黃純敏、李亞哲、陳柏宏（2015），『以維基百科為基礎之中文縮寫詞與同義詞庫建構』，*中華民國資訊管理學報*，第二十二卷，第二期，頁 117-140。

以維基百科為基礎之中文縮寫詞與同義詞庫建構

黃純敏*
國立雲林科技大學資訊管理學系

李亞哲
國立雲林科技大學資訊管理學系

陳柏宏
國立雲林科技大學資訊管理學系

摘要

雖然過去對於辨識縮寫詞已有不少研究，但其研究範圍並未包含概括縮詞，此外，面對不斷增長及變化的詞彙，已成為資訊檢索及詞庫維護最大的問題。有別於過去以統計方式處理，本研究以維基百科的內文組成結構為基礎，提出數項創新且輕量級同義詞配對識別法。由於同義詞並沒有絕對客觀的標準答案可資核對，為驗證本研究所提出方法是否有效，我們進行兩階段包含主客觀方式評量。實驗結果顯示本研究所提出的方法，除了能有效萃取出縮寫詞、異形同義及同形異義詞之外，還能夠識別出過去研究無法解決的概括縮詞。在第一階段評量平均精確率為 72%、召回率 82%，其中縮寫詞的精確率高達 92%，概括縮詞的召回率為 90%。第二階段評量結果，使用者接受度亦達 91%。在效率方面，平均找出 1 組同義詞只需要 0.01 秒。

關鍵詞：同義詞、縮寫詞、概括縮詞、維基百科、同形異義詞

* 本文通訊作者。電子郵件信箱：jennyhuang921@gmail.com
2014/03/24 投稿；2014/09/16 修訂；2014/11/05 接受

Huang, C.M., Li, Y.C. and Chen, P.H. (2015), 'Wikipedia-based Chinese abbreviation and synonym construction', *Journal of Information Management*, Vol. 22, No. 2, pp. 117-140.

Wikipedia-based Chinese Abbreviation and Synonym Construction

Chuen-Min Huang*
Department of Information Management, National Yunlin University of Science & Technology

Ya-Che Li
Department of Information Management, National Yunlin University of Science & Technology

Po-Hung Chen
Department of Information Management, National Yunlin University of Science & Technology

Abstract

Purpose－A synonym can be any part of speech with the same or similar meaning of another word. Broadly speaking, it covers abbreviations in its scope. By convention, authors tend to indicate their writing with high artistic qualities by using numerous synonyms in context. Due to the interchangeable feature and the rampant growth of new usages, synonyms increase the difficulty of Natural Language Processing (NLP) and vocabulary maintenance. Unlike traditional approaches failed in its fallacy outcomes due to the adoption of statistical methods to determine synonyms, this study aims to construct a comprehensive synonym database via lightweight methods which would also take update issue into serious consideration.

Design/methodology/approach－The study proposes a research framework based on the analysis of contextual structure of Wikipedia. Due to the lack of a recognized correct corpus to assess synonyms, we adopted a two-stage evaluation including subjective and objective ways. Taken the virtue of continuous user involvement and

* Corresponding author. Email: jennyhuang921@gmail.com
2014/03/24 received; 2014/09/16 revised; 2014/11/05 accepted

suggestion, the constructed synonym database will be synchronously updated accordingly.

Findings－The proposed methods not only can correctly identify abbreviations, synonyms, and homographs, it can also successfully extract generalized terms with its multinomial sub-terms which had never done before. This finding indicates that a greater deployment of the comma algorithm can be undertaken to other customized application. The precision and recall rates of the first-stage evaluation are 72% and 82%, respectively. The user acceptance rate conducted in the second-stage reaching 91% was very promising. As for the efficiency evaluation, it took only 0.01 seconds to extract one set of synonyms from the system.

Research limitations/implications－This study mainly focused on formal descriptions extracted from Wikipedia. It is suggested that future research may consider applying to confusion word set or social media to fill the gap.

Practical implications－This paper contributes to automatic synonym construction research in several ways with a couple of practical implications. First, it demonstrates that a statistics-free, lightweight method can effectively generate a comprehensive coverage of synonyms. Second, this method can work with search engines to conduct big data analysis. Third, this study depicts that synonym construction can be portrayed in terms of ontology architecture to guarantee the sustainability of knowledge and the growth of literacy competencies of users.

Originality/value－Even though there have been many researches towards synonyms, none of them proposed the resolution to identify the generalized term with its multinomial sub-terms. This study is the first of its kind to solve this problem. In addition, words will be labeled with their name entity such as names of people, places, and organizations. Search results will be displayed based on the ontology architecture in which the word association can be clearly visualized.

Keywords: synonym, abbreviation, generalized term, Wikipedia, polysemy

壹、緒論

在中文文件中，字詞常以縮寫型態出現，例如：「台灣鐵路局」縮寫成「台鐵局」。此外，字詞的用法也會隨著時間、文化以及使用頻率的不同而不斷地增加及改變，例如：在過去未出現的社群網站「Facebook」，現今許多人會直稱「臉書」、「FB」來代表。這些不斷被創造的同義詞，以及高度「可縮寫性」之用法，雖然為現代人爭取了時效及便利性，也豐富了情感上的表達，但對於字詞處理則是一大難題，影響所及包括搜尋引擎的檢索效果都受到很大的考驗。

同義詞一般指稱異形同義詞，廣義來說亦可包括縮寫詞，為行文方便，本論文在不造成混淆的情況下，有時將使用同義詞涵蓋縮寫詞。縮寫詞又可細分三種類別：縮減法（Reduction）、刪去法（Elimination）、概括法（Generalization）。在過去的相關研究中，自動萃取縮寫詞的範疇大多侷限於縮寫詞中的縮減法及刪去法（Huang & Yang 2005），所提出的方法多運用自然語言處理（NLP）搭配最大熵法則（Maximum Entropy），找出上下文章中縮寫詞所對應的原形詞及運用最大共同子字元（longest common subsequence; LCS）演算法比對出多個字串當中，全部共同擁有的字元，例如：「雲科大」每一個字元都依序出現在原形詞「雲林科技大學」之中，依此來判斷出縮寫詞是否為原形詞的子字串。在異形同義詞方面，過去研究多參考字詞共現（Word Co-occurrence）率以找出其間是否為相關（Huang & Yang 2005；黃純敏等 2007），然而計算共現率十分仰賴資料集，如：一篇新聞中「馬總統」及「馬英九」，同時出現的機率極高，因此電腦將二者視為同義詞。至於縮寫詞中的概括法所指稱的概括縮詞如：「八國聯軍」由「義大利」、「美國」、「法國」、「奧匈帝國」、「日本」、「德國」、「英國」、「俄國」等八個國家所組成，由於無組成規則可循，目前為止尚無任何研究提出可自動辨別的方法。因此如何解決概括縮詞與其子詞配對，成為本研究主要挑戰的議題。

過去對於同義詞研究雖已有不少成果，但其結果仍存在許多問題：(1)雖可獲取高精確率，但因無法知悉是否找全所有縮寫詞配對，因此無法檢驗召回率成效；(2)因極度仰賴資料集特性，僅能做少量的資料測試；(3)共現率計算易產生誤判，例如：總統「馬英九」召見大提琴演奏家「馬友友」引發媒體大量的報導，導致這段時間在同一篇新聞中共同出現的頻率非常高，此時電腦有可能會將「馬英九」和「馬友友」誤認為是同義詞。此外，過去對同義詞的研究評估準則多由研究者自行定義，缺乏客觀評斷，因此如何設計客觀的評估正確率準則，也成為本研究另一項焦點議題。

鑒於維基百科（Wikipedia）為現今最大的線上百科知識寶庫，總共有 285 種語言版本，擁有 2,200 萬篇的文章，此百科全書強調內容自由、共同編輯，截至目

前維基百科上大約有 3 千萬名登記註冊用戶，其中有 10 萬名積極貢獻者長期參與編輯工作，而整個網站的總編輯次數更是超越 10 億次之多。使得維基百科一方面能快速增長內容，另一方面藉由群眾力量，能把潛在錯誤迅速糾正。因此成為許多研究的研究素材（Nguyen & Cao 2008; Fu et al. 2012）。

維基百科本文內容首先會列出敘述條目的定義及描述一個實體或事件源由，並在特殊名詞以引號加註或附上超連結，以引導到描述此字詞的文章，其作用在於提供使用者補充相關資訊，幫助其更了解此文章的內容意思及源由。為充分運用其資源，本研究以探究維基百科的內文組成結構為基礎，作為同義詞的萃取來源。此外，由於相同的字詞在不同的時間點、不同人對其認知、接受程度也不同，因此本研究將實驗結果開放外界進行接受度評估，希望藉由群眾智慧來不斷地修正以提升整體同義詞庫的品質。研究結果發現，在第一階段評量平均精確率為 72%、召回率 82%，其中縮寫詞的精確率高達 92%，概括縮詞的召回率為 90%。第二階段評量結果，使用者接受度亦達 91%。在效率方面，平均找出 1 組同義詞只需要 0.01 秒。

本文第貳節探討過去相關研究，第參節描述本研究的研究架構及流程，第肆節提出評估方法，檢視所提出的概念成效收益。最後於第伍節提出總結，並對往後研究提出未來展望。

貳、文獻探討

一、中文字詞處理

在字詞處理中，詞為最小有意義且可以自由使用的最小單位，但中文與英文文件的不同處，在於英文有明顯的空白區隔，能區分出詞與詞之間的邊界，而中文因字詞間無間隔，必須利用標點符號將文章標示成獨立的字串，再將字串轉換成詞的組合。因此正確的中文斷詞已成為自然語言處理的重要基礎（Zhang et al. 2002）。

常見的三種中文斷詞方法，主要分為詞庫斷詞、統計與混合斷詞法。下面簡介這三種方法：

1. 詞庫斷詞法：詞庫斷詞法是以既有詞庫比對文件為取詞依據，此法擁有高品質取詞水準，優點在於運算快速，但無法處理新詞。
2. 統計斷詞法：統計斷詞（Zhang et al. 2000），其概念源自詞頻（term frequency; TF），當詞頻達到某一門檻值時，假設該詞是有意義的，可分為 2-Gram、3-Gram…等，以此類推至 N-Gram。如「我喜歡吃葡萄」用 2-Gram 作斷詞，可斷出：「我喜」、「喜歡」、「歡吃」、「吃葡」、「葡萄」等詞組；3-Gram 作斷詞，可斷出：「我喜歡」、「喜歡吃」、「歡吃葡」、「吃葡萄」等詞組。此種以

統計的方式分析字詞出現的頻率，不須仰賴詞庫，對於新詞彙萃取有很大的幫助；但其缺點為 N 值的設立問題，N 值加大，可萃取更多罕見詞，但是過大的值，將會增加系統運算成本，此外，後續以人工加入過濾有效字詞的成本也相當可觀（Zhang 2000; Kitet al. 2003; Kang & Hwang 2006）。

3. 混和斷詞法：混合斷詞則事先利用詞庫把字詞過濾出，繼而利用統計法處理未斷出之字詞，可兼顧詞之品質與萃取出未知詞之優點，並減少 N-Gram 所需處理的字詞資料量，為許多研究人員採用（Hong 2009; Tsai 2010）。

字詞分析處理中，以詞庫斷詞法最常被使用，因為斷詞速度快，可提升系統之效能。近年來有關中文字詞分析相關之研究，多使用中央研究院 CKIP 詞庫小組所研發的中文斷詞系統。該斷詞系統具備辨識未知詞與附加詞類標記之功能。然礙於 CKIP 斷出之字詞十分瑣碎，必須再經過多重的篩選、過濾及合併，才能獲致重要關鍵字詞，因此許多研究皆利用 CKIP 斷詞後，再做進一步的分析處理，而其最大的特色之一就是擁有詞性的標記，透過詞性標記後的詞彙特性，可進行詞性合併，以擷取出具有意義之特徵詞彙（陳良駒&陳日鑫 2010; Lin et al. 2012）。

二、中文縮寫詞相關研究

在中文文件中，字詞常以縮寫型態出現，目的在於讓結構複雜、音節較多的字刪減或省略，得以方便的被使用。過去對於縮寫詞的研究，常以最大共同子字元或稱字詞共現法（co-occurrence）作為判斷縮寫詞是否為原形詞的子字串。最大共同子字元為動態規劃演算法，其主要目的為在兩個序列中，找出最長共同的部分，舉例來說：若 X 表示字串“accdb”；Y 為字串“acdc”；Z 表示字串“ac”而 L 為字串“acd”。雖然 Z 跟 L 皆為 X 跟 Y 的子字串，但是 L 的字串長度大於 Z，因而選定 L 為 X 與 Y 的最長共同子字串。過去學者在單純以最大共同子字元為擷取依據的研究，共產出 51,000 組配對，精確率為 51.3%，成效並不顯著（Li & Yarowsky 2008）。

由於中文縮寫形態並沒有很明確的被定義出來，只能用經驗法則訂定。近幾年的研究將縮寫詞的形式大約歸類成三種類別：(1)縮減法、(2)刪去法、(3)概括法。實驗結果多使用召回率（Recall）及精確率（Precision）來評估。以下簡述這三種方法，並整理相關範例於表 1。

1. 縮減法：此方法為三種方法中最常見的一種，可從原形詞中的任意位置挑選字元出來，再將其重新排列組合。挑選的字元數必須少於原形詞，但通常也不會超過三個字。例如：原形詞「台灣大學」當中，挑選「台」、「大」出來組成縮寫詞「台大」。
2. 刪去法：此法與縮減法類似，一樣是將想要當成縮寫的字元挑選出來組合，

但在原形詞挑選出的字必須是連續的。如原形詞「技嘉電腦公司」挑選連續的「技」、「嘉」成為縮寫詞「技嘉」。

3. 概括法：概括法是指由概括詞（generalized term）指向其涵蓋的所有子詞（specialized term）集合。一般為將多個原形詞所共有的一個詞或語素抽取出來後，在它之前加上表示原形詞數目的數詞或數量短語，且省略其餘部分，如原形詞「陸軍」、「海軍」、「空軍」可以簡稱為「三軍」。

表 1：中文縮寫詞範例

縮寫方式	原形詞	縮寫詞
縮減法	台灣大學 資訊管理學系 第一核能發電廠	台大 資管系 核一廠
刪去法	華碩電腦公司 中華航空	華碩 華航
概括法	高血壓、高血糖、高血脂 多喝、多吃、多尿	三高 三多

過去縮寫詞研究多以共現理論為基礎，某些研究增加詞性考量以訓練規則，發現單一類別成效優於混合類別，原因推測為此種方式對某些變化性高的類別並不適用（Huang & Yang 2005）。其後的延伸研究，探究縮寫詞與原形詞相對位置與共現率的關係，企圖找出對應條件規則，實驗結果發現，在上下文句數差值及共現率兩個條件的限制下，縮寫詞及原形詞對應精確率平均可達 86%~93%，優於以詞性為特徵選取的 70%~80%，也印證單一類別成效確實優於混合類別（黃純敏等 2007）。該研究團隊最新的研究提出新的權重計算方式（MR value），用於多型縮寫詞的實作，目地是要找出縮寫詞與原形詞有一對多的關係，研究顯示其配對精確率可達 85%~98%。此項研究成果可提供資訊檢索時，增加召回率的參考（黃純敏&蕭明華 2012）。雖然上述所列縮寫詞的研究結果，精確率可以達到 70%~98%，但召回率數值因礙於無正確答案比對，因此僅能推估或無法檢測。且因研究方法的限制，對於測試文章的數量，也多僅能做少量的測試或侷限於測試資料集，無法進行同步更新。

部分研究以隱藏式馬可夫模型（HMM）進行縮寫詞配對，如：Chang 與 Lai（2014）使用 1,235 對詞組作為訓練資料集，進行兩項實驗，第一項實驗由原形詞推測縮寫詞，正確率達 72%；第二項實驗則由縮寫詞反推原形詞，正確率為 51%。另外 Chang 與 Teng（2006）也使用此模型以 94 個檔案作為實驗資料集，提出單字

元回復（Single Character Recovery）反覆訓練方法，進行縮寫詞推估原形詞的實驗，結果顯示訓練資料的精確率為 62%；測試資料則為 50%。可見以隱藏式馬可夫模型應用於同義詞庫建置，仍有不少探討的空間。

三、同義詞庫建置研究

（一）人工建置同義詞庫

在人工建置同義詞庫方面，以大陸的梅家駒等，在 1983 年所編撰的「同義詞詞林」最具代表性（梅家駒等 1982），目的是希望能夠幫助文字寫作以及翻譯。該詞庫收錄了近七萬的詞彙，全部按詞義編排，分成大、中、小三類，大類以大寫拉丁字母編號，分別是：1.人、2.物、3.時間與空間、4.抽象事物、5.特徵、6.動作、7.心理活動、8.活動、9.現象與狀態、10.關聯、11.助語、12.敬語，依序往下按各類不同的特點以小寫字母劃分 94 個中類，再往下以阿拉伯兩位數字劃分 1,438 個小類，而小類中再並列 3,933 個標題詞。

該詞庫雖然字彙豐富，但由於分類不夠完善，到 2009 年魯東大學以同義詞詞林為基礎重新編寫，並將之改名為「新編同義詞詞林」（馮志偉 2009），共分出 15 個大類、203 個中類、1,477 個小類，收錄了約 13 萬個字詞，並且期望能提供給自然語言、機器翻譯、資訊檢索、語言教學以及寫作使用，雖然「新編同義詞詞林」的詞量非常充足，也經常被用來與自動化建置的同義字庫比較、評估的對象，不過對於時代文化的多元，新字詞產生速度飛快，使得同義詞詞林雖然擁有許多正式的同義詞，但對於層出不窮的創新語彙，要能夠成功判斷字詞之間是否同義，則是束手無策，例如：知名小說家「九把刀」本名「柯景騰」，並未納入詞庫；此外形容不願工作，繼續依靠父母的照顧及經濟支援的年輕人，過去稱為「米蟲」，現在通稱「啃老族」，也未收錄。類此隨著時代變化而創造出新的詞彙及同義詞，可預見任何有形的詞庫其收錄範圍及更新頻率，將永遠趕不上新詞的產生速度。

（二）自動化建置同義詞庫

在自動化同義詞庫建置方面，陸勇等學者在 2009 年提出了以多策略混合方式（Multiple Hybrid Strategies method）建構中文同義詞庫（Lu et al. 2009），其方法首先進行字詞字面相似度計算，此步驟只考慮詞彙的字面結構等因素，沒有考慮語義、語境等因素。接著利用句法規則進行預定的特徵模式配對，最後從語義角度來計算字詞之間的關係，再將字詞之間語義相似度的判斷轉化成頁面排名（PageRank）。該研究以 5,000 個搜尋字作為資料集，利用多策略混合方式萃取線上百科全書（包括：維基百科、百度百科與互動百科）的中文同義字，最後評估結果顯示，F-Score 值可以達到 93%，相較於使用同義詞詞林的 62%高。由於線上百科全書有不斷更新內容的特性，這種萃取方式能夠解決過去傳統研究有關詞庫

時效性問題。但是以 PageRank 分析法，若是電腦要自動地找出一個搜尋字的同義詞就必須要查找出所有線上維基百科中出現這個搜尋字的頁面，這種方式雖然能夠提升整體研究精確率，但是長期下來會拖累整體研究環境之網路速度，如此高負荷的研究成本對於要能夠隨時更新的同義詞庫是非常不適合的。

雖然線上百科全書中可以萃取出許多非正式的同義詞，但是只依賴線上百科全書作為同義詞萃取來源是不足的，因為有許多擁有歧視、非法的字眼是記者常用的字詞，如：「波波」一詞指到波蘭或東歐念 4 年醫學院，缺乏臨床經驗，回台到醫院實習後，無法通過國考，只能在醫院裡當永遠的實習醫師。這些在線上百科全書並不會記載，如此會影響整體召回率。除此之外，對於概括縮詞，該研究並沒有提出任何的萃取方法，只能夠萃取出縮寫詞以及異形同義。因此要如何萃取出過去都沒辦法萃取的概括縮詞，如何以客觀的方法評估研究成果，是本研究要探討且要解決的重要議題。

參、研究方法

一、實驗流程概觀

本研究考量實驗語料庫取得之便利性及內容多樣化，以台灣 Yahoo!奇摩新聞網為取材對象，擷取 2013 年 4 月至 8 月新聞，包括政治、社會、地方以及生活四大類，各 3,000 篇，總數 12,000 篇新聞文件為資料集。實驗採多元方式建構同義詞詞庫，主要分成三個步驟進行：首先針對資料集進行斷詞及關鍵詞萃取，其後利用維基百科內文組成結構，以本研究提出的同義詞判斷法與概括縮詞萃取演算法，平行處理概括縮詞萃取、同義詞萃取、同形異義詞萃取，最後進行使用者評量。研究架構如圖 1 所示。鑑於同義詞會因隨時間而改變用法，亦註記取得時間，實驗結果開放使用者驗證、評量與建議，以達到活化同義詞庫及兼顧涵蓋率。以下就實驗流程分述如次：

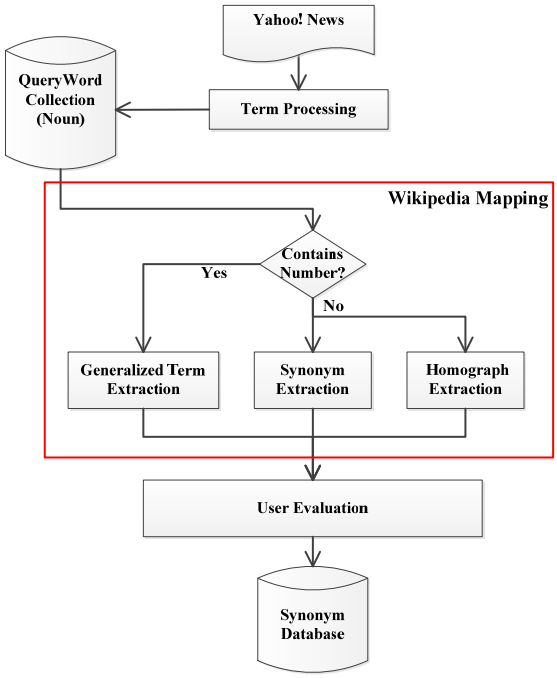


圖 1：研究架構概觀

（一）字詞處理（Term Processing）

此項處理事件主要目的為經由 CKIP 斷詞系統處理下載的新聞資料集，並挑選其中名詞並作為查詢字，篩選過程如下：

1. 萃取新聞內文：以「派遣工比例過高教育部挨轟」這篇新聞為例，我們使用 Apache 開發的 HttpClient 作為萃取 HTML 原始碼的工具，再以 Jsoup 剖析及萃取內文，如圖 2。

圖 2：Yahoo!新聞內文範例

2. 斷詞處理：透過 CKIP 斷詞後，回傳含所有詞性的字詞，本實驗從中篩選出名詞、過濾重複字詞、再將結果儲存，共取出 109,543 個名詞作為查詢字，如圖 3。

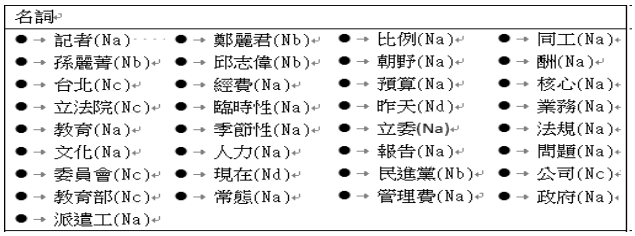


圖 3：擷取名詞作為查詢字片段資料列表

（二）概括縮詞萃取（Generalized term Extraction）

經觀察中文文章中的概括縮詞所涵蓋的子詞多使用連續的頓號「、」或是特定中文連接詞（如「或」、「和」、「以及」…等）來進行列舉及延伸，本研究據此設定規則，以頓號與連接詞為特徵，提出過去未有的概括縮詞萃取演算法。首先讀取全文字串，斷詞後所萃取出的查詢字，若有中文數字，如「三民主義」、「三軍」或「四大天王」…等，則利用維基百科進行概括縮詞萃取，如遇到符合的字元（包含「、」、「或」、「以及」、「和」等字元）則往後偵測，計算這些特定字元所出現的次數，若是符合條件則萃取出來，處理流程如圖 4。以查詢詞「三軍」為例，由於該百科全書對「三軍」的解釋有四種。最後抽取出來的結果為：「上軍、中軍、下軍」、「步兵、騎兵、戰車」、「前軍、中軍、後軍」以及「陸軍、海軍、空軍」，如圖 5。

```
public PairWords getPairWords(QueryWord){
    ArrayList<PairWords> result;
    if(Query Word contains figure numbers){
        int figureLength = the figure number - 1;
        while(Scan the Wikipedia Context){
            index = current Scan text;
            if(The index is a Comma mark){
                int count = calculate the count of all chinese conjunction mark(ex:或)
                etc before the other punctuation(ex:"，" or "：");
                if(The count is equal to figureLength+1){
                    String[] pairWords = the words between the Comma Mark;
                    int firstPairWordsLength = calculate the Mode of pairWords length;
                    String firstWord = retrieve the words from (index - firstPairWordsLength)
                        to index;
                    result.add(firstWord && pairWords);
                }
            }
        }
    }
    return result;
}
```

圖 4：概括縮詞萃取流程

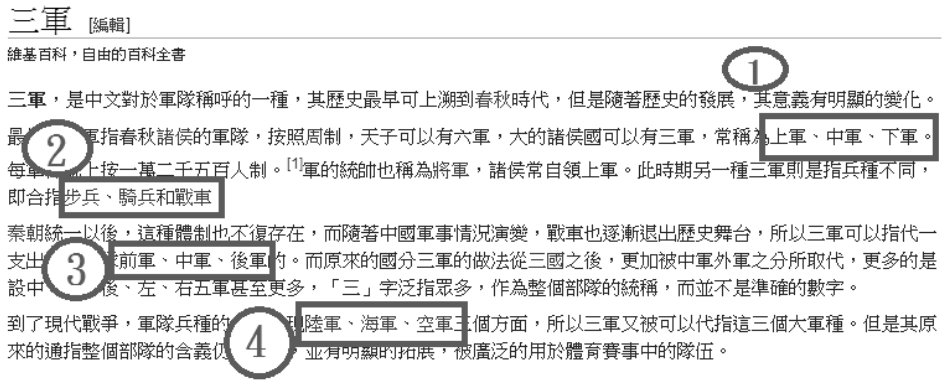

三軍 [編輯]
維基百科，自由的百科全書
三軍，是中文對於軍隊稱呼的一種，其歷史最早可上溯到春秋時代，但是隨著歷史的發展，其意義有明顯的變化。
最[illegible]指春秋諸侯的軍隊，按照周制，天子可以有六軍，大的諸侯國可以有三軍，常稱為上軍、中軍、下軍。
每軍[illegible]上按一萬二千五百人制。[1]軍的統帥也稱為將軍，諸侯常自領上軍。此時期另一種三軍則是指兵種不同，
即合指步兵、騎兵和戰車
秦朝統一以後，這種體制也不復存在，而隨著中國軍事情況演變，戰車也逐漸退出歷史舞台，所以三軍可以指代一
支出[illegible]前軍、中軍、後軍的。而原來的國分三軍的做法從三國之後，更加被中軍外軍之分所取代，更多的是
設中[illegible]後、左、右五軍甚至更多，「三」字泛指眾多，作為整個部隊的統稱，而並不是準確的數字。
到了現代戰爭，軍隊兵種的[illegible]陸軍、海軍、空軍三個方面，所以三軍又被可以代指這三個大軍種。但是其原
來的通指整個部隊的含義仍[illegible]並有明顯的拓展，被廣泛的用於體育賽事中的隊伍。

圖 5：「三軍」在維基百科的解釋

（三）同義詞萃取（Synonym Extraction）

依據維基百科的內文組成結構，本研究歸納出幾種輕量且有效的萃取模式，說明如下：

1. 首段首字詞萃取（First Head Term Extraction）：由於維基百科首段會列出最常見的描述語，如：使用者無論輸入「九把刀」或「柯景騰」都自動定向「九把刀」。內文首字詞則列出本名，其後才是內容的詳細介紹，因此我們將首字詞視為萃取對象，作為與查詢詞配對，萃取流程如圖 6。

```
if (searchKey.equals(firstHead) == false) {
Synonym synonym = new Synonym();
synonym.setQueryWord(originalSearchKey);
synonym.setMappingWord(firstHead);
synonym.setSourceType("FirstHead");
synonym.setSuggestPeople("Computer");
synonym.setTotalScore(100);
synonym.setEvaluatePeopleAmount(1);
synonym.setMappingTime(nowTime);
wikiObject.getSynonyms().add(synonym);
```

圖 6：首段首字詞萃取流程

2. HTML 粗體標籤配對（HTML Bold Tag Mapping）：由於維基百科對於描述語與簡稱均使用粗體字表示，因此我們藉由首字詞及粗體字特徵視為同義詞及其配對的線索，配對流程如圖 7。圖 8 說明「國立成功大學」與其簡稱「成大」使用粗體字表示，而成為本研究配對及萃取對象。

```
public ArrayList<Synonym> boldAlgorithm(WikiObject wikiObject) {
Element contentElements = wikiObject.getFirstContentHTML();
ArrayList<Synonym> result = new ArrayList<Synonym>();
for (Element e : contentElements.select("b")) { //搜尋<b>
    Synonym n = new Synonym();
    n.setQueryWord(wikiObject.getSearchKey());
    n.setMappingWord(e.text());
    Calendar c = Calendar.getInstance();
    n.setMappingTime(c.getTimeInMillis());
    n.setSourceType("boldAlgorithm");
    result.add(n);
}
return result;
}
```

圖 7：HTML 粗體標籤配對流程


國立成功大學 [編輯]

維基百科，自由的百科全書

國立成功大學（英語：**National Cheng Kung University**），簡稱**成大**，是中華民國一所國立大學，也是分享「邁向頂尖大學計畫」5年500億補助的研究型大學，已發展為學門規模全台第2的綜合大學[註 1]成大2013年在泰晤士高等教育世界大學排名第301至350名，在泰晤士高等教育亞洲大學排名則第47名，兩項排名都是在臺灣的各大學排第5名[1]。

圖 8：「國立成功大學」在維基百科中的解釋畫面

3. 模式對應（Pattern Mapping）：維基百科對於機構詞目多緊接英文翻譯，其後則為簡稱，如圖 9。為了讓電腦能夠自動辨識及檢出該簡稱，我們將此模式定義如圖 10，其中 QueryWord 是指每一個維基百科頁面的查詢字，MappingWord 是指要萃取的同義詞，藉此找出同義詞。例如：以查詢字「國立成功大學」及「國立政治大學」為例，可順利檢出配對字「成大」與「政大」。


國立成功大學 [編輯]

維基百科，自由的百科全書

國立成功大學（英語：**National Cheng Kung University**），簡稱成大，

國立政治大學 [編輯]

維基百科，自由的百科全書

國立政治大學（英語：National Chengchi University, NCCU），簡稱政大

圖 9：「國立成功大學」及「國立政治大學」在維基百科中的解釋畫面

```
QueryWord（英語：MappingWord），簡稱MappingWord，
```

圖 10：配對字模式定義例子

4. 資訊盒數值配對（Infobox Value Mapping）：經觀察維基百科的編排結構發現，該百科全書使用「Infobox」（資訊盒）的表格記載許多結構化的資訊，本研究藉由此種結構化特性，萃取相關欄位作為同義詞的來源之一，如維基百科中有關「台大」的解釋畫面，指出「台大」又稱為「帝大」、「臺灣第一學府」和「臺灣最高學府」。此外從所列示的網站名稱，「http://www.ntu.edu.tw」，可萃取「NTU」是「國立台灣大學」的英文縮寫，如圖 11。

圖 11：「台灣大學」在中資訊盒的解釋

（四）同形異義詞萃取（Homograph Extraction）

本研究主要探討同義詞及縮寫詞，但研究過程發現有一詞多義（同形異義）的情況，因此亦將其亦列入萃取範圍，未來納入詞庫可用以協助資訊檢索，但不包括於最後同義詞評量項目。同形異義詞指兩個詞，字面形式相同，但涵義卻大相逕庭。由於外形相同，因此常造成資訊檢索的檢出失誤。本研究透過維基百科查詢，若搜尋結果有兩個以上時，則加入句法規則判斷，如符合（Na+Nb）時，Na 視為 Nb 之限定語（qualifier）。如「小甜甜」一詞，將整理出「小甜甜」及「（藝人（Na）小甜甜（Nb）」，如圖 12。前者為漫畫，後者為某女藝人，加註限定語作為同形詞之區隔，可明顯區分出兩同形詞之不同，有助於搜尋引擎作為查詢區別。

圖 12：同形異義詞萃取解釋畫面

（五）使用者評估 （User Evaluation）

由於同義詞並沒有絕對客觀的標準答案可資核對，為驗證本研究所提出方法是否有效，也為提升詞庫的涵蓋率及精確率，我們將進行兩階段評量，並將使用者回饋納入同義詞庫建置的架構之中。實驗除了進行精確率與召回率計算外，也考量使用者主觀的接受度，作為評量參考，評量流程如圖 13。由於同義詞會因為時間因素而改變，因此在新增同義詞至資料庫之前，我們也將取得的時間註記至資料庫中。圖 14 臚列資料表欄位，包括編號、查詢詞、比對萃取結果、比對規則、取得時間、及接受度。實驗過程及結果討論於下節詳述。

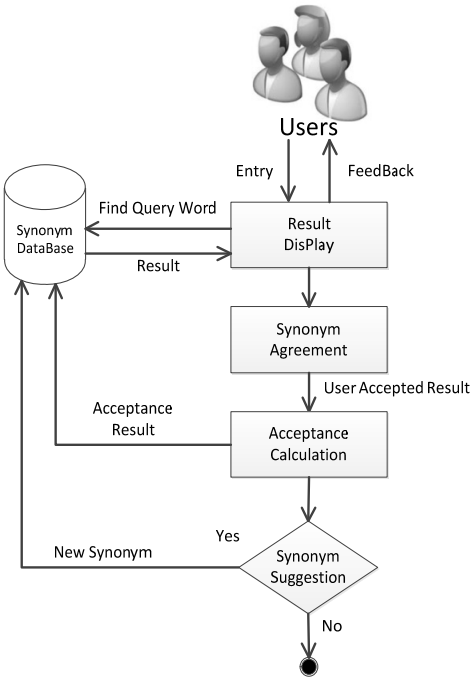


圖 13：使用者進行同義詞接受度評分示意圖

uid	queryWord	mappingWord	sourceType	mappingTime	acceptability
4994	三軍	三軍	FirstHead	2013-05-11	100
4995	三軍	陸軍、海軍、空軍	commaAlgo	2013-05-11	100
4996	三軍	上軍、中軍、下軍	commaAlgo	2013-05-11	81
4997	三軍	步兵、騎兵、戰車	commaAlgo	2013-05-11	86
4998	三軍	前軍、中軍、後軍	commaAlgo	2013-05-11	92
⋮					
5397	台大	國立臺灣大學	FirstHead	2013-05-15	100
5398	台大	臺大	boldAlgorithm	2013-05-15	100
5399	台大	National Taiwan University	en	2013-05-15	100
5400	台大	杜鵑花城	InfoBox	2013-05-15	70
5401	台大	台灣第一學府	InfoBox	2013-05-15	100
5402	台大	帝大	InfoBox	2013-05-15	86
5403	台大	臺灣大學	boldAlgorithm	2013-05-15	100

圖 14：同義詞配對存取範例

肆、實驗結果與討論

一、實驗評估標準

由於本研究以同義詞為探討對像，對於同形異義字則不列入評估項下。實驗使用精確率以及召回率作為萃取結果的衡量準則如表 2 及公式(1)、(2)所示。精確率是指系統所萃取出的同義詞，正確的結果所佔的比率，精確率愈高表示系統在檢出及配對能力愈好；召回率則是指系統能找出所有正確的詞組比率，如果召回率很高，則表示程式設計考量很周全，不輕易漏掉正確的資訊。例如本實驗所找出的某詞目的 8 個同義詞中，有 7 個正確，1 個錯誤，另外 2 個是受測者建議的，那麼精確率就是 87.5%，召回率則是 77.7%。

表 2：檢索結果差異判斷表

Results	Correct Match	Incorrect Match
Retrieved	A	B
Not Retrieved	C	D

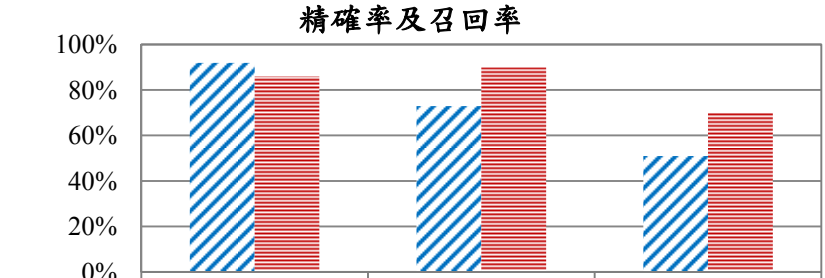
$$P = \text{Precision rate} = A/\ A + B \tag{1}$$

$$R = \text{Recall rate} = A/\ A + C \tag{2}$$

二、評估步驟

（一）階段一：評估精確率與召回率

為了解受測者在無參考資源輔助情況下，對同義詞組的直覺反應，本步驟首先以隨機方式但務必包含一般縮寫詞、概括縮詞、同義詞在內的 30 組配對詞，將之均分成 6 份，再分配給參與測試的 60 位雲科大資管系學生，使每一份相同問卷都有 10 人填寫。所有受測者被要求在問卷上圈選出不適當詞組及建議合適的同義詞。問卷回收結果再由 5 位資管研究生討論受測者的建議是否為合適，並上網查證有疑義者，再以多數決方式決定是否納入詞庫收錄範圍，並作為下一階段的標準答案。評估結果統計如圖 15。其中在精確率方面，我們發現表現最好的是縮寫詞，推測其原因可能基於多數縮寫詞是由原形詞縮減而成，因此即使受測者過去沒有見過此縮寫詞，也能夠從中推論及認可，使得精確率達到九成二；在召回率方面，表現較好的為概括縮詞，其原因可能為概括縮詞需一一列舉子詞，受測者在沒有網路資源輔助下，對某些概括縮詞是由哪些詞所組成，記憶不全，如：「八寶粥」此字詞，是由「糯米」、「薏仁」、「乾百合」、「白果」、「蓮子」、「紅棗」、「紅豆」、「龍眼肉」此八個子詞所組成，所以當系統列出此八個子詞，受測者多傾向同意系統所萃取出的結果，填入的建議詞也較少，因而召回率可達九成；但另一方面對於較深澀的子詞組合，多數傾向認為是錯誤的，因此使得精確率表現也並不突出。至於異形同義詞因不似縮寫詞有跡可循，可從中推論，也不像概括縮詞之艱深，受測者有較多表達意見的空間，可能因此降低了精確率與召回率。此階段共有 38 組使用者所建議的詞，經過五位研究生利用網路查證後，剩餘 27 個詞，當中包括概括縮詞組 3 個、同義詞組 14 個、縮寫詞組 10 個，範例如表 3。



	縮寫詞	概括縮詞	異形同義
精確率	92%	73%	51%
召回率	86%	90%	70%

圖 15：階段一評估結果

表 3：使用者建議詞組範例

概括縮詞組		同義詞組		縮寫詞	
三寶	人參、貂皮、烏拉草	馬英九	馬小九	雲林科技大學	雲林科大
三寶	勞保、健保、199 吃到飽	陳菊	花媽	健康檢查	健檢
三鐵	台鐵、高鐵、北高捷運	行車紀錄器	黑盒子	台灣大哥大	台哥
		蘇貞昌	電火球	人力資源	人資

（二）階段二：評估接受度

本階段將以主觀接受度作為評量訴求，最後再將修正結果回測系統效度。實驗做法為將前一階段的問卷包含受測者所建議的詞，設計成線上問卷，開放線上評量，並要求每一位受測者勾選「是否同意這組同義詞是正確的？」，也持續接受修正及建議，而所建議的詞都會被系統納入候選列中，如其接受度超過六成，則納入詞庫。本階段評估系統開放 72 小時，參與受測人數為雲林科技大學資管系大三、大四生，受測人數達 120 人。我們視每一位受測者所勾選的正面意見作為接受度評估標準，計算公式為（接受度＝贊成人數／評估人數），假如在 120 位使用者中，所有受測者對「台灣積體電路製造股份有限公司」的縮寫詞是「台積電」認為是合理，那麼這一組的同義詞之接受度是 100%。實驗結果顯示接受度大多在 80%~100%之間，其中以縮寫詞高達 98%，最受肯定，其餘如概括縮詞為 89%，異形同義詞 84%，也都表現不錯，總體平均達 91%，數據結果如圖 16。此階段評估結果後，取得 11 組使用者建議的詞，加上第一階段所建議且收錄的 27 詞，在經過接受度評估後，未超過六成的有 12 個，範例如表 4。進一步分析發現某些使用者建議的同義詞雖未包括在維基百科中，但是接受程度卻很高，如：「行車記錄器」，受測者認為同義詞還有「黑盒子」的意思，也受到高度肯定。由此說明同義詞之認定，確有其靈活與其無可掌握性。開放使用者建議，雖有誤植之可能，但經由群眾智慧的修正，可化解此種顧慮，也可活化同義詞庫及達到與時並進的涵蓋率，確實是未來建構的方向。在實驗最後，我們評估系統萃取同義詞的速度，平均找出 1 組同義詞只需要 0.01 秒。

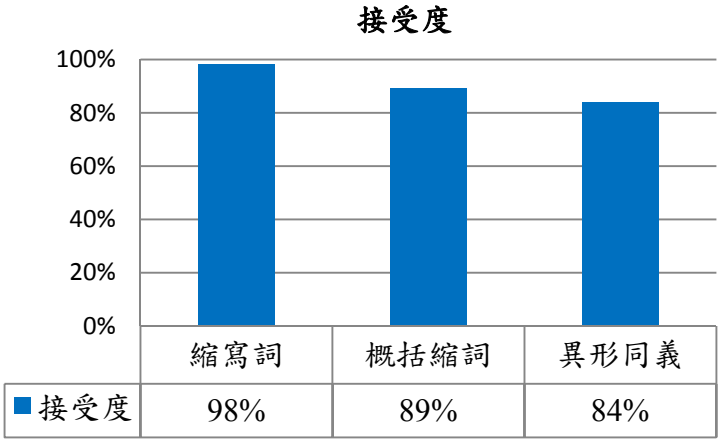


圖 16：階段二評估結果

表 4：使用者建議詞組未通過接受度門檻範例

查詢字	使用建議詞
蕭敬騰	歐陽盆栽
皮卡丘	電氣鼠
衝浪	滑浪
電子計算機	電子機
三寶	勞保、健保、199 吃到飽

（三）字詞萃取結果與討論

本實驗一開始將 Yahoo!新聞網站的政治、財經、社會以及地方四大類別的 12,000 篇新聞進行 CKIP 斷詞，取出 573,697 個名詞，經去除重複後，得出 109,543 個名詞作為查詢字。經本研究所提出的同義詞配對識別法，將查詢字平行處理及接受使用者建議，共取得 3,597 組字詞，其中概括縮詞組有 191 個、同義詞組 2,448 個、同形異義詞組 958 個、及使用者建議且通過接受度門檻者 26 組字詞，如表 5。各類字詞萃取結果顯示如表 6。

表 5：各類字詞組數列表

概括縮詞	同義詞（含縮寫詞）	同形異義詞	使用者建議
191	2448	958	26

表 6：各類字詞萃取結果列表

查詢字	縮寫詞	查詢字	概括縮詞
國立臺灣科技大學	臺科大；臺科	三從	從父、從夫、從子
宏碁股份有限公司	宏碁；Acer	三國	曹魏、蜀漢、孫吳
高速鐵路	高鐵	四德	婦德、婦言、婦容、婦功
國民中學學生基本學力測驗	國中基測	五代	後梁、後唐、後晉、後漢與後周
查詢字	同義詞	查詢字	使用者建議詞
關公	關羽;關聖帝君;關二哥;協天大帝	陳菊	花媽
九把刀	柯景騰;Giddens Ko	馬英九	馬小九
臺灣	福爾摩沙；寶島	行車紀錄器	黑盒子
紅樓夢	石頭記	蘇貞昌	電火球

回顧過去學者（詳述於文獻探討）曾分別以字詞共現、詞性、上下文句數、MR Value 權重計算、隱藏式馬可夫模型、或結合字面相似度、句法規則及語意考量的多策略混合方式，進行縮寫詞與原形詞配對自動建置及同義詞庫，經統整過去相關研究與本研究比較，臚列幾個面向如表 7。經比較各種萃取方法，發現單純以字詞共現率或使用隱藏式馬可夫模型表現較差，因此在表 8 所列出之本研究與相關研究之精確率與回現率比較，未將此二者列入。表列其他方法之精確率雖可以達到 80%~92%，但召回率數據多礙於資料集本身並無正確答案可資比對，因此多為推估之值，或直述無法檢測。反觀本研究進行兩階段評量，並將使用者回饋及接受度納入考量，因此字詞可以持續更新至同義詞庫內，提高字詞品質及正確率之估算，因此回現率之準確性具有公信力及說服性。

表 7：與相關研究之面向比較表

	相關研究（如文獻探討）	本研究
資料集來源	萃取來源為網路新聞或百科全書。	萃取及比對來源較多元，包括台灣 Yahoo!資料集、維基百科以及使用者建議。
研究焦點	針對刪去法及縮減法，做字面上的判斷。無法解決概括法辨識。	除了能解決刪去法以及縮減法之外，亦可有效解決概括法。
主要萃取方式	以統計方式為主，包括使用隱藏式馬可夫模型（HMM）、字詞共	分析維基百科的內文結構，以演算法進行比對及萃取。

	相關研究（如文獻探討）	本研究
	現率、 詞性判斷、上下句數、MR Value 等。	
時效性	未考慮同義詞時效問題，較新穎的同義詞沒有持續訓練。	時效性高且彈性，持續訓練，較新穎的字詞可以持續更新至同義詞庫內。
使用者回饋	未考慮使用者回饋，依賴電腦自動進行同義詞萃取。	互動性高，除了電腦自動化萃取外，也可讓使用者建議新同義詞。
同義詞庫架構	無特別針對同義詞庫欄位設計，無法判斷同義詞的型態。	針對同義詞庫設計適當欄位，讓電腦自動判斷同義詞形態、時效性以及合理度。

表 8：與相關研究之精確率與回現率比較表

相關研究（如文獻探討）	Precision rate	Recall rate
本研究（維基百科的內文組成結構）	92%	86%
共同子字串結合詞性判斷	80%	--
共同子字串結合上下句數判斷	92%	27%
共同子字串結合 MR Value	92%	94%
多策略混合方式	92%	95%

伍、結論

相較於過去相關的研究，其資料來源十分依賴有限的資料集主題特性，僅能做少量的資料測試，本研究萃取及比對來源較多元，包括：台灣 Yahoo!新聞文件、維基百科以及使用者建議。有別於過去研究的萃取方式多半使用統計方式，其共現率計算容易產生誤判，本研究以維基百科的內文組成結構為基礎，所提出的輕量型建構同義詞庫方法除了能讓電腦自動地萃取出縮寫詞、異形同義及同形異義詞之外，還能夠萃取出過去研究所無法解決的概括縮詞。

鑑於過去對成果評估方式常流於主觀，每組縮寫詞答案都是由研究作者、專家或是問卷設計定義出來的結果，本研究認為文字具有高自由度發揮的空間，因此同義詞除了大家較熟稔的用語，可以快速得到共識外，其實並沒有絕對標準答案，只要多數人能理解及接受該用法，就達到溝通的目的。這也就是在進行精確率與召回率評量時，不易獲得準確及客觀一致結果的原因。因此本實驗採兩階段評估方式，包含主客觀評量。除了使用精確率及召回率做為評估方式外，也考量

將使用者主觀的接受度，作為評量參考，並將使用者回饋納入同義詞庫建置的架構之中。實驗結果顯示本研究所提出的方法，在第一階段評量平均精確率為 72%、召回率 82%，其中縮寫詞的精確率高達 92%，概括縮詞的召回率為 90%。第二階段評量結果，使用者接受度亦達 91%。在效率方面，平均找出 1 組同義詞只需要 0.01 秒。

由於本研究所建構之同義詞庫比對來源為維基百科，該線上百科已有群眾智慧基礎，再加上開放使用者建議，可蒐集到遺漏、冷僻或新穎的用詞，其中雖有誤植之可能，但使用者所建議之字詞須通過同意門檻值方可納入同義詞庫。如未來字彙集累積愈多、使用者愈廣，詞庫的品質將愈穩定，也愈不容易受到少數意見影響，可達到活化同義詞庫及兼顧涵蓋率。

本研究主要探討同義詞及縮寫詞，惟在研究過程發現有一詞多義（同形異義）的情況，因此亦將其亦列入萃取範圍，為恐混淆受測者，不納入最後同義詞評量項目。未來研究可考量以維基百科為基礎，拓展探討範圍至混淆字集萃取與成效評量，甚者延伸至自動偵測錯誤與修正。由於維基百科不記載有歧視、非法的字眼，研究者也可考慮以民眾最常用的社交網站，如：Facebook、無名小站、新浪微博、噗浪（Plurk）等為資料來源，以補足這個缺點。

誌謝

本研究受行政院國家科學委員會計畫 NSC 102-2218-E-224-001 補助，特此致謝。

參考文獻

梅家駒、竺一鳴、高蘊琦、殷鴻翔（1982），*同義詞詞林*，上海辭書出版社。

陳良駒、陳日鑫（2010），『植基於詞彙數量關係探討軍事新聞主題－以青年日報為例』，*資訊管理展望*，第十二卷，第一期，頁 21-42。

馮志偉（2009），『語義互聯網與辭書編纂』，*暨南大學学华文学院学报*，第四卷，第四期，頁 88-94。

黃純敏、石朝元、張精哲（2007），『中文縮寫詞延伸研究』，*第十八屆國際資訊管理學術研討會論文集（ICIM 2007）*，銘傳大學，台灣，五月二十六日

黃純敏，蕭明華（2012），『改進中文縮寫詞與原形詞配對率』，*第十八屆兩岸資訊發展高峰論壇（CSIM 2012）*，實踐大學，臺灣，八月二十日。

Chang, J.S. and Lai, Y.T. (2004), ‘A preliminary study on probabilistic models for Chinese abbreviations’, *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing (SIGHAN 2004)*, Barcelona, Spain, July 25-26, pp. 9-16.

Chang, J.S. and Teng, W.l. (2006), 'Mining atomic Chinese abbreviation pairs with a probabilistic single character word recovery model', *Proceedings of the fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN 2006)*, Sydney, Australia, July 22-23, pp.17-24.

Fu, M.H., Peng, C.H., Kuo, Y.H. and Lee, K.R. (2012), 'Hidden community detection based on microblog by opinion-consistent analysis', *Proceedings of 2012 International Conference on Information Society (i-Society)*, London, UK, June 25-28, pp. 83-88.

Hong, C.M., Chen, C.M. and Chiu, C.Y. (2009). 'Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems', *Expert Systems with Applications*, Vol. 36, No. 2, pp. 3641-3651.

Huang, C.M. and Yang, C.P. (2005), 'Chinese Abbreviations and Expansion', *Proceedings of the National Computer Symposium (NCS 2005)*, Kuan Shan University, Tainan, Taiwan, December 15-16.

Kang, S.S. and Hwang, K.B. (2006), 'A language independent n-gram model for word segmentation', *Proceedings of the 19th Australian joint conference on Artificial Intelligence: advances in Artificial Intelligence (AUS-AI 2006)*, Hobart, Australia, December 4-8, pp. 557-565.

Kit, C., Xu, Z. and Webster, J.J. (2003), 'Integrating ngram model and case-based learning for Chinese word segmentation', *Proceedings of the second SIGHAN workshop on Chinese language processing (SIGHAN 2003)*, Sapporo, Japan, July 11-12, pp. 160-163.

Li, Z. and Yarowsky, D. (2008), 'Unsupervised translation induction for Chinese abbreviations using monolingual corpora', *Proceedings of the Association for Computational Linguistics (ACL-2008),* Clumbus, Ohio, June 15-20, pp. 425-433.

Lin, C.J., Zhan, J.C., Chen, Y.H. and Pao, C.W. (2012), 'Strategies of processing Japanese names and variant characters in traditional Chinese text', *Computational Linguistics and Chinese Language Processing*, Vol. 17, No. 3, pp. 87-108.

Lu, Y., Zhang, C. and Hou, H. (2009). 'Using Multiple Hybrid Strategies to Extract Chinese Synonyms from Encyclopedia Resource', *Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC 2009)*, Kaohsiung, Taiwan, December 7-9, pp. 1089-1093.

Nguyen, H.T. and Cao, T.H. (2008), 'Named entity disambiguation on an ontology enriched by Wikipedia', *Proceedings of the IEEE International Conference on*

Research, Innovation and Vision for the Future (RIVF 2008), Ho Chi Minh City, Vietnam, July 13-17, pp. 247-254.

Tsai, R.T.H. (2010), 'Chinese text segmentation: A hybrid approach using transductive learning and statistical association measures', *Expert Systems with Applications,* Vol. 37, No. 5, pp. 3553-3560.

Zhang, J., Nie, J.Y., Gao, J. and Ming, Z. (2000), 'On the use of words and N-grams for Chinese information retrieval', *Proceedings of the fifth International Workshop on Information Retrieval with Asian languages (IRAL 2000)*, Hong Kong, China, September 30-October 1, pp. 141-148.

Zhang, K., Liu, Q., Zhang, H. and Cheng, X.Q. (2002), 'Automatic recognition of Chinese unknown words based on roles tagging', *Proceedings of the first SIGHAN workshop on Chinese language processing (SIGHAN 2002)*, Taipie, Taiwan, August 31-September 1, pp. 1-7.