Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 161-180

結合長詞優先與序列標記之中文斷詞研究

A Simple and Effective Closed Test for Chinese Word Segmentation Based on Sequence Labeling

林千翔*、張嘉惠*、陳貞伶*

Qian-Xiang Lin*, Chia-Hui Chang*, and Chen-Ling Chen*

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，如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。然而當外部資源不易取得時，如何以簡易的方式達到準確的斷詞，則是本研究的目標。在本篇論文中我們以訓練資料所提供的詞彙建構一個辭典，並以長詞優先比對（Maximum Matching）提供正向及反向的斷詞結果做爲應用序列標記之機器學習特徵函數，用以提升隱藏式馬可夫模型（HMM）及條件隨機域（CRF）序列標記的準確率。我們發現，藉由長詞優先比對，得以在完全不修改模型之訓練及測試過程的前提下，透過辭典的遮罩（Mask）及特製化（Specialized）方式，改善斷詞的效能。實驗結果顯示，長詞優先可大幅改善馬可夫模型的斷詞效能（F-measure: 0.812→0.948）；而利用 Mask 方式則可將斷詞效能提升至 0.953;另挑選高錯誤率的字元做爲特製詞，則可再次提升斷詞效能至 0.963。若採用條件隨機域做爲序列標記模型，則僅需透過辭典遮罩，即可將系統斷詞效能提升至 0.963。

關鍵字: 自然語言處理，隱藏式馬可夫模型，中文斷詞，條件隨機域

*國立中央大學資訊工程學系
E-mail: chia@csie.ncu.edu.tw
The author for correspondence is Chia-Hui Chang.

Abstract

In many Chinese text processing tasks, Chinese word segmentation is a vital and required step. Various methods have been proposed to address this problem using machine learning algorithm in previous studies. In order to achieve high performance, many studies used external resources and combined with various machine learning algorithms to help segmentation. The goal of this paper is to construct a simple and effective Chinese word segmentation tool without external resources, that is, a closed test for Chinese word segmentation. We use training data to construct a vocabulary to combine maximum matching word segmentation results with sequence labeling methods including hidden Markov model (HMM) and conditional random fields (CRF). The major idea is to provide machine learning algorithm with ambiguity information via forward and backward maximum matching as well as unknown word information via vocabulary masking. The experimental results show that maximum matching and vocabulary masking can significantly improve the performance of HMM segmentation (F-measure: 0.812 → 0.948 → 0.953). Meanwhile, combining maximum matching with CRF achieves a performance with 0.953 and is improved to 0.963 via vocabulary masking.

Keywords: Chinese Word Segmentation, Maximal Matching, Hidden Markov Model, Conditional Random Field, Vocabulary Masking

1. 序論

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，大致可歸納爲兩點，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爲兩類：

- **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 皆爲辭典中的詞，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 稱爲「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／爲／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，因爲詞xy 是由x 與y等兩個不同的詞所組成，因此xy 稱爲「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，例如應用最大熵分類 Maximum Entropy (MaxEnt) (Xue, 2003)、向量支持機 Support Vector Machine (SVM) (Asahara, *et al*., 2003; Goh, *et al*., 2005)、Transformation-Based Learning Algorithm （TBL） (Lu, 2005)等分類演算法，另外以隱藏式馬可夫模型 Hidden Markov Model （HMM） (Asahara, *et al*., 2005; Lu, 2005; Xue & Shen, 2003; Zhang, *et al*., 2003)的序列標記演算法等等，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 (Asahara, *et al*.,2003; Lu, 2005; Xue & Shen, 2003; Zhang, *et al*., 2003)，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，分別爲擴充「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，爲了擴充觀測符號，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 (Li, *et al*., 2003) 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 (Zheng & Wu, 1999) 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 (Luo, *et al*., 2002) 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。中研院陳克健博士等人於 1997 年開始，提出了三篇關於解決未知詞問題的研究 (Chen & Bai, 1997; Chen & Ma, 2002; Ma & Chen, 2003) ，最早於 1997 的研究 (Chen & Bai, 1997)，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 (Chen & Ma, 2002) ，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爲未知詞。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 (Ma & Chen, 2003) 中，該研究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 (Zhang, *et al.*, 2002) 於 2002 年的研究，則使用類似詞性標示（part-of-speech tagging）的作法，稱爲「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，例如最大熵分類法 Maximum Entropy（ME）(Xue, 2003) 是將斷詞轉成字元分類問題（character classification），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 等人則使用支持向量機 Support Vector Machine（SVM）(Goh, *et al*., 2005) 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）(Li, *et al*., 2005) 的方法做斷詞，該篇研究認爲

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。

另外一類則是以序列標記（sequence labeling）問題來處理中文斷詞，尤其以隱藏式馬可夫模型爲主。不過單獨使用 HMM 本身的效能並不高（約 81%）(Asahara, *et al.*, 2003; Xue & Shen, 2003)，因此這兩篇研究將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 (Asahara, *et al.*, 2003)以及 TBL (Lu, 2005)來當成主要的演算法做斷詞，以達到較佳的斷詞結果。另外，條件隨機域 Conditional Random Fields（CRF）則是 2003 年後廣爲使用的資訊擷取（information extraction）及斷詞（segmentation）方法，如 Massechusset Amherst 大學 A. McCallum, F. Peng, F. Fang 等人在 Rocling 2004 的論文，運用了 24 個中文詞素如姓氏、國名、職稱字首、職稱字尾、地名、日期、單位、動詞、名詞、形容詞等，以及大量的辭典（Vocubulary），得以將中研院平衡語料庫(AS)的 closed test 達到 0.956；而 H. H. Tseng 等人在 Sighan Backoff 2005 的論文，則藉由罕見字的字首及字尾表所建構的特色，解決未知詞的問題，在 2,558,840 的特徵函數下，對中研院平衡語料庫的 closed set 可達到 0.97。

3. 系統架構

我們提出的系統架構如圖 1 所示，主要想法是利用訓練資料中已斷詞的文件（Segmented Texts），建立一個辭典（Vocabulary Construction），再利用長詞優先比對(Maximum Matching)提供正向及反向標記資訊，讓學習模組（Learning Module）得以學習最佳參數；實際斷詞時，即將未斷詞之文章（Unsegmented Texts），同樣利用長詞優先比對，產生與訓練資料相同的測試資料，藉由以訓練好的模型（Model），標記文件並得到斷詞結果（Segmented Texts）。我們使用兩種學習模組，一者爲隱藏馬可夫模型，一者爲條件機率域來解決中文斷詞的問題。

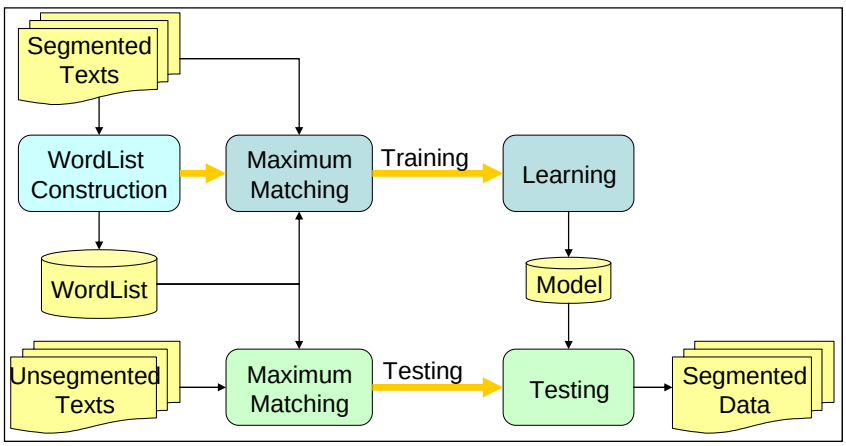

圖 1. 系統學習架構

3.1 BIES分類與序列標記問題

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爲位於詞的開始（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，因此也稱爲「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 1 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 1. 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

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。這也是我們之所以採用長詞優先比對提供 BIES 分類額外資訊的原因。

另外，BIES 字元分類雖然可以藉由前後幾個字的資訊，來完成斷詞，然而決定每個字元是類別是獨立的，因此並無法兼顧前後字元的標記，例如一個字元若被標記爲 B，則其後字元理應被標記爲 I 或 E，而不是 S 和 B，這也是序列標記問題希望能解決的問題。本篇論文採用分別採用隱藏馬可夫模型及線性條件隨機域模型做爲序列標記的演算法。

3.2 長詞優先法

長詞優先法（Maximum Matching Algorithm, MM）是最簡單也最爲廣泛使用的辭典比對式的斷詞方法，其斷詞的策略是由句子的一端開始，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爲止。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爲兩種不同的變形，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 2 中的第二個例子：「即將來臨時」字串，因爲「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 2. 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」。

3.3 隱藏式馬可夫模型

隱藏式馬可夫模型可以視爲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Π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。令 S 表示所有狀態的集合，$S=\{s_1,s_2,\ldots,s_N\}$，N 表示模型中所有狀態的個數，K 表示所有觀測符號的集合，$K=\{k_1,k_2,\ldots,k_M\}$，M 表示模型中所有觀測符號的數

目，則隱藏式馬可夫模型可由三個機率分佈Π, A, B 來描述：

- Π=(π_i)代表狀態初始的機率向量，π_i=P($q_1 = s_i$)，1≤i≤N，表示在 t=1 時，狀態爲 s_i 的機率，且需滿足Σπ_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≤*i,j*≤N，表示從狀態 s_i 到狀態 s_j 的機率，且滿足 a_{ij}≥0 和Σ$_j a_{ij}$=1。
- B=[$b_i(k)$] 代表觀測符號矩陣，$b_i(k)$=P($o_t = v_k \mid q_t = s_i$)，1≤*i*≤N 和 1≤*k*≤M，表示在狀態爲 s_i 時，觀測符號爲 v_k 的機率，且滿足Σ$_k b_i(k)$=1。

給定輸入之觀察序列 $O = o_1 o_2 \cdots o_n$ = (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= $Q = q_1 q_2 \cdots q_n$ (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 $P(Q_1^n \mid O_1^n)$ 爲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：

$$P(Q_1^n \mid O_1^n) = \prod_{t=1}^{n} P(q_t \mid q_{t-1}) P(o_t \mid q_t) = \pi_{q1} \prod_{t=1}^{n-1} A_{q_t, q_{t+1}} \prod_{t=1}^{n} B_{q_t}(o_t) \quad (1)$$

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。

隱藏式馬可夫模型當初所提出來的方法 (Rabiner, 1989) 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（因而稱之爲「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 (Manning & Schutze, 1999) ，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，又可稱爲「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」稱之。

3.3.1 觀測符號的擴充（FB+HMM）

隱藏式馬可夫模型原本的設計是只有單一個觀測符號，將長詞優先比對資訊加入隱藏式馬可夫模型，直接面臨的問題是如何計算多個觀測符號的機率。方法一是分別計算各個符號出現的機率再以觀測符號彼此獨立的假設來計算多個觀測符號的聯合機率；方法二則是直接記錄多個觀測符號的聯合機率；前者節省空間，後者機率估計較準。因此我們採用第二種方式，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，延伸爲「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 FB+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爲「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. FB+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	
HMM 訓練測試資料	觀測序列	狀態	觀測序列	狀態
	研-B-B 究-I-E 生-E-B 命-S-E 起-B-B 源-E-E	B E B E B E	結-B-S 合-I-B 成-E-E 分-B-B 子-E-E	? ? ? ? ?

3.3.2 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 (Kim, *et al*., 1999; Lee, *et al*., 2000) 所提出來的，之後於 2001 年到 2004 年間，A. Molina 及 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）(Pla & Molina, 2001; 2004)、淺層分析（shallow parsing）(Molina & Pla, 2002)、詞義消歧（word sense disambiguation）(Molina, *et al*., 2002) 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），將原來的狀態符號產生出新的狀態符號，特製化的過程以底下式子來說明：

$$f(\langle o_i, q_i \rangle) = \langle o_i, q_i.o_i \rangle \tag{2}$$

$<o_i, q_i>$代表觀測序列中的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，由此觀測符號加上原來狀態來產生新的狀態，經過特製化過程的隱藏式馬可夫模型又稱爲「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進行特製化，而是只針對某些較容易分類錯誤的觀測符號才做特製化，此過程則稱之爲「詞彙式的隱藏式馬可夫模型」（Lexicalized HMM），此過程是屬

於特製化過程的一種特例，也被稱爲詞彙化（lexicalization），正式說來：

$$f(\langle o_i, q_i \rangle) = \begin{cases} \langle o_i, q_i.o_i \rangle & \text{if } o_i \in W \\ \langle o_i, q_i \rangle & \text{if } o_i \notin W \end{cases} \quad (3)$$

其中W 爲特製詞（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。

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

在本篇論文中，透過第一階段將所有的觀測符號做延伸之後，我們進一步的以這些新的觀測符號來做詞彙化，也就是取某些特定的觀測符號來當成特製詞，將其對應的狀態做延伸的過程。舉例來說，如表 4 所示，假設觀測符號「生-E-B」、「起-B-B」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」及「B-起-B-B」。也是多了兩個新的狀態：一個是由觀測符號「生-E-B」所屬的新狀態「B-生-E-B」，以及由觀測符號「起-B-B」所屬的新狀態「B-起-B-B」。因此在新的訓練資料中，狀態符號被延伸了。

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，此兩種不同的準則分別說明如下：

- **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)，此門檻值是決定最合適的特製詞數量，我們會於實驗四中找出最佳斷詞效能的門檻值。

3.4 條件式隨機域

條件隨機域爲一種無向圖(undirected graphical)模型，可被用來估算給予一觀測序列,得到相對應的狀態序列的條件機率分佈。相對於 HMM 以生成模型（generative model）描述觀察序列如何經由狀態轉移及符號產生的過程，CRF 專注於狀態序列在給定觀測序列下的條件機率分佈，屬於一種鑑別式機率模型（discriminative model）。原則上，條件隨機場的圖模型佈局是可以任意給定的，一般常用的佈局是鏈結式的架構，鏈結式架構不論在訓練、推論、或是解碼上，都存在有效率的演算法可供演算。

鏈結式條件隨機域模型（如圖 2 所示）可以定義爲如下的問題：給定一組訓練資料樣本{X_1, X_2, …}，以及其相對的序列標資料{Y_1, Y_2, …}，監督式學習的目標是找到最佳的潛藏函數，使得最大似然度估計（likelihood）Π_i $P(Y_i|X_i)$最大化。

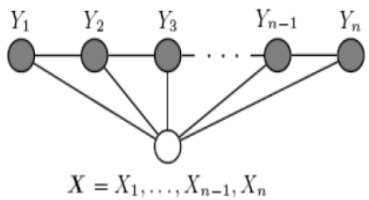

圖 2. 鏈結式條件隨機域模型

令 X^n 爲一長度爲 n 的觀測序列，則狀態序列 Y=y_1, y_2, …, y_n 的條件機率以隨機域表示如下：

$$P(y_1,...,y_n \mid X) = \frac{1}{Z_X}\prod_{i=1}^{n}\psi_i^v(y_i,X)\psi_i^e(y_i,y_{i-1},X) \tag{4}$$

其中 $\psi_i^v(y_i,X)$ 和 $\psi_i^e(y_i,y_{i-1},X)$分別爲觀測序列X在第i個位置節點及邊的潛藏函數值，Z_x則是觀測序列 X 的正規化常數，其目的在使$\Sigma_Y P(Y|X)=1$。大部份序列標記應用採用對數-線性（log-linear）的潛藏函數。

$$\begin{aligned}\psi_i^v(y_i,X) &= \exp\left(\sum_j \mu_j g_j(y_i,X,i)\right)\\ \psi_i^e(y_i,y_{i-1},X) &= \exp\left(\sum_k \lambda_k f_k(y_i,y_{i-1},X,i)\right)\end{aligned} \tag{5}$$

其中 $g_j(y_i,X,i)$及$f_k(y_i,y_{i-1},X,i)$分別爲節點及邊的觀測屬性與標記 y 結合所得的特徵函數（feature function），而μ_j及λ_k分別爲各函數的權重。雖然只是用單一指數模型來描述在給予觀察序列的條件下整個狀態序列的聯合機率，因此在不同的狀態中，不同的特徵函數所賦予的權重可以考慮到狀態彼此的情形。

在本篇論文中，我們採用 CRF++工具來做我們序列標記的工作（表 5 爲 CRF++的輸入範例），並利用其樣版來產生 g_j 及f_k 等特徵函數。CRF++樣板的類型有 Unigram 與 Bigram 兩種，決定是否僅由單一狀態符號或是兩個狀態符號，與觀測符號結合來產生特徵函數。我們主要採用 Unigram 的狀態符號樣版，假設觀測符號共 6030 個字，正向與反

向長詞優先比對各可得 BIES 四種標記，若三種觀測符號均取前後各一位及當前的觀測符號共三個 unigram，兩種長詞優先比對則以目前及前一位組成 bigram，則總共可產生{(6030+4+4)*3+(8+8)*1}*4=72,584 特徵函數，其中第一項爲三種觀測符號的 unigram，第二項爲兩種長詞優先比對的 bigram，最外層的 4 則代表與狀態符號的組合。另外 Bigram 的狀態符號樣版與兩種長詞優先比對觀測符號組成(4+4)*16=128 特徵函數，故一共產生 72,716 個特徵函數。不過扣除一些不可能的 bigram 組成，最後得到 72,280 個特徵。

表 5. CRF++的輸入範例

	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
CRF 訓練測試資料	研 B B 究 I E 生 E B 命 S E 起 B B 源 E E	B E B E B E	結 B S 合 I B 成 E E 分 B B 子 E E	? ? ? ? ?

4. 實驗

我們使用中央研究院平衡語料庫第 3.1 版，當成我們實驗的資料。此語料庫大小共 575 萬詞，不重覆的詞共 145,608 個，是第一個已斷好的詞並帶有詞類標記的現代漢語語料庫。我們取其中已斷詞的文章來當成我們的實驗對象，並且用隨機的方式依四比一的比例分割成兩個部分，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 等三個評估方式來驗證斷詞的效能，分別定義如下：

- $\text{Precision} = \frac{\text{系統正確斷出的詞數}}{\text{系統斷詞的總詞數}}$
- $\text{Recall} = \frac{\text{系統正確斷出的詞數}}{\text{真正的詞數}}$
- $\text{F measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

我們的實驗主要可分三階段來說明：第一階段結合長詞優先之斷詞主要分析辭典大小及未知詞對斷詞效能的影響；而第二階段則針對詞彙式隱式馬可夫模型，調整各策略使用的特製詞大小；最後階段則整體斷詞效能的比較。

4.1 結合長詞優先之斷詞效能

此部份的實驗，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的資訊，同時辭典大小也可控制未知詞的多寡，因此這部分的實驗，我們先實驗辭典大小對斷詞效能的影響，接著實驗沒有未知詞影響的情形下，序列標記演算法能解決多少斷詞歧義性的問題。

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，藉由調整兩個集合不同的分割比例，以產生出不同的數量的辭典，並對同一份測試資料下，驗證各自的斷詞效能。實驗結果如表 6 所示。

表 6. 辭典大小對斷詞效能的影響

	無未知詞	不同辭典大小				HMM
訓練資料比例(Set1/Set2)	100/0	80/0	60/20	40/40	20/60	0/80
辭典(Set 1)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
FB+HMM Recall	0.957	0.946	0.946	0.944	0.941	0.812
FB+HMM Precision	0.976	0.951	0.949	0.945	0.934	0.811
FB+HMM F-measure	**0.967**	**0.948**	**0.948**	**0.945**	**0.937**	**0.812**
BMM F-measure	**0.949**	**0.929**	**0.926**	**0.921**	**0.912**	**0.427**

表 6 的第二欄分割比例爲 100/0，相當於沒有未知詞情況下的效能；而第三～六欄使用不同比例的辭典大小，包括 80/0、60/20、40/40、20/60 等分割比例的結果，反應未知詞所佔的比例之不同時斷詞效能的表現；而最後一欄分割比例爲 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。最後一列則是反向長詞優先的斷詞效能。

實驗結果顯示，在減少辭典的詞數的情況下，FB+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，並不能有所做爲，因此我們設計 Mask 的實驗來解決此一問題。

4.1.1 Mask模擬未知詞實驗

上述實驗在減少長詞優先法所需之辭典詞數的做法，某種程度提供了訓練資料中可能會預見未知詞的情形，但是不免犧牲了長詞優先法的正確性。因此我們引用 Mask 的作法

(Wu, Chang, & Lee, 2006)，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。辭典 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，其作法是將訓練資料分割成 K 個部分 $S_1, S_2, \cdots, S_K$，，並且每個部分都建立各自的辭典 $D_1, D_2, \cdots, D_K$，令 D 爲各辭典的聯集（$D=\cup_i D_i$），則共可產生 K+1 個辭典。接著對每一部份的訓練資料 S_i 以總辭典 D 減去 D_i，做爲長詞優先比對的辭典，用來標示 FB+HMM 所需要的觀測符號。在這過程中，訓練資料 S_i 中有些字詞會因爲未知詞的關係，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，加上原先以總辭典 D 對所有訓練資料所做的標記，將此 K+1 個資料形成整個 Mask 的訓練資料。圖 3 爲 K＝3 的示意圖。

我們取 K=2 至 K=10 來檢試 Mask 的效能，實驗結果如圖 4 所示，其中 K=1 表示不做分割，也就是沒有使用 Mask 的結果。實驗結果顯示，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%）。

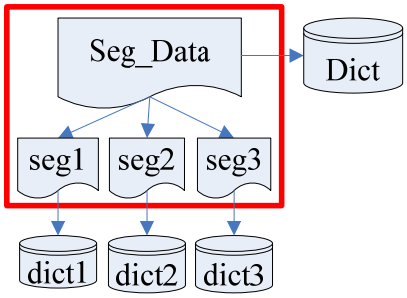

圖 3. Mask(K=3) 資料分割與建立辭典

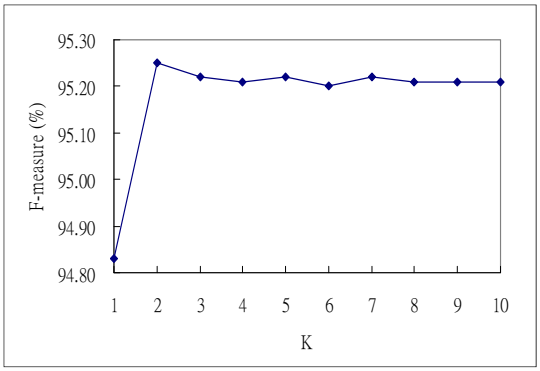

圖 4. Mask K=1 至 K=10 的實驗結果

4.1.2 斷詞歧義性的影響

接著我們針對沒有未知詞的情況下，檢視序列標記模型能解決多少歧義性問題。我們的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，由於實際的情況下一定會遇到未知詞問題，因此這部分的實驗屬於「完美情況」的實驗。其中實驗的基準（baseline）爲正向長詞優先法（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）。除了我們的系統 FB+HMM，即同時結合正向長詞優先法與反向長詞優先法資訊的實驗結果之外，我們也比較只結合正向長詞優先法資訊（F+HMM）以及只結合反向長詞優先法資訊（B+HMM）的隱藏式馬可夫模型之斷詞效能。表 7 爲實驗在無未知詞狀態下的斷詞效能。

如表 7 所示，隱藏式馬可夫模型若只有使用字元的資訊時，其結果不論是召回率、準確率或 F-measure 都只有 0.81 左右，而加入正向長詞優先法與反向長詞優先法的資訊之後，系統的斷詞效能 F-measure 便由 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準作法。而 CRF 在使用正反向長詞優先比對所提供的歧義性資訊，得到最高的斷詞效能 0.977，顯示鑑別式機率模型較生成式機率模型有更好的預測表現 。

表 7. 無未知詞狀態下的斷詞效能

	FMM	BMM	HMM	F+HMM	B+HMM	FB+HMM	FB+CRF
Recall	0.936	0.939	0.812	0.944	0.947	0.957	0.981
Precision	0.956	0.959	0.811	0.962	0.965	0.976	0.974
F-measure	0.946	0.949	0.812	0.953	0.956	0.967	0.977

4.2 詞彙化隱藏馬可夫模型

第二階段則是 Lexicalized HMM，係根據 SWF 與 SEF 兩種不同的詞彙化策略，來調整各策略使用的特製詞大小，目的是去找出使得模型能有最佳斷詞效能的特製詞數量的門檻值（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，因此這個實驗我們只取「訓練資料」來做實驗。爲了驗證這個部分的效能，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 FB＋HMM 模型，而剩下的 10% 則當成驗證效能的調整資料（validation set）。

由於 SWF 爲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% 的訓練資料，取出高頻率的詞做特製詞。而 SEF 爲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 FB＋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 5 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成特製詞時，兩者在調整資料下的斷詞效能都有顯著的上升，而當 SWF 準則取 292 個詞（出現頻率大於 4800 次）時，SEF 準則取 173 個詞（出現頻率大於 25 次）時，做詞彙化的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，這是因爲狀態增加，使得模型參數增加而導致準確率下降的緣故。

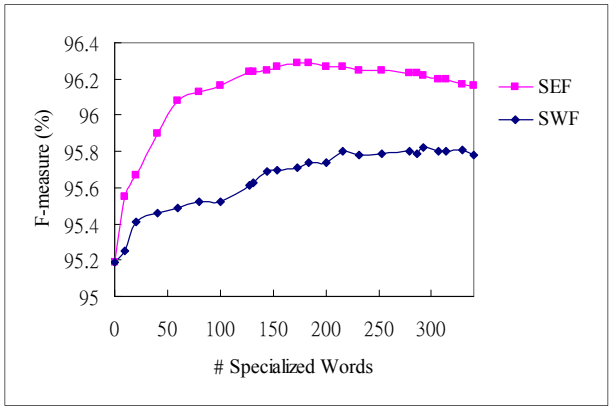

圖 5. 10% 驗證資料中 SEF 與 SWF 在不同特製詞大小下的斷詞效能

4.3 整體效能比較

最後我們比較不同學習模型在一般有未知詞情況下的斷詞效能，我們以正向長詞優先法（FMM）、反向長詞優先法（BMM）、以及只使用字元資訊之隱藏式馬可夫模型（HMM）等斷詞作法作基準，先與 FB+HMM 以及 FB+CRF 的結果做比較（如表 7 所示）。再取 Mask K=2 之最佳設定，產生訓練檔供 FB+HMM 及 FB+CRF 建立模型，同時針對特製最佳 SWF 與 SEF 特製詞（SWF 爲取 292 個詞作爲特製詞，而 SEF 則取 173 詞作爲特製詞）來實驗，以驗證 FB+HMM 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。

表 8. 有未知詞情況下 Mask 與 Specialized 效能比較

	Mask+FB+HMM	Mask +FB+CRF	SWF	SEF
Recall	0.947	0.966	0.958	0.963
Precision	0.958	0.961	0.962	0.964
F-measure	0.953	0.963	0.960	0.963

實驗結果顯示，應用長詞優先法比對可以有效的改進序列標記的效能。而在馬可夫 HMM 模型下而採用 Mask 方式產生訓練資料，可將 F-measure 由 0.948 提升至 0.953；若是 採用條件隨機域 CRF 模型，則更可將 F-measure 由 0.959 提升至 0.963。另外，不論使用 SWF 或 SEF 準則，詞彙化後的斷詞結果也可有效地將 F-measure 由 0.953 提升到 0.960 與 0.963 的結果，而且此兩種不同準則在相較之下，SEF 不但使用的特製詞數較少

且斷詞效能也較好。

表 9. 各斷詞模型的效率及使用成本分析

	HMM	FB+HMM	SEF=143	FB+CRF
Training Time	52 秒	1 分 9 秒	1 分 22 秒	37 分
Testing Time	8 秒	9 秒	58 分 6 秒	20 秒
Space	21MB	28MB	303MB	582KB

表 9 則是各種模型訓練及測試所需的時間、空間以及模型的大小。就訓練時間來說，我們可以看到 HMM 各種衍伸模型的訓練時間相較 CRF 都是相當快（65 倍）。其中 Specialized HMM 由於模型大小及測試時間遠大過其他模型，在資源有限的環境下並不實用；而 CRF 雖然訓練及測試的時間較長，但離線訓練所花的時間通常不是主要缺點，而測試時間也在合理範圍內（2 倍），在效能的考量下是實用上較好的選擇。

5. 結論

在本篇論文中，我們考慮僅從訓練語料中，如何建立一個不藉由外部資源的中文斷詞模型。首先我們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，結合長詞優先法可以大幅地提升馬可夫模型的斷詞效能（F-measure: 0.812→0.948）；而利用 Mask 方式也可進一步改善斷詞效能（F-measure: 0.948→0.953）；另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）。若採用條件隨機率學習模組，在同樣使用正反向長詞優先比對所提供的歧義性資訊情況下，CRF 則提供比 HMM 更好的斷詞效能，F-measure 可由 0.948 提升至 0.959；若是藉由 Mask 的模擬測試，則可再將斷詞效能提升至 0.963，顯示鑑別式機率模型較生成式機率模型有更好的預測表現。因此雖然 CRF 訓練時間較久，在模型大小差異不大、但效能較佳的情形下，CRF 反而是實用上會是更好的選擇。本篇論文的效能雖未超越 H.H. Tseng 等人所達到之 0.97，但使用的特徵值相較的減少很多（1/100），以此爲基礎，加入其他特色將有機會推進中文斷詞的效能。

參考文獻

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

Asahara, M., Goh, C. L., Wang, X., & Matsumoto, Y. (2003). Combining Segmenter and Chunker for Chinese Word Segmentation. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 144-147.

Chen, K. J., & Bai, M. H. (1997). Unknown Word Detection for Chinese By a Corpus-based Learning Method. *In Proceedings of ROCLING X*, 159-174.

Chen, K. J., & Liu, S. H. (1992). Word Identification for Mandarin Chinese Sentences. *Proceedings COLING '92*, 101-105.

Chen, K. J., & Ma, W. Y. (2002). Unknown Word Extraction for Chinese Documents. *In Proceedings of COLING 2002*, 169-175.

Goh, C. L., Asahara, M., & Matsumoto, Y. (2005). Chinese Word Segmentation by Classification of Characters. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(3), 381-396.

Kim, J. D., Lee, S. Z., & Rim, H. C. (1999). 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)*, 121-127.

Kudo, T. CRF++ 0.57: Yet Another CRF toolkit. Available from http://crfpp.googlecode.com/svn/trunk/doc/index.html

Lee, S. Z., Tsujii, J. I., & Rim, H. C. (2000). Lexicalized Hidden Markov Models for Part-of-Speech Tagging. *In Proceedings of 18th International Conference on Computa-tional Linguistics, Saarbrucken, Germany*, 481-787.

Li, M., Gao, J. F., Huang, C. N., & Li, J. F. (2003). Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 1-7.

Li, Y. Y., Miao, C. J., Bontcheva, K., & Cunningham, H. (2005). Perceptron Learning for Chinese Word Segmentation. *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 154-157.

Lu, X. (2005). Towards a Hybrid Model for Chinese Word Segmentation. *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 189-192.

Luo, X., Sun, M., & Tsou, B. K. (2002). Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information. *In Proceedings of COLING 2002*, 598-604.

Ma, W. Y., & Chen, K. J. (2003). A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 31-38.

Manning, C. D., & Schutze, H. (1999). Foundation of Statistical Natural Language Processing. Chapter 9-10. 317-380.

Peng, F., Feng, F., & McCallum, A. (2004). Chinese Segmentation and New Word Detection using Conditional Random Fields. *In Proceedings of International Conference on Computational Linguistic* (COLING), 562- 568.

Molina, A., & Pla, F. (2002). Shallow Parsing using Specialized HMMs. *Journal of Machine Learning Research 2*, 595-613.

Molina, A., Pla, F., & Segarra, E. (2002). A Hidden Markov Model Approach to Word Sense Disambiguation. *In Proceedings of the VIII Conferencia Iberoamericana de Inteligencia Artificial (IBERAMIA)*, 1-9.

Pla, F., & Molina, A. (2004). Improving Part-of-Speech Tagging using Lexicalized HMMs. *Natural Language Engineering*, 167-189.

Pla, F., & Molina, A. (2001). Part-of-Speech Tagging with Lexicalized HMM. *In proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE,* 77(22), 257-286.

Tseng, H. H., Chang, P. H., Andrew, G., Jurafsky, D., & Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 168-171.

Wu, Y. C., Chang, C. H., & Lee, Y.S. (2006). 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, 144-155.

Xue, N. (2003). Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese*, 29-48.

Xue, N., & Shen, L. (2003). Chinese Word Segmentation as LMR Tagging. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 176-179.

Zhang, H. P., Liu, Q., Zhang, H., & Cheng, X. Q. (2002). Automatic Recognition of Chinese Unknown Words Based on Roles Tagging. *In Proceedings of First SIGHAN Workshop on Chinese Language Processing*, 71-77.

Zheng, J. H., & Wu, F. F. (1999). *Study on segmentation of ambiguous phrases with the combinatorial type*. Collections of Papers on Computational Linguistics. Tsinghua University Press, Beijing, 129-134.