中華民國計算語言學學會
通訊 ACLCLP
The Association for Computational Linguistics and Chinese Language Processing
發行人：王駿發
主　編：曾淑娟
執行編輯：黃　琪
第十四卷第五期
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

語音資料庫工本費調降

語音資料庫 MAT-160Edu 及 MAT-2000Edu 自即日起調降工本費：

- MAT-160Edu 由原來之 2,000 元(US$70.-)降至 200 元(US$10)。
- MAT-2000Edu 由原來之 40,000 元(US$1,350.-)降至 20,000 元(US$700)。

國語語音資料庫開放申請

- MAT-400 學術版工本費 400 元 (USD20.-)
- MAT-2500ExtV 學術版工本費 10,000 元 (USD350.-)

資料庫簡介請參閱本訊第二　六頁。

開放技術報告委託發行

學會為本著促進計算語言學之發展及交流宗旨，開放代為發行學術單位研究人員之技術報告，委託代為發行之技術報告皆須經過本會之審查通過方可發行，申請的方式及索取申請書，請聯絡本會秘書處黃琪小姐。

中文資訊檢索標竿測試集-文件集開放申請

「中文資訊檢索測試集第一版(CIRB010)-文件集」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。工本費 NT$ 1,000(US$ 50.-)，詳細說明及申請方式請逕自上網查詢。

光華雜誌雙語語料庫試用版開放申請

本資料庫之來源來自光華雜誌，光華雜誌為一般性的雜誌，有多元的主題，引人入勝報導性的主體。目前遠流出版社透過學會發行研究用版本。完整收錄了 1976　2001 年的《光華雜誌》數位化文字資料，總共 312 期雜誌，4700 萬字，6000 篇文章，是台灣百科，更是認識台灣、學習英語、翻譯工作與拓展國際業務的珍貴資料庫。本資料庫可用於機器翻譯，跨語言資訊檢索，機器輔助語言教學等研究。申請試用版需先取得「光華雜誌語料庫」原始資料庫使用權方可申請，原始資料庫使用權之申請，請聯絡本會秘書處黃琪小姐。

學生出席國際會議補助

學生出席國際會議補助，原補助之會議只限 ACL 及 COLING，經本次理事會決議增加補助兩個會議：ACM-SIGIR 及 ICASSP，每個會議補助一　二名，補助金額由本會審查委員視區域及論文級別決議之，補助上限 US$1,000.-，申請日期為收到論文接受之通知二週內提出申請。詳細辦法請逕自上網查詢。

MAT Speech Database –MAT-400

1. MAT-400

These speech files are collected from 216 male and 184 female speakers through telephone networks. The speech signal is recorded under the following conditions.

sampling rate : 8kHz
bits per sample : 16
encoding type : 16-bit linear PCM
max. data length : 160,000 samples

They are grouped into five sub-databases. A detail description is given in the following table.

Sub-Databases	Number of Files	Prompting Item Numbers	Speaking Style	Description
MATDB-1	3600	1 - 9	spontaneous	short answering statements
MATDB-2	2000	10 - 14	read	numbers pronounced in five different ways
MATDB-3	4800	15 - 26	read	Mandarin syllables
MATDB-4	12000	27 - 56	read	words of 2 to 4 syllables
MATDB-5	4000	57 - 66	read	phonetically balanced sentences

Each recorded utterance is in a speech dat file. The file name has the format;

tnnmmiis.VAT,

where t -- database code in alphabet,
nn --content code in digits,
mm -- prompting sheet number,
ii -- item number on prompting sheet,
s -- sequence number.
VAT -- file extension name dedicated to MAT speech file

A Phonetic Symbol File (PTC file) is assigned to specify the contents of a set of speech data files. The contents in the PTC file are the designated utterances in sequence of item number. The file name of PTC files has the format;

tmm.PTC

where t -- database code in alphabet,
mm -- prompting sheet number,

For each item(ii) in a prompting sheet (mm), the format is,

(item no.) <Chinese characters>

<corresponding Chinese phonetical symbols> (“Ju4 In1 Fu2 Hau4”)
<corresponding phonetical symbol in Pin-Yin>.

2. Programs

Speech File Editing Program (VEDITOR)

This program runs under a Windows system of Chinese version. It requires the screen display size of 1024 x 768. It provides a tool for users to edit speech files. The file header parameters, as well as the waveform, are displayed on the screen. The user can edit the file header, modify the waveform, and playback the edited voice in an interactive mode.

File conversion (VATWAV)

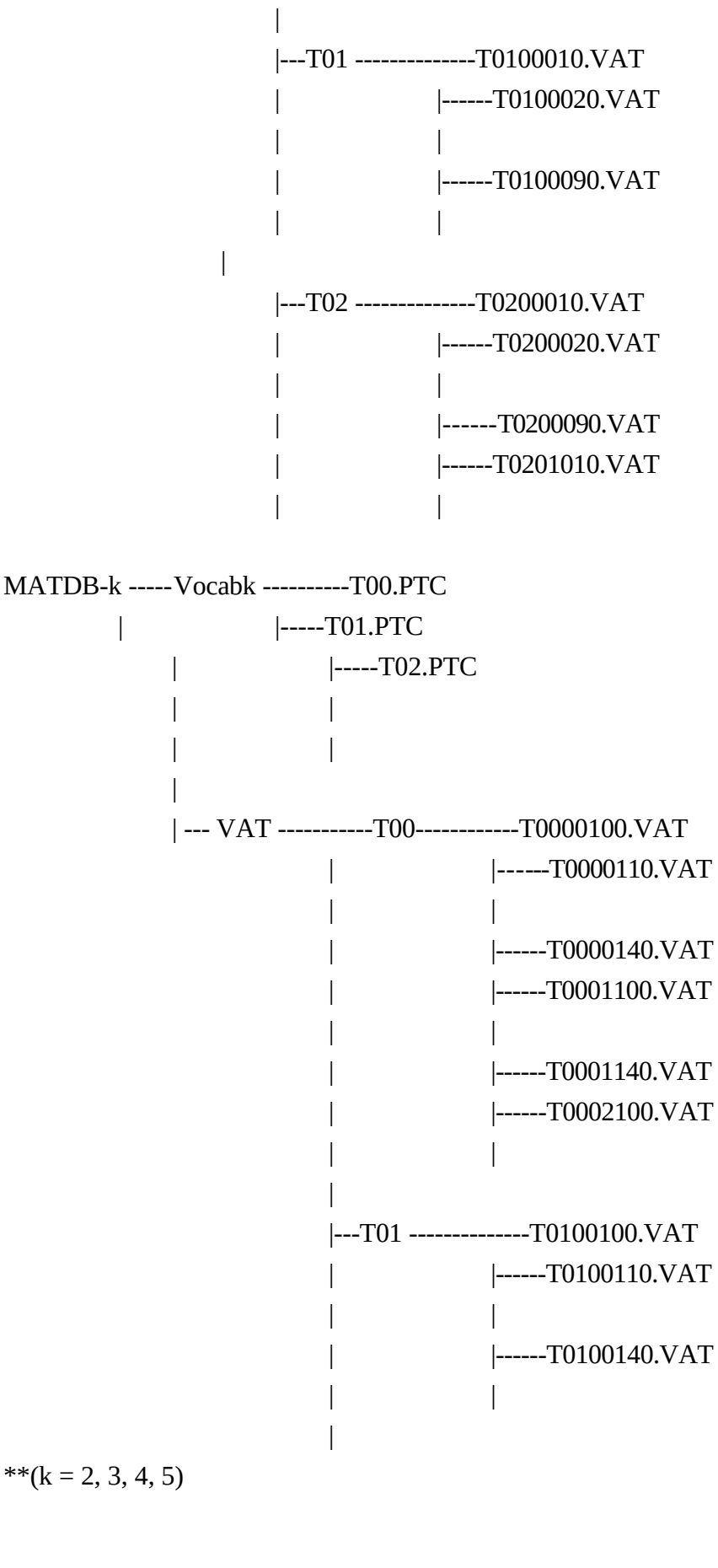
This program is for converting an MAT speech data file into *.WAV format. Then the facilities for *.WAV files can be applied to the recorded speech data.

3. File Structure

INSTRUCTION
- Mandarin-syllable-table.DOC
- MAT-400-Brief.DOC
- MAT-400-statistics.DOC
- MAT-400-Title-page.DOC
- MAT-file-format.DOC
- MAT-syllable-count-66.DOC
- Readme.TXT

PROGRAM
- VEDITOR.EXE
- POIN2.TAB
- SYL_IMF.TAB
- VATWAV.EXE

```
MATDB-1 --------   Quest   (nine questions)
                |
               | --- VAT --------T00------------T0000010.VAT
                           |                |------T0000020.VAT
                           |                |
                           |                |------T0000090.VAT
                           |                |------T0001010.VAT
                           |                |
                           |                |------T0001090.VAT
                           |                |
```


|
|---T01 ----------T0100010.VAT
| |------T0100020.VAT
| |
| |------T0100090.VAT
| |
|
|---T02 ----------T0200010.VAT
| |------T0200020.VAT
| |
| |------T0200090.VAT
| |------T0201010.VAT
| |
MATDB-k -----Vocabk ----------T00.PTC
| |-----T01.PTC
| |-----T02.PTC
| |
| |
|
| --- VAT ----------T00----------T0000100.VAT
| |------T0000110.VAT
| |
| |------T0000140.VAT
| |------T0001100.VAT
| |
| |------T0001140.VAT
| |------T0002100.VAT
| |
|
|---T01 ----------T0100100.VAT
| |------T0100110.VAT
| |
| |------T0100140.VAT
| |
|

**(k = 2, 3, 4, 5)

MAT Speech Database −MAT-2500ExtV

1. MAT-2500ExtV

This is a product of the joint validation project conducted by Association of Computational Linguistics and Chinese Language Processing and Computer & Communication Laboratories of Industrial Technology Research Institute, Taiwan. The original database is MAT-2500Ext where speech data are collected through telephone networks in Taiwan. The database contains files provided by 2573 speakers (1268 males and 1305 females). The speech signal is recorded under the following conditions.

sampling rate : 8kHz
bits per sample : 16
encoding type : 16-bit linear PCM
max. data length : 160,000 samples

They are grouped into four sub-databases. A detail description is given in the following table.

Sub-Databases	Prompting Item Numbers	Speaking Style	Description
MATDB-6	67 - 68	read	Numbers
MATDB-7	69 - 72	read	Special terms
MATDB-8	73 – 82	read	Words of 2 to 4 syllables
MATDB-9	83 - 90	read	Sentences

Each recorded utterance is in a speech data file. The file name has the format;

tnnxmiis.VAT,

where t -- database code in alphabet,
nn -- content code in digits,
x -- code of recording site
m -- prompting sheet number,
ii -- item number on prompting sheet,
s --sequence number.
VAT -- file extension name dedicated to MAT speech file

A Phonetic Symbol File (PTC file) is assigned to specify the contents of a set of speech data files. The contents in the PTC file are the designated utterances in sequence of item number. The file name of PTC files has the format;

tnnx.PTC

where t -- database code in alphabet,
nn -- content code in digits,
x -- code of recording site

For each item(ii) in a prompting sheet (nnx), the format is,

(item no.) <Chinese characters>
<corresponding Chinese phonetic symbols> ("Ju4 In1 Fu2 Hau4")
<corresponding phonetic symbol in Pin-Yin>.

2. Programs

Speech File Editing Program (**VEDITOR**)

This program runs under a Windows system of Chinese version. It requires the screen display size of 1024 x 768. It provides a tool for users to edit speech files. The file header parameters, as well as the waveform, are displayed on the screen. The user can edit the file header, modify the waveform, and playback the edited voice in an interactive mode.

File conversion (**VATWAV**)

This program is for converting an MAT speech data file into *.WAV format. Then the facilities for *.WAV files can be applied to the recorded speech data.

3. File Structure

INSTRUCTION
- Mandarin-syllable-table.DOC
- MAT-2500ExtV-Description
- MAT-2500ExtV-Brief.DOC
- MAT-2500ExtV-Title-page.DOC
- MAT-file-format.DOC

PROGRAM
- VEDITOR.EXE
- POIN2.TAB
- SYL_IMF.TAB
- VATWAV.EXE
- STRUCT.TXT

PTC-tnn1	PTC-tnn3
MATDB-6	MATDB-6
MATDB-7	MATDB-7
MATDB-8	MATDB-8
MATDB-9	MATDB-9
PTC-tnn4	PTC-tnn5
MATDB-6	MATDB-6
MATDB-7	MATDB-7
MATDB-8	MATDB-8
MATDB-9	MATDB-9
PTC-tnn6	PTC-tnn8
MATDB-6	MATDB-6
MATDB-7	MATDB-7
MATDB-8	MATDB-8
MATDB-9	MATDB-9

PTC-tnn9
MATDB-6
MATDB-7
MATDB-8
MATDB-9

文獻語料庫口語語料庫演講系列之二

地點: 中央研究院語言學研究所籌備處研究大樓 704 室
日期: 九十二年十二月十三日 (星期六)
主辦單位：中央研究院語言學研究所籌備處
議程：

14.00-15:30	Robin Lickley (Queen Margaret University College, Edinburgh, England)(1) The HCRC Maptask corpus: design and annotation(2) Using the HCRC Maptask corpus for psycholinguistic research on the production and perception of spontaneous speech
15:30-16:0	***茶敘***
16:00-17:30	Robert Eklund (Telia Research AB, Farsta, Sweden)A cornucopia of sine qua nons, provisos, caveats and aftermaths in a plethora of human--human and human--{machine|"machine"} corpora

THE HCRC MAP TASK CORPUS: DESIGN, ANNOTATION AND THE USE OF THE CORPUS IN PSYCHOLINGUISTIC RESEARCH ON THE PRODUCTION AND PERCEPTION OF SPONTANEOUS SPEECH

Robin Lickley
Queen Margaret University College, Edinburgh, England

Abstract

In the first part of this talk, I describe the design, collection and annotation of the HCRC Map Task Corpus.
The corpus consists of 128 dialogues between pairs of speakers (N=64), recorded on digital audio tape and video. In each dialogue, one speaker (Giver) describes a route through a "map" of an imaginary place consisting of illustrated named landmarks to another speaker (Follower), whose task it is to reproduce the route on their own version of the map. The two maps differ in certain respects (e.g. some landmarks are either not on both maps or are in different positions or have different names), which leads to confusion and discussion at various points in the exercise.
Landmark names were designed to illustrate various aspects of phonology.
Factors in the dialogue design include SPEAKER ROLE, EYE CONTACT, FAMILIARITY with the other speaker, task PRACTICE and speaker GENDER.
Each speaker performs as Giver twice, with the same map but different partners, and as Follower twice, with different maps and different partners (SPEAKER ROLE, PRACTICE).
In 64 dialogues, speakers are able to see each other, and in the remaining 64 dialogues, speakers are divided by a screen, so are unable to make EYE CONTACT.

Speakers interact twice with partners with whom they are familiar and twice with people that they had never met before the recordings were made (FAMILIARITY).
32 speakers in the corpus are male and 32 female (GENDER).
Transcriptions evolved in three main stages: first, transcribers worked from tape recordings producing word-level transcriptions; at a later date, signal processing software was used to produce transcriptions in which each word was time-stamped; in a third pass, when disfluencies were annotated, yet more accurate transcriptions were produced.

Annotation was performed on many levels: Dialogue structure (Transactions, Games and Moves); part of speech; syntactic structure; reference to landmarks (first and subsequent mentions); speaker gaze; prosody; disfluency. All the transcriptions and annotations were converted into XML, so that analysis on multiple levels is possible.

The HCRC corpus was recorded in Glasgow, Scotland, and most speakers were students with a local accent. The map task design has since been employed for other languages (e.g., Dutch and Italian). Other versions in English have been recorded by Lickley and colleagues (32 dialogues with the same maps with no written names for the landmarks; 12 dialogues with people who stutter) and for the ongoing "MONITOR" project, at Edinburgh and Glasgow Universities, where variables such as time pressure and feedback are used to test hypotheses about speech production processes.

In the second part of the talk, I describe some of the studies that have made use of the corpus. The main focus is on psycholinguistic studies of the production and perception of disfluencies.
Following on from earlier work which examined the point at which listeners could detect that an utterance was disfluent, the first perceptual study using Map Task utterances examined the "Transcriber Problem". In this work, we showed that spontaneously produced disfluencies could easily be missed by listeners and only transcribed poorly if perceived at all, even when they were specifically asked to listen carefully and give an accurate transcription of a short stretch of speech. We argue that a combination of the non-recognition of certain words and limitations of short term memory conspire to produce what appears to be a very useful failing by our speech perception mechanism. In a subsequent project, we attempted to model this behaviour in automatic speech recognition, with only limited success.

Other perceptual studies involving stimuli from Map Task dialogues include one which compares listeners' perceptions of the speech of people who stutter and people (in the original corpus) who do not stutter.
Disfluencies in the Map Task corpus have also been analysed from the point of view of production. As well as overall disfluency rates for the HCRC Map Task Corpus, we have analyses of disfluency types by dialogue move types: we describe how the rate of disfluency varies according to whether people have eye contact, whether they are male or female, whether they have the role of Instruction Giver or Follower, what type of dialogue move they are producing (e.g. more cognitively demanding moves produce a higher rate of disfluency than "easier" moves) and, following on from work on inter-turn intervals in the corpus, how these relate to the production of disfluencies.
For the Monitor project, we also examine the effects of visually presented listener feedback and time pressure on the production of disfluency.

The HCRC Map Task Corpus is available on CD-ROM from ELSNET and the LDC. Most of the annotation is available from the HCRC web pages.

A cornucopia of sine qua nons, provisos, caveats and aftermaths in a plethora of human–human and human–{machine|"machine"} corpora

Robert Eklund

Robert.Eklund@teliasonera.com

http://roberteklund.info

TeliaSonera, Voice Technologies, Sweden

Institutute of Computer Science, Linköping University, Sweden

Abstract

This paper describes methodological and linguistic aspects on the collection of a number of corpora spanning almost a decade of activities carried out at Telia Research/TeliaSonera, as well as private data collections. Both language and speech phenomena are treated, and differences as to channels, settings and instructions are

described.

1. Corpora

A short summary of the corpora included in this talk are given below.

1.1 Translation corpus …

As part of the *Spoken Language Translator* project (Rayner et al., 2001) —a speech-to-speech translation project running between 1993 and 1999, covering English, Swedish and French—language data was needed to build the Swedish language models. Before speech data were collected, a first corpus was created by translating existing American English speech data. At a first stage, a small numbers of researchers each translated a large number of sentences from English into Swedish, but in a second phase a small number of sentences were emailed out to a large number of Swedish speakers for translation, resulting in a huge number of translators. It was found that vocabulary size at the receiving end was positively affected as a function of number of translators.

1.2 …and a Wizard-of-Oz pilot

Since it is well-known that translations are always coloured by the source language (even in the case of professional translators), it was decided to stage a number of *Wizard-of-Oz* data collections. Wizard-of-Oz (WOZ) is a technique where you mimic the envisioned system (using mock-ups and actors), and have a number of subjects carrying out tasks with it this "system", while in fact the system is enacted by hidden humans (actors or researchers), thus approaching an authentic user–agent situation. Thus, a small WOZ pilot was staged, and a comparison with the translated data showed that even a very small WOZ corpus yielded idiosyncratic Swedish data that were not found even in huge amounts of translated data.

1.3 Wizard-of-Oz corpora (human–"machine")

During 1996 and 1997, two major WOZ simulations were made, resulting in around 7,500 utterances/54,000 words (the two collections pooled). These were subsequently used to train the speech models for the Nuance Swedish speech recognizer. Instructions in *WOZ-1* were given in partly verbal, partly iconographic form. Since it was shown that the verbal instructions were "copied" by the users, *WOZ-2* made use of only iconographic instructions.

1.4 Human–Human corpus

From a linguistic point of view, it is interesting to see what the differences between automatized systems and authentic human interaction might be. Thus a small number of subjects that had participated in WOZ-2 were given the exact same tasks to be carried out with two real-life, human, travel agents, working at the travel agency *Nymans*. This resulted in 1,734 utterances/9,250 words to be compared with the "same" dialogs with the believed-to-be machine in WOZ-2.

1.5 Bionic corpus (human–machine)

As the project went on, a up-and-running systems was created, and a final data collection were made with the functional automatized system. Thus, 1982 utterances/12,849 words were collected in the *Bionic* corpus.

1.6 Multimodal corpus

Since humans communicated not only by means of oral language, but also by dint of gestures, facial expressions, mouse and keyboard channels (in the case of computers) and so on, a multimodal setting was created in a joint project with the Royal Institute of Technology in Stockholm. The system handled apartments in downtown Stockholm, and accepted both keyboard, mouse and verbal input, and responded in the form of spoken language from an animated agent, graphical displays, text presentations and icons.

1.7 Tok Pisin corpus (human–human)

As a private project (not being financed by Telia), the researcher spent a couple of months in Papua New

Guinea during the period December 1999 to February 2000. During that stay, a corpus of completely authentic air travel dialogs was collected at the Air Niugini travel agency at that Kavieng Airport, New Ireland. This corpus was later compared to a similar corpus of similar Swedish speech data—part of the multi-site project *Swedish Dialogue Systems*—collected at two travel agencies in Lund, Sweden.

2. Linguistic analyses

A short description of some of the linguistic analyses carried out is given below.

2.1 Sundry observations

The corpora mentioned above have been under scrutiny from a wide variety of different angles, and have resulted in a number of articles over the years. These will be presented in the talk. Examples of such studies are e.g. a comparison of telephone data (Telia WOZ data) and multimodal data (AdApt), indicating differences both with respect to disfluency and syntax, and cross-linguistic observations, i.e. a comparison between Swedish and Tok Pisin data, and a study where Swedish data and American English disfluency data were studied (Eklund & Shriberg, 1998).

2.2 Disfluencies

The four WOZ simulations above have been used as the basis for the hitherto most exhaustive study of disfluency in Swedish (Eklund, forthcoming). Seven disfluency categories (unfilled pauses, filled pauses, segment prolongations, explicit editing terms, mispronunciations, truncations and repairs) have been studied, and comparison between the different channels and settings have been made. Also, comparison between the Swedish data and the Tok Pisin corpus have also been carried out. A fairly detailed description of the observations made will be given in the talk.

2.3 Ingressive speech

It has been claimed (Reeves & Nass, 1996) that, given that a system behave in any way like a human being, human users will interact with the systems as if it were a human being. However, when comparing the eight subjects that participated in both WOZ-2 and Nymans, it was found that ingressive speech (speaking on inspiratory airstream) was found only in the human–human setting (Eklund, 2002). Ingressive speech is a very common phenomenom in Swedish (and other languages), and is produced automatically. Eklund (2002) found that around 10% of all instances of affirmative "yes" were ingressive. While all subjects made use of ingressive speech in the human–human setting, not one made use of a single ingressive word in the human–"machine" setting, indicating that interaction with a machine is not viewed as equivalent to interaction with a human being.

3. Summary

The above description will be given in much more detail in the final talk, and other and additional observations, both methodological and linguistic, will be made. The talk will focus on both methodological issues dealing with crucial phenomena to take into consideration when designing a data collection session, with the goal to collect "as-good-as-possible" speech (and language) data, e.g. how instruction will inevitably affect the data yielded, but also what linguistic aspects are affected, both as a result of the data collection method proper, but also as a function of differences in settings (i.e. human–machine vs human–human, or simulated human–human vs authentic human–human).

國立台灣大學自然語言處理實驗室簡介

陳信希

國立台灣大學資訊工程學系

1. 研究目標

人類語言處理一直是人工智慧研究重要領域之一，如何讓電腦分析與生成人類語言，讓電腦與使用者以人的語言溝通是電腦科學終極目標之一。這個領域的研究近年來極為受到重視，麻省理工學院 2001 年元月/二月的科技評論就將其列為未來改變世界十大資訊科技之一；Gartner Group 在 2000 年年底也報導自然語言處理是未來十年最重要的十二項資訊技術之一。

國立台灣大學自然語言處理實驗室主要的研究目標如下：

(1) 語言工程基本理論和技術

語言工程是非常重要的知識工程，語言知識的表現、取得、運用、和整合是主要的研究方向。本研究也探討口語和書面語的差異，以及在語法、語意、和語用等不同層次上衍生的問題。這部份的成果支援大型應用系統的發展。

(2) 多語言資訊處理

網際網路突破空間距離，造成一個不分國界的資訊地球村，尤其透過全球資訊網，不同語言所呈現的資訊皆唾手可得。本研究主要在探討多語言資訊的分類、檢索、過濾、擷取、和摘要等問題，以其降低語言所帶來的障礙，發揮資訊的最大效能。

(3) 本土語言處理與應用發展

本土語言自動化處理有其獨特的問題，必須由我們自己去解決。在研究裡，我們不僅討論中文的斷詞、詞性標記、剖析、解譯等問題，而且也將成果擴及台語、客語、甚至於原住民語語料的分析和合成。第一套國台翻譯和台語語音合成系統已經完成。我們希望這個主題的研究，對台灣族群的融合、和文化的保存有積極的貢獻。

以下由三項研究目標中，分別各取一個主題，深入介紹台大自然語言處理實驗室過去的一些成果。

2. NE 擷取及應用

自然語言處理技術，已成功應用到資訊的擷取上，最有名的就是 DARPA 所支持的 Tipster Project，其下有一項 Message Understanding 技術評比(簡稱 MUC)，根據這項國際活動所開發出來的英文 named entities (NE)擷取技術，包括人名、地名、組織名、時間、日期、錢、百分號等文件重要成分，系統效能普遍已達 90%以上。NE 是文件的主要成份，掌握 NE 是理解文

件基本工作。不同的策略被提出來辨識 NE，例如語料庫為本的方法，用以擷取中文人名；規則式的方法用以擷取時間和日期、金錢和百分號等。過去的研究主要集中於一般領域，被運用於資訊檢索、問答系統、自動摘要等應用。

過去大部分的研究在單語 NE 的擷取，台大資訊系自然語言處理實驗室將其延伸到跨語言資訊檢索。我們首先提出以形素為基礎的中英專有名詞組音譯，接著進階到以音素為基礎。音素間相似度也由原來的人工設定到機器自動學習，中英文專有名詞組的組成規則和轉換規則也由語料庫中學習。

由於生物科技的快速發展，大量新的研究成果不斷地發表出來，如何掌握最新的理論和技術，避免重複的研究，開創新的成果，是生物科技研究人員一大挑戰。基本上科技論文是以人的語言(如英文)所書寫的文件，而自然語言處理技術就是處理文件的利器。運用自然語言處理技術，處理生物科技的論文，儼然成為生物資訊重要的研究項目之一。過去的研究探討如何由科技文章中辨識蛋白質名稱、基因名稱、細胞名稱、疾病名稱、藥物名稱，並嘗試找出其間的特定關係。

但是在生物資訊名稱辨識上，仍存在盲點，不是以簡易的 regular expressions 處理，就是採用 Gazetteer 策略，收集名稱表，以類似辭典查詢的方式為之。而名稱間的關係，幾乎每篇論文都是事先確定好，這種方式的優點是簡單不需要複雜的技術，但相對地也不易延伸到其他主題的資料庫。台大資訊系自然語言處理實驗室運用統計性自然語言處理技術，設計強健性名稱辨識技術，擷取重要的關係，建立重要成分間的關連圖，進而挖掘隱性知識，提供參考。

3. 自動摘要

自動摘要就是由文件、語音廣播、影片等不同媒體的資料中，自動擷取重要資訊，提供使用者參考，以節省閱讀大量資料所發的時間。其實不只是省時這項功能，網際網路檢索結果，通常也會擷取文件前幾個句子當摘要，提供使用者判斷文件相關的依準，這可降低某種程度網路塞車的現象。進一步的應用，則包括決策、導覽、規畫等方面的支援。自動摘要研究的歷史很早，是自然語言處理領域傳統研究課題之一。由於網際網路興起所衍生的許多應用，這項研究課題近年來備受重視。台大資訊系自然語言處理實驗室，由單一文件摘要研究到多文件摘要，並延伸到跨語言、跨媒體摘要。我們也研究特徵選擇對摘要的影響，及效能評估問題，和標題摘要產生的方式。

自動摘要研究，依摘要表現的方式可分擷取和摘錄兩種。擷取模型由原來的文件中抽取某個數量的句子，形成摘要。摘錄模型須對文件進行理解，抽取文件意涵，產生新的句子。單一中文文件自動摘要，以擷取模式為目標，因此如何挑選"重要"句子是必須考慮的要素，採用的技術、策略和目標歸納如下：

(1) 以段落為單位來挑選

一般人寫文章時，通常以段落來描述特定主題。如果要強調摘要結果句子的連貫性，

增加摘要之可讀性，由文件中選取”重要的”段落來作為摘要，是考慮的方向之一。但是文件段落長度差異很大，或沒有明顯的段落區隔時，如何以主題來切割文件，形成另一個問題。在挑選合適的段落上，可以把每個段落視為個別的小文件，利用檢索技術由段落集中挑選，以構成最後的結果。

(2) 以句子為單位來挑選

不管文件中有沒有段落分隔符號，這種策略把文件看成由一堆句子組成，因此沒有文件切割的問題。但是中文句子的定義，非常模糊。標點符號的運用，差異很大。缺項在中文非常普遍，文件中已經提過的觀念常常會被省略，這在電腦處理上的挑戰性很高。如何判定有意義的句子單位，如何處理指涉問題，以保持句子的連貫性，是中文自動摘要必須接受的挑戰。

(3) 挑選的原則

不管句子或段落的挑選，句子或段落的位置，是考慮的策略之一。以新聞稿子為例，為了吸引主編的注意，記者通常以倒金字塔型的方式來表達新聞內容，亦即最主要的內容會先在文件前面敘述，後面的部份用來補充說明，越後面越不重要，因此位置可反應重要性。其他挑選的參數有：文件名稱、重要詞組、詞頻、詞連慣性、詞共現性、人名、地名、公司名、組織名等線索。採用不同的線索，就得引進不同的處理方法。例如：要得到詞組，必須進行句子部分剖析；要取得詞彙間的關係，必須運用統計分析技術。

(4) 摘要的長度

在挑選句子或段落時，摘要的長度是考量的要素。我們知道摘要的功能之一是節省閱讀文件所發的時間，摘要越長發的時間越多，摘要越短覆蓋的主題有越少的傾向。因此在摘要選取時，如何兼顧到長度和覆蓋度兩者之間的平衡點，是必須探討的課題。

(5) 系統評估

評估非常重要，可以藉以判斷摘要策略、基本模組、和摘要程序設計的好壞。評估的要素有摘要長短、閱讀時間、系統運作時間、摘要的準確性等。最理想的狀態是系統在最短的時間，由文件中摘取最合適長度的資料，而讀者只要發很短的時間，就可以依不同的應用，做出最佳的反應。當然，在很多"最好"的要求下，如何調配各個影響參數，是研究的目標，而系統評估可以反應系統是否達到既定的目標。

過去在多文件自動摘要的研究是以線上新聞網站為例，將新聞資料彙整摘要。資訊的來源，可能是同一天的新聞，也可能是同一主題不同天的新聞。對於前者，我們將來自不同新聞網站的線上新聞匯集在一起，如何依談論的主題先作規類，是研究課題。接著如何分析新聞稿談論內容相同和相異之處，是另一重點。最後，進行摘要和結果呈現。對於後者，我們探討事件偵測和追蹤的問題，屬同一事件的新聞才作摘要。

當考慮的範圍擴大到多語言文件的自動摘要時，不僅需探討不同文件間的相同和相異處，因為不同語言的加入，挑戰性也加大。我們延續過去在機器翻譯系統的成果，運用詞彙歧異分析技術，來識別不同語言文件間的相同和相異處。使用者可以依其所熟悉的語言順序，來閱讀文件。

4. 本土語言處理

國內是多語並存的社會，國語、台語、客語、原住民語是我們平常使用的語言，在不同語言文化的保存極為重視的時代，如何運用資訊技術處理我們自己的語言是我們的責任。我們運用過去所開發出來的語言處理技術，設計一套本土語言互譯及語音合成系統(網址：http://nlg.csie.ntu.edu.tw/systems/TWLLMT/)，使用者以中文輸入，可以國語語音、台語語音及客語語音輸出，這對國內不同語言之學習及溝通有大的助益。

使用者在進入這套系統，輸入想要翻譯的中文詞或句子以後，可以勾選要聽國語、台語還是客語的語音輸出，每一種語言又會有不同的選項。輸入並選擇完成後，按下「傳送」以後，系統就會以該語言的翻譯結果呈現在網頁上。翻譯結果的網頁上還有一個按鈕，按下去後就可以聽到這句話的發音，這是語音合成的結果。

在台語部份，有漳州腔或泉州腔的發音或變調規則可供選擇，使用者可以自己試試看這兩種腔調有什麼特色。另外有文讀與白話兩種不同唸法，如果想用台語讀古文或詩詞，就要用文讀音來讀。

此外，在呈現台語的翻譯結果時，因為有些詞的中文字表示法還有爭議，所以這些詞直接用教會羅馬字的方式拼音出來，使用者可以看到中文字或全用羅馬拼音寫出的句子。因為台語很多情形都會變調，在勾選「變調後拼音結果」以後，使用者也可以了解句子中哪些字會變調、變成什麼調。

在台語語音合成部份有不錯的成果後，我們便與臺大客家社合作，發展屬於客語的翻譯與語音合成部份。由臺大客家社同學提供詞典、錄音及程式的寫作，台大資工所自然語言處理實驗室提供技術上的支援，完成了這套系統的客語語音合成部份。

在客語部份，一樣有「說話音」和「讀書音」的選擇。一般說話時的句子就用「說話音」來唸，古文詩詞就改用「讀書音」來唸。翻譯結果會用中文字和羅馬拼音各寫一次，使用者就可以自己對照學到每個詞的寫法，或是播放語音合成來學習唸法。

相關論文

A. NE 擷取及應用

Hsin-Hsi Chen and Jen-Chang Lee (1994). “The Identification of Organization Names in Chinese

Texts." *Communication of Chinese and Oriental Languages Information Processing Society*, **4**(2), Singapore, 1994, 131-142 (in Chinese).

Hsin-Hsi Chen and Jen-Chang Lee (1996). "Identification and Classification of Proper Nouns in Chinese Texts." *Proceedings of 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, August 5-9, 1996, 222-229.

Hsin-Hsi Chen and Guo-Wei Bian (1998). "White Page Construction from Web Pages for Finding People in Internet." *International Journal of Computational Linguistics and Chinese Language Processing*, **3**(1), 1998, 75-100.

Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding and Shih-Chung Tsai (1998). "Proper Name Translation in Cross-Language Information Retrieval." *Proceedings of 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Quebec, Canada, August 10-14 1998, 232-236.

Hsin-Hsi Chen, Yung-Wei Ding and Shih-Chung Tsai (1998). "Named Entity Extraction for Information Retrieval." *Computer Processing of Oriental Languages*, Special Issue on Information Retrieval on Oriental Languages, **12**(1), 1998, 75-85.

林偉豪，陳信希(2000). "**反向異文字音譯相似度評量方法與跨語言資訊檢索**." **第十三屆計算語言學研討會論文集，台北，台灣，**2000 **年八月二十四日-八月二十五日，**97-113。

Wei-Hao Lin and Hsin-Hsi Chen (2002). "Backward Machine Transliteration by Learning Phonetic Similarity." *Proceedings of 6th Conference on Natural Language Learning*, 31 August – September 1 2002, Taipei, Taiwan, 139-145.

Hsin-Hsi Chen, Changhua Yang and Ying Lin (2003). "Learning Formulation and Transformation Rules for Multilingual Named Entities." *Proceedings of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, July 12, Sapporo, Japan, 2003, 1-8.

Wen-Cheng Lin, Changhua Yang and Hsin-Hsi Chen (2003) "Foreign Name Backward Transliteration in Chinese-English Cross-Language Image Retrieval." *Proceedings of Workshop of Cross Language Evaluation Forum*, 21-22 August, Trondheim, Norway, 2003.

Hsin-Hsi Chen (2003). "Spoken Cross-Language Access to Image Collection via Captions." *Proceedings of 8th European Conference on Speech Communication and Technology*, September 1-4 2003, Geneva, Switzerland.

Wen-Juan Hou and Hsin-Hsi Chen (2004). "Enhancing Performance of Protein and Gene Name Recognizers Using Collocation." To appear in *Journal of Biomedical Informatics*, Special Issue on Natural Language Processing in Biomedicine: Aims, Achievements and Challenges.

B. 自動摘要

Hsin-Hsi Chen and Sheng-Jie Huang (1999). "A Summarization System for Chinese News from Multiple Sources." *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*. November 11-12, 1999, Academia Sinica, Taipei, Taiwan, 1-7.

Hsin-Hsi Chen and Chuan-Jie Lin (2000). "A Multilingual News Summarizer." *Proceedings of 18th*

International Conference on Computational Linguistics, July 31-August 4 2000, University of Saarlandes, 159-165.

Hong-Jia Wong, June-Jei Kuo and Hsin-Hsi Chen (2001). "Headline Generation for Summaries from Multiple Online Sources." *Proceedings of 6th Natural Language Processing Pacific Rim Symposium*, November 27-29 2001, Tokyo, Japan, 653-660.

June-Jei Kuo, Hung-Chia Wung, Chuan-Jie Lin and Hsin-Hsi Chen (2002). "Multi-document Summarization Using Informative Words and Its Evaluation with a QA System." *Proceedings of The Third International Conference on Intelligent Text Processing and Computational Linguistics*, Lecture Notes in Computer Science, LNCS 2276, 2002, 391-401.

Hsin-Hsi Chen, June-Jei Kuo and Tsei-Chun Su (2003). "Clustering and Visualization in a Multi-Lingual Multi-Document Summarization System." *Proceedings of 25th European Conference on Information Retrieval Research*, Lecture Notes in Computer Science, LNCS 2633, April 14-16, Pisa, Italy, 2003, 266-280.

Hsin-Hsi Chen, June-Jei Kuo, Sheng-Jie Huang, Chuan-Jie Lin and Hung-Chia Wung (2003). "A Summarization System for Chinese News from Multiple Sources." Accepted by *Journal of American Society for Information Science and Technology*.

C. 本土語言處理

林川傑 陳信希(1999). "**中文到閩南語之線上翻譯及閩南語之語音合成**." **數位博物館中之語文處理技術研討匯集刊**, 1999 **年** 5 **月** 8 **日**, **南港**, 11.1-11.4.

Chuan-Jie Lin and Hsin-Hsi Chen (1999). "A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan." *International Journal of Computational Linguistics and Chinese Language Processing*, **4**(1), February 1999, 59-84.