多語語碼轉換之未知詞擷取

Unknown Word Extraction from Multilingual Code-Switching Sentences

吳依倫　Yi-Lun Wu
元智大學資訊管理學系
Department of Information Management
Yuan Ze University
s986301@mail.yzu.edu.tw

謝佼彣　Chaio-Wen Hsieh
元智大學資訊管理學系
Department of Information Management
Yuan Ze University
s971641@mail.yzu.edu.tw

林瑋軒　Wei-Hsuan Lin
元智大學資訊管理學系
Department of Information Management
Yuan Ze University
s971636@mail.yzu.edu.tw

劉君毅　Chun-Yi Liu
元智大學資訊管理學系
Department of Information Management
Yuan Ze University
s971647@mail.yzu.edu.tw

禹良治　Liang-Chih Yu
元智大學資訊管理學系
Department of Information Management
Yuan Ze University
lcyu@saturn.yzu.edu.tw

摘要

在多語環境下，一段語句可能發生由一種語言轉換到另一種語言的現象，也就是說，語句由兩種或兩種以上的語言所組成，此即爲語碼轉換(code-switching)現象。以我國語言

使用的情況來說，國語夾雜台客英短語的現象在日常生活中已相當普遍，這些語言混用現象也造成了語言處理上的重大挑戰。有鑑於此，本論文收集中英、國台及國客夾雜之文字語料，並分析以國語爲主要語言之中英、國台及國客夾雜現象，接著提出以交互資訊(mutual information)與熵(entropy)爲基礎之未知詞擷取演算法，自動從多語夾雜語料中找出未知詞。實驗結果顯示本論文所提出的方法可藉由過濾無關的新詞提升未知詞擷取之精確度。

關鍵詞：語碼轉換、未知詞擷取、交互資訊、熵

一、緒論

語音及語言處理在人機介面應用中扮演相當重要的角色。近年來，在國內外研究學者共同努力下，語音及語言處理技術已有顯著的進步並逐漸實現在許多應用之中，如：語音辨識與合成(speech recognition and synthesis)、口語對話系統(spoken dialog system)及語音查詢與檢索(speech query and retrieval)等。然而，全球共有超過 6,900 種語言[1]，加上全球化趨勢與國際交流日益頻繁，人們對於多語(multilingual)服務的需求也逐漸增加，例如：出國旅遊時即可能需要多語點餐、導覽甚至緊急醫療服務；國際化企業亦可藉由多語電話客服系統協助其全球客戶解決問題。因此，現行系統如何支援多種語言便成爲目前語音及語言處理技術重大的挑戰之一。

在多語環境下，一段語句中可能發生由一種語言轉換到另一種語言的現象，也就是說，語句由兩種或兩種以上的語言所組成，此即爲語碼轉換(code-switching)或語言混合(language mixing)現象[2][3]。這種現象較常發生在使用雙語或多語的地區，一般語者受其文化及教育的影響，在全球化及現代化的過程中對於本地方言及外來語接受較高而成爲雙語或多語使用者，因此在使用語言時，常會因爲不同的習慣、場合及對象而產生多語夾雜或混用的語句。這些多語夾雜語句的特性之一就是以一種語言爲主要語言(primary language)，而其他次要語言(secondary language)以字詞或片語等短語的形式夾雜其中。以台灣地區來說，除了一般通行的國語(Mandarin)外，國人對於台語(Taiwanese)甚至是客語(Hakka)的使用也很普遍，加上中英雙語教學的推行及教育水準不斷提高，學習英語早已成爲全民運動，因此，日常生活中常會出現國語夾雜台語、客語及英語等混用現象，這些現象亦常出現在新聞媒體、報章雜誌及網路文件中。下列句子即是以國語爲主要語言，夾雜英語(E)、台語(T)及客語(H)短語之範例句 (資料來源：網路新聞與搜尋結果)。

(E1) 兩岸 **ECFA** 即將進入正式協商。
(E2) 享受樂活舒壓的 **SPA** 活動。
(T1) 選情緊繃，候選人四處趴趴走拜票。
(T2) 這裡有一家很傳統的柑仔店。
(H1) 客家對於天穿日非常重視。
(H2) 四炆四炒是客家菜的代表菜色之一。

在語碼轉換相關研究中，大部分仍著重從語言學或社會學的角度探討語碼轉換現象，對於自動處理語碼轉換語音及語言的相關研究並不多見，而國內近年來已有部分研究團隊積極朝此領域發展。在語料庫方面，有台師大、清大、交大、成大及台大共同錄

製的 EAT (http://www.aclclp.org.tw)語料庫，其中即包含中英夾雜句，而長庚大學亦有錄製國台客多語語音資料庫 ForSDat (Formosa Speech Database) [4]；在語音辨識方面，成大在中英混合語音辨識上已有成果發表[5][6]，長庚及元智大學亦有從事國台、國客混合語音辨識之研究[7][8]；其它地區則有學者針對廣東語-英語混合句進行語音辨識[9][10]，中英日等六種語言之聲學模型[11]，或是中英混合句的語音合成[12][13]。

上述多語語碼轉換相關研究著重於語料庫建立、語音辨識與合成之研究，對於詞典擴增(lexicon augmentation)、語言模型(language modeling)及語音辨識後語言理解(spoken language understanding)等語言處理層面的議題較少探討，然而這些議題在語音處理技術上亦扮演重要的角色，例如：在詞點擴增的議題中，(E1)中的 ECFA 即無法在英文詞典找到，(T1)與(T2)中的"趴趴走"與"柑仔店"也無法從台語字典找到，這些未知詞都會影響後續語言模型及語音理解的效果。有鑑於此，本論文收集中英、國台及國客夾雜之文字語料，並分析以國語爲主要語言之中英、國台及國客夾雜現象，接著提出以交互資訊(mutual information)與熵(entropy)爲基礎之未知詞擷取演算法，自動從多語夾雜語料中找出未知詞。本論文章節安排如下：第二章簡介多語夾雜語料庫及分析結果；第三章說明未知詞擷取演算法；第四章爲實驗結果；第五章爲結論。

二、多語夾雜語料收集與分析

(一) 多語夾雜語料收集

本論文所探討的語碼轉換現象限定於國台客英四種語言，並且以國語夾雜台語、客語及英語短語的混用現象爲主，分析這些現象可進一步瞭解多語夾雜語料庫的特性，包括容易發生夾雜現象的句型或語法結構等，亦有助於後續未知詞擷取、語言模型及語言理解模組之設計。中英及國台夾雜語料較爲常見，透過網路 BBS、Blogs、討論區等收集有關新聞時事、旅遊、美食等主題即可取得語料，至於國客夾雜語料收集上較爲困難，但網路上仍有部分專業網站提供客語學習等相關資源，例如：行政院客家委員會推動的「哈客網路學院」 (http://elearning.hakka.gov.tw/)，站內提供一系列的客語能力認證教材，並可超連結至「臺灣客語詞彙資料庫」(http://wiki.hakka.gov.tw/)，該資料庫將客語詞彙區分爲 30 大類，並提供四縣、饒平、六堆、海陸、美濃、詔安及大埔等七種腔調之字典檔，總計 35,605 個詞彙，每個客語詞彙均有音標以及國語及英語辭義解釋，更重要的是有提供客語造句範例及對應的國語譯句，例如：海陸腔詞彙檔中編號 01-013 的詞彙「風搓」，代表第一大類天文地理中的第十三個詞彙，其國語解釋爲颱風，客語例句及國語譯句如下。

人講：「一雷壓三搓」，𣍐看這擺个風搓怕毋會登陸咧。　　(客語例句)
俗話說：「一雷壓三颱」，我看這次的颱風可能不會登陸了。　　(國語譯句)

由於所謂國客夾雜語句係指以國語爲主要語言夾雜客語短語之語句，因此可將客語詞彙取代國語譯句中意義相同的詞彙得到國客夾雜語料，例如：上例中，若將客語詞彙風搓取代國語譯句中的颱風即可得到一國客夾雜句子，不過當字典檔中國客詞彙相同時，這種詞彙取代的方法將失效，因此，我們初步從海陸字典檔 4,959 個詞彙之例句中，扣除國客詞彙相同的例句，再以詞彙取代的方法產生共 1,275 句國客夾雜語料。

(二) 語碼轉換現象分析

在多語夾雜語句中，雖然台客英等短語可能出現在國語語句內任何位置，但在實際使用上並非完全沒有規則可循，目前已有學者針對中英夾雜語句分析英語短語出現的語法結構及樣式(pattern)[2][3]，我們依其整理出的樣式分析從網路收集到的國台客英夾雜語料，部分結果如表一所示。爲了進一步分析這些夾雜短語的詞性與型態，我們隨機選取 500 份網路新聞以人工方式找出夾雜的台語及英語短語，其詞性與型態分佈情形如表二及表三所示。由表二的統計結果發現，中文語句中所夾雜之英語短語約有 90%是名詞，並以人名、地名及組織名較爲常見，而動詞僅佔了約 10%，這顯示了在表達動詞時一般較常使用母語，而表達名詞時，因爲許多公司、餐廳及地方等名稱本來即是英文，直接以英文表達並不會造成溝通的不便，因此並不常翻譯成中文使用，甚至有些情形下直接以英文表達反而比翻譯更容易理解。至於國語語句中夾雜台語短語的分佈情形，由表 3 的統計結果可發現動詞約佔 70%，名詞僅佔約 24%，結果與英語短語的分佈相反，這顯示了以台語表達動詞較爲常見，例如：挫著等, 阿莎力, 趴趴走等，原因除了說話者個人的習慣之外，使用這些詞彙較能以通俗的方式表達語意也是可能的原因之一，至於名詞較少使用台語短語可能是使用台語命名的地名、組織名較少的原因。另一方面，不論是台語的名詞或動詞短語其型態較不一定，較難呈現分類上的趨勢，不像英語名詞短語較集中分佈在人名、地名及組織名，因此表三並未進一步細分台語短語各詞類之型態。

表一、國語夾雜台客英短語之句型樣式

編號	句型樣式	範例句
1	程度副詞(Dfa) + 短語(Adj) (e.g., 很、非常、相當、最)	這道料理**非常 smooth** 入口即化 這家餐廳的老闆**很阿莎力** 隔壁的小孩一點也不瘦，反而很**大箍** (客)
2	短語(Adj) + 的(DE)	這裡都沒有可以 **shopping 的**地方 終於體驗到甚麼是**足感心的**服務了 小孩子什麼都不肯吃，所以餵養得**瘦夾夾的** (客)
3	的(DE) + 短語(Noun)	紐約**的 pizza**，單片就幾乎比臉大 好想念劉文聰**的番仔火**跟雞蛋糕啊！ 媽媽**的黃瓠粄**做得很好吃 (客)
4	1. + 2. 2. + 3. 1. + 2. + 3.	介紹你一家我**很尬意的**火鍋店 哪一間 **hotel 的 view** 最棒？ 這附近有一些**很古早的柑仔店** 他是**很擄人的**小孩，總愛耍無賴 (客)
5	數量定詞(Neqa) + 短語(Noun) (e.g., 很多、許多、一些)	我們還有**一些 issue** 要解決 這夜市有賣**很多賊仔貨** 吃飯前不要吃那麼多**零嗒** (客)
6	短語(Noun) + 位置詞(Ncd) (e.g., 上、下、內、裡、旁)	我們約在 **lobby 旁**的水池見面 **烘爐地上**有一尊超級大的土地公神像 晚飯後，祖父常在**天墀坪裡**唱山歌 (客)

表二、中英夾雜語料英語短語之詞性與型態分佈

詞性		數量	比例		範例
名詞	人名	89	25.14%	**90.4%**	John Culver, Kobe, Paul Hertz
	地名	80	22.60%		Boston, London, Paris
	組織	70	19.77%		NASA, NIKE, NHK, LV, Sony
	單位	14	3.95%		cm, GHz, kg
	食物	13	3.67%		bagel, coffee, salad
	其他	54	15.25%		cartoon, CPR, I-phone, MSN, MVP
動詞	—	34	9.6%	**9.6%**	call-in, DIY, po, shopping
合計		**354**	**100%**		

表三、中英夾雜語料台語短語之詞性與型態分佈

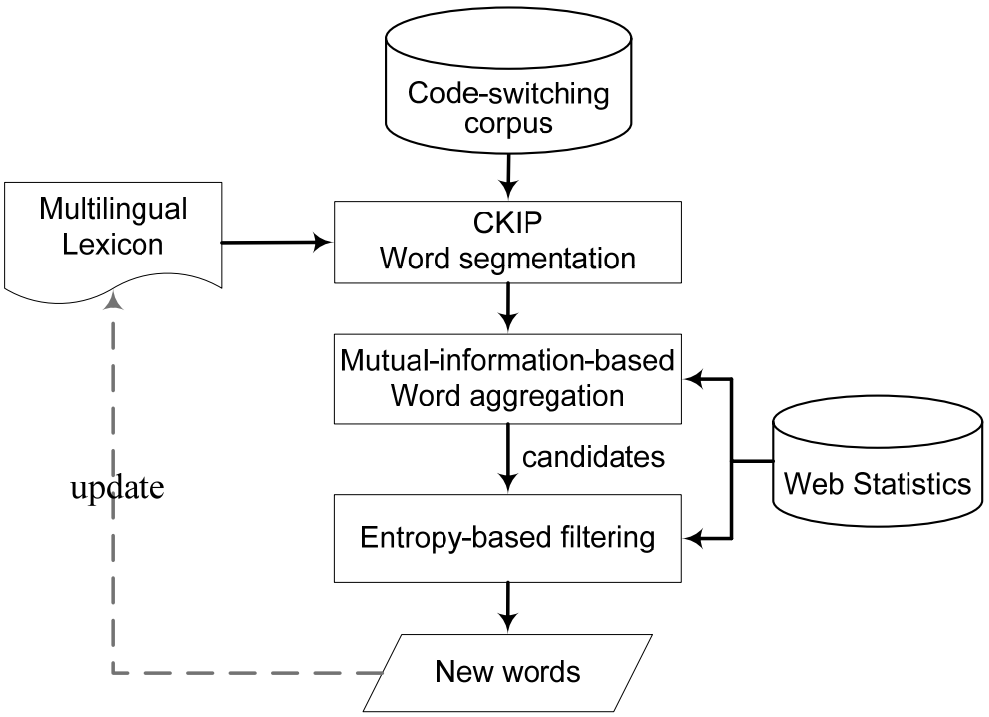
詞性	數量	比例	範例
名詞	17	23.61%	咱, 阮, 月娘, 運將, 囝囡, 代誌, 天公伯, 古早厝
動詞	50	69.44%	尬意, 讀冊, 拍謝, 假仙, 挫著等, 阿莎力, 趴趴走
副詞	2	2.78%	攏, 嘛
疑問詞	3	4.17%	安怎, 蝦米, 衝啥
合計	**72**	**100%**	

三、未知詞擷取

多語夾雜語料庫中夾雜的短語有些並未出現在目前的字典中，因此可能無法正確的斷詞，例如：”趴趴走”即無法被 CKIP 斷詞系統(http://ckipsvr.iis.sinica.edu.tw)[14]斷成一個詞，因此本年度另一項工作就是從多語夾語料庫中找出未知詞，尤其是夾雜短語的部分。我們提出的未知詞擷取演算上做法上是先對語料庫進行斷詞，此時未知詞將被切成數個單字詞或較短的詞，因此兩相鄰詞是否經常在語料中重複出現就成爲偵測新詞的重要依據。我們使用文獻中常用的點式交互資訊(Pointwise Mutual Information, PMI)[15]來衡量兩詞的內聚力，並以閥值(threshold)篩選有較大 PMI 值的相鄰詞爲候選新詞。由於 PMI 只考慮兩個詞是否經常相鄰出現，但經常相鄰出現的字合併後未必是新詞，有鑑於此，我們除了使用 PMI 之外，亦將使用前後文脈之 entropy 來過濾無關的新詞以提升精確度[16]，流程如圖一所示。

(一) CKIP 斷詞

我們收集的多語夾雜語料庫經過國台客英字典及 CKIP 斷詞，此時未知詞將以單字詞或短語的形式出現，而英文單字將被 CKIP 標記爲 FW。



圖一、未知詞擷取流程圖

(二) 以交互資訊爲本的聚合機制(Mutual-information-based word aggregation)

針對語料庫中斷完詞的句子，從左至右計算任兩相鄰詞的 PMI 分數，定義如下。

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} = \log \frac{C(w_i, w_j) \bullet N}{C(w_i)C(w_j)}, \quad (1)$$

其中 $C(w_i, w_j)$ 表示 w_i 與 w_j 在語料庫中相鄰出現的次數，而 N 代表語料中的字數且爲常數，在此我們取 Google 查詢所回傳的文件數做爲 $C(w_i, w_j)$ 、$C(w_i)$ 及 $C(w_j)$，由於無法知道正確的 N 值，我們以 10^{12} 代替之 (N 值大小並不影響 PMI 的排序結果)。當所有相鄰詞的內聚力計算完後以 PMI 值遞減排序，接著即可篩選有較大 PMI 值的相鄰詞爲新詞候選者。

(三) 以 Entropy 爲本的過濾機制(Entropy-based filtering)

前一個階段產生的候選新詞可能包含無關的新詞，其主要原因在於 PMI 只考慮詞與詞的內聚力，並未考慮該詞是否爲一個語意完整的單元。一般來說，統計式方法較難直接評估一個詞語意的完整性，不過一個語意完整的詞在語用的表現上卻有其特徵，也就是它可與許多其它的詞搭配使用形成更大的單元，因此，若某個詞與之相鄰的詞較少且集中在少數幾個時，即表示它與這些詞的相依性高，較有機會合併成爲新詞。根據上述特性，一個詞的語意完整性便可用與之相鄰詞的數量及分散程度間接衡量，原則如下所示。

相鄰詞數量多且分佈平均 → *語意完整性高* → *單獨使用*
相鄰詞數量少且分佈集中 → *語意較不完整* → *適合與其相鄰詞合併成爲新詞*

表四、“趴趴”之左右文脈情形(取前五個最常出現的詞)

趴趴			
詞頻	左文脈	右文脈	詞頻
8	愛	走	336
5	軟	GO	42
5	台灣	造	23
2	【	熊	21
1	「	照	19

舉例來說，”趴趴走”使用 CKIP 斷詞的結果爲“趴趴” “走”，因此，我們以“趴趴”爲例查詢 Google，並從回傳結果中擷取 999 筆含有“趴趴”的標題，分析緊鄰其左右的詞彙分佈情形，如表四所示。由結果可以發現，“趴趴”的左邊及右邊分別出現 212 及 58 個不同的詞，並且其右邊文脈分佈相當集中，“走”就佔了 336 次，而左文脈的分佈則較爲平均，這顯示“趴趴”與“走”的相依性高，較有可能向右合併形成爲新詞。

爲了以系統化的方法分析每個詞左右文脈分佈的集中程度，我們可用左右文脈的詞頻除以總詞頻的方式將左右文脈轉換爲機率表示法，再以 entropy 表示左右文脈分佈的集中程度。假設 $RC(w_t)=\{w_1,...,w_n\}$ 表示某個詞 w_t 的右文脈(right context)，也就是語料庫中緊鄰出現在 w_t 右邊的字詞集合，則 w_t 右文脈的 entropy 可定義爲

$$H_{RC}(w_t) = -\sum_{w_i \in RC(w_t)} P(w_i)\log_2 P(w_i), \tag{2}$$

其中 $H_{RC}(w_t)$ 爲 w_t 右文脈的 entropy，而 $P(w_i) = C(w_i)/N$ 爲 w_t 右文脈中某個詞出現的機率，其中，$C(w_i)$ 爲語料庫中 w_i 緊鄰出現在 w_t 右邊的次數，N 爲語料庫中所有緊鄰出現在 w_t 右邊的總次數。同理，$H_{LC}(w_t)$ 表示 w_t 左文脈的 entropy，計算方法同上。根據上述 entropy 的計算方式，文脈分佈愈集中，則 entropy 愈小表，而文脈分佈愈平均，則 entropy 愈大。因此，以 entropy 來衡量一個詞的語意完整性，其原則如下。

Entropy 大 → 語意完整性高 → 單獨使用

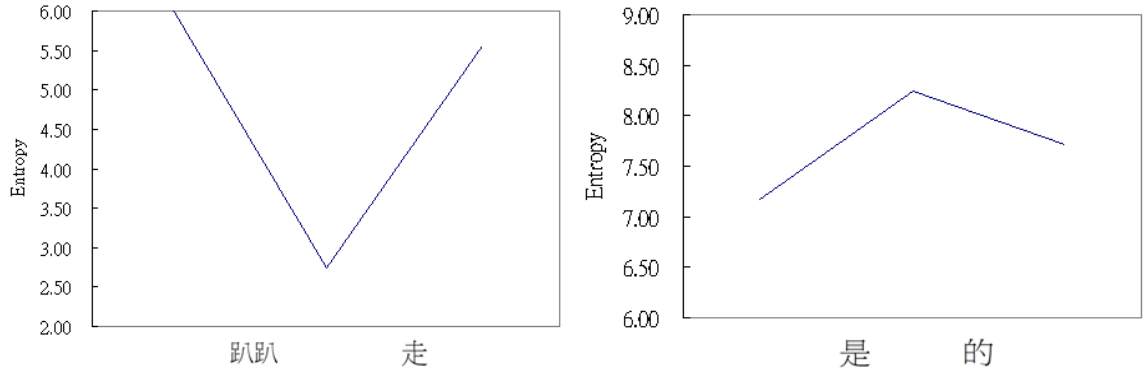
Entropy 小 → 語意較不完整 → 適合與相鄰詞合併成爲新詞

表五即爲趴趴走等詞左右文脈的 entropy。由表五可發現”趴趴”右文脈的 entropy 較低，表示其語意較不完整，適合與其右邊的詞”走”合併成爲一個新詞，至於”是”與”的”這兩個詞左右文脈的 entropy 都很大，表示這些單字詞已經是完整的語意單元，不需要與其它詞合併。由此觀察可發現當兩個詞要合併時，如果其中一個詞的左邊或右邊不完整時，便是一個良好的合併候選者，因此，我們定義兩個詞 w_i 與 w_j 合併前的 entropy 爲 w_i 右文脈與 w_j 左文脈 entropy 的最小值，如下所示。

$$H_{before}(w_i, w_j) = min(H_{RC}(w_i), H_{LC}(w_j)), \tag{3}$$

表五、合併前後左右文脈 entropy — 以趴趴走、是的爲例

	$H_{LC}(w_t)$	$H_{RC}(w_t)$		$H_{LC}(w_t)$	$H_{RC}(w_t)$
趴趴	6.07	**2.75**	是	7.17	8.24
走	4.85	5.56	的	8.34	7.72
趴趴走	6.07	5.56	是的	7.17	7.72



(1) (2)

圖二、兩詞合併前後 entropy 變化情形

表五的另一個發現爲兩詞合併後其左右文脈的 entropy 會變大，這是因爲合併後的新詞語意較爲完整之故，例如：趴趴走合併前的 entropy 爲 2.75，合併後若取"趴趴"的左文脈爲合併後的左文脈，"走"的右文脈爲合併後的右文脈，則合併後左右文脈的 entropy 分別變大爲 6.07 與 5.56，如圖二(1)所示，反之，兩個語意完整不需合併的詞如果合併，則合併後的 entropy 並不會明顯變大，甚至可能變小，如表五及圖二(2)中"是"與"的"的範例所示。由此可知，兩個詞合併前後 entropy 的變化便是判斷是否爲新詞的重要依據，因此，我們定義兩詞合併前後的 entropy 比值(ratio)，如下所示。

$$\lambda^{LC}_{w_i w_j} = \frac{H^{LC}_{after}(w_i, w_j)}{H_{before}(w_i, w_j)}, \tag{4}$$

$$\lambda^{RC}_{w_i w_j} \frac{H^{RC}_{after}(w_i, w_j)}{H_{before}(w_i, w_j)}, \tag{5}$$

其中 $\lambda^{LC}_{w_i w_j}$ 與 $\lambda^{RC}_{w_i w_j}$ 分別爲 w_i 與 w_j 兩個詞合併後左右文脈 entropy 與合併前的比值，其中 $H^{LC}_{after}(w_i, w_j) = H_{LC}(w_i)$ ，也就是取 w_i 的左文脈做爲 w_i 與 w_j 合併後的左文脈，同理，$H^{RC}_{after}(w_i, w_j) = H_{RC}(w_j)$ ，因此，當 $\lambda^{LC}_{w_i w_j}$ 與 $\lambda^{RC}_{w_i w_j}$ 皆大於 1 時代表合併後 entropy 較合併前大，可考慮將兩詞合併成爲新詞。

四、實驗結果

(一) 實驗設計

在實驗設計上，我們從雅虎新聞中隨機挑選 500 篇新聞，接著以人工的方式找出國台夾雜的句子，再以 CKIP 斷詞系統進行斷詞，如果某句所夾雜的台語短語無法被正確的斷詞，則表示該短語爲台語新詞，反之，可以被正確斷出的台語短語將不列入本實驗中。依此原則，本實驗共挑選出 40 句包含台語新詞的國台夾雜測試句，句中任兩相鄰詞皆爲合併候選者，共計 200 個，其中僅有 41 個台語新詞，即爲本實驗的標準答案。測試時，給定一斷完詞的測試句，首先透過 Google 的回傳結果從左至右計算句中任兩相鄰詞的 PMI 分數及左右文脈 entropy，接著針對有較高 PMI 的相鄰詞以合併前後的 entropy 比值，即 $\lambda^{LC}_{w_iw_j}$ 與 $\lambda^{RC}_{w_iw_j}$ 做爲閥值來評估是否合併成爲新詞，最後以召回率(recall)、精確率(precision)與 F-measure 來評估未知詞擷取演算法的效果，其中召回率是用來評估系統能從正確的 41 個新詞中找出幾個，精確率則是評估系統所建議的新詞中有多少是正確的，而 F-measure 則是召回率與精確率的綜合評估，即 2 * recall * precision / (recall+precision)。一般說來，調高 PMI、$\lambda^{LC}_{w_iw_j}$ 與 $\lambda^{RC}_{w_iw_j}$ 等閥值將可提系統的精確率，但召回率卻可能下降，反之，若閥值太低則可以找出更多新詞，但也會降低系統的精確度。

(二) 結果

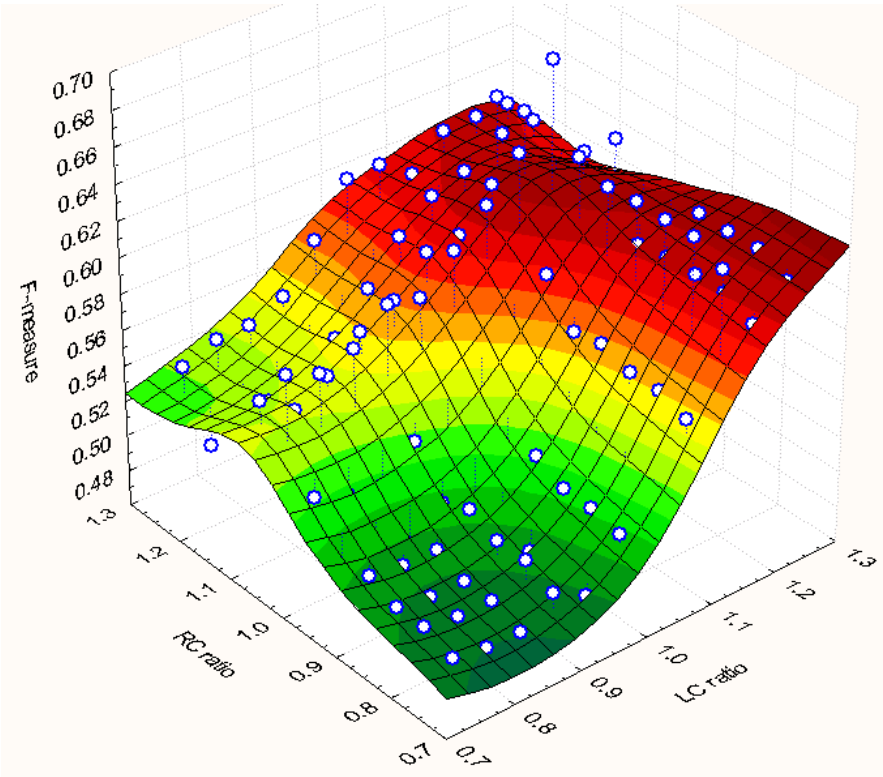
表六即爲部分候選新詞之 PMI 及合併前後左右文脈 entropy 之比值。結果顯示調高 PMI 等閥值可過濾許多不爲新詞的候選詞，但也可能損失如"破糊糊"(PMI 太小)與"皮皮挫"($\lambda^{LC}_{w_iw_j}$ 太小)等正確的新詞。本實驗藉由調整閥值(PMI=3，$\lambda^{LC}_{w_iw_j}$ =1.15，$\lambda^{RC}_{w_iw_j}$ =1.05)得到的最佳結果爲 F-measure=68.49%，Recall=60.98%，Precision =78.13%，如表七所列之部分結果，圖三即爲同時調整 $\lambda^{LC}_{w_iw_j}$ 與 $\lambda^{RC}_{w_iw_j}$ (0.75 ~ 1.25)之實驗結果。

表六、候選新詞之 PMI 及合併前後左右文脈 entropy 之比值

候選詞	PMI	$\lambda^{LC}_{w_iw_j}$	$\lambda^{RC}_{w_iw_j}$
瞭解 甚麼	12.69	1.11	1.00
碎碎 念	12.33	4.14	5.29
皮皮 挫	11.42	0.92	1.07
囝 囡	11.07	1.29	1.38
才 發生	3.90	0.79	0.96
好 山	0.80	1.13	1.18
破 糊糊	-2.14	1.33	1.14
就 是	-2.87	1.13	1.27

表七、不同閥值未知詞擷取效果之影響 (PMI=3)

$\lambda_{w_i w_j}^{LC}$	$\lambda_{w_i w_j}^{RC}$	Recall	Precision	F-measure
1.15	0.75	0.6585	0.6000	0.6279
1.15	0.80	0.6585	0.6136	0.6353
1.15	0.85	0.6585	0.6136	0.6353
1.15	0.90	0.6585	0.6136	0.6353
1.15	0.95	0.6341	0.6341	0.6341
1.15	1.00	0.6098	0.6757	0.6410
1.15	**1.05**	**0.6098**	**0.7813**	**0.6849**
1.15	1.10	0.5610	0.7667	0.6479
1.15	1.15	0.5366	0.8148	0.6471
1.15	1.20	0.4390	0.9000	0.5902
1.15	1.25	0.4390	1.0000	0.6102



圖三、兩詞合併前後 entropy 變化情形

五、結論

本論文提出結合交互資訊與熵之未知詞擷取演算法，自動從多語夾雜語料中找出未知詞，使用交互資訊之目的在計算詞與詞的內聚力並挑選較高者爲候選新詞，接著使用以熵爲基礎的過濾機制，根據候選新詞左右文脈的分佈情形過濾無關的新詞，實驗結果顯示本論文所提之以熵爲基礎的方法可藉由過濾無關的新詞提升未知詞擷取之精確度。未來，我們將研究機器學習的方法以期達到更精確的結果。

誌謝

本研究感謝王政欽及粘書豪同學在先期工作上所做的努力。此外，本研究承蒙國科會 NSC 99-2221-E-155 -036 -MY3 費補助特此致謝。

參考文獻

[1] P. Fung and T. Schultz, “Multilingual Spoken Language Processing,” *IEEE Signal Processing Magazine*, 25(3), pp. 89-97, 2008.

[2] L. Ge, “An investigation on English/Chinese Code-switching in BBS in Chinese Alumni's Community,” Master thesis, University of Edinburgh, 2007.

[3] Y. Liu, “Evaluation of the Matrix Language Hypothesis: Evidence from Chinese-English Code-switching Phenomena in Blogs,” *Journal of Chinese Language and Computing*, 18(2), pp. 75-92, 2008.

[4] R. Y. Lyu, M. S. Liang, and Y. C. Chiang, “Toward Constructing a Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin,” *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), pp. 1-12, 2004.

[5] C. H. Wu, Y. H. Chiu, C. J. Shia, and C. Y. Lin, “Automatic Segmentation and Identification of Mixed-language Speech using Delta-BIC and LSA-based GMMs,” *IEEE Trans. Audio, Speech, and Language Processing*, 14(1), pp.266-276, 2006.

[6] C. L. Huang and C. H. Wu, “Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis,” *IEEE Trans. on Computers*, 56(9), pp. 1225-1233, 2007.

[7] D. C. Lyu, R. Y. Lyu, Y. C. Chiang, and C. N. Hsu, “Speech Recognition on Code-switching among the Chinese Dialects,” in *Proc. of ICASSP-06*, pp. 1105-1108, 2006.

[8] W. T. Hong, H. C. Chen, I. B. Liao, and W. J. Wang, “Mandarin/English Mixed-Lingual Speech Recognition System on Resource-Constrained Platforms,” in *Proc. of ROCLING-09*, pp. 237-250, 2009.

[9] J. Y. C. Chan, P. C. Ching, T.Lee and H. M. Meng, “Detection of Language Boundary in Code-switching utterances by Bi-phone Probabilities,” in *Proc. of ISCSLP-04*, pp. 293-296, 2004.

[10] J. Y. C. Chan, P. C. Ching, T.Lee and H. Cao, “Automatic Speech Recognition of Cantonese-English Code-mixing Utterance,” in *Proc. of Interspeech*, pp. 113-116, 2006.

[11] C. M. White, S. Khudanpur, and J. K. Baker, “An Investigation of Acoustic Models for Multilingual Code-Switching,” in *Proc. of Interspeech*, 2008.

[12] Y. Zhang and J. Tao, “Prosody Modification on Mixed-Language Speech Synthesis,” in *Proc. of ISCSLP-08*, pp. 1-4, 2008.

[13] Y. Qian, H. Liang, and F. Soong, “A Cross-Language State Sharing and Mapping Approach to Bilingual (Mandarin–English) TTS,” *IEEE Trans. on Audio, Speech, and Language Processing*, 17(6), pp. 1231-1239, 2009.

[14] W. Y. Ma and K. J. Chen, “"Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff,” in *Proc. of ACL, Second SIGHAN Workshop on Chinese Language Processing*, pp. 168-171, 2003.

[15] K. Church and P. Hanks. “Word Association Norms, Mutual Information and Lexicography,” *Computational Linguistics*, vol. 16, no. 1, pp. 22-29, 1991.

[16] Z. Luo, and R. Song, “An Integrated Method for Chinese Unknown Word Extraction,” in *Proc. of the Third SIGHAN Workshop on Chinese Language Learning*, pp. 148-155, 2004.