中文摘要

一. 綜述：漢語自然語言處理的重要論題

黃居仁 陳克健
中央研究所

本文由計算語言學理論及漢語語法分析兩個觀點出發；討論漢語自然語言處理最重要的題目及其理論背景，並藉由這些討論來介紹本論文集中所收的九篇論文的貢獻及相關學術地位。本文中討論的幾個論題爲：一、「詞」在自然語言處理中基本地位及在中文分析中的特殊問題，二、中文剖析的大要素，包括 1) 詞彙與詞類分析，2) 語法功能之判定， 3) 多重詞義之解析及 4) 多重結構之解析，三、如何由結構導出意義，四、如何構建應用系統。本文以討論漢語自然語言之未來發展方向作爲總結。

二. 訊息爲本的格位語法 — 一個適用於表達中文的語法模式

陳克健 黃居仁
中央研究院

本論文提出一個以訊息爲本的語法模式，這個語法模式和其配合的剖析方法，可以很精確的表達及很有效率的分析中文句子。本語法模式採用詞彙爲中心的表達方式，將每一個詞的語法及語意訊息以特徵結構表示。詞彙結合爲片語，片詞建構成句子，以中心語驅動，結合時必須符合句子中詞彙所規定的語法限制。語意的合成假設具有結合性可以從詞彙的語意訊息聯併獲得。以一語法模式保存了聯併語法的所有優點而且兼顧了剖析的效率與語意的分析。

三. 以邏輯爲本的中文剖析

陳信希
國立臺灣大學

中文是一種使用非常彈性且前後文相關的語言，因此電腦很難處理中文語句。除此之外，由於中文句子語彙之間並沒有明顯的分隔符號，斷詞爲另一個困難的問題。這篇論文採用邏輯程式的技術，將斷詞視爲剖析的一部分。爲了處理中文空詞高自由度的使用，論文將c-command和subjacency兩項限制條件，放在整合的剖析-斷詞模型中，以決定那些成分被移走且/或刪除。論文也提出一種語法型式化語言，其具有均一處理移位現象及任意個數的移位、預先自動偵測語法錯誤、和清楚的敘述語等特點。剖析器產生裝置將語法規則轉換成程式碼，並作最佳化。圖形聯併支持多值、反面、離接等結構，在這個模型中，被採用來表示成分間的共存限制和資訊傳遞。許多常見的語言現象如主題-評論結構、把字句、被字句、關係子句、同位句、遞續結構等，都在這個環境中表示出來。最後，本文也討論中文長句的處理。

四. 日中機器翻譯系統之時貌認知處理

郭俊桔
松下研究中心

句時貌是一種綜合解析句中主動詞，時貌記號，副詞，主詞，受詞和其他構句要素的時貌意義函數而非只考慮主動詞之時貌意義的動詞時貌。本研究以情況形態(situation types)和時貌解釋細分類(further distinction of aspect meaning)來表示句中之時貌，情況形態是人類對於發生事件的時貌性質基於其本身之認知和理解力所做的情況分類。本論文中使用事件，狀態，習慣和一般來表示情況形態，事件又可以依其有無動作主再細分爲達到，過程，達成和動作。其中，達到和達成

是表示完成的時貌解釋，然而其他的情況則表示非完成的時貌解釋。進一步，完成的時貌解釋又可被細分爲結束，經驗和完成；非完成的時貌解釋則被細分爲習慣，進行，繼續，開始和反覆等。爲了驗證上述提案方法的有效性和進步性，本文中以日中機器翻譯之時貌處理爲例，並討論解析句中同時出現多數個時貌記號之問題點和相關解決方法。

五. 中文句中名詞串的歧義處理

李錫堅[1] 葉慶隆[2]
[1]愛丁堡大學 [2]國立交通大學

本論文提出一個法則導向的方式來解決中文句子中連串名詞結構(serial noun constructions)的岐異問題，中文句子中相連兩個名詞不一定具有修飾詞—首語(modifier-head)的關係或是位移的相鄰名詞，它們還可能是擁有名詞組(possessive noun phrase)、同位名詞組(appositive noun phrase)、連接名詞組(conjective noun phrase)。此外，超過兩個名詞組的階層結構，由於名詞間的不同組合方式，不一定是由左到右相接（left-to-right association)。由測試文章我們統計出串列名詞組發生率有20%以上，本論文將使用語法種類特徵和語意階層(semantic hierarchy)設計岐異解決法則。本論文亦將提出四種名詞—名詞組合的先後關係(precedence relation)，以解決串列名詞的階層關係。本論文提出的方法已結合聯幷基底(unification-based)的圖性剖析器(chart-parser)，我們將以一些例子作說明。

六. 中文裏的定量複合詞：構成律以及剖析程式

莫若萍 楊曜榮 陳克健 黃居仁
[1] 俄亥俄大學 [2] 中央研究院

本論文將提出剖析中文時如何處理定量式複合詞。像衍生性的複合詞一般，定量式複合詞也可不斷地衍生新詞，數量龐雜無法在辭典中一一列出。因此造成斷詞或者是剖析時歧異產生。但比起其他複合詞，定量式複合詞卻較容易歸納其衍生的規則，進而使其在剖析前即已辨認出來。

我們發現定量式的詞不但具有組合性同時也有階層關係，因此根據這種關係我們列出組合規則並將之應用於我們所設計的剖析系統中。結果發現，大部份的定量式複合詞皆可辨識出來，同時斷詞時產生的歧異性也大爲減低。

七. 統計式分詞法

江東輝 張景新 林銘裕 蘇克毅
臺灣新竹國立清華大學電機工程研究所

中文詞與詞之間並無類似空白符號之類的分隔符號，故在進行中文訊息處理之前，需先界定詞的界線。傳統的分詞方法主要是利用詞典訊息，輔以一些經驗法則，如長詞優先法，來找出中文的分詞點。由於中文構詞及句法相當複雜，這樣的作法，對於大型系統而言，未必能適用。

本文重點主要在於利用中文句中所有可資運用的特徵，發展一套一般化的中文分詞公式，從而推導出各種的統計分詞模式。在估計統計參數的估計值時，一般是以最大似然度作爲估計標準。但這種估計標準並未能反應出各種可能的分詞樣型間相對的排名順序。因此，我們採

用具有強健性的調適性學習法，來調整參數的估計值，以提昇系統的效能。

實驗結果顯示，我們所提議的分詞模式在各種情況下均能經濟而有效地達到分詞的效果。在使用詞長度訊息及應用強健性的調適性學習法於一簡單的統計模式之下，對測試語料而言，以詞爲單位的分詞辨認率達９９．３９％，以句爲單位的辨認率則達９７．６５％。

此外，在一般情況下，並非所有詞彙都可以在系統的詞典內找到。這類的「新詞」或「未知詞」往往嚴重影響分詞的辨認率。因此，我們也針對此一「未知詞問題」提出一些可行的解決方法。

八. 具備高效率語言模型技術的國語聽寫機

簡立峰　陳克健　李琳山
中央研究院

金聲一號（Golden Mandarin I）是國際上第一套可以即時辨認大字彙、無限文句的國語語音聽寫系統。這套系統的語言模型較爲簡單，因此語言處理能力較爲有限。爲了改進這項缺失，本文提出一項新的語言模型方法。這個方法利用一最佳優先的格狀詞組剖析演算法，成功地結合統計式馬可夫語言模型與聯併文法理論。實驗結果證實這項新方法所得的正確率優於原有語言模型，且如果剖析策略適當，辨認速度甚至可以更快。根據分析，這是因爲利用文法分析一些不合文法的詞彙組合可以先事先去除，而成功的剖析策略與語言模型機率可以導引正確搜尋方向。本文除提出這項新的語言模型方法外，對金聲一號國語語音聽寫系統的設計以及統計式語言模型與文法理論的特性差異也都會加以介紹討論，相信藉由這個新的語言模型方法，可以進一步提昇國語聽寫機的成效。

九. 從議論文體篇章到推理樹 — 基於語法標記的中文篇章摘要系統

鄒嘉彥 連興隆 何慶昌 黎邦洋
香港城市大學

議論文體的特色，在於通過邏輯推理，將作者想要傳遞的信息，用多層次互相結合的各個命題表達出來。在語言的表層結構上，這個推理過程，會使用包含各種事實或意見的修辭關係表達出來，例如前題、條件、推導及結論等等。基於語言的本質，這個多層次的推理結構，在篇章中只能表示爲直線串聯起來的各個命題，再用標點符號和具有特定功能的語法標記，標誌出各命題相互間的層次關係。

本研究提出一個篇章處理的方法，可以用來分析及獲知一篇議論文的篇章結構，及其論証的過程。這個處理方法經由對篇章進行語法及修辭結構的分析，最終推導出作者推論時所依據的推理法則〔即推理樹〕。其中修辭結構的分析，主要是基於對語法標記功能的辨別。

這個篇章處理方法其中一個重要的應用，就是自動化中文篇章摘要。通過一連串的實例，本文論証了如何利用篇章處理後得到的推理樹來生成涵蓋原文不同細節、或詳或簡的多個摘要。用戶可根據實際需要，指定所想看的摘要的長度，從而達到篇章摘要系統的主要目標。

十. BehaviorTran 英中機譯系統之計算模式

蘇克毅[1] 張景新[1] 王重乃[2] 張玉玫[2] 吳銘文[2]
[1] 臺灣新竹一國立清華大學電機工程研究所 [2] 致遠科技

本文詳述 BehaviorTran 機器翻譯系統所採行的「語料爲本．統計導向」(CBSO, Corpus-Based Statistics-Oriented)的設計理念。我們將簡略

地介紹 BehaviorTran 的一些特色。並說明在研發過程中，所發現的一些規則式系統及純統計式系統的問題。由於這些問題使得大型機器翻譯系統不易發展及維護，也不易延伸至不同的語言，及適應不同的使用者。因此我們發展出「語料爲本．統計導向」的設計理念。本文將闡述此一理念在發展大型實用化系統的必要性及可行性。同時介紹基於此一理念所獲致的一些研究成果。包括統計式的機器翻譯模式，分析模組的評分函數，統計式轉換暨生成模式，參數控制式的回饋控制模式，及雙向式翻譯知識抽取模式等技術。