反向異文字音譯相似度評量方法與跨語言資訊檢索

林偉豪，陳信希

國立台灣大學資訊工程學系

Page 97 ~ 113

Proceedings of Research on Computational Linguistics

Conference XIII (ROCLING XIII)

Taipei, Taiwan

2000-08-24/2000-08-25

反向異文字音譯相似度評量方法與跨語言資訊檢索

林偉豪　　陳信希

國立台灣大學資訊工程學系

Similarity Measure in Backward Transliteration between Different Character Sets and Its Application to CLIR

Wei-Hao Lin and Hsin-Hsi Chen

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, TAIWAN, R.O.C.

E-mail: b4506060@csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

Abstract

This paper classifies the problem of machine transliteration into four types, i.e., forward/backward transliteration between same/different character sets, based on transliteration direction and character sets. A phoneme-based similarity measure is proposed to deal with backward transliteration between different character sets. Chinese-English information retrieval is taken as an example. The experiments show that phoneme-based approach is better than grapheme-based approach. In a mate matching of 1,261 candidates, the average rank is 7.80 and 57.65% of candidates are ranked as number one.

摘要

本文首先根據音譯的方向是否跨不同文字系統，將機器音譯分成「正向同文字」、「正向異文字」、「反向同文字」與「反向異文字」等四種來討論。接著以相似度的比較作為音譯系統的基礎，將語音相似度分為物理聲音、音素、和形素三

個層級，並討論計算語音相似度的方式。最後，提出一個以音素相似度為基礎的方法，以中文和英文的音譯為例，進行反向異文字的音譯。實驗結果顯示在音素上的比較，比在形素上的比較來得有效。在一個 1,261 個人名的候選名單中，執行配偶配對實驗，平均排名是 7.80，其中 57.65%的排名為第一名。

1. 介紹

網際網路隨著電子商務的快速發展，更深入我們的日常生活中，也擴大了世界各國在網際網路上參與的熱度，不同語言所呈現的「內容」(content)在網際網路上傳播。舉凡許多廠商覬覦的中國大陸市場所使用的中文，或是整個歐盟成員國間的各種語言，已經讓網際網路從大部分以英文為主的內容，擴大成為多語言的內容。對網際網路的使用者、或是電腦應用系統(例如搜尋引擎、網路資源蒐集軟體、或是新聞自動摘要系統(Chen and Lin, 2000))來說，因為語言不同所形成的閱讀與處理障礙，也日漸增加。在這種多語的大環境下，機器翻譯(machine translation)與跨語言資訊檢索(cross language information retrieval)等相關自然語言處理系統研究，就極為受到重視。

所謂跨語言資訊檢索(Chen, 1997)就是以一種語言所表達的查詢(query)，去檢索另一種語言所呈現的內容。因為語言上的差異，通常需要將查詢轉換成跟內容一樣的語言。歧義分析(disambiguation)，是查詢翻譯(query translation)一項重要的研究(Bian and Chen, 2000)。根據 1995 年網路使用者，對 Wall Street Journal、Los Angeles Times 和 Washington Post 等新聞語料檢索的統計(Thompson and Dozier, 1997)，分別有 67.8%、83.4%、和 38.8%的檢索詞含專有名詞。我們知道辭典的覆蓋度，一直是查詢翻譯的重要問題，在專有名詞的翻譯更是挑戰。Chen 等人(1998)，Knight 和 Graehl(1998)，Wan 和 Verspoor(1998)都相繼提出機器音譯(machine transliteration)的方法，來處理這個問題。

音譯可以根據處理的方向，區分成正向音譯(forward transliteration)與反向音

譯(backward transliteration)。當一個語言的專有名詞，因為沒有適當或是不容易以意譯來表示時，會採用正向音譯，將其音呈現出來。例如義大利的觀光勝地Firenze，中文就音譯成「翡冷翠」，此為正向音譯。反過來說，當我們看到一個中文的音譯人名「阿諾史瓦辛格」，如果想要找出原文是 Arnold Schwarzenegger，就是反向音譯。一般來說，使用羅馬字母的拼音文字語言，會保持原詞語字母的拼法，然後以原語言的發音規則，或是自己語言的發音規則來發音。但如果在象形文字與拼音文字語言之間作音譯時，則需要將聲音由原語言盡量用另外一種語言相近的音素來表示，而且要符合目的語言(target language)的語音組合規則。很顯然地，拼音文字與象形文字之間的音譯處理相對來說較為困難，反向音譯比正向音譯更難。正向音譯允許某種程度的失真，所能夠接受的錯誤範圍較大；但反向音譯則不是。反向音譯較不允許錯誤，也就是在找出原文的過程中，必須要相當準確，否則反向音譯的結果應用性就較低。

本文第二節由音譯的正向和反向，以及是否跨文字來分析音譯問題，並介紹過去相關的研究。第三節由相似度的觀念，來執行機器反向音譯的程序。第四節提出一種以音素進行相似度比較的方法。第五節介紹實驗規畫，並對實驗結果進行討論，最後是結論。

2. 音譯分類與相關研究

根據音譯的方向，我們將音譯問題區分為「正向音譯」與「反向音譯」兩種。另外，根據音譯的原始語言與目標語言所採用的字母系統，還可以將音譯區分為「同文字系統間音譯」與「異文字系統間音譯」兩種。以下各小節，就針對這四種組合來介紹相關問題。

2-1 正向同文字間音譯

相同文字系統之間由於共用同一種文字，尤其以羅馬字母為基礎的拼音文字，不同語言在形素(grapheme)與音素(phoneme)間的組合規則雖不一樣，但是一個語言的詞語，要表達成另外一個同文字系統的語言，通常沒有問題。這類型的

音譯通常保持原始語言的文字拼法，而目的語言的使用者則以目的語言的發音規則，或是以原始語言的發音規則來發音。例如 Beethoven 雖然是德國名字，但是在英文的文本中，還是直接使用相同的文字拼法。即使在使用相同拼音字母的語言中，還是可能存在音譯。例如義大利觀光勝地 Firenze(義大利文)，英文則音譯為 Florence。

不同語言使用者在發音時，會採用自己語言的發音規則。例如英語使用者可能會依英語的發音規則來發音，這樣就跟原來德文的發音不同。但大體來說在音素上的發音較為接近，而且越來越多的人會選擇以原始語言來發音，以尊重原始語言。另外，日文中的漢字雖然與中文相通，但由於在發音上差距甚大，所以通常日文漢字翻譯成中文時，表面上與羅馬拼音文字一樣，保持原來日文漢字的寫法，但中文使用者通常會以中文的念法來對日文漢字發音。除非這位使用者學習過日文，才有辦法以正確的日文漢字來發音。

2-2 正向異文字間音譯

在正向異文字間音譯時，主要的工作在於將原始語言的音素，以目的語言的音素來呈現，並配合目的語言的組合規則表示。如果應用在書寫系統上，還要進一步將之前音譯後的結果，選擇目的語言適當的書寫文字，來呈現最後音譯的結果。Wan 與 Verspoor(1998)發展出一套自動將英文專有名詞，正向音譯成中文的系統。在將英文形素轉成音素的過程中，這個系統先將英文字母音節化(syllabification)。拆音節的方法主要有以規則為本(rule-based)，以及範例學習(instance learning)兩種。此系統採用規則為本的方式，但並不是利用上千條的規則來拆解音節，而是利用子音群(consonant cluster)與母音來當成音節的分界來拆解。由於中文為單音節的文字，且多為「子音＋母音」的結構，所以系統還要進一步將之前拆解的音節，做進一步的次音節化(sub-syllabification)。將沒有辦法以中文字發音的英文子音群拆開，並加上跟情境相關的母音，以兜成「子音＋母音」的音節。在將音素轉成目標語言(在這裡是中文)的文字過程時，Wan 與

Verspoor 的系統，先將拆解完成的音節查表轉換成漢語拼音，接著再查表將漢語拼音最後的中文音譯結果輸出。

2-3 反向同文字間音譯

如前所述，同文字系統間音譯，通常都是保持原來的詞彙組合與型態，所以並不需要做反向的音譯，來找出原始語言的詞彙到底為何。因此，這方面的處理比較簡單。

2-4 反向異文字間音譯

中文和英文間的轉換，是屬於反向且跨文字系統的音譯，這是本文所要討論的重點。在反向音譯(以後如果沒有特別說明，指的都是異文字間的反向音譯)的研究，有兩種不同的處理方式：一種是直接將音譯後目標語言的詞彙，利用某個模型反推出原始語言的詞彙；另一種是將音譯後目標語言的音譯字，與一串原始語言的候選字相比對，判斷何者可能是原來原始語言所使用的詞彙。

Knight 與 Graehl(1998)利用衍生模型(generative model)，設計一個反向音譯的系統，將音譯後的日文字反向音譯出原來的英文詞彙。當嘗試將英文(原始語言)專有名詞，音譯成日文(目的語言)片假名(katakana)時，衍生模型分成幾個階段處理，包括寫下要音譯的英文詞彙，用英文將該詞彙發音，將英文發音修改成日文可以發的音，將這個日文發音轉成片假名，並寫出片假名。假設我們有一個根據 P(w)機率分佈來產生英文字（word）的產生器，又假設我們有一個英文發音器。給定一個英文字時，發音器會依據 P(p|w)的機率來設定該字的發音（pronunciation）。對一個英文發音 p，如果我們想要找出這個發音可能的英文字時，我們就可以尋找看看哪一個英文字 w 可以讓 P(w|p)這個機率有最大值。根據貝式定理（Bayes' Theorem），這相當於尋找 P(w)・P(p|w)。這個系統用到如下五個機率分佈，其中 w 為英文字、e 為英文發音、j 為日文發音、k 為片假名、o 為光學辨識出來的字元：

(1) P(w)：產生英文詞彙。

(2) P(e|w)：英文詞彙發音。

(3) P(j|e)：將英文發音轉成日文發音。

(4) P(k|j)：將日文發音轉成片假名。

(5) P(o|k)：加入因為光學字元辨識所產生的錯誤。

當 OCR 取得一個片假名字串 o 時，反向音譯使用下面的公式，找出英文字串 w。

$$\arg\max_{w} P(w)\times P(e\mid w)\times P(j\mid e)\times P(k\mid j)\times P(o\mid k)$$

Chen 等人(1998)提出一個將英文音譯成中文(目的語言)的音譯字，反向音譯回英文(原始語言)的模組，並應用於中英跨語言資訊檢索系統。這個系統是將可能的音譯字辨識出來，再進行反向音譯。首先利用漢字羅馬拼音系統(例如 Wade Giles (威翟)，或是漢語拼音(Pinyin))，把可能的音譯字(中文)轉成羅馬字母。接著將這個詞彙與一串可能的專有名詞進行比對，藉此找出可能的原文(英文)。

3. 語音相似度

本篇論文把音譯問題視為相似度的衡量。正向音譯即是在不同語言之間，讓音譯後的結果能夠保持最大的相似度。在反向音譯，如果預先給予一份候選名單，則系統比較音譯字與候選名單上的詞彙，計算兩兩相似度。相似度的比對，可以分成三個層次：形素、音素、和物理聲音。

音譯後的詞彙與原詞彙之間，最直接的比較方式，就是請母語使用者發音，然後以物理上可以測量到的音波來比較。如果從人類可以發出的語音來看，音素集合是固定且有限的，我們可以嘗試在音素的層次來比較。兩個音素的發音位置，或是發音方式越相近，兩個聲音也會越相似。當我們以書寫文字來比較時，就是直接比較形素的相似度。如果書寫文字系統不同，例如中文的方塊字，與英文的羅馬拼音文字，就必須先轉換到相同的字母集合，才能進行比對。

在形素上的比較，Odell 與 Russell 的 Soundex 系統(Knuth, 1973)，是屬於同語言的羅馬拼音字母，利用子音來捕捉詞彙發音的特性。當兩個詞彙的子音位置

與發音相似時，表示這兩個詞彙的發音就可能越相似。而 Chen 等人(1998)的研究，可以視為在形素上比較相似度的反向音譯系統。由於所討論的中文音譯字，與原始語言英文的書寫系統不同，他們先將音譯字轉換成羅馬字母，這個動作稱為「羅馬拼音化」(romanization)。他們所採用的標準拼音系統，有威翟與漢語拼音，並加上一些經驗法則修正，來提高系統效能。

由於羅馬拼音系統，主要並不是考慮語音上的相近來設計，例如漢語拼音就用到了 Zh、Q 與 X 等羅馬字母，來表示與字母發音完全無關的漢語語音，所以英文音譯成中文的音譯字，在利用羅馬拼音系統轉換成羅馬拼音字母後，這些羅馬拼音字母，跟原來詞彙的拼音字母，在發音上並不十分相近。

有鑑於在形素層次上做羅馬拼音化時，非常需要一個以形素相近為出發點而設計的的羅馬拼音系統。例如在中文和英文這兩種書寫系統完全不同的語言，我們可以設計一個「自動建立羅馬拼音對照表」的系統。這個系統分為兩個階段：第一個階段是訓練，我們從已知的英-中音譯字與原文詞彙的配對中，學習英中音譯字所應該轉換的羅馬拼音字母。例如 Elton 與「愛爾頓」這個配對，先將中文代換成注音符號後，然後分別對兩個字做音節拆解的動作，得到「El・ton」與「ㄞ・ㄦ・ㄉㄨㄣ」，這裡忽略英文重音與中文聲調符號，而・為音節間隔符號。接著進一步將英文音節做次音節化後，我們就可以得到英文音節與中文字的音節對應共三組，包括「ㄞ→e」、「ㄦ→l」與「ㄉㄨㄣ→don」。第二個階段實際從事形素相似度衡量，系統根據前一個階段訓練所得到的對照表，將英-中音譯字轉換成英文詞之後，再與候選名單相比較。如前例，「愛爾頓」先轉換成注音符號「ㄞ・ㄦ・ㄉㄨㄣ」，然後查表後得到「e・l・don」。拿掉音節符號後就得到「eldon」，然後再做配偶配對(mate matching)。

表一列出上述例子的訓練結果。跟其他羅馬拼音系統來比較，我們可以發現：由這個系統所產生的對應，在形素上比其他拼音系統來得更接近實際情形。像是英-中音譯字中的ㄦ，例如貝爾（Bell）中的「爾」字，如果採用其他拼音系

統來做形素上的比較時，可以發現其他系統完全配對失敗（er≠l），只有經過訓練階段所產生對應才能正確配對（l=l）。換句話說，這個系統的對照表是比較有效的，所以能夠在形素層次上的相似度比較，有更好的效能。

表一・訓練結果與羅馬拼音系統

注音符號	威翟	耶魯	漢語拼音	注音符號第二式	「自動產生羅馬拼音對照表」系統的結果
ㄞ	ai	ai	Ai	Ai	e
ㄦ	erh	er	Er	Er	l
ㄉㄨㄟ	tun	dwei	Duan	duan	don

4. 音素相似度評量

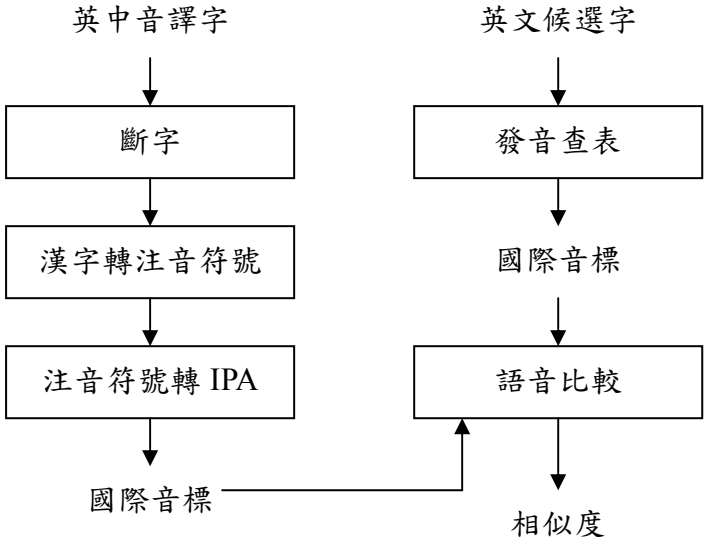
考量物理發音在跨語言資訊檢索的實用性，以及形素層次上比對的事前訓練，因此音素層次上的相似度比較易顯重要。而衡量兩個詞彙的音素相似度，我們提出一個以國際音標(International Phonetic Alphabet，IPA)為基準的比較，先將兩個詞彙的國際音標列出來，然後比較國際音標的相似度，進而達到反向音譯的目的。圖一顯示音素相似度比較的流程。我們先說明流程圖左邊的部分，也就是英中音譯字處理的部分：

(1) 斷字：在收到英中音譯字時，第一個步驟即是取出其中的中文字，也就是斷字。例如「亞瑟」這個音譯字，經過斷字後取出「亞」與「瑟」。

(2) 漢字轉注音符號：將前一個步驟斷出來的漢字，經查表後得到相對應漢字的注音符號。例如「亞」查表後，得到「ㄧㄚ」，在此我們忽略聲調符號。

(3) 注音符號轉 IPA：將前一個步驟中的注音符號，經查表二後，得到相對應注音符號的 IPA。表二列出母音和子音與 IPA 對照，這部份參考謝國平(1998)，並略作修正。例如「ㄧ」查表後，得到「ɪ」。一般 IPA 的表示必須配合特殊字體，才能顯現，CMU pronunciation dictionary 0.6 版(簡稱 CMU dict)

(ftp://ftp.cs.cmu.edu/project/fgdata/dict/)採用 ASCII 來表示，附錄列出 CMU dict 符號和 IPA 符號對照。例如，「ɪ」對應「IY」。



圖一・音素比對

表二・注音符號與 IPA 對照表

(a) 子音部份

注音符號	ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ	ㄐ	ㄑ	ㄒ	ㄓ	ㄔ
IPA	π	πH	μ	ϕ	τ	τH	ν	λ	κ	κH	ξ	τ˩	τ˩H	˩	τ♣	τ♣H
注音符號	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ											
IPA	♣	|	τσ	τσH	σ											

(b) 母音部份

注音符號	ㄧ	ㄨ	ㄩ	ㄚ	ㄛ	ㄜ	ㄝ	ㄞ	ㄟ	ㄠ	ㄡ	ㄢ	ㄣ	ㄤ	ㄥ	ㄦ
IPA	ι	υ	ψ	α	ο	Φ	ε	αι	ει	αυ	ου	αν	↔ν	αN	↔N	™

對候選名單中的音譯字，處理的方式也是類似。每一個英文詞彙，我們查表(CMU dict 0.6)，以得到該詞彙的發音。例如「Arthur」的發音，經查表後得到「AA R TH ER」(忽略重音標示)。

「語音比較」是整個流程最重要的部分，當我們拿到兩串 IPA 時，如何比較這兩個 IPA 字串的相似度呢？首先來看以下三個定義，由字母對齊相似度、到字

串對齊相似度，最後到字串相似度。

定義 1：字母對齊相似度

假設 S_1 與 S_2 這兩個字串的字母集合為 Σ，而 Σ'表示 Σ 加上「_」(_表示空白字元)，給予 Σ'中的兩個字元 x 與 y，s(x, y)表示 x 與 y 對齊後，所得到的分數，稱為字母對齊相似度。

定義 2：字串對齊相似度

假設 A 為字串 S_1 與 S_2 的某一種對齊方式(alignment)，S_1'與 S_2'為插入空白後的字串。如果 S_1'與 S_2'的長度為 l，則對齊方式 A 的分數如下：

$$\sum_{i=1}^{l} s(S_1'(i), S_2'(i)) \text{。}$$

我們以一個例子說明上述定義。如前例，「亞瑟」經查表後，得到的發音是「IY AA S r」，而 Arthur 的發音為「AA R TH ER」，所以此時的 Σ={AA, ER, IY, R, r, S, TH}，而音素間彼此的分數，以下面的對稱矩陣表示：

S	AA	ER	IY	R	r	S	TH	_
AA	5	0	0	-10	0	-10	-10	-5
ER	0	5	0	-10	8	-10	-10	-5
IY	0	0	5	-10	0	-10	-10	-5
R	-10	-10	-10	10	-10	-10	-10	-5
r	0	8	0	-10	5	-10	-10	-5
S	-10	-10	-10	-10	-10	10	8	-5
TH	-10	-10	-10	-10	-10	8	10	-5
_	-5	-5	-5	-5	-5	-5	-5	-5

下面這個對齊方式：

亞瑟	IY	AA	_	S	r
Arthur	_	AA	R	TH	ER

依定義 2 所給定的字串對齊相似度分數為：-5 + 5 + -5 + 8 + 8 = 11。

然後我們來定義字串相似度。

定義 3：字串相似度

給定一個字母集合 Σ'，和成對的分數矩陣。字串 S_1 與 S_2 的相似度，定義

成 S_1 與 S_2 的最佳對齊方式 A 的值，也就是最大的字串對齊相似度值。相似度跟相關的最佳對齊方式，可以用 dynamic programming 的方式來找出。Gusfield (1997)曾定義基底條件(base condition)為

$$V(i,0) = \sum_{1 \le k \le i} s(S_1(k), \_)$$
$$V(0,j) = \sum_{1 \le k \le j} s(\_, S_2(k))$$

一般的 recurrence 式可以寫成：

$$V(i,j) = \max[V(i-1,j-1) + s(S_1(i), S_2(j)),$$
$$V(i-1,j) + s(S_1(i), \_),$$
$$V(i,j-1) + s(\_, S_2(j))]$$

$0 \leqq i \leqq \text{length}(S_1)$，$0 \leqq j \leqq \text{length}(S_2)$，$V(0, 0) = 0$。其中 V(i, j)為 $S_1[1..i]$與 $S_2[1..j]$這兩個前字串(prefix)，最佳對齊方式的值。假設 S_1 與 S_2 的長度各為 n 與 m，則最佳對齊方式的值就是 V(n, m)。如果利用 dynamic programming 的方式來求，這個值可以在 O(nm)的時間內算出來。

在我們的反向異文字音譯語音相似度評量中，Σ'為 63 個 IPA 音標符號(含空白)，其中英文有 39 個，中文除了共用的之外，另外還有 24 個中文所獨用的符號，所以整個分數矩陣的大小為 63x63。我們對分數矩陣中的分數指定方式如下：

(1) 原則上，IPA 匹配(match)給 10 分，不匹配(mismatch)扣 10 分。但若匹配的為母音，則只給 5 分，而母音不匹配不扣分。這裡我們希望利用母音來捕捉音節的對齊，所以母音不對齊不扣分。但由於母音在不同語言間的匹配，意義較不顯著，因此相同的母音只給 5 分。

(2) 與空白字元(_)對齊，可以看做 insertion 或是 deletion。由於不匹配可以看成是一個 insertion 加上一個 deletion。例如 abcdfgh 和 abcdigh，其中 f 與 i 未匹配，當要對齊時，可以採用如下的方式：

abcdf_gh

abcd_igh

所以未匹配要扣的分數，跟兩個字元對上空白，亦即做一次 insertion 和一次 deletion 要相同，這樣才沒有偏好。因此，為了公平起見，我們讓 insertion 或是 deletion 的扣分，等於不相同配對的一半，也就是 10/2=5 分。另外，關於空白對空白分數還是設-5(參考分數矩陣範例)，原因是兩個字串 ab 與 ac 在對齊時，如果 a 對 a 匹配給 10 分，不匹配扣 10 分，則 ab 和 ac 字串對齊相似度為：10 + (-10) = 0 分。如果加上空白，再進行對齊，如 a_b 和 a_c，這樣的分數為 10 + (-5) + (-10) = -5 分。也就是在對列時，同時加上空白是沒有用的，只是會把分數拉低，所以空白對空白是-5 分。

(3) 其他根據發音位置與發音方式的相近，中英文在音譯上的習慣、中英文各自的發音特性、將某些音標之間的配對分數設為 8 分，如表三所列。

表三・其他音標之配對

理由	例子
中文不分清濁	P 與 B、D 與 T、F 與 V、G 與 K、S 與 Z
發音方式與位置相近	B 與 Ph、K 與 Kh、D 與 Th、P 與 Ph
發音位置相近	L 與 R、DH 與 Th
發音方式相近	CH 與 Tch、CH 與 TSch、H 與 Th、G 與 Tc、JH 與 Tc，L 與 R、M 與 ANG、N 與 AN、N 與 AHN、N 與 ANG、NG 與 ANG、NG 與 AN、NG 與 AHNG、S 與 Sc、S 與 c、S 與 TH、S 與 TS、Z 與 Sc、Z 與 TS、Z 與 TSc
音譯習慣以及跨語言所造成的音標空缺	K 與 Tc、L 與 e、R 與 e、TH 與 Th、ZH 與 Tch、ER 與 r、ER 與 L、ER 與 e、UW 與 V、JH 與 TSc、G 與 Tch
中文不分長短母音	IH 與 IY、UW 與 W
半母音與母音	IY 與 Y

5. 實驗結果

我們採用配偶配對(mate matching)的方法，來評估語音的相似度。方法如下所述：給予已知的原始語言詞彙 o_i，與音譯後的目的語言詞彙 t_i 的配對清單集合，$\{(o_1, t_1), (o_2, t_2), \ldots, (o_n, t_n)\}$。當讀入音譯後的目的語言詞彙 t_k 時，測量語音相似度的系統，對整個清單中的每個原始語言詞彙作相似性比對，並計算每一對相似度的分數。之後再看看正確的原始語言詞彙 i_k，落在依分數高低排序的配對結果中的名次。名次越高，表示語音相似度比較越準確。

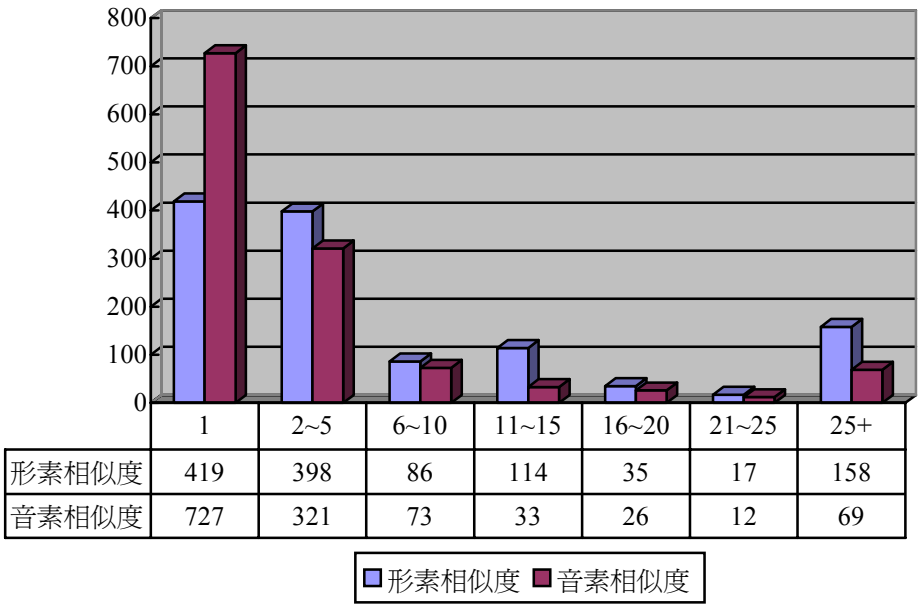
集合中的每一個目的語言詞彙，都設定名次後，我們就可以取這些名次平均值，作為一個語音相似度比較方法的評量標準。另外一個標準化的指標，是將這個平均的名次除以整個配對集合的個數 n。這表示當一個音譯後的詞彙輸入後，系統需要提出多少個候選詞彙，才會包括到正確答案。

根據這項評量語音相似度的方法，我們採用與 Chen 等人(1998)實驗相同的候選名單，共 1574 個人名。扣除無法找到發音的人名 313 個，合格的候選名單共 1,261 個人名。表三列出音素相似度和形素相似度的結果。本文所採用的音素相似度平均排名為 7.80，比 Chen 等人所採用的形素相似度平均排名 9.69，表現還要好。

表三・評估結果

	音素相似度	形素相似度
平均排名	7.80	9.69

圖二進一步列出排名分佈情況。



圖二・排名分佈

從以上結果我們可以清楚發現，在語音相似度的比較上，音素層次比形素層次表現的好。不僅平均排名上，音素相似度比形素相似度的平均排名好。如果進

一步觀察名次的分佈，音素相似度有 57.65%的結果都是最相似的，也就是正確答案。反觀形素相似度，只有 33.28%。

進一步觀察實驗結果中匹配失敗的配對，我們可以將失敗的原因歸類如下：

(1) 約定俗成但聲音並不相近的音譯：由於中文與英文是兩種不論書寫與發音都是相差甚遠的語言系統，因此不論對專業譯者或是一般作家，音譯並不是一件容易的事。但是一些已經約定俗成的翻法，例如 Bach（巴哈）、Caesar（凱薩）、John（約翰）等音譯，在音素上卻不十分相近，所以在音素層次上比對的效果不好並不令人意外。

(2) 英文非重音節的子音被忽略：英文中不在重音節的子音（通常是靠近結尾的部分），由於在中文使用者的語音知覺上並不明顯，所以音譯時經常就直接省略不翻，例如：Briand（白里安）中結尾的 d 在音譯成中文時就沒有被翻出來。

(3) 插入的母音造成混淆：由於中文字為單音節且多為子音加母音（CV）的結構，當英文字要轉換成中文時，勢必要在適當的地方插入母音才能構成 CV 結構。例如 Paul（保羅）與 Young（楊格）中結尾的ㄜ，原來是一個音節的字，到中文變成了兩個音節，也造成在音素層次上比對的困擾。

(4) 不是追求聲音接近的翻法：在一些特別的場合，特別是書寫的文本，翻譯者並不純粹追求聲音上相近的音譯方式，而可能為了與中文命名法相近（像 Gertrude，葛麗露）、或是為了簡潔（像 Gillian，姬兒），或是一味因襲傳統音譯方式卻忽略聲音上是否相近（像 Patricia，珮格麗特），這些在在都造成音素上的比對並不成功。

雖然如此，這些配對失敗的中文音譯字，比較音素相似度方法所找出來的最相似字，仍然反應音素上的相近。例如「保羅」雖然跟正確答案 Paul 並不十分相近，但系統比對得到的 Polo 在音素上其實是比 Paul 來得比「保羅」更接近；

又或「姬兒」雖然無法對到 Gillian，但是這個方法找到的 Jill 也是更接近「姬兒」，讓我們對這個方式在音素上比較相似度的能力深具信心。

6. 結論與未來的研究

機器音譯研究中，最具挑戰性，也最具實用價值的問題，就是在跨文字系統的反向翻譯。這種反向音譯在跨語言資訊檢索，或是機器翻譯時，都是一個不能忽略的問題。利用語音相似度的原理，從事反向音譯時，如果相似度的比較層次分為物理聲音、音素、與形素，而物理聲音無法進行時，我們發現音素層次上的比較，比之前在形素層次上的比較來得準確。

根據 Knight 與 Grahel(1998)對音譯系統的評量標準，這個以音素相似度來進行反向音譯作業的方式，相當接近人類在判斷音譯字是否相近，因為音素比較接近實際的聲音，而形素通常差距較大。而這個方法在應用到其他語言配對時，只要給定不同的配分矩陣就可以。最後，這個方法可以根據分數的高低，來提供一串可能的清單。所以，這個方法不管在理論與實際應用上都是深具價值。

機器音譯並不完全只是在語音上追求相等，有的專有名詞翻譯，因為歷史因素或是語言使用者的習慣，採取意譯而不是音譯。例如國家名稱 the United States，在大部分的中文文件中都是意譯成「美國」，而不採取音譯。同時，並不是所有專有名詞都採取意譯，例如 British Virgin Island 中的 Virgin，在中文音譯成「維京」，而不是採取意譯，Island 則直接翻譯成島。因此，在反向異文字音譯處理之前，先將地名送進雙語字典。如果已有現存的翻譯，就直接採用此翻譯。如果沒有，再檢查有沒有關鍵詞。關鍵詞查雙語辭典，其餘部份才經反向異文字音譯處理。

參考文獻

Bian, Guo-Wei and Chen, Hsin-Hsi (2000) "Cross Language Information Access to Multilingual Collections on the Internet," *Journal of American Society for*

Information Science, **51**(3), 2000, pp. 281-296.

Chen, Hsin-Hsi (1997) "Cross-Language Information Retrieval," *Proceedings of ROCLING Workshop on ED/MT/IR*, Academic Sinica, Taipei, June 2, 1997, pp. 4-1~4-27.

Chen, Hsin-Hsi; Huang, Sheng-Jie; Ding, Yung-Wei and Tsai, Shih-Chung Tsai (1998) "Proper Name Translation in Cross-Language Information Retrieval," *Proceedings of 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Quebec, Canada, August 10-14 1998, pp. 232-236.

Chen, Hsin-Hsi and Lin, Chuan-Jie (2000) "A Multilingual News Summarizer," *Proceedings of 18th International Conference on Computational Linguistics*, July 31-August 4 2000, University of Saarlandes.

Gusfield, Dan (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, 1997, Cambridge University Press.

Knight, Kevin and Graehl, Jonathan (1998) "Machine Transliteration," *Computational Linguistics*, Vol. 24, No. 4, 1998, pp. 599-612.

Knuth, Donald E. (1973) *The Art of Computer Programming, Volume 3, Sorting and Searching*, Addison-Wesley, Reading, Mass, 1973, pp. 391-392.

Thompson, P. and Dozier, C. (1997) "Name Searching and Information Retrieval," *Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, 1997.

Wan, Stephen and Verspoor, Cornelia Maria (1998) "Automatic English-Chinese Name Transliteration for Development of Multilingual Resources," *Proceedings of 17th COLING and 36th ACL*, 1998, pp. 1352-1356.

謝國平(1998)，*語言學概論*，三民書局，台北，1998 年 10 月。

附錄・CMU dict 符號與 IPA 符號對照表

本表根據 CMU dict 0.6 版所訂定， *表示該符號原來不在 CMU dict 中，我們為了中英音譯而增加。

cmu dict 符號	IPA 符號	cmu dict 符號	IPA 符號	cmu dict 符號	IPA 符號
AA	α	M	μ	*Tc	τ⌡
AE	Θ	N	ν	*Tch	τ⌡H
AH	℘ or ↔	NG	N	*c	⌡
AO	□	OW	o	*TSc	τ♣
AW	αυ	OY	oι	*TSch	τ♣H
AY	αι	P	π	*Sc	♣
B	β	R	ρ	*Zc	⏐
CH	τΣ	S	σ	*TS	τσ
D	δ	SH	Σ	*TSh	τσH
DH	Δ	T	τ	*r	Φ
EH	E	TH	T	*AIY	αι
ER	™	UH	Y	*EYIY	ει
EY	ε	UW	υ	*AUW	α
F	ϕ	V	ϖ	*OWUW	o
G	γ	W	ω	*AN	αν
HH	η	Y	φ	*AHN	↔ν
IH	I	Z	ζ	*ANG	αN
IY	ι	ZH	Z	*AHNG	↔N
JH	δZ	*Ph	πH	*e	™
K	κ	*Th	τH	*y	ψ
L	λ	*Kh	κH		