LANGUAGE AND LINGUISTICS 9.2:381-393, 2008
2008-0-009-002-000261-1

詞語義項的表徵及其可計算性研究*

張仰森[1,2]　吳雲芳[2]　俞士汶[2]
北京信息科技大學[1]
北京大學[2]

對漢語詞語義項的意義訓釋表徵、語義原語、概念依存、二維結構表等幾種表徵方式的優缺點及其可計算性進行了分析，爲計算機的自動詞義消歧的詞典知識源選擇提供了依據。

關鍵詞：詞義表徵方法，可計算性，自動詞義消歧

1. 引言

多義詞在自然語言中非常普遍，也是自然語言計算機處理中的主要障礙，研究多義詞的處理已經成爲自然語言處理研究中的熱點。所謂多義詞是具有兩個或兩個以上互相聯繫的不同意義的詞。產生多義詞的根本原因是不同現實現象之間存在著這樣那樣的聯繫，語言社會在認識到這些聯繫之後就會在兩類現象之間建立起一種聯想關係，並用表示甲現象的詞去表徵與它有聯繫的其他現實現象，從而使該詞具有多個意義。

義項是對詞語的不同意義的分項說明或解釋。不同義項的性質也不完全相同，有些是能夠獨立造句的詞義，有些則表現爲獨立造詞的語素義。在語言詞彙中，越是日常交際中常用且歷史悠久的詞，越可能成爲多義詞，且多義詞的義項個數也可能越多。因爲義項的增加一般都是在語言的運用和發展的過程中進行的。

儘管一個詞可能具有多個義項，但在一個特定的語言環境中，通常每個詞只有一個義項能和周圍的語言環境相適應。詞義排歧的任務就是根據多義詞所在的上下文環境，確定一個與之相適應的多義詞之義項，這就需要對多義詞的各個義

* 基金項目：國家 973 項目 (2004CB318102) 和 863 (2001AA114210, 2002AA117010) 項目，中國博士後研究基金 (2005038026)。

項與其所在上下文之間的接續組合能力進行計算。而要用計算機進行這樣的計算，就必須對多義詞的詞義表徵方法及其可計算性進行研究，以利於選擇可計算性較強的詞義表徵方法用於詞義標注與消歧的模型研究。本文試圖對目前漢語詞義的表徵方法進行研究和討論，並通過研究基於這些方法的語義可計算性來探討各種語義表徵方法的優缺點，探討如何利用現有的知識資源，進一步提高多義詞詞義區分的可計算性，從而使多義詞詞義排歧在計算機上更加容易實現。

2. 詞語義項的表徵方法及其可計算性

2.1 意義訓釋的表徵方法及其可計算性

2.1.1 意義訓釋的詞義表徵方法

這是傳統辭書中的詞語釋義方式，是以自然語言來定義和解釋詞語意義的。例如，《現漢》中對“儀表”的詞義解釋如下：

【儀表】①人的外表
②測定溫度、氣壓、電量、血壓等的儀器

由於自然語言本身的模糊性和歧義性，這種釋義方式往往使知識背景不同的人在詞義的理解上都存有差異，要讓計算機通過這樣的釋義理解詞義其難度可想而知。另外，這種釋義方式具有遞歸性，即在解釋多義詞的文本中又使用了多義詞。例如：

【把握】①握，拿

這裡，“握”和“拿”是用來解釋“把握”這個詞的，而“握”和“拿”本身又是多義詞，這就形成了一種遞歸現象。正因爲這些問題，當人應用詞典釋義時，可以根據豐富的世界知識來理解詞義，但要讓計算機應用它來進行排歧則可能會引發歧義膨脹和理解困難。

2.1.2 意義訓釋下的可計算性

基於傳統辭書釋義方式而建立的機讀詞典是 20 世紀 80 年代的一種流行知識資源，人們希望能從機讀詞典中自動獲取詞彙語義的知識庫，然而由於辭書釋義

方式的上述缺點，這一目標沒有完全實現。儘管如此，機讀詞典的出現還是爲詞語的語義解釋提供了一種信息來源，使得基於這種知識資源的相關研究迅速開展起來。當時的大多數自動詞義消歧研究就是直接利用詞典中詞義的釋義文本。最經典的工作是 1986 年 Lesk 的工作，他通過計算 *Oxford Advanced Learner's Dictionary* 中多義詞各個義項的釋義部分與其所在上下文詞語間的覆蓋度，來確定詞語在當前上下文中應選擇的義項，實現多義詞的詞義消歧與標注 (Lesk 1986)。然而，由於詞典對詞義的詮釋力求簡明，一般不會很長，這樣使得很多歧義詞的各個詞義解釋與上下文詞語的覆蓋度幾乎爲零，從而無法實現義項選擇的計算。Lesk 公布的該方法的正確率在 50~70% 之間，不是很理想。儘管有人提出通過對上下文詞語進行同義詞擴展，擴大計算的上下文的窗口，進而增大計算覆蓋度的成功率，但由於窗口擴展要用到同義詞計算甚至相似度計算，僅用普通的釋義詞典很難實現。可見，意義訓釋下的詞義表徵方式，其可計算性是比較弱的。正是由於其可計算較弱的原因，基於意義訓釋的詞義表徵方式，在面向內容計算的許多自然語言處理應用中顯得力不從心，爲此許多相關的語義知識資源開始建設，HowNet、WordNet 以及 CCD 就是這方面的代表。

2.2 基於義原的表徵方法及其可計算性

2.2.1 基於義原的詞義表徵方法

在現實世界中，概念及概念之間的關係是對世界知識的反映，而概念往往卻要用詞（字、詞組）來表示，一個詞可能會表示多個概念，與該詞的每個義項相對應，每個概念反映該詞的一個語義內容。然而，關於世界知識的概念是無窮的，如果直接用概念來表示詞義，必將導致計算關係的複雜化，無法實現文本內容的計算。能否找到一些最基本的、不易於再分割的、意義最小的語義單位，這種語義單位的數量有限，又可以通過各種組合運算關係實現對其他概念的表示，就類似於向量空間中的正交向量基，從而實現文本內容計算的簡化。我們把這種最小語義單位稱作詞語義原，通過詞語義原實現對詞語義項的表徵。典型的代表就是董振東先生的《知網》。《知網》是一個以漢語和英語的詞語所代表的概念爲描述對象，以揭示概念與概念之間以及概念所具有的屬性之間的關係爲基本內容的知識系統。與一般作爲語言處理資源的詞典相比，其有如下一些特點：(a) 設計並使用了一種知識詞典的描述語言 (KDML) 對所有概念進行定義，從而使概念定義形式化，有效地保證了描述的複雜度和描述的一致性；(b) 採用有限的義原集合來表示無限的概念，詞語的釋義空間是有限的。在對概念進行定義時，把若干

與被定義概念有關的義原按一定的規則列出來；(c) 概念定義方式具有很強的可計算性，便於計算機處理。

《知網》中的“概念”是對詞語意義的一種描述，每個“概念”對應著詞語的一個義項。其所設計的知識描述語言，包括 1,500 個左右的“義原”，以及一些符號（*, %, $, @, #……）和標點（=，，，{ }），按照一定的語法公式“義原＋順序＋特殊符號＋分隔符”進行描述。這裡的“義原”是經精心研究從漢語中提取出的最基本的、不易再分割的最小語義單位。可以利用這些有限的“義原”經過知識描述語言的運算，以表達現實世界中的無限概念。另外，《知網》本身不是一種義類詞典，而是一個能描述各種概念及概念屬性之間關係的語義網絡，它的語義描述呈縱橫交錯的網狀結構，從而爲基於這種表徵方式的語義推理提供了可能，這種可推理性爲面向文本內容的理解和機器學習提供了理論上的可行性。

2.2.2 義原表徵方法的可計算性

從上面的論述可知，在《知網》這種詞義表徵方式下，基於它所描述的語言的語義可計算性大大增強，那麼在面向語言信息計算的計算機處理中，如何實現這種可計算性呢？一般情況下，針對不同的自然語言處理應用目的，這種可計算性的實現方法也會有所不同。例如，在面向機器翻譯和信息檢索的應用中，這種可計算性主要體現在詞語相似度的計算上，而在面向句法結構消歧或詞義消歧與標注的應用中，則注重多義詞與其所在上下文的詞語間的關聯度計算，對於自動問答系統很可能會要求基於這種詞義表徵方式進行語義推理計算。根據《知網》的特點，它有一套運用有限義原表述各種概念（詞語義項）的語義公式，它使得概念的描述以一種義原表達式的形式實現，並且各概念之間的關係構成了一種可用於推理的語義網絡。對這種描述概念的義原表達式的不同理解和運用，就會形成不同的相似度、關聯度或推理求解的計算方法。如何通過義原表達式實現詞語義項相似度、關聯度和可區分度等的計算，是一個很值得研究的問題。

關於《知網》詞義表徵方式下的詞語相似度計算方法，劉群、李素建 (2002) 提出了將描述概念（詞語義項）的義原表達式分解爲幾個部分，再計算各部分的相似性，在計算部分相似的基礎上得到兩個概念相似度的計算結果，實驗結果表明應用其所提方法得到的相似度和人的判定是基本符合的。將一個實詞概念的語義表達式分解成四個部分：(a) 第一基本義原描述式；(b) 輔助基本義原描述式；(c) 關係義原描述式；(d) 符號義原描述式。第一基本義原描述式的值爲一個基本義原，將兩個概念的這部分相似度記爲 $sim_1(C_1, C_2)$；輔助基本義原對應於語義

表達式中除第一基本義原外的所有其他義原（或具體詞），其值爲一個主義原描述式以外的所有基本義原描述式，爲一個基本義原的集合，兩個概念的這部分相似度記爲 $sim_2(C_1, C_2)$；關係義原對應於語義表達式中所有關係義原描述式，其值爲一個特徵結構，兩概念這部分的相似度記爲 $sim_3(C_1, C_2)$；符號義原描述式對應於語義表達式中的符號義原，兩個概念這部分的相似度記爲 $sim_4(C_1, C_2)$。於是兩個概念的語義表達式的整體相似度可按下式計算（劉群、李素建 2002）：

$$(1)\quad sim(C_1, C_2) = \sum_{i=1}^{4} \alpha_i \prod_{j=1}^{i} sim_j(C_1, C_2)$$

其中，α_i $(1 \le i \le 4)$ 是可調節的參數，且有：$\alpha_1+\alpha_2+\alpha_3+\alpha_4=1$，$\alpha_1 \ge \alpha_2 \ge \alpha_3 \ge \alpha_4$。反映了 Sim_1 到 Sim_4 對於總體相似度所起到的作用依次遞減。對各部分的權值進行遞減限制，因爲第一基本義原描述式反映了一個概念最主要的特徵，所以應賦予其較大的權值，而輔助基本義原、關係義原和符號義原對兩個概念的相似度的貢獻不應大於第一基本義原。對式 (1) 中的每一部分相似度的計算可採用下列策略：

(a) 對第一部分 $sim_1(C_1, C_2)$，由於所有的義原根據上下位關係可以構成一個樹狀的義原層次結構，所以可以採用語義距離相似度的計算辦法，近似計算兩個基本義原的相似度：

$$(2)\quad sim(p_1, p_2) = \frac{\lambda}{d + \lambda}$$

式中，p_1 和 p_2 分別表示兩個概念的第一基本義原，d 爲兩個義原在樹狀義原層次結構中的路徑距離，λ 爲一個可調節的參數，表示兩個義原的相似度爲 0.5 時，其在義原層次結構樹中的路徑距離，是爲計算參數 d 而做的標定和補充。

(b) 對於第二部分 $sim_2(C_1, C_2)$，由於輔助基本義原的描述方式不止一個，因此，計算兩個概念的這部分相似度就比較複雜，因爲很難找到兩個表達式中的相互對應部分，爲此可考慮首先計算表達式中的任意兩個部分之間的相似度，將相似度較大的兩個部分對應起來，這樣就可以將表達式分解成幾個部分，通過加權求和求得 $sim_2(C_1, C_2)$。

(c) 對於第三部分 $sim_3(C_1, C_2)$ 和第四部分 $sim_4(C_1, C_2)$，通過把關係義原相同的描述式分爲一組和符號義原相同描述式分爲一組來實現其相似度的計算。

按照上述公式 (1)，取 $\alpha_1=0.6$、$\alpha_2=0.2$、$\alpha_3=0.1$、$\alpha_4=0.1$；計算“回答”和“回應”兩個詞的語義相似度，其值爲 1.0，若看兩個詞的語義表達式，均爲 {reply|答}，證明語義是相似的；若把“回答”和“回家”應用上述公式，“回家”的語義表達式選用 {GoBack|返回:LocationFin={family|家庭}}，則得到的相似度值爲

0.07619。

語義相似度主要反映的是詞語語義的相似程度。在機器翻譯中，可以理解爲兩個詞語在不同的上下文中可以互相替換使用而不改變文本的句法語義結構的程度（朱嫣嵐等 2006）。而語義相關度則不同於語義相似度，反映的是兩個詞語互相關聯的程度。可以用這兩個詞語在同一個語境中共現的可能性來衡量。一般情況下，兩個詞語的相似度高，其相關度也較高，但兩個詞的相關度高，其相似度卻未必高。例如“踢”和“球”的相關度高，但它們的相似度卻不高。《知網》中的義原分爲六大類，每一類都是樹狀結構，各類之間又通過解釋義原相互聯繫，義原樹中的上下位關係構成了義原的相似度，義原與解釋義原的關係形成了義原的關聯度（許雲等 2005）。這樣《知網》中的義原就構成了一個網狀結構，通過計算兩個詞語間的相關度，達到句法結構消歧或詞義標注消歧，朱嫣嵐等 (2006) 還通過相似度和相關度實現詞語褒貶傾向性的計算。兩個義原的關聯度可按下式計算：

$$(3)\quad \mathrm{rel}(C_1, C_2) = \max\left(1 - \frac{d(p_i, p_j)}{D}\right)(1 \le i, j \le 2, i \ne j)$$

式中，p_i 和 p_j 分別表示兩個概念的第一基本義原；D 爲橫向關聯影響深度，即某一義原向上第幾層的解釋義原對其特徵的影響，D 取一個合適的值，如果超過這個深度，則認爲其沒有影響。$d(p_i, p_j)$ 爲義原 p_i 出現在義原 p_j 的解釋義原中時，出現的向上數的層數。許雲等 (2005) 經過考慮實例影響因素，同時考慮詞語義項的相似度和相關度，給出了下面的計算詞彙相關度的計算公式：

$$(4)\quad R(W_1, W_2) = \max\left(\eta_1 \sum_{i=1}^{4} \alpha_i \prod_{j=1}^{i} \mathrm{sim}_j(C_1, C_2) + \eta_2\left(1 - \frac{d(C_1, C_2)}{D}\right) + \eta_3 \mathrm{Examp}(C_1, C_2)\right)$$

式中 $\eta_1+\eta_2+\eta_3=1$。第一部分考慮了兩個概念的相似度，第二部分考慮了兩個概念的相關度，第三部分則是實例影響因素，它是通過《知網》中的概念（詞語義項）描述實例中的信息進一步提高語義相關計算的合理性，計算公式如下：

$$(5)\quad \mathrm{Examp}(C_1, C_2) = \max \mathrm{Sim}(C_{ei}, C_j) \qquad (1 \le i, j \le 2, i \ne j)$$

這裡，C_{ei} 爲第 i 個義項的實例單詞集合中的任一個詞的義項。利用上述兩個公式，所計算的詞語相關度與人的觀察是比較符合的，說明了上述方法的有效性。

基於義原的表徵方法的可計算性，隨著應用的不同其實現計算的方法會有所不同。比如，在面向詞義消歧的研究中，有的研究直接通過計算多義詞所在語句

的義原同現概率，並經過互信息的選擇來確定多義詞的詞義，這種思路比上面的計算相似度的方法要簡單。而基於《知網》的自動問答系統研究除了相似度、關聯度等計算外，可能還會涉及到推理計算，採用什麼樣的方法進行計算，目前這方面的報導還不多見，但可以肯定的是，其計算的方法和途徑一定是很困難的。

2.3 基於概念依存關係的表徵方法及其可計算性

2.3.1 基於概念依存的詞義表徵方法

在人的大腦中所儲存的詞彙知識，就像一本詞典所載有的信息那樣，也規定詞的拼寫形式或發音形式、詞的意義。但與普通詞典不同的是，在人們的腦海裡，各個概念或詞語之間關係以一種不同於詞典的、極其複雜的方式進行表徵的。它不同於普通的詞典，普通詞典是按照字母順序來組織詞語，而人腦則是按詞義信息對詞語進行組織，通過這種組織關係建立起各種概念之間的依存關係，便於機器進行推理計算等。由 Princeton 大學、California 大學 Berkeley 分校和 Microsoft 公司開發的面向機器的語義知識庫 WordNet、FrameNet 和 MindNet，以及由北京大學計算語言學研究所針對漢語的特點，在 WordNet 框架的基礎上開發的中文概念詞典 CCD，都可以看作是一種表示語言中的各個概念之間的依存關係。

WordNet 的基本思想是利用關係表示詞彙語義，它使用同義詞集合代表概念，並且力圖在概念間建立不同的關係指針，表達不同的語義關係。從而將抽象的概念形式化、具體化，並通過詞彙意義對其進行計算和操作，進而可以建立起概念之間的多種語義聯繫和推理，從而爲面向自然語言處理的機器可計算性和自動語義推理奠定了良好的基礎。

2.3.2 概念依存關係表徵方法的可計算性

正像前面所討論的關於 HowNet 可計算性的實現方法一樣，儘管 WordNet 和 CCD 等語義詞典也是用來描述概念及其之間關係的知識資源，但它們又和 HowNet 有所不同（董振東、董強 2001），它們沒有像 HowNet 那樣的知識詞典描述語言，也不能用有限的義原來表示無限的概念。那麼基於 WordNet 的詞語可計算性在面向自然語言處理的應用研究中又如何實現呢？和 HowNet 一樣，我們認爲針對不同的應用目標可能會有不同的可計算性實現方法，不過，從目前現有的資料來看，大多數研究者主要集中在詞語相似度的計算上，由於 WordNet 被組

織成一種樹形圖，樹中的每個節點代表一個概念，兩個節點之間的路徑長度可被用來表示兩個概念的語義距離。有的研究者通過計算在 WordNet 中詞節點之間上下位關係構成的最短路徑來計算詞語之間的相似度；有的則考慮比較複雜的情況，根據兩個詞的公共祖先節點的最大信息量來衡量兩個詞的語義相似度；有的利用 WordNet 計算詞語的語義相似度時，除了節點間的路徑長度外，還考慮概念層次樹的深度或區域密度（朱靖波等 2001），這主要考慮了 WordNet 中概念描述的粗細程度不均勻等因素的影響。

設 s_1, s_2 爲詞語 w_1 和 w_2 在 WordNet 中對應的詞義，考慮 WordNet 中概念結點間的路徑長度以及各概念在樹中的深度，則詞語 w_1 和 w_2 間的語義距離 SD 可按下式計算：

$$(6)\quad SD(w_1, w_2) = \frac{1}{2}\left(\frac{Dis(s_1, ca)}{Dis(s_1, root)} + \frac{Dis(s_2, ca)}{Dis(s_2, root)}\right)$$

式中，ca 表示詞語 w_1 和 w_2 之義項 s_1, s_2 在 WordNet 中的共同祖先概念節點，Dis 函數表示兩個概念在 WordNet 中位置之間的路徑長度。根據詞義之間的距離可計算兩個詞語的相似度如下：

$$(7)\quad sim(w_1, w_2) = e^{-SD(w_1, w_2)}$$

由公式 (7) 可以看出，兩詞語的語義距離愈大，其相似度愈小。當兩個詞語的語義距離爲 0 時，其相似度爲 1，這時兩個詞爲絕對相似。

荀恩東、顏偉 (2006) 等利用 WordNet 的同義詞詞集 (synset)、屬類詞 (class word) 和意義解釋三個集合，從中抽取出候選同義詞的詞彙語義特徵，兩個概念之間的相似度可通過計算其在三個不同意義特徵空間中的距離來得到。所計算的距離越小，相似度越大。

2.4 二維結構表徵方法及其可計算性

一、二維結構詞義表徵方法：二維結構表徵方法是一種採用關係數據庫或其他技術實現的一種數字詞典的表徵方法，目前有代表性的詞典包括北京大學計算語言學研究所的《現代漢語語法信息詞典》、《現代漢語語義詞典》等。北京大學計算語言學研究所開發的漢語語法信息詞典和語義詞典採用成熟的數據庫技術，每一個數據庫文件都刻畫了屬於某一類的具體的詞語與它們的語法信息屬性或語義信息屬性的二維關係。填入的信息儘量以漢字或直觀明瞭的助記符表示。這種

詞典的結構性使得對詞語的語法屬性或語義屬性的提取更加容易，其中的“備註”字段提供了詞語不同義項的用法示例，可以用來解決基於實例學習方法獲取詞義消歧規則時的數據稀疏問題。語法信息詞典中的同形信息、詞類信息以及語義詞典中的語義類型信息都使詞義消歧的算法與模型更易建立，計算複雜度降低，可計算性增強。從另一個方面講，由於採用結構化建立詞典，一些面向結構化數據庫的數據挖掘算法就可以應用於語法信息詞典或語義詞典，從其中獲取相應應用所需要的知識。

二、語法信息詞典或語義詞典的結構性屬性字段可使隱含在文本中的語法信息和語義信息顯性化，通過比較不同詞語之間相同字段的差異，可以獲得相同詞語之間的語法或語義上的距離，儘管目前基於語法信息詞典或漢語語義詞典的詞語相似性或詞語相關性的研究還不多見，但我們計畫開展這方面的研究。本文在針對詞義消歧的研究中應用語義詞典的語義類等屬性字段來計算多義詞在上下文中的合適詞義。

3. 從詞義消歧的角度看詞義表徵方法

詞義消歧是自然語言處理應用研究的一個必要的中間步驟，其目的是利用上下文環境通過一定的計算確定多義詞在當前語境中的詞義。即：

設詞語 W 有 n 個詞義 $S_1, S_2, \ldots, S_n$，在特定的上下文環境 C 中只有 S' 是正確的詞義，每個詞義 S_k 和上下文 C 存在關係 $R(S_k|C)$，詞義消歧就是尋求同 C 關係最強的詞義 S'：

$$(8)\quad s' = \arg \max_{1 \le k \le n} (S_k \mid C)$$

顯然，要實現詞義消歧，就要獲得多義詞的上下文特徵，但基於意義訓釋、語義原語和語義關係的詞義表徵方式的上下文特徵提取計算方法還沒有很好解決，因而從目前的研究結果來看，僅僅依靠上述的詞義表徵方法的詞義消歧實踐結果都不是很理想。

Lesk 提出的利用釋義詞典中詞語義項的釋義與多義詞所在上下文之間的覆蓋度實現消歧的方法，由於詞典中相關義項的釋義內容較少，使得覆蓋度的計算所需要的特徵數據稀疏嚴重，會導致許多詞語的各義項的釋義與上下文的覆蓋度爲 0，從而導致詞義消歧的失敗。另外，由於現有詞典中一個詞的不同義項之間釋義的交疊性，使得人往往都難決定多義詞在上下文中所應選擇哪個義項，因此，

通過計算釋義與上下文的覆蓋度也很確定在當前語境中到底該選擇哪個義項，同樣會導致詞義消歧的失敗。第三，在計算釋義與上下文之間覆蓋度時，同樣涉及到詞語相似性或等價性的計算，而這些計算若光依靠釋義詞典本身，顯然是實現不了的，因此，就詞義消歧來講，應用 Lesk 的方法效果是不會很好的。

利用《知網》的特點和基於詞語相似度的計算方法，余曉峰等 (2004) 試驗了利用實詞概念相似度進行漢語詞義無導消歧的試驗，其基本思想是對多義詞 w 的 n 個義項 $k_1, k_2, \ldots, k_n$，取出其所在語句中其他的 m 個詞，它們分別有 $r_1, r_2, \ldots, r_m$ 個義項 ($r_i \geq 1,\ 1 \leq i \leq m$)，其中實詞 w_i 的 r_i 個義項分別爲 $k_1^i, k_2^i, \ldots k_{r_i}^i$ 。如果記多義詞 w 的 n 個概念與句中其他 m 個實詞 $w_1, w_2, \ldots, w_m$ 的 $r_1+r_2+\ldots+r_m$ 個義項相似度的最大值爲：

$$(9)\quad \mathrm{Max}(w) = \mathop{\mathrm{Max}}_{1\le i\le m}(w, w_i) = \mathop{\mathrm{Max}}_{1\le i\le m}(\mathop{\mathrm{Sim}}_{1\le a\le n, 1\le b\le r_i}(k_a, k_b^i)) = \mathrm{Max}(\mathop{\mathrm{Sim}}_{1\le i\le m, 1\le a\le n, 1\le b\le r_i}(k_a, k_b^i))$$

則取 Max(w) 所對應的詞的某個義項 $k_j (1 \leq j \leq n)$ 作爲詞義消歧的結果輸出，即：

$$(10)\quad k_j = \arg \mathrm{Max}(w)$$

(9) 式中的相似度計算採用劉群、李素建 (2002) 中所提供的方法。該文針對 10 個多義詞進行了詞義消歧試驗，當取多義詞上下文前後各一個實詞進行相似度計算時，詞義消歧正確率爲 37.3%，當取句中除多義詞以外的所有實詞參與相似度計算時，詞義消歧正確率爲 42.06%。楊爾弘等 (2001) 利用《知網》的義原表達式，通過計算多義詞各義項的義原與其上下文中其他詞的義原同現概率，利用互信息的計算結果確定多義詞在當前語句中的義項，這種方法基於《知網》的義原計算比較簡單，取得了 71% 的消歧正確率。

基於 WordNet 的詞義消歧計算的文章不是很多，朱靖波等 (2001) 在詞義消歧研究中，在應用基於統計方法的基礎上，爲彌補訓練數據的不足，通過應用 WordNet 進行詞語相似度計算以後，對詞義消歧模型進行數據平滑，以提高詞義消歧的正確率。

要取得好的詞義消歧效果，目前的大多數研究是集中在統計語言模型或基於實例的消歧方法上，但統計方法的缺點是沒有充分利用語句中的詞義信息以及數據稀疏問題，能否將統計語言模型和基於語義關係的 CCD 或基於義原表徵的 HowNet 等資源相結合，建立基於多種知識資源的詞義消歧與標注模型，以提高漢語詞義消歧的正確率是我們目前正在進行的研究。

4. 結論

通過上面的分析可以知道，基於意義訓釋的詞義表徵方式，由於其詞語意義是以自然語言形式定義的，各詞義定義間又相互獨立，所以計算機很難應用它進行文本內容的計算，更談不上基於它進行語義推理，這種詞義表徵方法的可計算性是比較弱的。

基於義原表徵方式的《知網》，由於其定義了一種詞義知識描述語言，從而以一種形式化的方式對詞義（《知網》上稱爲概念）進行表述，這種形式化的描述式可以看作是語法公式，它對經過精心研究從漢語中所提取出的 1,500 個義原進行運算，使得應用有限的義原經計算後可表示真實世界中的無限概念，把無限的（高維的）概念空間映射到有限的（低維的）義原空間，從而導致這種表徵方法具有了很強的可計算性。另外，《知網》本身不是一種義類詞典，而是一個能描述各種概念及概念屬性之間關係的語義網絡，它的語義描述呈縱橫交錯的網狀結構，從而爲基於這種表徵方式的語義推理提供了可能，這種可推理性爲面向文本內容的理解和機器學習提供了理論上的可行性。

WordNet 和 CCD 都是一種能夠描述概念之間語義關係的語義知識詞典，這種詞典可以把概念與概念之間的各種語義關係以網絡的形式進行表示，通過語義網中各個概念節點間的距離，實現概念距離的計算，進而實現詞語相似度等的計算，並且基於這種語義網絡也可以實現語義推理計算。上下位概念間的關係推理和距離計算在這類語義詞典中的計算是比較容易實現的。但由於這類語義詞典不像《知網》那樣，可以將無限的概念映射到有限的義原空間中，因而對那些非上下位概念間的語義距離計算，可能會導致計算複雜度的增大。

結構性表徵方式由於採用關係數據庫的結構，使得文本中詞語的語法或語義屬性顯性化，也便於採用成熟的數據挖掘技術獲取相關應用研究所需要的語言學知識，因此，採用結構性表徵方式的現代漢語語法信息詞典和漢語語義詞典都是在統計知識不足或數據稀疏時，獲取語法性規則知識和語義性規則知識的良好知識源。

引用文獻

Lesk, Michael. 1986. Automated sense disambiguation using machine-readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 1986 SIGDOC Conference*, 24-26. Toronto, Canada.

朱靖波, 李珩, 張躍, 姚天順. 2001.〈基於對數模型的詞義自動消歧〉,《軟件學報》12.9:1405-1412。

朱嫣嵐, 閔錦, 周雅倩, 黃萱菁, 吳立德. 2006.〈基於 HowNet 的詞彙語義傾向計算〉,《中文信息學報》20.1:14-20。

余曉峰, 劉鵬遠, 趙鐵軍. 2004.〈一種基於《知網》的漢語詞語詞義消歧方法〉, http://mitlab.hit.edu.cn/papers/2004/25.pdf。

李素建. 2002.〈基於語義計算的語句相關度研究〉,《計算機工程與應用》2002.7:75-83。

荀恩東, 顏偉. 2006.〈基於語義網計算英語詞語相似度〉,《情報學報》25.1:43-48。

許雲, 樊孝忠, 張鋒. 2005.〈基於《知網》的語義相關度計算〉,《北京理工大學學報》25.5:412-414。

楊爾弘, 張國清, 張永奎. 2001.〈基於義原同現頻率的漢語詞義排歧方法〉,《計算機研究與發展》38.1:833-837。

董振東, 董強. 2001.〈《知網》和漢語研究〉,《當代語言學》3.1:33-44。

趙軍, 金千里, 徐波. 2005.〈面向文本檢索的語義計算〉,《計算機學報》28.12:2068-2078。

趙應鐸. 1994.〈從詞與詞的組合上劃分多義詞的義項〉,《江淮論壇》1994.6:83-86。

劉揚, 俞士汶, 于江生. 2005.〈CCD 語義知識庫的構造研究〉,《小型微型計算機系統》26.8:1411-1415。

劉群, 李素建. 2002.〈基於《知網》的詞彙語義相似度的計算〉,第三屆漢語詞彙語義學研討會論文。台北：中央研究院。

[Received 23 December 2006; revised 29 August 2007; accepted 10 October 2007]

張仰森
北京信息科技大學計算機及自動化系
中國 100085 北京市海淀區清河小營東路 12 號
zhangyangsen@163.com

Study on Word Sense Expressing Methods and Their Calculability

Yangsen Zhang[1, 2], Yunfang Wu[2], and Shiwen Yu[2]
The Beijing Information Science and Technology University[1]
Peking University[2]

The merits and shortcoming of some Chinese word sense expressing methods such as word sense allocution, semantic primitive, and concept dependence are analyzed in this article, and the basis to select dictionary knowledge sources for computer automatic word sense disambiguation is offered.

Key words: word sense expressing method, calculability, automatic word sense disambiguation