黃純敏（2019），『基於語彙鏈、格律斷詞方法以主題模型』，*中華民國資訊管理學報*，第二十六卷，第三期，頁 275-306。

基於語彙鏈、格律斷詞方法以主題模型進行古詩詞探勘與分析

黃純敏 *
國立雲林科技大學資訊管理系

摘要

鑒於傳統白話文的斷詞技術對於古詩往往有扞格不入的缺憾，本研究分別以基於語句鏈提出的 CSCP 與基於詩詞格律提出的 CCPF 斷詞法，擷取詩詞關鍵語彙。實驗素材取自中國詩詞全盛時期的唐宋詩詞，共計 204633 首詩，建構潛藏狄利克雷分配（LDA）的特徵詞詞袋，再依朝代分別執行 CSCP-LDA 及 CCPF-LDA，產出四種唐、宋朝主題模型。所有主題採用 Gibbs Sampling 進行估計和推斷，參數的選擇採用原始的最佳預設 α 和 β 的值，並以 Perplexity 的最低值訂出 LDA 主題數量 110 與迭代數 600。研究發現唐宋詩主題詞以一字詞及二字詞居多，CSCP 斷詞取決於語句鏈分佈率，斷出的字詞屬於鏈結頻率較高者，因此詞數較 CCPF 少。實驗也發現即使唐詩數量遠低於宋詩，然而唐詩不重複的主題字詞數量比宋詩還多，表示唐詩的用詞中較多元、活潑、多樣化；宋詩則趨向保守、謹慎，推測或許是因為宋朝各派思想主流，如佛、道、儒各家的思想，已逐漸融合，成為一統局面，因此用字較趨一致。實驗結果顯示 CSCP 所斷出的主題字詞正確率雖不如 CCPF，但是 UMass Topic Coherence 以及專家的評量，都顯示 CSCP-LDA 主題凝聚程度優於 CCPF-LDA，也與原詩文極度相關，說明利用分佈率斷詞的 CSCP-LDA 有較高的機會凸顯詩詞主題。

關鍵詞：主題模型、主題凝聚、古詩分類、詩詞格律、中文語句鏈、潛藏狄利克雷分配

* 本文通訊作者。電子郵件信箱：jennyhuang921@gmail.com
2018/09/28 投稿；2018/12/11 修訂；2019/02/18 接受

Huang, C.M. (2019), 'Exploring and analyzing ancient poetry with LDA topic model based on lexical chain and regulated verse segmentation methods', *Journal of Information Management*, Vol. 26, No. 3, pp. 275-306.

Exploring and Analyzing Ancient Poetry with LDA Topic Model Based on Lexical Chain and Regulated Verse Segmentation Methods

Chuen-Min Huang*
Department of Information Mauagement, National Yunlin University of Science and Technology

Abstract

Purpose ─ To investigate the feasibility of applying Latent Dirichlet Allocation (LDA) to a large number of Chinese ancient poems. This study explores word usages, the connotation of poems, the topical association between poems, and to observe the changes in words between different dynasties.

Design/methodology/approach ─ Since term segmentation techniques of vernacular are often inadequate for classical Chinese poetry, this study proposes two methods - Chinese Syntactic Chain Processing (CSCP) and the Chinese Classic Poetic Formula (CCPF), to process poetry segmentation. The experimental material was collected from "The Complete Tang Poetry" and "The Complete Song Poems", totaling 204,633 pieces, constructing the word bag of the LDA, and then implementing CSCP-LDA and CCPF-LDA, producing four kinds of Tang, Song Dynasty topic model. All topics were estimated and inferred using Gibbs Sampling, and the parameters were chosen using the preset values of $\alpha = 0.5$, $\beta = 0.1$. The perplexity value is calculated and determined 110 as the LDA topic number, 600 as the iteration number.

Findings ─ The research result observes that even though the number of Tang poetry is much less than that of Song poetry, the number of unique words identified is more than that of Song poetry, indicating that Tang poetry is more pluralistic, lively and diversified; Song poetry tends to be conservative and cautious. The experimental results show that

* Corresponding author. Email: jennyhuang921@gmail.com
2018/09/28 received; 2018/12/11 revised; 2019/02/18 accepted

the correct rate of segmented word by CSCP is not as good as CCPF, but the evaluation of UMass Topic Coherence and experts indicates that the generated poetic theme of CSCP-LDA is better than that of CCPF-LDA.

Research limitations/implications — Although the correct rate of word segmentation of CCPF is effective, it cannot be applied to non-regulated verse poems, and the CCPF-LDA classification effect is not as good as CSCP-LDA. Future research is recommended to explore ancient poetry classification by using other approach, such as deep neural network approach.

Practical implications — Although literati distinguish the poets and poetry in different styles, the rules of the distinction are not obvious and generally recognized; therefore, it is difficult to generate the rules for the classification of poetry from critics' comments or from poetic writing alone. To our best knowledge, the CSCP is the first of its kind to analyze ancient poetry not relying on the rules of classical Chinese regulated verse. This study is also the only one applying LDA to analyze the meaning of verses. With the promising results of topic modeling of this study suggests that the traditional vernacular word segmentation method and the removal of single character are not suitable for the word processing of ancient poetry.

Originality/value — We proposed a new poetry segmentation method. The fundamental idea of building CSCP is a bottom-up concatenating process based on the intensity and significance degree of distribution rate to extract meaningful descriptors from a string by processing the direct link and the inverted link in parallel. The process will be iterated until no concatenation can be found.

Keywords: topic model, topic coherence, classical poem classification, Chinese classic poetic formula, Chinese syntactic chain processing, LDA

壹、前言

隨著資訊技術的發展與普及，各類古典漢學文獻紛紛電子化，有些提供網頁瀏覽或查詢（羅鳳珠等 2007），此種作法有助於推播中華傳統文化，讓社會大眾對於浩如煙海的文學作品能隨時一窺堂奧。在古典文學中，詩詞起於先秦，盛於唐宋，其間湧現許多著名詩人，詩人所具備豐富的情感與對事物敏銳的感知能力，讓詩、詞之美悠遊於中國文壇，留下無數膾炙人口的詩文。因此在漢語文化中詩詞堪稱是中國文學作品中最具代表的瑰寶。過去有不少對詩人的風格或詩詞的詮釋與分類之研究，提供後人欣賞、比較的絕佳參考。雖然專家的研究對詩詞釋義的解析具有舉足輕重的影響力，礙於人力卻也只能分析少量詩詞。但觀《全唐詩》便收錄五萬餘首詩詞、兩千餘位詩人，《全宋詩》所收錄之詩作數量更是《全唐詩》四倍之多，在浩瀚的詩詞文獻中要具備互相引證發明的能力，以少數專家的努力，即使經年累月、皓首窮經，亦無法克竟全功。此外，有些研究或受限於分析者的觀點，或因當初可查考的作品有限，可能無法提供全面的分析。如論詩人風格，李白多被歸類於浪漫詩派代表，但我們觀其詩作也常有社會情懷、田園等風格。

詩，分為近體詩和古體詩，古體詩有四言、五言、七言、雜言（長短句皆有），近體詩分為五言絕句、五言律詩、七言絕句、七言律詩，還有極少數的六言詩。詩詞格律之基礎奠定於中國歷史中的唐宋兩朝（王力 2002），創作上要符合用韻、平仄、對仗、字數等四大要素；古體詩體制句數不限、格律自由，無嚴格限制。根據中國國家圖書館所提供之唐（Liddy 1990）、宋（Jim Barnett et al. 1990）詩分析系統，指出唐宋兩朝詩作多集中在五言與七言詩，如表 1。詞一文體由於作風自由，使得其格律結構相較於詩字數通常參差不齊、變化多端，導致歸納格律較為困難，基於近體詩有固定的詞數格律，相較古體詩，比較規律不複雜，因此本研究只針對五言、七言的絕句與律詩進行探討。為行文順暢，將以詩、詩作、詩詞、詩文等作為詩的表述。

表 1：《全唐詩》、《全宋詩》詩作總數

詩體	唐	宋
雜詩	8487	13192
三言詩	129	107
四言詩	866	4114
五言詩	15706	124550
六言詩	188	2444
七言詩	16607	48269

過去相關詩文斷詞研究多先藉由人工標記或人工校正所整理之詞義、平仄規則、專有名詞詞庫，如：以單一著名詩人近體詩為例，建立詞彙知識庫（楊哲青等 2004）；針對近體詩作之格律規範，結合領域專家之專業分析彙整成格律規則集（王迺仁等 2005）；依據詩作文體的語言特性，輔以詞譜、典故、人名、地名等專有名詞語料庫，建立詩詞語言詞彙切分之規則，再經由領域專家過濾、篩選，產生詞彙集（羅鳳珠 2005）。若遇到一詞有多個斷詞結果時，仍需經由人工輔助校正，過程十分繁瑣且耗時。

資料探勘領域中的文件分類技術研究已久，相較於近代研究，多數文獻雖可藉由網際網路搜尋而取得，但卻少有研究針對古詩之字詞、風格、主題及其作者寫作風格進行分類。目前對古詩詞的相關研究，多著重於填詞輔助，如（Tosa et al. 2008; 羅鳳珠 2011a）與自動生成詩詞短語（Wang et al. 2016; Yan 2016）、（劉文蔚 1932）、（蔣銳瀅等 2015）。其中（蔣銳瀅等 2015）在格律詩自動生成過程，藉由機率潛藏語意分析（Probability Latent Semantic Analysis; PLSA）進行詞彙集擴展，藉以加強詩詞的主題及意境。雖然該研究宣稱 PLSA 能有效挖掘主題相關字詞，但過去諸多研究指出 PLSA 在變數多時會出現過度擬合（overfitting）的現象，進而導致其模型不得不框限於所有的訓練資料，加上對新進文件無法處理，因此越來越多人利用潛藏狄利克雷分配（Latent Dirichlet Allocation; LDA）來克服該問題。LDA 是一種貝氏機率的延伸應用，其基本概念是將一篇文檔當成字詞集合，每篇文檔可視為多主題的組合體，每個主題為多個相關字詞組成，具有可作為監督（Quercia et al. 2012）與非監督（馮時等 2013）學習以及可彈性延展的優點，其實驗成效已受到學者的肯定與重視（Séaghdha & Korhonen 2014）。

諸多研究顯示資訊擷取的良窳與字詞處理有密切關係，因此斷詞技術是一項基本且重要的前置處理。中文字串不同於英文，其詞串間無空白區隔，所以在區分字串與詞串並不容易。而現今多數文字處理研究多集中在處理白話文，對韻文與文言文的處理相對較少。台灣處理白話文以使用中央研究院 CKIP 斷詞系統最為普遍，除了高效率外，其最大的特色之一為所斷出的字詞附有詞性標註，研究者可選擇所需的詞類，或剖析句法進行詞性合併，以擷取出更有意義之複合詞彙。然而，詩詞是一種特殊形式的文學，古人多透過提煉語言的精粹度，藉由嚴謹的格律要求讓鬆散的句子凝聚，同時將情感寄託於詩詞歌賦之中，其句子雖短，但語意卻格外緊密，能將不同時空的事物、場景聚攏或共現，因此以常用的 CKIP 斷詞及詞性標註，並不適用於古文的處理。

本研究分別以基於語句鏈（lexical chain）提出的「中文語句鏈處理」（Chinese Syntactic Chains Processing; CSCP）與基於詩詞格律提出的「中文詩詞格律」（Classical Chinese Poetry Forms; CCPF）斷詞法，擷取詩詞關鍵語彙，建構 LDA 的特徵詞詞袋，其中 CSCP 是目前唯一提出以非格律方式分析古詩詞者，本

研究也是唯一利用 LDA 特性剖析詩句內涵者。實驗素材取自中國詩詞全盛時期的唐宋詩詞，再依朝代分別執行 CSCP-LDA 及 CCPF-LDA，產出四種唐、宋朝主題模型，詩文分類意象圖如圖 1。實驗結果顯示 CSCP 所斷出的主題字詞，不如 CCPF 正確，但是 UMass Topic Coherence 以及專家的評量，都說明 CSCP 主題凝聚程度十分優異，也與原詩文極度相關，說明了利用分佈率選詞的 CSCP 比斷詞正確高的 CCPF 有較高的機會凸顯 LDA 的詩文主題。本研究架構與方法不僅可從大量文件取得合語意的語詞，而且可剖析詩句主題內涵、進行跨詩句間的主題關聯、與探勘不同朝代間的詩作用語遞嬗。使得文字探勘跨越時空，藉助現代科技挖掘古之幽情的各種面貌與關聯。

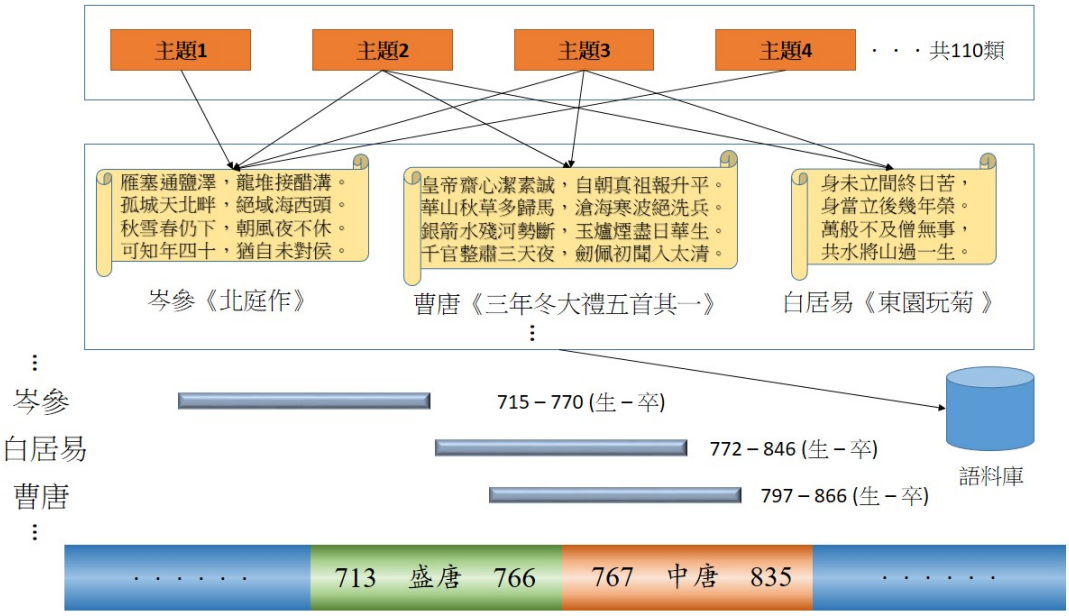

圖 1：詩文分類意象圖

貳、文獻探討

一、主題探勘

LDA 由 Blei 等人在 2003 年提出（Blei et al. 2003），其基本概念為，一篇文檔是由字詞所構成的一個集合，每篇文檔可以包含多個主題，詞與詞之間沒有順序及先後的關係，文檔中每一個詞都由其中的一個主題生成，這使得計算效率大為增加。其應用十分廣泛，如：機器學習、資訊檢索等相關領域。其發展的演進是

由 LSA（Deerwester et al. 1990）、PLSA（Hofmann 1999），最後再到 LDA。其中，LSA 藉由奇異值分解（singular value decomposition; SVD）將文本資料由高維度投影至低維度空間，然而 LSA 並非以統計模型為出發之方法，導致其計算成本與複雜度相對較高（Gao & Zhang 2003）。PLSA 將機率密度函式當作已觀察到的文件和字詞之間潛在語意關聯性的呈現方式，使用最大相似度估算法則（Maximum Likelihood）與最大期望演算法（Expectation-Maximization）推估潛在參數（latent variables）結果之機率模型。然而 PLSA 無法將機率分配給先前未出現的文件，導致其參數數量會隨文件數量線性成長，甚至產生過度擬合的現象，因此，後續研究多採用 LDA 來解決上述的問題。LDA 利用了狄利克雷先驗參數（Dirichlet Prior）分佈和多項（Multinomial）分佈的共軛（Conjugacy）性建構一貝氏模型（Bayesian model）來表達文件與主題關聯。

然而，後續相關研究（Mimno et al. 2011）發現該模型所產生的主題並不易解讀，與一般人的認知有顯著差距（Chang et al. 2009），為了處理這個問題，相關研究者結合外部資源來引導建模過程，但是這些資源都主要是針對特定實驗，缺乏廣泛應用的可能性（Chen et al. 2013）。因此，為了提高主題可解讀性的問題，有以知識為先驗基礎的主題模型（Knowledge-Based Topic Models; KBTM）它由使用者提供領域當中現有的知識來做為資料集，使得輸出的主題結果更為凝聚（coherent），聲稱考量文本兼容性（context compatibility）及主題凝聚性（topic coherence）。過去研究也認為如欲提昇 LDA 主題凝聚力，除考量字詞代表性外，主題數與 Dirichlet 先驗參數值的設定，必須審慎測試（Huang & Wu 2015）。在字詞處理方面，（Wallach 2006）以不同字詞組合測試 LDA 的結果，發現 Bigram 比 Unigram 有更高的主題正確率。不同於使用關聯規則萃取同時出現的字詞，（Huang & Wu 2015）認為緊鄰字詞（word concatenation）比鄰近字詞（word adjacency）更具有內容代表性，實驗以緊鄰字詞為考量，用一個月新聞量，比較中文單、複合詞組合用於 LDA 的成效，結果發現複合詞的主題凝聚力以及執行效率都優於單詞，但是 Perplexity 則較高。說明考量複合詞及字詞順序，有助提昇 LDA 的效果。

如何使用量化的評價方式衡量主題凝聚度是一個開放的問題。所謂主題凝聚，顧名思義是指在同一主題之下所有語詞表達意旨的集中程度，亦即這些語詞的語意具有高度相關性及可解釋性。諸多研究採用 Perplexity 值計算，但發現常與人的判斷相左，無法有效反應一個主題模型所產生的字詞語意是否足夠凝聚。Chang 等人（Chang et al. 2009）通過大量的實驗，發現主題模型雖然有好的 held-out likelihood，卻沒有好的語意表達，因此 Perplexity 值方法只能說明資料集與模型的符合程度。Mimno 等人（Mimno et al. 2011）提出一個 UMass Topic Coherence，做為評估主題品質更好的方法，此方法只仰賴於字詞在文件當中的共

現關係，並且不需要其他外部資源或是人工標註，主要考慮在主題當中高頻字詞的條件機率，該研究結果顯示，主題的凝聚分數越高，則與專家的標註有高度一致性，較高的主題凝聚分數表示主題有很高的質量且更具備可解讀性，然而研究也發現仍有部分不佳的主題字詞，獲得高的主題凝聚分數。Newman 等人（Newman et al. 2010）提出一種基於主題詞之間的成對逐點互信息 Pointwise mutual information（PMI），做為衡量主題凝聚的概念，其主要考量字詞之間的聯合機率分佈。Mimno 等人（Mimno et al. 2011）用對數條件概率代替 PMI，該方法記錄經常共同出現的詞，並在 Gibbs 採樣器中對新主題的分配，更新採樣之前和之後所有相關詞的計數。該研究宣稱比傳統 LDA 產出更一致的主題。

過去也有研究提出衡量主題可解釋性的各種方法，Xie 和 Xing（Xie & Xing 2013）邀請專家進行主題標註，討論主題標題及其所屬字詞是否相關，最後將所有標註者所註記為相關的單詞和候選詞的數量之間的比率作為主題模型的凝聚程度。Newman 等人（Newman et al. 2010）依據所觀察到的前 N 個主題詞的連貫性，而對主題進行評分。Chang 等人（Chang et al. 2009）提出了一種基於詞語入侵（word intrusion）的方法，將「入侵詞」隨機置入主題，並要求受測者識別入侵詞。該研究假設為：入侵詞在連貫主題中比在不連貫主題中更具識別性，因此可以透過測量入侵詞被識別出的容易程度，來估計主題的可解釋性。然而實驗結果與預期相反，他們發現Perplexity值與主題可解釋性竟然呈現負相關。儘管詞語入侵方法後來也被廣泛接受，然而該方法須高度依賴人為詮釋，俟後 Lau 等人（Lau et al. 2014）將該方法完全自動化，然而結果顯示自動化的處理成效略低於專家的表現。最終，所有這些方法都試圖評估主題的可解釋性。然而對於計算主題模型的語意可解釋性的最佳方法並沒有達成共識。

二、文本語彙處理

目前常見的三種詩詞斷詞方法主要分為詞庫斷詞、格律斷詞與統計斷詞。其中，詞庫斷詞法是藉由人工分類建立詞庫或專有名詞詞彙資料庫後，比對文件為取詞依據（羅鳳珠&曹偉政 2008），此方法執行速度快、容易，但由於這類詞庫多建立在人工的主觀分類，因此容易使詞彙的原貌與原義消失（王迺仁等 2005）；格律斷詞又可分為平仄對仗與句法規則，前者多需借助專家的格律知識建構斷詞模組（羅鳳珠等 1999），再從詩作中擷取出符合規則的名詞，而這類方法僅能判別詩作符合或不符合詩譜格律，因此有一字多韻的問題待解，後者則是藉由電腦運算結合詩詞規則（羅鳳珠 2011b），此方法雖能達到自動快速斷詞效果，但若要精確統計浩瀚詩作中的詞彙使用狀況，仍得依賴人工判斷，不易驗證；統計斷詞通常根據詩作中使用的字詞出現率為依據進行斷詞，但詩作中常利

用虛詞技巧作為修飾、連接之功能而不具特別意義，因此統計斷詞通常需仰賴其它演算法提升斷詞精確性（Yi et al. 2005; 王迺仁等 2005）使得運算成本相當可觀。在一般白話文的統計斷詞有些加入機器學習法：Maximum Entropy（Xue 2003）以及 Conditional Random Field（Tseng et al. 2005）等，這些學習演算法多以字元為單元，使用數種類似的特徵，如目前字元、加上前後各一字元、加上前後各兩字元等，來當作模型的屬性，此類演算法相對則較為簡易，但正確率不高。有些研究結合多種學習演算法，以加強斷詞效能，如在（Asahara et al. 2003）使用 HMM 結合 SVM，實驗結果顯示可提高斷詞正確性，然而處理過程十分複雜，也不易驗證。綜觀上述，具有豐富意義的詞彙是詩作風格分析的重要關鍵，而 LDA 的基本概念是建立主題與文件的分佈關係，進行 LDA 的文件取詞，應著重詞彙主題內涵的精確，然而古詩文的每個字都有其涵義，其字詞結構與白話文顯著不同。依此，取詞的代表性應足以影響主題凝聚力，亦是本研究的探索重點。

主題模型研究希望挖掘文本潛在主題，理想上，每個主題所包括的語彙應該是一系列語意相關的詞彙，此種想法和語句鏈頗為相似。語句鏈是指文件中具有相同意義或關聯的字詞所構成的集合，語彙之間無句法結構關係，卻能捕捉部分文本內聚結構，如聖誕節→聖誕老人→禮物→紅襪子。倘若將「霧霾」一詞放進語句鏈：「“霾害”、“廢氣”、“空汙”、“霧霾”」中，說明霾害是廢氣造成的空氣汙染現象；而在另一語句鏈：「“霾害”、“減碳”、“高碳排放”」，此霾害的語意則表達現實的環境條件。（Morris & Hirst 1991）最早引入語句鏈觀念建構文本關聯，該研究以少量資料（180 句）為實作對象，用人工一一標註每句的重要名詞、動詞、分詞，再參考 Roget’s Thesaurus，建立語彙關聯（同義詞、廣義詞、狹義詞等），提出依據鏈結的長度、同質指數、分佈率、共現率等計算權重。該研究發現語句鏈關係愈強，文本敘述愈相關，礙於當初 Roget’s Thesaurus 只有紙本，該研究並未實作。其後（Barzilay & Elhadad 1999）驗證（Morris & Hirst 1991）的理論，並針對其未處理一詞多義所產生語意混淆的缺失提出修正，以電子版 WordNet 為語意關聯條件設定的參考依據，再回溯語彙所在語句，擷取權重高的句子組成摘要。為協助人工判定語句鏈與主題的關聯度，該研究建構視覺化的語句鏈，以專家撰寫摘要為正確樣本進行評估，發現該研究產出的摘要，無論在召回率及精確率都高於 Microsoft Summarizer。

一般來說，建構語句鏈需選擇定義明確的詞彙庫，執行步驟有三：(1)篩選候選詞；(2)對每一個詞依據相關條件找出最合適的鏈；(3)如果找到，則置入該詞於其鏈，進行更新。一旦語句鏈明確，文章主題也就確定了，也可說語句鏈所表達的主題，決定了該文本的主題。透過語句鏈共同構成詞彙的上下文以提升文句的連貫性與內容凝聚力。過去的研究在篩選候選詞，由於英文沒有斷詞的考量，多

會先去除停用詞，再從文本中選擇鄰近相關字詞，而後進行關連計算。由於不考慮詞彙在文本的順序，就會出現：“the department chair couches offers”以及“the chair department offers couches”有同樣“department”和“chair”兩個關聯詞，其實是敘述不同主題。此外，停用詞有時候會扮演貫穿文句的重要角色，提早刪除，也常造成語言模型預測的失焦。本研究所提出的 CSCP 斷詞法基於語句鏈理念，回歸常人寫作用詞的習慣，亦即每個字元有其接續字元的機率，當蒐集足夠寫作的範例，即可預測接續字詞出現的機率。這也就是目前一般中文輸入法提供的字詞預測功能。有別於相關研究，本研究提出的語句鏈無須使用複雜的演算法，如 MEM（Chiong & Wei 2006）、HMM（Ageishi & Miura 2008），也不需大量範例訓練，僅用簡易字元相連機率，即使應用於單一文件，亦可萃取內文重要語詞。

三、詩詞格律

「格律」是指近體詩之創作需符合用韻、平仄、對仗、字數等四大要素。由於用韻、平仄、對仗與探討詩詞生成較為相關，需要藉由人工標記語意建立對應詞庫，才能用於後續資料探勘，字數規則較為單純簡單且不須人工事前作業，且容易撰寫演算法，作為詩詞斷詞的絕佳憑藉。因此本研究僅針對字數作為格律斷詞的依據。王迺仁與曾憲雄以詩作規則，針對五言與七言詩句之特性整理出切割樣板（王迺仁等 2005），如表 2。每一詩句套用樣板來切割詞彙，再根據贅字（stop word）資料庫刪除不必要的詞彙，接著刪除低詞頻者，最後再經由領域專家過濾、篩選，產生詞彙集。若遇到一詞有多個斷詞結果時，仍需經由人工輔助校正。

表 2：五言、七言切割樣板

詩體	切割樣板
五言	（2,3）、（2,2,1）、（2,1,2）
七言	（2,2,3）、（2,2,2,1）、（2, 2,1,2）

經檢視相關詩文斷詞研究，多仍需藉由人工引導或藉助專門詞庫，才能分割出具有意義的辭彙與建立相關之詞彙資訊，目前尚無任何研究提出無須仰賴專業詞庫或專家加註的自動斷詞方法。基於此，本研究根據（王迺仁等 2005）所制定格律斷詞（CCPF）規則，以長詞為優先斷詞，若長詞頻率未達門檻，則進一步分析雙字與單字的組合關係，毋須經過複雜的比對過程和人工驗證，即可自動切分重要詞彙。

參、研究方法

一、資料前處理與研究架構

在詩詞處理中，詞為最小有意義且可以自由使用的單位，有關中文字詞分析的相關研究，多使用中研院 CKIP 字詞小組所研發的中文斷詞系統。然而此斷詞系統適用於當代論述，對於古文詩詞則往往有扞格不入的缺憾。以唐朝張繼《楓橋夜泊》一詩為例，由圖 2 之斷詞結果可看出 CKIP 用於詩詞斷字雖具有標記字詞的能力，但有不少是零星、片段且無法正確切分，如「月落烏啼霜滿天」一句的切分方式，已失卻「月落」、「烏啼」、「霜滿天」之原意。Huang（Huang 2014）基於語句鏈理論，曾提出一種簡易、創新的單文件提取摘要的方法。該自動摘要方法無需詞庫，只考量字元前後的鏈結分佈率與次數，不斷合併單元詞，即可有效萃取重要敘述語及新詞，可用來協助註解圖片，實驗結果證實成效頗為優異。基於古詩詞字數簡短的特性，本研究修改該研究部分鏈結邏輯，使之適用於多文件的古詩斷詞，提出 CSCP 斷詞法。同時，本研究也基於過去學者所提出的格律斷詞規則（王迺仁等 2005），將格律斷詞程式化，透過詞頻擇選最適詞彙，勿需再經由專人輔助校正，提出 CCPF 斷詞法。

```
月(Na)  落烏(Na)  啼(VA)  霜(Na)  滿(Neqa)  天(Nf)  ，(COMMACATEGORY)
-----------------------------------------------------------------------
江楓(Nb)  漁火(Na)  對(P)  愁眠(VA)  。(PERIODCATEGORY)
-----------------------------------------------------------------------
姑蘇城(Nc)  外(Ncd)  寒山寺(Nc)  ，(COMMACATEGORY)
-----------------------------------------------------------------------
夜半(Nd)  鐘聲(Na)  到(P)  客船(Na)  。(PERIODCATEGORY)
-----------------------------------------------------------------------
```

圖 2：張繼《楓橋夜泊》CKIP 斷詞結果範例

本研究架構如圖 3 所示，實驗採用中國詩詞的黃金時代唐、宋兩朝之詩詞，包含初唐、盛唐、中唐、晚唐與南宋、北宋。由於唐宋詩集版本甚多，除了有文本抄錄的差異外，有些將同詩名及卷數合併，如：[橫吹曲辭。前出塞九首]，因計算方式不同，可能認定一首或九首的差異。在資料蒐集過程中，本實驗發現古代詩詞創作常有方言詞與古語詞等古字，在中國國家圖書館之系統中，這類古字多用部首拆解後以括號包起表示。然而，隨著詞彙的與時俱進，這類古字早已有新的樣貌甚至不再使用，僅有少部分仍保持原樣並被電子化，儘管如此，這類古字之編碼由於尚未普及導致實驗檢索困難，因此本實驗比對《同義詞詞林》之古字以及其它線上詩集資料庫之通用字，處理範例如表 3、表 4，其餘超出編碼的字

體，其詩作將不列入本實驗對象。經處理後，共刪除 499 首，將 204632 首詩作為研究素材，其中唐詩的數量僅占宋詩的 18.48%，細目如表 5。接著進行兩項萃取詩詞重要描述語的平行分項實驗：(1)以 CSCP 斷詞；(2)以 CCPF 斷詞。而後分別進行 LDA 主題分析，產出：唐 CSCP-LDA、唐 CCPF-LDA、宋 CSCP-LDA、宋 CCPF-LDA，藉以探討唐宋兩朝詩作之用詞風格及主題關聯，最後比較評估兩種斷詞方法的斷詞正確率、主題凝聚力。並依據實驗結果探索跨時代詩作的主題關聯、不同朝代間的詩作用語遞嬗。

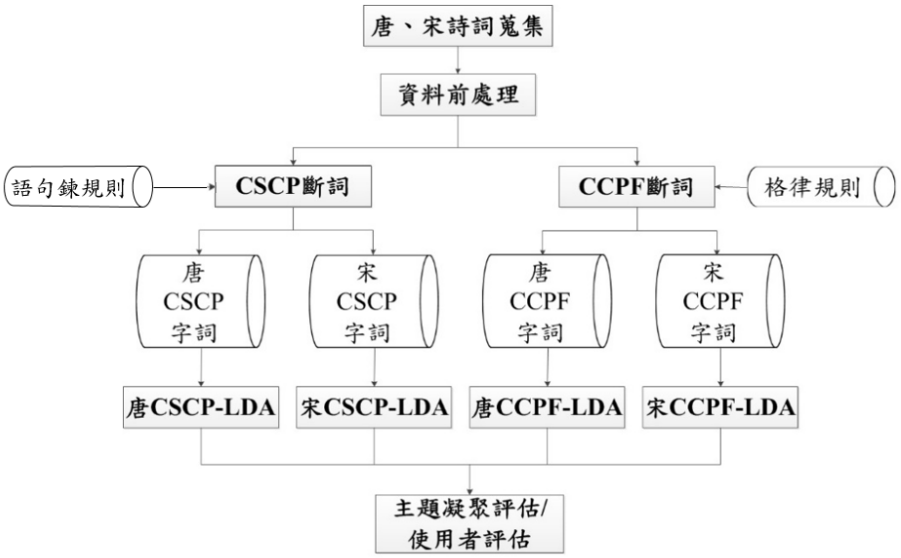

圖 3：研究架構圖

表 3：部分古字對照表

處理前	處理後
{言 寧}	譍
{革登}	鐙
{扌弃}	揳
{舟|竹/夆}	艟

表 4 部分古字通用字

處理前	處理後
{顧此=女}	嫮
{氵忄占}	聒
{犭术}	术
{骨孝}	髇

表 5：資料集數量統計

	唐詩		宋詩		
	五言	七言	五言	七言	加總
絕句	2859	9185	9556	68608	90208

律詩	12638	7240	38684	55862	114424
加總	15497	16425	48240	124470	204632

二、進行步驟

（一）中文語句鏈處理

本步驟以語句鏈方法萃取字串，首先將所有文件字詞串聯，將原本用於單文件圖片註解的概念（Huang & Chang 2013b），延伸為多文件特徵詞擷取。其處理概念為：將文件當中最小單位之字詞與前後字詞進行鏈結，利用分佈率統計每個相連語詞在所有連結字詞之比率，順向與逆向同時考量，此法不需參考詞庫，經不斷反覆鏈結，再考量子字串是否存廢，即可獲取足以代表該文章之重要代表語詞，步驟說明如下。

- 步驟 1：建立有向圖

以三首詩部分詩句為例，「姑蘇城外寒山寺，夜半鐘聲到客船。」、「夜行獨自寒山寺，雪徑泠泠金錫聲。」、「寒山吹笛喚春歸，遷客相看淚滿衣。」首先，依據標點符號斷句，去除重複字元並記錄該字元的連結關係，依據所有字元在內文中所出現的位置鏈結其前後字元，並於每一句句首、句尾各加入 S_Token 與 E_Token 作為空白字元串聯詩句。舉上述詩詞為例，「山」雖出現過三次，但僅保留單一端點（vertex），用以鏈結其在內文中所有出現位置之前後字元。其後將內文裡每個字元作為一個端點，再將各端點順向鏈結下一個字元，同時逆向鏈結上一個字元，建構出圖形的邊線（edge）。圖 4 實線為字元順向鏈結，虛線為逆向鏈結，最後產出一個有向圖形的資料結構圖。

- 步驟 2：計算平均分佈率與連線端點次數

此步驟主要計算有向圖的平均分佈率（Average Distribution Rate; ADR）。分佈率的定義為：某一文字（或詞句）之後出現的字的次數除以該文字（或詞句）之後出現的所有字的次數總和；或者，某一文字（或詞句）之前出現的字的次數除以該文字（或詞句）之前出現的所有字的次數總和。舉例來說，在整個文字檔案中，「我的」出現 4 次，「我們」出現 2 次，「我是」出現 3 次，「我家」出現 1 次，則「我的」的分佈率為 0.4（4/10），「我們」的分佈率為 0.2（2/10），「我是」的分佈率為 0.3（3/10），而「我家」的分佈率則為 0.1（1/10）。為了確定兩個端點是否可以連接，我們依據字詞順向與逆向個別分佈率計算出相連的比例，順向分佈率如公式(1)，逆向分佈率如公式(2)。D(i, j)表示端點 i 到端點 j 的順向鏈結分佈率；~D(p, q)表示端點 p 到端點 q 的逆向鏈結分佈率。Link(i,j)表示端點 i 到端點 j 的數量；$Diagram(Node_i)$為端點 i 所指向所有下一個端點之次數總和，$InvD(Node_p)$為端點 p 所指向所有上一個端點之次數總和。

$$D(i,j) = Link(i,j)/Diagram(Node_i) \quad (1)$$

$$\sim D(p,q) = Link(p,q)/InvD(Node_p) \quad (2)$$

首先計算順向鏈結分佈率，例如：假設端點「寒」指向端點「山」，且出現 3 次，並無其分支，因此可獲得實線[寒→山]分佈率為 1，在邊線上的數字[1, 3]，前者表示連接分佈率，後者表示端點連外分支總次數。接著，計算逆向分佈率，例如:端點「寺」逆向指向到端點「山」出現 2 次，亦無其他分支端點，因此獲得虛線之邊線[山←寺]，分佈率為 1，如圖 5。

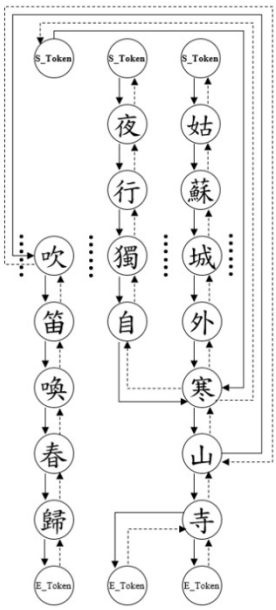

圖 4：有向圖

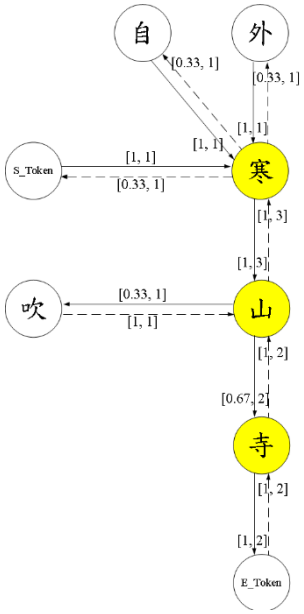

圖 5：分佈率說明圖

- 步驟 3：合併端點及建立新連結邊線

對於兩端點來說，利用公式(3)與公式(4)，計算出端點合併的標準。首先，各端點會以順向合併公式合併字元，如：「寒」及「山」兩個端點 Diagram($Node_i$)>1，因此「寒」及「山」進行合併，產生新的端點「寒山」，再與其對應之下個端點作出連結，如：「寒山→寺」。

$$D(i,j) \geq Diagram(Node_i) > 1 \quad (3)$$

$$\sim D(p,q) \geq InvD(Node_p) > 1 \quad (4)$$

如果只考量順向端點合併，當連結端點分支過多，將導致有意義的字詞因平均分佈率過低，而無法合併成為候選組合詞，因而同時考量逆向合併，將有助於

取得有意義的複合詞。如：「寒山寺」一詞，「寒」所能連接的字詞太過廣泛，因此「寒山」的分佈率會下降，假若「山寺」的分佈率較「寒山」來的高，就會將「寒山寺」斷成「寒」及「山寺」兩組字詞。而逆向合併如：「寺」一字逆向只有「寒山」一詞，分佈率高，故可擷取合併詞「寒山寺」，如圖 6。惟過去提出此方法僅用於單文件圖片註解（Huang & Chang 2013a），本計畫將其範圍擴充於古詩集斷詞。因此 CSCP 會產生較多的一字詞。

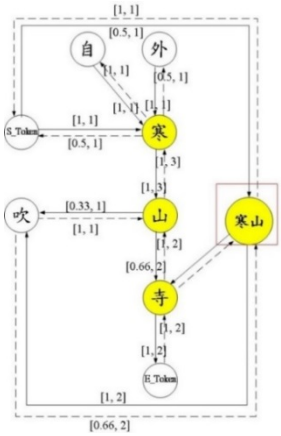

圖 6：合併端點及建立新連結邊線說明圖

- 步驟 4：反覆執行合併

為了萃取出更具意義的長字串，需重複執行「計算分佈率」與「合併」的步驟，直到沒有新的字詞端點可合併為止。在上述例子當中，先合併一元的端點後，可以獲得字串長度大於 1 的字詞，註記作為後處理參考，再依序不斷的反覆合併，每次合併均記錄合併前字詞狀態，此計算程序包括四個巢狀迴圈，一個 if 條件，共有 5 個循環複雜度（Cyclomatic Complexity），整個演算法的時間複雜度為 O(n4+n)=O(n4)，合併步驟虛擬碼如圖 7。

假若字串間有相同的字詞，以「寒山寺」為例，可分出「寒山」及「山寺」兩個字詞，利用公式。）為字詞估計意義的衡量，若「寒山」(a)的分佈率(f_a)為 1，而「山寺」(b)的分佈率(f_b)為 0.67，則以「寒山」做為斷詞端點。之後比對延伸的複合詞字串，如「寒山」延伸為「寒山寺」，需比較「寒山」及「寒山寺」的分佈率，若「寒山」的分佈率較「寒山寺」為低，則將「寒山寺」取代「寒山」，此方法用於合併多個字詞時處理方式亦同。

$$\text{remove}(a), \text{if } f_b \geq f_a$$

$$\text{remove}(b), \text{if } f_a \geq f_b \tag{5}$$

```
Input: poetry set
Output: term list
Procedure: iteration algorithm
(1)     for each poetry Pi in poetry_set do
(2)       for each sentence si in Pi do
(3)         for each term ti in sentence si
(4)           for each node ni in term ti do
(5)                 if node ni & ni+1≥ Diagram(Node_i) > 1 then
(6)                 Set ni to be equal to ni + ni+1
(7)           End for
(8)         End for
(9)       End for
(10)    End for
```

圖 7：合併步驟虛擬碼

（二）中文格律斷詞

此方法根據表 2 之詩詞格律斷字規則萃取詩作詞彙，由於（王迺仁等 2005）研究指出五言、七言詩句在句首、句中通常使用雙字詞，惟末三字則可能為 3 字、2+1 字、1+2 字的組合。為執行自動斷詞，本研究採用詞頻觀點，先判斷詩句末三字為 3 字詞的詞頻是否超過 1，若成立，則視為有意義的詞，如為五言詩，則採 2+3 之斷詞規則；七言詩則採 2+2+3 之斷詞規則，否則判斷末三字以 2+1、1+2 何者組合詞頻為高。以王維《鹿柴》部分詩句「不見人」為例，3 字組合為「不見人」，1+2 之組合為「不」「見人」， 2+1 為「不見」「人」，其中因「不見人」總詞頻未超過 1，因而分別計算上述四個字詞的詞頻，並以較高詞頻組為取詞依據，此例以 2+1 組合詞頻較高，故取「不見」「人」。此計算程序有二個迴圈，一個 if 條件，共有 3 個循環複雜度，演算法的時間複雜度為 O(n2+n)=O(n2)，斷詞虛擬碼如圖 8，斷詞結果正確範例如表 6、表 7 所示。

```
Input: poetry set
Output: poetry sentence list
Procedure: Rule based CCPF algorithm
(1)     for each poetry Pi in poetry_set do
(2)       for each sentence Si in Pi do
(3)         If Si length = 5 then
                Set W1, 2 & W3, 4, 5 & W3 & W4, 5 & W3, 4 & W5 in
                poetry sentence list
(4)         Else
                Set W1, 2 & W3, 4 & W5, 6, 7 & W5 &
                W6, 7 & W5, 6 & W7 in poetry sentence list
(5)       End for
(6)     End for
```

圖 8：格律斷詞虛擬碼

表 6：五言詩 CCPF 斷詞範例

五言詩規則	範例
2+3	白居易《問劉十九》 綠蟻 新醅酒 紅泥 小火爐 晚來 天欲雪 能飲 一杯無
2+2+1	王維《鹿柴》 空山 不見 人 但聞 人語 響 *返景 入深 林 復照 青苔 上
2+1+2	王維《相思》 紅豆 生 南國 春來 發 幾枝 願君 多 采擷 此物 最 相思

表 7：七言詩 CCPF 斷詞範例

七言詩規則	範例
2+2+3	杜牧《遣懷》 落魄 江湖 載酒行 楚腰 纖細 掌中輕 十年 一覺 揚州夢 贏得 青樓 薄倖名
2+2+2+1	韋莊《金陵圖》 江雨 霏霏 江草 齊 六朝 如夢 鳥空 啼 無情 最是 臺城 柳 依舊 煙籠 十里 隄
2+2+1+2	李益《夜上受降城聞笛》 迴樂 峰前 沙 似雪 受降 城外 月 如霜 不知 何處 吹 蘆管 一夜 征人 盡 望鄉

*說明：有不少詩詞專家將「返景入深林」斷為「2，1，2」。此種上下句不一致的斷句，也發生在李商隱的〈登樂遊原〉:「夕陽無限好，只是近黃昏。」

（三）LDA 處理

在給定詩詞文件$M = \{d_1, \cdots d_n\}$，字詞集$W = \{W_1, \cdots W_n\}$，分別從前述 CSCP 及 CCPF 斷詞過程產生。詩詞文檔 M 的每篇詩文 d_t 都假設含有若干潛在主題 z_k，每個主題都以字詞 w_i 的分佈表示。亦即語料庫字詞集 W 中的每個單詞 w_i 都假設是由潛在主題 z_k 生成，該潛在主題 z_k 則是從 K 主題的文檔特定分佈所提取的。所有主題採用 Gibbs Sampling 進行估計和推斷，參數的選擇採用預設的α和β的值，分別是 α=0.5、β=0.1。先計算 Perplexity 值，並利用 Perplexity 的最低值來訂定 LDA 主題數量 K 與迭代次數 N。由於本研究比較唐宋兩代詩詞用詞遞嬗，因此斷完詞之後，區分為唐/宋 CSCP、唐/宋 CCPF 字詞集，接著個別進行 LDA 主題模型處理。唐/宋 CSCP/CCPF-LDA 如圖 9，LDA 符號說明如圖 10，執行步驟說明於後。

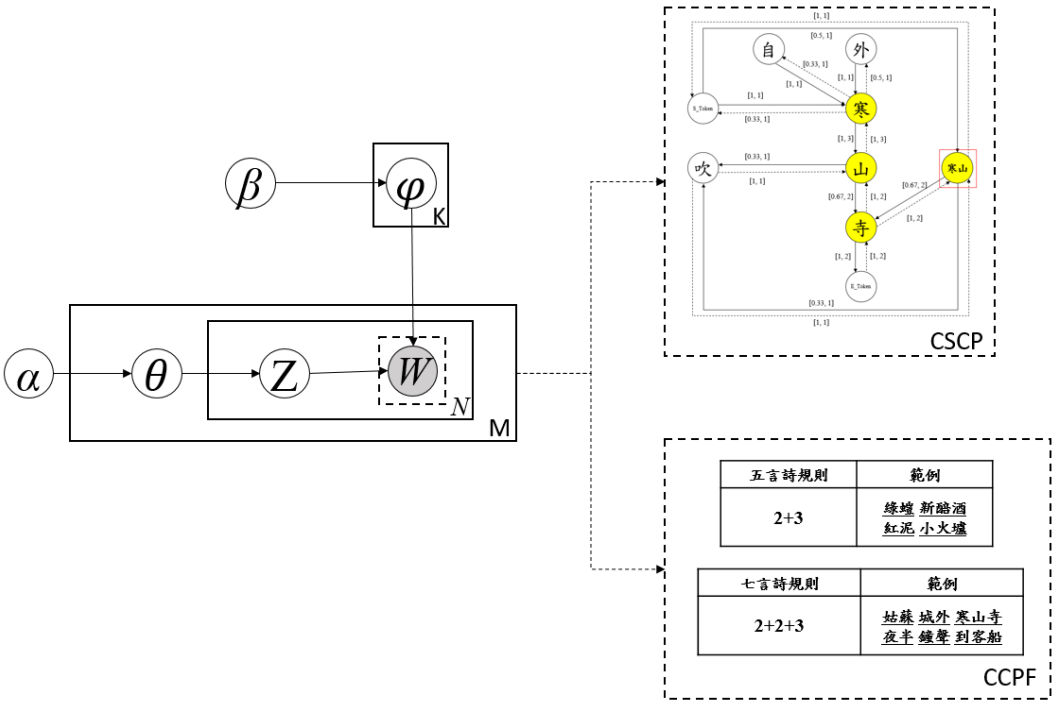

圖 9：唐/宋 CSCP/CCPF-LDA

M：詩文總數
K：主題總數
N：單篇詩詞中字詞總數
Z_{dn}: 詩文 d 中第 n 個單詞的主題
W_{dn}：某一字詞
α：每首詩主題分佈的 Dirichlet 之前的參數
β：每個主題詞在詩文分佈的 Dirichlet 之前的參數
θ_d：一首詩文 d 的主題分佈
φ_k：主題 k 的單詞分佈

圖 10：LDA 符號說明

詩文檔 M 表示為潛在主題的隨機混合，其中每個主題的特徵在於所有單詞的分佈。其機率總概率如公式(6)，LDA 生成一篇詩詞文檔聯合分佈生成過程為：

1. 從 Dirichlet 分佈 α 取樣生成詩文檔 d 的主題分佈 θ_d
2. 從主題 Z 的多項式分佈中取樣生成詩文檔 d 第 n 個從 CSCP/CCPF 斷出的字詞的主題
3. 從 Dirichlet 分佈 β 中取樣生成主題 Z_{dn} 的詞語分布$\varphi_{z_{dn}}$
4. 從詞語的多項式分佈$\varphi_{z_{dn}}$中採樣最終生成詞語 W_{dn}

5. 重複上述過程直到尋遍詩詞文件中的每一個從 CSCP/CCPF 斷出的詩文字詞

$$P(W,Z,\theta,\varphi,\alpha,\beta) = \prod_{k=1}^{K} P(\varphi_k|\beta) \prod_{d=1}^{M} P(\theta_d|\alpha) \prod_{n=1}^{N} P(Z_{d,n}|\theta_d) P(W_j|\varphi_{z_{dn}}) \quad (6)$$

（四）評估主題凝聚度

Perplexity 是一種平均分支係數的觀念，主要用來度量機率分布或機率模型的預測結果與樣本的契合程度，困惑度越低則契合越準確。使用於語言模型時，表示平均來說，我們預測下一個詞時有多少種選擇。舉個例子來說，對於一個長度為 N，由 a-z 這 26 個英文字母隨機組成的序列。由於這 26 個字母隨機出現，所以每個字母出現的概率是 1/26。於是，在看到另一個 perplexity 是 59 時，我們就可以直觀的理解為，平均情況下，這個語言模型預測下一個詞時，其認為有 59 個詞等可能可以作為下一個詞的合理選擇。Perplexity 值計算公式如公式(7)，其中 M 為文件數量，P(W_d)表示文件 d 中的詞 W 出現的機率。雖然過去有些研究發現以 Perplexity 評估主題凝聚度，Perplexity 然而其結果常與人的判斷相左，但是基於此方法對於資料集與模型符合程度的評量有一定的效度，本研究仍然採用作為衡量選擇主題數量的參考。此外，不少研究採用 UMass Topic Coherence，認在該度量中，計算來自主題的前 N 個代表性單詞的每兩個單詞組合的逐點互信息（PMI）分數。因此，主題的平均得分越高表明主題越連貫。UMass Topic Coherence 如公式(8)。其中 D(v)是有出現字詞 v 的文件次數，D(v, v')是字詞 v 與 v'同時出現在同一個文件的次數，$V^{(t)} = \left(v_1^{(t)}, \cdots, v_M^{(t)}\right)$是指在主題 t 當中有 M 個主題字詞，由於 PMI 分數是在詩文檔等級估算的，為了不使得 Log 為零，故添加了 1 以避免零的對數。此方法不需要其他外部資源或是人工標註，研究結果也顯示，主題的凝聚分數越高，則與專家的標註有高度一致性。

$$Perplexity = \exp\left\{-\frac{\sum_{d=1}^{M} logP(W_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (7)$$

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right)+1}{D\left(v_l^{(t)}\right)} \quad (8)$$

此外，基於過去研究亦顯示自動化的處理成效略低於專家的表現，然而對於計算主題模型的語意可解釋性的最佳方法並沒有達成共識。以人工進行主題凝聚度評估或主題標註是一件相當主觀且耗時耗力的任務，即使有相關經驗的專家參與，也會面臨意見不一的情況。由於本研究的詩詞的主題字詞多為一字詞，所含的詞意及意境也將成為衡量的基準，即使詩詞專家亦覺得進行主題標註有見解的主觀性與實質的困難度。因此本研究由詩詞專家針對 LDA 群聚的主題詞相關性進

行評量 Relevance Measure（RM）。我們隨機挑選 3 個主題，每個主題抽取唐、宋 CSCP/CCPF 出現機率前 5 的主題字詞，並就各類 LDA 取出的前 5 個字詞比對原出處詩詞，從中選擇符合較多主題字詞的 15 首不同的詩詞（共計 60 首詩），再請詩詞專家逐一評估個別主題字詞與詩詞的相關程度。最後分析唐/宋 CSCP-LDA, CCPF-LDA 四個主題模型的表現，以及人工與自動化處理的差異。

貳、實驗評估與討論

一、CSCP 及 CCPF 斷詞結果分析

由於宋詩的資料集數量為唐詩的 5.4 倍，詞數也呈現大致等比數量，如表 8。但數據也顯示唐、宋詩的 CSCP 詞數僅佔 CCPF 的 52%及 58%，但是詞頻則多出 1.2 倍左右（127%, 120%）。推測 CSCP 斷詞因取決於語句鏈分佈率，斷出的字詞多屬於鏈結頻率較高者，因此詞頻高、詞數少。統計結果也顯示 CSCP 有較高的標準差，意謂其詞頻歧異性很高，詞頻高者表示當代重要且常用的字詞。但是由於 CSCP 優先擇選分佈率較高的字詞，這表示高鏈結頻率的端點可能會因為端點本身連結太多端點，削弱被取出的機率，也同時增加取出下個端點與端點連結的可能，而造成本身被孤立成為一字詞。

為了解 CCPF 斷詞的成效，經隨機抽取唐、宋/五言、七言/絕句、律詩 8 種詩文末三字被斷為 3 字詞，以及 1+2, 2+1 各兩首，共 32 首，請詩詞專家逐一評定其正確率。專家認為只要字詞可適度解釋該詩意，即認為合用，因為特殊意涵字詞（如：「對愁眠」）或統稱意涵字詞（如：「對」、「愁眠」）並不影響詩中的涵義，若必須在個別字詞中選出其一，似乎嫌牽強；因為兩者都能用於解釋該詩文。因此作答是偏重所斷出的字詞「是否能成詞解釋該詩文」來考量。從可以看出以格律為考量的 CCPF，相較以語句鏈分佈率選詞的 CSCP 更能精準的斷出正確的詞，尤其是斷為 3 字的正確率可達到百分百，其錯誤率則發生在 1+2 或 2+1 組合。我們也將 CKIP 斷詞結果一併做比較，數據顯示 CKIP 的斷詞結果，相較之下，確實不佳。表 10 列出張繼《楓橋夜泊》以及方干《題寶林寺禪者壁》以三種斷詞方法之結果，斷詞結果可看出 CCPF 的正確率之高，則為三者之最；CSCP 和 CKIP 可斷出有意義的一字詞，但是 CKIP 的斷詞結果較無章法，正確率最低，也確實不適合作為古詩詞斷詞。

如前所述，CSCP 的實驗取順逆向分佈率較高者作為取詞依據，舉例來說，若「寒山」的順向分佈率較逆向分佈率高，則將順向分佈率做為「寒山」的代表分佈率，反之取逆向分佈率做為代表分佈率。若順向分佈率及逆向分佈率相等，則以正向詞為取詞代表。假若正負向分佈率皆為 1，意謂此複合詞字串僅出現在某特定詩詞中，如：「笊籬無柄」，其順逆向分佈率均為 1，詞頻過低，表示不具

特徵意義，予以排除。表 11 列出個別取詞詞數，從中可發現順向詞數與逆向詞數約為 6:4，意謂逆向取詞佔有相對的份量。若僅以順向分佈率作為 CSCP 的斷詞依據，將缺少將近四成的衡量基準，也將影響斷詞結果及成效。

表 8：CSCP 及 CCPF 字詞處理結果

詩詞數統計	CSCP 詞數	CCPF 詞數	CSCP 詞頻	CCPF 詞頻	CSCP 詞頻標準差	CCPF 詞頻標準差
唐詩	128,890	248,825	701,999	551,558	23.1	6.96
宋詩	581,696	998,940	3,511,667	2,924,629	41.35	17.03

表 9：CSCP/CCPF/CKIP 斷詞正確率

斷詞方法	CSCP	CCPF	CCPF	CKIP
末三字	-	1+2、2+1	3	-
正確率	61.76%	78.91%	100%	46.26%

表 10：CSCP/CCPF 末三字斷詞範例（含 CKIP 斷詞）

詩文	張繼《楓橋夜泊》		
斷詞方法	CSCP	CCPF	CKIP
斷詞結果	月落烏 啼 霜滿 天 江楓 漁火 對 愁眠 姑蘇城外 寒山寺 夜半 鐘聲 到 客船	月落 烏啼 霜滿天 江楓 漁火 對愁眠 姑蘇 城外 寒山寺 夜半 鐘聲 到客船	月 落烏 啼 霜 滿 天 江楓 漁火 對 愁眠 姑蘇城 外 寒山寺 夜半 鐘聲 到 客船
正確率	10/14=71.42%	12/12=100.00%	10/17=58.82%
詩文	方干《題寶林寺禪者壁 》		
斷詞方法	CSCP	CCPF	CKIP
斷詞結果	邃 巖 喬木 夏 藏寒 牀下 雲溪 枕 上看 臺殿 漸多 山 更重 却令 飛去 即 應難	邃巖 喬木 夏藏 寒 牀下 雲溪 枕上 看 臺殿 漸多 山更 重 却令 飛去 即應 難	邃巖 喬木 夏藏寒 牀 下雲溪枕 上 看 臺殿 漸 多山 更 重 却 令 飛去 即 應 難
正確率	11/17=64.71%	14/16=87.50%	6/18=33.33%

表 11：CSCP 斷詞順逆向字詞代表統計表

詩詞類別	CSCP 順向詞數	CSCP 逆向詞數
唐詩	82,695	46,195
宋詩	370,581	211,115

二、LDA 主題萃取結果字詞分析

由於實驗資料顯示各類 LDA 模型 Perplexity 值皆在主題數 80 後呈現穩定，其中 CSCP Perplexity 值在主題數 110 時達到最低點，因此本實驗選取主題數 110 做為主題數量設定。基於 LDA 主題詞分配特性，每首詩詞可分配至不同主題，如張繼《楓橋夜泊》在 CSCP-LDA 分配至 8 個主題，在 CCPF-LDA 分配至 4 個主題，茲列出所屬主題如表 12。該表只列出分配高的前 20 個字詞，從中可看出被分配的主題用語，除了框出的字詞出現在原詩文之外，主題之間有相近涵蓋的主題內容成分，如：「聲」「夜」、「冷」、「獨」、「霜」、「江」、「巫峽」等。可見藉由 LDA 主題模型，無須人工分類，即可群聚相關主題詩文。

表 12：張繼《楓橋夜泊》所屬主題

CSCP			
Topic 1	Topic 10	Topic 39	Topic 45
齊、西、低、泥、啼 迷、溪、青、北、蹄	聲、明、夜、情、聽 應、宿、生、半、東	楚、遠、臨、越、巴 江、蜀、千里、復、秦	明、望、到、發、日 曉、地、上、冷、凝
Topic 49	Topic 95	Topic 100	Topic 103
上、向、月、半、白、 天、臨、遙、映、樹	對、白、菊、落、秋、 晚、同、霜、黃、登	清、生、名、行、獨、 輕、聲、上、白、落	中、風、通、還、野、 山、近、閑、處、終

CCPF			
Topic 22	Topic 38	Topic 66	Topic 79
風雨、白雲、不肯、窗下 莫道、西南、愁殺、巫峽 溪邊、行處	猶自、花下、惆悵、春來 春風、少年、從今、笙歌 別有、宮女	珍重、殷勤、今日、分明 爲報、月落、由來、萬里 相看、燈前	歲月、月明、高僧、無一事 鐘聲、夜靜、爲問、蓮花 巴山、魚龍

為瞭解不同 LDA 模型主題字詞的特性及分佈情形，分別從四種 LDA 主題（共 110 個主題）擷取分佈率最高的 100 個字詞，共 11000*4 個詞，進行分析。如前所述，CSCP 優先擇選分佈率較高的字詞，而高鏈結頻率的端點，容易被孤立成為一字詞，因此資料顯示 CSCP 以一字詞的數量為多，而 CCPF 因為格律規則所限，絕大部分字詞都是二字詞。表 13 的數據顯示，唐宋詩主題詞確實以一字詞及二字詞為多。

表 13：各類 LDA 110 個主題之字詞詞數統計（每類 11000 個字詞）

斷詞類別	一個字	兩個字	三個字	四個字
Tang_CSCP	6353	4577	62	8
Song_CSCP	7695	3257	46	2
Tang_CCPF	676	9710	614	0
Song_CCPF	100	10602	298	0

除了觀察主題字詞詞數之外，對於字詞的使用唐宋是否有偏好或差異，也是一個十分值得探討的議題。經抽取各類 LDA 前 100 詞頻之主題字詞，發現唐詩數量雖然少於宋詩，但在詞頻高的主題字詞中，其使用不重複的字詞數量比宋詩還多，如圖 11。推測唐詩的用詞中可能較具新創、活潑、求變、多樣化；宋詩則趨向慎用、保守，這或許與宋朝各派思想主流，如佛、道、儒各家的思想，已逐漸融合，成為一統的局面，因此用字較趨一致。資料也顯示唐宋兩朝重複使用的字詞，從 CSCP 斷出的有 3010 個，比起兩朝個別獨有者還多，CCPF 也有 3501 個，比唐詩還多，表示唐詩所經常使用的字詞，有為數眾多仍然在宋詩也持續經常被使用。

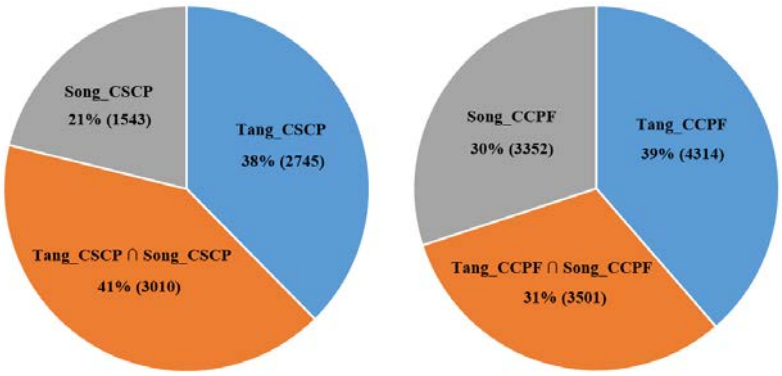

圖 11：各類 LDA 詞頻最高前 100 個不重複主題字詞統計表（每類 11000 個字詞）

顯示各類 LDA 前 20 詞頻之唐宋詩主題字詞統計，可以看出 CSCP 斷出的字詞中，「一」、「來」、「春」、「去」、「山」、「不」、「人」、「空」、「時」、「日」、「風」等字，同屬於唐宋 LDA 詞頻高者；同樣情形也出現在唐宋 CCPF 中的「今日」、「不知」、「春風」、「何處」、「萬里」、「人間」、「相逢」、「青山」等字詞。再比對唐 CSCP 前 20 高詞頻者如、「長」、「空」、「盡」、「上」、「新」、「在」、「入」、「秋」、「出」等字，在宋詩使用頻率雖然已下降，但仍在前 100 詞頻內。唐 CCPF 的「人」、「何事」、「惆悵」、「不」、「年年」、「今朝」、「秋風」、「無人」、「不

是」、「何人」、「明月」、「可憐」，雖然不在宋 CCPF 的前 20 詞頻，但下降幅度也有限。這或許也說明唐詩的璀璨，影響宋詩的創作，甚至延續至今。也意味著即使隨著朝代的改變，詩人慣用的字詞表述，仍然存有相當程度的重複性。

唐宋詩除了有為數不少的重複用詞之外，觀察其個別獨有用詞，也別具不同意義。唐詩獨有表示此類詞在宋詩已罕見，宋詩獨有字詞卻可能表示新生詞。經檢視所有 204632 首詞，整理出唐宋詩獨有主題字詞，如表 15。依詞頻順序列出其中 10 個。發現唐詩的「何處生春」、「艷歌」、「梨花」、「賤妾」、「歌舞人」，竟然在宋詩是罕見的。反觀宋詩的獨有字詞，確實比較中規中矩，由於可一窺兩朝的詩人的禮教束縛確有差異。此外，也發現唐詩獨有字詞「盥漱」一詞可追溯至禮記．內則：「子事父母，雞初鳴，咸盥漱。」，雖然在宋詩很罕用，但是被使用在紅樓夢．第四十九回：「寶玉此時歡喜非常，忙喚起人來，盥漱已畢。」。由此可看出字詞的使用，隨著年代的推進，某些字詞在沉浮幾世紀後，也可能再重新復出。

表 14：各類 LDA 前 20 詞頻之唐宋詩主題字詞

詩詞類別	CSCP	CCPF
唐詩	一、來、春、去、山 不、人、空、長、時 多、盡、日、上、新 在、入、風、秋、出	今日、不知、春風、何處、萬里 人間、人、何事、惆悵、不 年年、相逢、青山、今朝、秋風 無人、不是、何人、明月、可憐
宋詩	一、不、山、人、有 來、風、時、春、更 天、去、老、日、已 自、歸、今、猶、空	春風、人間、不知、平生、今日 萬里、歸來、梅花、當年、東風 相逢、西風、何處、青山、功名 風流、風雨、江湖、先生、千里

表 15：唐宋詩特有主題字詞

詩詞類別	CSCP	CCPF
唐詩	何處生春、豔歌、盥漱 城砧、本師、江島、梨花 楚色、亞相、前雲	紅兒貌、花宮、賤妾、殘鶯 爭那、盥漱、羅幌、存思 嬌歌、歌舞人
宋詩	愛吟詩、堯夫非是、故應 可人、祇今、寄聲、老成 聖賢、有子、彷彿	堯夫、造物、故應、愛吟詩 詩翁、胸次、祇應、祇今 端爲、等閒

三、唐宋詩詞語詞使用分析

歸納唐宋詩詞中，某些語詞使用率不同的原因可能有以下三種：

1. 戰爭帶來的痛苦：北宋自趙匡胤稱帝開始，就實行重文輕武政策，國力一直衰弱不振。從 1040 年起，至 1114 年，北宋與西夏發生五次戰爭。北宋與契丹（遼國）作戰，幾乎每戰必敗，簽定喪權辱國的條約。1127 年，北宋被女真金國消滅，宋徽宗第九子趙構在南京應天府即位，是為南宋高宗。1279 年崖山海戰，南宋軍被蒙古軍打敗，宋末帝趙昺隨陸秀夫背著跳海而死，南宋至此滅亡。就整個宋朝而言，除了岳飛等少數曾立下勝利的戰功之外，大多處在戰敗帶來的禍害和痛苦當中，人民家園喪失、親人離散；南宋開始，更有許多詩詞作品表達光復江北家園河山的渴望。
2. 詩詞風格的轉變：向來評論家一致認為「詩莊，詞媚」，唐朝以詩歌取勝，宋朝以宋詞聞名，在主流文體的感染之下，唐朝的詩和宋朝的詩詞在不知不覺中反映了「莊、媚」的不同特色。
3. 宋朝禮教十分嚴格：唐朝是豪放時代，尤其楊貴妃所帶動的女性開放風潮幾乎橫掃全國。宋朝女性受到嚴格的禮教束縛，一般閨秀不得拋頭露面，形同拘禁，例如歐陽修的「庭院深深深幾許」就是為此一社會現象抱屈，李清照特別表示感激。

以上三種社會背景和人生際遇，足以讓宋朝詩詞中，某些用語的使用率明顯比唐朝高出許多，例如：心存疑惑和不安定感：「不見」、「不是」、「不知」、「無人」「何處」、「何人」、「何事」、「誰知」。對於時空特別繫念和敏銳：「如今」、「今日」、「春風」、「東風」、「秋風」、「平生」、「千里」、「萬里」、「人間」、「青山」。抒發對於故人故國的思念，以及復國無望的感傷：「相逢」、「明月」、「可憐」、「故人」、「白髮」。

四、LDA 主題凝聚程度評估

（一）UMass Topic Coherence

圖 12 為利用 UMass Topic Coherence 評估主題凝聚程度結果，結果顯示唐詩無論以 CSCP 或 CCPF 斷詞都比宋詩的表現為佳，這表示唐詩所提取的主題字詞較能反映原詩詞所表述的意境及內容。而宋詩也可能因詩詞數量較多，稀釋了主題字詞的凝聚性，因此 Perplexity 值高於唐詩。

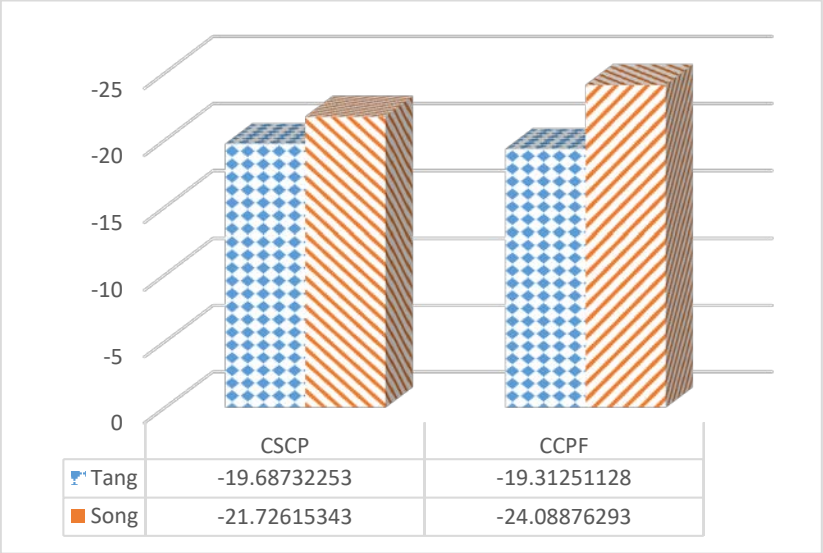

圖 12：UMass Topic Coherence 評估結果

（二）相關評量

為評估原詩文與 LDA 主題相關程度，我們從四種 LDA 隨機抽取 3 個主題，共 12 個主題。每一主題取前 5 個分佈機率較高的主題字詞，比對所有詩文，依序字詞出現在詩文擷取字詞最多，請受測詩詞專家就問卷所列出的詩詞原文與主題字詞，評估主題相關緊密程度，如表 16。評估結果如圖 13 顯示專家認為唐/宋詩的 CSCP 詩文與主題詞百分百相關，而唐/宋詩的 CCPF 相關程度分別是 87%與 73%。從 UMass Topic Coherence 以及專家的評量都說明 CSCP 表現優於 CCPF，雖然 CSCP 所斷出的主題字詞，不如 CCPF 正確，但是 CSCP 主題凝聚程度十分優異，也與原詩文極度相關，說明利用分佈率斷詞的 CSCP 有較高的機會凸顯詩詞主題。過去諸多研究顯示資訊擷取的良窳與字詞處理有密切關係，傳統白話文的斷詞方法以及捨棄一字詞的做法，證實並不適用於古詩詞的字詞結構分析。

表 16：主題字詞凝聚程度評估問卷範例表

詩詞名稱	詩詞內文	主題字詞
西陂	野渡扁舟自在浮，慣來江上不驚鷗。 一竿釣破滄浪月，喚入蘆花不點頭。	秋風、扁舟、滄浪、江上、春雨

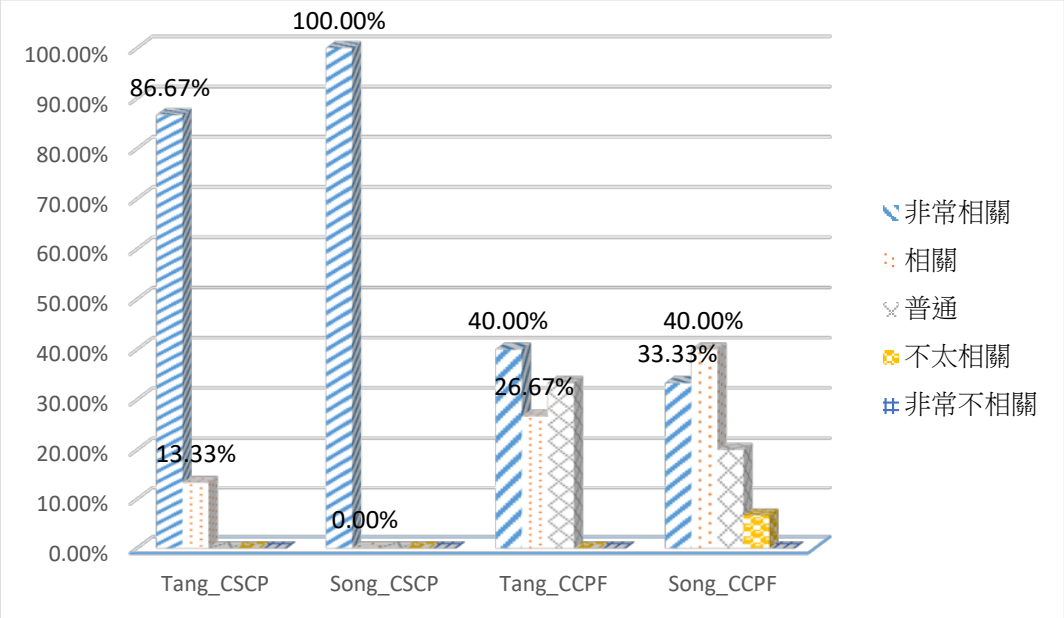

圖 13：相關評量評估結果

伍、結論與展望

如今資訊科技發達，網路資源隨處可得，需仰賴大量人力及時間的問題也可望逐漸解決。詩詞的創作往往著眼於詩人與各朝代間的史事、人文或藝術等主題。過去有不少對詩人的風格或詩詞的詮釋與分類之研究，提供後人欣賞、比較的絕佳參考。雖然專家的研究對詩詞釋義的解析具有舉足輕重的影響力，礙於人力卻也只能解析於萬一。目前對古詩詞的相關研究，多著重於填詞輔助與自動生成詩詞短語。此外，諸多研究顯示資訊擷取的良窳與字詞處理有密切關係，而現今多數斷詞處理多集中在處理白話文，對韻文與文言文的處理相對較少。由於古詩是一種特殊形式的文學，其句子雖短，但語意卻格外緊密，因此以常用的CKIP斷詞及詞性標註，過去並沒有被考慮在古詩詞的處理。經檢視相關詩文斷詞研究，多仍需藉由人工引導或藉助專門詞庫，才能分割出具有意義的辭彙與建立相關之詞彙資訊，目前尚無任何研究提出無需仰賴專業詞庫或專家加註的自動斷詞方法。本研究分別以 CSCP 與 CCPF 斷詞法，擷取詩詞關鍵語彙，建構 LDA 的特徵詞詞袋，其中 CSCP 是目前唯一以非格律方式分析古詩詞的斷詞方法，CCPF 則是根據格律規則，以長詞為優先斷詞，若長詞頻率未達門檻，則進一步分析雙字與單字的組合關係，毋須經過複雜的比對過程和人工驗證，即可自動切分重要詞彙。實驗結果顯示以格律為考量的 CCPF，相較以語句鏈分佈率選詞的 CSCP 更能精準的斷出正確的詞，尤其是斷為 3 字的正確率可達到百分百。單純就古詩文近

體詩斷詞而言，這是一項驚人的成果，至於末三字為 1+2 或 2+1 者，正確率也達 78.5%，未來探勘近體詩文者，可藉以做為最耗時間的資料前處理，減少大量人力成本。實驗也發現 CKIP 的斷詞結果較無章法，正確率最低，也確實不適合作為古詩詞斷詞。

本研究是目前唯一利用 LDA 特性探勘詩句內涵者。藉助 LDA 主題模型以統計概率分析角度切入的思考模式，分配詩文字詞產出對應的主題，再從主題追溯原詩文出處，無須耗費專家群過多時間及精力解析，也無須閱讀者從搜尋引擎以關鍵詞零散式的蒐羅整理，因此可以達到高效率分析以及公正之效果。本研究實驗素材取自中國詩詞全盛時期的唐宋詩文，經過 CSCP 及 CCPF 斷詞後，進行 LDA 主題模型分析，實驗結果顯示 CSCP 所斷出的主題字詞，正確率達 61.76%，雖仍不如 CCPF 的成效，然而 UMass Topic Coherence 以及專家的評量，都說明 CSCP-LDA 主題凝聚程度十分優異，也與原詩文極度相關，Perplexity 值相較格律斷詞的 CCPF 表現得更為平穩。這意味著利用分佈率選詞的 CSCP 比斷詞正確高的 CCPF 有較高的機會凸顯 LDA 的詩文主題，也顯示 CSCP 在古詩的處理更適合用於 LDA。

另外值得一提的是，透過 LDA 主題模型以群聚詩詞，從唐詩與宋詩的主題群聚結果，說明了不同朝代的詩詞有不同的用詞偏向，資料也顯示一首詩可分配多種主題，代表主題彼此間有某種程度相近涵蓋的內容。從主題字詞可看出詩所描繪的情境與朝代歷史背景有所關聯，經抽取各類 LDA 前 100 詞頻之主題字詞，發現唐詩數量雖然少於宋詩，但在詞頻高的主題字詞中，其使用不重複的字詞數量比宋詩還多，推測唐詩的用詞中可能較具新創、活潑、求變、多樣化；宋詩則趨向慎用、保守，因此用字較趨一致。唐宋兩朝各有其最常使用的字詞，推測除了受到詩人人生際遇影響外，也與當代社會背景息息相關。另外，唐宋兩朝也重複使用不少的字詞，表示唐詩所經常使用的字詞，有為數眾多仍然在宋詩也持續經常被使用。這或許說明唐詩的璀璨，影響宋詩的創作，也指出即使隨著朝代的改變，詩人慣用的字詞表述，或家戶喻曉的用語，仍然存有相當程度的重複性，甚至延續至今。本研究除了統計唐宋詩的重複用詞之外，也觀察其個別獨有用詞，唐詩獨有表示此類詞在宋詩已罕見，宋詩獨有字詞卻可能表示新生詞。研究發現唐詩獨有字詞多顯現歡樂、開放的氣氛，反觀宋詩的獨有字詞，則比較中規中矩，由於亦可一窺兩朝的詩人的禮教束縛確有差異。經歸納唐宋某些語詞使用率不同的原因可能有以下三種：1. 戰爭帶來的痛苦詩， 2. 詞風格的轉變， 3. 宋朝禮教十分嚴格，此三種社會背景和人生際遇，足以讓宋朝詩詞中，某些用語的使用率明顯比唐朝高出許多。

由於詩文為濃縮且精巧的語言，字字皆有其代表之意，與一般文件長篇大論的論述用詞不同，一首詩同字詞重複出現的頻率甚微，本研究將 CSCP 過去用於

新聞單文件斷詞以進行圖片註解的用途，轉為古詩集斷詞處理，取其斷詞結果作為 LDA 的詞袋。雖然其計算複雜度較高，但主題凝聚效果最佳。可見以語句鏈理論擷取順逆向語詞的做法，雖斷詞正確率不及 CCPF，但不影響詩文群聚結果，適用於古詩的主題群聚及後續詩詞探勘分析。未來實驗若提升硬體計算元件等級，可加速運算效率，也可考量採用深度學習 RNN-LSTM 將詩文正確斷詞結果讓機器學習，庶幾可提升斷詞正確率。至於 CCPF 基於格律判斷，雖然斷詞成效頗佳，但無法用於非格律的古體詩，且其 CCPF-LDA 分類效果不如 CSCP-LDA，未來進行古詩詞主題分析者，可自行衡酌採用何者為斷詞方法。

本研究實驗結果證實我們所提出的研究架構與方法不僅可從大量古詩集取得合語意的語詞，而且可剖析詩句主題內涵、進行跨詩句間的主題關聯、與探勘不同朝代間的詩作用語遞嬗。使得文字探勘跨越時空，呈現古之幽情的各種面貌與關聯。然而實驗也發現一字多型（如：黃、黃，帶、帶），因內碼不同，被視為不同字，也影響後續字詞的統計。建議未來研究可提出改善方案，也期許擴展至探勘更多朝代的詩文內涵，或嘗試進行大部頭書籍（如：四庫全書、古今圖書集成）的內文主題連結。

誌謝

本文接受行政院科技部專題研究計畫（MOST 105-2221-E-224-053）之補助研究經費，順利完成此篇著作之研究工作，僅此致謝。

參考文獻

王力（2002）*詩詞格律*，中華書局（香港）有限公司。

王迺仁、曾憲雄、楊哲青、蘇俊銘、羅鳳珠（2005），『詩風規則之研究－以唐朝近體詩為例』，*第二屆文學與資訊科技國際研討會*。

馮時、景珊、楊卓與、王大玲（2013），『基於 LDA 模型的中文微博話題意見領袖挖掘』，*東北大學學報：自然科學版*，第 34 卷，第 4 期，頁 490-494。

羅鳳珠、李元萍、曹偉政（1999），『中國古代詩詞格律自動檢索與教學系統』，*中文資訊學報*，第 13 卷，第 1 期，頁 36-43。

楊哲青、蘇俊銘、曾憲雄、羅鳳珠（2004），『詩作風格知識庫之研究－以蘇軾近體詩為例』，在羅鳳珠（主編），*語言、文學與資訊*，新竹：國立清華大學出版社，頁 263-295。

劉文蔚（1932），*詩學含英*，錦章圖書局。

蔣銳瀅、崔磊、何晶、周明、潘志庚（2015），『基於主題模型和統計機器翻譯方法的中文格律詩自動生成』，*電腦學報*，第 38 卷，第 12 期，頁 2426-2436。

羅鳳珠（2005），『詩詞語言詞彙切分與語意分類標記之系統設計與應用』，*第四屆數位典藏技術研討會*。

羅鳳珠（2011a），『以語言知識庫為基礎的智慧型作詩填詞輔助系統』，*教學科技與媒體*，95 期，頁 36-52。

羅鳳珠（2011b），『植基於中國詩詞語言特性所建構之語意概念分類體系研究』，*圖書與資訊學刊*，78 期，頁 63-86。

羅鳳珠、張智星、許介彥（2007），『植基於語意學及使用者認知觀點的資訊檢索系統設計：以全唐詩網站為例』，*第三屆文學與資訊科技國際研討會*，（日本學藝大學）。

羅鳳珠、曹偉政（2008），『唐宋詞單字領字研究』，*語言暨語言學*，第 9 卷，第 2 期，頁 189-220。

Ageishi, R. and Miura, T. (2008), 'Named entity recognition based on a Hidden Markov Model in part-of-speech tagging', Paper presented at the 2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT).

Asahara, M., Goh, C.L., Wang, X. and Matsumoto, Y. (2003), 'Combining Segmenter and Chunker for Chinese Word Segmentation', Paper presented at the Proceedings of Second SIGHAN Workshop on Chinese Language Processing, pp. 144-147

Barzilay, R. and Elhadad, M. (1999), 'Using lexical chains for text summarization', *Advances in automatic text summarization*, pp. 111-121.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research*, *3*(Jan), pp. 993-1022.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. and Blei, D. M. (2009), 'Reading tea leaves: How humans interpret topic models', Paper presented at the Advances in Neural Information Processing Systems Vancouver, British Columbia.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L. and Blei, D.M. (2009), 'Reading tea leaves: How humans interpret topic models', Paper presented at the Advances in neural information processing systems.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M. and Ghosh, R. (2013), 'Discovering coherent topics using general knowledge', Paper presented at the Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 209-218.

Chiong, R. and Wei, W. (2006), 'Named entity recognition using hybrid machine learning approach', Paper presented at the 2006 5th IEEE International Conference on Cognitive Informatics.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), 'Indexing by latent semantic analysis', *Journal of the American society for information science*, Vol. 41, No. 6, pp. 391-407.

Gao, J. and Zhang, J. (2003), '*Sparsification strategies in latent semantic indexing*', Paper presented at the Proceedings of the 2003 Text Mining Workshop.

Hofmann, T. (1999), 'Probabilistic latent semantic indexing', Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. Vol. 51, No. 2, pp. 211-218

Huang, C.-M. (2014), 'Applying A Lightweight Chinese Lexical Chain Processing In Web Image Annotation', Paper presented at the Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV).

Huang, C.-M. and Chang, Y.-J. (2013a), 'Applying a Lightweight Iterative Merging Chinese Segmentation in Web Image Annotation', *Lecture notes in computer science*, Vol. 7988, pp. 183-194.

Huang, C.-M., & Chang, Y.-J. (2013b), *Applying a lightweight iterative merging Chinese segmentation in web image annotation.* Paper presented at the International Workshop on Machine Learning and Data Mining in Pattern Recognition.

Huang, C.-M., & Wu, C.-Y. (2015), 'Effects of Word Assignment in LDA for News Topic Discovery', Paper presented at the The 4th International Congress on Big Data, New York, U.S.A.

Jim Barnett, K.K., Inderjeet Mani and Elaine Rich. (1990), Natural Language Processing, *Communication of the ACM*, Vol. 33, No. 8, pp. 50-71.

Lau, J.H., Newman, D. and Baldwin, T. (2014), 'Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality', Paper presented at the Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 530-539

Liddy, E.D. (1990), 'Anaphora in Natural Language Processing and Information retrieval', *Information Processing & Management*, Vol. 26, No. 1, pp. 39-52.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M. and McCallum, A. (2011), 'Optimizing semantic coherence in topic models', Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 262-272

Mimno, H.W., Talley, E., Leenders, M. and McCallum, A. (2011), 'Optimizing semantic coherence in topic models', Paper presented at the Proceedings of the 2011

Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), Edinburgh, UK.

Morris, J. and Hirst, G. (1991), 'Lexical cohesion computed by thesaural relations as an indicator of the structure of text', *Computational Linguistics*, Vol. 17, No. 1, pp. 21-48.

Newman, D., Lau, J.H., Grieser, K. and Baldwin, T. (2010), 'Automatic evaluation of topic coherence', Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100-108

Quercia, D., Askham, H. and Crowcroft, J. (2012), 'TweetLDA: supervised topic classification and link prediction in Twitter', Paper presented at the Proceedings of the 4th Annual ACM Web Science Conference. pp. 247-250

Séaghdha, D.O. and Korhonen, A. (2014), 'Probabilistic distributional semantics with latent variable models', *Computational linguistics*, Vol. 40, No. 3, pp. 587-631.

Tosa, N., Obara, H., & Minoh, M. (2008), 'Hitch haiku: An interactive supporting system for composing haiku poem', Paper presented at the International Conference on Entertainment Computing. Entertainment Computing-ICEC 2008 pp. 209-216

Tseng, H., Chang, P., Andrew, G., Jurafsky, D. and Manning, C. (2005), 'A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005', Paper presented at the Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing. pp. 168-171

Wallach, H. M. (2006), *Topic modeling: beyond bag-of-words*. Paper presented at the Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA.

Wang, Z., He, W., Wu, H., Wu, H., Li, W., Wang, H. and Chen, E. (2016), 'Chinese poetry generation with planning based neural network', *arXiv preprint arXiv: 1610.09889*.

Xie, P. & Xing, E.P. (2013), 'Integrating document clustering and topic modeling', *arXiv preprint arXiv: 1309. 6874*.

Xue, N. (2003), 'Chinese Word Segmentation as Character Tagging.' *International Journal of Computational Linguistics and Chinese*, Vol. 8, No. 1, pp. 29-48.

Yan, R. (2016), '*i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema.'*

Yi, Y., He, Z.-S., Li, L.-Y., Yu, T. and Yi, E. (2005), 'Advanced studies on traditional Chinese poetry style identification', Paper presented at the 2005 International Conference on Machine Learning and Cybernetics.