申請日期：093-02-26	IPC分類 G06F 17/20
申請案號：93104530	

(以上各欄由本局填註)

發明專利說明書 200529014

一、發明名稱	中文	處理中文自然語言文句的方法
	英文	Method for Processing Chinese Natural Language Sentence
二、發明人(共3人)	姓名(中文)	1.陳貽浚 2.張鳳玲 3.鄭華森
	姓名(英文)	1.Chen, Yi-Chun 2.Chang, Feng-Lin 3.Cheng, Hua-Sen
	國籍(中英文)	1.中華民國 TW 2.中華民國 TW 3.中華民國 TW
	住居所(中文)	1.台北市信義區松隆路125號八樓之4 2.台北市慶城街26-1號四樓 3.台北市德行東路104巷9弄2號5樓
	住居所(英文)	1.8-4th Fl.,No. 125, Song-Rong Rd., Taipei, Taiwan, R.O.C. 2.4th Fl., No.26-1, Ching-Chen St., Taipei, Taiwan, R. O. C. 3.5th Fl., No. 9, Lane 104, Deshieng E. Rd., Taipei, Taiwan, R.O.C.
三、申請人(共1人)	名稱或姓名(中文)	1.金揚資訊科技股份有限公司
	名稱或姓名(英文)	1.Simpleact Incorporated
	國籍(中英文)	1.中華民國 TW
	住居所(營業所)(中文)	1.台北市民生東路四段54號八樓801室 (本地址與前向貴局申請者不同)
	住居所(營業所)(英文)	1.Room 801, 8th Floor, No.54, Sec.4, Ming-Sheng E. Rd.,Taipei, Taiwan, R. O. C.
	代表人(中文)	1.梁敬建
	代表人(英文)	1.Liang, Jin-Chiang

一、本案已向

國家(地區)申請專利	申請日期	案號	主張專利法第二十四條第一項優先權

無

二、□主張專利法第二十五條之一第一項優先權:

申請案號:

無

日期:

三、主張本案係符合專利法第二十條第一項□第一款但書或□第二款但書規定之期間

日期:

四、□有關微生物已寄存於國外:

寄存國家:

寄存機構: 無

寄存日期:

寄存號碼:

□有關微生物已寄存於國內(本局所指定之寄存機構):

寄存機構:

寄存日期: 無

寄存號碼:

□熟習該項技術者易於獲得,不須寄存。

五、發明說明 (1)

【發明所屬之技術領域】

本發明是有關於一種處理中文自然語言文句的方法，即利用淺層文法解析(Shallow Parsing)，將中文句子轉換為三元表示式的自然語言處理方法。

【先前技術】

中華民國專利公報公告第526427號係一種功能性系統，其包含需要存取匯集式資源的一組功能，此系統包含附接來實做存取架構的界面，特徵為以預定方式歷經的複數個狀態，一狀態形成特定長度存取的可能性並根據可存取此匯集式資源的功能定義優先順序；此與本案根本不同；另中華民國專利公報公告第517191號係一種用於資料處理系統中建立處理電子訊息規則的方法，係用以檢測移動電子訊息到一個文件的用戶輸入，比較該電子訊息與該文件夾中的其他電子訊息的特徵，形成一種比較，和根據這一比較產生處理電子訊息的規則，亦與本案不同。

【發明內容】

請參考第1圖，為一般傳統上處理自然語言文句之方塊圖。一文句經由構詞分析(morphological analysis)102，取得文句中每一字詞的基本型態，例如英

文的動詞 studies，經過分析轉變為動詞原形 study，再經由句法分析 (syntactic analysis)104，得到文句的句法結構 (syntactic structure)，最後由語意解釋 (semantic interpretation)106產生一語意表示式 (semantic representation)。

但在處理中文文句時，此類方法卻有以下不足之處：

1.中文文句在書寫時，不像英文字詞間有空白區隔，因此處理中文文句時，必須先作斷詞處理，將文句中的每一字詞辨識出來。

2.中文在語言學上是屬於孤立語 (isolating language)，少有文法上的各種形式變化，如，名詞單複數及動詞語尾等變化。因此如英文等印歐語系的構詞分析，便無法套用於處理中文字詞。

3.傳統上的句法分析通常作整句剖析，取得最適當的句法結構，作自動化處理時錯誤率高，一般需要人工作後處理修正。

4. 對整個句子作語意解釋時，需要足夠的知識庫，此知識庫包含每一字詞的語意分類及功能，建立此類知識庫，需要大量的時間與人力。

有鑒於此，本發明提出一種處理中文自然語言文句的方法，針對中文的特性，先作斷詞處理，並以詞組層 (phrase-level)的淺層文法解析取代複雜的整句文法剖析，最後以簡易的三元表示式，表示文句的組成要素。

五、發明說明 (3)

【實施方式】

請參照第2圖，為本發明提出的中文自然語言文句處理方法之方塊圖，首先斷詞處理程序202，係採用長詞優先法則，將文句中的中文字詞斷出並標上詞性，產生標記詞性的字詞序列；字詞過濾程序204，係濾除不需要的字詞，以簡化詞組的複雜度；詞組剖析程序206，係以詞組層的淺層文法解析出文句中的詞組；三元表示式轉換程序208，係為產生對應於文句的三元表示式。

本發明提出另一種用於表現文句中子句(clause)的三元表示式，係將文句中由詞組組成的子句，以三元表示式表現之，其中，一個三元表示式包含三個組成成分：主詞(S)、述詞(P)及受詞(O)，每一成分為一詞組，意即，一個三元表示式由三個詞組所組成，表現出一子句中主詞、受詞及之間的關聯性。請參照範例一，"張三喜歡李四"以三元表示式表示為"[[張三], [喜歡], [李四]]"。三元表示式的定義請參照定義一。

範例一：

(1)張三 喜歡 李四。

(2)[[張三], [喜歡], [李四]]

定義一：

一個三元表示式T包括三個組成成分，T表示為[S, R, O]，其中：

1S係由一至多個名詞所組成的序列，其文法角色為子句中

的主詞。

1R係由一至多個動詞或介詞所組成的序列，其文法角色為子句中的述詞。

10係由一至多個名詞所組成的序列，其文法角色為子句中的受詞。

在定義一中，一個三元表示式包括S、R與O三個成分，分別代表一子句中的主詞、述詞與受詞。而一中文文句可能含有一至多個子句，因此一文句亦由一至多個三元表示式表現之。請參照範例二，為一個具有二個子句的中文句子，"張三參加比賽得到冠軍"，以三元表示式表示為"[[[張三], [參加], [比賽]], [[zero], [得到], [冠軍]]]"，其中，範例二(2)有二個三元表示式，而在第二個三元表示式中，"zero"表示一零代詞，為主詞省略，係中文經常發生的文法角色省略現象。

範例二：

(1)張三 參加 比賽 得到 冠軍。

(2)[[[張三], [參加], [比賽]], [[zero], [得到], [冠軍]]]

本發明又再提出另一種用於解析文句中子句結構的三元規則，三元規則包括三元生成規則與三元例外規則，其中，三元生成規則係為子句構成之基本句型(主詞－述詞－受詞)，三元例外規則係用以處理中文之零代詞現象。

三元生成規則，係包括四條規則，分別處理四種基本子句句型，請參照表格一，其中，規則1所處理的句型為主詞+

及物動詞詞組+受詞，規則2所處理的句型為主詞+不及物動詞詞組，規則3所處理的句型為主詞+介詞+受詞，規則4所處理的句型為一文句僅以一名詞詞組所構成。

表格一：三元生成規則

編號 規則

1 Trilpe1(S,P,O) a np(S),vtp(P),np(O).

2 Trilpe2(S,P,none) a np(S),vip(P).

3 Trilpe3(S,P,O) a np(S),prep(P),np(O).

4 Trilpe4(S,none,none) a np(S).

請參照表格一，"vtp(P)"表示述詞為及物動詞詞組，"vip(P)"表示述詞為不及物動詞詞組，"prep(P)"表示述詞為介詞。編號4規則為處理一文句僅以一名詞詞組所構成。

三元例外規則，係包括五條規則，用以處理中文經常發生的零代詞現象，請參照表格二，其中，規則1z1、2z1與3z1處理的是當零代詞出現在子句的主詞位置，規則1z2處理的是當零代詞出現在子句的受詞位置，規則1z3處理的是當零代詞同時出現在子句的主詞與受詞位置，如範例二(2)，在套用規則1z1後，"zero"表示一零代詞，為主詞省略。

表格二：三元例外規則

編號 規則

1z1 Trilpe1z1(zero,P,O)a vtp(P),np(O).

2z2 Trilpe1z2(S,P,zero)a np(S),vtp(P).

五、發明說明 (6)

3z3Trilpe1z3(zero,P,zero)a vtp(P).

2z1Trilpe2z1(zero,P,none)a vip(P).

3z1Trilpe3z1(zero,P,0) a prep(P),np(0).

為讓本發明之上述和其他目的、特徵、和優點能更明顯易懂，下文特舉較佳實施例，並配合所附圖式，作詳細說明如下：

圖式之簡單說明：

第 1圖繪示的是一般傳統自然語言文句處理之一方塊圖；

第 2圖繪示的是本發明提出的中文自然語言文句處理方法之方塊圖；

第 3圖繪示的是本發明之一詞組剖析流程圖；以及

第 4圖繪示的是本發明之一三元表示式轉換流程圖。

第 5圖繪示的是本發明之一流程圖。

第 6圖繪示的是本發明之一流程圖。

重要元件標號

102: 構詞分析程序

104: 句法分析程序

106: 語意解釋程序

202: 斷詞處理程序

204: 字詞過濾程序

206: 詞組剖析程序

208: 三元表示式轉換程序

步驟 s300至步驟 s306為本發明之一斷詞處理實施步驟

步驟 s400至步驟 s410為本發明之一字詞過濾實施步驟

步驟 s502至步驟 s518為本發明之一詞組剖析實施步驟

步驟 s602至步驟 s616為本發明之一三元表示式轉換實施步驟

較佳實施例

本發明所提出的中文自然語言文句處理方法，請參照第 2 圖，包括斷詞處理程序 202、字詞過濾程序 204、詞組剖析程序 206及三元表示式轉換程序 208等四個程序，本文就這些程序提出一較佳實施例。

請參照第 3圖，其繪示的是斷詞處理程序之步驟，包括：輸入一中文文句 s300，經由步驟 s302，自文句第一個字開始，與辭典中的詞相比對，找出所有符合的中文詞，接著在步驟 s304中，依長詞優先規則，從所有符合的中文詞挑出最長之字詞，並標上該詞的詞性，在步驟 s306中，若仍有剩餘之部分中文文句未比對，則繼續比對及挑出最長之字詞，直至中文查詢句子做完為止。

請參照第 4圖，其繪示的是字詞過濾程序之步驟，包括：輸入一已標定每個字詞詞性的字詞序列 s400，經由步驟 s402，檢查序列中第一個字詞字詞是否為名詞、動詞或介詞，若是，在步驟 s404中，將該字詞挑出併入新序列，若否，則在步驟 s406中，將該字詞移除，接著在步驟 s408中，檢查原序列中是否仍有剩餘字詞，若仍有剩餘字詞為處理，則回到步驟 s402繼續做處理，直至原字詞序列處理

完為止。

請參照第5圖，其繪示的是詞組剖析程序之步驟，包括：輸入一已標記詞性之字詞序列s502，由最左方的字詞開始處理s504，經由步驟s506檢查目前處理字詞是否與右方字詞具有相同的詞性，若是，則步驟s508結合右方字詞與目前處理字詞或序列為一新序列，接著在步驟s510，檢查目前處理序列是否與右方字詞具有相同的詞性，若是，回到步驟s508，處理下一個字詞，若否，在步驟s512，取出尚未處理之剩餘字詞回到步驟s504，在步驟s514中，檢查輸入的序列中是否還有未處理的字詞，若有，在步驟s516中，產生一僅包括一字詞之序列並將未處理剩餘字詞作為新的輸入，若無，則輸出剖析結果並結束詞組剖析程序。

請參照第6圖，其繪示的是三元表示式轉換程序之步驟，包括：輸入一詞組序列s602，在步驟s604中，由最左方的詞組開始處理，在步驟s606中，依據三元生成規則產生一三元表示式，接著在步驟s608中，檢查是否有產生三元表示式，若有，在步驟s612中檢查是否還有未處理之詞組，若無，則依據三元例外規則產生一三元表示式，接著在步驟s612中，若發現還有未處理之詞組，在步驟s610中，將未處理之剩餘詞組作為新的輸入，回到步驟s604，若所有的詞組都已經處理完成，則輸出結果並結束程序。

圖式簡單說明

四、中文發明摘要　(發明名稱：處理中文自然語言文句的方法)

五、(一)、本案代表圖為：第 __2__圖

(二)、本案代表圖之元件代表符號簡單說明：

202：斷詞處理程序

204：字詞過濾程序

206：詞組剖析程序

208：三元表示式轉換程序

一種處理中文自然語言文句的方法，利用淺層文法解析(Shallow Parsing)，將中文句子轉換為三元表示式。此自然語言處理方法採用詞性及中文句法等資訊，抽取出句子中的重要元素，並轉換為三元表示式，一三元表示式代表句子中的一個子句(clause)。由中文句子轉換至三元表示式是依據本文中所提之三元規則，三元規則包括三元生成規則與三元例外規則，為參考中文句法之基本原則(主詞－述詞－受詞)來建立，三元例外規則則是用以處

五、英文發明摘要　(發明名稱：Method for Processing Chinese Natural Language Sentence)

A method for processing Natural Language Chinese Sentence can transform a Chinese sentence into a Triple representation using shallow parsing techniques. The method is concerned with parsing Chinese sentence by employing lexical and syntactical information to extract more prominent entities in a Chinese sentence, and that is then transformed into a Triple representation by

四、中文發明摘要　(發明名稱：處理中文自然語言文句的方法)

理中文之零代詞(zero anaphor)現象。本文中處理中文文句的步驟包括斷詞處理、字詞過濾、詞組剖析及三元表示式轉換等四個程序，其中，斷詞處理程序採用長詞優先法則，將文句中的中文字詞斷出並標上詞性；字詞過濾程序為濾除不需要的字詞；詞組剖析程序為解析出文句中的詞組；三元表示式轉換程序為產生文句的三元表示式。

五、英文發明摘要　(發明名稱：Method for Processing Chinese Natural Language Sentence)

employing the Triple rules referring to elemental Chinese syntax - SVO (subject, verb, and object in order). The lexical and syntactical information in our method is referring a lexicon poessessed of part-of-speech (POS) information and pharse -level syntax in Chinese respectively. The Triple representation consists of three elements which are agent, predicate, and patient in a sentence.

六、申請專利範圍

申請專利範圍

1.一種處理中文自然語言文句的方法，係有關於將一中文自然語言文句轉換為一結構化表示式，其中包括：

一斷詞處理程序，係將一中文文句作斷詞處理，將文句中的字詞一一斷開，並標示每個字詞的詞性；

一字詞過濾程序，係將一中文文句經斷詞處理後，將不處理或不必要的字詞濾除；

一詞組剖析程序，係剖析中文文句，將文句中的詞組抽取出來形成一詞組序列；以及

一三元表示式轉換程序，係將一文句經由詞組剖析程序，產生詞組序列後，將文句中的子句，一一轉換為三元表示式。

2.如申請專利範圍第1項所述之字詞過濾程序，其中，將不屬於以下詞性的字詞濾除：名詞、動詞及介詞。

3.如申請專利範圍第1項所述之詞組剖析程序，其中包括抽取出以下詞組：名詞詞組、動詞詞組及單一介詞。

4.如申請專利範圍第1項所述之三元表示式轉換程序，其中，三元表示式的定義如下：

一個三元表示式T包括三個組成成分，T表示為[S, R, O]，其中：

S係由一至多個名詞所組成的序列，其文法角色為子句中的主詞；

R係由一至多個動詞或介詞所組成的序列，其文法角色為子句中的述詞；

六、申請專利範圍

0係由一至多個名詞所組成的序列，其文法角色為子句中的受詞。

5.如申請專利範圍第1項所述之三元表示式轉換程序，其中，三元表示式可表現之子句結構有以下四種：

主詞+及物動詞詞組+受詞；

主詞+不及物動詞詞組；

主詞+介詞+受詞；以及

一文句僅以一名詞詞組所構成。

6.如申請專利範圍第1項所述之三元表示式轉換程序，其中，三元表示式可表現之具有零代詞現象之子句結構有以下五種：

零代詞+及物動詞詞組+受詞；

主詞+及物動詞詞組+零代詞；

零代詞+及物動詞詞組+零代詞；

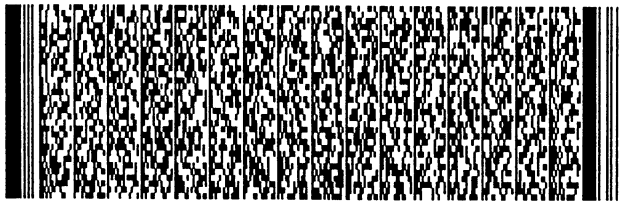
零代詞+不及物動詞詞組；以及

零代詞+介詞+受詞。

7.一種處理中文自然語言文句的方法，係有關於將中文文句中的子句，轉換為一三元表示式，此三元表示式包含三個組成成分，分別依序對應於子句中的主詞、述詞與受詞。

8.如申請專利範圍第7項所述之一三元表示式，其中第二個組成成分為一子句中主詞與受詞的關聯性。

9.如申請專利範圍第8項所述之一子句中主詞與受詞的關聯性，包括動詞詞組與介詞。

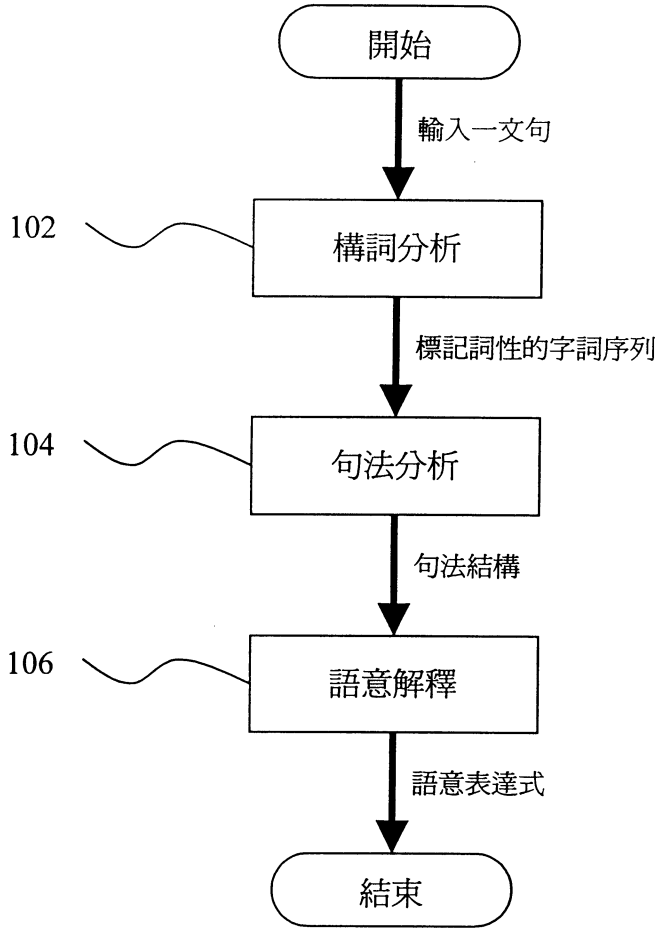


六、申請專利範圍

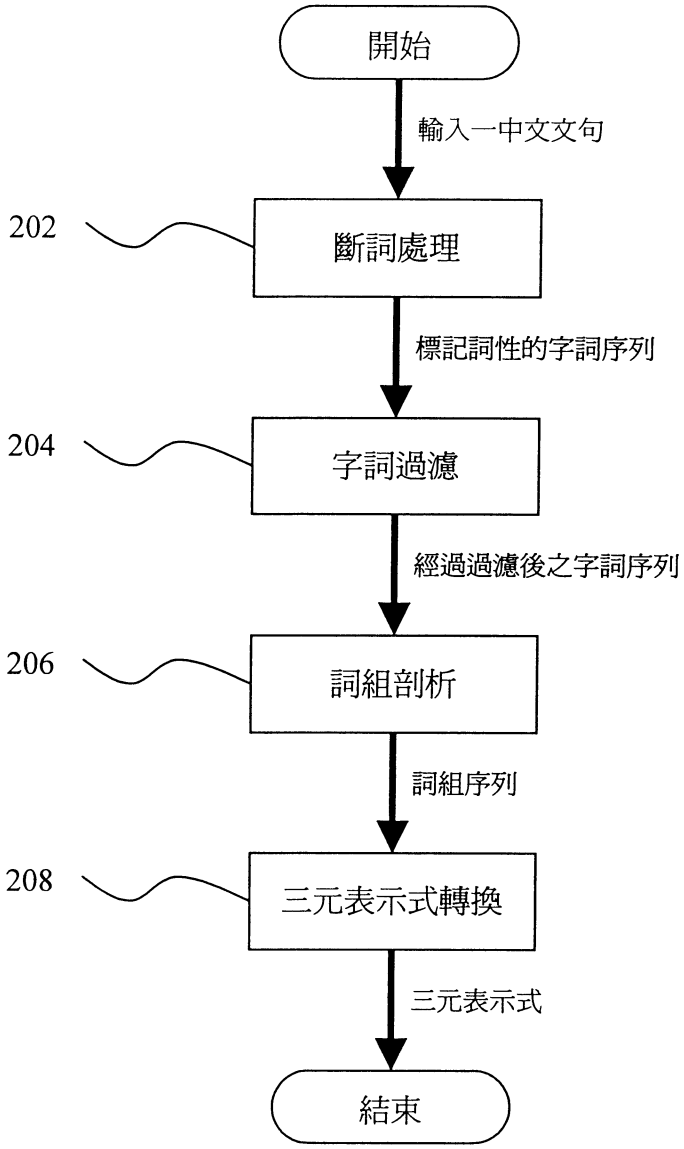
10.如申請專利範圍第7項所述之一三元表示式，其中，三元表示式對應的子句，包括零代詞出現於主詞與受詞位置的子句。

11.一種處理中文自然語言文句的方法，係有關於分析文句中所有的子句結構，以一三元表示式表現一子句的方式，進而表現整個文句。

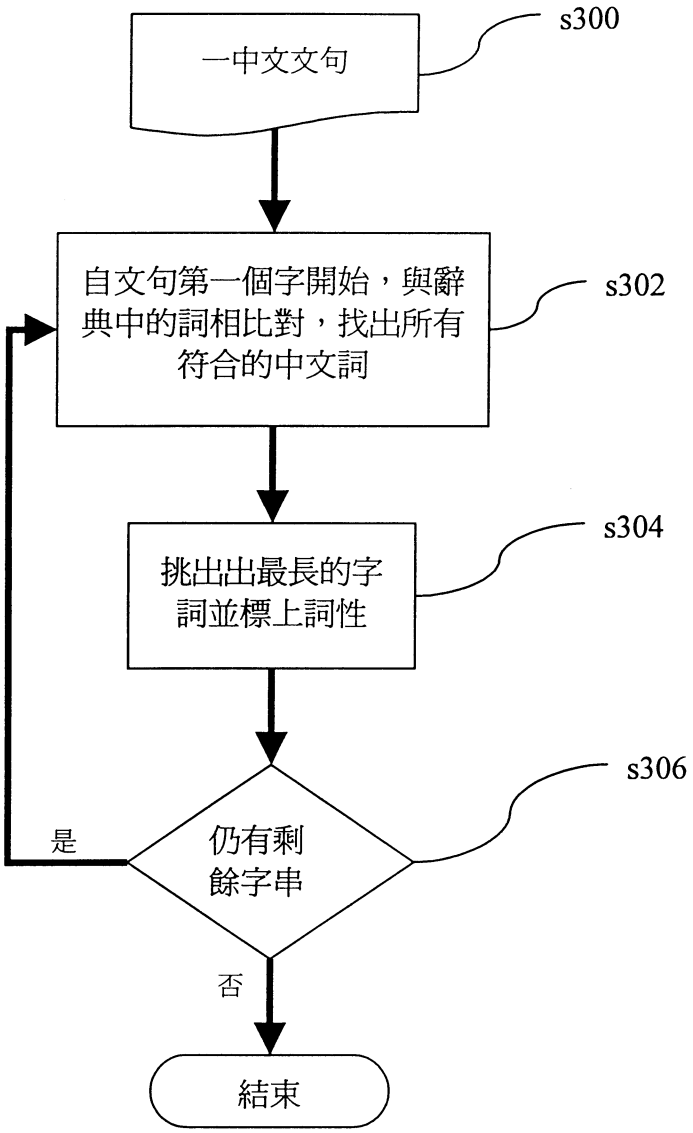
12.如申請專利範圍第11項所述之表現中文問句的方式，是以一至多個三元表示式表現一中文文句，其中一至多個三元表示式依照先後出現順序，對應至文句中的所有子句。



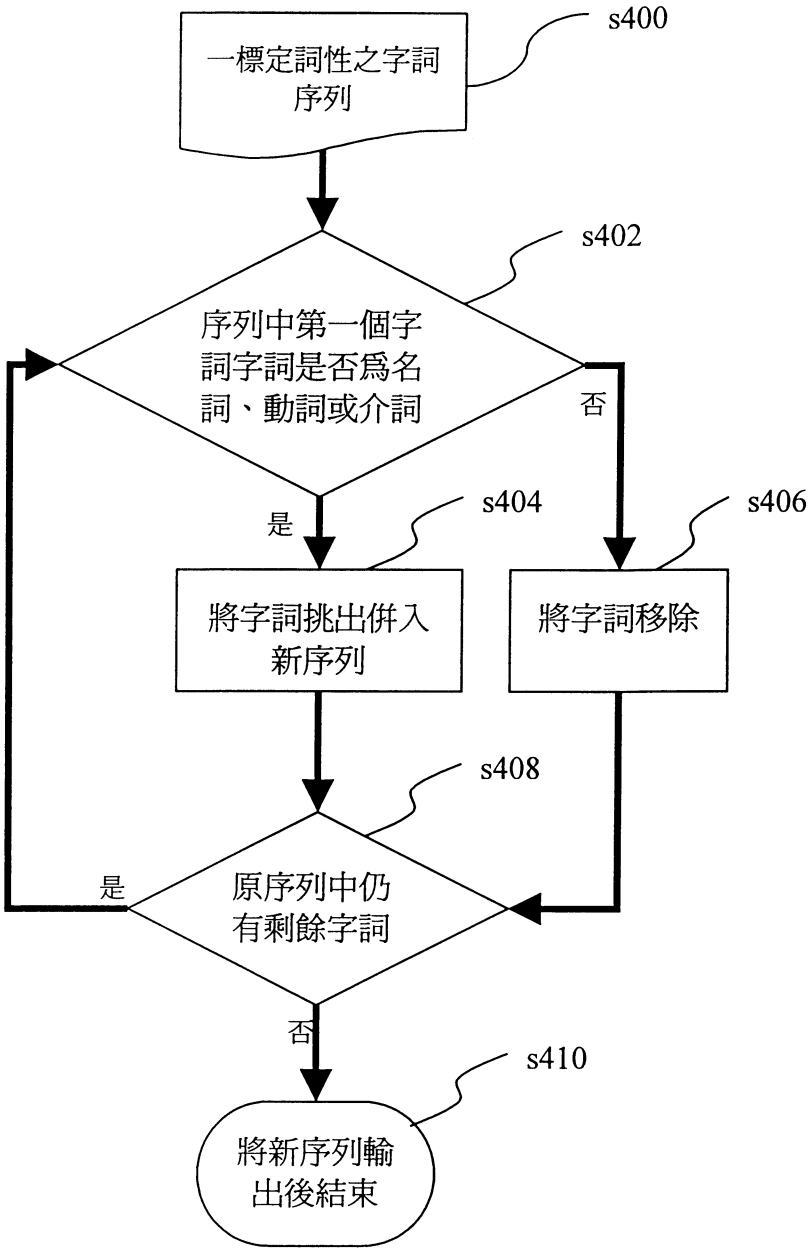
第 1 圖


開始
輸入一中文文句
202
斷詞處理
標記詞性的字詞序列
204
字詞過濾
經過過濾後之字詞序列
206
詞組剖析
詞組序列
208
三元表示式轉換
三元表示式
結束

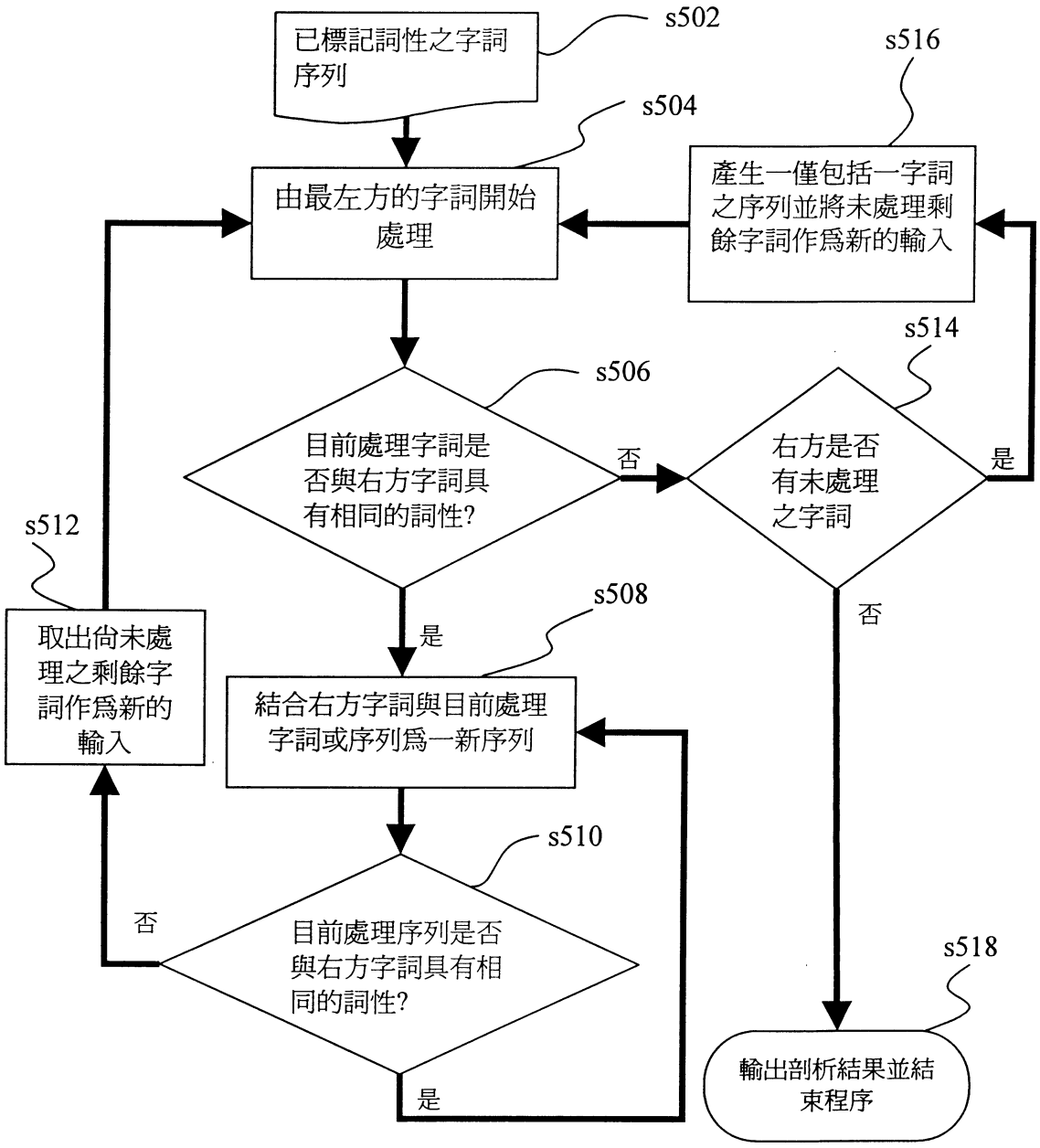
第 2 圖


一中文文句
s300
自文句第一個字開始，與辭典中的詞相比對，找出所有符合的中文詞
s302
挑出出最長的字詞並標上詞性
s304
仍有剩餘字串
s306
是
否
結束

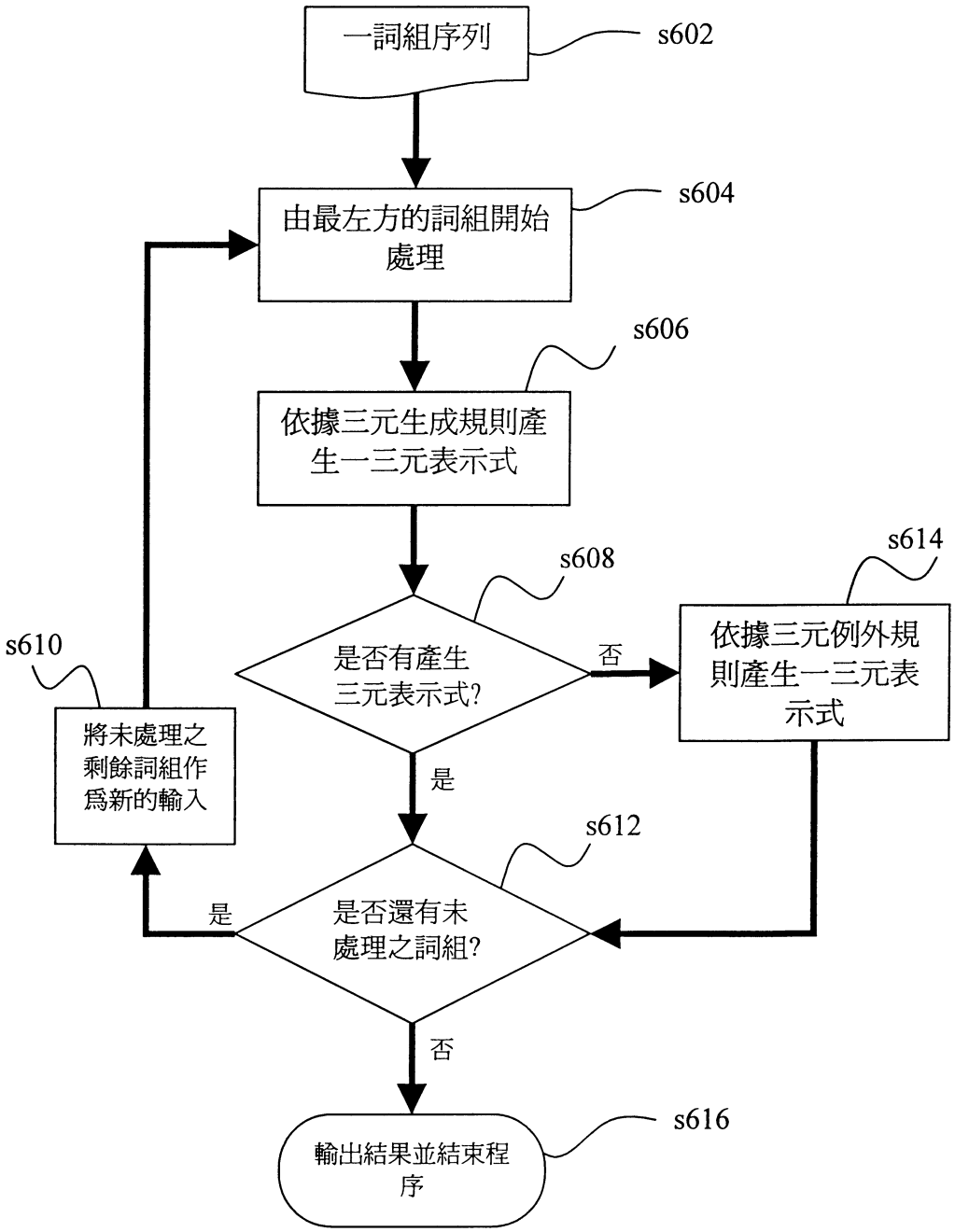
第 3 圖


s400
一標定詞性之字詞序列
s402
序列中第一個字詞字詞是否爲名詞、動詞或介詞
否
是
s404
s406
將字詞挑出併入新序列
將字詞移除
s408
是
原序列中仍有剩餘字詞
否
s410
將新序列輸出後結束

第 4 圖


已標記詞性之字詞序列
s502
s504
由最左方的字詞開始處理
s506
目前處理字詞是否與右方字詞具有相同的詞性?
否
是
s508
結合右方字詞與目前處理字詞或序列爲一新序列
s510
目前處理序列是否與右方字詞具有相同的詞性?
否
是
s512
取出尚未處理之剩餘字詞作爲新的輸入
s514
右方是否有未處理之字詞
是
否
s516
產生一僅包括一字詞之序列並將未處理剩餘字詞作爲新的輸入
s518
輸出剖析結果並結束程序

第 5 圖


一詞組序列
s602
由最左方的詞組開始處理
s604
依據三元生成規則產生一三元表示式
s606
是否有產生三元表示式?
s608
否
依據三元例外規則產生一三元表示式
s614
是
將未處理之剩餘詞組作爲新的輸入
s610
是否還有未處理之詞組?
s612
是
否
輸出結果並結束程序
s616

第 6 圖

六、指定代表圖