限制式滿足及機率最佳化的中文斷詞方法

張俊盛　陳志達　陳舜德

國立清華大學資訊所

摘 要

本文提出一個使用機率模式的系統，解決斷詞上的問題。首先將斷詞轉換成限制條件滿足問題，再以詞獨立出現機率做動態規劃的最佳化決定，運算時間與詞的長度成幾近線性的關係，詞的長度也不受限制。

一. 介紹

中文句子的結構中，單獨的漢字，並非句法上及語意上的最小元素。詞才是中文中能夠獨立出現，並自由運用的最小單位[5]。因此研究句法語意分析，必須以詞作爲基本單位。中文斷詞的目的，主要是爲了簡化自然語言處理系統的運算步驟，避免考慮太多不可能的斷詞情形。

西方語言並沒有斷詞問題，因爲西方語言的字彙單獨就具有獨立的語意。但要用中文表達一個最基本有意義的概念時，卻可能要用數個漢字來組成。例如相對於英語的grape，在中文要使用兩個漢字「葡萄」。再考慮三個字長的字串：好學生。在這個字串中，隱藏著五個詞彙：好、學、生、

好學、學生。總共可以找出三種組合，卻只有一種組合是合理的：好|學生。

將輸入句子的字串順序，轉換成詞語順序的過程就叫做斷詞。

目前以文法爲主要架構的中文自然語言處理系統中，斷詞是系統辨識中文輸入句子不可缺少的步驟。而愈來愈多的中文電腦應用領域，如光學字體辨識、語音辨識、文書校對工具、資料檢索、簡繁體互換以及中文輸入法等等，也可望利用斷詞做爲新的輔助方法。

雖然詞彙現象是屬於整個文法中的一部份，而且因爲中文的詞彙可以靈活運用，常常參雜了句法成份。但是大部份的情形，仍然可以不牽涉句法分析就決定詞彙在句子中的位置。

以往電腦斷詞的研究，可略分爲由構詞規則爲出發點的法則式斷詞法[1,2]，以及利用統計數據資料歸納爲判斷憑據的統計式斷詞法[3,9]。

法則式斷詞法強調的是語言現象。[1]就利用有些漢字大多只出現在詞彙的首個位置或末個位置的現象，來簡化斷詞步驟。[2]則提出看法，認爲斷詞只是自然語言的一部份，斷詞程式應該不能錯失掉任何可能的正確結果。雖然[1,2]都嘗試利用構詞現象來輔助斷詞，然而語言學界的詞彙研究並未提供簡易的方法，可用來做爲斷詞的依據。因此[1]使用的是『長詞優先』的規則，雖然這個規則在某些情況會失敗。[2]則嘗試利用『字詞的結合性』來解釋斷詞的現象，然而並未提出實際的作法。

相對的，統計式斷詞則著眼於大量資料的處理。認爲語言的性質，可以從大量的語料庫，經由數學模式獲得。此類系統目前以蔡文祥與范長康的鬆弛法[3]，及Sproat-Shih的統計式斷詞法[9]爲代表。[3]首先引進機率模式，利用影像處理常用的鬆弛法，成功地解決斷詞問題上相當複雜的情形。[9]使用的方法較爲特別，他將詞看做是一串相依出現機率特別高的漢字，從大量語料中得到二個漢字相依出現的機率，並利用一階馬可夫機率模式做爲斷詞依據。藉此方法Sproat-Shih無須使用人工預建的詞典，在處

理詞的長度不超二個漢字的情形就能有很好的效果。因爲忽略了構詞規則，所以統計式斷詞法，在正確率方面會有一定的瓶頸。另一方面，因爲方法的不同，前者須要反覆的運算，速度較慢。而後者則須建一個龐大的統計表，並且目前也侷限在只考慮詞長度不超過二個漢字的情形。

近來的做法也有將兩種方法合併的。[12]利用中文中自由詞素(free morpheme)或附著詞素(bound morpheme)的性質簡化斷詞步驟。並且利用字詞詞性的一階馬可夫機率依機率值大小排列所有的可能結果。然後再使用HPSG的剖析器(Head-driven Phrase Structure Grammar parser)，首次在斷詞中利用句子的文法及語意做爲斷詞依據。

雖然[3,9,12]都能達到95%的以上斷詞正確率，然而因爲速度或是記憶體容量的限制，這些系統在實用性上都有一些限制。[3,9]目前都只能處理詞長不超過二個漢字的情形。[3]的方法雖然並未限制詞的長度，然而這將讓原本已偏慢的速度更加惡化。[9]的方法則很難擴張到更高階的機率模式，因爲受限於統計表的大小必須隨著機率的階數呈次方關係增加。雖然[12]展示了一個翻譯系統下的斷詞子系統，但卻使用文法做爲判斷依據。因爲中文文法在自然語言處理上的諸多現象比斷詞還要複雜，應用在斷詞上可能會有實用性的問題。

本文所提出的斷詞系統則完全以機率值爲判斷依據。利用獨立的機率模式，實驗的結果顯示，系統斷詞的正確率並不亞於複雜的鬆弛法。我們將將斷詞轉換成限制式滿足問題(Constraint Satisfaction Problem)，再以詞獨立出現機率做動態規劃(Dynamic Programming)的最佳化決定(Statistical Optimization)。減少了運算時間，詞的長度也不受限制。未來，我們將進一步利用構詞法則，補充詞庫的不足。

以下在第二節中，我們將描述斷詞如何轉換成CSP的問題形式，並說明統計式的最佳化模式。第三節將說明系統架構，並提出實驗數據及結果。第四、五節分別爲本文的討論與結論。

二. 機率式斷詞

(i)斷詞的問題描述

CSP (Constraint Satisfaction Problem)是一種用來解決限制式滿足的問題。著色問題，圖形辨認問題以及排程問題(scheduling problem)都可以利用CSP的方法來解決[7]。在給定一組限制式後，CSP的目的就在找出能滿足所有限制式的一組變數解。

一個二元的CSP問題，其限制式只涉及兩個變數：

給定n個變數$X_1,X_2,...,X_n$，以及一組二元關係的限制式

$K_{i,j}$：$(X_i,X_j)\in K_{i,j}$

找出n個變數的一組值$(x_1,x_2,...,x_n)$滿足所有的限制式$K_{i,j}$。

斷詞問題可以用CSP來解決。

假設斷詞的輸入是由n個漢字所組成的句子，

$S=(C_1,C_2,..,C_n)$

令C_i與C_{i+1}相鄰漢字的間隔爲X_i。斷詞的目的就是將任一個X_i，標上「斷」或「不斷」的註記。而且，被任二個最接近的斷點間隔所分開的漢字，必須是一組詞。

爲了說明方便，我們各用X_0及X_n代表最前與最末的間隔，並用符號"^"與"="表示間隔「斷」或「不斷」。

| C_1 | C_2 | C_3 | .... | C_n |

X_0 X_1 X_2 X_3 X_{n-1} X_n

以下，我們對限制式，作更進一步說明：

對S中的任一串相連漢字$W_{i,j}=(C_i, C_{i+1}, ..., C_j)$，如果$W_{i,j}$被認可爲中文詞，則依下列狀況處理之：

(i). i=j

意味著C_i是一個單字詞。X_{i-1}及X_i的值可以都是"^"。設定限制式(^，^)$\in K_{i-1,i}$。

(ii). i<j

此種情形表示(C_i, C_{i+1}, ...,C_j)是一組多字詞。因此X_{i-1}, X_i, ..., X_j 的一組解可能是(^, =, =, ..., =, ^)

設定限制式：

(^,=) $\in K_{i-1,i}$

(=,=) $\in K_{i,i+1}$

(=,=) $\in K_{i+1,i+2}$

.....

(=,=) $\in K_{j-2,j-1}$

(=,=) $\in K_{j-1,j}$

因此當輸入一個句子之後，我們就可以逐一檢查句中任一個漢字爲首的右向鄰接漢字是否爲一組詞彙。如果是，就設定限制式的內容。

下面我們舉實例說明：

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
把	他	的	確	實	行	動	作	了	分	析
X_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}

句子中隱藏的所有詞彙，列出如下：(下面的說明是基於假設：漢字"析"不在詞彙庫中)

C_1 = 把　　(^,^)$\in K_{0,1}$

C_2 = 他　　(^,^)$\in K_{1,2}$

C_3 = 的 C_3C_4=的確　　(^,^)$\in K_{2,3}$ (^,=)$\in K_{2,3}$ (=,^)$\in K_{3,4}$

C_4 = 確 C_4C_5=確實　　(^,^)$\in K_{3,4}$ (^,=)$\in K_{3,4}$ (=,^)$\in K_{4,5}$

C_5 = 實 C_5C_6=實行　　(^,^)$\in K_{4,5}$ (^,=)$\in K_{4,5}$ (=,^)$\in K_{5,6}$

C_6 = 行 C_6C_7=行動　　(^,^)$\in K_{5,6}$ (^,=)$\in K_{5,6}$ (=,^)$\in K_{6,7}$

C_7 = 動 C_7C_8=動作　　$(\wedge,\wedge)\in K_{6,7}$ $(\wedge,=)\in K_{6,7}$ $(=,\wedge)\in K_{7,8}$

C_8 = 作 $(\wedge,\wedge)\in K_{7,8}$

C_9 = 了 $(\wedge,\wedge)\in K_{8,9}$

C_{10}= 分 $C_{10}C_{11}$=分析　$(\wedge,\wedge)\in K_{9,10}$ $(\wedge,=)\in K_{9,10}$ $(=,\wedge)\in K_{10,11}$

爲了避免不同詞的互相干擾，以上每一個詞所設定限制式中的"="，應與其他詞所設定者不同。不過爲了表達清晰起見，我們把所有的"="都寫成一樣。整理之後可列出限制式如下：

$K_{0,1}$ = { (^,^) }
$K_{1,2}$ = { (^,^) }
$K_{2,3}$ = { (^,^) (^,=) }
$K_{3,4}$ = { (^,^) (=,^) (^,=) }
$K_{4,5}$ = { (^,^) (=,^) (^,=) }
$K_{5,6}$ = { (^,^) (=,^) (^,=) }
$K_{6,7}$ = { (^,^) (=,^) (^,=) }
$K_{7,8}$ = { (^,^) (=,^) }
$K_{8,9}$ = { (^,^) }
$K_{9,10}$ = { (^,^) (^,=) }
$K_{10,11}$ = { (=,^) }

根據文獻中有關研究，CSP中的限制式，可以預先消去相互抵觸的限制條件，而且不會影響最後的結果。這種方法叫做Arc Consistency algorithm，簡稱AC[7,8]。

以上面的例子來說，經過AC簡化限制條件後，可以拿掉：

$K_{9,10}$={(^,^)}

這是因爲由$K_{10,11}$，我們知道X_{10}="="，因此$K_{9,10}$中的(^,^)不會爲最後的結果所滿足，將它消除並不會影響最後的結果。

文獻中二元CSP的研究也證明，若將變數視爲結點，限制式視爲連接結點的邊，若形成一樹狀結構，我們就可以直接從限制式中，找出解答，而不須要回溯(backtrack)。斷詞問題轉換成二元CSP後，形成線性串列的形態，爲樹狀結構的特例，因此也可以不經回溯，就直接找出答案。但是因爲合乎限制式的答案個數依然太多，所以必須另外使用評估方法，幫助找出最可能的結果。

(ii)以機率值決定斷詞

令輸入句子$S=(C_1, C_2, \ldots, C_n)$，通過CSP測試的一個斷詞結果是$(W_1, W_2, \ldots, W_k)$。其中任一組$W_i$都是詞典中記載的詞。查出每一個詞彙的出現機率，使用獨立機率，做爲評估斷詞可能性的依據。因此句子S，最可能的斷詞結果是：

$$\text{argmax}P(W_1,W_2,\ldots,W_k|C_1,C_2,\ldots,C_n)$$

$$\doteqdot \text{argmax}P(W_1)*P(W_2)*\ldots*P(W_k)$$

$$\doteqdot \text{argmax}\prod_{i=1,n}P(W_i)$$

其中$P(W_i)$是經大量語料統計的詞彙獨立出現機率

以獨立機率模式應用於斷詞問題，方法雖不複雜，卻隱含著以往研究在斷詞上的諸多看法，以下將獨立機率與其他方法比較：

(a)長詞優先性質

獨立機率模式隱含長詞優先性質，因爲詞彙的出現機率大多遠小於1，所以項數較少有助於機率乘積的增加，使得斷詞結果傾向於長詞。

(b)詞素的自由性與束縛性

詞素(morpheme)是語言上最小有意義的單位。可以單獨使用的詞素叫做自由詞素，必須和其他詞素連用的詞素叫做附著詞素。

附著詞素不會單獨出現，其獨立出現機率必然趨近於0，也因此降低在斷詞中出現的可能。

(c)與Sproat-Shih統計式斷詞的比較

Sproat-Shih以相鄰漢字的出現機率做爲斷詞依據，其作用相似於本系統使用的詞彙出現機率。但是Sproat在選擇斷詞時使用的是貪婪法策略(greedy method)，相對於此，獨立機率模式則著眼於整體的最佳化，有助於複雜文句的處理。

(d)與鬆弛法之比較

鬆弛法和獨立機率模式都利用詞彙出現機率做爲判斷依據，雖然鬆弛法和獨立機率模式在作法上有顯著差異，然而由於斷詞問題的性質，使用獨立機率模式簡單的計算而無須繁複的收斂過程，就能得到和鬆弛法相近的結果。

(iii)以動態規劃法增進速度

系統將斷詞問題轉化成爲CSP，字的間隔"斷"或"不斷"是用一組變數表示。對應輸入資料的順序，可以排定變數求值的先後順序。因爲每一個變數的值都只跟前面的變數值有關，而且只考慮機率乘積最大的解，此種性質正好符合使用動態規劃法的兩大要素：多階段的決策過程(multistage decision process)及最佳化原則(principle of optimality)。

由於使用動態規劃法，系統可以避免執行時間與句子長度呈指數函數的關係。因此相近於DeRose標註英文字彙詞性的效果[6]，系統可以在幾近線性的時間內，完成中文斷詞。

三.實驗說明及結果分析

(1)實驗說明

這個系統目前是在IBM/AT相容個人電腦上使用TURBO PROLOG發展的。利用TURBO PROLOG的external data base功能，系統預先建立一個六萬個詞的詞庫。詞彙的出現頻率取自劉英茂等編著的[常用中文詞的出現次數][4]。詞彙的來源，則由[常用中文詞的出現次數]及[電子辭典][10]合併而成。合併後，扣除部份重覆的詞彙後，系統共收藏有詞彙60274個。合併後的詞庫，依照詞的長度區分，列表如表一。

因爲劉英茂編著的[常用中文詞的出現次數]，是根據大量語料，統計出一百萬個詞彙中，出現頻率最高的四萬詞，並且分別記載其出現頻率。此一資料正好做爲斷詞判斷的依據。但是電子辭典上還有22018個詞彙，是不在常用的四萬個詞彙以內。我們假設這些詞的出現頻率都小於1，並且給予初始頻率值0.99。合併後的詞庫，依照詞彙的出現頻率，列表如表二。

系統的測試資料，隨機取自一個包括科學、評論、小說、散文等約30萬詞的語料庫。目前系統在斷詞方面，已經測試了擷取自不同類別的11篇文章，共43069個漢字的語料。測試語料分由機器及人工斷詞，再交由機器比對，核算出結果。

由於評估系統目地及方法上的不同，此處使用資料檢索文獻上常用的召回率(Recall)及精確率(Precision)做爲計算斷詞正確率的依據。如果令系統產生的斷詞結果爲M，人工斷詞的結果爲P，機器斷詞與人工斷詞一致的部份爲$P \cap M$，則召回率$= P \cap M / M$，精確率$= P \cap M / P$。

其中召回率的著眼點在得知，正確結果有多少比率可以由系統產生出來。系統平均可達到95.97%的斷詞召回率(正確的詞數/測試的總詞數)。精確率則可以衡量系統產生的結果，有多少比率是正確的。系統可以達到91.83%的精確率(正確的詞數/測試的總詞數)。字正確率則以漢字爲計算單位，系統平均可以達到93.89%的正確率。詳細測試結果列表如三。表四並根據召回率對錯誤情形做進一步分析。

由於Sproat-Shih的系統只考慮簡化的二字詞斷詞問題，爲了在比較上有所依據，我們另外統計二字詞的情形。在人工斷詞的資料中，找出二字詞有11312個，系統辨認到其中的10542個，召回率爲93.19%。而在機器斷詞的輸出中，共產生二字詞10755個，與人工斷詞相符的部份有10542個，精確率爲98.02%。詳細測試資料如表五。

系統斷詞的速度，目前每秒大約可以分別完成6.7組詞的測試。因爲受限於TRUBO PROLOG資料庫的速度限制，因此如果將它移轉到更佳的環境下，預料還可大幅改善其速度。

(2)結果分析

以下我們從測試結果中，摘選一些斷詞錯誤的例子。例句中畫有底線之處，爲錯誤的斷詞。句尾括弧中的編號，爲該例句的來源檔案編號。

(1) 缺乏¦ 讓人¦ 重¦ 讀¦ 的¦ 吸引力 (l-2)
(2) 攻擊¦ 人類¦ 並¦ 以¦ 人爲¦ 食 (j-14)
(3) 出現¦ 不¦ 同¦ 意識¦ 狀態 (j-10)
(4) 患¦ 精神¦ 官¦ 能¦ 異常¦ 的¦ 人 (j-10)
(5) 出¦ 現在¦ 遠離¦ 文明¦ 的¦ 沼澤 (l-4)
(6) 相¦ 對於¦ 國內¦ 食品¦ 廠家 (a-1)
(7) 從¦ 非洲¦ 的¦ 雨林¦ 草¦ 原 (j-14)
(8) 到處¦ 有¦ 鱷¦魚¦ 游¦ 來¦ 游¦ 去 (l-4)
(9) 昊¦ 昊¦ 的¦ 青天¦ 燦¦ 燦¦ 的¦ 白日 (g-1)
(10)經理¦ 小¦ 室¦ 德¦ 太¦ 郎¦ 離開¦ 大廳¦ 後 (l-2)

表一　詞彙長度分佈表

詞彙長度	詞彙個數
1	8097
2	39054
3	8046
4	4723
5	263
6	71
7	12
8	8
9	1
合計	60275

表二　詞彙頻率分佈表

詞彙出現頻率	詞彙個數
0.99	22018
1	16791
2	5929
3	3078
4	1949
5	1365
6	970
7	770
8	642
9	497
10	390
11 ~ 50	3909
50 ~ 54438	1967

表三　測試結果

檔案編號	總字數 A	人工斷詞總詞數 B	機器斷詞總詞數 C	機器與人工斷詞相符部份 D	總錯誤字數 E	召回率	精確率	字正確率
a-1	4314	2741	2867	2643	243	96.42%	92.19%	94.37%
b-1	5122	3291	3464	3137	346	95.32%	90.56%	93.24%
e-1	3178	2332	2463	2243	206	96.18%	91.07%	93.52%
g-1	4457	3201	3305	3104	205	96.97%	93.92%	95.40%
j-10	3361	2237	2372	2140	236	95.66%	90.22%	92.98%
j-14	3932	2734	2800	2666	146	97.51%	95.21%	96.29%
l-2	3588	2432	2554	2295	284	94.37%	89.86%	92.08%
l-3	3950	2720	2937	2517	412	92.54%	85.70%	89.57%
l-4	4709	3333	3466	3209	260	96.28%	92.59%	94.48%
l-8	3582	2546	2628	2469	169	96.98%	93.95%	95.28%
p-1	2876	2104	2155	2053	126	97.58%	95.27%	95.62%
合計	43069	29671	31011	28476	2633	95.97%	91.83%	93.89%

召回率 = D / B
精確率 = D / C
字正確率 = 1- E / A

表四　對召回率錯誤發生原因進一步分析

錯誤原因	因詞庫未收藏詞導致的錯誤				因機率值值而誤斷	錯誤詞數合計	正確詞數合計	人工斷詞總詞數	召回率
檔案編號	一般複合詞	純複合詞	重疊構詞	地名,人名或譯名					
a-1	21	51	0	18	8	98	2643	2741	96.42%
b-1	17	54	0	59	24	154	3137	3291	95.32%
e-1	47	26	0	0	16	89	2243	2332	96.18%
g-1	22	22	14	25	14	97	3104	3201	96.97%
j-10	29	14	0	21	33	97	2140	2237	95.66%
j-14	25	17	1	10	15	68	2666	2734	97.51%
l-2	5	8	1	98	25	137	2295	2432	94.37%
l-3	8	18	7	151	19	203	2517	2720	92.54%
l-4	8	26	3	64	23	124	3209	3333	96.28%
l-8	18	17	0	23	19	77	2469	2546	96.98%
p-1	12	20	3	3	13	51	2053	2104	97.58%
合計	212	273	29	472	209	1195	28476	29671	95.97%
百分比	0.71%	0.92%	0.10%	1.59%	0.70%	4.03%	95.97%	100.00%	95.97%

表五　二字詞統計資料

(i) 召回率

檔案編號	人工斷詞個數	與機器斷詞不相符的個數	召回率
a-1	1284	48	96.26%
b-1	1509	102	93.24%
e-1	696	41	94.11%
g-1	1054	64	93.93%
j-10	993	63	93.66%
j-14	1023	45	95.60%
l-2	989	93	90.60%
l-3	1012	169	83.30%
l-4	1163	80	93.12%
l-8	915	35	96.17%
p-1	674	30	95.55%
合計	11312	770	93.19%

(ii) 精確率

檔案編號	機器斷詞個數	與人工斷詞不相符的個數	精確率
a-1	1268	32	97.48%
b-1	1432	25	98.25%
e-1	678	23	96.61%
g-1	1004	14	98.61%
j-10	945	15	98.41%
j-14	988	10	98.99%
l-2	921	25	97.29%
l-3	865	22	97.46%
l-4	1106	23	97.92%
l-8	890	10	98.88%
p-1	653	14	97.87%
合計	10755	213	98.02%

斷詞錯誤的原因可略分爲兩類，第一類是因爲詞庫未收藏正確的詞彙，導致斷詞失敗。第二類則由於機率模式的限制，最佳機率值的句子並不是正確的結果。進一步分析錯誤發生的原因歸納如下：

1.因機率值導致的錯誤

此類錯誤在測試時共發生209次，約佔全部錯誤詞數的17%。

(1) 假設m個漢字長度的詞 $W = (C_1,C_2,...,C_m)$，其中包含著兩個詞：$W_1=(C_1,...,C_k), W2=(C_{k+1},...,C_m)$。

(i)如果 $P(W)>P(W_1)*P(W_2)$，系統會優先選擇W，而不會選擇 $W_1|W_2$。

雖然大部份的情形長詞優先短詞的規則會成立，但在少數的情形下，也有例外。 例如：例句(1,2)。

此種情形表示著長詞之內包含著短詞，這些短詞不但可以單獨使用，而且在長詞中的排列順序又正好符合句子的文法結構。主述、動賓等複合詞較易發生這種情形。

爲了要保留所有可能的情形，可以在詞典建構時，在這類詞彙上特別記載此種情形。因此，除了選出出現頻率較高的長詞外，還能在剖析回溯時優先考慮此種可能。

(ii)反之如果 $P(W) < P(W_1)*P(W_2)$，系統優先選擇短詞$W_1|W_2$，而非長詞W。在這種情形下，詞彙即使已被詞庫收藏，也不會被選到。例如：例句(3,4)。

產生這種現象的詞彙，詞彙中的漢字單獨出現的頻率一定非常高。這類的詞彙爲數並不多，也常常是非必要的。

以例句(3)的"不同"爲例，實際上的意義是"不+同"。類似的詞彙還有"這次"="這+次"，"一個"="一+個"。

但是例句(4)的"官能"，就無法輕易從漢字原來的組合看出意思來。爲了解決這類問題，系統可以提供次佳選擇來配合句法剖析器，讓錯誤可以彌補得回來。所以在系統優先提供機率值最高的結果後，仍然須要保留長詞，以便剖析程式回溯時可做爲考慮。

(2)假設m個漢字的一串漢字$W=(C_1,C_2,...,C_m)$，W_1及W_2是W中包含的一組詞彙，且$W_1=(C_1,...,C_j)$，$W_2=(C_{j+1},...,C_m)$。如果W中也包含另一組詞彙，$W_3=(C_1,..., C_k)$，$W_4=(C_{k+1},...,C_m)$，而且$P(W_1)*P(W_2) > P(W_3)*P(W_4)$，系統會優先選擇$W_1|W_2$，而非$W_3|W_4$。假如正確的斷詞結果是$W_3|W_4$就會發生錯誤。

系統因機率值而產生的錯誤，大部份屬於這一類。例如例句(5,6)是屬於這種情形。這是因爲出現的頻率雖然較高，局部卻不合句法。單獨地使用詞彙出現頻率做爲斷詞依據，是無法考慮到這種現象的。

以例句(5)爲例，"|出|現在|"的頻率雖然高過"|出現|在|"，但是前者卻可能較不合語法。對於此種情形，一個值得考慮的解決方式，可以加入詞性的相連出現機率來輔助斷詞。再以同一組詞爲例："出"是Va:【動作不及物動詞】，"現在"是Nd：【時間詞】，"出現"是Vh：【狀態

不及物動詞】，"在"是Pe：【時間或地方標誌,後接動作所發生的時間或場所】。因爲系統的詞性相連出現機率顯示，P(VaINd) < P(PeIVh)，表示後者的斷法較易出現在句子中，因此可能會是較正確的斷法。

2.因系統未收藏的詞彙導致錯誤

此類錯誤在測試中共發生986次，約佔全部錯誤83%，顯見此類問題的普偏。分析發生錯誤的詞彙，可歸納爲三大類：複合詞、構詞、地名人名及譯名，以下分別討論。

(i)複合詞

詞素依照本身的意義，還可區分成實詞素及虛詞素兩種。實詞素有實在的意義，虛詞素則沒有[12]。根據[11]的定義，複合詞是由兩個或以上的實詞素組合而成的單位。完全由自由詞素結合而成的複合詞就稱爲純複合詞(pure compound)。若是複合詞中包含有附著詞素，而且全部是實詞素，那麼就叫做一般複合詞(general compound)，否則如果包含了任一個虛附著詞素，就稱爲純詞(pure word)。詞素是虛或實，差別在語意，對斷詞系統而言並無差異。因此此處將純詞也歸類於一般複合詞中，和純複合詞相互比較。

附著詞素因爲須與其他詞素連用，不能獨用，可用來幫助減少一些組合情形的考慮，並可藉以找出包含有附著詞素的一般複合詞。例句(7)的"草原"就是純詞，因爲"草"和"原"都具有自由詞素的用法。然而例句(8)的"鱷魚"卻是一般複合詞，因爲"鱷"只具備附著詞素的用法。

由表四得知，實驗中986個因爲詞庫未收藏的詞彙導致的錯誤，有212個是屬於一般複合詞，273個是屬於純複合詞。(實驗並未實際針對每一個詞彙由語意分析其詞素成份。任一個詞素，只要詞庫中登錄有自由詞素的用法，就不再做附著詞素的考慮。)

這意味著實驗中的212個錯誤，佔全部測試資料的0.71%，可望透過此一簡單性質，而無須使用複雜的規則來幫助辨認。

不過目前這種作法仍隱含著兩個問題：

1. 附著詞素應與那些鄰接的詞素合併目前仍無法得知。
2. 盲目地強迫附著語素與其它語素合併成一個較長詞彙的作法，是否會造成更多的錯誤。

雖然附著詞素從語意上來看是不完整的，然而卻不知道附著詞素和句子中的那些部份形成一個詞，並且也不知道連成的詞會多長。

另一方面，因爲具有附著詞素用法的漢字數量很多，這些漢字多數又同時具有自由詞素的用法，如何不使原來當作自由詞素用法的詞彙被誤用爲附著詞素，仍是一個問題。

(ii) 重疊構詞

重疊構詞具有明顯的特徵，與其他構詞現象相比較，辨識時不須使用太多規則，即使盲目匹配也能得到很好的改進效果。例句(9)爲此種錯誤。因重疊構詞產生的錯誤有29個。

(iii) 地名、人名或譯名

與複合詞或構詞相比較，此類詞彙不具共同特徵，目前沒有有效的方法可用來輔助辨認。例句(10)爲此種錯誤。但注意到當系統未正確辨識詞彙時，斷詞結果往往產生許多的單字詞，而且單字的排列也不按語法，往往使斷詞或詞性標示的機率數值異常偏低。

此類錯誤非常普遍，共有472個，佔全部測試資料的1.59%。

四. 討論

1. 以詞庫做爲詞彙的查詢依據，將會因爲詞庫的容量限制，而影響斷詞的正確率。事實上除了構詞法則所衍生的詞彙以外，有一部份的詞彙是不可能完全收納於詞庫中。尤其在分析報紙社論時，發現根本無法一一將這些可能稍縱即逝的詞彙收集在詞庫內。這反應出一個現象，那就是中文詞彙的組成現象非常活躍。因此對於這個問題，如果不從詞素的層次來分析，還是會不斷遭遇新詞的困擾。[12]曾提出以語法律及語意關係，來判斷複合詞的存在並且預測詞性。
2. 即使存在無限大的詞庫，能夠包容所有詞彙，還是會有許多情形是目前的斷詞系統所無法解決的。例如漢語語法中的"併入現象"，若是發生在複合詞中，就會使詞彙產生變形，而無法從字典中查到。比方說述賓式複合詞的"生氣"，可以在動詞詞素及名詞詞素中併入名詞組，而說成"生你的氣"。這種現象無疑的表示，中文詞彙靈活運用的程度可以超越詞彙的層次。因此，與其在斷詞程式盲目使用簡陋的規則，還不如將這種問題提升到斷詞之後的構詞步驟再解決。
3. 中文的許多詞彙，例如人名或地名等專有名詞，若不在詞庫的收藏範圍之列，根本缺乏可資認定的規則，也很難單純從句法結構就認出此類詞彙。這類情形除了對句法的了解外，往往還需要前後文的語意以及常識輔助，才會有較正確的判斷。
4. 斷詞系統對複合詞的認定，向來沒有一定的標準。例如"棒球手套"或"棒球|手套"，"木棒"或"木|棒"，不同的斷詞方法以及不同的應用範圍，都可能會有不同的看法。這種情況在複合詞的組成份子都是自由詞素時，尤其紛歧。因爲當詞素組成一個較大單位的詞彙時，語意會跟著改變。因此，如果詞庫中並沒有收藏這個複合詞，那麼在由詞素合成詞彙時，必然要有某種方法來指示，合成過程所造成語意上的改變。

五.結論

獨立機率模式在中文斷詞上有很好的表現，可用來輕易架構在其他系統上，提供一種高效率的前置詞彙辨識處理方法。

機率方法可用來幫助語言處理系統大量減少繁複的規則，對於句子的表面結構，往往有出人意表的效果。但是由於機率的方法過於直覺，少量的規則可能無法避免，因此我們將進一步使用構詞法則，來幫助斷詞。

未來我們還將藉助機率模式，幫助句子中專有名詞的認定，並且利用詞性標示的結果輔助斷詞做更精確的判斷。

謝詞

本文研究得到國科會補助，計畫編號NSC80-0408-E011-07，謹此致謝。實驗中所採用的部份詞彙資料，來自工研院電通所技術移轉的國語日報電子辭典。感謝原始發展的中研院詞庫小組，以及電通所王明松先生的技術支援。清大語言所鄭縈小姐給予我們有關辭彙上的許多建議，助理林東游小姐幫忙語料庫的建立，以及區思萍、汪莉娟、陳瑋芬同學協助斷詞結果的人工校正，在此一併申謝。

參考書目

[1] 何文雄，中文斷詞的研究，碩士論文，國立台灣工業研究技術學院，1983.

[2] 陳克健、陳正佳、林隆基，中文語句分析的研究-斷詞與構詞，技術報告TR-86-004，中央研究院，1986.

[3] C.K.Fan and W.H.Tsai, 1987, "Automatic word identification in Chinese sentences by the relaxation technique," Proc. of National Computer Symposium, 1987, pp.423-431.

[4]劉英茂等，常用中文詞的出現次數，六國出版社，1975.

[5]趙元任，中國話的文法，中文大學出版社，1982.

[6] DeRose, S.J., Grammatical Category Disambiguation by Statistical Optimization, Computational Linguistics 14, 1988, pp.31-39.

[7] Dechter, R. and J. Pearl, Network-Based Heuristics for Constraint-Satisfaction Problems, J. of Artificial Intelligence 34, 1988, pp.1-38.

[8] Mackworth, A.K. and E.C. Freuder, The Complexity of Some Polynomial Network Consistency Algorithms for Constraint Satisfaction Problem, J. of Artificial Intelligence 25, 1985, pp.65-73.

[9] Richard Sproat and Chilin Shih, A Statistical Method for Finding Word Boundaries in Chinese Text, Coumputer Processing of Chinese & Oriental Languages, Vol. 4, March 1990.

[10]工業技術研究院電子工業研究所，中文電子詞細部設計手册，1990.

[11]中央研究院計算中心中文知識庫小組，國語的詞類分析，技術報告T0002，1989.

[12]陳克健等，國語中的複合詞和語言剖析，76年全國計算機會議論文集，1987，415-422頁.

測試語料來源

a-1.中東戰爭相關報導

中國時報 第9版80.1.19.

中國時報 第10版80.1.19.

工商時報　第2版80.1.19.

b-1.社論

中國時報 第四版80.1.13.

工商時報 第二版80.1.13.

工商時報 第二版80.1.19.

g-1.中國現代散文評析

蒲公英的歲月　余光中　長安出版P743-P748

e-1.天文望遠鏡使用與製作

丸山秀明　　世歲出版P130-151

j-11.焦慮與精神官能病

馬丁博士　桂冠出版P20-P27

j-14.野生動物世界

1.虎　2.豹　3.美洲獅　時代公司出版

l-2.推理雜誌69期 P.193-P.200

l-3.推理雜誌70期　創作推理 P.49-P.59

l-4.推理雜誌72期　違者必死 P.177-P185

l-8.樹與女· 冥府客棧　胡品清譯　爾雅出版P24-P31

p-1.當代中國大陸作家叢刊　女作家卷2

雨，沙沙沙-荒山之戀　王安憶 P192-198

(註:檔案的分類仿照BrownCorpus)