口語和書面語中漢語反身詞的分布情形與指涉關係*

漆聯成 陳沛麒

國立政治大學語言學研究所

摘要

本文以實際的語料檢視漢語反身詞"自己"在口語和書面語使用上的實際分佈情況及其與前行語的指涉關係。研究的數據顯示：首先，無論是口語或書面語，漢語反身詞及其前行語的約束關係大都超越了句子的層次。以約束關係的範疇而言，是言談約束多於近距約束多於長距約束。其次，就單純反身詞和複合反身詞的出現頻率而言，單純反身詞在書面語的出現頻率高達百分之八十七，而在口語中則降爲百分之五十。顯示單純反身詞在口語和書面語的使用情形有顯著的差距。

一、前言

漢語反身詞在近幾年受到多數句法學家的熱烈討論(Tang [19][20]; Huang and Tang [9]; Cole et al. [5]; Cole and Song[6]; Battistella and Xu [1]; Li [17])。此現象導因於漢語反身詞的約束關係在表面上明顯的違反了Chomsky 的約束原則A (Binding Principle A)(Huang [13])。爲了解釋漢語反身詞的長距約束現象(long-distance binding)，管束句法學家提出了漢語反身詞在不同的句法層次，例如邏輯形式(logic form)上，仍遵守局部約束(local bound)的限制。然而許多語言學家從不同的角度反駁管束句法學

* 二位作者對此研究都作了相同的貢獻。姓名的順序不表示何者的貢獻較多或較少。作者特別要感謝李櫻教授，感謝她不斷的鼓勵和指導以及在不同的階段提供重要的建議。此外，我們還要感謝兩位Rocling 的審稿者提供寶貴的意見，以及國立政治大學語言學研究所劉淑梅，曾惠鈴，和時雪煒同學提供部份口語語料。最後，我們對於葉瑞堂先生在電腦輸入上的協助表示感謝。當然，本文的任何缺失，仍是作者的責任。

派的論證。Chen [4]從功能語法的觀點，認爲漢語反身詞的約束關係是由[主題性]及[基點度]([high topicality] and [pivot])二個概念所控制。文中更指出了漢語反身詞有言談約束(discourse binding)的現象(見Chen [4])。[1] 語用學家Huang [11][12]則運用了新格來斯原則(Neo-Gricean Principle)來檢視漢語反身詞的功能和指涉關係。Chen [3] 也在計算語言學的範疇中討論照應詞的地位。

本文由一個全新的方式出發，以眞實的口語及書面語語料，檢視漢語反身詞的實際分布情況及其與前行語(antecedent)的關係。過去無論是管束句法學家，功能句法學家或語用學家多鐘情於漢語反身詞的長距約束和指涉關係的討論，卻少有人以實際語料檢視漢語反身詞“自己”在口語和書面語的分佈和指涉情形。[2] 經由本文的量化統計，我們可以更了解反身詞“自己”的分布和指涉情況。藉此檢驗之前相關文獻對反身詞“自己”的討論是否朝正確的方向發展，並彌補前人研究的不足。我們相信藉由本文提出的統計數據及研究結果，將可以對漢語反身詞的研究提供一個新的且合理可行的方向，以期未來對漢語反身詞的研究有更大的貢獻。

全文共分五節。在第一節前言之後，第二節討論漢語反身詞的現象，包括反身詞種類，分佈情形，和指涉關係，而第三節說明本文的研究方法。第四節是結果與討論。第五節是結論與建議。

二、漢語反身詞的現象

(一) 種類

---

[1] Chen [4]將言談約束定義爲，“自己”與其前行語並非出現在同句中，而其前行語爲整段言談的主題，並且跨越了句子的層次。(The antecedent of *ziji* are not in the same sentence as the reflexive. Rather, they are topics of the whole discourse paragraph that are across not only clausal, but also sentential boundaries.)

[2] 見 Huang [10]與 Li [15]. Li [15]以實際書面語語料提出漢語照應詞在篇章使用上的階層性結構。

漢語有兩種反身詞 ，一種是單純反身詞(bare reflexive)“自己” ，另一種是複合反身詞(compound reflexive)，即“{他/你/我}(們)自己”二種(湯 [21])。本文以單純反身詞做爲研究重點。

(二) 分布情形

漢語反身詞可以出現在任何名詞片語可出現的位置，茲列舉如下。

A. 主語

(1) 我希望(我)自己去。

B. 賓語

(2) 他瞧不起自己。

C. 間接賓語

(3) 小明送了自己一個禮物。

D. 名詞組主語

(4) 小華愛上了自己的同事。

E. 介詞組賓語

(5)李四把自己鎖在房裡。

此外，Li and Thompson (1981)還指出“自己”除了是一個反身詞之外，也可以是一個副詞。如(6):

(6) 我自己要去。

(三) 指涉特性

漢語反身詞的特性之一就是長距約束現象。[3] (7)句中，“自己”的前

---

[3] 一般認爲漢語反身詞的前行語必須是主語和有生(animate)，這就是平常熟知的主語取向(subject orientation)原則。

(i) 那件事給李四帶來了自己都意想不到的衝擊。(Kao 1993:51)

(ii) 張三告訴李四自己的過去。

在(i)句中，“那件事”非有生，故無法充當爲前行語。(ii)句中的“李四”不是主詞，所以不能做爲“自己”的前行語。本文暫時不討論主語取向的特性留待未來作近一步的研究。

行語可以是句中任何一個主詞。英語則無此現象，如例(8)。[4]

(7) 小華以爲小明知道小莉喜歡自己。

(8) ~~John~~ thinks that Michael loves himself.

三、研究方法

爲了更進一步了解“自己”長距約束關係和分佈情形，我們使用眞實語料，根據“自己”的分布狀況以及其前行語出現的領域做了統計，來檢驗這個現象。以下是我們的語料來源及分類標準：

(一) 語料來源

在口語資料部分，我們從大約一百分鐘(約三萬字)的對話錄音中(32分鐘的日常對話，45分鐘收音機節目訪談對話，及26分鐘的電視節目"女人女人"座談對話)找到62個“自己”。在書面語資料部分，我們採用1995年《大學報》五月份的第一、二、三週及四月分第四週週報，在大約五萬四千字中，我們找到119個“自己”。[5]

(二) 分類標準

A、複合反身詞與單純反身詞

我們的分類標準是將單純反身詞和複合反身詞分別計算，因爲複合反身詞的指涉現象異於單純反身詞。

B、分布狀況

我們將“自己”出現的位置分爲：(A)主語，(B)賓語，(C)間接賓語，(D)介詞組賓語，及(E)名詞組主語等五類。以本文所收集的語料舉例

---

[4] 底線代表相同指涉，刪除線代表非相同指涉。

[5] 以下的句型暫時未列入統計。如，自己管理自己，自己點給自己。 Li and Thompson (1981) 將此種句型歸類爲 generic 用法故不列入討論。因爲時間限制，以至於口語語料數量與書面語數量有差距，在未來的研究中會增加更多的語料。

說明如下：

(A)主語： 出現在主語位置。

(9a)而這次創作的靈感是在他上課上到一半時才想到，不過當時並沒有想出第三張圖的概念，但自己一直希望是四張卡片爲一套。

(9b) 我願意・・・燈起燈落 、幕升幕降 ，每當自己結束一個角色，再接演新戲時就是生命新里程的開始。

(9c) 對書很有感情的他，當初學得卻是醫學技術，畢業後覺得自己適合開書店。

(9d) 電腦就會幫你找出適合同學的姓名 ，或者是輸入自己欲查詢的星座。

(B)賓語：

(10) 自由社會是假設所有成年人都有能力照顧自己。

(C) 間接賓語：

(11) ・・・在演出前幾天會刻意減少練習 ，給自己多些時間思考樂曲的內涵。

(D) 介詞組賓語：

(12) 學生不要把自己當成弱勢族群。

(E) 名詞組主語：

(13) 學生已長大 ，已可表達自己的意見。

C、指涉關係

我們把反身詞的前行語出現的領域分爲：(A) 近距約束，(B)長距約束，和(C) 言談約束三類。

(A).近距約束 (Local Binding)

指涉關係是局部控制。

(14) 他們紛紛取出自己的作品放在太陽底下曬。

(B).長距約束 (Long-Distance Binding)

前行語和反身詞在同一句內，但在表層結構上並非受到局部控制。通常句中的主要動詞爲"知道"、"表示"、"以爲"、和"希望"等等。

(15) 他清楚地知道自己不會死。

(16) 黃議震表示這次展覽並不在炫耀自己的收藏。

(C).言談約束 (Discourse Binding)

如果前行語出現的位置並非和反身詞在同一句內，而是在整段言談中居於主題的地位，則將此前行語歸於這一領域。我們同時也規定反身詞的句內前行語必須是名詞組或是代名詞，才作長距和近距的分析。如果離反身詞最近的前行語是空號代詞(zero anaphor)，而且此空號代詞受到整段的主題所控制，這種情況就歸於言談約束。例如：

(17) 蔣爲文表示，$[e1]_{NP}$[6]當初參加學生台文促進會的泰雅營，$[e2]_{NP}$ 看到泰雅族人母語流失的情形相當嚴重，$[e3]_{NP}$ 有感而發，所以$[e4]_{NP}$ 才會用台文寫下自己的心情。

在(17)句中，離"自己"最近的前行語爲[e4]，從句法的角度來看，這是一個近距約束的關係。但是，此前行語[e4]爲空號代詞且受到此段的主題"蔣爲文"的控制。而此段的主題"蔣爲文"和[e1], [e2],[e3], 以及[e4]共同形成一個主題串(Topic chain)。[7] 此一情形即爲本文所稱之言談約束。[8]

---

[6] [e]表示空號代詞。

[7] Li [15]指出所謂主題串是由一組關於相同主題的子句所組成。其主題出現在一段言談的首句，其後子句主語皆爲空號代詞。 她進一步指出空號代詞用於子句群中形成單一主題串，而代名詞標示新的主題串以貫聯言談主題。(Zero anaphora is used between clauses which form a single topic chain, while pronominal anaphora occurs to mark the beginning of a new topic chain where topic continuity is preserved.)

[8] 在長距和近距中也可以做主題的分析，但此一問題本文留待未來再探討。 Her [8] 將言談層次和句法層次的主題加以區別定義(詳見 Her[8])。他將言談層次的主題稱 frame, 將句法層次的主題稱爲 topic 。句法層次的主題通常是言談層次的主題，當句法層次的主題未出現時，則通常主語佔有此言談主題的功能。本文採取 Her [8]的立場，不否認句中也有言談主題的存在。但是，在本文中，我們暫時將言談

四、結果與討論

(一) 指涉關係

(A) 反身詞“自己”在書面語中的指涉關係

根據語料統計結果，共找到 104個單純反身詞及15個複合反身詞。我們將這104個單純反身詞依分類標準統計列表說明如下。表一標示實際分佈個數，表二則將數據標準化。

表一、書面語中“自己”的分佈情形和指涉關係

書面語	主語	賓語	間接賓語	介詞組賓語	名詞組主語	總和
言談約束	19	1	1	1	25	47
長距約束	11			3	6	20
近距約束		8		1	28	37
總和	30	9	1	5	59	104

表二、表一標準化

書面語	主語	賓語	間接賓語	介詞組賓語	名詞組主語	總和
言談約束	18.2%	0.9%	0.9%	0.9%	24%	45%
長距約束	10.5%			2.8%	5.7%	19%
近距約束		7.6%		0.9%	26.9%	36%
總和	28.8%	8.5%	0.9%	4.8%	56.7%	

根據表二我們發現，在書面語中反身詞與其前行語的指涉關係頻率高低依序爲：

言談約束 (45%)> 近距約束 (36%)> 長距約束 (19%)

(B) 反身詞“自己”在口語中的指涉關係

根據語料統計結果共找到 31個單純反身詞及31個複合反身詞。我們將這31個單純反身詞依分類標準統計列表說明如下。表三標示實際分佈

層次的主題保留給句子以上的單位。

個數，表四則將數據標準化。

表三、口語中“自己”的分佈情形和指涉關係

口語	主語	賓語	間接賓語	介詞組賓語	名詞組主語	總和
言談約束	11	2		1	4	18
長距約束	3				1	4
近距約束		3	3		3	9
總和	14	5	3	1	8	31

表四、表三標準化

口語	主語	賓語	間接賓語	介詞組賓語	名詞組主語	總和
言談約束	35.4%	6.4%		3.2%	12.9%	58%
長距約束	9.6%				3.2%	12.9%
近距約束		9.6%	9.6%		9.6%	29.0%
總和	50%	14.2%	10.7%	3.5%	32%	

根據表四，口語中反身詞與其前行語的指涉關係頻率高低依序爲：

言談約束 (58%)＞ 近距約束 (29%)＞ 長距約束 (12.9%)

根據統計結果，不論在口語或書面語，漢語反身詞的指涉關係都是以言談約束爲最多。近距指涉居次。長距指涉最低。仔細檢視實際語料的長距約束關係。我們發現，長距約束關係的句子，在口語語料中有 4 句，出現比例約 12.9%；書面語語料中有 20 句，出現比例約 19%。本文研究提出二點重要的發現：第一，過去語言學家熱烈討論的長距約束現象，在實際的使用上不到二成。這顯示出漢語反身詞未來的研究方向，尚有很大的空間，例如，言談的原則。第二，值得注意與重新思考的是，先前的研究不斷以超過兩個以上局部領域的複雜長距約束關係作爲討論的對象，如上述例句(7)，重述於例(18)。

(18) 小華以爲小明知道小莉喜歡自己。

但是，不論在本文的口語或是書面語語料中，並未發現此種超越兩個局部領域的長距約束現象。茲舉口語例句(19)-(20)及書面與例句(21)-(26)

如下：

(19) 你們覺得當兵前跟當兵後自己有什麼最大的收穫・・・

(20) 所以媽媽是比較包容所有錯誤的人，所以他們如果有什麼覺得自己害羞的事情，不敢說的事情・・・

(21) 會計系二年級的朱書毓就覺得自己做得還不錯。

(22) 馬傳堯表示上大學學麥克筆字後自己也希望做些有創意的東西。

(23) 中山海洋環境工程系四年級張景翔說，他並不知道自己因爲沒有出席週會會被記申誡。

(24) 另一位主持人輔仁大學大傳系新聞組二年級林立綺說，爲了自己將來的飯碗問題，會更在意一些傳播生態上的問題。

(25) 學生還可以現場報名自己動手染衣服及印製圖案，現場到處可見穿著自己成品的學生。

(26) 他指出賽車是一項和自己競爭的活動，相較大部份參賽者在車子改裝上投下數十萬元，他只用不到五萬元改良底盤及輪胎定位・・・

語料顯示，在實際語料中的長距約束關係，不論“自己”的語法關係爲何，其指涉範圍並未找到超過二個局部領域的句子，並且我們發現這些在實際使用上的長距約束關係，其反身詞與前行語的指涉關係相當明顯，不會有潛在前行語過多的現象。Xu〔22〕指出漢語反身詞“自己”是個不折不扣長距離照應語，在表層結構上並不受區域性條件限制，至於選用哪個主語爲其前行語，Xu〔22〕並未具體指出如何選擇。然而，本文分析已經反映出語用事實。那就是，爲了避免溝通時的語意岐義，及潛在前行語過多，長距約束關係在實際使用的頻率非常低。即使使用長距約束關係，指涉範圍也會在兩個局部約束範圍內，以提供明確的指涉對象。

(二) 分佈情形

由表二得到書面語中漢語反身詞其語法關係的階層性排列爲：

(27) 名詞組主語 > 主語 > 賓語 > 介詞組賓語 > 間接賓語
(56.7%) (28.8%) (8.5%) (4.8%) (0.9%)

由表四得到口語中漢語反身詞其語法關係的階層性排列爲：

(28) 主語 > 名詞組主語 > 賓語 > 間接賓語 > 介詞組賓語
(50%) (32%) (14.2%) (10.7%) (3.5%)

(27)和(28)說明反身詞“自己”在口語和書面語語料中扮演不同的指涉角色及句法上的功能。比較(27)與(28)的階層性排列，我們發現一個很有趣的現象。在口語部份，反身詞“自己”最常以主語的語法關係身份出現，其比率高達45%。在書面語中，反身詞“自己”卻以名詞組主語的語法關係身份出現的頻率最多，高達56.7%。這說明了“自己”在口語和書面語中的確扮演了不同的功能角色。但是，相關文獻對反身詞的討論多半在於它作爲一個賓語時的現象(Sells et al. [18])。根據實際語料檢測結果，漢語反身詞“自己”以賓語的身份出現的次數，卻遠遠的少於主語和名詞組主語的功能。這是漢語反身詞“自己”的一項特色。

(三) 口語和書面語中單純反身詞和複合反身詞的比較

根據檢測結果發現，在口語中的62個“自己”，有31個單純反身詞及31個複合反身詞。在書面語中有104個單純反身詞和15個複合反身詞。我們將這個統計結果做比較，其出現頻率比較如下：

表五、單純反身詞和複合反身詞在口語和書面語中出現的頻率

	口語	書面語
單純反身詞	50%	87.3%
複合反身詞	50%	12.6%

這個統計表值得注意的地方是，在書面語方面，單純反身詞出現的頻率遠比複合反身詞來得高(87.3% vs. 12.6%)。在高達87.3%的統計數字顯

示，單純反身詞“自己”是比較偏向書面語文字，這表示，書面語的語境使得單純反身詞“自己”與前行語的指涉關係較爲明顯。而複合反身詞在書面語被使用的機率就非常低。但是在口語方面，單純反身詞和複合反身詞的使用頻率則是相當接近(50% vs. 50%)。這表示在口語的使用上，複合反身詞的使用有其特殊的功能。此現象透露出一個非常重要的訊息，那就是，漢語反身詞在口語和書面語的使用上，是採取不同的原則，扮演不同的功能角色。必須分別加以探討。

口語和書面語最大的不同在於，口語是瞬間即逝。它無法像書面語一樣有文字意像的儲存，也因此說話者和書寫者勢必有不同的方式去表達及追溯一段言談中較早出現的主題。[9] 以書面語而言，Fox [7]及Li [15]分別探討英文及中文照應詞在敘述文上呈現的篇章結構(discourse structure)。她們均發現名詞、代名詞、及空號代詞在篇章中的確有一定的階層性結構(hierarchical structure)，以達到結合相同主題的子句群(clauses)形成主題串(topic chain)，或銜接幾個主題串形成段落(paragraph)的功能(詳見Li [15])。同時她們亦分別根據口語和書面語語料檢視發現，口語比書面語使用更多的代名詞。也就是說，在口語中需要較多的代名詞標示開始另一主題串的功用(見註七)。

這些現象說明了表五所顯示的意義。在說話時因受限於說話的即時性 (spontaneity) 及記憶長度，說話者和聽話者都沒有足夠的時間回想和思考。心理語言學家實驗認爲，人類的說話是有一定的記憶長度。雖然確切的數字到目前尚無定論，但根據口語即時的特性，我們得以解釋口語中之所以需要較多的代名詞不斷標示其主題，甚至有贅述 (redundancy)的現象：因爲說話者要提供聽話者足夠的資訊，必須給予相當明確的指

[9] Chafe [2] 提出因語言處理在說話與寫字的速度不同,以及是否說話者與聽話者直接溝通與否,皆造成口語和書面語有不同的形式。

涉對象，以避免語意模糊 (ambiguity)。相較於口語的即時性，書面語有文字的儲存，能允許有較長的主題串，即使主題是在超過三到四個空號代詞引介的子句之外，讀者依然可以清楚地追溯 (refer back)，不會有混淆指涉對象的情況發生。表五的結果證明了以上論述。我們在口語語料發現複合反身詞有大量增加的現象。當代名詞+反身詞的結構出現時，說話者使用代名詞表達清楚的指涉關係，以協助聽話者掌握當下談話主題。這就是漢語反身詞的語用現象。

茲舉例說明如下： (A, B, and C 代表不同的說話者。)

(29)

A： 1 他開始爭取他音樂的天空，
2 也就是他覺得，
3 我才不要你的音樂概念呢！
4 我要的是在我的專輯裡頭呈現了我的音樂概念。

B，C： 5 對。

A： 6 比如說有名的例子。
7 我們看到 George Michael ，
8 他在 Wham 時代的音樂就和後來他自己作的那個 Father Figure 不太一樣。

B，C： 9 對，
10 不一樣。

上述對話中，第七句用名詞引出新的主題對象，英國流行歌星 George Michael 。第八句的主語用代名詞“他”指稱主題，隨後又用複合反身詞再次指涉主題。我們發現在書面語中可以出現空號代詞的位置，在口語中被代名詞取代，以縮短主題串的長度，表達明確的指涉關係。

五、結論與建議

Zribi-Hertz [23] 根據大量的書面語語料仔細檢視了英文反身詞的分佈與指涉情形。文中發現，即使是第三人稱單數反身詞(如himself/ herself)也並非全然遵守Chomsky的約束原則A，而是有超越一個句子範圍的指涉

情形。Zribi-Hertz 認爲，除了結構上的限制，對英語反身詞的討論必須加上言談的原則，反身詞理論的研究才能完整。在漢語研究部份，Chen [4]和Xu [22]指出，漢語反身詞的研究應建立在言談-語用原則(discourse-pragmatic principle)加以討論。本文的量化研究更具體反應反身詞的實際使用情形，以提供漢語反身詞一個重新思考的方向。

本文利用眞實的書面語和口語語料來檢驗漢語反身詞“自己”的分佈狀況以及其和前行語的指涉關係。由以上的討論可以得知，首先，無論是口語和書面語語料均顯示，反身詞的指涉範圍爲：

言談約束關係 > 近距約束關係 > 長距約束關係

值得注意的是，過去大部份反身詞研究皆鍾情的長距約束關係在實際語用的情形，其頻率不到20%。這表示過去反身詞的研究尚未能一窺全貌，在言談約束上仍有很大的研究空間。本文發現言談約束關係才是反映更多實際反身詞使用的現象。其次，在書面語上單純反身詞的使用佔了絕對的優勢(87.3%)。但是，在口語中單純反身詞和複合反身詞的使用頻率呈現了均勢的狀態(50% vs. 50%)。這個訊息透露出，漢語反身詞在口語和書面語中的使用情形是不同的。在口語上，複合反身詞的使用是有其絕對的意義，必須加以分開探討。過去反身詞的研究幾乎以單純反身詞在書面語中的特性爲主。本文發現，由於書面語與口語在形式與溝通需要的不同(Chen [4], Fox [7], Li [15])，反身詞“自己”也有不同的使用情形，反映不同的功能角色。

本文的研究爲未來漢語反身詞的研究提供了新的方向。以往對漢語反身詞的相關討論皆專注於單純反身詞的討論，而忽略了複合反身詞的重要性。對於漢語反身詞的探討不應僅局限在句內的長距指涉關係，而更須要探討漢語反身詞在實際使用上的分佈情形與指涉關係，對於漢語反身詞的研究才能完整。此外，除了反身詞在書面語中句法關係的討

論，其言談功能上所扮演的角色及實際使用上異於書面語的現象，都值得更進一步的探討。

參考書目

[1] Battistella, Edwin and Yonghui Xu. 1990. "Remarks on the Reflexives in Chinese." Linguistics 28: 205-240.

[2] Chafe, W. L. 1982. "Integration and Involvement in Speaking, Writing, and Oral Literary. Ed. Deborah, Tannen. Spoken and Written Language: Exploring Orality and Literary. Norwood, New Jersey: Ablex.

[3] Chen, Hsin-hsi. 1992. "The Transfer of Anaphors in Translation." Literary and Linguistic Computing 7: 231-238.

[4] Chen, Ping. 1992. "The Reflexive *ziji* in Chinese: Functional vs. Formalist Approach." Research on Chinese Linguistics in Hong Kong. Ed. T. Lee. Hong Kong: The Linguistic Society of Hong Kong.

[5] Cole, Peter et al. 1990. "Principles and Parameters of Long Distance Reflexives." Linguistic Inquiry 21: 1-21.

[6] Cole. Peter and Li-May Sung. 1994. "Head Movement and Long-Distance Reflexives." Linguistic Inquiry 25: 355-406.

[7] Fox, Barbara A. Discourse Structure and Anaphora: Written and Conversational English. Cambridge: Cambridge Univ. Press.

[8] Her, One-Soon. 1991. "Topic as a Grammatical Function in Chinese." Lingua 84: 1-23.

[9] Huang, C.-T. James, and Tang C.C. Jane. 1991. "The Local Nature of the Long-Distance Reflexives in Chinese." Long Distance Anaphor. Ed. Jan Koster. Cambridge: Cambridge.

[10] Huang, Shuanfan. 1992. "Getting to Know Referring Expression : Anaphora and Accessibility in Mandarin Chinese." Proceedings of Rocling V 25-52.

[11] Huang, Yan. 1991. "A Neo-Gricean Pragmatic Theory of Anaphora." Journal of Linguistics 27: 301-335.

[12] ---. 1994. The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese. Cambridge: Cambridge.

[13] ---. 1994. "Review of Long Distance Anaphora." Journal of Linguistics 22: 629-645.

[14] Kao, Rong-rong. 1993. Grammatical Relations and Anaphoric Structures in Chinese. Taipei: Crane.

[15] Li, Cherry Ing. 1985. Participant Anaphora in Mandarin Chinese. PH.d. Diss., Univ. of Florida.

[16] Li, Mei-Du. 1985. Reduction and Anaphoric Relation in Chinese. PH.d. Diss., Univ. of California San Diego.

[17] Li, Yafei. 1993. "What Makes Long Distance Reflexives Possible?" Journal of East Asian Linguistics 2: 135-166.

[18] Sells, Peter et al. 1986. "Reflexivization Variation: Relation Between Syntax, Semantics and Lexical Structure." Eds. M. Iida et al. Studies in Grammatical Theory and Discourse Structure: Interactions of Morphology, Syntax, and Discourse. CSLI Stanford Univ.

[19] Tang, C.C. Jane. 1985. A Study of Reflexives in Chinese, master's thesis, National Taiwan Normal University.

[20] ---. 1989. "Chinese Reflexives." Natural Language and Linguistic Theory 7: 93-121.

[21] 湯廷池. 1992. "漢語句法與詞法的照應詞." 清華學報 22: 301-3 50.

[22] Xu, Liejong. 1993. "The Long-Distance Binding of Ziji." Journal of Chinese Linguistics 21:123-141.

[23] Zrib-Herbi, Ann. 1989. "Anaphor Binding and Narrative Point of View: English Reflexive Pronouns in Sentence and Discourse. Language 65: 695-727.

適合大量中文文件全文檢索的索引及資料壓縮技術

簡立峰[1]， 古鴻炎[2]

1.中央研究院資訊科學研究所
E-mail: lfchien@iis.sinica.edu.tw
2. 國立台灣工業技術學院電機工程學系
E-mail: root@guhy.ee.ntit.edu.tw

摘要

本文主要是提出適合大量中文文件全文檢索的索引及資料壓縮技術。在全文索引部份其特色是利用先前發展特殊設計的特徵檔(Signature File)索引配合自動分析技術，文件所需之索引空間因而可隨不同需求而調整，一般而言索引只佔文件大小的 15% ~ 30% 左右，如此不僅索引建置速度極快同時也可獲致極佳檢索效率。另外在文件內容方面，我們以LZ77方法為基礎再經適當改進得到的良好的壓縮率(47.3% ~ 53.0%)同時也具備文件密碼化甚至加快檢索速度的功能。目前這些技術皆已經成功應用在一高效率的中文文件檢索系統－"尋易"(CSmart)。

壹、緒論

隨著網路的佈建及電子文件的成長，文件全文檢索(Full-text Searching)的需求日殷[1]。在英文文件搜尋方面已有許多檢索系統如 WAIS及BRS/Search成功的發展出來。另外考慮中英二者語言的差距，有關中文文件全文檢索技術的研發也日受重視並有不錯的進展[2,3,4,5]。然而過去的研究較少討論大量中文文件全文檢索的索引及資料壓縮技術，由於索引及資料壓縮對於儲存空間節省、光碟製作、文件密碼化、資料傳輸等關係密切，因此本文即以適合大量中文文件全文檢索的索引及資料壓縮技術為題，提出一套已經成功應用在"尋易"(CSmart)中文文件檢索系統的高效率方法，並說明資料壓縮技術對全文檢索的重要性。

以一般全文檢索系統而言，文件及文件索引是系統中最佔儲存空間的兩類資訊。通常索引空間需求大概是文件量的50%~300%左右，由於電腦儲存媒體容量大增且價格日益低廉，研究者對空間節省較不以為意，因此全文檢索中有關文件壓縮與索引壓縮的研究較少。但是就發展實用系統的角度來看，空間的節省仍然有其必要性。因為空間的節省也就代表經費的降低。以製作一個含300MB的文件及700MB索引的全文光碟資料庫為例，所需即是一1GB的光碟，但如果文件及索引可以適當壓縮，則也許只須使用500MB光碟即可，對資料量大且發行數目較多的全文光碟而言空間的節省仍有其價值。此外透過資料壓縮技術將所需空間減低同時也具備資料密碼化的效果，有助於降低非法拷貝的機會，另外壓縮後的資料傳輸較快也有助於加快檢索速度。

為了降低索引的空間需求並維持極佳檢索效率，我們曾發展出一項特殊設計的特徵檔(Signature File)索引方法[5~8]，其所需之索引空間可隨不同需求而調整，一般而言索引只佔文件大小的 15% ~ 30% 左右。此外最近我們也發展出自動測試技術，利用這項技術，系統可以自動產生測試查詢以檢測特徵檔的效率，自動決定索引大小及索引配置。由於這項技術使我們瞭解到系統除了提供索引壓縮功能，更重要是必須同時能分析壓縮後的檢索效率，而傳統檢索系統多缺乏類似功能。

另一方面，為了降低文件本身的空間需求並且將文件密碼化，我們以LZ77方法[9]為基礎再經適當改進得到的良好的壓縮率(47.3% ~ 53.0%)同時也具備極佳的文件密碼化甚至加快檢索速度的功能。而且根據我們實驗發現文件壓縮後可以減少磁碟存取次數可以進一步加快檢索速度。以上述技術為基礎我們發展出尋易(CSmart)中文文件檢索系統。這個系統的可提供多元化檢索功能且查詢速度極快，一般查詢皆可即時(Real-time)反應。此外，透過與 WWW (World-Wide Web)及MOZAIC的結合，可非常容易製作各種多媒體文件資料庫並提供網路查詢。利用此系統我們並開始製作新聞資料庫、電子詞典等在臺灣學術網路使用。

尋易系統的設計基本上是成功的整合我們發展的許多技術，部份技術我們已發表一些文章，包括整個尋易系統設計[5]，近似自然語言技術[6]，特徵檔技術[7]，以及語音檢索[8]，本文主要是著重資料壓縮技術，這是藉助本文第二位作者過去技術的基礎[11]，也是尋易系統最近的技術成果。以下各節，第貳節是有關尋易系統架構簡介，第參節是索引技術之介紹，第肆節說明文件壓縮技術，第伍節是結論。

貳、尋易系統架構

尋易檢索系統是針對中文文件特性所設計的的檢索程式，其角色與Internet上著名的WAIS系統接近，使用者可以透過WWW server直接存取儲存在尋易系統內的中文文件。尋易系統與WWW的結合方式可以如圖1所示。另外尋易檢索系統的整個架構主要是以特徵檔索引為基礎，其主要結構如圖 2 所示。

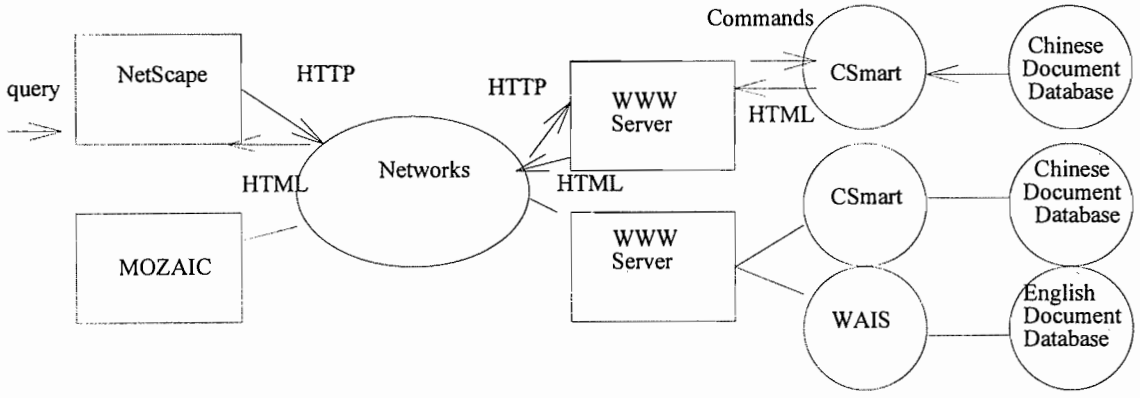


圖 1. 尋易檢索系統與WWW的整合應用方式

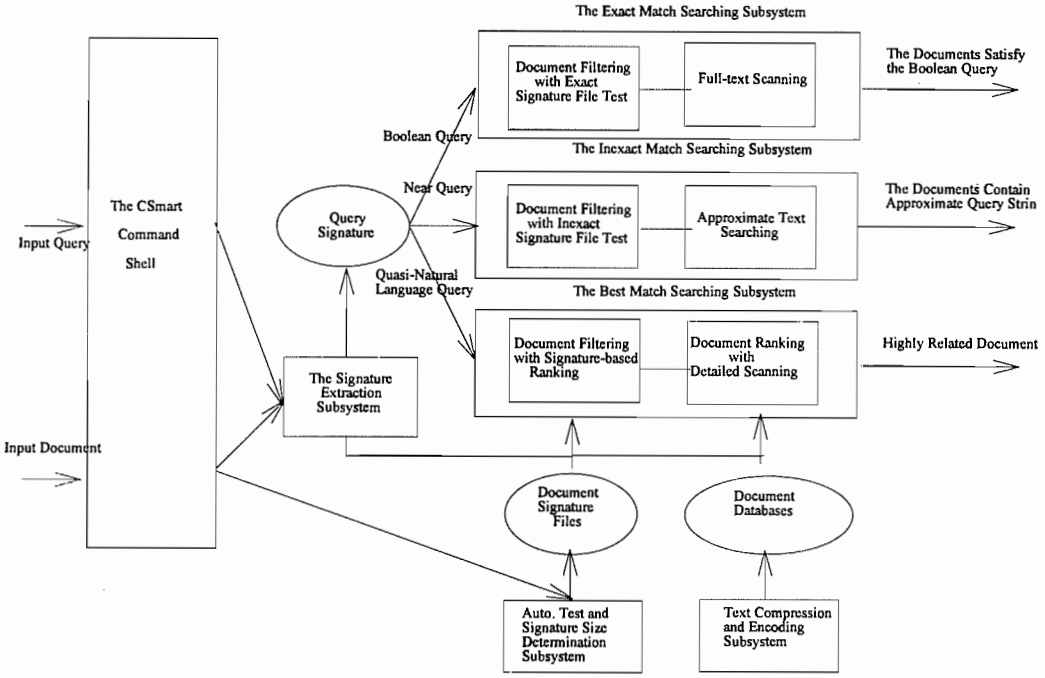


圖 2 尋易系統的系統架構

其中包括輸出入介面(Command Shell)、特徵檔產生(Signature Extraction)、精確比對搜尋(Exact Match Searching)、近似比對搜尋(Inexact Match Searching)、最佳比對搜尋(Best Match Searching)、自動測試及索引大小決定(Auto Test and Signature Size Determination) 以及文件壓縮解碼 (Text Compression and Encoding)等七個主要程序。

所有文件儲存在尋易系統內都須經過壓縮處理，當使用者欲瀏覽文件內文時才會解碼顯示。整個壓縮及解碼處理由文件壓縮解碼程序負責。任何文件欲加入資料庫必須先經特徵產生程序產生該文件的特徵並加入特徵檔(Signature File)。如果系統管理者希望所產生的特徵索引不論檢索速度與儲存空間大小都很有效率，也可執行自動測試及索引大小決定程序。尋易系統會參考現有資料庫內容、儲存空間需求、溢檢率(False Drops)的情形重新決定索引大小及特徵配置方式。

另外，使用者在查詢時可以選擇以一般布林查詢、近似字串查詢、近似自然語言查詢的方式表達查詢主題。不論是那一種查詢都必需先產生查詢特徵(Query Signature)。若是布林或近似查詢即交由精確或近似比對搜尋程式處理，這包括第一階段先將查詢特徵和所有文件特徵比對，未滿足該查詢特徵的文件將被濾掉。未被濾掉的文件內容在第二階段將會被讀出及與查詢仔細比對，眞正滿足該查詢文件才會檢索出。若使用者是以近似自然語言方式查詢文件，則將交由最佳比對搜尋程序處理。在第一階段該程序會將查詢特徵與所有文件特徵一一比對估算出其查詢與文件之初步相似度(Relevance Value)，相似度足夠高的文件才會在第二階段繼續處理。第二階段基本上是將這些文件內文讀出，仔細比對查詢句子中一些關鍵語出現在文件中的頻率及位置，以進一步判斷其相似度，最相似的一些文件才會檢索出。

根據前述說明，特徵檔的產生基本上是整個系統的核心且同時提供布林查詢、近似查詢及近似自然語言查詢之用。另外精確比對搜尋、近似比對搜尋和最佳比對搜尋皆是一種兩段式搜尋概念，其第一階段皆以快速的方式濾去絕大多數不相關的文件，第二階段才仔細比對餘下的文件，這種搜尋技術非常適合處理相當大量的文件。根據我們實驗發現由於我們所發展之特徵檔具極佳之過濾(Filtering)效果，在需快速反應時，往往只要施行第一階段搜尋(即比對特徵檔)即可得到很好的結果。以下第參節將進一步說明特徵檔的產生。

參、特徵檔的產生與自動調整

英文特徵檔技術

從前節介紹可知尋易系統以特徵檔為主要索引。特徵檔的主要是針對布林查詢、近似字串查詢及近似自然語言查詢而設計。這種特徵檔一方面必須將全文訊息儘量完整且清楚的記錄起來，另一方面又必須節省空間且方便存取(Access)。特徵檔的效率是以其在處理布林查詢及近似自然語言查詢時能過濾掉不相關文件的程度來衡量。由於中、西文文件特性大不相同，西文通常是以詞為單位的方式來產生文件特徵[1]。基本上西文系統會以 hash 方式令每一個詞有一固定長的 0/1 字串(其中有若干處設定為 1)為其詞特徵(Word Signature)，而文件特徵即是文件所有詞簽名之重疊處理(Superimpose)。如圖 3 所示。

signatures of words:

document	00100100
information	00100001
retrieval	10000100

signature of a document just contains the above three words:

10100101

圖 3. 西文系統以詞特徵及重疊處理(Superimpose)產生文件特徵舉例

因為英文 word 有豐富的詞尾變化，產生特徵時這些 word 都必須先進行原形化處理。然而即始經過原形化後詞的數目仍很多，考慮索引空間大小及每一個詞都必須有一個唯一的特徵值，勢必造成很多詞的特徵很接近，如只有一個 bit 值不同。因此由重疊後的文件特徵並不易還原出原先眞正組成的詞，所以這類特徵比較適合充當精確比對時判斷有無包含查詢字串的濾波器(Filter)，而不適合充當特徵向量(Feature Vector)，也即不適合應用在相關性估計(Relevance Estimation)。

以統計特性為基礎之特徵檔產生方式

為了發展適合中文特性的特徵檔並且同時兼具精確比對時的濾波器及最佳比對之特徵向量雙重效果，我們提出以統計特性為基礎之特徵檔產生方式[6,7]。我們以單字和雙字做為特徵檔產生之基礎，因為中文有分詞上的困難，我們捨去以詞為單位的處理方式，取而代之是以單字和相連雙字為單位。在台灣地區中文常用字在5,000 字左右，也就是對每一文件只要 5,000 bits 令每一bit分別代表一中文字之存在與否，如此一來即可清楚記錄這些字是否出現在文件內。這種方式的問題在於文件內許多字序及相鄰訊息(Positional Information)，如"日本"或"本日"無法藉此表達，因此如果能夠把雙字的特性同時考慮進來則可大為加強位置訊息，因此我們首先將特徵分成兩段，第一段代表常用單字出現的訊息，第二段代表雙字及罕用字出現的訊息，這每一段的長度是可以調整的。雖然常用字有 5,000 字，但許多字很少在一起聯用，因此我們提供一套統計分群方法。利用一些訓練語料(Corpus)，將統計特性差距大的一些字合為一群，在第一段特徵裡令其共用一個 bit 以記錄其出現之訊息。這個分群的群數自然和第一段特徵長度有關，如果第一段長度有 1,000 bits，則將 5,000 字分群 1,000 群。再者，因為中文罕用字及雙字數目太多，我們並不採用統計分群技術，而是利用 hash 方式令某些罕用字或雙字共用一個 bit，如此一來，文件特徵之產生即如下圖 4所示。

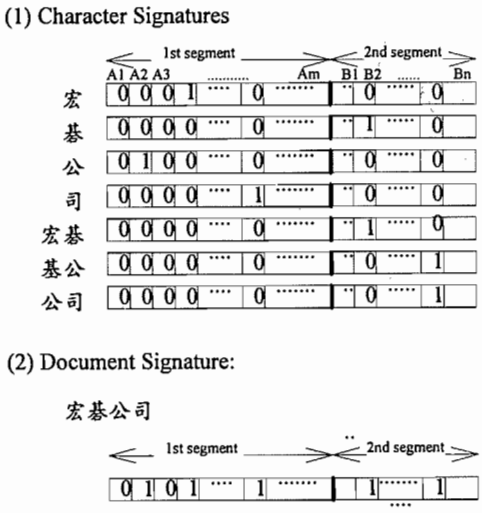


圖 4 尋易系統文件特徵(Signature)產生舉例

若文件內容是如(2)所示之"宏碁公司"之字串，且"宏"、"公"、"司"等三個常用字及"碁"這罕用字及"宏碁"、"碁公"、"公司"這三個雙字之特徵如(1)所示，則其最後文件特徵將如(2)之結果。根據我們的研究這樣的特徵結構比英文特徵檔基本上記錄很豐富的訊息不佔太多空間又易存取。此外特徵長度及每一個字或雙字的特徵可以隨文件特性加以調整。其進一步的說明可參閱[6,7]。

實驗結果

尋易系統目前已初步完成且基本上做了相當多的測試並陸續發表不少成果〔5~8〕，包括在精確比對方面與傳統逐字反檔(Character Inversion File)方法比較，其不論索引空間、索引時間、及查詢速度皆遠優於該方法。另外在溢檢率降低方面也有不少改進及分析。由於這些結果可以在前述文獻上參閱得到，本節只略列一些實驗結果供做參考，包括在新聞資料庫應用環境之索引時間、索引空間、查詢速度等。

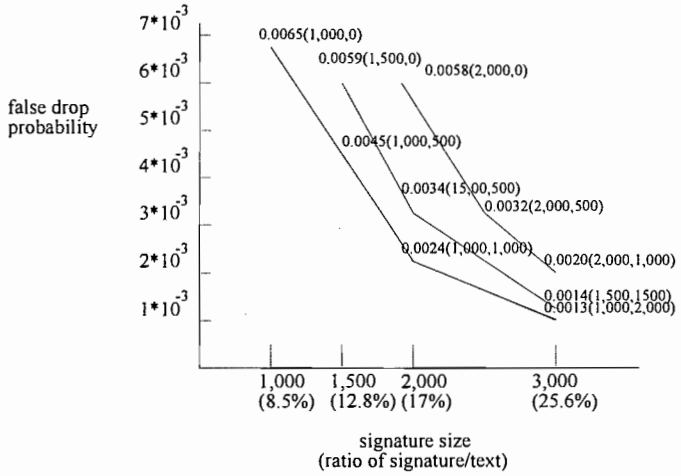


圖 5. 溢檢率與索引大小的關係，圖中括弧內第一個數字代表第一段特徵長度，而第二個數字代表第二段特徵長度。

首先在溢檢率與索引大小關係方面，由圖5所示可發現利用前述特徵檔方法可以得到不錯的溢檢率，特別是當第二段特徵長度較長時。舉例，當第一、二段特徵長皆為1,000 bits 時索引大小只佔文件的17%，而溢檢率只有0.0024，也即對任一查詢平均每1000件無關文件只有2.4件會被誤認，但這些誤認文件可利用進一步比對文件內文而濾掉。另外根據表1，我們可以發現當索引文件比維持25%時皆可獲致極佳的結果，不論工作站或PC，索引時間，平均字串搜尋時間。表1索引時間並不包括以下第肆節將介紹的壓縮時間。

Host	Main Memory	Document Size	No. of Documents	Index Space	Indexing Time	Average Search Time
PC486	8M	1GB	1.25M	250MB	55min	about 10sec
PC486	8M	100MB	0.12M	25MB	4min	1sec below
SPARC10	32M	1GB	1.25M	250MB	25min	3~5sec
SPARC10	32M	100MB	0.12M	25MB	2min	1sec below

表 1 實驗結果舉例

特徵檔自動分析方法

綜合上述實驗，我們發現利用前述特殊設計的特徵檔索引技術已經將索引空間降到最低，為了存取效率方便，如非必要索引不應再進一步壓縮。取而代之，因為所發展的特徵檔索引為可調式，我們決定發展自動分析技術。允許系統管理者自行決定所需索引文件比，尋易系統利用先前類似圖5的數據決定若干可能的索引大小，之後參考使用者已鍵入的查詢及資料庫內容，重新對每一種索引組合配置特徵，另外隨機產生一些測試查詢。進而，系統會自我分析，決定最適合系統管理者期望的索引文件比以及最低的溢檢率。根據我們初步實驗發現，由此方式產生的溢檢率可以降得更低。而且對系統管理者而言，管理上非常方便。利用此方式，我們創造了更有智慧的可調式索引，也使得索引壓縮可以配合檢索成效一併考慮，這是傳統索引壓縮不易做到的。

肆.中文文件壓縮方法

對大資料量(如 100mega bytes 以上)的中文文句(Chinese text)作全文搜尋(full-text searching)的應用裡，資料只能儲存在速度較慢的次要儲存設備(如硬碟、光碟等)裡，為了加快從硬(光)碟到主記憶體的傳輸速度、及減低記憶空間的需求，我們可先把原始資料作壓縮處理，然後才將壓縮後的資料存入硬(光)碟裡。不過，在全文搜尋及類似的資訊檢索應用裡，只考慮降低壓縮率(壓縮後資料長度除以原始資料長度)是不夠的，因為從硬(光)碟傳回的資料需先經過解壓縮處理，還原成原始資料後，才能去作搜尋，如果解壓縮的速度太慢，而資料量又很大，就會使整個全文搜尋系統的反應變慢到無法忍受。

因此，在這裡我們提出一種適用於大資料量中文全文搜尋應用的資料壓縮方法，事實上它是以 LZ77 方法[9,10]為基礎再經過修正而得到的，我們提出的主要修正是：

(1)擴大字符集(alphabet)

在 256 個 ASCII 字元之外，再將五大(Big-5)中文碼裡的 5401 個常用中文字，及我們定義的 32 個特殊字元(表示兩字串間成功比對的長度)含蓋進來，根據我們過去的經驗[11]，把表示一個中文字的兩個 bytes 當作一個字符集字元來處理，要比當成兩個分開的 ASCII 字元來處理，會得到更低 10% 以上的壓縮率。

(2)調適分群(adaptive grouping)

在把字符集裡的字元分成若干群後，隨著壓縮處理之進行，讓常出現的字元轉移到一個以較短的碼來編碼的字元群去，而將不常出現的字元排擠到一個以較長的碼來編碼的字元群去。這裡提出的調適分群想法，和以前被提出的固定式分群想法[12,13]是很不一樣的。

所以要將字符集裡的字元加以分群的原因是，當我們對前述之大字符集編一個變長碼(variable-length code)，再配合使用移前(move-to-front)[10,14]調適策略時，會發現字符集太大使得移前調適的處理速度變得更慢，這樣就無法達成快速解壓縮的目標，而經由分群再配合我們提出的一種調適方法，就可得到很高的解壓縮速度，詳細情形以下將進一步說明。

資料壓縮前段處理:字串比對與字串編碼

由於我們提出的修正作法是以 LZ77 壓縮方法[9]為基礎的，所以先回顧原始的 LZ77 壓縮方法，它的處理程序可配合圖6來作說明，圖6裡 P 指標所指示的前看區(lookahead buffer)用來儲存 32 個等待作壓縮的輸入字元，而P指標之前的已編碼區則是用來儲存最近 4684 個已做過壓縮處理的輸入字元，原始的 LZ77 壓縮方法就是，到已編碼區中去尋找一個可以和前看區字元串比對成功的最長字元串，然後將此字元串的位置(和指標 P 間的差距) d，長度 e，及前看區中未能繼續比對成功的字元 u 等三項訊息，以一個三重組(triple) <d, e, u> 記載下來(例如當圖6 裡Q指標所指的是比對成功的最長字元串時，就會得到 <3, 2, c>)，接著，把指標 P 往右移 e 個位置(相當於滑動窗裡的字元往左移動)，再反覆做前述的動作，如此就可獲得一序列的三重組 $<d_1, e_1, u_1>$, $<d_2, e_2, u_2>$, ...。至於對應的解壓縮器，它不必去做字串比對的動作，只需依據接收到的三重組來把對應的原始資料輕鬆地恢復回來，所以解壓縮的速度會比壓縮的快很多。

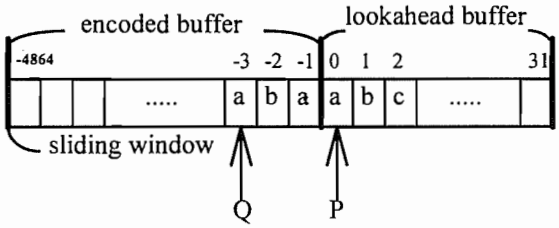


圖6. LZ77壓縮處理之滑動窗

前述的三重組序列，其實也可看成是一個二重組 $<d_i, e_i>$ 和單重組 $<u_i>$ 交替出現的序列，因此，後來 Bell 等人提出一種變形作法稱為 LZSS[10]，將二重組與單重組需交替出現的限制加以取消，但這需使用一個額外的 bit 來告知解壓縮器，接著出現的是一個二重組還是單重組，由它們的實驗顯示，取消限制的確可獲得更低的壓縮率。

研究了 LZ77 及 LZSS 的方法後，我們遂提出如下的修正作法，將 LZSS 裡需額外用一個 bit 去分辨二重組與單重組的作法加以改變，而將二重組的 d 組件與 e 組件的次序對調，再將二重組的 e 組件和單重組的 u 組件看成是來自於同一個字符集的不同字元(如 ASCII 字符集包含了控制字元與英文字母)，如此，就可將原先的特定用途 bit 融入到表示字元的 bits 去，而解壓縮器也可依據目前解得的字元去判斷，接下來要解出字元還是二重組裡的 d 組件。除了將二重組之 e 組件看成是字符集裡的字元，我們也把 Big-5 碼裡的常用中文字放到字符集去，而不常用的中文字則仍然看作是 2 個連接的 ASCII 字元，如此，我們的字符集裡便有6827個字元，使用這樣一個字符集，輸入的資料 bytes 很明顯地就不是直接放入圖6 裡的前看區，而需先經過構字分析(lexical analysis)，把二個 bytes 表示的常用中文字和一個 byte 表示的 ASCII 字元分別切割出來，並轉成字符集裡該字元的代號(2byte整數)，然後才將字元代號放入圖6 的前看區裡。

關於圖6 裡，要從已編碼區中找出一段能和前看區字串比對成功的最長字元串的問題，我們使用如下的解決辦法，當一個字元剛從前看區進入已編碼區時，就將此字元及其後的 31 個字元看成一個 key(可以跨越到前看區去)，然後依據此 key 的前二字元去作赫序(hashing)[15]，以決定它要插入(insert) 到 512棵二元搜尋樹中的那一顆，而當一個字元要從已編碼區的左端離去時，就將此字元及其後 31 個字元所構成的 key，從它所在的搜尋樹中加以刪除(delete)，這樣先將已編碼區中的字元依據赫序值建造成許多棵搜尋樹，而前看區裡 32 個字元構成的 key 也先經過赫序才到對應的搜尋樹去找具有最長共同子字串的 key，就可使字串搜尋的速度提高很多。前面提到的赫序，會使用一個key的最前面兩個字元去計算，這意謂搜尋後若發現最長共同子字串長度為 0，並不表示前看區的第一個字元就絕對未出現於已編碼區中，因此，最長共同子字串長度為 0 或為 1 時，我們都當作 0 值來處理，也就是要送出一個單重組。

資料壓縮後段處理:字元編碼與調適分群

每次當前段處理輸出一個二重組 <e, d> 或單重組 <u> 後，後段處理就接著拿去做字元編碼的動作，然後決定要不要將該字元換群，如果剛剛是對二重組的 e 組件編碼，則緊接著還需對 d 組件編碼。前述的字元編碼其實是分成兩部份去做的，即群編碼部份和元素編碼部份，其中群編碼是要產生一個碼用以表示 e 或 u 所在之字元群，我們使用變長碼(variable-length code)來表示不同的群，而元素編碼是要產生一個碼用以表示 e 或 u 目前被排在所屬字元群的第幾個元素上，我們使用定長碼(fixed-length code)來區分同一群內的各個元素。

在本研究裡，我們把字符集的 6827 個字元分成如下之 8 群：

A群:元素個數固定為 1,放置最常出現的字元，群碼為 $(010)_2$，不需元素碼；
B群:元素個數固定為 2,聚集次常出現的字元,群碼為 $(100)_2$,元素碼使用1個 bit；
C群:元素個數固定為 8,聚集次次常出現字元,群碼為 $(1110)_2$,元素碼使用3個 bits；
D群:元素個數固定為 32,群碼為 $(101)_2$,元素碼使用5個 bits；
E群:元素個數固定為 128,群碼為 $(011)_2$,元素碼使用7個 bits；
F群:元素個數固定為 512,群碼為 $(00)_2$,元素碼使用9個 bits；
G群:元素個數固定為 2048,群碼為 $(110)_2$,元素碼使用11個 bits；
H群:元素個數固定為 4096,聚集最罕出現字元,群碼為 $(1111)_2$,元素碼用12個 bits；

各群的元素個數是在群數固定為 8 時，經由實驗(此時群碼為固定長度)對不同的組合去量測、比較壓縮率後才決定下來的，然後再依實驗裡各群的出現頻率以霍夫曼(Huffman)編碼方法去決定群碼。

壓縮器(或解壓縮器)一開始動作時，它並不知道那一些字元較常出現或較不常出現，因此它是隨意地將字符集字元分成 8 群，如果能夠隨壓縮處理之進行，把較常出現的字元改分配到較前面(即靠近A群這邊)的字元群去，則可使整體的壓縮率降低。為了達成前述目標，我們嘗試了幾種調適分群的作法，最後發現一種效果最好的，詳細說來是，為字符集的每一個字元 s 設一個出現頻率計數變數 N(s)，並為每一個字元群 X 設一個先進先出指標 P(X)，以將一個字元群當作一個貯列

(queue)來控制群內元素之進出，然後，當一個剛被編完碼的字元 s 目前是字元群 X 的一個元素時，就進行如下的調適動作：

```
delete(X, s);      /* 將字元 s 從字元群 X 中刪除 */
N(s) <= N(s) + 1; /* 字元 s 之出現頻率值加 1 */
Y <= X - 1;        /* 將變數 Y 設定為變數 X 所代表之字元群的前一群 */
while( X ≠ 'A' and N(s) > N( front(Y) ) ) {
    w <= dequeue(Y);      /* 從貯列 Y 的最前端取出一個字元 */
    enqueue(X, w);    /* 將字元 w 插入到貯列 X 的最後端 */
    X <= Y;
    Y <= X - 1;    /* 將變數 Y 設定為變數 X 所代表之字元群的前一群 */
}
enqueue(X, s);    /* 將字元 s 插入到貯列 X 的最後端 */
```

在前述的調適程序(procedure)裡，函數 delete(X, s) 的用途是，把字元 s 從字元群 X 之貯列中強迫去除，不管 s 是否在貯列的最前端；函數 front(Y) 用以查詢貯列 Y 最前面的字元；至於函數 dequeue(Y) 與 enqueue(X, w) 的功用，就如程序裡的註解所說明的。其實，前述程序的想法是取自電腦作業系統裡的一種 CPU 排程(scheduling)的方法(即 multi-level round robin)，再配合我們的問題而修正得到的。

至此還未說明的是二重組裡 d 組件的編碼方法，d 的數值範圍是 0 到 4863 (表示差距 1 到 4864=76*64)，要對 d 編碼，我們先求 d 除以 64 之商數 dq 與餘數 dr，然後以變長碼來對 dq 編碼，而以定長碼來對 dr 編碼。對於 dq 之編碼，我們也是先作分群，然後以群碼結合元素編碼來作為 dq 之編碼，在 dq 的數值範圍(0至75)裡，我們把0至3放在第一群(群碼為00)，4到11放在第二群(群碼為110)，12到27放在第三群(群碼為01)，28到43放在第四群(群碼為111)，44到75放在第五群(群碼為10)，群碼是依據實驗得到之各群出現頻率的數值來訂定的。此外，由實驗結果來看，使用固定分群或調適分群去對 dq 編碼，所得到的壓縮率差異只在 0.05% 的數量級，因此我們就選取固定分群之作法。

實驗結果

我們已將所提出的方法寫成可實際使用之軟體(以C語言寫作)，在經過典型中文文句資料檔的測試後，發現所提方法的壓縮率(47.3% ~ 53.0%)一般說來比著名壓縮軟體 ARJ 或 PKZIP 的壓縮率(52.7% ~ 60.5%)還低 5.3% 以上，至於時間花費，以 486-33 個人電腦來說，我們的軟體的平均壓縮與解壓縮速度分別是 28.2K Bytes/sec 與 112.6K Bytes/sec，解壓縮的速度很明顯地比壓縮的快很多，雖然這樣的速度仍比 ARJ 的 178.6K Bytes/sec 慢，但是， ARJ 是不是以組合語言寫的？有無經過最佳化佈置？我們並不知道。

目前這項壓縮技術已應用在尋易系統中。對於尋易系統也產生若干影響，這也是我們將文件壓縮技術應用在中文全文檢索後所獲致的經驗。首先文件壓縮增加了索引建置時間，這幾乎是文件壓縮技術唯一的缺點，幸好多數應用並不要求太快的索引時間。其二，因為解壓縮速度快，加上原先溢檢率已低，因此壓縮技術的應用

並不增加搜尋時間。其實根據我們實驗，如果所使用的硬碟存取不快，利用壓縮技術反而加快硬碟I/O，甚至可加快檢索速度。其三，文件儲存空間可節省一半，就實際應用而言是非常重要的結果。其四，壓縮後造成自然的密碼化，也是所有文件檢索實用系統所關切的。雖然文件既然可瀏覽當然可拷貝，但許多應用裡，我們發現提供資料的單位皆希望辛苦建立的資料不應該太容易就被拷貝。最後，我們希望說明就一檢索系統而言自行發展文件壓縮的技術是絕對有其必要性。因為，任何公共使用的壓縮程式，都會失去密碼化的價值。

伍. 結論

本文主要是提出適合大量中文文件全文檢索的索引及資料壓縮技術。在全文索引部份是利用先前發展特殊設計的特徵檔(Signature File)索引以及本文所提自動分析方法，實驗證實索引已壓縮至極低，並不須進一步壓縮。反倒是，如何提供管理者同時考慮檢索速度與壓縮比才是最重要，這點本文所提方法已具相當不錯的成果。另外在文件內容壓縮方面，我們以LZ77方法為基礎再經適當改進得到的良好的壓縮率(47.3% to 53.0%)，同時也發現壓縮技術有助於文件密碼化甚至加快檢索速度的功能。本文證實了大量中文全文檢索對壓縮技術的需求。目前這些技術皆已經成功應用在一高效率的中文文件檢索系統－"尋易"(CSmart)。

參考資料:

[1]. Salton, G., "Introduction to Modern Information Retrieval", NY, McGraw-Hill, 1983.

[2]. Tseng, S. S., Yang, C. C. and Hsieh, Ching-chun, "On the design of Chinese Textual Database", Computer Processing of Chinese and Oriental Languages, 4, 1989, 240-271.

[3]. Liang, Tyne, Lee, S. Y. and Yang, W. P. "On the Design of Effective Chinese Textual Retrieval Based On the Signature Method", Computer Processing of Chinese and Oriental Languages, Vol. 8, No. 1, June 1994, pp. 87-96.

[4]. Wu, Zimin and Tseng, Gwyneth, "Chinese Text Segmentation for Text Retrieval; Achievements and Problems". JASIS, 44(9), 1994, 532-542.

[5]. Chien, Lee-feng, Huang, Tung-I, Lee, Shi-Bai and Lee, H. C., "尋易 (CSmart) -- A High Performance Chinese Document Retrieval System", submitted to ICCPCOL'95.

[6]. Chien, Lee-feng, "Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts", to appear on ACM conf. on SIGIR'95.

[7]. Chien, Lee-Feng, "A Model-Based Signature File Approach for Full-text Retrieval of Chinese Document Databases", To appear on Computer Processing of Chinese and Oriental Languages, 1995.

[8]. Lin, Sung-Chien, Chien, Lee-feng and Lee, Lin-Shan, "Unconstrained Speech Retrieval for Chinese Document Databases with very large Vocabulary", The 4th European Conference on Speech Communication and Technology, EuroSpeech"95

[9] Ziv, J. and A. Lempel, "A Universal Algorithm for Sequential Data Compression", IEEE trans. Information Theory, Vol. 23, No. 3, pp. 337-343, May 1977.

[10] Bell, T. C., J. G. Cleary and I. H. Witten, "Text Compression", Prentice-Hall, Inc., 1990.

[11] Gu, H. Y., "Adaptive Chinese Text Compression Using Large Alphabet, Markov Modeling, and Arithmetic Coding", National Computer Symposium(Chia-yi), pp. 568-575, 1993.

[12] Chang, C. C and W. H. Tsai, "A Data Compression Scheme for Chinese-English Characters", Computer Processing of Chinese & Oriental Languages, Vol. 5, No. 2, pp. 154-182, March 1991.

[13] Chang, K. C. and S. H. Chen, "A New Locally Adaptive Data Compression Scheme Using Multilist Structure", The Computer Journal, Vol. 36, No. 6, pp. 570-578, 1993

[14] Bentley, J. L., *et al.*, "A Locally Adaptive Data Compression Scheme", Commun. ACM, pp. 320-330, 1986.

[15] Weiss, M. A., "Data Structure and Algorithm Analysis in C", Benjamin/ Cummings Publishing Company, Inc., 1993.

THE NEW GENERATION BEHAVIORTRAN : DESIGN PHILOSOPHY AND SYSTEM ARCHITECTURE

*Yu-Ling Una Hsu and *+Keh-Yih Su

*Behavior Design Corporation
No. 5, 2F, Industrial East Road IV
Science-Based Industrial Park
Hsinchu, Taiwan 30077, R.O.C.

+Department of Electrical Engineering
National Tsing-Hua University
Hsinchu, Taiwan 30043, R.O.C.

Email: *una@bdc.com.tw, *+kysu@bdc.com.tw

ABSTRACT

In many conventional machine translation systems, the translation outputs are usually strongly affected by the syntactic information of the source sentences and thus tend to produce literal translations that are not natural to the native speakers. In this paper, we introduce the design philosophy and system architecture of the new generation BehaviorTran, which will enable an MT system to operate with high modularity and to acquire its translation knowledge from a bilingual corpus with a two-way training method. In such a paradigm, the knowledge bases only provide static descriptions on the legal forms of the constructs, while ambiguity resolution and preference evaluation are governed by sets of statistical parameters. This makes it easier to adapt the system to specific user styles and maintain different parameter sets for different customers. Thus, it is expected to be a promising paradigm for producing satisfactory translations.

1. Introduction to the First Generation BehaviorTran

The BehaviorTran English-Chinese Machine Translation System (formerally ArchTran) is the first of its kind research launched in Taiwan, and is among the few commercialized English-Chinese systems in the world (Chen[1], Wu[2]).

The research on BehaviorTran began as a joint project between National Tsing Hua University, Taiwan, and Behavior Tech Computer Corporation (BTC) in May, 1985. And as the scope of research gradually extended, the BehaviorTran was later transferred to Behavior Design Corporation (BDC) in Feb., 1988 to continue the improvements on the system.

After four years of research, the first generation BehaviorTran was released in 1989 and serves as the kernel of a value-added network (VAN)-based translation service. The system is running on the SUN workstation and written in C language. Its primary domain is computer manuals and related documents.

The overall translation strategy adopted in the first generation BehaviorTran is conventional transfer-based approach. In this approach, the whole translation process can be logically divided into three phases, namely, analysis, transfer, and synthesis. In the first generation BehaviorTran, the English analysis component consists of a set of ATN-style augmented context-free phrase structure rules, which will parse the input English sentences into corresponding syntactic trees (Hsu [3]). And the transfer and synthesis operations are encoded in a set of pattern-action pairs, called tranfer rules, to carry out a sequence of tree to tree mappings to reflect the changes in substructures and linear order in the source-target language pair (Chang [4]).

2. Motivation for the Revision of BehaviorTran

Since the release of the first generation BehaviorTran, BDC translation center has established a customer base of several internationally-renowned companies. From several years of practical experience and the feedbacks from posteditors and customers, we find some drawbacks of the original system which urge us to make a thorough revision to the first generation BehaviorTran. The major drawbacks are stated as follows:

First, the degree of modularity in the first generation BehaviorTran is low. The application of transfer rules and the selection of target translations are closely related to the output of source language analysis grammar. Thus, once the analysis grammar is modified, a great number of transfer rules or lexical information should be modified accordingly. That greatly increases the load of system maintenance. Moreover, since different components are intricately related, when a new source or target language comes into play, most parts of the original components cannot be reused. This drawback becomes more and more sailent since BehaviorTran intends to extend itself to a multilingual translation system.

Second, as mentioned previously, the transfer rules in the first generation of BehaviorTran are mainly based on the output of superficial syntactic parse trees of the analysis grammar. However, since parse trees are usually huge and branchy, and sentences similar in meaning may be presented in different surface syntactic structures, the transfer operations required for producing good translations are numerous and usually very complex and complicated. Thus, it is hard and costly to acquire a complete set of transfer rules and to ensure correct interactions among them.

Third, since the transfer operations in the first generation BehaviorTran are very complex, system designers usually tend to use minimal numbers of local adjustments in the transfer phase to get readable target translations. As a result, the output target structures usually retain a large portion of the source information, such as the part of speech of terminal words and the sentence patterns. The transfer mapping, thus, may minimize the required transfer operations, but may not optimize the translation quality. Consequently, literal translations which are not natural enough to native speakers are generated from time to time

Fourth, except the lexical tagger, which uses statistical information to solve lexical category ambiguities, most knowledge bases of the first generation BehaviorTran are written by linguists.

However, as the system scales up, this kind of rule-based approach suffers from many problems as indicated below:

- ❑ It is hard to maintain consistency of the large amount of fine-grained knowledge among different persons at different time.
- ❑ It is hard and costly to acquire the large amount of fine-grained knowledge with human intervention.
- ❑ It is hard for human to deal with complex and irregular knowledge in terms of formal and precise rules. Exceptions of rules occur from time to time.
- ❑ It is hard to maintain uncertainty knowledge due to the lack of objective preference measure.

Fifth, as the business of the BDC translation center grows, the translation domains of BehaviorTran extend from computer science to electrical engineering, mechanical engineering, aviation, navigation etc. and the number of customers and posteditors are increasing. It becomes more and more sailent that the special patterns and style in each subdomain should be taken into consideration to render satisfactory translations. Besides, the feedback from posteditors and customers should also be incorporated to improve the translation system. However, in the first generation BehaviorTran, the work of sublanguage knowledge acquisition and feedback analysis is labor-intensive and, thus, very time-consuming and not cost-effective.

Since fixing the drawbacks mentioned above requires a revolutionary change of the design philosophy and basic architecture of the first generation, we started to develop the new generation BehaviorTran.

3. Design Philosophy of the New Generation BehaviorTran

3.1 A Cooperative Approach Integrating Both Linguistic and Statistical Information

To avoid the shortcomings of rule-based approach, the design philosophy of the new generation BehaviorTran moves toward a corpus-based, statistical-oriented approach. With this approach, linguists are requested to construct the language model, corpus are used as the main information source and statistical techniques are used to learn model parameters and automatically acquire the knowledge from the corpus. The advantages of this approach are listed below:

- ❑ uncertainty or preference is interpreted objectively and consistently
- ❑ consistency can be easily maintained even in large scale systems
- ❑ automatic training is possible with least human intervention
- ❑ well-established statistical theories and techniques are available
- ❑ remove the burden of rule induction from linguists to machine
- ❑ easy to meet the desirable designing goals of wide coverage, robustness, adaptability, controllability, parameterization and cost-effectiveness

Our researches along this line of design philosophy include a bi-directional transfer model (Chang [5]), various kinds of score functions for selecting the best candidate (Su [6], Chang [7], Chiang [8], Lin [9]), semi-automatic grammar construction (Su [10]), compound extraction (Wu [11]), etc.

3.2 Introduction of Intermediate Normal Forms

Another major change in the new generation BehaviorTran is the introduction of the Normal Form (NF) levels. NFs refer to the intermediate structures between source language parse trees and target language output translations. With the introduction of NFs, we intend to set up a set of linguistically-justified intermediate levels which can separate the original transfer process into several independent phases, and can serve in a manner relatively independent of involved language paris and surface forms. This idea is close to traditional interlingua approach. However, our NFs are unlike interlingua since they are not universal representations of all languages. Instead, NFs are normalized language-specific representations minus the language-specific idiosyncracies, which are most troublesome in MT. NFs also contain the (near) universal representation of semantic roles and relations. In this sense, NFs are similar in spirit ot the reduced f-structre in Lexical Functional Gramar (LFG), which can be directly mapped to a (potential universal) semantic representation (Halvorsen [12]). Besides, another major difference between NFs and interlingua is that NFs do not involve the decomposition of lexical entries into semantic primitives (e.g. "kill" = "cause to become not alive" (Schank [13])). It has been pointed out in many MT systems (Bennett [14], Durand [15]) that the set of universal semantic primitives are hard to be clearly defined. And it is not obvious to us that the decomposition of words into primitives will improve the quality of translation.

A schematic view of the translation flow in the new generation BehaviorTran is shown in Figure 1 below. PT stands for the parse tree, NF1 stands for the first-level normal form, and NF2 stands for the second-level normal form. The subscripts 's' and 't' stand for the source and target language respectively. P(X|Y) represents the conditional probability for X to appear given that Y is observed. Such parameters (conditional probabilities) are used to assign preference scores for disambiguation.

We further assume that the parse trees are produced based on a phrase structure grammar G, the NF1 constructs are produced based on a set of normalization rules, NR1, and the NF2's are produced according to a second set of normalization rules, NR2. In addition, the reverse operations are directed by sets of generation rules of the various levels (GR2, GR1, and GR0), which specify the sets of legal $NF1_t$, PT_t and T's in the generation processes.

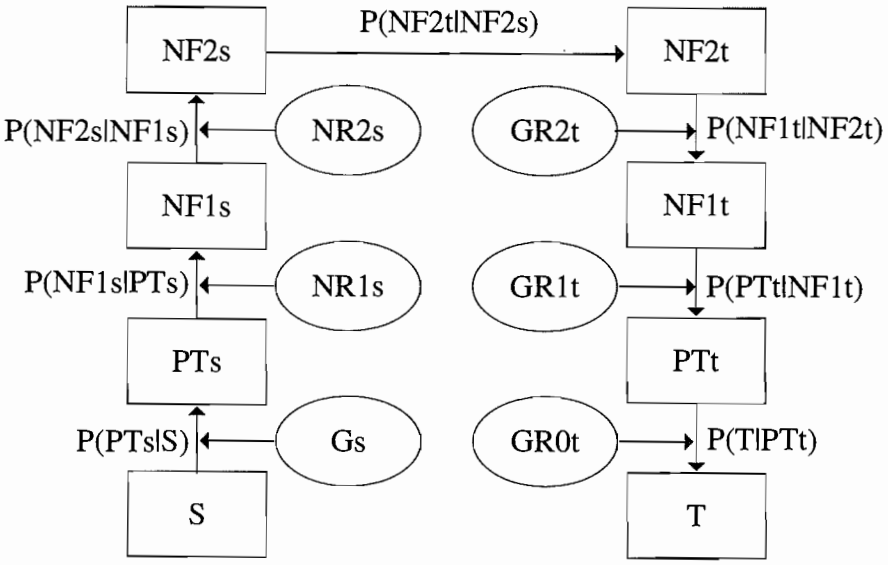


Figure 1 A Schematic View of the Translation Flow in BehaviorTran

Note that we introduce two intermediate structures in each language. NF1 is the level for syntactic normalization, in which all the elements that do not influence the cognitive meaning of a sentence will be eliminated. Thus the function words, punctuations, and unnecessary branchings and nodes are eliminated in NF1. NF2 is a semantically-oriented representation in which the basic constituents (i.e. governers, dependants, and modifiers) are marked with their semantic case roles (e.g. Agent, Theme, Time, Manner etc. (湯 [16], 詞庫 [17])), and some closed class elements (e.g. tense, aspect, modality, case markers, etc.) are extracted and recorded as a set of attribute-value pairs on the relevant nodes. Details about the NFs and their merits are illustrated in section 4.

3.3 Two-Way Training

The goal of a practical MT system is to produce fluent outputs that are natural to the native speakers of the target language. However, under the traditional transfer-based MT architecture, most output translations are strongly influenced by the sentence patterns of the source language and many literal translations are produced across the transfer phase (Somers [18], Su [19]). Such source-dependency is easily introduced to a transfer-based MT system in the one-way analysis, transfer and generation flow as mentioned earlier.

An alternative approach we propose is a two-way training approach which acquire the translation knowledge from a bilingual corpus. The bilingual corpus contain lots of well-polished source-target sentence pairs which are, undoubtedly, wonderful sources for transfer knowledge acquisition.

To change the system architecture from one-way design toward two-way design, the transfer knowledge should thus be trained from both properly normalized source and target knowledge representations, which should both fall within the range of the sentences that will be produced by the

native post-editors, according to the discourse context of the source language and the target language respectively. The following flow shows the general idea for training a two-way system. The bold arrows at the right hand side emphasize that the intermediate representations for the target language are directly derived from the target sentences in an aligned bilingual corpus.

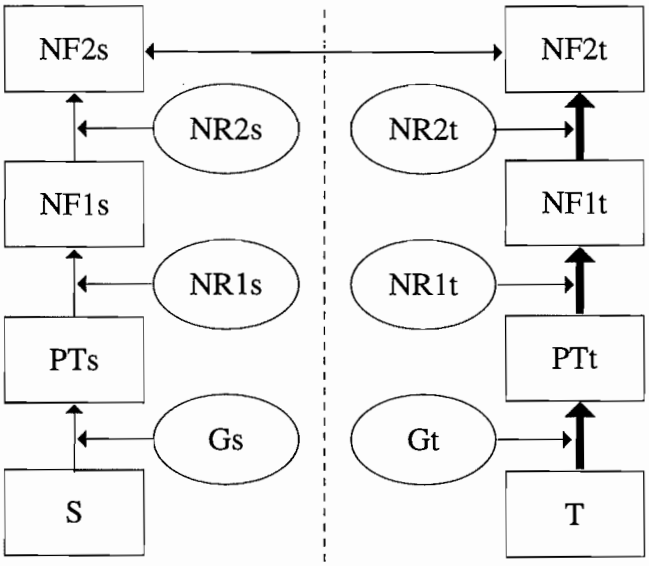


Figure 2 Two-Way Training Flow

Note that, the translation flow still follows the analysis, transfer and generation steps, but the training procedure for knowledge acquisition is different from the one-way design system. The arrow symbols indicate that the PT's, NF1's and NF2's for *both* the source and target sentences are derived from the source and target sentences respectively, based on their own phrase structure grammars and normalization rules. Thus, all such intermediate representations are guaranteed to fall within the range of the sentences that will be produced by the native speakers of the source and target languages; the transfer phase only *select* those preferred candidates among such constructs. In addition, the transfer parameters are estimated based on such intermediate representations and the transfer knowledge is derived from both the source and target sentences of an aligned bilingual corpus. Details about the formulations and training issues of the two-way training model can be found in Su [20].

4. System Architecture of the New Generation BehaviorTran

4.1 New System Architecture

With all the new ideas mentioned above, the new generation BehaviorTran gradually takes shape. The schematic veiw of the new architecture has been shown in Figure 1. Since the current interest of BehaviorTran is the language pair English-Chinese, detailed architectures of the English-to-Chinese and Chinese-to-English translation flow are presented in Figure 3 and Figure 4 respectively. The English-Chinese translation flow is briefly illustrated with an example in section 4.2 (許 [21]).

Figure 3 English-Chinese Machine Translation Flow

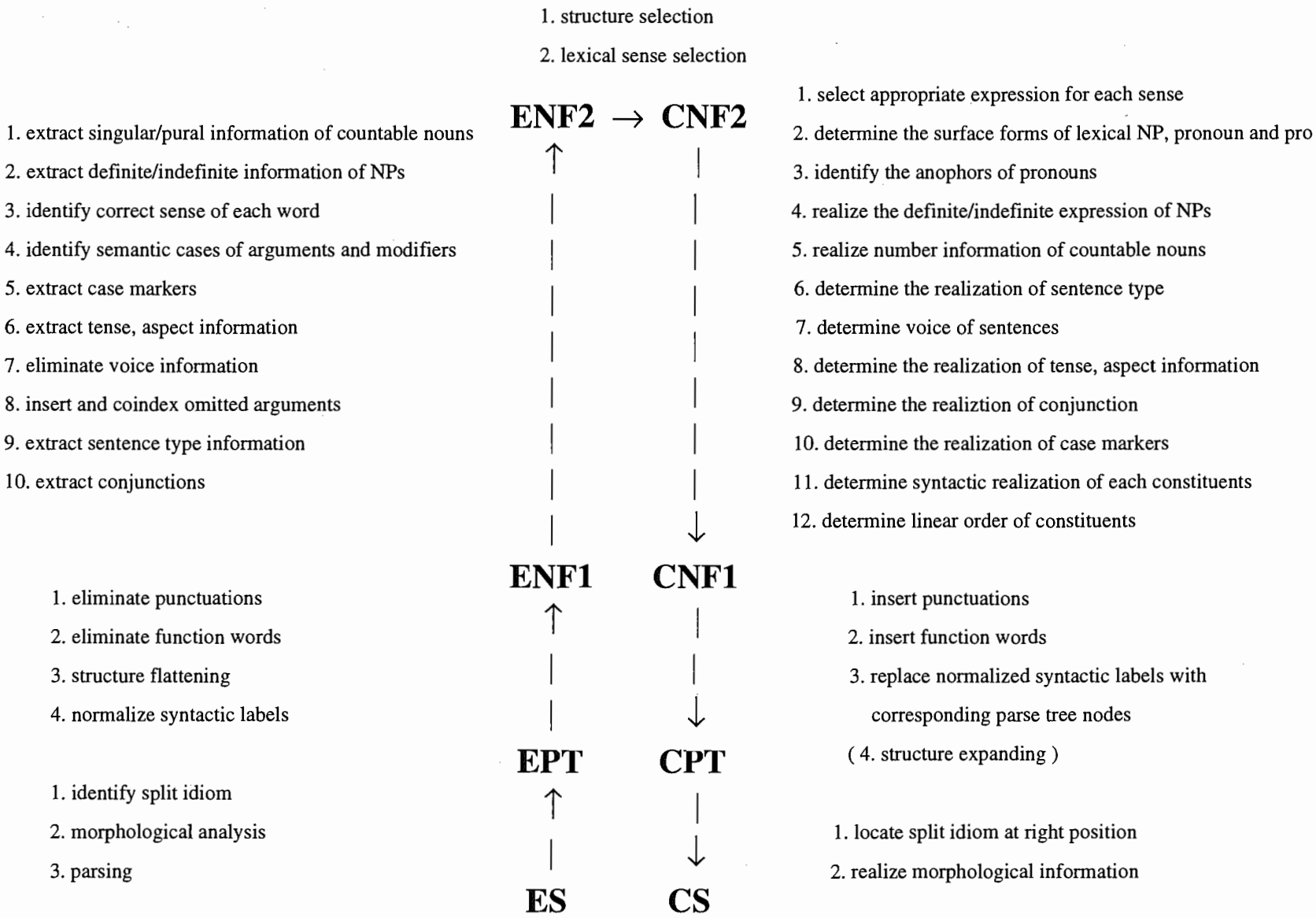
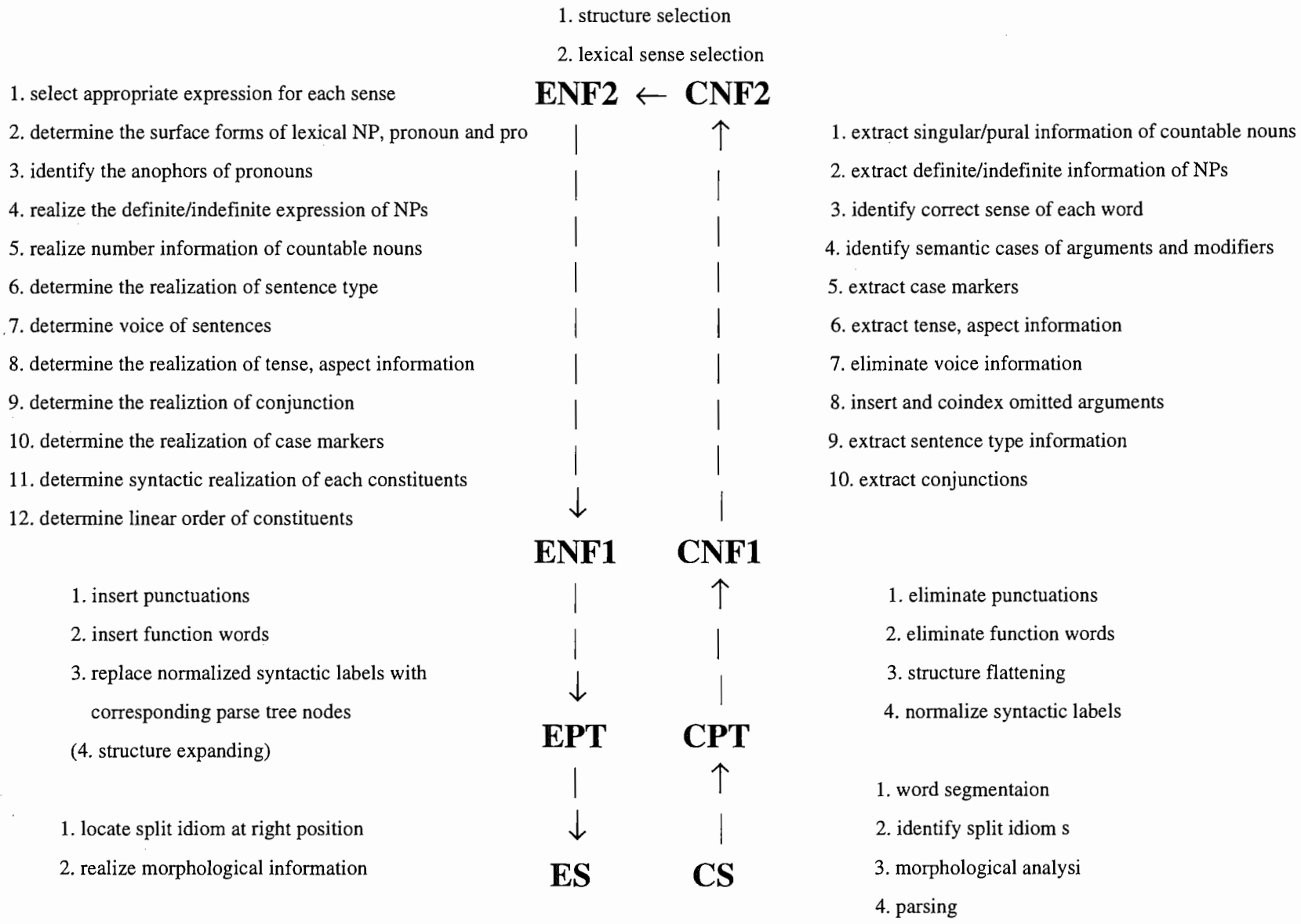


Figure 4 Chinese-English Machine Translation Flow



4.2 English-Chinese Translation Flow

A. Syntactic Parsing

The first module is a syntactic parsing module which produces the parse tree (PT) of the input sentence according to the phrase structure grammar of the source language. This module provides the syntactic information for the input sentences. An example of the parse tree is shown in the following figure. ("To meet spectrum analyzer specifications, allow a 30 minutes warm-up before attempting to make any calibrated measurements .")

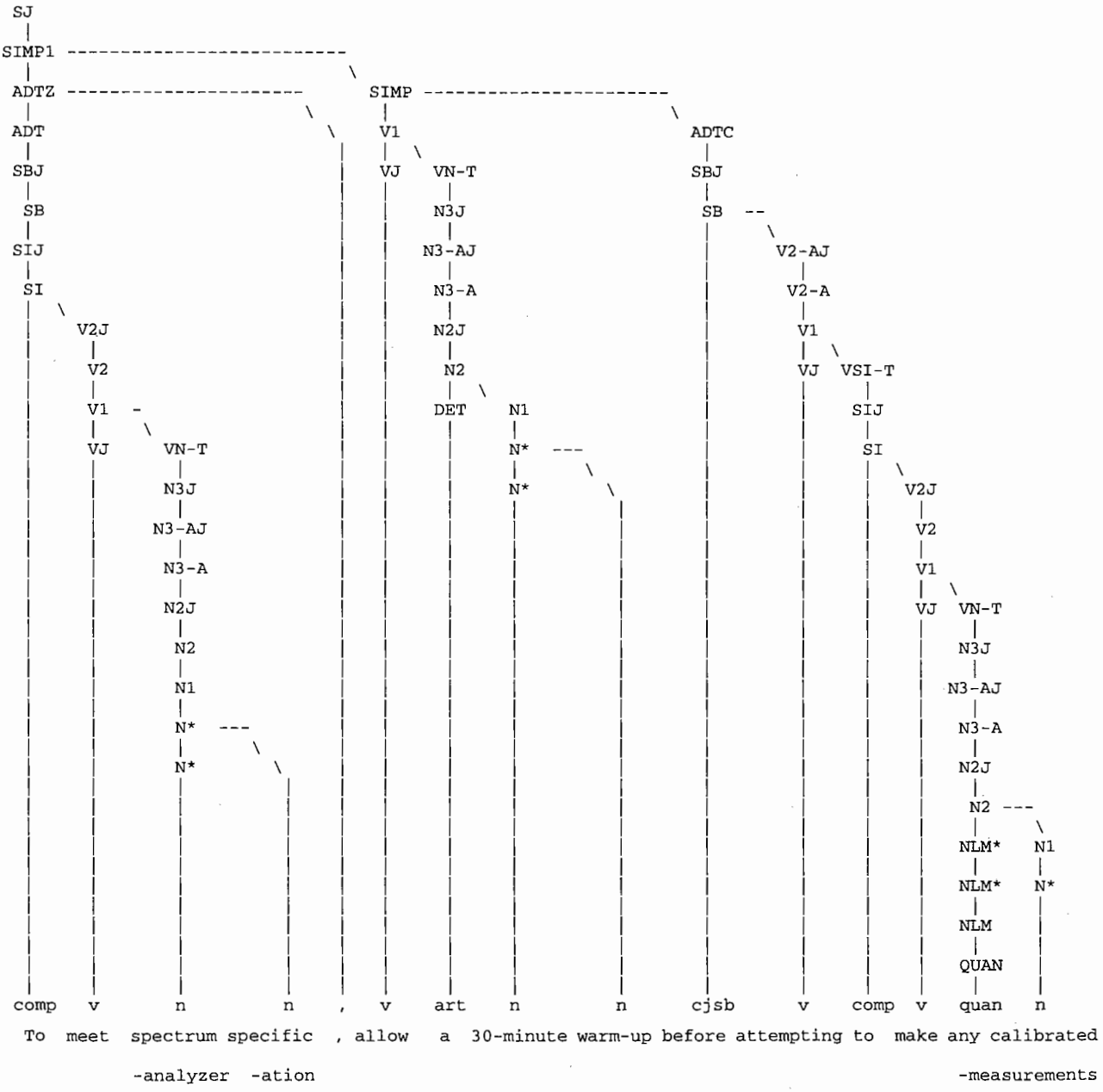


Figure 5 Example : A Source Parse Tree

Note that the parse trees produced by the phrase structure grammar of a large-scaled system are usually huge, branchy, and nodose. So it will be an arduous work to build the transfer grammar

directly from the parse tree constructs.

B. Level 1 Normalization(NF1)

In NF1 level, all the elements that do not contribute to the cognitive meaning of a sentence are eliminated. Those elements, such as punctuations, function words, and unbranching tree nodes, will not influence the choice of target translations and shall not be taken into consideration in subsequent stages. Besides, removing those redundant information will reduce the size of the possible parameter space and will simplify the process of further normalization.

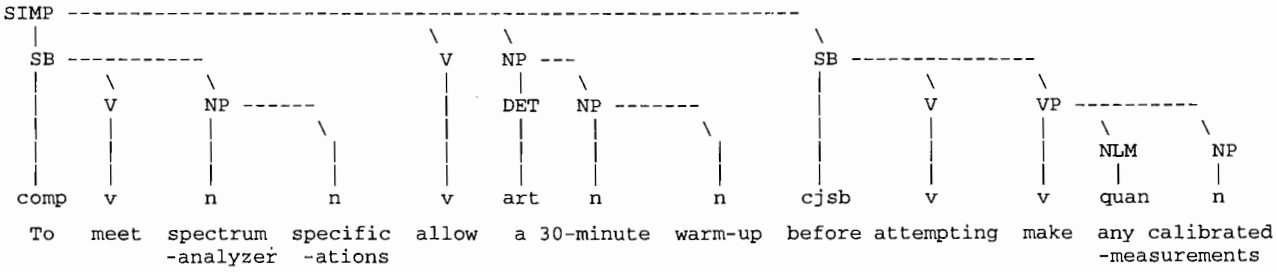


Figure 6 Example : A Source NF1 Tree

In the current example, the syntax tree is greatly compacted by retaining only the major syntax structure; a large number of nodes are compacted and re-labelled with representative node labels.

C. Level 2 Normalization (NF2)

The NF2 level is the level for semantic representation. A NF2 tree is an order-free dependency tree which specifies the semantic case roles of its governers (head), dependants (arguments), and modifiers (adjuncts), and is enriched with sets of feature-value pairs (such as tense, modality, voice, number etc.) on superier nodes.

The reason we perform the semantic-oriented normalization in NF2 is two-fold. First, as most MT researchers agree, what should be preserved in the process of translation is the semantic meaning of the source sentence instead of its syntactic structure. However, as mentioned earlier, in most traditional transfer-based MT systems, the transfer rules are constructed mainly based on the source syntactic trees, and therefore the translation outputs are usually strongly affected by the source sentence patterns and are often judged by the native speakers as "readable but not nature enough". Thus, elevating the intermediate representation from a syntactic parse tree to a semantic dependancy tree will make it possible to some extent to get rid of the tie from the syntactic information of the source sentence, and make it easier to render correct, fluent, and natural target translations.

Second, since NF2 involves feature extraction (i.e. remove some surface elements (e.g. modals, case markers, etc.) and record them as a set of feature-value pairs on superier nodes), some sentences that are different in their syntactic forms may be normalized to the same NF2 construct (e.g. active-passive pairs), and thus may further reduce the possible parameter space for statistical training. Since the parameters required to characterize the translation model may be numerous, the

compression and normalization of the intermediate constructs is a very important processes which actually makes the two-way training approach feasible.

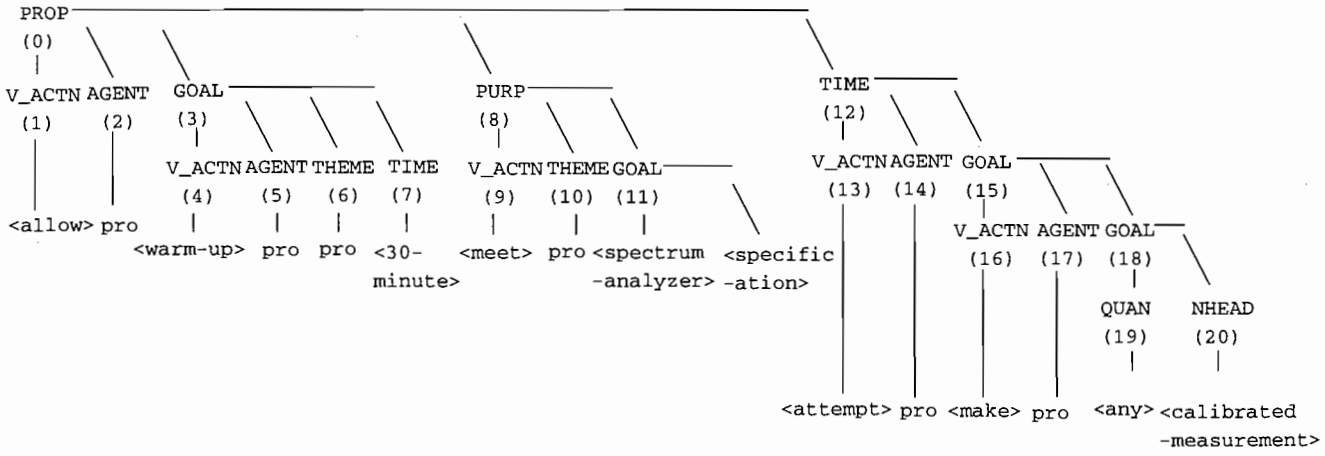


Figure 7 Example : A Source NF2 Tree

The example above is analyzed as a NF2 tree which specifies the Action (V_ACTN) being performed, the Agent (AGENT) who conducted the action, as well as the TIME, GOAL and PURPose for conducting the action (Extracted features are not presented here for simplification).

D. Target NF2 Selection

Given the NF2 tree of the source sentence, a proper NF2 tree of the target sentence could be selected among the set of target NF trees that are produced by the target analysis grammar. The selection could be made based on the parameters trained by the two-way training method, and can further incorporate the discourse and stylistic information. Note that the process is actually a 'selection' process rather than a 'derivation' or 'transfer' process from the source NF2 trees. By selection, the target NF2 is only selected from legal target NF2's, and therefore the output target NF2 will not be an illegal one.

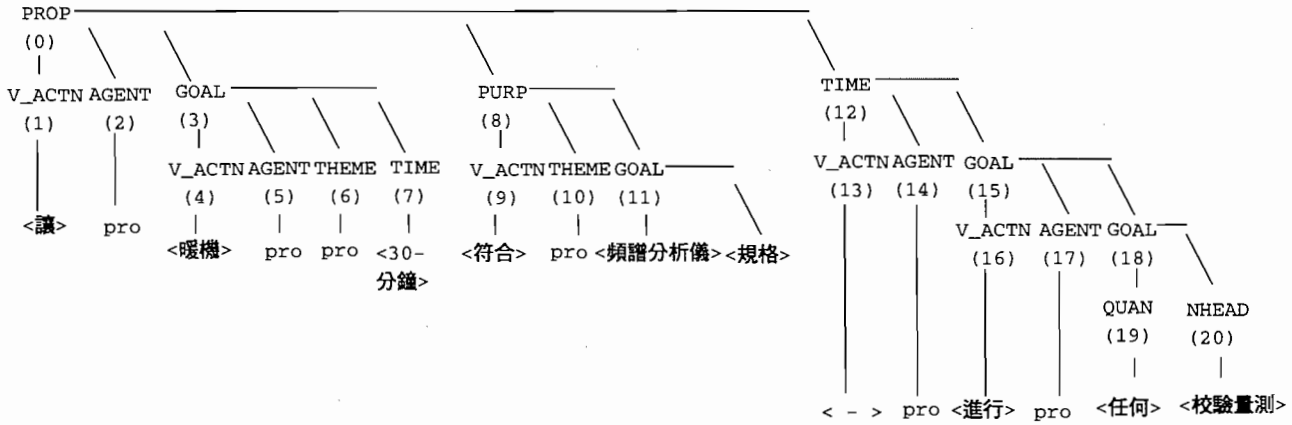


Figure 8 Example : A Selected Target NF2 Tree

In the above example, the selected target NF2 does not differ much from the source NF2 due to previous normalization. The major change here is the transfer of word senses (where a sense is

represented with a pair of angle brackets).

E. Normalized Structure Generation

Given a target NF2 that is derived from the target grammar, the next step would be to choose an appropriate normalized syntax tree for generation. As in the phase above, we may also include discourse and stylistic information in this phase to select the realization form of feature-value pairs (e.g. case markers, tense, voice, etc.) and to select the preferred linear order of constituents. An NF1 tree of the target sentence generated in this way will contain the skeletal syntactic structure for the target sentence, as shown below:

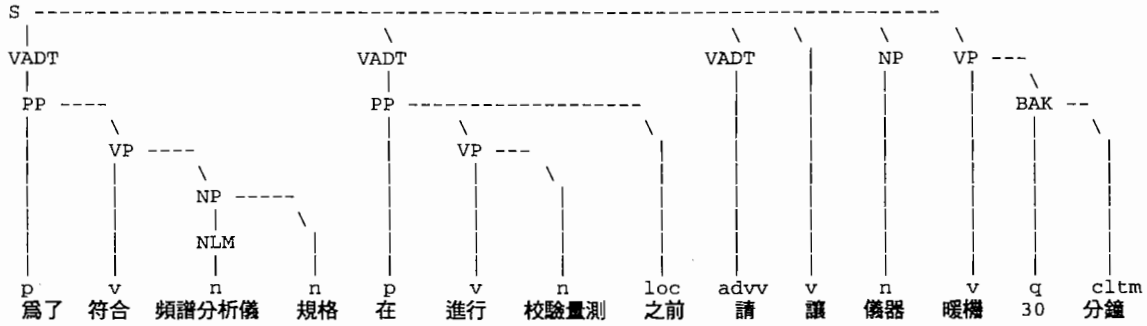


Figure 9 Example : A Selected Target NF1 Tree

Note again that the generation step here is actually accomplished by a selection process from a set of legal normalized syntax trees which are derived from the target grammar. Thus, the final translation output will not be deformed and produce unnatural translation.

F. Surface Structure Generation

After the NF1 tree is generated, the subsequent step is to determine whether some function words or punctuations should be added to improve its fluency and meet user-preferred style. The final syntactic tree of the example is shown below. (In this simple case, only two punctuation marks are patched here.)

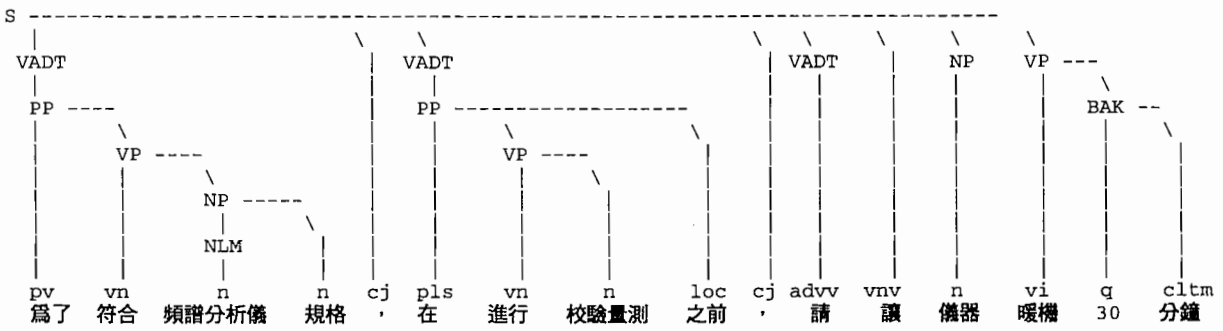


Figure 10 Example : A Selected Target Syntax Tree

G. Morphological Generation

Finally, the morphological generation is performed in the final step to generate the morphemes

required in the target language. In the above example, no specific tokens of this kind are inserted, and the final preferred translation is 〝爲了符合頻譜分析儀規格，在進行校驗量測之前，請讓儀器暖機30分鐘。〞 The Chinese-specific morpheme 「們」, which is used in conjunction with certain nouns to produce their plural form, such as 「同學們」, for example, is generated in this phase.

4.3 Merits of the New Architecture

- ❑ With the introduction of NF representations, the output of the analysis grammar for any source language can be used to synthesize any number of target languages without rewriting the analysis component, and vice versa the generation component. Thus, new language pairs may be added to the MT system with a minimum amount of development time.
- ❑ NFs separate the transfer process into several phases. Operations in each phase are independent to those in other phases. This greatly enhances the modularity of the system and lighten the burden of manipulation and maintenance in each phase.
- ❑ The target normal forms are directly derived from the target grammar, not a deformed version from the source grammar, and thus can eliminate the bias resulted from the source language.
- ❑ The mapping between the source and target normal forms can be easily tuned by the two-way training method to generate the sentences which reflect the preferred sentence patterns and styles encoded in the training corpus.
- ❑ The knowledge bases in this architecture only provide static descriptions on the legal forms of the constructs, while ambiguity resolution or preference evaluation is governed by sets of parameters. This makes it easier to adapt the system to specific user styles and maintain different parameter sets for different customers.

5. Concluding Remarks

In this paper, we present the design philosophy and architecture in the new generation BehaviorTran. With its superiority in knowledge acquisition, modularity, adaptability, and bidirectionality, this new architecture is expected to play an important role in designing the MT systems of the next generation. And all these new changes enable BehaviorTran to gain more flexibility and better performance and to move from a purely English-Chinese translation system toward a multilingual translation system.

References

[1] Chen, S.-C., J.-S. Chang, J.-N. Wang and K.-Y. Su, "ArchTran: A Corpus-Based Statistics-Oriented English-Chinese Machine Translation System," *Proceedings of Machine Translation Summit III*, pp. 33--40, Washington, D.C., USA, 1991.

[2] Wu, M.-W., J.-S. Chang and K.-Y. Su, "The Current Status of ArchTran: A Corpus-Based Statistics-Oriented English-Chinese Machine Translation System," *Proceedings of the 1991*

Workshop on Machine Translation, pp. 123-138, Nantou, Taiwan, June 24-26, 1991.

[3] Hsu, H.-H. and K.-Y. Su, "A Bottom-Up Parser in the Machine Translation System with the Essence of ATN", *Proceedings of International Computer Symposium (ICS)* 1986, Vol. 1 of 3, pp. 166-173, Tainan, Taiwan, R.O.C., Dec 17-19, 1986.

[4] Chang, C.-L. and K.-Y. Su, "A New Mechanism of Transfer for An English-Chinese MT System," Conf. on Translation Today, the Hong Kong Institute for Promotion of Chinese Culture, H.K., Dec. 17-21, 1987.

[5] Chang, J.-S. and K.-Y. Su, "A Corpus-Based Statistics-Oriented Transfer and Generation Model for Machine Translation," *Proceedings of TMI-93*, pp. 3--14, 5th Int. Conf. on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, July 14--16, 1993.

[6] Su, K.-Y. and J.-S. Chang, "Semantic and Syntactic Aspects of Score Function," *Proc. of COLING-88*, vol. 2, pp. 642--644, 12th Int. Conf. on Computational Linguistics, Budapest, Hungary, 1988.

[7] Chang, J.-S., Y.-F. Luo and K.-Y. Su, "GPSM: A Generalized Probabilistic Semantic Model for Ambiguity Resolution," *Proceedings of ACL-92*, pp. 177--184, 30th Annual Meeting of the Association for Computational Linguistics, University of Delaware, Newark, DE, USA, 1992.

[8] Chiang, T.-H., Y.-C. Lin and K.-Y. Su, "Robust Learning, Smoothing and Parameter Tying on Syntactic Ambiguity Resolution," to appear in *Journal of ACL*, 1995.

[9] Lin, Yi-Chung, Tung-Hui Chiang and Keh-Yih Su, "Discrimination Oriented Probabilistic Tagging," *Proceedings of ROCLING-V*, ROC Computational Linguistics Conference V, pp. 87--96, 1992.

[10] Su, K.-Y., Y.-L. Hsu and C. Saillard, "Constructing A Phrase Structure Grammar By Incorporating Linguistic Knowledge And Statistical LOG-Likelihood Ratio," *Proceedings of ROCLING-IV*, pp. 257--275, National Chiao-Tung Univ., Taiwan, August 18--20, 1991.

[11] Wu, Ming-Wen and Keh-Yih Su, "Corpus-based Automatic Compound Extraction with Mutual Information and Relative Frequency Count," *Proceeding of ROCLING VI*, pp. 207-216, Nantou, R.O.C., September 2-4, 1993.

[12] Halvorsen, Per-Kristian, "Semantics for Lexical-Functional Grammar," *Linguistic Inquiry* 14.4, pp. 567-615, 1983.

[13] Schank, R.C., "Identification of Conceptualizations Underlying Natural Language", in R.Schank & K. Colby (eds.) *Computer Models of Thought and Language*,, San Francisco:Freeman, pp. 187-248.

[14] Bennett, W.S., "METAL: Past, Present, and Future," *Proceedings of the 1990 Workshops on Machine Translation*, pp. 1-35, Nantou, R.O.C, June 24-26, 1991.

[15] Durand, J et.al. "The Eurotra Linguistic Specifications: An Overview." Machine Translation 6 (1991), vol. 6, no. 2, pp. 103 - 147.

[16] 湯廷池，「原則參數語法，對比分析與機器翻譯」，*漢語詞法句法四集*，學生書局，

頁 251-335。

[17] 中文詞知識庫小組，「中文詞類分析」，技術報告 (no. 93-05)，中研院資訊科學研究所。

[18] Somers, Harold, "Current Research in Machine Translation", *Machine Translation*, vol. 7, pp.231-247, Kluwer Academic Publishers, 1993.

[19] Su, K.-Y. and J.-S. Chang, "Why MT Systems Are Still Not Widely Used ? " *Machine Translation*, vol. 7, no. 4, pp. 285--291, Kluwer Academic Publishers, 1993.

[20] Su, K.-Y., J.-S. Chang and Y.-L. Una Hsu, "A Corpus-based Two-Way Design for Parameterized MT Systems: Rationale, Architecture and Training Issues," to appear in *Proceedings of TMI-95*, 1995.

[21] 許玉玲、郭明緯，「MT 新架構」，自然語言處理部門技術報告 (no. LG-94-01-NF-01)，致遠科技。

AUTOMATIC IDENTIFICATION OF COHESION IN TEXTS: EXPLOITING THE LEXICAL ORGANISATION OF ROGET'S THESAURUS

A.C. Jobbins & L.J. Evett

Department of Computing, Nottingham Trent University
Burton Street, Nottingham NG1 4BU, England
e-mail: amj@doc.ntu.ac.uk

Abstract

The identification of semantic relations between words can be applied to the subsequent identification of cohesion in texts. Groups of cohesive portions of text can be used to identify document structure and sub-topic areas. One method of automatically identifying the semantic relations between words is the utilisation of an existing lexical knowledge source, such as Roget's Thesaurus. A technique has been developed that exploits the lexical organisation of the thesaurus and this has been applied to the identification of semantically related words and of cohesion in texts. The development of this technique is outlined and the results from experiments conducted to investigate the application of the technique are presented and discussed.

Introduction

The term text is used in linguistics to describe a passage of written words of any length that forms a unified whole. How can this notion of a unified whole be recognised as existing and therefore forming a text? Human readers are able to determine whether a specimen in their native language constitutes a text because a reader has the ability to distinguish between a series of unrelated sentences and a series of related sentences which would form a text. This ability is based on the reader's knowledge of language and of the world.

A coherent text would be about particular subject areas and this is reflected in the language used. For example, a text about *Sailing* would contain words associated with this subject area, such as *jibe*, *mast*, *sail* and *wake*. If some of the words in a text are associated with certain subject areas then such words would not only be related to those subject areas but also to each other within those subject areas. Therefore, a text would contain groups of related words, for example, the words given associated with the subject area of *Sailing* are also related in meaning to each other. This relationship of meanings can also be expressed as a semantic relation. From this it can be stated that a text contains groups of semantically related words and it could be hypothesised that these

semantic relationships provide some of the information to the reader that enables them to identify a text as constituting a unified whole. The identification of semantic relations between words in a text could facilitate the recognition of a text that does constitute a unified whole. Where semantic relations between words exist across a text this provides information about the continuity of the use of the language in that text and therefore, continuity of the same context.

Halliday and Hasan proposed a theory of cohesion which identifies a passage as a coherent text [1]. Any piece of written language that is functioning as a unity, constitutes a text. It will display a quality of consistency that is defined in its grammatical structure and the meanings of the words used. Cohesion distinguishes a text from a set of unconnected sentences and identifies a text as a unified whole. They defined cohesion as follows:

> "*The concept of cohesion is a semantic one; it refers to relations of meaning that exist within the text*"

Halliday and Hasan identified cohesion in texts by locating chains of semantically related words. Their theory of cohesion has been incorporated into various work to analyse natural language texts e.g. [2], [3], [4], [5], [6].

Knowledge Sources

To identify cohesion in texts a means of locating semantic relationships between words is required. To automatically identify semantic relationships between words an existing electronic lexical knowledge source could be used, for example, Chodorow used on-line dictionaries [7], Rose et al. utilised text corpora [8] and Amsler used lexical knowledge-bases [9]. To elicit semantic relations from a dictionary would require analysis of the words used within the definitions. Analysis of text corpora could be used to extract words that frequently co-occur together and which could then be deemed as demonstrating a semantic relationship. However, this method of corpus analysis is based on statistical derivation of words and those cases of words that are semantically related but do not happen to occur in conjunction with each other in the given corpus could not be collected. The option of developing a lexical knowledge-base which could contain semantically related words requires a means of acquiring such information to form the knowledge-base. A source of lexical information that has so far not been exploited in depth in its electronic format for the extraction of semantic relations is the electronic version of Roget's Thesaurus and it was proposed to investigate this source further. The technique developed provides an automated method of extracting semantic information from the thesaurus.

Roget's Thesaurus

The thesaurus was first conceived by Peter Mark Roget in 1806 and it was actually finished in 1852. In his introduction he described his thesaurus as being the "*converse*" of a dictionary. A dictionary explains the meanings of words, whereas a thesaurus, given an idea or meaning, aids in finding the words that best express that idea. The third edition electronic version of Roget's Thesaurus is composed of 990 sequentially numbered and named categories. There is a hierarchical structure both above and below this category level. There are two structure levels above the category level and the top-most consists of eight major classes where each class is further divided into a number of subclasses and within this level there are the 990 categories. Under each of the 990 categories there are groups of words that are associated with the category heading given. The words under the categories are grouped under five possible grammatical classifications, namely: noun, verb, adjective, adverb and preposition. The paragraphs within categories and grammatical classifications are further subdivided into semi-colon groups which contain words that are even more closely related. Some semi-colon groups may have cross-references or pointers indicating to other related categories or paragraphs, these are given by a numerical reference to the category number followed by the related title in brackets. Figure 1 gives an example of a paragraph extract within category 373 and the grammatical classification of noun, in the thesaurus.

H00373.03.03.04092.00.00.%H Female
P00373.03.03.04093.01.00.%P N.
100373.03.03.04094.02.00.%T female,feminine gender,she,her,-ess; femineity,feminality,muliebrity;femininity,femininеness,the eternal feminine;womanhood 134 (adultness);womanliness,girlishness; feminism,women's rights,Women's Lib (or) Liberation;matriarchy, gynarchy,gynocracy,regiment of women;womanishness,effeminacy, androgyny 163 (weakness);gynaecology,gyniatrics;obstetrics 167 (propagation).

Figure 1: Roget's Thesaurus Category Extract

The thesaurus contains a collection of words that are grouped by their relation in meaning. Those words grouped together have a semantic relationship with each other and this information could be used to identify semantic relationships between words. For example, a semantic relationship between two words could be assumed if they occurred within the same category in the thesaurus.

The work of Sedelow and Sedelow supports the use of Roget's Thesaurus, where they claimed it to be an adequate representation of human knowledge and of English semantic space [10]. They considered the issue of multilocality of words in the thesaurus and the disambiguation of homographs by the application of a general mathematical model of thesauri. They demonstrated

that it is possible to develop algorithms that can elicit semantic structures from the thesaurus and from manual experimentation tested the semantic organisation. From these results they concluded:

> *"...any assertions that the Thesaurus is a poor representation of English semantic organization would be ill-founded and, given the depth of analysis, would have to be regarded as counterfactual."*

Morris & Hirst utilised Roget's International Thesaurus for the identification of lexical cohesion in a text as an indicator of text structure [4]. Using the thesaurus they devised a system of building lexical chains from the words of a text. These lexical chains occur due to a text being about particular subject areas and finding the structure of a text involves identifying areas of a text being about the same thing. They developed a method of determining whether two words demonstrated a cohesive link by using the information contained in the index of the thesaurus where they established five types of thesaural relations between words that constituted semantic relationships. This work involved the manual computation of lexical chains and a total of five texts were analysed [11].

Thesaural Connections

The application of the thesaurus for the identification of semantic relations between words required a means of determining what constitutes a valid semantic connection in the thesaurus between two words. For example, given words w^1 and w^2 how could the lexical organisation of the thesaurus be exploited to establish whether a semantic relation $\{w^1,w^2\}$ exists between them? Morris and Hirst identified five types of thesaural relations between words based on the index entries of Roget's Thesaurus [4]. For this approach four types of possible connections between words in the thesaurus were identified for the representation of semantic relations between words by considering the actual thesaural entries. This ensured the inclusion of all words located in the thesaurus, for example, those words that form part of a multi-word thesaurus entry may not be represented in an index entry. The connections that have been identified are considered between pairs of words and are outlined as follows:

(1) Same category connection is defined as a pair of words both occurring under the same category. Figure 2 gives an example of this connection type.

word [1]: *river*
word [2]: *tributary*

words [1] and [2] both occur under category 350

Figure 2: Same Category Connection

The words would be considered to be semantically related because they were found within the same category, where a category contains a group of associated words. This connection represents the strongest connection type of the four presented because the occurrence of words within the same category indicates they are highly related and therefore have been grouped within the same area of the thesaurus.

(2) Category to cross-reference connection occurs when a word has an associated cross-reference that points to the category number of another word. Figure 3 illustrates this connection type.

word [1]: *tide*
word [2]: *river*

word [1] occurs under category 350
word [2] has a cross-reference pointing to category 350

Figure 3: Category to Cross-Reference Connection

Cross-references occur at the end of semi-colon groups and point to other categories that closely relate to the current group of words. Therefore, the words contained under the group of words a cross-reference is pointing to are related to the current group of words that cross-reference is associated with.

(3) Cross-reference to category connection can be described as the inverse of the previous connection type given in (2). The cross-references associated with a word could be matched with the categories another word occurs under.

(4) Same cross-reference connection is defined as the cross-references of two words pointing to the same category number. Figure 4 gives an example of this connection type.

word [1]: *tide*
word [2]: *flood*

words [1] and [2] both have cross-references pointing to
category 350

Figure 4: Same Cross-Reference Connection

The association of a cross-reference with a group of words indicates that the category the cross-reference is pointing to contains words that are related to the current group. Therefore, if two groups of words both have the same cross-references associated with them this implies that the words within these two groups could also be related.

Semantic Relations

A semantic relation between two words could be predicted by the satisfaction of one or more of the four connection types identified in Roget's Thesaurus. The number of matches found between a pair of words for each of these connection types could be cumulated and this could provide a quantitative indication of the level of connectivity or semantic relatedness between the two words. However, the number of matches found between a pair of words would be influenced by the number of times those words appear in the thesaurus. For example, if a word had a high occurrence rate in the thesaurus, where it could appear under many different categories and could have many cross-references associated with it, this could distort the indications of connectivity. The probability of finding matches between words of a high occurrence would be greater than those of a low occurrence rate, due to the increased number of possible matches that could be made between these words. This could effect the accuracy of the assessment of the semantic relatedness between words, where a pair of words may have attained a high degree of matches simply because they had high rates of occurrences in the thesaurus and therefore, an increased probability of matches being found. Consequently, the number of matches found for each connection type between a pair of words were normalised. Figure 5 outlines the method of this normalisation process, where n is the number of matches found and max is the maximum number of matches that could have been made between a pair of words.

$$(n/max) \times 100$$

Figure 5: Normalisation of Number of Matches Found

Word Pairs Experiment

An experiment was conducted to determine whether the connections identified in Roget's Thesaurus could be successfully applied to the identification of semantic relations between words. This was carried out on a set of semantically related word pairs and on a corresponding set of unrelated word pairs.

Method: Forty word sets were used, each consisting of three words of between four to six characters long. The second member of each set was the primary associate of the first (a related word pair) and the third member of each set was a nonassociate of the first (an unrelated word pair). For example, for the word set: {*sweet,bitter,notice*}, *bitter* is an associate of *sweet* and *notice* is a nonassociate of *sweet*. The words and their associates selected were drawn from Postman and Keppel [12] and the nonassociates, acting as a control for each pair, were derived from an experiment conducted by Evett and Humphreys which investigated the type of information required for the lexical access of visual words [13]. The list of word sets used in this experiment are given at Appendix A.

Two approaches were taken where the same category connections were considered between pairs of words and then all four of the connection types identified were considered. For both sets of these results, to determine which pair of words in each word set (i.e. the related and unrelated word pairs) represented the strongest semantic relation, the number of matches attained were compared and the word pair that achieved the highest number of matches in each word set was selected. For example, if the word pair {*sweet,bitter*} attained a total of 20.5 matches and the word pair {*sweet,notice*} attained a total of 2.7 matches then the first word pair would be selected as representing the stronger semantic relation.

Results: Table 1 shows the results for the forty word sets, giving the overall percentage of related word pairs that attained a higher number of matches than the corresponding unrelated word pairs in each word set.

Connection Types Considered	Related word pairs more strongly semantically related
Same category	80
All four connections	87.5

Table 1: Percentage of Related Words Scoring Higher than Unrelated Words

Discussion: When considering only the same category connection 80% of the related word pairs attained a greater number of matches and therefore, were more strongly semantically related than the corresponding unrelated word pairs. When all four of the connection types were considered this result was improved upon where 87.5% of the related word pairs were correctly identified as representing a stronger semantic relation. When considering just the same category connection 40% of the unrelated word pairs failed to attain any matches and when considering all four connection types 35% of the unrelated word pairs failed to attain any matches. From these results

it can be observed that considering all four connection types yields a greater identification of semantic relatedness between a pair of words.

Cohesion in Texts

Halliday and Hasan's theory of cohesion proposed a classification of the semantic relations that exist between words within coherent texts [1]. It has been shown that the connections derived from Roget's Thesaurus can be utilised for the identification of semantic relations between words. The relations in the thesaurus represent many of the relations identified by Halliday and Hasan. The thesaurus method was extended to identify semantic relatedness or cohesion across an entire text. To assess the semantic relations found in an entire text, each word in a text was compared to every other word in that text. Therefore, if a text had n number of words then the total number of word pairs to be compared would be: n-1 + n-2 + n-3 +...+ n-n. The following algorithm, hereafter referred to as the cohesion algorithm, locates semantic relations between words across a text and provides an overall measure of cohesion for that text:

(1) Filter out the function words from the text[1];

(2) For each word in the text locate it in Roget's Thesaurus and extract the related information about categories and cross-references;

(3) Compare each word in the text to all the other words in the text and for each of these word pairs obtain the normalised number of matches found;

(4) For each word cumulate the total number of matches found and then calculate the average number of matches found for that word.

The average number of matches given for each word is used as an indication of the overall level of cohesiveness that word had with the rest of the words in the text. This figure ranges from 0 to 100 where the attainment of a 0 would indicate that word did not match with any other word in a text and 100 would indicate a successful match with every word in a text. The total number of matches found for every word in a text provides an overall measure of cohesion for that text.

1. For each of these documents the function words were removed leaving the remaining content word set. The function word set includes words such as *the*, *and*, *there*, etc., these words would be limited for the identification of semantic relations between words because of their generality of usage.

Cohesion Experiment

The experiments conducted by Halliday and Hasan to test their hypothesis of cohesion were manually executed [1]. The different aspects of cohesion were looked for in various texts and they subjectively determined whether cohesion existed within these texts. The cohesion algorithm developed produces a measure of cohesion and this measure provides a quantifiable level of cohesion in a text. This automatic technique will be consistent for all texts and not influenced by subjective decision-making about the existence of semantic relations between words.

Experiments were conducted that measured the amount of cohesion in a document and also measured the amount of cohesion in a control document, thereby providing a means to assess the success of this approach. The original document was a piece of coherent text and the corresponding generated control document represented a piece of 'incoherent' text. Fifty documents, each of at least 500 words in length, were selected at random from the Lancaster/Oslo/Bergen corpus [14]. For each of these fifty documents a control document was generated. The control documents were created with similar word characteristics to the corresponding original document. This was achieved by taking every word in the original document and randomly selecting from a lexicon a word of the same length and similar word frequency to create a control document.

To assess the level of cohesion in a text, pairs of words in that text must be compared. A text can be defined as being a piece of coherent language of any size and the comparison between word pairs could be done across an entire document or in smaller units within that document. Therefore, when determining cohesion, decisions need to be made about the search space adopted, for example, adjacent words, sentences or documents. Halliday and Hasan claimed that cohesion could exist within and across sentences:

> *"Since cohesive relations are not concerned with structure, they may be found just as well within a sentence as between sentences."*

Two experiments were conducted where content word pairs were compared across entire documents and within sentence boundaries and the cohesion algorithm was applied using two approaches where only the same category connection was considered and all four category connection types were considered.

Method: The fifty original documents and the fifty control documents detailed above were used in this experiment. The cohesion algorithm was applied to each of these documents, extracting the number of matches found for each word in a document with all the other words in that document and this was also conducted at the sentence level. Every word in a document would have an associated measure of cohesion, i.e. the average number of matches found for that word. For each document an overall measure of cohesion was produced by calculating the average of the total number of matches found. To assess whether the original document attained a higher level of

cohesion than the corresponding control document, the measure of cohesion produced for each document was compared. The document with the highest measure of cohesion was selected as the document attaining the higher level of cohesion.

Results: Table 2 shows the overall results attained for the correct selection of the original documents at the document and sentence level of analysis with both approaches to the connection types considered.

Connection Types Considered	Document Level	Sentence Level
Same category	98	92
All four connections	100	94

Table 2: Percentage of Coherent Texts Achieving Higher Scores than Controls

Discussion: When considering only the same category connection and analysing at the sentence level 92% of the original documents were successfully identified as demonstrating a higher level of cohesion than the control documents. When considering all four connection types this result was improved upon, where the number of original documents successfully identified was 94%. Analysis at the document level is shown to be more successful than analysis at the sentence level, where the consideration of just the same category connection identified 98% of the original documents and when considering all four connection types a 100% success rate was attained, where all 50 of the original documents were identified as demonstrating a higher level of cohesion than the corresponding control documents.

The experiments applying the cohesion algorithm were conducted on a large sample size of documents and for one of the approaches taken it successfully identified cohesive texts over non-cohesive texts for every set of documents investigated. This success is strong evidence for the robustness of the approach taken. This is because there are bound to be many relations between the words in the control documents, since there are so many words (500) involved. However, the relations in the texts would be expected to be more consistent and the technique successfully reflects this.

Summary and Discussion

Roget's Thesaurus is a lexical tool for language construction and understanding. It has a hierarchical structure where words are grouped by meaning, according to their semantic relations and then by grammatical categorisation. It was hypothesised that these groups of semantically

related words could be used to automatically identify semantic relations between pairs of words. A method was developed which utilised the lexical organisation of the thesaurus to identify semantic relationships between words. An experiment was conducted which applied this algorithm to pairs of associated and non-associated word pairs, identifying those word pairs that demonstrated a semantic relationship between them. It was found that for 87.5% of the associated words pairs a semantic relationship was correctly predicted over the non-associated word pairs.

Halliday and Hasan proposed a theory of cohesion which described the manner in which a text is cohesive [1]. They tested their theory through manual analysis of texts, subjectively identifying the cohesive links that existed between words in the texts they examined. Although essentially cohesion is identifiable via a series of word pairs, the theory of cohesion was proposed for identifying cohesion within units of text, whether this textual unit is a sentence, paragraph or an entire document. The technique developed that employed Roget's Thesaurus for the identification of semantic relations between words was successfully applied to the identification of cohesion across units of text. A cohesion algorithm was developed and experiments were conducted to measure the amount of cohesion found across sentences and entire documents. To validate this measure of cohesion, the amounts of cohesion across control sentences and documents were also collected and the results compared. It was found that in every case analysis at the document level attained a measure of cohesion in the original document greater than for the amount attained in the corresponding control document. The results show that this technique successfully identifies a coherent text by its level of cohesion attained relative to a control text. By calculating the average of the measures of cohesion attained for each of the coherent texts analysed, a threshold measure of cohesion could be produced which could then be used for the application of this technique to previously unseen texts.

Semantic relations between words in a text can provide much information about that text, whether it is about the overall cohesiveness of that text or the subject area of that text. Locating groups of semantically related words could be used to extract sets of words that could represent particular subject areas. These groups of words could then be applied to the problem of text subject classification. Further to this application the identification of groups of semantically related words in a text could elicit information about the structure of a text. If semantically related words adhere to particular subject areas then the identification of groups of such words throughout a text could indicate the areas in that text where certain subject areas are covered. This could provide an outline of a text's structure, where sub-topic subject area changes could be identified. For example, the identification of semantically related groups of words could cluster in different parts of a text. These clusters may represent different subject areas within that text and this could reveal the overall structure of that text. Some work has been conducted on the identification of text structure although investigations have been carried out on only a few texts. Hearst used word repetitions to divide texts into sub-topic areas [6] and Morris and Hirst used thesaural relations to generate,

manually, chains of related words [4]. Work is currently being carried out to use the present measure of cohesion to identify text structure automatically.

References

[1] M.A.K. Halliday & R. Hasan (1976) '*Cohesion in English*', Longman Group

[2] E. Ventola (1987) 'The structure of social interaction: a systematic approach to the semiotics of service encounters', *Open Linguistic Series*, Frances Pinter Publishers

[3] G. Myers & T. Hartley (1990) 'Modelling lexical cohesion and focus in written texts: popular science articles and the naive reader' in U. Schmitz, R. Schütz & A. Kunz (Eds) '*Linguistic Approaches to Artificial Intelligence*', Verlag Peter Lang

[4] J. Morris & G. Hirst (1991) 'Lexical cohesion computed by thesaural relations as an indicator of the structure of text', *Computational Linguistics*, Vol. 17, No. 1, pp. 21-48

[5] H. Kozima (1993) 'Text segmentation based on similarity between words', *Proceedings of the 31st Annual Meeting on the Association for Computational Linguistics*, pp. 286-288

[6] M.A. Hearst (1994) 'Multi-paragraph segmentation of expository texts', *Report No. UCB/ CSD 94/790*, University of California, Berkeley

[7] M.S. Chodorow, R.J. Byrd & G.E. Heidorn (1985) 'Extracting semantic hierarchies from a large on-line dictionary', *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, Illinois, pp. 299-304

[8] T.G. Rose, L.J. Evett and A.C. Jobbins (1994) 'A context-based approach to text recognition', *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval,* Las Vegas, Nevada, pp. 219-227

[9] R.A. Amsler (1989) 'Research towards the development of a lexical knowledge base for natural language processing', *Proc. 1989 SIGIR Conf. Assoc. for Computing Machinery*, pp. 242-249

[10] S.Y. Sedelow & A. Sedelow (1986) 'Thesaural knowledge representation', *Proceedings, 2nd Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Advances in Lexicography*, University of Waterloo

[11] J. Morris (1988) 'Lexical cohesion, the thesaurus, and the structure of text', *Technical report CSRI-219*, Department of Computer Science, University of Toronto

[12] L. Postman & G. Keppel (1970) '*Norms of Word Association*', Academic Press, New York

[13] L.J. Evett & G.W. Humphreys (1981) 'The use of abstract graphemic information in lexical access', *Quarterly Journal of Experimental Psychology*, 33A, pp. 325-350

[14] S. Johansson (1980) 'The LOB corpus of British-English texts: presentation and comments', *ALLC Journal*, 1

Appendix A

The following table gives the forty words sets used in the word pairs experiment, where the second word is an associate of the first word and the third word is a nonassociate of the first word.

First Word	Associate	Nonassociate
sweet	bitter	notice
butter	bread	class
smooth	rough	court
short	long	card
soft	hard	tray
chair	table	weeds
sand	dune	book
seeds	poppy	ruler
cats	dogs	pool
tree	forest	violin
never	always	tartan
cold	frost	point
pepper	salt	post
under	over	peel
thread	needle	wander
take	give	mask
apple	fruit	dress
band	brass	field
stars	moon	mind
nurse	doctor	bridge
bird	robin	class
dagger	cloak	tray
light	dark	card
horse	pony	pool
white	black	court
fast	slow	book
fish	tuna	mind

First Word	Associate	Nonassociate
plane	pilot	weeds
sleep	dream	ruler
fear	afraid	violin
round	square	tartan
nail	hammer	wander
water	bath	peel
sting	wasp	mask
face	nose	post
grass	green	point
church	priest	bridge
floor	carpet	notice
mother	child	field
lamb	sheep	dress

PROBABILISTIC LANGUAGE MODELING BASED ON MIXTURE PROBABILISTIC CONTEXT-FREE GRAMMAR

Kenji Kita and Tatsuya Iwasa

Faculty of Engineering
Tokushima University
Minami-josanjima, Tokushima 770, JAPAN

e-mail. {kita, iwasa}@is.tokushima-u.ac.jp

Abstract

This paper proposes an improved probabilistic CFG, called *mixture probabilistic CFG*, based on an idea of cluster-based language modeling. The basic idea of this model involves clustering a training corpus into a number of subcorpora, and then training probabilistic CFGs from these subcorpora. At the clustering, the similar linguistic objects (e.g., belonging to the same context, topic or domain) are formed into one cluster. The resulting probabilistic CFGs become context- or topic-dependent, and thus accurate language modeling would be possible. The effectiveness of the proposed model is confirmed both from perplexity reduction and speech recognition experiments.

1 Introduction

Recently, probabilistic language models have been shown effective in many natural language applications. One such application is automatic speech recognition. Speech inherently contains ambiguities and uncertainties that cannot be resolved by pure acoustic information. During recognition, many acoustically similar hypotheses are built. To effectively rank these hypotheses, the speech recognizer is required to rely on linguistic likelihood as well as acoustic likelihood. A probabilistic language model provides the basis for calculating linguistic likelihood.

One well-known probabilistic language model is a probabilistic context-free grammar (CFG), that is a grammar whose production rules have attached to them a probability of being used. These production probabilities are usually estimated from a training corpus under a probabilistic independent assumption, that the choice of a production rule is independent of the context. But, this simple assumption often results in a poor estimate of probability. Recently, more powerful language models beyond simple probabilistic CFGs have attracted considerable attention [1, 2, 3, 4]; some of them take context-sensitive probabilities into account.

This paper will describe an improved probabilistic CFG, called *mixture probabilistic CFG*, based on an idea of cluster-based language modeling. The basic idea of this model involves clustering a training corpus into a number of subcorpora, and then training probabilistic CFGs from these subcorpora. At the clustering, the similar linguistic objects (e.g., belonging to the same context, topic or domain) are formed into one cluster. The resulting probabilistic CFGs become context- or topic-dependent, and thus accurate language modeling would be possible.

This paper is organized as follows. Section 2 gives an overview of a probabilistic CFG. Section 3 describes a mixture probabilistic CFG. Section 4 contains evaluation experiments, including language model evaluation experiments from the viewpoint of perplexity reduction and speech recognition experiments. Finally, Section 5 presents our conclusions.

2 Probabilistic CFG: An Overview

A probabilistic CFG [5] extends a CFG so that each production rule is of the form $<A \rightarrow \alpha, p>$, where p is the conditional probability of A being rewritten into α. The probabilities of all A-productions (rules having A on the LHS) should sum to 1.

In the probabilistic CFG, the probability of a derivation can be computed as the product of the probabilities of the rules used. Suppose that

$$S \stackrel{r_1}{\Longrightarrow} \gamma_1 \stackrel{r_2}{\Longrightarrow} \gamma_2 \stackrel{r_3}{\Longrightarrow} \cdots \stackrel{r_n}{\Longrightarrow} \gamma_n = w \tag{1}$$

is a derivation of w from the start symbol S, then the probability of this derivation D is given by

$$P(D) = \prod_{i=1}^{n} P(r_i). \tag{2}$$

The probability of a sentence w is the sum of the probabilities of all possible derivations for w.

$$P(w) = \sum_{D} P(D) \tag{3}$$

The production probabilities are estimated from a training corpus as follows:

Definition of Symbols

$\{B_1, B_2, \ldots, B_I\}$ $\cdots$ A set of training sentences.

$\{D_1^i, D_2^i, \ldots, D_{n_i}^i\}$ $\cdots$ A set of derivations for the i-th sentence B_i. Here, n_i represents the number of derivations for B_i.

$N_j^i(r)$ $\cdots$ This function counts the number of rule occurrences (of its arguments) in the derivation D_j^i.

Training of the Probabilistic CFG

The conditional probabilities of rules in the probabilistic CFG were estimated using the following procedure [5].

1. Make an initial guess of $P(A \rightarrow \alpha)$ such that $\sum_\alpha P(A \rightarrow \alpha) = 1$ holds.
2. Parse the i-th sentence B_i and get all the derivations for B_i.
3. Re-estimate $P(A \rightarrow \alpha)$ by the following formula.

$$\overline{P(A \rightarrow \alpha)} = \frac{\sum_i C_A^i(\alpha)}{\sum_i \sum_\beta C_A^i(\beta)} \tag{4}$$

where

$$C_A^i(\alpha) = \sum_j \left(\frac{P(D_j^i)}{\sum_k P(D_k^i)} N_j^i(A \rightarrow \alpha) \right) \tag{5}$$

4. Replace $P(A \rightarrow \alpha)$ with $\overline{P(A \rightarrow \alpha)}$ and repeat from step 2.

3 Mixture Probabilistic CFG

3.1 Cluster-Based Language Modeling

There are two different approaches for cluster-based language modeling. The first approach addresses the data sparseness problem. In probabilistic language modeling, model parameters are usually estimated according to their frequencies in a training corpus. However, since the amount of available data is limited, many events are infrequent and do not occur in the corpus. To circumvent this problem, the training data is clumped into a number of clusters, which are then used to smooth probabilities of occurrence for infrequent events. A class-based n-gram model [6] is a typical example of this approach.

The second approach aims to increase the model precision. The basic assumption in this approach is that the language model parameters have different probability distributions in different topics or domains. The training corpus contains texts from various kinds of topics or domains. This approach first divides the training corpus into a number of subcorpora according to their topics or domains, and then performs topic- or domain-dependent language modeling. Works [7, 8] belongs to this category.

3.2 Mixture Probabilistic CFG

A mixture probabilistic CFG is based on the second approach. In a conventional manner, production probabilities are estimated using the whole training data. In a mixture probabilistic CFG, however, we divide the training corpus into N clusters, and estimate separate probability distribution for each cluster. Thus, as a result, we have N probability distributions for the CFG.

Now suppose that the training corpus T is divided into N clustsers $T_1, T_2, \cdots, T_N$. That is,

$$T = T_1 \cup T_2 \cup \cdots \cup T_N \tag{6}$$

$$T_i \cap T_j = \phi \quad (\text{if } i \neq j) \tag{7}$$

Let $P_i(S)$ denote the probability of sentence S using the probability distribution obtained from cluster T_i. Then, the mixture probabilistic CFG calculates the probability of S as follows:

$$P(S) = \sum_{i=1}^{N} q_i P_i(S) \tag{8}$$

In Equation 8, q_i is the probability of sentence S arising from cluster T_i and calculated as follows:

$$q_i = \frac{|T_i|}{\sum_j |T_j|} \tag{9}$$

Here, $|T_i|$ indicates the number of sentences in cluster T_i.

4 Evaluation Experiments

4.1 Corpus and Grammar

In our evaluation experiments, we used the ADD (ATR Dialogue Database) Corpus [9], which was created by ATR Interpreting Telephony Research Laboratories in Japan. The ADD Corpus is a large structured database of dialogues collected from simulated telephone or keyboard conversations which are spontaneously spoken or typed in Japanese or English.

Currently, the ADD Corpus contains textual data from two tasks (text categories); one consists of simulated dialogues between a secretary and participants at international conferences (Conference Task); and the other of simulated dialogues between travel agents and customers (Travel Task). In our experiments, we used the keyboard dialogues from the Conference Task.

In the experiments, we also used a Japanese intra-phrase grammar for the Conference Task. This grammar does not describe a sentence structure, but it describes constraints inside Japanese phrases. Figure 1 shows some productions in our grammar.

<start>	⟶	<bunsetu>
<bunsetu>	⟶	<interj>
<bunsetu>	⟶	<conj>
<bunsetu>	⟶	<np>
<bunsetu>	⟶	<vaux>
<bunsetu>	⟶	<quote>
	.....	
<np>	⟶	<n-suffix>
<np>	⟶	<n-suffix> <p-k-wa>
<np>	⟶	<n-hutu> <p-kaku-ga>
	.....	
<interj>	⟶	m o sh i m o sh i
	.....	

Figure 1: Example of CFG productions.

In Figure 1, the grammar symbols quoted by <> indicate nonterminal symbols. The start symbol, indicated by <start>, is rewritten into phrase category names. For example, <inter>, <conj> and <np> are nonterminal symbols for interjection words, conjunctional phrases and noun phrases, respectively. Our grammar was written for phone-based speech recognition, thus terminal symbols were phone names.

Table 1 shows the size of the grammar and the training/evaluation data.

Table 1: Size of the grammar and the training/evaluation data.

Number of productions	2,590
Number of words	1,591
Number of training data	34,301
Number of evaluation data	693

4.2 Corpus Clustering

Corpus clustering is required to derive probability distributions in a mixture probabilistic CFG. In our evaluation experiments, the clustering was conducted using

phrase category names such as <interj>, <conj> or <np>. We first segmented the training corpus into phrases, and then assign a phrase category name to each phrase. Category assignment was carried out by analyzing each phrase using the the intra-phrase grammar. In this way, the training corpus was devided into a number of clusters according to their phrase categories.

There is one thing that should be noted here. Since the parameters for the mixture probabilistic CFG are derived by statistical estimation from each cluster, the size of each cluster (the number of phrases belonging to each cluster) is largely reponsible for the quality of the model. In other words, in order to estimate the reliable probabilities, each cluster must have enough data. In our experiments, the intra-phrase grammar had 109 phrase categories. However, after clustering based on these 109 categories, some clusters had very few data. For the reliable statistical estimation, clusters having fewer than 10 phrases (32 clusters in total) were merged into one cluster. As a result, we had 78 clusters obtained.

4.3 Evaluation Results

To evaluate the quality of a mixture probabilistic CFG, we calculated the *test-set perplexity* [10]. As a comparison, we also calculated the test-set perplexity of a simple probabilistic CFG. The test-set perplexity is the information-theoretic average branching of words along the test sentences (test set), and is used as a measure of the difficulty of a recognition task relative to a given language model. In general, speech recognition performance is expected to increase as the test-set perplexity decreases. Thus, a language model with low perplexity is better.

As stated earlier, terminal symbols of the CFG were phone names. Therefore, we actually calculated the test-set perplexity per phone. A formula for the test-set perplexity per phone, PP, is given by:

$$PP = 2^{LP} \tag{10}$$

$$LP = -\frac{1}{N_W}\sum_{i=1}^{N_S}\log_2 P(S_i) \tag{11}$$

where N_S is the total number of phrases in the test-set, N_W is the total number of phones in all phrases, and $P(S_i)$ is the language model probability for the i-th phrase S_i. The results of perplexity measurements are summarized in Table 2, which supports the effectiveness of the mixture probabilistic CFG.

Table 2: Test-set perplexity

Simple probabilistic CFG	2.77 / phone
Mixture probabilistic CFG	2.47 / phone

4.4 Speech Recognition Experiments

We also conducted speech recognition experiments using three language models:

- Pure CFG (without production probabilities),
- Simple probabilistic CFG,
- Mixture probabilistic CFG.

As the speech recognition system, we used the *HMM-LR system* [11, 12], which is an integration of *hidden Markov models* (HMM) [13] and *generalized LR parsing* [14]. The HMM-LR system is a syntax-directed continuous speech recognition system. The system outputs sentences that the grammar can accept.

The speech recognition experiments were conducted under the speaker-dependent condition, using discrete-type, context-independent HMMs without duration control. The results reported in Table 3 compare three language models in terms of phrase recognition performance. The mixture model attains the best performance.

Table 3: Phrase recognition performance

Pure CFG (without production probabilities)	83.6%
Simple probabilistic CFG	86.4%
Mixture probabilistic CFG	89.0%

5 Conclusion

This paper proposed an improved probabilistic CFG, called *mixture probabilistic CFG*, based on an idea of cluster-based language modeling. The effectiveness of the proposed model was confirmed by perplexity reduction and speech recognition experiments.

References

[1] Su, K. Y. and Chang, J. S.: "Semantic and Syntactic Aspects of Score Function", *Proc. COLING-88*, pp. 642-644 (1992).

[2] Chitrao, M. V. and Grishman, R.: "Statistical Parsing of Messages", *Proc. DARPA Speech and Natural Language Workshop*, pp. 263-266 (1990).

[3] Magerman, D. M. and Marcus, M. P.: "Parsing the Voyager Domain Using Pearl", *Proc. DARPA Speech and Natural Language Workshop*, pp. 231-236 (1991).

[4] Black, E., Jelinek, F., Lafferty, J., Magerman, D. M., Mercer, R. and Roukos, S.: "Towards History-based Grammars: Using Richer Models for Probabilistic Parsing", *Proc. DARPA Speech and Natural Language Workshop* (1992).

[5] Fujisaki, T., Jelinek, F., Cocke, J., Black, E. and Nishino, T.: "A Probabilistic Parsing Method for Sentence Disambiguation", In *Current Issues in Parsing Technology*, Tomita, M. (Ed.), pp. 139-152, Kluwer Academic Publishers (1991).

[6] Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C. and Mercer, R. L.: "Class-based n-gram Models of Natural Language", *Computational Linguistics*, Vol. 18, No. 4, pp. 467-479 (1992).

[7] Carter, D.: "Improving Language Models by Clustering Training Sentences", (1994).

[8] Iyer, R., Ostendorf, M. and Rohlicek, J. R.: "Language Modeling with Sentence-Level Mixtures", *Proc. of the Human Language Technology Workshop*, pp. 82-87 (1994).

[9] Ehara, T., Ogura, K. and Morimoto, T.: "ATR dialogue database", *Proc. of the 1990 International Conference on Spoken Language Processing*, pp. 1093-1096 (1990).

[10] Lee, K. F.: *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers (1989).

[11] Kita, K., Kawabata, T. and Saito, H.: "HMM Continuous Speech Recognition Using Predictive LR Parsing", *Proc. ICASSP-89*, pp. 703-706 (1989).

[12] Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T. and Shikano, K.: "ATR HMM-LR Continuous Speech Recognition System", *Proc. ICASSP-90*, pp. 53-56 (1990).

[13] Huang, X. D., Ariki, Y. and Jack, M. A.: *Hidden Markov Models for Speech Recognition*, Edinburgh University Press (1990).

[14] Tomita, M. (Ed.): *Generalized LR Parsing*, Kluwer Academic Publishers (1991).

ARE STATISTICS-BASED APPROACHES GOOD ENOUGH FOR NLP?

A CASE STUDY OF MAXIMAL-LENGTH NP EXTRACTION IN MANDARIN CHINESE

Wenjie Li, Haihua Pan[*], Ming Zhou[†], Kam-Fai Wong and Vincent Lum
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
E-mail: {wjli,hpan}@se.cuhk.hk

Abstract

Statistics-based approaches became very popular in recent NLP researches, because of their apparent advantages over linguistics or rule-based approaches. Some even claimed that it would not be necessary to employ the latter approach at all. Thus, it seemed necessary to evaluate such claim and the applicability of the former to NLP in general.

Because of the usefulness of noun phrases (NPs) in many applications, in this paper, we present a simple statistics-based partial parser to detect the boundaries of maximal-length NPs in part-of-speech tagged Chinese texts. On the basis of our experimental results, we will show that statistics-based approaches with purely part-of-speech tags are not adequate for NP extraction in Chinese; they fail to handle cases with structural ambiguity. Our experiments suggest that syntactic and semantic checking is necessary to correctly mark the boundary of maximal-length NPs in Chinese. We conclude with possible solutions to the problematic cases for statistics-based approaches.

1 Introduction

Noun phrases are the basic building blocks of sentences in natural language. They are the basic means for representing concepts in human cognition. They are also

[*]Department of Chinese, Translation and Linguistics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[†]Department of Computer Science, Tsinghua University, Beijing, PRC.

the more appropriate translation units in language translation than words or part-of-speech classes, as argued in Van der Eijk [7]. Furthermore, noun phrases abound in our daily documents and conversation. Thus, extracting NPs from running texts is very useful for many applications such as verb frame characterization, document indexing, information retrieval, sentence parsing, machine translation, etc.

Traditionally, to obtain a noun phrase in a text means to parse the whole sentence first, and then extract the partial tree with NP labels. However, this whole-partial method is quite difficult and involves a great deal of complexity, since various ambiguities cannot be resolved by syntactic or even semantic information. So recently, phrase-oriented partial parser or phrase extractor is gradually explored in noun phrase extraction and preposition phrase attachment (Church [4], Rausch, Norrback and Svensson [11], Bourigault [2], Voutilainen [15], Chen and Chen [3], etc.). The majority of the literature on NP extraction prefer statistics-based approaches over rule-based approaches to avoid detailed and tedious linguistic engineering. Although there are several studies on extracting NPs in English and non-Asian languages using stochastic methods, studies on extracting Chinese NPs have not been reported thus far.

In this paper, a probabilistic partial parser is proposed to extract maximal-length noun phrases in Chinese, which will be used in an information retrieval system. Our research aims at examining the applicability of stochastic methods in parsing Chinese. On the basis of our experimental results, we argue that merely statistics-based approaches with part-of-speech tags are not adequate for maximal-length noun phrase extraction in Chinese, and it is necessary to employ syntactic and semantic information and some kind of rule-based techniques in detecting the boundary of NPs.

2 Previous Work

Church [4] proposes a part-of-speech tagger and a simple non-recursive noun phrase extractor. His noun phrase extractor brackets the "minimal-length noun phrase" (non-recursive) in part-of-speech tagged texts according to two probability matrices: starting NP matrix and ending NP matrix; the same methodology has been used by Garsde and Leech [9] in their probability parser. By calculating the probabilities of inserting an open or close bracket between all pairs of parts of speech, Church achieves a recall rate of 98%, i.e., only 5 out of 248 noun phrases are missed. Although the recall rate is pretty high, the test corpus is too small, and only minimal-length non-recursive NPs are tested.

Rausch, Norrback and Svensson [11] design a *nuclear noun phrase* extractor which takes part-of-speech tagged Swedish texts as input and inserts brackets around noun phrases, i.e., sequences of determiners, premodifiers and nominal phrase. Their system can identify 85.9% of all nuclear noun phrases in a 6,000 word long text collection with a precision of 84.3%.

Bourigault [2] reports a tool, *LECTER*, for extracting terminologies from French texts, and it can extract *maximal-length noun phrases.* His system can recognize 95% of all noun phrases, that is, 43500 out of 46000, from the test corpus. However, no figures are given on how much 'garbage' the system suggests as noun phrases.

Voutilainen [15] also announces an *NPtool* to acquire maximal-length noun phrases. It uses a lexicon with part-of-speech tags and head information, and two rule bases (one is NP-hostile; the other is NP-friendly) for the task. The two mechanisms produce two NP sets and the intersection set of them will be labeled as the final NP. The recall is 98.5-100% and the precision is 95-98% in different domains, which is validated manually by some 20000 words. But as pointed in Chen and Chen ([3]), the recall is only about 85% according to the sample text listed in his appendix.

Chen and Chen [3] design a new and more sophisticated mechanism by combining the statistical method and rule-based method for extracting simple English NPs based on the SUSANNE corpus [13]. They use a probabilistic partial parser with dynamic programming to find out the best liner chunk sequence for the tagged input sentences, and then assign a syntactic head and a semantic head to each chunk with the help of linguistic knowledge. Then the plausible maximal noun phrases are extracted and connected according to the information of syntactic head, semantic head and a finite state mechanism with only 9 states. The average precision is 95%. Due to the difficulty of distinguishing different NP types such as *maximal-length* NPs, *minimal-length* NPs, etc., the average recall is hard to measure, and Chen and Chen only give a suggestive recall of 96% for simple NPs which contain no prepositional phrases (except for the *of*-phrase) or relative clauses. That is, the recall will be much lower if all types of NPs are considered.

Since all the researches discussed so far deal with English and non-Asian languages and seem to suggest that statistics-based approaches are adequate for extracting NPs (except for Voutilainen [15], which employs rule-based methods), it is necessary to examine the applicability of statistics-based approaches to languages like Chinese. In the following, after briefly discussing the complexity of noun phrases, we

will present our experiments on extracting maximal-length NPs in Chinese using a statistics-based partial parser, and discuss the results and their implications to the viability of statistics-based approaches to natural language processing.

3 The Complexity of NPs

Noun phrases in English are usually composed of a determiner, an adjective, and a noun, though the first two elements are optional. They can also be modified by prepositional phrases (PPs) and relative clauses. When they are modified by PPs, a PP-attachment ambiguity may arise, as exemplified in (1), the PP in which can modify either the NP or the verb.

(1) John [$_V$ saw] [$_{NP}$ the girl] [$_{PP}$ with a telescope].

The PP-attachment problem is a very complex issue, and requires utilizing lexical, syntactic, semantic, and pragmatic information, so statistics-based approaches do not necessarily have advantages over rule-based approaches in dealing with such problem. NPs with relative clauses also increase the complexity of noun phrases. Because of the word order in English, it is not easy for a statistics-based parser to mark the boundary of relative clauses and the maximal-length NPs. There is also the structural ambiguity induced by the so-called garden-path sentences, as exemplified below:

(2) John told [$_{NP}$ the boy] [$_{RC}$ the dog bit] [$_S$ Sue liked him].

Simply using statistical information cannot rule out the possibility that *Sue* is analyzed as the object of the verb *bit*, which is the reason for a human parser to backtrack when the verb *liked* is encountered.

Since previous studies on English NP extraction employing statistical methods did not cover NPs with PP or relative clause modification (though Chen and Chen include *of*-phrases in the extracted NPs), they cannot provide solid evidence for the claim that statistical methods are superior to rule-based methods.

In Chinese all the modifiers of NPs precede the head noun. The PP-attachment problem and garden-path sentences induced by relative clauses are avoided in the language. Thus, Chinese presents itself as a testing case for us to examine whether the statistics-based approach can simplify the parsing problem and avoid the complexity of the whole-part method mentioned in the introduction section.

4 Extracting Maximal-length Chinese NPs

4.1 The Corpus

In our current experiments, we use a news report corpus of 30 files which contain 16660 words, 3278 NPs, and 750 sentences. On the average, there are 22 words in a sentence (not including punctuations). All the files have been tagged by TAGGER, a part-of-speech tagging system, developed by Tsinghua University, Beijing, China [1]. The tag set, designed by Beijing University, China [16], contains 24 general categories and 110 part-of-speech tags. The following shows a snapshot of the tagged corpus with marked NP boundaries, where symbol '$' marks the beginning of a sentence, and the English characters after symbol '#' indicate the part of speech of the Chinese word before '#'.[1]

...

$ [他 #rn] 对 #p 着 #utz [报话机 #ng] 拚命 #d 地 #usdi 喊 #vgo 着 #utz ， [大本营 #ng] 一时 #d 寂静 #a ， [整个 #b 绒布 #s 河谷 #ng] 回荡 #vgn 着 #utz [罗则 #npf 颤抖 #vg 的 #usde 声音 #ng] 。

$ [这 #rn 位 #qni 5 2 #mx 岁 #ng 矮 #a 壮 #a 汉子 #ng 的 #usde 眼 #ng] 里 #f 闪 #vgn 着 #utz [泪花 #ng] 。

...

4.2 Method

Our experiment consists of two parts: training and testing. Of the 30 files, we use 25 of them for training and close test, and the rest 5 for open test. First, we manually marked all the maximal-length NPs in the 25 files using "[" for left boundary and "]" for right boundary. We found conjoined NPs and many NPs with PP and/or relative clause modification in our corpus.

Second, we trained our NP extraction program (NPext) using the 25 files with all the maximal-length NPs marked to acquire statistical information about the probability of any two categories for marking left and right boundaries. Thereafter, NPext marked the maximal-length NPs in the 25 files without the boundary markers. Since NPext marked the left and right boundaries independently, we need to pair them, and several pairing methods were examined. Finally, we conducted the open test on the rest 5 files.

[1]The description of the part-of-speech tags is given in the appendix.

Table 1: Probability of Starting an NP

	a	ng	p	vgn
a	0	0.017	0	0
ng	0.031	0.021	0	0
p	0.650	0.728	0.833	0.139
vgn	0.804	0.723	0.333	0.438

4.3 Training

Following Church [4], we acquired two matrices which contain statistical information about the probability of having a left or right NP boundary between any two part-of-speech tags. Suppose that w_i and w_{i+1} are two adjacent words in the sentence, t_i and t_{i+1} are their part of speech, respectively, and NP_B and NP_E are the left and right boundaries. Then the probabilities are defined as below:

$$\begin{aligned} P(NP_B|t_i, t_{i+1}) &= \textit{probability of a left boundary} \\ &= \frac{freq(t_i, NP_B, t_{i+1})}{freq(t_i, t_{i+1})} \\ P(NP_E|t_i, t_{i+1}) &= \textit{probability of a right boundary} \\ &= \frac{freq(t_i, NP_E, t_{i+1})}{freq(t_i, t_{i+1})} \end{aligned}$$

A sample is shown in Tables 1 and 2 for the four common part-of-speech categories in the corpus: a (adjective), ng (general noun), p (preposition) and vgn (verb with an NP object). The first row is the t_{i+1}; the first column is the t_i; and the other entries are probabilities.

From Tables 1 and 2, we can see that "p" and"vgn" are most likely to start an NP, and "ng" to end an NP. Note that in Table 1 there are values larger than zero in the pairs "p" and "p", "vgn" and "p", and "vgn" and "vgn" in Chinese; this is different from English, as shown in Church [4]. The reason is that, unlike English, all the modifiers precede the head noun in Chinese. As a result, Chinese has NPs with the word order "PP N" or "Relative-Cl N", where "Relative-Cl" can start an NP with a verb of the category "vgn".

4.4 Testing

Using the knowledge acquired from the training phase, we conducted close tests on the 25 files used for training, and open tests on the 5 remaining files. In both tests, NPext

Table 2: Probability of Ending an NP

	a	ng	p	vgn
a	0	0	0	0
ng	0.57	0.028	0.744	0.837
p	0	0	0	0
vgn	0	0	0	0

Table 3: Results for Candidate Boundaries

		Close Test	Open Test
Correct No.	left	2717	494
	right	2722	523
Wrong No.	left	4040	770
	right	2882	510
NP No.		2723	555
Recall %	left	99.7	89.1
	right	99.9	94.2
Precision %	left	40.2	39.0
	right	48.6	50.6

first found the candidate boundaries of all NPs by marking left and right boundaries independently, and subsequently, it obtained the final NPs through pairing the left and right boundaries.

4.4.1 Finding Candidate Boundaries

When the probability is larger than a threshold, an appropriate boundary marker is inserted. For instance, for the word pair 在 #p 'at' and 学校 #ng 'school', if the threshold is set to 0.4, "[" will be inserted between 在 and 学校, but not "]", since P(NP_B|p,ng) is 0.728 which is larger than the threshold 0.4, and P(NP_E|p,ng) is zero, which is less than the threshold. Table 3 shows the results for candidate boundaries when the threshold is set to zero.

4.4.2 Pairing Left/Right Boundaries

Since the statistical method only depends on statistical information, the marked left/right boundary can be incorrect. Furthermore, there may be more left bound-

Table 4: Recall After Pairing

Combinations		Forward %		Backward %	
left	right	close	open	close	open
ML	ML	79.7	67.7	79.1	67.8
* MP	MP	81.9	69.4	81.8	69.1
ML	MP	79.6	67.6	79.8	67.8
MP	ML	80.7	69.1	80.6	68.7

aries marked than right boundaries, or the other way round. In order to get correct maximal-length NPs, two methods, *maximal length* (ML) and *maximal probability* (MP), were employed to pair the candidate left/right boundaries. The maximal probability method chooses the candidate boundary with the highest probability, while the maximal length method selects the left and right pair with the maximal length. For example, suppose that we have three left boundaries and two right boundaries marked for a candidate NP, then we will choose the outmost boundaries as the left and right boundaries, if we apply the maximal length method to both left and right boundaries. But, if we use the maximal probability, then the boundaries with the highest probability will be chosen as the left and right boundaries, respectively.

By varying the direction of pairing: forward and backward, and using different combinations of the two methods: ML and MP, we had eight ways of pairing the left and right boundaries. Tables 4 and 5, respectively, show the recall and precision of the final maximal-length NPs, where the candidate boundary set was acquired with a threshold of 0.1.[2]

The comparison of the eight pairing strategies leads to the following conclusions:

- There does not exist much difference between the two directions of pairing: forward and backward, which means that, for Chinese, the characteristic of starting an NP and ending an NP is almost the same.

[2]Note that our precision and recall were calculated based on the definitions given in Chen and Chen, [3] repeated below, where "a" represents the number of NPs marked by both NPext and the human evaluator, "b" the number of NPs marked by NPext only, and "c" the number of NPs marked by the human evaluator only.

$$Precision = a/(a+b) * 100\% \quad (1)$$

$$Recall = a/(a+c) * 100\% \quad (2)$$

Table 5: Precision After Pairing

Combinations		Forward %		Backward %	
left	right	close	open	close	open
ML	ML	77.1	68.9	77.3	70.3
MP	MP	78.0	67.3	77.3	69.7
ML	MP	76.9	68.8	77.2	70.3
* MP	ML	78.1	70.6	78.7	71.3

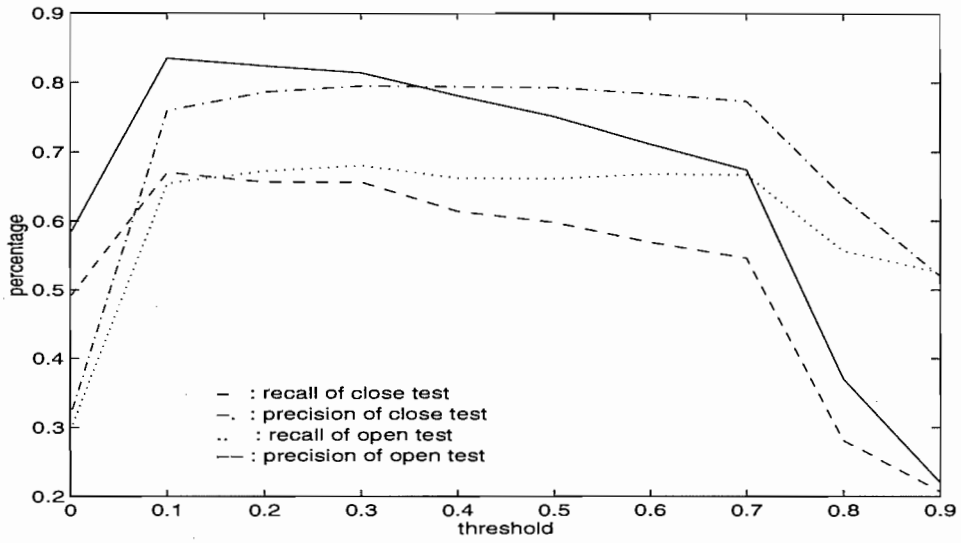


Figure 1: Recall and Precision With Different Thresholds

- Table 4 suggests that the combination of maximal probability for both left and right boundaries with the *forward* direction leads to the best recall, while Table 5 indicates that the combination of maximal probability for left boundary and maximal length for right boundary with the *backward* direction gives us the best precision; both are marked by '*' in the tables.

We also calculated the precision and recall of the close and open tests with thresholds varying from 0 to 1 for obtaining the candidate left and right boundaries. Figure 1 shows us the experimental results after pairing. We can see from Figure 1 that the threshold of 0.1 gives us the best precision and recall for both close and open tests.

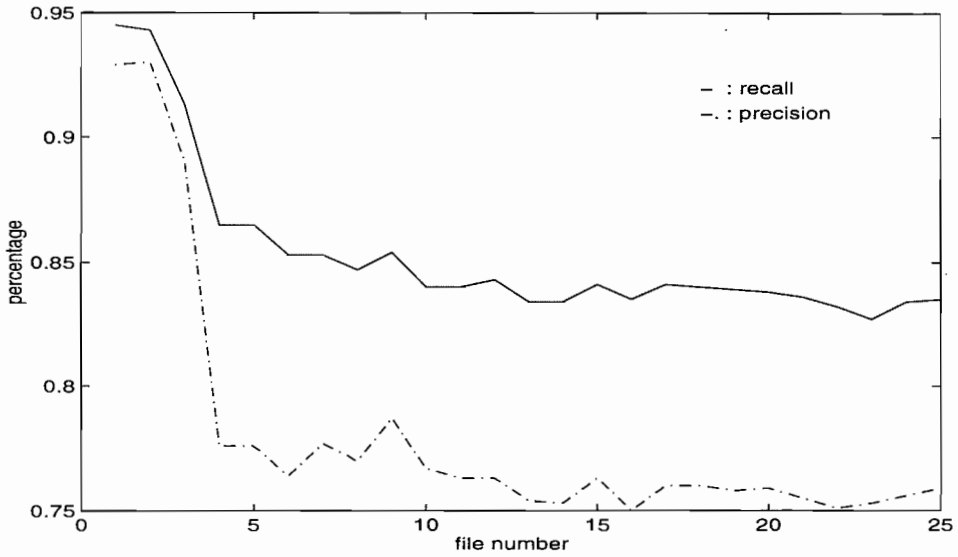


Figure 2: Precision and Recall for Close Test

4.5 Experiment Evaluation

This is the first attempt to find maximal-length Chinese noun phrases using statistical methods based on boundary probabilities. During the research, we found that the best recall and precision for close test are 81.9% and 78.7%, and the best ones for open test, 69.4% and 71.3%, as shown in Tables 4 and 5. Although the corpus size is relatively small, our results are reliable. This is justified by further experiments with varying training set. Results of these experiments show that the recall and precision for both close and open tests stablized within the 25 training files. The recall and precision for close tests emerged when the number of training files reached around 12, as shown in Figure 2. Similarly, those for open test stablized at around 22 training files, as shown in Figure 3.

Table 6 lists the distribution of the errors made by our NPext program. The error types are explained below with examples except for the "others" category; the errors in this type were mainly caused by wrong tags marked in the corpus and wrongly marked boundaries for training.[3]

A: The correct analysis should be two consecutive NPs, i.e., NP1 and NP2, but NPext combined them into one. Typical cases are the double subject and object

[3]We did not give the English translation for our examples, since it is not necessary to understand the content of the sentences for making our point; simply checking the tags and subscripts of the brackets is enough to verify our claims.

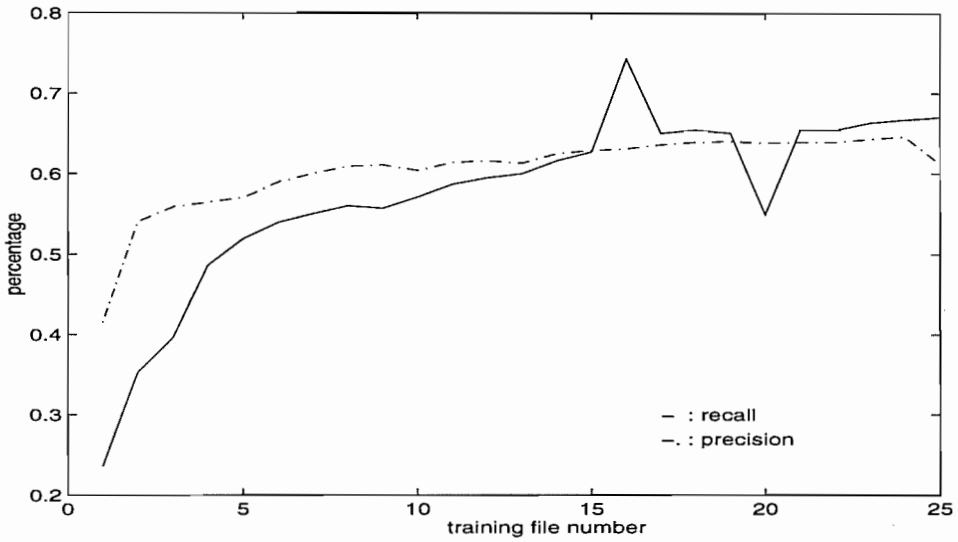


Figure 3: Precision and Recall for Open Test

Table 6: Error Distribution

Error types	NP missed		NP marked wrong	
	No.	%	No.	%
A	60	13.3	30	4.16
B	91	20.1	186	25.76
C	87	19.3	205	28.4
D	10	2.2	10	1.39
E	9	2.0	9	1.38
Others	195	43.1	282	39.1
Total	358	100	722	100

constructions in Chinese. Below are two examples with the correct analysis followed by the wrongly marked one.

Correct: [$_{NP1}$ 原材料 #j 价格 #ng] [$_{NP2}$ 上涨 #vg 幅度 #ng] 大 #a.
Wrong: [$_{NP}$ 原材料 #j 价格 #ng 上涨 #vg 幅度 #ng] 大 #a.
Correct: ... 给 #vgn 了 #utl [$_{NP1}$ 我们 #rn] [$_{NP2}$ 做人 #vg 的 #usde 权力 #ng]
Wrong: ... 给 #vgn 了 #utl [$_{NP}$ 我们 #rn 做人 #vg 的 #usde 权力 #ng]

The two NPs: NP1 and NP2 in the correct sentences above were wrongly merged into one NP, as indicated by the bracketed NPs in the wrong sentences.

B: This type is the opposite of type A, i.e., the correct analysis is one NP, but it was marked as two NPs: NP1 and NP2.

Correct: ... 是 #vy [$_{NP}$ 治理 #vg 整顿 #vg 的 #usde 关键 #ng 一 #mx 年 #ng].
Wrong: ... 是 #vy [$_{NP1}$ 治理 #vg 整顿 #vg 的 #usde 关键 #ng] [$_{NP2}$ 一 #mx 年 #ng].

Our program NPext incorrectly split the one NP in the correct sentence above into two NPs: NP1 and NP2.

C: The correct analysis is an NP containing a relative clause (RC), but NPext split it into a verb and an NP, so the result is a verb phrase (VP), not an NP.

Correct: ... 在 #pzai [$_{NP}$[$_{RC}$ 制造业 #ng 中 #f 居 #vgn 主导 #ng 地位 #ng] 的 #usde 汽车 #ng 工业 #ng] 尤 #d 为 #vi 明显 #a 。
Wrong: ... 在 #pzai 制造业 #ng 中 #f [$_{VP}$[$_{V}$ 居 #vgn] [$_{NP}$ 主导 #ng 地位 #ng 的 #usde 汽车 #ng 工业 #ng]] 尤 #d 为 #vi 明显 #a 。

D: This type involves compounds and sequences like "vgn" "vgn" "ng". The correct boundary should be between "vgn" and "vgn" for the sentence below.

Correct: ... 实施 #vgn [贴 #vgn 花 #ng 分割 #vg 等 #x 合理化 #ng 建议 #nvg]
Wrong: ... 实施 #vgn 贴 #vgn [花 #ng 分割 #vg 等 #x 合理化 #ng 建议 #nvg]

E: This type involves sequences like "p" "vgn" "ng", and the correct boundary should be between "p" and "vgn", but NPext wrongly marked it between "vgn" and "ng". That is, the result should be P + NP, but it was marked as P + VP.

Correct: ... 从 #p [$_{NP}$[$_{RC}$ 稳定 #vgn 社会 #ng] 的 #usde 大局 #ng] 出发 #vgo.
Wrong: ... 从 #p [$_{VP}$[$_{V}$ 稳定 #vgn] [$_{NP}$ 社会 #ng 的 #usde 大局 #ng]] 出发 #vgo.

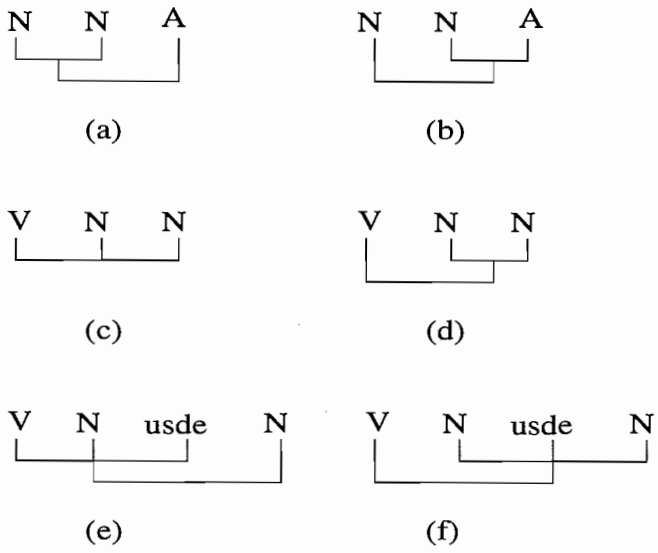


Figure 4: Structural Ambiguity

4.6 Discussion

By carefully examining the errors made by NPext, we see that most of them are structurally ambiguous. Figure 4 shows the possible structures for the sequences in (3) below, which are the possible sources for the error types mentioned earlier; note that (3a) and (3b) cover the error types A and B, and (3c) corresponds to error type C.

(3) a. N N A b. V N N c. V N usde N

Pattern (3a) has two possible structures (a) and (b) in Figure 4, (3b), (c) and (d), and (3c), (e) and (f). Note that (c) in Figure 4 is the double object configuration, but (d) is the single object configuration. The tag *usde* in (e) and (f) is a relative clause marker, or in general a modifier marker in Chinese. Structure (e) gives us an NP with a relative clause, but (f), a VP. Thus, we see that the sequences in (3) are ambiguous, but statistics-based approaches cannot differentiate them.

One of the reasons is that both structures for an ambiguous sequence are equally plausible, so using statistical information the system or program will have an equal chance in making wrong or correct prediction. For the sequence in (3c), the situation is even worse. The statistical information prefers structure (f) in Figure 4, since the probability for having a left boundary between a verb and an N is 0.723, as shown in Table 1, but the probability for a right boundary between an N and a *usde* is 0.005. Consequently, it is very unlikely for our statistics-based program to

favor structure (e) in Figure 4. That is, for sequences like (3c), an analysis of an NP with a relative clause is very unlikely. This is the reason why the C type error is very high, as shown in Table 6. Even though one could collect statistical information for more detailed tag classification, which may reflect semantic differences of some syntactic categories, we see no easy and clear ways to collect statistical information which could differentiate the structures of (e) and (f) in Figure 4 for sequence (3c).

Furthermore, although it is very difficult for a statistics-based parser to analyze the "V N" sequence in (3c) as a relative clause, it is relatively easy for a rule-based system to obtain that, since we can detect the patterns of relative clauses once a 'usde' is encountered. For instance, we can write a simple procedure to determine whether the sequence in (3c) contains a relative clause by appealing to relative clause rules or patterns.[4] The same conclusion can be applied to the sequence in (3b).

Hence statistics-based approaches are not adequate to make the necessary distinction, and some kind of rule-based approaches is necessary in extracting maximal-length NPs in Chinese. Therefore, statistics-based approaches and rule-based approaches are complimentary, and should both be employed in parsing natural languages.

5 Conclusion

In this paper, we have proposed a simple statistics-based maximal-length NP extractor for Chinese. Our experiments showed that statistics-based approaches were not adequate for maximal-length NP extraction in Chinese, since the best recall is 69.4% and the best precision, 71.3% for open tests. Therefore, it is not enough to have just the part-of-speech information and the probabilities of beginning an NP and ending an NP for NP extraction. Rule-based patterns, syntactic and semantic information should also be utilized in resolving structural ambiguities for the sequences of tags such as those, as listed in (3), which are the most problematic cases in NP extraction for Chinese, and thus a combination of statistics and rules and patterns should fare better than approaches which only employ one of them.

[4]Here, we ignore the possibility that rule-based approaches need to check semantic factors to decide whether a relative clause analysis is feasible. But it suffices to say that statistics-based approaches do not have any advantage over rule-based approaches on this matter.

Acknowledgement

The work reported in this paper is partially supported by the Hong Kong Research Grant Council under the 1994/95 Earmarked Grant for Research Initiative (RGC Ref no. CUHK 258/94E).

A The Part of Speech of Chinese Used by the System

nf	姓氏	npf	人名	npu	机构名	npr	其它专名
ng	普通名词	t	时间词	s	处所词	f	方位词
vg	一般动词	vgo	不带宾	vgn	带体宾	vgv	带动宾
vga	带形宾	vgs	带小句宾	vgd	带双宾	vgj	带兼语宾
va	助动词	vc	补语动词	vi	系动词	vy	动词“是”
vh	动词“有”	vv	来、去连谓	a	形容词	z	状态词
b	区别词	mx	系数词	mw	位数词	mg	概数词
mf	分数词	mb	倍数词	mm	数量词	mh	数词“半”
mo	数词“零”	qni	个体量词	qnc	集合量词	qnk	种类量词
qng	名量词“个”	qnm	度量词	qns	不定量词	qnv	容器量词
qnf	成形量词	qnt	临时量词	qnz	准量词	qvp	专用动量词
qvn	名动量词	rn	体词性代词	rp	谓词性代词	rd	副词性代词
p	介词	pba	把 (将)	pbei	被 (让, 叫)	pzai	在
d	副词	cf	连词前段	cpw	并连词	cpc	并连分句
cps	并连句子	cbc	分句词语间	cbs	句子间	usde	“的”
uszh	“之”	ussi	“似的”	usdi	“地”	usdf	“得”
ussu	“所”	ussb	“不”	utl	“了”	utz	“着”
utg	“过”	upb	被	upg	给	y	语气词
o	象声词	e	叹词	hm	数词前缀	hn	名词前缀
k	后缀	i	成语	j	简称语	l	习用语
x	其他	xch	非汉字	xfl	数学公式		

References

[1] Bai, Shuan-Hu. (1992) Studies and implementation of probability-based automatic part-of-speech tagging for Chinese corpora. Master's report, Tsinghua University, Beijing, China (in Chinese).

[2] Bourigault, Didier. (1992) Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of COLING-92*, pages 977–981, Nantes, France.

[3] Chen, Kuang-hua and Hsin-Hsi Chen. (1994) Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation.

[4] Church, K. (1988) A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, pages 136-143. Association of Computational Linguistics, Austin, Texas.

[5] Church, K. and P. Hanks (1989) Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22-29.

[6] Church, K. and W. Gale, et al. (1989) Parsing, word associations and typical predicate-argument relations. In *Proceedings of the 1989 DARPA Speech and Natural Language Workshop.*

[7] van der Eijk, P. (1993) Automating the acquisition of bilingual terminology. In *Proceedings of EACL'93*, Utrecht, the Netherlands.

[8] Feng, Zhiwei. (1988) The complex feature in the description of Chinese sentences. *Chinese Information Processing*, 4(3):20-29 (in Chinese).

[9] Garside, Poger and Geoffrey Leech. (1985) A probabilistic parser. In *Proceedings of Second Conference of the European Chapter of the ACL*, pages 166-170.

[10] Magerman, D. and M. Marcus. (1990) Parsing a natural language using mutual information statistics. In *Proceedings of the 28th National Conference on Artificial Intelligence.*

[11] Rausch, Norrback, and Svensson. (1992) Excerpering av nominalfraser ur löpande text. Ms., Stockholms Universiet, Institutionen för linfvistik.

[12] Salton, Gerard and Maria Smith. (1989) On the application of syntactic methodologies in automatic text analysis. *ACM.*

[13] Sampson, Jeofray. (1995) *English for the Computer: The SUSANNE Corpus and Analytic Scheme.* Oxford University Press, Oxford, UK.

[14] Sheridan, Paraic and Alan F. Smeaton. (1992) The application of morpho-syntactic language processing to effective phrase matching. *Information Processing and Management*, 28(3).

[15] Voutilainen, Atro. (1993) NPtool: a detector of English noun phrases. In *Proceedings of Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 48–57.

[16] Yu, S. W. (1992) The design of modern Chinese grammatical electronic dictionary. In *Proceedings of the International Conference on Chinese Information Processing* (in Chinese).

THEORETICAL AND EFFECTIVE COMPLEXITY IN NATURAL LANGUAGE PROCESSING

Jean-Yves MORIN
AI Program
Department of Linguistics and Translation
University of Montreal
P.O.B. 6128, Station "Centre Ville"
Montreal (Quebec)
H3C 3J7
CANADA

FAX: (1-514) 343-2284
email: morinjy@iro.umontreal.ca

Abstract

In this paper, we first review some *theoretical complexity* results relevant to NLP and we show both their interest and inherent limitations. We then argue for a notion of *effective complexity* and, we try to identify effective *sources of complexity* and *sources of determinism* in natural language processing. Finally, we show how the former can be tamed by using the latter in order to guarantee effectiveness of language computations.

Theoretical complexity

In the last few years, there has been some interest in applying the techniques of algorithmics, and especially *complexity theory* (CT) in order to characterize the computational properties of modern grammatical formalisms: LFG (Berwick, 1982), GPSG (Barton, 1985), Barton et al., 1987), Ristad, 1986a, b, c, d, 1990b), 2-level morphology (Barton, 1986, Barton et al., 1987), prosodic morpho(phono)logy (Ristad, 1990a, 1994), etc.

Here is a summary of some the results of this "language complexity game," as Ristad, 1993) has called it.

(1) The UNIVERSAL RECOGNITION PROBLEM (URP)[1] for lexical functional grammars (Bresnan, 1983 ed.) is NP-hard (Berwick, 1982), Barton et al., 1987, ch. 4).

(2) URP for two-level morphology (Karttunen, 1983), Koskenniemi, 1983) is NP-complete (Barton, 1986), Barton et al., 1987, ch. 5).[2]

(3) URP for ID/LP grammars is NP-complete (Barton, 1985), Barton et al., 1987, ch. 7).

(4) URP for unordered CFGs is NP-complete (Barton et al., 1987, appendix A).

(5) URP for classical GPSG (Gazdar et al., 1985) is EXP-POLY-hard (Barton et al., 1987, ch. 8), Ristad, 1986a, b, c, d, 1990b).

(6) URP for R-GPSG is NP-complete (Barton et al., 1987, ch. 9), Ristad, 1990b).[3]

(7) The problem of *morpheme sequence generation* and *morpheme sequence recognition* for prosodic morphology ("prosodic composition"

---

[1] Barton et al. (1987) contrast the FIXED RECOGNITION PROBLEM (FRP) and the UNIVERSAL RECOGNITION PROBLEM (URP). They consider URP to be more representative. In the case of FRP, the language is fixed and the grammar does not constitute a parameter of the problem. In the case of URP, the grammar is a parameter of the problem. They argue that grammars should constitute a parameter, because of their importance.

[2] Koskenniemi & Church (1988)

[3] R-GPSG (for Revised GPSG) is a restricted version of GPSG characterized by limitations on almost all components: limitations on the depth of syntactic categories (unit feature closure: category-valued feature can take only atom-valued features as value), limitations on the length of ID-rules, limitations on the interaction of metarules (unit closure), simple defaults replacing both (FCRs and FSDs) and limitations on UIPs (especially the Head Feature Convention).

and "prosodic recognition" , in Ristad's terminology) are NP-complete (Ristad, 1990a, 1994).[4]

But, the applicability of these results is fairly limited for many reasons.

(A) Coarseness of CT. Complexity theory gives us only a very coarse-grained classification of the complexity (or cost) of computational problems in terms of bounds (*orders of complexity*) in the *worst-case* : O(higher bound), Ω (lower bound) and Θ (both higher and lower bound). It would be more useful to have a fine-grained characterization of problems for example, in terms of *average* (or *most frequent*) case(s), but such a characterization is not (yet) available.[5]

(B) Descriptive complexity. In the case of natural language computations, *descriptive complexity* (the complexity of grammars) seems much more important than *algorithmic complexity* (the complexity of the algorithms using them).[6] There

---

[4] These problems are defined in the following way by Ristad (1994):

"The Morpheme Composition Problem for prosodic morphology ("Prosodic Composition") is to decide whether the phonological correlates of a given set of morphemes can be composed into an executable phonological structure, according to a given morphological dictionary". (Ristad (1994: 193)

"Therefore, the Possible Word Problem for prosodic morphology ("Prosodic Recognition") is to decide whether a given sequence of phonemes is subsumed by the phonetic correlate of some combination of the morphemes listed in the morphological dictionary of a particular human language". (Ristad, 1994: 196)

On prosodic morphology, see, for instance, McCarthy (1981) or McCarthy & Prince (1990).

[5] Perrault (1984) had already made this point. An analysis of overall cost (or even redeemable cost), taking into account possible optimizations of complex but frequent cases (precompilation, memoization) might also be interesting.

[6] For example, given a $O(|G|^2 * n^3)$ bound (which is the actual bound for Earley's algorithm) in any parser with substantial coverage, the size of the grammar |G| can easily be larger than 10^6 symbols (recall that in CFG-based algorithms, the lexicon must be entirely spelled out, with all inflectional and

have been many interesting developments in the field of descriptive (or Kolmogorov) complexity recently.[7] But, to our knowledge, application of Kolmogorov complexity or related approaches to natural language computations (as opposed to formal languages)[8] has been fairly limited.[9] The reasons for this should be obvious. Descriptive (Kolmogorov) complexity deals with the shortest possible descriptions of objects.[10] In the case of natural language, it is very hard, if not impossible, to prove that something is the shortest possible description, even for a single phenomenon.[11]

(C) Most of the results concern grammatical formalisms.

Partiality of grammatical formalisms. It should be obvious that any grammatical formalism is *partial*, in the sense that it will always be possible to write grammars in a formalism which are not possible grammars of human languages. Trivially,[12] this

---

derivational morphology expanded) will almost always dominate the n^3 factor, except for large values of n ($n \geq 100$) seldom if ever encountered in practice.

[7] Cf. Kolmogorov (1965), Solomonoff (1964), Chaitin (1987) for classical works and Li & Vitányi (1993) or Watanabe (1992 ed.) for a rich sample of recent developments.

[8] Cf. Li & Vitányi (1993) and Boekee et al. (1982).

[9] Cf. Rissanen & Ristad (1994) for an application of the MDL (minimum description length) principle to the acquisition of metrical phonology.

[10] Furthermore, the theory of descriptive complexity makes use of sophisticated mathematical tools with which most linguists (including the present author) are not thoroughly familiar.

[11] If one were to adopt a principle-and-parameters approach, then it might be possible to define the shortest description of a given vector of binary parameters. The problem with this kind of approach is that a lot is hidden in the *interpretation* of the parameters and this would have to be spelled out in order for Kolmogorov complexity to be applicable.

can be done for any grammar G by reducing its lexical component to only one lexical form *f*, and then associating with this form all the lexical types $\{t_1, t_2, ..., t_m\}$ defined by the original grammar: $f \in \{t_1, t_2, ..., t_m\}$ thus creating a one-word, perfectly ambiguous grammar, where all sentences are strings of *f*s.

(D) The results concerning grammatical formalisms are essentially *negative*.

It has been shown, for example, that the URP (universal recognition problem) is NP-hard for classical LFG (lexical functional grammar) and EXP-POLY-hard for classical GPSG (Barton et al., 1987), thus showing that these two formalisms are potentially intractable (i.e., not inherently efficient, qua formalisms).

(E) The reductions used in these demonstrations are not perfectly *faithful*.

(i) They are not always spelled out rigorously in full detail.[13]

(ii) They are based on artificially constructed data, which could never appear in actual natural languages or descriptions thereof.[14]

(iii) They make crucial use of *empty categories*.[15]

(iv) They deal only with abstract *grammar formalisms* and do not take into account *substantive constraints*, which effective grammars also respect.

---

[12] Assuming, uncontroversially, that any grammatical formalism will allow lexical information to be represented in some way, relating some representation of form (phonological, graphemic, etc.) with grammatical information.

[13] Manaster-Ramer (1994) makes much the same point.

[14] It could even be argued that the fact that such data present difficulties for a given formalism is a quality, not a defect.

[15] Which, it should be noted, are not an essential component of LFG, GPSG or HPSG, as opposed to GB.

Nonetheless, the mere fact that such reductions are possible within a grammatical formalism is indicative. It allows us to discover that some constraints (substantive or formal) are implicitly respected by effective descriptions but are not explicitly stated either as part of the formalism itself or as substantive constraints attached to it. For instance, for GPSG (Gazdar et al., 1985) and HPSG (Pollard & Sag, 1994), one can mention:

(a) the implicit limitation on the *length* of (the right-hand side of) ID-rules (GPSG) or ID-schema (HPSG);

(b) *functional* constraints on the contents of ID-rules or ID-schema[16]

(c) *endocentricity* of ID-rules or ID-schema[17]

(d) dispensability of *empty categories*[18]

Empty categories are not an essential component of information-based grammatical theories (as opposed to configurational theories, like GB). In information-based grammatical theories, global dependencies are linked to lexical expectations, which do not have to be computed but just searched. A trace empty node, which can be hypothesized just about anywhere in GB, corresponds to an element of a SUBCAT list, which is part of stored lexical entries and reduced only by actually occuring elements (complements, or fillers).

(e) *universal projection of lexical information*

The *universal projection of lexical information* states that all lexical categories are projected, not only major categories (but minor categories do not have *autonomous*

---

16 Daughter nodes in ID-constraints (rules or schema) are typed (e.g. lexical, head, complement, adjunct, filler, etc.)

17 Although this is not much of a constraint by itself, as Kornai & Pullum (1990) have demonstrated for all versions of X-bar theory, it is sufficient to exclude some perverse use of ID-rules.

18 Cf. Pollard & Sag (1994, ch. 9).

projections).[19] This simply captures the intuition that all lexical items in a string carry grammatical information, i.e., there are no *useless* words. It also avoids *projection paradoxes* (i.e. is a noun phrase an NP or a DP?) and the proliferation of *functional projections*, characteristic of GB approaches (where distinct information has to correspond to distinct nodes), with all the empty nodes they presuppose.

---

[19] That is, the projection of a minor category must be unified with that of a major category. For instance, in a sequence DET(erminer) QUANT(ifier) CLASS(ifier) N, all four categories have a projection DETP, QUANTP, CLASSP and NP, but only NP is autonomous. Therefore, there is only one NP node holding the grammatical information of all four projections :

DETP = QUANTP = CLASSP = NP.

Cf. Morin (1989).

(f) *off-line parsability* (or *bounded projection.*)

Off-line parsability (Kaplan & Bresnan, 1983) excludes non branching cyclic derivations, which could lead to undecidability. A grammar is off-line parsable if it does not allow derivations of the form: $A \Rightarrow^* A \Rightarrow^* \alpha$ (or, in terms of parse trees, cyclic trees of the form: $[_A \ldots [_A \alpha] \ldots]$, where ... contains only further brackets.) If such derivations (or parse trees) were allowed, the same string α could be assigned an infinite number of structures and parse trees could be infinitely deep for a given string.[20]

A grammar obeys bounded projection if and only if :

(i) any local tree admitted by the grammar is either a projection (i.e., $[_{X^i} \ldots X^{i-1} \ldots]$), an adjunction (i.e., $[_{\alpha'} \alpha \beta]$) or a coordination (i.e., $[_{\alpha'} \alpha_1 \alpha_2 \ldots \alpha_n]$) tree,

(ii) projections are bounded: there is a maximal value (2 in our model) of max for any projection X^{max}) and

(iii) it does not allow empty categories.

This constraint is much stronger than OLP, it thus also guarantees decidability.

Also, on the positive side, CT results help us identify some aspects of grammatical formalisms (e.g., empty categories, empty derivations) as potentially problematic. Empty categories allow a complex hypothesis space to grow indefinitely, independently of the length of the input. Empty derivations allow derivation trees to

---

20 Shieber gives a formal definition of OLP, which is more general than the traditional LFG (Kaplan & Bresnan, 1983: 266) one, while being applicable to abstract constraint-based grammars (where, informally, tree-nodes are labeled by trees). τ / <0> is, informally, the label of the root of tree τ and ρ is a monotonic weakening function (like subsomption).

"Definition 57 A grammar *G* is *off-line parsable* if and only if there exists a finite-ranged function ρ on models such that $\rho(M) \leq M$ for all M and there are no parse trees τ admitted by *G* such that $\rho(\tau$ / <0>) = $\rho(\tau'$ / <0>), for some τ' a sub-parse tree of τ with identical yield". (Shieber, 1992: 81)

Haas (1989) defines a constraint of depth-boundedness, which is stronger than OLP, but weaker than bounded projection.

grow indefinitely, independently of the length of the input. Moreover, it encourages us to consider natural language computations at a more abstract level than the usual *algorithmic level*, in Marr's (1980) terminology, namely the *computational level*, where problems are defined purely in terms of input and output, independently of the specific algorithms and data structures used.

There are also some results, concerning specific language computation problems, which purport to be defined more or less independently of any given formalism.

(I) URP for agreement grammars (simple grammars embodying both agreement and lexical ambiguity) is NP-complete (Barton et al., 1987, ch. 3, Ristad & Berwick, 1989).

(II) The anaphora resolution problem is NP-complete (Ristad, 1993).

These reductions seem to suggest that natural language computations are inherently NP-complete. How can we explain then (unless $P = NP$) that actual language processing by humans is normally quite efficient? One would have to resort to mysterious performance factors which would not degrade performance (like the more usual performance limitations), but, on the contrary, improve it, acting as they would as *oracles* or *accelerators*.[21]

---

[21] As a matter of fact, Ristad's position on this problem is not very clear (as Manaster-Ramer (1994) also notes in his review of Ristad, 1993). On the one hand, he virulently attacks the traditional competence-performance distinction, while, on the other hand, he uses something quite similar to account for the fact that natural language computations are not intractable, after all. A much simpler way out would be to assume that humans do not use only *linguistic* knowledge in language computations, but other sources of information, which act as sources of determinism, counteracting the sources of complexity present in natural languages.

Effective complexity

A different (but complementary) approach to the study of complexity tries to identify inherent *sources of complexity* in natural languages, as opposed to sources of complexity which are simply artifacts of the particular formalisms used.

It also tries to identify means to effectively cope with complexity problems and to reduce the disastrous effects of such complex computations by insulating them in precisely defined locations to avoid complex interaction dependencies, or by identifying *sources of determinism* that effectively constrain natural language computations.

Sources of grammatical complexity[22]

We can identify many inherent sources of grammatical complexity in natural languages.

First, there is *lexical complexity*.

The number of lexical items in any wide-coverage model is quite large ($\geq 10^n$, where $4 \leq n \leq 6$, using conservative estimates). The information associated with each of these items is complex.[23] Furthermore, lexical items can be ambiguous, the same form being associated with many types.[24] The structure of the lexicon itself can also be quite complex (with defaults, simple or multiple inheritance, lexical types, lexical rules, etc.). But an interesting feature of lexical information is that most of it can be

[22] There is no room here to discuss semantic and pragmatic complexity.

[23] In an explicit (but purely linguistic) lexicon, like the DEC for French, (Mel'cuk et al., 1984-...), for example, each lexical entry corresponds to pages of (fairly succinctly coded) information.

[24] Here, we use the word 'type' in an informal sense, to refer to any particular combination of grammatical information.

precompiled and stored.[25] Thus, at runtime, lexical retrieval can produce, in linear time, all the types associated with a given form. If a form is ambiguous, a disjunctive type will be retrieved.

Then, there is syntactic complexity.

At the level of phrase structure, any natural language exhibits a large number of grammatical constructions (local trees, in an information-based framework), with many possible values for each of the constituents (nodes in the local tree) of any construction. There is also phrase structure ambiguity: active ambiguity (many possible constituents for a given type of object) and passive ambiguity (many possible types of which a given type can be a constituent).

At the relational level, there are grammatical, thematic and rhematic relations, as well as binding relations (global dependencies, control, rection, anaphora, etc.).

Sources of determinism: natural partitions

If all these sources of complexity could interact freely, and thus combine multiplicatively, language computations would obviously be intractable. But, we would like to suggest that there are also inherent sources of determinism in natural languages, which make it possible to partition the space of objects and constraints in

---

[25] Most of it, but not all of it. For mildly inflected languages, like the Romance languages, storing precompiled inflected forms seems to be feasible and could result in an order of magnitude increase of the size of the lexicon. (We can mention that, at the end of the nineteenth century, Bescherelle hand-compiled all the inflected forms of some 8000 French verbs, resulting in a two volume dictionary of verbal forms.) Recall that most of the space in the lexicon is used for *grammatical information*, not for forms, and inflected forms share most of this information. For highly inflected languages, (like Latin, Russian, Basque, Finnish, etc.), precompilation of forms does not look like such a good idea, but precompilation of morphological processes could reduce on-line computations to very simple (deterministic) processes.

such a way that many of these sources of complexity could combine additively instead of multiplicatively.

Lexical objects and syntactic constraints

A first kind of partition, already implicit in traditional conceptions of language, is the partition of linguistic entities into *lexical objects* (stored in the lexicon) and *grammatical constraints* (represented in the grammar). In practical terms, since lexical objects are, for the most part, precompiled, this suggests a partition of parsing, for example, in two distinct phases.

First, lexical *initialization* (retrieving all stored lexical information for all the segments of a string $\omega = \omega_1\ \omega_2 \ldots \omega_m$

$$lex(\omega) = lex(\omega_1)\ lex(\omega_2) \ldots lex(\omega_m)$$

and then *parsing* proper (applying grammatical constraints to find an analysis for ω).

$$parse(lex(\omega_1)\ lex(\omega_2) \ldots lex(\omega_m)) = \sigma$$

Many parsing algorithms (including bottom-up filtering (Blache, 1990, Blache & Morin, 1990)) use such a partition.[26] Such a partition is useful inasmuch as lexical information is stored, and not computed. Even if a lot of ambiguous or disjunctive information is retrieved in this phase, it does not involve any computations. Therefore, trying to disambiguate at this point would simply reduce the size of the $lex(\omega_i)$'s, potentially removing information which will have to be recovered at a later point. It is just not worth the effort.[27]

---

[26] Lexical initialization could even be done in parallel, since lexical access for a form is completely independent of lexical access for the other forms. Cf. Sabot (1988) for the details of such a proposal.

[27] On the other hand, in some cases like speech recognition, or for languages like Chinese (where words are not separated in writing) or like Basque or Finnish (where morphological computations are needed), it might be the case that the lexical initialization part should itself be further decomposed. Cf. Gan (1994) for an interesting integrated model of word segmentation in written Chinese, where, instead of predefined partitions like the ones we are suggesting, partitions are defined on-line, by taking into account the "computational temperature of the system". When computational temperature is

Phrase structure constraints and functional constraints

Another type of partition, which is implicit in work inspired by LFG (Bresnan, 1983 ed.), or by GPSG (Gazdar et al., 1985), but much less so in HPSG (Pollard & Sag, 1994), is the one between phrase structure and functional constraints.

Maxwell & Kaplan (1993) discuss the interface between these two types of constraints, which have very different computational properties, CFG-phrase structure parsing being polynomial in the size of the input string, while known general constraint satisfaction algorithms are exponential in the size of the constraint system. They show that simple composition or simple interleaving (on-line pruning of phrasal edges not satisfying a set of functional constraints in an active chart) are both exponential in the worst case, while non interleaved pruning (caching the constraint solutions on each edge) is polynomial but involves a lot of copying overhead. What they suggest instead is *factored extraction*: extracting a concise set of functional constraints from the active chart and passing them to a constraint solver. First, a chart is built, based only on the context-free backbone grammar. Then a set of constraints is recursively extracted (starting at the root node) and combined conjunctively (except for ambiguous nodes, where they are combined disjunctively) and reduced using various classical techniques.

But what makes the strategy particularly interesting is that it uses specific linguistic knowledge in the reduction phase. Since heads and their projections share all constraints (this is itself a constraint, expressed in LFG by the equation $\uparrow = \downarrow$), head constraints are substituted for their projections. Therefore, in the case of ambiguous constituents with the same head, the disjunction can be reduced to only the constraints

coming from the effective differences.[28] This is a special case of what we call *propagation constraints* below.

Further partitions of PS-constraints

Immediate dominance and linear precedence constraints

Phrase structure constraints can themselves be further partitioned. A natural dividing line is between ID and LP constraints. Again, ID-rules or schema could be used directly to parse input and LP constraints to filter ungrammatical combinations. A variant of this general strategy, bottom-up filtering (Blache, 1990, Blache & Morin, 1990), precompiles LP relations in exclusion tables that act as prefilters on ID-rules.

Decomposition, adjunction and coordination constraints

It is a well-known fact that adjunction (and coordination which is just a particular kind of adjunction with tighter constraints) enormously complicate the search space of a parser. It might be interesting to separate the straight decomposition rules with lexical heads from both of these types. *D-rules* constrain the obligatory unification of minor category phrasal projections with permissible (and accessible) major category phrasal projections and the attachment of subcategorized complements. *A-rules* and *C-rules* constrain adjunction and coordination of lexical or phrasal categories. We can then have a partition of the *parse* function where we first do strict decomposition :

$$decomposition(lex(\omega_1)\ lex(\omega_2)\ \ldots\ lex(\omega_m)) \quad = \quad \sigma'$$

---

high, anything goes, so to speak, and low-level constraints are applied, more or less at random. When computational temperature cools down and some structures have crystallized, only high-level constraints are applicable, if this does not work, temperature goes up again and so on, so forth.

[28] They also discuss the necessity of moving some functional constraints into the context-free part in order for factored extraction to be efficient, since their strategy is very sensitive to the specific form or the grammar used, as demonstrated in their experiments. Propagation constraints are more general (and, hopefully, robust) in that respect.

and then apply *adjunction* and *coordination* in an interleaved manner, but only if needed.[29]

$$adjunction\text{-}coordination(\sigma') = \sigma$$

In other words, only *D-rules* are always active (and their applicability is bound by the number and nature of lexical forms $lex(\omega_i)$ in the representation to be parsed. *A-rules* and *C-rules* are only activated when no more *D-rules* are applicable and there are still unattached constituents. *C-rules* also need the presence of specific markers. In that way, *adjunction-coordination* never interferes with *decomposition* and the composition of *decomposition* and *adjunction-coordination* is additive, not multiplicative.

An interesting feature of D-rules is that they only need to refer to coarse- or medium-grained grammatical information : parts of speech, projection level, functional constraints (*SUBCAT*, *SPEC-OF*, *ADJUNCT-OF*, etc.) and all this information is directly accessible in lexical entries and does not have to be computed.

Furthermore, once we adopt the hypothesis of *universal projection of lexical information* and *bounded projection*, it becomes possible, to strictly bound the number n of possible nodes in a parse tree given a sequence of i lexical forms ($n < 3i$) ($n \leq 2i$ for decomposition nodes and $n \leq i-1$ nodes for adjunction-coordination nodes). Of course, this presupposes that global dependencies are never expressed through empty categories.

Propagation constraints and coherence constraints

Fine-grained grammatical information is treated only in propagation and coherence constraints.

---

[29] Adjunction is needed only if a constituent is intrinsically an adjunct (e.g. a clitic, a sentential Comp, etc.) or is left unattached by decomposition (e.g. a non selected PP, an appositive NP, etc.). Coordination is needed only if a conjunction is detected (some constituent must also have been left unattached, since this is a special case of adjunction).

Propagation constraints apply to local trees. They guarantee that some types of information are propagated from daughters to mother (and vice-versa) in a local tree (but never between siblings). Given type abstraction over objects and grammatical information, they have the following form:

$\kappa(\delta(\tau)) \quad = \quad \kappa(\delta'(\tau))$

where τ is a local tree,

δ, δ' are abstract types of nodes (e.g. MOTHER, DAUGHTER, FILLER-DAUGHTER, etc.) and

κ is an abstract category type (a path in more traditional terminology).[30]

For example:

$H(M(\tau)) \quad = \quad H(HD(\tau))$

The head (H) value of the mother (M) is identical with the head value of the head (HD).

$MINOR(M(\tau)) \quad \geq \quad MINOR(LEXD(\tau))$

The MINOR value of the mother is an extension of the MINOR value of the lexical daughters (LEXD).

Coherence constraints, on the other hand, guarantee that every node in the final product is coherently labeled. *Coherence constraints* correspond more or less to FCRs in GPSG. They have the following form.

$\alpha \quad \supset \quad \beta$

where α and β are elementary constraints on categories (disjunctive and negative combinations are excluded, but conjunctive and doubly implicative combinations are allowed, since they are deterministic).

For example:[31]

[LEVEL : phrasal] $\supset$

---

[30] In our approach, paths are invisible. They are named by abstract types. So, changes in the representation do not affect access to proper values.

[31] These rules are part of our description of quantified NP's in Chinese (Morin & Ren, 1992).

([CLASS : α] ⇔ [QUANT : β])

A (phrasal) object is classified if and only if it is quantified.

It should be noted that *propagation constraints* and *coherence constraints* not need not take into account the origin of the relevant grammatical information.[32] Constraints can thus be applied blindly and locally again reducing non determinism. Furthermore, coherence constraints never instantiate anything, they just check their input and filter it out if they are not satisfied (unless we allow constraint relaxation). There is no free instanciation, any value appearing in a structure is entirely constrained, either by lexical or by grammatical constraints.[33]

Conclusion

In this paper, we have discussed some notions of complexity and some sources of effective complexity in natural language processing. We tried to show that, once certain hypotheses are adopted, sources of determinism in natural languages become apparent, and it is possible to use these results to partition the space of grammatical and lexical constraints in such a way as to guarantee efficient parsing.

References

Aho, A. & J. Ullman (1972-1973) *The Theory of Parsing, Translation and Compiling, vol.1: Parsing; vol. 2: Compiling* . Englewood-Cliff, NJ: Prentice-Hall.

Barton, E. (1985) "On the complexity of ID/LP parsing," *Computational Linguistics*, 11, 4: 205-218.

Barton, E. (1986) "Computational complexity in two-level morphology," *ACL-24*: 53-59.

Barton, E. et al. (1987) *Computational Complexity and Natural Language*, Cambridge, Mass.: MIT Press.

Berwick, R. (1982) "Computational complexity and lexical-functional grammar," *AJCL*, 8, 3-4: 97-109.

Berwick, R. (1984) "Strong generative capacity, weak generative capacity, and modern linguistic theories," *Computational Linguistics*, 10: 189-202.

---

[32] This is the *chaptalization* hypothesis (Morin & Ren, 1992).

[33] All the problems related to the complexity of free instanciation are thus eliminated a priori. Cf. Ristad (1986a, b, c, d, 1990b), Barton et al. (1987, ch. 8) and Jutras (1990) for different analyses of the complexity of free instanciation and related linguistic and computational problems.

Boekee, D. E. et al. (1982) "On complexity and syntactic information," IEEE Transactions SMC-12: 71-79.

Blache, Ph. (1990) *L'analyse syntaxique dans le cadre des grammaires syntagmatiques généralisées: Interprétation et stratégies*, doctoral dissertation, Université d'Aix-Marseille II.

Blache, Ph. & J.Y. Morin (1990) "Bottom-up filtering, a parsing strategy for GPSG," *COLING-90*, 2: 19-23.

Bresnan, J. (1983 ed.) *The Mental Representation of Grammatical Relations*, Cambridge: MIT Press.

Carpenter R. (1992) *The Logic of Typed Feature Structures*, Cambridge University Press.

Chaitin, G. J. (1987) *Algorithmic Information Theory*, Cambridge: Cambridge University Press.

Gan, Kok Wee (1994) *Integrating Word Boundary Disambiguation with Sentence Understanding*, Ph. D. dissertation, National University of Singapore.

Haas, Andrew (1989) "A parsing algorithm for unification grammar," *Computational Linguistics*, 15, 4: 219-232.

Jutras, J.-M. (1990) *L'instanciation en grammaire syntagmatique généralisée*, M.A. thesis, Université de Montréal.

Kaplan, R. & J. Bresnan (1983) "Lexical-functional grammar: A formal system for grammatical representation," *in* Bresnan (1983 ed.).

Kasper, R. (1987) *Feature Structures: A Logical Theory with Applications to Language Analysis*, Ph. D. dissertation, Univ. of Michigan.

Kasper, R. & W. Rounds (1986) "A logical semantics for feature structures," *ACL-24*: 257-266.

Kolmogorov, A. (1965) "Three approaches to the quantitative definition of information," *Problems of Information Transmission*, 1: 1-7 (translated from Russian).

Kornai, A. (1994) "The generative power of feature geometry," *Annals of Mathematics and Artificial Intelligence*.

Kornai, A. & G. Pullum (1990) "The X-bar theory of syntax", *Language*, 66, 1.

Koskenniemi, K. (1983) *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*, doctoral dissertation, University of Helsinki.

Koskenniemi, K. & K. Church (1988) "Complexity, two-level morphology and Finnish", *COLING-88*.

Li, Ming & Paul Vitányi (1993) *An Introduction to Kolmogorov Complexity and its Applications*, Berlin, New-York: Springer-Verlag.

Manaster-Ramer, A. (1994) "Review of Ristad, E. S. (1993) *The Language Complexity Game*," *Computational Linguistics*, 21, 1: 124-131.

Marr, D. (1980) *Vision*, San Francisco: W.H. Freeman.

Maxwell, J. & R. Kaplan (1993) "The interface between phrasal and functional constraints," *Computational Linguistics*, 19, 4: 571-590.

McCarthy, J. (1981) A prosodic theory of nonconcatenative morphology, *Linguistic Inquiry*, 12: 373-418.

McCarthy, J. & A. Prince (1990) Foot and word in prosodic morphology,: the Arabic broken plural, *Natural Language and Linguistic Theory*, 8: 209-283.

Mel'čuk, I. et al. (1984-...) *Dictionnaire explicatif et combinatoire du français contemporain*, Montréal, Paris : Presses de l'Université de Montréal and Presses du CNRS, 3 volumes published.

Morin, J.-Y. (1985) "Théorie syntaxique et théorie du parsage: quelques réflexions," *Revue québécoise de linguistique*, 14, 2: 1-40.

Morin, J.-Y. (1989) "Particules et parsage universel," *in* Weydt, H. (1989 ed.) *Sprechen mit Partikeln.*, Berlin: De Gruyter, pp. 713-728.

Morin, J.-Y. & X. Ren (1992) "Classifier Features in Chinese: A GPSG Approach", *ICCL-1*.

Nagata, M. (1991) "An empirical study on rule granularity and unification interleaving, toward an efficient unification-based parsing system," *COLING-92*: 177-183.

Perrault, R. (1984) "On the mathematical properties of linguistic theories," *Computational Linguistics*, 10: 165-176.

Pinker, S. (1984) *Language Learnability and Language Development*, Cambridge: Harvard University Press.

Pollard, C.J. & I. Sag (1994) *Head-Driven Phrase Structure Grammar*, Chicago: University of Chicago Press.

Ristad, E. S. (1986a) "Computational complexity of current GPSG theory, *ACL-24*: 30-39.

Ristad, E. S. (1986b) "Defining natural language grammars in GPSG, *ACL-24*: 40-44.

Ristad, E. S. (1986c) "Sources of complexity in GPSG theory, *Theoretical Linguistics*, 13, 1-2: 105-124.

Ristad, E. S. (1986d) *Complexity of linguistic models: a computational analysis and reconstruction of generalized phrase structure grammar*. S.M. thesis, MIT.

Ristad, E. S. (1990a) "Computational structure of generative phonology and its relation to language comprehension," *ACL-28*: 235-242.

Ristad, E. S. (1990b) "Computational structure of GPSG models," *Linguistics and Philosophy*, 13, 5: 523-590.

Ristad, E. S. (1994) "Complexity of morpheme acquisition," *in* Ristad (1994 ed.: 185-198).

Ristad, E. S. (1993) *The Language Complexity Game*, Cambridge, Mass.: MIT Press.

Ristad, E. S. (1994 ed.) *Language Computations*, DIMACS, vol. 17, American Mathematical Society..

Ristad, E. S. & R. Berwick (1989) "Computational consequences of agreement and ambiguity in natural language," *Journal of Mathematical Psychology*, 33: 379-396.

Rounds, W. (1973) "A grammatical characterization of the exponential time languages," *Proceedings of the 11th Annual Symposium on Foundations of Computer Science*. New-York: IEEE Computer Society, p. 135-143.

Rounds, W. (1987) "Review of Barton, E., R. Berwick et E. S. Ristad (1987),*Computational Complexity and Natural Language*," *Computational Linguistics*, 13, 3-4: 354-356.

Rounds, W. (1988) "LFP: A logic for linguistic descriptions and an analysis of its complexity," *Computational Linguistics*, 14, 4: 1-9.

Rounds, W. (1991) "The relevance of computational complexity theory to natural language processing," *in* Sells, P., S. M. Shieber & T. Wasow (1991 ed.: 9-29).

Rounds, W. et al. (1986) "Finding natural languages a home in formal language theory, *in* Manaster-Ramer (1986 ed.) *Mathematics of Language*. New-York: John Benjamins.

Sabot, G. (1988) *The Paralation Model, Architecture-Independent Parallel Programming*, Cambridge: MIT Press.

Sells, P. et al. (1991 ed.) *Foundational Issues in Natural Language Processing*, Cambridge: MIT Press.

Shieber, S. M. (1983) "Direct parsing of ID/LP grammars," *Linguistics and Philosophy*, 7, 2: 135-154.

Shieber, S. M. (1985) "Using restriction to extend parsing algorithms for complex feature-based formalisms," *ACL-23*: 145-152.

Shieber, S. M. (1992) *Constraint-based grammar formalisms*, Cambridge, Mass.: MIT Press.

Solomonoff, R. (1964) "A formal theory of inductive inference," *Information and Control*, 7: 1-22 et 224-254.

Watanabe, O. (1992 ed.) *Kolmogorov Complexity and Computational Complexity*, Berlin, New-York: Springer-Verlag.

A UNIFYING APPROACH TO SEGMENTATION OF CHINESE AND ITS APPLICATION TO TEXT RETRIEVAL

Jian-Yun Nie [1]
Xiaobo Ren [2]
Martin Brisebois [1]

[1] University of Montreal,
BP.6128, succ. Centre-ville, Montreal, Quebec, H3C 3J7 Canada
[2] Center for Information Technology Innovation
1575 Bd. Chomedey, Laval, Quebec, H7V 2X2 Canada

In segmentation of Chinese, two competing approaches have been often used separately: the rule-based approach and the statistical approach. Each approach has its advantages and disadvantages. In this paper we describe a hybrid approach which unifies them in a single flexible segmentation process in which items stored in the dictionary or identified by heuristic rules are assigned a default probability. By varying the default probability value, the hybrid approach can cover a wide range of approaches from the purely statistical one to the purely rule-based one. Our experiments on two corpora show that by a proper setting of the default probability, the hybrid approach gives much better results than statistical or rule-based approaches alone. A text retrieval system is then adapted to the segmented Chinese texts. Preliminary results of the retrieval system are reported.

1. Introduction

Natural language processing is an important issue in many areas such as Information Retrieval (IR). IR systems for Indo-European languages are widely used in libraries, information centers and increasingly across the information web in computer networks. An IR system aims to select the texts from a corpus which are relevant to a given query [1]. Typically, a system determines the relevant documents according to the frequency of occurrences of the *words* of the query within the documents and the corpus. In Indo-European languages, the identification of words is a trivial task, but in Chinese, it is difficult because there is no separation between words in Chinese texts. Thus traditional approaches for IR cannot be directly applied to Chinese.

One might think that, as there is no available separation of words in Chinese, text retrieval can operate on a character string basis. This approach has been used in some experimental systems for Japanese text retrieval [2, 3] for which the same problem is encountered as for Chinese texts. However, this approach would lead to a great deal of

incorrect matching between queries and documents due to the almost free combination of characters in sentences. To take an example, if one wants to retrieve documents about 识别 (recognition), then it is possible to find a document containing the sentence 他认识别的人 (he knows other people) by the character-based approach. In addition, character-based retrieval would lead to an explosion of index file size due to the great number of character combinations as searching keys.

We believe that Chinese text retrieval should operate on segmented texts in order to gain efficiency and quality in the retrieval operation. Moreover, this approach can benefit much from the development of information retrieval for Indo-European languages.

The process of segmentation has been the subject of much intensive research in the area of computer-based analysis of Chinese for the past decade. These approaches may be classified into two main groups: the rule- and dictionary-based approach and the statistical approach. Approaches in the first group rely on knowledge defined by human experts (dictionary and heuristic rules) in segmentation. These approaches only make use of general knowledge on Chinese words: the words included in the dictionary are often the most usual ones, and the heuristic rules correspond to common word structures. On the other hand, approaches of the second group use specific statistical information about the corpus or application area. These two approaches have often been used separately in automatic segmentation processes, except in a few ones such as [4]. This does not correspond to the human segmentation process in which both general knowledge and specific information are used.

In this paper, we describe a hybrid approach for segmentation of Chinese which uses dictionary, heuristic morphological rules and statistical information in a single process. The basic idea is to consider general knowledge as background knowledge, and to place specific statistical information in front of it. This idea is achieved simply by assigning a default probability to items stored in the dictionary or identified by the heuristic rules.

This approach has a high flexibility: By varying the default probability value, the hybrid approach can cover a wide range of approaches from the purely statistical approach to the purely rule-based approach.

We tested our approach with two corpora. We have shown that for both corpora, the hybrid approach yields better results than the two competing approaches alone. We further adapted a general IR system, SMART, to our segmented Chinese texts. The performance of the IR system for Chinese is evaluated with respect to different segmentation approaches. It is shown that the segmentation quality has a great impact on the retrieval quality.

2. Statistical approach vs. Rule- and dictionary-based approach

Dictionary-based approaches [5-14] operate according to a very simple concept: a correct segmentation result should consist of legitimate words (in a restrictive sense, those in a dictionary). In general, however, several legitimate word sequences may be obtained from a Chinese sentence. The maximum-matching (or longest matching) algorithm is often used

then to select the word sequence which contains the longest (or equivalently, the fewest) words. This algorithm may be described as follows:

An input character string is compared with the contents of the dictionary so that all sequences of characters constituting recognized lexical items can be highlighted. Words are linked from beginning to end of the input string, with several candidate word chains being proposed. Among all possible word chains, the one with the fewest and thus the longest words is considered to be the best segmentation.

The above approach is often extended by a set of heuristic morphological rules [7]: a character string which is not stored in the dictionary, but may be derived from the rules, is also a possible word candidate. Typically, heuristic rules are set for identifying words having some common structures such as affix structure (大众化 - popularize) or nominal pre-determiner structure (一百个人 - hundred people).

Rule- and dictionary-based approaches have the advantage of being simple, general and often efficient: The heuristic knowledge built into the system corresponds closely to knowledge about linguistic phenomena occurring in Chinese words and this knowledge is represented in a straightforward way, allowing human experts to verify its correctness. It has been shown that a simple rule-based approach may often achieve a performance comparable to that of a sophisticated statistical approach.

However, a prerequisite for high-quality results in rule- and dictionary-based segmentation is a dictionary which is *complete*. It is unrealistic to suppose that a truly complete Chinese dictionary will be available because of the enormous *size* such a potential dictionary would imply, its *domain dependency* (certain strings may be words in some domains while not in others), and the fact that new words are constantly being produced (the *creative* aspect of language).

Although the maximum-matching algorithm may solve the major part of segmentation ambiguity, several possible segmentation results may still remain because they have equal lengths. To solve the remaining ambiguity, it has often been suggested that syntactic, semantic, or even pragmatic analysis should be used [6]. In practice, however, we do not have enough knowledge for the last two analyses to be feasible at the present time. Even for the syntactic analysis, although one succeeded in analyzing the core part of Chinese syntax [15], it still seems to lack of syntactic rules in Chinese that have a good coverage and are as rigorous as in Indo-European languages. In IR context, especially, as texts may be written in different styles and concern various areas, this solution is difficult to materialize now. Instead of using sophisticated linguistic analyses, we suggest to use statistical information as an alternative solution.

Statistical approaches [16-21] do not need pre-established dictionary and rules. They rely on statistical information such as word and character (co-)occurrence frequencies in the text which may be obtained automatically from training data set manually. One of the advantages of statistical approaches is their capacity to cope with the particularities of

application areas through the statistical information. The simplest statistical approach is as follows:

Given a manually segmented training document set, the probability of a character string S to be a word is calculated as follows:

$$p(S) = \frac{\text{number of occurrences of } S \text{ being segmented as a word in the training set}}{\text{number of occurrences of } S \text{ in the training set}}$$

Given an input string to be segmented, the best solution is composed of a sequence of potential words S_i such that $\Pi_i p(S_i)$ is the highest.

Many statistical approaches make use of more complex, typically first-order Markov, models. Although statistical approaches avoid the tedious task of establishing a dictionary and heuristic rules, they require a great deal of manually segmented texts to train the model. The training data are also difficult to set up (often not much easier than setting up a dictionary). Moreover, inconsistency is often unavoidable and difficult to check in manual segmentation, affecting the reliability of the statistical information obtained. In addition, the acquisition of statistical information is not cumulative. Probabilities need to be revised constantly. From this point of view, there is no clear advantage for statistical approaches on data preparation.

Through the above analysis, we can see that rule- and dictionary based approaches and statistical approaches have quite complementary properties: The former is general but application-insensitive; the latter is specific but the statistical information cannot be generalized. It is natural then to suggest a hybrid approach which combines them in a single approach in order to compensate the drawbacks of each approach with the advantages of the other.

Hybrid approaches have been used by a few researchers. Fan and Tsai [4], for example, describe a statistical approach which incorporates a dictionary. The probability of a dictionary entry is first assumed to be 1, then revised by a relaxation process using statistical information. However, the relaxation process will not apply to the words on which there is no statistical information, that is, the relaxation process may have a poor coverage. If uncovered words appear, the segmentation accuracy may be seriously affected. In our hybrid approach, a default probability much lower than 1 is assigned manually to all the lexical items in the dictionary. We do not have the problem of poor coverage. If statistical information is also available, it can be integrated readily with human established dictionary.

3. A hybrid segmentation approach

From a cognitive point of view, statistical data provide a sort of short term knowledge about the application context, whereas the vocabulary stored in a dictionary may be seen as long term knowledge generally accepted by people. When people segment Chinese texts, both

types of knowledge are used: Usually, a correct segmentation may be determined unambiguously by cutting the sentence into usual legitimate words. In some circumstances, however, unusual words or new words may be used. In this case, people usually look into the context (or application area) in order to determine whether an unusual or new string may be a word. Although in human examination of context, syntactic, semantic and pragmatic analyses may be appealed, statistical information about the utilization of words (in the same area) also provide useful indication. This latter information can be incorporated into a computer-based analysis. Our hybrid segmentation process works in a similar way:

A dictionary is used as a repository of background knowledge. Each entry in the dictionary is assigned a default probability. If statistical data are available, we can also establish a statistical dictionary which consists of a set of potential words together with their probability to be valid words in the given corpus. The two dictionaries can then be merged together in a statistical segmentation process such that both kinds of information are used.

Merging dictionary with statistic information

Although statistical approaches and rule-based approaches have often been seen as two competing ones, they are indeed compatible. In fact, a rule-based approach using longest-matching algorithm may also be seen as a special case of statistical segmentation: each potential word in the input string which is stored in the dictionary or derived from a heuristic rule, is assigned an equal probability (less than 1). Then the maximum-matching algorithm is equivalent to a statistical approach which chooses the segmentation result of the highest probability. For example, for the phrase 中国文学 (Chinese literature), there may be the following segmentation possibilities according to the dictionary:

中国 文学
中国 文 学
中 国文 学
中 国 文学
中 国 文 学

If each potential word is assigned equal probability values p (<1), then the first segmentation which contains the fewest words will have the highest probability p^2. The other possible results will have lower probabilities (p^3 or p^4). This result is the same as with the maximum-matching algorithm.

The rule- and dictionary-based approach being seen as a special statistical approach, it is then possible to combine them in a single hybrid segmentation process. In such a hybrid approach, if statistical information about a dictionary item is available, it is used in priority; otherwise, the default probability is assigned to that item. By varying the default probability value, we can change the relative importance of the statistical information and the dictionary. When the default probability value is set to 0, the hybrid approach will not take into account the words stored in the dictionary. Consequently, the hybrid approach becomes a purely statistical approach. On the other hand, when the default probability value is very

high (near 1, but <1), the hybrid approach will consider almost exclusively the words stored in the dictionary. Thus we obtain the rule-based approach in this case. We see that the hybrid approach can cover a wide range of approaches from the purely statistical approach to the purely rule-based approach, as illustrated by the following figure:

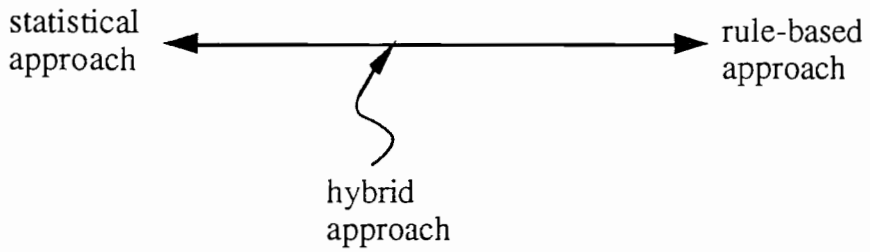


Figure 1. Comparison of the three approaches

In rule-based approaches, if a character is not grouped with its neighboring characters, that individual character is usually considered to be a word. In fact, a single character has much less chance to be a word than a compound string included in a dictionary, as noted by Bai [22]. Bai labels a single character not in the dictionary as a "semi-word" in order to distinguish it from a word in the dictionary. In his approach, the latter is used in preference to the former. In our approach, we apply the same principle: a single character is assigned the probability $p/2$ where p is the default probability assigned to dictionary items.

Heuristic rules

Apart from the dictionary, a set of heuristic rules is also incorporated into our segmentation process in order to identify and segment words which follow some rules (for example, numbers and dates). In this paper we only deal with the following two groups of morphological rules. More discussion about heuristic rules may be found in [7].

Nominal pre-determiner structure

Words corresponding to this structure frequently occur in Chinese, for example, 每一周 (*every week*), 这一回 (*this time*). In order to establish a set of heuristic rules for this structure, we first define the following categories of single characters:

- determiners: 这 (*this*), 那 (*that*) 此 (*this*) 该 (*this*) 其 (*its, his, her*)
  每 (*each*) 各 (*every*) 某 (*some*) 首 (*first*) ...
- ordinal-number markers: 第 (*number*)
- cardinal numbers: 零 (*zero*) 一 (*one*) 壹 (*one*) 二 (*two*) 贰(*two*)
  十(*ten*) 百(*hundred*) 半(*half*) ...
- classifiers: 班(*class*) 帮(*band*) 包(*bag*) 杯(*cup*) 辈(*generation*) 本(*book*)
  组(*group*) 次(*time*) 层 (*layer*) 年 (*year*) 月(*month*) 日(*day*)...

The following rules cover a major part of the words in this structure (where [...] indicates optional status and [...]* an optional arbitrary repetition):

ordinal cardinal [classifier] → pre-det	第一周 (*first week*) 第二 (*second*)
determiner [cardinal]* classifier → pre-det	这一回 (*this time*) 每层 (*every layer*)
cardinal [classifier] → pre-det	十一(*eleven*) 一九九一年(*in 1991*)
	一百本(*hundred books*)

Apart from these general rules, some special cases are also considered. For example, some determiners (各, 首) cannot be followed by an ordinal as in 首一次, 各一组, but can be followed by a classifier suck as 首次 (first time), 各组 (each group).

Affix structure

In our segmentation, for a word to be considered as having an internal affix structure, both of the following conditions should be true:

1. The first (last) character should be a possible prefix (suffix). For example:
   prefix: 大(big) 小(small) 总(general) 副(vice) ...
   suffix: 人(person) 们(plural mark) 权(right) 会(association) 化(-ize/-ization), ...
2. The remaining characters should form a known word.

Most internal affix structures fit these conditions. However, the second condition is not always true. For example the string 副总经理 (*vice general manager*) cannot be identified to be a single word having a duplicated prefix structure 副 + 总 + 经理 due to the second condition. The setting of this condition is to prevent classifying some strings incorrectly as a word such as for the string 保护人权 (*protect human rights*). Without the second condition, this string may be identified as a single word formed from the word 保护 by adding two successive suffixes 人 and 权, which may mean "the rights of protectors". This latter case occurs more frequently in our corpora than the former. Thus we keep the condition for a practical reason. However, we are aware that this condition should be replaced later by other more refined conditions.

4. Implementation

In order to give a more thorough view of our system, we describe several implementation details in this section.

Dictionary organization

Both manual dictionary and statistical information are stored in a run-time dictionary. In order to increase efficiency in dictionary look-up, this dictionary is organized as an open

hash table. The first Chinese character C_1 (2 bytes) of a word is used to calculate a unique location Hash(C_1) in the hash table. Each location in the hash table points to a list of words starting by the character. The following figure shows a fragment of the run-time dictionary (where i is the hash address for 人 and i+1 for 忍, i.e. Hash(人) = i and Hash(忍) = i+1; and the real number are probabilities of the words):

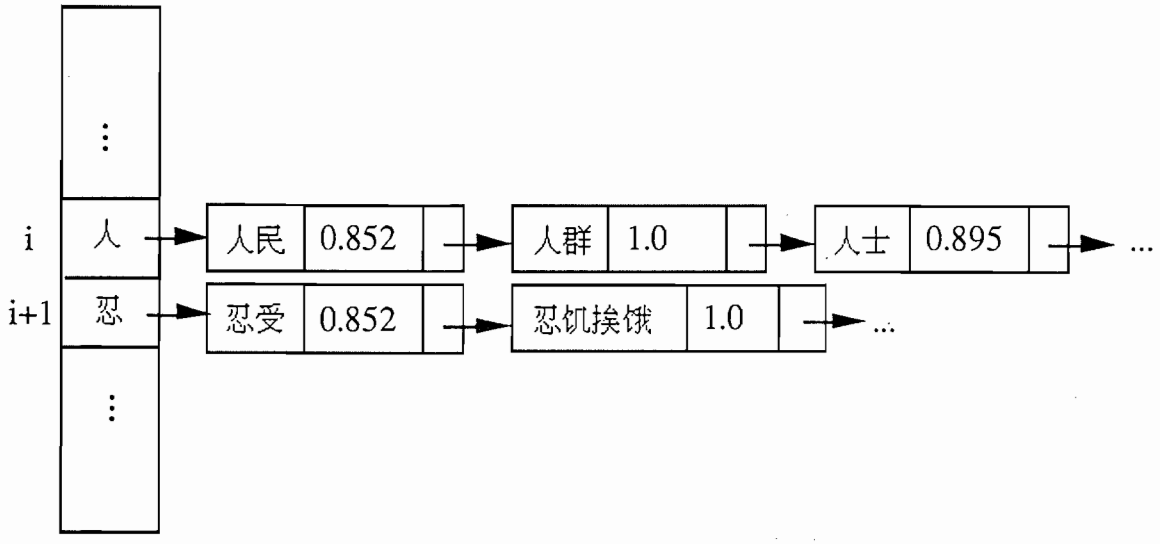


Figure 2. Organization of the run-time dictionary

Our manual dictionary contains over 91 000 entries. A few thousand new words are identified in the statistical information. These new words are mainly names of non-Chinese people and countries, or words that can be identified by heuristic rules and are not included in the manual dictionary.

The organization of the segmentation process

The segmentation process is similar to a purely statistical approach. Given an input string to be segmented, the following two main sub-processes are applied to it:

1. Dictionary look-up:
   This sub-process associates to each character in the input string a list of the candidate words, together with their probability, which are substrings of the input string starting from this position.
2. Find the best combination of the candidate words:
   This sub-process combines the word candidates to cover the entire input string and chooses those combinations that have the highest probability.

The first sub-process is quite straightforward. The complexity of the algorithm is mainly determined by the combination procedure. The following recursive algorithm is used:

Procedure best-combine($C_1...C_iC_{i+1}...C_n$);

/* $C_1...C_iC_{i+1}...C_n$ is the input string */

1. For each word candidate $C_1...C_i$ at the beginning:

   a) find the set of the best combinations for the remaining string $C_{i+1}...C_n$:

      R := best-combine($C_{i+1}...C_n$),

   b) for each S_k in R:

      - combine the word candidate $C_1...C_i$ with S_k,
      - assign the probability $p(C_1...C_i)*p(S_k)$ to the segmentation starting by the word $C_1...C_i$ followed by S_k;

2. Return the set of combinations covering the string that have the highest probability together with that probability.

We give some examples to illustrate the segmentation process. These examples show the actual process of the hybrid segmentation with the default probability set to 0.001.

Example 1: 大会决议和议程项目

1. After the dictionary look-up, the following word candidates, together with their probability, are associated to each character in the string:

大: 大会 1.000000, 大 0.016073
会: 会 0.029028
决: 决议 0.955782, 决 0.001081
议: 议和 0.001000, 议 0.000500
和: 和议 0.001000, 和 0.944933
议; 议程 1.000000, 议 0.000500
程: 程 0.001000
项: 项目 0.936073, 项 0.023973
目: 目 0.000500

2. The combination procedure is applied recursively to the input string such that word sequences are built from end to beginning. For the substring 目, there is only one possibility with probability = 0.005. For the substring 项目, two combinations are possible:

项目 0.936073
项 目 0.000013

Only the best one (项目) is used for further combination with characters before it. So for the substring 程项目, the only retained combination is 程 项目. For the substring 议程项目, we have again two possibilities, but only 议程 项目 will be retained. This combination process is to be applied until the first character of the input string has been combined. Finally, the following correct segmentation is chosen as the best result:

大会 决议 和 议程 项目

which is of the highest probability.

We notice that although there are several combinations for the substring 决议和议程 (决议 和 议程, 决 议和 议程, 决议 和议 程) that would all remain as possible solutions in a rule-based approach, our hybrid segmentation is able to determine the correct one using the statistical information: 决议 和 议程. This example shows the contribution of statistical information.

Example 2: 1993年8月17日第47/233号决议

In our implementation, special attention has been paid to the determination of complex pre-determiner strings that contain Chinese and ASCII characters as in this example. A string of ASCII numbers (and some other kinds of special strings) is considered as an inseparable token.

After the dictionary look-up, the following word candidates are associated:

1993: 1993年 0.001000, 1993 0.001000
年: 年 0.683775
8: 8月 0.001000, 8 0.001000
月: 月 0.935073
17: 17日 0.001000, 17 0.001000
日: 日 0.767800
第: 第47/233号 0.001000, 第 0.533917
47/233: 47/233号 0.001000, 47/233 0.001000
号: 号 0.869823
决: 决议 0.955782, 决 0.001081
议: 议 0.000500

The candidate words 1993年, 8月, 17日, 第47/233号, 47/233号 are all identified as pre-determiner structures. They are assigned the default probability.

Finally, the selected result is the following:

1993年 8月 17日 第47/233号 决议

5. Experiments

We tested our hybrid approach on two corpora, both from the United Nations. We segmented both corpora manually. Automatic segmentation results are compared with the manual one to evaluate their accuracy. Each corpus is split into a training set and a test set. The training set has been used to calculate the probability for potential words (see section 2 for the calculation). The characteristics of the corpora are highlighted in the following table:

Corpora	Size (Kbyte)	training set	test set
Corpus 1	164	149	15
Corpus 2	1 270	1 247	272

Table 1. Characteristics of the corpora

Different default probability values have been used in the hybrid segmentation. The following table shows the number of errors using the hybrid approach to segment the training set and the test set of corpus 1 (similar observations have been obtained on Corpus 2):

default probability p for items from manual dictionary	No. of errors in segmenting training set (34433 words)	No. of errors in segmenting test set (3487 words)
0	52	1346
0.00001	50	272
0.0005	50	105
0.001	50	104
0.005	62	103
0.01	73	101
0.02	106	109
0.05	152	105
0.1	196	103
0.2	292	99
0.3	381	112
0.4	479	133
0.5	552	142
0.9999	3405	324

Table 2. Influence of the default probability in the hybrid segmentation

We can see in this table that both competing approaches alone do not yield satisfactory results, either for the training set or for the test set. In the case of the pure statistical segmentation (p=0), the segmentation of the training set is very good. This is reasonable because the approach is trained by the same data. When the approach is applied to the test data, however, we observed a high ratio of error (38.6%). This is mainly because the training data do not completely cover the test set. This observation is consistent with the remark we made earlier that for a statistical approach to yield good results, it is essential that the training data has a good coverage of the application area.

On the other hand, in the case of the purely dictionary- and rule-based approach ($p\approx1$), the error ratio is almost the same for the training data and test data. The segmentation accuracy is around 90%. In comparison with the other reports of near 99% of accuracy using such an approach, we note that in our corpora, there are quite a number of names of non-Chinese people or countries. Our current segmentation does not incorporate rules for the detection of such names. This subject has been investigated in some other studies, for example [23].

In the case of truly hybrid approach (when the default probability is between 0 and 1 exclusively), better results are obtained. The best results correspond to the setting of the default probability between 0.001 and 0.1. In some cases (between 0.00001 and 0.001), we even observed a performance on the training data better than the purely statistical approach.

The following graph shows the variation of segmentation accuracy of the hybrid approach on the test data in both corpora. The default probability value varies from 0 to near 1. We can draw the conclusion that the hybrid approach is significantly better than the two competing approaches alone. When the default probability is set between 0.001 and 0.1, we obtain the best results for both corpora about 97% accurate).

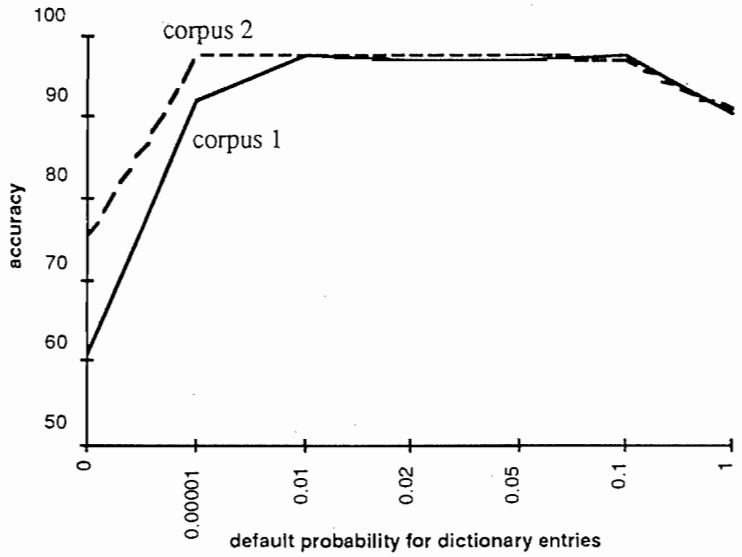


Figure 3. Segmentation accuracy for different approaches

6. Application to Text Retrieval

The problem of Chinese text retrieval has been investigated in [24, 25]. However, These studies mainly concerns the segmentation of Chinese texts rather than their retrieval. To our knowledge, there is no general IR system built for Chinese texts until now.

In this study, we try to build a complete IR system for Chinese texts. Note that when Chinese texts have been segmented, traditional IR approaches may be adapted to their retrieval. This is the approach we took: we adapted the SMART [26] system in our implementation. SMART is a text retrieval system developed in Cornell University. This system compasses a variety of tools for text tokenizing, word statistic measuring, and query evaluation.

Implementation

The application of SMART to index and retrieve segmented Chinese texts may seem to be easy and direct. However, as SMART is designed for English texts, it does not deal

with non-ASCII characters such as Chinese characters. To adapt it to Chinese texts, two solutions are possible:

1. extend the character set considered by SMART to cover non-ASCII characters;
2. encode Chinese texts by ASCII characters.

In our current implementation, the second solution is used. Chinese characters are encoded in HZ format in which each Chinese character is encoded by two ASCII characters. A Chinese character string is delimited within ~{ and ~} in order to make difference from ordinary (non quoted) ASCII characters. For example, the following string

SMART 信息检索系统

is encoded in HZ as the ASCII string: SMART ~{PEO"<lKwO5M3~}.

The problem with the encoded HZ texts is that Chinese characters are often encoded by symbols such as punctuation markers (?, !, . %, ...). As SMART checks for tokens according to English writing, Chinese characters are often incorrectly cut in the direct application. To solve this problem, we modified the SMART tokenizing program in order to deactivate the original tokenizing process and replace it with a new one which keeps the delimited Chinese codes together.

In indexing, SMART ignores the words which are considered as common-words. A list, called stop-list, of such words is set up for English. We enhanced the English stop-list by about 300 common Chinese words. These words are often adverbs and prepositions that are not important for IR purposes. We also included in the stop-list the Chinese symbols such as punctuation markers. Here are some items included in the stop-list:

按照, 把, 被, 比, 比较, 并, 并且, 不论, 不能, 才, 常, 除非, 此外.

The indexing process of SMART may now be applied to the Chinese texts (documents and queries) in order to extract important keywords from them.

Experiments

The adapted retrieval system has been verified by using the test set of Corpus 2. The test data are composed of 797 relatively independent paragraphs. We consider each paragraph as an independent document in our experiments. A set of 10 queries in Chinese in the domain of the these documents has been set up and manually evaluated by examining through the documents. The query evaluation of the system is compared with the manual evaluation in order to evaluate the system's performance in terms of precision and recall defined as follows:

$$\text{recall} = \frac{\text{the number of relevant document retrieved}}{\text{the number of relevant documents in the corpus}}$$

$$\text{precision} = \frac{\text{the number of relevant document retrieved}}{\text{the number of document retrieved}}$$

We applied the modified SMART system to the results of three different segmentation process: the purely statistical approach, the purely rule-based approach and the hybrid approach with default probability = 0.001. For document indexing, we used *tf*idf* scheme for keyword weighting [1]. Queries are evaluated using a simple Boolean retrieval method.

The following figure shows the variation of the precision ratio over the recall ratio for the three segmentation approaches.

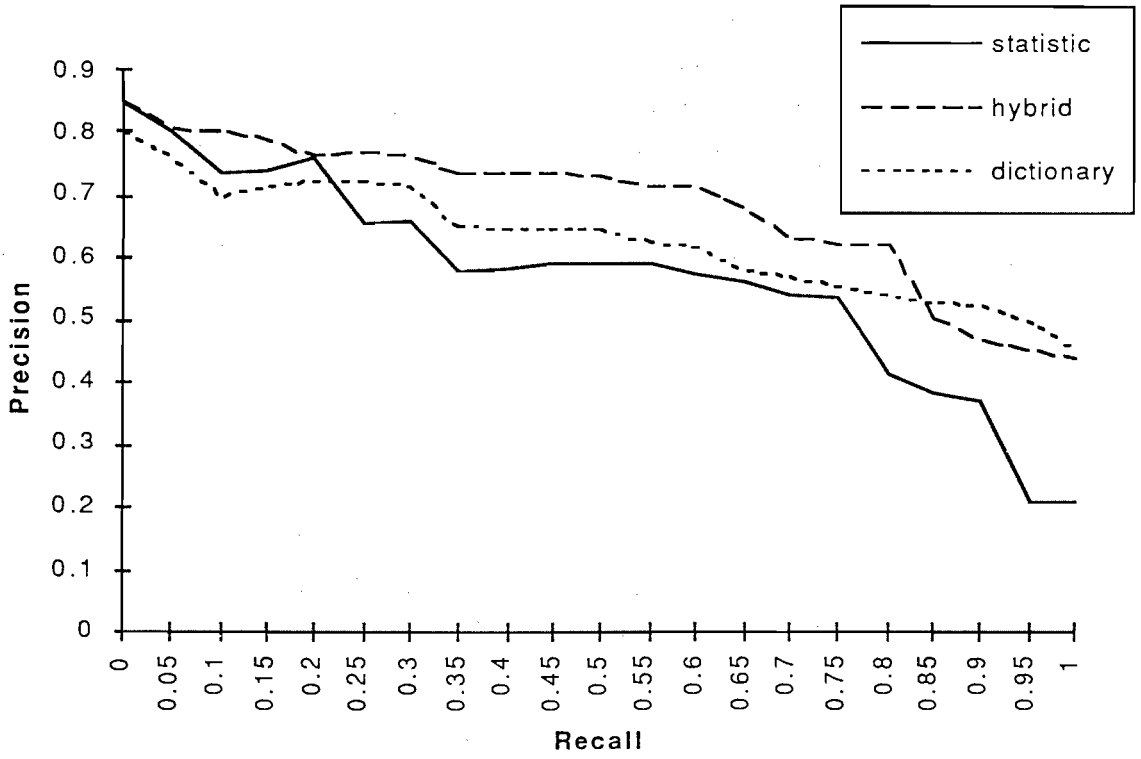


Figure 4. Evaluation of the retrieval performance

It can be seen that the hybrid segmentation leads to the best retrieval performance. This may be seen more clearly in the following table in which we give the *average precision* of the retrieval with respect to the three segmentation processes. The average precision is the common measure used for IR systems which is the average of precision ratios when the recall ratio = 0%, 5%, 10%, 15%, 20%, ..., 100% respectively. The following table shows the comparison of the system's performance with respect to the segmentation processes.

Segmentation approach	Average precision
statistical	56.79
hybrid (p = 0.001)	68.24
rule/dictionary-based	62.86

Table 3. Retrieval performance

We can compare this table with Figure 3 and see that the retrieval performance is strongly consistent with that of the segmentation. The same ranking is maintained for both segmentation and retrieval: the hybrid approach, the rule- and dictionary-based approach, and finally the statistic approach. This leads to the conclusion that Chinese texts should be segmented with a high quality segmentation process if one expects a high retrieval performance.

7. Future work

In this paper, we described a hybrid segmentation approach which makes use of both human-defined knowledge and statistical information. In comparison with other segmentation approaches, this approach is marked by its high flexibility: it can cover both the statistical approach and the rule-based approach by varying the default probability assigned to manually established lexical items. The hybrid framework allows us to see that statistical information and man-defined lexical knowledge represent two extreme cases in segmentation, but they are not incompatible, thus can be combine in a single process.

We also tried to adapt a general information retrieval system, SMART, to retrieve segmented Chinese texts. Our adaptation shows the feasibility of using IR systems designed for Indo-European languages to Chinese.

As one of the subjects for our future work, we plan to enhance our segmentation process by incorporating more heuristic rules, in particular, for dealing with proper names. In a previous work, we investigated this subject [27] but it has not been integrated into the present implementation.

On Chinese text retrieval, there is a lot to be done. In an attempt to obtain better recall we will investigate the application of word stemming to Chinese words in such a way that comparison becomes possible between 大众 and 大众性, 现代 and 现代的. This goal may be achieved by considering heuristic morphological rules as for segmentation.

References

[1] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*: McGraw-Hill, 1983.

[2] H. Fujii and W. B. Croft, "A comparison of indexing techniques for Japanese text retrieval," *Research and Development in Information Retrieval, ACM-SIGIR*, 237-246, 1993.

[3] Y. Ogawa, A. Bessho, and M. Hirose, "Simple word strings as compound keywords: An indexing and ranking method for Japanese texts," *Research and Development in Information Retrieval, ACM-SIGIR*, 227-236, 1993.

[4] C.-K. Fan and W.-H. Tsai, "Automatic word identification in Chinese sentences by the relaxation technique," *Computer Processing of Chinese and Oriental Languages*, vol. 4, pp. 33-56, 1988.

[5] N. Y. Liang and Y.-B. Zhen, "A Chinese word segmentation model and a Chinese word segmentation system PC-CWSS," *COLIPS*, vol. 1, pp. 51-55, 1991.

[6] W. Jin, "A Case Study: Chinese segmentation and its disambiguation," Computing Research Laboratory, New Mexico State University, Las Cruces, Technical report MCCS-92-227, 1992.

[7] K.-J. Chen and S.-H. Kiu, "Word identification for Mandarin Chinese sentences," *5th International Conference on Computational Linguistics*, 101-107, 1992.
[8] C.-L. Yeh and e. al, "Rule-based word identification for Mandarin Chinese sentences - A unification approach," *Computer processing of Chinese and Oriental Languages*, vol. 5, 1991.
[9] B.-I. Li and e. al., "A maximal matching automatic Chinese word segmentation algorithm using corpus tagging for ambiguity resolution," *R.O.C. Computational Linguistics Conference*, Taiwan, 135-146, 1991.
[10] Y.-X. Zhou and W.-T. Wu, "A Practical Method of Segmentation of Chinese -- A Method Based upon Chain Table," *Journal of Chinese Information Processing*, vol. 4, pp. 34-41, 1989.
[11] T.-S. Yao, G.-P. Zhang, and Y.-M. Wu, "A rule-based Chinese automatic segmentation system," *Journal of Chinese Information Processing*, vol. 4, pp. 37-43, 1990.
[12] H. Xu, K.-K. He, and B. Sun, "The implementation of a written Chinese automatic segmentation expert system," *Journal of Chinese Information Processing*, vol. 5, pp. 38-47, 1991.
[13] K.-K. He, H. Xu, and B. Sun, "The Design Principle for a Written Chinese Automatic Segmentation Expert System," *Journal of Chinese Information Processing*, vol. 5, pp. 1-14, 1991.
[14] L.-J. Wang, T. Pei, W.-C. Li, and L.-C. Huang, "A Parsing method for identifying words in Mandarin Chinese sentences," *12th International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1018-1023, 1991.
[15] Z.-D. Dong, "Chinese platform project of Chinese information processing and Chinese language research," *Communications of COLIPS*, vol. 3, pp. 79-88, 1993.
[16] J.-S. Chang and e. al., "Chinese word segmentation through constraint satisfaction and statistical optimization," *ROCLING-IV*, Taiwan, 147-165, 1991.
[17] R. Sproat and C. Shih, "A statistical method for finding word boundaries in Chinese text," *Computer Processing of Chinese and Oriental Languages*, vol. 4, pp. 336-351, 1991.
[18] M.-Y. Lin, T.-H. Chiang, and K.-Y. Su, "A preliminary study on unknown word problem in Chinese word segmentation," *ROCLING V*, 147-176, 1992.
[19] T.-H. Chiang and e. al., "Statistical models for segmentation and unknown word resolution," *5th R.O.C. Computational Linguistics Conference*, 123-146, 1992.
[20] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, pp. 61-74, 1993.
[21] M.-S. Sun and e. al., "Some Issues on the statistical approach to Chinese Word Identification," *3rd International Conference on Chinese Information Processing*, 246-253, 1992.
[22] S. Bai, ""Semi-word" method for Chinese word segmentation," *International Conference on Chinese Computing*, Singapore, 304-309, 1994.
[23] R. Sproat, C. Shih, W. Gale, and N. Chang, "A stochastic finite-state word-segmentation algorithm for Chinese," *ACL'94*1994.
[24] Z. Wu and G. Tseng, "ACTS: An automatic Chinese text segmentation system for full text retrieval," *Journal of the American Society for Information Sciencé*, vol. 46, pp. 83-96, 1995.
[25] Z. Wu and G. Tseng, "Chinese text segmentation for text retrieval: Achievements and problems," *Journal of the American Society for Information Science*, vol. 44, pp. 532-542, 1993.
[26] C. Buckley, "Implementation of the SMART information retrieval system," Cornell University, Technical report 85-686, 1985.
[27] J.-Y. Nie, W. Jin, and M.-L. Hannan, "A hybrid approach to unknown word detection and segmentation of Chinese," *International Conference on Chinese Computing*, Singapore, 326-335, 1994.

破音字發音的預測方法

王文俊†‡　黃紹華‡　李俊曉†　劉繼謚†

†交通部電信研究所　基本科技研究室
‡國立交通大學　電信工程研究所

email: wjwang%tl9000@tlrouter.motctl.gov.tw

摘要

目前在中文語音合成研究的主要方向大都偏重在如何提高清晰度及自然度，而字轉音的正確率問題則較少討論到。事實上語音合成的最主要目的是要讓聽者了解合成語音的意義，而不致產生誤解。因此字轉音正確率的提高實有其重要性，本文就是希望能利用有效的分析方法結合語言學的知識，針對破音字的發音提出解決的方法。本實驗所採用的方法是結合字的聯想與樹狀語言模式兩項處理。由實驗結果發現總合預測正確率約為 85% ，較傳統猜測法提高了約 20% 以上。其中字的聯想是一種類似句型的觀念，而樹狀分析模式則是為了能有效地運用詞類訊息。這些方法均具有延展性及擴充性，可適用於其它破音字的處理或相關的語音及語言處理。

1. 緒論

目前在中文語音合成研究的主要方向大都偏重在如何提高清晰度及自然度，而字轉音的正確率問題則較少討論到。事實上語音合成的最主要目的是讓聽者了解合成語音的意義，而不致產生誤解。因為就語音認知程序而言，聽者是藉由正確的音來組成相對應的字及詞，進而了解文句或文章的意義。因此為了達成此一目標，就必須降低字轉音的錯誤率。而造成國語字轉音錯誤的最大原因就是破音字的問題，因為錯誤的音形成不同意義的字及詞，導致誤解文句的原意。所以本文就是想利用有效的分析方法結合語言學的知識，針對破音字的發音提出解決的方法。

對於中文而言詞才是句法上及語意上的最小元素，因此在文字翻語音的處理中，輸入文字串必須先經過斷詞及構詞處理。而傳統上破音字的問題就是希望能在斷詞時一併解決，藉由事先已經經過檢查且存在詞庫中的各詞的相對注音決定各詞的發音，而破音字的發音也隨著詞發音的決定而跟著決定。這樣的處理除非在斷詞混淆或音異義異的破音詞情況下，的確可以解決絕大部份的破音字問題。而本篇文章所要探討的主題則正是經過上述斷詞處理後成為單字詞的破音字的發音問題，至於斷詞混淆或異音異義的破音詞問題則不在此討論。

2. 中文文法的特性

破音字並非中文固有的特性，相同的情況也出現在英文，稱之為 Homography 其義為同形異義字。對文句翻語音的系統而言，在處理過程中可資利用的資訊應就是上下文的文字及詞類。以「為」字為例從中研院中文詞庫小組所發表的資料〔1〕中可看出「為」字的發音可以根據詞類不同而有效的區別，如書中所述當「為」字在文句中的詞類為 P02 或 VG2 時此字應為第二聲，而當其詞類為 VJ1 或 P03 時此字應為第四聲。但事實上如此書所列之總數 178 類的詞類對電腦處理甚至非語言學家而言都並非很容易的事；而至於在縮小詞類數目下所發展出的詞類標示系統，其正確率亦尚無法達到百分之百同時也並未就此一破音字混淆問題做深入探討。因此本文的目的就是希望在容忍一些詞類標示混淆的情況下，仍能藉由其他相關的資訊幫助作破音字的分析，以建立一套有效的處理模式。本實驗初期的目標仍先選定由「為」字開始，因為在中研院所公佈的辭典中「為」字的詞頻是佔整個八萬詞目詞典中的第九位，而且其他大部分的破音字成詞的機會很大，形成單字詞造成混淆的機率反而較小。

決定破音字發音的困難究其原因主要來自斷詞及構詞的錯誤，使得應成詞的破音字在錯誤的詞位，而造成發音的混淆。另一點則是詞類標示有問題所造成，而詞類標示的問題則與下列兩點有關〔2〕：一為大部分的中文詞均可作為動詞、名詞或形容詞但卻沒有任何的構詞變化。二為中文無嚴謹的文法限制，在文句中允許各種詞類的組合，甚至不合文法的情況亦常存在，因而增加了所有語言處理的困難。以下所提出的方法即是希望能補償屬於中文文法的這兩項缺失。

3. 樹狀語言模式

文字翻語音的處理是由一整串的文字中去預測各字的發音，因此可資利用的資訊應包括文字及隱藏的詞類。本章就是要討論如何有效的運用這兩項資訊，首先處理的對象當然是上下文中的文字，對於此項字的處理，本文所提出的方法可稱為字的聯想（ lexical association ），所謂 lexical association 是一種根據相鄰字的出現頻率幫助決定破音字的發音，事實上和破音字成詞的字就可以視為是一種距離為 1 的 lexical association 的情況，而當把距離延長時也可以找出其他一些高頻率的字。這些字和破音字的組合可視為是一種常用句型，其中在這些常以組合方式出現的字中間可以擺上不同的字或詞，此種架構在文法上雖不屬於詞，但就實用而言在文章及口語中常會搭配出現，此種類似句型的觀念可用來解決詞庫過度膨脹的問題，而適當地運用句型及複合詞的文法更可有助於增加系統處裡文句的深度和廣度。不過由於中文的結構允許許多詞可以合併為新詞，因此在這種不一致的構詞處理下，此種句型架構的資料結構表示方式並不容易。

接下來要談到如何利用詞類訊息，雖然我們承認詞類標示的正確對所有語言處理都有很大的助益，但在此將不作如何提升詞類標示正確率的討論。我們秉持的是一種 partial parsing 的觀念，將不易作詞類標示易造成混淆的部份，由更高層次的處理來解決，在此我們所採取的分析方法是樹狀語言模式。此種分析模式已被廣泛利用在多方面的研究〔3〕〔4〕，且得到不錯的結果。樹狀語言模式可視為是 N-gram model 的變形，當 N=3 時此模式就變成所謂的三連文法模式。眾所週知 N-gram model 明顯地較三連文法模式精確，因為此種模式考慮到更多的資訊，但是在大部分的情況下，三連文法模式已經夠用，因而傳統的語言模式通常只用到雙連文法模式或三連文法模式及其之間的平滑模式。在〔5〕中就曾提及利用類似雙連及三連文法模式的條列式相似率比較法，來解決破音字問題。所謂條列式相似率的作法係以所欲處理的破音字在文句中不同位置的相鄰詞及其所具詞類為依據，並利用此破音字的正確注音作區分，計算出個別的相似值；在比較之後，留下一些具有高鑑別率的規則作為其後判斷的依據。這種作法的缺點是對詞的位置的處理缺乏彈性，同時只能建立簡單的分類規則。為了使破音字發音預測的正確率提高，我們必須考慮更多的相鄰字或詞，並且要有更有效的分析方法，因此 N 值就有必要提高並建立更有效率的樹狀語言模式。此模式的分析步驟為：首先必須整理出所有影響的因素形成 Questions set ，理論上這些影響「為」字發音的因素應由語言學家設計，但在本實驗進行時並無法找到相關的資料，而且即使有了類似資料其是否適於程式化執行亦未可知，因此本實驗所使用的 Questions 是在考慮系統複雜度及前節所述中文的特性和「為」字的語法關係之下所制定出來的，這些規則再經由檢測語料庫的步驟篩選出適當的規則組合以作為樹狀語言模式的產生依據。接著再以二分分裂法的方式持續地在所有 Questions set 中選擇最有效的 Question ，將資料分割成數個內部分佈更一致的小單元。本篇文章所討論的樹狀模式建立程序主要是利用 Greedy algorithm ，分裂的判斷依據則是採用 Gini criterion〔6〕。至於更詳細的樹狀結構的運算及修剪將不在此討論。

樹狀語言模式的優點是利用多變化的 Question 組合，可以用更有效的方式來實現 N-gram model ，另外在 N-gram model 下由於考慮語料中無法包含各種

可能的組合因此必須使用到各種平滑的技巧以避免機率值出現為零的這種考慮，在樹狀語言模式中可以被輕易地克服，原因是此系統具有容錯的能力在無符合分析時所有的情況下，依然能產生一個最接近的結果。

4. 實驗結果與討論

首先我們先介紹實驗所使用的詞庫及語料，詞庫的總數約為八萬詞目，其中包含單字詞、雙字詞、三字詞、四字詞及五字詞。訓練及測試的語料庫則是取自報紙新聞，整個語料庫總計約為 35 萬句 ，詞數約為 260 萬 ，字數約為 400 萬。所有的文句先經過初步的斷詞處理〔7〕後，整理出含有成詞的「為」字共 11400 句，而含有單字詞的「為」字共 12200 句，我們所要處理的就是這 12200 句的語料。整個處理的步驟如下：第一步是對所有語料中的破音字進行標示注音的工作，接著統計相鄰文字的出現頻率，得出表一的結果，由表一可看出滿足表中所列句型約 3000 句而「為」字發音正確率可達 98% ，其中屬於相鄰字的詞可以考慮加在系統的詞庫中，至於某些具有高頻出現率的詞，如「為了」為何不在原先之詞庫中，理由是我們所用的原始斷詞方法，希望能將某些具有規則的字留在構詞處理時解決。第三步則是建立屬於「為」字的樹狀語言模式，如圖一所示為分析後所得結果之示意圖，本系統所使用的詞類分類係參考〔8〕其中分類 0 至分類 11 為述詞，分類 39 為述補式複合詞，分類 20 為代名詞。在圖中所出現的幾個 Question 是經過上節所述的處理步驟而產生的，在此略述一下其所代表的意義：由字面意義分析可知，「為」字的單字意義可區分為「是」、「被」或「替」前兩者發第二聲，而後者則發第四聲；一般而言當「為」字有「是」的意義時，其前通常會緊接著動詞，而當「為」字有「替」的意義時，其後通常會有代名詞或動詞，而很多情況下原本屬於動詞的詞，在不同的文句中會轉化為其他詞類，因此有必要考慮這些特殊的情況。

表二所示可視為是傳統破音字發音猜測法和樹狀語言模式所得結果之比較，如表中語料庫資料分佈的數據，就正是傳統作法以在大量語料庫中出現機率最大者為主的預測正確率數據，而利用樹狀語言模式所得到的預測正確率則約為 80% ，若再加上由 lexical association 所得到的結果正確率更可提高為 85% 左右。

5. 結論

字轉音的正確率對語音合成系統而言可以視為是一個基本的要求，雖然對國語共約 1300 個基本單音及 5401 個常用字而言，破音字所佔的數量也許並不大，但錯誤的破音字發音對系統仍是實現語意了解的一大障礙。本文所提出的解決破音字發音的方法是結合字的聯想與樹狀語言模式兩項處理，由實驗結果發現總合預測正確率約為 85% 較傳統猜測法提高了約 20% 以上。其中字的聯想是一種類似句型的觀念，而樹狀分析模式是為了能有效地運用詞類訊息；這

些方法均具有延展性及擴充性，可適用於其它破音字的處理或相關的語音及語言處理。
未來的研究方向將朝向更詳細的樹狀結構分析以提高預測的正確率，另外也將擴大研究至所有的破音字，並考慮增加詞意的資訊來分辨如「一行人」或「一行字」這種相鄰詞類相同的情況以及包括音異義異的破音詞處理。

致謝

感謝交通部電信研究所所長王金土博士，副所長周義昌博士以及基本科技研究室主任鄭伯順博士對語音信號處理研究的持續支持與鼓勵。也感謝交通大學電信工程研究所教授陳信宏博士的指導與建議。

參考文獻

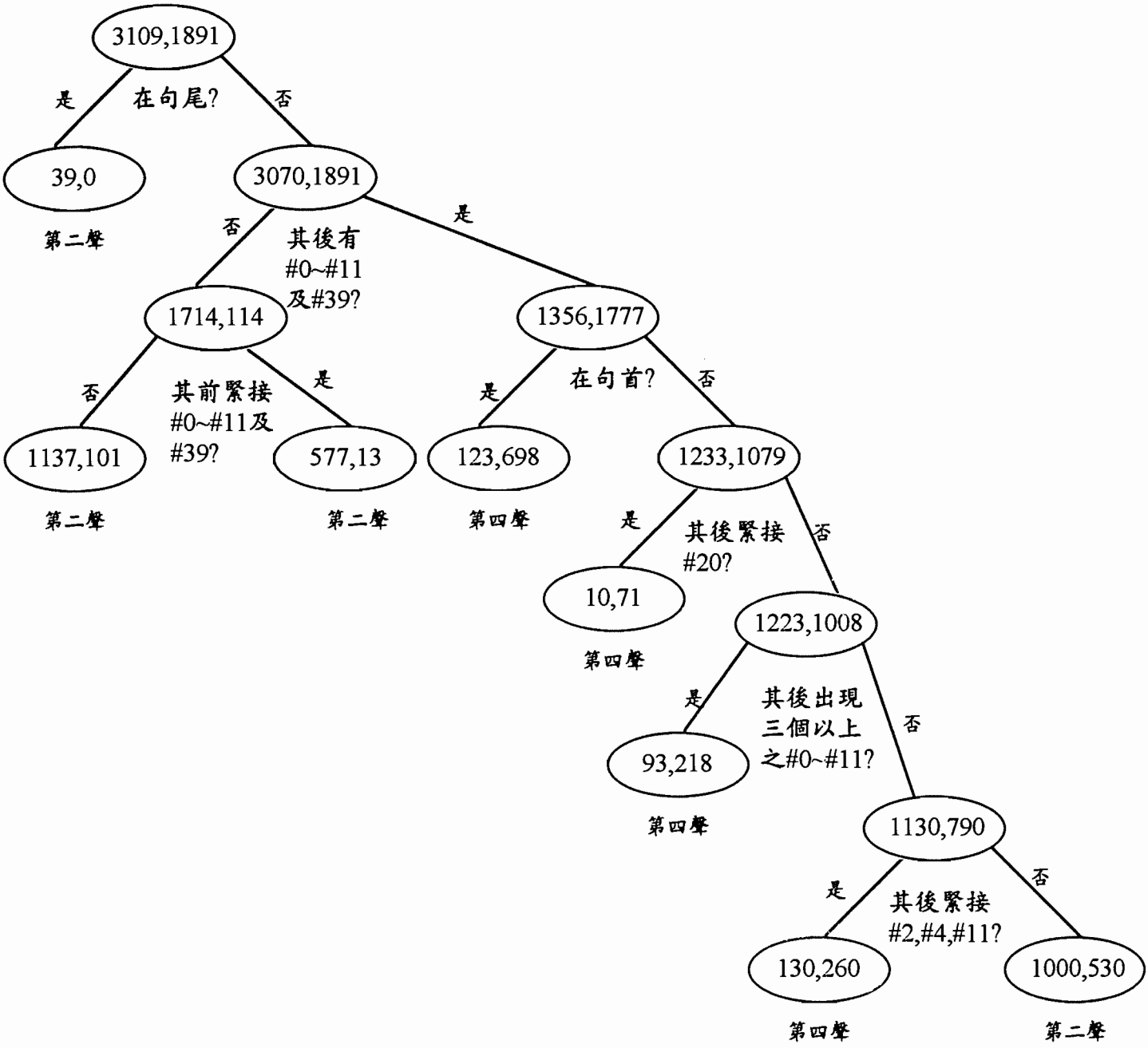
〔1〕 中央研究院，資訊科學研究所，「中文書面語頻率辭典」。
〔2〕 K.J.Chen, S.H.Liu, L.P. Chang and Y.H. Chin, '' A Practical Tagger for Chinese Corpora '', ROCLING 1994, pp.111-126.
〔3〕 Michelle Wang and Julia Hirschberg, '' Automatic Classification of Intonational Phrase Boundaries '', Computer Speech and Language 1992 vol.6, pp.175-196.
〔4〕 W.J.wang, N.Campbell, N.Iwahashi and Y.Sagisaka, " Tree-based Unit Selection for English Speech Synthesis ", ICASSP 1993, pp. 191-194.
〔5〕 王文俊，李俊曉，「以詞類分析作破音字處理」，交通部電信研究所，內部報告。
〔6〕 L.Breiman, J.H.Friedman, R.A.Olsen and C.J.Stone, " Classification And Regression Trees ", Monterey. CA: Wadsworth, 1984.
〔7〕 李俊曉，李俊仁，黃英峰，「中文文句翻語音語言處理系統—工作文件1 」，交通部電信研究所，內部報告。
〔8〕 蘇育新，「中文文句自動斷詞標詞類之研究與應用」，國立交通大學碩士論文。

表一、含「爲」字的句型整理

句　　型	樣本總數	注音爲第二聲	注音爲第四聲
...爲了...	1133	0	1133
...爲之...	91	89	2
...爲由...	124	124	0
...以...爲...	769	756	13
...爲...去...	10	0	10
...爲...而	228	13	215
...爲的(卻,就)是.	7	0	7
...稱爲...	41	40	1
...爲..了...	97	11	86
...爲(此,免,使)...	316	6	310

表二、使用分類樹對於含「爲」字語料的影響

語料項目	資料分佈	預測正確率
訓練語料	62.18%	80%
測試語料	62.60%	79.175%


3109,1891
是
在句尾?
否
39,0
第二聲
3070,1891
否
其後有
#0~#11
及#39?
是
1714,114
1356,1777
否
其前緊接
#0~#11及
#39?
是
是
在句首?
否
1137,101
第二聲
577,13
第二聲
123,698
第四聲
1233,1079
是
其後緊接
#20?
否
10,71
第四聲
1223,1008
是
其後出現
三個以上
之#0~#11?
否
93,218
第四聲
1130,790
是
其後緊接
#2,#4,#11?
否
130,260
第四聲
1000,530
第二聲

圖一、分類樹示意圖

以CELP為基礎之文句翻語音中韻律訊息之產生與調整

吳宗憲，陳昭宏，莊欣中

國立成功大學 資訊工程研究所

chwu@server2.iie.ncku.edu.tw

摘要

在本論文中，我們對於所收集的語料庫中的各個單音，分析其基週變化特性，藉由向量量化之觀念，歸納出十二組基週軌跡參考樣本，以代表對應於四聲及輕聲的音高週期之變化，並藉由一拜氏網路機率統計模型，對連續語音資料加以分析，以描述文句與語音韻律變化之關係，並在語音合成過程中，決定一適當的基週參考樣本以供作韻律上的調整。另外，我們提出了一套韻律訊息產生及調整的方法，供韻律調整模組對CELP語音合成器中之激發源脈衝加以調整。經過 20 位測試者評估之後，在平均可辨度方面達到 96.6%，而在自然度方面，評分在等級「可」以上的佔了 84.4%。

一、緒言

一般說來，語音合成的方式主要可分為兩類：第一類的作法是將可能使用的語音信號事先錄製下來，當系統欲說出某一文句時，僅須找出相對應的語音信號，將其輸出即可。這一類語音合成方式的複雜性低、運算量較小，但合成之語音易於達到自然、流利、清晰的要求，適合應用在少量文句的語音合成系統之中。另一類語音合成方式，是先將基本語音合成單元及合成規則存放於記憶體之中，利用這些基本語音合成單

元組合成與輸入文句相對應的語音信號，並配合語音合成規則加以調整音高(pitch)、音長(duration)、音強(energy)、停頓(pause)及句調(intonation)等音韻特徵。這一類語音合成方式的複雜性較高，但所需之記憶體空間較小，可以合成任意文句，因此應用範圍極為廣泛，本論文中所提之文句翻語音系統便是屬於此類合成方式。

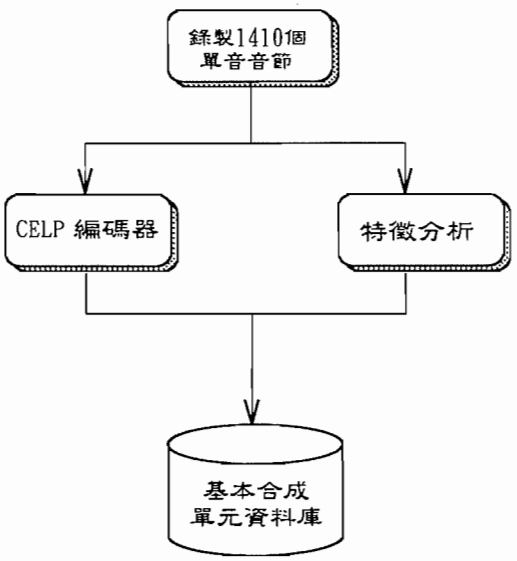
另外，語音編碼技術通常可以分為三種，即波形編碼（Waveform coding），參數編碼（Parameter coding）及混合式編碼（Hybrid coding）。波形編碼方式，一般應用在時域上（Time domain），如DM、DPCM、ADM 及 ADPCM 等，皆是於時域上去模仿語音波形，可以產生較高音質的語音，但壓縮率較低。參數編碼方式是應用在頻域上（Frequency domain），以模擬人類聲道特性為基礎，如 Linear Prediction Coding （LPC）等，雖然可以有較高的壓縮率，但所得到的音質是較不清晰的。而混合式編碼則是結合上面兩種方法的優點，如 Multi Pulse Excited Coding （MPE） 及 Code Excited Linear Prediction Coding (CELP) 等。

本論文所製作的文句翻語音系統，主要是以 408 個國語單音音節，配合聲調(tone)的變化，作為基本的語音合成單元，所以我們預先錄製了 1410 個由女性發聲的國語單音（含四聲及輕聲），利用碼本激發線性預測語音編碼技術 (CELP) 高壓縮率及其合成音質幾近原音之特性，將所有語音資料編碼壓縮後儲存，其壓縮率可達 13.3 倍。

過去的語音合成系統，通常是採用條列法則 (Rule-based) 來決定如何調整音韻的變化，但是必須藉由人工去分析大量的語音資料，這是一件非常費時且困難的工作。在本論文中，我們藉由一拜氏網路(Bayesian Network) 機率統計模型，對連續語音資料加以分析，以描述文句與語音韻律變化之關係。供韻律調整模組對語音合成器中所產生的激發源脈衝 (excitation pulse) 加以調整，以期使得輸出的合成語音更為自然、流利。

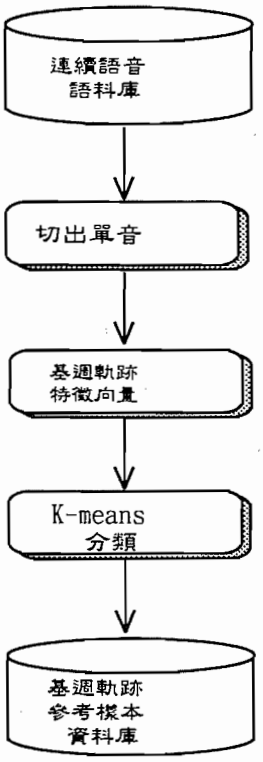
二、系統架構

基本合成單元資料庫的建立，如圖(一)所示，我們錄製了由女性發聲的 1410 個國語單音音節，包含了四聲及輕聲的變化，錄製時是以 11.025k 的取樣頻率，並控制各個單音音節的平均音量大小之誤差在 20% 以下，而且單音音節長度皆固定在 0.27 ms。接著將這 1410 個單音音節透過 CELP 語音編碼器加以處理，編碼後之參數儲存於基本合成單元資料庫中。另外，我們找出音節的母音段中（final），所有的基週中心點標記（pitch position），以及穩定區音框位置，並將這些資料亦存於資料庫中。



圖（一） 基本合成單元資料庫建立流程

在基週軌跡參考樣本方面，我們收集了約一百八十句的連續語音，然後由其中切分出各個單音音節，分析其基週變化之特性，透過 K-means 分類及向量量化之後，歸納出具代表性的十二組基週軌跡參考樣本，以對應四聲及輕聲的變化，之後，可供系統中韻律訊息產生及調整之用，見圖(二)。



圖(二) 基週軌跡參考樣本資料庫建立流程

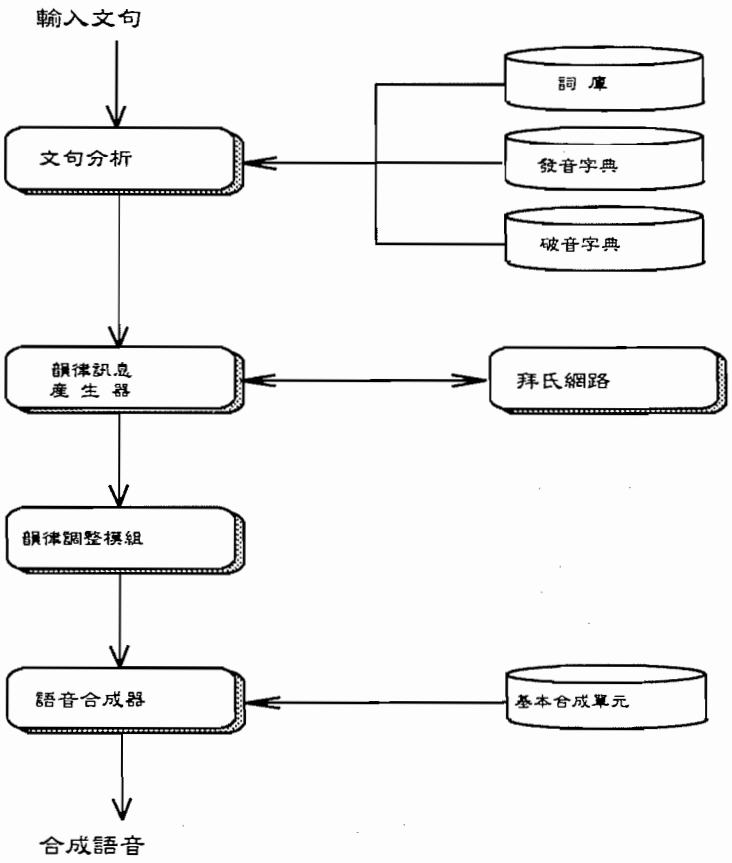
本系統的基本架構如圖(三) 所示，主要流程如下：

1. 文句分析：利用系統詞庫對輸入的文句字串進行斷詞及構詞的工作，找出詞邊界、詞類屬性、以及斷句邊界，並配合發音字典與破音字典，取得各個字相對應的注音符號。

2. 韻律訊息產生器：根據一些簡單規則，以及透過一拜氏網路機率統計模型，產生對於基週軌跡、音強、音長及停頓等韻律特徵的調整訊息參數。

3. 韻律調整模組：依據韻律訊息產生器之結果對語音合成參數進行調整修飾。

4. 語音合成器：利用碼本激發線性預測 (CELP) 語音合成器將調整後的合成參數轉換成語音信號輸出。

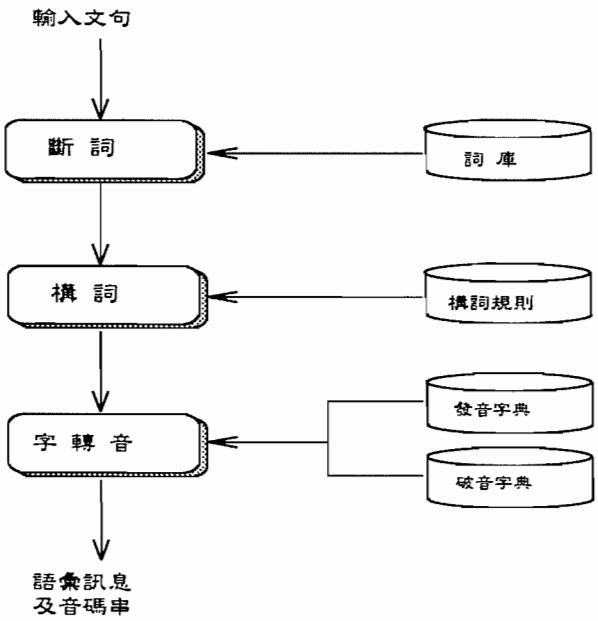


圖（三）文句翻語音系統流程

三、文句分析

整個文句分析的方塊圖如圖（四）所示。第一個步驟是斷詞（Word Identification）。主要有兩種方法，即統計學法與語言學法。統計學法是蒐集大量詞彙加以分析統計，這個方法較為簡單，不太需要語言學上的知識，但語料庫不可能涵蓋所有的詞彙，因此有可能遇到從未出現的語句而造成斷詞上的錯誤。而語言學法則是利用語言學上的知識，以制定一套完善的文句剖析法則，或是利用一些經驗法則來分析文句。兩種方法各有利弊，也都無法完全正確的斷詞。本系統是於事先建立一套詞庫，再配合一些簡單的規則，來處理這一部份的工作。

接下來我們將找出與每個字元相對應的正確發音。首先，我們先建立一個基本的發音字典，裡面記錄了與每個中文字相對應目前使用最普



圖（四）文句分析流程

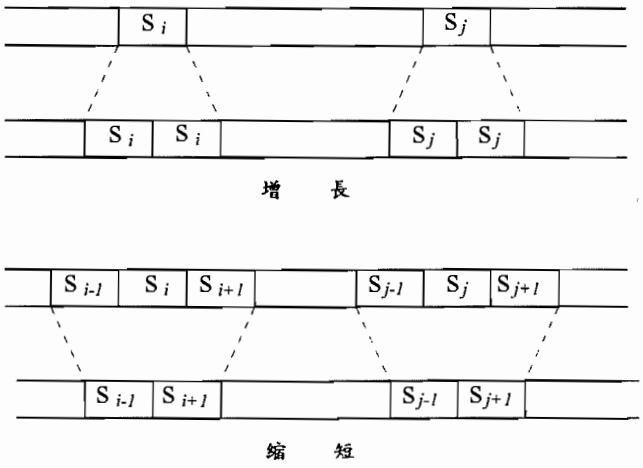
遍的發音；接著，必須對破音字加以處理，我們是對常出現的破音字，蒐集與其相關的破音詞，儲存於系統中的破音詞典中，若輸入的文句中含有破音字時，便將該部分斷詞之結果和詞典中的破音詞相比對，以決定該破音字的正確發音。另外，在我們的說話習慣中，對於某些情況的發音會有固定的變調，如下：三聲字接三聲字時，前一個三聲字會變為二聲；「一」、「不」的變調。

四、韻律調整模組

§4.1 音長、音量及停頓的調整

當音長需要增長或縮短時，主要是對基本合成單元的穩定區段加以調整，才不致使修改過的合成單元與原始合成單元相差太大。各個基本合成單元的穩定區的搜尋，我們已在系統資料的立流程中處理完成，並記錄下來。至於穩定區的找尋方法，說明如下：

首先對於每一個合成單元，以240點為一個音框(frame)，求取各音框的 LPC 係數，並轉換成倒頻譜（Cepstrum）係數。接著利用穩定區內相鄰音框之倒頻譜係數變化極小的性質，依序求出各音框與附近音框的差異性最小的兩個音框，並標示為穩定區，當我們需要調整音長時，便由語音解碼器所產生的激發源脈衝(Excitation Impulse)上，從穩定區增長或縮短，然後合成語音，見圖（五）。



圖（五）增加或刪除激發源脈衝的穩定區以改變音長

基本上音強的調整，是直接由時域上對語音波形乘上一個增益值即可。至於停頓的處理可由斷詞及構詞處理之後，在發音詞邊界處插入所需的停頓，也就是靜音，同時對各個標點符號也做不同的停頓處理，如表（一）所示。

標點符號	，	；	。	？	！
停頓時間	480ms	500ms	520ms	540ms	560ms

表（一）標點符號與停頓時間之關係

在執行時，依據以下的原則作整體的音長、音量及停頓的調整：

（１）模擬人的換氣動作，大約每六個字為一循環，遇到詞則提前或延後換氣。換氣前音量逐漸降低，換氣時略微停頓，換氣後音量較為提高。

（２）句尾要拉長音長，減小音量。

（３）遇到常用的介詞時，例如的、是、著、到、和、得等字，縮短音長、降低音量，並拉長前一個字的音長。

（３）詞之前、句尾亦加入適當之停頓。

（４）將文句分為直述句、疑問句、驚嘆句、命令句，分別調整其音長及音量整體走勢。

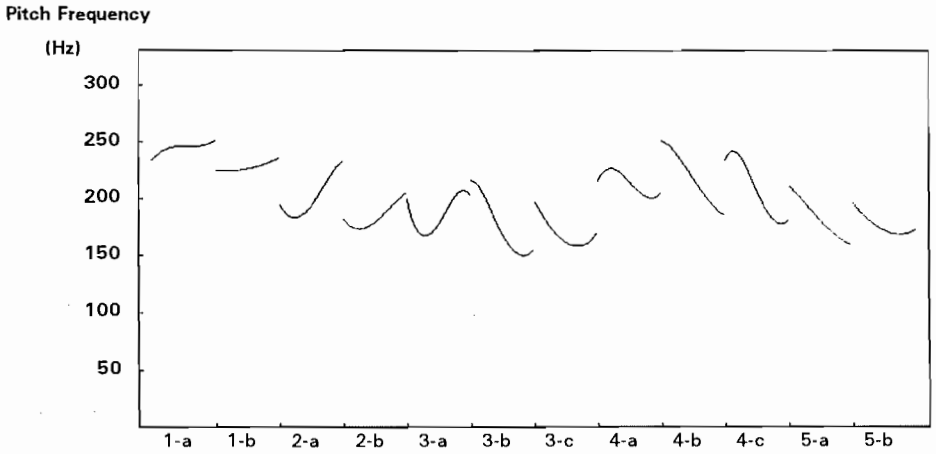
§4.2 音高的調整

基本上每個中文單音都具有四聲變化以及輕聲，但是在組成句子時，隨著連接狀況的不同，單音的聲調會有許多的變化，僅用四聲及輕聲來表示，是不夠完整的。因此，我們根據分析統計之後，取了12組的基週軌跡樣本（pitch contour pattern）來代表原來的五聲變化，其中一聲、二聲及輕聲分別包含二類基週軌跡樣本，三聲與四聲則分別包括了三類，參考圖（六）。

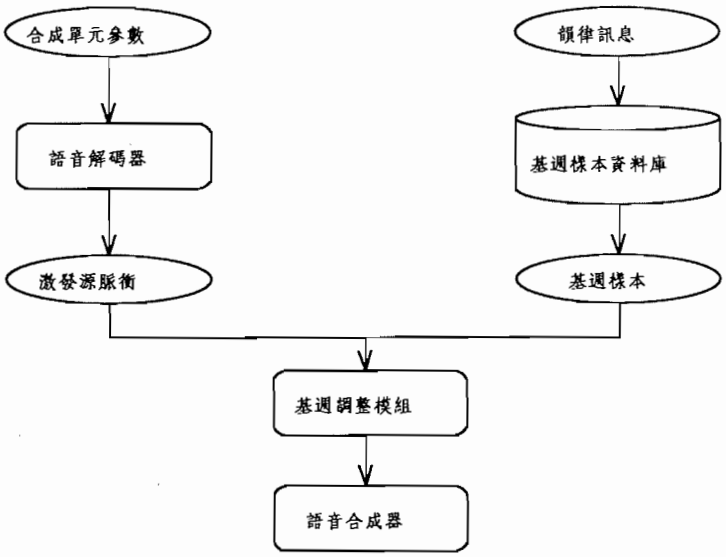
對於一個單音的基週軌跡，可利用正交化多項式展開法(Orthnormal Ploynomial Expansion)〔10〕將其轉換為一組四維的向量值 $\vec{a}=(a_0,a_1,a_2,a_3)$，來表示相似的曲線。因此，對於每個單音的基週軌跡，僅需用一個四維的向量來表示，即 $\vec{a}=(a_0,a_1,a_2,a_3)$。

接著再討論有關音高的調整，處理流程請參考圖（七），首先由韻律訊息產生器產生音高調整訊息，接著由資料庫中找出相對應的基週樣本，然後透過基週調整模組，對原合成單元的激發源脈衝，作基週上的調整。其中，基週調整的演算法是利用基週

同步重疊相加合成法(pitch synchronous overlap and add, PSOLA)〔11〕。



圖（六）十二組基週軌跡樣本(pitch contour pattern)



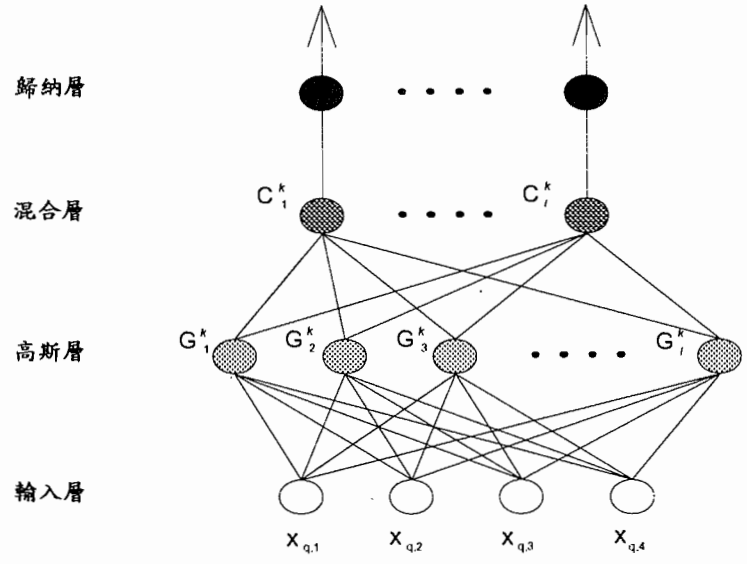
圖（七）音高週期調整流程

五、韻律訊息產生器

韻律變化的相關訊息，包括音高、音長、音量及停頓等調整參數，在過去的文句翻語音系統中，通常是採用條列式的法則

(Rule-based) 來決定韻律的調整參數，必須藉由人工去分析大量的文句與語音資料，這是一件極費時且困難的工作，另一方面，語音韻律的變化是非常繁多的，我們無法用少量的規則，即可涵蓋所有可能的變化，若使用大量規則，會造成處理速度變慢，且常會有使用規則混淆（ambiguous)的情況發生。因此我們採用機率統計觀念與分類法則，從大量的語料庫中找出一些具代表性的參考樣本，以描述文句與韻律變化之間的關係。在此我們採用拜式網路 (Bayesian Network)來建立參考樣本。

拜氏網路是一個以拜氏定理(Bayes' Theorem)為理論基礎的網路模型，其架構可分為四層：輸入層 (Input slab)，高斯層(Gaussian slab)，混合層 (Mixture slab) 及歸納層(Aposteriori slab)。圖（八）是拜氏網路的架構圖。



圖（八） 第 k 聲調的拜氏網路架構

基本上，拜式網路如同是一個拜式分類器再配合混合高斯機率分佈(Gaussian Distribution) 的觀念。首先對輸入的特徵向量分類，分類結果的每一類便代表一個高斯分佈，其平均值和變異數可經計算而

得，至於加權值則是根據某一特徵向量分佈在各類中的個數而定。拜式網路的主要目的是在求得某一輸入向量 X 在歸納層的某一類別 C_i 中出現的機率 $P(C_i|X)$，根據拜式定理 (Bayesian Theorem):

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

其中P(X): 向量 X 出現的機率；$P(C_i)$: 屬於類別 C_i 的機率；$P(X|C_i)$: 在類別 C_i 中，屬於向量 X 的條件機率；混合層的輸出是一個混合高斯機率，它是取高斯層中的各個分佈機率乘上個別的加權值當作輸入，而決定輸出的法則一般有兩種，一是 GAM ，取機率密度總和為輸出，另一為 PGAM，則是取最大的機率密度為輸出。高斯層是根據對輸入特徵向量分類後的結果而定的。

在基週訊息產生模組中，我們對於不同的聲調(tone)分別建立一個拜氏網路，其輸出是對於某一個欲處理的字，決定應選取那一類的基週軌跡樣本以供調整。其輸入則是考慮該字在文句中相關的特徵參數，共有四項，第一項參數是相連的前一字的基週軌跡，以一組經正交化函數轉換後的向量$\vec{a}=(a_0,a_1,a_2,a_3)$表示，第二項參數是考慮下一個字的基週軌跡，但由於下一字的基週軌跡尚未計算，因此取下一字聲調相對應的代表性基週軌跡，最後兩項參數則是考慮該字在文句中的位置及在詞中的位置。下面將描述我們所用的拜氏網路架構及機率估算流程，參考圖 (八)，首先定義下列變數：

k：第 k 聲調(tone)；

N：第 k 聲調的基週軌跡類別總數；

C_i^k：第 k 聲調的第 i 類基週軌跡樣本；

M：高斯層中的子類別個數；

G_j^k：高斯層中的第 j 個子類別；

Xq：第 q 組輸入特徵向量；

$x_{q,l}$：Xq 中的第 l 項參數；

w_{ij}：連接混合層中 Node i 與高斯層中 Node j 的加權值；

對於一輸入向量 Xq ，其對應高斯層中子類別 G_j^k 的機率 $p(X_q|G_j^k)$，我們用一個高斯機率密度函數來計算，如下：

$$p(X_q|G_j^k) = N\left[X_q, \mu_j^k, \sigma_j^k\right] = \frac{1}{(2\pi)^{D/2}\left(\prod_{d=1}^{D}\sigma_{j,d}^{k\ 2}\right)^{1/2}} \exp\left(-\sum_{d=1}^{D}\frac{(x_{q,d}-\mu_{j,d}^k)^2}{2\sigma_{j,d}^{k\ 2}}\right)$$

其中 D 表輸入特徵向量的維數當高斯層子類別的條件機率求出之後，再乘上高斯層與混合層之間的加權值，即可在混合層得到混合高斯機率，計算公式如下：

$$p(X_q|C_i^k) = \sum_{j=1}^{M} p(X_q|G_j^k)\, w_{ij} = \sum_{j=1}^{M} N\left[X_q, \mu_j^k, \sigma_j^k\right] w_{ij}$$

其中 w_{ij} 為類別 C_i^k 中的向量被分到子類別 G_j^k 的個數除以訓練時類別 C_i^k 中的向量個數所得之值。然後根據拜式定理，可由下面關係式求得歸納機率：

$$p(C_i^k|X_q) = \frac{p(X_q|C_i^k)p(C_i^k)}{p(X_q)}$$

因為對所有的類別而言，輸入向量的機率 p(Xq) 是相同的，所以上式可化簡為：

$$p(C_i^k|X_q) = p(X_q|C_i^k)p(C_i^k)$$

上式中的 $p(C_i^k)$ 等於類別 C_i^k 中訓練樣本的個數除以全部訓練樣本總數所得之值。最後，找出具有大輸出機率值的類別 C_i^k，即是我們欲選取的基週軌跡類別。

六、實驗結果與結論

§6.1 實驗過程與結果

我們分別就可辨度(intelligibility)與自然度(naturalness)這兩方面評估系統的效能。實驗對象共 20 人，過程說明如下：在硬體方面包括了 PC/AT 486個人電腦、8-bit 聲霸卡以及外接喇叭，軟體方面則分為系統主程式 (250k bytes) 、詞庫檔 (360k bytes) 與合成單元資料庫 (939k bytes)。在可辨度的評估方面，我們以本系統對輸入文句做即時處理，立即輸出合成語音。測試者必須在語音輸出之後，將所聽到的語音以聽寫方式寫出，最後統計正確文句與被測試者所寫下的結果之間的差異，以作為評估可辨度的方式。在自然度方面，則是將事先準備的測試文件先由原先錄製音檔的該名女性以自然方式唸出，當作滿分標準，再由未經過韻律處理的系統播放，最後由本系統播放同一份文件，測試者則根據個人觀點對此系統的接受程度進行評分，評分的等級分為「優」、「佳」、「可」及「劣」四種。實驗結果顯示在表（二）及 表（三）。在可辨度方面，使用12組基週軌跡樣本可以達到平均 96.6%的可辨度，，未使用12組基週軌跡樣本則可達到平均97.1%的可辨度。我們發現兩者的可辨度相差無幾。在自然度方面，使用12組基週軌跡樣本的系統，在等級「可」以上的佔了 84.4%，未使用使用12組基週軌跡樣本的系統，在等級「可」以上的僅佔 79.6%。仔細比較各表中自然度的等級所佔的比例，我們可以發現本系統所輸出的語音已讓人感覺有不錯的效果。

測試種類	數　量	可辨度
單音	1410	92.8%
二字詞	200	95.4%
三字詞	200	98.7%
四字詞	200	99.1%
句子	100	97.3%
平　　均		96.6%

(a)可辨度方面實驗結果

測試種類	數量	自　然　度			
		優	佳	可	劣
二字詞	100	45%	34%	8%	13%
三字詞	100	32%	37%	16%	15%
四字詞	100	30%	44%	9%	17%
句子	100	25%	36%	17%	21%
短文	5	20%	52%	16%	12%
平　　均		84.4%			15.6%

(b)自然度方面實驗結果

表（二）使用12組基週軌跡樣本的實驗結果

測試種類	數　量	可辨度
單音	1410	94.4%
二字詞	200	95.4%
三字詞	200	98.9%
四字詞	200	99.1%
句子	100	97.6%
平　　均		97.1%

(a)可辨度方面實驗結果

測試種類	數量	自　然　度			
		優	佳	可	劣
二字詞	100	39%	30%	15%	16%
三字詞	100	31%	34%	15%	20%
四字詞	100	26%	33%	19%	22%
句子	100	20%	33%	22%	25%
短文	5	19%	35%	28%	18%
平　　均		79.6%			20.4 %

(b)自然度方面實驗結果

表（三）未使用12組基週軌跡樣本的實驗結果

§6.2 結論

在本論文中，對於韻律訊息的產生及調整的研究，確實對合成語音的自然度及流利度有明顯的提昇，也可達到即時處理的要求，但是要讓一般大眾能夠廣泛接受，成為一套實用價值高的文句翻語音系統，仍有一些需要改進的地方。

首先，若能夠增加斷詞及構詞的正確性，有助於音韻調整訊息的產生，對於破音字的處理，如何能找出正確的讀音，也是目前需要改善的地方。接著，在語音合成器方面，由於為了降低語音資料所佔的記憶體空間，所以利用語音編碼技術將資料加以壓縮，但是解碼還原之後的語音與原音比較仍有些許誤差，而在經過韻律調整模組中，PSOLA 演算法的處理後，由於為了將兩段波形能夠平順的連接在一起，部分運算難免會造成波形頻譜上的失眞，導致雜訊產生。因此，我們在時域上加上一個後置低通濾波器(low-pass filter)，其主要的功能是將共振峰波谷內的雜訊降低，因為人類的聽覺遮蔽效應，使得語音中共振峰附近的雜訊會被遮蔽，而波谷處的雜訊便相對地特別明顯。另外，在我們的錄音過程中，是以一種類似標準化的方式來錄製每一個基本合成單音，這種方式雖然能使合成出來的語音，各個單音都能較為清楚，但是，它卻缺少了連續語音中，字與字之間該有的連音(coarticulation)資訊，所以，未來我們應找出適當的方法來彌補這一個缺失，以提高合成語音的流利度。另外，目前系統所使用供作分析的語料庫，似乎仍不夠完備，在分析並學習文句與語音韻律變化之關係上，應再收集更多自然語音資料，藉由更大量的資料訓練，將來在處理各方面的應用時，才能產生更客觀且完整的韻律訊息。

參考文獻

〔1〕Lin-shan Lee and Chiu-yu Tseng "Mandarin Speech Input/Output Techniques for Chinese Computers - The Stateof the Art," Proc. Natl. Sci. Counc. ROC(A), Vol. 11, No. 4, pp. 273-290, 1987.

〔2〕歐陽明、李琳山，「一套中文的文句國語音系統」，台大電機研究所碩士論文，1985年6月

〔3〕劉繼謚等，「以線性預測編碼為合成器的中文文句翻語音系統」，電信研究季刊，第19卷第3期，民國78年9月

〔4〕朱國華等，「一套以多脈衝激發語音編碼器為架構之即時中文文句翻語音系統」，電信研究季刊，第21卷第4期，民國80年12月

〔5〕Ngor-chi chan and Chorkin Chan, "Prosodic Rules for Connected Mandarin Systhesis," Journal of Information Science and Engineering 8, pp.261-281, 1992.

〔6〕John Choi, Hsiao-wuen Hon, Jean-luc Lebrun, Sun-pin Lee, Gareth Loudon , Viet-Hoang Phan, and Yoganathan S., "Yanhui, a Software Based High Performance Mandarin Text-to-Speech System," Proceedings of ROCLING VII, pp.35-50, 1994.

〔7〕李俊曉等，「中文文句翻語音系統的語言處理模組」，電信研究季刊，第21卷第4期，民國80年12月

〔8〕Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 bit/second Code Excited Linear Prediction (CELP), National Communications

System, Office of Technology and Standards, Washington, DC 20305-2010, 14 February 1991.

〔9〕Lin-shan Lee, Chiu-yu Tseng, and Ching-jiang Hsieh, "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System," IEEE Trans. on Speech And Audio Processing. Vol. 1, No. 3, July, 1993.

〔10〕S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Trans. Commun., Vol. 38, pp. 1317-1320, Sept. 1990.

〔11〕F. J. Charpentier and M. G. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," ICASSP, pp.2015-2020, 1986.

〔12〕Christian HAMON, Eric MOULINES, Francis CHARPENTIER, "A diphone synthesis based on time-domain prosodic modifications of speech," ICASSP, pp.238-241, 1989.

〔13〕L. S. Lee, C. Y. Tseng and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," IEEE Trans. Acoust, Speech, Signal Processing, Vol.37, No.9, pp.1309-1319, Sept. 1989.

〔14〕Michael S. Scordilis and John N. Gowdy, "Neural network based generation of fundamental frequency contours," ICASSP, pp.219-222, 1989.

〔15〕S. H. Chen, S. H. Hwang, and C. Y. Tsai, "A first study on neural net based gendration of prosodic and spectral

information for mandarin text-to-speech," ICASSP, pp.45-48, 1992.

〔16〕Jhing-Fa Wang, Chung-Hsien wu, Shih-Hung Chang, and Jau-Yien Lee, "A Hierarchical Neural Network Model Based on A C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition," IEEE tras. Signal Processing, Vol 39, No. 9, pp.2141-45, 1991.

〔17〕Michael S. Scordilis and John N. Gowdy, "Neural network control for a cascade/parallel formant synthesizer," ICASSP, pp.297-300, 1990.

〔18〕Yoshinori SAGISAKA, "On the prediction of global F0 shape for Japanese text-to-speech," ICASSP, pp.325-328, 1990.

〔19〕Shaw-hwa Hwang and Sin-hrong Chen, "Neural network synthesiser of pause duration for mandarin text-to-speech," ELECTRONICS LETTERS, Vol.28, pp.720-721, April, 1992.

應用於"音中仙"國語聽寫機之短語規則分析與建立

葉瑞峰 王駿發 許志興

e-mail address: wangjf@server2.iie.ncku.edu.tw

國立成功大學資訊工程研究所

摘 要

簡便而好用的人機介面是資訊研究的一個重要課題，鍵盤乃是針對歐美拼音文字而設計的，對於中文這種方塊文字，若非受過專業輸入訓練是很難普遍地利用鍵盤來做中文輸入，所以發展國語聽寫機技術對資訊中文化是有十分重大的影響。

而在聽寫機方面的研究已行之多年，在國外已完成的系統有Hearsay II、Harpy、BBN、TINA及Dragon 等系統，在國內則有台大與中研院聯合開發的"金聲系列"[1][2]以及成功大學所發展的"音中仙巨量詞彙輸入法"。

本文即是對於語音辨認後處理之自然語言處理提出方法，使中文的語音輸入技術在理論上及實用上都能兼顧的考量下發展，在本文中短語分析規則主要有兩類：

1.單一詞未知短語規則：指短語規則中，有一未知詞而其它詞已知稱之。例如"姓氏+校長"為一短語規則其中校長為已知，姓氏為未知。此類法則乃針對詞庫中未建之詞必需加以簡單組合之詞，利用大量語料庫做統計，再依據統計的輸出做為辨認系統構詞的法則權重，以解決斷詞含混與詞庫不足的問題。

2.多詞未知短語規則：指短語規則中有多詞未知，例如"某某市某某路某某巷"其中市、路、巷為已知，但縣市名稱、路名及巷名三個未知。此類法則所處理的主要對象是數量詞和住址或複合詞可以用狀態轉移表示的詞組。

一、 國語聽寫機概說

在本節中我們分別就訊號方面的辨認特性以及在中文語言特性的研究來討論。

語音在辨認上之特性

由於語音是一種高時變性的訊號，所以在辨認時會造成若干誤差。此外外在環境與輸入設備也可能對語音訊號造成相當程度的影響，我們就以語音透過電話線為例來測試語音訊號的穩定性。發現語音訊號中的穩定資訊是集中於母音部份，而子音部份不但在辨認上的特徵不穩定，同時也容易受到環境和輸出入介面的影響。

為了測試母音跟子音在辨認上的穩定性，我們利用電話的通道效應[25]來對語音作處理，之所以選擇電話作為干擾的原因主要是因為通過變動性大的電話除了背景噪音外，在頻譜上若干對應的增益衰減，此外每次不同的電話連接都會有不同的效應產生，所以是屬於比較客觀的測試環境。

首先我們利用麥克風輸入408個音節三遍，每遍盡量用不同的聲調（四聲變化），然後再經由電話線來測試，我們一樣透過電話線來唸一千多個單音節，不過我們是透過四個不同的電話機，六次不同的電話連接來測試的。

子音	麥克風輸入		電話線輸入	
	單項	累積	單項	累積
第一名	59.53	59.53	25.53	25.53
第二名	19.04	78.57	12.76	38.29
第三名	10.71	89.28	17.02	55.31
第四名	4.76	94.04	8.51	63.82
第五名	3.57	97.61	5.50	69.01

表 1 子音透過麥克風及電話線之辨識率

母音	麥克風輸入		電話線輸入	
	單項	累積	單項	累積
第一名	84.52	84.52	51.06	51.06
第二名	10.71	95.23	10.63	61.69
第三名	2.38	97.71	6.38	68.07
第四名	1.10	98.81	8.51	76.58
第五名	1.00	99.81	6.41	82.99

表 2 母音在透過麥克風及電話線之辨認率

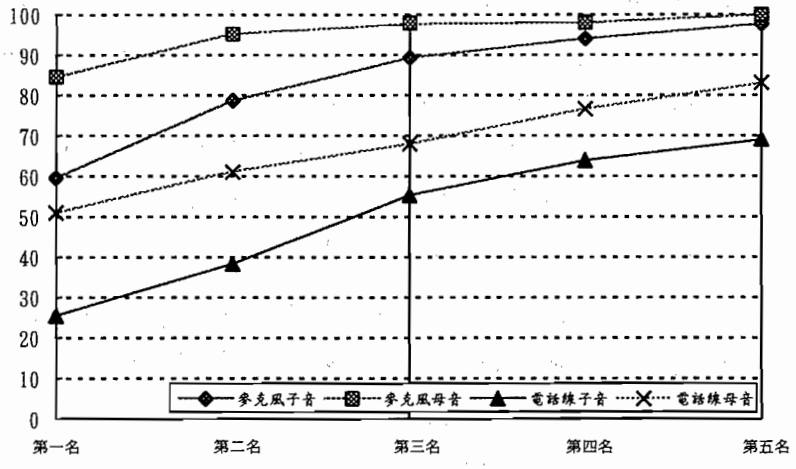


圖 1 透過麥克風及電話線後，子母音辨認率之比較

由以上實驗可知：母音在辨識上是十分穩定的，在未透過電話線的時候子母音的辨識率在前五名都達到九成五以上，可是在透過電話線通道效應干擾後，母音前五名累積辨識率仍能保持在八成以上，可是子音卻已經滑落至七成左右。所以在辨識上利用母音的穩定性是可以提高系統的包容性的。

國語的特性

中文語音是一字一音節，同音異形字[14]的字集相當多。此外由於中文語音存在有若干混淆集，如{八，搭，他，嘎，喀，....}，還有因發音習慣的不同所產生的音變現象 [13]如『倒塌』『他們』中『塌』和『他』皆是標示ㄊㄚ，但在台灣地區有相當比例的人唸出的音卻是不同的。

語言的最小基本語法單位都是詞，不過中文的詞在文句中並沒有像拼音文字一般以空格隔開，所以中文的自然語言處理便多了斷詞、配詞這個動作了。而且中文詞並沒有在字形或字音上表現出詞類變化的現象，所以很難直接從字形或字音得到所謂的詞類，所以中文詞性要絕對標示成為一件極為困難的事。

所以不論是國語語音辨認或是自然語言處理包容性(Robustness）都是十分重要的考量。

目前兩種實現國語聽寫機的方法

在目前關於國語聽寫機之操作模式大抵可以依處理之基本單元而分為兩類：

一.以句子為處理單元：由於句子含有最充足的語法訊息，所以有些聽寫機的設計是建構在整句處理的基礎上。以句子為基本輸入單元雖然可以避開『斷詞含混』的現象，但是由於以句子為處理單元，必須負荷比以詞為基本辨認單元更大的計算量，所以必須要有良好的硬體設備配合，而且只有在辨認正確率極高的情況，方能做有效處理，對於聲音較不穩定或者是發音習慣較為特殊者，以及一般無法提供相當硬體設備的情況下，可能都會造成實用上的不便。

二.以詞為處理單元：不論從語言學或心理學的觀點，詞皆是表示意念的最小單位，在自然語言的處理上，也是以詞為最小之語法單元，所以以『詞』為基本辨認單位看來也是十分合理，而且由於詞為處理單位可以容許辨識模組誤差，所以在現階段實用上是比較適當的。可是由於詞的定義一般人總是不能十分明確地定義出來，所以以詞為基本辨認單元的國語聽寫機必須要能克服斷詞含混的問題。

而在本文中即是就以詞為基本處理單元之運作模式加以研究，希望能透過一些方法在不影響系統即時處理以及保有系統實用性的原則下，將處理單元由詞提昇至詞組甚至未來可能由下而上發展至整句處理，以達成中文聽寫機的理想。

以下我們就以句子"在台灣大學生活像白紙"的詞絡（word lattice）來比較整句處理跟將句子作合理的斷句後其可能路徑的多寡，可以見得加入一些斷詞的知識是可以降低複雜度，使系統在處理上速度更快。

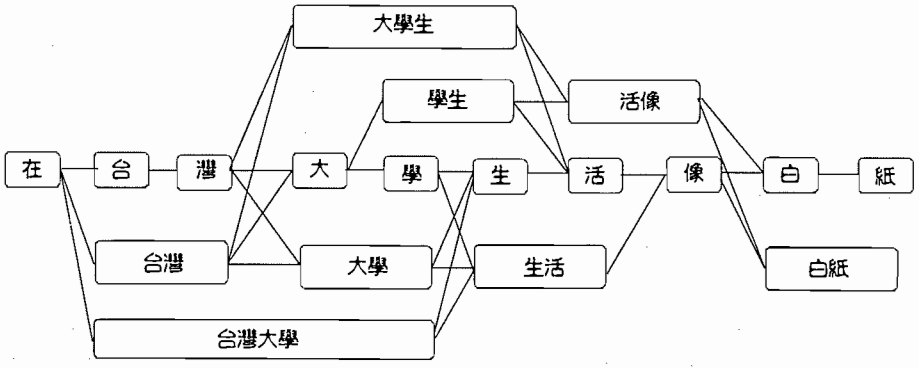


圖 2 以整句處理模式下的詞絡

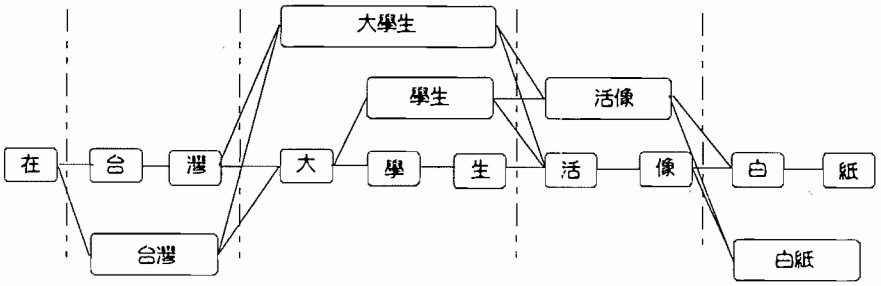


圖 3 存在合理斷點（詞）的詞絡

音中仙系統架構

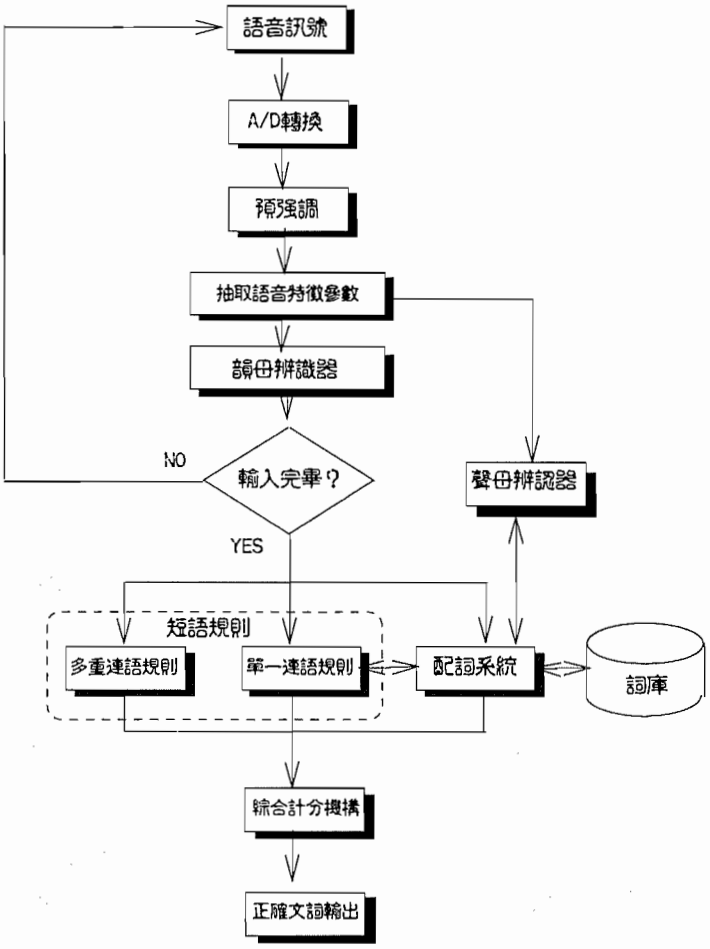


圖 4 音中仙國語聽寫機之系統流程

"音中仙"語音辨識器簡介

音中仙之辨識模組是利用拜氏網路[8]來估算參考樣本及測試樣本間的相異度，此網路是利用混合高斯機率密度之觀念來完成拜氏分類法則，以求得輸入特徵向量與參考類別間的相似機率值，其架構如下：

拜氏網路基本上是以拜氏定理為理論基礎，在架構上可分為輸入層、高斯層、混合層、歸納層。輸入層為待辨識的語音音框的特徵參數，高斯層是由統計訓練樣本的分佈情形所形成，混合層是一種混合的高斯機率分佈，而歸納層的輸出就是把混合層的機率轉換成距離輸出，在根據此一距離輸出來取辨認音節輸出。

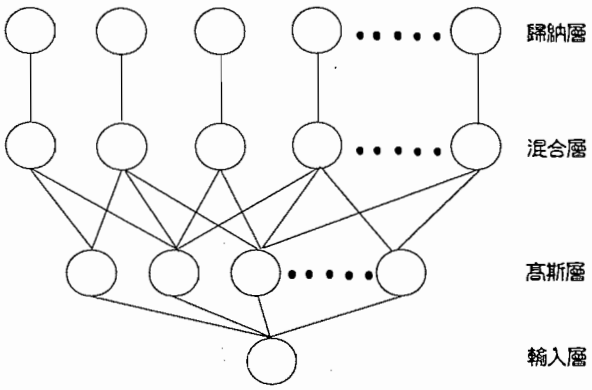


圖 5 拜氏網路

配詞機構

配詞機構的目的是在系統需要從辨識候選音節得到可能對應之文詞時做配詞的工作。由於音中仙具有大量詞彙處理能力的即時系統，所以必須具備快速準確的配詞能力。

在記憶體的考量下，我們不可能將所有整個詞庫載入系統中，所以音中仙的詞庫結構可分為兩部份，在圖中上方是用來配二字詞的結構，下方則是用來配長度三以上之長詞（含三字詞）。雖然在結構上有所不同，但是基本上還是以詞的前兩個音的韻母組合為指標開始搜尋，我們將這樣的指標以串列實現。而在系統執行時載入記憶體中也就只有一些串列跟指標而已，至於詞庫是儲存在磁碟中的。

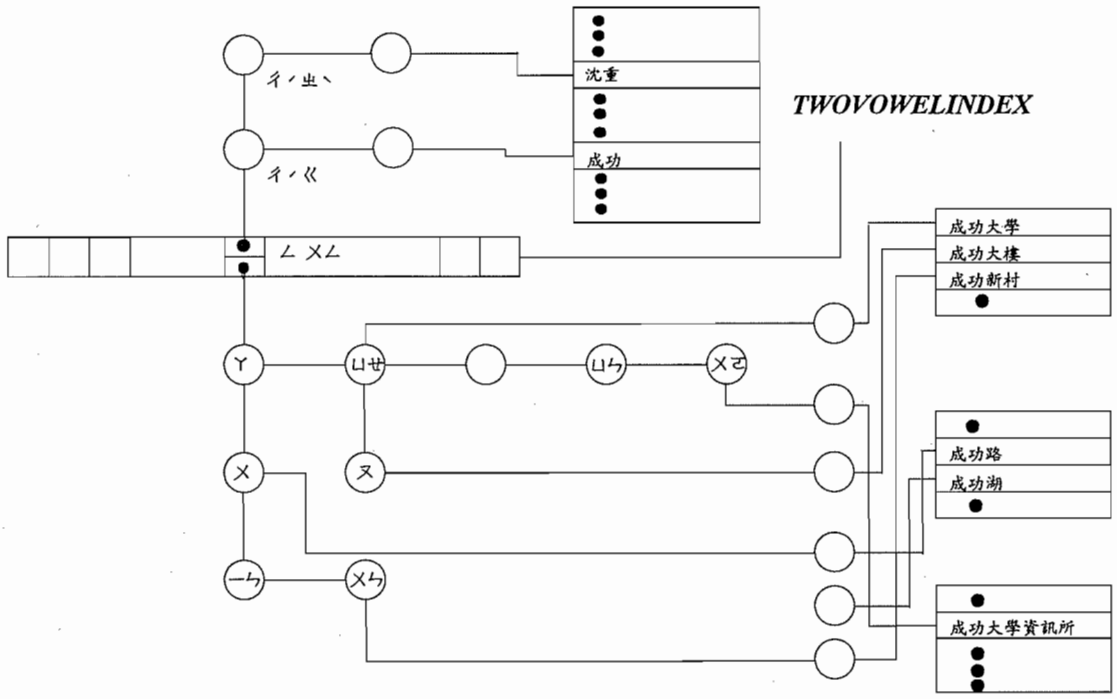


圖 6 音中仙系統之配詞機構

在一字詞語二字詞之辨認上，音中仙也做了一些處理，由於短詞語音所含的語言資訊有限，並且因為混淆音集的存在，我們要發展一套高包容性的系統，便需要對子音混淆集做歸納，而在系統處理時對辨認音節處理外也將其所屬混淆集的音節也一併處理，但是如此一來又可能會產生候選詞過多的窘態，所以對於音調的處理變成了重要的一環，不過音調處理在單字詞還算很不錯，不過在多字詞時由於轉折音的存在及發音習慣的不同，我們在做辨認後處理時也是必須考慮到音韻規則[10]才可以更準確地得到所想要輸入的字詞。關於四聲音韻在辨認上可以歸納如下：

聲調辨認結果	推測可能辨認結果
第一聲	第一聲，第二聲
第二聲	第二聲，第三聲
第三聲	第三聲，第四聲
第四聲	第四聲，第三聲

表 4 音韻處理表

二、 單一詞未知短語規則的分析與建立

短語規則主要是結合統計式[11][12]的文法觀念，利用電腦自動統計的方式產生，也就是在欲處理的規則中給予統計的頻率評分，如此可以便免統計式大量參數資料之不足[13]，也可以隨狀況分析隨時加入新的規則，而不像傳統剖析器一般必須在全盤考量後制訂，制訂後的修改又會往往造成系統維護的困擾。

在制訂規則時，必須考慮到一些問題如辨認模組的候選音過多，或輸入不是十分合法的字詞組合，新詞的產生或舊詞新用，甚至是隨著語言演化所產生之新文法新句型等，所以我們在制訂規則時，必須考慮到其包容性。

在雙連資訊的統計中，發現大部分的參數值都是零，連語現象(Collocation)〔15〕是十分強烈的。在作為語音辨認後處理時，可以利用樣本比對（Pattern matching）〔3〕的方法，逐步由下而上（bottom up）〔13〕地組織，如圖七所示為單一詞未知短語規則建立流程。

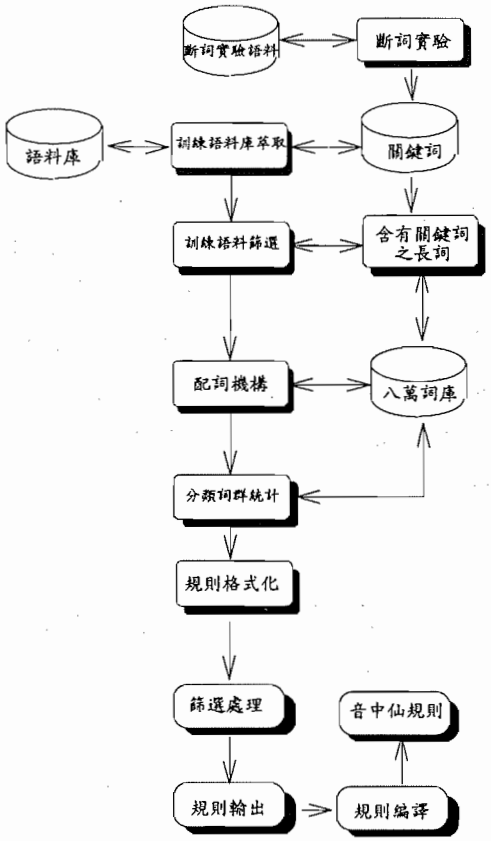


圖 7 單一詞未知短語規則之建立流程

斷詞實驗

誠如前面討論，以詞為辨認單元是可以降低系統複雜度以達成及時處理的程度，但是以詞為基本輸入單元可能因為個人斷詞習慣的不同而導致斷詞岐異的現象[4][6]，此外有些詞的數量是詞庫無法完全處理的如數量詞、複合詞等。針對斷詞歧異我們做了以下實驗：

我們分別就以下的五種文章，選取國小課本、報紙新聞部份（不含廣告）、小說、科學報導雜誌以及討論中國文學的『中國學術通義』的漢語文資料各約六百至八百字，分別讓國高中學生共六人，理工科大學生四人，文科大學生三人，還有研究生三人共十六人作斷詞分析。

以下所得之正確率乃指上述由人斷詞結果在音中仙系統詞庫中直接配詞，可以配得正確詞的漢字數除以所有測試漢字數總和。

	國小課本	報紙新聞	小說	科學雜誌	中國學術通義
正確率	89.76%	83.55%	81.03%	78.67%	79.98%

表 5 斷詞實驗之正確率

在實驗中，發現以詞為輸入單元，有相當之可行性。而發生斷詞錯誤之原因主要有下列幾個原因：

1.中文詞綴的引入，如『股票族』、『未登記』。

2.數量詞：定詞 + 量詞，如『三百五十公噸』。

3.介詞與連接詞：如『受折磨』，『戰爭與和平』。

4.複合詞：如 『國科會主委』。

5.人名：如 『王小明』。

6.古文：如 『猝然臨之而不驚』。

7.翻譯名詞：如 『關貿總協』。

8.外來語：如 『DNA』、『去氧核糖核酸』。

訓練語料庫萃取

根據斷詞實驗所得之結果，將一般斷詞容易發生錯誤的情況，將其關鍵詞訂出。根據所得的關鍵詞去訓練語料庫(計算語言學會提供之語料庫)中將含有關鍵詞的語料萃取出來，在此時我們所採的統計視窗大小為十二，即是在關鍵詞前後各有六個漢字，因為一般在四字詞以上的長詞不容易發生斷詞錯誤的情況。假如我們分析的關鍵詞是『受』這個詞則在語料庫萃取所得的訓練語料如下所示：

後桃園地檢署 受 理本案，關於

隔壁的老婆婆 受 無情的兒子拋

種不孝的行徑 受 大家指責。

訓練語料篩選

由於有些關鍵詞可能會是其他較長詞的一部份如：分析『受』這個字，訓練語料會將含有『感受』的語料也一併加入，所以我們必須將這些狀況排除，在這部份我們在系統詞庫中將含有關鍵詞的詞找出來，在根據上下文判斷是否是屬於我們分析的關鍵詞狀況。

另外也必須檢查關鍵詞位置是否為其他詞之局部所構成，這是補償前面可能的漏失。如關鍵詞為『經過』，而訓練語料為『...已經過去...』，此時我們將以『已經』及『過去』這兩個詞來和關鍵詞『經過』來比較，若相差不多，就以人力判斷，不過在實作的過程這樣的狀況非常少。

配詞

這是根據我們分析的關鍵詞前接詞跟後接辭去詞庫中配詞，這一階段配詞，首先在關鍵詞兩側設立絕對斷點然後對訓練語料其他部份斷詞，而以長詞優先，高頻詞優先的順序作為斷詞依據。

分類詞群

此外我們還必須對詞庫裡的詞作分類，我們以分類詞群[14]做為馬可夫統計模式的基礎主要的原因有二：

一、避免由於詞庫內詞數過多造成統計結果過於龐大。

二、可以藉分類詞群來調整語料庫不夠平衡的弊病。

而分類的標準主要是根據計算語言學會的詞庫之語法標示以及語意標示共分為兩百類。在分類詞群的類別數目上，我們發現若分類太多會造成短語規則過於冗雜，包容性會降低，但是所得的短語規則較為精確。若分類太少，會造成合於規則的輸出過多，容易造成混淆現象。由我們實驗的結果，發現有一些混淆的現象，可見分類數還不夠多。

馬可夫統計模式的引用

在自然語言處理中，馬可夫統計模式已經十分廣泛地被使用。在馬可夫模式中也是根據字與字之間或詞與詞之間的連接機率，在節省記憶體和處理速度的考量下，一般來說，一階馬可夫語言模式也就是雙連關係（bigram）由於在實作上容易，且在語音辨認後處理上能夠維持相當的水準，所已被廣泛接受。

輸入詞串： $W = w_1 w_2 w_3 \ldots\ldots w_n$

W的發生機率為：$P(W) = \prod_{i=1}^{n} P(w_i | w_1 w_2, \ldots. w_{i-1})$

雙連關係 ：$P(W) = \prod_{i=1}^{n} P(w_i | w_{i-1})$

其中在參數的訓練有人用純粹的機率統計模式，也有人用相對資訊（mutual information）來處理，而在本文中我們是用計算連接次數來處理，主要的原因是為了在日後有所調整時，不至於喪失原有的一些資訊。

單一詞未知短語規則之輸出格示

經過上述步驟之處理，系統的輸出格式如表 6 說明包括終止項(terminal term）以及非終止項(nonterminal term)，所謂終止項即是只在系統詞庫中可以配到詞的輸入，而非終止項所稱的就是經由短語規則處理之後可以得到的詞組輸入，也就是說短語規則的作用就是將非終止項轉為終止項的組合。至於混合項所描述的便是可以直接由配詞得到，也可以由多個詞經由短語規則所生成的單元。

【例一】

可愛的綿羊

{@可愛的$$ㄎㄜˇ ㄞˋ ㄉㄜ•〔(2)<53><893>〕〔(3)<53><122>〕}

【例二】

中華民國84年

{@中華民國$$ㄓㄨㄥ ㄏㄨㄚ ㄇㄧㄣ ㄍㄨㄛ〔(1)<0><23>〕〔(2)<0><210>〕〔3〕<0><189>〕#年$$ㄋㄧㄢˊ}

表 6 短語規則中輸出格式之符號定義:

{ }	：規則開始與結束
@ #	：關鍵詞國字部份起始位置
$$	：關鍵詞國字部份結束，注音部份開始
〔 〕	：非關鍵詞之分類詞群區塊
	〔(A)<x,y,z,...><X,Y,Z>〕
	其中 A 代表詞長
	x,y,z,.... 代表分類詞群
	X,Y,Z,.... 代表統計分數
	x,y,z,....與 X,Y,Z,....有相對應之關係

篩選處理

這樣產生出來的規則，是具有統計上的意義。不過系統可能會因規則產生的合理組合太多而導致速度或記憶體上的負擔，所以在這裡我們做了一些篩選動作。篩選的對象即是在規則中所佔分數比重較低者，以及在口語習慣上較不會出現之組合，例如在詞頭詞綴我們在規則中就專對其後接詞統計來做處理，而將其前接詞產生的規則略去。

音中仙短語規則之產生

在前一節所討論的都是屬於靜態單一連語規則的建立來討論，在本節是針對音中仙辨認模組與短語規則結合後的運作，加以觀察討論。

在音中仙國語聽寫機中所用的詞庫，是由計算語言學會提供的八萬詞目詞庫加以整理所得到的。而如此大量的資料，假如規則沒有加以處理，在做搜尋（serching）跟比對（matching）時都會造成系統在速度上嚴重的負擔，在詞庫方面我們是根據詞的音來做索引，而在規則庫方面我們用了一個規則庫編譯器，這個編譯器的作用主要是將我們所制訂的規則編譯成音中仙辨認模組所處理的型態，而這種型態的改變主要的目的便是加速搜尋跟比對。

而這個編譯器運作的原理主要可以分為兩部份說明：

關鍵詞語：在關鍵詞語的處理上主要是根據關鍵詞的注音與字的位置，將類似的規則放置在一起。而且如果關鍵詞語越長的，我們在評分機構中將會得到越高的評分。

非關鍵詞語：在非關鍵詞的比對中，首先我們的短語規則編譯器會把在規則中含有相同位置、相同詞長以及類似的分類詞群之非關鍵詞語的規則集中在一起，在辨認模組運作時，便可以省去大量比對的時間。

實驗結果

我們從中國時報二至四月份的報導中，取得1791個含有介詞之測試樣本，在原有系統中加入以六十四類共一百餘個介詞為關鍵詞的規則（按語言學會提供之八萬詞庫所標示詞性）共一千零七十五條，是否可以合理地組合出我們所要輸入的測試詞組，以下便是利用規則庫組合的辨認率：

	前三名	前五名	前十名
正確率	58.42	64.30	72.33

然後我們討論規則庫的加入對原有系統詞的影響，分別在有使用規則庫與沒有使用規則庫兩種情況來輸入中文的詞，這些測試的詞是由原詞庫中取得，而兩種情況所辨認的音串輸入是屬於同一個聲音，下面是實驗結果：

未加入介詞規則前原有系統詞之累積辨識率：

	第一名	第二名	第三名	第四名	第五名	第六名	第七名	第八名	第九名
三字詞	60.32	76.40	82.40	86.77	90.36	92.04	94.39	96.78	98.02
四字詞	62.12	76.33	82.02	85.52	89.98	91.03	94.00	96.20	97.18

加入介詞規則詞後原有系統詞之累積辨識率：

	第一名	第二名	第三名	第四名	第五名	第六名	第七名	第八名	第九名
三字詞	40.23	52.02	61.32	70.89	73.22	78.32	84.02	87.96	91.08
四字詞	57.08	67.01	75.18	81.56	86.00	87.26	91.06	93.01	94.63

由實驗發現在短語規則處理下，詞組輸出是十分可行的方法雖然有些詞的正確輸出排名會被擠到後面，原因分類詞群的多寡良窳對系統輸出的影響，這也說明了，我們分類詞群的數量還可以再區分為更多類，也就是說令每一類之間的特性更接近了，這樣便可以得到合理輸出又降低系統詞混淆的程度。不過在音很準(即每次皆只取第一名)的時候事實上這些問題都是不存在的，我們曾經就以正確音輸入，發現效果十分良好，不過在考慮實際狀況時，這樣的假設似乎有些嚴苛。

三、 多詞未知短語規則處理

不論是在詞庫或辭典的編定，有一些詞是不可能收錄完全，例如由定詞和量詞所合成的複合詞就具有無限制的常用性[9]。事實上我們可以利用一些法則從原有詞庫的詞來組合生成這些詞，在本節便是針對這一狀況利用狀態轉移的觀念來處理。

數字規則

數字可以分為"零"到"九"十個數字以及"十、百、千、萬..."兩類，而其中的轉移是有一定的規律的，在國語的數字系統中有其獨特的語法結構不同於西歐語言，經由我們觀察分析，中文數字是分為四位一小節，也就是說主要可以分析為兩階段，第一階段為在一萬以下的數字表示法，以及萬、億、兆...等，第二階段的表示法，也就是說我們可以將數字及千百十的組合規則定為一個狀態轉移，而這個狀態轉移將成為第二階段狀態轉移中的一個狀態，我們便可以利用兩階段的狀況轉移來表示中文中數字系統的組合。

而其中『零』會有兩種角色，一為單純數字中的零，一為十、百、千、萬、億、兆的位數省略表示，所以在狀態圖中將會對於這種省略表示的『零』特別給予一種狀態。，而這樣的觀念可以用下列狀態轉移圖來說明：

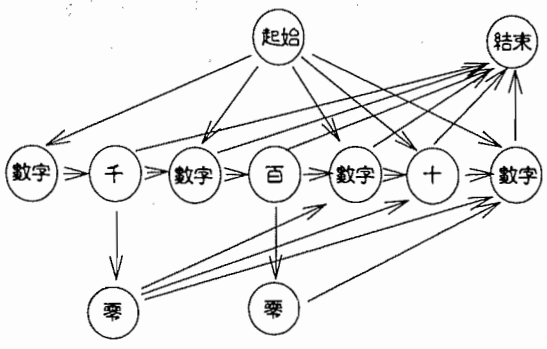


圖 8 第一階段數字狀態轉移圖

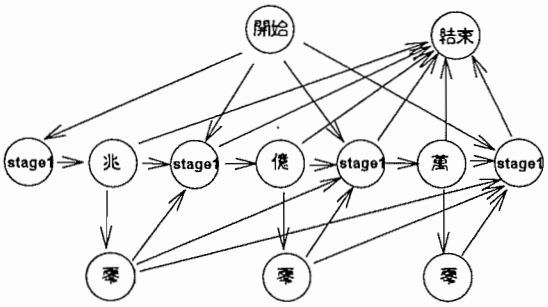


圖 9 第二階段數字狀態轉移圖

在前面所討論的數字系統是利用數字與十、百、千、萬等正規的寫法，事實上在日常生活中也有可能直接由數字來組合例如學號、電話號碼等等。

甚至有些狀況我們也習慣用數字串來取代前面所提的正規數字表示，而單純利用數字串的狀態轉移是十分簡單的如下圖所示：

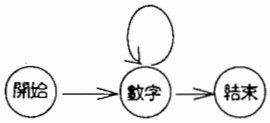


圖 10 數字串之狀態轉移圖

不過若要系統能兼顧正規數字表示以及數字串的表示方式，我們就要考慮到"四"跟"十"的混淆效應，基本上如果在數字表示所得的結果含有"十"之外的"百"、"千"、"萬"那就可以確定為正規數字表示，經由正規數字的狀態轉移便可以得到正確的結果。若在數字表示中找不到"百"、"千"、"萬"，也就是說我們並不確定輸入的是簡單數字串或是正規的數字表示時，我們就必須對四跟十混淆狀況討論。因為在四字以上的數字串我們可以明確地決定是屬於哪一種表示方式，所以首先我們在三個數字串的考量在三個數字串的合法組合中，如果在第二個數字產生混淆時，我們就依辨認音節給分，因為事實上"八十一"跟"八四一"都是合理的數字表示，如果"四"、"十"混淆不是產生在第二個數字也就是數字系統中的十位數，那我們就可以確定是四。而在二位數中如果不是十十的組合，基本上我們都認為是合理的，而一位數字更不用說四跟十都是合理的數字。而誠如前面所討論的作為語音辨認的後處理我們必須考慮到系統的即時處理能力，在本章中我們關於這類可以無限組合的詞組處理，我們採用最先最佳（first best）演算法來節省運算量，也就是我們從辨認模組中得到候選音節排名，根據該排名我們將音節應對至字，然後我們取屬於狀態的關鍵字前五名，之所以取前五名是因為在我們的實驗中，每個音取前五個字已經是綽綽有餘了，然後由每個音的第一名的字去組合，並進入狀態轉移去看看是否可以得到

合理的組合，如果合理便是我們要找的輸出（first best），而其他的組合因為速度的考量就不再進入狀態轉移中去運作了。如果在最佳路徑中我們無法得到合理輸出，我們會根據辨認音節權重去找第二條次佳運作的路徑，如此重複運作一直到找到一條合理輸出或是已經將關鍵字組合的路徑用完為止。不過若是屬於前面所提的"四"、"十"混淆，我們還是允許在輸出中可以得到兩條以上的輸出，因為事實上不僅訊號上類似，即使是人類也會常常誤判四跟十，所以再此我們特別將這一類混淆考慮進去。至於一般數字中常會混的"一七","六九"事實上由辨認模組輸出的加權便可解決，也就是說這樣的混淆在我們的辨認模組便以得到解決。

地址規則

在日常生活中地址是一項常用的資訊，舉凡個人資料的填寫、郵政金融等行政手續上，地址都是不可或缺的資訊，而數字更是一般常用的輸出入 項目。所以在這裡我們提出一套方法針對地址來處理。

在地址規則的制訂上，我們有必要瞭解在台灣地區的地址系統中可以區分為兩類，一類是必須項，如號；有一些是非必須項，如段、巷、弄以及幾號之幾。在非必須項中也有一定的關係，例如『弄』的出現，前面一定有『巷』，所以也是可以利用狀態轉移來處理的一種情況。

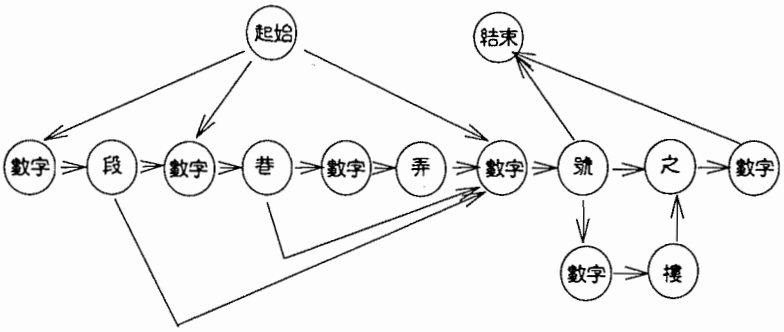


圖 11 地址狀態轉移圖

網狀規則

從人類的的斷詞經驗看來，有些句子可能會有多種合理斷詞，而這些合理斷詞中，一般而言是不會將一個詞切開，而是將幾個詞組合在一起。所以為了使我們的規則更具生命力，更具包容性，將短語規則所得之結果視為詞的種類，允許在規則中的非關鍵詞中還可以是其他的規則，這樣的網狀的規則(Rule Nesting），使得系統對於斷詞的包容性提升，也就是說原有規則產生詞組的生命力也隨之提升了。

而這觀念目前在系統中是用來實現單一連語規則以及多重連語規則的結合，而利用這樣結合的效果也解決了一些問題，未來如果規則庫建立完整後，便可以將規則分類，在歸入分類詞群，向上建構逐步增加處理的單位。

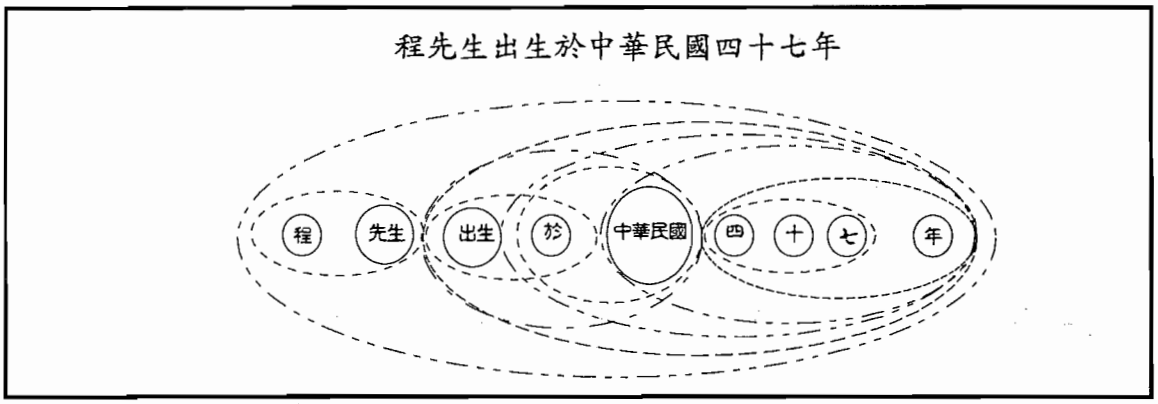


圖 12 網狀規則實例

定量複合詞

一般關於數量詞的描述，可以用定詞加量詞的語法分析來處理，事實上我們只要利用數字規則為非關鍵詞部份，就是說數字狀態規則是我們分類詞群標示中的一種，再利用量詞作為短語規則的關鍵詞，給予適當的分數，如此再依網狀規則的觀念，便可以處理定量式複合詞。例如我們說『五千元要買一頭牛』，其中『五千元』、『一頭牛』便可以利用定量式複合詞規則來組合輸出。而這樣的處理一來可以解決詞庫無法完全收錄定量式複合詞的問題，同時也較合乎一般人的斷詞習慣。

一般混合式規則

而在一般狀況下也有可能有某些字詞組合經常會夾雜數字串，例如"中華民國八十四年"、"第三課"等等，這一些我們也可能利用我們發展的兩種方式來產生其生成規則。首先還是針對前面第三章所描述的方法來處理所謂的關鍵詞部份，然後再利用數字狀態轉移方法來檢查非關鍵詞部份是否合乎條件，換句話說也就是說數字狀態轉移已經被列入我們的分類詞群中獨立為一類了。

實驗結果

【實驗一】我們分別就長度為三至九的正規數字，利用電腦亂數產生測試數字

所得之正確辨識率為：

	三字	四字	五字	六字	七字	八字	九字
第一名	88.10	85.76	86.02	85.30	83.76	81.14	78.02
第二名	6.52	6.06	1.24	2.32	1.68	1.02	1.68

表 7 正規數字狀態轉移辨識率

【實驗二】在一般純數字串的數字系統中，我們也是利用電腦亂數產生樣本，所測試的樣本數為

	三字	四字	五字	六字	七字	八字	九字
第一名	87.46	84.35	86.31	85.26	82.76	81.36	81.08
第二名	7.77	10.52	3.24	1.25	1.07	0.02	0.01

表 8 一般數字串之辨識率

從上面的實驗中我們可以得到幾個結論：

一）在音完全正確或是以鍵盤輸入的情況下，我們知道由狀態轉移的處理下，輸入串越長或是轉移過的狀態數越多，其出現在第一名的機率越大。不過在我們上面的實驗中發現在語音辨認中並不是如此，雖然隨著輸入音串的增長，合理的輸出會越少。不過也可能由於使用者在某個狀態的音不是準確的，而不能在狀態轉移中轉移至合法結束狀態，所以在這種情況下越長音串辨識率便越低了。我們曾經嘗試放寬條件來使得可以進入狀態轉移的組合多一點也就是在處理音的條件上放鬆了，不過這樣的結果雖然讓辨識率提升了，不過卻會使系統處理的速度降了下來，所以我們覺得在前面所提的條件在實用上是合理而可行的。

二）在上面兩個實驗中我們發現音串長度為四時，在第一名輸出的機率有降低的趨勢也就是說比音串長度為三小，同時也比音串長度為五來的小，不過在前二名的累積輸出機率便超越了音串長度為五的輸出。這樣的結果我們由實驗中的實作知道，這是因為在我們原有的詞庫中有些四字詞的成語是用若干個數字在詞當中，由於辨認音節權重的配分下，可能在配詞機構中會產生分數較高之輸出，所以就會把數字組合的名次往後擠了！這類的詞有『三頭六臂』、『四面八方』、『三人市虎』等，在我們去嘗試輸入這一些成語時，數字的組合也常常出現在我們的合理輸出中，不過由於辨認音節權重以及我們在狀態轉移時取最先最佳輸出（first best）的影響下，數字串的輸出並不會對系統詞造成妨礙。

四、 結論

在系統評估上，容錯率跟速度都是重要的考量，而我們在系統的發展上同時也考慮到這兩個特性，所以我們提出了將句子分為幾組詞組來處理的構想，希望從合理斷詞的想法上大量降低處理的複雜度，發展出一套在實用可行的系統。在單一詞未知短語規則的實驗中可以發現可以藉此來輸出詞組的組合，但是還是會有一些不良組合會影響系統輸出，這也表示分類詞群的研究還是有十分寬闊的發展空間。

而詞庫無法收錄的詞，從現實狀況觀察，歸納出一定的文法狀況，並藉狀態轉移的檢查篩選合理輸出。在這樣的考量下，我們從實驗中發現效果十分良好，而這一類的詞包括一些數字組合、一些複合詞甚至一些日常生活中時常會使用的樣板。

在這兩種短語規則的處理下，可以解決了中文詞綴、數量詞、介詞、連接詞以及一些複合詞的問題，關於古文、翻譯名詞、外來語以及新詞是可以利用整理詞庫的方式來解決，不過關於姓名的輸入一直是我們尚未能解決的問題。

參考文獻

[1] Ren-Yuan Lyu,Lee-Feng Chien,Shiao-Hong Hwang,Hung-Yun Hsieh , Rung-Chiuan Yang,Bo-Ren Bai,Jia-Chi Weng,Yen-Ju Yang,Shi-Wei Lin , Keh-Jiann Chen,Chiu-Yu Tseng,Lin-Shan Lee,"Golden Mandarin(III)- A User-Adaptive Prosodic-Segment-Based Mandarine Dictation Machine for Chinese Language with Very Large Vocabulary ,"ICASSP-95,pp.57-60,1995.

[2] Lin Shan Lee,Chiu-Yu Tseng,Hun-yan Gu,Fu-Hua Liu,Chen-Hao Chang, Yueh-Hing Lin,Yumin Lee,Shih-Lung Tu,Shew-Heng Hsieh,and Chian-Hung Chen, "Golden Mandarin (I)-A Real-Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary ,"IEEE Trans. Speech and Audio Processing, Vol 1,No. 2,pp158-179,1993.

[3] 許聞廉,陳克健,"『國音』智慧型輸入系統的語意分析",技術報告,中央研究院

[4] 陳克健,陳正佳,林隆基,"中文語句分析的研究-斷詞與構詞",技術報告,中央研究院,1986

[5] 謝子陵,"A Lingguistic Decoder for Mandarin Speech Recognition Using a Score Parser,"碩士論文,國立成功大學

[6] 張俊盛,陳志遠,陳順德,"限制式滿足及機率最佳化的中文斷詞方法",ROCLING IV,pp.147-165.

[7] Ming-Shih Chen,Tracy Yang and Hsiao-Chuan Wang,"Speaker Identification over telephone system based on channel-effect cancellation,"The Journal of the Chinese Institute of Engineers,IEEE,1993.

[8] Jhing-Fa Wang,Chung-Hsien Wu,Shih-Hung Chang,and Jau-Yien Lee, "A Hierarchical Neural Network Model Base on A C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition ,"IEEE Trans. Signal Processing,Vol 39,No. 9,pp.2141-2145,1991.

[9] 趙元任,"中國話的文法",中文大學出版社,香港,1980.

[10]羅肇錦,"國語學",五南出版公司,台北市,八十一年

[11]J.H. Wright,G.J.F.Jones and H.Lloyd-Thomas,"A Robust Language Model Incorporating A Substring Parser and Extended N-gram ," ICASSP,IEEE,Vol 1,pp.361-364,1994.

[12]J.H Wright,G.J.F.Jones and E.N.Wrigley,"Hybrid Grammar-bigram speech recognition system with first-order dependence model ," ICASSP-92,IEEE,Vol 1,pp.169-172,1992.

[13]Marie Meteer,and J. Robin Rohlicek,"Statistical Language Modeling Combining N-gram and Context-free Grammars ,"ICASSP- ,IEEE,1993.

[14]張元貞,林頌堅,簡立峰,陳克建,李琳山,"國語語音辨認中詞群語言模型之分群方法與應用",ROCLING VII,pp.17-34,1994.

[15]James Allen,"Natural Language Understanding,"The Benjamin/Cumming Publishment Comoany,Inc. ,Redwood City , CA, USA,1995.

[16]Lee-Feng Chien,Keh-Jiann Chen,and Lin-Shan Lee," A Best-First Language Processing Model Integrating the Unification Grammar and Morkov Language Model for Speech Recognition Application,"IEEE Trans. Speech and Audio Processing,Vol 1,No. 2,April 1993.

[17]McDonald, David D.,"An Efficient Chart-based Algorithm for Partial Parsing of Unrestricted Texts,"1992.

[18]Lee-feng Chien,K.J. Chen,and Lin-shan Lee,"An Augmented Chart Parsing Algorithm Integrating Unification Grammar and Markov Language Model for Continous Speech Recognition ,"ICASSP,IEEE, pp.585-589,1990.