以 HMM 作中文詞性之自動標注
Chinese Part-of-Speech Tagging using an HMM

張照煌　　陳正德
Chao-Huang Chang　Cheng-Der Chen

工業技術研究院 電腦與通訊工業研究所 尖端資訊技術中心
Computer & Communication Lab., Industrial Technology Research Institute

Abstract

We address the problem of Chinese part-of-speech tagging based on a first-order, fully-connected hidden Markov model. Part of the 1991 United Daily corpus of approximately 10 million Chinese characters is used for training and testing. A news article is first segmented into clauses, then into words by a Viterbi-based word identification system. The (untagged) segmented corpus is then used to train the HMM for tagging using the Baum-Welch reestimation procedure. A Viterbi decoder based on trained HMM parameters can tag an unsegmented Chinese sentence automatically. Experimental results show that the system is able to correctly tag approximately 96% of the words in the testing data.

摘　要

許多智慧型中文自然語言的應用，如中文語音識別、語音合成、全文檢索、及機器翻譯等，須要利用到標注中文詞詞性的中文語料庫 (Corpus)。然而語料庫動輒數十萬甚至數百數千萬詞，利用人工標注需要動用大量人力費時經年，標注結果又不一致，亟須借助電腦自動標注。本文介紹我們以一階互聯隱藏式馬可夫模式 (Hidden Markov Model, 簡稱HMM) 爲基礎製作中文詞性自動標注系統之理論基礎、系統設計、及實驗結果。

1. 引　言

語言學家把中文詞的用法根據語法和語意特性加以詳細分類，即所謂的「詞性」(Part-of-Speech)。最簡單的分類即如英文的八大詞類 (名詞、動詞、形容詞等)。但是在電腦自然語言處理技術上，分類少則數十種，也有多達數百種者。每一分類給予一個代號，稱爲詞性標記 (Part-of-Speech Tag)。標注中文詞詞性的中文語料庫 (Corpus) 可應用於許多智慧型中文自然語言的領域，如中文語音識別、語音合成、文字識別、全文檢索、及機器翻譯等。

由於人類所使用的語言常有一個詞多種用法的現象，所以許多詞的詞性標記並非唯一。所以語料庫的詞性標注必須將各詞根據其用法給予正確的標記，是件需要運用智慧判斷而難以完全自動化的工作。然而語料庫動輒數十萬甚至數百數千萬詞，利用人工標注需要動用大量人力費時經年，標注結果又不一致，亟須借助電腦自動標注。對於西方語言(如英文) 詞性自動標注的研究由來已久[2-6,8]. 然而，中文詞性自動標注的技術發展則尚在萌芽階段，原因可大致分析如下:

1. 中文「詞」的定義莫衷一是，沒有定論；例如「第一封信」究竟是一、二、三、或四個詞，另如「洗個澡」、「原委會」等詞仍有許多值得討論的問題；
2. 中文詞性標注和斷詞密切相關，真正高正確率的自動斷詞程式，由於未知詞、省略詞、使用詞庫等問題，仍不易達成，致使人工校正難以免除；
3. 中文詞性標記集如何選定，標記數的多寡，分類的粗細不一，難以決定；
4. 含完整標記訊息的中文詞庫並不可得；
5. 就算以人工標注，也是件很困難的工作；有許多詞的詞性即使語言學家也難以決定或仍有爭議；
6. 英文有現成由人工標注好的語料庫，如布朗 (Brown) 語料庫及 LOB 語料庫，可供開發系統之訓練或驗證，中文的相對語料庫則付之闕如。

這些問題糾纏在一起使得中文詞性自動標注，比起英文者，倍加困難。

本文嘗試解決其中一些問題，以一階互聯隱藏式馬可夫模式 (Hidden Markov Model, HMM) 爲基礎製作中文詞性自動標注系統，探討其理論基礎、系統設計、並提出實驗所得結果。這在所知文獻中，是相當新穎的一個嘗試。

2. 以 HMM 作詞性自動標注

隱藏式馬可夫模式[7] 在語音辨識技術中有著廣泛的應用。Kupiec [4] 提出以 HMM 爲基礎製作的英文及法文詞性自動標注系統；其最大的優點在於可以用未加標注的語料訓練其參數。除了基本的一階互聯隱藏式馬可夫模式之外，Kupiec 的標注系統有兩個新創特色：(1) 詞等值群 (Word Equivalence Class) 的觀念：一個「詞等值群」由擁有相同可能詞性標記的所有詞組成。例如，英文詞 "type" 和 "store" 同屬 *noun-or-verb* 這個詞等值群。此一觀念不僅大大地減少模式的參數個數，更使標注系統可以適應各種不同的操作環境 (Robust)。我們的中文標注系統也引用此一觀念。(2) 預先設定的附加網路：根據對標注錯誤的分析及語言特性的考慮，在一階模式上附加預先設定的網路。由於其實驗顯示預設網路效果有限 (整體錯誤率相差 0.2%)，在我們的系統並未採納。

以下說明如何利用 HMM 來自動標注詞性：一個基本的一階互聯 HMM 有 N 個狀態 (state)、M 種可能觀察 (observation) 則有三組參數：二維矩陣 A (NxN) 爲狀態轉換機率分佈 (state transition probability distribution)、二維矩陣 B (NxM) 爲觀察機率分佈 (observation probability distribution)、一維陣列 P (N) 爲最初狀態分佈 (initial state distribution)。對於一長度爲 T 的觀察串列 O，基於 A、B、P，有一些演算法，例如 Viterbi 最佳路徑搜尋法，以求得隱藏狀態串列 I。

對詞性自動標注問題而言，N 即是所用標記集的標記總數，M 則爲詞或詞等值群的總數。中文有十萬以上的詞而詞等值群的總數小於一千。簡單計算可知，等值群這個觀念可以把 B 減少一百倍以上，大量節省參數所須空間和訓練所須的語料及時間。如此「詞性自動標注」問題可歸納爲：已知一詞列 (觀察串列 O)，如何求其正確的詞性序列 (隱藏狀態串列 I)。

2.1 詞性標記集

我們所定義的詞性標記集 (tag set) 含 46 個一般標記，及 11 個特殊標記。一般標記包括 A0 (形容詞)，C0-C1 (連接詞)，D0-D2 (代詞)，I0 (感嘆詞)，M0 (量詞)，N0-N9 (名詞)，P0 (介系詞)，R0-R6 (助詞)，T0 (語尾詞)，U0-U4 (數詞)，V0-V4 (動詞)，X0 (象聲詞)，Y0-Y4 (複合詞)，Z0-Z2 (副詞)。特殊標記則爲標點符號(PAR, SEN, PCT, DUN, COM, SEM, COL)，未知詞(UNK)，外來語(ABC)，及

2.5 詞庫

系統中的中文通用詞庫含約 80,000 詞項，一詞項由組成該詞的漢字和其 EQC-id 構成。詞庫的原始資料來自電通所和中央研究院的合作：電通所收集需要的詞、詞音、及詞頻，而中央研究院提供語法及語意標記。本系統所須只用到詞和語法標記 (即詞性)。而上文也說明到詞性部分經過簡化和重組。

3. 實 驗 結 果

整個詞性自動標注系統，包括斷詞器、EQC-id轉碼器、 HMM 訓練模組、及 Viterbi 解碼模組，均以 C 語言在 Sun Sparc工作站製作完成，初步測試結果討論如下。

在已完成的實驗中，我們所利用的語料庫有二：(1) **Corpus1** (地方新聞，共 1,418 句，12,284 詞) 經過「語料庫之準備」各步驟處理，即已斷詞並標注詞性的語料庫，大部分實驗結果都是以此語料庫進行測試。(2) **Corpus3** (美聯社訊，共 3,784 句，35,849 詞)。

目前我們的系統有 338 個「詞等值群」：(1) 最常用的一百個混淆詞 (ambiguous word) 各自獨立成爲一「詞等值群」。(2) 其餘 238 個詞等值群則依定義分別由擁有相同可能詞性標記的所有詞組成。

測試時用到兩類詞庫，以資比較：(1)「通用詞庫」：即前述的電通所 80,000 詞通用詞庫；(2)「密閉詞庫」：利用測試語料 **Corpus1** 的詞及所標注詞性整理而得的詞庫。「密閉詞庫」可用以測試系統的上限。測試可分爲「內測試」(測試語料和訓練語料相同) 和「外測試」(測試語料必須不同於訓練語料) 兩種。此處限於篇幅，我們直接報告較有意義的「外測試」實驗結果。

3.1 外測試，通用詞庫

表1 顯示以通用詞庫進行之外測試 (outside test) 的結果－以正確率表示。正確率的定義很簡單：以標注正確的詞數除以總詞數即得。由於目前系統未處理詞庫中未記載的詞 (一律標爲未知詞 UNK)，若去除未知詞後再計算，則正確率可視爲系統的上限，由表中可知，約有超過半數的錯誤是由未知詞造成。另一種算法是只考慮混淆詞 (即多詞性的詞)，這是系統真正消除混淆的能力指標。根據我們的詞性標記集和詞庫，**Corpus1** 語料庫的 12,284 詞中，約有百分之卅五的詞是混淆詞。

爲了進行外測試，我們將語料庫分成兩部分：其一用以訓練，另一則用以測試。表之前兩行 (訓練、測試) 即分別爲訓練、測試語料的句數 (注意：並非詞數或字數)。測試結果顯示，和內測試比較，正確率如預期地下降了：已知詞下降了兩個百分點，混淆詞下降了五個百分點。一般而言，我們的系統可以正確地標注約九成二的已知詞或八成的混淆詞。

在最後一列中，我們利用一個較大的語料庫 **Corpus3** 訓練，而以整個 **Corpus1** 測試。由於文章種類造成用詞的差異，混淆詞正確率又下降了三個百分點。不過已知詞的正確率仍有百分之九十一・八三。這顯示了應用「詞等值群」這個觀念所帶來對環境 (文章種類) 的強固性 (Robustness)。

訓練	測試	所有詞	已知詞	混淆詞
800	618	85.80%	92.37%	78.16%
1,000	418	86.58%	92.83%	79.95%
1,200	218	86.90%	92.16%	79.40%
3,784	1,418	85.14%	91.83%	76.40%

表1 正確率 (外測試，通用詞庫)

訓練	測試	所有詞	已知詞	混淆詞
800	618	96.01%	96.01%	80.24%
1,000	418	96.20%	96.20%	82.27%
1,200	218	95.41%	95.41%	79.91%

表2 正確率 (外測試，密閉詞庫)

合成之數目(NUM, ARA)。基本上，這個詞性標記集是由台北中央研究院詞知識庫的分類[10] 簡化並重組而成。原來的分類細分至五個層次，以人工標注都嫌太細而不易使用。孫[9] 提出了一個三個層次的詞性標記集 TUCWS，計含 120 個詞性標記，供斷詞系統使用。不過，他們係以人工標注語料庫，而非自動標注，是以該標記集是否適用，須進一步探討。

2.2 語料庫之準備

所使用的 1991 年台灣聯合報語料庫，超過一千萬字，包括該年一到三月間約二十天的報紙文章內容。然而該語料庫由未經斷詞的原始檔案組成，所以須經斷詞之後，才能用以訓練或測試所設計的詞性標注系統。語料庫之準備可分爲如下幾個步驟：

1. 前處理： 清除輸入文章中不適用的部分，如標題、記者資料、圖、表等等。若文章大部分均爲不適用的部分，則刪除之。
2. 斷句： 把一篇文章分割成句子或子句；斷句之標點符號爲「,」「。」「?」「;」「:」等。
3. 自動斷詞： 利用基於詞庫以 Viterbi 演算法製作的自動斷詞程式將句子分割成詞串。
4. 人工修正： (不一定要) 以人工檢查斷詞結果，以修正斷錯的詞。大部分的錯誤是由於未知詞 (詞庫中未記載的詞，如人名地名) 的影響，也有因斷詞作法的缺陷者。此步驟並非必須，但對訓練語料則有相當幫助。
5. 詞等值群查表： 根據詞庫及詞等值群表，把詞串轉換成詞等值群代號 (EQC-id)。

經由以上步驟，一篇文章可轉換成一序列的 EQC-id 串列。

人工標注整個語料庫可能需要耗費幾百個人年，並不可行。然而，如 Merialdo [6]指出，標注過的語料庫對 HMM 參數初值的設定很有幫助，且爲評估系統所必須。因此我們以下列步驟標注一個較小規模的語料庫：(1) 以此較小規模的 (尚未標注) 語料庫訓練系統的 HMM 參數；(2) 利用訓練過的 HMM 標注該語料庫；(3) 以人工檢查並修正標注的結果。

2.3 模式參數之訓練

利用 Baum-Welch 重估計 (reestimation) 程序以及由 EQC-id 組成的未標注語料庫，即可訓練標注系統的 HMM 參數；由於語料庫係由多個句子組成，我們必須利用多觀察序列 (multiple observation sequence) 的再估計公式[7]。根據此訓練程序的特性，經由訓練後的 HMM 參數，可以最準確地預測訓練語料的最可能詞性標記序列。

2.4 自動標注詞性

經由訓練後的 HMM，即可利用標準的 HMM 解碼演算法 (如 Viterbi)，進行新語料的詞性自動標注。以一個句子爲單位，其詞性標注程序如下：

1. 自動斷詞： 利用上述自動斷詞程式將句子分割成詞串。
2. 詞等值群查表： 利用上述詞等值群查表轉碼程式，將詞串轉換成詞等值群代號 (EQC-id)。
3. Viterbi 解碼： 以轉得的 EQC-id 序列爲輸入 (或稱觀察)，利用Viterbi 解碼演算法以求得其最可能隱藏狀態序列，亦即其詞性標記序列。
4. 比對式後處理： 一階互斥隱藏式馬可夫模式並不足以描述預測詞性的局部限制。但是，更高階的 HMM 所須的參數數目呈指數成長，須用的空間，時間，及訓練用語料均大幅度增加。所以，Kupiec [4] 提議以預設附加網路來補充描述語言特性。但效果並不理想，且破壞了 HMM 整體的一致及優美性。我們則設計了一套「比對式後處理」作法，利用標注錯誤之原因分析，整理出一些規則，作爲樣式比對的基礎。以預設的樣式比對輸入的 EQC-id 序列，如果符合，就執行對應的修正動作。

3.2 外測試，密閉詞庫

表2 總結我們的系統利用「密閉詞庫」進行之外測試的正確率：對已知詞（所有詞均爲已知）正確率達百分之九十六，就混淆詞而言，則約爲八成。此正確率和 Kupiec 的英文標注系統（以 HMM 標注 Brown 語料庫）效果相當，毫不遜色。

4. 標注錯誤之原因分析

4.1 互淆矩陣

在辨識問題的討論，常以互淆矩陣 (Confusion Matrix) 來分析。互淆矩陣列出詞性 A 被認成詞性 B 的次數；可以看出兩種詞性互相混淆的程度。經由對一次內測試實驗結果之互淆矩陣仔細分析，我們可以發現以下幾個嚴重的問題。

【ANVZ問題】：中文的詞性沒有明顯的詞尾變化，所以一個詞雖然可能有許多不同的詞性，卻常常只有一種表示法，即其「原型」。因此，即使人工標注也常常碰到許多難以區分的情況。在英文中，常可以根據詞尾來區分一個詞的詞性：例如，-ing 表示現在分詞或動名詞，-ly 表示副詞，-tion 表示名詞，-en 表示過去分詞等等。中文則不然，如「分散」一詞，根據不同的上下文，可以是動詞 (V0) 'distribute'，名詞 (N0) 'distribution'，形容詞 (A0) 'distributive' 'distributing' 'distributed'，副詞 (Z0) 'distributively'。其中尤以動詞和名詞更易混淆；所以標錯的情況時常發生，例如 V0 常被標成 N0，而 N0 也常被標成 V0。

【RP問題】：如動詞和名詞等由於無法窮舉，稱爲開放詞，反之，如連接詞等，則稱爲密閉詞。各個詞性的詞數相差甚爲懸殊。通常，這對標注系統並不會造成問題。然而，在我們的詞性標記集裡，R5 (前級時貌) 只有三個詞：「在」(P0|R5|V0)，「將」(P0|R5|Z0)，及「正」(A0|R5|Z0)。前二者也是常見的介詞 (P0)。實驗發現，雖然「在」大部分均用爲介詞，卻常被標爲 R5。經仔細研究其訓練後的參數 A, B, P，發現在 B 矩陣中 R5 的機率，由於只有三個詞，值均相當大（在：0.683，將：0.227），而相對地 P0 的機率值均小。（由於機率的特性使然，各狀態，如 P0 或 R5，的類察機率和須爲一。）此外，A 矩陣值，由於是未監督學習 (unsupervised learning)，P0 和 R5 並沒有明顯的差異：對訓練而言，每一個「在」都有可能是 P0 或 R5。這是「未監督學習」可用未標注語料訓練所付出的相對代價。

4.2 幾種常見的錯誤樣式

標注錯誤常常成群出現；一個錯誤經常會造成相鄰混淆詞的標注錯誤。幾種常見的錯誤樣式 (Error Patterns) 如下：V0-V0 (被標成N0-N0)，Z0-V0 (A0-N0)，V0-N0 (C1-Z2)，V0-P0 (N0-R5)，P0-N0 (R5-V0)，P0-N1 (R5-V4)，N0-V0-N0 (U1-C1-Z2)。可大致分成下類三種類型：(1) ANVZ型：由上述 ANVZ 問題造成；屬合理的標注錯誤。(2) RP型：含 R5 的幾種錯誤樣式是由 RP 問題造成；可用後處理去除此類錯誤。(3) 慣用語型：有一些中文的慣用表示法由多個混淆詞組成。例如，『以 ... 爲 準』中的三個詞均爲混淆詞：以 (C1 N3 P0)，爲 (C1 P0 V0)，準 (A0 N0 Z2)。這也是 V0-N0 常被標成 C1-Z2 的原因。如果把未知詞也考慮進來的話，可以發現更多更長的標注錯誤群。一個未知詞不只造成本身的誤標，也連帶造成相鄰詞的誤標。

5. 結　論

本文介紹以一階互聯隱藏式馬可夫模式爲基礎，配合 Kupiec 氏的等值群觀念，製作中文詞性自動標注系統之理論基礎、系統設計、及實驗結果。在所作的測試顯示此系統的標注正確率高達百分之九十六，是頗

令人鼓舞的結果。然而，我們也發現此模式尚有一些弱點，以致混淆詞的處理，未臻理想。所設計的比對式後處理雖可以改善部份問題，仍須進一步改良。以下是我們未來的研究方向：(1) 利用更大量的語料測試驗證此自動標注系統；(2) 嘗試製作二階隱藏式馬可夫模式新版本；(3) 未知詞標注問題的研究；(4) 複合詞標注問題的研究；(5) 比對式後處理之規則收集；(6) 斷詞和詞性標注同時處理；(7) 和語音識別語言解碼器 (linguistic decoder) 之整合應用。

附　註

本文英文稿發表於1993年6月在美國舉行的 Workshop on Very Large Corpora: Academic and Industrial Perspectives [1]。係工業技術研究院電腦與通訊工業研究所執行經濟部委託之前瞻性資訊系統技術研究專案 37H2100 計畫成果之一。

參 考 文 獻

1. C.-H. Chang and C.-D. Chen. HMM-based Part-of-Speech Tagging for Chinese Corpora. In *Proc. of ACL-93: Workshop on Very Large Corpora*, pages 40–47, Columbus, Ohio, USA, June 1993.

2. K. Church. A stochastic parts program and noun phrase parser for unresticted text. In *Proc. of ICASSP-89*, pages 695–698, Glasgow, Scotland, 1989.

3. S. DeRose. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14:31–39, 1988.

4. J. Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6:225–242, 1992.

5. Y.-C. Lin, T.-H. Chiang, and K.-Y. Su. Discrimination oriented probabilistic tagging. In *Proc. of ROCLING V*, pages 87–96, Taipei, 1992.

6. B. Merialdo. Tagging text with a probabilistic model. In *Proc. of ICASSP-91*, pages 809–812, Toronto, 1991.

7. L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

8. B. Santorini. Part-of-speech tagging guidelines for the Penn Treebank project. University of Pennsylvania, Pennsylvania, March 1991.

9. M.S. Sun, T.B.Y. Lai, S.C. Lun, and C.F. Sun. The design of a tagset for Chinese word segmentation. In *First International Conference on Chinese Linguistics*, Singapore, June 1992.

10. 中央研究院中文詞知識庫小組. 國語的詞類分析 (修訂版). 技術報告 T0002, 中央研究院計算機中心, 台北, 1989.