中文資訊的處理（二）中文資訊處理的科際關係

科學月刊
第十八卷第三期
謝清俊

一、與語文學的關係

做中文資訊處理的研究固然必須具備計算機科學的知，然而，在中國語文方面的知識也是不能缺少的。數千年來，語文是世界上任何一個民族文化延綿的主要工具，也是人與人溝通的主要橋梁。自從計算機發明了以，語文開始兼作人與計算機之間溝通的媒體。這現象明顯地指出語文和計算機有密不可分的關係。然而，由於時下計算機的能力有限，無法像人一樣處理和應用我們的語言，所以只得設計一套更簡明的語言，讓計算機使用。這種語言就稱為人工語言或計算機語言，以便和我們日常使用的「自然語言」區別。

目前的計算機語言使用了自然語言所用的符號：譬如，英文字母（中文字）、標點符號、阿拉伯數字等等，和自然語言沒有什麼兩樣，然而在語法結構上卻比自然語言簡明許多，這是為了要配合時下計算機的能力來處理計算機語言的語法的緣故。當我們學習一種計算機語言時，常常覺得它刻板之至，所有的規定不容絲毫偏差，這不僅不通情理，而且顯得固執和愚蠢，這是計算機語言的語法太欠缺彈性所造成的。另一個原因是：在語意的處理上，計算機語文能做的更少；除了必要設定的語意在事先就硬性規定以某種形式表達以外，其他都忽略掉了，或是留給應用程式自己去看著辦！

雖然目前的計算機語言和自然語言之間有這麼大的差距，然而科學家們正努力在減少它們的距離，希望有一天，計算機能具有相當的語文程度，用自然語文和我們溝通。賦與計算機處理自然語文能力的努力，統稱為自然語文處理的研究，這是近年來十分熱門的研究方向之一。對我國而言，由於研究的對象

是我國的語文，自然有其獨特的問題。這方面的研究可名為「中國語文處理」，應該是中文資訊處理研究中極重要的一環。

語文與計算機的結合是件大事，在國外已行之有年。像計算機語言學（computationallinguistics）、計算機詞彙學（computer morphology）、文獻處理（text processing）以及自然語文處理等等新學科，皆在此結合下蓬勃發展。可是，國內的情形就顯得完全不一樣。在大學中，即使偶而開了上列的課程，也用英語作為對象，這和我們的社會、文化以及現實生活脫了節。為了要使計算機有處理中文語文的能力，或是要利用計算機來推動語文知識的應用，沿著上述的發展方向，針對中文語文的特色來發展中文的計算機詞彙學、中文的全文處理技術以及中文的自然語文處理等學科是必要的。

二、與人工智慧的關係

計算機使用得越普遍，對語文處理技術的要求就殷切，而且對處理能力和品質的需求就越高。要言之，語文處理的需求來自二方面：其一是藉以改善人與機器溝通的方式，其二是藉以提升對文獻和事務處理的能力。在日本第五代電腦計畫想發展的智慧型計算機中，自然語文處理和智慧型人機介面都是必要的功能。而智慧型人機介面所做的事，除了處理語音和字形丟表達的媒體以外，還是要依賴語文的處理來做好認別及產生的工作。

早期計算機的使用大多是處理一些表格化的語文資料。例如，處理各種單據、查詢一些欄位化的資訊等等。目前，這樣的使用方式已經無法滿足接踵而來的要求；人們要求處理完整的原始文獻，如信件、公文、法律條文與判例、新聞、會議紀錄、專利文獻、史籍檔案與文獻.....等等。處理的媒體亦多元化，包括聲音、影像及圖表等等。對處理的基本功能也起了變化，不再囿計計算、邏輯判斷等，而是擴大到一些較智慧型的能力，譬如：推理、計劃、識別、學習、理解等等。這也就是在上期中，本文所說的由資訊處理發展至知識處理的

過程。其實，語言、知識、智慧三者是相互交織密不可分的，當計算機有些處理自然語文的能力時，它自然將擁有一些知識以及具備某種程度的智慧能力，而這正是人們心目中追求的理想。再說，人類累積的知識和資料，絕大部分是用語文表達的，計算機和語文的結合，自自然然地增加了處理知識的能力，包括知識的表達、、取得、組織以及應用。對計算機來說，用機器來做這些事，就是人工智慧研究的主題。

再者，在研究中文的語意時，已經涉及意念和知識的表達。為了要認定什麼是知識？它的範圍和界定是什麼？知識如何產生？如何取得？如何分類？認知的過程如何？如何運用知識？.....等等問題時，將涉及知識論（哲學）、認佑科學（心理學）、分類學（圖書館學），乃至於腦神經醫學等相關的知識。而目前在國外發展的趨勢，也是朝這個科際大給合的方向邁進。

三、與其他科學的關係

前文已經提到，改進人機溝通的方式是做中文資訊處理研究的主要原動力之一。在考慮人機溝通因素時，不可避免地要用到一些專門學科的知識。例如：心理學、人機工學（erg-onomics）、傳播學等。為了使機器表達資訊的方式能為工作人員接受和喜愛，並且要維持一個良好的工作環境，不要讓工作人員受到傷害，這些專門學科的介入是必須的。有許多關於輸入輸出的工業標準，是經由人機工學的設計以及心理學對接受程度和使用效率的測試而訂定出來的。譬如：在不同媒體中字體點陣的大小標準；字的間距、行距和畫面的標準；鍵盤上符號的安排標準以及輸入方法的評估等等均是。

由以上的討論，我們已經了解：中文資訊處理的研究涉及許多專門學科，是相當複雜的一門科際整合形態的新研究領域。在這樣的情形下，我們還需要一些知識來駕御這些錯綜複雜的科際關係，才能做好中文資訊處理在科學上的研究和工程方面的發展。這些知識包括模控學（cybernetics）、系統科學以及管

理科學，此外在工程發展方面則須重視系統工程（system engineering）的學養。模控學和系統科學都是觀察和歸納複雜的系統現象，並以之推導出對系統的了解（知識）的學問。雖然它們之間有對自然系統和人工系統的分野和差異，然而對中文資訊處理的研究而言，二者的觀念和素養均為必需：因為中文資訊處理的研究不只涉及機器系統（人工的），更涉及到「人」以及在「人的社會組織形態下」，如何使這機器系統能夠良好的運作以充分發揮其效能。是故它涉及與人有關的系統（自然的）和管理科學的知識了。此外，由工程的立場來看，要開發這麼複雜的產品，系統工程的知識當然重要。

綜合以上所述，中文資訊處理的研究不僅是計算機科學和語文學的深入結合，還經常涉及下列三類的專門學科，是典型的科際整合的形態。

一、心理學、人機工學、傳播學（科技、語文傳播）；

二、智識論（哲學）、認知科學、圖書館學（分類學、資訊系統）、腦神經醫學；

三、模控學、系統科學、管理科學、系統工程。

基本研究問題的分類

中國語文有其特色。做中文資訊處理時依語文之特性不同，而演繹出各種的處理方法，所以中文資訊處理可以據此分類。依文獻結構的元素來分，處理的對象可分為：字、詞彙、片語、句子、段落、文章等。由文法結構的層次來說，則可分為語法和語意兩大部分。目前的商用系統，還停留在只能處理「字」的階段。詞彙或更複雜的形式，則無理論的模式可用，只能依應用問題的性質寫些依附資料（date dependent）的程式作特殊的解決之道。這樣的做法使得一個程式只能解決一個問題，而無法與同樣性質的問題共享，其投入之成本自然高漲，且對複雜的問題則無法做有系統的解法。至於以語法與語意二者，還只是研究中的對象。在研究室裡，上述各層面的問題雖均已有小小的涉

獵，可惜計畫之規模較小，且久缺長期的恒定性，以致於目前的成就不大，離實用仍有距離。

另一個角度的分法是由語文表達所用的媒體來區分。可分為三類：形、音以及碼。「形」是指人類以視覺功能處理的文字外觀，它包括各種印刷、顯示、或書寫的形態。以「音」表達者就是以人類聽覺可以處理的語文形態，稱為語音。「碼」是指以計算機可以閱讀的方式所作的表達。它包括數位化的字形點矩陣、數位化的語音訊號、字的認別碼、交換碼、檢索碼等等。

第三個角度的做法是以資訊處理的基本功能來分。計算機本質上是自動機（automata），所以由自動機的基本功能能來說，研究的問題可分為產生和識別二大類。若由應用的角度來細分其功能，則定義二中之各項運作都可視為資訊處理的基本功能。

上述的三個分類角度是各自獨立的，可以組成一個三度空間。在此空間中一個點則可代表一種中文資訊處理的基本研究問題。例如做字形產生的研究是位於「字」、「形」、以及「產生」的交會點；又如語音識別為「字、詞、句子」、「音」以及「識別」等交會處之總稱。此結構表現如附圖。圖中的三個坐標分別是：語文結構、語文表達的媒體、和資計處理之基本功能。以此分類，可以對中文資訊處理面臨的種種基本問題作一綱領式的了解。

以目前研究的情形而言，許多基本問題尚待努力，而研究範圍則局限於「句子」以下的簡單語文結構。若是我們把功能概略納入產生和識別兩類，把語文結構約略分為目前常見的處理對象－－字、詞、句，則附圖可化簡為分別以形、音以及碼之三個平面，而每個平面上則有「產生、認別」與「字、詞、句」等組成之六類問題。這個約略的分類可以含蓋了目前所有中文資訊處理研究的基本問題。（待續）