超長句之剖析

李炳煌，邱垂豐，黎偉權

工業技術研究院電子工業研究所

摘要

當剖析器剖析某些結構複雜或字數過多(如＞40字)的所謂「超長句」句子時，容易造成所需時間過長（如＞3分鐘），本文主要提出對此類句子的剖析對策，以及如何兼顧超長句在英中機譯系統之中文生成問題。此外，此一剖析對策亦可有效處理某些固定型態的英文句型句子。

1. 問題

所謂「超長句」就是該句子的剖析時間超過某種不可忍受的合理程度，如時間大於 3分鐘之謂，一般造成此一現象之原因主要有二：

①句子長度(字數)過長。
②句子結構複雜存在著許多可能的句子結構，此一現象導因於過多的連接詞、介詞片語、多詞類 ...等等。

由於句子的剖析可視為一種搜尋問題(Search Problem)，當分岔參數（branch factor，代表句子結構複雜度）及搜尋深度（Search depth，代表句子長度）增大時，很容易造成組合爆炸（Combinatory Explosion）。在 all-path bottom-up parser〔1〕來講，情況更是嚴重，雖然 some-path bottom-up parser〔1〕可經由縮小分岔參數稍微疏解其搜尋的時間，甚或是使用 beam search方式〔2，3〕限制其分岔參數為一上限，但搜尋時間基本上仍為指數成長(Exponential grow)，而且使用 beam search 仍須考慮錯誤回復（recovery）的問題。

另外根據富士通 ATLAS II〔4〕測試結果知平均字數在 33.5 以下譯文皆可理解，在 33.5~45.7 間譯文稍作修改即可使用，45.7 以上則譯文完全不可用，因此超長句的譯文改善也是問題之一。

2. 文句庫

ERSO自1989年 1月開始接受客戶進行 Beta-side 測試工作，主要以「資訊電腦」及「環境污染」領域文章為翻譯主題，初期抽取空氣污染方面相關論文之摘要作為翻譯題材，共 871 句，19539 字，平均句長 18.3，超過40 字數的句子約 42 句，佔 4.8%，詳見表 1。一般來說，摘要的句子結構都不很嚴謹、且長，甚至發現其中有少數罕見的語法，因此不似書本來得正式、嚴謹。

總 句 數	871
總 字 數	19539
平 均 句 長	18.3
＞40 字句數	42
＞40 字 %	4.8

表 1. 文句庫統計資料

ERSO-ECMT 翻譯速度如圖 1 所示，平均 34 字以下，其翻譯速度在 1分之內，超過 40 字則翻譯速度呈指數成長，因此，若要控制句子翻譯時間在可以忍受的合理範圍內，則需限制句子長度，也就是說除了降低其分岔參數外，仍須減少其搜尋深度。

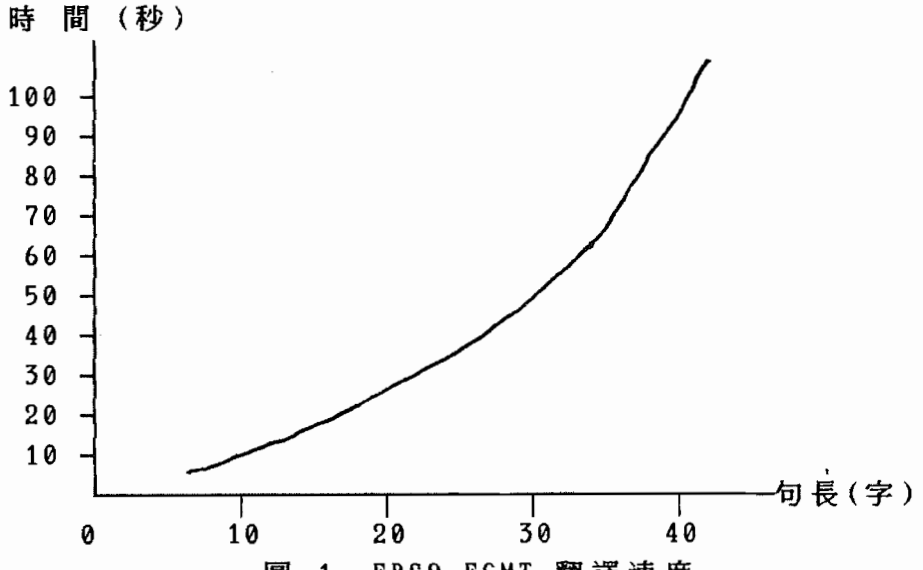

圖 1. ERSO-ECMT 翻譯速度

3. 作法

針對超長句， ERSO-ECMT 除了採取 some-path bottom-up 剖析方式，降低其分岔參數外，主要引用 Divide and Conquer〔5〕觀念，將超長句切斷成各自獨立有意義的單位: 句段，降低其搜尋深度，分別作剖析或翻譯，然後再組合剖析或是各自翻譯的結果，輸出中文句子。

ERSO-ECMT 整體系統流程如圖 2 所示，切句流程如圖 3 所示，當英文句子經由語形分析後，判斷該句是否需時 3 分以上，目前只以句長40字為判斷條件，當句長大於40字時，則進行切句處理，依據句子上的一些特定結構或功能字如：從屬介詞，對等結構之連接詞...等，來進行大結構(如：敘述句(SDEC)，名詞組(NP)，不定詞組(INF)，動詞組(VP)) 句子的切斷工作，切斷後的句段都由兩種方式處理：

① 分別剖析各句段，再重組成一完整句子結構的剖析樹，作轉換生成。
② 個別句段可直接轉換生成中文，然後再予以組合成完整中文句子輸出。

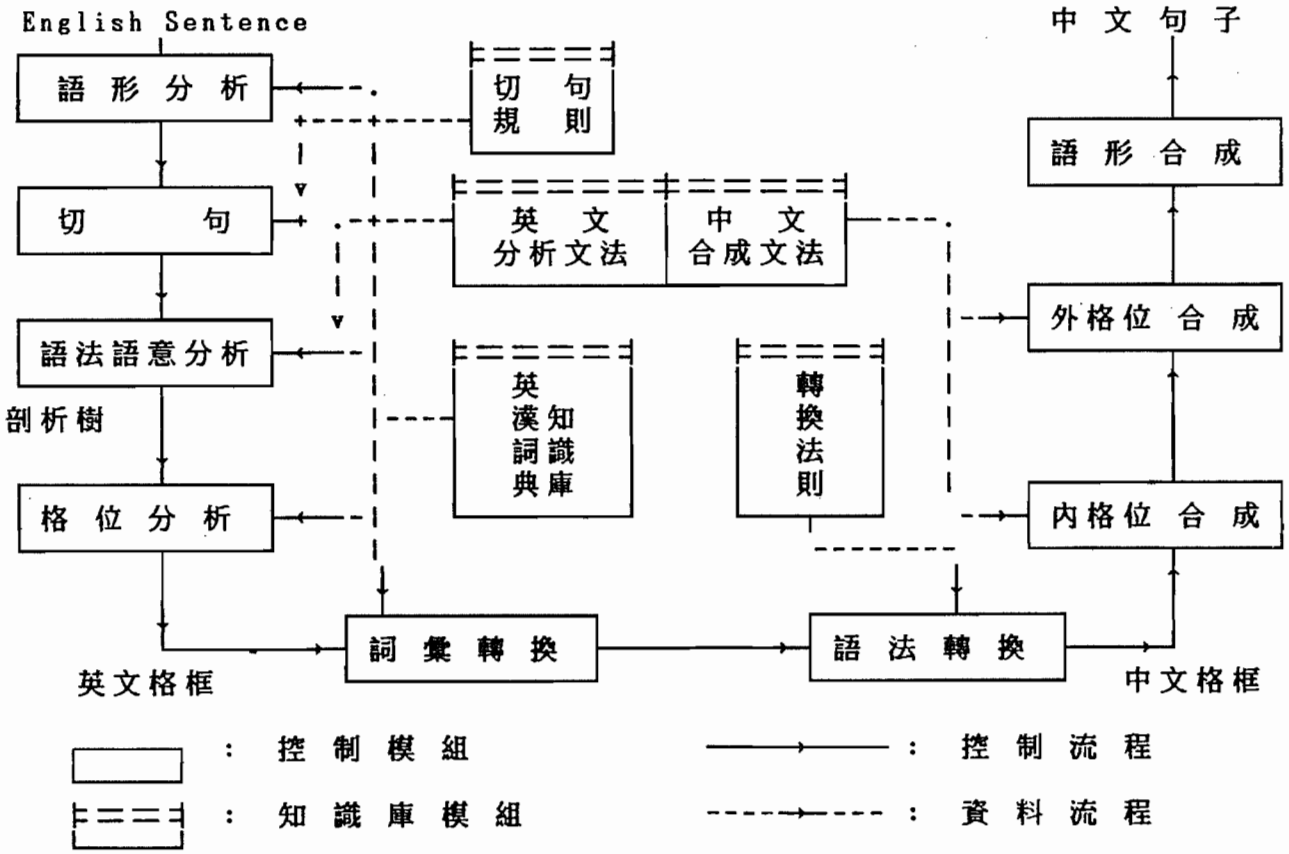

圖 2. ERSO英中機器翻譯系統處理流程圖

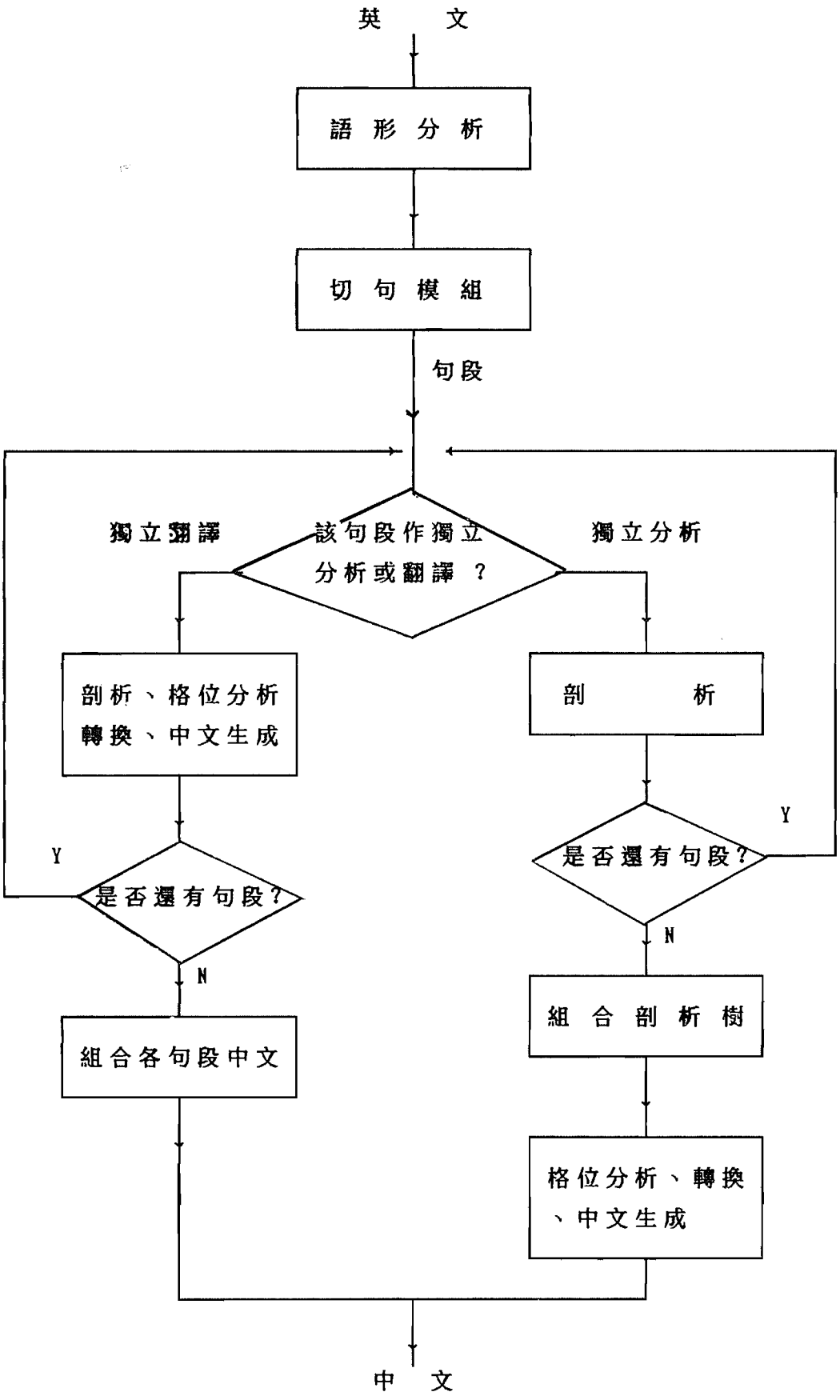

圖 3. 控制切句處理的流程圖

切句模組參考切句規則，以從頭到尾以樣板比對兼聯併(Pattern Match with Unification) 方式比對符合該句子的樣板，當找到符合樣板，則依據規則之右式做切斷及重組動作，句段本身句長若仍超過40字時，反覆進行比對樣板及切斷動作，直到各段可單獨進行剖析為止，其中切句規則包含左式及右式，左式以 Augmented Regular Expression (即 regular Expression 加入 test 功能) 方式表達之，右式則為 Action，表達法如下：

SEGRULE : = (LHS RHS)

LHS : = (E ... E)

E : = E1 | (E1 test)

E1 : = VAR | CAT | ving | ved | num | english | (closure LHS)
| (plus LHS) | (opt E1)

CAT : = a | z | x | art | b | c | h | m | n | p | r | v | u | w | wn | wu

RHS : = (A ... A)

A : = (parse VAR node) | (parse_transfer_synthesis VAR)
| chinese

VAR : = v_1 | ... | v_8

* 左括孤 "(" 及右括孤 ")" 為 Terminal。
test 為任意 LISP Expression。
english 為任意之英文單字。chinese 為任意之中文字串。
parse 為 Action，主要作剖析動作，node 為 grammar node。
parse_transfer_synthesis 亦為 Action，主要作剖析、轉換、生成。

其中 closure 及 plus 之比對方式採取最短匹配（Shortest Match）方式，如下例 %v3 所示，比對 "%v1 ving %v2 , (closure (ving %v3 ,)) c ving %v4" 之結構，以便取得最佳正確的句段：

[56] The air-use plans may be used as the basic framework for achieving the desired air quality by such means as limiting the emissions from individual sources, limiting the emissions from sources in certain areas, or disallowing new pollutant sources in overburdened areas.

[譯文] 空氣使用計畫可以藉著從個別的源限制放射物，從在某些領域的源限制放射物，或在過度負荷領域不允許新的污染源，用為達到需要的空氣品質的基本架構.

為了提供較大的中文生成彈性，可選擇：

①各句段分別剖析後，將剖析樹再組合，因此剖析器必須能提供除以單詞為輸入單位外也能接受子樹為輸入單位，如下例所示為一 "In order INF , SDEC" 結構，因此可單獨剖析 INF 及 SDEC，最後再將組合後剖析樹作中文生成。

[7] In order to evaluate or rank land use plans in terms of air quality, it is necessary for planners to be able to project emission density using only planning variables, because detailed source characteristics are not available at the time alternative plans are being developed and evaluated.

[譯文] 為了以空氣品質評估或排列土地使用計畫，對計畫者必要能只用計畫變數設計放射密度，因為詳細的原始特徵不可用，同時不同計畫被發展，並且被評估。

5. 優點

一般來說，超長句若不切斷處理，將會導致系統資源（時間、記憶體）消耗殆盡，引起不當的效果，因此與其讓系統耗去時間，不如多設定切句規則來強制性切斷句子，以便系統能順利運轉。

根據此次測試超長句的結果，如表 2 所示，得知只要能設定好的切句規則，經切斷後的句子不論在時間及正確率上都比沒做的好，也足以驗證以上的論點。此外超長句經過切斷處理後的譯文改善，也是優點之一。

＞40 字句數	42
切句正確 %	36
切對並譯正確句數	28
切對並譯正確 %	66.7
剖析長句平均時間	＜ 2 分鐘/句

表 2. ERSO-ECMT 超長句評估結果

當然 Fail-Soft 亦是切斷後，所獲致的一項好處，雖然某些句段存在一些句法錯誤，但在整個大結構上來講，亦可使其成功組合成一剖析樹，如下例為一 "%v1 (n of %v2 ,)) c n of %v3" 結構，若當其中一個 "n of %v2" 剖析不成功時，亦可強迫其為 NP，而組合成完整剖析樹：

[140] Refinement of the program should include certification of the Air Force Management Laboratory for hazardous waste analysis, documentation of waste analysis procedures, expansion of analytical capability, reduction of the numbers of used oil analysis, and operation of a solvent recovery unit at Cape Canaveral AFS.

[譯文] 程式的改良應包含 為危險的廢物分析的空軍管理實驗室的證明，廢物分析程序的文件，分析能力的擴大， 用過的油分析數的縮減，和在卡納維爾角 AFS 溶劑回收單位的操作.

6.結論

超長句在科技性文章比較常見，是不可忽視的處理對象，與其讓使用者來作前編輯，倒不如複雜的留給工程師。雖然初步獲致好的翻譯成效，但也激發對小句段的剖析研究，因為此一做法也較合乎人類在認知自然語言的過程，此方面的研發工作，電子所仍在進行中。

感謝

本文係工業技術研究院執行經濟部 78 年度計劃（計畫代號：3131500）所得成果的一部份。

參考文獻

〔1〕Bennett W.S. and Slocum J., "The LRC Machine Translation System", Computational Lingustics, Vol. 11, No. 2-3, Apr.-Sept., 1985, pp. 112-121.

〔2〕Winston P.H., "Artificial Intelligence", Addison-Wesley, Reading, MA., USA, 1984.

〔3〕Su Keh-Yih et. al., "A New Parsing Strategy in Natural Language Processing Based on The Truncation Algorithm", NCS 1987, Vol. 2, pp. 580-586.

〔4〕"Japanese Polish Support System for The Japanese-English Translation", 情報處理學會第 36 回(昭和 62 年後期)全國大會, pp. 1249-1250.

〔5〕Ellis Horowitz, "Fundamentals of Computer Algorithm", 儒林, 1978.

〔6〕"英語 200 句型", 新光書店, 1972.