中華民國計算語言學學會
通訊 ACLCLP
The Association for Computational Linguistics and Chinese Language Processing
第十五卷第一期
發 行 人：王駿發
主　　編：曾淑娟
執行編輯：黃　瑛
E-Mail：rocling@hp.iis.sinica.edu.tw　網 址：http://rocling.iis.sinica.edu.tw/ROCLING/
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
傳 真：(02)2788-1638
劃 撥：19166251

本 期 要 目

第十六屆自然語言與語音處理研討會

學會一年一度的「計算語言學研討會」自本屆(第十六屆)起將更名為「自然語言與語音處理研討會」，英文名稱則更名為「Conference on Computational Linguistics and Speech Processing」，縮寫仍維持「ROCLING」。本屆研討會謹訂於九月二日　三日假福華翡翠灣舉行，本次會議將合併數個 Workshop 舉行，大會亦將選出最佳論文，並於會中頒發獎項。徵稿啟事請參閱第二　四頁，歡迎踴躍投稿及參加。

中文資訊檢索標竿測試集第三版開放申請

「中文資訊檢索測試集第三版(CIRB030)」已開放申請，本測試集包含三個部分：文件集、問題集及相關判斷（答案集）。簡要說明請參閱第五　九頁，申請手續請逕自上網查詢或聯絡秘書處。

電子郵件帳號申請

本會為擴大對會員的服務，特免費提供本會會員電子郵件帳號使用，有意者請向本會秘書處提出申請。

職務異動通知

· 原任「中文計算語言學期刊」總編輯中研院資訊所陳克健博士任期至四月屆滿，五月起由成功大學資訊系吳宗憲教授接任。

· 原任「學術委員會」主任委員之成功大學資工系吳宗憲教授，即日起由中研院資訊所王新民博士接任。

學生出席國際會議補助

本會為擴大補助學生出席國際會議，特於 3/2 理監事聯席會議通過新修訂辦法，新辦法除了增加補助之會議外，並且放寬申請人資格，申請資格及辦法如下列：

· 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。
· 補助金額：由審查委員會依地區別及論文等級審定補助金額，每名補助金額上限為美金 1,000 元
· 補助名額：每個會議補助一　二名。
· 申請辦法：
  1. 日期：論文被接受發佈日二週內提出。
  2. 手續：申請人需將論文接受函、審查意見、學生證及論文全文等相關資料郵寄至本會秘書處。

ROCLING XVI:

Conference on Computational Linguistics and Speech Processing
第十六屆自然語言與語音處理研討會

September 2-3, 2004, Howard Pacific Green Bay, Taipei, Taiwan, ROC
http://www.aclclp.org.tw/rocling2004.html

CALL FOR PAPERS

The 16th ROCLING Conference will be held September 2-3, 2004 at Howard Pacific Green Bay in Taipei. Sponsored by Association for Computational Linguistics and Chinese Language Processing (ACLCLP), ROCLING is the most historied and comprehensive conference focused on computational linguistics, speech processing, and related areas in Taiwan. ROCLING XVI will be hosted by Institute of Information Science, Academia Sinica. The conference will feature invited lectures, tutorials, panel discussions, and lecture and poster sessions and two workshops: Workshop on Intelligent Web Technologies and Workshop on Computer Assisted Language Learning.

Papers are invited on substantial, original, and unpublished researches on all aspects of computational linguistics, including, but not limited to the following topic areas.

(a) cognitive linguistics	(l) parsing/generation
(b) discourse modeling	(m) phonetics/phonology
(c) document database/large corpora	(n) quantitative/qualitative linguistics
(d) electronic dictionaries	(o) speech analysis/synthesis
(e) information retrieval	(p) speech recognition/understanding
(f) language understanding	(q) spoken dialog systems
(g) language processing over Internet	(r) spoken language processing
(h) machine translation	(s) syntax/semantics
(i) NLP and educational applications	(t) Web information extraction
(j) morphology	(u) Web corpora
(k) computer assisted language learning	(v) others

Paper Submission:
Prospective authors are invited to submit full papers of no more than 25 A4-sized pages in pdf or MS Word format. Papers will be accepted only by electronic submission through the conference web site. Prospective authors without web access should contact the Program Committee Co-Chair (whm@iis.sinica.edu.tw) before the submission deadline. The submitted papers should be written in either Chinese or English, and in single column, double-spaced format. The first page of the submitted paper should bear the items of paper title, author name, affiliation and email address. All these items should be properly centered on the top, with a short abstract of the paper following.

Best Paper Award:
The best paper will be selected and announced at ROCLING XVI.

Important Dates:
Preliminary paper submission due:	July 5, 2004
Notification of acceptance:	July 26, 2004
Final paper due:	August 9, 2004

Sponsors:
Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
Institute of Information Science, Academia Sinica
Institute of Linguistics, Academia Sinica

Organizing Committee:
Pu-Jen Cheng, Lee-Feng Chien, Wei-Ho Tsai, Hsin-Min Wang, and Jeng-Haur Wang (Academia Sinica)
Qi Huang (ACLCLP)

Intelligent Web Technologies

Workshop at ROCLING XVI

Taipei, Taiwan, September 2, 2004

http://myweb.ncku.edu.tw/~whlu/iwt2004.htm

whlu@mail.ncku.edu.tw

Call For Papers and Participants

IMPORTANT DATES:

Abstracts due: July 2, 2004

Papers due: July 8, 2004

Acceptance notification: July 28, 2004

Camera ready due: August 8, 2004

OVERVIEW

The Web is becoming the largest data repository in the world and presents a key driving force for a large spectrum of information technology (IT). How to benefit intelligent Web-based information systems through the mining of diverse Web data resources is being studied in the emerging research area of Web knowledge discovery.

To develop effective and intelligent Web applications and services, it is critical to discover useful knowledge through analyzing large amounts of contents, hidden content structures, or usage patterns of Web data resources. To achieve such goal, a variety of techniques in diverse research areas are needed to integrated properly, including natural language processing, information extraction, information retrieval, information filtering, knowledge representation, knowledge management, machine learning, databases, data mining, Web mining, text mining, agent, human-computer interaction, and semantic Web. These integrated techniques must address the important challenges from the scale, the heterogeneous and dynamic nature of Web contents and usage patterns.

This workshop intends to bring together researchers and practitioners to foster the exchange of ideas and the dissemination of emerging techniques on intelligent Web technologies (knowledge discovery through Web usage, structure and content mining). The workshop will capture current important developments of new models, new methodologies and new tools for building a variety of embodiments of scalable, effective and intelligent Web-based information systems.

TOPICS

Original contributions are solicited in the following subjects (but not necessarily limited to):

- Web information extraction and wrapper generation
- Web content and structure mining
- Web Information retrieval and filtering
- Web mining for knowledge discovery, business intelligence and security

- Web data collection and analysis, including query logs, click streams, call center streams, and transactional data
- Web data preparation, including cleansing, transformation, and sampling
- Text Mining for Creating Metadata
- Learning Taxonomies and Ontologies from the Web
- Classification/clustering of Web pages (sites) and multimedia content
- Web search engines, meta-search engines and inference engine
- E-mail classification and spam filtering
- Semantic Web
- Intelligent Web agents
- Knowledge community formation and support
- Intelligent E-technology
- Intelligent human–Web Interaction
- Web-based personalized techniques

TIME SCHEDULE

The desired workshop will run in one day of approximately 6 hours of 1-2 keynote speeches and 10-12 paper presentations.

PAPER SUBMISSION

All papers must be submitted in either PDF or MS Word format to the co-chair Dr. Wen-Hsiang Lu whlu@mail.ncku.edu.tw.

ORGANIZING COMMITTEE

Vincent Shin-Mu Tseng (Co-Chair)
Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan
tsengsm@mail.ncku.edu.tw

Wen-Hsiang Lu (Co-Chair)
Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan
whlu@mail.ncku.edu.tw

CIRB030 資訊檢索測試集簡介

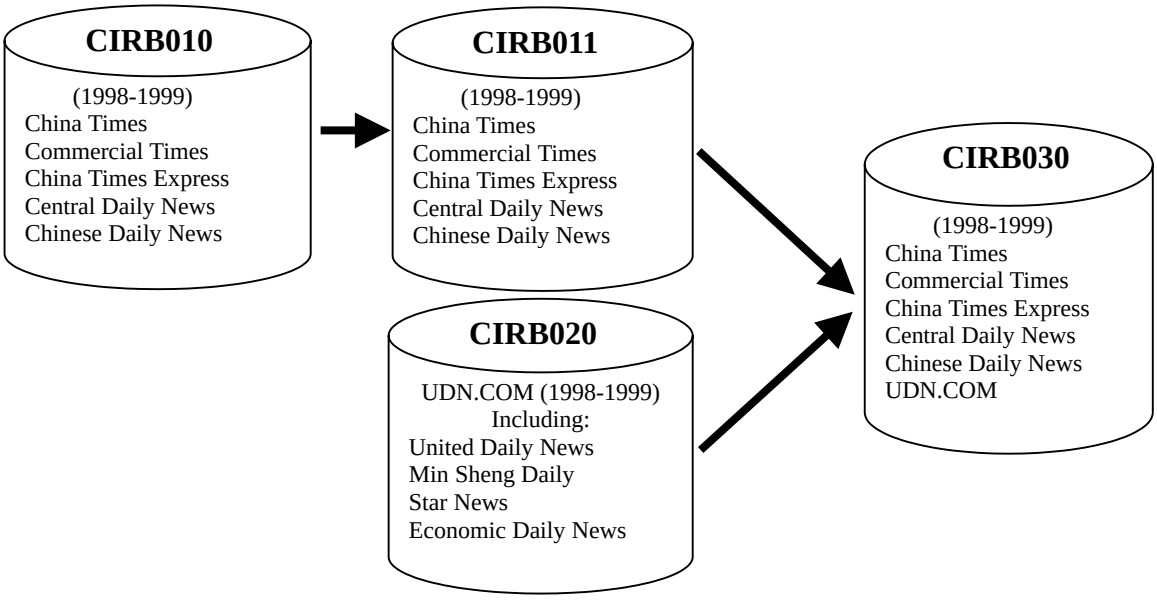
陳光華
國立台灣大學圖書資訊學系
khchen@ntu.edu.tw

陳信希
國立台灣大學資訊工程學系
hh_chen@csie.ntu.edu.tw

1. 序論

資訊檢索測試集是用於評估資訊檢索系統的績效。測試集對於資訊檢索系統開發過程的評量與資訊檢索系統的評量而言，都是一種極為重要且有效的工具。CIRB030 (Chinese Information Retrieval Benchmark, version 3.0)便是用於評估中文資訊檢索系統的測試集。

一般而言, 測試集包含三個部分：文件集（在 CD-ROM 的 DocSet 文件夾), 問題集（TopicSet 文件夾），相關判斷（答案集，AnswerSet 文件夾)。本文將為各位介紹 CIRB030 資訊檢索測試集, CIRB030 和用於 NTCIR3 Workshop 的 CIRB011 與 CIRB020 有些許不同, 基本上, CIRB030 的文件集是由 CIRB011 與 CIRB020 的文件集合組而成，但修改了部分文件亂碼的問題與部分文件 HEADLINE 與內文不符的問題，同時刪除無內文的文件，因此文件的數量與 CIRB011 及 CIRB020 的文件總數有些許出入。而 CIRB011 和先前發行的 CIRB010 僅有標記上的不同；CIRB011 與 CIRB020 的標記則完全相同。相關標記會於下文說明。這一次我們決定直接發行 CIRB030 而跳過 CIRB020，原因就是想整合 CIRB011 的文件集與 CIRB020 的文件集。因此，您無須擔心如何取得 CIRB020 的問題，它不會在台灣單獨發行，除非您曾參與 NTCIR 資訊檢索評估會議 [1]。為了更清楚地說明 CIRB 版本的演變情形，請您參考圖一。



圖一：CIRB 版本演變

另外必須要注意的是, CIRB030 的新聞文件已經整合為 7 個文件檔案, 它們是位於 CD-ROM 的 **DocSet** 文件夾之下的 cdn1998-1999 (中央日報), chd1998-1999 (中華日報), ctc1998-1999 (工商時報), cte1998-1999 (中時晚報), cts1998-1999 (中國時報), udn1998 (聯合報系 1998), 與 udn1999 (聯合報系 1999)；而這些文件在 CIRB011 與 CIRB020 時是各自獨立的，因此有381,681個文件檔案。下文將分別說明文件集、問題集與答案集三部分。

2. 文件集

CIRB030 的文件集是由不同的新聞機構合法取得的，表一說明這些新聞文件的來源與數量。新聞文件的內碼為 BIG5，且經過後續處理，加上適當的 XML 標記，這些標記是經過 NTCIR 執行委員會的討論而制訂的，表二羅列這些標記，並說明其意義。圖二則是一個標記後新聞文件的例子。

表一：文件集的組成

中央日報 (cdn1998-1999)	27,770
中華日報 (chd1998-1999)	34,728
中國時報 (cts1998-1999)	38,116
中時晚報 (cte1998-1999)	5,747
工商時報 (ctc1998-1999)	25,811
聯合報系 (udn1998 and udn1999)	249,203
總計	381,375

表二：CIRB030 文件的標記

必要標記		
<DOC>	</DOC>	The tag for each document
<DOCNO>	</DOCNO>	Document identifier
<LANG>	</LANG>	Language code: CH, EN, JA, KR
<HEADLINE>	</HEADLINE>	Title of this news article
<DATE>	</DATE>	Issue date
<TEXT>	</TEXT>	Text of news article
選擇標記		
<P>	</P>	Paragraph marker
<SECTION>	</SECTION>	Section identifier in original newspapers
<AE>	</AE>	Contain figures or not
<WORDS>	</WORDS>	Number of words in 2 bytes

3. 答案集

CIRB030 的問題集是由日本、韓國、台灣、以及 TREC [2] 共同製作的，換言之，問題集具有國際化的特色，而且，每一個問題皆有四個語言的版本，亦即中文、英文、日文、以及韓文。我們使用<SLANG>標記表明該問題的製作國家或機構，如<SLANG>CH</SLANG>表示該問題是由台灣製作的；<SLANG>EN</SLANG> 是由 TREC 製作的；<SLANG>JA</SLANG> 是由日本製作的；<SLANG>KR</SLANG> 則是由韓國製作的。<TLANG>標記則用於表明該問題目前的呈現語言。圖三展示一個 CIRB030 製作的問題的例子，而表三說明問題集使用的標記。

```
<DOC>
<DOCNO>udn_xxx_19980101_0001</DOCNO>
<LANG>CH</LANG>
<HEADLINE> 南華早報報導中共內部兩件與台灣相關新發展： </HEADLINE>
<DATE>1998-01-01</DATE>
<TEXT>
<P>香港英文南華早報今天報導，一九九八年的台灣新聞，可能會和一九九七年的亞洲金融風暴一樣成為最大的新聞事件。</P>
```

<P>該報說，中共在過去一周傳出兩件與台灣有關的新發展。一是以中共國家主席江澤民為組長的中央對台灣工作領導小組，將增加幾名重量級的文職和軍職成員　而軍方代表包括中央軍委兩位副主席張萬年和遲浩田。</P>
<P>無論從任何角度分析，中共計畫擴大中央對台領導小組，顯然都是旨在加快打破兩岸目前的僵局。</P>
<P>南華早報說，第二項發展是中共通過內部文件向各級幹部，特別是台灣事務幹部，傳達海峽兩岸僵局一兩年內一定會達成突破的信息。</P>
<P>該報說，中共將在三月召開過全國人民代表大會之後，宣布一項和一九九五年一月公布的江八點同等重要的建議。</P>
<P>消息人士說：「江澤民的條件似乎是李總統不要用「中華民國總統」即可。另外則是由於大陸面積較大，李總統可先訪問中國大陸，江澤民再訪台灣」</P>
<P>消息人士引述一名中共高幹的話說，突破有可能很快達成。他被引述說：「（一九九八）春天耕耘播種，冬天來到前即可收穫。」</P>
</TEXT>
</DOC>

圖二：新聞文件範例

<TOPIC>
<NUM>013</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>NBA labor dispute</TITLE>
<DESC>
To retrieve the labor dispute between the two parties of the US National Basketball Association at the end of 1998 and the agreement that they reached.
</DESC>
<NARR>
The content of the related documents should include the causes of NBA labor dispute, the relations between the players and the management, main controversial issues of both sides, compromises after negotiation and content of the new agreement, etc. The document will be regarded as irrelevant if it only touched upon the influences of closing the court on each game of the season.
</NARR>
<CONC>
NBA (National Basketball Association), union, team, league, labor dispute, league and union, negotiation, to sign an agreement, salary, lockout, Stern, Bird Regulation.
</CONC>
</TOPIC>

圖三：問題範例

最初，我們共製作 50 個問題，然而我們在 NTCIR 資訊檢索評估會議後，刪除了 8 個較不適用的問題，留下的合格的問題的編號為 1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、17、18、19、20、21、22、23、24、25、27、32、33、34、35、36、37、38、39、40、42、43、45、46、47、48、49、以及 50。因此，您可能會疑惑為何在 CD-ROM 的 **TopicSet** 文件夾下的問題似乎有遺漏的狀況，但是這是正確的，因為前述不適用的 8 個問題並沒有收錄於此次的 CIRB030 問題集。表四說明問題集的來源與數量。

表三：CIRB030 問題的標記

<TOPIC>	</TOPIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language code: CH, EN, JA, KR
<TLANG>	</TLANG>	Target language code: CH, EN, JA, KR
<TITLE>	</TITLE>	The concise representation of information request, which is composed of noun or noun phrase.
<DESC>	</DESC>	A short description of the topic. The brief description of information need, which is composed of one or two sentences.
<NARR>	</NARR>	A much longer description of topic. The <NARR> has to be detailed, like the further interpretation to the request and proper nouns, the list of relevant or irrelevant items, the specific requirements or limitations of relevant documents, and so on.
<CONC>	</CONC>	The keywords relevant to whole topic.

表四：CIRB030 問題集的組成

Created By	**Original**	**After filtering (This release)**
Japan	15	12
Korea	12	10
Taiwan	13	13
TREC	10	7
TOTAL	50	42

4. 答案集

相關判斷是由台灣、日本、韓國的工作同仁一起合作完成的，而台灣負責整合最後的相關判斷。CIRB030 的相關判斷分為四個層級：非常相關、相關、部分相關、與不相關，每一個層級都會設定代表的符號與數值，表五說明這些符號與數值。

然而，TREC_EVAL 程式採用的是二元層級的相關判斷，而 TREC_EVAL 被視為是資訊檢索評估的標準作法，因此我們決定製作兩組答案，一組為嚴謹相關（Rigid Relevance），也就是非常相關與相關視為相關；一組為鬆散相關（Relaxed Relevant），也就是非常相關、相關、部分相關皆視為相關。您可以在 CD-ROM 的 **AnswerSet** 文件夾找到二個檔案，CIRB030RJCH-Rigid 為嚴謹相關；CIRB030RJCH-Relax 為鬆散相關。

表五：相關判斷的層級

相關層級	**符號**	**分數**
非常相關（Highly Relevant）	S	3
相關（Relevant）	A	2
部分相關（Partially Relevant）	B	1
不相關（Irrelevant）	C	0

您可以使用 TREC_EAVL 程式評估您的資訊檢索系統產生的檢索結果，這個程式可以在 CD-ROM 的 **TREC_EVAL** 文件夾找到 Windows 版與 unix 版的程式。要注意的是，TREC_EVAL 程式要求固定的格式，檢索結果的格式如下：

qid iter docid rank sim runid

qid 代表問題編號；*docid* 代表文件標號；*rank* 為檢索出文件的排序；*sim* 為文件與問題的相似性；*runid* 為該次檢索結果的編號（您可以自行賦予）；*iter* 沒有特別的功用，設定為 1 或 0 即可。各個欄位以'TAB' (\x0A, \t) 字元分隔。

5. 結論

CIRB030 的文件集皆為中文新聞文件，因為我們僅獲得使用中文文件的權力。您可以運用這個測試集進行跨語資訊檢索，如 E-C、J-C、與 K-C，或是單語資訊檢索，僅有 C-C 。如果您想進行更複雜的跨語資訊檢索的評估，如 E-CJ、E-CK、C-JK、E-CJK 等等，最好的方法是參加 NTCIR 資訊檢索評估會議，您即可合法取得使用英文、日文、韓文的權力（部分有時間的限制），細節請參考 NTCIR 官方網站[1]。有關 CIRB 系列測試集使用於 NTCIR 資訊檢索評估會議的情形，可以參考[3]、[4]、與[5]等論文。

致謝

感謝中時報系、聯合報系、中央日報、中華日報提供新聞文件，感謝 TREC 與 CLEF 協助製作問題集，感謝相關判斷者的辛勤工作。

參考文獻

[1] NTCIR. http://research.nii.ac.jp/ntcir/

[2] TREC. http://trec.nist.gov/

[3] Kando, Noriko. Overview of the Second NTCIR Workshop. In *Proceedings of the Second NTCIR Workshop on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 51-72, Tokyo, 2001.

[4] Chen, Kuang-hua and Chen, Hsin-Hsi. The Chinese Text Retrieval tasks of NTCIR Workshop 2. In *Proceedings of the Second NTCIR Workshop on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 51-72, Tokyo, 2001.

[5] Chen, K. H., Chen, H. H., Kando, N., Kuriyama, K., Lee, S., Myaeng, S. H., Kishida, K., Eguchi, K. and Kim, H., Overview of CLIR Task at the third NTCIR workshop, *Working Notes of the Third NTCIR Workshop Meeting, Part II: Cross Lingual Information Retrieval Task*, pp. 1-38, 2002.

The 18th Pacific Asia Conference on Language, Information and Computation (PACLIC)

December 8 (Wednesday) - 10 (Friday), 2004
Waseda University, Tokyo
Organized by Logico-Linguistic Society of Japan
www.decode.waseda.ac.jp/PACLIC18/index.html

Preliminary Call for Submission and Participation

The Logico-Linguistic Society of Japan is pleased to announce that PACLIC 18 is scheduled to be held at the International Conference Center of Waseda University, December 8-10, 2004. This will be the latest installment of our long standing collaborative efforts among theoretical and computational linguists in the Pacific-Asia region in providing an opportunity to share their findings and interests in the formal study of syntax and semantics of natural and formal languages. The spectrum of topics covered ranges from theoretical and computational studies in syntax, semantics, and pragmatics to corpus linguistics and contrastive analyses of Pacific-Asian languages. The program will include several invited lectures by renowned experts in the field.

Submission details and deadlines will be announced and updated at the following web page:

www.decode.waseda.ac.jp/PACLIC18/index.html

Akira IKEYA,
Honorary Chair of PACLIC 18 Program Committee
President of the Logico-Linguistic Society of Japan
Dean of School of Humanities, Toyo Gakuen University

Schedule and Important Dates (All dates in JST)

- Submissions due: July 26 (Monday), 2004
- Notification of acceptance sent: August 30 (Monday), 2004
- Final versions due: October 25 (Monday), 2004

PACLIC Steering Committee

Chu-Ren HUANG, Academia Sinica, Taipei
Akira IKEYA, Toyo Gakuen University, Tokyo
Kiyong LEE, Korea University, Seoul
Kim Teng LUA, Chinese and Oriental Languages Information Processing Society, Singapore
Benjamin T'SOU, City University of Hong Kong, Hong Kong
Donghong JI, Laboratories for Information Technology, Singapore
Yasunari HARADA, Waseda University, Tokyo
Pornsiri SINGHAPREECHA, Thammasat University, Bangkok

PACLIC18 Program Committee

Honorary Chair
Akira IKEYA, Toyo Gakuen University, Tokyo
Co-Chairs
Chu-Ren HUANG, Academia Sinica, Taipei
Beom-mo KANG, Korea University, Seoul

第二屆語言學卓越營：語料庫與計算語言學

時間：2004 年 7 月 12 ~ 23 日
地點：中央研究院
主辦單位: 行政院國家科學委員會
承辦單位: 中央研究院語言學研究所 台灣語言學學會 中華民國計算語言學學會
協辦單位: 國立台灣大學語言學研究所　國立中正大學語言學研究所

本屆語言學卓越營邀請國內外中文語料庫與計算語言學學者以兩週的專業課程介紹中文語料庫與計算語言學基礎知識。 課程分三主題：中文計算語言學概論、中文語音處理與語料庫語言學。 配合課程，主辦單位並安排專題演講與座談會與學員分享中文語料庫與計算語言學相關領域最新研究狀況。

將於五月初開放報名，歡迎有興趣的學者和學生踴躍報名參加!

師資陣容:

中文計算語言學課程:

俞士汶	(北京大學計算語言所)
鄭錦全	(中研院語言所)
黃居仁	(中研院語言所)

中文語音處理課程:

石基琳	(美國伊利諾大學 Urbana-Champaign 分校語言所)
陳信宏	(交通大學電信所)

語料庫語言學課程:

薛念文	(美國賓州大學資訊所)
張俊盛	(清華大學資訊工程系)
陳克健	(中研院資訊所)

聯絡人: 楊元禎小姐
電話: 02-27863300 轉 300
傳真: 02-27856622
地址: 中央研究院語言學研究所　(115)台北市南港區研究院路二段 128 號
E-mail: tsil@gate.sinica.edu.tw
詳細活動內容將於四月公佈於中央研究院語言所網站 http://www.ling.sinica.edu.tw

中研院詞庫小組研究-答客問

陳克健　　黃居仁
中央研究院資訊所　　中央研究院語言所

這一篇小文是助理經常問我們的問題，也許對計算語言學有興趣的同學也會有相同的疑問，因此藉這個機會將一些計算語言學相關問題重新整理發表在學會通訊上，若能釐清某些研究上的觀念及提供一些有用的訊息，甚幸。

問：簡述一下詞庫小組的研究工作。

答：詞庫小組長期以來所立下的最終目標希望能使電腦具有了解及處理中文語言的能力。為了達成這個長遠的目標，詞庫小組採用漸進的方式，一方面研究中文的語法語意現象，一方面尋求語言知識的表達模型及計算機處理語言的技術。從對語言本身的了解進而尋求全面而完整的解決策略。這是詞庫小組一貫的方向及立場，因此詞庫小組以詞彙研究為出發點，以詞彙知識庫的構建為初期目標，希望能清楚了解語句中最小的有意義的單位「詞」，其語法、語意功能，並探討詞彙知識表達的模型。為了研究實際的詞彙語法語意現象，必須先建立研究的素材及工具，因而有中文語料庫的構建計劃。工作包括語料的收集，分類到分詞附加詞類及句結構標記，並儘量以自動化方式完成以上工作。語料是語言研究的素材，為了能進一步分析語料，更容易從語料中自動抽取語言訊息，因此有語料標記的過程，包括文件的屬性分類，文本的分詞及附加語法標記及句結構標記。這些加工過程利用了所發展的自動分詞、詞類標記及剖析技術，以電腦處理求得快速的結果及標記的一致性，再經人工修訂以保證成果的品質和正確性。這樣的一個處理過程，除了節省人力外，另一方面也是測試所發展出的分詞、標記、剖析等處理技術，做為系統改進及知識表達欠缺的參考，是相輔相承的。完成後的語料庫因為附加了詞類、結構及成分間語意關係資料，成為較佳的研究素材，可以容易而客觀的觀察、比較及抽取語言知識。

簡而言之，以上所描述的詞知識庫、語料庫、語言處理技術，形成了中文語言研究的基礎環境，也是電腦自動處理中文的發展平臺。我們相信要發展有用的自然語言處理系統，電腦除了有基礎語言知識外還必須有自動學習的能力，因為即使是一個特定領域的自然語言系統，被動式的人為輸入語言及領域知識是永遠無法滿足需要。要能及時調適語言變化，獲得新知識，電腦必須要有自動學習能力。為了達成此一目標，詞庫小組近程的研究工作包括：

(a)繼續構建句結構語料庫

以剖析系統分析標記語料，產生句結構樹，並加人工修正。完成的句結構樹用來抽取語言訊息，回饋剖析系統。

(b)語言知識自動學習

研究詞彙自動抽取、辨認分類的方法及語言自動分析，語法及語意關係自動抽取的機制，以特定領域文件為學習對象，完成自動學習系統。

(c)語言及知識表達研究

研究詞彙語意及語言知識的表達模型，並充實現有詞彙知識表達內容。

我們的遠程目標希望電腦真正能具備語言的能力，不但可以完成交談式的智慧型人機介面，同時經由語言的自動分析了解，可以在浩翰的網路世界中，自動整理出有用的知識，形成許多不同領域的專家系統。

問：花這麼多時間這麼多人力建立語言研究基礎環境方向正確麼？方法正確麼？

答：自然語言本身就是一個複雜的系統，為了能讓電腦處理語言必須先讓電腦具有語言知識及語言處理的技術。因語言本身的複雜性而有語言學的學門，為了研究電腦處理語言而有計算語言學的學門。這兩個學門一個是「知」一個是「行」恰為互補。因此整個詞庫小組的研究工作不是做單純的資料整理的工作，相反的資料整理只是一個必然的結果及必要的過程。詞庫小組研究計劃的理念是「了解語言現象後，才能做出正確的知識表達架構、語言知識內容及處理的技術」。語言知識的取得是研究的成果是分析的結果，而分析的方式是以電腦輔助或自動抽取的形式進行，期以最少的人力做出最精緻的成果。以實際的語料為處理對象，研究語言真正呈現的方式，及實用導向的系統設計，因此整個研究過程是建立在一個非常堅實的基礎上，一步一腳印，腳踏實地經過不斷討論方能把中文詞庫、語法、語料庫一一構建完成。目前的具體成果是以精簡的人力，以長時間的累積經驗完成的初步研究環境，包括一個八萬目詞以上的詞庫，五百萬詞的標記語料庫，中文句結構樹資料庫及分詞剖析系統。詞庫小組最終的理想是經由基礎語言知識的建立及處理技術的發展，電腦系統有朝一日可以自動學習新的語言知識，吸收常識，累積知識自動成長而不是被動的被灌入知識，因此目前的研究及資料整理工作是一個必經的歷程，為未來的目標奠定基礎。詞庫小組十幾年來以每年平均不到五百萬的研究經費，兼顧詞庫、語料庫、語法、處理技術的研究，比較日本 EDR 電子詞典計劃每年平均十四億日元，實微不足道。

中文資訊處理研究團隊已有十幾年合作的經驗，雖然限於人力經費的關係。許多重要的研究工作都是在極度困難的情況下勉力完成，所發展的成果，如詞典、語料庫、樹圖資料庫、分詞系統等。皆能透過計算語言學會或直接由網路下載推廣至學術界及工業界，成效良好。然而中文資訊研究的成果必須持續不斷的累積、發展、維護及推廣才能見到真正的成效。因此研究人員及經費維持長期性的穩定非常重要，目前面臨的困難依然是臨時助理較難維持長期工作，容易造成經驗及知識累積的斷層。已多次建議成立「中文語文中心」，解決此一問題。

問：詞庫為什麼要分析「高興」和「快樂」有什麼不同？（為什麼要做詞彙語意的研究？）

答：因為我們不清楚語句的語意是怎麼從詞彙的語意合成的。詞庫研究的目標是要如何使電腦俱備人類的語言能力，如果我們連句子語意是怎麼合成的都無法弄清楚，如何能設計電腦程式分析語句、了解語意呢？因此我們從詞彙語意研究著手，希望能了解詞彙和概念(concept)的表達關係，詞彙的多義性及意義的呈現機制，及意義延伸的方式，進而了解語意的合成機制，能發展一個有用的語言知識表達模型，電腦可以利用表達的語言知識，分析語言、了解語言，進而學習新的語言知識。

分析近義詞，例如「高興」和「快樂」，是一種研究這個問題的手段及技巧。藉由近義詞的相似及差異比對，可以比較容易能了解語意和語法之間的互動關係，及語意特徵之間的相容性互斥性及邏輯關係，找到語言知識及常識的分野及知識表達模型。

問：電腦如何學習語言知識？

答：讓電腦能有自動學習語言知識的能力，首要條件之一就是電腦必須俱備足夠的基礎語言知識及常識。電腦利用俱備的基本知識，從分析文件過程中發掘新的詞彙、新的句型、新的詞彙關係，以及新的常識。探索分析學習的過程所用到的語言處理技術及語言知識包括：

1. 詞彙分析技術，
2. 新詞自動分辨分類技術，
3. 語句剖析技術，及
4. 詞彙知識庫（如 CKIP dictionary 等)和知識本體架構(ontology，如 SUMO, WordNet, HowNet 等)。

新詞經常出現在不同領域文件中，專名出現頻率最高，包括人名、地名、公司名、書名、舟車名等等。動補、動賓、偏正並列式複合動詞、名詞也經常出現用來表達一些新的概念。電腦猜測新詞的語法語意屬性，所依賴的訊息包括構詞律、詞素的語法語意，及新詞所在的上下文訊息。如果電腦俱備以上訊息，再加上一些常識，一般的複合詞的語法語意屬性已能猜出十之八九。

構詞律及詞彙的訊息，比較固定，因此多由語言專家藉助電腦分析觀察語料以半自動的方式獲得。至於上下文裡的線索，則多藉重電腦從大量語料中不斷自動抽取而得。從已知詞的上下文可以抽取出意類和上下文之間的共現關係，做為判斷未知詞的線索。因此如果已知詞記錄越多，所能習得的線索也越多，線索越多判斷的正確性也越高。電腦以遞迴漸進的方式一方面充實知識一方面也由於知識的充實能有更強的語言分析能力，也就能習得更多的知識，這是整個研究的主要精神。

至於電腦如何習得新的句型、新的詞彙關係、新的語意關係，其實跟人的學習過程

沒有兩樣。語料中提供了這些知識，只是電腦能不能從中正確的分析出這些知識。我們的想法是，設計一個語句剖析系統，它能依據已有的知識分析語料，並抽取分析結果中所含的知識，譬如句型及詞彙語意關係。不斷分析並充實知識的過程中，我們假設正確的知識會一再重覆出現，而錯誤的分析只會偶而為之。在這個前提下，經過不斷的學習及充實知識，電腦的剖析系統會不斷的自我改進。因此本研究的主要重點是如何避免同樣的錯誤重覆的發生，其方法是採用語意為主、語法句型為輔的剖析策略，如此可以避免因錯誤句型而重覆產生錯誤的結果，而語法限制在輔助導引出正確的語意分析。

問：純粹統計的研究策略有哪些優點？哪些限制？

答：由於大量標記語料的出現使得統計式的語言處理方法因此盛行，並且在許多應用領域有突破性的進展。例如語音辨認利用了統計式的語言模型使得正確率大為提升。統計模型非常適用於決策選擇(decision making)，應用最廣的就是分類的選擇(classification problems) 。許多自然語言處理的問題都是要解決歧義，例如翻譯上的歧義，詞義詞類的歧義，標記的歧義，結構的歧義，同音字的的歧義，分詞的歧義，都可以歸納為分類問題，以統計機率模型計算不同歧義答案的機率選取機率最高的答案。這個策略一定不會錯，但不保證會成功。因為統計機率模型推算出的答案機率不會是 1 或 0，也不是真正的正確機率，而是以不完整的資訊作出的推估值。資訊不完整的原因主要是訓練語料不足及沒能掌握到真正的相關係數(dependent parameters)。訓練語料不足的問題比較不屬於研究層次的問題，想辦法增加語料就是了。沒能掌握到相關係數才是真正的問題所在。許多研究採用亂槍打鳥方式以不同的機率模型亂試，把 Markov Model, SVM, Neural Network, Bayesian Classifier, Maximum Entropy Model 試遍了，發現各個方法差異不大。問題不是在於方法不好，只是沒有找到正確的相關特徵係數。好的方法必須用的正確才能做出好的結果。也就是說分析語言了解語言特性應下的功夫不能少。如果有好的統計機率模型加上正確的相關特徵係數一定可以得到不錯的結果。

問：為什麼不和大陸的學者交流，或利用對方研究的成果？

答：我們也很希望能互相合作，但基於政治環境的現實考量大規模的合作並不可行。私下的交流並未有中斷。從計算語言學研究的成果而言，相信我們要比大陸先進。語料庫、詞庫的建立也比大陸更早且較完整。不過大陸的研究單位較多人員眾多，近年來研究成果在質和量方面進展迅速,觀念也比以前開放，已有許多方面超越我們值得我們借鏡。

基本上大陸的研究成果可以參考及使用的我們並不重覆整理，例如中文自然語言處理開放平台http://www.nlp.org.cn/就提供了許多免費的資料庫及技術。他們也成立了 CLDC（Chinese Linguistic Data Consortium）可以透過 CLDC 得到技術和資料的授權。另外大陸同義詞詞林的資料及架構已應用在我們的詞彙知識表達研究上。知網(HowNet)的知識架構也是一項詞彙核心知識表達的基礎工程，我們也正與大陸中科院董振東教授合作建立

繁體字版的知網(HowNet)，將來也會開放給各位使用。

問：如何取得中研院詞庫小組建構的資料或技術？

答：只要是學術研究，原則上都可透過計算語言學會得到授權資料。詳細資料及申請辦法可上網取得。

http://rocling.iis.sinica.edu.tw/ROCLING/

至於線上使用也很方便：

平衡標記語料庫：http://www.sinica.edu.tw/SinicaCorpus/
中文句結構樹資料庫：http://TreeBank.sinica.edu.tw/
中文分詞系統：http://ckip.iis.sinica.edu.tw/~ma/uwextract/

相關連結：
中央研究院語言學研究所_文獻語料庫研究室 http://corpus.ling.sinica.edu.tw/

最後叮嚀：有些訊息沒有提供網址，請用 google 查詢(http://www.google.com/)。