基於特製隱藏式馬可夫模型之中文斷詞研究

Chinese Word Segmentation using Specialized HMM

林千翔、張嘉惠

國立中央大學資訊工程學系

Email: pshivp@db.csie.ncu.edu.tw, chia@csie.ncu.edu.tw

摘要

中文斷詞在中文的自然語言處理上，是個相當基礎且非常重要的工作。近年來的斷詞系統較傾向於機器學習式演算法來解決中文斷詞的問題，但使用傳統的作法，隱藏式馬可夫模型在解決中文斷詞的問題上，無法達到較好的斷詞效能（F-measure 約 80%），所以許多研究都是使用外部資源或是結合其他的機器學習演算法來幫助斷詞。本研究的目的是使用「特製化」（specialization）的概念來提升隱藏式馬可夫模型的準確率，我們的作法是給予隱藏式馬可夫模型更多的資訊，在完全不修改模型之訓練及測試過程的前提下，透過兩階段特製化的方式，分別為擴充「觀測符號」，以及擴充「狀態符號」的方式，大大地改善了隱藏式馬可夫模型的斷詞準確性。第一階段中，我們使用長詞優先法，來增加額外的資訊於隱藏式馬可夫模型中，使得模型擁有更多的斷詞資訊做學習。於實驗結果發現，只使用這個最簡單的長詞優先斷詞方法，確實能大幅地提升隱藏式馬可夫模型的效能。而第二階段中，我們則使用詞彙式隱藏式馬可夫模型（Lexicalized HMM）的概念，也就是只根據某些特製詞（specialized words）來做特製化，將狀態做延伸，實驗結果也證明詞彙式隱藏式馬可夫模型可再次提升系統斷詞效能。

1. 緒論

中文斷詞在中文的自然語言處理上，是非常重要的前置處理工作。許多中文的自然語言相關的領域，例如：問答系統、自動摘要、文件檢索、機器翻譯、語音辨識...等，都需要先處理中文斷詞，可見中文斷詞是個相當基礎且非常重要的工作。

所謂的「中文斷詞」就是將一連串的中文「字串」轉換成「詞串」的組合。例如：「我昨天去台北」這個中文句子，透過中文斷詞的處理後變成「我／昨天／去／台北」，也就是將｛我、昨、天、去、台、北｝字串轉成｛我、昨天、去、台北｝的詞串組合。傳統上，處理中文斷詞會遇到的問題，大致可歸納為兩點，一是「歧義性」（ambiguity）問題，二是「未知詞」（unknown word）問題。歧義性問題即是同一個中文字串，於不同的文章當中，存在不同的斷詞結果，因此容易造成斷詞上的錯誤。歧義型態大致上可以分為兩類：

- 交集型歧義（overlapping ambiguity）

  令 x, y, z 代表中文字元所組成的字串，若 x、z、xy 與 yz 皆為辭典中的詞，則 xyz 的組合，於不同的文章中，可能會被斷詞成 xy/z 或 x/yz 等兩種不同的結果，則 xyz 稱為「交集型歧義字串」。例如：「不可以」三個中文字元所組成的字串，辭典中的詞含有「不、不可、可以」，「不可以」所組成的字串，在下列句子中，因其上下文的不同而產生不同的斷詞結果：「不／可以／忘記」、「不可／以／營利／為／目的」。

- 組合型歧義（covering ambiguity）

  令 x, y 代表中文字元所組成的字串，若 x、y、xy 都是辭典中的詞，xy 的組合中，可在不同的文章中，分別被斷詞成 xy 或 x/y，因為詞 xy 是由 x 與 y 等兩個不同的詞所組成，因此 xy 稱為「組合型歧義字串」。例如：「才能」二個字所組成的字串，辭典中的詞有「才、能、才能」，在下列句子中「才能」組成的字串，將產生不同的斷詞結果：「他／才能／非凡」、「只有／他／才／能／勝任」。

另外，「未知詞」則指辭典中未收錄的詞，包含了人名、地名、組織名、人名地名組織名之縮寫、衍生詞、複合詞、數字型態等，由於人類所使用的語言會隨著社會不斷改變，而持續地創造出新的用語，並且詞的衍生現象也非常地普遍，因此新詞會不斷的出現，辭典永遠無法因應新詞產生的速度，所以會出現未知詞問題，斷詞系統必須能夠處理未知詞，才可提高斷詞的正確性。

近年來的斷詞系統傾向於機器學習式（machine learning-based）演算法來解決中文斷詞的問題，例如 Maximum Entropy (ME) [22]、Support Vector Machine (SVM) [2, 6]、Transformation-Based Learning Algorithm (TBL) [11]、Hidden Markov Model (HMM) [2, 11, 23, 25] 等等，並且顯示了使用機器學習式演算法做中文斷詞，確實可以達到很高的斷詞準確率。

本研究使用隱藏式馬可夫模型（Hidden Markov Model, HMM）來解決中文斷詞的問題。雖然已有數篇研究同樣使用隱藏式馬可夫模型來處理斷詞問題 [2, 11, 23, 25]，但使用傳統的作法，隱藏式馬可夫模型在解決中文斷詞的問題上，無法達到較好的斷詞效能（F-measure 約 80%），因此這些研究 [2, 11, 23] 便結合了其他機器學習演算法，以增加斷詞的效能。我們的研究目的是希望只使用隱藏式馬可夫模型當成主要的演算法，並且應用「特製化」（specialization）的概念來提升隱藏式馬可夫模型的準確率。我們的作法是給予隱藏式馬可夫模型更多的資訊，在完全不修改模型之訓練及測試過程的前提下，透過兩階段特製化的方式，分別為擴充「觀測符號」，以及擴充「狀態符號」的方式，大大地改善了隱藏式馬可夫模型的斷詞準確性。

於第一階段中，為了擴充觀測符號，我們使用最簡單也最常被使用的辭典比對式斷詞演算法－「長詞優先法」（maximum matching algorithm），來增加額外的資訊於隱藏式馬可夫模型中，使得模型擁有更多的斷詞資訊做學習。第二階段擴充狀態符號的方式，我們則使用詞彙式隱藏式馬可夫模型（Lexicalized HMM）的概念，也就是只根據某些特製詞（specialized words）來做特製化，將狀態做延伸，來提升系統斷詞的效能。

2. 相關研究

中文斷詞的研究已有相當歷史，但在近幾年仍陸續新的方法提出，底下我們分別就解決歧義性及未知詞兩個問題分別做文獻回顧。

首先就斷詞歧義性問題，M. Li 等人 [9] 於 2003 年的研究中，提出一種非監督式（unsupervised）訓練的方法，藉由訓練 Naïve Bayes 分類器，來解決中文斷詞的交集型歧義問題，實驗結果可達到 94.13% 的準確率。另一方面，解決組合型歧義比解決交集型歧義更加困難，主要的原因是，要解決組合型歧義則需要依賴更多的內文資訊，如句法分析（syntactic）、語意分析（semantic） 以及前因後果的資訊（pragmatic information）等，才能正確的解決這類的歧義問題。1999 年 J. H. Zheng 等人 [26] 使用規則式（rule-based method）的作法來處理組合型歧義，並達到 85 % 的準確率。而 2002 年 X. Luo 等人 [12] 的研究，則是使用類似於自然語言處理領域中解決「詞義消歧」（word sense disambiguation）的問題，來解決組合型歧義問題，該篇研究使用 TF.IDF 權重計算的公式，重新定義新的 TF 與 IDF 的公式，以此方式來解決組合型歧義問題，達到 96.58 % 的準確率。

解決未知詞問題是做中文斷詞的另一個重要步驟，近年來也有數篇研究再處理未知詞問題。中研院陳克建博士（Chen）等人於 1997 年開始，提出了三篇關於解決未知詞問題的研究 [3, 5, 13]，最早於 1997 的研究 [3]，透過統計斷詞語料庫，產生所有單一字元之已知詞的偵測規則。此階段的研究只能偵測出所有的單一字元的結果，並未真正將未知詞擷取出來。2002 年的研究 [5] ，則是使用人工加上一些統計的方法來建立擷取規則，將所有被偵測出屬於未知詞部分的單一字詞，透過擷取規則以合併這些單一字詞而成為未知詞。實驗中測試 1,160 個未知詞，結果達到 89 % 的擷取準確率。另外於 2003 年的研究 [13] 中，同樣做擷取未知詞的研究，該研究中將所有種類的未知詞的構詞方式以 context free grammar 表示出來，並搭配 bottom-up merging algorithm 來解決大部分統計特性低的未知詞擷取問題。實驗效能達到 75 % 的擷取準確率。

其他解決未知詞問題的研究，如 Zhang 等人 [24] 於 2002 年的研究，則使用類似詞性標示（part-of-speech tagging）的作法，稱為「角色標示」（roles tagging），角色指的是在未知詞的組成成分、上下文以及句子中的其他部分，並且依據句子的角色序列來辨識出未知詞。實驗部分針對中國人名以及外國翻譯名等未知詞做測試，並且達到不錯的準確率以及召回率。

近年來的研究主要趨向於機器學習式的方法來處理中文斷詞，例如 Maximum Entropy（ME）[22] 以及 Conditional Random Field（CRF）[20] 等，這些統計式的學習演算法都是轉成字元分類問題（character classification）來處理中文斷詞問題，並且使用了數種類似的特徵，如目前字元、加上前後各一字元、加上前後各兩字元等，來當作模型的屬性。而 C. L. Goh 等人則使用 Support Vector Machine（SVM）[6] 來解決中文斷詞的問題，該篇研究結合辭典比對式方法－長詞優先法，利用長詞優先法的歧義性以及未知詞的資訊，來加強 SVM 的特徵屬性以改善斷詞效能。另外也有使用感知機（Perceptron）[10] 的方法做斷詞，該篇研究認為 Perceptron 方法雖然與 SVM 類似，不過效能卻較 SVM 差一些，但由於其訓練的速度非常快，因此他們系統提出的主要貢獻就是一個速度快且效能不至於差太多的斷詞方法。

有些研究為了加強學習演算法的斷詞效能，則是結合了數個學習模型，採用混合式作法來處理斷詞問題，如 M. Asahara 等人 [1] 以及 N. Xue 等人 [23] 的研究，為了加強 ME 斷詞結果，這兩篇研究則結合了 SVM、CRF [1] 以及 TBL [23] 等作法，使用混合式方法的結果提升斷詞準確率。另外，許多研究也使用隱藏式馬可夫模型來處理斷詞問題。如 HHMM [25] 系統，便使用了五層的隱藏馬可夫模型，根據不同的目的各自訓練出各個模型，最後再整合成斷詞系統。而 HMM+SVM [2]、HMM+TBL [11] 等兩篇研究，則使用隱藏式馬可夫模型的斷詞結果當成是一個屬性，並分別使用 SVM 以及 TBL 來當成主要的演算法做斷詞，以達到較佳的斷詞結果。此兩篇研究於實驗中也列出只使用隱藏式馬可夫模型做斷詞的效能，其 F-measure 的結果分別為 80.4 % 以及 81.4 %。因此，我們發現

隱藏式馬可夫模型需仰賴其他外部資源或是結合其他的學習演算法，才可以達到可接受的斷詞效能。

3. 系統架構

我們提出的系統是以隱藏式馬可夫模型來解決中文斷詞的問題，並且透過兩階段「特製化」（specialization）的方式來加強隱藏式馬可夫模型的斷詞效能。第一階段特製化，我們結合了長詞優先法的結果來增加觀測符號的資訊，以「擴充觀測符號」；第二階段特製化，則是透過詞彙式（Lexicalized HMM）的特製化過程，以「擴充狀態符號」。

因此我們的系統架構主要可分為兩個部分，第一部份：我們稱之為「M-HMM」，也就是結合長詞優先法於隱藏式馬可夫模型中，讓訓練之模型增加斷詞歧義性與未知詞的資訊，藉此以改善隱藏式馬可夫模型處理中文斷詞的正確性；第二部分：我們稱之為「Lexicalized M-HMM」，這部分透過兩種不同的準則（criteria）來決定特製詞（specialized words），並以屬於特製詞之觀測符號做特製化，透過擴充狀態符號而再次加強斷詞準確率。

3.1. 長詞優先法

長詞優先法（Maximum Matching Algorithm, MM）是最簡單也最廣泛使用的辭典比對式的斷詞方法，其斷詞的策略為由句子的一端開始，試著比對出在辭典中最長的詞，當作斷詞結果，接著去除此詞後，剩下的部分繼續做長詞優先法斷詞，直到句子的另一端結束為止。一般來說，如果所使用的辭典夠大，長詞優先法斷詞可達到超過 90 % 以上的斷詞準確率。

長詞優先法依照比對方向的不同又可分為兩種不同的變形，第一種是「正向長詞優先法」（Forward Maximum Matching, FMM），即由句子開頭的第一個字元開始，由左而右逐一掃瞄，比對出在辭典中最長的詞，以當作斷詞的結果，並直到句子的結尾而結束。相反地，另一種長詞優先法的變形則是「反向長詞優先法」（Backward Maximum Matching, BMM），由句子的最後一個字元開始掃瞄，從右至左依序比對辭典中的詞，比對到最長的詞當成反向長詞優先法的斷詞結果，並

直到句子的開頭而結束。

此兩種不同的長詞優先斷詞法，當斷詞的結果不同時，則表示發生交集型歧義，如表 1 中的第二個例子：「即將來臨時」字串，因為「將」可與「即」和「來」結合成｛即將、將來｝等不同的詞，因此屬於交集型歧義字串，正向長詞優先法會斷詞成「即將／來臨／時」，而反向長詞優先則斷詞成「即／將來／臨時」。

表 1　長詞優先法的不同變形

例句	**正向長詞優先**	**反向長詞優先**
即將畢業	即將／畢業	即將／畢業
即將來臨時	即將／來臨／時	即／將來／臨時

另外，由於長詞優先法屬於辭典比對式斷詞方法，只有在辭典中的詞才有可能正確斷出，所以無法解決未知詞問題。當遇到未知詞時，正向長詞優先與反向長詞優先都將斷詞成單一中文字元。例如：「鴻海董事長郭台銘」字串，由於辭典中未收錄｛鴻海、郭台銘｝等詞，因此正向長詞優先法與反向長詞優先法都同樣會斷詞成「鴻／海／董事長／郭／台／銘」。

3.2. BIES 分類問題

利用機器學習式演算法來解中文斷詞的問題時，一般的作法是將中文斷詞問題轉換成分類的問題，而最常被使用的方法就是轉換成字元分類問題（character classification problem），將每個字元都給予其對應的類別，透過字元類別來做分類，這些字元的類別由出現在中文詞當中的特定位置來決定，一個字元的位置可以分為位於詞的開始（beginning）、位於詞的中間 （intermediate）、位於詞的結尾（end）以及由單一字元組成的詞（single-character）等四種類別，因此也稱為「BIES 分類問題」。

理論上中文字元可以存在於中文詞的任何位置上，例如表 2 的例子，字元「中」可以存在於詞的開始（B）、詞的中間（I）、詞的結尾（E）、以及單一字元的詞（S）。所以 BIES 分類所要解決的問題也就是決定每個字元的正確類別。

在中文斷詞的問題上，一旦將欲斷詞字串中的所有字元都已分類完成，則也表示已經斷詞完成，例如：「今天是重要的日子」這個中文字串，利用分類問題將找出每個字元所對應的 BIES 標籤，在此例子中，也就是「BESBESBE」，則相當於是已經斷詞出｛今天、是、重要、的、日子｝等詞出來了，因此原來的中文字串便可以轉換成「今天／是／重要／的／日子」的斷詞結果。

表 2　字元「中」可出現在詞的任何位置

B	中醫
I	國民中學
E	集中
S	在　資料庫　中

3.3. 隱藏式馬可夫模型

隱藏式馬可夫模型可以視為一個雙層的隨機序列，包含了隱藏層的狀態序列（state sequence）和可觀察層的觀測序列（observation sequence）。隱藏層是無法直接觀察得到的，但可以從另一個可觀察的觀測序列之隨機過程的集合觀察得出。因此，隱藏式馬可夫模型是一個馬可夫鏈的機率函數，無法直接觀察的隱藏層就是一個有限狀態的馬可夫鏈，其初始的狀態機率分佈以及狀態之間的轉移機率由狀態初始機率向量$\prod$和狀態轉移機率矩陣A 來決定，另外還需定義觀測符號機率矩陣B ，儲存各個觀測符號在不同的狀態下的機率值。

隱藏式馬可夫模型可由 (S, K, N, M, Π, A, B) 等七個元素來表示，底下針對模型相關符號與參數做說明：

- S 表示所有狀態的集合，$S=\{s_1,s_2,...,s_N\}$。
- K 表示所有觀測符號的集合，$K=\{k_1,k_2,...,k_M\}$。
- N 表示模型中所有狀態的個數。
- M 表示模型中所有觀測符號的數目。
- $\Pi=(\pi_i)$ 代表狀態初始的機率向量，$\pi_i=P(q_1=s_i)$，$1\le i\le N$，表示在$t=1$時，狀態為s_i的機率，且需滿足$\sum\pi_i=1$的條件。

- $A=[a_{ij}]$ 代表狀態轉移機率矩陣，$a_{ij}=P(q_{t+1}=s_j \mid q_t=s_i)$，$1\le i,j\le N$，表示從狀態$s_i$到狀態$s_j$的機率，且滿足$a_{ij}\ge 0$ 和 $\sum_{j=1}^{N} a_{ij}=1$

- $B=[b_j(k)]$ 代表觀測符號矩陣，$b_j(k)=P(o_t=v_k \mid q_t=s_j)$，$1\le j\le N$ 和 $1\le k\le M$，表示在狀態為s_j時，觀測符號為v_k的機率，且滿足$\sum_{k=1}^{K} b_j(k)=1$。

給定輸入之觀察序列$O=o_1o_2\cdots o_n$（o_t表示在時間t所對應的觀測符號，且滿足$o_t\in K$）。隱藏式馬可夫模型的目的就是要選出一個對應於觀測序列之最佳的狀態序列$Q=q_1q_2\cdots q_n$（q_t表示在時間t所對應的狀態，且滿足$q_t\in S$），也就是找出$P(Q_1^n \mid O_1^n)$為最大機率值時的狀態序列。

由於在馬可夫基本假設下，第$t+1$的時間狀態只和第t的時間狀態有關，與其他任何以前的時間狀態無關，即$P\{q_{t+1}=s_k \mid q_1,q_2,...,q_t\}=P\{q_{t+1}=s_k \mid q_t\}$，且隨機過程中的機率轉移不隨時間改變，因此$P(Q_1^n \mid O_1^n)$的計算可簡化成：

$$P(Q_1^n,O_1^n)=\prod_{t=1}^{n} P(q_t \mid q_{t-1})P(o_t \mid q_t)=\pi_{q_1}\prod_{t=1}^{n-1} A_{q_t,q_{t+1}}\prod_{t=1}^{n} B_{q_t}(o_t)$$

而取得此最大值的狀態序列Q_1^n，則是使用維特比（Viterbi）演算法計算得到。

另外於訓練過程中，隱藏式馬可夫模型當初所提出來的方法 [19] 是使用非監督式的學習方法（unsupervised approach）做訓練，也就是從未標示狀態的文件中做訓練，因而稱之為「隱藏式」，訓練的方法則是使用 Baum-Welch 演算法做參數的更新。而近年來許多領域都已發展出大量已標示的語料庫（corpus）可供訓練，隱藏式馬可夫模型同樣可以在已標示狀態的文件中來做監督式（supervised approach）訓練 [14]，訓練過程則直接利用最大概似估計法（maximum likelihood estimation）計算出模型參數則此模型，又可稱為「可見式馬可夫模型」（Visible Markov Model, VMM）或「語言模型」（Language Model）等，但絕大部分的研究仍然稱「隱藏式」馬可夫模型。於我們的系統中，我們使用監督式的方法來訓練模型，在本論文中也直接以「隱藏式馬可夫模型」做系統的說明。

3.4. 特製隱藏式馬可夫模型

隱藏式馬可夫模型的特製化（specialization）概念，最早是由 J. D. Kim 等人於 1999 年與 2000 年等兩篇研究 [7, 8] 所提出來的，之後於 2001 年到 2004 年間，A. Molina & F. Pla 等兩位學者，更是將此概念成功的應用到許多不同的領域上，如詞性標示（part-of-speech tagging）[17, 18]、淺層分析（shallow parsing）[15]、詞義消歧（word sense disambiguation）[16] 等問題上。

特製化的過程是指在不修改隱藏式馬可夫模型的訓練以及測試過程的前提下，透過狀態的延伸使得模型增加更多資訊，以提升模型準確率。其主要的作法就是給予一個特製化函式（specialization function），以產生出新的狀態，特製化的過程以底下式子來說明：

$$f(\langle o_i, q_i \rangle) = \langle o_i, q_i \cdot o_i \rangle$$

$\langle o_i, q_i \rangle$代表某個觀測符號以及其對應的狀態，新的狀態符號經過特製化的過程中，由原來的觀測符號加上原來狀態來產生，此特製化的隱藏式馬可夫模型又稱為「特製隱藏式馬可夫模型」（Specialized HMM）。而如果不將所有的觀測符號所對應的狀態都做特製化，而是只在特定的觀測符號下，才做特製化的過程則稱為「詞彙式的隱藏式馬可夫模型」（Lexicalized HMM），此過程屬於特製化過程的一種特例，又被稱為詞彙化（lexicalization），此過程以底下式子來說明：

$$f(\langle o_i, q_i \rangle) = \begin{cases} \langle o_i, q_i . o_i \rangle & \text{if } o_i \in W \\ \langle o_i, q_i \rangle & \text{if } o_i \notin W \end{cases}$$

其中 W 為特製詞（specialized words），只有屬於特製詞的觀測符號才會做特製化處理，而特製詞的選擇又有許多不同的準則來選取。

3.5. M-HMM

在 BIES 分類問題中，由於一個字元可出現在詞的不同位置，而導至所對應的 BIES 標籤不只一個，一旦類別標示錯誤，連帶會使得斷詞結果錯誤。但此種斷詞歧義性在 HMM 模式下，並無特殊處理方式。由於正向長詞優先與反向長詞

優先在做斷詞時，遇到歧義性的句子會產生不同的斷詞結果，因此如能將正向長詞優先與反向長詞優先的資訊同時加入 HMM 模型中，相當於提供歧義性的資訊，並且長詞優先法屬於辭典比對式斷詞法，雖無法直接提供未知詞的資訊，但可間接的調整辭典大小來反應未知詞多寡。

將隱藏式馬可夫模型（HMM）改成 M-HMM 的過程，主要是將正向長詞優先（FMM）與反向長詞優先（BMM）之斷詞結果(即所得的 BIES 標籤)，與原來的「字元」組成的新的觀測符號，延伸為「字元-FMM-BMM」等三個資訊結合而成的觀測序列。表 3 中以一個例子來針對 M-HMM 訓練以及測試過程做個說明，在訓練階段中，原始的觀測符號序列為「研、究、生、命、起、源」，加入了長詞優先法的資訊後，新的觀測符號序列便被轉換成「研-B-B、究-I-E、生-E-B、命-S-E、起-B-B、源-E-E」。這些中文字元旁的 B、I、E、S 標籤即是由正向長詞優先與反向長詞優先法所標示的，因此新的觀測符號種類相當於增加了 16 倍，在此狀態種類並未做改變。

表 3　M-HMM 的例子

	訓練過程		測試過程	
原始句子	研究／生命／起源		結合成分子	
	觀測序列	狀態	觀測序列	狀態
HMM 訓練測試資料	研-B-B	B	結-B-S	?
	究-I-E	E	合-I-B	?
	生-E-B	B	成-E-E	?
	命-S-E	E	分-B-B	?
	起-B-B	B	子-E-E	?
	源-E-E	E		

3.6. Lexicalized M-HMM

這部分為隱藏式馬可夫模型特製化的第二階段，透過第一階段 M-HMM 的過程，將觀測符號延伸之後，此階段以新的觀測符號來做詞彙化，也就是取特定的觀測符號當成特製詞，來做詞彙式(Lexicalized)的特製化過程。此階段的特製化過程描述如下。對於每一個特製詞中的觀測符號 w_i 及其對應狀態為 s_i，則詞彙

化的過程是新增一個狀態「$s_i \cdot w_i$」，而原本的狀態「s_i」仍由其他觀測符號所擁有。此過程也相當於是將訓練資料中屬於特製詞的觀測符號給予新的類別，而使得新的訓練資料不再只有原來的 B、I、E、S 四個類別。

我們以一個例子來做說明，如表 4，假若觀測符號「生-E-B」、「起-B-B」是屬於特製詞，則經過詞彙化的過程之後，觀測符號 「生-E-B」 以及 「起-B-B」所對應的狀態就被轉換成「B-生-E-B」、「B-起-B-B」了。在觀測符號 「生-E-B」中，原來的狀態 B 便被分割成兩個不同的狀態：一個是由觀測符號「生-E-B」所屬的狀態「B-生-E-B」以及其他未分割的觀測符號（如觀測符號「研-B-B」）之狀態「B」。因此在新的訓練資料中，狀態符號被延伸了。

表 4　特製詞集合｛生-E-B, 起-B-B｝做詞彙化產生新的狀態

觀測符號	原來的狀態	新的狀態
研-B-B	B	B
究-I-E	E	E
生-E-B	B	B-生-E-B
命-S-E	E	E
起-B-B	B	B-起-B-B
源-E-E	E	E

此特製化過程也將牽扯到一個問題：由於隱藏式馬可夫模型的三個主要參數都與「狀態符號」有關，因此這階段的特製化過程，將增加隱藏式馬可夫模型的參數大小，因此計算量也就會跟著增加，而且過多的特製詞不見得能一直提升準確率。所以我們必須根據訓練資料來決定特製詞的大小。

特製詞的選擇方式，我們是使用兩種不同的準則（criteria）來選取，說明如下：

- SWF: (the Words with High Frequency)

  取在訓練資料中屬於最高頻率的觀測符號，當成特製詞。

- SEF: (the Words with Tagging Error Frequency)

  取具有高測試錯誤率 (或稱標示錯誤率) 的詞，當成特製詞。

不論是使用 SWF 或是 SEF 準則來選取特製詞，都需要決定一個門檻值（threshold），此門檻值是決定特製詞的大小，我們會於實驗四中找出最佳斷詞效能的門檻值。

4. 實驗

於系統實驗中，我們使用中研院平衡語料庫第 3.1 版，當成我們實驗的資料。此語料庫，共有 575 萬詞，是第一個已斷好的詞並帶有詞類標記的現代漢語語料庫。我們將其中已斷詞的中文文章來當成我們的實驗對象，並用隨機的方式分成兩個部分，取其中的 80% 當作訓練語料，用來訓練隱藏式馬可夫模型。而剩下的 20% 則當成我們系統的測試語料。斷詞的評估方式則是使用準確率（Precision）、召回率（Recall）以及 F-measure 來驗證斷詞效能，分別定義如下：

$$\text{Precision} = \frac{\text{系統正確斷出的詞數}}{\text{系統斷詞的總詞數}}$$

$$\text{Recall} = \frac{\text{系統正確斷出的詞數}}{\text{真正的詞數}}$$

$$\text{F measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}$$

由於我們的系統分成 M-HMM 與 Lexicalized HMM 兩部分，因此在實驗的部分，我們也由此兩部分來做實驗。M-HMM 實驗的部份為實驗一、二、三；而 Lexicalized HMM 實驗的部分則為實驗三與實驗四。

4.1. M-HMM 實驗（實驗一、實驗二）

M-HMM 實驗的部分，主要驗證隱藏式馬可夫模型結合長詞優先法之後，在觀測符號中加入更多資訊之前與加入之後的斷詞效能的比較。由於長詞優先法可以提供斷詞歧義性與未知詞等資訊，因此這部分的實驗，我們是先驗證歧義性的斷詞效能，再驗證未知詞資訊多寡之斷詞效能的比較。

實驗一驗證斷詞歧義性效能，方法是取所有訓練資料與測試資料中的所有詞，來當成長詞優先法所使用的辭典的詞（共有 145,608 個詞），使得在測試過程中不會出現未知詞。表 5 為 M-HMM 解歧義性的斷詞效能。其中實驗的基線（baseline）作法為正向長詞優先法（FMM）、反向長詞優先法（BMM），以及只

使用字元資訊當成觀測符號的隱藏式馬可夫模型（HMM）。除了 M-HMM 的實驗結果之外，同時我們也比較只結合正向長詞優先法資訊（FMM+HMM）以及只結合反向長詞優先法資訊（BMM+HMM）的隱藏式馬可夫模型之斷詞效能。

表 5　實驗一：M-HMM 解歧義性的斷詞效能

	FMM	BMM	HMM	FMM+HMM	BMM+HMM	**FMM+BMM+HMM (M-HMM)**
Recall	0.936	0.939	0.812	0.944	0.947	0.957
Precision	0.956	0.959	0.811	0.962	0.965	0.976
F-measure	0.946	0.949	0.812	0.953	0.956	0.967

實驗顯示，隱藏式馬可夫模型只使用字元的資訊時，其斷詞結果只有 0.81 左右，而加入正向長詞優先法與反向長詞優先法之後，系統的斷詞效能 F-measure 由 0.812 大幅地提升到 0.967，並且斷詞結果也勝過正向長詞優先法與反向長詞優先法等兩種基線作法。因此，實驗結果證明了長詞優先法所提供之歧義性資訊的確可提升隱藏式馬可夫模型的效能。

實驗二主要是驗證長詞優先法所使用的辭典，對 M-HMM 斷詞系統的影響，也就是實驗未知詞的斷詞效能。由於辭典是由訓練資料產生，因此實驗時我們將訓練資料隨機分割成兩部分：訓練集合 1（set 1）以及訓練集合 2（set 2），辭典只由訓練集合 1 來產生，藉由調整訓練資料不同的分割比例，以產生出不同的辭典數量，在相同的測試資料下以驗證各自的斷詞效能。實驗結果如表 6 所示。

表 7　實驗二：M-HMM 解未知詞的斷詞效能

	不含未知詞	**未知詞（實驗二）**				HMM
訓練資料比例 (Set1/Set2)	100/0	80/0	60/20	40/40	20/60	0/80
辭典中的詞數	145,608	132,273	116,428	96,780	69,446	0
Set2 中的未知詞數	0	0	17,418	45,212	103,990	All
測試資料中的未知詞數	0	14,415	17,323	22,524	34,573	All
Recall	0.957	0.946	0.946	0.944	0.941	0.812
Precision	0.976	0.951	0.949	0.945	0.934	0.811
F-measure	0.967	**0.948**	0.948	0.945	0.937	0.812

在此實驗的第一個部分，分割比例為 100/0，相當於實驗一歧義性的效能，而未知詞實驗的部分，共實驗 80/0、60/20、40/40、20/60 等分割比例的結果，由未知詞所佔的比例之不同來驗證斷詞效能，而最後一個部分，分割比例為 0/80，代表完全不從訓練資料中建立辭典，也就是測試資料中所有的詞都屬於未知詞，並且在訓練的過程中完全沒有從正向長詞優先法或反向長詞優先法中得到任何資訊，只依賴字元的資訊做斷詞。

實驗二的結果可得知，隨著增加未知詞的資訊，也就是在減少字典的詞數的情況下，M-HMM 的斷詞效能跟著減低，但是降低的幅度並不大，顯見只要有基本詞彙，即可提升 HMM 斷詞效能，但對於未知詞問題，並不能有所做為，因此我們將於實驗三設計 Mask 的實驗來解決此一問題。

4.2. Mask 實驗（實驗三）

由於實驗二是透過減少訓練資料中的詞，來建立長詞優先法所需之辭典的方法以提供未知詞資訊，但是犧牲了長詞優先法的正確性。因此我們引用 Mask 的作法 [21]，在不犧牲訓練資料的詞的前提下，產生具有未知詞資訊的訓練資料。Mask 的概念是讓訓練過程中也有機會碰到未知詞，也就是仿造測試時真正的情形，其作法如下：

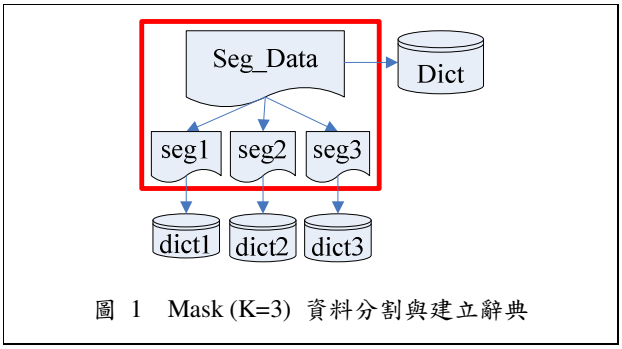

圖 1 Mask (K=3) 資料分割與建立辭典

首先將訓練資料分割成 K 個部分，並且每個部分都建立各自的辭典，因此可產生 K+1 個辭典，如圖 1 所示，此 K+1 個資料便可建立 K+1 個訓練資料。我們先以所有辭典的聯集(Dict=dict1＋dict2+dict3)來標示 M-HMM 所需要的觀測符號，這也相當是原始的訓練資料。接著每次遮住一個部份辭典，也就是產生一個

較小的辭典 Dict-dict(i)，來標示 M-HMM 所需要的觀測符號。在這過程中，有些字詞會因為未知詞的關係，會被錯標成單一字詞 S，但其狀態符號，可以讓 HMM 知道正確的標籤；如果標示結果與原來相同時，則可直接省略，以避免在一個狀態所見到的觀測符號機率不公平的增加，如此重複 K 次最後將此 K+1 個資料形成整個 Mask 的訓練資料。

實驗三為使用 Mask 方法所做的 M-HMM 的實驗，取 Mask K=2 至 K=10 來驗證結果，而 K=1 表示不做分割，也就是沒有使用 Mask 的結果，實驗如圖 2 所示。實驗三結果顯示，使用 Mask 的方法可提供隱藏式馬可夫模型更多未知詞資訊，使得斷詞效能有所提升，並且在 K=2 時，達到最佳的斷詞效能（F-measure = 95.25%）。

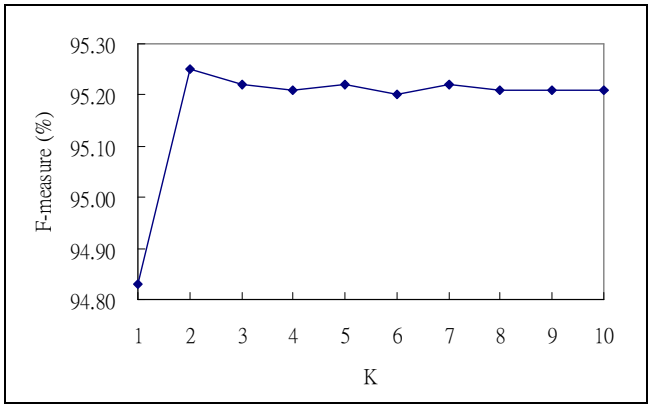

圖 2　實驗三：Mask K=1 至 K=10 的實驗結果

4.3. Lexicalized M-HMM 實驗（實驗四、五）

實驗四是根據 Lexicalized M-HMM 的 SWF 與 SEF 兩種不同的詞彙化策略下，用來調整各自使用的特製詞大小，以找出使得模型能有最佳斷詞效能的門檻值（threshold）。由於這個實驗是用來調整系統用到的特製詞，而不是做斷詞效能的實驗，因此我們只取「訓練資料」來做此實驗。我們將全部訓練資料（佔全部資料 80%）分割成兩部分，依 7 比 1 的比例來分割（分別佔全部資料的 70% 與 10%），其中 70% 的資料（轉換成具有長詞優先法資訊的資料）用來訓練 M-HMM 模型，而剩下的 10% 則當成驗證效能的調整資料（tuning data）。

由於 SWF 為取訓練資料中出現頻率最高的詞當成特製詞，因此我們統計 70%

的資料，取出高頻率的詞做特製詞。而 SEF 為取高測試錯誤率的詞當成特製詞，因此我們先從 70% 的資料建立 M-HMM 模型，並且於調整資料中做測試，根據調整資料中高測試錯誤率的詞做特製詞。取得 SWF 與 SEF 之特製詞後，接著驗證在不同的門檻值下，調整資料的斷詞效能。實驗數據如圖 3 所示。

實驗結果顯示，我們使用 SWF 與 SEF 兩種不同的詞彙化策略，在剛開始取較少的詞當特製詞時，兩者在調整資料下的斷詞效能都有顯著的上升，而 SWF 在取 292 個詞（出現頻率大於 4800 次）時，SEF 取 173 個詞（出現頻率大於 25 次）時，斷詞效能達到最佳結果，並且再繼續隨著特製詞數的增加，斷詞結果便開始往下降，這是因為狀態數增加，使得模型計算量增加而導致準確率下降之緣故。

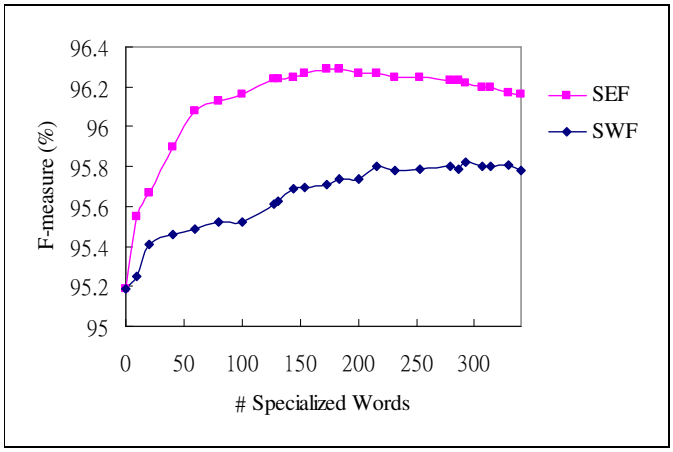

圖 3 實驗四：在不同特製詞大小下，SEF 與 SWF 準則在調整資料下的斷詞效能

實驗五則是測試最佳特製詞結果的 SWF 以及 SEF 準則之 Lexicalized M-HMM 斷詞效能，實驗的設定使用 Mask K=2 之 M-HMM 的設定以及最佳 SWF 與 SEF 特製詞（SWF 為取 292 個詞作為特製詞，而 SEF 則取 173 詞作為特製詞）來做此實驗，並且也與正向長詞優先法（FMM）、反向長詞優先法（BMM）、只使用字元資訊之隱藏式馬可夫模型（HMM）等基線斷詞作法及 M-HMM 的結果作比較，以驗證本系統在狀態延伸前與延伸後的斷詞效能作比較。實驗結果如表 8 所示。實驗結果顯示 Lexicalized M-HMM 不論使用 SWF 或 SEF 準則，其斷詞結果都比 M-HMM 的斷詞效能較好，F-measure 由 0.953 提升到 0.960 與 0.963，

而且使用 SEF 準則與使用 SWF 準則相較之下，SEF 不但特製詞較少且斷詞效能也較好。

表 8　實驗五：Lexicalized M-HMM 的斷詞效能

	FMM	BMM	HMM	M-HMM	**SWF M-HMM**	**SEF M-HMM**
Recall	0.925	0.928	0.812	0.947	0.958	0.963
Precision	0.928	0.930	0.811	0.958	0.962	0.964
F-measure	0.926	0.929	0.812	0.953	**0.960**	**0.963**

5. 結論

在本篇論文中，我們應用隱藏式馬可夫模型之特製化的概念來提升中文斷詞的效能，我們系統的最大的優點，就是完全不需要對隱藏式馬可夫模型的訓練過程以及測試過程做任何修改，只需將訓練資料根據特製化函式來做轉換即可。我們使用兩階段的特製化過程逐步的改良隱藏式馬可夫模型的斷詞效能，在第一階段中結合了長詞優先法的資訊，使得觀測符號增加更多的資訊，於實驗結果顯示，結合長詞優先法在沒有未知詞的情況下，可以大幅地提升隱藏式馬可夫模型的斷詞效能（F-measure: 0.812→0.967），而在有未知詞的情況下，利用 Mask 方式也些微改善斷詞效能（F-measure: 0.948→0.953）。而第二階段使用詞彙式的特製化方式，挑選高錯誤的字元使得狀態增加，實驗也證明能再次提升斷詞效能（F-measure: 0.953→0.963），實驗中發現使用 SEF 準則的結果會比 SWF 準則不但使用的特製詞較小且又能達到更好的斷詞結果。

誌謝

本研究由國科會編號 NSC 94－2213－E－008－020 贊助。

參考文獻

1. M. Asahara, K. Fukuoka, A. Azuma, C. L. Goh, Y. Watanabe, Y. Matsumoto, T. Tsuzuki. “Combination of Machine Learning Methods for Optimum Chinese Word Segmentation,” *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 134–137, 2005

2. M. Asahara, C. L. Goh, X. Wang and Y. Matsumoto. “Combining Segmenter and Chunker for Chinese Word Segmentation,” *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 144–147, 2003
3. K. J. Chen and M. H. Bai. “Unknown Word Detection for Chinese By a Corpus-based Learning Method,” *In Proceedings of ROCLING X*, pp. 159–174, 1997
4. K. J. Chen and S. H. Liu. “Word Identification for Mandarin Chinese Sentences,” *Proceedings COLING '92*, pp. 101-105, 1992
5. K. J. Chen and W. Y. Ma. “Unknown Word Extraction for Chinese Documents,” *In Proceedings of COLING 2002*, pp. 169–175, 2002
6. C. L. Goh, M. Asahara and Y. Matsumoto. “Chinese Word Segmentation by Classification of Characters,” International Journal of Computational Linguistics and Chinese Language Processing Vol. 10, No. 3, pp. 381-396, 2005
7. J. D. Kim, S. Z. Lee and H. C. Rim. “HMM Specialization with Selective Lexicalization,” *In Proceedings of the joinSIGDAT Conference on Empirical Methods in Natural Lan-guage Processing and Very Large Corpora(EMNLP-VLC-99)*, pp. 121-127, 1999
8. S. Z. Lee, J. I. Tsujii and H. C. Rim. “Lexicalized Hidden Markov Models for Part-of-Speech Tagging,” *In Proceedings of 18th International Conference on Computa-tional Linguistics, Saarbrucken, Germany*, pp.481-787, 2000
9. M. Li, J. F. Gao, C. N. Huang and J. F. Li. “Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation,” *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 1–7, 2003
10. Y. Y. Li, C. J. Miao, K. Bontcheva and H. Cunningham. “Perceptron Learning for Chinese Word Segmentation,” *In Proceedings of Fourth SIGHAN Workshop on*

Chinese Language Processing, pp. 154–157, 2005

11. X. Lu. "Towards a Hybrid Model for Chinese Word Segmentation," *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 189–192, 2005
12. X. Luo, M. Sun and B. K. Tsou. "Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information," *In Proceedings of COLING 2002*, pp. 598-604, 2002
13. W. Y. Ma and K. J. Chen. "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 31–38, 2003
14. C. D. Manning and H. Schutze. "Foundation of Statistical Natural Language Processing," Chapter 9-10. pp. 317-380, 1999
15. A. Molina and F. Pla. "Shallow Parsing using Specialized HMMs," Journal of Machine Learning Research 2, pp. 595–613, 2002
16. A. Molina, F. Pla and E. Segarra. "A Hidden Markov Model Approach to Word Sense Disambiguation," *In Proceedings of the VIII Conferencia Iberoamericana de Inteligencia Artificial, IBERAMIA 2002*, pp. 1-9, 2002
17. F. Pla and A. Molina. "Improving Part-of-Speech Tagging using Lexicalized HMMs," Natural Language Engineering, pp. 167-189, 2004
18. F. Pla and A. Molina. "Part-of-Speech Tagging with Lexicalized HMM," *In proceedings of International Conference on Recent Advances in Natural Language Processing(RANLP2001)*, 2001
19. L. R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE, Vol.77, No.22*, pp. 257-286, 1989
20. H. H. Tseng, P. H. Chang, G. Andrew, D. Jurafsky, and C. Manning. "A

Conditional Random Field Word Segmenter for Sighan Bakeoff 2005," *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 168–171, 2005

21. Y. C. Wu, C. H. Chang and Y. S. Lee, "A General and Multi-lingual Phrase Chunking Model Based on Masking Method," Lecture Notes in Computer Science (LNCS): Computational Linguistics and Intelligent Text Processing, Vol. 3878, pp. 144-155, 2006
22. N. Xue. "Chinese Word Segmentation as Character Tagging," International Journal of Computational Linguistics and Chinese, pp. 29–48, 2003
23. N. Xue and L. Shen. "Chinese Word Segmentation as LMR Tagging," *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 176–179, 2003
24. H. P. Zhang, Q. Liu, H. Zhang and X. Q. Cheng. "Automatic Recognition of Chinese Unknown Words Based on Roles Tagging," *In Proceedings of First SIGHAN Workshop on Chinese Language Processing*, pp. 71-77, 2002
25. H. P. Zhang, H. K. Yu, D. Y. Xiong and Q. Liu. "HHMM-based Chinese Lexical Analyzer ICTCLAS," *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 187–187, 2003
26. J. H. Zheng and F. F. Wu. "Study on segmentation of ambiguous phrases with the combinatorial type," Collections of Papers on Computational Linguistics. Tsinghua University Press, Beijing, pp. 129-134, 1999