應用混淆音矩陣之中英文音譯詞組自動抽取

郭金喜[1,2]　　楊英魁[2]

[1]中華電信研究所　　[2]國立台灣科技大學電機系

jskuo@cht.com.tw　　ykyang@mouse.ee.ntust.edu.tw

摘要

機器音譯(Machine Transliteration)是機器翻譯中重要的一環，因爲許多文章中常有人名、地名及組織名等專有名詞夾雜其中，雖然經由查閱預先整理之詞典可以解決部分的問題，但是這些專有名詞數量隨時間不斷的增加及成長，而辭典的整理既費時又費力，透過音譯詞組自動抽取(Transliterated-Term Pair Extraction)，可動態補充辭典內容之不足。有足夠的中英文音譯詞組做爲訓練語料之後，則可建立一中英文音節對應(Syllable Mapping)系統，應用於中英文詞組音譯，但問題是該如何快速獲取足夠的中英文音譯詞。本文提出一方法，自網頁中抽取出大量的中英文音譯詞組，利用中文語音辨認系統在辨認過程所產生的混淆矩陣(Confusion Matrix)來克服發音變異(Pronunciation Variation)。從實驗結果發現本文所提出的方法可達到 32.26%的檢出率(Recall)及 95.23%的準確率(Precision)，足以證明所用方法確實可有效的應用於音譯詞組自動抽取。

1. 簡介

當國際交流日益頻繁，各國間的資訊傳遞也更加迅速，許多的媒體必須在短時間將所收到的外國資訊儘可能完善的翻譯成本國文字，以滿足讀者的需求，在現今媒體開放、競爭的台灣這種步調更加快速。這些外國資訊常包含有許多的專有名詞(Proper Noun)如人名、地名及組織名等夾雜其中，同一名詞出現在不同文章中但由同一人員翻譯可能會出現不同的譯名，同一名詞由不同人員翻譯也可能會出現不同的名稱。這些問題主要是因爲所接收的外國資訊涵蓋非常廣泛，發生的地點及所使用的語言更是廣佈於全世界，實在是非單一個人可以準確的音譯出由不同語言所發聲的的專有名詞。專有名詞的音譯並不在本文的探討範圍，但音

譯詞組的自動抽取卻是建構機器音譯系統不可或缺的一步。

機器音譯常用來處理人名、地名等，其作法乃是將這些專有名詞經由發音方式自一語言轉換至另一語言。它是機器翻譯中重要的一環，因爲在許多文章中常有人名、地名及組織名等專有名詞夾雜其中，雖然經由查閱預先整理之詞典可以解決部分的問題，但是這些專有名詞數量隨時間不斷的增加及成長，而辭典的整理既費時又費力，透過音譯詞組自動抽取可動態補充辭典內容之不足。

想要自動抽取音譯詞組，必須要能自足夠大的語料庫中抽取出多樣且量多的音譯詞組，一般測試用語料庫大多無法滿足這樣的需求。網際網路是現今世界上最大的分散式資料庫，其所包含的資料雖然缺乏有系統的整理 (Systematically Organized) ，但卻包羅萬象而且源源不絕不斷有新的內容產生，這樣具有動態特性的資料是許多研究不可或缺的素材。本文的目的是要自這些網頁資料中抽取出許多可能的中英文音譯詞組，做爲未來發展機器音譯的基礎。

英文是目前國際上最通用的語言之一，許多資訊是透過英文翻譯或音譯至其他語言去，有許多的名詞先被引進至英文，其他語言的使用者，透過再從英文引進這些名詞。因此造成許多語言自英文引進的外來語，其原來的字源(Word Origin)[Llitjos2001] 並非來自英文，因此若不了解外來語的字源，常會有發音不一致的其行產生。例如義大利地名 Firenze 及其英文音譯 Florence [Lin2000]，究竟應該採用哪一種讀法，實在很難決定。即使對常用的英文字如 Mary/meɪri/ 、marry/mæri/及 merry/mɛri/ ，有些人把這三者均唸成不同或某兩者相同，但大多數的美國卻把這三者均唸成/mɛri/[Jurafsky2000]。這表示有發音變異的問題存在。而機器音譯則用來將人名、地名等專有名詞經由發音方式自一語言轉換至另一語言，所以爲了克服不同人的發音變異問題，必須抽出足夠的音譯詞組，進而建構不同語言間的音節轉換關係。

鄰近的日本與韓國也極力引進及吸收國外資訊，日本文字[NIHOGO90] 中有片假名用於表達外國人的國名、地名、人名(中國、韓國人名 除外)、外來語及專門術語等等，因此英文及日文中的音譯詞組抽取[Brill2001] 可以較清楚的區分何者爲外來語。韓文也引進大量的外

來語 ，因此有許多人致力於英文及韓文間之音譯研究[Jeong97][Lee98][Jung2000][Kang2000][Oh2002]。韓文中雖然沒有像日文中特別以片假名來表示外來語，但韓文中的外來語在其字尾多包含有 “Josa” 及 “Eomi” 等符號。這些特性是中文中所沒有的，也增加了中英文音譯詞組自動抽取的困難度。

有關中文方面的音譯詞組自動抽取研究，曾有研究[Xiao2002]利用某些特定的外國人名所出現的音譯中文字如『夫』、『斯』、『基』等找尋其他的中文音譯詞，這些字也許常出現於外國人名的中文音譯中，特別是斯拉夫民族的人名，但是中文人名也有可能會用到這些字，如『李斯』及『郭李建夫』等。[Lee2003] 則使用統計音譯模型於中英文雙語語料庫(Parallel Corpora)音譯詞組抽取，由於雙語語料庫的資料量相對小於散佈於網際網路上的網頁資料，故能夠取得的中英文音譯詞組數量相對較少。

根據機器音譯模組化學習法(Modular Learning Approach)[Knight98]，在一堆不同語言的音譯詞組候選詞中，最後的音譯詞 $\hat{C}$ 可由(1)決定

$$\hat{C} \equiv \underset{C_w}{argmax}\ p(C_w|E_w) = \underset{C_w}{argmax}\ p(E_s|E_w)\,p(C_s|E_s)\,p(C_w|C_s) \qquad (1)$$

，其中 C_w 及 E_w 分別爲目標語言(Target Language)及來源語言(Source Language)文字串， C_s 及 E_s 是目標語言及來源語言文字串所對應之音素(Phoneme)串， $p(C_w|E_w)$ 爲一條件機率函數 ，下標 C_w 是在所有音譯詞中使得 $p(C_w|E_w)$ 最大的音譯詞。在本文中來源語言是指英文，目標語言則是指中文。公式(1)意思爲要從一個英文字串所對應的中文候選詞中挑出一個最有可能的中文字串，乃是找出一個中文字串使得英文字轉音的機率、英文對中文音轉音的機率以及中文的音轉字的機率三者連乘最大。本文的目的在於音譯詞組抽取，故著重於音素相似程度的評估。由於中英文隸屬於不同語系，且並無類似韓文EKSCR(English-to-Korea Standard Conversion Rules)[Oh2002] 的規則可供遵循，而且一個英文字(Word)可有多個音節，但一個中文字則只有一個音節，其對應關係應不易決定。

基於上述理由，本文提出基於語音辨認的混淆音矩陣，來解決中英文音譯詞自動抽取過程中必須克服的發音變異以及中英文音節對應不易問題。音譯詞組自動抽取的目的在於蒐集

足夠的中英文音譯詞組，以處理中英文音譯的音節轉換問題，這樣經由大量統計而產生的中英文音節轉換系統，納入了許多中英文音節的各種對應關係，對於進一步的中英文音譯研究有很大的裨益。

本文內容第二節將討論如何自動抽取中英文音譯詞組，第三節爲實驗結果與討論，最後是結論。

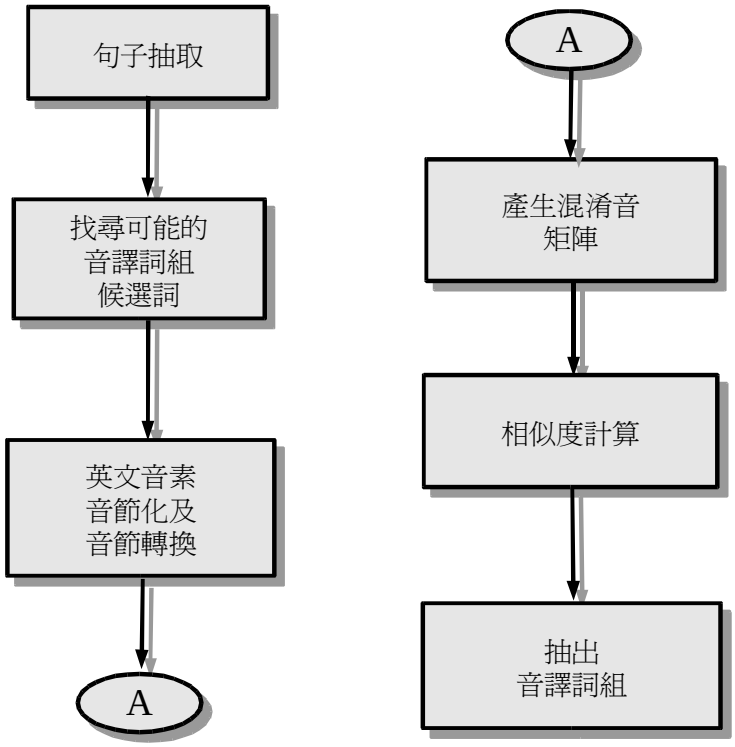

圖 1 中英文音譯詞組自動抽取流程圖

2. 使用方法

網際網路是現今世界上最大的分散式資料庫，每年以極快速的速度成長，其內容五花八門而且不斷有內容更新及產生。本文的目的是要自這個最大的資料庫中抽取出許多可能的中英文音譯詞組，處理中英文音譯的音節轉換問題，做爲進一步發展機器音譯的基礎。這個準結構化的(Semi-Structured)資料庫的確包羅萬象、應有盡有，但如何能夠在其中有效率且快速的獲取所需要的資訊，訂出適當的搜尋或比對範圍，減少不必要的計算，過濾不必要的雜訊，進而抽取出大量可能的中英文音譯詞組，是一項艱難但必須克服的問題。

音譯詞組係由不同語言的文字串所對應而成，如果要在非雙語對應(Non-Parallel)的網頁中抽取出音譯詞組，則表示網頁中包含有混合兩種語言以上的文字資料。因此重要的是要找到分屬目標語言與來源語言的對應文字串，如果無法找到非常明確的對應文字串，則必須縮小範圍或以其他有效的特徵過濾不必要的文字雜訊。

本文在中英文音譯詞組自動抽取的方法上，如圖 1. 中英文音譯詞組抽取流程圖所示，主要可分爲四個步驟即 1) 找尋可能的音譯詞組候選詞，2) 英文音素音節化(Syllabification)及音節轉換，3) 產生混淆音矩陣，4)相似度計算。中英文音譯詞組自動抽取模組首先自龐大的文字語料庫中找到一個由標點符號隔開的句子，並在此句子中找到一連續的英文字串，這個英文字串可能包含一個或一個以上的英文字，再以此字串爲中心往兩旁延伸，訂出中文詞尋找範圍，並經由音節相似度比對，過濾掉不符合要求的候選詞。在音節相似度比對之前，必須將英文字串中的每一個英文字，經由英文字轉音及音素音節化程序，將每一個英文音節轉化至中文音節，並產生相對應的混淆音矩陣，這些中文化後的英文音節再與中文候選詞的音節進行相似度比對，找出適當的中文候選詞，並決定是否爲可能的中英文音譯詞組。

以下分別就每一步驟進行更詳細的說明：

2.1 找尋可能的音譯詞組候選詞

[Nagata2001]提出自近乎雙語(Partially Bilingual)的網頁中抽取出翻譯詞組，他們觀察到在日文的網頁中有許多英文詞與日文詞夾雜，大部分的日文相對應翻譯詞即英文詞被包括在括號內，而且緊接在日文詞後面。此種現象不僅出現在日文中，也同時出現在東方語系的中文及韓文中。有許多情形的確是如此，而且括號也常暗示強烈的詞組對應關係，但並非所有的翻譯詞組或音譯詞組皆以這種對應型態出現。以下面這段取自報章的文章說明，

『...MP3 所引起長久以來「版權」的問題，訴訟不斷，爭議不休，始終沒有一個確定的解決方案，經營 Kuro 庫洛 P2P 音樂交換軟體的飛行網，3 日發表 P2P 與版權爭議的解決方案—C2C(Content to Community)，希望能在使用端、科技業者與唱片業者三方中間找到一

個平衡點。...』

在文章中，出現的中英文詞的關係可分爲以下幾種情形，1) 用來形容或補充說明相關詞，如 P2P 與音樂交換軟體的關係。2)日常生活中常用之英文詞，如 MP3。3)無適當的翻譯詞或音譯詞，如 C2C。同時在此段文章亦出現有一音譯詞組『 Kuro 庫洛』，此音譯詞組在文章是緊鄰在一起，但並非以括號突顯音譯詞的型態出現；相反的，文中出現以括號突顯的翻譯詞『 C2C(Content to Community) 』，其關係恰如 1) 用來進一步說明前面所出現的詞。

因此本文參考[Nagata2001]的觀察，使用類似的方法，但並不僅限於處理括號所給予的提示。假設自語料庫中抽出一句子 S=（$s_1 s_2 ... s_m$），其中每一個 $s_{i,i=1..m}$ ，是一中文字(Character)或英文字元(Alphabet)，自 S 中先找到可能的英文字串(Word String) EWS ，EWS 包含一個或一個以上的字元，EWS 可表示爲 EWS ∈ S, EWS =（$t_1 t_2 ... t_n$），其中個別 $t_{i,i=1..n}$ 是以空白斷開來或強迫被切開的英文詞(Word 或 Token)，也是後續處理的基本單位。而可能的中文候選字串則可沿著 EWS 左右兩邊尋找而得，若 l_{ss} 和 l_{se} 分別爲 S 之起始與結束之位置(Location)， l_{es} 和 l_{ee} 分別爲 EWS 之起始與結束之位置，lnc_l 和 lnc_r 分別爲沿 EWS 左右兩邊所遇到的第一個非中文字的位置(括號在此是可被忽略的)。故 EWS 左邊的中文候選字串 CW_l 的範圍爲從 $max(l_{ss}, lnc_l)$ 至 l_{es} ，而 EWS 右邊的中文候選字串 CW_r 的範圍爲從 l_{ee} 至 $min(l_{se}, lnc_r)$ 。這個方法不需經過斷詞程序，即可找到適當的中文候選字串 ，再透過後面所敘述的音節相似度計算，進一步找到正確的音譯詞組。但缺點是可能會將一些音節相似的候選字串納入考慮，不只增加計算量，也使得錯誤率提高。

以上一段文章中部分字串『經營 Kuro 庫洛』例，先找到這個字串所屬的句子，繼而找到句子中的來源語言(即英文)字串『Kuro』，找到英文字串後則沿此英文字串的左右兩傍找尋目標語言(即中文)候選字串，故可找到『經營』以及『庫洛』兩中文候選字串，做爲 Kuro 可能的音譯候選字串。而依上述的方法，雖可找到英文字串『 C2C 』，但卻無法

找到相對應的中文候選字串。

如果要決定『經營』以及『庫洛』兩候選字串是否爲 Kuro 的音譯候選詞，則必須經過相似度計算。但來源語言字串與目標語言字串分屬於不同的語言，該如何計算彼此的相似度變成一項問題。直接輸入兩種不同語言的音素資料，並試著找出兩者的關係是一種方式，但問題是如何找出不同語言的音素資料關係。當抽取出大量的音譯詞組時，這樣的轉換關係是很容易可以獲得的；另一種方式則是將兩種語言的音素資料轉換至其中一方或其他第三種表示方式，使得資料表示形式一致，相似度計算相對變成較簡單。待解決的問題則是不同語言的音素資料轉換以及如何對應轉換過程的一對多或多對一關係。

2.2. 英文音素音節化及音節轉換

英文的音素約分爲四種類型，即子音(Consonant)、母音(Vowel)、半母音(Semi-Vowel)及鼻音(Nasal) ，其中子音約有 17 個，母音約有 16 個，半母音有 4 個，鼻音有 3 個[Jurafsky2000]。由這些音素所組成的英文音節總數可達數千種，而中文僅有約 414 個音節。由中文音節對應至英文音節則會有一對多的對應問題，使得對應關係更加複雜。而由英文音節對應至中文音節會有多對一的對應問題，但一個中文音節可拆成聲母(Initial)及韻母(Final)兩部份，透過子音與聲母及母音與韻母的對應，可較簡化其對應的複雜度。中英文音節對應關係則可應用現有語言學相關資料[NTNU82] 。

要達到音節轉換之前，必須先將前一節所找到的英文字串 EWS 中的每一個英文詞經由英文字轉音系統 MBRDICO[Pagel98] 轉換成一串的音素，經由 MBRDICO 所產生的音素係以 SAMPA(Speech Assessment Methods Phonetic Alphabet)表示法表示，隨後則將這些音素轉換至以 IPA(International Phonetic Association)表示法表示。將音素轉至以 IPA 表示法表示的目的，是希望能沿用 IPA 與中文聲母與韻母的關係，這種使用 IPA 與中文音節對應關係的方式，將來也可應用至其他語言與中文間的音譯詞組自動抽取。

爲求找到中英文音譯詞組，先將轉換後的英文音素音節化，音節化後的英文音素係以子音母音對(Consonant-Vowel Pair) 的方式呈現，[Wan98] 曾類似的音節化演算法，但其方

法伴隨著英文字轉音系統，並非直接針對音素處理。因此本文的音節化程序則直接以音素爲主。

以英文字 Kuro 爲例，首先利用英文字轉音系統轉成音素串/kurə/，然後利用音節化程序將此音素串切割成/ku/ 及 /rə/ 等音節，這些切割後的音節則再利用語言學上的規則，將每一個音節中的音素轉換成相對應的中文聲母或韻母。

如前所述，音譯過程乃是將文字經由發音的方式自一語言轉換至另一語言，但發音時常會因爲腔調或發音部位的不同，導致對於不同的翻譯人員對同一字詞有不同的發音，如 /rə/ 可能會有人發成 /lo/ 或 /ra/，這樣會使得音譯時所處理的音很接近但實際上不相同，更使得相似度比對時無法達到預期效果，音譯詞組自動抽取的成效連帶受到影響。可能的解決方式是設法收集相關混淆音並建立這些混淆音之間的關係，既可處理確實相關的混淆音，又可排除不相關的雜訊音。若以人工方式收集並區分這些混淆音，既曠日又費時，而且不知該從何處著手。如果能夠充分運用電腦的計算能力，快速的收集到這些資訊並建立彼此的關係，對於音譯詞組自動抽取的進行必有很大的幫助。

2.3 產生混淆音矩陣

語音辨認系統可被視爲一有雜訊的通道[Wang2002]，當一信號輸入至辨認系統時，原本應該辨認到正確的信號，卻可能因爲雜訊與輸入的信號混合，使得混合後的信號被誤認爲其他的信號，導致辨認結果錯誤。這些雜訊對於語音辨認效影響極大，因此通道雜訊的消除是語音辨認研究中重要的課題。混淆音矩陣是語音辨認過程的副產品，它表列了某一音常與某些音混淆在一起，這些混淆音可能是原本正確的音，也可能是常被誤認的音，因此常被用來分析辨認結果，進而改善辨認效能。

取自語音辨認系統的混淆音矩陣，可用來解決如何避免人工方式收集並區分混淆音費時費力的問題，這是因爲這些混淆音矩陣本身即表列了正確及常被混淆的音。但問題是要如何去控制混淆音矩陣品質，將容易混淆的音納入，而將確實因爲雜訊原因產生的音排除。採用同時考慮主音與混淆音是否同時出現及彼此的相依程度的方式，可達到上述目標。

由語音辨認系統產生的混淆音矩陣，並不適合直接拿來應用於音譯詞組抽取上，原因是部份混淆音並非取自良好的語料，剔除這些不良語料後，可較準確的計算出相對應的混淆音。本文所使用的混淆音矩陣有兩種，一是根據全音節計算而得的全音節混淆音矩陣。另一是將全音節混淆音矩陣分別拆成以聲母及韻母爲主的音素混淆音矩陣。因爲發音變化時可能只有聲母、韻母或者兩者同時產生變化，將音節拆成由聲母及韻母分別的對應關係，可以用來克服這種問題。

因爲發音方式不同所產生的問題，可以引用混淆音來解決。除此之外，還有在來源語言中會有發音變異的問題，例如 /t/ 和 /d/ 在英文中的發音，在子音之前或在一串子音群時常會被忽略而不發聲[Jurafsky2000]，這種問題在音譯詞組抽取時必須加以考慮，因爲如果這時候不納入考慮，則會影響到以後中英文音節對應的研究。這種發音變異的問題相當不易處理，原因是在來源語言的音素或音節中這些音是存在的，只是在發音的時候，它因人而異可以發音也可以不發音。如何規範這些不確定性是很大的挑戰。

2.4 相似度計算

公式(1)是機器音譯模組化學習法中用來處理機器音譯的數學模型，但本文的重點是在音譯詞組自動抽取，而且因爲使用 MBRDICO 英文字轉音系統，因此重點將是公式(1)的音素計算 $\boldsymbol{p}(\boldsymbol{C}_s|\boldsymbol{E}_s)$ 上。 [Brown93]曾提出一系列的統計式機器翻譯模型應用於機器翻譯詞的相似度計算，這些統計模型稍加修改也可適用於機器音譯上，但問題是該如何將發音變異問題同時納入於相似度計算上。而在相似度計算時，中文候選字串只是大約訂出大約範圍，並不知道英文詞所對應中文候選詞的真正位置。在發音變異問題時，某些被轉換至中文音節後的英文音節可能被忽略，因此必須將英文音節的所有可能組合列出，而對應中文候選詞音節的選取則採取滑動視窗(Sliding Window)的方式，滑動視窗的大小即爲英文音節的數目，中英文候選詞選出後再進行音節間的相似度計算。

在相似度計算之前，先定義以下符號：

EWS：爲一英文字串。

EW：爲 ***EWS*** 中之一英文詞(Word)。

ES：爲 *EW* 所對應的音素串。

ECS：爲 *ES* 被轉換至中文音節的音節串。

ECS_i：爲 *ECS* 第 *i* 個音節子集合。

E_{ij}：爲 ECS_i 中第 *j* 個音節。

CWS：爲抽取音譯詞組時的中文候選字串。

CS：爲 *CWS* 所對應的音節串。

CW_i：爲 *CWS* 第 *i* 個字串子集合。

CS_i：爲 CW_i 所對應的音節串。

C_{ij}：爲 CS_i 中第 *j* 個音節。

$$\boldsymbol{p}(\boldsymbol{CWS}|\boldsymbol{EW})=\sum_{\boldsymbol{CW}_i}\boldsymbol{p}(\boldsymbol{CW}_i|\boldsymbol{EW}) \tag{2}$$

音譯詞組抽取時是以計算音節相似度做爲抽取與否的依據，故直接將中英文字串轉換成相對的音素或音節，做進一步的比對。則公式(2)變成

$$\boldsymbol{p}(\boldsymbol{CWS}|\boldsymbol{EW})\approx\sum_{\boldsymbol{CS}_i}\boldsymbol{p}(\boldsymbol{CS}_i|\boldsymbol{ES}) \tag{3}$$

爲了使音節相似度計算的進行，可以在同一基準上，故將英文音素經音節化並轉換至中文音節。並同時將發音變異問題納入考慮，則公式(3)變成

$$\begin{aligned}&\boldsymbol{p}(\boldsymbol{CWS}|\boldsymbol{EW})\\&\approx\sum_{\boldsymbol{CS}_i}\boldsymbol{p}(\boldsymbol{CS}_i|\boldsymbol{ECS})\\&=\sum_{\boldsymbol{CS}_i}\sum_{\boldsymbol{ECS}_j}\boldsymbol{p}(\boldsymbol{CS}_i|\boldsymbol{ECS}_j)\end{aligned} \tag{4}$$

而使得公式(4)機率最大的中文詞 $\hat{\boldsymbol{C}}$ 及英文詞 $\hat{\boldsymbol{E}}$，$\hat{\boldsymbol{J}}=(\hat{\boldsymbol{C}},\hat{\boldsymbol{E}})$ 則可下式決定

$$\hat{\boldsymbol{J}}\approx\underset{\boldsymbol{CS}_i\ \boldsymbol{ECS}_j}{\boldsymbol{argmax}}\ \boldsymbol{p}(\boldsymbol{CS}_i|\boldsymbol{ECS}_j) \tag{5}$$

$$\begin{aligned}&=\underset{\boldsymbol{CS}_i\ \boldsymbol{ECS}_j}{\boldsymbol{argmax}}\ \boldsymbol{p}(\boldsymbol{C}_{i1}\boldsymbol{C}_{i2}\ldots\boldsymbol{C}_{in}|\boldsymbol{E}_{j1}\boldsymbol{E}_{j2}\ldots\boldsymbol{E}_{jn})\\&\approx\underset{\boldsymbol{CS}_i\ \boldsymbol{ECS}_j}{\boldsymbol{arrmag}}\prod_{k=1}^{n}\boldsymbol{p}(\boldsymbol{C}_{ik}|\boldsymbol{E}_{jk})\end{aligned} \tag{6}$$

，其中 $\boldsymbol{p}(\boldsymbol{C}_{ik}|\boldsymbol{E}_{jk})$ 爲 $\boldsymbol{C}_{ik}$ 與 $\boldsymbol{E}_{jk}$ 兩個音節間的機率 。

公式 (6) 中的 $\boldsymbol{p}(\boldsymbol{C}_{ik}|\boldsymbol{E}_{jk})$ 是兩個中文音節的機率，這個機率的計算可直接由混淆音矩陣得到，若同時將全音節混淆音矩陣(ASR-Syllable, AS) 、音素混淆音矩陣(ASR-Phoneme, AP) 以及根據語言學規則[NTNU82] 所訂定的音素混淆音矩陣(Rule-based

Phoneme, RP) 納入考慮，則 $p(C_{ik}|E_{jk})$ 可寫爲，

$$p(C_{ik}|E_{jk})=\alpha t_s(C_{ik}|E_{jk})+\beta t_p(C_{ik}|E_{jk})+\gamma t_r(C_{ik}|E_{jk}), \alpha+\beta+\gamma=1, \quad (7)$$

，其中 $t_s(C_{ik}|E_{jk})$ 可直接利用 AS 求得， $t_p(C_{ik}|E_{jk})$ 可利用 AP 求得， $t_r(C_{ik}|E_{jk})$ 是可利用 RP 求得， α 、 β 及 γ 則分別爲 $t_s(C_{ik}|E_{jk})$ 、 $t_p(C_{ik}|E_{jk})$ 及 $t_r(C_{ik}|E_{jk})$ 的權重(Weighting)。

因爲AP是將一個中文音節拆成聲母及韻母，假設聲母及韻母的產生彼此沒有關聯，所以 $t_p(C_{ik}|E_{jk})$ 可由下式求得，

$$t_p(C_{ik}|E_{jk}) \approx p(CI_{ik}|EI_{jk})\, p(CF_{ik}|EF_{jk}) \quad (8)$$

，其中 CI_{ik} 及 EI_{jk} 分別爲 C_{ik} 及 E_{jk} 之聲母部份， CF_{ik} 及 EF_{jk} 分別爲 C_{ik} 及 E_{jk} 之韻母部份。RP 是根據的語言學規則所訂定的音素混淆音矩陣，例如國音聲學可以發音部位及發音方法分成兩類，若以發音部位爲分類依據，b、p及m(以IPA表示)都受到上下唇而發音，是屬於同部位的音，故可視爲同一群[NTNU82]。故 $t_r(C_{ik}|E_{jk})$ 可定義爲，

$$t_p(C_{ik}|E_{jk}) \equiv 1 \text{ ，} \textit{如果} C_{ik} \textit{與} E_{jk} \textit{在RP的同一混淆音群中}$$
$$\equiv 0 \text{ ，} \textit{其他} \quad (9)$$

3. 實驗結果與討論

實驗結果是取自台灣區域的繁體中文網頁，過濾後的純文字檔案大小約有 500MB，自其中抽出 80,094 個句子，其檔案大小約爲 5MB，最後共抽出 10,225 個可能的音譯詞組(Transliterated-Term Pair)。在計算檢出率(Recall) 及準確率(Precision)時，採用隨機選擇(Randomly Selected) 方式，自 80,094 個句子中挑出 200 個句子爲評估樣本，在 200 個句子中共產生 488 個候選音譯詞組，可抽出 21 個音譯詞組，經人工確認，其中有 20 個相關。此外經人工確認在 488 個候選音譯詞組中，有 62 個相關的音譯詞組。故檢出率爲 32.26%，準確率爲 95.23%。

表 2 所示爲部份的實驗結果，其中括號內的數字代表出現的次數，這也可得知對某一英文字而言，大部分相對的中文音譯詞爲何。有趣的是許多中文詞僅出現一次，這些低

頻詞並無法以詞共現(Word Co-occurrence)的翻譯詞抽取方法抽取出來，但透過以發音爲特徵的音譯詞組抽取是可以將這些低頻詞抽取出來。

由表 2 的實驗結果可發現發音變異的問題確實存在於音譯處理中，例如英文的 /t/ 常對應至中文的『t』或『t'』(以 IPA 符號表示)，而英文的/d/ 則也可能會對應到中文的『t』或『l』，這表示應用於混淆音矩陣於音譯詞組抽取是合理的。另外表 2 中的 Charles 對應至其中的一個中文音譯詞『策略師』，從以發音爲特徵的相似度計算的角度來看，並沒有錯誤，但對於音譯的角度而言卻是個錯誤。這個錯誤是因爲在找尋可能的音譯詞組時範圍設定過寬，導致將相連接的其他文字也納入範圍之內，再加上相似度計算彼此特性相近，故將此錯誤納入。

英文字	音譯中文詞 1	音譯中文詞 2	音譯中文詞 3	音譯中文詞 4
Robert	霍伯德(1)	羅伯特(4)		
Charles	查理斯(1)	查爾斯(5)	察爾斯(1)	策略師(1)
Michael	麥可(4)	麥克(6)	邁可(1)	邁克(1)
Richard	李查德(1)	李查羅(1)	李察(1)	理查(3)

表 2 經由音譯詞組自動抽取程序所得的部份實驗結果

在公式(7) 中使用了 AS、AP 及 RP 三種方法，但三者的效能如何尚不得而知。表 3 爲分別單獨使用 AS、AP 及 RP 三種方法來計算音節相似度，個別所產生的數量及佔總數的比重，其中第二列初步抽出數量(Raw Count)是表示音譯詞組抽取過程中未先以一般詞典先行濾除常用的英文字，故有一些非人名地名的詞彙被抽出，如『Homework』即有相對應的音譯詞『洪沃客』，這可以顯現出網路中無奇不有的特性。第三列含獨特(Unique)英文詞的音譯詞組是指僅計算不重複的英文詞時所得的音譯詞組數量及比重，因爲一個音譯詞組包含一個中文詞及一個英文詞，僅計算不重複英文詞數量，可了解平均一個英文詞所對應到的中文詞數量。第四列含獨特中英文詞的音譯詞組是指同時計算不重複的中英文音譯詞組數量及比重，以了解平均一個中英文音譯詞組的重複狀況。

從表中可看出使用語言學規則所歸納而得的音素混淆音矩陣(RP)的效果最好，而由語音辨認所得的混淆音矩陣並不如預期來的好，由語音辨認結果所產生的混淆音矩陣中分析

發現 ，若中文音節符號以 IPA 表示，如果 A 是 B 的混淆音，並不保證 B 一定是 A 的混淆音，例如『p'』常被念成『p』，因此『p'』常是『p』的混淆音，但『p』並不見得會常被念成『p'』，所以『p』不見得會是『p'』的混淆音。這也使得應用混淆音矩陣於音譯詞組抽取時，有些正確的音譯詞組可能無法被抽出。但使用語言學規則所歸納而得的音素混淆音矩陣是經過不斷的微調所產生的結果，例如上述之非雙向規則，就可以在這時候由人工直接加入規則中。

項目及方法	三種方法均使用	AS	AP	RP
初步抽出數量	10,225	4,964(485.%)	8,254(80.7%)	9,175(89.7%)
含獨特(Unique)英文詞的音譯詞組	3,742	1,887(50.4%)	3,086(82.5%)	3,412(91.2%)
含獨特中英文詞的音譯詞組	4,779	2,400(50.2%)	3,798(79.5%)	4,224(88.4%)

表 3 分別單獨使用 AS，AP 及 RP 三種方法所得的數量及比重

爲了進一步瞭解 AS，AP 及 RP 三者的影響如何。表 4 爲交叉驗證分別單獨使用 AS，AP 及 RP 三種方法所得資料之數量及比例，以了解三種方法彼此間的差異程度，其中可發現 AP 與 RP 在含獨特英文詞的音譯詞組與含獨特中英文詞的音譯詞組交集部份兩者重疊性均很高(分別爲 AP 的 98.3%及 RP 的 88.9%)，雖然在含獨特英文詞音譯詞組部分 AP 差集僅佔 1.7% ，RP 差集僅佔 22.4% ，但在含獨特中英文詞的音譯詞組 AP 的差集則上升至佔 2% ，RP 的差集則下降至 12% 。就音譯詞組自動抽取的角度而言，由 AP 來取代經過微調過的 RP 應該是可以合理的，這是因爲微調過的 RP 須具備許多的語言學知識，而且也必須耗費許多時間觀察許多可能的情形。此外雖然 RP 在含獨特英文詞的音譯詞組交集部份可以包含 86%的 AS，在含獨特中英文詞音譯詞組交集部份可以包含 82.6%的 AS，但 AS 在含獨特英文詞的音譯詞組仍有 13.9% 的獨特詞組，在含獨特中英文詞的音譯詞組仍有 17.4% 的獨特詞組，顯示 AS 不應被完全捨棄。

項目及方法	AS vs AP	AP vs RP	RP vs AS
含獨特(Unique)英文詞的音譯詞組(交集)	1,514 (80.2% to AS) (49.1% to AP)	3,034 (98.3 % to AP) (88.9% to RP)	1,624 (86% to AS) (47.6% to RP)
含獨特英文詞的音譯詞組(差集)	373(19.8% to AS) 1,572(51% to AP)	52(1.7% to AP) 764(22.4% to RP)	263(13.9% to AS) 1,788(52.4 to RP)
含獨特(Unique)中英文詞的音譯詞組(交集)	1,824 (76% to AS) (48% to AP)	3,721 (98% to AP) (88.1% to RP)	1,982 (82.6% to AS) (46.9% to RP)
含獨特中英文詞的音譯詞組(差集)	576(24% to AS) 1,974(52% to AP)	77(2% to AP) 503(12% to RP)	418(17.4% to AS) 2,242(53.1% to RP)

表 4 交叉驗證分別單獨使用 AS，AP 及 RP 三種方法所得資料之數量及比例

在[Lin2000] 曾討論到中英文詞組音譯失敗的原因，其中有一項是約定成俗但聲音不相近的音譯，由本文實驗中可以發現，由於許多人名或地名係先被引進至英文，在處理中英文詞組音譯詞組抽取時，若能把字源因素納入考慮，以原始語言的發音規則發音，則有較大的機會被抽取出來。以 Bach (巴哈)例，因爲 Bach 是一個德國人名，若能以德語發音，則更能貼近音譯是以發音方式將一詞從一個文字轉換到另一文字的特性。

4. 結論

機器音譯常用來處理文章中許多人名、地名、組織名等專有名詞，是機器翻譯中重要的一環。本文提出基於語音辨認系統所產生的混淆音矩陣，用來解決中英文音譯詞組自動抽取所面臨的發音變異問題，並自網頁中抽取出大量的中英文音譯詞組。由實驗結果發現，本文所提出的方法確實有效的處理發音變異問題。中英文音譯詞組自動抽取是研究中英文詞組音譯的第一步，未來將繼續進行中英文音節的自動轉換等相關的中英文詞組音譯研究。

5. 致謝

中華電信研究所王文俊博士提供語音辨認系統所產生的混淆音矩陣，由於這個混淆

音矩陣資料，使得中英文音譯詞組自動抽取變成可行。中華電信研究所賴玟杏小姐提供許多語言學的資訊。中央研究院資訊所簡立峯博士提供許多寶貴的意見。在此一併致謝。

6. 參考文獻

[Brill2001] Eric Brill, Gary Kacmarcik, Chris Brockett, "Automatically Harvesting Katakana-English Term Paris from Search Engine Query Logs", In *Proceedings of NLPRS'2001*

[Brown93] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation", Computational Linguistics, 19(2), pp.263-311, 1993

[Jeong97] Kil-Soon Jeong, Yun-Hyung Kwon, and Sung-Hyun Myaeng, "Construction of Equivalence Classes of Foreign Words through Automatic Identification and Extraction, NLPRS'97

[Jung2000] SungYoung Jung, SungLim Hong, and Eunok Paek, "An English to Korea Transliteration Model of Extended Markov Windows", In *Proceedings of COOLING'2000*

[Jurafsky2000] Daniel Jurafsky and James H. Martin, Speech and Language Processing, pp. 156-163, Prentice-Hall, New Jersey, 2000

[Kang2000] In-Ho Kang and GilChang Kim, "English-to-Korean Transliteration Using Multiple Unbounded Overlapping Phoneme Chinks", In *Proceedings of COLING'2000*, 2000

[Knight98] Kevin Knight and Jonathan Graehl, "Machine Transliteration", Computational Linguistics, 24(4), pp. 599-612, 1998

[Lee2003] Chun-Jen Lee and Jason S. Chang, "Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts Using a Statistical Machine Transliteration Model", In *Proceedings of NAACL*, pp. 96-103, 2003

[Lee98] Jae-Sung Lee and Key-Sun Choi, "English to Korea Statistical Transliteration for Information Retrieval", Computer Processing of Oriental Languages, Vol. 12, No. 1, pp. 17-37, 1998.

[Lin2000] Wei-Hao Lin and Hsin-Hsi Chen, "Similarity Measure in Backward Transliteration between Different Character Set and Its Application to CLIR", In *Proceedings of Computational Linguistics Conference XIII*, pp. 97-113, 2000 (in Chinese)

[Llitojos2001]Ariadna Font Llitjos and A. Black, "Knowledge of Language Origin improves Pronunciation Accuracy of Proper Names, In *Eurospeech'2001* Vol. 3, Aalborg Denmark, pp.1919-1922, 2001

[Nagata2001] Masaaki Nagata, Teruka Saito, and Kenji Suzuki, "Using the Web as a Bilingual Dictionary", In *Proceedings of ACL'2001 DD-MT Workshop*, 2001

[Oh2002] Jong-Hoon Oh and Key-Sun Choi, "An English-Korea Transliteration Model Using Pronunciation and Contextual Rules, In *Proceedings of COLING'2002*, Taipei, Taiwan, 2002

[Pagel98] Vincent Pagel, Kevin Lenzo, and Alan W. Black, “Letter to Sound Rules for Accented Lexicon Compression”, In *Proceedings of the ICSLP'98*, Sydney, Australia, 1998

[Wan98] Stephen Wan and Cornelia Maria Verspoor, “Automatic English-Chinese Name Transliteration for Development of Multilingual Resources”, In *Proceedings of 17th COLING and 36th ACL*, pp. 1352-1356, Montreal , Quebec, Canada, 1998

[Xiao2002] Jing Xiao, Jimin Liu and Tat-Seng Chua, “Extracting Pronunciation-translated Names from Chinese Texts Using Bootstrapping Approach”, In *Proceedings of COLING'2002*, Taipei, Taiwan, 2002

[NIHOGO90] 日本語知識百科，和風語言雜誌，豪風出版社，1990

[NTNU82] 國音學，國立台灣師範大學國音教材編輯委員會，正中書局，1982

[Wang2002]『雜訊語音辨認技術』，王文俊等，中華電信研究所技術報告編號 91-31-002，2002