中文詞彙語意資料的整合與擷取：詞彙語意學的觀點

高照明
zmgao@ntu.edu.tw
台灣大學外國語文學系

摘要

本文從詞彙語意學理論的觀點整合知網(Hownet) 、現代漢語分類辭典、教育部國語辭典等資源，並利用 Wordnet 和漢英辭典，擷取上述不同來源的中文詞彙語意訊息。我們透過整合後的訊息發展一套系統，使用者輸入兩個詞可以找出兩個詞之間的詞彙語意關係包括（一）同義關係（二）反義關係（三）上下位關係（四）部件與整體關係 （五）相同事件（六） 相同領域(domain) （七）相同語意特徵（八）相同的語意類別（九）事件與語意角色。

關鍵詞：詞彙語意關係、詞彙知識庫、知網(Hownet)、義元、語意特徵、語意角色、事件角色轉換、Wordnet、現代漢語分類辭典、重編國語辭典修訂本、同義詞、反義詞、上位詞、下位詞、全體詞、部分詞

一 前言

詞彙語意學的發展與資訊科學及人工智慧有相當密切的關係。六零年代語言學家 Fillmore (Fillmore 1968)提出語意角色的理論架構格理論(case theory)對於語意學及句法學產生深遠的影響，同一時期 Wilks (Wilks 1968)從人工智能的角度研究語意知識的表達。七零年代 Shank (Shank 1975)提出腳本理論將詞彙知識與常識具體化程序化，作爲自然語言理解的基礎。而 Sowa 等人(Sowa 1984)則從事 conceptual graph 的研究。七零年代末期 John Sinclair (參考 Sinclair 1987) 首創以語料庫及計算機研究詞義和搭配語並編纂辭典（Collins Cobuild English Dictionary）。八零年代，利用機讀辭典研究語法與詞彙語意開始興起，其中最多研究人員使用的資源是 Longman Dictionary of Contemporary English(LDOCE) (參考 Boguraev and Briscoe (eds) 1989)。九零年代隨著英國國家語料庫(http://www.natcorp.ox.ac.uk/)及相關檢索軟體(SARA, Xaira) 的完成，研究人員開始有龐大的語料庫及檢索工具研究詞彙語意。而 Wordnet 計畫(http://wordnet.princeton.edu/)推出(semantic concordancer)以 Wordnet 詞項的意義標示語料庫中的詞的詞義，爲計算詞彙語意學奠定了深厚的基礎。近年來越來越多標注詞彙語意訊息的語料庫出現，如標記論元結構（argument structure）及語意角色訊息的 FrameNet、 VerbNet 、PopNet。計算詞彙語意學研究的重心轉爲利用語料庫及統計演算法，例如 Church 首創以互見訊息(mutual information)和 t-score 來擷取搭配語(參考 Church and Hanks 1990) Church et al. (1991) Church et al. (1994)。Hearst (1992)透過句型擷取上下位詞。Grefefenstette (1994)以語法剖析器和統計擷取同義詞。Jones (2002)透過語料庫擷取反義詞。Turney (2006), Girju 等 (2007) 更進一步以統計及機器學習演算法研究詞彙語意關係，這與傳統透過詞彙知識庫擷取與判定詞彙語意關係的方法大異其趣。以大量語料結合統計或機器學習演算法的優點是不需要詞彙知識庫即可從語料中擷取一些語意關係，缺點是擷取的資料不夠精確與完整必須透過專家來校對與補充。本文的目的在於整合現有的各種中文詞彙知識庫，並利用這些資料庫截長補短

來擷取最多的語意關係，作爲未來評估機器學習演算法擷取詞彙語意關係研究的平台。

詞彙語意關係與語意網(semantic web)及本體論(ontology)息息相關。Tim Berners-Lee (2000)提出語意網的概念描繪了下一代網際網路的遠景。語意網的成功仰賴本體論，也就是必須能清楚的表達通用或某一特定領域知識的詞彙的意義並顯示相關概念之間的語意關係。目前絕大部分的本體論都是透過人工建立及人工標示。如果能夠半自動建立及標記對於語意網的發展將有很大的幫助。建立大規模詞彙語意關係的資料庫可以大幅提升本體論發展的時間及品質。

詞彙語意關係的研究也能夠廣泛應用於資訊檢索與擷取，機器翻譯系統等，例如，資訊擷取時需要辨識文章裡面人、事、時、地、物。檢索時需要利用同義詞，而自然語言理解更必須判斷語意角色，

本研究利用現有的中英文詞彙知識庫發展一個能夠截長補短自動找出中文詞彙語意關係的系統。英文部分我們主要利用 Wordnet 以及漢英辭典，中文方面我們使用知網(Hownet)，現代漢語分類辭典，教育部重編國語辭典修訂本等資源。我們的系統能夠判斷下列的詞彙語意關係（一）同義關係（二）反義關係（三）上下位關係（四）部件與整體關係 （五）相同事件（六） 相同領域(domain) （七）相同語意特徵（八）相同的語意類別（九）事件與語意角色。

二 語意關係的理論

傳統上詞彙之間的關係可以分成組合的關係(syntagmatic relation)及聚合的關係(paradigmatic relation)。組合的關係就是構成詞組的關係(水平的關係)，依據語法依存關係可以分成：修飾與被修飾的關係、主詞與謂語的關係、動詞與受詞的關係。在 Chomsky 的句法理論中，主詞與謂語的關係、動詞與受詞的關係是透過動詞來選擇它的主詞或受詞的語義 (請參考 Pustejovsky 2000 對組合關係的理論)。聚合的關係可以看成在某一種句式裡面代換的關係（垂直的關係），通常是同義詞、反義詞的關係、上下位詞的關係。

魯川(2001)將語意組合關係分成語意關係和語意依附兩種。語意關係分爲並列關係、選擇關係、等同關係、同現關係、加合關係、配合關係、接合關係。而語意的依附關係分成爲事件的依附（包括語氣、話語、與情態），述謂的依附(時態、動貌、語態、程度)，指稱意結的依附（數）。

Koenig (1999)將詞彙的關係概分成兩種，一種是分類的關係(classificatory)，另一類是構詞的關係(morphological)，例如相同詞性的關係就是分類的關係，而 bird 和 birds 之間屬於構詞的關係。Koenig 利用中心語驅動詞組語法(Head-driven Phrase Structure Grammar, HPSG)的架構提出 Type Underspecified Hierarchical Lexicon)解釋詞的癖性(idiosyncrasy)與滋生性(productivity)。

Chaffin and Herrmann (1988)認爲語意關係有兩種理論，網路理論（network theory）及關係元素理論(relation element theory)，處理時前者計算連結兩個概念節點的路徑，而後者計算所表示的元素的異同。他們歸納出五大類共三十一種的語義關係，這五大類分別是對比(contrasts) 、相似(similars) 、類別包含(class inclusion) 、格關係(case relation)（亦稱爲語意角色關係）、部分全體關係(part whole)。

Calzolari (1988)建議應該合併辭典與同義詞辭典成爲一個大的詞彙知識庫，這裡面應該包含下列幾種語意關係：上下階層關係、同義關係、IS-A 關係、論元關係、詞彙場(lexical fields)，搭配語關係、術語庫、衍生詞的關係。

當代語言學理論中對詞彙語意關係發展最詳盡者是 Melcuk、Zholkvsky、Apresyan 等人所提出的 Explanatory Combinatorial Dictionary (ECD) (Melcuk 1988)其中提出數十個詞彙函數

(lexical functions)。以 Magn 這個詞彙函數爲例，它的功能是[intensifier]。輸入一個詞就可以得到這個函數對應的值，如 Magn(to condemn) = strongly。

Byrd (1994)描述 I.B.M.發展的詞彙知識庫 ComLex 及相關工具，可以根據一個詞的詞義找出各種關係，包括拼字相同、上下位詞、同義詞、典型的論元(argument) 、例如：領養 adopt 典型的受詞是別的父母親的孩子(a child of other parents)，及語意選擇限制(selectional restrictions)，例如領養的主詞及受詞都是人(person)。這些關係都由現有的辭典裡面（如 Merriam Webster Dictionary, Longman Dictionary of Contemporary English (LDOCE)的訊息抽取出來。Klavans (1994)也以及 Merriam Webster Dictionary 裡面的定義及同義詞辭典抽取出來相關詞的語意網(semantic net)。

近年有關於詞彙語義表徵最重要的理論架構堪稱 Jackendoff 與 Pustejovsky 的著作。Jackendoff (1983, 1990)提出抽象的概念結構(conceptual structure)用以解釋許多的語言現象。

Pustejovksy (1995)提出衍生詞彙理論(Generative Lexicon)，理論有四層表徵，論元結構、(argument structure) 、事件結構(event structure) 、 屬性結構(qualia structure)、及詞彙繼承結構(lexical inheritance structure)。論元結構註明事件參與者的語義角色.。事件結構(event structure)將事件分成狀態(state)過程(process)過渡(transition)及更小結構。屬性結構(qualia structure)將名詞屬性分成 formal(語意類別) 、 constitutive(與組成這個東西的關係) 、 telic (東西的功用) 、agentive(東西如何產生) 、詞彙繼承結構(lexical inheritance structure)。透過 type coercion、 selective binding、 co-composition 等理論的機制衍生詞彙理論可以有系統解釋許多詞彙語意的現象，例如：enjoy a book 和 enjoy a meal 前者是 enjoy reading a book，後者是 enjoy eating a meal. Jackendoff 和 Pustejovsky 的理論，由於相當抽象，並不容易實做。

中文有關詞彙語意關係的專著除了魯川(2001)的「意合網路」之外，以董政東與董強(Dong and Dong 2006)所發展的知網(Hownet) 最重要。我們將在後面詳細介紹 Hownet。

從上述各種詞彙語意學理論的介紹不難發現語意的類別、語意特徵、同義詞、反義詞、上下位詞、部分與全體關係、語意角色、事件關係，這些是絕大部分詞彙語意學理論和詞彙知識庫都包括的訊息。

三國內外著名詞彙知識庫的介紹

最著名的詞彙知識庫首推普林斯頓大學所發展出 Wordnet (http://wordnet.princeton.edu/)，有大量的自然語言處理研究依靠這個知識庫。透過 Wordnet 可以查出詞的定義，例句，同義詞，上位關係詞，下位關係詞，部分關係詞，全體關係詞等。Euro Wordnet 是歐盟幾個國家以 Wordnet 爲基礎合作發展多語詞彙知識庫。

美國南加大整合許多的知識庫和發展了 Ontosaurus (http://www.isi.edu/isd/ontosaurus.html) 如同 Wordnet 它註明了詞項的定義，同義詞，及語意的類別。

Sumo Search Tool (http://sigma.ontologyportal.org:4010/sigma/WordNet.jsp)這個工具將 Wordnet 的 sense 直接對應到 IEEE Suggested Merged Upper Ontology (SUMO)。優點是 Wordnet 的詞義類別更清楚的表示出來，例如 buffalo 這個詞對應到 SUMO 有 buffalo 水牛，city 城，meat 肉。透過 SUMO ontology 我們可以得到許多詞彙之間的關係，例如 SUMO 裡面包括 subclass 和 disjoint 這類的關係，在 SUMO 裡面有諸如 Buffalo is a subclass of hoofed mammal Buffalo is disjoint from DomesticAnimal 這類對於自然語言理解和推論非常重要的知識。DOLCE 與 SUMO 類似，也是一個非常重要的 ontology，並提供與 Wordnet 的對應。

卡內基美侖大學利用 Wordnet 及其它包括百科全書在內的許多電子資源，發展 Lexical Freenet (http://www.cinfn.com/doc/)。這是到目前爲止我們所知最最完整的詞彙關係資料庫。能夠找出的詞彙關係遠遠超過 Wordnet。例如，我們輸入 Taipei, Taiwan 兩個詞得到一些相當

有趣的結果，包括 Taipei 是 Taiwan 的一部份，而反義詞是中國的 China's。

柏克萊加州大學所發展的 FramNet (http://framenet.icsi.berkeley.edu/)所發展的檢索介面，依據不同的語意框架 frame 詳細探討每一個語意框架常用的詞彙(lexical element)語意角色(semantic role)和對應的語法功能(grammatical function)。

具有語意訊息的中文資料庫有中研院語言所黃居仁教授的中文詞彙網路，這個系統結合了 Wordnet，SUMO ontology，以及中研院詞彙知識庫裡面的詞性標記。

大陸商務出版社所出版的「同義詞詞林」編排的方式是按照語意階層由大類到小類分類。類似同義詞詞林但是分類更細的是大陸出版的現代漢語分類辭典。共有三層的語意分類，最上層的總目有十二類分別是 A 人 B 物 C 時間空間 D 抽象事物 E 特徵F 動作 G 心理活動 H 活動 I 現象與狀態 J 關聯 K 助詞 L 敬語。總類是最大語意類，下面有次類，次類下有次分類。

教育部重編國語辭典修訂本(http://140.111.34.46/dict/)，除了解釋，並有例句，相似詞，相反詞。

大陸董振東先生獨力發展出來的知網 Hownet (http://www.keenage.com)也是一個非常重要的詞彙知識庫(參考 Dong and Dong (2006)。知網 Hownet 包含的訊息相當的多，是一個雙語的知識庫，可以表達概念與概念之間的常識關係。細節將在下一節介紹。

四 Hownet 語意訊息的表達與擷取

我們逐一介紹如何利用知網(Hownet)，,現代漢語分類辭典，重編國語辭典修訂本等資源，並利用 Wordnet 和漢英辭典這些知識庫來找出語意關係，及每一個知識庫的限制。最後我們將這幾個詞彙資源整合成一個綜合性詞彙語意關係知識庫來找出具有語意關係的詞。由於 Hownet 的結構最複雜，語意的關係也最詳細，我們先探討如何應用這個具備雙語詞彙語意表達模式。Hownet2002 比 Hownet 2000 的語意表達方式更豐富，特別是語意角色的部分更清楚，對自然語言理解更有幫助，但是因爲結構相當複雜處理不易，所以我們針對新舊版的不同語意表達模式設計不同的應用程式。我們分別建立 Hownet 2000 及 Hownet2002 的資料庫。這兩版的基本差異可以從下列的表達模式看出來。

{human|人, #occupation|職稱,*curel醫治, medicall醫}
{human|人:HostOf={Occupation|職位},domain={medicall醫},{doctorl醫治:agent={~}}}

第一個是舊版的 Hownet 的表示法，其中的詞由義元表達，義元本身沒有內部結構，第二個是新版的表示法，義元有內部結構。舊版的 Hownet 語意角色不清楚，例如＊符號可能表示在某個事件擔任 agent, experiencer，或其它的語意角色，但在新版的 Hownet 裡面哪個義元在哪個事件中擔任什麼語意角色非常明確。換言之，新版詞與詞之間的語意關係相當明確。舊版 Hownet 的義元沒有內部結構缺點是對語意角色的判定較不容易，但相對的程式處理起來較簡單，檢索義元的速度也較快速。我們利用這個特色設計一個工具程式。輸入一個中文或英文詞，程式會將這個詞的所有義元列出來，使用者可以選擇其中一個或多個義元，並且可以用 AND 或 OR 來查詢包含這些義元的詞。

例如；輸入車這個中文詞，程式會顯示這個詞 Hownet 義元的表示法，有交通工具、切割、人的姓等幾個不同的意義。使用者選擇其中一個意義，再選擇要檢索一個或多個義元並選擇這些義元之間的關係是 AND 或 OR，程式就會列出包含這些義元的相關中文詞。由於 Hownet 是中英雙語，我們的程式也可以輸入英文用相同的方法找出相關的英文詞及其中文翻譯。這個工具對於全自動或半自動建立中英雙語或中英跨語言的本體論(ontology)系統非常有用，例如輸入車選擇 LandVechiclel車這個義元可以得到不同的車如板車，叉車，餐車，彩車，巴士，長途汽車等。

詞彙與詞彙之間的關係透過 Hownet 的意元表達模式，原本不清楚的詞彙語意關係變得比較清楚。例如：餐車的義元是 {LandVehicle|車，@eat|吃}透過 Hownet 的知識表達法，餐車與吃的語意關係得以連結。但是@究竟代表什麼意義或哪一個語意角色在 Hownet2000 裡面並沒有明確的交代。而在 Hownet 2002 裡面就比較清楚，{LandVehicle|車:{eat|吃:location={~}}}顯示餐車是吃的地點。

醫生在 Hownet2002 裡面有三個英文翻譯 doctor, surgeon, doctor，它們的義元表示都是{human|人:HostOf={Occupation|職位},domain={medical|醫},{doctor|醫治:agent={~}}}。由於義元內部具有結構，因此我們先撰寫一個剖析器將內部結構剖析再來比較義元。義元是一種表達語言知識的 meta language,醫生的義元表示醫生是一個人，具有職位，是醫學領域，且是醫治事件裡面扮演主事者的語意角色。 而醫療這個詞只有一個義元{doctor|醫治}。我們的程式（http://140.112.185.57/~denehs/compare.html）可以把具有相同的義元找出來，或是把相同事件裡面不同的語意角色找出來。例如輸入醫生和醫療會得到下面訊息。

醫生(doctor)和醫療(doctor)同屬於醫治(doctor)事件
醫生(doctor)是 agent
醫療(doctor)是事件

護士(nurse) 的義元是{human|人:HostOf={Occupation|職位},domain={medical|醫},{TakeCare|照料:agent={~}},{doctor|醫治:agent={~}}}，所以輸入醫生和護士會得到。

醫生(doctor)和護士(nurse)同屬於醫治(doctor)事件
醫生(doctor)是 agent
護士(nurse)是 agent

醫生(physician)	護士(nurse)	共同點：人:domain={medical|醫}}	{human|	共同語意特徵：{medical|醫}

病人(patient) 的義元是{human|人:domain={medical|醫},{SufferFrom|罹患:experiencer={~}},{doctor|醫治:patient={~}}}，所以輸入醫生和病人會得到。

醫生(doctor)和病人(patient)同屬於醫治(doctor)事件
醫生(doctor)是 agent
病人(patient)是 patient

醫生(physician)	病人(patient)	共同點：人:domain={medical|醫}}	{human|	共同語意特徵：{medical|醫}

醫院(hospital)的義元是 {InstitutePlace|場所:domain={medical|醫},{doctor|醫治:content={disease|疾病},location={~}}}，所以輸入醫生和醫院會得到
醫生(doctor)和醫院(hospital)同屬於醫治(doctor)事件
醫生(doctor)是 agent
醫院(hospital)是 location

醫生(doctor)	醫院(hospital)	共同點: 不屬於相同的語意類別	共同語意特徵：{medical|醫}

五 語意資料的整合、擷取、與評估

在本節中我們敘述如何整合知網(Hownet),現代漢語分類辭典,教育部國語辭典等資源,並利用 Wordnet 和漢英辭典,擷取上述不同來源的中文詞彙語意訊息。由於大規模質與量的評估非常困難,我們隨機選擇十四組詞來測試這些詞彙知識庫。

處方在 Hownet 裡面有兩個意義,一個是動詞,一個是名詞,所以有兩個義元。

處方(prescription) {documentl文書:domain={medicall醫},{orderl命令:ResultEvent={preparel準備:content={medicinel藥物}},instrument={~}}}

處方(prescribe) {writel寫:ContentProduct={documentl文書:{orderl命令:ResultEvent={preparel準備:content={medicinel藥物}},instrument={~}}},domain={medicall醫}}

這裡可以看出 Hownet 的一些問題。理論上 Hownet 應該將開處方這個詞與醫療這個事件相連結,並表示開處方的 agent 爲醫生,但是 Hownet 並沒有這樣的訊息,所以醫生和處方的相同特徵只有{medicall醫}。

我們再看其它的例子,藥在 Hownet 裡面有三個意義,所以有三個不同的義元,

藥(certain chemicals) {chemicall化學物}
藥(kill with poison) {killl殺害:instrument={physicall物質:modifier={poisonousl有毒}}}
藥(medicine) {medicinel藥物}

輸入醫生和藥得到沒有共同的語意類別或特徵。原因是 Hownet 裡面並沒有連接藥{medicinel藥物}與醫療這個事件。理論上藥{medicinel藥物}應該列爲與醫療這個事件的工具 instrument。

另外 Hownet 的義元表示法常有不一致的情形,例如悲哀和痛苦,Hownet 的表示法如下:

悲哀(sorrowful) {sorrowfull悲哀}
痛苦(pain) {experiencel感受:CoEvent={unfortunatel不幸}}
痛苦(agony) {unfortunatel不幸}

它們是跟感情情緒有關的近義詞,但在 Hownet 裡面卻沒有任何的關係。此外,Hownet 沒有明顯的同義和近義關係,必須完全靠義元比對,看似一個簡單的工作,實際上卻相當複雜,原因是 Hownet 裡面有相當多的義元,用來定義所有詞彙的每一個義元的重要性並不相等,而有時近義的詞卻用完全不同的義元來表示。所以直接以義元比對有時並不能找到同義詞。事實上 Hownet 的問題是所有以 meta language 來表示語意的理論所必須面對的共同問題。不管是 meta language 或語意特徵,都很難用一套有限的元素來定義所有的詞。例如輸入老闆和老闆娘可以發現 Hownet 的義元沒有表示老闆和老闆娘之間有配偶的關係。

老闆(boss) {humanl人:{employl雇用:agent={~}}}
老闆娘(shopkeeper's wife) {humanl人:modifier={femalel女},{employl雇用:agent={~}}}

老闆(boss)和老闆娘(proprietress)同屬於雇用(employ)事件
老闆(boss)是 agent

老闆娘(proprietress)是 agent

老闆(boss)	老闆娘(proprietress)	共同點：{human|人}	共同語意特徵：{employ|雇用:agent={~}

如果我們再看 Hownet 裡面男人和女人的表示，就會發現中間存在許多不一致的情形。

男人(husband) {human|人:belong={family|家庭},modifier={male|男}{spouse|配偶}}
男人(man) {human|人:modifier={male|男}}
女人(wife) {human|人:belong={family|家庭},modifier={female|女}{spouse|配偶}}
女人(women) {human|人:modifier={female|女}}

男人(husband)	女人(wife)	共同點：{human|人:belong={family|家庭}}	共同語意特徵：{family|家庭} {spouse|配偶}
男人(man)	女人(women)	共同點: {human|人}	共同語意特徵: (無)

男人與女人有配偶的意義，有共同語意特徵: {family|家庭} {spouse|配偶}，但是老闆與老闆娘同樣有配偶的意義，Hownet 卻沒有相對的義元表示。

如前所述 Hownet 並沒有明確的同義關係，同義關係必須另外寫義元比對的程式。同樣的，Hownet 也沒有明確的反義關係。例如，買與賣的 Hownet 表示法如下，兩者不但沒有任何相同的義元也沒有明確的反義關係。

HowNet 義元
買(buy) {buy|買}
賣(betray) {betray|背叛}
賣(sell) {sell|賣}

事實上，Hownet 還有獨立的檔案描述事件角色轉換關係(Event Role Shift)，例如
欠與有這兩個義元的關係是 implication，也就是如果 X 欠 Y 一樣東西(target)，Y 就是這一樣東西(target)的 possessor。

owe|欠(X) [implication]←→own|有(Y);
  target OF owe|欠=possessor OF own|有;
  possession OF owe|欠=possession OF own|有.

取與得到這兩個義元的關係是 consequence，也就取的結果是得到。而取的 agent 就是得到的 possessor.
take|取←→obtain|得到 [consequence];
  agent OF take|取=possessor OF obtain|得到;
  possession OF take|取=possession OF obtain|得到.

同理，偷與取這兩義元的關係是 hypernym 的關係，我們從這這些例子可以看出來在事件角色轉換關係裡面，語意的關係如 [implication]， [consequence] ， [hypernym]在左邊或右邊表

示不同的關係。例如 owe|欠(X) [implication]←→own|有(Y) 表示欠 imply 有，steal|偷←→take|取 [hypernym]表示偷是取的下位詞。下面是 Hownet 事件轉換關係的一些例子。

steal|偷←→take|取 [hypernym];
agent OF steal|偷=agent OF take|取;
possession OF steal|偷=possession OF take|取;
source OF steal|偷=source OF take|取.

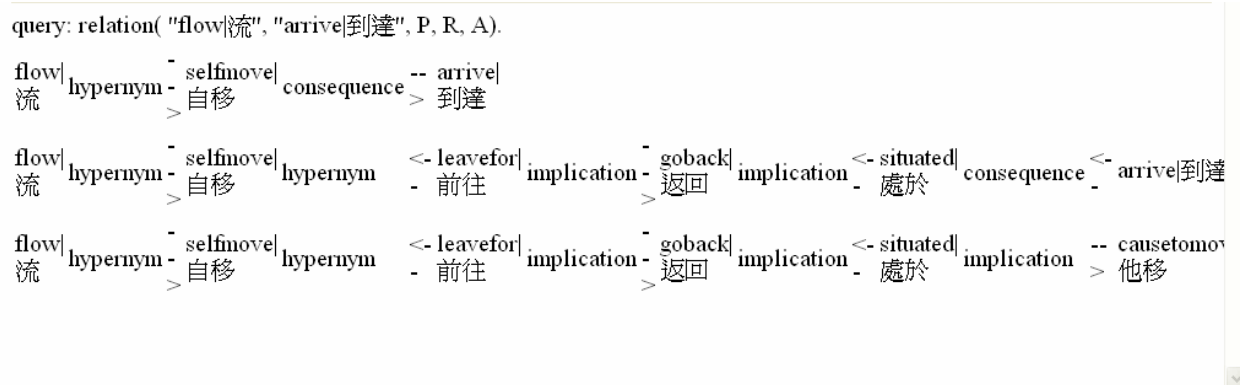
rob|搶←→take|取 [hypernym];
agent OF rob|搶=agent OF take|取;
possession OF rob|搶=possession OF take|取;
source OF rob|搶=source OF take|取.

earn|賺←→take|取 [hypernym];
agent OF earn|賺=agent OF take|取;
possession OF earn|賺=possession OF take|取;
source OF earn|賺=source OF take|取.

buy|買←→take|取 [hypernym];
beneficiary OF buy|買=agent OF take|取;
possession OF buy|買=possession OF take|取;
source OF buy|買=source OF take|取.

我們利用資訊科學常用的 acylic graph (http://en.wikipedia.org/wiki/Directed_acyclic_graph)來表示這些關係。我們利用 Prolog 程式能夠很方便的使用 predicate calculus 自動推論的優點結合 Perl 程式字串處理能力，透過 perl 的 Prolog 模組 (Fandino 2006)，以 perl 程式處理字串後直接在 perl 程式內呼叫 Prolog 程式。如下圖輸入兩個詞彙選擇其中的意義後會找出兩個詞彙之間的語意關係。兩個詞之間的語意的路徑不是唯一的。

圖一 輸入兩個詞並選擇意義後可以得到兩個詞之間的語義關係的路徑



總之，Hownet 裡面沒有明確的同義詞和反義詞。同義的關係必須靠義元比對。反義的關係則被事件角色轉換關係裡面的 mutual precondition 等所部分取代。上下位關係也在事件角色轉換關係裡面記載。至於部分與全體的關係在 Hownet 的裡面透過 whole，part 等義元來表示。例如，輪胎和汽車的義元分別爲

輪胎(tire) {part|部件:whole={part|部件:PartPosition={leg|腿},whole={LandVehicle|車}}}

汽車(automobile) {LandVehicle|車}

輪胎是汽車的一部份這個關係可以透過剖析 Hownet 義元的結構得到。

上面花了相當多的篇幅探討如何利用 Hownet 找出詞彙語意的關係。Howent 雖然提供許多的語義訊息，但是有些地方不一致，而且對於同義詞，反義詞並沒有清楚的記載。英文 Wordnet 裡面則詳細記載了同義關係(synset),反義關係，部分關係，全體關係，上位關係，下位關係。我們利用 Wordnet ::QueryData 和 Wordnet::Similarity 這兩個 Perl 模組及漢英辭典，輸入兩個中文詞，利用漢英辭典將中文詞轉換成英文詞，再利用前述兩個 Perl 模組，即可得到兩個詞之間是否爲同義關係,反義關係，部分關係，全體關係，上位關係，下位關係。例如輸入男人與女人可以找到兩者之間是反義詞。

男人 = boy/dick/joe/man/buck/hombre/blighter/menfolk/husband
女人 = jane/dame/women/judy/donah/wife/womenfolk/frow/hen/tomato/frau/woman/female
女人(woman) is 男人(man)'s antonyms
男人(man) is 女人(woman)'s antonyms
女人(wife) is 男人(husband)'s antonyms
男人(husband) is 女人(wife)'s antonyms

輸入汽車與救護車可以得到汽車是救護車的上位詞，救護車是汽車的下位詞。

汽車 = auto/car/machine/motorcar/motor/autocar/automobile
救護車 = ambulance
汽車(car) is 救護車(ambulance)'s hypernyms
救護車(ambulance) is 汽車(car)'s hyponyms

透過 Pedersen Wordnet::Similarity 這個模組，我們可以得到兩個詞之間語意相似度。如同 Hownet 的事件角色轉換關係，兩個詞的語義路徑不止一條，通常最短的那一條較符合我們的直覺。

汽車 = auto/car/machine/motorcar/motor/autocar/automobile
救護車 = ambulance

WordNet::Similarity

auto#n#1 - ambulance#n#1 : 0.96
auto#n#1 -- motor_vehicle#n#1 -- car#n#1 -- ambulance#n#1

car#n#1 - ambulance#n#1 : 0.96
car#n#1 -- ambulance#n#1

car#n#2 - ambulance#n#1 : 0.782608695652174
car#n#2 -- wheeled_vehicle#n#1 -- self-propelled_vehicle#n#1 -- motor_vehicle#n#1 -- car#n#1 -- ambulance#n#1

car#n#3 - ambulance#n#1 : 0.5
car#n#3 -- compartment#n#2 -- room#n#1 -- area#n#4 -- structure#n#1 -- artifact#n#1 --

instrumentality#n#3 -- container#n#1 -- wheeled_vehicle#n#1 -- self-propelled_vehicle#n#1 -- motor_vehicle#n#1 -- car#n#1 -- ambulance#n#1

但 Wordnet 本身的限制造成應該找到的關係卻沒有找到，例如醫生和病人，員工和雇主並沒有找到反義的關係。

醫生 = medic/aesculapius/physician/surgeon/hakeem/medico/doctor/housestaff
病人 = in-patient/invalid/valetudinarian/patient/inpatient/case
No relationship

員工 = staff/personnel
雇主 = hirer/employer/gaffer
No relationship

翻譯可能出錯及 Wordnet 本身的不一致，即使找到語意關係不見得是正確的。例如我們輸入汽車和輪子，應該找到汽車是輪子的全體詞，卻找到汽車是輪子的上位詞。

汽車 = auto/car/machine/motorcar/motor/autocar/automobile
輪子 = wheel
汽車(machine) is 輪子(wheel)'s hypernyms
輪子(wheel) is 汽車(machine)'s hyponyms

而輸入汽車和輪胎卻找不到任何關係。

汽車 = auto/car/machine/motorcar/motor/autocar/automobile
輪胎 = tire
No relationship

由於 Wordnet 也有一些不一致的地方，我們進一步整合其它的資源。將教育部重編國語辭典修訂本的內容剖析後找出包括解釋、同義詞、反義詞等三類重要語意訊息。教育部重編國語辭典修訂本裡面的相似詞和相反詞事實上就是同義詞和反義詞，不過各個詞條之間內容並不一致，例如輸入醫生得到下面的同義詞與反義詞。

醫生的同義辭 : 醫師,大夫,
醫生的反義辭 : 病人,

但是輸入大夫卻沒有反義詞。我們將所有的同義詞，反義詞連結起來，去掉不一致的地方，重新建立一個資料庫。

除了教育部重編國語辭典修訂本外，我們也將現代漢語分類辭典建成資料庫。現代漢語分類辭典提供比同義詞詞林更詳盡的語義類別訊息。例如：犯罪，犯法的上一層語意是罪刑，刑罰，再上一層是政治，法律，軍事。同屬最下一層的詞是同義或近義詞。如果我們輸入犯罪和法律可以得到這兩個詞共有的類別。

分類辭典: 不屬於相同的語義類別, 共同的部份: 政治●法律●軍事
犯罪 : 政治●法律●軍事→罪行●刑罰→犯罪・違法

法律 : 政治●法律●軍事→法律（一般）→法（一般）

由於每一個詞彙知識庫都有優缺點我們把 Hownet、Wordnet、教育部國語會辭典、現代漢語分類辭典這四個資源完全整合起來。程式在 (http://nlp2.csie.org/~denehs/word_relation.html)。

輸入桌子，椅子，我們得到下面結果。

教育部重編國語辭典修訂本: 不是同義詞
桌子的同義辭 : Not found.
椅子的同義辭 : Not found.
教育部重編國語辭典修訂本: 不是反義詞
桌子的反義辭 : Not found.
椅子的反義辭 : Not found.

分類辭典: 不屬於相同的語義類別, 共同的部份: 飲食●衣服●居住●財產→家用器物

桌子 : 飲食●衣服●居住●財產→家用器物→桌子
椅子 : 飲食●衣服●居住●財產→家用器物→坐具・椅・凳

Hownet

桌子 V.S. 椅子

HowNet 義元
桌子(table) {furniture|家具:{put|放置:location={~}}}
椅子(chair) {furniture|家具:{sit|坐蹲:location={~}}}

Events
(No Event Match)

Relationship
桌子(table) 椅子(chair) 共同點: {furniture|家具} 共同語意特徵: (無)

Wordnet
桌子 = desk/table
椅子 = chair
No relationship

我們再輸入桌子，家具測試這幾個詞彙知識庫。

教育部重編國語辭典修訂本: 不是同義詞
桌子的同義辭 : Not found.
家具的同義辭 : Not found.
教育部重編國語辭典修訂本: 不是反義詞
桌子的反義辭 : Not found.

家具的反義辭 : Not found.

分類辭典: 不屬於相同的語義類別, 共同的部份: 飲食●衣服●居住●財產→家用器物

桌子 : 飲食●衣服●居住●財產→家用器物→桌子
家具 : 飲食●衣服●居住●財產→家用器物→家具
Hownet

桌子 V.S. 家具

HowNet 義元
桌子(desk) {furniture|家具:{put|放置:location={~}}}
家具(furniture) {furniture|家具}

Events
(No Event Match)

Relationship

桌子(desk) 家具(furniture) 共同點: {furniture|家具} 共同語意特徵: (無)

Wordnet
桌子 = desk/table
家具 = furniture/movable
家具(furniture) is 桌子(table)'s hypernyms
桌子(table) is 家具(furniture)'s hyponyms

我們可以從不同的詞彙知識庫得到不同的訊息，但是對於因果關係，目前我們建立的四個詞彙知識庫仍然無法找到。例如輸入犯罪和入獄兩個詞得到下列輸出結果。

教育部重編國語辭典修訂本: 不是同義詞
犯罪的同義辭 : 違法,違警,坐法,犯法,犯科,犯警,
入獄的同義辭 : 下獄,坐牢,
教育部重編國語辭典修訂本: 不是反義詞
犯罪的反義辭 : 立功,犯案,
入獄的反義辭 : 出獄,

分類辭典: 不屬於相同的語義類別, 共同的部份: 政治●法律●軍事→罪行●刑罰
犯罪 : 政治●法律●軍事→罪行●刑罰→犯罪・違法
入獄 : 政治●法律●軍事→罪行●刑罰→關押・監禁

Hownet
HowNet 義元
犯罪(commit a crime) {do|做:content={fact|事情:modifier={guilty|有罪}}}
入獄(put in prison) {suffer|遭受:cause={guilty|有罪},content={detain|扣住},domain={police|警}}

Events
(No Event Match)

Relationship

犯 罪 (commit crime)	a 入 獄 (put in prison)	共同點: 不屬於相同的語意類別	共同語意特徵: (無)

Wordnet
犯罪 =
crime/malefaction/misdeed/sin/maleficent/commitment/misdoing/transgress/perpetration/delinquency/guilt/guilty/wrongdoing/trespass
入獄 = be jailed
No relationship

四個詞彙知識庫只有在現代漢語分類辭典裡面找到兩個詞有相同的語意類別政治●法律●軍事→罪行●刑罰。但是兩者之間因果的關係並無法得到。

上面的 14 組測試資料顯示雖然有不少關係可以透過我們整合後的資料擷取出來，但是扔然有不少關係無法得到。

六 結論

我們利用了四個大規模的詞彙知識庫，並開發了不少的工具程式，輸入任兩個詞彙可以找出下列的語義關係（一）同義關係（二）反義關係（三）上下位關係（四）部件與整體關係 （五）相同事件（六） 相同領域(domain) （七）相同語意特徵（八）相同的語意類別（九）事件與語意角色。

從我們的測試資料顯示單純只靠詞彙知識庫無法得到所有的詞彙語義關係。下一階段將會進一步加入用現有的各種資源如中研院句法樹庫資料，FrameNet，VerbNet，PopNet，Sketch Engine，充分結和語料庫和詞彙知識庫的優點，將統計，語意，語法結合起來擷取更多的詞彙知識。

致謝

本研究得到國科會計畫「詞彙語意關係之自動標注─以中英平行語料庫爲基礎(I)（II）（III）」NSC91-2411-H-002-080 NSC92-2411-H-002-061 NSC93-2411-H-002-013 經費補助，特此致謝。本研究建構的系統由台大資工系高紹航，黃子桓，江加恩，台大資管系戴士強程式設計一併致謝。

參考文獻

Berners-Lee, Tim. (2000) Weaving the Web : the original design and ultimate destiny of the World Wide Web by its inventor. New York : HarperBusiness.

Boguraev, Branimir. and Briscoe, Ted. (1989) Computational Lexicography for Natural Language Processing. Longman: Harlow.
Boguraev, Branimir and Pustejovsky, James (eds.) (1996) Corpus Processing for Lexical Acquisition, MIT Press.
Chaffin, Roger and Illerrmann, Douglas. (1988) The Nature of Semantic Relations: a Comparisons of Two Approaches. In Evens (eds) (1988), pp. 289-334.
Church, K. and Hanks, P. (1990) "Word Association Norms, Mutual Information, and Lexicography." Computational Linguistics, Vol. 16, No. 1, pp. 22-29.
Church, K. et al. (1991) "Parsing, Word Associations, and Typical Predicate-Argument Relations." In Tomita (ed) *Current Issues in Parsing Technology*, Kluwer.
Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. (1994) 'Lexical Substitutability,' in Atkins and Zampolli (eds.) Computational Approaches to the Lexicon, pp. 153- 177. Oxford, Oxford University Press.
Cruse, Allan. (1986) Lexical Semantics. Cambridge: Cambridge University Press.
Dong, Zhendong and Dong, Qiang. (2006) Hownet and the Computation of Meaning. World Scientific.
Evens, Martha. (eds.) (1988) Relational Models of the Lexicon: Representing Knowledge in Semantic Networks. Cambridge University Press.
Fillmore, Charles. (1968) The Case for Case. In E. Bach and R. T. Harms, eds., Universals in Linguistic Theory, Holt, Riinehart and Winston, New York, 1-88.
Koenig, Jean-Pierre. (1999) Lexical Relations. CSLI , Stanford University.
Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007), SemEval-2007 Task 04: Classification of Semantic Relations between Nominals, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic, pp. 13-18.
Grefefenstette, Gregory. (1994) Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers.
Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 539–545, Nantes, France.
Jackendoff, Ray. (1983) Semantics and Cognition. Cambridge, Mass.: MIT Press.
Jackendoff, Ray. (1990) Semantic Structures. Cambridge, Mass.: MIT Press.
Jones, Stevens. (2002). Antonymy: A Corpus-based Perspective. London ; New York : Routledge, 2002
Levin, Beth. (1985) 'Introduction,' in B. Levin (ed.) Lexical Semantics in Review, Lexicon Project Working Papers 1, Center for Cognitive Science, MIT, pp. 1-62.
Melcuk, Igor. (1988) 'The Explanatory Combinatory Dictionary,' in M. Evens (ed.) (1988), pp. 41 - 74.

Pedersen, Patwardhan, and Michelizzi (2004) WordNet::Similarity - Measuring the Relatedness of Concepts - Appears in the *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pp. 1024-1025, July 25-29, 2004, San Jose, CA (Intelligent Systems Demonstration)
Pustejovsky, James, Sabine Bergler, and Peter Annick (1993) 'Lexical Semantic Techniques for Corpus Analysis,' Computational Linguistics, Vol. 19, No. 2, pp. 331 - 358.
Pustevojsky, James. (1995)The Generative Lexicon. The MIT Press.
Pustevojsky, James. (2000) Syntagmatic Processes. in Handbook of Lexicology and Lexicography, de Gruyter, 2000.
Resnik, Phillip. (1992) 'WordNet and Distributional Analysis: A Class-based

Approach to Lexical Discovery,' in Workshop Notes, Statistically-Based NLP Techniques, American Association for Artificial Intelligence, pp. 109 - 113.
Schank, Roger. (1975) Conceptual Information Processing. Amsterdam: North-Holland.
Sinclair, John. (eds). (1987) Looking up. Glasglow: Collins.
Sowa, John F. (1984) Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley.
Turney, P.D. (2006), Expressing implicit semantic relations without supervision, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL-06)*, Sydney, Australia, pp. 313-320.
Wilks, A. Yorick (1968) On-line Semantic Analysis of English Texts. Machine Translation, Vol. 11, pp. 59-72.

董大年(主編)(1998)現代漢語分類辭典。上海：漢語大辭典出版社。

魯川 (2001) 漢語語法的意合網路。北京：商務印書館。

梅家駒(主編)(1984)同義詞詞林。北京：商務印書館。

軟體

British National Corpus http://www.natcorp.ox.ac.uk/

DOLCE ontology http://www.loa-cnr.it/DOLCE.html

FrameNet

Hownet http://www.keenage.com/

Language::Prolog::Yaswi

http://search.cpan.org/~salva/Language-Prolog-Yaswi-0.14/Yaswi.pm
Lexical Freenet http://www.cinfn.com/doc/
Ontosaurus http://www.isi.edu/isd/ontosaurus.html
PopNet http://verbs.colorado.edu/~mpalmer/projects/ace.html

Sketch Engine http://www.sketchengine.co.uk/
Wordnet http://wordnet.princeton.edu/

VerbNet http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
Wordnet::Similarity http://www.d.umn.edu/~tpederse/similarity.html
Wordnet ::QueryData http://people.csail.mit.edu/jrennie/WordNet/
中文詞彙網路 Chinese Wordnet (CWN) http://cwn.ling.sinica.edu.tw/
教育部重編國語辭典修訂本 http://140.111.34.46/dict/

中文單詞之韻律模式研究

A Study on Prosodic Modeling for Isolated Mandarin Words

陳啓風　Chi-Feng Chen
國立交通大學電信工程學系
Department of Communication Engineering
National Chiao Tung University
linuxe.cm94g@nctu.edu.tw

江振宇　Chen-Yu Chiang
國立交通大學電信工程學系
Department of Communication Engineering
National Chiao Tung University
gene.cm91g@nctu.edu.tw

王逸如　Yih-Ru Wang
國立交通大學電信工程學系
Department of Communication Engineering
National Chiao Tung University
yrwang@cc.nctu.edu.tw

陳信宏　Sin-Horng Chen
國立交通大學電信工程學系
Department of Communication Engineering
National Chiao Tung University
schen@mail.nctu.edu.tw,

摘要

在本文中，我們對中文單詞提出了以音節爲基本單位的基頻軌跡及音節長度的韻律模型。在基頻軌跡模型中，我們考慮了聲調、音節在詞的位置以及前後音節連音的三種影響因素，並假設這些影響因素彼此獨立相加而組成音節基頻軌跡。在音節長度模型中，我們考慮了聲調、音節在詞的位置、基本音節以及前後音節連音的四種影響因素，我們同樣假設這些影響因素爲彼此獨立且具加成性。我們使用一個含 107,936 個詞的單一女性語者的語料庫來評估所提方法是否有效，我們並用決策樹來分析音節長度如何受音節音素結構的影響，也用決策樹來分析音節間 pause 的長度和前後音節音素結構的關係，實驗結果顯示訓練後此兩韻律模型的影響因素都符合我們對中文韻律的知識。

Abstract

In this paper, syllable-based prosody modelings of pitch contour and syllable duration for

isolated Mandarin words are proposed. In the syllable pitch contour model, three main affecting factors of tone, syllable position in word, and inter-syllable coarticulation are considered. These three affecting factors are assumed to be independent and additive. Similarly, in the syllable duration model, four affecting factors of tone, syllable position in word, base-syllable, and inter-syllable coarticulation are considered. We also assume that these affecting factors are independent and additive. A large single female-speaker speech database containing 107,936 words was used to evaluate the performance of the proposed methods. After well-training, the decision tree method was used to analyze the 411 affecting factors of base-syllable and to explore the relationship between inter-syllable pause duration and the nearby linguistic features. Experimental results showed that all these affecting factors conformed to our knowledge about Mandarin prosody.

關鍵詞：韻律模式，基頻軌跡，影響因素，連音

Keywords: Prosody modeling, Pitch contour, Affecting factor, Coarticulation

一、緒論

文句轉語音系統要能合成出自然流利的語音，關鍵在於韻律的變化是否自然順暢。韻律的變化包括音調的高低起伏、音量的強弱、發音的長短及停頓的時機、長度等。目前韻律的合成方法大致分爲規則法 [1,2]、類神經網路 [3,4]和統計法。規則法是以語言學的方法，歸納出一些發音的規則，利用這些規則來產生合成語音的韻律。但是人類說話的方式變化複雜，不容易掌握。類神經網路是利用一組複雜的網路來模擬人腦的記憶與學習功能，其學習方法是採用漸進式的修正錯誤與更新記憶的方式，需經由長時間的學習訓練，雖有不錯的效果，但無法分析影響韻律的因素。本文以統計法的方式，可從大量的語音資料中統計出韻律變化，利用所考慮影響韻律的因素加總後，控制韻律變化，並分析各個影響因素對韻律訊息的影響程度。

本文著重於以中文單詞語料庫爲基礎之韻律模式的研究，探討音節的基頻軌跡及長度模式，考慮幾個主要的影響因數，希望藉此了解中文單詞的音節基頻軌跡及長度如何變化，以作爲未來中文語音合成系統產生韻律信息之用，期望合成出自然流暢的中文單詞聲音。本論文在接下來的第二部份會介紹我們所提出的韻律模型，第三部份介紹模型的訓練方法，實驗結果在第四部份討論，最後於第五部份對於本研究給予一個結論。

二、韻律模式

韻律模式以音節爲單位，在給予特定的語言資訊後，使之預測音節的基頻軌跡及長度，做爲韻律訊息，並分析音節基頻軌跡及長度在各個因素的影響程度。考慮主要影響因素分別爲：聲調(tone)、音節在詞的位置(word-position)、基本音節(base syllable)、音節間的連音狀態(inter-syllable coarticulation state)。

（一）、基頻軌跡之韻律模型

假設所有影響因素可用累加的方式來表示音節的基頻軌跡，如式子(1)：

$$\mathbf{sp}_n = \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{w_n} + \boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{c_n,tp_n} + \boldsymbol{\mu}^p \quad (1)$$

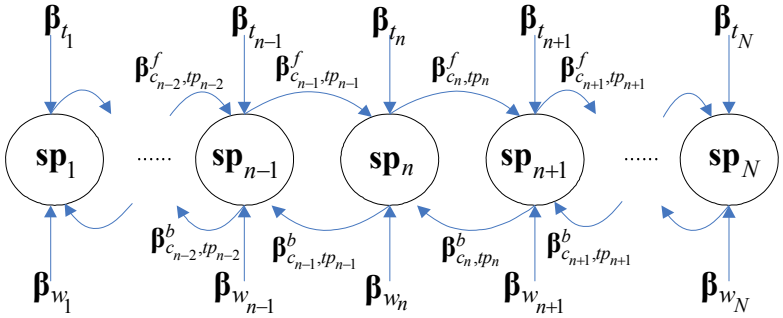
其中 $\mathbf{sp}_n$ 、$\mathbf{sp}_n^r$ 、$\boldsymbol{\beta}_{t_n}$ 、$\boldsymbol{\beta}_{w_n}$ 分別爲 pitch 模型中第 n 個音節的基頻軌跡參數向量、基頻軌跡參數向量殘餘值(residual)、聲調及詞中位置影響因素；$\mathbf{sp}_n$ 爲由一段音節基頻軌跡轉化爲四個正交參數表示的向量，轉換方法參見 [5]；$w_n \in \{(2,1),(2,2),..(j,k),..(8,8)\}$ 代表音節在詞的位置，其中 (j,k) 代表 j 字詞中的第 k 個音節；c_n 、tp_n=(t_n,t_{n+1}) 分別爲在第 n 個音節與第 n+1 個音節間的連音狀態、聲調組合(tone pair)，在這裡音節間的連音狀態 $c_n \in \{c1, c2, c3\}$ 代表音節間的連音程度，c1 、c2 、c3 分別爲強連音(tight)、正常連音(normal)、弱連音(loose)；$\boldsymbol{\beta}^b_{c_n,tp_n}$ 爲第 n 個音節受第 n+1 個音節的後向影響因素(backward affecting factor)；$\boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}}$ 爲第 n 個音節受第 n-1 個音節的前向影響因素(forward affecting factor)；$\boldsymbol{\mu}^p$ 爲基頻軌跡參數的整體平均(global mean)。

（二）、音節長度模型

假設所有影響因素可用累加的方式來表示音節的長度，如式子(2)：

$$sd_n = sd_n^r + \gamma_{t_n} + \gamma_{w_n} + \gamma_{sy_n} + \gamma^f_{c_{n-1},fi\_in_{n-1}} + \gamma^b_{c_n,fi\_in_n} + \mu^d \tag{2}$$

其中 sd_n 、sd_n^r 、γ_{t_n} 、γ_{w_n} 、γ_{sy_n} 分別爲 duration 模型中第 n 個音節的長度、長度殘餘值、聲調、詞中位置及基本音節影響因素；c_n 、fi_in_n 分別爲在第 n 個音節與第 n+1 個音節間的連音狀態及第 n 個音節韻母類別與第 n+1 個音節聲母類別之組合(final-initial class pair)；$\gamma^b_{c_n,fi\_in_n}$ 爲第 n 個音節長度受第 n+1 個音節的後向影響因素(backward affecting factor)；$\gamma^f_{c_{n-1},fi\_in_{n-1}}$ 爲第 n 個音節長度受第 n-1 個音節的前向影響因素(forward affecting factor)；μ^d 爲音節長度的整體平均(global mean)。音節韻律與影響因素的關係示意圖以 pitch 模型爲例，如圖一：



圖一、音節基頻軌跡參數向量與影響因素關係圖

我們分別假設 $\mathbf{sp}_n^r$ 及 sd_n^r 呈 $N(\mathbf{sp}_n^r;\mathbf{0},\mathbf{R}^p)$ 及 $N(sd_n^r;0,R^d)$ 的高斯分佈(Gaussian distribution)，因此 $\mathbf{sp}_n$ 與 sd_n 可表示成數學式如式(3)及(4)：

$$P(\mathbf{sp}_n \mid t_n,w_n,c_{n-1},c_n,tp_{n-1},tp_n) = N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{w_n}+\boldsymbol{\beta}^f_{c_{n-1},tp_{n-1}}+\boldsymbol{\beta}^b_{c_n,tp_n}+\boldsymbol{\mu}^p,\mathbf{R}^p) \tag{3}$$

$$P(sd_n \mid t_n,w_n,sy_n,c_{n-1},c_n,fi\_in_{n-1},fi\_in_n) = N(sd_n;\gamma_{t_n}+\gamma_{w_n}+\gamma_{sy_n}+\gamma^f_{c_{n-1},fi\_in_{n-1}}+\gamma^b_{c_n,fi\_in_n}+\mu^d,R^d) \tag{4}$$

三、模型的訓練

爲了求取韻律模式的各個參數，我們採用 sequential optimization 的方法以及最大相似度法則(Maximum likelihood criterion)的條件來訓練模型，我們首先定義相似度函數(likelihood function)如下式：

$$L^{P}=\sum_{n=1}^{N}\log N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{w_n}+\boldsymbol{\beta}^{f}_{c_{n-1},tp_{n-1}}+\boldsymbol{\beta}^{b}_{c_n,tp_n}+\boldsymbol{\mu}^{p},\mathbf{R}^{p}) \tag{5}$$

$$L^{d}=\sum_{n=1}^{N}\log N(sd_n;\gamma_{t_n}+\gamma_{w_n}+\gamma_{sy_n}+\gamma^{f}_{c_{n-1},fi\_in_{n-1}}+\gamma^{b}_{c_n,fi\_in_n}+\mu^{d},R^{d}) \tag{6}$$

pitch 以及 duration 模型獨立訓練各自的參數，且訓練方法類似，訓練的過程可分爲兩大部分，第一部分爲參數的初始化，第二部份爲以疊代法的 sequential optimization。以下以訓練 pitch 模型爲例：

（一）、參數的初始化(Initialization)

(a) 直接平均所有音節的 $\mathbf{sp}_n$，求出整體 pitch 平均(global pitch mean) $\boldsymbol{\mu}^{p}$

(b) 以下式求取聲調影響因素的初始值：

$$\boldsymbol{\beta}_t=\frac{\sum_{n}\left((\mathbf{sp}_n-\boldsymbol{\mu}^{p})\delta(t_n=t)\right)}{\sum_{n}\delta(t_n=t)}\quad,\text{for } t=1,2..5 \tag{7}$$

(c) 以下式求取詞位置影響因素的初始值：

$$\boldsymbol{\beta}_w=\frac{\sum_{n}\left((\mathbf{sp}_n-\boldsymbol{\beta}_{t_n}-\boldsymbol{\mu}^{p})\delta(w_n=w)\right)}{\sum_{n}\delta(w_n=w)}\quad,\text{for } w=(2,1),(2,2)..(8,8) \tag{8}$$

(d) 以下列的條件，標記音節間的連音狀態 c_n

I. 若兩音節間基頻軌跡相連接，表示兩連音互相影響程度強，連音狀態標記爲強連音 c1 。

II. 兩音節間的基頻軌跡不相連接，但音節間的間隔區間內最低能量較大(大於一個臨界值)，連音狀態標記爲正常連音 c2 。

III. 不滿足以上條件者，則連音狀態標記爲弱連音 c3 。

(e) 求取前後音節影響因素的初始值，如下式：

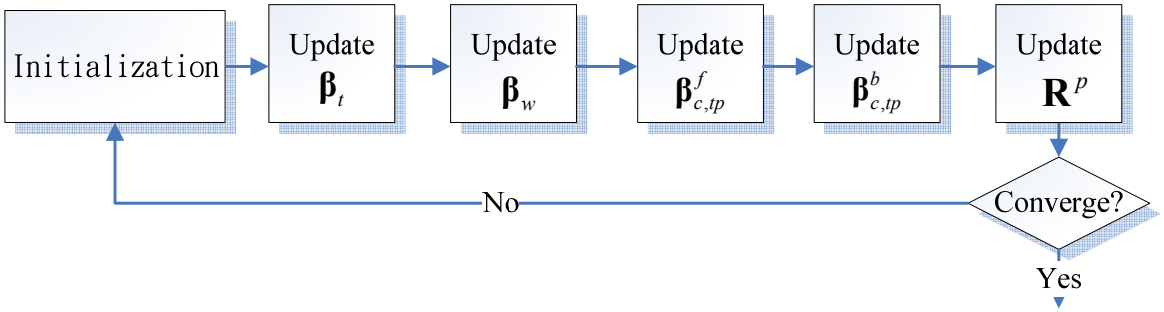
$$\boldsymbol{\beta}^{f}_{c,tp}=\frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta(c_{n-1}=c)\delta(tp_{n-1}=tp)}{\sum_{n=1}^{N}\delta(c_{n-1}=c)\delta(tp_{n-1}=tp)}-\frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta(c_{n-1}=c)\delta(t_n=j)}{\sum_{n=1}^{N}\delta(c_{n-1}=c)\delta(t_n=j)} \tag{9}$$

$$\boldsymbol{\beta}^{b}_{c,tp}=\frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta(c_n=c)\delta(tp_n=tp)}{\sum_{n=1}^{N}\delta(c_n=c)\delta(tp_n=tp)}-\frac{\sum_{n=1}^{N}\mathbf{sp}_n\delta(c_n=c)\delta(t_n=i)}{\sum_{n=1}^{N}\delta(c_n=c)\delta(t_n=i)} \tag{10}$$

For c = 1~3 and tp=(i,j)

（二）、以疊代法的 sequential optimization

各個影響因素初始化後依序將聲調($\boldsymbol{\beta}_t$)、word-position($\boldsymbol{\beta}_t$)、受前後音節等影響因素($\boldsymbol{\beta}^f_{c,tp}$, $\boldsymbol{\beta}^b_{c,tp}$)及 covariance matrix($\mathbf{R}^p$)的參數值更新，然後使用更新後的參數值，算出整個訓練語料的目標函數值，一直重覆更新參數值及目標函數值，直到含數值收斂，如圖二之流程圖：



圖二、訓練流程圖

而 duration model 各個參數更新方法與 pitch model 類似，而參數更新的順序爲聲調(γ_t)、word-position(γ_w)、基本音節(γ_{sy})、受前後音節等影響因素(γ^f_{c,fi_in}, γ^b_{c,fi_in})及 covariance matrix(R^d)。

四、實驗結果與分析

實驗語料庫之單詞來自於『NCTU 文句分析器』的詞典選擇而來，以聲調平衡爲主要的選擇條件，總共有 107936 個詞，277218 個字，其中詞長最短爲二字詞、最長八字詞，詞長統計如表一，聲調統計如表二，語料庫是由專業的女性廣播人員以流利的方式唸出錄製，錄音場所爲一般安靜房間。

表一、詞長數量之統計

詞長	二字詞	三字詞	四字詞	五字詞	六字詞	七字詞	八字詞
數量	64872	26026	16062	797	124	49	6

表二、聲調數量之統計

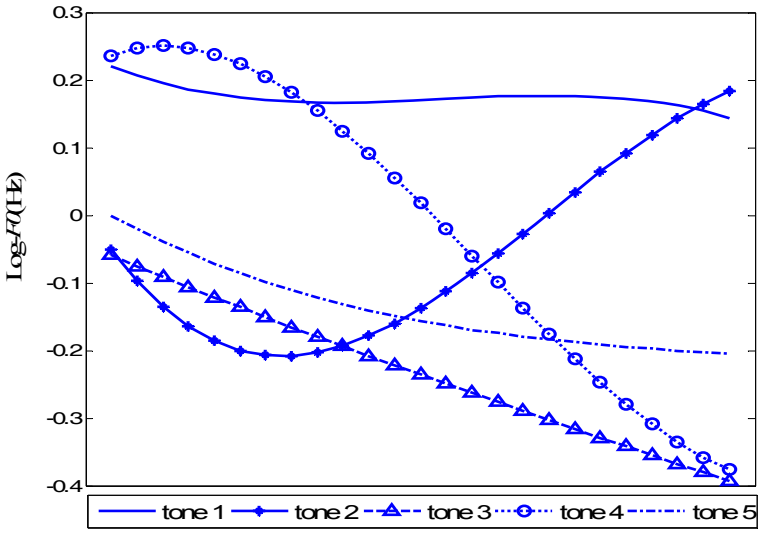
聲調	一聲	二聲	三聲	四聲	五聲
數量	62349	69278	48904	94786	1901

接下來我們依序分析基頻軌跡、音節長度韻律模型以及預測 pause 長度。

（一）、基頻軌跡韻律模型

1、聲調影響因素(Tone affecting factor) $\boldsymbol{\beta}_t$

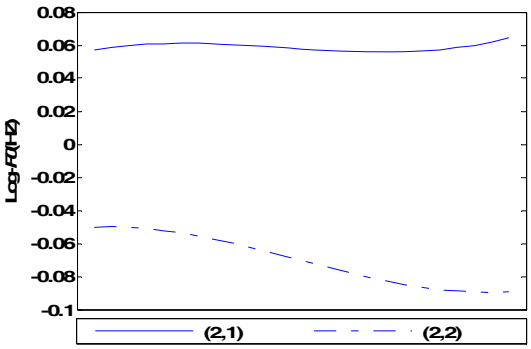
聲調影響因素的基頻軌跡如圖三所示，觀察得知，由模型所訓練出來的聲調影響因素符合我們所認知的聲調基頻軌跡。



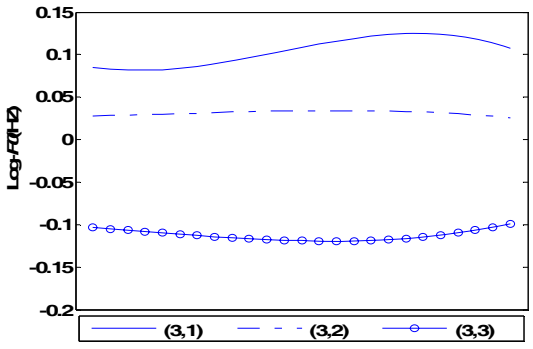
圖三、聲調影響因素基頻軌跡

2、音節在詞的位置影響因素(Word position affecting factor) $\boldsymbol{\beta}_w$

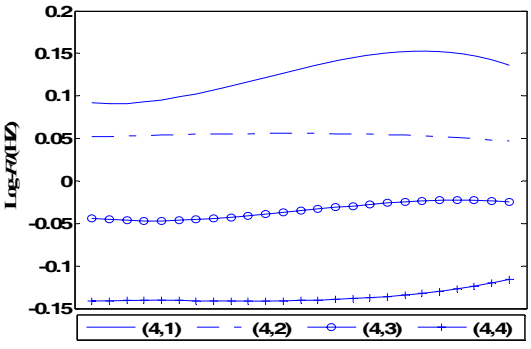
Word position affecting factor 的基頻軌跡影響如下圖所示，分別有二字詞至五字詞位置影響因素(圖四~七)，同時將詞首與詞末的基頻軌跡整合比較(圖八、九)，可觀察得，音節在詞的位置愈接近詞首，基頻愈高，愈接近詞末愈低，且詞首有上仰趨勢(圖八)，在詞尾均有微幅的上揚(圖九)，同時我們也發現到詞長越長，則基頻軌跡變化的動態範圍愈大，如二字詞的動態範圍在 0.06 ~ -0.08，而五字詞的動態範圍在 0.15 ~ -0.17 之間。



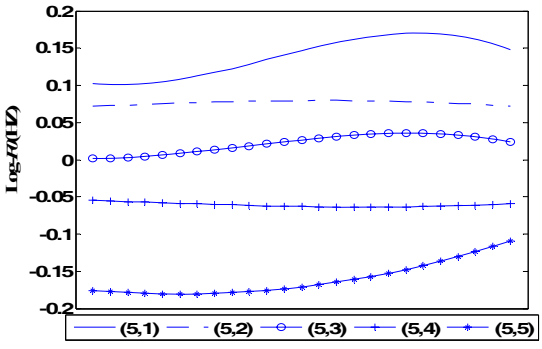
圖四、二字詞影響因素基頻軌跡



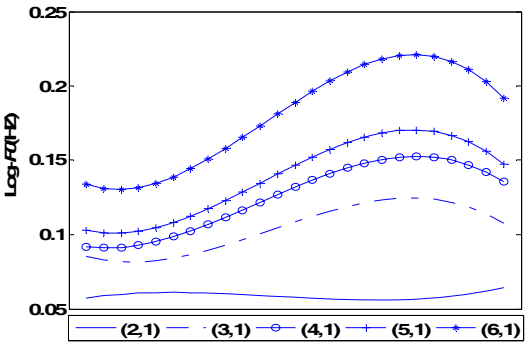
圖五、三字詞影響因素基頻軌跡



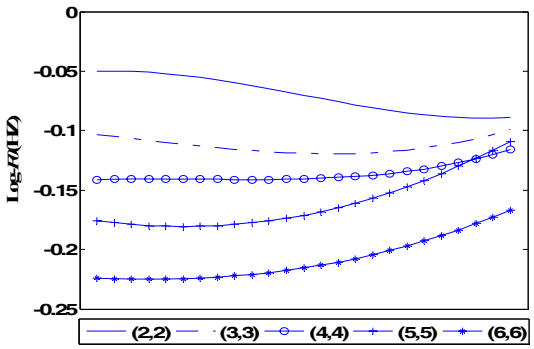
圖六、四字詞影響因素基頻軌跡



圖七、五字詞影響因素基頻軌跡



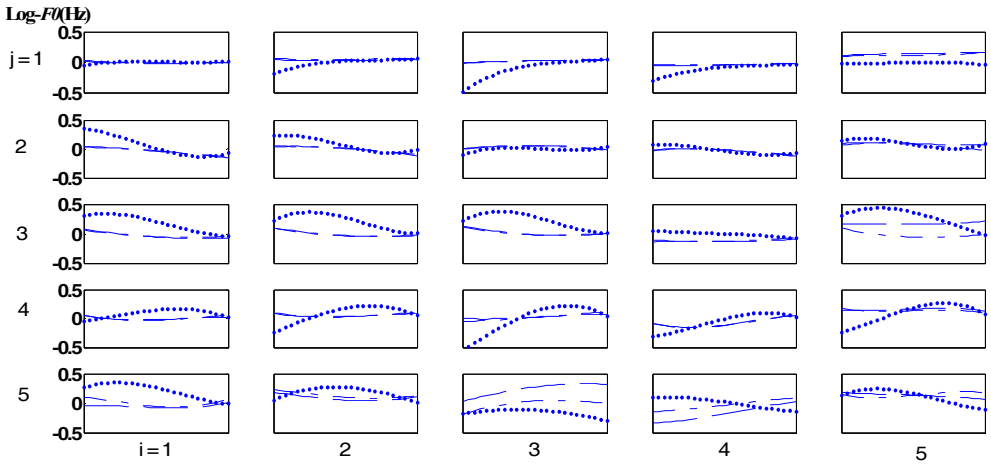
圖八、整合詞首基頻軌跡



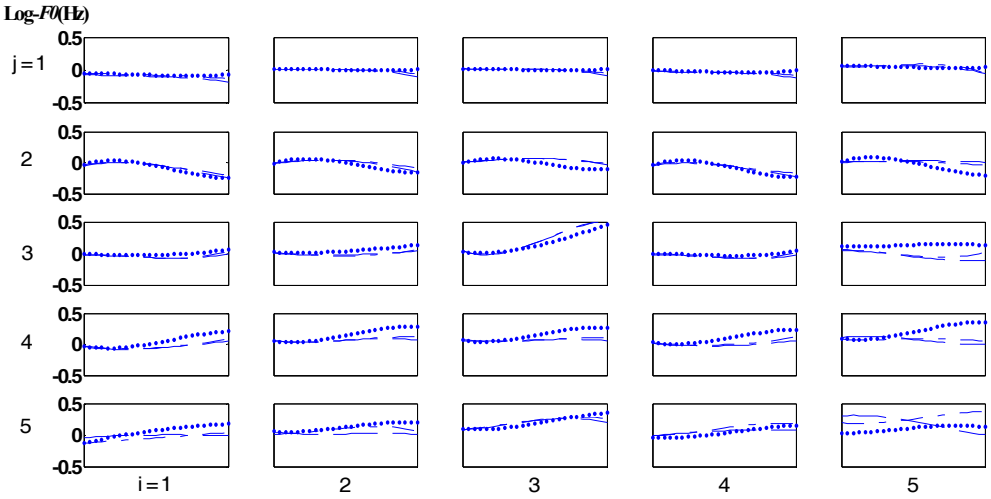
圖九、整合詞末基頻軌跡

3、受前後音節影響因素(forward and backward affecting factor)

圖十爲五個聲調受前音節中各個聲調在各個連音狀態影響的基頻軌跡。可觀察得c1明顯比c2及c3易受影響。以c1爲觀察對象，目前音節的基頻軌跡前端高度，受前一個音節基頻軌跡後端高度影響，若前音節基頻軌跡後端比受影響的前端聲調基頻軌跡還高，則受影響的前端基頻軌跡會往上改變，反之往下改變。圖十一爲受後音節的影響，其影響原理類似於受前音節的影響因素，可觀察得受後音節較受前音節的基頻軌跡影響相對較小。



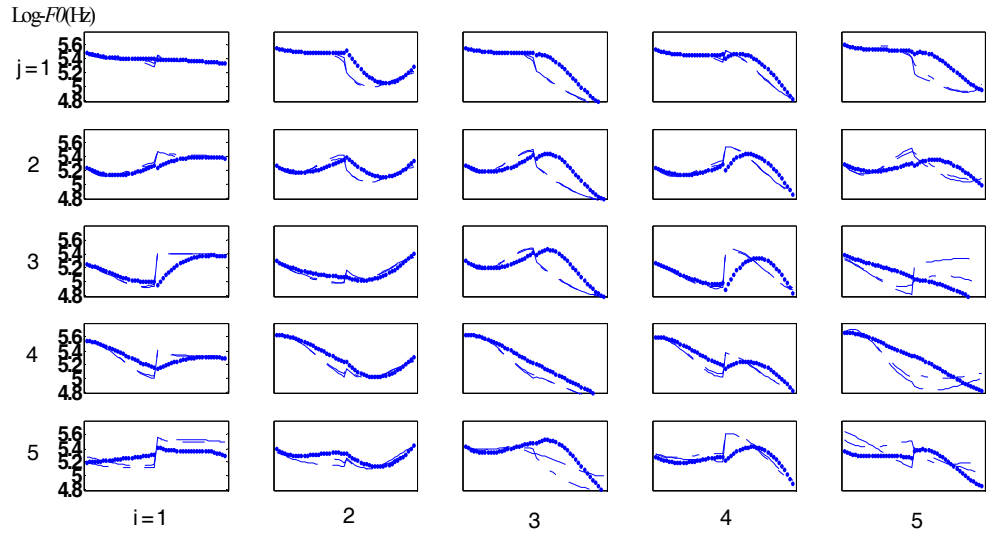
圖十、Forward affecting factor $\boldsymbol{\beta}^{f}_{c,tp}$ ， $c=\{c1,c2,c3\}$ ， $tp=(i,j)$ ，其中點線(…)爲c1、點虛線(--・)爲c2、虛線(--)爲c3



圖十一、Backward affecting factor $\boldsymbol{\beta}^{b}_{c,tp}$，$c=\{c1,c2,c3\}$，$tp=(j,i)$，其中點線(…)爲 c1、點虛線(--・)爲 c2、虛線(--)爲 c3

4、二字詞基頻軌跡預測

圖十二爲各種二字詞聲調組合(tone pair)的預測結果，假設二字詞音節基頻軌跡相連且相互影響，可觀察得 c1 明顯易受影響，而且 c1 在受前後音節影響下，兩音節基頻軌跡相連續且平滑。而三聲接三聲變二聲接三聲的語言特性也可由下圖證實。

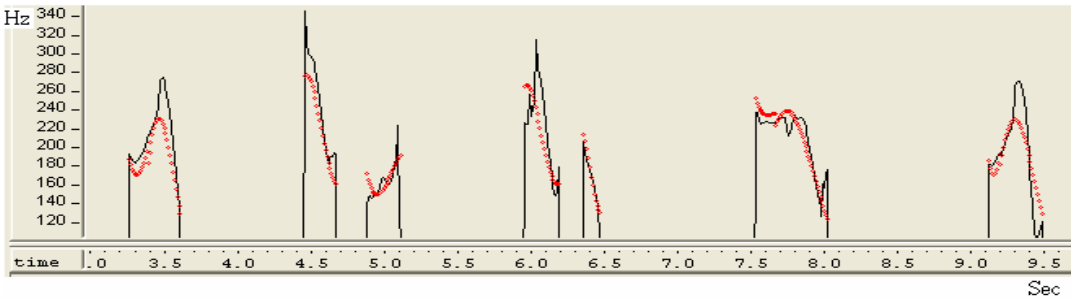


圖十二、二字詞基頻軌跡變化圖，其中點線(…)爲 c1、點虛線(--・)爲 c2、虛線(--)爲 c3，其 j 爲第一音節，i 爲第二個音節

5、預估之基頻軌跡範例

訓練模型後，pitch 模型預測音節基頻軌跡結果如下圖所示，其中黑色線(實線)爲原始音

節的基頻軌跡，紅色線(點線)爲 pitch 模型所預測的基頻軌跡，可觀察得，不僅個別音節的基頻軌跡走勢相似，且強連音音節基頻軌跡也預測的不差。



圖十三、Pitch 模型預測基頻軌跡

6、Covariance matrix 比較

Covariance matrices 在訓練前(covariance matrices of the original syllable F0)與訓練後(covariance matrices of the normalized syllable F0)分別爲 $\mathbf{R}^{P}_{original}$ 及 $\mathbf{R}^{p}$。可觀察得訓練模型後 covariance 明顯有意義下降。

$$\mathbf{R}^{P}_{original} = \begin{bmatrix} 0.040124 & 0.0053695 & -0.0020751 & -0.00075677 \\ 0.0053695 & 0.018669 & 0.0020441 & -0.0016243 \\ -0.0020751 & 0.0020441 & 0.0037581 & 8.3657\times10^{-5} \\ -0.00075677 & -0.0016243 & 8.3657\times10^{-5} & 0.0011703 \end{bmatrix}$$

$$\mathbf{R}^{p} = \begin{bmatrix} 0.011843 & 0.0021328 & 9.2796\times10^{-5} & -0.00035788 \\ 0.0021328 & 0.0052484 & 0.0011326 & -0.00030867 \\ 9.2796\times10^{-5} & 0.0011326 & 0.0022113 & 0.00033545 \\ -0.00035788 & -0.00030867 & 0.00033545 & 0.00090293 \end{bmatrix}$$

（二）、Duration 模型

1、聲調影響因素(Tone affecting factor)

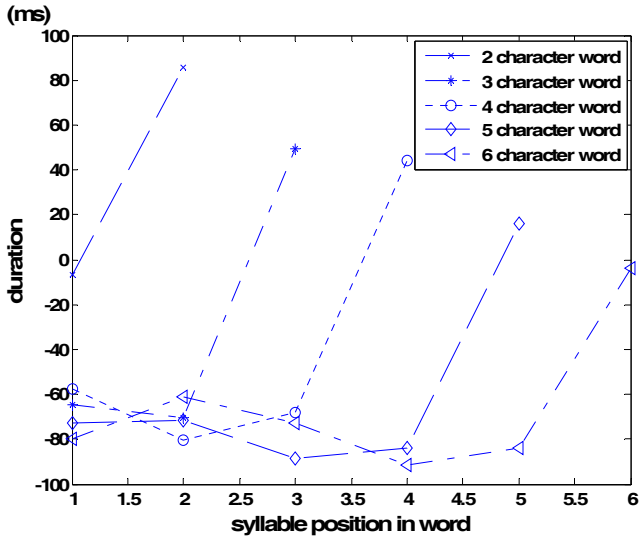
聲調對音節長度影響程度如表三，觀察得五聲音節最短，二聲音節最長。

表三、聲調對音節長度影響

聲調	一聲	二聲	三聲	四聲	五聲
長度	6.9ms	35.0ms	-22.5ms	-17.2ms	-82.2ms

2、詞的位置影響因素(Word-position affecting factor)

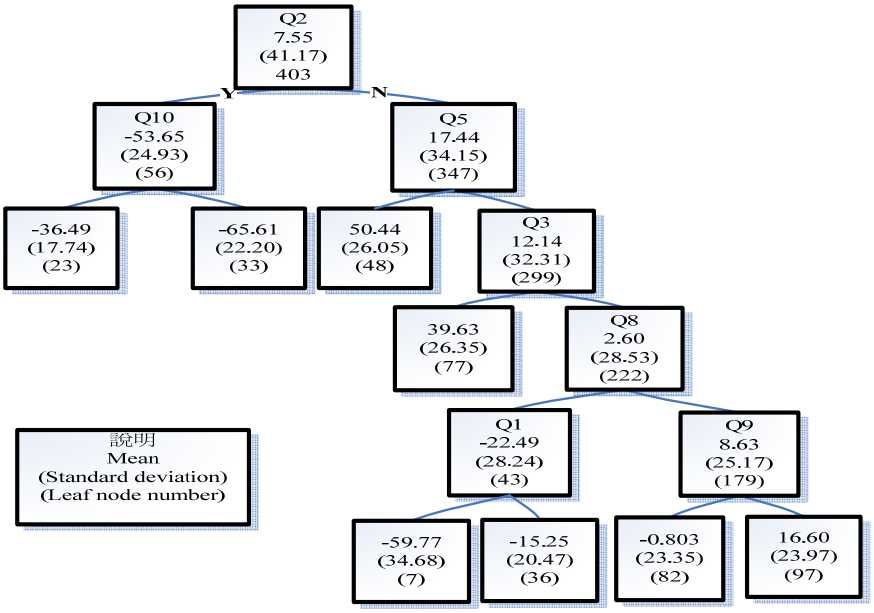
音節在詞的位置對長度影響程度如圖十四，可觀察在詞的字末位置音節長度較長，觀察五字詞、六字詞可得知詞愈長愈容易產生較短的音節。



圖十四、詞的位置影響因素

3、音節影響因素(syllable affecting factor)

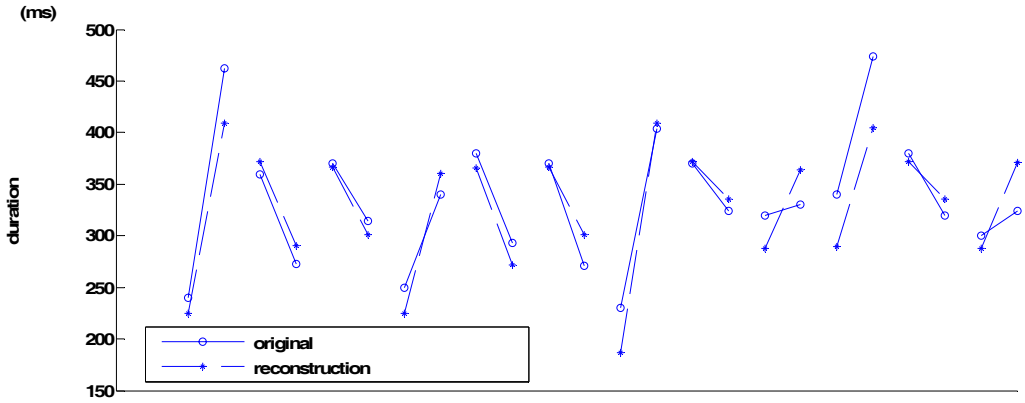
我們使用決策樹觀察音節影響因素對音節長度影響程度，如圖十五。由決策樹可觀察出聲母爲空聲母(Q1)、{b、d、g}(Q2)或韻母類別爲單母音(Q8)這類音節長度較短，而鼻音結束(Q10)、{f,s,sh,x,h}(Q3)及{c,ch,q}(Q5)這類摩擦音的音節長度會較長。



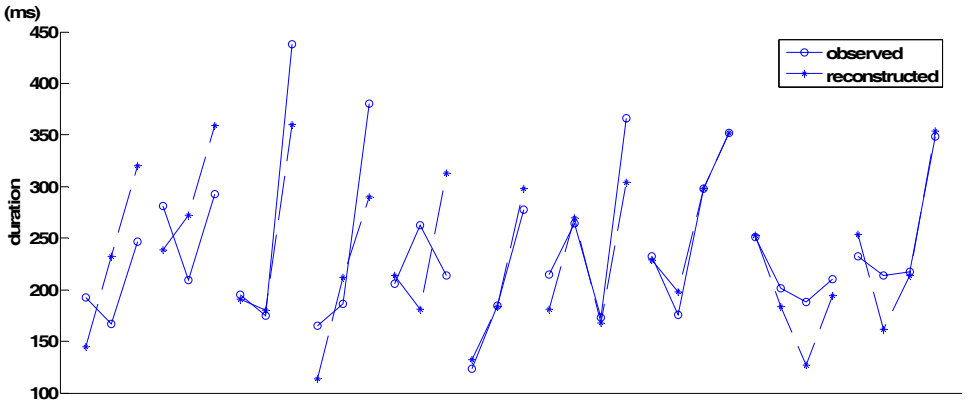
圖十五、以決策樹分析音節影響因素分類結果

4、音節長度預估之範例

Duration 模型預測音節長度與實際長度比較，如圖十六、十七分別爲二字詞及三、四字詞的比較圖。其中 observed 爲音節實際長度，reconstructed 爲音節預測長度，其每線段爲一單詞，大致有不錯的預測結果。



圖十六、二字詞音節實際長度與模型預測長度之比較圖



圖十七、三字詞及四字詞音節實際長度與模型預測長度之比較圖

5、Variance 比較

訓練模型後，以訓練前 variance 爲 9304.5 與訓練後 variance 爲 2494.7。可觀察得使用模型後變異量明顯有意義下降。

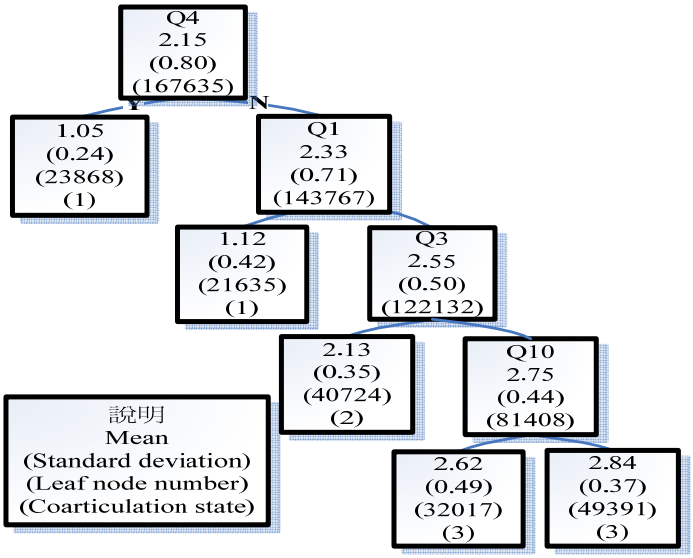
（三）、以決策樹預測 pause 長度

在下面小節中，利用決策樹預測音節間的連音狀態及 pause 長度，其應用於 question set 中的聲母類別和韻母類別分別爲兩音節的間隔區間相鄰的聲母類別及韻母類別。

1、預測音節間之連音狀態

以連音狀態做爲決策樹分類的目標，由 c1 設值爲 1， c2 設值爲 2， c3 設值爲 3。由分類

結果觀察得聲母爲 NULL (Q1)及{m,n,l,r}(Q4)分別爲 mean=1.1173 及 mean = 1.047，意指分佈較集中於c1；而聲母爲{f,s,sh,x,h}(Q3)，此類 mean=2.1341，是指分佈較集中於c2 ，其他類別假設爲c3 。如圖十八所示。



圖十八、以決策樹分析語料庫連音狀態的分類結果

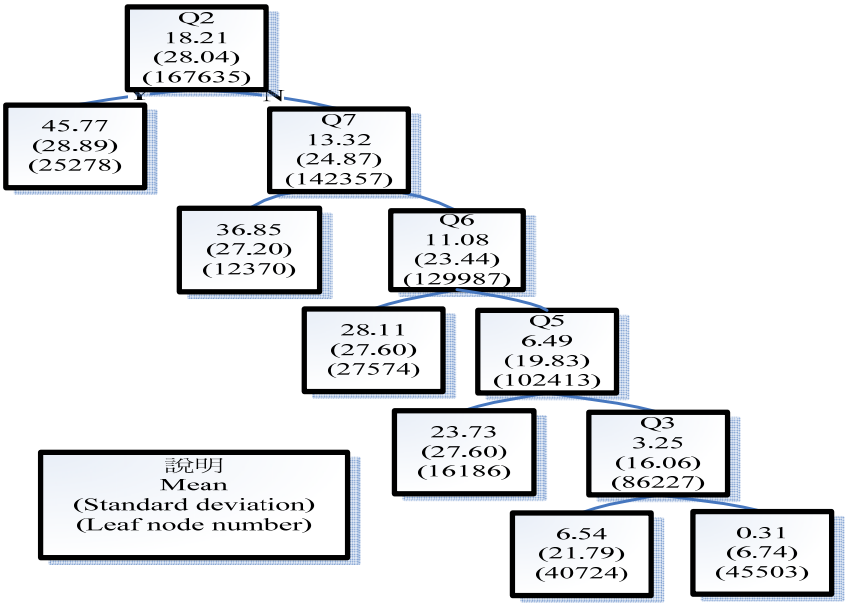
2、預測音節間之 pause 長度

由語料庫中 pause 的長度視爲目標值，其單位爲毫秒(ms)。 由圖十九觀察得 Pause 長度分類明顯與 pause 相鄰的聲母類別有關。將 Pause 長度與聲母類別關係整理如下：
爆破音_不送氣 > 爆破音_送氣 > 塞擦音_不送氣 > 塞擦音_送氣 >摩擦音_清音 >m,n,l,r 及空聲母。

假設預測 pause 長度可直接利用相鄰的聲母類別判斷，由圖十九可觀察得 pause 長度分類的 mean 設爲預測的 pause 長度，整理於表四，其中聲母類別爲 NULL 及{m,n,l,r}的 mean 爲 0.31028，可假設 pause 長度爲 0ms，滿足由圖十八預測此種類別爲連音狀態最明顯的c1 。

表四、pause 長度選取表

類別	聲母	Pause 長度
1	ㄇ、ㄋ、ㄌ、ㄖ、空聲母 (鼻音_濁音)	0ms
2	ㄏ、ㄒ、ㄕ、ㄈ、ㄙ (摩擦音_清音)	7ms
3	ㄅ、ㄉ、ㄍ (爆破音_不送氣)	46ms
4	ㄐ、ㄓ、ㄗ (塞擦音_不送氣)	28ms
5	ㄆ、ㄊ、ㄎ (爆破音_送氣)	37ms
6	ㄑ、ㄔ、ㄘ (塞擦音_送氣)	24ms



圖十九、以決策樹分析語料庫 pause 長度的分類結果

再假設預測 pause 長度與 pause 在詞的位置有相關性，利用 pause 在詞的位置以決策樹分析整理於表五及表六：

表五、二字詞第一個 pause 位置的 pause 長度選取表

類別	1	2	3	4	5	6
(2,1)	0ms	8ms	56ms	34ms	44ms	29ms

表六、三字詞至四字詞 pause 在詞位置的 pause 長度選取表其中(詞長 ，詞中第幾個 pause)。

類別 / 位置	1	2	3	其他類別
(3,1)	0ms	5ms	36ms	23ms
(3,2)	0ms	7ms	45ms	28ms
(4,1)	0ms	6ms	33ms	20ms
(4,2)	0ms	7ms	41ms	26ms
(4,3)	0ms	8ms	47ms	28ms

由表六可觀察得主要影響 pause 長度主要爲聲母類別有 1、2、3 類，且在詞中較後面的 pause 有較長的停頓時間。

五、結論

使用各種影響因素相加，來預測各種韻律訊息，及討論各種影響因素對韻律訊息的分析，由實驗結果證實各種影響因素相加便可預測各種韻律訊息，其各種影響因素分析也

符合語言特性。相信未來應用於語音合成，可明顯自然流暢許多。最後，本文所提出的方法，不因語言不同而有所改變，所以未來可朝向建立一套整合國、台、客語的韻律產生器及韻律分析邁進。

參考文獻

[1] L.-S. Lee, C.-Y. Tseng, and M. Ouh-Young, ‘‘The Synthesis Rule in a Chinese Text-to-Speech System,’’ IEEE Trans. Acoust, Speech, Signal Processing, vol.37, no.9, pp.1309-1319, Sep. 1989.

[2] L.-S. Lee, C.-Y. Tseng, and C.-J. Hsieh, ‘‘Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System,’’ IEEE Trans. Speech and Audio Processing, vol.1, no.3, pp.287-294, July 1993.

[3] S.-H. Chen, S.-H. Hwang, and Y.-R. Wang, ‘‘An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech,’’ IEEE Trans. Speech and Audio Processing, vol.6, no.3, pp.226-239, May 1998.

[4] 黃紹華 ，”中文文句翻語音系統中韻律訊息產生器之研究”，國立交通大學博士論文，民國八十五年六月。

[5] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, ‘‘A Statistics-base Pitch Contour Model for Mandarin Speech,’’ J. Acoust. Soc. AM. 117(2), Feb. 2005, pp. 908-925

[6] C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, ‘‘On the Inter-syllable Coarticulation Effect of Pitch Modeling for Mandarin Speech,’’ Proc. of Interspeech 2005, Lisbon, Portugal, pp. 3269-3272

[7] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, ‘‘A New Duration Modeling Approach for Mandarin Speech,’’ IEEE Trans. On Speech and Audio processing, vol. 11, no. 4, July

以中文十億詞語料庫爲基礎之兩岸詞彙對比研究

洪嘉馡
國立台灣大學語言學研究所
jiafei@gate.sinica.edu.tw

黃居仁
中央研究院語言學研究所
churen@gate.sinica.edu.tw

許銘維
中央研究院語言學研究所
javan@gate.sinica.edu.tw

摘要

近幾年來，由於兩岸交流頻繁，兩岸使用的詞彙，也因此互相影響甚重，語言學界對於漢語詞彙的研究，不論在語音、語義或語用上的探討，發現兩岸對使用相同漢語時的詞彙差異有著微妙性的區別。而兩岸卻又的確是使用漢字體系的書寫系統，只有字形上有可預測的規律性對應。本文在以兩岸使用共同文字系統的原則上，來比對兩岸使用詞彙的特性與現象，以探究與語義對應與演變等相關的議題。

首先，在 Hong and Huang (2006) [1]的對應上，藉以英文 WordNet 爲比對標準，藉由比較北京大學的中文概念辭典(Chinese Concept Dictionary (CCD))與中央研究院語言所的中文詞網(Chinese Wordnet (CWN))兩個 WordNet 中文版所使用的詞彙，探討兩岸對於相同概念詞彙的使用狀況。本文進一步使用中文概念辭典與中文詞網所使用的詞彙，在 Gigaword Corpus 中繁體語料與簡體語料的相對使用率，探究兩岸對於使用相同詞彙，或使用不同詞彙的現象與分佈情形。

關鍵詞：
CCD; CWN; WordNet; Gigaword Corpus; 兩岸詞彙; 詞義; 概念

一、 緒論

兩岸使用詞彙的差異問題，在目前兩岸人民的各種交流中，早就已經呈現出許多無法溝通、理解困難，或是張冠李戴，表達不合宜的錯誤窘境。探討兩岸使用詞彙的差異性時，不僅讓大量使用詞彙的記者們，感受到兩支的差異 (如：華夏經緯網 [2]、南京語言文字網 [3]、廈門日報 [4])，亦成爲漢語詞彙學與詞彙語義學上研究的重要課題(如：王 等

[5]、姚 [6]、)。

以往對於這個議題的研究，不論語言學學者或文字工作者注意到這個問題時，僅能就所觀察到特定詞彙的局部對應，來提出分析與解釋而缺乏全面系統性的研究。本文的研究方法，第一是延續 Hong and Huang [1]、洪 等 [7]的研究方法，先以 WordNet 做詞義概念的判準，比對中文概念辭典與中文詞網裡，概念相同、語義相同的詞彙使用狀況；第二、是以有大量兩岸對比語料的 Gigaword Corpus 作爲實證研究的基礎，驗證中文概念辭典與中文詞網對於相同概念語義的詞彙，使用上，確實有其差異性的存在。這是一個具有完整性、全面性、概括性的整合研究。

又 Miller 認爲他們可以使用同義詞集來表現詞彙概念和描述詞彙的語義內容，所以他們建立了 WordNet，近年來，也有不少研究團隊在處理以 WordNet 爲出發點的不同語言翻譯。

值得一提的是，同屬於漢語詞彙系統的繁體中文系統與簡體中文系統，在中央研究院語言所與北京大學計算語言所的研究團隊裡，也針對此議題，做了不少相關的研究，因此，本文想要探討的是，相同概念的漢語詞彙語義，在繁體中文與簡體中文的使用狀況。

另外，對於繁體中文系統與簡體中文系統的對應，我們以 WordNet 當作研究語料的基礎，是爲了可以建立一套符合詞彙知識原則並能運用於英中對譯的系統，如此一來，即可比對出兩個中文系統，在語言使用上的差異性。

二、 研究動機與目的

自兩岸交流日趨頻繁之後，本屬於同文同種的漢語系統，確有不少知識與信息交流的障礙，造成這樣的原因，莫過於兩岸詞彙使用的差異。相同的詞形，卻代表不同的詞義；或相同的語義，卻有兩種不同的表達詞彙。這種問題，已經讓許多文字工作者費盡心思，試圖來解決這樣的窘境；而語言學者對於這種現象，也試圖從語音、語義、語用等方面著手，希望從各種與語言相關的角度，來探究兩岸詞彙的差異。

在研究議題上，光是觀察到兩岸選擇以不同的詞形來代表相同的語義，如下述兩例，是不夠的。在詞彙語義學研究上，我們必須進一步追究，這些對比的動機，語言的詞彙與詞義演變的動力是否相關，對比有無系統性的解釋等。Chinese GigaWord Corpus 包含了分別來自兩岸的大量語料，包括 7 億餘字中央社資料，及近 4 億字新華社資料，可以看出台灣和大陸對於同一概念而使用不同詞彙的實際狀況與分佈。以 Gigaword Corpus 的語料呈現台灣/ 大陸使用的狀況，如下：

(1) 台灣的「煞 (155/ 65)」、大陸的「非典 (354/ 33504)」
（「Sars (SEVERE ACUTE RESPIRATORY SYNDROME)」「嚴重急性呼吸道綜合症」的翻譯）

(2) 台灣的「計程車 (22670/ 68)」、大陸的「出租車 (422/ 5935)」

如要追究動機與解釋等理論架構問題，當然不能只靠少數觀察到的例子，而必須建立在數量較大的語料庫上，以便做全面深入的分析。以上兩個對比爲例，其重在大陸與台灣

的語料中，都有相當多的變例出現。

三、 WordNet 和中文詞網(CWN)

WordNet 是一個電子詞彙庫的資料庫，是重要語料來源的其中一個語料資料庫，WordNet 的設計靈感源自於近代心理語言學和人類詞彙記憶的計算理論，提供研究者在計算語言學，文本分析和許許多多相關的研究(Miller et al. [8]、Fellaum [9])。在 WordNet 中，名詞、動詞、形容詞、副詞，這四個不同的詞類，分別設計、組合成同義詞集(synsets)的格式，呈現出最基本的詞彙概念，在這當中，以不同的語義關係連結各種不同的同義詞集，串成了 WordNet 的整個架構，也呈現了 WordNet 整個全貌。

自從 Miller et al.和 Fellaum 發展 WordNet 以來，WordNet 就持續不斷地更新版本，目前最新的版本是 WordNet 3.0 版，這些版本間的差異，包括了同義詞集的量和他們的詞彙定義。然而，對於拿 WordNet 來做研究語料的學者，多數還是以 WordNet 1.6 版爲最多，因爲這個版本是目前最多計算語言學學者使用的。在 WordNet 1.6 版裡，有將近 100,000 的同義詞集。

我們知道，雙語領域分類，可以增加我們各種領域詞彙庫的發展，同樣的，在上一段的內容，我們也提到關於以 WordNet 爲基礎，發展出繁體中文系統(Chinese Wordnet, CWN)與簡體中文系統的對譯，我們使用雙語詞網，作爲詞彙知識資料庫來實現、支持我們在詞彙概念上的研究。

在中英雙語詞網中，每一個英文的同義詞集，我們都會給予三個最適合且對等中文翻譯，而這些翻譯，如果不屬於真正的同義詞，我們也會標註他們的語義關係(Huang et al. [10])，又這些雙語詞網，也在中研院語言所詞網小組團隊的發展，將每一個同義詞集都與 SUMO 概念節點連結，進而開發出 Academia Sinica Bilingual Ontological Wordnet (Sinica BOW, Huang and Chang, [11])。當我們無法直接取得中英相對應的詞彙，我們在雙語詞網的資料庫裡，可以利用這些語義關係，進而發展並預測領域分類。

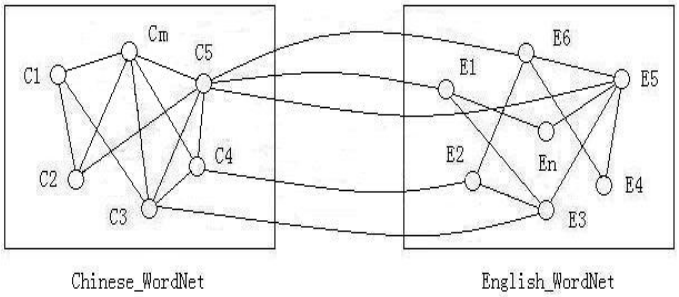
四、 WordNet 和中文概念辭典(CCD)

CCD，中文概念辭典(Chinese Concept Dictionary)，是一個中英雙語的詞網，整個架構發展也是來自於 WordNet [12]、[13]、[14]。在 CCD 的發展手冊裡記載，研究團隊描述這些詞義的首要條件，是不可以破壞原本 WordNet 對於同義詞集定義概念與其語義關係的架構。另一方面，CCD 的研究團隊考量到可以存在許多在中文與英文的不同描述架構，所以，他們不止表現對中文詞彙內涵的表達，也發展了中文詞彙語義與概念的關係性，以利於強調中文的特質。

CCD 的研究團隊專注在整個 CCD 的架構，提出同一概念的同義詞集的定義，其所呈現的概念、定義和概念網的上下位語義關係，每一個同義詞集都有其基本關係，彼此之間亦有語義關係的存在。至於 CCD 的邏輯推演原則在語義網上的呈現，是運用到數學的形式而來的，是可以幫助研究者在中文語義分析上的使用。

自從 2000/09 開始，北京大學計算語言學研究所就已經開始著手以 WordNet 爲基準，研究 CCD，並建立一個中英雙語的詞網，一個可以提供各種不同研究的詞網，如機器翻譯(MT)，訊息擷取(IE)…等等。

基於 WordNet 英文概念與 CCD 中文概念是屬於兩個不同知識背景，也因此 CCD 中，他們兩者間的相互關係與概念，是非常複雜、繁瑣的。CCD 包括了大量且繁雜的成對、成組的小網絡，大致上，差不多有 10^5 的概念節點和 10^6 的成組小網絡的概念關係，他們的關係，呈現如下圖：



圖一、WordNet 小網絡中複雜的關係結構

五、 文獻探討

對於兩岸詞彙對比的探討，過去的研究，多半著重在語言特徵的區別。如列舉語音方面、詞彙方面的對比(南京語言文字網 [3])；或以語音、詞彙、語法及表達方式等方面來分析語言差異的現象 (如：王 等[5]、姚 [6]、許 [15]、戴 [16])。

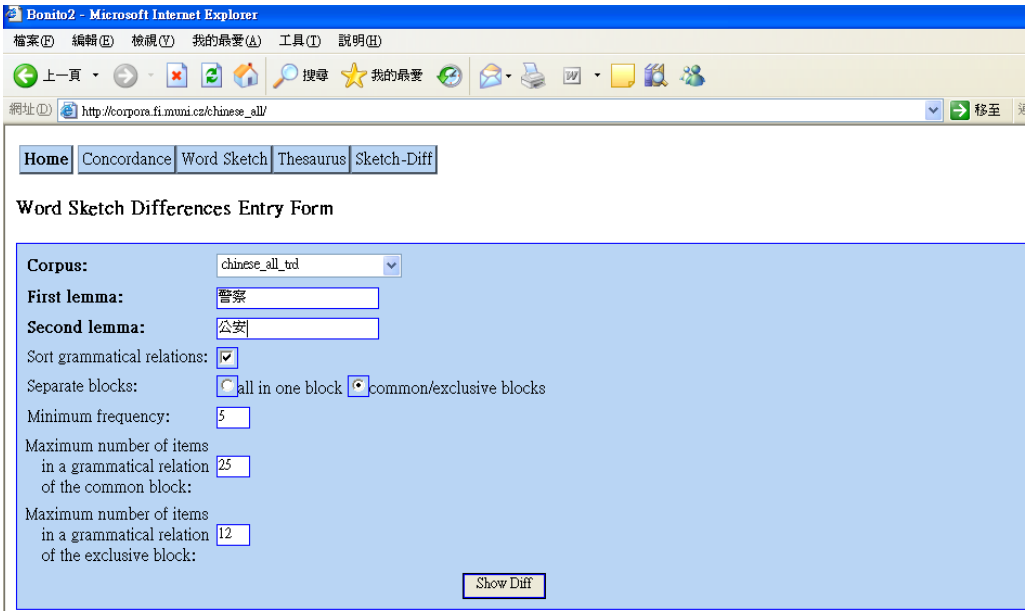
近年來，對於兩岸詞彙對比的研究，比較新的研究方法，則是以 WordNet 爲基礎，取兩岸語料庫資料作比較，進而分析兩岸詞彙的對比(如：Hong and Huang [1])；或以 Gigaword Corpus 爲基礎，探索兩岸對於漢語詞彙在使用上的差異現象，例如：相關共現詞彙(collocation)的差異、台灣或大陸獨用的差異、特定語境下的特殊用法的差異、語言使用習慣的差異等等(如：洪 等 [7])。

六、 研究方法

本研究以英文的 WordNet、繁體中文系統的中文詞網(CWN)、以及簡體中文系統的中文概念辭典(CCD)等三大資料庫爲主，對於繁體中文系統的英中對譯與簡體中文系統的英中對譯，我們先進行比對，試圖在比對中，尋找出兩者之間的差別與使用分佈。

相同的概念，本歸屬於一個同義詞集，但因兩岸在詞彙使用上的差異，而有所不同，儘管如此，仍舊有一些兩岸使用相同的詞彙來表達相同的概念語義。本文中將從繁體中文系統與簡體中文系統的英中對譯資料裡，集中探究同一個同義詞集，在兩岸使用的詞彙是完全相同、完全不同的狀況。然後，再將這些完全相同、完全不同的詞彙，以 Gigaword

Corpus 爲基礎，分析這些詞彙在這個語料庫裡，所呈現出兩岸使用的狀況。
接著，本文再以語料庫爲研究出發點，是以約十一億字的 Chinese Gigaword Corpus 爲主要語料來源，以中文詞彙素描爲搜尋語料工具 [17]、[18]、[19]。Chinese Gigaword Corpus 包含了分別來自兩岸的大量語料，包括近四億字新華社資料，及七億餘字中央社資料。因此，可以提供兩岸詞彙差異的大量詞彙證據。
而中文詞彙素描則提供了語料庫爲本，詞彙差異比較的工具。可以看出兩岸對於同一概念而使用不同詞彙的實際狀況與分佈，也可以看出同一語義詞彙在兩岸的實際語料中，所呈現的相同點與差異性。我們主要利用中文詞彙素描中詞彙素描差異(word sketch difference)的功能。詞彙描素差異的實際介面畫面如圖二：



圖二、中文詞彙素描 Engine 下的詞彙描素對比

在此功能下，我們將已經比對過 CCD 與 CWN 對應不同對譯的詞彙，進一步探究兩詞彙的使用狀況與分佈。在本文中，主要是以比對兩岸詞彙詞頻爲主，倘若在 CCD 與 CWN 的對應中，確實是相同語義，卻在兩岸使用完全相同或完全不同的詞彙，那麼其各自使用的詞彙，在 Gigaword Corpus 裡繁體語料與簡體語料交叉比對後所得的詞頻，也應當會有近似的分佈現象，藉此數據，不但可以證明 CCD 和 CWN 在英中對譯上，繁體中文系統與簡體中文系統，是有差別的，也可以證明，確實有兩岸使用不同詞彙來表達相同概念與義的用法，進而了解兩岸詞彙的實際現象，以進行本研究的分析。

七、 語料分析

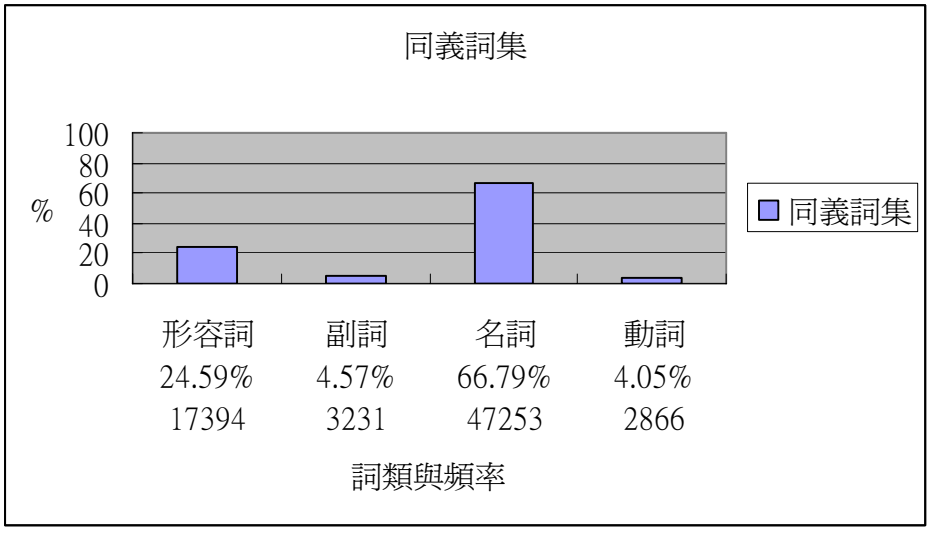
繁體中文系統的英中對譯(CWN)與簡體中文系統的英中對譯(CCD)，依不同詞類，區分成：名詞、動詞、形容詞和副詞四大類來進行對比，以 WordNet 爲主，檢測在同一個 Synset 中，繁體中文系統的對譯詞彙和簡體中文系統的對譯詞彙，然後再進行比對。

在四大詞類中，我們可以清楚得知，在同一個 Synset 中，繁體中文系統，可能有多個相對應的對譯詞彙，同樣地，簡體中文系統也可能有個相對應的對譯詞彙。在這些對譯詞彙裡，又有可能是兩邊使用的對譯詞彙完全一樣，稱之「完全相同」；如果，兩邊使用的對譯詞彙，沒有一個相同的，稱之「完全不同」，也就是「真正不同」；或者，只有使用其中一個或一個以上對譯詞彙，這個狀況，稱之「部份相同」，而在「部份相同」的對譯詞彙，如果兩邊的對譯詞彙使用的詞首相同，稱之「詞首相同」，如果只是使用到相同的字，則稱之「部份字元相同」，如：

Synset	CCD 對譯詞彙	CWN 對譯詞彙	
bookshelf	書架、書櫃、書櫥	書架、書櫃、書櫥	完全相同
lay-off	下崗	解雇	完全不同
immediately	立即	立刻	詞首相同
according	據報	根據	部份字元相同

表一、CCD 和 CWN 對譯的各種分佈狀況

對於 CWN 與 CCD 的對比，總共有 70744 個 Synset 是對譯相同的，分屬於形容詞、副詞、名詞和動詞這四個詞類當中，其中，以名詞在 CWN 與 CCD 的完全相同對譯中，所佔比例最高，有 66.79%；反之，動詞所佔比例最低，僅有 4.05%，其詳細的分佈情況，如下圖顯示：



圖三、CCD 和 CWN 依不同詞類呈現翻譯完全相同的分佈

而 CWN 與 CCD，使用完全不相同的情況，依各詞類的分佈情形，如下圖顯示：

CCD 和 CWN翻譯不同的分佈狀況

	形容詞	副詞	名詞	動詞	總數
同義詞集數量	17915	3575	66025	12127	*99642*
不同對譯數量	521	344	18772	9261	*28898*
	2.91%	**9.62%**	**28.43%**	**76.37%**	*29.00%*
	最少			最多	

表二、CCD 和 CWN 翻譯不同的分佈狀況

值得一提的是，在 CCD 和 CWN 翻譯不同的分佈狀況裡，很清楚得看到，「動詞」在兩岸的使用狀況，有極大的差異性，然而，在我們使用漢語時，常會以同類近義詞或語義相近相關詞來取代原本的詞彙，在此，我們又更進一步，更仔細地分析，試圖將每一個詞類中，有這樣的使用情形分類出來，以得到真正兩岸使用不同詞彙的現象。
我們以「詞首相同」、「部份字元相同」和「真正不同」，這三大類別來分析 CCD 和 CWN 在每一個詞類中，翻譯不同的分佈狀況，如下表顯示：

形容詞的分佈

類別	詞首相同	部份字元相同	真正不同	總數
同義詞集	169	175	177	521
	344			
百分比	32.44%	33.59%	33.97%	100%
	66.03%			

名詞的分佈

Category	Shared compound head	Shared compound element	None	Total
Synset	7113	7843	3816	18772
	14956			
Percentage	37.89%	41.78%	20.33%	100%
	79.67%			

副詞的分佈

Category	Shared compound head	Shared compound element	None	Total
Synset	77	114	153	344
	191			
Percentage	22.38%	33.14%	44.48%	100%
	55.52%			

動詞的分佈

類別	詞首相同	部份字元相同	真正不同	總數
同義詞集	3269	3316	2676	9261
	6586			
百分比	35.30%	35.80%	28.90%	100%
	71.10%			

表三、CCD 和 CWN 在每個詞類中，翻譯不同的分佈狀況

從表一到表三，我們可以清楚知道對於每個詞類，CCD 和 CWN 在翻譯不同的詞彙裡，仍然有些算是語義相近的相關詞彙，扣除這些相關詞彙後，兩岸詞彙在使用上的真正不同，就可清楚呈現。至於，上文中，所提及關於「動詞」是兩岸詞彙中，使用最多不同的狀況，我們從表三的分析得知，在動詞的使用上，因爲較常出現同類近義詞或語義相近相關詞來取代原本的詞彙的狀況，所以「詞首相同」和「相同字元」這兩類佔了很大

的因素，在 9261 個詞彙裡，就有 6586 個詞彙，大約是 71.10%，其真正兩岸對於動詞的不同使用，則有 2676 個詞彙，大約是 28.90%。
我們將以圖三和表三中，四個詞類裡，使用完全相同的詞彙與真正不同的詞彙，藉由 Gigaword Corpus 來分析兩岸人民對於詞彙使用的實際狀況。

八、 實驗設計與分析

本研究將繁體中文系統與簡體中文系統的英中對譯資料，經比對過後，設計了一個查詢介面，以利於相關研究學者，對於兩岸詞彙的實際使用，進行搜尋，這個搜尋介面，主要是以 WordNet 的英文同義詞集爲基礎，查詢相同概念、相同語義在兩岸對譯的詞彙情況，其介面首頁呈現如下圖四，另外，在這個搜尋介面中，目前完成的查詢功能有兩岸翻譯完全相同、完全不相同與部份相同，如下圖五～圖七所示：

圖四、以 WordNet 爲中心的兩岸詞彙對譯查詢介面

圖五、兩岸翻譯完全相同呈現在兩岸詞彙對譯查詢介面上

該詞彙在兩岸翻譯爲 完全不相同 ，詳細查詢內容如下:

查詢詞彙	bow_id	釋義	cwn翻譯	ccd翻譯
book	00457832V		定下, 雇用,	預訂,

圖六、兩岸翻譯完全不相同呈現在兩岸詞彙對譯查詢介面上

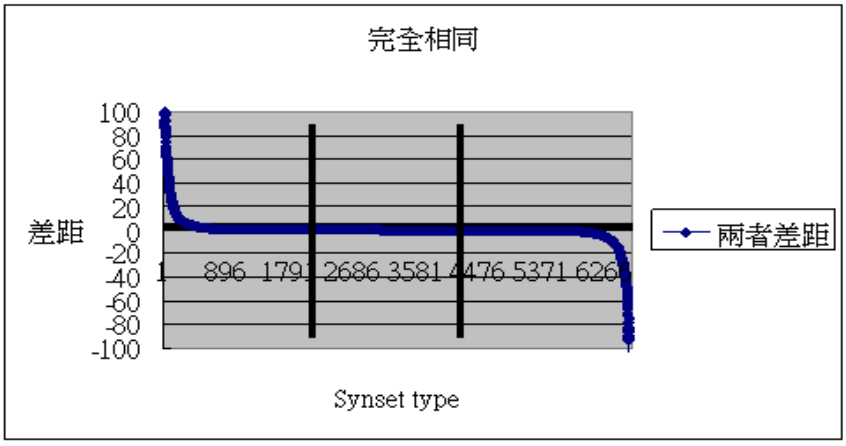
該詞彙在兩岸翻譯爲 部份相同 ，詳細查詢內容如下:

查詢詞彙	bow_id	譯義	相同	不同	
				cwn	ccd
negatively	00003199R		消極,		否定,

圖七、兩岸翻譯部份相同呈現在兩岸詞彙對譯查詢介面上

根據 CCD 和 CWN 以 WordNet 爲中心，比對出來的對譯有完全相同、完全不同與部份相同等三大類，在此，本研究僅就前兩類的資料，再以 Gigaword Corpus 爲依據，檢測實際語料中所呈現的狀況。

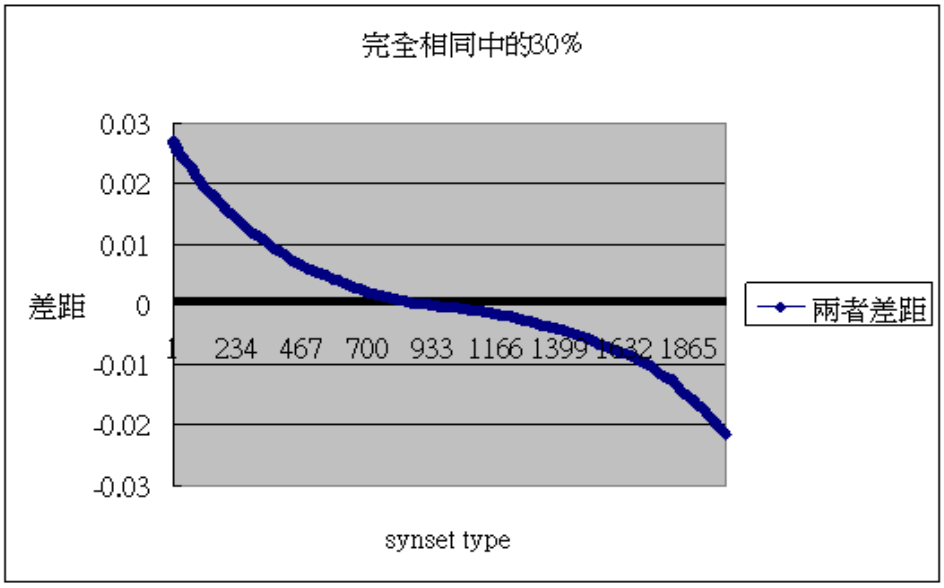
首先，我們先取兩岸使用詞彙「完全相同」的資料，檢測這些資料在 Gigaword Corpus 中，分屬在繁體中文與簡體中文的使用頻率，再計算每個詞彙的頻率在繁體中文資料與簡體中文資料裡所佔的比例，如此，即可知道每一個詞彙，在繁體中文與簡體中文裡，出現和使用的情形。理想的想法，如果一個詞彙在兩岸使用的情況是非常接近的，其兩者詞頻比例的差距，應該是非常小的。我們試著將同一詞彙在兩岸使用的詞頻比例相減，以便檢測這些使用上完全相同的詞彙，又因其差距的數值過小，所以我們以放大 100000 倍後的數值來呈現，其分佈情形，如下圖所示：



圖八、兩岸使用完全相同詞彙的分佈情況

從圖八來看，曲線彎曲的前後兩端，代表兩者的差距較大，靠左邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠右邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。在使用完全相同詞彙中，仍有些使用差異明顯的語料，這是值得我們深入探討的議題。在 6637 筆兩岸使用完全相同詞彙中，在 Gigaword Corpus 中中央社/新華社語料中使用分佈差異的平均值爲 0.0143%。使用差異分佈在這個平均值內的詞彙，共計有 5880 筆。換句話說，兩岸使用完全相同的詞彙裡，使用狀況較爲相近的有 5880 筆，使用狀況較爲不相同的仍有 757 筆。在中央社與新華社語料中分別有 354 筆和 403 筆。這 757 筆資料是「同中有異」的詞語，值得我們將來進一步分析。

在 6637 個兩岸使用完全相同詞彙中，圖八中雖然顯示其頻率差距幾乎是零。但是，如果我們由差距最小的第 3076 個詞，依前後各取 30%的（就是第 2153 個詞彙取到第 4144 個詞彙），將差距在放大呈現如圖九。從圖中的差距數值顯示，是非常非常小的。一方面證明兩岸使用這些詞彙的情形，是非常非常接近的。另一方面，當間距放大後，我們看到差異分佈呈平滑的 S 字型，這也與預期中自然語料分佈的狀況相符。

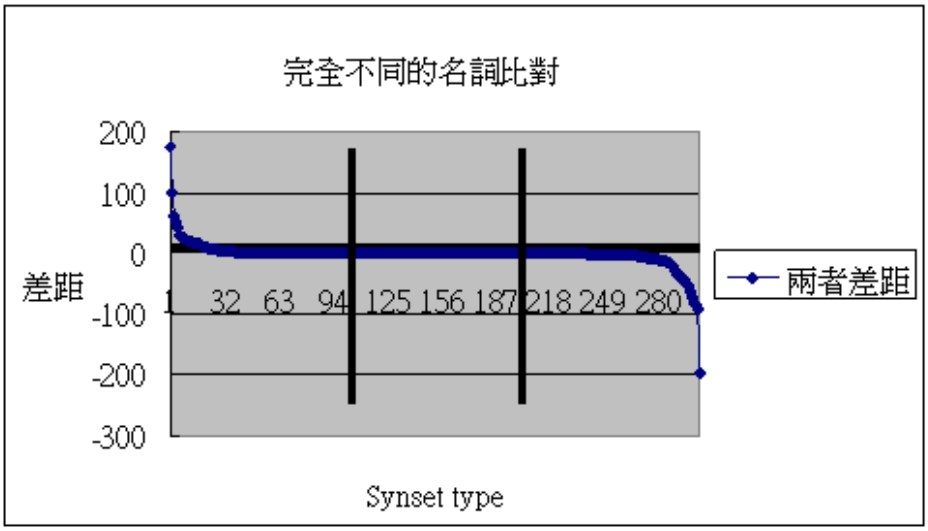


圖九、兩岸使用完全相同詞彙中，差距最小的 30%的分佈情況

詞彙	詞頻		附註
	CNA (繁體中文)	XIN (簡體中文)	
酒桶	32 (0.157 μ)	20 (0.155 μ)	使用狀況非常接近
絲瓜	1380 (6.78 μ)	96 (0.748 μ)	使用狀況有差異
柳葉刀 雙刃小刀	2 (0.00982 μ)	273 (2.13 μ)	使用狀況有顯著差異

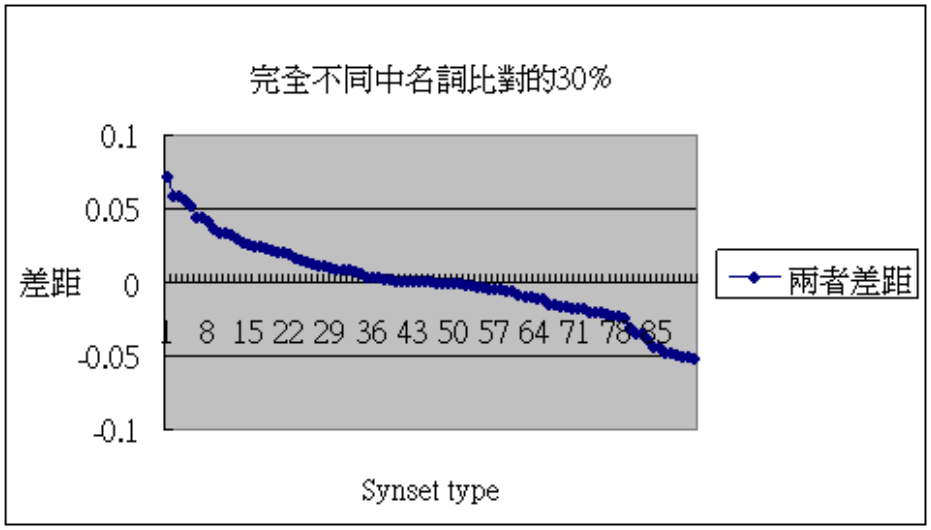
表四、兩岸使用完全相同詞彙的分佈狀況示例

接著，我們以相同的實驗方法與步驟來檢測兩岸使用完全不同的詞彙，檢測這些資料在 Gigaword Corpus 中呈現的分佈，在此，本文僅取數量較大的名詞和動詞來做比對，並且擷取語料的原則是 CCD 的使用詞彙在 XIN 的詞頻大於 CNA 的詞頻；反之，CWN 的使用詞彙在 CNA 的詞頻大於 XIN 的詞頻，其分佈情形，如下圖所示：



圖十、兩岸使用完全不同的名詞詞彙的分佈情況

在兩岸使用完全不同的名詞詞彙裡，共計有 302 筆資料，靠右邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠左邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。我們一樣採取兩者差距最小的 30%來檢測，其計有 91 筆資料，從圖十一的差距數值來看，可以證明，這些不同的詞彙，在所屬的語言系統裡，其使用狀況的獨特性，換言之，同一個詞彙，在繁體中文系統裡，使用的頻率較高，在簡體中文系統裡，使用的頻率較低，反之亦然，而呈現相對之分佈狀態，這樣的情形，在圖十一的差距數值和表五例子中得到驗證。

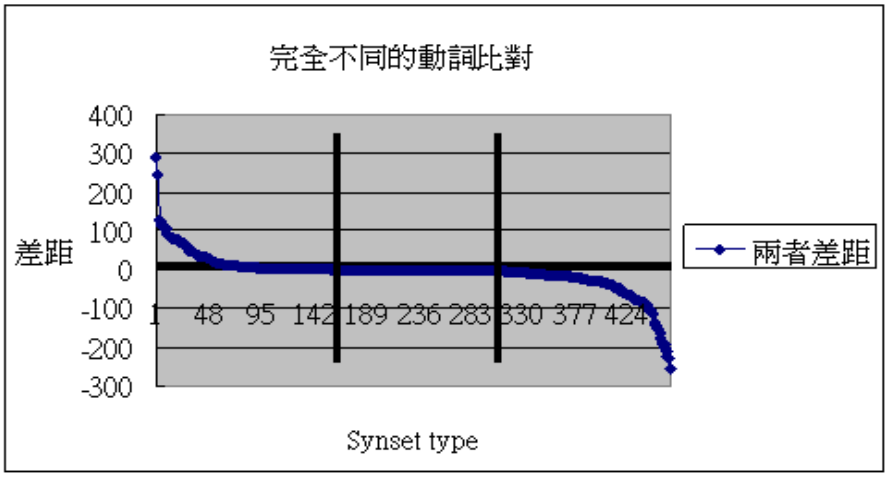


圖十一、兩岸使用完全不同的名詞詞彙中，差距最小的 30%的分佈情況

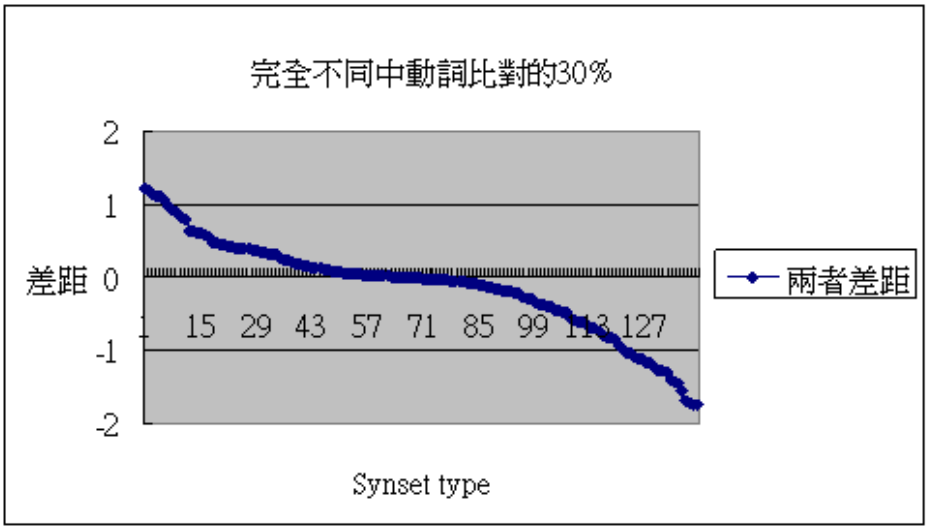
詞彙		詞頻				附註
CCD	CWN	CCD		CWN		
		XIN	CNA	CNA	XIN	
風帽	頭罩	10 (0.0779 μ)	2 (0.0098 μ)	101 (0.4963 μ)	37 (0.2882 μ)	使用狀況對比明確
雙休日	週末	1383 (10.7736 μ)	25 (0.1228 μ)	17194 (84.4908 μ)	6105 (47.558 μ)	使用對比較不明確
屏幕 CRT 屏幕	映像管	3086 (24.04 μ)	118 (0.5798 μ)	427 (2.0983 μ)	1 (0.0078 μ)	使用對比較不明確

表五、兩岸使用完全不同的名詞詞彙的分佈狀況示例

至於在兩岸使用完全不同的動詞詞彙裡，共計有 461 筆資料，靠右邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠左邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。我們採取一樣的方式來進行檢測，其 30%的資料，共計有 140 筆，從圖十二、圖十三的差距數值來看，確實可以證明這些使用不同的動詞詞彙，在繁體中文系統與簡體中文系統，有其使用狀況的對比性。



圖十二、兩岸使用完全不同的動詞詞彙的分佈情況



圖十三、兩岸使用完全不同的動詞詞彙中，差距最小的 30%的分佈情況

兩岸使用完全相同詞彙的平均值是 0.0143%，那麼，理論上，兩岸使用不同詞彙的比例，應該大於這個平均值，倘若小於這個平均值，則有可能是兩岸使用相同概念詞彙時，產生混用的現象。在使用不同的名詞詞彙中，以大陸獨有詞的比例來排序，發現有 174 筆資料小於這個平均值；以台灣獨有詞的比例來排序，則有 168 筆資料小於這麼平均值；在使用不同的動詞詞彙中，以大陸獨有詞的比例來排序，發現有 320 筆資料小於這個平均值；以台灣獨有詞的比例來排序，則有 348 筆資料小於這麼平均值。這個數據證實了一個直覺的觀察，就是說兩岸詞彙互相影滲透響的現象日益顯著。以目前的數據看來，台灣的用法影響大陸略強於於大陸的用法影響台灣。我們在最後的版本中將對這些相互滲透影響的詞彙，提出深入的分析與解釋。

九、 結論

兩岸詞彙在使用上的相同、不同或些許的差異，在交流頻繁的情形下，已經日趨明顯，如何區分並釐清兩岸詞彙的個別語義架構，又能在其架構下，增加我們對於漢語詞彙語義系統性演變脈絡的理解，是我們從事語言研究者的不容忽視的議題。本文藉由 WordNet 所發展出的繁體中文系統 CWN 與簡體中文系統 CCD，進行兩岸詞彙的比對，再將比對過後的詞彙，以收集實際大量語料的 Gigaword Corpus 爲基礎，檢測兩岸在詞彙上使用的現象與分佈狀況；亦可由 Gigaword Corpus 所呈現的狀況，證明繁體中文系統 CWN 與簡體中文系統 CCD 在比對上的正確度與可靠性；也證實了 CCD 和 CWN 將兩岸詞彙對比的使用狀況質化呈現，而 Gigaword Corpus 則是以實際語料來驗證兩岸詞彙對比的使用狀況量化呈現。我們更進一步發現了兩岸共用詞彙有「同中有異」的現象，而對比詞彙也產生了互相滲透影響的現象。值得更深入探討研究。

參考文獻

[1] Hong Jia-Fei, Chu-Ren Huang. 2006. WordNet Based Comparison of Language Variation —A study based on CCD and CWN. Presented at Global WordNet (GWC-06). Pp. 61-68. January 22-26. Jeju Island, Korea.

[2] 華夏經緯網，2004，<趣談海峽兩岸詞彙差異>，http://www.huaxia.com/wh/zsc/00162895.html.

[3] 南京語言文字網，2004，<兩岸普通話大同中有小異>，http://njyw.njenet.net.cn/news/shownews.asp?newsid=367.

[4] 廈門日報，2004，<趣談兩岸詞彙差異>，http://www.csnn.com.cn/csnn0401/ca213433.htm.

[5] 王鐵昆、李行健，1996，<兩岸詞彙比較研究管見>，《華文世界》，第 81 期，台北。

[6] 姚榮松 1997. 《論兩岸詞彙差異中的反向拉力》，第五屆世界華語文教學研討會，世界華語文教育協進會主辦，1997.12 月 27-30 日，台北劍潭。

[7] 洪嘉馡, 黃居仁, 馬偉雲. 2006. 語料庫爲本的兩岸對應詞彙發掘. 第七屆漢語詞彙語義學研討會(CLSW-7). 交通大學. 2006.5.22-24.

[8] Miller G. A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. 1993. “Introduction to WordNet: An On-line Lexical Database,” In Proceedings of the fifteenth International Joint Conference on Artificial Intelligence.

[9] Fellbaum C. 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.

[10] Huang, Chu-Ren. Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy. 2003. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. Languages and Linguistics. 4.3. (2003)509-532.

[11] Huang, Chu-Ren, and Ru-Yng Chang. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO”. Presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May (2004).

[12] 于江生, 俞士汶。2004。中文概念詞典的結構。中文信息學報(Journal of Chinese Information Processing), vol. 16 No. 4 (2004)12-21。

[13] 于江生, 劉揚, 俞士汶。2003。中文概念詞典規格說明。Journal of Chinese language and Computing, 13(2) 177-194。

[14] 劉揚, 俞士汶, 于江生。2003。CCD 語義知識庫的構造研究。2003 中國計算機大會(CNCC’2003)。

[15] 許斐絢，1999，《台灣當代國語新詞探微》，國立台灣師範大學華與文教學研究所碩士論文，台北。

[16] 戴凱峰，1996，《從語言學的觀點探討台灣與北京國語間之差異》[A Linguistic Study of Taiwan and Beijing Mandarin]，政治作戰學校外國與文學系碩士論文，台北。

[17] Chinese Word Sketch Engine: http://wordsketch.ling.sinica.edu.tw/

[18] Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, and David Tugwell. 2005. 中文詞彙素描. ASIALEX 2005: Words in Asian Cultural Context. June 1-3. Singapore.

[19] Lexical Data Consortium. 2005. Chinese Gigaword Corpus 2.5.: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T14

VOT productions of word-initial stops in Mandarin and English: A cross-language study

Li-mei Chen*
Kuan-Yi Chao**
Jui-Feng Peng***

Department of Foreign Languages and Literature
National Cheng Kung University

* Associate Professor, leemay@mail.ncku.edu.tw
** Lecturer, chaoky@mail.ncku.edu.tw
*** Postgraduate student, tn670124@mail.tn.edu.tw

Abstract

Voice Onset Time (VOT) is considered as one of the best methods for examining the timing of voicing in stop consonants and has been applied in the study of many languages. The present study is designed to examine VOT production for phonetically voiceless stops in Mandarin and English by native Chinese speakers. Thirty-six Taiwanese Chinese speakers recruited from National Cheng Kung University participated in this study. The results indicate the following. 1) Based on the three universal categories proposed by Lisker and Abramson (1964), for phonetically voiceless stops, Mandarin and English occupy the same place along the VOT continuum. 2) The mean VOT value for the apical stop /t/ is slightly lower than the mean value for the labial stop /p/. This does not conform to the general consensus, which states that the further back the place of articulation the longer the VOT. Very similar findings were also observed in previous studies. 3) The difference between the mean VOT values of the English /p/ and /t/ produced by Chinese speakers was subtle, while it reached significance for native English speakers. This suggests that a first language could be a crucial factor in L2 production. Future studies might examine variations in L2 production both for the same persons over time and for different speakers.

Keywords: voice onset time (VOT), voiceless stops, place of articulation

1. Introduction

Voicing contrast in stops has been discussed in phonetics and phonology for the past few decades. Beginning with Lisker and Abramson (1964), in their well-known cross-language study, voice onset time (VOT) has been widely used to differentiate stop categories across languages. Since then, VOT has come to be regarded as one of the best acoustic cues for discriminating three general stop categories, especially in word-initial position. In contrast with the considerable number of studies investigating stop voicing contrast in a variety of

languages, only a few have examined Mandarin word-initial stops, not to mention comparing VOT patterns in Mandarin and English. Therefore, the purpose of this present study is threefold. First, it is intended to provide information for a general VOT pattern of Mandarin word-initial stops. By analyzing VOTs in stop consonants, linguists have concluded that for most languages, VOT values get longer as the place of articulation moves backward (Lisker & Abramson, 1964; Cho & Ladefoged, 1999; Gósy, 2001). However, there are some exceptions, such as Mandarin, which does not follow the general rule (Lisker & Abramson, 1964; Cho & Ladefoged, 1999; Chao, Khattab & Chen, 2006). The second purpose is to explore the possible effects of this phenomenon. Vowel context is also examined to determine whether there is a correlation between VOT and subsequent vowels. Moreover, to date no study has focused on comparing the in-depth differences between Mandarin and English, except for Chao et al. (2006) who pinpoints the existence of subtle differences between the two languages. Thus, the third aim is to compare VOT patterns of the two languages and observe L2 production (i.e. English production) by native Chinese speakers.

2. Literature review

2.1 Voice onset time

Lisker and Abramson (1964) conducted a cross-language investigation of word-initial stops in 11 languages and define voice onset time as the temporal interval from the release burst of the stop consonant to the onset of the first formant (F1) frequency that reflects glottal vibration. Following their study, VOT has been widely used to examine voicing contrast in stops in many languages (Keating, Linker, and Huffman, 1983; Rochet & Fei, 1991; Cho and Ladefoged, 1999; Gósy, 2000; Khattab, 2000; Zheng & Li, 2005; Riney, Takagi, Ota, and Uchida, 2006). In addition to investigating phonetic characteristics of voiced and voiceless stops in various languages, some researchers have studied VOT with respect to place of articulation, speaking rate, bilingual language learners, and vowel environment (Kewley-Port, Pisoni, and Studdert-Kennedy, 1983; Port and Rotunno 1979; Kessinger and Blumstein 1997; Benkí, 2001; Kehoe, Lleó, and Rakow, 2004). Thus, VOT is one of the main acoustic cues used to measure the timing of voicing in stops.

Although VOT is now used across the world as a linguistic cue, some researchers, however, challenge its role and importance as a reliable measure for separating phonemic categories. In their study examining voicing contrast among French-English bilinguals, Caramazza, Yeni-Komshian, Zurif, and Carbone (1973) argue that voice onset time is ineffective at differentiating stop categories. Bohn and Flege (1993) also question its importance to the perception of stop voicing. Docherty (1992) indicates that VOT narrowly concentrates on word-initial stops. Moreover, Klatt (1975) even suggests five other acoustic cues that are equally important to voice onset time: that is, low frequency energy in subsequent vowels, burst loudness, fundamental frequency, pre-voicing, and segmental duration. Even if VOT does have limitations, it is still regarded as one of the most important acoustic parameters for distinguishing voicing contrast, especially for word-initial stops.

2.2 VOT category

In Lisker and Abramson's 1964 study, all stops are classified into three groups depending on the number of stop categories in each language. VOT ranges for the three stop categories are -125 to -75ms, 0 to +25ms, and +60 to +100ms. Cho and Ladefoged (1999) also provide VOT ranges for occlusives, particularly in voiceless aspirated and unaspirated stops. Rather than three categories, they distinguish four: unaspirated, slightly aspirated, aspirated, and highly aspirated. The approximate mean VOT values for each category are 30 ms, 50 ms, 90 ms, and over 90 ms, respectively. In agreement with Lisker and Abramson's (1964) categorization, on the basis of Cho and Ladefoged's (1999) categorization, stops in Mandarin and English are found to occupy the same place along the VOT continuum, whereas stops in the two languages do not belong to the same range along the continuum, especially for voiceless aspirated occlusives. Chao, Khattab, and Chen's (2006) findings confirm Cho and Ladefoge's classification and reveal that for voiceless aspirated stops, Mandarin falls into the 'highly aspirated' region while English belongs to 'highly aspirated' category. A comparison of the different stop categories in Mandarin and English is given in section 2.4, below.

2.3 Effect on VOT

2.3.1 Place of articulation

Some researchers have reported a significant link between place of articulation and voice onset time. Cho & Ladefoged (1999) propose some possible relations including 1) the further back the closure, the longer the VOT; 2) the more extended the contact area, the longer the VOT; and 3) the faster the movement of the articulator, the shorter the VOT. Of these three suggested links, the present study focuses on the first in connection with Mandarin. In addition to this first principle, it may be stated that the velar stop /k/ has the longest VOT duration and bilabial stop /p/ the shortest, with the alveolar stop /t/ in between the two (Lisker & Abramson, 1964). Factors used to explain why VOT is longer when articulation takes place nearer the back of the mouth include aerodynamics, articulatory movement velocity, and differences in the mass of the articulators (Cho & Ladefoged, 1999).

The size of the supraglottal cavity behind the constricted points should be taken into consideration when considering the impact of aerodynamics. The cavity behind the velar stop has a smaller volume than that behind the alveolar and bilabial stops. In other words, the velar stop is under greater pressure when airflow is released; therefore, it might take longer to produce a velar stop, and the VOT value for the velar stop might be longer than either the alveolar or the bilabial stop. As for articulatory movement velocity, Cho and Ladefoged (1999) claim that the tip of the tongue and the lips move faster than the back of the tongue; moreover, the tongue tip moves faster than the lower lip. This may explain why in many languages velar stops have longer VOT than labial and alveolar stops. However, articulatory movement velocity does not affect alveolar and bilabial stops in this way in all languages, which implies that other factors are involved. In reference to the extent of articulatory contact area, Cho and Ladefoged (1999: 211) claim that, "In general, stops with a

more extended articulatory contact have a longer VOT." In summary, it is indubitable that velar stops have longer VOT than the two other stops. However, no final conclusion may be reached in the case of labial and alveolar stops.

Although there is general agreement that the further back the place of articulation, the longer the VOT, there are still some exceptions. Lisker and Abramson's (1964) study reports that unaspirated stops in Tamil and aspirated stops in Cantonese and Eastern Armenian do not follow this rule. It is found that the VOT of alveolar /t/ is shorter than bilabial stop /p/, but the velar stop /k/ still has the longest VOT. Studies by Rochet and Fei (1991) and Chao et al. (2006) arrive at similar results. Investigating Mandarin Chinese, they conclude that the VOT duration for /t/ does not confirm the predictions; on the contrary, it is shorter than the VOT for /p/. The cause of this phenomenon is still unknown.

2.3.2 Vowel context

How vowels influence the VOT of preceding stops is still an open question. Lisker and Abramson (1967) propose that following vowels have no significant influence on VOTs, while other researchers apply similar research methods, but more systematically, and find that VOTs are longer when followed by tense high vowels (Klatt, 1975; Weismer, 1979). Similar results are obtained in Port's (1979) study, which analyzes VOT for English word-initial stops, and in Gósy's research, which examines Hungarian voiceless plosives. Rochet & Fei (1991) also reach similar findings with respect to Mandarin stops, claiming that "the nature of the vowel had a significant effect on the VOT values of the preceding consonants" (p. 105). In other words, word-initial stops have longer VOT values when followed by either of the high vowels /i/ or /u/ than when followed by the low vowel /a/. This accords with the results presented in Chao et al's (2006) study which examines the Mandarin Chinese of Taiwanese speakers. By contrast, however, Fant (1973) finds that for Swedish aspirated stops, VOTs are longer when stops are followed by /a/ than /i/ or /u/. Although the finer points of the issue are still undecided, a general conclusion that may be made is that vowel context does have some effects on voice onset time.

2.4 Mandarin and English stops and VOT patterns

In Lisker and Abramson's (1964) study, VOT measurements occurring before the release burst are said to have negative values, called 'voicing lead', whereas 'voicing lag' refers to measurements occurring after the release burst and are assigned positive values. Following these definitions, Keating (1984) subdivides the voicing lag dimension into 'short lag' (20–35ms) and 'long lag' (over 35ms). On the basis of this classification, stops are divided into three phonetic categories: voiced, voiceless unaspirated, and voiceless aspirated. Mandarin and English are said to contain two stop categories; detailed descriptions of the stops in these two languages are elaborated in the following sections.

2.4.1 English stops

Although, as Keating (1984) mentions, English has a great deal of positional variation, in the

present study only syllable initial stops are discussed. English is known to contrast voiced and voiceless phonemes in word-initial position, while voiced stops are said to have two possible phonetic realizations, voiced or voiceless unaspirated (Keating, Linker, & Huffman, 1983; Keating, 1984; Docherty, 1992). Lisker and Abramson (1964) provide two sets of VOT values for English voiced stops (/b, d, g/), one with a positive short lag, and the other with a negative voicing lead. They further suggest that only a single type of phonetic representation is produced by each native speaker. Klatt (1975) measures VOT values for English stops and reports positive values for both voiced /b, d, g/ and voiceless unaspirated stops /p, t, k/. Keating (1984) also points out that English voiced stops are sometimes pronounced with some lead values but mainly with short lag and long lag. Table 1 shows mean VOTs for English stops, as reported by Lisker and Abramson (1964), Klatt (1975), and Docherty (1992).

Table 1. Mean VOTs for English stops

	Lisker & Abramson, 1964 (AE)	Klatt, 1975	Docherty, 1992 (BE)
	Mean	**Mean**	**Mean**
/p'/	58	47	42
/t'/	70	65	64
/k'/	80	70	62
/p/		12	
/t/		23	
/k/		30	
/b/	1/-101	11	15
/d/	5/-102	17	21
/g/	21/-88	27	27

(AE=American English; BE=British English. All measurements are in milliseconds (ms). Note: /p', t', k'/ represents voiceless aspirated stops, while /p, t, k/ refers to voiceless unaspirated stops.

2.4.2 Mandarin stops

It is known that all Mandarin stops are phonetically voiceless and that aspiration is the only distinctive phonetic feature, differentiating two phonemic categories: voiceless unaspirated /p, t, k/ and voiceless aspirated /p', t', k'/. Unlike in English, stops in Mandarin occur only in word-initial position. Moreover, Mandarin stops fall into short lag versus long lag patterns.

Table 2 juxtaposes mean Mandarin VOTs, as measured by different researchers. As well as Rochet and Fei's (1991) study of Mandarin Chinese, Liao (2005) and Chao et al. (2006) focus on Taiwanese Chinese accents. Two points are of note. First, as the table shows, VOT values for Mandarin /p', t', k'/ are obviously higher than their equivalents in English. This may imply that for voiceless aspirated stops, especially for the velar /k'/, Mandarin and English may occupy different areas along the VOT continuum. Secondly, all values for /t'/ production are close to, but slightly lower than, the values for /p'/. It is interesting to note the possible effect of not conforming to the general pattern with respect to place of articulation.

Table 2. Mean VOTs in Mandarin

	Rochet & Fei, 1991 (MC)	Liao, 2005 (TC)	Chao et al., 2006 (TC)
	Mean	**Mean**	**Mean**
/p'/	99.6	75.4	82
/t'/	98.7	71.4	81
/k'/	110.3	98.8	92
/p/		17.9	14
/t/		18.6	16
/k/		28	27

(MC=Mandarin Chinese; TC=Taiwanese Chinese accent. All measurements are in milliseconds (ms). Note: Rochet & Fei only provide the mean VOT for voiceless aspirated /p', t', k'/.

3. Methodology

3.1 Aims of the experiment

As mentioned above, some studies have examined VOT in Mandarin Chinese, but few have attempted to compare Mandarin and English VOT patterns, particularly with respect to voiceless aspirated stops. To the best of our knowledge, so far only Chao et al. (2006) have compared VOT patterns in these two langauges, and they found that there are indeed subtle differences in VOT production between Mandarin and English. Therefore, the aim of the present experiment is to compare Mandarin and English VOT patterns.

3.2 Stimuli

It is known that, in Mandarin, stops occur only in the word-initial position; moreover, all stops are phonetically voiceless and they are only distinguished by aspiration. The present experiment examines only voiceless stops in the initial position. Klatt (1975) finds that the differences in VOT values relate to the environment of the following vowel. Therefore, in this experiment each of the stops is augmented by three peripheral vowels; that is, two high vowels, /i/ and /u/, and one low vowel, /a/. The Mandarin word list consists of 16 words (excluding /k'i/ and /ki/, as no meaningful lexical items for /k'i/ and /ki/ exist in Chinese). Note that compound words (two or three characters side by side forming a 'word') are used rather than single characters because they are more complete and more sense to the subjects.

Two procedures are used to create an English word list. First, only voiceless aspirated stops /p', t', k'/ in the word-initial position are examined here due to the debatable implementation of English voiced stops; moreover, a CVCV sequence is used to ensure the target stop is stressed. Velar /k'/ followed by the high vowel /i/ is not included, as no corresponding words are found in Mandarin. Secondly, analogous to the Mandarin stimuli, disyllabic and not monosyllabic words are used to design the English word list.

3.3 Subjects

Thirty-six native speakers of Taiwanese Chinese were recruited from various departments at National Cheng Kung University in southern Taiwan. Subjects include 21 staff (mean age= 40 years) and fifteen students (mean age= 22 years), aged from 20 to 50 (mean age for all

subjects= 32 years). All of the subjects were born and raised in Taiwan, have no marked regional accent, and reported no sophisticated knowledge of linguistics at the time of testing.

3.4 Procedures

Each subject was scheduled to record the word lists in a soundproof booth, using a high-quality microphone (AKG C1000S) and a professional 2-channel mobile digital recorder (MicroTrack 24/96). The target words for both languages were randomised in order not to be predictable. The recording was made when the subjects indicated they were ready. The subjects were first asked to read each word on the Mandarin and English word lists at a normal speed and repeat the whole lists twice in a row. All speakers were allowed to ask questions and practice words with which they were unfamiliar, but they were not informed of the purpose of the experiment. After the recording, they were asked to fill in a short questionnaire relating to their linguistic background.

3.5 Measurements and analyses

Wavesurfer software was used to make acoustic measurements of the speech material. Spectrograms and waveforms are displayed on screen and a manually controlled cursor is used for durational measurements, as shown in figure 1. VOT values were obtained by measuring the interval between the beginning of the release burst and the onset of the first formant visible in the frequency region. Target sounds that were obviously mispronounced are not included in the final analysis. Mean VOT values, standard deviations (SD), and graphical representation were made using EXCEL and SPSS. ANOVA tests were used for all statistical analyses, including the comparison of results and calculation of significance.

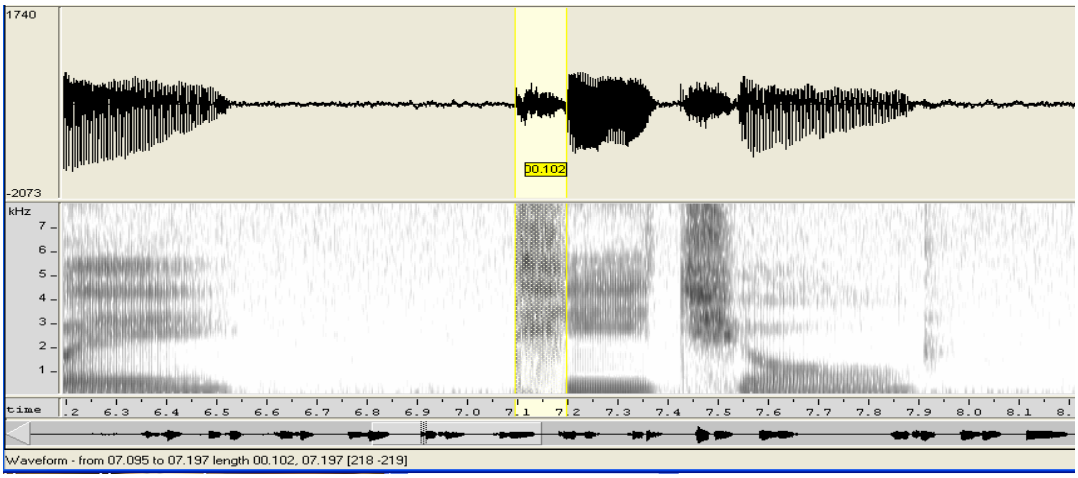


Figure 1: Spectrogram and waveform for the Mandarin word, “ti qiu”

4. Results

Mandarin VOT patterns for voiceless stops are discussed in section 4.1 below. Owing to the debatable phonetic implementations for English voiced stops, only voiceless aspirated stops (/p’, t’, k’/) in Mandarin and English are compared. Vowel quality is also taken into consideration in section 4.1.2, below.

4.1 Mandarin VOT

4.1.1 VOT means and distribution

The mean VOT values for six Mandarin stops are shown in figure 2, and detailed measurements including standard deviation (SD) are presented in table 3. Compared with the data reported by other researchers (Rochet & Fei, 1991; Liao, 2005; Chao et al., 2006), the VOT means for Mandarin stops presented in this study are relatively low, especially for the voiceless aspirated /k'/. Overall, VOT values for velar stops /k'/ and /k/ are significantly higher than those for bilabial and alveolar stops [F (2, 835) = 15.917, p= .000< .05]. Regarding the relation between place of articulation and VOT value, it is interesting to note that among voiceless aspirated stops, /t'/ has a higher value than /p'/, which does not conform to the general rule that VOT values rise as the place of articulation moves further back. The AONOVA test shows that the difference between /p'/ and /t'/ does not reach significance [F (1,627) = 1.885, p= .170 > .05]. However, this finding is only relevant to the voiceless aspirated /p', t'/, and not to the voiceless unaspirated /p, t/. In addition, as table 3 indicates, , contrary to English VOT patterns, the mean VOTs for Chinese bilabial and alveolar stops are much closer to each other, both for aspirated and unaspirated stops. The two main results of the present study are in accordance with studies by three other researchers (Rochet & Fei, 1991; Liao, 2005; Chao et al., 2006). The only difference is that for aspirated stops, Liao (2005) reports /p'/ with a significantly higher value than /t'/ [F (1, 19) = 7.464, p= .013< .05], while the two other studies showed no significant difference.

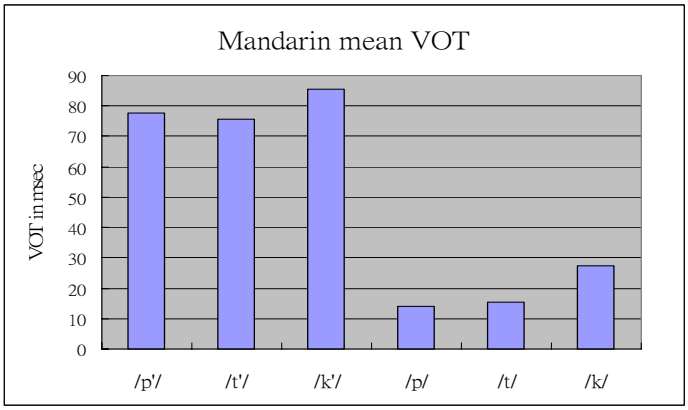


Figure 2. Mean VOT values for Mandarin stops

Table 3. General VOT means (ms) and standard deviation (SD) for all Mandarin stops

	/p'/	/t'/	/k'/	/p/	/t/	/k/
General means (in ms)	77.8	75.5	85.7	13.9	15.3	27.4
Standard deviation (SD)	23.7	18.4	19.4	6.6	5.7	9.6

Figures 3 and 4 show the VOT distribution for all Mandarin stops. Looking first at the voiceless aspirated stops, it can be seen that VOT ranges for /p', t', k/ are centralized around 63–90ms, 65–87ms, and 74–98ms, respectively. The values of standard deviation (SD) presented in table 3 also imply that /p'/ (SD=23.7 ms) allows more variation than /t/ (SD=18.4 ms) and /k'/ (SD=19.4 ms). As for voiceless unaspirated stops, the VOT ranges are centered around 10–18ms, 12–18ms, and 20–33ms, respectively. Unlike voiceless

aspirated stops, the unaspirated /k/ (SD=9.6 ms) shows more variation than the two other stops and it may also be seen that the VOT range for /t/ is smaller than those for /p/ and /k/.

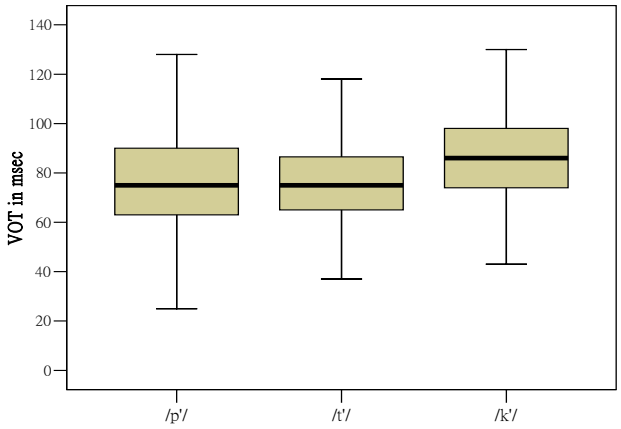


Figure 3. Boxplot for Mandarin voiceless aspirated stops

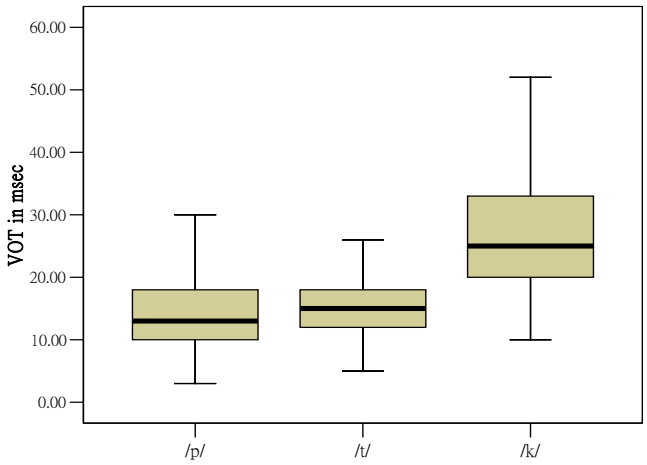


Figure 4. Boxplot for Mandarin voiceless unaspirated stops

4.1.2 Vowel context

Although there is an exception (Fant, 1973), it is widely accepted that word-initial stops have longer VOT values when followed by high vowels than by low vowels (Klatt, 1975; Weismer, 1979; Port, 1979; Rochet & Fei, 1991; Chao et al., 2006). In addition, Chao et al. (2006: 33) report that "all the stops, except /t/ which does not yield significance, have significantly longer VOTs when the following vowel is /i/ or /u/ than when it is /a/." Figures 5 and 6 show VOTs for voiceless stops followed by one of the three vowels, /i, u, a/. As the figures indicate, the VOTs for the unaspirated stops /p, t, k/ and the aspirated stops /p', t', k'/ are shorter when followed by the low vowel /a/ than by the high vowels /i/ and /u/. When doing t-test, the result also reveals that vowels, high or low, have significant effect on the VOTs for stops [$p < .05$].

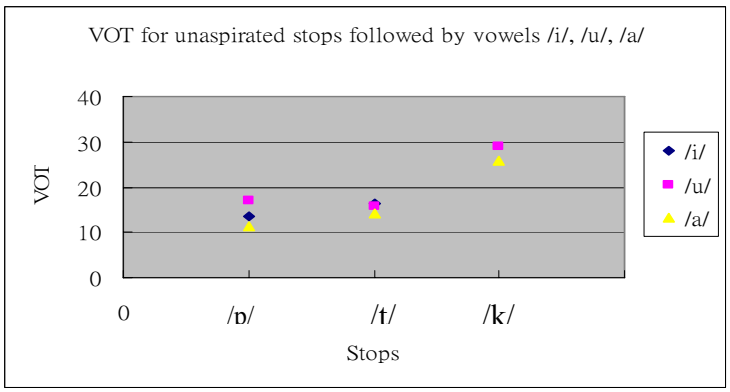


Figure 5: VOT for unaspirated stops followed by vowels /i/, /u/, /a/

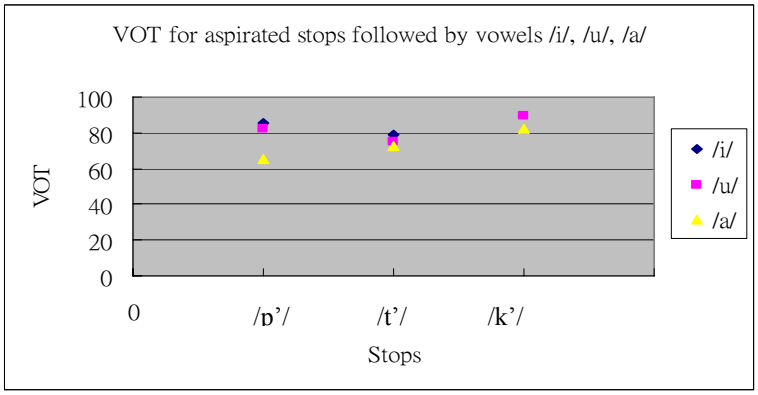


Figure 6: VOT for aspirated stops followed by vowels /i/, /u/, /a/

4.2 Comparing Mandarin and English VOT

As mentioned at the beginning of section 4, only phonetically voiceless aspirated stops are involved in the comparison of Mandarin and English VOT patterns. Figure 5 presents the mean VOTs for /p', t', k'/ in the two languages. The English mean VOTs are adopted from Lisker and Abramson's (1964) influential cross-language study. Visual inspection of the figure shows that Chinese speakers generally produce higher VOTs for /p', t', k'/ than English speakers. It should be noted that the differences between Mandarin and English VOTs are not stark but subtle, which raises the question whether L2 learners are aware of the slight differences between the two languages and are capable of producing them with authentic L2 production. This issue will be further discussed in section 5, below.

Apart from the differences mentioned above, place of articulation is another point which is worth noting. It is widely known that the further back the place of articulation, the longer the VOT, and there seems to be a general consensus on this. However, as figure 7 indicates, the mean VOTs for English /p', t', k'/ follow this rule, whereas Mandarin /p'/ and /t'/ do not. Moreover, the VOT values for aspirated bilabial and alveolar stops are closer to each other in Mandarin than in English VOT patterns.

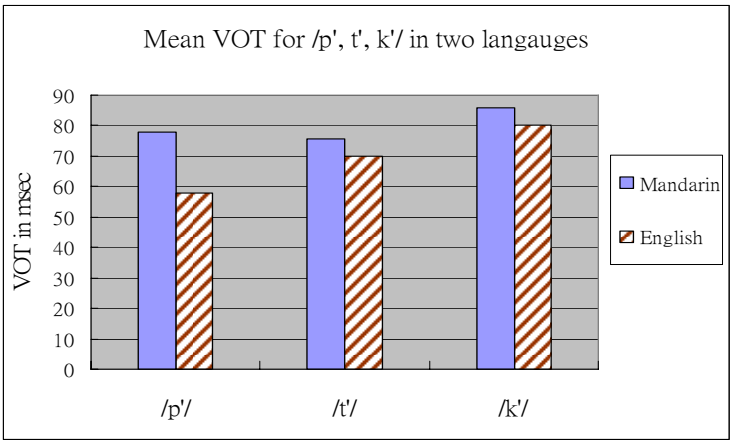


Figure 7. Mean VOTs for voiceless aspirated stops in Mandarin and English

4.3 English VOT in native Mandarin speakers

In section 4.2, it was mentioned that there are slight differences between VOT productions for voiceless aspirated stops in Mandarin and English. Since the two languages share similar VOT patterns with only subtle differences, it is worth investigating how native Chinese speakers produce English voiceless aspirated /p', t', k'/. Chao and Chen (2006) find that native Chinese speakers often produce English /p', t', k'/ with 'compromise' values. Thus, it is interesting to observe the English VOT patterns of the L2 learners (i.e. Chinese learners of English) in this study.

4.3.1 VOT means and distribution

Figure 8 shows the mean VOT durations for English /p', t', k'/ produced by native Chinese speakers; detailed measurements including SD are presented in table 4. As the figure shows, the velar /k'/ has a highly significantly longer VOT than the three voiceless aspirated stops [F (2,831) =106.450, p= .000< .05]. Nevertheless, VOT values for /p'/ and /t'/ still do not reach significance (p> .05). This result is similar to that found for Mandarin VOT patterns that differ in place of articulation. As mentioned above, the VOT values for Mandarin /p', t'/ do not increase as the place of articulation moves further back. However, Chinese speakers' L2 production accords with the general rule.

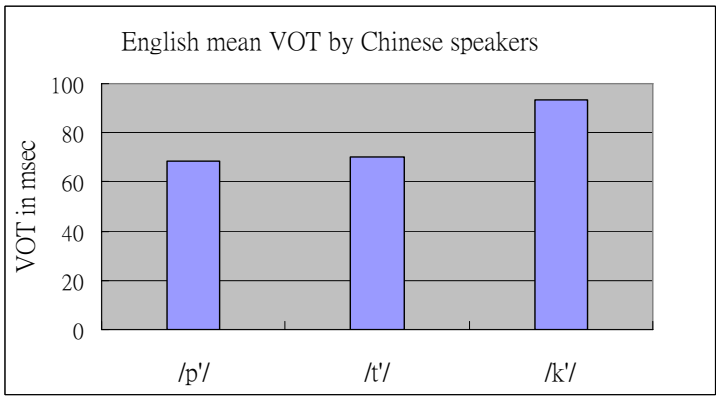


Figure 8. English VOT means for voiceless aspirated stops by Chinese speakers

Table 4. English VOT means (ms) produced by native Chinese speakers; standard deviation (SD) for voiceless aspirated stops

	/p'/	/t'/	/k'/
General means (in ms)	68.7	70.2	93.4
Standard deviation (SD)	21.8	19.2	20.5

Figure 9 compares the mean VOTs for English productions by native Chinese speakers, native Mandarin productions, and native English productions. Looking at the figure, it may be noted that native Chinese speakers produce intermediate VOT values only for English aspirated /p'/, by comparison with native speaker productions for either language. One may also notice that in their production of aspirated velar /k'/, Chinese speakers produce far higher English VOTs than in their corresponding Chinese production and than the English mean produced by native speakers. As for /t'/ production, it is interesting to observe that English mean VOTs for native speakers and Chinese subjects are almost the same (VOT= 70ms for the former; VOT= 70.2ms for the latter). Individual variations among Chinese native speakers should also be taken into account when forming comparisons.

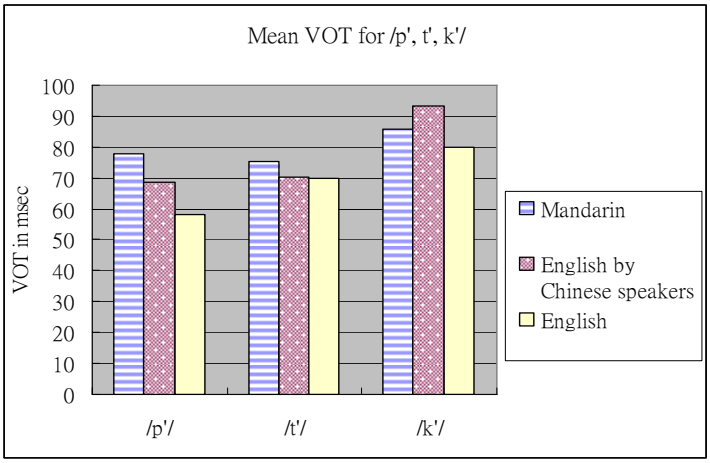


Figure 9. Mean VOTs for voiceless aspirated stops in Mandarin, English produced by Chinese speakers, and English produced by native speakers

5. Discussion

Three important conclusions may be derived from the present study and will be discussed in detail below.

With respect to Mandarin VOT patterns, the VOT means obtained for the six Chinese stops are somewhat lower than the data reported in previous studies (Rochet & Fei, 1991; Liao, 2005; Chao et al., 2006), as shown in table 5, below. The lowness of these values may be explained as follows. First, disyllabic words were used in the present study, rather than the monosyllables which were examined in the study by Rochet & Fei (1991). Using disyllables creates a more natural context for the subjects and likely obtains more accurate VOT values. Methodological differences may be another reason for the lower values. A

third explanation is the number of subjects tested: the experiment for the present study used more subjects than the previous two studies (Liao, 2005; Chao et al., 2006), which means the values obtained are probably more reliable.

As for the comparison of English and Mandarin VOT patterns, the results indicate that for voiceless aspirated stops, both English and Mandarin belong to the long-lag category, contrary to previous findings. Chao et al. (2006) claim that Mandarin /p', t', k'/ fall into the highly aspirated category and suggest that the aspirated category should not be considered as a single long continuum. The VOT means reported by Rochet and Fei (1991) also imply that Mandarin and English occupy different regions of the VOT continuum. Although both languages share similar stop category, there are still differences between them. Comparing /p'/ productions first, it can be seen that Chinese speakers produce much longer VOT values than native English speakers. Among the three voiceless aspirated stops in each language, only the alveolar /t'/ has a value close to the others. This accords with the results presented by other researchers who examined Chinese voiceless stops (Rochet & Fei, 1991; Liao, 2005; Chao et al., 2006). Place of articulation is another factor worth noting. Although the general rule states that the further back the place of articulation, the longer the VOT, this is not the case for Mandarin voiceless aspirated stops. The results indicate that the VOT for the aspirated alveolar stop /t'/ is shorter than that for the aspirated labial stop /p'/, except when they are followed by the low vowel /a/. The results for stops followed by the low vowel /a/ are consonant with the results of Chen, Tsay, and Hong's study (1998). The cause of this result is complicated and requires further discussion. In addition to Mandarin, Lisker and Abramson (1964) report that unaspirated stops in Tamil and aspirated stops in Cantonese and Eastern Armenian do not follow the general rule either. Of these four languages, both Cantonese and Mandarin are tone languages. Whether tone affects VOT values is still a controversial question. Some researchers have claimed that there is no significant influence (Chen et al. 1998; Ran, 2005), whereas in a study by Liu et al. (*Article in Press*) it is found that "VOT values associated with high-level and high-falling tones were shorter than those associated with mid-rising and falling-rising tones." The test stimuli used in the present study are not in the same tone; therefore, if tones do influence VOT values, it is possible that some of the results may be explained in this way.

Table 5. Mean VOT values (ms) for Mandarin voiceless stops

	Rochet & Fei, 1991 (monosyllables)	Liao, 2005 (disyllables)	Chao et al., 2006 (disyllables)	Present study (disyllables)
	Mean	**Mean**	**Mean**	**Mean**
/p'/	99.6	75.4	82	77.8
/t'/	98.7	71.4	81	75.5
/k'/	110.3	98.8	92	85.7
/p/		17.9	14	13.9
/t/		18.6	16	15.3
/k/		28	27	27.4

As for vowel context, it is found that the VOTs for stops, both unaspirated and unaspirated, are longer when followed by the high vowels /i/ and /u/ than by the low vowel /a/. This supports the findings of many studies (Port, 1979; Gósy, 2001; Rochet & Fei, 1991; Chao et al., 2006). Although there are some exceptions (Lisker & Abramson, 1967; Fant, 1973), more and more studies support the view that high/low vowel quality influences the VOT value of preceding stops. Front/back vowel quality has no significant influence on VOT.

Since the differences between Mandarin and English VOTs are subtle, it is worth observing the English VOT performance of Chinese speakers. Chao and Chen (2006) propose that native Chinese speakers often produce English /p', t', k'/ with 'compromise' VOT values. Whether these speakers are able clearly to distinguish the subtle differences between the two languages, or whether their L2 productions are influenced by their first language (i.e. Mandarin), is an interesting issue for further discussion. The present findings reveal that, except for /k'/, Chinese speakers' L2 productions of /p'/ and /t'/ are either intermediate or close to English native speakers' productions. To understand the exception of /k'/ values, language proficiency could be taken into consideration. Liao (2005) suggests that proficiency has a certain influence on interlanguage production of stop consonants. According to Liao (2005), L2 learners with a higher level of proficiency have greater accuracy than those with a lower level. 21 of the staff members observed in this study are classified as having a low level of proficiency, which may be one of the reasons for their striking /k'/ production. It should also be noticed that the mean VOT values of English /p'/ and /t'/ by Chinese speakers are close to each other. Previous studies have examined this phenomenon and provided various suggestions for factors affecting L2 production. On the one hand, it is suggested that first language (L1) effect on L2 plays a crucial part in L2 learners' VOT productions (Thompson, 1991; Flege et al., 1997). On the other hand, Flege and Hammond (1982) also claim that speakers actually produce intermediate phonetic categories between their native language and a foreign language. Variations in L2 production both for the same persons over time and for different speakers could be examined further and taken into consideration in future studies.

References

[1] A. Caramazza, G. Yeni-Komshian, E. Zurif, and E. Carbone, "The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals," *Journal of the Acoustical Society of America*, vol. 54, pp. 421–28, 1973.

[2] B. L. Rochet and Y. Fei, "Effect of consonant and vowel context on Mandarin Chinese VOT: production and perception," *Canadian Acoustics*, vol. 19, no. 4, pp. 105, 1991.

[3] D. H. Klatt, "Voice Onset Time, frication, and aspiration in word-initial consonant clusters," *Journal of Speech and Hearing Research*, vol. 18, pp. 686–706, 1975.

[4] D. Kewley-Port, D. B. Pisoni, and M. Studdert-Kennedy, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *Journal of Acoustical Society of America*, vol. 73, no. 5, pp. 1779–1793, 1983.

[5] G. H. Yeni-Komshian, A. Caramazza, and M. S. Preston, "A study of voicing in Lebanese Arabic," *Journal of Phonetics*, vol. 7, pp. 197–204, 1977.

[6] G. Khattab, "VOT production in English and Arabic bilingual and monolingual children," *Leeds working papers in linguistics and phonetics*, vol. 8, pp. 95–122, 2000.

[7] G. Weismer, "Sensitivity of voice-onset-time (VOT) measures to certain segmental

features in speech production," *Journal of Phonetics*, vol. 7, pp. 197–204, 1979.
[8] G. J. Docherty, *The timing of voicing in British English obstruents.* New York: Foris, 1992.
[9] H. Liu, M. L. Ng, M. Wan, S. Wang, and Y. Zhang, "The effect of tonal changes on voice onset time in Mandarin esophageal speech," *Article in Press.*
[10] I. Thompson, "Foreign accents revisited: the English pronunciation of Russian immigrants," *Language Learning*, vol. 41, pp. 177–204, 1991.
[11] J. E. Flege and R. M. Hammond, "Mimicry of non-distinctive phonetic differences between language varieties," *Studies in Second Language Acquisition*, vol. 5, no. 1, pp. 1–17, 1982.
[12] J. E. Flege, O.-S. Bohn, and S. Jang, "Effects of experience on non-native speaker's production and perception of English vowels," *Journal of Phonetics*, vol. 25, pp. 437–470, 1997.
[13] J. R. Benkí, "Place of articulation and first formant transition pattern both affect perception of voicing in English," *Journal of Phonetics*, vol. 29, pp. 1–22. 2001.
[14] K. Chen, J. Tsay, and G. Hong, "Duration of initials in Mandarin: Fundamental acoustic research and its clinical significance," *Journal of Speech Language-hearing Association*, vol. 13, pp. 138–149, 1998.
[15] K.- Y. Chao, G. Khattab, and L.- M. Chen, "Comparison of VOT Patterns in Mandarin Chinese and in English," in *Proceedings of the 4th Annual Hawaii International Conference on Arts and Humanities*, 2006, pp. 840–859.
[16] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: acoustical measurements," *Word*, vol. 20, pp. 384–422, 1964.
[17] L. Lisker and A. S. Abramson, "Some effects of context on voice onset time in English stops," *Language Speech,* vol. 10, pp. 1–28, 1967.
[18] M. Gósy, "The VOT of the Hungarian voiceless plosives in words and in spontaneous Speech," *International Journal of Speech Technology*, vol. 4, pp. 75–85, 2001.
[19] M. M. Kehoe, C. Lleó, and M. Rakow, "Voice onset time in bilingual German-Spanish children," *Bilingualism: Language and Cognition*, vol. 7, pp. 71–88, 2004.
[20] O.-S. Bohn and J. E. Flege, "Perceptual switching in Spanish/English bilinguals," *Journal of Phonetics*, vol.21, no. 3, pp. 267–290, 1993.
[21] P. A. Keating, "Phonetic and phonological representation of stop consonant voicing," *Language*, vol. 60, pp. 286–319, 1984.
[22] P. A. Keating, W. Linker, and M. Huffman, "Patterns in allophone distribution for voiced and voiceless stops," *Journal of Phonetics*, vol. 11, pp. 277–90, 1983.
[23] Q. B. Ran, "Experimental studies on Chinese obstruent consonants: with the emphasis on standard Chinese," Ph. D. dissertation, Nankai University, Mainland China, 2005.
[24] R. F. Port and R. Rotunno, "Relation between Voice-Onset Time and vowel duration," *Journal of the Acoustical Society of America*, vol. 66, pp. 654–62, 1979.
[25] R. H. Kessinger and S. E. Blumstein, "Effects of speaking rate on voice-onset time in Thai, French, and English," *Journal of Phonetics*, vol. *25,* pp. 143–168, 1997.
[26] S. J. Liao, "Interlanguage production of English stop consonants: A VOT analysis," M. A. thesis, National Kaohsiung Normal University, Kaohsiung, Taiwan, 2005.
[27] T. Cho and P. Ladefoged, "Variation and universals in VOT: evidence from 18 languages," *Journal of Phonetics*, vol. 27, pp. 207–29, 1999.
[28] T. J. Riney and N. Takagi, "Global foreign accent and voice onset time among Japanese EFL speakers," *Language Learning*, vol. 49, no. 2, pp. 275–302, 1999.
[29] T. J. Riney, N. Takagi, K. Ota, and Y. Uchida, "The intermediate degree of VOT in Japanese initial voiceless stops," *Journal of Phonetics*, vol. 35, no. 3, pp. 439–443, 2007.
[30] X.-R. Zheng and Y.-H. Li, "A contrastive study of VOT of English and Korean Stops," *Journal of Yanbian University*, vol. 38, no. 4, pp. 99–102, 2005.

台灣共通語言

Taiwan Common Language

余明興 Ming-Shing Yu
國立中興大學 資訊科學與工程學系
Dept. of Computer Science and Engineering
National Chung-Hsing University
Taichung, Taiwan, 40227
msyu@nchu.edu.tw

摘要

本篇論文討論理想的漢字書寫方式，我們以國語、台語、和英語混合使用的情境作為書寫方式所要描述的對象，這涵蓋了現在最常使用的書寫內容和將來極為可能的書寫內容。 我們會討論書寫方式所應具備的功能、目前漢語書寫方式的優缺點、以及書寫時要遵循的原則等。 綜合上述的諸多條件之後，我們提出一種新的漢字書寫方式，以及拼音字母和聲調符號的設計。 這種書寫方式不僅易學易懂，而且提供漢字一種進化的機制，使其能隨著時代的變遷而調整。 就母語（主要是台語、客語等）書寫的需要而言，本套書寫方式可以看成是用音來補強漢字不足的地方，不需要太多的學習就能很容易的溝通。 國語轉母語的工作會變的很簡單，我們可以很容易地擁有大量的母語文獻。 到時母語的推廣和資訊處理就會有足夠的文獻可以使用。

關鍵詞：台灣共通語言、漢字書寫方式、拼音、聲調。

誌謝： 本研究承國科會計畫”台灣共通語言環境建置”補助，計畫編號為 93-2213-E-005-016 和 94-2213-E-005-002，謹此誌謝。

§1. 緣起

法國科學家法布爾(Jean-Henri Fabre 1823-1915)曾經觀察到一種有特殊習性的蟲子，這種蟲子喜歡跟著前一隻同類的後面前進，我們就稱這種習性為「跟隨者」的習性[張晴雨、李翰祥 2004]。 那麼，如果將這些蟲子頭尾相接排成一圈，會有什麼事發生呢？ 難道會一直繞圈嗎？ 法布爾就真的抓了很多這種蟲子，把他們首尾相接放在一個花盆的周圍，並且在不遠處放置牠們喜歡吃的食物。 結果，這些蟲子只是不停地繞著花盆轉圈子，一星期之後，牠們都累死(或餓死)在花盆四週。

就漢字的使用而言，我們是否也像上述“跟隨者”的蟲子一樣？ 就算是現在使用的漢字系統有許多缺點，我們仍然不加以改進，而繼續沿用？ 要知道語言文字的使用效率如果不高的話，至少會浪費掉我們的寶貴時間。 如果一個人一天浪費了十分鐘，那麼全台灣一天要浪費多少時間？ 更有甚者，如果某些思想、智慧因為書寫方式功能不足而無法記錄下來，那麼損失會有多少呢？

即使你不覺得國語書寫有困難，至少台語（或客語）的書寫總沒有那麼容易吧！ 我們在這裡提出的書寫方式是以國語、台語、和英語混合使用的情境作為書寫方式所要描述的對象。我們希望不必經過太多學習，文章可以很容易寫，也很容易懂。 假設我們現在要寫一句台語，我們設定的目標是：對於懂台語的讀者，我們希望他能夠明白該句的意義和語音；對於只懂國語的讀者，我們希望他能夠明白該句的意義。

本論文組織如下。 我們將於第二節討論漢字的主要缺點，以及我們識字的能力和合理的學習時間。 在第三節我們簡單討論書寫方式在處理本土語言的重要性。 在第四節提出我們台灣共通語言的書寫原則。 在第五節我們提出拼音字母和聲調符號的設計。 第六節是結語。

§2. 漢字的主要缺點

在本節我們要討論漢字的主要缺點，才知道要改進的地方在哪裡。 我們認為漢字的主要缺點有三個：字太多、沒斷詞、不表音。 而漢字也有一個大優點：表意。 我們將在用到之前加以說明。

§2.1. 字太多

漢字的字數若只計算教育部所公佈的繁體正體字約有 3 萬(29,892)個字，另有異體字 76,338 字[教育部異體字字典 2004]。 字這麼多當然難以學習和記憶。 為了容易區分，字

的筆劃數勢必不可能都很少，那些筆劃數多的字通常都很難寫。 那麼這 3 萬字中，我們能夠學到多少？，又能夠用出(寫出)多少字？ 曾志朗曾經統計過金庸小說中所出現過的不同的字，不到 5 千個[曾志朗 2002]。 並且推測人的用字極限約為 4300 個。 因此我們可以說，這 3 萬字絕大部分都是我們幾乎用不出來的。

在這 3 萬個漢字中，目前我們”只”需要學 5 千多個漢字。 學 5 千多個漢字需要花費我們 12 年的時間，從小學一直學到高中，尤其在小學階段最為辛苦。 為了學習這麼多的漢字，我們的小學生需要花很多的時間將這些字寫很多遍，很多人寫字的手指頭都有一顆小突起，就是寫字寫出來的。 但是在小學階段的學童，需要的是跑、跳、爬等全身性的大肌肉運動，花太多時間寫字會影響健康和發育[蔣為文]。

冤枉的是，這 5 千多個字約有一半是用不太到的。 我們曾對所收集到的含有 85905204 字（書一頁若以 500 字計算，約相當於 200 頁的書 859 本）的語料做統計，最常用的 948 個字可以涵蓋 90%的內容，最常用的 2469 個字可以涵蓋 99%的內容。 假如我們少學一半，只學最常用的 2500 個字，那麼我們的負擔就大大的減少。 所會遇到的情況是，平均 100 個字才會遇到一個不認識的字。 這個字我們就把它的音寫下來，在絕大多數的情況下（人名、公司名等專有名詞除外），我們是可以看得懂的。

§2.2. 沒斷詞

話說小明最近運氣不太好，去找算命仙算命。 算命仙跟他說”大富大貴，沒有大災難，要小心”，小明就很放心的回去了。 過了幾天小明卻出了車禍，還好只是輕傷，他就回去找算命仙理論。 算命先說，我明明跟你交代的很清楚”大富大貴沒有，大災難要小心”，是你自己不夠小心的。 當然，很多人也聽過”下雨天留客天留我不留”的例子。

上面的誤解或各說各話就是來源於中文的書寫中並沒有清楚的斷詞。 或許有人會說，我在閱讀時並沒有感受到斷詞問題的嚴重，絕大部分的情形沒有斷詞並不會造成什麼困擾。 我要說，其實我們有此功力是練了許多年的結果，我們已經花費了太多的時間在做這種練習。 如果有適當的斷詞，我們可以省下時間做更多的事情或學習。 這就是說，漢語在這方面的”效率”不夠高。

或許有人會說我講的太誇張，大家可以看以下英文的例子。 我們如果把詞和詞中間的空白去掉會怎樣？ 容易讀嗎？
“Afterenrollment,studentsareinformedoftheconditionoftheuniversityandtheeffortsstudentsshouldmakeinthefuturethroughthefreshmencounselingservice.”

這是我從某個大學的英文網頁有關學生的部份抓下來的一句話。 有興趣的人可以試試看，你要花多少時間才能得到以下的原始內容：

“After enrollment, students are informed of the condition of the university and the efforts students should make in the future through the freshmen counseling service.”

大家應該可感覺到，”效率”很低是不是？ 其實外國人學中文時，斷詞就是一項難題。我們已經花下許多時間去學習，才變的不是問題。

§2.3. 不表音

我們先描述中文在表達讀音方面的能力，再來討論它在語言學習的過程中的情況。 像中文這種文字，基本上一個字都帶有某種涵義，通常稱為「意符文字」。 意符文字的好處是當我們學得夠多的字和它們的涵義後，看到某些不認識的詞(字的組合)時，常常可以猜出這個詞的意思。 意符文字通常字數多(可多達數萬)、連帶的筆畫複雜字形類似，難以學習。 中文的另一個特徵是”字”通常不帶有”音”的訊息。 我們如果不認識某一個字，通常是無法讀出它的音，而且可能差的很遠，完全讀不出來。

拼音文字就是用一些表示音的字母來拼出所要描述的語音，一個詞通常由若干個字（字母）構成。 這種文字通常看到詞可以讀出語音，至少是相似的語音，不會差太多。 拼音文字通常字數很少，只有幾十個或一、兩百個，容易學習。 拼音文字的一個缺點是說你如果看到一個未曾聽過或見過的詞，那麼就很難猜測它的意義了。 嚴格說起來，任何一種文字都帶有”意”和”音”的訊息；意符文字有讀音，拼音文字也有意義。 上述的”意符文字”和”拼音文字”描述的是它們的主要特徵或來源。

一般人學習語言的過程有四個階段：聽、說、讀、寫。 出生開始先聽人家說，試著去了解人家所說的話。 而說的人則盡量以現場的景物配上肢體動作等，讓聽者能明瞭話的意思。 等了解到夠多的語彙（以語音形式表現的詞彙）和語法之後，就可以開始說話了。在學到了夠多的字詞（書寫形式）之後，就可以開始閱讀。 而在閱讀時，只要閱讀內容不超出閱讀者所能理解的範圍(也就是生活知識和語言知識等)，閱讀者就能理解。 在讀過夠多的文章之後，就可以照著開始學寫文章。

我們以中文(國語)和英語的學習來比較。 在聽和說的階段，兩種語言的學習進展情形是差不多的。 而要進入閱讀階段就有較大的區別，因為中文要認識和記憶到足夠的字詞需要較長的時間（由於中文字不表音的關係)，因此中文要開始閱讀就需要較長的準備時間。 英文在閱讀的時候比較能夠將文字的音讀出來，只要閱讀內容不超出閱讀者所能理解的範圍，閱讀者就能理解。 比較而言，中文要開始閱讀需要較長的準備時間。 因為漢字難學難記，所以我們的小學生必須將大部分的時間放在漢字的練習，而比較忽略文學作品的欣賞和人文素養的培養。

不表音的另一個壞處是很多的”語言”，例如台語和客語等母語，難以書寫；這就是一般

說的「有音無字」的現象。 母語(mother tongue)指的是人們所在的族群講的族語；在自然的情況下，和母親所講的話是相同的。 用母語來溝通是最自然、最親切的方式。 小孩從小就應該跟他講母語，要求他使用母語。 用母語讓他體驗人們的關愛，使他對所處的社會有深厚的感情。 進而喜歡這個社會，以這個社會為榮[楊維哲 1997]。

用母語來學習效率也最好，因為在學習的時候語音配合意義，使得學習效果好，腦部的連結和發展也好。 楊維哲教授就曾這樣說過：「學習語言，就是人生的第一步，就是小孩學習其他一切事物的基本工具。 我們用族語教他，他自然學到很多有關的文化、價值觀、分析力。這些東西，你若不是用你的族語去教他，他就只是學到「陽春」的字句而已。 不管你的族語是什麼語言，就是因為你教他這一項語言，他的效率才會這麼高。」

任何一種語言都是人類寶貴的資產。 用母語來溝通是最自然、最親切的方式。 聯合國教科文組織在「全球瀕臨消失危機的語言概況」文章中提到：一個語文的消失，意味著人類思想與知識寶庫的萎縮與枯竭。」 因此我們的書寫方式必須要能夠很容易的來書寫母語。

§3. 計算語言學相關問題

從計算語言學的角度來看，如果書寫的方式不為多數人接受，那麼有許多語言處理的工作將會遭遇到困難。 首先是語言的翻譯將沒有適當的對應形式，例如將國語的文句轉到台語的文句時，要如何表達台語的文句？ 再者是語料的收集整理和使用將會很困難，相信許多從事台語處理的人都會有語料不足的經驗。 由此可以看出書寫方式在計算語言學中的重要性。

在處理本土語言的時候，如果不能有適當的書寫方式，很多問題就無法繼續做下去。 我們最近正在研究的一個主題是要將國語文句轉成台語語音，暫且稱其為「國文轉台音系統(Mandarin Text to Taiwanese Speech System)」。 這樣的一個系統的建構流程我們以一個例子來說明，這個例子是從中研院的平衡語料庫中找出來的： 我們 真 希望 科學家 能 發明 打 下去 不會 痛 的 針。

我們構想的解決步驟如下：

1. 轉成台語文句： 阮 真 希望 科學家 能夠/eiˇdangˇ/ 發明 注 下去 不/bhei_/ 痛 的 射。
2. 變調處理： 阮/ghun¯/ 真/jin_/ 希望/hi_bhangˇ/ 科學家/ke_hakˇga¯/ 能夠 /eiˇdang`/ 發明/huat`bhing_/ 注/zuˇ/ 下去/loˇ-kiˇ/ 不/bheiˇ/ 痛/tiannˇ/ 的

/ei_/ 射/sia_/。

3. 韻律訊息： 停頓、音調、音長、音量等。
4. 語音處理： 輸出所要的語音。

首先我們就需要把國語文句轉成台語文句，因此台語文句需要有適當的寫法。 至少要解決不同人使用不同詞彙的問題，例如上例中的「能」，有人會寫成「能夠」，也有人會寫成「會當」。 我們認為第一個步驟是非常重要的步驟，如果有適當的書寫方式，又能夠做的很好的話，那麼台語很快就可以有許多文章可供閱讀、教學，也有很多語料可供研究。 台語的書寫方是目前有全漢字、全羅馬字、以及漢羅並用三大類。 因為書寫的方式差異甚大，使得溝通不易，語料的收集和累積也很慢，阻礙了推廣和進一步的研究（如文法等）。

§4. 台灣共通語言的書寫原則

我們在書寫時，最重要的就是要讓讀者能夠明瞭我們所寫的文句的意義（意），其次是希望讀者能夠讀出我們所寫文句的語音（音）。 假設我們現在要寫一句台語，我們設定的目標是：對於懂台語的讀者，我們希望他能夠明白該句的意和音；對於只懂國語的讀者，我們希望他能夠明白該句的意。 我們希望能夠盡量使用較常用的字，因為前面有提到過，我們只需要 2 千 5 百字就能涵蓋 99%的用字。 消極來說，至少不要造字。 因為目前字已經多到無法盡學了，再加字給大家學幾乎是不可能的事情。

我們在前面提到，目前的漢字書寫系統有許多缺點，因此需要改進。 但是我們也應該運用漢字的優點，漢字最大的優點就是因為它主要有表意的功能，因此可以跨越時空。 從空間上看，在中國，相同的文字在不同的地區可能有不同的讀法，但所代表的意義卻是一樣。 就時間而言，以前的文章或許古人的讀法和我們不一樣，但是意義仍然可以了解。 我們的書寫原則有三個：字和音併用、以詞為書寫單位、和外來語直接引用；以下就分三小節來說明。

§4.1. 字和音併用

一篇文章中若是只有少數漢字用拼音取代，並不會影響我們對文章的了解，除非是人名等專有名詞可能無法知道正確的字。 反而是有些語言中找不到字（或沒有較為公認的字、或忽然間寫不出來的字）的音可以很容易的拼寫出來。 我們也可以採用字加音的方式來避免讀者讀錯音。 以下是一些例子。

(a) 這樣的場面讓人很 gan¯ga`。

(b) 我們/ghun¯/明天要去爬山，你要去嗎？

(c) 我們/lan¯/明天要去爬山，你有準備什麼東西/miˇgiann_/？

(d) 番茄/ta_ma¯do˙/ 是很好的水果。

在上述四句中，我們在(a)句用拼音取代文字，而在(b)到(d)句是把拼音附加在文字旁邊。句(a)是一句國語。 在(a)中，如果”尷尬”忽然間寫不出來，用拼音應該也可以讀的懂。 句(b)句是一句台語。 在(b)中，”我們”並不包含聽者，因此要唸成/ghun`/（台語第二聲），在這裡我們標的音是變調後的台語一聲音。 如果你擔心讀者唸不出來正確的音，除了字之外，可以把音也加上去。 句(c)是一句台語。 在(c)中，”我們”有包含聽者，因此要唸成/lan`/，在這裡我們標的音是變調後的台語一聲音。 類似地，如果你擔心讀者把”東西”唸成/dong_sai¯/，就可以把/miˇgiann_/這個音加上去。 在這裡”東西”有人可能會主張寫成”物件”，如此義音兼具。 我們認為如果讀者是精通台語的人，應該可行。 可是如果讀者只懂國語，要看懂就比較困難，這樣文章就比較不易流通。 句(d)是一句台語。 在(d)中，我們把番茄的台語音/ta_ma¯do˙/加上去，這樣意和音都有了。 就算是只懂國語的人，除了看的懂之外，也可以順便學習番茄的發音。 在上面的例子中，我們的拼音加上了聲調符號。對於大部份的二字或更多字的詞，或者是作為補充拼音使用的情況，並不一定需要加上聲調。沒加聲調的句子就如下列，但音節之間要加上連音符號。

(e) 這樣的場面讓人很 gan-ga。
(f) 我們/ghun/明天要去爬山，你要去嗎？
(g) 我們/lan/明天要去爬山，你有準備什麼東西/mi-giann/？
(h) 番茄/ta-ma-do/ 是很好的水果。

§4.2. 以詞為書寫單位

詞和詞之間可以視需要加以斷開（建議用空白符號），至少在語義有停頓的地方可以加上空白。 在我們前面的例子句(d)中，番茄/ta_ma¯do˙/後面就有加上空白。 另外我們建議像英文一般，在句和句之間(即句點、問號、和驚嘆號後面)加上空白。 以下我們再舉一些例子。

(i) 下雨 天留客 天留 我不留。
(j) 下雨天 留客天 留我不 留。
(k) 你們和他們 什麼時候/dang-si/ 要比賽籃球？

(l) 地震/dei-dang/後 很多壁 都有裂縫/bit-sun/。
(m) 地動(地震)後 很多壁 都有裂縫。

在上述五句中，(i)和(j)句是常見的國語句子，用來寫台語也適用；(k)到(m)句是台語的例子。 在句(k)中，我們標示”什麼時候（英文為 when）”要唸成/dang-si/，不是要唸成/sia-mi-si-hao/。 這是一個明顯的以詞為書寫單位的例子，我們把”什麼時候”當成一個詞，它的讀音是/dang-si/。 在句(l)中，我們把”地震”和”裂縫”都標上了音。 如果覺得”地震”不會被唸成別的音，也可以不標音。 如果你希望把”裂縫”唸成/bit-sun/，就可以把這個音標上去，因為它也可能會被唸成/li-pang/。 在句(m)中，”地動”是”地震”較早的漢字，以台語來說意音皆合。 但”地動”現在較不通用，若怕別人看不懂，可用括號註明是地震的意思。 如果不在意”裂縫”的唸法，可以不用標音。

以詞為書寫單位的另一個好處是可以免除”找字”的困擾。 例如”蟑螂”的台語音是/ga-zua/，以詞為書寫單位就是說”蟑螂”這個詞讀成/ga-zua/這個音；並不是說”蟑”讀成/ga/，而”螂”讀成/zua/。 以前的觀念是為每個台語音(音節、單音)去找字，例如為/ga-zua/的/ga/和/zua/各別去找字。 結果有人用了很不常見的字，更有人去造字；而這種字通常是很難流通的。 以”男生”這個詞來說，大家一看就知道它的意思，大部份會台語的人也知道要唸成/za-bo/。 可是如果要為台語的/za-bo/找字的話，結果就精采了，寫法有：查埔、查甫、者夫、諸父、諸夫、這夫、丈夫、乾夫、慈父、打捕等，甚至造字(打不出來)。到底哪一個才是/za-bo/呢？ 我們認為寫”男生”就可以了，再不然加上音寫成”男生/za-bo/”，雖然長一點，但是意和音都很清楚。

以詞為書寫單位還有一個好處是字的數目和音節的數目可以不一樣。例如”明天”的台語音是/mi-a-zai/，即兩個字對應到三個音節；我們不必有”明”唸成/mi/，而”天”唸成/a-zai/這類的對應。 再舉一例，”什麼時候”的台語是/dang-si/，台語的/dang-si/漢字可以寫成”什麼時候”，不必再去為/dang/尋找漢字。

§4.3. 外來語直接引用

我們主張文中如果有外來語時，應該盡量直接引用。 這樣做容易書寫又容易閱讀，也可以擴充我們的詞彙。 各語言可以獨立和外語溝通，不必透過另外一種語言，可以促進語言的地位平等。 不要再像以前一翻再翻的做法：以前有一位美國總統叫做 Reagon，國語翻成「雷根」，台語再唸成/lui-gin/，這樣的做法誤差太大。 我想當你跟美國人提起/lui-gin/的時候，他大概很難知道你是在說誰。 以下是我們的例子。

(n) 避免 SARS，請常洗手和量體溫。
(o) 美國總統 George Bush 關心台灣的安全。
(p) Toyota 的高級車 Lexus 風評很好。

(q) 我們實驗室有一台 HP 的 Server。

上述四句是國台語通用的例子。 句(n)的 SARS 若要寫成”嚴重急性呼吸道症候群”，恐怕是寫的人費力，讀的人也費時。 句(o)中，Bush 要翻成「布希」嗎？ 在中國是翻成「布什」。 句(p)中，Lexus 的中文又該是什麼？ 句(q)中，HP 有人或許還寫的出”惠普”，如果是 IBM 呢？ 要寫得出來就很難了，而且就算寫出來，看的人也不一定知道是 IBM。

§5. 書寫方式相關問題探討

若要書寫母語，會有用字如何選擇的問題。 要使用拼音，就要有拼音方式和聲調符號的設計。 種種這些相關問題，我們就分在不同的小節來探討。

§5.1. 本字

在書寫母語時，一般認為如果能找到本字是最理想的情況。 本字的意思是意音皆合的字。但是尋找本字並不是一件容易的事，不同人找到的本字可能差異甚大，例如「事情」一詞的台語，曾被提出的寫法就有「代誌」、「代志」、「載誌」、「事誌」、「事志」等。 我們認為如果有足夠語料的話，可以從中尋找最常用者推薦給大家。 如果沒有這類使用資訊，我們建議就寫成「事情/dai-ji/」，這樣意音皆備，大家也不用學那麼多新詞彙。 其實在大部分情形下（例如是在講白話音的情況下），「事情」並不會被念成/su-jing/，這時不標上音也沒關係。

其次我們認為尋找本字的時候，詞意是很重要的，不要因為要求音正確而妨礙了意。 例如曾看過一個主張是把「打人」的「打」寫成「拍」，說這樣音才對。 我們認為「打人」(可能會受傷)和「拍人」(可能是安慰或鼓勵)的意思差很多，寫成「拍人」賺到音卻賠了意，並不划算。

最後我們再提一個重要原則就是盡量要使用常用的字，最好是在一般常用的 5 千多字內去找本字，因為前面已經提過，我們用字的極限就差不多在 5 千多字。 千萬不可以造字，因為除非要大家再去學這些新字，根本無法流傳，而且在電腦上也很難處理。

§5.2. 拼音方式

以「共通」（即共用、流通）的要求來說，就是對於所要共通的語言（例如台灣漢語系的國、台、客語），要用同一套拼音方式。 拼音應該是相同的音要用相同的拼音符號，而不同的音要用不同的拼音符號（即一種拼音只對映到一種讀音）。 以共通的觀點而言，教育部在

2003年公佈的「中文譯音使用原則」和「客語拼音系統」，將拼音「p、t、k」對應到「ㄆ、ㄊ、ㄎ」。 而教育部在2006年公佈的「臺灣閩南語羅馬字拼音方案」，卻將拼音「p、t、k」對應到「ㄅ、ㄉ、ㄍ」[臺灣閩南語羅馬字拼音方案使用手冊 2006]。 這樣的拼音方式很難在國台客語共通使用。

拼音方式有兩大類，一類是對於需要拼音的對象（語言）另行創造拼音符號，例如我們有在使用的注音符號（ㄅㄆㄇㄈ…）就是特別設計來標記國語的音。 另一類是使用已有的符號來標記，例如前述用羅馬字來標記。 在使用已有符號來標記的方式中，我們又發現了兩種做法。 一種是接近英語的拼音，例如將拼音「p、t、k」對應到「ㄆ、ㄊ、ㄎ」；另一種是接近IPA(International Phonetic Alphabet，國際音標)的拼音，例如將拼音「p、t、k」對應到「ㄅ、ㄉ、ㄍ」。

我們認為拼音方式若能接近美式英語的發音，應該是最方便於在台灣的人使用，因為大家最熟悉。 依此觀點，我們也設計了一套拼音，稱為美式拼音。 國台語美式拼音的子音部份列在表5.1，其它若干拼音方式也列出來對照。

表5.1. 國台語美式拼音的子音表。台語羅馬字只列出教育部2006年公佈的部分。

注音符號	美式拼音	通用乙式	漢語拼音	注音二式	台語羅馬字
ㄅ	b	b	b	b	p
台帽	bh	bh			b
ㄆ	p	p	p	p	ph
ㄇ	m	m	m	m	m
ㄈ	f	f	f	f	
ㄉ	d	d	d	d	t
ㄊ	t	t	t	t	th
ㄋ	n	n	n	n	n
ㄌ	l	l	l	l	l
ㄍ	g	g	g	g	k
台鵝	gh	gh			g
ㄎ	k	k	k	k	kh
ㄏ	h	h	h	h	h
ㄐ	ji	ji	j	ji	
ㄑ	chi	ci	q	chi	
ㄒ	si	si	x	shi	
ㄓ	j	jh	zh	j	
ㄔ	ch	ch	ch	ch	
ㄕ	sh	sh	sh	sh	
ㄖ	r	r	r	r	
ㄗ	tz	z	z	tz	ts
ㄘ	ts	c	c	ts	tsh
ㄙ	s	s	s	s	s
台字人如	z	zz			j
零韻	-ih	-ih	-i	-ih	
台姆(台□)	m(mh)	m(mh)			
台秧(台嗯)	ng(ngh)	ng(ngh)			ng

我們提出的美式拼音是依照通用拼音乙式來修改，在子音部分共有 5 個不同，分述如下。我們用/chi/ 取代/ci/ 來標示ㄑ，用/j/ 取代/jh/ 來標示ㄓ，用/tz/ 取代/z/ 來標示ㄗ，用/ts/ 取代/c/ 來標示ㄘ，用/z/ 取代/zz/ 來標示台語「字人如」的子音部份。 這五個取代我們認為較符合美式英語的發音，前四個取代其實早在注音二式就已經發展出來了。 第五個取代使得美式拼音和通用乙式不相容(也和漢語拼音不相容)，因為/z/在通用乙式是用來標記ㄗ，若為了相容也可以不改變而維持/zz/。

§5.3. 聲調符號

聲調符號常見的有三種：數字編號、調值編號、和調形符號。 數字編號在我們的國語是編成第 1、2、3、4 聲；在台語是編成第 1、2、3、4、5、6、7、8 聲，其中第 2 聲和第 6 聲是同一個調。 因為在不同語言中編號相同的聲調調形(值)並不相同，例如國語第 2 聲聲調上揚，而台語第二聲聲調卻是下降，所以不容易編號成共通的形式。 當然硬要用編號也行（強力推動即可），可是從編號還是無法看出其調形。 因此我們認為數字編號不易達到共通的效果。

調值編號是把聲調對應到主音的 Do、Re、Mi、Fa、So，用簡譜 1、2、3、4、5 來代表，再把調值的變化情形寫出來，例如國語第三聲的調值符號是 214。 調值編號可以很準確的描述一個音節的聲調變化，也不會有不同語言編號不同的問題。 但是它的缺點是寫起來較麻煩，而且不同專家給的調值也有些許不同。 如表 5.2 所示的若干台語調值，幾乎沒有兩個人給的台語調值是相同的。 所以我們認為，調值編號也不容易達到共通的目的。

表 5.2. 若干台語調值。

來源	1,陰平	2,陰上	3,陰去	4,陰入	5,陽平	6,陽上	7,陽去	8,陽入
董峰政:通用拼音教材	55	53/51	22/21	32/33	24	53/51	33	44
吳守禮:台語注音符號的溯源	55	51	21/11	32	25	51	33	4/5
楊青矗:台華雙語詞典	44	52	21	32	24	52	33	43
王啓陽:台灣通用拼音	55	53	21	21	24	53	33	53

調形符號是把調值變化用簡單的上揚和下降等符號來表示，例如國語的二三四輕聲用ˊˇˋ•的符號來表示，這種方式的好處在於容易「望符生調」。 我們認為這種符號可以適用於各種語言，而且近似的調值可以用同一符號，經過適當的設計之後可以有良好的共通性。 以國、台語的整合而言，我們提出下述的設計。

1. 將國、台語近似的聲調用同一符號，我們認為下列配對的聲調很近似，可以用同一調形符號：（國一&台一）、（國二&台五）、（國三&台三）、（國四&台二）、（國輕&台四）。 也就是說國語一聲可以和台語一聲共用同一調形符號，依此類推。 若加上台語的一個特別高升的調，即三連音的第一個音，我們共需 8 個聲調符號。

2. 調形符號擺放的位置有兩類：和主要母音擺在一起（例如：把 bǎ）或和音節中的子母音分開放（例如：把 baˇ）。 我們認為應該和子母音分開放，因為聲調作用的範圍是整個音節，不是只在於主要元音。 而且分開放的時候空間較大（高度較夠），容易設計調形。理想上應該是像ˇ（國語三聲）的符號，但是設計成半形字，比較美觀。 在 Unicode 中，我們找到了意義符合調形符號的碼，以下將它們的字型(Lucida Sans Unicode)、Unicode

的十六進位碼、和 Unicode 中的意義列出。

（國一&台一）：¯(02C9): High level tone, Mandarin Chinese first tone、

（國二&台五）：ˊ(02CA): Mandarin Chinese second tone、

（國三&台三）：ˇ(02C7): Mandarin Chinese third tone、

（國四&台二）：ˋ(02CB): Mandarin Chinese fourth tone、

（國輕&台四）：˙(02D9): Mandarin Chinese fifth tone、

（台七）：_(02CD): Low level tone、

（台八）：˚(02DA): Ring above、

（台高升）：'(02C8): Primary Stress, downstep。

在上面的 Unicode 中，只有標示台語第八聲的意義不太一致，但它的字型很容易和台語第四聲區別。 我們只要仔細為這些碼設計一下合適的半形字體，就可以有很好的調形符號。 在字型尚未設計好之前，就先用 Lucida Sans Unicode 的字型。

3. 我們主張任何一個聲調都應該標示調號。 現在有些拼音方式一聲不標調號的做法會產生混淆。 不標調號應該表示調號沒標出，任一聲調都有可能。 我們認為將來中文會有更多的多字詞(多音節詞)，對許多多字詞而言，不標調號對讀音的判別並不會造成困擾，因此會有作者選擇不標聲調，因此我們認為沒標調號不應該用來表示標成一聲。 而且當國台客語混合書寫的時候，要標音調難度還蠻高的。

我們在附錄 A 列出若干台語八音的拼音例子。 用的是我們主張的拼音符號和聲調符號，希望能增進大家的了解。 我們所主張的美式拼音的五個和通用乙式不同的子音有四個列在附錄 A，其中/j/沒列出，因為/j/代表國語的ㄓ，在台語中沒有此音。

§6. 結語

依照我們所提議的書寫方式，目前台灣各漢語系(國語、台語、客語)都可以用一套共同的寫法來描述。 學會一種語言之後，要再學習另一種語言會變的很容易。 這對於了解其他族群的文化，促進族群的和諧會有很大的作用。 至於原住民各族的語言，因為原住民語言目前基本上是「有音無字」，所以我們認為可以在和漢語系語言拼音相容的原則下，設計出原住民語言的拼音系統。

照我們所設計的台灣共通語言的書寫方式，我們可以發展出一套容易學、容易寫、容易了解的漢字書寫系統。 這種簡單易用的特性，不只台灣人愛用，相信海外華人，乃至於外國人在學習中文時都較能接受。 我們期待這樣的書寫方式可以讓中文更為普及。

參考文獻

1. 教育部異體字字典，正式五版，2004。 教育部國語推行委員會 編著，http://140.111.1.40/bian/fbian5.htm。
2. 曾志朗，2002，”人腦可突破認知極限？”科學人，2002 年 5 月號。
3. 張晴雨、李翰祥編著，2004，【人生不加框：會思考才會贏】，培育文化。
4. 楊維哲，”台語之美：建設台灣語文的運動”，1997。 自立晚報 1997 年 1 月 22 日第 3 版。
5. 蔣為文，”救救我們的孩子：從人文的觀點談台灣需要的語言文字改革”。
6. 臺灣閩南語羅馬字拼音方案使用手冊，2006。 教育部國語推行委員會，http://www.edu.tw/EDU_WEB/Web/MANDR/index.php。

附錄 A. 子音為/chi、tz、ts、z/的一些台語字。 □代表作者沒找到該音的漢字。

蚩 chi¯ 取 chi` 試 chi˘ 閃 chih˙ 持 chi´ 取 chi` 市 chi_ 扼 chih°

千 ching¯ 請 ching` 槍 ching˘ 冊 chik˙ 榕 ching´ 請 ching` 穿 ching_ 摶 chik°

充 chiong¯ 昶 chiong` 倡 chiong˘ 促 chiok˙ 戕 chiong´ 昶 chiong` 匠 chiong_ 搦 chiok°

叉 tsa¯ 吵 tsa` 吒 tsa˘ 插 tsah˙ 材 tsa´ 吵 tsa` 喳 tsa_ 眨 tsah°

摻 tsam¯ 慘 tsam` 讒 tsam˘ 插 tsap˙ 慚 tsam´ 慘 tsam` 潛 tsam_ 眨 tsap°

餐 tsan¯ 孱 tsan` 粲 tsan˘ 察 tsat˙ 田 tsan´ 孱 tsan` 孱 tsan_ 賊 tsat°

叉 tsei¯ 扯 tsei` 刷 tsei˘ 冊 tseih˙ 蹉 tsei´ 扯 tsei` 找 tsei_ 唶 tseih°

查 tza¯ 早 tza` 炸 tza˘ 帶 tzah˙ 查 tza´ 早 tza` 攔 tza_ 昨 tzah°

沾 tzam¯ 斬 tzam` 跺 tzam˘ 雜 tzap˙ 巉 tzam´ 斬 tzam` 站 tzam_ 十 tzap°

曾 tzan¯ 囋 tzan` 棧 tzan˘ 節 tzat˙ 殘 tzan´ 囋 tzan` 綻 tzan_ 塞 tzat°

遭 tze¯ 佐 tze` 作 tze˘ 作 tzeh˙ 皂 tze´ 佐 tze` 坐 tze_ 昨 tzeh°

這 tzei¯ 姐 tzei` 制 tzei˘ 節 tzeih˙ 齊 tzei´ 姐 tzei` 坐 tzei_ 絕 tzeih°

煎 tzein¯ 剪 tzein` 荐 tzein˘ 折 tzeit˙ 前 tzein´ 剪 tzein` 賤 tzein_ 捷 tzeit°

租 tzo¯ 祖 tzo` 阼 tzo˘ 作 tzoh˙ 皂 tzo´ 祖 tzo` 助 tzo_ 昨 tzoh°

妝 tzong¯ 總 tzong` 葬 tzong˘ 作 tzok˙ 崇 tzong´ 總 tzong` 狀 tzong_ 昨 tzok°

專 tzuan¯ 轉 tzuan` 篆 tzuan˘ 拙 tzuat˙ 全 tzuan´ 轉 tzuan` 撰 tzuan_ 絕 tzuat°

尊 tzun¯ 准 tzun` 俊 tzun˘ 卒 tzut˙ 船 tzun´ 准 tzun` 陣 tzun_ 拭 tzut°

遮 zia¯ 惹 zia` □zia˘ 跡 ziah˙ □zia´ 惹 zia` 偌 zia_ □ziah°

□zim¯ 忍 zim` 任 zim˘ 追 zip˙ 壬 zim´ 忍 zim` 刃 zim_ 入 zip°

□ziong¯ 冗 ziong` 釀 ziong˘ 追 ziok˙ 茸 ziong´ 冗 ziong` 讓 ziong_ 弱 ziok°

中文詞義全文標記語料庫之設計與雛形製作

[1]柯淑津、[2]黃居仁、[3]洪嘉馡、[1]劉詩音、[1]簡卉伶、[2]蘇依莉

[1]東吳大學資訊科學系

{ksj, ms9405, ms9504}@cis.scu.edu.tw

[2]中央研究院語言所

{churen, isu}@gate.sinica.edu.tw

[3]台灣大學語言學研究所

jiafei@gate.sinica.edu.tw

摘要

詞義標示語料庫對自然語言處理佔有很重要的地位，尤其反映在訊息特徵及自然語言理解之研究上，但目前大規模之中文詞義標示語料庫卻付之闕如。本文設計出一個超過 11 萬詞的大規模中文詞義全文標示語料集，以中研院平衡語料庫爲標示對象，從中摘錄 56 篇完整文章，利用 N-Gram 與搭配資訊等語言知識，並結合機器學習技巧以及機率模式的方式作爲處理自動詞義標示的前置作業工作，最後爲達高精確度之效果，再將自動產生之標示結果經由人工校訂而成。

關鍵詞：詞義辨識，詞義標記語料庫，自然語言處理，誘導式作法

Keywords: Word Sense Disambiguation, Sense Tagged Corpus, Natural Language Processing, Bootstrap Method

一、簡介

語料庫對自然語言處理研究佔有相當重要的地位。尤其是統計式計算語言處理，常需仰賴語料庫所蘊藏的豐富資源，作爲計算的依據。隨著數位文獻之普及，語料庫的種類越來越多，內容也越來越豐富。標示訊息越完整的語料庫對研究的幫助越大。有些語料庫只呈現原始文本內容，有的則再加上詞性標示或詞義等相關資料。目前，標示詞性的語料庫有不少，例如：中文方面有中研院的一千萬詞平衡語料庫[1]及中文十億詞語料庫(Chinese Gigaword Corpus)[2]等，至於標示詞義的語料庫不論中英文都很少。標示詞義(Sense)或語意(Semantic)的語料資源，在理論語言學上，將提供詞彙語意學研究之豐富材料與基本架構；在計算語言學上，將對自然語言處理的核心工作如：WSD 多重詞義辨析研究、自然語言理解等，都將會有關鍵性的突破。另外，這些標示語料經統計處理後所獲得之資訊，可應用於資訊檢索、資訊擷取、文件摘要、自動問答等議題之研究。

大量精確的詞義標示資料，可提供多項計算語言相關研究的豐富素材。但是，中文語料庫詞義標記主要的瓶頸爲缺乏足供自動標記參考的資料，而人工標示需要昂貴成本，造成語料庫標示語意工作的難產。近年來的許多研究，顯示出對大規模詞義標示語料集的大量需求，這些資源在建構上是否完備，往往會影響整個研究進行方向以及研究結果的正確性。在某些語言上，已具有較代表性的詞義標示語料集存在，例如英文語料集 SemCor [3]以及 Senseval [4] 提供之多國語言的全文標示語料集，例如捷克語、荷蘭語，義大利文及英文。反觀中文，目前大量中文標示詞義語料集卻一直付之闕如，只有少數幾個規模不大的中文詞義標示語料集，例如 Senseval-2 的中文詞義標示語料集包括 15 個中文詞彙的詞義標示，Senseval-3 的中文詞義標示語料集包含 20 個中文詞，其規模與真實的語言環境存在相當大的差異。製作大規模語料庫所遇到最大瓶頸爲成本，藉由人工標示雖然可以得到高正確率，但所需的詞義標示成本卻非常昂貴，而且可以標示詞義的專家也不多見。爲了克服此問題，本文提出一套半自動詞義標示方法，作爲標示詞義的前置作業，再經由專門人士校訂。語料庫製作以中研院平衡語料庫爲對象，從中摘錄文章，並對摘錄出之文章中之詞做詞義標示的動作，設計製作出一個大規模的中文詞義標示語料集以供自然語言處理研究使用。

二、詞義區分詞典

本文標記中文詞義語料庫所使用的詞義區分詞典爲中央研究院資訊所、語言所詞庫小組所製作的『詞義區辨小辭典』第三版[5]，詞典的內容以中頻詞爲主，共包含 5047 個詞形，9400 個詞義。詞典所收錄的詞條(entry)，以現代漢語通用語詞爲範圍，不列入現今已不用或罕用的詞彙。而收錄的中文詞彙條目，包含單字詞、雙字詞和多字詞。詞典中提供各詞條豐富的訊息，除詞目(lemma)、標音（拼音與注音）、義項、詞類、例句等內容外，還包括有各詞義對應至英文詞網 WordNet 2.0 (http://wordnet.princeton.edu/)之同義詞集及其編號。圖一爲詞彙「瘋狂」在字典中的訊息，共有兩個詞義，其中第一個詞義「形容人因精神錯亂而舉止失常。」對應於詞網的同義詞集{crazy}，第二個詞義「形容行爲或事物無節制，超乎平常的程度。」對應於詞網的同義詞集則爲{madly}，「瘋狂」的兩個詞義各自都可又細分爲兩個義面。

三、標示語料來源

本文使用『中央研究院現代漢語平衡語料庫』 (Sinica Corpus) [1] 作爲語料標的。語料中的每個文句都已依詞斷開，並標示詞性。本研究爲求表達出文脈結構與前後文關係等訊息之完整性，標示語料的選擇以全文爲單位。詞義的標示以全文標示爲原則，但因本文所使用的詞義區分字典目前仍在編撰中，因此並非每個出現在語料庫中的詞彙都收錄在詞義區分字典裡，對於字典尚未收錄之詞彙，我們以其詞性做爲標示。標記詞性可以降低詞義的岐義度，因此詞性的標示可視爲粗的詞義標記。

我們依字典詞彙出現於文章內容中的覆蓋率以及文章長度作爲選擇之依據，共選出 56 篇文章，總長度爲 114,066 個詞彙，148,863 個字元。語料集中的文章主題分佈統計結果如表一，以文章數目計算，文學類最多共有 35 篇, 若以文章長度計算，則以生活類的佔有率最高。

瘋狂 feng1 kuang2 ㄈㄥ ㄎㄨㄤˊ

詞義1：【不及物動詞，VH；名詞，nom】形容人因精神錯亂而舉止失常。｛crazy, 00872382A｝

義面1：【不及物動詞，VH】形容人因精神錯亂而舉止失常。

例句：片中一名〈瘋狂〉殺手，拿著剃刀。

例句：石門五子命案的父母與其說是迷信，不如說是〈瘋狂〉。

義面2：【名詞，nom】形容人因精神錯亂而舉止失常。

例句：因此，石門五子命案的〈瘋狂〉，其實也正是我們社會瘋狂的一粒種籽啊！

詞義2：【不及物動詞，VH；名詞，nom】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。｛madly, 00045197R｝

義面1：【不及物動詞，VH】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。

例句：他〈瘋狂〉的愛上一個女孩子。

例句：每年都有不計其數的台灣客前往香港〈瘋狂〉大採購。

例句：當時少棒青少棒在台灣很〈瘋狂〉，連我們城市的小孩子也愛打棒球。

義面2：【名詞，nom】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。

例句：經過一陣〈瘋狂〉後，大家都累了，個個都喊著喉嚨痛、腳痛。

例句：死了七九百餘人的人民教室案，也使人想到愈來愈多的宗教〈瘋狂〉事件。

例句：只要幅度不超過，則多頭仍然大有可爲，但仍切忌一味追高的〈瘋狂〉舉動。

圖一 詞彙「瘋狂」在『詞義區辨小辭典』詞典範例

表一、語料集所含文章主題分佈

主題	文章篇數	文章長度	
		詞彙	字元
哲學	4	1451	1976
社會	5	27385	35918
生活	12	57605	74710
文學	35	27625	36259
總計	56	114066	148863

整體而言我們標示詞義之目標詞彙，以單一詞性(詞形與詞性之配對)爲單位做統計，被收錄於字典之單義詞共 863 個，出現頻率爲 12,124 次，多義詞有 650 個，出現頻率 23,521 次，統計其詞性分佈如表二。多義詞詞義數量由 2（如：自然 D、堆 Nf 和喜 VK 等）至 27(如：吃 VC）不等，平均詞義數爲 2.97。以詞性區分，平均而言詞義數最多之詞性爲語助詞，平均達 4.83 個詞義數。詞義數最少之詞性爲感嘆詞，平均爲 1.32 個詞義數。若是以詞形爲單位做統計，不考慮其詞性之差異，在本研究使用的 56 篇文件中共有 598 個詞形收錄於詞典中，每個詞形的平均詞義數爲 4.53。

表二、標示語料集詞性分佈

詞類	詞彙數	詞例數	範例
不及物動詞	231	3,317	對 VH, 跑 VA, 走 VA
介詞	51	1,854	在 P, 跟 P, 到 P
及物動詞	373	5,733	說 VE, 沒有 VJ, 開始 VL
名詞	321	5,070	人家 Nh, 感覺 Na, 下 Ncd
形容詞	21	45	一般 A, 原 A, 定期 A
定詞	55	3,175	那 Nep, 前 Nes, 多 Neqa
後置詞	31	455	上 Ng, 裡 Ng, 當中 Ng
副詞	287	8,892	就 D, 又 D, 起來 Di
連接詞	69	1,554	就是 Cbb, 而 Cbb, 或 Caa
量詞	81	976	回 Nf, 份 Nf, 間 Nf
語助詞	47	4,574	啊 T, 喔 T, 哇 I
總數	1567	35,645	

四、詞義標示語料庫

我們製作的詞義標示語料集的型態爲 XML 格式檔，所使用的標籤結構如圖二所示，標籤使用說明請見表三。語料庫內含多篇文件，每篇文件以<doc>標籤區隔，也就是在標籤<doc>及</doc>範圍內容爲同一篇文件。文件內容再往下細分爲句子，每個句子以<sent>標籤區隔，句子內容依詞彙出現順序呈現，其間以<w>標籤作爲詞彙區隔，其內又再細分爲三個標籤：word, pos, tag1 分別呈現詞彙、詞性、以及詞義標示等資訊。

在詞義標示部分，依標示類別分爲三種，第一種爲詞義代碼標示，採用 Huang et al. [6]之定義，爲四位數整數，前兩位爲詞義序號，表明標示詞義出現在字典中之詞義順序。第三碼是詞形標碼，第四碼爲義面編碼（如圖三）。第二種爲標點符號之處理，對於標點進行標示詞義，不具意義，因此，我們將標點符號的詞義代碼直接設定爲其符號本身。至於，第三種是針對未知詞（包含辭典未收錄、尚未有詞義分析之詞彙）的部分，我們以該詞彙之詞性作爲其詞義標示。圖四是部分標示語料範例，爲語料庫中編號 101664 之文章第 18 句內容，第一個詞『灰灰』，是未知詞，標示詞義爲其詞性「Nb」。第二個詞『說』的詞義標示爲「0111」，表示在此處詞義「以口語媒介引述或陳述訊息。」爲「說 1」的第 01 個詞義的第 1 個義面。另外，第 3, 5, 8 個詞是標點符號，因此詞義代碼爲符號本身。

整個標示語料庫共含有 114066 個詞彙，我們依標示種類作爲統計，其結果如表四。其中標點符號有 27530 個，成功標示出詞義代碼之詞彙數共有 35645 個，對於未知詞或是字典尚未收編處理之詞，我們以詞性作爲詞義標碼的則有 50891 個詞例。我們進一步分析發現這部分資料，包括有：文章中有些英數字或是專名，例如：(二)、CPU、清華大學等，在我們的標示語料庫中共有 4258 個詞例。另外，因爲本研究所使用之詞義區分字典尚在建構中，有些詞彙目前未收錄於我們的詞義字典中，這部分共有 4541 個詞

彙，在語料中共出現 31730 次。至於，剩餘的 14903 個詞例，雖然是已收錄於字典中之詞彙，但是不在本次規劃標記範圍內，因此先以詞性標記，加上詞性標記可降低多義詞之岐義度。

表三、語料庫使用標籤說明

標籤名稱	內容說明	例子
<corpus>	語料庫起始	<corpus>
<doc id=>	文件起始及編號	<doc id="100863">
<sent id=>	句子起始及編號	<sent id="1">
<w id=>	詞彙資料及編號	<w id="1">
<word>	詞彙	<word>人家</word>
<pos>	詞性	<pos>Nh</pos>
<tag1>	詞義標示	<tag1>0122</tag1>

表四、 語料集標示種類統計

標示種類	詞彙數		說明
標點符號	27530		標點符號無需標示詞義
詞義代碼	35645		已完成詞義標示
詞性標示	50891	4258	不需標示（英數字、專名）
		31730	詞彙(4541 個)目前未收錄於字典中
		14903	未處理
總和	114066		

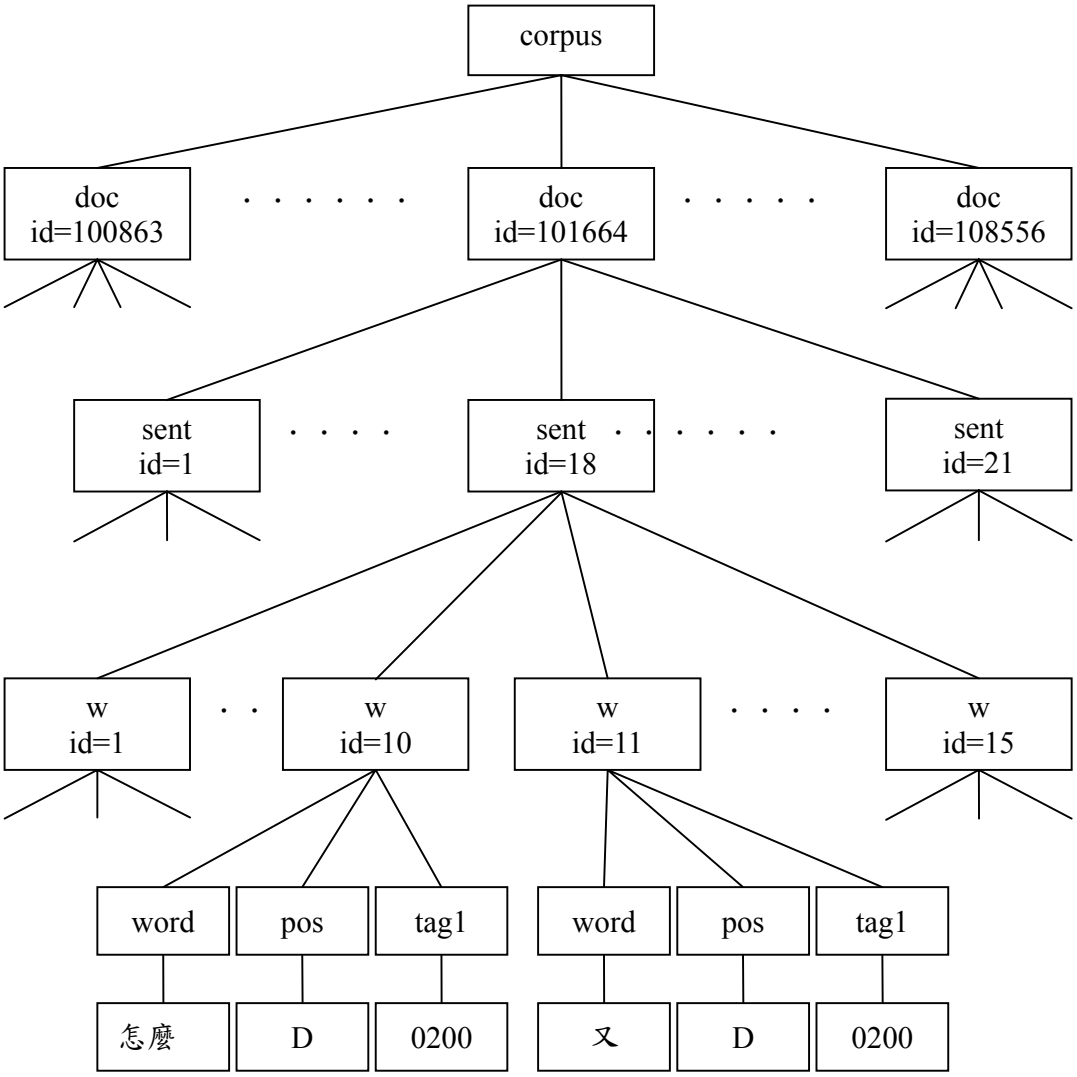
五、詞義標示方法

標示詞義之工作需仰賴大量的人力，因此，爲了節省成本，本文設計出一套半自動標示詞義之方法[7]，先利用此標示方法對語料作初步的詞義標示處理，以作爲人工標示之前置作業。

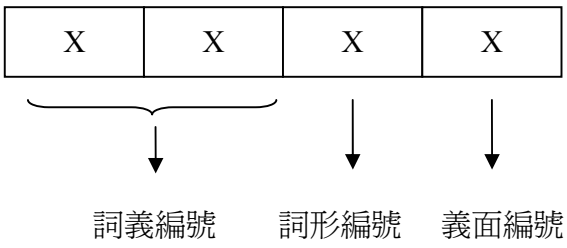
我們所設計之半自動標示詞義的方法，採用誘導式方法（bootstrap）逐步放寬標示條件，來擴增標示語料，其系統組織圖如圖五所示。首先蒐集一些已被標示過詞義的資料作爲詞義標示工作的種子訓練資料。其來源主要有兩個部分，第一個部分爲詞義區分詞典中之例句，第二個部分爲辭典編撰小組，在搜尋整理詞義過程中所標示的語料庫部分內容。若來自這兩部分的例句數量不足時，我們會隨機.從研究院語料庫中選出部分文句，交由人工標示詞義後加入成爲種子標示句。將上述已標示資料合爲訓練集，以本文選出來的 56 篇標示集文章，則作爲測試集。

自動標示詞義的第一階段採用 N-gram 模式，將標示出詞義的資料加入訓練集中，以作爲第二階段的訓練語料。本文利用 N-gram 處理詞義標示是基於下面的假設：存在

包圍目標詞彙前後 N 個詞彙完全相同的兩個子句，我們推論它們應擁有一樣的詞義[8]。在此使用 N-gram 有兩項主要目的，第一是擴大訓練集，因語料庫中常可見到相似之子句。第二個目的是過濾訓練資料集的雜訊，以此檢驗人工標示資料之不一致性。



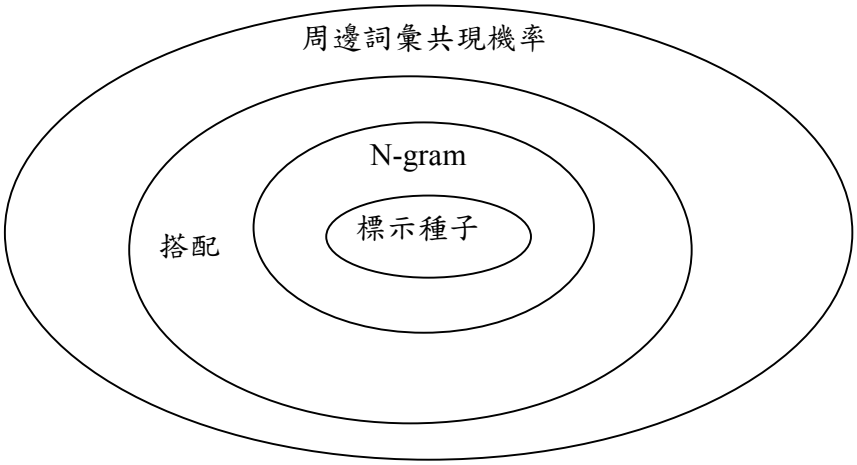
圖二、語料集標籤結構及範例



圖三、詞義代碼編碼方式說明

```
<doc id="101664">
  <sent id="1">
        :
        :
  <sent id="18">
    <w id="1"> <word>灰灰</word> <pos>Nb</pos> <tag1> Nb </tag1> </w>
    <w id="2"> <word>說</word> <pos>VE</pos> <tag1>0111</tag1> </w>
    <w id="3"> <word>：</word> <pos>COLONCATEGORY</pos> <tag1>：</tag1> </w>
    <w id="4"> <word>白白</word> <pos> Nb </pos> <tag1> Nb </tag1> </w>
    <w id="5"> <word>，</word> <pos>COMMACATEGORY</pos> <tag1>，</tag1> </w>
    <w id="6"> <word>剛剛</word> <pos>D</pos> <tag1>D</tag1> </w>
    <w id="7"> <word>見面</word> <pos>VA</pos> <tag1>VA</tag1> </w>
    <w id="8"> <word>，</word> <pos>COMMACATEGORY</pos> <tag1>，</tag1> </w>
    <w id="9"> <word>你</word> <pos>Nh</pos> <tag1>Nh</tag1> </w>
    <w id="10"> <word>怎麼</word> <pos>D</pos> <tag1>0200</tag1> </w>
    <w id="11"> <word>又</word> <pos>D</pos> <tag1>0200</tag1> </w>
    <w id="12"> <word>要</word> <pos>D</pos> <tag1>D</tag1> </w>
    <w id="13"> <word>走</word> <pos>VA</pos> <tag1>0400</tag1> </w>
    <w id="14"> <word>了</word> <pos>Di</pos> <tag1>0110</tag1> </w>
    <w  id="15">  <word> ？ </word>  <pos>QUESTIONCATEGORY</pos>  <tag1> ？ </tag1>
</w>
  </sent>
</doc>
```

圖四、XML 格式之標示語料範例



圖五、標示詞義系統組織圖

第二個階段我們使用搭配資訊來增加標示集數量，搭配資訊是一種很強的語言關係，能決定目標詞彙之詞義[9]。我們先以詞頻、搭配詞與目標詞彙距離變異量等條件作爲選擇搭配詞彙之初步依據，最後再經過相互資訊 MI 計算來檢驗搭配詞與目標詞彙之間的相關程度。

經過 N-gram 及搭配資訊兩個階段的處理工作，我們將訓練語料標示量做了實質擴增。接著，再經過機率模式計算，盡可能將大部分詞彙標上詞義資訊。最後爲求標示語料之高精準度，我們將經由自動標示詞義處理過後的整個標示語料，再交由原字典編撰小組成員進行人工校正處理。

整個自動標示部分之實驗結果我們分爲兩部分說明，第一部份詞義標示以詞義下再細分至義面爲準，結果如表五所示，整體的正確率爲 57.47%。至於，第二部分我們將詞義標示處理至詞義爲止，不再細分義面，結果如表六所示，整體的正確率可提升至 64.51%。

六、結論

詞義標示語料庫對自然語言處理佔有很重要的地位，尤其反映在計算語言學研究上，常需語料庫所提供的豐富資訊來作計算，但目前存在的中文詞義標示語料集的數量少之又少，因此，我們設計出一個包含約十萬詞大規模的中文詞義標示語料集，以供自然語言處理相關研究使用。標示詞義之步驟爲先使用自動詞義標示作爲人工詞義標示之前置工作，再將結果交由人工校訂。自動詞義標示方法爲利用周邊詞彙提供的資訊，經由 N-gram，搭配資訊以及機率模式計算出最有可能的詞義。未來藉由詞義區分詞典的漸趨完備，期望能達到對中央研究院現代漢語平衡語料庫五百萬詞全文標記的目標。

表五 詞義標示至義面為準的實驗結果

詞性	詞彙數	詞例數	詞例比率	正確詞例	錯誤詞例	正確%	錯誤%
A	6	22	0.10%	14	8	63.64%	36.36%
Caa	2	38	0.16%	37	1	97.37%	2.63%
Cbb	7	231	1.00%	37	194	16.02%	83.98%
D	64	3454	14.98%	2135	1319	61.81%	38.19%
Da	7	22	0.10%	21	1	95.45%	4.55%
Dfa	5	202	0.88%	200	2	99.01%	0.99%
Dfb	1	1	0.00%	1	0	100.00%	0.00%
Di	7	1146	4.97%	934	212	81.50%	18.50%
Dk	2	5	0.02%	5	0	100.00%	0.00%
I	15	693	3.00%	307	386	44.30%	55.70%
Na	98	648	2.81%	563	85	86.88%	13.12%
Nb	5	18	0.08%	18	0	100.00%	0.00%
Nc	8	29	0.13%	27	2	93.10%	6.90%
Ncd	13	283	1.23%	209	74	73.85%	26.15%
Nep	6	2227	9.66%	615	1612	27.62%	72.38%
Neqa	2	38	0.16%	32	6	84.21%	15.79%
Nes	6	128	0.56%	118	10	92.19%	7.81%
Neu	3	127	0.55%	114	13	89.76%	10.24%
Nf	40	228	0.99%	212	16	92.98%	7.02%
Ng	13	147	0.64%	102	45	69.39%	30.61%
Nh	8	1668	7.23%	140	1528	8.39%	91.61%
P	33	1659	7.19%	1136	523	68.48%	31.53%
T	13	2838	12.30%	1648	1190	58.07%	41.93%
VA	28	451	1.96%	328	123	72.73%	27.27%
VAC	1	4	0.02%	3	1	75.00%	25.00%
VB	9	14	0.06%	14	0	100.00%	0.00%
VC	76	1177	5.10%	1061	116	90.14%	9.86%
VCL	5	174	0.75%	103	71	59.20%	40.80%
VD	19	170	0.74%	128	42	75.29%	24.71%
VE	26	1703	7.38%	1370	333	80.45%	19.55%
VF	5	20	0.09%	11	9	55.00%	45.00%
VG	9	170	0.74%	103	67	60.59%	39.41%
VH	66	1940	8.41%	661	1279	34.07%	65.93%
VHC	2	13	0.06%	13	0	100.00%	0.00%
VI	3	4	0.02%	4	0	100.00%	0.00%
VJ	19	326	1.41%	206	120	63.19%	36.81%
VK	11	63	0.27%	55	8	87.30%	12.70%
VL	5	160	0.69%	140	20	87.50%	12.50%
V_2	1	823	3.57%	433	390	52.61%	47.39%
nom	1	1	0.00%	1	0	100.00%	0.00%
	650	23065	100.00%	13259	9806	57.47%	42.51%

表六 詞義標示僅處理至詞義爲準的實驗結果

詞性	詞彙數	詞例數	詞例比率	正確詞例	錯誤詞例	正確%	錯誤%
A	6	22	0.10%	14	8	63.64%	36.36%
Caa	2	38	0.16%	37	1	97.37%	2.63%
Cbb	7	231	1.00%	37	194	16.02%	83.98%
D	64	3454	14.98%	2158	1296	62.48%	37.52%
Da	7	22	0.10%	21	1	95.45%	4.55%
Dfa	5	202	0.88%	200	2	99.01%	0.99%
Dfb	1	1	0.00%	1	0	100.00%	0.00%
Di	7	1146	4.97%	934	212	81.50%	18.50%
Dk	2	5	0.02%	5	0	100.00%	0.00%
I	15	693	3.00%	307	386	44.30%	55.70%
Na	98	648	2.81%	570	78	87.96%	12.04%
Nb	5	18	0.08%	18	0	100.00%	0.00%
Nc	8	29	0.13%	27	2	93.10%	6.90%
Ncd	13	283	1.23%	209	74	73.85%	26.15%
Nep	6	2227	9.66%	642	1585	28.83%	71.17%
Neqa	2	38	0.16%	32	6	84.21%	15.79%
Nes	6	128	0.56%	118	10	92.19%	7.81%
Neu	3	127	0.55%	114	13	89.76%	10.24%
Nf	40	228	0.99%	212	16	92.98%	7.02%
Ng	13	147	0.64%	102	45	69.39%	30.61%
Nh	8	1668	7.23%	1549	119	92.87%	7.13%
P	33	1659	7.19%	1143	516	68.90%	31.10%
T	13	2838	12.30%	1660	1178	58.49%	41.51%
VA	28	451	1.96%	347	104	76.94%	23.06%
VAC	1	4	0.02%	3	1	75.00%	25.00%
VB	9	14	0.06%	14	0	100.00%	0.00%
VC	76	1177	5.10%	1065	112	90.48%	9.52%
VCL	5	174	0.75%	107	67	61.49%	38.51%
VD	19	170	0.74%	128	42	75.29%	24.71%
VE	26	1703	7.38%	1475	228	86.61%	13.39%
VF	5	20	0.09%	11	9	55.00%	45.00%
VG	9	170	0.74%	103	67	60.59%	39.41%
VH	66	1940	8.41%	664	1276	34.23%	65.77%
VHC	2	13	0.06%	13	0	100.00%	0.00%
VI	3	4	0.02%	4	0	100.00%	0.00%
VJ	19	326	1.41%	206	120	63.19%	36.81%
VK	11	63	0.27%	55	8	87.30%	12.70%
VL	5	160	0.69%	140	20	87.50%	12.50%
V_2	1	823	3.57%	433	390	52.61%	47.39%
nom	1	1	0.00%	1	0	100.00%	0.00%
Total	650	23065	100.00%	14879	8186	64.51%	35.49%

參考文獻

[1] Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. *Proceeding of the 11 th Pacific Asia Conference on Language, Information and Computation.* Seoul:Kyung Hee University. pp.167-176.

[2] Wei-yun Ma, and Chu-Ren Huang. 2006. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Presented at the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy. 24-28 May.

[3] SemCor, http://multisemcor.itc.it/semcor.php

[4] Senseval, http://www.senseval.org/

[5] 黃居仁，主編。中文的意義與詞義。中央研究院文獻語料庫與詞庫小組技術報告 06-03。南港，中研院, 2006。

[6] Huang, Chu-Ren**,** Chun-Ling Chen, Cui-Xia Weng, Hsiang-Ping Lee, Yong-Xiang Chen and Keh-jiann Chen. 2005. The Sinica Sense Management System: Design and Implementation, *Computational Linguistics and Chinese Language Processing.* Vol. 10, No. 4, pp. 417-430.

[7] 柯淑津、黃居仁、陳振南，2004，全語料庫中文詞義標記的初步研究，第五屆詞彙語意研討會，北京。

[8] 柯淑津、陳振南，2007, 結合機器學習與語言知識的全語料庫中文詞義標示方法,第八屆詞彙語意研討會，香港。

[9] Yarowsky, 1993, One Sense Per Collocation, In Proceedings of ARPA Human Language Technology Workshop, Princeton.

以文件分類技術預測股價趨勢

Predicting Trends of Stock Prices with Text Classification Techniques

陳俊達 Jiun-Da Chen
國立政治大學資訊科學系
Department of Computer Science
National Chengchi University
g9414@cs.nccu.edu.tw

王台平 Tai-Ping Wang
真理大學資訊管理學系
Department of Information Management
Aletheia University
tpwang@email.au.edu.tw

劉昭麟 Chao-Lin Liu
國立政治大學資訊科學系
Department of Computer Science
National Chengchi University
chaolin@nccu.edu.tw

摘要

股價的漲跌變化是由於證券市場中眾多不同投資人及其投資決策後所產生的結果。然而，影響股價變動的因素眾多且複雜，新聞也屬於其中一種，新聞事件不但是投資人用來得知該股票上市公司的相關營運資訊的主要媒介，同時也是影響投資人決定或變更其股票投資策略的主要因素之一。本研究提出以新聞文件做爲股價漲跌預測系統的基礎架構，透過文字探勘技術及分類技術來建置出能預測當日個股收盤股價漲跌趨勢之系統。

本研究共提出三種分類模型，分別是簡易貝氏模型、*k* 最近鄰居模型以及混合模型，並設計了三組實驗，分別是分類器效能的比較、新聞樣本資料深度的比較、以及新聞樣本資料廣度的比較來檢驗系統的預測效能。實驗結果顯示，本研究所提出的分類模型可以有效改善相關研究中整體正確率高但各個類別的預測效能卻差異甚大的情況。而對於影響投資人獲利與否的關鍵類別"漲"及類別"跌"的平均預測效能上，本研究所提出的這三種分類模型亦同時具有良好的成效，可以做爲投資人進行投資決策時的有效參考依據。

Abstract

Stocks' closing price levels can provide hints about investors' aggregate demands and aggregate supplies in the stock trading markets. If the level of a stock's closing price is higher than its previous closing price, it indicates that the aggregate demand is stronger than the aggregate supply in this trading day. Otherwise, the aggregate demand is weaker than the aggregate supply. It would be profitable if we can predict the individual stock's closing price level. For example, in case that one stock's current price is lower than its previous closing price. We can do the proper strategies(buy or sell) to gain profit if we can predict the stock's closing price level correctly in advance.

In this paper, we propose and evaluate three models for predicting individual stock's closing price in the Taiwan stock market. These models include a naïve Bayes model, a *k*-nearest neighbors model, and a hybrid model. Experimental results show the proposed methods perform better than the NewsCATS system for the "UP" and "DOWN" categories.

關鍵詞：股價預測，簡易貝氏模型，*k* 最近鄰居模型，混合模型。

Keywords: Stock Price Prediction, naïve Bayesian models, *k*NN models, hybrid models.

一、緒論

股價漲跌趨勢的預測是個令人感興趣的研究議題，然而影響股價變動的因素眾多且複雜。許多的相關研究使用技術分析或基本分析法[11]來做爲股價趨勢預測的特徵項目選取方式[6][8][12]。基本分析法著重於長期面的經濟因素變化，而對短期的證券市場的變動較不在乎。技術分析則著重於證券市場本身的變化，主要是透過圖表或技術指標的歷史資料及研究分析，從中找出規則並藉此來預測未來股價可能的趨勢變化；而不考慮其他可能也會對股價產生某種程度影響的外部因素，如經濟、政治、國際情勢…等其他各種方面的潛在影響。

對投資大眾而言，新聞是日常生活中非常容易接觸到的一種資訊來源；新聞是屬於會影響股價變化的非結構化資料，也是投資人可以用來得知該股票上市公司的相關資訊的主要媒介之一。新聞事件的本身也是影響投資人決定或變更其股票投資策略的其中一種考量因素；使其投資策略由賣方變爲買方，或由買方變爲賣方，導致交易市場中買賣雙方的力量發生變化，更進而影響股票價格的變動。因此，新聞文件除了是投資人在決定其投資策略的重要參考依據外，同時也隱藏著具有影響股價變化的可能性[13][14][17][18][21]。

對於新聞文件這種屬於非結構化的資料[15]，我們需要進行相關技術的處理才能將之轉化成半結構化或結構化資料，也才能從中進一步擷取出有用的資訊。然而，相關研究對於結合新聞與股價預測的研究議題上卻相對地著墨較少，且其在預測的成效上亦有其限制[14][17][21]。因此，本研究提出以結合文字探勘技術及分類技術來針對非結構化的新聞事件進行分析，建置出一個能預測個股當日收盤股價漲跌趨勢的系統，可同時改善相關研究中整體正確率高但各個類別的預測效能卻彼此差異甚大的情況提出改善之道，並可做爲輔助投資人進行投資決策時的有效參考依據。

本研究提出以預測個股股票當日收盤股價的漲跌趨勢變化來作爲本研究及系統建置的重心；透過整合股價資料與財經新聞事件，並結合文字探勘技術及分類技術來建置出預測台灣股市之個股當日收盤股價漲跌趨勢預測之系統模型，以提供投資人在股票交易時間內進行投資決策的參考依據。

我們提出三種不同分類模型來建立以新聞事件爲基礎的股價預測系統，分別是簡易貝氏模型、k 最近鄰居模型以及混合模型。在系統模型中，新聞事件的資料分析及處理是透過文字探勘技術來進行；而分類器則是用來整合個股股價資料與個股財經新聞事件，在一筆新的新聞事件發佈後，分類器可以自動將之分類並進行股價漲跌趨勢的預測。我們將新聞事件資料集分割成訓練資料及測試資料兩大類，利用訓練資料來訓練分類器，並用測試資料來驗證該分類器的成效好壞。透過整合股價資料與財經新聞事件，並結合文字探勘技術及分類技術來建置一個對證券市場中個別股票的當日收盤股價進行漲跌趨勢預測之系統模型。

本研究的實驗結果不但顯示新聞事件的本身是影響股價漲跌變化的主要因素之一，同時也顯示我們所提出的簡易貝氏模型、k 最近鄰居模型以及混合模型可以有效改善相關研究中整體正確率高，但各個類別的預測效能卻差異甚大的情況。而對於影響投資人獲利與否的關鍵類別"漲"及類別"跌"的平均預測效能上，本研究所提出的這三種分類模型亦同時具有良好的成效，可以做爲投資人進行投資決策時的有效參考依據。

在第二節中，我們將簡要地回顧相關文獻，說明股價預測的相關技術，及以新聞爲資料

來源的股價預測相關研究。第三節中則是說明我們的研究方法及系統架構。第四節中則是本研究的實驗及結論，我們將所收集的樣本資料集合進行相關實驗並分析實驗結果，最後提出本研究的結論。

二、文獻探討

本章節說明以新聞爲主要資料來源的股價漲跌趨勢預測相關研究文獻。

（一）、新聞對股價指數的預測

Wuthrich 等人[21]針對全球五個主要證券交易市場的股價指數(Dow Jones Industrial Average, Financial Times 100 Index, Nikkei225, Hang Seng Index, Straits Times Index)來進行當日股價指數漲跌趨勢的預測；透過收集並分析當日交易日開盤前的相關財經新聞網站所發佈的文章及新聞內容，並利用專家所建立的關鍵詞組資料庫及文字探勘技術，將該關鍵詞組利用權重方式設定其對股價的上漲或下跌的潛在影響力大小以進行證券市場股價指數的漲跌趨勢預測。該研究以正確率(Accuracy)做爲評估其系統效能的指標，實驗結果顯示 Wuthrich 等人的系統對全球五個主要證券交易市場股價指數的平均預測正確率爲 43.6%；實驗結果並顯示在假設市場中的交易成本爲零時，五個主要證券交易市場股價指數的平均報酬率 5.9%，而 Wuthrich 等人所建置的系統之平均報酬率達 20.8%。

（二）、新聞對個別股票股價的預測

Gidófalvi[14]則是探討新聞事件發佈後對相關股票即時股價變化的影響，其研究的基本假設是認爲在交易時間內所發佈的新聞事件會在其發佈後的某一段時間內對相關個股的股價具有影響力(window of influence)，而導致股價的變動，並提出新聞事件對股價變化的影響力時間間隔爲該新聞發佈的前後 20 分鐘之內。Gidófalvi 結合即時股價資料及即時新聞資料，並透過簡易貝氏文件分類器(naïve Bayesian text classifier)來對交易時間內所發佈的新的一筆新聞來進行分類，並預測該新聞可能對股價變化的影響。Gidófalvi 將新聞事件發佈後對股價的影響分爲三個類別: 「上漲(Up)」、「不變(Unchanged)」、及「下跌(Down)」，並透過這三個類別標籤來建立該筆新聞事件與股價變動程度之間的關係。

在 Mittermayer[17]研究中，其所提出的 NewsCATS(News Categorization and Trading System)是一個可以對新聞進行自動化的分析與分類的系統，該系統並可以主動提出投資策略的建議。實驗結果顯示 NewsCATS 投資策略建議具有比以隨機方式決定買賣投資策略更好的成效，隨機方式的每筆平均投資報酬率爲 0%，而 NewsCATS 的每筆平均投資報酬率則爲 0.11%。Mittermayer 認爲在 NewsCATS 中以新聞分類的方式可以提供比新聞本身更多的資訊來進行股價趨勢的預測。

（三）、中文文件前處理

中文斷詞方式主要可分成下列兩種方法[9]：詞庫比對法(Dictionary-Based Approach)以及統計分析法(Statistical Approach)。詞庫比對法是指透過事先建立的詞庫，對輸入文件中的詞彙進行比對，再擷取出文件中出現的詞彙，完成斷詞程序。統計分析法則是透過大量文件分析，經由分析結果取得統計參數後，擷取出統計參數滿足某些條件的詞，這些統計參數可以是詞彙發生的頻率，但此方法的缺點在於當關鍵詞出現的頻率極少時，可

能無法被擷取出來。本研究對中文斷詞的處理方法是選擇採用詞庫比對法；透過由中央研究院中文詞知識庫小組中文詞庫來進行新聞文件字句的中文斷詞處理程序。

三、研究方法

本章節說明我們的研究目標、步驟、系統架構，以及本研究所提出的簡易貝式模型、*k* 最近鄰居模型以及混合模型這三種分類器。

（一）、研究目標及系統架構

本研究的目標是結合文字探勘技術及分類技術來建立一個能預測股票當日收盤股價漲跌趨勢之系統模型，以提供投資人在股票交易時間內進行投資決策的參考依據。舉例來說，若系統預測當日該股票的收盤股價會是上漲時，則在當日交易時間內的股價波動若低於前一交易日的股價時，則可建議投資人買進，倘若當日該股票的收盤股價也確實是上漲時，則投資人將會因而獲利；反之亦然。

本研究所提出的系統模型架構如圖 1 所示。我們提出三種不同分類模型來建立以新聞事件爲基礎的股價預測系統。在系統模型中，新聞事件的資料分析及處理是透過文字探勘技術來進行；而分類器則是用來整合個股股價資料與個股財經新聞事件，在一筆新的新聞事件發佈後，分類器可以自動將之分類並進行股價漲跌趨勢的預測。我們將新聞事件資料集分割成訓練資料及測試資料兩大類，利用訓練資料來訓練分類器，並用測試資料來驗證該分類器的成效好壞。透過整合股價資料與財經新聞事件，並結合文字探勘技術及分類技術來建置一個對證券市場中個別股票的當日收盤股價進行漲跌趨勢預測之系統模型。

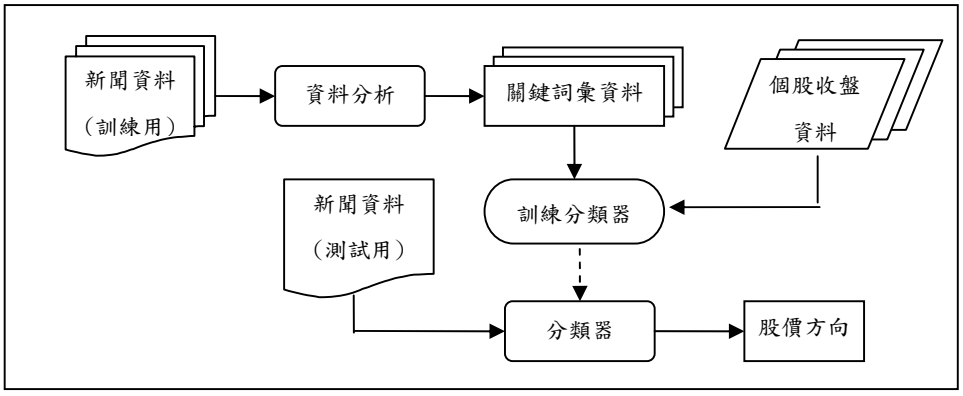


圖 1. 系統架構

（二）、分類器

分類(classification)[15][19]是指透過分類器將未知類別的資料依據其屬性值的不同來完成對該資料分派類別標籤的過程。分類器會先透過事先提供的訓練資料來學習分類規則，分類器訓練完畢後便可針對新的一筆未知類別的測試資料進行自動分類並建立該筆資料的類別標籤。在分類器的選擇上，在本研究中我們共設計三種不同的分類模型做爲分類器的核心，分別是簡易貝氏模型(naïve Bayes models)[15][20]、*k* 最近鄰居模型(*k*NN models)[15][20]以及混合模型(hybrid models)，以下分別敘述之。

1、簡易貝氏模型

簡易貝氏模型是以貝氏定理(Bayes' theorem)爲基礎，透過交換事前(prior probability)、事後機率(posterior probability)的方式來將未知類別的測試資料分派到類別機率最大的類別。簡易貝氏模型會先根據訓練資料的樣本來建立各類別機率表，對於之後所給予的測試資料則會依其的屬性值計算其歸屬於各個類別的機率值，並將具有最高機率值的類別作爲該測試資料的類別標籤。說明如下。

假設目前存在某一個特徵值x，且在樣本空間中可能出現的類別總共有k個$\{C_1, C_2, \ldots, C_k\}$，每個類別彼此間均互斥。$P(C_1)$、$P(C_2)$、…、$P(C_k)$分別爲其事前機率，則P(x)表示如下(公式 1)。

$$P(x) = \sum_{i=1}^{k} P(x \cap C_i) \quad \cdots\cdots(\text{公式 1})$$

條件機率(conditional probability)是指在已知出現某一特徵值x的條件下，某個類別C_i發生的機率記爲$P(C_i|x)$，其計算公式如下(公式 2)。

$$P(C_i \mid x) = \frac{P(C_i \cap x)}{P(x)} \quad \cdots\cdots(\text{公式 2})$$

事後機率是指當該特徵值x出現時，屬於類別C_i的機率，表示爲$P(C_i|x)$，其公式如下(公式 3)。

$$P(C_i \mid x) = \frac{P(C_i \cap x)}{P(x)} = \frac{P(C_i) \bullet P(x \mid C_i)}{P(x)} \quad \cdots\cdots(\text{公式 3})$$

假如目前是一組彼此互相條件獨立的特徵值$(x_1, x_2, \ldots, x_d)$時，則當給定某個類別C_i時，其條件機率可以表示如下(公式 4)。

$$P(x_1, x_2, \ldots, \mathrm{x_d} \mid \mathrm{C_i}) = \prod_{j=1}^{d} P(x_j \mid C_i) \quad \cdots\cdots(\text{公式 4})$$

依循上述的模式，我們可以得到 k 個類別中，包含 d 個特徵值的簡易貝式分類器模型，其公式如下(公式 5)。

$$P(C_i \mid x_1, x_2, \ldots, x_d) = \frac{P(C_i) \bullet \prod_{j=1}^{d} P(x_j \mid C_i)}{\sum_{i=1}^{k} \left(P(C_i) \bullet \prod_{j=1}^{d} P(x_j \mid C_i) \right)} \quad \cdots\cdots(\text{公式 5})$$

在特徵值選取的方法主要有資訊增益值(IG, Informatin Gain)、資訊增益比(Gain Ratio)、Gini-index、距離度量(Distance Measure)、J-measure、G 統計、$\chi 2$ 統計、最小描述長度(MLP)、正交法(Ortogonality Measure)、Relief…等，不同的度量方法有不同的分類效果，特別是對於高度分支的特徵值屬性(highly branching attributes)。在本實驗中我們嘗試以資訊增益值來做爲特徵值選取的方法，在未來我們認爲可以去嘗試不同的特徵值選取方法，以選取出更適當的特徵值來增進系統的分類效果。以下是信息增益值的簡要介紹。

資訊增益值主要是運用熵值(Entropy)的概念來做爲屬性選擇的評估依據，資訊增益值的計算方式是將未分類之前所獲得的資訊量減去分類後的資訊量，並以增益值的大小來做

爲特徵值選取的評估依據[19][20]，其計算公式如下(公式 6)。Ex 是原始樣本資料集合，H(Ex)是原始類別的熵值，H(Exla)則是考慮特徵值 a 後其不同屬性值下的熵值加總。

$$IG(Ex,a) = H(Ex) - H(Ex \mid a) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 6)}$$

因此，在簡易貝氏模型及 *k* 最近鄰居模型中的特徵值是以將新聞資料集合進行中文斷詞處理後所取得的關鍵詞，分別計算其資訊增益值後取出前 d 個關鍵詞來做爲其特徵值。而對於 d 值的設定上我們是透過對取樣的新聞資料集合進行初步實驗來決定，我們任意選取 3 個數量來比較其對系統效能的差異，分別是 25、50 及 100 個特徵值數量，並從中選出一個相對較好的來做爲 d 值的設定，以聯電新聞資料集之一爲例，實驗標的爲聯華電子股份有限公司，新聞資料來源爲台灣 Yahoo!奇摩股市新聞資料庫，資料取樣期間爲民國 95 年 2 月至民國 96 年 4 月，資料來源爲台灣 Yahoo!奇摩股市新聞資料庫，該新聞資料集經中文斷詞處理後共有 2890 個關鍵詞，透過上述方式的初步實驗結果顯示，以我們任意設定的這三個特徵值數量的整體效能而言，25 及 50 是差異不大，100 則相對較差一些，因此對於聯電新聞資料集之一的 d 值我們設定爲 50 個特徵值；對於本實驗其餘四組新聞取樣期間較短的新聞資料集合的 d 值設定則爲 25 個特徵值。

在本實驗中對於每一個特徵值的可能值只有兩種，出現(True)或未出現(False)，並假設這些特徵值彼此是條件獨立。對於某個特徵值在某一筆新聞中是否出現，我們可以透過檢查在該筆新聞文件中的該特徵值所代表的關鍵詞之詞頻是否大於 0，若該關鍵字至少出現一次則該特徵值的值爲出現，否則在該筆新聞文件中該特徵值的值則視爲未出現。在本研究中的類別數共有三個，分別是類別"漲"、類別"跌"及類別"持平"，透過訓練資料我們可以計算出每個類別出現的機率，表示爲P(C="漲")、P(C="跌")及P(C="持平")；並可以分別計算出在已知類別下第i個特徵值(x_i)出現及未出現的機率，表示爲P(x_i=出現|C="漲")、P(x_i=未出現|C="漲")、P(x_i=出現|C="跌")、P(x_i=未出現|C="跌")、P(x_i=出現|C="持平")、P(x_i =未出現|C="持平")。然而，在實際上可能會發生在已知類別下個某個特徵值都未出現而導致零機率的現象，進而在計算事後機率時(公式 5)會因爲該機率值爲零而導致無論其它機率值多大都還是會使機率相乘的結果爲零的絕對否定現象，因此我們採取將某個特徵值的所有可能值的出現次數都加上 $\frac{1}{\text{特徵值總數}}$ 的方式來避免零機率的現象[20]。

對於一筆未知類別但已知特徵值(x_1,x_2,…,x_d)的測試資料時，我們可以透過公式 5 來計算每個類別的事後機率，表示爲P(C="漲"| x_1,x_2,…,x_d)、P(C="跌"| x_1,x_2,…,x_d)及P(C="持平"| x_1,x_2,…,x_d)，並將該筆未知類別的分類爲具有最大機率值的類別，而在實際計算上，由於分母都是相同的，因此我們可以僅計算分子並比較其大小即可。

2、*k* 最近鄰居模型

k 最近鄰居模型所根據的基礎是 *k* 最近鄰居分類法(*k*NN, *k*-Nearest Neighbor Algorithm)[20]。最近鄰居分類法是指相同一類的物件彼此應該會聚集在一起，即所謂的「物以類聚」。若以向量空間中的點來表示，則對於同一類別物件的這些點彼此間的距離應該會比較接近。所以對於一個未知類別的測試資料，我們只需要在訓練資料中找出和此筆資料最接近的點，就可以最近鄰居分類法來判定此筆未知類別的測試資料的類別應該和其最接近的點的類別是相同的。然而，在多數情況下若只有使用最近鄰居來決定類別可能並不恰當。因此，常見的做法是先求取最接近的 *k* 個資料點，再根據對應的 *k*

個類別資訊來進行投票，來決定最後的類別，這種方法稱爲 k 最近鄰居分類法，也就是以 k 個最靠近的鄰居來投票決定自己的類別，至於最好的 k 值，完全是取決於資料而定。

k 最近鄰居模型是根據 k 最近鄰居分類法來找出所有的訓練資料中和該筆測試資料距離最近的 k 個鄰居，並比較這 k 個鄰居的類別標籤何者類別爲最多數後，將以此類別做爲該筆測試資料的類別，即以多數決的方式將該筆測試資料歸類爲 k 個最近鄰居中所屬的類別中票數最高的類別。

本研究中對於距離的計算方式是採用歐幾里得距離(Euclidean distance)[20]，假設在 n 維的向量空間中有兩個點 $P=(p_1,p_2,...p_n)$ 、 $Q=(q_1,q_2,...,q_n)$ ，則歐幾里得距離的計算公式如下(公式 7)。

$$D_{Euclidean}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 7)}$$

在建置k最近鄰居模型過程中，我們將整個新聞資料集合進行中文斷詞及TFIDF處理後可以建立一個關鍵詞對應於新聞文件的矩陣，假設共有d個關鍵詞，則我們可將此矩陣視爲一個d維的向量空間，每一筆新聞文件代表此d維向量空間中的一個點。因此，對於對於一筆未知類別但已知特徵值($x_1,x_2,\cdots,x_d$)的測試資料時，我們會去計算和每一筆訓練資料的距離，透過公式 7 我們可以找出k個與該測試資料最近的鄰居，並以其中最多數的類別做爲該測試資料的類別。

3、混合模型

我們並同時提出一種結合簡易貝氏模型及 k 最近鄰居模型的混合模型，透過設定門檻值 ε 的方式使得混合模型可以判別並分派一筆新的測試資料到其所適合的分類模型中來進行分類。細節說明如下。

當一筆測試資料以混合模型來決定其類別時，我們會先分別計算該筆測試資料在簡易貝氏模型中類別機率最高及次高的機率值，分別以C_i及C_j代表之。接著，混合模型會去檢查該類別機率最高及次高的類別兩者機率值差距的比例大小Δp是否大於混合模型中所事先設定的門檻值 ε ，Δp的計算公式如下(公式 8)。

$$\Delta P=\frac{P(C_i)-P(C_j)}{P(C_i)}\text{，}C_i,C_j\in\{C_1,C_2,...,C_k\}\ldots\ldots\ldots\ldots\ldots\ldots\text{(公式 8)}$$

在Δp小於 ε 的情況下，代表在簡易貝氏模型中具有最高機率值的類別(C_i)和次高機率值的類別(C_j)對該筆測試資料而言是不相上下的，也就是說該測試資料的類別不是很明顯地應該被分類爲C_i，因爲該筆測試資料歸屬於C_j類別的機率也不小。因此，在Δp小於 ε 的情況下，混合模型會選擇以k最近鄰居的模式來將該筆測試資料進行分類；而在Δp大於 ε 的情況下，混合模型就會選擇以簡易貝氏的模式來將該筆測試資料進行分類，見公式 9。

$$\text{混合模型}=\begin{cases}\text{若 } \Delta P>\varepsilon\text{，則適用簡易貝氏模型}\\ \text{若 } \Delta P\le\varepsilon\text{，則適用}k\text{最近鄰居模型}\end{cases}\ldots\ldots\ldots\ldots\text{(公式 9)}$$

透過這種機制，我們可以避免在簡易貝氏模型中只能將新聞的類別標籤分派給具有最高機率的類別，即使是在最高與次高類別的機率值差距微乎其微時，此情況下所隱含的意義是該新聞並非可以被明顯歸類於具有最高機率的類別，而本研究所提出的混合模型就

可以在此種情況下提供另一個客觀的比較依據。

四、實驗及分析

本章節是介紹本研究的實驗資料來源、實驗設計、評估方法，以及實驗的結果與分析。

（一）、資料來源及實驗設計

本研究的實驗標的爲台灣證券市場的股票上市櫃公司，我們針對電子類股中的半導體類股及發光二極體類股中選出四家公司來進行本研究相關實驗，並收集與該公司相關的財經新聞料及該個股收盤股價資料，總共取得四組新聞資料集合(聯電新聞資料集、光寶新聞資料集、晶電新聞資料集、以及立碁新聞資料集)，總計 747 筆財經新聞資料。新聞資料來源爲台灣 Yahoo!奇摩股市新聞資料庫[1]，資料取樣期間爲民國 95 年 9 月至民國 96 年 4 月；個股收盤股價資料來源則爲台灣證券交易所的個股日收盤價資料庫[4][10]。

本研究的實驗標的則是針對相同的取樣期間但新聞樣本的股票取樣標的不同的資料樣本集，實驗目的則是用來檢驗並比較本研究所提出的簡易貝氏模型、*k* 最近鄰居模型以及混合模型是否具有提升系統整體預測效能的共通性，並比較不同樣本資料集合對於不同股票標的下的系統效能差異程度。在實驗中，我們採用三折式交叉驗證分析法(3-fold cross-validation)[20]來做實驗，也就是將所收集的資料樣本平均切成三等分，其中三分之二的資料樣本做爲訓練資料，三分之一的資料樣本作爲測試資料，並分別計算出該測試資料的精確率及召回率；之後將訓練、測試資料輪流對換，此步驟共執行 3 次，直到讓每一筆資料都當過一次測試資料，如此可得到整體偏差值較小且客觀的數據。

（二）、評估方法

本實驗的評估方法是採用精確率(Precision)、召回率(Recall)、以及 F-measure 來做爲評估系統成效的指標[19]。符號定義如下：

TP：文件實際爲該類別，而系統也正確地將文件分類爲該類別之個數
FP：文件實際非屬該類別，但系統將文件分類爲該類別之個數。
TN：文件實際非屬該類別，系統也正確地將文件分類成非該類別之個數。
FN：文件實際爲該類別，但系統將文件分類成非屬該類別之個數。

精確率是計算分類系統預測其爲某一類別時，且系統正確預測的百分比，其公式如下。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{(公式 10)}$$

召回率是計算分類系統補捉到正確分類的百分比，其公式如下。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{(公式 11)}$$

F-measure 是依據精確率和召回率兩個指標加以綜合而成的評估指標，其計算公式見公式 12，其中 α 值是用來設定在 F-measure 中精確率、召回率重要程度高低的調整參數，在本研究中，我們將 F-measure 的 α 值設爲 1，也就是將精確率及召回率對於 F-measure 影響力的重要程度視爲是均等的。

$$\text{F - measure}_{\alpha} = \frac{(1+\alpha) \bullet \text{Precision} \bullet \text{Recall}}{\alpha \bullet \text{Precision} + \text{Recall}},\ \alpha \geq 0 \quad \text{.....(公式 12)}$$

此外，我們並以「所有類別平均」及「類別"漲"、"跌"平均」來做爲各屬性值離散化區間下的整體平均效能評估指標，以及「類別標準差」來做爲比較類別彼此間差異程度的評估指標[5]。

「所有類別平均」是指類別"漲"、類別"跌"、以及類別"持平"的這三個類別的整體平均精確率、整體平均召回率及整體平均 F-measure 值，此評估指標可以用來衡量系統整體的預測效能，指標值越高代表系統整體的平均預測效能越佳，對於投資人在進行投資決策時的參考價值也越高，其公式分別如公式 13、公式 14、公式 15 所示。

$$\text{所有類別平均}_{\text{精確率}} = \frac{\sum_{i=1}^{3} \text{Precision}_{\text{類別}i}}{3},\ \text{類別}i = \{\text{"漲","持平","跌"}\} \quad \text{.....(公式 13)}$$

$$\text{所有類別平均}_{\text{召回率}} = \frac{\sum_{i=1}^{3} \text{Recall}_{\text{類別}i}}{3},\ \text{類別}i = \{\text{"漲","持平","跌"}\} \quad \text{.....(公式 14)}$$

$$\text{所有類別平均}_{\text{F-measure}} = \frac{\sum_{i=1}^{3} \text{F - measure}_{\text{類別}i}}{3},\ \text{類別}i = \{\text{"漲","持平","跌"}\} \quad \text{.....(公式 15)}$$

「類別"漲"、"跌"平均」則是指類別"漲"及類別"跌"這兩個類別的平均精確率、平均召回率及平均 F-measure 值，此評估指標可以用來衡量系統對於真正會影響投資獲利與否的類別"漲"及類別"跌"這兩個類別的平均預測效能，指標值越高代表系統對於類別"漲"及類別"跌"這兩個類別的平均預測效能也越佳，除了可提供投資人在進行買進或賣出時的投資決策輔助外，更是影響投資人獲利與否的關鍵指標，其公式分別如公式 16、公式 17、公式 18 所示。

$$\text{類別"漲"、"跌"平均}_{\text{精確率}} = \frac{\sum_{i=1}^{2} \text{Precision}_{\text{類別}i}}{2},\ \text{類別}i = \{\text{"漲","跌"}\} \quad \text{.....(公式 16)}$$

$$\text{類別"漲"、"跌"平均}_{\text{召回率}} = \frac{\sum_{i=1}^{2} \text{Recall}_{\text{類別}i}}{2},\ \text{類別}i = \{\text{"漲","跌"}\} \quad \text{.....(公式 17)}$$

$$\text{類別"漲"、"跌"平均}_{\text{F-measure}} = \frac{\sum_{i=1}^{2} \text{F - measure}_{\text{類別}i}}{2},\ \text{類別}i = \{\text{"漲","跌"}\} \quad \text{...(公式 18)}$$

「類別標準差」是用來衡量各類別彼此間的差異程度大小的指標，類別標準差越小表示該系統對於各類別的預測效能越一致及穩定，越能提供投資人進行投資決策時的有效參考依據。反之，若類別標準差越大則代表該系統對於各類別的預測效能落差較大，較容易發生對於某一個類別的預測效能很高，但對另一個的預測效能卻可能非常低的情況發生，此情況會導致該系統的預測結果並不能提供投資人進行投資決策時的有效參考依據。對於分類器的整體預測效能而言，我們會希望其「類別標準差」越小越好，代表該

分類器對於各類別的預測效能較爲穩定及可靠。類別標準差的計算公式分別如公式19、公式 20、公式 21 所示。

$$\text{類別標準差}_{\text{Precision}} = \sqrt{\frac{\sum_{i=1}^{3}(\text{Precision}_{\text{類別i}} - \overline{\text{Precision}_{\text{所有類別平均}}})}{3}}\text{，類別}i=\{\text{"漲","持平","跌"}\}$$

……(公式 19)

$$\text{類別標準差}_{\text{Recall}} = \sqrt{\frac{\sum_{i=1}^{3}(\text{Recall}_{\text{類別i}} - \overline{\text{Recall}_{\text{所有類別平均}}})}{3}}\text{，類別}i=\{\text{"漲","持平","跌"}\}$$

……(公式 20)

$$\text{類別標準差}_{\text{F-measure}} = \sqrt{\frac{\sum_{i=1}^{3}(\text{F-measure}_{\text{類別i}} - \overline{\text{F-measure}_{\text{所有類別平均}}})}{3}}\text{，類別}i=\{\text{"漲","持平","跌"}\}$$

……(公式 21)

（三）、模擬 NewsCATS 系統

本研究的重心是透過分結構化的新聞資訊來進行相關個股的當日收盤股價漲跌趨勢預測的研究。然而，在我們目前爲止所能找到的股價預測相關研究中，滿足同樣是透過分析新聞資訊並針對相關個股進行股價漲跌趨勢預測的研究限制者中，僅以 Mittermayer 的研究和本研究最爲接近。因此在本研究中的實驗比較標的爲 Mittermayer 所提出的 NewsCATS 系統。雖然 NewsCATS 的整體預測效果優異，且對於類別"持平"的平均預測精確率高達 98%，但對於類別"漲"、"跌"的平均預測精確率卻分別只有 5%及 6%，預測效果較類別"持平"差距非常顯著。這個現象顯示 NewsCATS 的系統限制是僅能提供投資人對於類別"持平"的預測來做爲其投資決策的參考依據，對於投資人更爲重視且影響其獲利與否的類別"漲"及類別"跌"這兩個類別上的預測效果，NewsCATS 系統卻不能提供投資人有效及可靠的預測參考依據。

基於 Mittermayer 的研究和本研究仍有許多不同之處，如：新聞資訊的語言不同、證券市場不同、新聞取樣期間不同…等諸多差異點。且受限於難以取得 Mittermayer 當時的樣本資料及其分類器的重建時的相關設定，因此，我們採取模擬 NewsCATS 的方式來做爲實驗的比較基礎。在之後進行相關的實驗中，本研究會以模擬的 NewCATS 系統來代替實際的 NewsCATS 系統，並和我們所提出的簡易貝氏模型、*k* 最近鄰居模型以及混合模型來進行彼此系統預測效能的比較。以下我們會以模擬的 NewsCATS 系統以代替實際的 NewsCATS 系統來做爲實驗比較的標的。

（四）、實驗結果及分析

本實驗的重心在於探討在相近的取樣期間下，對於不同新聞標的股票下的不同系統彼此間的整體效能差異程度爲何，並進一步探討本研究在不同的資料集合下是否仍具有提升系統整體預測效能的適用性。本實驗的樣本資料集共有四個，分別是聯電新聞資料集、光寶新聞資料集、晶電新聞資料集、以及立碁新聞資料集。

表 1 中分別爲聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在建置簡易貝氏模型、*k* 最近鄰居模型與混合模型時的最適參數設定值。

表 1. 相近取樣期間不同樣本資料集的最適參數設定值

新聞樣本資料集合	最近鄰居數	門檻值設定	資料取樣期間	資料筆數
聯電新聞資料集	k= 1	ε = 20%	4 個月	197
光寶新聞資料集	k= 1	ε = 10%	5 個月	187
晶電新聞資料集	k= 1	ε = 90%	7 個月	291
立碁新聞資料集	k= 7	ε = 90%	5 個月	72

我們針對聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集這四組取樣期間相近且針對不同新聞標的股票的新聞樣本資料集合進行彼此間系統效能的比較與分析探討。在這四組新聞樣本資料集合中，我們以「所有類別平均」、「類別"漲"、"跌"平均」及「類別標準差」來評估在本研究所提出的簡易貝氏模型、k 最近鄰居模型、混合模型與模擬 NewCATS 系統的整體效能；「所有類別平均」可以顯示出系統整體的平均預測效能，「類別"漲"、"跌"平均」則可以顯示影響投資人獲利與否的類別"漲"及類別"跌"的平均預測效能，而「類別標準差」則可以顯示系統內的類別間彼此預測效能的差異程度。

在表 2、表 3 及表 4 中，分別顯示聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在簡易貝氏模型下的系統預測精確率、召回率及 F-measure 值，其中的模擬 NewsCATS 系統欄位的數值是將每個新聞樣本資料集下在其系統下所模擬的 NewsCATS 系統分別加總平均，也就是在這四組新聞樣本資料集合下所模擬的平均 NewsCATS 系統效能。

實驗結果顯示，當以「所有類別平均」來做爲系統整體評估指標時，除了晶電電新聞資料集在精確率及 F-measure 較模擬 NewsCATS 系統分別低 2.66%及 0.61%外，其餘的新聞樣本資料集合的系統整體平均預測效能都是比模擬 NewsCATS 系統高。

而對於影響投資人獲利與否的類別"漲"及類別"跌"的評估指標「類別"漲"、"跌"平均」，聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集這四組新聞樣本資料集合的系統預測效能都是比模擬 NewsCATS 系統高；其中，精確率比模擬 NewsCATS 系統高 9.71%到 33.00%之間，召回率比模擬 NewsCATS 系統高 7.28%到 32.61%之間，F-measure 值比模擬 NewsCATS 系統高 17.45%到 27.36%之間。

當以「類別標準差」來評估在這四組新聞樣本資料集合下的簡易貝氏模型中的類別間彼此預測效能差異程度時，實驗數據顯示模擬 NewsCATS 的類別間的預測效能差異程度最大，顯示該系統的預測並不能有效投資人進行投資決策時的參考依據。

表 2. 相近取樣期間不同樣本資料集下之簡易貝氏模型精確率比較

精確率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	49.24	73.81	53.03	24.07	80.41
類別"跌"	47.87	39.17	34.39	62.50	11.48
類別"漲"	45.70	44.62	24.44	42.93	27.95
所有類別平均	47.60	52.53	37.29	43.17	39.95

類別"漲"、"跌"平均	46.79	41.89	29.42	52.71	19.71
類別標準差	1.79	18.63	14.51	19.21	39.24

表 3. 相近取樣期間不同樣本資料集下之簡易貝氏模型召回率比較

召回率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	49.54	18.67	33.90	77.78	75.72
類別"跌"	55.35	16.85	62.13	18.52	20.94
類別"漲"	36.51	91.72	25.94	39.39	22.43
所有類別平均	47.13	42.41	40.66	45.23	39.70
類別"漲"、"跌"平均	45.93	54.29	44.03	28.96	21.68
類別標準差	9.65	42.71	19.02	30.06	40.33

表 4. 相近取樣期間不同樣本資料集下之簡易貝氏模型 F-measure 值比較

F-measure(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	49.39	29.80	41.36	36.77	77.70
類別"跌"	51.34	23.56	44.27	28.57	14.61
類別"漲"	40.59	60.03	25.17	41.08	19.74
所有類別平均	47.37	46.93	38.90	44.17	39.51
類別"漲"、"跌"平均	46.35	47.29	35.27	37.38	19.93
類別標準差	5.73	19.51	10.29	6.36	38.71

在表 5、表 6 及表 7 中分別顯示聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在 *k* 最近鄰居模型下的系統預測精確率、召回率及 F-measure 值。

實驗結果顯示，當以「所有類別平均」來做爲系統整體評估指標時，這四組新聞樣本資料集合的系統預測效能都是比模擬 NewsCATS 系統高；其中，精確率比模擬 NewsCATS 系統高 2.62%到 10.81%之間，召回率比模擬 NewsCATS 系統高 1.04%到 11.56%之間，F-measure 值則比模擬 NewsCATS 系統高 2.12%到 11.50%之間。而對於影響投資人獲利與否的類別"漲"及類別"跌"的評估指標「類別"漲"、"跌"平均」，除了立碁新聞資料集在召回率上較模擬 NewsCATS 系統低 2.24%外，其餘都是較模擬 NewsCATS 系統的預測效能高。當以「類別標準差」來評估在這四組新聞樣本資料集合下的 *k* 最近鄰居模型中的類別間彼此預測效能差異程度時，實驗數據也顯示模擬 NewsCATS 的類別間的預測效能差異程度最大，顯示該系統的預測並不能有效投資人進行投資決策時的參考依據。

表 5. 相近取樣期間不同樣本資料集下之 *k* 最近鄰居模型精確率比較

精確率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	46.04	57.39	56.12	16.59	80.41

類別"跌"	46.52	33.54	44.52	61.11	11.48
類別"漲"	52.63	61.36	40.99	50.00	27.95
所有類別平均	48.40	50.76	47.21	42.57	39.95
類別"漲"、"跌"平均	49.57	47.45	42.76	55.56	19.71
類別標準差	3.67	15.05	7.92	23.17	39.24

表 6．相近取樣期間不同樣本資料集下之 *k* 最近鄰居模型召回率比較

召回率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	48.00	57.33	43.03	83.33	75.72
類別"跌"	46.32	34.62	38.98	33.33	20.94
類別"漲"	49.21	61.83	57.97	5.55	22.43
所有類別平均	47.84	51.26	46.66	40.74	39.70
類別"漲"、"跌"平均	47.76	48.23	48.47	19.44	21.68
類別標準差	1.45	14.59	10.00	39.42	40.33

表 7．相近取樣期間不同樣本資料集下之 *k* 最近鄰居模型 F-measure 值比較

F-measure(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	47.00	57.36	48.71	27.67	77.70
類別"跌"	46.42	34.07	41.57	43.13	14.61
類別"漲"	50.86	61.59	48.03	10.00	19.74
所有類別平均	48.12	51.01	46.93	41.63	39.51
類別"跌"、"漲"平均	48.65	47.84	45.44	28.80	19.93
類別標準差	2.42	14.82	3.94	16.58	38.71

在表 8、表 9 及表 10 中分別顯示聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在混合模型下的系統預測精確率、召回率及 F-measure 值。

表 8．相近取樣期間不同樣本資料集下之混合模型精確率比較

精確率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	50.87	82.50	56.14	18.29	80.41
類別"跌"	54.45	43.89	44.46	56.67	11.48
類別"漲"	47.99	46.91	40.99	55.56	27.95
所有類別平均	51.10	57.77	47.20	43.50	39.95
類別"漲"、"跌"平均	51.22	45.40	42.73	56.11	19.71

類別標準差	3.23	21.47	7.94	21.84	39.24

表 9．相近取樣期間不同樣本資料集下之混合模型召回率比較

召回率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	53.44	24.00	42.22	91.67	75.72
類別"跌"	50.26	19.41	40.17	25.92	20.94
類別"漲"	42.86	93.11	57.97	11.36	22.43
所有類別平均	48.85	45.51	46.78	42.98	39.70
類別"漲"、"跌"平均	46.56	56.26	49.07	18.64	21.68
類別標準差	5.43	41.29	9.74	42.78	40.33

表 10．相近取樣期間不同樣本資料集下之混合模型 F-measure 值比較

F-measure(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	52.12	37.18	48.19	30.50	77.70
類別"跌"	52.27	26.92	42.21	35.57	14.61
類別"漲"	45.28	62.39	48.03	18.87	19.74
所有類別平均	49.95	50.91	46.99	43.24	39.51
類別"漲"、"跌"平均	48.78	50.25	45.68	27.99	19.93
類別標準差	3.99	18.25	3.41	8.56	38.71

在本實驗中我們針對聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集這四組取樣期間相近但新聞取樣標的爲不同股票進行彼此間整體效能差異程度的比較，並和模擬 NewsCATS 系統進行比較，藉以檢驗在實驗 A 中對於本研究所提出的簡易貝氏模型、*k* 最近鄰居模型及混合模型是否同樣適用於其它的樣本資料集合，且同樣能具有改善模擬 NewsCATS 系統的整體預測效能。本實驗顯示，不同的新聞的資料集合雖然彼此的預測效能並不會完全相同，不過透過最適參數的設定可以使其在本研究所提出的簡易貝氏模型、*k* 最近鄰居模型及混合模型中具有較 NewsCATS 系統爲佳的系統預測效能。

五、結論

股價的漲跌變化是由於證券市場中眾多不同投資人及其投資決策後所產生的結果。然而，影響股價變動的因素眾多且複雜，新聞也屬於其中一種，新聞事件不但是投資人用來得知該股票上市公司的相關營運資訊的主要媒介，同時也是影響投資人決定或變更其股票投資策略的主要因素之一。本研究提出以新聞文件做爲股價漲跌預測系統的基礎架構，透過文字探勘技術及分類技術來建置出能預測當日個股收盤股價漲跌趨勢之系統。本研究共提出了簡易貝氏模型、*k* 最近鄰居模型以及混合模型這三種分類模型，並透過實驗來檢驗系統的預測效能。

實驗結果顯示本研究所提出的簡易貝氏模型、*k* 最近鄰居模型以及混合模型這三種分類模型對於系統的整體平均預測效能及對於影響投資人獲利與否的類別"漲"及類別"跌"的平均預測效能都是比相關研究的系統預測效能爲佳，顯示本研究所提出的分類模型可以提供提資人穩定及可靠的預測品質。

參考文獻

[1] Yahoo!奇摩股市，http://tw.stock.yahoo.com/。

[2] 中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/。

[3] 中央研究院資訊科學所中文組實驗室中文詞知識庫小組，http://godel.iis.sinica.edu.tw/CKIP/index.htm。

[4] 中華民國證券櫃檯買賣中心，http://www.otc.org.tw/。

[5] 方世榮，*統計學導論*，華泰書局，頁 39-81、215-231，1993。

[6] 王春笙，*以技術指標預測台灣股市股價漲跌之實證研究－以類神經網路與複迴歸模式建構*，台灣大學資訊管理研究所碩士論文，1996。

[7] 王疏艷，基於決策樹方法的分類規則的挖掘，海鼎出版，2002，http://hd123.com/asprun/Message/MessageList.asp?gid=17658。

[8] 施正宏，*結合總體經濟指標及個股財報資料以預測個股漲跌－以台灣電子類股爲例*，中原大學資訊管理學系碩士論文，2004。

[9] 曾元顯，"關鍵詞自動擷取技術與相關詞回饋"，*中國圖書館學會會報 59 期*，頁 59-64，1997。

[10] 臺灣證券交易所，http://www.tse.com.tw/。

[11] 謝德宗，*投資學*，華泰書局，頁 235-253、324、403-418，1997。

[12] H. Braun and J. S. Chandler, "Predicting Stock Market Behavior through Rule Induction: An Application of the Learning-from-Example Approach," *Decision Sciences*, volume 18, number 3, pp. 415-429, 1987.

[13] G. P. C. Fung, J. X. Yu and W. Lam, "News Sensitive Stock Trend Prediction," *Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 289-296, 2002.

[14] G. Gidófalvi, "Using News Articles to Predict Stock Price Movements," *Technical Report: CSE 254*, Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA, 2001.

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kaufmann, pp. 614-626, 2006.

[16] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM of Research and Development*, pp. 159-165, 1958.

[17] M.-A. Mittermayer, "Forecasting Intraday Stock Price Trends with Text Mining Techniques," Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences, Track 3, p. 30064b, 2004.

[18] R. P. Schumaker and H.-C. Chen, "Textual Analysis of Stock Market Prediction Using Financial News Articles," *Proceedings of the Twelfth Americas Conference on Information Systems*, Acapulco, Mexico, 2006.

[19] Wikipedia,http://www.wikipedia.org/.

[20] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, pp. 76-80, pp. 88-96, pp. 149-151, pp. 296-304, 2000.

[21] B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, and J. Zhang, "Daily Stock Market Forecast from Textual Web Data," *Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2720-2725, 1998.

基於階層架構資訊及關鍵詞語義擴展的階層式目錄整合研究

洪誠澤　陳英祥　吳秉蓉　楊正仁
元智大學資訊工程學系
Department of Computer Science and Engineering
Yuan Ze University
{chris,sean,pjwu,czyang }@syslab.cse.yzu.edu.tw

摘要

目錄整合的議題近年來受到不少研究的注意。針對攤平式與階層式的分類目錄，分別有不同研究利用來源端目錄所隱含的目錄資訊，有效的提升整合的精確度效能。然而在目前的文獻回顧裡，我們尚未看到使用外部語義庫的資訊來提升階層式目錄整合效能的相關研究。在本論文中，我們探討如何利用外部語義庫與目錄階層架構關係的資訊，使得目錄整合效能可以進一步被提升。在初步實驗中，我們使用最大熵（Maximum Entropy）模型來實作 KSE-ME 整合機制，並用實際的 Web 目錄來進行測試，同時與使用支持向量機核心的 ECI-SVM 一起評估。實驗的結果顯示如果能同時利用階層架構資訊及關鍵詞語義擴展資訊，將可得到良好的整合效果。

關鍵詞：階層式目錄整合，最大熵模型，關鍵詞語義擴展，階層架構索引典資訊

一、緒論

在資訊的處理與組織上，常常會以階層式目錄的架構來分類相關的資訊。而在不同的環境中，也常需要將不同的目錄內容整合在一起。例如，在電子商務市場上，一些大型的網路書店，像亞馬遜書店(Amazon) [1]，需要整合一些下游的零售商的產品目錄，以提供客戶更多元的網路購物選擇。一些大型的商業公司，也須要與它下游廠商提供的零件與產品目錄進行整合，以加強企業內部資源整合。在學術領域上，一些聯合性的學術數位資源網站也會需要整合多個數位圖書館的數位內容目錄，以提供相關研究學者使用這些學術資源。此外，一般使用者最常使用的入口網站與新聞網站，將來也會面臨和其他同質性網站進行目錄內容整合的需求。

從各式分類目錄的實際應用環境顯示，透過分類目錄呈現資訊內容的方式，確實需要一套有效的機制，以提供網路資訊的準確整合和交換。然而，由於網際網路上分類目錄的資訊量愈加龐大與內容日趨廣泛，採用人工整合的方式不僅耗時費力且成本過高，長期維護下來，非常不符合經濟效益。此外，若目錄的規模不斷地成長，人工的整合方法也無法滿足實際的需求。有鑑於此，面對如此日趨龐大的目錄資訊，如何建立準確有效的自動化目錄整合機制，已成爲目錄整合的重要議題。

過往已經有許多研究討論如何進行目錄整合[3,6,9,10,13,14,16,17]。然而當中有許多研究偏重在單純的攤平式目錄整合機制之上，討論將目錄類別全部扁平化以後，整合到攤平式的目的目錄中，並不考慮直接整合到階層式目錄中[3,10,14,16,17]。因此這樣的過程中無法考慮到目的目錄中，類別與子類別之間的從屬關係。由於許多實際的目錄皆是階層式的架構，因此階層式目錄整合機制很需要進一步探討。

過往階層式文件分類研究顯示[8,11]，利用階層式架構確實能有效的加強原本攤平式分類的準確性。其中，McCallum 等人發現在階層式架構下，利用機率模型與聚斂(shrinkage)的方法，可以有效地提升文件分類的準確性。他們的實驗結果並顯示當資料集中的特徵資訊數量足夠多時，階層式架構的分類效果明顯高於採用攤平式架構[11]。然而，他們的研究也顯示，shrinkage 的效果並非絕對能提升階層式分類的準確性。當訓練資料量較大時，反而可能造成某些目錄的準確率下降[11]。

在階層式目錄整合的相關研究上，有不少相關研究皆證實階層式架構在目錄整合上的好處[6,7,9,13]。其中，Doan 等人透過擷取階層式架構中，相鄰節點的相關資訊特徵，來加強本體知識的語意對應[7]。Rajan 等人則是採用最大相似可能性（maximum likelihood）模型，並且充分地討論階層式目錄整合的各種對應情形[13]。他們的方法中更進一步提出，可以將來源目錄的類別新增到目的目錄中，建立新的階層式架構。Chen 等人和 Ho 等人並利用階層式架構索引含的索引典關係（thesaurus），以支持向量機（Support Vector Machines）建立起一個整合機制 ECI-SVM，有效地提升階層式目錄整合的成效[6,9]。

然而過往階層式目錄整合研究只注意到利用階層式架構的好處，我們發現事實上目錄中的語義關係可以進一步被用來加強整合效果。在[15]的研究中顯示，利用外部語義庫（例如 WordNet）自動建立文件叢集的標題時，能有效地建立具語義概念代表性的主題標籤。因此在本研究中，基於過往 ECI-SVM 的研究，探討如何利用[15]的研究成果，找出可以代表文件的關鍵詞，並透過外部語義庫來擴展出輔助語義資訊（Keyword Semantic Expansion），提升目錄整合的效能。

由於統計物理模中的最大熵模型（Maximum Entropy Model，ME）可以將所有特徵的機率條件一起考慮，預測結果不容易受到單獨表現不顯著的資料條件所影響，因此過往研究顯示 ME 在自然語言處理與文件分類上有很好的表現[4,5,12]。另一方面，如果資料中擁有代表性的明顯的機率條件，其最後的結果卻會依據此具代表性的條件來提升整合的效果。因此，我們在研究中採用 ME 做爲階層式目錄整合的模型，並與過往研究中的 ECI-SVM 來分析比較整合效能。我們分別實作以 ECI 爲基礎，但分類器改爲 ME 的 ECI-ME，以及在 ECI-ME 之上加入關鍵字語義擴展（Keyword Semantic Expansion）的 KSE-ME。我們以兩個實際的 Web 階層目錄，進行目錄整合實驗。實驗結果顯示，在階層式架構下，ECI-ME 的準確率表現平均優於 ECI-SVM，並且引用 InfoMap [2]的外部語義庫資訊，KSE-ME 能進一步加強階層式目錄整合的整體成效。

本論文其餘的章節安排如下：第二節將介紹過往目錄整合相關研究，以及外部語義庫來加強文件資訊的研究。在第三節中，我們將簡略介紹最大熵模型分類器，以及如何將關鍵詞語義擴展運用在 ECI 的方法之中。第四節將說明我們的實驗結果。最後，第五節是本研究的結論。

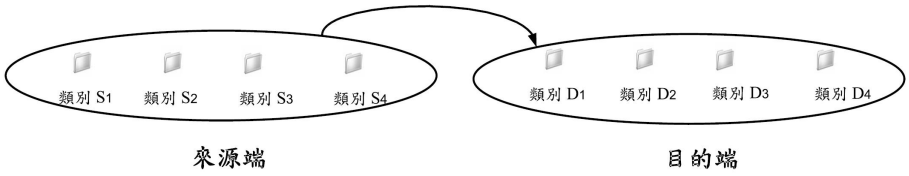
二、相關研究

（一）、目錄整合

目錄進行整合時，最簡單的方式是用基本的文件分類方式來進行，然而過往研究發現可以利用其他相關的資訊來有效提升目錄整合的準確性。以下我們將目錄整合分爲攤平式目錄整合與階層式目錄整合，然後依這二大類分別介紹相關研究。

1、攤平式目錄整合

在目錄整合的研究中，最初僅考慮單純的攤平式目錄整合。如圖一所示，來源端的目錄與目的端的目錄中，分別包含了類別 $S_1,\ldots, S_m$ 以及 $D_1,\ldots,D_n$，這些類別之間相互沒有關連，也沒有上下之間的階層關係。在大多數的早期相關研究中，如果所處理的類別之間有目錄階層關係，便會將這些類別的子類別全部合併成如圖一的攤平式架構來處理。並不討論上下之間的階層關係。



圖一、攤平式目錄整合

例如在 2001 年，Agrawal 和 Srikant 所提出的 Enhanced Naive Bayes (ENB)方法[3]。這個方法主要是挖掘出原始目錄架構裡所具備的隱藏資訊，再利用機率模型中的 NB 分類器(Naive Bayes classifier)，將這些架構隱藏資訊學習出來，最後利用 NB 分類器分析這些資訊，用以提高整合的準確率。實驗的結果顯示，只利用 NB 的原本方法來進行目錄整合時，所達到的正確率爲 47.4%。但是透過 Agrawal 和 Srikant 所提出的 ENB 方法，卻能提升到 61.7%。證明隱含資訊的使用能有效提昇目錄整合正確性。但在他們研究中，僅略微提到可以擴展成階層式的整合方式，卻沒有進一步討論相關作法。

Sarawagi 等人在 2003 年提出一個交互學習(Cross-training, CT) 的方法[14] 來改進目錄整合的正確性，並利用 SVM 分類器和 Expectation Maximization (EM) 分類器來實作整合機制。CT 利用交互學習的方式，將來源端的目錄資訊先擷取出來加入目的端的分類器，如此一來目的端的目錄便可學習來源端的目錄資訊，並加強目錄整合的效果。在他們的實驗中，實驗結果顯示 CT 的機制能夠有效的改進 EM 分類器。但對於 SVM 分類器而言，卻只能對約半數的目錄有改進的效果，不能全面地加強目錄整合的準確性。而他們所討論的整合對象，也還都是攤平式的目錄。

除了上述的相關研究之外，後續還有許多相關研究在探討攤平式目錄整合的應用，提出不同的輔助演算法[10,17]。在 2004 年，Lee 和 Zhang 使用 SVM 並且配合 cluster shrinkage 的方法，以及利用 co-bootstrapping 等兩種機制，來加強攤平式目錄整合的效果。根據他們的實驗結果顯示，這些方法皆能有效地提升攤平式目錄整合的準確性。然而，由於這些加強方法都是基於攤平式的方法來進行目錄整合，並沒有討論如何針對階層式目錄來直接整合。

Wu 等人在 2005 年，利用來源目錄的階層式架構資訊，並使用最大熵模型來進行目錄整合 [16]。其研究結果顯示，整合的效果明顯地改善原有的 ENB 方法。在這個研究當中，主要改進的方法，就是透過來源目錄的階層式架構資訊加強攤平式目錄整合的效果。雖然此研究只使用了目錄的階層架構資訊，並未討論如何對階層式目錄進行整合，但是從他們的研究結果顯示，階層架構資訊確實可以被用來提升目錄整合的效能。

雖然已有研究討論如何使用階層架構資訊，然而它們仍是對攤平式目錄來整合，並未進一步討論如何有效地直接整合階層式目錄。因此雖然這些研究顯示來源端目錄資訊可以提升整合效能，但是對於階層式目錄整合的實際應用卻缺乏討論。因此，如何將兩

個階層式目錄，在不經過扁平化的過程而直接整合，將是一個重要的研究議題。

2、階層式目錄整合

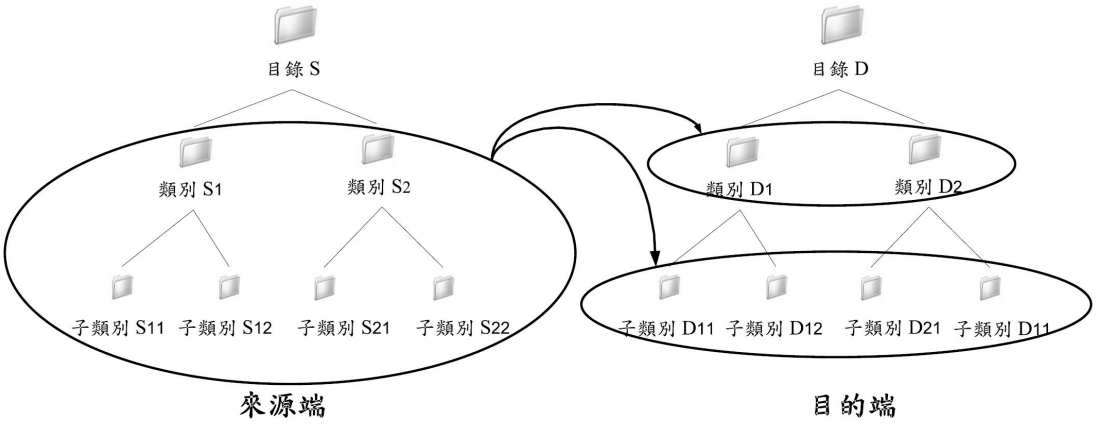
針對目錄整合問題，我們在這裡所討論的階層式目錄整合，與攤平式目錄整合的最大差別，就在於這些階層式目錄並不先被扁平化成攤平式目錄，而是直接進行兩個目錄的整合。如圖二，目錄中的文件存在子類別當中。在進行整合時，會先參考到子類別的父類別資訊，也就是 S_1 或 S_2 類別。在整合至目的端時，會依據文件及階層資訊的特性，可能被整合到類別目錄 D_1 或 D_2 當中，也有可能被整合到子類別目錄當中。

在 2005 年，Rajan 等人提出了一個兩階段的對應與整合方法，並且討論了階層式目錄整合中的四種情形[13]。Rajan 等人有效地利用階層式架構與 maximum likelihood 的方法來提升目錄整合的成效。他們的實驗結果顯示，maximum likelihood 在一對多的階層式架構目錄整合效能非常顯著，並且此研究也顯示目的目錄的階層式架構在目錄整合上能有效地幫助準確率的提昇。

在 2006 年時，Chen 等人[6]和 Ho 等人[9]陸續地進一步探討如何從來源端目錄與目的目錄中，從他們的階層關係中抽取出類似於索引典(thesaurus)的詞彙關係，加強階層式目錄整合的成效。在他們的研究中，提出了一個 ECI 的階層目錄整合方式，採用 SVM 分類器並配合 one-against-rest 的分類機制，讓原本是二元分類法的 SVM 模型能解決多類別的問題。Ho 等人的實驗結果進一步顯示，在 Yahoo！和 Google 上的五個實際的階層式目錄中，利用來源目錄與目的目錄階層式結構的關係，在大多數的攤平式和階層式目錄下，皆能明顯地提升目錄整合的成效[9]。在本研究中，我們並基於 ECI 的方式，另外再考慮加上語義資訊，來提升整合效能。

（二）、外部語義庫

在 2006 年，Tseng 等人[15]提出了一個在相關叢聚的文件中，找出較具有代表性的特徵詞，並從中運用這些特徵詞到外部的語義庫 WordNet 尋找出最具代表性的一個詞，來表示這些特徵詞。Tseng 等人應用 Correlation Coefficient (CC)特徵詞的篩選方法，並且透過所提出的機制來自動選擇 WordNet 相關特徵詞中合適的上位詞，以代表相關文件的標題。例如：找出的特徵詞是 apple、banana、orange，作者進一步利用 WordNet 上位詞關係，找出 fruit 這個詞來當作標籤。實驗結果顯示，此方法可以自動地選擇出



圖二、階層式目錄整合

合適的特徵詞與適切的叢聚文件標題。

從過往階層式目錄整合的研究中，我們發現若能適當地運用來源目錄與目的目錄之間的階層式架構關係，並且透過合適的語義詞典將兩階層式目錄的語義關係加強，將能有效地提升階層式目錄整合的準確性。因此，本論文進一步探討如何結合來源目錄與目的目錄的階層式架構關係，並參考 Tseng 等人[15]的研究，運用合適的外部語義庫來擷取相關語義詞彙與叢集標籤，以加強階層式目錄整合的準確性。

三、階層目錄整合方法

（一）、問題描述

在目錄整合的過程中，我們預設使用者乃是針對兩個同質性比較高的目錄來進行整合。所謂「同質性較高」是指兩個目錄具有類似的分類意義，且兩者之間有部分文件是相同的。兩個目錄分別是來源端目錄 S 與目的端目錄 D。S 當中有 n 個相似而不同的類別目錄 $S_1,S_2,...,S_n$ ，而每一個類別底下或者有最多 m 個子類別目錄 $S_{11},S_{12},...,S_{1m}$，其中 m 是該層子類別的最大值。其他層次以此類推。另一方面，D 當中有 p 個相似而不同的類別目錄 $D_1,D_2,...,D_p$，而每一個類別底下或者有最多 q 個子類別目錄 $D_{11},D_{12},...,D_{1q}$。其他層次以此類推。

階層式目錄整合的目的，就是將來源端目錄 S 底下的子類別目錄當中的所有文件能夠正確整合到目的端目錄 D 底下的子類別目錄當中。例如將 Google 目錄的 autos 類別中的文件，整合到 Yahoo!目錄 automotive 的類別裡。在這裡有幾個整合上的議題需要特別討論。首先，由於在真實的目錄環境裡，一份文件可能會被歸類在一個以上的目錄類別中。針對這種情況，我們採用 one-against-rest 的分類機制，以保留真實目錄環境中一份文件可以同時存在兩個以上的類別的情形。第二，兩個實際的目錄架構很有可能並不一樣，爲了簡化評估上的複雜度，因此我們在本研究中，先討論對稱架構上的目錄整合問題，也就是將兩個目錄先簡約成相同的對稱架構，進行階層式目錄整合上的討論。

（二）、最大熵模型

我們使用 Maximum Entropy (ME) 模型 [5]爲整合機制的分類器核心。ME 是一個統計物理學模型技術，用來測量既有資料最大複雜度的條件機率。ME 模型能測量所有可能的分佈，並呈現出最貼近已知條件的資料分佈情形，在文件分類上的運用都能夠有相當不錯的效能[12]。在本研究中，ME 主要是用來計算來源目錄內的文件被整合到目的目錄中的機率分佈。

1、機率分佈的 Entropy

在 Maximum Entropy 模型中，Entropy 用來表示在一個不確定性的情況下，機率分佈的複雜度。對一個 alphabet 集合 X，若它有一組機率分佈模型 $P(x) = \{p(x_1), p(x_2), \ldots, p(x_n)\}$, $x_i \in X$，則對機率 p 而言，它的 Entropy $H(p)$ 被定義爲：

$$H(p) = -\sum_{\forall x_i \in X} p(x_i)\log p(x_i) \qquad (3.1)$$

如果考慮一組條件機率分佈模型 $p(y|x)$的 Conditional Entropy，則可定義爲：

$$H(p) = -\sum_{\forall x \in X} \tilde{p}(x) p(y \mid x) \log p(y \mid x) \qquad (3.2)$$

其中 $\tilde{p}(x)$ 是由實際現象所觀察出之經驗機率（empirical probability）。

2、Maximum Entropy 的限制

在找出一組使得熵值爲最大的機率前，必須定義出特徵函式（feature function）$f(x,y)$ 來表示所要觀察的現象。例如式(3.3)就是一個常見的二元表示法。若其中 x 爲特徵詞，y 代表文件的集合，則 $f(x,y)=1$ 時所滿足的條件是表示特徵詞 x 出現在文件集合 y 中。

$$f(x,y) = \begin{cases} 1 & \text{如果(x,y)滿足條件} \\ 0 & \text{其它} \end{cases} \qquad (3.3)$$

針對所給予的資料集，我們可依需求決定出 $f(x,y)$之滿足條件。在此資料集中，我們希望針對經驗機率 $\tilde{p}(x,y)$ 算出特徵期望（feature expectation），如式(3.4)。但實際觀察所得到的條件分佈則如式(3.5)的近似函式，其中 $\tilde{p}(x)$ 是由訓練文件所觀察出之經驗機率。特徵期望應與經驗期望一致，因此必須滿足 $E_p\{f\} \equiv E_{\tilde{p}}\{f\}$ 之限制。此外，每一組計算出來的機率值總合必須爲 1，如式(3.6)。

$$E_p\{f\} \equiv \sum_{x,y} \tilde{p}(x,y) f(x,y) \qquad (3.4)$$

$$E_{\tilde{p}}\{f\} \equiv \sum_{x,y} \tilde{p}(x) p(y \mid x) f(x,y) \qquad (3.5)$$

$$\sum_{y} p(y \mid x) = 1 \qquad (3.6)$$

3、最大熵模型之解

Maximum Entropy 的原理是從一個受限制之條件機率分佈集合 C 中，找出一個機率模型 p^*，使 Entropy 得到最大值。式(3.7) 的 p^* 即爲 Maximum Entropy 的解法。

$$p^* = \arg\max_{p \in C} H(p) \qquad (3.7)$$

因此只要確定出 p^* 就可得到 Maximum Entropy。從 p^* 的式(3.7)與 Entropy 本身的兩個限制，帶入 Lagrange Multipliers 來處理（推演過程可參考[5]），可以得到式(3.8)來計算文件分類的條件機率。透過 Maximum Entropy 計算 $p(y|x)$的機率值，其中 y 是文件集合，x 爲特徵詞的集合，$f_i(x,y)$表示是第 i 個特徵函式，$z(x)$的計算如式(3.9)。

$$p(y \mid x) = \frac{1}{z(x)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x,y) \right) \qquad (3.8)$$

$$z(x) = \sum_{y} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x,y) \right) \qquad (3.9)$$

因此只要計算出最合理的 λ 值，就可以得到最大的 Entropy 的機率值。機率模型中的每個特徵詞都會有一個 λ 值，其權重由一個 Improved Iterative scaling（IIS）的方程

式計算所得。主要是爲了要使每個 λ 值滿足以下方程式，相關內容可參考[4]。圖三爲 IIS 的演算法，在一開始的時候，給 λ_i 一些隨機產生趨近於 0 的數值。接著在迴圈中重複作微分的動作，直到結果收斂爲 0。因此，IIS 演算法的中需要調整 λ_i 的值，使其滿足微分式能等於 0。當微分結果收斂爲 0 時，就將 λ_i 代入 $\lambda_i + \delta_i$，並產生預測的 λ_i 值與最大熵值。

IIS Algorithm

1.Start with some value for each λ_i

2.Repeat until convergence:

Find each δ_i by solve the equation: $\frac{\partial B(\Delta)}{\partial \delta_i} = 0$

Set $\lambda_i \leftarrow \lambda_i + \delta_i$

圖三、Improved Iterative scaling演算法

期望值方程式(3.9)的功能是要滿足每個條件機率，使式(3.8)計算出來的值，在式(3.6 加總爲 1。因此 Maximum Entropy 分類器能夠在目錄整合時，保證每個特徵詞能夠滿足機率值總和爲 1。

（三）、關鍵詞語義擴展

1、目錄整合流程

目錄整合的流程如圖四。目錄整合的步驟包含網頁文件處理，特徵詞選取。然後從特徵值中進行關鍵詞的語義擴展，自外部語義庫中取出適當的上位詞（hypernyms），並加入階層架構的索引典資訊，與文件原有特徵詞共同組合成擴展文件特徵。最後將此擴展文件特徵轉換爲 ME 分類器格式，進行目錄整合。以下將進一步說明各個步驟。

2、文件處理

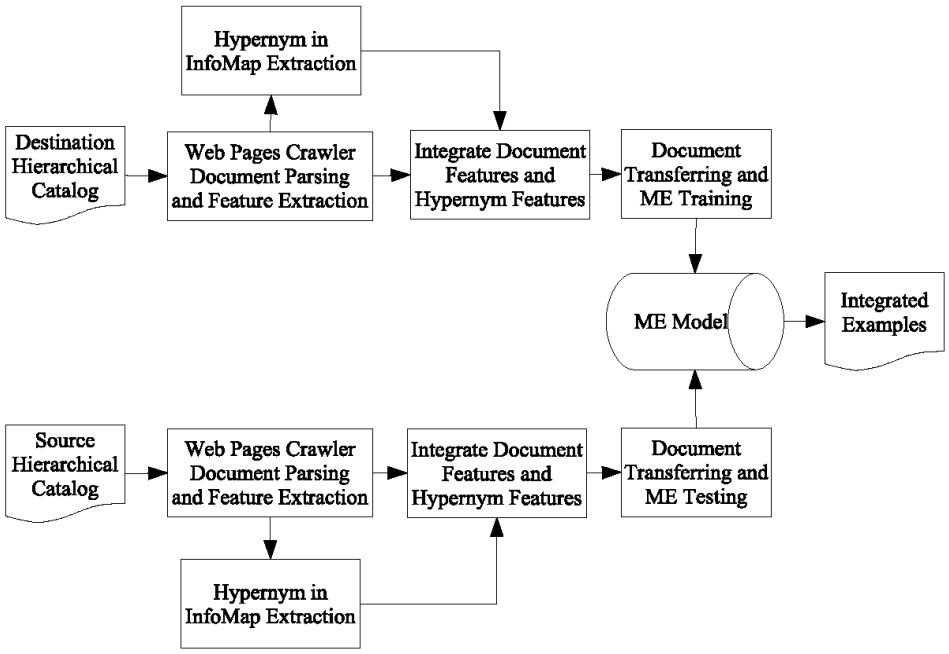
在此步驟中，我們將對 Web 網頁文件做一些前置處理。這些前置處理包含：移除 Web 網頁文件中的 HTML 標籤，script 網頁執行碼和文字。移除文章中的 stop word，並用 Porter 的演算法對每個詞做 stemming 處理。處理後的單詞做爲文件的特徵詞。

3、文件中特徵詞權重計算

特徵詞權重的計算方式，一般在實作上有 TF-IDF、TF 與 TF/*sum*(TF)等方法。我們考慮到 TF-IDF 在計算時會因目錄的變動而常常需要更新，在實際使用上將會花費許多計算時間，因此在本研究中，我們使用如式 (3.10)來計算 TF/*sum*(TF) 特徵詞權重計算。

$$f_i = \frac{TF(w_i, d)}{\sum_{i=0}^{n} TF(w_i, d)} \qquad (3.10)$$

其中 w_i 表示是第 i 個特徵詞，d 爲文件，$TF(w_i,d)$爲 w_i 這單字或片語在文件 d 中所出現的頻率，總共有 n 個特徵詞。



圖四、引用外部語義庫進行階層式目錄整合流程圖

4、類別關鍵詞選取

如果將文件中所有的特徵詞都將其上位詞加入，將會使文件充滿過多無關重要的資訊而影響整合效果。因此在進行語義擴展的時候，必須先對文件中的特徵詞篩選出重要的關鍵詞，再針對這些具有代表性的關鍵詞來進行語義擴充，也就是將他們的上位詞加入特徵詞集合當中，來輔助整合效果。同時，爲避免在同一類別中，不同文件之間的關鍵詞仍可能存有分歧，因此我們針對一個目錄類別來選取可以代表該類別的關鍵詞。

在過往研究中可以發現，以 Correlation Coefficient 的方式，可以擷取到具有代表性的關鍵詞 [15]。因此我們利用 Correlation Coefficient 的方式來抽取關鍵詞，再進行語義擴展。也就是是針對同一個目錄底下所有文件中的特徵詞，透過底下方程式(3.11)計算出能夠代表此目錄當中每一個特徵詞的權重。在 (3.11) 式當中的 TP 表示類別 C 之文件含有特徵詞 T 的文件數，FP 表示 C 以外其他類別之文件含有特徵詞 T 的文件數，FN 表示 C 以外其他類別之文件不含有特徵詞 T 的文件數，TN 表示類別 C 之文件不含有特徵詞 T 的文件數表示在目錄底下有包含或者沒有包含此特徵詞的正負向文件個數。所算出來的 Co(T,C) 表示在類別 C 中特徵詞 T 與類別 C 的 Correlation Coefficient。透過 Correlation Coefficient 方法可以精確的選擇出一些針對此目錄較具代表性的特徵詞。

$$\mathrm{Co}(\mathrm{T},\mathrm{C}) = \frac{(\mathrm{TP}\times\mathrm{TN}+\mathrm{FN}\times\mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})(\mathrm{FN}+\mathrm{TN})}} \qquad (3.11)$$

5、關鍵詞語義擴展

經由以上式 (3.11) 所計算出來的權重，反映出該特徵詞在該類別中的關連性，因此我們擷取出權重最高的 5 個特徵詞，視爲可以代表該類別的關鍵詞，查詢他們在外部

語義庫 InfoMap 的上位詞資訊，進而將這些上位詞算出權重後加入特徵值空間中。例如某一目錄類別最高的 5 個特徵詞爲 output, signal, circuit, input, frequency，透過 InfoMap，我們可得到 signal, signaling, sign, communication, abstraction, relation 等 6 個上位詞，再用下式 (3.12) 計算出第 i 個上位詞的權重 HW_i。由這些權重，我們可以對一個文件 d 來決定它的上位詞特徵向量 H_d。

$$HW_i = \frac{HF_i}{\sum_{i=0}^{n} HF_i} \qquad (3.12)$$

其中 HF_i 代表從這 5 個特徵詞所取出的所有 n 個上位詞中，第 i 個上位詞出現的頻率。

6、階層架構的索引典資訊計算

在過往研究中發現，階層目錄中每一類別的標籤資訊架構，例如 Google 目錄中的 “recreation/autos/” 該類別的說明，可視爲該目錄的一個索引典（thesaurus），在目錄整合上相當有幫助 [6,9]。此資訊可以是該類別的名稱，亦或者是該類別說明。在實驗中，我們以該類別的說明來組成其索引典。

因此，針對一個文件 d，我們將它在目錄的索引典特徵向量 L_d，與經由關鍵詞語義擴展後所得的上位詞特徵向量 H_d，以及原先文件中的特徵向量 F_d 一起整合，如式 (3.13) 所示。此外，透過 λ 、α 來取得權重的平衡，此方法所計算出來的特徵向量 FE_d，即可加以提升整合正確性。當中的 λ 主要是調整 [6,9] 所提出加入階層式目錄資訊。如表一中，文件所存在的目錄資訊權重爲 $1/2^1$，再上一層的目錄資訊權重爲 $1/2^2$，以指數遞減的方法來計算。另一方面，α 是調整文件當中本身特徵詞與從外部語義庫所取出的上位詞資訊。將每個特徵詞的權重計算出來，在目錄整合時，可以依照較高權重的特徵詞加強來源端的資訊，進而提升整合效能。

$$FE_d = \lambda \frac{L_i}{\sum_{i=0}^{n} L_i} L_d + (1-\lambda)\left[\alpha \times H_d + (1-\alpha) \times F_d\right] \qquad (3.13)$$

表一、階層式目錄標籤權重

Hierarchical	**Label Weight**
Document Level L_0	$1/2^0$
One Level Upper L_1	$1/2^1$
Two Level Upper L_2	$1/2^2$
.	.
.	.
.	.
nLevel Upper L_n	$1/2^n$

7、目錄整合

最後，將上述的特徵值轉換成 Maximum Entropy 模型分類器的資料格式，進行目錄整合工作。每一份文件都會有自己本身的目錄資訊和＜*feature*, *weight*＞配對而成立。因此每一份文件被定義爲＜*line*＞=＜*target*＞＜*feature*＞： ＜*value*＞...＜*feature*＞：＜*value*＞。當中的 target 定義爲 1（表示是正向文件）或是－1（表示是負向文件），＜*feature*

＞爲文件當中的特徵詞，＜*value*＞爲特徵詞的權重。

四、目錄整合實驗

爲了瞭解這些改進方式的效能，我們以真實的目錄來進行實驗，分別自 Google 和 Yahoo!取得部份目錄。在實驗中，我們所採用的外部語義庫是 InfoMap。在實驗上，我們依據 ECI 方法 [9] ，並加以運用到 SVM (ECI-SVM) 與 ME (ECI-ME) 分類器中。此外，在 ECI-ME 上，再進行關鍵詞語義擴展 (KSE-ME)，以下將進一步說明各項細節。

（一）、實驗環境

1、資料集

在實驗的部份，我們從兩個實際目錄：Google 和 Yahoo!，分別取得 5 個分項階層目錄來當作資料集，表二和表三是這些階層目錄的根節點目錄名稱。表四是這 5 個階層目錄當中所擷取下來文件的數量和類別數量。其數量依 5 個目錄 Autos、Movies、Outdoors、Photo 和 Software 來區分。

表二、Yahoo!中目錄的根節點位址

Category	**Matched URL**
Autos	http://dir.yahoo.com/recreation/automotive/
Movies	http://dir.yahoo.com/entertainment/movies and film/
Outdoors	http://dir.yahoo.com/recreation/outdoors/
Photos	http://dir.yahoo.com/arts/visual arts/photography/
Software	http://dir.yahoo.com/computers and internet/software/

表三、Google 中目錄的根節點位址

Category	**Matched URL**
Autos	http://dir.google.com/top/recreation/autos/
Movies	http://dir.google.com/top/arts/movies/
Outdoors	http://dir.google.com/top/recreation/outdoors/
Photos	http://dir.google.com/top/arts/photography/
Software	http://dir.google.com/top/computers/software/

透過每個網頁目錄節點中所包含的次目錄和對外連結，依此建立目錄之間的上下層關係和向外的連結。整合的過程會依向外的連結，取得文件。以 Google 目錄爲例，我們進行實驗的文件當中，差集的部份｜Google－Yahoo!｜是訓練文件（G-Y），交集的部份｜Google∩Yahoo!｜是測試文件（G Test）。

表四、實驗中所用到的目錄類別數，以及訓練與測試文件數量

	Yahoo!			**Google**		
	Y-G	Y Class	Y Test	G-Y	G Class	G Test
Autos	1681	24	436	1096	12	426
Movies	7255	27	1344	5188	26	1422
Outdoors	1579	19	210	2396	16	208
Photo	1304	23	218	615	9	235
Software	1876	15	710	5829	27	641
Total	13965	108	2918	15124	90	2932

2、測量方式

在實驗當中我們的測量整合時的精確率 (P，precision)與召回率 (R，recall)，以及 $F_1 = 2PR/(P+R)$，來比較不同系統的成效。在目前實驗中，我們允許一份文件可以被整合到多個目錄類別中。因此 Precision 的算法是（正確分類的文件數/所有分至該類的文件數），Recall 的算法是（正確分類的文件數/應該分至該類的文件數）。Recall 也可看成是過往研究中的整合精確度。

除此之外，我們也同時考量多個類別的整體表現，因此也使用 micro-average 與 macro-average 兩種平均方法。Micro-average 由於是全部文件一起累加統計，不分類別，因此容易受到佔大多數的大件類別影響。相對地，Macro-average 考慮每個類別的成效後再做平均，因此容易受到大量的小類別而影響。

3、外部語義庫

在目前研究中，我們使用的外部語義庫是 InfoMap [2] 來尋找上位詞。事實上，其他的外部語義庫也可以使用，例如 WordNet，然而在過往研究中發現，使用 InfoMap 所擴展的結果與使用 WordNet 的結果所得的成效相當接近 [15]。因此換用 WordNet 可能也會有與目前類似的分類表現。未來在我們的研究計畫中，預備將進一步探討外部語義庫的品質對於整合品質的影響。

4、分類器

在分類器上，我們使用 SVM^{light} 來實作 ECI-SVM 的整合機制，使用 linear kernel 以及預設的參數，版本爲 5.00 版。ME 分類器則使用 Edinburgh 大學的 ME 工具。所使用的 ME 工具的版本爲 20041229 版。

（二）、實驗結果與討論

在實驗中，我們設定不同的 λ 值。爲了解不同 λ 值的影響，在來源目錄中，我們將 λ_s 值設定從 0.00 到 1.00，而在目的目錄中，λ_d 值設定爲 0.00，0.01，0.05 分別來看它們的成效如何。如此取的原因是，在過往研究中發現 λ 值過大其實會將 Recall 值降低 [6,9]。在我們實驗中也確實有如此情況。另一個參數 α 值，我們目前只作了一些初步測試，由於篇幅的關係，此處僅報告 α=0.1 的情況(KSE-ME)。在表八與表九的結果中，我們可以看到隨著 λ_s 的增加，Macro-Recall 與 Micro-Recall 都同時下降。

表六到表十一顯示由 Google 到 Yahoo!階層式目錄的效能。在表六與表七當中，可發現 Macro-Precision 與 Micro-Precision 的效果都不夠良好。這是因爲在實驗中，我們允許一份文件可以分至多個目錄類別所致。所以在 λ_s 值較低時，由於缺乏階層架構索引典的輔助，False-positive 的比例會普遍升高，使 Precision 表現不佳。但我們可以看出，隨著 λ_s 值提高，False-positive 的比例逐步下降，使 Precision 逐步升高。同時，我們也可以發現，採用關鍵詞語義擴展的 KSE-ME，在 Precision 上有最好的表現。

在表八與表九中，我們可以發現，KSE-ME 在 Recall 的表現上，普遍都要比 ECI-ME 爲佳。至於 ECI-SVM，雖然其 Precision 的表現不如 ECI-ME 與 KSE-ME，但在 Recall 的表現上，反而普遍有最好的表現。但可從表中看出，當 λ_s=0.05 這個常在實驗中使用的值時，KSE-ME 依然有領先的表現。

若從 F_1 的表現上來看，表十和表十一顯示 KSE-ME 都有不錯的成效，比 ECI-ME 與 ECI-SVM 的 F_1 表現都來得好。因此，從目前初步的實驗可以得知，利用階層架構索引典資訊以及關鍵詞語義擴展，階層式目錄整合的效能可以有效的提升。

表六、階層式目錄整合的 Macro-Precision

		λ_d								
		ECI-SVM			ECI-ME			KSE-ME		
macroP		0.00	0.01	0.05	0.00	0.01	0.05	0.00	0.01	0.05
	0.00	1.08%	1.09%	1.24%	1.11%	1.12%	1.21%	1.32%	1.32%	1.42%
	0.10		1.18%	3.12%		1.28%	2.77%		2.36%	9.05%
	0.20		1.34%	8.37%		1.74%	14.80%		4.93%	26.55%
	0.30		1.82%	13.18%		3.04%	30.04%		10.09%	36.34%
	0.40		3.50%	15.32%		5.98%	35.96%		17.62%	42.42%
λ_s	0.50		6.87%	16.25%		12.98%	38.62%		22.82%	47.55%
	0.60		9.45%	16.81%		20.91%	40.42%		30.08%	49.37%
	0.70		12.68%	17.20%		23.85%	41.34%		33.96%	50.12%
	0.80		14.69%	17.52%		24.95%	41.81%		35.31%	51.01%
	0.90		16.86%	17.90%		24.38%	42.17%		35.52%	51.96%
	1.00		17.64%	17.93%		24.93%	42.40%		36.05%	52.72%

表七、階層式目錄整合的 Micro-Precision

		λ_d								
		ECI-SVM			ECI-ME			KSE-ME		
microP		0.00	0.01	0.05	0.00	0.01	0.05	0.00	0.01	0.05
	0.00	1.08%	1.09%	1.23%	1.11%	1.12%	1.18%	1.29%	1.30%	1.40%
	0.10		1.18%	2.43%		1.28%	2.26%		1.75%	4.80%
	0.20		1.34%	6.34%		1.71%	9.82%		2.84%	18.52%
	0.30		1.80%	9.86%		2.84%	23.47%		4.81%	29.10%
	0.40		3.29%	11.33%		5.58%	30.47%		6.94%	35.71%
λ_s	0.50		6.21%	12.24%		10.93%	34.10%		11.32%	39.81%
	0.60		8.90%	12.87%		18.58%	37.24%		19.35%	40.81%
	0.70		11.71%	13.46%		20.91%	38.79%		25.17%	41.34%
	0.80		13.27%	13.95%		21.36%	39.32%		26.97%	41.97%
	0.90		14.76%	14.26%		20.29%	39.67%		27.37%	42.41%
	1.00		15.27%	14.33%		20.33%	39.92%		28.13%	43.61%

表八、階層式目錄整合的 Macro-Recall

		λ_d								
		ECI-SVM			ECI-ME			KSE-ME		
macroR		0.00	0.01	0.05	0.00	0.01	0.05	0.00	0.01	0.05
	0.00	94.66%	94.66%	88.48%	95.80%	96.05%	94.42%	97.43%	97.69%	97.69%
	0.05		95.83%	94.31%		95.96%	92.09%		96.99%	94.80%
	0.10		95.67%	92.44%		92.98%	85.16%		93.80%	89.98%
	0.20		92.41%	86.91%		87.57%	79.65%		89.21%	83.12%
	0.30		89.58%	83.66%		84.40%	76.92%		85.06%	81.02%
	0.40		86.54%	81.54%		81.89%	76.40%		81.68%	78.32%
λ_s	0.50		84.11%	80.12%		79.85%	75.54%		78.66%	76.99%
	0.60		80.72%	79.28%		78.20%	74.87%		76.79%	76.46%
	0.70		79.37%	78.53%		76.70%	74.35%		75.42%	76.12%
	0.80		77.69%	77.89%		76.22%	73.49%		74.08%	76.07%
	0.90		77.18%	77.82%		74.67%	72.94%		73.49%	76.00%
	1.00		76.56%	77.79%		73.22%	72.83%		73.31%	75.93%

表九、階層式目錄整合的 Micro-Recall

		λ_d								
		ECI-SVM			ECI-ME			KSE-ME		
microR		0.00	0.01	0.05	0.00	0.01	0.05	0.00	0.01	0.05
	0.00	94.76%	94.79%	87.91%	96.20%	96.44%	94.59%	97.26%	97.40%	97.60%
	0.05		96.23%	95.14%		96.85%	94.35%		97.33%	96.27%
	0.10		96.37%	93.63%		95.51%	90.00%		95.68%	92.36%
	0.20		94.69%	89.69%		91.88%	85.41%		91.92%	86.06%
	0.30		92.74%	87.39%		88.69%	82.46%		87.87%	83.59%
	0.40		90.10%	85.58%		86.47%	82.01%		84.65%	81.02%
λ_s	0.50		88.08%	84.52%		84.96%	80.92%		81.67%	79.86%
	0.60		85.17%	83.66%		83.86%	80.40%		79.72%	79.48%
	0.70		83.56%	82.91%		82.87%	79.82%		78.07%	79.10%
	0.80		81.71%	82.19%		82.53%	79.24%		77.05%	79.07%
	0.90		81.06%	82.05%		81.64%	78.66%		76.50%	78.93%
	1.00		80.54%	82.01%		79.38%	78.59%		76.26%	78.76%

表十、階層式目錄整合的 Macro-F_1

		λ_d								
		ECI-SVM			ECI-ME			KSE-ME		
macroF		0.00	0.01	0.05	0.00	0.01	0.05	0.00	0.01	0.05
	0.00	2.13%	2.15%	2.44%	2.19%	2.21%	2.38%	2.60%	2.61%	2.81%
	0.10		2.32%	5.92%		2.52%	5.27%		4.60%	16.44%
	0.20		2.62%	14.59%		3.39%	23.59%		9.34%	40.24%
	0.30		3.54%	21.99%		5.80%	41.84%		18.04%	50.17%
	0.40		6.65%	24.85%		11.01%	47.65%		28.99%	55.03%
λ_s	0.50		12.42%	25.99%		21.91%	50.03%		35.37%	58.79%
	0.60		16.57%	26.66%		32.55%	51.53%		43.23%	60.00%
	0.70		21.62%	27.14%		35.89%	52.26%		46.83%	60.44%
	0.80		24.34%	27.52%		37.01%	52.48%		47.82%	61.07%
	0.90		26.99%	28.00%		36.04%	52.67%		47.89%	61.72%
	1.00		27.81%	28.05%		36.36%	52.83%		48.34%	62.23%

表十一、階層式目錄整合的 Micro-F_1

		λ_d								
		ECI-SVM			ECI-ME			KSE-ME		
microF		0.00	0.01	0.05	0.00	0.01	0.05	0.00	0.01	0.05
	0.00	2.14%	2.16%	2.43%	2.19%	2.22%	2.33%	2.54%	2.56%	2.77%
	0.10		2.33%	4.74%		2.53%	4.40%		3.44%	9.13%
	0.20		2.63%	11.85%		3.35%	17.61%		5.51%	30.48%
	0.30		3.54%	17.72%		5.50%	36.54%		9.13%	43.17%
	0.40		6.35%	20.01%		10.48%	44.43%		12.83%	49.57%
λ_s	0.50		11.60%	21.38%		19.37%	47.98%		19.88%	53.13%
	0.60		16.11%	22.31%		30.42%	50.90%		31.14%	53.93%
	0.70		20.54%	23.16%		33.40%	52.21%		38.07%	54.30%
	0.80		22.84%	23.86%		33.93%	52.56%		39.95%	54.83%
	0.90		24.97%	24.30%		32.50%	52.74%		40.31%	55.17%
	1.00		25.67%	24.40%		32.37%	52.94%		41.10%	56.13%

五、結論

隨著網路資訊蓬勃發展與快速整合和交換的需求，目錄整合在許多領域已經成爲重要的議題，在過往研究中，已從初步的攤平式目錄整合，逐漸深入到階層式目錄整合的討論。因此在本論文中，我們針對階層式目錄整合，再進一步討論整合效能加強的方式。我們架構於之前研究的階層架構索引典資訊的加強方法，另外提出使用關鍵詞語義擴展

的方式，來進一步增進階層式目錄整合效能。

在我們的初步實驗中可以看出，使用階層架構索引典資訊與關鍵詞語義擴展這兩個方式的 KSE-ME，在 Precision 上有很好的表現，在 Recall 上雖不能普遍比 ECI-SVM 來得好，但也還是普遍比 ECI-ME 來得佳。在綜合考量的 F_1 的評估上，KSE-ME 具有最好的表現。

從我們目前的研究成果可以發現，如何在不降低 Recall 表現的同時，還能夠減少 False-Positive 的整合技術，仍待進一步討論。如此，當 Precision 也提高的時候，也將是階層式目錄整合技術趨於成熟，能夠運用在實際環境中的時候。未來我們也計畫進一步討論外部語義庫的品質，對於整合效能影響的關鍵因素，期待能探索出有效提升目錄整合效能的方式。

致謝

本研究感謝國科會計畫 NSC-96-2221-E-155-067 的支助，並感謝論文審查委員寶貴的建議。

參考文獻

[1] http://www.amazon.com.

[2] “Information mapping project,” Computational Semantics Laboratory, Standford University. [Online]. Available: http://infomap.stanford.edu/.

[3] R. Agrawal and R. Srikant. “On Integrating Catalogs.” In *Proceedings of the 10th WWW Conference. (WWW10)*, pp. 603–612, Hong Kong, May 2001.

[4] A. Berger. “The improved Iterative Scaling Algorithm: A Gentle Introduction.” *Technical report*, 1997.

[5] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. “A Maximum Entropy Approach to Natural Language Processing.” *Computational Linguistics,* pp. 39–71, 1996.

[6] I.-X. Chen, J.-C. Ho, and C.-Z. Yang. “On Hierarchical Web Catalog Integration with Conceptual Relationships in Thesaurus.” In *Proceedings of the 29th International ACM SIGIR (SIGIR 2006)*, pp. 635–636, Seattle, Washington, USA, 2006.

[7] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. “Learning to Map between Ontologies on the Semantic Web.” In *Proceedings of the 11th WWW Conf. (WWW2002)*, pp.662–673, Honolulu, Hawaii, 2002.

[8] S. Dumais, and H. Chen. “Hierarchical Classification of Web Content.” In *Proceedings of the 23rd Annual ACM Conf. on Research and Development in Information Retrieval (SIGIR’00)*, pp.256–263, Athens, Greece, 2000.

[9] J.-C. Ho, I.-X. Chen, and C.-Z. Yang. “Learning to Integrate Web Catalogs with Conceptual Relationships in Hierarchical Thesaurus.” In *Proceedings of the 3rd Asia Information Retrieval (AIRS 2006)*, pp. 217–229, Singapore, 2006.

[10] W. S. Lee and D. Zhang. “Web Taxonomy Integration through Co-Bootstrapping.” In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pp. 410–417, 2004.

[11] A. MaCallum, R. Rosenfeld, T. Mitchell, and A. Ng. “Improving Text Classification by Shrinkage in a Hierarchy of Classes.” In *Proceedings of the 15th International Conference on Machine Learning (ICML-98),* pp. 359–367, Madison, Wisconsin, 1998.

[12] K. Nigam, J. Lafferty, and A. McCallum. “Using Maximum Entropy for Text Classification.” In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61–67, Oct. 1999.

[13] S. Rajan, K. Punera, and J. Ghosh. “A Maximum Likelihood Framework for Integrating Taxonomies.” In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, pp.856–861*,* Pittsburgh, Pennsylvania.

[14] S. Sarawagi, S. Chakrabarti, and S. Godbole. “Cross-Training: Learning Probabilistic Mappings between Topics.” In *Proc. of the 9th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining*, pp. 177–186, 2003.

[15] Y.-H. Tseng, C.-J Lin, H.-H Chen, and Y.-I. Lin. “Toward Generic Title Generation for Clustered Documents.” In *Proceedings of the 3rd Asia Information Retrieval (AIRS 2006)*, pp. 145–157, 2006.

[16] C.-W. Wu, T.-H. Tsai, and W.-L. Hsu, “Learning to Integrate Web Taxonomies with Fine-Grained Relations: A Case Study Using Maximum Entropy Model.” In *Proceedings of the 2nd Asia Information Retrieval Symposium 2005 (AIRS 2005)*, pp. 190–205, Jeju Island, Korea, Oct. 2005.

[17] D. Zhang and W.S. Lee. “Web Taxonomy Integration using Support Vector Machines.” In *Proceedings of the 13th WWW Conference (WWW2004),* pp.472–481, New York, NY, May 2004.

Speech recognition of mandarin syllables using both linear predict coding cepstra and Mel frequency cepstra

黎自奮　Tze Fen Li
明道大學管理研究所
Institute of Management
Ming Dao University
tfli@mdu.edu.tw

張水清　Shui-Ching Chang
僑光技術學院資訊管理系
Department of Information Management
The Overseas Chinese Institute of Technology
monet@ocit.edu.tw

Abstract

This paper is to compare two most common features representing a speech word for speech recognition on the basis of accuracy, computation time, complexity and cost. The two features to represent a speech word are the linear predict coding cepstra (LPCC) and the Mel-frequency cepstrum coefficient (MFCC). The MFCC was shown to be more accurate than the LPCC in speech recognition using the dynamic time warping method. In this paper, the LPCC gives a recognition rate about 10% higher than the MFCC using the Bayes decision rule for classification and needs much less computational time to be extracted from speech signal waveform, i.e., the MFCC needs computational time 5.5 time as much as the LPCC does. The algorithm to compute a LPCC from a speech signal much simpler than a MFCC, which has many parameters to be adjusted to smooth the spectrum, performing a processing that is similar to be adjusted to smooth the spectrum, performing a processing that is similar to that executed by the human ear, but the LPCC is easily obtained by the least squares method using a set of recursive formula.

Key words: Bayes decision rule, linear predict coding, Mel-frequency cepstrum coefficient, signal processing, speech recognition.

1. Introduction

A speech recognition system basically contains extraction of features and classification of an utterance of an acoustical word. The measurements made on the speech waveform include energy, zero crossings, extrema count, formants, LPC cepstrum (LPCC) [1-4] and the Mel frequency cepstrum coefficient (MFCC) [5-8]. The LPC method provides a robust, reliable and accurate method for estimating the parameters that characterize the linear, time-varying system which is recently used to approximate the nonlinear, time-varying system of the speech waveform. The MFCC method uses the bank of filters scaled according to the Mel scale to smooth the spectrum, performing a processing that is similar to that

executed by the human ear. The filters with Mel scales spaced linearly at low frequencies and logarithmically at high frequencies are used to capture phonetically the characteristics of speech [8]. For recognition, Davis and Mermelstein [5] used the dynamic time warping algorithm to show that the performance of the MFCC was better than the LPCC.

In this paper, we use a simple technique [9] for speech data compression of the sequence of MFCC vectors and the sequence of LPCC vectors to obtain a matrix of feature values respectively. For speech recognition, we simply use a simplified Bayes decision rule with weighted variance, where each step is a simple calculation and which has the minimum probability of misclassification. In our study, there are two speech recognition experiments. In the first experiment, since both LPCC and MFCC are said to be robust and reliable to noise and estimation errors, our speech experiment is implemented in a noisy environment to test which feature is better on speech recognition. Pick up 9 female and 10 male students and each pronounces 10 digits once using a common (not high-quality) microphone. Some students pronounce mandarin syllables not very clearly, since we have several types of accents to pronounce the same mandarin syllables. In the second experiment, there are 87 students to pronounce the mandarin syllables in a quiet classroom, which are most commonly used in usual conversations. Our speech experiment is done like natural talking. Hence our speech system can be commonly used for all peoples and in all environments. The recognition rate using LPCC is significantly better than the rate using MFCC and the LPCC needs much less computational time to be extracted from speech signal waveform.

2. Bayes Decision Rules

Let $X = (X_1, ..., X_k)$ be the input feature vector of a speech data, which belongs to one of m categories (syllables) c_i, $i = 1, ..., m$. Consider the decision problem consisting of determining whether X belongs to c_i. Let $f(x \mid c_i)$ be the conditional density function of X given category c_i. Let θ_i be the prior probability of c_i such that $\Sigma_{i=1}^{m} \theta_i = 1$, i.e., the θ_i is the probability for the category c_i to occur. Let d be a decision rule. A simple loss function $L(c_i, d(x)), i = 1, ..., m,$ is used such that the loss $L(c_i, d(x)) = 1$ when $d(x) \neq c_i$ makes a wrong decision and the loss $L(c_i, d(x)) = 0$ when $d(x) = c_i$ makes a right decision. Let $\tau = (\theta_1, ..., \theta_m)$ and let $R(\tau, d)$ denote the risk function (the probability of misclassification) of d. Let $\Gamma_i, i = 1, ..., m$, be m regions separated by d in the k-dimensional domain of X, i.e., d decides c_i when $X \in \Gamma_i$. Then

$$R(\tau, d) = \sum_{i=1}^{m} \theta_i \int L(c_i, d(x)) f(x \mid c_i) dx$$
$$= \sum_{i=1}^{m} \theta_i \int_{\Gamma_i^c} f(x \mid c_i) dx \qquad (2.1)$$

where Γ_i^c is the complement of Γ_i. Let D be the family of all decision rules which

separate m categories. Let the minimum probability of misclassification be denoted by

$$R(\tau) = \inf_{d \in D} R(\tau, d) \tag{2.2}$$

A decision rule d_τ which satisfies (2.2) is called the Bayes decision rule with respect to the prior distribution τ and is given in (2.3) [10]. We state the Bayes decision rule in the following theorem.

Theorem 2.1. [10] The Bayes decision rule with respect to τ is defined by

$$d_\tau(x) = c_i \quad \text{if} \quad \theta_i f(x \mid c_i) > \theta_j f(x \mid c_j) \tag{2.3}$$

for all $j \neq i$,i.e., $\Gamma_i = \{x \mid \theta_i f(x \mid c_i) > \theta_j f(x \mid c_j)\}$ for all $j \neq i$.

Note that if $\theta_i = 1/m$, $i = 1, ..., m$, the Bayes decision rule (2.3) become a ML classifier.

3. Feature Extraction

3.1 Preprocessing Speech Signal

Since our speech recognition experiment is implemented in a noisy environment, the speech data must contain noise. We propose two simple methods to eliminate noise. One way is to use the sample variance of a fixed number of sequential samples to detect the real speech signal, i.e., the samples with small variance does not contain speech signal. Another way is to compute the sum of the absolute values of difference of two consecutive samples in a fixed number of sequential speech samples, i.e., the speech data with small absolute value do not contain real speech signal. In our speech recognition experiment, the latter provides slightly faster and more accurate speech recognition.

3.2 Mel-Frequency Cepstrum Coefficient (MFCC)

The MFCC is a representation defined as the real cepstrum of a windowed short-time signal derived from the fast Fourier transform of the speech signal. In the MFCC, a nonlinear frequency scale is used, which approximates the behavior of the auditory system. The discrete cosine transform of the real logarithm of the short-time energy spectrum expressed on this nonlinear frequency scale is called the MFCC. Davis and Mermelstein [5] showed the MFCC representation to be beneficial for speech recognition. We detail the MFCC as follows [8]:

Let $s[n]$ denote the N samples of a speech waveform. The discrete Fourier transform (DFT) $X[k]$ of the speech signal is defined by

$$X[k] = \sum_{n=0}^{N-1} s[n] e^{-j2\pi nk/N}, \qquad 0 \le k < N \tag{3.1}$$

We define a filterbank with M filters $(m = 1, ..., M)$, where filter m is a triangular filter given

$$\begin{aligned}
&H[m,k] = 0 && if \quad k < f[m-1] \\
&H[m,k] = (k - f[m-1])/(f[m] - f[m-1]) && if \quad f[m-1] \le k \le f[m] \\
&H[m,k] = (f[m+1] - k)/(f[m+1] - f[m]) && if \quad f[m] \le k \le f[m+1] \\
&H[m,k] = 0 && if \quad k > f[m+1]
\end{aligned} \tag{3.2}$$

which satisfies $\Sigma_{m=1}^{M} H[m,k] = 1,\ k = 0, 1, ..., N-1$.

Let f_l and f_h be the lowest and highest frequencies of the filterbank in H_z and let F_s be the sampling frequency in H_z. The boundary points $f[m]$ are uniformly spaced in the mel-scale:

$$f[m] = (\frac{N}{F_s}) B^{-1}(B(f_l) + m \frac{B(f_h) - B(f_l)}{M+1}) \tag{3.3}$$

where $B(f) = 1125 ln(1 + f/700)$ and $B^{-1}(b) = 700(e^{(b/1125)} - 1)$. The log-energy is computed by

$$S[m] = ln\{\sum_{k=0}^{N-1} |X[k]|^2 H[m,k]\}, \qquad 0 < m \le M. \tag{3.4}$$

The MFCC is then the discrete cosine transform of the M filters outputs:

$$c(n) = \sum_{m=0}^{M-1} S[m] cos(\pi n(m - 0.5)/M) \qquad 0 \le n < M \tag{3.5}$$

For speech recognition, normally, the number M of filters is from 10 to 20 and the MFCC produced from the first few filters are the most effective in recognition. In our experiment, we use $M = 12$

3.3 Linear Predict Coding Cepstrum (LPCC)

The MFCC was proved to be better than the LPC cepstrum for recognition by using the dynamic time warping (DTW) method [5], but the computational complexity for the MFCC is much heavier than that of the LPC cepstrum. The LPC coefficients can be easily obtained by Durbin's recursive procedure [11-13] and their cepstra can be quickly

found by another recursive equations [11-13] without computing the discrete Fourier transform (DFT) and the inverse DFT, which are computationally complex and time consuming.

The LPC method can also provide a robust, reliable and accurate method for estimating the parameters that characterize the linear and time-varying system [3, 11-13]. The following is a brief discussion of LPC method. It is assumed [13] that the sampled speech waveform $\hat{s}(n)$ can be linearly predicted from the past p samples of $s(n)$. Let

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{3.6}$$

where p is the number of the past samples and let E be the squared difference between $s(n)$ and $\hat{s}(n)$ over N samples of $s(n)$, i.e.,

$$E = \sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2. \tag{3.7}$$

The unknown a_k, $k = 1, ..., p$, are called the LPC coefficients and can be solved by the least square method. The most efficient method known for obtaining the LPC coefficients is Durbin's recursive procedure [3, 11-13]. Here in our experiments, $p = 12$, because the cepstra in the last few elements are almost zeros.

Both LPCC and MFCC are the method to compress or simplify the huge speech data $s(n)$ of a syllable into a simple data without loss of speech information. The LPCC is more or less like the sufficient statistics of a random samples in statistics [14]. The LPC coefficients a_k, $k = 1, ..., p$, are actually the least squares estimators of the regression coefficients, i.e., the minimum variance linear estimators a_k of the regression coefficients [14]. The huge data of a frame are well-represented by the LPC coefficients unless LPC coefficients are too small, i.e., the estimates a_k of the regression coefficients are not significant as compared with noise. On the other hand, to produce a MFCC, one has to obtain the DFT of a frame of the huge data and after the Mel filter banks smooth the spectrum, performs the inverse DFT on the logarithm of the magnitude of filter bank output. It seems to us that the formula in (3.1)-(3.5) to produce a MFCC are a little arbitrarily or artificially or experimentally adjusted for human ears. There is no theoretical theory to support the MFCC to well represent a syllable without loss of information. Hence in this paper, we create a huge database from common mandarin sentences to obtain the recognition rates using the LPCC and MFCC respectively.

3.4 Feature Extraction [9]

Our method to extract the feature from LPCC (MFCC) is quite simple. Let $x(k) = (x(k)_1, ..., x(k)_p)$, $k = 1, ..., n$, be the LPCC (MFCC) vector of size $p = 12$ for the k-th frame of a speech waveform, where n is the length of the LPCC (MFCC) sequence and p is the number of LPC coefficients in each frame. Normally, if a speaker does not intentionally elongate pronunciation, a mandarin syllable has 30-70 vectors of LPCC (MFCC).

Since an utterance of a syllable is composed of two basic parts: stable part and feature part. In the feature parts, the vectors have a dramatic change between two consecutive vectors, representing the unique characteristics of the syllable utterance and in the stable parts, the vectors stay about the same. Even if the same speaker utters the same syllable, the duration of stable parts of the sequence of LPCC (MFCC) vectors changes every time with nonlinear expansion and contraction and hence the duration of the portion of feature vectors and duration of stable parts are different every time. Therefore, the duration of stable parts is contracted such that the compressed speech waveform has about the same length of the sequence of the vectors. Li [9] proposed several simple compression techniques to contract the stable parts of the sequence of vectors. We state a simple one with good recognition rate as follows:

Let $x(k) = (x(k)_1, ..., x(k)_p),\ k = 1, ..., n$, be the k-th vector of a LPCC (MFCC) sequence with n vectors, which represents a mandarin syllable. Let the difference of two consecutive vectors be denoted by

$$D(k) = \sum_{i=1}^{p} | x(k)_i - x(k-1)_i |, \quad k = 2, ..., n. \tag{3.8}$$

In order to accurately identify the syllable utterance, a compression process must first be performed to remove the stable and flat portion in the sequence of vectors. A LPCC (MFCC) vector is removed if its absolute difference $D(k)$ from the previous vector $x(k-1)$ is too small. In this study, a squared difference criterion is also used to remove the stable and flat portion of the sequence. The criterion is expressed as follows:

$$D(k) = \sum_{i=1}^{p} [x(k)_i - x(k-1)_i]^2, \quad k = 2, ..., n. \tag{3.9}$$

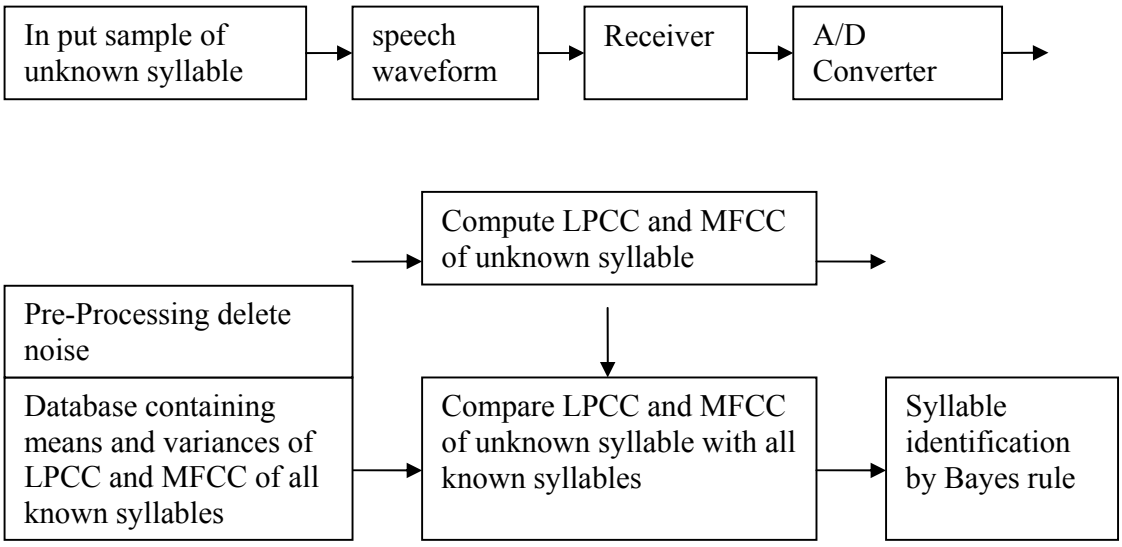
Let $x'(k),\ k = 1, ..., m(< n)$, be the new sequence of LPCC (MFCC) vectors after deletion. We think that the first part (about first 40 vectors) of an utterance of a mandarin syllable contains main features which can most represent the syllable and the rest of the sequence contains the "tail" sound, which has a variable length. If a speaker intentionally elongates pronunciation of a syllable, the speaker only increases the tail part of the sequence. The length of the feature part stays about the same. As in [9], we partition the feature part (first 40 vectors of the new sequence) into 8 equal segments and partition the tail part with variable length into two equal segments. If the length of the new sequence of vectors representing a syllable is less than 40, we neglect the tail sound and partition the new sequence into 10 equal segments. The average value of the LPCC (MFCC) in each segment is used as a feature value. Note that the average values of samples tend to have a normal distribution. This compression produces 12×10 feature values for each mandarin syllable.

4. Experimental Results

There are two speech recognitions implemented in our study. One is the digit recognition in a noisy environment and the other is the speech recognition on the mandarin monosyllables which are most commonly used in general conversations.

The following is a flow chart to show the speech recognition on a syllable.

Figure 1. Flowchart of a syllable recognition



4.1 The Digit Recognition

The digit recognition is implemented in a noisy environment, a classroom with windows open, which has noise from students inside classroom and from students and autos on the street outside classroom. The database of 10 mandarin digits is created by 19 persons (9 female and 10 male students) who pronounce 10 digits (0-9) once. The speech signal of a mandarin monosyllable is sampled at $10\,kHz$. A Hamming window with a width of 25.6 *ms* is applied every 12.8 *ms* for our study. A 256 point Hamming window is used to select the data points to be analyzed.

In our experiments, we use this database to produce the LPCC (MFCC) and obtain a 12×10 matrix for each digit sample. On the average, the time to produce a MFCC using DFT and formula in Section 3.2 is 5.5 times as much as to produce a LPCC. Among 19 samples (pronounced by 19 students) of each mandarin digit, pick up one sample (from one student) for recognition and the rest of 18 samples (from the other 18 students) of the digit is used for training, i.e., the rest of 18 samples of this digit is used to estimate the parameters which represent the digit. Hence each of 19 students has to be tested, i.e., there are 19 testing samples for each digit.

Since the average value of samples tends to be normally distributed. In order to reduce computation for classification, we assume that all elements in the 12×10 matrix of feature values are stochastically independent. It was proved [15] that using weighted variance in the Bayes decision rule for each class may increase the recognition rate. Hence, the conditional normal density given syllable c_i with weighted variance c can be represented as

$$f(x_1,...,x_k \mid c_i) = \left[\prod_{l=1}^{k} \frac{1}{\sqrt{2\pi} c\sigma_{il}}\right] e^{-\frac{1}{2}\sum_{l=1}^{k}(\frac{x_l-\mu_{il}}{c\sigma_{il}})^2} \tag{4.1}$$

where $i = 1,...,m = 10$, k=12×10 and c is a weighted factor for the variance. Taking logarithm on both sides of (4.1), the Bayes decision rule (2.3) with equal prior on each syllable becomes

$$l(c_i) = \sum_{l=1}^{k} \ln(c\sigma_{il}) + \frac{1}{2}\sum_{l=1}^{k}(\frac{x_l-\mu_{il}}{c\sigma_{il}})^2,\ i = 1,...,m. \tag{4.2}$$

The Bayes decision rule (4.2) decides a syllable c_i with the least $l(c_i)$ to which the feature matrix $x = (x_1,...,x_k)$ belongs. For the Bayes decision rule, 18 samples of the syllable c_i are used for estimating its mean μ_{il} and variance σ_{il}^2. The weighted factor c is selected from 0.8 to 1.3.

Note that in the Bayes decision rule, a matrix of feature values representing the testing digit pronounced by one student is compared with 10 matrices of means representing 10 digits' parameters. The means are computed from the feature values pronounced by the rest of 18 students. Hence the feature values of the digits pronounced by the student to be tested are independent of the feature values of the digits pronounced by the other 18 students and in the training data to produce 10 matrices of means (each matrix represents one digit's parameters μ_{il}, $l = 1,...,k = 12\times10$), the feature values of 10 digits between any two persons of the other 18 students are mutually independent. Therefore, the Bayes rule uses simple normal distributions. Table 4.1 shows that the number of correct digits of 190 testing samples and the recognition rates are obtained using LPCC and MFCC features with absolute difference and squared difference criteria.

Table 4.1 Correct digit recognition rates

	absolute difference criterion		squared difference criterion	
	LPCC	MFCC	LPCC	MFCC
	total testing samples = 190			
19 students	181	178	182	179
	(95.3%)	(93.7%)	(95.8%)	(94.2%)
	total testing samples = 100			
10 students (pronounce most clearly)	100	96	100	96
	(100%)	(96%)	(100%)	(96%)

Table 4.1 also shows the misclassified digits pronounced by the 10 students who

pronounce most clearly and distinctly. Since all mandarin syllables are monosyllables, the speech wave for each monosyllable is short. If the monosyllables are not pronounced clearly, it is difficult to recognize by the human ear. Hence, to test the recognition ability of the Bayes decision rule, which should not be damaged by the ambiguous pronunciation, we select 10 students (4 female and 6 male) among 19 students, who pronounce most clearly and distinctly. As in the first speech experiment, 10 digits pronounced by each student are used for testing and 90 samples (9 samples for each digit) from the other 9 students are used for training the means and the variances of each digit. There are 10 testing samples for each digit. From the classification in digits, the LPCC for speech recognition is lightly better than the MFCC for two criteria (absolute and squared differences). After compression of a sequence of LPCC and MFCC vectors, the two compression criteria give about the same recognition rates, but the squared difference criterion takes less time to compute. The same speech recognition experiment was implemented in a quiet environment [15] and gave the correct digit recognition rate 98.6%. For the robustness to the noise, our results show that the LPCC gives a recognition rate no less than the MFCC. This contradicts to the results obtained by Davis and Mermelstein [5] in a quiet environment.

4.2 The Speech Recognition

In this speech recognition experiment, 87 students participate in the experiment. Each pronounces loudly and clearly several sentences of mandarin syllables, which are commonly used in the usual conversation in our life. We cut these sentences into single words (syllables). We select the syllables which have at least 9 samples, i.e., each syllable as a candidate for speech recognition should appear in the sentences at least 9 times. Hence there are 102 different syllables to be classified. The 102 syllables appear in the sentences from 9 to 45 times. There are totally 1644 samples for 102 syllables to be tested. This experiment is designed as in the digit recognition in the first experiment. Each of 1644 samples is tested and the other 1643 samples are used for training, i.e., the syllables with 9 samples have 8 samples for training and the syllable with 45 samples has 44 samples for training. To compress the speech wave of a syllable into a 12×10 matrix of feature values, we only use the absolute difference criterion, since in the first experiment on digit recognition, there are no difference on recognition rates between the absolute difference and the squared difference criteria. The simplified Bayes decision rule with $m = 102$ and the weighted factor $c = 1.2$ in (4.1) and (4.2) is used to classify 102 different mandarin syllables. Table 4.2 shows the results. The table shows that the LPCC feature has the recognition rate 0.9057 better than the rate 0.8102 obtained by the MFCC feature. The total time needed to compute the MFCC of 1644 samples based on the formula in Section 3 is 5.5 times as much as that needed to compute the LPCC of the same 1644 samples.

Both recognition rates are not high enough, since some syllables having only 8 samples for training have poor rates. Hence we increase the minimum number of samples for training to 10, i.e., we select the syllables from the sentences, which should have at least 11 samples (to appear at least 11 times in the sentences) as candidates for speech recognition. This restriction results in 91 different mandarin syllables with a total 1523 samples to be tested in speech recognition, i.e., each of 1523 samples is used for testing and the remain of 1522 samples are used to train 91 different syllables. The recognition rates are increased to 0.9140 for the LPCC and 0.8188 for the MFCC. The recognition results for 91 syllables are shown in Table 4.2.

The above recognition rates all show that the LPCC features a little higher than the MFCC. Hence we make a statistical hypothesis testing in our study. We adopt two nonparametric methods (McNemar test and Cochran Q-test) [16] to test if the LPCC is better than the MFCC. The McNemar test is to compare two rates provided by the LPCC and MFCC individually. We obtain the approximate standard normal *z*-value 7.4593 for 102 syllables and 7.3899 for 91 syllables. Both are strongly significant at the level $\alpha = 0.0001$.

The Cochran Q-test is to compare two features (MFCC and LPCC) if they are equally effective in classification. We obtain the approximate Chi-square (df =1) Q value 55.6411 for 102 syllables and 54.6104 for 91 syllables, which are both strongly significant at the level $\alpha = 0.0001$. Both tests show in Table 4.3. Obviously, the two nonparametric tests make a decision to favor the LPCC.

Table 4.2 Correct syllable recognition rates pronounced by 87 students

features	LPCC	MFCC
total samples=1644 for 102 different syllables		
correct samples	1489	1332
correct rates	90.57%	81.02%
total samples=1523 for 91 different syllables		
correct samples	1392	1247
correct rates	91.40%	81.88%

Tables 4.3 Statistical testing hypotheses

Mc Nemar test :

H_0 : two recognition rates are equal

$z - \text{value} = 7.4593$ for 102 syllables. $= 7.3899$ for 91 syllables

$p - values < 0.0001$ for both 102 syllables and 91 syllables

decision : reject H_0 for both tests at the level $\alpha = 0.0001$

Cochran's Q – test :

H_0 : LPCC and MFCC are equally effective in classifiction

Q – value = 55.64 for 102 syllables. = 54.61 for 91 syllables

$p - values < 0.0001$ for both 102 syllables and 91 syllables

decision : reject H_0 for both tests at the level $\alpha = 0.0001$

Discussions and Conclusion

In this paper, we have used two speech recognition experiments to test if the LPCC feature has a higher ability in classification of the mandarin monosyllables than the MFCC. The speech waveform of a mandarin syllable is extracted into a sequence of LPCC (MFCC) vectors and the sequence of vectors is then compressed into a matrix of LPCC (MFCC) values, which tend to have a normal distribution. Using the Bayes decision rule, we have found that in the first digit experiment, the mandarin digit recognition rate using LPCC feature is no less than the rate using the MFCC feature. In the second speech recognition experiment, we build a large amount of mandarin syllables, which are the most commonly used in usual conversations. From the nonparametric statistical analysis, the LPCC has a significant higher ability in classification than the MFCC. Furthermore, the LPCC feature needs much less computational time to be extracted from speech signal waveform than the MFCC.

References

[1]. S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-22(2), 135-141, 1974.

[2]. B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Amer., 50, 637-655, 1971.

[3]. J. Makhoul and J. Wolf, Linear Prediction and the Spectral Analysis of Speech, Bolt, Baranek, and Newman, Inc., Cambridge, Mass., Rep. 2304, 1972.

[4]. J. Tierney, "A study of LPC analysis of speech in additive noise", IEEE Trans. Acoust., Speech, Signal Processing, 28(4), 389-397, 1980.

[5]. S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoust., Speech, Signal Processing, 28(4), 357-366, 1980.

[6]. W. Q. Zhang, L. He, Y. L. Chow, R. Z. Yang, and Y. P. Su, "The study on distributed speech recognition system", IEEE 2000 ICASSP, 1431-1434.

[7]. T. Fukuda, M. Takigawa, and T. Nitta, "Peripheral feature for HMM-based speech recognition", IEEE 2001 ICASSP, 129-132.

[8]. X. D. Huang, A. Acero, and H. W. Hon, Spoken Language Processing-A guide to theory, algorithm, and system development, Prentice Hall, PTR, Upper Saddle River, New Jersey, USA, 2001.

[9]. T. F. Li, "Speech recognition of mandarin monosyllables", Pattern Recognition, 36, 2713-2721, 2003.
[10]. K. Fukunage, Introduction to Statistical Pattern Recognition, New York: Academic Press, 1990.
[11]. Sadaoki Furui, Digital Speech Processing, Synthesis and Recognition, Marcel Dekker, Inc., New York and Basel, 1989.
[12]. J. Durbin, The fitting of time-series models, Rev. Inst. Int. Statist., 28(3) (1960) 233-243.
[13]. L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, Englewood Cliffs, New Jersey, 1993.
[14]. S. S. Wilks, Mathematical Statistics, New York: J. Wiley and Sons, 1962.
[15]. T. F. Li and T. F. Lin, On probability distribution of feature values for speech digit recognition, Technique Report, Department of Applied Mathematics, Feng Chia University, Taichung, Taiwan, 1994.
[16]. W. W. Daniel, Applied Nonparametric Statistics, Georgia State University, 1979.

應用文件重排序與局部查詢擴展於中文文件檢索之研究

Improving Retrieval Effectiveness by Document Re-ranking and Local Expansion

王文祺、林伯慎

國立台灣科技大學資訊管理系

Email: M9409112@mail.ntust.edu.tw, bslin@cs.ntust.edu.tw

摘 要

在資訊檢索領域中，查詢擴展(Query Expansion)已成爲提升檢索系統效能的重要技術之一。而查詢擴展的技術中又以局部查詢擴展(Local Expansion)對於檢索效能的提昇最爲顯著。局部查詢擴展是分析初始檢索結果的前幾篇文件，從中選出擴展詞，但此方法有一項缺點，就是當這些文件中與查詢句不相關的文件較多時，所選出的擴展詞可能與查詢句不相關，若把這些擴展詞加入查詢句做檢索，會造成查詢偏移(Query Drift)的情形，降低檢索效能。

因此，在本論文中針對局部查詢擴展的缺點做改進，使用文件重排序(Document Re-ranking)的方法，在作擴展前先將初始檢索輸出的文件重新排序，讓相關的文件盡量往前排，以提升局部查詢擴展的精確性。本論文研究了三種文件重排序法，實驗結果顯示其皆能有效的提昇檢索效能。且我們更進一步地探討這三種重排序法間的互補性，將其作結合，實驗結果顯示重排序法間的結合能更有效的提昇檢索效能，將檢索的平均精確率(MAP)從 0.3727 提升至 0.3956，前 10 篇文件的精確率(P@10)從 0.4929 提升至 0.5595。

關鍵字：資訊檢索、局部查詢擴展、查詢偏移、文件重排序

1. 序論

在現今資訊量極爲龐大的時代，要在大量的資料中找出符合使用者需求的資料，是資訊檢索領域困難的課題。目前使用者與檢索系統之間的互動，多是以詞爲主，系統也多是以詞作爲檢索的單位。然而使用者下達檢索詞時，在表達特定

的檢索概念，可能會因爲使用不同的詞，而檢索出不同的結果，有些符合查詢概念的文件，可能會因爲字詞的使用不同，而無法被檢索出來。這都是「字詞差異」(Word Mismatch)造成的問題。

查詢擴展(Query Expansion)爲解決上述問題的方法，其基本概念是把一些與查詢主題相關的詞－稱作擴展詞(Expanded Words)－加入到原始查詢句中，擴展原查詢句的概念，以提高一些原本未被檢索到的文件被檢索出來的機率，使查詢結果更爲精確。在本論文中使用的查詢擴展方法是局部查詢擴展(Local Expansion)[7、10、11、13、14、18]，其概念是先對使用者所下達的查詢句做第一次的檢索，篩選出排名前面的文件進行分析，將這些文件中重要性較高的詞作爲擴展詞，加到原查詢句中，再進行檢索。由於擴展詞增加了查詢概念的涵蓋度，因此可以提昇檢索效能。

然而，局部查詢擴展有一樣弱點，就是當第一次的檢索結果不佳時，排名前面的文件中，包含相關文件的比率相對地比較低，如果從這些文件選取擴展詞，很可能會選出與查詢主題不相關的詞，而造成查詢的偏移(Query Drift)。爲了彌補局部查詢擴展的弱點，可以使用文件重排序(Document Re-ranking)的技術[7、9、12、15、16、17、18、19、20]，在第一次檢索結果輸出後，用更精確的演算法對輸出的文件重新排序，讓排序在前面的文件中，能涵蓋較多與查詢相關的文件，以改進擴展詞的品質，進而提升檢索的效能。

本論文的研究主要是利用文件重排序來改進局部查詢擴展。在論文中提出了三種文件重排序演算法，分別是概念查詢法、文件分群法與局部鏈結法。而實驗結果顯示，這三種文件重排序都能夠有效地提昇檢索效能，而其中又以局部鏈結法效果最佳。另外，利用重排序法之間的互補性而加以結合，可以進一步達到更好的效能。在同時結合三種方法的情況下，平均精確率可以從 0.3727 提升至 0.3956，前 10 篇文件精確率則可以從 0.4929 提升至 0.5595。

此外，我們也探討擴展詞的過濾，希望藉由過濾的方法，將與查詢句較不相關的擴展詞過濾掉，以提昇檢索效能。實驗結果也顯示了擴展詞的過濾確實對於檢索效能有正面的效益。

本論文的第二章中爲實驗語料介紹與基礎檢索模型。第三章中將介紹三種文件重排序的方法與重排序方法間的結合。第四章爲擴展詞過濾方法。第五章爲爲結論與未來研究方向。

2. 研究方法

2.1 實驗語料與評估準則

本論文使用的實驗語料爲中華民國計算語言學學會所發行的「CIRB030 中文資訊檢索測試集」(NTCIR-3 中文語料部份)，此測試集共包含三個部分：問題集、文件集和答案集。文件集皆爲一般的新聞文件，包含了七個部份，一共有 381375 篇新聞文件。問題集一共包含了 42 個查詢問題。圖 1 爲一個查詢的範例。在本

論文中的實驗均是以<DESC>標籤的內容作爲查詢句。在答案集中每份文件均標記了和查詢問題的相關度，分爲四個層級：非常相關、相關、部分相關與不相關。檢索輸出文件的判別方式分爲兩種，一種爲嚴謹相關（Rigid Relevance），也就是把「非常相關」和「相關」視爲相關，其它視爲不相關；另一種爲寬鬆相關（Relax Relevance），則是把「非常相關」、「相關」和「部分相關」均視爲相關。本論文中均以“寬鬆相關”作爲判別標準。

```
<TOPIC>
<NUM>001</NUM>
<SLANG>CH</SLANG>
<TLANG>CH</TLANG>
<TITLE>漢代文物大展</TITLE>
<DESC>
查詢故宮博物院所舉辦之千禧漢代文物大展相關內容
</DESC>
<NARR>
台灣的故宮博物院是著名的典藏中國寶物的博物館，有關漢朝的典藏品展現了西元前206年到西元220年，中國漢朝的強盛與偉大。對於故宮博物院所舉辦之千禧漢代文物大展之說明，例如展出的文物種類、對於展出文物之介紹、展出時間、故宮的籌畫過程、合作單位等，以及展出後的成果與民眾的反應視為相關。非本次展覽內容之漢代文物介紹，以及其他展覽活動之介紹視為不相關。
</NARR>
<CONC>
漢代，文物大展，故宮博物院，歷史
</CONC>
</TOPIC>
```

圖 1 CIRB030 問題集之範例

在評估方法上本論文採用 TREC 所定的標準評估程式(TREC_EVAL) [1]來評估檢索效能，使用其中的「平均精確率(MAP)」和「前 N 篇文件的精確率(Precision: At N docs)」作爲效能指標，其定義如下：

• 平均精確率(MAP)：先對各個查詢問題，計算檢索出文件的平均精確率(AP)，再對所有查詢問題作平均。其計算方法如下公式：

$$AP=\frac{1}{r}\sum_{i=1}^{r}\frac{i}{Doc(i)} \tag{1}$$

$$MAP=\frac{1}{Q}\sum_{j=1}^{Q}AP_j \tag{2}$$

r 是檢索系統對該查詢問題，所檢索出文件中相關文件的數目，$Doc(i)$ 則是檢索出來的第 i 篇相關文件的排名值。Q 是查詢問題的總數。AP_j 是第 j

個查詢問題的平均精確率。MAP 表示所有查詢問題的平均精確率的平均。

• Precision: At N docs：表示在檢索出 N 篇文件時的精確率。

對所有查詢問題的前 N 篇文件精確率計算如下式：

$$P@N = \frac{1}{Q} \times \sum_{j=1}^{Q} p_j \tag{3}$$

p_j 表示第 j 個查詢問題的前 N 篇文件精確率。

2.2 檢索系統架構

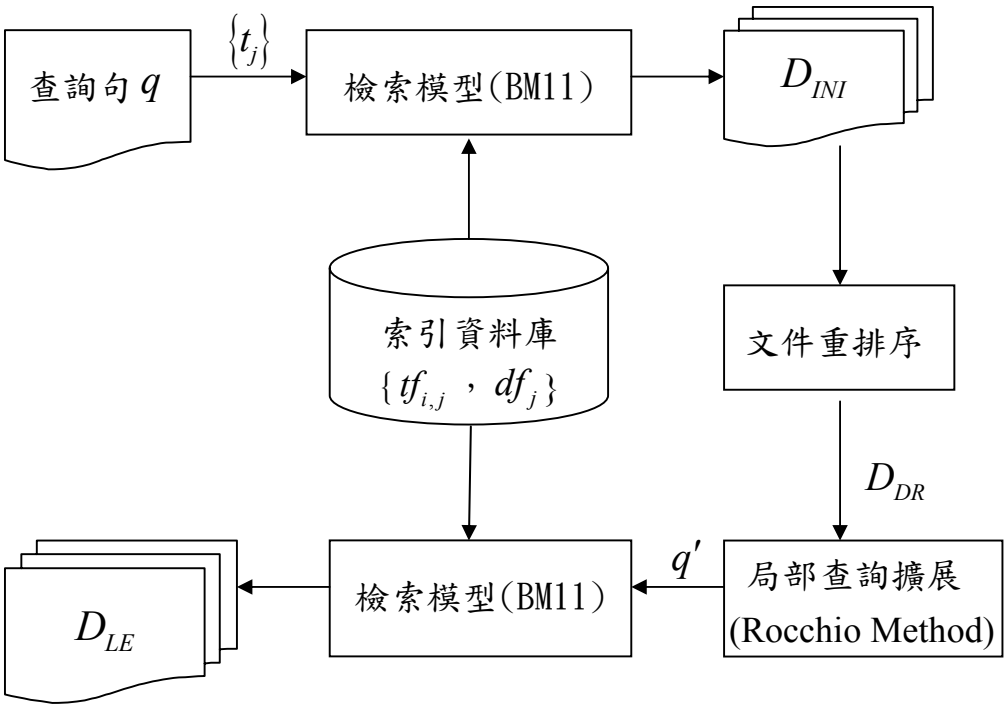


圖 2 檢索系統架構圖

本論文的檢索系統架構如圖 2 所示。由於中文的詞與詞之間並沒有明顯的間隔，而檢索系統的演算法是以「詞」爲基礎單位，因此所有檢索文件或查詢句都必須要先經過斷詞處理。本實驗中的斷詞工具是採用中研院的線上斷詞系統[8]。我們把新聞文件庫中的每份新聞文件進行斷詞後，只保留表 1 中所示特定詞類的詞，並計算詞頻($tf_{i,j}$)和文件頻(df_j)，存入索引資料庫中。

檢索時，查詢句q經斷詞後可以得到查詢詞集($\{t_j\}$)，再和索引資料庫中各篇文件以 BM11 模型進行相似度比對，可初步篩選出和查詢句最相關的前 N 篇文

件(D_{INI})。接著進行文件重排序，重排序後的文件(D_{DR})挑選前 R 篇作局部查詢擴展，找出擴展詞加入原查詢句，再用擴展後的查詢句q'進行檢索，而產生最後的文件的排序(D_{LE})。

本論文所使用的檢索模型爲 Robertson and Walker 提出的 OKAPI BM11 檢索模型[2]，其計算式如下：

$$BM11(d_i, q) = \sum_{j=1}^{T} c_j \times tf'_{i,j} \times idf'_j$$

$$tf'_{i,j} \quad \frac{tf_{i,j}}{tf_{i,j} + \frac{dl_i}{\overline{dl}}}$$

$$idf'_j \quad \log(\frac{n - df_j + 0.5}{df_j + 0.5}) \tag{4}$$

c_j 是關鍵詞 j 在查詢句中出現的次數，n 是文件總數，dl_i 是文件 d_i 的長度，$\overline{dl}$ 是所有文件的平均長度。上式的 $tf'_{i,j}$ 和 idf'_j 是 BM11 模型對詞頻和反文件頻的修正。

在局部查詢擴展的方法上，本論文採用「Rocchio Blind Feedback」[3][4]。Rocchio 演算法中將關鍵詞的權重定義如下：

$$w_j = \frac{1}{R} \times \sum_{i=1}^{R} tf'_{i,j} - \beta \times \frac{1}{S} \times \sum_{i=1}^{S} tf'_{i,j} \tag{5}$$

R 指的是排名前 R 篇的文件，S 則是其餘的文件（$S = n - R$，n 代表總文件數)，β 爲可調整的參數。如果第 j 個關鍵詞在前 R 篇文件中出現頻率高且在其餘文件中出現頻率低，此時 w_j 高，表示這個關鍵詞對於檢索出前 R 篇文件具有鑑別力。因此，權重 w_j 可視爲關鍵詞鑑別力的度量，局部查詢擴展即可用它來挑選最具鑑別力的詞作爲擴展詞。

普通名詞	Na
專有名詞	Nb

地方詞	Nc
時間詞	Nd
外文標記	FW
動作不及物動詞	VA
動作類及物動詞	VB
動作及物動詞	VC
動作接地方賓語動詞	VCL
動作句賓動詞	VE
狀態不及物動詞	VH
狀態使動動詞	VHC
狀態類及物動詞	VI
狀態及物動詞	VJ

表 1 斷詞後保留的詞類標記

3. 文件重排序

本論文共研究了三種文件重排序的方法，分別是：

1. 概念查詢法。
2. 文件分群法。
3. 局部鏈結法。

下面章節將詳細介紹此三種方法。

3.1 概念查詢法

概念查詢的方法 Qiu[5]提出，最初是應用在全域查詢擴展上。其想法是：不應該將每個查詢關鍵詞個別獨立地看待，而應該把整個查詢句當成整體的查詢概念。如圖 3 的這個例子所示，如果用查詢關鍵詞「黑澤明」和「電影」個別地去擴展，所擴展出的詞很有可能會偏離整個查詢的主題。但是，若把查詢句中的所有關鍵詞合併成爲一個查詢概念，所擴展出的關鍵詞就會比較接近查詢主題。

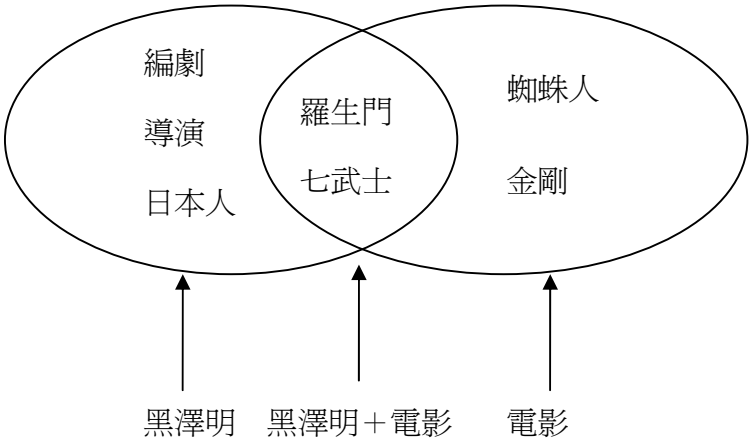


圖 3 概念查詢擴展示意圖

我們應用概念查詢的演算法，篩選出與查詢概念關聯度較高的一些關鍵詞來建立動態的關聯詞典，並利用此關聯詞典來進行文件重排序。步驟如下：

1. 對初步篩選的前 N 篇文件中所有的詞，建立詞向量 $\overrightarrow{t_j}$

$$\overrightarrow{t_j} = \overrightarrow{d_j} \Big/ \left|\overrightarrow{d_j}\right|$$

$$\overrightarrow{d_j} = (d_{1,j}, d_{2,j}, ..., d_{N,j})^T$$

$$d_{i,j} = (0.5 + 0.5\frac{tf_{i,j}}{\max_i(tf_{i,j})}) \times \log(\frac{m}{|d_i|}) \qquad (6)$$

N 表示重排序文件數，$tf_{i,j}$ 表示詞 j 在文件 i 的詞頻，$\max_i(tf_{i,j})$ 表示詞 j 在此 N 篇文件中出現過的最大次數，m 表示在此 N 篇文件中的詞的總數，$|d_i|$ 表示在文件 i 中相異詞的數目(並非詞的總個數)。

2. 建立「概念查詢向量」(Concept Query Vector)。

$$\overrightarrow{q_c} = \sum_{t_j \in q} q_j \cdot \overrightarrow{t_j} \qquad (7)$$

q 爲查詢句，$\vec{t}_j$ 爲公式(6)所定義的詞向量，q_j 爲關鍵詞 t_j 的權重，在本實驗中以詞頻作爲權重。

3. 計算「概念查詢向量」與各關鍵詞向量的關聯度。

$$r(\overrightarrow{q_c}, \overrightarrow{t_j}) = \frac{\overrightarrow{q_c}^T \cdot \overrightarrow{t_j}}{\sum_{t_j \in q} q_j} \qquad (8)$$

$\vec{t}_j$ 表示公式(6)所定義的詞向量，q_j 爲關鍵詞 t_j 的權重，$\vec{q}_c$ 爲公式(7)所定義的概念式查詢向量。

4. 選出與查詢概念關聯度較高的一些關鍵詞，建立動態關聯詞典 C。關聯詞典中必須記錄每個關聯詞與查詢概念的關聯度 $r(\overrightarrow{q_c}, \overrightarrow{t_j})$。

有了關聯詞典後，我們便可根據下式修正排序分數。

$$s_i' = \alpha \times s_i + (1-\alpha) \times \sum_{t_j \in (C \cap d_i)} r(\overrightarrow{q_c}, \overrightarrow{t_j}) \qquad (9)$$

s_i 表示文件 d_i 在第一次檢索得到的分數，s_i' 表示更新後的文件 d_i

的分數， C 代表關聯詞典，$C \cap d_i$ 代表同時出現在關聯詞典 C 與文件 d_i 的關鍵詞。α 代表原始排序分數 s_i 在重排序分數 s_i' 中所佔的比率，$0 \le \alpha \le 1$。

3.2 文件分群法

本方法的概念是希望藉由分群的分法把內容相似的文件分在同一群集中，再利用群集與查詢句的關聯度來修正排序分數[20]。在圖 4 的查詢範例中，我們看到文件 A 因爲包含了“漢代、文物、大展”這三個查詢關鍵詞，所以會得到較高的檢索分數。而文件 B 雖然看起來是相關的文件，卻因沒有包含查詢句中的關鍵詞，以致於檢索分數爲 0。然而，文件 A 和文件 B 其實相當地類似，也包含一些相同的關鍵詞(例如：東漢、馬王堆)。利用文件分群的方法，文件 A 和文件 B 因爲內容相似，可能會被分在同一群集中。如果此群集和查詢句的關聯度高，文件 B 就可以靠著所屬群集的關聯度分數而提升其排名。

本實驗的分群演算法是採用 K-means 分群法[6]，在大量高維度的資料中，找出具有代表性的 K 個資料點，稱爲群中心(centroids)。之後每份文件就可以計算其所屬群集和查詢句的關聯度(群中心和查詢向量的 cosine 夾角)，用來修正排序分數。

實驗演算法如下：

1. 建立查詢句向量 $\vec{q}$ 與每份文件的文件向量。

2. 利用 K-means 演算法，計算出各群集的群中心向量 $\overrightarrow{c_k}$。K-means 演算法步驟：
   a. 隨機選取 K 個文件向量，這 K 個文件向量就爲初始的群中心向量。
   b. 對每一文件向量計算其與 K 個群中心向量的距離，找出距離最接近的群中心，並分群進此群集中。
   c. 全部文件分群完畢後重新計算各群集的群中心向量。
   d. 重複 b、c 步驟直到所有群集內的資料皆不再變動爲止。

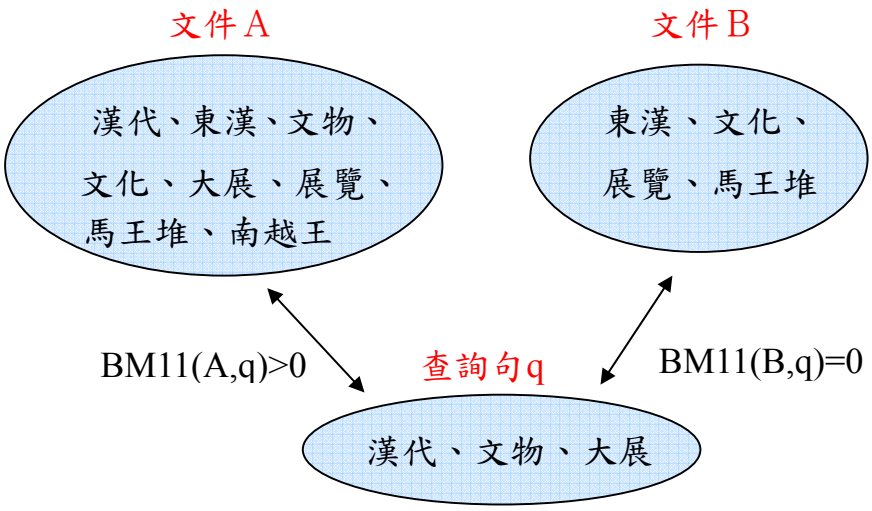


圖 4 文件檢索分數示意圖(BM11)

3. 計算查詢句與群集 k 的關聯度。

$$\cos(\vec{q},\overline{c_k}) = \frac{\vec{q}\cdot\overline{c_k}}{|\vec{q}|\times|\overline{c_k}|} \tag{10}$$

$\overline{c_k}$ 表示第 k 群的群中心的向量。特別注意的是公式(10)中，屬於同一群集的文件，並不個別計算其與查詢句的關聯度，而是共用所屬群集的關聯度。

4. 每篇文件重排序的分數則可以用原排序分數和群集關聯度兩者加權得到，如下公式：

$$s_i' = \alpha \times s_i + (1-\alpha)\times\cos(\vec{q},\overline{c_k}) \tag{11}$$

s_i' 、s_i 、α 之定義和公式(9)中相同。文件 i 屬於群集 k 。

3.3 局部鏈結法

在 BM11 的檢索模型中，分析查詢句與各篇文件的關聯時，僅是計算個別查詢關鍵詞對於文件的權重並加總起來。但是，各個關鍵詞之間可能有語意上的關聯或限制，若分別去計算權重，可能導致只包含部份查詢概念的文件其關聯度反而會比包含全部查詢概念的文件來得高。例如：我們欲查詢「漢代文物大展」的相關文件時，可能會有「汽車大展」的文件因「大展」這個詞出現很多次而排名較前面之情形，造成檢索精確率的下降。爲了修正這種不合理的現象，我們利用

查詢詞之間「局部鏈結」的統計來修正排序分數。局部鏈結的統計以查詢句中相鄰的兩個關鍵詞形成的「關鍵詞組」(keyword pair)做爲統計的單位，例如，查詢句“漢代、文物、大展”就可被分爲“漢代、文物”和“文物、大展”兩個詞組。每一詞組中的兩個詞(例如“漢代”與“文物”)如果在同一篇文件中出現的位置足夠接近，稱之爲一個局部鏈結(Local Link)。這兩個詞在文件中的距離以區間(frame)來限制，例如，區間數設爲 50 就表示同一組內的兩個詞距離必須小於 50 才算是一個局部鏈結。

實驗演算法爲：

1. 對查詢句 q 中每個關鍵詞組(t_j, t_k)，統計其在文件 i 中的局部鏈結數${}_iL_{j,k}$以及局部鏈結文件頻$df_{j,k}$。局部鏈結數${}_iL_{j,k}$的計算方式爲，若(t_j, t_k)出現在文件 i 中的區間一次，則局部鏈結數加 1。局部鏈結文件頻$df_{j,k}$則是計算(t_j, t_k)之局部鏈結共在多少文件中出現過。

2. 對原始文件分數作加權。

$$s_i' = \alpha \times s_i + (1-\alpha) \times \left(\sum_{(t_j,t_k)\in q} {}_iL_{j,k} \times idf_{j,k} \right)$$

$$idf_{j,k} = \log \frac{n}{df_{j,k}} \qquad \text{(公式 12)}$$

(t_j, t_k)表示查詢句中某一關鍵詞組。n爲總文件數。

s_i'、s_i、α之定義和公式(9)中相同。

3.4 文件重排序方法之結合

本論文的三種文件重排序方法運用了不同的概念：概念查詢法是先用查詢概念產生出關聯詞(非原查詢詞)，並計算其關聯度以改進文件的排序；文件分群法是利用「群集的關聯度」來改進文件的排序；局部鏈結法則是利用「查詢詞局部共現」特性，讓查詢語意更準確的文件可以提升排名。前兩種方法重在查詢概念的「擴展」，將一些相關但可能不包含查詢詞的文件有機會往前排，其差異是擴展的方式不同。第三種方法則重在查詢語意的「準確」，所影響的文件是那些含有查詢詞的文件。由於這些方法的設計理念和作用的範圍各有不同，我們想進一步探討這些方法是否能彼此互補，並適當地結合。我們將重排序方法結合的方式如下：

$$R_i = \beta \times R_{i,1} + (1-\beta) \times R_{i,2}$$

$$R_i' = \gamma \times R_i + (1-\gamma) \times R_{i,3}$$

$$s_i' = \alpha \times s_i + (1-\alpha) \times R_i' \quad (13)$$

$R_{i,1}$ 、 $R_{i,2}$ 、 $R_{i,3}$ 爲文件 d_i 在任兩種文件重排序方法所得到的修正部份分數(公式 9、11、12)。 β 爲權重參數。

3.5 實驗結果

首先對於檢索系統進行實驗，發現在局部查詢擴展中取前 10 篇文件當作是相關文件(R=10)，並取權重排名前 80 個詞當作擴展詞加入查詢句(E=80)，可達到最佳的檢索效能，故本節實驗數據皆以此設定。實驗之系統架構圖參考圖 2。

表 2 文件重排序方法間之比較(無局部查詢擴展)

		MAP	P@10	P@100
Baseline(BM11 only)		0.2429	0.4476	0.1907
加入文件重排序	概念查詢法	0.2426	0.4643	0.1950
	文件分群法	0.2600	0.4643	**0.2195**
	局部鏈結法	0.2520	0.4690	0.2040
	三種結合	**0.2723**	**0.4786**	0.2150

表 3 文件重排序方法間之比較(有局部查詢擴展)

		MAP	P@10	P@100
BM11 + Rocchio		0.3727	0.4929	0.2767
加入文件重排序	概念查詢法	0.3821	0.5262	**0.2838**
	文件分群法	0.3831	0.5238	0.2795
	局部鏈結法	0.3855	0.5405	0.2757
	三種結合	**0.3956**	**0.5595**	0.2767

表 2 顯示了文件重排序方法間之比較。對 Baseline 而言，應用文件分群的重排序方法有最佳的平均精確率。而應用局部鏈結的重排序方法，則有最佳的前 10 篇文件的精確率。應用概念式查詢的重排序法和另外兩種方法比較起來則相對比較差。表 3 結合局部查詢擴展後，三種文件重排序法的平均精確率，與 Baseline 比較起來，都有不錯的提昇效果，而提升的效能則以局部鏈結的重排序法最佳。對於前 10 篇文件的精確率來說，局部鏈結的重排序法，有相當明顯的提昇效果，爲三種文件重排序方法中最佳的。

且由表 2 得知，文件重排序方法間的結合確實能夠有效的提昇檢索平均精確率和前 10 篇文件的精確率，且結合的方法確實比單用任一種文件重排序的方法

達到的效果要來的好。表 3 也顯示出結合局部查詢擴展後更能大幅提昇檢索的精確率，且全部文件重排序方法的結合所展現的檢索效能比單用任一種文件重排序法或是任兩種文件重排序方法的結合都要來的好，爲所有方法中最佳的，也證明了文件重排序方法之間確實具有互補的效果。

4. 擴展詞過濾方法

4.1 擴展詞的過濾

Rocchio Method 的擴展詞權重是計算對於前 R 篇文件的「鑑別力」，並未考量這些詞是否真的與查詢主題相關，可能因而擴展出了與查詢主題不相關的詞。因此，我們進一步地研究擴展詞的過濾方式，期望藉由過濾法來篩選掉與查詢主題不相關的擴展詞，以提升檢索效能。我們定義了擴展詞和查詢句的相關度，過濾掉相關度較低的擴展詞，以降低查詢偏移(Query Drift)的可能性，提高檢索的效能。其演算法如下：

1. 計算擴展詞 e_j 與查詢關鍵詞 q_k 的相關係數(Correlation Coefficient)。

$$r_{e_j,q_k} = \frac{P_{e_j,q_k} - P_{e_j} \times P_{q_k}}{\sqrt{P_{e_j} \times (1-P_{e_j})} \times \sqrt{P_{q_k} \times (1-P_{q_k})}}$$

$$P_{e_j,q_k} = df_{e_j,q_k} \Big/ n$$

$$P_{e_j} = df_{e_j} \Big/ n$$

$$P_{q_k} = df_{q_k} \Big/ n \qquad (14)$$

P_{e_j,q_k} 表示擴展詞 e_j 與查詢關鍵詞 q_k 共同出現的機率。P_{e_j} 、P_{q_k} 表示擴展詞 e_j 、查詢關鍵詞 q_k 在所有文件中出現的機率。df_{e_j,q_k} 表示擴展詞 e_j 與查詢關鍵詞 q_k 共同出現的文件數。df_{e_j} 表示擴展詞 e_j 的文件頻。df_{q_k} 表示查詢關鍵詞 q_k 的文件頻。n 表示總文件數。

2. 計算擴展詞 e_j 對查詢句的相關度。

$$r_{e_j} = \sum_{q_k \in q} r_{e_j,q_k} \qquad (15)$$

3. 設立一門檻值來過濾掉與查詢句相關度(r_{e_j})較低的擴展詞。

4.2 實驗結果

表6 爲使用<DESC>查詢內容作過濾對於平均精確率(MAP)及前10 篇文件精確率(P@10)的影響。No-Filter 表示未將擴展詞作過濾的動作。Filter 表示有將擴展詞過濾。由表中數據可知，用<DESC>的查詢句來過濾，所得到的MAP、P@10 和未作過濾時差不多。造成這種差別是因<DESC>中的查詢句較長，且包含一些與查詢主題不相關的詞。

表6 DESC 查詢內容作過濾對於MAP 及P@10 的影響

		No-Filter		Filter	
		MAP	P@10	MAP	P@10
BM11 + Rocchio		0.3727	0.4929	0.3727	0.5024
加入文件重排序	概念查詢法	0.3821	0.5262	0.3820	0.5238
	文件分群法	0.3831	0.5238	0.3830	0.5214
	局部鏈結法	0.3855	0.5405	0.3846	0.5429
	三種結合	0.3956	0.5595	0.3951	0.5619

表7 爲使用<TITLE>查詢內容作過濾對於平均精確率及前10 篇文件精確率的影響。與DESC 的結果作個對照，我們發現雖然作擴展詞過濾後，並不能提升平均精確率，但是對於前10 篇文件的精確率卻可以有效地提升，且以局部鏈結重排序法及三種重排序法間的結何提升幅度最大。由此可知本實驗的擴展詞過濾方法在查詢句沒有贅詞或是較少贅詞時，對於檢索效能的提升確實有幫助。

表7 TITLE 查詢內容作過濾對於MAP 及P@10 的影響

		No-Filter		Filter	
		MAP	P@10	MAP	P@10
BM11 + Rocchio		0.3341	0.4762	0.3350	0.4857
加入文件重排序	概念查詢法	0.3377	0.4595	0.3380	0.4667
	文件分群法	0.3366	0.4881	0.3377	0.4929
	局部鏈結法	0.3381	0.4833	0.3399	0.5048
	三種結合	0.3425	0.4905	0.3449	0.5094

5. 結論與未來研究方向

5.1 結論

本論文主要研究三種文件重排序的方法，希望藉由重排序的演算法將相關文件往前排序，以改良局部查詢擴展的缺點，進一步地提升檢索效能，而實驗結果也證明此三種重排序法確實對於檢索效能有所貢獻。在 3.4 節中，我們加深探討這三種重排序法之間是否具有互補的特性，將重排序法之間作結合，以期望能更進一步地提升檢索效能，而實驗結果也證明重排序法間確實互補，能有效提升檢索效能，將平均精確率從 0.3727 提升至 0.3956，前 10 篇文件精確率從 0.4929 提升至 0.5595。

在第四章中，我們對於擴展詞的選取方式作更加深入的探討。對擴展詞作過濾的方法，實驗結果顯示其確實對於檢索效能有所提升。

5.2 未來研究方向

本論文的局部鏈結重排序法僅探討查詢句中相鄰的兩個查詢詞之間的語意關係，將包含較多語意鏈結的文件往前排序，但此方法容易將沒有語意關係的兩個關鍵詞作配對，降低重排序的效能。未來實驗可利用句法剖析(Parsing)，更精確地抽取出查詢句語意，以預期能達到更佳的效能。

而本論文結合的方式是利用人工的方式做參數調整，以達到最佳化，若能訂立一套自動的參數估測方式，對於檢索系統的改進將會更有幫助。

參考文獻

[1] 陳光華(2004). “資訊檢索的績效評估”. 2004年現代資訊組織與檢索研討會.

[2] S. E. Roberson & S. Walker. “Some simple effective approximations to the 2-Possion model for probabilistic weighted retrieval”. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994.

[3] Rocchio, J.J. Jr.(1971).“Relevance feedback in informationRetrieval”. In the Smart system – experiments in automaticdocument processing, 313-323. Englewood Cliffs, NJ : Prentice Hall Inc.

[4] Gerard Salton and Chris Buckley. “Improving retrieval performance by relevance feedback.” Journal of the American Society for Information Science. 1990.

[5] Qiu,Y. & Frei, H. P. “Concept based query expansion”. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993, pp.160-169.

[6] J. B. MacQueen (1967). “Some methods for classification and analysis of

multivariate observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probabilit", Berkeley, University of California Press.

[7] L.P. Yang, D.H. Ji(2005). "Chinese information retrieval based on terms and relevant terms." ACM Transactions on Asian Language Information Processing. Vol.4,Issue 3(2005). pp.357-374

[8] http://ckipsvr.iis.sinica.edu.tw/

[9] Xu J., Croft W.B., "Query expansion using local and global document analysis." Proceeding of the 19th annual international ACM SIGIR conference on research and development in information retrieval, 1996, pp.4-11.

[10] Yuen-Hsien Tseng, Da-Wei Juang and, Shiu-Han Chen. "Global and Local Expansion Term Expansion for Text Retrieval." Proceedings of the Fourth NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, June 2-4,2004,Tokyo,Japan.

[11] Yuen-Hsien Tseng, Yu-Chin Tsai, and Chi-Jen Lin."Comparison of Global Term Expansion Methods for Text Retrieval." Proceedings of NTCIR-5 Workshop Meeting, Deceber 6-9,2005,Tokyo,Japan.

[12] K.S. Lee, Y.C. Park, and K.S Choi. "Document Re-ranking Model Using Clusters." Information Processing & Management, v37 n1 p1-14 Jan 2001.

[13] Harman, D. (1992 June). "Relevance feedback revisited." Paper presented at the Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, New York.

[14] ZHANG, M., SONG, R., LIN, C., MA, S., JIANG, Z., LIU, Y., et al. (2002). "Expansion-based technologies in finding relevant and new information." Paper presented at the TERC.

[15] Xu J., Croft, W.B. "Improving the Effectiveness of Information Retrieval with Local Context Analysis." ACM Transactions on Information Systems, 2000.

[16] M. Mitra., A. Singhal. And C. Buckley. "Improving Automatic Query Expansion." In Proc. ACM SIGIR'98.

[17] Qu, Y.L., Xu, G.W., Wang J.2000. "Rerank Method Based on Individual Thesaurus." Proceedings of NTCIR2 Workshop.

[18] Kamps, J. 2004. "Improving Retrieval Effectiveness by Reranking Documents Based on Controlled Vocabulary." The 21th European Conference on Information Retrieval.

[19] Yang Lingpeng, Ji Donghong, TangLi. 2004. "Document Re-ranking Based on Automatically Acquired Key Terms in Chinese Information Retrieval." In Proceedings of the COLING'2004, pp. 480-486

[20] Luk, R.W.P., Wong, K.F.2004. "Pseudo-Relevance Feedback and Title Re-ranking for Chinese IR." In Proceedings of NTCIR4 Workshop.

針對數學與科學教育領域之電腦輔助英中試題翻譯系統

呂明欣
國立政治大學資訊科學系
94753007@nccu.edu.tw

高照明
國立台灣大學外國語言學系
zmgao@ntu.edu.tw

劉昭麟
國立政治大學資訊科學系
chaolin@nccu.edu.tw

張俊彥
國立台灣師範大學地球科學系
changcy@ntnu.edu.tw

摘要

由國際教育學習成就調查委員會(IEA)統一命題之國際數學與科學教育成就趨勢調查測驗(TIMSS)，爲便於台灣中小學生施測與理解，英文原文試題內容需要經過許多人工討論及翻譯時間。爲了增進翻譯內容一致性及其效率，我們設計一套符合測驗試題的輔助翻譯系統，將不同格式的試題文件，經執行語法分析式的片語擷取和字典查詢，透過使用者介面，選擇合適的片語詞彙翻譯選項和詞序調整，以及提供目前常用之線上翻譯服務、回顧翻譯類似句、以及加減詞彙等功能。爲了能提昇翻譯詞彙的選擇正確性，我們記錄翻譯者選詞動作，讓翻譯者能回顧過去曾處理過的翻譯類似句，並且按照系統提供之選詞頻率資訊、科學領域的期刊語料之詞頻統計，以及利用統計式中英詞彙對列和語言模型，更改選詞的優先順序。我們嘗試以過去試題爲實驗對象，按年級及學科區分 6 大試題類別，搭配 4 種選詞策略，透過 BLEU 及 NIST 之翻譯評估指標比較線上翻譯系統和本系統，實驗結果顯示在各實驗組的評估上均有優於線上翻譯系統的效果。

關鍵詞：自然語言處理，電腦輔助教學，受限語言，試題翻譯，機器翻譯，TIMSS

一、緒論

TIMSS(Trends in International Mathematics and Science Study) [16]，由國科會委託國立台灣師範大學科學教育中心（以下稱師大科教中心）負責，針對國小四年級及國中二年級學生執行數學與科學領域之試題翻譯及測驗工作。主要工作流程包含：從 IEA 取得英文試題內容，師大科教中心決議進行翻譯工作分配、中文試題交換審稿校正及翻譯問題討論，最後將中文翻譯試題定稿。其中完成上述說明之工作流程，依照 TIMSS 官方網頁歷年統計至少需要一個月以上的準備時間。主要原因在於試題翻譯的原則上須避免在翻譯過程中由於個人主觀的因素（如：爲具體表達題意，加入不適當詞彙，或者是同一題目中，主要詞彙因在文章前後翻譯方式不一致而產生的語意混淆）或不同的翻譯者因不同的翻譯標準使得翻譯內容有所出入，間接影響試題在題意表達上的品質。故爲了能夠完整表述且不影響題意的正確性，過去投入了不少的人力資源及時間成本。因此，若有一個好的翻譯輔助系統[3]協助翻譯工作的話，除了可以提供在字詞翻譯上一致性的標準以外，利用電腦及機器翻譯的技術取代人工檢查字詞翻譯亦有可能提升翻譯的準確及效率。

另外，TIMSS 試題在內容上，依題型種類區分爲選擇題及問答題，句型結構多半以直述句和問句構成問題的題幹，加上誘答選項，因此語法結構和主題較一般翻譯文句明確簡短，亦不需考慮太過複雜的修辭方式。

有鑒於試題內容較一般文章結構精簡的特性，在機器翻譯的領域中有所謂子語言(sublanguage)的翻譯類型。即翻譯需求限制於某一特定領域上，減少翻譯過程所發生的

歧義、修飾文法及實作的複雜度。現今的研究多採用受限語言(controlled-language)的方式，限定專業領域的語彙和語法結構的範圍，減少詞彙歧義[4]和複雜性，以及加入人機互動式的機器翻譯模式。因此，我們在系統設計重心上，以尋找專有名詞，將試題內容經由斷詞及文法分段，利用查詢詞典，輔以統計式翻譯模型[1]找出所有可能適合原文詞彙的翻譯結果，並提供使用者介面，以人工編輯或半自動的方式，透過中文句型與英文句型的對應找到合理的詞序關係。

二、TIMSS 試題翻譯系統

基於緒論的中英試題翻譯之問題定義及系統設計重心，我們完成的試題翻譯系統，在執行的流程主要區分爲 3 部分。

i. 將 TIMSS 英文試題透過文件轉檔程式轉成文件檔
ii. 從系統翻譯模組選擇線上翻譯模式，或字典翻譯模式，配合修正翻譯模組產生中文翻譯結果
iii. 將翻譯結果輸出至使用者介面

首先，試題資料的檔案格式可能爲純文字檔、Microsoft Word 的 DOC 檔或 Adobe Systems 制定的 PDF 檔等非純文字檔格式，故要能正確讀取文字內容之前必須先進行轉檔動作。我們呼叫 JACOB PROJECT[13]的應用程式套件進行純文字檔轉檔，目前，利用轉檔的缺點在於當 PDF 檔的內容爲中文內容，且原先文書檔之存檔方式是以特殊字型的方式存檔（如利用 LaTeX 文書編輯工具定義之字型），這時中文檔案會由於編碼還原問題產生亂碼的情形，而英文內容之檔案則無此問題。接著，存入暫存檔之後再讀取其文字內容，依題號、試題內容及誘答選項，並輸出題目選單於題目選擇及新增修改面板供翻譯者選擇欲翻譯之試題內容。

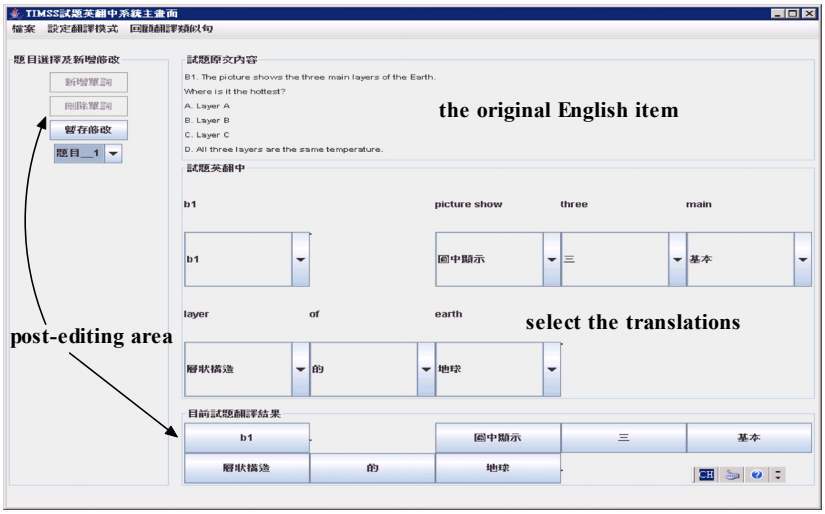
系統將 TIMSS 試題檔轉換成純文字檔後，接著視翻譯者選擇可切換**線上翻譯模式**及**字典翻譯模式**。

i. **線上翻譯模式**：我們考慮到目前谷歌所推出的 Google Translate[12]以及雅虎線上翻譯系統[19]，實作 Http Connection Interface ，輸入翻譯句子後傳回譯文的 html 文字檔，經過字串剖析，將 html 語法去除，逐行檢查並判斷該行是否存在具有翻譯結果字串之左右相鄰識別文字，若符合則取出左右相鄰識別文字中間之翻譯結果字串，再透過介面顯示如圖一之翻譯訊息，系統會經翻譯者確認，自動暫存英文原文和中文翻譯的結果，待最後儲存動作完成後存入檔案中。

圖一・線上翻譯模式介面

ii. **字典翻譯模式**：翻譯者按題號選擇所需翻譯內容後，系統會讀取該題目，在取得欲翻譯之英文句子後，會先行判斷目前系統翻譯模組的選擇，若翻譯者選擇字典翻譯模式，則會經由字典查詢單字的方式，將所有單字和詞組的所有可能的中文翻譯，輸出至使用者介面，讓使用者進行後續之調整翻譯詞序和中文翻譯選詞

等動作，如圖二所示。



圖二．字典翻譯模式介面

然而，翻譯詞序和中文翻譯選詞的動作，就翻譯者本身來說仍然需要許多人工檢查的時間，因此，爲增進選詞和調整詞序的效率，我們需要足夠多的中英試題翻譯資訊，利用機器學習的技術，找出中英試題在翻譯和文法結構的關連性。然而 IEA 每隔四年舉行一次 TIMSS 測驗，語料來源上，我們目前僅有 1999 年及 2003 年所舉辦的 TIMSS 中英對照測驗題目（在之後章節，我們分別以 TIMSS1999 及 TIMSS2003 來表示 1999 年及 2003 年所出題之測驗題目），在試題資料部分並不足以能利用機率統計的方式，依照語境準確選出適當表達語意的中文詞組，並還原中文翻譯的詞序。因此，需要透過其他外部資訊或中英語料庫的輔助，做爲替代的方案。在第三章和第四章，我們將會介紹如何做到翻譯的一致性，詳細說明中文詞組翻譯及選詞順序調整的策略。

三、翻譯的一致性

一篇好的中文翻譯內容，能掌握翻譯前後間內容的一致性，即完整陳述原文的意思，在語彙上使用正確詞義表達，表達的文法形式也要能和原文一致，並且流暢易讀。以 TIMSS 試題來說，要能正確翻譯出試題陳述部分，包含內容出現的問句片語、專有名詞，和提示考生在做題目時常用詞彙，以適切引導考生閱讀試題內容。此外，問句主體在翻譯內容前後亦不可出現不同的翻譯結果。

爲避免不同翻譯者會有不同的翻譯習慣，翻譯者經常需要檢討需要統一翻譯的詞彙。以提示考生的常用詞彙來說，如 check one box 會統一翻譯成“勾選一項”、as shown below 統一翻譯成“如下所示”，故針對上述現象，系統亦需提供翻譯者可隨時動態調整其中英對照翻譯規則之彈性。

另外，傳統市售的翻譯系統，或線上翻譯系統所提供之翻譯結果通常爲語料統計中最有可能的翻譯結果，但對於特定領域或一詞多義的原文內容，不能完全保證其翻譯結果之正確性，翻譯者也無法直接針對特定領域知識之原文詞彙，任意修正其翻譯結果，讓系統經過修正後，選擇最好的解釋。故系統在使用者介面的設計上，要能提供翻譯者選擇符合語境的翻譯選項及合理的詞序調整方式，並且在接受翻譯者調整翻譯後，能夠根據先前的調整，自動修正類似句之翻譯詞彙。

本系統在翻譯過程中針對需一致性翻譯的詞彙或詞組，定義以下檔案，讓翻譯者能自行定義詞組的中英翻譯規則。

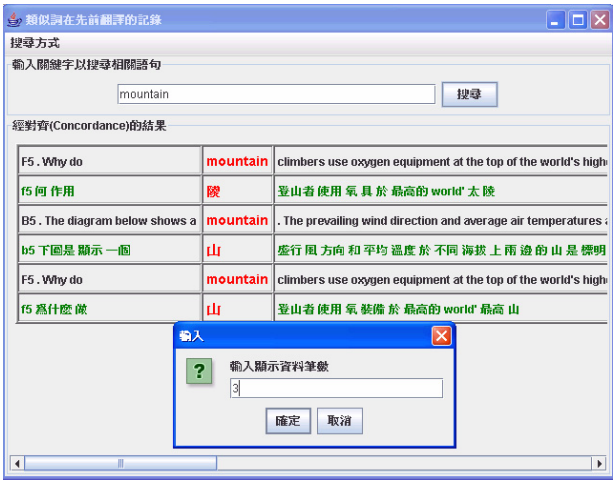
i. **規則詞典檔**：將試題翻譯小組所決議之統一翻譯詞組以人工的方式建立中英對照詞典檔。

ii. **片語句型檔**：爲常用的名詞片語、動詞片語及形容詞片語等詞組的中英翻譯對照檔。和規則詞典檔不同的是，規則詞典檔所記錄的詞組是連續的詞組，如 in order to（爲了）或 build up（建立，構成），而片語句型檔所記錄的詞組則爲非連續的詞組，如 place~ on~（散佈某物於某處）或 carry~away（載走某物）。

另外，爲了減少討論單字翻譯的正確性及適切程度的時間，讓翻譯者檢索過去已有的翻譯內容，我們提供回顧翻譯似句之功能，將檢索的翻譯結果以中英文字對齊（concordance）的方式呈現於使用者介面。以下爲主要之檢索執行流程。

i. **從中英對照詞組檔建立倒轉檔（inverted file）**：當翻譯者利用檢索視窗輸入英文檢索單字或連續性片語，系統會從過去的倒轉檔內容，找尋所有和檢索單字或片語有關的中英對照詞組檔之存放位置，再逐一搜尋檔案之內容，檢查檔案內所有中英對照結構，其英文單字片語是否和檢索詞相同，若相同，則將目前比對位置的所有中英對照結構、英文原文句子及中文翻譯結果儲存於一緩衝區（buffer）。

ii. **產生中英文字對齊（concordance）結果**：從緩衝區儲存的中英對照結構，找出檢索詞的中英的對照結構，接著把英文原文句子及中文翻譯結果，以檢索詞的中英的對照結構爲中心，將英文原文句子及中文翻譯結果切成左右兩部分，最後根據翻譯者所輸入的資料筆數，檢查緩衝區符合檢索詞的句子數目，以及檢查所有檢索詞中文翻譯，將英文原文句子及中文翻譯結果，依照相同中文翻譯句子最多的順序爲前的排序方式，輸出至結果視窗，如圖三所示。



圖三 中英文字對齊介面

四、中文詞組翻譯及選詞順序調整

除了利用自行定義詞組的中英翻譯規則，以及回顧翻譯句的方式，掌握部分詞組的中文翻譯之外，我們還定義一套詞組翻譯及翻譯修正流程，以下爲詳細步驟。

i. 詞性標記及 stemming

我們呼叫 MXPOST[15]詞性標記工具將原文加入標記，並根據 Porter 演算法[17]，實作還原各詞性（名詞、動詞、形容詞、副詞）的詞幹變化（如~ly、~ing、~ed 等）。對於某些特殊詞幹變化的單字（如動詞 break，其過去式爲 broke，被動式爲 broken，以及名詞單複數型態）比較難以用演算法處理的部分，我們利用 WordNet[25]，依照詞性

做字典檔搜尋找到 stemming 後的結果。

ii. 斷詞處理及詞組翻譯

英文句子部分包含單字（word）、詞組（term）及標點。其中，詞組即爲數個單字的組合、片語（phrase，如名詞片語、動詞片語、形容詞片語）或複合名詞。

在單字和詞組的語彙表達程度上，詞組往往能掌握比較完整的動作表達或描述。若單純只用一般字和字之間的空白做斷詞，便無法利用斷出來的單字決定詞組的意思。

因此，我們希望在翻譯上能做到英文單字的字意能完整對應中文的單字字意，以及做到詞組式的翻譯，使得文句的語意能捕捉得更完整。而第一步要先有辦法斷出正確的字詞，接下來才能找出正確的中文詞組翻譯。

在斷詞方法上，首先將英文句子，利用字和字之間的空白做基礎斷字的動作。但這時要先考慮標點符號的情形，因爲標點符號在英文文句中通常是連在單字右方，因此先檢查文句有標點符號時再插入空白字元，再利用空白斷詞。接下來進入詞組斷詞部分，我們分別利用英文剖析器 MINIPAR[14]及片語句型檔斷詞方法，將英文句子內片語或詞組找到，再用字典找出其片語翻譯，其中字典部分我們使用牛津現代英漢雙解詞典[29]（Concise Oxford English Dictionary，收錄 39429 個詞彙）。在字典查詢完兩種尋找片語或詞組方法之翻譯後，接著進入長詞詞組篩選階段，首先，排除重覆找到的詞組翻譯，再來檢查若片語句型檔分詞或 MINIPAR 分詞之中出現片語結構類似，片語長度不同的情況，則選擇較長的片語翻譯結果，最後輸出完成片語翻譯之句子。

MINIPAR 分詞上，MINIPAR 以解析其語法解析結構，擷取最常出現之名詞片語（即 NP 或 NPL）及動詞片語（VP 或 VP+PP）及形容詞片語（ADJP）。

片語句型檔斷詞方法上，我們以句型比對的方式，除了按照相鄰片語做搜尋外，利用“*”代表任一 word，“<>”代表左右詞彙限制範圍內所有之 word，都可以利用該演算法擷取其片語。故像要抓取 “pulse rate”、“success rate”、“～rate”等複合名詞，可利用“* rate”，代表抓取 rate 前一個 word 當做複合名詞；若需要抓取 “as ～ as possible”或“as ～ as”等句型時，可利用“as <> as possible”及“as <> as”之描述方式，來抓取這些類似片語。惟須注意的是，比對的句型若有相似結構但不同長度的字串樣式，如一英文句子爲“…the diagram shown below…”，同時滿足片語句型檔內的 “the diagram shown”和 “the diagram shown below”片語句型，則我們會選擇長度較長的“the diagram shown below”而不是選擇“the diagram shown”加上“below”做爲分詞的結果。

iii. stop word filtering

執行英文翻譯時，會有某些單字因無法單獨存在，需搭配其他單字出現，因而出現不需翻譯或無法翻譯詞彙之情形。例如冠詞 the、an、a，常加於單數名詞之前，當不須特別強調單數時，便不需要特別翻譯冠詞。介系詞 for、to、of，當出現於疑問詞之後，或片語之中，其意義表達助動詞如 do、does，爲了不讓後續翻譯動作翻譯出上述之詞彙，我們將這些詞彙建立一集合，並針對英文句子並做出以下判斷。

a) 冠詞部分，the 直接去除，a 和 an 則判斷出現位置是否爲句首，並檢查 a 後面是否有其他標點符號（如括號或句點，代表有可能是編號），若以上條件皆不符合，則予以去除。
b) 介系詞部分大致保留，惟需判斷前一字元是否爲 what、how、who、when、why 等疑問詞，若條件符合，則予以去除。另外，to 出現於句首時亦予以去除。
c) 助動詞部分，則判斷前一字元是否爲 what、how、who、when、why 等疑問詞，同介

係詞判斷方式，若符合則予以去除。

iv. 單字和專有名詞翻譯

在斷詞處理及片語翻譯階段，我們完成了片語的中文翻譯，剩下來的單字及專有名詞的部分，我們則再次利用查詢字典及修正翻譯模組的中英對照規則詞典檔的方式，找出每個單字及專有名詞所有可能的中文翻譯結果，做爲中文選詞之選項。另外，針對英文姓名轉換成中文命名部分，我們參考 20000-NAMES.COM 網站，將網頁所列的男性和女性英文姓名按性別分類，分別存成文字檔，接著將待翻譯的英文句子所切分出來的詞彙，和先前儲存的所有英文姓名加以比對，找出相同姓名後，我們利用事先建立好的中文姓名庫，按性別回傳適當名字。例如，當英文名詞出現“Bob”時，屬於男性姓名，故從中文姓名庫的男性姓名分類中，以亂數方式，任意取得 “小華”當作對應的姓名翻譯結果，並記錄目前“Bob”是和 “小華”相互對應，以便同題內容中再出現“Bob”姓名時可做到人名一致性的翻譯。

西元表示部分，我們則是判斷句中的詞彙是否包含西元年月日等詞彙，作爲自動轉換民國表示的標準。

v. 翻譯結果修正

在完成單字及片語翻譯之後，最後就是針對各單字翻譯出來的中文翻譯選項，調整適當的選項排列順序，以及詞序調整。

在第二章已說明目前 TIMSS 試題資料並不足以利用機率統計的方式，依照語境選出適當表達語意的中文詞組，並且還原中文翻譯的詞序。因此我們利用詞頻高低決定選詞重要程度的概念，記錄翻譯者過去選詞之詞頻數，並且利用中文語料庫統計詞頻，透過中英語料庫，以統計式翻譯爲基礎，建立中英詞彙對列及語言模型。以下將針對**翻譯者選詞詞頻**、**中文語料庫詞頻**及**中英詞彙對列和語言模型**詳細介紹。

（一）翻譯者選詞詞頻

中文選詞順序調整部分，我們考慮試題本身的敍述部分及問句所運用到的修詞方式不多，大部分以直接翻譯爲主，故常用的詞組其對照的中文詞組種類大致不會變動太大。在這邊我們使用漸進式的訓練學習方法，翻譯者可利用介面選擇中文詞組選項，藉由統計最常被翻譯者選擇的選項做爲日後選詞的優先順序，其他的選項則是再按次常選擇順序依此類推調整。爲了達到這個目的，我們首先將每次經翻譯者確認翻譯無誤的中英文對照詞組統一記錄至中英對照詞組檔內，如表一所示。

表一、中英對照詞組檔

英語題目：
L5. When male wolves place their scent on trees, they most likely are doing this in order to
經前處理及對照翻譯後儲存的中英對照詞組檔（以 BNF 表達儲存範例）：
<English-Chinese Form> := [<English Word>|<English Phrase> : <Chinese Word>|<Chinese Phrase>] | [<Symbol>]
When male wolves place their scent on trees ,
[when:當][male:雄][wolf:狼][place:置於][their:他們的][scent:氣味][on:上][trees:樹木][,]
they most likely are doing this in order to
[they:他們][most:最][likely:可能][be:是][do:做][this:此][in order to:爲了]

英文句子經過翻譯後，按下儲存動作，會自動將所有中英詞彙對照翻譯的結果，依照表一之格式，儲存在中英對照詞組檔內，接著，將目前儲存於中英對照詞組檔所有中英對照的結構，和修正翻譯模組的多義索引修正檔所有過去已有記錄之中英對照結構比對，若可以找到相同的結構，則將目前結構的選詞次數欄位加 1，若無相同結構，代表

過去的記錄無該中英翻譯對照之記錄，則將該結構儲存，並將選詞次數設定爲 1。完成更新多義索引修正檔次數後，在下一次翻譯新的句子時，就可以依照新的選詞次數高低，適當調整翻譯詞彙的選項順序。

除了利用多義索引修正檔調整翻譯詞彙的選項順序，針對目前翻譯句子中待調整選詞順序之單字，取出該單字所有中英翻譯選項及目前該單字前後詞彙之中英翻譯，即緊鄰詞彙，接著，從所有中英對照詞組檔中找出所有同爲目前單字中英翻譯選項的前後詞彙，比較目前翻譯句子找到的緊鄰詞彙和過去留存在中英對照詞組檔的緊鄰詞彙是否相同，若相同或包含，則以當時所對應的中文翻譯選項爲優先。如不相同，則取該字的中文翻譯中選擇次數最多的詞組爲優先。例如：我們假設在“wolves place their scent”中，place 需要調整翻譯選項，我們知道 place 可以翻譯成“置於”或 “地方”，並且確定 place 的緊鄰詞彙：wolf 翻譯成 “狼”，their 翻譯成 “他們的”。接著，從所有中英對照詞組檔找尋過去 place 的所有緊鄰詞彙中，是否存在 wolf 翻譯成 “狼”和 their 翻譯成 “他們的”記錄，假使發現存在，並且當時 place 翻譯成 “置於”我們便決定 place 的翻譯即以 “置於”爲優先。

（二）中文語料庫詞頻

爲了彌補目前資料量過少的語料來源，我們透過尋找有關數理、科學教育領域的中文語料，以補足語料庫的不足。之後，系統依此統計資料，計算中文詞組之詞頻高低，檢查英文詞彙所對應的所有中文翻譯，若該中文翻譯在語料統計資料上有相對應的詞頻高低值，則依詞頻調整其翻譯詞彙順序。

然而，英漢字典的中文翻譯詞彙，並不必然存在於語料庫之中，使得英漢字典詞彙，無法順利找出相對應語料庫詞彙之詞頻（如：科學領域專有詞彙），加上字典詞彙在描述上和語料庫詞彙可能不盡相同，可能具有意思相近或同義（synonymous）關係。例如：在字典詞彙中，[test： 測試]表示英文在翻譯 test 一詞的中文詞彙爲“測試”，但在語料庫句子中，有可能不使用“測試”一詞，而使用和“測試”相似的詞彙，例如 “試驗”、“實驗”、 “試”、 “嘗試”等詞彙，這樣的情況下，就無法直接找出並統計 “測試”一詞在語料庫上的詞頻數，而必須考慮和“測試”相似的詞彙在語料庫上統計的詞頻數，間接反映“測試”在使用上的重要程度。故我們先將語料庫所有中文詞彙找出其相關之近義詞彙集合，之後，針對無法直接比對字典詞彙和語料庫詞彙找出其詞頻高低的詞彙，可檢查字典詞彙是否存在於語料庫詞彙之近義詞彙集合，若存在，代表可間接找出和該語料庫詞彙對應的詞頻，最後，建立字典詞彙對應語料庫詞彙之詞頻關係。

我們的中文語料庫來源參照國立科學教育館的期刊語料[9]（40 卷第 7 期至 46 卷第 2 期共 111 篇）以及師大科教中心教學教育月刊語料[10]（240 到 285 期共 218 篇），共 329 篇。首先，將上述語料經過中研院 CKIP 斷詞系統輔助，獲得約 73 萬 4 千個中文詞彙，接著統計重覆出現的中文詞彙次數，產生語料庫詞彙詞頻表，共有 23422 個中文詞彙，平均每篇詞彙個數約 2800 字，字典來源爲牛津現代英漢雙解詞典[11]共 39429 個詞彙。

爲了取得語料庫與字典詞彙的近義關係，我們首先將斷詞後的語料庫詞彙，分別輸入至 HowNet[8]近義詞尋找方法，及中研院一詞泛讀系統[2]，接著將這兩種方法所找到的近義詞集合，取交集篩選，最後輸出爲經過篩選過後之近義詞詞彙集合。

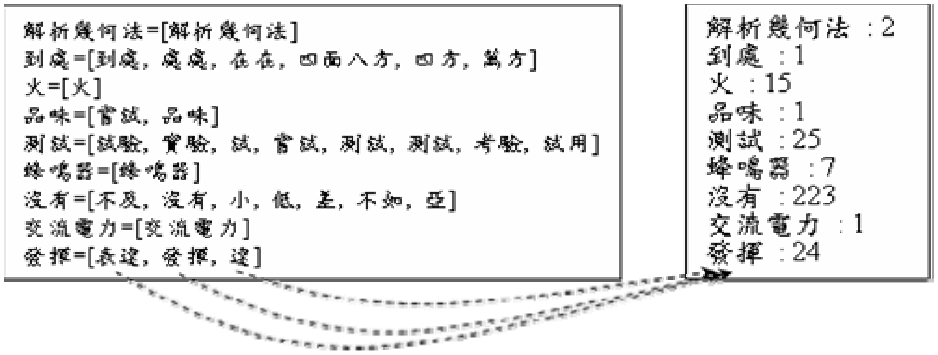
經過上述步驟，可建立**字典和語料庫相似字對應檔**和透過原先經過斷詞統計詞頻之**語料庫詞彙詞頻表**，建立字典和語料庫相似字對應檔和語料庫詞彙詞頻表之目的在於建

立字典詞彙對應語料庫詞彙之詞頻關係。建立流程如下。

i. 直接比對所有字典詞彙與語料庫詞彙詞頻表所記錄的語料庫詞彙，若相同，則記錄目前字典詞彙及所對應之詞頻數，若無，則記錄目前字典詞彙及其詞頻數爲 0。

ii. 針對目前字典詞彙及其詞頻數爲 0 的情況，利用字典和語料庫相似字對應檔，檢查字典詞彙是否出現於近義詞集合，若有，代表目前字典詞彙可間接對應近義詞集合所指向之語料庫詞彙，並找出其對應詞頻數，若無，則記錄目前字典詞彙及其詞頻數爲 0。

針對上述流程 2 部分，我們以圖四爲例，假設語料庫內有“發揮”一詞，查詢相似字對應檔發現其近義詞集合爲“表達，達”，若英漢字典不存在“發揮”中文詞，而有“表達”或“達”之詞彙，則根據近義詞對應關係，可以間接指向語料庫詞彙詞頻表之“發揮”所對應之詞頻（24 次）。因此，透過查找相似字的方式，找出所有相似字的詞頻，建立屬於字典詞彙的詞頻統計表，可做爲判斷中文詞彙重要程度的選詞參數。



圖四、字典和語料庫相似字對應檔(左)及語料庫詞彙詞頻表(右)

（三）中英詞彙對列和語言模型

由於語料庫統計詞頻和使用者選詞次數的方式調整選詞順序，並未考慮前後文之間的搭配關係，因此我們利用 word alignment 的技術，以科學人雜誌之中英對照雙語平行語料庫爲來源，以 GIZA++[6]爲分析工具，找出雙語詞組或詞彙的對照關係及其機率值，並透過以下機率模型計算以下的條件機率。

$$\Pr(c_i \mid c_{i-1}, e_i) \cong \Pr(c_i \mid e_i) * \Pr(c_i \mid c_{i-1}) \tag{1}$$

公式(1)中定義 c 爲中文翻譯詞彙，e 爲英文詞彙，利用前一個中文翻譯選詞的結果，即 $\Pr(c_i \mid c_{i-1})$，找出和目前中文翻譯詞 c_i 共現的機率，以及中英詞彙對列機率，找出兩者相乘之最大機率分數，以近似 $\Pr(c_i \mid c_{i-1}, e_i)$ 的數值，做爲選擇中文翻譯詞 c_i 的可能值。

不過，根據公式(1)所計算出來的機率分數，僅決定 bi-gram 的翻譯機率，若需要計算整句中文翻譯的機率分數，即考慮中文句子長度爲 N 時，相乘所有 bi-gram 的翻譯機率，結果如下。

$$\prod_{i=1}^{N} \Pr(c_i \mid c_{i-1}, e_i) \cong \prod_{i=1}^{N} \Pr(c_i \mid e_i) * \Pr(c_i \mid c_{i-1}) \tag{2}$$

綜合語料庫統計詞頻資訊、使用者選詞次數，以及利用雙語語料統計的機率值，可以將這些衡量詞彙選詞重要程度的方法當做參數，定義字典原文詞彙 e 的所有中文翻譯選項爲 $w_1, w_2, \ldots w_i$，以及定義語料庫統計詞頻資訊爲 $W_f(w_i)$、使用者選詞次數 $G_f(w_i)$ 以及根據公式(2)利用雙語語料統計的機率值 $P(w_i)$，作爲原文詞彙 e 之中某一翻譯選項

w_i 輸入至三種方法的函數名稱，其相加的分數定義 $S(w_i)$爲選詞計分函數的計算結果，如下表示。

$$S(w_i) = \alpha W_f(w_i) + \beta G_f(w_i) + \gamma P(w_i) \quad (3)$$

其中，$S(w_i)$所定義的 α、β、γ 爲各別決定 $W_f(w_i)$、$G_f(w_i)$和 $P(w_i)$計算之權重，由於目前研究仍無法決定 α、β、γ 的最佳值，因此設定 α=β=γ=1，代表三者的重要性相同。

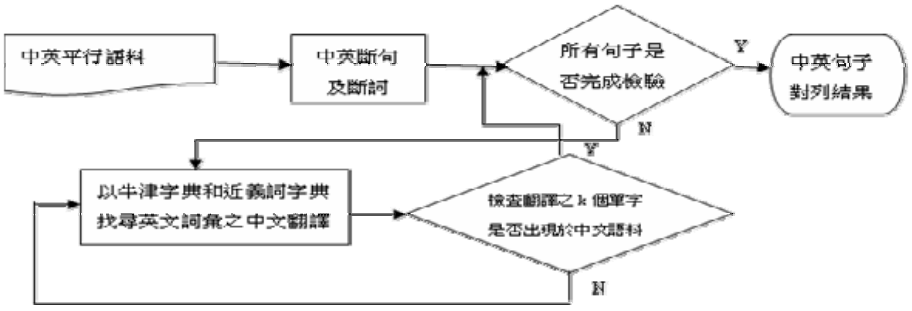
接下來，我們將利用 GIZA++[12]及 mkcls[11]等工具，介紹如何將利用科學人雜誌提供的中英雙語語料，經過簡單的中英語句對列（sentense alignment）技術，接著將中文語料部分，利用中研院 CKIP 斷詞系統加以斷詞，英文語料部分，則是經過英大小寫轉換及利用字和字之間空白斷詞，再來利用 word clustering 技術，建立詞彙分群，並加以編號，將文字語料轉換成純文字資訊，最後輸入至 GIZA++產生詞彙對列結果，並評估其實驗效果。

我們抽樣科學人雜誌 2002 年 3 月至 2006 年 12 月共 110 篇文章，句對數、中英詞彙數及中英總詞彙個數如表二所示。

表二、實驗語料來源統計

句對數		詞彙數	總詞彙個數(tokens)
中文	2780 句	9313 字	301747 個
英文		10608 字	390020 個

首先，由於科學人雜誌文章內容格式爲 html 網頁格式，加上其中英對照網頁的內容僅利用分欄的方式，將中英內容各別置入左右兩欄內，而未標註其中英句子相互對照的位置，以致於不能符合訓練資料需滿足句子對列（sentence alignment）的條件。然而，以人工方式針對中英語句判斷是否句子對列，需要大量的時間和精神，故我們依照圖五之簡易句子對列流程，找出可能的中英句子對列結果。



圖五　簡易句子對列流程

GIZA++計算 alignment 的機率，爲目標語言 t 在長度 m 以及來源語言 s 在長度 n 條件下語彙對列的條件機率，即計算 $\Pr(t_1^m \mid s_1^n) = \prod_m \Pr(t_j \mid s_{a_j})$。然而，實際檢查語料庫，根據語料庫提供的統計資料進行公式計算時，會因爲語料庫所提供的部分資訊不足，使得條件機率的值爲 0 的情形，即所謂的稀疏資料（sparse data）問題。因此，爲了能提昇 alignment 的品質，增加日後計算詞彙翻譯機率的正確性，Och 等人在 1999 年於“An Efficient Method for Determining Bilingual Word Classes”[5]提出一種 word clustering 方法，將分佈稀疏的語料資訊進行分配（partition），使得經常同時出現的單一語言詞彙（monolingual word）或雙語詞彙（bilingual word）能在同一類，即使來源詞彙和目標詞彙的對照資訊不足，也可改以檢查來源詞彙所屬的類別和目標詞彙的對照資訊，來增加其機率值。同時，Och 也提出了一套最佳化機制，在調整目標詞彙類別 T*（target word class)或來源詞彙類別 S*（source word class）時，檢查其詞彙翻譯條件機率值是否符合

$\arg\max_{S^*,T^*} \Pr(s_1^n \mid S^*) \Pr(t_1^m \mid s_1^n, S^*, T^*)$ 條件，確保 clustering 結果之正確性。

以 word clustering 概念爲基礎，我們利用 Och 所開發 mkcls 工具[11]，設定需要執行最佳化（optimize）的次數執行 5 次及 word class 個數設定爲 80 個，之後輸入中英語料給 GIZA++執行，透過 EM algorithm、HMM model 及 IBM-model 1~5 之訓練過程，反覆訓練 5 次，輸出中英詞彙對列結果及中英詞彙對照翻譯機率表。

在 2780 句中英句對中，我們發現在中英詞彙對照翻譯機率表裡，針對中文 9313 個詞彙對應至英文 10608 個詞彙共出現 18181 種不同詞彙對列組合。接著，我們將中英詞彙對照翻譯機率表裡出現過 18181 個英文詞彙，扣除相同英文詞彙對應多個中文詞彙的次數，可得到 5224 個不同英文詞彙的接受配對個數，和原先 10608 個英文詞比較起來，發現 5384 個爲 GIZA++判定爲英文詞彙對應至 NULL，即語料中無任何合適中文詞彙可供配對。若扣除 5384 個未出現於中英詞彙對照翻譯機率表中的詞彙，接著，以人工檢查的方式，檢查中英詞彙對照翻譯機率表內正確配對的中英詞對共有 3372 筆記錄，扣除掉相同英文詞彙對應多個中文詞彙的情況，尚有 2761 個不同英文詞彙是配對正確的中文詞彙解釋。

接著，根據公式(1)，計算中文字的組合影響到最後中文詞彙選擇的機率，因此需考慮中英詞彙對列機率及中文雙連語言模型，即 $\Pr(c_i \mid c_{i-1})$。

其中，我們建立中文雙連語言模型的方式爲利用 SRI Speech Technology and Research Laboratory 所開發的自然語言工具 SRILM（網址爲 http://www.speech.sri.com/projects/srilm/），輸入中文語料檔後，根據統計各中文詞彙及下一個中文詞彙的出現次數，計算其出現機率，最後輸出中文語言模型機率表，如圖六所示。

-2.150164	運動 消耗
-2.139987	運動 能力
-2.153002	運動 健將
-2.146435	運動 模式
-2.142739	運動 模型
-2.153002	運動 模擬儀
-2.153002	運動 器官
-2.144583	運動 聯盟
-1.341907	運算 ，

圖六、中文語言模型機率表

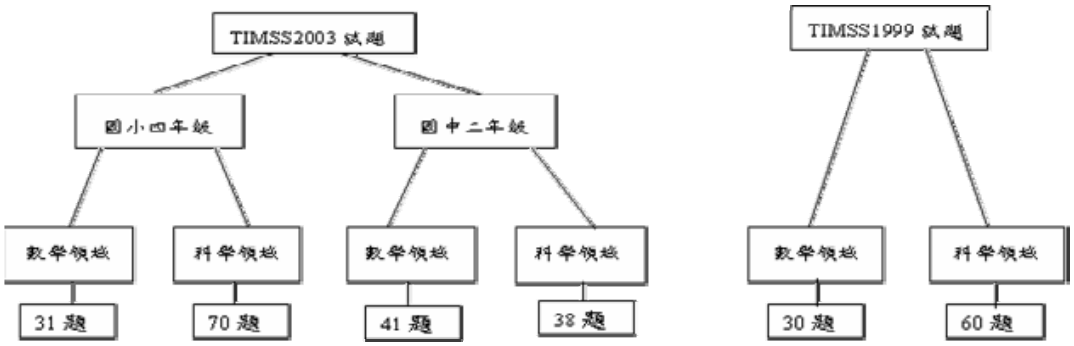
由圖三可觀察到，第一個欄位和第二個欄位分別表示爲各 bi-gram 詞彙下對應的 log probability（以 $\log_{10}$ 爲基底）。因此，我們以“sports leagues”爲例，依照公式(1)改寫的公式，計算當目前英文爲 “sports leagues”時， 我們確定 “sport”翻譯爲 “運動”，而“leagues”翻譯成 “聯盟”的機率爲。

$$\Pr(\text{聯盟} \mid \text{運動}, leagues) \cong \Pr(\text{聯盟}|leagues) * \Pr(\text{聯盟}|\text{運動})$$

其中，Pr(聯盟|leagues)可經由中英詞彙對照翻譯機率表查出 leagues 機率最大的中文翻譯，Pr(聯盟|運動)則是經由圖一中文語言模型機率表，找出中文 bi-gram 共現的條件機率值。

五、實驗和評估

我們實驗主要的資料來源爲 TIMSS1999 及 TIMSS2003 的試題，在試題分類上，依照 TIMSS 師大科教中心網頁，說明在 1999 年所出題之試題內容，主要區分數學及科學類別，並且以國中二年級爲考試對象。而在 2003 年所出題的題型內容上，除了依照 1999 年的試題類別外，在對象上則是以國小四年級及國中二年級爲考試對象。因此，我們按照上述之試題分類，將網頁所提供的中英試題內容，下載並轉檔爲純文字檔，接著將試題分類及題目個數統計如下。



圖七、TIMSS 試題分類及題數統計樹狀圖

從圖七來看，由於試題依照數學及科學領域的範圍內容，以及不同年級分別，在難易程度上會比照年級有所不同。例如國小四年級的數學會著重於數的認識、單位比較或基本四則運算，而國中二年級則除了前述之基本觀念外，還增加代數、分數四則運算等概念。在科學領域部分也有類似的學習進度分配，例如從國小四年級針對地科及生物等觀念，到國中二年級則進入理化等觀念的配合。另外，在專有名詞的使用上，以及題目內陳述的內容長短，國中二年級的內容較國小四年級的用詞上較爲艱深。

除了衡量以試題內容爲翻譯對象，測試線上翻譯系統及本系統在數學及科學領域翻譯程度的好壞，同時也應該考慮不同年級的試題內容，驗證翻譯系統在翻譯能力上，是否按照愈低年級其翻譯正確愈高的趨勢。故實驗的比較上按照年級別、試題種類分類（以數學領域爲 A 代號，科學領域爲 B 代號）及年分可大致區分如表三之組別表，以實驗國中二年級試題和國小四年級試題對於翻譯效果的影響程度。

表三、TIMSS 試題實驗組別表

中二 A 組	中二 B 組	小四 A 組	小四 B 組	中二 AB 組	小四 AB 組
TIMSS1999 及 TIMSS2003 國中數學領域試題	TIMSS1999 及 TIMSS2003 國中科學領域試題	TIMSS2003 國小數學領域試題	TIMSS2003 國小科學領域試題	TIMSS1999 TIMSS2003 數學及科學領域試題	TIMSS2003 數學及科學領域試題

我們實驗主要比較的翻譯系統爲谷歌（以下稱 Google Translate）和雅虎線上翻譯系統（以下稱 Yahoo!），以及本系統。其中本系統根據公式(3) $S(w_i)$之選詞計分公式調整，計有利用語料庫統計詞頻資訊爲 $W_f(w_i)$、使用者選詞次數 $G_f(w_i)$以及根據公式(2)利用雙語語料統計的機率值 $P(w_i)$等選詞策略。我們的實驗目的，在於評估本系統在不同選詞策略之下，經由和線上翻譯系統比較的結果，找出各別選詞策略，或不同選詞策略的搭配之下，其翻譯水準的差異性，並藉由各選詞策略實驗出來的結果，以排名的方式，試圖找出對本系統最有利的選詞策略組合。

有鑒於此，我們考慮以下不同選詞策略的組合，系統依照這些組合共可區分爲。

i. 隨機選詞：即系統不考慮任何選詞策略，以任意選詞的方式，輸出其選詞結果，

之後，藉由此方式評估並實證系統在最差情況的翻譯水準。

ii. 考慮翻譯修正階段選詞 $W_f(w_i)$模式：即考慮翻譯者選詞次數，來決定選詞順序。由於翻譯者選詞次數需要經由翻譯者長時間使用系統，才可進行實測，故在本研究中，無法針對 $W_f(w_i)$部分進行實證。

iii. 考慮中文語料庫詞頻選詞 $G_f(w_i)$：即透過收集來的語料庫詞頻高低，決定選詞順序。

iv. 考慮中英詞彙對列機率及語言模型 $P(w_i)$：即利用 GIZA++及 SRILM 工具完成的中英詞彙對照翻譯機率表和中文語言模型機率表等資訊，經由公式(2)計算最佳選詞路徑。

v. 同時考慮 $G_f(w_i) + P(w_i)$模式：在無法得知 $W_f(w_i)$資訊之情況下，計算 $S(w_i)= G_f(w_i) + P(w_i)$的選詞計分結果。

在評估翻譯效果的標準上，我們採用目前以 n-gram 匹確正確率爲基礎的 BLEU 和 NIST 評分策略爲標準，其評估方式如下。

i. BLEU 部分：

BP: 長度懲罰因子
L_{ref}: 參考翻譯句子長度
L_{sys}: 系統翻譯句子長度
$\delta_n(e_s, r_s)$ 系統翻譯句子e_s與參考翻譯句子r_s匹配的n-gram個數
$c_n(e_s, r_s)$ 系統翻譯句子e_s中n-gram的個數

$$\text{score}=\text{BP}*\exp\left(\sum_{n=1}^{N} W_n \log p_n\right) \quad (4)$$

$$\text{BP}=\min\{1,\exp(1-\frac{L_{ref}}{L_{sys}})\} \quad (5)$$

$$p_n=\frac{\sum_{s=1}^{S}\delta_n(e_s,r_s)}{\sum_{s=1}^{S}c_n(e_s,r_s)} \quad (6)$$

ii. NIST 部分：

$$\text{Info}(w_1\ldots w_n)=\log_2(\frac{\text{number of occurrences of } w_1\ldots w_{n-1}}{\text{number of occurrences of } w_1\ldots w_n}) \quad (7)$$

$$\text{score}=\sum_{n=1}^{N}\{\sum_{\substack{\text{all } w_1\ldots w_n \\ \text{that co-occur}}}\text{Info}(w_1\ldots w_n)/\sum_{\substack{\text{all } w_1\ldots w_n \\ \text{in sys output}}}\max(c_n(e_s,r_s),1)\}*\exp\{\beta\log^2[\min(\frac{L_{sys}}{L_{ref}},1)]\} \quad (8)$$

首先，我們針對系統依上述隨機選詞、$G_f(w_i)$、$P(w_i)$以及 $G_f(w_i) + P(w_i)$等選詞策略下，依照 cumulative 4-gram scoring 之 BLEU 和 NIST[7]評分標準，將表三 6 類的英文原文試題，連同師大科教中心所提供的 6 類原文試題的系統標準翻譯，以及本系統根據 6 類英文原文翻譯的系統建議翻譯，輸入至美國國家標準與技術局（NIST）所開發的 mteval-v10 之 BLEU 和 NIST 評分工具(網址爲 http://www.nist.gov/speech/tests/mt/scoring/index.htm)，再將計分結果輸出至檔案，如表四所示。

表四、本系統各選詞策略之 NIST 及 BLEU 值比較表

組別	中二 A 組		中二 B 組		小四 A 組		小四 B 組		中二 AB 組		小四 AB 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU
隨機選詞	1.6573	0.0201	1.2879	0.0000	0.8693	0.0000	0.6351	0.0096	1.4902	0.0000	0.7191	0.0093
$G_f(w_i)$	4.6169	0.2130	3.8632	0.1505	3.4706	0.2099	4.5275	0.2229	4.4453	0.1866	4.6349	0.2202
$P(w_i)$	2.6173	0.0662	2.2518	0.0613	2.0501	0.0848	0.9885	0.0223	2.4986	0.1330	1.5621	0.0501
$G_f(w_i)$ + $P(w_i)$	3.0430	0.0786	2.1371	0.0962	3.3754	0.1991	1.0977	0.0290	4.4453	0.1866	2.7722	0.1317

從表四針對各實驗組以 NIST 和 BLEU 評測的結果，可以發現隨機選詞和其他系統比較上明顯要低，尤其在中二 B 組、小四 A 組及中二 AB 組上的 BLEU 分數出現零分的情形，代表系統在無任何選詞方法調整之下，其翻譯效率是相當低的。在 $G_f(w_i)$語料庫選詞部分，和隨機選詞、P(wi)和 $G_f(w_i)$ + $P(w_i)$的系統相比，還是明顯要優於其他選詞策略。

我們接著以目前本系統評測最佳的 $G_f(w_i)$選詞策略之 BLEU 和 NIST 分數，和 Google Translate 和 Yahoo!線上翻譯系統加以比較，其結果如表五所示。

表五、本系統與線上翻譯系統之 NIST 及 BLEU 值比較表

組別	中二 A 組		中二 B 組		小四 A 組		小四 B 組		中二 AB 組		小四 AB 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU
Google Translate	2.8204	0.0964	2.2163	0.0677	2.1460	0.0993	1.1145	0.0171	2.5948	0.0834	1.6255	0.0523
Yahoo!	3.3721	0.1051	2.7817	0.0842	2.1312	0.1111	1.3517	0.0332	3.1779	0.0962	1.8199	0.0705
$G_f(w_i)$	4.6169	0.2130	3.8632	0.1505	3.4706	0.2099	4.5275	0.2229	4.4453	0.1866	4.6349	0.2202

從表五可以觀察得到，系統在 $G_f(w_i)$選詞策略之下，各組的 BLEU 值和 NIST 值皆比 Google Translate 和 Yahoo!線上翻譯系統要高，顯示建立一致性翻譯的詞彙或詞組，以及科學與數學相關之中文語料詞頻的確能提高選詞的正確性。也同時反應 Yahoo!和 Google Translate 翻譯系統設計的目的是將廣泛領域知識（domain genernal）的翻譯內容，以語料統計中最可能的選詞方式，採用直接翻譯或 phrase-based 翻譯將內容輸出。因此，選詞策略上較無法以限制特定領域知識（domain spasific）範圍的做法，即加入更多數學及科學領域的片語句型或專有名詞的規則轉換條件，使得反應在系統建議翻譯的翻譯詞彙與 TIMSS 試題的參考翻譯的結果有所差距，由於參考的翻譯僅有師大科教中心針對 TIMSS1999 及 2003 年版本的翻譯內容，使得 Google Translate 或 Yahoo!爲了增加翻譯結果流暢度所加入之詞彙，可能因爲缺乏其他參考對象，而影響最後的翻譯的品質或正確性。

比較有趣的是，從表四和表五可發現考慮中英詞彙對列機率及語言模型 $P(w_i)$在分數上十分接近 Google Translate 的表現，並且在 $G_f(w_i)$ + $P(w_i)$的選詞分數上，亦與 $G_f(w_i)$的分數相差不多。不過從這兩點的比較上，皆發現加入 $P(w_i)$的結果並不能使 BLEU 與 NIST 的分數提升，反而在分數上較 Google Translate 以及 $G_f(w_i)$略低。由於 $P(w_i)$以科學人雜誌爲雙語語料統計之機率選詞模型，在語料類型上是接近 TIMSS 試題的內容，屬於科學與數學範疇，因此理論上在 BLEU 和 NIST 分數上會較同爲統計式翻譯爲基礎，但語料來源不完全針對科學領域統計的 Google Translate 要好，但仍比僅考慮個別中文翻譯選項之詞頻高低的 $G_f(w_i)$之分數來得低，主要原因可歸納幾點。

i. **統計語料在數量上的不足**：由於以 GIZA++詞彙對列工具所找出之正確中英詞彙對列個數僅有 2761 個，因此在預測適合之中文翻譯選項時明顯無法找出牛津字典 39429 個英文詞彙中所有中文翻譯選項的可能對列機率，因此可能會有選詞上的誤差。

ii. **系統產生的中文翻譯句子未經任何詞序上的調整**：由於系統翻譯的模式是按照英文詞彙原有的詞彙順序，將單字與片語詞組切分後查詢牛津字典之中文翻譯結果，因此中文詞序上和一般中文句子不同，而 $P(w_i)$在語言模型的機率統計上，是依照一般中文詞序統計，因此實際將語言模型之統計結果，套用至本系統時，會因爲許多中文 bi-gram 在 $P(w_i)$的語言模型未出現過，使得機率值爲 0 的情形，雖然在做法上，我們利用乘上極小數（我們預設以 10^{-20}）的方式，避免因機率值爲 0，影響中英詞

彙對列機率的結果，但系統本身的中文詞序，仍然會造成 $P(w_i)$對照語言模型的機率計算上出現不小的誤差。

針對 2 的推論，我們再度針對 $P(w_i)$公式的結構，區分新的實驗組別，進一步探討 $P(w_i)$公式各部分對於整體 BLEU 和 NIST 值的變化，並驗證 2 推論的正確性。由前述之公式(2)可得知 $P(w_i)$爲

$$\prod_{i=1}^{N} \Pr(c_i \mid c_{i-1}, e_i) \cong \prod_{i=1}^{N} \Pr(c_i \mid e_i) * \Pr(c_i \mid c_{i-1})$$

從公式(2)右半部可得知$\Pr(c_i \mid e_i)$爲中英詞彙對列機率模型，$\Pr(c_i \mid c_{i-1})$爲中文雙連語言模型，而表四完成的 $P(w_i)$是經由公式(2)計算的選詞結果，因此我們分別比較單獨利用中英詞彙對列機率模型（以 Align 取名）和中文雙連語言模型（以 Bigram 取名），和 $P(w_i)$結果上的差異性，如表六所示。

表六、P(wi)與各別機率模型之 NIST 及 BLEU 值比較表

組別	中二 A 組		中二 B 組		小四 A 組		小四 B 組		中二 AB 組		小四 AB 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU
$P(w_i)$	2.6173	0.0662	2.2518	0.0613	2.0501	0.0848	0.9885	0.0223	2.4986	0.1330	1.5621	0.0501
Align	2.6940	0.0686	2.6165	0.0926	3.3518	0.1923	1.0719	0.0242	3.1006	0.1142	2.7707	0.1239
Bigram	2.3934	0.0588	2.4677	0.0895	2.4126	0.0988	0.9760	0.0229	2.8480	0.0957	2.2474	0.0824

從表六各組別 NIST 和 BLEU 值可觀察出，由於 $P(w_i)$的分數是按公式(2)由 Align 和 Bigram 的機率分數相乘而得，受到 Bigram 分數較低的影響，和 Align 計算的機率分數相乘會使得最後 $P(w_i)$的機率分數更低，因此其 NIST 和 BLEU 分數也相形較低。並且，從 Align 的大部分分數（除中二 AB 組）高於 $P(w_i)$的情形來看，可說明 Align 和 Bigram 相結合的 $P(w_i)$之後其 NIST 和 BLEU 分數的確會有減分的效果，亦證實 2 的推論爲正確。

六、結論

在本研究中，我們爲了減少師大科教中心以人工方式進行 TIMSS 試題翻譯、校稿及討論的時間，我們嘗試實作一輔助試題翻譯系統，在評估翻譯效率實驗過程當中，我們透過 BLEU 及 NIST 爲評估指標，進行 TIMSS 試題翻譯的結果，發現即使在某些選詞爲任意選詞策略之下，系統詞彙翻譯之正確度上能亦能有接近線上翻譯之成果。

針對翻譯結果的流暢度上，雖然實作出利用使用者介面，進行新增詞彙及刪除詞彙等動作，來達到翻譯結果符合語法之描述，但未處理有關自動加減詞彙之機制，因此，可透過統計式翻譯模型的輔助，找出翻譯詞彙中最大相鄰之可能詞彙，或者找出某些詞彙因相鄰關係而使翻譯詞彙產生減少詞彙之關係，來增進並改善翻譯的流暢程度。

在翻譯詞序的調整部分，可利用語法剖析的方式，利用中英剖析樹的對應，找出類似的結構對應關係，在判斷新的翻譯句子時，利用其對應關係的相似程度，找出最大可能之詞序組合，其中，需要克服的條件，在於語法剖析樹在結構的種類上，視描述方法不同可能會有許多不同的組合，因此需以分類及相似度比對的方式，先找出最大相似之子樹結構，之後在判斷未知句子結構時，才能有效減少比對相似之維度，同時，在中英詞性種類的對應上，由於詞性判斷的標準不一，例如“美麗（Beautiful）”一詞，英文詞彙爲形容詞，而在中文詞性判斷上，不同語境之條件下可能其中文詞彙可能會判讀成

副詞或形容詞，因此，要先能解決詞性判別之歧義問題，才可進行上述分類及相似度比對之方法，在這同時，翻譯系統執行時如何有效率執行結構的比對，也是進一步思考的範圍之一。

參考文獻

[1] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, A Statistical Approach to Machine Translation, Computational Linguistics, 16:2, 79-85, 1990.

[2] Cheng, Chin-Chuan, Word-focused Extensive Reading with Guidance, Selected Papers from the Thirteenth International Symposium on English Teaching, 24-32, 2004.

[3] Fons Van de Vijver and Ronald K. Hambleton, Translating Tests: Some Practical Guidelines, European Psychologist, 1, 89-99, 1996.

[4] Ide Nancy, Véronis Jane and Word Sense Disambiguation: The State of the Art, Computational Linguistics, 24:1, 1-40, 1998.

[5] Franz Josef Och, An Efficient Method for Determining Bilingual Word Classes, Proceedings of European Chapter of the Association for Computational Linguistics, 71-76, 1999.

[6] Franz Josef Och, Hermann Ney, Improved Statistical Alignment Models, Proceedings of the Thirty-eightth Annual Meeting of the Association for Computational Linguistics, 440-447, 2000.

[7] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., BLEU: a method for automatic evaluation of machine translation, Fourtyth Annual meeting of the Association for Computational Linguistics, 311-318, 2002.

[8] Yang Xiaofeng and Li Tangqiu, A Study of Semantic Disambiguation Based on HowNet, International Journal of Computational Linguistics & Chinese Language Processing, 7:1, 47-78, 2002.

[9] 國立科學教育館科學研習月刊，http://www.ntsec.gov.tw/publish/pdf.asp

[10] 師大科教中心科學教育月刊，http://140.122.147.172/journal/(new)journal.htm

[11] Concise Oxford English Dictionary，http://stardict.sourceforge.net/Dictionaries_zh_TW.php

[12] Google Translate, http://www.google.com/translate_t

[13] JACOB PROJECT, http://danadler.com/jacob/

[14] MINIPAR HOME PAGE, http://www.cs.ualberta.ca/~lindek/minipar.htm

[15] MXPOST, http://ccc.kmit.edu.tw/wiki/advanceprogramming/index.htm

[16] TIMSS 中文版官方網頁, http://timss.sec.ntnu.edu.tw/timss2007/news.asp

[17] The Porter Stemming Algorithm, http://www.tartarus.org/martin/PorterStemmer/

[18] WordNet API, http://nlp.stanford.edu/nlp/javadoc/wn/

[19] YAHOO! 雅虎線上翻譯, http://tw.search.yahoo.com/language/

Word sense induction using independent component analysis

Petr Šimon
Institute of Linguistics
Academia Sinica, Taiwan
sim@klubko.net

Jia-Fei Hong
Graduate Institute of Linguistics
NTU, Taiwan
jiafei@gate.sinica.edu.tw

Abstract

This paper explores the possibilities of using independent component analysis (ICA) for features extraction that could be applied to word sense induction. Two different methods for using the features derived by ICA are introduced and results evaluated. Our goal in this paper is to observe whether ICA based feature vectors can be efficiently used for word context encoding and subsequently for clustering. We show that it is possible, further research is, however, necessary to ascertain more reliable results.

1 Introduction

Word senses are known to be difficult to discriminate and even though discrete definitions are usually sufficient for humans, they might pose problems for computer systems. Word sense induction is a task in which we don't know the word sense as opposed to more popular word sense disambiguation.

Word sense can be analyzed by observing behaviour of words in text. In other words, syntagmatic and paradigmatic characteristics of a word give us enough information to describe all it's senses, given that all it's senses appear in the text.

Based on this assumption, many techniques for word sense induction have been proposed. All are based on word co-occurrence statistics. There are two

strategies for creating the vectors that encode each word: global encoding strategy, which encodes co-occurrence of word types with other word types and local encoding strategy which encodes co-occurrence of word tokens with word types. The global encoding strategy is more popular, because it provides more information and does not suffer from data sparseness and most of the research has focused on sense analysis of words of different forms, i.e. on phenomena like synonymy etc. However, by encoding word types, we naturally merge all the possible sense distinctions hidden in word's context, i.e. context of a token. For more details cf. (3; 11; 10).

Problem of high dimensionality that would be computationally restricting, us usually solved by one of several methods: principal component analysis (PCA), singular value decomposition (SVD) and random projection (RP) and latent semantic analysis, also known as latent semantic indexing is a special application of dimensionality reduction where both SVD and PCA can be used. See (1; 2) for overview and critical analysis.

The classical approach to word context analysis is a vector space model, which uses simple the whole co-occurrence vectors when measuring word similarity. This approach also suffers from a problem similar to data sparseness, i.e. the similarity of words is based on word forms and therefore fails in case where synonym rather than similar word form is used in the vector encoding (11; 10).

Major problem with the classical simple vector space model approach is the superficial nature the information provided by mere co-occurrence frequency, which can only account for seen variables. One of the most popular approaches to word context analysis, latent semantic analysis (LSA), can improve this limitation, by creating a latent semantic space using SVD performed on word by document matrix. Frequency of occurrence of each word in a document represents each entry w_{ij} in the matrix, thus, the whole document serves as a context. Document is, naturally, some sort of meaningful portion of text. SVD then decomposes the original matrix into three matrices: word by concept matrix, concept by concept matrix and concept by document matrix. The results produced by LSA are, however, difficult to understand for humans (9), i.e. there is no way of explaining their meaning.

2 ICA

Independent component analysis (ICA) (7) is a statistical method that takes into account high order statistical dependencies. It can be compared to PCA in the sense that both are related to factor analysis, but PCA uses only second-order statistics, assuming Gaussian distribution, while ICA can only be performed on non-Gaussian data (6). Comparison with SVD is provided by (12) on word context

analysis task.

ICA is capable of finding emergent linguistic knowledge without predefined categories as shown in (4; 5) and others.

As a method for feature extraction/dimensionality reduction it provides results that are approachable by humans reader. Major advantage of ICA is that it looks for factors that are statistically independent, therefore it is able find important representation for multivariate data.

ICA can be defined in a matrix form as $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ represents the independent variables, components, and the original data is represented by $\mathbf{x} = (x_1, x_2, ..., x_n)^T$, which can be decomposed into $\mathbf{s} \times \mathbf{A}$, where $\mathbf{A}$ is a $n \times n$ square mixing-matrix.

Both the mixing-matrix $\mathbf{A}$ and independent components $\mathbf{s}$ are learning by unsupervised process from the observed data $\mathbf{x}$. For more rigorous explanation see (7).

We have used FastICA algorithm as implemented in R language[1].

3 Data collection

The context matrix has been constructed from words from Sinica Corpus of a frequency higher than 150. This restriction yielded 5969 word types. We have chosen this limited lexicon to lower the complexity of the task.

The whole corpus was stripped from everything but all words whose word class tag started with N, V, A or D. This means that our data consisted of nouns (N, including pronouns), verbs (V), adjectives (A) and adverbs (D) [2].

Then we collected co-occurrence statistics for all words from window of 4 preceding and 4 following words, but only if these were within a sentence. We defined sentence simply as a string of words delimited by ideographic full-stop, comma, exclamation and question mark (◦ , ， , ！ and ？). In case of context being shorter than 4 words, the remaining slots were substituted by zero indicating no data available.

We have normalized the data by taking log of each data point a_{ij} in context matrix. Since this is a sparse matrix and lot of data points are zero, one has been added to each data point.

After the extraction of the independent components, we have encoded contexts of word tokens for each word type selected for analysis using these independent components. Thus we are able to provide reliable encoding for words, which is based on global properties. Note that there is no need to pursue orthogonality of different word types that are sometimes required in the context encoding. The

[1] http://www.stats.ox.ac.uk/ marchini/software.html

[2] For complete list see: http://wordsketch.ling.sinica.edu.tw/gigaword_pos_tags.html

similarities between different word types are based on the strength of independent components for each word type and therefore much better results of similarity measure can be expected than one would get from binary random encoding as introduced in (8).

We could experiment with several strategies to context matrix construction: different word classes in the context and different sizes of feature vectors. Context in our experiments is defined by four words that precede and four words that follow each keyword. Then we study the feature similarities across different words. To aid the analysis, a hierarchical clustering is used to determine closeness of relation among feature vectors of specified dimension. This step is to find most reliable feature vector dimension for subsequent experiments. As mentioned before, the features can be traced back and their nature determined, i.e. they can be labelled.

Due to the time constraints, we have predetermined feature vector size beforehand. We've extracted 100 and 1000 independent components and used them in two separate experiments.

Having determined the size of feature vectors, we use original word contexts for each word token and encode the context using these vectors. That means that each word in the context of particular keyword is replaced by it's respective feature vector, a vector of quantified relations to each of the independent components that has been extracted by ICA from the global co-occurrence matrix.

We than use maximum-linkage hierarchical clustering to find related words and based on the features present in the vectors we determine their characteristics that will provide clues to their word senses.

4 Results

We ran two experiments, one with 100 independent components and the second with 1000 components. For the experiment we have manually selected 9 words, which we expected to be easier to analyze. We have, however, failed to find in Chinese word that would allow for such obvious sense distinctions as English *plant*,*palm*, *bank* etc. Such words are typically used in word sense related task to test the new algorithms. The failure to find words that would have similarly clear-cut sense distinctions, might have influenced our initial results. The words we have selected are (number in bracket indicates the number of senses according to Chinese Wordnet)[3]: 山頭 (3), 名牌(2), 犯規(2), 約談(2), 措辭(2), 堅硬(2), 富有(2), 屬下(2), 天氣(3).

[3]http://cwn.ling.sinica.edu.tw

4.1 Independent components

When ICA algorithm retrieves the specified number of independent components, each of them can be labelled by creating a descending list of those words that are most responsive for each of the components (5; 4). Only the most responsive word could be assigned to each of the components as a label, but this way we would not be able to determine characteristics of the components with sufficient clarity. As we will see, even listing several items from the top of the list of the most responsive words, won't always provide clear explanation of the nature of the component in question. This is due to the fact that the independent components are not yet very well understood, that it is not yet entirely obvious how the components are created (5).

Bellow are few examples independent components and labels assigned to each of them. We list up to 20 most responsive words for each component to provide information for human judgment. These are examples from the 100 independent components experiment. For future research, perhaps an automatic way of determining different number of labels required to explain each independent component might be proposed using time series analysis, but for that, more research has to be provided to better understand the nature of independent components in order to justify such step.

First ten independent components can be seen in 1. As we can see, independent components cannot we regarded as synsets as known in WordNet, since they clearly contain words from multiple classes. We can perhaps call them collocation sets, colsets. But this term will have to be revised based on the subsequent research on the nature of independent components.

Table 4.1 shows an example how a particular word type is encoded. The independent components in this example are are sorted by the most important features. We can see how the encoding in Table 4.1 contrasts with Table 4.1, which shows ten least salient features for word type *yuyan* 語言.

4.2 Sense clustering

We have used maximum-linkage hierarchical algorithm from Pycluster package[4] to cluster word token contexts. The use of hierarchical clustering is motivated by the attempt to provide gradual sense analysis where subsenses could be identified within partial senses.

Our goal in this paper is to observe whether ICA based feature vectors can be efficiently used for word context encoding and subsequently for clustering. Clustering results were evaluated by native speaker with linguistics knowledge, who labelled all the sentences according to Chinese Wordnet and in this paper,

[4]http://bonsai.ims.u-tokyo.ac.jp/ mdehoon/software/cluster/software.htm

Label	IC	Responsive words (descending order)
TIME	0	時間 年 月 小時 天 段 半 經過 期間 週 分鐘 後 星期 久 持續 內 日 之後 工作 結束
TIME	1	三十 二十 五十 一百 四十 十 公尺 公里 歲 十五 以上 六十 超過 約 大約 左右 · 分鐘 十二 八十
FAMILY	2	媽媽 母親 孩子 爸爸 父親 女兒 父母 歲 兒子 家 小孩 弟弟 回家 妹妹 哥哥 帶 太太 照顧 回來 家人
COMPARATIVE	3	項 不同 電腦 好 什麼 系統 孩子 以上 後 路段 肯 設置 系所 當地 考 救 參與 最近 專線 事件
POPULATION	4	成長 去年 增加 今年 人數 減少 營收 達 預估 成長率 季 佔 比例 期 衰退 同 高達 明年 人口 營業額
GAIN	5	得到 獲得 受到 受 肯定 給予 尊重 給 重視 關心 鼓勵 能 支持 表現 關懷 意見 太 都 獲 照顧
MULTIMEDIA	6	媒體 電視 新聞 廣告 報導 節目 雜誌 報紙 廣播 記者 電台 傳播 電視台 宣傳 製作 電子 大眾 電話 報 刊登
WAR	7	伊拉克 軍事 飛彈 部隊 攻擊 美國 中共 戰爭 蘇聯 武器 科威特 波斯灣 行動 聯合國 美 美軍 以色列 國防部 海珊 中東
WARNING	8	注意 應 不要 特別 避免 重要 應該 結果 提醒 最好 要 點 選擇 準備 安全 小心 健康 保持 飲食 呼籲
COMPETITION	9	選手 比賽 冠軍 運動 屆 中華 錦標賽 女子 亞運 參加 世界 協會 金牌 體育 男子 球員 國 我國 國際 教練
PRODUCTION	10	生產 技術 工業 製造 設備 工廠 產業 科技 產品 機械 材料 電子 廠 研發 化學 原料 知識 科學 加工 農業
ECONOMY	11	元 經費 費用 補助 預算 筆 錢 美元 負擔 金額 支出 收入 支付 新台幣 成本 貸款 每 資金 給 花費
RESEARCH	12	資料 調查 報告 結果 統計 顯示 分析 做 研究 份 進入 依據 指出 數據 預測 專家 地震 發現 評估 正確

Table 1: Independent components: 100 IC set, first 10 IC

Feature strength	Responsive words (descending order)
7.55055952072	用 字 聽 語言 首 英文 句 唱 音樂 詞 表達 歌心 國語 獲得 寫 聲音 使用 詩 歌曲
6.93665552139	特色 具有 具 原住民 特殊 文化 獨特 語言 風格 色彩 豐富 背景 不同 特性 表現 很多 當地 傳統 歷史 最
6.20834875107	教學 英語 國小 國中 教育 學習 老師 課程 教師 小學 高中 學校 孩子 小朋友 學生 家長 教材 數學 教科書 英文
3.42819428444	她 得 我 他 快 孩子 玩 吃 深 態度 全 起來 父親 父母 跑 家庭 母親 共同 相當 一起
3.34706568718	品質 提高 高 提升 水準 成本 降低 效率 達到 低 提昇 改善 安全 整體 保障 服務 過 國民 享受 考量

Table 2: Partial example of encoded word 語言 (five most salient features)

Feature strength	Responsive words (descending order)
0.157807931304	申請 規定 昨天 不得 下午 取得 法院 任何 辦理 證明 行為 許 同意 違反 上午 是否 接受 機關 多 凌晨
0.157353967428	了解 不同 觀察 去 思考 看 分析 重新 調整 看看 深入 調查 重要 較 瞭解 面對 一下 探討 從 體會
0.152415782213	起 九月 三月 七月 六月 一日 五月 四月 二月 十二月 自 十月 民國 八月 至 十一月 底 一月 止 十五日
0.0953392237425	很 最 非常 相當 較 太 比較 更 十分 愈 比 越 極 得 那麼 這麼 一點 愈來愈 越來越 甚
0.0753756538033	選手 比賽 冠軍 運動 屆 中華 錦標賽 女子 亞運 參加 世界 協會 金牌 體育 男子 球員 國 我國 國際 教練

Table 3: Partial example of encoded word 語言 (five least salient features)

犯規 IC^{100}		
Cluster	Sense	Count
a	0	5
	1	28
b	0	3
	1	1

犯規 IC^{1000}		
Cluster	Sense	Count
a	0	1
	1	0
b	0	6
	1	30

Table 4: Results for word 犯規

措辭 IC^{100}		
Cluster	Sense	Count
a	0	9
	1	1
b	0	1
	1	8

措辭 IC^{1000}		
Cluster	Sense	Count
a	0	0
	1	1
b	0	9
	1	9

Table 5: Results for word 措辭

number of sense were also determined this way. Then we have assigned sense label to each cluster according to most prevalent sense in the cluster.

For example, word *fangui* 犯規 has two sense in Chinese Wordnet. We cut the tree produced by hierarchical clustering algorithm into two and our expectation is that word tokens manually labelled as sense 1 will be in one of the clusters and word tokens labelled as sense 2 will be in the other. Naturally some incorrect classifications can be expected as well and therefore we assign sense label according to the label most frequent in the particular cluster. In case we get both clusters labelled the same, the sense induction has failed.

In this experiment we have not pursued correct classification of all the words, therefore we leave the evaluation of those results out.

For reference we include tables with results of several words.

約談 IC^{100}		
Cluster	Sense	Count
a	0	22
	1	67
b	0	1
	1	12

約談 IC^{1000}		
Cluster	Sense	Count
a	0	1
	1	0
b	0	21
	1	80

Table 6: Results for word 措辭

山頭 IC^{100}		
Cluster	Sense	Count
a	0	39
	1	21
	2	11
b	0	7
	1	10
	2	0
c	0	2
	1	1
	2	0

山頭 IC^{1000}		
Cluster	Sense	Count
a	0	32
	1	39
	2	9
b	0	0
	1	3
	2	0
c	0	5
	1	2
	2	1

Table 7: Results for word 措辭

Word	IC100	IC1000
山頭	0	0
名牌	0	0
犯規	1	1
約談	0	1
措辭	1	1
堅硬	0	1
富有	0	0
屬下	0	1
天氣	0	0

Table 8: Overall results

5 Conclusion

The major advantage of our approach is that it uses global characteristics of words based on their co-occurrence with other words in the language, which are then applied to derive local encoding of word context. Thus we retrieve reliable characteristics of word's behaviour in the language and don't loose the word sense information, which allows us to analyze semantic characteristics of similar word forms.

Our current results are not very satisfying. I can be observed, however, from Table8 that increased number improves the sense induction considerably. We will pursue this track in our subsequent research. On the other hand, this result is not surprising. Considering the nature of independent components, which are rather symbolic features similar to synonymic sets, synsets, or rather collocation sets, collsets, it can be expected that much larger number of these components would be required to encode semantic information.

6 Future work

With manually semantically tagged word tokens we will try to automatically estimate the sufficient number of independent components that would improve precision of sense clustering.

Another approach we intend to try is to add feature vectors of all the context words and cluster the resulting vectors. This approach should emphasize more important features in given contexts.

We will also do more carefull preprocessing and also apply dimensionality reduction (typically done by PCA) before running ICA as has been done in some of the previous studies.

References

[1] E. Bingham. *Advances in Independent Component Analysis with Applications to Data Mining*. PhD thesis, Helsinki University of Technology, 2003.

[2] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, New York, NY, USA, 2001. ACM Press.

[3] S. Bordag. Word sense induction: Triplet-based clustering and automatic

evaluation. In *11 th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*, pages 137–144, 2006.

[4] T. Honkela and A. Hyvärinen. Linguistic feature extraction using independent component analysis. In *Proc. of IJCNN 2004*, 2004.

[5] T. Honkela, A. Hyvärinen, and J. Väyrynen. Emergence of linguistic features: Independent component analysis of context. In A. C. et al., editor, *Proceedings of NCPW9*, 2005.

[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley, 2001.

[7] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4):411–430, 2001.

[8] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1, pages 413–418. IEEE Service Center, Piscataway, NJ, 1998.

[9] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 2001.

[10] D. B. Neill. Fully automatic word sense induction by semantic clustering. Master's thesis, Cambridge University, 2002.

[11] R. Rapp. A practical solution to the problem of automatic word sense induction. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 194–197, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] J. Väyrynen and T. Honkela. Comparison of independent component analysis and singular value decomposition in word context analysis. In *AKRR'05*, pages 135–140.