中文剖析的問題與對策

Keh-jiann Chen

陳克健

Academia Sinica

中央研究院

Proceedings of ROCLING I(1988)
R.O.C. Computational Linguistics Workshops I pp 19-28

中華民國第一屆計算語言學研討會論文集 19-28 頁

中文剖析的問題與對策

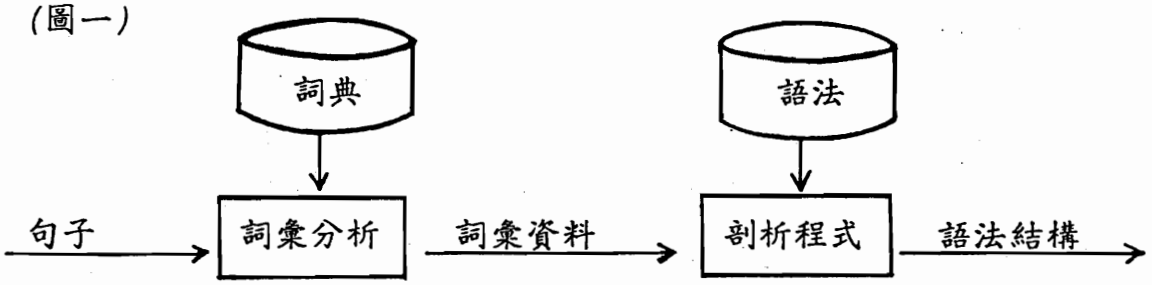
中央研究院　資訊科學研究所

陳克健

第一節　中文剖析簡介

電腦剖析自然語言大致上分爲兩個階段，第一個階段是詞彙分析(lexical analysis)，第二個階段是語法結構分析(syntactic analysis)。詞彙分析的過程，主要工作是將待分析的句子中的詞彙相關語法語意資料，利用詞典及詞彙規律找出來。這些語法語意資料成爲第二階段語法分析的輸入資料，電腦利用語法規律分析輸入資料得到句子的語法結構。這個過程可以簡單的用圖一表示。

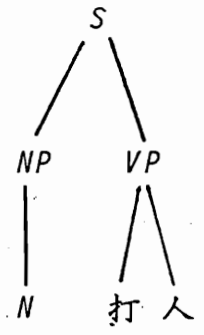
(圖一)



張三打人

張三 . 打 . 人

N　V　N



張三　V　N

在剖析句子的過程當中，因各國語言不盡相同，處理的細節部分就不太一樣。中文可以說是比較特殊的一種語言，它有一些特性和其他西方語言很不相同，因此不能一成不變的將處理西方語言的方法拿來處理中文。以下我們對剖析中文有關的特性做一個簡單的說明。

(1) 中文的句子是由字串所組成，詞與詞中間沒有詞間標記。因此在詞彙分析時多了一層斷詞及構詞的工作。由於複合詞的合成非常容易，使得斷詞不僅僅是查詞典的工作，構詞律也變得非常重要。

(2) 中文幾乎沒有詞類變化、格位變化、單複數變化及時態變化，比較難從符號上或形式上得到語法訊息。這是分析中文句子語法結構比西洋語言困難的理由之一。

(3) 中文句子結構複雜，且鮮有格位標記，較難掌握主要動詞及格位角色。

(4) 中文詞彙多義的情形比較少。因此，一旦得到結構再分析語意時，會比西方語言容易。

本文將針對中文的這些特點，討論電腦剖析文句的困難及對策。

第二節　詞彙分析－－ 斷詞與構詞

詞彙分析的階段中，電腦接受一列中文字串，必須在輸入的字串中找出構成此一句子的詞彙，並提供句中每一詞彙的相關語法資料。在此一階段中所遭遇的主要問題有：

(1) 詞典的設計

(2) 斷詞的方法

(3) 複合詞與構詞

(4) 斷詞含混的問題

2.1 詞典的設計

詞典中應包括所有的詞 (words)，但是辦不到。因爲詞集合是一個開放集合 (open set)，可以隨時增減。新創的詞彙或複合詞極易產生，因此需要一個詞典的維護系統，隨時增減詞彙。詞彙集合(lexicon)應包含字、詞(包括專有名稱、簡稱、成語等)及部分的複合詞，特別是很難用規律產生的複合詞或詞義詞類語法訊息無法從組合成分得到者，均應列入詞典中。單字包含單字詞及非自由語素(bound morpheme)，非自由語素爲構詞成分，當然應該列入。

2.2 斷詞的方法

斷詞的基本方式大多採用匹配的方法，將詞典中的詞彙和輸入的句子匹配，找出句子裡頭隱藏著的詞彙。詞頭字及詞尾字的使用，有時可以增進處理的速度及正確率。最長匹配法則，大多時候應用無誤，但不是百分之百正確。比較嚴重的兩個問題是：(1) 詞典中找不到匹配的詞 (2) 一個字串有許多種不同的斷詞方式。

詞典中找不到匹配的詞，可能的原因有二：一個可能是複合詞，另一個可能是新詞或專有名稱。如果複合詞能夠處理得很好，詞典隨時增加新詞，剩下來的成分八九不離十是專有名稱(或簡稱)。

2.3 複合詞和構詞

複合詞的組成非常自由，只要語意和語言習慣允許，任何字串都能很自由的組合成複合詞。因此我們能夠處理的方式是，非規律性的組合都放入詞典中，特別是複合詞中不符合詞尾中心律者 (除動補、動賓結構外)，皆應收入詞典中。有規律者一部分應在詞彙分析時，以構詞處理，例如數字、時間、動詞重疊等。一部分應在詞組律中處理，例如名方式、定量式、動賓式、動補式、名名式等。

2.4 斷詞含混的問題

由於中文構詞容易，可能的斷詞方式會很多：例如"組合成分子時"相鄰任何兩個字都可以成詞，如何找出正確的斷詞並不容易，這個和語法、語意的知識有相當密切的關係。也有利用統計機率的方法——鬆馳法 (relaxation)，有用簡單規律(rule-based)幫助判斷的方法，不過基本上這些方法都和語法、語意有關。如何增加電腦這方面的知識，應該從實際處理著手，慢慢發現問題，尋找對策，漸漸的改善系統的正確率。

2.5 參考資料

有關詞典的設計，可參考[13,16,19,21](機器翻譯的詞典設計和此處討論者不同，不在本文討論之範圍)。斷詞的方法有[15,16,18,19,21]，字詞定義可參考[9,23,25]，複合詞的討論包括[8,9,22,23,25]。解決斷詞含混的問題可參考[15,18,21]。另外一個有趣的題目是電腦自動選詞的問題，目前沒有很好的答案但有嘗試的結果[5]。

第三節　語法與句子剖析

中文幾乎沒有詞類變化，詞的分類與語法的表示方式可能與其它語言不盡相同，本文將針對剖析過程可能遭遇的主要問題做一個分析，試圖找出可能的答案。至於語法上的細節問題不在討論範圍之內。以下我們分別討論詞的分類問題及中文語法表達問題。

3.1 詞的分類問題

如何分類？基本上分類的作用是將具有相同語法特徵的詞彙分爲一類，目的是爲了描述上的簡便性。然而中文欠缺詞類變化，如果純粹採用語法特徵做爲分類標準，可能產生許多不必要的多重分類。例如動詞的主要功能是扮演句子的謂語，但也可以名物化做爲主賓語，同時也能成爲名詞的修飾語。形容詞和動詞幾乎有相同的特性。名詞則少有謂語性，但成爲名詞的修飾語是毫無問題的，因此三大詞類的分類就是一個頭痛的問題。我們的看法是實詞的劃分應依照語意爲中心，以語法特性爲輔的分類方式。至於功能詞當然以語法特徵爲分類標準。各個詞項的語法特徵除了分類特徵之外，應該還有其它的非分類特性，應該用特徵集合(feature structure)來表示，以避免太複雜的次分類問題。例如動詞的格框(case frame)幾乎公認是分析中文不可缺的特徵，名詞與量詞的配合，詞的語意分類等，也是不可缺的特徵項目。

如果三大詞類的劃分依照語意爲中心而不以語法功能分類的結果，和一般大家心目中的動詞、名詞、形容詞不會有太大差異，但對電腦而言，因各個詞類有多重語法功能的傾向，如何以語法律來描述合法的句子或片語結構成爲一大問題。解決的方法必須依賴語意的幫助。例如動詞或形容詞的名物化，只可能在某些動詞的主賓語出現。這些動詞的主賓語的語意限制可能是事件、題目，狀態等。幾乎每一個動作動詞都可能形成一個事件，只是有些動詞習慣上必須有主語或賓語，以句子或動詞片語的形式出現。有些動詞可以獨立形成事件，例如'游泳'、'開會'、'打架'、'結婚'等。至於形容詞大多數是描述一種狀態，因此像動詞'保持'的賓語的語意限制就是狀態，因此像'美麗'、'清潔'、'完整'等就會名物化，成爲'保持'的賓語。

我們的結論是有規律的語法特性不必重覆分類，除了以上的例子外，再舉一個明顯的例子。形容詞幾乎都可以有副詞性，因此也不必再多重分類爲副詞。

3.2 語法與剖析

語法表達及剖析問題的產生，主要的原因是:一、詞的多重分類。二、句子的複雜結構。多重分類產生大量的可能片語或句子結合方式，增加剖析的複雜度，甚而產生含混的語法結構。句子複雜性的來源有:(1)主題化 (2)賓語前提 (3)連動、並列、兼語、動補等複雜句型(4)謂語論元的增減(5)名詞片語結構複雜等。另外標點符號濫用也造成剖析的困難。因以上幾個因素產生的問題有:

(1) 剖析的處理單位是什麼?

由於欠缺句子的定義及明確界限，剖析系統的處理單元可能是逗點、句點或分號之間的片斷;也可能是一個句子或一個片語。因此,也許系統的設計應配合實際的現象，考慮處理的單元爲句子或片語。剖析完成後，再經過較高層次的分析，得到複雜句或段落的分析。

(2) 語法表達的問題

由於句子結構的複雜性及多變性，如何選用語法模式能簡明的表達中文句型，是一個值得研究的問題。近代語法模式理論所採用的非序性支配律(ID rules)及序性律(linear precedence rules) 將成分結構及成分排序的規律分離的方式，可能是可以參考的原則，另外聯併化 (unification)及詞彙中心的處理方式及分離必要成分和非必要成分(Adjuncts)的方法也可以簡化語法表面的複雜程度。

另外再强調一點語法模式宣稱性(declarative) 的重要性。由於語言學家和計算機專家之間欠缺共通的背景知識，很難溝通，因此語法的表達必須和電腦處理程序無關 (independent)。語言學家只要依據語法模式表達語法，計算機專家只要依據模式結構設計剖析程式,語法內容的改變不影響剖析程式的設計。

(3) 角色的指定問題

主語、賓語的語法功能角色在中文裡很難加以明確的定義。一般比較傾向於如何指定語意功能角色，如主事者、受事者、工具、地方等角色的指定，涉及語法及部分的語意分析。困難的地方在於如何找到正確的成分結構及角色的指派。

角色的指派和動詞、角色的位置、角色的意義都有關係。因此必須從動詞的格框、句型結構，詞彙語意分類著手。 另外，介詞片語時間成分的分析也是缺一不可。

(4) 語法結構上的含混

語法結構含混產生的因素很多，有因詞多重分類而產生的，有因詞類多重功能而產生的，有原始語法含混的。例如"瘋狂的攻擊"可以是副詞修飾動詞，或者是形容詞修飾名詞。這個例子可能是多重分類或多重功能的結果。又例如 N-V-N-V-N 可以組成至少七種以上不同的句子結構，產生的原因是多重功能及原始語法含混。

爲了解決結構上含混的問題，除了語法的整理及詞彙訊息的加强之外，很顯然的，部分語意分析是必然不可缺的。另外尋找或發展適當的語法表達、處理方法及系統設計上的對策也十分重要。

3.3 參考資料

中文剖析系統比較全面性探討的文章有[14,20,21]。詞彙分類的問題，語言學家談論的很多。而針對剖析而談的，包括分類的評估標準有[11]，分類的原則有[3]，分類的類別有[1,4,20,23]。至於特徵集合的詞彙特性表示方式，可以參考[7]，詞彙的分析結果[17,23]。語法模式的研究，聯併化(unification)爲主的語法研究，可以參考本研討會另一篇研習講題[24]有豐富的參考資料。中文剖析的問題及解決的技巧，散見在不同的論文中，例如[2,4,10,14,20]。一般性的剖析問題及方法也在本研討會的另一個研習講題談及[12]。關於自然語言處理的參考書目可以從[6]獲得詳細的資料。

本文之研究得國科會專題研究計劃(NSC-77-0408-E001-01)及中央研究院與工業技術研究院電子工業研究所合作之「中文詞知識庫第三期及中文語句剖析系統第一期合作研究與開發計畫」之部分補助，特此申謝。並感謝中央研究院計算中心協助打字及排版。

4. 參考書目

1. Chang, L.L. et al. "Classification and Cooccurence Restrictions in Chinese Simple Noun Phrases," ICCPCOL'87, Chicago , pp107-110.

2. Chen, C.G. K.J. Chen, L.S. Lee, "A Model for Lexical Analysis and Parsing of Chinese Sentences," Proceedings of 1986 International Conference on Chinese Computing, Singapore, pp33-40,1986.

3. Chen, K.J., C.R. Huang, L.L. Chang, "Word Classification and Their Relations to Grammatical Representations in Chinese," in preparation.

4. Chen, K.J., L.L. Chang, C.R. Huang, and C.C. Hsieh, "A Classification of Chinese Verbs for Language Parsing," ICCPCOL '88, Toronto, pp414-417.

5. Choi, A., C.H. Cheng, and Y.L. Ko, "Word Extraction from Chinese Documents by Occurrence Counts," ICCPCOL'88, pp488-491, 1988.

6. Gazdar, G., A. Frang, K. Osborne, R. Evans, Natural Language Processing in the 1980s, CSLI Lecture Notes #12, 1987.

7. Gazdar, G. et al. "Category Structures," CSLI report #102, 1984.

8. Huang, Shuanfan, "Chinese Morphology:anatomy of a double-headed language," the Second International Conference on Sinology, 1986.

9. Li and Thompson 著，黃宣範譯，漢語語法，台北文鶴書局，1980。

10. Lin, L.J., J. Huang, K.J. Chen, and L.S. Lee, "A Chinese National Language Processing System Based upon the Theory of Empty Categories," Proceedings of AAAI'86, pp1059-1062, 1986.

11. Shiu, Y.L. and K.Y. Su, "Criteria for the Classification of Lexical Categories in a Syntax-First Parsing System," Proceedings of ROCLING I, 1988.

12. Su, K.Y., "Principles and Techniques of Natural Language Parsing : A Tutorial, " Proceedings of ROCLING I, 1988.

13. Tseng, S.S., M.Y. Chang, C.C. Hsieh and K.J. Chen, "Approaches on an Experimental Chinese Electronic Dictionary, "ICCPCOL'88, Toronto, pp371-374, 1988.

14. Yang, Yiming, "Studies on an Analysis System for Chinese Sentences ," Ph.D. Dissertation, Kyoto University, 1985.

15. Yeh,C.L., H.J. Lee, "Rule-Based Word Identification for Mandarin Chinese Sentences," ICCPCOL'88, Toronto, pp432-436, 1988.

16. 何文雄，中文斷詞的研究，國立工業技術學院，工程技術研究所，碩士論文，1983。

17. 呂叔湘，現代漢語八百詞，第三版，香港:商務印書館，1987。

18. 范長康、蔡文祥，"以鬆弛法作中文斷詞"，全國計算機會議論文集,台北，pp423-431，1987。

19. 梁南元，"書面漢語自動分詞系統-CDWS"，中文信息，pp44-52，1987。

20. 陳克健、林隆基、陳正佳，中文語句分析系統的研究-語法及剖句，中央研究院資訊所，技術報告 TR-86-006，1986。

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

22. 陳克健、張麗麗、張莉萍、謝清俊，"國語中的複合詞和語言剖析"，全國計算機會議論文集，台北，pp415-422，1987。

23. 張麗麗等，國語的詞類分析(修訂版)，技術報告0002，中央研究院計算中心，1988。

24. 黃居仁，"聯併(Unification):語法理論與剖析"，ROCLING I，1988。

25. 趙元任著，丁邦新譯，中國話的文法，香港中文大學出版，1980。