International Journal of

Computational Linguistics & Chinese Language Processing

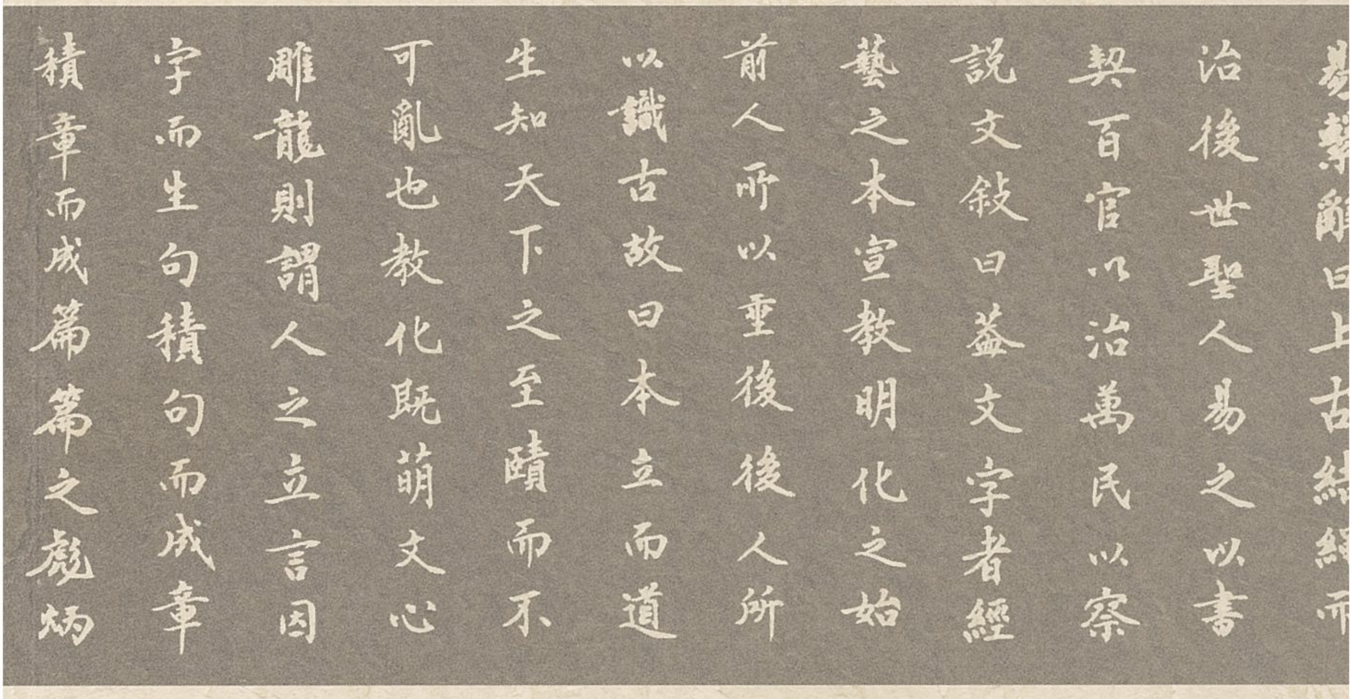
中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

Special Issue on “Selected Papers from ROCLING XXVI”

Guest Editors: Jen-Tzung Chien, Hung-Yu Kao, and Chia-Hui Chang



Vol.19 No.4 December 2014 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

Executive Editor: *Abby Ho*
English Editor: *Joseph Harwood*
The Association for Computational Linguistics and Chinese Language Processing, Taipei

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 19, no. 4, December 2014

Contents

Special Issue Articles:

Selected Papers from ROCLING XXVI

Papers

Special Issue Article:

Forewords

The 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014) was held at National Central University, Jhongli, Taiwan on Sep. 25-26, 2014. ROCLING, which sponsored by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), is the leading and most comprehensive conference on computational linguistics and speech processing in Taiwan, bringing together researchers, scientists and industry participants from fields of computational linguistics, information understanding, and speech processing, to present their work and discuss recent trends in the field. This special issue presents extended and reviewed versions of five papers meticulously selected from ROCLING 2014, including 3 natural language processing papers and 2 speech processing papers.

The first paper from National Central University focused on constructing a large POI (Point-of Interest) database. They solve problems of Taiwan address normalization, store name extraction, and the matching of addresses and store names by training a statistical model. They obtain 0.791 F-measure for store name recognition on search snippets. The second paper from National Taiwan University extracted policy positions from data collected from recent highly-debated Cross-Strait Service Trade Agreement (CSSTA) to predict the electoral behavior from this information. They used the keywords of each position to do the binary classification of the texts and count the score of how positive or negative attitudes toward CSSTA. The proposed approach saves human labor of the traditional content analysis and increases the objectivity of the judgment standard. The third paper from National Tsing Hua University constructed a system to aid academia paper writing. They used writing common patterns to train a writing sentence classifier. This system can also provide hints for users to guild their paper writing.

The last two papers are speech processing papers. The paper from National Taiwan Normal University investigated and developed language model adaptation techniques for use in ASR (automatic speech recognition). The proposed approach measured the relationships between a search history and an upcoming word. Their language models can offer substantial improvements over the baseline N-gram system and some state-of-the-art language model adaptation methods. The paper from Academia Sinica, Taiwan presented study examines prosodic characteristics of Taiwan (TW) English in relation to native (L1) English and TW speakers’ mother tongue, Mandarin. By examining prosodic patterns of word/sentence, similarity analysis in this paper suggests that between-speaker similarity is greater when they are in the same speaker group in both word and sentence layer.

The Guest Editors of this special issue would like to thank all of the authors and reviewers for sharing their knowledge and experience at the conference. We hope this issue provide for directing and inspiring new pathways of NLP and spoken language research within the research field.

Guest Editors

Jen-Tzung Chien

Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

Hung-Yu Kao

Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan

Chia-Hui Chang

Department of Computer Science and Information Engineering, National Central University, Taiwan

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 1-18

POI 擷取:商家名稱辨識與地址配對之研究

POI Extraction from the Web: Store Name Recognition and Address Matching

林育暘、張嘉惠

Lin Yu-Yang[*] and Chang Chia-Hui[*]

摘要

行動化是 2014 的趨勢之一，而適地性服務（Location-based Service）在這波趨勢中具有至關重要的地位，因為裝置行動化的因素，大量查詢需求因此誕生，例如：路線導航、查詢附近餐廳、加油站等。適地性服務要能廣泛的提供服務，通常需要有一個完整的 POI（Point of Interest）資料庫，而整個網路就是最大的資訊來源。這些資料源自於網站管理者、群眾外包（crowdsourcing）或個人使用者所分享的資訊，包括了地址、名稱、電話、評論等資訊。現在雖然有各種擷取地址相關資訊的方法，但經常面臨無法取得明確POI的名稱，在資訊檢索上受到很大的限制。

在本篇論文中，我們提出一個商家名稱辨識的方法，藉由收集網路上包含地址的網頁，建立一個具有商家名稱與地址關聯性的資料庫，以提高地址相關資訊檢索的效果，讓使用者在使用行動裝置查詢時，能直接輸入店家名稱或關鍵字查詢地址之服務，提供便利的查詢功能。其中，在商家命名實體辨認上，本篇論文提出了商家與組織名稱在命名上的共同特性，利用此共同特性當作特徵加入 CRF 模型，以提供 N-Gram 與詞性之外的特徵。

關鍵詞：商家地理資訊檢索、商家名稱擷取、商家名稱與地址配對、序列標記、條件隨機域

[*]國立中央大學資訊工程學系
Department of Computer Science and Information Engineering, National Central University
E-mail: 101522052@cc.ncu.edu.tw; chchang8ncu@gmail.com

Abstract

Mobility is one of the trends in 2014. According to the report of IDC (International Data Corporation), the worldwide shipments of tablets have exceeded PCs in 2013 Quarter 4, while smart phones has already exceeded other devices in unit shipments and market ratio. With this trend, many location-based services (LBS) have been proposed, for example, navigation, searching restaurants or gas stations. Therefore, how to construct a large POI (Point-of Interest) database is the key problem. In this paper, we solve three problems including Taiwan address normalization, store name extraction, and the matching of addresses and store names. To train a statistical model for store name extraction, we make use of existing store-address pair to prepare training data for sequence labeling. The model is trained using common characteristics from store names in addition to POS tags. When testing on search snippets, we obtain 0.791 F-measure for store name recognition.

Keywords: POI, Store Name Extraction, Name-address Matching, Sequence Labeling, Conditional Random Field

1. 緒論

根據國際數據資訊 IDC 於 2013 年 9 月調查報告顯示，平板電腦的出貨量在 2013 年第四季首次超過個人電腦，而智慧型手機不論在出貨量或市佔率早就遠遠超過桌上型電腦和可攜式電腦的總和，IDC 甚至預測平板電腦的出貨量將在 2015 年超過桌上型電腦和可攜式電腦的總和。這顯示了行動裝置的普及是一種不可抵擋的趨勢。

行動裝置的普及造就大量地域性查詢的需求，其中最常見的一種查詢，就是尋找附近的餐廳或加油站，根據 Google 於 2013 年第一季台灣智慧型手機使用行為調查(多選題)，搜尋內容依序為產品資訊（60%）、餐廳、酒館和酒吧（51%）、旅遊（49%）、工作機會（29%）以及購屋、租屋資訊（28%）。然而當使用者在電子地圖上搜尋這些地點名稱 (POI，Point of Interest)時，經常無法找到，因為電子地圖上雖有地點名稱標註，但是相關資訊不足，而這些資訊其實大多可以在網頁中找到。因此使用者大多必須開啟瀏覽器搜尋商家名稱找出地址，並把地址輸入至電子地圖查詢路線。但行動裝置螢幕小，且輸入文字不便利，如果要反覆查詢將是一件耗時耗力的工作。如果這時候有一個商家地理資訊系統能事先將網路上的商家資訊進行整合，最後提供一個 APP 直接讓使用者查詢，將可以大幅度減少使用者與裝置間的互動次數，有效的提供搜尋的便利性。

為建構商家地理資料庫，Chuang 等人(Chuang *et al.*, 2014)對於包含地址網頁提出以廣度優先搜尋、黃頁爬蟲、與地址樣版查詢三種抓取程式，並利用 Chang 等人(Chang *et al.*, 2012)的地址擷取程式，取得大量中文地址。Li 與 Chang(2009)並定義地址相關資訊擷取問題，希望藉此豐富每個 POI 的相關資訊，提高地理資訊檢索(Geographical Information

Retrieval, GIR)的召回率。然而不論是 Li 與 Chang(2009)或 Chang 等人(Chang *et al.*, 2012)或 Chuang 等人(Chuang *et al.*, 2014)的相關資訊擷取方法都僅能從多筆地址網頁擷取資訊，所得資訊有限，對於單筆地址網頁的相關資訊擷取仍尚無研究。

本研究從地址擷取的角度出發做為商家辨識的標記，利用已抓取大量包含地址的網頁，先找出網頁中的地址，再藉由地址找出對應的商家名稱進行配對。換言之，給定一個已知地址，我們希望能透過網路資料擷取出該地點的名稱（如：商家名稱、政府單位…等）。舉例而言：當我們已有地址「新北市板橋區中山路二段 88 號 3F」，我們希望能知道這個地址對應的名稱「大鈞醫學美容診所」，如此即可進一步藉由地址、名稱並利用搜尋引擎收集更多額外商家資訊。這些額外資訊不僅可以有效提昇地圖上搜尋也就是地理檢索系統 GIS 的召回率，也可提昇商家分類的準確率(陳宜勤 等，2013)。

在辨識商家名稱的部分，本篇論文使用了條件隨機域 (Conditional Random Field)當作學習演算法。目前有許多關於中文組織名稱辨認的研究 (Zhang *et al.*, 2007) (Yao, 2o11) (Ling *et al.*, 2012) (Wu *et al.*, 2008)，可以從新聞或一些較正式的文章中萃取出組織名稱，但是並沒有嘗試以一個CRF-Model直接對各種網站中的整個網頁內容進行中文組織名稱辨認。這兩者之間不同處在於新聞類文章屬於較正式的文章體裁，因此容易出現行政機關與正式的組織名稱，例如：行政院和維德食品有限公司，但是整個網路上商家組織名稱的命名方式傾向則不同，例如：吼牛排、努哇克咖啡、阿嬤祖傳菜包肉粽仙草…等，都是商家組織名稱。另外，一個完整的網頁內容有結構與非結構化的資訊交錯呈現，雖然結構化資訊會造成自然語言文字內容的破碎，但這些結構也隱含有可利用的資訊。

為了使商家辨識能以最少人力進行自動化學習，本研究使用自動標記方式建立訓練資料，我們先針對部份的黃頁網站(如 104 求職網、愛評網、工商名錄網站)撰寫 Parser 取得大量商家名稱與地址的組合，並以這些已經取得的商家名稱對網頁語料進行自動標記，再利用自動標記後的語料訓練 CRF 序列標記模型。然而一個地址可能出現在多個網頁之中，僅只仰賴其中一個網頁也有失之偏頗之慮，因此我們也收集了 Google Snippets 當作訓練資料進行比較。本篇論文的第二個主題則是商家地址的配對，由於一個網頁可能包含多個商家名稱，我們對網頁以簡單的規則進行分類後，使用了啟發式（heuristic）的配對規則，利用各類型的網站所具有的表達特性，對地址與商家名稱進行配對。

本研究承續 (Su, 2012) (Chuang *et al.*, 2014)之研究，經由爬取網頁上包含地址的大量網頁（包括 Yellow Page 與 Surface Web）進行商家名稱擷取。其中 Yellow Page 提供了大量商家名稱以及地址與商家的配對資料，而 Surface Web 則利用 (Chang *et al.*, 2012)之地址擷取模型擷取出了可能含有台灣地址的網頁與地址清單。本篇論文以已知可能含有台灣地址的中文網頁、每筆網頁的地址清單、大量商家名稱清單以及已知的地址與商家名稱配對資料為基礎，提出了一個商家名稱擷取系統，方法分為三大步驟：地址網頁的前處理、商家名稱命名實體辨認、及地址－商家名稱匹配。本研究在三個模型聯合標記商家名稱的方式下，地址與商家名稱的平均配對正確率為 0.57。

本論文共有五個章節，第一節是緒論，說明研究動機與背景；第二節是相關研究，

介紹中文組織名稱辨認和地址相關資訊擷取的相關研究。第三節是方法，會詳細介紹如何對地址-網頁分類、中文組織名稱辨認以及地址與商家組織名稱的配對。第四節是我們針對現有的網頁中，依據我們的分類，每類隨機抽取網頁進行的實驗與結果分析。最後是我們的結論以及未來的展望。

2. 相關研究

擷取地址相關資訊牽涉到三個領域，資訊擷取（Information Extraction）、自然語言處理（Natural Language Processing）與資訊檢索（Information Retrieval）。這三者彼此間互相交錯，很難精確切割出各自所屬的範疇。大致上來說，資訊擷取主要是從各種結構化資料與非結構化文字萃取出特定資訊的方法，而自然語言處理則屬於人工智慧領域的一個分支，目的在於自動化的理解並處理人類所使用的語言。資訊檢索則是從大量資料中以機率統計模型對資料進行排序（rank）、建立索引，快速找出使用者目標文件的方法。

本研究相關的主要技術，分別為如何有效爬取包含地址之目標網頁、地址相關資訊擷取與命名實體辨認。地址相關資訊擷取是在得知地址資訊後，從含有地址的網頁中擷取出與該地址相關的資訊，如：電話、網址、電子郵件、評論…等資訊。命名實體辨認則是為了辨認文句所提到的特定種類概念，如：人名、地名、組織名稱。本章中將依序介紹這些技術的相關研究。

2.1 包含地址的網頁抓取與地理資訊檢索

這裡所謂的地理資訊檢索，是從網路上爬取包含地點或地址的網頁，萃取地理資訊並利用此資訊排序與建立索引，提供快速檢索的服務。目前的搜尋引擎，像是 Google 和 Yahoo 也分別從 2005 與 2002 年開始提供電子地圖的服務。而這些服務需要藉由使用者的標記等群眾外包的方式建立 POI 資訊。(Dirk & Susanne , 2007)等人提出了一個以位置資訊為基礎的搜尋引擎，可以自動從網路資源中取得與空間相關的文句，而在他們最近的研究中 (Ahlers, 2013a; 2013b)，則專注在如何從深度網頁例如黃頁與 Wikipedia 擷取出位置命名實體。由於地址是 POI 的明確指標，因此 Chuang 等人(Chuang *et al.*, 2014)提出以廣度優先搜尋、黃頁爬蟲與地址樣版查詢三種策略爬取含有地址的網頁。實驗結果顯示雖然爬取黃頁網頁可以較快取得大量地址，然而地址樣版查詢可以補足黃頁涵蓋度不足之處，也是建立商家查詢服務不可或缺的方法。

2.2 地址與相關資訊擷取

地址擷取是因應地址資訊檢索所產生的需求，目的是從網路上大量的網頁中，擷取取出地址資訊，在 2009 年 Li 的研究中 (Li, 2009)，Li 以序列標記（Sequence Labeling）和 CRF 模型對美國地區的英文地址進行訓練與測試。Li 利用該地區地址的特性建立了 14 種特徵，並使用 BIEO 標記法，實驗結果 F-measure 達到了 0.913 的準確率。2011 年 Huang 延續了 Li 的研究 (Chang *et al.*, 2012)，利用 17 種台灣地址特徵和 BIEO 及 IO 兩種標記法，其中 IO 標記法因為邊界偵測能力較弱，需搭配極大分數子序列（Maximal Scoring

Subsequence）進行修正。BIEO 標記法的實驗結果 F-measure 約在 0.96 至 0.99 之間，IO 標記法則在 0.94 至 0.96 之間。

相關資訊擷取是地址擷取的延伸研究，目的是針對已知的地址擷取出與該地址有關的訊息，如：電話、網址、電子郵件、評論⋯等資訊。主要的作法是針對已經成功擷取出的地址，找出可能的上下邊界、劃出資料範圍作為該地址的相關描述，可以視為一種深度網頁資料擷取（Deep Web Data Record Extraction）的一種特例。在 Li 的研究中，主要是把所有地址所在的文字葉節點（Text Leaf Node）當作起點，利用這些節點走訪至根節點過程中，Html Tag 的變化當作邊界點。但是 Li 的方法對於網頁中擁有兩種以上的地址相關資訊排版無法有效擷取，為了解決此問題，Huang 會先針對各地址路徑的相似度作出分類，再針對各類實行 Li 的方法。在最後英文地址相關資訊擷取的實驗中，Li 的相關資訊擷取的 F-measure 達到了 0.8689，而加入了 Huang 的改進則提昇 0.0233。

2012 年 Su (Su, 2012)發現 Li 與 Huang 的做法過度簡化各筆紀錄（Record）的產生模版（Template），Li 與 Huang 的做法中，只要模版中有任何一筆選擇性資料（Optional Data），就會發生連鎖錯誤。為了解決此問題，Su 將 2010 年 Wei Liu 所提出基於視覺（Vision-Based）的資料紀錄（Data Record）擷取演算法套用在地址相關資訊擷取的研究中，並重作 Li 的實驗，將 F-measure 由 0.7912 提昇至 0.9504。

Li、Huang 和 Su 的研究皆專注於資訊擷取的效果上，但其前提是網頁中存在多個地址字串。若提及地址相關資訊的網頁內不存在地址字串，則無法得知網頁內含與 POI 相關的資訊，更不可能有後續萃取資訊的過程。因此，本研究試圖擷取出地址的商家組織名稱，以利後續的相關資訊萃取與檢索。

2.3 中文組織命名實體辨認

命名實體辨認屬於資訊萃取與自然語言的一個共同分支，此研究起因於任何系統皆無法窮舉出所有的詞彙與代表的意義，因為再大的詞庫都會有沒收錄的詞彙（OOV word，Out-of-Vocabulary），且同樣的詞彙在不同的內容中很可能代表不同的意義。目前的主要方法是利用序列標記配合機率統計模型計算出最可能的標記。

目前已經有許多中文組織名稱辨認的研究 (Zhang *et al.*, 2007) (Yao, 2011) (Ling *et al.*, 2012)，2007 年 Zhang 等人以人民日報 的新聞當作訓練資料，將數個 CRF 模型串連起來進行辨識，採用的特徵有：是否為前級輸出的各種命名實體（is Named-Entity）、常見的組織名稱開頭、內容與結尾、N 元文法（N-gram）。在 Zhang 所做的實驗中，F-measure 達到 0.9794。

2011 年 Yao (2011)則是將中文組織名稱分為三段：前置詞（Prefix words）+中間詞（middle words）+記號詞（mark words）（例如：中國+移動通訊+公司）且不採用現有的模型，使用自行設計的統計方法，考慮組織名稱的頻率、詞性與長度，配合以下假設進行計算：「記號詞能完全收錄」、「前置詞與中間詞為名詞、形容詞、序數或位置⋯等」、「記號詞大部分為名詞」和「組織名稱小於等於 10 個字」，最後的實驗使用了人

民網的語料進行訓練，以人民網、新華網 和北京郵電大學網站首頁的新聞 當作測試資料。平均準確率最高達到 0.959，平均召回值則達到 0.8724，皆超過隱藏馬可夫模型（HMM）與最大熵模型（ME）。

2012 年 Ling 等人 (Ling *et al.*, 2012) 以規則式的辨認方法（Rule-based Named-Entity Recognition）辨識人民日報與新浪網的新聞，Ling 首先將語料經過斷詞並將中文組織名稱拆解為多個修飾詞（Modifiers）+核心特徵詞（Core Feature Word）。在統計訓練資料後，找出常用的核心特徵詞，建立核心特徵詞庫當作組織名稱的結尾，並找出 6 種左邊界特徵（left-border features）判斷組織名稱的起點。在取得組織名稱候選者之後，利用該系統的常見錯誤模式（Debugging Patterns）進行修正。最後的實驗結果顯示，Ling 的方法的 F-measure 最高達到了 0.8573。

然而上述研究皆著重新聞語料之命名實體擷取，對於非新聞文件的一般網頁擷取並未著墨。事實上網頁的自由度使得命名實體擷取相對較為困難，這也是本篇論文的挑戰之處。

3. 商家名稱擷取與地址配對系統

本研究承續 (Su, 2012) (Chuang *et al.*, 2014) 之研究以及 (Chang *et al.*, 2012) 之地址擷取系統，經由爬取網頁上大量含有地址的網頁（包括 Yellow Page 與 Surface Web）進行商家名稱擷取。我們從這些網頁中過濾出含有台灣地址的可用網頁，進行商家名稱擷取，之後利用網站的特性如清單網頁、深度資訊網頁、註腳網頁、及自由文字網頁等為每一個地址配對商家名稱。

3.1 商家名稱辨認

本研究試圖對網頁內容擷取出所有的商家名稱，這裡所指的商家名稱涵蓋了各種範圍：明確的興趣點（POI，Point of Interest）、實際的組織名稱和產品的廠商名稱。目前在命名實體辨認的領域，通常使用序列標記法（Sequence Labeling）透過條件隨機域（CRF）模型進行辨認，然而監督式學習需仰賴大量的訓練資料，為減少人工標記的負荷，本文利用已知的商家名稱對網頁內容進行自動標記，並以標記後的網頁文字當作 CRF 的訓練資料。當 CRF 訓練完畢後，即可對網頁內容進行商家名稱辨識，建立商家名稱清單。下面將分別介紹本研究的自動標記、以及訓練資料的準備方式。

藉由 Web 上的黃頁網站所提供的商家資訊，我們可以取得「地址-商家名稱對」清單，對訓練網頁進行自動標記。然而由於網頁總數達 39.6 萬筆，而不重複的商家名稱總數高達 68.8 萬，基於執行時間的考量，無法對所有的網頁的每個句子都檢查是否存在已知商家名稱。因此，我們以每筆網頁已知的地址清單來加快標記速度：也就是說，系統只會依據網頁所擁有的地址查詢對應的商家名稱，並對網頁內容掃描這些對應的商家名稱是否存在，若存在就會以特殊的標籤（Tag）來標註這些商家名稱。圖 1 左圖即是自動標記自動產生訓練資料的流程圖。

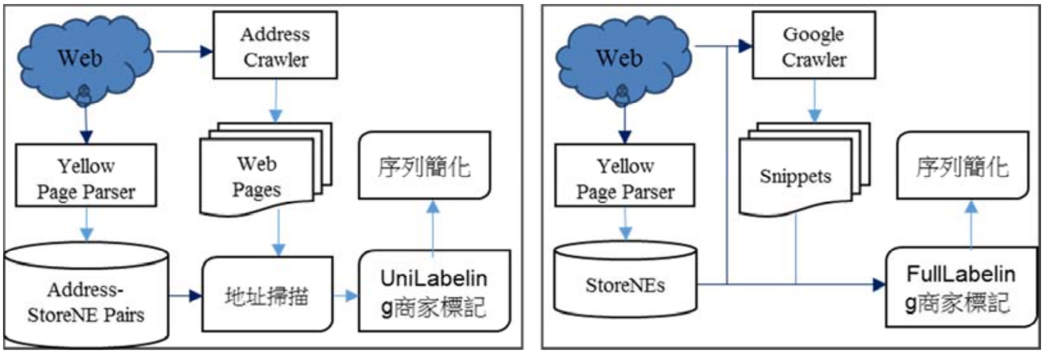


圖 1. 個別完整網頁的自動標記流程(左)與 Snippets 自動標記流程(右)

本研究另外以商家名稱當作關鍵字收集 Google 搜尋引擎提供的 20 筆 Snippets，並以所有的已知商家名稱對這些 Snippets 中的句子進行標記，試圖降低個別網頁資料的複雜度與標記不完整的問題。如圖一右圖所示，以 Google Snippets 為資料來源的處理流程，主要差異點在自動標記不僅只用單一的商家名稱來協助標記（稱之為 UniLabeling），而是採用所用商家名稱來進行標記（稱之為 FullLabeling），其餘皆與以整個網頁為資料來源的處理方式相同。訓練資料處理流程如下所述：

- 前處理

針對每一個原始網頁後，本系統首先使用 Apache Tika™ (Apache License, 2004)將網頁內容連同標題擷取成文字內容後才進行後續步驟。為了使序列單元（Tokens）特徵的強度增強，系統會先將所有全形符號轉換成半形符號，圓弧型的括號「(、（、﹙」統一轉成「(」，因為此種括號通常含有補充說明的意義。非圓弧型的括號「[、{、〔、｛、〈、…」統一轉成「[」，因為此種括號通常具有強調的意思。第二步是將換行符號、地址電話、時間…等以正規表示法取代成特殊的序列單元，這些取代動作能有效加強邊界特徵，縮短序列的長度，提昇辨識效果。

- 樣本序列

完整的網頁內容與一般的文章相比，不同的地方在於網頁會利用結構化資訊、排版等表達方式將文字內容傳達給使用者，因此很少有完整的句子，而是直接把項目、名稱、屬性⋯等資訊以列表或依序列出等方式呈現。若我們採取傳統的句子樣本單元（Training or Testing Examples），進行訓練與測試，很難有好的成果。因此我們將網頁內容轉成文字後，移除空白類字元、以連續三個換行符號當作分隔符號（Delimiter），將文字切為許多區塊（Block），以含有商家名稱的區塊加上前後區塊，以連續三區塊為一個訓練樣本，這樣的好處是盡可能讓訓練樣本涵蓋商家名稱，也能有較多的非商家名稱範例。同樣地在測試時，也採用三行文字為一個單位當作樣本單元進行測試。

- 序列單元與標記

一般說來，在人名辨識中，雖然人名有大量的組合與可能性，但是依然會有所謂的常用字，「菜市場名」就是一種很好的例子。但是商家組織名稱中除了結尾部份的常用詞外，在主要名稱上幾乎沒有任何規範，例如：「土地」、「阿嬤祖傳菜包肉粽仙草」中所有詞皆為常用詞彙，「18 度 c 巧克力工坊」、「591 租屋」為中英數字元交錯出現，「努哇克咖啡」、「蕾克爾烘培坊」為音譯詞，「蘆薈花園雲南食府」、「三峽歷史文物館」為地名。僅管如此，這些商家組織名稱的詞性卻有常見序列，如名詞+名詞或動詞、專有名詞+名詞或動詞、數字或英文+名詞或動詞 …等，所以詞性是一種不可忽略的重要特徵。因此在序列單元（Tokens）的選擇上，我們利用 Stanford Segmenter 及 POS Tagger 將網頁的文字內容經過斷詞及詞性（POS，Part of Speech）標記，以詞為單位進行訓練與測試。經過斷詞的序列，再以 B、I、E、O 四種標記代表商家名稱的起始、中間、結尾、以及非商家名稱。

- 特徵

一般人在判斷一段文字是否是商家名稱時，會依靠兩類特徵，第一種是外部特徵（Outside Feature），這種特徵落在商家名稱的左右，但是此種特徵無法準確判斷商家名稱，只能進行推測上的輔助。第二種則是內部特徵（Inside Feature），內部特徵能提供強烈的判斷資訊，因為絕大多數的商家名稱都是由三個部份所組成：真名（Real Name）、產品或服務（Service or Product）、地標性詞彙（Landmark），舉例來說：「燦坤 3C 量販店」可以拆成「燦坤/3C/量販店」或「燦坤/3C 量販/店」。即使是非常短的商家名稱都會有這種結構，例如：「麗嬰房」是「麗/嬰/房」，其中嬰是指提供兒童用品。

我們統計已知的商家名稱，對組織、建築、房間、地標建立清單，例如：會、城、房、站⋯等，當每個序列單元（Tokens）是以此清單中的文字為結尾，就表示具有地標性詞彙的特徵（Landmark Feature）。另外，我們也收集了黃頁網站的服務、產品建立清單，如果序列單元（Tokens）含有此清單中的詞彙，就表示具有產品服務特徵（Service/Product Feature）。

當我們有了上述兩種特徵，問題就簡化成如何找出真名（Real Name）的部份，網頁內容與一般文章不同的地方在於名稱更傾向於單獨出現，而鮮少存在於一段完整的句子中 ，所以一段文字意思的起點就變成很重要的特徵：如果一個序列單元（Tokens）是樣本單元的起點或前一個序列單元屬於符號類，就具有開始特徵（Start Feature），反過來說，當序列單元是樣本單元的結尾或下一個序列單元屬於符號類，就具有結尾特徵（End Feature）。例如：網頁中的標語「[阿嬤祖傳菜包肉粽仙草]有阿嬤的精神傳承製作出客家傳統米食好滋味!」與網頁標題「阿嬤祖傳菜包肉粽仙草」中，前者的「阿嬤」的前一序列單元為符號，後者為序列單元的起點所以皆具有開始特徵，「仙草」則具有結尾特徵。系統最後選擇對商家名稱具有強烈判斷資訊的內部特徵加入訓練模型，所有原始特徵列於表 1。

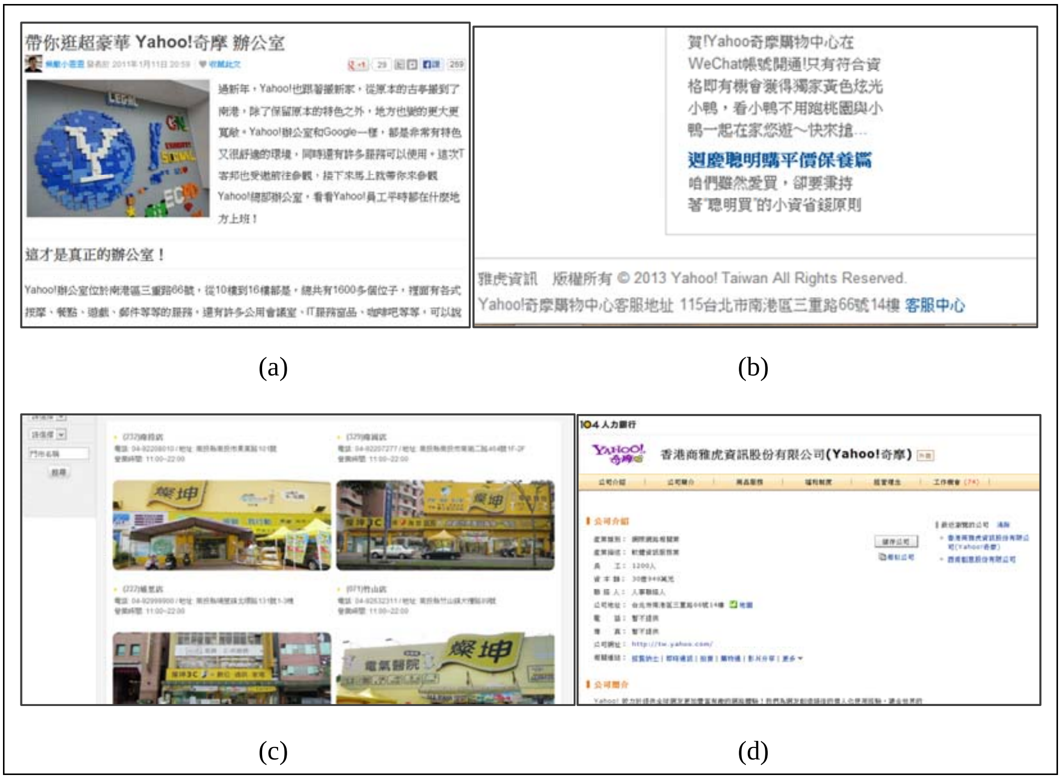
表 1. 本研究所使用的原始特徵

NO.	Feature	Explanation
1	Token	個別詞 Individual Word, e.g. 591, 租屋
2	isPOS	詞性 Part of Speech, e.g. NR, NN, CD
3	isStart	樣本序列開頭 或 短語開頭
4	isSymbol	屬於符號詞, e.g. (, [, breakline, !, :
5	isService/Product	屬於服務/產品詞, e.g. 3C, 壽司, 出租, 通信
6	isLandmark	屬於地標詞, e.g. 廟, 莊, 公司, 店
7	isEnd	樣本序列結尾 或 短語結尾

3.2 地址-商家名稱匹配

當我們有了地址與商家名稱後，便可以開始進行配對。由於各類別的網頁特性差異很大，所以系統會針對各類別設計各自的啟發式（heuristic）的配對方式。首先我們依照網站將網頁分成不同群組，接著依網站中的地址資訊將網頁分成四類。

- 自然語言網頁：當地址字串所在的文字節點（Text Node）有超過 50 個字就會歸類至自然語言網頁（請參考圖 2），因為會這個長度相當於一小塊片段文字（Snippet）。位於此種網頁的商家名稱左右大多接有能意會到該處為商家名稱的訊息，例如：「走進 edia cafa 店裡一眼望去」、「我昨天去了燦坤 3C 買東西」。這也是網頁中唯一接近一般文章的類別。通常具有外部特徵 （Outside Feature）。
- 註腳資訊網頁：當一個「網站」內超過 80%的網頁都有相同的地址與文件物件樹路徑（DOM Tree Path），這些地址就會歸類至註腳資訊網頁。此類別中的所有網站，商家名稱周圍的文字資訊都有很高的相似度，經常會有：「本網站為…」「…版權所有」、®、©、地址、電話…，這些資訊在 N 元文法（N-Gram）的特徵上，能提供有用的資訊。
- 清單網頁：當一個網頁內包含超過 3 筆地址有相同的文件物件樹路徑（DOM Tree Path），這些地址就歸類為清單類型。清單型的商家名稱雖然不像自然語言網頁中，商家名稱的左右具有描述性的文字，但取而代之的是周圍具有換行符號、電話、地址、時間等資訊，藉由事先用正規表示法取代這些字串後，亦能利用 N-Gram 取得此特性。
- 深度資訊網頁(Detail Pages)：當一個網站內不同網頁的地址有相同的文件物件樹路徑（DOM Tree Path），但是地址字串卻不相同，這些地址就歸類為深度資訊網頁，當我們從多個網頁來看時，地址和商家名稱通常擁有同樣的文件物件樹路徑（DOM Tree Path），我們可以透過此特性進行商家名稱的修正。



帶你逛超豪華 Yahoo!奇摩 辦公室

過新年，Yahoo!也跟著搬新家，從原本的古亭搬到了南港，除了保留原本的特色之外，地方也變的更大更寬敞。Yahoo!辦公室和Google一樣，都是非常有特色又很舒適的環境，同時還有許多服務可以使用。這次T客邦也受邀前往參觀，接下來馬上就帶你來參觀Yahoo!總部辦公室，看看Yahoo!員工平時都在什麼地方上班！

這才是真正的辦公室！

Yahoo!辦公室位於南港區三重路66號，從10樓到16樓都是，總共有1600多個位子，裡面有各式按摩、餐點、遊戲、郵件等等的服務，還有許多公用會議室、IT服務當品、咖啡吧等等，可以說

(a)

賀!Yahoo奇摩購物中心在WeChat帳號開通!只有符合資格即有機會獲得獨家黃色炫光小鴨，看小鴨不用跑桃園與小鴨一起在家悠遊～快來搶...

週慶聰明購平價保養篇

咱們雖然愛買，卻要秉持著"聰明買"的小資省錢原則

雅虎資訊　版權所有 © 2013 Yahoo! Taiwan All Rights Reserved.

Yahoo!奇摩購物中心客服地址 115台北市南港區三重路66號14樓 客服中心

(b)

(c)

104 人力銀行

香港商雅虎資訊股份有限公司(Yahoo!奇摩)

公司介紹

員　　工：1200人

公司網址：http://tw.yahoo.com/

公司簡介

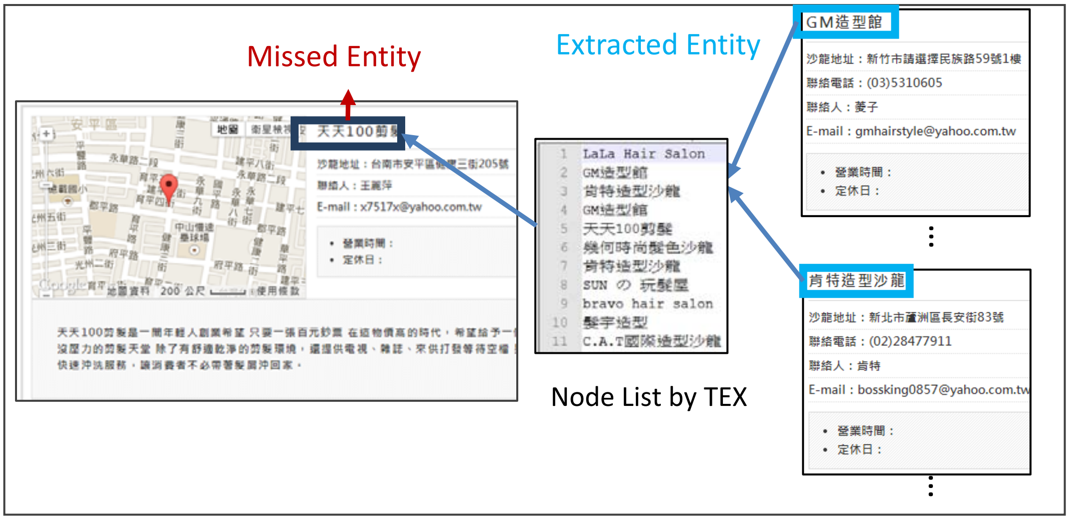
(d)

圖 2. (a)　自然語言網頁範例　(b)　註腳網頁範例　(c)　清單網頁範例　(d)深度網頁範例

對於第一和第二類網頁而言，地址所對應的商家名稱通常落在：網頁標題、地址前、地址後或高頻商家名稱。若只有一個地址，則第一順位是網頁標題中的商家名稱。其次，以靠近地址的商家名稱為優先配對對象，「地址前」的配對方式是將地址與所在位置的前五行內的商家名稱列為配對候選者，而「地址後」則是將地址與所在位置的後兩行內的商家名稱列為配對候選者，當多個候選者距離相同時，會以網頁中出現較多次的商家名稱為優先，若次數完全相同則選擇位於地址前方的商家名稱。

至於第三和第四類網頁，因為網頁通常由模板（Template）和紀錄（Record）所組成，而相同類型的紀錄會放置在類似路徑下，所以存在一個專門的研究領域稱為 Wrapper Induction，目的是透過參考一個或多個網頁內容反向推導出模板與紀錄。本研究中使用了 TEX 作為輔助工具(Hassan & Sleiman, 2013)，TEX 是一個 Deep Web Crawling Tool，可以將多個網頁的原始檔文字內容當作輸入（作者稱為 TextSet），透過尋找各文件所擁有的共享樣式（Shared Pattern）當作紀錄的分隔點，經過反覆尋找共享樣式與切割後，找出最後的資料節點。藉由 TEX (Hassan & Sleiman, 2013) 擷取出網頁中具有同性質的資料節點，當有一定數量的同類節點被認為是商家名稱且商家名稱長度佔節點內容的 20%以上時，則把同類的非商家名稱節點也視為商家名稱進行配對。舉例而言，圖 2 中的「天天 100 剪髮」並沒有被 CRF 辨識出來，但是在同網站的其他網頁中，此節點的內容

「GM 造型館」、「肯特造型沙龍」…等已被成功辨識為商家名稱，所以系統也會將「天天 100 剪髮」視為商家名稱。本系統中，門檻值為 0.2，即該節點有 20%以上的內容被認為是商家名稱，則其餘網頁的該節點也會被認為是商家名稱。



天天100剪髮
沙龍地址：台南市安平區健康三街205號
聯絡人：王麗萍
E-mail：x7517x@yahoo.com.tw
- 營業時間：
- 定休日：

天天100剪髮是一間年輕人創業希望 只要一張百元鈔票 在這物價高的時代，希望給予一
沒壓力的剪髮天堂 除了有舒適乾淨的剪髮環境，還提供電視、雜誌、來供打發等待空檔
快速沖洗服務，讓消費者不必帶著髮屑沖回家。

```
1  LaLa Hair Salon
2  GM造型館
3  肯特造型沙龍
4  GM造型館
5  天天100剪髮
6  幾何時尚髮色沙龍
7  肯特造型沙龍
8  SUN の 玩髮屋
9  bravo hair salon
10 髮宇造型
11 C.A.T國際造型沙龍
```

GM造型館
沙龍地址：新竹市請選擇民族路59號1樓
聯絡電話：(03)5310605
聯絡人：菱子
E-mail：gmhairstyle@yahoo.com.tw
- 營業時間：
- 定休日：

肯特造型沙龍
沙龍地址：新北市蘆洲區長安街83號
聯絡電話：(02)28477911
聯絡人：肯特
E-mail：bossking0857@yahoo.com.tw
- 營業時間：
- 定休日：

圖 3. 深度資料網頁配對範例

當我們利用路徑找出所有可能的商家名稱後，將開始進行實際配對。清單型網頁與深度資訊網頁的配對方式大致相同：以每筆地址的上方全部內容與下方兩行內當作配對候選，離地址近的優先配對，當距離相同時，以地址前方的商家名稱為優先。但清單型網頁會以地址為界線，在挑選配對候選者時，不會越過地址進行配對。

另外當我們以地址為關鍵字收集 Google Snippet 後，這些 Snippets 中的網頁結構資訊較弱，但是可以同時參考大量與近期相關的網頁提高可信度，所以當我們以商家名稱的 Snippet 訓練出 CRF 模型後，就直接以某地址為關鍵字所得到的所有 Snippets 中，出現最多次的商家名稱和該地址進行配對。

4. 實驗

因為本研究是先進行商家名稱辨認，再將地址與已知的商家名稱進行配對，所以實驗部份也依照這兩個階段來進行。第一階段的實驗為商家名稱辨識率，資料來源有兩種，第一種是以[9]所取得的約 50 萬個可能含有地址的網頁當作原始資料，在經過前處理後，過濾出約 39 萬個含有台灣地址的網頁，經過地址正規化後含 19 萬筆台灣地址（請參考表 2）。在經過網頁分類後，我們隨機挑選各類中的 100 個網站，每個網站中各隨機抽取 1 個網頁進行實驗，但 Detail Pages 因為配對方法需參考多個網頁，所以隨機挑選了 11 個網站，每個網站抽取 10 個網頁。最後對這 410 個網頁人工標記了 10,457 個商家名稱當作測試資料。而訓練資料則隨機挑選了 30,000 個訓練樣本，包含 51,775 個以自動標記法標記的商家名稱。

表 2. 以個別完整網頁為資料來源的訓練語料與測試資料

	Training Corpus		Testing Data				
	Raw	Preprocessing	FreeText	Foot	Detail	List	Sum
# Sites	-	13,224	100	100	11	100	311
# Web Pages	508,038	396,093	100	100	110	100	410
# Addresses	272,987	190,180	219	156	467	807	1,649
# Stores	-	-	1,841	1,975	3,855	2,786	10,457

第二種資料來源是使用 Google 搜尋引擎所取得的網頁內容片段（Snippets，請參考表 3），在訓練資料的部份，我們以 11,138 筆商家名稱進行查詢，以自動標記的方式產生了兩種訓練資料：SnippetUniLabeling 和 SnippetFullLabeling，在 SnippetsUniLabeling 中，我們僅以關鍵字的商家名稱對 Snippets 中的句子進行標記，共標記了 222,121 個商家名稱，而 SnippetsFullLabeling 中，則是以所有已知的商家名稱對 Snippets 中所有句子進行標記，共標記了 390,113 個商家名稱，藉由不同的標記方式產生不同程度的雜訊，以了解雜訊對辨識率的影響。在測試資料的部份則以 6,963 筆地址為關鍵字，收集每筆地址排名前 20 的搜尋結果（Snippets），以自動標記的答案進行最後 NER 效能評估。最後再對兩類資料進行交叉測試。

表 3. 以 Search Snippets 為資料來源的訓練資料與測試資料

	Training Data		Testing Data	
	# of Store Queries	Tag Stores	# of Address Queries	Stores (Auto Labeling)
Snippet Uni Labeling	11,138	222,121	6,963	70,449
Snippet Full Labeling	11,138	390,113	6,963	70,449

第二階段為地址與商家名稱配對的正確率，針對不同資料來源以各自的方式進行配對，第一種是針對不同網頁類別以各自的啟發式（heuristic）規則進行配對，第二種是以 Snippets 中各商家名稱的最高出現次數進行配對。

標記比對的評估方式如下：雖然我們有明確訂出商家組織名稱的判定規則，但很多時候依然難以準確定出邊界標準，例如：「飯店名稱：西門星辰大飯店」中，「西門」二字該不該列入商家名稱中有許多意見分歧的情況，由於商家名稱主要提供後續的地理資訊檢索，因此系統標記結果若能包含正確答案(Gold)，我們即認定正確，若是僅為正確答案的部份，則給 0~1 之間的分數，並依此分數計算 Precision、Recall、F-measure：

$$Gold, SysTag\text{比對分數} = \begin{cases} 1, & if\ SysTag\text{包含}Gold \\ \dfrac{TagNELength}{GoldNELength}, & if\ Gold\text{包含}SysTag \end{cases}$$

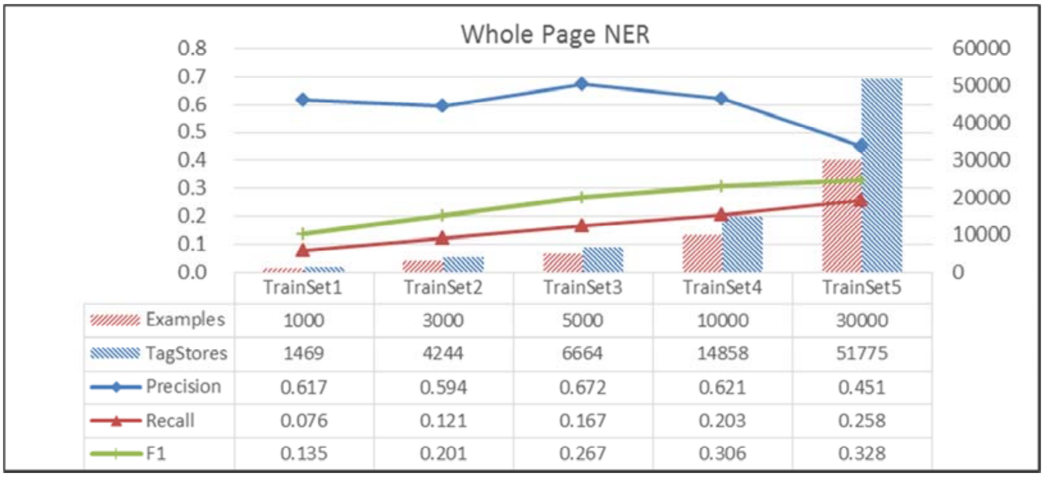
$$Precision = \frac{SysTag\text{辨識出的所有商家名稱與}Gold\text{進行比對的分數總和}}{SysTag\text{所辨識出的所有商家名稱數量}}$$

$$Recall = \frac{SysTag\text{辨識出的所有商家名稱與}Gold\text{進行比對的分數總和}}{\text{人工標記的所有商家名稱數量}}$$

此種評估方式，可以解決當 CRF 辨識出的商家名稱邊界包含地名、百年老店…等難以判定是否屬於商家名稱的一部分的問題。

4.1 商家名稱辨識率

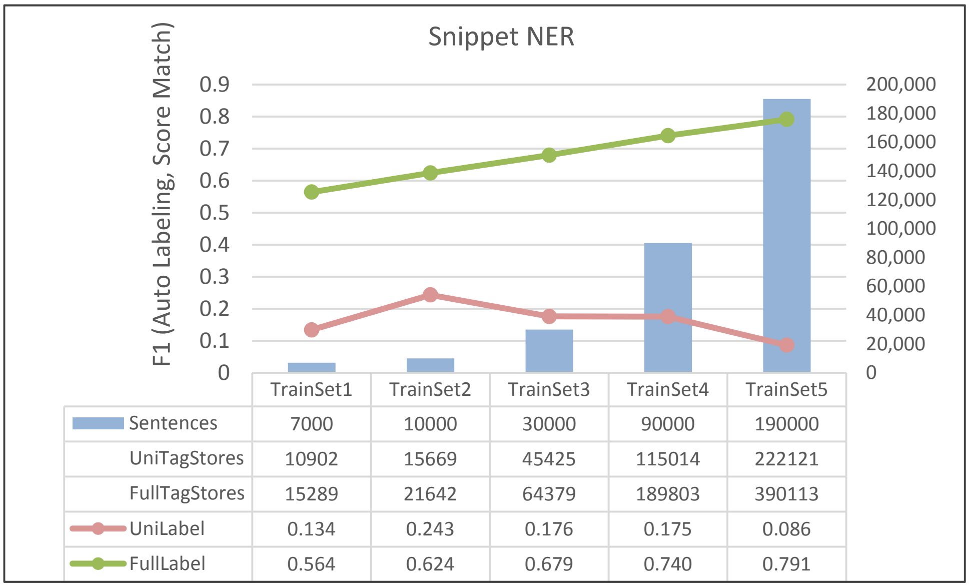
我們首先以個別完整網頁為資料來源，實驗了訓練資料數量對辨識效能的影響。接著我們以 Snippet 為資料來源，分別實驗了 Uni-Labeling 和 Full-Labeling 的效能，以了解在自動標記中，雜訊對辨識效能的影響，然後對兩種資料來源所訓練出的模型進行交叉測試，觀察不同來源的訓練資料所訓練出的模型，應用在不同測試資料時的表現。最後是本研究的在商家辨識部份的最後輸出。



	TrainSet1	TrainSet2	TrainSet3	TrainSet4	TrainSet5
▨ Examples	1000	3000	5000	10000	30000
▨ TagStores	1469	4244	6664	14858	51775
◆ Precision	0.617	0.594	0.672	0.621	0.451
▲ Recall	0.076	0.121	0.167	0.203	0.258
+ F1	0.135	0.201	0.267	0.306	0.328

圖 4. 完整網頁中，訓練資料數量對 F1 的影響

圖 4 是訓練資料數量對 Precision、Recall、F1 影響的趨勢圖，圖中顯示當訓練資料數量達到 30,000 樣本序列時，辨識效果依然只有 0.328，雖然 Recall 獲得提昇，但是 Precision 也較大幅的下降。主要的原因可能在於我們使用自動標記產生訓練資料時，並沒有使用所有已知的商家名稱進行標記，所以造成了大量標記錯誤（應為 B/I/E/S、卻標成 O）。因此在 Search Snippets 實驗中，我們嘗試探討降低語料複雜度與標記不完全兩個問題。



	TrainSet1	TrainSet2	TrainSet3	TrainSet4	TrainSet5
▬ Sentences	7000	10000	30000	90000	190000
UniTagStores	10902	15669	45425	115014	222121
FullTagStores	15289	21642	64379	189803	390113
●─ UniLabel	0.134	0.243	0.176	0.175	0.086
●─ FullLabel	0.564	0.624	0.679	0.740	0.791

圖 5. 以 Snippets 為資料來源，雜訊與訓練資料數量對效能的影響

在 Snippets 方面的實驗，我們測試了訓練資料數量與標記品質對辨識效能的影響，以了解在自動標記中，雜訊對辨識效能的影響。如圖 5 所示，在 UniLabeling 模型中，當資料增加時，訓練資料含有的雜訊(標記不完全)更為嚴重，使得效能下降；而 FullLabeling 模型因為使用所有的商家名稱進行標記，所以雜訊大幅減少，在資料增加的情況下可大幅度提昇效能，FullLabeling 模型的效能最高為 0.791。

不過在 Search Snippets 的測試資料中並非使用人工標記的答案進行驗證，而是使用自動標記的答案。為了了解使用某一語料所訓練出的模型是否能應用在另一不同語料的測試資料，我們對個別網頁與 Search Snippets 進行了交叉測試，我們以完整網頁為訓練資料所訓練出的模型對 Snippet 的測試資料進行測試，同時也以 Snippet 中兩種訓練資料所訓練出的模型對 410 個網頁進行測試。

實驗結果如表 4 所示，圖中顯示不論是何種測試資料類型，由 SnippetFullLabeling 所訓練出的模型都具有比較好的辨識效果，甚至比個別完整網頁所訓練出的模型用在測試同類資料還要高，可見在自動標記中，只使用部份已知的商家名稱所產生的訓練資料，並不是一個好的方式，會大幅度受到雜訊與樣本數限制的影響。

綜合以上實驗結果來看，我們認為影響辨識效能的主要的原因有三個：第一是因為商家組織名稱屬於變異性較大的一種命名實體，在訓練階段中，資料的準備能否盡可能的涵蓋各類商家組織名稱的特性。第二，網頁屬於一種結構複雜的資料來源，以此種資料來源再以自動標記進行訓練，可能造成訓練樣本的品質不佳，因此對商家名稱這種變化性極大的命名實體，較難辨識出正確的答案，因此需要更多的特徵與提昇標記品質。

表 4. 交叉測試

	Whole Page	Search Snippets
Whole Page Model	0.305	0.473
Snippets Model (Full Labeling)	0.310	0.791

第三，當我們利用已知的商家名稱進行標記時，這些已知資料可能存在不正確、不齊全或是歧義性等問題，造成自動標記的第一次錯誤，而且擁有大量的已知名稱和網頁時，無法對所有網頁中的所有字串都檢查是否存在商家組織名稱，只能利用地址查詢是否存在對應的商家名稱，造成第二次的錯誤，若要在合理的執行時間內解決此問題，可能需要使用 Hadoop 或是其他分散式系統，以所有已知的商家名稱進行標記以提昇標記品質。

4.2 地址-商家名稱 匹配正確率

圖 6 是 SnippetFullLabeling 以不同訓練資料所訓練出的模型對配對正確率的影響，圖中顯示當 NER 的效能大幅提高時，Match 雖然跟著上升，但僅有微幅成長。而在完整網頁為資料的實驗中，雖然我們無法辨認出所有的商家名稱，但經由啟發式（heuristic）的配對規則，可以提昇配對的正確率。圖 7 是以完整網頁為資料來源，地址-商家名稱配對正確率的實驗結果。以單一類別來看，在深度資訊網頁的實驗中，利用文件物件樹路徑的相似度後，可以將配對準確率提昇至 0.951，平均正確率則為 0.573。

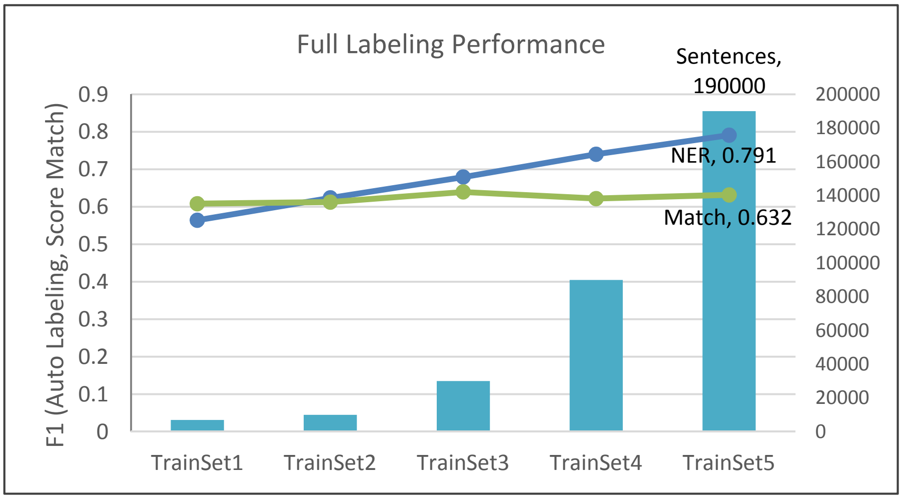


圖 6. SnippetFullLabeling 不同訓練資料數量的模型中，NER 對 Match 的影響

5. 結論

2014 是一個行動裝置的時代，大量的適地性服務（LBS）因此誕生，而 POI 資料庫在這波以行動裝置為主流的趨勢中具有至關重要的地位，建立一個完整的 POI 資料庫，可以讓使用者在地圖上提供更為便利的查詢。地址是 POI 的重要指標如果能找出地址所代表

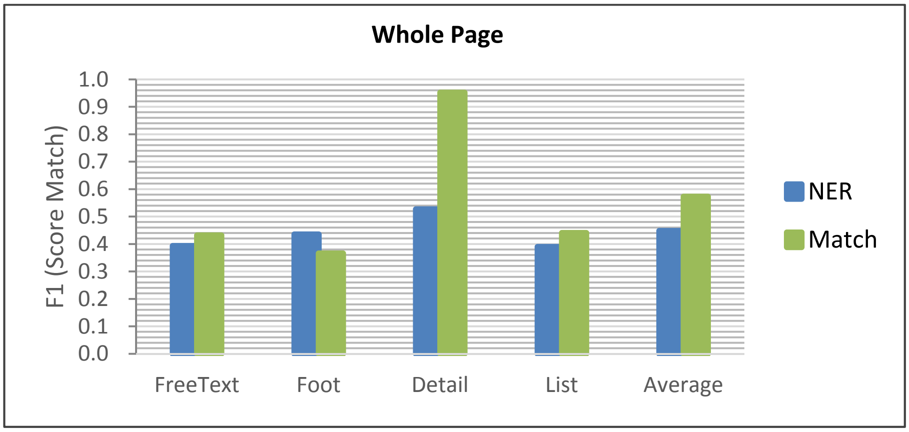


圖 7. 以完整網頁為資料來源的配對正確率（訓練樣本數：4,398）

的商家名稱，再以商家名稱當作搜尋引擎的關鍵字取得 POI 的相關資訊，就可以成功建立 POI 資料庫。

過去命名實體以新聞報導中的人名、地名、組織名擷取為主軸，目的在了解新聞中的事件，但對於網路上的興趣點 POI 的收集較少著墨。本研究試圖直接對整個網頁進行辨認，雖然受限於標記的不全，在命名實體辨認的效果並不好，但是在深度資訊網頁（也是含有最多地址的網頁類型）的地址-商家名稱配對中，利用網頁間的相似度可以取得 0.9514 的準確率，而平均正確率則為 0.5726。而 Google Snippets 的方法中，NER 效能最高為 0.791，配對正確率最高為 0.632。

在實驗過程中，我們發現啟發式的配對規則雖然可以提昇 Detail Pages 的配對正確率，但是其餘類型依然很仰賴命名實體的辨認結果。若要更進一步提昇商家名稱的辨識結果，我們覺得可以朝兩個方向進行，第一，必須將外部特徵加入特徵矩陣中，因為外部特徵雖然不能明確指出商家名稱，但是依然是進行推測的重要提示，在未來我們希望能把外部特徵和詞頻加入 CRF，提昇商家名稱的辨識效果。第三是利用分散式系統的速度，完整標記已知（大量）已知的商家名稱，解決自動標記產生的訓練資料品質不佳的問題。

參考文獻

Ahlers, D. (2013). Business entity retrieval and data provision for yellow pages by local search. *Integrating IR technologies for professional search, ECIR*, 2013.

Ahlers, D. (2013). Lo major de dos idiomas － cross-lingual linkage of geotagged Wikipedia articles. *Advances in Information Retrieval*, 2013, 668-671.

Apache Tika (2004). The Apache Software Foundation, [Online]. Available: http://tika.apache.org/.

Chang, C.-H., Huang, C.-Y., & Su, Y.-S. (2012). Chinese Postal Address and Associated Information Extraction. *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, 2012.

Chuang, H.-M., Chang, C.-H., & Kao, T.-Y. (2014). Effective Web Crawling for Chinese Addresses and Associated Information. in *EC-Web*, Munich, Germany, 2014.

Dirk, A., & Susanne, B. (2007). Location-based Web search. *Advanced Information and Knowledge Processing 2007*, 55-66.

GeoNames. [Online]. Available: http://www.geonames.org/.

Hassan, R. C., & Sleiman, A. (2013). TEX: An efficient and effective unsupervised Web information extractor. *Knowledge-Based Systems*, 2013, 109-123.

John, L. D., Andrew, M., & Fernando, N.C. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML Proceedings of the Eighteenth International Conference on Machine Learning*, 282-289.

Li, S.-Y. (2009). *Application and Extraction of Postal Addresses and Related Information*, National Central University, 2009.

Ling, Y., Yang, J., & L. He. (2012). Chinese Organization Name Recognition Based on Multiple Features. *Pacific Asia conference on Intelligence and Security Informatics, 7299*, 136-144.

Liu, W., Meng, X., & Meng, W. (2010). ViDE: A Vision-Based Approach for Deep Web Data Extraction. *Transactions on Knowledge and Data Engineering*, *22*(3), 447-460.

The Stanford NLP (Natural Language Processing) Group. Stanford NLP Group, [Online]. Available: http://nlp.stanford.edu/software/segmenter.shtml.

Su, Y.-S. (2012). *Associated Information Extraction for Enabling Entity Search on Electronic Map*, National Central University, 2012.

Wu, C.-W., Tsai, R. T.-H., & Hsu, W.-L. (2008). Semi-joint labeling for Chinese named entity recognition. In Proceedings of the 4th Asia information retrieval conference, *4993*, 107-116.

X. Yao. (2011). A Method of Chinese Organization Named Entity Recognition Based on Statistical Word Frequency, Part of Speech and Length. *Broadband Network and Multimedia Technology (IC-BNMT)*, 637-641.

S. Zhang, & X. Wang. (2007). Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields. *Natural Language Processing and Knowledge Engineering 2007*, 229-233.

教育部重編國語辭典修訂本－主站。中華民國教育部，[Online]. Available: http://dict.revised.moe.edu.tw/.

陳宜勤、賴郁婷、莊秀敏與張嘉惠(2013)。加入 Google Snippets 改善網頁商家多標籤分類, *The 18th Conference on Artificial Intelligence (TAAI 2013)*, 6-8.

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 19-28

Public Opinion Toward CSSTA: A Text Mining Approach

Yi-An Wu* **and Shu-Kai Hsieh***

Abstract

Extracting policy positions from the texts of social media becomes an important technique since instant responses of political news from the public can be revealed, and also one can predict the electoral behavior from this information. The recent highly-debated Cross-Strait Service Trade Agreement (CSSTA) provides large amounts of texts, giving us an opportunity to test people's stance by the text mining method. We use the keywords of each position to do the binary classification of the texts and count the score of how positive or negative attitudes toward CSSTA. We further do the trend analysis to show how the supporting rate fluctuates according to the events. This approach saves human labor of the traditional content analysis and increases the objectivity of the judgement standard.

Keywords: Policy Position, Opinion Mining, Politics, Social Media, Trend Analysis

1. Introduction

Deriving reliable estimates of public opinions is central to the study of electoral behavior and policy positions. Among different methods, linguistic strategy has been one of the most widely used approaches in related studies in the field of political communication. For instance, budge *et al*. (1987) utilizes discourse-level opinion interpretation and stance recognition; while laver *et al*. (2003) and klemmensen *et al*. (2007) treated the words as “data”' encoding information about the political position of the texts' author. In addition to theoretical surveys, there are also numerous appealing applications on the political positions such as *abgeordnetenwatch*[1], where citizens are able to ask the members of parliament questions and express their attitudes through surveys, and the members of parliaments also respond to the questions. The dynamic design often attracts large organizations and political parties to keep a close eye on how the public form and represent their political stance, thus enhancing the

* Graduate Institute of Linguistics, National Taiwan University
E-mail:

[1] http://www.abgeordnetenwatch.de

transparency and accountability in the development of democracy.

Over the past few years, the production of huge volume of textual data has become an essential part of our current social life. In this context, there have been growing interests in applying text mining techniques to support Natural Language Processing applications in social and political domains, ranging from subjectivity and opinion mining, to ontologies and knowledge discovery. More and more attentions have been paid to the analysis and prediction tasks from the social media (Tumasjan *et al.*, 2010; Conover *et al.*, 2011; Bermingham & Smeaton, 2011), which set a new scene for the data-driven research paradigm for social and political domains.

Recently, the public of Taiwan has had a heated debate on the issue of Cross-Strait Service Trade Agreement (CSSTA). After months of simmering tensions between ruling party and opposition party strongly backed by the student-led Sunflower Movement, the debate has finally reached a breaking point on March 18, 2014, at which students occupied the Legislative Yuan. This action of "Occupy Taiwan Legislature" marked the beginning of a series of different political negotiations and efforts on this topic for both sides till April, 7. During this period, as well-recognized by many, using of novel communication technologies - Facebook sharing, instant messaging, sparking discussions on PTT, cloud documentation, etc - have reshaped the social movement not only domestically, but also globally.[2]

The uncertainty among members of society over the implementation of CSSTA is palpable. Due to its nature of easy access and instant response, the social media has become the dominant source in opinion shaping and the accompanying sentiment spread. The extraction and tracking of uprising political opinions and events has thus become one of the most important topics that must be now be reckoned with. Though the task of analyzing and interpreting the social and political texts has gained its popularity in NLP-aided Social Science related fields, with the huge amounts of texts, it is not possible to analyze them manually. Instead, we propose to use the text mining approach, which automatically extract opinion and information profiles from the texts. In addition, this approach also strengthens the objectivity, for the norms are set *a priori*, thus human bias is reduced.

Our work is motivated by the compelling study of Junqué de Fortuny *et al*. (2012) which analyzed political opinions in Belgium by text mining of the newspapers. They used sentiment analysis to detect the opinion of the texts, and found the trends over timeline. Gelbukh *et al*. (1999) also used text mining techniques to analyze the Internet and newspaper news. They extracted the information of the texts by three steps: finding the topic of the document,

[2] Interested readers can refer to the cloud folder at http://hackfoldr.org/congressoccupied/ and the popular forum at http://www.reddit.com/r/IAmA/comments/21xsaz/we_are_students_that_have_taken_over_taiwans

extracting the opinion paragraphs by pattern matching, and matching topics with opinion paragraphs. They intended to discover how society interests are changing and to identify important current topics of opinion.

As a pioneering work in the context of Taiwan society, this research aims to trace the public opinion toward CSSTA from the perspective of text mining. The approach involves the manually extracting of *political stance* related keywords and phrases, supervised machining learning, and a statistical model of the trend. We focus on the individual posts on PTT rather than news since they are more representative. The potential political or commercial applications are valuable. One can discover the public opinion and response in a short time.

This paper is organized as follows: first, we introduce some backgrounds of the studies of policy positions in section 2. Our approach to this topic and also the materials we used is described in section 3. The validity of our approach and the results are shown in section 4. Section 5 concludes the paper and suggests future works.

2. Previous Works

There is a growing body of studies on the topic of analysis policy positions. One traditional approach is content analysis, such as the Comparative Manifestos Project (CMP) (Budge *et al.*, 1987; Benoit & Laver, 2007; Slapin & Proksch, 2008), where thousands of manifestos over 50 countries are interpreted by human decoders. However, this approach is so costly that it requires a huge amount of human labor. Another approach is computerized coding schemes (Kleinnijenhuis & Pennings, 2001), which match the texts to coding dictionaries. Laver and Garry (2000) created a dictionary of policy position which contains the predefined categories of political issues and the corresponding words. However, the approach also require much human labor on building dictionaries, and the words are insensitive to the contexts.

A variant of the second approach is the research of Laver *et al*. (2003), where they compared words in two different types of texts. One is the reference texts whose policy positions are defined *a priori*, and the other is the virgin texts whose policy positions are unknown but need to be found out. This approach is similar to the conventional **keyness** calculation where the *salient* keywords in target texts are measured and weighted statistically in comparing with the reference texts. However, as mentioned in (Klemmensen *et al.*, 2007), the validity of the positions obtained by the this approach is “dependent on the choice of reference text and the quality of the a priori scores attached to these reference texts.” This poses a challenge for us because of the lack of representative reference corpus that can reflect the current language usage.[3] In this study, we adopt the second approach with a little variation, i.e. we also built the dictionary and tested its validity. More detailed procedures are explained

[3] Note that Sinica Corpus had ceased to update around 17 years ago.

in the next section.

3. Methodology

3.1 Materials

The material we used in this experiment includes a list of manually created seed words and phrases representing the pro-and-con political polarity, respectively. 8 linguistic graduate students from NTU were asked to compile the list based on their observations on the texts with CSSTA debate. It is noted that the keywords may be a word, a phrase, or a sentence. After some preprocessing, there are in total 350 terms for supporting CSSTA and also 350 terms for opposing CSSTA. We also use the texts on the website "服貿東西軍"[4] to be our gold standards of supporting and opposing texts. The selected texts are used to do the evaluations of our keywords.

Another resource we used in this work is the PTT corpus, a social corpus which has been constructed and dynamically updated by LOPE lab at National Taiwan University[5]. As an online bulletin board favored by many of the youth, PPT is doubtless the largest public forum and social media in Taiwan, with more than 1.5 million registered users and over 150,000 users online during peak hours. Many newest information are posted instantly on the Gossiping board. We analyzed every post on Gossiping board from January 1, 2014 to July 1, 2014, in total around 150,000 posts.

3.2 Procedures

Basically, we follow the text mining techniques suggested by Gupta Gupta and Lehal (2009), e.g. feature extraction, search and retrieval, categorization, and summarization. The detailed procedures are described as follows.

- Extract features.

  We arranged the works of every person with the unified format, which includes the keywords and the corresponding texts. Then we save the data in CSV files.

- Open-sourced Chinese word segmentation with custom dictionary.

  In order to flexibly fit the target texts, we extend an open-sourced Chinese word segmentation system[6]. There are many long keywords in the texts, which needs to be reserved in segmentation, so we first create the user dictionary of every keyword and load it to Jieba before word segmentation.

[4] http://ecfa.speaking.tw/imho.php
[5] http://140.112.147.131/PTT/
[6] https://github.com/amigcamel/Jseg

- Establish the model for the classifier.

After segmentation, each text is saved as an document (a vector of features and weights). The weighting scheme of the model is TFIDF and the classifier is a SVM classifier, which separates the documents in a high-dimensional space by hyperplanes.

- Use cross validation for evaluations.

N-fold cross-validation performs N tests on a given classifier, each time partitioning the given dataset into different subsets for training and testing. The indices for evaluations are accuracy, precision, recall, F1, and standard deviation.

- Calculate the information gain from the classification model.

Information gain is a measure of a feature's predictability for a class label. Some features occur more frequently with definite type of texts, so they are more informative. The information gain is defined as

$$IG(T,a) = H(T) - H(T|a),$$

where H is the Information Entropy

$$H(X) = -\sum_i P(X_i)\log_2 P(X_i)$$

The information gain is the entropy reduced by adding the new feature *a*.

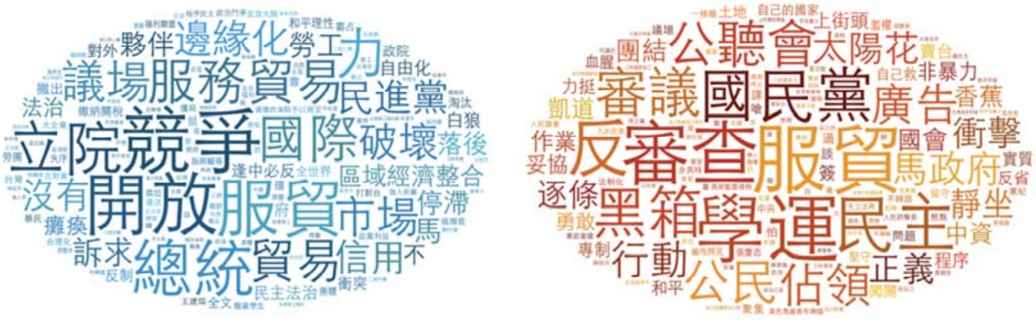
- Use the information gain to evaluate the texts from the PTT corpus.

We search the keywords of every post. Each keyword has the weight of the information gain. We sum over the information gain to judge the stance of the post, and then the scores of every post are further summed up in a day in order to observe the daily trend.

4. Results and Discussion

4.1 Keywords

We choose the keywords as the first step since many terms can potentially reveal one's attitude. For instance, the supporter for CSSTA would call students “霸佔”, *occup*y, the parliament, while the opponent would use “留守”, *stay*, in the parliament. The supporter emphasized “信用”, *credibility*, “經濟”, *economy*, and “秩序”, *social order*, while the opponent would stress the “黑箱”, *black box*, “行動”, *action*, and “正義”, *justice*. The following are the word clouds for two types of keywords.



(a) Supporting keywords.

(b) Opposing keywords.

Figure 1. Word clouds for supporting and opposing keywrods.

It is worth noting here that although opinion mining and sentiment analysis are often considered synonymous in many studies, it is necessary to draw the line between these two concepts. Following (Xu & Li, 2013), opinion is "a statement of the personal position or beliefs regarding an event, an object, or a subject (opinion target), while sentiment is the author's emotional state that may be caused by an event, an object, or a subject (sentiment target)". So as reflected in the lists of keywords, we may find words representing certain opinions may be associated with a sentiment (e.g., "破壞", *destroy*), but there are cases with standalone opinions (e.g., "開放", *open*).

4.2 Classifier

We use these keywords as features to train the classifier. The gold standards of the texts are chosen from the "服貿東西軍" website. The cross-validation yields the results in the table 1.

Table 1. Cross-validation tests for the classifier.

Accuracy	Precision	Recall	F-score	Std. Dev.
0.850	0.850	0.859	0.855	0.040

There are about one third of keywords which can be found in our testing data. (supporting keywords: 116/350, opposing keywords: 136/350) The results show that 85 percent of the texts can be correctly classified as positive or negative opinion toward CSSTA by these keywords. Therefore, with the validity of our keywords selection, we are able to use the information gain of keywords to do the trend analysis.

4.3 Information Gain

From the classification model, we also obtain the information gain of each keyword. The information gain means to what degree the keyword contains the political polarity. The larger the information gain of a word, the greater probability of distinguishing two types of texts by the word. Some samples are shown in table2. Some keywords can distinguish the texts better like “競爭”, *competition*, and “反服貿”, *anti-CSSTA*, and thus they have more weights in classifying the texts.

Table 2. Information gains for two types of keywords.

Type	Keyword	IG	Type	Keyword	IG
support	競爭(Competition)	0.1242	oppose	反服貿(Anti-CSSTA)	0.1013
support	總統(President)	0.0862	oppose	學運(Movement)	0.1013
support	邊緣化(Marginalization)	0.0628	oppose	國民黨(KMT)	0.0996
support	破壞(Destroy)	0.0603	oppose	審議(Deliberation)	0.0804
support	落後(Fall behind)	0.0444	oppose	民主(Democracy)	0.0638
support	貿易夥伴(Trading partners)	0.0412	oppose	跳針(Skipping)	0.0628
support	利大於弊(Good than harm)	0.0402	oppose	行動(Action)	0.0528

4.4 Trend Analysis

While sentiments are always polar, it is not always the case for opinions. So instead of aiming to do binary classification of political texts only, we turn to use the information gain to do the trend analysis. First, we sum keywords of each post, and sum over the posts of the same day. In other words, the score of each date is calculated as the following equation:

$$\text{Score} = \sum_{i}\sum_{w} IG(w) * C(w), i = \text{post index}, w = \text{word}$$

where $IG(w)$ denotes the information gain of a word w, $C(w)$ denotes the word count of w, and the summation first sum over the word w in a post, then sum over the post i in a day. The reason why we sum up the values of IG's is that since IG is the change in information entropy, we can add up the entropy changes to see the tendencies of a text in the topic of CSSTA. Higher IG value means closer relations to the topic. The results are shown in the Figure 2. The corresponding events are listed in the Table 3. The figure demonstrates the popularity of this topic of each day, and the top spike remarkably indicates that the discussion on CSSTA increases abruptly from March 18, which was the date that protesters occupied Taiwan Legislative chamber, to the March 23, which was the date that some protesters further occupied the Executive Yuan.

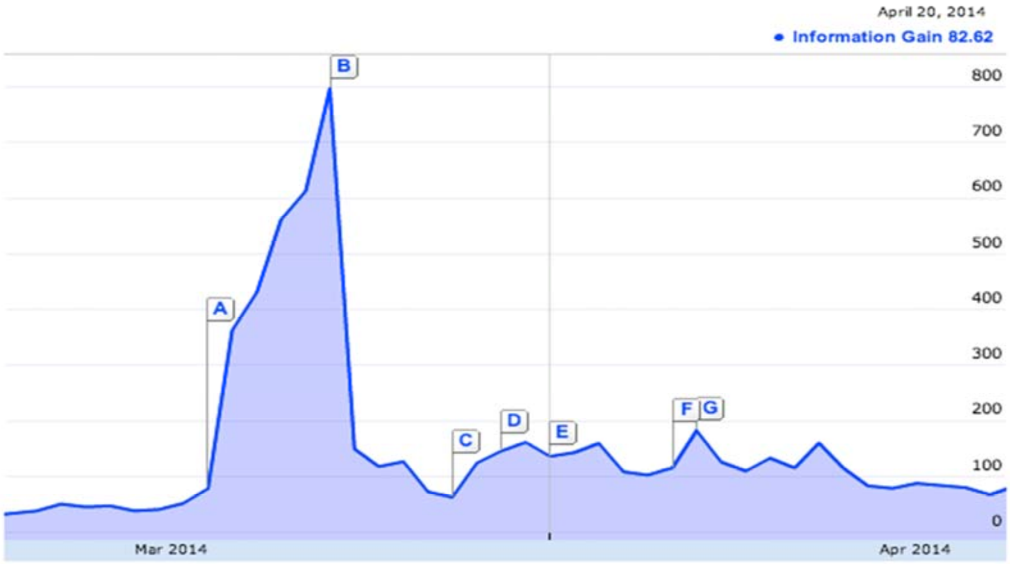


Figure 2. The trend of the topic popularity. (For the interactive figure, please click here.)

Table 3. Important events of Sunflower Student Movement.

Label	Date	Event
A	Mar. 18	Occupation of the Legislative Yuan
B	Mar. 23	Occupation of the Executive Yuan
C	Mar. 28	Rejection of the appeals by the Premier Jiang
D	Mar. 30	Demonstration
E	Apr. 1	March of the supporters
F	Apr. 6	Declaration of the President of the Legislative Yuan
G	Apr. 7	Announcement of the evacuation by the student leader Lin

The Figure 3 shows the ratio of supporting CSSTA from the analysis of posts. We calculate the supporting information gain over the total information gain, and also sum over the posts in one day. We can add the information gains like the previous analysis since the IG's are entropy changes. The information gains are added in both supporting and opposing aspects, and are compared to show the polarity of a text. The figure shows that the trend of supporting rate of CSSTA. The supporting rate drops on March 19, because of the Sunflower student movement. The supporting rate fluctuates for two possible reasons: the quantity of posts differs every day, and also the content of posts varies drastically. Thus the scores of the keywords varies in a wide range, which lead to the fluctuation of the supporting rate. But in general, we can see the tendency of the change.

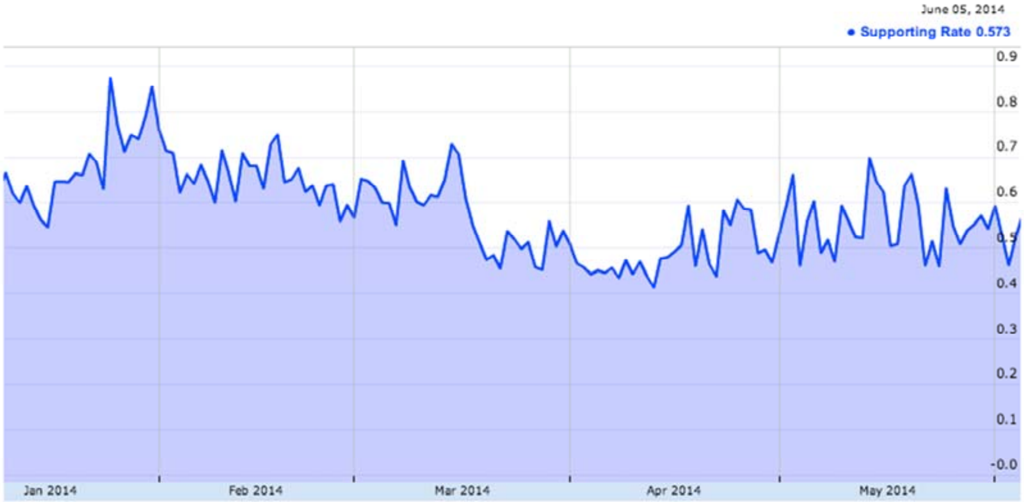


Figure 3. The trend of the supporting rate. (For the interactive figure, please click here.)

This method can be implemented on the coming election. The dynamic process of supporting rate for each candidate can be revealed by the texts on the social web, which is more efficient that the traditional telephone survey. Moreover, we can do more fine-grained analysis since the data is producing every day, and the We can ask, for example, how the event or the speech of the candidates affect their supporting rate. There are huge potential of the political interests.

5. Conclusion

Mining and tracking political opinions from texts in the social media is a young yet important research area with both scientific significance and social impact. The goal of this paper is to move one step forward in this area in Chinese context. We started from the manually created keywords and key phrases of CSSTA, used them to build a classifier and calculated their information gain, and then did the trend analysis of the PTT corpus. This approach involves interdisciplinary fields including information retrieval, data mining, statistics, machine learning, and computational linguistics. We hope that this text mining approach could discover the public opinion toward CSSTA, and further reveal political stances. Future works include more sophisticated language processing techniques applied to more broad domain of political topics, as well as developing dynamic tracking system gearing up for year-end election 2014.

References

Benoit, K., & Laver, M. (2007). Estimating party policy positions: Comparing expert surveys and hand-coded content analysis. *Electoral Studies*, *26*(1), 90-107.

Bermingham, A., & Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011*, 2-10.

Budge, I., Robertson, D., & Hearl, D. (1987). *Ideology, strategy and party change: spatial analyses of post-war election programmes in 19 democracies*. Cambridge University Press, 1987.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter." In ICWSM 2011, 89-96.

Gelbukh, E. F. , Gelbukh, E., Sidorov, G., Guzmán-arenas, A. *et al*. (1999). Text mining as a social thermometer," In *Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99*. Citeseer, 1999.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), 60-76.

Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications*, *39*(14), 11616-11622.

Kleinnijenhuis, J., & Pennings, P. (2001).11 measurement of party positions on the basis of party programmes, media coverage and voter perceptions. *Estimating the policy positions of political actors*, 2001, 162.

Klemmensen, R., Hobolt, S. B., & Hansen, M. E. (2007). Estimating policy positions using political texts: An evaluation of the wordscores approach. *Electoral Studies*, 26(4), 746-755.

Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 619-634.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02), 311-3313.

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3), 705-722.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10, 178-185.

Xu, G., & Li, L. (2013). *Social Media Mining and Social Network Analysis*. IGI Global, 2013.

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 29-46

學術論文簡介的自動文步分析與寫作提示

Automatic Move Analysis of Research Articles for Assisting Writing

黃冠誠*、吳鑑城*、許湘翎*、顏孜曦*、張俊盛*

Guan-Cheng Huang, Jian-Cheng Wu, Hsiang-Ling Hsu,

Tzu-Hsi Yen, and Jason S. Chang

摘要

學術論文是一種特殊的文體，有外顯、制式化的結構，如「簡介」、「相關文獻」、「方法」、「結果」、「討論」等。在各節中，又透過所謂文步的隱藏性修辭結構，有條不紊地呈現研究的背景、動機、內容。因此，在學術論文寫作的教學，分析文步扮演了重要的角色。在本論文中，我們提出了一個方法，將所給予的學術論文的每一個句子，標示所隱含的文步（moves），藉以幫助英文非其母語學生，寫作學術論文。我們採取透過常見寫作樣板（common patterns）取得訓練資料的研究路線。而我們的方法涉及擷取常見寫作樣板、標示樣板的文步、產生標示文步的訓練資料、設計分類特徵值、訓練一個文步分類器。在執行時，我們將句子轉換成特徵向量，運用分類器預測句子的文步。我們提出一個雛型系統 WriteAhead，應用分類的句子的資料，提示學習者，如何寫作各種文步的句子。

關鍵詞: 學術英文寫作、電腦輔助語言學習、修辭學、文脈分析

Abstract

Rhetorical moves are a useful framework for analyzing the hidden rhetorical organization in research papers, in teaching academic writing. We propose a

[*] 國立清華大學資工系
NTHU NLPLAB
Email: {cheng, hsiang, joe}@nlplab.cc; wujc86@gmail.com; jschang@cs.nthu.edu.tw

method for learning to classify the moves of a given set sentences in a academic paper. In our approach, we learn a set of move-specific common patterns, which are characteristic of moves, to help annotate sentences with moves. The method involves using statistical method to find common patterns in a corpus of research papers, assigning the patterns with moves, using patterns to annotate sentences in a corpus, and train a move classifier on the annotated sentences. At run-time, sentences are transformed into feature vectors to predict the given sentences. We present a prototype system, MoveTagger, that applies the method to a corpus of research papers. The proposed method outperforms previous research with a significantly higher accuracy.

Keywords: Academic English Writing, Computer-assisted Language Learning, Rhetoric, Context Analysis

1. 簡介

近年來，英文逐漸變成全世界學術研究最主要的溝通的媒介。而學術英文寫作，也成為非常重要的研究與教學的領域。學者也很重視，如何透過電腦的輔助，幫助一般性的語言學習，甚或特定性的學術寫作。學術寫作包含許多的文章類型，包括學術論文、計畫申請書、回顧與評論文章等（Swales, 1990）。其中，研究論文占有最重要的角色。

在學術論文中，「簡介」是絕大部分論文都有的第一個小節。現今，幾乎沒有學術論文，沒有「摘要」與「簡介」，而直接詳細地描述研究的目的、方法、結果。而且，對寫者和讀者而言，「簡介」在學術論文中都扮演非常重要的角色。一篇好的簡介，要能為整篇論文定調，抓住讀者的興趣，提供論文的扼要資訊。換言之，「簡介」 肩負重大責任——吸引讀者注意，讀完全文。

因此，有一些研究開始分析論文簡介如何達成其溝通的任務。Graetz (1985) 發現論文簡介似乎有共同的「問題—解法」修辭結構，依序包括問題（problem）、方法（solution）、評估(evaluation)、結論（conclusion）等部分。

Swales (1990) 分析大量的論文簡介，歸納出一套修辭的動機與模式：「創造研究空間」（Create A Research Space, CARS）。Swales 認為論文爭取研究得到讀者的認同，有如環境中生物爭取生存空間。為此，大部分作者依循三個修辭的步驟——也就是文步（moves）——來說服讀者。如圖 1 所示，這三個文步包括了「界定研究範圍」、「建立利基」、「佔據利基」。在每一個文步下，又需要描述若干必要或選項的內容。另外，美國國家醫學圖書館，也主張醫學論文作者，應提供分段有標題（labeled sections）的結構化摘要（structured abstract）[1]。

[1] 詳見 www.nlm.nih.gov/bsd/policy/structured_abstracts.html

CARS 文步	子文步與資訊內容
文步 I	1. 聲明研究領域的重要性，及/或
界定範圍	2. 聲明研究課題的廣泛性與普及性，及/或
	3. 回顧與評論前人研究
文步 II	1A. 提出與前人不同的聲明，或
建立利基	1B. 指出前人研究的缺口（gap），或
	1C. 提出本論文的研究議題（research question），或
	1D. 說明本研究所根據的典範與傳統
文步 III	1A. 概述本論文的目的，或
佔據利基	1B. 概述本論文的方法
	2. 宣布本論文的主要結果與發現
	3. 指出本論文的結構

圖 1. Swales (1990) 提出的 CARS 模式的文步與資訊內容

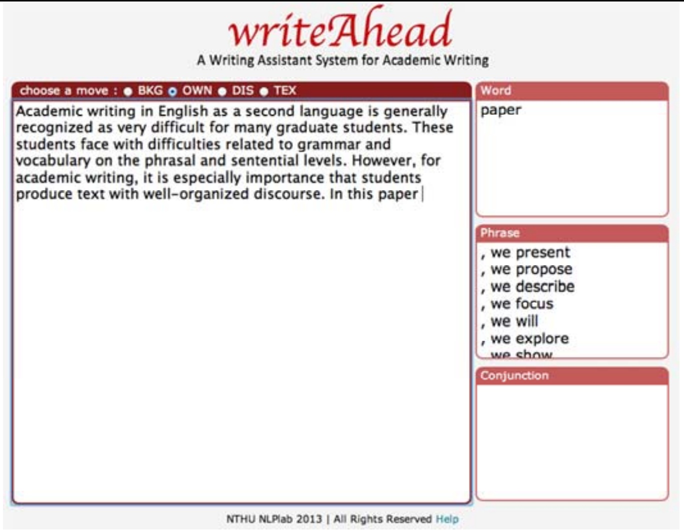
目前已經有許多學術寫作教材，透過文步分析來教導英文非母語的學生，如何寫作學術論文（如 Swales & Feak, 2004; Glasman-Deal, 2010）。也有研究者開發軟體系統（例如，Marking Mate: writingtools.xjtlu.edu.cn:8080/mm/markingmate.html），分析學生的作文並自動產生批改的建議與評分。但是很少有系統能夠在學生寫作中，依照文步的推進，適時地提供寫作提示與輔助。直覺上，如果我們能將大量的論文簡介加以處理，自動化分析其中每句的文步，繼而分析特定文步句子的常見片語或句型，我們將可以在寫作的過程，有效地協助學生。

然而，過去所提出的自動文步分析方法，都需費時費工標註大量論文。有鑑於此，我們提出新方法，以降低人工標註的工作量，且標注之資料將運用於訓練統計式分類器，來預測論文簡介中句子的文步，並藉以開發一個線上輔助寫作系統 WriteAhead。在 WriteAhead 的開發過程，我們採用了比 CARS 更簡單的文步分類，如圖 2 所示。用了此一分類方式，除系統較易於自動分類文步外，使用者亦比較容易掌握並使用於寫作過程。

我們期望此一自動文步分析工具，以及 WriteAhead 系統，有助於提升英文非母語者（non-native speakers, NNS）寫作學術論文的能力。在本論文中，我們提出了一套監督式機器學習的方法，能夠自動地學習如何將語料庫內的簡介句子，大略地分類為幾個文步。有了分類的句子之後，我們就可以統計各文步的 N 連詞（ngrams）詞頻。在 WriteAhead 系統，即可參考使用者選擇的文步，以及游標之前的內容，提示單字以及接續片語。

WriteAhead 文步	資訊內容	對應之 CARS 文步
背景（BKG）	領域：重要性、術語定義、缺口	文步 I-1,2,3, 文步 II-1B
	引用與評論前人研究	文步 I-3
本論文（OWN）	目的：輸入、輸出、條件	文步 III-1A, 文步 II-1C
	方法：研究路線、典範、依據、步驟	文步 III-1B, 文步 II-1D
	結果：實作、實驗、評估、結果、發現	文步 III-2
討論（DIS）	比較本論文與前人研究的相同之處	
	對照本論文與前人研究的相異之處	文步 II-1A
	未來研究方向	
文本組織（TEX）	提供全文的節大綱（目次表）	文步 III-3
	提供節內細分子節的大綱	
	指示圖表（編號）	
	回顧之前資訊、預告之後資訊	

圖 2. WriteAhead 採用文步與 CARS 模式文步之對照


writeAhead
A Writing Assistant System for Academic Writing
choose a move : ● BKG ○ OWN ● DIS ● TEX
Word
Phrase
Conjunction
NTHU NLPlab 2013 | All Rights Reserved Help

Academic writing in English as a second language is generally recognized as very difficult for many graduate students. These students face with difficulties related to grammar and vocabulary on the phrasal and sentential levels. However, for academic writing, it is especially importance that students produce text with well-organized discourse. In this paper

paper

, we present
, we propose
, we describe
, we focus
, we will
, we explore
we show

圖 3. WriteAhead 系統操作範例

圖 3 顯示 WriteAhead 系統的操作實例。在圖中，使用者已經介紹了研究背景（BKG 文步），接著使用者選擇了「本論文文步」（OWN），繼而輸入“In this paper” 等字。根據這些資訊，WriteAhead 顯示了適合此一脈絡的提示如下，作為繼續寫作的參考：

, we present	**, we describe**	**, we explore**
, we propose	**, we will**	**, we show**

WriteAhead 能夠提供與排列這些提示，是因為 WriteAhead 透過大量的論文原始資料以及少量的人工標示，學習如何辨識 OWN 文步的句子，並進而統計這些句子內的常見片語及其頻率。我們將在第三節詳述 WriteAhead 所運用的文步分類器的訓練過程。

本論文接下來的部分，安排如下。我們在下一節回顧相關的研究。接著，我們描述如何學習自動將論文簡介句子標註文步（第三節）。我們繼而描述如何將所提出的方法，實際製作成一個考慮文步類別進行寫作提示的雛形系統，以及相關的實驗設定、評估指標、以及實驗結果（第四節）。最後，我們指出未來研究方向，並作結論（第五節)。

2. 相關文獻

學術英文研究與教學（English for Academic Purpose）為相當重要的研究領域。近年來，學者對於研究計劃書，以及學術會議與期刊論文，都有深入的研究（Connor & Mauranen, 1999; Swales & Feak, 2004）。這些研究通常針對論文逐句逐段進行人為分析，經過歸納後，提出一套論文修辭的分析架構。在本研究中，我們則針對學術論文的「簡介」這一個部分，提出一套自動化的結構分析方法，並開發一套能夠讓學生一面寫作，一面獲得寫作提示的電腦輔助寫作系統。我們也討論如何在句子中，擷取能反應修辭結構的特徵，以有助於產生訓練資料，將句子歸類。

許多學者都指出，在表面上以及小節分段上，研究論文大致上有共通的簡單結構——IMRD 結構，即簡介（introduction）、方法（method）、結果（results）、討論（discussion）。也有學者進一步闡述 IMRD 的修辭結構，就像上下寬大，中間狹窄的沙漏：開始時先廣後專（from general to specific），結尾時由專而廣（from specific to general）。Swales (1990) 更為簡介這一個小節，提出了所謂的 CARS 模式（亦即「創造研究的空間」"Create a Research Space"）。CARS 模式歸納了典型的學術論文簡介修辭的動機與模式。CARS 模式提出之後，廣泛地為學者採用作為分析論文「簡介」節的寫作修辭策略 (例如，Cooper, 1985; Hopkins, 1985; Crookes, 1986; Samraj, 2002, 2005)。也有學者沿用 CARS 模式來分析「結果」節（Thompson, 1993），以及「討論」節（如，Hopkins & Dudley-Evans, 1994），以及醫學論文的摘要（Salager-Meyer, 1990, 1991, 1992）。與上述研究不同，我們採用人工督導與機器學習的方式，自動化分類與標註「簡介」節中句子的文步。

在自然語言處理的研究領域，Anthony & Lashkia (2003) 收集了近 700 篇論文摘要，並運用了 CARS 模式，人工標示摘要中每句的文步。之後，再透過機器學習方法，發展出自動文步標示系統 *MOVER*。Anthony 運用 *MOVER* 於學術寫作教學，發現可以幫助學生閱讀、分析、寫作摘要，讓學生有信心地寫出摘要的草稿，突破沒有使用輔助系統時，容易猶豫不決，久久難以下筆的障礙。然而，Anthony 發現 *CARS* 的文步劃分太細，造成 *MOVER* 標示文步的精確度不高。他建議合併相關易混淆文步。如此，可以大幅度提高 *MOVER* 文步分類的正確度，也不至於過於影響 *MOVER* 的效用。我們也將 *CARS* 的 3 大文步共 11 小文步，合併為 4 個文步，以提昇分類正確度，同時也減低使用者的認知負擔。

不同的學術領域的社群有不同文化與溝通的模式。醫學領域的編輯認為摘要應分成有標題的區段，亦即所謂結構化摘要（structural abstracts）。結構化摘要可以讓作者寫出的摘要，資訊完整、流暢易讀（Harley, 2000）。其實這些有標題的一到三句的小段，和文步的觀念是一致的。Shimbo *et al.* (2003) 運用了 MEDLINE 醫學文獻資料庫中標注區段或文步的摘要，開發一套分區檢索的文件資訊檢索系統。該系統運用支撐向量機（Support Vector Machine, SVM），將摘要中的句子劃分為「目的」、「方法」、「結果」」「結論」四種文步。Yamamoto & Takagi (2005) 也開發出類似的 SVM 系統，可將句子分為「背景」在加上以上四類的文步。Hirohata *et al.* (2008) 則是利用 CRF 系列分類器，來標示整個摘要。這些系統通常利用片語、動詞時態、句子位置、前後句特徵，做為分類依據。

近來，學者運用了許多不同的統計式分類的方法，開發文件或文步分類的作法。這些方法包括簡易的貝氏模型（Naïve Bayesian Model, NBM) (Anthony, 2003)，支撐向量機（Support Vector Machines, SVM) (McKnight & Arinivasan, 2003; Shimbo *et al.*, 2003; Yamamoto & Takagi, 2005)，隱藏式馬可夫模型（Hidden Markov Model, HMM) (Wu *et al.*, 2006; Lin *et al.*, 2006)，以及條件式隨機場（Conditional Random Fields, CRFs) (Hirohata *et al.*, 2008)。大部分的研究都是針對摘要，只有 Teufel (2000)、Teufel & Moens（2002）這一系列的研究，是針對全篇論文的分析。

和我們最相關的研究，應屬 Teufel（2000）的博士論文研究。Teufel 回顧評論文獻的各種文節分析的架構，並自行提出一套全篇論文的文步分析架構。Teufel （2000）和 Anthony（1993）有相同的意見，主張文步不宜做過度精細的分類，以免人工分類標示時，標示者難以達成共識。她提出兩層式的分類：大分類先分成「背景」、「前人研究」、「本論文」等三個文步。之後，本論文再細分為「目的」、「本論文」、「組織」；而前人研究再細分為「對照」、「立論根據」、「引用前人」，總共七個文步。本研究和 Teufel 的主要區別是，Teufel 利用人工分析，得到一組常見的句型，藉以分析文步，而本研究則透過自動化的語料庫分析得到一組常見的片語與句型。在訓練資料方面，Teufel 依賴對文節的直接標示，而我們透過常見片語與句型，間接標示句子的文步。而我們所採用的分類架構也有所不同，包括了「背景」（含領域、缺口、前人研究），「本論文」（含目的、方法、結果），「討論」（和前人研究的比較與對照），和文節結構（含論文組織、圖表的指示、內容的預告與回顧）等四種文步。

相較於文步分析的文獻中前人的研究，我們提出一套系統，能以較低的人工標示成本，自動學習如何產生訓練資料，進而學習文步的分類。我們並呈現一套實作的系統，其中利用自動化文步標示，輔助英語非母語學生寫作論文的簡介。總體而言，我們利用論文寫作中的常態與常用的表達方式，以及自然語言處理的技術來達到呈現論文修辭現象，並輔助學生寫作的目的。

3. 方法

為了能夠針對學生寫作論文過程中，所想表達的資訊（文步），提供適當的寫作提示，我們需要大量標示文步標籤的句子。人工逐句直接標示文步，無疑地非常費時耗工，絕非最好的作法。比較有潛力省時省力的方法，是先擷取一些論文少量常見句型（例如，“Recently, there have been ...”），透過人工檢視這些句型。決定句型是否大都表達特定的文步（如，背景與重要性）。如果句型有表達特定文步的傾向，就可以保留句型，並標註所屬文步。

最後，再以標示文步的句型來比對句子，產生大量標示文步的句子，以產生統計式文步分類器的訓練資料。

3.1 問題陳述

我們試圖收集大量學術論文，並對其中簡介部分的每個句子都標註修辭文步。之後，我們再利用這些標示資料，開發一個寫作輔助系統。這個系統要能接受學生設定的文步，提供適當的寫作提示。我們觀察學術論文中表達特定修辭文步時，常常用幾個相當特定句型。而一個常見的句型，在一份一萬篇論文的語料庫，出現可達數百次。所以，我們只要能夠擷取與標示這些句型，就可以得到大量的有標示的句子，當做訓練資料，運用於開發出一套文步分類器。運用此分類器，自動標示論文句子文步，我們就可以開發寫作輔助系統。我們現在正式地提出問題陳述。

問題陳述：給定 n 個學術論文「簡介」的句子 $S = s_1, s_2, ..., s_n$，我們的目標將 S 標註上一序列對應的文步標籤 $M = m_1, m_2, ..., m_n$，其中 m_i 為 s_i 的文步類型。為此，我們從 S 計算出常見的 k 個句型 $P = p_1, p_2, ..., p_k$，並人工標註對應文步 $T = t_1, t_2, ..., t_k$，而人工標示句型 p_i 為 t_i 時，必須確認符合 p_i 句型的句子大都表達 t_i 文步的資訊。

在本節的其餘小節，我們將描述我們對此一問題的解決方法。首先，在第 3.2.1 節，我們描述如何從網路收集學術會議與期刊的論文，並擷取其中的「簡介」此一節。接著，我們在第 3.2.2 節描述，如何從簡介中，統計常見的句型，以及人工標記常見句型之文步（第 3.2.3 節），進而產生標示文步之訓練資料（第 3.2.4 節）。最後，我們描述如何在訓練資料上，附加特徵值（第 3.2.5 節），以及訓練統計式機器學習模型（第 3.2.6 節）。

3.2 學習將論文句子標注文步

我們試圖找到一組各種文步的常見句型，藉以產生標示文步句子之訓練資料，以訓練一套文步分類器。我們的訓練過程如圖 4 所示。

(1) 從網路收集研究論文簡介	（第 3.2.1 節）
(2) 從論文簡介中統計常見句型	（第 3.2.2 節）
(3) 人工標記常見句型之文步	（第 3.2.3 節）
(4) 產生有文步標示之訓練資料	（第 3.2.4 節）
(5) 訓練資料附加特徵值	（第 3.2.5 節）
(6) 訓練機器學習模型	（第 3.2.6 節）

圖 4. 訓練模組的流程

3.2.1 從網路收集學術論文簡介

在訓練過程的第一步，我們收集大量的研究論文，以訓練文步分類器。為此，我們選擇有彙整論文可供直接下載的學會網站，且取得經過 PDF 檔案轉換或光學字元識別（OCR）處理的論文文字檔。然而，通常檔案都未標明節資訊。我們利用簡單規則，大致上辨識出節標題，並擷取論文「簡介」的部份。

3.2.2 擷取簡介常見句型

在訓練的第二步，我們利用現有的句子分割程式，將前一步驟取得的論文簡介，分割成一句一句。然後，再逐句進行切割詞彙（tokenization）、標示詞性（part of speech tagging）與基底片語（base phrases 或 chunks）擷取的預處理作業。

由於專有名詞（如作者名）以及數字（例如年度，或節、圖表編號）變化性大，以及名詞（如 method, approach 等）之前，常有各式的形容詞（如 new, novel）。這些現象都會導致句型發散，不易歸類成常見句型。為了有效歸納常見句型，對於句子內的詞彙，我們做以下的處理：

- 專有名詞、數字詞替換為其詞性標籤（即 NE, CD）
- 名詞片語、動詞片語，去除修飾語的部份，只留下中心語
- 複數名詞替換為單數名詞
- 不同時態的動詞替換為原形動詞

例如，我們會將原始的句子 (1) 替換為 (2) 之後，擷取 N 連詞（ngram）。除了考慮 N 連詞頻率，我們也計算相鄰詞語詞之間的相互資訊（mutual information），篩選所得的常見句型與片語，大都有修辭的功能，而且直覺上對寫作很有幫助的多字詞語（multiword expressions）或短詞串（lexical bundles）。

(1) **Researchers** have successfully **applied** ANN techniques **across** abroad **spectrum of** problem **domains** .

(2) **researcher apply technique across spectrum of domain** .

3.2.3 人工標記常見句型之文步

在訓練的第三步驟，我們挑選一些高頻且文步特性明顯的片語並手動地標記上文步。在此階段，我們將文步分為背景（BKG）、本論文（OWN）、討論（DIS）、文本（TEX）四種類型。 BKG 部分描述領域、課題、缺口、文獻，OWN 部分描述本論文之方法、結果，DIS 部分討論本論文與前人之優劣異同，TEX 部分描述全文或節的目的與組織。 表 1 顯示標了文步的片語範例，以及標籤的簡單定義。所以這個階段的標註對象是處理過後的片語。人工標註的過程中，很難控制標註的品質，因此標註者之間的一致性，需經反覆的核對，調解有衝突的標記 。

表 1. 有文步標記之句型範例

文步	句型	解釋
TEX	in section , we review work	文本：描述全文或節的目的與組織
BKG	research support in part by NE	背景：描述領域、課題、缺口、文獻
DIS	it be important to note that	討論：討論本論文與前人之優劣異同
TEX	rest of paper structure as follow	
OWN	in paper , we propose approach	本文：描述本論文之方法、結果
BKG	follow NE (CD) ,	

3.2.4 產生有文步標示之訓練資料

在訓練的第四步驟，我們利用有標記的句型去匹配大量論文簡介句子，並將句型的文步標註到句子上面。匹配的原則是愈長的句型愈優先。我們利用句型來產生大量有標記文步的句子，用以做為之後模組的訓練資料。表 2 為匹配成功的句子的範例。這個階段的標註範圍是單句。

表 2.句型對應句子的範例

文步	句型	匹配句子
TEX	in section , we review work	In the next section, we will first review some related works.
BKG	in year , there be	In recent years, there has been a rapid growth of interest in the sociological study of childhood.
OWN	in paper , we propose approach	In this paper, we propose a novel unsupervised approach to query segmentation, an important task in Web search.

3.2.5 附加訓練資料之特徵值

在訓練的第五階段，我們要附加特徵值到訓練資料以用來訓練標記文步模型。我們從句子中所抽出 N 連詞特徵值。表 3 為 N 連詞特徵值的例子。為了讓特徵值更能反應文步，我們也加入詞類、語意分類（Word class）的特徵值。我們利用 Teufel（1999）中人工編輯的一組學術論文的分類詞彙。表 4 為我們所使用的 語意分類（Word class）的特徵值。

表 3. 輸入句“In this paper , we will describe a method …”的 N 連詞特徵值

N-gram	Features
Surface unigram	in this paper we will describe a method
Surface bigram	in_this this_paper paper_, ,_we we_will will_describe describe_a a_method
Lemma unigram	in this paper we will describe a method
Lemma bigram	in_this this_paper paper_, ,_we we_will will_describe describe_a a_method
Chunk head unigram	in paper we describe method
Chunk head bigram	in_paper paper_, ,_we we_describe describe_method

表 4. 分類詞類集範例

詞類名稱	詞性	詞彙
AFFECT	v	afford, believe, decide, feel, hope, imagine, regard, trust, think
COMPARISON	v	compare, compete, evaluate, test
TEXT	n	paragraph, section, subsection, chapter

3.2.6 訓練機器學習模型

目前有許多機器學習方法可以處理分類的問題。基本的監督式的方法需要正確的分類資訊，非監督式方法則不需要有正確答案。在本研究中，我們採用監督式訓練方法，但是我們並不直接人工標註正確答案。我們透過標註少量句型，間接地自動產生大量的標記句子，作為監督式機器學習方法所需的訓練資料，並使用最大熵模型（Maximum Entropy, ME）來訓練文步分類器。

訓練完成後，我們就運用此一分類器，將語料庫內所有的論文句子，加以分類，標註上適當的文步。之後，我們就可以運用這些附有文步標籤的句子，來統計各種文步的常見 N 連詞。之後，WriteAhead 系統在輔助寫作時，將參照使用者設定的文步，並根據輸入的內容，查詢適當的片語提供給學習者參考。

4. 實驗與結果

我們設計 WriteAhead 的初衷，是為了提示使用者接著可以寫的數個字詞，以輔助學習者寫作學術論文的「簡介」。因此，我們擷取經過審查、編輯的程序，發表的學術論文，來實作我們提出的方法，以及開發寫作輔助系統。本節中，我們描述模組訓練的實驗設定（第 4.1 節），以及初步實驗的效能評估與結果（第 4.2 節）。

4.1 實驗設定

我們從密西根大學的計算語言學及資訊檢索組（Computational Linguistics And Information Retrieval Group, CLAIR) 設計維護的計算語言學會（Association for Computational Linguistic）會議與期刊論文典藏網站 *ACL Anthology Network*（AAN, clair.eecs.umich.edu/aan），我們擷取 AAN 學會的會議與期刊論文，共四萬多篇的論文的文字檔案。 這些檔案主要是由 PDF 格式的檔案，透過轉檔（類似於 OCR 辨識）所得到的文字檔案。因此，這些檔案有著各式的雜訊，像是殘留的換行連字符號、單字辨識錯誤等。 我們透過設計及分析規則，設定簡單的條件，辨識出節的標題， 並挑選了標示很清楚的論文將近一萬篇。之後，我們根據標題的編號，標題的內容，抽取「簡介」部分來做為研究的訓練資料，以及系統開發的資料。

表 5. 有匹配句型之句子文步分布情形

文步	句數
BKG	3,333
OWN	7,199
DIS	1,572
TEX	5,687
總計	17,791

我們逐篇處理這一萬篇論文簡介。我們利用 Python/NLTK[2] 的分割英文句子、詞彙的工具，將一篇篇論文分割成句子，再將句子分割成詞彙與標點（tokens）。有了句子與詞彙後，我們接著使用 Genia Tagger[3] 標註詞性與基底片語（base phrase 或 chunks）。之後，當所有的緒論單字都被斷詞和標記詞性以及區塊後，我們利用統計方法獲得若干的句型。我們人工的挑選了五百個句型後，手動濾掉文步特性不明顯得的片語並把剩下的句型都標上文步，剩下近約四百個有文步標記的句型。我們在利用這些標記過的句型去匹配一萬篇的論文簡介。我們得到大約一萬八千個句子，其文步的分佈如表 5 所示。再將標記好的句子附加上特徵值 N-gram、詞語分類後，讓 ME 模組做訓練，獲得文步標註模組。

[2] http://www.nltk.org

[3] http://www.nactem.ac.uk/GENIA/tagger/

我們藉由訓練所得的文步標註模組，對一萬篇簡介中的每一句進行文步標註。最後我們統計各種文步中的 N 連詞資訊，我們繼而將一萬多篇簡介內的句子，逐句做文步的分類，運用於 WriteAhead 寫作輔助系統。

4.2 評估與討論

如前所述，WriteAhead 的設計目標是輔助學習者寫作學術論文的「簡介」，所以應該評估各種寫作情境下,使用者覺得 WriteAhead 的提示,是否有助於寫作出更好的「簡介」。然而，一般而言，凡是涉及使用者的評估都是非常困難。退而求其次，我們目前僅針對文步分類器部分，評估其分類正確性。由於論文的文步是依序推移，所以我們針對「簡介」 的整個節，來評估文步的標註是否正確。

表 6. 總共 50 篇簡介之句子標示文步與預測文步與預測正確率

文步	標示句數	預測句數	正確句數	精確率
BKG	621	470	402	.86
OWN	238	259	144	.56
DIS	312	461	241	.52
TEX	117	98	75	.76
總計	1,288	1,288	862	.67

為了達成能自動的為論文簡介句子標註文步此一目標，我們從 *ACL Anthology Network* 中隨機挑選五十篇論文簡介的句子,做為我們文步標註模組的評估資料。表 6 顯示評估的結果。整體的文步預測正確率 67%，還有改善的空間。就個別的文步來看，背景文步 （ BKG）的正確率達 86% 而文脈文步（TEX）達 76%，這可能是因為背景、文脈文步兩者都有比較固定的表達方式。相對的，本論文（OWN）、討論（ DIS）兩種文步的精確率僅僅略高於 50%，這當然是因為表達的方式比較分歧，不易透過常見句型來加以掌握，未來可能還需要發掘比較有效的特徵值。

個別句子的分類正確率並不高，這可能歸咎於幾個原因。首先，標註資料太少，而且標註的正確性也不是非常理想。另外，表達同一類的文步，用字遣詞的差異性很大，很難用有限的資料來掌握，相反地字詞也有不小的詞彙語意歧義。

雖然個別句子的分類正確性不理想，我們觀察統計後的各分類之高頻 N 連詞還算合理。受限於時間，我們尚未評估 WriteAhead 運用各分類高頻 N 連詞，對於提示使用者的效果。不過我們認為，高頻 N 連詞的精確率可能遠高於文步標示的精確率。

本論文所使用的分類器是 Maximal Entropy ，未來也將考慮採用 SVM 或是 CRFs 。本論文所提出的方法，是基於跨領域的論文修辭研究，應該不會受不同學術專業領域特殊性的影響。但是，個別領域表達的方式在用字遣詞仍然有不小的差異，受限於資料，本系統應該對非資訊領域（例如文學、管理學、教育學）的適用性應該不是很理想，需要另外蒐集資料，依照學科建置不同的系統。

5. 結論

對於如何改善我們所提出的系統，我們預見許多可能的未來研究方向。例如，可以運用既有的自然語言處理技術，擷取更具效果的特徵值，來提升文步分類的正確率。例如，我們可以自動產生寫作文體之分類詞彙群。並且，根據分類詞彙群，擷取詞群式的常見樣板（class-based patterns），用來幫助分類的正確性，以及提供富含資訊的寫作提示。另外一個有潛力的研究方向，是讓使用者在另一個文字框，輸入母語（如中文、日文）草稿，而系統參考這些母語草稿，來調整提示的英文句型與片語。另外，我們也可以讓使用者選取部分沒有把握的 2-5 個字，系統提示正確或錯誤的機率，以及其他可以替換的表達方式。

總而言之，我們介紹了一套方法，能處理所搜集到的學術論文，將每一個句子標示上適當的文步（move），並統計各類文步的常見片語，藉以幫助英文非其母語學生，寫作學術論文。 我們的方法涉及擷取常見寫作句型、標示句型的文步、產生大量已標示文步的句子以及特徵值，作為訓練資料來開發文步分類器。我們藉由此一分類器，預測句子的文步。我們提出一個雛型系統 WriteAhead，應用分類的句子與常見片語的資料，提示學習者，如何寫作各種文步的句子。

致謝詞

本研究承蒙科技部補助研究經費，計畫標號 NSC 100-2511-S-007 -005 -MY3。

參考文獻

Anthony, L., & Lashkia, G. V. (2003). Mover: A machine learning tool to assist in the reading and writing of technical papers. *IEEE Trans. Prof. Commun.*, *46*, 185-193.

Connor, U., & Mauranen, A. (1999). *Linguistic Analysis of Grant Proposals: European Union Research Grants*.

Della Pietra, S., Della Pietra, V., Lafferty, J., Technol, R., & Brook, S. (1997). Inducing features of random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19*(4), 380-393.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing: Machinery, 16*(2), 264-285.

Graetz, N. (1985). Teaching EFL students to extract structural information from abstracts. In Jan M. Ulijn and Anthony K. Pugh, editors, R*eading for Professional Purposed: Methods and Materials in Teaching Languages*, pages 123-135. Acco, Leuven, Belgium.

Hirohata, K., Okazaki, N., Ananiadou, S., Ishizuka, M., & Biocentre, M. I. (2008). *Identifying Sections in Scientific Abstracts using Conditional Random Fields*.

Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006). Generative Content Models for Structural Analysis of Medical Abstracts. In *Proceedings of th HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BioNLP'06)*, 65-72.

McKnight, L., & Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *AMIA Annual Symposium Proceedings* (Vol. 2003, p. 440). American Medical Informatics Association.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., ... & Veuthey, A. L. (2007). Using argumentation to extract key sentences from biomedical bstracts. *International journal of medical informatics, 76*(2), 195-200.

Shimbo, M., Yamasaki, T., & Matsumoto, Y. (2003). *Using sectioning information for text retrieval: a case study with the MEDLINE abstracts.*

Swales, J.M. (1990). Genre analysis: English in Academic and Research Settings. *Cambridge University Press.*

Teufel, S. (1999). Argumentative Zoning: Information Extraction from Scientific Text. *PhD thesis, University of Edinburgh*.

Teufel, S., & Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics, 28*(4), 409-445.

Wu, J. C., Chang, Y. C., Liou, H. C., & Chang, J. S. (2006). *Computational analysis of move structures in academic abstracts*.

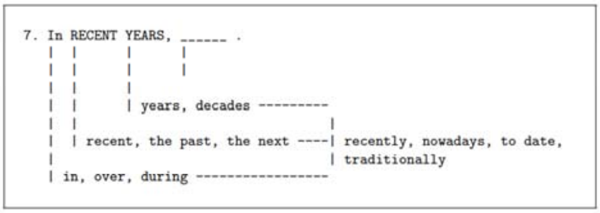
Yamamoto, Y., & Takagi, T. (2005). A sentence classification system for multi-document summarization in the biomedical domain. In *Proceedings of International Workshop on Biomedical Data Engineering*, 90-95.

附錄 A　整合常見句型的寫作樣板

我們擷取常見句型標示文步之後，發現許多句型很類似，只有少數的幾個字變動。我們可將這些句型聚集起來，歸納整合成為正規式樣板（regular expression patterns）。這些樣板避免羅列許多句型的不便，一目了然——既代表了寫作的常態，也呈現了各種變化。運用在教學上讓學生學習很有效果，寫作時也容易加以模仿、改寫。

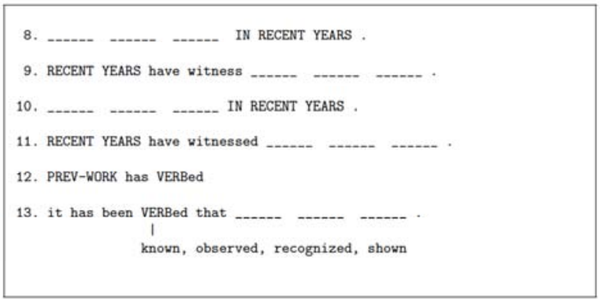
例如，從附錄 B 中我們可以看到下面左邊這些和時間有關的句型。經過觀察與歸納，我們可以得到下面右邊的樣板及其變化型：

recently , al (CD)
currently , there be
at present ,
over year ,
over decade ,
recently , method
in year ,
there be work
to date ,
in decade ,
currently ,
traditionally ,
recently ,



```
7. In RECENT YEARS, ______ .
   |  |      |      |
   |  |      |      |
   |  |      |
   |  |      | years, decades ---------
   |  |                                |
   |  | recent, the past, the next ----| recently, nowadays, to date,
   |                                   | traditionally
   | in, over, during -----------------
```

變化句型



```
 8. ______  ______  ______  IN RECENT YEARS .

 9. RECENT YEARS have witness ______  ______  ______ .

10. ______  ______  ______ IN RECENT YEARS .

11. RECENT YEARS have witnessed ______  ______  ______ .

12. PREV-WORK has VERBed

13. it has been VERBed that ______  ______  ______ .
                 |
                known, observed, recognized, shown
```

附錄 **B** 各種文步的常見句型

B.1 背景文步

follow NE (CD) ,
NE (CD) show that
NE (CD) demonstrate that
NE (CD) propose model
it be , however ,
there be , however ,
to knowledge , there be
to good of knowledge ,
in case , however ,
NE (CD) present
NE (CD) describe
however , in case ,
to knowledge , this be
collection comprise CD
in practice , however ,
recognition (NE) be
NE (CD) propose
as matter of fact ,

on hand , approach
currently , there be
this , however ,
first of all ,
however , for language
approach , however ,
research support by NE
however , there be
however , while
study show that
difficulty be that
currently , system
there be also
most of method
challenge be that
recently , model
however , they
at present ,
in general ,

it know that
as alternative ,
over year ,
this be important
much of work
over decade ,
however , if
however , unlike
recently , method
in year ,
it observe that
they show that
there be work
however , when
to date ,
most of system
to knowledge ,
this be task
it recognize that

however , since
in decade ,
however , study
however , approach
unfortunately ,
difficulty be
problem with
challenge be
they describe
currently ,
traditionally ,
in year
while approach
unlike method
recently ,
recently

B.2 「本論文」文步

in paper , we propose approach
in work , we focus on
in paper , we report on
in paper , we show that
in paper , we present approach
in paper , we present mothed
in paper , we present system
in particular , we show that
in paper , we focus on
in study , we focus on
in paper , we show how
in paper , we describe system
in paper , we propose

focus of paper be on
goal of research be to
aim of paper be to
in paper , we explore
in paper , we introduce
in paper , we use
purpose of paper be to
in paper , we consider
in paper , we describe
in paper , we address
in work we focus on
in work , we use
in paper , we study
in paper , we propose
goal of work be to
in paper , we investigate
goal of paper be to
goal in paper be to
in paper we describe

in study , we
paper address problem of
result show that model
in work , we
result show that method
to address problem ,
result show that approach
in paper we present
focus of paper be
we propose that
in study ,
paper focus on
we demonstrate that
in paper we
paper describe system
purpose be to
therefore , we
solution be to
idea be to

we start with
we hypothesize that
aim be to
in paper ,
we argue that
hypothesis be that
goal be
motivation for
in study
in paper
in work
paper present
purpose of
focus be
aim of
paper describe
we demonstrate
paper provide
we evaluate

method
in paper , we argue that
in paper , we propose model
in paper we focus on
in paper , we present
in paper we show that
in paper we describe system
work present in paper
in paper , we
we also show that
paper propose method for
in paper we discuss
in paper we investigate
in paper we propose
to achieve goal ,
in paper , i
thus , method
finally , result
experiment show that
work focus on
goal be to
claim be that
result indicate that
therefore , method
in work ,
evaluation show that
result show that
we evaluate approach
we show that

B.3「討論」文步

it be important to note that
this be due to fact that
contribution of paper be as follow
however , we believe that
advantage of approach be that
contribution of work be :
in order to do this
view express endorse by sponsor
as it turn out ,
reason for this be that
it be worth note that
contribution of paper be :
to overcome problem ,
for example , name
in particular , it
in contrast , model
it be obvious that
it turn out that
contribution of paper be
reason for this be
to knowledge , work
we also show how
in contrast , system
first , it
as result of
contribution be :
by contrast ,
in comparison ,
for reason ,
in practice ,
reason be that
specifically , it
this be problem
this lead to
as consequence ,
that be why
intuition be that
analysis show that
this mean that
we believe that
in principle ,
on contrary ,
example show that
difference be that
in short ,
we then discuss
unlike NE ,
it note that
among them ,
in sum ,
this be because
we note that
this suggest that
contribution be
advantage of
observation be
we believe
although approac

B.4「組織」文步

in section , we review work
remainder of paper organize as follow
in CD , we describe model
rest of paper structure as follow
in CD , we present model
remainder of paper structure as follow
rest of paper organise as follow
part of paper organize as follow
in CD , we present approach
we discuss result in CD
in CD we describe how
paper structure as follow :
in remainder of paper ,
in CD we discuss work
we discuss work in CD
next , in CD ,
in CD we present experiment
finally , CD conclude paper
section present and discuss result
CD present result of experiment
finally , we draw conclusion
in CD we present
in CD we discuss
in what follow ,
result show in CD
finally , CD present
article organize as follow
finally CD conclude paper
in section CD ,
paper organise as follow
in rest of paper
finally , in CD
work discuss in CD
discussion present in CD
paper organize as follow
in section that
we then present
CD describe model
CD present method
CD describe result
CD discuss result
CD review work
CD describe method
CD show result
plan of paper
finally , we
CD describe system
CD present result
CD present work

remainder of paper organise as follow
in CD , we describe system
rest of paper organize as follow
outline of paper be as follow
paper organize as follow : CD
structure of paper be as follow
in CD , we describe method
paper organize as follow : in
finally , we conclude in CD
in CD , we describe corpus
in CD , we review work
organization of paper be as follow
finally , CD present conclusion
finally , in CD ,

in CD we present result
for example , CD show
in section of paper ,
paper organize as follow :
in CD we show that
we conclude paper in CD
in rest of paper ,
finally , we present result
in section , we describe
CD show example of
finally , CD conclude
as we see ,
CD give overview of
result report in CD
result present in CD
as we show ,
paper proceed as follow
we conclude in CD
result discuss in CD

approach describe in CD
in CD we introduce
paper structure as follow
in CD we describe
conclusion draw in CD
result give in CD
after that ,
CD present evaluation
structure of paper
CD conclude paper
CD report result
CD describe algorithm
CD present algorithm
CD present experiment
CD introduce model
CD introduce method
CD present model
CD show example
CD describe how

CD describe experiment
CD describe setup
in section ,
CD show how
CD describe work
CD describe approach
CD give result
CD discuss work
in section
CD describe
CD introduce
CD conclude
CD show
CD detail
CD explain
CD present
CD discuss

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 47-60

使用概念資訊於中文大詞彙連續語音辨識之研究

Exploring Concept Information for Mandarin Large Vocabulary Continuous Speech Recognition

郝柏翰、陳思澄、陳柏琳

Po-Han Hao[*], Ssu-Cheng Chen*, and Berlin Chen*

摘要

語言模型是語音辨識系統中的關鍵組成，其主要的功能通常是藉由已解碼的歷史詞序列資訊來預測下一個詞彙為何的可能性最大，以協助語音辨識系統從眾多混淆的候選詞序列假設中找出最有可能的結果。本論文旨在於發展新穎動態語言模型調適技術，用以輔助並彌補傳統 *N* 連(*N*-gram)語言模型不足之處，其主要貢獻有二。首先，我們提出所謂的概念語言模型(Concept Language Model, CLM)，其主要目的在於近似隱含在歷史詞序列中語者內心所欲表達之概念，並藉以獲得基於此概念下詞彙使用分布資訊，做為動態語言模型調適之線索來源。其次，我們嘗試以不同方式來估測此種概念語言模型，並將不同程度的鄰近資訊(Proximity Information)融入概念語言模型以放寬其既有詞袋(Bag-of-Words)假設的限制。本論文是以中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)為任務目標，以比較我們所提出語言模型調適技術與其它當今常用技術之效能。實驗結果顯示我們的語言模型調適技在以字錯誤率(Character Error Rate, CER)評估標準之下，對於僅使用 *N* 連語言模型的基礎語音辨識系統皆能有明顯的效能提升。

關鍵詞： 語音辨識、語言模型、概念資訊、模型調適

[*] 國立臺灣師範大學資訊工程學系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {ie965225, boe20211}@gmail.com; berlin@csie.ntnu.edu.tw
The authors for correspondence are Ssu-Cheng Chen and Berlin Chen.

Abstract

Language modeling (LM) is part and parcel of automatic speech recognition (ASR), since it can assist ASR to constrain the acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the final output hypothesis given an input utterance. This paper investigates and develops language model adaptation techniques for use in ASR and its main contribution is two-fold. First, we propose a novel concept language modeling (CLM) approach to rendering the relationships between a search history and an upcoming word. Second, the instantiations of CLM are constructed with different levels of lexical granularities, such as words and document clusters. In addition, we also explore the incorporation of word proximity cues into the model formulation of CLM, getting around the "*bag-of-words*" assumption. A series of experiments conducted on a Mandarin large vocabulary continuous speech recognition (LVCSR) task demonstrate that our proposed language models can offer substantial improvements over the baseline *N*-gram system, and achieve performance competitive to, or better than, some state-of-the-art language model adaptation methods.

Keywords: Speech Recognition, Language Model, Concept Information, Model Adaptation

1. Introduction

語言模型(Language Models, LM)已被廣泛地使用於語音辨識、機器翻譯、資訊檢索以及文件摘要等各種任務之中，並成為關鍵的組成(Rosenfeld, 2000; Bellegarda, 2004)。在語音辨識任務上，其主要的功能通常是藉由已解碼的歷史詞序列(Word History)資訊來預測下一個詞彙(Upcoming Word)為何的可能性最大，以協助語音辨識系統從眾多混淆的候選詞序列假設(Candidate Word Sequence Hypotheses)中找出最有可能的結果(Furui *et al.*, 2012; O'Shaughnessy *et al.*, 2013)。最重要也最為常用的語言模型是 *N* 連(*N*-gram)語言模型，諸如二連(Bigram)與三連(Trigram)語言模型。*N* 連語言模型被用來估測每一個待預測詞彙在其先前緊鄰的 *N*-1 個詞彙已知的情況下出現的條件機率；由此可知，*N* 連語言模型是假設每一個詞彙出現的機率僅與它緊鄰的前 *N*-1 個詞彙有關，並以多項式分布(Multinomial Distribution)表示之。然而 *N* 連語言模型仍存在著許多缺點需要改善，至少有三點：(1)*N* 連語言模型限制了 *N* 的大小，僅能擷取短距離的詞彙規則資訊，無法考慮長距離的語句或篇章資訊；(2)當 *N* 增加時不僅會使模型參數量呈現指數性的遞增，造成空間與時間複雜度快速增加，也容易遭遇資料稀疏、無法為每一種詞序列的排列組合估測出準確的機率值的問題；(3)*N* 連語言模型極容易面臨訓練語料與測試語料不匹配(Mismatch)而造成的估測誤差。有鑑於此，近十幾年來有許多動態語言模型調適技術被提出，用以發展有效的語言模型輔助並彌補傳統 *N* 連(*N*-gram)語言模型不足之處。常見的有快取模型(Cache Model)(Kuhn, 1988)，以及源自於資訊檢索領域的主題模型(Topic

Model)(Blei & Lafferty, 2009)等；而主題模型在語音辨識任務的實作上，又以機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)(Hofmann, 1999)以及其延伸狄利克里分配(Latent Dirichlet Allocation, LDA)(Blei *et al*., 2003)最普遍被使用。

本論文旨在於發展新穎動態語言模型調適技術，用以輔助並彌補傳統 *N* 連(*N*-gram)語言模型不足之處。首先，我們提出所謂的概念語言模型(Concept Language Model, CLM)，其主要目的在於探詢隱含在歷史詞序列中語者內心所欲表達之概念，並藉以獲得基於此概念下詞彙使用分布資訊，做為動態語言模型調適之線索來源。其次，我們嘗試以不同模型架構與估測方式來建立此種概念語言模型，並將不同程度的鄰近資訊(Proximity Information)融入概念語言模型以放寬其既有詞袋(Bag-of-Words)假設的限制。本論文是基於公視電視新聞語料庫來進行中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition)實驗，以比較本論文所提出語言模型調適技術與其它當今常用語言模型調適技術之效能。

本論文的後續安排如下：第二節回顧當今常見的語言模型調適技術；第三節介紹本論文提出的概念模型以及不同模型估測方式，並嘗試將鄰近資訊(Proximity Information)融入概念語言模型；第四節介紹實驗語料、實驗設定以及實驗結果分析；第五節比較不同的語言模型；第六節則是結論及未來展望。

2. 常見的動態語言模型調適技術

動態語言模型調適宗旨在於希望在語音辨識過程中動態調整語言模型對於詞彙出現的預測機率，以獲得最好的語音辨識效能。本節將扼要回顧在語音辨識領域常被使用的動態語言模型調適技術。

2.1 快取模型

快取模型(Cache Model)是在二十多年前年首次被提出(Kuhn, 1988)，用在語音辨識過程中動態來輔助或調整 *N* 連語言模型於預測詞彙出現的機率。其基本概念是如果我們講了一些詞彙，則一段時間內這些詞彙再次出現的機率會很高。我們因此可以利用此線索在語音辨識過程中不斷地產生一個語言模型(例如單連快取模型)，並透過線性組合的方式與原始 *N* 連語言模型(例如三連語言模型)結合來動態地調適語音辨識所需的語言模型：

$$\hat{P}_{\text{Trigram}}(w_i \mid w_{i-2}w_{i-1}) = \lambda \cdot P_{\text{Trigram}}(w_i \mid w_{i-2}w_{i-1}) + (1-\lambda)\cdot \frac{n(w_i, H_i)}{|H_i|} \tag{1}$$

其中 $|H_i|$ 代表詞彙 w_i 對應的歷史詞序列 H_i 中的總詞數；$n(w_i, H_i)$ 是 w_i 在 H_i 出現的次數。過去許多研究亦實驗了二連快取(Bigram Cache)模型、三連快取(Trigram Cache)模型等更高階的快取模型，但由於歷史詞序列可能存有許多辨識錯誤資訊，以歷史詞序列來建立模型調適基礎 *N* 連語言模型的效果通常不是很顯著。

2.2 觸發對模型

觸發對模型(Trigger-Pair Model)模型可視為快取模型的延伸(Lau *et al.*, 1993; Troncoso & Kawahara, 2005)，其概念簡單來說是由訓練語料來統計出當任一詞彙 w_x 出現後，在同一文件中的一定間隔內會伴隨著另一詞彙 w_y 出現的可能性為何，這種伴隨關係稱之為「觸發對」(Trigger-pair)，其中 w_x 稱之為觸發項， w_y 稱之為被觸發項。觸發項與被觸發項的統計資訊可以藉由訓練語料中，統計、收集兩兩詞序列之間的平均交互資訊(Mutual Information)量多寡或是使用詞頻數(Term Frequency)與反文件頻數(Inverse Document Frequency)的關係來決定是否形成一個觸發對，以及其對應的條件機率 $P(w_y \mid w_x)$。觸發對模型運用於語言模型時，是由待預測詞彙 w_i 對應的歷史詞序列 H_i 中尋找詞彙 w_i 的可能的觸發項 $h_1, h_2, \cdots, h_{L_i}$ (假設歷史詞序列 $H_i = h_1, h_2, \cdots, h_{L_i}$ ，而每一個歷史詞彙 h_l 對於詞彙 w_i 的觸發機率為 $P(w_i \mid h_l)$)，並將這些觸發項分別預測的條件機率 $P(w_i \mid h_l)$ 動態線性組合而成為觸發對模型：

$$P_{\text{Trigger}}(w_i \mid H_i) = \frac{1}{L_i - 1} \sum_{l=1}^{L_i - 1} P(w_i \mid h_l) \tag{2}$$

而式(2)動態產生的觸發對模型亦可再透過線性組合方式與原始 N 連語言模型結合來動態調適語音辨識所需的語言模型(如式(1)的結合方式)。

2.3 主題模型

通常在資訊檢索任務上，主題模型藉由一組潛藏主題分布用來描述“詞彙-文件”共同出現的特性(Blei & Lafferty, 2009)。當主題模型被應用至語音辨識過程時，待預測詞彙 w_i 與其對應歷史詞序列 H_i (在此可視為一篇文件)之相互關係其有一組潛藏的主題分布用來描述歷史詞序列 H_i 與待預測詞彙 w_i 共同出現關係，不再是單純地經由計算 w_i 在 H_i 的出現頻率而估測，而是透過 w_i 出現在不同潛藏主題分布的頻率以及 H_i 產生這些潛藏主題的可能性來決定，是某種程度上的概念比對(Concept Matching)。機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)(Hofmann, 1999)以及其延伸狄利克里分配(Latent Dirichlet Allocation, LDA)(Blei *et al.*, 2003)是最常被使用的主題模型實例。在此舉機率式潛藏語意分析為例來作說明，當其被用至語音辨識來進行語言模型調適時，基於歷史詞序列 H_i 來預測詞彙 w_i 的發生機率可表示為(Gildea & Hofmann, 1999)：

$$P_{\text{PLSA}}(w_i \mid H_i) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid H_i) \tag{3}$$

其中 T_k 為某一個潛在主題，而 $P(w_i \mid T_k)$ 與 $P(T_k \mid H_i)$ 分別表示詞彙 w_i 發生在主題 T_k 的機率以及歷史詞序列 H_i 產生此主題的機率。我們假設每一個潛藏主題產生候選詞的機率 $P(w_i \mid T_k)$ 不因詞序列搜尋及拓展過程而變動，可先藉由最大化調適(或訓練)語料發生機率而求得；但由於歷史詞序列在語音辨識之前不能事先決定，而且數量非常多並且會隨語音辨識過程演進而改變，每一個歷史詞序列對於主題分布的權重必須在語音辨識過程使用期望值最大化(Expectation Maximization, EM)演算法(Dempster, 1977)來進行線上

(動態)估測。機率式潛藏語意分析的優點是在決定待預測詞彙 w_i 發生的機率時，不僅會考慮整個歷史詞序列 H_i 的主題分布特性，而且會隨語音辨識候選詞序列搜尋的演進，動態調整詞序列所含有的潛藏主題分布資訊。而式(3)動態地產生的機率式潛藏語意分析模型亦可再透過線性組合方式與原始 N 連語言模型結合來動態調適語音辨識所需的語言模型(如式(1)的結合方式)。

另一方面，狄利克里分配擁有與機率式潛藏語意分析相似的數學表示式，可視為後者之延伸，而且狄利克里分配在許多語音辨識任務上都展現了不錯的效用(Tam & Schultz, 2005)。兩個模型間的主要差異在於機率式潛藏語意分析假設其模型參數在參數空間上是固定和未知向量，而狄利克里分配對於模型參數多了先備限制(a Priori Constraints)，認為參數向量本身也是隨機變數，遵循著某種狄利克里分布特性。由於狄利克里分配模型的最佳化較為困難、不容易達到正確的估測，許多近似的估測演算法像是變動性貝氏近似(Variational Approximation)演算法或是吉卜森取樣(Gibbs Sampling)演算法因此被提出來估測狄利克里分配之模型參數(Blei & Lafferty, 2009)。關於主題模型的回顧與近期發展，可以參考(Blei, 2014; Kim *et al*., 2013; Potapenko & Konstantin, 2013)。

3. 概念語言模型

在本論文我們提出所謂的概念語言模型(Concept Language Model, CLM)來實踐語言模型調適，其主要假設是認為每一句的語句都是用來代表語者內心隱含而欲傳達的概念，並藉由語言(及語音)來具體表達相對應的概念。而概念模型最主要的目的則是希望能夠獲取使用者欲表達的概念，並假設在同一概念之中歷史詞序列中所有詞彙以及待預測詞彙具有共同的關係，進而藉此共同關係達到預測詞彙出現機率的目的。在實作上，概念模型會使用(搜尋)與初步語音辨識結果近似同領域文件(或調適語料)內表述的若干概念，用以近似語者內心欲傳達的真正含意，並基於此來建立概念語言模型。在本論文之中，概念模型的建立是分兩個面向來探討，分別是「詞彙」面向與「群聚」面向；以下將依序做介紹。

3.1 以詞彙面向建立概念語言模型

在我們想要表達某一特定概念時，我們常常會利用一組具有代表性的「概念關鍵詞」(Concept Words)來表達我們對事物的看法，而在同一概念底下用來描述事物的概念關鍵詞之間則具有相當高的關聯程度。例如在馬致遠的【天淨沙・秋思】之中，連續使用了「枯藤」、「老樹」等多個名詞串接，並藉由這一組連續的詞彙之組合來描述秋天蕭瑟荒涼的景象。基於此，本論文提出所謂詞概念語言模型(Word-based Concept Language Model, WCLM)，並應用於語言模型調適。在建構詞概念語言模型時，我們期望能夠針對每一語句不同的語言意涵，在調適語料的若干文件中挑選一組具有代表性的概念關鍵詞組 $\mathbf{c}$，藉以描述任一對歷史詞序列中所有詞彙與待預測詞彙之間的相依關係，如式(4)所示：

$$
\begin{aligned}
P_{\text{WCLM}}(w_i \mid H_i, W) &= \frac{P(w_i, H_i \mid W)}{P(H_i \mid W)} \\
&= \frac{\sum_{c\in\mathbf{c}} P(w_i, H_i \mid c)P(c \mid W)}{\sum_{c'\in\mathbf{c}} P(H_i \mid c')P(c' \mid W)} \\
&= \frac{\sum_{c\in\mathbf{c}} P(w_i \mid c)\prod_{l=1}^{L_i} P(h_l \mid c)P(c \mid W)}{\sum_{c'\in\mathbf{c}} \prod_{l'=1}^{L_i} P(h_{l'} \mid c')P(c' \mid W)}
\end{aligned}
\tag{4}
$$

其 W 代表語者所講語句所欲表達的語言資訊，在此我們先以語音辨識初步(第一階段)所產生的詞圖(Word Graph)(Ortmanns *et al*., 1997)來近似(詞圖包含所有可能的候選詞序列)；而 **c** 代表與 W 所欲表達的語言資訊有關的一組概念關鍵詞組。從式(4)的推導可看出詞概念語言模型欲模型化(紀錄)當某個概念關鍵詞 c 出現的情況下，待預測詞彙 w_i 與其歷史詞序列 H_i 共同出現的關係。同時，考量模型估測之可行性，式(4)進一步假設當某一個概念關鍵詞 c 出現的情況下，待預測詞彙 w_i 與其歷史詞序列 H_i 中任意的詞彙之間是彼此獨立的，也就是所謂的詞袋(Bag-of-Words)假設。而式(4)中 $P(w_i \mid c)$ 與 $P(h_l \mid c)$ 可從調適語料庫裡概念關鍵詞 c 所出現處的鄰近資訊(Proximity Information)，或者說是出現處上下文的詞彙分布而估測得； $P(c \mid W)$ 可透過適當方式計算 W 與 c 之相似度而求得。

實務上，我們首先遭遇到的問題就是「如何挑選具代表性的關鍵詞組？」。為此，本論文在挑選概念關鍵詞時運用了兩階段的挑選方式，如圖 1 所示。在第一階段時，我們利用了在資訊檢索領域之中常使用的虛擬關聯回饋 (Pseudo-Relevance Feedback, PRF)(Baeza-Yates & Ribeiro-Neto, 2011)，並利用基於庫爾貝克—萊伯勒差異量(Kullback-Leibler Divergence, KL-Divergence)之查詢與文件模型化技術(Kullback & Leibler, 1951; Zhai, 2008)，以詞圖 W(含有欲表達的詞彙和語意資訊)為查詢從調適語料的文件集檢索出一組較為相關的文件子集，稱這些文件為虛擬關聯文件(Pseudo-Relevance Documents)，並假設這些文件含有與所欲表達的語言資訊有關的概念。

在第二階段時，我們進一步從虛擬關聯文件子集裡挑選出一組一定數量的概念關鍵詞組，然後藉由這組概念關鍵詞組來量化(機率化)歷史詞序列中所有詞彙與待預測詞彙在此概念關鍵詞組下的共同出現關係。關於概念關鍵詞挑選準則，我們可以基於詞頻與反向文件頻率分數(TF-IDF Score)(Baeza-Yates & Ribeiro-Neto, 2011)。詞頻與反向文件頻率分數是一項常被用於資訊檢索以及文字分析領域中的技術，其公式可以表示如下：

$$
w_{j,m} = \begin{cases} (1+\log f_{j,m}) \times \log(N/n_j) & if \quad f_{j,m} > 0 \\ 0 & ohterwise \end{cases}
\tag{5}
$$

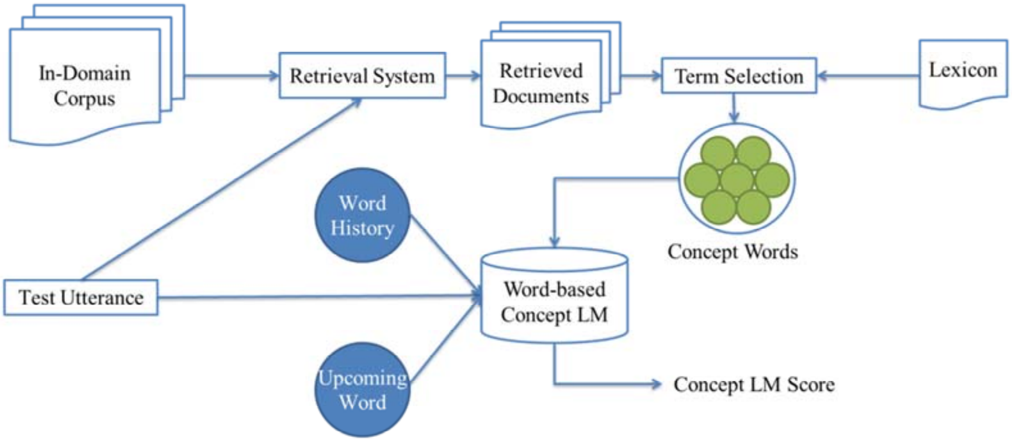


圖 1. 詞概念語言語言模型流程圖

上述的詞頻與反向文件頻率分數主要可分為兩個主要部分：第一部分為 $(1+\log f_{j,m})$，其中的 $f_{j,m}$ 則代表詞彙 w_j 在此文件 d_m 中所出現的次數，稱之為詞頻(Term Frequency, TF)，可以解釋為具越高詞頻的詞彙對文件來講越重要；第二部分為 $\log(N/n_j)$，其中 n_j 之則是代表詞彙 w_j 出現在所有虛擬關聯文件的文件個數，稱之為反向文件頻率(Inverse Document Frequency, IDF)，當某一詞彙出現僅出現在少數的文件之中，則此詞彙越具有獨特性。我們期望透過式(5)能找出具有重要性與獨特性的詞彙做為概念關鍵詞。

3.2 以群聚面向建立概念語言模型

群聚概念語言模型(Cluster-based Concept Language Model, CCLM)假設在調適語料的文件集內之文件可以由一組概念類別 **C** 來表示，藉由語者講的語句所欲表達的語言資訊 W 與這些概念類別的個別關聯程度來獲得語句可能的概念分布，並做為語言模型預測的根據：

$$P_{\text{CCLM-1}}(w_i \mid H_i, W) = \frac{\sum_{C\in\mathbf{C}} P(w_i, H_i \mid C)P(C \mid W)}{\sum_{C'\in\mathbf{C}} P(H_i \mid C')P(C' \mid W)} = \frac{\sum_{C\in\mathbf{C}} P(w_i \mid C)\prod_{l=1}^{L_i} P(h_l \mid C)P(C \mid W)}{\sum_{C'\in\mathbf{C}} \prod_{l'=1}^{L_i} P(h_{l'} \mid C')P(C' \mid W)} \tag{6}$$

其中概念類別的求取可透過一般分群演算法諸如 *K*-Means 演算法(Baeza-Yates & Ribeiro-Neto, 2011)而求得；$P(C \mid W)$ 可基於將語言資訊 W 與每一個概念類別 C 表示成向量形式，計算 W 與 C 之(餘弦)相似度而求得；$P(w_i \mid C)$ 代表概念類別 C 預測詞彙 w_i 的單連語言模型機率，可透過最大化相似機率估測而得(Zhai, 2008)。從式(6)的推導可看出群聚概念語言模型欲模型化(紀錄)當某一個概念類別 C 出現的情況下，待預測詞彙 w_i 與其歷史詞序列 H_i 共同出現的關係。

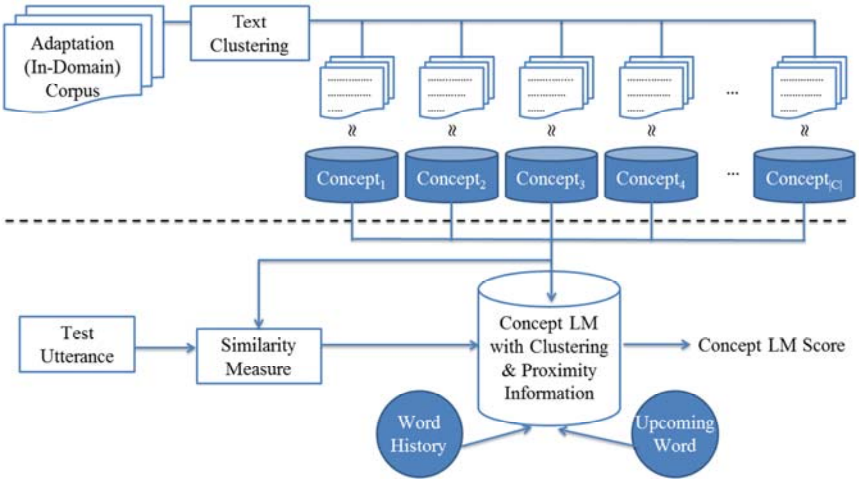


圖 2. 群聚概念語言模型示意圖

我們可以將式(6)中概念類別 C 預測詞彙 w_i 的語言模型延伸成為雙連(Bigram)或者三連(Trigram)語言模型，而可分別得到下面兩個表示式：

$$P_{\text{CCLM-2}}(w_i \mid H_i, W) = \frac{\sum_{C\in\mathbf{C}} P(w_i \mid h_L, C) P(h_1 \mid C) \prod_{l=2}^{L_i} P(h_l \mid h_{l-1}, C) P(C \mid W)}{\sum_{C'\in\mathbf{C}} P(h_1 \mid C') \prod_{l'=2}^{L_i} P(h_{l'} \mid h_{l'-1}, C') P(C' \mid W)} \tag{7}$$

$$P_{\text{CCLM-3}}(w_i \mid H_i, W) = \frac{\sum_{C\in\mathbf{C}} P(w_i \mid h_{L-1}, h_L, C) P(h_1 \mid C) P(h_2 \mid h_1, C) \prod_{l=3}^{L_i} P(h_l \mid h_{l-2}, h_{l-1}, C) P(C \mid W)}{\sum_{C'\in\mathbf{C}} P(h_1 \mid C') P(h_2 \mid h_1, C') \prod_{l'=3}^{L_i} P(h_l \mid h_{l-2}, h_{l-1}, C') P(C' \mid W)} \tag{8}$$

如此一來，概念語言模型可以同時考慮詞彙間出現的先後規則性或者是鄰近資訊(Proximity Information)，可以免除以詞袋(Bag-of-Words)假設的限制。最後，式(4)、式(6)、式(7)與式(8)動態產生的各種不同概念語言模型亦可再透過線性組合方式分別與原始 N 連語言模型結合來動態調適語音辨識所需的語言模型(如式(1)的結合方式)。圖 2 為群聚概念語言模型之示意圖。

4. 實驗設定與結果討論

4.1 實驗語料

本論文的語音辨識實驗是使用台師大所自行研發的大詞彙連續語音辨識系統(所使用詞典大小約為 7 萬 2 千詞)(Chen *et al*., 2004)以及公視新聞的公視電視新聞語音語料庫(Mandarin Across Taiwan Broadcast News, MATBN)(Wang *et al*., 2005)。此新聞語音語料庫是由中央研究院資訊所口語小組耗時三年(2001~2003)與公共電視台合作錄製完成。我

們初步地選擇外場採訪記者語料作為實驗題材，將其中約 25 小時收錄於 2001 年 11 月至 2002 年 12 月期間的語料作為最小化音素錯誤(Minimum Phone Error, MPE)聲學模型訓練的語料以建立聲學模型(Acoustic Models)(Liu *et al*., 2007)。另外，本論文以 2003 年所蒐集的語料中挑選各約 1.5 個小時作為發展集語料(Development Set)以及測試集語料(Test Set)，分別包含了 292 與 307 句的語句；我們以發展集語料來最佳化語言模型訓練所需之參數設定，然後據此作用在測試集語料。

表 1. 語音辨識實驗所使用之發展集語音語料以及測試集語音語料統計資訊

語料	句數	長度(小時)	說話速度
發展集語料	292	約 1.5	8.52 字/秒
測試集語料	307	約 1.5	8.50 字/秒

表 2. 語言模型估測所使用背景文字語料以及調適文字語料統計資訊

語料	詞數	句數
調適語料	約 1,000,000	3,643
背景語料	約 80,000,000	2,068,991

在語言模型的估測上，我們使用自 2001 至 2002 年中央通訊社(Central News Agency, CNA)的文字新聞語料，內含有約一億五千萬個中文字(經由斷詞之後約有八千萬個詞)做為背景語料庫用來訓練三連語言模型(Trigram Language Model)，此語言模型是使用 SRI Language Modeling Toolkit (SRILM)(Stolcke, 2000)訓練而得，採用 Good-Turning 平滑化方法來解決資料稀疏的問題。另一方面，我們亦蒐集同為公視電視新聞語料庫中的同領域文件做為調適語料庫，用來估測本論文中所探討的各式做為調適之用的語言模型，總共約三千六百多文句。本論文實驗所使用之語音語料庫以及文字語料庫的扼要統計資訊分別如表 1 與表 2 所示。

4.2 基礎實驗結果

在第一組實驗，我們於表 3 列出基礎語音辨識系統(使用背景三連語言模型，表示成 Background Trigram)以及一些常用語言模型調適技術在測試集語料的語音辨識字錯誤率(Character Error Rate, CER)結果，包括了觸發對模型(記作 Trigger)、機率式潛藏語意分析(記作 PLSA)以及狄利克里分配(記作 LDA)。值得一提的是，機率式潛藏語意分析以及狄利克里分配所使用潛藏主題數目設為 128；而這些語言模型調適技術都是作用在語音辨識的第二階段，也就是詞圖候選詞序列的語言模型重新排列(Word Graph Rescoring)。由表 3 我們可以觀察出三個現象。首先，觸發對模型(Trigger)似乎未能對基礎語音辨識系統的效能有顯著的提升。其次，機率式潛藏語意分析(PLSA)以及狄利克里分配(LDA)獲

得相同的語音辨識效能；相對於基礎語音辨識系統而言，能有約 4.6%的相對字錯誤率降低。第三，狄利克里分配雖使用較複雜的模型參數分布假設與估測演算法，但在我們的實驗裡並沒有獲得比機率式潛藏語意分析明顯較好的成果。

表 3. 語音辨識字錯誤率(%)：分別使用背景三連語言模型以及其它常見語言模型調適技術

Trigram	Trigger	PLSA	LDA
20.08	20.02	19.15	19.15

表 4. 語音辨識字錯誤率(%)：使用不同概念語言模型，包括 WCLM、CCLM-1、CCLM-2、CCLM-3。

WCLM	CCLM-1	CCLM-2	CCLM-3
19.30	19.26	19.18	19.03

表 5. 語音辨識字錯誤率(%)：群聚概念語言模型(CCLM-1、CCLM-2、CCLM-3)使用不同概念類別(群聚)數目。

概念類別(群聚)數目	CCLM-1	CCLM-2	CCLM-3
8	19.24	19.18	19.03
16	19.26	19.11	19.11
32	19.33	19.24	19.21
64	19.26	19.13	19.09
128	19.37	19.31	19.27

4.3 概念語言模型實驗結果

接著，我們評估本論文所提出兩類概念語言模型的語音辨識效能：詞概念語言模型(記作 WCLM)與群聚概念語言模型(記作 CCLM)。其中群聚概念語言模型因為單連語言模型、雙連語言模型和三連語言模型的使用(參考第三節)可以有三種變形(分別記作 CCLM-1、CCLM-2、CCLM-3)。基於在發展集語料所得出的最佳模型設定，在此 WCLM 共使用 128 個概念關鍵詞，而 CCLM-1、CCLM-2、CCLM-3 所使用的概念類別(群聚)數目分別為 16、8 與 8。它們在測試集語料的語音辨識字錯誤率結果列於表 4。基於表 4 的結果，我們有下列幾個觀察。首先，詞概念語言模型(WCLM)能較基礎語音辨識系統有一定的效能提升(約 3.8%的相對字錯誤率降低)，但其效用較機率式潛藏語意分析(PLSA)以及狄利克里

分配(LDA)來的稍差。其次，群聚概念語言模型在使用雙連語言模型和三連語言模型做為其組成模型(Component Models)時(参考式(7)與式(8))，能達到與機率式潛藏語意分析以及狄利克里分配差不多甚至更好的效果，例如 CCLM-3 能較基礎語音辨識系統有約5.2%相對字錯誤率降低。值得注意的是我們所提出的詞概念語言模型與群聚概念語言模型僅需要在進行詞圖候選詞序列的語言模型重新排列之前，執行一次文件檢索或者與概念類別(群聚)相似度估算，並不需像機率式潛藏語意分析以及狄利克里分配一樣在詞圖候選詞序列之語言模型重新排列時重新估算其組成模型，所以執行速度上會來得較快。另一方面，我們也嘗試結合詞概念語言模型與群聚概念語言模型，透過線性組合方式同時來調適基礎語音辨識系統所用之背景三連語言模型，而能讓字錯誤率下降至 18.98%。

最後，由於群聚概念語言模型能在上述實驗中獲得相當具競爭力的結果，我們因此進一步觀察它的變形(CCLM-1、CCLM-2、CCLM-3)在測試集語料使用不同概念類別(群聚)數目的表現，如表 5 所示。當我們比較表 4 與表 5 時可以發現，使用基於發展集語料所得最佳概念類別(群聚)數的各種群聚概念語言模型實際上在測試集語料上都有相當好的效能；顯示利用發展集語料所求得的模型(複雜度)參數稍後在測試集語料都能有一致的效能表現。

5. 結論與未來展望

在本論文，我們比較了一些常見語言模型調適技術在中文大詞彙連續語音辨識的效能。此外，我們提出所謂的概念語言模型(Concept Language Model, CLM)，其主要目的在於近似隱含在歷史詞序列中語者內心所欲表達之概念，並藉以獲得基於此概念下詞彙使用分布資訊，做為動態語言模型調適之線索來源。再者，我們嘗試以不同模型架構以及估測方式來實作此種概念語言模型，包括了將不同程度的鄰近資訊(Proximity Information)融入概念語言模型以放寬詞袋(Bag-of-Words)假設的限制。在基於公視電視新聞語料庫所進行的實驗顯示，我們所提出建構在概念語言模型之上的語言模型調適技術與其它當今常用技術相比，都夠達到具競爭性甚至較好的效能。關於未來研究方向，我們希望能結合或使用其它較新穎的模型，諸如遞迴式類神經網路語言模型(Recurrent Neural Network Language Model, RNNLM)(Mikolov *et al*., 2010; Deng & Yu, 2014)，來實現概念語言模型所欲擷取的詞彙和語意使用資訊。同時，我們亦希望能將其它在資訊檢索領域以發展相當不錯的新穎語言模型(Blei, 2014; Chen *et al*., 2004; Kim *et al*., 2013; Zhai, 2008)應用到中文大詞彙連續語音辨識的任務上。

致謝

本論文之研究承蒙教育部－國立臺灣師範大學邁向頂尖大學計畫(102J1A0800)與行政院科技部研究計畫(MOST 103-2221-E-003-016-MY2, NSC 103-2911-I-003-301, NSC 101-2221-E-003-024-MY3 、 NSC 101-2511-S-003-057-MY3 、 NSC 101-2511-S-003-047-MY3 和 NSC 102-2221-E-003-014- MY3)之經費支持，謹此致謝。

參考文獻

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: the Concepts and Technology behind Search*, Addison-Wesley Professional.

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech Communication*, *42*(11), 93-108.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, *1*, 203-232.

Blei, D. M, & Lafferty, J. (2009). Topic models. in Srivastava, A., & Sahami, M., (eds.), *Text Mining: Theory and Applications*, Taylor and Francis, 2009.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Chen, B., Kuo, J.-W., & Tsai, W.-H. (2004). Lightly supervised and data-driven approaches to Mandarin broadcast news transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, 777-780.

Chen, K.-Y., Liu, S.-H., Chen, B., Wang, H.-M., Hsu, W.-L., Chen, H.-H., & Jan, E.-E. (2014). Leveraging effective query modeling techniques for speech recognition and summarization. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Dempster, A. P., Laird , N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, *39*(1), 1-38.

Deng, L., & Yu, D. (2014). *Deep Learning: Methods and Applications*, Foundations and Trends in Signal Processing, Now Publishers.

Gildea, D., & Hofmann, T. (1999). Topic-based language models using EM. In *Proceedings of the European Conference on Speech Communication and Technology*, 2167-2170.

Furui, S., Deng, L., Gales, M., Ney, H., & Tokuda, K. (2012). Fundamental technologies in modern speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 16-17.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceeding of the ACM Special Interest Group on Information Retrieval*, 50-57.

Kim, D.-k., Voelker, G. M., & Saul, L. K. (2013). A variational approximation for topic modeling of hierarchical corpora. In *Proceedings of the International Conference on Machine Learning*.

Kuhn, R. (1988). Speech recognition and the frequency of recently used words: A modified Markov model for natural language. In *Proceedings of International Conference on Computational Linguistics*, 348-350.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79-86.

Lau, R., Rosenfeld, R., & Roukos, S. (1993). Trigger-based language models: a maximum entropy approach. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, 45-48.

Liu, S.-H., Chu, F.-H., Lin, S.-H., Lee, H.-S., & Chen, B. (2007). Training data selection for improving discriminative training of acoustic models. In *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding*, 284-289.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 1045-1048.

Ortmanns, S., Ney, H., & Aubert, X. (1997). A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech and Language*, 11, 43-72.

O'Shaughnessy, D., Deng, L., & Li, H. (2013). Speech information processing: Theory and applications. *Proceedings of the IEEE*, *101*(5), 1034-1037.

Potapenko, A., & V. Konstantin. (2013). Robust PLSA performs better than LDA. In *Proceedings of the European Conference on Information Retrieval*, 784-787.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? *Proceedings of IEEE*, *88*(8), 2000, 1270-1278.

Stolcke, A. (2000). *SRI Language Modeling Toolkit*. Available at: http://www.speech.sri.com/projects/srilm/.

Tam, Y., & Schultz, T. (2005). Dynamic language model adaptation using variational Bayes inference. In *Proceedings of the Annual Conference of the International Speech Communication Association,* 5-8.

Troncoso, C., & Kawahara, T. (2005). Trigger-based language model adaptation for automatic meeting transcription. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 1297-1300.

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: a Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics & Chinese Language Processing*, *10*(1), 219-235.

Zhai, C. X. (2008). Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, *2*(3), 137-213.

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 61-76

Some Prosodic Characteristics of Taiwan English Accent

Chao-yu Su[*+], Chiu-yu Tseng[+] and Jyh-Shing Roger Jang[#]

Abstract

The present study examines prosodic characteristics of Taiwan (TW) English in relation to native (L1) English and TW speakers' mother tongue, Mandarin. The aim is to investigate 1) how TW second-language (L2) English is different from L1 English by integrated prosodic features 2) if any transfer effect from L2s' mother tongue contributes to L2 accent and 3) What is the similarity/difference between L1 and L2 by prosodic patterns of word/sentence. Results show the prosody of TW L2 English is distinct from L1 English; however, TW L2 English and TW Mandarin share common prosodic characteristics which differentiate from L1 English. Analysis by individual prosodic feature shows distinct L2 features of TW English which might attribute to prosodic transfer of Mandarin. One feature is less tempo contrast in sentence that contributes to different rhythm; another is narrower loudness range of word stress that contributes to less strong/weak distinction. By examining prosodic patterns of word/sentence, similarity analysis suggests L1 and L2 speakers produce prosodic patterns with great within-group consistency respectively but their within-group patterns are distinct to counterpart group. One pattern is loudness of sentence and another one is timing/pitch patterns of word. The above prosodic transfer effect and distinct TW L2 patterns of prosody are found in relation to syntax-induced narrow focus and lexicon-defined word stress which echo our previous studies of TW L2 English and could be implemented to CALL development.

Keywords: Prosody, L1, L2, Mandarin, English, Contrast, Lexical Prosody, Narrow Focus.

[*] Institute of Information System & Application, National Tsing Hua University, Taiwan
E-mail: morison@gate.sinica.edu.tw

[+] Institute of Linguistics, Academia Sinica, Taipei, Taiwan
E-mail: cytling@sinica.edu.tw

[#] Department of Computer Science & Information Engineering, National Taiwan University, Taiwan
E-mail: jang@mirlab.org

1. Introduction

Computer assistant language learning (CALL) offers many advantages which differ from a traditional classroom setting where one teacher is responsible for a group of students. CALL allows learners to decide and adjust the level and pace of learning individually by. Another advantage that the classroom setting could not provide is unlimited access of on-line high-quality comparison between speech produced by a learner and a native speaker. By far the most popular CALL systems are computer-assisted pronunciation teaching (CAPT) system based on automatic speech recognition (ASR) outcome. The goals of CAPT are automatic diagnosis of pronunciation including specific or global error (Witt & Young, 2000; Coniam, 1999; Moustroufas & Digalakis, 2007), but the focus has been on segmental errors. However, in recent years studies focusing on suprasegmentals have shown that in addition to segmental information, prosodic information is in fact indispensable. Specifically, when detailed information of the consonant and vowel segments in the speech signal is removed, results show how listeners pay attention to prosodic features such as the pitch variation, rhythm alternation, loudness change as well as intonation. The resulting speech without any segmental and lexical content suggests that listeners are also sensitive to prosodic information (Scruton, 1996; Trofimovich & Baker, 2006; Munro, 1995). This has led to more research attention to investigate prosody in relation to comprehensibility and accent of native vs. non-native speech; and a more balanced understanding regarding the contribution from both the segmental and suprasegmental aspects of language (Derwing & Munro, 1997; Anderson-Hsieh *et al*., 1992; Munro & Derwing, 1999, Celce-Murcia *et al*., 1996; Derwing *et al*., 1998). Reported studies that applied prosodic training for second-language (L2) learners have demonstrated that computer-assisted prosody training systems did improve the overall comprehensibility of L2 speech (Hardison, 2004; Hirata, 2004). These studies showed prosody training with a real-time pitch display could improve both prosody and segmental accuracy, as judged by native speaker raters, while similar effect is found for English-speaking learners of Japanese. Another study demonstrated that aligning Mandarin English duration patterns with native English using resynthesis technology and dynamic time warping also brought significant increase in intelligibility (Tajima *et al*., 1997). Complementary findings are studies that showed how incorrect timing and stress patterns are often cited as major contributors to intelligibility deficit (Benrabah, 1997; Anderson-Hsieh *et al*., 1992). However, it appears that considerable gap does exist between research findings and software development. CALL systems are usually criticized as not necessarily “linguistically and pedagogically sound” (Derwing & Munro, 2005; Neri *et al*., 2002). For example, a study specifically states that most CALL programs were developed with little understanding of phonology and how to apply phonological knowledge to teaching (Pennington, 1999). In short, there is less understanding of L2 prosody, and even less CALL systems that have applied features of L2 prosody into the system.

The present study is developed from the above discussed background and aims to analyze prosodic characteristics of TW L2 English accent supported by linguistic knowledge. The speech data used in the present study is AESOP-ILAS (Asian English Speech cOrpus Project collected by the Institute of Linguistics, Academia Sinica) representing accent of Taiwan L2 English, which is part of AESOP that was designed and constructed to represent to include various kinds of L2 English spoken in Asia (Visceglia *et al*., 2009) with built-in linguistic knowledge (Anderson-Hsieh *et al*., 1992). Built-in linguistic knowledge in the corpus design is to elicit characteristics which are predicted to be present in L2 English speech. Our previous studies have catalogued a series of TW L2 features that may impede intelligibility. The series of studies to TW L2 accent started from prosodic under-differentiation which is not only found in syntax-elicited narrow focus but also in lexicon-defined word stress. Acoustic analysis of syntax-elicited narrow focus also showed that TW L2's production of narrow focus is less robust in F0 and amplitude than L1 (Visceglia *et al*., 2011; Visceglia *et al*., 2012). Further investigations of lexical-stress prosody showed the degree of contrast in F0 and amplitude is again less robust, making word stress in TW L2 English less differentiable (Tseng *et al*., 2012). The above two studies showed that lack of pitch and loudness contrasts is one of major feature of TW L2 accent in both word and sentence prosody. Further analysis revealed more complex L1s' features in words that may be difficult for TW L2 speakers (Tseng & Su, 2014). Native (L1) speakers may choose to realize word stress through binary stress/no-stress contrast anchored by the position of primary stress. Post-primary syllables are reduced to near-tertiary stress while pre-primary syllables are elevated to near-primary magnitude in F0. The 3-way primary/secondary/tertiary contrast is merged into a binary stress/no-stress contrast with robust prosodic contrast between the primary stress and its following syllable(s). As expected, the position-related merge of the secondary word stress is difficult for TW L2 speakers.

In addition to the above prosodic difference found between L1and TW L2 English, we also compared TW L2 accent and TW Mandarin, the target L2 speakers' mother tongue, and found in what ways TW L2 accent could be attributed to their L1 Mandarin features (Nguyen *et al*., 2008). Following this line of research, TW Mandarin is also included in the present study to further examine if and how some TW L2 English accent can further be attributed to Mandarin.

The present study aims to incorporate prosodic features found to contribute to TW L2 accent, and try to conduct prosody classification among L1 English, L2 English and Ll Mandarin by machine learning technology. The aim is to test if L1 English, L2 English and Ll Mandarin could be discriminated from each other by integrated prosodic features elicited by syntax-induced narrow focus and lexicon-defined word stress. Further discrimination analysis compares distinct prosodic characteristics of TW L2_Eng and TW L2_Eng-L1_Man shared characteristics of prosody to verify if prosodic features of TW L2_Eng are in relation to Mandarin. In addition, speaker-pair similarity by prosodic patterns is computed to test (1) difference between L1 English and TW L2 English groups and (2) cohesion within L1 English/TW L2 English group.

2. Speech Data

Read speech of Native English (L1_Eng), Taiwan L2 English (L2_Eng), Taiwan Mandarin (L1_Man) are used in present analysis. The materials of English speech are 5 reading tasks from the AESOP-ILAS recoded by 9 L1 (4M&5F) and 9 L2 (5M&4F) speakers. These 5 tasks are designed to elicit production of English segmental and suprasegmental characteristics including: (1) word-level features such as segmental by target words in carrier sentence; (2) phrase boundary phenomena such as declarative falls and interrogative rises by target words at phrase boundaries (3) form, timing and location of pitch accents, which are used to create phrasal and sentential prominence (broad and narrow focus) by target words in narrow focus position. 20 target words with 2-, 3- and 4-syllable of all possible stress patterns (Appendix A) are embedded in Task1 to Task 3. (4) function words in stressed and unstressed positions and (5) prosodic disambiguation of syntactic structures.

In section 3.1 and 3.2, the sentences in task 1 to task 5 are used for prosody classification among L1_Eng, L2_Eng and Ll_Man. In section 3.3, lexicon-defined prosodic similarity among speakers is computed by 20 stress-balanced target words in carrier sentence, Task1, to eliminate effect from higher level. An example of target word marked in boldface in carrier sentence is as follow.

- I said **SUPERMARKET** five times.

The sentences with broad and narrow focus in task 3 are used to test syntax-elicited prosodic similarity among speakers. An example of sentence in which broad and narrow focus are embedded is as follow. Narrow focus and broad focus are marked in boldface and italic respectively.

Context: Do you buy fruit at the farmer's market?

- No. I *usually* buy *fruit* at the **SUPERMARKET** because they stay *open later*.

After selecting sentences with acceptable F0 extraction, 369 L1_Eng and 434 L2_Eng sentences are used in present analysis.

The material of L1_Man is intonation balanced speech corpus (3441MB, 31:10) in SINICA COSPRO (Tseng *et al.*, 2003) which aims to examine role of intonation with respect to prosodic grouping in Mandarin speech. 3 types of sentences including declarative, interrogative and exclamatory with balanced POS combination are designed and collected in this corpus. In order to compare with English materials (task1 and task3 in AESOP-ILAS) in which all sentences are declarative, only declarative sentences are included in present analysis. Speech of one male and one female with good recording quality are chosen for analysis. After further selecting sentences with acceptable F0 tracking, 288 L1_Man declarative sentences are used in present analysis. Prosodic words in Mandarin are adopted as units of word-layer segmentation and corresponding feature extraction.

2.1 Annotation

All data were pre-processed automatically for segmental alignment using the HTK Toolkit, which was then manually spot-checked by trained transcribers for accuracy. F0 values were extracted and measured using a semitone scale.

3. Feature Extraction & Classification

3.1 Feature Extraction

Prosodic features used in present study are F0, duration, intensity. Each feature is z-normalized by sentence first then each sentence is encoded as a feature vector representing prosodic characteristics with hierarchical structure by sentence and word layer. The higher-level features, namely sentence-level features are derived by average of features in subsidiary units, namely word while word-level features are computed by subsidiary phoneme. In addition to conventional 6 types of general feature representation including mean, standard deviation, maximum, minimum, range and pairwise contrast referring to PVI (Grabe & Low, 2002) by each feature and each layer, histogram representation is also adopted to show more detailed properties of feature distribution. The adoption of histogram representation also could overcome inconsistent dimension among sentences which derived from varied number of words and phonemes thus requirement of consistent dimension could be fulfilled for classifier input. Two prosodic features encoded by histogram representation are mean and pairwise contrast by subsidiary units in sentence and word layer. Present histogram representation encodes prosodic features with 7 bins in which distribution of units is normalized to 100%. Normalized duration and F0 values were further refined to remove intrinsic physical properties based on previous knowledge. The intrinsic physical property for duration denotes segmental duration of each phoneme and intrinsic physical property for F0 denotes intonation of each sentence. 200 prosodic features in total are used in the present study.

3.2 Classification

Two popular classifiers for prosody classification among L1_Eng, L2_Eng and Ll_Man used are introduced as follows.

3.2.1 KNNC

The principle of k-nearest-neighbor classifier coded as KNNC (Cortes & Vapnik, 1995) is based on concept that data instances of the same class should be nearer in the feature space. As a result, for a given unknown data point x, the class is determined by K nearest points of x. The principles compute the distance between x and all the data points in the training space to decide K which is used for assign/predict class of unknown data point x.

3.2.2 SVM

Given a set of data with each example in data marked by binary categories, a support vector machine (SVM) (Coomans & Massart, 1982) training algorithm builds a model that assigns examples into one category or the other as accurate as possible while examples of the separate categories are divided by a clear gap that is as wide as possible. Unknown data points are then predicted to belong to a category based on which side of the gap they fall on.

3.3 Discrimination Analysis by Prosodic Features

Discrimination analysis is conducted between pair of speaker group by 200 prosodic features described in section 3.1. P value (Lehmann, 1997) is adopted as discriminative indicator between pair of speaker group. In a statistical test, sample results are compared to likely population conditions by way of two competing hypotheses: the "null hypothesis" is a neutral statement about "no difference" between two groups; the other, the "alternative hypothesis" is the statement that the person performing the test would like to conclude if the data will allow it. The *p*-value is the probability of obtaining the observed sample results when the null hypothesis is actually true. It could be quantified by the conditional probability $\Pr(X|H)$ (X is a random variable representing the observed data and H is the statistical hypothesis under consideration) which gives the likelihood of the observation if the hypothesis is assumed to be correct. If this *p*-value is very small, it suggests that the observed data is different from the assumption that the null hypothesis is true, and thus that hypothesis must be rejected and the other hypothesis accepted as true.

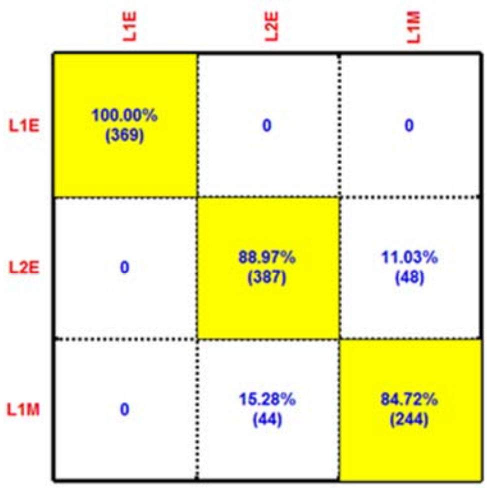
3.4 Similarity Comparison by Prosodic Patterns

The similarity is defined by cosine measure between any two of L1/L2 speakers by prosodic patterns of word/sentence. The value of point (i, j) in the matrix denotes cosine distance between speaker i and speaker j. In following section, the matrix is represented by a plot with i×j grids in which shading value of each grid denotes value of point (i, j). The darker the color is, the more similar between speakers i and j.

4. Results

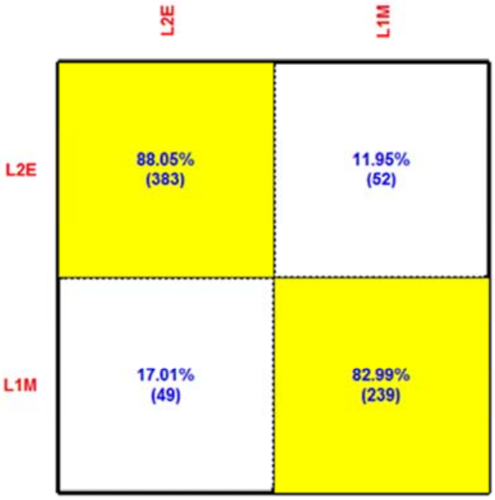
4.1 Prosody Classification among L1_Eng, L2_Eng and Ll_Man

In order to test if L1 English, TW L2 English and TW L1 Mandarin could be identified from each other by prosody, classification is conducted and performance is computed by 2 classifiers, SVM/KNNC. Average recognition rate is 91.57% by SVM and 81.86% by KNNC respectively. Figure 1 shows recognition rate in form of confusion matrix by best classifier, SVM and results suggest L1_Eng with most distinct characteristic with the others, L2_Eng and L1_Man. L1_Eng could be 100% identified from L2_Eng and L1_Man; however, only 88.97% of L2_Eng and 84.74% of L1_Man could be recognised from the others. Further binary classification is conducted between L2_Eng and L1_Man and shows best recognition rate 86.03% by SVM. Figure 2 shows confusion matrix which demonstrates only 88.05% of L2_Eng and 82.99% of L1_Man could be identified from each other.



	L1E	L2E	L1M
L1E	100.00% (369)	0	0
L2E	0	88.97% (387)	11.03% (48)
L1M	0	15.28% (44)	84.72% (244)

Figure 1. The recognition rate among L1_Eng, L2_Eng and Ll_Man by prosodic features and SVM



	L2E	L1M
L2E	88.05% (383)	11.95% (52)
L1M	17.01% (49)	82.99% (239)

Figure 2. The recognition rate between L2_Eng and Ll_Man by prosodic features and SVM

4.1.1 Discussion

The above results suggest that L1_Eng could be differentiated from L2_Eng and L1_Man; however, confusion is found between L2_Eng and L1_Man. In other words, L1_Eng is distinct from L2_Eng and L1_Man prosodically; on the other hand, L2_Eng and L1_Man share some common prosodic characteristics which differentiate from L1_Eng. In the following section, discrimination analysis is conducted by prosodic features to show distinct prosodic characteristics of L2_Eng from L1_Eng and common prosodic characteristics between L2_Eng and L1_Man.

4.2 Discrimination Analysis by Prosodic Features

Table 1 shows most distinct prosodic characteristics between L2_Eng and L1_Eng. After pairwise discrimination analysis between L2_Eng and L1_Man is conducted by each prosodic feature, the most discriminative features are computed and listed in Table1. Results show most discriminative prosodic features by lowest 5 p-values in L2_Eng vs. L2_Eng are 'mean by normalized F0', 'minimum by normalized F0', 'mean by normalized volume', 'maximum by normalized volume' and 'stand deviation by normalized duration' in sentence layer and maximum/PC/stand deviation/range/histogram_dimension#3 by normalized volume in word layer.

Table 1. The most distinct prosodic characteristics between L2_Eng and L1_Eng by p-value

Speech Pair Layer	L2_Eng vs. L1_Eng
Sentence Layer	'NorF0_Mean'
	'NorF0_Min'
	'NorVol_Mean'
	'NorVol_Max'
	'NorDur_STD'
Word Layer	'NorVol_Min'
	'NorVol_PC'
	'NorVol_STD'
	'NorVol_Range'
	NorVol_hisBySubMean_D3'

Table 2. The most similar prosodic characteristics between L2_Eng and L1_Man by p-value

Speech Pair Layer	L2_Eng vs. L1_Man
Sentence Layer	'NorVol_DisBySubPC_D5'
	'NorDur_DisBySubPC_D1'
	'NorDurWOIntri_DisBySubMean_D5'
	'NorDur_DisBySubPC_D3'
	'NorF0_PC'
Word Layer	'NorF0_Mean'
	'NorVol_Range'
	'NorF0Res_DisBySubMean_D2'
	'NorF0_DisBySubPC_D6'
	'NorVol_DisBySubPC_D7'

Table 2 shows common prosodic characteristics between L2_Eng and L1_Man. Pairwise discrimination between L2_Eng and L1_Man is conducted by prosodic feature and most similar features are listed in Table 2. Results show most similar prosodic features by highest 5 p-values by L2_Eng vs. L1_Man are 'histogram_dimension#5 by pairwise contrast of normalized volume', 'histogram_dimension#1&3 by pairwise contrast of normalized duration', ' histogram_dimension#5 by normalized duration without intrinsic properties' and 'pairwise contrast by normalized F0' in sentence layer and 'mean by normalized F0', 'range by normalized volume', 'histogram_dimension#2 by f0 without intonation effect', 'histogram_dimension#6 by normalized F0'and 'histogram_dimension#7 by normalized volume in word layer.

4.2.1 Discussion

The results show F0/duration/volume in sentence layer and volume in word layer contribute to TW L2 accent. By discrimination analysis between L2_Eng and L1_Man, results demonstrate F0/duration/volume in sentence layer and F0/volume in word layer are shared L2_Eng-L1_Man prosodic properties. We further assume that distinct features of L2 accent might attribute to prosodic characteristics borrowed from their mother tongue, namely L1_Man thus distinct features of L2Eng are compared with L2Eng-L1Man shared features. The results show distinct L2_Eng features do overlap with L2Eng-L1Man common features. Comparison by sentence layer shows similar features found coexisting in L2Eng-L1Eng distinct features and L2Eng-L1Man common features (green in Table 1 and Table 2) are stand deviation by normalized duration in L1Eng-L2Eng distinct features and histogram_dimension#1&3 by pairwise contrast of normalized duration in L2Eng-L1Man common features. Pairwise contrast is defined by between-phone variation and the property is similar to stand deviation representing global variation; thus we could regard them as overlap. In summary, the results suggest tempo contrast by syntax-elicited narrow focus in sentence layer and loudness range by lexicon-defined word stress in word layer are distinct L2 features of TW English which might attribute to prosodic transfer of Mandarin, namely L2s’ mother tongue.

4.3 Similarity Comparison by Prosodic Patterns

In addition to analysis by individual prosodic feature in section 3.2, similarity is computed between any two of L1/L2 speakers by prosodic patterns of word/sentence. After between-speaker similarity is derived, we examine if between-speaker similarity is greater when they are in the same speaker group. The aim is to test if consistency within each speaker group (L1/L2) and discrimination between L1 and L2 could be found.

4.3.1 Similarity in Word Prosody

Figure 3, 4 and 5 show similarity matrix between any two of L1/L2 speakers by prosodic patterns of word. First row by normalized duration in Figure 3 demonstrates by color lightness, first L1 speaker is more similar with speaker 1 to speaker 9 than speaker 10 to speaker 18 which represent L1 speakers and L2 speakers respectively. In addition, the left-top block by green dotted cross demonstrates L1 speakers with more consistency within group than the other blocks. It suggests L1 with greater cohesion/consistency than right-top (L1 vs. L2), left-bottom (L1 vs. L2) and right-bottom (L2 vs. L2). Right-bottom (L2 vs. L2) block also shows secondary consistent which is darker than right-top (L1 vs. L2), left-bottom (L1 vs. L2). It suggests L2s' prosodic patterns are consistent as well. Normalized duration without intrinsic properties in Figure3 further shows that removing intrinsic duration could further help to discriminate L1 and L2.

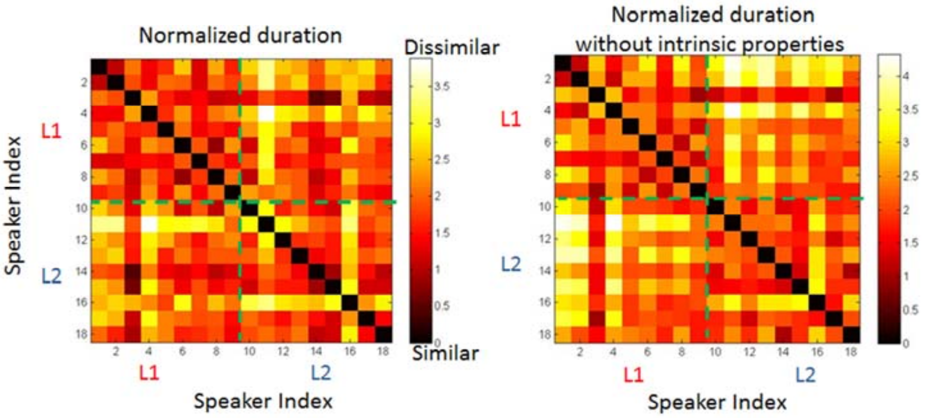


Figure 3. The similarity between any two of L1/L2 speakers by duration patterns in word layer. Color bars show the more dark the color, the more similar between two speakers. The value of point (i,j) in the matrix represents cosine distance between i and j that diagonal indicates self-similarity with darkest color. The green dotted cross represents boundary between L1 and L2 speakers.

Figure 4 also shows great cohesion within speaker group (L1&L2) respectively and great difference between speaker group (L1 vs. L2) by normalized F0 and normalized F0 without intonation effect; however, removing intonation appears not to improve L1-L2 discrimination significantly.

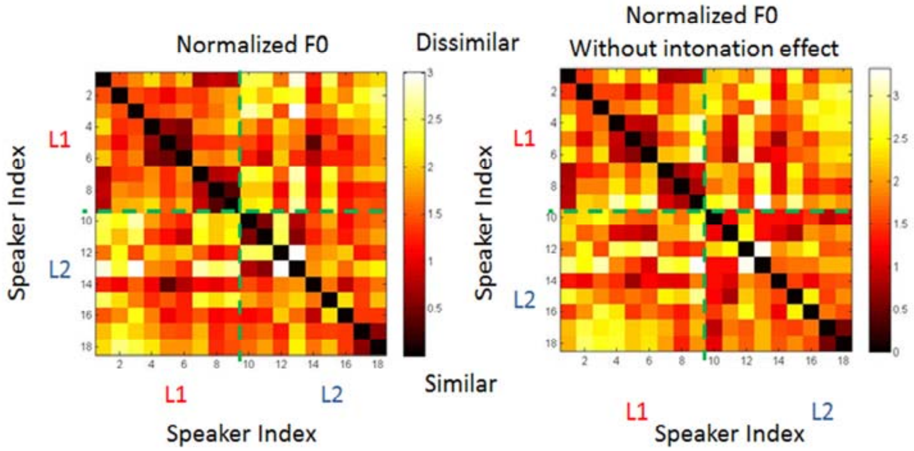


Figure 4. The similarity between any two of L1/L2 speakers by F0 patterns in word layer.

Figure 5 shows similarity matrix by normalized intensity. Results show no significant discrimination found between L1 and L2.

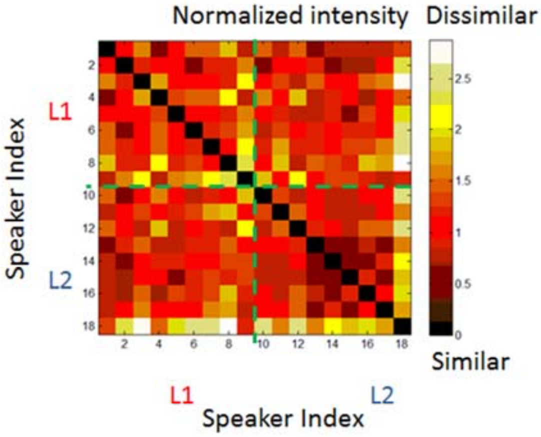


Figure 5. The similarity between any two of L1/L2 speakers by intensity patterns in word layer.

4.3.1.1 Discussion

By between-speaker similarity of word by duration/F0, the two distinct blocks by shading value representing L1s' and L2s' patterns are found. It suggests between-speaker similarity by word layer is greater when they are in the same speaker group. In other words, L1 and L2 produce respective timing/pitch patterns of word with great within-group consistency but within-group features are distinct from counterpart group. Between-group discrimination and within-group consistency is not found by loudness patterns. The results suggests timing/pitch patterns elicited by lexicon-defined word stress in word layer are distinct L2 features of TW English.

4.3.2 Similarity in Sentence Prosody

Figure 6, 7 and 8 show similarity matrix between any two of L1/L2 speakers by prosodic patterns of sentence. By Figure 6 and 7, no significant discrimination between L1 and L2 is found by normalized duration, normalized duration without intrinsic properties, normalized F0 and normalized F0 without intonation.

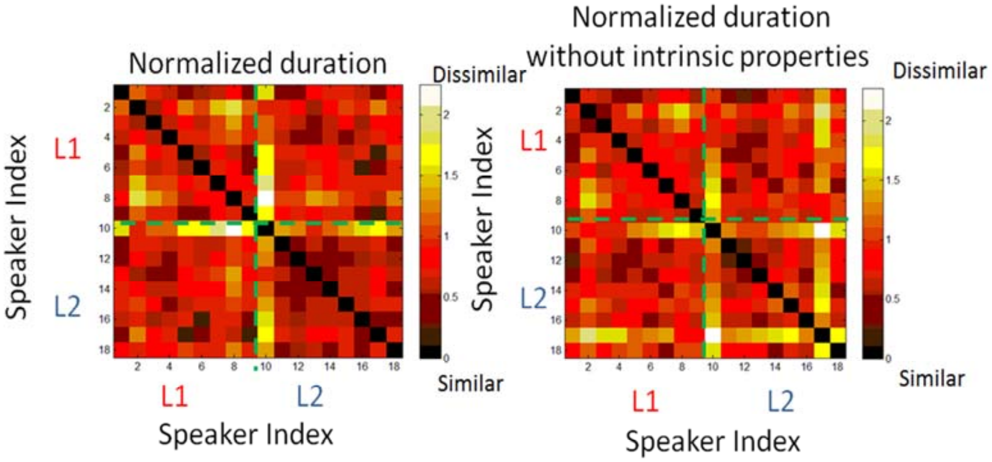


Figure 6. The similarity between any two of L1/L2 speakers by duration patterns in sentence layer.

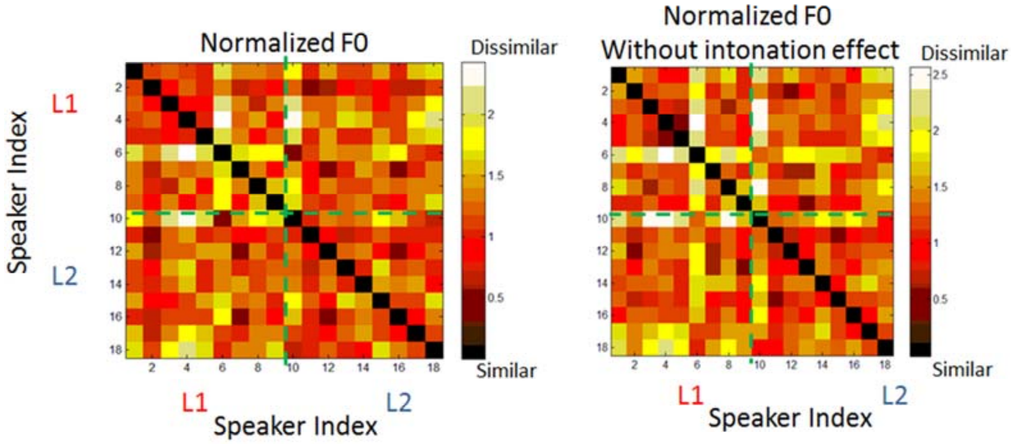


Figure 7. The similarity between any two of L1/L2 speakers by F0 patterns in sentence layer.

Figure 8 shows intensity patterns of sentence with great within-group cohesion and great between-group difference in both L1 and L2.

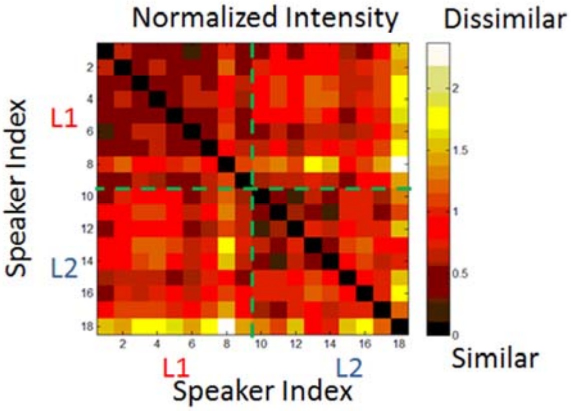


Figure 8. Similarity between any two of L1/L2 speakers by intensity patterns in sentence layer.

4.3.2.1 Discussion

By intensity similarity of sentence, the two distinct blocks by shading value representing L1s' and L2s' patterns are found. It suggests between-speaker similarity by intensity of sentence is greater when they are in the same speaker group. In other words, L1 and L2 produce respective prosodic patterns with great within-group consistency but within-group features are discriminative to counterpart group. Between-group discrimination and within-group consistency is not found by timing/pitch patterns. The results suggest loudness patterns elicited by syntax-induced narrow focus in sentence layer are distinct L2 feature of TW English.

5. Discussion and Conclusion

The present study examines prosodic characteristics of Taiwan English in relation to native English and Mandarin, mother tongue of TW speakers. Prosody classification among native English, TW L2 English and TW Mandarin is conducted by machine learning technology and results show Taiwan L2 English is found to be distinct from L1 English in prosody. However, TW L2 English and Taiwan Mandarin share some common prosodic characteristics which differentiate them from L1_Eng. Further comparison by each prosodic feature shows distinct L2 features of TW English can be attributed to prosodic transfer of Mandarin is tempo contrast elicited by syntax-induced narrow focus in sentence layer and loudness range by lexicon-defined stress in word layer. By examining prosodic patterns of word/sentence, similarity analysis suggests that between-speaker similarity is greater when they are in the same speaker group in both word and sentence layer. In other words, L1 and L2 speakers produce respective prosodic patterns with great within-group consistency but their within-group patterns are discriminative to counterpart group by loudness patterns in sentence layer and timing/pitch patterns in word layer. We believe the above study with incorporated linguistic knowledge not only sheds light on better understanding of TW L2 English, but can also be applied CALL system implementation. Future works will include providing prosody evaluation matrix of L2 by word and by sentence with degree measures of similarity and improvement scoring so that L2 learners will become more sensitive to prosody features.

Reference

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*, 529-555.

Benrabah, M. (1997). Word-stress: A source of unintelligibility in English. *IRAL*, XXXV(3), 157-165.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge, England: Cambridge University Press.

Coniam, D. (1999). Voice Recognition Software Accuracy with Second Language Speakers of English. *System*, *27*(1), 49-64.

Coomans, D., & Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*, *136*, 15-27.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1-16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379-397.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393-410.

Grabe, E. & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In Gussenhoven, C. & Warner, N. (eds.) *Papers in Laboratory Phonology 7*, Berlin, Mouton de Gruyter, 515-546.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *8*, 34-52. http://llt.msu.edu Retrieved from

Hirata, Y. (2004). Computer-assisted pronunciation training for native English speakers learning Japanese pitch and duration contrasts. *Computer Assisted Language Learning*, *17*, 357-376.

Lehmann, E. L. (1997). Testing Statistical Hypotheses: The Story of a Book. *Statistical Science*, *12*(1), 48-52.

Moustroufas, N., & Digalakis, V. (2007). Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech and Language*, *21*(1), 219-230.

Munro, M. J. (1995). Nonsegmental factors in foreign accent: ratings of filtered speech. *Studies in Second Language Acquisition*, *17*, 17-34.

Munro, M. J., & Derwing, T. M. (1999), Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*(Supp. 1), 285-310.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, *15*(5), 441-467.

Nguyen, T. A. T., Ingram, J., & Pensalfini, R. (2008). Prosodic transfer in Vietnamese acquisition of English contrastive stress patterns. *Journal of Phonetics*, *36*(1), 158-190.

Pennington, M. (1999). Computer-aided pronunciation pedagogy: Promise,limitations, directions. *Computer Assisted Language Learning*, *12*(5), 427-440.

Scruton, R. (1996). The eclipse of listening. *The New Criterion*, *15*(30), 5-13.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, *25*, 1-24.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1-30.

Tseng, C.-Y., & Su, C.-Y. (2014). Prosodic Differences between Taiwanese L2 and North American L1 speakers—Under-differentiation of Lexical Stress. *Speech Prosody 2014*, Dublin, Ireland.

Tseng, C-Y., Su, C.-Y., & Visceglia, T. (2013). Underdifferentiation of English Lexical Stress Contrasts by L2 Taiwan Speakers. *Slate 2013*, 164-167. Grenoble, France.

Tseng, C.-Y., Cheng, Y.-C., Lee, W.-S. & Huang, F.-L. (2003). Collecting Mandarin speech databases for prosody investigations. *Oriented COCOSDA 2003*. Sentosa, Singapore.

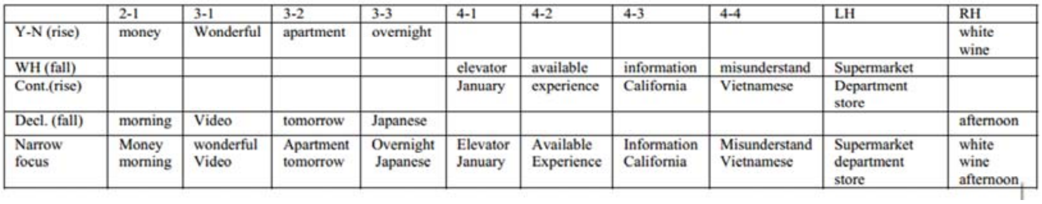
Visceglia, T., Tseng, C. Y., Kondo, M., Meng, H., & Sagisaki, Y. (2009). Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project). *Oriental COCOSDA 2009*. Beijing, China.

Visceglia, T., Tseng, C. Y., Su, Z. Y. & Huang, C. F. (2011). Realization of English Narrow Focus by L1 English and L1 Taiwan Mandarin Speakers. *the 7th International Congress of Phonetic Sciences*. Hong Kong, China.

Visceglia, T., Su, C. Y., & Tseng, C. Y. (2012). Comparison of English Narrow Focus Production by L1 English, Beijing and Taiwan Mandarin Speakers. *Oriental COCOSDA 2012*, 47-51. Macau, China.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2-3), 95-108.

Appendix A. Target words by syllabicity, stress type and experimental condition



	2-1	3-1	3-2	3-3	4-1	4-2	4-3	4-4	LH	RH
Y-N (rise)	money	Wonderful	apartment	overnight						white wine
WH (fall)					elevator	available	information	misunderstand	Supermarket	
Cont.(rise)					January	experience	California	Vietnamese	Department store	
Decl. (fall)	morning	Video	tomorrow	Japanese						afternoon
Narrow focus	Money morning	wonderful Video	Apartment tomorrow	Overnight Japanese	Elevator January	Available Experience	Information California	Misunderstand Vietnamese	Supermarket department store	white wine afternoon

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 77-90

Quantitative Assessment of Cry in Term and Preterm Infants: Long-Time Average Spectrum Analysis

Li-mei Chen[*]

Abstract

Long-time average spectrum (LTAS) was used to analyze the cry utterance of 26 infants under four months old; 16 of them were full-term and the other 10 infants were preterm. The results of first spectral peak (FSP), mean spectral energy (MSE), spectral tilt (ST), high frequency energy (HFE) were used to compare the cry production between term and preterm infants. In addition, cry duration and percent phonation were also compared. According to previous studies, cry production of term and preterm infants show significant differences because immature neurological development of preterm infants. Major findings in this study are: 1) no significant difference in unedited cry duration across groups; 2) no significant difference in percentage of cry utterance across groups; 3) no significant difference in FSP across groups, and higher FSP in term infants; 4) no significant difference in MSE across groups, and a decrease of MSE in both groups over time; 5) no significant difference in ST across groups, and a quicker reduction of energy with larger ST in preterm infants over time; 6) no significant difference in HFE across groups, and a significant decline of HFE over time in both groups. Systematic characterization of infant cry can help to estimate health condition of infants in order to provide appropriate care.

Keywords: Long-time Average Spectrum, Infant Cry, Preterm Infants

1. Introduction

Previous studies show that preterm infants are prone to immaturity of neurological development which leads to their sensitiveness toward pain stimulation, and the greater pain they suffer would reflect on cry production. If a set of distinctive measures can be identified, it might be possible to differentiate infant cries due to organic pathology and cries in the spectrum of normative behavior, including infant colic which is frequently found in infants

[*] Department of Foreign Languages and Literature, National Cheng Kung University, TAIWAN
Email: leemay@mail.ncku.edu.tw

younger than 4 months of age. The measures can thus be used to support doctors' diagnosis to identify if the unknown cries are caused by just infant colic or other more complicated factors in order to provide appropriate care. Cry utterances were analyzed with long-time average spectrum (LTAS) in two groups of newborn infants in this study. Non-partitioned cry episode and the 3 equal-length partitions (P1, P2, P3) were analyzed. First spectral peak, mean spectral energy, spectral tilt, and high frequency energy, as well as unedited cry duration and percent phonation were measured.

Colic strikes infants who are under four months old, and it makes the infants cry in the evening on a daily bases or at the moment of waking up (Lester *etal.*, 1990). The cause of this pain is still unknown (Zeskind & Barr, 1997). Colic occurs when infants are around one month old and it often disappears without a reason when infants are older than three months (Clifford, 2002). It is a universal and commonly-seen phenomenon which is the cause for excessive cry behavior. Though previous studies suggested that higher fundamental frequency and a larger percentage of dysphonation in cry could be found in the pain cries of infants who suffered from colic, no standard acoustic features in cry utterance of infants with colic was established (Zeskind & Barr, 1997). Long-time average spectrum might provide an option to investigate if there are any significant characteristics in the cries of infants with colic.

Though infants are not able to talk, they can express their feelings and emotions through cry, facial expression, and body movement. Diseases are able to be discovered by some characteristics in cry production (Radhika *et al.*, 2012). For example, different pain stimuli would lead to different fundamental frequencies in infant cry utterance (Radhika *et al.*, 2012). If more specific characteristics are found in certain diseases, it would be more effective in prescribing and curing. Sometimes parents can differentiate why their babies cry by their various cry production (Soltis, 2004). As for the way of eliciting cries, Johnston, Stevens, Craig, and Grunau (1993) proposed two different ways: the heel-stick procedure and injection. In this current study, injection was used as the only standard method to elicit cry to avoid any nuances that might caused by the different types of pain stimuli. However, even though there are measures to quantify the pain intensity infants endure, the experience of pain is quite subjective and is not merely related to physiological but also psychological factors (Qiu, 2006). Moreover, since infants use cry to arouse caregivers' attention, it can be expected that infants' cry utterance differs with and without their caregivers around them (Greenet *et al.*, 1995). Usually, the responses from caregivers bring cry behavior to a halt (Green *et al.*, 1995). Cry is thus a way of drawing others' attention to help infants get rid of the uncomfortable situation or meet their needs (LaGasse *et al.*, 2005). Therefore, cry is not only an independent behavior but also plays an important role in social interactions between infants and their caretakers (Green *et al.*, 1995).

Because of the immature development of nervous systems caused by premature birth, cry production of preterm infants is believed to reveal different characteristics from that of term infants whose nervous system is comparatively well-developed. Premature infants were reported to have higher fo in their cry utterance, and it might be due to the immature, and shorter vocal folds (Johnston *et al.*, 1993). Or as Zeskind (1983) stated that high-risk infants were not able to perfectly control their cry production and that they tended to react more intensely towards pain stimuli than did low-risk infants. Infants react differently to the same stimulus pain whether they are healthy or born at risk. However, while some studies reported that preterm infants were more sensitive to pain stimuli, others found that some premature infants had less intense reactions towards pain than normal infants (Qiu, 2006).

The main objective of this current study is to find out how the cry production between term and preterm infants differs from each other. The findings might help in detecting infants' health conditions. Moreover, if the difference of the cry utterance can be systematically characterized, the measurements can be further applied to identify features in neonate cry due to infant colic.

2. Method

2.1 Participants

Previous studies indicated that gender did not lead to significant differences in first spectral peak, mean spectral energy, spectral tilt, and high frequency energy (Goberman & Robb, 1999; Goberman *et al.*, 2008). Therefore, gender was not controlled in this study. There were 26 infant participants; 16 were term infants and the other 10 were preterm infants. The infants were all under four months old for both term infants and preterm infants according to their gestational ages. All of the infants in this study were considered to have normal hearing according to interview with parents.

2.2 Data Collection

For collecting cry utterance of both preterm and term infants, TASCAM wave recorder and RODE uni-directional microphone were used in audio recording. The microphone was held near the infants' mouth. All infants were in the supine position while receiving the injection. This can also avoid influence of different postures in acoustic properties, for example, fundamental frequency (Lin & Green, 2007). The cry production of both groups of infants was recorded during and after they received the injection in the hospital. The pain stimulus was thus the same in both groups of infants.

2.3 Acoustic Analysis

The analysis in this current study was mainly based on Goberman and Robb (1999). A cry episode of infants was defined as the duration of the continuous cry utterance, beginning with the first audible cry utterance after the pain stimulus, and an episode was completed as soon as the infants stopped cry. The non-voiced parts of a cry episode were first edited out in the cry utterance, making a “non-partitioned cry episode” (Goberman & Robb, 1999). In this current study, the inspiratory cry was eliminated, and only the phonatory parts were analyzed. Then, a non-partitioned episode was divided into three partitions with the same length of durations (P1, P2, P3). P1, P2, P3 are regarded as the early, middle, and late sections of the cry episode, respectively, corresponding to the attack, cruise, and subdual phases of a cry episode as suggested by Truby and Lind (1965). Unedited cry duration, percent phonation, first spectral peak, mean spectral energy, spectral tilt, and high frequency energy were measured.

- First spectral peak (FSP): the first amplitude peak across the LTAS display.
- Mean spectral energy (MSE): the mean amplitude value from 0 to 8000 Hz. Average energy from 0 to 8000 Hz - first peak energy
- Spectral tilt (ST): the ratio of energy between 0-1000 Hz, and 1000-5000 Hz. Average energy from 1000 to 5000 Hz / average energy from 0 to1000 Hz
- High frequency energy (HFE): the sum of amplitudes from 5000 to 8000 Hz. Average energy from 5000 to 8000 Hz *(8000-5000) / the bandwidth of LTAS

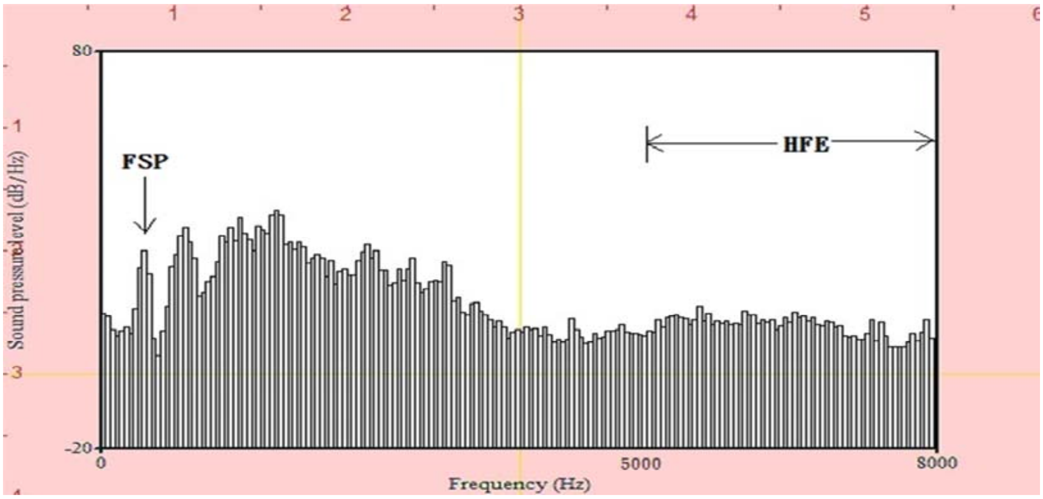


Figure 1. Typical LTAS display showing the location of the first spectral peak (FSP) and high frequency energy (HFE) between 5000Hz and 8000Hz.

3. Results & Discussion

3.1 Unedited Cry Duration

Cry duration reveals respiratory capability, and term infants were thus expected to have longer cry duration than preterm infants (Cacace *et al.*, 1995; Michelsson *et al.*, 1982; Thoden *et al.*, 1985). In this current study, the average duration of cry episodes for the 16 term infants was 42.27s (SD = 31.27s), and for the 10 preterm infants was 36.21s (SD = 30.93s). As expected, term infants had longer average duration of cry episodes. However, a *t* test was performed to examine whether cry duration differed statistically between these two groups, and indicated no significant difference between term and preterm infants, *t*(24) = 0.48, two-tailed, *p* = 0.63. The result is the same as that of Goberman and Robb (1999).

3.2 Percent Phonation

The amount of cries in term infants was reported to be larger than that in preterm infants (Cacace *et al.*, 1995; Michelsson *et al.*, 1982; Thoden *et al.*, 1985). The percentage of cry utterance in a long-term non-partitioned, unedited cry episode was calculated in this current study. However, no significant difference in percent phonation was found between these two groups in this current study. The average percent phonation across the cry episodes of the 16 term infants and the 10 preterm infants was 67.25% (SD = 17.04) and 67% (SD = 13.98) respectively. That is, 67% of the unedited cry episode contained cry production. Like what was found in Goberman and Robb (1999), there was no significant difference across groups in the percentage of cry utterance, *t*(24) = 0.039, two-tailed, *p* = 0.97.

3.3 First Spectral Peak (FSP)

The non-partitioned and partitioned first spectral peak values of the 16 term and the 10 preterm infants are listed in Table 1 and illustrated in Figure 2.

Table 1. First spectral peak from the non-partitioned episodes (NP) and three partitioned cry episodes with equal length (P1, P2, P3) in the term and preterm infants

		FSP (Hz)			
Group		NP	P1	P2	P3
Term	Mean	182.07	135.88	184.79	149.46
	SD	139.06	113.24	142.45	119.31
Preterm	Mean	130.44	104.35	117.40	139.14
	SD	71.74	52.06	67.36	82.11

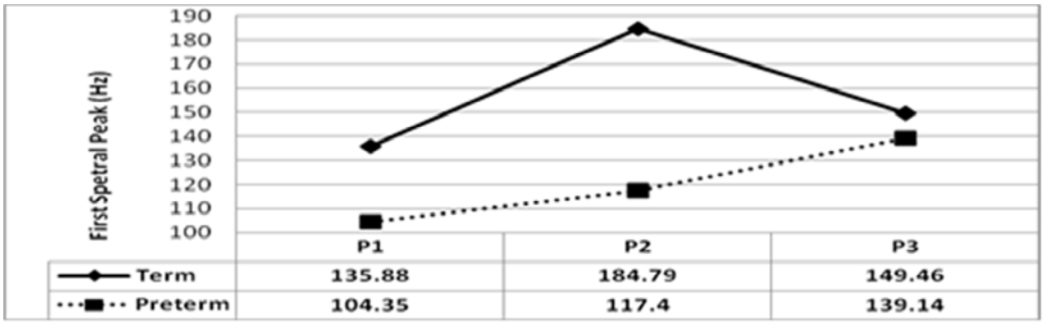


Figure 2. First spectral peak in term and preterm infants over time (P1, P2, and P3 are three equal-length partitioned cry episodes.)

A two-way analysis of variance (ANOVA) was performed to calculate if there were significant differences in FSP values between the two groups (term factor), and whether there was significant variation between the three equal-length cry durations (P1, P2, P3) in each group (partition factor). The results indicated no significant term by partition interaction (p = 0.64), no significant main effect for term status (p = 0.17), and no significant main effect for partition (p = 0.56). Despite the fact that there was no significant difference in statistical tests, from overall observation, term infants demonstrated higher FSP in non-partitioned and the three partitioned episodes than that in preterm infants. Moreover, term and preterm infants displayed different trends of FSP in P1, P2, and P3. Term infants' cry episode involved more distinct phases with decrease of FSP in P3, whereas FSP kept increasing from P1 to P3 in preterm infants.

While the infants were receiving injections, the sharp pain stimulated them and all the infants burst out to cry. According to the previous studies (Johnston *et al*., 1993; Goberman & Robb, 1999), preterm infants were expected to have higher FSP because preterm infants were thought to be more sensitive and would react more intensely to pain. Intensive cry causes the increase of the subglottal pressure and the stiffness of the vocal folds. Premature infants, compared to term infants, were thus reported to have higher *fo* in their cry phonation due to tension of the larynx. However, this difference was not found in this current study. The mean FSP of the term infants turned out to be higher than that of the preterm infants, in both the non-partitioned episode and the three equal-length episodes. Nevertheless, the difference between these two groups was not statistically significant as mentioned above. More data with controlled methodology in future studies can verify the discrepancy of the findings.

Another distinction between these two groups was the changes of FSP across three partitions. The trend of increase followed by decrease of FSP in term infants was not found in preterm infants. FSP kept increasing in preterm infants over time. This distinction was also found in Goberman and Robb (1999), in which FSP decreased significantly in term infants and there was no reduction of FSP in preterm infants.

3.4 Mean Spectral Energy (MSE)

The mean spectral energy values of non-partitioned and partitioned episodes of the 16 term and the 10 preterm infants are shown in Table 2 and Figure 3.

Table 2. Mean spectral energy from the non-partitioned episodes (NP) and three partitioned cry episodes with equal length (P1, P2, P3) in the term and preterm infants

		MSE (dB)			
Group		NP	P1	P2	P3
Term	Mean	19.368	19.982	19.507	14.323
	SD	9.627	9.523	11.158	11.266
Preterm	Mean	22.801	25.201	18.695	15.628
	SD	5.785	6.409	7.963	6.153

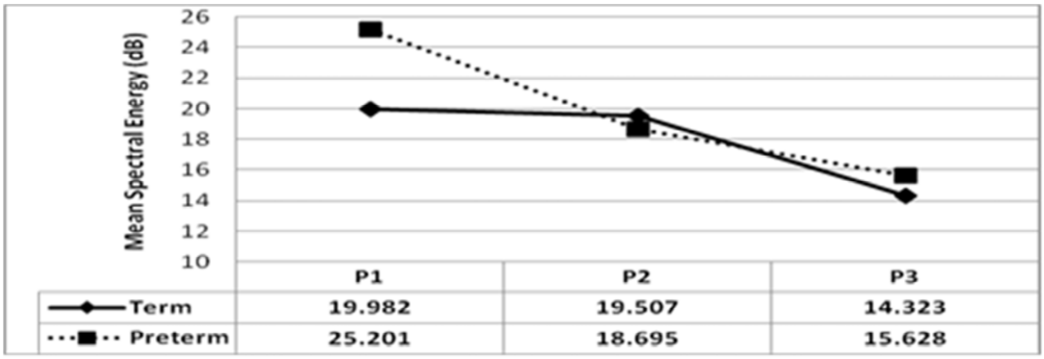


Figure 3. Mean spectral energy in term and preterm infants over time (P1, P2, and P3 are three equal-length partitioned cry episodes.)

A two-way analysis of variance (ANOVA) was performed to investigate if there were significant differences between term and preterm infants (term factor), as well as across P1, P2, and P3 (partition factor) in each group. The results indicated no significant term by partition interaction (p = 0.36). There was a significant main effect for partition (F = 6.47, p = 0.003), yet there was no significant main effect for term, p = 0.52. One-way ANOVA tests were then performed in each group to check the changes of MSE in P1, P2, and P3. In term infants, P2 was significantly higher than P3 (p = 0.029). In preterm infants, P1 showed significantly higher energy than P2 (p =0.042) and P3 (p = 0.012).

MSE refers to the average energy in the frequency range of 0-8000 Hz, which was indicated to correspond to tension of the laryngeal musculature (Fuller & Horii, 1988). In this current study, although no significant difference could be identified, preterm infants showed higher MSE in non-partitioned episode and the three equidurational cry episodes. This shows that during the cry duration, the preterm infants' laryngeal muscles were tighter and they had a more severe reaction toward pain stimulus. The tighter laryngeal muscles suggested a more intense cry production. This finding was also indicated in Goberman and Robb (1999). Moreover, a decrease of MSE over time could be observed in both term and preterm infants. This might suggest that the laryngeal muscles of both groups of infants loosened by phase, especially in preterm infants. There was a sharper decrease of MSE from P1 to P3 in preterm infants. The trend seemed to correspond to the distinct phases in a cry episode indicated in Truby and Lind (1965) with the attack phase (high amplitude) and the cruising phase followed by the subdual phase (the lowest period of stress).

3.5 Spectral Tilt (ST)

The spectral tilt values of non-partitioned and partitioned cry episodes of the two groups are listed in Table 3 and displayed in Figure 4.

Table 3. Spectral tilt from the non-partitioned episodes (NP) and three partitioned cry episodes with equal length (P1, P2, P3) in the term and preterm infants

		ST			
Group		NP	P1	P2	P3
Term	Mean	1.381	2.242	1.423	1.118
	SD	0.307	3.207	0.387	0.300
Preterm	Mean	1.839	1.935	2.811	3.218
	SD	0.685	0.659	2.795	5.326

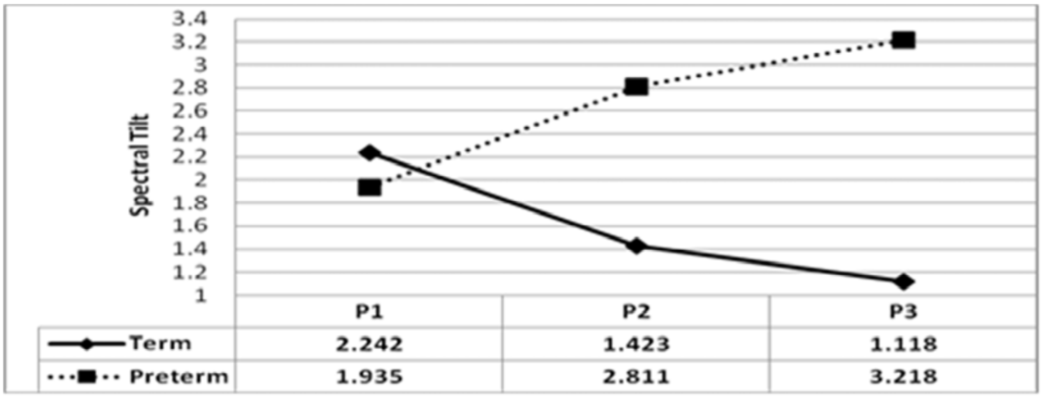


Figure 4. Spectral tilt in term and preterm infants over time (P1, P2, and P3 are three equal-length partitioned cry episodes.)

In order to evaluate if there were significant differences of ST between the two groups and whether there were significant variations between the three equal-length cry durations (P1, P2, P3) in each group, a two-way analysis of variance (ANOVA) was performed. There was no significant term by partition interaction (p = 0.223), no significant main effect for partition (p = 0.994), and no significant main effect for term (p = 0.123). To investigate changes in ST across partitions within each group, separate one-way ANOVA tests were performed for term and preterm infant groups. In term infants, post hoc comparisons identified a significantly higher ST in P2 than in P3 (p = 0.003), but no significant difference in ST across partitions in the preterm infants.

Spectral tilt measures the ratio of low frequency energy and high frequency energy, revealing how quickly the energy declines over time. The quicker the decline is, the larger the ratio. Overall, the term infants showed higher ST values at the onset of cry production which decreased across partitions, whereas the preterm infants had lower ST values at the onset, which increased over time. That is, there was a quicker reduction of energy across partitions in preterm infants. The ST of term infants did not increase over time as mentioned in Goberman and Robb (1999), on the contrary, the increase of ST was found in preterm infants. A higher ST value was reported to be related to hypoadduction of the vocal folds, and a lower ST reflects a hyperadduction of the vocal folds (Mendoza *et al.*, 1996). In this current study, hyperadduction was observed in the decrease of ST in term infants, whereas hypoadduction was observed in the increase of ST in preterm infants.

3.6 High Frequency Energy (HFE)

The high frequency energy values of non-partitioned and partitioned cry episodes of the two groups are listed in Table 4 and illustrated in Figure 5.

Table 4. High frequency energy from the non-partitioned episodes (NP) and three partitioned cry episodes with equal length (P1, P2, P3) in the term and preterm infants

		HFE (dB)			
Group		NP	P1	P2	P3
Term	Mean	1672	1703	1543	1272
	SD	582	552	720	590
Preterm	Mean	1737	1807	1511	1227
	SD	469	514	546	509

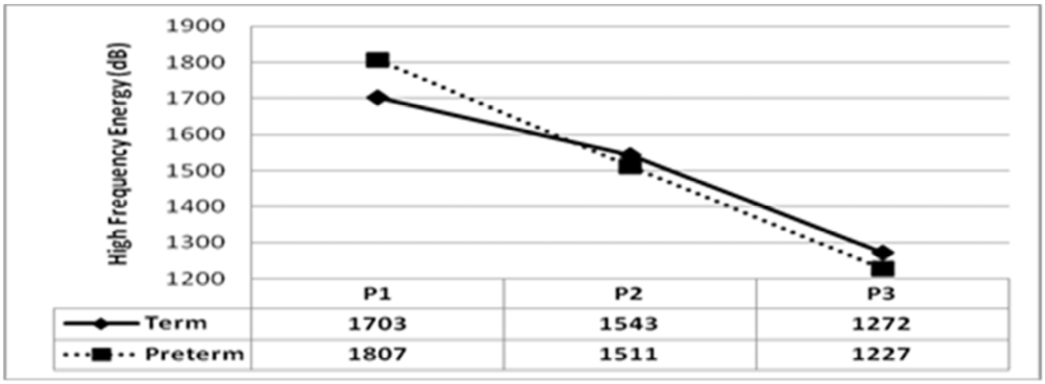


Figure 5. High frequency energy in term and preterm infants over time (P1, P2, and P3 are three equal-length partitioned cry episodes.)

In order to identify if there was significant variation of HFE between term and preterm infants, and whether there were significant variations between the three equal-length cry durations (P1, P2, P3) in each group, a two-way analysis of variance (ANOVA) was performed. No significant term by partition interaction (p = 0.805) was found. Like in Goberman and Robb (1999), there was no main effect for term (p = 0.962). That is, there was no significant difference in HFE across the two groups. There was significant main effect for partitions (F = 8.29, p = 0.001). One-way ANOVA tests were then performed to check changes in HFE across partitions within each group. Significant differences in HFE were found across partitions for both term infants (F = 3.91, p = 0.031) and for preterm infants (F = 4.57, p = 0.025). There was a significantly higher P1 in HFE than P3 in both infant groups (p = 0.029 in term infants, and p = 0.02 in preterm infants). In both groups, HFE decreased over time. The HFE of term infants did not change drastically over time; however, in preterm infants, the HFE showed a steep descent, crossing from 1807 to 1227.

HFE measures the energy in the range of 5000-8000 Hz, which was indicated to be related to the noise elements in phonation (e.g., irregular cry utterance). It was reported that dysphonation in infant cry was very likely related to neurological disorders (Mende, Herzel, & Wermke, 1990). However, no significant difference of HFE between groups was found in this current study. Further studies with more data from both term and preterm infants might verify the correspondence of HFE and its physiological bases.

4. Summary and Suggestion for Future Studies

Cry productions of 16 term infants and 10 preterm infants under 4 months of age were analyzed with long-time average spectrum (LTAS). Major findings were:

1. There was no significant difference between term and preterm infants in cry duration. However, term infants had longer overall cry duration, which corresponded to better

respiratory capability to support phonation;

2. There was no significant difference across groups in the percentage of cry utterance although previous studies indicated that the amount of cries in term infants was larger than that in preterm infants;
3. No significant variation was found between these two groups in FSP. Term infants showed overall higher FSP, which is different from previous findings. Moreover, FSP in term infants involved more distinct phases across three partitions, declining toward the end of cry episode;
4. There was no significant difference of MSE between term and preterm infants. Overall, preterm infants showed higher MSE, which corresponded to tighter laryngeal muscle and intense cry production. A decrease of MSE was found in both groups over time;
5. No significant variation was found between these two groups in ST. There was a quicker reduction of energy with larger ST in preterm infants over time, which revealed hypoadduction of the vocal folds;
6. There was no significant difference in HFE between two groups, and there was a significant decline of HFE over time in both term and preterm infants.

Some of the results in this current study did not match the findings in previous studies. The differences could be due to a few discerning variables. First, although the uni-directional microphone was used in this study, the environmental noises could not be completely controlled because the nurses were required to explain the procedure to the caregivers. Moreover, there was unavoidable overlapping from noises of cry from other infants. Once infant cry overlapped with adults’ voice or cry from other infants, the partitions could no longer be used for further analysis. Second, all the infants receiving injections had their caregivers around. Both term and preterm infants might use more strength in cry, hoping their caretakers would alleviate their pain. This caused inevitable interaction between adults and infants, bringing unexpected disturbance to the results. Third, some caretakers tended to soothe the infants as soon as they started to cry, which would significantly change the natural cry episode since the soothing and consolation from the caretakers might influence their cry production. The infants might feel safe and stopped crying. This might cause incomplete early, middle, and late sections in a cry episode, as Goberman and Robb (1999) mentioned. In further studies, the environmental noise (e.g., from nurses, parents, and other infants around) should be controlled. Moreover, video recording should be implemented in order to identify whether the infants stopped crying spontaneously or their attention was drawn by things around. By controlling disturbance, future study can acquire sufficient data to identify systematic distinction in the pattern of cry production between term and preterm infants. Furthermore, LTAS analysis utilized in this study for cry analysis can be automatically processed and more features can be incorporated in the analysis in future studies.

Acknowledgements

This research was supported by a grant from National Science Council (NSC102-2410-H-006 060) and the Ministry of Education, Taiwan, R.O.C. The Aim for the Top University Project to the National Cheng Kung University. We thank the infant participants and their families for their time and cooperation.

References

Cacace, A., Robb, M., Saxman, J., Risemberg, H., & Koltai, P. (1995). Acoustic features of normal-hearing preterm infant cry. *International Journal of Pediatric Otorhinolaryngology*, *33*, 213-224.

Clifford, T. (2002). *Infant colic: A prospective, community-based examination*. Unpublished doctoral dissertation. The University of Western Ontario, Canada.

Fuller, B., & Horii, Y. (1988). Spectral energy distributions in four types of infant vocalizations. *Journal of Communication Disorders*, *21*, 251-261.

Goberman, A. M. & Robb, M. P. (1999). Acoustic examination of preterm and full-term infant cries: The long-time average spectrum. *Journal of Speech, Language, and Hearing Research*, *42*, 850-861.

Goberman, A. M., Johnson, S., Cannizzaro, M. S., & Robb, M. P. (2008). The effect of positioning on infant cries: Implications for sudden infant death syndrome. *International Journal of Pediatric Otorhinolaryngology*, *72*, 153-165.

Green, J. A., Gustafson, G. E., Irwin, J. R., Kalinowski, L. L., & Wood, R. M. (1995). Infant crying: Acoustics, perception, and communication. *Early Development and Parenting*, *4*(4), 161-175.

Johnston, C., Stevens, B., Craig, K., & Grunau, R. (1993). Developmental changes in pain expression in premature, full-term, two- and four-month-old infants. *Pain*, 52, 201-208.

LaGasse, L. L., Neal, A. R., & Lester, B. M. (2005). Assessment of infant cry: Acoustic cry analysis and parental perception. *Mental Retardation and Development Disabilites*, *11*, 83-93.

Lester, B., Boukydis, C., Gracia-Coll, C. & Hole, W. (1990). Colic for developmentalists. *Infant Mental Health Journal*, *11*(4), 321-333.

Lin, H. C., & Green, J. A. (2007). Effects of posture on newborn crying. *Infancy*, *11*(2), 175-189.

Mende, W., Herzel, H., & Wermke, K. (1990). Bifurcations and chaos in newborn infant cries. *Physics Letters-A*, *145*, 418-424.

Mendoza, E., Munoz, J., & Naranjo, N. (1996). The longtime average spectrum as a measure of voice stability. *Folia Phoniatrica*, *48*, 57-64.

Michelsson, K., Raes, J., Thoden, C., & Wasz-Hockert, O. (1982). Sound spectrographic cry analysis in neonatal diagnostics: An evaluative study. *Journal of Phonetics*, *10*, 79-88.

Qiu, J. (2006). Does it hurt? *Nature*, *444*, 143-145.

Radhika, R. L., Chandralingam, S., Anjaneyulu, T. & Satyanarayana, K. (2012). A suggestive diagnostic technique for early identification of acyanotic heart disorders from infant's cry. *International Journal of Electrical and Electronics*, *1*(3), 32-38.

Soltis, J. (2004). The signal functions of early infant crying. *Behavioral and Brain Sciences*, *27*, 443-458.

Thoden, C., Jarvenpaa, A., & Michelsson, K. (1985). Sound spectrographic cry analysis of pain cry in prematures. In B. Lester & C. Boukydis (Eds.), *Infant crying: Theoretical and research perspectives* (pp. 105-118). New York: Plenum Press.

Truby, H., & Lind, J. (1965). Cry motions of the newborn infant. In J. Lind (Ed.), *Acta paediatrica Scandanavica: Newborn infant cry* (Suppl.163), 7-58.

Zeskind, P. & Barr, R. (1997). Acoustic characteristics of naturally occurring cries of infants with "colic". *Child Language Development*, 68, 394-403.

Zeskind, P. (1983). Production and spectral analysis of neonatal crying and its relation to other biobehavioral systems in the infant at-risk. In T. Field & A. Sostek (Eds.), *Infants born at-risk: Physiological and perceptual processes*. New York: Grune & Stratton.

The individuals listed below are reviewers of this journal during the year of 2014. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Guo-Wei Bian

Jing-Shin Chang

Tao-Hsing Chang

Yu-Yun Chang

Tai-Shih Chi

Hong-Jie Dai

Jeih-weih Hung

Yun-ju Hsueh

Mike Tian-Jian Jiang

Bor-Shen Lin

Suguru Matsuyoshi

Samir Rahman

Yu Tsao

Jia-Ching Wang

Jui-Feng Yeh

2014 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 19

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2014.

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of paper or other item, and its location, specified by the publication volume, number, and inclusive pages. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication volume, number, and inclusive pages.

AUTHOR INDEX

SUBJECT INDEX

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member： NT$ 1,000 (US$50.-)
group membership： NT$20,000 (US$1,000.-)
life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#：______________

Name：______________________________ Date of Birth：_____________

Country of Residence：_________________ Province/State：_________________

Passport No.：________________________ Sex: _____________________

Education(highest degree obtained)：____________________________________

Work Experience：____

____

Present Occupation：____

Address：____

____

Email Add：____

Tel. No：_________________________ Fax No：_______________________

Membership Category：□ Regular Member □ Life Member

Date：_____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member	：	US$ 50.- （NT$ 1,000）
Life Member	：	US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502　　傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw　網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號			（由本會填寫）
姓　　名		性別		出生日期	年　　月　　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
				申請人:	（簽章）
			中　華　民　國	年	月　　日

審查結果:

1. 年費：

終身會員：	10,000.-
個人會員：	1,000.-
學生會員：	500.-（限國內學生）
團體會員：	20,000.-

2. 連絡處：

地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
PAYMENT FORM

Name: ____________________________________(Please print) Date: _______________

Please debit my credit card as follows: US$ _____________

□ VISA CARD □ MASTER CARD □ JCB CARD Issue Bank:___________________

Card No.: ________ -_________-_________ -_________ Exp. Date:____________(M/Y)

3-digit code: _________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: ____________________________________

Phone No.: __________________________E-mail: _____________________________________

Address: ____

PAYMENT FOR

US$ _____ □ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: ______

US$ _____ □ Journal of Information Science and Engineering (JISE)

Quantity Wanted: ______

US$ _____ □ Publications:_________________________________

US$ _____ □ Text Corpora: _______________________________

US$ _____ □ Speech Corpora:______________________________

US$ _____ □ Others: _____________________________________

US$ _________ □ Membership Fees □ Life Membership □ New Membership □Renew

US$ _________ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP
% IIS, Academia Sinica
Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會
信用卡付款單

姓名: _____________________________(請以正楷書寫)　　日期:：_____________________

卡別：❑ VISA CARD　　❑ MASTER CARD ❑ JCB CARD　　發卡銀行：______________

信用卡號：__________-__________-__________-__________　　有效日期：__________(m/y)

卡片後三碼：______________（卡片背面簽名欄上數字後三碼）

持卡人簽名：____(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：_____________________________E-mail：__________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________ ❑ 中文計算語言學期刊(IJCLCLP) ________________

NT$__________ ❑ Journal of Information Science and Engineering (JISE)

NT$__________ ❑ 中研院詞庫小組技術報告______________________

NT$__________ ❑ 文字語料庫 _____________________________________

NT$__________ ❑ 語音資料庫 _____________________________________

NT$__________ ❑ 光華雜誌語料庫1976~2010

NT$__________ ❑ 中文資訊檢索標竿測試集/文件集

NT$__________ ❑ 會員年費：❑續會　　　❑新會員　　　❑終身會員

NT$__________ ❑ 其他: ______________________________________

NT$_______________ = 合計

填妥後請傳真至 02-27881638 或郵寄至:

11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收

E-mail: aclclp@hp.iis.sinica.edu.tw

Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

			AIR	AIR		
		Surface	(US&EURP)	(ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： □ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息爲本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期） 年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人：　黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：　______________________　收據抬頭：___________________________

地　　址：____

電　　話：________________________　E-mail：__________________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.

1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.

2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.

3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.

4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript

6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.

7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

> Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

> Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).

Please visit the following websites for details.

(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.19 no.4 December 2014

Contents

Special Issue Articles:

Selected Papers from ROCLING XXVI

Papers

Regular Issue Article:

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing