中華民國計算語言學學會
通訊 ACLCLP
發行人：陳信希
主　編：王新民
執行編輯：黃　琪
The Association for Computational Linguistics and Chinese Language Processing
第二十卷第一期
E-Mail：aclclp@hp.iis.sinica.edu.tw
Website：http://www.aclclp.org.tw
傳 真：(02)2788-1638
地 址：台北市研究院路二段128號中研院資訊所
電 話：(02)2788-3799 ext:1502
劃 撥：19166251

本 期 要 目

《中文計算語言學期刊》通告

《中文計算語言學期刊》原任總編輯成功大學吳宗憲教授已於 2008 年卸任，自 2009 年起總編輯與執行編輯的工作，將由臺灣師範大學曾元顯教授及臺灣大學陳光華教授兩位教授接任，總編輯任期四年(2009~2012)。

2009~2012 年諮詢委員 (Advisory Board)：

Jason S. Chang
National Tsing Hua University, Taiwan

Hsin-Hsi Chen
National Taiwan University, Taiwan

Keh-Jiann Chen
Academia Sinica, Taiwan

Sin-Horng Chen
National Chiao Tung University, Taiwan

Ching-Chun Hsieh
Academia Sinica, Taiwan

Chu-Ren Huang
The Hong Kong Polytechnic University, H.K.

Lin-Shan Lee
National Taiwan University, Taiwan

Jian-Yun Nie
University of Montreal, Canada

Richard Sproat
University of Illinois at Urbana-Champaign, USA

Keh-Yih Su
Behavior Design Corporation, Taiwan

Chiu-Yu Tseng
Academia Sinica, Taiwan

Hsiao-Chuan Wang
National Tsing Hua University, Taiwan

Jhing-Fa Wang
National Cheng Kung University, Taiwan

Kam-Fai Wong
The Chinese University of Hong Kong, H.K.

Chung-Hsien Wu
National Cheng Kung University, Taiwan

獎助學生出席國際會議

獎助會議：

1.COLING, 2. ACL, 3. ACM SIGIR, 4. ICASSP

獎助說明：

1. 申請人須同時具備下列資格：
   (1) 被接受論文之第一作者(指導教授不計)。
   (2) 本會會員。
   (3) 投稿時為國內在學學生。
2. 獎助金額：由審查委員會依地區別及論文等級審定獎助金額，每名獎助金額上限為美金 1,000 元
3. 獎助名額：每個會議補助一～二名。

申請辦法：

1. 申請期限：論文被接受發佈日起兩週內提出。
2. 申請手續：申請人需將論文接受函、審查意見、學生證、論文全文及申請書（請至本會網頁下載）等相關資料郵寄至本會秘書處。

受獎助人義務：

1. 出席會議發表論文。
2. 論文全文必須以書面同意投稿至本會期刊。
3. 代學會攜去宣傳品及帶回相關資料。

2009 APSIPA Annual Summit and Conference APSIPA ASC 2009
October 4 - 7, 2009
Sapporo Convention Center, Sapporo, Japan
Call for Papers

2009 APSIPA Annual Summit and Conference is the inaugural event supported by the Asia-Pacific Signal and Information Processing Association (APSIPA).

The APSIPA is a new association and it promotes all aspects of research and education on signal processing, information technology, and communications. The field of interest of APSIPA concerns all aspects of signals and information including processing, recognition, classification, communications, networking, computing, system design, security, implementation, and technology with applications to scientific, engineering, and social areas.

The topics for regular sessions include, but are not limited to:

1.	Signal Processing Track	3.	Computer/Information Track
1.1	Audio, speech, and language processing	3.1	Database and data mining
1.2	Image, video, and multimedia signal processing	3.2	Ubiquitous and mobile computing
1.3	Information forensics and security	3.3	Computer Graphics and visualization
1.4	Signal processing for communications	3.4	Computer vision and pattern recognition
1.5	Signal processing theory and methods	4.	Circuits and Systems/VLSI Track
2.	Communications Track	4.1	Biomedical circuits and systems
2.1	Communication and information theory	4.2	Nanoelectronics and gigascale systems
2.2	Information and network security	4.3	Intelligent systems and applications
2.3	Wireless communications and networking	4.4	VLSI systems and applications
2.4	Standards and emerging technology	4.5	Embedded systems

Sapporo and Conference Venue:

One of many nice cities in Japan, Sapporo is always recognized as a quite beautiful and well-organized city. With a population of 1,800,000, Hokkaido's largest/capital city, Sapporo, is fully serviced by a network of subway, streetcar, and bus lines connecting to its full compliment of hotel accommodations. Sapporo has already played host to international meetings, sports events, and academic societies. There are a lot of flights from/to Tokyo, Nagoya, Osaka et. al and overseas cities. With all the amenities of a major city yet in balance with its natural surroundings, this beautiful northern capital, Sapporo, is well-equipped to offer a new generation of conventions.

Important Due Dates and Author's Schedule:

Proposals for Special Session:	March 1, 2009
Proposals for Forum, Panel and Tutorial Sessions:	March 20, 2009
Deadline for Submission of Full-Papers:	March 31, 2009
Notification of Acceptance:	July 1, 2009
Deadline for Submission of Camera Ready Papers:	August 1, 2009
Tutorial Session (Hokkaido University) Date:	October 4, 2009
Conference dates:	October 5 - 7, 2009

Submission of Papers:

Prospective authors are invited to submit either long papers, up to 10 pages in length, or short papers up to four pages in length, where long papers will be for the single-track oral presentation and short papers will be mostly for poster presentation. The conference proceedings will be published, available, and maintained at the APSIPA website.

Detail Information:

WEB Site : http://www.gcoe.ist.hokudai.ac.jp/apsipa2009/

Oriental-COCOSDA 國際學術會議十週年活動

鄭秋豫

中央研究院語言學研究所

本文的主旨，是向國內計算語言學同行介紹 Oriental-COCOSDA 這個區域性國際學術組織（http://www.milab.is.tsukuba.ac.jp/o-cocosda）以及該組織最重要的活動：年度國際學術會議 Oriental-COCOSDA（簡稱 O-COCOSDA）十週年慶。本文將著重報導 Oriental-COCOSDA 這個學術組織的主要任務、活動，以及在台灣學界從該學會成立以來，所扮演的角色及地位。並希望經由這篇報導，能獲得國內更多的同行、後進的注意，積極參與 Oriental-COCOSDA 及類似學術組織的活動與服務。

一、Oriental-COCOSDA 學會沿革

Oriental-COCOSDA 是國際學術組織 COCOSDA 下屬的亞洲分支組織，正式成立於 1997 年。母組織 COCOSDA 爲「International *Co*mmittee for *Co*-ordination and *S*tandardization of Speech *Da*tabases and Assessment Techniques 口語語音資料庫協調暨標準化國際委員會（網址http://www.cocosda.org ）」的縮寫，成立於 1991 年。當時語音科技開發的相關研究，進入新的階段，以資料驅動(data driven)並統計測試爲基礎的創新研究方法，成爲研究主流。資料驅動研究方法不可或缺的配套，就是先建立口語語音資料庫及評估標準並將資源分享。但當時除了歐美的先進國家外，許多國家都還認定資料收集不是研究工作，以建置資料庫申請經費非常困難。而口語資料庫的建置在資料的量、建構標註系統等方面，比文字資料庫更困難，建立跨國、跨語言的語音科技評估的標準與平台，協調跨國口語處理研究計畫，評估的標準等原則性的指標都還不存在，有關資料庫的文獻也極度不足，因此一個國際性建立的協調組織便因需要而誕生。COCOSDA 成立時，非歐美語系的語言，只擴及中文和日文而已。

COCOSDA 的組織包括一位全球性召集人及 11 個區域，召集人總理會務，各區域代表則負責在所屬區域推廣 COCOSDA 相關業務，並定期向 COCOSDA 報告會務發展情況。每區域二人代表，包括英國、美國、法國、德國、荷蘭、義大利、南非、日本、台灣及近年才加入的中國與希臘，共計 22 人。COCOSDA 成立以來，不收任何會費，所有代表均屬服務性質。由於學會本身沒有經費，成立至今的主要活動形式，大致不出每年於大型國際學術會議 Eurospeech，ICASSP 或 LREC（Language Resource and Evaluation Conference）時舉行代表會議，其餘會務均以電子郵件方式進行。台灣地區代表爲清華大學電機系王小川教授及台大

電機系李琳山教授。首位召集人爲日本筑波大學教授 Shuichi Itahashi，鄭秋豫教授自 2006 年起接任該組織召集人至今。

COCOSDA 成立至今十七年來，下屬區域性分支組織僅有 Euro-COCOSDA 及 Oriental-COCOSDA，直到 2008 年才開始籌組 African-COCOSDA，會務的拓展並不如預期，自 2006 年起才又比較活躍（詳見下文）。

Oriental-COCOSDA 於 1993 年起由日本東京大學教授藤崎博也 Hiroya Fujisaki 發起，當時他鑑於亞洲的地區語言與文字與歐美國家有許多語言類型方面的差異，口語語音資料庫的建置及平台規劃的重要性更大，語言問題也非以歐美語言爲主平台所能處理。亞洲地區語言的主要特性包括：1. 語言多元性─本地區的語言及語族極爲繁多。亞洲地區的總人口約 38 億人（資料來源：維基百科，http://en.wikipedia.org/wiki/ ），語言類型方面，較歐美爲多，包括藏緬語系（中國境內方言、越南、泰國、緬甸、寮國等）、阿爾泰語系（日語、韓語）、南島語系（台灣、菲律賓、馬來西亞、印尼、東印度群島至夏威夷）。亞洲地區遍佈聲調語言，含中國境內各方言、越南、泰國、緬甸、寮國等。以聲調語言爲母語的人口約 15 億 5 千 2 百萬人，大幅超過非洲地區 8 億 4 千萬人（資料來源：維基百科，http://en.wikipedia.org/wiki/ ）。2. 文字多元性─本地區的文字包括中文的漢字、韓國的音節拼音、日本的假名、泰文、及印度的 22 種官方語言及文字。未有文字書寫系統的語言也甚多，遍佈南島語系（含台灣、菲律賓、馬來西亞、印尼直到大洋洲的南島語系及分支等），藏緬語系（含川西與西藏、蒙古、越南、泰國、緬甸、寮國、尼泊爾等），南亞印度語系（含印度境內非官方語言、巴基斯坦、斯里蘭卡等）。3. 文字系統多元性─如中文是獨有的音節文字漢字、日本是漢字與假名並列，蒙古、越南、泰國、緬甸、寮國、尼泊爾、菲律賓、馬來西亞、印尼均爲拼音文字。4. 拼音系統多元性─本地區的拼音系統繁多，如漢語拼音、日本的羅馬字等。

此外，九〇年代初期的亞洲地區，除日本外，經濟成長也大幅落後先進國家，語音科技方面的研究，有許多國家尚未起步，促進亞洲各國的語音科技相關研究，也刻不容緩。推動口語語音資料庫的建置及語音科技的開發比歐美更重要，而且亞洲地區語言的問題，恐由亞洲人自己來處理才最爲恰當。COCOSDA 的成員中以歐美爲主，以歐美語言爲基礎的語音平台未必能爲亞洲語言所用。

Fujisaki 教授在 Eurospeech93 會議期間，以上述訴求尋求亞洲地區研究語音科技開發學者的支持，建議成立 COCOSDA 在亞洲地區的分支組織 Oriental-COCOSDA，立刻獲得與會亞洲學者的一致認同。經過三年的籌畫，Oriental-COCOSDA 學會於 1997 年 3 月 21 日正式成立於香港，至今十一年。

二、Oriental-COCOSDA 組織及任務

學會的主要任務，除了交換意見、共享資訊、討論亞洲地區口語語料庫的建立、使用、評估、傳播及推廣外，並自 1998 年起每年定期舉辦同名的區域性國際學術會議 Oriental-COCOSDA—International Conference on Speech Database and Assessment，主要會務是在亞洲地區推廣 COCOSDA 的基本理念，及推動口語語料庫的建置、評估測試的平台，及推廣語音科技相關研究。

Oriental-COCOSDA 組織分爲諮詢委員會與區域代表二部份，設召集人一位，總理會務，向亞洲地區推廣本組織、吸收新成員、協調各國相關組織及代表，並確保每年由各國競標，輪流舉辦學術會議。諮詢委員三位，指導及監督會務；區域代表則總理區域性相關業務，在其所屬地區推廣相關業務，每年收集整理該地區與 Oriental-COCOSDA 任務相關的活動，於年度會議時報告，並負責輪流舉辦年度國際學術會議。召集人一職於學會成立之時，由當時四個區域代表公推 Shuichi Itahashi 教授擔任，他任此職至 2005 年底，在任內大力推展會務，建立了學會現今的規模。Itahashi 教授於 2005 年退休，召集人一職由會員於 2005 年十二月在印尼雅加達舉辦的 Oriental-COCOSDA 會議時，推舉中央研究院語言學研究所研究員鄭秋豫教授擔任。自 2006 年一月起，擔任召集人一職至今。

Oriental-COCOSDA 諮詢委員有三位，分別爲發起人 Hiroya Fujisaki 教授, 中國科學院聲學所研究員張家騄教授，及韓國的 Sougil Ann 教授。十一年來，區域性代表已由 1997 年成立時的四個地區增至十四個國家或地區，按照國家的字母順序分別爲：1.中國—李愛軍、鄭方；2.香港—P. C. Ching、李丹；3.印度：—Shyam S. Agrawal, K. Samudravijaya； 4.印尼—Arry Akhmad Arman, Hamman Rizza; 5.日本—Nick Campbell, Shuichi Itahashi, Satoshi Nakamura, 6.韓國—Yong-Ju Lee, Yung-Hwan Oh; 7. 馬來西亞—Ambigapathy Pandian, Syed Ariff, Zurai Mohd Don; 8.蒙古—Dahtseren Erdenebat, Dawa Idomuso (蒙古語代表); 9. 尼泊爾—Bhim Regmi; 10.巴基斯坦—Sarmad Hussain; 11.新加坡—賴金定、李海洲；12.台灣—王小川、李琳山；13.泰國—Virach Sornlertlamvanich, Thanaruk Theeramunkong；14. 越南—Luong Chi Mai。鄭秋豫接任召集人後，2006 年至 2008 年加入的國家有四個，分別爲馬來西亞、尼泊爾、巴基斯坦、越南；2009 年開始，菲律賓亦將加入。

三、代表性的學術活動：年度國際學術會議 Oriental-COCOSDA—International Conference on Speech Database and Assessment

自 1998 年起，學會每年召開區域性國際學術會議，由各國家或區域輪流舉辦的同名的國際性會議 Oriental-COCOSDA—International Conference on Speech

Database and Assessment，至今不曾間斷，是 COCOSDA 學會中最具學術活力及能見度的組織。

每年，主辦單位接受相關論文，並有健全的審查制度，每篇論文送二位審查人匿名審查。第一屆會議於 1998 年由日本竺波大學舉辦，規模五十餘人，論文約二十餘篇。隨後分別於 1999 年在台北、2000 年在北京、2001 年在韓國、2002 年在泰國、2003 年在新加坡、2004 年在印度、2005 年在印尼、2006 年在馬來西亞、2007 年在越南、2008 年在日本，定期舉行會議。如今會議論文規模已逾三十餘篇，與會人士約六十五人左右，不但會議規模穩定成長，且不斷有新成員加入。2009 年會議，現已預定由新疆大學在烏魯木齊主辦；2010 年的會議則將由尼泊爾語言學會主辦，地點爲加德滿都。

會議特色除口頭論文發表外，每次大會必同時將「區域報告」列入議程，每個國家或地區必須提出當年的相關活動報告。所以，經由這些區域報告，可以得知每個地區的工作重點以及口語語料庫的數量及特性，區域間彼此也有競爭性，互相激勵。

四、慶祝 Oriental-COCOSDA 學術會議十週年活動

1. Oriental-COCOSDA 2008 會議規模擴大

Oriental-COCOSDA 2008 國際會議，由日本 NICT(National Institute of Communication Technology)主辦，於 2008 年 11 月 25-27 在京都舉行。適逢成立十週年，除了在會期前後相關慶祝活動外，此次會議還擴大會議規模，增加了論文數量及與會國家數。除了原有的口頭論文場次外，還增加了壁報論文場次，總計此次會議共發表口頭論文 31 篇，壁報論文 14 篇，其中台灣地區佔有 4 篇論文；參與的學者除 Oriental-COCOSDA 的 13 個國家外，今年並有美國、法國、不丹的參與，共有 16 個國家。

2. 出版專書

爲慶祝 Oriental-COCOSDA 活動十週年，Oriental-COCOSDA 2007 會員大會決議：集結十年來的論文，由作者更新後，預訂於 2009 年出版專書一冊，爲 1998 至 2007 這十年間亞洲地區的口語語音資料庫及相關語音科技發展做總整理，提供一本跨亞洲語言的工具書，也爲學術發展留下歷史性的紀錄。書名定爲：*Resources and Standards of Spoken Language Systems-Advances in Oriental Spoken Language Processing*，由先後二任召集人 Shuichi Itahashi 教授和鄭秋豫擔任主編。本書將於 2009 年由德國 Springer 公司出版及發行。

全書包含 9 個章節與 1 章附錄。第 1 章〈導論〉，將詳述 Oriental-COCOSDA 的學術任務；第 2 章〈東亞語言概論〉，介紹漢語、印度語、印尼語、日語、韓

語、馬來語、蒙古語、尼泊爾語和越南語的音系學、語音學與韻律特性；第 3 章〈資料中心與語料庫〉，簡述若干管理語料庫的學術機構的組織與活動；第 4 章〈東亞語言語料庫〉，按照語音類型介紹東亞各國所蒐集的語料庫，如：獨白、對話或會話、會議語音、車內語音、語言學習者語音、情緒或表達性語音、韻律語料庫、多語語料庫、同步口譯語料庫、多模組語料庫、聲音與雜訊語料庫與編碼轉換等；第 5 章〈語音合成與辨識系統的評估〉，概括地說明語音合成與辨識系統性能評估方法（performance evaluation method）的準則；第 6 章〈標註與標記〉，摘要式地說明音位標記、韻律標記、情緒標記、分段(segmentation)與對齊；第 7 章〈語音軟體工具〉，介紹不同的語音處理、語音辨識、語音合成與韻律模型的軟體工具，以及軟體工具間的比較；第 8 章〈文字轉寫（orthographic）與 Romanization 系統〉，探討日語、韓語與泰語的文字轉寫和 Romanization 系統，同時也提及北印度、標準馬來語與蒙古語從音素到音位的轉換；第 9 章〈結論〉，綜述 Oriental-COCOSDA 的學術活動與未來計畫;〈附錄〉則爲 Oriental-COCOSDA 的歷史性敍述，包括：歷屆召集人、代表、工作坊與照片集等。

3. Oriental-COCOSDA 2008 會議前後相關慶祝活動

A. A-Star (The Asian Speech Translation Advanced Research)年度會議

A-Star(http://www.slc.atr.jp/AStar/)是一跨國性研究計畫，由日本 ATR/NICT 資深學者 Satoshi Nakamura 教授主持，邀集亞洲地區頂尖學者共九個單位參與，除日本的 ATR/NICT 外，尚有 JST（Yoshio Kitamura 博士）、台灣台大電機系（李琳山教授），中國科學院自動化科學所（徐波教授），新加坡 I2R（李海洲教授）、印尼 BPPT（Hammam Riza 教授），印度 CDAC（Shyam S. Agrawal 教授），韓國 ETRI(Jun Park 教授)，泰國 NECTEC(Chai Wutiwitatchai 博士)，越南科學院 IOIT（Luong Chi Mai 博士）。11 月 24 日舉行年度會議，當日共有 19 位學者參加，鄭秋豫教授以 Oriental-COCOSDA 召集人的身份應邀出席，提供跨國研究平台的參考意見，由於適逢 Oriental-COCOSDA 成立十週年，因此增加此項周邊活動，以示慶祝。

B. International Symposium on Asian Speech Resources

NII (National Institute of Informatics)爲慶祝 Oriental-COCOSDA 成立 10 週年，在東京舉辦了 International Symposium on Asian Speech Resources，安排了七場特邀報告，其中六個報告是由出席 Oriental-COCOSDA 2008 的國際學者，包括來自印度、中國、泰國、韓國、印尼、越南的學者進行報告，與會人員主要是日籍學者。

五、台灣學者參與 Oriental-COCOSDA 之情況

Oriental-COCOSDA 學會一路走來，台灣學者積極參與，扮演了不可或缺的角色，十餘年來參與最積極的三位學者李琳山、王小川、鄭秋豫教授，與日本學界有著良好互動關係。三位分工，推動亞洲地區及全球的跨語言、跨國口語語音資料庫開發、評估標準的建立，資源分享規劃，並因此將中文語音科技開發及相關研究成果推向國際社群。2005 年執行國科會跨國研究計畫「新一代語音科學與技術－由基礎到應用」之初，研究團隊前往日本參訪，與日本的研究同行進行較深入學術交流，也因同是參與 Oriental-COCOSDA 社群活動成員，而與日本學界有極深的淵源，Oriental-COCOSDA 學會即扮演了關鍵性的角色。而 Oriental-COCOSDA 學會所代表的研究社群，也因此與台灣學界有了更多的互動。印尼的"Ministry of Communication And Information Technology Republic of Indonesia" 於 2008 年 8 月 25 日舉辦 "Language Computing Localization" 研討會，邀請鄭秋豫擔任主題報告講員，便是一例。

六、結語

十一年前 Oriental-COCOSDA 成立之時，亞洲地區的語音資料庫極其有限，語音科技開發落後歐美先進國家。1998 年 Oriental-COCOSDA 舉辦第一次年度學術會議時，主要的議題是如何推廣語音資料庫的建置、分享經驗、提供共同平台的考量，以促進語音科技研究的進展及語音科技產品的開發。自此，這個會議成爲亞洲地區一個分享經驗、交換訊息的論壇。作爲一個學術組織，讓 Oriental-COCOSDA 不收會費，也沒有經費，所有的工作都是志願性，各國輪流舉辦年度會議都需自籌經費，卻仍爭相舉辦年度性學術會議的主因是：爲促進國內語音科技產業的蓬勃發展，需要一個具國際視野的學術論壇，交換訊息、開展與他國共享平台及學術資源的可能性。這個區域性的學術社群裡，國民所得的差距甚大，國民所得較低國家對經費籌措極其費心，參加大型以先進國家研究成果爲主的國際學術會議，光是旅運註冊便所費不貲，因此，這些國家寧願將資源投注於建置語音資料庫、發展自己的語音科技及參與區域性的學術活動。從十年前舉行第一次會議以來，亞洲地區的每一個國家都建置了各種語音資料庫、開發了語音科技產業。尤其是文化多元、官方語言多樣的國家如印度和印尼，語音科技產業更獲國家政策的加重考量。

值得欣慰的是，經由連續十年未曾間斷的學術會議，這些國家的研究並未與國際學術走向脫節，水準也不差；各國語音相關研究自然地產生良性競爭與比較，因此促進了各國的相關研究及成立語音聯盟的作法，如韓國、中國、泰國。此外，Oriental-COCOSDA 也促成了跨國研究團隊的成立，如 A-STAR、AESOP，

都是極具前瞻性、共享資源的國際合作研究計畫。

台灣學界從 Oriental-COCOSDA 成立以來，由於少數學者長期的積極參與國際學術事務，並以其研究成就、社群服務獲得國際間極大的肯定；但國內學術界參與國際學術事務的趨向，或仍以歐美先進國家爲主、或以兩岸交流爲主，區域性的參與相對較少。反觀日本學術界全方位的參與，無論在歐美、兩岸及區域性，不但無役不與，而且在不同領域都居領導地位─有領導性的學者，也參與領導性的事務─甚至比歐洲學術界都積極，適足以我們效法與學習的對象。本世紀亞洲的興起，是全球的共識，希望國內更多的同行能更重視區域性的學術團體及活動，積極參與 Oriental-COCOSDA 及類似學術組織的活動與服務，進而提升能見度，加強影響力。

TermMine[1] - 辭彙資訊之資料探勘與應用

吳鑑城、張俊盛
國立清華大學資訊工程系

研究背景介紹

詞彙翻譯的需求普遍見於生活各處，如外語書籍之閱讀、文件翻譯、語言學習以及跨語言資訊檢索(Cross Language Information Retrieval)等等。而此一需求常須透過學者專家殫精竭智地爲大眾編撰各類型辭典來獲得滿足。然而身處資訊爆炸的時代，每分每秒都可能有新詞彙的出現。尤其，不同的語言以及專業領域所使用的辭彙有時呈現互相重疊卻又另有解釋的現象，造成辭彙具有一詞多義(一詞多譯)，更增添詞彙涵蓋上的困難。此一現象，導致耗費大量人力、物力以及時間所製作的“最新辭典”仍然有許多未收錄的新詞彙（out-of-vocabulary, OOV）。

所幸，越來越多的雙語甚至多語文件出現（如雙語雜誌等），加上現行統計式機器翻譯技術日益成熟，可從這些平行語料(parallel corpora)中，自動地擷取翻譯資訊，有效地協助辭典編撰的過程。但美中不足的是，平行語料的數量仍嫌不足或過於特定領域（如官方文件），使得辭典的涵蓋範圍依舊有所限制。

而近年來，由於網路的盛行，全球資訊網(WWW)上有著龐大且多樣的資訊。因此也有研究者構思如何透過這豐富的資源進行詞彙翻譯的驗證或是擷取。Nagata et al. (2001) 認爲可從包含多種語言的網頁（見圖一），擷取互爲翻譯的日英詞彙。根據他們所做的實驗顯示，各領域之詞彙以及其翻譯，約有 17%~60% 的比例，以不同型式出現於網頁之中。他們透過詞彙及翻譯之間距離的特徵性，擷取互爲翻譯的日英詞彙，並且建議未來可參考資訊檢索(Information Retrieval)領域中常用之樣式(pattern)，來提高擷取之效能。

[1] 系統網址：http://termmine.cs.nthu.edu.tw/TermMine/

圖一：包含日英詞彙翻譯資訊之網頁資訊

樣式	出現次數	準確率	範例
FE	3036	28.1%	亞特拉斯 ATLAS
EF	1925	45.9%	Elton John 艾爾頓強
E(F	1485	59.7%	Austria(奧地利
F（E	1251	71.2%	亞特拉斯（Atlas
F(E	361	74.6%	亞特拉斯(Atlas
F.E	203	76.5%	Peter Pan. 小飛俠
EwF	197	78.3%	加州 Northern California
E,F	153	79.7%	Mexico, 墨西哥
F》(E	137	81.0%	鐵達尼號》(Titanic
F」(E	119	82.1%	亞特拉斯」(Atlas

表一：原文與譯文之間常見之樣式

其後，Wu et al. (2005)發現，文件中使用譯名時，除常於該譯名首次出現時，於其後以「(　)」加註原文外，尚有許多可能之樣式，如「/」或「,」(見圖一)。爲避免人工條列樣式容易有所遺缺，因此，他們透過有系統地從網頁資料中取得原文與譯文之間的樣式，驗證各樣式的準確率（見表一），並進一步從網頁資訊中，取得高準確度的英中翻譯。另，Kwok et al. (2005) 則透過樣式以及網路資訊來取得音譯名詞，如人名、地名等，並且建立了一線上服務系統 Chinet[2]。而這些研究主要都是透過搜尋引擎回傳的摘要，進行翻譯之擷取。不過由於搜尋引擎並非爲了搜尋翻譯而設計，因此也容易遇到回傳的摘要並無所需的翻譯資訊，導致

[2] http://xpkk.cs.qc.edu:8080/chinet/ (但發稿時測試無法連線)

召回率(recall)較低，所以研究者也開始思索著如何有效提升召回率。

Su (2006) 透過首次搜尋所取得的摘要進行目標語言的相關詞擷取，並作爲第二次搜尋摘要時的擴充查詢詞。Wu et al. (2007) 則利用預先訓練的音譯對照表作爲擴充查詢詞，有效地增加音譯資訊出現於搜尋所得摘要中的機率。此外，爲了能成功地尋找同一詞彙（如 ”option”）於不同領域的翻譯(如電腦領域的 “選項” 或金融領域的 “選擇權”)，Wu et al. (2008) 從 Wikipedia[3]中取得訓練資料，先行訓練各領域的相關詞彙，作爲輔助查詢之用，明顯地提升了尋找個別領域特有之翻譯的效能。而除了透過字串表面樣式來進行翻譯擷取的方式外，網頁亦有其他的資訊可供利用。例如 Lu et al. (2002, 2003)則是藉由連結錨文本（Anchor Text）以及超鏈結(hyperlink)，來進行多語詞彙翻譯之擷取，效果也相當優異，同時亦提供一即時線上翻譯搜尋系統 LiveTrans[4]供大眾使用。

TermMine 系統背景簡介

正因詞彙翻譯查找的需求日益增加，線上辭典如「Yahoo!奇摩字典」[5]，「Google Dictionary」[6]、或是像國立編譯館所編撰之「學術名詞資訊網」[7]等辭典庫也因應而生。可惜的是，這些辭典仍舊有著更新不易或是分類資訊不足之缺點。爲求能彌補這方面的缺憾，清大自然語言實驗室[8]和叡揚資訊[9]共同開發一套透過群體智慧共同創作的分類式專有名詞翻譯系統 – TermMine。

此系統以清大既有的網路搜尋翻譯研究爲基礎，進行文字探勘（Text Mining）之研究，以求發展辭彙蒐集、辭彙翻譯、辭彙分類等相關問題之解決方案。並透過社會加註（Social Tagging）與委眾（crowdsourcing）等做法，設計並建立雛形系統。而分類部份，則參考國立編譯館的分類項目，以及對應之 Wikipedia 分類架構與辭彙，逐步發展成爲集體共享之辭彙管理平台。在此系統中，融入了如 del.icio.us[10] 般社會加註（social tagging）與集體過濾（collective filtering）概念，是古代典籍集注（variorum）的今日網路版，落實協力建構與知識分享的理想。

[3] http://www.wikipedia.org/
[4] http://wkd.iis.sinica.edu.tw/LiveTrans/demo.html
[5] http://tw.dictionary.yahoo.com/
[6] http://www.google.com/dictionary
[7] http://terms.nict.gov.tw/
[8] http://nlp.cs.nthu.edu.tw/
[9] http://www.gss.com.tw/
[10] http://delicious.com/

圖二：TermMine 主畫面

TermMine 系統功能說明

目前 TermMine 的基礎詞庫是由清大自然語言實驗室透過將國家圖書館之碩博士論文資料，對各論文之中英文關鍵詞進行文字對應，取得高準度的中英對照關鍵詞表外，並納入國立編譯館名詞術語庫資料，現共具有中、英各 70 萬個詞，100 萬個翻譯，以及 180 個領域分類。

於系統主畫面中（見圖二），可得知近期內最熱門的前 10 個標籤(tag)，以及高頻率被查詢的中英文詞彙。使用者也可直接輸入欲查詢的中英文詞彙，搜尋系統中已收錄的詞彙翻譯。在輸入關鍵字時會同時顯示可選用的詞彙及已具有幾種翻譯資訊，提供使用者參考。在輸入查詢詞彙（以 database 爲例）後，將顯示人氣翻譯、國立編譯館、以及自建辭典的查詢結果（見圖三）。所謂人氣翻譯乃是列示出最多人收錄的翻譯，及其翻譯適用的分類（標籤資訊），可提供給使用者參考並可收錄適用者至自建辭典中。以此例中，使用者將可得知現有三種不同的翻譯，頻率由高到低分別是「知識庫」、「資料庫系統」以及「資料庫」，且大都屬“資訊”領域。目前此處資料乃是以國家圖書館所收錄的碩博士論文關鍵詞作爲起始資料。呈現同時也包括：網路所蒐集之摘要資料，其中摘要資料會顯示於右方區塊，作爲使用者進一步確認翻譯是否適用的參考資訊。而未來的人氣翻譯則將是由大家持續的新增翻譯與收錄，藉由群體智慧來繼續成長。而國立編譯館部分則呈現由專家學者所給予的翻譯以及其所屬的類別，作爲使用者的另一參

考。若該詞已曾被查詢之使用者收錄於自建辭典，也將同時顯示其自行已收錄的翻譯資訊。

圖三：TermMine 查詢"database"之結果

若目前尚無可參考的翻譯，或是所見翻譯皆不符合使用者的預設結果。使用者尚可透過網路即時搜尋翻譯的功能，且爲了搜尋到更精確的翻譯，您可以指定搜尋的領域(domain)及查詢的地區，搜尋結果之相關摘要亦會列示於右方區塊。目前提供生物、化學、物理、電腦資訊、經濟…等 26 個領域。網路搜尋將取得跟查詢詞彙相關的目標語言詞彙，若經確認爲所需翻譯，即可將其收錄到個人辭典之中，同時也分享給其他使用者此一翻譯（見圖四）[11]。

[11] 自建辭典、網路即時搜尋翻譯等部份功能僅供註冊會員使用。歡迎大家免費註冊已獲得完整服務功能。

圖四：TermMine 網路搜尋"Sinica"之結果

結語

本系統目的在於開發網路爲本的辭彙資訊之資料探勘技術，蒐集各類辭彙及相關資訊（如翻譯與解釋），提供個人間與社群內辭彙資訊分享與整合的功能，以建構一個跨越時間、空間、語言限制的辭彙管理平台，以期有助於知識管理所需之關鍵詞擷取、文件分類技術之發展。而系統功能，除可直接作爲一般使用者翻譯查詢之用途外，群策群力所累積之成果除可用於自然語言處理系統，更將有助於國內知識管理軟體，電子辭典，數位學習，數位內容產業的發展。

參考書目

Wen-Hsiang Lu, Lee-Fung Chien, and Hsi-Jian Lee. 2002. Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing*, 1(2): 159-172.

Wen-Hsiang Lu, Lee-Feng Chien, and Hsi-Jian Lee. 2003. LiveTrans: Translation Suggestion for Cross-Language Web Search from Web Anchor Texts and Search Results. In *Proceedings of Research on Computational Linguistics Conference XV* (ROCLING), 57-72, 2003.

K.L. Kwok, P. Deng, N. Dinstl, H.L. Sun, W. Xu, P.Peng, and J. Doyon. 2005. CHINET: a Chinese namefinder system for document triage. In *Proceedings of 2005 International Conference on Intelligence Analysis*.

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 2001. Using the web as a bilingual dictionary. In Proceedings of the workshop on Data-driven methods in *machine translation*, 1-8.

Chuan-Yao Su. 2006. Bilingual proper nouns extraction through web mining. Master thesis, National Chao Tung University, Taiwan.

Jian-Cheng Wu, Tracy Lin, and Jason S. Chang. 2005. Learning source-target surface patterns for web-based terminology translation. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, 37-40.

Jian-Cheng Wu and Jason S. Chang. 2007. Learning to Find Englishto Chinese Transliterations on the Web. In *Proceedings of EMNLP-CoNLL-2007*. pp.996-1004. Prague, Czech Republic.

Jian-Cheng Wu, Peter Wei- Huai Hsu, Chiung-Hui Tseng, and Jason S. Chang. 2008 Mining the Web for Domain-Specific Translations. In Proceedings of AMTA 2008.