

本期要目

壹. 會員大會會議記錄	第二 五頁
貳. 中文資訊檢索標竿測試集第一版說明	第六頁
參. 中研院語言所籌備處演講通告	第七 九頁
肆. 學術活動 – IJC-NLP-04	第十 十二頁
伍. 自然語言處理技術與生物醫學文獻- 梁婷教授	第十三 十六頁

第八屆理監事選舉結果

第八屆第一次會員大會已於 9/18 日於清華大學國際會議廳圓滿結束，會中除了報告學會今年度工作會務及明年度工作計畫外，並順利改選第八屆理事、監事及海外理事，共選出理事十一名，候補理事二名；監事五名，候補監事一名；海外理事三名。當選名單如下：

理事：

1. 鄭秋豫博士(中央研究院語言所籌備處研究員)
2. 張景新教授(暨南國際大學資工系助理教授)
3. 柯淑津教授(東吳大學資訊科學系副教授)
4. 王新民博士(中央研究院資訊所副研究員)
5. 許聞廉博士(中央研院資訊所研究員)
6. 陳信希教授(台灣大學資訊工程學系教授)
7. 簡仁宗教授(成功大學資訊工程系副教授)
8. 陳信宏教授(交通大學電信工程系教授)
9. 王駿發教授(成功大學工學院院長)
10. 吳宗憲教授(成功大學資訊工程系教授)
11. 簡立峰博士(中央研究院資訊所副研究員兼副所長)

附註：以上順序依候選時之編號排序

候補理事：

1. 張照煌博士(工研院電通所前瞻技術中心資深研究員兼副主任)
2. 劉美君教授(交通大學外文系教授)

附註：依序候補

監事：

1. 陳克健博士(中央研究院資訊所研究員)
 2. 蘇克毅博士(致遠科技股份有限公司總經理)
 3. 鄭伯順副所長(中華電信研究所副所長)
 4. 余孝先博士(工研院電通所前瞻技術中心主任)
 5. 張俊盛教授(清華大學資訊系教授)
- 附註：以上順序依候選時之編號排序

候補監事：

1. 沈家麟博士(台達電子研發中心經理)

海外理事：

2. 舒維都教授 Victor Zue (Massachusetts Institute of Technology, Cambridge)
3. 鄒嘉彥教授 Benjamin K. T'sou (香港城市大學教授及語言資訊研究所中心主任)
4. 李錦輝博士 Chin-Hui Lee (Georgia Institute of Technology, USA)

附註：以上順序依候選時之編號排序

第八屆理事長、監事主席當選人

經由第八屆第一次理監事聯席會議中新任理事及監事投票推選出本屆理事長、副理事長及監事主席，當選名單如下：

理事長：王駿發教授(成功大學工學院院長)

副理事長：陳信宏教授(交通大學電信工程系教授)

監事主席：張俊盛教授(清華大學資訊系教授)

第八屆理事及監事任期自九十二年九月十九日起至九十四年九月十八日止。

中華民國計算語言學學會

第八屆第一次會員大會會議記錄

一、 會議時間：九十二年九月十八日(13:00~14:00)

二、 會議地點：新竹市清華大學國際會議廳

三、 出席人員：

應出席人數：166 人

出席：87 人

請假：79 人

四、 主席：張俊盛教授

秘書長：曾淑娟博士

紀錄：黃琪小姐

五、 報告事項

(一). 主席報告(張俊盛教授)

1. 本次大會召開之目的除了進行今年度會務之工作報告及 91 年學會收支決算、92 年收支預算提請通過外，最重要的是選出第八屆理監事。
2. 本屆之 Rocling 研討會為擴大舉辦，除了合併三個 Workshop 外，同時也增加學生論文發表及國科會研究計畫之展示。
3. 學會將持續的進行研究資料之收集並授權學術界研究使用，同時希望相關研究單位亦能提供資料庫之交換使用，共同促進學術之交流及發展。

(二). 秘書處報告(曾淑娟博士)

1. 學會網頁改版：

為了提高學會網頁更多服務及功能，學會之網頁作了大幅度的改版，並已申請新網址，新網址為 WWW.aclclp.org.tw，英文版已陸續即將建置完成，諸位會員若擬向國外友人提供學會相關訊息者，請提供新網址，在中文版尚未建置完成前，舊網址仍繼續開放。

2. 通訊改版：

學會通訊首頁版面將作適當之修改，下一期通訊各位會員即可收到新版之通訊，對於通訊之版面或內容若有任何提議，亦請不吝賜教。

3. 出版品：

學會目前之出版品包含「中文計算語言學期刊」、「ROCLING 論文集」、「COLING-2002 論文集」及由「中研院資訊所與語言所籌備處詞庫小組」共同出版之技術報告。有興趣者可於休息時間至櫃臺服務處參閱。另外，學會提供了歷年曾主辦之研習會或研討會之講義及論文集，將免費贈送與會者，有興趣的同仁可至櫃臺索取，數量有限送完為止。

4. 學會財務：

- (1) 學會投資海外基金投資金額 US\$40,000.-，自 87/10/13 起至 92/8/31 止，台幣盈餘為 \$111,243。
- (2) 學會 92/8/31 財務概況，請參閱附錄第 2~4 頁之收支決算表及資產負債表，若有任何疑問請提出。

(三). 學術委員會報告(吳宗憲教授)

辦理碩士論文獎評審，今年共八篇申請，評審結果為：優等從缺，佳作二篇，分別為成功大學資工所楊敦淇同學，指導老師簡仁宗教授及蔡銘峰同學，指導老師陳信希教授。

(四). 會員委員會報告(簡仁宗教授)

1. 今年度工作報告 博碩士論文摘要集製作報告：

本年度會員委員會之工作，將持續製作「博碩士論文摘要集」，91 年已製作完成「88~90 年博碩士論文摘要集」，並已分贈學會會員。本年度將增錄 91 學年度博碩士論文摘要，預計十月底前製作完成，屆時將由秘書處免費提供學會會員。

2. 九十三年度工作計畫報告 提供免費會員電子信箱帳號服務：

為配合學會新爭取到的網址 aclclp.org.tw，本委員會擬協助建立學會之電子信箱伺服器，預計於明年度開放會員申請電子信箱，凡為本會之會員者，學會將可免費提供電子信箱帳號，屆時學會所有相關訊息可透過此帳號及其原始帳號發送與會員。

(五). 口語處理組報告(陳信宏教授)

1. 7 月 31 日理監事會議通過，學會目前發行的語音語料庫 TCC300 將對國內學術界免費授權(酌收工本費)，作為學術研究使用。本語料庫是由台灣大學李琳山教授、成功大學王駿發教授、交通大學陳信宏教授提供的 300 人麥克風語料。
2. 目前正由國科會支持進行廣播、交談、電台節目等 spontaneous-speech database 以及台語 database 的錄製及處理，以及車內語料的錄製及處理。在告一段落後，希望完成的語料庫也能以同樣方式提供授權使用。
3. 希望有興趣的會員能提供其他語料庫或參與相關語料庫的製作。
4. 未來希望擴大到提供各種軟體或其他資料的免費學術使用授權。

(六). 資訊檢索組報告(簡立峰博士)

本組過去一年辦理過兩此研討會分別是 91 年 9 月與國科會協辦 Semantic Web 研討會，以及本次與計算語言學會年會合辦的「資訊檢索技術研討會」。此次研討會承蒙師大陳柏琳教授負責規劃，邀請多位老師進行專題演講，未來 93 年度也將舉辦 1~2 次研討會針對重要學術議題邀請專家者進行專題演講，同時也將輪流由不同學校或研究單位主辦，持續推動資訊檢索研究在國內的發展。

(七). 期刊編輯報告(陳克健博士)

1. 期刊進度報告

本會期刊—「中文計算語言學期刊 Computational Linguistics and Chinese Language Processing」, 從上次會期至今已出版了第七卷第一期與第二期(發行份數各 500 本)、及今年八月份同時出版第八卷第一期和第二期(發行份數各 400 本)。為了增加稿源及加快出版的速度, 本期刊增加了一些新的作法:

- (1) 在第七卷第二期裡, 增加收錄了 Project report 的部份。
- (2) 於第八卷第二期中, 已將 Special issue articles 及 Regular issue articles 合併出版。

本刊第九卷第一期, 預訂將於 2004 年 2 月出版, 發行份數為 400 本。此期目前預定要出刊的論文, 將會是從本次投稿至 ROCLING 的論文中收錄。至於第九卷第二期, 預訂於 2004 年 8 月出刊, 發行份數亦為 400 本。此期將出版專刊—New Trends of Speech Signal Processing, 並感謝王駿發老師幫忙邀稿, 在此甚表感謝。

本期刊目前仍有稿源不足的問題, 本期刊在此誠摯邀請大家多多投稿。

2. 語言學門國內相關學術期刊排序

由國科會人文處委託國立台灣大學黃宣範及張顯達教授主持的『語言學門國內相關學術期刊排序』, 本期刊在二十種期刊中綜合排序的結果得到第 5 名。本期刊在『期刊引用率』排名第二, 在『期刊主觀評價』排名第五, 然而在『形式要件』評比中僅得 65 分排名第十一。此一評比項目較無爭議, 可依所列計分項目一一加總即可。經自我評鑑加總後應得 71 分與實際分數 65 分有相當差距。已去函要求更正。若經更正實際名次應為第四。(昨日已收到黃宣範教授回覆, 先前所做『形式要件』評比數據確實有誤, 並已予以更正, 因此學會期刊在本次評比排名更正為第四名)

評比結果對本期刊是一種鼓勵也是一種激勵, 評比標準極具參考價值, 將做為本期刊未來改進之依據。

六. 提案

1. 追認九十一年度「收支決算表」、「資產負債表」、「財產目錄表」、「現金出納表」、「基金收支表」提請通過。
決議：無異議通過
2. 追認九十二年度收支預算表提請通過。
決議：無異議通過

七. 臨時動議

八. 選舉第八屆理監事及海外理事

監票：王小川、張景新

發票：黃琪

唱票：王新民、簡仁宗

計票：余明興、郭人璋

選舉結果：

有效票數:56 票

廢票：1 票

理事當選名單：

1. 鄭秋豫博士 (中央研究院語言所籌備處研究員)
2. 張景新教授 (暨南國際大學資工系助理教授)
3. 柯淑津教授 (東吳大學資訊科學系副教授)
4. 王新民博士 (中央研究院資訊所副研究員)
5. 許聞廉博士 (中央研院資訊所研究員)
6. 陳信希教授 (台灣大學資訊工程學系教授)
7. 簡仁宗教授 (成功大學資訊工程系副教授)
8. 陳信宏教授 (交通大學電信工程系教授)
9. 王駿發教授 (成功大學工學院院長)
10. 吳宗憲教授 (成功大學資訊工程系教授)
11. 簡立峰博士 (中央研究院資訊所副研究員兼副所長)

候補理事：

1. 張照煌博士 (工研院電通所前瞻技術中心資深正研究員兼副主任)
2. 劉美君教授 (交通大學外文系教授)

監事當選名單：

1. 陳克健博士 (中央研究院資訊所研究員)
2. 蘇克毅博士 (致遠科技股份有限公司總經理)
3. 鄭伯順副所長 (中華電信研究所副所長)
4. 余孝先博士 (工研院電通所前瞻技術中心主任)
5. 張俊盛教授 (清華大學資訊系教授)

候補監事：

1. 沈家麟博士 (台達電子研發中心經理)

海外理事當選名單：

1. 舒維都教授 Victor Zue (Massachusetts Institute of Technology, Cambridge)
2. 鄒嘉彥教授 Benjamin K. T'sou (香港城市大學教授及語言資訊研究所中心主任)
3. 李錦輝博士 Chin-Hui Lee (Georgia Institute of Technology, USA)

九 散會

中文資訊檢索標竿測試集第一版(CIRB010)

中文資訊檢索標竿測試集第一版 (CIRB010) 之建構是依據資訊檢索評估的相關理論，期望成為中文資訊檢索領域中一項重要的測試資源，從事中文資訊檢索或是跨語資訊檢索研究的學者專家均可使用該測試集，以評量所研發之檢索系統的績效。本測試集共分為三部分：文件集 (CIRB010DocumentSet)、查詢主題(問題集) (CIRB010TopicSet) 以及相關判斷 (答案集) (CIRB010RelevanceJudgment)。本「測試集」所有權為台灣大學圖書資訊學系語言資訊處理系統實驗室所有，擬授權學會發行，待完成授權手續後，將以 e-mail 方式及學會網站公布申請辦法，「測試集」簡要說明如下：

1. 文件集 (CIRB010DocumentSet)

共有 132,173 篇新聞文件，是由中國時報(38,116 篇)、中時晚報(5,747 篇)、工商時報(25,812 篇)、中央日報(27,770 篇)、中華日報(34,728 篇)等五家報紙之新聞資料，皆已取得正式使用權。

2. 查詢主題 (CIRB010TopicSet)

本測試集目前共有 50 個查詢主題 (CIRB010TopicEN001-050.xml 英文版，CIRB010TopicZH001-050.xml 中文版)，可視為資訊檢索者的資訊需求。每個查詢主題由數個欄位結合而成，並使用不同標記加以識別，其格式為「 < 欄位名稱> [欄位內容] </欄位名稱>」。主要呈現查詢主題內容的欄位為<title>、<question>、<narrative>、<keywords>，每部分均包含不同層面的主題資訊。另外，亦有<topic>與<number>二個識別性欄位。

3. 相關判斷 (CIRB010RelevanceJudgment)

本測試集相關判斷的實施，是由三位相關判斷者同時對候選文件集中之文件與查詢主題進行關聯程度之決策，將其量化後利用公式結合三者之結果，產生介於 0 與 1 之間的文件相關度 (R)，若 R 值愈接近 1 表該文件與查詢主題愈相關，反之則愈不相關。

判斷者對 50 個查詢主題的相關判斷結果，依查詢主題之編次分別列舉於 50 個純文字檔案中 (CIRB010RelevanceJudgment001 ~ CIRB010RelevanceJudgment050)。檔案中條列該查詢主題之候選文件集中，文件相關度大於 0 之文件 (依相關度由小至大排列)，並列出其相關度值 (R)，未列舉於其中者均為不相關文件。以下為相關判斷結果呈現之範例，其中「查詢主題編號」為查詢問句中<number> CIRB010TopicXXddd</number>之 ddd。

查詢主題編號	文件識別號	文件與查詢主題之相關度
	001	cdn_foc_0004196 0.333
	001	cdn_foc_0004185 0.444
	001	cts_pol_0003217 0.778
	001	chd_pol_0005892 0.889
001	cdn_foc_0005643	1

中研院語言所籌備處演講通告

Place: Room 704, Institute of Linguistics (Preparatory Office), Academia Sinica

Date: October 27, 2003

- 14:00-15:30 Kikuo Maekawa (National Institute for Japanese Language, Tokyo)
Japanese Spontaneous Speech Corpus and its application to the study of linguistic variation
- 15:30-16:00 Coffee Break
- 16:00-17:30 Jane Tsay (National Chung Cheng University, Min-Hsiung)
Tone acquisition in Taiwanese: a corpus-based study

Kikuo Maekawa

Japanese Spontaneous Speech Corpus and its application to the study of linguistic variation

Abstract:

Recently, there is a growing consensus among linguists and speech engineers that spontaneous speech is one of the most important objectives of the coming decades. To study spontaneous speech is not an easy task, however. The biggest hurdle, among many others, is the need for large reliable database. Study of spontaneous speech requires large amount of data, because spontaneous speech is by far more diverse and variable compared to read speech.

With a view to tackling this hurdle, the joint research group of the National Institute for Japanese Language, Communications Research Laboratory, and Tokyo Institute of Technology started compiling a large corpus of spontaneous Japanese called CSJ, or Corpus of Spontaneous Japanese in 1999. CSJ contains about 660 hours of spontaneous speech corresponding to about seven million words. The content speech is mostly monologue (623 hours), but small amount of spontaneous dialogue (16 hours) and read speech (20 hours) are involved also.

The primary application area of the CSJ is the development of probabilistic language data (like N-gram) for automatic speech recognition, but we tried to endow the corpus with as much annotations as possible for the linguistic and/or phonetic study of spontaneous speech.

Digitized speech (16kHz, 16bit), two-way transcription, two-way POS analysis, and speaker information are provided for the whole corpus to meet the requirements of speech recognition study. In addition to these, segmental labels, intonation labels are provided for a true subset of the corpus called the Core, which contains about 500,000 words (or 44 hours).

Moreover, prompt progress of the compilation work enabled us to annotate the Core with respect to the following additional information: location and strength of clause boundaries, dependency-structure of clauses, and, location and strength of discourse segment boundaries. These additional annotations are currently underway, and will be finished by the end of 2003. The whole corpus will be publicly available in the spring of 2004. Visit the following URL for more information about the CSJ (<http://www2.kokken.go.jp/~csj/public/index.html>).

From a point of view of linguistic research, an important design characteristic of the CSJ consists in its treatment of “speaking style”. Because the corpus was supposed to be a resource for the study of linguistic variations, it is equipped with multiple devices for the assessment of speaking style.

Firstly, CSJ contains two different types of monologue speech: academic presentation speech (APS, 299 hours) and simulated public speech (SPS, 324 hours). The former is the live recording of real academic presentation done in a dozen of academic meetings, and the latter is the studio recording of layman’s speech in front of small friendly audience on everyday topics like “My most delightful memory”. It turned out that the speaking style of SPS is clearly lower than that of APS, as expected at the time of corpus design.

Secondly, all APS and SPS talks were impressionistically evaluated with respect to their perceived speaking style using five-scale rating. It is expected that this subjective evaluation provide us with fine grading of speaking style within each speech type and across speech types. As a matter of fact, preliminary analyses of the CSJ revealed that the impressionistic rating of speaking style (and also that of spontaneity) was significantly correlated with all six linguistic variations analyzed so far.

Lastly, it is possible to make inter-speaker comparison of different speaking modes, i.e., monologue (APS and/or SPS), dialogue (on four different topics), and, reading. Although the cross-modal comparison is possible with 16 speakers only, this should provide precious information about the phonetic nature of so-called “tone-of-voice”.

Jane Tsay
National Chung Cheng University, Chiayi, Taiwan
Tone acquisition in Taiwanese: a corpus-based study

Abstract:

Research on phonological development has focused primarily on physiologically and perceptually motivated patterns, described in current phonological theories with markedness constraints. However, learning phonology also requires learning the particular sound patterns found in the adult language's specific lexicon. In this paper, we report preliminary analyses of tone acquisition data from the Taiwan Child Language Corpus (TaiCorp), addressing markedness, lexical factors, and their interaction.

TaiCorp is a 2.3 million word corpus based on more than 300 hours of recordings of spontaneous speech of children learning Taiwan Southern Min (Taiwanese) as their first language. The corpus is transcribed and coded using CHILDES (Child Language Data Exchange System, MacWhinney 1995).

Regarding markedness, physiologically motivated constraints on tone include (i) low tone is more difficult to produce than high tone and (ii) rising tone is harder to produce than falling tone (e.g. Ohala 1978, Maddieson 1978). Perceptually motivated constraints on tone include (iii) tones with closer F0's are perceived as more similar (e.g. Gandour & Harshman 1978). Based on longitudinal data (2;1-2;3) of seven children in TaiCorp, we found that (i) low tone has a much higher error rate than high tone; (ii) rising tone has a higher error rate than falling tone; (iii) high tone was often confused with mid tone but never with low tone, low tone was often confused with mid tone but almost never with high tone, and mid tone was often confused with both low and high tones. In other words, tone errors generally involve making tones less marked.

Regarding the lexical aspect of phonological acquisition, we found that the error rates of lexical tone show semantic transparency effects, i.e. semantically transparent disyllabic words (composed of two full-toned monosyllabic morphemes) showed higher error rates than semantically opaque ones (Tsay, Myers & Chen 2000). This pattern implies that phonological development may be sensitive to transitional probabilities across morphemes or syllables, an important lexical factor (see e.g. Church & Hanks 1990).

Our current research focuses on the interaction between markedness and lexical factors by examining word frequency effects (see also Tyler & Edwards 1993, Gierut & Storkel 2002). Based on longitudinal data from one child (1;7-2;0), we found that among words produced with tone errors, error rates were lower in higher frequency words. Since as noted above, tone errors usually involve making tones less marked, the frequency effect implies that tonal development involves learning to suppress innate tendencies through experience with adult models.

IJCNLP-04 Newsletter No.1
The 1st International Joint Conference of Natural Language Processing
organized by the Asia Federation of NLP associations (AFNLP)
Website: www.cipsc.org.cn/IJCNLP-04/

Date :

Main Conference: March 22-24, 2004

Workshops: March 25, 2004

Venue:

Sanya, Hainan island, China

Land's End - Hainan is so remote on the sea that ancient people, while believing that earth is square, really thought it is where the land ends

Honorary Chair:

Makoto Nagao (Kyoto)

General Co-Chairs

Guangnan Ni (Beijing)

Benjamin K. Tsou (Hong Kong)

Local Host:

Chinese Information Processing Society of China

Chair of the Local Organizing Committee

Youqi Cao (Beijing)

Vice-Chair:

Maosong Sun (Beijing)

Sponsoring Organizations:

Chinese Information Processing Society of China

Association for Natural Language Processing of Japan

Association for Computational Linguistics

Paper Submission

The information on paper submission will soon appear in the website

<http://www-tsujii.is.s.u-tokyo.ac.jp/ijc-nlp04/submission.html>

The important dates are as follows.

Paper submission deadline: November 15, 2003

(Note that we abolish the paper registration deadline of November 8)

Notification of acceptance: December 23, 2003

Camera ready papers due: January 24, 2004

Thematic Sessions

A Thematic Session provides a good occasion to focus peoples with the same special interest, and let them meet each other at a specific time-space to discuss and exchange ideas.

The following proposals have been accepted as thematic sessions. Please note that the deadline and procedure for submitting papers to these sessions are the same as those for general sessions. Also, the same quality standard will be applied to evaluate various submissions across general sessions and thematic sessions. You will find the detailed submission procedure in our website.

TS-1: Natural Language Learning using Both Labeled and Unlabeled Data

Organizer : *Hang Li* (Microsoft Research Asia, Beijing)

Recently, a new trend has arisen in the field of Natural Language Processing (NLP): the development of machine learning technologies that use both labeled and unlabeled data for training. Methods that have been proposed under this paradigm include co-training, EM learning, transductive learning, and other semi-supervised learning techniques.

For many NLP tasks, existing data are by their nature unlabeled and manually labeling them is prohibitively expensive. Effective utilization of both unlabeled and labeled data in learning is also a challenging but important issue. The goal of this thematic session is to bring together researchers working on this issue from different perspectives, in order to share their latest research results and to discuss future directions. We think that this session will advance research not only in exploiting unlabeled data but also in other natural language learning issues.

TS-2: Natural Language Technology in the Text Processing User Interface

Organizers: *Michael Kuehn* (Universitaet Koblenz-Landau, Koblenz)

Kumiko TANAKA-Ishii (University of Tokyo, Tokyo)

The emergence of applications like mobile text processing, communication aids and authoring support require sophisticated methods of text processing under challenging conditions. We invite researchers to discuss language technologies such as (but not restricted to) language modeling, analysis, summarization and disambiguation, in order to assist the user at the text processing front-end.

TS-3: Mobile Information Retrieval

Organizer: *Mun-Kew Leong* (Institute for Infocomm Research, Singapore)

One of the strongest impacts in recent information technology is the way mobility has changed computer applications. The rapid rate of handphone adoption, the ubiquitous PDA, and the low cost of wireless adoption has created new problems, new challenges, and new opportunities to researchers in many disciplines. One common thread through all these applications is the necessity for

information retrieval in one form or another. Another characteristic is the limited screen size of mobile devices and the consequent ramifications on input and output. The use of NLP plays an integral part in creating better user interfaces, better analysis of results for precise display, and greater understanding in the iterative interaction (dialogue) between user and mobile device. We propose this workshop to explore user oriented and theoretical limits and characteristics of NLP and IR within the context of mobile devices.

TS-4:Text mining in Biomedicine

Organizers: *Sophia Ananiadou* (Salford University, Manchester)
Jong C. Park (KAIST, Daejeon)

With biomedical literature expanding so rapidly, there is an urgent need to discover and organise knowledge extracted from texts. Although factual databases contain crucial information the overwhelming amount of new knowledge remains in textual form (e.g. MEDLINE). In addition, new terms are constantly coined as the relationships linking new genes, drugs, proteins etc. As the size of biomedical literature is expanding, more systems are applying a variety of methods to automate the process of knowledge acquisition and management. These include a variety of techniques such as statistics, machine learning, SVMs, deep or shallow linguistic or domain knowledge etc. Some NLP related topics are challenging in biomedicine such as: dynamic terminology management, named-entity recognition , integration with non-textual resources, discovery of named relationships, populating and updating existing ontologies / taxonomies. The aim of this thematic session is to examine issues and challenges in the area of biomedical text mining.

自然語言處理技術與生物醫學文獻

梁婷

交通大學資訊科學系

「使用語言，是人類最獨特，也是最顯著的能力」。 - Steven Pinker

一、前言

自古以來文獻為人類傳遞知識的主要媒介之一。今日科技的日新月異帶來大量的研究成果，也產生了無以數計的科技文獻。因此如何將這些以文字記錄的知識有系統並無誤地的從文獻中自動萃取，分析、整理，並呈現出，實有賴於有效的自然語言處理技術的開發與應用，以進行文獻內涵的探勘和訊息萃取。

近年生物醫學研究蓬勃發展，相關文獻快速累積。以往人工精心打造的知識庫的更新與建立耗時耗力，因此極需自然語言處理技術和文件探勘技術，來加速這方面知識的萃取和管理。這種知識庫系統的建立無疑地將可促進資訊的整合、交流和更新，甚至帶來生物醫學發展的突破和新的發現。

二、應用範圍

生物文獻的增長快速以果蠅資料庫參考文獻而言在近一百年間几乎是呈指數型的增長。自然語言處理技術應用在生物醫學資訊知識管理上可從關鍵詞預測、新詞的辨識、詞典的建構、語意網路的建立、文件及網站的分類、自動摘要、生物醫學資料庫的自動化建立、至生物醫學知識數位學習等。在這些應用中最基本的工作就是生物物件的名稱(named entities)和其間的關連性的辨識、分類和萃取，例如蛋白質、基因物件的辨識和其間的作用關係(interaction)的辨識。這也是近來發表在自然語言處理或生物資訊學相關期刊和會議中許多論文發表的主題。在本篇文中我們將究自然語言處理技術在資訊萃取工作上的應用來做一簡單的介紹。

三、研究機構

這方面的研究在許多國家已成立了專屬部門或大型計劃來加強研發能力，如美國的 National Center for Biotechnology Information (NCBI), 英國的 European Bioinformatics Institute (EBI), 瑞士的 Swiss Institute of Bioinformatics (SIB), 日本的 Japanese GenomeNet 等。大型的計劃如英國的 [Sheffield University](http://www.dcs.shef.ac.uk/research/groups/nlp/pasta/) 的 Protein ActiveSite Template Acquisition (PASTA) (<http://www.dcs.shef.ac.uk/research/groups/nlp/pasta/>) 可以說是其中一個較先 (1998-2001) 利用自然語言處理技術將生物文件資料作自動萃取，再轉換成格式化的知識庫的系統。另外東京

大學 Tsujii 教授的實驗室執行的 GENIA project (<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>)也是目前這方面有名的計劃之一。詳細資料可參考他們的網站。Tsujii 教授去年也應邀來台在 COLING2002 會議前給一個這方面研究的專題演講。

國內的研究單位也紛紛重視生物資訊的研究。目前國家衛生院(<http://www.nhri.org.tw>)和國家高速網路與計算中心(<http://www.nchc.org.tw>)都提供相當豐富的生物資訊的資源如重要的生物網站維護、生物知識庫和生物資訊系統的建立。各大學也陸續增設生物資訊學的學習課程以促進這方面的研發能力。

三、語料資源

文字是人類記錄和傳遞知識的主要工具。因此文獻的收集整理是知識蒐集的第一步。在生物醫學的領域中由美國衛生院醫學圖書館所建的線上文獻資料庫 MEDLINE 即收集了自 1966 年至今約一千兩佰多萬個文獻摘要和索引等，內容包括醫學、護理、保健等，涵蓋約 4600 多個生物醫學相關期刊。目前可透過國家生物資訊中心(NCBI)的搜尋引擎 pubmed (<http://www.ncbi.nlm.nih.gov/PubMed/>) 免費擷取 MEDLINE。每筆 MEDLINE 的文獻摘要目前都以數個 MeSH(Medical Subject Headings) 詞彙做資訊檢索用。MeSH 詞庫是由主題專家所更新。2003MeSH 包含超過 21000 個 main headings 可自由下載 (<http://www.nlm.nih.gov/mesh/meshhome.html/>)。另外醫學圖書館也發展多用途的 UMLS (unified medical language system, <http://www.nlm.nih.gov/research/umls/>) 包括 UMLS Metathesaurus, SPECIALIST Lexicon, 和 UMLS Semantic Network。這些資源可作為使用者建立醫學領域知識資源和相關詞彙應用程式的重要工具。

MEDLINE 的文件本身是未經標記的原始語料，尚需經過人工標記才能作為一般以機器學習方式的分類法的訓練語料。在資訊萃取處理上，目前廣為所用的已標記語料有 Bio1，它包含有 100 篇標記好的 Medline 摘要以及所用的 taxonomy 是由 Tateishi et al. 於 2000 年所建的。另外東京大學 Tsujii 教授的實驗室執行的 GENIA project 也建立了可供免費下載的 GENIA corpus (<http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/>)，到今年八月為止所提供的 version 3.02p 包含了 2000 篇詞性及語意標記好的 MEDLINE 語料。GENIA corpus 提供生物名稱語義標記上(含有三十多種語義標記)一個很好的訓練和測驗用的語料。生物醫學領域的相關資源非常豐富，搜尋引擎如 BioHunt (<http://www.expasy.ch/BioHunt/>) 和 BioMedNet (<http://biomednet.com/>) 亦提供網站搜尋和評鑑，可讓使用者快速地找到合適的網站。

四、主要議題

生物資訊的萃取目前主要的研究議題有兩項，一個是生物物件名稱(named entities)的辨識和分類，另一個是物件間的關係辨識和抽取。在名稱的辨識上如同新聞語料中所面臨的挑戰包括詞界、新詞、命名的不規則與不一致性、語義的多樣性、省略詞彙、縮寫、指代現象處理等問題。由於名稱的組成往往包含了三個以上的詞(以 GENIA Corpus 而言每個體名平均有

3.8 個詞) 是以詞界辨識的問題在名稱分類前需先予以解決。目前名稱的辨識有專注於單類物件如蛋白質名稱到多種類物件的辨識。

至於生物物件之間關係抽取的挑戰性在於句型語意表示方法的多樣性和關係存在的複雜度，諸如肯定關係、否定關係、未定關係、隱藏關係、崎異關係的確認。再者生物文件中，如 Genia Corpus，單句所含的生物物件平均數有 5.28 個。因此物件的單一或多重關係的處理需要進一步的辨識。至於跨語句的關係處理像 PASTA 系統是以處理指代現象的方式解決。這項議題需要多一些專家知識庫的協助來作關係確認。

五、處理過程

一般資訊萃取的流程包括文件語法和語意處理。語法處理有前置處理、詞性標記、語法剖析、省略(ellipsis)處理等。文件前置處理包括斷句、斷詞處理。簡單的方法只是將文件以空格和標點做依據。較複雜的方法包括同義詞名詞片語的概念訊息抽取。其次是詞根形態處理以減少詞彙量，最後是停詞處理。接下是詞性標記和語法剖析。語法處理已有許多開發好的處理工具，需要一些訓練來應用到生物醫學文件。省略現象的處理目前多以法則來解決。

語意處理有語意標記、指代現象處理、關係辨識和抽取整合。語意標記乃將詞彙標記成定義好的物件名稱類別如基因、蛋白質、藥名等。物件名稱辨識目前處理有兩種方法一種是規則法；另一種是利用機器學習未產生規則。手建的規則需要專家建立，故缺乏擴充性和可移植性。機器學習為主的辨識則需要大量的標記好的語料以達到可接受的成效。依據 IdentiFinder System 的結果分析顯示新聞語料中其名稱辨識結果與訓練語料量成對數(log)的增長。因此對機器學習的方法首要的挑戰之一包括如何簡易地產生足夠量的訓練語料。

關係辨識部分目前多著重在物件關係存在與否。因此可以視為單一句子的分類工作，一般可借助統計式的模組來進行大量的辨識處理。相對的關係抽取多倚賴語法剖析器的協助和主要動詞為主的語法模型比對。目前主要的關係辨識如 Genia Project (<http://www-tsuji.is.s.u-tokyo.ac.jp/GENIA/>)，是對蛋白質間的交互作用關係進行辨識與萃取。這類的訊息萃取極需許多知識的資源協助，包括語法剖析器、知識本体、索引詞庫、專業詞典、和監督式學習模組的開發。

六、結論

二十一世紀可以說是資訊網路與生物科技產業的世紀。其中生物科技又被譽為為希望工程，許多學術研究機構莫不積極發展。本文謹簡單介紹自然語言處理技術在生物科技知識管理的一些應用。希望藉以能引起有心從事這方面研究者的興趣一起投入相關技術的開發。

七、參考資料

相關期刊:

Bioinformatics, Genome Informatics, Nucleic Acids Research, Computational Linguistics, Machine

Learning, Natural Language Processing, IEEE Transactions on Information Theory.

相關學術會議:

Applied Natural Language Processing *ANLP*, International Conference on Computational Linguistics *COLING*, Conference on Natural Language Learning *CoNIL*, Message/Document Understanding Conference *DUC/MUC*, Association for Computational Linguistics *ACL*, National Conference on Artificial Intelligence *AAAI*, Pacific Symposium of Biocomputing *PSB*, International Conference on Intelligent Systems for Molecular Biology *ISMB*.