數位典藏與數位人文 12 期（2023 年 10 月）
Journal of Digital Archives and Digital Humanities No. 12, pp. 43-82 (October, 2023)
DOI:10.6853/DADH.202310_(12).0003

詞嵌入應用於佛學研究——兼論詞嵌入模型評估

黃淑齡[1]、王昱鈞[2,*]

摘　要

詞嵌入是利用語料庫自動產生語義向量的方法，本論文的目標為探索詞嵌入在 Comprehensive Buddhist Electronic Text Archive（CBETA）漢文佛典中的可能應用面向。為取得適用於佛學研究的詞嵌入最佳模型，本文利用莊春江辭典、丁福保辭典和 Digital Dictionary of Buddhism 辭典建立實驗資料集，並設計偵測同義詞及干擾詞等兩種評估實驗來取得模型優化的基線。結果發現 Word2Vec CBOW（continuous bag-of-words）、Dimension 400、Window 10、Epoch 10 為最佳超參數組合，驗證正確率為 0.87，測試正確率為 0.86。據此，我們將 CBETA 語料分類訓練出不同詞嵌入模型，再跑出依據年代、譯者及部類的不同範圍語料對比詞表，並進行實際應用分析。本論文的主要貢獻有三：一、建置適用於漢文佛典研究之詞嵌入同義詞資料集；二、找出適於漢文佛典文本之詞嵌入超參數；三、探討與分析詞嵌入於漢文佛典研究之實例，包括可用於判斷譯詞的語義核心演變、能用於界定不明確的語義、能透過語義類比找出相關概念、能找出各部類的核心概念、能藉以拓展研究廣度和深度，以及可用於驗證傳統研究結果等面向。

關鍵詞：詞嵌入、漢文大藏經、佛學研究、語義關係、語義類比

投稿日期：2022 年 3 月 7 日；通過日期：2022 年 7 月 25 日。
[1] 法鼓文理學院博士生。
[2] 法鼓文理學院助理教授。
[*] 通訊作者：王昱鈞，Email: ycwang@dila.edu.tw

壹、前言

近年來，自然語言處理學界提出詞嵌入（word embedding）的概念，它是藉由計算詞彙的空間向量來理解詞義，並驗證了詞的意義與伴隨它們出現的上下文有關，上下文愈相似的詞語，語義關係愈接近（Bengio, Ducharme, Vincent, & Jauvin, 2003）。目前詞嵌入方法已廣泛應用於許多計算語言學問題，如語意分析、知識擷取、語法剖析、問答系統及機器翻譯等（B. Wang, Wang, Chen, Wang, & Kuo, 2019）。在人文領域亦利用詞嵌入方法考慮搭配詞的差異性，來進行不同詞彙的語義類比（Mikolov, Sutskever, Chen, Corrado, & Dean, 2013），或是進行語義解歧（Schmidt, 2015），也用來進行歷時性的語義變遷研究（Hamilton, Leskovec, & Jurafsky, 2016; Hengchen, Ros, Marjanen, & Tolonen, 2021; Kutuzov, Øvrelid, Szymanski, & Velldal, 2018），或從文學文本中提取社會網絡（Wohlgenannt, Chernyak, & Ilvovsky, 2016）等。若將文本視為詞彙之組合，即常用之詞袋（bag of words）概念，再計算所有詞彙的空間向量，將詞語間的共現關係建模成空間矩陣，就能觀察到具有相似語法行為的詞彙群聚成哪些主題；或者，也可以利用詞嵌入建立種子詞列表，進而成為一種監督式的主題建模方式（Schmidt, 2015）。詞嵌入的應用範圍可從詞彙語義層面進一步擴大到篇章語義層面，例如：可進行文本的主題分類或情感分析等等（Le & Mikolov, 2014; Leavy, Wade, Meaney, & Greene, 2018）。如上所述，詞嵌入在數位人文領域的應用愈來愈多元，已成為一重要的研究方式。

Chinese Buddhist Electronic Text Association（CBETA）中華電子佛典集成從 1998 年開始建置，至今已收集了包含《大正藏》、《卍新續藏》、《嘉興藏》、《漢譯南傳大藏經》、《高麗大藏經》等典籍，是全球最大的漢文電子佛典資料庫（侯坤宏、卓遵宏，2014）。CBETA 文本的特性包括：一、長時期，收集由東漢到民國約 2,000 年的資料；二、多學派，包含不同時期及地區的佛教主要學派思想；三、多譯者，常有同經異譯的版本；四、多來源語，除了在中國寫成的典籍外，還包含由梵語、中亞語、巴利語、藏語及日語譯成漢語的佛典（辛島靜志，2007）；五、內涵豐富，包括義理闡述、史傳地志、目錄版本、詩歌文學及音義詞典等（曾昭聰，2009）；六、多文體，兼具白話、文言、方言、韻文各種體裁。CBETA 多元的內容，對佛學研究者而言是極其寶貴的研究資源。然而，

過去佛學研究多針對人物、義理和語言等進行局部研究，未能充分利用 CBETA 文本特性的優勢進行知識抽取或大規模比較研究。在數位科技發達的今日，佛學在 CBETA 的支援下應有更廣闊的研究空間。這套電子佛典足以讓研究者從印刷時代的近距離閱讀，轉向超閱讀（如：過濾關鍵字、略讀、超連結，以及碎片化等等）和機器閱讀（如：自動問答、自動摘要、文件分類及自動評分等等）。然而要達到此目的，我們必須將傳統文本的敘事，從字句的書寫轉變成以數位資料庫來表達（Hayles, 2012）。詞嵌入就是其中初步的嘗試，因為詞嵌入可以將同義詞彙[1]以向量方式群聚，進而實現語義計算、主題判定與擷取文本，邁向超閱讀及機器閱讀的目標（Kamlovskaya, 2018）。CBETA 提供相當完整的後設分類，包括部類、經號、經名、卷數、作譯者及著書年代等等，其中不同類型或同類型中不同範圍的資料可作為詞嵌入演算結果的對比材料，得以藉由詞嵌入方法有效觀察詞彙於佛典文本中的關連，故本文採用 CBETA 作為詞嵌入應用的文本。我們希望先以不同時期或譯者在翻譯佛典時採用的詞彙差異性、不同部類的經典所反映出的思想取向，或是從幫助研究者拓展及深化研究內涵等角度，作為詞嵌入技術應用於佛學研究的初步目標，未來再擴展到探討主題分類的應用。

然而，由於目前多數研究提出的詞嵌入超參數（hyperparameters）[2]均基於西方語言之上（Burns, Brofos, Li, Chaudhuri, & Dexter, 2021; Leavy et al., 2018），並非針對 CBETA 文本所提出的最佳模型，而且也尚未有能評估不同模型優劣的標準。因此，為了給模型優化建立一條基線，本文嘗試提出佛典同義詞集，並設計驗證實驗來調校超參數，為佛典應用研究建立更準確的詞嵌入模型。本論文的主要貢獻有三：一、建置適用於漢文佛典研究之詞嵌入同義詞資料集；二、找出適於漢文佛典文本之詞嵌入超參數；三、探討與分析詞嵌入於漢文佛典研究之實例與相關面向。

---

1 由於多數學者（Tayor, 2003; Saeed, 2009）認為完全同義詞（perfect synonyms）應指語義相同，且可在任何語境中互相替換的詞，因此完全同義詞的數量極少。一般同義詞所指是意義相同或相近，但詞形不同的詞彙。例如《同義詞詞林》（梅家駒、竺一鳴、高蘊琦、殷鴻翔編，1983）所彙編的即是此類包含意義相同或相近的中文詞組，此觀點也為後來的編輯者所承襲，如《哈工大信息檢索研究室同義詞詞林擴展版》（哈爾濱工業大學信息檢索研究室編，轉引自 Ganlantree，2007）、《新編同義詞詞林》（亢世勇編，2015）。本文使用「同義詞」一詞，所指亦是包含同義詞及近義詞在內的意涵。

2 超參數是在模型訓練之前，事先根據經驗給定的參數，不同的超參數，訓練出來的模型也不相同。

貳、文獻回顧

一、傳統佛學的詞彙研究

佛經在東漢時期傳入中國，在音韻、詞彙和語法上對漢語產生了重大影響，相關的研究非常多。就詞彙上來說，主要研究類型大略可分為語義考釋、詞構分析、梵漢翻譯、語義變遷、風格考偽等（王冰，2011；竺家寧，2006；曾昭聰，2009）。以下將略述其研究方法。語義考釋是義理研究的基礎，相關著作最多，早期研究多以廣泛閱讀經本或其注疏來考察詞義的內涵，如李維琦（2003）提到考釋佛教的疑難詞語必須用到八種方法，即利用古注、翻檢辭書、與中土文獻對勘、從佛經本身求解、從眾多的使用同一詞語的語句中歸納、以經證經、揣摩文例、比照非漢文佛典等等。這些都是傳統常見的研究法，通常研究者需窮經皓首、博覽群經才能觸類旁通、多方引證。其中，與中土文獻對勘、從眾多的使用同一詞語的語句中歸納等方法已有利用共時語料釐清詞義的想法；曾昭聰（2005）參照《漢語大詞典》，列舉中古佛經或上古、近代漢語用例來釐清「名」和「字」的語義，則已兼及歷時語料的面向。竺家寧（1998）認為研究佛經語言學，除了要有精密的歷史觀念、清晰的辨偽過程及漢語語言學知識外，大量地羅列材料、窮盡式地探索也是要素之一。因此，佛典語料庫的出現在佛學研究上可謂一大突破。例如陳秀蘭（2018）研究四部具代表性的漢譯佛經，將其與對應的平行梵文本進行對勘，同時運用數理統計的方法作窮盡性的調查，研究它們共同的語言現象，並總結其中規律來說明這一時期的中外語言接觸對於漢語的影響。朱慶之（2019）通過與平行梵文語料的對讀，發現支謙譯《維摩詰經・菩薩品》的「是」字後置特殊判斷句可能是仿譯的產物，而非均是判斷句。陳淑庭（2021）以身體詞中的毛皮骨血類為例，透過統計中古時期的語料，發現這一類詞在組成雙音結構時具有封閉性強、基本規律穩定、引申義類型廣博且涵蓋範圍廣闊等特徵。陳鳳櫻（2021）則從《法華經》與《阿含經》語料中，試圖闡明中古漢譯佛經達成動詞和瞬成動詞與動詞前後之「已」字的互動關係。觀察近年論文發現，佛學應用語料庫的研究法多採用數理統計、平行對比、搭配詞（collocation）分析、關鍵字索引（key word in context, KWIC）觀察上下文等方法，至於人工智慧領域的方法學則較少觸及。

二、詞嵌入模型及超參數研究

Leavy 等人（2018）認為詞嵌入的應用已非常廣泛，然而我們在文學和歷史文本上構建這些模型時仍缺乏對超參數設定的研究和說明，多是沿用其他研究的預設值（Sculley & Pasanek, 2008）。因此，作者希望能以實際例子來證明評估參數的重要性。他們取用了 1700 到 1899 兩百年間包含多元主題的英文語料，試圖利用詞嵌入技術辨識出其中的醫療文本和疾病相關主題。過程中作者以 Word2vec 不同參數組合進行訓練（見表 1），藉以觀察不同設定的結果差異。評估方式是先從 19 世紀的醫學辭典中找出 10 個種子詞，然後針對每個嵌入模型抽取出與種子詞最相似的前 20 個詞。同時，取出文本中約 20,000 個檔案作為驗證資料集，其中人工標記為醫療文本者占 20%。以這些文本驗證每個詞嵌入模型所抽出的 20 個詞，愈多屬於醫療文本的詞表示該組模型的參數設定效能愈好。研究顯示，不同的參數組合所產出的詞表一致性非常低，平均數僅有 0.31。這代表針對不同語言和任務設定適合的詞嵌入參數是很重要的預備工作，亦即唯有經過定量驗證的參數設定才能為後續的分析提供較有意義的解讀。例如在這個 18、19 世紀英語醫療文本的辨識任務中，顯示以 Word2vec CBOW（continuous bag-of-words）、Dimension 100、keep Stopwords 為最佳組合，與一般的預設值不同。

表 1　Leavy 等人（2018）一文的參數評估設計

Parameters	Values
Model	Skip-Gram/CBOW
Dimension	25/100/400/800
Preprocessing	Lowercase/Stopwords

資料來源：Leavy 等人（2018）。

Hamilton 等人（2016）在歷時性語義變化研究中應用了詞嵌入，他們認為詞嵌入是劃時代的方法，但仍需要仔細的評估才能成為一種穩健的工具。因此，這個研究採用了跨越四種語言（英語、德語、法語和中文）和兩個世紀（1800–1999）的 6 個歷史語料庫，通過已知的歷史語義變遷來檢測三種詞嵌入模型 Positive Pointwise Mutual Information（PPMI）、Singular Value Decomposition （SVD）、Word2vec 的效果。在超參數部

分，主要遵循 Levy、Goldberg 與 Dagan（2015）的建議，使用 Window 4、Dimension 300 的設定。研究結果顯示，PPMI 明顯比其他兩種方法差，而 Skip-Gram with Negative Sampling（SGNS）和 SVD 詞嵌入效能在評估任務中表現相似，但作者最後還是選擇使用 Word2vec 的 SGNS 嵌入，理由是它們對詞頻和語義變化之間的關係提供了更好的估算。

B. Wang 等人（2019）進一步針對應用於語言分析的多種詞嵌入模型進行評估（包括 Word2vec SGNS、CBOW、GloVe、FastText、Ngram2Vec、Dictvec），並實證了評估詞嵌入模型的方法及指標。他們認為評估工作是以定量的、有代表性的指標來比較不同的詞嵌入模型的屬性，因此須考慮以下面向：（一）測試資料應客觀可靠；（二）應測試詞嵌入模型的多種屬性；（三）應與該模型後續所要處理的任務密切相關；（四）應在計算上具有高效性；（五）應具備統計學意義，即分數分布之間有足夠的差異以便進行區別。他們並將方法細分為外部評估和內部評估，外部評估是指以自然語言學問題，如：詞類標記（part-of-speech tagging）、組塊分析（chunking）、命名實體識別（named-entity recognition）、情感分析（sentiment analysis）、神經機器翻譯系統（neural machine translation）等來測試哪一種詞嵌入模型表現較好。內部評估則是單純以詞嵌入本身的成效來看，評估指標包括詞的相似度（word similarity）、詞的類比性（word analogy）、概念分類的正確性（concept categorization）、干擾詞的偵測率（outlier detection）等等。作者的結論是，在效能表現上不管是外部評估或內部評估，基於 Word2vec SGNS 的模型整體表現最好，且詞的相似性、詞的類比和概念分類是比較好的內在評估指標。更重要的是，他們指出沒有一個在所有任務中都表現良好的詞嵌入模型，可見隨著研究需要而選用模型、調校參數乃是必要工作。

Burns 等人（2021）則以詞嵌入進行拉丁文互文（intertextuality）研究。所謂互文性，是指文學中語言或語義上的相似性，如直接引用典故或僅是主題相似都包含在內。早期的互文研究多依賴統計重複出現的詞彙和短語匹配，或是使用序列對齊來進行。詞嵌入技術問世不久就成為此研究領域的重要方法，因此學界爭相進行拉丁語文獻的 Word2vec 模型訓練及競賽（Bjerva & Praet, 2015），以優化研究所需的工具。同時 Sprugnoli、Passarotti 與 Moretti（2019）也發展出拉丁文的同義詞資料集（synonym

selection dataset）作為實驗資料集，使競賽的結果更具公信力。Burns 等人評估了包括 Word2vec、FastText 和 nonce2vec 等方法。評估方式是上文提到的內部評估，第一個實驗是從 3 個干擾詞中將 1 個拉丁語單詞的同義詞找出來；第二個實驗是從語料中找出拉丁語詞典中 1,910 個詞的同義詞。結果 Word2vec 模型在兩項實驗中都表現最好。

三、中文領域的詞嵌入研究

在中文領域的詞嵌入研究方面，關於詞嵌入模型之比較，曾千蕙（2018）比較了八種詞嵌入方法，指出以內在評估而言，Word2vec 的 CBOW 和 Skip-Gram 表現最佳，其次是 GloVe；若以外在評估來看，在命名實體識別的任務中表現最好的是 Skip-Gram 和 Hyperspace Analogue to Language（HAL）。除此之外，目前中文的詞嵌入研究方向主要有二，一是利用詞嵌入計算來擴展詞彙或計算關鍵詞的相似度；另一則是搭配深度學習的其他方法，將詞嵌入作為預訓練詞彙向量的工具。

第一類論文如謝吉隆、楊芯淳（2018）探討國內風災新聞的報導演變，他們先進行關鍵詞分析，接著利用詞嵌入計算關鍵詞間的相似度，以達到詞彙擴展或觀察詞彙間在不同時期相似性的目的。黃泰霖、宋傳欽、姜志銘、譚克平與高桂惠（2019）則利用詞嵌入方法將一首詩轉換成一個向量，藉以衡量兩首詩之間的相似度；之後再利用成分分析法來探討促使唐詩流通的主要原因。張簡宇傑（2020）在發展工程文件自動摘要系統時，將含關鍵字詞之文句以 Word2vec 模型轉換成向量，並利用 TextRank 進行相似文句的重要度排序，再以重要性高的文句以及包含各類關鍵字詞的文句自動組成摘要。陳克威（2020）則是利用 Word2vec 找出內容具有相似特徵的學術論文並進行推薦。他也將學術論文中的參考文獻視為相關論文，藉以評估模型的表現，而實驗結果也證明了 Word2vec 的有效性。

第二類論文的數量較多，例如林昆賢、蔡俊明（2019）指出詞嵌入是銜接深度學習其他程序的重要步驟，因為詞嵌入比以往的詞編碼方式更能提升模型預測正確率，而且訓練後的詞向量也蘊涵了較為豐富的訊息。他以翻譯實驗為例，結果發現經預訓練詞向量者，翻譯成果之 Bilingual Evaluation Understudy（BLEU）評分普遍優於未經預訓練詞向量者。羅

文君（2019）採用 Word2vec 結合基於雙向長短期記憶網路（Bidirectional Long Short-Term Memory, BLSTM），進行詞性標記任務。實驗結果顯示，使用此模型在中國大陸《人民日報》1998 年 1 月分語料庫的詞性標記之整體準確率為 96.28%，與未加入詞嵌入的基線模型相比提升 0.76%；且未知詞的詞性標記之準確率為 81.51%，與基線模型相比提升 10.81%。曾元顯、許瑋倫、吳玟萱、古怡巧與陳學志（2020）則將詞嵌入應用於幽默對話系統的建置，他們利用 Word2vec 將大量的人類對話語料轉為詞向量，訓練出從文字序列產生另一種文字序列（seq2seq）的人工神經網路以回應使用者詢問的問題。其次，由於使用者的輸入可能是一句跟其當時情境相關的語句或詞彙，如「找出中秋節相關的笑話」或是僅有「中秋節」一詞。為求快速的回應，也採用 Word2vec 詞向量進行查詢擴展，如端午節、元宵節等，以提升檢索的成效。詹麒正（2020）以 PTT 八卦版為例，預測網路輿論的熱度風向，他以 Word2Vec 搭配 Long Short-Term Memory（LSTM）模型進行文章意見的分類預測，並使用文章及回覆等其他的特徵值作為輸入資料，配合特徵平衡等方法，對不同主題的輿論話題熱度進行模型訓練，結果發現預測結果有所收斂，表示其預測模型在一定程度上可預測 PTT 之輿論態勢。Hu、Zhao（2021）建構電影推薦系統時，使用 Word2Vec+CNN 的方法來提取電影名稱的特徵，並在「消費者—電影評分」矩陣的基礎上構建一個「使用者特徵—電影特徵」的融合矩陣。他們在「消費者—電影評分」矩陣上也考慮使用者的興趣因時間推移而發生的變化，因此使用 Word2Vec 模型對移動資料的標籤資訊進行訓練，獲得標籤之間的相似度，並根據相似度向使用者推薦影片。

綜上所述，由於多數論文在應用時選擇 Word2vec，或指出在數位人文領域中 Word2vec 的成效最好，Schmidt（2015）也認為 Word2vec 的 SGNS 是一條強大的基線，雖然它可能不是每項任務的最佳方法，但它在任何情況下都不會表現得太差。因此本文將評估的範圍限定於 Word2vec 的參數組合上。在評估方法上，我們根據前文的建議採取內部評估的兩種作法，說明如下：

（一）偵測同義詞（synonym detection）：給定一個詞以及另外一些詞的集合，其中該詞的同義詞包含在集合之中，再利用詞嵌入找出該詞的同義詞。

（二）偵測干擾詞（outlier word detection）：給定一個包含數個詞所構成的集合，再利用詞嵌入找出集合中最不相關的詞。

參、研究方法

一、Word2vec

本論文中我們利用 Mikolov 等人（2013）提出之 Word2vec 作為詞嵌入之方法，包含連續型詞袋模型（CBOW）與跳躍式模型（Skip-Gram）兩種模型，其將前饋式類神經網路的隱藏層移除，使用階層軟式最大化以及負例採樣方法來提高訓練的速度並改善訓練後詞向量的表示能力。以下為詳細說明。

（一）CBOW 連續型詞袋模型

CBOW 模型與前饋式類神經網路類似，不同之處在於 CBOW 將非線性隱藏層（non-linear hidden layer）移除，並且在輸入層的所有單詞皆共享隱藏層。如圖 1 所示，此模型包含三層，分別為輸入層、投影層、輸出層。

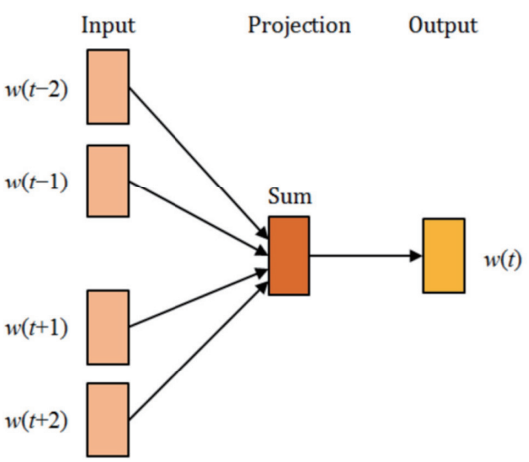

圖 1　連續型詞袋模型（CBOW）

資料來源：陳思澄、洪孝宗與陳柏琳（2015）。

已知當前詞 w_t 的上下文 w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2} 的情況下預測當前詞 w_t 出現的機率。從形式上看，CBOW 模型是由前 n 個詞語和後 n 個詞語去預測當前詞的模型。

（二）Skip-Gram 跳躍式模型

Skip-Gram 與 CBOW 相反，使用當前的詞來預測周圍的詞。在已知當前詞 w_t 的情況下，預測其上下文 w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2} 的機率。Skip-Gram 模型之概念如圖 2 所示。

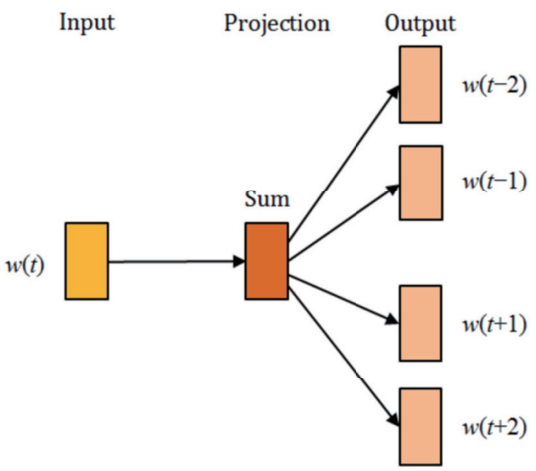

圖 2　跳躍式模型（Skip-Gram）

資料來源：陳思澄、洪孝宗與陳柏琳（2015）。

對於給定一個詞 w_t，可預測詞彙表中每個詞出現於上下文的機率，其算法如下：

$$P(w|w_t) = softmax\left(v_{w_t}^T v'_{w_{t+j}}\right) = \frac{exp\left(v_{w_t}^T v'_{w_{t+j}}\right)}{\sum_{w' \in V} exp\left(v_{w_t}^T v'_{w'}\right)} \quad (1)$$

其中 v_w 表示詞 w 在輸入詞嵌入矩陣中的詞向量，v'_w 表示詞 w 在輸出

詞嵌入矩陣中的詞向量。從形式上看，Skip-Gram 模型是反過來由目標詞語去預測前後 n 個詞語。而方程式中的 v_{w_t} 即是該詞 w_t 之詞嵌入向量。

無論 CBOW 或是 Skip-Gram 模型，有兩個超參數必須預先設定。其一為圖 1、圖 2 之 Projection 層的神經元數目，此決定了最終每個詞的詞嵌入向量維度。其二為上下文的 Window 大小，即前述之 w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2} 的數量。此二超參數對於 Word2vec 詞嵌入的效果有相當重大之影響，下節將詳述如何決定此超參數之實驗方式。此外，詞 w 可能是一字詞、二字詞或三字詞等中文詞彙，根據我們採用的自動分詞系統[3]所切分的結果為準。

二、實驗步驟

如文獻探討所述，針對不同語言和任務設定適合的詞嵌入參數是重要的先行工作，而要檢測參數是否有效，驗證及測試資料集的建置便很重要，如 B. Wang 等人（2019）所說，測試資料必須客觀可靠，還要跟所處理的任務密切相關。Leavy 等人（2018）是以人工標記文本類型作為測試集，Sprugnoli 等人（2019）則發展出拉丁文的同義詞資料集。本文採取後者的作法，因為後者更客觀的呈現詞語關係，也與我們後續要進行的分析較相關。

（一）建立實驗資料集

這部分分為三階段進行，首先建立同義詞集，再從中篩選出實驗資料集，然後由其中隨機抽選出驗證集（validation set）和測試集（test set）。我們從開放的佛學辭典——莊春江辭典、丁福保辭典和 Digital Dictionary of Buddhism 辭典（簡稱 DDB）——中抽取同義詞集。首先，我們觀察兩部中文佛學辭典的釋義形式，再利用關鍵詞過濾出同義詞，以簡化人工檢視的程序。例如：莊春江辭典利用「另譯為」這個關鍵詞過濾，可抽取出同義詞組如 { 穢 , 隨煩惱 , 隨染 , 染污 , 垢穢 , 鏽 }。丁福保辭典則擴大利用「譯作、譯為、又名、梵名、譯曰、又作、舊譯、又曰、梵言、譯言、舊稱、梵云、梵語、梵語曰、翻為、古云、或云」等詞過濾，再經人工檢視刪除錯誤，得到同義詞組如 { 三摩地 , 三昧 , 三摩提 , 三摩帝 , 三摩底 ,

3 我們採用法鼓文理學院針對 CBETA 開發的自動分詞系統進行分詞，其 F-score 為 0.9396（Y.-C. Wang, 2020）。

三麼地 , 三昧地 , 定 , 等持 , 正定 , 一境性 }。DDB 辭典是梵漢對譯辭典，其中漢語部分常有多個譯文，如 *adattâdāna* 譯為 { 不與取 , 偷盜 , 劫盜 , 強盜 , 盜 , 盜竊 , 罪 } 等，故我們刪掉梵文，將包含一個以上譯文的詞抽取出來作為同義詞組。這三種來源的同義詞組各自建置成同義詞集，其中，詞及詞組的數量統計如表 2。在這些同義詞集中，包含了不同類型的同義詞組合，例如：1. 不同音譯組合，如 { 一闡提 , 阿闡底迦 }；2. 音譯和義譯組合，如 { 伊帝目多伽 , 本事 }；3. 意義上的直譯和轉譯組合，如 { 等持 , 一心 , 定 }；及 4. 異體字的翻譯組合，如 { 鄰近 , 隣近 } 等多樣形態。這些型態常混合出現，且能凸顯漢語及梵漢翻譯的語言特徵，因此很適合用來驗證 CBETA 詞嵌入模型效能。

接著，我們將上述三種同義詞集合併、去除重複後再篩選出適合本實驗的資料集，篩選的原則是刪除罕見詞及單音節詞。原因是由於詞集中包含罕見的人、地、物名，如 { 漚辰那拘尼摩 , 漚辰那拘束摩 }（佛的字號）、{ 虛宿 , 阿濕毘儞 }（星宿名）及 { 于闍羅 , 于闍那 }（木名）等，如果這些詞出現在評估題目中，很可能因為在分類語料中找不到詞而無法評估。其次，同義詞集中的單音節詞歧義性太高，如 { 一 , 妒 }，也不利評估。故為避免評分干擾，實驗資料集僅保留那些在 CBETA 詞頻高於 10 的雙音節以上詞。實驗資料集的詞及詞組數量統計如表 3，所有同義詞組的數量統計如表 4。數量最多的前兩名同義詞組分別有 20 個詞和 16 個詞，20 詞內容為 { 一心 , 信心 , 信樂 , 善意 , 增上 , 增上心 , 心念 , 志意 , 志樂 , 意樂 , 樂欲 , 欲樂 , 正信 , 正心 , 深心 , 發心 , 直心 , 至心 , 至誠 , 誠心 }，16 詞內容為 { 厭離 , 對治 , 捨離 , 擇滅 , 斷盡 , 斷除 , 棄捨 , 永斷 , 消滅 , 滅惡 , 滅盡 , 滅除 , 遠離 , 除斷 , 除滅 , 頓斷 }。最後，我們從中抽取出驗證題目 2,000 筆，以及測試題目 2,000 筆，題目的設計如下節所述。

表 2　同義詞集中詞與詞組的數量統計

來源	同義詞數（去重複）	同意詞組			
		數量	最多詞數	最少詞數	平均詞數
莊春江辭典	1,006	324	11	2	3.28
丁福保辭典	8,489	4,019	19	2	2.61
DDB	11,690	6,791	40	2	3.08

資料來源：作者自行整理。

註：DDB = Digital Dictionary of Buddhism。

表 3　實驗資料集中詞與詞組的數量統計

來源	同義詞數（去重複）	同意詞組			
		數量	最多詞數	最少詞數	平均詞數
莊春江辭典、丁福保辭典、DDB	5,038	3,498	20	2	2.59

資料來源：作者自行整理。

註：DDB = Digital Dictionary of Buddhism。

表 4　實驗資料集中所有詞組的數量統計

同義詞組所含詞數	2	3	4	5	6	7	8	9	10	11	12	15	16	20
同義詞組數量	2,391	649	238	107	40	33	16	10	7	1	3	1	1	1

資料來源：作者自行整理。

（二）設計評估實驗

本研究的兩個實驗，一是偵測同義詞，一是偵測干擾詞。偵測同義詞的實驗中每題包含 1 個詞作為題目，選項中包含 1 個同義詞及 3 個非同義詞，我們要測試系統是否能從 4 個選項中挑出正確同義詞。例如：

寒林：(1) 五旬 (2) 首寂 (3) 屍陀林 (4) 傳說

其中 (3) 為正確答案。

為了自動產生題目，我們從實驗資料集中隨機抽取一同義詞組，再從中隨機抽取 2 詞作為題目和答案，錯誤答案則隨機從實驗資料集中取出 3 個詞，但此 3 詞必須不是題目的同義詞，如表 5 所示。正確率的計算公式如下：正確率 ＝ 答對正確同義詞題數 ÷ 所有題數。系統答題的方式是針對題目和 (1) 至 (4) 欄詞彙進行相似度計算，以相似度最高的為答案。

偵測干擾詞的實驗中，每個題目包含 3 個同義詞及 1 個非同義詞，我們要測試系統是否能在 4 個詞當中挑出非同義詞。例如：

(1) 阿鼻地獄 (2) 無間地獄 (3) 阿鼻獄 (4) 滿願子

其中 (4) 為非同義詞。

自動產生題目的作法是先挑出實驗資料集中詞彙數超過 3 的組合，每組再從中隨機挑選 3 個詞，然後隨機從實驗資料集中抽出 1 個詞作為誘答

詞，但此詞必須不是前面 3 詞的同義詞，如表 6 所示。正確率的計算公式如下：正確率 ＝ 答對正確干擾詞題數 ÷ 所有題數。系統答題的方式是將 (1) 至 (4) 每一詞彙與其他 3 個詞進行相似度計算後取其平均值，平均值代表該詞與其他 3 詞的相似度。取其中相似度最低者作為答案。

根據上述作法，本實驗共產生偵測同義詞 2,000 題，1,000 題作為驗證資料，另 1,000 題作為測試資料；偵測干擾詞 2,000 題，1,000 題作為驗證資料，另 1,000 題作為測試資料。

表 5　本文「偵測同義詞」實驗設計舉例

編號	題目（同義詞）	答案（同義詞）	誘答詞	誘答詞	誘答詞
1	寒林	屍陀林	五旬	首寂	傳說
2	蘇達梨舍那	善見	胚胎	念法	麤惑
3	尸羅律儀	戒行	遍行	兩邊	野馬
4	自說	憂陀那	耶舍	偈陀	瓦師
5	婆須達多	財施	無性	有所得	四等心
6	夭沒	終沒	振動	布灑	休息
7	周遍	周匝	悅意	旃簸迦	室哩
8	無生法忍	無生忍	開導	印解	癡冥
9	遍知	了智	大光	沙彌	律藏
10	無邊際	無窮	奉獻	妙觀	抖擻

資料來源：作者自行整理。

表 6　本文「偵測干擾詞」實驗設計舉例

編號	同義詞	同義詞	同義詞	誘答詞
1	阿鼻地獄	無間地獄	阿鼻獄	滿願子
2	霜佉	商佉	餉佉	優婆斯
3	達磨	達麼	達摩	宣說
4	迦多衍那	文飾	迦旃延	瞿夷
5	踰繕那	由旬	俞旬	勸勉
6	陵蔑	輕陵	凌蔑	自然
7	聖道	道諦	道聖諦	相符
8	摧伏	降伏	摧滅	不異
9	正覺	等正覺	偏覺	千萬億
10	五神通	般遮旬	五通	禪頭

資料來源：作者自行整理。

肆、結果與討論

一、實驗結果

在 Leavy 等人（2018）的實驗中，針對英文大小寫轉換（lowercase）及停用詞（stopword）的設定與否也進行了評估，他分析指出「進行文學和歷史文本詞嵌入時，參數的選擇非常重要；但與標準作法相反，即不將所有文本轉換為小寫字母和保留停用詞會帶來更好的效能」。據此，本實驗不進行文本前處理的評估，僅就模型參數本身進行調校。但相較於 Leavy 等人，我們增加了 Window 及 Epoch 兩種屬性的調校，前者乃設定語言模型決定詞時前後要看幾個詞的範圍，後者則設定訓練的總回數。完整的參數設計請見表 7。

表 7　本文的參數評估設計

參數	設定值	說明
Model	Skip-Gram/CBOW	設定使用語言模型
Dimension	25/100/400/800	設定每一個詞用幾個維度來表示
Window	5/10/15/20	設定語言模型決定詞時前後要看幾個詞的範圍
Epoch	1-(80/100)	設定訓練的總回數
Min_count	2	設定排除文本中出現次數少於 2 次的詞

資料來源：作者自行整理。

根據表 7 的參數組合，在總計 32 回的調校訓練中，結果以 CBOW、Dimension 400、 Window 10、Epoch 10 的參數組合表現最好。在偵測同義詞和干擾詞的實驗中，兩項的平均驗證正確率為 0.869。再以此最佳模型進行測試實驗，兩項的平均測試正確率為 0.856。詳細結果請見表 8，實驗結果中的平均正確率（accuracy）是實驗一和實驗二分數的平均值，排名（rank）則是依分數進行的排序，分數愈高排名愈前面。對比曲線圖請見圖 3。

表 8　本文參數評估實驗結果

參數設定					實驗結果			
No.	Model	Dimension	Window	Epoch	Min_count	Best_Epoch	Accuracy	Rank
1	Skip-Gram	25	5	1-100	2	100	0.8270	26
2	Skip-Gram	25	10	1-100	2	99	0.8155	30
3	Skip-Gram	25	15	1-100	2	100	0.8150	31
4	Skip-Gram	25	20	1-100	2	100	0.8005	32
5	Skip-Gram	100	5	1-100	2	98	0.8580	13
6	Skip-Gram	100	10	1-100	2	100	0.8565	15
7	Skip-Gram	100	15	1-100	2	99	0.8495	22
8	Skip-Gram	100	20	1-100	2	100	0.8430	24
9	Skip-Gram	400	5	1-100	2	92	0.8665	2
10	Skip-Gram	400	10	1-100	2	98	0.8645	5
11	Skip-Gram	400	15	1-100	2	100	0.8605	9
12	Skip-Gram	400	20	1-100	2	100	0.8555	19
13	Skip-Gram	800	5	1-100	2	80	0.8590	11
14	Skip-Gram	800	10	1-100	2	97	0.8575	14
15	Skip-Gram	800	15	1-100	2	99	0.8545	20
16	Skip-Gram	800	20	1-100	2	100	0.8560	17
17	CBOW	25	5	1-80	2	38	0.8320	25
18	CBOW	25	10	1-80	2	15	0.8270	26
19	CBOW	25	15	1-80	2	16	0.8210	28
20	CBOW	25	20	1-80	2	38	0.8195	29
21	CBOW	100	5	1-80	2	52	0.8545	20
22	CBOW	100	10	1-80	2	9	0.8595	10
23	CBOW	100	15	1-80	2	22	0.8590	11
24	CBOW	100	20	1-80	2	11	0.8560	17
25	CBOW	400	5	1-80	2	11	0.8655	4
26	CBOW	400	10	1-80	2	10	0.8690	1
27	CBOW	400	15	1-80	2	30	0.8615	8
28	CBOW	400	20	1-80	2	36	0.8565	15
29	CBOW	800	5	1-80	2	42	0.8660	3
30	CBOW	800	10	1-80	2	12	0.8620	7
31	CBOW	800	15	1-80	2	29	0.8645	5
32	CBOW	800	20	1-80	2	4	0.8495	22

資料來源：作者自行整理。

圖 3 分別列出 Skip-Gram 和 CBOW 兩種模型的表現，圖例中的編號是表 8 最左欄的參數組編號。藉由表 8 和圖 3，我們分別觀察模型、維度及取詞視窗的差別會帶來那些影響。首先，Skip-Gram 的平均排名是 18，

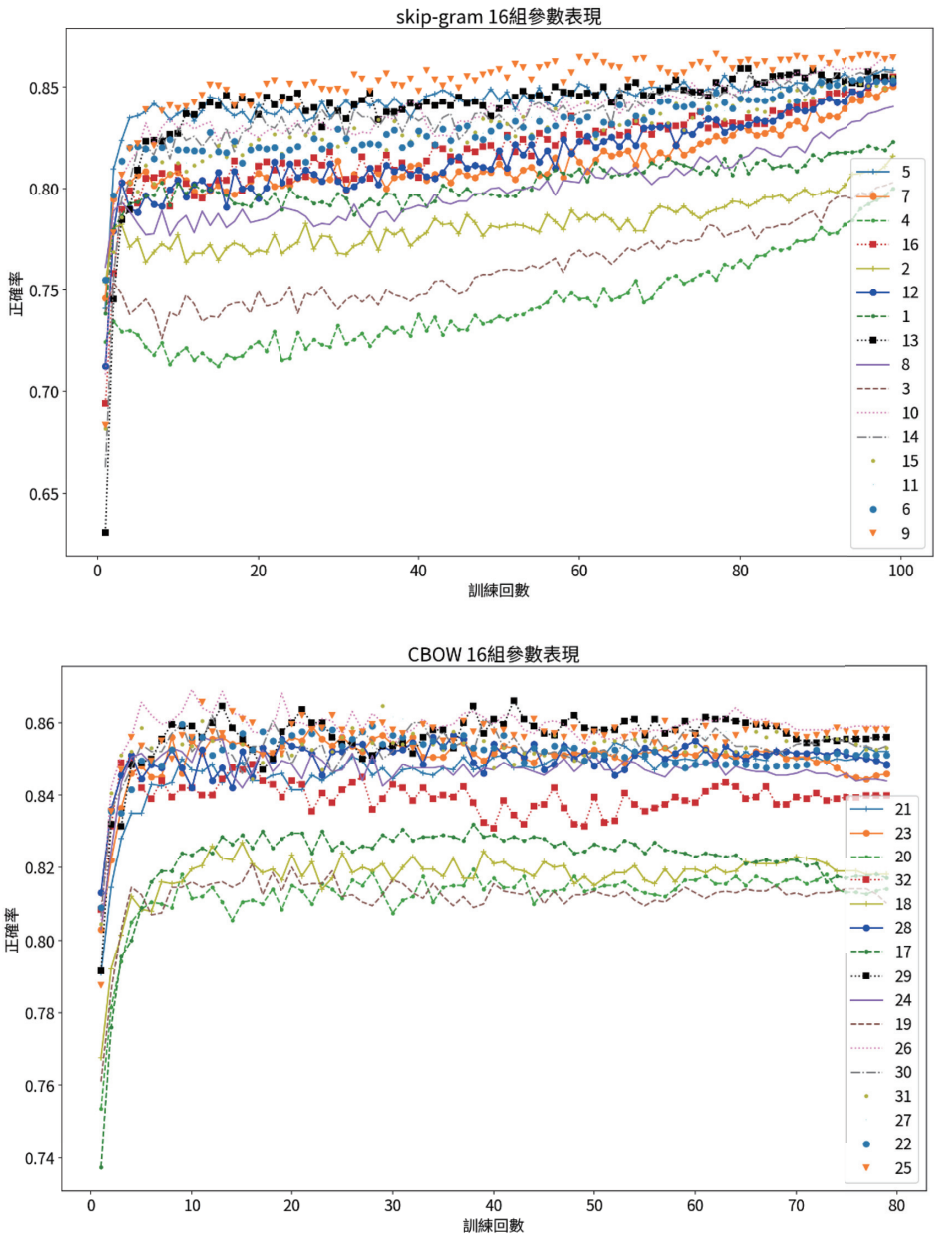

圖 3　本文參數評估實驗結果曲線圖

資料來源：作者自行整理。

CBOW 是 14，後者表現較好。其次，我們觀察到 Skip-Gram 正確率曲線隨著訓練回數增加而有向上爬升的趨勢，所以我們訓練到 100 回，以確認其最佳參數組合的正確率已達頂端；相對的，CBOW 模型的正確率曲線是

平坦的，因此我們只訓練到 80 回。第三，Skip-Gram 的 1-4 及 CBOW 的 17-20 是相對應的參數組（即 1 和 17 除模型外，其他參數相同，2 和 18 以下依此類推），在圖中皆是綠色系的線條，表示不管那種模型，維度 25 的正確率都最低。Skip-Gram 維度以 400 最好，CBOW 則 400/800 表現都不錯，顯示較高維度表現較佳。最後，從取詞視窗來看，如果固定模型及維度，會發現在大多數情況下取詞視窗 5-10 的表現都比 15-20 好。具體而言 Skip-Gram 以視窗 5 最好，CBOW 則視窗 5 或 10 高低互見。

二、語料分組及訓練模型

我們以 CBETA 全部語料進行訓練，得到最佳參數組合後，再對 CBETA 語料中年代、譯者、部類三個不同的類型，分別以詞嵌入訓練詞語間的關係，訓練參數即上節所得到之最佳參數組合 CBOW、Dimension 400、Window 10、Epoch 10。具體作法是先對上述三個類型進行分組，分組之方式如表 9 所示。接著，以上述參數分別用各組的語料來訓練詞嵌入模型，產出各組詞彙的詞嵌入向量，再進行對比分析。首先，針對每個詞，依據詞嵌入結果從不同群組中計算出與它相似度最高的前 20 詞，相似度計算公式如下：

$$similarity(w_1, w_2) = \cos(\vec{w_1}, \vec{w_2}) = \frac{\vec{w_1} \cdot \vec{w_2}}{||\vec{w_1}|| \, ||\vec{w_2}||} \tag{2}$$

其中 w_1, w_2 代表相異的兩個詞，$\vec{w_1}$, $\vec{w_2}$ 代表這兩個詞的詞嵌入向量。然後，再兩兩群組比較同一詞彙之同義詞的重疊率，重疊率高者表示相似度高，也就是語義變化不大的意思；重疊率低者表示在這兩個群組間，該詞語義可能發生變遷。例如，表 10「凡夫」在鳩摩羅什時期，和更早期的安世高相比，在詞嵌入模型跑出的 Top 20 同義詞中，交集有 4 個，分別是 { 愚癡 , 迷惑 , 賢聖 , 顛倒 }，則重疊率為 $\frac{\text{交集詞數}}{\text{總詞數}} = 4/20 = 0.2$。

一般來說，要計算詞彙語義相似度可直接計算其餘弦（cosine）相似度，然而，我們無法直接計算同一個詞在不同詞嵌入模型下的語義相似度，因為如果將所有語料混在一起訓練，則同一個詞只會得到一組向量；亦即必須分開來訓練同一個詞才能得到不同的向量。但是分開來訓練所得到的向量空間並不相同，無法直接進行比較。為解決此問題，Hamilton 等人（2016）提出正交普魯克（orthogonal Procrustes）分析來對齊向量坐標軸，以比較歷時性的詞嵌入向量。這種方法可以看出同一詞在不同詞嵌入模型語料中所產生的向量變化，但是只有數字的呈現。為了直接觀察同一詞嵌入模型所訓練出的同義詞，Leavy 等人（2018）利用同義詞的重疊率計算來達成。這種方法可以不受向量空間的限制，並且方便人文學者觀察同一詞在不同類型語料中各有哪些同義詞，以確認其語義變遷的具體軌跡。本文也採取這種作法。

下節將討論這些對比詞表所顯示的意義，以及它們在佛學研究上帶給我們什麼啟發。

表 9　CBETA 不同類型與範圍的語料分組

分類	對比項目	詞數
年代	漢魏晉南北朝（67–580 CE）	23,032,086
	隋唐五代（581–1100 CE）	68,266,824
	宋元明清（1101–1911 CE）	55,644,877
譯者	安世高、支婁迦讖、支謙、竺法護（東漢魏晉）	2,067,292
	鳩摩羅什、法顯、菩提流支、真諦（東晉南北朝）	5,069,021
	玄奘、義淨、實叉難陀、闍那崛多（唐貞觀至貞元）	7,946,217
部類	阿含部類	2,415,385
	般若部類	7,172,032
	密教部類	10,992,426
	律部類	14,056,791
	瑜伽部類	8,756,172
	禪宗部類	21,430,360

資料來源：作者自行整理。

註：表內「分類」取自 CBETA online reader（https://cbetaonline.dila.edu.tw/）的「經文選擇」，分別依據部類、作譯者、時間查詢所得。其中「譯者」部分又根據梁啟超（1998）分類中各期的代表人物選定。梁啟超將佛經翻譯分為三期：東漢至魏晉為第一期，東晉南北朝為第二期，唐貞觀至貞元為第三期；「詞數」是指法鼓文理學院自動分詞系統所切分的詞數。

三、詞嵌入應用於佛典研究的面向

在詞嵌入的應用中，最常見的是詞彙的歷時語義變遷、詞類比及語義相似度比較（Kamlovskaya, 2018; Schnabel, Labutov, Mimno, & Joachims, 2015），本文除將從這三個部分開展佛典研究應用外，也嘗試將詞嵌入結果作為義理研究的基礎，詳細分析如下文。

（一）從年代和譯者看出譯詞的語義核心演變

以「凡夫」為例，它的梵語可拆解為 *bāla-pṛthag-jana*，直譯對應到漢語的「愚－異－生」三字。根據《佛光大辭典》（釋慈怡編，1990），此詞指凡庸之人，「凡夫以無明之故，隨業受報，不得自在，墮於諸趣之中，遂產生種種類別之眾生，故應正譯為異生」（頁 730）。但從表 10 的對比來看，早期安世高等人多採「愚」義來翻譯，如「愚戇、愚冥、愚癡、愚騃、無明」等為其同義詞，它的反義則是「賢聖、聖賢」；到了鳩摩羅什等人「生（新生）」的意思才被帶入，所以同義詞中「毛道」名列第三，但主要意義仍偏向「愚、惑」之義；玄奘之後，「生」的意義似又削弱，仍以「愚」義為主。或許「凡夫」的翻譯用語也受時代影響，故我們擴展到更大的時空範圍來看，前述譯者都屬於表 11 魏晉南北朝到初唐年代的人，當時「凡夫」多譯為「愚」；直到隋唐五代後就可以看到「異」和「生」的意義大量的出現了，「異生」這個直譯詞甚至出現在表中第二位，「凡小」等帶有「新生」意義的詞也出現在列；宋元明清時，我們現在常用來指涉「凡夫」的「眾生」一詞也出現了，它不再帶有早期濃厚的「愚、惑」意味。由「愚人」到「眾生」，應可視為「凡夫」這個詞在漢語不同階段中的語義發展趨向。

表 10　從譯者對比「凡夫」詞嵌入演算結果（依相關度高低排序）

（東漢魏晉）安世高、支婁迦讖、支謙、竺法護	（東晉南北朝）鳩摩羅什、法顯、菩提流支、真諦	（唐貞觀至貞元）玄奘、義淨、實叉難陀、闍那崛多
愚戇	聖人	愚癡
愚冥	二乘	愚夫
賢聖	毛道	智慧
邪見	愚癡	纏垢
顛倒	凡夫人	邪法

表 10　從譯者對比「凡夫」詞嵌入演算結果（依相關度高低排序）（續）

（東漢魏晉）安世高、支婁迦讖、支謙、竺法護	（東晉南北朝）鳩摩羅什、法顯、菩提流支、真諦	（唐貞觀至貞元）玄奘、義淨、實叉難陀、闍那崛多
度世	迷惑	黑闇
正見	顛倒法	真見
愚癡	惑倒	燒惱
有為	賢聖	灌輪
愚	誑惑	阿曲
士	學人	棄背
緣覺	顛倒	蔽
世法	執著	虛偽
無明	虛妄	羂網
愚騃	聖法	盲冥
聖賢	愚人	炬
塵勞	妄	薄弱
聲聞	凡人	輪迴
俗法	貪愛	賢聖
迷惑	妄取	昏闇

資料來源：作者自行整理。

表 11　從年代對比「凡夫」詞嵌入演算結果（依相關度高低排序）

漢魏晉南北朝（67–580 CE）	隋唐五代（581–1101 CE）	宋元明清（1101–1911 CE）
二乘[4]	二乘	二乘
愚戇	異生	凡小
凡夫人	凡愚	凡愚
聖人	愚夫	凡夫法
凡夫法	凡小	博地
學人	凡失	異生
聖法	聖人	愚夫
毛道	二乘心	凡夫
顛倒	沈空	權小
凡愚	凡夫法	凡外

4　從表 11 中可以看到「二乘」一詞是各時代「凡夫」的 Top1 同義詞，但和表 9 比較，安世高時期和玄奘時期「二乘」卻都未出現在詞表中。關於這個問題我們進行三項觀察，首先在詞嵌入演算下，「二乘」出現在安世高時期的第 394 個同義詞，鳩摩羅什時期出現在第 2 位，玄奘時期出現在 495 位。其次，「二乘」的詞頻在安世高時期是 12，鳩摩羅什時期是 526，玄奘時期是 597。第三，佛學中有佛、菩薩、二乘（聲聞及緣覺）、凡夫這樣的階位概念，在鳩摩羅什時期，「二乘」的使用語境似較偏向「凡夫」一邊，例如「菩薩聞外道惡人及二乘惡人說佛法中非法非律」（鳩摩羅什《梵網經》），而玄奘時期則偏向「聖人」那邊，例如「此十惡法尚礙善趣二乘聖道」（玄奘《大般若波羅蜜多經》）。雖然詞嵌入計算不區分近反義，但是不同語義會影響語境，進而影響演算結果。據上所述，我們推測安世高時期「二乘」出現在凡夫的同義詞列表的後段是因為數量太少；之後用量增加而成為鳩摩羅什時期凡夫的 Top2 同義詞；到了玄奘時期，因為它和凡夫的語義偏離，影響到語境，因此它和凡夫的相關度又降低到後段。

表 11　從年代對比「凡夫」詞嵌入演算結果（依相關度高低排序）（續）

漢魏晉南北朝（67–580 CE）	隋唐五代（581–1101 CE）	宋元明清（1101–1911 CE）
凡人	凡人	八導
虛妄法	愚癡	毛道
愚癡	凢	聲聞
妄取	愚人	凢
凡	欣滅	倒見
迷惑	愚痴	眾生
欲界	有學	品愚
聲聞	迷倒	有漏
倒想	業繫	外道心
緣覺	凢夫	聲聞見

資料來源：作者自行整理。

（二）用已知語義界定新詞語義

中古時代漢譯佛典中出現許多新的複合形容詞，有些很難從字面上瞭解其語義，這時我們可以透過詞嵌入的方法，藉由已知的語義來定義這些尚不熟悉的新詞。以「端嚴」為例，表 12 顯示安世高等人使用「端嚴」時，與「麗妙、細妙、顯赫、高大、鮮澤、弈弈、煒曄」等語義接近，這些詞有的形容光采，有的說明形狀，大多可修飾物品。到鳩摩羅什等人時，「端嚴」的語義擴大了，與「殊特、殊妙、淨潔、端正、鮮白、香潔、可愛、鮮潔」等語義相關，這些詞部分是形容物品，部分則是形容人的相貌及能力，特別與「新鮮、潔淨」有關。到了玄奘時期，「端嚴」的同義詞變為「端正、超絕、殊妙、美妙、姝妙、超倫、妙好」，逐漸往形容人的相貌方向發展，從它的反義詞「醜陋、鄙陋」也可驗證這一點。表 12 可說呈現了三個不同時期譯者對「端嚴」一詞的看法，同時也提供了「端嚴」語義核心變化的清楚軌跡。

（三）以語義類比找出相關概念

早在 Mikolov 等人（2013）提出 Word2vec 工具時，他們即開始推廣詞嵌入向量的類比（word analogy）功能。它的目的是對於給定之樣本關係 $a : b$，以及一詞 x，其需找到一個詞 y，讓 $x : y$ 最類似於樣本關係 $a : b$。

表 12　從譯者對比「端嚴」詞嵌入演算結果（依相關度高低排序）

（東漢魏晉）安世高、支婁迦讖、支謙、竺法護	（東晉南北朝）鳩摩羅什、法顯、菩提流支、真諦	（唐貞觀至貞元）玄奘、義淨、實叉難陀、闍那崛多
麗妙	殊特	端正
月氏國	殊妙	形貌
軒窓	淨潔	容貌
顯赫	閻浮檀金	顏貌
白毫	威光	醜陋
高大	端正	超絕
龍象	容貌	顏容
鮮澤	金山	殊妙
南吳	鮮白	面貌
細妙	顏貌	容儀
弈弈	香潔	儀容
煒曄	照曜	儀貌
歸崇	可愛	美妙
山石	五色	形容
器樹	橋津	姝妙
漆	相當	超倫
出沒	面貌	鄙陋
衣樹	身色	八十種好
環珮	鮮潔	妙好
連綿	街巷	三十二相

資料來源：作者自行整理。

例如我們問「女人之於皇后，相當於男人之於什麼？」由詞嵌入運算可得到「國王」的結果。詞類比一度推動了詞嵌入的研究熱潮，但近來學界也開始探討它的應用面狹隘或存在偏見等問題（Gonen & Goldberg, 2019）。由於詞類比的驗證必須以人工設計如上述 4 個一組的推論集，其中很難不含主觀意見，如「哪些類比才算是詞類比？何謂詞類比的正確答案？」等都被質疑帶有偏見。因此很少研究以這種語義關係去調校超參數。因此除非刻意去設計，大部分的類比推理結果都不好，學界也持續在尋求改進方法（Nissim, van Noord, & van der Goot, 2020）。本文以 CBETA 全部語料作為詞類比運算的資料，發現的確如上所述，它的實際應用不如預期。然而，我們仍發現一些有趣的結果，以表 13 所示為例。

表 13　CBETA 類比關係詞嵌入演算結果

	樣本關係		推論關係		
編號	*a*	*b*	*x*（題目）	*y*（演算結果）	*y*（預設答案）
1	天臺	智顗	華嚴	惠光	法藏
2	慧遠	廬山	玄奘	義淨	長安
3	佛陀	世尊	觀音	觀世音	觀世音
4	文殊	智慧	阿難	多聞	多聞
5	泥犁	地獄	傍生	餓鬼	畜生道

資料來源：作者自行整理。

表 13 中我們基於不同的語義關係進行問答，包括詢問宗派創始者、主要傳教地、稱號、專長、翻譯等。例如第一題可理解為「天臺與智顗的關係，相當於華嚴與誰的關係？」演算的答案是「惠光」，而我們預設的正確答案是「法藏」。在五個問答題中，三、四兩題關於稱號和專長的問題得到正確解答。第一題演算結果得到「惠光」，惠光雖曾注釋過《華嚴經》，但一般將他歸為律宗大師。第二題主要傳教地得到的不是預期中的地點，而是人名「義淨」。然而觀察排名序列，Top2 以下的選項為 { 玉華宮 , 玉華 , 玉華殿 , 召延 , 法王城 , 弘福寺 , 閩嶺 , 弘法院 }，多是地點，且玉華宮是唐太宗首次召見玄奘的地方，該地與弘福寺、弘法院等處都是玄奘後來的譯經地，這些答案十分接近正確。至於第五題「泥犁」是三惡道中「地獄」的音譯，「傍生」則是三惡道中畜生道的直譯，但詞嵌入演算結果給的答案依序是 { 餓鬼 , 鬼界 , 畜生 , 畜生道 }，也是第三個答案才正確。由此看來，演算的結果雖不如預期但仍有參考價值。我們必須重新設計一組驗證資料集來調校超參數，才有可能得到堪用於詞類比的詞嵌入模型；但驗證及測試資料集的設計有主觀性的問題，故此應用目前仍有此兩難的研究限制。

（四）找出各部類的核心概念

佛教不同學派之間對相同主題常有不同見解或著重處，我們很好奇在詞嵌入演算下會如何呈現這個現象，故表 14 以「真實」為例進行不同部類的演算，看看他們各自著重的「真實」為何？我們觀察到以下兩個重點：1. 獨有詞（即其他部類未出現的詞）常標示出此部類的核心義理。例如阿含部類的「緣起」，般若部類的「無生性」，密教部類的「法性」，瑜伽

部類的「圓成實性」，[5] 禪宗部類的「心性、本性」等皆為獨有詞，各自代表該部類探討真實時的重要概念。2. 足以見出大乘佛學重本體的特色。表中除了阿含部之外的所有部類，討論真實時明顯更偏重於「體性」的探討，而它們都是大乘佛學的代表。

表 14　從部類對比「真實」詞嵌入演算結果（依相關度高低排序）

阿含部類	般若部類	密教部類	瑜伽部類	禪宗部類
虛妄	實	實	真實義	如實
不異	真性	實義	虛妄	了知
顛倒	虛妄	真	真	實
出離	言說	法句	實諦	真正
真諦	第一義	聖性	四聖諦	究竟
緣起	虛妄	如如	實相	虛妄
實	無生性	真心	俗諦	通達
隨順	實義	真體	實性	本性
我見	世俗法	本空	圓成	證得
幻	俗諦	實語	真如	遠離
平等	虛誑	法性	勝義諦	殊勝
法說	世諦	世諦	真實性	心性
虛偽	推求	真實語	理門	無所得
邪見	隱覆	甚深	增益	本空
正見	名相	虛妄	勝義	正見
善說	實性	如實	徧計	參學
非	體性	一切智智	第一義	此
常見	虛	生滅法	聖智	不可思議
正行	了達	言教	五俗	言說
誑	勝義諦	勝義諦	顛倒	世間

資料來源：作者自行整理。

（五）藉以拓展研究廣度和深度

詞嵌入除了可以輔助觀察不同部類切入主題面向的差異，還可藉以拓展主題的廣度和深度。例如「煩惱」是佛教中常探討的主題。我們從表 15 中看到各部類「煩惱」的同義詞可細分為以下幾類，分別是 1. 以障礙來表示煩惱，如 { 三毒 , 二障 , 五蓋 , 結使 , 五欲 }；2. 強調煩惱是未來煩惱的

5　由於分詞標準的關係，表 14 的「圓成／實性」被切分為兩詞。

根源，如｛隨眠, 習氣, 業果｝；3. 根據煩惱的類型來表述，如｛根本, 現行, 隨煩惱｝；4. 根據煩惱的樣貌來表述，如｛穢污, 麁重, 罪垢, 客塵, 稠林, 繫縛, 纏, 漏｝；5. 根據煩惱的原因來表述，如｛無明, 迷理, 智障, 思惑, 我慢, 掉悔｝；及 6. 根據煩惱的本質來表述，如｛智慧, 菩提｝等等。其中各細類下的每一概念都是佛教的專門術語，可再深入挖掘其意義。故藉著詞嵌入詞表的統整及提示，實有助於拓展佛學研究的深度及廣度。

表 15　各部類「煩惱」詞嵌入演算結果（依相關度高低排序）

阿含部類	般若部類	密教部類	律部類	瑜伽部類	禪宗部類
斷除	結使	隨煩惱	結使	煩	無明
愛欲	習氣	煩惱障	惑	惑	習氣
結使	三毒	結使	欲界	煩惱障	障
結縛	麤重	三毒	無明	惑障	三毒
貪欲	永斷	隨眠	三毒	二障	結使
永	惑	無明	結縛	永離	客塵
欲愛	折薄	集諦	渴愛	隨眠	斷除
漏	貪欲	障	斷除	繫縛	塵勞
穢污	纏	思惑	貪欲	智障	瞋恚
無色愛	繫縛	結縛	三界	結使	永斷
疑惑	麁惑	二障	生死	障染	繫縛
有漏	現行	罪垢	我慢	貪愛	增長
障礙	九十八	三障	滅除	三毒	貪愛
除	煩惱障	業障	五蓋	永	根本
五蓋	隨眠	業苦	欲貪	斷	菩提
結	永	染污	隨煩惱	染污法	貪欲
貪愛	斷除	覆蔽	惑業	障	智慧
掉悔	業果	重障	越度	隨煩惱	稠林
愛結	麁重	惑業	垢	永斷	五欲
永斷	迷理	現行	隨眠	永害	攀緣

資料來源：作者自行整理。

註：障礙義：粉色；根源義：藍色；類型：棕色；樣貌：綠色；原因：紫色；本質：黃色。

（六）用於驗證傳統研究結果

詞嵌入也可用於與基於語料庫所做的歷時語義研究對比。以高婉瑜（2014）「一旦」研究為例，他指出「一旦」這個詞的意義隨著時代演變而有所不同，最早是時間短語「一個早晨」，後來轉喻為「某一天」，再

轉變為表未定的時間副詞「突然」，最後才有假設連詞「如果」的意義。他特別指出「如果」一義出現的很晚，至少在《大正藏》中「一旦」都還沒有這個意義。根據詞嵌入演算結果，我們發現「一旦」的名詞語意「一朝」在魏晉南北朝就出現了。由於這個名詞語義的同義詞很少，故在此也看出詞嵌入演算的局限，表 16 第一欄中多數是「一旦」的搭配詞而非同義詞，稍後我們會談到其原因。到了隋唐五代，「倏、奄忽、溘」已出現在詞表中，可見時間副詞「突然」的語義已經出現。直到宋元明清時期，「一旦」作為時間短語和時間副詞的情況變得更常見，故「一朝」排在第一位，「忽爾、奄忽、忽、忽然、倏爾」等副詞都排在 Top10 內。「如果」之義則始終未出現，符合高文所說《大正藏》中的「一旦」還沒有轉變為假設連詞。由此，詞嵌入演算結果驗證了高文提出的結論。未來我們似乎也可以反過來，用詞嵌入詞表所呈現的結果來推估詞彙語義的變遷過程。

表 16　從年代對比「一旦」詞嵌入演算結果（依相關度高低排序）

漢魏晉南北朝 67–580 CE	隋唐五代 581–1101 CE	宋元明清 1101–1911 CE
孤露	恃怙	一朝
飢寒	衰老	忽爾
履藉	悲號	奄忽
崩亡	離祖	大限
酷	榮華	終沒
戀	兒女	饑寒
罹	悲慟	忽
離別	倏	忽然
喪	骨肉	倏爾
恃怙	末限	玲瑯
喪失	死王	骨肉
憂思	刑剋	功名
室家	孤單	棄
喪亡	悲傷	歲月
捐棄	喪亡	殞
逃走	衣祿	悔悟
憐	奄忽	伶俜
親戚	恩愛	黽勉
少壯	溘	拮据
一朝	朝露	繼晷

資料來源：作者自行整理。

四、詞嵌入超參數設定之必要性與限制

B. Wang 等人（2019）指出我們很難準確理解嵌入空間如何編碼語言關係，因為詞嵌入獲取語言關係和屬性的方式與人類學習語言的方式大不相同。因此，為了生成有效解決特定任務的詞嵌入模型，必須針對該項任務進行超參數訓練。我們的實驗也驗證了調校參數的必要性。表 17 以禪宗部類的「真實」一詞為例，表 18 以隋唐五代的「凡夫」一詞為例，從表中可看出沿用其他研究的預設參數所運算出來的詞表，和經過我們調校後的最佳參數跑出來的詞表相比，內容完全不同，且後者明顯更具有參考價值。以此可以驗證調校過的超參數確實提升了正確率，同時也說明以整個漢文佛典 CBETA 為內容所訓練出的詞嵌入超參數，對於其中細分類，如時間分類、內容分類等均具有適用性。

事實上，本文利用 CBETA 訓練了 12 組詞嵌入模型（詳見表 9），為何不針對各別模型調校適合的超參數，而選擇以整個 CBETA 所訓練出的超參數套用到其子模型中呢？此為本研究的限制之一，因為各個分類下的最佳參數，都需要有一個同義詞集來驗證，我們目前無法為各分類建置驗證同義詞集，故只能以實例進行實驗後評估，如上述利用表 17 及表 18 來評估整體的超參數應具有通用性。這也是目前許多數位人文領域詞嵌入研究的共同限制，如 Hengchen 等人（2021）研究歷時性的國族觀念語義變遷，就說明他們僅能直接採用 scikit-learn package 中的預設超參數進行實驗，而且由於無法建立驗證集，故僅能利用過去的歷史研究結果與實驗結果對比來進行人工評估。

表 17　禪宗部類「真實」參數調校前後詞嵌入結果比較

預設參數				
Model	Dimension	Window	Epoch	產出詞表
Skip-Gram	100	10	10	禮誼 , 切當 , 真寔 , 談辯 , 究竟 , 招紆 , 那件 , 參貴 , 實 , 邊論 , 真叅 , 頂髻 , 契得 , 要頓 , 第患 , 體究 , 那根源 , 節貴 , 真履 , 要知
調校後參數				
Model	Dimension	Window	Epoch	產出詞表
CBOW	400	10	10	如實 , 了知 , 實 , 真正 , 究竟 , 虛妄 , 通達 , 本性 , 證得 , 遠離 , 殊勝 , 心性 , 無所得 , 本空 , 正見 , 參學 , 此 , 不可思議 , 言說 , 世間

資料來源：作者自行整理。

表 18　隋唐五代「凡夫」參數調校前後詞嵌入結果比較

預設參數				
Model	Dimension	Window	Epoch	產出詞表
Skip-Gram	100	10	10	二乘 , 嬰愚 , 邪信 , 寂否 , 染世 , 邪劣 , 相縱 , 凡愚 , 癡倒 , 迷夢 , 欣真 , 偏覺 , 真滅 , 聖人身 , 性相心 , 尊尚 , 十定性 , 凡迷 , 大僻 , 凡小
調校後參數				
Model	Dimension	Window	Epoch	產出詞表
CBOW	400	10	10	二乘 , 異生 , 凡愚 , 愚夫 , 凡小 , 凡失 , 聖人 , 二乘心 , 沈空 , 凡夫法 , 凡人 , 愚癡 , 凡 , 愚人 , 欣滅 , 愚痴 , 有學 , 迷倒 , 業繫 , 凡夫

資料來源：作者自行整理。

此外，我們還觀察到詞嵌入應用的其他限制，例如：在詞嵌入表中可能混入某些搭配詞（collocation），因為取詞範圍（window）若設定的較寬，緊鄰的搭配詞也可能被誤為是相似詞，如表 15 中的動詞 { 斷除 , 永 , 永斷 , 永害 } 都是搜尋詞「煩惱」的搭配動詞而非近／反義詞；但是較寬的取詞範圍也可能優化詞嵌入的表現。本文發現在 Window 5–20 之間，以 5–10 的取詞範圍最佳，故詞表中不可避免的會參雜部分搭配詞。此外，詞嵌入的前處理還包含分詞（word segmentation），中文分詞較西方語言分詞複雜（Y.-C. Wang, 2020），佛典的分詞還涉及音譯和外來術語，難度加深，因此分詞錯誤也可能在詞表中造成干擾，如表 17 未調校參數前的詞表中出現 { 那件 , 第患 , 那根源 } 等詞，都是明顯由分詞錯誤所生成。上述限制均有賴於詞表使用者憑藉本身的語言及專業知識自行判讀。

伍、結論

詞嵌入在現今數位人文研究之中已是一重要的研究工具，然而若要實際應用於人文研究，仍需對詞嵌入訓練之參數進行調校，方能產生有意義的研究結果。本論文建置適於詞嵌入訓練之佛典實驗資料集，並設計驗證實驗來調校超參數，再以測試實驗來評估結果，分別得到 0.87 和 0.86 的表現。我們據此訓練出不同的佛學子類詞嵌入模型，以產生對比詞表來探討詞嵌入於漢文佛典研究的應用面向。由本文研究結果來看，詞嵌入應用在人文領域中有許多有趣的發現，我們提出了六個可能應用於佛學研究的面向，包括：一、從年代和譯者看出譯詞的語義核心演變；二、用已知語

義界定新詞語義；三、以語義類比找出相關概念；四、找出各部派的核心概念；五、藉以拓展研究廣度和深度；六、用於驗證傳統研究結果。這些面向仍偏向詞彙語義的應用，如前言提及，詞嵌入在主題判定和文本分析上尚有發展的潛力，許多有趣的應用亟待開發，也是未來我們繼續努力的方向。[6]

致謝

本文承蒙科技部 109–110 年度「基於深度學習之漢文佛典知識擷取方法之研究」專題研究計畫（109-2221-E-655-001-MY2）補助支持，最初發表於 2021 年 12 月的數位典藏與數位人文國際會議。感謝匿名審查委員提供修改建議，謹此致謝。

6 本文相關程式已公開在 GitHub 網站（https://github.com/DILA-edu/cbeta-word2vec）。

參考文獻

Ganlantree（2007 年 10 月 26 日）。哈工大《同義詞詞林》共用版的若干改進〔部落格文章〕。取自 https://blog.csdn.net/ganlantree/article/details/1845788

亢世勇編（2015）。*新編同義詞詞林*。上海：上海辭書出版社。

王冰（2011）。三十年來國內漢譯佛經詞彙研究述評。*華夏文化論壇*，*6*，169-174。

朱慶之（2019）。從平行梵本看支譯《維摩詰經．菩薩品》所謂「『是』後置特殊判斷句」的真實句法語義結構。*佛光學報*，*5*(2)，39-76。

辛島靜志（2007）。早期漢譯佛教經典所依據的語言（徐文堪譯）。在四川大學漢語史研究所、四川大學中國俗文化研究所編，*漢語史研究集刊*（第十輯，頁 293-305）。成都：巴蜀書社。

李維琦（2003）。考釋佛經中疑難詞語例說。*湖南師範大學社會科學學報*，*32*(4)，121-125。

林昆賢、蔡俊明（2019）。基於深度學習的自然語言處理中預訓練 Word2Vec 模型的研究。*國教新知*，*66*(1)，15-31。

竺家寧（1998）。認識佛經的一條新途徑：談談「佛經語言學」。*香光莊嚴*，*55*，6-13。

竺家寧（2006）。佛經語言研究綜述——詞彙篇。*佛教圖書館館刊*，*44*，66-86。

侯坤宏、卓遵宏（2014）。中華電子佛典協會（CBETA）、數位人文（1998 年 -，44 歲 -）。在*六十感恩紀——惠敏法師訪談錄*（第三篇第七章，頁 249-314）。臺北：國史館。

高婉瑜（2014）。漢文佛典「一旦」的詞類與演變問題。在漢譯佛典語言研究編委會編，*漢譯佛典語言研究*（頁 93-100）。北京：語文出版社。

陳克威（2020）。*基於 Word2vec 的學術論文推薦系統*（未出版之碩士論文）。國立東華大學資訊工程學系，花蓮。

陳秀蘭（2018）。*基於梵漢對勘的魏晉南北朝佛經詞彙語法研究*。上海：復旦大學出版社。

陳思澄、洪孝宗、陳柏琳（2015）。*使用詞向量表示與概念資訊於中文大詞彙連續語音辨識之語言模型調適*。論文發表於 The 2015 Conference on Computational Linguistics and Speech Processing。新竹，臺灣。

陳淑庭（2021）。*中古漢語毛皮骨血類雙音組合研究*（未出版之碩士論文）。東海大學中國文學研究所，臺中。

陳鳳櫻（2021）。*中古漢譯佛經中達成、瞬成動詞與動前動後「已」的互動：以妙法蓮花經，阿含經為例*（未出版之碩士論文）。國立清華大學語言學研究所，新竹。

梅家駒、竺一鳴、高蘊琦、殷鴻翔編（1983）。*同義詞詞林*。上海：上海辭書出版社。

梁啟超（1998）。佛典之翻譯。在*佛學研究十八篇*（第十章）。臺北：臺灣中華書局。

張簡宇傑（2020）。*基於文字探勘之智慧工程文件摘要系統*（未出版之碩士論文）。國立清華大學工業工程與工程管理學系，新竹。

曾千蕙（2018）。*詞向量化方法之比較與應用*（未出版之碩士論文）。國立臺灣大學資訊管理學研究所，臺北。

曾元顯、許瑋倫、吳玟萱、古怡巧、陳學志（2020）。基於檢索方法的中文幽默對話系統之建置應用與評估。*圖書資訊學刊*，*18*(2)，73-101。doi:10.6182/jlis.202012_18(2).073

曾昭聰（2005）。中古佛經詞義抉要。*咸陽師範學院學報*，*20*(1)，69-72。

曾昭聰（2009）。佛典文獻詞彙研究的現狀與展望。*佛教圖書館館刊*，*50*，58-65。

黃泰霖、宋傳欽、姜志銘、譚克平、高桂惠（2019）。唐詩流通度之探討。*中國統計學報*，*57*，263-285。

詹麒正（2020）。*網路輿論熱度風向策略研究——以 PTT 八卦版為例*（未出版之碩士論文）。國立交通大學資訊學院資訊學程，新竹。

謝吉隆、楊苾淳（2018）。從「應變自然」到「社會應變」：以文字探勘方法檢視國內風災新聞的報導演變。*教育資料與圖書館學*，*55*，285-318。doi:10.6120/JoEMLS.201811_55(3).0022.RS.BM

羅文君（2019）。*基於深度學習之中文詞性標記研究與實現*（未出版之碩士論文）。國立臺灣大學工程科學及海洋工程學研究所，臺北。

釋慈怡編（1990）。*佛光大辭典*。北京：北京圖書館出版社。

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

Bjerva, J., & Praet, R. (2015). Word embeddings pointing the way for late antiquity. In K. Zervanou, M. van Erp, & B. Alex (Eds.), *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (pp. 53-57). Beijing, China: Association for Computational Linguistics. doi:10.18653/v1/W15-3708

Burns, P. J., Brofos, J. A., Li, K., Chaudhuri, P., & Dexter, J. P. (2021). Profiling of intertextuality in Latin literature using word embeddings. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, ... Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4900-4907). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.389

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 609-614). Minneapolis, MN: Association for Computational Linguistics. doi:10.18653/v1/N19-1061

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1489-1501). Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1141

Hayles, N. (2012). *How we think: Digital media and contemporary technogenesis*. Chicago, IL: University of Chicago Press.

Hengchen, S., Ros, R., Marjanen, J., & Tolonen, M. (2021). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*, *36*(Suppl. 2), ii109-ii126. doi:10.1093/llc/fqab032

Hu, C., & Zhao, B. (2021). Movie recommendation system based on deep learning. *International Core Journal of Engineering*, *7*(9), 289-296. doi:10.6919/ICJE.202109_7(9).0043

Kamlovskaya, E. (2018). Word embeddings in humanities. Retrieved from https://dhh.uni.lu/2018/12/11/word-embeddings-in-humanities/

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384-1397). Santa Fe, NM: Association for Computational Linguistics.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of Machine Learning Research*, *32*, 1188-1196.

Leavy, S., Wade, K., Meaney, G., & Greene, D. (2018). *Navigating literary text with word embeddings and semantic lexicons*. Paper presented at the Workshop on Computational Methods in the Humanities 2018. Luasanne, Switzerland.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211-225. doi:10.1162/tacl_a_00134

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26, pp. 3111-3119). New York, NY: Curran Associates.

Nissim, M., van Noord, R., & van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*, 487-497. doi:10.1162/coli_a_00379

Saeed, J. I. (2009). *Semantics* (3rd ed.). Oxford, UK: Wiley-Blackwell.

Schmidt, B. (2015, October 25). Vector space models for the digital humanities [Blog post]. Retrieved from http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 298-307). Lisbon, Portugal: Association for Computational Linguistics. doi:10.18653/v1/D15-1036

Sculley, D., & Pasanek, B. M. (2008). Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, *23*, 409-424. doi:10.1093/llc/fqn019

Sprugnoli, R., Passarotti, M., & Moretti, G. (2019). *Vir is to moderatus as mulier is to intemperans: Lemma embeddings for Latin*. Paper presented at Proceedings of the Sixth Italian Conference on Computational Linguistics. Bari, Italy.

Taylor, J. R. (2003). Near synonyms as co-extensive categories: "High" and "tall" revisited. *Language Sciences*, *25*, 263-284. doi:10.1016/S0388-0001(02)00018-9

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C.-C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, *8*, E19. doi:10.1017/ATSIP.2019.12

Wang, Y.-C. (2020). Word segmentation for classical Chinese Buddhist literature. *Journal of the Japanese Association for Digital Humanities*, *5*(2), 154-172. doi:10.17928/jjadh.5.2_154

Wohlgenannt, G., Chernyak, E., & Ilvovsky, D. (2016). Extracting social networks from literary text with word embedding tools. In E. Hinrichs, M. Hinrichs, & T. Trippel (Eds.), *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (pp. 18-25). Osaka, Japan: The COLING 2016 Organizing Committee.

數位典藏與數位人文 12 期（2023 年 10 月）
Journal of Digital Archives and Digital Humanities No. 12, pp. 43-82 (October, 2023)
DOI:10.6853/DADH.202310_(12).0003

Word Embedding in Buddhist Studies: On the Basis of Evaluation of Word Embedding Models

Shu-Ling Huang[1], Yu-Chun Wang[2,*]

Abstract

Word embedding is a method to automatically generate semantic vectors using corpora. This paper aims to explore the possible applications of word embedding in the Chinese Buddhist database (Comprehensive Buddhist Electronic Text Archive, CBETA). In order to obtain the best model of word embedding for Buddhist studies, we compile an experiment dataset using Chunjiang Zhuang's dictionary, Fubao Ding's dictionary, and Digital Dictionary of Buddhism dictionary; and designs two evaluation experiments for detecting synonyms and outlier words to obtain a baseline for model optimization. It is found that Word2vec CBOW, Dimension 400, Window 10, Epoch 10 is the best set of parameters. The validation score is 0.87 and the test score is 0.86. Accordingly, we categorize the CBETA corpus to train different models; and then run comparative word lists for different chronologies, translators, and schools of Buddhism; then further demonstrated the applications in real cases. The main contribution of this paper is threefold: 1. to build a synonym collection for word embedding used in the study of Chinese Buddhism; 2. to identify the hyper-parameters of word embedding for the study of Chinese Buddhism; 3. to explore and demonstrate the results of word embedding in the Chinese Buddhist studies, including the ability to determine the semantic core evolution of translated words, to define new words, to identify related concepts through semantic analogy, to identify the core concepts of each school, and to expand the scope of researches. In addition, it can be used to verify the results of traditional research.

Keywords: word embedding, Chinese Tripitaka (CBETA), Buddhist studies, word relations, word analogy

---

Manuscript received: March 7, 2022; Accepted: July 25, 2022
[1] PhD Student, Dharma Drum Institute of Liberal Arts.
[2] Assistant Professor, Dharma Drum Institute of Liberal Arts.
[*] Email: ycwang@dila.edu.tw

Extended Abstract

Word embedding is a method to generate semantic vectors using corpora automatically. This paper aims to explore the possible applications of word embedding in the Comprehensive Buddhist Electronic Text Archive (short for CBETA) builds the largest Chinese database of Buddhist texts in the world, and it has the features of (1) spanning over two thousand years, (2) including multiple schools, (3) possessing different translating versions of the same sutra, (4) written with multiple source languages, (5) rich in content of thought, literature, geography, linguistics, etc., and (6) having multiple styles of writing. In the past, the study of Buddhism has mainly focused on the local investigation of a particular person, book, or school. However, it has not been able to take full advantage of the CBETA texts for knowledge extraction or large-scale comparative study. In this paper, we aim to extract the correlation of vocabularies in Buddhist texts through the word embedding method, to help researchers expand or deepen their research connotations.

Many studies have pointed out that word embedding in humanities research requires subject-specific hyperparameter tuning in order to obtain meaningful results. Therefore, in the first part of our research, the hyperparameter tuning of word embedding is conducted; in the second part, the tuned hyperparameters are applied to the different types and scopes of the CBETA corpus for comparison. Detail steps are as follows.

The first part of the paper addresses how to perform an evaluation experiment to optimize hyperparameters. It comprises four procedures: create a synonym set, develop it into an experimental data set, design validation and testing questions, and fine-tune hyperparameters . Each step is automated and is described as follows.

1. Create a Synonym Set

We extract the synonym set from the open dictionaries of Buddhism, including Chunjiang Zhuang's dictionary, Fubao Ding's dictionary, and Digital Dictionary of Buddhism dictionary (DDB for short).

2. Develop an Experimental Data Set

Following the first step, we combine the three synonym sets mentioned

above by removing the duplicates, rare words, and monosyllabic words. The reason for deletion is that if the experimental set contains rare names of people, places, and objects, the evaluation may not be able to give scores because the words cannot be found. Similarly, monosyllabic words are too ambiguous for evaluation and therefore need to be deleted.

3. Design Validation and Testing Questions

We then establish 2,000 questions for detecting synonyms and 2,000 questions for detecting noise words by randomly selecting words from the experimental data set. Of which 1,000 questions were used as validation data, and the other 1,000 questions were used as test data. In the synonym detection experiment, each select question contains one word as the question, and the options include one synonym and three noise words. In this experiment, we want to test whether the system can pick the correct synonym of the question from the four options. In the noise detection experiment, each select question contains three synonyms and one noise word, and we want to test whether the system can pick out the noise word among the four options.

4. Fine-Tune Hyperparameters

To establish a baseline for model optimization with the evaluation questions, we use two word embedding methods of Word2vec to fine-tune the hyperparameter combinations among different models, dimensions, Epoch, Window, and min-count. The best performance in the total of 32 tuning exercises is the hyperparameter combinations of Model CBOW, Dimension 400, Epoch 10, Window 10, and Min-count 2. We achieve a validation accuracy of 0.869 and a test accuracy of 0.856 for the evaluation questions.

In the second part of the paper, we explain how to apply word embedding to Buddhist researches. Firstly we categorize the CBETA corpus to train different embedding models with the best hyperparameter set; and run comparison word lists for different chronologies, translators, and schools of Buddhism; then further demonstrate the applications in real cases. It can be summarized in the following six aspects.

(1) A comparative word list produced by word embedding can be used to observe the word meaning changing through different eras and translators. Take *fan2fu1* (凡夫, ordinary guy) as an example. According to the Sanskrit

of its origin, *fan2fu1* can be broken down into the words of bāla-pṛthag-jana (fool-dissimilarity-birth). In the early stage of Chinese translations of Buddhist texts, it was translated as *yu2ren2* (愚人 , fool), but later evolved into the meaning of *zhong4sheng1* (眾生 , sentient beings).

(2) With the comparative word list, the sense of new words can be defined by known words. Take the new word *duan1yan2* (端嚴) as an example. In the early days, it was similar to words such as *gao1da4* (高大 , tall) or *xian1ze2* (鮮澤, fresh) which are often used to modify objects; but later its synonyms became *duan1zheng4* (端正, upright) or *shu1miao4* (殊妙, exquisite), which is more often used to describe the appearance or ability of a person.

(3) The word embedding method is well-known for its semantic analogy ability. To inspect the performance of our data, we try to identify the related concepts in Buddhist texts. The results were not as good as expected in our analogous experiment. However, it is not entirely ineffective. For example, according to the relationship between Manjushri and Wisdom, the model successfully finds the relationship between Ananda and Most-learned.

(4) With the help of comparative word lists produced by word embedding, the core concepts of each school of Buddhism can be recognized. We compare five schools of Buddhism and find those unique words to a particular school are usually the core concepts of that school. For example, if we look up the word *zhen1shi2* (真 實 , Reality) in the five schools, we find the words *yuan2qi3* (緣 起 , conditional arising) in the āgama school, *wu2sheng1xing4* (無生性 , nirvana) in the Prajna school, *fa3xing4* (法性 , Dharma nature) in the Tantra school, *yuan2cheng2shi2xing4* (圓成實性 , perfectly accomplished nature of reality) in the Yoga school, and *xin1xing4* (心性, nature temperament) in the Chan school are unique words that represent essential concepts in the exploration of reality in that school.

(5) Embedding calculation can also help us expand the study's breadth and depth. For example, the synonyms of *fan2nao3* (煩 惱 , annoyance) show that they have multiple focuses, such as (a) expressing obstacles, e.g., *san1du2* (三 毒 , three poisons); (b) emphasizing annoyance is the source of future annoyance, e.g., *sui2mian2* (隨眠 , inclination); (c) describing the existing state the annoyance, e.g., *xian4xin*g2 (現行, occurring); (d) describing the appearance of annoyance, e.g., *hui4wu1* (穢污 , dirt); (e) clarifying the cause of annoyance, e.g., *wu2ming2* (無 明 , ignorance); and (f) denoting the essence of annoyance, e.g., *zhi4hui4* (智慧 , wisdom).

(6) Comparative word lists produced by word embedding can also help us verify the results of traditional researches which mostly rely on corpus analysis. For example, we confirmed that the usage of *yi1dan4* (一旦, someday, suddenly) in the Taisho Collection did not contain the most common use today of *yi1dan4*, i.e., the usage of hypothetical conjunction *if*, and which is Kuo Wan-Yu proposed in 2014 with corpus analysis method.

To sum up, the main contribution of this paper is threefold: (1) to build a synonym collection for word embedding used in the study of Chinese Buddhism; (2) to identify the hyperparameters of word embedding for the study of Chinese Buddhism; (3) to explore and demonstrate the results of word embedding in the Chinese Buddhist studies.