信息結構化方法與自然語言處理機制研究

趙文銀

北京乾坤化物數字技術有限公司 北京

kingque@163.com

摘　要： 從本質上講，無論是人腦或者計算機，所謂的自然語言處理實際上就是對符號的處理。不同區域的人群由於采用了不同的符號以及不同的符號連接習慣，因此出現了不同的語言。自然語言是隨意和習慣的產物。形成自然語言的關鍵在於信息結構化技術，該技術把任何一個輸入信息轉換到結構樹上的唯一結點上，本文將通過雲計算產品來說明該理論在計算機上處理自然語言的過程，並由此假定人腦也是采用這種方式處理自然語言的。

關鍵詞：詞匯，自然語言，雲計算，信息結構樹，信息關聯通道

Structured Method of information and Research on Mechanism of Natural Language Processing

Wenyin Zhao

Beijing QianKunHuaWu Digital Technology, LLC, Beijing

kingque@163.com

Abstract: In essence, for either the human brain or computers, the so-called natural language processing is symbol processing. Because people in different regions use different symbols and different connecting symbols, there are different languanges.. Natural language is the result of randomness and habits. The key to natural language is the structured technology of information. The technology converts any input into the structure of a node on the tree, this paper will demonstrate the theory using cloud computing products We hence assume that the human brain is using this approach for natural language processing.

Key words: Vocabulary, natural language, cloud computing, structured tree of information, associated channels of information.

1　前言

我們目前無法知道人類的語言是什麼時候、什麼原因產生的。或許是當人類出現的時候就有了語言；或許是為了生存，人類需要交流信息，因為一種特殊的原因而產生的。我們能夠確定的是，語言是通過人的大腦來完成信息組織，然後通過口語或者書面語等形式進行傳遞的。

人類的生活至少離不開兩個世界，物理世界和虛擬世界，物理世界是我們的身體所寄存的世界，虛擬世界是大腦的產物，自然語言是在虛擬世界裏完成的。由

於受物理世界自然規律的約束，人類目前無法通過物理實驗的方式獲得人腦處理信息的准確機制，然而計算機的出現給人類提供了新的處理方法，通過模擬結果的方式來了解人腦處理自然語言的機制。

本文觀點和結論的依據來源於一項申請的發明專利《文字語言結構樹的構建方法》，該方法表明，在以計算機為載體的虛擬世界裏，任何信息都可以通過一個數學方法唯一地映射到一個結構樹的結點上。作為探索和研究，同樣可以假定，在以大腦為載體的虛擬世界裏，任何信息都可以通過一種方法唯一地映射到一個大腦神經元結點上。

2 重新認識自然語言處理機制

目前研究自然語言處理方法時，絕大多數研究人員會假定自然語言是一個有規律的體系，然後設計一些複雜的方法來處理這些規律，比如統計、搜索、比較等方法。而事實上人腦在處理語言信息時，幾乎是在瞬間完成的，不可能進行搜索、比較等操作。

自然語言處理的定義

自然語言處理是指在具有特殊組織結構的裝置裏，存儲了數量龐大的信息，通過信息之間的關聯關系，采用非線性方式，對信息進行重新組織以便獲得新信息的過程。

從這個定義裏可以發現，自然語言處理必須具備以下幾個基本條件：

具有特殊的組織結構。

存儲了龐大的信息。

每個信息都有大量的關聯信息。

信息內容和存儲位置的統一。

（二）語言的形成

群居動物為了獲得更好的生存環境和生存質量，需要約定一些固定的信息來表示具體的事物，這些固定的信息存儲在對應的神經元結點上，當某個動物發出這個信息時，其它動物就會從對應的神經元結點上獲得事物的圖象，這就是語言。

由於一個信息對應一個神經元結點，所以神經元的數量決定了語言的豐富程度。人類之所以有豐富的語言，是因為有數量龐大的神經元。

語言是人類在生存發展過程中逐步發展起來的，不同區域的人群，其語言也是不同的。

3 人腦是如何處理自然語言的

人腦或許是人類認識的東西裏最神秘的了：信息是如何編碼的？記憶是怎麼被存儲然後又被提取的？對未來的規劃是如何形成的？到底什麼是意識？為什麼大腦會休眠和做夢？

盡管目前人們不能在物理世界裏通過實驗方式探索人腦的運行機制，但是我們可以在以計算機為載體的虛擬世界裏通過模擬人腦的方式來分析研究這些現象。

假定物理世界的人腦和虛擬世界的人腦具有相同的運行機制，然後根據物理世界人腦產生的結果推導出可能的實現方法，在虛擬世界的"人腦"中應用，如果產生同樣的效果，那麼可以認為這個假定是合理的。

（一）人腦是如何處理詞匯的

人腦本身是一個比較簡單的計算和存儲裝置，其特殊的應用功能來自兩個方面：一個是把信息轉換為神經元結點位置的方法，另外一個是神經元的組織結構。

當人腦從外部獲得一個"詞匯"時，就會把這個"詞匯"轉換為神經元結點位置，然後在該結點位置存儲或者提取相關信息。

人腦通過學習獲得與該"詞匯"關聯的信息，並建立該"詞匯"和關聯信息的連接通道，從而形成複雜的組織結構。

（二）早期的學習促進了大腦的發育

在大腦早期的發育過程中，獲得新的詞匯以及其它信息是有利於新的神經元產生的。剛出生的嬰兒，大腦神經元的數量是有限的，並沒有達到最大容量，因此在出生後，接收更多的不同類型的信息刺激，會促進神經元的生長。因為當大腦獲得一個新的信息後，會按照規則轉換為神經元結點位置，當這個結點位置不存在時，就會生長出新的神經元。

（三）大腦如何把詞匯組織成語言

大腦把詞匯關聯起來後，就形成了語言。所以只有當人腦存儲了足夠多的詞匯，並且建立了詞匯的關聯關系後，才能形成語言。

學習的詞匯越多，學習的語句越多，存儲在大腦神經元的信息就越多，那麼其語言表達能力就越強。

4 如何用計算機處理自然語言

豐富的語言能力是人類區別其他動物的重要特征之一。在所有生物中，只有人類才具有語言能力。人類的多種智能都與語言有著密切的關系。人類的邏輯思維以語言為形式，人類的絕大部分知識也是以語言文字的形式記載和流傳下來的。

盡管到目前為止，人們一直認為人腦是一個神秘的東西，但是卻從來沒有放棄過探索和模仿人腦功能的想法。自然語言是人類生存、發展的最重要的基本工具之一，那麼自然語言是如何形成的？計算機能夠象人一樣處理自然語言嗎？

本文將結合一個具體的雲計算產品——"中華語言知識雲"，介紹一種通過構建類似人腦結構的系統處理自然語言的方法。

（一）核心技術及產品特征

實現人腦處理自然語言能力的最簡單、最直接的方法就是創造出和人腦結構一樣的存儲處理裝置，即在計算機裏創建一個網狀的、存儲結點數量至少和人腦神

經元數量相等的體系。

中華語言知識雲是一個強人工智能的學習系統，目的是作為一個具有高度智慧的"教授"來"傳道授業解惑"。該體系具有以下的技術特征：

網狀體系是由一棵數量龐大的結構樹，以及每個樹結點同其它結點連接的邏輯通道構成。

任意一個來自外部（或者內部意識產生）的信息，通過一個數學方法換算，都可以對應一個唯一的樹結點位置。

系統不直接對所接收的外部（或者內部意識產生）信息進行處理，而是首先把信息轉換為樹結點位置，然後對結點位置進行處理。

由於是對結點位置進行處理，所以對任何信息的處理速度基本一樣。處理信息所消耗的時間和信息數量無關。

學習是系統形成的重要步驟。學習的過程就是樹結點的形成和完善過程。新的信息是通過學習獲得的，而已經存在的樹結點，在學習的過程中不斷調整完善和其它樹結點之間的關聯關系。

當存儲的信息以及關聯信息的樹結點達到一定數量時，自組織行為使得系統可以創造出新的信息，而這些新的信息同樣可以產生新的樹結點。

該雲計算產品說明了以下幾個觀點：

自然語言是可以通過混沌系統來產生的。這可能是最簡單的，也是無法預測結果的方法。

當存儲信息的位置和信息本身變成一個問題的時候，其處理方法也就變得簡單了。

當位置按照結構樹來管理時，其管理方法也就變得簡單、高效了。

（二）詞匯的結構樹處理方法

現在所使用的自然語言越來越複雜，但是人類最早通過語言進行交流時，必定是最簡單的符號組合。從漢字的演變可以說明這點。最早的漢字數量很少，基本上是和生存有關的事物的符號信息。

所以計算機對自然語言的處理也是從簡單到複雜的處理過程。

如同人一樣，首先需要讓計算機認識字、詞、詞組、句子。因此需要把這些固定的信息導入到"中華語言知識雲"系統，在系統裏建立最原始、最簡單的結構樹，使系統具備初級的認識能力。當外部信息輸入到計算機後，系統對信息的處理就可以轉換為對樹結點的處理。

對於任意輸入的信息，比如"中"，系統經過計算就可以得到一個唯一的位置數據，這個位置數據不是人為規定的。下面是幾個把信息換算到樹結點位置的例子：

中：bsnnifbdsibfisfffb

華：ffbnfibdfkbiknfffb

雲：nifkfibdfkbisaffdb

中華：bvdnibbfivbbividib

語言：isbdibbfknbbdkidib

知識：fzkbibbfiabdsiiddb

中華語言：dzfbibbffabdkaidfb

中華語言知識雲：kadkdbdiikbbdafddb

這是一個 9 層結構樹，每層 64 個結點，容量為 64 的 9 次方，存儲結點數量超過 2 億億個，數量大大超過了人腦的 1 千億個神經元結點。

（三）詞匯的關聯

當最基本的結構樹創建以後，開始建立樹結點和關聯信息樹結點之間的關系。比如“中華語言知識雲”和“中華”、“語言”、“知識”、“雲”有直接關聯關系，“中華”和“中”、“華”有直接關聯關系。“雲”作為最基本的字，和“雲計算”、“雲彩”、“雲端”、“雲母”、“雲遊”等有關聯關系。同時在學習的過程中，會出現新的關聯關系。

關聯關系不僅包含直接使用的信息，同時也包括其它信息，比如“中華語言知識雲”和“北京乾坤化物數字技術有限公司”有關聯關系，因為該系統是該公司開發的雲計算產品；“中華語言知識雲”和“劉兆玄”、“中華語言知識庫”有關聯關系，因為劉兆玄會長提出“中華語言知識庫”也是基於雲計算的產品；“中華語言知識雲”和“中國辭書學會”有關聯關系。

關聯關系所包含的內容，采用文件方式存儲在樹結點目錄下。

關聯方式分為有方向關聯和無方向關聯。有方向關聯分為上行關聯和下行關聯兩種。

（四）區域的劃分

人腦對信息的存儲處理是按照信息的類型來進行劃分的，盡管目前還不能准確地獲得詳細的區域劃分，但是科學家們已經知道了一些區域，比如聲音、圖象、邏輯、數字、符號、交際等。同樣，模擬人腦的智慧系統也會劃分同樣的區域。目前所開發的“中華語言知識雲”系統只是由符號所形成的自然語言處理系統，因此分區會少一些，比如詞條區、語言區、情感區等。

同一個信息有可能存儲在不同的區域。

（五）學習是什麼

學習就是為信息建立關聯關系，然後把信息存儲到對應結點位置的過程，正確地學習方法是存儲該信息的盡可能多的關聯信息。

中華語言知識雲系統的學習過程如下：

強行灌輸。比如導入字典、各種詞典（醫藥詞典、計算機詞典、生物詞典）等。

人工建立關聯關系。在初期，由於系統裏存儲的信息是孤立的，不是一個混沌系統，因此需要人工幹預，幫助系統建立信息的關聯關系。

機器學習。當系統變成混沌系統以後，就可以出現自學習的現象了，因此可以給機器提供大量的資料，通過對信息的處理，建立更多的信息結點以及關聯關系，

同時可以在學習的過程中修改調整已經存在的關聯信息。

（六）自然語言的形成機制

當存儲信息的結點達到一定的數量，並且建立了信息之間的關聯關系時，語言就可以產生了。盡管這個結論太簡單了，有點不可思議，但是其中所包含的一些技術效果是已經得到了科學界的認同：

混沌效應。龐大數量的信息結點構造的一個確定性的系統，由於關聯結點的差異，以及關聯信息的激活順序不同，其得到的結果具有了不可預測、不可重複的特性。

自組織行為。不同結點的信息按照最合理的方式進行組織，獲得對給定信息的響應。

非線性關系。信息按照不規則的方式組織，因此對同一個信息的響應，有可能在不同的時間裏得到不同的結果。

當系統接收到一個信息指令的時候，系統首先會把信息轉換為存儲結點位置，然後從該位置提取合適的關聯信息並激活關聯信息對應的結點，然後把相關信息組織在一起並且輸出，就形成了語言。

如果接收的信息在系統裏沒有對應的存儲結點，或者存儲結點裏沒有關聯信息，那麼將無法獲得合適的輸出信息，因此語言交流將受到阻礙。

存儲的信息數量越多、信息之間的關聯關系越豐富、語言的表達能力就越豐富，語言表達的准確度也就越高。

可見，自然語言的形成過程本質上就是存儲的信息的非線性提取過程，語言的豐富程度由系統的混沌程度決定，語言的准確程度由系統的自組織能力決定。

5 自然語言的語法是什麼

在現代漢語規範詞典裏對語法的解釋是：語言的結構規則。包括詞法和句法。

在自然語言的形成過程中，本身並沒有遵循“語法”規則，是一種隨意和習慣的產物。語法是當人類知識累積到一定的程度後，人們為了探索和了解自然語言而進行的規範和分類。

無論人們對語法的研究程度有多深，制定的規則有多少，都不會對自然語言本身產生多大的改變。自然語言是通過人腦的特殊裝置形成的，人類總結和制定的語法可以作為信息來影響語言的組織，但是不會有本質上的改變。比如對於一個完全不符合語法規則的語句，人們仍然可以很方便地理解。對於“兔子紅色公路跑”、“公路跑紅色兔子”這樣的自然語言，人腦很容易理解，本文所引用的雲計算產品同樣很容易理解，但是這些句子不符合語法規則。

在“中華語言知識雲”系統裏，沒有任何預先設定的語法規則，只是單純的對符號進行處理，自然語言是隨意和習慣的產物。“語法”僅僅是處理結果的外在表現形式。

6 中文和英文以及其它語言的區別

據估計,世界上大約有 5600 多種語言，對於人腦和計算機來說，這些語言沒有任何差別，都是對符號的處理。人們也可以通過約定符號的方式，隨意創造出新的語言。

7 結論

從雲計算產品的研究開發中 ,發現自然語言的形成機制本身所采用的數學方法比較簡單，其複雜性來自於組織結構。比如同一個信息可能分布在不同的區域，同時激活才能得到結果；同一個信息不僅和同一區域的很多信息關聯，也可能和不同區域的很多信息關聯。

自然語言不是一個具有嚴密邏輯推理的產物 ,而是一個隨意和習慣性的符號的組合。

具有混沌效應的網狀結構裝置是形成自然語言的基礎。

不斷學習是增加信息結點和調整信息結點的重要手段。

輸出的語言信息只能來自存在的信息結點上 ,不可能出現語言的某個構成信息在結構裝置的結點上不存在的現象。自然語言只是存在的信息的自組織。