「國音」智慧型輸入系統的語意分析

「脈絡會意法」

許聞廉，陳克健

台北市南港中央研究院資訊所

Abstract

Phonetic input is the most popular Chinese keyboard input method among non-professionals. To ease the burdon of homophone selection, we have developed an automatic phoneme-to-character conversion expert system (with tones). Based on lab tests using a corpus of eight million Chinese characters from newspapers, the system exhibits an average hit ratio close to 95%. The best feature of this system is that it can be continually improved without any conceivable limit.

The system is constructed based on an assumption about human understanding. A pattern matching approach is adopted to process the semantic and syntactic structures of a sentence; the weights and marginal values of each pattern is calculated statistically.

摘要

在軟硬體進步神速的今天，電腦中文化最大的瓶頸就在於沒有一種免學易用的「輸入法」了。目前市面流行的輸入法不下十餘種。但供非專業人員使用的拼音輸入法重碼率較高，使用者兩眼奔波於螢幕與鍵盤間，不勝其苦。我們以四年的鑽研，建構了一套自動辨認同音字的專家系統。在八百萬字的報紙資料上測試的結果，平均正確率達到百分之九十五，使拼音輸入（帶聲調）進入實用的階段。。這個系統最大的特色是能夠不斷地修正、改進，不像一般以字、詞頻爲主的統計方法有其無法突破的瓶頸。

我們根據一些對人類「理解」系統的假設，使用Pattern matching 的方法，建立各個字、詞出現的環境特徵，並以統計決定各個樣版的強度與邊際效用。

一、 背景

中文在自然語言的處理上是相當複雜的。一般人在使用中文時相當地靈活；然而，中文文法所提供的訊息經常含混不清，造成了電腦處理上的莫大困擾。我們發展了一套處理中文自然語言的方法，稱爲「脈絡會意法」，非常適合處理電腦上的中文資訊。脈絡會意法可以運用的範圍包括（一）中文同音字的自動辨認；（二）中文字轉音以及語音合成系統；（三）語音辨認的後處理（音轉字以及容錯系統）；（四）OCR 、 OLCR 的後處理系統； （五）各類字形輸入法同碼字的自動選取系統；（六）中文句型剖析（PARSING）以及斷詞系統。 本文僅就「同音字辨認」系統爲例闡釋脈絡會意法的原則。以往有關的研究請參閱〔1〕~〔7〕。其中，除了〔4〕使用一部份語法外，大多是以字、詞頻的運算爲主。

中文字、詞的使用在文法結構上不像英文那樣有明顯的標示；所以，句型的分析單用「語法」成效不彰，必須借重上下文的「語意」分析，才能避免混淆。最明顯的就是，人們在對話時常需要靠上下文才能辨別同音字。這件事，由於潛意識已經習以爲常，我們反而不清楚自己的腦筋是如何作這些辨認工作的。這也就造成了自然語言研究上的一大困擾。 一般說來，上下文所提供的訊息相當的多，而且經常有相互矛盾之處；所以，如何能抽絲剝繭，去蕪存菁，歸納出最合適的結論才是最重要的。但是，語意分析牽涉甚廣，所需要的專業知識浩瀚無邊，即令是一般常識也有無從下手

之感。因此，時下絕大部分的句型剖析系統都利用統計頻率（譬如，用在斷詞部份）以及語言結構的一般文法規則去作無意識的「運算」，以達成一種概括性的解決方法。以下，我們就兩者之間做一個比較。

二、語意分析與詞彙配對的區別

一般以詞彙配對、頻率爲主的分析方法不外乎存詞、統計字詞的出現頻率以作爲同音詞及搶詞的判斷依據（通常都使用動態規畫的方法）。在使用時可以發現一個現象：字、詞的修正都是在您剛剛輸入或修改的字旁邊的一、兩個字。無法影響到六、七字以前的部份。也可說，目前其他系統所處理的只達到「詞」的層次。「脈絡會意法」的分析則可以達到「句子」（SENTENCE）甚至「言談」（DISCOURSE）的層次。譬如，使用語意分析辨別下列「同音字」（「隻」和「枝」）在詞彙輸入法就很難達到：「一隻可愛的小貓」，「一枝可愛的鉛筆」。以下，我們列出脈絡會意法與單純使用詞以及統計頻率方法的最明顯的區別：

一、統計頻率隨著不同的文章題材而有不同的分佈，如果在同一篇文章中摻雜著不同的題材，使用起來就非常地不方便。相對的，脈絡會意法的基本原則就是在不同的領域內發掘出不同的規則，所以規則之間極少有互相矛盾的（只是變得更細膩）。因此，同樣的規則資料庫可以很容易地適用於不同的使用者。

二、使用統計頻率所能達到的轉換正確率通常很難超過 90% 。 然而，脈絡會意法則可以將正確率不斷地提昇至理論的極限。

三、統計的優點在於大部分的分析皆可以電腦爲之，人力上的需求不高，也因此在理論上可以改進的可能性很低。相對的，脈絡會意法必須對各個不同的領域做深入的探討研究才能歸納出合乎其題材的規範。因此，需要大量的腦力分析，進行地毯式的研究，但其成效也相當地明顯。

四、單純的詞彙頻率收集以及比對用在「音轉字」系統上所能達到的正確率只有 80% 到 85% 左右。這是因爲「同音詞」本身所產生的混淆以及詞的「界線」不清所致。這些問題不能僅靠存詞解決，有些甚至是詞存得愈多問題愈大。

三、國音系統語意分析的基本原則---「脈絡會意法」

傳統的「句型剖析」法將所有可能的句子分解方式一一列出，再加上語意的匹配以決定最後合適的對應國字。這樣的作法速度非常慢，而且需要相當多的語言學知識輔佐。國音系統設計的目標是要儘量模擬人類的「理解」（UNDERSTANDING）系統。以下是我們對人類使用語言的一些觀察：

一、我們在聽人講話或看書時，經常一筆帶過，不見得很清晰地捕捉到每一個音或字，但並不會影響到「瞭解對方」。可見，人類具有相當高強的「提綱挈領」能力。

二、當我們在學習一個新的字或詞的用法時，往往需要觀察許多這個字、詞出現的句子後才有相當的把握。 在學習英語時尤其如此。

三、中國人寫的英文即使合乎文法，但是看在老美的眼裡往往格格不入；這經常是因爲他們不「習慣」這種寫法。

四、新的字、詞，如果以聯想的方式將之記下，則記憶可留存較久。

五、無意義、單調的組合、串連（譬如電話號碼）非常的難記；反倒是越複雜越有規律的事物記起來比較容易。

人類的常識判斷與理解過程是人工智慧研究中最不可捉摸的。國音的語意分析基本上就是儘量模擬人類的思考方式，將常識性的邏輯判斷轉變成電腦可以理解的符號運算。我們先看看下面這個例子：「台北市一位小孩昨天走失了」。這句話如果只打出它的注音，則可能組成的詞就相當的多。讓我們暫且看一下前面五個「音」所能構成的詞： 台北市，台北，事宜，適宜，一位，一味，移位。下面這個例子顯示，有時長詞並不一定「優先」：「台北是一個美麗的城市」，「台北市一個美麗的女孩走失了」。顯而易見的，單純地以字、詞之間的關係來「計算」永遠無法讓電腦瞭解句子的意思。

從語言學看來，中文並不像英文那樣有明顯的標記（譬如：英文的單複數、動詞時態等等）。表面上，這會造成電腦分析中文很大的困擾。然而有趣的是，由於上下文的輔助，標記的缺乏並沒有造成人們「理解」上太大的障礙。對於一個名詞在某些地方被當成形容詞，或動詞被當成名詞，我們似乎習以爲常，毫不在意。當我們觀察一個字、詞出現方式的經驗累積到相當的程度時，大多能隨性使用而不逾矩。這類經驗通常並不表示我們對這些詞類的「文法學理」有了深邃的認知（因爲絕大多數的人並沒有受過嚴格的語言學訓練），只能代表我們對它們的「應用規範」有了足夠的體認。如果後者足以構成「理解」的充分條件，我們也許可以從語言「現象」的分析得致「理解」所需的大部分要素。「國音」系統「脈絡會意法」的基本假設是：

人類理解的方式主要是依靠「樣版」(TEMPLATE)的記憶、聯想及推論

在自然語言中，有關一個字、詞的「樣版」就是這個字、詞所有出現的「情況」，也是綜合「語法」、「語意」的特徵規則。以下我們以「音轉字」系統爲例闡釋「樣版」理論的概念：首先，我們以「一隻非常可愛的貓」爲例，說明有關「隻」的樣版形式。首先，將各個字、詞的詞類標示在其下方。

一　隻　非常　可愛　的　貓

【數詞】【量詞】【副詞】【形容詞】【助詞】【名詞】

從這樣的詞類次序關係，我們大略可歸納以下的樣版：（其中圓括弧內的副詞表示可有可無；「｜」 的記號表示緊鄰。）「【數詞】｜隻｜(【副詞】)｜【形容詞】｜的｜【動物】」。當然，樣版中的「形容詞」和「的」也可以省略。如此可得「【數詞】｜隻｜【動物】」。另外，也有一種用法是「他買了小貓兩隻」；可以得到「【動物】｜【數詞】｜隻」。

類似這樣的樣版可以規範出「隻」的使用規則。當我們收集了足夠多「隻」的樣版後，「隻」字就能掌握自如了。讓我們再回頭仔細地想想「樣版」的功用。當我們輸入了一個單音「ㄓ」時，有許多可能對應的字；但是當上下文裡陸續出現了許多其他的訊息時，國音系統內存的樣版就會一一地與之對應，如果其中一個樣版對上了，「隻」字就會出現。以統計機率來說，單音的混淆程度最大，雙音次之（譬如同音二字詞或字串）；但是這兩者光靠頻率都無法有效地予以確認。樣版，因爲是一種有意義的複雜組合，能夠造成混淆的機率就非常之低了。而且，即使樣版之間有可能混淆，我們也可以很容易地在它們的強度上稍做區別，予以避免。國音系統就是利用這樣的原理，在上千萬的資料庫中，分析收集了幾萬的樣版，並且適當地調整各個樣版

的強度，以解決樣版之間相互的衝突，使得使用時修改錯誤的次數降至幾可忽略的程度。我們深信，「樣版」的使用是最能有效地捕捉上下文語意和語法結構的一種方式。基於「樣版」的假設，我們對前面提到的幾個現象有以下的解釋：

一、大部分樣版的複雜度都很高，混淆的可能性極低，甚至於通常只需要其中兩、三個訊息就已經能意識到它的存在。所以，即使遺漏了其中少數的訊息，對我們的瞭解（如果看成是樣版的認定過程）也不會有太大的影響。

二、背英文單字，生吞活剝總是遺忘得快。如果這個單字能經常在閱讀中出現或者我們能觀察許多這個單字的例句，漸漸形成了有意義的樣版後，就不易淡忘了。

三、合於文法的樣版並不一定就是母語使用者「慣用」的樣版。

四、「聯想」基本上就是借用舊的樣版， 所以不構成記憶上的太大負擔，相形之下，記憶保存的時間也比較長。

五、一個樣版的形成，不是因它本身具有某種意義，就是因它使用的頻率非常之高。所以，我們對一個陌生的電話號碼無法很快地形成樣版，深植腦中；相反地，對自己喜歡的歌曲甚至冗長的交響樂卻都能毫不費力地朗朗上口。

許多人也許會認爲：類似這樣的樣版林林總總，個數可能不下數十餘萬，甚至百萬，以目前國音系統少量的樣版，如何能保證「高人一等」的正確率呢？其實，樣版的個數雖多，但類型變化卻相當有限（否則，一般人如何能記得了？）。所以，我們將同型的樣版歸成一類（頗類似人類的「聯想力」）；分類之後，再以統計方式計算出各類樣版的強度和邊際效用；最後，在考量整體正確率、系統記憶容量及速率的因素下，將最主要的樣版建入國音系統。

以這樣嚴謹方式建構的系統，您可以很容易瞭解爲何我們有「知錯能改」和「不貳過」的自信了。任何一個新的錯誤都提供了我們建立新樣版或者解除舊樣版衝突的機會。當一個新樣版有足夠的出現頻率，我們就會將之納入系統；至於解決樣版間的衝突，國音系統內也提供了開發人員非常多的「樣版修改」程序，大致都不會造成任何問題。總之，在「樣版理論」的假設之下，國音系統將「辨認同音字」的問題轉變成「樣版的比對」； 將文字學上的一個基本問題用演算法(algorithm) 的理論來處理。不但避免了句型剖析的困擾以及統計頻率的瓶頸，也提供了一個新的句型剖析方式。

四、「樣版」的重要性質

以下，我們討論樣版的形態、蒐集以及檢索的問題：

「樣版」之基本形態大致有：「名詞片語」：一隻非常可愛的貓；「【數詞】｜隻｜(【副詞】)｜【形容詞】｜的｜【動物】」；「動詞片語」：洗了一個很舒服的澡；「洗｜(了)｜【定詞】｜(【副詞】)｜【形容詞】｜的｜澡」；「簡單句子」：他用斧頭把這枝樹幹劈成柴火；「【人】｜用｜【工具】｜把｜【物體】｜【動詞】｜成｜【物件】」。

脈絡會意法可以應用的範圍極廣，樣版的蒐集方式也不盡相同，完全視需要而定。譬如說在「音轉字」系統中，樣版蒐集的對象主要是在於區別同音字、詞以及幫助斷詞的規則。在「字轉音」系統中樣版則主要在幫助斷詞點的確定。樣版蒐集的來源主要

是各種「語料庫」，可以是文字的、或語音的。蒐集方式則是利用電腦統計以及語意分析師的專業判斷。

一個樣版通常依附在（或記錄於）其最重要的「成分」（ KEY）上。如果一個樣版的重要成分多於一個時，就有可能被記錄多次。這些成分主要是由「連續的」字串或字和語意的組合。其中，字串是指一些習慣用的字組，當然也包括了我們通常所謂的「詞」。譬如：我們的字串中有「有時」和「有十」以及「台北市」和「台北是」等等。以前述的名詞片語的樣版爲例，「一隻可愛的貓」的樣版可記錄在「隻」上，也可記錄在「貓」上，需要視實際情況而定。我們可以對這些主要成分事先加以「排序」，以加速檢索。

兩個樣版中可能有部份重疊，造成互相抵觸的現象，這時系統內可事先將這兩個樣版的「強度」予以標定，預先決定當兩者同時出現時，何者優先被使用。譬如，在醫學名詞中有一個樣版：「【器官名稱】｜科」；在職業名稱上有一個樣版：「【姓】｜【職業名稱】」。這兩個樣版就可能造成以下的同音字「互搶」：皮膚科、柯醫生。當「皮膚科醫生」出現時分析師就被告知應該將第一個樣版的強度提高，使其「勝過」第二個樣版，「科」字才會辨認正確。當然，如此的強度調適有可能在另一種情形產生不合適的效果。這時，這兩個樣版就可能需要再加以細分，使其更爲精確，避免重疊。這些改變的取捨原則可以由統計決定之。

另外一種方法是將這兩個詞分別與鄰近的字、詞構成片語、子句，進而合成整個句子。在此過程中如果其中某一個詞無法構成有意義的組合，則自然被淘汰掉。由於這類方法耗時較常，只有在必要時使用之。

五、「脈絡會意法」的應用範圍

一、語音輸入的後處理系統：語音辨認的技術目前尚未完全成熟；事實上，「語音」本身就存在著許多容易混淆的盲點。我們發現：一般人在聆聽他人說話，都經常需要藉助上下文的判斷，以確定其相對應的「字」而不致產生誤解。「音轉字」系統不但在音確定的情況下能將之轉換爲正確的字，並可以進一步地在音不確定時有效地幫助其判斷正確的轉換（也就是所謂的「容錯」）。此外，在有限詞彙不特定語者的應用上更適合使用脈絡會意法。

二、語音合成的輔助系統：這類「字轉音」系統最大的困難就在於詞的界線不清。所以，音轉字系統所歸納出的各項有關詞的規則正好可以協助其確定斷詞點。

三、OCR 字形辨認後處理系統： 其原理就如同輔助語音辨認的後處理一般，只是現在所要分辨的不是同音字，而是同形字。當然，所使用的規則本身因爲競爭的對手不同會有很大的改變。

四、「倉頡」及「大易」輸入法之輔助系統：這些以「拆字」爲主的鍵盤輸入法經常遇到因字碼記得不十分正確而影響輸入的速度；此外，「大易」輸入法的重碼率也較高。這些都可以使用脈絡會意法建立「容錯」專家系統加以改進。

五、脈絡會意法應用在句型剖析（PARSING）上： 樣版比對的方式可以直接用在確認簡單的名詞、形容詞片語以及句型上面。在一個複雜的句子中，我們先利用樣版比對在第一個循環找出其中的簡單的名詞、形容詞片語以及簡單子句。在下一個循環的樣版比對就從這些片語、子句出發，一層一層地將複雜的句型予以簡化，

直到最後剩下一個簡單句子，用單一的樣版就可以確認的地步爲止。對於變形句子，則以樣版事先決定其對應的基本句子，再以上法分析之。

六、中、英文檢錯系統：當中文系統使用的樣版收集到相當的數量，這些樣版就可以拿來和任何已經打好的文章作比對，並提醒使用者「不恰當」的語句用法。此外，樣版理論可以應用在任何語言，所以也可拿來做英文的檢錯系統。

六、系統正確率的評估

我們評估正確率的資料主要來自「計算語言學會」約千萬字的新聞語料庫。外加一百萬字的口語資料，三十萬字的公文信件資料，以及八十萬字的小學課本、作文等資料。我們首先將語料庫內的字轉成音（經斷詞將破音字找出）。然後用國音系統將這些音轉換成字，再和原文比對。注音有一千三百多種變化，我們對每一個音隨機選擇五千個含有此音的句子。以上法將這些句子轉換後，統計這個音對應之同音字的正確率，作爲此「音」的轉換正確率。再根據每個音出現的頻率加權計算出平均正確率約爲 94.78 % 。 如果去除新聞資料中的專有名詞錯誤比率，正確率可以再提高百分之一、二左右。在一般的通俗文章上，相信正確率可以更高，完全符合實用的要求。

七、結論

語言學上的構詞、句法以及各種的分類，對於「脈絡會意法」的規則建立有莫大的幫助。然而，過於細微的分類有時也會造成無所適從的困擾，主要在於中文常有「一詞多用」的情況。有時爲了瞭解某個詞的確切用法，必須觀察相當長的一段上下文才不致混淆；這也大大地增加了規則建構的困難度。還有許多情況，精確的語意描述可能大爲提高電腦分析花費的時間。此時，計算理論中的一些技巧能將問題適度地轉換，使之較爲容易處理，並且與原來的結果相差無幾。因此，在系統建構的過程中，求得一個高精確度與高效率之間的最佳平衡點，就變成一個最重要的考量了。

參考文獻：

1. Kuo, J.J., et al., "The development of New Chinese Input Method - Chinese Word String Input Method," Proceedings ofInternational Computer Symposium, Taipei, (1986).
2. Chen, S.I. et al., "The Continous Conversion Algorithm of Chinese Character's Phonetic Symbols to Chinese Chacters,"Proceedings of National Computer Symposium, Taipei, (1086).
3. Fan, C.K. and W.H. Tsai, "Disambiguation of Phonetic Chinese Input by Relaxation-based Word Identification,"Proceedings of ROCLING I, (1988), 145-160
4. Hsieh, M.L., T.T. Lo and C.H. Lin, "Grammatical Approacht o Converting Phonetic Symbols into Characters," Proceedings of National Computer Symposium, Taipei, (1989), 453-461.
5. Sproat, R., "An Application of Statistical Optimization with Dynamic Programming to Phonemic-Input-to-Character Conversion for Chinese, " Proceedings of ROCKLING III,(1990), 379-390.
6. Gu, H.Y., C.Y.Tseng and L.S. Lee, "Markov Modeling of Chinese Language for Linguistically Decoding the Mandarin Phonetic Input, Proceedings of National Computer Symposium,Taipei, (1989), 759-767.
7. Chang, J.S., S.D. Chern and C.D. Chen, "Conversion of Phonemic-Input to Chinese Text Through Constraint Satisfaction," Proceedings of ROCKLING IV, (1991), 30-36.