Computational Linguistics and Chinese Language Processing
Vol. 7, No. 2, August 2002, pp.1-36

情境——組織/存放辭彙語義知識的恰當框架

Situation – A Suitable Framework for Organizing and Positioning Lexical Semantic Knowledge

陳祖舜*、周強*、趙強*

Zusun Chen，Qiang Zhou，Qiang Zhao

摘要

作爲符號系統的自然語言其最大特點，也是優點是它有一個組織與存放概念知識的邏輯框架，就是它的辭彙體系。自然語言靠它實現了凝聚、吸收、組織、存放概念知識的功能，從而使語言內部逐漸形成一個極其龐大複雜的概念知識體系。語言的這項功能是它另兩項功能，即實現交際的媒介與體現思維的實體，的基礎。自然語言的語義學必須再現語言的這三項基本功能及它們之間的關係，因此自然語言的語義學必須以它的辭彙語義學爲核心與基礎。在這裏，語義詞典成了核心中的核心。

詞語是概念的符號體現。概念產生於特定的認知圖式。概念，或標識它的詞語的義項，只有在產生它的特定圖式中才能描述、定義清楚。概念的使用則是在使用環境中對照、還原、引用產生它的圖式的過程。我們用情境做認知圖式的數學模型，把情境理論當作上述的辭彙語義學和建基於其上的語義學的統一理論的框架，於是得出情境理論的一系列新課題。本文討論其中的初步問題：用情境表示圖式，用概念定義情境，和用情境定義/描述概念，後者是重點；建立情境代數以刻畫情境間的關係、變換與運算，實現用代數演算體現概念思維，建立情境網以實現圖式的結構、概念的組織方式。側重點落在語義詞典的構成與組織方面。本文主要是使用實例說明做法，後面的文章將討論相關的數

* 智慧技術與系統國家重點實驗室　清華大學電腦科學與技術系，北京 100084
State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science and Technology, Beijing 100084
E-mail: (czs, zhouq, zq)@s1000e.cs.tsinghua.edu.cn

學理論。隨後還有一系列文章討論建設上述的辭彙語義學和其上的語義學的各方面問題，以及相伴的，作爲理論依據的情境理論的建設。

關鍵字：概念、辭彙義、情境、情境代數、情境網、語義詞典、辭彙語義學

Abstract

The characteristic and an advantage of natural language is that, as a symbolic system, it has an internal logical framework for organizing and positioning conceptual knowledge, which is its lexicon system. This framework implements the fundamental function of natural language to condense, absorb, organize and position conceptual knowledge, and creates progressively a very large and complex build-in knowledge system in the language. It is also the basis of the other two fundamental functions of natural language; i.e., it serves as a tool for communication and as a medium for conceptual thought. The natural language semantics should reproduce the basic framework of natural language in their theoretic realms to represent these three functions and their relationships. The lexical semantics thereby become their core.

A word is the symbolic embodiment of a concept, and a concept is generated in a peculiar cognition scheme, which will be called its generating scheme. We cannot describe and define a concept clearly unless we put it into its generating scheme. Meanwhile, the implementation of the concept involves a procedure that contrasts, restores, and refers to its generating scheme in a special environment, which will be called its application scheme.

We propose to use the situation as a mathematical model to describe a cognition scheme. Therefore, the situation theory can serve as a unified theoretical framework for constructing the lexical semantics and the natural language semantics built upon it, as mentioned above. Therefore, many new viewpoints are proposed. In this paper, only some elementary questions about them are discussed, including: 1) using a situation to express a scheme and using a situation to describe a concept (this is the key point of the paper); 2) formulating the situation algebra for describing relations, transformations, and operations for situations so as to simulate conceptual thinking by means of algebraic calculus; 3) constructing a situation network to implement a scheme structure and conceptual structure, where the key point is the constitution and organization of a semantic dictionary. We use some practical cases to illustrate these methods. The mathematical theory relevant to them will be presented in our future papers.

Keywords: Concept, lexical meaning, situation, situation algebra, semantic dictionary, lexical semantics

1. 引言

約在 20 年前，邏輯學家 J.Barwise 與心理學家 J. Perry 創建了情境語義學[Barwise,J.,*et al.* 1983(1999)], 當即引起極大關注，Stanford 大學專門成立了語言與資訊研究中心(CSLI,Center for Study of Language and Information)從事探索，開展了情境理論與情境語義學（STASS group,Situation Theory And Situation Semantcs）建設[見[Barwise,J.1987]及CSLI 出版物]。對此它的創始人之一的 Barwise 在 1986 年 [Barwise,J., 1989] 曾指出，當前資訊時代急需資訊與資訊加工的理論。這種資訊理論的核心就是資訊的語義學。這是一種全新的，內涵式的語義學理論，是一種面向內容的資訊與交際理論[Barwise,J.1987]. 他希望他們的新理論[Barwise,J.,*et al.* 1983[1999]]就是這種核心理論。約在 10 年以前，邏輯學家、情境語義學另一位創始人 K. Devlin 也指出，沒有關於資訊與交際的數學理論，人工智慧理論就不成其爲理論[Devlin, K.,1991]。然而經過近廿年的轟動與輝煌，情境語義學漸漸失去了動力，而人工智慧理論竟也仍然是幾乎空白[比如見[趙海 等人 2002][1]。] 可以說，現如今這種理論貧乏的窘況與 20 年前的狀況竟毫無二致，人們對相關的理論的渴求依然如故[請見[鍾義信 1998][魏宏森 1998][趙海 等人 2002].]究其原因，我們認爲主要是情境理論在基礎設定上，特別是其哲學-心理學的基礎，存有問題，致使它未能找准前進的方向。

其實新理論的基礎問題從一開始就備受關注。Barwise 與 Perry 在 [Barwise,J.1987]《Linguistics and Philosophy》[vol.8, 1985][Barwise,J. *et al.* 1985] 等，與重版的 [Barwise,J.,*et al.* 1983[1999]] 中都曾經多次評述、研討他們的新理論的哲學-心理學基礎與數學基礎，並斷言他們選定的哲學-心理學基礎是沒有問題的。只是其數學工具（所謂的 KPU 集合論）選得不好，後來又提議改用非良構的集合論(non-well-founded set theory)爲基礎。Devlin[1991]也強調了哲學-心理學基礎與數學基礎在情境理論中的根本性。哲學、心理學、邏輯學、自然語言語義學等學科的學者曾圍繞情境語義學的基礎問題進行過多次熱烈的討論。Barwise,Perry,Devlin 等幾位創始人，以及參與爭論的持對立意見的許多著名學者都對情境理論的基礎做過深入的研究，圍繞著語言、思維、認知、交際等的本質、實質進行過針鋒相對的論戰，彼此間有過尖銳的批判[比如見《Linguistics and Philosophy》[vol.8, 1985].]儘管爭論中不乏真知灼見和深刻的、細緻的分析，但卻未能產生令人信服的理論[2]。其原因是各派理論在基礎上也都有自身無法克服的根本性的困難。

---

[1] 有趣的是，儘管在理論目標上[趙海 等人 2002]與我們的不盡相同，卻與我們一樣認爲人工智慧理論貧乏，急需發展；也一樣的把自然語言理解的理論問題當作是人工智慧理論的最困難的、最高級的問題。

[2] 此後學界仍不停地在探索、論述。以至在 1999 年[Barwise,J.,*et al.* 1983(1999)]重版之時作者仍要重提、重載當年他們的論爭文章。

若遵照它們的哲學-心理學理論來建設語言與思維的理論、資訊與資訊加工的理論等，一定也會像情境語義學理論一樣地迷失方向、陷入困境的。創建新語義學最重要也最困難的是確立它的哲學與心理學基礎，其次就是數學基礎。其實創建任何新理論也都如此。

我們曾在[陳祖舜, 1995]中分析過，問題出在情境語義學認定資訊存在於外部世界，個體或物種資訊接受者獲取資訊的能力得自於自身適應環境的進化結果。這就忽略了資訊接受者主體方面的複雜的而又是主導的關鍵性的因素。主體僅僅被當作被動的接受者，其知識構成、價值取向等因素都被忽略了。主體的能動的反映行爲、主動的實踐行爲被降格爲不過是從外部接受資訊做出回應的被動的適應環境的行爲。於是該理論把重點放在對外部世界的情境的描述與分類上。在這樣的框架裏，沒有了生成資訊的主體-客體相互作用:客體主體化與主體客體化兩個生動過程[3]。客體-主體間行進的資訊流顯得十分貧乏而且問題多多：思維有時可以沒有語言[見[Barwise,J.,*et al.* 1983(1999)]]，語言在思維中的至關重要的作用被否定了；語言的組織與存放概念知識的重要功能不見了，難怪該理論不去考察辭彙體系與它背後的概念結構；等等。

[陳祖舜，1995]中曾借公式：資訊=信號+解釋　表述我們的觀點。詳細講就是信號存在於客體世界，經過主體的能動的加工而得到資訊，並存在主體的認知器官（大腦）之中。在這裏，接受者的解釋機制是重要的主導因素。根據資訊的性質，也即解釋機制的性質，應當建立不同的語義學理論[4]。具體到自然語言符號系統，其解釋機制就是主體的概念知識與運作概念知識（即概念思維）的體系，以及概念思維的表達系統。我們要做的就是建立關於概念知識組織、運作（概念思維）與言語交際的語義學理論。語言作爲認識人的概念思維規律的視窗，是主要的考查物件。語言有三大功能[5]：即，交際的媒介，思維[6]的介質，和凝聚、組織、存儲與（支援）使用概念知識的框架。其中後一項功能是前兩項功能的前提與基礎。語言主要是用它的辭彙體系實現該項功能的。人腦中的許多概念組成一定結構，叫概念結構。辭彙體系是它的外顯形式[黃昌甯 等人, 1988]。過去我們說過[黃昌甯 等人，1988]，概念結構基本上是個網，概念的意思就存在於、體現在該網中它與其他概念的聯繫、關係之中。

任何概念都必須用詞語稱謂它。於是上述的概念結構就誘導出辭彙的一種結構（叫辭彙體系）。可見，辭彙體系以概念結構爲內核，而成爲它（概念結構）的外顯。人們

---

[3] 借用 Piaget 的術語、說法[見[雷永生 等人 1987]]。

[4] 據我們看法，一個完整的資訊語義學理論至少應包含三部分：以人或機器實現的概念思維與話語交際（中的資訊活動）爲物件的認知語義學研究；以人、動物或機器的行爲模式爲物件的行爲模式語義學研究；和以細胞社會爲物件的胞內與胞間生物大分子資訊傳遞爲內容的細胞社會語言語義學研究。混淆界限蠻幹必然得不償失。這種教訓在語義學中可謂多矣（不乏先例）。

[5] 據我們所知，現行的語言學理論似乎都只承認語言有兩項基本功能，即交際的工具與體現思維的符號表示系統。未見有把凝聚、組織、存放概念知識做爲其基本功能之一的。我們認爲由辭彙體系實現的這項基本功能乃是前兩項功能的基礎，因而更基本，是語言的核心功能。

[6] 本文中思維專指概念思維，也即抽象思維；知識專指概念知識。下同。

直接使用的只能是符號（詞語）。正因爲如此，才使辭彙體系成爲凝聚、組織、存儲與使用概念知識的唯一邏輯框架。

基於以上認識，我們認爲辭彙語義學研究應當以概念義爲本位[陳祖舜, 1995]。首先研究人腦中的概念結構，然後再去考查它的外顯----辭彙體系。因爲用詞語稱謂概念並不是簡單地貼個標簽。辭彙一旦形成體系就獲得了相對的獨立性，有其自身的發生、發展、與消亡的規律。諸如，造字、構詞、韻律、文野之分等方面都有其自身歷史演變的制約，不具有任意性且也相當複雜。何況哪些概念詞語化，哪些沒有，其間難有什麼規律可循。這樣，首先考察概念，可以在純態中把握它們，暫時撇開各種複雜因素來探討概念結構的相關的理論。而且概念義也確實是詞語\詞義的核心內容。有了概念的定義再用來描述相應的詞語，就可以把有關詞語的零星知識附加在其概念義上面。這種做法確有許多好處。

人腦中的概念結構常常也稱作（抽象的）概念詞典或語義詞典。儘管我們尚不知道人腦中的概念結構是否確以詞語與詞語間的聯繫來實現的[7]。[陳祖舜, 1995]曾設想這種概念詞典由底層的日常用語詞典爲基礎，與建於其上的專業詞典組成。專業詞典的主要成分，即它的衆多詞條，是用基礎詞典的詞條（概念）陳述的一個個知識包，再加上這些專業詞條間的聯繫網。[陳祖舜, 1995]討論了它們的構成。它們的構成與運行方式是辭彙語義學研究的中心物件。基礎詞典加上其上的常識知識庫，專業詞典再加上其上的專業知識庫一起形成了我們的知識總匯，是人腦解釋信號的機制。它們是語義學研究的物件。一個良好的關於概念思維與言語交際的語義學必須至少能在理論上再現上述的語言的三大功能。考慮到關於語言使用的系統知識也是概念知識，而使用環境、情境因素必須間接地通過使用者對其認識才能起作用[陳祖舜, 1995].，這就把辭彙體系推到了最基礎也是最重要的境地。特別是日常用語的辭彙體系[8]，佔據最核心最根本的地位。因此，合邏輯的結論就必然是：這種語義學理論一定是以（日常用語的）辭彙語義學爲核心與基礎的。於是可以把我們心目中的語義學簡括成：這將是一種統一的，把辭彙、短語、語句的意思與語境以及通常歸於語用範疇的許多因素結合在一起來考慮的，全新的，內涵式的語義學理論[陳祖舜, 1995]。這就是我們需要的。

［陳祖舜，1995]考察了這種設想的理論的各個基本面，特別是作爲其核心與基礎的語義詞典的構成。它更像是一篇宣言書，只是一個高度概括的綱要，需要對其各個方面的各個層面做進一步研究與闡發。本文及隨後的一系列文章意在逐步展開它的各個方面。首先要探究的當然是作爲主體的解釋機制的核心要素——語義詞典，考查它的構成與運行機制,建立它們的理論。本文作爲其中的第一篇，著重探討概念與概念結構，研究概念和概念間的聯繫的描述或叫定義方法，爲探求相應的數學性質奠定基礎。

---

[7] 神經語言學研究似乎支援這種看法。

[8] 下面我們就用辭彙體系稱呼日常用語的辭彙體系，用辭彙語義學稱呼日常用語的辭彙語義學，等等。

本文的意圖是要給出概念的真正的描述性定義。回顧地看，我們爲此做了一系列假設：首先假定辭彙體系是人腦中的概念結構的外顯形式，概念結構是辭彙體系的內核[[黃昌甯 等人 1988] [陳祖舜, 1995]]。接著我們假定概念的意思、內涵就存在於、就體現在它與其他衆多概念的聯繫的總和之中 [[黃昌甯 等人 1988] [陳祖舜, 1995]]。在此，需要區分本質聯繫與附帶聯繫。本文認爲概念首先是在產生它的那個情境（叫做它的定義情境，是產生該概念的認知圖式[9]的數學模型與有限近似）中與相關的其他概念建立起本質聯繫的；情境與情境之間還有錯綜複雜的聯繫，它們誘導出概念之間的附加聯繫。本文給出了用情境(包括情境運算式)定義概念、概念的性質，以及概念間的關係等的方法與工具，略微討論了情境間的聯繫，等。限於篇幅，還有許多直接相關的內容，特別是相關的數學理論，只能放到續篇裏了。

情境語義學[[Barwise,J.,*et al.* 1983(1999)][Barwise,J., 1989][Devlin, K.,1991]等等]提出的問題以及許多見解是很有見地的；在語義學中情境及其數學描述等是情境語義學的首創 [[Barwise,J.,*et al.* 1983(1999)][Barwise,J., 1989][Devlin, K.,1991][Barwise,J.1987]等]. 認知圖式與概念生成機制則是哲學-心理學中熟知的結論[見[雷永生 等人 1987]等]。在此,本文主要貢獻僅在：提出了把情境當作認知圖式的數學模型和在概念的定義情境中定義、描述概念與概念間的關係的做法[10],等。並爲此提煉了一套描述工具(它們是在情境理論的基礎上做成的)。隨後的文章會證明它們有堅實的數學理論的支援。情境語義學建立起來的成套的情境理論，特別是它的數學理論，無疑是極其重要的、寶貴的理論工具[[Barwise,J.,*et al.* 1983(1999)][Devlin, K.,1991][Barwise,J.1987]等等。]我們的工作順帶也驗證了這一點。

當然，這些都是爲了構作一個自主式的語義詞典，使辭彙體系成爲組織概念知識的框架，從而使語言能成爲支援、體現思維與實現交際的工具。這種將語言的凝聚、組織概念知識的功能作爲語言的主要功能的基礎，認定語義學要以辭彙語義學爲基礎與核心等的主張,也許是我們獨特的,但未必正確的見解。我們以此爲線索來看相關的研究。

向來的語義學都不曾把能動的語義詞典當作、取做運作概念思維，包括自然語言理解，的核心機制。傳統的語義學，比如邏輯語義學[參見其集大成者 Montague 語義學比如[Thomason,H.1974(1979)]]就不研究也研究不了辭彙體系的構成。除了描述零星的詞語的語義外它不處理大面積的辭彙，也處理不了。即使對個別詞語的意思的描述也存在不少難以解決的致命的困難。新誕生的語義學理論，比如[Jackendoff, R.1990]的概念語義學，它只承認數量極有限的幾個語義函數，而且當作是先驗的，顯然無法描述辭彙體系所承載的極其豐富的內容。它對詞語的意思的描述只能倒退到義素分解與標注法。根本無法產生一個能動的語義詞典作它的基礎與核心。等等。這樣的語義學自然不是也不能

---

[9] 借用 Piaget 的發展認識論中的術語。該理論和能動的反映論的認識論一致，是我們所主張的哲學-心理學基礎。

[10] 與此相配的是，在概念的運用情境中展開它（該概念）的定義情境。（後續的文章中將要討論。）

置於辭彙語義學基礎之上的。等等。總之，現有的語義學理論的架構都無法在理論的層面上自然地再現語言的三大功能，或它們認可的兩大功能。自然語言符號系統的最重要的特點也是優點，是符號系統本身凝聚著龐大而複雜的已有的常識知識體系。它顯然不是形式化的符號系統。也不能任意地簡化。現有的語義學理論似乎沒有尊重這一點，所作的理論建設根本不是本著這點進行的。

近年來，由於語言工程需求的驅動，出現了許多描述辭彙體系的方法與理論，也建立了一些系統。比如：(1)義素分解，[Jackendoff, R.1990] ；(2) 義場分解，[賈彥德 1999][張普 1995] ；(3) 語義分類樹，[陳群秀 等人 1995][張普 1991.3][EDR 1993] ；(4) 同義詞集標示，[Miller,G. A. *et al.* 1993]中的名詞的描述；(5)格框架，[張普 1991.3][魯川 1995]及其近期發展：框架網[Baker, C.F. *et al.* 1998]，以及[董振東 1997]的動詞部分；(6)詞語搭配/配價關係，[陳群秀 等人 1995][袁毓林 1998][濬陽 等人 1995][林杏光 等人 1997] ；(7)原語集標注[Chengming Guo (郭承銘) 1995] ；(8)多個、多層次網多重標注，[黃曾陽 1998] ；(9) 特徵義（核心語義特徵）標注，[董振東 1997] ；(10) 關係網，[黃昌甯 等人, 1988][Richardson, S.D. *et al.* 1998] ；(11) 專家系統，[Lenat, D.B. *et al.* 1989-1990].等等，就是從不同角度對詞義的描述理論、方法與系統進行的艱苦探索。

這些方法有些是陳舊的，早已證明是無效的，如(1)(2)；有些只是爲特定的目的而設的，如(3)(4)(5)(6)；有些只針對一方面詞語或一方面屬性，如(4)(5)(6)；真正想做成通用的語義詞典的，試圖成爲組織概念知識的框架，涵蓋常識知識，使能成爲自然語言理解的核心的(7)(8)(9)(10)(11)，看來又都缺乏系統的語義學理論的支援，要想進而成爲能動的解釋機制恐怕一時還不行。也看不出由它們如何直接建成它們的辭彙語義學和建基於其上的它們的語義學。至少還嫌太弱。此外，上述種種方法中多數給出的僅僅是標記法，並非語義描述。少數描述方法給出的也僅僅是區分性描述，而非內容描述。而給出內容描述的方法卻又難以數學化。等等。看來不可能成長成我們所期盼的語義詞典。必須“另起爐竈”這就是本文意圖所在。所幸情境語義學、反映論的認識論已爲我們準備好了非常合用的理論工具。

本文組織如下：§2 基於情境的概念描述方法。論述定義和描述概念的一種新方法，考察情境內部的構成。§3 情境代數與情境網。討論情境間的關係、變換與演算，和體現這些聯繫的組織方式----情境網。§4 關於情境描述。§5 結束語，概述了本文的要點，並提出今後的工作設想。最後是鳴謝與文獻

2. 基於情境的概念描述方法

2.1 情境與情境描述

哲學、心理學、語言學、乃至一般認識論科學中，所謂情境是指主體從認知的目的所把握的客體的那個部分。客體並不直接就是情境，只有當它成爲認知物件並爲主體所把握

時才是情境。此時主體用自己的概念工具在腦中描述、再現了這個情境。我們稱之爲抽象情境。抽象情境是，也只能是所描述的情境的有限近似：在空間與時間上，深度與廣度上和正確程度上的有限近似，甚至可能包含有錯。爲了區分，我們把客體世界的情境暫叫做客體情境。易見，對應同一個客體情境，可有多個抽象情境。這些同源的抽象情境之間有一些有趣的關係，以後的文章將論及它們。這裏暫且不提。爲了便於使用，在此做一些簡單推廣，即把在概念世界中產生的對真實客體情境的有所偏離或歪曲的描述，甚或虛構的，並不直接對應某客體情境的相關描述，也都叫做抽象情境。爲了區分，我們常把與客體情境相符的抽象情境標上“真實的”，而把其他的標上“虛擬的”記號、標簽。由於我們下面主要的是與抽象情境打交道，就把抽象情境簡記做情境。當用到真實世界中的客體情境時，除非上下文明白，都一律用客體情境來稱呼。

遵照[Barwise,J., *et al.* 1983(1999)]，情境描述的基本單元叫資訊元，它由四類基本量：時空場合、關係、個體、與定值元構成。基本形式爲《r,l:Loc,i_1,i_2,…,i_n;p》.下面稍作解釋：

作爲描述物件的情境，其內容大體包括：物件，它的存在，具有的性質，它與其他物件之間的聯繫；事件及其發生、存在（演變、發展）與消亡，以及人們對這些事物的認識：肯定與否定的判斷，好與惡的評價，贊同與反對的表態之類，等等。概括地講，這些內容都或多或少地涉及上述的四個基本量。（可能還會有別的基本量。）其中

時空量作爲事物存在形式的抽象，也是一種個體[11]。因爲它突出的重要作用而單列出來。爲了區分，在句法上用記號 l：Loc 表示。有時只用到時間區段（簡作時段）、空間處所（簡作區域），約定分別用 t,t_1,t_2,…； s,s_1,s_2,… 表示它們。它們是時空場合的組成成分與特例。值得一提的是，我們以時段和區域爲時間與空間的基本量。時間點、空間點作爲導出量，是數學的近似[12]。有時 l 可分解成時間與空間的卡氏積：l=t×s，我們就把時空場合 l 表示成 t×s：Tempo×Spat，或寫成 t：Tempo，s：Spat。這裏 Tempo 與 Spat 分別表示時間型式與空間型式。有時只涉及時間或空間，就只寫出時間項或空間項，還有一些關係與時空無涉，或可忽略其存在時空，就完全不寫。

關係作爲資訊元的主要構件，可以用來表示事物間靜態的聯繫和動態的作用。關係有其存在的時間與空間。在這裏，概念（包括關係）不再用它的外延集解釋。關係像是個函數運算元，當它作用在不同的個體列上時得到的是它的不同的“實現”。而它的外延則理解成是另一個函數，其值（外延集）依所在情境而定。

在語義學領域，個體這個概念是情境語義學提出來的[見[Barwise,J.,*et al.* 1983(1999)][Devlin, K.,1991]]。個體不同於通常的原子概念。語義學中原子通常含有本原

[11] [Barwise,J.,*et al.* 1983(1999)]中把時空場合作個體對待。[Cooper, R.1986]還論證了這樣做的合理性。但其後在[Devlin, K.,1991]等一系列論文中都又改回用時間區段與空間處所的卡氏積了。我們覺得時空場合當作個體常更好用，故仍用[Barwise,J.,*et al.* 1983(1999)]的約定。

[12] 我們將在另一篇文章中論述與此處觀點相配的時間空間結構，及其上定義的函數與關係等。

的、不可再分的、無內部結構的之類含義。情境語義學中個體是指在思維中可以當作一個整體的物件，能在思維中把它從其存在環境（情境）中割裂、剝離出來的，相對獨立、相對穩定的，能在思維中保持其質的規定性的物件。因此個體就其本身而言可以是有結構的，由其他一些個體組成，可能有很複雜的結構。照此理解，任何概念和從概念定義出的（幾乎任何）物件都能當作個體。是否當作個體不取決於物件自身的特性，而取決於思維如何把握、對待它。只要也只有當它被當作相對完整的整體時才是個體。資訊元中個體（列）協助關係形成資訊的內核（叫陳述相）。

定值元相當於通常的真值[13]。作爲資訊元的構成要素，它幫助最後形成資訊。改變定值元（其他要素不變）將得到一族同源陳述。根據 p 的取值不同，資訊元有不同含義：

當 p=1 時表示在時空場合 l 上個體 $i_1,\cdots,i_n$ 之間關係 r 成立。

當 p=0 時其含義是在 l 上不存在上述關係。[14]

當 p=⊥ 時含義是不能確定上述關係是否成立。所獲得的資訊少於上述兩種情況。

當 p=⊤ 時含義是該資訊元含矛盾，即自身是冗餘資訊。

情境描述是對情境的一個有限的近似的描述。我們稱 $\delta=\{\sigma_1, \sigma_2,\ldots,\sigma_n\}$爲一個最簡單的情境描述，這裏諸 σ_i 是若干個如上面定義的基本資訊元。

該描述實際上枚舉了所描述的情境的若干屬性，相當於它們的並。複雜點的情境（和情境運算式）的描述可能要用到更複雜的資訊元或諸資訊元之間不是並的運算等。我們後面文章會有所論述。目前只考慮最基本的情況。[15]

幾乎任何事物，其存在的時空都是相對的，有限的。在描述情境時我們常常把它的存在時空突出出來。寫成$\{l: \sigma_1,\sigma_2,\ldots,\sigma_n\}$，若 $l=t\times s$，就寫成$\{t\times s: \sigma_1,\sigma_2,\ldots,\sigma_n\}$.如果只有時間或空間項，則寫成$\{t: \sigma_1,\sigma_2,\ldots,\sigma_n\}$或$\{s: \sigma_1,\sigma_2,\ldots,\sigma_n\}$.

下面通過一個例子來說明情境描述的具體方法。該例子是最簡交易情境的一個描述：x，y 兩個人在時空場合 l 進行了一次交易，x 付給 y 貨幣 m，y 交付 x 貨物 g. 其中 x 交款場合是 l_1，y 付貨場合是 l_2。

---

[13] 情境語義學中稱作極性元，只取 0，1 兩個值。我們將它們稍作擴張，改成四元格，叫定值元域。以後可能還會用到其他形式的結構作定值元的域。

[14] 應該說這個表述不夠清晰。光表示了“在 l 上 $i_1,\cdots,i_n$ 之間不存在關係 r”，並未講明在別的時空場合上怎樣。特別是未說明是否排除了在 l 的“子段”上以及在含 l 爲“子段”的其他場合上可能存在該關係的情況。但若把時空場合項豐富一下，比如引進 in l,on l,upon l,at l 等運算元，可使表示更細緻一些。

[15]我們用符號 $\sigma\models\delta$ 表示描述 δ(陳述的內容)在情境 σ 中成立。情境語義學中符號 $\sigma\models\delta$ 叫做命題([Barwise,J.,*et al.* 1983(1999)][Barwise,J. 1987]). ⊨是情境與資訊元之間的一個關係。對最簡描述 $\delta=\{\sigma_1, \sigma_2,\ldots,\sigma_n\}$而言, $\sigma\models\delta$ 等價於 $\forall i.\ \sigma\models\sigma_i$

c_最簡交易（l,x,y,m,g;l_1,l_2） ⇦

{Arg: l:[c_時空場合],x,y:[c_人],m:[c_貨幣],g:[c_貨物]

Internal Arg: l_1,l_2: [c_時空場合]

Kernel:

l:

《r_他動-遷移,l_1:Loc,x,擁有權#m,x,y;1》

《r_他動-遷移,l_2:Loc,y,擁有權#g,y,x;1》

《l_contain,l,l_1;1》

《l_contain,l,l_2;1》

《coend,Proj(l,Tmpo), Proj(l_1,Tmpo);1》

《coend,Proj(l,Tmpo), Proj(l_2,Tmpo);1》

End_Kernel

Pre-conditions:

《r_擁有,Pre(l_1):Loc,x,擁有權#m;1》

《r_擁有,Pre(l_2):Loc,y,擁有權#g;1》

《r_需要,Pre(l_2):Loc,x,擁有權#g;1》

《r_需要,Pre(l_2):Loc,y,擁有權#m;1》

End_Pre-conditions

Post-conditions:

《r_擁有,Post(l):Loc,x, 擁有權#g;1》

《r_擁有,Post(l):Loc,y, 擁有權#m;1》

End_Post-conditions

Rationality:

《r_寧願,Pre(post(l)):Loc,x,擁有權#g, 擁有權#m;1》

《r_寧願,Pre(post(l)):Loc,y,擁有權#m, 擁有權#g;1》

End_Rationality

Superiors:

c_合作行爲: …

c_合同行爲: …

…

End_Superior

/*此處填寫從上位繼承來的各種量和對其新加的約束等，以及其他內容。暫略。*/

End_Internal_Arg

End_Arg

}

在上述描述中，我們把“x 交給 y 貨幣 m，並從 y 處取得貨物 g”這個簡單的交易活動看成兩個遷移，即“x 在時空場合 l_1 把對 m 的擁有權轉讓給了 y”，和“y 在時空場合 l_2 把對 g 的擁有權轉讓給了 x”。x，y 分別是這兩個遷移的施動者。並且 x，y 能進行轉讓的前提條件是有擁有權。這是社會行爲規範要求的。x，y 能分別完成各自的轉讓行爲，從而從整體上完成這次交易則是合作行爲所保證的。轉換的結果是互換了擁有權。之所以要進行這種轉換則是因爲雙方都是更願意擁有對方的東西。這些內容連同它們的推論都可用公理的形式附在 r_擁有，c_擁有權與 r_寧願等概念物件處。等。

描述中我們主要引用了兩個抽象概念：“擁有權”與“它動-遷移”。前者是個部分函數[16]，記成“擁有權#”。這裏‘#’是函數記號，叫做#型函數，“擁有權”是它的標識。運算式“擁有權#g”代表“對 g 的擁有權”。後者是個五目關係，關係式《 r_ 他 動 - 遷移,l:Loc,z, m,x,y; 1》表示在時空場合 l,z 將某物件 m 從 x 處移到了 y 處。這裏涉及到了人與物之間的一種“擁有-歸屬”關係。這是一種原關係。即不能再用其他概念來定義的。

此外，概念“r_寧願”與“r_需要”是屬於心理世界的，是兩種心智狀態，都可歸入價值論範疇。其中“r_需要”可取做原語。而關於需要的知識可以任意選用一種需求理論模式，比如可用 Engels 的三層結構模式，也可用 Maslow 的五層結構模式，或其他別的模式，用公理的形式陳述出來，表述成情境間的約束等。“r_寧願”是一種價值判斷，涉及到價值比較與取向，它像本例一樣要用一個情境來刻畫。同樣，也可放進許多知識在裏面。限於篇幅，一概從略。總之可以看出，情境可以用來放置許多相關的知識，可以揭示概念的本質屬性，提供該概念與其他概念的本質聯繫，等等。

描述式中的“l_contains”是時空場合論域中的關係，“coend”是時間論域中的關係，意思自明。“post”與“pre”是時間論域上的函數(不定函數).Post(t)與 pre(t)分別表

[16] 我們用函數形式來陳述物件的附加屬性而不把它們直接附在物件本身上面。這種技術處理的好處是，可使物件自身完整，而其附加屬性可隨時增減。我們共引用兩種屬性附加函數：dot 型函數與#型函數。前者用來附加物件自身固有的屬性，後者用來附加由其他關係誘導出的屬性。

示與某個時間段 t 銜接且在其後，和在其前的那種時間段。這裏的不定函數理解成：存在一個這樣的時間段。Proj(l,Temp)是時空場合 l 在時間域上的投影。類似的，Proj(l,spat)是 l 在空間域上的投影。我們在時空場合上已定義有一些關係與運算。能夠表示需要的各種時空場合。有關內容將另文討論。（見[Chen Zushun *et al.*unpublished]）

另外，我們承認上下位概念之間的繼承機制。比如關於“合作行爲”、“合同行爲”和“經濟行爲”這幾個更大的概念也需要在別處描述。它們都從“社會行爲”處繼承一些性質，後者又從“行爲”處繼承性質，等等。這樣既便於知識的組織，又便於引用。利用上下位的繼承機制，可使圖式體系變得很緊湊。（上位資訊由 Superior：引導）。

2.2 用情境定義概念

概念產生於一定的情境中。新概念用來吸收、凝聚對這個新情境的新認識，以形成一個相對穩定的、能獨立引用的個體作爲進一步認識的立足點。此時常需用一個詞語（或片語）稱謂它，以便日後在思維與交際中引用。該概念的最基本、最主要的性質就是在此定義情境中給出的。概念（除了原語）只有放到產生它的那個情境中去才能解釋清楚。易見，這種本質性質，就體現在由其定義情境建立起來的它與其他概念之間的聯繫上。

就以簡單交易情境爲例。首先買者、賣者、商品、貨款、價格、交易場所、交易時間等諸多概念，以及買者賣者關係、貨物價格關係等，只有用交易事件才能說清楚。再如付款、交貨這兩個事件，還有購買、銷售這兩個行爲，若不從交易事件來說好像也很難說清楚。至於交易合同生效時間，買者賣者與貨物和貨幣的關係等等，離開交易又從何談起？ 其次，情境頭“c_簡單交易”這個概念當然是要用交易這個情境來定義的。情境頭“c_簡單交易”看作函數可以用來形成一個簡單交易情境實例物件。該概念還可用作資訊元中的關係（記作 r_簡單交易）以陳述關於某個交易的資訊。此外，有了交易情境，也就有了一個模式。由此還能得到租賃、借貸、賒購、預訂等相關情境及它們定義的概念。而且還能建立起它們和它們定義出的概念與交易情境及交易情境定義出的概念之間的聯繫。順便指出，運用下面陳述的情境運算，這些情境還可以很簡單地從交易情境“計算”出來。

由此可見，情境是刻畫概念本質屬性、建立概念間本質聯繫、彙聚相關的概念知識的最自然的框架。僅從上述的例子就可以至少概括出用情境定義概念的下述六種情況：① 抽象出存在於該情境裏的，在該情境中擔任特定角色的量，我們稱之爲角色，是一種條件參量[17]，如買者、賣者等;② 抽象出在該情境中建立起來的特定關係，如買賣關係，貨物價格關係等，我們稱之爲情境誘導的關係;③表述一個情境進入到另一個情境中（成爲其子情境）因此而有了新義。比如交付當它作爲子情境含於交易情境中時，就變成了付款或交貨了。這就是“嵌入” ；④提及情境時可以有意忽略它的一些方面，人們在引

[17] 角色與條件參量是情境語義學引入的，分別參見[Barwise,J.,*et al.* 1983(1999)][Devlin, K.,1991]與[Barwise,J. 1987].

用情境時常是這樣。比如購買與出售不過是交易的部分情境，我們稱此變換爲“遮罩”；⑤只突出情境的一部分參量，即取出它的涉及其部分參量的那部分的內容，常也得到一個情境，這就是“投影變換”。比如從交易可用投影得出交貨、付款兩個情境。當然投影變換的功效遠不止於此。對一些複雜的情境講我們常能用投影得出一些非常有用的結構（不一定是情境）成爲構造其他物件（包括情境）的重要成份。此外投影運算在情境理論上也很有用（我們以後文章會談及）；⑥在已有的情境中增刪若干資訊往往又得出另一種情境或有用的結構。這就是情境的一些演算（和結果運算式）。由之常能得出一類相近的情境。比如從交易情境得出租賃借用等等情境。從而產生一叢相近的概念及它們的聯繫。等等。

以上是對一個情境而言的。幾個情境（可能還要再配上一些條件）聯合一起，常能引出（定義出）更多的概念來。這就要引入情境運算式了。這種運算式（的結果）有時又是個情境，或反過來講，有些情境可看成是由幾個更基本的情境，或再輔以若干附加條件，結合而成的。比如交易可看成是交付物品與交付錢幣兩個情境再附加若干條件做成的複合情境。等等。可以預見，基於情境的概念描述方法確實是個有力而又實用的方法。本文及後續文章意在說明只需簡單工具就能描述上述各項內容。往後還會看到，這些描述內容具有良好的數學理論作基礎，足以建立起優美的情境理論來。

限於篇幅，我們略去數學描述工具的定義[詳見附錄]，只給出用一個或幾個情境定義概念和誘導出關係的幾個具體實例，再輔以簡單解釋。

2.2.1 概念與其型式的定義形式

許多概念都可以用條件參量來表述。下面是一些實例：c_買者$\stackrel{def}{=}$ x|c_簡單交易(l,x,y, m,g;l_1,l_2),c_賣者$\stackrel{def}{=}$y|c_簡單交易(l,x,y,m,g;l_1,l_2),c_購貨款$\stackrel{def}{=}$m|c_簡單交易(l,x,y,m, g;l_1,l_2),c_商品$\stackrel{def}{=}$g|c_簡單交易(l,x,y,m,g;l_1,l_2),c_交易場合 $\stackrel{def}{=}$ l|c_簡單交易(l,x,y,m, g;l_1,l_2)等。還可以定義出它們的專屬型式[18]與專有屬性, 如 c_買者的專屬型式爲：[c_買者]=[x|c_簡單交易(l,x,y,m,g;l_1,l_2)]，該型式的專有屬性是 γ[x|c_簡單交易(l,x,y,m, g;l_1,l_2)].

另外，也可定義出一些暫時尚未形成概念的資訊內容,比如,c_買賣雙方$\stackrel{def}{=}$ <x,y>|c_簡單交易(l,x,y,m,g;l_1,l_2)等。“c_買賣雙方”實際上是買賣兩個量的序對：<x,y>|c_簡單交易(l,x,y,m,g;l_1,l_2)= <x|c_簡單交易(l,x,y,m,g;l_1,l_2),y|c_簡單交易(l,x,y,m,g;l_1, l_2)>。 相應的型式

[18] 我們把 type 譯成“型式”，有意與現成的譯名“類型”相區分。兩者其實指稱同一的物件，只是類型普遍地用在了句法領域，我們這裏則用在語義領域，故在術語上先區分一下。型式抽象運算元是[Devlin, K.,1991]、[Devlin,K. 1990]和[Barwise,J. 1987]首先引入的，儘管只是針對特定的類型的物件定義的。[Devlin, K.,1991]中對型式做了專門研究，得出有趣的結論：一切命題都可等價地表述成如下形式:p:T.[Devlin, K.,1991]用條件參量定義出兩種型式抽象，即物件抽象與情境抽象。我們要求一切參量都可抽象出一個型式，即它的專屬型式。爲此我們引入了結構參量概念。這樣做的好處是使我們的描述語言（本文未論及）是強類型的，而且可以在型式論域上建立結構。以後的文章會論證，這是一個非常重要的理論工具。

等式爲：[<x,y>|c_簡單交易(l,x,y,m,g;l_1,l_2)]=[<x|c_簡單交易(l,x, y,m,g;l_1,l_2),y|c_簡單交易(l,x,y,m,g;l_1,l_2)>]=<[x|c_簡單交易(l,x,y,m,g;l_1, l_2)],[y|c_簡單交易(l,x,y,m,g;l_1,l_2)]>= [x|c_簡單交易(l,x,y,m,g;l_1,l_2)]×[y|c_簡單交易(l,x,y,m,g;l_1,l_2)].

很多情況是由幾個情境聯合在一起引出一個新概念。下面是一個例子。

c_轉銷商 ≝z|{ c_簡單交易(l_1,z,y,m_1,g) ⊕ c_簡單交易(l_3,x,z,m_3,g) ⊕ c_擁有(l_2,z,g),l_1∩l_2∩l_3 }，

life_time(c_轉銷商)= l_1∩l_2∩l_3.（多個情境聯合定義的概念要求顯式給出其生存期。）

由 l_1∩l_2∩l_3 可推知 l_1*l_2*l_3.這裏結合符⊕是兩個情境的半加運算,後面有介紹；結合符∩表示兩個銜接的時空場合合成一個的運算，關係符*表示兩個時空場合前後銜接。顯然都滿足結合律。

2.2.2 關係抽象

引進新關係的關鍵工具就是上小節所謂的屬性抽象：若 T 是一個型式，我們用 γT 表示 T 的特有屬性（即 T 的所有元素都有的屬性）。引入新關係是極端重要的功能，我們再用一些例子來闡述該定義的內容實質，以幫助理解背後的想法。

1. 若 α 是個簡單參量，則 α 的特有屬性就是它專屬的型式[α]的屬性 γ[α]。

2. 若 α=σ′(β)是一個複合參量，則屬性 γ[σ′(β)]是 σ′(β)的特有屬性，具有該屬性的類，其型式是 [σ′(β)].這裏是把 σ′(β)當作個體參量對待的。屬於該類的物件 χ 應具有如下結構：χ=τ(δ)，滿足 τ:[σ]且 δ:[β].

比如，個體參量 c_簡單交易(l,x,y,m,g)的特有屬性 γ[c_簡單交易(l,x,y,m,g)]應至少包括：它指稱一種社會行爲，是發生在場合 l 的 x 與 y 之間的一種合作行爲、合同行爲。是經濟行爲，是買賣活動，至少涉及五個參量：在時空場合 l，其間 x 與 y 交換了貨物 g 與貨幣 m 等等。因爲是合作行爲就要受一定的規程制約，因爲是合同行爲和經濟行爲又要受一定的法律(經濟法)制約等等。(這些屬性雖不一定直接含在γ[c_簡單交易(l,x,y,m,g)]中，但可從它的上位型式等處的屬性獲得。）

與此相應地有：r_簡單交易= λ(l,x,y,m,g).γ[c_簡單交易(l,x,y,m,g)].[19]它是個關係，用在陳述《r_簡單交易,l:Loc,x,y,m,g:1》中，含義是 l,x,y,m,g 之間有"r_簡單交易"所言的關係。

3. 若 α 是個條件參量。我們則通過關係抽象（或叫 γ 抽象）來進行操作。比如 γ[m|c_簡單交易(l,x,y,m,g)]與 γ[g|c_簡單交易(l,x,y,m,g)]，就分別是簡單交易賦予的貨款與商品

[19] 這裏 λx.f(x)是 Church 的函數抽象記號，用以從具體的函數對應中抽象出一個函數。λ 也叫做函數抽象運算元。通常語言學上多用一元函數，並與句法結構相關聯。我們這裏用了多元函數，把句法資訊暫時放在了一邊。

的屬性。 γ[x,y|c_簡單交易(l,x,y,m,g)]是由簡單交易建立起的買者與賣者關係。

有時，要表達的關係要複雜一些，需要引用別的工具。但關係抽象已經提供了基本要素爲引用其他工具準備了基礎。比如貨款與貨物之間的等價交換關係——商品價格。因情況複雜要用更複雜一些的表述。我們稍微解釋一下。

r_商品價格[20] ≝γ[<id.,nominal_value.>(g,m|c_簡單交易(l,x,y,m,g))]= γ[g,nominal_value.m|c_簡單交易(l,x,y,m,g)][21].這裏的<id.,nominal_value.>是個二元運算元列,第一個是個恒等運算元 id.,第二個是“名義值”運算元 nominal_value.,作用在二元條件參量列(g,m)上,得到二元條件參量列(id.g,nominal_ value.m) =(g,nominal_value. m). 這裏的(g,m)當然是指情境 c_簡單交易(l,x,y,m,g)中的參量(g,m).而 dot 函數 nominal_value. 作用在貨幣 m 上的結果是貨幣 m 的名義值,也即票面值,因爲該屬性是貨幣的固有屬性,用 dot 函數表示。該條件參量列的型式抽象是[g,nominal_value.m|c_c_簡單交易(l,x,y,m,g)], 再對它做屬性抽象就得到在情境 c_簡單交易(l,x,y,m,g)中的 g 與 m 的名義值的元偶的屬性,也即二元關係,γ[g, nominal_value.m|c_簡單交易(l,x,y,m, g)].

應該注意,屬性抽象運算元 γ 作用在型式上，而不是作用在參量上。

3. 情境代數與情境網

上節討論了情境內部概念間的聯繫。以及如何由情境定義出關係、個體、個體元組與時空場合等量。本節我們來看情境間的聯繫。這種聯繫也間接地建立起了由它們定義的概念之間的聯繫。我們將用情境關係來表述這方面內容。此外，由一個或幾個情境經過運算轉變成另一個情境(或別的物件)也是常見的現象。它們可用映射來表示。結合起上節的內容可見情境上可定義一個代數,當然是個部分代數(partial algebra),而且其大多數關係與映射是十分“稀疏”的。這種稀疏性直接影響詞典資料的組織。對數學理論的建設是否產生影響，目前還不知道。建立情境代數或其他數學結構的目的是用數學中的關係反映概念結構中的聯繫，等等。以期最後實現用數學演算反映概念思維，包括從詞典構造直到言語交際與認知過程中的資訊提取，其中許多環節都必須實現在同一個數學結構中。可以想見這種關係與聯繫會非常之多，本文僅考查幾個與詞義描述直接有關的聯繫與運算。

3.1 情境變換與情境運算

---

[20] 我們約定用 r_作關係概念的標誌頭.這樣做的目的是爲了增加可讀性.因爲自然語言中通常用同一個詞語(術語)指稱關係和具有該關係的物件等.

另外,經常會考慮關係運算式中固定一個參量的情況.因此我們要用 λ 運算式,比如用 λ(x,y,⋯,z).γ[x,y,⋯,z|⋯], 以便利用現成的 Curry 化變換,甚至可直接寫成 γ(x,y,⋯,z).γ[x,y,⋯,z|⋯] 或乾脆寫成 γ(x,y,⋯,z).[x,y,⋯,z|⋯]等.(暫不用)

[21] <α,β>是 Bacus 的序列運算元，作用在等長的運算元列上，得到等長的結果列。此例長度爲 2.

由一些情境稍作變更就轉化成另一種情境。先來看一些最簡單的情況。

1）遮罩

實際上遮罩並不產生新情境，並不是情境變換，而只是涉及情境的引用。在交際與思維中往往並不需要完整地引用一個情境而只是引用它的一個側面，即隱蔽一部分參量不提。這部分參量自然仍存在，只是在交際或思維中無需提到它們。這種用法已反映在詞語上了。我們約定用 \α.σ(β) 表示含參物件 σ(β)中遮罩掉參量列 α（可能只含一個參量，下同。）中的所有參量。

舉例講，c_購買(l,x,m,g) ≝\y.c_簡單交易(l,x,y,m,g)，c_出賣(l,y,g,m) ≝\x.c_簡單交易(l,x,y,m,g)，可以作爲買賣兩種情境的定義，而 c_購買≝ λ(l,x,m,g).(\y.c_簡當交易(l,x,y,m,g)), c_銷售≝ λ(l,y,m,g).(\x.c_簡當交易(l,x,y,m,g)) 可作爲這兩個詞的定義。(只是它們的一種定義。因爲買賣有時還可不提及貨款等。因而還可有其他形式的定義。比如 c_購買 $_1$(l,x,g) ≝\y,m.c_簡單交易(l,x,y,m,g)等。)

下述關於遮罩運算元的性質是顯而易見的。

設 α =〈a,b,c,…〉是個參量列，π 是個排列運算元，σ(β)是個情境。顯然有：

·\πα. σ(β)=\α. σ(β)；
·\a. \b. \c. \…. σ(β)=\α. σ(β)；（左式也簡記作\abc…. σ(β).）
·\a\a\b\c\…. σ(β)= \a\b\c\…. σ(β)；

·etc.

在上述定義中我們有意使用“含參物件”這個含糊的術語，使它既能運用於情境也可以運用於情境的描述等物件。

2）增刪

有些情境彼此十分相似，只是在某些要素上稍有差異。這樣的一族情境可以設想成是由一個中心情境經不同修正逐個形成的。這就引出了增、刪、替代(換)運算。我們所謂的修改是對情境的內部結構的改動。設 τ,σ 是兩個情境，e 是個資訊元，我們用 τ=σ@e 表示 τ 是在 σ 上加進資訊元 e 的結果。相反的運算記作 σ=τ®e. 下面是簡單的例子。

通常認爲贈送他人一件東西, 應當是自願[22]給與對方所需要的東西。c_贈送(l,x,g,y) ≝c_給與(l,x,g,y)@《r_認爲,Pre(l):Loc, x,c_需要(Pre(l),y,g);1》[23]或反過來,有 c_給與(l,x,g,y) ≝c_贈送(l,x,g,y)®《r_認爲,Pre(l):Loc, x,c_需要(Pre_l,y,g);1》(從常理講，後一表述不太自然。儘管單從運算講兩式等價。)

---

[22] 下述的定義式忽略了“c_自願”含義。增加一個指稱自身的符號（比如用 THIS）就能做到。（暫略。）

[23] 此式也許該用《r_認爲,Pre(l):Loc, x,《r_需要,(Pre(l),y,g;1),;1》.這些屬於句法上的取捨還需再考慮。

下述增刪運算的性質是簡單的。

設 σ,τ 是兩個情境，設 δ,ε 是兩個描述，e 是個資訊元。

·σ@e@e=σ@e； δ@e@e=δ@e；
·σ®e®e=σ®e； δ®e®e=δ®e；
·若 σ⊨δ, 則有 σ@e⊨δ@e；

若更有 e⊨δ,則有 σ®e⊨δ®e; etc.

與增刪運算一樣，替換運算也是很重要的建立情境間聯繫的方法。我們用 σ[b\a]表示用 b 替換 σ 中出現的所有 a 得到的新物件 τ。

先看資訊元的替換。下面是個例子：

如果交易中轉讓的不是貨物的擁有權而是貨物的使用權,就成了租賃了:

c_租賃(l,x,y,m,g) ≝ c_簡單交易(l,x,y,m,g)[《r_他動-遷移,l_2:Loc,y, 使用權#g,y,x;1》\《r_他動-遷移,l_2:Loc,y,擁有權#g,y,x;1》]

[《r_需要,Pre(l_2):Loc,x, 使用權#g 1》\《r_需要,Pre(l_2):Loc,x,擁有權#g;1》]

[《r _擁有,Post(l):Loc,x, 使用權#g;1》\《r _擁有,Post(l):Loc,x, 擁有權#g;1》]

[《r_寧願,pre(post(l)):Loc,x, 使用權#g, 擁有權#m;1》\《r_寧願,pre(post(l)): Loc,x, 擁有權#g, 擁有權#m;1》].

後式也可寫成成組替換的形式：

c_租賃(l,x,y,m,g) ≝ c_簡單交易(l,x,y,m,g)[《r_他動-遷移,l_2:Loc,y,使用權#g, y,x;1》,《r_需要,Pre(l_2):Loc,x, 使用權#g 1》,《r_擁有,Post(l):Loc,x,使用權#g;1》,《r_寧願,pre(post(l)):Loc,x,使用權#g, 擁有權#m;1》\《r_他動-遷移, l_2: Loc,y,擁有權#g,y,x;1》,《r_需要,Pre(l_2):Loc,x,擁有權#g;1》,《r _擁有, Post(l): Loc,x, 擁有權#g;1》,《r_寧願,pre(post(l)):Loc,x, 擁有權#g, 擁有權#m;1》].

或直接寫成 c_租賃(l,x,y,m,g) ≝c_簡單交易(l,x,y,m,g)[使用權#g\擁有權#g]，如果知道情境 c_租賃(l,x,y,m,g)的有關結構的話。爲了能涵蓋這種情況,我們把替換演算定義成較寬的形式.並把情境中的替換僅當作一種特例。[詳見附錄。]

最常見的情況是 σ 是個情境，所得到的 τ 也是情境。或 σ 是個描述，τ 也是個描述。易見，對任意資訊 u，若有 σ⊨u，則有 σ[b\a] ⊨u[b\a]。式中 u[b\a]表示在資訊運算式 u 中出現的 a 全替換成 b 的結果。

下面是另一例子：

如果在交換物品的情境中一方付給的是貨幣，那就是我們的簡單交易了：

c_簡單交易(l,x,y,d,e) = c_互換物品(l,x,y,d,e) @ (arg:d: [c_貨幣])

替換也有一些簡單性質。限於篇幅，從略。

3）投影與嵌入

設 σ 是個情境，e 是個資訊元，滿足 σ⊨e.常有這種情況：需考慮是否存在和取出 σ 的一個部分(叫做 σ 的部分情境)τ，使支援關係 τ⊨e 成立[24]。比如 e 所含的參數只牽涉到 σ 的一部分,可考慮 σ 的含這些參量的部分 τ.

爲此我們考查一種構造部分情境的方法：由它的只含某一部分參量的部分資訊元做成的情境，叫作它在這幾個參量上的投影。我們用 Proj(σ,(i_1,⋯,i_k))表示 σ 中含有參量列(i_1,⋯,i_k)上的參量的那個部分，叫做 σ 在參量列(i_1,⋯,i_k)上的投影。

這種投影可稱之爲最小投影。再看下述的定義：我們用 Views(δ(l,j_1,⋯,j_n), (i_1,⋯,i_k)) 記一個新情境描述，它由 δ(l,j_1,⋯,j_n)的定義中的那些直接和間接含有屬於(i_1,⋯,i_k)中的參數的資訊元組成，叫做 δ(l,j_1,⋯,j_n)在(i_1,⋯,i_k)上張開的（或支起的）部分描述（稱作視景）。它顯然是含(i_1,⋯,i_k)的最大的部分描述。

情境描述 δ(l,j_1, ⋯ ,j_n) 在 (i_1, ⋯ ,i_k) 上的投影 Proj(δ (l,j_1, ⋯ ,j_n), (i_1, ⋯ ,i_k)) 與 Views(δ(l,j_1,⋯,j_n), (i_1,⋯,i_k))都當作獨立於其母體 δ(l,j_1,⋯,j_n)的子情境的描述。需要考慮其出處時可用 within δ(l,j_1,⋯,j_n)來限制。

當然還可定義別的投影，只要有需要。

下面性質說明 Proj 與 View 都是投影運算元。

簡單性質. 設 δ(l,j_1,⋯,j_n)是個情境描述。(i_1,⋯,i_k)是(j_1,⋯,j_n)的一個子列。顯然有

Proj(δ(l,j_1,⋯,j_n),(i_1,⋯,i_k))⊆δ(l,j_1,⋯,j_n);
Views(δ(l,j_1,⋯,j_n),(i_1,⋯,i_k))⊆δ(l,j_1,⋯,j_n);

Proj(Proj(δ(l,j_1,⋯,j_n),(i_1,⋯,i_k)),(i_1,⋯,i_k))=Proj(δ(l,j_1,⋯,j_n), (i_1,⋯,i_k));
Views(Views(δ(l,j_1,⋯,j_n),(i_1,⋯,i_k)),(i_1,⋯,i_k)) = Views(δ(l,j_1,⋯,j_n), (i_1,⋯,i_k)).

投影可以看成是引入新情境（部分情境）的方式。與此相反的是一個情境嵌入到另一個情境中，從而獲得新義的做法。我們用 τ within σ 表示“處在 σ 中的 τ”，讀作“囿於情境 σ 的情境 τ”。比如付款，就是在交易情境中交付（給與）貨款的行爲。交貨就是交易中的交付（給與）貨物的行爲。此時付款與交貨都是交易的子情境（部分情境），而簡單交易則可看成是由交貨與付款兩部分組成的。離開了交易，這兩個交付（給與）就不再是付款與交貨了。

“付款”與“交貨”這兩個概念只有在交易情境中才能定義出來。下面是定義式。

---

[24] 特別要提到的是，對作爲廣義情境的“真實世界”而言,恐怕只能考慮取用其一部分。因爲[Devlin, K.,1991]中已有論證，該世界不是情境，只能簡化地看作情境。我們做的相當於把它的部分定義成情境。

c_付款≝λ(l_1,x,m,y).(c_付給(l_1,x,m,y) within c_交易(l,x,y,m,g;l_1,l_2))

reminder {contains(l,l_1),coend(Proj(l,Tmpo),Proj(l_1,Tmpo))};

c_交貨≝λ(l_1,y,g,x).(c_付給(l_2,y,g,x) within c_交易(l,x,y,m,g;l_1,l_2))

reminder {contains(l,l_2),coend(Proj(l,Tmpo),Proj(l_2,Tmpo))}.

這兩個概念既從“付給”情境中獲取\繼承資訊，又從“交易”情境中獲取\繼承資訊。

注意，我們未要求嵌入的情境與被嵌入的情境所在的時空場合完全一樣，有時需加說明，以便應用。這裏是用提醒算符 reminder 與隨後的資訊來完成的。其作用是明白給出嵌入時必須滿足的條件。儘管此條件在被嵌入的情境中能查到。 換一種視角看，嵌入是在作爲基相的背景情境中突出其一個部分，也許稱作突顯（saliencing,salientify）更貼切。這在日常思維與表達中非常常見，不限定在辭彙層中採用。

投影與嵌入，以及下面幾個運算，也都有若干簡單性質，限於篇幅，一律從略。

4）半加⊕（聯）與相容或

§2 中說過，簡單交易可粗略看成是兩個給與的合成。現在我們給出它的符號表述。

設 σ ,τ 是兩個情境。我們用 σ ⊕τ 表示它們的聯合體，稱作 σ 和 τ 的半加或聯立。

設 δ,ε 是兩個基本描述。我們用 δ⊕ε 表示它們的集合並 δ∪ε 做成的描述。

下面的性質可以看成是情境聯合體的意思的直覺解釋。

簡單性質：設 σ,τ 是兩個情境，設 δ,ε 分別是它們的描述。則 δ⊕ε 是 σ ⊕τ 的描述。

情境的聯合體不一定是個情境，因爲沒有指定其生存時空。但已可構成一個情境的主體。舉例講，c_贈送(l_1,x,d,y)⊕c_贈送(l_2,y,e,x)實際上給出了“互贈物品”情境的主體部分。由此可以給出互贈情境的定義如下：

c_互相贈送(l,x,y,d,e) ≝ (c_贈送(l_1,x,d,y)⊕c_贈送(l_2,y,e,x)) with l = span(l_1,l_2).

這裏 with 是個附加時空場合(或時間)的算符。它用隨其後的時空場合(時間)來替換位於其前的情境的時空場合(時間)。Span 是時空場合型式上的函數，span(l_1,l_2)的含義是由 l_1 與 l_2 張開的時空場合，即，包含 l_1 與 l_2 的最小時空場合。

下面是另一個更有意思的例子。

簡單交易(l,x,y,m,g;l_1,l_2) ≝ (((c_給予$_1$(l_1,x,m,y) ⊕ c_給予$_1$(l_2,y,g,x))with l = span(l_1,l_2))@{

《coend,Proj(l_1,Tmpo), Proj(l_2,Tmpo);1》

Superiors:

c_合作行爲: …

c_合同行爲: …

…

End_Superior

}

在陳述條件參量時，曾引用過情境聯立的情況[Cooper, R.1986]，通常還要附加一些額外條件，做法與上面的一樣。與此相似地，情境σ,τ 的“相容或” $\sigma|\tau$，也是很常用的，另一種情境的聯合體。同樣也可附上一定條件用來定義條件參量或引入概念等。就不舉例了。

對情境聯合體還可進行增刪替換運算，只要把前面的定義稍微修改一下即可，暫略。

5）提升(概括、抽象)

我們僅給出三種提升，即個體化，諢名與資訊元化。

- 個體化. 把一個複雜的有結構的物件個體化是使它能轉換成參量的前提。在我們的描述中當其他量要作爲個體進入資訊元時就有一個個體化變換。個體化的含義是忽略它的結構，從整體上來把握它。日常語言中頻繁出現的名物化就是個體化現象。

若 e 是一個非個體的物件。我們可用 ιe 表示把它轉換成個體，即忽略它的結構，從整體上把握它。由於在資訊元中個體的位置明顯的表明了它們的個體身份，因此在句法表示上我們常常忽略這個變換。但在詞義/詞性標示時卻常要用到。

- 諢名. 人們經常是用物件的某個特徵來稱謂該物件。這就把它在特定的情境中的角色/屬性帶出了那個特定的情境，即一種泛化現象。我們用諢名運算元(nickname)來記錄這種用法。比如，我們常把用作進行教學活動的房間叫做教室。儘管當時並未進行教學活動。又如，我們把行車情境中駕馭車輛的人叫作司機(駕駛員) 。但也常用司機指稱其人，儘管此時他並未駕馭車輛，等等。這就是用物件的一項特徵稱謂它的做法。其描述爲：

c_教室≝nickname(Proj(l|c_教學(l,…),Spat)),

c_司機≝nickname(x|c_駕駛(l,x,v)).(比較:c_駕駛員/者≝x|c_駕駛(l,x,v)).

諢名運算元 nickname 是一種泛化運算元。當它作用在條件參量 x|Σ 上或是作用在角色[25]x|⟨σ⟩ 上，就是將 x 在 Σ 或 σ 中扮演的特定角色用作 x 的稱謂，從而可在其有效範圍之外稱呼它、引用它: nickname(x|Σ), nickname(x|⟨σ⟩),可見是非常強的手段！

[25] 條件參量其條件式是只含一個情境的就叫做角色。

在這層意思上諢名有點像摹狀詞,它們的聯繫可不嚴格地敍述成:

nickname(x|⟨σ⟩)={ιx|∃τ:[σ]∃y. ρx *contains* y∧(y *position* *in* τ ≈ x *position* *in* σ)[26]}

右式是說，x 是滿足下述條件的物件:存在與 σ 同型式的情境 τ 與個體 y, y 在 τ 中的角色與 x 在 σ 中的角色一樣且 x 的歷程(經歷)的一部分曾是 y 在 τ 中的那個角色。其中 ρx 表示 x 的經歷,即過程地看，x 歷經了哪些變遷，ρ 叫過程化運算元。[下一節有解釋。]

- 資訊元化. 將一個情境或含參物件轉換成資訊元的做法我們在上面的例子中已經遇到了。這相當於將客體中真實地或虛構地存在的事物轉換成對它們的認識、表述，這種轉換及其逆至少在句法上是很有用的。例如，設有一個情境 c_駕駛(l,x,v)，我們用

info(c_駕駛(l,x,v)) =《r_駕駛,l,x,v;1》

表示與該情境相對的一個資訊元，即言及它的、指稱它的資訊元。反過來也常要求把資訊元轉換成它所言及的、它所指稱的物件。比如用

deinfo《r_駕駛,l,x,v;1》= c_駕駛(l,x,v)

表示把資訊元陳述的資訊內容對應成爲它所言及的、所指的物件。*Info* 與 *deinfo* 分別叫做資訊元化運算元與去資訊元化運算元。它們的句法性質與語義性質，因篇幅，從略。

3.2 過程化（概要）

過程化是與生存期密切相關的概念。“一切事物都是過程。”即，個體與事件都是過程。“一切過程都生活在一定的（自己的）時空裏。”即，個體與事件都有自己的生存時空。

若 σ 是個情境，x 是在其中存在的某個個體，e 是其中的某個事件。我們用 π(x, σ) 與 π(e,σ)分別表示將它們看成是在 σ 中它們在各自的生存期上歷經的過程(叫歷程)。或許不用指定所在的情境，就簡化地分別用 πx 與 πe 表示，即看成是生存期上的時間的函數,叫做 x(或 e)的歷程。

事物的歷程是個非常重要的概念。儘管通常是給不出歷程的細緻描述的。但常常是，只要能給出它的部分描述就是很重要的資訊。如何描述事物的歷程，我們還沒有一套有效的方法。我們猜想區分兩種歷程也許是必要的：即，事物內在因素的演進與事物與外部環境的關係的變遷。前者可想象成是對用情境表示的物件在時間上的演進的表示（即對該情境的細化），後者可看作是對個體在其環境中的角色作用的細化。故用符號區分開來。

設 x 是個個體，e 是個情境。我們用 ρx 表示 x 在其生命歷程中所扮演的角色的經歷，用 μe 表示 e 在其生存期（存在時段）中自身經歷的演變。

---

[26] 這僅僅是示意式，嚴格的表述將在後面的文章中給出。

分階段描述可以看成是對事物歷程的一個近似描述。對 ρx 與 μe 都合用。其想法如下。

設 x 爲任意事物。若把 x 的生存期 t 分割成若干時段 $t_1,t_2,\cdots,t_k$.設 $St_1,St_2, \cdots ,St_k$ 分別是 x 在這些時段上的狀態。則我們把事態描述〈t_1:St_1〉〈t_2:St_2〉…〈t_k:St_k〉叫做對 x 的歷程的階段的一個描述，簡稱做 x 的分階段描述，如果對任意 j 而言, 〈t_j:St_j〉都是對 x 在時段 t_j 上的情況的描述。每個〈t_j:St_j〉就叫做 x 的一個階段(j=1,2,…,k) 。

比如講，如果 x 是用情境描述的事件，那麼就可以把 x 的事態描述〈t_1:St_1〉〈t_2:St_2〉…〈t_k:St_k〉看成是它的更精細的描述。

通常只能得到事物的階段描述的一部分，或叫部分階段描述。儘管只是這個物件的生命歷程的部分階段描述，也常是重要的。通常的語言邏輯分析工作可以佐證。

同一事物的歷程可有各種不同描述。不同描述之間可有各種關係。比如，階段描述、全局描述、部分描述、完整描述等。因爲同一事物同一階段的不同描述應該是相容的（所謂橫向相容性）。描述之間可以比較精細程度，按信息量的多寡可建立序關係。上例其實已反映了精細關係的序的基本想法。在此序關係下衆多描述組成一個定向集。可有一定的拓撲結構。對之我們可以建立相應的數學理論。我們以後將用收斂的定向描述集定義情境。從而用定向集的聚點定義可認識的事物。此外，對歷程的分階段描述也引來情境的接續運算與接續條件，即縱向相容性關係等的考慮。縱橫兩個方向各自的相容與互斥性交互作用可產生複雜的情景，反映情境在時間空間上的複雜結構。等等。所有這些都是重要而又有趣的。

過程進入情境理論考查的視野一下子就呈現出十分生動、十分豐富的特性來。儘管目前描述的手段還很粗糙，但因其本身重要，相信理論很快會充實發展起來的。

3.3 情境間關係（簡述）

不言而喻情境間的關係也是極豐富又極重要的。限於篇幅，我們不能詳細敍述它們。只粗略地看看它的大體面貌。先看情境語義學的有關內容。

情境語義學一開始[Barwise,J.,*et al.* 1983(1999)]就極其重視客體情境之間的關聯,稱作約束(constraint),認爲是客體情境攜帶資訊的機制[27]，並用關係 involving/involves 來概括/陳述這種聯繫。這裏 involves 是一類關係，[Barwise,J.1987]把它們分成：反映自然定律的名義約束(nomic constraint),反映本體上必然聯繫的必然約束(necessary constraint),與反

[27] 情境語義學的重要信條就是資訊存在於客體情境的約束(constraint)之中。[Barwise,J.,*et al.* 1983(1999)][Devlin, K.,1991][Barwise,J. 1987]

映強弱兩種邏輯聯繫的邏輯約束(logical constraint)⊳與≫[28] .

參照上述內容，結合語義詞典需要，我們隨意列舉如下一些實例，希望能解釋清楚情境關係既重要，不可或缺，又普遍存在，處處遇到。隨後的文章還將說明，在我們的框架裏很容易描述與研究、討論它們。

情境變換與運算引出部分情境、子情境等關係，它們是包攝關係的特例。橫向的相容與互斥關係可引出情境的正交分解等的討論。比如講，正交分解用在條件參量中的條件運算式時就是剖分情況的依據，從而成爲剖分情境的方式之一；類似地，縱向的相容與互斥則可引入（事件與個體的）歷程的正交分解等。它們在考查事物的可共存性、事件的可接續性、事物歷程的正交分解等方面的問題上是基本的。縱橫兩向的相容性與互斥性的交互作用爲我們考察事物經歷及其可能的發展提供了基本工具。

情境之間的上下位關係也比通常的義類分類法更靈活。傳統的同義、反義、近義、對義等關係往往界限模糊，用情境描述可以給出更確切的刻畫。從情境描述出發我們能給出，相仿關係，它們可作爲所謂"隱喻"的數學模型，等等。

對至關重要的因果關係也可提供便利的表示方式。首先，對已確認的因果關係可有顯式表示。對於與因果鄰近的其他關係，比如，相伴、互補、引發、中止、終結等關係也能給出明確界定。下面我們以誘導與引發兩個關係爲例說明實現方式。

誘導運算元是產生各種新關係的重要手段。若干情境的共存常使它們涉及的時空、個體間產生聯繫，從而在情境間引起誘導關係。新關係的建立往往有明確的時間起止和空間範圍。一個情境或一組情境的發生與存在往往又是另一情境出現、誕生或消亡的開始。這種類似因果的關係稱之爲引發。即以親族關係爲例。x 與 y 有夫妻關係，y 與 z 有母子關係，就誘導出 x 與 z 有父子關係。類似地，可以定義（狹義）同胞關係，姑嫂、叔侄、祖孫等等關係。加上前面介紹的各種運算還能定義，比如，兄弟、兄妹、姐弟、姐妹等等關係。如果要表示這種種關係何時（何處）誕生，就要用到引發關係。比如，y 生產 z（生產情境）的事件就引發了 y 與 z 的母子關係與 x 與 z 的父子關係，更引發了 z 的生命史（存在情境）。等等。凡此種種都能夠給出足夠精確的描述[見附錄]。

設 e_1,e_2 是兩個情境(事件)。如果情境(事件)e_1 的完成就意味著情境(事件)e_2 的開始，我們說情境(事件)e_1 引發了情境(事件)e_2。記做《initiation,e_1,e_2;1》。如果 t_1,t_2 分別是 e_1 與 e_2 的生存時間，則有《meet$_t$, t_1,t_2;1》。意即 e_2 緊接在 e_1 之後發生。

類似的，如果情境(事件)e_1 的完成就意味著情境(事件)e_2 的完成，我們說情境(事件)e_1 終止了情境(事件)e_2，記做《termination,e_1,e_2;1》。 如果 t_1,t_2 分別是 e_1 與 e_2 的生存時間，此時有《$=_e$, t_1,t_2;1》。意即 t_2 與 t_1 終端相同，也即 e_2 與 e_1 同時結束。

---

[28] 含意分別是 1.若α是事實，且 $\alpha \triangleright \alpha_1$，則 α_1 是也事實與 2. 設 σ 是某個客體情境, 若有 $\alpha \gg \alpha_1$ 且命題 $\sigma \models \alpha$ 爲真,則 $\sigma \models \alpha_1$ 也爲真。等等。當然都是對客體世界而言的。

注意，我們用了兩個情境。意即，當引發的是個關係或個體時也都是當作過程來對待的。關係與事件、事物一樣地是有其生存時空的，至少是有生存期。我們可以就此定義出各種“時間邏輯”算符。比如講 till(until,unless),during(when,while)等。

歸結到一點，可以說關於情境可以建立起內容豐富的代數理論和拓撲理論。

3.4 關於情境網

情境按相互聯繫構成一個網，叫情境網。該網的主幹是上下位關係。上下位用於知識的繼承，遵循鄰近優先原則。選定一些情境作基本情境，叫基本節點。其他情境是能用基本情境通過運算生成的，叫派生節點。基本節點上有情境的定義。派生節點上只有生成它的運算式。選做基本節點的情境應是基本的，或者是使用頻度高的。每個節點處附有類似公理/規律/規則形式表述的相關的知識。情境定義中陳述的該情境與其他情境的聯繫在該節點處要有啓動機制，以備需要時用。情境間的其他各種關係建立起的聯繫也有相應要求，使用時往往要配有搜尋。因此情境要做一些歸類。我們有一個初步的分類體系，大約是把事物或其各個方面按物質、精神、符號、人際社會四大部類及其相互作用分成各種類別。這裏就不談了。上述種種，包括未提到的，都旨在實現資訊的聯想。這裏面要用到一定的推理機制，但不限於此。這裏不細討論。我們的做法是，把需要的知識放在相關的局部節點上，使用時由當時的資料根據需要驅動。

4. 關於情境描述

情境描述是情境的資訊結構的一個有限近似。它可以看作是(人腦中的抽象)情境的一個模型。情境描述給出了情境型式的資訊結構的一個詮釋，儘管只是近似的，但它卻是具體給出的。在引入描述之前，我們關於型式、參量、典型元素、特有屬性等概念，在談到它們所共有的資訊結構時只能是很空泛地、很抽象地表述。有了描述的概念，就可給出它們的資訊結構所可能有的部分主要內容。以後我們還要論證描述和由它定義出的各種量的收斂性，其中就有對應的資訊結構的收斂性。

我們關於情境 c_簡單交易(l,x,y,m,g;l_1,l_2)的描述(見第 2 節)δ(l,x,y,m,g;l_1,l_2)同時也就給出了概念 c_買這等的近似定義。一個情境可有許多描述，它們之間有一種協調關係，這也使它們定義出的各量之間有協調關係，等等。上述的收斂性在此就很重要了。由上面的例子可以看到，描述一個情境總是要用到一些概念的，通常還要用到其他情境。而我們知道概念又是用情境定義的，這就涉及情境描述 的體系結構問題了。我們假定有一些基本概念，叫做語義原語，它們的意思假定是自明的，人們對之沒有異議，不含歧義。有一些基本型式，也含意明晰，大家無異議，它們相當於原始語義範疇。由這些原語與原始型式可以定義出一些情境描述，由這些情境描述又可定義出一些新概念，新概念加入已有概念又可進一步定義新情境描述，如此反復不已，得到一個情境描述網和一個概念結構，是對我們腦中的抽象情境與概念體系之網的近似描述。由上面的例子還可以看

到，由一些資訊常能推演出另一些資訊，比如由“擁有”可推演出“支配權”的歸屬，而由“支配權”又能得出相應行爲的合法性。等等。這些資訊間的聯繫是與背景文化有關的，用我們的話說，它們是與所在世界有關的。它們可以用公理形式放在適當的概念或資訊（元）處。此外，情境與情境之間也有種種聯繫，也是與所在世界相關的，它們能誘導出概念間的更多更重要的聯繫來。這些聯繫應放在情境網的適當位置上。這又産生了各種各樣的推演，是屬於語義詞典的。可見情境描述應當封閉於某些個推演。每個描述實際是個閉包，是可以由我們給出的有限描述生成的閉包。這樣，如何組織上述的網，恰當安放合用的推演等資訊就至關重要了。後續文章將試圖探討這種組織方法，和它們所倚的數學理據。由情境的描述可以給出該情境定義的各種量的描述。上面已講到同一個情境可有許多不同的描述。由它定義出的各種量也就有許多描述。同一個物件的不同描述之間應當是協調的。合在一起仍然是該物件的一個描述，而且是一個更豐富的描述。如何“演算”出這種“複合”描述，如何保證與論證這種協調性，沒有強有力的數學作後盾是不可想像的。在後續文章中我們將引入信息量概念，並論證這種不斷豐富的過程的收斂性問題。等等。所有這些內容不僅有趣，而且有用。順變提一句，今後我們的理論將以描述爲基本構件來建設。

5. 結束語

將情境作爲認知圖式的數學模型，在概念生成的情境中定義、描述概念，優點來自情境是組織與存放概念知識的最佳框架。與此同時情境理論（只要稍加改造）提供了現成的良好的理論工具。建立情境代數意義重大。它是實現我們的目標：把語義學建基於辭彙語義學之上的必不可少的中樞理論之一。我們隨後的工作還將進一步延伸它、發展它。後面文章還要說明：在使用情境中對照、落實、引用、還原概念的定義情境的做法。那時情境代數的功用就會顯得更清楚，當然，可能主要要用它的另一部分。那時情境作爲辭彙語義的組織與存儲概念知識的最佳框架的作用也才能得以顯現出來。這裏具有統一的代數理論（以及別的理論）就非常重要。

我們的目標是建立概念、概念組織（結構）與概念思維的數學理論。從而使我們能用概念演算，特別是代數演算(包括邏輯演算)反映概念思維。當然，語言與語言的使用（言語）的數學理論將作爲特例包含、嵌入在其中。這是任何理論學科成長的必由之路。如此建立的辭彙語義學理論才可望成爲建立語義學的基礎。在其中，本文著重在概念與概念聯繫的描述方法上，順帶也給出了描述所用的元語言的雛形。後者還需要嚴格定義。

我們的描述體系可以隨意提煉時間、空間，可以根據需要不斷地定義和生成新的關係，形成個體(用角色等)並引入條件化參量、複合參量，等有結構的參量，可以通過型式抽象引入新的型式。可用個體化、資訊化隨意提升，可用過程化轉而考查個體或事件的進程中的情況等等，研究事物在其生命期中的許多現象。這和只限定在預先選定的幾種關係和若干固定的元物件上的做法不一樣。和它們相比，同樣是用概念間的關係來描述概念，我們的做法有很大的靈活性和較強的生成與描述能力。我們能定義諸如“元概

念”、“亞(次)概念”、“超概念”等類物件，研究概念的詞語化，概念組織方式的語法化現象等等，從而獲得極大的自由來研究概念體系及其外顯形態問題等。此外，我們把對情境、概念、關係等物件的描述取作爲基本數學物件，使我們的工作能有良好的數學理論作後盾。這些將在隨後的文章中展開。

在隨後的論文中，我們還將重點考查：情境間的關係與演算；信息量的序關係及相關問題；原語的理論問題；時間結構；關於語義辭典組織；句法理論是語義規律的抽象的設想等。進行相關的理論研究，包括：關於描述，關於型式體系以及型式理論，特別是它與論域理論的關聯等。

鳴謝

本項研究得到國家自然科學基金（專案號：60173008）、國家 973 基金（專案號：G1998030507）和國家 863 計劃（專案號：2001AA114040）資助，謹在此致謝！

兩位審稿的專家提出了寶貴的意見，消除了原文中的錯誤，和不精確之處，對改進本文質量極有好處。謹致謝！

參考文獻

Baker, C.F., C.J.Fillmore and J.B.Lowe, "The Berkeley FrameNet Project", *COLING-ACL'98,* 1998, pp. 86-90.

Barwise,J. and J.Perry, "Shifting Situations and Shaken Attitudes", *《Linguistics and Philosophy》* vol.8, 1985, pp. 105-161

Barwise,J., "Information and Circumstance", in *The Situation In Logic*, Stanford: CSLI Publications, 1989

Barwise,J., "Recent Developments in Situation Semantics", in *Language and Artificial Intelligence*, edited by M.Nagao, Elsevier Science Publishers B.V.(North Holland), 1987, pp. 387-399

Barwise,J., J.Perry, *Situations & Attitude*, MIT Press,1983; Re-issued by CSLI Publications,1999

Chengming Guo (郭承銘), "Driving a Natural Set of Semantic Primitives from Longman Dictionary of English", *Machine Tractable Dictionary: Design and Construction*, Ablex Publishing Corporation, Norwood New Jersey, 1995, pp. 295-312

Cooper, R., "Tense and Discourse Location in Situation Semantics", *《Linguistics and Philosophy》* vol.9. No.6, 1986, pp. 17-36

Devlin, K. , *Logic and Information*, Cambridge: Cambidge University Press, 1991

Devlin,K., "Infons and Types in an Information-based Logic", in *《Situation Theory and Its Applications》* R.Cooper, K.Mukai, and J.perry(eds), 1990, pp. 79-96

EDR, *EDR Electronic Dictionary Specification Guide* , (TR-041) , 1993

Jackendoff, R., *Semantic Structures*, Cambridge:The MIT Press,1990

Lenat, D.B., R.V.Guha, *Building large knowledge-based systems:representation and inference in the CYC Project*, Reading,MASS.:Addison Wesley Pub., Co., 1989-1990

Miller,G. A., R. Beckwith, Ch. Fellbaum, D.Gross,and K.Miller, *WordNet: an On-line Lexical Database*, (revised)1993, 8.

Richardson, S.D., W.B.Dolan, and L.Vanderwende, "MindNet: acquiring structuring semantic information from text", *COLING-ACL'98*, 1998, pp. 1098-1102

Thomason,H.(ed.) *Formal Philosophy, selected papers of Richard Montague*, Nrw Havens and London: Yale University Press, 1974; third printing 1979

《Linguistics and Philosophy》 vol.8, 1985

陳群秀, 張普, "資訊處理用現代漢語語義分類體系:屬性分類", *《中文資訊處理平臺工程》*, 陳力爲, 袁琦 主編, 1995, pp. 206-219

陳祖舜,"資訊語義學: 一個新的計算語義學的構想", *《電腦科學》*vol.22 No.6, 1995,pp. 1-6,

董振東, "知網", *《計算語言學文集》*, 黃昌甯, 董振東 主編, 1997, pp. 19-24

黃昌甯,陳祖舜, "關於語義辭典構造的一些初步設想"，*《中文資訊學報》*vol2., No.3,1988, pp. 1-9

黃曾陽, *HNC(概念層次網路)理論: 電腦理解語言研究的新思路*, 北京:清華大學出版社, 1998

賈彥德, *漢語語義學* ，北京：北京大學出版社，第二版，1999

雷永生, 王至元, 杜麗燕, 李浙生, 高爾強, 陳曉希, *皮亞傑發生認識論述評*, 北京:人民出版社,1987

林杏光, 張慶旭, "現代漢語槽關係研究", *《語言工程》*陳力爲, 袁琦 主編, 1997,pp. 19-24

魯川, "現代漢語的語義網路", *《中文資訊處理平臺工程》*陳力爲, 袁琦 主編, 1995, pp. 232-252

瀋陽, 鄭定歐, *現代漢語配價語法研究*, 北京: 北京大學出版社, 1995

魏宏森, "申農資訊理論的科學貢獻/申農資訊理論的方法論意義", *《科技日報》*, 1998.7.18. 第 4 版

袁毓林, *漢語動詞的配價研究*, 南昌市:江西教育出版社, 1998

張普, "論語義場" , *《中文資訊處理平臺工程》*, 陳力爲, 袁琦 主編, 1995, pp. 183-194

張普, "資訊處理用現代漢語語義分析的理論與方法",*《中文資訊學報》*vol. 5, no. 3, 1991.3, pp. 7-18

另載: *《中文資訊處理平臺工程》*, 陳力爲, 袁琦 主編, 1995, pp. 195-205

趙海，王永成，王傑，馬穎華, "基於人工意識概念的人工智慧科學的重構"，*《模式識別與人工智慧》*，v.15.no.2, 2002, pp. 155-160

鍾義信, "從'統計'到'理解'，從'傳輸'到'認知'"，*《電子學報》*vol.26.no.7, 1998, pp. 1-8

Chen Zushun, Ma Liangrong, Guo Chengming, Ma Zhenhua, "Time Structure", 1996 (unpublished)

附錄 相關數學量的定義

文章[Barwise,J.1987]附有情境理論參考手冊（Situation Theory Reference Manual），可參看。

本附錄是就本文範圍（及進一步討論）要用到的諸概念的便覽。其中有些內容就源自上面手冊[29]，當然是變更了相應的基礎的,且也已有所更動與添加。

A. 若干概念的數學定義

本節介紹本文用到的與情境有關的定義。

A1 型式與參量、常量和變數[30]

定義 A1.1　型式是一種類型，即一種有共同屬性的（個體）物件的抽象。在我們這裏，共同屬性指的是有相同的（資訊）結構[31]。

若型式 T 是某個元素類 L 的抽象，通常也用 T 泛指這個類 L.

定義 A1.2如果 T 表示某個型式，T 中的元素（即抽象出它來的那個類 L 中的元素）所特有的那種屬性則稱作 T 所特有的屬性，或 T 的屬性。用 γT 表示。

定義 A1.3設 x 是某個個體物件，我們用運算式 x:T 來表示 x 有 T 特有的屬性 γT，並稱 x 是屬於 T 的，也說 x 是 T 的元素等。在一定條件下（指如果型式可作爲個體），x:T 可用《type_of, x, T;1》表示。此式與《γT, x;1》等效[32]（這裏是把 γT 當作關係來對待，屬性可以當作一種一元關係）。

定義 A1.4[33]　參量是型式的特有屬性的化身、具體化、實體化，可以看作是恰有上面講到的那種特有屬性（指一類元素所共有的資訊結構）的一種抽象的量，是這個類的典型元素、抽象元素，是其代表。

若 T 是某個型式，x 是恰有 T 的特有屬性 γT 的參量，此時我們說參量 x 恰屬型式 T，也說型式 T 是參量 x 的專屬型式。同時也稱參量 x 是型式 T 的典型參量。該關係記做 x::T. 易見，若 x::T 則有 x:T.

事情常有相反的一面，即先知道某類物件的（資訊）結構，需要求出該類和它的抽象——某型式。

---

[29] 詳細情況請見下面有關的註腳、注釋。

[30] 這些都是情境理論中的基本概念。

[31] 參看§4 關於情境描述

[32] x：T,《type_of, x, T;1》,與《γT,x;1》這三個式子是不完整的陳述。從前面的例子與後面的論述可看出，一般而言，一個物件具有某個屬性或屬於某個型式（類型）不是無條件的，而是就一定的時空而言的。因此在這裏應當加上存在的時空。

[33] 參量與型式是情境理論中的基本量。一切參量 x 都可抽象出一個型式，即該參量的專屬型式[x]；每個型式 T 都有一個特有屬性 γT，這些是我們加的。

定義 A1.5如果 x 是個參量，我們用[x]表示 x 的專屬型式。這裏[ ]是個算符，叫型式抽象。含義是從參量 x（即從 x 的資訊結構）抽象出的型式。

顯然，總有 x::[x]. 且若 y::[x]，則有[y]=[x]，因此也有 x::[y]. 而 γ[x]就是參量 x 的專有屬性。顯然總有《γ[x], x;1》如果認可這個表示式的話。

除了參量，我們還用到了其他兩個基本量：常量與變數，並約定所有出現的基本量都要滿足一定的型式要求。即存在型式 T，使 x:T. 其中，

常量是用來指稱實際存在的具體的物件[34]的。這種物件的內涵一般講是潛在無限的，因此常量的內涵也是潛在無限的。用上面定義的型式來表述，就是常量沒有專屬的型式，它所屬的型式是無限多的。常量只在具體的世界中存在，不管是真實世界或假想的、虛構的世界。一個世界中除了有一些個體常量外，可能還有一些關係常量、情境常量等高階的量。當然，時空場合常量總假定是有的，而且就是本文所選定的那種結構(參見[Chen Zushun *et al*.unpublished])。如果我們運用的是不同時空結構，那就需要明白定義出來。比如在考察行爲模式的語義問題時常遇到的那樣。

我們約定,除了需要時引入各種可能的虛構世界外我們固定四個世界,它們是現實世界 real,它對應客體世界的真實發生了的事與物的情況；理想世界 ideal，它反映我們當前所把握的所有真理的最高境界，即相對真理世界；絕對真理世界 God\Omnisciece,無所不包的終極真理，它的存在僅僅是爲了表述當前認識可能也含有錯誤,和供修正之用；系統實現（世界）System,系統中真正實現了的那個有限的近似。不把 real 與 ideal 合併是想把不具普遍性的真知從理論王國中劃分出去。

定義 A1.6（世界）一個參量 x 在某個世界 σ 中叫做適定的，或是有定義的，是說 σ 中含有型式爲[x]的量。

Devlin, K.[1991]曾論證道，現實世界不能等價於一個情境。我們將假定它的，以及任何世界的，一個投影可以當作一個情境（主要是指可有有限近似，並能無限逼近）。暫略。

變數和參量一樣是形式量。在我們目前的體系中，變數只是用來協助陳述函數並進行關係(函數)抽象的（通常用在引入新的關係時）。變數在函數中使用時有型式要求，即在給它們定值時要滿足型式要求。

正統情境語義學沒有提到變數，但實際是要用到的。做參量與做變數本是同一量的不同身份，用途又都很專一。本文僅僅是引用它們，就沒有在記號上再區分它們了。

---

[34] 注意，這裏的“實際存在的具體物件”除了指真實世界中切實存在的具體物件外，還包括虛構世界中的具體物件，比如在小說故事中談論的具體物件，或在言談話語中談論的虛擬的或具體的事物等。我們也認爲這些物件的內涵是潛在無限的，只是小說作者或談話者等未能全交待出來而已。另外，對常量，我們也常只知道它的部分資訊。但它是確定的，並不因資訊不足而不確定。這點與參量是不一樣的。

至此，應該說所有的內容還大體上沒有脫出數學中關於常量、變數與參量的貫常用法。 下面我們介紹有結構的參量，即結構參量。

A2 結構參量

數學中廣泛地使用有結構的量，比如向量、矩陣、集合、函數等。作爲個體時它們能作爲常量、變數與參量出現。使用這種高階的抽象物件能使思維簡潔而清晰，有利於把握物件間的主要聯繫。在情境語義學中已經引入了類似向量、集合與函數等高階量及相應的型式。據此可以定義出相應的有這些結構的常量、變數與參量。本文再介紹兩種結構參量：複合參量與條件參量，以及相應的型式。

定義 A2.1 複合參量是含有參量的參量，即含參量的物件當作(有結構的)參量。

我們只討論形如q(…)的含參物件。其中物件名q是個標識頭(函數符)，參量列“…”是型式爲 $T=T_1\times T_2\times\ldots\times T_n$ 的（向量）參量。而含參物件 q(…)當作參量時（即出現在型式抽象式或在別的物件中作參量用時）就稱作複合參量，其標識頭 q 可看作是個函數參量。作爲參量，其專屬型式爲 $T\to T_0$.此時複合參量 q(…)[35]的型式爲$(T\to T_0)\cdot T=T_0$，此處 $T_0=[q(\cdots)]$是個已知型式。

所謂條件參量，顧名思義，是指滿足一定附加條件的參量。意即，該參量除了滿足它自身原來的結構要求外，還另外要滿足一些別的條件，通常是用情境運算式（下篇文章要介紹）陳述的條件。在我們的體系中，主要定義了三種條件參量：

定義 A2.2

(1) 設 x 是個參量，Σ 是個情境運算式。我們用 x|Σ 表示一個條件參量，叫 B-條件參量[36]。相應的型式記作[x|Σ]

(2) 設 x 是個參量，σ 是個情境（參量），I 是個資訊運算式[37]。我們稱運算式 x|σ⊨I 與 σ|σ⊨I 分別爲 1 型與 2 型 D-條件參量。相應的型式記作[x|σ⊨I]與[σ|σ⊨I]。

(3) C-條件參量形如 x|(Σ,I)（各符號含義如上。）如果 I=Σ=∅，就得到無條件參量 x；如果單有 I=∅，就得到 B-條件參量 x|Σ；如果單 Σ=∅ 且用在情境 σ 中時，就得到 1 型 D-條件參量 x|σ⊨I（2 型的 D-條件參量要另行定義，不過也很容易，不贅述。）

B-條件參量的條件是用情境定義的，屬於抽象的情境世界。而 D-條件參量的條件是用命題定義的，屬於描述的世界。 B- 與 D-條件參量表現形式相似，實質也一樣。（可以證明實用上它們是等效的。）根據實際需要，我們這裏引入了一個新形式，叫 C-條件參量，取兩者之間之意。實質上仍一樣。

---

[35] 後面會看到，我們把它處理成一個函數映射的結果。這裏的標識頭就作函數識別字用。

[36] B-條件參量與 D-條件參量分別指根據 Barwise[Barwise,J.,*et al*. 1983(1999)]與 Devlin[Devlin, K.,1991]做出的條件參量定義。

[37] 由資訊元做成的運算式叫資訊運算式，其定義將在下篇文章中給出。在此可理解爲一組資訊元表述的條件。

條件參量往往有一個適用時空場合,特別是適用時段(生命期),可用 with 短語來陳述。

最簡單的情況是用一個情境表述的條件。對此要區分兩種情況。一種是所定義的參量（概念）只在該情境中有效。這樣的條件參量叫角色，意指在該情境中當任的語義角色。

定義 A2.3　設 σ 是一個情境，　x 是個參量。我們用 x|⟨σ⟩ 即一種特殊的條件參量，叫做 x 在 σ 中的角色，簡稱做角色，其型式記作[x|⟨σ⟩] ．角色的特殊處在，它的有效時空就是該情境的所在時空。

另一種是，所定義的參量（概念）可能越出該情境所在時空（目前尚未發現這樣的例子。），其有效時空需用 with 短語另附說明，與其他條件參量一樣。

A3 錨定函數[38]

結構參量的語義主要體現在它被替換成常量時的要求上。爲了幫助理解，先給出錨定的定義。

定義 A3.1　錨定函數是一種將參量映射成常量的函數。分下面四種情況敍述：

(1) 最基本的錨定是對一個簡單參量進行的。設 a 是個專屬型式爲 T 的參量，σ 是個世界，x 是 σ 中的常量，滿足 x:T. 如果我們（在 σ 中）把 a 替換成 x，就說 a 在 σ 中被錨定爲 x 了。如果用函數 g 標記這個對應 a↦x，則可將這個錨定過程記作 a[g]=x.

(2) 複合參量的錨定，分爲以下兩種情況：

當含參物件 α(β)當作（有結構的）複合參量用時，我們把它記成 α′(β)以與含參物件相區分。如果 σ 是個世界，α′(β)在 σ 中的錨定就是指在 σ 中用某個常量 x 來替換它。意即用上述的錨定函數 g 把 α′(β)映成 α′(β)[g]=x。易見，此時應有 x :[α′(β)]。即常量物件 x 屬於參量 α′(β)專屬的那個型式。後一句話的意思是：應當有常量 u 與 v，　使 x=u(v)，且有 u :[α]，v:[β]。這是 α(β)的整體錨定情況。

設 α(β)只是個含參物件，即不當作複合參量。 α(β)中出現的參量集合記做 V(α)。設 σ 是個世界。g 是一個錨定函數。如果 α(β)在 σ 中是有定義的（即 α 與 β 在 σ 中是存在的）。 我們用 g 把 α(β)中的相應參量映成相應的常量，所得的結果物件記作 α[g]。此時應假定這些參量被 g 映成的結果常量在 σ 中是存在的[39]，且滿足相應參量在 α(β)中的型式的要求。需要注意的是，這裏我們並未要求 α(β)的所有參量都被 g 映成常量（即 V(α) ⊆dom(g)），甚至未要求有參量被映成常量（即 V(α) ∩ dom(g) ≠ ∅）。如果 α 的所

---

[38] 錨定函數是情境理論提出來的。情境語義學[Barwise,J.,*et al.* 1983(1999)][Barwise,J. 1987][Devlin, K.,1991]等中稱之爲 anchor, anchoring，主要對條件參量、角色定義的。我們把它簡單地延伸到一般結構參量，並譯爲“錨定”。

[39] 這不是十分明確的說法。但在實際使用過程中，涉及到話語情境、背景情境、與話語所描述的情境等等時，是可以給出更明確的表述的。大體上是這樣：上述諸情境歸屬於一定世界，在該世界中某物件有定義，某些常量存在等。

有參量都被錨定了，即 $V(\alpha) \subseteq dom(g)$，我們也說 $\alpha(\beta)$被錨定了，稱 g 是 $\alpha(\beta)$的錨定函數。此時所得的結果是個常量物件 $\alpha(\beta)[g]$.易見，這是上述複合參量錨定的一種特例。即有 $u=\alpha$，$x=\alpha[g]$是型式$[\alpha(\beta)]$中的常量。如果 $V(\alpha)$中有參量，但不是所有參量，被映成了常量（即 $V(\alpha) \cap dom(g) \neq \varnothing$），我們就說 $\alpha(\beta)$被部分錨定了。

(3) 條件參量的錨定，分爲以下三種情況：

A. 設 $x|\Sigma$ 是個 B-條件參量，τ 是個世界。函數 g 映射 x 成 τ 中的常量（有與 x 相同的結構）。如果 g 能擴張成 Σ 在 τ 中的錨定函數 f，且使 $\tau \vDash \Sigma[f]$成立，意即常情境運算式 $\Sigma[f]$在 τ 中是真實的，就稱 g 是 B-條件參量 $x|\Sigma$ 在 τ 上的錨定函數。 映射結果記作 $(x|\Sigma)[g]$.當 Σ 只是個情境 σ 時，情況一樣。此時 $\tau \vDash \sigma[f]$表示常情境 $\sigma[f]$在情境 τ 中是真實的，即在情境 τ 中是真實存在的。

B. 設 $x|\sigma \vDash I$ 與 $\sigma|\sigma \vDash I$ 分別爲 1 型與 2 型 D-條件參量。τ 是某個世界（通常是個情境）設 h, g 分別是映射 x 與 σ（視作參量!）爲 τ 中常量的函數。如果 g 能擴張成 σ 與 I 在 τ 中的錨定函數 e（即既是 σ 的，也是 I 的錨定函數），且在 τ 中命題 $\sigma[e] \vDash I[e]$成立，我們稱 g 是 1 型 D-條件參量 $x|\sigma \vDash I$ 的錨定函數。錨定結果記作$(x|\sigma \vDash I)[g]$.類似地，如果 h 能擴張成 I 在 τ 中的錨定函數 f，且在 τ 中命題 $\sigma[f] \vDash I[f]$成立，我們稱 h 是 2 型 D-條件參量 $\sigma|\sigma \vDash I$ 的錨定函數。錨定結果記作$(\sigma|\sigma \vDash I)[h]$。

C. 對 C-條件參量 $x|(\Sigma,I)$，映射函數 g 是它在情境 τ 中的一個錨定函數的條件是，g 爲映射 x 爲 τ 中常量的函數，且能擴張成 Σ 與 I 在 τ 中的錨定函數 f，同時還滿足 $\tau \vDash \Sigma[f]$與 $\tau \vDash I[f]$。

(4) 當結構參量成爲別的物件的參量時上述的“參量提升”轉換仍可進行。這種提升可以無限制地遞迴進行下去。

由此可見，所謂條件參量實際是對該參量的適用錨定函數加以限制。

下面的簡單例子可以幫助理解條件參量與錨定函數的定義。

設 σ 是個世界，f 是一個把條件參量“x|c_簡單交易（l,x,y,m,g）”映射到 σ 中的錨定函數。這句話已暗含情境參量 c_簡單交易（l,x,y,m,g）在世界 σ 中是適定的，意即:在世界 σ 中已有人、貨幣、貨物、擁有權、寧願之類的概念，也已有簡單交易的活動了。設 f 把參量(l,x,y,m,g)映成常量(l_0,x_0,y_0,m_0,g_0).此時應有：$\sigma \vDash$c_簡單交易(l_0,x_0,y_0,m_0,g_0).此式的含義是： σ 中存在常量 c_簡單交易(l_0,x_0,y_0,m_0,g_0).也就是在情境 σ 中存在時空場合 l_0，和兩個人 x_0,y_0，以及貨幣 m_0 與貨物 g_0，它們具有如下的關係： 在 l_0 中 x_0 付給 y_0 貨幣 m_0 並且 y_0 交給 x_0 貨物 g_0.於是 f 把 x 映射成的常量 x_0 就成爲參與上述的經濟活動的行爲主體的一方了。因而具有按定義陳述的他在該項行爲活動中所具有的一切條件、活動、權利與義務等。當該條件參量作爲個體進入別的物件中時，任何一個錨定函數都會將它映射成上述這樣的常量，而不僅僅是映射成一個人。

B.情境描述中的幾個基本量的定義

本節介紹情境描述[40]的有關定義。

B1.情境描述

定義 B1.1 資訊元（infon）[41]是一條資訊單元的陳述。它的基本形式爲《r,l：Loc,i_1,i_2,…,i_n； p》.其中 Loc 是時空場合型式,l：Loc 用來標示該處的 l 是 r 的存在場合,不作通常的個體用。

定義 B1.2 我們稱 $\delta=\{\sigma_1,\sigma_2,\ldots,\sigma_n\}$爲一個最簡單的情境描述，這裏諸 σ_i 是上面定義的若干基本資訊元.

B2.情境定義的量

定義 B2.1（誘導）設 $\alpha= x|(\Sigma,I)$是個條件參量，我們稱 $\gamma x.\gamma[x|(\Sigma,I)]$爲$(\Sigma,I)$誘導的 x 上的關係，或叫 γ 抽象。當 x 是單個參量時也叫做(Σ,I)誘導出 x 上的屬性、性質。

定義 B2.2 設 α 是個參量列。我們稱\α 爲遮罩運算元。其含義是:用它作用在一個含參物件 σ(β)上的結果\α.σ(β)是使參量列 β 中屬於 α 的那些參量在外部不可見了。(但在\α.σ(β)內部就和在 σ(β)中一樣依然是存在的.)

定義.(初步) 設 σ,τ 是兩個情境，e 是個資訊元。如果對任意 δ(最簡)描述總滿足：(1)只要命題 σ⊨δ 成立，就有 τ⊨δ∪{e}；(2)只要 τ⊨δ∪{e},且 e∉cl[δ][42],就有 σ⊨δ.我們稱 τ 是由 σ 增加 e 的結果，並記 τ=σ@e 和 σ=τ®e.稱 σ 是由 τ 刪除 e 的結果。描述 δ∪{e}也記作 δ∪e.

定義 B2.3 設 σ 是個含參物件。a,b 是兩個同型式的量(即存在型式 T,滿足：a:T,b:T.此式也寫成 a,b:T).我們用 σ[b\a]表示用 b 替換 σ 中出現的所有 a 得到的新物件 τ.

定義 (嘗試) 設 σ 是個情境，$(i_1,\cdots,i_k)$是個參量列。設它們含於 σ 之中(意即 σ 的內容牽涉到它們。)我們用 Proj(σ,$(i_1,\cdots,i_k)$)表示 σ 中含有參量列$(i_1,\cdots,i_k)$上的參量的那個部分，叫做 σ 在參量列$(i_1,\cdots,i_k)$上的投影。

定義 B2.4 (描述的投影) 設 $\delta(l,j_1,\cdots,j_n)$是個情境描述。$(i_1,\cdots,i_k)$ 是$(j_1,\cdots,j_n)$的一個子列。我們用 Proj(δ $(l,j_1,\cdots,j_n)$, $(i_1,\cdots,i_k)$)記一個新情境描述,它是由 $\delta(l,j_1,\cdots,j_n)$的定義中刪除那些含有除時空場合及參量 $i_1,\cdots,i_n$ 之外的別的參量的那些資訊元後剩下來的部分內容組成。顯然, Proj(δ $(l,j_1,\cdots,j_n)$, $(i_1,\cdots,i_k)$)是 δ $(l,j_1,\cdots,j_n)$的部分情境,叫做 δ $(l,j_1,\cdots,j_n)$在參量列$(i_1,\cdots,i_k)$上的投影。

定義 B2.5 (描述的視景) 設 $\delta(l,j_1,\cdots,j_n)$是個情境描述。$(i_1,\cdots,i_k)$ 是$(j_1,\cdots,j_n)$的一個子列。我們用 Views($\delta(l,j_1,\cdots,j_n)$, $(i_1,\cdots,i_k)$)記一個新情境描述，它由 $\delta(l,j_1,\cdots,j_n)$的定義中的

---

[40] 情境理論[Barwise,J.,*et al.* 1983(1999)]中把客體世界中的情境（可能的與真實的）叫做事件(event)、事態(soa,state of affairs)、事實(fact)，而把我們叫做情境描述的物件叫做情境。

[41] 資訊元及其基本結構是情境理論的最基礎的概念。我們以後要稍加拓廣。

[42] 我們用 cl[δ]表示有限描述 δ 生成的閉包。參見§4 關於情境描述。

那些直接和間接含有屬於($i_1,\cdots,i_k$)中的參數的資訊元所組成。所謂某個資訊元 e 間接含有某個參數 i 是說，存在一個資訊元列 $e_1,\cdots,e_k$.滿足：e_k=e，且 e_1 含有參數 i,e_{j+1} 含有 e_j 中的參數，(j=1,⋯,k-1)。

定義 B2.6 (嵌入) 設 σ ,τ 是兩個情境，且 τ 是 σ 的部分情境。我們用 τ within σ 表示“處在 σ 中的 τ”，讀作“囿於情境 σ 的情境 τ”.

定義 B2.7 設 δ,ε 是兩個描述，滿足 δ ⊂ ε.我們用 δ within ε 表示囿於描述 ε 的描述 δ.

簡單性質. 設 σ,τ 是兩個情境，設 δ,ε 是兩個描述。滿足 σ 是 τ 的子情境,且 δ ⊂ ε. 顯然 δ within ε 是情境 σ within τ 的一個描述。

定義 B2.8 設 σ ,τ 是兩個情境。我們用 $\sigma \oplus \tau$ 表示它們的聯合體，稱作 σ 和 τ 的半加或聯立。

定義 B2.9 設 δ,ε 是兩個描述。我們用 δ⊕ε 記它們的集合並 δ∪ε 做成的描述.

B3.若干基本轉換

下面的陳述略去對應的情境描述的那部分。

定義 B3.1 符號 ι 叫做個體化運算元。若 e 是個非個體的物件。我們可用 ιe 表示把它轉換成個體，即忽略它的結構，從整體上把握它。

定義 B3.2 諱名運算元 nickname 是一種泛化運算元。當它作用在條件參量 x|Σ 上或是作用在角色 x|⟨σ⟩ 上，就是將 x 在 Σ 或 σ 中的(所扮演的)特定角色用作 x 的稱謂，從而可在其有效範圍之外稱呼它、引用它: nickname(x|Σ), nickname(x|⟨σ⟩).

定義 B3.3.資訊元化與退資訊元化.（略）

定義 B3.4 (過程化運算元 π)若 σ 是個情境,x 是在其中存在的某個個體,e 是其中的某個事件。我們用 π(x, σ)與 π(e, σ) 分別表示將它們看成是在 σ 中，它們各自在其生存期上歷經的過程(叫歷程), 即看成是生存期上的時間的函數,叫做 x(或 e)的歷程。

定義 B3.5 (過程化運算元 π)設 x 是個個體,e 是個事件。我們用 πx 與 πe 分別表示將它們看成是它們各自在其生存期上歷經的過程(叫歷程), 即看成是生存期上的時間的函數,叫做 x(或 e)的歷程。

定義 B3.6 (分階段描述) 設 x 爲任意事物。若把 x 的生存期 t 分割成若干時段 $t_1,t_2,\cdots,t_k$。 設 $St_1,St_2, \cdots ,St_k$ 分別是 x 在這些時段上的狀態。則我們把事態描述〈t_1:St_1〉〈t_2:St_2〉⋯〈t_k:St_k〉叫做對 x 的歷程的階段的一個描述,簡稱做 x 的分階段描述,如果對任意 j 而言, 〈t_j:St_j〉都是對 x 在時段 t_j 上的情況的描述。每個〈t_j:St_j〉就叫做 x 的一個階段,(j=1,2,⋯,k)

定義 B3.7 設 x 是個個體。我們用 ρx 表示 x 在其生命歷程中所扮演的角色的經歷。

定義 B3.8 設 e 是個情境。我們用 μe 表示 e 在其生存期（存在時段）中經歷的演變

定義 B3.9(“引發”與”終止”)設 e_1,e_2 是兩個情境(事件)。如果情境(事件)e_1 的完成就意味著情境(事件)e_2 的開始，我們說情境(事件)e_1 引發了情境(事件)e_2。記做《initiation,e_1,e_2;1》。如果 t_1,t_2 分別是 e_1 與 e_2 的生存時間，則有《meet$_l$, t_1,t_2;1》。意即 e_2 緊接在 e_1 之後發生。

類似的，如果情境(事件)e_1 的完成就意味著情境(事件)e_2 的完成，我們說情境(事件)e_1 終止了情境(事件)e_2。記做《termination,e_1,e_2;1》。 如果 t_1,t_2 分別是 e_1 與 e_2 的生存時間,此時有《=$_e$,t_1,t_2;1》.意即 t_2 與 t_1 終端相同,也即,e_2 與 e_1 同時結束。

定義 B3.10 (可認識的)事物。(有關的拓撲結構很有特色,頗值得注意。)(暫缺)

Computational Linguistics and Chinese Language Processing
Vol. 7 , No. 2, August 2002, pp. 37-58

A Study on Word Similarity using Context Vector Models

Keh-Jiann Chen*, Jia-Ming You*

Abstract

There is a need to measure word similarity when processing natural languages, especially when using generalization, classification, or example-based approaches. Usually, measures of similarity between two words are defined according to the distance between their semantic classes in a semantic taxonomy. The taxonomy approaches are more or less semantic-based that do not consider syntactic similarities. However, in real applications, both semantic and syntactic similarities are required and weighted differently. Word similarity based on context vectors is a mixture of syntactic and semantic similarities.

In this paper, we propose using only syntactic related co-occurrences as context vectors and adopt information theoretic models to solve the problems of data sparseness and characteristic precision. The probabilistic distribution of co-occurrence context features is derived by parsing the contextual environment of each word, and all the context features are adjusted according to their IDF (inverse document frequency) values. The agglomerative clustering algorithm is applied to group similar words according to their similarity values. It turns out that words with similar syntactic categories and semantic classes are grouped together.

1. Introduction

It is well known that word-sense is defined by a word's co-occurrence context. The context vectors of a word are defined as the probabilistic distributions of its left and right co-occurrence contexts. Conventionally, the similarity between two context vectors is measured based on their cosine distance [Alshawi and Cater, 1994; Grishman and Sterling, 1994; Pereira *et al*., 1993; Ruge, 1992; Salton, 1989]. However, the conventional measurement

* Institute of Information Science, Academia Sinica
E-mail: kchen@iis.sinica.edu.tw; swimming@hp.iis.sinica.edu.tw

suffers from the following drawbacks. First of all, the information in the context vectors is vague. All co-occurrence words are collected without distinguishing whether they are syntactically or semantically related. Second, the coordinates are not pair-wise independent (i.e., the axes are not orthogonal), and it is hard to apply singular value decomposition to find the orthogonal vectors [Schutze, 1992]. In this paper, we propose to use only syntactic related co-occurrences as context vectors [Dekang Lin, 1998] and adopt information theoretic models to solve the above problems. In our study, the context vectors of a word are defined as the probabilistic distributions of its thematic roles and left/right co-occurrence semantic classes. The context features are derived from a treebank. All context features are weighted according to their TF × IDF values (the product of the term frequency and inverse document frequency) [Salton, 1989]. For the context features, the Cilin semantic classes (a Chinese thesaurus) are adopted. The Cilin semantic classes are divided into 4 different levels of granularity. In order to cope with the data sparseness problem, the weighted average of the similarity values at four different levels will be the similarity measure of two words. The weight for each level is equal to the information-content of that level[Shannon, 1948; Manning and Schutze 1999].

A agglomerative clustering algorithm is applied to group similar words according to the above defined similarity measure. Obviously, words with similar behavior in the corpus will be grouped together. We have compared the clustering results to the Cilin classifications. It turns out that words in the same synonym class and with the same syntactic categories have higher similarity values than the words with different syntactic categories.

2. Data Resources

Ideally, to derive context vectors, a large corpus with semantic tags is required. Furthermore, to extract co-occurrence words along with their exact syntactic and semantic relations, the corpus structure has to be annotated. However, such an ideal corpus does not exist. Therefore, in this paper we will adopt the resources that are available and try to derive a useful but imperfect Chinese tree bank. Since the similarity measure based on the vector space model is a rough estimation, minor errors made at the stage of context vector extraction are acceptable.

2.1 Sinica Corpus

The Sinica corpus contains 12,532 documents and nearly 5 million words. Each sentence in the corpus was parsed by a rule parser [Chen, 1996]. The parsed trees were tagged with the structure brackets, syntactic categories and thematic roles of each constituent [Huang *et al*., 2000] as exemplified below, (Sinica corpus: http://www.sinica.edu.tw/ftms-bin/kiwi.sh):

Original sentence: 小 'small' 狗 'dog' 跳舞 'dance'

Parsed tree : S(Agent:NP:(Property:Adj:'小 small' | Head: N: '狗 dog')|Head:V: 跳舞 dance')

Although these labels may not be exactly correct, we believe that, even with these minor errors, the majority of word-to-word relations extracted from the trees are correct. However, the semantic label is not provided for each word in the parsed trees. In this paper, we will use Cilin classifications for semantic labeling.

2.2 Cilin- a Chinese Thesaurus

Cilin provides the Mandarin synonym sets in a hierarchical structure [Mei *et. al.*, 1984]. It contains 51,708 Chinese words, and 3918 classes. There are five levels in the Cilin semantic hierarchy, denoted in the format L_1-L_2-L_3-L_4-L_5. For example, the Cilin class of the word 我們 'we' is "A-a-02-2-01". In level 1, "A", denotes the semantic class of human; in level 2, "a", indicates a group of general terms; level 3, "02", means pronouns in the first person, and in level 4, "2" represents the plural property. In level 5, "01" represents the order rank in the level 4 group. This means that "01" in level 5 is the first prototypical concept representation of "A-a-02-2". In the rest of this paper, only the first four levels will be used. The fifth level is for sense disambiguation only (section 2.3).

2.3 Sense Disambiguation

A polysemous word has more than one Cilin semantic class. In order to tag appropriate Cilin classes, we have designed a simple sense tagging method as follows [Wilks, 1999]. The sense tagging algorithm is based on the facts that the syntactic categories of each word in the tree bank are assigned uniquely, and that each Cilin class has its own major syntactic category. If a word has multiple Cilin classes, we select the sense class whose major syntactic category is the same as the tagged category of this word. For example, 計畫 "Jie-Hwa" has two meanings. One is for "project" as a noun and the other is "attempt", therefore, if 計畫 "Jie-Hwa" was tagged with a noun category, we will assign the Cilin class whose major category is "project". Sense ambiguity can be distinguished by measure of syntactic properties for most words. However, there are still cases in which the syntactic category constraints cannot resolve the sense ambiguities. Then, we simply choose the prototypical sense class, i.e, the word that has the highest rank in this sense class with respect to all its sense classes in Cilin.

2.4 The Extraction of Co-occurrence Data

The extracted syntactically related pairs have either a head-modifier relation or head-argument relation. For instance, two syntactically related pairs extracted from the example in section 2.1 are:

(<Thematic role > <Cilin> <word1>), (<Thematic role> <Cilin> <word2>)

(agent Bi-07-2 狗'dog'), (Head(S) Hh-04-2 跳舞'dance')

(property Ea-03-3 小'small'), (Head(NP) Bi-07-2 狗'dog')

The context data of the word$_1$ 狗 "dog" consists of its thematic role "agent" and the Cilin class "B-i-07-2" ; the word$_2$ 跳舞 'dance' consists the thematic role "Head(S)" and the Cilin class "H-h-04-2" and so on. The word 小 "small" and 跳舞 "dance" are not syntactically related even though they co-occur. Therefore , they will not be extracted.

3. Context Vector Model

There are three context vectors of a word: role vector, left context vector and right context vector. The role vector is a fixed 48-dimension vector, and each dimension value is equal to the probabilistic distribution of its thematic roles. The left/right context vectors are closer to the probabilistic distributions of its left/right co-occurrence words and their semantic classes. The role vector characterizes a word based more on syntax and less on semantics, but the left/right context vectors are just the opposite. The cosine distance between their context vectors is a measure of the similarity of the two words. We will illustrate the derivations of context vectors and their similarity rating with a simplified example using (貓 "cat", 狗 "dog"). The role vector of "dog" is $\{127, 207, 169\ldots, 0\}_{48}$, which represents the values of "agent", "goal", "theme"… and "topic" respectively, generated by Equation (1). The role vector of "cat" is $\{28, 73, 56\ldots, 0\}_{48}$, which is also acquired by Equation (1).

Role vector of word $W = \{ V_1, V_2,\ldots,V_{48} \}_{48}$

$$Vi = \text{Frequency}(R_i) \times \log(1/P_i) \quad (1)$$

R_i: We label thematic roles "agent", "goal"… and "topic" from R_1 to R_{48} listed in Table2 in the Appendix.

Frequency (R_i): The frequency of R_i played by word W in the corpus.

P_i: = Total frequency of R_i in the corpus / Total frequency of all roles in the corpus.

$\log(1/P_i)$: The information–context of R_i [Shannon, 1948; Manning and Schutze, 1999]

The derivation of left/right context vectors is a bit more complicated. The syntactically related co-occurrence word pairs are derived first as illustrated in section 2.4. We will illustrate the derivation of the left context vector only. The right context vector can be derived similarly. The left co-occurrence word vector of word W is generated from frequency(word$_i$), where word$_i$ precedes and is syntactically related to W in the corpus. Due to the data sparseness problem, the feature dimensions of context vectors are generalized into Cilin classes instead of co-occurrence words. The generalization process reduces the effect of data sparseness. On the other hand, it also reduces the precision of characterization since each

word has different information content and two words that have the same co-occurrence semantic classes may not share the same co-occurrence words. In order to resolve the above dilemma, when we compare the similarity between $word_X$ and $word_Y$, 4 levels of left context vectors and right context vectors for $word_X$ and $word_Y$ are created[1]. The weighting of each feature dimension is adjusted using the TF*IDF value if $word_X$ and $word_Y$ have shared context words. Equation (2) illustrates the creation of the 4th level left context vector of $word_X$. The other context vectors for $word_X$ and $word_Y$ are created by a similar way.

Left context vector of $word_X = \{f1, f2, \ldots, f_{3918}\}_{L4}$

Where fi = Sum of

$\{ TF(word_j) \times IDF(word_j)$ if $word_X$ has the same neighbor $word_j$ with $word_Y$

$TF(word_j)$ if $word_X$ does not have the same neighbor $word_j$ with $word_Y$

} ; for every left co-occurrence $word_j$ with the *i*th Cilin semantic class. (2)

$TF(word_j)$: *the frequency of pair ($word_j$, Cilin($word_j$)) in wordx's co-occurrence context.*

$IDF(word_j)$: *-log(the number of the documents that contains the $word_j$/total document number of the corpus)*

In Equation (2), we adjust the term weight using TF × IDF, which is commonly used in the field of information retrieval [Salton, 1989] to adjust the discrimination power of each feature dimension. We will examine the difference in the adjustment of weights using TF × IDF and TF in section 4.2. We will next give a simplified example. Assume that the word 狗 "dog" has only three left syntactically related words: (小 "small" Ea033) with frequency 30, (可愛 "cute" Ed401) with frequency 5 and (養 "raise" Ib011) with frequency 10; and assume that the word 貓 "cat" has only two left syntactically related words: (黑 "black" Ec043) and (養 "raise" Ib011). Assume that we are measuring the similarity between 狗 "dog" and 貓 "cat". Then, we can compute the left context data of 狗 "dog" as {TF(Aa011),…,TF(Ea033),…,TF(Ed041),…,TF(Ib011) × IDF(養'raise')[2], …, TF(La064) }$_{L4}$[3]

---

[1] IDF(養'raise') = 4.188, the IDF values of all words range from 0.19 to 9.12.

[2] The granularities of the 4 levels of semantic classes are partially shown in Figure2. The four left context vectors and their dimensions are shown below and the right context vectors are similarly derived.

<LeftCilin1>$_{L1}$ A vector of 12 dimensions from "A" to "L".

<LeftCilin2>$_{L2}$ A vector of 94 dimensions from "Aa" to "La".

<LeftCilin3>$_{L3}$ A vector of 1428 dimensions from "Aa01" to "La06".

<LeftCilin4>$_{L4}$ A vector of 3918 dimensions from "Aa011" to "La064".

$= \{0_{Aa011},\ldots,30_{Ea033},\ldots,5_{Ed041},\ldots,10 \times 4.188_{Ib011},\ldots 0_{La064}\}_{L4}$

$= \{0_{Aa011},\ldots,30_{Ea033},\ldots 5_{Ed041},\ldots,41.88_{Ib011},\ldots 0_{La064}\}_{L4}$, since ,they share only the same left context word 養 "raise". The other levels of left context vectors of 狗 "dog" are $\{0_{Aa01},\ldots,30_{Ea03},\ldots 5_{Ed04},\ldots,41.88_{Ib01},\ldots 0_{La06}\}_{L3}$, $\{0_{Aa},\ldots,30_{Ea},\ldots 5_{Ed},\ldots,41.88_{Ib},\ldots 0_{La}\}_{L2}$, $\{0_{A},\ldots,35_{E},\ldots,41.88_{I},\ldots 0_{L}\}_{L1}$. The value of the E dimension is 35 because it is the sum of the values of Ea(30) and Ed(5) from $\{\ldots,30_{Ea},\ldots 5_{Ed},\ldots\}_{L2}$.The right context vectors of $<\text{RightCilin1}>_{R1}$ to $<\text{RightCilin4}>_{R4}$ are derived in a similar way.

3.1 Similarities between Two Context Vectors

Once we know the feature vectors of these two words, we can calculate the cosine distance of two vectors as shown in Equation (3).

vector A= $<a1,a2\ldots,an>$,vector B= $<b1,b2\ldots,bn>$

$$\cos(A,B) = \frac{\sum_{i=1}^{n} ai \times bi}{\sqrt{\sum_{i=1}^{n} ai^2} \times \sqrt{\sum_{i=1}^{n} bi^2}} \quad \ldots \qquad (3)$$

Therefore, the similarity of the two words x and y can be calculated as the linear combination of the cosine distances of all the feature vectors as shown in the Equation 4. The weight of each feature vector can be adjusted according to different requirements. For instance, if the syntactic similarity is more important, we can increase the weight w. On the other hand, if the semantic similarity is more important, the weights w1 to w4 can be increased. If more training data is available, the level 4 vector will be more reliable. Hence, the weight w4 should increase.

$$\begin{aligned}
&\text{similarity}(x, y) = w \times \cos(<\text{role vector}> x, <\text{role vector}> y) \\
&+ w1 \times \{ w11 \cos(<\text{LeftCilin1} > x, <\text{LeftCilin1} > y) + w12 \cos(<\text{RightCilin 1} > x, <\text{RightCilin 1} > y)\} \\
&+ w2 \times \{ w21 \cos(<\text{LeftCilin2} > x, <\text{LeftCilin2} > y) + w22 \cos(<\text{RightCilin 2} > x, <\text{RightCilin 2} > y)\} \\
&+ w3 \times \{ w31 \cos(<\text{LeftCilin3} > x, <\text{LeftCilin3} > y) + w32 \cos(<\text{RightCilin 3} > x, <\text{RightCilin 3} > y)\} \\
&+ w4 \times \{ w41 \cos(<\text{LeftCilin4} > x, <\text{LeftCilin4} > y) + w42 \cos(<\text{RightCilin 4} > x, <\text{RightCilin 4} > y)\}
\end{aligned} \qquad (4)$$

$$wk1 = \frac{|< \text{LeftCilin K} > x| + |< \text{LeftCilin K} > y|}{|< \text{LeftCilin K} > x| + |< \text{LeftCilin K} > y| + |< \text{RighttCilin K} > x| + |< \text{RightCilin K} > y|}$$

$$wk2 = \frac{|< \text{RightCilin K} > x| + |< \text{RightCilin K} > y|}{|< \text{LeftCilin K} > x| + |< \text{LeftCilin K} > y| + |< \text{RighttCilin K} > x| + |< \text{RightCilin K} > y|}$$

$k = 1,2,3,4$

$w + w1 + w2 + w3 + w4 = 1$

$|<$ vector $>|$ means the vector length

In the experiments, w = 0.3, w1 = 0.1×0.7, w2 = 0.1×0.7, w3 = 0.4×0.7, and w4 = 0.4×0.7.

4. Similarity Clustering

Because of the lack of objective standards for evaluating of similarity measures, a agglomerative clustering algorithm is applied to group similar words according to a similarity value. It turns out that words with similar syntactic usage and similar semantic classes are grouped together. We will evaluate our algorithm by comparing the automatic clustering results with manual classifications of Cilin.

4.1 Clustering Algorithm

To evaluate the proposed similarity measure, we tried to group words according to various parameters. We adopted bottom-up agglomerative clustering algorithm to group words. In order to compare the clustered results with Cilin classifications and reduce the data sparseness, we picked the 1000 highest frequency words in Cilin for testing. First of all, we produced a 1000×1000 symmetric similarity matrix called SMatrix, where SMatrix (x, y) = similarity ($word_X$, $word_Y$), for all x < y. The rest of the matrix was set to - INFINITY. Below are the details of the clustering algorithm.

Bottom-up Agglomerative Clustering (the greedy algorithm):

Initialize:

Assign the threshold; (a value ranging from 0.1 to 0.85)

Assign each word to its own group named Group(word)

Loop

Find the entry [x,y] of SMatrix with the maximal value and let the value be M = SMatrix[x][y];

If M is less than the threshold

exit loop

else

Grouping ($word_X$, $word_Y$)

Recalculate SMatrix

End Loop

Grouping ($word_X$, $word_Y$)

Merge Group($word_Y$) to Group($word_X$)

Recalculate SMatrix

{ Smatrix (i)(y) = Smatrix (y)(j)= -INFIITE, where j != y, i != y

$$SMatrix[x][i] = \frac{\sum_{p=1}^{m}\sum_{q=1}^{n} similarity(word_p, word_q)}{m \times n}$$

$0 < x < i$

Group (wordx) contains $word_p, 1 \le p \le m$

Group (wordj) contains $word_q, 1 \le q \le n$

$$SMatrix[j][x] = \frac{\sum_{p=1}^{m}\sum_{q=1}^{n} similarity(word_p, word_q)}{m \times n}$$

$0 < j < x$

Group (wordx)contains $word_p, 1 \le p \le m$

Group (wordj) contains $word_q, 1 \le q \le n$

}

4.2 Clustering Results vs Cilin Classification

We will make a comparison between the clustering results and Cilin classifications. There are two simple examples shown in Figure 1 and Figure 2 representing the clustering results and Cilin classifications, respectively.

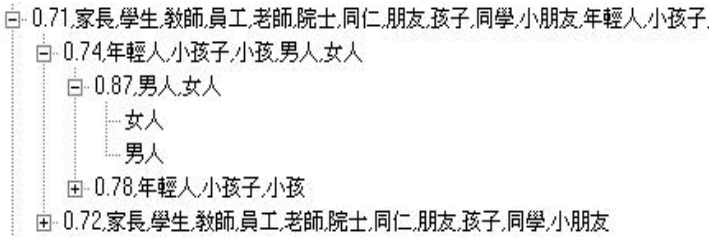


Figure 1 Clustering results with threshold = 0.7

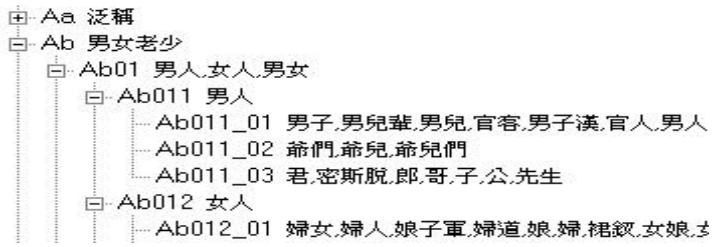


Figure 2 Examples of Cilin classification

After the clustering algorithm is applied, the words are distributed into m groups, i.e., $Group_1$, $Group_2$, $Group_3$…, $Group_m$. Then, we can define the recall and precision of the classification as follows.

$$\text{recall } k = \frac{\sum_{i=1}^{m} Gki}{\text{The number of words that are clustered in the kth level of Cilin}}, (5)$$

$$\text{precision } k = \frac{\sum_{i=1}^{m} Gki}{\text{The number of words that are clustered by this greedy algorithm}}, (6)$$

Gki = representing the maximum number of words in Groupi that are classified in the same Cilin class in level k.

Among our 1000 testing words, the number of words that were clustered in the 4th level of Cilin was 658; i.e., they were labeled with 459 different level-4 Cilin classes and among

them, 342 classes contained only one testing word, and the classes with multiple testing words contained a total of 658 testing words. With the threshold=0.7, our method clustered 830 words, and only 167 words of them were clustered in the correct Cilin class. Therefore, by Equation (5), recall 4 = 167/658 = 0.25 and with Equation (6) precision 4 = 167/830 = 0.20.

We adopted two methods for measuring similarity; one used Equation TF × IDF, and the other used Equation TF. The results are shown in Figure 3 to Figure 6 in the Appendix. We measured the performance by computing the F-score, which is (recall+precision)/2. We discovered that the best F-score of level1 was that 0.7648 located at a threshold equal to 0.65, the best F-score of level2 was that 0.5178 located at a threshold equal to 0.7, the best F-score of level3 was that 0.3165 located at a threshold equal to 0.8, and that the best F-score of level4 was that 0.2476 located at a the threshold equal to 0.8. All were obtained using TF × IDF strategy. Hence, we can see that the TF × IDF equation achieves better performance than the TF equation does. We list the detailed F-socre data for various parameters in appendix. Although the clustering results didn't fit Cilin completely, they are still alike to some degree. From the results, we find that they are similar to syntax taxonomy under a lower threshold and close to semantic taxonomy under higher thresholds.

5. Cilin classifications re-examined

To examine the practicability of our proposed method, we also inspected the similarity values of these 658 testing words which were clustered into 117 4th level Cilin classes. For each semantic class, the average similarity between words in the class and their standard deviation was computed. The results are listed in Table1 in the Appendix. We expected that synonyms would have high similarity values, but this was not always the case.

According to the assumption noted above, synonyms might have similar syntactic and semantic contexts in language use. Therefore, the average similarity should be pretty high, and the standard deviation should be quite low. However, some of the results didn't follow the assumption. We analyzed the data offer explanations in the following.

a) Synonyms with different POS: Words with the same semantic classification in Cilin could have different parts of speech (POS). (as shown below.)

Word set	Average similarity	Stand deviation
思想,考量 考慮,思考	0.544195	0.229534

The contexts of the noun (思想, "thinking") and the verbs (考慮,考量,思考, "think", "consider", "deliberate") were quite different. As a result, the average

similarity value was quite low, and the standard deviation was very high. After we removed the noun from the word-set, we recomputed the values and obtained the table shown below:

Word set	Average similarity	Stand deviation
考量,考慮 思考	0.770616	0.0429353

The results conform to our assumption. They also reveal that the context of synonyms may vary from POS to POS.

b) Error in Cilin Classification: The classifications in Cilin could be arbitrary. For example, the three words, 數量 "quantity", 多少 "how many" and 人數 "the number of people", were classified in a Cilin group. They might be slightly related, but grouping them together seems inappropriate according to the following table:

Word set	Average similarity	Stand deviation
數量,多少 人數	0.379825	0.253895

c) Different uses: Differences in their usage cause synonyms to behave differently. For example, when we measured the similarity of 美國 "America" to 日本 "Japan" and to 中國 "China", the results we obtained were 0.86 and 0.62,respectively, for each pair. Accroding to human intuition, they simply refer to names of countries and should not have such different similarity values. The reason for these result is that the corpus we adopted is an original Taiwan corpus. As a result, the usage of 中國 "China" is different from that of 美國 "America" and 日本 "Japan".

d) Polysemy: The word senses that Cilin adopted were not those frequently used in the corpus. See the following table:.

Word set	Average similarity	Stand deviation
十分,非常, 特別	0.45054	0.305209

Although the three words, 十分 "very/ten points", 非常 "very" and 特別"special, extraordinary" might seem to be very close in meaning to "very", the polysemous word 特別 "special, extraordinary" is different in its major sense. This influenced the result.

e) Words with similar contexts might not be synonyms: A disadvantage does exist when the context vector model is used. Words that are similar in terms of their contexts might not be similar in meaning. For example, the similarity value of 結婚 "marry" and 長大 "grow" is 0.8139. Although the two words have similar contexts, they are not alike in meaning. Therefore, the vector space model should incorporate the taxonomy approach to solve this phenomenon.

6. Conclusions

In this paper, we have adopted the context vector model to measure word similarity. The following new features have been proposed to enhance the context vector models: a) The weights of features are adjusted by applying TF × IDF. b) Feature vectors are smoothed by using Cilin categories to reduce data sparseness. c) Syntactic and semantic similarity is balanced by using related syntactic contexts only.

The performance of our method might have been influenced by the small scale of the Chinese corpus and accuracy of the extracted relations. Further more, Cilin was published a long time ago and has not been update recently, which may have influenced our results. However, our experimental results are encouraging. They supports the theory that using context vectors to measure similarity is feasible and worthy of further research.

References

Alshawi and Carter (1994) "Training and scaling preference functions for disambiguation." *Computational Linguistics*,20(4):635-648

Chen, K.J. (1996) "A Model for Robust Chinese Parser", *Computational Linguistics and Chinese Language Processing*, Vol. 1, pp.183-204.

Grishman and Sterling (1994) "Generalizing automatically generated selectional patterns." *In Proceeding of COLING-94*, pages 742-747, Kyoto, Japan.

Huang, Chu-ren, F.Y. Chen, Keh-Jiann Chen, Zhao-ming Gao, and Kuang-Yu Chen (2000) " Sinica Treebank: Desigm Criteria, Annotation Guidelines and On-line Interface", *Proceedings of ACL workshop on Chinese Language Processing*, pp.29-37.

Lin, Dekang(1998) "Automatic Retrieval and Clustering of Similar Words" *COLING-ACL98*, Montreal, Canada.

Manning, Christopher D. & Hinrich Schutze (1999) *Foundations of Statistical Natural Language Processing*, the MIT Press, Cambridge, Massachusetts.

Mei, Gia-Chu etc., 1984 同義詞詞林.(Cilin - thesaurus of Chinese words). Hong Kong, 商務印書館香港分館.

Miller, George A., and Walter G. Charles (1991) " Contexture Correlates of Semantic Similarity," *Language and Cognitive Processes* 6:pp.1-28.

Salton, Gerard (1989) Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Reading, MA: Addison Wesley.

Ruge (1992) "Experiments on linguistically based term associations." *Information Processing & Management*, 28(3):317-332

Schutze, Hinrich (1992) " Context Space", In Robert Goldman, Peter Norvig, Eugene Charniak, and Bill Gale (eds.), *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pp. 113-120, Menlo Park, CA. AAAI Press.

Shannon, Claude E. (1948) "A Mathematical Theory of Communication", *Bell System Technical Journal* 27: pp. 39-423, 623-656.

Wilks, Yorick (1999) "Is Word Sense Disambiguation just one more NLP Task?" arXiv: CS. CL/9902030 vl.

Appendix

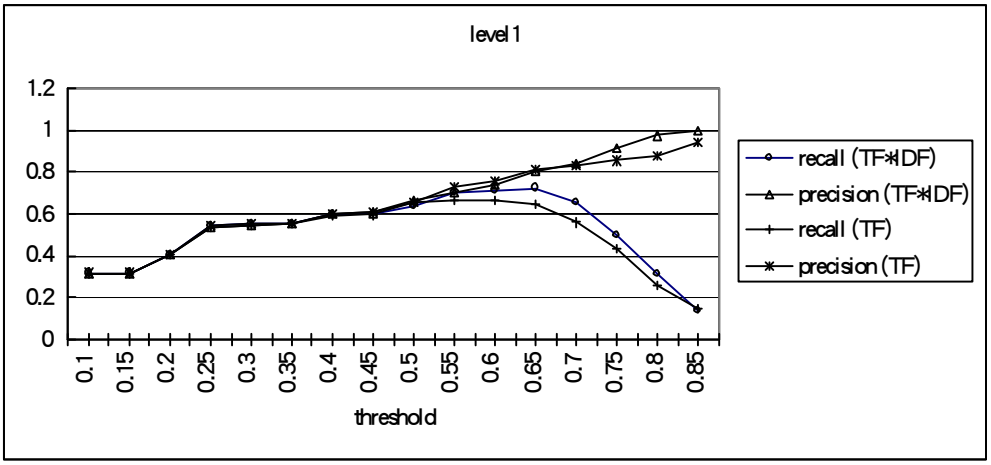


Figure 3 Clustering recall and precision at levels one of Cilin's semantic hierarchy

A partial data of Figure 3

	0.55	0.6	0.65	0.7	0.75	0.8	0.85
recall (TF*IDF)	0.706	0.709	0.725	0.657	0.498	0.313	0.138
precision (TF*IDF)	0.700738	0.736726	0.804756	0.840602	0.909853	0.973881	1
recall (TF)	0.665	0.666	0.643	0.559	0.434	0.261	0.148
precision (TF)	0.727085	0.757479	0.810962	0.829545	0.854202	0.874302	0.938776
F-Score (TF*IDF)	0.703369	0.722863	0.76488	0.748801	0.703927	0.643441	0.569
F-Score (TF)	0.696043	0.71174	0.726981	0.694273	0.644101	0.567651	0.543388

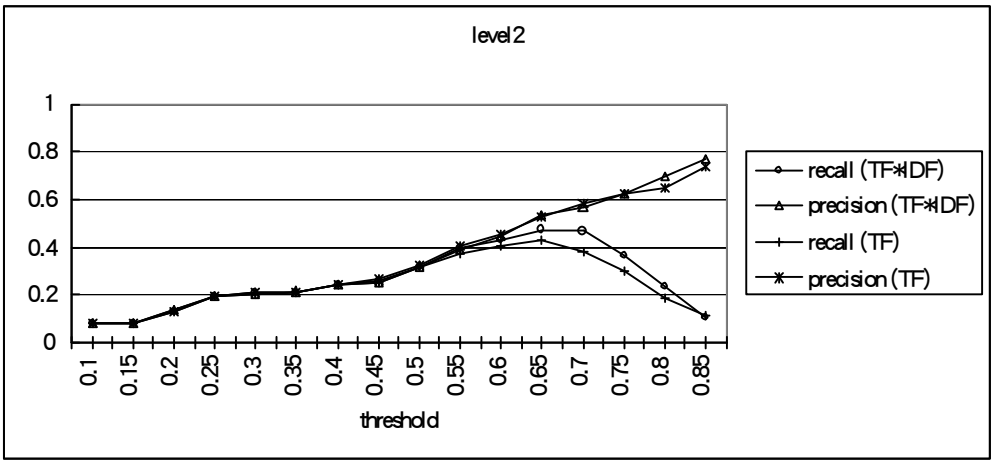


Figure 4 Clustering recall and precision at levels two of Cilin's semantic hierarchy

A partial data of Figure 4

	0.55	0.6	0.65	0.7	0.75	0.8	0.85
recall (TF*IDF)	0.394763	0.429003	0.475327	0.467271	0.367573	0.234642	0.108761
precision (TF*IDF)	0.389884	0.446903	0.53567	0.568421	0.624738	0.697761	0.77027
recall (TF)	0.372608	0.406848	0.431017	0.380665	0.300101	0.188318	0.114804
precision (TF)	0.403708	0.455128	0.527964	0.585859	0.626072	0.650838	0.734694
F-Score (TF*IDF)	0.392324	0.437953	0.505499	0.517846	0.496156	0.466202	0.439516
F-Score (TF)	0.388158	0.430988	0.479491	0.483262	0.463087	0.419578	0.424749

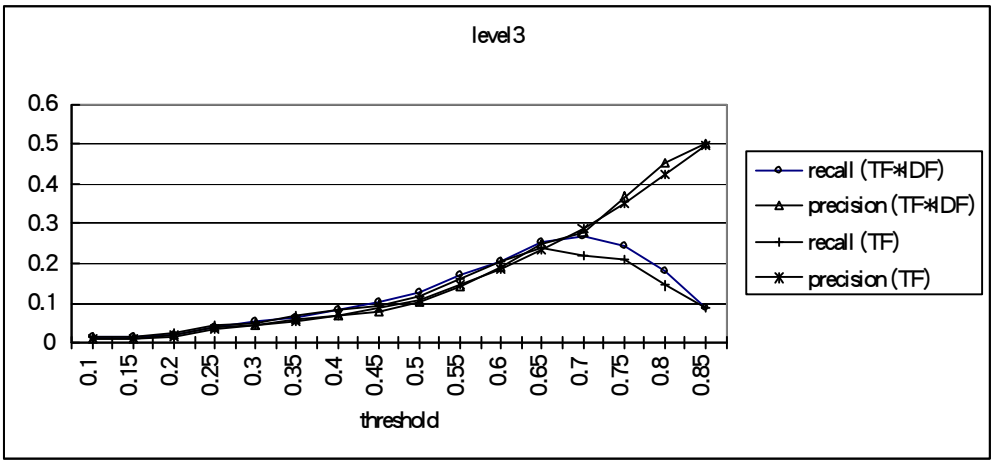


Figure 5 Clustering recall and precision at levels three of Cilin's semantic hierarchy

A partial data of Figure 5

	0.55	0.6	0.65	0.7	0.75	0.8	0.85
recall (TF*IDF)	0.169654	0.205496	0.252091	0.270012	0.243728	0.181601	0.087216
precision (TF*IDF)	0.142255	0.190265	0.249061	0.278195	0.366876	0.451493	0.5
recall (TF)	0.16129	0.205496	0.237754	0.221027	0.20908	0.144564	0.088411
precision (TF)	0.146241	0.183761	0.236018	0.285354	0.349914	0.424581	0.496599
F-Score (TF*IDF)	0.155955	0.197881	0.250576	0.274104	0.305302	0.316547	0.293608
F-Score (TF)	0.153766	0.194629	0.236886	0.253191	0.279497	0.284573	0.292505

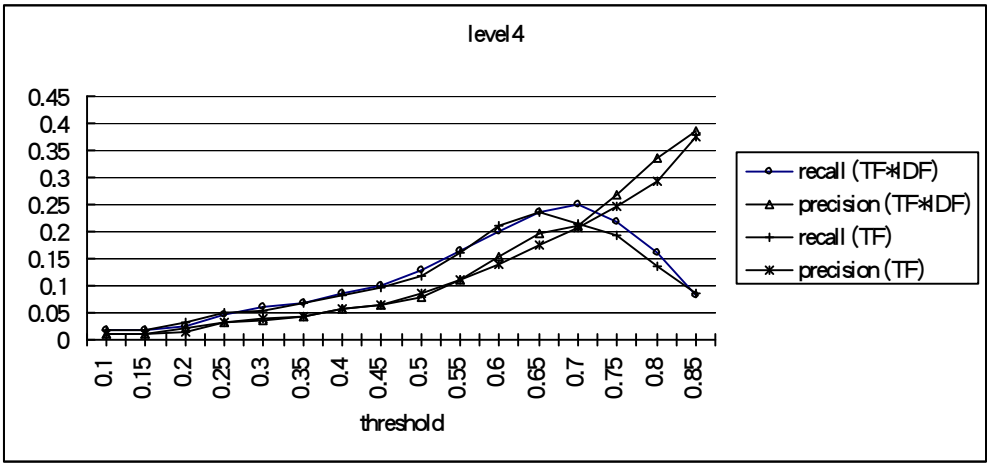


Figure 6 Clustering recall and precision at levels four of Cilin's semantic hierarchy

A partial data of Figure 6

	0.55	0.6	0.65	0.7	0.75	0.8	0.85
recall (TF*IDF)	0.164134	0.200608	0.237082	0.25076	0.218845	0.159574	0.083587
precision (TF*IDF)	0.110643	0.153761	0.195244	0.21203	0.268344	0.335821	0.385135
recall (TF)	0.159574	0.211246	0.237082	0.214286	0.194529	0.136778	0.086626
precision (TF)	0.111226	0.141026	0.174497	0.208333	0.246998	0.293296	0.37415
F-Score (TF*IDF)	0.137389	0.177185	0.216163	0.231395	0.243595	0.247698	0.234361
F-Score (TF)	0.1354	0.176136	0.20579	0.21131	0.220764	0.215037	0.230388

Table 1

GroupId	Word	average	sd
1	首先,第二,第一	0.391749	0.324147
2	狀態,大概,狀況,情形	0.400941	0.307061
3	十分,非常,特別	0.45054	0.305209
4	穩定,一定,固定	0.314694	0.262506
5	大量,太多,那麼,很多,許多	0.409695	0.256464
6	數量,多少,人數	0.379825	0.253895
7	具備,所有,具有,擁有	0.562212	0.24363
8	思想,考量,考慮,思考	0.544195	0.229534
9	大家,人們,民間	0.494836	0.214664
10	檢討,檢查,方案	0.34861	0.20496
11	類似,如同,好像,一樣,一般,接近,似乎	0.295764	0.203913
12	可以,良好,不錯,理想	0.35752	0.198352
13	明白,知道,理解,看出,發現,清楚,了解,意識,掌握	0.584519	0.195064
14	或許,也許,可能,是否	0.631396	0.186908
15	可以,肯定,同意	0.424513	0.186726
16	相同,同時,同樣,一致,一樣	0.432104	0.184799
17	以後,將來,之後,後來	0.525829	0.18295
18	體會,經驗,感受	0.513465	0.179421
19	科技,科學,統計	0.514208	0.173151
20	運動,走向,活動	0.472324	0.17241
21	不易,科學,正確	0.417904	0.17094
22	這樣子,這樣,如此,這麼	0.53811	0.170594
23	自動化,成為,變成	0.586634	0.167871
24	需要,要求,需求	0.584609	0.165922
25	在一起,共同,一起	0.559024	0.16552
26	上課,教學,教授	0.368009	0.164008
27	標準,專業,正式,規範	0.379336	0.163315
28	適合,相當,適當	0.301937	0.163089
29	以前,過去,之前	0.510083	0.161643
30	負責人,院長,主任,家長,領導,經理,主管,校長	0.54561	0.16104
31	透過,通過,經過	0.497587	0.1587
32	自然,當然,本來	0.450807	0.157253
33	基本,基礎,根本	0.303086	0.148239
34	組成,形成,構成,組織	0.558954	0.14715
35	方面,方向,走向	0.425739	0.146068
36	作為,表現,行為	0.500081	0.142414
37	城市,香港,都市	0.523753	0.142381
38	發現,發覺,感受,感覺,感到,覺得	0.587801	0.142334
39	實在,十分,真正	0.438013	0.140809

40	危險,事情,現實,事實,機關,活動,實際,行政,新聞	0.393595	0.139102
41	使用,分享,利用,採用,應用,運用	0.588138	0.132851
42	以為,認為,看看	0.684297	0.130785
43	裡面,期間,之間	0.343657	0.125277
44	博物館,媒體,中心	0.607148	0.124048
45	確定,決定,規定	0.566784	0.123687
46	維持,支持,保持	0.680776	0.123566
47	做法,辦法,方法,方式,措施,藝術,作法	0.619291	0.122881
48	根本,為主,基本,關鍵,重要,重點,主要	0.295473	0.122236
49	視為,當成,當作,當做,作為	0.747904	0.121471
50	可是,只是,但是,然而,不過	0.77375	0.121439
51	地區,地方,所在,社區,區域	0.591823	0.120499
52	消失,失去,沒有	0.632536	0.120176
53	非常,一定,特別,特殊	0.219618	0.119577
54	當中,中央,中心	0.303831	0.11493
55	成就,成功,完成	0.482776	0.112033
56	美國,中國,日本,大陸,台灣,國際,國家,我國,伊拉克,新加坡	0.614028	0.110559
57	其中,內部,裡面	0.498764	0.107996
58	告訴,報導,報告	0.425274	0.107627
59	開放,著手,開始,出發	0.677721	0.106543
60	有的,一般,部分	0.480778	0.100469
61	學會,社團,團體,協會,組織	0.660711	0.100259
62	增加,加上,豐富	0.480646	0.0995899
63	小朋友,兒童,小孩子,小孩,孩子	0.740326	0.0986638
64	一般,通常,平常	0.317419	0.0984778
65	完整,完全,整體	0.265376	0.0977068
66	工業,貿易,行業,企業,商業,交通	0.552969	0.0961678
67	計畫,計劃,設計,規劃	0.717254	0.0956848
68	規劃,設計,計劃,計畫	0.717254	0.0956848
69	系統,雙方,上面,世界	0.419015	0.0940401
70	人員,東西,個人,人口,人士,人物,份子,人類	0.473114	0.0934926
71	支援,幫忙,幫助,補助,協助,支持	0.645829	0.0932524
72	只有,只是,不過	0.698804	0.0926325
73	練習,作業,答案	0.399523	0.0902104
74	處理,安排,因應,從事	0.647893	0.0880783
75	呈現,展現,表現	0.628407	0.0859708
76	生態,動物,生物,植物	0.650257	0.0853656
77	太太,小姐,女性,女兒,女人,婦女,女孩	0.670354	0.0851898
78	前往,過去,走到	0.468615	0.0849678

79	吸引,引發,引起	0.749471	0.0834028
80	回家,回到,回來,回去	0.752641	0.0789966
81	現在,目前,今天	0.729664	0.0781769
82	圖書館,教室,房間,辦公室	0.687431	0.0758984
83	主張,主持,堅持	0.507651	0.0751646
84	委員會,機構,小組,單位,部門,組織	0.682376	0.0736292
85	系統,系列,體系	0.592294	0.0724394
86	期望,想要,願意,希望	0.750708	0.072335
87	高興,喜歡,快樂	0.71814	0.0691603
88	取得,爭取,得到,獲得	0.772821	0.0680632
89	見到,看見,看到,看看	0.789036	0.067304
90	明顯,肯定,明白	0.601009	0.0658739
91	不能,不要,不可	0.716424	0.0655924
92	高中,學院,研究所,大學,學校	0.731575	0.0652677
93	設置,舉辦,設立	0.717764	0.0621858
94	模式,程式,形式	0.669031	0.0601233
95	投入,進入,參加,加入	0.699484	0.0595996
96	作用,力量,功能,意義	0.763711	0.0577985
97	男子,先生,男人	0.632963	0.0562316
98	訓練,練習,教練	0.39034	0.0558793
99	確實,真正,實在	0.712853	0.0544218
100	學期,時期,階段,時代	0.681406	0.0539913
101	預算,計算,統計	0.377922	0.0531331
102	成果,成就,成績	0.646048	0.0524472
103	標準,規則,規範,原則	0.684752	0.050314
104	心態,觀念,概念,理念,思想,心理	0.7056	0.0490335
105	教師,教練,博士,教授,老師	0.709024	0.0489227
106	想法,意思,思想	0.69697	0.0466416
107	怎麼樣,如何,怎麼	0.482305	0.0457521
108	作業,工作,業務	0.741531	0.043475
109	年輕人,青年,青少年,少年	0.716108	0.0425544
110	體制,結構,架構,組織	0.67785	0.0397466
111	自己,本身,自我	0.57837	0.0380475
112	到底,算是,終於	0.111716	0.035438
113	環境,氣氛,條件	0.656953	0.0284353
114	缺乏,不足,緊張	0.594345	0.0275148
115	現場,市場,場所	0.470312	0.0273891
116	人家,兄弟,個人	0.405597	0.023535
117	全部,所有,一切	0.650397	0.0225594

Table2

Role index	
Role ID	Role
R1	agent
R2	apposition
R3	benefactor
R4	causer
R5	CHINESE
R6	companion
R7	comparison
R8	complement
R9	condition
R10	conjunction
R11	degree
R12	deontics
R13	DUMMY
R14	DUMMY1
R15	DUMMY2
R16	duration
R17	epistemics
R18	evaluation
R19	exclusion
R20	experiencer
R21	frequency
R22	goal
R23	Head[GP]
R24	Head[NP]
R25	Head[PP]
R26	Head[S]
R27	Head[VP]
R28	imperative
R29	instrument
R30	interjection
R31	location
R32	manner

Role index	
Role ID	Role
R33	negation
R34	particle
R35	possessor
R36	predication
R37	property
R38	quantifier
R39	quantity
R40	range
R41	reason
R42	receipient
R43	source
R44	standard
R45	target
R46	theme
R47	time
R48	topic

Computational Linguistics and Chinese Language Processing
Vol. 7, No. 2 , August 2002, pp. 59-76

基於《知網》的辭彙語義相似度計算[1]

Word Similarity Computing Based on How-net

劉群[*] 、李素建[+]

Qun LIU , Sujian LI

摘要

詞義相似度計算在很多領域中都有廣泛的應用，例如資訊檢索、資訊抽取、文本分類、詞義排歧、基於實例的機器翻譯等等。詞義相似度計算的兩種基本方法是基於世界知識（Ontology）或某種分類體系（Taxonomy）的方法和基於統計的上下文向量空間模型方法。這兩種方法各有優缺點。

《知網》是一部比較詳盡的語義知識詞典，受到了人們普遍的重視。不過，由於《知網》中對於一個詞的語義採用的是一種多維的知識表示形式，這給詞語相似度的計算帶來了麻煩。這一點與 WordNet 和《同義詞詞林》不同。在 WordNet 和《同義詞詞林》中，所有同類的語義項（WordNet 的 synset 或《同義詞詞林》的詞群）構成一個樹狀結構，要計算語義項之間的距離，只要計算樹狀結構中相應結點的距離即可。而在《知網》中辭彙語義相似度的計算存在以下問題：

1. 每一個詞的語義描述由多個義原組成；
2. 詞語的語義描述中各個義原並不是平等的，它們之間有著複雜的關係，通過一種專門的知識描述語言來表示。

我們的工作主要包括：

1. 研究《知網》中知識描述語言的語法，瞭解其描述一個詞義所用的多個義原之間的關係，區分其在詞語相似度計算中所起的作用；我們採用一種更

[1] 本項研究受國家重點基礎研究計畫（973）支持，項目編號是 G1998030507-4 和 G1998030510。
[*] 北京大學計算語言學研究所 & 中國科學院計算技術研究所 E-mail: liuqun@ict.ac.cn
Institute of Computational Linguistics, Peking University &
Institute of Computing Technology, Chinese Academy of Science
[+] 中國科學院計算技術研究所 E-mail: lisujian@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences

爲結構化的方式改寫了《知網》中詞的定義（DEF），其中採用了“集合”和“特徵結構”這兩種抽象資料結構。

2. 研究了義原的相似度計算方法、集合和特徵結構的相似度計算方法，並在此基礎上提出了利用《知網》進行詞語相似度計算的演算法；

3. 通過實驗驗證該演算法的有效性，並與其他演算法進行比較。

關鍵字：《知網》 辭彙語義相似度計算 自然語言處理

Abstract

Word similarity is broadly used in many applications, such as information retrieval, information extraction, text classification, word sense disambiguation, example-based machine translation, etc. There are two different methods used to compute similarity: one is based on ontology or a semantic taxonomy; the other is based on collocations of words in a corpus.

As a lexical knowledgebase with rich semantic information, How-net has been employed in various researches. Unlike other thesauri, such as WordNet and Tongyici Cilin, in which word similarity is defined based on the distance between words in a semantic taxonomy tree, How-net defines a word in a complicated multi-dimensional knowledge description language. As a result, a series of problems arise in the process of word similarity computation using How-net. The difficulties are outlined below:

1. The description of each word consists of a group of sememes. For example, the Chinese word “暗箱(camera obscura)” is described as: “part|部件, #TakePicture|拍攝, %tool|用具, body|身”, and the Chinese word “寫信(write a letter)” is described as: “write|寫, ContentProduct=letter|信件”;

2. The meaning of a word is not a simple combination of these sememes. Sememes are organized using a specific knowledge description language.

To meet these challenges, our work includes:

1. A study on the How-net knowledge description language. We rewrite the How-net definition of a word in a more structural format, using the abstract data structure of *set* and *feature structure*.

2. A study on the algorithm used to compute word similarity based on How-net. The similarity between sememes, that between *sets*, and that between *feature structures* are given. To compute the similarity between two sememes, we

use the distance between the sememes in the semantic taxonomy, as is done in Wordnet and Tongyici Cilin. To compute the similarity between two *sets* or two *feature structures*, we first establish a one-to-one mapping between the elements of the *sets* or the *feature structures*. Then, the similarity between the *sets* or *feature structures* is defined as the weighted average of the similarity between their elements. For *feature structures*, a one-to-one mapping is established according to the attributes. For *sets*, a one-to-one mapping is established according to the similarity between their elements.

3. Finally, we give experiment results to show the validity of the algorithm and compare them with results obtained using other algorithms. Our results for word similarity agree with people's intuition to a large extent, and they are better than the results of two comparative experiments.

Keywords: How-net, Word Similarity Computing, Natural Language Processing

1. 引言

自然語言的詞語之間有著非常複雜的關係，在實際的應用中，有時需要把這種複雜的關係用一種簡單的數量來度量，而詞義相似度就是其中的一種。

詞義相似度計算在很多領域中都有廣泛的應用，例如資訊檢索、資訊抽取、文本分類、詞義排歧、基於實例的機器翻譯等等[Gauch&Chong 1995，LI, Szpakowicz & Matwin 1995，王斌，1999，李涓子，1999]。本文的研究背景是基於實例的機器翻譯。在基於實例的機器翻譯中，詞語相似度的計算有著重要的作用。例如要翻譯“張三寫的小說”這個短語，通過語料庫檢索得到譯例：

1）李四寫的小說／the novel written by Li Si

2）去年寫的小說／the novel written last year

通過相似度計算我們發現，“張三”和“李四”都是具體的人，語義上非常相似，而“去年”的語義是時間，和“張三”相似度較低，因此我們選用“李四寫的小說”這個實例進行類比翻譯，就可以得到正確的譯文：

the novel written by Zhang San

如果選用後者作爲實例，那麼得到的錯誤譯文將是：

* the novel written Zhang San

通過這個例子可以看出相似度計算在基於實例的機器翻譯中所起的作用。

在基於實例的翻譯中另一個重要的工作是雙語對齊。在雙語對齊過程中要用到兩種語言的詞義相似度計算，這不在本文所考慮的範圍之內。

2. 詞語相似度及其計算的方法

2.1 詞語相似度的含義

詞語相似度是一個主觀性相當強的概念，沒有明確的客觀標準可以衡量。脫離具體的應用去談論詞語相似度，很難得到一個統一的定義。

本文的研究主要以基於實例的機器翻譯爲背景，因此在本文中我們所理解的詞語相似度就是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。兩個詞語，如果在不同的上下文中可以互相替換且不改變文本的句法語義結構的可能性越大，二者的相似度就越高，否則相似度就越低。

相似度這個概念，涉及到詞語的詞法、句法、語義甚至語用等方方面面的特點。其中，對詞語相似度影響最大的應該是詞的語義。

在本文中，相似度被定義爲一個 0 到 1 之間的實數。

詞語距離與詞語相似度之間有著密切的關係。實際上，詞語距離和詞語相似度是一對詞語的相同關係特徵的不同表現形式，二者之間可以建立一種簡單的對應關係。對於兩個詞語 W_1 和 W_2，我們記其相似度爲 $Sim(W_1,W_2)$，其詞語距離爲 $Dis(W_1,W_2)$，那麼我們可以定義一個滿足以上條件的簡單轉換關係：

$$Sim(W_1,W_2)=\frac{\alpha}{Dis(W_1,W_2)+\alpha} \qquad \cdots\cdots(1)$$

其中 α 是一個可調節的參數。α 的含義是：當相似度爲 0.5 時的詞語距離值。

這種轉換關係並不是唯一的，我們這裏只是給出了其中的一種可能。

在很多情況下，直接計算詞語的相似度比較困難，通常可以先計算詞語的距離，然後再轉換成詞語的相似度。

詞語相關性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。詞語相關性和詞語相似性是兩個不同的概念，二者沒有直接的對應關係。

2.2 詞語相似度的計算方法

詞語距離有兩類常見的計算方法，一種是根據某種世界知識（Ontology）或分類體系（Taxonomy）來計算，一種利用大規模的語料庫進行統計。

根據世界知識（Ontology）或分類體系（Taxonomy）計算詞語語義距離的方法，一般是利用一部同義詞詞典（Thesaurus）。一般同義詞詞典都是將所有的詞組織在一棵或幾棵樹狀的層次結構中。我們知道，在一棵樹狀圖中，任何兩個結點之間有且只有一條路徑。於是，這條路徑的長度就可以作爲這兩個概念的語義距離的一種度量。

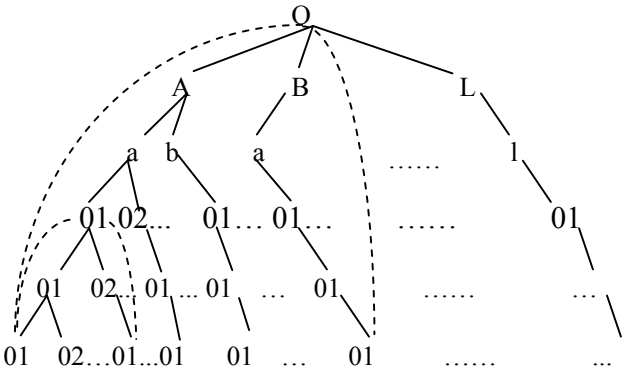


圖 1 《同義詞詞林》語義分類樹狀圖

[王斌，1999]採用這種方法利用《同義詞詞林》來計算漢語詞語之間的相似度（如圖 1 所示）。有些研究者考慮的情況更複雜。[Agirre & Rigau 1995]在利用 Wordnet 計算詞語的語義相似度時，除了結點間的路徑長度外，還考慮到了其他一些因素。例如：

概念層次樹的深度：路徑長度相同的兩個結點，如果位於概念層次的越高層，其語義距離較大；比如說：“動物”和“植物”、“哺乳動物”和“爬行動物”，這兩對概念間的路徑長度都是 2，但前一對詞處於語義樹的較高層，因此認爲其語義距離較大，後一對詞處於語義樹的較低層，其語義距離較小；

概念層次樹的區域密度：路徑長度相同的兩對結點，如果一對位於概念層次樹中低密度區域，另一對位元於高密度區域，那麼前者的語義距離應大於後者。引入區域密度的原因在於，有些概念層次樹中概念描述的粗細程度不均，例如在 Wordnet 中，動植物分類的描述極其詳盡，而有些區域的概念描述又比較粗疏，這會導致語義距離計算的不合理。

另一種詞語相似度的計算方法是用大規模的語料來統計。例如，利用詞語的相關性來計算詞語的相似度。事先選擇一組特徵詞，然後計算這一組特徵詞與每一個詞的相關性（一般用這組特徵詞在實際的大規模語料中在該詞的上下文中出現的頻率來度量），於是，對於每一個詞都可以得到一個相關性的特徵詞向量，然後利用這些向量之間的相似度（一般用向量的夾角餘弦來計算）作爲這兩個詞的相似度。這種做法的假設是，凡是語義相近的詞，他們的上下文也應該相似。[李涓子，1999]利用這種思想來實現語義的自動排歧；[魯松，2001]研究了如何利用詞語的相關性來計算詞語的相似度。[Dagan *et al.* 1995,1999]使用了更爲複雜的概率模型來計算詞語的距離。

這兩種方法各有特點。基於世界知識的方法簡單有效，無需用語料庫進行訓練，也比較直觀，易於理解，但這種方法得到的結果受人的主觀意識影響較大，有時並不能準確反映客觀事實。另外，這種方法比較準確地反映了詞語之間語義方面的相似性和差異，

而對於詞語之間的句法和語用特點考慮得比較少。基於語料庫的方法比較客觀，綜合反映了詞語在句法、語義、語用等方面的相似性和差異。但是，這種方法比較依賴於訓練所用的語料庫，計算量大，計算方法複雜，另外，受資料稀疏和資料雜訊的幹擾較大。

本文主要研究基於《知網（Hownet）》的詞語相似度計算方法，這是一種基於世界知識的方法。

3. 《知網（Hownet）》簡介

按照《知網》的創造者——董振東先生自己的說法[杜飛龍，1999]：

《知網》是一個以漢語和英語的詞語所代表的概念爲描述物件，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常識知識庫。

《知網》中含有豐富的辭彙語義知識和世界知識，爲自然語言處理和機器翻譯等方面的研究提供了寶貴的資源。不過，儘管《知網》提供了詳細的檔案[董振東，董強，1999]，但《知網》檔案的形式化和規範化程度都不高。

本節中，我們將主要通過對《知網》的知識描述語言的分析，利用集合、特徵結構等抽象資料形式，將《知網》的知識描述語言表示成一種更爲直觀、更爲結構化的形式，以便於後面的相似度計算。

3.1 《知網》的結構

《知網》中有兩個主要的概念：“概念”與“義原”。

“概念”是對辭彙語義的一種描述。每一個詞可以表達爲幾個概念。

“概念”是用一種“知識表示語言”來描述的，這種“知識表示語言”所用的“辭彙”叫做“義原”。

“義原”是用於描述一個“概念”的最小意義單位。

與一般的語義詞典[如《同義詞詞林》或 Wordnet]不同，《知網》並不是簡單地將所有的“概念”歸結到一個樹狀的概念層次體系中，而是試圖用一系列的“義原”來對每一個“概念”進行描述。

《知網》一共採用了個 1500 義原，這些義原分爲以下幾個大類：

1) Event|事件

2) entity|實體

3) attribute|屬性值

4) aValue|屬性值

5) quantity|數量

6) qValue|數量值

7) SecondaryFeature|次要特徵

8) syntax|語法

9) EventRole|動態角色

10) EventFeatures|動態屬性

對於這些義原，我們把它們歸爲三組：第一組，包括第 1 到第 7 類的義原，我們稱之爲“**基本義原**”，用來描述單個概念的語義特徵；第二組，只包括第 8 類義原，我們稱之爲“**語法義原**”，用於描述詞語的語法特徵，主要是詞性（Part of Speech）；第三組，包括第 9 和第 10 類的義原，我們稱之爲“**關係義原**”，用於描述概念和概念之間的關係（類似於深層格語法中的格關係）。

除了義原以外，《知網》中還用了一些符號來對概念的語義進行描述，如下表所示：

表 1: 《知網》知識描述語言中的符號及其含義

,	多個屬性之間，表示“和”的關係
#	表示“與其相關”
%	表示“是其部分”
$	表示“可以被該‘V’處置，或是該“V”的受事，物件，領有物，或者內容
*	表示“會‘V’或主要用於‘V’，即施事或工具
+	對 V 類，它表示它所標記的角色是一種隱性的，幾乎在實際語言中不會出現
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“V”的空間或時間
?	表示可以是“N”的材料，如對於布匹，我們標以“?衣服”表示布匹可以是“衣服”的材料
{}	(1) 對於 V 類，置於 [ ] 中的是該類 V 所有的“必備角色”。如對於“購買”類，一旦它發生了，必然會在實際上有如下角色參與：施事，佔有物，來源，工具。儘管在多數情況下，一個句子並不把全部的角色都交代出來 (2) 表示動態角色，如介詞的定義
()	置於其中的應該是一個詞表記，例如，(Chinal中國)
^	表示不存在，或沒有，或不能
!	表示某一屬性爲一種敏感的屬性，例如：“味道”對於“食物”，“高度”對於“山脈”，“溫度” 對於“天象”等
[]	標識概念的共性屬性

我們把這些符號又分爲幾類：一類是用來表示語義描述式之間的邏輯關係，我們稱之爲“**邏輯符號**”，包括以下幾個符號：, ~ ^ ；另一類用來表示概念之間的關係，我們稱之爲“**關係符號**”，包括以下幾個符號： # % $ * + & @ ? ! ；第三類包括幾個無法歸入以上兩類的“**特殊符號**”：{} () [] 。

我們看到，概念之間的關係有兩種表示方式：一種是用“**關係義原**”來表示，一種是用表示概念關係的“**關係符號**”來表示。按照我們的理解，前者類似於一種深層格關係，後者大部分是一種深層格關係的“反關係”，例如“$”我們就可以理解爲“施事、物件、領有、內容”的反關係，也就是說，該詞可以充當另一個詞的“施事、物件、領有、內容”。

義原一方面作爲描述概念的最基本單位，另一方面，義原之間又存在複雜的關係。在《知網》中，一共描述了義原之間的 8 種關係：上下位關係、同義關係、反義關係、對義關係、屬性-宿主關係、部件-整體關係、材料-成品關係、事件-角色關係。可以看出，義原之間組成的是一個複雜的網狀結構，而不是一個單純的樹狀結構。不過，義原關係中最重要的還是上下位關係。根據義原的上下位關係，所有的“基本義原”組成了一個義原層次體系（如圖 2）。這個義原層次體系是一個樹狀結構，這也是我們進行語義相似度計算的基礎。

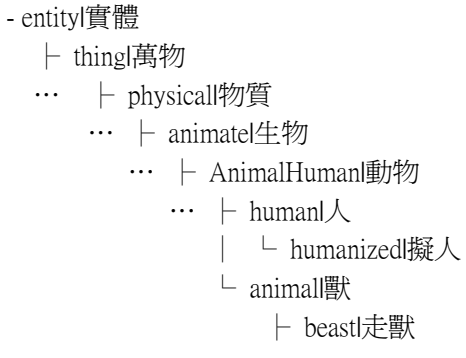


圖 2 樹狀的義原層次結構

雖然《知網》和其他的語義詞典（如《同義詞詞林》和 Wordnet）一樣，也有一個反映知識結構的樹狀層次體系，但實際上有著本質的不同。在《同義詞詞林》和 Wordnet 中，概念是描寫詞義的最小單位，所以，每一個概念都是這個層次體系中的一個結點。而在《知網》中，每一個概念是通過一組義原來表示的，概念本身並不是這個層次體系中的一個結點，義原才是這個層次體系中的一個結點。而且，一個概念並不是簡單的描述爲一個義原的集合，而是要描述爲使用某種專門的“知識描述語言”來表達的一個語義運算式。也就是說，在描述一個概念的多個義原中，每個義原所起到的作用是不同的，這就給我們的相似度計算帶來了很大的困難。下面我們就對這個描述概念的知識描述語

言進行一些考察。

3.2 《知網》的知識描述語言

《知網》通過一種知識描述語言對詞語的語義進行描述。在《知網》的文檔中，對知識描述語言做了詳盡的介紹。不過，由於該文檔過於偏重細節，不易從總體上把握。本節中我們試圖對於這種知識描述語言給出一個簡單的概括。

我們看幾個例子：

表 2：《知網》知識描述語言實例

詞	概念編號	描述語言
打	017144	exercise|鍛練,sport|體育
男人	059349	human|人,family|家,male|男
高興	029542	aValue|屬性值,circumstances|境況,happy|福,desired|良
生日	072280	time|時間,day|日,@ComeToWorld|問世,$congratulate|祝賀
寫信	089834	write|寫,ContentProduct=letter|信件
北京	003815	place|地方,capital|國都,ProperName|專,(China|中國)
愛好者	000363	human|人,*FondOf|喜歡,#WhileAway|消閒
必須	004932	{modality|語氣}
串	015204	NounUnit|名量,&(grape|葡萄),&(key|鑰匙)
從良	016251	cease|停做,content=(prostitution|賣淫)
打對折	017317	subtract|削減,patient=price|價格,commercial|商,(range|幅度=50%)
兒童基金會	024083	part|部件,%institution|機構,politics|政,#young|幼,#fund|資金,(institution|機構=UN|聯合國)

我們將這種知識描述語言歸納爲以下幾條：

1) 《知網》收入的詞語主要歸爲兩類，一類是實詞，一類是虛詞；

2) 虛詞的描述比較簡單，用“**{句法義原}**”或“**{關係義原}**”進行描述；

3) 實詞的描述比較複雜，由一系列用逗號隔開的“語義描述式”組成，這些“語義描述式”又有以下三種形式：

基本義原描述式：用“基本義原”進行描述；

關係義原描述式：用“關係義原=基本義原”或者“關係義原=(具體詞)”或者“(關係義原=具體詞)”來描述；

關係符號描述式：用“關係符號 基本義原”或者“關係符號(具體詞)”加以描述，我們還注意到，可以有多個關係符號描述式採用同一個關係符號；

4) 在實詞的描述中，第一個描述式總是一個**基本義原描述式**，這也是對該實詞最重

要的一個描述式，這個**基本義原**描述了該實詞的最基本的語義特徵。

根據以上分析，我們將《知網》對一個實詞的義項描述重新表示如下：

$$\text{實詞概念}:\begin{bmatrix}\text{第一基本義原描述}=\text{基本義原}_a\\ \text{其他基本義原描述}=\{\text{基本義原}_b,\text{基本義原}_c,\ldots\}\\ \text{關係義原描述}=\begin{bmatrix}\text{關係義原}_1=\text{基本義原}_x\mid\text{具體詞}_x\\ \text{關係義原}_2=\text{基本義原}_y\mid\text{具體詞}_y\\ \ldots\ldots\end{bmatrix}\\ \text{關係符號描述}=\begin{bmatrix}\text{關係符號}_1=\{\text{義原}_u\mid\text{具体詞}_u,\text{義原}_v\mid\text{具体詞}_v,\ldots\}\\ \text{關係符號}_2=\{\text{義原}_s\mid\text{具体詞}_s,\text{義原}_t\mid\text{具体詞}_t,\ldots\}\\ \ldots\ldots\end{bmatrix}\end{bmatrix}$$

在上面的運算式中，“〔……〕”表示特徵結構，“{……}”表示集合，“|”表示“或”。特徵結構和集合是這個運算式中使用的兩種抽象資料結構，也是下面我們進行相似度計算時面對的主要問題。

4. 基於《知網》的語義相似度計算方法

從上面的介紹我們看到，與傳統的語義詞典不同，在《知網》中，並不是將每一個概念對應於一個樹狀概念層次體系中的一個結點，而是通過用一系列的義原，利用某種知識描述語言來描述一個概念。而這些義原通過上下位關係組織成一個樹狀義原層次體系。我們的目標是要找到一種方法，對用這種知識描述語言表示的兩個語義運算式進行相似度計算。

利用《知網》計算語義相似度，一個最簡單的方法就是直接使用詞語語義運算式中的第一基本義原描述式，把詞語相似度等價於第一基本義原的相似度。這種方法好處是計算簡單，但沒有利用知網語義運算式中其他部分豐富的語義資訊。

[Li Sujian *et al.* 2002]中提出了一種詞語語義相似度的計算方法，計算過程綜合利用了《知網》和《同義詞詞林》。在義原相似度的計算過程中，不僅考慮了義原之間的上下位關係，還考慮了義原之間的其他關係。在計算詞語相似度時，加權合併了《同義詞詞林》的詞義相似度、《知網》語義運算式的義原相似度和義原關聯度。由於《同義詞詞林》和《知網》採用完全不同的語義體系和表達方式，詞表也相差較大，因此這種演算法中把它們合併計算的合理性值得懷疑。另外，我們前面介紹過，詞語相關度和相似度是兩個不同的概念，把語義關聯度加權合併計入義原相似度中，是不合適的。

4.1 詞語相似度計算

對於兩個漢語詞語 W_1 和 W_2，如果 W_1 有 n 個義項（概念）：S_{11}，S_{12}，……，S_{1n}，W_2 有 m 個義項（概念）：S_{21}，S_{22}，……，S_{2m}，我們規定，W_1 和 W_2 的相似度是各個概念的相似度之最大值，也就是說：

$$Sim(W_1,W_2)=\max_{i=1..n,j=1...m} Sim(S_{1i},S_{2j}) \qquad \cdots\cdots(2)$$

這樣，我們就把兩個詞語之間的相似度問題歸結到了兩個概念之間的相似度問題。當然，我們這裏考慮的是孤立的兩個詞語的相似度。如果是在一定上下文之中的兩個詞語，最好是先進行詞義排歧，將詞語標注爲概念，然後再對概念計算相似度。

4.2 義原相似度計算

由於所有的概念都最終歸結于用義原（個別地方用具體詞）來表示，所以義原的相似度計算是概念相似度計算的基礎。

由於所有的義原根據上下位關係構成了一個樹狀的義原層次體系，我們這裏採用簡單的通過語義距離計算相似度的辦法。假設兩個義原在這個層次體系中的路徑距離爲 d，根據公式(1)，我們可以得到這兩個義原之間的語義距離：

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \qquad \cdots\cdots(3)$$

其中 p_1 和 p_2 表示兩個義原（primitive），d 是 p_1 和 p_2 在義原層次體系中的路徑長度，是一個正整數。α 是一個可調節的參數。

用這種方法計算義原相似度的時候，我們只利用了義原的上下位關係。實際上，在《知網》中，義原之間除了上下位關係外，還有很多種其他的關係，如果在計算時考慮進來，可能會得到更精細的義原相似度度量，例如，我們可以認爲，具有反義或者對義關係的兩個義原比較相似，因爲它們在實際的語料中互相可以替換的可能性很大。對於這個問題這裏我們不展開討論。

另外，在知網的知識描述語言中，在一些義原出現的位置可能出現一個具體詞（概念），並用圓括號( )括起來。所以我們在計算相似度時還要考慮到具體詞和具體詞、具體詞和義原之間的相似度計算。理想的做法應該是先把具體詞還原成《知網》的語義運算式，然後再計算相似度。這樣做將導致函數的遞迴調用，這會使演算法變得很複雜。由於具體詞在《知網》的語義運算式中只占很小的比例，因此，在我們的實驗中，爲了簡化起見，我們做如下規定：

具體詞與義原的相似度一律處理爲一個比較小的常數（γ）；

具體詞和具體詞的相似度，如果兩個詞相同，則爲 1，否則爲 0。

4.3 虛詞概念的相似度的計算

我們認爲，在實際的文本中，虛詞和實詞總是不能互相替換的，因此，虛詞概念和實詞概念的相似度總是爲零。

由於虛詞概念總是用“**{句法義原}**”或“**{關係義原}**”這兩種方式進行描述，所以，虛詞概念的相似度計算非常簡單，只需要計算其對應的**句法義原**或**關係義原**之間的相似度即可。

4.4 實詞概念的相似度的計算

從前面的分析可知，《知網》的知識描述語言可以通過義原和集合、特徵結構這兩種抽象資料結構來表達。義原之間的相似度計算問題已經解決，剩下的問題就是集合和特徵結構的相似度問題了。

我們的基本設想是：整體相似要建立在部分相似的基礎上。把一個複雜的整體分解成部分，通過計算部分之間的相似度得到整體的相似度。

假設兩個整體 A 和 B 都可以分解成以下部分：A 分解成 A_1，A_2，……，A_n，B 分解成 B_1，B_2，……，B_m，那麼這些部分之間的對應關係就有 m×n 種。問題是：這些部分之間的相似度是否都對整體的相似度發生影響？如果不是全部都發生影響，那麼我們應該如何選擇發生影響的那些部分之間的相似度？選擇出來以後，我們又如何得到整體的相似度？

我們認爲：一個整體的各個不同部分在整體中的作用是不同的，只有在整體中起相同作用的部分互相比較才有效。例如比較兩個人長相是否相似，我們總是比較它們的臉型、輪廓、眼睛、鼻子等相同部分是否相似，而不會拿眼睛去和鼻子做比較。

因此，在比較兩個整體的相似性時，我們首先要做的工作是對這兩個整體的各個部分之間建立起一一對應的關係，然後在這些對應的部分之間進行比較。

還有一個問題：如果某一部分的對應物爲空，如何計算其相似度？我們這裏採用一種簡單的處理辦法：

將任一非空值與空值的相似度定義爲一個比較小的常數（δ）；

下面我們分別考慮集合和特徵結構的相似度計算問題。

4.4.1 特徵結構的相似度計算

特徵結構可以理解爲一個“屬性：值”對（Attribute-Value Pair）的集合，我們將一個“屬性：值”對稱爲一個“特徵”（Feature）。在一個特徵結構中，每個“特徵”的“屬性”是唯一的。

計算兩個特徵結構的相似度，首先要在兩個特徵結構的特徵之間建立起一一對應的關係。由於每個特徵結構的各個特徵都具有不同的屬性，因此這種一一對應關係通過特徵的屬性很容易建立起來：屬性相同的特徵之間一一對應，如果沒有屬性相同的特徵，那麼該特徵的對應物爲空。

這樣，特徵結構的相似度就轉化爲各個特徵的相似度的加權平均。其中的權值反映出該屬性在特徵結構中的重要程度。在目前我們認爲所有特徵具有相同的重要性。

剩下的問題就是計算兩個特徵的相似度。特徵由“屬性”和“值”組成。由於“屬性”相同，於是，兩個特徵的相似度可以等價於其“值”的相似度。

4.4.2 集合的相似度計算

集合的相似度計算比特徵結構更爲複雜，因爲集合的元素是無序而且平等的，因此首要任務是要在兩個集合的元素之間建立一一對應關係。

兩個集合的相似度計算模型，必須滿足我們對於集合相似度計算的一些直觀要求。這裏我們列出以下兩條：

1. 一個集合和它本身的相似度爲 1；

2. 假設兩個集合都有 n 個元素，其中 m（m<n）個元素相同，又假設兩個元素的相似度只能是 0（不同）或 1（相同），那麼這兩個集合的相似度應該是 m/n。

要計算兩個集合的相似度，最容易想到的方法是首先計算兩個集合的所有元素兩兩之間的相似度，然後再進行加權平均。但是這樣會帶來一個問題，就是一個集合和它本身的相似度可能不爲 1，除非它的任意兩個元素之間的相似度都爲 1。這個結果當然是不合理的。這也從另一個角度說明我們先前定義的原則（首先在兩個集合的元素之間建立一一對應關係）的合理性。

在本文中，我們採用以下演算法來爲兩個集合的元素之間建立一一對應關係：

1. 首先計算兩個集合的所有元素兩兩之間的相似度；
2. 從所有的相似度值中選擇最大的一個，將這個相似度值對應的兩個元素對應起來；
3. 從所有的相似度值中刪去那些已經建立對應關係的元素的相似度值；
4. 重複上述第 2 步和第 3 步，直到所有的相似度值都被刪除；
5. 沒有建立起對應關係的元素與空元素對應。

根據上述演算法建立起兩個集合元素的一一對應關係後，我們就很容易計算兩個集合的相似度了：集合的相似度等於其元素對的相似度的加權平均。又因爲集合的元素之間都是平等的，所以我們可以將所有的權值取成相同的，於是：集合的相似度等於其元素對的相似度的算術平均。

4.4.3 實詞概念相似度的計算

由前面的分析我們知道，在《知網》中對一個實詞的描述可以表示爲一個特徵結構，該特徵結構含有以下四個特徵：

第一基本義原描述：其值爲一個基本義原，我們將兩個概念的這一部分的相似度記爲 $Sim_1(S_1,S_2)$；

其他基本義原描述：對應於語義運算式中除第一基本義原描述式以外的所有基本義原描述式，其值爲一個基本義原的集合，我們將兩個概念的這一部分的相似度記爲 $Sim_2(S_1,S_2)$；

關係義原描述：對應於語義運算式中所有的關係義原描述式，其值是一個特徵結構，對於該特徵結構的每一個特徵，其屬性是一個關係義原，其值是一個基本義原，或一個具體詞。我們將兩個概念的這一部分的相似度記爲 $Sim_3(S_1,S_2)$；

關係符號描述：對應於語義運算式中所有的關係符號描述式，其值也是一個特徵結構，對於該特徵結構的每一個特徵，其屬性是一個關係義原，其值是一個集合，該集合的元素是一個基本義原，或一個具體詞。我們將兩個概念的這一部分的相似度記爲 $Sim_4(S_1,S_2)$。

於是，兩個概念語義運算式的整體相似度記爲：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2) \qquad \cdots\cdots(4)$$

其中，β_i（$1\le i\le 4$）是可調節的參數，且有：

$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\ge\beta_2\ge\beta_3\ge\beta_4$

後者反映了 Sim_1 到 Sim_4 對於總體相似度所起到的作用依次遞減。由於第一基本義原描述式反映了一個概念最主要的特徵，所以我們應該將其權值定義得比較大，一般應在 0.5 以上。

在實驗中我們發現，如果 Sim_1 非常小，但 Sim_3 或者 Sim_4 比較大，將導致整體的相似度仍然比較大的不合理現象。因此我們對公式(4)進行了修改，得到公式如下：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \qquad \cdots\cdots(5)$$

其意義在於，主要部分的相似度值對於次要部分的相似度值起到制約作用，也就是說，如果主要部分相似度比較低，那麼次要部分的相似度對於整體相似度所起到的作用也要降低。且可以保證一個詞和它本身的相似度仍爲 1。

下面我們再分別討論每一部分的相似度。

第一基本義原描述：就是兩個義原的相似度，按照公式(3)計算即可；

其他基本義原描述：其值爲一個集合，轉換爲兩個基本義原集合的相似度計算問題；

關係義原描述：其值爲一個特徵結構，轉換爲兩個特徵結構的相似度計算問題。而這個特徵結構中特徵的值就是基本義原或具體詞，因此這兩個特徵結構的相似度計算也可以最終還原到基本義原或具體詞的相似度計算問題。這裏，由於無法區分關係義原之間的重要程度，我們將對各個特徵的相似度取算術平均；

關係符號描述：其值爲一個特徵結構，轉換爲兩個特徵結構的相似度計算問題。而這個特徵結構中特徵的值又是一個集合，集合的元素才是基本義原或具體詞，因此這兩

個特徵結構的相似度計算也可以最終還原到基本義原或具體詞的相似度計算問題。同樣，由於無法區分關係符號之間的重要程度，我們將對各個特徵的相似度取算術平均；

到此爲止，我們已經討論了基於《知網》的詞語相似度計算的所有細節，具體的演算法我們不再詳細說明。

5. 實驗及結果

根據以上方法，我們實現了一個基於《知網》的語義相似度計算程式模組。

詞語相似度計算的結果評價，最好是放到實際的系統中（如基於實例的機器翻譯系統），觀察不同的相似度計算方法對實際系統的性能的影響。這需要一個完整的應用系統。在條件不具備的情況下，我們採用了人工判別的方法。

我們使用了三種方法來計算詞語相似度，並把它們的計算結果進行比較：

方法 1：僅使用《知網》語義運算式中第一基本義原來計算詞語相似度；

方法 2：Li Sujian et al. (2002) 中使用的詞語語義相似度計算方法；

方法 3：本文中介紹的語義相似度計算方法；

在實驗中，根據在多次嘗試中取得的經驗，我們將幾個參數值設置如下：

$\alpha = 1.6$

$\beta_1 = 0.5$, $\beta_2 = 0.2$，$\beta_3 = 0.17$，$\beta_4 = 0.13$

$\gamma = 0.2$

$\delta = 0.2$

實驗結果如下表所示：

表 3：實驗結果（一）

詞語 1	詞語 2	詞語 2 的語義	方法 1	方法 2	方法 3
男人	女人	人,家,女	1.000	0.668	0.861
男人	父親	人,家,男	1.000	1.000	1.000
男人	母親	人,家,女	1.000	0.668	0.861
男人	和尚	人,宗教,男	1.000	0.668	0.861
男人	經理	人,#職位,官,商	1.000	0.351	0.630
男人	高興	屬性值,境況,福,良	0.016	0.024	0.048
男人	收音機	機器,*傳播	0.186	0.008	0.112
男人	鯉魚	魚	0.347	0.009	0.209
男人	蘋果	水果	0.285	0.004	0.171
男人	工作	事務,$擔任	0.186	0.035	0.112
男人	責任	責任	0.016	0.005	0.126

考察方法 3 的結果，我們可以看到，“男人”（取義項“人，家，男”）和其他各個詞的相似度與人的直覺是比較相符合的。

將方法 3 和方法 1、方法 2 的結果相比較，可以看到：方法 1 的結果比較粗糙，只要是人，相似度都爲 1，顯然不夠合理；方法 2 的結果比方法 1 更細膩一些，能夠區分不同人之間的相似度，但有些相似度的結果也不太合理，比如“男人”和“工作”的相似度比“男人”和“鯉魚”的相似度更高。從可替換性來說，這顯然不合理，至少“男人”和“鯉魚”都是有生命物體，而“工作”只可能是一個行爲或者一個抽象事物。方法 2 出現這種不合理現象的原因在於其計算方法把部分語義關聯度數值加權計入了相似度中。另外，方法 2 的結果中，“男人”和“和尚”的相似度比“男人”和“經理”的相似度高出近一倍，而方法 3 的結果中，這兩個相似度的差距更合理一些。

表 4 中給出另外一些測試結果，供讀者參考：

表 4：實驗結果（二）

詞語 1	詞語 2	相似度	詞語 1	詞語 2	相似度
工人	教師	0.722	粉紅	紅	1
工人	科學家	0.576	粉紅	紅色	1
工人	農民	0.722	粉紅	綠	0.861
工人	運動員	0.722	粉紅	顏色	0.059
教師	科學家	0.576	綠	顏色	0.059
教師	農民	0.722	十分	非常	1
教師	運動員	0.722	十分	特別	0.624
科學家	農民	0.576	思考	考慮	1
科學家	運動員	0.6	思考	思想	0.074
農民	運動員	0.722	考慮	思想	0.074
中國	美國	0.936	跑	跳	0.444
中國	聯合國	0.136	跑	跳舞	0.127
中國	安理會	0.114	跑	運行	0.444
中國	歐洲	0.733	運行	跳舞	0.151

可以看到，絕大部分結果還是比較合理的，但也有部分結果不夠合理，例如“中國”和“聯合國”、“中國”和“安理會”的相似度都過低，這是因爲，“中國”、“聯合國”、“安理會”在《知網》中的第一基本義原分別是“地方”、“機構”、“部件”。“跑”和“跳”的相似度也較低，這是因爲這兩個詞被簡單定義爲兩個基本義原，而缺少其他資訊。這也從一個側面反映了知網的某些定義不合理或不一致之處。

需要聲明的是，上述試驗中，每個詞都只取了一個最常見的義項，而不是考慮所有義項。

6. 結論

與傳統的語義詞典不同，《知網》採用了 1500 多個義原，通過一種知識描述語言來對每個概念進行描述。

爲了計算用知識描述語言表達的兩個概念的語義運算式之間的相似度，我們採用了“整體的相似度等於部分相似度加權平均”的做法。首先將一個整體分解成部分，再將兩個整體的各個部分進行組合配對，通過計算每個組合對的相似度的加權平均得到整體的相似度。我們具體討論了特徵結構和集合這兩種抽象資料結構中各個組成部分的組合配對方式。通過對概念的語義運算式反復使用這一方法，可以將兩個語義運算式的整體相似度分解成一些義原對的相似度的組合。對於兩個義原的相似度，我們採用根據上下位關係得到語義距離並進行轉換的方法。

實驗證明，我們的做法充分利用了《知網》中對每個概念進行描述時的豐富的語義資訊，得到的結果與人的直覺比較符合，詞語相似度值刻劃也比較細緻。

參考文獻：

Agirre E. and Rigau G., “A proposal for word sense disambiguation using conceptual distance”, *Proc. of International Conference Recent Advances in Natural Language Processing (RANLP)* , 1995, pp. 258-264, Tzigov Chark, Bulgaria.

Dagan I., Marcus S., et al. , “Contextual Word Similarity and Estimation from Sparse Data”, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* , 1993, pp. 164-171

Dagan I., Lee L. and Pereira F., “Similarity-based models of word cooccurrence probabilities”, *Machine Learning, Special issue on Machine Learning and Natural Language,* 34(1-3) , 1999, pp. 43-69

Gauch S. and Chong M. K., “Automatic Word Similarity Detection for TREC 4 Query Expansion”, *Proc. of TREC-4: The 4th Annual Text REtrieval Conf.*, Nov. 1995, Gaithersburg, MD, 1995, pp. 527-536

LI Sujian, ZHANG Jian, HUANG Xiong and BAI Shuo, “Semantic Computation in Chinese Question-Answering System”, *Journal of Computer Science and Technology* 17(6) , 2002, pp. 993-999

LI Xiaobin, Szpakowicz S., and Matwin S., “A WordNet-based algorithm for word sense disambiguation”, *Proc. of the Twelth International Joint Conference on Artificial Intelligence (IJCAI)*. 1995, pp. 1368-1374

李涓子，“漢語詞義排歧方法研究”，清華大學博士論文，1999

王斌，“漢英雙語語料庫自動對齊研究”，中國科學院計算技術研究所博士學位論文，1999

魯松，“自然語言中詞相關性知識無導獲取和均衡分類器的構建”，中國科學院計算技術研究所博士論文，2001

董振東，董強（1999），“知網”, http://www.keenage.com

杜飛龍 (1999)，《知網》辟蹊徑，共用新天地——董振東先生談知網與知識共用，《微電腦世界》雜誌，1999 年第 29 期

Computational Linguistics and Chinese Language Processing
Vol. 7, No. 2, August 2002 , pp. 77-88

基於組合特徵的漢語名詞詞義消歧[1]

A Study on Noun Sense Disambiguation Based on Syntagmatic Features

王 惠*

WANG Hui

Abstract

Word sense disambiguation (WSD) plays an important role in many areas of natural language processing, such as machine translation, information retrieval, sentence analysis, and speech recognition. Research on WSD has great theoretical and practical significance. The main purposes of this study were to study the kind of knowledge that is useful for WSD, and to establish a new WSD model based on syntagmatic features, which can be used to disambiguate noun sense in Mandarin Chinese effectively.

Close correlation has been found between lexical meaning and its distribution. According to a study in the field of cognitive science [Choueka, 1983], people often disambiguate word sense using only a few other words in a given context (frequently only one additional word). Thus, the relationships between one word and others can be effectively used to resolve ambiguity. Based on a descriptive study of more than 4,000 Chinese noun senses, a multi-level framework of syntagmatic analysis was designed to describe the syntactic and semantic constraints of Chinese nouns. All of these polyseme nouns were surveyed, and it was found that different senses have different and complementary distributions at the syntax and/or collocation levels. This served as a foundation for establishing an WSD model by using grammatical information and a thesaurus provided by linguists.

---

[1] 本研究得到中國 973 重點基礎研究項目“面向新聞領域的漢英機器翻譯系統”(G1998030507-4）的支持。

* 北京大學計算語言學研究所，北京，100871 Email: whui@pku.edu.cn
Insitute of Computational Linguistics, Peking University, Beijing 100871, P.R.China

The model uses *the Grammatical Knowledge-base of Contemporary Chinese* [Yu Shiwen *et al.* 2002] as one of its main machine-readable dictionaries (MRDs). It can provide rich grammatical information for disambiguation of Chinese lexicons, such as parts-of-speech (POS) and syntax functions.

Another resource of the model is *the Semantic Dictionary of Contemporary Chinese* [Wang Hui *et al.* 1998], which provides a thesaurus and semantic collocation information of more than 20,000 nouns. They were employed to analyze 635 Chinese polysemous nouns.

By making full use of these two MRD resources and a very large POS-tagged corpus of Mandarin Chinese, a multi-level WSD model based on syntagmatic features was developed. The experiment described at the end of the paper verifies that the approach achieves high levels of efficiency and precision.

Key words: Word Sense Disambiguation, syntagmatic features, noun sense, Chinese Language Information Processing

1. 詞義消歧（WSD）概述

由於自然語言中一詞多義現象普遍存在，因此，要讓電腦正確地分析和理解自然語言，一個重要的前提就是能夠在某個特定上下文中，自動排除歧義，確定多義詞的意義。這就是通常所說的詞義消歧（Word sense disambiguation。

詞義消歧是大多數自然語言處理任務的一個必不可少的中間層次，使用帶詞義標注的文本可以提高資訊檢索中的查全率和查準率，實現基於概念的檢索；可以對漢語句法分析中類序同形的歧義問題的解決提供必要的語義信息，爲自動句法消歧提供幫助；在機器翻譯中有利於選擇可以恰當表達語句中詞的目標詞，以提高翻譯的準確性；利用大規模帶詞義標注的語料庫還可以建立基於語義類的語言模型，爲語音識別、手寫體識別和音字轉換提供幫助。因此，詞義消歧研究在自然語言處理領域具有重要的理論和實踐意義。從 50 年代初期開始就一直備受計算語言學家的關注[Ide, 1998]。

1.1 詞義消歧的知識源

早期人們所使用的詞義消歧知識一般是憑人手工編制的規則。但手工編寫規則費時費力，存在嚴重的知識獲取的“瓶頸”問題，只能處理爲數有限的個別詞，無法勝任處理大規模文本的詞義標注工作。

20 世紀 80 年代以後，詞典成爲人們獲取詞義消歧知識的一個重要知識源。Lesk[1986]、Luk[1995]根據《Oxford Advanced Learner’s Dictionary》中的釋義文本來判斷多義詞在上下文中的詞義。Dagan[1991]、Gale[1993]利用雙語對照詞典來幫助多義詞

消歧。Voorhees [1993]、Resnik [1995] 從不同角度利用 WordNet 中的上下位關係、同義關係進行英語詞義消歧探索。Yarowsky[1994]提出一種基於義類詞典《Roget's International Thesaurus》的詞義消歧方法。使用詞典作爲詞義消歧知識源的優點在於電腦可以從詞典中自動獲取識別多義詞的各個詞義的一些重要知識。但這種方法對詞的上下文不能進行預測，而且，對詞義消歧有幫助的一些組合特徵沒有在詞典中完全體現出來。

近年來，隨著電腦存儲容量和運算速度的飛速提高，通過使用各種機用資源和大規模語料庫，電腦能夠自動獲得各種動態的搭配知識及其統計資料，以此解決規則方法中的知識空缺問題。因而，詞義消歧研究中湧現出許多基於語料庫統計的方法。比如，Gale & Church[1992,1993]等利用雙語語料庫對英語多義詞進行訓練和測試。但使用雙語語料庫的主要問題是：獲得多義詞消歧知識的前提是一個多義詞在另一種語言中具有不同的翻譯詞，並且翻譯詞在另一種語言中必須是單義詞，這樣必然限定了多義詞的處理範圍。其次，雙語語料庫的規模和多樣性都很有限，大量多義詞或多義詞的某個詞義在語料中可能從未出現；而且由於現在雙語語料對齊技術尚不能達 100%的正確，也使得這種方法只能限定在小規模的實驗中。

總的來說，不管是基於規則的方法，還是基於詞典的方法，或者基於大規模語料庫的方法，任何詞義消歧系統都離不開詞義消歧時所用知識的資料源，詞義消歧知識庫的質量已成爲詞義消歧系統成敗的關鍵。英語詞義消歧研究已有多年的歷史，但大部分工作都由於缺少足夠的詞義知識，從而被限制在一個較小的規模（幾個或十幾個詞），大規模英語語料庫進行詞義標注的工作迄今尚未見到。

1.2 漢語詞義消歧研究

漢語詞義消歧研究從 20 世紀 90 年代以後才開始，主要是利用詞典提供的語言知識。清華大學童翔[1993]利用《同義詞詞林》中的語義分類，對漢語合成詞中的單字進行義項標注。此後，上海復旦大學曾使用《同義詞詞林》的中類語義編碼人工標注 5 萬語料，然後用一個二元模型進行訓練和測試，進行文本標注研究，正確率在 85%左右[轉引自李涓子 1999：18]。LAM[1997]利用《現代漢語詞典》的釋義文本和《同義詞詞林》的義類代碼，對實詞多義詞進行詞義消歧，平均正確率爲 45.5%。李涓子[1999]利用《同義詞詞林》、《現代漢語辭海》以及從大規模“人民日報”語料庫中獲取的詞語動態搭配知識，對文本中的每個詞進行詞義標注，平均正確率達到 84.77%，多義詞消歧的正確率爲 52.13%。此外，山西大學、哈爾濱工業大學、廈門大學也分別對漢語全文檢索、英漢機器翻譯等限定領域中的詞義消歧方法分別進行了探索[劉開瑛 1995；劉小虎 1998；Yang Xiaofeng 2002]。

漢語詞義消歧雖然在較短的時間內取得了令人鼓舞的進展，但它與英語詞義消歧一樣面臨著詞義知識獲取的“瓶頸”問題。現有的各種方法所利用的知識一般僅限於具體

的詞語搭配和義類信息（後者主要來自於《同義詞詞林》和“知網（Hownet）”）。由於詞典和語料庫中不可能包括每個詞的所有搭配實例；而有些低頻詞，在語料中出現次數也不多，很難搜集到它們的上下文環境，因而知識獲取中普遍存在著資料稀疏以及自動學習演算法的參數空間太大等問題。

2. 基於組合特徵的漢語詞義消歧

我們知道，詞義和詞的分佈之間具有密切的關係。一個詞無論包含多少種意義（sense），在一定語句中起作用的，往往只是其中某一個意義。詞的不同意義往往會在句法或辭彙搭配層面上表現出不同的組合特徵。人們之所以能夠在一定的上下文中理解多義詞的不同意義，正是借助於這些彼此獨立並且呈互補分佈的特徵。認知語言學家 Choueka[1983]的研究表明，人們通常僅僅利用上下文中的一個詞或少數幾個詞就能夠識別出多義詞的詞義。因此，完全可以根據詞與詞之間的組合關係來有效地分化多義詞。

對於電腦來說，要真正有效地提高詞義消歧的水平，不僅需要獲取詞的釋義和分類信息，而且更重要的是，綜合利用現有的語言知識資源，在詞類劃分基礎上，增加詞義的語法功能分析和語彙搭配描寫，從多知識源中提取多義詞的每個意義在不同層級上相互區別的組合特徵。

本文在北京大學計算語言學研究所開發的“現代漢語語法信息詞典”[俞士汶等, 2002]、“現代漢語語義詞典”[王惠等, 1998]和大規模語料庫的基礎上，提出了一種基於多級組合特徵的現代漢語詞義消歧策略。

2.1 利用詞類標記進行詞義消歧

從語言資訊處理角度來看，詞的組合特徵可以分爲兩大類，一類是詞類標記，一類是詞在上下文中的詞義搭配限制。漢語中有些多義詞的不同意義屬於不同的詞類，如“補貼”的❶義是動詞，❷義是名詞：

【補貼】❶貼補：～家用｜～糧價。❷貼補的費用：福利～｜副食～。

據筆者所作的調查，《現代漢語詞典》的 20513 個名詞中共有多義詞 3989 個，其中像“補貼”這樣包含不同詞類的意義的名詞有 932，占多義名詞的 23.4%。對 200 萬字的《人民日報》語料[1998 年 1 月]的統計結果與此相近，22744 個名詞中共有多義詞 2196 個，其中意義詞類不同的有 592 個，占 27%。這也就是說，僅僅利用詞類標記就可以消除超過 1/5 的歧義。

由於現有的漢語詞類標注工具已經可以達到 96%的正確率[李涓子 1999：30]，因此，對於詞類不同的意義，電腦可直接借助於語料中的詞類標記進行判斷。比如，遇到下面經過自動切詞、詞類標注[代碼解釋參見附錄 1]的文本：

[1]這/r 將/d 由/p 國家/n 予以/v 補貼/v。

[2]生活/n 補貼/n 很/d 快/a 發到/v 災區/n 人民/n 手/n 裏/f。

電腦可以很容易地根據詞類標注判斷出是例[1]中的“補貼”是❶義，例[2]中的“補貼”是❷義，從而給出正確的語義標注或英語譯文：

[1] This will be subsidized by the state.

[2] Living allowances were quickly handed out to the people in the stricken area.

2.2 詞類相同，則利用更細緻的語法功能與詞義搭配差異進行詞義消歧

如果一個詞的幾個意義都屬於名詞，詞性標記就無能爲力了。這時，可以根據更細緻的組合特徵來區分詞義。就現代漢語名詞而言，不僅數量巨大，而且據筆者統計，《現代漢語語法信息詞典詳解》所包含的 3491 個名詞中，有 23%是多義詞，單字詞中多義詞的比例更是高達 47.5%。單字詞平均有 2.8 個意義，雙音節詞有 2.2 個，三音節詞有 2 個。如：

【辦公室】❶辦公的屋子。❷機關、學校、企業等單位內辦理行政性事務的部門。

多義名詞內部的詞義關係也是錯綜複雜的，比如，有的是“部分～整體”關係，有的是比喻關係，“辦公室”的❷義則是從❶義引申而來的。因此，如何選取恰當的詞義組合特徵來把握數目龐雜的名詞，成爲問題的關鍵。

本文在對 4000 餘個名詞義項具體分析的基礎上，提出了一個多级的現代漢語名詞詞義組合分析框架：首先，考察名詞充當主語、賓語、定語、中心語等句法成分的能力及其所結合的詞類；然後，進一步揭示它在每個語法位置上的語義搭配限制。

這個分析框架把系統的語法分析與零散的辭彙語義搭配有機地結合在一起。利用它，我們可以對不同的名詞都可以採用統一的方法和步驟進行組合特徵分析。比如，“辦公室”的❶義指建築物，❷義是人（某種部門），把它們放入該框架，可清楚地顯示二者各自的組合特徵及其在分佈空間上的差異：

表 1 “辦公室”的兩個意義組合特徵對比

語法功能		❶義	❷義
直接作主語	～＋動詞	～改暗房	～提出/～發表聲明/～說
	～＋形容詞	～十分寬敞/～空了/～安靜	/
直接作賓語	動詞＋～	趨向動詞＋～： 闖進～/走進～/回到～/進～/到～/去～/走出～/離開～ 特定搭配： 調換～/坐～	特定搭配： 成立～/設立～
	介詞＋～	在～/從～（走過來）	

直接作定語	～＋名詞	～＋具體物： ～門/～窗戶/～玻璃	～＋人： ～主任/～秘書/～人員 特定搭配： ～工作
	～＋方位詞	～裏/～前/～內/～後面	/
	～＋處所詞	～門口/～門前	/
直接 作中心語	名詞＋～	身份＋～： 教員～/老師～/會計～/醫生～/個人～/主任～ 職位＋～： 校長～/所長～/廠長～/ 院長～/總理～/總統～ 組織機構＋～： 市政府～/紡織局～/外文系～	非指人名詞＋～： 縣誌～/國務院新聞～/外事～/港澳事務～/交易會～ 職位＋～： 校長～/場長～/所長～/廠長～/ 院長～/總理～/總統～ 組織機構＋～： 歷史系～/專案組～兒童村～
	數量詞＋～	一間～/一個～	一個～
	動詞＋～	/	就業安置～/消費指導～/春運～/住房解困～/糖業生產～
	人稱代詞＋的＋～	我的～/你的～/他的～	/

更重要的是，由於“現代漢語語法信息詞典”中已經對 35000 個名詞充當主語、賓語、定語、中心語等句法成分的能力及其所結合的詞類做了詳細的描寫，“現代漢語語義詞典”則進一步爲它們一一標注了語義類，並刻畫了它們在每個語法位置上的語義搭配限制。因此，通過查詞典，電腦就可獲得上述知識。

利用表 1 中的組合特徵，消歧系統可以對實際文本中出現的多義名詞的詞義進行判斷。比如[以下例句中的詞類代碼參見附錄]：

[1]國務院/n 僑務/n 辦公室/n 主任/n 郭東坡/nr 向/p 海外/s 同胞/n 和/c 國內/s 歸僑/n、僑眷/n、僑務/n 工作者/n 發表/v 新年/t 賀詞/n。

[2]去年/t，市/n 再/d 就業/v 辦公室/n 提供/v 了/u 3 萬/m 元/q 貸款/n。

[3]他們/r 衣著/n 鮮亮/a，一看便知/l 是/v 從事/v 辦公室/n 工作/n 的/u。

[4]職工們/n 跑進/v 廠長/n 辦公室/n, 興奮/a 的/u 神態/n 難以言表/l。

[5]每/r 間/q 辦公室/n 都/d 是/v 玻璃/n 拉門/n。

[6]在/p 簡陋/a 的/u 辦公室/n 裏/f，鄭朝銓/nr 副/b 廠長/n 表示/v 了/u 謹慎/a 的/u 樂觀/an。

[7]曾/d 在/p 辦公室/n 將/p 25 萬/m 日元/n 的/u 餐費/n 單據/n 交予/v [三和/nz 銀行/n]nt 職員/n 要求/v 報銷/v。

由於目前的漢語自動句法分析研究還尚未達到實用階段，難以給出一個詞的句法功

能信息。因此，詞義組合特徵的選擇與判斷，首先應著重依據搭配詞的詞類標記，保證選擇出的上下文信息與該多義詞盡可能存在句法關係。

對於“辦公室”來說，❶義可以後接方位詞和處所詞，也可跟在介詞“在、從”的後面；❷義則可受動詞修飾，即：

❶義　*右組合*：　～+方位詞　～+處所詞　　*左組合*：　介詞+～

❷義　*左組合*：　動詞+～

據此，電腦可以很有把握地判斷出例 6 及例 7 中的“辦公室”都是❶義，例 2 中的“辦公室”是❷義。

如果詞類串相同，則可進一步觀察兩個意義的辭彙搭配限制。比如，❶義、❷義都可以直接修飾名詞，但❶義後面的名詞通常表示無生命的具體物質，而❷義的修飾物件是人或者“工作、事務”等抽象名詞，　即：

❶義　*右組合*：～+名詞（具體物“門、窗戶、玻璃……”）

❷義　*右組合*：～+名詞（人“主任、秘書…”.or. 抽象物“工作、事務…”）

根據這個條件，詞義標注系統可以正確判斷出例 1、例 3 中的“辦公室”都是❷義。再如，❶義、❷義都可以與個體量詞組合，但❶義可以與“間、個”搭配，❷義則只能與“個”搭配，即：

❶義　*左組合*：量詞（“間、個”）+～

❷義　*左組合*：量詞（“個”）+～

因此，例 5 中的“辦公室”是❶義。

如果詞語在某個語法位置上的詞類串相同，辭彙搭配也相同，則需要進一步考察其他組合特徵。如“辦公室”的兩個意義都可以受表示“身份”與“組織機構”的名詞直接修飾，　“廠長辦公室”、“林業局辦公室”中的“辦公室”既可能是❶義，也有可能是❷義。但在具體的具體句子中，比如例 4 中，根據“廠長辦公室”前面的趨向動詞“進”，則可以判斷出其中的“辦公室”指❶義；在下面這句話中，由於“辦公室”後面跟有名詞“主任”，因而它肯定是指❷義：

[8]鹽池縣/nt 林業局/n 辦公室/n 主任/n 說/v

❶義、❷義都可以直接作動詞的賓語，但❶義前面的動詞通常是趨向動詞，而❷義前面的動詞僅限於“成立、設立”等。

由以上分析我們看到，詞性標記相同的多義詞各個意義的歧義消解，實際上是利用了詞義的兩個不同層次的組合特徵：（1）詞義與其他詞語組合構成的詞類串；（2）在每個詞類串中所能搭配的語義類或具體詞語。

初步試驗結果表明，《人民日報》1998 年 1 月中共出現 62 個“辦公室”，依靠組合詞類串，電腦可正確地判斷出其中 15 個表示❶義，7 個表示❷義；依靠搭配物件的語義類或特徵詞，電腦可以準確地判斷出其中 16 個是❶義，22 個是❷義。

3. 現代漢語名詞的自由義和非自由義

在詞義組合特徵描述的基礎上，詞義消歧知識庫中如果加入詞義的組合自由度信息，將會更加提高消歧系統的效率。

現代漢語名詞在句法分佈中並不是完全自由的，而是或多或少地要受到一些限制。比如，有些可以充當多種句法成分，有些則只能出現在其中一兩個位置上。筆者對《現代漢語語法信息詞典詳解》中 3500 個名詞（4319 個義項）的語法功能進行了統計，結果表明：

表 2　現代漢語名詞的句法功能

句法功能		數目	所占比例
單作主語		3926	94.8%
單作賓語		4011	97.5%
作謂語		3	0.1%
作補語		0	0
作狀語	直接修飾動詞	1	0.05%
作定語	直接修飾名詞	3210	74.7%
做中心語	受數量詞修飾	3745	86.8%
	受名詞直接修飾	3299	76.7%
	受動詞直接修飾	964	22.5%
	受人稱代詞直接修飾	351	5.8%
	受數詞直接修飾	138	2.2%

由表中可以清楚地看到，沒有一項語法功能是全體名詞都具備的。名詞作賓語、主語的能力最強，作中心語（受數量詞、名詞直接修飾）次之，作定語（直接修飾名詞）也在 70%以上；而能作謂語、狀語、或受動詞、人稱代詞、數詞直接修飾的都只有極少數名詞。因此，我們可以把前 5 項分佈看作是現代漢語名詞的優勢分佈。具有全部這 5 項優勢分佈的名詞義稱爲名詞的自由義，否則，是非自由義。如：

【樓】❶樓房：一座～｜大～｜教室～｜高～大廈。

❷樓房的一層：一～（平地的一層）｜一口氣爬上十～。

“樓”的❶義是自由義，❷義分佈範圍比❶義狹窄得多，只能受基數詞、量詞“層”修飾，或者作動詞“上、下”的賓語，因而是非自由義。如：

[1]報館在三層樓，電梯外面掛的牌子寫明到四樓才停。

[2]洋老鼠在裏面踩車、推磨、上樓、下樓，整天不閑著，——無事忙。

一般來說，多義名詞的各個義項中只有一個自由義，其餘都是非自由義。由於自由義的分佈範圍和出現頻率都要遠遠高於非自由義，因此，電腦可以把多義詞中的自由義作爲預設值。比如，1998 年 1 月份的《人民日報》語料中，“樓”這個詞共出現 67 次，詞義消歧系統首先都假定它是❶義：

[1]他/r 決定/v 帶/v 新/a 領導/n 到/v 這/r 座/q 樓/n 看看/v。

[2]樓/n 高/a 了/y ，老百姓/n 的/u 生活/vn 環境/n 改善/v 了/y。

[3]他/r 走進/v 樓/n 內/f ，樓道/n 十分/m 昏暗/a。

只有上下文中出現了“樓”❷義的典型搭配特徵時，如下面例 5 中“樓”前面有動詞“下”，例 6 中“樓”前面有動詞“上”，例 7 中的“樓”前有數詞“11”和“8”，例 8 中的“樓”前有量詞“層”，電腦借助於這些特徵詞才將系統的預設值取消，判斷出這幾個“樓”都是❷義。如：

[4]媽/n 老/a 了/y，腿腳/n 不/d 利索/a 了/y，懶得/v 下/v 樓/n 啦/y！

[5]羅/nr 科長/n 親自/d 從/p 11/m 樓/n 將/p 師傅/n 扶到/v 8/m 樓/n。

[6]他/r 竟然/d 沒有/d 看到/v 一/m 棟/q 兩/m 層/q 樓/n 的/u 房子/n。

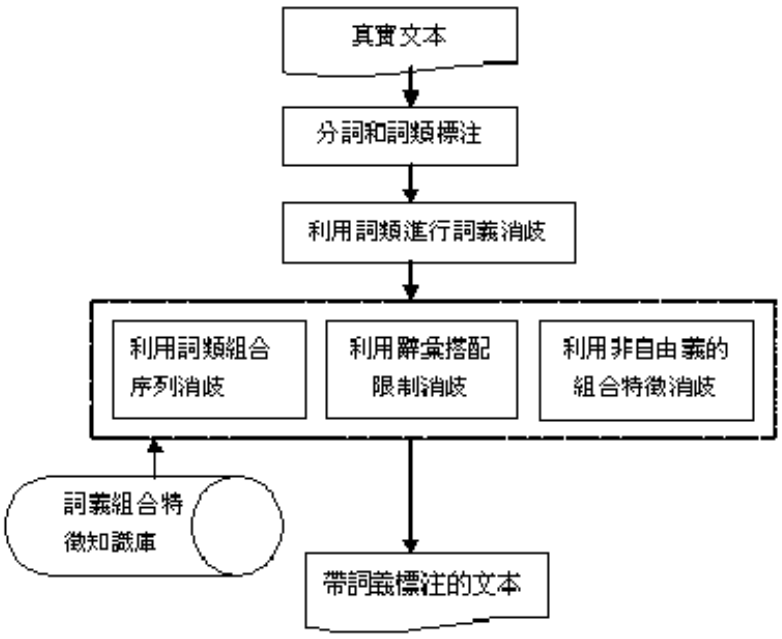
利用這種辦法，電腦迅速指出 67 個“樓”中有 23 個表示❷義。經檢查只有下面 1 例錯誤，其他全部正確。

阿西·賽德克已/nr 冒/v 著/u 漫天/z 飛雪/n 趕往/v 烏魯木齊市/ns 八/m 樓/n 附近/f 去/v 簽訂/v 1998 年/t 的/u 房屋/n 承包/vn 合同/n。

由上面的分析我們可以清楚地認識到，詞義組合特徵分析確實可有效地提高詞義消歧知識庫的質量，滿足漢語名詞詞義自動消歧的需要。但問題是這樣一個詞義知識庫規模究竟多大才能夠達到基本的實用水平呢？根據《現代漢語頻率詞典》[北京語言學院出版社，1985：492-514]的統計，1144 個高頻詞對語料的覆蓋程度約爲 75%，而且其中六成以上是多義詞。可見，數量不多的高頻多義詞是影響漢語真實文本詞義消歧準確率的關鍵。如果我們在詞義組合分析基礎上，對高頻多義詞的各個意義的組合能力進行集中研究和詳細描述，不僅可以有效地提高詞義知識庫的質量，而且也可以指導自動學習演算法的參數設計，將會十分有助於解決消歧語義知識獲取的瓶頸問題。

4. 結語

任何詞義消歧系統都離不開詞義消歧時所用知識的資料源。本文提出了一種充分利用現有資源，把語法功能、語義搭配等不同層面的知識統一起來分級描寫的詞義組合特徵庫的設計原則，並給出了一個基於詞義組合特徵的詞義消歧模型：



本文工作的最基本思想是分層次描寫漢語詞義的組合能力。目前，主要是對名詞的組合特徵分析及其在詞義消歧中的應用進行了一些試驗性的探索。初步的實驗結果是令人欣慰的，我們希望在積累了更多的實踐經驗後，能進一步完善這一詞義組合分析框架，並將這種思路應用於動詞、形容詞的詞義知識庫構造之中，同時努力實現由電腦輔助抽取詞義的組合特徵。

參考文獻

Choueka, Y. and S. Lusignan, "A Connectionist Scheme for Modeling Word Sense Disambiguation". *Cognition and Brain Theory*. 6 (1) 1983, pp.89-120

Dagan, Ido, Alon Itai, and Shaul Markovitch. "Two Languages Are More Informative Than One". In: *The 29th Annual Meeting of Association for Computational Linguistics*, Berkeley, CA: ACL, 1991. pp 130-137

Gale, William A, Kenneth W. Church, and David Yarowsky. "Using bilingual materials to develop word sense disambiguation methods". In: *The International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992. pp 101-112

Gale, William A, Kenneth W. Church, and David Yarowsky. "A Method for Disambiguation Word Senses in a Large Corpus". *Computer and the Humanities*. (26) 1993. pp 415-439

Ide, Nancy; Jean Véronis. “Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art”, *Computational Linguistics*. Vol.24, No.1, 1998. pp1-40

LAM SZE-SING, KAM-FAI WONG, and VINCENT LUM. “LSD-C –A. linguistic-based word-sense disambiguation algorithm for Chinese”. *Computer Processing of Oriental Languages*, Vol. 10, No. 4, 1997, pp 409-422

Lesk, Michal. “Automatic sense disambiguation: How to tell a pine from an ice cream cone”. In: Association for Computing Machinery, eds. *The 1986 SIGDOC Conference*. New Yark, ACM. 1986. pp24-26

Luk, Alpha K. “Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions”. In: ACL eds. *The 33rd Annual Meeting of ACL*, Cambridge, Massachusetts. 1995. pp181-188

Resnik, Philip. “Selection and Information: A Class-Based Approach to Lexical Relation”. [Ph. D. Dissertation], USA: University of Pennsylvania. 1993. pp 23-54

Towell, Geoffrey; Ellen M. Voorhees. “Disambiguating Highly Ambiguous Words”. *Computational Linguistics*, Vol.24, No.1, 1998. pp125-145

Yang Xiaofeng, Li Tangqiu. “A Study of Semantic Disambiguation Based on HowNet”. Computational Linguistics and Chinese Language Processing. Vol.7, No.1, 2002, pp47-78

Yarowsky, David. “Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French”. In: ACL eds. *The 32nd Annual Meeting of Association for Computational Linguistics*. Las Cruces, NM: ACL, 1994. pp 88-95

李涓子. “漢語詞義排歧方法研究” [博士學位論文]. 清華大學圖書館. 1999.

劉開瑛. “漢語全文檢索中的義項標注技術研究”. 《計算語言學進展與應用》. 北京：清華大學出版社，1995.

劉小虎. “英漢機器翻譯中詞義消歧方法的研究” [博士學位論文]. 哈爾濱工業大學. 1998.

童翔. “漢語真實文本的義項標注” [碩士學位論文]. 清華大學圖書館. 1993

王惠, 詹衛東, 劉群. “現代漢語語義詞典的設計與概要”.《1998 中文信息處理國際會議論文集》. 北京: 清華大學出版社. 1998. pp361~367

俞士汶, 朱學鋒, 王惠, 张化瑞. 《現代漢語語法信息詞典詳解》, 北京: 清華大學出版社. 2002.

附錄

語料庫標注詞類代碼表

代碼	詞類名稱	代碼	詞類名稱
a	形容詞	nz	其他專有名詞
b	區別詞	Ng	名語素
c	連詞	o	象聲詞
d	副詞	p	介詞
e	嘆詞	q	量詞
f	方位詞	r	代詞
h	前綴	s	處所詞
i	成語	t	時間詞
j	縮略語	u	助詞
k	後綴	v	動詞
l	習慣用語	vd	動副詞
m	數詞	vn	動名詞
n	名詞	x	非語素字
ns	地名	y	語氣詞
nt	機關團體名称	z	狀態詞

Computational Linguistics and Chinese Language Processing
Vol. 7, No. 2, August 2002, pp. 89-100

《現代漢語新詞語資訊電子詞典》的研究與實現[1]

Development and Study of the "Modern Chinese New Words Information Electronic Dictionary"

亢世勇*

Shiyong Kang

摘要

本文從四個方面介紹了我們正在開發中的《現代漢語新詞語資訊電子詞典》：（1）現代漢語新詞語的界定，（2）新詞語詞典的開發思想，（3）新詞語的採集與新詞語屬性資訊的描述，（4）近四萬新詞語的歸類實踐。我們認定的新詞語是指 1978 年以來通過各種途徑產生的、具有基本詞彙沒有的新形式、新意義或新用法的語文詞語。除了詞形、詞義或用法任何一個方面“新”外，還要求必須是人們日常生活中普遍、廣泛使用的語文詞語，人名、地名以及專科術語都不屬於我們所說的“新詞語”。我們堅持開放的原則，儘量全面的採集收錄新詞語，用人機兩用的研究理念，以北京大學計算語言學研究所的《現代漢語語法資訊詞典》爲模型打造一部收詞全面、資訊豐富、資源高度共用的現代漢語新詞語電子詞典，爲新詞語的研究、中文資訊處理的研究提供一個寶貴的資源。目前已收錄新詞語近 4 萬，首先我們按照現代漢語詞類的“優勢語法”功能，給這四萬新詞語分類並歸類，然後，利用成熟的關聯資料庫（在 ACCESS 環境下實現）詳細地描述了每個詞語的屬性資訊。設立總庫一個，語法資訊庫三個，包括名詞庫、動詞庫、形容詞庫，另外還設立了構詞法庫，舊詞庫、外來詞庫、簡略詞庫。總庫和其他各庫通過“詞語、拼音、義項”三個欄位聯繫起來，構成了一個具有上下位關係的有機系統，便於資訊的提取。這些庫總共設立屬性欄位 200 多個，包括每個詞語的語音資訊、語義資訊、來源

[1] 本項研究得到中國國家哲學社會科學規劃專案（01CYY002）支援；
本文於 2002 年 4 月在臺北舉行的“第三屆中文辭彙語義學會議”上宣讀，會後根據專家的意見作了修改，謹致謝忱。

* 山東煙臺師範學院中文系（264025） E-mail:kangsy46@sohu.com Tel:0535-6672439
Shandong of China: Yantai Normal College - Chinese Language Dept. (264025)

資訊、構詞法資訊、句法資訊和部分語用資訊。本詞典是目前國內收詞量最大、描寫資訊最多的一部新詞語詞典。

關鍵字：中文資訊處理　新詞語　電子詞典

Abstract

We introduce the development of the Electronic Lexicon of Contemporary Newborn Chinese Words: (1) the definition of a newborn word, (2) the main principle behind constructing the lexicon, (3) the collection of newborn words and their feature descriptions of them, and (4) the classification of 40,000 newborn words. In our opinion, a new bornword is a character string that appeared after 1978 in a new form, with a new meaning and with a new usage. In addition, it must be frequently used and accepted, but the names of men and places are not newborn words according to our definition. The approach to collecting newborn words is quite unrestricted, that is, the more the better. Based on the Contemporary Chinese Grammatical Knowledge Base of the Institute of Computational Linguistics at Peking University, we have finished compiling a lexicon of almost 40,000 newborn words semi-automatically. The lexicon, we believe, is a worthy resource for research on Chinese word-building rules and Natural Language Processing. Firstly, classification is done based on the preponderant grammatical characteristics of each word, and then the detailed features are described in the database of ACCESS. The lexicon contains a total base and three grammatical bases (i.e., a noun base, verb base and adjective base); what's more, it also has an old word base, a loanword base and a acronym base. The entire base is related to the sub-bases through the fields of word, phonetic notation and semantics fields, which form a hypernymy hierarchy that is quite convenient for searching. Totally, there are more than 200 fields in the bases that give information regarding phonetic notation, semantics, sources, word building, syntax and pragmatics. Without doubt, this lexicon is one of the largest domestic lexicons available with the most detailed descriptions of newborn Chinese words.

Keywords: Chinese information processing, New words, Electronic dictionary

1. 引言

2001 年我們獲得了中國國家社科規劃專案“《現代漢語新詞語資訊電子詞典》的開發與應用”（專案編號：01CYY002）。一年來，我們已按照規劃做了大量的工作，專案進展順利。本文從四個方面介紹《現代漢語新詞語資訊電子詞典》（以下簡稱“新詞語詞

典”）的基本情況：（1）現代漢語新詞語的界定（2）新詞語詞典的開發思想（3）新詞語的採集與新詞語詞典所描述的屬性資訊（4）近四萬詞語的歸類實踐

2. 現代漢語新詞語的界定

對於“新詞語”目前學術界有不同的看法，在全面考察了近 4 萬個新詞語並且借鑒、吸收了學術界新詞語研究成果的基礎上，我們認爲新詞語可以定義爲：通過各種途徑產生的、具有基本詞彙沒有的新形式、新意義或新用法的語文詞語。新詞語的特點在於“新”，“新”具體表現在詞形、詞義和詞語的用法上。鑒定新詞語的參照系是現代漢語基本詞彙的詞形、詞義和用法。只要在這三個方面的任何一點上與現代漢語基本詞彙不同，我們就認爲它是新詞語。基本詞彙的代表是北京商務印書館出版的 96 版的《現代漢語詞典》和《漢語大詞典》。“新”還有時間的限定，即 1978 年以來出現的新詞語。我們認定的新詞語既有“新”的特點，同時強調了新詞語的使用範圍，即必須是在社會生活中廣泛使用的語文性質的新詞語，可以進入普通辭彙的新詞語，那些新出現的專業術語沒有增加新的普通辭彙意義的，不在我們認定的新詞語範圍內。我們認定的新詞語具體如下：

(1) 新造詞語。比如“打假、扶貧、股盲、展銷、股市、高開、低走、哇噻、彩票、足彩、辣妹、酷裝、新新人類、哈韓族、哈日族、知本家、黑哨”等等。

(2) 舊詞新用。這類詞語詞形是原有的，“新”主要表現在產生了新意義或有了新的運用。具體分爲三種情況：A、原有的詞語增加了新的意義，如“下課、上課、氣候、跳槽、起飛、紅娘、窗口、下崗、亮相、新登場、跟進、充電、輸血、造血”等；B、原有的詞語有了新的用法。比如“結構”本來是名詞，但用爲動詞，如：你爲我結構人生；“運氣”原爲名詞，用爲形容詞，如：你這人很運氣。“火”原爲名詞，用爲形容詞，形容事物或人有聲勢，受歡迎。如：組織者們真沒想到晚會竟然這麼“火”。C、原有的詞語很長一段時間不用，又重新啓用，比如：“高就、賞光、黑道、綁票、撕票、夜總會、小姐、太太、金婚、銀婚”等。其中有些意義也發生了一些變化，比如“高就、賞光、太太、小姐”等原來主要用於地位比較高的人，有特指性，現在已經泛化，不論地位高低都可以用，變成了一種普通的說法。

(3) 方言辭彙進入普通話辭彙。如“炒魷魚、發燒友、埋單、的士、連鎖店、服裝城、跳樓價、大出血、娛樂圈、拍拖、三級片、主打、金曲、勁歌、勁舞、搞笑、爽、靚、馬子、二奶、套磁、磁實、貓膩、腕兒、搓、傍大款、侃大山、膀爺”等。

(4) 外來詞，從外族語借來的詞，又有：A、音譯詞如“的士、巴士、歐佩克、可口可樂、丁克、克隆、基因、託福、卡拉 OK、拜拜、酷(cool)、蔻(cute)、秀(show)、脫口秀(talk show)、血拼（shopping）、派對（party）、伊妹兒（E-mail）”

等；B、意譯詞，如“熱點（hot spot）、音樂電視（music television）、熱狗（hot dog）、超級市場（supermarket）”；C、音譯兼意譯詞，如“鐳射、呼啦圈、桑拿浴、迷你裙、吧女、酒吧、”等；D、直接使用日語的詞語，如：“放送、慰安婦、物語、寫真、人氣”等。

(5) 簡略詞，在原有詞語的基礎上縮略而成的詞語。分爲三種情況：A、簡稱詞，如“博導（博士研究生導師）、澳網（澳大利亞網球公開賽）、超市（超級市場）“；B、略語詞，如“嚴打（嚴厲打擊犯罪活動）、打假（打擊假冒僞劣商品）、防僞（防止假冒僞劣產品）、台資（臺灣人投入的資本）”；C、縮語詞，如“三講、三個代表、三假、三陪、三金”等。

(6) 修辭用法穩定下來構成的新詞語。主要有：A、比喻引申，如“豆腐渣工程、枕頭風、撒胡椒麵、下毛毛雨、泡沫經濟、朝陽產業、白色消費、下海、撈人”等；B、借代，如：“菜籃子工程、白髮世界、白條案、老人頭”等。C、仿擬，比如：“煙民、股民、彩民、網民”，“空姐、海姐、吧姐、呼姐、網姐、空嫂、海嫂、吧娘、呼嫂”，“文盲、科盲、股盲、舞盲、網盲”、“網民、網友、網哥、網姐、網迷、網蟲、網蠅”等等。

(7) 專用術語意義泛化、轉移，擴大使用範圍，轉爲普通辭彙。如“軟體、硬體、啓動、熱處理、冷處理、黃牌、主旋律、套牢、觸電、放電”等。

(8) 字母詞。主要有三類：A、純粹的字母詞，整個詞由英文字母構成，如“CT、IBM、CIA、TOFEL、GRE、CEO、ATM、CFO、BBS、CVD、DVD、VS、IT、IN、Q、VIP”等等；B、字母和漢字的組合，如“BP 機、BP 族、CALL 機、E 時代、E 人類、IT 界、IT 業、夠 IN、VIP 卡、很 Q”等等；C、數位和字母的組合，如“3D、3C、3S”等等。

3. 《現代漢語新詞語資訊電子詞典》的開發思想

3.1 新詞語研究的局限

現代漢語新詞語的研究受到了國內外的廣泛關注，學者們也做了大量的研究，產生了一些引人注目的研究成果。出版了新詞語詞典及詞語集三十多種、新詞語研究專著兩本，但是這些著作對新詞語的研究都有一定的局限。主要表現在以下方面：（1）這些研究成果都是印刷品，沒有有效的電子版成果，不能實現資源高度共用。（2）這些成果都是爲人用的，而沒有考慮到機器使用，應用範圍受到了限制。（3）由於受到研究技術和研究條件的限制，各種詞典收詞量有限，詞語的解釋及引例都有欠妥之處，更重要的是詞典提供的信息量極其有限。由於以上的不足，造成現有的各種新詞詞典應用價值不高。

3.2 《現代漢語新詞語資訊電子詞典》開發的目標

(1) 希望創建現代漢語新詞語研究的基礎平臺，實現資源高度共用，獲得較高的應用價值。本項研究利用電腦資料庫技術和相關的語料庫技術進行現代漢語新詞語的跟蹤研究，研究成果形式爲有效、實用的電腦資料庫軟體，其中包括新詞語電子詞典和大規模的相關語料，這樣可以實現資源的高度共用，使其具有較高的應用價值。

(2) 希望在漢語研究和中文資訊處理研究方面做出積極的貢獻。以往漢語的研究的資料和手段限制了漢語大規模的實用化的研究，由此造成的直接後果是嚴重制約了中文資訊處理的發展。本項研究利用電腦技術進行，積累了大量的機器可讀文件，爲大規模的實用的漢語研究奠定了基礎，其研究成果——新詞語屬性資訊電子詞典以及新詞語的構詞規律可以直接應用於中文資訊處理的未登錄詞語識別，有利於提高中文資訊處理技術的水平。

3.3 《現代漢語新詞語電子詞典》的開發具體思路

介於目前有關新詞語的研究比較零散，而且新詞語的研究又有十分重要的作用，我們擬對新詞語進行大規模的比較完備的研究。具體思路爲：

(1) 儘量窮盡地收集現有的新詞語，做到全面、準確。目前已收錄新詞語近 4 萬，收錄了我們所能見到的所有新詞語。

(2) 按照人機兩用的研究理念，打造一部適合於“人讀”和“機讀”的電子詞典。增加詞典的信息量，擴大詞典的使用範圍，提高其應用價值。

(3) 以北京大學計算語言學研究所的《現代漢語語法資訊詞典》爲模型，採用分類與屬性描述相結合的方法，在粗分詞類的基礎上對每個詞語語法語義屬性資訊進行詳細描述。具體採用成熟的關聯資料庫形式描述詞語和語法、語義屬性的二維關係，成果爲資料庫文件格式的電子詞典。

(4) 一部開放的詞典。本詞典在新詞語的收集及屬性的描述方面均堅持開放的原則，將跟蹤漢語辭彙的發展變化和漢語資訊處理的發展，不斷地收集、增加新詞語，增加新詞語屬性資訊的描述，以滿足實際需要。

4. 《現代漢語新詞語資訊電子詞典》詞語的採集與所描述的屬性資訊

4.1 新詞語的採集

首先利用我們自己開發好的《新詞語詞典資訊庫》和語料庫整理出一個新詞語詞表，然後按照我們的收詞原則——全面性原則、規範性與描寫性相結合原則、必要性原則、普

遍性原則、穩定性原則、音節原則等，從詞表中遴選出新詞語 3 萬多個，形成了新詞語詞典的基礎。此後，我們利用語言資訊處理技術不斷地從網上抓取新詞語及相關的例句集，不斷地擴充新詞語詞典。確定新詞語詞典中的詞目後，利用新詞語詞典資訊庫和包含《人民日報》1978 年以來的語料、《南方周末》創刊以來的語料以及人民日報報系其他報紙、人民網、光明日報、新民晚報等近年來語料的超大規模語料庫建立包含該詞語的例句集，考察這些詞語的意義和用法，描述其義項、語法屬性、語義屬性以及其他資訊等，從而開發出《現代漢語新詞語資訊電子詞典》。這些工作很大程度上利用電腦語料庫管理技術，在大規模機讀語料庫的支援下進行，能夠比較全面地考察每個新詞語的分佈環境，提高新詞語採集、收錄的合理性和資訊描述的準確度和覆蓋範圍，從而提升詞典的質量。

4.2 新詞語詞典屬性資訊的確立

新詞語詞典開發主要是爲了學習、研究新詞語，特別是爲中文資訊處理提供一個基本資源。爲了達到這一目的，新詞語詞典屬性資訊包括了語音資訊、來源資訊、語法資訊和部分語義、語用資訊，涉及了新詞語形、音、義以及用法的主要方面。

新詞語詞典描述的主要屬性資訊包括以下方面：

(1) 詞的常規資訊。包括詞的讀音、義項、音節、例句等。

(2) 語法資訊。按照北京大學計算語言學研究所的《現代漢語語法資訊詞典》的規格描寫新詞語的語法資訊。詞類體系沿用《現代漢語語法資訊詞典》的 18 個基本類，再加上成語、慣用語。詞類標記與其相同。各類詞語法屬性的設立在《現代漢語語法資訊詞典》基礎上有所改動，使其更加優化。

(3) 構詞法資訊。構詞法主要分爲單純構詞法和合成構詞法兩類。單純構詞法又分爲單音單純詞、多音單純詞。多音單純詞又分爲聯綿詞、音譯詞和疊音詞等。聯綿詞又分爲雙聲、疊韻其他等。合成詞又分爲複合式、重疊式、附加式三類。複合式又分爲聯合式、偏正式、補充式、動賓式和主謂式等。附加式又分爲兩種類型：“字首+詞根”、“詞根+尾碼”等。對於複合詞將構成複合詞的幾部分分解開來，分別標上該語素所屬的“詞性”，以便進一步考察由語素按照一定的構詞方法構成的新詞語的詞性的規律。

(4) 産生途徑。根據我們的考察主要包括：新造詞，舊詞新用，方言詞進入普通話的辭彙，外來詞，簡略詞，修辭用法穩定下來構成新詞，術語擴大使用範圍産生新義。

(5) 應用領域。應用領域的劃分是一個比較棘手的問題，我們大體上分爲政治、經濟、法律、軍事、文化、科技、教育、衛生、體育、商業、工業、農業、生活、通用等，暫時作爲工作規範，以後再逐漸調整。

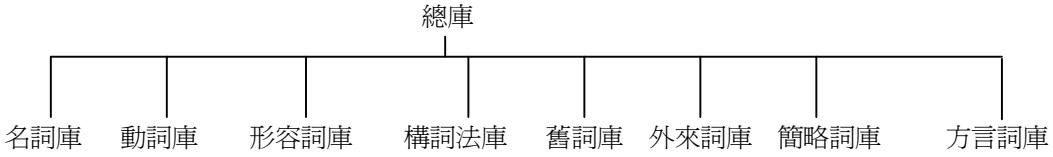
(6) 來源資訊，即該詞從那本詞典或哪些語料中來，如果很多本詞典都收錄了該詞，則說明該詞的複現率比較高，新詞語的身份更加確定。

(7) 時間資訊。該詞語大致產生的時間，以詞典的引例時間爲准；使用時間，以詞典出版時間爲准。

4.3 新詞語詞典的結構與各個庫的主要屬性資訊

4.3.1 新詞語詞典的總體結構

新詞語詞典採用成熟的關聯資料庫技術（在 access 軟體下實現）。填入的資訊儘量以直觀明瞭的漢字、字母、數位表示。根據新詞語屬性的確立，資訊庫總體上包括三個方面五個庫。總庫一個，語法資訊庫三個（名詞庫、動詞庫、形容詞庫），構詞法庫一個。另外還設立了舊詞庫、外來詞庫、簡略詞庫和方言詞庫，對新詞語當中的舊詞新用、外來詞、簡略詞、方言詞等的有關資訊進行了描述。這幾個庫通過“詞語、拼音、義項”三個欄位連接。構成一個上下連接的有機系統，便於資訊的提取。新詞語詞典的總體結構如下：



4.3.2 各個庫所描述的主要屬性資訊

總庫主要描述的資訊有：詞語、拼音、義項、詞性、音節、產生途徑、領域、時間、來源等。

其他各庫共有的屬性資訊有“詞語、拼音、義項”，均從總庫中繼承，其他屬性資訊如下。

語法資訊庫中名詞庫主要描述了與名詞搭配的各種量詞，名詞的子類，能不能直接受數詞、數量詞、其他名詞、動詞的修飾，能受哪些代詞直接或加“的”後修飾，前接或後接成分，能不能作定語、主語、賓語，能不能直接或加“地”作狀語，以及能不能重疊、臨時充當量詞等。動詞庫主要描述的資訊有：動詞的子類——係詞、助動詞、趨向動詞、補助動詞、形式動詞、自主動詞、非自主動詞、內外動詞、存現動詞、離合詞等；構成的句式——“把”字句、“被”字句、兼語句、雙賓句、存現句等；充當的成分——定語、名詞性結構的中心語、單作謂語、賓語、狀語；後帶的成分——體謂准賓語、動時量補語、結果補語、趨向補語等；動詞自身形態的變化——前受“不、沒、很、正”的修飾、後跟“著、了、過”、VV、AABB、V 一 V、V 了 V、V 了一 V、VVO 等。賓語、結果補語、趨向補語的詳細資訊將另行描述。形容詞庫描述的主要資訊有：子類、

直接作定語或加“的”後作定語、作謂語、補語、狀語或加“地”後作狀語或再加“很”後作狀語、作準謂賓、“有”的賓語、名詞性結構的中心語、AA 重疊及重疊後的詞性、ABAB、A 裏 AB、帶“著了過”、准賓語、趨向補語等。

構詞法庫描述的主要資訊有：1、構詞部件，分爲“成分 1”“成分 2”“成分 3”，分別填入構成該詞語的成分的類別，其中有的是語素、有的是詞。2、構詞法，考察該詞語的構詞方式，主要分爲：主謂、動賓、狀中、定中、補充、聯合、加字首、加尾碼等。3、詞性，填入詞語的詞性，從總庫中繼承來。4、音節，填入該詞語的音節數，從總庫中繼承來。

舊詞庫描述的主要資訊有：1、舊義，填入該詞語原來的意義；2、新義，填入該詞語新意義或新用法；3、詞性，填入該新詞語的詞性，如果詞性與原詞語詞性一致，則標詞性標記；如果改變了詞性則特別標明，如不及物動詞變爲及物動詞，標爲：Vt；及物動詞變爲不及物動詞，標爲：Vi。4、詞義演變途徑，考察由舊詞語演變爲新詞語的詞義的演變途徑，主要有 36 類：（1）同用相比，（2）同果相喻，（3）同質相喻，（4）同狀相喻，（5）特定代普通，（6）具體到抽象，（7）同位相喻，（8）語素換義，（9）泛化，（10）個體代全體，（11）普通代特定，（12）使動化，（13）同感引申，（14）物件更換，（15）客體更換，（16）指稱物件擴大，（17）工具帶本體，（18）特徵、標誌代本體，（19）專化，（20）同形相喻，（21）以果代因，（22）同所相喻，（23）部分代全體，（24）動靜引申，（25）所在代（比如：山頭、大哥大），（26）主體擴大，（27）功用代本體，（28）語用（小姐），（29）社會原因（草業），（30）時空引申，（31）正反引申，（32）句法影響，（33）抽象到具體，（34）特指化，（35）本體代特徵，（36）現象代本體。（根據羅正堅的《漢語詞義引申導論》和徐國慶的《現代漢語辭彙系統論》歸納出來）5、演變類型，詞義演變的類型，主要有 9 類：（1）轉移，（2）擴大，（3）虛化，（4）轉類，（5）縮小，（6）貶降，（7）揚升，（8）弱化，（9）深化。

外來詞詞庫描述的主要資訊有：1、途徑，外來詞進入漢語的主要途徑，主要有：（1）音譯，（2）諧音，（3）音譯加漢語語素，（4）音兼意譯，（5）按照外語詞語的意義創造一個漢語詞語；2、語音變化資訊，主要有：（1）音素的替換，（2）音節的增減；3、意義變化，主要有（1）擴大，（2）縮小，（3）轉移，（4）保持原意；4、縮略，考察外來詞語是否有縮略；5、應用領域。

簡略詞詞庫描述的主要資訊有：1、原詞語，填入簡略詞的原型。2、簡略的類型：（1）簡稱，（2）縮語，（3）略語，（4）准縮略語。3、構成方式：將原詞語劃段，根據實際情況描述如何進行縮略的，比如“北京大學”—“北大”，其構成方式描寫爲“a1b3”。4、同形：如有同形詞語，則有幾個填相應的數位。5、縮略方式：（1）縮合，如：奧林匹克運動會——“奧運會”；（2）節縮，如：電視連續劇——“連續劇”；（3）提取，如：中國高技術研究發展計劃綱要——“863 計劃”（該計劃的提出是 1986 年 3 月）；（4）其他，包括：A、用同義、近義詞語替換，如：浮式起重機——“浮吊”；

B、用上位詞語代替下位詞語，如：中華人民共和國教育委員會——“國家教委”；C、用英文中的字母縮略，如“MTV”。

舊詞新用庫、外來詞庫、簡略詞庫的開發主要是爲了研究新詞語的産生途徑以及主要原因。

各個庫屬性資訊的具體描述方法，請參閱拙作《<現代漢語新詞語資訊（電子）詞典>的開發應用》、《<現代漢語新詞語資訊（電子）詞典>的結構》。

5. 近四萬詞語的歸類實踐

進行語法資訊的描述首先要對新詞語進行分類和歸類。新詞語詞典所堅持的語法理論及詞類體系繼承了北京大學計算語言學研究所《現代漢語語法資訊詞典》所堅持的語法理論和詞類體系——片語本位語法體系。針對漢語詞類多功能的特點，我們堅持以“優勢語法功能”作爲詞類劃分和詞語歸類的標準。爲了明確漢語詞類的優勢語法功能，我們以《現代漢語語法資訊詞典》爲基礎進行統計，總結出漢語詞類語法特徵的分佈狀況以及優勢語法功能。漢語詞類優勢語法功能如下：

名詞：能受數量詞修飾，能作主賓語。

時間詞：修飾名詞、直接修飾“指量名”結構、作介詞“在”的賓語。

處所詞：作“在”的賓語、直接修飾名詞構成定中結構、能用“這兒、哪兒、那兒指代。

區別詞：能加“的”後或直接修飾名詞作定語。

動詞：能受“沒”或“不”修飾、能帶時態助詞“著、了、過”、能單獨作謂語。

形容詞：作謂語、受“很”“不”修飾、作定語。

狀態詞：不受“不”“很”修飾、加“的”後修飾名詞、加“的”後修飾“數量名結構、帶“的”後作謂語。

副詞：作狀語而不作定語。

根據以上“優勢語法功能”，對新詞語進行了分類，並對四萬新詞語逐一進行了考察，歸了類。在歸類中，我們注意到：1、這些優勢語法功能不是“對內具有普遍性”，因此不能包打天下，只是具有相對的普遍性，還有一些例外，需要用其他特徵協助判斷。2、很多優勢語法功能也不是“對外具有排他性”，一些優勢功能也可能是兩類詞共有的，比如能受“不”的修飾是動詞和形容詞共同的特徵，而不能受“不”的修飾又是名詞、時間詞、處所詞、區別詞、狀態詞的共同特徵。這種情況提醒我們給新詞語分類或歸類不能按照單一標準，而要綜合運用多個標準。

按照這些標準給現有新詞語分類和歸類情況如下：

名詞有：愛蟲、愛嬌、愛意、安居工程、安樂死、按摩小姐、奧星、奧運戰略、八卦新聞、吧女、吧台、吧蠅、霸氣、白金唱片、白領、白領犯罪、白領麗人、白判、白色公害、白色收入、板寸、傍姐、保護傘、保健菜、波霸、長期飯票、超級恐龍、炒家、搞笑片、AA 制、BBS 、e-Book 、SOHO 族、等等 24874 個。

動詞有：挨宰、暗箱操作、拔份、把脈、罷網、白領化、擺平、扮蔻、扮酷、扮靚、棒殺、傍、傍大款、包二奶、包二爺、包裝、煲電話粥、煲網、保級、暴跌、保廉、暴走族、曝光、爆炒、爆棚、蹦迪、蹦極、逼宮、飆、飆車、飆價、炒樓花、觸網、豐胸、搞掂、Call 、等等 12006 個。

形容詞有：霸氣、暴露、倍兒棒、慘、火、火爆、酷、靚、帥、爽、靚麗、IN 、Q 、Cool 等等 1002 個。

區別詞有：非常、海量、候鳥型、綠色、新銳、主打、袋裝、獨資、兩栖、落地式、程式控制、等等 119 個。

副詞有：慘、好、絕、斃、倍兒、倍加、等等 43 個。

成語有：築巢引鳳、引咎辭職、招商引資、一網情深、優化組合、友情出演、友情客串、心靈雞湯、霧裏看花、閃亮登場、強強聯合、牽線搭橋、錢權交易、夢中情人、美麗凍人、快樂老家、高開低走 236 個，

慣用語有：愛誰誰、愛情走私、愛心大放送、把蛋糕做大、別理我，煩著呢、找不著北、有沒有搞錯、玩兒深沈、沒事偷著樂、老鼠愛大米、空手套白狼、跟著感覺走、第一次親密接觸、常回家看看、瀟灑走一回、玩的就是心跳、等等 799 個。

歎詞有：哇噻、yeah2 個。

參考文獻

亢世勇．《現代漢語新詞語資訊（電子）詞典》的開發應用，《辭書研究》，2001（2）：55—63。

亢世勇．《現代漢語新詞語資訊（電子）詞典》的結構，載於《資訊網路時代中日韓語文現代化國際研討會論文集》，香港：香港文化教育出版社，2000：276—281。

亢世勇．語料庫技術在新詞語詞典開發中的具體應用，載於《中國辭書論集 2000》，北京：中國大百科全書出版社，2001：291—300。

王鐵昆．新詞語的判定標準與新詞新語詞典編纂的原則，《語言文字應用》，1992（4）：14—20。

劉一玲．尋求新的色彩，尋求新的風格——新詞語產生的重要途徑，《語言文字應用》，1993（1）：85—90。

季恒銓．新詞新語詞典編纂的新收穫，《語言文字應用》，1993（1）：77—80。

語用所“新詞新語新用法研究”課題組·整理漢語新詞語的若干思考,《語言文字應用》，1993（3）：65—76。

張志毅、張慶雲·新時期新詞語的趨勢與選擇，《語文建設》，1997（3）：15—18。

羅正堅·《漢語詞義引申導論》，南京：南京大學出版社，1996。

俞士汶·《現代漢語語法資訊詞典詳解》，北京：清華大學出版社，1998。

徐國慶·《現代漢語辭彙系統論》，北京：北京大學出版社，1999。

姚漢銘·《新詞語·社會·文化》，上海：上海辭書出版社，1999。

於根元·《網路語言概說》，北京：中國經濟出版社，2001。

Computational Linguistics and Chinese Language Processing
Vol.7, No.2 , August 2002 , pp. 101-112

基於詞彙語義的百科辭典知識提取實驗

An Experiment on Knowledge Extraction from an Encyclopedia Based on Lexicon Semantics

宋柔*、許勇+

Song Rou, Xu Yong

摘要

本文研究百科辭典釋文信息提取方法，設計了一個基於詞彙語義屬性和關係的形式系統。在對百科辭典的詞目按語義分類的基礎上，對釋文的線性詞串進行簡單的語義屬性匹配，便可提取文本中的簡單知識。在一項百科辭典信息提取的實驗中，這一方法的有效性得到了初步的驗證。

關鍵詞：知識提取，詞彙語義

Abstract

The typical approaches to extracting text knowledge are sentential parsing and pattern matching. Theoretically, text knowledge extraction should be based on complete understanding, so the technology of sentential parsing is used in the field. However, the fragility of systems and highly ambiguous parse results are serious problems. On the other hand, by avoiding thorough parsing, pattern matching becomes highly efficient. However, different expressions of the same information will dramatically increase the number of patterns and nullify the simplicity of the approach.

Parsing in Chinese encounters greater barriers than that in English does. Firstly, Chinese lacks morphology. For example, recognition of base-NP in Chinese is more difficult than that in English because its left boundary is hard to discern.

* 北京語言大學計算機系 Beijing Language and Culture University
E-mail: songrou@blcu.edu.cn
+ 北京工業大學計算機學院 Beijing Polytechnic University
E-mail: hopexy163@163.com

Secondly, there are many stream sentences in Chinese which lack subjects and cause parsing to fail. Finally, in Chinese, the absence of verbs is also pervasive. Sentential parsing centering on verbs, which is used with English, is not always successful with Chinese.

We are engaged in research on knowledge extraction from the Electronic Chinese Great Encyclopedia. Our goal is to extract unstructured knowledge from it and to generate a well-structured database so as to provide information services to users. The pattern-matching approach is adopted.

The experiment was divided into two steps: (1) classifying entries based on lexicon semantics; (2) establishing a formal system based on lexicon semantics and extracting knowledge by means of pattern matching.

Classification of entries is important because in the text of the entries of different categories there are different kinds of patterns expressing knowledge. Our experiment demonstrated that an entry of the encyclopedia can be classified precisely merely according to the characters in the entry and the words in the first sentence of the entry's text. Some specific categories, e.g., organization names and Chinese place names, can be classified satisfactorily merely according to the suffix of the entry, for suffixes are closely related with semantic categories in Chinese.

The formal system designed for knowledge extraction consists of 4 kinds of meta knowledge: concepts, mapping, relations and rules, which reflect lexicon semantic attributes. The present experiment focused on the extraction of knowledge about various areas from the texts regarding administrative places of China (how large is a place or its subdivisions). The results of the experiment show that the design of the formal system is practical. It can accurately and completely denote various expressions of simple knowledge in a Chinese encyclopedia. However, when the focus of knowledge changes, e.g., from administrative areas to habits of animals, it is a labor-intensive task to renew the formal system. Therefore the study of auto or semi-auto generation of this kind of formal system is required.

1. 問題背景

以信息技術爲基礎的在線知識服務是信息產業的發展方向。目前已經出現具有初步實用價值的在線知識服務，其中最具生命力的是問答服務（Q&A），而這一服務的關鍵技術

之一是文本知識的自動提取，它是爲各種各樣的問題提供答案的基礎。

目前，網絡技術的發展使得人們能輕易地獲取幾乎無窮無盡的文本。但由於網絡文本範圍太廣，涉及的語言現象太複雜，全自動、高準確率的信息提取和知識提取困難較大，短期內難以實用。百科辭典是一種受限的文本，知識含量高，知識表述比較規範。無論是從理論的角度看還是從應用的角度看，從百科辭典中自動獲取知識可當作文本知識自動提取的突破口。

文本知識提取的方法主要有兩種：基於語句分析的方法和基於模式的方法[Tsujii, J. 2000]。理論上說，文本知識的提取需要在徹底理解的基礎上進行。因此，句法分析和語義分析技術很自然地使用於這一領域，但它有脆弱性和多歧義問題。基於模式的方法可以避免對語句進行徹底分析，效率較高，但同一信息的不同表達形式會使模式數量大爲膨脹。[Tsujii, J. 2000，Hull, R. *et al.* 1999 and Soderland, S. G. 1996]處理的是英語或日語文本，主要使用基於語句分析的方法。其中[Hull, R. *et al.* 1999]的工作是基於大容量的知識庫，採用部分分析、語義解釋和推理的步驟；[Soderland, S. G. 1996]也是進行句法分析，包括名詞短語分析、同位關係識別（Appositive Recognition）和同指消解（Coreference Analysis）；[Tsujii, J. 2000]本身的工作主要使用語句分析的方法，但吸收了模式方法的優點。

漢語文本的知識提取使用語句分析方法比英文問題更大。首先是因爲漢語缺乏形式標誌。比如基本名詞短語的識別在英語中並不困難，但在漢語中由於難以確定其左邊界而識別率較低。其次，漢語常有缺主語的流水句，會造成句法分析的失敗。此外，英語句子的句法分析和語義分析一般都以動詞爲核心，而相當一部分漢語的句子沒有動詞，如“昌平縣面積 1352 平方公里，人口 43 萬”。如果照搬英語中的做法做句法分析（或淺層句法分析）、找動詞的語義格，其效果不會好。

漢語文本知識提取的工作已發表的並不多。[Gu, F. *et al.* 2001]的工作也是從百科辭典中提取知識，它的結果是一個框架結構的知識庫，可以提供實用的知識服務。但爲了得到這個知識庫，需要先設計一個形式語言，並用它對辭典文本進行人工標注。

本文研究漢語百科辭典的知識提取。我們的目標也是把百科辭典中的無結構的知識提取出來，生成帶結構的數據庫，向用戶直接提供知識服務。這項工作當然只能在一個受限範圍內通過人機結合的方式來完成。但是，我們希望使人的勞動集中於詞彙的語義屬性研究和詞庫中詞彙的語義屬性標注，避免人工標注語料所需的巨大勞動量。由於上述漢語分析中的困難，我們不採用常規的句法語義分析，而嘗試關鍵詞語爲核心的模式匹配的方法，其中關鍵詞語不一定是動詞，但具有信息提示的功能（如“面積”提示其後面有關於面積數量的信息），模式匹配主要依靠詞語的語義屬性。

我們的處理對象是《中國大百科全書》（光盤版），工作步驟是：（1）根據詞目確定題材類別，根據題材類別確定知識提取的目標；（2）建立基於詞彙語義的形式系統，用詞語模式匹配的方法提取知識。本文介紹了相關研究的一些實驗，測試結果證明這一方法是有效的。

2. 百科辭典詞目的分類

2.1 百科辭典詞目按題材分類的試驗

爲了提取知識的方便，首先需要把按領域分卷的《中國大百科全書》中的詞目進行分類。

這裏所說的詞目的類別，不是按專業領域劃分，而是按題材劃分的。比如，人物和概念是不同的題材。《中國大百科全書》美術卷中人物"徐悲鴻"釋文與數學卷中人物"華羅庚"釋文的風格相似，所表達的信息內容的類型十分接近，但與同在美術卷中概念"油畫"的釋文風格和信息內容的類型完全不同。

題材的異同取決於詞目的語義類。所有類別的釋文第一句話總是對於詞目給出一個概括性的說明，指出它的最重要的特徵，如人物的國籍和歷史地位，行政區劃的行政隸屬和政治經濟地位，動物的目科屬種等。第一句話以後，不同語義類的釋文有不同的信息內容。比如，人物的釋文包括人物的生卒時間和地點、生平事蹟、主要成就等，行政區劃的釋文包括該地區的面積、人口、沿革、地形、氣候、經濟、特產、名勝等，動物的釋文包括動物的體形、各部位的形狀大小顏色、分佈區域、生活習性、繁殖方式、與人類的關係等。

從百科辭典知識提取的使用目標出發，我們目前採用的詞目分類系統中的大類是人物、行政區劃、自然地理、動物、植物、機構組織、事件、裝置、其他。之所以採用這樣的分類體系，一是因爲這些類的詞目和釋文有比較明顯的特徵，知識抽取相對容易；二是因爲這些類在整個百科辭典中所占比重較大，詞目較多，有條件使用統計方法進行信息提取。有些大類下面還要分小類，如自然地理類中包括山脈、河流、湖泊、沙漠、島嶼等等，分小類的目的是使同類釋文的信息特徵更加一致。

我們使用現代漢語通用分詞系統 GWPS 的專名識別功能將詞目中的人名、地名（包括行政區劃名、自然地理名、古地名、景點設施名）、機構名挑選出來，實驗對象是美術卷、外國文學卷、世界地理卷、中國地理卷。我們只使用詞目內部的用字信息和釋文第一句話最後兩個詞的信息，識別結果如下：

	實有詞目	識別詞目	遺漏	誤識	準確率	召回率
美術卷人名	935	935	1	1	99.9	99.9
外國文學卷人名	2470	2471	0	1	99.96	100
世界地理卷地名	1153	1154	5	6	99.5	99.6
中國地理卷地名	1498	1500	0	2	100	99.9
美術卷機構名	98	98	0	0	100	100
合計	6154	6158	6	10	99.8	99.9

如果不用釋文信息，只用詞目內部的用字信息，則對機構名和中國地名影響不大，對人名和外國地名來說召回率大大降低。如此時美術卷人名詞目識別結果爲：

實有詞目	識別詞目	遺漏	誤識	準確率	召回率

935	779	166	10	98.7	82.2

原因是有些人名詞目使用的是法號（如“法常”）、綽號（如“泥人張”），有些是 GWPS 不具有識別能力的日本人名（如“奧村土牛”）。人名釋文的首句最後一個詞絕大部分是“身份詞”（“畫家”、“建築師”等），大部分首句前還帶有說明生卒年代的括號，所以利用了釋文首句的信息後，召回率大大提高。

地名詞目中，不同小類的詞目的釋文風格差別仍然很大。比如，行政區劃名釋文的主要信息是行政隸屬關係和政治經濟地位、面積、人口、沿革、地形、氣候、經濟、特產、名勝等，自然地理名的釋文中沒有這些內容。行政區劃名和自然地理名中還需分更小的類，因爲行政區劃中，關於國家的釋文同關於城市的釋文在詳盡程度上很不同，信息內容的類型上也有區別。自然地理名中，關於山脈的要介紹山脈地理分佈、走向、山峰高度、地質歷史等，關於河流的要介紹河流發源地、走向、流域面積、經濟功能等。爲此，我們對於中國地理卷中的地名進行了細分類試驗。對於行政區劃詞目，分爲省、自治區、地區、自治州、市、區、縣（包括自治縣）、鎮，共 8 類；對於自然地理詞目，挑選出江河、湖泊、山嶺、山脈、盆地、沙漠、平原、高原、丘陵、草原、島嶼，共 11 類。我們試驗完全依據詞目後綴進行識別。行政區劃詞目所用後綴和識別結果如下：

類名	省	市	地區	自治州	區	縣	鎮	合計
後綴	省	市	地區	自治州	區*	縣	鎮	
實有	23	385	5	9	22	275	36	755
標識	23	385	5	9	25	275	36	758
誤識	0	0	0	0	3	0	0	3
漏識	0	0	0	0	0	0	0	0
準確率%	100	100	100	100	88	100	100	99.6
召回率%	100	100	100	100	100	100	100	100

注：“區”類要從後綴“區”中去掉後綴“自治區”、“地區”、“風景區”、“風景名勝區”、“自然保護區”、“灌區”。該類的 3 個誤識錯誤是“皖西山區”、“皖南山區”、“神農架林區”。簡單地把“山區”當作後綴從“區”類中去掉是不行的，因爲上海有“寶山區”，北京曾有“燕山區”，等等。

自然地理詞目所用後綴和識別結果如下：

類名	河流	湖泊	山嶺*	山脈	島嶼	盆地	沙漠	平原	高原	草原	丘陵	合計
後綴	江,河*, 溪, 水*	湖, 錯, 池, 海*	山, 峰,	山脈	島*	盆地	沙漠*	平原	高原	草原	丘陵	

			嶺									
實有	144	65	162	19	20	14	5	19	11	2	8	466
標識	137	59	162	19	20	14	5	19	11	2	8	456
誤識	3	1	1	0	0	0	1	0	0	0	0	6
漏識	7	7	1	0	0	0	1	0	0	0	0	16
準確率%	97.81	98.31	99.38	100	100	100	80	100	100	100	100	98.68
召回率%	93.06	89.23	99.38	100	100	100	80	100	100	100	100	96.57

注：以“河”爲最後一個字但不是河流的詞目是“三河”；以“水”爲最後一個字但不是河流的詞目是“中國的地表水”和“中國的地下水”；漏識的河流是以“布”和“曲”爲後綴的西藏地區河流；以“海”爲湖泊詞目的後綴，需要人爲地去掉渤海、黃海、東海和南海，但仍然有一個誤識：“中國的近海”；“山嶺”類的誤識是“中國的火山”，漏識是“神農頂”；“島嶼”類要從後綴“島”中去掉後綴“半島”；漏識的湖泊是“月亮泡”、“大布蘇泡”和以“茶卡”爲後綴的西藏地區鹹水湖；誤識的沙漠是“中國的沙漠”，漏識的沙漠是“毛烏素沙地”。

2.2 關於詞目分類方法的結論

我們的試驗說明，僅根據詞目的用字構成和詞目釋文的首句用詞，就可以對於百科辭典詞目的主要題材類別進行分類，準確率和召回率可達到實用要求。對於某些類別，比如機構名和中國地名，則僅使用詞目後綴就能達到相當好的識別效果，其原因是漢語後綴成分與語義類別緊密相關。

3 百科辭典釋文知識提取實例

3.1 一個基於詞彙語義屬性的形式系統

我們把處理對象限定爲行文規範的百科辭典，目前只提取比較易於形式化的信息。我們的基本思想是：建立起一個基於詞彙語義的屬性和關係的形式系統，其中的屬性和關係同欲提取的信息緊密相關；使用屬性模式匹配的方法在線性詞串中提取信息。

我們首先做的是中國行政地名詞目釋文中面積信息的提取。

大部分面積信息的表述中有“面積”二字，但是在成串的說明中，有省略的情況；“填海”、“種植”等動詞帶數詞和面積量詞表示面積的情況下，有時也不使用“面積”。如關於

香港的釋文中有：

陸地面積 1071.8 平方公里。其中香港島 75.6 平方公里，九龍 11.1 平方公里，“新界”（包括大嶼山島等周圍 230 多座島嶼）975.1 平方公里，另新填土地 9.2 平方公里。

此外，中國行政地名詞目的釋文中，有 4 處“面積”的錯別字：2 處錯成“南積”，1 處錯成“面和”，1 處錯成“面只”。

作爲信息提取的初步研究，我們只考慮出現“面積”二字時的情況。

在 755 個中國行政地名詞目的釋文中，“面積”出現了 1668 次，其中 38 個“大面積”和 2 個“單位面積”用作修飾成分，如“形成眾多的鹽湖和大面積沼澤”和“樹木種類多,單位面積蓄積量高”，其餘 1628 處“面積”確實表達面積信息。

利用面向語言教學研究的文本檢索工具 CCRL 作爲輔助工具,我們用人工分析研究了這些“面積”的上下文。

與“面積”相關且帶有數值的信息可以看成是某些關係：

數量關係。論元爲主體、數值、度量單位。如“海壇島面積 323 平方公里”，“海壇島”爲主體，“323”爲數值，“平方公里”爲度量單位。

比例關係。論元爲分子主體、分母主體、比例數。多比例關係則涉及多個比例主體和多個比例數。如“青海……天然草場面積約占全省土地總面積的 46.39%”，“天然草場”爲分子主體，“全省土地”爲分母主體，“46.39%”爲比例數。

變化數量關係。論元爲主體、擴縮標記、數值、度量單位。如“……城區面積擴大了 15 平方公里”，“城區”爲主體，“擴”爲擴縮標記，“15”爲數值，“平方公里”爲度量單位。

變化比例關係。論元爲主體、擴縮標記、倍數或比值。如“貴州……茶園面積較 50 年代初擴大 20 多倍”，“茶園面積”爲主體，“擴”爲擴縮標記，“20 多倍”爲倍數。

變化數量關係和變化比例關係還應當涉及變化前時間和變化後時間。變化前時間往往顯式地給出，變化後時間有時省略，其實就是百科全書資料收集的時間。如上面最後一例，變化前時間是“50 年代初”，變化後時間爲百科全書資料收集的時間，文中省略。數量關係和比例關係也應當涉及時間，被省略的時間也是百科全書資料收集的時間。這些關係往往帶有修飾成分，如“***約*** 10 公頃”，“***不到*** 30%”，“5 倍***以上***”，“***擴大至*** 23 平方公里”等。這些也應當作爲論元加入到各關係中。

信息提取的任務就是確定這些關係中的論元在文本中所指的內容。其中，數值、比例數、倍數或比值、度量單位、擴縮標記、修飾比較容易確定，因爲它們形式規範，位置比較固定，而且後三者的集合基本上是封閉的。時間論元也有形式標記，包括“世紀”、“年代”、“年”、“月”等，表示朝代或事件的詞語後面加上“初”、“末”、“前”、“後”、“期間”等時間方位詞。確定時間論元的主要困難在於出現位置不固定。我們的策略是從其它論元出現的位置往前看 6 個逗號或句號，找到了時間論元特徵就可以提取出來，找不到就歸結爲省略，即時間論元是百科全書資料收集的時間。

最大的困難在於各種面積主體的確定。爲此，我們從實例中提取了一個基於詞彙語

義屬性的形式系統，它的內容包括 4 類元知識：

概念：

行政區劃 xq，往往是當前詞目本身，也可能是當前詞目所代表的行政單位的上級單位。

詞目替代詞 td，包括“省境”、“全省”、“市境”、“全市”、“區境”、“全區”、“縣境”、“全縣”。

行政區劃的分部 fb，包括“市區”、“城區”、“郊區”、“海域”、“陸域”、“陸地”，還包括方位分部如“東部”、“西北部”。

具有面積屬性的名詞性詞語 mc，包括“草原”、“平原”、“耕地”、“土地”、“陸地”、“森林”、“荒地”、“荒山”、“喀斯特地貌”、“茶園”、“桑園”、“果園”等。（注：“山嶺”、“山脈”、“河流”等不具有面積屬性。）

具有面積屬性的動詞性詞語 dc，包括“種植”、“播種”、“養殖”、“淡水養殖”等。

與具有面積屬性的動詞關聯的名詞性詞語 md，如與種植和播種關聯的有“作物”、“經濟作物”、“糧食作物”，以及具體的作物名稱“水稻”、“小麥”、“棉花”、“茶葉”、“菸草”、“甜菜”、“橡膠”等；與養殖有關的有“魚”、“蝦”等。

具有面積屬性的專名 zm，其類型包括“農場”、“林場”、“風景區”、“自然保護區”以及各種建築物等。

行政區劃類型 xl，包括“省”、“自治區”、“市”、“地區”、“自治州”、“區”、“縣”、“鎮”。

映射：

{td→xq}，由詞目替代詞到詞目本身，如在“江蘇省”釋文中，“全省”映射爲“江蘇省”。

{xq→xl}，由行政區劃名到它本身的行政區劃類型，如由“江蘇省”映射爲“省”。

{xq→xq}，由行政區劃名到它的上級行政區劃名，如由“蘇州市”映射爲“江蘇省”。

{md→dc}，由名詞到與它關聯的具有面積屬性的動詞，如由“棉花”映射爲“種植”。

{mc→mc}，由名詞到它的上級語義名詞，如由“糧食作物”映射爲“作物”。

關係：

數量關係：sl(time, body, number, area-unit, modifier)，即時間、主體、數值、面積單位、修飾成分滿足數量關係。

比例關係：bl(time, body-numerator, body-denominator, ratio, modifier-before, modifier-after)，即時間、分子主體、分母主體、比值、前修飾成分、後修飾成分滿足比例關係。

變化數量關係：bsl(time-before,time-after, body, extend-reduce, number, area-unit, mordify)，即變化前時間、變化後時間、主體、擴縮標記、數值、面積單位、修飾成分滿足變化數量關係。

變化比例關係：bbl(time-before,time-after, body, extend-reduce, ratio, mordify) ，即變化前時間、變化後時間、主體、擴縮標記、比值、修飾成分滿足變化比例關係。

其中面積主體 body 的構成方式爲：

xq [fb[fb]] [(zm | mc | md {md→dc}md)]

式中 (｜) 表示選擇，[ ] 表示可有可無。

規則：

規則的作用就是從文本中的適當位置抽取關係中論元所指的內容。規則的形式是：文本模式→關係，其中文本模式列出關係中各論元所指內容在文本中的相對於“面積”的位置。同一個規則中的同一個變元若重複出現，則代表同一個內容。

下面列出一些常用的規則。其中，text-begin 表示篇首，dot-comma 表示句號或逗號，no-area-string 表示一個句串，其中不出現“面積”。string 表示一個句串，其中每個標點句的首詞不帶有 xq、fb、mc、md、zm、time 屬性，而且句串中包含的標點句不超過 6 句。我們這裏所說的句串就是一串標點句，而標點句就是文本中以逗號、句號、分號、嘆號、問號分隔的字串。

text-begin no-area-string dot-comma [總] 面積 [modifier] [爲] number area-unit

→sl(nil, xq, number, area-unit, modifier)

例如：“阿克蘇市”釋文的開始幾句是：

新疆阿克蘇地區轄市和行署駐地,新疆重點墾區。位於塔里木盆地西北部。面積 1.83 萬平方公里,人口 38.13 萬。

匹配規則的條件部分後，得到的數量關係是：

sl(nil，阿克蘇市，1.83 萬，平方公里，nil)

這個關係的 5 個數據“nil”、“阿克蘇市”、“1.83 萬”、“平方公里”、“nil”分別存放在 sl 數據庫的 5 個字段 time、body、number、area-unit、modifier 下，表示在該百科辭典編制時阿克蘇市面積恰爲 1.83 萬平方公里。

dot-comma time string fb 面積 [modifier] [爲] number area-unit

→sl(time, xq td, number, area-unit, modifier)

例如：“安順市”釋文中有：

20 世紀 50 年代以前，城區面積僅 1.4 平方公里，人口 2.4 萬人。

匹配規則的條件部分後，得到的數量關係是：

sl(20 世紀 50 年代以前，安順市城區，1.4，平方公里，僅)

這個關係的 5 個數據“20 世紀 50 年代以前”、“安順市城區”、“1.4”、“平方公里”、“僅”分別存放在 sl 數據庫的 5 個字段 time、body、number、area-unit、modifier 下，表示在 20 世紀 50 年代以前安順市城區面積僅 1.4 平方公里。

dot-comma td string mc 面積 [modifer-before] 占 [td][[總]面積] [的] ratio [modifier-after]

→bl(nil, {td→xq}td mc, {td→xq}td , ratio, modifier-before，modifier-after)

例如：“安達市”釋文中有：

市境地形平坦,平均海拔 150 米。草原面積占 51.5%以上，宜發展畜牧。

匹配規則的條件部分後，得到的比例關係是：

bl(nil，安達市市境草原，安達市市境，51.5%，nil，以上)

這個關係的 6 個數據“nil”、“安達市市境草原”、“安達市市境”、“51.5%”、“nil”、“以上”分別存放在 bl 數據庫的 6 個字段 time、body-numerator、body-denominator、ratio、modifier-before，modifier-after 下，表示在該百科辭典編制時安達市市境草原面積占安達市市境面積 51.5%以上。

由於這些被提取出來的信息以關係數據庫的形式存放，所以可以借助數據庫檢索工具來檢索。

4. 測試與討論

我們檢查了中國行政地名詞目按漢語拼音排序 a-d 的 107 個詞目的釋文，這裏面出現“面積”176 次，帶有數量的面積信息 153 條，其中有些是行政區劃本身的面積，有些是行政區劃內部某個分區的面積。使用該系統的規則能夠正確提取信息的 141 條，準確率約爲 92%。其中，上述第 1 條規則使用 94 次，第 2 條規則使用 22 次，全部正確。特別是，107 個詞目釋文中，有 103 個提到了該詞目所代表的行政區劃的面積，它們都出現在靠近篇首的位置，其中 102 條可以用規則將面積信息提取出來，94 條用上述第 1 條規則，8 條用第 2 條規則。發生錯誤的大都是行政區劃內某一部分中某種特定地域的面積，主要問題是面積主體過於複雜。如“安徽省”釋文中有：

皖中丘陵水旱作物過渡區。以水稻、小麥爲主的水旱兼作、一年兩熟區。位於淮河以南、江淮分水嶺—滁河一線以北，土地面積占全省 23.7%，……

該例中，第一句是個小標題，後面幾句是對該標題所涉地區的說明。最後一句中的“面積”的主體是“安徽省皖中丘陵水旱作物過渡區土地”。這一主體的構成方式過於複雜，難以識別。

從這一實驗中可見，

(1) 由概念、映射、關係和規則組成的形式系統可以比較全面準確地表示一些簡單知識在百科辭典文本中的形式，這一個基於語義屬性的形式系統的框架設計是成功的。

(2) 爲了構造這一形式系統需要做大量的人工調查、分析、標注工作。若信息提取的焦點不變而僅僅換掉文本（把中國大百科全書換成其他百科辭典文

本），則由於各種百科辭典中同種知識的表達形式基本上是有限多種，所以增大的人工工作量不會太大，這是這種做法的優越性所在。但當信息提取的焦點改變（比如從提取行政區劃的面積知識轉而要提取動物的生活習性知識）時，人工的投入量仍然會相當大。爲此，必須研究這類形式系統自動（或半自動）生成的方法，這將是我們的下一步工作。

鳴謝

本文得到中國國家自然科學基金（60141001）和國家高技術計劃（2001AA114111）的資助，謹在此致謝。

參考文獻

Tsujii, J., “Generic NLP Technologies: Language, Knowledge and Information Extraction”, *Proc. of ACL2000*, 2000, pp.11-18.

Hull, R., and Gomez, F., “Automatic acquisition of biographic knowledge from encyclopedic texts”, *Expert Systems with Applications* 16(1999), pp.261-270.

Soderland, W. D., Fisher, J. Aseltine, and W. Lehnert, “CRYSTAL: Inducing a Conceptual Dictionary”, *Proc. of the International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995, pp. 1314-1319.

Gu, F. and Cao, C., “Biological Knowledge Acquisition From the Electronic Encyclopedia of China”, *Proc. of ICYCS’2001*, 2001, pp.1199-1203.