Computational Linguistics and Chinese Language Processing
Vol. 7, No. 1, February 2002, pp. 47-78

一種基於知網的語義排歧模型研究

A Study of Semantic Disambiguation Based on HowNet

楊曉峰*，李堂秋*

Yang Xiaofeng, Li Tangqiu

摘　要

本文提出了機器翻譯中句法分析的一種語義排歧模型，該模型以《知網》爲主要語義知識源。《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,它爲我們的排歧提供了豐富的語義資訊。排歧模型結合了基於規則及基於統計的方法，應用於分析所產生的中間結構中，從“優選”的角度進行詞義及結構的排歧。

排歧模型首先利用大規模的語料庫獲取義原的同現集合，該語料庫未進行任何的語義標誌，因此獲取過程是無指導的。然後它根據轉換模板構造出義原的語義限制規則。《知網》中的詞語義項由義原組成，義項的語義限制規則可以由其構成義原的語義規則得到。

在語義排歧階段，我們首先確定輸入句的每個實義詞的上下文相關詞集。由於實義詞的語義關係在對當前句子的語法結構確定及各詞語詞義的選擇起著相當重要的作用，我們對一個句子的評價就建立在對該句中實義詞的評價基礎之上。把詞語的當前上下文相關詞集與詞語各義項的限制規則所描述語義特徵資訊進行比較，根據比較的相似度選擇最合適的義項。同時將相似度的最大值作爲該詞語的評價值。中間分析結果中各實義詞的評價分值可以成爲評價此中間結果的依據，以此在多個中間結構中選出最佳的結果。這樣，我們在解決詞義歧義的基礎上同時也解決了結構歧義。

本文所提出的語義排歧模型已在機器翻譯系統中具體地實現。實驗例句的測試表明該排歧模型對解決句法分析中的辭彙歧義、結構歧義是有效的，並且優於傳統的 YES/NOT 的方法。

* 廈門大學計算機系，廈門，361005
Department of Computer Science, Xiamen University, Xiamen ,361005

本文首先提出了排歧模型的主要思想，並簡要介紹了《知網》。然後給出了從語料庫中抽取義原同現資訊及將其轉化成語義限制規則的方法。接著文章詳細介紹了排歧演算法，包括構建上下文相關詞集，義原間、語義規則和上下文詞集間的相似度計算。最後文章給出了模型的試驗實例結果。

關鍵字: 語義排歧、知網、中間語言、相似度、模式匹配、語料庫、語義限制規則、語義環境

Abstract

This thesis presents a description of a semantic disambiguation model applied in the syntax parsing process of the machine translation system.

The model uses *Hownet* as its main semantic resource, which is a common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. It can provide rich semantic information for our disambiguation.

The model makes the word sense and structure disambiguation in the way of "preferring". "preferring" is applied in the results produced by the parsing process. It combines the rule-based method and statistic based method.

First we extract from a large the co-occurrence information of each sense-atom. The corpus is untagged so the extracting process is unguided. We can construct restricted rules from the co-occurrence information according to certain transfer template. The semantic entry of a word in the Hownet is made of sense-atoms, so we can make out the restricted rules for each entry of any word.

During the course of disambiguation, the model constructs the context-related words set for each notational word in the input sentence. The semantic collocation relations between notional words can play a very important role in the syntax structure disambiguation. Our evaluation of some candidates is based on the degree of tightness of match between notional words in the structure. We compare the context-related words set of the word in the current structure with all the restricted rules of the word in the lexicon, and find the best match. Then the entry with the best match is taken as the word's explanation. And the degree of similarity shows how the word in the structure matches with other notional words in it, so it can be taken as the reference of the notional words. Because the discrepancy of different candidate parses of a structure, the same word has different content-related words set, and so will get different scores. We can calculate the best match according to

the score of all the notional words of the sentence. In this way we can solve the most of word sense disambiguation and structural disambiguation at the same time.

The semantic disambiguation model proposed in this thesis has been implemented in MTG system. Our experiment shows that the model is very effective for this purpose. And it is obviously more tolerant and much better than traditional YES or NO clear cut method.

In this thesis we first put forward the general idea of the method and give a brief introduce to *the Hownet Dictionary*. Then we give the methods of extracting co-occurrence information for each sense-atom from the corpus and transferring this information to restricted rules. Then the algorithm of disambiguation is proposed with detail, which includes constructing context-related words set, the calculation of the similarity between atom-senses, and between restricted-rules and the context-related sets. The experiment result given in the end of the paper shows that the method is effective.

Keywords: Word Sense Disambiguation, *Hownet,* InterLigua, Sense Atom, Corpus, Semantic Environment

1. 前言

1.1 文本分析的歧義消解問題

歧義是自然語言中普遍存在的現象。其研究可追溯至古希臘時期的亞裏斯多德，他在《工具論・辯謬篇》中就探討了自然語言的歧義問題。1930 年，恩普森(W.Empson)發表了《歧義的七種類型》(Seven Types of Ambiguity)一書，開始從語言理論的角度研究歧義問題。科艾(J.G.Kooij)於 1971 年發表的《自然語言的歧義》則標誌著自然語言的歧義進入了系統化的研究階段。在現代語言學的發展史上，“歧義問題總是成爲某個新的語言學派崛起時向傳統陣地進擊的突破口”[呂叔湘 1984, 馮志偉 1995]

歧義就是同一形式與不同的意義產生聯繫。在自然語言處理中，歧義是一個不能回避而且也無法回避的問題，它成爲自然語言的自動分析的巨大障礙之一。

漢語的歧義一般可以分爲以下兩種類型[馮志偉 1995, 苑春法 等]：

(1) 詞的多義，又稱辭彙歧義，即同一詞語可能具有多個不同的義項；如“打”一詞在“打字”、“打酒”、“打球”、“打地基”、“打人”中就有不同的意義；

(2) 短語的同形異構，又稱結構歧義，即同種組合卻含有不同的句法功能結構。如“VP+的+是+NP”就是一個有歧義的結構：

“扮演的是一個演員”

這句話可以理解爲“一個演員扮演了劇中某個角色”(“扮演的”是施事)，也可以理解爲“被扮演成一個演員”(“扮演的”的是受事)。

還有“N1+N2+N3”，可以被理解爲((N1+N2)+N3)，也可以理解爲(N1+(N2+N3))。這類的歧義結構在漢語中有很多，它一直是語法學家研究的熱點問題。

在機器翻譯中，辭彙歧義表現爲譯文會有多種的選擇，而結構歧義表現爲句法分析中，一個詞語或片語可能會產生一個以上結構不同的分析結果。

歧義在特定的語義及常識下，並不一定都能夠成立，例如在“打球”中，根據“打”的受事物件我們可以知道“打”只能選擇“Play”的譯文；而在“反對的是戰爭”中，我們也可以知道“戰爭”是反對的受事體而不可能是施事體。消除這類不符合語義知識的歧義過程稱爲歧義的消解，也稱排歧。我們在機器翻譯的句法分析過程中，必須要引入語義的知識才能夠更好地完成歧義的自動消解。

1.2 歧義消解的方法

歧義消除的方法多種多樣，有的方法在辭彙量、詞法結構和句型上對源語言文本進行限制，從而避免大多數的分析及選詞上的歧義；有些學者提出利用語義關聯網進行排歧[Dan Roth 1998]；還有的研究人員嘗試了基於學習的自動消歧方法，如 DAN ROTH 使用 Winnow 學習方法來進行拼寫校正、語料標注[趙鐵軍 等 2000]。面前市場上流行的“雅信 CAT 漢英雙向翻譯系統”，則採用了用戶互動式消歧方法，讓用戶自己來決定結構及譯詞的選擇。

辭彙消歧是結構消歧的基礎。大多數的排歧方法都是以多義詞的詞義排歧爲切入點。短語、句子、篇章都是由最基本的詞語構成，如果一個句法結構中的詞語意義尚且不能確定，整個結構的意義的把握更無從談起了。同時，詞義選擇需要足夠多的語義知識及上下文知識，而這些知識也爲進一步解決結構歧義問題提供了良好的依據。本文的研究也主要是從詞義消歧入手，在詞義消歧的過程中進行結構的消歧。

詞義消歧方法分爲三類：基於 AI 的方法，基於知識的方法，基於語料庫的方法。按詞義消歧的智慧程度又可分爲有指導與無指導的方法[Wilks *et al.* 1998, Philip *et al.*, 董振東 等]。

基於 AI 的方法包括符號主義方法和連接主義方法。如利用神經網路等進行詞義選擇(Collins)。這類的方法在實際上對語言理解並不實用。

基於知識的方法主要包括基於義類詞類詞典和基於規則的方法。前者的代表爲西班牙研究者基於 WordNet 提出的應用概念密度的詞義消歧方法。而後者則在基於轉換的機器翻譯系統中被廣泛地使用，如 Wilks 提出的應用選擇限制來詞義消歧。

基於語料庫的方法分爲基於統計和基於實例的兩種方法。基於統計的方法經常統計詞與詞、詞義與詞義的搭配，利用搭配消除歧義。基於實例的方法是根據輸入句與實例

的相似度計算來選擇最佳的匹配。這類方法中比較成功的系統有新加坡的LEXAS(Hwee)。

以上的方法各有其優缺點。基於知識的方法可以很好地處理確定的，大粒度的知識，語義和語法的知識比較豐富，但這些知識通常由專家組織，因此有很大的主觀性，並且知識的一致性、擴充性、完備性都難以很好地實現；而基於統計的方法可以較好地處理語言中的不確定的、小粒度的知識，靈活性好，但卻難以反映自然語言中具有普遍性的語法規律和語義知識。

1.3本文的主要工作

本文主要研究目標是在機器翻譯的文本分析中如何引入語義知識進行有效的詞義消歧，進一步進行結構消歧。在課題的研究過程中，本文主要在以下幾個方面進行了探討：

1：利用《知網》爲語義知識源，從《知網》中抽取出必要的語義資訊，並將之轉化成爲方便系統實現的表示結構。

2：從基於優選的角度對分析生成的中間語言進行排歧處理。本文先利用大規模的語料庫獲取義原的同現集合，並根據轉換模板構造出義原的初始限制規則，再通過手工的方式對初始規則進行修改與調整，以得到一個較完善的規則集。義項的語義限制規則可以由其構成義原的語義規則得到。排歧演算法將義項的語義規則與義項所在的語義環境進行相似度的計算，並根據計算結果進行義項選擇和結構的語義搭配的評價，從而進行詞義排歧與結構排歧。本文提出了這種排歧方法的詳細演算法與實現步驟。

3：將上述的思想及演算法在機器翻譯系統中具體地實現。

2. 系統結構與基礎知識

2.1 系統結構描述

結構消歧的難度很大，各種各樣的歧義結構還有待語法學家的進一步發現與總結。而辭彙消歧是結構消歧的基礎。短語、句子、篇章都是由最基本的詞語構成，如果一個句法結構中的詞語意義尚且不能確定，整個結構的意義就無法把握，排歧則更無從談起了。同時，詞義選擇需要足夠多的語義知識及上下文知識，而這些知識也爲進一步解決結構歧義問題提供了良好的依據。本文中語義排歧模型的指導思想就是首先解決多義詞的辭彙歧義，然後在此基礎之上進行結構歧義的消解。爲此本文中提出了一種語義排歧的模型，該模型以《知網》爲主要語義資源，以“優選”的方法來實現詞義與結構的消歧。

“優選”運用于分析所生成中間語言中，這種方法把詞語的當前語境與詞語各義項的限制規則所描述語義特徵資訊進行比較，根據比較的相似度選擇最合適的義項。同時將相似度的最大值作爲該詞語的評價值。中間分析結果中各實義詞的評價分值可以成爲

評價此中間結果的依據,以此在多個中間結構中選出最佳的結果(注意本文中提到的“最佳”都是相對於演算法而言的，是在當前演算法下最好的解，但這並不一定總是實際正確的解）。這樣，我們在解決詞義歧義的基礎上同時也解決了結構歧義。“優選”排歧將基於規則與基於統計的排歧方法相結合，排歧中所使用的義項限制規則的獲取方法是從大規模的語料庫中統計出義原的同現集合，再按一定的轉換模板半自動地生成的。使用這種方法可以大大減少手工編制規則的工作量。

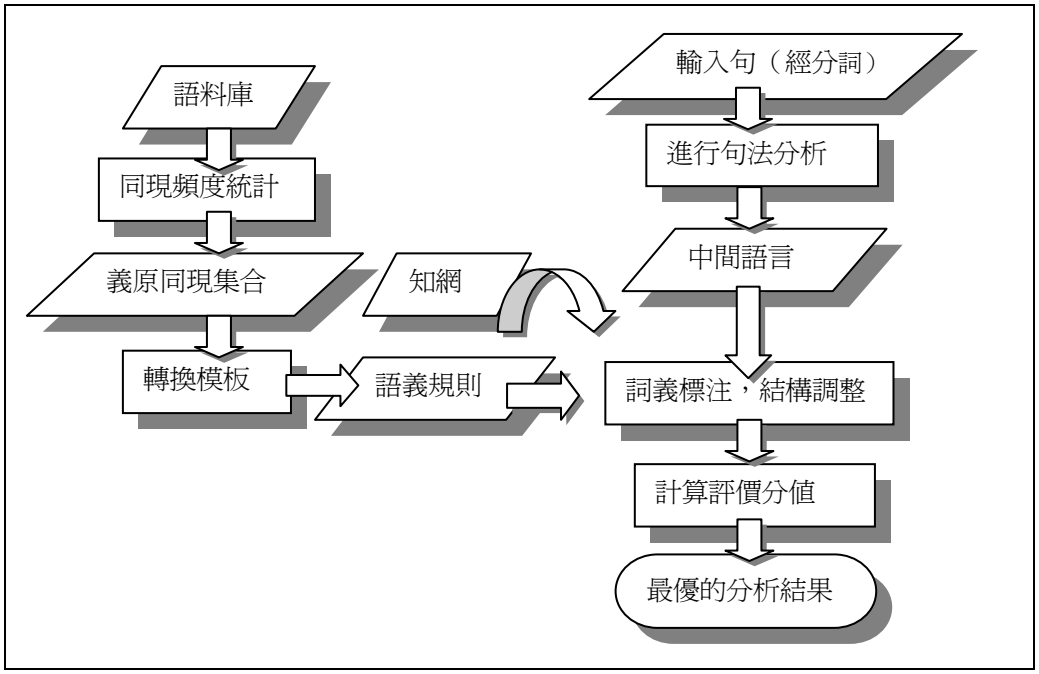


***圖 2.1** 含有詞義排歧模組的 PARSING 流程*

含有詞義排歧模組的漢英機譯系統分析部分的工作流程如圖 2.1 所示。

2.2 《知網》介紹

《知網》(英文名稱 *HowNet*)是其創建人董振東先生花費逾十年研究心血的重要成果。《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，它是一個網狀的有機的知識系統[李涓子 等 1999]。

語義詞典是知網系統的基礎文件。在這個文件中每一個詞語義項的概念及其描述形成一個記錄。目前詞典中提供漢英雙語的記錄，每一種語言的每一個記錄都主要包含 4 項內容。其中每一項都由兩部分組成，中間以“=”分隔。每一個“=”的左側是資料的功能變數名稱，右側是資料的值。它們排列如下：

NO.= 詞或短語序號

[W_X= 詞或短語
G_X= 詞或短語的詞性
E_X= 詞或短語的例子]+
DEF= 概念定義

其中的 W_X、G_X、E_X 構成每種語言的記錄，X 用以描述記錄所代表語種，X 爲 C 則爲漢語，爲 E 則爲英語。每個詞語由 DEF 來描述其概念定義，DEF 的值由若干個義原及它們與主幹詞之間的語義關係描述組成。義原是知網中最基本的、不易於再分割的意義的最小單位，知網通過對約六千個漢字進行考察和分析來抽取了 800 多個義原，並總結了如部分、主體、客體、從屬、時空、材料等若干種義原間的語義關係，這些關係在知網中用義原前附加如“%”、“@”、“$”等相對應符號來表示，因此我們把這些語義關係稱之義原的字首語義關係，而對應的符號爲義原的字首語義關係符。這些符號的意義在《知網》中有詳細的定義說明，表 2.1 列舉了本文出現的一些字首符號及其對應語義關係。具體的符號定義可參看文獻 9。

表 2.1 *義原的部分字首符號及其對應語義關係*

字首符號	對應的語義關係
#	相關
%	部分
$	事件的受事、目標、所有
*	事件的施事、體驗者、工具
+	蘊含
&	從屬
~	可能性
@	時空
?	材料
^	否定

下面我們用形式化的語言來對 DEF 進行定義：

DEF = [Mark]Atom[,[Mark]Atom]*

Mark = * | @ | ? | ! | ~ | # | $ | % | ^ | &

ATOM = $atom_1|atom_2|\ldots|atom_k$

所有這些義原及其關係應能對《知網》中出現的任何詞語概念進行定義。

下例是動詞“打”作“打球”解的義項在詞典中的定義：

W_C=打
G_C=V
E_C=~網球，~牌，~秋千，~太極，球~得很棒
W_E=play

G_E=V
E_E=
DEF=exercise|鍛練,sport|體育
通過 DEF 的定義可以知道在“打球”中“打”和“體育”與“鍛練”有關。

又例如“面”這個詞語包含有如下兩個義項：

W_C=面	W_C=面
G_C=N	G_C=N
W_E=noodles	W_E= face
G_E=N	G_E=N
DEF= food|食品	DEF= part|部件,%AnimalHuman|動物,skin|皮

第一個義項的“面”作“麵條”解，它的義項定義是“fool|食物”；第二個義項的“面”作“臉面”解，它的義項定義說明在這個義項中“面”是動物的部件，是“皮”。

除了語義詞典外，知網還提供了義原分類樹，分類樹把各個義原及它們之間的聯繫以樹的形式組織在一起，父子結點的義原具有上下位的關係。我們可以通過義原分類樹計算義原間的語義距離。在知網中存在 ENTITY、EVENT 等幾棵分類樹，如下圖是詞網中表示事件義原關係的 EVENT 分類樹：

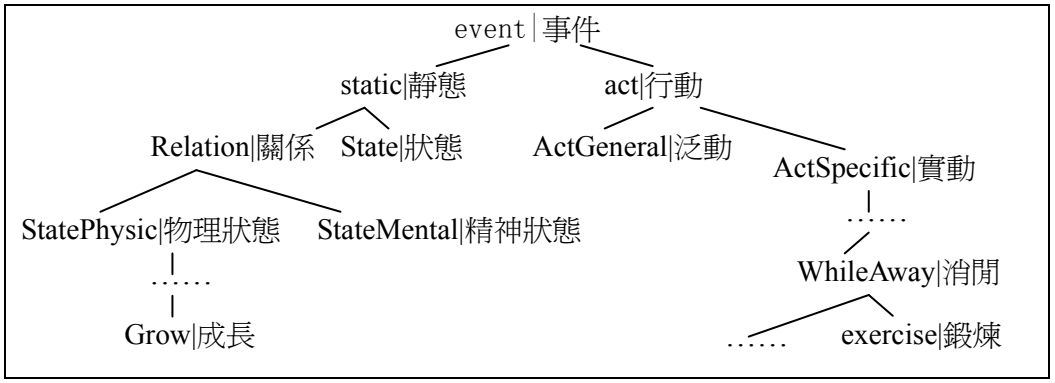


圖 2.2　EVENT 義原分類樹的樹結構表示

知網的詞典用文本的方式保存的，要事先把它轉換成方便系統實現的格式，考慮到系統用 LISP 實現，本文把詞典用表的形式來表示。詞條中除 DEF 外的項都被表示爲以該項爲首位元素的子表，而 DEF 項被拆分成一系列以語義關係爲首位元素的子表，DEF 中所有的具有相同語義關係的義原都歸於相應的子表中。如果義原無字首語義關係，可以都歸爲 PROPERTY 子表。例如我們可以把“面”的義項 2 表示爲：

```
( (W_C  面)(G__C N)(W_E face)(G_E N)
   ( (PROPERTY part|部件 skin|皮)
     (PARTOF AnimalHuman|動物) )
 )
```

3. 基於優選的詞義排歧

3.1 “優選”演算法的總體思路

本演算法將基於規則與基於統計的排歧方法相結合。首先利用統計的方法從未標注的大規則語料庫中抽取出義原的同現語義資訊，並根據轉換模板構造出義原的初始限制規則，再通過手工的方式對初始規則進行修改與調整，以得到一個較完善的規則集，繼而獲得詞語義項的語義限制規則。在對一個中間結構進行詞義排歧的過程中，我們按一定的演算法確定各詞語的上下文語境，將該詞語各義項的限制規則和其語境一一進行相似度比較，根據比較的結果確定它及其語境中各詞語的意義。最大的相似度可作該詞語的評價值，而中間結構的評價值就是根據其構成詞語的評價值計算得到的。在最後的結構排歧中，演算法以評價值大小爲優選依據從多個侯選的中間結構中挑出一個最優的結果。這樣我們在解決詞義的消歧的同時也可解決結構的消歧。

3.2 從語料庫中獲取義原的同現集合

語言學家 FIRTH 在對詞義辨識的描述中提出“觀其伴而知其意”(You shall know a word by the company it keeps)，他認爲，詞語的意義只能在上下文中才能得以辨識。如果一個詞語的某一詞義在語料庫中出現多次，我們在其出現的上下文中可以發現某些詞語出現的頻率很高，而這些詞語與該詞義之間有著比較密切的搭配的關係。詞義由義原構成，同樣地，與一個特定義原在上下文中共同出現次數越多的義原也與該義原有越密切的語義關聯，而這些的搭配義原就可以爲我們制定這個義原的語義限制規則提供很好的參考依據。動詞、形容詞的詞義對句子的語義影響最大，而動詞、形容詞的主要構成是動作類義原和屬性值及數量值類義原，因此我們希望得到這兩類義原的同現集合。

義原 a 的同現集合定義爲：

$S(a)=\{(b\ Prob(a,b)) \mid b \in ATOMSET\}$

集合 S(a)中的每個元素都是一個二元組(b Prob)，它的第一個分量是義原 b，第二個分量 Prob(a,b) 表示 a,b 義原的同現概率，即含有分別 a,b 義原的詞語在同一指定上下文窗口中出現的可能性。ATOMSET 是《知網》中所有義原的集合。

“同現”是針對一定詞語範圍而言的，我們把這一範圍稱爲上下文窗口，如果兩個義原所在的詞語在同一上下文窗口中出現，我們稱這兩個義原在當前前窗口中“同現”。上下文窗口的長度選取也是影響到統計結果好壞的重要因素。窗口選取得太小，例如只選擇當前詞語的前後一兩個詞是不夠的，漢語中的許多複雜句式語義相關詞的距離比較長，如

“人口普查的資料對國家制定長遠的經濟規劃具有十分重要的意義”

在這句話核心詞語是“資料”、“具有”、“意義”，它們之間相間的詞語數多達 6 個。如果視窗過小，句子中的關鍵搭配資訊就很可能超出了視窗的範圍而不能被統計。但反之如果窗口選擇太大，窗口內會含有過多與搭配不相關的詞語，如果把它們都進行統計，也會導致統計結果的不準確。按照一般的經驗，本文選取的窗口大小爲前後 7 個詞[梅家駒 1999]。

爲降低統計雜訊，視窗內詞語的選擇也應當考慮詞性、語法等資訊，應儘量選取存在語法關係的辭彙。本文是根據當前詞語的詞性與其他詞性可能組成的各種句法關係來選擇視窗的詞語。動詞及形容詞可以與其他詞性組成的句法關係如表所示[李涓子 等 1999]：

表 3.1 *不同詞性的可能搭配組合表*

詞性	相對位置	詞性組合	詞法關係
動詞	前	名+~	主謂關係
		動+~	並列關係
		名+~	定中關係
	後	~+名	動賓關係
		~+名	定中關係
		~+動	並列關係
		~+形	動補關係
形容詞	前	名+~	主謂關係
		形+~	並列關係
	後	~+名	定中關係
		~+動	並列關係

本文主要從搭配的角度來進行排歧，因此暫不考慮並列的語法關係。這樣從表中可以看到，對動詞來說，前窗口應選擇詞性爲名詞的詞,後視窗應爲詞性集合{名詞,形容詞}中的詞。形容詞的前後窗口都是詞性爲名詞的詞。詞語的前後出現的詞語所起的語法及語義關係差別較大，所以有必要就詞語的前後窗口分別選擇同現詞語，統計出來的義原同現集合也應相應地分成兩部分。我們令 S-(A)、S+(A)分別代表義原 A 的前、後同現集合。

我們的統計語料來源是《讀者 20 年文集》，規模大小約爲 1,100 萬字。語料庫沒有經過任何的語義標注，因此我們無法確定語料庫中一個句子裏某個詞語的詞義究竟是什么，如果把它的所有的義項都參與同現統計，勢必會使統計結果含有較大的語義噪音。爲此上下文窗口中的詞語應選取無詞義歧義的，即只對單義詞進行處理，而不考慮多義詞。注意這裏的“單義”是相對於詞性而言的，詞語可能具有多個詞性，如果詞語 W 在某個詞性 C 中的意義是唯一的，我們就稱 W 在詞性 C 下是單義的。在漢語的詞語中，單義詞的數量占很大的比重，在《知網》中的總詞語數爲 53332，而單義詞數目爲 47188，

占總數的 88.47%；而從實際的運用來考查，單義詞出現的頻度也不低。經統計，在語料庫中，單義詞的數目占 44.18%。這表明單義詞的搭配資訊具有相當的代表性，利用單義詞的搭配資訊來進行多義詞詞義選擇的思想是完全可行的。

義原的同現集合獲取過程是無指導的。但是我們事先要對統計語料庫進行自動分詞及詞性標性，才能正確地選擇詞語的上下文窗口。

演算法 3.1：義原的同現集合獲取演算法

定義：設 Total 爲單義詞的計算器; 陣列 COUNT-F,COUNT-B 是兩個二維陣列，COUNT-F [a,b]是 b 對於 a 的前同現次數；而 COUNT-B [a,b]是 b 對於 a 的後同現次數。

初始化：Total=0;

for each a∈動作類義原集∪屬性值類義原集∪數量值類義原集,
  for each b∈AtomSet
    Count-F[a,b]=Count-B[a,b]=0;

處理過程：

對於語料庫中出現每個單義詞語 W, Category(W) ∈{v,adj }，執行：

{ Total=Total+1;
  確定 W 的義原定義集合爲 Atoms(W)；
  確定 W 的前、後窗口 Window-F(W)，Window-B(W)
  for each a∈Atoms(W)，處理前同現集合：
  {
   for each b∈ $\bigcup_{k\in Window-F(W)} Atoms(k)$
     COUNT-F[a,b]=COUNT-F[a,b]+1；
  }
  for each a∈Atoms(W)，處理後同現集合：
  {
    for each b∈ $\bigcup_{k\in Window-F(W)} Atoms(k)$ 都做
      COUNT-B[a,b]=COUNT-B[a,b]+1；
  }
}

在對語料庫中的句子處理完畢後，可以計算出每個義原的同現義原集

S-(A)=

{(b freq(a,b)) | freq(a,b)=γ_1*count-F[a,b]/Total, b∈ATOMSSET }

S+(A)=

{(b freq(a,b)) | freq(a,b)=γ_2*count-B[a,b]/Total, b∈ATOMSSET }

其中 γ_1, γ_2 分別是前後同現集合中同現概率的放大係數。太小的概率難以表現同現義原間的頻度差異，演算法將概率值同時乘以一個放大係數，這並不影響它們之間的大小關係，而且適當地放大了相互的差異值。

統計語料庫經自動獲取得到的動作類義原同現集合 703 條，屬性值類及數量值義原同現集合 446 條。下例分別是動作類義原“EAT|吃”和屬性值類義原“HAPPY|福”的前後同現集合：

EAT|吃 ：

前同現集合：

{(HUMAN|人 0.56) (THIRDPERSON|他 0.41) (MALE|男 0.28) (MASS|衆 0.23)
(PLACE|地方 0.22) (BIRD|禽 0.20) (PROPERNAME|專 0.20) …}

後同現集合：

{(MEDICINE|藥物 0.52)(PART|部件 0.38)(ATTRIBUTE|屬性 0.26)
(HUMAN|人 0.20) (DESIRED|良 0.20) (FOOD|食品 0.15) …}

HAPPY|福 ：

前同現集合：

{(HUMAN|人 0.56) (THIRDPERSON|他 0.41) (MALE|男 0.28) (MASS|衆 0.23)
(PLACE|地方 0.22) (BIRD|禽 0.20) (PROPERNAME|專 0.20) …}

後同現集合：

{(MEDICINE|藥物 0.52)(PART|部件 0.38)(ATTRIBUTE|屬性 0.26)
(HUMAN|人 0.20) (DESIRED|良 0.20) (FOOD|食品 0.15) …}

通過對同現集合的分析可以發現，集合中有些義原的意義很接近，語義相似度比較小。我們希望能把同現義原集進行聚義處理，即語義接近的義原合併成一個義原，使同現集合的長度得以減少。下面我們給出聚類的演算法：

演算法 3.2 同現集合聚義演算法

設 S 爲義原 Atom 經統計得到同現集合；

初始化新集合 L=Φ;

將 S 中的義原按所在分類樹的層數高低排序，使上位義原總是位於其下位義原之前；

WHILE($S \neq \Phi$)

{

從 S 中 POP 出首位元素 A =(a $Prob_a$);

T = 0;

for each B∈S,B=(b $Prob_b$)

{ if 義原 a,b 的相似度 r 小於閾值

{ T=T+$Prob_b$*r;

將 B 從 S 中移出;

}

}

將元素(a T)PUSH 至集合 L;

}

將 L 中的義原按同現概率的數值從大到小進行排序，截取集合前 10%位元的義原；

用集合 L 代替原同現集合 S。

演算法先將同現集合中的義原按其在分類樹的層數排序，這樣可以保證在一個語義相似的義原集群中，義原總是向著最上位的義原，即最抽象的概念進行聚義。由於同現概率高的義原對語境的描述起著更爲重要的影響，因此演算法在聚義完畢後，又將同現集合按義原同現概率的大小排序，捨棄了小同現概率的義原，只保留序列前部 10%的元素。經過演算法 3.2 的處理，新的同現集合的長度將比原來的有顯著減少。

3.3 產生語義限制規則

3.3.1 義原限制規則的定義

利用用演算法 3.1 可以得到每個動作類及屬性值、數量值類義原的上下文同現集合。實際上到這一步我們已經可以根據同現集合與測試句中詞語的上下文語境進行相似度的計算，將相似度最高的義項作爲該詞的在當前句子中的詞義。這種方法對於簡單的句式結構有比較高的正確率，但對於具有特殊的語法性質的詞語則排歧結果不是很理想。如對動詞來說[楊曉峰 2001]，帶複雜賓語(如小句賓語和兼語賓語)的多義詞的詞義排歧結果會差於帶簡單賓語的多義詞。這是由於在複雜語式裏中心詞語與搭配詞語距離較遠，搭配詞語或是超出中心詞語的上下文窗口範圍，或是與中心詞語之間有過多的干擾詞語。

在機器翻譯的語法分析階段成生了源語句的中間語言，它可以更爲準確地描述詞語所在上下文的語境；同時我們也爲義原定義語義限制規則，它描述了含有該義原的詞語的期望出現的語義環境。這樣我們可以根據詞語實際所處的語義環境與義原規則中描述的語義環境進行相似度的計算，將比較結果作爲詞義排歧的依據。

本文中採取的中間語言模型在第二章中已給出了詳細的說明。如測試句：

“我怕你把筆尖給弄斷了。”

經分析的中間語言框架爲：

```
(   (CROOT 怕) (CAT V)
   (AGENT ((HUMAN +) (CAT PRON)(AGREE SG) (PERSON FIRST) (CROOT 我)))
   (THEME (    (CROOT 弄)
               (AGENT ( (HUMAN +)(CAT PRON)
                        (AGREE SG) (PERSON SECOND) (CROOT 你)))
               (THEME ((CAT N)  (CROOT 筆尖)))
               (RESULT ((CROOT 斷) (CAT V)))
          ))
)
```

在格結構中動詞或形容詞的語義環境是由 Agent、Theme、Result、Clause 等格資訊來描述。由於缺少語義資訊，在分析階段給出的語義格資訊並不一定正確，我們可以在詞義排歧階段進行自動的語義格調整。

具體的義原的語義限制規則的模型表示爲：

Rule = (SenseAtom Rule-Items)

Rule-Items=(Logic-Op {(Rule-Items)}+) | {(Sem-Case Logic-Item)}+

Logic-Item = (Logic-Op {Logic-Item}+) | { Sense-Item}+

Sem-Case = Agent | Theme | CO-THEME | Result | Clause | · · ·

Logic-Op = *AND* | *OR* | *NOT*

Sense-Item=*NOT* | (SenseAtom Prob) | (Relation {(SenseAtom Prob)}+)

SenseAtom= EAT|吃 | HAPPY|福 | · · ·

從上面的形式化描述中我們可以看出義原的語義規則與第三章的模式規則有相似之處。規則是用表的形式來表示的，表的首位元素指明規則所屬的義原。表的第二項即表的規則體。規則體由一套或多套子規則體構成。每一套子規則又是由一系列格的限制描述組成。格限制描述說明了含有當前義原的詞語在實際句子的語義環境中希望出現特徵資訊，包括義原及其同現概率，這些義原可能會帶上語義關係成分如 Partof，Material 等。格限制描述這些特徵資訊的邏輯語義關係。這些邏輯關係用*AND*、*OR*、*NOT*等符號來表示。*AND*表示在在實際格中出現的詞語義項應同時滿足指定的特徵資訊；*OR*表示只要能滿足特徵資訊的任意一項即可；而*NOT*則規定詞語不應出現任何指定的特徵資訊。邏輯運算不僅可以作用在格描述的特徵資訊上，也可以作用於不同套子規則中。在這一級上的邏輯算符一般使用*OR*，表示規則可以選用給出的若干套子規則中任意一套。帶有邏輯運算的限定規則可以表述各種複雜的語義資訊。

3.3.2 義原語義限制規則的生成

爲了將 3.2 中得到的義原同現集合的資訊充分運用到義原的限制規則定義中，本文採用了以下的轉換模板，它可根據同現集合構造出一個初始的義原語義規則：

（1）對於屬性值、數量值類義原，前同現集合中實體類的義原作爲“EXPERIENCE”格。
（2）對於屬性值、數量值類義原，後同現集合中實體類的義原作爲“THEME”格；
（3）從動作類義原的前後同現集合中取出“implement|器具”及其所有的下位義原，將它們作爲“INSTRUMENT”格；
（4）從動作類義原的前後同現集合中取出“earth|大地”、“place|地方”、“space|空間”及其所有的下位義原，將它們作爲“LOCATION”格；
（5）從動作類義原的前後同現集合中取出“time|時間”及其所有的下位義原，將它們作爲“TIME”格；
（6）從動作類義原的前後同現集合中取出“degree|程度”、“range|幅度”、“frequency|頻率”及其所有的下位義原，將它們分別作爲“DEGREE”、“RANGE”、“FREQUENCEY”格；
（7）從動作類義原的後同現集合中取出的屬性值、數量值類義原，將它們作爲“RESULT”格；
（8）將動作類義原的前同現集合餘下的實體類義原作爲“AGENT”格；
（9）將動作類義原的後同現集合中的餘下的實體類義原作爲 THEME 格；

例如對於動作類義原“eat|吃”，利用轉化模板得到的義原語義限制規則爲：

```
(eat|吃
 (AGENT (*OR* (HUMAN|人  0.56) (THIRDPERSON|他  0.41) (MALE|男  0.28)
              (MASS|衆  0.23) (PLACE|地方  0.22) (BIRD|禽  0.20)
              (PROPERNAME|專  0.20) … )
      )
 (THEME (*OR* (MEDICINE|藥物  0.52)(PART|部件  0.38) (HUMAN|人  0.20)
              (FOOD|食品  0.15) … )
 (RESULT (*OR* ((ATTRIBUTE|屬性  0.26)  (DESIRED|良  0.20) )
)
```

轉換模板爲我們構造了一個初始規則庫，規則庫中定義了義原 Agent、Theme、Result、Instrument 等格的限制描述。對於語法性質簡單的義原，這些格描述已經足夠。但是對於大多數義原而言，這些自動生成的規則就過於簡單了，因此我們需要在初始規則的基礎上手工對其進行修改與調整，加上必要的格描述、剔除錯誤的特徵義原。例如，對於如“URGE|促使”，由該義原定義的詞語如“促使”、“推動”、“鼓勵”等詞語，一般都帶有兼語，因此我們要爲其增加兼語表示的格(EVENT)的限制描述；還有如由“GIVE|給”定義的詞語，一般都帶有雙賓語，我們也要爲“GIVE|給”定義 CO-THEME(間接賓語)格的限制。另處我們也需爲“EXPECT|期望”等義原規則定義 CLAUSE 格描述。

在修改義原的語義限制規則時，本文從語料庫中爲每個義原選取一定數量的例句，

這些例句中都含有由該義原定義的詞語。我們根據參考例句對初始規則進行修改完善。由於義原的語義規則規模有限，並且事先有一個自動生成的初始規則集，因此手工制定與修改規則所花費的工作量並不大。通過在初始規則的基礎上進行人工調整的方法，我們可以得到一個較爲完善的語義規則集。

3.3.3 義項語義限制規則的確定

在爲動作類義原及屬性值、數量值類義原定義限制規則後，依據《知網》中對義項的一些規定，我們可以自動地生成動詞或形容詞的任意一個義項的語義限制規則。

對於動作類動詞，DEF 項的第一位置只能是事件類規定的主要特徵；因此可以直接將義項第一位置上的義原的限制規則作爲該義項的限制規則。

打 1：buy|買,commercial|商
打 2：exercise|鍛練,sport|體育
吃 1：eat|吃
吃 2：destroy|消滅,military|軍

在上面幾個動詞義項，我們分別將“buy|買”、“exercise|鍛練”、“eat|吃”、“destroy|消滅”的限定規則作爲打 1，打 2，吃 1，吃 2 的限定規則。

而對於形容詞，它們的義項主要由屬性值類義原及數量類義原構成。“屬性值”是所有屬於屬性值概念的唯一的主要特徵，“數量值”是所有屬於數量值概念的唯一的主要特徵，它們分別是形容詞的各義項的首位標識；屬性值類義原和數量值類義原除首位標識外必須還包含有一個次要特徵。在第二位元上一定要標注該屬性值或數量值所指向的屬性或數量特徵；而通常絕大多數情況下在第三位置上標注該屬性值或數量值的具體值，而這些具體值正是我們所感興趣的。有時在 DEF 第三位置後還有一些輔助特徵，它們只是進一步對關鍵義原進行補充說明，對義項的語義影響很小，因此我們在定義形容詞的限制規則中，只需考慮第三位的義原。例如下面是幾個形容詞的義項定義：

巨大 1：DEF=aValue|屬性值,size|尺寸,big|大

巨大 2：DEF=QValue|數量值,amount|多少,many|多

香 1：DEF=aValue|屬性值,circumstances|境況,flourishing|興,desired|良

香 2 ：DEF=aValue|屬性值,odor|氣味,fragrant|香,desired|良

在上面的例子中，我們將選擇“big|大”、“many|多”、“flourishing|興”、“fragrant|香”的義原限制規則作爲對應形容詞義項的語義限制規則。

對於量詞我們也可以自動生成限制規則，規則中規定量詞修飾的詞的語義特徵。在《知網》中，名量詞的定義裏用“&”標注其指向的屬性或事物的類型；例如：

本： DEF=NounUnit|名量,&publications|書刊
輛： DEF=NounUnit|名量,&LandVehicle|車

於是我們就可以直接把“&”標注的義原作爲義項的限定規則。如上例中，義項“本 1”與“輛 1”的規則分別爲

本 1：(THEME publications|書刊)
輛 1：(THEME LandVehicle|車)

與此同時，我們還可以爲某些語義格定義通用的規則，以處理中間結構中不能根據搭配關係進行排歧的詞語。這些語義格通常含有較明顯的語義特徵，如 LOCATION 格中一般含有地點等資訊，而 TIME 格中一般含有時間等資訊，因此，我們可以爲這兩個格定義通用的限定規則如下：

LOCATION：(*OR* PLACE 地方)
TIME： (*OR* 時間)

3.4 詞義排歧演算法的詳細描述

上一節介紹了如何獲得詞語義項的語義限制規則。本節將討論如何運用義項的語義規則在給定的中間結構中進行詞義排歧。

3.4.1 確定詞語的語義環境

本文第二章給出了機譯系統中間語言的表示方法，爲了方便地獲得某一實義詞的上下文相關詞，我們先將中間語言轉放成爲依存關係樹的形式。樹的根結點是句子的核心詞，其他受核心詞支配的附屬成分就作爲根結點的子樹，這些子樹分別也是以各附屬成分爲根結點而建立起來的依存關係樹。例如輸入句“維修/圖書館/的/空調”的兩個可能的句法結構樹如下：

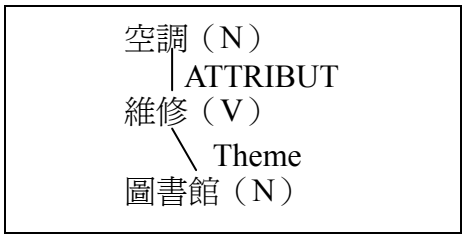


圖 3.1 ((維修 圖書館)的 空調)的樹結構

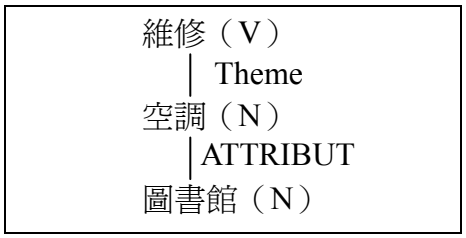


圖 3.2 (維修(圖書館的 空調))的樹結構

在確定詞語的上下文語境前我們定義詞語的限定關係：設有詞語 A，B，如果在句子中 A 修飾、支配 B，則稱在當前句中 A 是 B 的限定詞，B 是 A 的被限定詞。這裏的限定關係與傳統的依存關係有所不同：依存語法認爲發生依存關係的一對詞中，如果詞 A 修飾詞 B，則 B 爲主詞，A 爲從屬詞，A 是 B 的附屬成分，在依存關係樹中體現爲 A 是 B 的子結點；而在我們定義的限定關係中被限定詞則是限定詞的修飾支配物件。如對於主謂賓結構的句子來說，主動詞就是句子的詞語主語、賓語的限定詞；ADJ＋NP 的形

容詞短語中 ADJ 是 NP 的限定詞。如果限定詞與被限定詞之間是偏正的修飾關係，如 ADJ+NP,NP+NP 等，限定詞充當 ATTRIBUTE、MANNER 等附屬成分，則在依存關係樹中體現爲限定詞是被限定詞的子結點；如果它們之間是主謂結構，如 NP+VP+NP(主謂賓)、NP+ADJP 等，則限定詞在樹中充當被限定詞的父結點。

我們之所以要對詞語區分限定詞與非限定詞是由於這兩類詞的評價演算法不相同。根據定義，限定詞在詞語之間的搭配關係中起著主要的修飾支配作用，它的義項選擇及評價分數影響著當前分析結構的總體評價值。在具體的演算法實現上，限定詞的評價是通過各個義項的實例集與它的被限定詞語集的比較獲得的；而非限定詞的評價則要在其限定詞義項確定之後進行，選取與這個義項的實例的最高比較值作爲評價依據。對一棵關係樹進行評價就是建立在對句中充當限定詞的詞語結點的評價上的[詹衛東 1996]。

對於一個可充當限定詞的詞語，我們定義它的上下文語境，也稱爲上下文窗口，是所有被其限定的詞語並上所具有的格資訊。給定句子的一個句法結構，通過該結構對應的依存樹，我們可以很方便地得到各詞的上下文語境。下面是詞語在當前依存樹中的語境選取原則：

(1) 如果詞語是動詞，則選取其子結點的 AGENT、THEME、MANNER、TENSE、TOOLS、LOCATION、TIME 等附屬成分；
(2) 如果詞語爲其父結點的 PROPOSTION(從句)、ATTRIBUTE 等附屬成分，則需並上父結點並將其格注爲 Parent；
(3) 如果動詞的語義格爲 EVENT，這時動詞做兼語句裏的兼語，此時動詞所在兼語句中的賓語應充當其 AGENT 格；
(4) 如果動詞的語義格爲 THEME，這時動詞做賓語或小句賓結構的賓語從句。如果該動詞的 AGENT 語義格不存在，則加入父結點語境中的 AGENT 格。如果該動詞是被動語態，要將 AGENT 格改爲 THEME 格；
(5) 如果動詞的語義格爲 SUBEVENT，這時動詞做連動句，則加入父結點的主語詞及其所充當的語義格。如果該動詞是被動語態，要將 AGENT 格改爲 THEME 格；
(6) 如果詞語是形容詞且爲 ATTRIBUTE 格，則選取其父結點並將其語法格注爲 THEME；否則如果形容詞爲 PREDICATE 格，即形容詞作表語，則選取兄弟結點中的 Experiencer 格；
(7) 如果詞語是量詞且爲 Quatity 格，則選取其父結點並將其語法格注爲 THEME。

3.4.2 詞語的評價值計算

在 3.3.3 中介紹了動詞、形容詞及量詞的義項的語義限定規則獲取方法。上一節描述如何得到一棵中間語言的結構依存樹中各支配詞的上下文語境。現在可以根據詞語所在語義環境及其各義項的限定規則中所描述的語義環境進行相似度的計算。

1)詞語義項與規則的格限制描述的相似度

《知網》中提供了動作類，實體類、屬性值類等義原分類樹，義原間的語義距離與義原

在分類樹對應結點間的最短路徑的邊數有關：最短路徑越長，則表示兩個義原的語義距離越遠，它們的相似程度越小。並且，位於分類樹下部的一對父子結點對應的義原間的語義距離應當小於位於其上的任意一對父子。因此，在計算語義距離時應對分類樹上的邊加權。此外，計算中還應體現出不同類型的義原(如實體類與動作類)之間語義的不可比性。根據上述分析，義原間語義距離和語義相似度的定義如下：

設義原 A、B，它們分別位於分類樹的第 a 和第 b 層(規定根結點爲第 0 層)，它們的最近共同祖先(可以是 A 或 B 本身)位於第 c 層，則 A 與 B 的語義距離爲：

$$\text{DISTANCE-ATOM}(A,B) = \begin{cases} \text{MAXVALUE} & \text{;如果父原A,B不在同一棵分类树上} \\ (\sum_{i=c+1}^{a} Weight(i) + \sum_{j=c+1}^{b} Weight(j))/2 & \text{;否则} \end{cases} \quad \textbf{(3.1)}$$

其中權值函數應是一個單調遞減的函數。設當前分類樹的樹高爲 Depth，我們定義的權值函數如下：

$$Weight(i) = 2*(Depth - i) / (Depth*(Depth+1)) \quad \textbf{(3.2)}$$

而義原 A，B 的語義相似度爲：

$$\text{SIM-ATOMS}(A,B) = \begin{cases} 0 & \text{;如果父原A,B不在同一棵分类树} \\ (1 - \text{DISTANCE-ATOM}(A,B))*100 & \text{;否则} \end{cases} \quad \textbf{(3.3)}$$

其中 MAXDISTANCE 是義原 A 所在分類樹的最長的語義距離。

以上定義的語義距離與語義相似度中的 A，B 是可交換的，即 A 與 B 的語義相似度等於 B 與 A 的語義相似度。但是我們將會看到，SIM-ATOMS 中比較的是模式義原 A 與實際義原 B 之間的相似度。如果實際義原 B 是模式義原 A 的下位，則它們的語義距離應比較小，如果規則義原在實際義原的下位或是它們只是擁有某個相同祖先的兩個結點，則語義距離應較大。這樣我們應對語義距離函數修改如下：

$$\text{DISTANCE-ATOM}(A,B) = \begin{cases} \text{MAXVALUE} & \text{;如果父原A,B不在同一棵分类树上} \\ (m * \sum_{i=c+1}^{a} Weight(i) + \sum_{j=c+1}^{b} Weight(j))/2 & \text{;否则} \end{cases} \quad \textbf{(3.1')}$$

在實際計算中，我們可以讓 m 的值取得足夠大，使得當義原 B 是義原 A 的子孫結點時語義距離較小，否則將得到一個較大的語義語離。根據實驗，M 值設定爲 5 能夠取得比較好的排歧效果。

2)詞語義項與規則的格限制描述的相似度

義項的語義限制規則中定義了某一語義格可能出現的特徵義原的邏輯組合，而詞語的義項是由義原及語義關係構成的。我們希望能能夠判定一個義項滿足規則的格限制描述的

程度，即義項與規則的格描述的相似度。

設某義項語義規則中規定了某格的限制描述爲 C，
$C=(OP\ (R_1\ S_1)\ (R_2\ S_2)\ldots(R_m\ S_m)\ ,CR_1,CR_2,CR_z)$。
其中對於 $1\le i\le m$,$R'_i \in RelationSet$，S_i 是描述中字首語義關係爲 R'_i 的特徵義原。描述中可能含有無字首語義關係的義原，它們本身就是格描述的特徵語義屬性，爲了保持計算的統一性，不妨假定這些義原的字首語義關係爲 Property。OP 是邏輯組合運算符，包括 *AND*,*OR*,*NOT*，它們的功能在前面已說明過。$CR_i(1\le i\le z)$是帶有邏輯組合運算符的嵌套語義描述集。

現有一詞語義項 $Entry=((R'_1\ E_1)\ (R'_2\ E_2)\ldots(R'_n\ E_n))$ 對於 $1\le i\le n$ 有 $E_i \subset AtomSets$，$R_i \in RelationSet$。同理，對於義項中無字首語義關係的義原，我們也假定其字首語義關係爲 Property。

設(R S)是 C 中的一個元素，其中 $S=(\ (atom_1\ prob_1)\ (Atom_2\ Prob_2)\ldots(Atom_m\ Prob_m)\ )$，定義函數

$$SIM\text{-}ENTRY\text{-}RELATIONITEM((R\ ,S)\ ,Entry)=$$
$$\begin{cases} \underset{(atom\ prob)\in S, a\in E'_i}{MAX} SIM\text{-}ATOMS(atom,a)\times prob & ;\text{如果存在} i, 1\le i\le n, \text{有} R'_i = R \\ 0 & ;\text{否則} \end{cases} \quad (3.4)$$

要得到 C 與 Entry 的相似度，可以將 OP 的運算元依次與 Entry 進行相似度的計算，並根據 OP 的值從計算結果中挑選出合適的結果。注意對於 $CR_i(1\le i\le z)$，需要遞迴地進行計算。

我們定義集合 RS 是 OP 的每一個運算元與 Entry 比較的相似度集合，即

$$RS=\{SIM\text{-}ENTRY\text{-}RELATIONITEM(\ (R_i\ ,S_i),\ Entry)\mid 1\le i\le m\}$$
$$\cup\{\ SIM\text{-}Entry\text{-}SC(Entry\ ,\ CR_j)\ \mid\ 1\leqslant i\leqslant z\ \} \quad (3.5)$$

如果 OP 爲*AND*，根據 3.3.1 中的定義，只要 RS 集合中有一個元素的比較值較小，返回的值就應該小，此時應從 RS 中選取一個最小值；如果 OP 爲*OR*，則表示集合可取任何一比較值，這樣此時應從 RS 中選取一個最大值；如果 OP 爲*NOT*，表示集合中只有一個元素，並且如果該元素值比較大，返回的值反而應該小，反之元素的值比較小的話，返回的值應該大。因此，我們可以用 1 與元素值的差來計算比較相似度。

綜上所述，可以定義 C 與 Entry 的相似度爲

$$SIM-ENTRY-SC(Entry,CaseRule)=\begin{cases} \mathbf{MAX}\ RS & ;\text{当OP= *OR* 时} \\ \mathbf{MIN}\ RS & ;\text{当OP= *AND* 时} \\ \mathbf{1-MAX}\ RS & ;\text{当OP= *NOT* 时} \end{cases} \quad (3.6)$$

3)詞語的評價演算法

演算法 3.3：詞語的評價演算法

輸入：詞語 Word 及 Word 的上下文語境 ENV，其中

ENV＝((CASE'$_1$ WORD$_1$ CAT$_1$)(CASE'$_2$ WORD$_2$ CAT$_2$)…(CASE'$_n$ WORD$_n$ CAT$_n$))

輸出：Word 及 Wordi(1≤i≤n)在 ENV 中的最佳義項及評價值，它們以如下表的形式輸出：

((Word Mark BestEntry)(word$_1$ Mark$_1$ BestEntry$_1$)…(word$_n$ Mark$_n$ BestEntry$_n$))

定義：EnvBestMark,EnvBestEntry 分別記錄 word$_i$ 在 ENV 中的當前評價值及最佳義項；

WordBestMark,WordBestEntry 分別記錄 word 的當前評價值及最佳義項；

二維陣列 Best-Mark-Entry[word$_i$,Entry]用以記錄 word$_i$ 對於 Word 的 Entry 義項的評價值及最佳義項；

函數 GET-ENV-CASE-WORD(ENV，CASE)用以返回 ENV 中格爲 CASE 的項；

步驟：for each Entry1 in Entries-of-Word(Word)

```
{ 令累計總分數 T=0；
  確定 Entry1 的語義限制規則集合 RuleSet;
  for each R∈RuleSet
   { 設 R=((CASE1  CASESC1)(CASE2 CASESC2) …(CASEm CASESCm))；
     for each (CASE CASESC) in R
     { 設 GET-ENV-CASE-WORD(ENV，CASE)的返回值爲(CASE Wordi CATi)；
       EnvBestMark=MINIMUM;
       for each Entry2 of Entries-of-Word(Wordi)
         if  Entry2 的詞性爲 CATi  then
             { curMark=SIM-ENTRY-SC(Entry2,CASESC);
               if curMark>EnvBestMark then
                 {  EnvBestMark=curMark;
                    EnvBestEntry=Entry2;
                 }
             }
        Best-Mark-Entry[wordi,Entry1] = (wordi EnvBestMark EnvBestEntry);
        T = T + EnvBestMark;
     }
   T=T / |R|;
   if  T> WordBestMark then
      WordBestMark=T; WordBestEntry=Entry1;
}}
```

令 Result = {(Word WordBestMark WordBestEntry)}

U {Best-Mark-Entry[w,WordBestEntry] | (CASE w CAT) ∈ENV}

返回 Result;

演算法首先確定 word 各義項的語義規則集，注意義項的語義規則可能有多套。將

各規則中格的限制描述與 word 的語義環境 ENV 中各詞語的義項進行相似度的計算，這些詞語相對於 word 當前義項的最大相似度值及對應義項記錄於一個二維陣列中。接著演算法根據 ENV 中各詞語的最大相似度計算 word 當前義項的評價值。最後，演算法把 word 的具有最大評價值的義項作爲 word 的最佳義項 BestEntry，這一最大評價值作爲 word 的評價值。同時，演算法從二維陣列中取出 ENV 中各詞語相對於 BestEntry 的最大相似度及義項，並將其作爲該詞語在 word 的語境中的評價值及最佳義項。

3.4.3 詞語的義項選擇

在演算法 3.3 給出了詞語的評價值計算方法。該演算法可以得到中間結構中各限定詞詞的評價值及最佳義項，同時還可以確定受該詞語限定的各詞語在該詞語的語境中的評價值及最佳義項。本節中介紹如何在中間語言結構中進行詞義的選擇，即詞義排歧。

本文按照詞性的語法特性來以一定的先後順序對不同詞語進行義項選擇。一般來說，動詞的搭配關係對結構的語義語義影響最大，其次是形容詞。根據這一原則，我們提出了以下詞義選擇的步驟演算法：

1：創建一個 RESULT 哈希表，它的入口是結構依存樹中各結點對應的詞語，每個入口項的值就是入口詞語的侯選義項集合及結點的評價分值。如果詞語在分析階段已經計算出義項集及評價分數，就將它們設爲相應的項值，否則初始的候選義項爲詞語的所有義項，而詞語評價分值爲 MINIUM。

2：按至底向上的順序對結構依存樹中的動詞結點進行詞義選擇，將選擇結果集合(可能含有一個或多個的元素)及評價分值填入 RESULT 哈希表的相應項中。對受該動詞限定的所有詞語，我們可以得到在該動詞下的評價分值及最優義項，將這個評價分值與 RESULT 表中對應項的評價值進行比較，如果更大，則將此最優義項集合與分值替代原來的內容。

3：與第 2 步類似，對樹中所有形容詞及其限定詞進行詞義選擇並更改相應的 RESULT 表項。

4：與第 2 步類似，對樹中所有量詞及其限定詞進行詞義選擇並更改相應的 RESULT 表項。

5：對依存樹中還未進行過詞義排歧的結點，如果其所在語義格有通用的限制規則，則根據通用規則進行詞義選擇，更改相應的 RESULT 表項。

6：對於依存樹的每個詞語結點，到 RESULT 表查出對應項的義項集合，將這些義項的所有英文對譯詞作爲分析結構的詞語譯文。

3.4.4 詞義排歧演算法示例

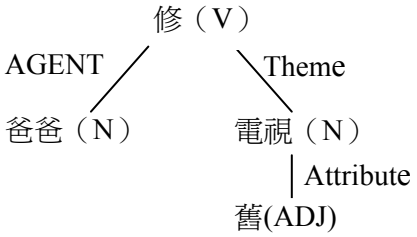
下面我們來看看如何運用本章給出的演算法進行詞義的排歧，設有輸入句

“爸爸正在修那台舊電視呢。”

分析後得到的中間結果爲：

```
((CAT V) (CROOT 修)
 (AGENT  ((CAT N) (CROOT 爸爸)))
 (THEME ((CAT N) (CROOT 電視)
          (ATTRIBUTE  ( (CAT ADJ) (CROOT 舊)))
         )))
```

(1) 將中間結果表示爲依存關係樹的形式：



(2) 確定動詞的上下文語義環境：

（修 ((AGENT 爸爸 N) (THEME 電視 N)))

（舊 ((THEME 電視 N)))

(3) 確定中間語言結構中各語義的義項規則

表 3.2 *中間語言結構中各語義的義項資訊表*

詞語	義項資訊		
	義項定義	義項規則	對應譯文
修	PutInOrder|整理	((AGENT(*OR* (HUMAN|人 0.21)..) (THEME(*OR* (SPACE|空間 0.33) (FACILITIES|設施 0.27) …)))	prune, trim
	Repair|修理	((AGENT(*OR* (HUMAN|人 0.23)..) (THEME(*OR* (PART|部件 0.37) (TOOL|用具 0.26) …)))	mend, repair, overhaul
	Study|學	((AGENT(*OR* (HUMAN|人 0.31)..) (THEME(*OR* (EDUCATION|教育 0.14) (KNOWLEDGE|知識 0.09) …)))	cultivate, study
爸爸	Human|人, family|家, Male|男	NIL	DAD, father, papa
電視	Image|圖像, Shows|表演物	NIL	television
	Tool|用具	NIL	television -set , TV-set
舊	Used|舊	(THEME (*OR* (clothing|衣物 0.29) (TOOL 用具 0.23) …))	old, used, worn
	Past|昔	(THEME (*OR* (TIME|時間 0.29) …)	old, past, bygone
	Original|原	(THEME (*OR* (PHYSICAL|物質 0.21) …)	former, onetime

(4) 對“修”各義項及其語義環境的詞語進行評價及選詞。結果如下：

表 3.3 “修”各義項在語境中的評價分值最對應義項

義項	語義環境中的詞語評分				義項評價分數
	爸爸		電視		
	評價分	最佳義項	評價分	最佳義項	
PutInOrder|整理	100	Human|人	83.92	tool|用具	91.96
Repair|修理	100	Human|人	91.07	tool|用具	95.53
Study|學	94.64	Human|人	0	NIL	47.32

從表中可得到“修”最大義項的評價分數爲 95.53，對應的最佳義項是“Repair|修理”

與該義項對應的“爸爸”在當前環境中的評價分數爲 100,最佳義項爲“Human|人”

“電視”在當前環境中的評價分數爲 91.07,最佳義項爲“tool|用具”

(5) 對“舊”及其語義環境的詞語進行評價及選詞。結果如下：

表 3.4 “修”各義項在語境中的評價分值最對應義項

義項	語義環境中的詞語評分		義項評價分數
	電視		
	評價分	最佳義項	
used|舊	100	tool|用具	100
past|昔	19.64	tool|用具	19.64
original|原	75.26	tool|用具	75.26

從表中可得到“舊”最大義項的評價分數爲 100，對應的最佳義項是“used|舊”

與該義項對應的“電視”在當前環境中的評價分數爲 100,最佳義項爲“tool|用具”

(6) 確定各詞語的詞義：

表 3.5 句子中各詞語的對應譯文

詞語	對應譯文
修	mend, repair,overhaul
爸爸	DAD, father, papa
電視	television-set ,TV-set
舊	old, used, worn

3.5 對中間結果進行語義處理

上一節中我們對詞義的排歧的演算法作了詳細的介紹。在詞義排歧的基礎上，我們可以進一步進行中間語言的結構排歧。爲了提高結構排歧的正確率，在詞義排歧的過程中，應該要同時對中間結果進行語義處理，調整句法分析錯誤的語義框架，或確定中間結構

中某些複雜的語義修飾關係。本節將給出中間結果的語義調整及語義關係確定的處理演算法。

3.5.1 調整分析結果的語義框架

3.4 中介紹了如何利用詞語所處的語義環境中與義項的規則進行相似度比較進行詞義排歧。語義環境由受轄詞及其語義格構成。當時我們假設了中間語言結構給出的是正確的語義格資訊，然而在實際的句法分析過程中，由於缺少語義資訊，中間結構的語義格並一定總是正確的。如下面的例句：

“被子叠得整整齊齊”

“我把被子叠得整整齊齊”

在第一個句子中，形式主語“被子”就是“叠”的受事體和邏輯賓語，因爲根據常識，“被子”是不可能折叠什么其他的東西，它只能是被叠。這類形式主語爲受事體的句子可以看成是“被”字句的省略形式，如這句話可以寫成“被子被叠得整整齊齊”。而第二個句子中“我”是“叠”的形式主語，也是施動者和邏輯主語，因爲一般情況下人不會被叠而可能是叠別的東西。在句法分析時不能獲得這些語義的資訊，因此難以正確地決定施事者與受事體，第一句的分析結果很可能會出現如下的錯誤：

ENV: (叠 (AGENT 被子)(RESUTL 整整齊齊))

在詞義排歧時獲得的語義資訊使我們能夠對結果的語義框架進行適當的調整。

“被子”的義項爲：

被子 1：tool|用具,*cover|遮蓋,#sleep|睡

而“叠”的義項的限制規則爲

叠 1：((AGENT (HUMAN|人 0.426))(THEME (TOOL|用具 0.138)))

它規定了 Agent 人的語義特徵是 HUMAN|人，如果用這一限制描述與“被子”的義項進行比較值會得到一個比較低的相似度值。這啓發我們可以採用如下的語義格測試調整方法：

設詞語 Word 的語義環境爲

$W=((CASE_1\ word_1)(CASE_2\ word_2)\ldots(CASE_i\ word_i)\ldots(CASE_n\ word_n))$

W 的一個義項規則爲 R；

如果詞語含有(CHANGECASE $CASE_i$ $CASE'_i$)的警告標記，同時，W 中存在 $Case_i$ 格且不存在 $Case'_i$ 格，則我們分析 $Case_i$ 的詞語 $Word_i$ 在 W 下的評價分值，如果小於某一預定義的最低閾值，則我們試著將 W 替換爲

W’=((CASE$_1$ word$_1$)(CASE$_2$ word$_2$)…(CASE’$_i$ word$_i$)…(CASE$_n$ word$_n$))

並再次將 W 與義項進行相似度計算，得到了一個新的 Word 的評價分值，如果新的評價分值大於預定義的調整閾值，我們就認爲在分析結構中的詞語 Word$_i$ 的語義格應調整成爲 Case’$_i$。

如處理“被子叠得整整齊齊”一句時，可以首先在語法分析階段的語法詞典中爲“叠”這樣具有主語充當受事體的動詞做上標誌(CHANGECASE AGENT THEME)，預選定義最低閾值與調整閾值分別爲 20 與 85。

在語義排歧時我們可以發現原語義環境 ENV 中詞語“被子”在“叠”下的評價分值很低，我們用新的語義環境

ENV’：(叠 (THEME 被子)(RESUTL 整整齊齊))

重新與“叠 1”的規則進行比較計算，調整後“被子”在“叠”下的評價分值達到了 100 分，高於調整閾值，我們就可以認爲在此結果框架中的 AGENT 應調整爲 THEME。

3.5.2 確定分析結果的語義關係

漢語中存在某些特殊的句型，如“VP+NP”、“VP+的+是+NP”或“VP+的+NP”，或是“NP+VP”、“NP+的+VP” 等偏正結構，其中的 NP 在 VP 的語義環境中所充當的語義關係格不是固定的。我們把 NP 當作 VP 語義環境的 Parent 格詞語。NP 可以作 VP 的任何一個不存在的語義成分。例如：

1：裝修圖書館的工人整整忙了一天。(“工人”做“裝修”的 AGENT)
2：天下哪有白吃的午餐？(“午餐”作“吃”的 THEME)
3：我們以前住的地方現在是片廣場了。(“地方”作“住”的 LOCATION)
4：她的打擾會使他感到厭煩。（“她”作“打擾”的 AGENT）
5：感謝你對藝術事業的大力支持！(“事業”作“支援”的 THEME)

從上面的例子我們可以看到，NP 在 VP 中可以充當的成分是很豐富的[1]。在進行詞義選擇時，我們根據詞語的語義環境與義項限制規則進行相似度計算，需要確定這個 NP 在 VP 語義環境中的語義格。

與 3.5.1 的思路相同，我們也可以採用測試比較法來解決 NP 的成分確定問題。下面主要針對“VP+DE+NP”或“VP+NP”類的定中結構進行討論，對於“NP+VP”或“NP+的+VP”型的結構，也可同樣進行處理。

設 VP+DE+NP 或 VP+NP 結構中 VP 的中心動詞 Word 的語義環境爲

W＝((PARENT PWORD)(CASE$_1$ word$_1$)……(CASE$_n$ word$_n$))

其中 PWORD 是 VP 短語的修飾詞語。

現有 W 的一個義項規則 R：

$R=((CASE'_1 \quad CASESC_1)(CASE'_2\ CASESC_2)\ \ldots(CASE'_m\ CASESC_m))$

我們可以將 W 中的 PARENT 格依次替換爲在 R 中存在而在 W 中不存在的格，當然替換的格必須不違反 W 中的某些對成分要求的語法限制，例如當 W 是不及物動詞時，Parent 就不能替換成爲 THEME 格。Parent 替換後得到的新語義環境 W'與 R 進行相似度的計算。設當 Parent 換成 $Case_i$ 時取得最大的相似度值 V，且 V 大於一個預定義的閾值，則我們可以確定 Parent 在 VP 中充當 $Case_i$ 的語義成分。

如果 $\{CASE'_1, Case'_2, \ldots,Case'_n\}-\{Case_1,Case_2,\ldots,Case_m\}=\Phi$ ，或者 V 小於閾值，則表明被修飾的 NP 在 VP 不能充當合適的語義成分，或語義成分不能確定。這種情況如

1：他中獎的消息立刻傳開了。(“消息”爲“他中獎”的同指)
2：創作方法很重要。 (“方法”實際上是“創作”的“方式”成分，但創作的規則中沒有對“方式”的限制描述)

對於上述這兩種的情況，可以將詞語與通用格規則進行匹配，具有最大匹配相似度所對應的格可作爲被修飾詞在 VP 中所充當的語義成分。

3.6 利用詞義排歧進行結構排歧

一個測試句經過語法分析後有可能產生多個中間結果，這就需要對中間結果進行評價，從中優選出一個最優結果。即進行結構排歧。

漢語的結構歧義錯綜複雜，許多的漢語言文學研究學者都對其進行深入的研究，並總結了許多的歧義短語組合格式。

本文對結構歧義的消除演算法是建立在詞義排歧的基礎之上。最佳的中間結果應是最符合語義與常識的，而中間結果“優選”的原則也應是選擇最滿足語義的結構。在“約束”排歧及本章前面介紹的詞義排歧中，我們對詞語進行義項選擇的同時還得到了詞語的評價分值，它們反映了詞語在當前語境中符合語義的程度。前文中定義了詞語的限定關係，那些起限定作用的詞語在當前結構中起著關鍵的搭配作用，它們對結構的語義具有最直接的影響，因此我們在對一個中間結果進行詞義排歧後，可以將排歧得到的各限定詞的評價分值的總和作爲當前分析結果的評價值。評價值的大小是優選的根據，具有最高評價值的中間分析結構就作爲最終的結構排歧結果。

例如：“vp+np+的+np”結構可能存在兩種歧義：

1： (vp (np 的 np)) 如“修理爸爸的自行車”
2： ((vp np) 的 np) 如“修理自行車的爸爸”

這兩句話可能存在 4 種歧義結構，對它們分別進行詞義排歧，得到的評價情況如下：

表 3.6 各分析結構的評價分數

分析結構	各詞語評價分數			分析結構評價分數
	修理	爸爸	自行車	
((修理 爸爸)的 自行車)	0	0	25.0	0.0
(修理 (爸爸 的 自行車))	89.28	0	89.28	89.28
((修理 自行車) 的 爸爸)	94.64	100	89.28	94.64
(修理(自行車 的 爸爸))	0	0	0	0·0

從表中可以確定(修理 (爸爸 的 自行車))與((修理 自行車)的 爸爸)分別是句 1 與句 2 的最佳分析結果。

3.7 實驗結果及討論

3.7.1 實驗結果

在義原的同現集合獲取中我們使用《讀者 20 年文集》作爲統計語料庫，測試時也使用同類的語料。測試目的是檢驗本文提出的排歧演算法的是否有效，同時通過中間語言的結構優取及調整的正確率來考查詞義評價模型是否合理。我們從語料庫中選取了 2，000 個測試句進行排歧實驗。下表是測試的指標及測試結果：

表 3.7 測試結果

測試指標	指標描述	測試值
詞義排歧的正確率	詞義判斷正確的詞語數/測試語料中歧義詞的總數	0.92
結構排歧的正確率	結構選擇正確的句數/測試語料中有多個候選分析結構的句數	0.82
結構調整的召回率	進行結構調整的結果數/測試語料中需進行調整的分析結果數	0.95
結構調整的正確率	經自動調整後正確的結果數/進行自動結構調整的結果數	0.98

測試結果表明，利用語料庫的同現義原來構造義原的語義限制規則，並以此進行詞義排歧的思想是合理的。並且在詞義排歧的過程中同時也能實現高正確率的結構排歧、結構調整。

3.7.2 演算法存在的問題

(1)：演算法找出的義原規則是具有普遍搭配關係的詞義排歧規則，我們通過義原的限制來構造義項的規則，有可能義原規則限制的粒度比義項應有語義限制大；這就造成了產生的義項規則語義粒度過大。例如象“找”一詞，它在作“找了兩元錢”時的義項的

定義是“return|還”，這一義項中的搭配受事體一般來說是和錢財有關的，但“return|還”的義原規則中的受事體的語義特徵不一定是和錢有關，這就有可能導致如“找書”中的“找”也會誤選爲“Return”。

(2)：《知網》中詞語含有的一些文言用法對干擾了詞義的正確選擇。如“去”一詞，在《知網》中有“leave|離開”的義項解釋，這種用法在現代文中很少出現，而它的存在對會影響“去”的常見義項“go|去”的選擇。

3.8 小結

本演算法具有如下特點：

(1) 在《知網》中詞語的義項由多個義原定義，不象傳統的分類詞典中義項只是一個類代碼，這樣對詞語義項的意義描述更加全面，豐富。

(2) 利用義原的規則與當前詞語所在語義環境進行相似度的比較進行詞義排歧，可以提高複雜句式結構的詞義排歧正確率。

(3) 將排歧知識建立在義原的基礎上，義原的數目是有限的，這樣避免了手工編制大規模詞義排歧知識的繁重勞動。同時義原的排歧知識是參考義原的同現集合，而這一集合是通過對語料庫無指導學習獲取的。這樣知識的獲取的工作量進一步的減少了。

(4) 利用詞語義項的評價演算法，在詞義排歧過程中可以對中間語言的語義框架做適當調整。

(5) 在進行詞義排歧的同時解決多個分析結果的結構排歧。

4. 總結

漢英機器翻譯中要解決分析產生的辭彙歧義、語義歧義，得到一個比較好的句法分析結果，必須要引進語義知識，進行語義分析。本文構造了一個用於機器翻譯文本分析的語義排歧模型，它能夠結合語義知識進行有效的詞義消歧，進一步進行結構消歧。

本文提出的語義排歧系統有以下特點：

1：在句法分析過程中進行的約束排歧大大減少了中間語言的生成數目，減輕了針對中間結果進行的“優選”工作。

2：在“優選”的方法中我們利用了從語料庫中構造義項語義限制規則的方法，減輕了人工制定語義知識的工作量。同時語義規則也避免出現單純利用統計進行語義排歧時對複雜句式的處理效果不理想的現象。

3：排歧過程中給出的義項評價值可以使得系統在完成排歧的同時可以進行別的與

語義有關的處理，如語法格的調整等。

本文所提出的語義排歧模型已在機器翻譯系統中具體地實現。實驗例句的測試表明這一排歧對處理辭彙歧義、結構歧義是有效的。

由於研究的時間關係，本文的排歧模型從功能上來看還只是一個實驗系統，還有不少可以改進的地方。例如：

1：本文主要是在詞歧消歧的基礎上進行結構消歧，因此只考慮了常用的幾種歧義格式，還有很多其他的歧義結構有待總結與處理。同時一些如“吃食堂”，“吃大餐”等具有特殊的語義格轉換歧義的現象，需要進一步的深入研究。

2：本文的語義排歧模型主要針對實義詞進行，對於虛詞詞語歧義問題沒有過多地考慮。

3：本文處理都是單句漢語的歧義排歧，未涉及到篇章級的上下文語境理解。而實際上不少歧義需要放在上下文中才能得以消除。

以上這些問題，都有待在後繼的工作中不斷地加以補充與改進，使用語義排歧模型更加有效、實用。

參考文獻

馮志偉 <論歧義結構的潛在性>《中文資訊學報》，1995，第 9 卷(4)

呂叔湘 <歧義類型>《中國語文》1984 年第 5 期

馮志偉 <歧義消解策略初探>《計算語言學進展與應用》清華大學出版社，1995

苑春法，黃錦輝，李文捷 <基於語義知識的漢語句法結構排歧> 《中文資訊學報》第 13 卷第 1 期

Dan Roth. “Learning to Resolve Natural Language Ambiguities: A Unified Approach.” AAAI-98, 1998

趙鐵軍等《機器翻譯原理》·哈爾濱：哈爾濱工業大學出版社，2000

Wilks, Y. Stevenson, M. “Word Sense Disambiguation Using Optimized Combinations of Knowledge Sources”, In *Proceedings of joint COLING-ACL'98*. 1998. Montreal, Canada.

Philip Resnik ,David Yarowsky. A Perspective on Word Sense Disambiguation Methods and Their Evaluation. *Proceedings of the SIGLEX Workshop* "Tagging Text with Lexical Semantics: What, why and how?", pp. 79-86, Washington, D.C.

董振東，董強·《知網》·http://www.how-net.com

李涓子，黃昌寧 <基於轉換的無指導詞義標注方法>《清華大學學報》（自然科學版），1999 年第 39 卷(7)

梅家駒 《現代漢語搭配辭典》漢語大詞典出版社. 1999 年 12 月第 1 版

李涓子，黃昌寧 ＜一種無指導的詞義排歧模型＞《計算語言學文集》北京：清華大學出版社，1999

楊曉峰，李堂秋，洪青陽 ＜基於實例的漢語句法結構分析歧義消解＞《中文資訊學》報，2001 年第 15 卷

詹衛東 ＜現代漢語 VP 結構定界各結構關係判定＞《北京大學碩士學位論文》，北京大學，1996