

自然語言人機介面的知識庫建構輔助工具

蔡妙真 孫三為 許仁榮

交通部電信總局電信研究所

摘要

隨著自然語言處理技術的發展，以自然語言為人機介面的遠景十分看好，舉凡各種資料庫的查詢或是一些設備的控制，都是可能的目標，而知識的取得和知識庫與資料庫的格式或資料結構的訂定，是發展人機介面最大的問題，也是一個完整的自然語言對話系統發展的瓶頸。

本文首先探討在人機對話系統中自然語言人機介面相關知識庫的建立方式，並介紹本所發展中的知識庫建構輔助工具：語料分析模組提供一個語料分析的模式，配合語料分析編輯介面，提供建構人員一個良好的分析環境；詞典建構輔助模組則提供一個詞彙輸入介面及詞彙使用範例，輔助詞彙的建立；語意樣板分析輔助模組自分析後的語料中找出所有句子的語意樣板；對話庫建構輔助模組則提供與對話有關的資料的蒐集；至於頻率統計輔助模組則提供各種統計方面的訊息。

一、簡介

「自然語言人機介面」有別於傳統上以「菜單驅動」(menu-driven)方式的人機介面，乃是指以人類日常生活中所使用的語言做為與機器交談的工具。由於自然語言十分的口語化，且在使用上不需經過如鍵盤輸入方式般的訓練，再加上在對話的過程中常可省略一些訊息，因此十分靈活，應用也較廣。

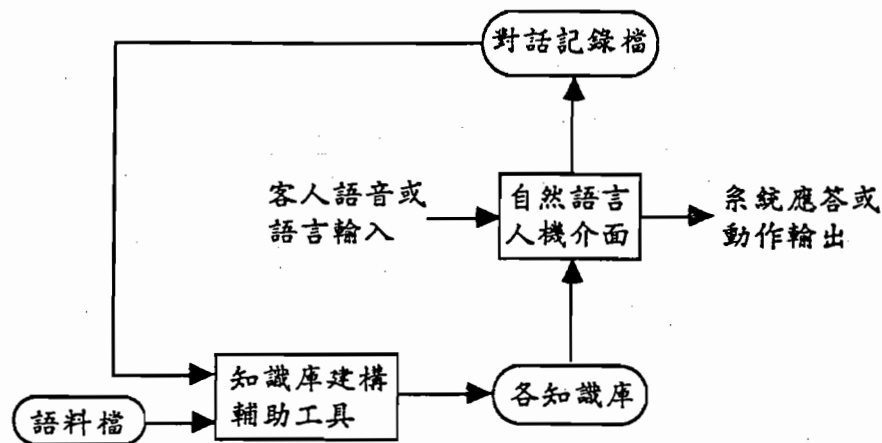
以自然語言為人機介面的研究在國外已有多年的經驗，以GUS為例[1]，它以英文與旅客對話，幫助旅客達成訂位的目的；在國內，以

自然語言爲人機介面的研究則是一個尚待開發的領域，由於語音技術持續的發展，因此以自然語言爲人機介面的遠景十分看好，舉凡各種資料庫的查詢或是辦公室的自動化，都是可能的目標。如何在眾多潛在的應用領域中加快人機介面的發展，是一個重要的課題。

建構以自然語言爲對話系統人機介面所遭遇的問題主要有二部分，一部分是整個對話系統所牽涉到的知識庫；另一部分則是程序性的，主要包括如何利用知識來決定應對。程序性的部分主要包括句意解譯與中間表現的解譯。句意解譯的主要憑藉是詞典，但是它與詞典的格式關係密切，與詞典的內容則較無關，因此即使要處理的領域換了，只要新領域詞典的格式不變，句意分析的部分就不需要大的修改；中間表現解譯器主要是解釋中間表現的內容，並產生應對語句，因此它與中間表現、語境、資料庫與對話庫等的內容及格式都有關聯。由以上的分析可知，知識的取得和知識庫與資料庫的格式或資料結構的訂定，是發展人機介面最大的問題，也是一個完整的自然語言對話系統發展的瓶頸，如何透過一些輔助工具幫助語料的分析以加速系統的發展，此即「知識庫建構輔助工具」的發展由來。

圖一就是知識庫建構輔助工具與自然語言人機介面的關係圖。由圖中可知，語料是知識的最大來源，而語料則可能是由實際對話蒐集而得或是在實際應用人機介面後所蒐集到的對話記錄檔。經由實際人機對話記錄的回餽、分析，並更新相關的知識庫，可提高系統的效率。由於知識庫的更新可能是一個經常性的工作，因此如何縮短整個回餽更新所需的時間就有賴知識庫建構輔助工具的幫助。

本文第二節首先探討在人機對話系統中自然語言人機介面相關知識庫的建立方式；第三節介紹建構輔助工具中的語料分析模組；第四節則討論各種訊息蒐集及詞典建構的輔助工具；最後則是結論與未來展望。



圖一、知識庫建構輔助工具與自然語言人機介面的關係

二、人機介面知識庫

2.1 自然語言人機介面相關知識庫的建立方式

自然語言人機介面相關知識庫的建立方式可分為幾個步驟：

1. 擬定系統目標

在建構系統之前首先必須確定系統的主要方向，決定系統的用途。例如，我們必須先決定該系統是要用於人事資料的查詢、104資料的查詢或是從事總機值機員的工作。

2. 蒐集語料

系統目標決定後就必須展開語料蒐集的工作。語料的來源主要有兩方面，一個是由日常生活中的實際對話蒐集而來，例如本所現有的電話轉接語料庫以及104電話號碼查詢語料庫，都是由實際對話錄音而得；另一種蒐集語料的方式則可用問卷，或由工作同仁記錄常用於系統目標的對話而得。這兩種方式以前者為佳，因為由此蒐集的語料是真實的語料，可反應出實際對話的真實情況，但是由於這種語料的取得十分困難，對整個系統來講通常也不完整。以電話轉接語料庫為例

，雖然它包含十卷錄音帶（共十小時），約 2500 個對話的語料[2,3]，但是對系統而言，它仍不能視為是一個完整的語料庫。

「人造語料」是蒐集語料的另一種方式。由人造的語料常易鑽牛角尖，也往往不易反應或模擬一些真實但細微的情況，因此實用性較真實語料差，但是透過嚴格篩選之後，大體上仍能反應出一些用法。在語料真實語料不足甚至缺乏的情形下，人造語料仍是不失為一可行之道。

3. 訂定系統功能

擬定系統目標並完成語料的蒐集之後，就必須訂定系統欲提供的功能。系統可提供的功能大部分可以由語料中看出，但是由於蒐集語料時，扮演系統角色的是人，能掌握外界的訊息，因此他所能處理的事多而複雜，要完全仿照人是一件不容易的事；另一方面，電腦難以掌握外界的變化但是卻能記憶比人更多的資料，可用以開發新功能。因此，在訂定系統功能時可截長補短，做適當的增刪。以電話轉接系統為例，由語料中可發現一些屬於服務台式的功能，例如值機員打電話告訴同仁「午飯送來了」，類似這一類的服務可加以刪除；又如「話機跟隨」的功能則可因為使用電腦的緣故，而擴充為「同仁話機及狀態的設定」。

4. 語料的增刪

系統欲提供的功能訂定後，在分析語料之前必須對語料做一些篩選與新增。語料中必須刪除的部分主要是與系統欲提供的功能無關的對話；而要新增的語料則屬於人造語料，這主要是基於「真實語料的不完整性」與「因應新增功能」等兩項的考量。

5. 訂定語料中句子所代表的意義

為滿足客人需求，系統與客人之間必須有一些對話，對於客人所說的話都必須做適當的應對，而瞭解客人所說的句子意義是系統能

做適切回應的前提，因此在分析語料之初，必須先就語料中客人所說的話做一個初步的歸類，訂出句子所代表的意義。

6. 對系統語料所顯現的一些問題做系統化的整理與策略性的考量

對話是互動的，對話的雙方常可藉由自己所說的話來影響對方接下來的談話，因此一個對話系統的應答就顯得十分重要。以電話轉接系統為例，由系統真實的語料中可發現一些人們在對話過程中常見的「冗贅」與「善變」，而這些情況都造成系統在處理上的困擾。下面就是構建電話轉接系統時必須考量的一些問題：

a. 冗贅的處理

語料中可發現一些不必要的應對情況，下列就是<<1A-44>>的部分對話：

Op: 研究所。
Cus: 喂！電話斷掉。廿。
Op: 推斷掉。
Cus: 厂廿、。

由對話中可發現，總機值機員的不當應對造成客人與總機值機員之間不必要的對話，這種不恰當的應答有時甚至會造成不可預知的輸入內容，因此，系統該如何應對是處理冗贅所必須解決的問題。以上述對話為例，下列的對話中總機值機員的應對顯然是較理想的。

Op: 研究所。
Cus: 喂！電話斷掉。廿。
Op: 這樣子啊！請問你要找那位？

b. 系統應對不一致的問題

在真實語料中，值機員對客人同一要求的處理方式可能因時間、客人聲音清楚的程度、甚至他當時的興緻，而有所不同。以轉接要求為例，值機員的處理方式就有下列幾種：無任何回答，直接轉接；應答後，轉接；重複客人所說人名或分機，而後轉接；重複客人所說人名或分機，等客人確認後，再轉接；重複客人所

說人名或分機，等客人確認，再應答後，轉接。因此在建構系統時，必須對系統運作的方式做一選擇。

除了上述問題外其他如：總機值機員或客人聽不清楚；客人所給訊息不完整；客人所給訊息有誤；不合理的對話等狀況的處理方式，都必須在建構系統時訂定。

7. 語料分析

語料是系統知識的主要來源，它提供了系統詞典及對話庫的內容，也提供系統在處理各種狀況時的依據，在語音辨認方面，語料也提供一些有用的訊息，而這些訊息的取得全靠良好的語料分析。

語料分析牽涉語料分析的模式。大部分的語料分析模式都局限於各個句子，這種各句獨立分析的方式有其缺陷，因為它無法考慮到語境，這在一個對話系統中尤其重要。除了句子的表面意義、深層意義、和語境之外，在對話系統中，系統的回答和所採取的步驟都是重要的資訊，因此一個好的語料分析模式至少必須能含括這些部分。

8. 知識蒐集

語料經分析後必須能將所分析的結果擷取出來供對話系統使用，此即知識的蒐集。

9. 擬定中間表現

自然語言的處理常常是多個層次的，尤其在系統的功能逐漸增加變得複雜後，系統很難由句子的字就知道該採取什麼行動，因此中間表現有其重要性；此外，由於系統功能可能有所增刪，因此中間表現必須具有可擴充性，對不同意義的描述也必須能有所區分。

中間表現依系統的需要也有不同的考量，以電話轉接系統為例，由於它所要處理的範圍有限，且相關的資料庫均需自己建立，因此中間表現的格式可自己訂；反之，對104查號系統或一般公司的物料查詢系統而言，由於這些資料的數量龐大，且多已建立，因此對話系統的中間表現最好與該資料庫的「查詢語言」(query language)的格式相同，至

少必需有這兩種格式的轉換程式，以便能直接利用資料庫查詢的功能及已存在的資料。

2.2 自然語言人機介面相關知識庫的建構輔助工具

在整個知識庫的建立過程中，語料分析與知識的蒐集是兩個最重要的步驟，由於這兩部分的難度較高、重複性大、所需的人力時間也較多，因此建立相關的建構輔助工具是提升效率的最有效的方式之一。圖二就是本文所提出的可應用於自然語言人機介面相關知識庫的建構輔助工具的系統架構。

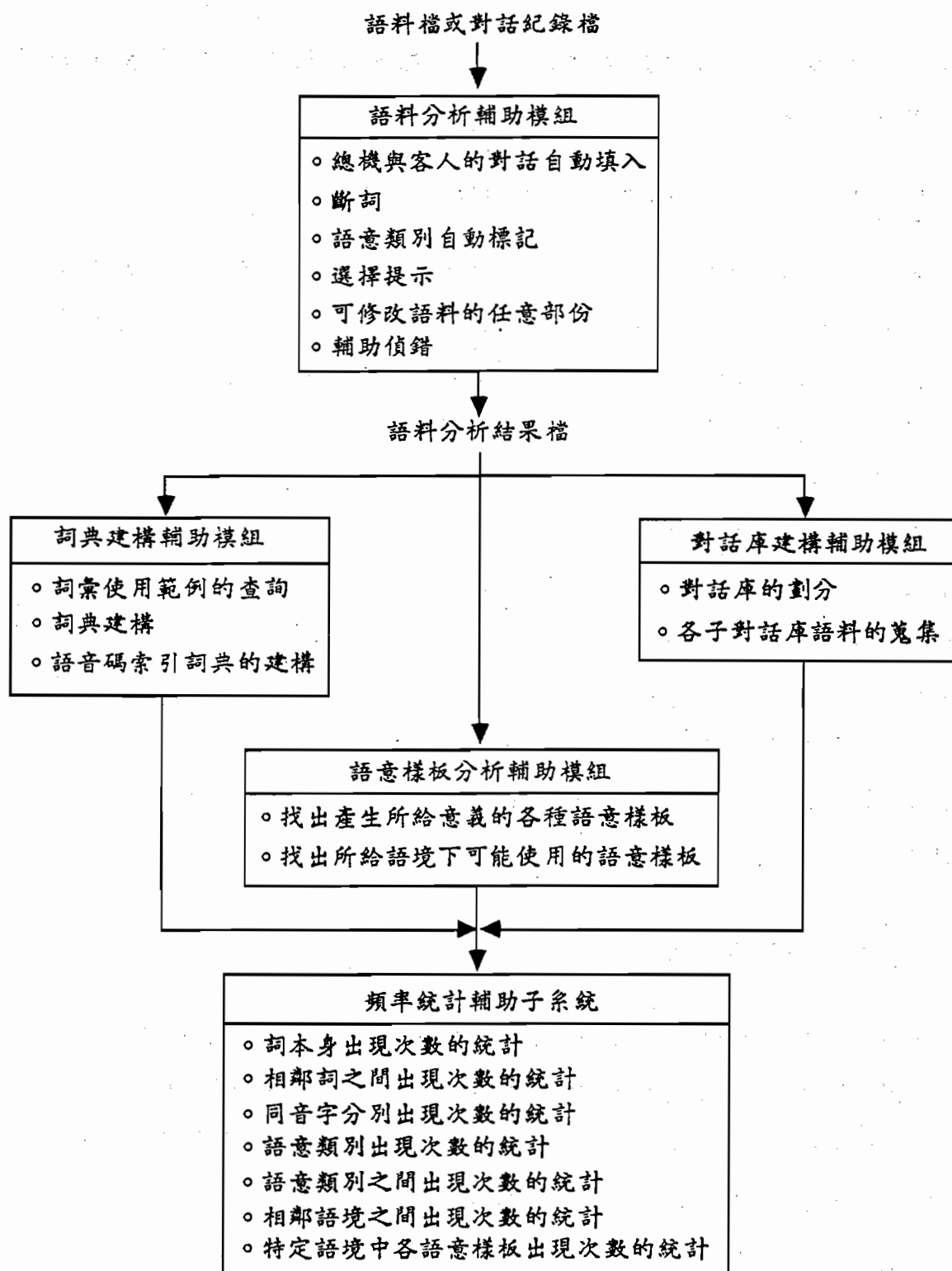
圖二中包括五個模組。語料分析輔助模組將讀入的語料檔或是對話記錄檔經過斷詞、語意類別標記之後，由分析人員確認並填入其他資料，產生語料分析結果檔，這是語料分析的部分；詞典建構、語意樣板分析、對話庫建構、及頻率統計等輔助模組則是屬於知識蒐集的部分，它們依各自之需要擷取語料分析結果檔中的各種訊息，供建構人員利用。本文將在第三、四節分別介紹語料分析輔助工具及知識蒐集輔助工具。

三、語料分析輔助工具

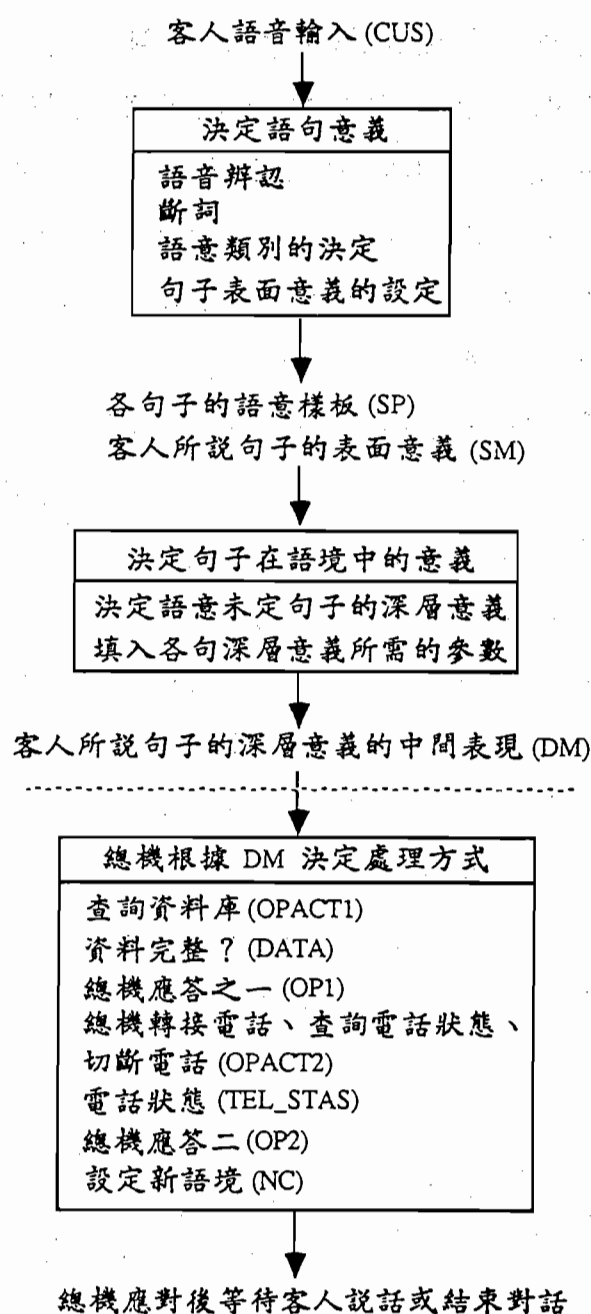
如同上述，語料是系統知識最大的來源，而且獲得困難。但是即使給了充分的語料，語料中能提供那些知識，如何分析語料以擷取這些知識？這都是一個重要的課題。本節提出本系統的語料分析模式，並說明此語料分析輔助工具所具備的特性。

3.1 語料分析模式

在一個對話系統中，由於人機對話間的互動，語料的分析已不能侷限在單一個句子裡，而必須考慮到同一個對話中的語境以及該對話系統的應用環境，圖三就是以電話轉接查詢為例，系統模擬總機值機員處理



圖二、知識庫建構輔助工具



圖三、總機轉接查詢系統的模擬流程

客人對話的步驟。在圖中，虛線以上是句意了解的部分：客人語音輸入(CUS)，經句意分析可得到客人所說句子的語意樣板(Semantic Pattern, SP)及句子的表面意義(Surface Meaning, SM)，經由語境(Context, CT)分析，得到客人的深層意義(Deep Meaning, DM)。圖三中虛線以下的部分是應對處理的部

分：依據客人所說的話，系統大都先查詢資料庫(OPACT1)，根據資料查詢的結果(DATA)，或先產生應答(OP1)再執行轉接等動作或直接動作(OPACT2)；有時轉接時電話在講話中或是沒人接(TEL_STAS)，系統必須將此訊息告知客人(OP2)。由於系統的應答有時改變了客人原先的意圖，而引導客人到一個新的語境，此客人與系統互動的結果也是一個重要的資訊，必須記錄下來，就是 NC(New Context)。

根據以上的分析，語料分析的螢幕格式如圖四所示。在圖四中，輔助模組首先填入語料檔檔名及語料檔中對話的總數，由於每一個對話通常都是由值機員開頭，我們以 OPINIT 記錄它；其次一個對話常包含好幾個「值機員－客人」的配對(CUS-OP)，由於螢幕無法同時容納，因此每一螢幕只描述一個 CUS-OP 對，而提供一些功能鍵做對話之間的切換；由於系統以代號記錄資料，「提示選擇區」提供建構者這些代號的意義。除了上述提到的 CUS、SP 等之外，每一個 CUS-OP 對尚包含客人與值機員所說的話的注音代碼，圖五就是以 <<1A-2>> 為例的分析結果，當游標指在 CUS 時，螢幕顯示 CUS 所對應的語音，透過游標的移動與空白鍵等功能鍵，建構者可選則或更改任意字的讀音。下面就是一個對話的資料結構：

```

(對話編號  OPINIT
  ( (CUS    (CUSER ...) (PHONE ...) (STACK ...) (SP ...)
    (SM ...) (CT ...) (DM ...) (WS ...))
    (OP     (OPER ...) (OPPHONE ...) (OPSTACK ...) (OPACT1 ...)
    (DATA ...) (OP1 ...) (OPACT2 ...) (TEL_STAS ...)
    (OP2 ...) (NC ...)))
  ...
  ( (CUS ...)
    (OP ...)))

```

其中 PHONE、STACK 和 OPPHONE、OPSTACK 分別記錄客人與值機員所說的話的語音代碼，WS 則是是否做過斷詞的標記。

3.2 語料分析輔助工具的特性

綜合上述的語料分析模式及設計理念，本語料分析輔助工具具有

語料檔檔名：		語料分析		對話總數：	
對話編號：	語料編號：	OPINIT：	CUS-OP 編號：		
CUS:					
SP					
SM		CT		DM	
OP:					
OPACT1 DATA		OP1	OPACT2	TEL_STAS	OP2 NC
選擇提示區					
輸入狀態					

圖四、語料分析的螢幕顯示格式

語料檔檔名: TELSC1.0K		語料分析		對話總數: 70	
對話編號: 1	語料編號: <<1A-2>>	OPINIT: 研究所.	CUS-OP 編號: 1		
CUS: 請接二五二.%					
SP					
4 7 9 SM		CT		DM	
7				7	
OP: 二五二是啊%,稍等一下啊.%					
OPACT1	DATA	OP1	OPACT2	TEL_STAS	OP2 NC
1	1	9 29	2 3	4	
<—△✓ ㄣ—ㄝ 儿、 ×✓ 儿、					
插字					

圖五、游標指在 CUS 時，螢幕顯示 CUS 所對應的語音

下列的特性：

- (1). 自動填入：由於語料都已存放在電腦中，因此在語料分析模式中的 Cus 和 Op 的欄位資料可由語料庫中直接獲得填入。
- (2). 自動斷詞：客人與值機員的應答常是一個句子或是片語，語料分析輔助工具在讀入未經斷詞的句子時（WS 未經標記）就自動斷詞，並加上斷詞標記 \$。系統只利用「長詞優先」的斷詞方法，語料分析人員必須檢查系統斷詞的結果做更正，必要時需做斷句（加入 % 的符號）的工作。
- (3). 語意範疇自動標記：本文所提之語料分析模式以語意為主，每個詞都有自己的語意範疇（Semantic Category, SC）。在自動斷詞後，系統以語意範疇相鄰出現的頻率為基礎自動標記各詞的語意範疇，同樣的，這個自動標記的結果仍需經過語料分析人員的檢查。
- (4). 選擇提示：在分析語料之初，系統所欲提供的功能和各個句子的意義都已分類，因此在輸入這些欄位的資料代碼時，可將各代碼所代表的意義依序列出，供分析人員選擇，以節省時間。例如在語料分析模式中，CUS 和 OP 的提示是注音符號，SP 的提示是語意範疇，SM、CT、DM、NC 和的提示則是句子意義的類別。圖六就是在輸入語意範疇時，語料分析的螢幕格式。
- (5). 語料的新增、刪除和修改：在分析語料時，系統提供分析人員線上新增、刪除和修改語料的功能。
- (6). 輔助偵錯：在分析語料的過程中，分析人員可能會誤填或是歸類錯誤，系統提供了兩個功能可幫助語料分析人員找出錯誤。第一項功能是詞意類別中詞的查詢。此項功能可將某一詞意類別中的所有詞找出，這可提供語料分析人員一個檢視語意類別分類的適當與否和各詞歸類的正確性的方法，圖七之1就是一個找尋出來的結果。另一項輔助偵錯的功能是詞的某一語意類別出現位置的查詢。以圖七之1為例，歸在請求語的「您好」及「你好」顯然有問題，透過查

語料檔檔名: TLOSC1.OK

語料分析

筆話總數: 70

對話編號: 1

語料編號: <<1A-2>>

OP INIT: 研究所.

CUS-OP 編號: 1

CUS: 請接二五二.%

SP

4 7 9

SH

CT

DM

7

7

OP: 二五二是啊%,稍等一下啊.%

OPACT1

DATA

OP1

OPACT2

TEL_STAS

OP2

NC

1

1

9 29

2 3

4

1.聽不懂

2.招呼語

3.問候語

4.請求語

5.幫助語

6.意向語

7.轉接語

8.指稱語

9.數字

10.人代稱語

11.人

12.姓

13.名

14.職稱

15.稱謂

16.單位泛稱

17.單位名稱

18.地點

19.地址泛稱

20.地址標識

21.地址

22.電話

23.確認語

24.否決語

25.否定語

[插字]

圖六、輸入語意範疇時的顯示螢幕

語意類別 = 3.問候語

您好 你好 你早 早

語意類別 = 4.請求語

請 麻煩 您好 好不好 你好 可不可以 可以 好嗎
拜託 好吧 能不能 麻煩一下 拜託一下 勞駕

語意類別 = 5.幫助語

幫 跟 給

語意類別 = 6.意向語

要 想 看

1.找出詞意類別中的所有詞

找尋的詞= 你好 語意類別 = 4.請求語
TLOSC1.OK <<1A-19>>找尋的詞= 您好 語意類別 = 4.請求語
TLOSC1.OK <<1A-15>>

2.找出錯誤歸類出現的語料結果檔和對話編號

圖七、系統的輔助偵錯功能

(7). 易於轉換到不同的應用領域：雖然本文的語料分析模式以電話轉接詢問對話為例，但是由於所有資料查詢的對話模式相近，因此只要修改系統的選擇提示，可很容易的運用到其他的領域。

語料分析人員利用語料分析模組填入對語料分析的結果，而知識蒐集的輔助工具則是將這些資訊依各自的需要分別找出，供語料分析人員做更深入的分析，這包括詞典建構輔助模組、語意樣板分析輔助模組、對話庫建構輔助模組、頻率統計輔助模組等

在詞典建構時，語料分析人員最需要的是各個詞在語料中出現的位置及其用法，因此詞典建構輔助模組中提供了一項詞彙查詢的功能，將同一個詞但具不同讀音(PH)、不同語意(SC)、以及該詞出現的前後三個具有不同語意的詞彙、語意（相關詞彙與意樣板）及該句子的語境(CT)與深層意義(DM)顯示在螢幕上，圖八就是「請問」的詞彙使用範例。以第二個句子為例，描述了「請問」的讀音為1（即くーん×ん），其語意範疇是「詢問事務語」（SC=41），而該句子出現在客人確認本所之後（CT=5），欲表達的是轉接同仁要求（DM=8）。

1. WORD：存放該詞之內碼。

詞彙查詢

找尋的詞= 請問 <-2VX7、;				
語意類別 = 41. 詢問事務語				
語境類別 = 2. 未定				
8. 轉接同仁 4. 問候 5. 確認本所 7. 轉接分機				
10. 問本所地址 13. 問本所事務 14. 問同仁狀態				
15. 問同仁分機 16. 問同仁單位 17. 問同仁地址 20. 問沒人接				
PH	SC	CT	相關詞彙與語意樣板	對話編號
1	41	5	你10那13那13	8 <<4A-79>>
1	41	5	能不能4 你10一下11你10	10 <<3B-78>>
1	41	5	是不是42億聚17	5 <<6A-71>>
1	41	8	喂2	2 <<10B-124>>
1	41	8	簡12志誠13先生15	8 <<9B-101>>
1	41	8	劉12政13	8 <<8B-18>>
1	41	8	吳12碧環13在不在47	8 <<4A-143>>
1	41	29	他10他10 他10大概57幾144	12 <<1A-119>>
1	41	29	那1 鄭12成深13鄭12	8 <<10B-10>>

PageUp: 往上一頁 PageDown: 往下一頁 F5: 找尋另一個詞彙 ESC: 離開

圖八、「請問」的詞彙使用範例

2. WORD_COUNT: 存放該詞在特定領域中出現的次數。
3. PHON: 存放該詞的語音編碼串列。
4. PHON_COUNT: 存放該詞發這個音的次數。
5. SEM: 用來存放語意訊息，其下分成TEST、SC、ASSIGN、LR、SEM_COUNT

五項。分述如下：

- (1) TEST: 存放供辨義用的條件。
- (2) SC: 即「語意範疇」(Semantic Category)。
- (3) ASSIGN: 指定該詞對本系統而言的服務類別及詞彙本身的特徵。
- (4) LR: 即詞彙衍生規則(LEXICAL RULE)。
- (5) SEM_COUNT: 同一詞而有不同的語義(SC)時，其間的頻次也有不同，SEM_COUNT即是記錄同一詞不同語義之間的頻次。

案使用範例、詞項格式及詞彙衍生規則，以利詞項的建立，圖九就是「請問」的詞項螢幕格式。

詞典建構					
找尋的詞=請問 <-<✓×ㄣ、;					
語意類別=41.詢問事務語					
語境類別=2未定 4問候 5確認本所 7轉接分機					
8轉接同仁 10問本所地址 13問本所事務 14問同仁狀態					
15問同仁分機 16問同仁單位 17問同仁地址 20問沒人接					
PH	SC	CT	相關詞彙與語意樣板	DM	對話編號
1	41	5	你10那邊18那邊18	10	<<5B-8>>
1	41	5	那個8王12副所長14	8	<<4A-79>>
1	41	5	能不能4你10一下1你10	10	<<3B-78>>
1	41	5	是不是42億悉17	5	<<6A-71>>
1	41	5	喂2	2	<<10B-124>>
PHON: <-<✓×ㄣ、 SEM_NO: 1					
TEST:					
SP: 41					
ASSIGN: (JOB_TYPE QUERY2)					
LR: (5)					
1.A->A了 2.A->AA 3.A->A看看 4.A->AA看 5.A->A一下					
6.A->A-A 7.A->A過					

圖九、「請問」的詞項建構螢幕顯示範例

如圖九「請問」的使用範例所示，其讀音只有一種（即<-<✓×ㄣ、），雖然它可出現在不同的語境下，但其意義都相同（SC=41），因此建構人員即可填入TEST（無條件成立）與SC，並設定其特徵是一詢問詞（JOB_TYPE=QUERY2）；由於「請問」可利用第五條詞彙衍生規則衍生成「請問一下」，因此在LR部分可填入(5)。

4.2 語意樣板分析輔助模組

本系統以語意範疇為詞歸類的依據，由語料分析模式中的SP、CT和DM三個欄位，我們可做各系統功能的語意樣板分析，這相對於以語法分析所找出的文法規則，而這些資料可用於語音辨認。語意樣板分析模組主要提供兩項資料的查詢，一項是找出產生所給意義的各種語

和DM三個欄位，我們可做各系統功能的語意樣板分析，這相對於以語法分析所找出的文法規則，而這些資料可用於語音辨認。語意樣板分析模組主要提供兩項資料的查詢，一項是找出產生所給意義的各種語意樣板，另一項則是找出所給語境下可能出現的語意樣板。以電話轉接系統為例，要知道客人使用那些語句表達轉接分機的要求，可將語料中所有DM為7（代表要求轉接分機）的資料取出，即可得知所有可能用來表達轉接要求句子的語意樣板；要知道客人在要求轉接分機之後，接下來會使用的句型，可將的所有CT為7的語料取出即可，這描述了客人所用句型、語意和對話語境的關係。

4.3 對話庫建構輔助模組

對話庫是系統在了解客人的輸入之後，用以執行動作產生應對的依據，對語料分析模式中的SM、Ct、DM、Op1、OpAct、Op2和NC等欄位的訊息做更精確的描述，可加快系統對話庫的建構，甚至可用以自動產生對話庫。本節就以<<1A-46>>和<<1B-3>>兩對話來描述對話庫的建立方式。

<<1A-46>>和<<1B-3>>對話的內容如下：

<<1A-46>>:

Op: 喂。
Cus: 麻煩接二八八。
Ding-T
Op: 電話在講話中。
Cus: 還在講話中啊。
Op: 對，嘿。
Cus: 好，謝謝。

<<1B-3>>:

Op: 研究所。
Cus: 喂。
Op: 找那一位。
Cus: 周萬民。
Op: 周萬民啊？
Cus: 是。
Op: 電話在講話，等會兒打來吧。
Cus: 等一會打來啊？
Op: 嘿嘿，對。
Cus: 好，謝謝。

利用語料分析輔助工具，對應於<<1A-46>>的語料分析過程如下所示：

Cus	SP	SM	Ct	DM	OPACT1	DATA	Op1	OpAct2	Op2	NC
麻煩接二八八	4 7 9	7	0	7	1	1		2 3	26	29
還在講話中啊	57 35 52 42	19	29	19			13			29
好，謝謝	23 26	24 27	29	29 27	4					

而 <<1B-3>> 的語料分析過程如下所示：

Cus	SP	SM	Ct	DM	OPACT1	DATA	Op1	OpAct2	Op2	NC
喂	2	4	0	4			7			8
周萬民	12 13	2	8	8	1	1	10			
是	23	24	8	8				2 3	26 33	26
等一會打來啊	48 29 42	26	26	26			13			29
好，謝謝	23 26	24 27	29	29 27	4					

其中SM、Ct、DP、和NC欄位中的2、4、7、8、27、和29等分別代表「要求不定」、「問候」、「轉接分機」、「轉接同仁」、「致謝」、和「結束」；Op欄又細分為Op1和Op2，記載的主要是系統對客人所說的話直接的回應或反問，如13代表「確認前句」，直接肯定客人所表達的意思，7代表「你要找誰」，要求客人提供完整資料；Op2則是系統依當時電話狀態所主動提出的建議，如26的「電話在講話中」、和33的「等會兒打來吧」；至於OpAct欄位中的1、2、3、4，分別代表系統在應對的過程中所做的事，其中1代表資料庫查詢，2代表轉接電話，3代表偵測電話狀態，4則代表系統掛斷電話，結束對話。在這個語料分析模式中，由於限於版面，因此TEL_STAS並未列出，而這些省略的訊息將在下面篇幅中補列。

由於對話庫所記錄的資料眾多，因此在建立對話庫之前，我們先對對話庫做一劃分，將其分為許多子對話庫。對話庫的劃分可以是一個語境的語料構成一個子對話庫，也可以由數個語境的語料構成一個子對話庫。以上述<<1A-46>>和<<1B-3>>為例，可分成主子對話庫、問候子對話庫、轉接子對話庫、電話狀態子對話庫和結束子對話庫，其所處理的語境分別是{0}、{4}、{7 8}、{19}、和{26 27 29}。

在對話之初無任何語境，因此可將語料庫中Ct是0（即對話的棟一句）的語料找出，即構成主劇本的資料，而每個子對話庫其語料的來源如下：

對任一子對話庫，所描述的語境集N，其語料的來源為：
 {語料 | 語料中的Ct欄位 $\in$ N 或 $(DM \in N \wedge DM \neq Ct)$ }

以上面的<<1A-46>>和<<1B-3>>為例，可得到五個對話劇本：

□主對話庫：Ct=0

SM	Ct	DM	Opact1	data	Op1	OpAct2	tel_sta	Op2	NC
7	0	7	1	1		2 3	1	26	29
4	0	4			7				8

□問候子對話庫：Ct=4 or (DM=4 and DM $\diamond$ Ct)

SM	Ct	DM	Opact1	data	Op1	OpAct2	tel_sta	Op2	NC
4	0	4			7				8

□轉接子對話庫：Ct $\in$ {7 8} or (DM $\in$ {7 8} and DM $\diamond$ Ct)

SM	Ct	DM	Opact1	data	Op1	OpAct2	tel_sta	Op2	NC
7	0	7	1	1		2 3	1	26	29
2	8	8	1	1	10				
24	8	8				2 3	1	26 33	26

□電話狀態子對話庫：Ct=19 or (DM=19 and DM $\diamond$ Ct)

SM	Ct	DM	Opact1	data	Op1	OpAct2	tel_sta	Op2	NC
19	29	19			13				29

□結束子對話庫：Ct $\in$ {26 27 29} or (DM $\in$ {26 27 29} and DM $\diamond$ Ct)

SM	Ct	DM	Opact1	data	Op1	OpAct2	tel_sta	Op2	NC
24 27	29	29 27	4						
26	26	26			13				29
24 27	29	29 27	4						

主對話庫是用來做為選擇子劇本的工具，並不產生應對，因此只需利用SM和DM兩個欄位。在這五個對話庫中，主對話庫描述了「若客人所

說句子的是7，則叫用轉接子對話庫；若客人所說句子的是4，則叫用問候子對話庫」。子對話庫中描述的則是系統根據客人所說的話及語境進行處理，並產生應對的過程，有時它也描述了對話庫間的轉移關係。以轉接子對話庫中的第一列為例，「客人以分機號碼要求轉接(DM=7)，總機值機員首先檢查該分機是否合理(Opact1=1)，因該分機號碼在合理範圍內(data=1)，所以總機值機員就執行轉接(Opact2=2)，結果發現電話在講話中(Opact2=3, tel_sta=1)，因此總機值機員將此訊息告訴客人(Op2=26)，並跳到結束狀態子對話庫(NC=29)。」

除了各對話庫語料的取得外在建立對話庫時還應注意語料的合併和對話庫的轉移。由於語料眾多，因此難免有重複或是相近的情形，我們可將這些語料加以合併，減少冗贅。以結束子對話庫為例，第一列和第三列就可加以合併。而轉移對話庫的方式及時機如下：

C t = 空的 → 保留原對話庫，跳到DM所指的對話庫

N C ≠ 空的 → 保留原對話庫，跳到N C所指的對話庫

C t ≠ DM → 離開原對話庫，跳到DM所指的對話庫

4.4 頻率統計輔助模組

由語料分析模式中的各欄位我們可做下列頻率統計的工作，而這項分析的結果可用於語音辨認及其他相關的研究：

1. 相鄰語境之間出現的次數 $F(\text{Context}_i | \text{Context}_j)$ ：利用語料分析結果中的CT和DM兩欄位，可找出相鄰語境之間出現的次數。在不同的語境下，客人所可能說的句子不同，用詞也不一樣。例如一開始對話時客人決不會說「這樣子啊」或是「好」等句子，而是要求轉接、詢問或是問候等語境的句型。那種語境的句子出現較多較多？這可供語音辨認後處理時利用。

2. 特定語境下各種語意樣板出現的次數 $F(SP_i | Context)$: 利用語料分析結果中的CT和SP兩欄位, 可找出特定語境下各種語意樣板出現的次數。在同一語境下可能出現的句型及其次數也是一個有用的資料。
3. 語意範疇之間相鄰出現的次數 $F(SC_i | SC_j)$: 利用語料分析結果中的SP欄位, 可找出語意範疇之間相鄰出現的次數。不同的語意範疇能出現的位置不同, 能相鄰出現的範疇也有限。
4. 語意範疇出現的次數 $F(SC_i)$: 利用語料分析結果中的SP欄位, 可找屬於某一語意範疇的所有的詞出現次數的總和。
5. 詞之間相鄰出現的次數 $F(W_i | W_j)$: 不同的詞之間相鄰出現的頻率有助於選擇下一個詞。
6. 詞出現的次數 $F(W_i)$: 不同的詞出現的頻率不同, 可幫助決定詞的邊界。
7. 同音字分別出現的次數 $F(W_i | WORD_SYLLABLES)$: 同一個音可能對應到許多的字, 每一個字出現的次數為何? 這有助於字的選擇。

大量的訊息統計提供了許多的資訊供各種自然語言人機介面的處理之用, 這是本系統的特點。

五、結論與未來展望

本文提出了一套發展中自然語言人機介面的知識庫建構輔助工具, 包括一個具有螢幕編修並且提示選擇的語料分析輔助模組; 一個提示各個詞的使用範例以及構詞規則以幫助詞典建構的詞典建構輔助模組; 語意樣板分析輔助模組則用以蒐集各句子的語意樣板; 對話庫建構輔助模組則提供對話過程的資訊; 頻率統計輔助模組可用以找出各種類型的統計資料, 提供語音辨認與語言了解之利用。

爲了更方便的應用於不同的應用領域, 此輔助工具仍有不足之處。首先, 不同領域的提示選擇、語意分類都不同, 因此提示選擇最好能線上建立與修改; 其次, 本系統的斷詞與語意標記程序十分簡單, 未

來仍需加強，以提高其正確性，至於對話庫訊息的濃縮更是未來的發展方向。

致謝

本文係交通部電信總局電信研究所81年度計劃（計劃代號：81313）中自然語言處理所得成果之一。感謝呂所長學錦、王副所長金土、資訊室盧主任清松，對本文各項先導工作的支持。

參考文獻

- [1] Bobrow, D.G., Kaplan, R.M., Kay, M., Norman, D.A., Thompson, H., & Winograd, T., 1977, "GUS, A Frame-Driven Dialog System," *Artificial Intelligence*, Vol.8, pp.595-604, 1977.
- [2] 胡海旭編，李臻儀、胡祖櫻、許仁榮著，1990，"電話總機值機員和客人對話分析研究報告，第一卷：「對話模式分析」以及「客人表達轉接電話要求的句子分析」，1990a"，中華民國交通部電信總局電信研究所。
- [3] 胡海旭編，李臻儀、胡祖櫻、胡海旭、唐憲章、許仁榮語料謄寫及校對，1990，"電話總機值機員和客人對話分析研究報告，第二卷：「對話原始語料」第一冊至第五冊，1990b"，中華民國交通部電信總局電信研究所。
- [4] 蔡妙真、許仁榮、孫三為著，1992，"一個適用於自然語言語音查詢系統的詞彙表示法與服務類別架構研究報告"，中華民國交通部電信總局電信研究所。