学校编码：10384　　分类号＿＿＿＿密级＿＿＿＿
学　　号：199928001　　UDC＿＿＿＿＿＿＿＿＿＿

学　位　论　文

# 面向机器翻译的汉语短语语义模式规则研究

郑 旭 玲

指 导 教 师：李堂秋　教授
申请学位类别：硕　　士
专 业 名 称：计 算 机 应 用
论文提交时间：2002 年 5 月
论文答辩时间：2002 年 6 月
学位授予单位：厦 门 大 学
学位授予时间：2002 年　月

答辩委员会主席：＿＿＿＿＿＿
评　　阅　　人：＿＿＿＿＿＿

2002 年 5 月

# 摘　要

在汉语中，短语具有特别重要的地位。全面而系统地研究汉语短语的构造原则，不仅能为短语分析提供有效的指导，提高短语句法语义分析的正确率，而且还能辐射对汉语词和句子的研究，促进汉语文本分析质量的全面改善。

本文以《知网》为主要语义知识源，对汉语短语构造的语义规律的表示、获取以及应用作了系统的研究。本文在深入研究《知网》知识词典描述语言和汉语短语构造的语义规律基础上，设计了用于形式化地描述这些规律的语义模式规则表述方法。在此基础上，本文设计并实现了基于语料库的二元语义模式规则自动挖掘和优选算法，该算法先采用数据挖掘中元规则制导的交叉层关联规则挖掘方法，自动发现汉语短语熟语料库中词语两两组合的语义规律，再根据统计结果自动优选后转换生成候选二元语义模式规则集。然后，依据人工归纳的汉语短语组合规律进一步对此规则集进行优化调整和扩充，以增强其词义排歧和结构排歧能力，得到一个较完善的语义模式规则库。结合课题组已开发的汉英机器翻译系统 XMMT 的特点，本文设计并实现了基于语义模式规则的语义分析排歧算法，从而在句法分析过程中引入汉语短语构造的语义规律进行词义和结构排歧。

实验表明：1）本文设计的语义模式规则能够较准确地刻画汉语短语构造的语义规律；2）本文提出的基于语料库的二元语义模式规则自动挖掘和优选算法是切实可行的，它大大减少完全由人工从大规模语料库中总结规则的工作量，避免了纯人工编制规则的主观性和片面性；3）本文提出的语义分析排歧算法能够有效消解短语分析中的词义歧义和结构歧义。

**关键词：**自然语言处理、机器翻译、语义分析、语义规则、关联规则、数据挖掘、模式匹配、语料库、消歧、知网

# Abstract

Phrase plays an important role in Chinese. The investigation in the construction of Chinese phrase can not only provide effectual guidance for phrase parsing, raise the accuracy of the syntax and semantic parsing, but also facilitate the research of Chinese word and sentence, improve the quality of Chinese text parsing.

In this paper, the representation, acquirement and application of the construction of Chinese phrase were investigated systematically. *HowNet* is used as main semantic resource. Based on thorough study of the Knowledge Dictionary Mark-up Language (KDML) of *HowNet* and the construction of Chinese phrase, a presentation of semantic pattern rules was designed to formalize the construction. Upon this foundation, a corpus-based algorithm was designed and implemented to acquire and filter binary semantic pattern rules automatically. In the algorithm, a data mining method for cross-level association rules is adopted, which is guided by metarule, to find the semantic laws of word combinations in Chinese phrase corpus. Then statistic results are used to filter the findings. In the end, the remains are transformed into binary semantic pattern rules. In order to enhance their abilities of word sense disambiguation and structure disambiguation, these rules were optimized and expanded according to rules of word combinations concluded artificially, and a set of semantic pattern rules was acquired. Combining the feature of the Chinese-English machine translation system XMMT, a semantic based disambiguation algorithm was designed and implemented. With the algorithm, word sense disambiguation and structure disambiguation can be done by semantic pattern rules matching during syntax parsing.

The experiment result indicates that: (a) The presentation of semantic pattern rules can formalize the construction of Chinese phrase quite well; (b) The corpus-based algorithm for acquiring and filtering binary semantic pattern rules is effective, and it can reduce the human labor, avoid subjectivity and unilateralism caused by writing rules manually; (c) The semantic based disambiguation algorithm can achieve satisfactory effects.

**Keywords:** natural language processing, machine translation, semantic analysis, semantic rule, association rule, data mining, pattern matching, disambiguation, corpus, *HowNet*

# 目　录


摘　要 ..... I
Abstract ..... III

第一章 绪论 ..... 1
1.1 课题的提出 ..... 1
1.1.1 提高短语分析质量的意义 ..... 1
1.1.2 短语分析中的歧义消解问题 ..... 1
1.1.3 语义知识在短语分析排歧中的运用 ..... 2
1.2 国内外相关的研究工作 ..... 2
1.3 本文的主要工作 ..... 3

第二章 系统结构与语义模式规则的表示 ..... 5
2.1 系统结构描述 ..... 5
2.2 系统主要语义知识源简介 ..... 6
2.2.1 《知网》简介 ..... 6
2.2.2 《知网-中文信息结构库》简介 ..... 7
2.3 语义模式规则的表示形式 ..... 8

第三章 二元语义模式规则的自动获取 ..... 11
3.1 总体思路 ..... 11
3.2 关联规则挖掘的基本概念 ..... 11
3.3 汉语短语语料库的构建和数据预处理 ..... 13
3.3.1 汉语短语语料库的构建和语料加工 ..... 13
3.3.2 面向关联规则挖掘的数据预处理 ..... 14
3.4 二元语义模式规则挖掘问题描述 ..... 15
3.5 发现子目标模式 ..... 16
3.5.1 由(k-1)-子目标模式集合 $L_{k-1}$ 生成候选 k-模式集合 $C_k$ ..... 17
3.5.2 压缩事务数据 ..... 20
3.5.3 子目标模式发现算法 ..... 21
3.6 由子目标模式产生二元语义模式规则集 ..... 23
3.7 实验及结果分析 ..... 26

第四章 语义模式规则的优化和扩充 ..... 28
4.1 增强语义模式规则的词义排歧能力 ..... 28
4.1.1 主要策略：引入受限变量 ..... 29
4.1.2 “mp np”中的词义消歧——一个典型例子 ..... 29
4.2 扩充多元语义模式规则 ..... 31
4.2.1 多元语义模式规则扩充原则 ..... 31
4.2.2 多元语义模式规则举例 ..... 31

第五章 基于语义模式规则的语义分析排歧 ..... 33

5.1 语义分析排歧的总体思想.....33
5.2 在句法分析规则中加入语义限制.....33
5.2.1 句法分析规则的形式.....33
5.2.2 在句法分析规则中嵌入语义评价函数.....34
5.3 基于语义模式规则的语义评价算法.....35
5.3.1 语义模式规则匹配度评价.....35
5.3.2 候选结果的语义评价.....37
5.3.3 语义评价算法描述.....37
5.4 语义分析排歧示例.....38
5.5 实验结果.....40

第六章 结束语.....41

攻读硕士学位期间发表的论文.....42
参考文献.....43
致 谢.....45

# 第一章 绪论

## 1.1 课题的提出

语言是人类社会文化的主要载体和信息交流的重要工具。随着经济的全球化、计算机的普及和 Internet 的迅猛发展，世界上不同地区的经济文化联系日益密切，人们日常工作生活的信息化和国际化程度不断提高，语言的差异已成为信息交流中面临的严重障碍。单纯依靠人工翻译，工作量大且速度有限，人们对实用的机器翻译或机器辅助翻译系统的需求急剧增长。尽管经过无数机器翻译专家们的执著研究和不断探索，机器翻译无论在理论技术还是在实际应用方面都取得了长足的进步，但不可否认的是，现有机器翻译系统的翻译质量还难以满足实用需要。此外，文本自动分类、文献检索、信息提取、自动文摘、语音识别与合成等自然语言处理技术的发展，也对自然语言自动分析质量提出了越来越高的要求。

### 1.1.1 提高短语分析质量的意义

构成自然语言文本的最基本单位是文字（如字母和汉字），而由文字构成书面语言的过程并不是简单的堆砌，而是可以大致划分成下面这些语法层次：字母（汉字）——词素（语素）——词语——短语（词组）——句子——句群——段落——篇章。语言的这些语法层次的客观存在为自然语言的自动分析提供了方便。实际上，几乎所有的自然语言分析算法都是按照词——短语——句子这三个层次顺序进行的。

宏观地看，汉语文本自动分析技术已经走过了字处理阶段，分词和词性标注（词处理阶段）也有了基本可以实用的成果[1]。20 世纪 90 年代以来，汉语分析理论在句法、语义、语用多个层面，以及从词到短语结构、到句子句式句型，乃至篇章话语等各级语言单位，形成了多层面、多角度、全方位的研究态势。然而，由于现有技术条件和语言学研究成果在形式化及可操作性上的限制，就总体发展情况而言，汉语分析的实用技术研究还处在句处理阶段。而如何提高短语分析质量正是现阶段迫切需要解决的重点问题之一。

从汉语语法体系的特点看：（1）正如朱德熙先生指出的“汉语句子可以看作是由词组实现得到的，大多情况下构造跟词组也基本一致”[2]，即汉语句子的构造原则与短语的构造原则基本一致；（2）汉语词、短语、句子这三级语法单位形式上存在连续性，没有天然分隔界限；（3）短语可以由简单而复杂套叠生成更大的短语[3]，语言结构的递归性在短语一级上体现得最为鲜明。这些因素综合决定了处于中间的短语的特殊地位。正是由于对词的属性判断可以直接在分析短语的过程中得到检验，对句子的分析理解也可以转化为对短语的层层剖析，因此全面系统地研究汉语短语的构造原则，不仅能为短语分析提供有效的指导，从而提高短语结构和语义分析的正确率，而且还能辐射对汉语词和句子的研究，促进汉语文本分析质量的全面改善。

### 1.1.2 短语分析中的歧义消解问题

歧义是自然语言中普遍存在的语言现象，它实质上是意义与形式之间的矛盾问题。同一形式与不同的意义相联系，就必然会产生歧义，这是自然语言不同于人工语言的特点之一[4]。歧义现象的存在给包括机器翻译在内的自然语言处理设置了巨大的无法回避的障碍，解决歧义问题是机器翻译研究的核心任务之一。单就短语自动分析而言，需面对的主要歧义有词汇歧义（lexical ambiguity）和结构歧义（structural ambiguity）。

词汇歧义，即单词的同形歧义，也就是一词多义。确定特定语境中的多义词的含义，即词义消歧，需要涉及词法、句法、语义和语用等多层次的知识，始终是自然语言处理研究中的一个难题。尽管在短语分析过程中缺乏全局的句法语义和语境信息，但由于多义词的不同含义往往显现在它跟不同词的搭配中，而且词与词组合成短语时要受到语义和句法搭配关系的制约，因此，在短语分析过程中，可以利用短语构造的句法语义规律来消解一部分词汇歧义。

结构歧义，即短语的同形歧义，指的是同一词类序列映射成不同结构层次、不同句法关系或不同语义关系的语言结构。语言学家从句法形式、语义关系等层次对结构歧义的类型、歧义现象产生的过程等问题作了大量的研究工作[4,5,6,7]，但结构歧义及其消解问题在语言学理论上尚未获得完全解决。而且这些研究大多是从人的角度来考虑的，与从自然语言的机器自动处理角度来考虑又有很大的不同。对自动处理而言，指出歧义只是解决实际问题的起点而不是终点，还必须找到切实可以用来排除歧义的因素，并表示成可操作的形式化知识。

从一个歧义格式[5]的不同结构层次划分对周围环境（context）的影响程度不同着眼，可分为外显型歧义格式和内含型歧义格式两种不同的歧义类型[8]。外显型歧义格式取不同结构划分时，它与周围句法单元的组合能力发生显著变化，如“vp np <的> np”、“v n”等等，大多可通过其与上下文的相互制约关系来解决。而内含型歧义格式的内部结构变化对其外部组合能力影响不大，如“ap np np”、“mp mp <的> np”等等，常常是由内在组成成分之间的制约关系来决定。因此，在分析短语时，应重点解决内含型歧义，为后继的整句乃至整个篇章的分析扫除不必要的干扰。

### 1.1.3 语义知识在短语分析排歧中的运用

要解决自动分析汉语短语时遇到的歧义问题，语义知识是不可或缺的。

首先，语义知识的运用有助于消减短语中的词义歧义。词语搭配除了要受句法上的限制之外，还要受语义上的限制。如果我们认定待处理的句子和短语都应该是有意义的话，那么，我们就可以运用短语内部组成成分之间的语义搭配限制来鉴别并剔除那些无意义（即不符合语义）的分析。这与句法分析时运用句法限制来排除不合法的分析是一致的。

其次，语义知识的运用有助于理清短语的句法语义结构。从句法层面看，短语的内含型歧义对于周围的句法单元基本上是封闭的，句法手段对其无能为力，而且其它外部环境的制约条件也更不确定[8,9]，因此，我们更需要关注的是它的内部组成成分之间的语义搭配规律，例如“新英汉字典”和“现代汉语字典”。另一方面，尽管外显型歧义大多可通过它与上下文的句法语义制约关系来解决，但对于大部分并非真正有歧义的实例往往在短语分析时就可运用语义知识来排歧了，例如“削苹果的刀”和“削苹果的皮”这两个短语，无需考察短语所处的上下文，依靠关于“苹果”、“皮”和“刀”的语义知识就可以消歧。

综上所述，如果我们能在短语句法分析过程中引入语义知识，具体说来，就是关于汉语中具有什么语义的词语可以相互组合、以怎样的方式组合成怎样的短语的语义知识，就可以尽早将那些无意义或不正确的分析剔除出去，从而有效改善短语分析质量并加速整个分析进程。不过，就目前理论研究的水平而言，我们认为比较适宜走以句法为主，语义为辅的路线。

## 1.2 国内外相关的研究工作

国外的自然语言处理研究，近一二十年来在范畴知识的研究上进展比较显著，其中语义方面尤为突出。

关于自然语言的语义分析理论，大致有语义场理论（Semantic Field）、义素分析理论（Componential Analysis）[10]、配价理论（Valence Grammar）[11]、格语法（Case Grammar）[12]、

论旨理论（Theta-Theory）[13]、概念依存理论（Concept Dependence）、语义网络（Semantic Network）以及蒙太格语法[14]（Montague Grammar）等。不同的语义理论在理论背景和侧重点上各有不同，比如：配价理论出自依存语法，重视语义的句法对应性；格语法在转换生成语法之后提出，重视纯粹的动名语义关系的发掘；论旨理论作为生成语法理论的一个子系统，把动词的句法搭配限制和语义搭配限制结合到一起统一处理；概念依存理论侧重表达概念之间的依存关系；蒙太格语法侧重通过演算得到句子的逻辑语义表达式。

从大规模描述自然语言语义知识的工程实践来看，以英语和其它主要欧洲语言为描述对象的研究中，影响较大的有 WordNet、MindNet 和 FrameNet 等计算机用语义词典的开发，基于统计方法的研究工作如 Chodorow[15]和 Ide[16]等人所作的有关从机器词典中自动抽取语义知识的研究工作，以及黄居仁[17]等人从汉语名量搭配词典中自动抽取汉语名词语义分类的研究。在面向实用系统的研究中，用复杂特征集以及合一运算来组织语言规则知识，广泛地吸收各种形式化表示方法的长处，体现实用特色，其代表性工作如 Jenson Karen 等人的研究。

20 世纪 80 年代朱德熙先生提出的词组本位语法体系[18]比较全面地建立了汉语短语结构层面的句法范畴，同时也引入了一定的语义范畴。20 世纪 90 年代以来，国内汉语研究界广泛讨论了“三个平面”的语法观[19]，此外，国内学者开始采用格语法、配价语法、语义指向分析、依存语法、GB 理论、话语分析等新方法进行汉语研究的新尝试。

与此同时，学术界也开始了语言知识基础工程的研究和建设，突出表现在对汉语句法范畴和语义范畴的符号化及大规模的词典化上，具体成果以“现代汉语语法信息词典”、“信息处理用现代汉语语义词典”、“现代汉语述语动词机器词典”、HowNet 等机器可读词典为代表。

与上述范畴知识的研究工作相比，规则知识的研究相对薄弱一些。国内在直接面对中文信息处理的研究工作中，具体规则的研究方面如马真、陆俭明[20]、孙宏林[21]等。此外，相当多的汉语研究工作虽然不是直接面向中文信息处理，但其中有不少研究实际上可以看作时跟发现汉语的语法规则紧密相关的，只不过这些研究通常都带有比较强的描写色彩，如沈阳基于配价理论和生成语法的空范畴理论对汉语动词句位结构以及 NP 所做的全面分析，袁毓林[22]等人关于汉语动词的配价研究等。

相比之下，对规则知识作大规模的系统研究还比较少见，如：詹卫东[23]侧重从句法层面上归纳现代汉语短语结构的组合规则，董振东和董强[24]通过对信息结构的描述来认识中文是如何描述诸如万物、部件、属性等等概念的，或如何由简及繁地表达意义的，从而揭示中文的语言结构的规律。

总体说来，尽管国内外直接以在句法语义层面发现结构组合规则为目标的研究不多，但国内外学者所作的这些研究工作都为本课题的具体研究提供了广阔的参考。有关的具体研究工作，在下文讨论具体问题时还会提及。

## 1.3 本文的主要工作

本文的主要研究目标是侧重从语义层面上归纳汉语词语组合的语义规律，并将其直接引入到机器翻译系统的文本分析过程，用于消解分析汉语短语时遇到的词义歧义和结构歧义，提高短语乃至文本整体分析质量。

本文主要从以下几个方面对课题的研究工作进行探讨：

1. 为形式化地描述汉语中词语组合的语义规律，设计了语义模式规则表示方法。
2. 构造了一个汉语短语语料库，并以《知网》为主要的语义知识资源，对语料进行了词性、义项、语义关系等标注。
3. 将基于规则和基于统计的方法相结合，设计并实现了二元语义模式规则自动挖掘算法和优

选算法。先采用数据挖掘中元规则制导的交叉层关联规则挖掘方法，自动发现汉语短语熟语料库中词语两两组合的语义规律；再根据统计结果进行自动优选后转换生成候选二元语义模式规则集。

4. 结合语言学家归纳的汉语词语组合的语义规律和实际测试结果，对自动获取的二元语义模式规则进行调整和优化，并适当扩充多元语义模式规则，最终形成一个较完善的语义模式规则库。
5. 设计并实现了基于语义模式规则的语义评价算法，并将其嵌入相应的句法分析规则内，使机器翻译系统能在句法分析过程中同步消解词义歧义和结构歧义。

# 第二章 系统结构与语义模式规则的表示

为了消解分析汉语短语时遇到的词义歧义和结构歧义，从而提高短语乃至文本整体分析质量，需要对汉语词语组合的语义规律进行全面而系统的研究和归纳，并将这些规律形式化地描述为可操作的规则，实际应用到汉语分析系统中。本文以《知网》为主要语义知识源，对汉语短语构造的语义规律的形式化表示——语义模式规则的表述、获取以及应用作了系统的研究。

本章将介绍语义模式规则的表示形式、语义模式规则获取模块的工作流程及引入语义分析排歧算法后的 XMMT 系统结构，同时还简要介绍了本文研究中不可或缺的语义知识资源——《知网》。

## 2.1 系统结构描述

本文的研究工作是建立在课题组已开发的汉英机器翻译系统 XMMT 基础之上的。XMMT 系统采用的是基于中间语言的机器翻译方法，由汉语分析和英文生成两部分组成。分析部分吸取了短语本位语法体系的思想，在充分发掘语法分析潜力的基础上，运用“约束”和“优选”相结合的语义排歧方法来实现词义与结构的消歧[25,26]。先期的工作对句法语义在汉语分析消歧中的运用作了有益的尝试，为本文的研究奠定了良好的基础。

然而，原系统在运用语义知识对短语分析进行“约束”消歧方面只做了一些初步探索和尝试[26,27]，研究的深度还不够，尤其是在词义和语义关系排歧上。指导消歧的语言知识（尤其是语义知识）的不足以及语义评价算法的误差，不仅制约了“约束”消歧作用的发挥，而且也对分析结果的“优选”造成了干扰甚至误导。因此，本文的主要研究目标就是针对 XMMT 系统在短语分析排歧上的不足，扩充汉语词语组合的句法语义知识，改进相关的算法，使其能更好地消解词义歧义和结构歧义，提高分析质量。

为了对短语分析排歧所需要的语义知识进行全面而系统的研究和归纳，本文在 XMMT 系统上外挂了一个语义模式规则获取模块。该模块采用基于规则与基于统计相结合的方法，其工作流程大致是：先依据语言学家提供的初始语言知识，对汉语短语语料库（由《知网－中文信息结构库》中的实例和从真实文本中抽取的实例组成）进行半自动加工（标注了各词语的词性、义项和词语间的句法结构、语义组合关系）；再运用数据挖掘中元规则制导的交叉层关联规则挖掘方法来自动发现熟语料库中词语两两组合的语义规律，并根据统计结果经优选算法自动筛选后转换生成候选二元语义模式规则集；然后根据语言学家归纳的语言知识和实际测试结果对候选语义模式规则进行人工优化调整和扩充，最终形成短语分析排歧所需的语义模式规则库。该方法大大减少完全由人工从大规模语料库中总结规则的工作量，避免了纯人工编制规则的主观性和片面性，也弥补了纯统计方法难以处理复杂和特殊语言现象的不足。

总结在语义模式规则研究过程中的体会和发现，本文修改了原系统的语义模式规则的表示方法，提出了新的基于语义模式规则的语义分析排歧算法，并改进了相关的实现机制。修改后的语义模式规则的形式化表示将在本章的第 3 节给出，而语义模式匹配算法和实现机制将在第五章中详细探讨。

此外，本文在修改 XMMT 系统的过程中注意到，原系统在对汉语短语进行句法语义分析时，对于每种句法分析结果仅保留当前局部分析结果中最优的一种语义取值。这种做法过于武断。汉语缺乏形态变化属于意合分析型语言，因而无论是词义排歧还是结构消歧往往需要涉及词法、句法、语义和语用等多层次的知识。当汉语分析过程出现歧义时，经常无法立即获得排歧所需的足够信息，如果匆忙作出判断，难免导致排歧错误。为此，本文借鉴“渐进歧义消解法”（incremental

disambiguation）的思想，对 XMMT 系统分析部分的结构及消歧处理流程作了修改。修改后的分析排歧流程大致如下：在进行句法语义分析时，如果根据短语各组成成分的句法语义限制足以对被分析的短语的词义歧义或结果歧义进行判断，就立即做出判断消解歧义，否则就把无法消解的歧义留在中间分析结果中，等待后继分析得到的排歧信息递增到足以对其作出判断时再消解；倘若当前的句子无法提供足够的排歧信息，才在最后对分析结果进行“优选”时根据综合评价的得分选出最佳结果（本文中提到的“最佳”都是相对本文所采用的评价算法而言的，并不一定总是实际上的正确解，也不一定唯一）。

综合上述修改后的 XMMT 汉英机器翻译系统分析部分的工作流程如图 2.1 所示。其中，虚线框内的部分为语义模式规则获取模块的工作流程。

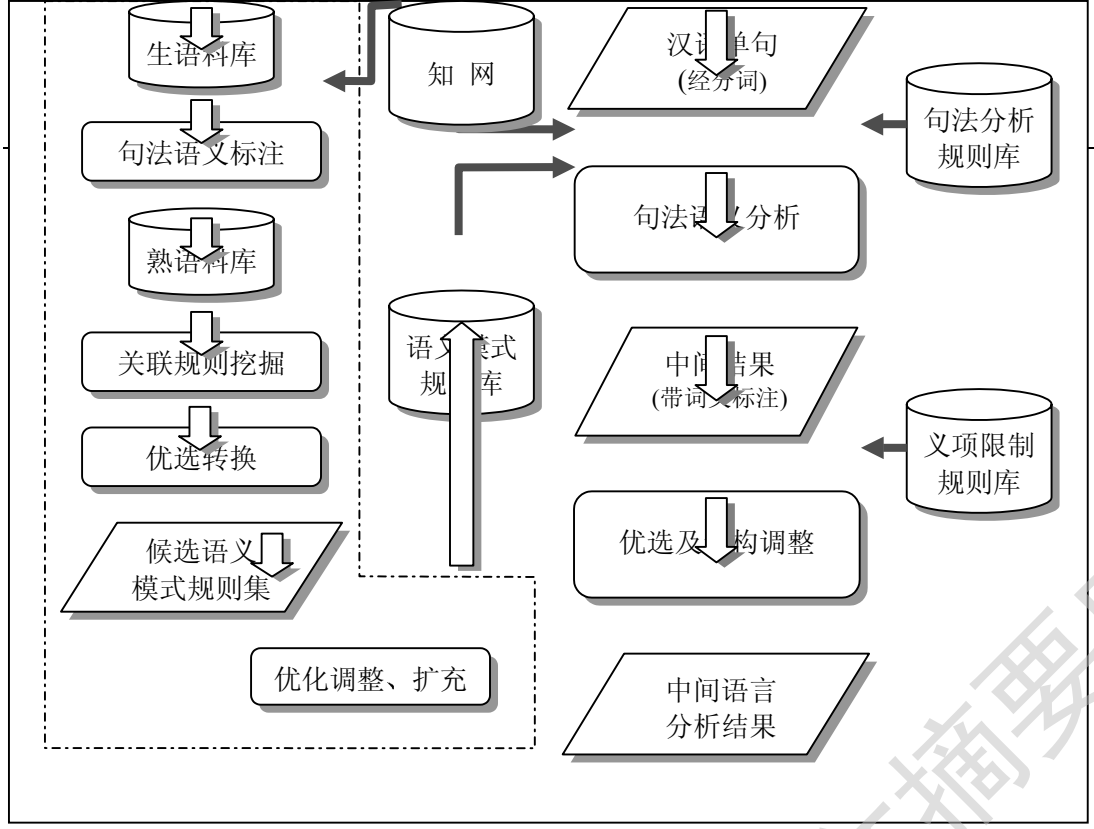
从上面对系统结构的描述可以看出，语义模式规则在本文所提出的短语分析排歧模型中的重要地位。为了方便对语义模式规则表示方法的讨论，我们先对本文研究工作中使用的主要语义知识资源——《知网》作简要的介绍。

## 2.2 系统主要语义知识源简介

董振东先生花费十余年心血构建的《知网》在语义知识的形式化描述方面开展了卓有成效的工作，为本文的研究工作提供了重要的语义知识支持。

### 2.2.1 《知网》简介

《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[28]，是一个网状的有机的知识系统。知网利用义原、动态角色、属性以及它们之间的语义关系来对知网语义词典中出现的所有词语项进行概念定义。义原是知网中最基本的、不易于再分割的意义的最小单位，用“thing|万物”、“event|事件”等中英文双语标记，目前共有 1503 个。知网将这些义原按照事件、事物、属性、属性值等分类，分别组织成一个个树型的层级网络（以下简称分类树），通过层级关系来反映义原间的上下位关系。此外，义原之间还存在的部分、主体、客体、从属、时空、材料等语义关系，用“%”、“*”、“@”、“?”、“$”等语义关系符来标记。




中英双语知识词典是《知网》的基本文件之一，其中每个词语项的概念定义由若干个义原及它们与主干词之间的语义关系描述组成。概念定义可形式化地描述为：

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

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

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

为了保证概念定义的复杂度和一致性的统一，知网制订了专门的知识词典的描述语言（Knowledge Dictionary Mark-up Language，KDML）。KDML 中对概念定义的义原、动态角色、属性和语义关系符等的使用以及排列顺序有严格的规定，例如：“洗衣机”的概念定义为“tool|用具,*wash|洗涤,#clothing|衣物”，其中“洗衣机”这种“用具”借助于“*”表示其为“洗涤”的施事，而“衣物”借助于“#”表示其为“洗涤”的受事，它们之间的顺序是不可以颠倒的。

### 2.2.2 《知网-中文信息结构库》简介

《知网-中文信息结构库》（以下简称《结构库》）的基础是《知网》，它是《知网》这一知识系统向中文研究延伸的产物。它通过对信息结构的描述，来认识中文是如何描述诸如万物、部件、属性等等概念的，或如何由简及繁地表达意义的，从而也将揭示中文的语言结构的规律[29]。

《结构库》中信息结构的描述对象是《知网》所规定的用于描述万物、部件、属性、属性值、事件、时间和空间等义原，描述内容是中文词语的各个组成部分之间的、由《知网》所规定的动态角色关系或属性。《结构库》中提出了句法分布式、句法结构式和信息结构模式三个概念。句法分布式是指由词性代表的词语基本单元的排列，例如“N + N”、“A + N”等。句法结构式是指由词性代表的词语基本单元的排列以及它们之间的管辖关系，例如“N <-- N”、“A <-- {V <-- N}”等，其中箭头指向被管辖者。而信息结构模式则是用于无歧义地描述由义元代表的词语基本单元的排列以及它们之间的管辖关系，它的描述由四部分内容构成，其排列如下：

SYN_S= 句法结构式

SEM_S= 信息结构模式

{ Query：问题

Answer：答案的句法分布式 }$^+$

例子：{符合该信息结构模式的真实语料的实例}$^+$

其中，Query 和 Answer 表明了该信息结构模式所传达的真正信息，以及由此可产生的问与答。信息结构模式的一个具体实例如下：

SYN_S=N <-- N
SEM_S={(物质/事情/事务) [受事] <-- <事件,行动,处理>} <-- [施事] (组织/场所)
Query1: 那是什么地方？/ 那是什么单位？
Answer1: N + N
Query2: 什么是 N + N?
Answer2: “管”N1“的”
例子：人事-局，人事-处，干部-部，国务-院，税务-局，税务-所，劳动-局，
邮政-局，邮电-部，电信-局，林业-部，林业-局，财政-厅，财政-部，
粮食-局，粮食-厅，文化-部，文化-局，交通-部，交通-部门，.....

目前公布的《结构库》包含 47 种句法分布式、57 种句法结构式和 268 种信息结构模式，并附带着一万多实例。它的素材来源于实际语料，又经过了人工精心筛选整理，它覆盖面宽但又能避免统计价值不高的重复，可作为中文信息处理的袖珍型经典语料库[29]。

## 2.3 语义模式规则的表示形式

在目前的技术水平上，任何语言研究的具体成果要能够为计算机所用，都必须兼顾两方面要求：一是要能够形式化，二是要有普遍的可操作性。针对《知网》知识词典描述语言 KDML 和汉语短语构造的语义规律的特点，本文采用语义模式规则来形式化地描述短语分析排歧所需的关于汉语中具有什么语义的词语可以相互组合、以怎样的方式组合成怎样的短语的语义知识。考虑到 XMMT 系统的开发平台是 GNU Common Lisp，语义模式规则以表型结构书写。语义模式规则的 BNF 形式化描述如图 2.2 所示。

```
<Rule> ::= (<POS-Serial>  <Syntax-Struction-Type>
                          <Semantic-Relation-List>  <DEF-Pattern> {<DEF-Pattern>}+)

<POS-Serial> ::= (<POS>  {<POS>}+)
<POS> ::= n | v | a | p | m | ……

<Syntax-Struction-Type> ::= <Tag>
<Tag> ::= {0 | 1 | 2 | … | 9}+

<Semantic-Relation-List> ::= ({<Semantic-Relation>}+)
<Semantic-Relation> ::= 合成 | 修饰 | 限定 | ……

<DEF-Pattern> ::= (<First-Item-Pattern> {<Sememe-Pattern>}*)

<First-Item-Pattern> ::= (FIRSTITEM  <First-Item>)
<First-Item> ::= <Main-Sememe> | (*HYP* <Main-Sememe> ) | <Variable>
<Main-Sememe> ::= {Hownet 中定义的概念的主要特征}

<Sememe-Pattern> ::= (<Relation>  {<Sememe-or-Variable>}+ )
<Relation> ::= Partof | Material | Feature | Property | ……
<Sememe-or-Variable> ::= (*DEFVAR* <Varialbe> <Domain>) |
                                                  <Sememe-Item>
<Domain> ::= ({<Sememe-Item>}+)
<Sememe-Item> ::= <Sememe> | ([*HYP*] <Sememe>) | <Variable>
```