# 建置專業領域詞表，讓電腦來助一臂之力

【語文教育及編譯研究中心副研究員　吳鑑城】

若將學習語言視為興建一座建築物，文法就像是建築物的骨架，詞彙則是骨架上一片片的磚瓦，骨架固然重要，詞彙量若不足，語言能力仍難以提升，並連帶影響表達以及理解的能力。隨著學習者對專業領域（如商務、觀光等）的語用需求，專業英語（English for Specific Purpose, ESP）、專業華語（Chinese for Specific Purpose, CSP）等藉由教師依照專業領域及情境，教導學習者具備特定領域溝通之能力的專業語言教學概念也因應而生。而考量詞彙運用在不同領域的差異，研究者或教學者也開始建置各專業領域的詞表，以使教學時能更有成效。然而，若僅靠人工進行詞表編輯，不僅耗時費力，也容易有所遺漏。因此，本研究嘗試透過電腦輔助，從約 3 億 8,100 萬詞的 5 年份新聞資料所構成的語料庫中，自動蒐集特定專業領域詞彙的候選詞，供後續領域詞彙表編輯之參考。

本研究透過先蒐集既有的領域詞彙並經過篩選後，作為運用詞向量（word embedding）技術進行自動化蒐集時的種子詞彙，進而尋找其他同領域的詞彙。詞向量簡單而言，就是將每個目標詞利用一個向量（vector）去表示，而目標詞的向量則是由目標詞在文章中前後文共同出現過的詞彙所建構而成。因此，當兩個詞彙的詞向量越相似，表示著彼此常鄰近的詞彙越相近，也意味著他們使用的情境越加相近。

## 一、蒐集種子詞彙

本研究首先自政府網站以及大專院校之教學網站，蒐集商務及觀光領域現有之領域詞表，並進一步透過兩項指標進行種子詞彙的篩選：**（一）詞表收錄數**：若單一詞彙收錄於較多份領域詞表，應表示該詞彙穩定度較高及使用度較為領域所認可。**（二）語料詞頻數**：透過日常生活中常接觸的新聞語料進行使用度分析，詞彙頻率代表著在生活中可見度。經過篩選後的領域種子詞彙範例可見表 1、表 2。從表中可觀察到，詞彙的詞表收錄數跟詞頻並非完全正相關。應是因為收錄數代表的是領域對於該詞彙的領域認同度，而語料庫所提供詞頻則代表的是該詞於語境上的常用度。而在研究結果中發現，利用領域認同度越高的詞彙作為種子詞彙越能夠找到同領域詞彙。

表 1、商務領域種子詞彙範例

| 詞彙 | 詞表數 | 詞頻 |
|---|---|---|
| 關稅 | 20 | 16,033 |
| 合併 | 15 | 15,383 |
| 折扣 | 15 | 6,088 |
| 公司 | 14 | 218,131 |
| 企業 | 14 | 124,613 |
| 出口 | 14 | 31,970 |
| 文件 | 14 | 14,544 |
| 董事會 | 14 | 12,053 |

表 2、觀光領域種子詞彙範例

| 詞彙 | 詞表數 | 詞頻 |
|---|---|---|
| 觀光 | 18 | 52,963 |
| 民宿 | 18 | 11,829 |
| 旅館 | 17 | 11,446 |
| 旅客 | 16 | 40,330 |
| 觀光客 | 16 | 10,683 |
| 旅行社 | 16 | 9,668 |
| 觀光產業 | 16 | 3,786 |
| 溫泉 | 15 | 13,714 |

## 二、建置自動化蒐集領域詞彙模組

本研究利用新聞語料訓練詞向量模型。此模型可接受輸入指定詞彙，並回饋一組關聯詞，並提供這些關聯詞與指定詞彙之間的向量相似值。表 3 展示了跟「關稅」及「觀光」兩詞彙的關聯詞範例。

表 3、關聯詞彙範例

| 詞彙 | 關聯詞彙（相似值） |
|---|---|
| 關稅 | 懲罰性關稅(0.84)、進口關稅(0.83)、關稅稅率(0.80)、報復關稅(0.65)、鋼鋁稅(0.65)、進口稅(0.64)、反傾銷稅(0.63)、關稅壁壘(0.63)、徵稅 |

| | |
|---|---|
| | (0.62)、邊境稅(0.61)、反傾銷關稅(0.59)、貿易制裁(0.59)、關稅政策(0.58)、貿易逆差(0.56)、重稅(0.55)、貿易壁壘(0.55)… |
| 觀光 | 觀光產業(0.82)、觀光旅遊(0.75)、觀光業(0.68)、旅遊觀光(0.67)、觀光市場(0.66)、國際觀光(0.63)、旅遊業(0.63)、文化觀光(0.63)、旅遊(0.62)、深度旅遊(0.61)、觀光行銷(0.60)、城市觀光(0.60)、夜間觀光(0.59)、國內旅遊(0.58)、生態觀光(0.58)、休閒觀光(0.58)、運動觀光(0.57)… |

考量專業領域詞彙蒐集過程中，一開始既有的詞彙量往往不多，若是期待等到蒐集許多的領域詞彙方進行擴充，在實務方面的實用性將會降低。因此，本研究嘗試了不同種子數量所獲得的成果，包括利用 1 個種子進行領域詞彙擴充，並利用第一次獲得的關聯詞組作為第二輪的蒐集起點，最後依照關聯詞重複出現數作為候選列表排序（見表 4）。

表 4、單一種子經兩輪擴充所獲得的領域候選詞彙表

| **領域** | **領域候選詞彙（重疊數）** |
|---|---|
| **商務** | 關稅(63)、進口關稅(50)、懲罰性關稅(35)、貿易制裁(31)、報復關稅(25)、鋼鋁稅(25)、進口稅(24)、關稅壁壘(24)、...、智財權(4)、非關稅貿易障礙(4)、保護主義政策(4)、外銷(4)、順差(4)、轉口(4)、逆差(4)、貿易出超(4)、輸陸(4)、入超(4)、貿易總額(4)、貿易額(4)、...、固定成本(1)、產量(1) |
| **觀光** | 觀光(54)、觀光產業(44)、觀光旅遊(27)、旅遊觀光(25)、深度旅遊(22)、觀光市場(21)、旅遊業(21)、…、自由行(16)、陸客(15)、城市觀光(14)、旅行業者(14)、夜間觀光(13)、運動觀光(13)、產業觀光(13)、國內觀光(13)、...、溫泉資源(1)、天然資源(1)、人文資源(1)、物產(1)、自然資源(1)、林相(1) |

本研究為克服領域種子詞彙可能稀少的實務困難，所提出應用詞向量技術的領域詞彙蒐集模組，經實驗觀察發現即使只使用一個具領域代表性的種子詞彙，也能夠達到相當不錯的效果，若考慮輔以人工進行部分篩選，更有著快速蒐集大量領域詞彙的可行性。此外，本研究僅須使用一般性的新聞媒體語料庫，而不須特別建置領域文本語料庫，即可在不同的專業領域上，都能有效地協助蒐集領域詞彙。因此除可直接運用於商務及觀光領域詞編輯外，更可推廣至其他領域詞表的編輯工作中，提升詞表的編輯效能。

## 資料來源

吳鑑城（2020）。**華語文專業領域詞彙自動化發展研究。**國家教育研究院研究計畫，計畫編號：NAER-2019-029-C-1-1-B5-03。執行日期：2019-08-01 至 2020-12-31。連結網址：https://rh.naer.edu.tw/handle/2gxy3