圖書資訊學研究 11：2 (June 2017)：139-179

基於詞性組合規則結合維基百科進行中文命名實體辨識與消歧義

Using Part-of-Speech Tagging Based Approach with Wikipedia to Assist Chinese Named Entity Recognition and Disambiguation

黃純敏

Chuen-Min Huang

雲林科技大學資管系副教授
Associate Professor
Department of Information Management
National Yunlin University of Science and Technology

【摘要 Abstract】

傳統命名實體辨識多採用規則與機率的方法，然而礙於語義混淆特性與未知詞的增長，精確率難以有效提高。本研究藉由詞性組合定義命名規則，並加入姓名鏈結演算法及透過維基百科文本編輯特性，以協助辨識及消歧義。研究發現應用姓名鏈結機率公式結合句法規則，可大幅提高人名辨識精確率；對於「地名」/「組織名」，由於二者命名規則相似，過去研究需藉助詞庫及特殊詞幹集區別，本研究透過簡易地名規則並結合維基輔助分歧。實驗結果顯示本研究在精確率、召回率、F-measure 分別達 86.32%、75.33%、80.33%，相較於其他大規模規則的判斷研究，及採用人工標註結合

作者：黃純敏 jennyhuang921@gmail.com
投稿日期：2017 年 5 月 16 日；接受日期：2017 年 6 月 23 日

HMM 機器學習的研究，本研究所歸納的規則不僅精簡，整體表現亦毫不遜色，尤其以精確率最為突出。

Traditional Named Entity Recognition (NER) adopts rule-based and/or probabilistic models in morphological analysis, while it still exists the problem of low accuracy due to the problem of semantic ambiguity and the growth of unknown words. In this study, we applied syntax rules of names and places to process Chinese NER, and extracted features from Wikipedia to assist disambiguation and thereby help to improve recognition accuracy. Our study found that the recognition accuracy is raised because of a combination of syntax rule with name algorithm. In addition, since the location and organization names usually follow some particular verbs, we only configured basic rules for location ER and referred to the geographical directories of Infobox in Wikipedia to verify their identities. In our overall system evaluation, the precision rate achieves 86.32%, recall reaches 75.65%, and F-measure reaches 80.4%. Compared with other automatic rule construction and quasi-machine learning methods, we have a better performance particularly on the precision rate.

【關鍵字 Keywords】

命名實體辨識、命名實體消歧義、詞性組合、句法規則、維基百科

Named Entity Recognition; Name Entity Disambiguation; POS Combination,; Syntax Rules; Wikipedia

壹、前言

隨著網路的普及化，瀏覽網頁獲取即時訊息已經成為人們日常生活必要的活動。據統計，目前網路每天可產生 2.5 quintillion（10^{18}）位元組的資料，而其中光是近兩年所累積的，就佔了九成，包括以感測器收集的氣候資訊、社群媒體的發文、數位圖片和視頻、交易記錄、GPS 信號等等（IBM, 2015）。網路分享時代所快速產出的的巨量資料，最難處理的不在於量，而在於它的異質性與複雜性，若能有效解決其異質與複雜性，也就能挖掘出其潛在的價值。因此如何在過多的資訊中，有效過濾整理、整合資訊成為當務之急。以社會大眾最常閱讀新

聞資訊為例，人們想了解的是，新聞中有哪些人（Who），發生的時間（When）、地點（Where），發生什麼事（What），為什麼發生這些事情（Why），以及事情是如何發生的（How），藉由對人、事、時、地、物特徵字詞的掌握，即可快速了解新聞事件的重點。因此在此項資訊擷取，命名實體辨識（Named Entity Recognition, NER）就是其中最重要的關鍵技術。為行文方便，以下將以 NER 或「命名實體辨識」或「命名實體」表述同樣意涵，也將視需要簡稱為「命名」。

NER 的議題不算新穎，過去也已經有不少研究成果（Chen & Lee, 1996; Chen, Ding, & Tsai, 1998; Fu & Luke, 2005），因此近年來此議題已逐漸停歇，很少有新的研究成果。然而基於語文本身具有的模糊及不完全的特性，加上新生詞的組合不斷推陳出新，因此對於未知詞（unknown word）的命名偵測及語意混淆（同形異義、一詞多義）的釐清或消歧義（disambiguity），一直都還有探索的空間。因為中文的命名變化多端，過去提出的句法規則，仍需不斷檢驗其精確（precision）率及召回（recall）率。尤其以 NER 為基礎的加值研究，如：資訊檢索（Chen et al., 1998）、自動摘要（Furu, Wenjie, & Yanxiang, 2007）、主題模型的先驗知識等的實驗成效，都與 NER 的正確率有著高度的相關。基於上述理由，本研究認為此議題有進一步的探討的必要。

在 NER 的研究中，除了時間、日期、比率、貨幣較具有規律性，因此容易識別外，人名、地名、組織名之組成，則常隨時間演化而難以指認，相關新生詞的產出速率，不僅無法透過詞庫斷詞擷取，也不易規則化。以此句為例：「昨日報載誤指陳大同捐了伍佰萬元給雲林縣國民黨」，專有名詞標記需將：「陳大同」標註為人名，「雲林縣」標註為地名，「國民黨」標註為組織名。若斷詞或標註技術不佳，可能會將該句斷出「日報」、「指陳」、「伍佰」等詞彙，以致影響重要人事物的指認。過去研究在規則的訂定動輒達萬條，在「人名」辨識，多以百萬姓名詞庫為訓練基準，考量單複姓、詞頻及男女區別；「地名」與「組織名」因規則相近，則藉助特定詞庫、特殊詞幹集、及特殊動詞進行區別。然而即使在巨量規則下，也使用大量的關鍵詞協助辨識，但由於命名實體的組合型態太多，精確率始終難以提高。因此，對於如何能更準確辨識文本特徵詞，一直是資訊擷取技術中的極具挑戰的議題。

NER 大抵可分成句法規則以及機器學習模式。句法規則模式多參考句子組成規則作為擷取命名的依據，此種模式所建構的規則需考量文法（如：詞性標註）、語句分析（如：詞彙位置）、字詞特徵（如：

大寫），並結合字辭典輔助。例如：專有名詞接在頭銜後，則該專有名詞很可能是人名。此種方式通常需要投入大量人力建構詳細規則及專用詞庫（Chen & Lee, 1996; Chen, Ding, Tsai, & Bian, 1998），執行上也十分有效率，然而當詞庫未收錄時，則無法辨識；當句法規則顧慮不周時，容易引起規則之間的相互衝突，倘若因此增加過多判斷規則，則可能造成過度擬合（overfitting）的現象，導致每條規則只適用少數情況，失卻規則訂定的本意。

機器學習模式需透過大量訓練資料集，以建立語言模型達到辨識的效果，常見的方法包括：最大熵（ME）（Bender, Och, & Ney, 2003）、支持向量機（SVM）（Takeuchi & Collier, 2002）、隱藏式馬可夫（HMM）（Fu & Luke, 2005）、條件隨機域（CRF）（Chen, Zhang, & Isahara, 2006）等。上述除了 Fu 與 Luke（2005）是以簡體中文語料庫為處理對象外，其他均為英文。此種學習模型，除了需要龐大的訓練語料外，也經常無法跨領域識別。因此有些研究如 Verma 等（2013）建議縮小學習範圍，僅標註有用的例子，以減少大量標註的成本耗費（Verma, Sikdar, Saha, & Ekbal, 2013）。然而對於如何定義有用的例子，則沒有共識。

由於網路論壇、部落格、社群網站等 Web2.0 互動網站已蓬勃發展，研究人員開始重視群眾資源（crowdsourcing）的知識探勘。Mihalcea 與 Csomai（2007）藉之擷取有用的資訊，以協助監督式或半監督式機器學習（Mihalcea & Csomai, 2007）。其中維基百科（Wikipedia）即為群眾智慧最出色的產物，可稱是當今最大型的線上百科知識庫。此百科全書有 10 萬名積極貢獻者長期參與編輯工作，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。其編輯體例包含許多內文關聯結構，如同義詞以重定向頁（redirect page）導引、同形異義詞附有消歧義頁、相關詞目則有內文超連結，作用在於補充相關資訊，幫助釐清混淆或類似詞目。因其內容豐富，過去若干研究利用維基百科協助命名辨識，實驗結果都肯定在提升命名辨識率的正面效果（Bunescu & Paşca, 2006; Kazama & Torisawa, 2007; Nguyen & Cao, 2008），然而上述研究處理素材都以英文為限，尚未有應用於中文命名辨識的研究。最近黃純敏、李亞哲與陳柏宏（2015）結合維基百科作為輔助中文縮寫詞與同義詞建構的研究，且成效不錯（黃純敏、李亞哲、陳柏宏，2015）。因此本研究也藉助維基語料內容豐富、持續更新之特點，比對其特定結構化欄位，作為協助取詞後驗證與消歧義。

由於中文命名沒有標準語料庫，過去研究紛紛提出個別取詞方法，礙於文長篇幅限制，無法詳列規則，以致難以一一印證結果。為此本研究採有正確答案的訊息理解會議（Message Understanding Conference, MUC-7）競賽中 MET-2 中文測試語料為驗證資料集。考量「時間」命名之規則明確，改善空間有限；「事物」範圍過廣、不包括在 MUC-7 命名實體定義範圍，因此以人名、地名、組織名為研究範圍。過去以句法規則進行命名辨識，即便推導方式較接近人的想法，執行上有效率，但涵蓋面仍無法周全，精確率難以有效提高。此外，其生成規則較為耗時，多依附於語言及特定知識領域，當增加過多判斷規則，可能造成過度擬合（overfitting）的現象，導致命名實體萃取的錯誤。本研究以辨識新聞內文重要組成因素為主，比較著重取出之命名精確性。句法規則以常用的詞性組合為考量，共歸納出人名規則 54 條，地名、組織名規則 11 條。本研究的「人名」命名規則，基於現代人取名多已朝向中性命名，因此不再探討男女姓名區分；為提高辨識率，加入人名語彙鏈公式協助分析。研究結果發現，不惟可精確萃取人名，也可有效提取頭銜，可作為分辨同名不同人，或同一人在不同時期的職位更迭。「地名」、「組織名」則藉由維基百科協助取詞後驗證與消歧義。實驗結果整體精確率達 86.32%、召回率 75.65%，F-Measure（2P&R）為 80.4%。相較於採用句法規則及大量詞庫（Chen & Lee, 1996; Chen et al., 1998）以及 HMM 機器學習（Fu & Luke, 2005）的研究，本研究所提出的規則數量雖少，整體表現十分優異，尤其以精確率最為突出。

本文第二節回顧過去相關研究，第三節描述本研究的研究架構及流程，第四節提出評估方法，檢視所提出的概念成效，並與相關研究比較，最後於第五節提出結論，並對往後研究發展提出未來展望。

貳、文獻探討

一、中文字詞處裡

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。隨著社會演變，新詞不斷出現，未知詞的辨識可說是斷詞處理最棘手的議題。在 Chen 與 Bai（1998）的研究，使用中研院 350 萬字詞語料庫其中 300 萬為訓練樣本，15 萬為測試樣本，執行未知詞偵測（Chen & Bai, 1998）。實驗結果發現絕大多數的未知詞都會被斷成一連串短詞，而每字串都至少有一個單字

詞，亦即多數單字詞可能是未知詞的一部分，複合詞則無法分辨出。Ma 與 Chen（2003）有鑒於之前的未知詞擷取，大多只能處理特定類型，因此限定未知詞組合中至少要有一個以上的單音節語素（Ma & Chen, 2003）；該研究實驗取材於網路文章，以由下而上的合併方法，利用統計、句法決定最小單元語素是否可與鄰近語素結合，結果有 75% 的精確率與 57% 的召回率。

常見的斷詞方法，主要分為詞庫斷詞和統計斷詞法，其中單純詞庫斷詞法是以既有詞庫比對為取詞依據，此法執行速度快、容易，但無法處理新詞。台灣以使用中央研究院 CKIP 斷詞系統最為普遍，該系統除了高效率外，最大的特色之一就是所斷出的字詞附有詞性標註，方便研究者後續剖析處理。在中研院詞庫小組的「中文詞類分析」技術報告手冊中（中文詞知識庫小組，1993），將中文詞分為八種詞類：述詞（V）、非謂形容詞（A）、體詞（N）、副詞（D）、介詞（P）、連接詞（C）、語助詞（T）與感嘆詞（I）。體詞包含了名詞、定詞（例如：這個、其他、一些等）、量詞（例如：一「件」、二「棟」、三「根」）、方位詞、代名詞等。述詞包含動作詞、狀態詞（例如：瀰漫、矗立、擅長等）以及其他有明確訊息卻無主體的詞。其他詞性的字詞大多為修飾、連接、表達語氣或態度的功能，不具特別具體的意義。因此，一般研究多選擇以體詞與述詞兩種詞性，篩選出具有事物概念的字詞集合。由於相同字詞因用法不同，其標註亦不同，因此透過 CKIP 所擷取出的字詞標註，常有一詞多註的現象，而對於未知詞，也仍多混淆、歧異或判斷錯誤（黃純敏等，2015）。因此，斷詞後仍需進一步檢測、分析與修正，以免影響後續處理的正確率。

單純統計斷詞為 N-Gram 的斷詞方式，其概念源自詞頻（Term Frequency, TF），其基本假設為，當詞頻達到某一門檻值時，該詞有較高機率是有意義的。依據不同字元數可分為 Bi-Gram、Tri-Gram... 等，推算至 N-Gram。如：「我昨天去台北」用 Bi-Gram 處理將產生「我昨」、「昨天」、「天去」、「去台」、「台北」等詞組，依此再篩選出詞頻較高的語詞。此法雖然不須仰賴詞庫，但 N 值設定越大，系統運算成本也相對增加，加上高詞頻仍可能為雜訊，仍需再透過人工過濾，因此單純 N-Gram 很少被單獨採用。近年來有些研究加入機器學習法，例如：最大熵（ME）（Xue, 2003）以及條件隨機域（CRF）（Tseng, Chang, Andrew, Jurafsky, & Manning, 2005）等，這些學習演算法多以字元為單元，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，此類演算法相對較為簡易，但正確率不高。有些

研究使用 SVM 訓練，並加入詞典比對，以長詞優先法以及未知詞來增強特徵屬性（Asahara et al., 2005）；也有些研究結合多種學習演算法，以加強斷詞效能，如 Asahara 等（2003）使用 HMM 結合 SVM（Asahara, Goh, Wang, & Matsumoto, 2003）以及 Lu（2005）使用 HMM 結合 TBL（Lu, 2005），實驗結果均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。

二、命名實體辨識

NER 又稱專名辨識，主要為辨識文章中特定的專有名詞，起源於 1990 年代美國國防部高等研究計劃局（Defense Advanced Research Project Agency, DARPA）發起並資助的 Tipster 文件計劃。其中 MUC 主要項目包含：資訊擷取，主題偵測與追蹤（Topic Detection and Tracking, TDT）等。該計畫從 1987 年到 1998 年舉辦七次會議，在第六屆會議（MUC-6）中，加入了 NER 的評比標準與測試集，主要包括中文、英文、日文等三種語言的評測，也定義了命名為人名、組織名、地名、時間、日期、比率、貨幣（Grishman & Sundheim, 1996）。由於 NER 正確率對於後續研究的成效，十分具有影響性，因此 MUC 鼓勵使用新方法改善資訊內容的萃取，並將原始文件轉換成核心資訊，提供資料庫的建立及後續的加值處理，如：資訊檢索（Chen et al., 1998）、自動摘要（Furu et al., 2007）、主題模型的先驗知識。

由於語文具有模糊以及不完全的特性，過去對於 NER 的研究，也有藉由機率模式並加入規則，利用狀態間的轉移機率來處理命名實體，實驗結果發現精確率可達到八成以上，惟在語文方面，多以英文為處理對象，同時缺點為語言模型確定後，倘若運用在不同的領域上，識別效果有限（Ageishi & Miura, 2008; Todorovic, Rancic, Markovic, Mulalic, & Ilic, 2008）。

以句法規則辨識，一般需參考句子文法、本文線索詞、字詞特徵（如：大寫），並結合關鍵字詞典輔助作為擷取命名實體的依據。中文 NER 方法可分成兩類：機器訓練模型與規則辨識。其中機器訓練模型因著重演算法之改善，只需藉由語料庫訓練出語言模型，不需要透過詞庫的方式進行配對，即能達到不錯的成效，因而受到較多關注（Sun, Gao, Zhang, Zhou, & Huang, 2002; Wu, Zhao, & Xu, 2003）。在簡體中文 NER 研究，Fu 與 Luke（2005）以人民日報為實驗素材，先進行雙字元取詞，再用人工方式逐一標註內容為獨立命名體（ISE）、詞組開頭（BOE）、詞組中間（MOE）以及詞組結尾（EOE），如：「溫

家寶 / 總理」被斷成 <BOE> 溫 </BOE><MOE> 家 </MOE><EOE> 寶 </EOE><ISE> 總理 </ISE>，接著利用已標註的語料庫以 HMM 運算，以最大相似估算（MLE）選取條件機率值最高者，進行訓練，藉由相連字詞組成的機率，計算成為命名實體的機率值。研究發現用詞連結比用字連結，效果更好，平均召回率可達 82.24%，精確率為 67.99%（Fu & Luke, 2005）。另一項 Wang 等（2012）簡體中文 NER 研究以 HMM 結合分群與分類兩種方式，實驗結果 F-Measure 可達 76%（Wang, Li, Wong, & Chao, 2012）。

Chen 與 Lee（1996）的研究結合句法規則與統計方法，以四個語料庫（包括：平衡語料庫、新聞集、外國譯名、本國人名）作為訓練資料集，取自由時報六類新聞為測試資料集，針對中文「人名」、「外國譯名」、「組織名」，分別提出了不同的解決方法。其人名辨識以姓氏為基礎，分析其組成方式、以連續三個字元出現的各種機率組合，做為評估基準（baseline），並用線索字（clue words）：頭銜、關聯詞、標點符號、性別詞、詞頻計數，及表達動詞：表示、發言、說…等做為提高召回率的輔助處理。實驗結果發現除了國際、經濟類外，其餘表現都不錯。透過輔助處理，精確率及召回率都有提升，分別為 88.04% 及 92.56%。對樣式較不固定的外來譯名，則以譯名的發音是否符合外來人名的規則做為判別；以 2692 個常用音譯字元進行訓練，但當專有名詞未包括在詞庫者，則無法正確辨識，如立陶宛被視為人名。實驗結果有 20% 外國人名辨識錯誤，14% 筆名、暱稱無法指認，平均精確率僅 50.62%，召回率為 71.93%。其中以社會類及經濟類表現最差，精確率分別僅 6.82% 及 17.11%。組織名的辨識，則透過組織結尾關鍵詞如：省、縣、市、航空公司、電台等，及前置動詞如：前往、來自，飛往等，判斷其命名實體。該研究雖使用大量的關鍵詞協助辨識，但由於組織名組合型態太多，無法一一兼顧，尤其對於新穎的命名實體，多半無法正確指認。組織名精確率平均 61.79%，召回率 54.50%（Chen & Lee, 1996）。

前述研究的接續實驗增加地名、時間、金錢、比率命名實體，並以 MET-2 資料集為辨識驗證，由於該資料集為簡體中文編碼（GB），因此先進行繁體中文編碼（Big5）轉換。在中國人名部分，辨識規則與之前方法相似，考量頭銜（建立 476 個）、句子位置、表達動詞（如：發言、說出）及詞頻記數。其長度限制 2-6 字，從百萬人名庫提取 598 個姓氏，去除罕用姓後保留 541 個，區分男女姓名。外國譯名以特殊字元集為準，再考量頭銜、指稱詞（如：叫做）、特殊動詞、詞間區

隔點（.）、詞頻等進行辨識。未包括在人名庫或頭銜詞集者則無法指認。如：醫生卡庫，因為“醫生”未列入頭銜詞集。實驗顯示外來譯名僅有 50.62% 的精確率與 71.93% 的召回率。地名以 16442 個地名詞典比對，另加入 45 個關鍵詞，如：' 山 ', ' 中心 ', ' 公路 ', ' 以北 ', ' 以西 ', ' 以東 ', ' 以南 ', ' 半島 ', ' 半球 ' 等，此外也藉助特殊動詞，如：來自，前往協助辨識。組織名藉由名稱與組織型態的組合，例如：「台北市」+「政府」，再輔以上下文的關連性、字元的獨立性與詞性特徵、詞頻 2 次以上，作為判斷的依據。並蒐集 776 個名稱與 1059 個關鍵字進行組合。實驗結果，人名、地名、組織名辨識的精確率與召回率分別為（74%，91%）、（69%，78%）、（85%，78%），F-Measure（2P&R）為 77.88%（Chen et al., 1998）。可見即使以大量句法規則辨識，但涵蓋面仍無法周全，精確率難以有效提高。因此，如何隨著時代演進，更新句法分析，以準確辨識文本專有名詞，確實是資訊擷取技術應持續不斷研究的議題。

三、使用維基百科於命名實體辨識

過去有若干以維基百科為語料庫進行命名辨識與消歧義的研究，其中 Bunescu 與 Paşca（2006）藉由維基本文與目標詞目相鄰的敘述語、重定向頁、及內文結構（標題、類別、超連結等），進行 cosine 相似度計算，並以監督式 SVM 進行訓練，實驗結果肯定維基有提升命名消歧義的作用（Bunescu & Paşca, 2006）。其後 Kazama 與 Torisawa（2007）以 CoNLL 2003 新聞語料庫為實驗資料集，採納前述 Bunescu 與 Paşca 所歸納的本文結構為辨識原則，另加入本文開端在“be”之後的第一句名詞片語，以 CRFs 作為結構預測模型。該研究雖藉助維基辨識，然而實驗結果的辨識正確率僅提升 1.21%，效果並不顯著（Kazama & Torisawa, 2007）。Nguyen 與 Cao（2008）的研究則同時考量辨識與消歧義，其基本假設是字詞若同時出現在類似的文章，討論的主題比較相近（Nguyen & Cao, 2008）。該研究以機器學習模型探勘內建的知識庫以擷取命名實體，並藉助網路開源碼工具“KIM ontology”建構命名的知識本體架構，包括上位語、下位語等，結果顯示維基對於消歧義有顯著成效。然而上述研究均以英文為處理考量，鑑於中文命名辨識在巨量資料的知識探勘，扮有引路之重要性角色，激發本研究採用維基內文加入中文命名辨識及消歧義的實驗。

參、研究方法

一、研究架構

基於過去研究是以大量人力釐訂巨量規則並蒐集或自建關鍵詞庫，以期找出所有可能的命名，本研究參考過去文獻，找出研究的切入口，以辨認新聞重要關鍵詞為主，著重精確率的提高。句法規則之釐訂以精簡、常見的詞性組合為主，對於詞性組合之歸納以涵蓋面廣又不失正確為原則。本研究下載 Yahoo! 奇摩新聞網站 2010-2011 年新聞文件共 32,000 篇，隨機抽取 3,000 篇文件，逐篇逐句拆解寫作結構及檢視詞性組合，以人名、地名、組織名為研究範圍，歷經三年討論及修訂而成，共歸納出人名規則 54 條，地名、組織名規則 11 條。由於人名命名雖富有彈性，仍有規則可循，為提高辨識率，因此加入人名語彙鏈公式協助分析；鑒於地名與組織名之詞性及句法規則相似，不易區分，過去多利用蒐集或自建詞庫比對並配合命名規則擷取，本研究則利用維基百科協助驗證詞性組合。實驗結果進行驗證並與過去知名的研究比較。研究架構如圖 1。

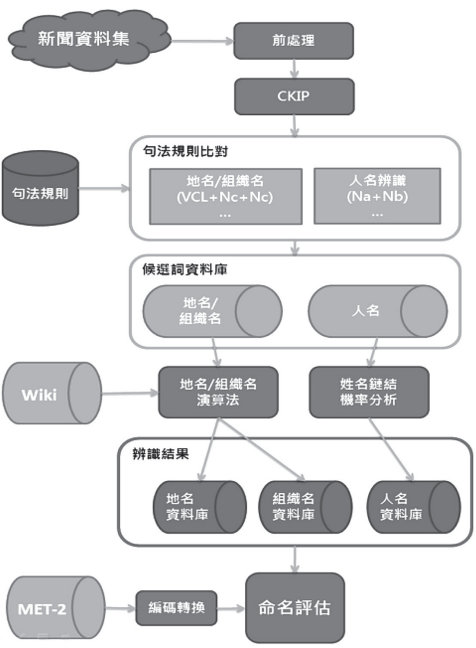

圖 1　研究架構圖

（一）句法規則說明

本研究所有詞性標記均參考中研院詞庫小組的「中文詞類分析」技術報告手冊（中文詞知識庫小組，1993）之註記。基於中文詞類複雜，一般研究多以體詞與述詞及與內容相關的詞性類別作為主要考量。據此，本研究亦從中篩選出常見的名詞、動詞，外加一個對等連接詞為詞性組合元件，使用詞性臚列如表 1。為了解採用詞性數量多寡與辨識結果的關係，在人名前處理階段，先以兩篇新聞做為前測，實驗一保留句子所有詞性進行人名規則比對，實驗二以過濾後詞性進行比對。規則中以“+”作為隔開詞性之符號，也可區隔詞性出現的前後位置；以“→”符號表示規則處理後所取出的詞性；規則中若有重複詞性（如：兩個 Nb），則以“( )”符號標示出所擷取的對象。以下舉例說明：範例 1：「不過黃國昌也回擊」，CKIP 詞性標註為 Cbb+Nb+D+VA，不符合任何人名詞性組合規則（參見表 3），經刪除關聯連接詞（Cbb）、副詞（D）後，留下 Nb+VA，符合 Nb+VA →取 Nb 為人名的規則，因而取出黃國昌。範例 2：「黃國昌卻因為密集跑行程」，CKIP 詞性標註為 Nb+D+Cbb+VH+VA+Na，不符合任何人名詞性組合規則，經過濾詞性後，留下 Nb+VH+VA+Na，符合規則 Nb+VH →取 Nb 為人名的規則，而取出黃國昌。範例 3：「李慶華、黃國昌兩人戰火激烈」，CKIP 詞性標註為 Nb+ PAUSECATEGORY +Nb+Neu+Na+Na+VH，不符合任何人名詞性組合規則，經過濾詞性後，符合 Nb + Nb →取 Nb 為人名的規則，亦順利取出李慶華和黃國昌。

範例 1：不過 (Cbb)　黃國昌 (Nb)　也 (D)　回擊 (VA)
範例 2：黃國昌 (Nb)　卻 (D)　因為 (Cbb)　密集 (VH)　跑 (VA)　行程 (Na)
範例 3：李慶華 (Nb)、(PAUSECATEGORY) 黃國昌 (Nb) 兩 (Neu) 人 (Na) 戰火 (Na) 激烈 (VH)

實驗結果發現以過濾後詞性比對，無論是精確率或召回率均高於用所有詞性比對的表現，如表 2。精確率是指所萃取出的命名實體，正確判斷的結果所佔的比率，精確率愈高表示所提出的方法在檢出及辨識能力愈好；召回率則是指能找出所有正確的命名實體的比率，如果召回率很高，則表示設計考量很周全，不輕易漏掉正確的資訊。

表 1
使用詞性列表

詞性	CKIP 詞類標記	詞性	CKIP 詞類標記	詞性	CKIP 詞類標記
Caa	對等連接詞，如：和、跟	VA	動作不及物動詞	VH	狀態不及物動詞
Na	普通名詞	VAC	動作使動動詞	VHC	狀態使動動詞 /
Nb	專有名稱	VB	動作類及物動詞	VI	狀態類及物動詞
Nc	地方詞	VC	動作及物動詞	VJ	狀態及物動詞
Ncd	位置詞	VCL	動作接地方賓語動詞	VK	狀態句賓動詞
Nd	時間詞	VD	雙賓動詞	VL	狀態謂賓動詞
Neu	數詞定詞	VE	動作句賓動詞	V_2	有
Nes	特指定詞	VF	動作謂賓動詞		
Neqa	數量定詞	VG	分類動詞		

表 2
人名辨識前處理結果（正確姓名數＝ 48）

	辨識總數	正確數	錯誤數	精確率	召回率
實驗一（所有詞性）	38	30	8	78.9%	62.5%
實驗二（常用詞性）	47	43	4	91.5%	89.6%

本研究句法規則演算法如圖 2，步驟說明如圖 3。一般而言，句法規則所包含的詞性組合愈精簡，涵蓋面愈廣，召回率愈高，然而精確率也將隨之降低，以單純 Na+Nb 人名規則為例，雖可順利提取許多人名，但遇到大象（Na）林旺（Nb），則無法判別當中的「林旺」並非人名，必須仰賴前後輔助詞性以資判別。依此，倘若延伸過多詞性，又將造成過度擬合的現象，導致每條規則只適用少數情況，失卻規則訂定的本意。因此規則長短的合理性，需不斷測試以取得最佳的狀態組合。

Input: sentence_set, entity_rule_set, and entity_result_set
Output: extraction_list
Procedure: Rule based name entity algorithm
Pattern with length **n**
(1) Extract the terms from either side of pos
(2) For each sentences S_j in sentences_set do
(3) For each terms t_{i+x} in S_j do
If terms t_i exists in entity_rule_set then
repeat /*Once an entity is found, exit for loop*/
Examine if the term t_{i+x} is an entity pos
Set x to be equal to x + 1.
until the rule match and an entity is found
End if
(4) Insert the results of algorithm into entity_result_set
(5) End for
End for

圖 2　NER 句法規則演算法

(1) 依序讀取已標註詞性的文件集之文件
(2) 以標點符號分段，依序讀取文件句子
(3) 先輸入一個詞性 t_i，將第一個詞性與現有的規則資料集進行比對，檢查該詞性是否為規則資料集，某個規則的第一個詞性，若不符合即跳過換下一個詞性 t_{i+1})，若找到符合的規則的詞性則依序輸入詞性，由該字元起向後比對 X 個字，直到找到最長的規定值為止
(4) 依據規則回傳命名實體結果並存入資料庫
(5) 重覆以上步驟，直到沒有新的文件

圖 3　NER 句法規則執行步驟

(二) 人名命名規則剖析

人名規則可歸納為三大組合，分別是以 Na+Nb 為中心者 4 條；以 Nb+Nb 為中心者 11 條；以 Nb+V 為中心者 38 條，另外有 1 條連接詞

之組合。以 Na+Nb → Nb 為例，此處 Na 為普通名詞，Nb 為專有名稱，當 Na 直接接續 Nb 時，Na 則多為身分或職稱，Nb 則為人名。例如：「總統（Na）馬英九（Nb）」，可取得「馬英九」為人名。如再往前後延伸詞性，可視為人名註解。比如，當 Nc+Na+Nb，此時 Nc 多為描述職稱的所在機構，而非地名，例如：「總統府（Nc） 發言人（Na）陳以信（Nb）」之「總統府（Nc）」主要為描述陳以信工作的組織。當 Nc 接在 Nb 之後，此時 Nc 絕大多數為組織名，如：「甘迺迪（Nb）國際（Nc）機場（Nc）」，少數為姓名之名，如「陳（Nb）雲林（Nc）」、「林（Nb）濁水（Nc）」。人名命名辨識規則及範例如表 3。本研究所訂定的人名詞性組成，若干看似有包含的關係，每一項卻皆有其存在的意涵，不能歸納成一項。如：Na+Nb 包含在 Nb+Na+Na+Nb 之中，若只考量 Na+Nb，將則只擷取：「得主（Na） 劉曉波（Nb）」，雖可取得「劉曉波」為人名，卻無法獲得：「諾貝爾和平獎得主劉曉波」全貌。透過多重詞性組合比對，不但可萃取其工作崗位、職稱組合，更能提升萃取人名特徵字詞的精確率。又如：Nb+Nb 包含在 Nb+Nb+V 及 Nb+Nb+Nb 之中，若未逐一檢定其後所接續的詞性，將無法擷取所有的人名。再者：Nb+VJ+Nb+VH（「杰夫（Nb）延續（VJ）普京（Nb）務實（VH）」），此處將擷取前後兩個 Nb 為人名；但是 Nb+VJ+Nb+VH+VC（「林瑟肯（Nb）發揮（VJ）賽揚（Nb）強（VH）投（VC）」），則只有前者是人名。可見中文句法因接續的詞性不同，指稱亦將變異。

表 3
人名命名實體辨識部分規則範例

	規則	範例
Na + Nb	Na+Nb → Nb	總統 (Na) 馬英九 (Nb)
	Nc+Na+Nb → Nb	總統府 (Nc) 發言人 (Na) 陳以信 (Nb)
	Nb+Na+Na+(Nb) → Nb	諾貝爾 (Nb) 和平獎 (Na) 得主 (Na) 劉曉波 (Nb)
	Nc+Na+Na+Nb → Nb	國防部 (Nc) 新聞 (Na) 發言人 (Na) 楊宇軍 (Nb)
	Nb+(Nb) → Nb	民進黨 (Nb) 蔡英文 (Nb)
	Nb+(Nb)+VC → Nb	國民黨 (Nb) 朱立倫 (Nb) 推動 (VC)

Nb + Nb	Nb+(Nb)+VG+VHC → Nb	國民黨 (Nb) 邱毅 (Nb) 做為 (VG) 團結 (VHC)
	Nb+(Nb)+VE+VE → Nb	國民黨 (Nb) 洪秀柱 (Nb) 想 (VE) 說 (VE)
	Nb+Nb+Nb → each Nb	蔡依林 (Nb)、許茹芸 (Nb)、王心凌 (Nb)
	Nb+Nb+Nb+Nb → each Nb	趙天麟 (Nb)、洪平朗 (Nb)、黃淑美 (Nb)、康裕成 (Nb)
	Nb+Nb+Nb+Nb+Nb → each Nb	金允珍 (Nb)、韓惠珍 (Nb)、李敏貞 (Nb)、孔孝珍 (Nb)、崔江姬 (Nb)
	Nb+(Nb+Nb)+VE+(Nb)+VC+VA → (Nb+Nb)；Nb	武騎常侍 (Nb) 司馬 (Nb) 相如 (Nb) 澄清 (VE) 黃崇右 (Nb) 接受 (VC) 應訊 (VA)
	Nb+(Nb+Nb)+VH+VCL → (Nb+Nb)	武騎常侍 (Nb) 司馬 (Nb) 相如 (Nb) 離職 (VH) 回到 (VCL)
	Nb+(Nb)+VE → Nb	象隊 (Nb) 周思齊 (Nb) 表示 (VE)
	Nb+(Nb)+VE+VH → Nb	特助 (Nb) 郭德志 (Nb) 獲判 (VE) 無罪 (VH)
Nb + 動詞	Nb+VCL → Nb	王世堅 (Nb) 到 (VCL)
	Nb+VG → Nb	柯文哲 (Nb) 就任 (VG)
	Nb+VH+VC → Nb	徐耀昌 (Nb) 也 (D) 特別 (VH) 為 (P) 裁員 (VC)
	(Nb)+VC+Nb → Nb	朱立倫 (Nb) 拿 (VC) 諾貝爾獎 (Nb)
	Nb+Nc+VE → Nb	曾薏蘋 (Nb) 台北 (Nc) 報導 (VE)
	Nb+VA → Nb	吳敦義 (Nb) 答詢 (VA)
	Nb+VA+VC → Nb	劉曉波 (Nb) 得獎 (VA) 發表 (VC)
	Nb+VA+VC+VC → Nb	張志軍 (Nb) 來訪 (VA) 做出 (VC) 回應 (VC)
	Nb+VA+VCL → Nb	范永奕 (Nb) 捷足先登 (VA) 抵達 (VCL)
	Nb+VA+VG → Nb	李建昇 (Nb) 應訊 (VA) 稱 (VG)
	Nb+VC → Nb	姚啟鳳 (Nb) 接受 (VC)
	Nb+VC+Nb+VA → Each Nb	鍾承佑 (Nb) 取代 (VC) 林智勝 (Nb) 上場 (VA)
	Nb+VC+VA → Nb	鄭秀文 (Nb) 推薦 (VC) 刮痧 (VA)
	Nb+VC+VA+VE → Nb	梁光烈 (Nb) 發表 (VC) 講話 (VA) 闡述 (VE)

Nb + 動詞	Nb+VC+VC → Nb	孫振宇 (Nb) 接受 (VC) 訪問 (VC)
	Nb+VC+VE → Nb	曾銘宗 (Nb) 補充 (VC) 指出 (VE)
	Nb+VC+VH+VC → Nb	李健熙 (Nb) 接受 (VC) 鎮定 (VH) 治療 (Na)
	Nb+VC+VH+VH → Nb	茱莉亞羅勃茲 (Nb) 創 (VC) 久違 (VH) 賣座 (VH)
	Nb+VC+VJ+VE → Nb	曾銘宗 (Nb) 補充 (VC) 歡迎 (VJ) 報導 (VE)
	Nb+VCL+VC → Nb	梁英斌 (Nb) 親臨 (VCL) 主持 (VC)
	Nb+VD+VA → Nb	陳介山 (Nb) 頒發 (VD) 認證 (VA)
	Nb+VE → Nb	郭添財 (Nb) 指出 (VE)
	(Nb)+VE+(Nb) → Each Nb	田秋堇 (Nb) 提醒 (VE) 李述德 (Nb)
	Nb+VE+Nb+VE → Each Nb	陳淞山 (Nb) 轉述 (VE) 陳水扁 (Nb) 主張 (VE)
	Nb+VE+VC → Nb	李念龍 (Nb) 坦承 (VE) 收養 (VC)
	Nb+VE+VH → Nb	郭德志 (Nb) 獲判 (VE) 無罪 (VH)
	(Nb)+VF+(Nb) → Each Nb	馬英九 (Nb) 勸 (VF) 陳長文 (Nb)
	(Nb)+VF+VB+(Nb)+VH → Each Nb	陳士華 (Nb) 打算 (VF) 先 (D) 道歉 (VB) 再 (D) 等 (P) 李念龍 (Nb) 息怒 (VH)
	Nb+VH → Nb	劉曉波 (Nb) 獲獎 (VH)
	Nb+VH+VC+VC → Nb	劉曉波 (Nb) 獲獎 (VH) 接受 (VC) 祝賀 (VC)
	Nb+VH+VC+VJ+VH+VC+VC → Nb	劉曉波 (Nb) 欣喜 (VH) 接受 (VC) 歡迎 (VJ) 獲獎 (VH) 祝賀 (VC) 表彰 (VC)
	(Nb)+VH+VC+(Nb)	李念龍 (Nb) 欣喜 (VH) 祝賀 (VC) 李小龍 (Nb)
	Nb+VJ → Nb	馬英九 (Nb) 延續 (VJ)
	(Nb)+VJ+(Nb)+VH → Each Nb	杰夫 (Nb) 延續 (VJ) 普京 (Nb) 務實 (VH)
	(Nb)+VJ+Nb+VH+VC	林瑟肯 (Nb) 發揮 (VJ) 賽揚 (Nb) 強 (VH) 投 (VC)
	Nb+VJ+VCL → Nb	王議賢 (Nb) 歡迎 (VJ) 前往 (VCL)
	Nb+VK → Nb	吳葶萱 (Nb) 希望 (VK)

	Nb+VK+VA → Nb	黃錫薰 (Nb) 涉嫌 (VK) 串供 (VA)
*	(Nb)+Caa+(Nb) → Each Nb	陳大同 (Nb) 和 (Caa) 張清芳 (Nb)

註 * Caa 連接詞組合

由於 CKIP 只能辨認部分知名人士，並給予詞性 Nb，如：馬英九（Nb），否則多將人名拆解為 2-3 個部分，如：張（Nb）花冠（Na）、陳（Nb）曉明（Nb）、林（Nb）濁水（Nc）、蔣明（Nb）恩（Na）、李（Nb）應（D）全（Neqa）。此外，Nb 也可扮演多種角色，它可能是部落名（如：阿美族）、或是歷史事件（如：二二八事件）等，透過本研究所提出的人名命名規則搭配人名組合公式（1），可有效辨識。例如：李大雄（Nb）獨（D）愛（VL）哈士奇（Nb）玩偶（Na），在去除 [獨（D）] 不常用詞性後，當中的「哈士奇」雖為 Nb，但不符合句法規則，因此不會將之認定為人名。

由於人名命名有很大的彈性空間，沒有任何人名辭典可將全部姓名全部囊括，通常以百大姓氏開頭，連接一至二個名字，如：李鵬、湯民國；而外譯姓名則無特別字元長短，但存在著特徵標註，如「巴拉克・歐巴馬」以音界號的方式呈現。過去研究多以 2-6 字為考量，然而根據內政部 103 年統計的姓氏統計數量為 1510，其中單姓 1396（92.45%），複姓 114（7.55%），三字姓和四字姓的數量為零，在姓氏之後，人名出現的頻率多為二個字，因此姓名字數集中在 2-4，少數為 5-6，逾 6 字者，僅千餘人（內政部，2014）。在 Cheung、Tsou 與 Sun（1995）的調查也指出大陸與香港人名組合以單姓和複名為主，其分配如表 4、表 5。基於上述理由，本研究在人名部分僅考量 2-4 字元為研究對象，也大為縮小計算成本。

表 4
中文單複姓分配統計

	大陸（%）	香港（%）
單姓	99.92	99.56
複姓	0.08	0.44

表 5
中文單複名分配統計

	大陸（%）	香港（%）
單名	29.07	2.13
複名	70.93	97.87

人名語彙鏈演算法如圖 4。首先匯入已斷詞的句子（SentencePOS），依序提取出字詞的詞性（termPOS t_i），進行人名規則（person_rule_

set）比對。此演算法係參考顯著估算 Significance Estimation（SE）公式修改，該公式是求兩個子字串的最大重疊交集。本研究加入頻率考量，如果某一字詞詞性為 Nb 且為單一或二個字元，則進行與後續字元結合的頻率計算，如公式（1），以擷取可能之姓名鏈結組合。

NER_person algorithm

Input: SentencePOS and person_rule_set
Output: personName

(1) Function_person(SentencePOS, person_rule_set)
(2) **For**(each termPOS ti in SentencePOS) **do**
(3) **If**(termPOS t_i is Nb) then
(4) **If**(t_i.length=1) then //t_i is a person name candidate
(5) **If**(Frequency Lc(t_i,t_{i+2})/Frequency Lc(t_i,t_{i+1}) >=1) **then**
(6) return Lc((t_i,t_{i+2}) as personName
(7) **else**
(8) return Lc(t_i,t_{i+1})as personName
(9) **else If**(t_i length=2) **then**
(10) **If**(Frequency Lc(t_i,t_{i+1})/Frequency Lc(t_i) >=1) **then**
(11) return Lc(t_i,t_{i+1}) as personName
(12) **else**
(13) return Lc(t_i)as personName
(14) **else**
(15) return t_i as personName

圖 4　人名語彙鏈演算法

$$P(PER)=\frac{Freq\ Lc(i,i+2)}{Freq\ Lc(i,i+1)}\ if\ \begin{cases}P(PER)<1, PER\in Lc(i,i+1)\\ \overline{P(PER)<1, PER\in Lc(i,i+2)}\end{cases},\ i=0 \qquad (1)$$

上述公式中 *P(PER)* 為人名組成機率。*Freq Lc(i,i+1)* 代表長度為單一或二個字元的 Nb 與其後接續第一個字元結合之詞頻，*Freq Lc(i,i+2)* 表示與其後接續的二個字元結合之詞頻。當 Lc(0,2) 出現詞頻小於 Lc(0,1) 時，*P(PER)*<1，表示應結合後一字元為人名；反之，就表示應結合後兩字元為人名。如果 Nb 長度不為 1 或 2 時，那就表示此 Nb 可直接判定為人名。例如：[司機（Na）李（Nb）應（D）全（Neqa）]，由於此處 Nb「李」僅為一字元，必須與後置一到兩字元組合，當「李

應全」（Lc(0,2)）出現頻率小於「李應」（Lc(0,1)）時，人名判定為「李應」（Lc(0,1)）。反之，則取「李應全」（Lc(0,2)），組合範例如表 6。以此項人名姓名結合法，可成功解決 CKIP 無法處理的部分未知詞，如：「縣長（Na）張（Nb）花冠（Na）」、「縣長（Na）李（Nb）朝（P）卿（Nh）」。前述「陳（Nb）雲林（Nc）」也可透過此公式合併出人名「陳雲林」。

表 6
中文姓名組合範例

姓氏組合	範例	可能組合	結果	符合判斷式
單姓	李	單姓 + 單名	李鵬	P(PER)<1
	湯	單姓 + 雙名	湯民國	P(PER) ≥ 1
複姓（雙姓）	諸葛（黃陳）	複姓（雙姓）+ 單名	諸葛亮（黃陳嬌）	P(PER)<1
	司馬（陳張）	複姓（雙姓）+ 雙名	司馬玉嬌（陳張宏）	P(PER) ≥ 1

（三）地名、組織名命名規則剖析

地名為一個地區的名稱，其範圍相當廣泛，可分為自然地理的山脈、海洋、湖泊與河流，人文地理的部分由一地方名稱加上地理行政單位所組成，如國家、城市、省份、縣市、鄉鎮、機場、基地等。相較於人名而言，地名與組織名的組成更無規則可循。Chen、Ding 與 Bian（1998）的研究在地名的辨識上利用地理關鍵字如山、中心、以南、以東、以西、市、鄉等，然而中文字詞在組合上有很大的隨意性，如返鄉、偏鄉，儘管詞語中都有包含「鄉」，但都不是地名，有些外國地名或是本國地名在作者撰寫文章時不一定會加註其行政單位如「多倫多市」或者「台北市」。如前述，人名常以地名命名，同樣的，地名也經常因名人命名。如：「甘迺迪機場」，CKIP 將之標註為「甘迺迪（Nb）機場（Nc）」，視為兩個獨立名詞，無法正確指認出標的物，失卻成為複合詞的意義。

對於組織名的辨識，Chen 等（1998）的研究以名稱結合組織型態，再輔以上下文的關連性、字元的獨立性與詞性特徵指認。然而此種做法並無法臚列所有規則，由於關鍵字變化很大，隨著時間的推移、不同的資料集、縮寫詞、別稱等而有所不同，很難正確指認。若

因此增加過多判斷規則，則可能產生過度擬合的結果，因此精確率及召回率難以提高。本研究觀察「組織名」與「地名」與上下文關係，認為其組成仍有模式可循。不僅詞性都是 Nc，句法規則也十分相似，多用於特定動詞之後或是與名詞（Na、Nb、Nc）的串連而形成。因此基本上可憑藉其與所接續的特定動詞（VC、VCL）或是串連的名詞（Na、Nb、Nc）進行擷取，如：「領導 阿根廷」及「帶領 長榮」都符合「（VC+Nc）→ Nc」、「直抵 台北」及「前往 教育部」都符合「（VCL+Nc）→ Nc」，然而前者為地名；後者為組織名。又如：「甘迺迪 機場」為（Nb+Nc）、「美國 職棒 大聯盟」為（Nc+Na+Na）、「台北 市府 轉運站」為（Nc+Nc+Nc），都可藉由合併名詞詞性取得地名或組織名。本研究歸納二者句法規則共 11 條，可分為三類，分別是 2 條與特殊前置動詞相關、8 條與合併名詞相關和 1 條與連接詞相關的規則 1 條，範例如表 7。然而同樣的規則其配對結果可為地名或組織名，故仍須搭配地名組織名演算法來加以指認。

表 7
地名、組織名辨識規則範例

	規則	類別	範例
動詞 + Nc	VC+Nc → Nc	地名	領導 (VC) 阿根廷 (Nc)
		組織名	帶領 (VC) 長榮 (Nc)
	VCL+Nc → Nc	地名	暢遊 (VCL) 台東 (Nc)
		組織名	前往 (VCL) 聯合國 (Nc)
合併連續名詞	(Nb+Nc)	地名	溫布利 (Nb) 足球場 (Nc)
		組織名	國民黨 (Nb) 縣黨部 (Nc)
	(Nb+Nc+Nc)	地名	奧黑爾 (Nb) 國際 (Nc) 機場 (Nc)
		組織名	中共 (Nb) 中央 (Nc) 政治局 (Nc)
	(Nc+Na)	地名	東勢 (Nc) 客家庄 (Na)
		組織名	新竹 (Nc) 縣政府 (Na)
	(Nc+Na+Na)	地名	亞洲 (Nc) 傳統 (Na) 市場 (Na)
		組織名	美國 (Nc) 職棒 (Na) 大聯盟 (Na)
	(Nc+Na+Na+Nc)	地名	楊梅鎮 (Nc) 陽光 (Na) 生活 (Na) 藝術館 (Nc)
		組織名	台北 (Nc) 商業 (Na) 技術 (Na) 學院 (Nc)

	(Nc+Nc)	地名	澎湖 (Nc) 馬公 (Nc)
		組織名	中區 (Nc) 國稅局 (Nc)
	(Nc+Nc+Na+Nc)	地名	台北 (Nc) 德國 (Nc) 文化 (Na) 中心 (Nc)
		組織名	中國 (Nc) 長城 (Nc) 工業 (Na) 總公司 (Nc)
	(Nc+Nc+Nc)	地名	台北 (Nc) 市府 (Nc) 轉運站 (Nc)
		組織名	財政部 (Nc) 基隆 (Nc) 關稅局 (Nc)
*	(Nc)+Caa+(Nc)	地名	台灣 (Nc) 與 (Caa) 德國 (Nc)
		組織名	交通部 (Nc) 和 (Caa) 經濟部 (Nc)

註 * （Caa 連接詞組合）

（四）維基百科辨識地名、組織名

由於維基百科為自由編輯之百科全書，新的命名實體產生後隨即透過知識共享的方式可新增編撰或是修改調整，彙集眾智能把潛在錯誤迅速糾正。其質量已為眾所肯定。其編排結構使用表格式「資訊盒」（Infobox）及開放類別（Open Category），記載結構化的資訊，例如：地名的特性描述為：成員國、聯邦、島國……等，組織名的特性描述為：組織、成立的公司、行政部門、機構……等，茲整理如表 8、表 9。本研究萃取相關欄位，並進行比對。演算法如圖 5，首先輸入候選詞至演算法當中，每一個候選詞將會與其對應的維基百科頁面欄位進行比對，如果所檢測的候選詞頁面包含地名特徵則認定該候選詞為地名，其他則認定為「組織名」。

表 8
維基百科資訊盒列表範例

命名實體	資訊盒列表項目				
人名	性別	出生	國籍	政黨	…
地名	人口	面積	首府	行政區	…
組織名	總部	成員	公司類型	成立	…

表 9
維基百科開放類別列表範例

命名實體	開放類別列表項目				
人名	在世人物	黨員	校友	出生	…
地名	成員國	聯邦	島國	行政區	…
組織名	組織	成立的公司	行政部門	機關	…

NER_place&organization algorithm

Input:place&organization_candidate
Output:placeName or organizationName
1: Function_place_organization_entity(place&organization_candidate)
2: **For**(each place&organization_candidate) **do**
4: **HtmlInfo=getDocFromURL("http://zh.wikipedia.org/wiki/"+term** T_i)
5: **If**(HtmlInfo contain place_info_set) then
6: **return** placeName
7: **else If**(HtmlInfo contain organization_info_set) then
8: **return** organizationName
9: **else**
10: **return** organizationName

圖 5　維基百科協助辨識地名組織名演算法

(五)NER 評估

MET-2 語料集主要新聞文章為中國國際廣播電台及新華社的報導，時間為 1994-1996 的文章，大多描述國外的事件。此資料集提供命名實體標準答案檔（稱為 key），可供實驗比對。其標示如下：
＜ENAMEX TYPE= "LOCATION" ＞莫斯科＜/ENAMEX＞
＜TIMEX TYPE= "DATE"＞1月3日＜/TIMEX＞電＜TIMEX TYPE=“DATE”＞今
＜/TIMEX＞＜TIMEX TYPE= "TIME" ＞晨＜/TIMEX＞
＜ENAMEX TYPE= "LOCATION" ＞俄羅斯＜/ENAMEX＞

由於該資料集為簡體中文編碼（GB），本研究為進行繁體中文研究，故需先轉換成繁體中文編碼（Big5），但是除了編碼之外，簡繁

體對同一事物甚至是專有名詞常有不同的表達，因此也需要轉換，否則會有斷詞錯誤的情況產生，若僅轉換編碼，俟後進行斷詞可能產生不同的結果。如：「斯威士兰」經轉換繁體再斷詞的結果為「斯威士（Nb）蘭（Na）」，與「史瓦濟蘭（Nc）」相去甚遠。本研究曾測試 Fu、Qin 與 Liu 所使用的 Stanford-NER 的轉碼功能，該工具以 Parallel Corpus 中文的語料庫，據稱透過有效的訓練可優化訓練模型提升辨識率（Fu, Qin, & Liu, 2011），但經過測試發現錯誤率其實很高。最後採用兼具轉碼及轉表述兩項特點的微軟 AppLocale Rightkey 1.1 進行字碼轉換，再進行斷詞，並以標點符號『，』、『。』、『？』作為斷句的依據。為進行後續句法分析，保留詞性標註。

肆、實驗評估與討論

一、命名辨識結果

(一) 人名

本項實驗透過句法規則、外譯姓名音界號及長詞優先之條件，進行人名辨識，如：「副省長（Na）馬麟（Nb）」、「交通部（Nc）副部長（Na）劉松金（Nb）」及「傑里（Nb）·奧利森（Nb）」、「格雷格（Nb）·庫克（Nb）」等。研究發現外國音譯人名，儘管利用明顯的「音界號」協助判斷，但由於音譯人名長度不固定，只有部分找回。此外，由於 CKIP 對於不普遍為人所熟知的外譯語詞之詞性，多一律給定 Nb，如：「城市（Na）卡杜納（Nb）」。此處「卡杜納」為 Nb，因字元超過二字元，且無其他輔助詞性可資分辨，因此依據句法規則 Na+Nb → Nb 誤判了「卡杜納」為人名。

至於應用姓名鏈結機率公式，確實可達到有效偵測姓名組合，如：「記者（Na）賈（Nb）燕京（Nc）」，經由結合姓名配對後，成功辨識該人名為「賈燕京」，不因「燕京」是 Nc 而誤判其為地名。過去 Chen 等的研究因姓氏未涵蓋在辭庫，而無法判別者，如：「庄霞琴」以及姓名誤判為譯名，如：溫克剛→溫克、賈西平→賈西（Chen & Lee, 1996），也可透過本研究的詞性分析「庄（Nb）霞琴（Na）」、「溫克（Nb）剛（D）」「賈西（Nb）平（VH）」，而正確取得姓名。實驗結果顯示有結合姓名鏈結機率配對者，無論在精確率或召回率都高於無結合配對，尤其精確率提升最多，統計數據如圖 6。

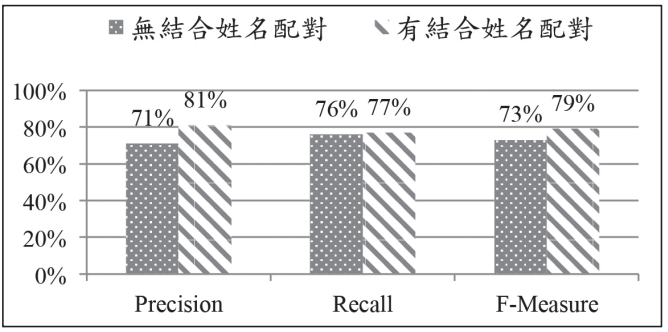

圖 6　有無結合姓名鏈結機率結果

（二）地名

在本研究中，地名為單詞者，占全部地名命名實體句法規則 70%，其中 Na 占 18%、Nb 占 1%、Nc 占 81%，如：墨西哥城（Na）、溫布利（Nb）、以色列（Nc），詞性分布如圖 7。在所有標示 Nb 詞性者，有接近兩成的比例為地方名如：「盧旺達（Nb）這（Nep）個（Nf）國家（Na）」、「東非（Nc）小（VH）國（Na）布隆迪（Nb）」、「溫布利（Nb）足球場（Nc）」。以本研究的人名詞性組合規則將排除上述三詞為人名的可能性，而列入地名候選詞。經檢視地名為雙詞組合者，以（Nb+Nc）出現頻率不少，如上述「溫布利（Nb）足球場（Nc）」及「賈夫納（Nb）半島（Nc）」、「盧塞（Nb）納港（Nc）」等，由於 Nb+Nc 也可能是人名組合，需再透過 Nb 字元數加以分辨。至於連續名詞，如：「肯尼迪（Nb）國際（Nc）機場（Nc）」，可藉由合併 Nc 辨識為地名，至於無 Nc 的連續名詞，如：「哈巴（Na）羅（Nb）夫斯克（Na）」則難以判斷為地名，可能會透過人名公式誤判「羅（Nb）夫斯克（Na）」為人名「羅夫斯克」。實驗評估顯示地名辨識精確率超過九成，召回率也幾近八成，說明了本研究提出的句法規則，確實發揮成效，結果如圖 8。

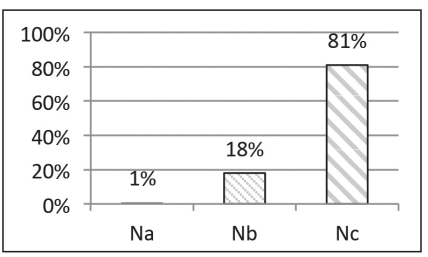

圖 7　地名單詞出現比例

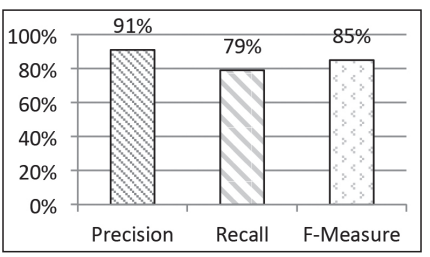

圖 8　地名辨識結果

（三）組織名

過去 Chen 與 Lee（1994）曾使用形態規則和上下文來識別組織的名稱，礙於組織名稱極不具有規則性，該實驗結果獲致精確率及召回率只有五、六成（Chen & Lee, 1994）。鑒於組織名與人名同樣具有多型及彈性的特性，卻又更無規則可循，以致想藉由簡易規則提高精確率及召回率，似乎難以體現。此外，組織名與地名因具有同樣詞性 Nc，更增加二者的分辨難度。有鑑於此，本研究僅訂定基本地名命名規則，希藉由簡易規則所萃取出的候選詞，透過維基百科協助辨識組織名。經統計組織名以 Nc 詞性最多，占 86%，單詞僅占全部組織名 27%，多數為連續 Nc，如，連續兩個名詞：「聯合國（Nc）」、「公安部（Nc）」，至連續多個名詞，如：「中國（Nc）航天（Na）工業（Na）總公司（Nc）」、「國務院（Nc）台灣（Nc）事務（Na）辦公室（Nc）」等。雖由不同名詞組成，最後一個名詞都是 Nc；與地名不同之處，在於 Nb+Nc 之組合例子不多。

維基對於組織名的縮寫詞，還提供了重定向頁，藉此可擷取其全稱，如：「北約」一詞，可擷取「北大西洋公約組織」頁面。經由維基內容比對所獲致之結果，可協助判定為地名或組織名。結果顯示精確率可達八成五，召回率也有七成。此項結果說明了即使單純依附地名命名規則辨識，透過維基百科協助驗證，也可獲致優異的成效。如圖 9。

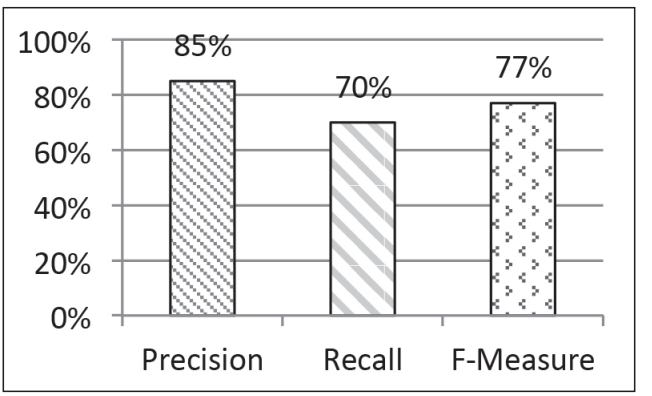

圖 9　組織名辨識結果

二、實驗結果比較

由於過去對於命名實體辨識已有不少研究成果，相較於 Chen 與 Bai（1998）基於中研院語料庫對未知字檢測的學習，自動建構規則的研究（Chen & Bai, 1998）在初始階段彙整出 1,455,633 條規則，於刪除出現率小於 3 的情況，共保留 215,817 條規則。本研究所歸納出的以詞性結合為主的句法規則十分精簡，僅包括人名規則 54 條、地名規則 11 條。Chen 等（1998）的命名實體研究（Chen et al., 1998），人名的部分從一百萬個人名的資料庫當中取出 541 個中文姓氏作為基礎並計算姓氏以及名字之間的組成機率，然而人名的組合容易隨著時代的演進而有不同的組合方式，加上對於中文姓氏並不容易指認音譯的人名，甚至有誤判的可能，因此該研究對於音譯人名又另外採用 280 個起始字元，然而外文姓名可能僅稱呼姓、名、或全稱，為提高召回率勢必將又採用更多的起始字元。

本研究在人名部分，除了基本詞性組合規則，並藉由姓與名之鏈結關係，發掘可能之組合，實驗以最常出現的 2-4 字元為實作考量，利用此人名組合公式，可在不增加句法規則的計算負擔下，增加原有規則的包覆範圍，對於音譯的名詞，藉由其專有名詞之詞性標註，仍可有效的偵測頭銜、組織，藉以分辨同名不同人，未來亦可依據時間先後推估同一人職位之變動，因此讓規則更具有延展彈性及正確性。

在地名部分，Chen 等（1998）在地名方面的研究採用了 45 個地名的關鍵字如：山、中心、公路等，然而這些關鍵字有可能只是一般普通名詞，而非指特定地名，如：火山、城鎮中心、快速公路等。音譯地名也可能不會包含上述關鍵字在其中。本研究發現外國地名音譯結果，結構十分類似人名，對此 CKIP 多給予 Nb 的詞性，因此，若無搭配句法分析，很可能被誤認為人名，藉由本研究的精簡規則，即可有效分辨出。此外，該研究在組織名的部份使用 776 個組織名以及 1059 組織關鍵字協助分辨，基於組織名增加速度太快，未來維護除了要不斷增加關鍵字之外，又要避免關鍵字與地名關鍵字形成重複，將是一大難題與負擔。

相較於採用人工方式逐一標註內容，再處理 HMM 運算的研究（Fu & Luke, 2005），其優點是，只需要藉由人工建立的語料庫訓練出的語言模型，即能達到辨識命名實體的成效，不需要透過詞庫的方式進行配對。其缺點為語言模型確定後，倘若運用在不同的領域上，識別效果有限。本研究則不需要額外的訓練成本，適用於不同類型文章，且

使用維基百科語料庫，以 Web2.0 精神，其知識庫建置會隨著群眾的智慧編輯而增長並持續修正，以達到協助提升命名實體辨識正確率的效果。

圖 10 顯示本研究的人名辨識精確率達 81%，明顯高於其他兩個研究，召回率較低的原因，推測是本研究規則十分精簡，以致無法網羅所有獨特案例，卻也達到 77% 不錯的水平。為有效分辨詞性與地名相近但變異性更大的組織名，我們借力使力，透過維基百科資訊盒和開放類別中定義之特徵協助指認。圖 11 顯示本研究雖然只列出簡單的 11 條地名，而辨識地名的精確率卻高達 91%，遠超過其他兩個研究；召回率也與兩個研究相近；辨識組織名的召回率略遜，但精確率也達 85%，與 NTU System 同，如圖 12。

圖 13 為實驗成果整體比較圖表，由於我們無法獲得其他研究在人名、地名與組織名的個別數量，因此僅以三種命名實體的結果直接加以平均，雖是如此，亦可看出本實驗在精確率與召回率分別達 85.67%、75.33%，F-measure 為 80.33%，與自動建構大規模規則的判斷研究（Chen et al., 1998）及採用人工標註結合 HMM 機器學習的研究（Fu & Luke, 2005）之成果比較，本研究的召回率仍算不錯，尤其以精確率最為突出。

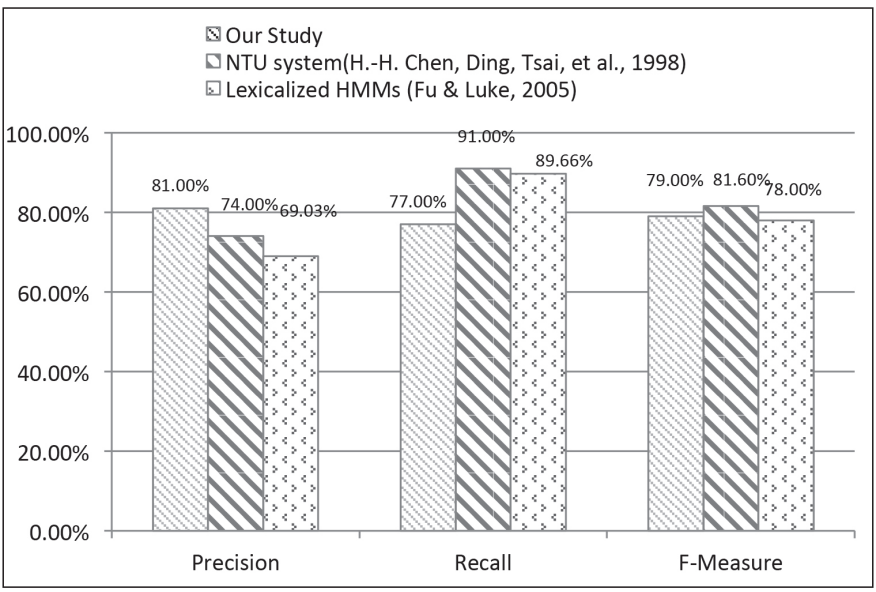

圖 10　人名命名實體辨識比較

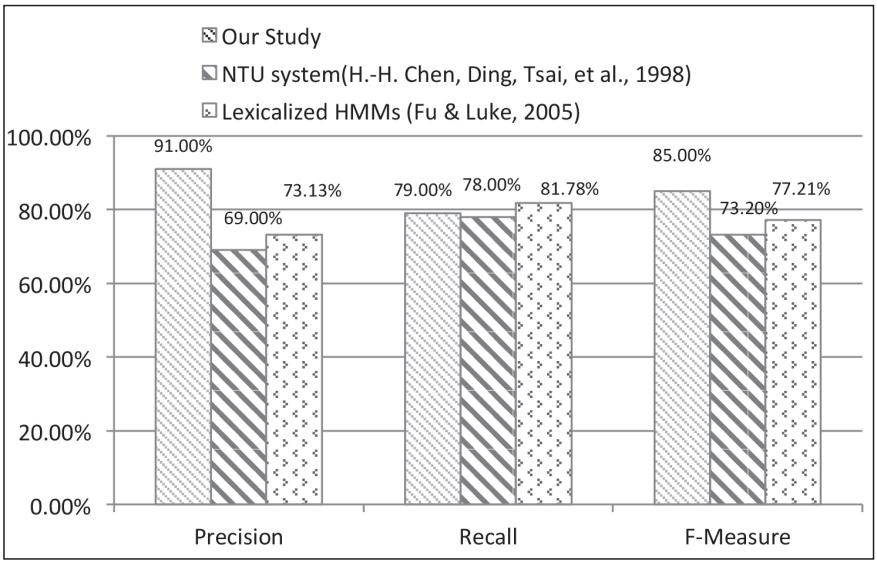

圖 11　地名命名實體辨識比較

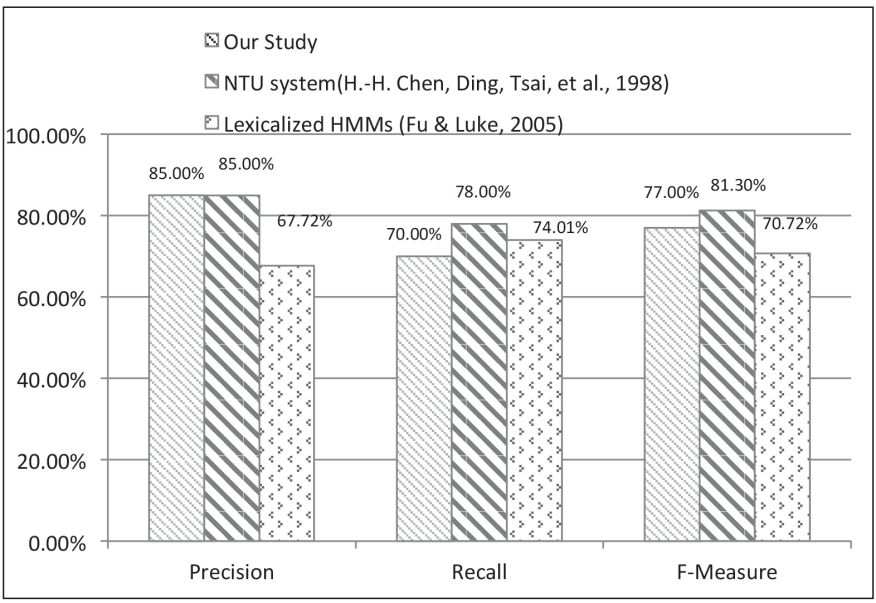

圖 12　組織名命名實體辨識比較

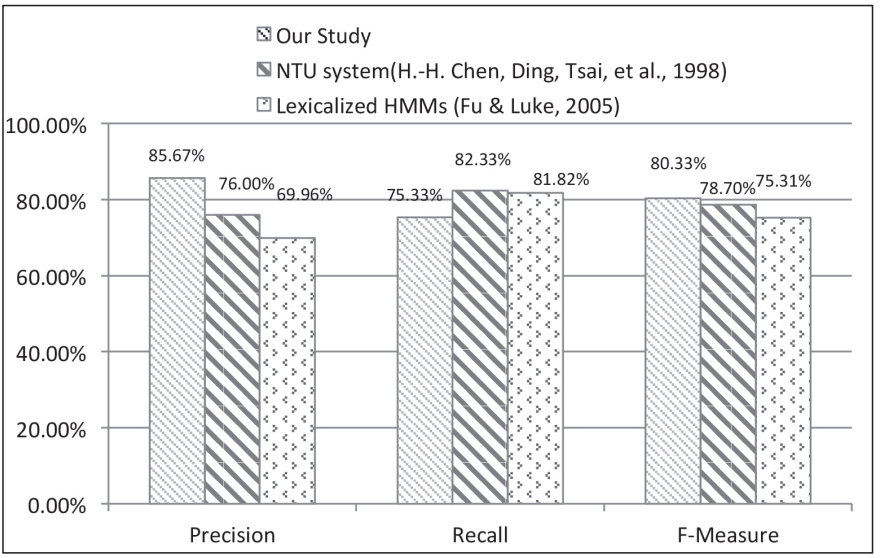

圖 13　實驗結果比較評估

伍、結論與展望

隨著科技的進展，巨量資料產生的速度只會增加而不會停歇，對於量的處理，可藉助先進的硬體設備解決，而其中最難處理的，則在於它的異質性與複雜性。隨著網路新聞資訊來源的多樣化、新創詞彙的快速與自由度，傳統詞庫斷詞法對於新穎的命名實體，或未知詞多半無法或未能正確指認其詞性。對於中文命名實體研究，仍以簡體中文為主，並著重於以人工建立規則配合統計機率模式推估。對於以維基百科為語料庫進行命名辨識與消歧義的研究並不多，若干研究結果都認定維基對提升命名辨識率的正面效果，只可惜處理素材都以英文為限。

由於中文命名實體變化多端，如能有效辨識與分歧人、事、時、地、物等特徵字詞，將可協助讀者在巨量資料中，快速掌握重要訊息。由於人工研擬規則，可能有涵蓋不全，以致召回率低，但精確率高的特點。過去 Chen 與 Bai（1998）對未知詞辨識的研究，以人工制定出 139 條規則，再依據粗訂原則，由程式跑出數十萬條細則，其後經過統合整理出數萬條。本研究使用詞性組合簡易句法規則，分別提出新

的辨識方法，並結合維基百科查詢專頁中資訊盒及開放類別之特性描述，以協助辨識。實驗結果有多項發現與突破：1. 證實以簡易句法規則結合維基百科即可有效進行命名實體辨識與消歧義；2. 以長詞優先處理人名命名規則，可有效合併頭銜，藉以分辨同名不同人；3. 應用姓名鏈結機率公式結合句法規則，可大幅提高人名辨識精確率；4. 在姓名字數方面，過去研究以 2-6 字為考量，本研究經查考在姓氏之後，人名出現的頻率多為二個字，姓名字數集中在 2-4，少數為 5-6，逾 6 字者，僅千餘人，因此本研究在人名部分僅考量 2-4 字元，大為縮小計算成本。5. 透過簡易地名規則可偵測 CKIP 誤判詞性 Nc 為 Nb 的失誤；6. 實驗證明複雜結構的組織名，透過維基百科資訊盒和開放類別中定義之特徵，可有效協助指認。實驗結果顯示本研究整體表現以精確率最為突出，至於召回率雖表現不稱優異，卻也達七成五水平，推測召回率不高的原因可能是本研究對於詞性組合之歸納以涵蓋面廣又不失正確為原則，以致無法網羅所有獨特案例。倘若再擴充規則，預期召回率可再提高，但精確率有可能相對降低。基於本研究以辨認新聞重要關鍵詞為主，著重精確率的提高，權衡之下，仍以提高精確率及可接受的召回率為訂定原則。

在實驗過程中，我們也發現某些地名誤判為人名，因受限於 CKIP 對詞性之定義，以及無其他輔助詞性可資分辨。現階段雖然無法辨識，也不易一一條列例外。未來的研究，希望能提出更好的方法解決。

參考文獻

中文詞知識庫小組（1993）。技術報告。中文詞類分析（編號：93-05）。台北市：中央研究院資訊科學研究所。

內政部（2014）。全國姓名統計分析。台北市：內政部。

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。資訊管理學報，*22*（2），123-147。

Ageishi, R., & Miura, T. (2008). *Named entity recognition based on a Hidden Markov Model in part-of-speech tagging*. Paper presented at the First International Conference on the Applications of Digital Information and Web Technologies.

Asahara, M., Fukuoka, K., Azuma, A., Goh, C.-L., Watanabe, Y., Matsumoto, Y., & Tsuzuki, T. (2005). *Combination of machine learning methods for optimum chinese word segmentation*. Paper presented at the Proc. Fourth

SIGHAN Workshop on Chinese Language Processing.

Asahara, M., Goh, C. L., Wang, X., & Matsumoto, Y. (2003). *Combining segmenter and chunker for Chinese word segmentation*. Paper presented at the Proceedings of Second SIGHAN Workshop on Chinese Language Processing.

Bender, O., Och, F. J., & Ney, H. (2003). *Maximum entropy models for named entity recognition*. Paper presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Edmonton, Canada.

Bunescu, R., & Paşca, M. (2006). *Using encyclopedic knowledge for named entity disambiguation*. Paper presented at the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy.

Chen, H.-H., Ding, Y.-W., & Tsai, S.-C. (1998). Named entity extraction for information retrieval. *Computer Processing of Oriental Languages, 11*(4), 75-85.

Chen, H.-H., & Lee, J.-C. (1996). *Identification and classification of proper nouns in Chinese texts*. Paper presented at the Proceedings of the 16th conference on Computational linguistics - Volume 1, Copenhagen, Denmark.

Chen, H. H., Ding, Y. W., Tsai, S. C., & Bian, G. W. (1998). *Description of the NTU System used for MET2*. Paper presented at the Proceedings of 7th Message Understanding Conference (MUC-7), Fairfax, VA.

Chen, H. H., & Lee, J. C. (1994). The identification of organization names in Chinese texts. *Communication of COLIPS, 4*(2), 131-142.

Chen, K.-J., & Bai, M.-H. (1998). Unknown word detection for Chinese by a corpus-based learning method. *International Journal of Computational linguistics and Chinese Language Processing, 3*(1), 27-44.

Chen, W., Zhang, Y., & Isahara, H. (2006). *Chinese named entity recognition with conditional random fields*. Paper presented at the 5th SIGHAN Workshop on Chinese Language Processing, Australia.

Cheung, L., Tsou, B. K., & Sun, M. (1995). Identifying Chinese names in unrestricted texts. *Journal of Chinese Information Processing, 9*(2), 28-35.

Fu, G., & Luke, K.-K. (2005). Chinese named entity recognition using

lexicalized HMMs. *ACM SIGKDD Explorations Newsletter, 7*(1), 19-25. doi:10.1145/1089815.1089819

Fu, R., Qin, B., & Liu, T. (2011). *Generating Chinese named entity data from a parallel corpus*. Paper presented at the IJCNLP.

Furu, W., Wenjie, L., & Yanxiang, H. (2007). *Measuring relevance with named entity based patterns in topic-focused document summarization*. Paper presented at the International Conference on Natural Language Processing and Knowledge Engineering.

Grishman, R., & Sundheim, B. (1996). *Message understanding conference-6: A brief history*. Paper presented at the Proceedings of the 16th conference on Computational linguistics, Copenhagen, Denmark.

IBM. (2015). Big data at the speed of business. Retrieved from http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Kazama, J. i., & Torisawa, K. (2007). *Exploiting Wikipedia as external knowledge for named entity recognition*. Paper presented at the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Lu, X. (2005). *Towards a Hybrid Model for Chinese Word Segmentation*. Paper presented at the Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing.

Ma, W.-Y., & Chen, K.-J. (2003). *A bottom-up merging algorithm for chinese unknown word extraction*. Paper presented at the Proceedings of Second SIGHAN Workshop on Chinese Language Processing.

Mihalcea, R., & Csomai, A. (2007). *Wikify!: linking documents to encyclopedic knowledge*. Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal.

Nguyen, H. T., & Cao, T. H. (2008). *Named entity disambiguation on an ontology enriched by Wikipedia*. Paper presented at the IEEE International Conference on Research, Innovation and Vision for the Future.

Sun, J., Gao, J., Zhang, L., Zhou, M., & Huang, C. (2002). *Chinese named entity identification using class-based language model*. Paper presented at the Proceedings of the 19th international conference on Computational

linguistics, Taipei, Taiwan.

Takeuchi, K., & Collier, N. (2002). *Use of support vector machines in extended named entity recognition*. Paper presented at the proceedings of the 6th conference on Natural language learning.

Todorovic, B. T., Rancic, S. R., Markovic, I. M., Mulalic, E. H., & Ilic, V. M. (2008). *Named entity recognition and classification using context Hidden Markov Model*. Paper presented at the The 9th Symposium on Neural Network Applications in Electrical Engineering

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & 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.

Verma, M., Sikdar, U., Saha, S., & Ekbal, A. (2013). *Ensemble based active annotation for biomedical named entity recognition*. Paper presented at the International Conference on Advances in Computing, Communications and Informatics (ICACCI).

Wang, L., Li, S., Wong, D. F., & Chao, L. S. (2012). *A joint Chinese named entity recognition and disambiguation system*. Paper presented at the Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing.

Wu, Y., Zhao, J., & Xu, B. (2003). *Chinese named entity recognition combining a statistical model with human knowledge*. Paper presented at the Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition, Sapporo, Japan.

Xue, N. (2003). Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing, 8*(1), 29-48.

Using Part-of-Speech Tagging Based Approach with Wikipedia to Assist Chinese Named Entity Recognition and Disambiguation

Chuen-Min Huang

Associate Professor
Department of Information Management
National Yunlin University of Science and Technology

Introduction

Traditional Named Entity Recognition (NER) adopts rule-based and/or probabilistic models in morphological analysis, but there still exists the problem of low accuracy due to semantic ambiguity and the rapid growth of unknown words. Individual writing habits also make it impractical if not impossible to lay out all the syntax rules, or collect every named entity in one thesaurus. In the study of NER, entities like date/time, monetary and percentage expressions are relatively regular and easy to identify; the scope of "things" is too wide and not included in the (Message Understanding Conference, MUC)-7 named entity definition range; therefore, they were excluded from this study. Identifying the entities of person, place, and organization is a challenging yet valuable task, since the three types of entities are often affected by the evolution of time, while they are the major constituents in a document, therefore, they were selected as the targets in this study.

In the past, a couple of studies employed different types of information from different levels of text to extract Chinese named entities, including character conditions, statistical information, titles, punctuation marks, location and organization keywords, speech-act and locative verbs, cache and n-gram models, with the number of generated syntax rules usually reaching in to the thousands. From combining too many named entities, accuracy is persistently difficult to improve. Many studies also conducted personal NER adopting various name-formulation rules trained from a

personal name corpus containing over 1 million Chinese personal names. Additionally, baseline, titles, positions and special verbs were used to enhance the recognition rate. As for location and organization NER, previous studies usually referred to a specific thesaurus, a special stem set, and a special verb. It is believed that the structures of location and organization names are more complex than those of personal names.

Rule-based approaches usually requires a lot of human effort to construct detailed rules, therefore, the rule generation is very time-consuming, and more dependent on the language and specific knowledge domains. The increase in excessive rules may cause an overfitting phenomenon, resulting in erroneous entity extraction results. The major indicated errors result from propagation errors, keyword sets, character sets, rule coverage, and so on. Therefore, the question of how to more accurately identify textual features has remained a challenging topic in information capture technology.

Design/methodology/approach

In this study, we randomly selected 3,000 documents from a total of 32,000 news articles collected from Yahoo! News 2010-2011, for which we analyzed the writing structure and observed each possible syntactic rule from the composition of most commonly used part of speech (POS) to its extended POS variation. In general, the shorter the combination of POS is, the more coverage it will have, and that raises the recall rate, but reduces the accuracy rate. Taking the simple personal name rule [Na+Nb→Nb] as an example, where Na is a common noun, Nb is a proper noun; by applying this rule, we will extract many personal names and more other mistakenly identified entities. That's why we need to add more POS from the adjacent context to help clarify and disambiguate. In this way, if we add too many POS extensions, it will cause an overfitting phenomenon, resulting in each rule only applying to a few cases. Thus, we have to iteratively test each situation to decide the optimal rationality of the length of the rules. Finally, we summarized 54 rules of personal names, and 11 rules of location and organization names. As for the MET-2 Chinese test corpus of MUC-7 which has correct answers, we adopted it as the validation data. Since the documents from MET-2 are simplified Chinese characters (predominantly

used in mainland China), we had to transform simplified Chinese characters in GB coding set to traditional Chinese characters in Big-5 coding set (as used in Taiwan) before testing.

In recognizing personal names, we applied the 54 rules to obtain candidate features, from which to further identify the monosyllabic or two-character combinations of surnames and given names based on a name concatenation formula. For practical considerations, our experiment only deals with the most frequent 2-4 characters of personal names. Applying the combination of extended POS composition, we not only successfully extracted persons' jobs and job portfolio, but also identify more related features to enhance the accuracy of the names. To improve recognition accuracy of location and organizations, we applied the 11 rules and also referred to geographical directories of Infobox in Wikipedia, the largest online encyclopedia for the time being, to assist in disambiguation. In our overall system evaluation, the precision rate achieves 86.32%, recall reaches 75.65%, and F-measure reaches 80.4%. Compared with other automatic rule construction and quasi-machine learning methods, we have a better performance particularly for precision.

Findings

Our study found that the recognition accuracy is raised because of a combination of syntax rules with name concatenation formula. With the help of POS tags, the transliteration of proper nouns can also be effectively detected from the extracted job titles to distinguish the same name from different persons. This study also found that the transliteration of foreign location names is very similar to personal names, and that is why the CKIP usually mistakenly identifies them as Nb. Thus, if there is no syntactic analysis, it is very likely to be mistaken for a personal name. The experiment showed that this problem can be effectively solved by our rules. While the location and organization names usually follow some particular verbs, previous studies showed that they are difficult to be differentiated even with the help of a specific thesaurus. Although we listed only 11 rules of location names, the accuracy rate of recognition is as high as 91%, which is far beyond other studies.

In summary, this study has the following findings: 1. It demonstrates

Chinese NER can be effectively fulfilled by applying simple syntax rules as a base and then incorporating feature extraction as verification from Wikipedia; 2. It shows a name concatenation formula can be effectively used to identify the monosyllabic or two-character combinations of surnames and given names; 3. An appropriated combination of multiple parts of speech (POS) can extend basic syntax rules to identify persons' job titles and job portfolio for further reference; 4. Some mislabeled location tags can be detected and corrected by applying our rules.

Research limitations/implications

This study highly relied on the POS tags generated by CKIP, which may not correctly identify some foreign locations if the tag is unexpectedly mislabeled. On the other hand, in this study, we only focused on the most popular 2-4 character combinations of surnames and given names, thus the name concatenation formula may not be applicable to other considerations. It is suggested that future research expands on our studies by considering the mentioned issues to fill the information gap.

Practical implications

Previous studies demonstrated that even with a large number of syntactic rules, the coverage is still not comprehensive, and the accuracy rate has remained difficult to effectively improve. This study illustrates how a light and shorter rule-based approach can be better than a heavyweight one. That means a lightweight syntax rule approach combined with Wikipedia can carry out a promising Chinese NER. Furthermore, the proposed method can work with news websites to extract and display associated key features of an event, which help to cluster similar events and conduct content comparison.

Originality/value — Previous studies adopting syntax rules to conduct name entity recognition usually have their limitations in comprehensive coverage and accuracy enhancement. This study focuses on identifying the core constituent elements of the news, thereby focusing more on the accuracy than coverage. Finally, we generated 54 rules of personal names,

and 11 rules of location and organization names. A couple of studies adopted Wikipedia to assist NER, while none of them focus on Chinese text processing. This study is of the first its kind to apply syntax rules combined with Wikipedia to conduct Chinese NER. The method is light-weighted and domain independent with the flexibility to be extended. Compared with other automatic rule construction and quasi-machine learning methods, we have a better performance particularly for precision.

ROMANZIED & TRANSLATED REFERENCE FOR ORIGINALTEXT

Chinese Knowledge Information Processing Group. (1993). Zhong Wen Ci Lei Fen Xi. (Tecnical Report NO. 93-05). Taipei : Academia Sinica Institute of Information Science

內政部（2014）。全國姓名統計分析。台北市：內政部。【Ministry of the Interior. (2014). Quan Guo Xing Ming Tong Ji Fen Xi. Taipei : Ministry of the Interior.】

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。資訊管理學報，*22*（2），123-147。【Huang, C. M., Li, Y. C., & Chen, P. H. (2015). Wikipedia-based Chinese Abbreviation and Synonym Construction. *Journal of Information Management*, *22*(2), 123-147.】

Ageishi, R., & Miura, T. (2008). *Named entity recognition based on a Hidden Markov Model in part-of-speech tagging*. Paper presented at the First International Conference on the Applications of Digital Information and Web Technologies.

Asahara, M., Fukuoka, K., Azuma, A., Goh, C.-L., Watanabe, Y., Matsumoto, Y., & Tsuzuki, T. (2005). *Combination of machine learning methods for optimum chinese word segmentation*. Paper presented at the Proc. Fourth SIGHAN Workshop on Chinese Language Processing.

Asahara, M., Goh, C. L., Wang, X., & Matsumoto, Y. (2003). *Combining segmenter and chunker for Chinese word segmentation*. Paper presented at the Proceedings of Second SIGHAN Workshop on Chinese Language Processing.

Bender, O., Och, F. J., & Ney, H. (2003). *Maximum entropy models for named entity recognition*. Paper presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003,

Edmonton, Canada.

Bunescu, R., & Paşca, M. (2006). *Using encyclopedic knowledge for named entity disambiguation*. Paper presented at the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy.

Chen, H.-H., Ding, Y.-W., & Tsai, S.-C. (1998). Named entity extraction for information retrieval. *Computer Processing of Oriental Languages, 11*(4), 75-85.

Chen, H.-H., & Lee, J.-C. (1996). *Identification and classification of proper nouns in Chinese texts*. Paper presented at the Proceedings of the 16th conference on Computational linguistics - Volume 1, Copenhagen, Denmark.

Chen, H. H., Ding, Y. W., Tsai, S. C., & Bian, G. W. (1998). *Description of the NTU System used for MET2*. Paper presented at the Proceedings of 7th Message Understanding Conference (MUC-7), Fairfax, VA.

Chen, H. H., & Lee, J. C. (1994). The identification of organization names in Chinese texts. *Communication of COLIPS, 4*(2), 131-142.

Chen, K.-J., & Bai, M.-H. (1998). Unknown word detection for Chinese by a corpus-based learning method. *International Journal of Computational linguistics and Chinese Language Processing, 3*(1), 27-44.

Chen, W., Zhang, Y., & Isahara, H. (2006). *Chinese named entity recognition with conditional random fields*. Paper presented at the 5th SIGHAN Workshop on Chinese Language Processing, Australia.

Cheung, L., Tsou, B. K., & Sun, M. (1995). Identifying Chinese names in unrestricted texts. *Journal of Chinese Information Processing, 9*(2), 28-35.

Fu, G., & Luke, K.-K. (2005). Chinese named entity recognition using lexicalized HMMs. *ACM SIGKDD Explorations Newsletter, 7*(1), 19-25. doi:10.1145/1089815.1089819

Fu, R., Qin, B., & Liu, T. (2011). *Generating Chinese named entity data from a parallel corpus*. Paper presented at the IJCNLP.

Furu, W., Wenjie, L., & Yanxiang, H. (2007). *Measuring relevance with named entity based patterns in topic-focused document summarization*. Paper presented at the International Conference on Natural Language Processing and Knowledge Engineering.

Grishman, R., & Sundheim, B. (1996). *Message understanding conference-6: A brief history*. Paper presented at the Proceedings of the 16th conference on Computational linguistics, Copenhagen, Denmark.

IBM. (2015). Big data at the speed of business. Retrieved from http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Kazama, J. i., & Torisawa, K. (2007). *Exploiting Wikipedia as external knowledge for named entity recognition*. Paper presented at the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Lu, X. (2005). *Towards a Hybrid Model for Chinese Word Segmentation*. Paper presented at the Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing.

Ma, W.-Y., & Chen, K.-J. (2003). *A bottom-up merging algorithm for chinese unknown word extraction*. Paper presented at the Proceedings of Second SIGHAN Workshop on Chinese Language Processing.

Mihalcea, R., & Csomai, A. (2007). *Wikify!: linking documents to encyclopedic knowledge*. Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal.

Nguyen, H. T., & Cao, T. H. (2008). *Named entity disambiguation on an ontology enriched by Wikipedia*. Paper presented at the IEEE International Conference on Research, Innovation and Vision for the Future.

Sun, J., Gao, J., Zhang, L., Zhou, M., & Huang, C. (2002). *Chinese named entity identification using class-based language model*. Paper presented at the Proceedings of the 19th international conference on Computational linguistics, Taipei, Taiwan.

Takeuchi, K., & Collier, N. (2002). *Use of support vector machines in extended named entity recognition*. Paper presented at the proceedings of the 6th conference on Natural language learning.

Todorovic, B. T., Rancic, S. R., Markovic, I. M., Mulalic, E. H., & Ilic, V. M. (2008). *Named entity recognition and classification using context Hidden Markov Model*. Paper presented at the The 9th Symposium on Neural Network Applications in Electrical Engineering

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & 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.

Verma, M., Sikdar, U., Saha, S., & Ekbal, A. (2013). *Ensemble based active annotation for biomedical named entity recognition*. Paper presented at the International Conference on Advances in Computing, Communications and Informatics (ICACCI).

Wang, L., Li, S., Wong, D. F., & Chao, L. S. (2012). *A joint Chinese named entity recognition and disambiguation system*. Paper presented at the Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing.

Wu, Y., Zhao, J., & Xu, B. (2003). *Chinese named entity recognition combining a statistical model with human knowledge*. Paper presented at the Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition, Sapporo, Japan.

Xue, N. (2003). Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing, 8*(1), 29-48.