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。

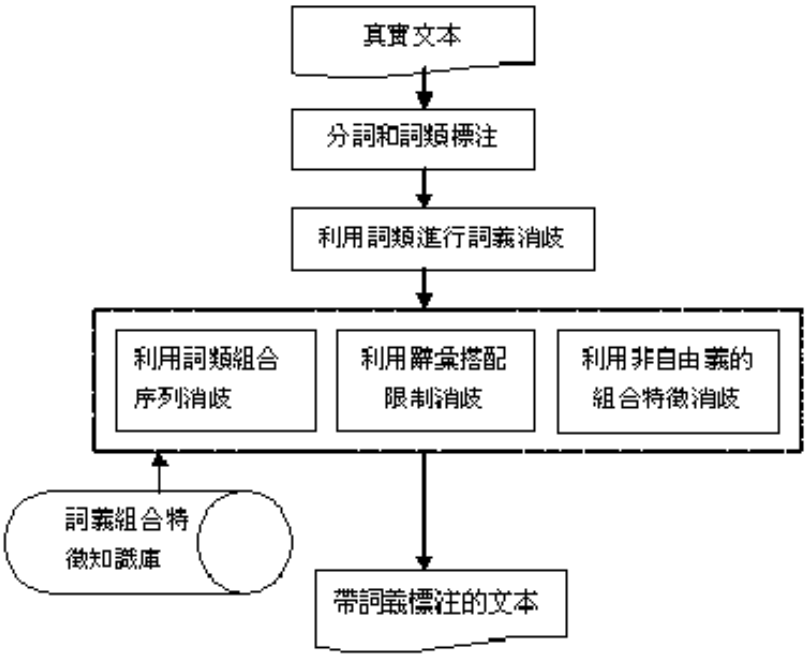
利用這種辦法，電腦迅速指出 67 個“樓”中有 23 個表示❷義。經檢查只有下面 1 例錯誤，其他全部正確。

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