = f^{d\prime\prime}(\mathbf{o}_{i-1}, \mathbf{s}'_i) \tag{11}$$

全域向量(Global Embedding)

首先，我們將文件編碼器的輸入語句與其對應的聲學特徵串接，經過編碼後可得到新的文件向量，將 (2) 修改成 (7) 的方程式。我們認為此種做法同時考慮整份文件的聲學特徵，因此我們所得到的文件向量便可包含其聲學特徵，所以稱之為全域向量。

局部向量(Local Embedding)

其次，我們亦嘗試將語句對應的聲學特徵與語句向量串接後，直接用於摘要選取器之輸入，可修改 (4) 為 (8)。此方法使得聲學特徵向量能直接作用於摘要選取時的判斷，卻僅只作用於當前時間點及未來時間點，所以我們稱其為局部向量。

選擇向量(Selective Embedding)

最後一種方式與前面兩種比較不同，我們的想法來自(Zhou, Yang, Wei & Zhou, 2017)的選擇機制 (Selective Mechanism)，其概念主要是希望在生成摘要前可以先進行選擇的動作，預先篩選出可能成為摘要的語句，之後便能找到更準確的摘要。而在本論文中，我們希望透過聲學特徵能預先篩選出可能的摘要句。如 (9) 所示，我們將文件編碼器的輸出$\mathbf{h}_i$及對應之輸入語句的聲學特徵 $\mathbf{a}_i$ 串接，並作為 $g(\cdot)$ 的輸入。$g(\cdot)$ 是一個三層的前向式類神經網路，會得到 $\mathbf{sGate}_i$，其數值範圍在 0~1 之間，可視為語句被選的機率或權重。最後我們將語句 $\mathbf{s}_i$ 和 $\mathbf{sGate}_i$ 相乘後可得到新的語句向量 $\mathbf{s}'_i$，如 (10) 所示，並將其取代 (4) 的輸入 $\mathbf{s}_i$ 如 (11)，因此我們將此種方法稱為選擇向量。

3.2.5 次詞向量 (Sub-word Information)

在語音文件摘要中，常見的語音辨識錯誤大多是因為辨識時將詞彙辨識成同音的其他詞語，而使用此辨識結果進行摘要擷取時，會因為其中的詞彙錯誤導致上下文含義被誤判，因而找不到正確的文件主旨。因此本論文提出使用次詞向量輔助模型學習文件特徵表示以避免詞彙辨識錯誤導致之影響，原本的模型中是使用詞向量為最小單位組成文章，然而詞彙的辨識錯誤亦會影響到斷詞的結果，而語句中的特徵表示較容易受到錯誤的詞向

量影響，因語句中含有的詞彙相對較少；若改以次詞向量進行訓練，同時可以學習到詞彙的語意，亦能減緩受到詞彙錯誤的影響。過去亦有研究(Bojanowski, Grave, Joulin & Mikolov, 2017; Chen, Xu, Liu, Sun & Luan, 2015; Kim *et al.*, 2016)表示使用次詞向量亦能有效地表達文件，且能輔助詞向量訓練。

在本論文中，我們改良基本模型架構，加入一個輔助的語句編碼器（如圖 4），其中的設置與原有的語句編碼器相同。為了方便區隔，我們可將原有的語句編碼器稱為詞階段語句編碼器，而我們使用的次詞向量是字向量，可稱之為字階段語句編碼器。而前述之語句向量為 $\mathbf{s}_i$，我們將其表示為 $\mathbf{s}_i^w$，字階段語句向量則定義為 $\mathbf{s}_i^c$。在此架構中，我們希望能以字向量輔助詞向量訓練語句表示，因此我們定義以下方程式以更好地融合字與詞的向量資訊：

$$\mathbf{s}_i^* = f_s(\mathrm{W}_s^w \mathbf{s}_i^w + \mathrm{W}_s^c \mathbf{s}_i^c + \mathbf{b}_s) \tag{12}$$

其中 $\mathbf{s}_i^*$ 表示詞與字階段的語句向量融合後的語句表示，而W_s^w, W_s^c 和 $\mathbf{b}_s$ 為訓練用之參數，$f_s(\cdot)$ 為一單層的前向式類神經網路，能夠簡單地結合 $\mathbf{s}_i^w$ 和 $\mathbf{s}_i^c$，最後我們可將 (2) 和 (4) 的 $\mathbf{s}_i$ 代換成新的語句向量 $\mathbf{s}_i^*$ 進行摘要選取。

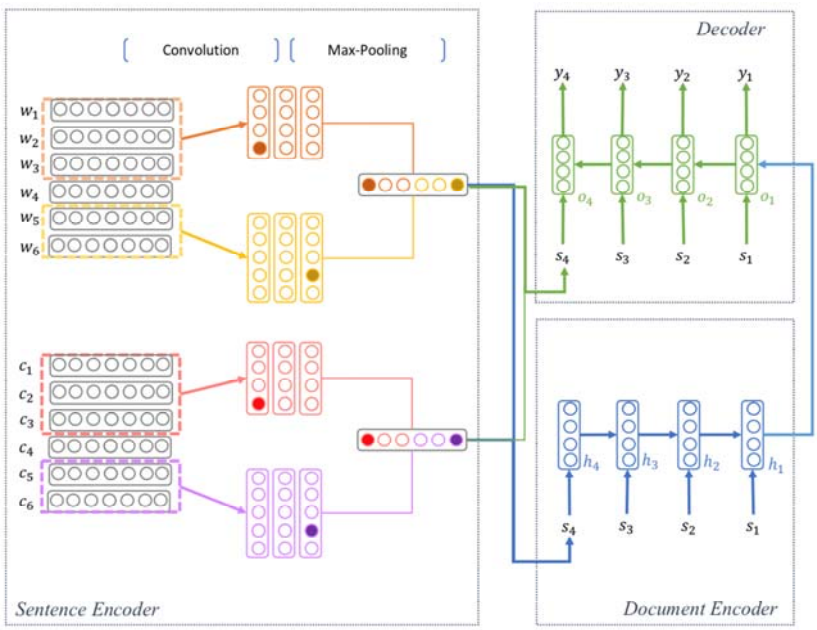


圖 4. 階層式類神經摘要模型 - 結合次詞向量
[Figure 4. Basic architecture with sub-word information]

3.2.6 注意力機制 (Attention Mechanism)

過去曾有學者(Ren *et al.*, 2017)表示，摘要主要是文件的簡短描述，而文件中的其他非摘

要語句則能夠細部地解釋摘要。文件的撰寫可概分為三種可能，第一種「通用到特定(General-to-specific)」所指的是文件開頭便簡短地描述文件內容，之後的內文皆是針對文件的細部闡述；第二種為「特定到通用(Specific-to-general)」，文件中先針對每個重點針對性地討論，最後作總結，因此主旨會落在文件後半部；最後一種則是「特定到通用到特定 (Specific-to-general-specific)」，所指的是文件先做細部討論，然後在中段破題點出文件主旨，之後再繼續討論細部的內容。從這三種情況中，我們可以知道文件中的語句和摘要句都有一定的關聯性，因此要找到摘要，文件中其他語句亦是必不可少。

對於文件摘要任務而言，值得注意的是摘要結果應該盡可能包含更多原文件中重要的資訊。因此，若我們希望摘要能夠包含更多重要資訊，應該要擷取出那些和文件中每個語句都有一定關聯性的語句，所以我們嘗試在我們的架構中加入注意力機制(Attention Mechanism)(Bahdanau *et al.*, 2015)。注意力機制可以找到每個語句與其他句的關聯性，因此我們可以將模型改良成如圖 5 的架構。

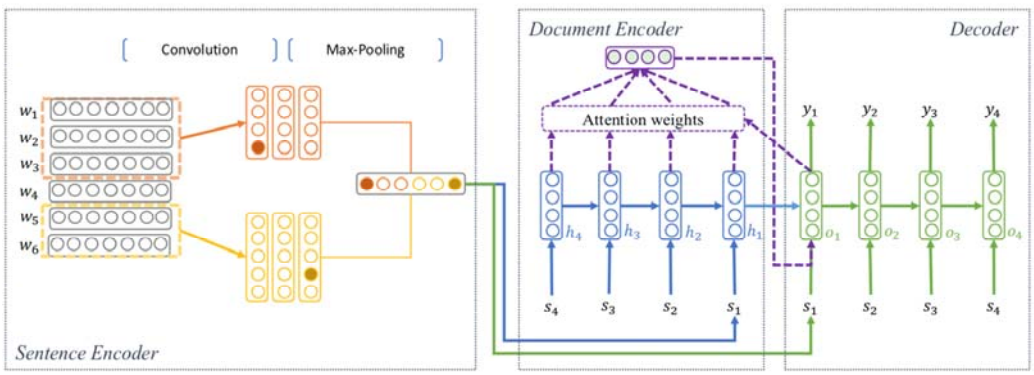


圖 5. 階層式類神經摘要模型 - 結合注意力機制
[Figure 5. Basic architecture with attention mechanism]

為了結合注意力機制，可以先簡單定義我們的摘要任務如下式：

$$p(\mathbf{y}_i|s_i, D, \theta) = m(\mathbf{s}_i, \mathbf{o}_i, \mathbf{c}_i) \tag{13}$$

其中 $\mathbf{c}_i$ 是透過注意力機制計算出的上下文向量，而 $\mathbf{o}_i$ 則是摘要選取器的隱藏層資訊，$m(\cdot)$ 代表整個摘要選取器，此式是表示摘要選取器的目標，主要是要預測語句的摘要類別機率 $p(\mathbf{y}_i|s_i, D, \theta)$。由於我們在摘要選取時結合注意力機制，因此可以重新定義 (4) 為下式：

$$\mathbf{o}_i = f^d(\mathbf{o}_{i-1}, \mathbf{s}_i, \mathbf{c}_i) \tag{14}$$

在每次 RNN 的計算中都會參考前一個時間的隱藏層資訊 $\mathbf{o}_{i-1}$ 、當前的語句向量表示 $\mathbf{s}_i$ 和該語句的上下文向量 $\mathbf{c}_i$，其中上下文向量主要是對文件編碼器的隱藏層資訊 $(\mathbf{h}_1, \dots, \mathbf{h}_n)$ 進行加權：

$$\mathbf{c}_i = \sum_j^N \alpha_{ij} \mathbf{h}_j \tag{15}$$

而 α_{ij} 是文件編碼器的隱藏層向量 $\mathbf{h}_j$ 對應的權重，此權重是透過下式計算：

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k^N \exp(e_{ik})} \tag{16}$$

$$e_{ij} = a(\mathbf{o}_{i-1}, \mathbf{h}_j) \tag{17}$$

其中 e_{ij} 用來計算語句 $\mathbf{s}_j$ 跟語句 $\mathbf{s}_i$ 的關聯性，若 $\mathbf{s}_i$ 為摘要句，則其跟其他語句都應有一定的關聯性，而非僅跟部分有關。在 (17) 中 $a(\cdot)$ 為一簡單的前向式類神經網路用於計算語句間的關聯性分數，再經過一個 softmax 函數將其轉化為一 0~1 的數值如 (16)。因此在預測摘要語句的機率 $p(\mathbf{y}_i|s_i, D, \theta)$ 時，α_{ij} 能夠反應出語句之間的相關性，因而判定該語句是否被認定為摘要。

3.2.7 強化學習 (Reinforcement Learning)

傳統的摘要模型訓練目標一般都是使用最大似然評估 (Maximum Likelihood Estimation, MLE)，也就是要最大化 $p(y|D, \theta) = \prod_{i=1}^n p(y_i|s_i, D, \theta)$，因此會選擇交叉亂度 (Cross Entropy) 計算損失 (loss)，目標函式可定義為下列方程式：

$$L(\theta) = -\sum_{i=1}^{n} \log p(y_i|s_i, D, \theta) \tag{18}$$

但是此種方法有兩個主要的缺點，第一是因為我們所使用的評估指標與損失函數的定義不同，模型的訓練目標是要最大化似然性，但卻使用 ROUGE 來評估摘要的好壞。其中似然性的定義主要是根據出現的機率決定，而 ROUGE 則是比較模型摘要和參考摘要之間詞彙覆蓋率，兩者的定義完全不同，且大多數評估指標函式是無法進行微分的，因此不適用於訓練參數；第二則是因為我們定義節錄式摘要為語句分類問題，可是其通常被視為單類別分類問題 (One Class Classification, OCC)(Tax, 2001)，主要能被模型學習到的大部分是摘要句，而非摘要句其實不太能辨識（圖 6），因而造成訓練上的困難。

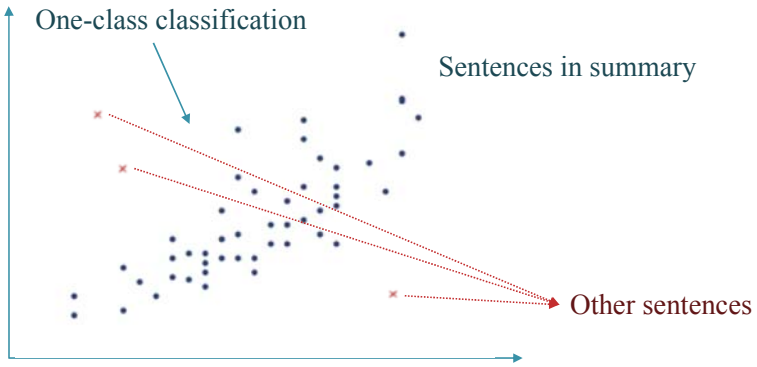


圖 6. 單類別分類問題示意圖
[Figure 6. One-class classification]

因此，我們使用強化學習(Sutton & Barto, 1998) 輔助模型訓練，由於基本的強化學習機制需要獎勵函數 (Reward Function)，此函數主要是用來判斷當前模型所預測的結果是否為正確，若正確則鼓勵訓練，反之則會懲罰。而獎勵函數的設定不像損失函數那麼嚴苛，因此我們使用摘要評估指標 ROUGE 作為獎勵函數，而訓練目標則可改成最小化獎勵期望值：

$$L(\theta) = -\mathbb{E}_{\hat{y}\sim p_\theta}[r(\hat{y})] \tag{19}$$

其中 p_θ 是指 $p(y|D,\theta)$，$r(\cdot)$ 是獎勵函數，而 $\hat{y}$ 是經過取樣 (Sample) 後得到的預測摘要。但是預測摘要 $\hat{y}$ 的可能性有無限多種，我們無法每次訓練都找到所有可能且計算其期望值來調整參數，這是很耗費成本的。因此我們將 (19) 改成 (20)，每次訓練只取一個樣本加速其訓練，並可將梯度 (Gradient) 函式改成 (21)，使其訓練上更為容易：

$$L(\theta) \approx r(\hat{y}) \tag{20}$$

$$\nabla L(\theta) \approx -r(\hat{y}) \sum_{i=1}^{n} \nabla \log p(\hat{y}_i|s_i, D, \theta) \tag{21}$$

4. 實驗結果 (Experimental Results)

4.1 實驗語料 (Corpus)

我們主要使用中文廣播新聞語料庫 (Mandarin Benchmark broadcast news corpus, MATBN)(Wang, Chen, Kuo & Cheng, 2005)。MATBN 是一個公開且常被應用於一些自然語言處理相關的任務上，如語音辨識(Chien, 2015)、資訊檢索(Huang & Wu, 2007)以及自動摘要(Liu *et al.*, 2015; Tsai, Hung, Chen & Chen, 2016)等。此資料集其中有 205 篇廣播新聞文件適用於摘要實驗，我們挑選其中的 20 篇作為測試集，餘下的 185 篇則為訓練集。

資料亦分成兩種，TD 為經過人工標註的文件，而 SD 則為經過自動語音辨識後產生的文件，因此 SD 會有部分的語音辨識錯誤。表 1 是對訓練集及測試集作的一些基本統計資料。此外，語音文件的聲學特徵類型列於表 2 中，是利用 Praat 工具擷取的結果，總計有 36 個特徵。

表 1. 用於摘要之廣播新聞文件的統計資訊[Tsai et al., 2016]
[Table 1. The statistics of MATBN]

	訓練集	測試集
文件數	185	20
每文件平均句數	20	23.3
每句平均詞數	17.5	16.9
每文件平均詞數	326.0	290.3
平均詞錯誤率	38.0%	39.4%
平均字錯誤率	28.8%	29.8%

此外，我們所使用之聲學特徵列於表 2 中，是利用 Praat 工具擷取的結果，總計有 36 個特徵，可簡單分為四種類型介紹：

- **Pitch 音高**：

  當我們在說話時，講到重點的時候，音高就會比較高來吸引注意，反之則會維持相對較低的音高。

- **Energy 能量**：

  能量一般是指語者的說話音量，通常都會被視為一項重要的資訊。當我們要強調某件事情時，除了音高會提高外，音量也會自然地放大，因而能幫助模型分辨重要資訊。

- **Duration 持續時間**：

  持續時間有點類似於一個語句中的詞彙數量，當持續時間越長沒有間斷時，則表示這句話包含的資訊相對較多。

- **Peak and Formant 峰與共振峰**：

  共振峰是頻譜中的峰值，主要用來描述人類聲道內的共振情形。如果聲音比較低沈，則共振峰會比較明顯，聽到的內容亦會較清晰；反之若聲音太過高亢，則共振鋒會比較模糊，同時聽到的內容也會比較模糊難辨。

表 2. 語音文件中每個語句對應的聲學特徵
[Table 2. List of acoustic features in MATBN]

聲學特徵	1. Pitch (min, max, diff, avg) 2. Peak normalized cross-correlation of pitch (min, max, diff, avg) 3. Energy value (min, max, diff, avg) 4. Duration value (min, max, diff, avg) 5. 1st formant value (min, max, diff, avg) 6. 2nd formant value (min, max, diff, avg) 7. 3rd formant value (min, max, diff, avg)

4.2 實驗結果 (Results)

首先本論文先比較過去的摘要方法於我們的資料集上的成效，之後在針對我們提出的架構和不同組合的摘要成果差異。

4.2.1 基礎實驗(Baseline)

過去 MATBN 資料集曾應用在各種不同的摘要方法上，從傳統的摘要方法(VSM, LSA)、非監督式類神經網路架構（SG, CBOW）到監督式類神經網路架構（DNN, CNN）都曾有學者使用。因此我們將過去的研究表現作為本論文比較的基礎實驗，結果列於表 3 中。

表 3. 基礎實驗結果
[Table 3. Results of baseline]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
VSM	0.347	0.228	0.290	0.342	0.189	0.287
LSA	0.362	0.233	0.316	0.345	0.201	0.301
SG	0.410	0.300	0.364	0.378	0.239	0.333
CBOW	0.415	0.308	0.366	0.393	0.250	0.349
DNN	0.488	0.382	0.444	0.371	0.233	0.332
CNN	0.501	0.407	0.460	0.370	0.208	0.312
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319

首先我們可以從表中發現傳統的向量空間模型（Vector space model, VSM）在文字文件和語音文件上的效果沒有差異太大，但文字文件的效果仍是比語音文件優異；另外我們可以將 VSM 跟 LSA 作一個簡單的比較，可以發現 LSA 的結果能很明顯的看出文字文件和語音文件的差異，同時也比 VSM 的效果好許多。

接著我們從非監督式類神經網路架構的結果觀察，SG(Skip-gram)和 CBOW 應用於

訓練詞向量的差異其實不大，因此在整體的摘要效果上兩者的差異其實並沒有很大，但 CBOW 相較於 SG 是較優異的，而此二者方法的效能亦超越傳統的向量模型許多。

最後我們針對監督式類神經網路架構作討論，DNN 是最基本的多層類神經網路架構，而 CNN 則是使用摺積式類神經網路架構，Refresh 是與本論文相似的階層式架構。其中在文字文件的效果上，可以很明顯地發現三者都超越了非監督式的方法，尤以 CNN 的效果最好，可能是因為 CNN 比 DNN 更能抓到重要資訊，而參數量又比 Refresh 少，較易於訓練；但在語音文件的成效上，三者都比非監督式的效果差，可能是因為其太過於依賴文件中的詞彙資訊，因而受到語音辨識錯誤的影響較為嚴重，導致其效果較差。

後續章節我們將以 Refresh 的數據與本論文提出之架構進行比較及分析。

4.2.2 階層式類神經摘要模型實驗 (Our models)

在實驗結果分析中，我們前面章節介紹模型時提到的副架構分開實驗，以下會列出不同實驗設置的效果，以及結果討論與分析。

I. 次詞向量

首先，我們先比較詞向量和字向量用於模型中的效果，如下表所示，可以看出單獨使用詞向量的結果在語音文件上的效果反而比單獨使用字向量的時候優異，但在文字文件上反而相反，這樣的結果與我們的假設有些許出入，可能是因為訓練文件中錯誤的字比較集中，因而無法透過周圍的資訊來學習正確的詞彙資訊；此外，若使用融合向量於我們的模型中，在語音文件的結果上可以有很明顯的進步，但在文字文件上僅於 ROUGE-2 有進步，因而我們認為字向量和詞向量之間可能仍有相輔相成的作用。

表 4. 階層式類神經摘要模型-次詞向量結果
[Table 4. Results of our model with sub-word information]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
詞向量	0.526	0.473	0.520	0.380	0.262	0.370
字向量	**0.544**	0.473	**0.535**	0.363	0.242	0.351
融合向量(詞+字)	0.543	**0.481**	0.533	**0.392**	**0.266**	**0.380**

II. 強化學習

承上所述，我們認為融合向量於語音摘要上有相當大的可能性，因此我們嘗試同時使用融合向量和強化學習於模型上，從表 5 中可以很明顯的看到強化學習於我們的方法中有一定的成效在，不過在文字文件摘要上有比較多的進步，主因可能是在於參考摘要不包含語音辨識錯誤，因此沒有辦法完全解決語音辨識錯誤的影響，若能將聲學特徵亦加入強化學習的獎勵函數中或許能改進此情況。

表 5. 階層式類神經摘要模型-強化學習
[Table 5. Results of our model with reinforcement learning]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
融合向量	0.543	0.481	0.533	0.392	0.266	**0.380**
融合向量+強化學習	**0.555**	**0.479**	**0.543**	**0.395**	**0.269**	0.379

III. 聲學特徵+強化學習

經過前面兩項實驗比較，我們可以發現融合向量可以解決部分的語音辨識錯誤影響，而強化學習則比較專注於摘要資訊性。因次我們嘗試於模型上結合聲學特徵與強化學習的方法，從表 6 中，我們可以發現在語音文件摘要上，效果比較顯著的是使用局部向量的方式結合聲學特徵；然而在文字文件摘要中，比較好的結果是使用全域向量。因此我們可以推論出聲學特徵對於人類轉寫的文字文件效用不彰，而對於自動辨識的語音文件上，還是有不錯的效果，但可能需要讓聲學特徵直接參與摘要選取的階段才能有效的提升效能。然而，整體的數據上仍是比前面的實驗差了許多，可能是模型上還需作更多細部的調整，或結合其他機制。

表 6. 階層式類神經摘要模型-聲學特徵+強化學習
[Table 6. Results of our model with acoustic features and reinforcement learning]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
無聲學特徵	0.479	**0.400**	0.469	0.352	0.226	0.342
全域向量	**0.486**	**0.400**	**0.473**	0.350	0.222	0.336
局部向量	0.478	0.399	0.469	**0.384**	**0.264**	**0.370**
全域向量+局部向量	0.464	0.373	0.453	0.350	0.224	0.336
選擇向量	0.448	0.371	0.439	0.350	0.213	0.334

IV. 次詞向量+注意力機制

因前一個實驗結果發現聲學特徵和強化學習共同訓練時效果相對較差，因此我們這次比較結合次詞向量和注意力機制的實驗結果。從表 7 中可以發現同時使用融合向量和注意力機制的效果在文字文件上較為優異，而在語音文件上仍是以詞向量的結果比較好。雖然整體的效果皆比之前的結果好，但可能是因為注意力機制訓練的主要是文件中語句之

間的關聯性，而對於語音文件而言，若辨識錯誤的太多，比較難找到語句間的語意關聯性，因而導致結果相對較差。

表 7. 階層式類神經摘要模型-次詞向量+注意力機制
[Table 7. Results of our model with sub-word information and attention mechanism]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
詞向量+注意力機制	0.523	0.472	0.519	0.401	**0.290**	**0.392**
字向量+注意力機制	0.535	0.477	0.529	0.368	0.245	0.356
融合向量+注意力機制	**0.567**	**0.496**	**0.557**	**0.402**	0.278	0.389

V. 次詞向量+注意力機制+強化學習

接續前一個實驗，我們加入強化學習機制於訓練中，實驗結果如表 8 所示。從結果可以發現，不管是文字文件還是語音文件，加入強化學習機制後，皆是在輸入為詞向量時會得到較好的效果。這有可能是因為我們的強化學習中獎勵函數使用 ROUGE 分數，而 ROUGE 計算時主要是以詞為基本單位，因而導致在其他情況下結果相對較差。

表 8. 階層式類神經摘要模型-次詞向量+注意力機制+強化學習
[Table 8. Results of our model with sub-word information, attention mechanism and reinforcement learning]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
詞向量+注意力機制+強化學習	**0.543**	**0.491**	**0.539**	**0.350**	**0.226**	**0.337**
字向量+注意力機制+強化學習	0.525	0.451	0.515	0.342	0.221	0.329
融合向量+注意力機制+強化學習	0.518	0.448	0.502	0.347	0.209	**0.337**

VI. 綜合比較

最後，我們將前述提到之架構做一個綜合比較，實驗結果如表 9 所示。其中我們可以發現當強化學習機制和注意力機制同時使用的情況下，不管是在文字文件還是語音文件上效果都相對較差。此種情況有可能是因為我們的注意力機制主要針對的是摘要資訊性提升，而強化學習中由於使用 ROUGE 分數作為獎勵函數，而 ROUGE 也是計算摘要資訊

性，因此當兩者同時訓練時，雖然都是針對資訊性，但可能因為太過注重而造成反效果。

其次，我們也嘗試結合注意力機制和聲學特徵的應用，如表 9 的最後兩列，由於前面的討論中發現使用局部向量方式結合聲學特徵在語音文件上會有較佳的效果，因此此部分實驗亦採用局部向量。實驗結果顯示加入聲學特徵在文字文件摘要上有些許的提升，但於語音文件摘要中沒有太大的影響，然而跟未加入聲學特徵訓練的實驗數據相比較，我們發現數據其實差異不大，此情況可能是因為此部分的實驗受到注意力機制的影響較顯著，聲學特徵對於此部分實驗不是其訓練的重點，因此沒有顯著的提升。

表 9. 階層式類神經摘要模型-綜合比較
[Table 9. Comprehensive comparison of our models]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
融合向量+注意力機制+強化學習	0.518	0.448	0.502	0.347	0.209	0.337
融合向量+注意力機制	0.567	0.496	0.557	**0.402**	0.278	0.389
融合向量+注意力機制+聲學特徵+強化學習	0.532	0.455	0.521	0.336	0.220	0.326
融合向量+注意力機制+聲學特徵	**0.569**	**0.507**	**0.561**	0.401	**0.288**	**0.394**

VII. 視覺化注意力

另外，我們亦針對注意力機制中的權重進行分析（圖 7），圖中每個列和行代表代表文件中的語句，每個列的語句標號旁括弧內的數值為 $p(\mathbf{y}_i = 1|s_i, D, \theta)$ ，即該句被辨識為摘要的機率。若該列中每欄的顏色越深，則代表該句和其他句的關聯性越大，則該句也被視為摘要，其中被紅框圈起的列為參考摘要。從紅框的部分看可以很明顯的發現，我們的摘要系統選出的摘要大部分和參考摘要相同，因此可驗證我們的注意力機制於摘要任務上真的有一定的成效。

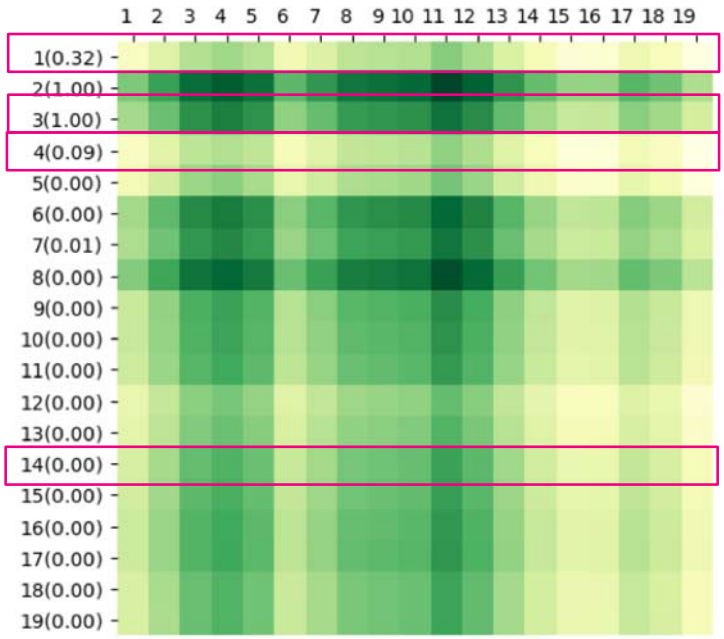


圖 7. 注意力機制權重視覺化
[Figure 7. Visualization of attention weight]

簡單總結整體實驗結果，我們提出之模型架構確實可有效提升語音文件摘要的成效，然而對於避免語音辨識錯誤的影響上，次詞向量和聲學特徵的效果仍有待加強；而注意力機制和強化學習等方法對於文字文件的效果仍比較顯著。因此若要實質性地提升語音文件摘要的成效，我們認為仍須從語音辨識的部分著手，若能不經過轉寫直接擷取摘要，或許更符合語音文件摘要，亦能有較優異的成效。

5. 結論與未來展望 (Conclusion & Future Work)

過去有關自動文件摘要的研究主要仍著重於文字文件摘要；而近年來由於大數據及機器學習技術的蓬勃發展，使得多媒體文件相關研究更為容易，因而逐漸有多媒體文件摘要相關研究的出現。雖然多媒體技術進步快速，但大多數的語音文件摘要方法仍多半由文字文件摘要方法延伸而來。直至近期隨著深層學習技術漸趨成熟，多媒體文件摘要技術也隨之成長。

順應深層學習的浪潮，本論文提出一種階層式類神經網路架構來從事語音文件摘要，同時亦適用於一般文字文件摘要。文件摘要任務可概分為節錄式與重寫式摘要。本論文旨在探討節錄式語音文件摘要方法。其中為了提升摘要資訊性及連貫性，我們加入注意力機制及強化學習技術；另外我們亦嘗試使用聲學特徵及次詞向量於模型訓練中，以避免計算摘要時受到過多語音辨識錯誤影響。經由一系列的實驗分析與討論，首先我們發

現注意力機制和強化學習皆可提升摘要資訊性，但同時使用時效果會相對較差；其次在避免語音辨識錯誤的部分，次詞向量與聲學特徵皆有不錯的成效，尤以次詞向量的效果較為顯著；最後對於摘要連貫性，我們的方法雖然有學習排序，但資料集中的參考摘要不包含排序資訊，因此無法完整地學習到語句間的連貫性。因此透過初步的實驗結果，足以證明我們提出的架構對於語音文件摘要有不錯的成效，但主要都反應於文字內容上，若要實質性的改善語音文件摘要的缺點，仍需更深入的探討。

承上所述，未來的研究我們可以針對幾個面向繼續深入。首先是應用預訓練語言模型於摘要研究上，改善語句或文章的語意表示，由於最近有許多預訓練語言模型已經使用相當大量的資料及高效能的設備進行訓練，且已被證明在許多任務上有相當亮眼的成績，僅需針對應用微調即可，或許可以嘗試進行深入研究；其次是重新整理資料集，因為摘要連貫性對於摘要亦是相當重要的指標，若成本允許，則可以僱請專家幫忙為資料進行重新標註，除了標註摘要語句外，同時亦加入摘要語句的順序，更有利於後續的摘要排序相關研究；再者，節錄式摘要亦可能發生語句間語意重複的情況，然而鮮少學者針對節錄式摘要重複性進行研究，因此為了減少節錄式摘要之重複性，或許可將重寫式摘要研究中常見之減少冗余的機制改良並應用於我們的方法上，應能得到更具意義的摘要結果；最後也最重要的是需要避免語音辨識錯誤影響語音文件摘要效果，從我們的實驗可以得出，現今的方法仍有所侷限，而為了有效地提升語音文件摘要準確性，或許我們能嘗試使用語音特徵如 Fbank 和 MFCC 等作為摘要系統之輸入，應可得到較原始的語音內容，亦能減少遇到辨識錯誤的情況，且因節錄式摘要是進行語句選擇，因此不需再進行轉寫，因而能使得摘要同為語音形式，但此想法需要多加考慮的部分在於難以評估結果正確與否，也相較兩階段的方法難實現，因此較少學者投入這方面的研究，若能實現我們的構想，應可使語音文件摘要技術達到新的高度，亦造福日後的學者們。

參考文獻(References)

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

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

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *TACL*, *5*, 135-146.

Chen, B., Kuo, J.-W., & Tsai, W.-H. (2004). Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing 2004*. doi : 10.1109/ICASSP.2004.1326101

Chen, B., Kuo, J.-W., & Tsai, W.-H. (2005). Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription. *International Journal of Computational Linguistics and Chinese Language Processing*, *10*(1), 1-18.

Chen, X., Xu, L., Liu, Z., Sun, M., & Luan, H. (2015). Joint Learning of Character and Word Embeddings. In *Proc. of IJCAI 2015*, 1236-1242.

Chen, Q., Zhu, X., Ling, Z., Wei, S., & Jiang, H. (2016). Distraction-based neural networks for modeling documents. In *Proc. of IJCAI 2016*, 2754-2760.

Cheng, J. & Lapata, M. (2016). Neural summarization by extracting sentences and words. In *Proc. of ACL*, 484-494. doi: 10.18653/v1/P16-1046

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., …Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. of EMNLP 2014*, 1724-1734. doi: 10.3115/v1/D14-1179

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proc. of NAACL-HLT 2016*, 93-98. doi: 10.18653/v1/N16-1012

Chien, J.-T. (2015). Hierarchical Pitman-Yor-Dirchlet language model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(8), 1259-1272. doi: 10.1109/TASLP.2015.2428632

Collobort, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, M., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493-2537.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780.

Huang, C.-L. & Wu, C.-H. (2007). Spoken Document Retrieval Using Multilevel Knowledge and Semantic Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *15*(8), 2551-2590. doi: 10.1109/TASL.2007.907429

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2014). On using very large target vocabulary for neural machine translation. In arXiv preprint arXiv: 1412.2007.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modeling sentences. In *Proc. of ACL 2014*, 655-665. doi: 10.3115/v1/P14-1062

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proc. of EMNLP 2014*, 1746-1751. doi: 10.3115/v1/D14-1181

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *Proc. of AAAI 2016*, 2741-2749.

Lei, T., Barzilay, R., & Jaakkola, T. (2015). Molding CNNs for text: Non-linear, non-consecutive convolutions. In *Proc. of EMNLP 2015*, 1565-1575. doi: 10.18653/v1/D15-1180

Liu, S.-H., Chen, K.-Y., Chen, B., Wang, H.-M., Yen, H.-C., & Hsu, W.-L. (2015). Combining Relevance Language Modeling and Clarity Measure for Extractive Speech Summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(6), 957-969. doi: 10.1109/TASLP.2015.2414820

Nallapati, R., Zhai, F., & Zhou, B. (2017). SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proc. of AAAI 2017*, 3075-3081.

Nallapati, R., Zhou, B., dos Santos, C., Gülçehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proc. of CoNLL 2016*, 280-290. doi: 10.18653/v1/K16-1028

Narayan, S., Cohen, S. B., & Lapata, M. (2018). Ranking Sentences for Extractice Summarization with Reinforcement Learning. In *Proc. of NAACL 2018*, 1747-1759. doi: 10.18653/v1/N18-1158

Narayan, S., Cardenas, R., Papasarantopoulos, N., Cohen, S. B., Lapata, M., Yu, J., …Chang, Y. (2018). Document Modeling with External Attention for Sentence Extraction. In *Proc. of ACL 2018*, 2020-2030. doi: 10.18653/v1/P18-1188

Narayan, S., Papasarantopoulos, N., Cohen, S. B., & Lapata, M. (2017). Neural extractive summarization with side information. In arXiv preprint arXiv: 1704.04530.

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. In arXiv preprint arXiv:1705.04304.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proc. of EMNLP 2015*, 379-389. doi: 10.18653/v1/D15-1044

Ren, P., Chen, Z., Ren, Z., Wei, F., Ma, J., & de Rijke, M. (2017). Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model. In *Proc. of SIGIR 2017*, 95-104.

See, A., Liu, P., & Manning, C. (2017). Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL 2017*, 1073-1083. doi: 10.18653/v1/P17-1099

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th Advances in Neural Information Processing Systems*, 3104-3112.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tan, J., Wan, X., & Xiao, J. (2017). Abstractive document summarization with a graph-based attentional neural model. In *Proc. of ACL 2017*, 1171-1181. doi: 10.18653/v1/P17-1108

Tax, D. M. J. (2001). *One-class classification: Concept learning in the absence of counter-examples.* Unpublished doctoral dissertation, Technische Universiteit Delft.

Torres-Moreno, J. M. (2014). *Automatic text summarization*. Hoboken, New Jersey: John Wiley & Sons. doi: 10.1002/9781119004752

Tsai, C.-I., Hung, H.-T., Chen, K.-Y., & Chen, B. (2016). Extractive Speech Summarization Leveraging Convolutional Neural Network Techniques. In *Proceedings of IEEE SLT 2016*. doi: 10.1109/SLT.2016.7846259

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer Networks. In *Proceedings of Advances in Neural Information Processing Systems 2015*.

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese Broadcast News Corpus. *International Journal of Computational Linguistics & Chinese Language Processing*, *10*(2), 219-236.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of Advances in Neural Information Processing Systems 2015*, 649-657.

Zhou, Q., Yang, N., Wei, F., & Zhou, M. (2017). Selective Encoding for Abstractive Sentence Summarization. In *Proc. of ACL 2017*, 1095-1104. doi: 10.18653/v1/P17-1101

Computational Linguistics and Chinese Language Processing
Vol. 25, No. 1, June 2020, pp. 57-80

應用多模式特徵融合的深度注意力網路進行謠言檢測[1]

Rumor Detection Using Deep Attention Networks With Multimodal Feature Fusion

王正豪*、黃靖幃*

Jenq-Haur Wang and Chin-Wei Huang

摘要

隨著社群平台蓬勃的發展，許多謠言與假訊息也充斥在社群媒體之中。現今各大社群平台大多是透過人工的舉報或統計的方式來進行謠言的分辨，這在資訊快速傳播的時代，非常缺乏效率。本論文提出一個結合圖像描述模型的多模式特徵融合方法，並透過深度注意力網路來進行謠言檢測。從 Tweets 中擷取出圖像、文字內容、與發文者的社群特徵後，首先，我們將圖像輸入圖像描述模型，透過 CNN 與 Seq2Seq 模型產生能描述該圖像的語句；其次，這些語句與文字內容串接，經過 word embedding 編碼後，以 Early 及 Late Fusion 兩種特徵融合方式，進一步結合社群特徵。最後，我們設計了多層 (Multi-layer)及多單元 (Multi-cell)雙向遞迴式神經網路 (BRNN)，並結合注意力機制賦予每個特徵不同的權重，以找出最重要的特徵並進行分類。實驗結果顯示，以 Early Fusion 融合所有特徵，使用基於 GRU 的多單元 (Multi-cell) BRNN 架構能達到最佳效果，F-measure 達 0.89，驗證了本論文所提出謠言檢測方法的有效性，未來將以更大量的資料進行實驗。

[1] 本論文承蒙審查委員提供諸多建議，謹此致謝。
* 國立台北科技大學資訊工程系
Department of Computer Science and Information Engineering, National Taipei University of Technology
E-mail: jhwang@csie.ntut.edu.tw

Abstract

With the rapid growth of information, browsing social media on the Internet is becoming a part of people's daily lives. Social platforms give us the latest information in real time, for example, sharing personal life and commenting on social events. However, with the vigorous development of social platforms, lots of rumors and fake messages are appearing on the Internet. Most of the social platforms use manual reporting or statistics to distinguish rumors, which are very inefficient. In this paper, we propose a multimodal feature fusion approach to rumor detection by combining image captioning model with deep attention networks. First, for images extracted from tweets, we apply Image Caption model to generate captions by Convolutional Neural Networks (CNNs) and Sequence-to-Sequence (Seq2Seq) model. Second, words in captions and text contents from tweets are represented as vectors by word embedding models and combined with social features in tweets with early and late fusion strategies. Finally, we design Multi-layer and Multi-cell Bi-directional Recurrent Neural Networks (BRNNs) with attention mechanism to find word dependency and learn the most important features for classification. From the experimental results, the best F-measure of 0.89 can be obtained for our proposed Multi-cell BRNN based on Gated Recurrent Units (GRUs) with attention using early fusion of all features except for user features. This shows the potential of our proposed approach to rumor detection. Further investigation is needed for data in larger scales.

關鍵詞：謠言檢測、遞迴式神經網路、注意力機制、圖像描述、特徵融合

Keywords: Rumor Detection, Bi-directional Recurrent Neural Networks, Gated Recurrent Unit, Self-attention Mechanism, Multimodal Feature Fusion

1. 緒論 (Introduction)

隨著社群網路快速發展，人們可以即時從各大社群平台獲取最新訊息。然而謠言及假訊息充斥其中，如何分辨訊息的真假，避免人們被誤導，是現今各大社群平台所面臨的重大問題。較著名的社群網站，如 Facebook、Twitter 等，針對謠言的辨別都有相應的處理機制。Facebook 利用公正的第三方機構對訊息進行人工驗證，得知訊息的真偽；而 Twitter 則利用自動評估系統與人工標記，標示具爭議或誤導資訊。然而，第三方驗證與人工標記等方法無法即時進行辨識，並阻止假訊息繼續傳播。因此如何快速又準確的辨識假訊息或謠言，已成為近年來熱門的研究議題。

在社群網路謠言檢測的相關研究中，大致可分為針對發文內容，以及透過分析社群網路的傳播結構兩種方法。首先，社群網路發文內容，包含有: 文字、圖像等，可以透過深度學習方法，例如: 遞迴式神經網路 (Recurrent Neural Networks, RNNs)，或卷積神

經網路 (Convolutional Neural Networks, CNNs)，來進行分類以辨識假訊息，但因為發文內容簡短，如果文件數量不足，其學習效果有限。其次，社群網路使用者之間可能有不同的關係，如: 好友、追隨、等，透過社群網路分析方法只著重在發掘關係結構，容易忽略文件內容所表達的資訊，導致謠言的辨識率不佳。有鑒於此，本論文提出一個結合圖像描述模組的多模式特徵融合方法，並透過深度類神經網路架構搭配注意力機制，以提升謠言偵測的準確率。首先，我們提出了利用圖像描述模型 (Image Captioning Model)，以 CNN 擷取圖像特徵，並透過 Sequence To Sequence (Seq2Seq)概念 (Sutskever, Vinyals & Le, 2014)，將圖像轉換為能夠表達圖像內容的文字。其次，我們設計了多層 (Multi-layer)以及多單元 (Multi-cell) 兩種雙向遞迴式神經網路(Bi-directional RNNs, BRNNs)，結合自注意力機制 (self-attention)，並利用 Early 及 Late Fusion 兩種特徵融合方法結合文字、圖像、及社群特徵，以提升分類準確率。雖然過去相關研究已經有論文提出多模式特徵融合的深度神經網路，結合文字、圖像、及社群等特徵進行謠言檢測，如: Jin 等人的作法 (Jin, Cao, Guo, Zhang & Luo, 2017)，其中文字內容透過 Long Short-Term Memory (LSTM)及注意力機制 (attention)，擷取特徵並計算出 attention 權重；而圖像內容則是直接以 CNN 架構取出特徵，並且將 attention 權重與圖像特徵直接進行 elementwise multiplication。然而，這樣的相乘並沒有具體可解釋的實質意義，因為 CNN 所取出的特徵向量與 LSTM 後 atttention 的特徵向量之間，維度不相同且兩向量各維度並沒有任何關聯。 而且現有方法的深度神經網路架構僅採用 LSTM 及 attention，隨著更多深層的神經網路模型不斷進步，仍有進一步改善的空間。因此在本論文所提出的方法中，我們結合了圖像描述模型，如: Vinyals 等人 (Vinyals, Toshev, Bengio & Erhan, 2015) 和 Xu 等人 (Xu *et al.*, 2015) 所提方法，先將圖像內容轉換成最可能的文字描述，以提升圖像特徵的語意，然後再與其他文字，進行 word embedding 及特徵融合；同時我們設計了多層(Multi-layer)及多單元 (Multi-cell)雙向遞迴式神經網路(Bi-directional RNNs, BRNNs)，結合自注意力機制 (self-attention)，並以 Gated Recurrent Unit (GRU)取代 LSTM，以提升分類效果。本文主要的貢獻為:

(1) 我們是第一個結合圖像描述模型的多模式融合謠言偵測方法，讓圖像內容的融合具有意義。比起現有作法，能有效提升準確率。

(2) 我們提出創新的多單元雙向遞迴式神經網路 (Multi-cell BRNN): 在 forward 及 backward 雙向的 RNN，以多個記憶單元 (memory cells)，同時進行序列資料的記憶與學習，能進一步提升效果。

實驗結果顯示，使用基於 GRU 的多單元雙向遞迴式神經網路 (Multi-cell BRNN)搭配注意力機制，可以使分類結果的 F-measure 達到 0.816；在進一步以 Early Fusion 融合社群特徵後，能達到最佳的謠言檢測率，F-measure 可達 0.89，驗證了本論文所提出方法的效果。後續我們在第二章介紹相關研究，第三章詳述研究方法，第四章則描述實驗結果及分析，第五章則是結論。

2. 相關研究 (Related Work)

隨著社群平台上大量使用者產生內容 (user generated content)的快速出現，謠言檢測已成為不可忽略的議題。不管是在 Facebook 或 Twitter，都提供錯誤資訊的檢測機制，以驗證使用者發文的真實性。Facebook 透過使用者與第三方檢查機構協助，對不實訊息進行標註，被標註的訊息會經過 FactCheck.org 和 Snopes.com 等第三方事實查核機構驗證。若經驗證確認為假訊息，則該訊息將會被公開。Twitter 則透過使用者的標註，以自動評估系統賦予每一則推文可信度等級。若可信度等級過低或該訊息內容被一定程度的用戶標示為假訊息，則判斷該訊息為謠言。然而，在這個資訊傳播快速的時代，第三方驗證與人工標記方法都無法即時辨別假訊息並阻止其繼續傳播。如何快速又準確的進行謠言檢測，即是本論文主要探討的議題。

謠言檢測的特徵來源主要可分為兩大類: 發文內容，以及傳播路徑。透過發文內容特徵擷取，以及發文被分享及再分享等行為，作為謠言檢測的特徵，並以機器學習方法訓練模型。例如: Castillo 等人 (Castillo, Mendoza & Poblete, 2011) 根據 tweets 的文字內容，再加上使用者的發文與 retweet 行為，以及引用外部來源等特徵，以 decision tree 來判斷 Twitter 的資訊可信度 (information credibility)。Gupta 等人 (Gupta, Zhao & Han, 2012) 以類似 PageRank 的方式進行 authority propagation，並且依據相似事件應該有相似可信度的想法，計算出可信度的值。

近年來人工智慧再度受到重視，大多謠言檢測相關論文都使用深度學習方法。例如: Ma 等人 (Ma *et al.*, 2016) 利用 RNN 來檢測 Weibo 與 Twitter 推文是否為謠言； Yu 等人 (Yu, Liu, Wu, Wang & Tan, 2017)和 Chen 等人 (Chen, Li, Yin & Zhang, 2018) 分別提出基於 CNN 的錯誤訊息識別卷積法(CAMI)與深度注意力機制，嘗試在發文早期判斷該推文是否為假訊息；Ma 等人 (Ma, Gao & Wong, 2018) 將立場檢測任務與謠言檢測任務整合，試圖透過判別訊息的立場來輔助假訊息的判斷；Jin 等人 (Jin *et al.*, 2017) 則是結合社群網路上的多媒體訊息，如: 文字、圖像、及社群特徵，其中文字內容透過 LSTM 及注意力機制，擷取特徵並計算出注意力權重；而圖像內容則是直接以 CNN 架構取出特徵，並且將注意力權重與圖像特徵直接進行 elementwise multiplication。然而，這樣的相乘並沒有具體可解釋的實質意義，因為 CNN 所取出的特徵向量與 LSTM 後注意力權重向量之間，維度不相同且兩向量各維度並沒有任何關聯。同時在該論文中，對於文字特徵的處理，僅使用 LSTM 進行特徵擷取，隨著更多深層的神經網路模型不斷進步，仍有進一步改善的空間。因此本論文針對以上兩點問題進行改善：首先，針對圖像特徵部分，我們使用一句短語來表達該圖像的內容，比起直接用 CNN 特徵向量來代表圖像，更能表達該圖像的語意。其次，針對文字特徵部分，我們使用雙向遞迴式神經網路(BRNN)結合注意力機制來取得發文內容字詞之間的關係，並試圖找出重點字詞，使後續謠言分類效果得以提升。

隨著電腦運算能力的提升與圖形處理器（Graphics Processing Unit, GPU）的發展，深度學習方法特別是各種類神經網路架構成為熱門的研究方法。CNN 最初是由 Yann

LeCun 等人提出 (LeCun, Bottou, Bengio & Haffner, 1998)。其概念是透過卷積網路層（Convolutional）與池化網路層（Pooling）使輸入的訊息可以保留更多的特徵，不像基本的神經網路只能取得輸入資料一個維度的特徵。CNN 通常用在處理圖像相關的任務，目前已經有許多不同變異的架構應用在各領域，例如著名的 VGG Net (Simonyan & Zisserman, 2015)與 GoogleLeNet (Szegedy *et al.*, 2015) 等架構都在解決傳統卷積層在特徵傳遞過程中，因為某些特徵不夠明顯而被忽略的問題。RNN 最初是由 Elman 所提出 (Elman, 1990)，後來被 Mikolov 等人 (Mikolov, Karafiát, Burget, Černocký & Khudanpur, 2010) 應用在自然語言處理中。RNN 的主要架構如圖 1 所示，是由單層隱藏層的神經網路不斷遞迴而成的。

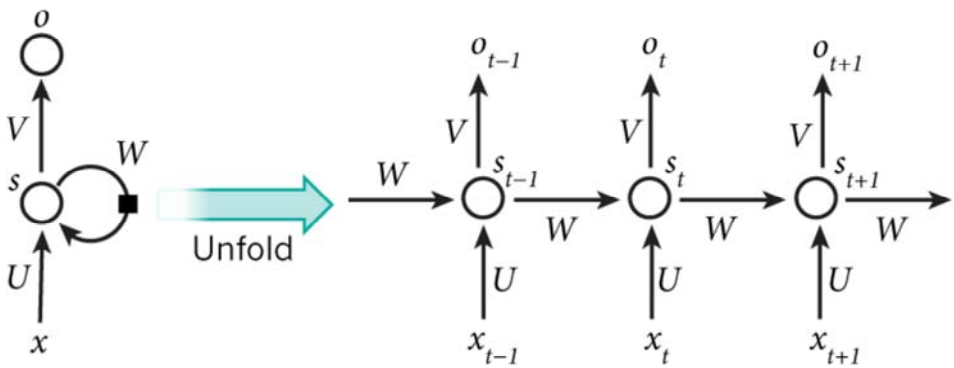
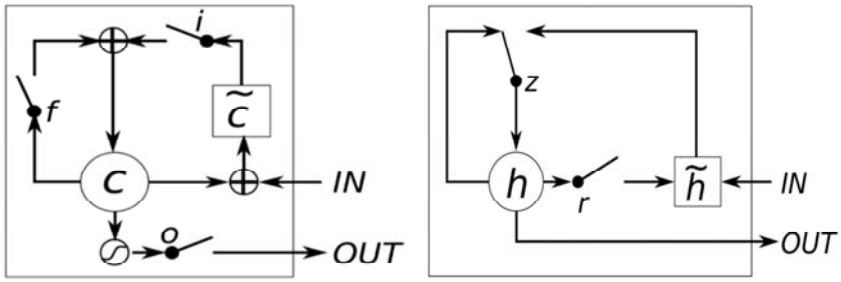


圖 1. 遞迴式類神經網路架構圖 (LeCunn et al., 2015)
[Figure 1. The architecture of recurrent neural networks (LeCunn et al., 2015)]

由上圖可得知，若輸入的資料是一串序列，則資料將會按照時間順序依次輸入至隱藏層，並將上一時間點隱藏層的輸出作為下一時間點隱藏層的輸入。透過這樣的方式，可以使每個時間點的輸出都能與上個時間點的輸入有關，讓神經網路能對整個序列順序進行記憶並學習。

然而，由於 RNN 是以 Back-Propagation Through Time（BPTT）的方式進行訓練與傳遞特徵，容易因為特徵權重的大小，影響下一層隱藏層輸出的資訊，進而導致神經網路可能無法學習到長時間的訊息，其問題稱之為梯度爆炸或梯度消失。目前已經有許多方式來解決該問題，其中最常見的方法就是利用長短期記憶神經網路（LSTM）進行改善。透過三個 Gate：Input Gate、Forget Gate、Output Gate，控制資訊的流動，確保特徵不會因為權重太小而被神經網路忽略。Cho 等人 (Cho *et al.*, 2014) 提出一個嶄新的架構，稱為 Gated Recurrent Unit (GRU)，則進一步簡化處理單元。經過 Chung 等人(Chung , Gulcehre, Cho & Bengio., 2014)的實驗與探討，發現 GRU 其不僅與 LSTM 一樣可以解決遞迴式神經網路的梯度爆炸與梯度消失問題，時間效率也比 LSTM 更好。LSTM 與 GRU 之架構分別如圖 2(a)及 2(b)所示：



(a) *(b)*

圖 2. (a) LSTM Cell 與 (b) GRU Cell 架構圖 (Chung et al., 2014)
[Figure 2. The architectures of an LSTM Cell and a GRU Cell (Chung et al., 2014)]

GRU 透過圖 2 中的 z（update gate）與 r（reset gate）共同控制當前時間點的隱藏狀態。Update gate 負責決定要將多少訊息傳遞到下一個時間點；Reset gate 負責決定要遺忘多少過去的訊息。GRU 現今已經廣為採用，成為解決遞迴式神經網路梯度消失的主流辦法。由於在各種不同的任務中，GRU 與 LSTM 常有不同的表現，而且 GRU 使用較少的 gates，架構簡單效率較佳，因此在本論文中，我們將比較基於 GRU 與 LSTM 的 RNN 架構，對於謠言偵測的效果。

Sequence To Sequence（Seq2Seq）概念最早由 Sutskever 等人所提出 (Sutskever *et al.*, 2014)，被應用在機器翻譯的任務。將輸入的句子（Sequence）經過學習，產生另一個句子（Sequence）。Seq2Seq 架構主要是由兩個遞迴式神經網路所組成，分別稱為 Encoder 與 Decoder，其架構如圖 3 所示:

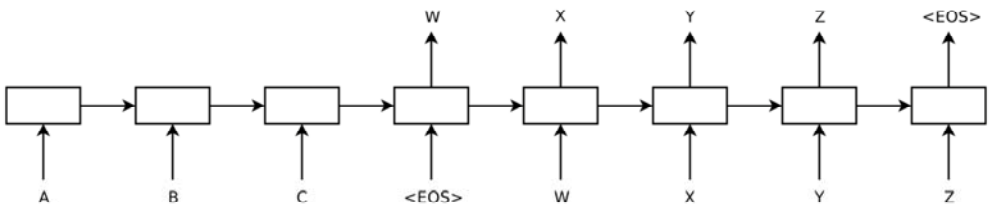


圖 3. Seq2Seq 架構示意圖，輸入”ABC”以產生”WXYZ” (Sutskever et al., 2014)
[Figure 3. The architecture of Seq2Seq Model, which outputs “WXYZ” for input “ABC” (Sutskever et al., 2014)]

在 Encoder 階段，RNN 不斷學習輸入 sequence 中的特徵，在遇到終止符號（<EOS>）時，類神經網路停止編碼並開始 Decoder 的階段，根據前面的記憶，產生一個代表該句子的向量（W），稱之為 context vector，再將它傳入 Decoder，訓練神經網路輸出最接近對應文件的向量，直到出現終止符號。Seq2Seq 架構被應用在許多情境，例如: Facebook 團隊的 Gehring 等人提出 ConvSeq2Seq (Gehring, Auli, Grangier, Yarats & Dauphin, 2017)，將 CNN 與 Seq2Seq 結合，以提升文件翻譯時的速度與準確率；Xing 等人(Xing *et al.*, 2017)

提出一個主題感知的 Seq2Seq 模型，並應用在聊天機器人中，為聊天機器人生成更多訊息豐富且有趣的回應。Zhao 等人 (Zhao *et al.*, 2018) 則基於 Seq2Seq 架構，結合 CNN 的圖像 encoder 與 LSTM 的文字 decoder，進行圖像描述。本論文應用類似想法，產生圖像所對應的文字描述特徵，以進行謠言檢測。

與 RNN 類似的，Seq2Seq 架構也會因為訊息過長而導致梯度消失的問題。雖然 LSTM 常被用來解決該問題，但其效果有限。透過注意力機制 (attention)，可以使神經網路在進行計算時，加強關注與輸入資訊相關的重點特徵，而不只是侷限在經過 RNN 計算後的隱藏向量。Mnih 等人 (Mnih, Heess, Graves & Kavukcuoglu, 2014) 首度將注意力的概念與 RNN 結合，應用在圖像分類任務之中。Bahdanau 等人 (Bahdanau, Cho & Bengio, 2015) 首先將注意力機制用在自然語言處理的任務上。結合雙向 RNN 之注意力機制如圖 4 所示：

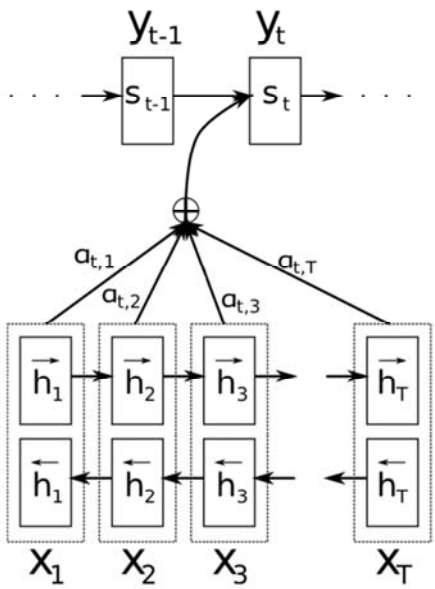


圖 4. 結合雙向 RNN 之注意力機制架構 (Bahdanau et al., 2015)
[Figure 4. The architecture of attention mechanism combining Bidirectional RNNs (Bahdanau et al., 2015)]

如圖 4 所示，輸入文件$X_1, X_2, \dots, X_T$之後，首先，先透過雙向遞迴式神經網路(BRNN)得到各個隱藏層的狀態$h_1, h_2, \dots, h_T$，其中h_j=$\overrightarrow{h_j^T}, \overleftarrow{h_j^T}$。假設當前 Decoder 的狀態為$S_{t-1}$，則輸入與輸出之間的關係可以表示為：

$$\overrightarrow{e_t} = (a(S_{t-1}, h_1), a(S_{t-1}, h_2), \dots, a(S_{t-1}, h_T)) \quad (1)$$

其中a為計算相關性的函數，例如內積或加權內積等。其次，透過 Softmax 函數，對 $\overrightarrow{e_t}$ 進行正規化，即得到注意力權重α_{tj}，定義為：

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})} \quad (2)$$

最後，將注意力權重與各個隱藏層狀態h_j進行加權運算，得出 Encoder 的輸出向量（context vector）$\overrightarrow{c_t}$，並傳入 Decoder 中，其公式為：

$$\overrightarrow{c_t} = \sum_{j=1}^{T} \alpha_{tj} h_j \quad (3)$$

RNN Decoder 的 hidden state 為 S_t，最後輸出為 y_t，S_t 由前一個 hidden state S_{t-1}，前一個輸出 y_{t-1}，以及$\overrightarrow{c_t}$經過函數 f 計算而得。本論文將 GRU 及 LSTM 等 RNN 架構結合注意力機制，以提升神經網路對重要特徵的關注程度，使謠言偵測的準確度得以提升。

3. 研究方法 (The Proposed Method)

本論文提出的方法主要可分為五大步驟，分別為：特徵擷取 (Feature Extraction)、圖像描述 (Image Captioning)、特徵融合 (Feature Fusion)、遞迴式神經網路 (Recurrent Neural Network)、注意力機制 (Attention Layer)，如圖 5 所示。

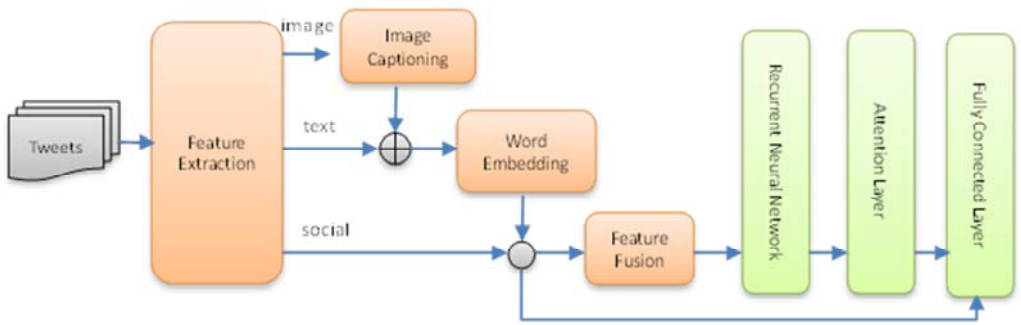


圖 5. 系統架構圖

[Figure 5. The system architecture of the proposed approach]

如圖 5 所示，Twitter 上的推文先經過 Feature Extraction，取得文字內容、圖像、與社群特徵。首先，將圖像特徵輸入圖像描述模組，經過卷積神經網路（Convolution Neural Network）與 Sequence to Sequence (Seq2Seq) 神經網路架構計算後，產生出描述該圖像的語句。其次，語句與文字內容串接，經過 Word Embedding 編碼，透過 Feature Fusion 與社群特徵融合。接著，融合後的特徵向量傳入雙向遞迴式類神經網路層（Bi-directional Recurrent Neural Network, BRNN），找出文字內容中各字詞之間的關係。我們以單層 BRNN 為基礎，設計出兩種不同的堆疊方式，分別為多層雙向遞迴式神經網路（Multi-layer BRNN）、多單元雙向遞迴式神經網路（Multi-cell BRNN）。最後，透過注意力機制（Attention Layer）的計算，加強推文中重要字詞的權重，並輸入一個全連接層（Fully Connected Layer），以進行假訊息的分類。

3.1 特徵擷取 (Feature Extraction)

一篇推文 (Tweet)中通常包含了文字敘述、圖像訊息以及社群資訊。首先，我們擷取出推文中的文字敘述，透過各種 RNN，嘗試找出文字內容的上下文關係。其次，通常推文常含有與文字內容相關的圖像，因此我們利用圖像描述模組，擷取圖像特徵並產生描述該圖像訊息的短句，以找出圖像中隱含的語意。此外，我們考慮各種社群特徵，包括推文的情緒極性、推文中的標籤(hashtag)、及發文者的使用者特徵等。使用者在推文中常會表達個人的意見或情緒，因此我們透過情緒分析模組，採用 SentiWordNet (Esuli & Sebastiani, 2006)對字詞進行情緒極性的擷取，透過計算出推文中每個字詞的情緒分數，加總平均後得出該推文所表達的意見傾向，包括: 正面、中立、負面。社群使用者也常透過 hashtag 標示本文重點主題，對分類可能有幫助。此外，我們也考慮使用者之間的互動作為使用者特徵，包括: 追隨、發文、與回覆等，為了與相關論文進行公平比較，在使用者特徵的部份我們採用與 Jin 等人 (Jin *et al.*, 2017) 相同的特徵，包括: 使用者在 Twitter 的朋友數量、追隨數量、追隨數量中是朋友的比例、總發文數量與是否有被 Twitter 認證等。最後結合使用者特徵、情緒特徵、與標籤特徵便構成社群特徵。

3.2 圖像描述 (Image Captioning)

我們參考 Vinyals 等人所提出的架構 (Vinyals *et al.*, 2015)，使用結合 CNN 與 LSTM 組成的 Seq2Seq 網路架構，產生出能描述該圖像的文字敘述。模型架構如圖 6 所示：

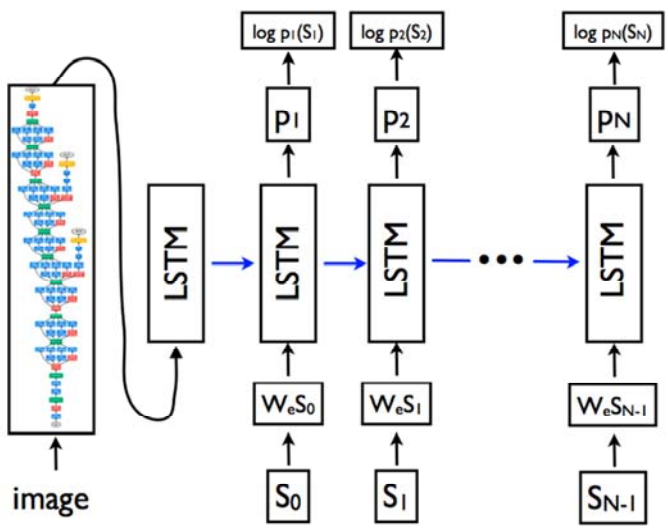


圖 6. 圖像描述模型架構圖 (Vinyals et al., 2015)
[Figure 6. The architecture of image captioning module (Vinyals et al., 2015)]

如圖 6 所示，圖像的部分採用 Google Inception Net V3 的 CNN 架構。此架構共有 42 層，共使用了 4 種不同維度大小的卷積核，可以取得圖像在不同尺度下的特徵，避免

一些細微特徵被忽略。經過 Inception 模組結取出來的圖像特徵向量會與經過 one-hot 編碼的文字敘述一同輸入 LSTM 運算。運算過程中圖像的特徵向量只會輸入一次，之後每個時間點依序輸入文字敘述中的字詞(S_i)，並輸出相關分數 p_i 最高的 k 個候選詞。而這些候選詞會分別輸入到下一個時間點，與當時所輸出的候選詞結合後，從中選出相關分數最高的字詞，再傳入下一個時間點。經過迭代後，最後一個時間點將會輸出一個融合了圖像訊息的完整文字描述。

3.3 特徵融合 (Feature Fusion)

在擷取了三種多模式的特徵，包括: 文字特徵、圖像特徵與社群特徵之後，我們提出特徵融合的方法來整合不同特徵。推文的文字特徵採取 one-hot 編碼的方法，用 300 維的向量來表示每個字詞的特徵。圖像在經過圖像描述模型轉換成描述語句後，將該語句轉換成與推文的文字特徵同樣 300 維的向量。我們也從推文的文字訊息中擷取出標籤(hashtag）與該推文所表達的情緒。在推文情緒方面，我們根據擷取的情緒分數，分為三種類別：正面、中立、負面。若該推文的情緒分數總分為大於 1，則它視為正面；若情緒分數總分小於 0，則視為負面；若情緒分數總分介於 0 到 1 之間，則視為中立。最後將上述的文字特徵、圖像特徵、以及情緒與 hashtag 兩者經過 one-hot 編碼後的向量進行串聯(concatenate)，即得到所有特徵的向量。

由於社群特徵與其他特徵差異很大，我們考慮兩種不同的特徵融合策略: 早期融合(early fusion)，和晚期融合 (late fusion)。在早期融合的策略中，我們同樣利用 one-hot 編碼，將社群特徵轉換成向量。為了讓社群特徵與圖文特徵能有相當的重要性，我們利用一個 autoencoder，將社群特徵壓縮為 300 維，並且串接在圖文特徵之後，以訓練分類器。而在晚期融合的策略中，我們先以圖文特徵輸入 RNN 和注意力機制，得到一系列的輸出，然後再將社群特徵以 one-hot 編碼轉換成向量，與圖文輸出結果一起輸入 Fully Connected Layer 進行分類。

3.4 遞迴式神經網路 (Recurrent Neural Networks)

本論文在 RNN 模組中使用 GRU Cell 取代傳統的 LSTM，並設計多層的 BRNN 堆疊的架構，以探討謠言偵測的效果。

單層雙向遞迴式神經網路（Bi-directional Recurrent Neural Networks, or BRNNs）最早是由 Schuster 等人提出 (Schuster & Paliwal, 1997)，分別將遞迴神經網路中每一個訓練序列分成向前傳遞（forward pass）與向後傳遞（backward pass）。兩者分別是獨立的單向 RNN，且兩個神經網路都連接到同一層輸出層，如圖 7 所示：

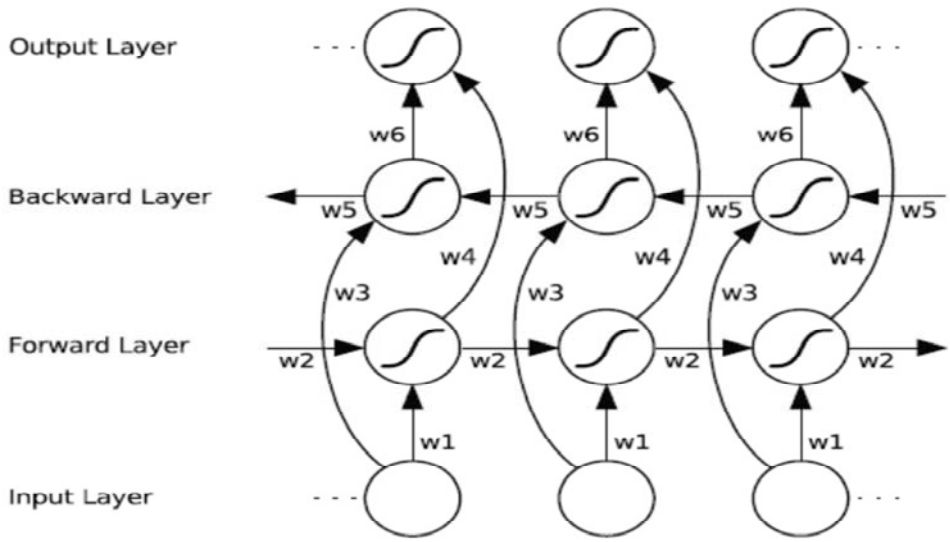


圖 7. 雙向遞迴式網路架構圖 (Graves, 2012)
[Figure 7. The architecture of bi-directional recurrent neural networks (Graves, 2012)]

對於雙向遞迴式神經網路的結果，我們透過串接的方式來組合向前與向後傳遞的輸出以表示一個字詞。

基於單層的雙向遞迴式神經網路，本論文設計了以下兩種不同的方式，進行多層雙向網路的堆疊。首先，多層雙向遞迴神經網路 (Mutli-layer BRNN)，是透過讓文字訊息經過多個回合的雙向遞迴式網路計算，強化文件中字詞之間的相互關係。架構如圖 8 所示：

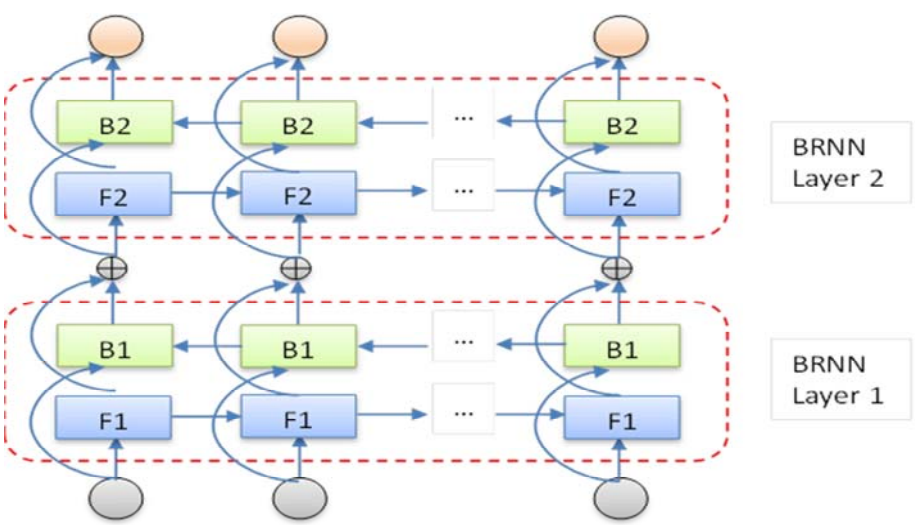


圖 8. 多層雙向遞迴式神經網路架構圖
[Figure 8. The architecture of Multi-layer BRNNs]

多層雙向遞迴式神經網路的輸入與輸出與單層 BRNN 一樣，透過在輸入層依序輸入文件中的字詞，得出代表該字詞的數值向量。其中，當每個字詞經過第一層的 BRNN 後，其得到的輸出已經包含了文件中向前傳遞與向後傳遞的資訊，因此當第一層 BRNN 的輸出傳入第二層時，不僅保留了原始文件中字詞之間的關係，每個字詞向量也記憶了經過第一層計算後的特徵。當下一層網路在計算時，由於其字詞之間的關係已經在上一層被找出，故不必再重新計算，使得網路能更快速的收斂，提升整體模型的計算效率。

其次，我們設計了另一種堆疊雙向網路的方法: 多單元雙向遞迴式神經網路 (Multi-cell BRNN)，透過增加 BRNN 中每個方向的單元數量，進行更深入的計算，當前 Cell 的輸出會作為下一層 Cell 的輸入，同一個神經元中的多個 Cell 同時進行序列資料的記憶與學習。架構如圖 9 所示：

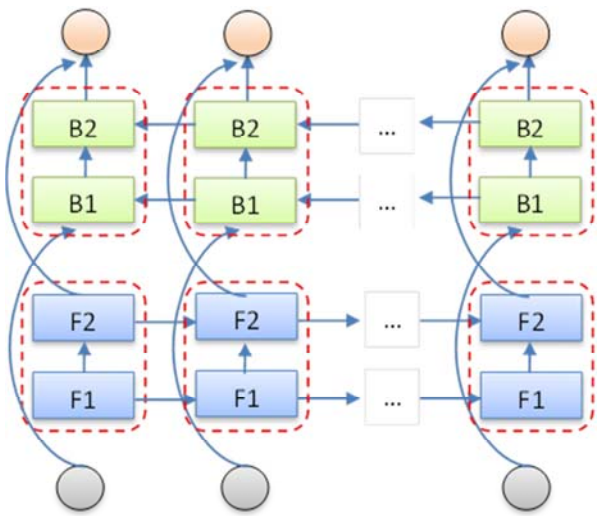


圖 9. 多單元雙向遞迴式神經網路架構圖
[Figure 9. The architecture of Multi-cell BRNNs]

多單元雙向遞迴式神經網路架構，在輸入層依序輸入文件中的字詞，並在輸出層得出代表該字詞的數值向量，如同上述的兩種架構。但由圖 8 可知，對於向前傳遞與向後傳遞而言，每個方向同一時間點的輸入經過一個 Cell 計算後，會再傳入下一個 Cell 繼續計算。該架構與多層雙向遞迴式神經網路不同的是，每個時間點的輸入只有考慮到單一方向的影響，向前傳遞與向後傳遞相對於整體架構來說，還是兩個獨立的神經網路架構，直到最後輸出時才進行整合。此方法能比多層 BRNN 更深入的計算每個字詞之間的關係，但相對的需要花費更多的資源與時間才會讓神經網路收斂。

3.5 注意力機制 (Attention Layer)

注意力機制常被用在遞迴式神經網路，加強關注與輸入資訊相關的重要特徵。在上述的各種神經網路架構中，我們結合自注意力機制 （Self-Attention）來計算文中每個字詞之

間的關係，以取得每則 Tweet 中文字特徵的重要資訊。Self-Attention 是一種注意力機制，與傳統的 Attention 機制差別在於，Self-Attention 不需要透過引入外部的資訊來找出較為重要的訊息，僅需要通過自身的訊息就能更新權重與參數，找出較重要的資訊。它的核心概念是 scaled dot-product attention 架構，是一種 dot-product attention 的變形，如圖 10 所示。

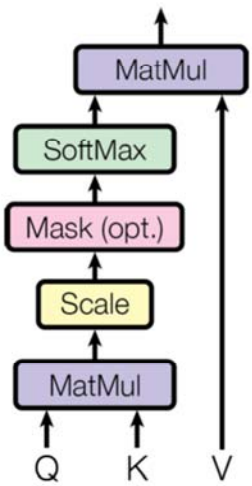


圖 10. Scaled Dot-Product Attention 示意圖 (Vaswani et al., 2017)
[Figure 10. Scaled Dot-Product Attention (Vaswani et al., 2017)]

經過 Vaswani 等人 (Vaswani *et al.*, 2017) 與 Tan 等人 (Tan, Wang, Xie, Chen & Shi, 2018) 的探討與比較，已證實該內積（乘法）注意力機制比使用單層神經網路的標準注意力機制 (Bahdanau *et al.*, 2015) 更有效率。

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

其中 d_k 為 key 的維度。當維度越大，Q 與 K 的內積也會越大，因此除以一個調整數 $\sqrt{d_k}$ ，防止該數值結果過大，最後透過 softmax 函數將結果正規化，將獲得的權重與 V 相乘，更新其向量的數值。

為了加速運算，我們採用 Multi-Head Attention 架構，如圖 11 所示：

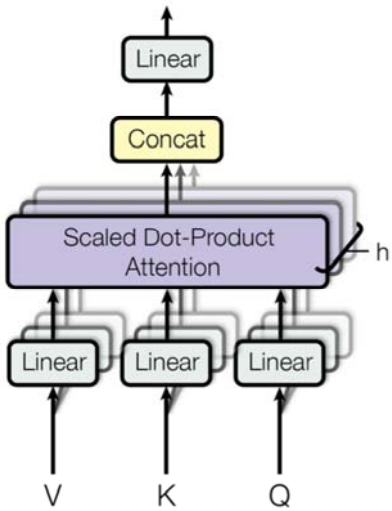


圖 11. Multi-Head Attention (Vaswani et al., 2017)
[Figure 11. Multi-Head Attention (Vaswani et al., 2017)]

經過 h 個不同投影線性轉換後，可以將 h 個 scaled dot-product attention 神經網路進行平行運算，並將每一次的結果進行串接，最後再經過一層線性轉換得到 multi-head attention 的結果。如下所示：

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O$$
$$where\ head_i = Attention\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right) \quad (5)$$

其中 $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W^O \in \mathbb{R}^{hd_v \times d_{model}}$。$i$ 表示為第幾個 scaled dot-product attention 網路，W_i^Q, W_i^K, W_i^V 皆為經過訓練的權重矩陣，W^O 為線性轉換層的權重矩陣，而 $d_k = d_v = d_{model} / h$ 則表示為矩陣的維度。

因為 self-attention 機制是對每個輸入的字詞與所有字詞進行計算，學習文件內部的結構與字詞之間的依賴關係，故計算每個字詞的最大路徑長為 1，即每個字詞都會被計算一次，不會像遞迴式神經網路一樣發生因訊息傳遞路徑過長導致過小的特徵被忽略，進而產生的梯度消失或梯度爆炸問題。在本論文架構中，input 經過各種 RNN，以及 self-attention 機制後，會計算出一連串的 output 向量，我們再將所有向量輸入到全連接層 (Fully Connected Layer)，進行最後的分類。

4. 實驗與討論 (Experiments and Discussions)

本論文採用的資料集分為兩大部份：圖像描述資料集與謠言檢測資料集。首先，在圖像描述方面，我們使用 Microsoft COCO 2014 (Tan *et al.*, 2018) 資料集，它在圖像相關任務中廣泛被使用，包含圖像識別與圖像特徵點檢測，如: Vinyals 等人 (Vinyals et al, 2015) 與 Xu 等人 (Xu *et al.*, 2015)。該資料集的每一張圖像資料都含有 5 句短語進行描述，每一句短語在資料集中皆為唯一。

其次，在謠言檢測方面，由於資料取得不易，每筆訊息都需要經過第三方機構公開認證其真偽，才能確定該訊息是屬於謠言或事實。本實驗採用 MediaEval 2015、2016 任務中所提供的 Twitter 謠言檢測資料集，已經過 Twitter 官方認證，其中也包含了每則推文的多媒體訊息與發文者的相關特徵。兩個資料集的分布狀況如表 1 與表 2 所示：

表 1. 圖像標註資料分布統計
[Table 1. Data distribution in image captioning dataset]

資料集	圖像數量/描述短句數量
Training Data	82783 / 413915
Test Data	36454 / 182270.

表 2. 謠言資料集資料分布統計
[Table 2. Data distribution in rumor dataset]

資料集	推文數量 (event)
Training Data	Real: 189 / fake: 157
Test Data	Real: 21 / fake: 24

在圖像描述的相關實驗中，我們採用雙語互譯評估（Bilingual Evaluation Understudy, BLEU）評估方法來評量圖像描述模型，BLEU 最早是由 IBM 的 Papineni 等人所提出的 (Papineni, Roukos, Ward & Zhu, 2002)，主要是用來評價模型的翻譯結果與參考文件是否相似。BLEU 的定義為: modified n-gram precision 的幾何平均 (geometric mean)，

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

$$BP = \begin{cases} 1 & if\ \ c > r \\ e^{1-r/c} & if\ c \le r \end{cases} \tag{6}$$

其中c表示譯文的長度，r表示參考文件的長度。

Modified n-gram precision 為 clipped n-gram 個數除以所有 n-gram 個數，

$$P_n = \frac{\sum_{C\in\{Candidates\}}\sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in C'} Count_{clip}(n-gram')} \tag{7}$$

其中 clipped n-gram 個數計算方式如下:

$$Count_{clip} = \min(Count, Max\_Ref\_Count) \tag{8}$$

而在謠言檢測相關實驗中，主要著重在二元分類任務，因此採用準確率（Accuracy）、精確率（Precision）、查全率（Recall）與 F-Measure 進行評估，並利用 T 檢定來比較不

同模型的差異程度。

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

在以下謠言檢測的實驗中，我們主要的 baseline 比較對象皆為 Jin 等人所提出的方法 (Jin *et al.*, 2017)。

4.1 圖像描述的效果 (The Effects of Image Captioning)

首先，為了探討圖像描述模型的效果，針對 MSCOCO 2014 的每一筆圖像資料經過圖像描述模型後，我們利用訓練資料集中的 3 句短文進行模型的訓練，其他 2 句短文進行驗證。為了實驗對照，我們也使用 Mediaeval 2015, 2016 中的圖像訊息進行訓練，BLEU 的實驗評估結果，如圖 12 所示:

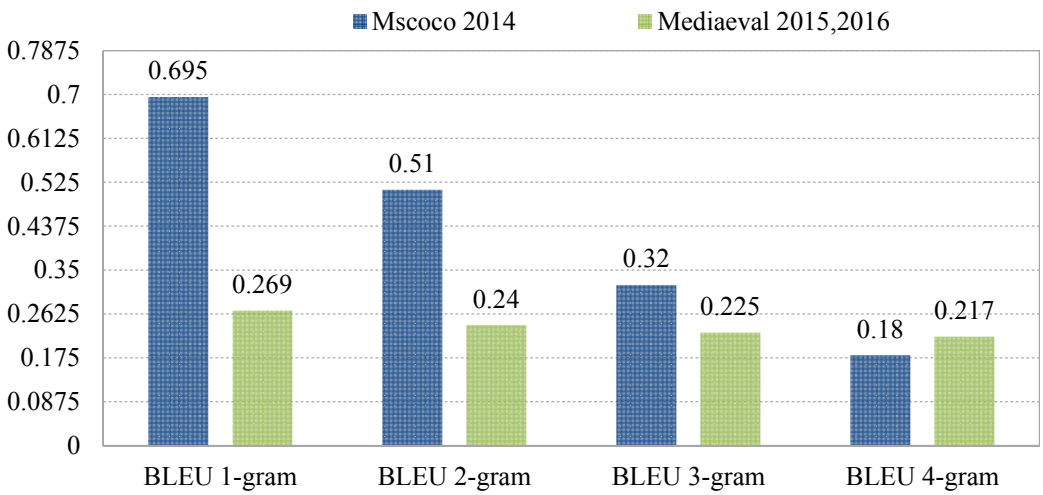


圖 12. 圖像描述實驗評估結果
[Figure 12. Experimental results for image captioning]

如圖 12 所示，以 MSCOCO 2014 資料集訓練的翻譯模型其 BLEU-1, BLEU-2 評估分數分別為 0.695 和 0.51，明顯優於 MediaEval 2015, 2016 所訓練的模型，且相當接近 Xu 等人的結果 (Xu *et al.*, 2015)。由於 Mediaeval 2015, 2016 資料集屬於 Tweets，礙於 Twitter 的資料特性，Tweets 中的文字訊息並不一定是在描述其中的圖像資訊，且發文者與回文者也不一定是客觀的對圖像訊息進行描述，這會使得參考文件不完整，無法訓練出良好的翻譯模型，因此導致最後其 BLEU 分數都偏低，故後續實驗將採用 MSCOCO 2014 資料集訓練出翻譯模型做為謠言檢測的圖像描述模組。

4.2 Word Embedding的效果 (The Effects of Word Embedding)

為了探討文字訊息的向量編碼方法對於謠言偵測的效果，我們使用兩種常見的方法進行比較：隨機初始化法，以及預訓練好的 Word2Vec 字典。前者是從 -1 ~ 1 之間隨機產生代表該字詞的數值向量，之後經過神經網路的訓練進行調整；後者則採用 GoogleNews 預訓練的 Word2Vec 字典對應的字詞向量進行訓練並更新。實驗結果如圖 13 所示：

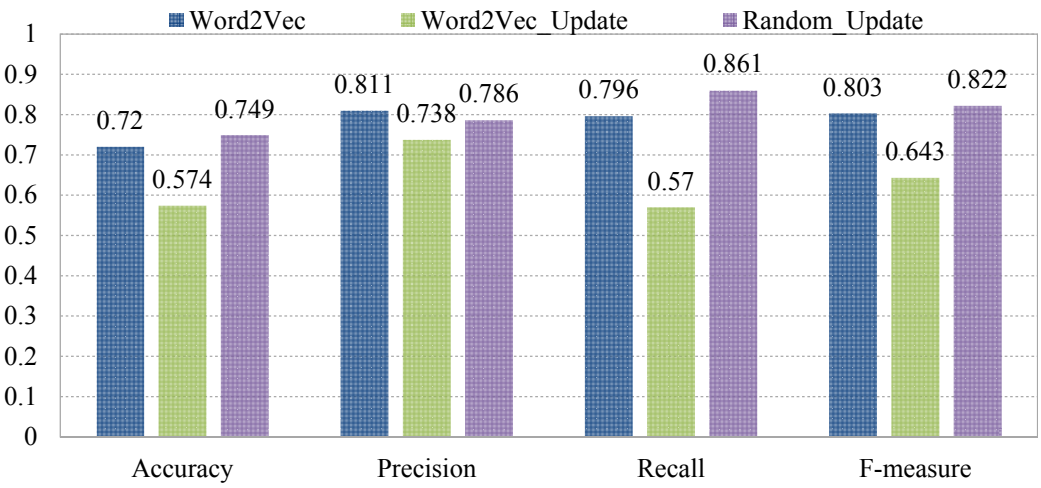


圖 13. Embedding Layer 實驗結果
[Figure 13. Experimental results for embedding layer]

由圖 13 所示，本論文所提方法針對謠言檢測資料集進行訓練時，利用隨機初始化法產生字詞向量會有較好的結果，其 F-measure 高達 0.822。經過多次實驗與探討，發現主要由兩個因素影響此實驗結果。第一，使用 Google News 預訓練的 Word2Vec 字典，每個字詞向量是由許多新聞文章訓練產生，在整個向量空間中彼此都有關聯。若在訓練時使用字典中的字詞向量，並不斷的在 RNN 中更新字詞向量，會使得該字詞在向量空間中的意義被改變，失去與其他字詞的關係，導致最後模型的精準度下降。第二，在謠言檢測資料集中，有一些字詞並未出現在 Google News 的 Word2Vec 字典裡，其字詞像向量為零，導致該字詞被神經網路忽略，進而降低模型的準確率。透過圖 13，我們也發現，使用 Word2Vec 字典但不更新字詞向量的效果較佳，也驗證了預訓練字典裡，若以不同資料來訓練並更新字詞向量，將會失去其原本的意義。

4.3 遞迴式神經網路架構比較 (The Effects of Recurrent Neural Networks)

在此實驗中，我們先僅以文字特徵進行不同 RNN 架構之謠言偵測效果比較，如下圖所示:

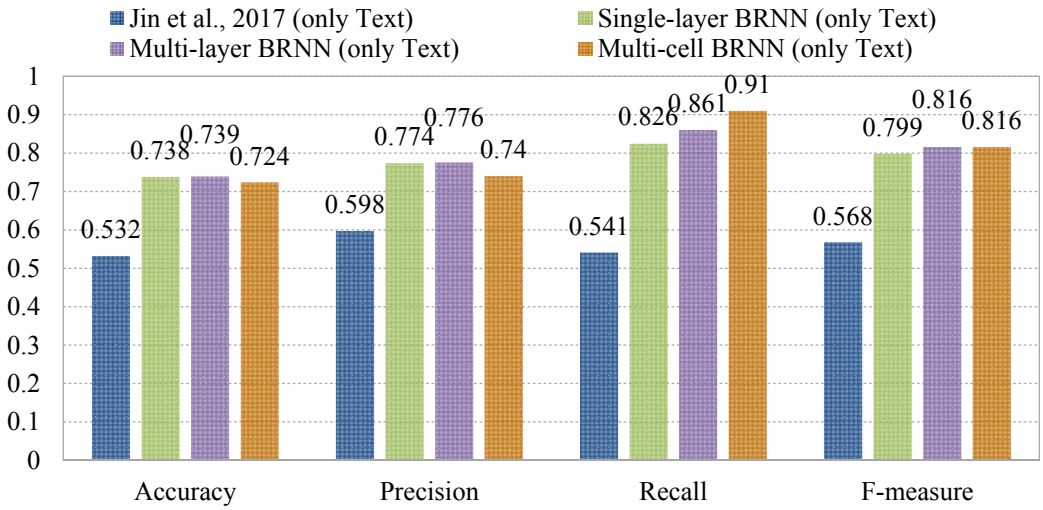


圖 14. 不同遞迴式神經網路架構之比較（文字特徵）

[Figure 14. Experimental results for different recurrent neural networks (Text feature)]

由圖 14 所示，當只考慮文字特徵時，多層 BRNN 與多單元 BRNN 架構的 F-measure 都達到 0.816，效果皆優於 Jin 等人所提出的方法 (Jin *et al.*, 2017)。而由圖 14 也可以得知，在這三種不同的 BRNN 架構中，若只看文字特徵，三者並沒有明顯的優劣，F-measure 都接近 0.8。

4.3.1 特徵融合的效果 (The Effects of Feature Fusion)

接著我們探討結合文字、圖像、與社群特徵，對分類結果的影響。由於社群特徵包含了標籤 (hashtag)、情緒、及使用者，我們首先比較不同組合特徵選取的實驗結果，如圖 15 與圖 16 所示:

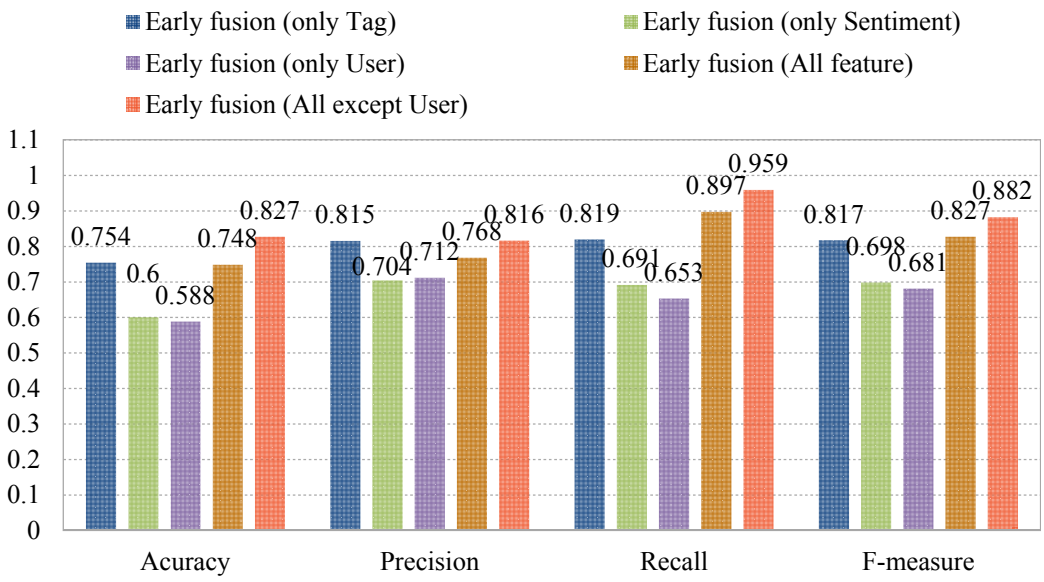


圖 15. Early Fusion 之特徵選取效果比較
[Figure 15. Experimental results for early fusion]

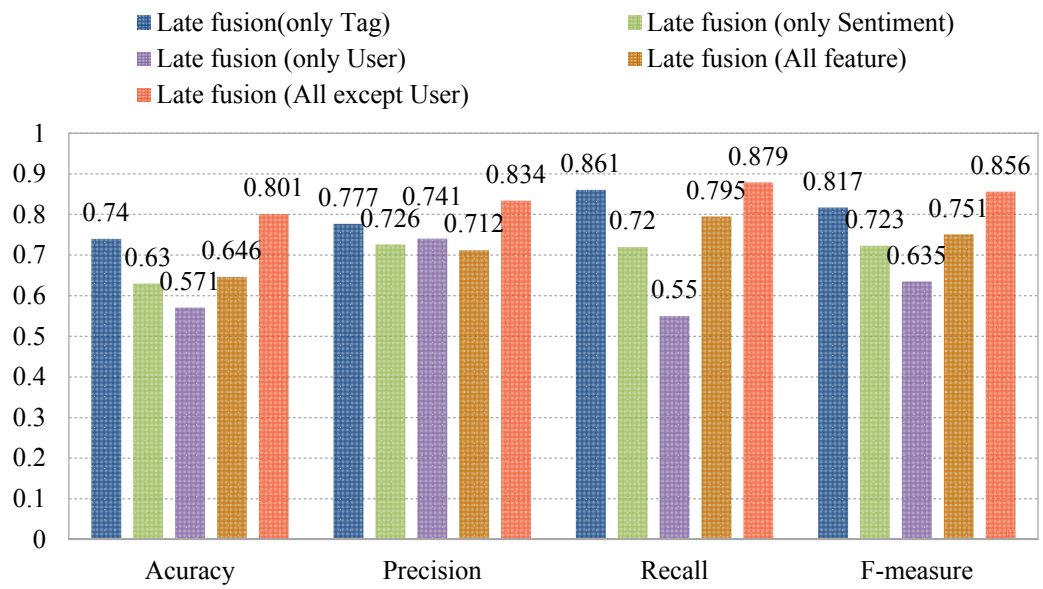


圖 16. Late Fusion 之社群特徵選取效果比較
[Figure 16. Experimental results for late fusion]

由圖 15 與圖 16 所示，不論使用 Early Fusion 或 Late Fusion 特徵融合策略，比起情緒與使用者特徵，加入標籤 (hashtag) 特徵具有較佳的效果。實驗結果顯示加入使用者

特徵反而會降低分類效果，在不採用使用者特徵的情況下，使用 Early Fusion 特徵融合策略，F-measure 最高分別可以達到 0.882 及 0.856，比納入全部特徵時的效果分別提升了 5.5%及 10%。經過觀察，我們發現使用者特徵對於該推文是否為謠言並沒有太大的關係，因為並不會因為該使用者朋友數量或總發文數的多寡，而影響其發出的推文為謠言或事實。

4.3.2 不同RNN架構的效果 (The Effects of RNN Architectures)

在使用 Early Fusion 特徵融合的情形下，我們進行不同 RNN 架構之實驗比較。實驗結果如圖 17 所示:

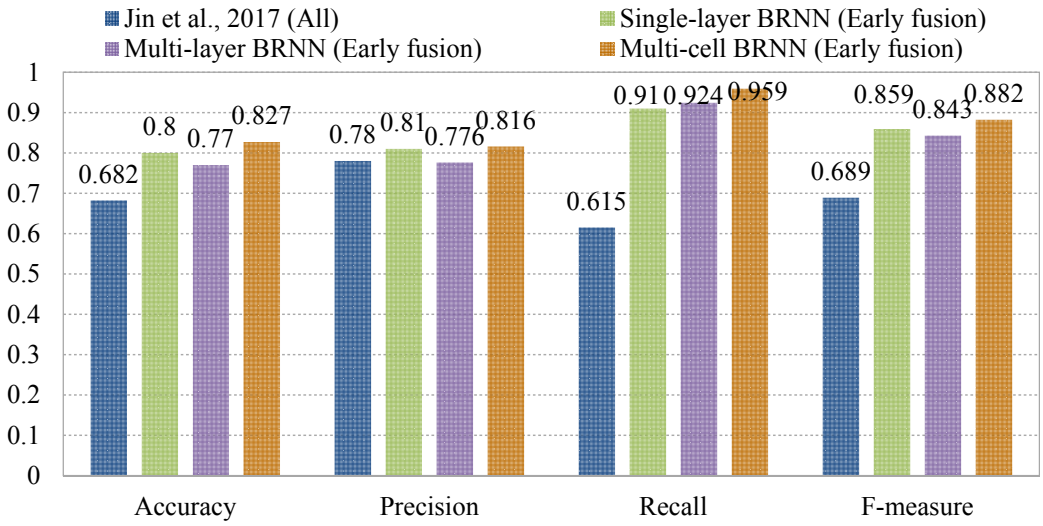


圖 17.不同遞迴式神經網路架構之比較（除使用者之外的所有特徵）

[Figure 17. Experimental results for different recurrent neural networks (All features except user feature)]

如圖 17 所示，在結合除了使用者之外所有特徵後，多單元 BRNN 在所有的評估標準下最為突出，F-measure 達到 0.882，其次為單層與多層 BRNN。經過反覆實驗與探討，我們發現使用多層 BRNN 時，經過第一層 BRNN 後，特徵向量已經過內部神經元運算，找出所有字詞的關聯，並進行了調整，故其輸出的特徵向量已經與原先的輸入不同。且由於該向量與所有特徵的關聯性已經被確定，後續再進入下一層 BRNN 時，該特徵向量並不會再有太大的變動，所以使用單層 BRNN 才會與使用多層 BRNN 的結果相近。為了驗證推論的正確性，本實驗進一步採用 T 檢定，分別針對 F-measure 及 accuracy，進行兩種架構的評比，判斷其差異是否為常態。經過計算，兩種架構針對 F-measure 及 accuracy 的 p-value 分數皆為 0.006，皆小於 0.01，證實其實驗結果並非偶然，具有統計意義。

實驗中我們也發現，多單元 BRNN 雖然類似於單層 BRNN，但是其向前傳遞與向後傳遞神經網路，分別經過了多層的隱藏層計算，透過加深內部 cell 的數目，強化了每個特徵向量與特徵序列（向前傳遞與向後傳遞）之間的關係。最後的實驗結果也顯示了多

單元 BRNN 比其他方法效果好，F-measure 可以達到最高的 0.882。我們也採用多單元 BRNN 與多層 BRNN 進行 T 檢定，針對 F-measure 及 accuracy，兩者的 p-value 數值為 0.04 及 0.014，皆小於 0.05，證實該結果具有統計意義。

接著我們比較兩者特徵融合策略，對最後分類結果的影響。實驗結果如圖 18 所示：

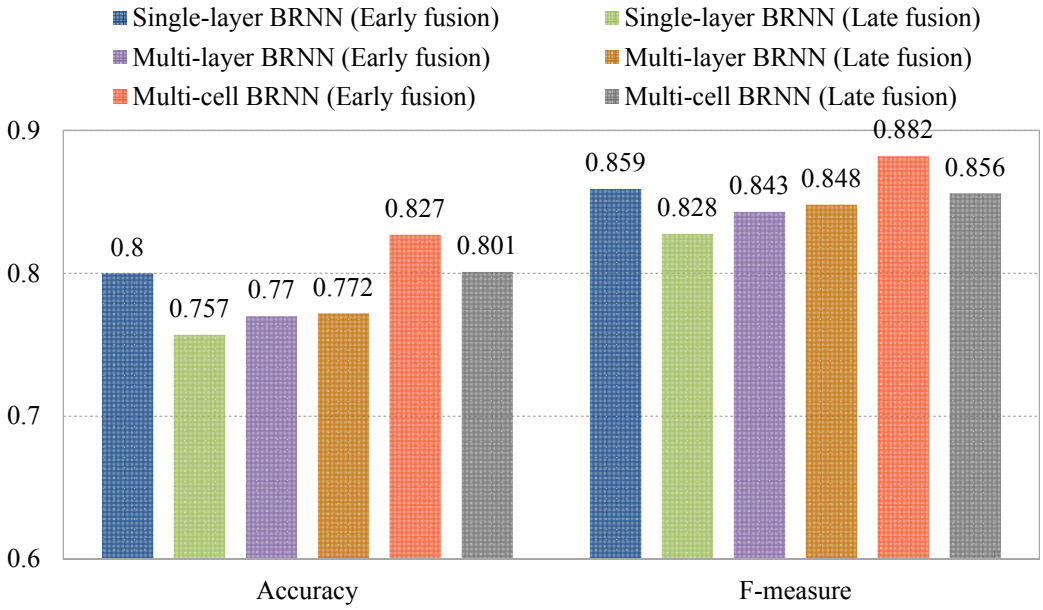


圖 18.不同特徵融合策略之比較

[Figure 18. Experimental results of different feature fusion strategies]

如圖 18 所示，在使用 Late Fusion 特徵融合策略時，不論是多層或多單元 BRNN，兩者的效果非常相近。而 Early Fusion 特徵融合策略對於單層以及多單元 BRNN，能大幅提升其效果。經過多次實驗與探討，我們發現其原因是，透過 BRNN 與注意力機制，社群特徵在 Early Fusion 階段融合，也能像其他特徵一起進行向量權重的調整，找出其中最能代表謠言或非謠言的重要特徵，提升分類準確率。

4.4 LSTM與GRU的比較 (The Effects of LSTM vs. GRU)

最後我們探討 RNN 架構中，採用 LSTM 或 GRU 不同處理單元對分類結果的影響。由於謠言檢測資料集資料量較小，為了降低資料分布的影響，我們採用 5-fold cross-validation。同時由於原 MediaEval 資料集的組成是根據不同事件 (event) 區分為 real 或 fake，並將該事件相關發文與回應 tweets 跟著列為 real 或 fake，並且已經區分為 training 與 test data，因此在進行 5-fold cross-validation 的資料 partition 時，我們將 training 及 test data 分別隨機切為 5 份，分別取 4 份 training 及一份 test，進行 5 次實驗後取其平均。實驗結果如圖 19 所示:

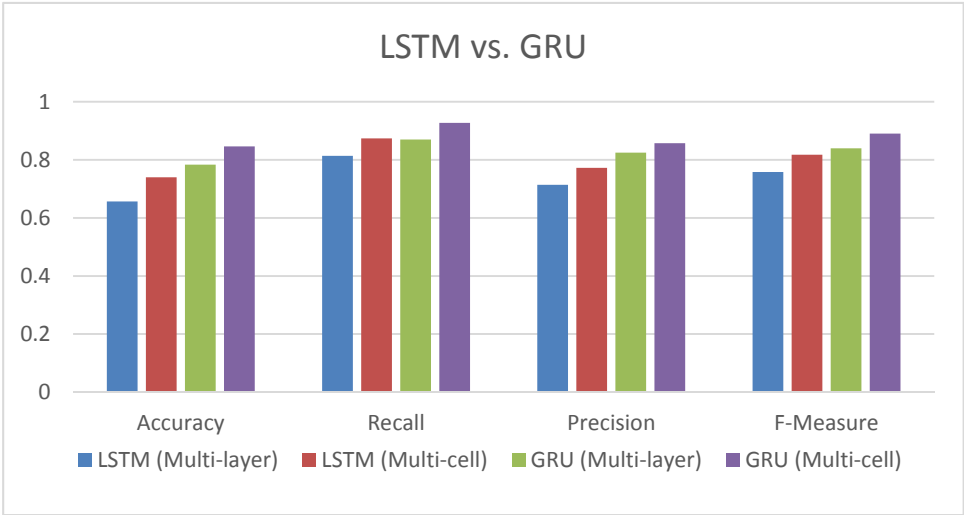


圖 19. LSTM 與 GRU 之比較
[Figure 19. Experimental results of LSTM and GRU]

如圖 19 所示，不論是 accuracy 或 F-Measure，基於 GRU 的 BRNN 架構 (BiGRU) 都比基於 LSTM 的 BRNN 架構 (BiLSTM)效果要好。同時，多單元(Multi-cell) BRNN 也都比多層(Multi-layer) BRNN 效果要好，最佳的效果為 Multi-cell BiGRU，F-Measure 達到 0.89。因此，這也驗證了 LSTM 與 GRU 並沒有絕對的優劣，不同任務須採用不同架構的設計才能獲得較佳的效果。

5. 結論 (Conclusions)

本論文提出一個基於多模式特徵融合的深度類神經網路架構進行謠言檢測。透過圖像描述模型能將圖像轉為敘述文字，有效發掘圖像中的語意，並與文字內容串接，進行 Word Embedding；針對社群特徵，我們採取了 Early 及 Late Fusion 不同特徵融合策略。我們也設計了多層與多單元兩種不同的雙向遞迴式神經網路(BRNN)架構，並結合注意力機制，以提升分類效果。實驗結果顯示，使用基於 GRU 的多單元(Multi-cell) BRNN 架構 (Multi-cell BiGRU)，以 Early Fusion 方式融合社群特徵，並結合文字特徵、圖像描述模組，能有效提升謠言檢測效果, 最佳 F-measure 達 0.89。

本論文所提出的方法仍有限制。首先，我們的方法主要是針對 Twitter 社群平台上的推文進行謠言檢測，由於每則推文的長度並不會太長，所以通常不能直接適用於長文件的謠言檢測。其次，本論文擷取圖像特徵的方式是將圖像先經過圖像描述模型，轉換成有意義的文字訊息。雖然經過評估該模型有一定的精確度，結果經過比對也大致符合圖像所表達的意義，但部分結果還是有落差，因此如何提升圖像描述模型的效果，有待進一步探討。最後，社群特徵並非全部有助於提升分類效果，其中的情緒分類僅採用情緒字典的比對，未來將採取不同的情緒分析方法，以進一步提升謠言偵測的效果。

參考文獻(References)

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

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of WWW 2011*, 675-684. doi: 10.1145/1963405.1963500

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining 2018*, 40-52. doi: 10.1007/978-3-030-04503-6_4

Cho, K., Merriënboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., … Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734. doi: 10.3115/v1/D14-1179

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of NIPS 2014*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211. doi: 10.1016/0364-0213(90)90002-E

Esuli, A. & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 417-422.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning 2017*, 1243-1252.

Graves, A. (2012). *Supervised sequence labelling with Recurrent Neural Networks.* (p.26). German, Heidelberg: Springer.

Gupta, M., Zhao, P., & Han, J. (2012). Evaluating event credibility on Twitter. In *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, 153-164. doi: 10.1137/1.9781611972825.14

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia 2017*, 795-816. doi: 10.1145/3123266.3123454

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436-444. doi: 10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. doi: 10.1109/5.726791

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., … Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of IJCAI 2016*, 3818-3824.

Ma, J., Gao, W., & Wong, K.-F. (2018). Detect rumor and stance jointly by neural multi-task learning. In *Proceedings of The Web Conference 2018*, 585-593. doi: 10.1145/3184558.3188729

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of INTERSPEECH 2010*, 1045-1048.

Mnih, V., Heess, N., Graves, A. & Kavukcuoglu, K. (2014). Recurrent models of visual attention. In *Proceedings of neural information processing systems 2014*, 2204-2212.

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

Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681. doi: 10.1109/78.650093

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR 2015*.

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

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2015*, 1-9. doi: 10.1109/CVPR.2015.7298594

Tan, Z., Wang, M., Xie, J., Chen, Y., & Shi, X. (2018). Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence 2018*, 4929-4936.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., …Polosukhin, I. (2017). Attention is all you need. In *Proceedings of neural information processing systems 2017*, 5998-6008.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2015*, 3156-3164. doi: 10.1109/CVPR.2015.7298935

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., … Ma, W.-Y. (2017). Topic aware neural response generation. In *Proceedings of Thirty-First AAAI Conference on Artificial Intelligence 2017*, 3351-3357.

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

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. In *Proceedings of IJCAI 2017*, 3901-3907. doi: /10.24963/ijcai.2017/545

Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R., … Qiao, Y. (2018). A Multi-task Learning Approach for Image Captioning. In *Proceedings of IJCAI 2018*, 1205-1211. doi: 10.24963/ijcai.2018/168

Computational Linguistics and Chinese Language Processing
Vol. 25, No. 1, June 2020, pp. 81-102

Linguistic Input and Child Vocalization of 7 Children from 5 to 30 Months: A Longitudinal Study with LENA Automatic Analysis

Chia-Cheng Lee[*], Li-mei Chen[+], and D. Kimbrough Oller[#]

Abstract

This study examined longitudinal changes in linguistic input, conversational turns, and child vocalizations in Chinese-speaking families using the computerized LENA (Language Environment Analysis) software, a system that captures audio data in children's natural environment and parses out speech data automatically. All-day home recordings (11-16 hours) from seven typically developing Chinese-learning children (two males and five females) at the ages of 5, 10, 14, 21, and 30 months were analyzed. Adult word count (AWC), conversational turn count (CT), and child vocalization count (CV) of 70 recordings (i.e., 7 children x 5 ages x 2 recordings) were retrieved from the LENA software. These recordings included times when families were asleep. As a result, the present study also compared the results with and without LENA-determined silence time (i.e., quiet and sleep time). The results showed that the percentage of silence in the recordings decreased with age, indicating that the children's awake time increased as they age. When the children were awake, they listened to an average of 1734 adult words, engaged in 39 conversational turns, and produced 150 vocalizations per hour from 5 to 30 months of age. The CV and CT increased with age, while the AWC did not show a clear pattern, which was similar to English normative estimates from Gilkerson and Richards (2008). The CT was also found to be a more effective contributor to the number of CV than AWC, indicating that speech produced in temporal proximity

[*] Department of Speech and Hearing Sciences, Portland State University, USA
E-mail: jiajiajanejane@gmail.com
[+] Department of Foreign Languages and Literature, National Cheng Kung University
E-mail: leemay@mail.ncku.edu.tw
The author for correspondence is Li-mei Chen.
[#] School of Communication Sciences and Disorders, The University of Memphis, USA
E-mail: koller@memphis.edu

to children's vocalizations or directed to children played an important role in eliciting child vocalizations.

Keywords: LENA, Adult Word, Conversational Turn, Child Vocalization, Longitudinal Study, Cross-language Comparison

1. Introduction

Child speech samples have traditionally been collected by visiting children's homes or inviting families into a research laboratory. LENA (Language Environment Analysis) software, a system that collects audio data without research assistants' presence and parses out audio data into several categories automatically, was developed in 2004 in the United States (LENA Research Foundation, 2020). The software has been used for observing English-speaking individuals (Gilkerson & Richards, 2008; Greenwood, Thiemann-Bourque, Walker, Buzhardt & Gilkerson, 2011; Suskind *et al*., 2013), Chinese-speaking families (Gilkerson *et al*., 2015; Lee, Jhang, Relyea, Chen & Oller, 2018; Zhang *et al*., 2015), preterm infants (Caskey, Stephens, Tucker & Vohr, 2011, 2014), multilingual speakers (Liu & Kager, 2017; Oller, 2010; Orena, Polka & Srouji, 2018), individuals with disorders (Ambrose, VanDam & Moeller, 2014; Charron *et al*., 2016; Oller *et al*., 2010; Thiemann-Bourque, Warren, Brady, Gilkerson & Richards, 2014; VanDam, Ambrose & Moeller, 2012; Warren *et al*., 2010), and older adults (Li, Vikani, Harris & Lin, 2014). The number of studies on the quantity of linguistic input, conversational turns, and child vocalizations in Chinese-speaking home environments have been limited. The present study observed changes in the quantity of linguistic input, conversational turns, and child vocalizations which occur between 5 and 30 months of age in Chinese-speaking families using LENA.

Research has shown that linguistic input, including the quantity and quality of caregiver speech and turn taking sequences, plays an important role in the child's vocal development (Caskey *et al*., 2011; Hart & Risley, 1995; Rowe, 2012; Suskind *et al*., 2013). This in turn serves as a strong predictor of their later vocabulary growth (Hart & Risley, 1995; Ramírez-Esparza, García-Sierra & Kuhl, 2014). Studies have also found that early vocal production is associated with future speech and language development. Rescorla *et al*. (2000) indicated that some children who were identified as late talkers at two years of age continued to exhibit language delay and were identified as children with Specific Language Impairment at three years of age. Gilkerson *et al*. (2018) also showed that school-age language and cognitive outcomes (9-13 years old) and quantity of adult talk and adult-child interaction during 18 to 24 months of age are related.

1.1 Linguistic Input and Conversational Turn

Linguistic input from adults or siblings is identified as one of the largest influences on children's verbal performances, including that of preterm infants (Caskey *et al*., 2011). Children understand five times more words than the words they produce (Ingram, 1989), suggesting that a substantial number of words need to be heard before a child speaks. Roy *et al*. (2009) reported that adult word input frequencies and age of acquisition of words is highly correlated. Adult word input between 10 and 36 months of age has been found to be related to a child's IQ at 3 years (Hart & Risley, 1995). Gilkerson and Richards (2009) also found that children who scored higher on language assessments tended to have talkative parents. The number of words parents spoke to children between two and six months of age predicted language ability at two years of age. Parents who earned at least a bachelor's degree talked more to their children than less educated parents. Also, first-born children were spoken to more than later born children.

Children may be at risk of learning languages if they do not have sufficient language exposure (Velleman & Vihman, 2002). Many scholars have claimed that language acquisition takes place even when the linguistic input that children are exposed to is addressed to them indirectly (Akhtar, Jipson & Callanan, 2001; Oshima-Takane, 1988; Oshima-Takane, Goodz & Derevensky, 1996). Other scholars argued that speech addressed directly to children has a stronger effect on children's language learning (Oller, 2010; Pearson, Fernandez, Lewedeg & Oller, 1997; Shneidman, Arroyo, Levine & Goldin-Meadow, 2013; Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013). The same phenomenon has been posited by Shneidman *et al*. (2013) and Shneidman and Goldin-Meadow (2012), who found that direct speech has a more important role in early word learning than indirect speech in children who grew up in communities where indirect speech was the major linguistic input.

In addition to receiving speech and language input, children also respond to the input (Hart & Risley, 1995). Mother-child vocal interactions have been discussed in several studies (Gratier *et al*., 2015; Gros-Louis, West, Goldstein & King, 2006; Jaffe *et al*., 2001). From 3 to 4 months of age, infants start to use pragmatic, semantic, and syntactic factors to predict when a conversational turn will end and begin (Gratier *et al*., 2015). However, studies on linguistic input and conversational turn-taking in Chinese-speaking environments, especially vocalizations produced in home environments, are, as of yet, few in number. Studies investigating the relationship among linguistic input, conversational turns, and children's vocalizations should shed some light on our understanding of the relationship between different types of linguistic input and language development.

1.2 Assessing Vocal Development Using an Automated Approach

Although the LENA system was mostly utilized in American-English environments, the system has yielded valid and reliable speech and language estimates in other languages (French: Canault, Le Normand, Foudil, Loundon & Thai-Van, 2016; Spanish: Weisleder & Fernald, 2013, Chinese (Mandarin and Shanghai dialect): Gilkerson *et al*., 2015; Zhang *et al*., 2015; Korean: Pae *et al*., 2016; Dutch: Busch, Sangen, Vanpoucke & van Wieringen, 2018; Vietnamese: Ganek & Eriks-Brophy, 2018). After comparing Chinese speech samples analyzed by the LENA system with the same samples transcribed by a native Chinese transcriber, Gilkerson *et al*. (2015) indicated that the validity of the LENA system in identifying and estimating adult words, child vocalizations, and conversational turns is reasonably accurate. Zhang *et al*. (2015) observed 22 Chinese-speaking families and their typically developing children between 3 and 23 months of age in Shanghai for a period of 6 months. A total of 19 recordings were made by each family. The 22 families were divided into two groups based on the speech output of the first three recordings. One group of families had fewer adult words (Group A), while the other group had a higher rate of adult words (Group B) in their first three recordings. The authors provided monthly feedback to the families regarding strategies to increase their linguistic input to and interaction with their children. The results overall showed that adult words and conversational turns increased during the first three months, but decreased during the last three months. However, Group A showed increased number of adult words in the last few recordings, which was not observed in Group B. The study indicates that the LENA system can be used to track children's vocal, speech, and language development and/or treatment progress. The authors also found that their number of conversational turns correlated positively with the MacArthur-Bates Communicative Development Inventories - Verbal (Fensen *et al*., 2007) and Minnesota Child Developmental Inventory Expressive Language (Ireton, 1992) scores for the change from baseline to 3 months. LENA estimates have also shown reliable and valid results when compared with scores of standardized assessments (Richards *et al*., 2017), including – Preschool Language Scale – 4th Edition (Zimmerman, Steiner & Pond, 2002) and the Receptive-Expressive Emergent Language Test – 3rd Edition (Bzoch, League & Brown, 2003).

Table 1 shows adult word count (AWC), conversational turn count (CT), and child vocalization count (CV) per hour from various ages, settings, and population. Depending on the children's age and the recording environment, children received different linguistic input and produced different number of words. AWC ranged from 889 to 1966. CT ranged from 17 to 75. CV ranged from 73 to 188 per hour. Gilkerson and Richards (2008) examined a corpus of spontaneous speech data in English-speaking families and created normative estimates for CV and CT each month when children were between 2 and 48 months of age. Here only

values measured at 5, 10,14, 21, and 30 months are listed in Table 1.

Table 1. Studies reported AWC, CT, and CV per hour in families with 0-3-year-old children

First author & Year	Population, *n* (male/female)	Age	Language	AWC per hour	CT per hour	CV per hour
Ambrose (2014)	Hard of hearing, *n*=28 (10/18)	12-36 months (mo)	English	1429	59	Not Applicable (NA)
Gilkerson (2008)	Typically Developing (TD), *n*=329 (167/162)	2-48 mo	English	NA	5 mo 17 10 mo 23 14 mo 27 21 mo 36 30 mo 40	5 mo 73 10 mo 95 14 mo 102 21 mo 145 30 mo 184
Greenwood (2013)	TD, *n*=30 (NA/NA)	12-20 mo	English	1095	38	143
Thiemann-Bourque (2014)	Down syndrome (DS), *n*=9 (3/6), and age- and gender-matched TD, *n*=9 (3/6)	9-54 mo, young DS 9-11 mo, old DS 25-54 mo	English	Yong DS 889 Old DS 1044 TD NA	Young DS 18 Old DS 19 TD 44	Young DS 102 Old DS 64 TD 179
Warren (2010)	Autism, *n*=26 (22/4), and age- and gender-matched TD *n*=78 (66/12)	16-48 mo	English	Autism 1079 TD 1138	Autism 35 TD 4	Autism 134 TD 188
Zhang (2015)	TD, *n*=22 (10/12)	3-23 mo	Shanghai dialect and Mandarin	Baseline 1758 1 mo 2174 1-3 mo 1966 4-6 mo 1711	Baseline 63 1 mo 75 1-3 mo 66 4-6 mo 56	NA

1.3 The Present Study

Because of the laborious coding required for estimating linguistic input from the ambient environment, studies focusing on child speech development are usually based on a limited set of recordings. To our knowledge, only three studies (Gilkerson *et al*., 2015; Lee *et al*., 2018;

Zhang *et al*., 2015) reported observations in Chinese-learning children's natural environments using LENA. In view of this, LENA was adopted for data collection and processing in the present study. This paper explores the relationship among children's vocalization, the linguistic input children received, and amount of interaction adults and children had per hour (e.g., total number of AWC/total length of a recording). However, the recordings included times when families were asleep. Thus, the present study investigated the research questions using the total length of the recording without LENA-determined silence time (i.e., quiet, sleep time) to calculate another set of average numbers of AWC, CT, and CV per hour (e.g., total number of AWC/(total length of a recording without silence time in the recording)). Periods of silence were removed to ensure that the analysis only included times when children were most likely to be awake. Analyzing results by removing periods of silence time from LENA recordings has also been reported in several other studies (Marchman, Martínez, Hurtado, Grüter & Fernald, 2017; Sacks *et al*., 2013). Since children at 0-2 years old sleep an average of 12.7 hours a day and children at 2-3 years old sleep an average of 12 hours a day (Galland, Taylor, Elder, & Herbison, 2012), the results of the present study could have been influenced by long sleeping times. Therefore, the present study aimed to compare the results when silence time was included with the results when silence time was removed from the analyses.

The present study investigated the following questions:

1. Do adult word count (AWC), conversational turn count (CT), and child vocalization count (CV) increase as children grow older?
2. Are there different patterns in AWC, CT, and CV when LENA-determined silence time is removed?
3. Are both AWC and CT effective contributors to the number of CV at 5, 10, 14, 21, and 30 months?
4. Do AWC, CT, and CV show cross-language differences?

2. Methods

2.1 Participants

Seven Chinese-speaking families and their children (two males and five females) participated in the study. The families lived in Tainan, Taiwan, an environment where Mandarin Chinese and Southern Min (Taiwanese) were mostly spoken. All the children were born full-term without hearing or neurodevelopmental disorders. Table 2 shows demographic information of the participants.

Table 2. Demographic information of the participants

Child	Gender	Birth order	Mother's education
A	F	1st	M.A.
B	F	1st	B.A.
C	F	1st	B.A.
D	F	1st	B.A.
E	M	1st	B.A.
F	M	2nd	B.A.
G	F	2nd	B.A.

2.2 Recording Procedure

The digital language processor (DLP), a recording device developed along with the LENA Pro system (LENA Research Foundation, 2020), was used to collect data. Before each recording session started, a child wore a specially designed vest with a DLP (Figure 1). The caregiver turned the DLP on to start a recording session and switched the DLP off after 16 hours of recording. The recording file was automatically uploaded and processed (Figure 2) once the DLP was connected to a computer with the LENA Pro software. The LENA Pro software identified speech and other sounds from each recording and generated counts at 5-minute, hour, day, and month intervals. The authors retrieved the counts/reports (Figure 3) from the software for further analysis.

Figure 1. The LENA digital language processor (DLP) placed in the pocket of a vest

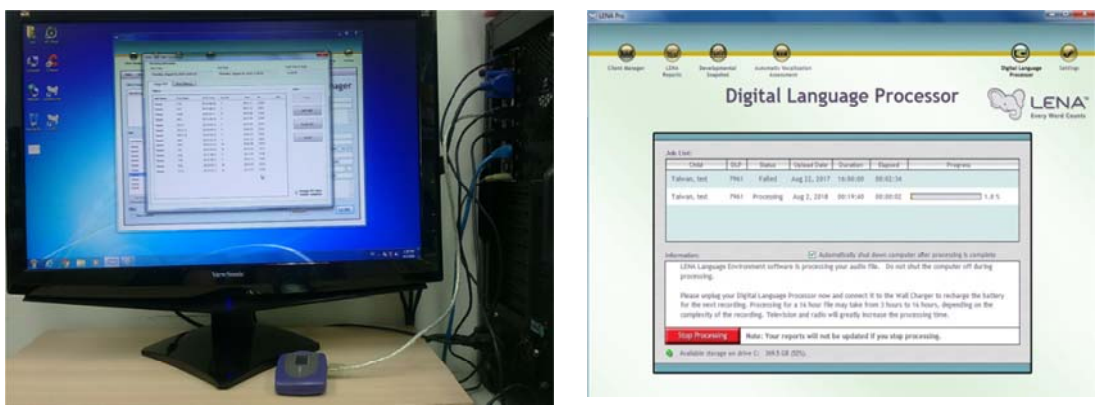
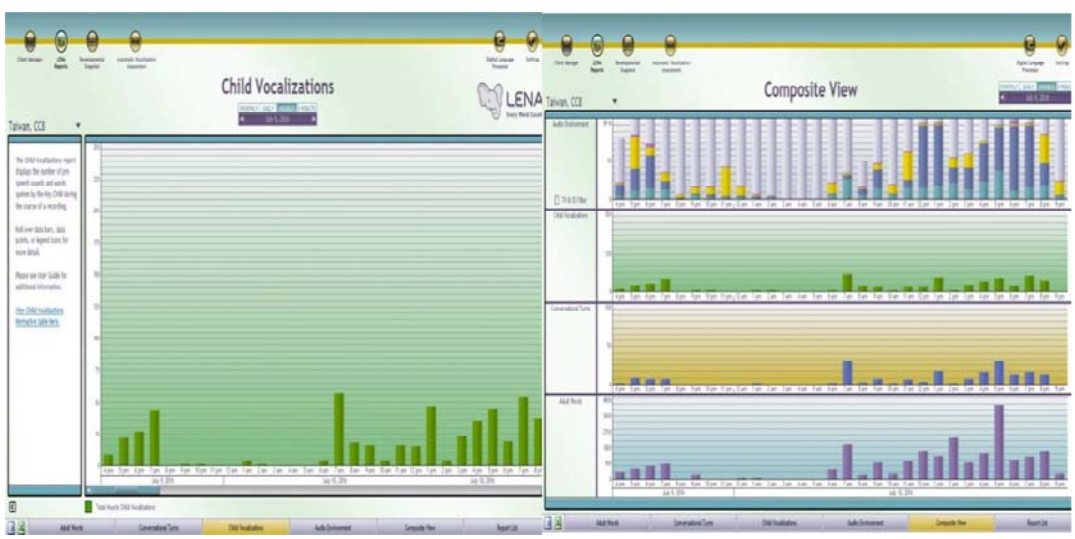


Figure 2. Data transfer from a DLP to the LENA Pro software



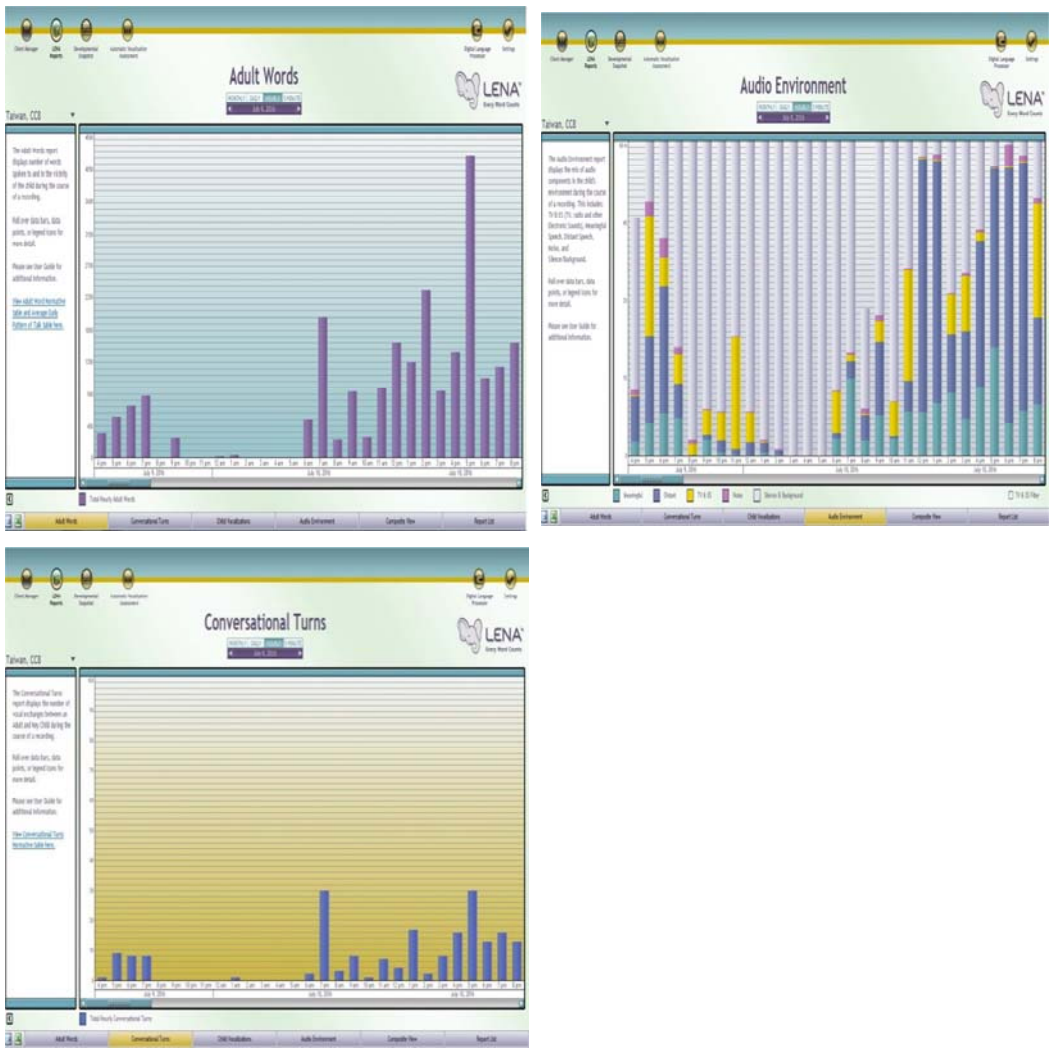


Figure 3. Reports from the LENA Pro software

A set of two recordings were made at each age: 5, 10, 14, 21, and 30 months old. A total of 70 recordings were analyzed (i.e., 7 children x 5 ages x 2 recordings). All the recordings were 16 hours in length except for 6 of the recordings due to insufficient power of the device used on the recording day. The 6 recordings were between 11 and 14 hours in length.

2.3 Data Processing by the LENA Software

The audio data was processed and categorized by the LENA Pro software into eight sound categories: (1) the key child who wore a vest with the DLP, (2) other child, (3) adult male, (4) adult female, (5) overlapping sounds, (6) noise, (7) electronic sounds (e.g., TV), and (8)

silence (i.e., silence, quiet, or vegetative sounds such as sneezes, coughs, or snores). Each category was further identified as clear and unclear (i.e., quiet and distant) subcategories. After the eight sound categories were identified, the LENA system determined adult word count (AWC), communication turn count (CT), and child vocalization count (CV).

2.3.1 Adult Word Count (AWC)

AWC measured the total number of words spoken around the key child. Using acoustic features in speech signal (e.g., formants, pitch, segment duration, silence duration), adult sounds were identified as phones using American-English phone parsing models. Speech segments were identified based on differential acoustic energy patterns, and no specific adult words were identified. AWC included both speech directed to the key child and speech directed to others. In Mandarin Chinese, one syllable represents one spoken syllable, whereas one word may contain one or more spoken syllables. For example, 窗戶 chuang hu (*window*) has two spoken syllables but counts as one word. Gilkerson *et al*. (2015) compared syllable count (e.g., 窗戶 chuang hu = two syllables) and word count (窗戶 chuang hu = one word) transcribed by a trained native Chinese human transcriber with AWC and found that both comparisons showed valid and reliable estimates of adult word count. The authors suggested that since the comparisons were both reliable, researchers can use LENA-determined AWC (syllable count) in future studies. The authors also indicated that since all languages have phonemes and syllables, and the acoustic features of consonants and vowels are similar across languages, using acoustic information to estimate adult word count should not be affected by language differences.

2.3.2 Conversational Turn Count (CT)

Conversational turn count (CT) refers to the total number of conversational turns the child engaged in with other speakers. A conversational turn is defined as a child speaking and an adult or a child responding, or an adult or a child speaking and the child responding within 5 seconds. Both intentional and unintentional vocal production and responses can be counted as turns.

2.3.3 Child Vocalization Count (CV)

Child vocalization count (CV) is the total number of speech-related vocalizations the child produces. A CV would be identified if there was a 300 millisecond or longer vocal break between the key child's vocalization. Cries, laughs, and vegetative sounds such as sneezes, coughs were excluded from child vocalization count. Similar to AWC, the LENA system did not identify specific words or syllables in utterances. If a child says "ma" or "I want that I want that I want that" without pauses between words, each utterance is counted as one

vocalization.

2.4 Data Analyses

Five categories retrieved from the LENA reports were used for further analyses in the present study: (1) the length of each recording, (2) adult word count (AWC), (3) communication turn count (CT), (4) child vocalization count (CV), and (5) length of silence in each recording. The total number of words or sounds in each recording may differ depending on the length of the recording. Since the length of each recording was different, the average number of AWC, CT, and CV per hour retrieved from each recording was first calculated. Next, the average number of AWC, CT, and CV per hour retrieved from each recording without silence were calculated. Two sets of statistical measures were then analyzed. First, six one-way repeated measure ANOVAs were performed to explore whether there were any changes in the three variables (the average number of AWC, CT, and CV per hour) across time as well as when silence was included or excluded. Next, ten multiple regressions were performed at the ages of 5, 10, 14, 21, and 30 months to examine how much AWC and CT contribute to CV at each age and whether or not silence was included.

3. Results and Discussion

3.1 Changes of AWC, CT, and CV Overtime

Figure 4A shows the average number of adult word count (AWC), conversational turn (CT), and child vocalization (CV) per hour and their standard deviations from the recordings made at 5, 10, 14, 21, and 30 months. The average number of AWC per hour shows an increase from 5 to 10 months and a gradual decrease from 10 to 30 months. However, the differences among the five ages are not statistically significant, which is similar to the finding of Gilkerson and Richards (2008). The authors stated that AWC and chronological age in English-speaking families were not significantly correlated. The results in the present study also showed that the number of child vocalizations increased slowly with age, even when the child received a fair amount of linguistic input from the environment. That is, children heard an average of 412 to 752 adult words per hour from 5 months to 30 months old. However, the average number of child vocalizations only increased from 27 to 90 vocalizations per hour from 5 to 30 months.

The average number of CT per hour also shows a gradual increase from 5 (5 per hour) to 30 (23 per hour) months. The differences among the five ages are statistically significant [$F(4, 24) = 3.318$, $p < .05$]. A post hoc analysis indicates that the average number of CT per hour at 21 months (18 per hour) is significantly higher than at 5 months (5 per hour) [$t(6) = 3.716$, $p < .05$]. The increased number of CT indicates that the adults became more and more responsive to their children's utterances, and vice versa. The adults may have initiated the

conversation when they thought that their children were ready to talk, or responded to their child utterances right away. The children may have also learned to gain other people's attention by producing sounds. Or, they may have learned to respond to adults' speech right away as they grew older.

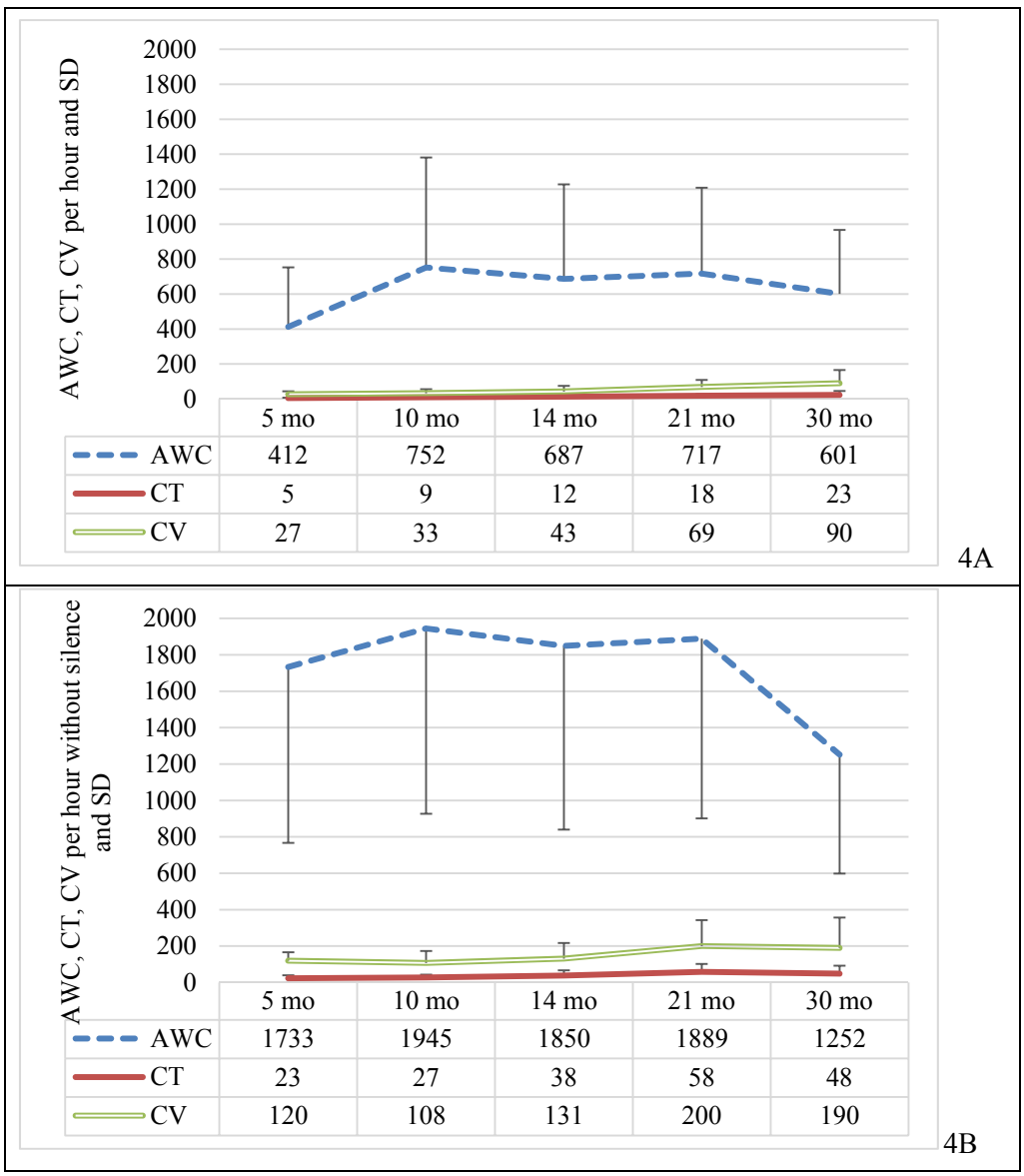


Figure 4. Average adult word (AWC), conversational turn (CT), child vocalization (CV) per hour with and without silence and standard deviations at 5, 10, 14, 21, and 30 months

3.2 Changes of AWC, CT, and CV Overtime after Removing Silence

Periods of silence were removed from recordings to ensure that only the times when children were most likely to be awake were included in the analysis. Figure 4B shows the average number of AWC, CT, and CV per hour and their standard deviations after removing the periods of LENA-determined silence from the recordings. The standard deviations of the average AWC per hour was high at all five ages as shown in both Figures 4A and 4B. However, the variability across families is even higher after the periods of silence were removed. The percentage of silence (i.e., (silence time/total length of recording) x 100) decreased with age (i.e., 5 mo: 73%, 10 mo: 66%, 14 mo: 62%, 21 mo: 59%, 30 mo: 48%), which was in line with Galland's *et al*. (2012) finding that children' s sleep time decreased with age.

As expected, the mean number of the three variables was at least twice as high without silence as with silence. Without silence time, the average number of CT and CV per hour also gradually increased from 5 (CT: 23; CV: 120 per hour) to 30 months (CT: 48; CV: 190 per hour). But, the differences among the five ages were not statistically significant. The average number of AWC per hour showed an increase from 5 (1733 per hour) to 10 (1945 per hour) months and a gradual decrease from 10 to 30 (1252 per hour) months. Yet, the differences among the five ages were not statistically significant either. Also, the average number of CT per hour was significantly different across ages before silence was removed, but was not significant after silence was removed. The average number of CT (i.e., increased with age), and the periods of silence (i.e., decreased with age) may account for the change.

In addition, the AWC and CT in the present study from the data across the five ages with silence removed (AWC: 1734; CT: 39 per hour) were more similar to Chinese-speaking data from Zhang *et al*. (2015) (AWC baseline: 1758; CT baseline: 63 per hour) than the results with silence included (AWC: 634; CT: 14 per hour). Zhang *et al*.'s (2015) results were more similar to results when silence was excluded in the present study because the authors instructed their Chinese-speaking families to record for 12 hours during the daytime. The finding also suggests that LENA-determined silence was identified as reasonably accurate.

3.3 Relationships among AWC, CT, and CV

Multiple regressions were performed at each age to explore the relationship among AWC, CT, and CV. The results showed that the numbers of AWC and CT could predict the numbers of CV at 10 months and 30 months. At 10 months, the results of the regression indicated that the model explained 88.1% of the variance and that the model was a significant predictor of the number of CV, $F(2,4) = 23.306$, $p = .006$. While the number of CT contributed significantly to the model (B = 3.677, $p = .003$), the number of AWC did not (B = -.008, $p = .222$). That is, the increase of one unit of CT could contribute to the increase of 3.677 units of CV. At 30 months,

the results of the regression indicated that the model explained 95% of the variance and that the model was a significant predictor of the number of CV, $F(2,4) = 57.9$, $p = .001$. While the number of CT contributed significantly to the model (B = 3.899, $p = .002$), the number of AWC did not (B = -.044, $p = .266$). That is, the increase of one unit of CT could contribute to the increase of 3.899 units of CV.

3.4 Relationships among AWC, CT, and CV after Removing Silence

Multiple regressions were performed at each age to explore the relationship among AWC, CT, and CV after the removal of the silence. The results showed that the numbers of AWC and CT could successfully predict the numbers of CV at 10 months, 21 months and 30 months. At 10 months, the results of the regression indicated that the model explained 85.4% of the variance and the model was a significant predictor of the number of CV, $F(2,4) = 18.614$, $p = .009$. While the number of CT contributed significantly to the model (B = 4.194, $p = .004$), the number of AWC did not (B = -.017, $p = .168$). That is, the increase of one unit of CT could contribute to the increase of 4.194 units of CV. At 21 months, the results of the regression indicated that the model explained 91.3% of the variance and that the model was a significant predictor of the number of CV, $F(2,4) = 32.397$, $p = .003$. While the number of CT contributed significantly to the model (B = 3.656, $p = .001$), the number of AWC did not (B = -.054, $p = .058$). That is, the increase of one unit of CT could contribute to the increase of 3.656 units of CV. At 30 months, the results of the regression indicated that the model explained 93.9% of the variance and that the model was a significant predictor of the number of CV, $F(2,4) = 47.429$, $p = .002$. While the number of CT contributed significantly to the model (B = 4.077, $p = .01$), the number of AWC did not (B = -.028, $p = .664$). That is, the increase of one unit of CT could contribute to the increase of 4.077 units of CV. Both sets of analyses indicated that speech directed to children or speech spoken right before or after child vocalizations (i.e. CT) imposed stronger effects to children's vocalizations than speech that was not spoken in temporal proximity to children's vocalizations.

3.5 Cross-language Comparison

With silence time included, the average number of AWC, CT, and CV across the five ages was 634, 14, and 52 per hour (i.e., 634*12 hr=7608, 14*12 hr=168, 52*12 hr=624 per 12-hour day) respectively. Compared with the English normative percentile estimates for AWC, CT, and CV in Gilkerson and Richards (2009), the Chinese-speaking families' AWC in the present study were at the 10^{th}-20^{th} percentile, and CT and CV were below the 10^{th} percentile. With silence excluded, the average number of AWC, CT, and CV across the five ages was 1734, 39, and 150 per hour (20808, 468, 1800 per 12-hour day) respectively. Compared with the English normative percentile estimates for AWC, CT, and CV in Gilkerson and Richards (2009), the

Chinese-speaking families' AWC in the present study were at the 80th-90th percentile, and CV and CT were at the 40th-50th percentile, which were much higher than when silence was included. As discussed earlier, the results with silence excluded were more similar to Zhang *et al.*'s (2015) AWC and CT baseline values; the results with silence excluded can be compared to the results in Gilkerson and Richards (2009). These results showed that the Chinese-speaking caregivers in the present study were on the talkative end of the English normative estimates. However, the Chinese-speaking adults and children were not vocally engaged at similar rates as AWC because the percentile of CT and CV were much lower than percentile of AWC. Gilkerson and Richards (2009) found that children who were first-born, were girls, or had parents with higher education tended to receive more adult talk each day. In the present study, the three factors might have also contributed to high AWC in the present study: 1) All seven mothers were highly educated, having received at least a bachelor's degree, 2) five out of the seven children were first born, and 3) five of the seven children were girls. However, unlike the results reported in Gilkerson and Richards (2009), the talkative caregivers in the present study did not have talkative children.

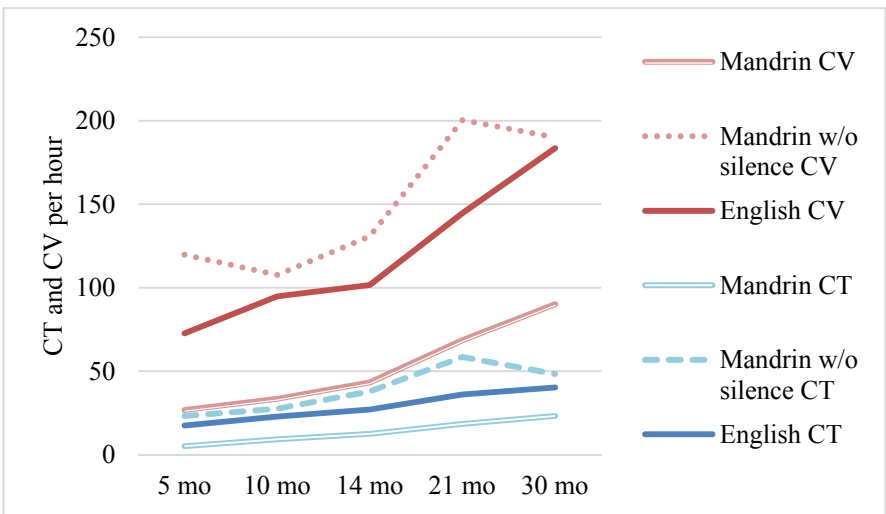


Figure 5. Average adult word (AWC), conversational turn (CT), child vocalization (CV) per hour from the present study and Gilkerson and Richards (2008)

Figure 5 shows longitudinal CT and CV changes in the English-speaking families from Gilkerson and Richards (2008) and the Chinese-speaking families from the present study. Both groups of families showed a gradual increase with age. When silence was included, the Chinese-speaking families showed overall lower CT and CV than the English-speaking families. However, when silence was removed, the Chinese-speaking families showed higher values than the English-speaking families. The group differences could be explained by the fact that the LENA-determined silence not only included times when families were sleeping

but also when families were awake but quiet. The results of the two sets of data would be more comparable if the English samples also exclude LENA-determined silence. Another possible reason for the group differences is sample size. More participants and detailed analyses are needed to explore possible cultural differences or confirm the results.

3.6 Limitations and Future Directions

Limitations were identified in the present study and can be addressed in future research. First, a differentiation of the number of child-initiated conversational turns and adult-initiated conversational turns would help examine parent-child interaction patterns and identify the relationship between CT and CV. Now, CT consists of both when a child speaks and an adult responds, and when an adult speaks and the child responds. The LENA Advanced Data Extractor (ADEX, LENA Research Foundation, 2020) would be useful in future research because it provides a more detailed output, including utterances or words of male adults, female adults, the key child, and other children.

Second, to ensure that the key child is really taking turns with another speaker or vice versa, the content of the adult words and child vocalizations requires human coding because the LENA system does not identify the content of the speech sample. For example, it is possible that a parent was holding the key child while talking to another person, but the LENA system may count this parent's utterances as if she or he were talking to the key child. Third, regarding the unit of speech samples, the LENA system categorizes adult and child speech samples in different units. AWC refers to the number of individual words adults speak, while CV means the number of speech-related utterances produced by the children. When a child produces prelinguistic sounds in a sequence or one breath, the LENA system may count these sounds as one CV. However, when the child starts to produce words or a mixture of babbling and words, the LENA system may still recognize those word strings/vocalizations as one CV. Again, human coding of the recording would be able to identify children's utterances in word or syllable units when the child starts to produce words.

Furthermore, the results of the present study were only compared with the English normative estimates because Chinese normative estimates using LENA are not available. Developing a Chinese version of the LENA normative estimates would enhance people's understanding of the effects of early vocal development and adult-child interactions on later development in the Chinese-learning children. Including a larger cohort of participants (i.e., with different socio-economic status, later-born children, male children) to collect a corpus would best represent the Chinese-learning children's speech capacity at the age.

4. Conclusion

The LENA automated approach has provided researchers with a new recording method that has automatic parsing capacities. The researchers investigated longitudinal changes in the average AWC, CT, and CV with and without silence time, relationship among the three variables, and cross-language comparison in Chinese-learning families with children ranging in age from 5 to 30 months. The percentage of LENA-determined silence decreased with age, indicating that the children's awake time increased as they age. The results also showed that a typically developing Chinese-learning child in the present study listened to an average of 1734 adult words, engaged in 39 conversational turns, and produced 150 vocalizations per hour from 5 to 30 months of age when he or she was awake. Child vocalizations and conversational turns increased over time, but adult word count did not show a clear pattern. When the periods of silence were included, the number of AWC and CT predicted the numbers of CV at 10 months and 30 months. After the periods of silence were removed, the results showed that the numbers of AWC and CT predicted the numbers of CV at 10, 21, and 30 months. This result suggests that the speech produced in temporal proximity to children's vocalizations or directed to children exerted a stronger influence on the number of child vocalizations than the quantity of adult words.

Acknowledgement

This research was supported by Chiang Ching-Kuo Foundation for International Scholarly Exchange in Taiwan to Li-mei Chen for international collaboration with Dr. Kim Oller at the University of Memphis. A special thank you is extended to the families of the children in this longitudinal study for their support of this project.

References

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning words through overhearing. *Child Development*, *72*(2), 416-430. doi: 10.1111/1467-8624.00287

Ambrose, S. E., VanDam, M., & Moeller, M. P. (2014). Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear and Hearing*, *35*(2), 139-147. doi: 10.1097/AUD.0b013e3182a76768

Busch, T., Sangen, A., Vanpoucke, F., & van Wieringen, A. (2018). Correlation and agreement between Language ENvironment Analysis (lenaTM) and manual transcription for Dutch natural language recordings. *Behavior Research Methods*, *50*(5), 1921-1932. doi: 10.3758/s13428-017-0960-0

Bzoch, K. R., League, R., & Brown, V. L. (2003). *Receptive-expressive emergent language test*. Austin, TX : PRO-ED.

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENATM) in European French. *Behavior Research Methods*, *48*(3), 1109-1124. doi: 10.3758/s13428-015-0634-8

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2011). Importance of parent talk on the development of preterm infant vocalizations. *Pediatrics*, *128*(5), 910-916. doi: 10.1542/peds.2011-0609

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2014). Adult talk in the NICU with preterm infants and developmental outcomes. *Pediatrics*, *133*(3), e578–e584. doi: 10.1542/peds.2013-0104

Charron, C., Fitzpatrick, E. M., McSweeney, E., Rabjohn, K., Somerville, R., & Steacie, P. (2016). Language ENvironment Analysis (LENA) with children with hearing loss: A clinical pilot. *Canadian Journal of Speech-Language Pathology & Audiology*, *40*(1), 93-104.

Fensen, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories*. Baltimore, MD: Paul H. Brookes.

Galland, B. C., Taylor, B. J., Elder, D. E., & Herbison, P. (2012). Normal sleep patterns in infants and children: A systematic review of observational studies. *Sleep Medicine Reviews*, *16*(3), 213-222. doi: 10.1016/j.smrv.2011.06.001

Ganek, H. V., & Eriks-Brophy, A. (2018). A concise protocol for the validation of Language ENvironment Analysis (LENA) conversational turn counts in Vietnamese. *Communication Disorders Quarterly*, *39*(2), 371-380. doi: 10.1177/1525740117705094

Gilkerson, J., & Richards, J. A. (2009). *The power of talk. Impact of adult talk, conversational turns and TV during the critical 0-4 years of child development* (LENA Foundation Technical Report LTR-01-2). Retrieved from https://www.lena.org/wp-content/uploads/2016/07/LTR-01-2_PowerOfTalk.pdf

Gilkerson, J., & Richards, J. A. (2008). *The LENA Natural Language Study* (LENA Foundation Technical Report LTR-02-2). Retrieved from https://www.lena.org/wp-content/uploads/2016/07/LTR-02-2_Natural_Language_Study.pdf

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics*, *142*(4), e20174276. doi: 10.1542/peds.2017-4276

Gilkerson, J., Zhang Y., Xu D., Richards J. A., Xu X., Jiang F., …Topping K. (2015). Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech, Language, and Hearing Research*, *58*(2), 445-452. doi: 10.1044/2015_JSLHR-L-14-0014

Gratier, M., Devouche, E., Guellai, B., Infanti, R., Yilmaz, E., & Parlato-Oliveira, E. (2015). Early development of turn-taking in vocal interaction between mothers and infants. *Frontiers in Psychology*, *6*, 1167. doi: 10.3389/fpsyg.2015.01167

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly*, *32*(2), 83-92. doi: 10.1177/1525740110367826

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*(6), 509-516.doi: 10.1177/0165025406071914

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H Brookes Publishing.

Ingram, D. (1989). *First Language Acquisition: Method, Description and Explanation*. New York, NY: Cambridge University Press.

Ireton, H. (1992). *Minnesota Child Development Inventory*. Minneapolis, MN: Behavior Science Systems.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., Jasnow, M. D., Rochat, P., …Stern, D. N. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the Society for Research in Child Development*, *66*(2), i-149.

Lee, C. C., Jhang, Y., Relyea, G., Chen, L. M., & Oller, D. K. (2018). Babbling development as seen in canonical babbling ratios: A naturalistic evaluation of all-day recordings. *Infant Behavior and Development*, *50*, 140-153. doi: 10.1016/j.infbeh.2017.12.002

LENA Research Foundation. (2020). L*ENA Research Foundation*. Retrieved from http://www.lena.org/

Li, L., Vikani, A. R., Harris, G. C., & Lin, F. R. (2014). Feasibility study to quantify the auditory and social environment of older adults using a digital language processor. *Otology & Neurotology: Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology*, *35*(8), 1301-1305. doi: 10.1097/MAO.0000000000000489

Liu, L., & Kager, R. (2017). Perception of tones by bilingual infants learning non-tone languages. *Bilingualism: Language and Cognition*, *20*(3), 561-575. doi: 10.1017/S1366728916000183

Marchman, V. A., Martínez, L. Z., Hurtado, N., Grüter, T., & Fernald, A. (2017). Caregiver talk to young Spanish-English bilinguals: Comparing direct observation and parent-report measures of dual-language exposure. *Developmental Science*, *20*(1), e12425. doi: 10.1111/desc.12425

Oller, D. K. (2010). All-day recordings to investigate vocabulary development: A case study of a trilingual toddler. *Communication Disorders Quarterly*, *31*(4), 213-222. doi: 10.1177/1525740109358628

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, …Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354-13359. doi: 10.1073/pnas.1003882107

Orena, A. J., Polka, L., & Srouji, J. (2018). Evaluating the LENA recording system for investigating speech input in a French-English bilingual context. *The Journal of the Acoustical Society of America*, *143*(3), 1871-1871. doi: 10.1121/1.5036141

Oshima-Takane, Y. (1988). Children learn from speech not addressed to them: The case of personal pronouns. *Journal of Child Language*, *15*(1), 95-108. doi: 10.1017/S0305000900012071

Oshima-Takane, Y., Goodz, E., & Derevensky, J. L. (1996). Birth order effects on early language development: Do secondborn children learn from overheard speech? *Child Development*, *67*(2), 621-634. doi: 10.2307/1131836

Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richards, J. A., Ma, L., …Topping, K. (2016). Effects of feedback on parent–child language with infants and toddlers in Korea. *First Language*, *36*(6), 549-569. doi: 10.1177/0142723716649273

Pearson, B. Z., Fernandez, S. C., Lewedeg, V., & Oller, D. K. (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics*, *18*(1), 41-58. doi: 10.1017/S0142716400009863

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880-891. doi: 10.1111/desc.12172

Rescorla, L., Dahlsgaard, K., & Roberts, J. (2000). Late-talking toddlers: MLU and IPSyn outcomes at 3;0 and 4;0. *Journal of Child Language*, *27*(3), 643-664. doi: 10.1017/S0305000900004232

Richards, J. A., Xu, D., Gilkerson, J., Yapanel U., Gray S., & Paul, T. (2017). Automated assessment of child vocalization development using LENA. *Journal of Speech, Language, and Hearing Research*, *60*(7), 2047-2063. doi: 10.1044/2017_JSLHR-L-16-0157

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*(5), 1762-1774. doi: 10.1111/j.1467-8624.2012.01805.x

Roy, B. C., Frank, M. C., & Roy, D. K. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, *31*(31), 2106-2111.

Sacks, C., Shay, S., Repplinger, L., Leffel, K. R., Sapolich, S. G., Suskind, E., …Suskind, D. (2013). Pilot testing of a parent-directed intervention (Project ASPIRE) for underserved children who are deaf or hard of hearing. *Child Language Teaching and Therapy*, *30*(1), 91-102. doi: 10.1177/0265659013494873

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*(3), 672-686. doi: 10.1017/S0305000912000141

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659-673. doi: 10.1111/j.1467-7687.2012.01168.x

Suskind, D., Leffel, K. R., Hernandez, M. W., Sapolich, S. G., Suskind, E., Kirkham, E., …Meehan, P. (2013). An exploratory study of "quantitative linguistic feedback": Effect of LENA feedback on adult language production. *Communication Disorders Quarterly*, *34*(4), 199-209. doi: 10.1177/1525740112473146

Thiemann-Bourque, K. S., Warren, S. F., Brady, N., Gilkerson, J., & Richards, J. A. (2014). Vocal interaction between children with Down Syndrome and their parents. *American Journal of Speech-Language Pathology*, *23*(3), 474-485. doi: 10.1044/2014_AJSLP-12-0010

VanDam, M., Ambrose, S. E., & Moeller, M. P. (2012). Quantity of parental language in the home environments of hard-of-hearing 2-year-olds. *The Journal of Deaf Studies and Deaf Education*, *17*(4), 402-420. doi: 10.1093/deafed/ens025

Velleman, S. L., & Vihman, M. M. (2002). Whole-word phonology and templates: Trap, Bootstrap, or Some of Each? *Language, Speech, and Hearing Services in Schools*, *33*(1), 9-23. doi: 10.1044/0161-1461(2002/002)

Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., …Gray, S. (2010). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*, *40*(5), 555-569. doi: 10.1007/s10803-009-0902-5

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143-2152. doi: 10.1177/0956797613488145

Zhang, Y., Xu, X., Jiang, F., Gilkerson, J., Xu, D., Richards, J. A., … Topping, K. J. (2015). Effects of quantitative linguistic feedback to caregivers of young children: A pilot study in China. *Communication Disorders Quarterly*, *37*(1), 16-24. doi: 10.1177/1525740115575771

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool Language Scale* [4th ed]. San Antonio, TX: The Psychological Corporation.

Computational Linguistics and Chinese Language Processing
Vol. 25, No. 1, June 2020, pp. 103-122

應用多跳躍注意記憶關聯於記憶網路之研究

A Research of Applying Multi-hop Attention and Memory Relations on Memory Networks

詹京翰*、劉立頌*、李俊宏+

Jing-Han Zhan,Alan Liu, and Chiung-Hon Lee

摘要

機器學習與深度學習近年發展越來越迅速，在自然語言處理任務上取得相當大的突破。透過類神經網路可以實現複雜的語言任務，如文章分類、摘要提取、問答任務、機器翻譯、圖片說明生成等。本論文以記憶網路做為研究目標、問答任務作為驗證應用。模型將先驗知識保存於記憶中，再透過注意力機制(Attention Mechanism)找出與問題相關的記憶，並推理出最終答案。問答任務數據集採用 Facebook 所提供的 bAbI 數據集，其中共有 20 項不同種類的問答任務，可驗證模型在不同任務的準確率。此研究透過記憶間的關聯計算，縮減記憶關聯的數量，除了下降 26.8%權重的計算量外，也能提高模型的準確率，於實驗中最多可提高約9.2%左右。同時實驗採取較小的數據量作為驗證目標，改善即使在數據集不足的情況也能達到相當程度的改善效果。

Abstract

With the rapid advancement of machine learning and deep learning, a great breakthrough has been achieved in many areas of natural language processing in recent years. Complex language tasks, such as article classification, abstract extraction, question answering, machine translation, and image description generation, have been solved by neural networks. In this paper, we propose a new

*國立中正大學電機工程學系
Department of Electrical Engineering, National Chung Cheng University
+南華大學資工系
Department of Computer Science and Information Engineering, Nanhua University
E-mail: hoe8624@gmail.com; aliu@ee.ccu.edu.tw; chlee@nhu.edu.tw
The author for corresspondence is Chiung-Hon Lee.

model based on memory networks to include a multi-hop mechanism to process a set of sentences in small quantity, and the question-answering task is used as the verification application. The model saves the knowledge in memory first and then finds the relevant memory through the attention mechanism, and the output module reasons the final answer. All experiments have used the bAbI dataset provided by Facebook. There are 20 different kinds of Q&A tasks in the data set that can be used to evaluate the model in different aspects. This approach reduces the number of memory associations through the calculation of associations between memories. In addition to reducing the calculation weight of 26.8%, it can also improve the accuracy of the model, which can increase by about 9.2% in the experiment. The experiments also used a smaller amount of data to verify the system for improving the case of insufficient data set.

關鍵詞：記憶網路、多點跳躍網路、關係網路、注意力機制

Keywords: Memory Networks, Multi-hop Networks, Relation Networks, Attention Mechanism

1. 緒論 (Introduction)

深度學習研究近年大幅成長，其中記憶網路模型於文字相關的自然語言任務受到了不少關注。在自然語言領域中，聊天機器人、問答任務等，都具有序列資料的特性，也就是文句字詞有時間先後的關係，因此計算的過程需要給與模型詞序資訊，或是依照順序輸入至模型內。其中記憶模型使用外部記憶的方式儲存文本或先驗知識，推理時再從中找出與問題關聯性高的記憶內容，可以避免在計算過程中損失重要的資訊。其中結合了注意力機制的應用，使輸出模組可以根據現在的問題關注重要的記憶內容，推理得出正確的答案。

應用於語言模型領域中的記憶網路模型有單跳躍注意機制(Single-hop attention) 與多跳躍注意機制(Multi-hop attention) 的推理方式，因記憶與記憶間相互獨立，在數據量足夠大的狀況下已可學習到不錯的效果。但在數據量較小的前提下則較為無法學習數據內的資訊。為提高模型學習與推理的能力，提高記憶儲存效率與模型的推理方法則顯得相當重要。

本研究結合自然語言常用資料集與深度學習的工具，嘗試結合不同理論，強化語言理解與推理能力，並分析實驗結果的表現。實驗採取較小的數據量作為驗證目標，改善即使在數據集不足的情況也能達到相當程度的改善效果。目的歸納如下兩點。

(一) 研究多跳躍注意機制對於記憶網路預測的表現。

(二) 研究記憶網路記憶關聯提取對於推理能力的提升。

本論文研究在小數據集的前提下，不同的機制對於問答系統模型的影響。在研究中我們將關係網路的概念，以關聯記憶的形式與記憶模型結合，於 bAbI 數據集(Weston,

Bordes, Chopra & Mikolov, 2016)20 項任務中驗證，準確率最多可提高約 9.2%左右的準確率。關聯提取的部分還有降低權重的功用，相比於保存所有關聯計算，平均每項任務可下降 3 萬個權重數量，整體下降 26.8%權重的計算量。

我們在接下來的小節討論與整理記憶網路相關領域研究文獻；第三節為研究方法與設計，對於本論文研究方式與方法做一系列的整理與說明；第四節則為實驗結果與分析，比較改善前後模型的表現，驗證所採用方法的可行性與價值；最後一節為結論與建議，總結本論文所採用方法的優缺點，以及未來可嘗試的方向。

2. 文獻回顧 (Literature Review)

記憶網路(Memory Networks)主要運用於問答任務與情感分析等應用上，採用外部記憶的方式儲存先驗知識，透過注意力機制找到與問題相關的記憶內容，再利用推理模組從問題與相關記憶得出最終答案。記憶網路由許多模組組合而成，各個部分可由設計者採用不同方式實現，本小節介紹記憶網路相關研究，以及語言模型的相關理論。

2.1 注意力機制 (Attention Mechanism)

注意力機制(Attention mechanism)最早應用於圖像領域，論文(Bahdanau, Cho & Bengio, 2015)結合類神經網路模型，將其運用於機器翻譯任務上，首次將注意力機制應用於自然語言處理領域上。

自從編碼器解碼器架構(Encoder-Decoder) (Cho *et al*., 2014)的提出，改善了單個 RNN (Recurrent Neural Networks) (Elman, 1990)長期記憶的不足，並提升了在自然語言處理領域各種任務的效果。但因輸出解碼過程在不同時間步所輸入的編碼向量相同，導致在轉換的過程中容易損失許多訊息，若提高向量大小則會導致計算量增加，而注意力機制的提出可以有效提高模型編碼的效率。

2.2 記憶網路 (Memory Network)

記憶網路(Memory Network, MemNN) (Weston, Chopra & Bordes, 2014)由Facebook人工智慧實驗室所提出，目的在提高類神經網路模型長期記憶能力，應用於序列性資料上。如保存問答任務的先驗知識、聊天的語境訊息等。過往在處理序列性資料時，RNN 可以有效的處理短期時間先後關係，每個時間步都會參考上一個時間步輸出的結果，但其只透過記憶單元儲存重要資訊，隨時間步推移更新記憶單元內容，對於長序列的訓練過程中可能會有梯度消失(gradient vanishing) 與梯度爆炸(gradient exploding) 的問題發生，造成 RNN 在長期記憶中表現不是很好。即使後來長短期記憶模型(Long Short-Term Memory, LSTM) (Hochreiter & Schmidhuber, 1997)相對提高了長期記憶能力，但對於更大的序列仍然有其限制存在。

記憶網路的限制在於訓練過程需要透過強監督的方式學習(Strong-Supervised Learning)，訓練用數據需要提供與查詢相關的標註句子，然而並非所有數據集都有支持

事實的標註，並不利於將此模型應用到不同的數據集或不同任務上。端對端記憶網路(End-to-End Memory Networks, MemN2N)模型(Sukhbaatar, Szlam, Weston & Fergus, 2015)，在記憶網路模型的基礎上修改與完善，使其能以端對端的方式完成學習。透過弱監督方式(Weak-Supervise Learning) 即可完成訓練，有利模型的擴展並應用到不同的任務或資料集上。此模型利用軟性注意力機制(Soft Attention Mechanism) 來估計每條記憶與問題相關的程度，並使用相關性高的記憶計算出最後的輸出。

動態記憶網路(Dynamic Memory Networks, DMN)模型(Kumar *et al*., 2016)，將大部份自然語言處理領域的任務視為問答任務的一種。該模型以記憶網路為基礎進行改善，可透過端對端學習完成訓練，應用於問答任務、情感分析以及詞性標註等。模型架構與記憶網路模型相似，主要由四個模組所組成，分別為輸入模組、問題模組、情景記憶模組(Episodic Memory Module) 與應答模組。與前述記憶網路的不同在於編碼方式。此模型採用門控循環單元模型(Gate Recurrent Unit, GRU) (Chung, Gulcehre, Cho & Bengio, 2014)編碼，隨著時間步的推移更新隱藏狀態。相較於單純使用詞袋(Bags of word, BOW) 更可以表示出字詞之間的順序關聯。

在問答系統中加入知識庫(Knowledge Bases, KBs)可以有效的提高模型的知識儲存量，但其並不夠完整，無法支持不同類型的答案，由於數據的稀疏性，較難創建包含所有領域的 KB，不利於擴展到不同的領域。鍵值記憶網路(Key-Value Memory Networks)模型(Miller *et al*., 2016)使用鍵值(key-value) 的方式將文章中的編碼存取下來，架構基於端對端記憶網路模型，針對先驗知識的儲存提出不同方式編碼，應用於自然語言中問答的相關領域。

鍵值記憶網路模型與端對端記憶網路模型相似，最大的不同在於記憶的儲存方式。端對端記憶網路透過不同的嵌入矩陣對文本編碼，而鍵值記憶網路則透過鍵值的方式表示，以鍵記憶(key memory)與值記憶(value memory)形式儲存。鍵值記憶網路優點為在訓練網路之前，可先對先驗知識進行適合的編碼。即使是不同領域的知識，使用者也可選擇編碼方式，而不單純依賴於詞嵌入矩陣的訓練，在使用上有了更多的彈性。

遞歸實體網路(Recurrent Entity Networks, EntNet)模型(Henaff, Weston, Szlam, Bordes & LeCun, 2017)，記錄世界的實體與狀態於記憶中，當有新資訊輸入時，則根據輸入訊息更新相對應記憶單元，可應用於自然語言中的閱讀理解與問答任務中，其在 bAbI-10k 數據集與 Children'sBook Test (CBT)數據集 single hop 訓練中，較更早提出之方法表現為優。

我們依照論文(Sukhbaatar *et al*., 2015) (Henaff *et al*., 2017)中之方法，以一千筆訓練資料進行實驗發現，最初記憶模型有效改善長期記憶關係，可在 bAbI 資料集中通過 16/20 項任務。但需要透過強監督方式進行訓練，並不利於擴展應用。而弱監督訓練則只能通過 2/20 項任務，且錯誤率大幅增長。而端對端記憶網路模型透過弱監督方式訓練，相較於弱監督記憶網路，通過任務比例提昇，也大幅下降平均錯誤率。

而綜合前述論文所提供的數據，以一萬筆訓練資料為前提實驗，相較於前述以一千

筆資料訓練，模型相對通過更多的任務。端對端記憶網路模型通過了 17/20 項任務；動態記憶網路模型通過了 18/20 項任務。而首先通過所有任務的模型為遞歸實體網路模型，其平均錯誤率降低至 0.54%。如表 1 所示。

表 1. 不同模型於 10k 數據量實驗結果
[Table 1. Experimental results of different models with 10k data]

Model	**MemNN**	**MemN2N**	**DMN**	**EntNet**
Mean Error	39.2	4.2	6.395	0.54
Failed Tasks(error>5%)	17	3	2	0

雖然在表 1 中 EntNet 在 10k 數據量，錯誤率小於 5%的標準中通過所有任務，但其在部份任務中的錯誤率還是大於 4%。而且若是利用較少的 1k 的數據量進行訓練的話，正確度則會從原本的 99.5%降到 89.1%。因此在較少數據的情況下，模型的準確性還有可以改進的空間。若是模型能提高少數據下訓練的效果，可以減少訓練時間，與資料收集的成本。而若是要應用到其他的資料量較少的情況，也能有比較好的效果。

2.3 多跳躍注意機制 (Multi-hop Attention)

多跳躍注意(Multi-hop Attention)機制於端對端記憶網路模型中所提出，透過不斷比對問題與記憶得出問題答案，這個過程模擬人類在推理過程的思考方式，後續模型透過不同的方式實現多跳躍機制，用以強化模型的推理能力。

問題簡化網路模型(Question Reduction Networks, QRN) (Seo, Min, Farhadi & Hajishirzi, 2017) 模型架構為 RNN 的一種，可有效處理短期與長期序列關係。透過多輪讀取機制，逐漸簡化問題，達到深入語意理解的效果，最後推理得出最終答案並轉化為自然語言輸出。此外 QRN 模型中所提出的公式允許在遞歸神經網路時間軸上並行化，提升訓練與推理部分的效率。

AOA Reader 模型(Attention-over-Attention) (Cui *et al*., 2017)，應用於填空任務(Cloze-style questions)。與過往模型最大的不一樣在於透過不一樣的注意力機制組合，計算出權重預測最後的結果，而非使用單一一種注意力機制計算方法。模型透過雙向門控循環模型對先驗知識與問題進行編碼，將編碼向量點積相乘，經過 softmax 計算出詞彙的機率，此注意力機制的計算方法為許多模型通用方法，而此論文創新的地方為其不僅計算 document-to-query 的注意力數值，也計算 query-to-document 的注意力權重，最後利用兩者矩陣相乘得到最後注意力機制數值，並透過後續模型進行推理。

論文(Trischler *et al*., 2016)提出了 EpiReader 神經網路模型，用以解決自然語言任務中的填空問題。EpiReader 模型分為兩個部分，第一部分為提取模組(Extractor)，通過淺層文本與問題的比對，提取出若干個問題的可能候選答案；第二部分為推理模組(Reasoner)，通過更深層的語意比較候選答案與問題之間的關聯。提取模組從大量可能性中篩選出小部分候選答案，而推理模組則處理更精確的推理匹配部分。

神經語意編碼器模型(Neural Semantic Encoders, NSE) (Munkhdalai & Yu, 2016)架構，在過往多輪讀取機制多為固定步數，但並非所有的問題需要相同推理的步數。有些問題只需要簡單的詞句比對即可得出結論，有些問題則需要複雜的語意理解與深度推理，因此 NSE 利用動態步數調整模型以解決此問題。

整理論文(Sukhbaatar *et al*., 2015) (Henaff *et al*., 2017) (Seo *et al*., 2017)實驗數據以表格呈現，首先表 2 實驗在 bAbI 數據集上，以一千筆資料訓練，透過表格中可發現透過多跳躍機制可提高通過的任務數量或是降低平均錯誤率，不同多跳躍步數也會影響結果。

表 2. QRN 與 MemN2N 模型不同 hop 實驗數據
[Table 2. Different hop experimental data of QRN and MemN2N models]

Model	MemN2N			QRN	
	1 hop	2 hop	3 hop	2r	3r
Mean Error	9.58	8.45	8.15	9.9	11.3
Failed Tasks(error>5%)	17	11	11	7	5

以兒童圖書測試數據(Children's Book Test, CBT)為實驗數據，表 3 整理幾種單跳躍與多跳躍模型實驗結果，目前效果最優為多跳躍模型，因此多跳躍機制成為目前研究領域的趨勢，而本論文研究也將嘗試以不同方式將多跳躍注意機制結合單跳躍模型。

表 3. Single 與 Multi hop 不同模型於 CBT 數據及實驗結果
[Table 3. Different models of Single and Multi hop experimental results]

Model		**Named Entities**	**Common Nouns**
Single Pass	Kneser-Ney Language Model + cache	0.439	0.577
	LSTMs (context+query)	0.418	0.560
	Window LSTM	0.436	0.582
	EntNet (general)	0.484	0.540
	EntNet (simple)	0.616	0.588
Multi Pass	MemNN	0.493	0.554
	MemNN+self-sup	0.666	0.630
	EpiReader	0.697	0.674
	AoA Reader	0.720	0.694
	NSE	0.732	0.714

2.4 關係網路 (Relation Network)

關係網路(Relation Network) (Santoro *et al*., 2017)目的在於透過加入對物件、實體或是語句之間的關係計算，提供更多訊息給後續推理模組進行推理。關係網路應用於圖像問答(Visual Question Answering, VQA)，使用簡單的模型來建構物體之間的聯繫，核心概念在於最終答案與成對的對象具有一定的關聯性，而問題也會影響對成對對象的查詢。其透過神經網路計算任意對象兩兩之間的潛在關係。

關係網路的優勢在於其架構簡單，使用彈性大，可將其插入於不同的模型裡，提高了模型推理的能力，可應用於關係推理相關任務上。前述論文在實驗中使用問答相關資料集做為驗證，在 bAbI dataset 二十個任務中通過了十八個，而在 Sort-of-CLEVR 中取得最優的結果，且超過人類所能達到的分數。

RelNet (Bansal, Neelakantan & McCallum, 2017)中將計算兩兩物件之間關係的概念帶入遞歸實體網路中，此論文將關係概念用來計算記憶與記憶之間的關聯。過往遞歸實體網路模型與記憶網路相關模型，記憶的儲存相互之間獨立，而 RelNet 模型透過關係計算將記憶彼此連結起來。計算記憶狀態儲存的公式與原模型相同，差別在多加上了計算不同記憶之間的關聯，並應用於最後的推理計算上。

遞歸關係網路(Recurrent Relational Networks, RRN)模型(Palm, Paquet & Winther, 2017)運用節點關係解決數獨的問題，在 9*9 的數獨內共有 81 個節點，每個節點都需要考慮同一行、同一列與同個方框內的訊息，不能出現同樣的數字。此模型對每個節點初始化的狀態為{$h1$,$h2$,…,$h81$}，透過多層感知器(MLP) 計算每個節點之間的關聯，將計算出的關係數值相加，用以更新結點的狀態，每個節點的更新考慮上一個時間步的狀態、輸入以及關係數值。此模型除了應用於數獨上，也在 bAbI 數據集、Pretty-CLEVR1 表現優秀。

以(Cui *et al*., 2017) (Trischler *et al*., 2016)兩篇論文實驗所提供的資料為基礎，使用 bAbI 數據集中 10k 數據量訓練，並平均 20 項任務的錯誤率比較結果顯示於表 4。由實驗可以了解到，加入關係計算能有效的提升模型的訓練結果與計算。

表 4. relation 相關模型比較
[Table 4. A comparison of relation models]

Method	Mean Error Rate (%)
RRN	0.46±0.77
RelNet	0.29
EntNet	9.7±2.6

2.5 預訓練模型 (Pre-trained Model)

近年來的語言模型研究使用大量文章預訓練(Pre-train)通用語言模型，然後再根據具體應用，用 supervised 的訓練資料，微調(Fine-tuning)模型，使之適用于具體應用，來提昇模

型的效能。

其中 BERT 模型(Devlin, Chang, Lee & Toutanova, 2018)以及後續發表，讓 BERT 更小，訓練更快的 Albert 模型(Lan *et al*. 2019)，被廣泛地應用於問答任務，且得到相當優異的成果。

這種結合預訓練模型再加上後續的微調訓練的方式，可以讓許多的自然語言處理任務得到極大幅度的效能提升，也讓我們可以用更小的資料就能訓練出極好的效果。

3. 研究方法 (Research Method)

記憶網路透過外部記憶的保存，強化長期記憶的能力，經過注意力機制找尋與問題相關的記憶槽，推理出對應的答案。記憶槽與記憶槽之間相互獨立運作，針對不同的實體保存相關訊息，對於需要多項記憶交互推理的複雜任務，推理模組較無法使用足夠的訊息輸出正確答案。本研究則透過記憶之間的關聯計算與多跳躍機制推理，嘗試提高模型記憶儲存與推理能力，並以問答相關任務作為驗證，模型需要具備語言理解以及推理的能力。接下來我們將介紹本論文整體模型的架構，並介紹各模組內的設計。

3.1 模型架構 (Model Architecture)

本論文之模型架構以 EntNet 模型(Henaff *et al*., 2017)為基礎，用固定大小記憶單元保存輸入數據的實體與相關屬性，記憶內容則隨著輸入句子即時更新。模型架構在記憶槽(memory slot)之間加上記憶關聯的計算與提取，保存於關聯槽(relation slot)內。原模型記憶槽與記憶槽之間相互獨立運作，但記憶與記憶之間應具有一定的關聯，透過關聯計算可將各記憶槽內容聯繫起來。模型主要可分為三個部分，分別為 Encoder 模組，負責將輸入的自然語言詞句編碼為向量的形式，以利電腦後續計算；動態記憶模組在每次句子輸入時，更新記憶槽內所儲存的資訊，再透過計算記憶槽彼此間的關聯來更新關聯槽，記憶槽大小與關聯槽大小相等；最後輸出模組根據問題，從記憶槽與關聯槽中推理出最後的答案。整體架構如下方圖 1 模型架構圖所示。相較 EntNet，在本架構中我們加入了 Relation memory 的部份，以嘗試透過結合記憶關聯計算，增加記憶間的聯繫。

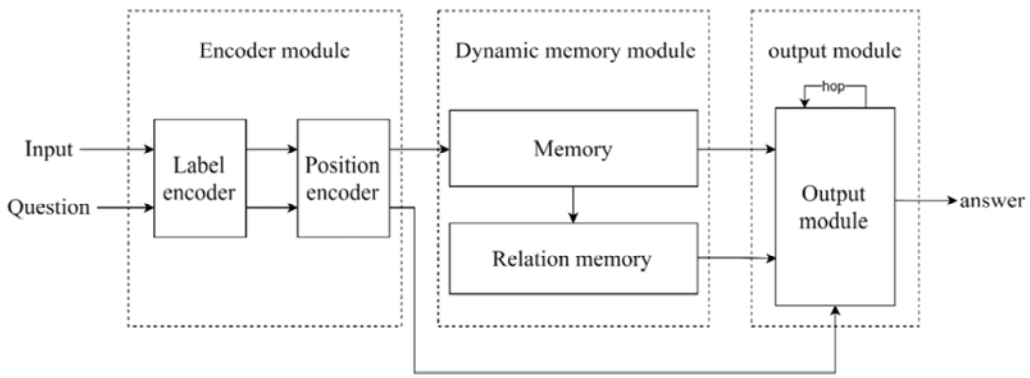


圖 1. 模型架構圖.
[Figure 1. Diagram of the model architecture]

3.2 Encoder模組 (Encoder Module)

此模型應用於問答任務上，所使用的數據皆為自然語言形式，自然語言無法直接輸入至電腦計算，因此在輸入至模型訓練前須先轉換為編碼的形式，以利後續的運算，此模組分為兩個步驟編碼，透過 Label encoding 初步將句子轉換為數字，再經過 Position encoding 給予字詞於整體句子的相對位置資訊。首先建立詞彙庫，將數據集中所有用到的詞彙對應到一個固定的數字，詞彙庫中詞彙量與數字量相等，不會新增多於欄位的詞彙。完成詞彙庫的建立後，將數據集根據詞彙庫轉換為數字的形式表示。範例如下所示:{}內為不同詞彙所對應的編號，[]為每個句子依照詞彙庫轉換為對應編號。

{hallway:1,John:2,the:3,to:4,went:5,.:6}

John went to the hallway.→[2,5,4,3,1,6]

基本數值轉換後，雖然句子都以數字形式表示，但編碼並無相對應的意義，以上面例子為例，hallway 數值為 1、John 數值為 2，hallway 的兩倍為 John，這並無法解釋詞與詞之間的關係，所以這些數字將會再次轉換為模型訓練出來的向量，而數字是為了將相同的詞彙轉換為相同的向量。首先根據詞彙庫的大小，建立與詞彙量相等量的可訓練向量，每個詞彙有對應的向量，並在整體模型訓練的過程中一起更新向量數值，透過使用自然語言相關數據集，訓練詞彙對應的向量。

位置編碼（Position encoding）目的在於賦予字詞之間順序的關係，自然語言語意會根據詞彙的順序而有所不同，若是使用 BOW 的方式編碼，詞彙出現在句子任意位置對於編碼並沒有不同，但於實際語言相同詞彙於不同位置，對於語意理解可能會有很大程度的不一樣，如下方範例所示，相同用詞於不同位置所得出的語意相差甚大。

John likes Mary.≠Mary likes John.

本實驗位置編碼採用訓練的方式達成，透過 mask 的方法為語句加入順序關係，如下方公式(1)所示。$\{e_1,\cdots,e_k\}$為句子中每個詞彙的編碼向量，$\{f_1,\cdots,f_k\}$是需要學習的 multiplicative mask，為可訓練的向量。使用這個 mask 的目的在於加入位置資訊。透過訓練的過程更新權重，當相同詞彙於不同的位置時，所乘上的f_k向量也會有所不同。透過這樣的方式將位置的訊息加入至編碼中，最後將其加總表示整體句子的向量。

$$s_t = \sum_i f_i \odot e_i \tag{1}$$

3.3 動態記憶模組 (Dynamic Memory Module)

動態記憶模組由兩個部分組成，分別為記憶儲存槽與關係儲存槽，記憶槽以 key-value 的形式保存，key 保存實體、value 保存狀態，更新完記憶後再依據記憶與記憶之間的關聯更新關係槽內容，記憶槽與關係槽數量相等，架構如下方圖 2 所示。本架構與 EntNet 的差別為本研究加入了關係儲存槽 r。新加入部份在圖中以較粗之線條繪出。

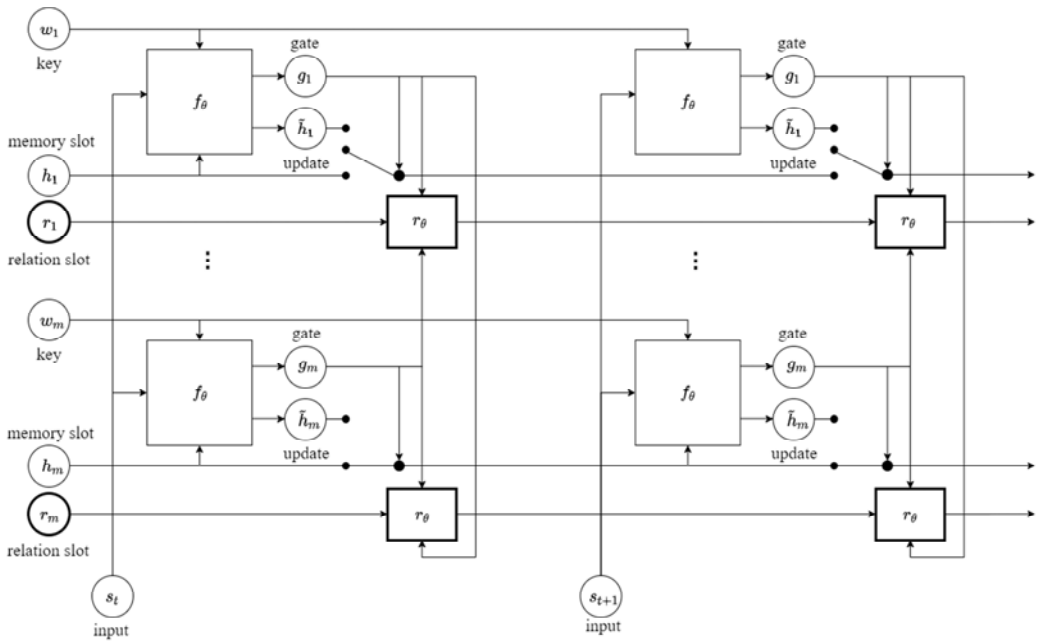


圖 2. 動態記憶模組架構圖
[Figure 2. Diagram of the dynamic memory module architecture]

數據編碼完後以向量形式表示每個句子s_t。t 為不同時間步的句子，依照順序輸入至模型內更新記憶槽與關係槽。每個記憶槽由 key 和 value 組成,分別為 wi 和 hi，以 key-value 的形式保存資訊。key 負責保存實體、value 負責保存狀態，範例如下方所示。範例中 key 保存了 John 這個實體，value 保存了 John 所做的動作，每個記憶槽都有自己的 key 與 value 向量，透過輸入數據與 key-value 的比對可找到此次狀態更新應該更新於哪個記憶槽。

John went to hallway.=>{key:John,value:went to hallway}

當每個句子輸入至模型內時，系統透過公式(2)計算句子與 key、value 之間的關係。σ 表示 sigmoid activation function，g_i 是 gate，輸出數值將介於 0~1 之間，此數值為門控機制，用以決定更新與保存多少記憶內容。g_i 由 w_j 和 h_j 決定。前者表示與關鍵字的匹配程度，後者表示與 memory 內容的匹配程度。與此記憶槽實體越相關的語句，所計算出的數值會越高。公式(3)為 RNN 的計算公式，用以計算出輸入句子的內容。$\widetilde{h_j}$表示需要新增到已有的 memory 中的狀態值。 ϕ 可以是任意的 activation function，實驗進行時使用的是 PReLU。U、V、W皆為可訓練權重,並且所有的 gated RNN 共享這些引數，於整體模型訓練時一起更新。公式(4)更新每個記憶槽 h_j 內容，將原記憶槽內容門控與新記憶相加，透過門控數值控制更新的幅度。公式(5)用以遺忘不必要資訊，若是不斷將新記憶加入記憶槽內，向量數值將會越來越大，透過除上 normalization 數值，保持記憶向量數值範圍。

$$g_i \leftarrow \sigma\left(s_t^T h_j + s_t^T w_j\right) \tag{2}$$

$$\widetilde{h_j} \leftarrow \emptyset\left(Uh_j + Vw_j + Ws_t\right) \tag{3}$$

$$h_j \leftarrow h_j + g_j \odot \widetilde{h_j} \tag{4}$$

$$h_j \leftarrow \frac{h_j}{\|h_j\|} \tag{5}$$

模型中每個關係槽保存對應記憶與所有其他記憶的關係。更新完記憶槽將會得到此次輸入對於每個記憶的門控數值，將門控數值兩兩相乘計算彼此間的關聯，相同的門控計算加總於相同關係門控數值g_i^r，如公式(6)所示。此公式目的在於將相同關聯對象保存在一起。其中g_i^m，g_j^m依據相同關聯對象，選擇相關的記憶槽。例如模型有 20 個記憶槽，記憶槽 1 與其他所有的記憶槽計算出 19 個關係，將這些關係保存於第一個關係槽。如此，後續推理時可直接從對應關係槽找出與其他記憶的關聯，如公式(7)所示。

$$g_{ij}^r = g_i^m g_j^m \sigma\left(< s_t, r_{ij} >\right) \tag{6}$$

$$g_i^r = \sum_{i,j} g_{ij}^r \tag{7}$$

公式(8)計算關係更新內容，向量 A、B 為可訓練權重，根據原本關係槽內容r_{ij}與輸入句子s_t所需要更新的內容，最後以 PReLU 為 activation function。公式(9)為關係更新。將關係門控數值乘上新的關係內容，並加上原本的關係槽內容用以更新關係槽資訊。

$$\tilde{r}_{ij} \leftarrow PReLU(Ar_{ij} + Bs_t) \tag{8}$$

$$r_{ij} \leftarrow r_{ij} + g_{ij}^r \odot \tilde{r}_{ij} \tag{9}$$

3.4 輸出模組 (Output Module)

動態記憶模組更新完記憶槽h_i與關係槽r_i，將最後狀態保存給輸入模組推理使用。公式(10)將同個記憶槽h_i與關係槽r_{ij}向量並接在一起，並乘上可訓練權重C計算出記憶m_i。然後再透過注意力機制計算與 query 相關的p_i數值，數值越高代表相關性越高，如公式(11)所示。

$$m_i = C\left[h_i; r_{ij}\right] \tag{10}$$

$$p_i = Softmax(q^T m_i) \tag{11}$$

將注意力數值乘上對應記憶，越相關記憶數值相對會越高如公式(12)所示，以保留與問題相關重要資訊。系統最後根據公式(13)推理出最後問題的答案，其中 R 跟 H 為參數矩陣。query 問題會依照訓練時的方式被編碼成 k 個維度的向量 q。本研究使用數據為問答任務，系統根據詞彙庫輸出最有可能的答案 y。

$$u = \sum_i p_i m_i \tag{12}$$

$$y = R\emptyset(q + Hu) \tag{13}$$

從第二章文獻探討可發現，多跳躍機制有助於提升模型推理能力，本研究嘗試將此概念加入推理模組，將上方記憶與注意力權重相乘加總的向量 u，與原 query 向量相加，作為新的 query 向量，重複公式(11)(12)計算，每多一次計算 hop 數增加 1，原本推理模組為 hop1，重複一次計算為 hop2，依此類推，如公式(14)所示。

$$q = q + u \tag{14}$$

3.5 討論 (Discussion)

本研究以 EntNet 模型為研究基礎，嘗試透過結合記憶關聯計算，增加記憶間的聯繫，而非不同記憶槽獨立運作。在 3.1 節中介紹整體模型架構。主要 Encode 模組、動態記憶模組以及輸出所組成。3.2 節介紹文字如何轉換為向量形式表示。從建立基本詞彙庫到訓練詞彙向量的過程。3.3 節中介紹動態記憶模組細節。負責記憶保存與更新的部分，除了原記憶模型的記憶槽外，將關係的計算加入模型內，使不同的記憶槽可透過關聯計算，計算彼此的關係，記憶間的關聯計算隨著記憶槽數量而快速增長，將其提取為同樣記憶數量的關聯槽，可降低權重與計算量。3.4 節為輸出模組的細節。當記憶模組將先驗知識保存後，輸出模組針對問題從記憶中取出相關的部分，並推理出最後的答案。研究方法中的關係計算與多跳躍的推理方法，可泛化應用到不同的記憶網路架構，或是具有記憶保存的架構的模型上。

4. 實驗 (Experiments)

本研究所有實驗選擇以 bAbI dataset 做為實驗驗證的數據集，此數據集為自 Facebook AI Research (FAIR)所提供的綜合閱讀理解和問答資料集。選擇此數據集驗證目的有四點，分別如下：

(一) 數據集包含了二十種任務，可從不同面向測試模型的優勢與劣勢。

(二) 為問答與自然語言理解常用數據集，有許多不同模型實驗數據可參考比較。

(三) 包含英文、印地語與改組(人類不可閱讀) 等數據，可了解語言模型應用於不同自然語言之效果。

(四) 於 20 項任務中提供 1k 資料量與 10k 資料量，可實驗數據量多寡對於模型學習的影響。本研究目標為於數據集 1k 的前提下，提升模型訓練效果。

以下實驗為求準確性，以交叉驗證方式，透過不同訓練數據做驗證，並平均多次訓練結果。實驗使用 1k 數據量訓練模型，並將 10k 數據切分多份 1k 檔案，透過多次實驗驗證改善效果。

4.1 實驗一(多點跳躍訊息推理) (Experiment 1: Multi-hop Reasoning)

實驗目的：

由前面的實驗及討論中，我們可以看出多跳躍針對複雜的問答推理，普遍相較於單跳躍對於推理結果效果更好，而 EntNet 模型屬於單跳躍模型，本實驗嘗試將多跳躍的概念應用於輸出模組中，嘗試增強遞歸神經網路推理能力。

實驗內容：

本實驗將引入端對端記憶網路的多跳躍推理公式，選用此方法原因在於端對端網路與 EntNet 模型相似，都具備記憶單元保存資訊，其他模型架構方法差異較大。輸出模組負責推理答案，透過注意力機制找尋與此次問題相關的記憶，依據記憶內容推理出最終答案，而經過一次注意機制的計算為單跳躍，本實驗嘗試增加跳躍數量。如 3.4 小節中的公式(14)，注意力機制計算出的數值與計算所使用的 query 相加，做為新的 query 再次與記憶做注意力機制的計算，每多做一次跳躍數增加 1，實驗將比較雙跳躍與原先單跳躍的差異。實驗結果整理於表 6 內之 Multi hop 欄位。

由實驗數據中可以看出，增加跳躍數量並沒有增加模型的準確率，甚至部分的準確率相較於原模型有下降的趨勢，平均錯誤率反而上升。分析訓練出的模型於訓練資料、測試資料的表現，可以看出有過擬合的趨勢，複雜化推理模組無法提升效果。推測為記憶模組所保存的資訊不足，無法提供足夠的資訊給與推理模組進行後續的推理。因此設計實驗二透過記憶關聯的計算，與關係槽的保存提升模型外部記憶保存的能力，嘗試保存更關的資訊是否能提升模型的效果。

4.2 實驗二(記憶關聯) (Experiment 2: Memory Relation)

實驗目的：

此實驗主要應用於 EntNet 模型的動態記憶模組，根據實驗一的實驗結果，可發現複雜化推理模組無法提升推理能力，因此實驗二將關聯計算加入記憶之間。相較於原本記憶分別獨立保存訊息，記憶關聯的計算能將相關記憶串連起來。如同人類記憶保存並非把所有部分完全獨立，透過聯想可聯繫到不同的想法或記憶。本實驗額外保存記憶與記憶的間的關聯，用以增加推理模組可用訊息，增加模型推理能力。

實驗內容：

本實驗將記憶槽兩兩根據公式(6)計算出關係門控數值，用以決定此次數據輸入對於關係槽更新多寡。實驗記憶槽數量與原模型相同，使用 20 個記憶槽保存重要資訊。另外額外加入關係槽用以保存對應記憶槽的關係，如第一個記憶槽與其他所有記憶槽關聯計算，保存於第一個關係槽，關係槽數量與記憶槽數量相等，確保關係保存不會隨記憶槽增長而大量增加。關聯計算透過 Cm 2 排列組合計算，如公式(15)所示。20 個記憶槽計算出

190 個關係，再將 190 個關係分別保存到對應的關係槽內。實驗結果如表 6 中之 Relation slot 欄位所示。

$$number\ of\ relation = C_2^m \tag{15}$$

bAbI 數據集於不同任務有不同的推理難度，有些任務需要結合多項先驗知識交叉推理才能得出答案。從實驗數據中可以看出，透過關聯計算能有效降低平均錯誤。相較於原先記憶獨立保存，此方法可以更有效地從數據中學習詞句間的關聯。但也並非所有任務都有明顯改善。任務 2 改善效果最為明顯，數據特性是找到兩項支持事實的句子，才能推理出問題的答案，而關聯的計算剛好是兩兩記憶槽的計算，效果顯著於提升任務 2。但任務 3 更多的支持事實句子卻無法提升。推論為數據集太小以及關聯計算的方法的影響。

另於表 5 比較保存所有關聯計算(如 RelNet 模型即是採用此方法)，與關聯提取兩種方法。對於權重的使用量，權重數量越多代表所需要 GPU 所需要的計算量越大。本實驗中 20 個記憶槽將會計算出 190 個關係，此實驗將 190 個關係數值提取於 20 個關係槽保存，目的只保存重要的資訊。如此可以大為減少模型權重的數量 6 萬個，較原使用所有關聯的權重數量降低了 26.8%。實驗結果的數據可以發現在不同任務提升效果不同，也有部分準確率是些微下降，但整體仍以提升為主。

表 5. 保存所有記憶關聯與提取方法權重數量比較
[Table 5. A comparision between all relation method and relation slot method]

Task	All relation method	Relation slot
Task 1: Single Supporting Fact	110000	80000
Task 2: Two Supporting Facts	112900	82900
Task 3: Three Supporting Facts	113400	83400
Task 4: Two Argument Relations	109400	79400
Task 5: Three Argument Relations	115200	85200
Task 6: Yes/No Questions	113400	83400
Task 7: Counting	115100	85100
Task 8: Lists/Sets	115100	85100
Task 9: Simple Negation	111100	81100
Task 10: Indefinite Knowledge	111500	81500
Task 11: Basic Coreference	111600	81600
Task 12: Conjunction	110400	80400
Task 13: Compound Coreference	111600	81600
Task 14: Time Reasoning	111700	81700

Task 15: Basic Deduction	109700	79700
Task 16: Basic Induction	109500	79500
Task 17: Positional Reasoning	111100	81100
Task 18: Size Reasoning	110700	80700
Task 19: Path Finding	113200	83200
Task 20: Agent's Motivations	113600	83600
Sum of all task parameters	2240200	1640200
Mean parameters	112010	82010

4.3 實驗三(自我記憶關聯) (Self memory Relation)

實驗目的：

實驗二中將計算出的 190 個關係提取為與記憶槽數量相等的關係槽，透過這樣的方式輸出模組不需要將所有關係都計算過，在動態記憶模組進行保存時，即可篩選出重要的關聯資訊進行保存，並非所有關係數值都需要被保存，輸出可以只專注於重要的資訊推理答案。此實驗嘗試將記憶關聯直接更新於原記憶槽內，而不另外透過關係槽保存關係資訊，實驗是否透過記憶的自我關聯更新，即可提升記憶槽保存內容的品質。

實驗內容：

關聯計算方法同實驗二，差別在於實驗三更新的目標，為記憶槽本身所保存的內容，原輸入關聯槽的部分改成記憶槽本身，輸出所更新的目標也是記憶槽。如公式(16)~(18)所示，計算此次輸入句子在兩兩記憶槽間的關係，方法與實驗二相似，而公式(19)透過門控數值決定此次更新的多寡，而更新的目標是記憶本身。實驗結果如表 6 中之 Self memory 欄位所示。

$$g_{ij}^r = g_i^m g_j^m \sigma(< s_t, h_i >) \tag{16}$$

$$g_i^r = \sum_{i,j} g_{i,j}^r \tag{17}$$

$$\tilde{r_i} \leftarrow PReLU(Ah_i + Bs_t) \tag{18}$$

$$h_i \leftarrow h_i + g_i^r \odot \tilde{r}_i \tag{19}$$

表 6. 原模型與多跳躍、關係計算、自我關聯更新實驗結果(錯誤率)
[Table 6. Error rate of different models]

Task	**Original model**	**Multi hop (hop2)**	**Relation slot**	**Self memory**
Task 1: Single Supporting Fact	0.00%	0.00%	0.00%	0.00%
Task 2: Two Supporting Facts	20.80%	28.40%	11.60%	52.30%
Task 3: Three Supporting Facts	58.70%	56.70%	62.90%	62.10%
Task 4: Two Argument Relations	0.10%	0.20%	0.00%	0.00%
Task 5: Three Argument Relations	1.20%	1.20%	1.40%	17.20%
Task 6: Yes/No Questions	3.60%	3.50%	1.90%	11.40%
Task 7: Counting	10.10%	10.10%	6.90%	23.40%
Task 8: Lists/Sets	1.30%	2.20%	1.70%	9.10%
Task 9: Simple Negation	0.40%	0.00%	0.00%	35.60%
Task 10: Indefinite Knowledge	0.50%	3.70%	0.80%	3.80%
Task 11: Basic Coreference	8.90%	8.00%	4.20%	7.50%
Task 12: Conjunction	0.00%	0.00%	0.00%	0.60%
Task 13: Compound Coreference	5.60%	5.60%	6.20%	5.80%
Task 14: Time Reasoning	20.50%	21.30%	20.60%	55.90%
Task 15: Basic Deduction	5.10%	29.70%	0.00%	45.80%
Task 16: Basic Induction	50.00%	50.70%	51.00%	51.20%
Task 17: Positional Reasoning	41.20%	39.00%	37.70%	39.40%
Task 18: Size Reasoning	8.00%	7.60%	6.20%	8.50%
Task 19: Path Finding	87.80%	86.80%	85.30%	87.60%
Task 20: Agent's Motivations	0.90%	0.20%	0.90%	2.90%
Mean Error	**16.24%**	**17.74%**	**14.96%**	**26.00%**
Failed Tasks(error>5%)	**11**	**11**	**9**	**15**

相較於實驗二加入關聯計算的提升，此實驗平均準確率反而下降不少。推論為關聯計算雖能提升模型推理效果，但若是直接更新記憶槽本身，反而會造成記憶保存的效果下降，目前仍是將兩者分開保存效果較好。但根據表 7 權重的數量比較，可以發現自我記憶關聯更新可以下降不少權重運算，單個任務可下降 1 萬權重量，所有任務為 20 萬權重。

表 7. 實驗二關聯槽方法與實驗三自我關聯更新權重比較
[Table 7. A comparision between relation slot and self memory update]

Task	Relation slot	Self memory update
Task 1: Single Supporting Fact	80000	70000
Task 2: Two Supporting Facts	82900	72900
Task 3: Three Supporting Facts	83400	73400
Task 4: Two Argument Relations	79400	69400
Task 5: Three Argument Relations	85200	75200
Task 6: Yes/No Questions	83400	73400
Task 7: Counting	85100	75100
Task 8: Lists/Sets	85100	75100
Task 9: Simple Negation	81100	71100
Task 10: Indefinite Knowledge	81500	71500
Task 11: Basic Coreference	81600	71600
Task 12: Conjunction	80400	70400
Task 13: Compound Coreference	81600	71600
Task 14: Time Reasoning	81700	71700
Task 15: Basic Deduction	79700	69700
Task 16: Basic Induction	79500	69500
Task 17: Positional Reasoning	81100	71100
Task 18: Size Reasoning	80700	70700
Task 19: Path Finding	83200	73200
Task 20: Agent's Motivations	83600	73600
Sum of all task parameter	1640200	1440200
Mean parameter	82010	72010

4.4 實驗總結 (Experiment summary)

前面的三個實驗中主要探討兩個方向：準確率與權重計算量。從準確率方面來看，實驗一於輸出模組中做更動，透過重複性的注意力機制計算，嘗試提升複雜任務的推理能力。而實驗一的實驗結果平均錯誤率反而略為提升，推論為記憶保存的內容不足以支撐複雜的推理過程。

實驗二於動態記憶模組中做改善。透過記憶間的關聯，計算連結步動記憶間的關係。

從實驗數據中可看出實驗二的改善效果較為明顯，特別是任務 2 的準確率大幅提升。

實驗三效果下降最多。將關聯計算與本身保存的記憶同時更新於同個記憶槽，反而造成模型整體效果下降。推論將關聯計算更新記憶槽，會造成記憶保存的混亂。目前方法仍是需要分開保存關聯資訊與記憶本身，但也不代表記憶的關係自我更新不可行，而是需要詳細研究記憶槽與關聯槽的內容與特性，從而找出更好的記憶保存方法。

從權重計算量方面來看，實驗一權重使用量最少。因模型尚未加進記憶關聯的計算，且因模型重複使用相同注意力機制重複計算，權重用量與原模型差異不大。

而實驗二關聯提取與所有關聯計算比較，權重下降幅度最多，所有任務整體下降了 60 萬權重數量，較實驗一下降了 26.8%的權重量。實驗三雖準確率不高，但相較實驗二整體任務又下降了 20 萬權重。

5. 結論 (Conclusions)

本論文透過關係的計算使記憶內不同記憶槽具有關聯，如同人類記憶中不同實體並非單獨保存各自的訊息。例如實體的訊息或屬性可以聯想到其他實體或事件。關係的概念首先由 Google Deepmind 團隊於論文(Santoro *et al.*, 2017)中所提出，應用於圖像問答任務(Visual Question Answering, VQA)，計算兩兩物體間的關係。而記憶網路的概念目的在於透過不同的記憶保存與推理方式，提升模型的長期記憶能力，RelNet 模型首先將關係的概念帶入記憶網路中，提升模型的準確性，但其缺點也很明顯:大量的提高模型的權重與計算量。本論文提出關係提取的概念可以大量減少權重的計算量。

實驗中所採用的問答任務使用 20 個記憶槽，關係的縮減從 190 個關係提取到 20 個關係槽中，即使是小型任務也可以發現計算量大幅下降。而在越大型自然語言任務所運用的記憶槽數量相對越高，其中兩兩相對的關係計算也會大量增長。將關係提取出重要資訊於關係槽內，可以大量減少記憶儲存所需要占用的大小，以及輸出模組所需要推理的的計算量。

透過實驗二實驗結果可以看出整體準確率可以有效的提升，而這樣的方法不侷限於實驗所使用的 EntNet 模型，也可以運用於不同記憶網路架構內。而非記憶網路模型也可以用同樣的概念計算物體、數據、詞句以及時間上的關係。

以本論文為基礎，未來還可朝關係計算方法進行改善。從實驗二中可以看出不同任務提升的效果不同，其中任務 2 提升效果最為明顯。與任務本身的特性有關，透過更多關聯計算或許可提升其他任務的準確率。本研究的關係計算所採用的是兩兩記憶槽的計算，但現實世界不同的實體關係可以是兩個、三個或是群體間具有一定的關聯，例如籃球、羽毛球與足球三者皆屬於球類。bAbI 數據集中的任務 3 也需要需要更多先驗知識交互推理。未來若是關聯計算能帶入群組關聯計算，增強記憶間的連結性，應能再次提升模型記憶保存與推理能力。

本論文所改善的部分皆落於記憶保存與推理的部分，編碼的部分應能透過預訓練的方式改善。近幾年的語言模型研究多為預訓練(Pre-train)與微調(Fine-tuning)的方法，透過

大量文本資料來訓練自然語言詞句的關係。透過未標註的大量資料訓練，使編碼的向量可以更準確的表示詞句的意思。

本研究的實驗編碼方式都是與整體模型一起訓練，包含編碼、動態記憶模組以及推理模組的權重，數據量的不足較無法深入學習詞句意涵，而目前網路文本資料量大，未來可將模型的 Encoder 模組經過預訓練，提升編碼效果，或是針對編碼方式做改進，提高整體模型預測的效果。

參考文獻(References)

Bahdanau, D., Cho, K., & 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)*.

Bansal, T., Neelakantan, A., & McCallum, A. (2017). RelNet: End-to-End Modeling of Entities and Relations. In arXiv preprint arXiv:1706.07179.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., …Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734. doi: 10.3115/v1/D14-1179

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In arXiv preprint arXiv:1412.3555.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2017). Attention-over-Attention Neural Networks for Reading Comprehension. In *Proceedings of the 55th Annu. Meet. Assoc. Comput. Linguist*, *1*, 593-602. doi: 10.18653/v1/P17-1055

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In arXiv preprint arXiv:1810.04805.

Elman, J. L. (1990) Finding structure in time. *Cogn. Sci.*, *14*(2), 179-211. doi: 10.1016/0364-0213(90)90002-E

Henaff, M., Weston, J., Szlam, A., Bordes, A., & LeCun, Y. (2017). Tracking the World State with Recurrent Entity Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, *9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., …Socher, R. (2016). Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, *48*, 1378-1387.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In arXiv preprint arXiv:1909.11942.

Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., & Weston, J. (2016), Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1400-1409. doi: 10.18653/v1/D16-1147

Munkhdalai, T., & Yu, H. (2016). Reasoning with Memory Augmented Neural Networks for Language Comprehension. In arXiv preprint arXiv:161006454.

Palm, R. B., Paquet, U., & Winther, O. (2017). Recurrent Relational Networks. In arXiv preprint arXiv: 171108028 Cs.

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., … Lillicrap, T. (2017) A simple neural network module for relational reasoning. In arXiv preprint arXiv:170601427.

Seo, M. J., Min, S., Farhadi, A., & Hajishirzi, H. (2017). Query-Reduction Networks for Question Answering. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-End Memory Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, *2*, 2440-2448.

Trischler, A., Ye, Z., Yuan, X., Bachman, P., Sordoni, A., & Suleman, K. (2016). Natural Language Comprehension with the EpiReader. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 128-137. doi: 10.18653/v1/D16-1013

Weston, J., Bordes, A., Chopra, S., & Mikolov, T. (2016). Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *Proceedings of the ICLR2016*.

Weston, J., Chopra, S., & Bordes, A. (2014). Memory Networks. In arXiv preprint arXiv:14103916.

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 IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

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

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

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

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

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

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

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

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

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

Here shows an example.

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

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

Please visit the following websites for details.

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

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

No page charges are levied on authors or their institutions.

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

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

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

International Journal of
Computational Linguistics & Chinese Language Processing
vol.25 no.1 June 2020

Contents

Papers

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing