TR-86-004

中文語句分析的研究—斷詞與構詞

陳克健　　陳正佳　　林隆基

中華民國七十五年二月

第一章　引言

自然語言處理的目標在使計算機具有使用人類語言的能力。自有計算機以來就有許多人在從事這方面的研究，隨著硬體技術的日益進步以及人工智慧、認知科學、計算語言學等方面專家的研究所得，自然語言處理目標的達成益見可能。

到目前爲止，已經有很多自然語言處理系統被建造出來〔Dy83, He77, Ka73, La82, Ro82〕。它們分別被應用在文章閱讀分析、文章自動摘要、文句翻譯、問題諮詢系統、資料詰取等方面。此外，也有系統被用作爲語言學、心理學等方面的研究工具。

雖然讓不懂英文或不了解電腦的中國人直接經由中文與電腦溝通是一個很理想的辦法，但是在現在爲止，有關中文的語言處理的研究仍然非常少。爲了讓電腦能更深入而普遍的被國人使用，研究出各種中文語言處理系統是非常有必要的。

對於中文而言，一個句子是由字(character)所組成。而西洋語言是由詞(words)所構成。字只是書寫的單位而已，本身並不具文法意義，因此它並不是句法分析的基本單位。句法分析的基本單位是詞。一個詞語是由一個或一個以上的字元所組成。所以，我們必須有一個部門負責把輸入的字元串 (character sequence) 轉換成對應的詞語串，這是斷詞所要做的工作。

斷詞工作會面臨許多問題。試考慮下面特別造出來的句子：

1 開發中國家不知道路有幾條。

從這個句子裡，我們發現了下列問題：

1. 什麼是一個詞？

到底"開發中國家"當成一個詞，還是"開發" "中" "國家"當作三個詞。

2. 如果"開發中國家"做爲一個詞。但是這五個字也恰好是"開發" "中" "國家"三個詞或"開發" "中國" "家"三個詞依序排列而成的字序列，斷詞是不是也應該把這六個詞都產生出來？或者有些是贅詞應該除去，如何除去？

3. "不知道路"該對應成"不知" "道路"還是該對應成"不知道" "路"？

上面的例子與問題點出了斷詞必須解決的三個問題：

1. 如何決定什麼是一個詞語
2. 如何減少斷詞時產生的贅詞數目
3. 如何解決一個字元同時與前後字元合成詞語的問題。

我們提出一些方法，它可以去除所有在斷詞時就能知曉的贅詞，並且由於這些方法的使用，使得斷詞部門所能接受的文句不再限於白話文，即使是文言文或文言白話夾雜的句子，它也一樣能處理。

語句經過斷詞步驟之後，希望能得到一個最佳的唯一分段及每一段（即一個詞）所代表的詞類及詞性，事實上要做到唯一的解並不可能〔黃 84〕。例如，老王買好酒，及老王買好酒。兩種分段結果都說得通，不同的分段結果有不同的文法結構和不同的語意。因此斷詞的程式不能只產生一個分段結果。又例如，國民大會代表，

可以是國民大會代表，這裏的代表指的是動詞，也可以是'國民大會代表'一個詞，這裏的代表是名詞，國民大會代表是兩個名詞組成的複合詞。因此不只要有不同的分段，每一段還可能有多重的詞類及詞性。

基本上每個詞的詞類、詞性及其它文法資料，形成一個詞的資料檔，叫做字典。字典形成最簡單的方法就是把所有的詞全部收集到字典裏，但是這是一個不可行的方法。因爲詞集是一個開集合(open set)，嚴格的說它可以產生無限多的詞。例如數詞，一，二，三，……。就是無窮多而且新的詞語也因時勢的需要而不斷產生。再例如，任何一個國家或地區名稱（如：中國、美國、台北），都可以和'人'結合成一個詞，把所有這類的詞都放入字典也不恰當，因此構詞規則也是一個重要的研究題目。

第二章　詞語的處理

依照構詞學的原理詞是由詞素 (morpheme) 所構成，是不是只要把每一個中文字看成是詞素，由詞素結合成詞，由詞結合成句子。分析句子時，也以字爲基本單元呢？本章從計算機處理的觀點來看詞語的處理方式，上面所提的方式是行不通的。因此提出一個可行的折中辦法。部分的詞語以詞檔的方式建成字典，部分以構詞的方式產生。

以下的討論僅就一些已知的文法現像加以歸納整理，無法得到週全的結果，希望從實作得到回饋進而能了解完整中文語詞的文法現像。

2.1　斷詞與構詞

斷詞的過程所牽涉的領域正好是語言學上構詞 (morphology) 研究的範圍。在構詞方面，語言學家爲我們提出了許多斷詞時必須考慮到的問題，但是，基本上本系統的斷詞與語言學上的構詞，在對問題的研究觀點上，並不完全相同。

在斷詞 (recoghition of words) 的階段，我們重視的是如何把一串字元轉換成與這些字串相對應的詞。這些字串一開始並不被認爲具有意義，只有當與他們對應的詞找到之後，我們才認爲這串字元是一個詞，這時它才算具有意義。但是，構詞學探討的則是詞語的內在結構，它假設每個詞語都是由更小的具有意義的單位組成（我們稱這種具有意義的最小單位爲詞素），然後他們從詞語的組成詞

素的功能上，來描述出每個詞語的結構，對於中文而言，大部分的詞素都是由一個字元組成，只有少數是由二個字元或更多字元組合而成。當然，我們也可以把斷詞程序再分解成兩個部分，一個部分負責找出字元所對應的詞素，另一個部份則負責把詞串轉換成詞串，但是由於下列理由，使得這種想法不切實際。

1. 詞素的辨認因人而異〔趙，70〕，對於像"豆腐"，"姊夫"這些詞，我們曉得各別字元都具意義，因此是一個詞素，對於"彷彿""菩薩""玫瑰"這些詞，由於各別字元不具意義，因此這些詞本身就是一個詞素，可是像"如果""組織""立即""麻煩"，這些詞是否可以再加以分析成更小的單元，不同的人可能就會有不同的答案，這使得我們在詞素的決定上會發生困難。

2. 並沒有一套完備的詞素合成規則，使得我們能夠在分析完詞素串中各詞素間的關係之後，產生對應的詞串。

3. 以機器處理的眼光來看，即使2式能行得通，它也不及"盲目匹配" (blind match) 這種辦法來得有效。

所謂的"盲目匹配"是指我們不再企圖由詞語的組成詞素合成的詞語，而是我們直接以詞語的組成字串爲鍵字 (key)，然後把有關這個詞語的資料放入字典內，將來如果輸入字串中含有這個子字串 (substring)，匹配程式自然會把這個詞語找出，而跳過構詞分析的步驟。

雖然直接把構成詞語的字串放到字典，是可以避開許多構詞的問題，但是由於下列原因，這種方法尚需要做深入的考慮。

1 如果把所有成詞的字串通通列到字典內，將會使得字典所佔的記憶容量變得非常的龐大，這種過大的字典也間接會影響字典搜尋的詞語的速度。

2 有些詞語的生成方式是非常有規律的，同時它們也可能具備有高度衍生性，有時甚至是無限的衍生性，在這種情況下，想要把它們窮舉出來列入字典，如果不是行不通也是非常不智，因爲只要能夠掌握住這些類詞語的生成規則，我們可以在斷詞時或者以後的階段，把它們從組成詞素或組成詞語中合成出來，而避免掉字典裡面訊息的大量重複。例如：中文中可以和 "人","家", "學" 等結合成詞的詞素非常多，如：中國人，美國人，……化學家，數學家，……統計學，生理學，……。他們的結合規則又很簡單，這些合成詞都不應收到字典內。

2.2 字典收集詞語的原則

由於把所有的詞語通通收集到字典裡，並不是一個很好的辦法，有些具有規則的詞語，我們希望它可由存放在字典內的相關詞語合成而出，而不是重複的把這些可推演而出的詞語放到字典上。

當一個詞語同時滿足下列三項條件之時，我們就應該是用程式將它產生出來，而不是列入字典。

(1) 它具有明確的規則，讓我們能夠從相關詞語中合成出來。

(2) 它的語法、語義特性，也必須能夠由相關的詞語中得出（否則，我們將無法做語法、語義分析）。

(3) 即使(1)(2)的條件均滿足，但是如果只有少數的幾個詞語滿足

這種規則，或者這個詞語在句子中的出現頻率非常高，那麼爲了減少系統的複雜，增進語詞的合成速度，我們可能也把它列到字典中，而不是產生它。最具代表性的例子是"我們""你們"這些詞，雖然他們滿足"名詞+"們"複數名詞"這種簡單規則，很容易就可以合成出來，但是由於它們的出現率非常高，因此我們把它放到字典，以增進斷詞的速度。

第三章　構詞方法

基於前一章的討論，我們將例舉出得自〔趙,70〕與〔Li,81〕上的一些中文詞語構詞方式，然後再探討以機器處理的眼光來看，這些構詞方式是否具有規則可尋，系統應該如何處理這些詞語。

根據〔Li,81〕與〔趙,70〕在中文中，較有規則的構詞方式有：1.重疊，2.附加，3.複合三種。

3.1　重疊詞語的處理

所謂重疊是指一個詞語的詞素重複而衍生出一種新詞出來，衍生出來的詞語在語法或語義上都可能不同於原先的詞語。

重疊的類型主要有：

(A)意志動詞的重疊

所謂意志動詞指的是當動作者施行動詞，表示的動作時是有意志的（例如："教" "看" "吃"等），它們可以重複以表示動作者，"稍微地"做某事，例如：

(1) a.　請你教他英文

　　b.　請你教教他英文

像這一類的例子有：

嚐嚐、看看、想想、說說、寫寫

請教請教、打聽打聽、評斷評斷

當這一類的動詞是單音節時，我們還可以有×一×，×一×看，×看看這一類的型式（其中×代表單音節動詞）。例如：做一做

，寫一寫看，寫寫看。

但是當意志動詞是可以分離的動賓複合詞時，例如：睡覺、打仗、開刀，則只有第一個成分需要重疊，例如：睡睡覺、打打獵、開開刀。

(B)形容詞與副詞的重疊

形容詞可以重複而使得它的語意效果較原來的意義更加生動，而副詞的重複則有加重語氣的強調意思。

(2) a. 漂亮的臉

b. 漂漂亮亮的臉

(3) a. 他的確見過張三

b. 他的的確確見過張三

然而並不是所有的形容詞均可重複，我們可說：

簡簡單單、平平凡凡、活活潑潑、老老實實、確確實實、的的確確、平平白白、靜靜、圓圓、胖胖。

但是我們卻不能說：

* 偉偉大大　*抽抽象象　*賢賢明明　* 友友善善

* 一一定定

到目前爲止還沒有人提出規則，說明那些形容詞或副詞可以重疊，那些不行。

(C)量度分類詞的重疊

大多數單音節的量度詞或分類詞可以重疊，以表示"每一個"的意思，例如：

磅磅肉　　條條新聞　　件件消息　　個個蘋果

篇篇文章　　天天年年

但是月、分、秒則不可重疊成：

＊月月　　＊秒秒　　＊分分

(D)親屬用語的重疊

例如：爸爸、媽媽、哥哥、姊姊、弟弟、公公、舅舅、婆婆、姑姑

但是卻不可說：　＊姨姨　　＊兄兄

同時，也不是所有的親屬用語均可重疊，例如：姪子、外甥、孫子等。

對於A, B, C類的重疊詞語，我們都不把它收入字典，而是在斷詞的某一階段，利用特別的程式將它們合成出來。由於這一種構詞方式非常具有規則，我們只要用很簡單的程式，即可產生這些重疊詞語，但是因爲並不是所有的動詞、形容詞、或量度詞皆可重疊，因此，對於可重疊的這些詞語，我們會在字典裡註上特別標記，將來這些標記即可引導適當的程式來處理重疊問題。

由於D類的詞語並不很多，而且它也具有一些反例，因此，這一類詞都直接的被列入字典而不再加以分析。

3.2　附加詞語的處理

所謂的附加詞素是指：添加到其他詞語以形成較大單位的附著詞素，這些附加詞素大多無法獨自構成一個詞語，依照附加詞素在詞語中出現的位置，我們可歸納爲三類：

(A)詞首

(B)詞尾

做爲詞首的附加詞有：禁，可，好，難，舍，家，貴，弟，令，堂，表，單，多，汎，僞，不，反，老，小，第，初。

例如：禁穿，禁看，可愛，可笑，好吃，好看，難聽，家父，家兄，敝處，敝人，敝姓，貴庚，貴姓，貴國，令堂，令兄，堂哥，表哥，多目標，多音節，僞君子，僞政府，老王，老張，老二，老三，初一，初五，小張，第三，第五十，……

這些詞首大多具有一個特徵，那就是它只會附加到少數幾個詞語上而已，而且我們也不容易得出附加後的詞語的語法、語義特徵。因此，除了"老""小""第""初"這四個較有規則，它們所附加而成的詞語，應由程式合成以外，其他詞語可以通通列入字典。

出現在詞尾的附加詞素，主要的有：兒，們，學，家，子，頭，的，化，性，論，觀，率，法，界，炎，員等，與這些詞尾合成出來的詞語的合成規則，一般而言都是片斷而不很一致的，處理這一類詞的原則大概是這樣的：

針對這些詞尾找出可以與它結合的詞語，然後觀察這些詞語是否具有某些規則性（語義上），然後再研究這些合成後的詞語在文法上是否一樣具有規則性，如果上述條件通通滿足，我們便可把這些資料附著到這些詞尾上，將來一旦句子中有些詞尾出現，它便會觸動適當的程式到詞尾這個地方找尋所需的資料，然後與前面的詞

語比較，以決定是否可以合成爲一個詞語。

例如：以"學"爲例，我們可能會有這種規則：

專門學科名詞1＋"學"⟶名詞,表有關名詞1的專門學問

於是將來一旦字串內有"學"這個字，附著在"學"內的資料，就終究會使程式檢查"學"前面的詞語是不是一個名詞，如果是的話，它就會把這個新詞合成出來。但是，有一點值得注意的是，這種合成並非一定正確的，因此，我們仍然必須保留舊的兩個詞語。

以"經濟學台北"這句話來說，如果我們把"經濟""學"合成"經濟學"，而去除掉了"經濟"與"學"這兩個舊詞語，將來的語法分析就得不出正確的結果。

除了上述的詞尾以外，時貌記號"了"（表完成），"著"（表持續），"過"（表經驗）等都是附加在動詞詞尾的附加詞，它們具有影響動詞特性的功能，因此在目前我們是把它放在系統的語法分析部門處理。

這些詞首、詞尾的附加詞還有一個重要的功能就是做爲斷詞的記號，中文中可用的斷詞記號包含：(1)標點符號，(2)詞首字，(3)詞尾字〔何，82〕，(4)獨立字。所謂詞首字（詞尾字），指的是這些字組合成詞的時候永遠在詞首（詞尾）出現，因此可以視爲一種斷詞記號。附加詞大部分都屬於這一類。

3.3 複合詞的處理

第三種構詞方式是複合，所謂複合大致是指兩個或兩個以上的

詞語或詞根合成而爲一個新詞，有關複合詞的明確定義可參考〔趙，70〕。

根據趙元任的分類，複合詞可依照它的內在文法結構，主要有五種：

1. 主謂複合詞 (subject-predicate compounds)

   "年輕" "路過"

2. 主從複合詞 (subordinate compounds)

   "熱心" "瓜分"

3. 並列複合詞 (coordinate compounds)

   "依靠" "貴重"

4. 動一賓複合詞 (verb-object compounds)

   "破產" "失業"

5. 動一補複合詞 (verb-complement compounds)

   "推開" "進來"

在這五項複合詞之中的 1 2 3 類以及大多不屬於這五類的複合詞，絕大多數已成成語，他們都是孤立的成語，因此除了將它們通通列入字典以外，並沒有其他的辦法可以由它的組成詞素或詞語中衍生出來了。在中文中比較值得探討應該如何處理的複合詞，除了 4. 5. 項以外，還有名詞性複合詞。

3.3.1 名詞性複合詞的處理

所謂名詞性複合詞是指兩個名詞合成而爲一個新名詞，這種複合詞的最大特徵是它具有無限的衍生性與創造性，並且有大部分名

詞性複合詞，它的語法語義都可由組成分子的語法語義推導出來，由於上述原因，使得很多這類複合詞都可以不列入字典，而讓程式利用規則產生出來。

下面是一些名詞性複合詞的例子，（我們假設 N_1 是名詞性複合詞的第一個名詞，N_2 則是第二個）

1. N_1 是球類名稱，N_2 是運動器具

   棒球手套

   籃球框

   網球拍

   足球架

2. N_1 是國家或地區種族名稱，N_2 是 "人" " 話" 或政府機關組織

   中國人　　義大利人　　中國話

   英國人　　阿拉伯人　　英國話

   美國政府　　台灣省政府

3. N_1 是 N_2 的一部分或組成分子

   烏爪　　客廳沙發　　學校教員　　學校學生

   牛尾　　辦公室桌子　　大學校長　　台大教授

4. N_1 是 N_2 的組成材料

   大理石桌子　　皮鞋　　棉衣

   銅像

   金屬盒子

這類複合詞的處理辦法，主要是利用他們的語意特徵來判別，

我們可以在系統裡寫下很多名詞複合規則，例如：

1. 動物名稱＋身體部分的名稱→表動物身體部分的名稱
2. 機關組織名稱＋機關組織成員→表某機關組織的某種成員
3. 種族名稱、地區名稱、國家地區＋ "人" → 表某一種族、地區、或國家的成員
4. 專門學科或技術名稱＋ "專家" →某一學科或技術的專家

然後，我們再在字典裡每一個詞語的特徵欄中，存入每一個詞語的語意特徵，將來到了句法分析時，控制程式每當看到兩個相鄰名詞時，它就會檢查這兩個名詞的語意特徵構成的語意特徵序列，是否滿足某條名詞複合規則，而決定是否合成一個複合詞。

例如：以 "張三很喜歡吃台灣香蕉 "這句話而言，當控制程式看到 "台灣" 與 "香蕉" 時，它由 " 台灣 " 與 " 香蕉 " 的特徵欄裡拿出語意特徵值，而組成" 地區名稱＋水果名稱 "的序列，然後發現這個序列滿足某條複合規則，於是 "台灣香蕉" 這個詞語就被衍生出來。

以這種方式處理名詞性複合詞的最大好處，是它可以大量的減少字典裡的字彙儲存個數，因爲這一類的複合詞在中文中實在太多了，雖然這個處理變法會稍微降低處理速度。

3.3.2 動一補複合詞 (verb-complement verb)

這一類複合詞又叫結果式動詞複合詞 (resultative verb compounds) ，它是由兩個要素所組成，第一個要素是一個動詞，而第二個要素則表第一個要素動作或過程的結果，根據〔 Li,81 〕第

二個要素可以表示各種不同的結果：

(1)原因：

(a)我把茶杯打翻了

(b)窗戶已經被關上了

(2)達成：

(a)請把字寫清楚

(b)我已經找到書了

(3)方向

(a)他跳過去了

(b)他們跑進來了

(c)我要收回我的東西

(4)階段

(a)我的錢用光了

(b)張三已經把歌唱完了

動補式複合詞具有下列特性：

1.它可得在詞語中間插入“得”或“不”以表能力，例如：

(a)他跳過去了

(b)他跳得過去

(c)他跳不過去

2.它不可重疊，例如：我們可說

“嚐嚐”“活動活動”，但卻不可說“拉開拉開”以表暫時貌。

3.除了方向動詞以外，動－補語之間不插入“得”“不”以外的任何

詞語：

例如：我們不能說“看了到”“看過到”

到目前爲止，我們除了處理含有“得”或“不”的能性動賓複合詞以外，所有其他動賓複合詞，我們都把它列入字典。動賓複合詞的處理，較名詞性複合詞的處理要來得困難，因爲：

1.合成後的複合詞文法特性難以預測，例如：“坐”和“壞”都是不及物動詞，可是“坐壞”卻是一個及物動詞，例如：張三坐壞了一把椅子。

2.並無明確的規則可以使我們曉得那些詞或詞素可以複合成動賓複合詞，例如：以“到”爲例，我們有“看到”“想到”“夢到”“做到”，卻沒有“站到”“笑到”“跑到”。

以“見”爲例，我們則只有“看見”“夢見”，而沒有“想見”“做見”。

3.3.3 動賓複合詞的處理

第三種值得討論的複合詞是動賓複合詞，所謂的動賓複合詞是指這個複合詞的兩個組成成分具有動詞一賓語的文法關係。

下列例子都是動一賓複合詞：

1.(a)革命、關心、懷疑、注意、出版、提議、得罪。

(b)開刀、開玩笑、行禮、溜冰、跳舞、幽默、打主意。

動賓複合詞具有如下特性：

(1)整個單位的語意具有成語性，例如：開刀、作秀。

(2)組成要素有些不可分離，有些可做有限的分離，例如：

1.a 中的動賓複合詞都是不可分離的。

1.b 中的動賓複合詞則可有限分離。

由於動賓複合詞大都具有成語性，但是在文法上它却又可分開，宛如兩個獨立的詞語，我們面臨了這些資料該如何存放在字典的問題。

我們處理這類複合詞的原則是這樣的：

(1)對於不可分離的動賓複合詞，由於它與一般的詞語並無兩樣，因此我們將它收入字典。

(2)對於可分離的動賓複合詞，我們不但把這個複合詞收入字典，同時也把它的兩個組成分子列入字典，以“幽默”爲例，我們不但把“幽默”列到字典，同時也把“幽”“默”收入字典，這樣做的原因如下：

(a)由於這種複合詞具成語性，無法合成。因此，我們必須有地方存放它的語義，所以複合詞要列入字典。

(b)由於在實際句子當中，動賓複合詞大都不分離，直接把動賓複合詞存入字典，可以減少無謂的文法分析時間。

(c)字典中只列動賓複合詞，而不列入它的組成成分，會使斷詞的工作變得非常複雜，考慮句子“張三幽了李四一默”。

在此句中，如果“幽”不獨列出來，它在字典內就沒定義，此時控制程式勢必要找出句子中的所有字元，來和“幽”拼湊以決定“幽”是不是某個動賓複合詞的動詞部份，這樣做是很沒效率的。

利用上述原則處理含有動賓複合詞的句子時，當分離開來的動賓複合詞的二個組成分子，經過文法分析之後，它們會以詞組的型態組合在一起，此時他們都是沒有語意的“空動詞”或“空受詞”，但是到了語意分析階段程式會檢查出這種情形，然後把它們合成爲一個詞語，再到字典中把這個複合詞的語意拿出，因而圓滿解決動賓複合詞語意上不可分離，語法上可分離的問題。

第四章　斷詞方法介紹

4.1　斷詞基本原則

中文輸入的基本單位是字而不是詞，因此在句法分析前，必須先由斷詞程式把輸入進來的字串轉成詞串，然後才有辦法做句法分析。

但是，由於中文字的特性，有時一個字即可構成一個詞（例如：水、吃、打），有時，一個字可以同時和字串的前後文構成不同的詞（例如："知道理"這個字串中的"道"可和前文"知"合成"知道"，也可和後文"理"合成"道理"）。有時一個字串本身雖已經是一個詞，但它卻只是句中更大字串所構成的詞的一部份而已（例如：在"我愛中華"和"中華民國"之中的"中華"，在前者"中華"是一個獨立名詞，可是在後者"中華"卻只是"中華民國"的兩個組成字元罷了），有時一個字串到底是一個詞還是兩個詞，我們根本無法決定，例如"國民大會"這個字串，可斷成"國民大會"，也可斷成"國民""大會"，因爲"國民"與"大會"都是有意義的詞，在這種情形下，我們面臨了該如何斷詞的問題。

最直截了當的辦法就是把輸入字串中的任何片斷所可能對應的詞語通通找出來。以"美國人"爲例，這種方法就會把"美"、"國"、"人"、"美國"、"國人"、"美國人"通通找出來，以"台灣大學"爲例，就會把"台"、"大"、"學"、"台灣"、"大學"

、"台灣大學" 找出來。這種辦法的最大問題是，它會產生許許多多沒用的贅詞，以致於影響了剖句的速度與正確性。但是我們也不能只保留最大片段構成的詞，以"台灣大學"爲例，固然在大部分的句子當中，它都是用來指稱某一所大學，但是也不能排除它是兩個詞恰好靠在一塊的可能，例如：我們可說"在台灣大學很多"，在此時."台灣" 與"大學" 是兩個獨立的名詞。

斷詞程序只是整體語言處理程序的一部分，在這個階段不可能完全解決上述問題，亦即我們不可能要求斷詞程式做到每一輸入字串都只對應出一串正確的詞串，但是我們也不可讓斷詞程式產生太多沒用的詞語，因此一個斷詞程式原則上要遵守下面二個原則：

1 斷詞要完全：

一個斷詞程式必須保證不會錯失掉任何正確詞串中的任何詞語，否則會造成錯誤。

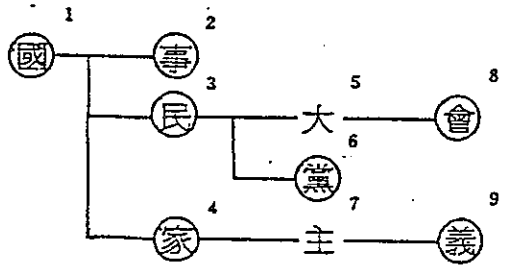
2 儘量減少贅詞數量：

在完全性的前提下，我們必須儘可能的把在斷詞階段就可確定是沒有用的詞語去除，以免妨害剖句效率。

4.2 詞樹介紹：

本系統的斷詞程序牽涉到詞語在字典裡的存放方式，因此我們首先介紹詞語在字典的存放格式。

中文詞語在字典中是以詞樹〔何，82〕的方式存放，一棵詞樹可以把所有以某個中文字元爲詞頭的所有中文詞語通通收集起來，予以有組織的結構安排，以便系統程式取用。圖一便是一顆以“國”字爲起頭的詞所組成的詞樹。



圖一：

在詞樹裡，每一個節點（ node ）都有三欄（ field）資料。他們分別爲匹配欄、資料欄和子樹群欄。

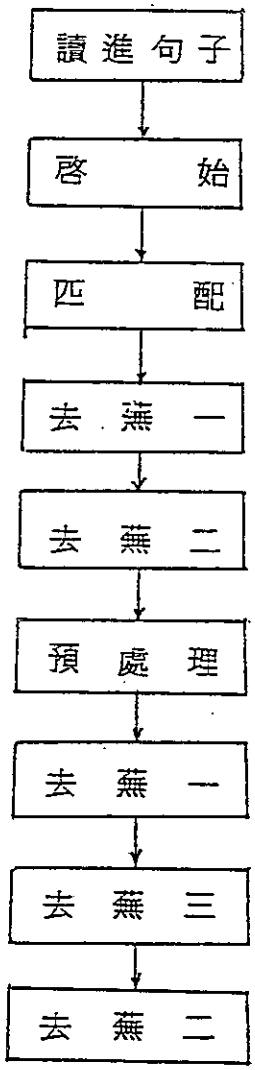
匹配欄：存放一個中文字元，主要是用做鍵語（ key ）以利程式找到詞語資料。

資料欄：存的是某個詞語的所有詞語資料，而這個詞語就是由樹根到這一節點的路徑內所有匹配欄內的字元，依序排列出來的字串所組成的詞，以圖一節點.6.而言，他的資料欄存的就是 1 ， 3 ， 6 節點內匹配欄內的字元，依次排列出來的字串組成的詞的詞語資料，也就是“國民黨”這個詞的資料，但並不是所有節點的資料欄內都有資料。圖二節點 5 就沒有，因爲“國民大”並非一個詞，所以它沒有詞語資料。在圖二中，凡是劃有圓圈的節點都是資料欄有資料的節點。

子樹群欄：存的是這個節點的所有子樹，對於末端節點，例如節點 2 ， 6 ， 8 ， 9 ，它們的子樹群欄爲空欄。

4.3 斷詞程序介紹

詞語的主要流程如圖二：



圖二：

其程序依序大致如下：讀進句子。建立系統工作區。找出輸入字串中任何片斷所可能構成的詞語，將它們掛到詞語收集圖。去除贅詞。利用掛在詞語收集圖端點上的特別訊息合成新詞語。再做一次贅詞去除工作。其中各單元的功能，以下分予描述：

4.3.1 啓始程式

啓始程式的作用在建立系統工作區的起始狀態，它的主要工作有三項：

1. 建立向量化輸入文句區
2. 建立詞語收集區
3. 建立剖句訊息收集區

其中向量化輸入文句區是一個一維陣列，裡面的元素存放的是輸入字元。

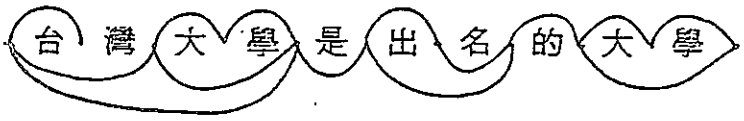
詞語收集區與剖句訊息收集區，也是一個一維陣列，每個陣列裡存放的一個代表端點的錄，這個錄裡收集了所有以這個端點爲起點或終點的離弦與入弦，在啓始的時候，這些錄都是空的。假設輸入是“台灣大學是出名的大學”，在啓始程式執行完後，詞語收集區的形狀變成如下：

· 台 · 灣 · 大 · 學 · 是 · 出 · 名 · 的 · 大 · 學 ·

圖三：

4.3.2 匹配程式

匹配程式的功能是它會把輸入字串中的任何片斷所可能對應的詞語通通找出來，圖三的情形，經過匹配程式處理之後，就變成圖四的情形：



圖四：

在圖四中，每一根弦表示在字典內可以找到一個詞語，它的組成字元恰好是這根弦所包含的字串。

匹配程式的程序如下：

1. 設定詞樹列串爲空列串（nil），片斷記錄列串爲空列串（nil）。

2. 拿進下一個輸入字元（假設爲C）。

3. 如果字典內沒有以C爲起頭的詞樹，則：

把這個字元的編號推入（push）片斷記錄列串；

否則就取出C爲起頭的詞樹，將它推入一個詞樹收集區（我們稱之爲詞樹列串）。

4. 對於存放在詞樹列串中的每一棵詞樹做：（假設現在是針對詞樹T）

4.1 如果T的資料欄不是空欄，則：

4.1.1 如果資料欄內有特別資料欄，則將這些資料取出，送到詞語收集區中，這個詞語的起始端點上。

4.1.2 如果資料欄內還有資料，則將這些模板原型取出，掛到詞語收集區中，這個詞語的兩個端點上。

4.2 如果T的子樹欄內有一棵子樹，它的匹配欄內的字元是C，則取出這個子樹，以取代T。

5.如果輸入字元尚未拿完，則回到步驟1.。

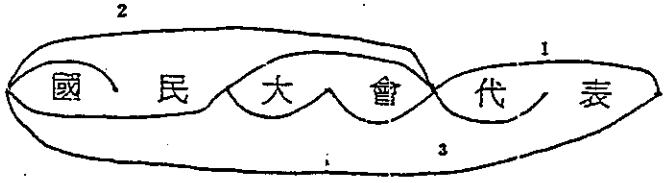
在上述程序中，詞樹列串的用途是收集到目前為止，還可繼續匹配下去的詞樹。

而片斷記錄列串的用途，則是記載"中斷點"，所謂的中斷點是指沒有其他弦跨過這個端點，例如圖四的點5、點6都是中斷點，可是點3、點10都不是。片斷記錄列串的資料將來會被去蕪程式用到。

4.3.3 去蕪程式

4.3.3.1 去蕪程式一

去蕪程式一的目的，在消除一個長字串構成的詞語中，可能附帶產生的贅詞，例如："國民大會代表"，經匹配程式後產生如下的結構（假設在字典裡，"民"和"表"不被收集為詞）：



圖五：

在這個詞語圖表裡，我們認爲只有弦 1 、弦 2 、弦 3 值得保留，其他都是贅詞。

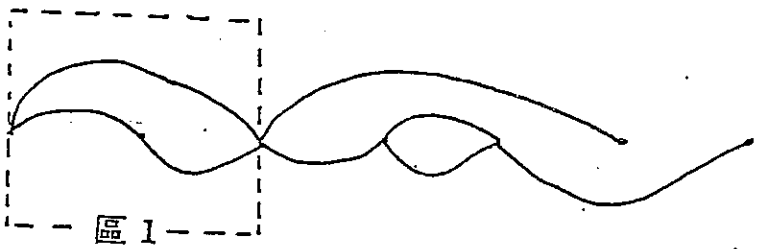
在說明去蕪程式一之前，我們先介紹一個名詞“片段”：

一個圖形（ Graph）的連續子圖形是一個片段，如果：

(1)片段外的弦不會跨進或通過這個片段。

(2)存在一根弦恰好連著這個片段的兩個端點。

例如：在圖六之中只有區 1 是一個片段



圖六：

去蕪程式一的程序如下：

1 針對詞語收集區內的每一片段做：

如果存放在這個段落的終止端點上的注意欄內有　保留程式，則取出這個程式計算出那些弦要保留，分別在這些弦上做標記。

除了連接片段兩端點的弦，以及標記需保留的弦以外，所有的弦通通去除。

例如圖五中的" 國民大會代表" 是一個片段，依照步驟 1 2 除了連接片段兩端點的弦，及標記需保留的弦以外，所有的弦通通去除，因此爲了保留弦 1 和弦 2 ，在弦 3 字典內的資料欄必須標記，"國民大學" 及 " 代表" 兩詞保留。

去蕪程式一是一個很有用的程式，在大部份情況下，它都可以刪除掉許多沒有用的贅詞。

4.3.3.2 去蕪程式二

除了構成片段的圖形，可以讓我們找出許多贅詞以外。另外一種找尋贅詞的辦法就是找出那些無法連到詞語收集區的兩端點的弦，他們都是沒用的弦，考慮圖七的圖形：

圖七：

假設"解"與"答"都不是一個詞，因此，字典沒有收錄，則圖七便是一個經過匹配程式，去蕪程式一處理過後的圖形。

在這裡，我們發現"了解"這個詞是贅詞，因爲它不是某一完整詞串中的一個組成分子，因此，必須予以去除，去蕪程式二負責執行這件工作。

去蕪程式二的程序如下：

1.由詞語收集圖形的起始端點到終結端點，依序針對每一端點V做：

1.1 除了起始端點外，如果V沒有任何入弦，則刪除V的所有出弦。

2.由詞語收集圖形的終結端點到起始端點，依序針對每一端點V做：

2.1 除了終結端點外，如果V沒有任何出弦，則除V的所有入弦。

4.3.3.3 去蕪程式三

在中文的構詞上，有所謂的“連用”（bound form）與獨用（free form），依照趙元任〔趙,70〕的說法；當一個詞素可以獨自存在出現時，我們就稱這個詞素是獨用，否則，便稱他爲連用。問題是在口語中，或許會有許多連用的詞素，可是當斷詞所接受的輸入文句是文言文或文言白話夾雜的句子時，連用詞素已不復存在，因爲幾乎所有的詞素都可獨用。

當上述說法成立時，去蕪程式二的作用即告消失，以圖六的情形而言，此時“答”不再不能獨自成詞，我們可說：“我就是不願答”，在此時“答”是一個動詞，它可獨自存在。

解決這個問題的徵結在於（假設目前的詞素序是ABCD）：

⑴雖然大部份的詞素都可獨用，但是，當它們與前後詞素可以構成一個更大的詞時，它們大都變成連用，本身不獨自成詞，例如：當“梨”和“子”一道出現時，這兩個詞素必定會合在一起，成爲一個詞，而“梨”本身構成的詞，則成了贅詞。如：“梨子不好吃”中，“梨”成了贅詞。但是在其他句子當中，“梨”可以獨自成詞，如“我喜歡吃梨”。

⑵當詞素B可同時分別和詞素A，C合成爲一個詞時，此時B到底該和A或C合成爲一個詞，主要取決於A，C兩個詞素和B結合力的高低。

當兩個詞素組成的詞語出現在句中，在此限制下，如果組成詞素仍然可以獨自成詞，則我們稱此詞素在這個詞語的結合力弱，反之，則稱它結合力强。

例如：“了解”中，“了”的結合力弱，因爲我們可以說“我說了解答”，此時“了解”出現，但“了”卻可獨自成詞，同樣的“答”對“解答”也是結合力弱，因爲我們有“了解答問題的方法”這種句子，但是“答案”中的“案”，道理中的“道”都是結合力强。

因此，當ＡＢ和ＢＣ都構成一個詞時，如果Ｃ的結合力弱，Ａ的結合力强，就該取ＡＢ，Ｃ，如果Ａ，Ｃ結合力都弱，就取ＡＢ，ＢＣ，兩個結合力都强時，表示這個句子有錯誤。

(3)當ＡＢ，ＢＣ，ＣＤ均可結合成詞時，此時毫無疑問的，我們該取ＡＢ，ＣＤ，而非Ａ，ＢＣ，Ｄ或其他。

例如：“了解答案”分解成“了解”“答案”，“知道理由”分解成“知道”“理由”。

去蕪程式三的程序如下：

1.針對每一個同時與前文，後文結合成詞的字（假設爲Ｂ）做：

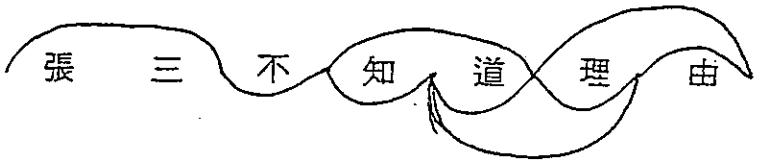
1.1 如果Ｂ是一個詞，則將Ｂ在詞語收集區中的弦拿掉。

1.2 如果Ｂ後面的字（假設爲Ｃ）也同時和前後文結合成詞，則將ＢＣ在詞語收集區中的弦拿掉。否則：

(A)如果B前面的字（假設爲A）結合力强，則刪除BC在詞語收集區中的弦。

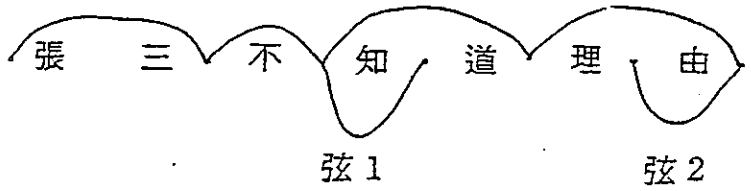
(B)如果C的結合力强，則刪除AB在詞語收集區中的弦。

以“張三不知道理由”爲例，在去蕪程式三執行前，它的情況如下：



圖八：

經過去蕪程式三處理後，變成圖九的情況：



圖九：

其中，弦1，弦2將來還可被去蕪程式二去除。

4.3.3.4 預處理程式

預處理程式的作用乃在於處理一些剖句程式無法處理或不易處理的文法現象，目前，可以在這個階段處理的事有：

1.時間詞串

例如：明末清初，民國二十六年七月七日。

2.數字詞串

例如：三百多，二千五百六十。

3.名詞性複合詞

例如：學生宿舍，老師辦公室。

4.名稱、頭銜

例如：台大資訊工程學系，美國總統雷根先生。

5.住址

例如：台北市羅斯福路315巷20號。

6.時貌詞尾的附加

例如：說過、看了、坐著。

7.動詞的重複

例如：高興高興。

8.動賓複合詞的複合

例如：穿起來、聽起來，吃完、做完。

9.動詞＋"一"＋動詞＋〔看〕表示嘗試

例如：吃一吃，吃吃看，吃一吃看。

10.動詞＋"看"＋"看"

例如：吃看看。

11.動詞＋"不"＋動詞　表示詢問

例如：看不看，知道不知道，知不知道。

12.形容詞的重疊

例如：高高興興。

預處理程式處理前述現象的方式，大致如下：

我們把負責處理這些現象的特別程式的名稱，附在字典中相關詞語上，當匹配程式做初步斷詞時，它會將這些程式的名稱拿出，放到詞語收集區的適當位置（主要是放在端點上的注意欄內），而預處理程式的主要工作就是把這些程式名稱拿出來呼叫執行。

預處理程式的程序如下：

1.針對詞語收集圖形中的每一端點，依序做（假設此時端點爲V）：

1.1把V注意欄內的每一個程式名稱拿出來呼叫執行。

在目前階段，我們爲預處理部門所寫的特別程式，均只具窺視前後有限項字元或詞語的能力，因此，它尚無法處理數字、住址、頭銜、名稱，這些需具狀態觀念的詞語串。

第五章　結論

中文的構詞文法規則雖不完全，但是採用字典的方式可以解決大多數的問題。而斷詞的成功與否又和字典的好壞息息相關。如何建立一個適用的字典成爲一件非常重要的工作，也是一件非常困難的工作。字典的內容除了詞樹以外，應包含一些有關每一個詞的附屬資料，如詞類、詞性、對應字碼、文法特性等等。這些資料的建檔，工程浩大，非一朝一夕可以完成的。應設計一個線上更新系統，以漸進的方式建立之。何文雄等〔何，82〕採用劉英茂先生〔劉，75〕所訂的四萬餘目詞的斷詞系統。此一系統尚未考慮句子文法分析之用故未把詞性、詞類資料建入字典中。四萬目詞對一般的白話文尚足以適用，對特殊的專有名詞可能就力有未逮。據統計中文詞彙數目超過十萬以上，如果字典網羅所有的詞，不僅數量過於龐大影響處理速度，且常有新詞產生，應如何更新也是一大問題。因此可能僅就常用詞彙建入字典內，把斷詞的程序加強，使無法辨認的小段都視爲專有名詞（或特殊詞類），這可能是一個比較可行的辦法。

中文斷詞不僅在語句分析上扮演著重要角色，同時在中文輸入的研究上，有重要地位。張系國〔張，73〕提出採用詞語注音做爲中文輸入的方式，主要用到了一個觀念就是詞可以解決字的輸入碼重覆問題。例如：ㄓㄨㄥ可以是"中" "衷" "鐘" ……，但是ㄓㄨㄥ ㄍㄨㄛˊ的ㄓㄨㄥ只有一個，就是"中"。雖然注音的方式並非一個很好的中文輸入方式。但是對其他的輸入方式，如輸入字形，我們仍然可以利用

詞檢的功能去解決輸入字碼的重號問題，希望能做到輸入員鍵入整句中文字串，由斷詞程式分斷出正確的詞，免去輸入員必須一邊鍵入輸入碼，一邊斷詞的負擔。將來如果語音辨認的能力足以分辨出每一個字音，則斷詞又是解決同音字不可缺少的工具。當然這只是一個粗略的構想，離實際的完成還有重重的困難必須克服的。

參考資料

1. 何文雄，中文斷詞的研究，國立台灣工業技術學院，工程技術研究所，1983.
2. 張系國等，中文輸入輸出系統，參考資料第二集，台北，中央研究院，數學研究所，1973.
3. 黃正德，中文生成語法概要，中央研究院資訊所演講稿，1984.
4. 趙元任，中國話的文法，台北，學生書局，1980.
5. 劉英茂等，常用中文詞的出現次數，台北，六國出版社，1975.
6. Dyer, Michael G., In-depth Understanding - A Computer Model of Integrated Processing for Narrative Comprehension, MIT Press, 1983.
7. Hentrix, G., "LIFER: A natural language interface facility", SIGART Newsletter 61, 25-26, 1977.
8. Kaplan, R. M., "A general syntactic processor," in Rustin (Ed) Natural Language Processing, Prentice-Hall, 193-241, 1973.
9. Lawson, Veronica (Ed), Practical Experience of Machine Translation, North-Holland, 1982.
10. Li, C. N. and Tompson S. A., Mandarin Chinese: A Functional Reference Grammar, Berkeley, University of California Press, 1981.
11. Robinson, Jane J. "DIAGRAM: a grammar for dialogues", CACM 25, 27-47, 1982.