# 以遺傳演算法為基礎的中文斷詞研究

陳稼興　謝佳倫　許芳誠

智慧型資訊系統實驗室　智慧型資訊系統實驗室

國立中央大學資訊管理系　國立中央大學資訊管理系　真理大學資訊管理系

## 摘 要

斷詞在中文自然語言處理上，是個非常重要的前期作業。本研究提出以遺傳演算法為基礎的中文斷詞模型，用以處理中文斷詞。在我們提出的模型中，詞庫是自動建立的，除了避免人為介入導致的不客觀性外，也避免浪費寶貴的人力資源。在斷詞處理上，則是利用詞庫中的「詞出現次數」和「詞長」兩個因子編成適應函數，作為遺傳演算法演化的依據。一般斷詞方法，在斷短詞上的效果不錯，一旦遇到長詞，正確率就會大幅下降；但是若改採長詞優先，則因長詞可能包含短詞，導致短詞可能斷不出來。本研究模型的特色是，長詞有較大的機會被優先斷出，而任何短詞只要在文章中出現的次數夠多，還是有機會被斷出。此外，在模型中我們運用遺傳演算法進行中文斷詞，由於遺傳演算法可以讓我們保留最好的前三個（或更多）斷詞結果，而不是僅僅保留一個斷詞結果，讓後階段的中文處理有更多的選擇，這樣的特性有助於處理「斷詞的歧義性(ambiguity)」的問題。為驗證模型的效益，我們採用中時電子報下載的電子檔案為樣本進行實驗。實驗分析結果顯示，本研究模型確實已達可接受水準。

**關鍵詞：中文斷詞、遺傳演算法、中文自然語言處理**

# A Study on Chinese Word Segmentation: Genetic Algorithms Approach

Jiah-Shing Chen, Chia-Lun Hsieh and Fang-Cheng Hsu

Intelligent Information Systems Lab

Department of Information Management

National Central University

## Abstract

For Chinese natural language processing systems, word segmentation is a very important pre-processing step. In this study, a genetic algorithm-based word segmentation model is proposed. In the model, a dictionary for word segmentation is automatically generated from the training articles. GA's population search feature makes it easy to find several better segmentation candidates, which are helpful to the following steps in Chinese language processing. Experimental results on 300 articles show that our GA-based approach to Chinese word segmentation is highly feasible.

**Keywords:** Chinese word segmentation, genetic algorithms, Chinese language processing

# 1. 緒論

中文的句法(syntactic)和語意(semantic)基本單位是「詞」而非「字」[許菱祥, 1986]，單獨的中文字未必是語句分析的最小單位。但在書寫或印刷的中文句子中，往往只有字的界線而無詞的界線。在此前提下，任何中文自然語言處理，例如：文件檢索、中文輸入、光學字體辨識、語音辨識、機器翻譯等，都需先對中文句子進行「斷詞」，才能進行下一步的處理。由於斷詞結果的正確性及完整性對後續的處理動作有關鍵性的影響，這使得中文斷詞變成一件非常重要的工作。

對大多數歐美語系國家所使用的語言(例如：英文)而言，它們的句法和語意基本單位是「字(word)」，雖然每個字都是由多個字母(letters)組成，但由於每個字與字之間都有明顯的空白做為分隔，所以可以容易的判斷出單字。換言之，這些類型的語言不需煩惱如何從眾多字母中分割出有意義的字，因此沒有斷詞的困擾。

所謂「中文斷詞」，是將一連串的中文「字」，轉換成 「詞」的組合，其中每個「詞」是由「一個字」或者「多個字」所組成。如何將由「字」所組成的「句子」切割成一個個的「詞」，是中文斷詞的主要使命，也是本論文的研究核心。

過去已有許多學者提出各種斷詞演算法，但迄今尚未有任何演算法能斷出完全正確的結果。傳統中文斷詞演算法通常很難處理長詞[Nic, 1996]，抑或即使可以處理長詞，但大多受限於被要求先建立詞庫，在新興資訊急速增加的情形下，新詞也不斷地出現，因此詞庫往往無法因應新詞出現的速度。另一方面，中文字超過七萬五千多個[黃大一， 1990]，即使一般常用字也將近六千個，這些「字」可以組合出的有意義「詞」數，包括二字詞、三字詞、四字詞、甚或五字詞、六字詞等等，可能是天文數字，若單以人工方式建立詞庫，勢必耗費很大的工夫，還可能不適用。此外，因詞庫中的長詞可能包含短詞，因此也可能導致短詞斷不出來。

本研究提出以遺傳演算法(Genetic algorithms)為基礎的中文斷詞模型，用以處理中文斷詞的問題。在我們所提出的模型中，詞庫是自動建立的，這樣做除可避免因人為介入導致的不客觀性外，也避免浪費寶貴的人力資源。而斷詞的處理則是依據詞庫中「詞出現次數」和「詞的長度」等資訊，作為遺傳演算法演化的依據，進而找出正確的斷詞。本研究模型的最大特色是，長詞有較大的機會被優先斷出，而任何短詞只要在文章中出現的次數夠多，還是有機會被斷出。

在本研究提出的模型中，我們運用遺傳演算法進行中文斷詞，由於遺傳演算法可以讓我們保留最好的前三個（或更多）斷詞結果，而不是僅僅保留一個斷詞結果，讓後階段的中文處理有更多的選擇，這樣的特性有助於處理「斷詞的歧義性(ambiguity)」的問題。另外，本研究模型的詞庫是從訓練文章中自動產生，除

對長、短詞問題的處理有獨特之處外，對專有名詞、複合名詞等複合詞問題的處理也有一定的成效。

為驗證模型的效益，我們採用中時電子報下載的 300 篇的電子檔案為樣本，進行實驗。實驗分析結果顯示，本研究所提模型對中文斷詞的「詞召回率」約為 87%，「詞的精確率」約為 85%。而最高的詞召回率達 88%，最高的詞精確率可達 87%，確實已達可接受水準。

本文的其他內容如下，第二節簡要介紹中文斷詞的問題與相關的斷詞遺傳演算法，第三節為遺傳演算法簡介與本研究模型之說明，第四節除了說明本研究的實驗設計和實驗步驟外，並對實驗結果進行個案分析與討論，第五節為本研究之結論與後續研究方向。

# 2. 中文斷詞的問題與斷詞相關方法

## 2.1. 中文斷詞的問題

一個中文句子往往有相當多可能的斷法，且不同的斷法可能有不同的意義，因此「斷詞的歧義性(ambiguity)」便成為中文斷詞的問題之一。斷詞的歧義基本上可分四種類型：句子結構歧義、詞彙歧義、詞類歧義、詞間歧義。其中的前三者，是中、西方語言常會有的歧異狀況，而「詞間歧義」則是中文特有的歧義類型[陳永德，1997]。

當一個句子中的歧義種類越多，則可能的斷詞組合也就越多，斷詞工作也就越難。例如：「閒雜人等不得跨越紅線」，可以斷成「閒雜人｜等｜不得｜跨越｜紅線」，也可以是「閒雜人｜等不得｜跨越｜紅線」。以此句子而言，就同時包含了結構歧義、詞類歧義與詞間歧義。

除了斷詞的歧義性外，中文斷詞面對的另一個困難是複合詞的問題。因為複合詞，例如：人名、地名等專有名詞，或定量詞、複合名詞等等，可以無限衍生，所以根本無法將所有的可能複合詞都列舉出來，更無法將它們全部納入詞庫中，因而常常出現無法解決的未知詞，成為中文斷詞成效降低的另一個重要因素。

多數斷詞演算法都需藉助於詞庫。理論上詞的長度可以是無限，實作上詞的長度則有其上限。詞庫允許的詞長上限越大，詞庫也將越大。太大的詞庫往往可能導致演算法的效率低至無法接受的情形，但若為了縮小詞庫而降低詞長的上限，則可能因為無法斷出長詞，而使得結果的正確性受到影響。由於中文的詞數眾多，詞庫必然無法包含所有的長、短詞，因此如何設計適當的詞庫也是中文斷詞必須解決的問題。

## 2.2. 過去的中文斷詞方法

過去十餘年間，已有許多研究[王良志、貝子勝、黎偉權、黃麗卿, 1991; 范

長康、蔡文祥, 1987; 陳克建、陳正佳、林隆基, 1986; Chen and Liu, 1992; Fan and Tsai, 1988; Gan, Palmer and Lua, 1996; Leung and Kan, 1996; Nie, Hannan and Jin, 1995; Sporat and Shih, 1990; Sporat, Shih, Gale and Chang, 1996; Wu and Tseng, 1995; Yeh and Lee, 1991]提出各種中文斷詞方法。基本上，這些中文斷詞法基本上可分成三種：統計式(statistical methods) [范長康、蔡文祥, 1987; Fan and Tsai, 1988; Sporat and Shih, 1990] 、法則式(heuristic rule-based methods) [王良志、貝子勝、黎偉權、黃麗卿, 1991; 陳克建、陳正佳、林隆基, 1986; Chen and Liu, 1992] 與結合法則式與統計式兩種方法的混合式(hybrid methods) [Nie, Hannan and Jin, 1995; Sporat, Shih, Gale and Chang, 1996; Yeh and Lee, 1991]斷詞法。由於已有學者對這些方法進行詳細的分析[Wu and Tseng, 1993, 1995]，因此本文僅作簡要介紹。

### 2.2.1. 統計式斷詞法

統計式斷詞是藉由語料庫(corpus)的資料來歸納語言現象，利用一組數學模式達到斷詞的目的，主要是依機率統計值來決定斷詞的位置。例如，[范長康、蔡文祥，1987]利用機率模式，直接計算各詞的出現頻率來反覆求得各詞的機率分佈，進而找出最佳的斷詞組合。[Sporat & Shih, 1990]經由大量的語料中統計出句子內每兩個相鄰字的相對強度，用已決定詞的邊界，並藉由一階馬可夫機率模式來作為斷詞依據。

統計式斷詞的優點是執行效率高，但由於統計式斷詞大多只能處理二字詞和單字詞，所以當詞長大於二時，則斷詞效率會大幅降低，且斷詞的正確率不高[Nic, 1996]。此外，大量的語料取得不易、統計資料相當佔空間、詞頻會因語料庫的建構者而異等也是統計式斷詞法的缺點。

### 2.2.2. 法則式斷詞法

法則式斷詞法通常需配合詞庫或辭典一起運作。法則式斷詞主要是根據一些規則，逐步排除不可能的詞語組合，以達到較好的斷詞結果。最具代表性的法則式斷詞法是「長詞優先法(maximum matching method or longest matching method)」[Li et al., 1988; Liang, 1990]，此方法建立在一個經驗法則上：在一個中文句子中，最有意義的詞通常是許多有意義的連續中文字串當中最長的字串。因此在一個句子中，此演算法將優先斷出最長的有效詞。例如「長江」及「長江三峽」皆存在詞庫中，但「長詞優先法」將優先斷出「長江三峽」，以保留最完整的語意。

另外，[陳克建等，1986]利用詞素(morpheme)構成的詞典來搜尋句子裡所有可能的詞，再以構詞規則、詞的結合力來排除一些不可能的組合。[王良志等，1991]提出一個結合斷詞和語法分析，以剖析為導向的斷詞法。由具有測試及評分的 Tomita 剖析器，和增強型上下文無關的文法(Augmented Context-Free Grammar, ACFG)寫成的中文構詞規則所組成。[Chen and Liu, 1992]所提的斷詞法則以應用構詞規則與六條經驗法則組成。構詞規則主要是用來處理複合詞和專有名詞，經驗法則則用以解決歧義性的問題。

法則式斷詞法的主要缺點是受到詞庫品質的影響很大。當句子中出現新生的詞彙，將使得斷詞正確性降低。若為了提高斷詞正確性，而不斷新增詞庫的詞彙，

則可能大幅降低斷詞的效率。

### 2.2.3. 混合式斷詞法

上述兩類斷詞法各有優劣，因此有學者提出綜合的斷詞法。例如，[Yeh and Lee, 1991]提出以聯併(Unification)為基礎的斷詞法，先利用詞典搜尋可能的斷詞組合，接著利用構詞規則簡化斷詞組合，再以一階馬可夫機率模式排列出所有可能的結果，然後依照機率值排列所有可能的斷詞組合，最後使用 HPSG 剖析器(Head-driven Phrase Structure Grammar Parser)逐一過濾這些斷詞組合，確認該斷詞組合是否合於文法。

# 3. 中文斷詞與遺傳演算法

## 3.1. 遺傳演算法

遺傳演算法是 Holland[1975]受達爾文的「物競天擇，適者生存」天擇說啟發而發展的演算法則。遺傳演算法採用一組特別的字串模擬各種生物的染色體(chromosome)，並計算所有染色體對環境的適應度(fitness)，在每個世代之間讓各個染色體以隨機的方式進行交配(crossover)與突變(mutation)來產生下一代，再根據該染色體的適應度選擇(selection)是否讓其生存。這個演化交替的動作會一直持續到達成最終目標（例如事先決定的演化代數）為止。在 Holland 的文獻[1975]中所描述的最基本演算法則，稱為 simple genetic algorithm (SGA)。SGA 大致的演算法如下：

```
SGA ( )
{
    隨機設定初始群組
    計算染色體的適應度
    當尚未達到最終目標
    {
        選擇好的個體
        進行染色體間的交配以及突變
        計算染色體的適應度
    }
}
```

遺傳演算法特別適用於搜尋解答空間很大、非線性、複雜、可能有雜訊、而且無法預測可能解的問題，這是傳統決定性最佳化技巧(deterministic optimization)或是貪婪法則(greedy heuristics)所無法做到的。遺傳演算法提供了一個相當簡單的系統架構、運作流程，卻能產生強大的解答搜尋能力；而且又具有高問題獨立性，和必須要依附問題模式的傳統演算法有明顯不同，這種彈性也是其他方法所不及的；同時遺傳演算法靠著群組間的各點可以同時探索不同的區域，再伴隨著世代演化交替、隨機搜尋的特性，這種平行處理的能力使它不容易陷入局部最佳解(local optimum)的困境，而向整體最佳解(global optimum)收斂；這些特點都讓遺傳演算法成為目前各領域的新寵。

遺傳演算法中的重要因子有：染色體編碼、適應函數、選擇方式、交配及突變。

1. 編碼(Encoding)

遺傳演算法中，唯一用來表示問題特性的就是染色體編碼。大部分的最佳化問題都有固定數目的變數，因此最普遍的編碼方式就是將這些變數對應到某個字元或整數，再把其編成固定數目的位元，將這些位元組合起來就成為一個染色體。

2. 適應函數(Fitness Function)

適應函數是用來評估每個染色體所代表之解答的好壞，即其適應度。通常適應函數即為最佳化問題的目標函數。

3. 選擇(Selection)

遺傳演算法中的選擇機制是模擬自然界適者生存的現象，適應度高的染色體存活率較高，而適應度低的染色體存活率相對就較低。因此，適應度較高的染色體所擁有後代有可能比較多。如果有某一個染色體適應度明顯高於其他的染色體，就有可能藉著世代替換，而逐漸成為這個族群的主體。

4. 交配(Crossover)

選擇完後，獲選的父代染色體將進行隨機交配，最簡單的交配方式是一點交配(one-point crossover)，隨機在兩個染色體 (A, B) 中挑一個切割點，將 A 的前半段與 B 的後半段組合，將 B 的前半段與 A 的後半段組合，再用來取代原來的 A 與 B。另外亦有兩點或多點交配，但最常用的還是一點及兩點交配。

5. 突變(Mutation)

突變運算是模仿自然界中生物基因的隨機突變，通常是依據一個很小的機率(如 0.001)來將某位元反轉。突變運算有助於遺傳演算法脫離局部最佳解。

## 3.2. 以遺傳演算法為基礎的中文斷詞模型

### 3.2.1. 系統架構

本研究所提的模型主要可分成兩個子系統：一個是詞庫訓練系統，另一個是斷詞系統，如圖 1 所示。

詞庫訓練系統為斷詞的前置作業系統，目的在於訓練並產生系統所需的詞庫。換言之，本研究模型的詞庫是由經由訓練文章自動產生的。當訓練文章被輸入，並經過斷句處理後，所有句子中的字將被分別以二字詞至 N 字詞(N>2)的方式累計其次數，並將所有字詞的內碼與累計次數資料記錄於詞庫中，作為詞庫的內容。若輸入的句子中有 N 個中文字，則系統將自動產生(N-1)個二字詞、(N-2)個三字詞、…、2 個(N-1)字詞。以「參選台北市市長」句子為例，將產生 6 個二字詞：「參選」、「選台」、「台北」、「北市」、「市市」、「市長」，5 個三字詞：「參選台」、「選台北」、「台北市」、「北市市」、「北市長」，4 個四字詞：「參選台北」、「選台北市」、「台北市市」、「北市市長」，餘依此類推。

當詞庫中的詞與新產生的詞完全相同時，該詞的累積次數就加 1。累積次數高，表示這個詞的內聚力較強。也就是說，既然這個詞常常聯袂出現，表示此詞在文章裡，很有可能是重要詞彙的一部份，才會常常出現；若次數偏低，表示此詞的內聚力較弱，也就是說，此詞應不為詞彙的一部份，只是剛好相鄰的排列在一起。

斷詞系統則是本模型的的核心，目的在於利用先前訓練出來的詞庫和遺傳演算法找出理想的斷詞結果。斷詞系統在接受輸入的測試文章，並經過斷句處理後，把句子交與遺傳演算法系統，並根據特定的適應函數比對詞庫，便可不斷演化衍出最佳的斷詞結果。

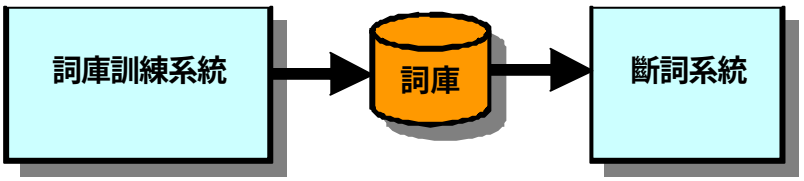


圖 1: 系統架構

如果詞庫中的詞非常大量時，搜尋工作將很耗時。為提昇效率，本研究的詞庫檔案結構採用雜湊法(hashing)[Robert et al., 1993]製作。如果有詞素的雜湊結果是相同時（即發生碰撞情形時），則在相同的儲存格上，利用串連法(chaining)的方式，往另外一個維度做鏈結，以解決碰撞問題。

## 3.2. 以遺傳演算法為基礎的中文斷詞基本程序

本模型的運作程序基本上可分為三個步驟：「斷句」、「製作詞庫」、「用遺傳演算法斷詞」。各步驟分別詳述如下：

步驟一：斷句

首先取得一些文章，除去所有的標點符號、阿拉伯數字，只留下純中文字。接著在連續兩個刪除點之間，切出一串串短句。將文章切成純中文的短句，主要目的是為下階段製作詞庫作準備。以下面的文章片段來為例，若輸入系統的原始文章如下：

> 【記者樊嘉傑台北報導】針對近來北高直轄市長選舉，國民黨、民進黨市長提名人日益激烈的「口水戰爭」，民進黨中央並不樂見。黨內高層人士在分析選情時指出，民進黨市長候選人在爭取婦女票方面，相較於國民黨候選人處於弱勢，引發「口水戰爭」，對婦女選票的吸收更為不利。年底「三合一」選舉，民進黨中央的基本策略是以直轄市長選舉帶動立委、市議員選舉的聯合作戰。因此對扮演「火車頭」角色的北高市長選舉極為慎重，黨中央召開的選戰會議，特別要求北、高市長提名人陳水扁、謝長廷派代表參加。

經過斷句步驟後，文章將被分割為下列一連串只包含中文字的短句：

記者樊嘉傑台北報導?
針對近來北高直轄市長選舉?
國民黨?
民進黨市長提名人日益激烈的?
口水戰爭?
民進黨中央並不樂見?
黨內高層人士在分析選情時指出?
民進黨市長候選人在爭取婦女票方面?
相較於國民黨候選人處於弱勢?
引發?
口水戰爭?
對婦女選票的吸收更為不利?
年底?
三合一?
選舉?
民進黨中央的基本策略是以直轄市長選舉帶動立委?
市議員選舉的聯合作戰?
因此對扮演?
火車頭?

角色的北高市長選舉極為慎重?
黨中央召開的選戰會議?
特別要求北?
高市長提名人陳水扁?
謝長廷派代表參加?

步驟二：製作詞庫

將步驟一產生的短句，分別以二字詞、三字詞、四字詞、五字詞（甚至六字詞）等，加上累計每個 N 字詞(N>1)出現的次數製作出詞庫。換言之，詞庫中每個紀錄中的資料，包含了所有語詞的內碼、以及該些語詞在文章中出現的累積次數、及下一個儲存格的位置。

步驟三：用遺傳演算法斷詞

假設欲接受斷詞的中文短句有 n 個中文字。為方便 GA 斷詞的運作，我們設計對應於短句的染色體長度為 n-1（也就是染色體含有 n-1 個基因），並讓這些基因分別對應 n 個字的 n-1 個間隔。在基因值的設計上，我們假設當基因值為 1，表示此間隔是詞與詞之間的分隔處；當基因值為 0，表示此間隔兩邊的字是詞的一部份，是屬於同一個詞裡的詞素，所以該處不應打斷。例如：短句「強調將堅決支持」有 7 個中文字，染色體的長度應為 6。若染色體的基因值為 100110，則表示該短句被斷成四個詞：「強」、「調將堅」、「決」、「支持」（圖 2）。

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ |
|---|---|---|---|---|---|---|---|
| 句子： | 強 | 調 | 將 | 堅 | 決 | 支 | 持 |
| 詞： | $W_1$ | | $W_2$ | | $W_3$ | | $W_4$ |
| 染色體： | 1 | 0 | 0 | 1 | 1 | 0 | |

圖 2: 斷詞染色體結構

本例句中有 7 個字，分別為 $C_1$ 到 $C_7$，我們將句首到第一個基因值為 1 的基因所斷開的一字詞「強」表示為詞 $W_1(C_1)$、最後一個基因值為 1 的基因到句尾所斷出的二字詞「支持」表示為詞 $W_4(C_6,C_7)$、其餘被最接近的兩個基因值為 1 的基因所斷開的詞「調將堅」、「決」分別表示為詞 $W_2(C_2,C_3,C_4)$，以及詞 $W_3(C_5)$。

接著根據斷出來的詞 $W_{I=1...y}$（在上例 y=4），計算出染色體的適應度。本研究所設計的適應函數(F)考慮「詞的累積次數」與「詞長」兩因素。基本上，一個詞在文章中出現的次數越多，表示此詞素可能是詞彙的一部份，因此我們以語詞在詞庫中累積次數表示詞的內聚力。

由於詞庫的設計，是將文章分解成以二字詞至 N 字詞的方式儲存，所以短

詞會被包含在長詞中，而且短詞出現的次數一定不會比長詞少。例如：假設「中央大學」的累積次數為 3，則「中央」和「大學」的累積次數也一定是 3，但若文章中還有「中央點」或「大學生」等詞，那麼「中央」和「大學」的累積次數就會大過 3。如此一來，專有名詞「中央大學」就無法被斷出來了。因此若不考慮詞長因素，則將無法斷出較長的詞。我們不但在適應函數中加入詞長，還將詞長取平方，主要是考量因為，累積次數的效應往往遠大於詞長的效應，為了加強詞長的權重才將詞長再做一次乘積。

$$F = \sum_{i=1}^{y} [T(W_i) \times L(W_i)^2] \ , \ i = N$$

其中：

$T(W_i)$ : 詞 $W_i$ 在詞庫中的累積次數

$L(W_i)$ : 詞 $W_i$ 的長度

$y$ : 詞的個數

若詞 $W_i$ 可以在詞庫裡找到，則直接以詞庫中所記錄的累積次數乘上 $W_i$ 詞長的平方，計算適應度。利用上述適應函數，使用者可以指定詞長的上限(M)。若 GA 所切出來的詞 $W_i$ 的長度大於 M 則可令 $T(W_i) = -1$，使適應函數的值降低，以鼓勵斷詞。另外，若詞 $W_i$ 長度小於等於 M，但在詞庫中找不到，則為鼓勵未知詞的合併，我們將它視為未知詞並令 $T(W_i)= 1$。

藉由遺傳演算法的三個基本操作：複製、交配、突變，以及適應函數的導引，適應度較高的染色體得以被保留，也就是較正確的斷詞解得以保留；而適應度較低的染色體將被淘汰。如此反覆演化，GA 將可找到最佳的斷詞方式。就上面的例子而言，GA 最後應可演化出最佳的染色體 011010，此染色體所對應的斷詞為：「強調」、「將」、「堅決」、「支持」（圖 3）。

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ |
|---|---|---|---|---|---|---|---|
| 句子： | 強 | 調 | 將 | 堅 | 決 | 支 | 持 |
| 詞： | $W_1$ | | $W_2$ | $W_3$ | | $W_4$ | |
| 染色體： | 0 | 1 | 1 | 0 | 1 | 0 | |

圖 3: 最佳斷詞的染色體結構

# 4. 實驗設計與結果分析

## 4.1. 實驗設計

本研究的實驗平台是 Pentium-200 PC 上的 Windows 95 作業系統，在 Borland

C++ V4.52 的程式環境下，利用 Hunter 教授所提供的 GA Package SUGAL[Hunter, 1990]實作以遺傳演算法為基礎之中文斷詞系統。

實驗樣本取自中時電子報的 300 篇電子檔文章。並將文章分為二類：訓練文章與測試文章。其中 285 篇(95%)作為訓練文章，其餘 15 篇(5%)為測試文章。如前所述，訓練文章的目的是製作詞庫，而測試文章主要則是用來測試斷詞結果的成效。

我們採用資訊檢索上的常用的召回率(recall rate)與精確率(precision rate)作為評估斷詞正確性的依據。若令應有的正確詞集合為 H（從測試文章中以人工斷詞方式取得），系統斷詞的詞集合為 S，則系統所斷出來的結果符合於應有的正確詞的總詞數，就是系統斷出的正確詞數|H? S|。

G詞的召回率＝系統斷出的正確詞數/應有的正確詞數 = |H? S| / |H|

G詞的精確率＝系統斷出的正確詞數/系統斷詞的詞數 = |H? S| / |S|

本研究的實驗程序如圖 4 所示，基本上可分為四部分：「訓練文章」的斷句、訓練詞庫、「測試文章」的斷句、用 GA 斷詞。

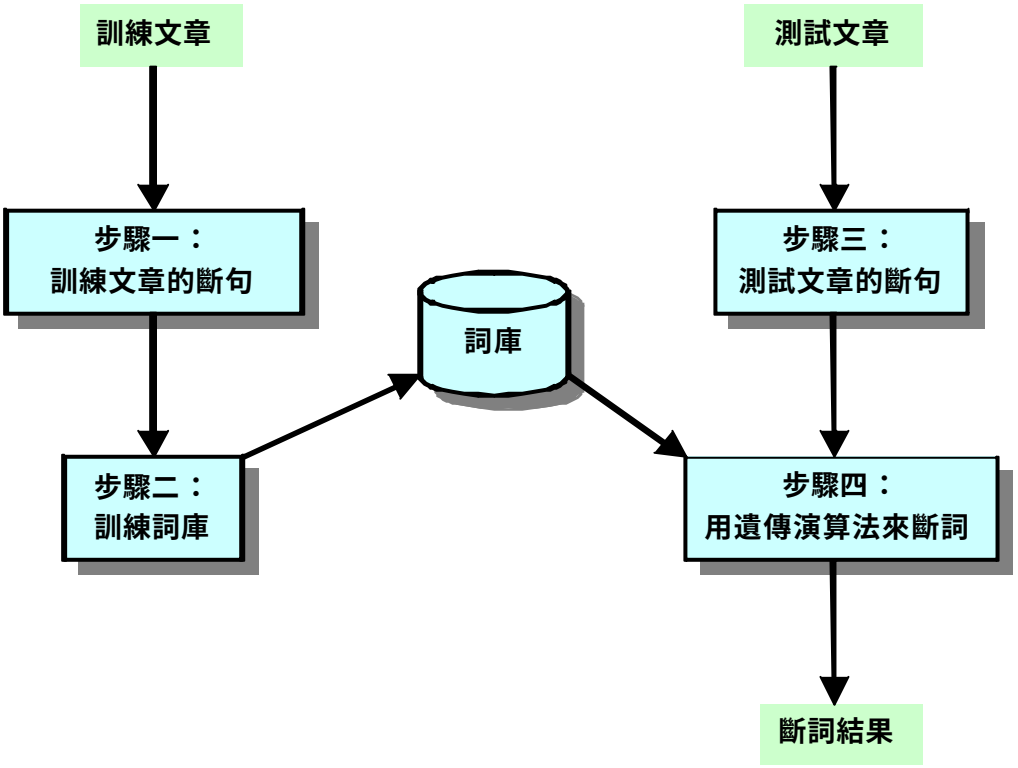


圖 4: 本研究實驗步驟

**實驗步驟一：「訓練文章」的斷句**

先除去所有「訓練文章」中的標點符號、阿拉伯數字，只留下中文字。並在連續二個刪除點間切出一連串短句。

**實驗步驟二：訓練詞庫**

利用前一步驟所產生短句製作詞庫。詞庫中包含二連字、三連字、四連字、五連字（甚至六連字）等詞的內碼、各個詞在文章中出現的累積次數和下一個儲存格的位置等資訊。

**實驗步驟三：「測試文章」的斷句**

完成詞庫後，接著整理「測試文章」以便進行隨後的斷句測試工作。首先除去「測試文章」中的標點符號、阿拉伯數字，並切出一連串純中文的短句（如實驗步驟一）。這部分的作法與實驗步驟一是相同的，只是所輸入的文章是測試結果用的而非訓練詞庫用的。

**實驗步驟四：用遺傳演算法來斷詞**

將步驟三處理過的一連串中文短句，一句句用遺傳演算法進行斷詞，並找出最佳解，並與正確解比較。實驗採用的遺傳演算法參數設定，如表 1 所示：

表 1: 遺傳演算法之相關參數設定

| 參　數 | 設 定 內 容 | 參　數 | 設 定 內 容 |
|---|---|---|---|
| 基因型態 | 二進位（0 或 1） | 評估方式 | 最大適應度 |
| 染色體長度 | 句子長度-1 | 交配方式 | 兩個切點 |
| 起始族群 | 隨機產生的染色體 | 突變機率 | 每個基因有 0.001 的機會 |
| 演化族群 | 20 個染色體 | 複製下一代的方式 | 期望值法＋菁英保留法 |
| 演化代數 | 45 代 | | |

## 4.2. 實驗結果

本實驗使用的訓練文章，每篇文章平均約 46 個句子，每句約 10 個字。285 篇訓練文章，若最多做到四連字，經實驗步驟一與步驟二後，詞庫中約有 3 萬多個相異詞。另外，本研究將測試文章分成三組，每一組有 5 篇文章。其中第一組測試文章，平均每句有 15.22 字；第二組測試文章，平均每句 14.31 字；第三組測試文章，平均每句 13.5 字。除了上述三組測試文章外，我們又從訓練文章中

挑選 5 篇文章作測試分析（令之為第四組），目的是為了比較外測與內測的結果差異。

基本上，經過演化後，最後一代中適應度最高的染色體應該就是最佳解。但有時符合系統的最佳斷詞組合，未必完全正確，所以我們決定保留適應度最高的前三名染色體當作斷詞結果。為求實驗的穩定性，實驗重複進行 10 次，穩定度為 0.7。實驗結果如表 2 所示。

表 2: 外測資料與內測資料的評估結果

| 組別 / 評估指標 | 外測資料 | | | | 內測資料 |
|---|---|---|---|---|---|
| | 第一組 | 第二組 | 第三組 | 平均 | 第四組 |
| 平均召回率 | 0.76 | 0.79 | 0.81 | 0.79 | 0.87 |
| 平均精確率 | 0.71 | 0.75 | 0.79 | 0.75 | 0.85 |
| 最高召回率 | 0.80 | 0.83 | 0.84 | 0.84 | 0.88 |
| 最高精確率 | 0.75 | 0.78 | 0.83 | 0.83 | 0.87 |

從表 2 可看出，前三組測試資料，因為並非用於訓練詞庫的文章，屬於外測資料，所以斷詞的正確性會較低，而且較不穩定。詞的平均召回率約為 79%，平均精確率約為 75%。其中最高召回率達 84%，最高精確率達 83%。

而第四組資料，由於是從訓練文章中抽取出來的，系統先前處理過，屬於內測資料，因此斷詞正確性會較高，而且較穩定。詞的平均召回率約為 87%，詞的平均精確率約為 85%。而最高的召回率更達 88%，最高的精確率則高達 87%。

總和上述實驗結果，可以看出本研究所提出的模型確實對中文斷詞有其可行性。

## 4.3. 實驗分析

為進一步瞭解遺傳演算法用於中文斷詞的情形，我們選取二個個案進行個案分析。希望透過此分析能有助於掌握本模型的運作細節，並提供日後進一步研究的改善依據。

**個案一：** 對短句「兩人並未因為日前陳水扁陣營指控新黨」進行斷詞。

**分析　：**

此句有 17 個字，因此染色體含有 16 個基因。本句斷詞的期望正解是: 兩人｜並未｜因為｜日前｜陳水扁｜陣營｜指控｜新黨。GA 找出最佳解和期望正解一樣（圖 5）。

兩 人 並 未 因 為 日 前 陳 水 扁 陣 營 指 控 新 黨

| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

染色體

圖 5: GA 對個案一的斷詞結果

我們將詞庫中與此句相關的所有斷詞組合之累積次數整理成如表 3 所示。可以很明顯發現，對子句「日前陳水扁陣營」的斷詞出現很大的歧義。儘管如此，GA 還是找到正確解。GA 找出的正確解可從表三看出，20 個染色體族群經過 45 代演化後，以第一種斷詞組合所獲得的適應度($16*2^2+353*3^2+137*2^2=3789$) 最高，所以 GA 以此為最佳解。

對此句子進行斷詞的解答空間是 $2^{16}=65536$，比 20 個染色體演化 45 代所產生過的染色體群組數目(20*45=900)大了 73 倍之多。而 GA 能在 1/73 的解答空間裡找出最佳解來，表示 GA 搜尋的效率與正確性的良好。而且對於大於二字詞的問題，即使是複合詞如人名，GA 也可將之正確的斷開來。

表 3：個案一斷詞組合之累積次數

| 斷詞 / 斷詞組合 | 兩 人 | 並 未 | 因 為 | 日 前 陳 水 扁 陣 營 | 指 控 | 新 黨 | 適應度 |
|---|---|---|---|---|---|---|---|
| 1. | 11 | 8 | 42 | 16 \| 353 \| 137 | 30 | 144 | **3789** |
| 2. | 11 | 8 | 42 | 0 \| 3 \| 137 | 30 | 144 | 597 |
| 3. | 11 | 8 | 42 | 0 \| 3 \| 49 | 30 | 144 | 469 |
| 4. | 11 | 8 | 42 | 16 \| 0 \| 45 | 30 | 144 | 785 |
| 5. | 11 | 8 | 42 | 16 \| 45 \| 0 | 30 | 144 | 785 |
| 6. | 11 | 8 | 42 | 16 \| 45 | 30 | 144 | **1189** |

**個案二：對長句「國民黨台北市長參選人馬英九與新同盟會會長許歷農昨日在一場公開活動短暫晤談」進行斷詞。**

**分析　：**

本句斷詞的期望正解是：國民黨｜台北｜市長｜參選人｜馬英九｜與｜新同盟會｜會長｜許歷農｜昨日｜在｜一場｜公開｜活動｜短暫｜晤談。

GA 找出最佳解是：「國民黨」「台北」「市長」「參選人」「馬英九」「與新同盟會」「會長」「許歷農」「昨日」「在」「一場公開」「活動」「短暫」「晤談」（圖 6）。在本個案中，GA 所斷出的詞中有二個詞（「與新同盟會」和「一場公開」）與期望正解不符。

國民黨台北市長參選人馬英九與新同盟會會長許歷農昨日在一場公開活動短暫晤談

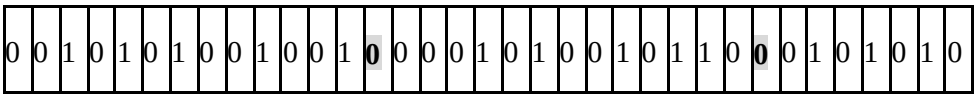


圖 6: GA 對個案二的斷詞結果

為何 GA 沒有找出最佳解來？我們分析主要原因在於此句子的解答空間是 $2^{35}$?30 億，遠大於(大約是 3000 萬倍)GA 演化 45 代所產生過的染色體群組數目(20*45=900)，染色體產生演化不完全的情形。儘管如此，GA 斷對的機率仍然高達 94%。

綜合上節整體實驗的結果與上述個案分析，我們歸納出造成斷詞錯誤的原因主要有三：(1) 句子太長，(2) 詞庫錯誤，(3) 未知詞。

系統在長句的表現上比短句差，主要是因為句子越長，字與字之間的關係就越複雜，或句中的歧義狀況就越多，因此系統斷詞出錯的機會便相對的增加。另外，因為句子越長，其解答空間也越大。如果測試句子的解答空間遠大於 GA 的搜尋空間，染色體演化不完全的情形就很可能會發生，此時找出的最佳解可能就不是最好的斷詞方式。在我們的實驗中有 50%的錯誤是因為句子太長所致。

適應函數是根據詞長與詞庫中詞的累積次數兩個資訊導引斷詞，而詞庫的來源是經由訓練文章轉換而來，因此詞的累積次數完全依靠訓練文章而得。當測試句子裡的詞，以及詞庫中的詞內聚力不同時，就會有斷詞斷錯的情形。特別是當訓練樣本內出現相似的詞，但句子結構有所不同時，這種干擾會更大。

另外，因為詞庫中詞的累積次數是用 N 連字的方式製作，因此會出現斷詞次數高估的錯誤。例如，「市長參選人馬英九」，正確的斷詞方式是：「市長｜參選人｜馬英九」，但因為在訓練文章中除了『候選人』，還有『候選人是』、『候選人馬』等詞出現次數頻繁，使得『選人』二字的次數 305 遠大於『參選人』的次數 93，導致在適應函數的評估上，『參選人』就會被斷開了成「參｜選人」，整個句子就會斷成「市長｜參｜選人｜馬英九」。在我們的實驗中有 30%的錯誤源於詞庫資訊錯誤。

雖然在適應函數的設計上，略有考慮到未知詞的處理，但是，對於連續兩個或兩個以上的未知詞相鄰時，系統可能無法將這些未知詞斷開。例如，「台北市政府門口請願區集合出發」中的「門口」與「請願區」，這兩詞在詞庫中並沒有紀錄（在詞庫訓練時只出現過一次，因此未納入詞庫），因此形成兩個相連未知

詞。理論上這句話應該斷詞為：「台北市政府｜門口｜請願區｜集合｜出發」，但因出現相連未知詞，導致 GA 將之斷詞成「台北市政府｜門口請願區｜集合｜出發」。在我們的實驗中約 20%的錯誤源於出現此類未知詞。

從另一個角度，由於 GA 搜尋能力與適應函數有很大的關係。若適應函數設計得當，確實涵蓋問題的解答方向，則 GA 就可以在解答空間裡搜尋出最佳解，甚至有很高的搜尋效率。如果適應函數設計不當，使得無法演化出適當的染色體，終將無法達成預期的最佳解。我們利用第四組測試資料評估 GA 的搜尋能力，結果顯示約有 77%的句子，具有最高適應度的染色體，所對應的正是最佳斷詞，由此可見本模式所提供的適應函數是可以被接受的。

# 5. 結論

本研究所提的斷詞模型，可自行從文章中產生詞庫，其優點為可避免人為介入的不客觀性，也可避免浪費寶貴的人力資源。而運用遺傳演算法的斷詞方式，亦有助於減少人為的錯誤判定、且有更自由的空間去尋找問題的答案。

一般斷詞方法，在短詞上的效果不錯，但大都侷限於二字詞或三字詞，一旦遇到長詞，正確率就會大幅下降。但若是考慮長詞優先，如果較長的詞包含了較短的詞，則較短的詞就有可能會斷不出來。而本研究所提出的斷詞模型，則可以減少上述的困擾。從模型提供的適應函數看，基本上較長的詞有較大的機會被優先斷出，具備長詞優先的優點；但卻沒有長詞優先的缺點（無法斷出短詞），事實上任何短詞只要在文章中出現的次數夠多，還是有機會被斷出，次數越高，機會越大。這是因為適應函數中還有另一個主要的因子：詞的累積次數。另外，如果訓練詞庫時，我們將詞長的上限設為 M 字詞，則基本上系統最多也可以斷出 M 字詞。運用本研究所提斷詞模型斷詞最大的特色是，根本不需考慮長詞或是短詞優先的問題，也不需要考慮詞性的問題，遺傳演算法會主動依詞長和語詞出現的次數，找出好的斷詞結果。

一般傳統的斷詞法只能找出一組斷詞組合，但實際上在交予後階段的語法、語意分析時，所謂「最好的」斷詞方式，在後處理階段其結果可能並不理想。本研究用遺傳演算法斷詞的特點是，可以同時找出多個斷詞結果，供後續中文自然語言處理之用。我們可以將傳演算法找出的斷詞結果，保留最好的前三名，如此，讓後階段的自然語言處理，可以有不只一種的選擇，進而有更好的處理結果。除了打破傳統的中文斷詞方式外，這種可以處理非絕對唯一解的特性，正適合處理自然語言的現象。

許多時候，因為不知句子前後文為何，因此同一個句子會出現多個斷詞組合都是合理的斷法。在前節的實驗中，為了評估系統成效，我們必須給系統一個所謂的期望正解，但基於可能存在不只一種合理斷詞的事實，本模式的斷詞成效，其績效事實上應該是大於上節的評估結果。

本研究模型對於未知詞並未多做處理，是未來可以再加強的部分。而在適應函數的設計上，也可以再做調整。因為適應函數的適當與否，對遺傳演算法的演

化結果，有著很重要的影響。所以要改善斷詞結果，應該也可以從改善適應函數設計的方向著手。其中包括參數的修正，或增加評估的條件，例如加上「詞在眾文章中的廣度」，藉以提高適應函數的可信度。

由於本研究的詞庫是以中文內碼來當作儲存的根據，因此無法分辨同字不同字形的中文字。例如，「台」與「臺」、「体」與「體」、還有「群」的「君」可以在「羊」的左邊也可以在上面、「峰」的「山」可以在「夆」的左邊也可以在上面。因內碼不同，故本研究模型將上述同字不同形的字皆視為相異的二個字，導致各個語詞的累積次數比實際小。日後如果可以多加一個異形字對照表，應可以讓斷詞系統更加精準。

詞庫是以連字的方式來累計詞的出現次數，因此會有短詞次數高估的現象，並導致斷詞錯誤。所以如何改進詞庫的設計，或如何再加入更多的訊息以供斷詞系統參考都是未來可以努力的地方。另外，我們的模型和一般演算法一樣，通常都是建立在「先完成斷詞處理後，才進行下一階段中文處理」的基礎上。未來可朝利用遺傳演算法同時進行語法或語意分析工作和斷詞作業，以提高正確率及應用價值。

# 參考文獻


[1] 王良志、貝子勝、黎偉權、黃麗卿，「以剖析為導向的中文斷詞法」，電子發展月刊，163 期，頁 40-45，民 80 年。

[2] 范長康、蔡文祥，「以鬆弛法作中文斷詞」，全國計算機會議論文集，頁 423-431，民 76 年。

[3] 許菱祥，「中文文法」，大中國圖書公司，民 75 年。

[4] 陳克建、陳正佳、林隆基，「中文語句分析的研究-斷詞與構詞」，中央研究院資訊所技術報告，TR86-004，民 75 年。

[5] 陳永德，「中文斷詞中長詞優先、詞頻比對與前詞優先規則之使用」，國立台灣大學心理學研究所博士論文，民 86 年。

[6] 黃大一，「中文字碼」，長松文化事業有限公司，民 79 年。

[7] Chen K. J. And S. H. Liu, “Word Identification for Mandarin Chinese Sentences,” *Proceeding of COLING-92, 14th Int. Conf. On Computational Linguistics*, pp. 101-107, 1992.

[8] Fan, C. K. and W. H. Tsai, “Automatic Word Identification in Chinese Sentences by the Relaxation Technique,” Computer Processing of Chinese and Oriental Languages, Vol. 2, No. 4, pp. 33-56, 1988.

[9] Gan, K. W., M. Palmer and K. T. Lua, “A Statistically Emergent approach for language Processing: Application to Modeling Context Effects in Ambiguous Chinese Word Boundary Perception,” Computational Linguistics, pp. 531-553, 1996.

[10] Goldberg, David E. *Genetic Algorithms in Search Optimization and Machine Learning*, Addison Wesley, 1989.

[11] Holland, J. H. *Adaptation in natural and artificial systems*. The University of Michigan Press, Ann Arbor, 1975.

[12] Hunter, A. SUGAL, http://osiris.sounderland.ac.uk/ahu/sugal/home.htmlSUGAL University of Sunderland, England, 1990.

[13] Li, G. C., K. Y. Liu and Y. K. Zhang, "Identifying Chinese Word and Processing Different Meaning Structures," Journal of Chinese Information Processing, Vol. 2, pp. 45-53, 1988.

[14] Liang, N. Y. "Knowledge of Chinese Word Segmentation," Journal of Chinese Information Processing, Vol. 4, pp. 42-49, 1990.

[15] Leung, C. H. and W. K. Kan, "A Statistical Learning Approach to Improving the Accuracy of Chinese Word Segmentation," Literary and Linguistic Computing, pp. 87-92, 1996.

[16] Nic, J. Y. and M. Briscobois, "On Chinese Text Retrieval," *Proceeding of SIGIR*, pp. 225-233, 1996.

[17] Nie, J. Y., M. L. Hannan and W. Jin, "Combining Dictionary, Rules and Statistical Information in Segmentation of Chinese," Computer Processing of Chinese and Oriental Languages, Vol. 9, pp. 125-143, 1995.

[18] Robert, L. K., P. L. Bruce and L. T. Clovis, *Data Structures and Program Design in C*, Prentice Hall, 1993.

[19] Sporat R. and C. Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," Computer Processing of Chinese and Oriental Languages, Vol. 4, No. 4, pp. 336-351, 1990.

[20] Sporat R., C. Shih W. Gale and N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chines," Computational Linguistics, Vol. 22, pp. 377-404, 1996.

[21] Wu, Z. and G. Tseng, "Chinese Text Segmentation for Text Retrieval: Achievement and Problems", Journal of the American Society for Information Science, Vol. 44, No. 9, pp. 532-542, 1993.

[22] Wu, Z. and G. Tseng, "ACTS: An Automatic Chinese Text Segmentation System for Full Text Retrieval," Journal of the American Society for Information Science, Vol. 46, pp. 83-96, 1995.

[23] Yeh C. L. and H. J. Lee, "Rule-Based Word Identification for Mandarin Chinese Sentences-A Unification Approach," Computer Processing of Chinese and Oriental Languages, Vol. 5, No. 2, pp. 97-118, 1991.