國　立　交　通　大　學

電信工程學系碩士班

碩士論文

中文斷詞器之改進
An improvement on Chinese Parser

研究生：江振宇

指導教授：陳信宏　博士

中華民國九十三年七月

中文斷詞器之改進

An Improvement on Chinese Parser

研 究 生：江振宇　　Student： Chen-Yu Chiang

指導教授：陳信宏 博士　　Advisor： Dr. Sin-Horng Chen

A Thesis
Submitted to Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
In Partial Fulfillment of the Requirements
For the Degree of
Master of Science
In
Electrical Engineering
July, 2004
Hsinchu, Taiwan, Republic of China

中華民國九十三年七月

中文斷詞器之改進

研究生：江振宇　　指導教授：陳信宏 博士

國立交通大學電信工程學系碩士班

中文摘要

在本論文中，我們設計了中文斷詞器的基本架構，並實現了此中文斷詞器，以模組化的設計方法，使得整個斷詞器的架構更加系統化，可以成為一個語音合成系統的軟體開發元件，改善了先前中文斷詞器的架構問題。整個最核心的斷詞單元，採用規則法斷詞，並使用詞典樹增加詞典比對速度。構詞單元，我們採用中研院提供的構詞規則以及自行整理出之規則應用，並使構詞單元之程式處理效率最佳化。對於特殊符號的語音讀法，我們設計了文字正規化單元，解決特殊符號的讀法問題。為了瞭解斷詞器之性能，我們以〈中研院平衡語料庫 3.0 版〉做為測試語料，測試結果顯示斷詞的召回率達到 0.78，精確率達到 0.87，而詞類標記的精確率可以達到 0.96。最後我們分析本斷詞器之斷詞結果，探討斷詞錯誤之可能更正方法。

關鍵字：中文斷詞器、語音合成、斷詞單元、構詞單元、詞類標記、文字正規化

An Improvement on Chinese Parser

Student: Chen Yu Chiang Advisor: Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

In this thesis, a Chinese word tagger for text-to-speech (TTS) is implemented. It contains four basic modules. They are word identification module, word combination module, POS (part of speech) tagging module, and text normalization module. In word identification module, we adopt a word matching algorithm with 6 heuristic rules proposed by the Chinese Knowledge Information Processing group (CKIP), Academia Sinica, to identify words from input Chinese character string. The word combination module groups words into compounds using 95 determinative-measure (DM) compound rules and 10 reduplication rules. The POS tagging module gives POS tags to words identified by the word identification module. To transform from written form to spoken form, we design the text normalization module. Lastly, the Sinica Corpus published by CKIP is used to evaluate the performance of our system. We achieve a recall rate of 0.78, a precision rate of 0.87 in word identification, and a precision rate of 0.96 in POS tagging. We also analyze word identification results to give advices in future works.

Keywords: Chinese word tagger, Text-to-Speech, Word identification, Compound words, POS tagging, Text normalization

誌謝

首先，非常感謝指導教授陳信宏老師兩年來的指導，指引我研究的方向以及做事情的方法，也十分感謝王逸如老師、廖元甫老師平日給我之教導。雖然碩士班僅僅兩年，但對我來說卻是學習的精華時刻，好像把小學、中學、高中、大學濃縮了一樣，充實極了！

實驗室裡總是一片歡愉；輝哥、郭威志學長、傅振宏學長，不厭其煩的讓我問問題，「我兒阿樹」、「我兒阿樹的叔叔俊良哥」、「苦瓜的孫諺哥(小太陽)」、「同窗六年的小 Z」、「我高中同學的鄰居嘉俊」、「跟我一樣是食神迷的阿德」、「超乖的祺翰」、「欺負我的學弟金翰」、「來打 CS 的 LUBO」、「鋼琴王子希群」、「長的超像劉文聰的張隆勳」、「HTK 很強的順哥」、「飛天小女警佩穎」，謝謝你們陪伴我，讓你們忍受我兩年來持續的冷笑話。

最後，我要感謝常叮嚀我不要熬夜的家人，沒有你們 24 年來的支持，不會有現在的我，另外也要感謝室友羅正倫、林宏曄、高中死黨彭康硯、學弟繼堯、一直是大學實驗課夥伴的姿斐。寫到這裡，暫時停筆，應該是人生的轉換點到了吧！還等什麼？

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

中文斷詞器應用的範圍很廣，包括資訊擷取、自然語言處理、中文文句翻語音系統(Text-To-Speech system，簡稱 TTS)等方面。本論文研究的中文斷詞器最主要應用在 TTS system，而一個最基本的 TTS system 有三大模組：文句分析器、韻律訊息產生器以及語音合成器，如圖 1-1 所示

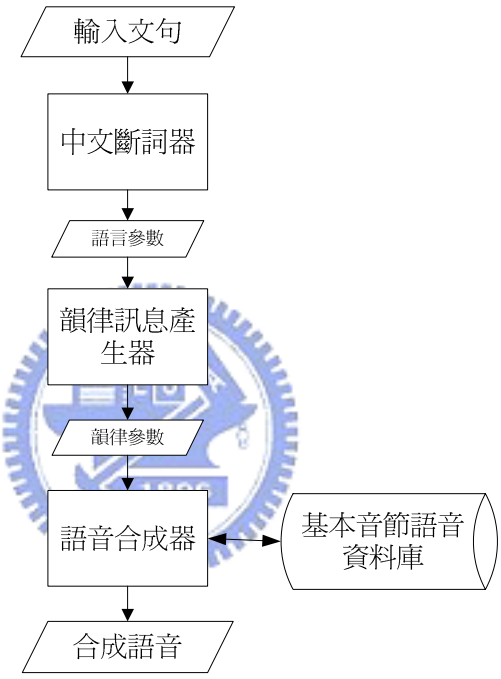


圖1-1 TTS system

中文斷詞器輸出的為語言參數，這些語言參數包括詞、詞長、詞類、基本音節資訊等，韻律訊息產生器利用這些語言參數產生語音合成所需要的韻律參數，這些韻律參數是基頻軌跡(pitch contour)、音長(duration)以及音量(energy)，最後語音合成器用 PSOLA 的技術，依據韻律參數將基本音節語音調整合成出語音。

顯而易見地，中文斷詞器為整個 TTS system 的最前面一級，如果斷詞出來的結果不正確，將會影響合成的結果甚深，造成合成之韻律不佳，所以如何改進斷詞之錯誤，是我們的目標。再者，由於中文之語法複雜、詞的集合可以說是無

限，要建構不同應用領域的 TTS system，斷詞器的詞典需要時常更新，構詞規則也要有所增加，許多輸入的特殊符號必須轉為語音讀法，另外，面對在未來要建構的 Corpus-Based TTS system，以及語言學和韻律之間關係的研究，一個可以作為研究用基本雛形的中文斷詞器是必要的，因此，考慮到實用性以及研究功能的中文斷詞器設計，也是我們希望達到的目標。

1.2 研究方向

本論文之研究重點，在於設計一套模組化、兼具實用性以及研究功能的中文斷詞器，基於這樣的設計之下，分析斷詞之錯誤，研究如何解決這些斷詞錯誤。首先，我們採用中研院提出的六條斷詞規則【1】，並以詞典樹的資料結構儲存詞典，設計了基本的斷詞單元，對於定量複合詞以及重疊詞，我們參考中研院提出的構詞規則【2,3】以及唐大任【4】，設計了構詞單元，另外，以語音合成為考量，為了解決輸入文句中特殊符號的讀法，設計了文字正規化單元。最後，我們以〈中研院平衡語料庫 3.0 版〉做為詞類標記單元的雙連文模型以及斷詞測試的訓練/測試語料庫，給予錯誤之斷詞詳細的分析，提出可能的解決辦法。

1.3 章節概要

第一章 緒論：介紹本論文之研究動機與方向。

第二章 中文斷詞器概述：說明中文斷詞器設計的基本架構，且定義分詞之標準，給予整個斷詞器設計之指南。

第三章 中文斷詞器之設計：介紹斷詞單元、構詞單元、詞類標記單元、文字正規化單元的原理及設計方法。

第四章 實驗結果與分析：以〈中研院平衡語料庫 3.0 版〉做為斷詞的測試語料，對於錯誤之斷詞給予分析且提出可能之解決辦法。

第五章 結論與未來展望。

第二章 中文斷詞器概述

2.1 分詞標準

這裡提出的中文斷詞器，最主要是以語音合成的角度來設計，希望斷出的詞適合於語音合成的單位，但由於中文詞的定義較模糊，我們首要的工作便是定義分詞的單位，也就是訂定「分詞標準」；對於資訊處理而言根據中研院【5】:「詞為一個具有獨立意義，且扮演特定語法功能的字串應視為一個詞」，然而對於語音合成而言，會有不同的分詞單位，例如「一個球隊」，以資訊處理的標準，會斷為「一個」、「球隊」，「一個」含有了球隊的數目資訊，「球隊」含有了某個事物的名稱資訊，但對於語音合成，「一個球隊」應該屬於一個分詞單位，可以由以資訊處理標準斷詞的結果「一個」、「球隊」再構詞成一個語音上的分詞單位。

由上面的例子可以知道，如果要達到以語音合成考量的斷詞結果，可以利用「資訊處理標準」的斷詞結果再經由一些規則，將斷詞結果提升到適用於語音合成，所以這裡提出的中文斷詞器，將斷詞處理依據不同的「分詞標準」分為不同的前後級處理，分述如下：

第一級：由「構詞單元」及「斷詞單元」構成，詞集合為「詞典」以及「構詞單元產生的詞」，由此級斷出的詞希望能達到「資訊處理」的標準，使得斷詞結果含有充分的語法和語意資訊，以供後級利用。

第二級：由「後置構詞單元」及「未知詞構詞單元」組成。由於第一級產生的斷詞結果含有充分的資訊，第二級可以利用這些資訊再加上一些規則或統計資訊，把衍生詞或未知詞(人名、專有名詞等)斷出。

第三級：就是「以語音合成考量的構詞單元」。這一級的「分詞標準」是以語音合成為考量，利用前兩級所給予的資訊，將語音合成上為一個分詞單為的詞結合出。

2.2 中文斷詞器系統架構

目前已完成的系統模組已達到「第一級的分詞標準」，系統概述如下：

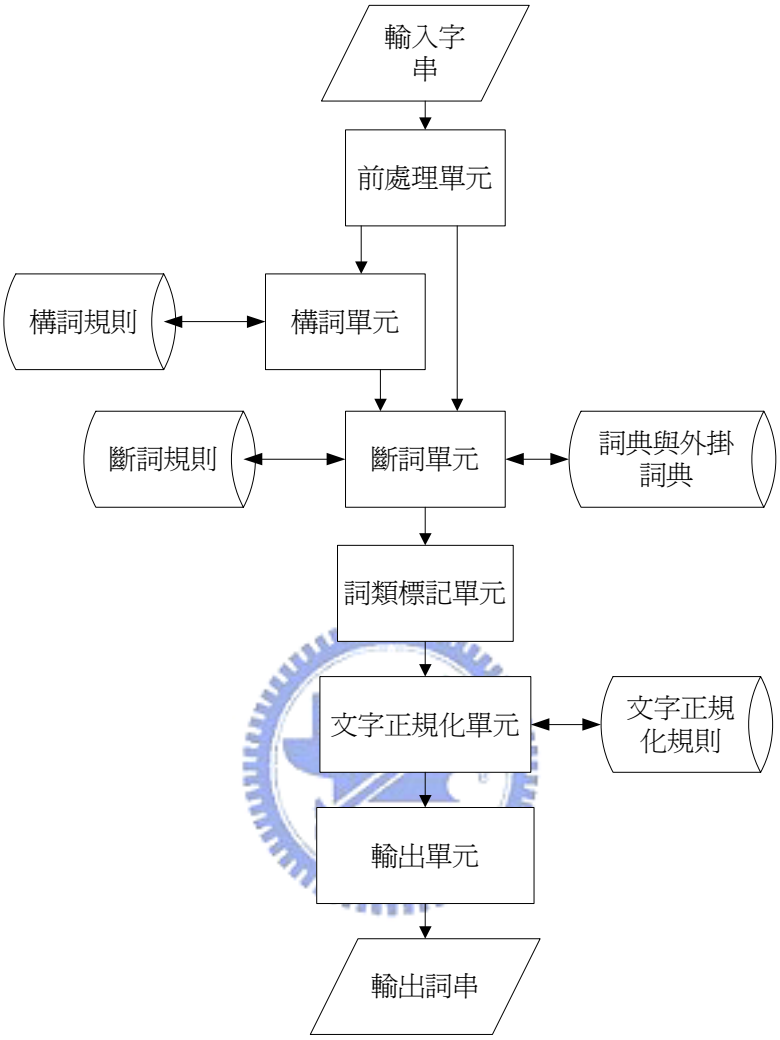


圖2-1:中文斷詞器架構圖

(1)前處理單元：

由於輸入的字串裡可能含有 ASCII code 或 Big-5 code，為了使系統處理的字串格式統一，在進入斷詞器之前，首先我們必須把所有的 ASCII code 轉為 Big-5 code，例如字串「空屋率從 10.11% 下降至 9.53% 。End」會先轉為 Big-5 字串「空屋率從１０．１１％下降至９．５３％。Ｅｎｄ」。

(2)斷詞單元：

這個單元是中文斷詞器最核心的部分，目的是將輸入字串做適當的斷詞。此

單元分為兩大步驟(1)建立候選詞組(2)挑選候選詞組。建立候選詞組是將輸入字串利用詞典和構詞規則以 Matching Algorithm 找出所有可能的詞串組合，然而這些詞串組合之中只有一個是含有適當斷詞結果的詞串，所以必須經由一些規則或是統計的方法，挑選候選詞組，斷出最適當的詞。由於「斷詞單元」為最前面的分詞單元，接下來的「詞類標記單元」、「文字正規化單元」，以及未來的「後置構詞單元」及「未知詞構詞單元」都是利用此單元的斷詞結果再做構詞。

(3)詞典及外掛詞典

由於中文斷詞的方法是將輸入文句和詞典做比對，詞典的好壞會影響「斷詞單元」的斷詞結果。目前詞典的詞數統計如下表 2-1。

	詞數
一字詞	13110
二字詞	64894
三字詞	26043
四字詞	16066
五字詞	999
六字詞	155
七字詞	65
八字詞	9
	總計 121,341

表 2-1:詞典統計表

另外，當遇到特定使用環境的時候，會遇到輸入文句中有很多的特殊詞，我們可以利用增加這些詞於外掛詞典，解決特殊詞的斷詞。

(4)構詞單元

由於中文斷詞的方法是將輸入文句和詞典做比對，但是要將所有可能的詞列於詞典之中是不可能的，這些不能詳列於詞典的詞像是定量複合詞及重疊詞等，它們的組成是有規律的，可以由輸入的文字串中經由構詞規則結合出來，這樣子構出的詞就相當於比對詞典的動作。由此模組構出的詞，會留下其構詞的結構，

以便後級模組使用(如文字正規化模組)。

(5)詞類標記單元

此單元將利用機率統計模型，給予斷出詞對應的詞類標記(Part of Speech, POS)。

(6)文字正規化單元

以語音合成為考量，在輸入文句之中，有些阿拉伯數字、詞或符號必須由寫法轉為語音讀法，這個過程稱為文字正規化，此模組利用斷詞單元及構詞單元留下的詞結構來進行文字正規化，舉例來說「９０％」應該讀為「百分之九十」。

(7)輸出單元

這個單元輸出韻律產生器需要的語言參數：詞、詞類及音碼。

第三章 中文斷詞器之設計

3.1 斷詞單元

斷詞單元是中文斷詞器最核心的部分，是整個中文斷詞器第一級的分詞單元，目的是將輸入字串做適當的斷詞，達到「資訊處理」的標準，詞集合為「詞典」以及「構詞單元產生的詞」。在這裡首先說明斷詞單元的工作原理，再說明斷詞單元的設計。

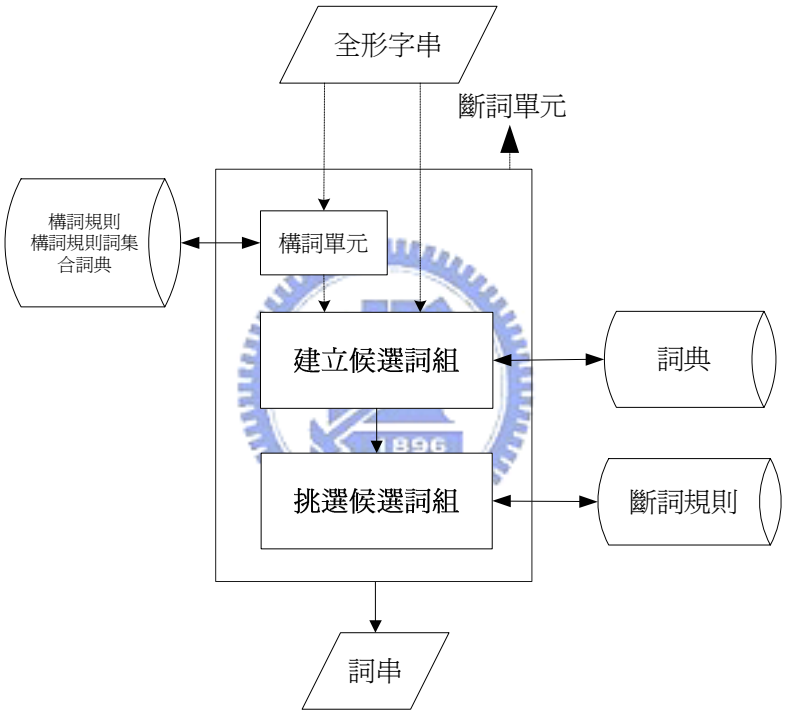


圖3-1-1 斷詞單元工作流程圖1

3.1.1 斷詞單元的工作原理

斷詞單元的工作分為兩大步驟(1)建立候選詞組(2)挑選候選詞組，簡單來說。建立候選詞組就是把輸入字串與詞典做比對的動作(word matching)，若輸入字串中有詞典裡的詞，便把字串斷為一個個的詞，然而斷出的詞串可能不只一個，造成斷詞混淆的現象，由於所有可能的詞組之中，只有一組是含有適當斷詞

結果的詞串，我們依據斷詞規則挑選候選詞組，解決斷詞混淆的現象。詳細的工作方法詳述於下：

(1)建立候選詞組

首先從輸入字串的第一個字開始，連續組成三個詞，成為一個候選詞組，並將所有可能的候選詞組找出來，且候選詞組中的詞必須是(a)詞典中的詞或(b)由構詞單元構出的詞。例如：【高高興興地唱歌】這字串，可建立如下五組候選詞組：

【高　高　興】　【高　高興　興】　【高高　興　興】

【高高興興　地　唱】　【高高興興　地　唱歌】

若是字串中的字，不足以挑選三個詞來構成詞組，則將不足的部分，其詞長視為零，來構成詞組。如【高興】這例，不足以構成詞組的有【高　興】與【高興】，因此視【高　興】之後的一個詞，其詞長為零，使之能構成一個詞組，視【高興】之後的兩個詞，其詞長為零，使之能構成一個詞組。

由於構詞單元構出的詞可能與詞典詞重複，例如：「十二萬分」可由構詞單元中構出，同時在詞典中也存有「十二萬分」，我們以詞典中的詞優先選擇，不以構詞單元的詞為詞組中的詞，因為這些存於詞典中詞通常有特別的語法及語意資訊，或是有多的詞類，我們必須保留這些資訊給予後級處理。整個建立候選詞組的流程圖見附錄 1。

(2)挑選候選詞組

建立所有候選詞組後，依序由斷詞規則一開始，一條經一條規則，來挑選候選詞組，一旦剩下一個詞組，則不必經由以下的規則處理，直接取此詞組的**第一個詞**，輸出至後置構詞單元。根據中研院詞庫小組提出的 6 條斷詞規則【1】，在不考慮專有名詞與衍生詞所造成的錯誤，其正確率可達 99%左右，因此我們採

用這 6 條斷詞規則並修改第一條斷詞規則，幫助斷詞單元斷詞，以下依序說明這些規則：

(1) 斷詞規則一：長詞優先

僅保留候選詞組中詞長和為最大，或詞組中詞長為零的個數最多的詞組。若滿足斷詞規則一的詞組數不只一個，則將這些候選詞組，留給斷詞規則二處理；相反地，若只有一個詞組滿足斷詞規則一，則輸出此詞組的第一個詞為斷詞結果。

例如：將【有一張桌子】輸入斷詞器，會產生下列三個候選詞組，

(a) 有　一　張

(b) 有　一張　桌

(c) 有　一張　桌子

根據此斷詞規則，候選詞組(c)具有最大的詞長和(五個字)，所以選擇詞組(c)。

(2) 斷詞規則二：標準差小的優先

由斷詞規則一處理後的候選詞組中，選擇詞長具有最小標準差的詞組，相當於選擇 V(C)值為最小的詞組，若僅有一詞組其詞長具有最小標準差，則選擇此詞組的第一個詞輸出為斷詞結果；反之，具有最小標準差的詞組不只一個，則將這些有相同最小標準差的候選詞組，繼續進行下面的斷詞規則。

定義：$V(C)=(L(W_1)-Mean)^2+(L(W_2)-Mean)^2+(L(W_3)-Mean)^2$

其中　W_i：　第 i 個詞

$L(W_i)$：　第i個詞的詞長

$$Mean=\frac{1}{3}\sum_{i=1}^{3}L(Wi)$$

例如：將【研究生命起源】輸入斷詞器，經斷詞規則一處理後，餘下(a)(b)兩個候選詞組，

(a) 研究　生命　起源

(b) 研究生　命　起源

其中候選詞組(a)的 V(C)值為 0，比候選詞組(b)的 V(C)值 2 還小，根據此斷詞規則，選擇(a)詞組。

(3) 斷詞規則三：附著語素最少者優先

附著語素為**很少單獨出現在語句中的一字詞【4】**，為了使斷出來的詞含有較少的一字詞，所以我們可選取含有較少附著語素的詞組為優先，來減少斷出一字詞的機率。

例如：將【協調上手續】輸入斷詞器，經斷詞規則一、二處理後，餘下(a)(b)兩個候選詞組，

(a) 協調　上　手續

(b) 協調　上手　續

其中候選詞組(a)中的【上】不是附著語素，而候選詞組(b)中的【續】為附著語素，根據此斷詞規則，所以選擇(a)詞組。

(4) 斷詞規則四：候選詞組中定量複合詞字數合最少者優先

例如：將【機身重三十噸】輸入斷詞器，經前三條斷詞規則處理後，餘下(a)(b)兩個候選詞組，

(a) 機身　重　三十噸

(b) 機　身重　三十噸

其中候選詞組(a)中的【三十噸】為構出的定量複合詞，字數為三個字，而

候選詞組(b)中的定量複合詞，則有【三十噸】和【身重】，字數為五個字，根據此斷詞規則選擇(a)詞組。

(5) 斷詞規則五：一字詞詞頻最高者優先

經由前四條斷詞規則處理後，若餘下的詞組中均僅有一個一字詞，我們則優先選取具有最高一字詞頻的詞組。

例如：將【傑出的人才】輸入斷詞器，經第前四條斷詞規則處理後，餘下(a)(b)兩個候選詞組，

(a) 傑出　的　人才

(b) 傑出　的人　才

其中候選詞組(a)中的【的】的詞頻，比候選詞組(b)中的【才】的詞頻還高，根據此斷詞規則選擇(a)詞組。

(6) 斷詞規則六：總詞頻最高者優先

選擇此三個詞的總詞頻和為最高的詞組，相當於選擇 F(C)值為最高的詞組。

定義：$F(C) = \log(F(W_1)) + \log(F(W_2)) + \log(F(W_3))$

其中　W_i：　第 i 個詞

$F(W_i)$：　第i個詞的詞頻

例如：將【總辦事處秘書組主任】輸入斷詞器，經第前五條斷詞規則處理後，餘下(a)(b)兩個候選詞組，

(a) 總辦事處　秘書組　主任

(b) 總辦事處　秘書　組主任

其中候選詞組(a)的總詞頻和比候選詞組(b)的總詞頻和還高，根據此斷詞規則選擇(a)詞組。

3.1.2 斷詞單元的設計

在這裡我們依據斷詞單元的工作原理，設計此單元，圖 3-1-3 為斷詞單元的工作流程圖，以下為分項說明：

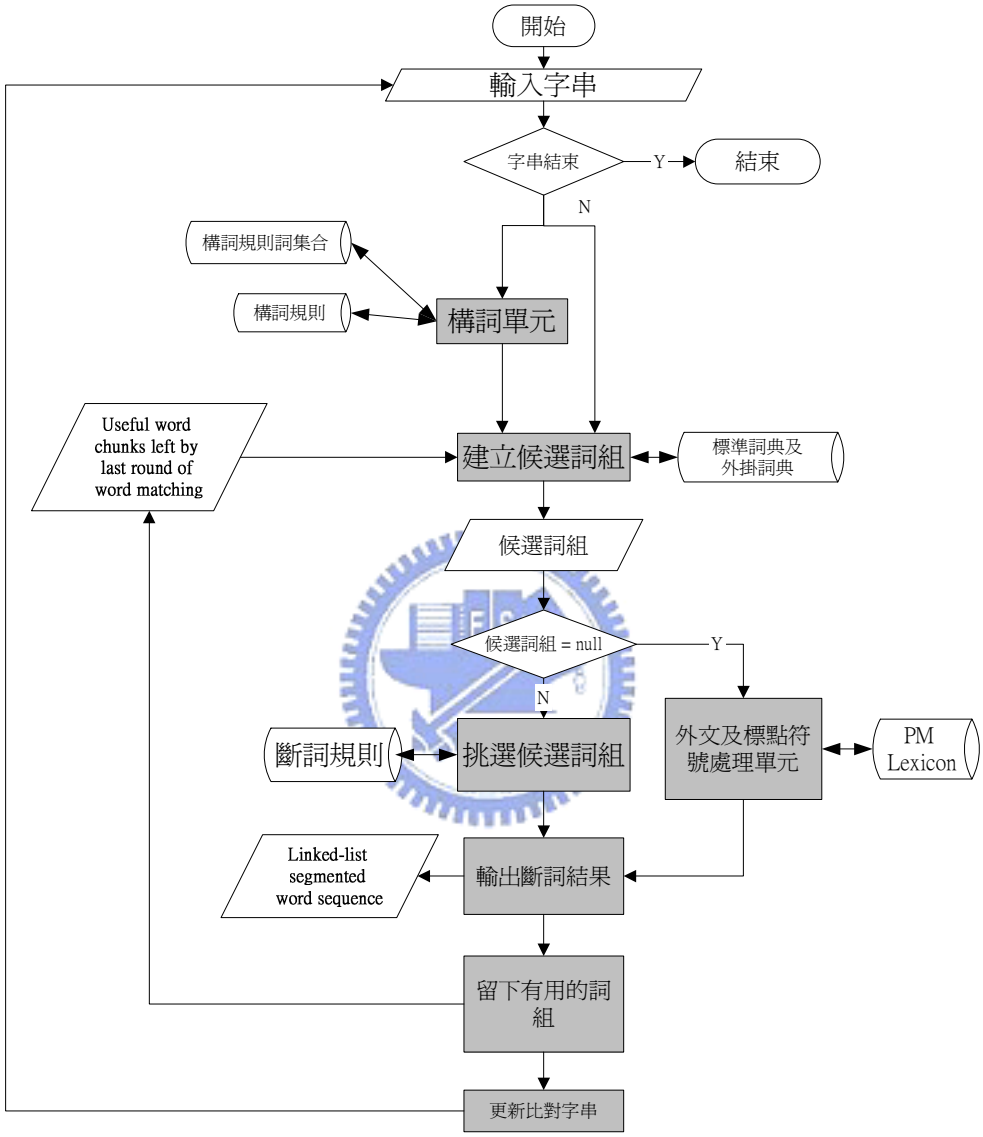


圖3-1-2: 斷詞單元工作流程圖2

(1)輸入字串

輸入字串必須為經過前處裡的全形字串。

(2)構詞單元

構詞單元將不能列舉於詞典中且合乎構詞規則的詞，由輸入字串中構出，這

些詞是「定量複合詞」、「數詞定詞」、「數量定詞」、「時間詞」、「地方詞」及「重疊詞」，這些構出的詞，同樣與詞典中的詞為建立候選詞組的詞集合。構詞單元詳細的內容請參閱「3.2 構詞單元」的說明。

(3)詞典及外掛詞典

詞典及外掛詞典為建立候選詞組的詞集合。詞典裡的詞為最基本的詞集合，裡面包含了詞、音碼、詞頻、詞類的資訊。外掛詞典是解決當遇到特定使用環境的時候，會遇到輸入文句中有很多的特殊詞，我們可以利用增加這些詞於外掛詞典，解決特殊詞的斷詞。

(4)建立候選詞組

建立候選詞組為斷詞最主要的第一個步驟，輸入為字串，輸出為所有可能的候選詞組，工作方法如上一小節「斷詞單元的工作原理」所述。設計上為了節省輸入字串與詞典比對的時間，這裡是利用詞典樹的方法將輸入字串與詞典比對，詞典樹請參閱附錄 2。另外由於這些產生的候選詞組數目不一定，我們將它們以動態記憶體的資料結構方法儲存在記憶體中，如圖 3-1-3 所示。

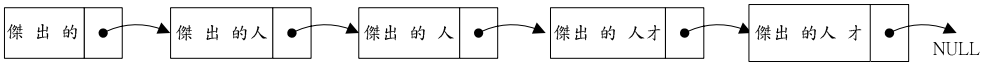


圖3-1-3: 候選詞組的資料結構

在這裡另外要注意的是當輸出的候選詞組為空集合時，代表輸入字串為「標點符號」、「外文」或「特殊符號」，它們將送到「外文及標點符號處理單元」處理。

(5)挑選候選詞組

挑選候選詞組為斷詞最主要的第二個步驟，是將適當斷詞結果由候選詞組中找出來，工作原理如上一小節「斷詞單元的工作原理」所述，但在實際設計斷詞單元時，我們要將中研院提出的六條斷詞規則，進行部份的修改及增訂，以增加斷

詞的效率，這些規則是 Rule1：長詞優先 Rule2：標準差小的優先 Rule3：附著語素最少者優先 Rule4：候選詞組中定量複合詞字數合最少者優先 Rule5：一字詞詞頻最高者優先 Rule6：總詞頻最高者優先 Rule7：任意選擇候選詞組

Rule 1～6 為中研院提出原本的斷詞規則，我們稍微對這些規則做了修改，舉輸入串為「傑出的人才」例子說明：

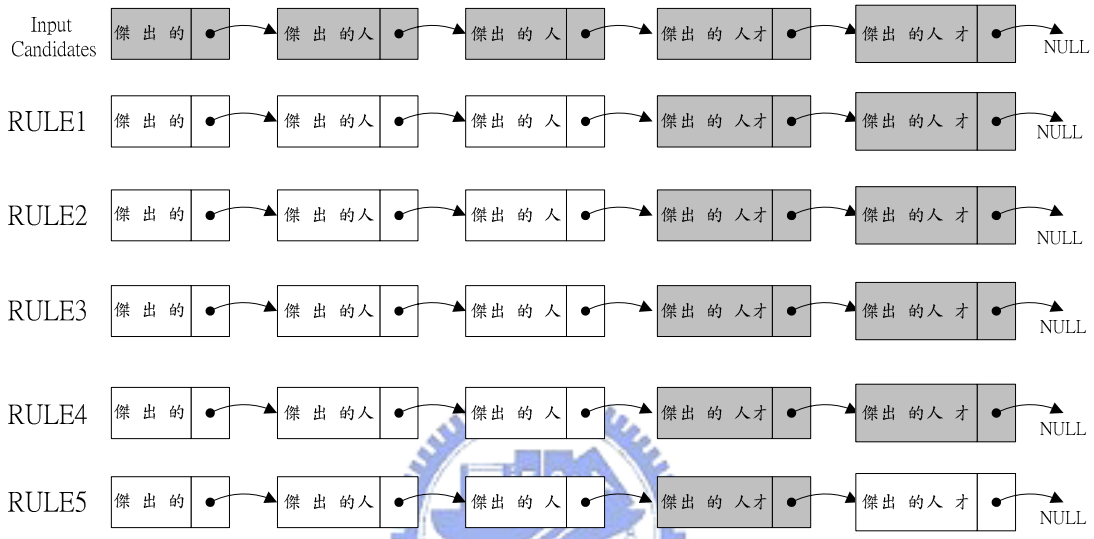


圖3-1-4: 挑選候選詞組示意圖

圖 3-1-4 中的灰色底的候選詞組，代表符合某條斷詞規則的詞組，「傑出的人才」的例子，我們很清楚的知道，其實詞「傑出」並不用做到「斷詞規則五：一字詞詞頻最高者優先」才能把詞給斷出，由於斷詞結果是選擇詞組中的第一個詞，在經過「斷詞規則一」後，詞組「傑出 的 人才」與詞組「傑出 的人 才」，第一個詞都是一樣的，沒有必要再做之後的斷詞規則選擇詞組，我們可以直接在 Rule1 後輸出適當的斷詞結果，如圖 3-1-5 所示：

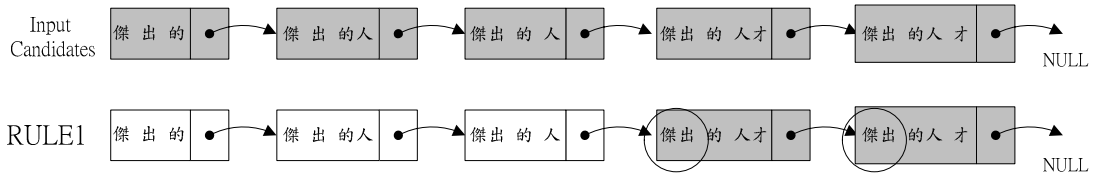


圖3-1-5: 修改之斷詞規則示意圖

因此我們可以對於這些斷詞規則進行稍微的修改，只要在經過某一條斷詞規則

之後，如果合乎斷詞規則的候選詞組，它們的第一個詞都是一樣的，我們直接將斷詞結果輸出，而不再往下做以後的規則。

Rule7 是為了解決 Rule1～Rule6 不能選出詞組的情形，Rule7 將隨意選擇候選詞組中任一個詞組的第一個詞作為斷詞結果輸出。

(6)外文及標點符號處理單元

「標點符號」、「外文」及「特殊符號」，在這裡給予它們詞類標記以及音碼。

(7)輸出斷詞結果

在這裡我們將經過「挑選候選詞組」斷出的詞，存入一個 linked list 資料結構，「挑選候選詞組」每斷出一個詞，便把新的斷詞結果接在 linked list 上。以「傑出的人才留在台灣」為例，如圖 3-1-6 所示

圖3-1-6: linked-list Output words

(8)留下有用的詞組

由於三個連續的詞為一組候選詞組，但經過斷詞規則斷詞後，只輸出符合斷詞規則的詞組中第一個詞為斷詞結果，那麼另外後面兩個詞仍然可以留下來，當作斷出下一個詞的候選詞組中第一和第二個詞，只要再向詞典與構詞比對出第三個詞，新的候選詞組又再產生了，不必一次搜尋三個連續的詞作為候選詞組，節省許多時間及運算量，因此，只要是候選詞組中第一個詞與斷出詞是一樣的，此候選詞組留下，為了斷出下一個詞，只要再向詞典或構詞規則比對得一詞便可以建立新的候選詞組，以「傑出的人才留在台灣」為例，如圖 3-1-7 所示

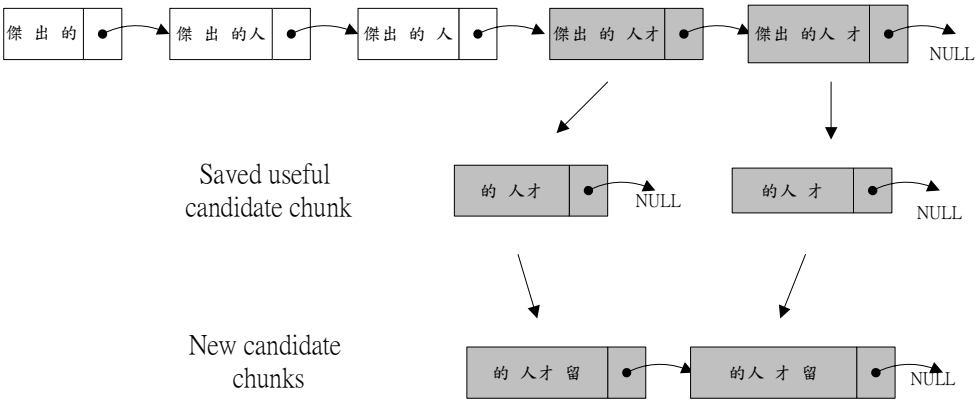


圖3-1-7: 留下有用的詞組示意圖

(9) 更新比對字串

在斷出一個詞之後，我們必須更新與詞典及構詞規則比對的字串開始位置。

3.2 構詞單元

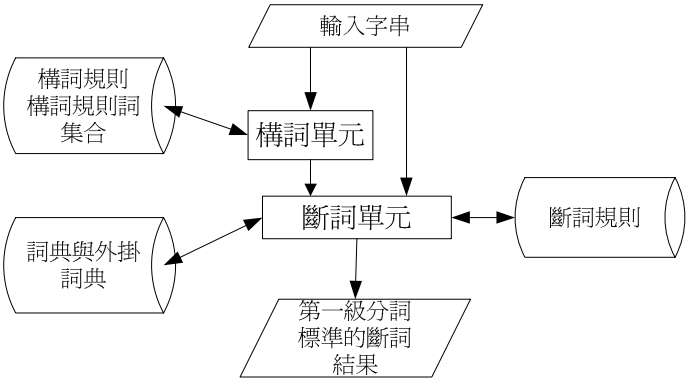


圖3-2-1 構詞單元在斷詞器中的位置

由於在候選詞組進入斷詞單元前，我們必須以查詞典的動作來建立候選詞組，然而詞典無法收錄所有的詞，這些未收錄的詞，有些是具有規則的，我們可以利用構詞規則，把輸入文句中符合規則的詞構出，所以由構詞規則構出的詞就相當於查詞典的動作，這些詞是「定量複合詞」、「數詞定詞」、「數量定詞」、「時間詞」、「地方詞」、「重疊詞」，希望能夠經由構詞單元，把這些詞由輸入文句詞組中合併出來，相同於查詞典的地位，因此構詞單元構出的詞與詞典中的詞皆為建立候選詞組的詞集合。

我們將構詞單元的構詞規則，根據處理方法不同，分為兩大類來說明：(1)量複合詞構詞單元(包含數詞定詞、數量定詞、時間詞、地方詞及部分的重疊詞的構詞)、(2)重疊詞構詞單元。

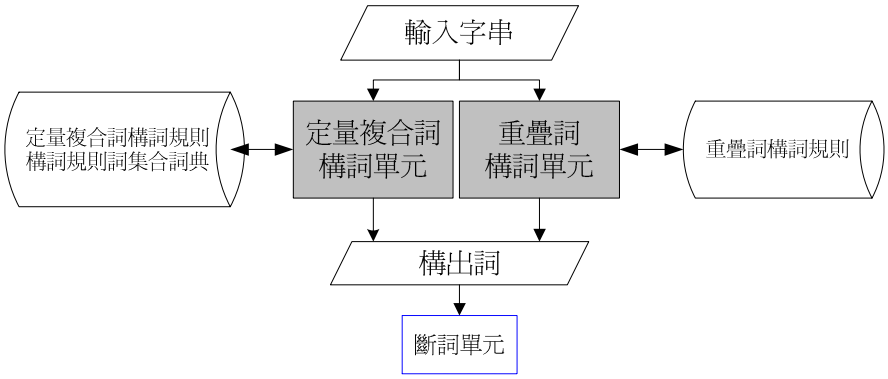


圖3-2-2:構詞單元示意圖

3.2.1 定量複合詞構詞單元的設計

在唐大任「中文斷詞器之研究」【4】中，將所有由定量複合詞構詞規則構出的詞，都給予定量複合詞的「詞類標記 58,DM」，然而其中有些構出的詞，詞類為「數詞定詞」、「數量定詞」或「時間詞」，我們希望由「定量複合詞構詞單元」構出的詞，能給予它們正確的詞類，因此本單元的設計，會把構出的詞類給予細分，某一條構詞規則構出的詞，會對給予這個詞特定的詞類。

3.2.1.1 將說明將說明構詞規則的表示法，3.2.1.2 將說明「定量複合詞構詞單元」在程式語言中構詞的演算法。

3.2.1.1 定量複合詞構詞規則的表示法

由【2】我們知道樹狀結構的構詞規則可以用 Regular Expression 表示，然而以 Regular Expression 表示定量複合詞構詞規則，是容易給人看懂規則，為了要把規則應用在程式語言中，我們必須把 Regular Expression 轉為以「規則標記」表示的 Chomsky Normal Form 表示方法，將某規則**所有可能的**樹狀結構(general tree)表示出來，應用在程式語言中，好處是如果對於某一項規則要修改或是要增加新規則，我們只要對這些「規則標記」進行簡單的修改便可。

我們以表 3-2-1 來說明。「規則標記」-2～113 所對應的詞，為構詞規則的詞集合，它們就是構詞規則中最基本的元素，而同一個規則標記的詞是具有相同特性(詞類)或是句法一致的特性，詳細的規則標記請參閱附錄 3。「規則標記」199～275 所對應的為「定量複合詞的構詞規則」，它們的構成為「規則標記序列」，這些序列表示某構詞規則的「子規則」，同時每一條子規則都有對應的詞類，詳細的定量複合詞構詞規則請參閱附錄 4。

表 3-2-1：構詞規則表示法

規則標記	集合	詞類
-2	忠,孝,仁,愛……	101
0	一,二,兩,三,四,五,……	17
3	大,小,整	20
4	多,餘,半,出頭,好幾,開外,整,正,許,足,之多	21
…	…	…
22	半	20
…	…	…
27	年	101
28	班	101
…	…	…
36	另外,近,將近	18
…	…	…
90	個,對,雙,支……	22
…	…	…

Regular Expression	規則標記	Comsky Normal Form	詞類	範例
IN1 -> NO*	199	0*	17	一百
…	…	…	…	
NOP_1 -> IN1 (DESC) {半}	205	199 3 22	20	四大半
…	…	…	…	
NOP -> IN1 (DESC) ({半}) M	210	205 90	58	四大半個
…	210	199 90	58	一百個
NOP3 -> IN1 M PNM	212	199 90 4	58	一百個之多
…	…	…	…	
CNP -> IN1 {年} {IN1,ON,N} {班}	237	199 27 199 28	14	一年一班
	237	199 27 2 28	14	一年甲班
	237	199 27 -2 28	14	一年忠班
…	…	…	…	
OSP3 -> {另外,近,將近} {NOP3}	243	36 212	58	將近一百個之多

3.2.1.2 定量複合詞構詞的演算法

對於一串輸入的中文字串，應該找出所有定量複合詞組合的可能，並將所有的可能全部送入斷詞單元，整個定量複合詞構詞單元的工作流程，表示在圖3-2-3。

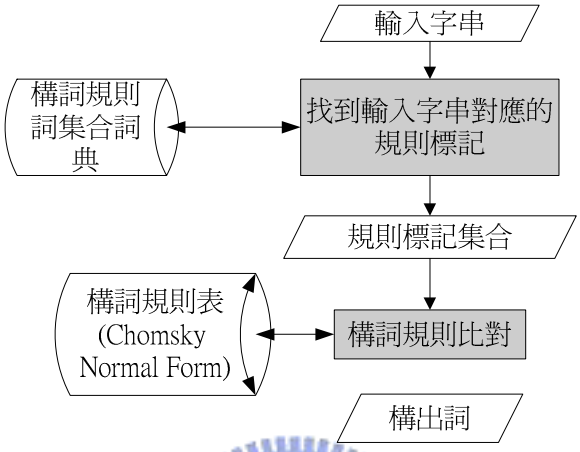


圖3-2-3:定量複合詞構詞工作流程圖

以下皆以定量複合詞「將近一百個之多」為例，整個構詞的流程分為兩大步驟：

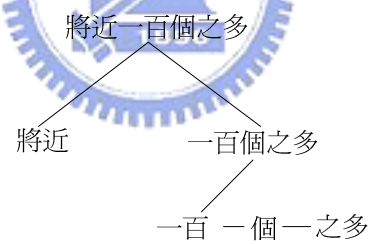


圖3-2- :定量複合詞將近一百個之多

(1)找到輸入字串(詞串)對應的構詞規則標記

這一步驟的目的，是將所有可能為定量複合詞的詞串找出來，並給予詞串對應的「規則標記」與詞類。

類似於斷詞單元中建立候選詞組的動作，首先從輸入字串的第一個字開始，將輸入字串與「構詞規則詞集合詞典」做比對，組成所有可能的詞串，找到詞串對應之所有可能規則標記，在下一步驟「構詞規則詞比對」的時候，需要規則標記來進行構詞規則比對。

經過這步驟將構成「將近　一　百　個 之多」這樣含有構詞規則標記的詞串，且這個詞串對應可能的規則標記如表 3-2-2 所示：

詞串	將近	一	百	個	之多
構詞規則標記	36	0	0	90	4
		6			
		42			

表 3-2-2:「將近一百個之多」之規則標記集合表

在找到所有規則標記的集合之後，便可以進行構詞規則比對的動作。

(2)構詞規則比對

這一步驟的目的，是要把所有可能符合構詞規則的詞，經由構詞規則的比對，進而合併構成新詞，且由於構詞規則是以樹狀結構表示，一個樹代表一個構詞規則，且規則與規則之間可能再以其他規則結合，並不只是單純的規則標記序列比對，必須將規則標記集合對應構詞規則**所有可能的**樹狀結構找出，以來構出某一類的詞。構詞規則比對的步驟如下：

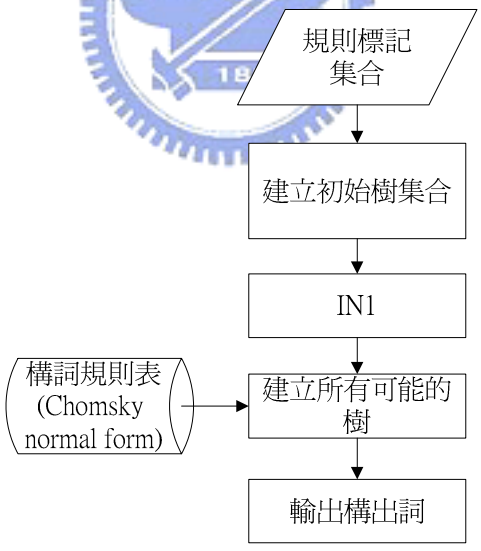


圖3-2-4:構詞規則比對流程圖

a. 建立初始樹集合

由於定量複合詞構詞規則是以樹狀結構表示的，我們必須將輸入的規則標記集合，先轉為以樹狀結構表示的資料結構，如圖 3-2-5。

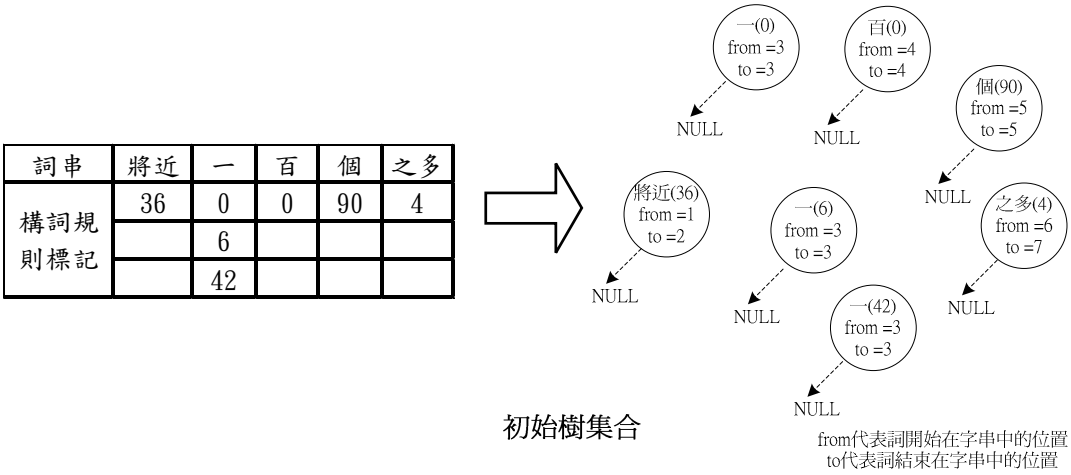


圖3-2-5:建立初始樹集合

b. IN1

IN1 為數詞定詞的構詞規則 IN1 -> {NO1, NO2}*，如：一百、一百五十等等，由於此項規則有簡單的遞迴特性，我們可以直接將符合此規則的詞構出來，相同的我們也將構出之數詞定詞建立成以樹狀結構表示的資料結構，並給予構出的詞數詞定詞的詞類標記。

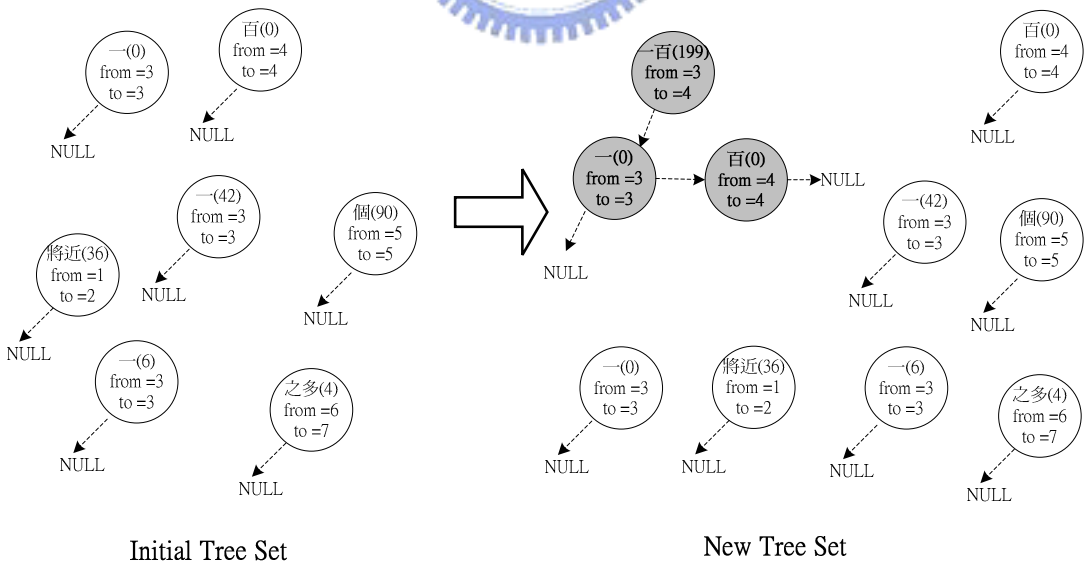


圖3-2-6:IN1規則構詞示意圖

c. 建立所有可能的樹

這個步驟是構詞規則比對的核心，目的是把所有可能的樹狀結構組合找出來(可能的構詞組合)，若樹集合中的某一棵樹可以和樹集合中其他的樹以某個構詞規則結合，我們便建立一棵新的樹(新的詞)，並給予新的樹「規則標記」(代表符合哪項規則)和詞類(一條規則對應到一個特定的詞類)，並把新建的樹加入樹集合中，當樹集合中的每一棵樹無法再繼續以構詞規則構出新詞，代表構詞結束。圖 3-2-7 為此步驟的流程圖：

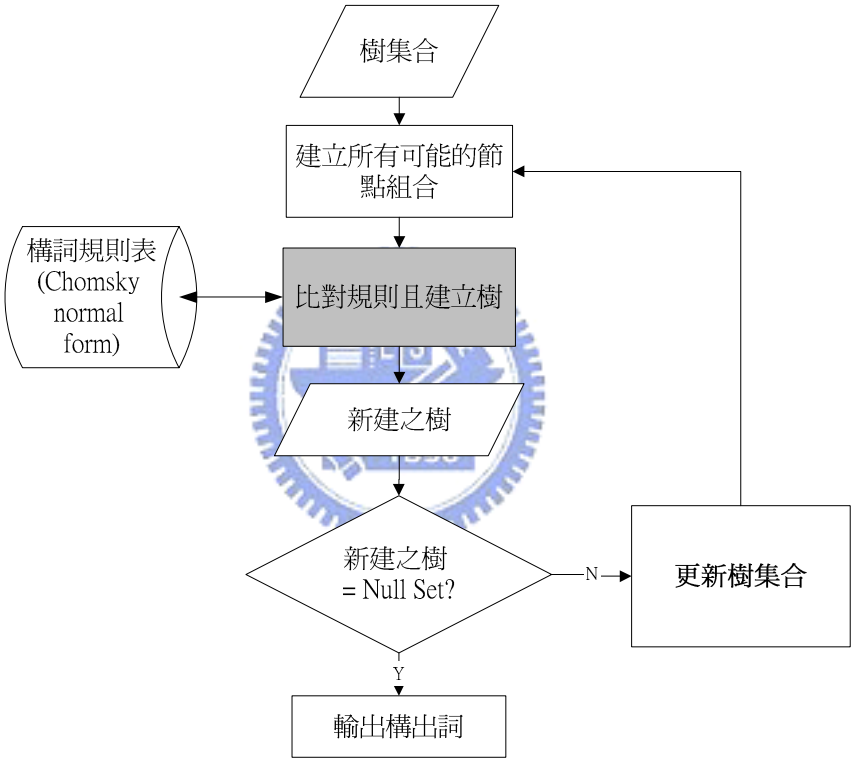


圖3-2-7: 建立所有可能的樹

分項說明

(c.1)建立所有可能的樹組合

將所有可能的節點組合找出，如圖 3-2-8 所示

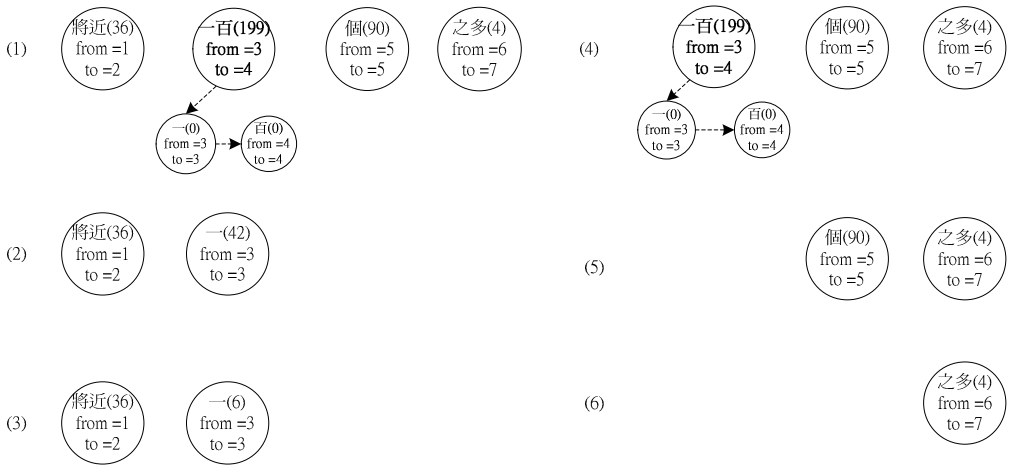


圖3-2-8:所有可能之節點組合

(c.2)比對規則且建立樹：

將每個節點組合向構詞規則進行比對，若符合某一項規則，便產生新的一棵樹，這個步驟十分類似斷詞單元中查詞典的動作，就是將所有可能的規則標記序列組合向構詞規則表進行比對，為了增加規則比對的速度，我們將定量複合詞構詞規則相同於詞典樹的做法，將規則儲存成樹狀結構表示，如圖 3-2-9 所示

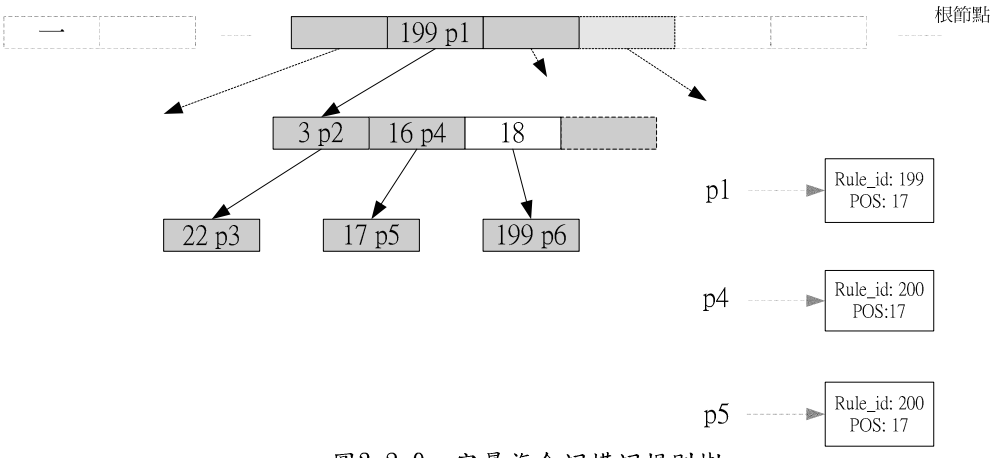


圖3-2-9: 定量複合詞構詞規則樹

整個規則樹為一個 general tree，同一層的節點依照規則標記的數字大小

排序好，是一個記憶體動態改變大小的陣列，每增加一個新的規則序列便會插入樹狀結構之中，深色底的部分代表規則的終端點，代表由根節點走到這個終端點經過的節點為一個構詞規則序列，在這個終端點也會紀錄這個規則對應的規則資訊(規則標記、詞類)的記憶體指標(p*)，而非終端點不為一個規則，沒有對應的規則資訊。在比對規則的時候，以輸入規則序列「199(一百) 16(多) 17(萬)」為例，「199」先由根節點開始搜尋(二分法搜尋)，之後「16」由節點「199」的子節點繼續搜尋，以下類推，一但為終端點，便符合一向規則，以這個例子會符合三條規則「199」、「199 16」及「199 16 17」，並得到相對應的規則資訊。

(c.3)更新樹集合：

這個步驟將新建立的樹集合，加到原本的樹集合裡。如果在經過比對規則後(步驟c.2)，沒有產生新的樹集合，代表構詞結束。

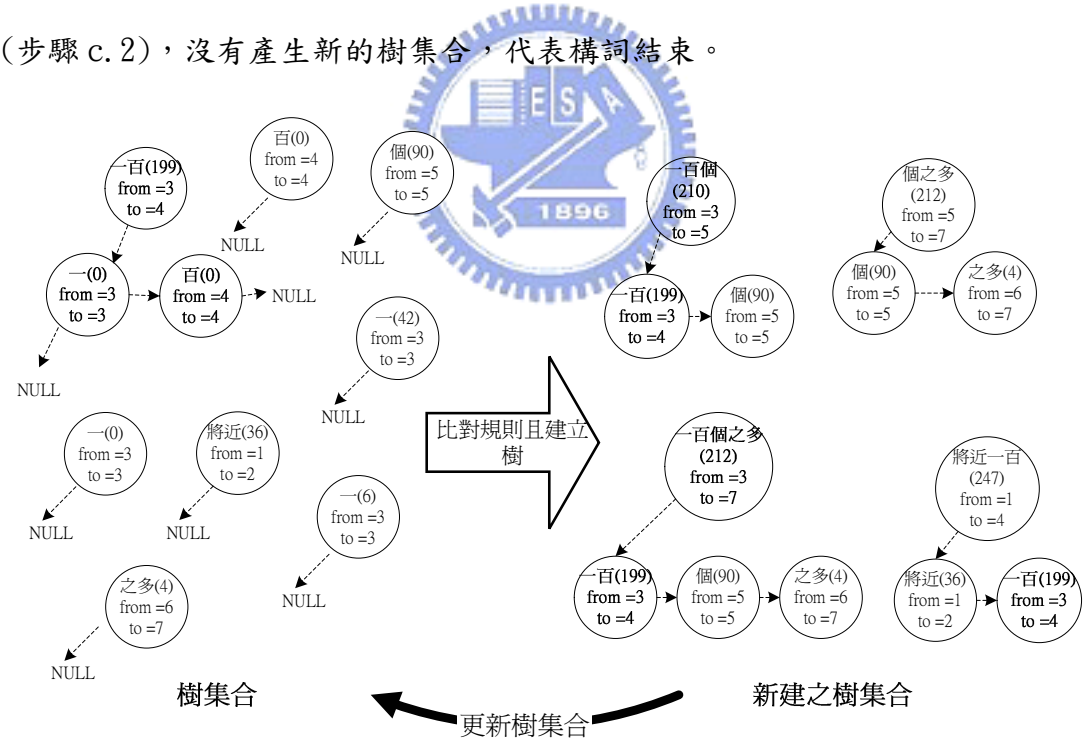


圖3-2-10: 更新樹集合

d.輸出構出詞

這一步驟是將步驟 3 樹集合中構出的詞輸出，由於構詞規則可能會有幾條規則會構出一樣的詞，這樣造成構出的詞是一樣的，但規則標記與詞類可能不一樣，然而這些詞類與規則標記都是這個構出詞**可能的**規則標記與詞類，因此我們必須把所有的規則標記與詞類資訊保留，記載在這個構出詞中，下舉一例：

兩千年		
符合的規則	rule_id	詞類
NOP1 -> IN1 (DESC) ({半}) M	210	58 定量複合詞
TDM2 -> IN1 {年}	262	16 時間詞

3.2.2 重疊詞構詞單元

在唐大任「中文斷詞器之研究」【4】中，只有考慮四字疊詞的構詞，經由參考中研院【3】提出的重疊詞構詞規則，我們將重疊詞構詞規則重新修改，詳列於下表 3-2-3。至於構出的重疊詞詞類，我們給予他們原本的詞類，例如：「開一開眼界」，「開眼界」原本的詞類是「VA 動作不及物動詞」，相同的我們給予「開一開眼界」「VA 動作不及物動詞」。

表 3-2-3：重疊詞構詞規則表

型態	XX
regular expression	RD1 -> X ({一,了}) X
合乎此規則的條件	(1)X 的詞類 = VA | VB | VC | VD | VE | VF
範例	(1)**聞一聞**野花的香味

型態	XXY(Z)
regular expression	RD5 -> X ({一,了}) XYZ
合乎此規則的條件	XY(Z) 的詞類 = VA | VB
範例	(1)請他來**評評理**(2)希望民眾來**開開眼界**

型態	XX
regular expression	RD2 -> XX {的,地}
合乎此規則的條件	X的詞類 = VH
範例	肚子**空空的**

型態	XXYY
regular expression	Re1 -> XXYY
合乎此規則的條件	XY = VA | VAC | VB | VC | VCL | VH | VHC | VI | VJ | VK | VL | A | D | Na | Nc | Ncd | Neu
範例	心情**上上下下**起伏

型態	XXXX
regular expression	Re2 -> XXXX
合乎此規則的條件	XX = VA | VAC | VB | VC | VCL | VH | VHC | VI | VJ | VK | VL
範例	下起了**綿綿綿綿**細語

型態	XYXY
regular expression	Re3 -> XYXY
合乎此規則的條件	XY = VA | VAC | VB | VC | VCL | VE | VF | VH | VHC | VI | VJ | VK | VL | D | Na
範例	讓我**想像想像**一下

3.3 詞類標記單元

詞類可以代表一個詞在句子中的語法角色，而詞類標記單元的功能，就是將斷詞單元產生的詞串，予以標記詞類，一般來說，語言學者可以用一些規則法來標記每一個句子中的詞，然而如果要自動化地標記詞類，規則法過於繁雜，因此利用大量的統計資訊，化簡為馬可夫模型，並以 Viterbi Decoding 求最佳詞類串結果。

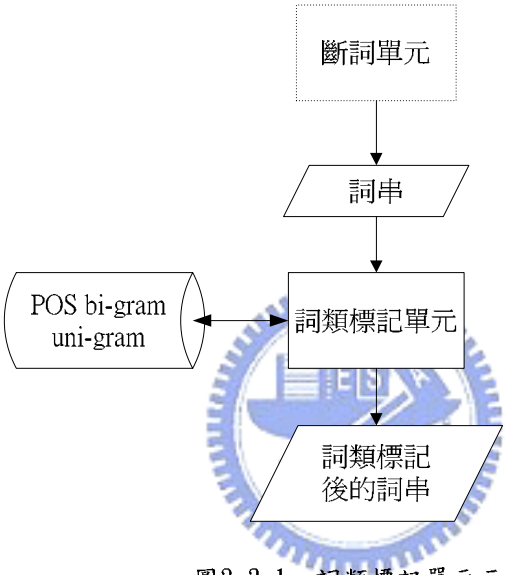


圖3-3-1: 詞類標記單元示意圖

3.3.1 詞類標記單元的工作原理

假設一中文詞串為

$$W_1^n = W_1, W_2, \ldots\ldots W_n$$

對應的詞類串為

$$pos_1^n = pos_1, pos_2, \ldots\ldots, pos_n$$

其中 n：詞數

W_i：此詞串中第 i 個詞 $(1 \le i \le n)$

pos_i：此詞類串中第 i 個詞的詞類 $(1 \le i \le n)$

此中文詞類串的構成機率可表示為下式：

$$P(pos_1^n \mid w_1^n)$$

$$= P(pos_1 \mid w_1^n)\prod_{i=2}^{n} P(pos_i \mid pos_1^{i-1}, w_1^n) \qquad (3\text{-}3\text{-}1)$$

由式(3-3-1)可以觀察出句子中的每一個詞類出現機率，與詞串(w_1^n)及之前出現的詞類(pos_1^{i-1})有關，且 pos_1^{i-1} 之組合隨 $i-1$ 之增加成次冪增加，這使我們建立語言模型上產生相當大的困難，我們必須做合理的假設來簡化求取 $P(pos_1^n \mid w_1^n)$ 的複雜度，我們假設：

1. 詞類串合乎馬可夫序列，且為穩態(Stationary)， $P(pos_1^n)$ 可簡化為 N 階的馬可夫模型 (Markov Model with Nth Order)，也就是(N+1)連文的語言模型，如下式：

$$P(pos_1^n) = P(pos_1)\prod_{i=2}^{n} P(pos_i \mid pos_{i-N}^{i-1}) \qquad (3\text{-}3\text{-}2)$$

2. 詞類 pos_i 與詞 w_k 有關，假如 $i = k$

因此式(3-3-1)可化減為式(3-3-3)如下：

$$P(pos_1^n \mid w_1^n)$$

$$= P(pos_1 \mid w_1)\prod_{i=2}^{n} P(pos_i \mid pos_{1-N}^{i-1}, w_i) \qquad (3\text{-}3\text{-}3)$$

也就是說，句中某一個詞類的出現機率只和現在是什麼詞以及過去的前 N 個詞之詞類有關，假設我們使用 V 種不同的詞類來標記詞串，當 N 越大，一個(N+1)連文的語言模型，需要的參數量($\sum_{i=1}^{N+1} V^i$)也會越大。即使我們擁有足夠大的訓練語料庫，可以估算出相當可靠的詞類 N 連文參數，但仍得面對這些資料在儲存

空間與執行速度上的困難，針對這個問題，我們選擇 $N = 1$，採用效果不錯且所需參數量較少的雙連文模型(bigram language model)。

$$P(pos_1^n \mid w_1^n)$$

$$= P(pos_1 \mid w_1)\prod_{i=2}^{n} P(pos_i \mid pos_{i-1}, w_i) \tag{3-3-4}$$

我們再對(3-3-4)式中的 $P(pos_i \mid pos_{i-1}, w_i)$ 進行整理

$$P(pos_i \mid pos_{i-1}, w_i)$$

$$= \frac{P(pos_i, pos_{i-1}, w_i)}{P(pos_{i-1}, w_i)} = \frac{P(pos_i, pos_{i-1}, w_i)}{P(pos_{i-1})P(w_i)} = \frac{P(pos_i, w_i \mid pos_{i-1})}{P(w_i)}$$

$$= \frac{P(w_i \mid pos_{i-1}, pos_i)P(pos_i \mid pos_{i-1})}{P(w_i)} = \frac{P(w_i \mid pos_i)P(pos_i \mid pos_{i-1})}{P(w_i)}$$

$$= \frac{P(pos_i \mid w_i)P(w_i)P(pos_i \mid pos_{i-1})}{P(pos_i)P(w_i)}$$

$$= \frac{P(pos_i \mid w_i)}{P(pos_i)} P(pos_i \mid pos_{i-1}) \tag{3-3-5}$$

將(3-3-5)式代入(3-3-4)式，此中文詞類串的構成機率可表示為下式：

$$P(pos_1^n \mid w_1^n)$$

$$= P(pos_1 \mid w_1)\prod_{i=2}^{n}\left[\frac{P(pos_i \mid w_i)}{P(pos_i)} P(pos_i \mid pos_{i-1})\right] \tag{3-3-6}$$

將詞類串機率表示成(3-3-6)之後，我們假設輸入的詞串為 $< W(1), W(2), ..., W(N) >$，其對應可能的詞類串為 $< pos_i(1), pos_i(2), ..., pos_i(N) >$，我們可利用 Viterbi Search 找出最有可能的詞類標記連接方式，如下式

$$< pos^*(1), pos^*(2), ..., pos^*(N) >$$

$$= \arg \max_{\forall\{pos_i(n)\}} P(pos_i(1) \mid W(1))\prod_{n=2}^{N}\left[\frac{P(pos_i(n) \mid W(n))}{P(pos_i(n))} P(pos_i(n) \mid pos_j(n-1))\right] \tag{3-3-7}$$

其中

$pos_i(n); i = 1,...,V$ ：詞$W(n)$對應的詞類，V 為詞類數目

$P(pos_i(n) \mid pos_j(n-1))$ ：詞類雙連文模型

3.3.2 詞類雙連文模型的建立

要應用 Viterbi-search 來求取最佳的詞類串，我們必須求取 (1) 詞類單連文模型 $P(pos_i)$、(2) 詞類雙連文模型 $P(pos_i \mid pos_{i-1})$以及(3) 機率 $P(pos_i \mid w_i)$

我們所使用的詞類種類共有 58 類，為詞庫小組標示〔中研院平衡語料庫 3.0 版〕用的 57 個詞類(見附錄 5)，再加上定量複合詞的第 58 類，建立模型所用的語料庫為〔中研院平衡語料庫 3.0 版〕，以來統計詞類雙連文模型所要的各個參數。

(1) 求取單連文模型 $P(pos_i)$，其數學式為

$$P(pos_i) = \frac{Count(pos_i)}{\sum_{pos_i=1\sim58} Count(pos_i)} \qquad (3\text{-}3\text{-}8)$$

pos_i ：一詞類串中第 i 個詞之詞類(pos_i =1~58)

$Count(pos_i)$：表 pos_i 在〔中研院平衡語料庫 3.0 版〕出現的次數

(2) 求取詞類雙連文模型 $P(pos_i \mid pos_{i-1})$，其數學式為

$$P(pos_i \mid pos_{i-1}) = \frac{Count(pos_{i-1}, pos_i)}{Count(pos_{i-1})} \qquad (3\text{-}3\text{-}9)$$

$Count(pos_{i-1}, pos_i)$：表示先出現 pos_{i-1} 再接 pos_i 的這樣組合，在〔中研院平衡語料庫 3.0 版〕出現的次數

限於我們從有限的語料庫求取以上參數，難免會遇到$Count(pos_{i-1}, pos_i)$ 為 0，使得 $P(pos_i \mid pos_{i-1})$ 式中分母為 0 的情況，但是不表示實際上先出現 pos_{i-1} 再接 pos_i 的這樣組合不會出現，因此必須採用平滑法(Smoothing)【7,8】，來將機率

重新分配，求出所有的 $P(pos_i \mid pos_{i-1})$，最後可得我們的詞類雙連文模型。以下敘述我們使用平滑法來建立詞類雙連文模型的作法。

a. 當 $Count(pos_{i-1}, pos_i) = 0$

$$P(pos_i \mid pos_{i-1}) = \beta(pos_{i-1}, pos_i) \times \frac{P(pos_i)}{\sum_{pos_i : Count(pos_{i-1}, pos_i)=0} P(pos_i)} \quad (3\text{-}3\text{-}10)$$

b. 當 $1 \le Count(pos_{i-1}, pos_i) \le k$

$$P(pos_i \mid pos_{i-1}) = d_{Count(pos_{i-1}, pos_i)} \times \frac{Count(pos_{i-1}, pos_i)}{Count(pos_{i-1})} \quad (3\text{-}3\text{-}11)$$

c. 當 $Count(pos_{i-1}, pos_i) > k$

$$P(pos_i \mid pos_{i-1}) = \frac{Count(pos_{i-1}, pos_i)}{Count(pos_{i-1})} \quad (3\text{-}3\text{-}12)$$

其中 $\beta(pos_{i-1}, pos_i) = 1 - \frac{1}{A}\sum_{a \ge 1} d_a \cdot c_a \cdot a$ (3-3-14)

$$d_a = \frac{a - m}{a} \quad (3\text{-}3\text{-}15)$$

$$m = \frac{c_1}{\sum_{a=1}^{A} a \cdot c_a} \quad (3\text{-}3\text{-}16)$$

$$A = \sum_{a \ge 1} c_a \cdot a \quad (3\text{-}3\text{-}17)$$

k =5

c_a ：發生 a 次的事件數目

d_a ：discounting coefficient

A ：發生之所有事件數

(3)求取機率 $P(pos_i \mid w_i)$，其數學式為

$$P(pos_i \mid w_i) = \frac{P(pos_i, w_i)}{P(w_i)} = \frac{Count(pos_i, w_i)}{Count(w_i)} \tag{3-3-18}$$

3.4 文字正規化單元

以語音合成為考量，在輸入文句之中，有些阿拉伯數字、詞或符號必須由寫法轉為語音讀法，這個過程稱為文字正規化，舉例來說「９０％」應該讀為「百分之九十」。經由蒐集整理的結果，我們發現到大部分的定量複合詞如果含有阿拉數字及特殊符號，都需要被正規化為讀法，而且讀法與定量複合詞的結構關係密切，圖 3-4-1 解釋了其關係。

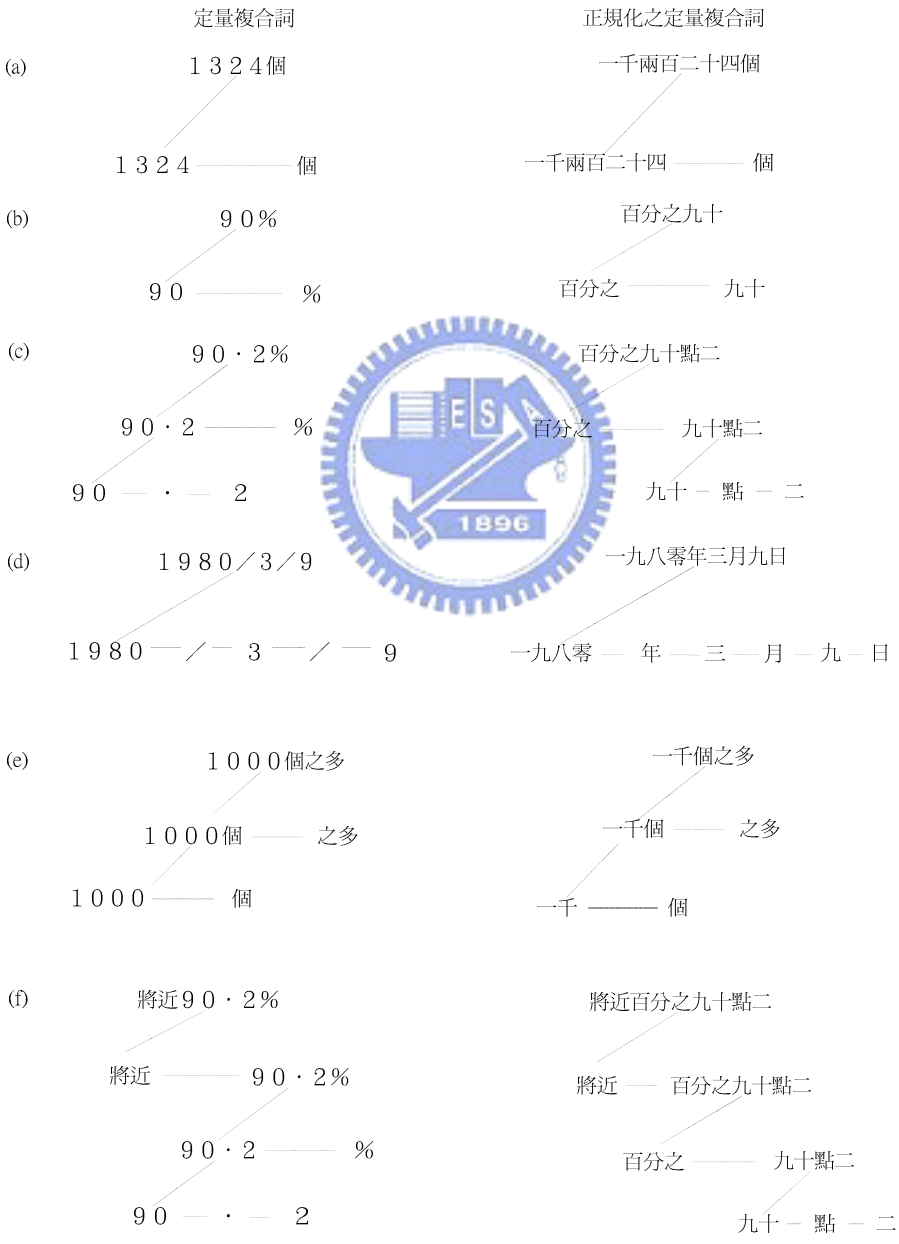


圖3-4-1: 定量複合詞與文字正規化之關係

3.4.1 文字正規化的原理

在這裡我們先定義文字正規化的一些基本單位：

構詞基本元素：指複合詞樹狀結構中的子節點，也就是構詞的基本單位。

正規化元素：指構詞基本元素所對應的語音讀法。

正規化元素規則：構詞基本元素轉換為正規化元素的規則

比較圖 3-4-1 中(a)的定量複合詞「１３２４個」與正規化之定量複合詞「一千兩百二十四個」，其中的**構詞基本元素**(定詞)「１３２４」直接經由數字正規化對應為**正規化元素**「一千兩百二十四」，而且正規化之後的樹狀結構與原本的定量複合詞之結構一樣。觀察圖 3-4-1 中(b)：「９０％」經由正規化後，正規化元素「九十」對應到「９０」，「百分之」對應到「％」，正規化後的對應順序與樹狀結構改變。再來看圖 3-4-1 中(c)，「百分之」對應到「％」，「９０．２」也經由正規化對應到「九十點二」。

圖 3-4-1 中(a)(b)(c)中的每個構詞基本元素都會對應到一個正規化元素。圖 3-4-1 中(d)：「１９８０／３／9」在正規化後，兩個「／」分別對應到「年」、「月」，注意到「日」並沒有對應到任何原本定量複合詞的基本元素，是此正規化的**附加的成分**。

圖 3-4-1 中(d)以及(e)其中有些子節點需要被文字正規化，而整個定量複合詞正規化的結果，就是把需要正規化的與不需要正規化的部分，依據原本的樹狀結構組合而成。

3.4.1.1 文字正規化的特性：

經由以上例子的觀察，我們知道文字正規化有 5 種特性：

1. 一般定量複合詞的正規化，其中阿拉伯數字(定詞)直接經由數字正規化，樹狀結構不改變，如圖 3-4-1 中(a)。
2. 一般定量複合詞的正規化，正規化元素與構詞基本元素的順序不同，如圖

3-4-1 中(b)。

3. 需要正規化的定量複合詞中又含有需要正規化的子節點，相同於構詞規則中的遞迴特性，如圖 3-4-1 中(c)。
4. 正規化元素與構詞基本元素並非一對一的關係，必須加上附加的成分，如圖 3-4-1 中(d)。
5. 定量複合詞中，部分的子節點需要被文字正規化，而整個定量複合詞正規化的結果，就是把需要正規化的與不需要正規化的部分，依據原本的樹狀結構組合而成，如圖 3-4-1 中(e)和(f)。

3.4.1.2 文字正規化規則的表示：

依據文字正規化的特性，對於符合特性 1 的定量複合詞節點，我們直接對其中的定詞進行數字正規化，對於符合特性 2、3、4 的正規化，我們可以將文字正規化規則寫成一條條的樹狀規則表示，格式如下：

Form 1:

rule_id(DM_unit1:TN_unit1:Code1:order1|DM_unit2:TN_unit2:Code2:order2……)

Form 2:

rule_id(DM_unit1:TN_unit1:Code1:order1|……),Dummy*x*:C_unit*x*:Code*x*:order*x* …

Form 3:

DM_unit:TN_unit:Code:order

符號說明：

「(、)」：樹狀結構代表同一層的邊界。

「|」：樹狀結構中 sibling 的分界。

「:」：構詞基本元素、正規化元素、正規化元素對應的音碼、正規化順序的分界

符號。

「,」：分隔「規則欄位」與「附加成分欄位」。

欄位說明：

rule_id：符合此項正規化規則的定量複合詞規則編號。

DM_unit*x*：構詞基本元素，可以為 rule_id，也可為引號”” 中的詞

TN_unitx：正規化元素規則，有三種類別：

1. NTN：代表數字正規化。
2. 詞：以引號”” 中的詞為正規化元素。
3. rule_id：若 rule_id 也符合某項正規化規則，正規化元素就是此節點正規化後的讀法，若 rule_id 不符合任何正規化規則，正規化元素相同於原本此節點對應的構詞基本元素。

Codex：正規化元素對應的音碼，有三種類別：

1. NTN：給予數字正規化後的音碼
2. 音碼：直接給予引號”” 中的音碼
3. rule_id：給予此節點 rule_id 對應的音碼

order*x*：正規化元素組成正規化詞的排列順序。

C_unit*x*：附加成份的正規化元素，由於附加成份的正規化元素沒有對應到任何原本複合詞的基本元素，這裡的正規化元素只可以為詞，以引號”” 中的詞表示。

Form 1 為基本的「規則欄位」，用於表示構詞基本元素與正規化元素有一對一關係時，Form 2 包含逗號前的「規則欄位」以及逗號後的「附加成分欄位」用於表示構詞基本元素與正規化元素非一對一關係，Form 3 用於單純的符號轉換。詳細的所有文字正規化規則，請參附錄 6，以下舉幾個正規化規則說明(省略音碼欄位 Codex)：

Form 1:

A. 201(199:NTN:0|99:"點":1|199:199:2)

符合此項規則的如「９０．２」，這個詞會先由構詞單元構出，且含有樹狀結構的資訊如圖，構詞基本單元「９０(rule_id=199)」所對應的正規化元素規則為「NTN」數字正規化，構詞基本單元「．(rule_id=99)」所對應為「"點"」正規化為詞，「２(rule_id=199)」所對應為 rule_id「199」，正規化元素相同於原本此節點對應的構詞基本元素。

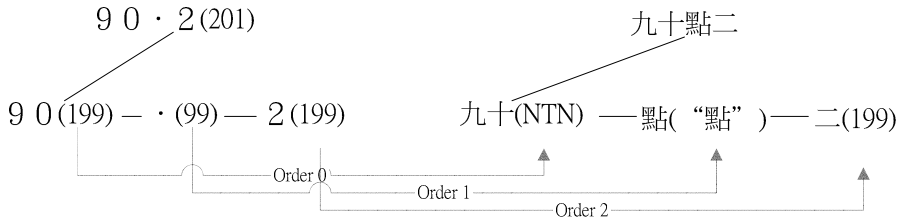


圖3-4-2: 文字正規化９０．２

B. 202(199:NTN:1|100:"百分之":0)

符合此項規則的如「９０％」

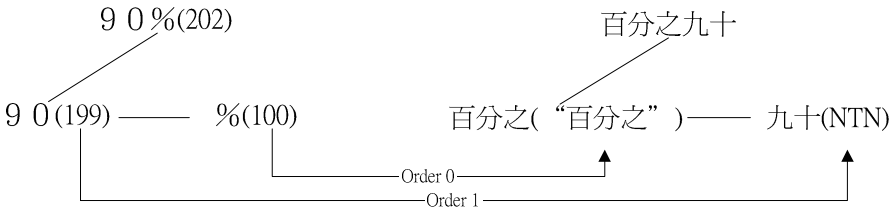


圖3-4-3: 文字正規化９０％

C. 202(201:201:1|100:"百分之":0)

符合此項規則的如「９０．２％」，要注意的是「９０．２」符合在範例 A 中的正規化規則，因此在此項正規化會再呼叫 rule_id=201 對應的正規化規則。

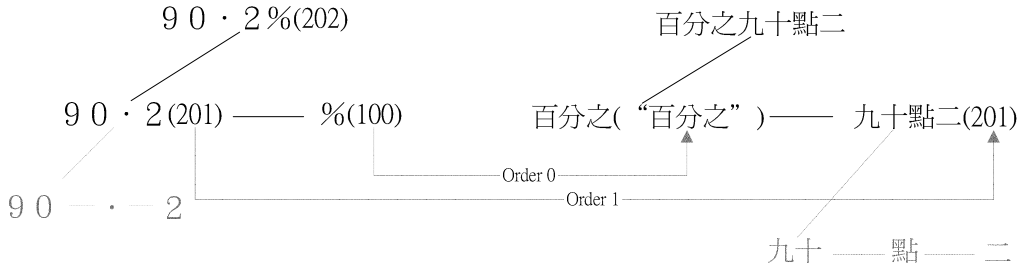


圖3-4-4: 文字正規化９０・２%

Form 2:

D. 262(199:199:0|102:"年":1|199:NTN:2|102:"月":3|199:NTN:4),0:"日":5

符合此項規則的如「１９８０年３月９日」

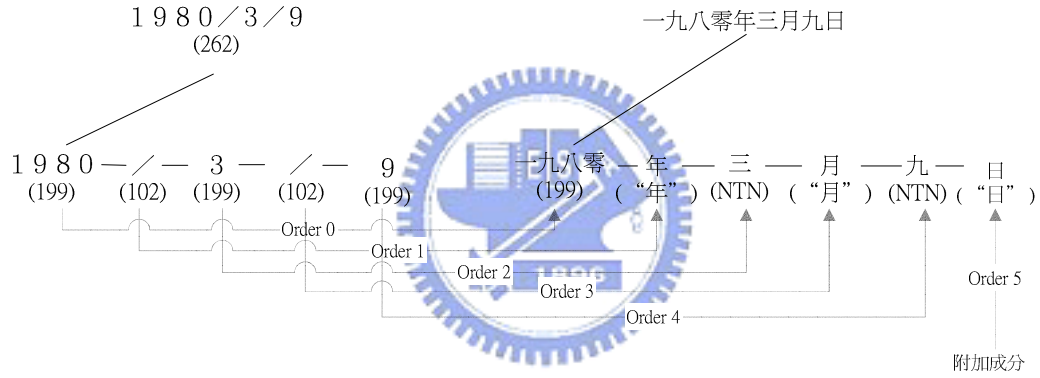


圖3-4-5: 文字正規化１９８０／３／９

Form 3:

E. "ＴＥＬ":"電話":0

此項規則直接將字串"ＴＥＬ"轉為中文的字串"電話"

3.4.2 文字正規化模組的設計

正規化模組需要複合詞的樹狀結構以及斷詞結果，因此文字正規化模組必須放置在斷詞單元之後，示意圖如圖 3-4-6。

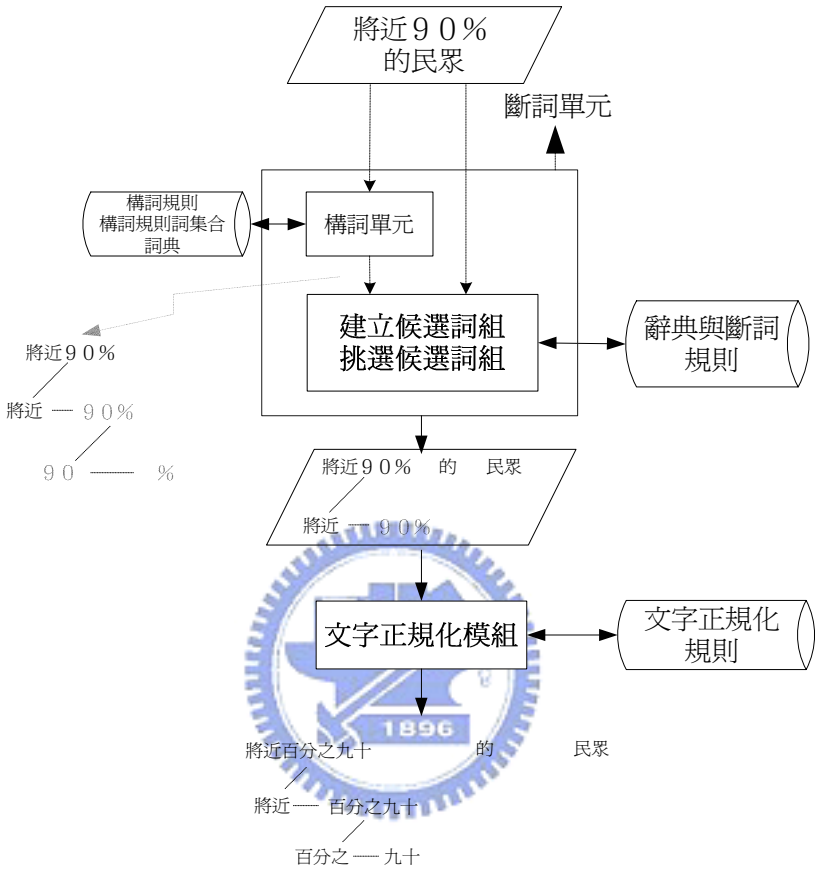


圖3-4-6: 文字正規化單元在斷詞器的位置

依據特性 5，我們知道定量複合詞的節點，有些是需要正規化，而有些是不是不需要的，所以我們對於定量複合詞樹狀結構的每一個節點，都要做過正規化規則的比對，而進行正規化的動作，最後再將每個節點正規化後的結果，依據原本定量複合詞的樹狀結構，組合成正規化的讀法，如圖 3-4-6 中所示，「將近９０％」中只有「９０％」需要文字正規化，最後文字正規化模組依據原本定量複合詞的樹狀結構，組合成正規化的讀法「將近９０％」，文字正規化模組的大致流程圖如圖 3-4-7。

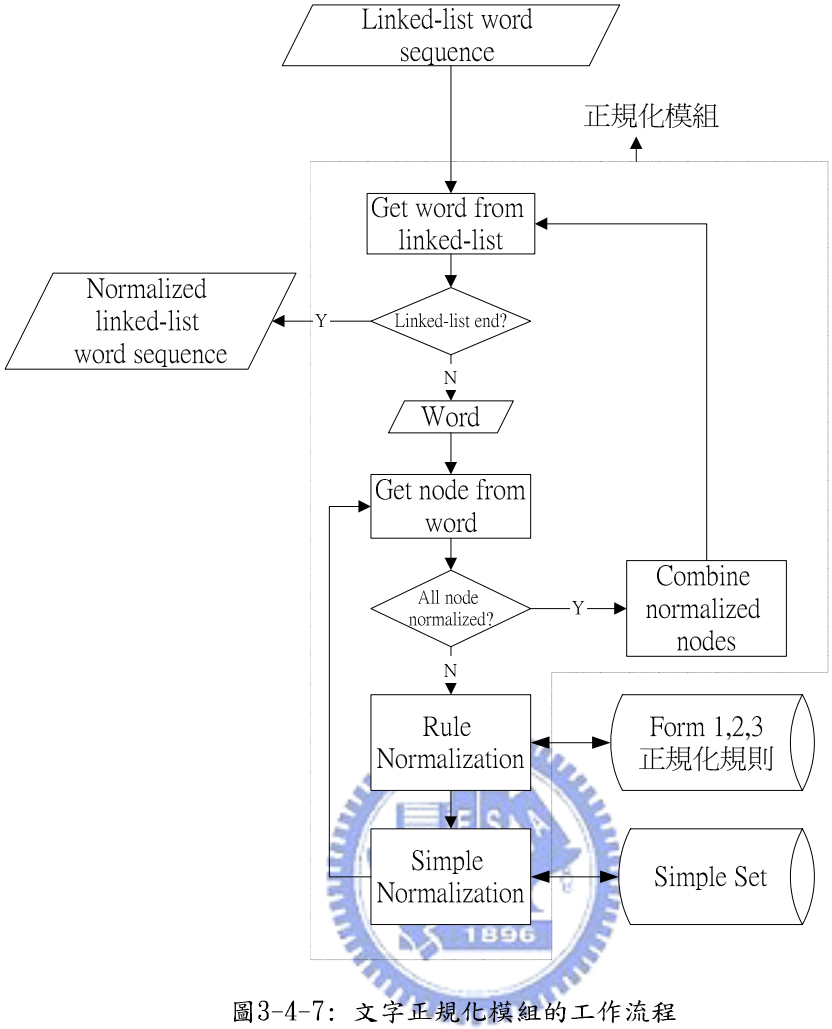


圖3-4-7: 文字正規化模組的工作流程

分項說明：

1. Linked-list word sequence：

輸入正規化模組的為由斷詞單元產生的斷詞結果詞串，詞串是用 linked-list 相連起來，如果詞是由構詞單元構出的，便含有文字正規化所需要的樹狀結構。

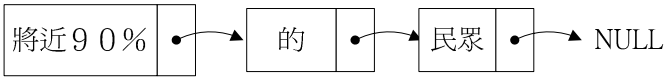


圖3-4-8: Linked-list word sequence

2. Get word from linked-list：

文字正規化單元一次處理一個詞，當正規化完一詞之後就換 linked-list 的下一個詞做正規化。

3. Get Node from word：

對於複合詞樹狀結構的每一個節點，如前面所述，都要做過正規化規則的比對，而進行正規化的動作，如果某一節點可以被文字正規化，這個節點的子節點便不需要再進行正規化規則的比對。

4. Rule Normalization：

符合特性 2、3 及 4 的複合詞節點，我們將正規化寫成一條條的正規化規則(Form 1、2、3)，構詞基本元素與正規化元素排列的關係紀錄在文字正規化規則裡，處理過程較複雜。詳細的演算法在後面說明(3.4.3.1 小節)。

5. Form 1,2,3 正規化規則：

用於 Rule Normalization 且以樹狀結構表示的正規化規則，表示法詳見**文字正規化規則的表示**，所有的規則列於附錄 6。

6. Simple Normalization：

符合特性 1 的複合詞節點，如定量複合詞規則 210 -> 199 90 (如:１００個)，我們簡單的將節點中的阿拉伯數字定詞直接正規化為語音讀法，構詞基本元素與正規化元素的排列是一樣的。在這裡我們另外注意，如果輸入的節點可以經由之前的 Rule Normalization 進行文字正規化，此模組不會處理正規化的動作。

7. Simple Set：

符合以 Simple Normalization 來進行文字正規化的定量複合詞，都會列於 Simple Set 中，如定量複合詞規則 210 -> 199 90。

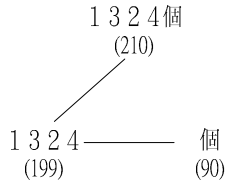


圖3-4-9: Simple Set Normalization

8. Combine normalized nodes：

當複合詞中的所有節點都經過文字正規化後，依據原本定量複合詞的樹狀結構，組合成正規化的讀法

圖3-3-10: Combine normalized nodes

9. Normalized linked-list word sequence：

輸入的詞串與輸出的詞串，它們的資料結構是一樣的，經過文字正規化後，會將資料結構中的語音讀法欄填上正規化後的文字與音碼。

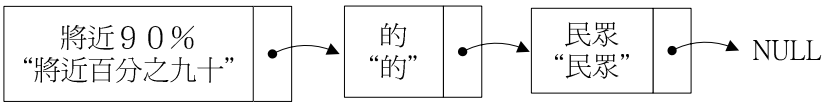


圖3-4-11: Normalized linked-list word sequence

3.4.3 文字正規化演算法

經過以上的敘述，其演算法整理如下

參數說明

Link: linked-list word sequence

Word: word in linked-list

NULL: end of linked-list

Node: a node in tree-structure Word.

Rule: Rule Normalization module

Simp: Simple Normalization module

演算法 TEXT_NORMALIZATION(Link)

Step 1: Get *Word* from *Link*. If *Word* is *NULL*, go to Step 6.

Step 2: Get non-normalized *Node* from *Word*. If all *Node*(s) are normalized, go to Step 5.

Step 3: Normalize *Node* by *Rule* or *Simp*.

Step 4: Go back to Step 2.

Step 5: Construct normalized word; Combine normalized *Node*(s) and then go back to Step 1.

Step 6: Terminate.

3.4.3.1 Rule Normalization 的演算法

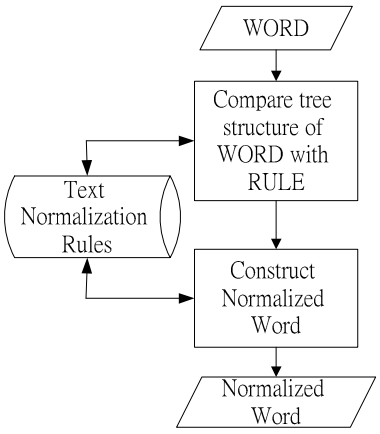


圖3-4-12: Rule Normalization演算法

Rule Normalization 最主要分成兩個步驟，第一部分是比較詞與正規化規則的樹狀結構，如果它們的樹狀結構一樣，才進行第二部分正規化的動作，以下舉一些例子：

Example 1：

「９０%」，它符合 202(199:NTN:1|100:"百分之":0）此項規則。

(1)比較詞與正規化規則的樹狀結構：

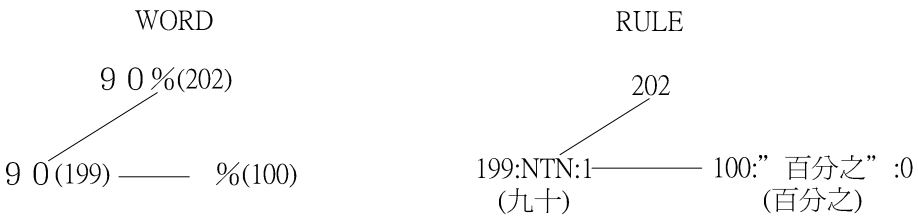


圖3-4-13: 文字正規化example 1

在比較樹狀結構的過程中，我們同時紀錄**構詞基本元素**與**正規化元素**對應的順序關係，並依據**正規化元素規則**產生對應的**正規化元素**

正規化元素的排列順序 Order	構詞基本元素 DM_unit	正規化元素規則 TN_unit	正規化元素 Word.TN
0	%	“百分之”	百分之
1	９０	NTN	九十

(2)正規化

如果在前一步驟樹狀結構的比較是相同的，我們可以將正規化元素的結果依正規化元素排列順序，構成正規化的文字

百分之九十 -> 百分之　九十

Example 2：

「１９８０／３／９」符合規則：

262(199:199:0|102:"年":1|199:NTN:2|102:"月":3|199:NTN:4),0:"日":5

(1)比較詞與正規化規則的樹狀結構：

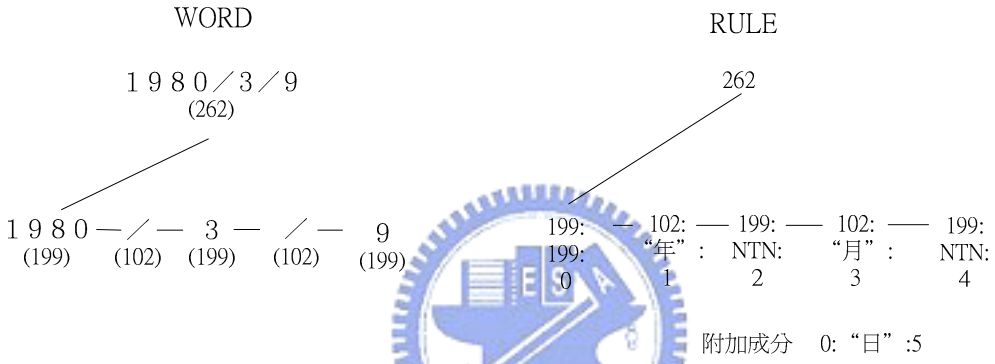


圖3-4-14: 文字正規化example 2

正規化元素的排列順序 Order	構詞基本元素 DM_unit	正規化元素規則 TN_unit	正規化元素 Word.TN
0	１９８０	199	１９８０
1	／	“年”	年
2	３	NTN	三
3	／	“月”	月
4	９	NTN	九
5		“日”	日

(2)正規化

１９８０年三月九日-> １９８０　年　三　月　九　日

在這個例子裡，規則的表示為 Form 2，含有附加成分，比較是否合乎此項規則，只比較樹狀結構部份，而在產生正規化單元時，才考慮附加成分。

Example 3:

「９０．２％」符合規則：202(201:201:1|100:"百分之":0)

(1)比較詞與正規化規則的樹狀結構：

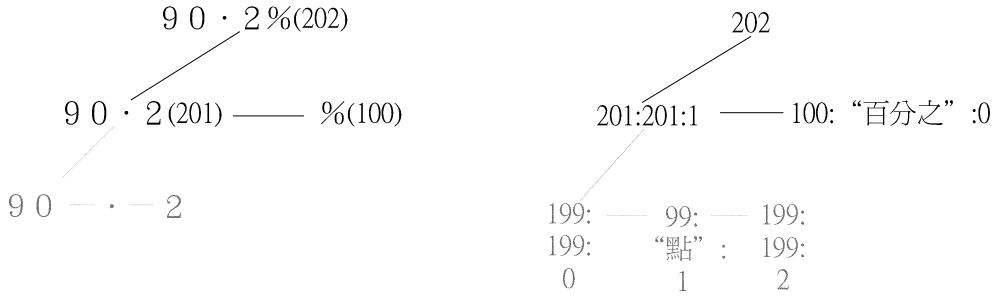


圖3-4-15: 文字正規化example 3

正規化元素的排列順序 Order	構詞基本元素 DM_unit	正規化元素規則 TN_unit	正規化元素 Word.TN
0	％	“百分之”	百分之
1	９０．２	201	九十點二

在這個例子裡要注意的是子節點「９０．２」對應的正規化元素規則，也就是另一項文字正規化規則：201(199:NTN:0|99:"點":1|199:199:2)，因此正規化元素「九十點二」是由再呼叫一次正規化單元而得來。

正規化元素的排列順序 Order	構詞基本元素 DM_unit	正規化元素規則 TN_unit	正規化元素 Word.TN
0	９０	NTN	九十
1	．	“點”	點
2	２	199	二

九十點二 -> 九十　點　二

(2)正規化

百分之九點二 -> 百分之 九點二

在圖示法說明整個 Rule Normalization 模組文字正規化的過程後，其演算法整理如下：

參數定義：

WORD：待文字正規化的樹狀結構複合詞

WORD.DM：WORD 裡的構詞基本元素的樹狀結構

WORD.TN：WORD.DM 對應的正規化元素

NOR_WORD：正規化後的詞

RULE：文字正規化規則集合(Form 1,2,3)

rule：某一個文字正規化規則

rule.DM_unit*x*：規則裡構詞基本元素的樹狀結構

rule.TN_unit*x*：規則裡的正規化元素規則

rule.order：規則裡正規化元素對應構詞基本元素的排列順序

rule.complment：規則的附屬成分

演算法 RULE_NORMALIZATION(*WORD*)

For each *rule* in *RULE*

Step 1：Compare tree structure of *WORD* with *rule*; Compare each *WORD.DM* with *rule.DM_unitx* . If they are the same in tree structure, go to Step 2, otherwise, go to Step 5.

Step 2：Construct *WORD.TN* according to *rule.TN_unitx*. Remember the order of *WORD.TN* according to *rule.order*.

Step 3：Construct *NOR_WORD*; Combine *WORD.TN* and *rule.complement* in the order specified by *rule.order* to construct *NOR_WORD*.

Step 4：Terminate; Normalized text *NOR_WORD* has been constructed in Step 3. Exit RULE_NORMALIZATION(*WORD*).

Step 5：Take next *rule* and go to Step 1

3.4.3.2 Simple Normalization 的演算法

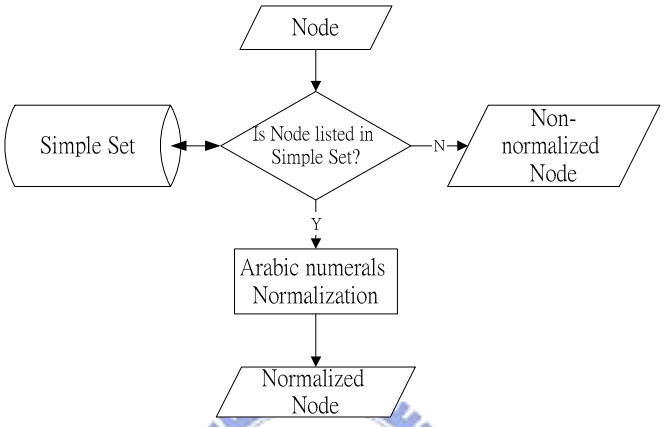


圖3-4-16: Simple Normalization 演算法

Simple Normalization 分為兩大步驟，首先我們先比對輸入的節點是否需要行數詞定詞的數字正規化；這些需要數字正規化的定量複合詞節點，將會列於 Simple Set 之中，若節點必須進行數字正規化，就會繼續下一個步驟「數字正規化」，反之則跳出。

第二步驟便是進行「數字正規化」，就是簡單地把輸入節點或子節點的阿拉伯數字，進行「數字正規化」，由於符合特性 1 的節點才會進入此模組處理，因此正規化元素與構詞基本元素的排列是一樣的，依據原本節點的樹狀結構，組合成正規化的讀法。下面舉一些例子：

Example 1：

「１００個」符合特性 1，也列於 Simple Set 名單之中，其樹狀結構如下圖：

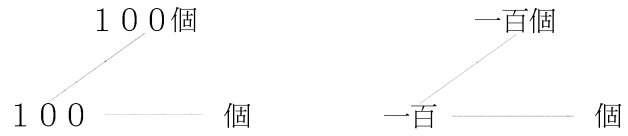


圖3-4-17: Simple Normalization example 1

１００個 -> １００ 個　　一百個 -> 一百 個

演算法可歸納如下：

參數說明：

Node：輸入待正規化的節點

Simple Set: 需要被數字正規化的集合

演算法 Simple_Normalization(Node)

Step 1: If *Node* is listed in *Simple Set*, go to Step 2, or go to Step x.

Step 2: Search for Arabic numerals node in *Node*.

Step 3: Normalize Arabic numerals.

Step 4: if all nodes in *Node* are normalized, go to Step 5, or go back to Step 2.

Step 5: Construct normalized node; Combine all nodes with normalization result.

第四章 實驗結果與分析

4.1 測試語料

我們採用〈中研院平衡語料庫 3.0 版〉做為測試語料，語料庫已經過正確之斷詞與詞類標記，以下為此測試語料的統計資訊：

表 4-1 中研院平衡語料庫 3.0 版語料庫統計

文章篇數	9286
總詞數	5841942
總中文詞數	4883661
中文專有名詞數	94121
外文詞數	27502
標點符號數	930779

4.2 實驗項目

為了測試此中文斷詞器的表現，在這裡設計了三個實驗來進行測試，它們分別是(a)實驗 A：僅詞典之斷詞(b)實驗 B：加上構詞單元之斷詞(c)實驗 C：加上構詞單元之斷詞，輸出時將定量複合詞拆開為定詞與量詞。以下為分項說明：

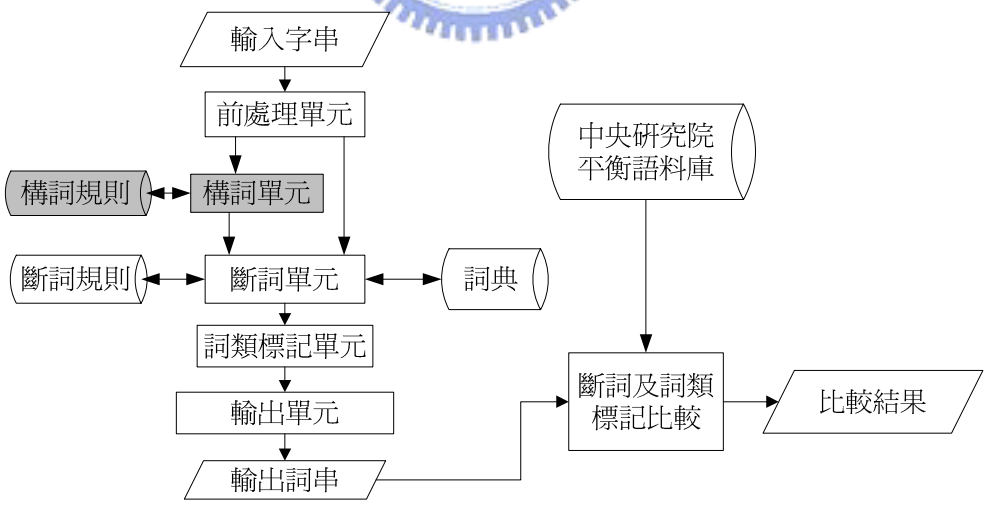


圖4-2-1: 實驗架構圖

(a)實驗 A：僅詞典之斷詞

斷詞單元為整個斷詞器的核心，在這個實驗裡只用詞典中的詞為建立候選詞組的詞集合，不將構詞單元產生的詞列入詞集合，斷詞結果做為其他實驗的基本

比較，實驗之架構圖如圖 4-2-1，注意在此實驗中，我們將構詞單元這模組的功能關閉。

(b)實驗 B：加上構詞單元之斷詞

構詞單元將定量複合詞、時間詞以及重疊詞等構出，與詞典同為候選詞組的詞集合，因此我們將圖 4-2-1 的構詞單元功能開啟。

(c)實驗 C：加上構詞單元之斷詞，輸出時將定量複合詞拆開為定詞與量詞

由於中研院平衡語料庫中的定量複合詞會斷為定詞與量詞，為了分析構詞單元的表現，在輸出單元輸出時，我們將構出之定量複合詞拆解成定詞與量詞的部分，成為與平衡語料庫一樣的分詞結果，如圖 4-2-4 所示；一個定量複合在拆解為定詞和量詞的組合後，若能夠與平衡語料庫有一致的斷詞結果與詞類，便代表此定量複合詞的構出為正確的。相同於實驗 B，我們將圖 4-2-1 的構詞單元功能開啟。

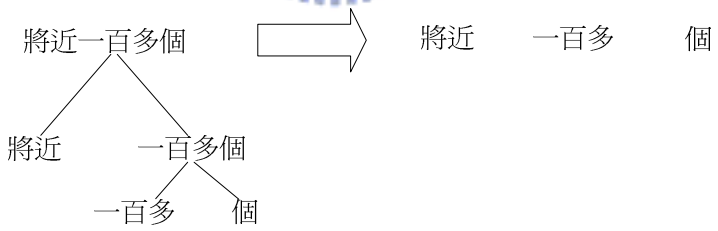


圖4-2-2: 定量複合詞的拆解

要注意的是，有些定量複合詞構詞單元構出的詞，同時有可能是定量複合詞或時間詞，而詞類是由詞類標記單元給予的，例如「１２點」，在平衡語料庫中，若詞類為定量符合詞則斷為「１２　點」，若為時間詞則斷為「１２點」，因此在實驗 C 的輸出，如果詞類標記標為定量複合詞則輸出為「１２　點」，如果標為時間詞，輸出為「１２點」。

4.3 斷詞及詞類標記結果之評量標準

由於專有名詞在整個平衡語料庫裡，佔了將近 2%，因此在評量斷詞器效能時，同時觀察含與不含專有名詞的斷詞結果，在這裡我們定義召回率(recall)及精確率(precision) 做為斷詞結果及詞類標記結果的評量標準，定義如下：

a. 在計算所有中文詞的條件下

$N1$ =（平衡語料庫的中文詞數）

$N2$ =（經斷詞器斷出之中文詞數）

$N3$ =（經斷詞器斷詞且與平衡與料庫一致的中文詞數）

$N4$ =（斷詞結果與平衡與料庫一致，且詞類標記正確的中文詞數）

b. 在不考慮專有名詞的情況下

$N1$ =（平衡語料庫的中文詞數）－（平衡語料庫中的專有名詞數）

$N2$ =（經斷詞器斷出之中文詞數）

－（斷出之詞對應平衡語料庫為專有名詞的詞數）

$N3$ =（經斷詞器正確斷出之中文詞數）

－（正確斷出且對應平衡語料庫為專有名詞的詞數）

$N4$ =（正確斷詞，且正確標記詞類的中文詞數）

－（正確斷詞、標記詞類，且對應平衡語料庫為專有名詞的詞數）

$$\text{斷詞召回率} = \frac{N3}{N1}$$

$$\text{斷詞精確率} = \frac{N3}{N2}$$

$$\text{詞類標記召回率} = \frac{N4}{N1}$$

$$\text{詞類標記精確率} = \frac{N4}{N3}$$

4.4 斷詞結果之比較

各個實驗的結果列於表 4-4-1，很明顯地可以看到實驗 B 的精確率(含或不含專名)，比實驗 A 中的精確率還要高，但是召回率較低，這是因為在加入構詞規則之後，許多的定量複合詞由構詞單元構出成為較長的詞，造成許多與平衡語料庫不一致的結果，而實驗 C 的召回率以及精確率都比實驗 A、B 高，這是因為

表 4-4-1 斷詞結果

	實驗A 僅以詞典斷詞		實驗B 加入構詞單元		實驗C 將定詞與量詞分開	
	含專名	不含專	含專名	不含專	含專名	不含專
*N*1	4883661	4789540	4883661	4789540	4883661	4789540
*N*2	4773152	4679443	4434692	4341015	4654548	4560848
*N*3	4008905	3964142	3844914	3799378	4061316	4015805
召回率	0.821	0.828	0.787	0.793	0.832	0.838
精確率	0.84	0.847	0.867	0.875	0.872	0.880

實驗 C 將構出之定量複合詞，拆解成定詞與量詞的部分，成為與平衡語料庫一樣的分詞標準。另外我們發現，在不計算專有名詞的情況下，每一個實驗下都可以得到較高的召回率以及精確率，可見專有名詞，雖然僅僅在整個語料庫佔了 2%，對於整個斷詞的結果是有影響的，如果專有名詞在語料庫佔有更高的比例，影響一定會更加顯著。

而各個實驗斷詞之詞長分佈見表 4-4-2，很明顯地實驗 A 的平均詞長 1.649 比平衡語料庫的 1.627 長一些，這是由於在我們的詞典中，將許多單字詞合成二字詞，例如：「就-是、不-是」，一字詞加二字詞合成三字詞，例如：「的-時候、自己-的」，二字詞加二字詞合成四字詞，例如：「師範-大學、研究-報告」。而實驗 B 的平均詞長為 1.715，較實驗 A 及平衡語料庫長了許多，主要是因為構詞單元構出的詞總數為 141220 個，佔了斷出詞的 3.18%，且平均詞長為 2.91，

表 4-4-2 斷詞之詞長分佈

	中研院平衡語料庫	實驗A僅以詞典斷詞	實驗B加入構詞單元	實驗C將定詞與量詞
一字詞	2247793	2100131	1743888	1922717
二字詞	2297404	2334963	2326053	2358310
三字詞	265917	257859	266473	275003
四字詞	59395	75896	83738	86140
五字詞	8298	3353	8601	8709
六字詞	2256	641	3034	1991
七字詞	1045	294	1326	882
八字詞	765	15	704	390
平均詞	1.627	1.649	1.715	1.691

而實驗 C 的平均詞長 1.691，比實驗 B 的詞長較接近中研院平衡語料庫的結果，是由於實驗 C 將定量複合詞拆開成較小的定詞與量詞。

實驗 B 雖然有較高的精確度，但並沒有辦法證明加入構詞規則，對斷詞有所幫助，因此表 4-4-3 列出了不算構詞規則構出詞的召回率與精確率，我們可以看到實驗 B 可以得到較高的召回率以及精確率，同時正確斷出的詞數 $N3$，實驗 B 的結果也較實驗 A 多，因此由這樣的統計結果，加入構詞單元，除了把定量複合詞、時間詞以及重疊詞等等構出，也提升了斷詞的正確性。

表 4-4-3 不算構詞規則構出詞之斷詞結果

	實驗A僅以詞典斷詞		實驗B加入構詞單元	
	含專名	不含專	含專名	不含專
N1	4613498	4520544	4613498	4520544
N2	4419603	4337396	4387618	4296066
N3	3798555	3754534	3811579	3768914
N4	3649654	3609498	3663663	3622883
召回率	0.823	0.831	0.826	0.834
精確率	0.859	0.866	0.869	0.877

表 4-4-4：斷詞規則對系統的影響

	實驗A				
斷詞規則	K2	K3	貢獻率 =K2/K1	精確率(K3/K2)	召回率(K3/K0)
一	4692303	3924542	0.9832	0.836	0.80361
二	20875	12451	0.0044	0.596	0.00255
三	47833	38952	0.0100	0.814	0.00798
四	0	0	0.0000	0.000	0
五	26710	17972	0.0056	0.673	0.00368
六	13404	8671	0.0028	0.647	0.00118
七	204	172	0.0001	0.843	0.00003
	實驗B				
斷詞規則	K2	K3	貢獻率 =K2/K1	精確率(K3/K2)	召回率(K3/K0)
一	4314581	3751135	0.9729	0.869	0.76806
二	28379	18714	0.0065	0.659	0.00388
三	50519	42584	0.0114	0.843	0.00872
四	3105	2206	0.0007	0.710	0.00045
五	26447	18187	0.0060	0.688	0.00372
六	11410	8032	0.0026	0.704	0.00164
七	251	216	0.0001	0.861	0.00004

*K*0：平衡語料庫的中文詞數 = 4883661

*K*1：由實驗斷出之中文詞數；實驗 A：*K*1=4773152，實驗 B：*K*1= 4435304

*K*2：經由某個斷詞規則斷出之詞數

*K*3：經由某個斷詞規則正確斷出之詞數

表 4-4-4 列出每一條斷詞規則對於整個斷詞結果的影響，由表中可以發現到，不管是實驗 A 或 B，斷詞規則一：長詞優先，貢獻率都在 0.97 以上，幾乎決定了所有的斷詞結果，因此將構詞單元放置在斷詞單元之前，把較短的詞合為較長的單位，再將所有可能的詞組組合，由斷詞規則來選擇斷詞結果，是適當的做法。

表 4-4-5 與唐【4】斷詞結果之比較

	過去系統		實驗A	
	僅以詞典斷詞		僅以詞典斷詞	
	含專名	不含專	含專名	不含專
*N*1	4854197	4763310	4883661	4789540
*N*2	4781417	4736527	4773152	4679443
*N*3	3989085	3944195	4008905	3964142
召回率	0.822	0.828	0.821	0.828
精確率	0.834	0.833	0.840	0.847
	過去系統		實驗B	
	加入構詞單元		加入構詞單元	
	含專名	不含專	含專名	不含專
*N*1	4854197	4763310	4883661	4789540
*N*2	4543405	4497910	4435304	4341627
*N*3	3798239	3752744	3844952	3799416
召回率	0.782	0.789	0.787	0.793
精確率	0.836	0.834	0.867	0.875

另外我們將斷詞結果，再與唐【4】的系統比較，如表 4-4-4 所示，實驗 A 的結果，與過去的系統僅斷詞單元的結果相近，而實驗 B 的結果就比過去系統加入構詞單元，在召回率以及精確率都有進步，主要的原因是構詞單元構詞的輔助，在稍後會進行構詞單元的分析。

4.5 斷詞結果分析

我們比較實驗 B 與中研院平衡語料庫的斷詞結果，我們發現到有 21.3%的不一致，這些不一致不外乎 1.較中研院斷詞結果長 2.較中研院斷詞結果短 3.搶詞，以下分為這幾項來說明分析

1. 較中研院平衡語料庫之斷詞結果長

這種情形佔所有不一致(21.3%)的 85.8%，而在這 85.8%之中，73.6%是由於詞庫已收錄這些長詞，26.4%是由構詞單元構出，如表 4-5-1、4-5-2 所列出的例子

表 4-5-1：範例-詞庫收錄較長之詞

類型	斷詞器詞庫收錄之長詞	中研院
的+單字詞	的人	的 人
	的是	的 是
單字詞+的	新的	新 的
雙字詞+的	自己的	自己 的
二字詞+二字詞	網際網路	網際 網路
定量複合詞	這個	這 個

表 4-5-2：範例-構詞單元構出較長之詞

類型	器構詞單元構出	中研院
定量複合詞	一個	一 個
時間詞	六月二十七日	六月 二十七日
重疊詞	吃吃看	吃吃 看
重疊詞+的	靜靜的	靜靜 的

我們認為這些斷詞結果，較中研院平衡語料庫佳。

2. 較中研院平衡語料庫短

這種情形佔所有不一致的 10.5%，造成斷出的詞較中研院還要短的主要原因

是詞典中未收錄某個長詞所造成的，表 4-5-3 列出幾個例子

表 4-5-3：範例-構詞單元構出較長之詞

類型	中研院	斷詞器
中文人名	李文秀	李　文　秀
專有名詞	微軟	微　軟
英譯名	蘇魯克	蘇　魯　克
地方詞	廣告系	廣告　系
	波希尼亞	波　希　尼　亞

建議之解決辦法：

(1)增加專有名詞、中文人名的模組於斷詞單元之後

(2)增加新詞於詞典

(3)增加衍生詞構詞單元於斷詞單元之後

3. 搶詞

這種情形佔所有不一致的 3.7%，造成的原因有(a)斷詞規則造成(b)構詞單元造成(b)詞庫未收錄詞造成。

(a)：斷詞規則造成的錯誤，如表 4-5-4 所示：

表 4-5-4：範例-斷詞規則造成的錯誤

造成搶詞之斷詞規則	中研院	斷詞器
斷詞規則一	在 校園 中	在校 園中
斷詞規則二	三百多　人	三百 多人
斷詞規則三	一 個 人 的	一 個人 的
斷詞規則四	十日　內	十　日內
斷詞規則五	的確　是　正面	的　確是　正面
斷詞規則六	台上 演到 台 下	台 上演 到 台 下

(b)：構詞單元造成之錯誤，如表 4-5-5 所示：

表 4-5-5：範例-構詞單元造成之錯誤

中研院	斷詞器
只 剩下 一萬多 隻	只剩 *下一萬多* 隻
為什麼 會 這麼	為 *什麼會* 這麼

在範例中，構詞單元構出不適當的詞，以致於斷詞錯誤。

(c)：詞庫未收錄詞造成之錯誤，如表 4-5-6 所示：

表 4-5-6：範例-詞庫未收錄詞造成之錯誤

中研院	斷詞器
李文秀 眼 中	李 文 *秀眼* 中

在範例中，由於詞典未收錄人名「李文秀」，以致於斷詞錯誤。

以上三種不一致的現象，其中較中研院平衡語料庫長的斷詞，是比較好的結果，這些就已經佔不一致的 85.8%，因此只有剩下 14.2%的不一致(佔所有語料的 3.0%)，需要我們進一步改進；對於較中研院短的結果，我們尚可以在斷詞單元後面再加上專有名詞或衍生詞構詞單元等，將較短的詞合為一個長詞，然而對於搶詞的結果，在斷詞單元之後再挽救是不可能，表 4-5-7 列出由各斷詞規則造成的搶詞的比例。

表 4-5-7：各斷詞規則造成的搶詞的比例

造成搶詞之斷詞規	比例
斷詞規則一	26.40%
斷詞規則二	28.20%
斷詞規則三	20.60%
斷詞規則四	1.20%
斷詞規則五	16.10%
斷詞規則六	7.40%
斷詞規則七	0.10%

接下來我們對於斷詞規則造成的搶詞進行分析：

- 斷詞規則一：

長詞優先所造成的錯誤佔所有搶詞的 26.4%，我們將它分類說明。

(a)詞典未收錄某個衍生詞所造成之搶詞，如表 4-5-8 所示

表 4-5-8：範例 – 詞典未收錄某個衍生詞所造成之搶詞

中研院平衡語料庫	斷詞器之搶詞結
電信所 及 中研院	**電信 所**及 中研院
課務組 主任	**課務 組**主任

由於詞典中沒有「電信所」以及「課務組」，只有「電信」以及「課務」，這樣子造成在斷詞規則一就決定了斷詞結果，但如果在詞典裡有「電信所」以及「課務組」，斷詞可由之後的斷詞規則再決定，還有機會得到正確的斷詞結果。但由詞典大小的考量來看，收錄大量的衍生詞造成詞典過大，因此，以類似定量複合詞構詞的規則法來進行構詞，並且把這個衍生詞構詞單元放置在斷詞單元之前，是可以考慮解決方法。

(b)詞典未收錄專有名詞、地方詞造成的搶詞，如表 4-5-9 所示

表 4-5-9：範例 – 詞典未收錄專有名詞所造成之搶詞

類型	中研院平衡語料庫	斷詞器之搶詞結果
專有名詞	張榮發 覺得	**張 榮 發**覺 得
	昇陽 電腦	**昇 陽**電 腦
	和 約克夏 布丁	和**約 克 夏** 布丁
地方詞	愛國東路 口	**愛國 東 路**口

對於這一類的搶詞狀況，解決的辦法是將常用的專有名詞收錄於詞典之

中，另外也要製作一個專有名詞構詞單元，放置於斷詞單元之前，而對於地方詞，我們必須蒐集且建立地方詞的詞典，方能解決此搶詞狀況。

(c)詞典過於合詞造成的搶詞

表 4-5-10：範例 – 詞典過於合詞造成的搶詞 1

中研院平衡語料庫	斷詞器之搶詞結果
生產 費用 的 一個	生產費 **用的** 一個
攝取來 的	攝取 **來的**
升 高中 的	升高 **中的**
有 機會 在	有機 **會在**
晉兵 的 人數 連 秦兵	晉兵 **的人** 數連 秦兵

這一類的錯誤，至少佔斷詞規則一搶詞的 8.6%，出現最多的是「單字詞 + “的”」這樣的搶詞，其實大部分「單字詞 + “的”」的二字詞，在分詞成單字詞以及“的”之後，語意並不改變(除了少部分的詞，如「目的」)，因此在這裡建議將詞典中分詞後不改變語意的「單字詞 + “的”」詞去除，以減少這樣的搶詞，這些可以考慮分詞的二字詞有「用的」、「來的」、「中的」、「長的」、「似的」、「準的」等，若要考慮將 “的”與前面的詞合併為一個更長的詞單位，應該把此項合詞動作放置在斷詞單元之後。相同的情形也發生在「 “的” + 單字詞」，這些可以考慮分詞的二字詞有「的人」、「的是」、「的心」等。

- 斷詞規則二：

標準差最小所造成的錯誤，佔所有搶詞的 28.2%

由之前的表 4-4-4：斷詞規則對系統的影響所示，可以知道斷詞規則二的精確率是所有斷詞規則中最低的，下面我們將這些搶詞錯誤分類說明：

(a)詞典過於合詞造成的搶詞

表 4-5-11：範例 – 詞典過於合詞造成的搶詞 2

中研院平衡語料庫	斷詞器之搶詞結果
演變成 了	演變 **成了**
走過去 了	走過 **去了**
顯示出 了	顯示 **出了**
董事長 的	董事 **長的**
開發出來 的	開發出 **來的**
委員會 中	委員 **會中**
心理學 上	心理 **學上**

對於這一類的錯誤，佔斷詞規則二搶詞的 20.7%，我們必須將過於合詞的二字詞從詞典中移除，如同之前表 4-5-10 的例子，方能解決此問題。

(b)重疊詞及定量複合詞造成的搶詞

表 4-5-12：範例 – 重疊詞及定量複合詞造成的搶詞

中研院平衡語料庫	斷詞器之搶詞結果
堂堂正正 的	**堂堂 正正**的
這 一剎那	這一 **剎那**

在範例中「堂堂正正」、「正正的」、「堂堂」皆可由重疊詞構詞單元構出，但由於斷詞規則二的緣故，斷出了「**堂堂 正正**的」這樣的結果，但如果加上規則「XXYY“的”」，或是刪除規則「XX “的”」，便可以解決這樣的錯誤。另外，範例中的「這 一剎那」被斷成「這一 剎那」，是由於定量複合詞構詞單元構出不適當的詞，才造成搶詞，因此對於定量複合詞的構詞，因該要加上更嚴格的文法或語意的限制，使得定量複合詞構詞單元構出更適當

的詞。

(c)語意錯誤之搶詞

對於這一類的錯誤，如果能夠加上語法及語意資訊輔助斷詞，這些斷詞的錯誤是可以避免的。

表 4-5-13：範例 – 語意錯誤之搶詞

中研院平衡語料庫	斷詞器之搶詞結果
投 手榴彈	投**手　榴彈**
四百九十多 人	**四百九十　多**人
１９世紀 末	**１９　世紀**末
八十八年 前	**八十八　年**前
大學生 還 比較	大學 **生還** 比較

- 斷詞規則三：

附著語素少者優先，佔所有搶詞的 20.6%

表 4-5-14：範例 – 斷詞規則三造成之搶詞

中研院平衡語料庫	斷詞器之搶詞結果
{一, 二, 三..} 個	{一, 二, 三..} 個人
第一 **大**	第　一大
大 熊貓	大熊　貓
一起　**來**　學習	一　起來　學習

範例中的「{一, 二, 三..} 個人」，這樣的搶詞錯誤佔了絕大部分，而正確的斷詞應該是「{一, 二, 三..}個　人」，但是由於「人」為附著語素，造成在斷詞規則三搶詞，對於這樣的錯誤，我們可以簡單地把定量複合詞「{一, 二, 三..}個」設定較高的斷詞優先順序。

對於其他斷詞規則三造成的搶詞錯誤，我們必須檢討附著語素的求取以及定義，或是增加語法及語意的資訊以幫助斷詞，減少這一類搶詞的現象。

- 斷詞規則四：

候選詞組中定量複合詞字數合最少者優先，僅佔所有搶詞的 1.2%

表 4-5-15：範例 – 斷詞規則四造成之搶詞

中研院平衡語料庫	斷詞器之搶詞結果
二十多 年 前	**二十 多年**前
二十幾 個 人	**二十 幾個**人
及 第六 腦神經	及**第 六** 腦神經

- 斷詞規則五：

一字詞詞頻最高者優先，佔所有搶詞的 16.1%

表 4-5-16：範例 – 斷詞規則五造成之搶詞

中研院平衡語料庫	斷詞器之搶詞結果
怪 石頭	怪石 頭
及 其他	及其 他
的確 是	的 確是

- 斷詞規則六：

總詞頻最高者優先，佔所有搶詞的 7.4%

表 4-5-17：範例 – 斷詞規則六造成之搶詞

中研院平衡語料庫	斷詞器之搶詞結果
台中縣 警局 昨天	台中 縣警局 昨天
風 吹走 了	風吹 走 了
台北 市政府	台北市 政府

對於斷詞規則四至六造成的搶詞，我們發現需要的是語意及語法的輔助，簡單的規則法無法解決這些錯誤。另外由表 4-4-4 所示的斷詞規則對系統的影響分佈表，我們發現到幾乎所有的斷詞(97%)，都在「斷詞規則一：長詞優先」決定斷詞結果了，要解決斷詞規則一後面的幾條斷詞規則造成的錯誤，除了增加語法以及語意資訊幫助斷詞以外，將更多的構詞模組加入構詞單元，如人名的構詞及衍生詞的構詞等，從輸入文句之中構出更長的詞，使得在斷詞單元斷詞時，應用

到斷詞規則一的機率更高，然而大量的構詞規則，可能會造成構出大量不適當的詞，同時也增加了系統的複雜度，因此是否增加構詞規則於斷詞單元之前，需要再進行更深入的討論。

4.6 定量複合詞構詞單元效能之分析

在這裡我們將觀察在實驗 B、C 結果中，斷詞單元斷出且為定量複合詞構詞單元產生的詞，以來分析定量複合詞構詞單元的效能，因此我們依照這些構詞規則的編號，列出某一個由規則構出且斷詞器斷出的詞數目，如表 4-6-1 所示

表 4-6-1：定量複合詞構詞規則使用之分佈

規則標記	構出詞數 N1	N2	rate = N2/N1	規則標記	構出詞數 N1	N2	rate = N2/N1
199	12405	9838	0.793	234	643	597	0.928
200	0	0	0.000	235	476	434	0.912
201	0	0	0.000	236	46	31	0.674
202	0	0	0.000	237	1099	1084	0.986
204	497	479	0.964	238	1603	1330	0.830
205	985	13	0.013	239	958	901	0.941
206	0	0	0.000	240	11	9	0.818
207	0	0	0.000	241	6125	6018	0.983
208	2	0	0.000	242	3069	2985	0.973
209	3723	3289	0.883	243	858	816	0.951
210	52824	51314	0.971	244	2884	2557	0.887
211	1108	321	0.290	245	2299	2275	0.990
212	0	0	0.000	246	446	419	0.939
213	0	0	0.000	247	350	307	0.877
214	1267	1131	0.893	248	824	738	0.896
215	489	477	0.975	249	3961	3757	0.948
216	204	181	0.887	250	94	76	0.809
217	759	720	0.949	251	407	359	0.882
218	75	69	0.920	252	29	26	0.897
219	7303	7031	0.963	253	2	2	1.000
220	4045	3796	0.938	254	0	0	0.000
221	247	104	0.421	255	0	0	0.000
222	104	102	0.981	256	0	0	0.000
223	0	0	0.000	257	475	436	0.918
224	0	0	0.000	258	43	3	0.070
225	38	36	0.947	259	62	0	0.000
226	63	56	0.889	260	0	0	0.000
227	849	732	0.862	261	853	575	0.674
228	3	3	1.000	262	362	213	0.588
229	102	101	0.990	263	7730	5340	0.691
230	382	107	0.280	264	1938	1809	0.933
231	342	341	0.997	265	2485	1921	0.773
232	939	799	0.851	266	2543	2494	0.981
233	1	1	1.000	267	2299	2274	0.989

規則標記	構出詞數 N1	N2	rate = N2/N1	規則標記	構出詞數 N1	N2	rate = N2/N1
268	456	340	0.746	278	414	356	0.860
269	1248	1109	0.889	279	9	9	1.000
270	17	12	0.706	280	52	52	1.000
271	141	134	0.950	281	1455	1237	0.850
272	653	578	0.885	282	61	32	0.525
273	570	438	0.768	283	0	0	0.000
274	624	468	0.750	300	381	348	0.913
275	78	76	0.974	301	676	669	0.990
276	8	8	1.000	302	645	638	0.989
277	2	2	1.000				
Total	141220	127333	0.902				

表中 N1 為實驗 B 中，由某一規則構出之詞數，N2 為定量複合詞中定詞與量詞分開後與中研院平衡語料庫一致的數目，也就是將實驗 C 的結果與中研院平衡語料庫比較的結果，我們以 rate=N2/N1 來評量此構出詞的正確率。由表 4-6-1，我們知道構出詞的總數為 141220 個，在實驗 B 中佔所有斷詞結果的 3.18%，其中有 127333 個與中研院平衡語料庫是一致的，大約 0.902 的正確率，被應用最多的為規則標記 210，也就是構詞規則 NOP1 -> IN1 (DESC) ({半}) M，如「一個」、「五支」，大約佔構出之定量複合詞的 37.4%，正確率大約 0.971，第二多的是規則標記 199 的數詞定詞，如「一百六十」，大約佔構出之定量複合詞的 8.8%，正確率只有 0.793，主要原因是由於數詞定詞出現在括號中造成的不一致，例如「(八十五)」，中研院斷為「(八十五)」，而我們斷為「(　八十五　)」，將括號與數詞定詞分開，這樣的不一致是可以忽略的。

被應用第三多的規則，是規則標記 263 的時間詞(IN1 “年”)，佔所有構出詞之 5.8%，如「13 年」，它的正確率只有 0.691，主要不一致的原因是「13 年」有可能為定量複合詞或時間詞，在進行詞類標記時若標為時間詞，在實驗 C 會輸出為「13 年」，但如果實際上的詞類為定量複合詞，「13 年」應該被輸出為「13　-年」，因此錯誤是由詞類標記單元造成的。

另外在這裡我們分析在搶詞的情況之下，定量複合詞構詞單元構出詞受到各斷詞規則造成之搶詞的比例，如表 4-6-2

表 4-6-2：定量複合詞詞受各斷詞規則造成之搶詞比例

斷詞規則	搶詞數量	比例
斷詞規則一	1370	35.2%
斷詞規則二	1911	49.1%
斷詞規則三	147	3.8%
斷詞規則四	322	8.3%
斷詞規則五	43	1.1%
斷詞規則六	70	1.8%
斷詞規則七	28	0.7%
總數	3891	100.0%

定量複合詞的搶詞(3891 筆)佔所有搶詞(38468 筆)的 10.1%，我們以造成搶詞前三名的斷詞規則來說明：

- 「斷詞規則一：長詞優先」的搶詞，佔 35.2%

中研院平衡語料庫	斷詞規則一下的搶詞結
幾 個 年長 的	**幾個年** 長的
一 回到 家	**一回** 到家
民國 二十八 年	**民國二十八年** 次
五月 初五	**五月初** 五

範例中的「幾個年長的」以及「一回到家」，這種情形是搶了詞典的詞，唯有靠語法及語意的判斷方能解決此一錯誤於斷詞規則一之前，而對於「民國二十八年次」以及「五月初五」的搶詞錯誤，是因為構詞規則不足所造成的，若我們有構詞規則「民國 + 二十八 + 年次」以及「五 + 月 + 初 + 五」，這種錯誤是可以避免的

- 「斷詞規則二：標準差最小優先」的搶詞，佔 49.1%

中研院平衡語料庫	斷詞規則一下的搶詞結
該 鄉公所	**該鄉** 公所
二千多 人	**二千** 多人
二十四日 前	**二十四** 日前
兩百多 年 來	**兩百** 多年來

- 「斷詞規則四：候選詞組中定量複合詞字數合最少者優先」，佔 8.3%

中研院平衡語料庫	斷詞規則一下的搶詞結
四十多 年 來	**四十　多年來**
六十多 年 前	**六十　多年前**

斷詞規則二、四總共造成的的搶詞，佔了定量複合詞搶詞的一半以上，若是加上語意及語法資訊來輔助，這些錯誤是可以避免的。

4.7 重疊詞構詞單元效能之分析

由重疊詞構詞單元構出的詞總共為 4676 個，只佔實驗 B 中斷詞結果的 0.11%，比例十分的少，各個重疊詞構詞規則的構詞結果如表 4-7-1 所示，M1 代表某規則構出之詞數，M2 代表構出之詞與平衡語料庫一致之詞數，由表中我們

表 4-7-1：重疊詞構詞單元的構詞結果

規則編號	型態	M1	M2	M3	精確率 =M2/M1	詞類標記精確率
1000	XX	805	365	110	0.45	0.30
1001	X{"一"}X	313	1	1	0.00	1.00
1002	XXY	155	45	42	0.29	0.93
1003	X{"一"}XY	9	0	0	0.00	0.00
1004	XXYZ	31	5	3	0.16	0.60
1005	X{"一"}XYZ	3	0	0	0.00	0.00
1006	XX{"的"}	1931	1	1	0.00	1.00
1007	XXYY	986	928	767	0.94	0.83
1008	XXXX	23	4	1	0.17	0.25
1009	XYXY	420	84	50	0.20	0.60
總計		4676	1433	975	0.31	0.68

知道整個重疊詞構詞單元的精確率只有 0.31，這是由於平衡語料庫將大部份重疊詞分成短詞而造成的不一致，我們認為斷詞器的分詞結果較佳，以下為一些例子：

規則編號	型態	斷詞器	中研院平衡語料庫
1000	XX	步步	步　步
1001	X{"一"}X	看一看	看　一　看
1002	XXY	說說話	說　說話
1003	X{"一"}XY	定一定神	定　一　定神
1004	XXYZ	修修指甲	修修　指甲
1005	X{"一"}XYZ	伸一伸舌頭	伸　一　伸舌頭
1006	XX{"的"}	亮亮的	亮亮　的
1007	XXYY	辛辛苦苦	辛辛　苦苦
1008	XXXX	對對對對	對　對　對　對
1009	XYXY	如何如何	如何　如何

4.8 詞類標記結果

詞類標記單元的結果列於表 4-8-1，相對於唐【4】的詞類雙連文模型結果(列於表 4-8-2)，詞類標記的結果在召回率以及精確率有明顯的提升，我們首先比較詞類標記現在採用的數學式(3-4-6)以及唐【4】的詞類雙連文模型結果(4-8-1)

表 4-8-1 詞類標記結果

	實驗A		實驗B		實驗C	
	僅斷詞單元		加入構詞單元		將定詞與量詞分開	
	含專名	不含專名	含專名	不含專名	含專名	不含專名
$N1$	4883661	4789540	4883661	4789540	4883661	4789540
$N3$	4008905	3964142	3844914	3751237	4061316	4015805
$N4$	3831661	3786898	3686340	3640804	3882335	3836824
詞類標記召回率	0.803	0.809	0.831	0.839	0.834	0.841
詞類標記精確率	0.956	0.955	0.959	0.958	0.956	0.955

表 4-8-2 唐【4】之詞類標記結果

	唐[4]	
	加上構詞單元	
	含專名	不含專名
$N1$	5781132	5690249
$N3$	4724539	4679044
$N4$	3859785	3826880
詞類標記召回率	0.668	0.673
詞類標記精確率	0.817	0.812

$$P(pos_1^n \mid w_1^n)$$

$$= P(pos_1 \mid w_1)\prod_{i=2}^{n}\left[\frac{P(pos_i \mid w_i)}{P(pos_i)}P(pos_i \mid pos_{i-1})\right] \qquad (3\text{-}4\text{-}6)$$

$$P(pos_1^n)$$

$$= P(pos_1)\prod_{i=2}^{n} P(pos_i \mid pos_{i-1}) \qquad (4\text{-}5\text{-}1)$$

式 3-4-6 多加了詞本身的資訊 $P(pos_i \mid w_i)$，對於某一個詞，它可能的詞類，也有不一樣的機率，然而式 4-8-1 沒有考慮到這個特性，舉一個很明顯的例子：單字詞「的」，可能的詞類如表 4-8-3，可以發現到「的」最有可能的詞類是「DE」，然而唐【4】的系統很容易將「的」的詞類標為「D」副詞或「T」語助詞，是因為沒有考慮機率 $P(pos_i \mid w_i)$ 以及機率 $P(pos_i)$ 這兩項，所以在式 4-8-1 可以說是把每一種詞類發生的機率是視為一樣的，只由詞類雙連文模型來決定詞類，很容易造成這一類詞類標記的錯誤。

表 4-5-3：「的」可能的詞類

	平衡語料庫出現的數目Count("的",pos)	P(pos|"的")
時態標記(Di)	2	0.00
副詞(D)	1	0.00
普通名詞(Na)	6	0.00
語助詞(T)	11581	0.04
地 的 之 得(DE)	285811	0.96
Count("的")	297401	

在這裡我們也統計了平衡語料庫裡，平均每一個中文詞所擁有的詞類數目為 1.06 個，以及 perplexity 如式 4-5-2：

$$perplexity = 2^{H} \qquad (4\text{-}5\text{-}2)$$

$$H = \sum_i P(w_i)\left[\sum_j P(pos_j \mid w_i)\log_2(P(pos_j \mid w_i)\right]$$

我們得到 $perplexity$ =1.40，也就是說如果我們是以亂猜的方法，來決定一個正確斷詞的詞之詞類，應該可以得到詞類標記精確率 1 / 1.40 =0.714，而經過我們詞類標記單元的標記後，可以達到平均將近 0.96 的精確率，所以只要斷詞正確，詞類標記單元就可以有效地給予正確的詞類。

第五章　結論與未來展望

經由第四章的諸多實驗結果與討論，本章我們提出下列的結論以及日後努力的目標：

(1)在斷詞結果的表現方面，精確率可以達到 0.87，召回率可以達到 0.79，斷詞規則一：長詞優先，幾乎決定所有的斷詞結果(佔 97%)。大約有 21%與平衡語料庫不一致的斷詞結果，在這些不一致之中，85.8%是因為斷詞結果較平衡語料庫長，在這 85.8%中詞典收錄長詞貢獻了 73.6%，構詞單元構出長詞貢獻 26.4%，我們認為這些結果較平衡語料庫之結果佳。而斷詞結果較平衡語料庫短的結果，佔所有不一致的 10.5%，這些大不部分被斷為一字詞或二字詞。另外，搶詞僅佔所有不一致 3.7%。

(2)對於斷詞結果較平衡語料庫短的結果(佔不一致的 10.5%)，主要原因是詞典未收錄某個長詞所造成的，這些大部分是專有名詞及地方詞，少數是衍生詞或常用詞，因此製作專有名詞以及衍生詞的構詞單元是必要的，而對於常用的地方詞及專有名詞，我們應該持續更新詞典，將它們納入詞典。

(3)造成搶詞的結果(佔不一致的 3.7%)原因有許多，對於斷詞規則一造成的搶詞，大部分是因為詞典未收錄專有名詞、地方詞或衍生詞所造成，因此是否要將專有名詞和衍生詞的構詞放置在斷詞單元之前，是可以考慮的方法。至於斷詞規則二至七所造成的搶詞，這些的搶詞結果造成語意錯誤，因此如果能在斷詞規則一之後加上更多的語意以及語法資訊幫助斷詞，這些錯誤是可以避免的。另外我們發現到，有些搶詞是因為詞典收錄過於合詞的二字詞所造成，如「用的」、「來的」、「中的」、「長的」、「似的」、「準的」、「的人」、「的

是」、「的心」等，建議將這些詞從詞典中移除。

(4)定量複合詞的構詞，我們對於每一個規則給予相對應的詞類，同時也保留了定量複合詞的樹狀結構，在實驗 C 中，我們將定量複合詞的斷詞結果，拆為與平衡語料庫一樣的分詞標準，一致性可達到 90.2%，造成搶詞的機會低，僅佔所有搶詞的 10.1%，大部分是因為斷詞規則造成的錯誤，因此加上語意語法的輔助是必要的，以來減少搶詞。另外我們也必須持續增加及修改這些定量複合詞之構詞規則，使之更加完善。

(5)對於重疊詞的構詞，雖然在測試語料中十分少，但我們還是必須持續收集整理這些重疊詞的規則，納入更多對話中會出現的重疊詞，例如「對對對」、「是是是」、「對不對」、「是不是」等等。

(6)破音字的處理對於語音合成來說，是很重要的工作，因此未來將破音字處理的模組加入整個斷詞器是必要的工作。

(7)對於文字正規化，我們已提出一個系統化的做法，將文字的正規化與定量複合詞的結構建立其關係，未來必須持續蒐集整理這些文字正規化的規則，使得其功能更加完善。

參考文獻

【1】 Chen, Keh-jiann, Shing-Huan Liu,”Word Identification for Mandarin Chinese Sentences,”Proceedings COLING '92, pp.101-105, Nantes, France,1992

【2】 Mo, Ruo-ping Jean, Yao-Jung Yang, K. J. Chen, Chu-Ren Huang,”Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation,”Proceedings of ROCLING V (R.O.C. Computational Linguistics Conference) pp. 215-223,1992

【3】 Chen, Feng-Yi, Ruo-Ping Jean MO, Chu-Ren Huang and Keh-Jiann Chen, “Reduplication in Mandarin Chinese: Their Formation Rules, Syntactic Behavior and ICG Representation,”Proceedings of ROCLING V (R.O.C. Computational Linguistics Conference) pp. 215-223,1992

【4】 唐大任，“中文斷詞器之研究”，國立交通大學電信工程學系碩士論文，民國九十一年七月

【5】 黃居仁，陳克健，陳鳳儀，魏文真，張麗麗， “「資訊處理用中文分詞規範」設計理念及規範內容”，中央研究院歷史語言研究所，中央研究院資訊科學研究所

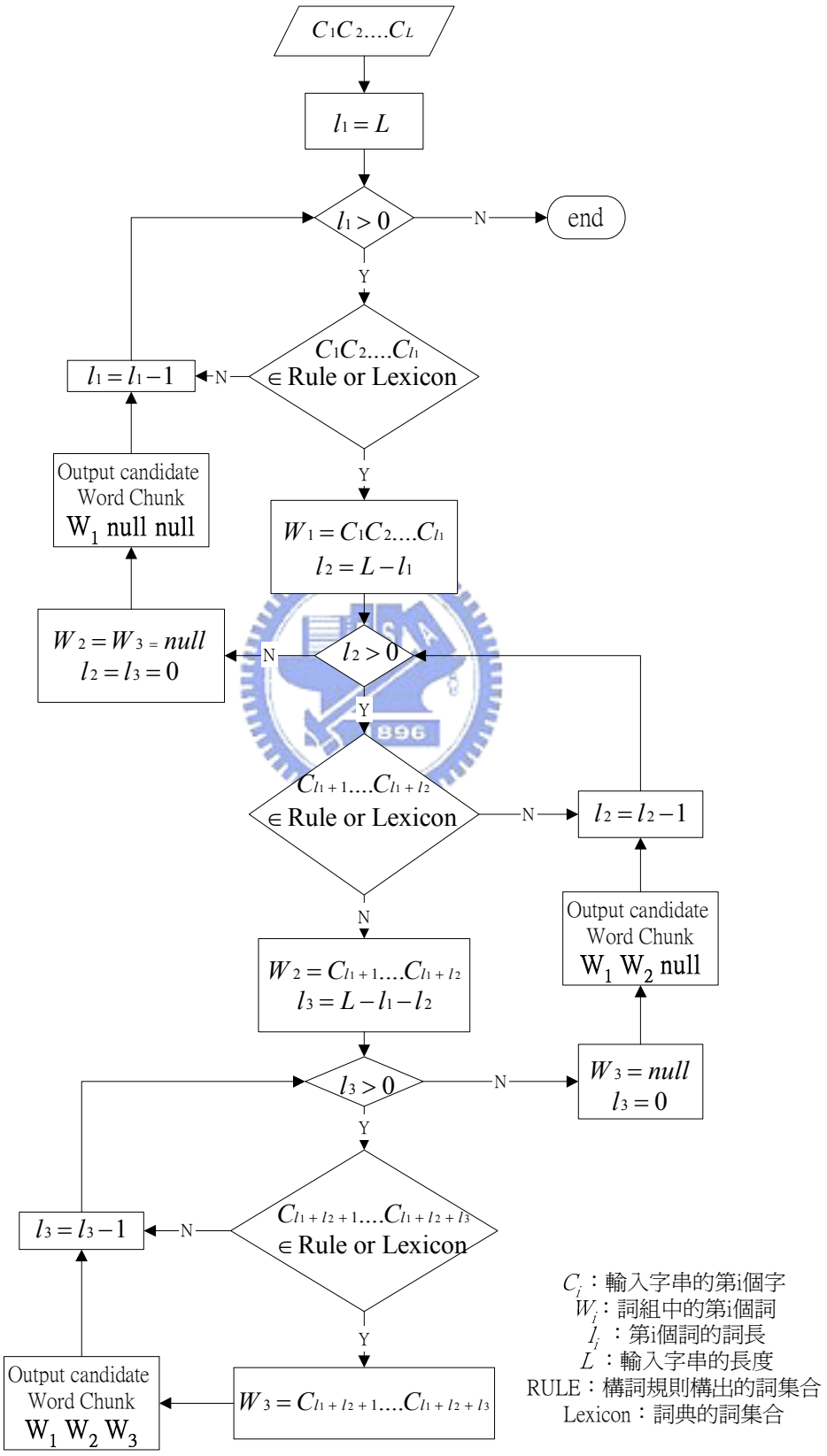
【6】 葉政育，以韻律產生器之效能為基礎設計精簡與高效率之文句分析器，國立台北科技大學電機工程學系碩士論文，民國九十一年六月

【7】 S.M. Katz, “Estimation of probabilities from sparse data for the language

model component of a speech recognizer", IEEE Transactions on Acoustics, Speech, and Signal Processing 1987, Volume: 35 , Issue: 3 , Pages:400 – 401

【8】 S.Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, "The HTK Book (for HTK Version 3.2)", Pages 192 - 199

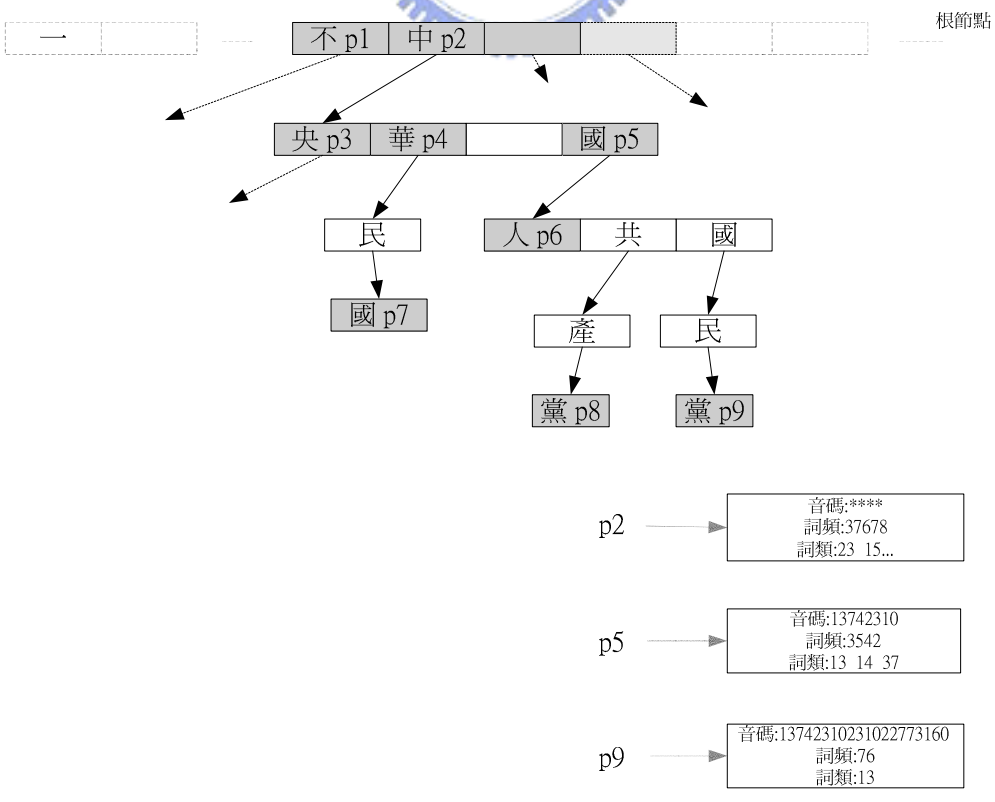
附錄1：建立候選詞組流程圖



C_i：輸入字串的第i個字
W_i：詞組中的第i個詞
l_i ：第i個詞的詞長
L ：輸入字串的長度
RULE：構詞規則構出的詞集合
Lexicon：詞典的詞集合

附錄 2：詞典樹

在進行建立候選詞組時，輸入字串必須向詞典進行比對，為了增加詞典比對的速度，當系統將詞典讀入記憶體時，便會將詞典儲存成詞典樹的資料結構如下圖，整個詞典樹為一個 general tree，同一層的節點(字)依照 Big-5 碼的大小排序好，是一個記憶體動態改變大小的陣列，每增加一個新的詞便會插入樹狀結構之中，深色底的部分代表詞的終端點，代表由根節點走到這個終端點經過的節點為一個詞，在這個終端點也會紀錄這個詞對應詞典資訊(音碼、詞類)的記憶體指標(p*)，而非終端點不為一個詞，沒有對應的詞典資訊。在比對詞典的時候，以輸入字串「中國國民黨」為例，「中」先由根節點開始搜尋(二分法搜尋)，之後「國」由節點「中」的子節點繼續搜尋，以下類推，一但為終端點，便輸出為一個詞，以這個例子會有三個詞「中」、「中國」及「中國國民黨」為詞典比對結果，並得到相對應的詞典資訊。



附錄 3：構詞規則標記

構詞規則標記 (rule_id)	集合
-2	忠,孝,仁,愛,信,義,和,平,智,勇,溫,良,恭,簡,讓
0	一,二,兩,三,四,五,六,七,八,九,十,廿,卅,百,佰,千,仟,萬,億,兆,零,幾,０,１,２,３,４,５,６,７,８,９
1	壹,貳,參,肆,伍,陸,柒,捌,玖,拾,佰,仟,萬,億,兆,零,幾,几
2	甲,乙,丙,丁,戊
3	大,小,整
4	多,餘,半,出頭,好幾,開外,整,正,許,足,之多
5	半,多,許,整,正
6	一,全,滿,整,成,一切,所有
7	多少,若干,幾多
8	很,挺,怪,真,好,極,滿,更,再,頂,最,太,忒,多,夠,非常,異常,十分,尤其,有點,略爲,稍微,比較,不大,過
9	多,許多,許許多多,有些,好些,幾許,有的,少許,多數,少數,大多數,泰半,不少,部分,部份,個把
10	半,若干,有的
11	這,那,哪
12	上,下,前,後,頭,末,次,首,某,另,同
13	本,貴,敝,什麼,啥,諸,何,別,旁
14	不到,上下,左右,不等,以上,以下
15	多,少
16	多,餘,來
17	萬,億,兆
18	點
19	又
20	分之
21	強,弱
22	半
23	雙
24	整整,滿滿
25	好幾
26	數
27	年
28	班
29	他
30	國,省,州,縣,鄉,村,鎮,鄰,里,郡,區,站,巷,弄,段,號,地,樓,街,市,洲
31	該
32	第
33	每
34	各
35	逐
36	另外,近,將近
37	此
38	其它,其他,其餘
39	任何
40	成
41	不到
42	一
43	這,那
44	平方,立方
51	個
52	分
53	秒
54	時,點
55	點
56	小時
57	刻
58	元
59	年
60	月
61	日,號
62	號
63	月份
64	上,下,每,本,元,正
65	元,正
66	度
67	段
68	巷
69	弄
70	之
71	號
72	樓
73	攝氏,華氏
74	華氏
75	零下

76	清晨,凌晨,早晨,早上,晚上,上午,中午,下午,晨間,午間,晚間,半夜,午夜,晨,午,晚,傍晚,深夜,昌午,子時,丑時,寅時,卯時,辰時,巳時,午時,未時,申時,酉時,戌時,亥時
77	咸豐,昭和,道光,正統,元狩,嘉靖,嘉慶,宣統,德裕,建安,元嘉,元朔,雍正,寶元,明治,德光,開皇,開元,開運,洪武,貞觀,正大,正德,正隆,正朔,元始,永樂,永曆,永嘉
77	民國,中華民國,西元,公元
79	春天,春,冬天,冬,夏天,秋,夏,冬季,夏季,春季,秋季,炎夏,春秋,嚴冬,寒冬,秋天,炎暑,秋日,仲夏,初冬,盛夏,早春,孟冬,孟秋,孟夏,孟春,暮秋,暮春,大月,烈暑,隆冬,隆暑,窮冬,夏令,新春,仲冬,仲秋,仲
80	星期一, 星期二, 星期三, 星期四, 星期五, 星期六, 星期日, 星期天,禮拜一, 禮拜二, 禮拜三, 禮拜四, 禮拜五, 禮拜六, 禮拜日, 禮拜天,週一, 週二, 週三, 週四, 週五, 週六, 週日
81	當年,半晌,新年,前夕,老年,前期,暑假,乾季,來年,年關,來世,寒假,開春,假日,三伏,課餘,後葉,花期,中葉,餘暇,雨季,半衰期,冰河期,末代,梅雨季節,暮歲,風季,淡季,當世,檔期,太平盛世,髫年,齠年,停經期,年假,年下,年終,農閒,農隙,農時,來生,糧季,零週期,例假,亂世,過渡期,過渡時期,觀光季,公餘,後期,寒暑,花甲,花季,會期,婚期,黃金時代,舊曆年,今生,今世,經期,青年期,青春期,邪世,宣
82	,今天,昨天,今,當天,元宵,今日,春節,星期天,國慶,下旬,星期日,中旬,除夕,光復節,聖誕,雙十國慶,音樂節,元旦,陰天,潑水節,端午節,勞動節,夏日,星期四,星期一,重陽,耶誕,防空節,復活節,大年初一,當日,燈節,端午,冬至,例假日,狂歡節,驚蟄節,清明,中秋,初一,聖誕節,自由日,安息日,兒童節,元宵節,白露,兵役節,蒲節,芒種,母親節,暮節,佛誕,返校日,父親節,婦女節,大寒,大雪,大暑,單日,端陽,端陽節,冬節,臺灣光復節,團圓節,年節,農曆年,臘八,臘日,六甲,良辰吉日,良日,禮拜天,禮拜日,立冬,立秋,立夏,立雪,立春,感恩節,國定假日,國慶日,鬼節,穀雨,開年,開國紀念日,空軍節,黑道日,黑色星期五,後日,寒食,火曜日,黃道吉日,黃道日,教師節,九九重陽,酒神節,九三軍人節,金曜日,禁煙節,驚蟄,莒光週,莒光日,七七,七夕,秋分,清明節,晴天,情人節,暇日,夏至,下元,下雨天,小
90	本,把,瓣,部,柄,床,處,期,齣,場,朵,頂,堵,道,頓,錠,闋,棟,幢,檔子,檔,封,幅,發,分,份,服,個,根,根兒,管,行,戶,件,家,架,卷,具,節,句,屆,箇,捲,劑,隻,尊,盞,張,枝,支,樁,幀,只,株,折,灶,軸,口,棵,款,客,輛,粒,片,輪,枚,面,門,幕,匹,所,艘,扇,首,乘,襲,頭,條,長條,台,臺,挺,堂,帖,顆,座,則,冊,任,尾,味,位,頁,
90	通,口,頓,盤,局,番
90	對,雙,宗,番,畦,餐,行,身,列,長列,系列,排,長排,副,付,套,筆,串,長串, 掛,幫,房,批,組,窩,綑,捆,群,胎,桌,嘟嚕,部,種,類,樣,樣兒,派,路,趟,落,伙,夥,束,簇,席,疊,紮,色,票,叢,隊,攤,式,蓬,項
90	些,分,分兒,團,堆,泡,綹,撮,把,股,灘,汪,陣,口,口兒,塊,滴,欄,捧,抱,層,重,帶,截,長截,長截兒,截兒,節,節兒,長節,長節兒,段,段兒,長段,長段兒,絲,絲兒,點,點兒,片,縷,部份,部分,坨,匹,疋,階,抔,波,
90	盒,盒子,匣,匣子,箱,箱子,櫃子,櫥,櫥子,籃,籃子,簍,簍子,爐子,包,包兒,袋,袋兒,池子,瓶,桶,聽,罐,罈,盆,鍋,籠,盤,碗,杯,勺,勺子,匙,湯匙,筒,擔,瓶子,桶子,罐子,罈子,盆子,鍋子,籠子,盤子,杯子,筒
90	筆,劃,劃兒,橫,豎,直,撇,捺,挑,剔,鉤,鉤兒,拐,點,格,格兒,國,省,州,縣,鄉,村,鎮,鄰,里,郡,區,站,巷,弄,段,號,地,街,樓,衖,市,洲,部,司,課,院,科,系,級,股,室,廳,會,會兒,陣,陣子,世,輩,輩子,代,學期,學年,年代,下子,版,冊,編,回,章,面,小節,集,卷,面兒,方面,邊,邊兒,頭,頭兒,方,拍,板,眼,程,作,倍,成,分,
90	度,輪,回,次,遍,趟,下,下兒,遭,番,聲,聲兒,響,圈,圈兒,步,把,仗,覺,頓,關,手,手兒,腳,掌,巴掌,拳,拳頭,眼,口,刀,仞,槌,槌子,板,板子,鞭,鞭子,棒,棍,棍子,陣,針,箭,槍,槍矛,砲,場,周,曲,跤,記,回合,票
91	身,頭,臉,鼻子,嘴,肚子,手,腳,桌子,院子,地,屋子,池,腔,家子
92	公厘,公寸,公分,公尺,公丈,公引,公里,市尺,公釐,營造尺,台尺,吋,呎,碼,哩,浬,海浬,庹,噚,尺,里,釐,寸,丈,米,厘,厘米,海哩,海里,英尺,英里,英呎,英寸,米突,米尺,微米,毫米,英吋,英哩,光年
93	公畝,公頃,市畝,營造畝,坪,畝,分,甲,頃,英畝
94	公克,公斤,公噸,市斤,台兩,台斤,日斤,盎司,盎斯,磅,公擔,公衡,公兩,克拉,斤,兩,錢,噸,克,英磅,英
95	公毫,仟克,公撮,升,公升,市升,營造升,台升,日升,品脫,加侖,蒲式耳,公斗,公石,公秉,公合,公勺,斗,
96	微秒,釐秒,秒,秒鐘,鐘,分鐘,刻,刻鐘,點,點鐘,時,小時,更,夜,旬,紀,世紀,輪,天,日,星期,週,周,禮拜,
97	角,毛,元,文,文錢,圓,塊,塊錢,先令,盧比,法郎,法朗,辨士,馬克,鎊,盧布,美元,美金,便士,里拉,日元,日圓,台幣,港幣,人民幣,刀,打,令,綸,籮,大籮,焦耳,千卡,仟卡,燭光,仟瓦,千瓦,伏特,馬力,爾格,瓦特,瓦,卡路里,卡,仟赫,千赫,兆赫,赫,赫茲,位元,莫耳,歐姆,法拉第,安培,分貝,居里,微居里,毫居里,
90	公厘,公寸,公分,公尺,公丈,公引,公里,市尺,公釐,營造尺,台尺,吋,呎,碼,哩,浬,海浬,庹,噚,尺,里,釐,寸,丈,米,厘,厘米,海哩,海里,英尺,英里,英呎,英寸,米突,米尺,微米,毫米,英吋,英哩,光年,公畝,公頃,市畝,營造畝,坪,畝,分,甲,頃,英畝,公克,公斤,公噸,市斤,台兩,台斤,日斤,盎司,盎斯,磅,公擔,公衡,公兩,克拉,斤,兩,錢,噸,克,英磅,英兩,公錢,毫克,毫分,,公毫,仟克,公撮,升,公升,市升,營造升,台升,日升,品脫,加侖,蒲式耳,公斗,公石,公秉,公合,公勺,斗,毫升,夸,夸特,夸爾,石,斛,西西,微秒,釐秒,秒,秒鐘,鐘,分鐘,刻,刻鐘,點,點鐘,時,小時,更,夜,旬,紀,世紀,輪,天,日,星期,週,周,禮拜,月,月份,季,年,年份,歲,載,號,晚,宿,週年,周年,周歲,角,毛,元,文,文錢,圓,塊,塊錢,先令,盧比,法郎,法朗,辨
99	．,•,．
100	%
101	,
102	／
103	：
104	ＦＡＸ：,ｆａｘ：
105	ＴＥＬ：,ｔｅｌ：
106	ＡＭ,ａｍ
107	ＰＭ,ｐｍ
108	$

附錄 4：系統採用之定量複合詞構詞規則

類別	regular expression	POS	範例
NOP_	IN1 -> NO1*	17	一百
	IN2 -> NO2*	17	壹佰
	IN3 -> {IN1,IN2} {多,餘,來} ({萬,億,兆})	17	一百多萬 一百多
	DN -> (IN1) {點} IN1 ;	17	一點三
	FN2 -> (IN1 {又}) IN1 {分之} {IN1,DN} ({強,弱})	20	一又三分之一
	DN_1 -> IN1 {成} (IN1)	20	三成五
	NOP_1 -> IN1 (DESC) {半}	20	四小半 四半
	NOP_1 -> IN1 DESC	20	四大
	NOP_2 -> DESC {半}	20	大半
	NOP_3 -> IN1 PNM	20	一百整 一百多
NOP	NOP1 -> IN1 (DESC) ({半}) M	58	三大個 四小半個 五個
	NOP2 -> DESC ({半}) M	58	大半個
	NOP3 -> IN1 M PNM	58	一百個之多
	NOP4 -> M PNM	101	個之多
	NOP5 -> {IN3,DN,FN2,雙} M	58	一百多萬個
	NOP_6 -> {IN1, IN3,NOP_3,DN,FN2} {平方,立方} Nfg_1	58	一百平方公分
	NOP_6 -> {IN1, IN3,NOP_3,DN,FN2} Nfg_2	58	三畝
	NOP_7 -> {IN1,IN3,NOP_3,DN,FN2} Nfg_1	58	一百公分
	NOP_8 -> {IN1,IN3,NOP_3,DN,FN2} Nfg_3	58	一百多公斤
	NOP_9 -> {IN1, IN3,NOP_3,DN,FN2} Nfg_4	58	一公升
	NOP_10 -> {IN1, IN3,NOP_3,DN,FN2} Nfg_5	58	一小時
	NOP_11 -> {IN1,IN3,NOP_3,DN,FN2} Nfg_6	58	一元
	ONP -> ON M	58	甲方 乙方
RD	RNOP2 -> {半} M	101	半個
	RNOP3 -> {DESC,成} M	101	大個
	RD10 -> ({NOP1,NOP_1 }({又}) {NOP1,NOP_1} (({又}) {NOP1,NOP_1})*	58	一個又一個
	RD11 -> RNOP2 RNOP2	58	半個半個
	RD12 -> RNOP3 RNOP3	58	大個大個
	RD14 -> {這,那} {一} M M	58	這一支支
	RD13 -> {一} M M	58	一個個
WQP	WQP -> WQ M	58	整個
	WQP -> WQ Nff	58	全身
	WQP -> {整整,滿滿} {NOP1, NOP2,NOP3,NOP5,NOP_6, NOP_7,NOP_8,NOP_9,NOP_10,NOP_11}	58	整整一百個
WQP_	WQP_ -> {整整,滿滿} {IN1,IN3,NOP_3,DN,FN2}	17	整整一百
QQP	QQP -> QQ {NOP4,M}	58	若干個之多 若干個
DQP	DQ2 -> DFa {多,少}	20	很多
	DQP1 -> {好幾} {M,NOP1, NOP2,NOP3,NOP5,NOP_6, NOP_7,NOP4,NOP_8,NOP_9,NOP_10,NOP_11}	58	好幾十個 好幾大半個 好幾個之多 好幾個 數個之多
	DQP2 -> {DQ1,DQ2} M	58	那麼多個
DQP_	DQP_1 -> {好幾} {IN1,IN3,NOP_2,NOP_3}	20	好幾百
PQP	PQP1 -> {數} {M,NOP1,NOP2,NOP3,NOP4,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10 ,NOP_11}	58	數百個 數大半個 數百個之多 數個
	PQP2 -> PQ {NOP4,M}	58	若干個之多 半個之多
PQP_	PQP_1 -> {數} {IN1,IN3}	17	數百 數十餘萬
CNP	CNP -> IN1 {年} {IN1,ON,N} {班}	58	一年一班 一年甲班 一年忠班
DSP	DSP1 -> DS M	58	本項 貴班
	DSP1 -> {他} {國,省,州,縣,鄉,村,鎮,鄰,里,郡,區,站,巷,弄,段,號, 地,樓,衖,市,洲} ;	58	他國
	DSP2 -> {該} {M,NOP1,NOP2,NOP3,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_ 11,PQP1}	58	該一支 該數十支
DSP_	DSP_2 ->{該} {IN1,IN3,DN,FN2,NOP_2,NOP_3,PQP_1}		該數十

OSP	OSP1 -> {第} {NOP1, NOP2,NOP3,NOP5,NOP_6, NOP_7,NOP4,NOP_8,NOP_9,NOP_10,NOP_11}	58	第一千個
	OSP2 -> {每} {PQP,NOP1,NOP2,NOP3,NOP5,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10, NOP_11}	58	每一百個
	OSP2 -> {各} {DESC,M,雙,XQP,NOP}	58	各項
	OSP2 -> {逐} M ;	58	逐項
	OSP3 -> {另外,近,將近} {PQP,NOP1,NOP2,NOP3,NOP5,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10, NOP_11} ;	58	將近數百個 將近一百個
	OSP4 -> OS {PQP,NOP1,NOP2,NOP3,NOP5,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10, NOP_11} ;	58	前一百個
OSP_	OSP_1 -> {第} {IN1,IN3,DN,FN2,NOP_3}	17	第一百
	OSP_2 -> {各} {IN1,IN3,DN,FN2,NOP_3}	17	各一百
	OSP_2 -> {每} {IN1,IN3,DN,FN2,NOP_3}	17	每一百
	OSP_3 -> {另外,近,將近} {IN1,IN3,DN,FN2,NOP_3} ;	17	將近數百 將近一百
	OSP_4 -> OS {IN1,IN3,DN,FN2,NOP_3} ;	17	前一百
DDP_	DDP_1 -> DD	58	這一百項
	DDP_2 -> {此} {WQP_,DQP_1,DQP_2,PQP_1,IN1,IN3,DN,FN2,NOP_3} ;	58	此一百個
OHSP	OHSP -> ({其它,其他,其餘}) {任何} {PQP1,NOP1,NOP2,NOP3,NOP5,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_11}	58	任何其餘一百個
OHSP_	OHSP_ -> ({其它,其他,其餘}) {任何} {IN1,IN3,DN,FN2,NOP_3} ;	58	其餘任何一百個
TIME	TDM -> {IN1} {個} ({多,餘,來,半}) Nfg_5	58	一個多月
	STDM -> IN1 {分} {IN1 {秒} (IN1)}	58	一分十五秒三
	TDM1 -> IN1 {小時} (STDM) (TPNM)	58	一小時一分秒
	TDM2_ -> IN1 {時,點} IN1 {刻} (TPNM)	16	一點一刻整
	TDM2 -> IN1 {小時} IN1 {刻} (TPNM)	58	一小時一刻整
	TDM2 -> ({Ndaac,Ndaad}) {元} {年} ({元} {月} (IN1({日,號})))	16	民國元年元月十號
	TDM2 -> ({Ndaac,Ndaad}) IN1 {年} ({元} {月} (IN1({日,號})))	16	民國三十年元月十號
	TDM2 -> ({Ndaac,Ndaad}) {元} {年} (IN1 {月} (IN1({日,號})))	16	民國元年四月十號
	TDM2 -> ({Ndaac,Ndaad}) IN1 {年} (IN1 {月} (IN1({日,號})))	16	民國三十年四月十號
	TDM3 -> ({Ndaac,Ndaad}) {元} {年} {元} {月份} ;	16	民國元年元月份
	TDM3 -> ({Ndaac,Ndaad}) IN1 {年} {元} {月份} ;	16	民國七十年元月份
	TDM3 -> ({Ndaac,Ndaad}) {元} {年} IN1 {月份} ;	16	民國元年三月份
	TDM3 -> ({Ndaac,Ndaad}) IN1 {年} IN1 {月份} ;	16	民國五年三月份
	TDM4 -> IN1 {月} (IN1 ({日,號})) ;	16	一月一日
	TDM4 -> {元,正,上,下,每,本} {月} (IN1 ({日,號})) ;	16	本月四號
	TDM5 -> IN1 {日,號};	16	一號
	TDM7 -> Ndabd1 (Ndabe) TDM1_ ;	16	星期一傍晚五點
	TDM8 -> Ndadb1;	16	星期一傍晚
	TDM9 -> Ndabe TDM1;	16	傍晚五點
	TDM10 -> {每,上,下,本} ({個}) {TDM7,TDM8};	16	每個禮拜五
LLP	LLP -> IN1 {度} (IN1 {分} (IN1 {秒}))	15	一度一分一秒
ADP	ADP -> (IN1 {段}) (IN1 {巷}) (IN1 {弄}) IN1 ({之} IN1) {號} (IN1 {樓}) ({之} IN1)	14	一段一號
TDP	TDP -> {攝氏,華氏} ({零下}) {IN1,DN} {度}	58	攝氏五度
BD	BD2 -> {NOP1,NOP2,NOP3,NOP4,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_11} BD	58	一百個左右
	BD2 -> {不到}{NOP1,NOP2,NOP3,NOP4,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10, NOP_11}	58	不到一百個
BD_	BD_2 -> {IN1,IN3,NOP_2,NOP_3,STDM,LLP,TDP,STDM,TDM1~10…..etc} BD	17	一百左右
	BD_2 -> {不到} {IN1,IN3,NOP_2,NOP_3,STDM,LLP,TDP,STDM,TDM1~10…..etc}	17	不到一百
MONEY	MON -> $ {INI,IN3,DN}	58	$100

附錄 5：詞類表

編號	標記	詞類	編號	標記	詞類
1	A	非謂形容詞	**24**	Nh	代名詞
2	Caa	對等連接詞	**25**	I	感嘆詞
3	Cab	連接詞，如：等等	**26**	P	介詞
4	Cba	連接詞，如：的話	**27**	T	語助詞
5	Cbb	關聯連接詞	**28**	VA	動作不及物動詞
6	Da	數量副詞	**29**	VAC	動作使動動詞
7	Dfa	動詞前程度副詞	**30**	VB	動作類及物動詞
8	Dfb	動詞後程度副詞	**31**	VC	動作及物動詞
9	Di	時態標記	**32**	VCL	動作接地方賓語動詞
10	Dk	句副詞	**33**	VD	雙賓動詞
11	D	副詞	**34**	VE	動作句賓動詞
12	Na	普通名詞	**35**	VF	動作謂賓動詞
13	Nb	專有名詞	**36**	VG	分類動詞
14	Nc	地方詞	**37**	VH	狀態不及物動詞
15	Ncd	位置詞	**38**	VHC	狀態使動動詞
16	Nd	時間詞	**39**	VI	狀態類及物動詞
17	Neu	數詞定詞	**40**	VJ	狀態及物動詞
18	Nes	特指定詞	**41**	VK	狀態句賓動詞
19	Nep	指代定詞	**42**	VL	狀態謂賓動詞
20	Neqa	數量定詞	**43**	V_2	有
21	Neqb	後置數量定詞	**44**	DE	的，之，得，地
22	Nf	量詞	**45**	SHI	是

23	Ng	後置詞	**46**	FW	外文標記

表 3-3-2：標點符號及其他詞類標記

編號	標記	說明
47	DASHCATEGORY	—
48	ETCCATEGORY	…
49	COMMACATEGORY	，
50	PERIODCATEGORY	。
51	QUESTIONCATEGORY	？
52	COLONCATEGORY	：
53	SEMICOLONCATEGORY	；
54	EXCLANATIONCATEGORY	！
55	PARENTHESISCATEGORY	「」( )【】
56	PAUSECATEGORY	、
57	SPCHANGECATEGORY	‖
58	DM	定量複合詞
100	BM	附著語素

附錄 6：文字正規化規則表

範例	語音讀法	規則表示法
.9	點九	201(99:"點":"3262":0|199:199:199:1)
90.9	九十點九	201(199:NTN:NTN:0|99:"點":"3262":1|199:199:199:2)
90%	百分之九十	202(199:NTN:NTN:1|100:"百分之":"306111481001":0)
90.9%	百分之九十點九	202(201:201:201:1|100:"百分之":"306111481001":0)
3:21:3am	上午三點二十一分三秒	269(199:NTN:NTN:1|103:"點":"3262":2|199:NTN:NTN:3|103:"分":"1148":4|199:NTN:NTN:5|106:"上午":"41523197":0),0:"
10:2:6pm	下午十點二分六秒	269(199:NTN:NTN:1|103:"點":"3262":2|199:NTN:NTN:3|103:"分":"1148":4|199:NTN:NTN:5|107:"下午":"42253197":0),0:"
2004/3/9	兩千零四年三月九日	262(199:NTN:NTN:0|102:"年":"2264":1|199:NTN:NTN:2|102:"月":"4387":3|199:NTN:NTN:4),0:"日":"4004":5
Fax	傳真	"ＦＡＸ":"傳真":"23391134":0
Tel	電話	"ＴＥＬ":"電話":"42624301":0
$100	一百元	276(108:"元":"2393":1|199:NTN:NTN:0)
$100.1	一百點一元	276(108:"元":"2393":1|201:201:201:0)