動態詞典及其與馬可夫中文語言模型之整合

Dynamic Dictionary and its Integration with Markovian Chinese Language Models

古鴻炎
Hung-yan Gu

國立台灣科技大學電機系
Department of Electrical Engineering
National Taiwan University of Science and Technology, Taipei
E-mail: root@guhy.ee.ntust.edu.tw

摘要

在中文輸入有關的應用裡，都需用到一個中文語言模型以將輸入資料轉換成中文。本文就針對中文語言模型會碰到的新詞與慣用語彙的處理問題，提出以動態詞典機制來解決，並且對動態詞典應該採用的結構作了探討；然後，我們進一步提出一種把動態詞典機制整合到馬可夫中文語言模型的方法，以使整合後的模型具有原先馬可夫語言模型的快速、逐音節漸近處理的能力，同時還具有非監督式記錄與調適的能力。對於這個整合模型，我們已將它製作成軟體程式，然後以測試語料去測試、驗證，結果顯示整合模型的確會依照預期方式運轉，並且正確轉換率比原先馬可夫模型的高而且穩定。

關鍵詞：動態詞典，馬可夫模型，中文語言模型

Abstract

In many applications for Chinese-character inputting, a processing component of Chinese language model is needed in order to convert the input into its corresponding Chinese sentence. In this paper, the problems faced by a Chinese language model, i.e., fresh word and word-usage habit, are considered and a mechanism of dynamic dictionary is proposed. In addition, a method is proposed to integrate the mechanism of dynamic dictionary into a Markovian Chinese language model. This integration is intended to enable unsupervised recording and adaptation while the fast and incremental processing capability of the original Markovian language models are kept. For the integrated model, we have implemented it as a software module and then conducted several tests with it. The test results show that the integrated model can operate in the predicted way, and the obtained conversion rates are stable and better than the rates from the original Markovian language models.

Keywords: dynamic dictionary, Markov model, Chinese language model

1、前言

在許多應用裡，如：國語語音辨識、線上手寫中文字辨識、光學中文字辨識、及以鍵盤輸入中文等，都可使用中文語言模型來提高辨識的正確率，而使輸入速度加快。過去被提出的語言模型可概分爲：查詞典式[1]、文法分析式[2,3]、及統計式[4,5,6,7,8]等作法，依據過去我們對馬可夫(Markovian)中文語言模型(一種統計式的語言模型)所做的研究[5,9,10]得知，查詞典式作法其實可看作是馬可夫語言模型的一個特例，另外，馬可夫語言模型與文法分析式作法比起來，其優點是模型之訓練有一套公式化的程序，且一定可找出一個最大機率的文句作爲輸出，不必操心不合規則的語句及規則擷取的問題，但是它也有自己的問題，其中一個很重要的是，在訓練階段所用的訓練語料中未出現過的語彙(如"戴歐辛"、"鮮酪乳")，如果所用的靜態詞典(原先馬可夫模型裡用以查循輸入資料可能對應的多字詞的詞典)裡也沒有收錄，那麼這個語詞就很難被正確轉換出來，對於這樣的問題，過去我們曾提出了一種解決方法[10]，不過，這裡我們提出另一種對使用者更方便的非監督式(unsupervised)的機制來解決，稱爲動態詞典機制，並且我們也研究了它與馬可夫中文語言模型作整合的方法，以使整合後的模型具有原先馬可夫語言模型的快速、逐音節漸近處理的能力，並具有非監督式記錄與自動調適(adaptation)的能力。所謂"非監督式"是指使用者不需特別下指令，叫語言模型去將某一個片語當作新詞而放入詞典裡；"自動調適"是指語言模型會在使用者輸入文句時作線上漸進的調整，以改善原先的馬可夫語言模型會對不同的文章類型而有明顯的轉換率差異的問題。

關於訓練語料中未出現過且靜態詞典裡未收錄的語詞，使得馬可夫語言模型轉換錯誤的問題，我們過去提出的一種解決方法是，建造一種複合式的馬可夫語言模型，以支援使用者主動作線上新詞登錄的動作，"複合式"模型指的是，將"字"爲狀態的一階馬可夫語言模型與"詞"爲狀態的零階馬可夫語言模型結合運用，不作此種結合就很難同時保有較高的正確轉換率，及能夠支援主動式的新詞登錄的動作[10]。由

於此種方法需要使用者主動下命令作新詞登錄的動作，所以至少有三個缺點：(a)對使用者來說不方便且費力；(b)使用者如何決定某個片語要當作是詞而放入詞典裡；(c)一直作登錄的動作時，詞典占據的記憶空間會一直成長下去。爲了一舉解決前述的問題，所以本文研究了動態詞典機制。

雖然，動態詞典的觀念很早就被應用於作資料壓縮[11,12]，即詞典式編碼方法，但是本文將此觀念加以擴充及具體化，而使它能夠被應用於中文語言模型裡。另外，動態詞典的觀念和快取(cache)的觀念雖然有一些類似的地方，且過去已有研究者將快取機制與馬可夫英文語言模型作結合[7]，以提高對同一詞類(如名詞)中成員詞的預估出現機率的準確性，但是，本文的動態詞典機制，在結構上與功能上都和他們的快取機制不相同。就結構上說，他們的快取機制是給每一種詞類一個 cache (即相同詞類的單詞要放在同一個 cache)，並且每個 cache 的大小不能太大(如設爲 200)，然而本文的動態詞典是共用的，並且詞典的大小是愈大愈好。就功能上來說，他們的快取機制是要用來掌握英文單詞的動態使用情形，因爲在一篇文章裡，通常跟主題有關的術語、用語會突然變得經常出現，可是馬可夫語言模型裡的參數，是由百萬字級的訓練語料中計算出來的，並不能完全反應欲輸入之文章的本地(local)特性，所以需要 cache 且 cache 不能太大，以便掌握短期的、動態的使用情形；至於本文的動態詞典機制，主要賦予的功能是記錄相鄰中文字的共出現(co-ourrence)關係，記錄專有名詞，記錄使用者的習慣用語，所以詞典要愈大愈好，以便記錄得愈多愈詳盡，而當發生詞典中有同音詞時或文章主題更換時，則以最晚進入詞典的優先選用，這樣也可達到掌握動態使用情況的目標。

由於我們目前只想尋找一種製作上簡單(考慮所花的處理時間與記憶空間的 overhead)卻又實用的機制，所以就決定研究動態詞典機制及它與常用的馬可夫中文語言模型的整合的問題。除了動態詞典的觀念之外，事實上還存在有其它的觀念可供借鏡，如 blending of finite-context models [11]。

2 、動態詞典的功用及其結構設計

2.1 動態詞典的功用

本文所以要將動態詞典的觀念應用到中文語言模型裡，其動機是來自於，比較中文與英文時至少可觀察到下列三點差異：

(a) 中文裡不同的名詞都有其慣用的量詞，如：一"匹"馬、一"隻"狗、一"支"雨傘、...等，那麼要如何將量詞與名詞之間的共發生(co-occurrence)關係，自動地記錄下來。

(b) 中文句子裡的詞序限制比英文寬鬆多了，如："飯我吃了"、"我飯吃了"、"我吃飯了"，在句法(syntax)上都是對的，所以要使用句法規則來篩選出合法的語句，不會像英文那麼有效。前面所說的句法規則包括以詞類爲狀態之馬可夫模型，即以前一或二個詞類來預測下一個詞類的出現機率。

(c) 一個中文字唸一個音節，且國語裡不同的音節個數(含聲調)只有 1300 個左右，不像英語裡不同音節的個數超過 10,000 個以上，使得以注音或語音來輸入中文時，發生了嚴重的同音字、同音詞(如：食物、時務、實物)的問題，這樣的問題需要較多的前後文形成的片語來幫忙分辨，可是一般常用的馬可夫中文語言模型是使用詞爲狀態、一階(由前一詞預測下一詞)的模型，前後文的資訊僅限於詞與詞之間的關聯性，而且中文句子裡單字詞又經常出現，不像英文裡的詞大多是多音節詞，所以中文的馬可夫模型的分辨能力會較差。那麼，如何建造一種機制來將較長的片語資訊考慮進來！

依據前面的觀察，這裡我們就來檢視動態詞典的觀念是否能支援所需要的功能，並且探討動態詞典應該用什麼樣的結構去建造，我們先從簡單的結構來看，最簡單的動態詞典結構就是貯列(queue, first-in-first-out)的結構[13]，對於使用者輸入的、已經過同音錯字更正後的中文句子(不含音標資料)，按次序從貯列的一端放入，當記憶空間不足時，就從另一端將最早放進的句子拿出。

假設下列的文句已被輸入且放入貯列結構裡，接著當輸入

"除了一支花雨傘，一枝花之外，他還帶著一隻貓。"

/ㄓ ㄏㄨㄚ/兩音節時，則語言模型應該要輸出"枝花"，因爲，我們假設所用的靜態詞典(即原先馬可夫語言模型裡用的)裡沒有一個雙字詞是唸/ㄓ ㄏㄨㄚ/，而由/ㄓ/查出的單字詞和由/ㄏㄨㄚ/查出的單字詞的所有可能組合出的雙字詞(稱片語較恰當，但本文都將稱爲詞)中，貯列裡可以找到"支花"、"枝花"等二個，其中"枝花"在貯列的後面，所以我們可訂出如下的規則：

(R1) 動態詞典結構中有多個選擇時，則選取時間上最晚的。

當再輸入一個音節/ㄩㄚˇ/而成爲/ㄓ ㄏㄨㄚ ㄩㄚˇ/時，語言模型要將"枝花"改成"支花雨"，雖然靜態詞典裡沒有"支花雨"這樣一個三字詞，但是從查出的單字詞組合而成的三字詞中， 貯列裡可以找到一個，再依據如下的規則：

(R2) 當動態詞典中可找出不等長度的片語時，令長度較長的有較高的出現機率。

所以"支花雨"會被選取，而不會選擇"枝花雨"。

如果動態詞典機制依循(R1),(R2)兩規則來處理，再令組詞單元先組出長詞並去動態詞典中尋找看有無出現過，則這樣的機制可被用來記錄量詞與名詞之間的關聯性，也就是將來再輸入"二支花雨傘"或"五隻貓"的注音時，"支"與"隻"就不會轉換錯誤。可是，"四枝花"的"枝"就不一定正確，因其與前後文形成的片語較短，而依規則(R1)可能選到"隻"或"支"，另外，輸入"三隻小貓"的注音時，"隻"可能會被轉成"芝"，因爲動態詞典裡沒有"..隻小貓.."的片語存在，而靜態詞典裡卻有"三芝"這個詞(假設此詞的出現頻率夠大)，不過沒有關係，只要將錯字改正然後存入動態詞典裡， 下一次就不會再錯了。同樣道理，當第一次輸入詞典裡未收錄的一個名詞、術語時，語言模型會輸出錯誤的字，但改正後將此詞語放入動態詞典，

則下一次就不會再錯了。由此可知，我們的動態詞典機制需要的記憶空間大小是愈大愈好，以儘量記錄靜態詞典裡未收錄的詞語，以及使用者個人的習慣用語。

2.2 動態詞典的結構設計

前述的簡單貯列結構會碰到的一個問題是與破音字有關的，如果貯列裡存有"時間的寶貴"之片語，則當輸入"實踐的重要"的注音時，轉換出的文句會是"時間的重要"，因爲"間"是破音字，由/ㄐㄧㄢˋ/去查單字詞時也會找到"間"，而由/ㄕˊ/可找出"實"、"時"、...等單字詞，然後由組合出的三字詞去查貯列時會查到"時間的"。這樣的破音字問題，原先的馬可夫中文語言模型也會碰到，並且對馬可夫模型來說是很難加以改進的，因爲訓練模型時使用的語料，通常只有文字資料而無注音資料。對於動態詞典機制來說，則可將使用者輸入的注音資料一併存入貯列裡，如此將來到貯列查循時，注音資料也要符合才算比對成功，如此就可解決破音字帶來的問題了。

前面我們說到動態詞典裡去查看某一個三字詞或更多字的詞存不存在，只使用簡單的貯列結構在實做上是不可行的，因爲一個音節若有 10 個同音字的話，則 5 個連續音節可能組合出的五字詞就有 100,000 個，不用說六、七或更多字組成的詞了，再加上我們希望動態詞典愈大愈好(如容量爲 20,000)，則字串比對的運算量及所需的時間是可想而知的。所以，我們就將簡單貯列之結構改換成赫序表(hash table)[13]與貯列的複合結構，以赫序表的功能來加快搜尋的速度，而以貯列的功能來控制元素的插入與刪除。爲了配合赫序表的操作，實作上我們就把一個中文句子的組成字(如:ABCDE)，拆成相鄰字組成的雙字詞序列(如:AB, BC, CD, DE)，然後將這些雙字詞插入赫序表，可是如此做就無法迅速地確定雙字詞之間的時間次序(因爲要依貯列的 link 循序下去找)，例如由兩音節/ㄓ ㄏㄨㄚ/去查出"支花"與"枝花"都在赫序表裡時，如何知道那一個是較晚存入的，我們採用的一個解決辦法是應用郵戳(time stamp)的觀念，即每次要將一個雙字詞存入赫序表時(不管是新插入的或是更新的)，就將目前時間計數器的值取出一起存入到赫序表裡，然後將時間計數器加一。

採用赫序表之結構也意味由單字詞去組合出多字詞的方式必需跟著改變，如要查片語 XYZ 是不是存在於動態詞典裡，就必需先將 XYZ 分解成 XY 與 YZ 再各別去查 XY 與 YZ 在不在赫序表裡，這樣也就引起了張冠李戴的可能性，如先前已輸入了語句"一隻黑貓帶著兩支黑拐杖並撐著一支花雨傘"，且已被存入動態詞典裡，則接著輸入 "一隻黑貓" 的四個音節時，會轉換出來的是"一支黑貓"，這是因爲 "一支" 的郵戳比 "一隻" 晚，而 "支黑" 比 "隻黑" 晚，乍看之下，也許會認爲何不依據郵戳值是否連續去判斷原先是否是接在一塊的，不過，這樣做是不可行的，例如若輸入 "一隻黑狗和二隻黑貓" 後，那是不是使得"隻黑" 與 "黑狗"不被認爲是連在一塊的，因爲"隻黑"的郵戳值會被第二次出現的"隻黑"所更新，對於這樣的問題，我們的解決方法是再增加一個赫序表，用以儲存中文句子裡相鄰三字所形成之三字詞，例如要將中文句子 ABCDE 存入動態詞典，就相當於把 AB,BC,CD,DE 等雙字詞及對應的注音資料存入第一個赫序表，而把 ABC, BCD, CDE 等三字詞存入第二個赫序表，然後，當我們要查三字詞 XYZ 存不存在於動態詞典裡時，就可先到第一個赫序表去看 XY 和 YZ 的字與音是否可比對成功，通過後再到第二個赫序表去看 XYZ 是否存在，而當要查一個四字詞 WXYZ 存不存在於動態詞典裡時，也是先到第一個赫序表去看 WX,XY,YZ 可否比對成功，然後再到第二個赫序表去看 WXY 與 XYZ 存不存在，至於五、六、...等多字詞的檢查可依此類推，如此，前述例子之"一隻黑貓"就不會轉換錯誤了。關於檢查多字詞是否在動態詞典裡，實作上我們是以漸進方式來作，詳細說明在第三節。

3、動態詞典機制與中文馬可夫模型之整合

3.1 整合的方法

由前面的說明可知動態詞典機制的功用在於補充中文馬可夫模型的一些不足的地方，如靜態詞典或訓練語料裡收錄之名詞、術語不夠詳盡；階數太低，無法應用較長的片語資訊來作多重選擇時分辨的依據。這裡我們就實際來看如何把動態詞典的功能整合到馬可夫模型裡，也就是說要以馬可夫模型爲基礎，而把動態詞典的查閱方式轉變成爲合乎馬可夫模型要求的處理方式。很明顯的，我們無法反過來把馬可夫模型整合到動態詞典機制裡，因爲馬可夫模型是一種機率模型比動態詞典機制複雜多了。

馬可夫模型的組成元素有兩種，其中一種稱爲狀態(state)，另一種稱爲狀態轉移機率(state transition probability)，簡稱爲轉移機率[14]。建立中文的馬可夫模型的一種作法是，把中文裡的詞當作是馬可夫模型的狀態，然後用一個詞典來收錄馬可夫模型的所有可能的狀態(也就是"詞")，接著才能依據這個狀態集合(即詞典)，去訓練語料中求取那些用來計算轉移機率數值的參數。在我們以前提出的複合馬可夫模型裡，多了一種元素，稱爲狀態機率(state occupying probability)，表示一個中文詞不管前後文獨立被使用的機率。

分析馬可夫模型的組成後，我們整合動態詞典機制進來的第一步是，把那些從動態詞典裡查到的多字(兩字及兩字以上)詞(片語)也當作是馬可夫模型的狀態，這樣的觀念雖然簡單，但卻有實作上的問題要考慮。在原先的馬可夫模型裡，我們可以由所用的靜態詞典而知道最長的詞的長度(如爲 5)，因此到靜態詞典查詢輸入音節可能對應的詞時，就只需考慮相連的一至五個音節的組合情形，可是對於從動態詞典裡去查可能對應的詞時，我們希望詞長是愈長愈好，然而我們如何預先知道一個詞會長到多長，所以對於動態詞典裡查出的詞，必須再加入一種機制，以讓它能夠從基本的單字詞成長爲雙字詞，再成長爲三字詞，如此繼續下去。

爲了讓動態詞典裡查出的詞能夠動態增長下去(可能的話)，所以我們整合動態詞典機制到馬可夫模型的第二步是，利用馬可夫模型裡從一個狀態轉移到另一個狀態的轉移機率值，來攜帶前後二狀態中相鄰

的兩個中文字是否有相連地出現在動態詞典裡的資訊。實際上的作法是，當要計算馬可夫模型裡由一個狀態轉移到另一個狀態的轉移機率時，就先檢查此二狀態的相鄰的兩個中文字是否有相連地出現在動態詞典裡，如果有的話，就設定一個高於 0.99 的值作爲狀態轉移機率值的求取結果，即令

$$P(Y|X) = Pc(X,Y) = 0.99 + 0.01(0.06\frac{Ts(X,Y)}{Tm} + 0.94*2^{-\left\lfloor\frac{Ts(X,Y)-Tc}{64}\right\rfloor}) \quad (1)$$

其中 X 表示前一個狀態的詞尾字， Y 表示後一個狀態的詞頭字， Ts(X,Y) 表示由動態詞典裡查出的郵戳值， Tm 表示最大之郵戳值， Tc 表示目前時間計數器裡的值， 2 的冪次項 2^{-X} 用以讓新進入動態詞典的詞的影響力隨著時間而衰減， Ts(X,Y)/Tm 項用以分辨動態詞典裡的詞的時間次序。如果檢查得知兩相鄰字未曾相連地出現在動態詞典裡的話，就仍然使用原先馬可夫模型裡的求取方法[15,16,]，我們使用的修正過的公式[10]是

$$P(Y|X)=Pm(X,Y)= \begin{cases} (1-Pe)*N(X,Y) / N(X), & \text{if } N(X,Y) > 0 \\ Pe*(N(Y)+1) / (Nt+10000), & \text{if } N(X,Y)=0 \end{cases} \quad (2)$$

$$Pe = (Ns(X) + 1) / (N(X) + 2) \quad (3)$$

其中 N(X,Y) 表示在訓練語料中 Y 緊跟著 X出現的次數， N(X) 表示 X 在訓練語料中出現的次數， Ns(X) 表示具有 N(X,Y)=1 之不同的 Y 的個數， Nt 表示訓練語料的總字數，而 Pe 則表示逃脫機率。這裡我們假設由原馬可夫模型求出的狀態轉移機率值不超過 0.99 ，這樣的假設我們以 20,000 字以上的測試語料來驗證，結果只發現一個例外情形，就是由單字詞"什"到單字詞"麼"的轉移機率會大於 0.99，對待此例外情形，我們可以想像"什麼"有存在於動態詞典裡就好了，雖然"什"與"麼"會組合成雙字詞，但若要繼續成長爲三字詞"什麼 X"或"X 什麼"，則必須在第二個赫序表裡要能夠找到它們才行。所以，我們可依據轉移機率值是否大於 0.99 ，來判斷相鄰的兩個中文字是否有相連地出現在動態詞典裡，此外還可以比較兩個轉移機率 P(Y1|X1)與 P(Y2|X2)的大小，而得知兩者(X1,Y1)、 (X2,Y2)在動態詞典裡的時間次序(如果兩者都有在動態詞典裡) 。當發現從一個單字詞狀態 X 到另一個單字詞狀態 Y 的轉移機率 P(Y|X) 大於 0.99 時，才要進行組合成長詞的處理，若有一者(X 或 Y)爲多字詞就不作此種處理，即利用原馬可夫模型裡的單字詞狀態作爲整合動態詞典機制進來的媒介。

3.2 動態規畫法找最佳路徑之回顧

在說明機動地組合、增長動態詞典中的詞的作法時，也要同時看最佳路徑(句子)是如何被找出的，因爲這兩種動作都得作漸進式處理而有密切的關係，漸進式處理就是每次使用者輸入一個音節後，就要作一部份的處理(包括查靜態詞典、組合增長動態詞典中的詞、及找目前的最佳路徑)，不必等到整句話的注音輸入後才開始處理，而徒增使用者的等待時間。這裡我們先回顧在複合馬可夫模型裡，以動態規劃找最佳路徑的方法[9,10]，令 $S_1,S_2, ..., S_t$ 代表使用者已輸入的音節序列， V_{tk} 代表由 k 個音節 $S_{t-k+1}, S_{t-k+2}, ..., S_t$ 去查靜態詞典所得到的詞語的集合， W_{tkj} 代表 V_{tk} 的第 j 個元素，接著令 Q(t,k,j) 表示從起始點到詞語 W_{tkj} 爲終點之間的一條最佳路徑的機率值，而所謂的最佳路徑是以複合馬可夫模型來評定的，另外， K 表示詞典裡最長詞的詞長，則 Q(t,k,j) 可以如下之遞迴公式

$$Q(t,k,j) = P(W_{tkj})[\underset{h=1}{\overset{K}{MAX}} \underset{i=1}{\overset{|V_{t-k,h}|}{MAX}} Q(t-k,h,i)\cdot P(C_{tkj}|D_{t-k,h,i})] \quad (4)$$

去求值，公式(4)裡，$|V_{tk}|$ 表示集合 V_{tk} 的元素個數， C_{tkj} 代表 W_{tkj} 之詞頭字，而 D_{tkj} 代表 W_{tkj} 之詞尾字。接著，我們可依如下公式

$$Q(t) = \underset{h=1}{\overset{K}{MAX}} \underset{i=1}{\overset{|V_{t,h}|}{MAX}} Q(t,h,i) \quad (5)$$

去找出最佳路徑的機率值 Q(t)，然後由最佳路徑的終點詞回溯(backtrack) 回去到起始點，去將音節序列 $S_1,S_2, ..., S_t$ 最可能對應的中文句子找出來。例如已輸入的音節序列是 /ㄧˋ ㄓ ㄏㄨㄚ ㄩˇ ㄙㄢˇ/，則從靜態詞典查出的可能被對應的詞語以及可能連結出來的路徑就如圖 1 所顯示的情形，依據圖 1 來

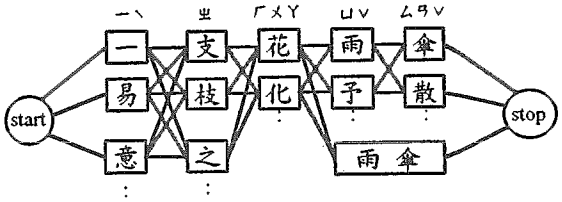

圖1 可能連結之路徑

看，前面公式裡的 $W_{5,1,1}$ 就代表"傘"， $W_{4,1,1}$ 代表"雨" $W_{5,2,1}$ 代表 "雨傘"， $V_{5,1}$ = {傘、散、...}，可能的連結路徑如： (一)(枝)(花)(雨傘)、(一)(支)(花)(雨)(傘)、...等；至於目前的最佳路徑是那一條，則需將有關的狀態機率 P(X)與轉移機率 P(Y|X) 的值帶入公式(4)、(5)去決定，我們只說"目前的"最佳路徑，因爲，這五個音節之後可再繼續輸入如 /ㄉㄜ。 ㄏㄢˊ ㄧˋ/ 等音節，也就是說公式(4)、(5)可被用於作漸進式處理。

3.3 合成長詞

在 3.3 和 3.4 節的說明裡，將假設語句"除了一支花雨傘、一枝花之外，他還帶著一隻貓"及對應的注音資料已被存入動態詞典裡。當使用者輸入/ㄧˋ ㄓ/這兩音節後，我們如果一邊以公式(4)去計算 Q(t,k,j) 的數值，一邊去檢查 $P(C_{tkj}|D_{t-k,h,i})$ 的數值是否大於 0.99，就會發現到 P("支"|"一") 和 P("枝"|"一") 的數值都大於 0.99 ，並且轉移機率的前後兩個狀態都是單字詞，所以 "一支" 和 "一枝" 要被組合成雙字詞。如果只考慮雙字詞之合成，則公式(4)可修正成

$$Q(t,k,j) = \max_{h=1}^{K} \max_{i=1}^{|V_{t-k,h}|} U(t,k,j,h,i) \quad (6)$$

$$U(t,k,j,h,i) = \begin{cases} \dfrac{Q(t-k,h,i)}{P(W_{t-k,h,i})} * \dfrac{[P(C_{tkj} \mid D_{t-k,h,i})-0.99]*100{,}000+102}{1{,}000{,}000}, \\ \quad \text{if } P(C_{tkj} \mid D_{t-k,h,i}) > 0.99 \\ \quad \text{and } k=1,\ h=1 \\ Q(t-k,h,i)*P(C_{tkj} \mid D_{t-k,h,i})*P(W_{tkj}), \quad \text{otherwise} \end{cases} \quad (7)$$

其中公式(7)的下半部跟公式(4)是相同意義的，新加入的是公式(7)的上半部，其意義是把所合成雙字詞的詞頭字的狀態機率去除掉，再乘上所合成雙字詞被定義的狀態機率，所以要如此處理，可考慮路徑 (一)(之) 和 (一)(支) 的機率，即 P(一之) = P("一") * P("之"|"一") * P("之")，和 P(一支) = Q(1,1,1) / P("一") * P("一支")，其中 Q(1,1,1) = P("一")。至於我們將合成的雙字詞的狀態機率作這樣的定義，如公示(7)的上半部右邊，目的是要讓合成的雙字詞的出現頻率當以靜態詞典查出之雙字詞的的觀點來看時是 1 到 1001 之間，我們用以定義靜態詞典查出之詞的狀態機率的公式是

$$P(X) = (N(X) + 1) / 1{,}000{,}000, \quad (8)$$

其中 N(X) 表示出現頻率。

如果考慮要能夠由雙字詞再去組合成三字詞，則必須再增加一個變數 R(t,k,j) 以記錄當公式(6)的 Q(t,k,j) 達到最大時對應的公式(7)裡的 $P(C_{tkj} \mid D_{t-k,h,i})$ 的值，至於用 pathh(t,k,j) 和 pathi(t,k,j) 分別去記錄 Q(t,k,j) 達到最大時的 h 和 i 的下標值，則是原先的馬可夫模型本來就要用到的，因爲作回溯時必須用到這樣的變數，然後，我們就可將公式(7)改成

$$U(t,k,j,h,i) = \begin{cases} Q(t-k,h,i)*P(C_{tkj} \mid D_{t-k,h,i})*3, \\ \quad \text{if } P(C_{tkj} \mid D_{t-k,h,i}) > 0.99 \text{ and } R(t-k,h,i) > 0.99 \text{ and} \\ \quad k=1 \text{ and } h=1 \text{ and } pathh(t-k,h,i)=1 \text{ and char. sequence} \\ \quad W_{t-2,1,pathi(t-k,h,i)}, W_{t-1,1,i}, W_{t,1,j} \text{ be in 2nd hash tab.} \\ Q(t-k,h,i)*Pm(D_{t-k,h,i}, C_{tkj})*P(W_{tkj}), \\ \quad \text{else if } P(C_{tkj} \mid D_{t-k,h,i}) > 0.99 \text{ and } R(t-k,h,i) > 0.99 \text{ and} \\ \quad k=1 \text{ and } h=1 \text{ and } pathh(t-k,h,i)=1 \text{ and char. sequence} \\ \quad W_{t-2,1,pathi(t-k,h,i)}, W_{t-1,1,i}, W_{t,1,j} \text{ not in 2nd hash tab.} \\ \dfrac{Q(t-k,h,i)}{P(W_{t-k,h,i})} * \dfrac{[P(C_{tkj} \mid D_{t-k,h,i})-0.99]*100{,}000+102}{1{,}000{,}000}, \\ \quad \text{else if } P(C_{tkj} \mid D_{t-k,h,i}) > 0.99 \\ \quad \text{and } k=1 \text{ and } h=1 \\ Q(t-k,h,i)*P(C_{tkj} \mid D_{t-k,h,i})*P(W_{tkj}), \quad \text{otherwise} \end{cases} \quad (9)$$

在公式(9)的最上列，是條件符合組合成三字詞以上之多字詞的情況，乘上數值接近 1 之 $P(C_{tkj} \mid D_{t-k,h,i})$ 是要將在動態詞典裡的時間次序考慮進來，而乘上 3 的目的是，讓詞長每增加一個字時整個詞的狀態機率就放大三倍；公式(9)的第二列，表示要組合成三字詞的三個字未曾在第二個赫序表裡相連地出現的情況，所以不能讓他們組合成三字詞，爲了不讓他們組成三字詞，我們就強迫取消後兩個字組成雙字詞的機會(雖說後二字有相連地出現在第一個赫序表裡)，因此第二列的式子是乘以原馬可夫模型之轉移機率 $Pm(D_{t-k,h,i}, C_{tkj})$ 及狀態機率 $P(W_{tkj})$，由於第二列式子要強迫取消後兩個字組成雙字詞，所以在設定變數 R(t,k,j) 的值時，若是由本列裡的 $P(C_{tkj} \mid D_{t-k,h,i})$ 數值去設定，就必須以 $Pm(D_{t-k,h,i}, C_{tkj})$ 數值取代；至於公式(9)裡的第三、四列，其含意就如公式(7)裡的。

3.4 路徑尋找演算法

當以公式(9)來處理使用者輸入的三個音節 /ㄧ ㄟ ㄕ ㄏㄨㄚ/ 時，由圖一的可能連結之路徑，可知道 (一)(枝)(花) 和 (一)(支)(花) 等由單字詞串接成的路徑會被組合成三字詞(令動態詞典裡已存有 3.3 節提到的語句)，而由於 "..一枝花.." 比 "..一支花.." 晚存入動態詞典，所以"一枝花"會被選取，此時由"花"狀態往回指之路徑變數 pathh(3,1,1) 和 pathi(3,1,1) 會分別被設成 1 與 2，即指向"枝"，而由"枝"狀態往回指之路徑變數會被設成指向"一"。接著，當使用者再輸入第四個音節 /ㄩㄢˇ/ 時，得到的輸出會是"一枝花雨"，而不是所想要的"一支花雨"，造成這種結果的原因是，當輸入到第三個音節時，依據公式(6)來進行的動態規畫之路徑選取，就會認定從起始狀態走到"花"的最佳路徑是 (一)(枝)(花)，而把其餘的可能路徑都排除掉了，這樣說是因爲公式(6)裡只取一個最大值，並且由狀態"花"往回指之路徑變數只有一組 pathh(3,1,1)、 pathi(3,1,1) 用以存最大值時的下標值，因此就無法循其他路徑回溯回去。

由前述的例子可知道，一開始不具有最大機率的路徑，可能後來會變成是具有最大機率的路徑，如"一支花" 和 "雨" 可以組成四字詞，"一枝花" 和 "雨" 則不能組成四字詞，而依據公式(9)第一、二列式子的定義，可知四字詞的機率會大於三字詞與單字詞的連結，所以，我們必須修正公式(6)以順應公式(9)，此時如果仍要找循整合模型所定義的最佳路徑，則所需花用的計算量與記憶空間都會隨輸入的音節個數成指數成長，因此我們的修正方式是，只讓進入到一個狀態的前幾個具有較大機率值的路徑的資料(累乘機率值、回溯指標)保留下來，其餘路徑的資料就捨棄掉，也就是要以 beam search 方法[17] 來找尋計算量限制內的最佳路徑，令 M 爲 beam 的寬度(即每個狀態上要保留的路徑個數)，則公式(6)要更改爲

$$\text{IndexSet} = \{ (h,i,m) \mid 1 \le h \le K,\ 1 \le i \le |V_{t-k,h}|,\ 1 \le m \le M \} \quad (10a)$$

$$Q(t,k,j,1) = \max_{\text{IndexSet}} U(t,k,j,h,i,m); \qquad \lambda_{t,k,j,1} = \underset{\text{IndexSet}}{ARGMAX}\ U(t,k,j,h,i,m) \quad (10b)$$

$$Q(t,k,j,2) = \underset{\text{IndexSet-}\lambda_{t,k,j,1}}{MAX} U(t,k,j,h,i,m);$$

$$\lambda_{t,k,j,2} = \underset{\text{IndexSet-}\lambda_{t,k,j,1}}{ARGMAX} U(t,k,j,h,i,m) \quad (10c)$$

:

$$Q(t,k,j,M) = \underset{\text{IndexSet} - \lambda_{t,k,j,1} - \lambda_{t,k,j,2} - \cdots - \lambda_{t,k,j,M-1}}{MAX} U(t,k,j,h,i,m) \quad (10d)$$

其中 Q(t,k,j,n)裡的下標 n 表示 Q(t,k,j,n)是所有可能的 U(t,k,j,?,?,?)中的第 n 大者；至於這裡的 U(t,k,j,h,i,m)變數比公式(9)的 U(t,k,j,h,i)變數多了一個次元，就表示公式(9)裡相互遞迴參考的 Q(t,k,j)和 U(t,k,j,h,i)變數，都要再增加一個表示候選路徑名次的次元以配合公式(10)；而公式(9)裡用到的 R(t,k,j)變數，也要增加一個次元而成爲 R(t,k,j,n)，以記錄公式(10)裡求得 Q(t,k,j,n)數值的那個 U(t,k,j,h,i,m) 變數所對應的轉移機率 $P(C_{tkj} \mid D_{t-k,h,i})$的值；再者，回溯指標變數 pathh(t,k,j)和 pathi(t,k,j)也要再加入一個次元而成爲 pathh(t,k,j,n)和 pathi(t,k,j,n)，以配合新增的另一個回溯指標變數 pathm(t,k,j,n)，來共同指示時刻 t 、詞長 k 、第 j 個同音詞上的第 n 個候選路徑的前一站在那裡(即時刻 t-k 、詞長 pathh(t,k,j,n)、第 pathi(t,k,j,n)個同音詞上的第 pathm(t,k,j,n)個候選路徑)，由此可知公式(10)裡的 $\lambda_{t,k,j,n}$ 變數就是表示一個包含 pathh(t,k,j,n)、 pathi(t,k,j,n)、 pathm(t,k,j,n)等三個成員的向量；除此之外，公式(9)裡的其它變數 W_{tkj} 、 C_{tkj} 、 $D_{t-k,h,i}$ 就不必更動了。

依據公式(10)裡 Q(t,k,j,n) 的定義，原先用以找出最佳路徑的機率值 Q(t)的公式(5)就要改成

$$Q(t) = \underset{h=1}{\overset{K}{MAX}} \; \underset{i=1}{\overset{|V_{t,h}|}{MAX}} Q(t,h,i,1) \quad (11)$$

4 、測試實驗

在第三節裡，我們提出了一種整合動態詞典機制到複合馬可夫模型的方法，這裡就以實驗的方式來驗證其正確性。另外，動態詞典機制所提供的功能，是否能使音節至中文字的正確轉換率變得較平穩，或者會陰錯陽差而產生反效果呢？這也是我們需要驗證的。

在依據我們的整合方法來運轉的馬可夫模型裡，動態詞典機制達成的功能是否仍然如原先所想的(如第二節裡的敘述)，這裡，我們以定性實驗的方式來探討，首先把整合後的模型製作成軟體程式，然後輸入如前面提到的例句 "除了一支花雨傘，一枝花之外，他還帶著一隻貓" 的注音給這個程式，程式接著顯示準備輸出的文句，其中若有錯字我們就按鍵加以更正，然後按一個控制鍵表示要把文句正式輸出及存入動態詞典，之後，我們就再逐個音節輸入"一支花雨傘"的注音，看輸出的文句是否會依照"一"、"一隻"、"一枝花"、"一支花雨"的次序出現，結果的確是如我們所想的；然後，在輸入一些其它的文句後，再依序輸入"一支花雨傘"的注音，顯示出來的文句仍是同樣的變化次序，所以我們提出的整合方法確實可把動態詞典的功能熔合進來。

關於動態詞典的功能，是否會陰錯陽差而產生反效果(音至字轉換率變壞)，這裡我們就以定量的實驗來探討，在以下敘述的測試實驗裡，我們使用了一個約有 52,000 個詞的靜態詞典，用以查詢最多連續 5 個音節所可能對應的語詞。關於原馬可夫模型的訓練，我們分成三種訓練條件來作比較，第一種訓練條件稱爲 TR1 ，使用了國小國語課本之課文(約 90,200 個中文字)，報紙社論文章(約 75,000 個中文字)，以及短篇小說、寓言(約 105,500 個中文字)，來統計出那些有關的計次參數的數值，以便代入公式(2)及(3)去估計狀態轉移機率的數值；第二種訓練條件稱爲 TR2 ，使用了 TR1 的語料外，再加入 8.9 Mbytes(剔除了 ASCII 碼部分及例行的氣象預報)由台灣學術網路上蒐集到的新聞報導短文(1995 年)；第三種訓練條件稱爲 TR3 ，使用了 TR2 的語料外，再加入 21.3 MBytes(剔除了 ASCII 碼部分及例行的氣象預報)由網路上另外蒐集到的新聞報導短文(1996 年)。至於實驗裡作爲測試語料的文章，依其來源分爲三類，第一類取自於 22 篇小學生的作文[18]，共 9,119 個音節，第二類取自於 13 篇晚報社論，共 7,783 個音節，第三類則是取自於報紙上的 11 篇關於醫藥知識的報導，共有 7,522 個音節，這些測試語料都未被用於訓練馬可夫模型。

爲了比較有無加入動態詞典機制對不同測試語料、及正確轉換率的影響，我們設定了三種測試條件。第一種是最基本的情況，將馬可夫模型裡的狀態轉移機率都直接設爲 1 ，並且不使用動態詞典機制，相當於零階、詞爲狀態之馬可夫語言模型；第二種測試條件是，依公式(2)與(3)去計算狀態轉移機率，但不使用動態詞典機制，相當於使用我們以前提出的複合式馬可夫模型；第三種測試條件是，使用整合了動態詞典機制的複合式馬可夫模型。我們設定動態詞典的大小爲，第一個赫序表含有 16,000 個 entries ，第二個赫序表含有 24,000 個 entries ，而 beam search 的寬度 M 則設爲 5 。第三種測試條件的前提是，每次輸入一句中文的音節後，自動音轉字處理輸出的中文字，如果有轉換出錯誤的字，輸入者自己要立即進行錯字更正的操作，而不是留給另外一個校對者去操作。

依據前述的實驗條件，讓實現各個條件的程式分別去處理三類的測試語料，也就是把測試語料對應的注音資料送給程式去轉換出中文字，結果我們得到了如表 1 所示的正確轉換率的數值，我們定義轉換率

表 1 不同實驗條件與測試語料下得到的轉換率

測試語料 / 音轉字模型	小學作文 (9119 字)	報紙社論 (7783 字)	醫藥報導 (7522 字)
零階、詞狀態馬可夫模型	92.8	90.1	86.5
複合式馬可夫模型(TR1)	94.4	91.5	88.6
複合式馬可夫模型(TR2)	94.1	93.2	89.0
複合式馬可夫模型(TR3)	93.6	93.8	89.5
整合動態詞典之馬可夫(TR1)	94.9	93.5	93.3
整合動態詞典之馬可夫(TR2)	94.5	94.7	93.7
整合動態詞典之馬可夫(TR3)	94.3	95.0	93.8

爲，正確轉換出來(未作錯字更正前)的中文字字數除以輸入的音節個數，將此表裡第 2,3,4 列的轉換率和 5,6,7 列的轉換率分別作比較，可看出加入動態詞典機制之整合模型得到的轉換率，都比未加入動態詞典機制的模型的轉換率高，所以動態詞典機制的確可用以提升轉換率，並且使得文章種類不同而引起的轉換率變化緩和許多，對於三類的測試語料，整合模型的轉換率都相當平穩地維持在 93% 以上，不像原先的複合式馬可夫模型的轉換率會隨文章種類而有大幅度的變化，例如由第 2 列可看出此模型的轉換率會由 94.4% 變化到 88.6%。另外，比較第 2,3,4 列的轉換率，可看出馬可夫模型的訓練語料變多，對同類型文章的測試語料的轉換率會有不少幫助(93.8% - 91.5%)，可是對不同類型文章的測試語料的轉換率的幫助很少或有害(93.6%- 94.4%)，不過，整合動態詞典機制後，這種對比情形就會緩和一些，如第 5,7 列所顯示的 95.0%-93.5%與 94.3%-94.9%。再從垂直方向來看表 1 ，可發現小學生作文的文章類型，轉換率隨不同模型的改進相對地較小(94.3% - 92.8%)，因爲此類文章較簡單且一開始的轉換率就較高，再者，轉換率會隨馬可夫模型的訓練語料變多而些微下降，因爲 TR2 與 TR3 的訓練語料的文章類型不是小學生作文，至於晚報社論文章的轉換率改進，很明顯地反應了訓練語料之數量的影響，並且有飽和的趨勢，而醫藥報導的文章類型，前面的模型獲得的轉換率都很低(都低於 90%)，因爲馬可夫模型的訓練語料未包含此類型的文章，不過，當使用整合動態詞典機制的模型後，轉換率就可提升到 93%以上，這是因爲醫藥報導這一類的文章含有相當多的術語、名詞(如"異位性"、"攝護腺")未收錄在靜態詞典裡，而經由動態詞典機制的調適能力，可使轉換率提升許多。

5 、結語

動態詞典的觀念雖然很早就被其他研究領域所應用，不過，本文以中文的特性和需求爲基礎，將動態詞典的觀念加以擴充和具體化，以應用到中文語言模型裡。就動態詞典的結構來說，本文使用了兩個赫序表，並且赫序表的大小是要愈大愈好；就功能上來說，動態詞典機制除了用以掌握各詞語的動態使用情形之外，本文還賦予動態詞典機制其它的功能，如記錄相鄰中文字的共出現(co-ourrence)關係，記錄專有名詞，記錄使用者的習慣用語等。我們認爲，只要搜尋速度不會慢到不能接受，赫序表的記憶容量可以儘量加大，如此，整合了動態詞典機制的語言模型，就會隨著使用者的愈常使用，而變得愈善解人意。

關於動態詞典機制與馬可夫中文語言模型的整合，我們提出了一種整合的方法，使得動態詞典裡查出的詞，可以在馬可夫模型裡由單字詞機動地成長爲雙字詞、三字詞、...等等，而最佳路徑(句子)的搜尋，則由原先馬可夫模型裡使用的動態規畫法修正爲 beam search 法，以便在熔入動態詞典機制後，仍然能維持原馬可夫模型的快速、漸進的處理方式。我們已將整合後的模型製作成軟體程式，然後以測試語料去測試、驗證，結果顯示整合模型的確會依照預期的方式運轉，並且正確轉換率會比原先的馬可夫模型的高而且穩定。

參考文獻

[1] Kuo, J. J., J. H. Jou, M. S. Hsieh, and F. Maehara, "The Development of New Chinese Input Method -- Chinese Word-string Input System", Proceedings of International Computer Symposium (Tainan, Taiwan), pp. 1470-1479, 1986.

[2] Hsieh, M. L., T. T. Lo and C. H. Lin, "Grammactical Approach to Converting Phonetic Symbols into Characters" , Proceedings of National Computer Symposium (Taipei) , pp. 453-461 , 1989.

[3] Chien, L. F., Some New Approaches for Language Modeling and Processing in Speech Recognition Applications, Ph.D. Dissertation, Department of Computer Science and Information Engineering, National Taiwan University, Jan. 1991.

[4] Lin, M. Y. and W. H. Tsai, "Removing the Ambiguity of Phonetic Chinese Input by the Relaxation Technique", Computer Processing of Chinese and Oriental Languages, pp. 1-24, 1987.

[5] Gu, H. Y., C. Y. Tseng and L. S. Lee, "Markov Modeling of Mandarin Chinese for Decoding the Phonetic Sequence into Chinese Characters", Computer Speech and Language, Vol. 5, No. 4, pp. 363-377, 1991.

[6] 林頌堅等，"國語語音辨認中多領域語言模型之訓練、偵測與調適"，第九屆計算語言學研討會論文集(台南)，第 159-182 頁， 1996 年。

[7] Kuhn, R. and R. De Mori, "A Cache-Based Natural Language Model for Speech Recognition", IEEE trans. Pattern Analysis and Machine Intelligence, Vol. 12, No. 6, pp. 570-583, June 1990.

[8] Rao, P. S., M. D. Monkowski and S. Roukos, "Language Model Adaptation via Minimum Discrimination Information", Proceedings of ICASSP(Detroit, USA), Vol. I, pp. 161-165, 1995.

[9] Gu, H. Y., A Study on a few Relevant Problems about Machine Dictation of Mandarin Speech, Ph. D. Dissertation, Department of CSIE, National Taiwan University, Jan. 1990.

[10] 古鴻炎、陳志耀，"使用新式注音鍵盤及複合馬可夫語言模型之中文輸入系統"，中華民國電腦學會電腦學刊，第七卷，第三期，第 1-9 頁， 1995 年。

[11] Bell, T. C., J. G. Cleary and I. H. Witten, Text Compression, Prentice-Hall Inc., Engle-wood Cliffs, New Jersey, 1990.

[12] Gu, H. Y. "A New Chinese Text Compression Scheme Combining Dictionary Coding and Adaptive Alphabet-character Grouping", Journal of Computer Processing of Oriental Languages, Vol. 10, No. 3, pp. 321-335, 1997.

[13] Weiss, M. A., Data Structures and Algorithm Analysis in C, Addison-Wesley, 1997.

[14] Ross, S. M., Introduction to Probability Models, third edition, Academic Press, Inc., 1985.

[15] Katz, S. M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE trans. Acoust., Speech, and Signal Processing, pp. 400-401, March 1987.

[16] Witten, I. H. and T. C. Bell, "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression", IEEE trans. Information Theory, Vol. 37, pp. 1085-1094, 1991.

[17] Nilsson, N., Principles of Artificial Intelligence, Tioga: Palo Alto, 1980.

[18] 鄭博眞編著，小學生作文寶典習作篇，小叮噹圖書公司， 1993 年六月。