# 數碼時代下的文言學習與研究——文白對譯的機器學習芻議

# Research and learning of Classical Chinese in the digital era – A brief discussion on machine learning for translation between classical and modern Chinese

林葦葉
香港大學教育學院副教授
LAM Wai Ip
Associate Professor, Faculty of Education, The University of Hong Kong

摘要:
在中國與漢字文化圈國家的教育中，文言和文言文一向是中文或漢文學習的重要內容。文言作為語言系統，文言文作為情思哲理，學習文言儘管不完全等同於學習文言文，卻只能通過學習文言文獲致。如何在白話文已成主流的時代下，學生通過學習數量畢竟有限的文言文，有效遷移文言能力，以理解未學習過的文言文，從來是文言學習的宏旨和挑戰。

本報告首先介紹已經發展並實踐了十年的「文白對譯」學與教方法。文白對譯脫胎自王寧教授收錄在《訓詁學原理（第一版）》的〈 中學语文课本文言文注释研究〉，語文注釋與文意注釋不同，語文注釋旨在尋找與文言準確對當，而且學生能夠理解的白話，即嚴格對譯，以成為學生積累文言詞彙和句法等構式的庫存，「揆之本文而協，驗之他卷而通」，實現學習遷移。

文言對譯建立在文白一貫、略有差異的第一原則上，即文言和白話相同者多，而相異者少而且系統；學習文言，因此聚焦於文白之間的系統性的少數差異上，便文言成為「可以學習」(learnable) 的內容。

根據這原則，文白對譯提出六項對譯體例：

一、字字落實，只有極少情況下，文言毋須對譯成白話，則設符號〈 〉表示，如結構助詞「之」的運用；
二、不增字衍譯，所有白話對譯，必定在文言原文中找到根據，只有極少情況下，白話對譯增補文言省略或沒有的成份，則設符號（ ）表示，如容易造成誤解的主賓語省略；
三、只對譯字詞句等構式的語言意義，不作上下文意解釋，只有極少情況下，字詞句等構式意義與上下文意解釋距離過遠，不相連貫，則設符號＝表示特定構式在上下文中的解釋，如「為」字的具體意義，常常無法不按上下文意解釋；

以上三項，文白一貫為主，文白差異為輔。以下三項相反，文白差異為主：

四、文白語序不一致者，設符號〔 〕表示，如文言修飾語後置，白話修飾語前置，中小學課程中，文白語序不一致的常見差異僅七類；
五、文言活用者，設符號｛ ｝表示，唯能不以活用解釋，盡量不用，以免活用泛濫，如動詞用為名詞，按沈家煊名動包含說，便不算活用；
六、通假、古今字、錯字者，分別設 〇、□、△ 圈起白話對譯表示。

文白對譯的原則、體例以及所用符號，有助學生注意 (notice) 文白之間的系統性差異，而實踐證明，白話對譯本身可以理解 (comprehensible)，是學生從文言原文到語譯理解之間的有效輔翼工具 (scaffolding)。

文白對譯例子，如：

子　曰：「　　內　省　不 疚　，
先生 說：「（道德高尚的人）（向） 內心 反省，不 痛苦，
1-1, 2-2, 3-5, 4-6, 5-7, 6-8

夫 何 憂　　何 懼　　？」
那　擔憂〔甚麼〕　害怕〔甚麼〕？」
1-1, 2-3, 3-2, 4-5, 5-4

由於文白對譯之間一致者多，差異者少，因此便於利用機器學習的技術，訓練文言原文和白話對譯的翻譯，並且發展出雙文本詞語對應 (Bitext word alignment) 的數據，如上文第一句的 1-1, 2-2, 3-5, 4-6, 5-7, 6-8，即第 1 個文言原文詞彙對應第 1 個白話對譯，第 6 個文言原文詞彙對應第 8 個白話對譯，以預測未經訓練文言原文的白話對譯。

本研究嘗試在 R 環境中，先運用 word2vec 的文本挖掘技術，把十篇香港教育局公布的指定文言學習篇章及其對譯轉變成詞嵌入 (word embedding) 數據，然後運用 Keras 的深度學習技術訓練，以了解文白對譯在數碼時代下用作機器學習的可能性。由於訓練的文本語料數據不大，即使經文本增強技術 (text augmentation) 變換，依然有限，但結果能為日後更大規模發展提供具試驗意義的啟示。

關鍵詞：文言學習、文言白話對譯、機器學習、文言白話有效遷移