應用句型結構與部份樣本樹於對話行爲之偵測

Dialogue Act Detection Using Sentence Structure and Partial Pattern Trees

梁維彬、蕭育丞、吳宗憲
Wei-Bin Liang, Yu-Cheng Hsiao, and Chung-Hsien Wu

國立成功大學資訊工程學系
Department of Computer Science and Information Engineering, National Cheng Kung University
E-mail: liang@csie.ncku.edu.tw ychsiao9@gmail.com chwu@csie.ncku.edu.tw

摘要

本論文提出一使用部份樣本樹及句型結構於對話行爲之偵測。爲了建構具強健性的對話行爲偵測模型，我們針對語音辨識之輸出語句，使用部份樣本樹來產生多重候選句，以避免語音辨識錯誤所衍生句子錯誤之問題。而後再經由剖析器得到候選句所對應之語法規則。而再針對每一類對話行爲所包含的規則做句型分類，來降低對話行爲之間的混淆。最後，利用潛在對話行爲矩陣來描述語法規則和意圖之間的關係。另外，在對話系統應用中，我們採用部份觀察馬可夫決策程序從對話歷程中訓練出之最佳對話策略，以增進對話系統的可用性。在實驗中，我們建立一個旅遊資訊諮詢對話系統，作爲實際應用測試平台。而在測試時，分別就每項對話行爲做測試。相較於應用語義表格(semantic slot)方法達到之 48.1%正確率，本論文所提之方法可得到整體正確率爲 81.9%，提升了 33.8%的正確率。由實驗可知論文所提之方法在實際應用上能有明顯的效能提升。

Abstract

This paper presents a dialogue act detection approach using sentence structures and partial pattern trees to generate candidate sentences (CSs). A syntactic parser is utilized to convert the CSs to sentence grammar rules (SRs). To avoid the confusion between dialogue intentions, the *K*-means algorithm is adopted to cluster the sentence structures of the same dialogue intention based on the SRs. Finally, the relationship between these SRs and the intentions is modeled by a latent dialogue act matrix. Moreover, for the application to a travel information dialogue system, optimal dialogue strategies are trained using the partially observable Markov decision process (POMDP) for robust dialogue management. In evaluation, compared to the semantic slot-based method which achieves 48.1% dialogue act detection accuracy, the proposed approach can achieve 81.9% accuracy, with 33.3% improvement.

關鍵詞：對話行爲、部份樣本樹、句型結構、部分馬可夫決策程序

Kewords: Dialogue act, partial pattern tree, sentence structure, POMDP

一、 緒論

在本論文中，我們以建構一套旅遊資訊查詢系統爲主要目標。因此，以下我們將對國內外的針對對話系統之相關研究做一文獻回顧。首先，針對旅遊資訊對話系統的相關研究[1]，在國外方面，有美國麻省理工學院(MIT)的餐廳導覽系統[2]、美國電信電報公司

(AT&T)的線上服務系統[3]、日本國家資訊通信科技研究機構(NICT)的旅遊導覽系統[4]，以及 Philips 公司所開發的火車時刻票價查詢系統[5]。在國內方面，台大有銀行電話查詢系統[6]、交大有汽車導覽系統[7]、工研院則有智慧型總機、氣象查詢系統的實現[8]，和成大智慧型醫療服務對話系統[9]。在意圖偵測部份的相關研究，Choi 等學者[10]將對話語料中每句話標記其意圖，及對整個對話過程標記其爲對話過程的開始、結束、正在對話等，再使用機器學習(machine learning)的方法來建立其模型以判斷意圖，但這一部分的方法對於語音辨識錯誤造成系統回應錯誤的問題並未考慮。而在對話管理部份，目前的研究有應用有限狀態機(finite state machine) [4]與部份觀察馬可夫決定程序(partial observation Markov decision process, POMDP)[11]來實現。這一部分的研究主題爲系統與使用者互動中，先判斷使用者的意圖，再對此意圖作出最適當的回應。

近幾年來，口述語言對話系統已經有顯著的進步，尤其是建構於填表式(slot-filling)的資料庫查詢方法已進入應用的階段。然而，因爲語音辨識錯誤造成表格填入錯誤，進一步使得自然語言理解產生錯誤及誤判使用者的意圖，導致系統回應錯誤。這類型的問題尚未成功地解決。因此，如何有效地在具語音辨識錯誤的條件下仍能得到好的意圖偵測結果是我們的研究主要目標。除此之外，我們亦希望能得到良好的人機互動，所以一個有效的系統回應機制以避免對話發散之窘境，讓使用者不致於對系統產生排斥甚至厭惡進而提高其可行性，也是我們所考量的部份。

在本論文的其他段落安排如下。第二節描述我們爲了本實驗所蒐集的旅遊相關資訊語料及其對話、語義類別、對話行爲和對話行爲所對應的行動標記。然後，第三節介紹論文核心的對話行爲偵測模型與其訓練方法將被描述。下一段的第四節爲對話管理決策的訓練。第五節的實驗說明了對話行爲偵測器中語音辨識器元件的訓練、數種對話行爲行爲偵測比較實驗和相關統計資料。最後，結論與未來展望將被討論於第六節。

二、 語料收集

2.1 語料錄製與標記

在實驗室環境下，使用 audio-technica AT9940 數位錄音麥克風，以 16 位元 16KHz 的取樣頻率將發音人的語料錄於單聲道。錄音情境爲，發音人面對電腦自行輸入譯文、操作錄音及辨識過程，錄音計劃負責人操作修改回應的部分。總共收錄到 144 個對話回合(dialogue turn)，總數爲 1,586 句的語料。錄音完成後，以人工方式進行對話語料相關資料標計(如 Dialogue 編號和 Turn 編號)、對話行爲(dialogue act, DA)。

2.2 語義類別(Semantic Class)

在以填表(slot-filling)方式爲基礎的對話系統中，若關鍵字詞過於散亂將間接導致系統效能不彰的問題，相較之下，若將關鍵字提升到語義類別，不僅可在對話行爲偵測上可避免偵測類別過多的問題，對於語料庫的擴增與維護也有較良好的管理。因此，在語料庫的資料分析過程中，我們將標記爲關鍵字的詞彙作進一步的整理，即如表 1 中所呈現的內容。最後，我們蒐集的語料庫總共包含 27 種語義類別。

2.3 對話行爲(Dialogue Act)和其系統回應行動(Action)

當語者說出了一句話，這句話本身的文字有其意義，但語者之所以會說出這句話有各種目的。這樣就可用對話來做一個行動，這就是對話行爲。舉例來說，「請問安平古堡的

票價是多少？」這句話被表達出來的 DA 爲詢問票價。因此，語言學家稱所有類型的溝通行爲爲"對話行爲"[12]。完成語料收集後，依據系統提供的任務(task)，我們分析對話語料並且在每一種系統任務中設計了該任務所含括的表單(slot)值及每個表單可填入值，如表 2 所示。在我們所收錄的語料中，可分爲三大任務，分別爲查詢系統服務、查詢景點相關資訊和查詢交通相關資訊，其中第 j 個任務所包含的 DA 數表示爲：

$$\text{Task}_j\text{的 DA 數} = \prod_{i=1}^{\text{Task}_j\text{包含的 slot 數}} \text{Slot}_i\text{可填入值數} - 1 \quad (1)$$

其中-1 是因爲未填值可能在其他任務中有填值，例如：任務 1 的“我想查詢高鐵”和任務 3 的 “跟我說高鐵的時刻表”。 其他的意圖包括歡迎、結束、無意圖，總共 38 個。當對話系統偵測出使用者的 DA 後，對話系統應作出合理系統回應行爲以達到和使用者之間的互動。因此，我們根據 DA 的內容整理出如圖 2 最右欄位所示的系統回應。大致可分爲系統詢問未填值的表格資訊和回答資訊的行動，總共有 20 種 [13]。

表 1：語義類別範例及其對應的關鍵詞彙範例。

編號	語義類別	詞彙	編號	語義類別	詞彙	編號	語義類別	詞彙
1	日期	日,月,星期,禮拜	10	感謝語	謝謝,感謝	19	開車	開車
2	城市	台北,台中,台南	11	疑問詞	什麼,怎麼	20	高鐵	高鐵
3	地點	安平古堡,成功大學	12	肯定	好,沒錯,了解	21	火車	火車
4	…	…	13	…	…	22	…	…

表 2：系統任務分類及其表單和可填入值之對照表範例。

Task	Category	Slot	可填入值				DA個數
1	查詢系統服務	系統服務	查詢地點	查詢車站	查詢服務	未填值	4-1=3
2	查詢景點相關資訊	地點表單	有地點	無地點			2*7-1=13
		地點資訊	地址	介紹	…	未填值	
2	查詢交通相關資訊	交通方式	公車	高鐵	火車	…	5*2*2-1=19
		目的地	有目的地	無目的地			
		出發地	有出發地	無出發地			
無	其他	歡迎					3
		…					
總計							38

表 3 ：DA 列表、例句和其對應的 Action。

Task	DA 編號	DA	Example	Action(編號)
1	1	查詢服務	你有什麼服務可以查詢？	回答系統能提供的服務(1)
	…	…	…	…
2	4	地點	安平古堡。	詢問使用者想查詢地點的何種資訊(4)
	5	查詢地址有地點	我想查詢安平古堡的地址。	回答地點地址資訊(5)
	…	…	…	…
3	17	公車有出發地有目的地	怎麼搭公車從成功大學前往赤崁樓？	回答搭乘公車方式(12)
	…	…	…	…
	35	有出發地無目的地	我從台北出發。	詢問使用者想要何種交通方式(11)
無	36	歡迎	您好。	系統無任何反應行動(18-1)
	37	結束	謝謝。	系統無任何反應行動(18-2)
	38	無意圖	無此種語料	系統無任何反應行動(18-3)

三、 系統架構

本論文的系統架構如圖 1 所示，虛線以上爲使用者部份，包括發音 U 和接收系統回應訊息所產生的聲音 U′；而虛線以下爲對話系統部分，主要分爲三個部份，包括輸入處理(input processing)、對話管理(dialogue management, DM)和輸出處理(output processing)。在輸入處理部分，語者的發音 U 經由麥克風傳送至自動語音辨識器(automatic speech recognizer, ASR)並產生辨識結果 W，此一辨識結果將送至口述語言理解(spoken language understanding, SLU)單元進行潛在對話行爲偵測而得到第 c 類對話行爲 DA_C，而 SLU 也是本論文的核心，我們將逐一介紹 SLU 的訓練方法和採用的技術。然後，DA_C 將傳送到對話管理並根據由 POMDP 所訓練而得的對話策略(strategy)和對話意圖歷史記錄(dialogue act history) DA_H 來採取合適的回應(action) a_t。當系統做出回應後，系統將從我們蒐集而來的旅有資訊資料庫(travel information database)中查詢對應的資料並輸出文字 $Context_t$ 至語音合成器(text-to-speech synthesizer, TTS)產生語音資訊 U′傳達給使用者。以下我們將逐一介紹各個部份。

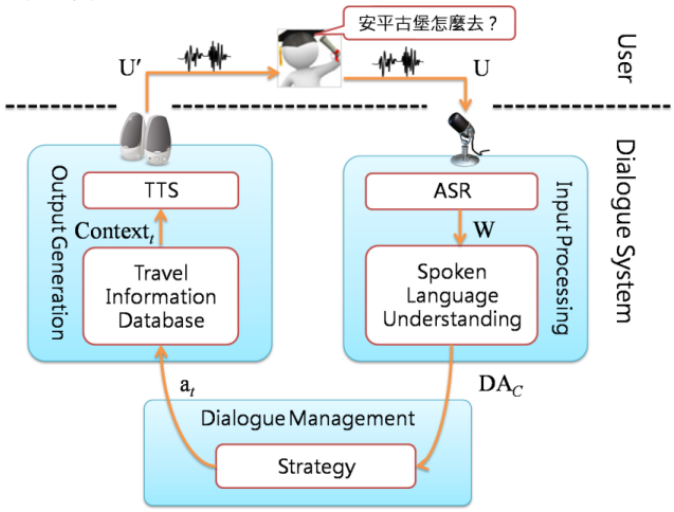

圖 1 ：對話系統架構圖。

3.1 口述語言理解(SLU)

在潛在對話行爲偵測過程中，系統必須根據使用者發音 U 和對話意圖歷史記錄 DA_H 來偵測最佳意圖 DA*，則此偵測法則(detection criterion)定義爲式子：

$$\mathrm{DA}^* = \underset{\mathbf{DA}}{\operatorname{argmax}}\, P(DA_C \mid \mathrm{U}, DA_H) \tag{2}$$

其中 **DA** 爲所有可能的 DA 集合，DA_C 爲發音 U 被辨識爲第 c 類 DA。第 i 種可能辨識字串 W_i 爲 U 經過辨識後所得到的可能辨識結果文字。然而，我們僅取最佳的辨識結果 $\hat{W}$，因此式子(2)改寫爲式子(3)，進一步展開爲式子(4)：

$$\begin{aligned}\mathrm{DA}^* &= \underset{\mathbf{DA}}{\operatorname{argmax}} \sum_{W_i} P(DA_C, W_i \mid \mathrm{U}, DA_H) \\ &\approx \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} P(DA_C, W_i \mid \mathrm{U}, DA_H) \qquad (3) \\ &= \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} P(DA_C \mid W_i, \mathrm{U}, DA_H) P(W_i \mid \mathrm{U}, DA_H) \qquad (4)\end{aligned}$$

假設辨識結果 W_i 與輸入語音 U 代表相同意義且語音辨識結果 W_i 與對話意圖歷史記錄

DA_H 獨立，則我們可以得到式子(5)。此外，P($DA_C|W_i$)爲經由貝氏決策法則(Bayes' decision rule)而來，因此我們將式子進一步改寫爲式子(6)：

$$\mathrm{DA}^* = \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} P(DA_C \mid W_i) P(DA_C \mid DA_H) P(W_i \mid \mathrm{U}) \tag{5}$$

$$= \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} \frac{P(W_i \mid DA_C) P(DA_C)}{P(W_i)} P(DA_C \mid DA_H) P(W_i \mid \mathrm{U}) \tag{6}$$

其中， $P(W_i)$可被省略，$P(DA_C)$在本論文假設爲均等事前機率(equal prior)，所以最後得到式子：

$$\mathrm{DA}^* \approx \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} P(W_i \mid \mathrm{U}) P(W_i \mid DA_C) P(DA_C \mid DA_H) \tag{7}$$

其中 $P(W_i|\mathrm{U})$爲自動語音辨識器從語音 U 所得到的辨識字串 W 的機率；$P(W_i|DA_C)$爲辨識字串被偵測爲第 c 個 DA 的偵測機率(probability of DA detection)，即對話行爲偵測的部份；$P(DA_C|DA_H)$爲對話意圖歷史(dialogue history)機率，用來防止系統跳脫使用者的意圖，例如使用者正在查詢高鐵時刻表，系統卻進入旅遊景點的相關資訊查詢功能。

3.2 對話行爲之偵測(Dialogue Act Detection)

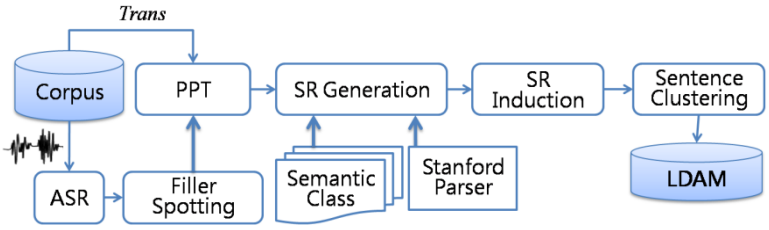

圖 2：對話行爲偵測的潛在語義矩陣模型訓練流程。

圖 2 展現了本論文 DA 偵測所使用的潛在意圖矩陣模型的訓練流程。無論文字類型的譯文檔或者譯文檔所對應的語音皆在訓練過程中被採用。爲了克服惱人的語音辨識錯誤問題，部份樣本樹用來產生多條候選句。在語句規則(sentence rule, SR)產生步驟裡，候選句會參考語義類別進行替換，然後使用 Stanford 剖析器[14]進行規則的產生。我們把語料庫中所有能產生的語句規則和所有的 DA 以矩陣模式建構兩者之間的關係，並進行歸納(induction)的動作。由於不同的 DA 之間可能擁有相同的語句規則，爲了避免 DA 之間的混淆，我們採用修改過的 K-means 演算法對同一類別 DA 之句型進行分類。我們將逐一詳細說明各流程步驟。

3.3 部份樣本樹(Partial Pattern Trees, PPT)

在本研究中，我們將對話語句視爲數個功能性詞彙(optional phrase, OP)與至少一個主要關鍵詞(main phrase, MP)之間的組合。因此我們從一句複雜的句子中擷取出部份樣本句(partial pattern, PP)，可看做類似對這句句子作部份分解，而 MP 則隱含著語句的語義，爲了保留語句的主要語義，因此每一句 PP 必需包含著 MP。相反的，OP 則可能在辨識結果中因刪除型錯誤(deletion error)而被省略。所以 PP 就是完整句子因爲某些 OP 在辨識結果中被刪除型的句子，而 PPT 是根據 PP 所建立的語句模組。

當在辨識對話語句時，一個常遭遇到的問題就是語句模組在建立時並沒有考慮到對話語

句常雜夾著一些無意義的贅語(例如：”嗯”和”喔”)和各種可能的語音辨識錯誤，這個問題進而造成 DA 偵測錯誤。在某些研究[15][16]中顯示出，一個詞彙出現在這些不流利的贅語之後的平均機率會小於出現在無贅語之後的情況。這指出我們很難去預測，當一個詞彙出現在多餘贅語之後的機率。因此根據這些觀察，我們將省略掉贅語或沒被辨識出來的詞彙而產生的句子，稱之爲 PP。更進一步探討，部份樣本樹一個很重要的應用，就是針對替代性錯誤做修正，而之前的研究對於錯誤的回復(recovery)通常都是利用句型規則在眾多的候選句中找出最符合語法句型的句子[17][18]。然而這些方法所產生的句子在句型上雖然非常符合，但是卻可能在語義上的意義是不足的。所以針對此點我們所定義的每一句部分樣本句，都需保留原句中的主要關鍵詞以維持句子的語義，然而功能性詞彙則有可能被省略。因此根據上述觀察，我們提出了利用 PP 來建立效能更佳的語句模組，在這裡我們將句子 $Trans_i$ 視爲一連串的 OP 與一 MP 的組合，表示成：

$$Trans_i = \{OP_1^i, OP_2^i, ..., OP_{NB_i}^i, MP^i, OP_{NB_i+1}^i, ..., OP_{NB_i+NA_i}^i\} \tag{8}$$

其中 NB_i 和 NA_i 分別爲在 MP 之前與 MP 之後的 OP 數。根據上述定義，PP 爲包含 MP^i 的子序列，其中每一個 OP 都有可能被省略，在本文中，被省略的 OP 我們替換成 Filler。替換成 Filler(F)的原因是我們想保持句子原本的句型，且也可以假設爲語音辨識錯誤時可能會發生的情況。舉例說明若有一句子爲”ABC”且 A，C 爲 OP 而 B 爲 MP 彙，則共有四句 PP 分別爲”ABC”，”ABF”，”FBC”和”FBF”。

3.4 填充字擷取(Filler Spotting)

即使 ASR 的辨識結果採用詞圖(word-graph)[19]爲基礎的重新計分(rescoring)方式，依然可能存在語音辨識錯誤的問題。這是因爲 word-graph 在重新計分時倘若辨識結果中夾雜著其他無助於 SLU 效能的文字，則除了從譯文檔產生 PP，我們也將譯文檔所對應的語音進行辨識以增加句型的多樣性(diversity)，彌補文字語料無法產生的句型，使得 SLU 單元能辨認更多種類的句型。所以我們對每個語音辨識結果的詞彙作填充字擷取。本論文採用統計方式的卡方檢定(χ^2-test)來進行填充字擷取。我們記錄被正確辨識詞彙所對應的分數，接著對每個詞彙計算其分數的平均值 (mean) μ 與標準差(standard deviation) σ。而填充字擷取的依據，我們使用卡方檢定，數學式爲：

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - \mu)^2}{\sigma} \tag{9}$$

藉此判斷語音辨識結果的每個詞彙要接受或拒絕，拒絕的詞彙我們替換成 Filler。替換成 Filler 的原因是我們想保持句子原本的句型，且也可以假設爲語音辨識錯誤時可能會發生的情況。

3.5 句型規則產生(Sentence Rule Generation)

在句型規則產生部分，首先，我們需要一個語義剖析器來處理訓練語句，並建立對應的語義樹狀結構，得到其句型規則，本研究利用史丹佛大學所研究開發的剖析器來達成此目的。史丹佛的剖析器[14]是基於 PCFG (Probabilistic Context Free Grammar) 的觀念所建立而成的剖析器。所謂的 PCFG 是一種隨機語言模型 (Stochastic Language Models, SLM)，而 SLM 的主要目的之一是根據訓練語料的統計資料來提供足夠的機率資訊以運用在語音辨識的構句處理上，不僅能有效提高的辨識正確率，更可藉由搜尋路徑的限制，節省計算時間，而應用在文句剖析上則能提供正確性較高的句法結果。關於史丹佛剖析

器主要的核心概念可以參考文獻[20][21]。圖 2 流程中的句型規則產生，在剖析前，我們先利用事先定義好的語義類別將語句中的詞彙替換成語義類別，再透過史丹佛剖析器得到剖析結果。替換的目的是降低句型規則的複雜度，讓相同語義的詞彙屬於同一條規則，例如：「NP → NN 安平古堡」和「NP → NN 億載金城」皆屬於「NP → NN 地點」這條規則。圖 3 範例是語句「怎麼去安平古堡」經過語義替換爲「疑問詞 路線 地點」，經由剖析器可得到一顆文法樹，其包含的句型規則包括：(1) Root → IP、(2) IP → VP、(3) VP → ADVP VP、(4) ADVP → AD 疑問詞、(5) VP → VV 路線 NP 和(6) NP → NN 地點。我們便以這六條規則代表這句話。

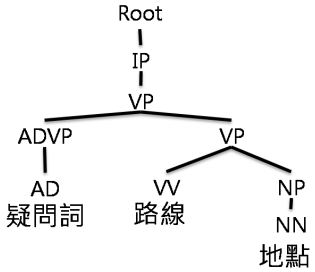

圖 3：剖析得到的文法樹範例。

3.6 句型規則歸納(Induction)

假設從所有語料中得到的句型規則可被表示爲維度爲 L 的規則向量 **Rule**，每個維度對應著一條句型規則。則此規則向量和所有的 **DA** 可構成一個矩陣來建立句型規則與意圖之間的關係，此關係可定義爲：。

$$\mathbf{\Phi}_{L\times Q}=\begin{bmatrix}\phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,Q}\\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,Q}\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{L,1} & \phi_{L,2} & \cdots & \phi_{L,Q}\end{bmatrix} \tag{10}$$

其中$\mathbf{\Phi}_{L\times Q}$是維度爲 $L\times Q$ 的文法結構資訊矩陣，L 代表訓練語料所有句型規則的個數，Q 代表意圖的總數。矩陣中每個元素 $\phi_{l,q}$ 代表著第 l 條文法規則 $Rule_l$ 在第 q 個 DA 中所佔的重要性。因此本研究中定義 $\phi_{l,q}$ 的估計法如下：

$$\phi_{l,q}=(1-\varepsilon_l)P(Rule_l \mid DA_q) \tag{11}$$

其中，$P(Rule_l \mid DA_q)$是該條規則佔該句語法結構的比重，該項可以寫爲：

$$P(Rule_l \mid DA_q)=\frac{C(Rule_l, DA_q)}{\sum_k C(Rule_k, DA_q)} \tag{12}$$

且 C($Rule_l$, DA_q)表示句型規則 $Rule_l$ 出現在 DA_q 中的次數。另外，(1-ε_l) 是利用量度文字亂度 (Entropy) 的方法來量度某條規則在該語料中是否具有鑑別性並賦予該元素的權重，則ε_l 可定義爲：

$$\varepsilon_l=-\frac{1}{\log Q}\sum_{q=1}^{Q}\frac{C(Rule_l, DA_q)}{\sum_{i=1}^{Q} C(Rule_l, DA_i)}\log\frac{C(Rule_l, DA_q)}{\sum_{i=1}^{Q} C(Rule_l, DA_i)} \tag{13}$$

3.7 語句分群(Sentence Clustering)

使用者的 DA 可能以不同語句表達，不同語句意味著他們可能蘊含著不同的句型規則，因此，同一個 DA 下可能包含著多種句型規則導致和其他意圖之間造成混淆。例如：兩段屬於同樣 DA_1 的語音分別包含語句規則{1,2,3}和{4,5,6}，另有一段屬於 DA_2 的語音包含語音規則{3,4,7}，則此語音可能會被誤判爲 DA_1。因此，爲了避免意圖之間的混淆，語句需要分群。對於傳統的分群方法，如 *K*-means 演算法，必須計算各資料點和 centroid 之間的距離。然而，句型規則的 centroid 不具有任何物理意義。因此爲適應本論文的需求，我們先將屬於第 q 個 *DA* 的第 i 個譯文檔 $Trans_i$ 表示爲：

$$\Phi_q^{Trans_i} = (\phi_{1,q}\delta_1, \phi_{2,q}\delta_2, \ldots, \phi_{L,q}\delta_L) \tag{14}$$

其中 $\phi_{l,q}$ 的定義等同於上述句型規則歸納矩陣 $\mathbf{\Phi}_{L\times Q}$ 中的 $\phi_{l,q}$ 。而 δ_l 指出若 $Trans_i$ 使用到 $Rule_l$，則其值爲 1，反之爲 0。另外，我們選擇一個特殊函式作爲最大化群之內相似度，此特殊函式表示爲：

$$(G_1, G_2, \ldots, G_K)^* = \arg\max \sum_{k=1}^{K} \sqrt{\sum_{\Phi_q^{Trans_i}, \Phi_q^{Trans_j} \in G_k} Similarity(\Phi_q^{Trans_i}, \Phi_q^{Trans_j})} \tag{15}$$

其中 K 爲欲分群之數量，G_k 表示屬於第 k 群的譯文檔集合，而 *Similarity*(•)爲兩則譯文檔之間的相似度計算，我們採用 Cosine Measure 數學式表示爲：

$$Similarity(\mathbf{\Phi}_q^{Trans_i}, \mathbf{\Phi}_q^{Trans_j}) = \frac{\mathbf{\Phi}_q^{Trans_i} \cdot \mathbf{\Phi}_q^{Trans_j}}{\left\|\mathbf{\Phi}_q^{Trans_i}\right\| \cdot \left\|\mathbf{\Phi}_q^{Trans_j}\right\|} \tag{16}$$

3.8 潛在對話行爲矩陣模型(Latent DA Model, LDAM)

經由部份樣本句與填充字擷取後，我們將得到的句型規則與意圖句型分類後的類別建立關係矩陣，而產生的句型規則比原本訓練的文字語料所產生的句型規則包含更多意涵，所以我們稱此矩陣爲潛在意圖矩陣且定義爲：

$$\mathbf{LDAM}_{L\times M} = \begin{bmatrix} \nu_{1,1} & \nu_{1,2} & \cdots & \nu_{1,M} \\ \nu_{2,1} & \nu_{2,2} & \cdots & \nu_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{L,1} & \nu_{L,2} & \cdots & \nu_{L,M} \end{bmatrix} \tag{17}$$

其中意圖矩陣 $\mathbf{LDAM}_{L\times M}$ 爲一個維度爲 $L\times M$ 的文法結構資訊矩陣，L 代表訓練語料經過語句規則產生步驟後所有的句型規則個數，M 代表意圖句型群聚後的類別總數。矩陣中每個元素 ν_{lm} 代表著第 l 條句型規則在第 m 個 DA 中所佔的重要性。因此本研究中定義 ν_{lm} 的估計法如下：

$$\nu_{lm} = (1-\varepsilon_l) P(Rule_l \mid DA_m) \tag{18}$$

其中 $P(Rule_l|DA_m)$是該條規則佔該句語法結構的比重，該項可以寫爲：

$$P(Rule_l \mid DA_m) = \frac{C(Rule_l, DA_m)}{\sum_k C(Rule_k, DA_m)} \tag{19}$$

而 C($Rule_l$, DA_m)表示句型規則 l 出現在第 m 個意圖句型分類後的類別中的次數。另外，(1-ε_l)是利用量度文字亂度 (Entropy) 的方法來度量該條規則在語料中的鑑別性，當作矩陣中該元素的權重，ε_l 可定義爲：

$$\varepsilon_l = -\frac{1}{\log C}\sum_{c=1}^{C}\frac{C(Rule_l, DA_m)}{\sum_{i=1}^{C} C(Rule_l, DA_i)}\log\frac{C(Rule_l, DA_m)}{\sum_{i=1}^{C} C(Rule_l, DA_i)} \tag{20}$$

其中大寫 C 爲由上述 K-means 分群而得最終 DA 數量，而 DA_C 所表示的就是 **LDAM** 第 c 個 column 的內容。最後，我們得到 SLU 偵測語義所需要的模型。

3.9 對話行爲型態偵測

在對話行爲 DA 偵測中， $P(W_i \mid DA_C)$項因爲一個句子包含語義成分及語法成分，所以我們對語音辨識的結果進一步拆解爲：

$$P(W \mid DA_c) \approx P(\mathbf{Rule}_W \mid DA_c)P(SC_W \mid DA_c) \tag{21}$$

其中 $\mathbf{Rule}_W$ 爲記錄辨識結果 W 所使用的句型規則。則 $\mathbf{Rule}_W$ 對潛在意圖矩陣中第 c 個類別的相似度，我們採用 Cosine Measure 來計算，定義爲：

$$P(\mathbf{Rule}_W) \mid DA_c) = \frac{\mathbf{Rule}_W^T \bullet DA_c}{\|\mathbf{Rule}_W\| \times \|DA_c\|} \tag{22}$$

而 SC_W 爲將辨識結果 W 轉換爲語義類別的函式。在語義成分分數的計算，我們經由統計文字語料，得到每個對話行爲 DA 出現每個意圖類別的機率，數學式可寫爲：

$$P(\mathrm{SC}_W \mid DA_c) = \prod_{w_n \in W} P(\mathrm{SC}_{\mathrm{j}}^{w_n}) \tag{23}$$

其中 $P(\mathrm{SC}_{\mathrm{j}}^{w_n})$ 表示辨識字串 W 的第 n 個字 w_n 屬於第 j 個語義的機率。這可以從語料庫中離線預先估測而得。

3.10 對話行爲歷史記錄

對話行爲歷史記錄方面的目的是爲避免使用者在查詢其中一項任務時，在尚未完成任務卻因爲 ASR 錯誤造成對話行爲誤判而轉爲詢問使用者其他任務的內容。假設 DA_t 定義爲目前語音所得到的 DA，即 DA_C，而過去歷史紀錄 DA_t^{t-1} 定義爲 DA_H，倘若我們假設對話行爲只與前一個對話行爲有關，則數學表示法可定義爲：

$$P(DA_t = DA_C \mid DA_1^{t-1} = DA_H) = P(DA_t \mid DA_1, DA_2, \cdots, DA_{t-1}) = P(DA_t \mid DA_{t-1}) \tag{24}$$

四、 對話管理決策

對話管理是基於對話行爲偵測結果而採取適當回應以與使用者進行互動，而適當回應仰賴於對話策略的規劃。在本研究中，我們採用 POMDP 作爲對話策略規劃的工具。POMDP 將系統所處的狀態視爲隱含變數(hidden variable)，因此必須使用一個相信函數(belief function)來假設系統所處的狀態並定義了五個變數值組(tuples) {**S**,**A**,**R**,**T**,**O**}，分別代表狀態組成的集合 **S** 並用一個相信函數(belief function) b 來控制(maintain)，在本研究中，定義爲前文所提的 DA；回應使用者的方式所組成的集合 **A**；獎勵函數 $R(s,a)=r$，表示在狀態 s 採取回應使用的方式 a，系統所得到的獎勵爲 r；轉移機率 $T(s,a,s')=P(s_{t+1}=s'|s_t,a_t)$爲系統在時間點 t，在狀態 s 採取使用的方式 a，而在時間點 t+1，狀態將會變成 s'的機率；觀察(observation)所組成的集合 **O**，描述著 POMDP 能接收的訊息而觀察機率 $P(o'|s',a)$表示系統在時間點 t 採取使用的方式 a，及在時間點 t+1 系統所處的狀態 s'，所觀察到的觀察的機率。綜合以上變數，POMDP 應用於本研究的概念圖如圖 4 所示。

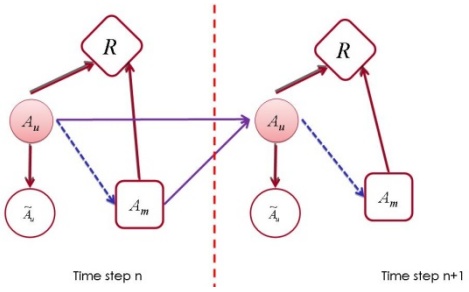

圖 4：馬可夫決策程序概念圖。

在相信函數的更新部份，我們引用文獻[11]所推導的公式。

$$b'(s') = P(s'|o',a,b) = k \cdot P(o'|s',a)\sum_{s\in S} P(s'|s,a)b(s) \tag{25}$$

其中 $b'(s')$ 爲更新的相信函數，$P(o'|s',a)$ 爲觀察機率，$P(s'|s,a)$ 爲轉移機率，$b(s)$ 爲相信函數。部分觀察馬可夫決策程序的最佳值函數(Optimal value function)爲：

$$V^*(b) = \max_{a\in A}[\sum_{s\in S} r(s,a)b(s) + \gamma \sum_{o',s,s'} p(o'|s',a)p(s'|s,a)b(s)V(b'(s'))] \tag{26}$$

4.2 對話策略學習

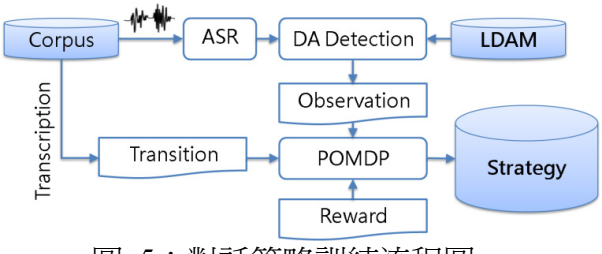

圖 5：對話策略訓練流程圖。

在訓練對話策略之前，我們必須定義系統的狀態、觀察及獎勵函式。在本論文中，對話策略學習的狀態和狀態所對應的回應行爲爲第二節所提到的 slot 和 action 的部份。而對話策略訓練流程圖如圖 5 所示。首先，從我們收錄的語料中，由文字的每個對話(dialogue)得到轉移(transition)機率，及從文字相對應的音檔經由對話行爲偵測後得到觀

察機率，及定義獎勵函式爲：

$$r=\begin{cases}+10 & ,if\ \text{系統採取正確回應}\\ -10 & ,if\ \text{系統採取錯誤回應}\\ -5 & ,if\ \text{重複詢問問題}\\ -100 & ,if\ \text{系統採取正確回應歡迎發生在開始之外}\\ +100 & ,if\ \text{系統結束}\end{cases} \tag{27}$$

在觀察機率，假設觀察爲對話行爲型態偵測後的假設結果，可表示爲：

$$P(o'\mid s',a)\equiv P(DA_{o'}\mid DA_{s'},a) \tag{28}$$

假設觀察與上一次系統回應無關，所以觀察機率爲：

$$P(DA_{o'}\mid DA_{s'},a)\cong P(DA_{o'}\mid DA_{s'})=\begin{cases}P(DA_C\mid S,DA_H)(1-p_{err_C})\\ (1-P(DA_C\mid \mathrm{U},DA_H))\cdot\dfrac{p_{err_C}}{|\mathrm{DA}_u|-1}\end{cases} \tag{29}$$

其中 P_{errc} 透過訓練語料求得每個對話行爲型態的正確率。此觀察機率包含著對話行爲型態偵測分數 $P(DA_C|\mathrm{U}, DA_H)$及此對話行爲型態偵測結果的可信度(1-P_{errc})。最後我們經由 POMDP 軟體[22]訓練我們的對話策略。

五、 實驗

5.1 自動語音辨識器(ASR)的建立

ASR 的基本建立步驟，包含了聲學特徵參數的萃取(feature extraction)、聲學模型訓練(acoustic model, AM)和語言模型(language model)的訓練。我們採用劍橋大學所開發的工具 HMM Tool Kit(HTK)來建立本研究的 ASR。爲了建立一個較爲可靠的 ASR，我們先採用麥克風語料庫 TCC300 進行種子(seed) AM 訓練，包括 115 個右相關(right-context dependent) initial 次音節和 38 個獨立(right-context independent) final 次音節，分別使用 3 個和 5 個狀態(state)，每個狀態最多 32 個 mixtures。，再以我們所錄製而來的旅遊相關語料進行調適(adaptation)。特徵參數爲 39 爲度的梅爾倒頻譜特徵參數(MFCC)，其中預強調係數爲 0.97，其餘相關參數設定可參考 HTK Book 說明。在語言模型部分，我們採用 TCC300 語料庫來建立語言模型的部分，並嘗試進行調適。然而，經由實驗發現，語言模型在本系統似乎作用性不高，這是因爲我們所收錄的旅遊相關語料內容對於 TCC300 而言屬於微量，因此許多旅遊景點相關字詞無論在 uni-gram 和 bi-gram 都只是極小值。就語音辨識器而論，我們所建立的語音辨識器對於所蒐集的旅遊相關語料庫有高達 84.33%的正確率。經語者調適後，更可達 93.12%的正確率。

5.2 對話語料之分析

將收集而來的語料，經過整理挑選出適合的語料作爲訓練語料之用，語料總共有 144 個對話回合，總數爲 1586 句，爲了了解對話用句分佈情形，本論文針對語料做了兩種分析，分別爲對話之意圖分佈圖和對話的長度分析。在表 3 中，我們定義了 38 種 DA 而圖 6(a)呈現出語料裡各種 DA 分佈的情形。由分佈圖可知，使用者會根據他的需求查詢他所想要的資訊，所以每個意圖出現頻率不同，而「結束」出現頻率遠高於其他意圖說明了在對話結束時使用者習慣性說告別用語。說明了語料裡對話回合次數分佈的情形，

所謂對話回合次數即是一次對話需要來回多少次(turns)。就如圖 6(b)所示，每一次對話的句數大都分布在 3~15 句之間。而 3~5 句代表著使用者只使用一項任務便結束系統，其他則表示使用者可能同時查詢了好幾項資訊才結束系統。

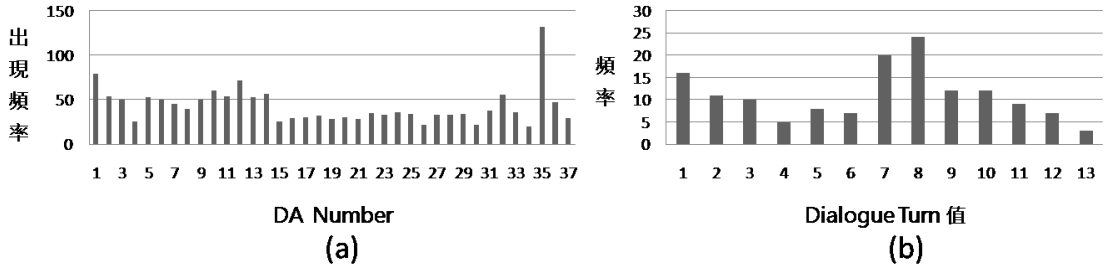

圖 6：對話意圖分佈和對話回合次數分佈。

5.3 系統評估分析

系統評估方面，我們把評估方式分爲二種，第一種是對各類 DA 做評估，觀察 DA 偵測模組對於每類 DA 的運作效能。第二種是對話回合(turn)次數的評估，此評估目的在於了解 POMDP 對於整個對話流程其影響結果。評估語句總數爲 912 句。評估 DA 偵測模組時，分別考慮(1)僅使用語義表格(semantic slot, SS)，依據使用者填入的 semantic slot 判斷使用者的意圖，(2)使用 SS 和史丹佛剖析器(Stanford Parser, SP)建立潛在意圖矩陣 LDAM 的偵測方式，(3)使用 SS、SP 和部分樣本樹(PPT)建立 LDAM 的偵測方式和(4)使用 SS、SP、PPT 和句型分類(sentence clustering, SC)建立 LDAM 的偵測方式四種對話行爲模組的評估方法。各方法的對話行爲偵測準確率如圖 7，平均正確率分別爲 49.6%、76.2%、81.6%和 82.9%。圖 7 標示爲 DA 的欄位爲 37 種 DA，缺少「無意圖」DA 是因爲無法收集無意圖 DA 的句子，故無法評估其準確率，無意圖 DA 是用來當系統無法判斷使用者的意圖時所做出的回應。評估方法(1)的結果，雖然在某些 DA 上能有可接受的表現，但我們可以發現屬於任務 3 的表現出現落差，因爲 DA 中有幾種彼此會互相混淆，造成很多意圖的準確率是 0%，例如詢問交通方式的 DA。使用評估方法(2)的結果，相較於(1)有明顯的改善，但依然無法解決因爲語音辨識錯誤而造成 DA 偵測的問題。使用評估方法(3)的結果，在測試時，語句的辨識結果必須經過 PPT 產生樣本候選據來進行意圖偵測，我們可以看到標示爲「無出發地有目的地」的 DA 有非常顯著的改善，這類的句子如「怎麼到安平古堡」容易在評估方法(1)和(2)被判別爲其他類似的對話行爲型態，如「火車無出發地有目的地」DA。最後，評估方法(4)，這個組合方式即爲論文中 LDAM 模型的訓練方式，經實驗證明，在所有的方法中，我們所建構的 LDAM 模型是可行的。另外，查詢地點(2)、查詢車站(3)和系統歡迎(36)因爲問題本身簡單，以致於四種評估方法皆有相同的效能。在查詢交通方式的火車和高鐵部分，因爲本身的關鍵字詞重疊性太高，以致於沒有明顯的改善表 4 爲評估由 POMDP 訓練而得的對話策略模式對於我們所蒐集語料的效用，我們可以發現 POMDP 的確能降低對話回合次數，但由於收錄的語料辨識結果還算正確且對話行爲型態判斷大部份也是正確，所以降低的對話回合次數沒有顯著的提升。在另一方面，本來對話行爲型態偵測錯誤的句子，若是用人工訂定的回應可能會產生奇怪的回應，例如使用者詢問地址而系統卻回答票價，若使用 POMDP 則可能產生較爲正確的回應，例如上述例子中，系統回應將變爲詢問使用者意圖。雖無減少對話回合次數但卻使得使用者覺得系統回應更爲人性化，所以 POMDP 確實有它的效能。

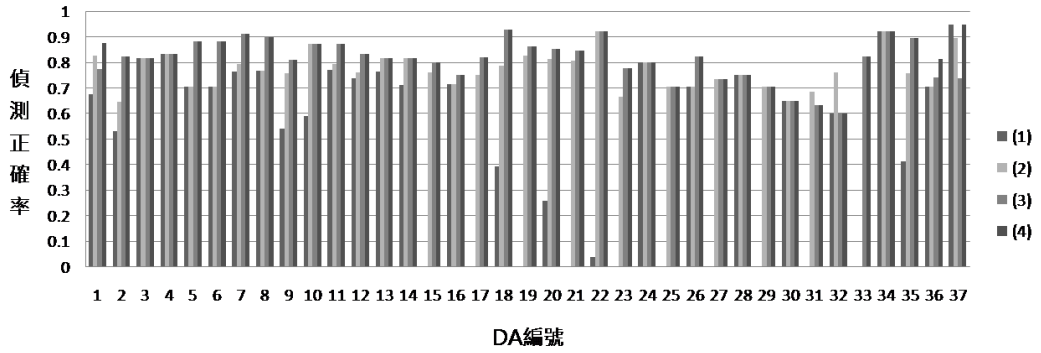

圖 7：各方法對話行爲型態準確率比較：(1)僅使用 SS 的偵測，(2)使用 SS 和 SP 偵測，(3)使用 SS、SP 和 PPT 的偵測，(4)使用 SS、SP、PPT 和 SC 的偵測。

表 4 ：對話回合次數之比較。使用 POMDP 訓練而得的策略降低了對話回合數。

	without POMDP	with POMDP
Average #(Turns)	8.6	8.4

六、結論與未來展望

本論文提出了利用部份樣本句與句型規則建構出潛在意圖矩陣，藉此判斷對話行爲型態，並有效改善因語音辨識錯誤造成對話行爲型態錯誤的問題。此外本論文也加入對話歷史的概念，考慮對話語義脈絡來幫助對話理解。爲了增加人機之間的互動，在系統決策管理方面也運用 POMDP 以求取最佳策略，使得系統產生最佳回應，減少系統與使用者之間無法完成任務的情況。透過實驗的證明，本論文所提出的方法在潛在對話行爲偵測上平均準確率爲 81.9%，相較於單純使用表單填格方式的意圖偵測平均準確率 48.1%，使用本方法可提升了 33.8%之準確率，本論文所提出的方法的確是有效的方法。雖然本論文所提出的方法可達到不錯的成效，但仍有不少地方有待改進，以下我們將逐一說明可改進的地方：(1)如何自動找尋應用領域的對話行爲型態以減少人爲介入，爲理解系統另一個研究議題。(2)在本論文中所使用的 POMDP，可以進一步將狀態假設爲許多不同值域的集合，使系統能更細緻地與使用者互動。(3)可以定義獎勵函數，根據不同的填值狀況作不同的獎懲。

參考文獻

[1] X.-D Huang, Alex Acero, H.-Wd Hon, “Spoken Language Processing”, Prentice-Halln, Inc. 2001

[2] Ji.-J. Liu, Y.-S. Xu, S. Seneff, and Victor Zue, “Citybrowser Ⅱ: A multimodal restaurant guide in Mandarin”, in Proc. International Chinese Spoken Language Processing, 2008.

[3] AT&T(2002) How May I Help You? [Online]. Available: http://www.research.att.com/~algot/hmihy/

[4] C. Hori, K. Ohtake, T. Misu, H. Kashioka, S. Nakamura, “Dialog Management using Weighted Finite-State Transducers”, Interspeech, 2008

[5] S. Bennacef, L. Devillers, S. Rosset, and L. Lamel, “Dialogue in the RAILTEL Telephone-Based System”, in Proc. of ICSKP’96, vol. 1, pp. 550-553, 1996

[6] C.-J. Lee, E.-F. Huang, and J.-K. Chen, “A Multi-keyword Spotter for the Application of the TL

Phone Directory Assistant Service", in Proc. Workshop on Distributed System Technologies & Applications, pp. 197-202, 1997

[7] 蔡金翰, "語音對話系統和對話策略之研究," 國立交通大學電信工程學系碩士論文, 2005

[8] T.-H. Chiang, C.-M. Peng, Y.-C. Lin, H.-M. Wang and S.-C. Chieh, "The Design of a Mandarin Chinese Spoken Dialogue System", in Proc. COTEC'98, Taipei 1998, pp.E2-5.1~E2-5.7

[9] 陳銘軍, 葉瑞峰, 吳宗憲, "以知識概念模型爲基礎之多主題對話管理系統", in Proc. ROCLING XV, Hsinchu, Taiwan, 2003.

[10] W.-S. Choi, H. Kim, and J.-Y. Seo, "An Integrated Dialogue Analysis Model for Determining Speech Acts and Discourse Structures," the Institute of Electronics, Information and Communication Engineers (IEICE), 2005

[11] J. D. Williams, and Steve Young, "Partially Observable Markov Decision Processes for Spoken Dialog Systems," Computer Speech and Language, 2007.

[12] David R. Traum, "Speech Act for Dialogue Agents," Kluwer Academic Publishers, 1999.

[13] Y.-C. Xiao, "MHMC Annotation of MHMC Travel Corpus," 2009. [Online]. Avaliable: http://chinese.csie.ncku.edu.tw/~liang/MHMC_Annotation_of_Travel_Corpus.pdf

[14] Stanford Parser [Online]. Available: http://nlp.stanford.edu/software/lex-parser.shtml

[15] E. Shriberg and A. Stolcke, "Word Predictability After Hesitations: A Corpus-Based Study", in Proc. International on Conference Spoken Language Processing (ICSLP), pp. 1868-1871, 1996.

[16] M Siu, M. Ostendorf, and H. Gish, "Modeling Disfluencies in Conversational Speech" , in Proc. International on Conference Spoken Language Processing (ICSLP), vol 1, pp. 386-389, 1996.

[17] T.R. Niesler and P.C. Woodland, "Variable-Length Category N-gram Language Models", Computer, Speech and Language, vol. 21, pp. 1-26, 1999.

[18] J. S. Hamaker, "Towards Building a Better Language Model for Switchboard: the POS Tagging Task," in Proc. International Conference on Acoustics, Speech, and Signal Processing(ICASSP), pp. 579-582, 1999.

[19] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 3, pp. 288-298, 2001

[20] Dan Klein, and C. D. Manning, "Fast Exact Inference with a Factored Model for Natural Language Parsing," in Advances in Neural Information Processing Systems, 2003.

[21] Dan Klein, and C. D. Manning, "Accurate Unlexicalized Parsing," in Proc. the 41st Meeting of the Association for Computational Linguistics, pp. 423-430, 2003.

[22] POMDP 軟體 [Online]. Available: http://staff.science.uva.nl/~mtjspaan/software/approx/