Computational Linguistics and Chinese Language Processing
Vol. 28, No.1, June 2023, pp. 35-52

台語—國語神經網路翻譯系統

Taiwanese-Mandarin Neural Machine Translation

周廷軒 *、林川傑*

Ting-Hsuan Chou and Chuan-Jie Lin

摘要

本論文提出一個台語—國語互譯的神經網路機器翻譯系統，以目前能找到的大型台語語料庫做為訓練資料，測試以字或詞為翻譯單位、自訓練或預訓練嵌入模型、各種未知詞處理策略，來找尋各種翻譯方向的最佳系統。最後建立出的國語翻台語系統效能過贏過目前已知的所有台語翻譯系統，在新聞類文章翻譯效能 BLEU 分數來到 75.02 分，散文類文章則是 38.13 分。此外，我們也第一次建立台語翻國語的系統，在上述兩種文本類別測試的 BLEU 分數分別是 73.38 及 35.32。

關鍵詞：台語、機器翻譯、神經網路、嵌入向量

Abstract

This paper proposes a Taiwanese-Mandarin neural machine translation system trained by all available Taiwanese corpora and translation datasets. The unit of translation can be either Chinese words or characters. Embedding will be either self-trained or pre-trained. And some strategies will be proposed to handle OOV output. The final Mandarin-to-Taiwanese outperforms all known Taiwanese MT systems. The best BLEU score evaluated on news articles is 75.02, and the best score on literature articles is 38.13. We also built the first Taiwanese-to-Mandarin NMT system in the world, which achieves BLEU scores of 73.38 and 35.32 on those two genres of articles.

Keywords: Taiwanese, Machine Translation, Neural Network, Embedding Model

* 國立臺灣海洋大學資訊工程學系
Department of Computer Sciene and Engineering, National Taiwan Ocean University
E-mail: {10957013, cjlin}@email.ntou.edu.tw

1. 緒論 (Introduction)

台灣最常使用的幾種語言，包括了台語及國語。關於國語、也就是中文 (或稱華語) 自然語言處理已有大量的研究成果，在台語方面的開發卻非常地少。近來神經網路的成功，讓我們有機會用來開發台語的機器翻譯系統。不過神經網路須倚賴非常大量的文本資料來做為訓練資料，這對於資源不夠豐富的台語來說是個問題，本論文也希望觀察用這種數量的資源能達成怎樣的翻譯效能。

本論文中 "台語" 這個詞指的是台灣地區所使用的閩南語，因為已經和大陸閩南地區使用的閩南語有些區別，因此常被稱為台灣閩南語，本論文為方便簡稱為 "台語"。

台語最常見的兩種書寫形式是漢字系統和拼音系統。以漢字書寫可加快閱讀速度，但實際上許多人在不清楚該寫成哪個漢字時會直接寫出拼音，這種交雜的寫法在本論文會稱為「台語漢羅」。也有學者支持全部以羅馬拼音來書寫，在本論文稱為「台語全羅」寫法。台語拼音系統常見有台羅拼音和白話字 (教會羅馬拼音) 兩種，本論文找到的資料集大多採白話字拼音。台語書寫範例請見第 2 節。

台語機器翻譯在 NLP 領域的相關研究實在不多，已知第一個國台語機器翻譯系統是由 Lin and Chen (1999) 所提出，翻譯方法僅靠查詢字典，但是加上了語音輸出結果。而且為了語音輸出，還處理了台語各種變調情形。

薛丞宏 (2014) 以程式自動整理、補齊了許多台語語料及國台語翻譯資料集，並提出一個統計式的機器翻譯系統。國語翻台語實驗最佳結果以詞為評估單位為 19 分、以字為評估單位為 32 分

黃志超 (2015) 則提出了一個以範例為本的國台機器翻譯系統，範例來源是蒐集自台語電視劇對白的一萬多個國台語對應句子。翻譯時在範例中找尋與國語輸入句最相似的句子，再修正對應的台語範例句、處理未知詞之後做為翻譯輸出。最佳結果正確率為 73.31%。

蔡育霖等人 (2018) 提出的是台語古詩朗讀系統，是以台語文讀音將古詩中的每個字唸出，須注意變調的位置。這個題目與機器翻譯較無關聯。

Li *et al.* (2022) 提出了建立利用語法及語意資訊來增進翻譯順暢度的修正規則的方法。該論文雖然提出修正翻譯的方法，評估時並不是直接評估翻譯效能，而是看錯字率 (WER)，無法基於相同測試資料及評估公式做比較。

其他台語翻譯研究大多和語音辨識 (Chen *et al.*, 2020) 及語音合成 (陳世翔, 2015; 張佑竹, 2017; 許文漢等人, 2020; Liao *et al.*, 2022) 有關，與本論文進行文字翻譯方向略有不同。本論文裡國語翻台語拼音的工作和台語語音合成中決定漢字或未知詞發音的工作相同，但未處理變調問題。

Meta 在 2022 年 10 月 19 日發布了一款閩南語與英文互相翻譯的 AI 語音翻譯系統 [1]

[1] https://huggingface.co/spaces/facebook/Hokkien_Translation

(Chen *et al*., 2022)，翻譯輸入、輸出都是語音形式，英文翻閩南語的 BLEU 分數 7.8 分，表示閩南語資源數量不足確實是個大問題。

本論文預計採用 transformer 架構來訓練國語—台語的翻譯系統，會使用多個台語語料庫來訓練台語語言模型，並以國台語翻譯資料集來訓練翻譯系統，以字典資源處理未知詞翻譯等等。希望可以製作出一個效果不錯的國台語相互翻譯系統。

2. 國台語語料庫與字典 (Corpora and Dictionaries)

本論文會用到的實驗資源，包括用來訓練翻譯系統的國、台語平行語料庫，用來訓練語言模型的台語語料庫，以及處理翻譯未知詞的國語及台語字典等等。以下介紹各種語言資源。

2.1 國、台語平行語料庫 (Taiwanese-Mandarin Parallel Corpora)

2.1.1 iCorpus臺華平行新聞語料庫 (iCorpus: Taiwanese-Chinese News Corpus)

iCorpus 臺華平行新聞語料庫 [2] 收集了 3,266 篇民視新聞國語文稿，再由專家翻譯成台語文句。台語以教會羅馬字拼音方式書寫，並以數字表示調號，範例請見圖 1。依換行符號切分對齊後共有 83,544 句國台語對應句。


```
{
    "台語": "Obama toa7-seng3 Bi2-kok thau5-chit8-ui7 ou-lang5 chong2-thong2\n chu3-
            bi2 tek8-phai3-oan5 Cho5-am7-phang hoa5-hu2 po3-to7…",
    "文號": 1,
    "日期": "2014-04-04",
    "華語": "Obama 大勝 美國 首位 黑人 總統\n駐美 特派員 曹郁芬 華府 報導…"
},…
```


圖 1. iCorpus 臺華平行新聞語料庫範例
[Figure 1. An Example from iCorpus: Taiwanese-Chinese News Corpus]

薛丞宏 (2014) 提出了台語斷詞及自動補齊全漢或全羅模式的整理方法，也用來產生 iCorpus 臺華平行新聞語料庫的漢羅版 [3] (台語改採用台羅拼音)。不過 2014 之後的新聞沒有處理到，因此有補上漢羅版本的只有 64,121 句。圖 2 為新聞語料庫的國語、全羅及漢羅資料對應範例。

[2] https://github.com/sih4sing5hong5/icorpus
[3] https://github.com/sih4sing5hong5/huan1-ik8_gian2-kiu3

```
{
    "國語": "Obama 大勝 美國 首位 黑人 總統",
    "全羅": "Obama toa7-seng3 Bi2-kok thau5-chit8-ui7 ou-lang5 chong2-thong2"
    "漢羅": "Obama 大勝 美國 頭一位 烏人 總統"
},
{
    "國語": "他 在 芝加哥 的 勝選 演說中 ，",
    "全羅": "i ti7 Chicago e5 seng3-soan2 ian2-soat-tiong ，"
    "漢羅": "伊 佇 Chicago 的 勝選 演說 中"
},
```


圖 2. 新聞語料庫國語、全羅及漢羅對應範例

[Figure 2. Sentences from iCorpus Written in Mandarin and Taiwanese (in Romanization or Chinese Characters)]

2.1.2 TGB通訊 (Tai-Bun Thong-Sin; Taiwanese News & Variety Magazine)

TGB 通訊 [4]是學生台灣語文促進會發行的刊物，發表的文章主要是以台語漢羅或是國語來書寫，許多文章甚至同時提供國、台語版本。台語文章中的羅馬拼音為教會羅馬拼音原始調號形式，沒有斷詞符號。

本論文使用的 TGB 語料採用薛丞宏 (2014) 整理過後的資料，整理工作包括將拼音改為台灣羅馬拼音、自動產生台語全羅的部分、將國語、漢羅、全羅加上斷詞邊界。薛丞宏總共收集了 1,015 篇雙語對應文章，共有 35,025 句，範例請見圖 3。不過我們有觀察到一些句子對齊錯誤、或是僅用單語書寫因此不該有對齊句的狀況，經人工檢查處理過後本論文實驗採用其中的 33,158 句，以下簡稱為 TGB 語料。

由於 TGB 通訊主要是刊錄以台語寫作的文章，用語會比較接近台語的自然語感，這和 iCorpus 語料庫以國語語感為主、用語及語序相近、國台語斷詞一致的特性完全相反，也常會有較大幅的改寫、漏詞、漏句等情形發生。

[4] https://taioanchouhap.pixnet.net/blog/category/583492

漢羅：

自有記 tì 以來, góa ē 記得細 hàn kan-ta^{n} 做過 1 pái 花童, 5 歲 ê 時, góa kap 表小妹良慈 tàu-tīn 做 3 叔 hām 3 嬸結婚時 ê 花童. Góa ē 記得 goán 2 ê hông chng-thā^{n} kah súi-tang-tang, 面 boah 粉, 2 pêng chhùi-phóe boah 紅粉, chhùi--nih chhat 紅 kì-kì ê ian-chi, tō án-ne 第 1 pái mā 是唯一 1 pái 做花童.

國語：

有記憶以來，我只記得小時候只當過一次花童，在五歲的時候，我和表妹良慈一起當三叔、三嬸結婚時的花童。我記得我們倆被打扮得很漂亮，臉上還塗了粉，兩頰也塗上腮紅，嘴唇擦上紅支支的口紅，就這樣第一次也就這麼唯一的一次當了花童。

圖 3. TGB 通訊文章範例
[Figure 3. An Article in TGB Corpus]

2.2 台語語料庫 (Taiwanese Corpora)

本論文翻譯系統採用的神經網路架構是 transformer，其 encoder 端跟 decoder 端的輸入皆有嵌入層，可以替換為來源語言及目標語言預訓練的語言模型。國語語言模型已由本實驗室以中文維基百科訓練而得，台語語言模型則需要再找到大型台語語料庫才能訓練。

以下是目前能找到的幾個大型台語語料庫，分別介紹它們的資料量、拼音系統及書寫形式等等。

(1) **教育部臺灣閩南語字詞頻調查工作** [5]：教育部在 2008 年委託學術單位進行的台語字詞頻調查統計計畫，總共蒐集了 1,093 個檔案。提供了漢羅及全羅兩種書寫形式，拼音系統是白話字，調號為數字版。

由於本論文希望基於完整文句來建立台語語言模型，因此只選擇了語料庫中分類為論文、小說、報導文學、以及散文這四類的文章，共有 314 篇、9,074 段文字。

(2) **台語文數位典藏資料庫** [6]：由「台灣白話字文學資料蒐集整理」計畫建立的資料庫，蒐集到一千餘本的白話字書刊，總共有 64,281 段文字。提供了漢羅及全羅兩種書寫形式，拼音系統是白話字，調號為正式版，但我們找到的 JSON 版本 [7]已經將調號轉為數字版。

(3) **台灣白話字文獻館** [8]：來自《台灣教會公報》白話字全羅拼音的文章，再由台灣白話字文獻館補上漢羅部分，總共有 37,984 段資料。原調號為正式版，我們以程式轉為數字版。

[5] https://github.com/Taiwanese-Corpus/Ungian_2009_KIPsupin

[6] https://db.nmtl.gov.tw/site3/dindex

[7] https://github.com/Taiwanese-Corpus/nmtl_2006_dadwt

[8] http://pojbh.lib.ntnu.edu.tw/script/index.php

2.3 字典與詞頻列表 (Dictionaries and Lexicons with Word Frequency)

翻譯工作的前、後處理有幾個步驟需要用到字典及詞頻的資訊，包括字詞對齊、未知詞翻譯處理等等。此處先介紹我們所使用的國台語字典或字詞頻列表，後面章節再介紹如何應用在翻譯工作中。

(1) **台華線上辭典** [9]：由師大華語研究所鄭良偉教授所建立，提供有台語詞漢羅寫法及白話字拼音，調號為數字版，及其對應國語詞，共有 62,046 個台語詞條。

(2) **臺灣閩南語常用詞辭典** [10]：由中華民國教育部主編，收錄了 20,605 筆台語詞條與 14,980 個例句，提供有台語詞漢羅寫法及台灣羅馬字拼音，調號為正式版，及其對應國語詞。由於其他語料庫及詞典的拼音系統都採用白話字，調號都是數字版，我們也以程式轉換了此辭典的拼音寫法。

    我們將以上這兩個詞典合併，並用臺灣閩南語常用詞辭典提供的斷詞版例句來統計詞頻，後面稱之為**國台雙語詞典**。

(3) **台語單字音典**：列出每個中文字所有可能的台語讀音，為本實驗室所整理的資料 (Lin & Chen, 1999; 蔡育霖等人, 2018)，資料來源是國台雙語辭典 (楊青矗, 1992) 及彙音寶典 (沈富進, 1954)。中文字各種讀音的頻率統計自台華線上辭典及臺灣閩南語常用詞辭典的台語漢羅拼音對應。

(4) **國語詞頻列表**：整理自中研院平衡語料庫 3.1 版及 4.0 版 [11]、中文維基百科 (詞條標題)[12] 這兩種資源，合併後共有 1,796,343 個國語詞條，再使用中研院平衡語料庫來統計國語詞的詞頻。

3. 翻譯訓練與未知詞處理 (Training and Out-of-Vocabulary Handling)

本論文是以一般常見的流程來訓練國台語機器翻譯系統，也就是準備好大量的國語及台語對譯句子、國語及台語各自的詞嵌入向量語言模型，用 transformer 神經網路 (Vaswani *et al*., 2017) 來訓練出機器翻譯系統，這時翻譯效能主要決定自訓練資料的數量與品質。

不過在研究過程中，發現在台語與國語之間翻譯的這個特別情況下，還有兩種選項可以選擇，而且和翻譯效能有蠻大的影響。一者是翻譯基本單位可選擇字或是詞，一者是訓練翻譯時不考慮低頻詞，但需額外處理輸出句裡的未知詞。

能有上述兩種選項的主要原因，是因為台語及國語同屬於漢語系，中文字本身就攜帶有語意，兩者大部份的詞彙都可以用漢字來書寫，所以有可能可用 “字” 做為翻譯基本單位。而且兩語言有不少詞彙使用了相同的中文字，因此有機會提出處理輸出句未知詞的方法。以下分別詳述。

---

[9] https://github.com/Taiwanese-Corpus/Tinn-liong-ui_2000_taihoa-dictionary

[10] https://twblg.dict.edu.tw/holodict_new/ ; https://github.com/g0v/moedict-webkit/

[11] https://ckip.iis.sinica.edu.tw/

[12] https://zh.wikipedia.org/

3.1 翻譯基本單位 (Basic Unit of Translation)

選擇以 "字" (character) 或是以 "詞" (word) 做為翻譯基本單位，表示不論是翻譯的訓練資料集、或是語言模型的訓練資料集都需要準備成相對應的基本單位分斷版本。

以詞為基本單位時，iCorpus 臺華平行新聞語料庫的國語句已經是斷好詞的版本。台語句以教會羅馬字拼音方式書寫時也是以空白做為分詞符號 (以連字號 – 連接詞中間各字的音節)，漢羅句也是斷好詞的版本，因此不用特別處理。TGB 通訊語料庫的國語句則是以中研院斷詞系統 (Li *et al*., 2020) 來斷詞。

所有台語語料庫，包括 TGB 通訊語料庫的台語句、教育部臺灣閩南語字詞頻調查工作語料庫、台語文數位典藏資料庫、台灣白話字文獻館，因為全羅句本身就帶有分詞符號，他們的漢羅句就能以對齊漢字與拼音 (查詢單字音典) 的方式來決定漢羅句的斷詞版本。例如：

漢羅原始：羊仔母欲去樹林挽果子

全羅原始：iunn5-a2-bu2 beh khi3 chhiu7-na5 ban2 koe2-chi2

漢羅斷詞：羊仔母 欲 去 樹林 挽 果子

以字為基本單位時，所有國語句裡的中文字都會斷開，數字或是英文詞則保留原詞字串。所有台語全羅句則是依音節 (syllable) 斷開 (也就是將連字號都改為空格)。至於台語漢羅句，則是會斷成中文字或拼音音節。以下是所有斷詞及斷字後的例子。

國語斷詞：2000 年 她 開始 以 英文 寫 回憶錄

國語斷字：2000 年 她 開 始 以 英 文 寫 回 憶 錄

漢羅斷詞：2000 年 伊 開始 用 英文 寫 回憶錄

漢羅斷字：2000 年 伊 開 始 用 英 文 寫 回 憶 錄

全羅斷詞：2000 ni5 i khai-si2 iong7 eng-bun5 sia2 hoe5-ek-liok8

全羅斷字：2000 ni5 i khai si2 iong7 eng bun5 sia2 hoe5 ek liok8

在處理過程中，我們常碰到漢羅句與全羅句內容不一致、中文字與拼音無法完全對齊的情形，此時我們只好先處理能對齊的部份，而無法對齊的部份查字典做長詞優先斷詞。未來可能需要花工夫好好整理台語語料庫內容。

在建構神經網路翻譯系統時，使用的預訓練嵌入向量資料必須和翻譯實驗資料的翻譯單位一致才行。也就是說，進行以字為翻譯單位的實驗時，需採用訓練自斷字版語料庫的字嵌入向量語言模型；進行以詞為翻譯單位的實驗時，需採用訓練自斷詞版語料庫的詞嵌入向量語言模型。

3.2 輸出句未知詞之處理 (OOV Handling)

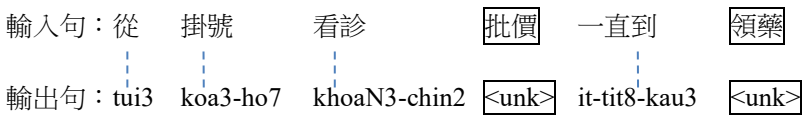
本論文採用的 transformer 程式 [13] 中有個參數 min_freq 可以設定詞頻門檻值。當一個詞在訓練資料中出現次數低於詞頻門檻值時，該詞會被改為 <unk> 來進行後續的機器翻譯訓練，因而翻譯輸出句有時會出現 <unk> 這個字串。因為國語與台語皆屬於漢語群，其詞彙與使用漢字相近，因此我們額外提出處理未知詞輸出的翻譯方法。

3.2.1 尋找未知詞對應之來源詞 (Aligning an Unknown Word to its Source Word)

首先第一步是尋找未知詞在輸入句中對應的來源詞為何，尋找方法是採用最長子序列演算法 (LCS) 來對齊輸入句及輸出句的詞。但由於兩個句子是不同語言 (甚至可能一者是漢字、一者是拼音)，比對時必須考慮翻譯的對應。我們將國台語詞彙間配對分數定義如下：

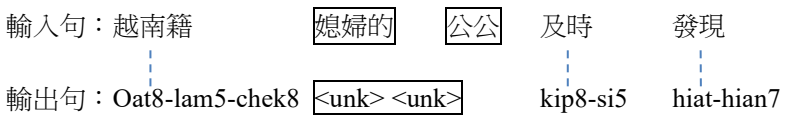
- 完全相同的字串 (中文或非中文字串) 配對分數為 1 分
- 國台雙語詞典所收錄的翻譯配對分數為 1 分
- 國語詞與台語漢羅詞之間的翻譯配對分數是以相同的漢字字數來計算 Dice 分數。例如國語詞 "刀子" 與台語詞 "刀仔" 共同字串是 "刀"，因此兩者間的 Dice 分數是 2×1 / (2+2) = 0.5
- 國語詞與台語全羅詞之間的翻譯配對分數是比對漢字與其台語發音來計算 Dice 相似分數。例如國語詞 "一直到" 與台語詞 "it-tit8-kau3"，漢字 "一" 可讀為 "it"、"直" 可讀為 "tit8"、"到" 可讀為 "kau3"，因此共同子字串長度為 3，Dice 分數是 2×3 / (3+3) = 1

找尋對齊之前會先將連續的 <unk> 合併。產生共同子序列對齊後，若 <unk> 在輸入句相對應位置有未被對齊的詞，就視為此未知詞對應之來源詞。例如：



其中虛線表示已對齊的配對，因此句中第一個 <unk> 對應的國語詞是 "批價"，第二個 <unk> 對應的國語詞是 "領藥"。

若是 <unk> 在輸入句的對應位置有多個未對齊的詞，則這些詞都會產生翻譯，並採用輸入句的斷詞方式。例如：



[13] 程式改寫自 bentrevett 的 GitHub 專案 https://github.com/bentrevett/pytorch-seq2seq

上例中輸出句兩個 <unk> 相鄰出現故而合併，在輸入句的對應位置有 "媳婦的" 和 "公公" 兩個國語詞，因此會在該位置產生兩個輸出翻譯詞。

請注意以上演算法都只處理 <unk> 的翻譯做法。若是輸出句有未被對齊的詞，會保留在輸出句原位置不會更改。

3.2.2 處理國語翻台語未知詞 (OOV Translation from Mandarin to Taiwanese)

接著說明依照對應的國語詞產生台語未知詞翻譯的方法，其中國語翻漢羅及國語翻全羅步驟大致相同，只要選擇需要的台語寫法當作輸出即可。

首先查詢國台雙語詞典，若有收錄對應國語詞，則以詞典中提供的台語 (漢羅或全羅寫法) 做為未知詞的翻譯。若該國語詞有多種台語說法，則選擇詞頻較高的台語詞 (參見第 2.3 節)。像是上例中"公公"這個國語詞在詞典裡有"ta-koaN"、"lau7-a-peh"、"an2-kong"等等多種台語翻譯，其中以"ta-koaN"的詞頻最高，所以選為此未知詞的翻譯輸出。

如果國台雙語詞典查不到，會將國語詞以中研院斷詞系統進行斷詞，再重查國台雙語詞典。在我們拿到的國台翻譯實驗資料中，已斷好詞的國語句常常和中研院的斷詞標準不太一致，像是上例中 "媳婦的" 這個詞依中研院斷詞標準應再被斷成"媳婦"、"的"，查詞典可得到台語翻譯"sin-pu7"、"e5"，再串接成"sin-pu7-e5"做為未知詞的翻譯結果。

如果斷出來的詞也未收錄在詞典中，就會直接照字面翻譯。國語翻台語漢羅就直接採用國語的漢字當作輸出。國語翻台語全羅會改查詢台語單字音典，選擇最常唸的讀音做為輸出，如此就能將"拱範宮"這種專有名詞的讀音"kiong2-hoan7-keng"串接出來當做未知詞的翻譯結果。

3.2.3 處理台語翻國語未知詞 (OOV Translation from Taiwanese to Mandarin)

接著說明對應的台語詞產生國語未知詞翻譯的方法，其中漢羅翻國語及全羅翻國語步驟大致相同，只需在查詢詞典時使用不同的台語寫法欄位。

首先查詢國台雙語詞典，若有收錄對應台語詞，則以詞典中提供的國語詞做為未知詞的翻譯，有多種選擇時則選詞頻較高的國語詞。

如果一個台語全羅詞未收錄在詞典中，它有一定的可能性是個專有名詞或是較文言、較艱深的中文詞，字典少有收錄，會依漢字直接照字唸。為此我們多增加了一個流程來判斷台語全羅詞是否為已知國語詞的讀音。

首先將台語全羅詞利用連字號斷出各個音節，利用單字音典查詢每個台語音節對應的漢字，窮舉出所有漢字組合並查詢第 2.3 節介紹的國語詞頻列表，選擇詞頻較高者做為翻譯輸出。以"Hai2-seng-koan2"為例，第一個音節"hai2"對應的漢字只有"海"，第二音節"seng"對應的漢字有"升"、"生"、"先"…，第三音節"koan2"對應的漢字有"管"、"館"…，進行組合後有"海升管"、"海生管"、"海生館"…等組合，其中"海生館"出現在國語詞頻列表中，選為翻譯輸出。

其餘未收錄在詞典中的台語詞 (漢羅或全羅)，會以國台雙語詞典進行長詞優先斷詞，並選擇詞頻較高的對應國語詞做為輸出。若是斷出單字不在詞典中，就會直接照字翻。漢羅的漢字部分就直接採用，漢羅和全羅的羅馬拼音部份則會查詢台語單字音典，選擇最常對應的國語字做為輸出。

以"真好食"為例，依國台雙語詞典、長詞優先會斷成"真好"、"食"，其常見對應國語詞是"很好"、"吃"，再串接成"很好吃"做為未知詞的翻譯結果。再以"chhiong3-pang3-liau2"為例，會斷成"chhiong3"、"pang3"、"liau2"。其中"chhiong3"在國台雙語詞典查不到，改查台語單字音典、選擇最常對應的國語字"縱"。而"pang3"、"liau2"在國台雙語詞典裡最常對應的國語詞是"放"、"了"，三者串接成"縱放了"做為未知詞的翻譯結果。

3.2.4 以字翻字系統處理詞翻詞系統輸出句之未知詞 (Using Character-Based MT System to Translate OOV in Word-Based Translation)

在後續的實驗會發現，以字為翻譯單位所訓練出來的翻譯系統在逐字翻譯的效能上會贏過以詞為單位的翻譯系統。因此我們採用一種新方法，就是以字翻字系統去處理詞翻詞系統輸出句的未知詞。

每次找到未知詞對應來源詞的時候，直接以字翻字系統翻譯這個對應來源詞，翻譯結果拼接成一個詞再填入未知詞出現的位置。若有多個對應來源詞，會各自取得翻譯結果。

在後續的實驗結果可看到字翻字系統確實能成功地處理不少未知詞的翻譯，但分析常翻錯的情形發現大多是能在字典找到的翻譯詞條。因此我們又提出一個新的混合式策略：如果字典查得到對應來源詞，就以字典提供的翻譯最高頻者做為輸出，否則就以字翻字系統來產生翻譯輸出。請注意，全羅翻國語的查字典步驟還會多考慮一種情形：窮舉出所有拼音可對應的漢字組合並查詢國語詞頻列表。

3.2.5 以詞為單位來評估字翻字系統輸出的方法 (Approximated Evaluation of Character-Based MT System in the Unit of Words)

因為字翻字系統的輸出結果沒有斷詞邊界，無法以詞為單位進行評估。為了能夠與已知的翻譯系統效能進行比較，我們提出兩種斷詞的方法，分別是利用來源句或是正解來進行斷詞。這些方法只是暫行的策略，用來觀察效能比較。未來可以直接使用更精準的台語斷詞系統來斷詞。

這裡先說明以來源句決定斷詞邊界的方法。簡單來說就是以斷詞版來源句裡所有詞、各詞在字典中的翻譯、漢字詞能窮舉出的所有拼音組合來對輸出句進行最多對齊的斷詞。若輸出句中還有未決定斷詞邊界的字串區間，就以國台雙語詞典做長詞優先斷詞。

以正解句來幫字翻字系統的輸出結果進行斷詞的做法是：將出現在正解句裡的所有來源詞視為一個小型的詞典，對字翻字系統的輸出句做最多對齊的斷詞。由於實際上拿不到正解句資訊，這評估分數主要為 upper bound。

4. 實驗結果與討論 (Experiments and Discussion)

4.1 實驗資料與評估方法 (Experimental Data and Evaluation Metrics)

本論文國語－台語機器翻譯系統採用 transformer 模型，翻譯方向有國語翻台語及台語翻國語，台語又分為漢羅寫法以及全羅寫法。翻譯單位可選擇以詞或字為單位，嵌入向量採用自訓練或預訓練嵌入向量語言模型。

翻譯實驗使用第 2.1 節介紹的兩個資料集，分別為 iCorpus 臺華平行新聞語料庫 (以下簡稱 iCorpus 語料) 以及 TGB 通訊語料庫 (以下簡稱 TGB 語料)。iCorpus 語料中，國語與台語全羅對譯的句子有 83,544 句，國語與台語漢羅對譯的句子有 64,121 句。TGB 語料國語與台語對譯的句子有 33,158 句。因為 iCorpus 語料與 TGB 語料本質上很不相同 (參見第 2.1 節)，兩套資料集會各自進行實驗。

本論文的國語預訓練詞嵌入向量是利用整部中文維基百科當作訓練語料，約三千多萬句國語句子，已先用本實驗室開發的斷詞系統產生斷詞版本。訓練工具為 fastText (Bojanowski *et al*., 2017)，採 skip-gram 的方式，維度設為 250。

台語的預訓練詞嵌入向量是利用台語文數位典藏資料庫、白話字文獻館資料以及教育部臺灣閩南語字詞頻調查工作語料庫所有文章當作訓練語料，總共約十一萬多個台語句子，斷詞做法如第 3.1 節介紹。也是使用 fastText、skip-gram 的方式訓練，維度也設為 250。請注意，漢羅寫法與全羅寫法各會訓練一套嵌入向量語言模型。

評估方法採用十等份交叉驗證法 (10-fold cross-validation)。翻譯評估公式為 BLEU，並採用 Smoothing 5 計算方法來避免短句翻譯不易有長詞組重疊的情形。

以詞為評估單位時 Smoothing 5 的 BLEU 分數會計算 unigram 到 5-gram。如果以同一個程式來計算以字為評估單位的 BLEU 分數，等於最長只會看到連續 5 個字，與 5 個詞所涵蓋的長度不太相同，不太公平。經統計，國、台語詞平均長度是 1.76 字，5 個詞大約涵蓋 9 個字。因此以字為評估單位的 BLEU 分數會改計算到 9-gram。

4.2 基本版翻譯實驗 (Baseline MT Experiments)

首先以 transformer 來建立基本版國台翻譯系統，做為後續各系統的比較基準。訓練時國台語皆以詞為翻譯基本單位，嵌入層測試了自訓練模型(初始值設為亂數) 與預訓練模型 (參見第 4.1 節) 兩種。以詞為單位的 BLEU 評估結果列在表 1，其中「國語→全羅」表示由國語翻譯至台語全羅，以下類推。

表 1. 各種詞嵌入模型翻譯實驗結果 (以詞為評估單位之 BLEU 分數)
[Table 1. Results of NMT with Different Corpora and Embedding Models (Word-Based, BLEU Scores)]

語料庫	iCorpus		TGB	
詞嵌入模型	自訓練	預訓練	自訓練	預訓練
國語→全羅	58.68	68.08	31.86	35.72
國語→漢羅	49.88	61.34	31.24	34.96
全羅→國語	51.90	59.11	30.72	33.01
漢羅→國語	49.59	59.64	31.01	32.81

由表 1 可以看到，加入預訓練詞嵌入向量模型的系統，不論在何種語料庫、何種翻譯方向或台語書寫形式，翻譯效果都比自訓練模型好。不過在 TGB 語料上的提升沒有像在 iCorpus 語料那般明顯。後續實驗若沒特別標明，都是採用預訓練嵌入模型。

此外，這實驗僅使用最一般的翻譯系統訓練流程，翻譯效能 BLEU 分數就已經可以達到 60 左右 (TGB 則是 34 左右)，分數都比先前各台語機器翻譯研究還高，表示神經網路確實可以訓練出效果較好的台語翻譯系統。

4.3 翻譯單位實驗 (Translation Unit Experiments)

接著測試翻譯單位對翻譯結果的影響。以詞為單位的翻譯系統與第 4.2 節相同。以字為單位訓練時國台語翻譯系統時，預訓練嵌入模型是以斷字版的台語語料庫 (請見第 2.2 節) 來訓練，翻譯訓練資料也是採用斷字版。結果列在表 2，因為輸出結果也是斷字版，評估時改以字為評估單位。

由表 2 可以看到，改以字為翻譯單位後，各系統不論在 iCorpus 或 TGB 語料庫的測試結果都上升不少，可能是因為照字翻比較能克服專有名詞這類未知詞的翻譯問題。

表 2. 翻譯單位實驗結果
[Table 2. Results of Translation Unit Experiments]

語料庫	iCorpus		TGB	
翻譯單位	詞	字	詞	字
國語→全羅	81.08	88.62	49.76	57.75
國語→漢羅	74.72	88.51	48.39	58.26
全羅→國語	69.36	86.79	42.55	50.34
漢羅→國語	74.82	89.98	42.76	54.81

4.4 未知詞處理實驗 (Out-of-Vocabulary Handling Experiments)

在翻譯訓練過程中若設定詞頻 1 次以下者視為未知詞，系統輸出就會出現 <unk> 未知詞標籤。第 3.2 節介紹了各種未知詞的處理策略，實驗結果呈現在表 3 及表 4。其中各種未知詞處理策略包括：

- 無：詞頻門檻值設為 1，不會有未知詞輸出
- 未：詞頻門檻值設為 2，不處理輸出句裡的未知詞
- 規：規則式策略，以第 3.2.2、3.2.3 節所提方法處理未知詞
- 翻：以字翻字系統處理詞翻詞系統輸出之未知詞 (第 3.2.4 節)
- 混：混合式策略，先查詢字典，再採用字翻字系統翻譯結果

表 3 是在 iCorpus 語料庫的評估結果。可以看到不論是以詞或是以字為評估單位，採用混合式未知詞處理策略的詞翻詞系統能達到相當高分的效能，不論何種翻譯方向及台語書寫形式都是全論文最高分 (數字以粗體呈現)。唯一例外是台語漢羅↔國語互譯時，以字為評估單位會是字翻字系統效能最好。

表 4 是在 TGB 語料庫的評估結果。可以看到以詞為評估單位時，不論是何種翻譯方向及台語書寫形式，採用混合式未知詞處理策略的詞翻詞系統會是全論文最佳系統，以字為評估單位時則是以規則式未知詞處理策略的字翻字系統效能最好。

表 3. 未知詞處理實驗結果 (iCorpus 語料庫)
[Table 3. OOV Handling Experiments with iCorpus]

評估單位	詞					字				
未知詞處理	無	未	規	翻	混	無	未	規	翻	混
國語詞→全羅詞	68.08	61.30	68.82	74.81	**75.02**	81.08	74.11	85.58	90.56	**90.73**
國語詞→漢羅詞	61.34	54.27	64.26	73.70	**74.54**	74.72	67.30	84.56	90.51	90.57
全羅詞→國語詞	59.11	55.00	71.90	72.52	**73.38**	69.36	64.25	86.47	85.68	**87.99**
漢羅詞→國語詞	59.64	54.43	68.40	70.56	**71.51**	74.82	68.12	87.54	89.07	89.86
國語字→全羅字	*65.11*	*55.13*	*66.54*	--	--	88.62	89.48	89.59	--	--
國語字→漢羅字	*61.59*	*55.67*	*64.53*	--	--	88.51	90.45	**90.71**	--	--
全羅字→國語字	*59.22*	*52.68*	*65.95*	--	--	86.79	87.49	87.62	--	--
漢羅字→國語字	*60.01*	*55.65*	*68.65*	--	--	89.98	90.35	**90.62**	--	--

表 4. 未知詞處理實驗結果 (TGB 語料庫)
[Table 4. OOV Handling Experiments with TGB Corpus]

評估單位	詞					字				
未知詞處理	無	未	規	翻	混	無	未	規	翻	混
國語詞→全羅詞	35.72	36.82	37.14	37.48	**37.56**	49.76	50.78	52.77	52.72	52.79
國語詞→漢羅詞	34.96	37.37	37.82	37.29	**38.13**	48.39	51.04	53.77	52.95	53.83
全羅詞→國語詞	33.01	33.14	34.91	34.81	**35.32**	42.55	42.79	48.12	46.02	48.58
漢羅詞→國語詞	32.81	34.18	35.11	35.09	**35.30**	42.76	44.00	48.35	47.59	48.84
國語字→全羅字	*29.24*	*29.25*	*29.44*	--	--	57.75	57.77	**57.87**	--	--
國語字→漢羅字	*29.09*	*29.14*	*29.84*	--	--	58.26	58.30	**58.48**	--	--
全羅字→國語字	*27.53*	*27.86*	*27.88*	--	--	50.34	50.36	**51.21**	--	--
漢羅字→國語字	*33.18*	*33.51*	*34.57*	--	--	54.81	55.16	**55.40**	--	--

第 3.2.5 節介紹過如何對字翻字系統的輸出句斷句，方便進行以詞為單位的評估，數字列在表 3 及表 4，以斜體呈現。可看到雖然字翻字系統常在以字為評估單位時獲得較好的分數，斷詞後分數卻輸給詞翻詞的系統。其中一大主因是台語斷詞錯誤，另一個則是實驗資料的台語斷詞規則與詞典不一致所致。

為了觀察在最佳斷詞狀態、字翻字系統能達到怎樣的效能，第 3.2.5 節也討論了用正解來斷詞的做法，用以觀察所能達成效能上限。數據列在表 5 中，其中“詞_混合”表示詞翻詞系統、以混合式策略處理未知詞，“字_規則”表示字翻字系統、以規則式策略處理未知詞、並以來源句出現的詞進行斷詞，“字_正解”表示字翻字系統、以規則式策略處理未知詞、並以正解句的詞彙進行斷詞。

表 5. 最佳系統比較表 (以詞為評估單位)
[Table 5. Performance of the Best NMT Systems (Evaluating by Words)]

語料庫	iCorpus			TGB		
翻譯系統	詞_混合	字_規則	字_正解	詞_混合	字_規則	字_正解
國語→全羅	75.02	*66.54*	*70.34*	37.56	*29.44*	*45.57*
國語→漢羅	74.54	*64.53*	*71.21*	38.13	*29.84*	*45.76*
全羅→國語	73.38	*65.95*	*68.93*	35.32	*27.88*	*40.84*
漢羅→國語	71.51	*68.65*	*74.21*	35.30	*34.57*	*45.21*

以正解句斷詞對兩個翻譯實驗資料集的效果完全不同。在 iCorpus 語料即便是用正解斷詞，仍無法贏過目前最佳系統，只有台語漢羅翻國語的方向有機會再提升 2.7 分。然而 TGB 語料這邊以正解斷詞時，所有翻譯方向的 BLEU 分數可再上升 5~10 分，表示

未來若有正確率高的台語斷詞系統可使用，最佳系統會變成是字翻字、以規則式策略處理未知詞的翻譯系統。

5. 結論 (Conclusion)

本論文提出一個台語—國語互譯的神經網路機器翻譯系統，神經網路採用 transformer 架構，測試了以字或詞為翻譯單位、自訓練或預訓練嵌入模型、各種未知詞處理策略，來找尋各種翻譯方向的最佳系統。翻譯訓練資料採用 iCorpus 臺華平行新聞語料庫以及 TGB 通訊語料庫。論文主要貢獻：

(1) 以 transformer 架設出效能最佳的台語—國語神經網路機器翻譯系統

(2) 若測試資料集是以台語書寫的文章再加上國語對應文字，以「字」做為翻譯單位會比以「詞」為翻譯單位獲得更好的翻譯效能。

在嵌入向量模型的設定上，不論何種翻譯單位，預訓練嵌入模型效果幾乎都比自訓練嵌入模型還要好。在翻譯單位實驗中，以詞為評估單位時，詞翻詞系統效果比較好。以字為評估單位時，字翻字系統效果比較好。以上實驗不論是何種翻譯方向、何種台語書寫形式、何種翻譯單位、使用哪個實驗資料集皆如此。

未知詞處理方法有查字典的規則式策略、以字翻字系統做翻譯、以及兩者的混合式策略。測試後發現，以詞為評估單位時皆是詞翻詞系統、混合式策略處理未知詞的效果最好，以字為評估單位時則大多是字翻字系統、規則式策略處理未知詞效果較好。

整體來說，iCorpus 語料庫因為國、台語的文句較相似，句中新聞用語也比較多，因此翻譯分數較高。以詞為評估單位時，BLEU 分數在 71.51~75.02 之間。以字為評估單位時，BLEU 分數在 87.99~90.73 之間。表示我們的系統已能翻譯新聞這類較嚴謹的文章。

而 TGB 語料庫因為台語文句較自然，國、台語文句不一定完全對齊，因此翻譯分數較低。以詞為評估單位時，BLEU 分數在 35.30~38.13 之間。以字為評估單位時，BLEU 分數在 51.21~58.48 之間，都贏過現在已知的台語翻譯系統。

未來如果有品質很好的台語斷詞系統，以詞評估的 BLEU 分數有機會上升至 40.84~45.76 之間。因此未來的重要工作之一是開發台語斷詞系統。

此外，研究過程中也發現，國、台語文句偶有斷詞不一致的狀況，也會有漢字或拼音錯誤的情形 (許多來自薛丞宏的自動整理結果)。未來若能修正這些錯誤，相信能再提高翻譯效能。

參考文獻 (References)

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135 - 146.

Chen, P.-Y., Wu, C.-H., Lee, H.-S., Tsao, S.-K., Ko, M.-T., & Wang, H.-M. (2020). Using Taigi Dramas with Mandarin Chinese Subtitles to Improve Taigi Speech Recognition. In *Proceedings of the 23rd Conference of the Oriental COCOSDA*, 71 - 76.

Chen, P.-J., Tran, K., Yang, Y., Du, J., Kao, J., Chung, Y.-A., Tomasello, P., Duquenne, P.-A., Schwenk, H., Gong, H., Inaguma, H., Popuri, S., Wang, C., Pino, J., Hsu, W.-N., & Lee, A. (2022). Speech-to-Speech Translation for a Real-World Unwritten Language. https://arxiv.org/abs/2211.06474

Li, P.-H., Fu, T.-J., & Ma, W.-Y. (2020). Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 34(05), 8236-8244.

Li, Y.-H., Young, C.-P., & Lu, W.-H. (2022). Using Grammatical and Semantic Correction Model to Improve Chinese-to-Taiwanese Machine Translation Fluency. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, 75 - 83.

Liao, Y.-F., Hsu, W.-H., Pan, C.-M., Wang, W.-J., Pleva, M., & Hládek, D. (2022). Personalized Taiwanese Speech Synthesis using Cascaded ASR and TTS Framework. In *Proceedings of the 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*, 1 - 5.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 6000 - 6010.

張佑竹 (2017)。以知識表徵方法建構台語聲調群剖析器。**中文計算語言學期刊**，**22**(2)，73-86。[Chang, Y.-C. (2017). A Knowledge Representation Method to Implement A Taiwanese Tone Group Parser. *International Journal of Computational Linguistics and Chinese Language Processing*, *22*(2), 73 - 86.]

蔡育霖、黃兆湘、 林川傑 (2018)。台語古詩朗誦系統。**第三十屆自然語言與語音處理研討會**, 184 - 198.[Tsai, Y.-L., Huang, C.-H., & Lin, C.-J. (2018). A Taiwanese Text-to-Speech System for Ancient Poems. In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018)*, 184 - 198.]

薛丞宏 (2014)。**漢語間統計式機器翻譯語料處理－用臺灣閩南語示範**。國立交通大學碩士論文。[Sih, S.-H. (2014). *Corpus Preprocessing for Statistical Machine Translation between the Chinese Languages-Using Taiwan Southern Min as Examples*. Master thesis. National Chiao Tung University.]

許文漢、曾證融、廖元甫、王文俊、潘振銘 (2020)。基於深度學習之中文文字轉台語語音合成系統初步探討。**中文計算語言學期刊**，**25**(2), 69-84. [Hsu, W.-H., Tseng, C.-J., Liao, Y.-F., Wang, W.-J., & Pan, C.-M. (2020). A Preliminary Study on Deep Learning-based Chinese Text to Taiwanese Speech Synthesis System. *International Journal of Computational Linguistics & Chinese Language Processing*, *25*(2), 69 - 84.]

陳世翔 (2015)。**華台語文轉音系統中未知詞發音決策**。國立中興大學碩士論文。[Chen, S.-H.. (2015). *Decision for Pronunciation of Out-of-Vocabularies in a Mandarin to Taiwanese Text-to-Speech System*. Master thesis. National Chung Hsing University.]

黃志超 (2015)。**範例為本的國語—台語翻譯之研究**。國立臺灣海洋大學碩士論文。[Huang, C.-C. (2015). *A Study on Example-Based Mandarin-Taiwanese Machine Translation*. Master thesis. National Taiwan Ocean University.]