國立清華大學 電機工程學系
實作專題研究成果摘要

Automatic Spoken Language Translation and Captioning for NTHU Course Videos

《清華大學網路課程自動語音翻譯與字幕標記》

專題領域：系統組
組　　別：A18
指導教授：劉奕汶
組員姓名：丁文淵、柳奕丞
研究期間：2019 年 7 月 1 日至 2020 年 6 月底止，計 11 個月

Abstract

因為 NLP、人工智慧研究的興起，帶動學者透過研究字詞間與上下文的關係[1]，以及設計新的模型進行翻譯[2]，希望可以改善現今翻譯、解讀的品質。此外，與我們學生最為相關的其中一項就是教學部份，若是可以透過設計之機器翻譯模型，結合網路上的語音辨識軟體進行翻譯與字幕標記，將會有助於外籍學生於本校的學習，也會有助於學校發布開放式課程的進行。

Introduction

本專題主要以 Python 建構「影音翻譯系統」，包含：影音下載、轉檔、語句切割、語音辨識，並將辨識之中文內容輸入至我們以 Keras、Tensorflow 建造之機器翻譯模型轉為英文內容，最後一樣透過此系統產生字幕檔 (.srt)。字幕檔可以透過 MKVToolNix 等軟體將最後翻譯之英文字幕添加回原影片。

針對以上設計流程，將會以以下數點描述設計個部分之細節：

1. 影音處理 – 語句處理與強度偵測

針對語音內容，我們使用 pytube, pydub 等 packets 進行下載、轉檔，並且透過 parselmouth 這個 packet 產生音訊時間與頻率、強度之間的關係圖，如圖 1 所示，方便我們判斷合適之音訊強度進行語句分割。

因為音訊的產生有很多可能性，有可能是背景雜音、突發性雜音，或是最主要希望擷取的內容—主講者音訊，因此我們針對音訊強度和圖 1 的關係圖設計「語音強度偵測演算法」，藉此辨識一句話的起點與終點時刻，輸入網路語音辨識系統—Google Recognizer 取得中文語音內容。

2. 自然語言處理 – 深度學習模型建立與翻譯

取得中文語音內容後，我們將之輸入模型進行訓練。此模型為 Seq2Seq model，由中文語句轉到英文語句的模型。模型的架構如圖 2 所示。然而，因為中文語句和英文語句並非同樣以一個單字為單位，因此輸入中文內容至模型前，我們需要進行「資料

預處理」，我們透過 jieba、[1]Keras 的 Tokenizer、Model、GRU[4]、Embedding 改良並完成之。

接著，我們將處理過的內容送進模型。最上層為 Embedding 層，我們藉此產生中文詞向量。至於 Decoder 端的[2]Embedding 層，我們採用 Stanford University 研發之 GloVe 英文詞向量[5]。Embedding 層以下則是 3 層 GRU，如圖 3 所示，Encoder 和 Decoder 都各有 3 層，分別負責不同任務。Encoder 負責將整句話的摘要濃縮在一個向量當中，並作為 Decoder 的 GRU 初始狀態（initial state）。

至於 Decoder 的部分，透過上層 Embedding 層取得英文詞向量，結合 Encoder 送入之初始狀態輸出我們期望的英文翻譯內容。

3. 自然語言處理–訓練資料搜尋與爬蟲程式實作

因為我們希望翻譯的內容為專業科目，需要用到大量學科專有名詞，因此我們透過「國家教育研究院」的「雙語詞彙、學術名詞暨辭書資訊網[3]」，以及許多字典網站，如「查查線上辭典[4]」、「聽力課堂[5]」，設計爬蟲程式取得中英對照句作為平行資料(Parallel Data)，希望可以藉由這些資料讓最後翻譯的內容更加準確。

4. 影音處理 – 自動產生字幕檔

標準字幕檔採用「.srt」格式，其中必須包含語句編號、起始時間與終點時間、語句內容。三者需要依照一定格式呈現，否則之後無法作為字幕檔添加字幕回原影片。因此，我們一樣設計演算法整理翻譯後的內容、時間點，以符合字幕檔要求之格式。

---

[1] https://github.com/Hvass-Labs/TensorFlow-Tutorials/blob/master/21_Machine_Translation.ipynb

[2] https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/?fbclid=IwAR14AeDS3aSqr7y2hnrlWAIP6SJnhK1Tn2FI9itdv40TQnQYVev2Oho9Scs

[3] http://terms.naer.edu.tw/

[4] https://tw.ichacha.net/

[5] http://fy.tingclass.net/

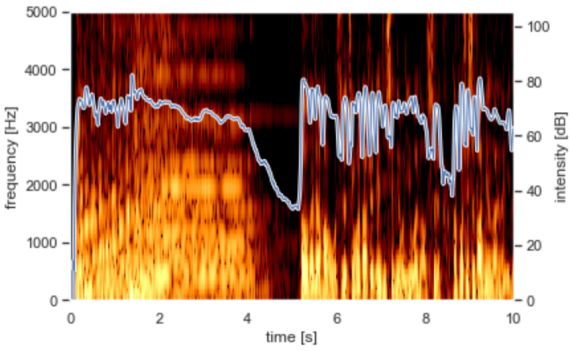

圖 1 音訊強度、時間、頻率關係圖

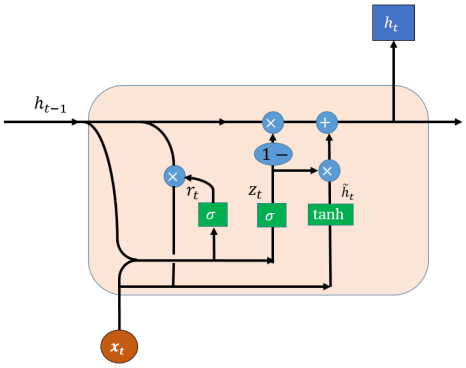

圖 3 使用之 GRU 架構圖
(圖片來源：[6] Fig.1)

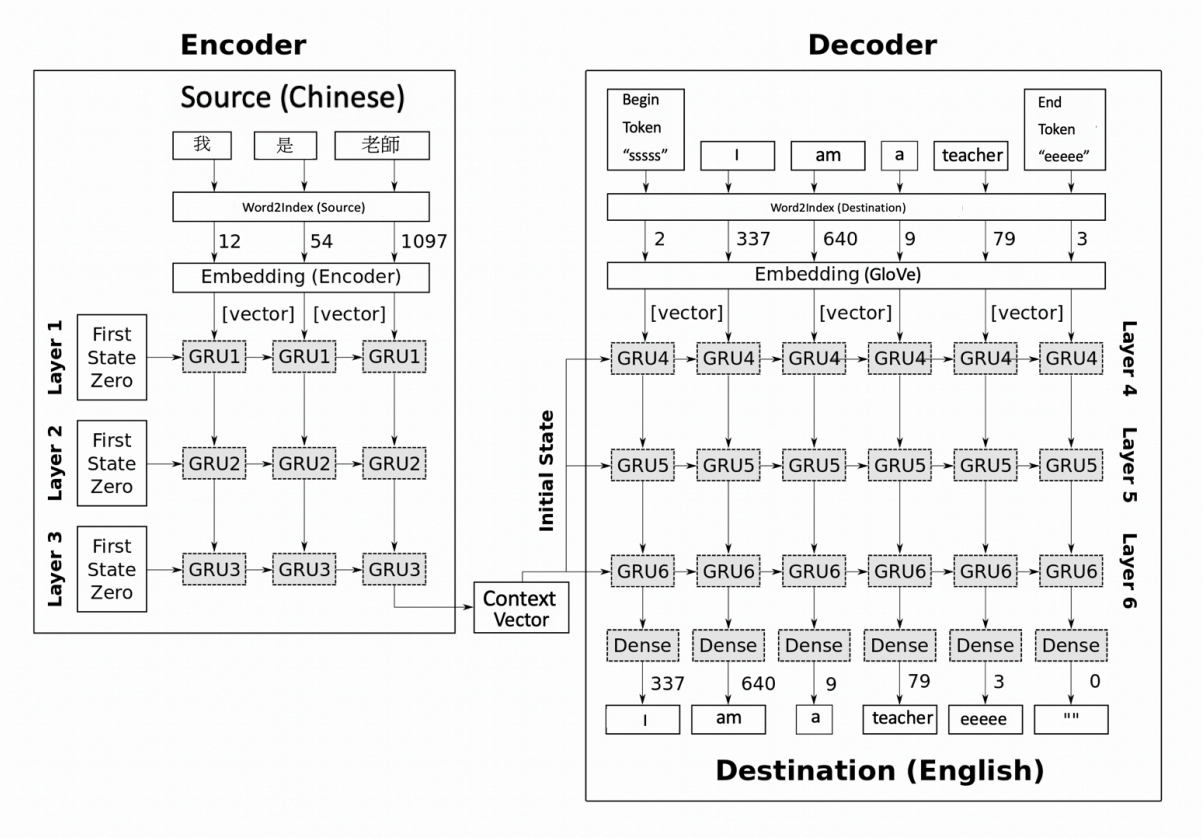

圖 2 模型架構與翻譯流程示意圖

最後，我們將1165025組中英文對照句當作機器翻譯模型的 training set，再利用一萬組中英文對照句當作模型的 validation set，跑了10個 epochs，batch size 設定為512。訓練結果如圖4所示。

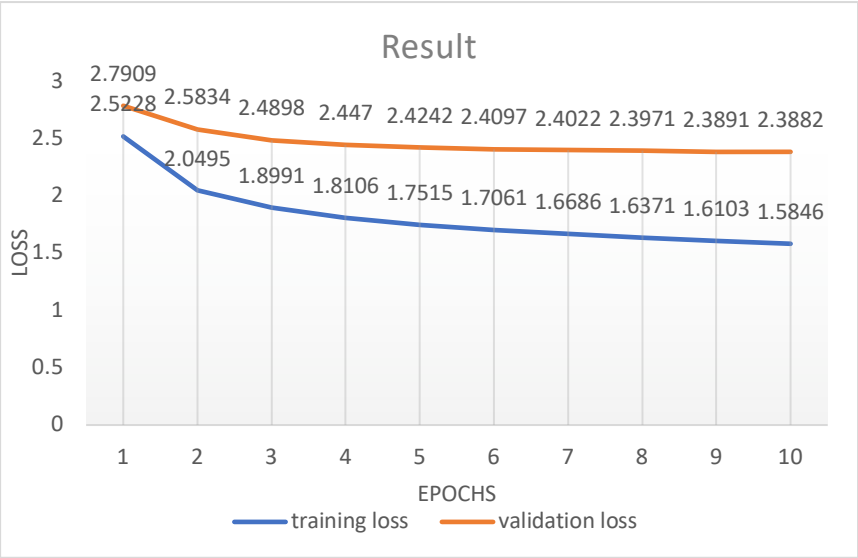

圖 4 training loss 和 validation loss 與訓練 epoch 數之關係圖

Validation Loss 最後訓練至2.3882，且有數個翻譯成功例子，圖5為其中一個成功翻譯的例子。

```
In [741]: translate_to_english(chinese = "這個圖是一條曲線")
          this graph is a curve
```

圖 5 中英 Machine Translation 成功範例

此外，為了清楚表現添加字幕的表現方式，我們運用線上資源翻譯的內容（中文）和模型翻譯出的內容（英文）添加回原影片並上傳至 YouTube，連結如下所示：

1. 影片翻譯平台之系統實作 微積分介紹
https://www.youtube.com/watch?v=DmXv5bzNRNc&t=56s
2. 影片翻譯平台之系統實作 數位聲訊與分析介紹
https://www.youtube.com/watch?v=ZelWcooumqo

Reference

[1] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 2227-2237, 2018.

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

[3] I. Sutskever, O. Vinyals, and Q.V. Le, Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems 27*, pages 3104-3112, 2014.

[4] R. Dey, and F.M. Salem, Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks, arXiv preprint arXiv:1701.05923, 2017.

[5] J.Pennington, R. Socher, and C. Manning, GloVe: Global Vectors for WordRepresentation. *EMNLP*, pages 1532-1543, 2014.

[6] R. Dey and F. M. Salem, "Gate-variants of Gated Recurrent Unit (GRU) neural networks", 2017 *IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston, MA, 2017, pp. 1597-1600, doi: 10.1109/MWSCAS.2017.8053243.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, Attention Is All You Need. *Advances in Neural Information Processing Systems 30*, pages 5998-6008, 2017.

心得感想

丁文淵：

透過這一次的實作專題，我學到了不少新技能，像是：如何架構一個簡單的Seq2Seq模型、如何有效率地蒐集網路資料等等。在熟悉這些新技能的過程中，我學到了不少有關深度學習、自然語言處理的知識。在架構我們的Seq2Seq模型時，我們也嘗試過添加更多更複雜的架構，如：Attention[7]。但我們在有限的時間內並沒有將那些複雜結構成功地加進我們的模型。雖然有點可惜，但在研究這些架構的過程中，我真的學到了不少。我非常感謝我的隊友，柳奕丞。這段時間若沒有他的努力，我們是無法完成此作品的。我也很感謝我的教授，劉奕汶教授。兩學期下來，教授不斷地給我們鼓勵與信心，並提出很多很好的建議，以確保我們的專題能順利進行下去。最後，我也想感謝帶我們做專題的學長，戴強麟學長。學長每一周都會關心我們的進度，並願意幫助我們解決各種專題上的障礙。總而言之，這次的專題非常的精彩，收穫良多。希望我未來能學到更多這方面的知識，利用深度學習設計出更多更好的作品。

柳奕丞：

於這次專題當中，我認為最大的收穫莫過於自主研究、分析論文、實作設計的能力，透過前人的研究，我們可以更快瞭解現今自然語言處理的發展，並且有幸透過此次專題實作翻譯系統並對學校的課程資源有直接的貢獻，個人感到欣喜且有成就感。其中，最

為感謝的就是專題指導教授，劉奕汶教授，謝謝老師在每次的例行會議中可以提供我們寶貴的意見，使我們在進行專題時能有一定方向且能及時修正問題點。另外，我也要感謝電機系碩士班學長，戴強麟學長，願意在每週開設特定時間和我們討論專題的進度，並且耐心講解一些重要的觀念。當然，還有我的組員，丁文淵，這份專題的完成必定需要兩方的配合和互相支援，缺一不可，每次的討論當中都有不同的啟發和了解。我對這兩位教授、學長，以及我的組員表達莫衷的感謝。不論未來的研究、工作內容是否和自然語言處理有直接相關性，但我相信「人生中沒有白走過的路。」每一次遭遇問題，就是了解自己的不足之處，並提醒我們應該努力之處；每一次的問題解決，就是表現自己對此次問題的了解，而未來仍有許多挑戰等著我們！