基于语义规则的汉语短语结构分析排歧初探

郑旭玲　　李堂秋　　杨晓峰　　陈毅东

厦门大学计算机科学系，厦门，361005

E-mail：tqli@jingxian.xmu.edu.cn

摘要：本文提出了一种基于语义规则的制约和优选相结合的汉语短语结构分析排歧方法。文中首先概述了这种方法的总体思想，并对其语义知识资源——《知网》和《知网-中文信息结构库》作了简要介绍，然后较详细地描述了语义规则的表述方式和基于语义规则匹配度的主要排歧算法，最后用一个实例对整个排歧流程作了说明。

关键词 排歧 语义规则 知网 中文信息结构库

Research on Chinese Phrase Structure Disambiguation Based on Semantic Rules

Zheng Xuling　　Li Tangqiu　　Yang Xiaofeng　　Chen Yidong

Department of Computer Science, Xiamen University, Xiamen, 361005

E-mail：tqli@jingxian.xmu.edu.cn

ABSTRACT: In this paper, a new approach based on semantic rules to resolve Chinese phrase Structure ambiguities is proposed. We first present the main idea of this approach and give a brief introduction to its semantic knowledge resource — *Hownet* and *HowNet - Chinese Message Structure Base*. We then describe in details the standard of semantic rules and the main disambiguation algorithm based on the matching degree of semantic rules. Finally, we instance the whole disambiguating procedures.

Keywords Disambiguation, Semantic Rules, *HowNet*, *HowNet - Chinese Message Structure Base*

1　引言

自然语言的歧义问题，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[1]。为了从理论上概括语言中同形歧义结构的类型，朱德熙教授提出了“歧义格式”这一概念。歧义格式是指具有容纳多种句法结构或语义关系能力的词类序列，如“np1 np2 np3”、“ap np1 np2”、“vp1 vp2 np”等都是汉语中常见的歧义格式。然而，正如文献[1]中指出的，在现实的语言表达中，当我们将具体的单词代入到歧义格式中使其实例化为具体的短语时，并不是所有短语都有歧义。例如，当我们把歧义格式“ap np1 np2”实例化为“大 眼睛 男孩”和“新 英汉 辞典”时，它们的句法结构都是唯一的，分别为“((大 眼睛) 男孩)”和“(新 (英汉 辞典))”，而且各成分间的语义关系也是确定的，前者“大”修饰“眼睛”，“大眼睛”属于“男孩”这个整体，后者“英汉”限定了“辞典”，“新”修饰了“英汉辞典”。但是计算机在自动分析此类无歧义短语时，却往往由于无法正确判定其句

法结构而造成了歧义（下面称之为准歧义）。本文探讨的短语结构分析排歧方法所要消除的就是准歧义。如果我们无法消解短语结构分析中的准歧义，这势必影响整个句子的分析效果。相反的，如果我们在分析阶段不仅能够消解短语结构的准歧义，而且还能判别各成分间的语义关系，这将显著提高分析质量，并为后继的生成阶段提供更多有用信息。

然而，传统的基于语法规则的方法和基于统计的方法都无法有效消解短语结构的准歧义。基于统计的方法在分析句法结构方面先天不足，而传统的基于语法规则的方法凭借词类或具体词语的前后修饰成分等语法知识虽然可以剔除那些不合语法的歧义结构，但是对那些符合语法的歧义结构却无能为力。只有引入语义知识，从语义层面上考察各成分间语义组合的合法性，才能有效地消解符合语法的结构歧义。因此，如何在短语的句法分析中引入可用于排歧的语义知识成为一个值得研究的问题。

本文所要探讨的是一种基于语义规则的制约和优选相结合的短语结构排歧方法。

2 基于语义规则的短语结构排歧方法

要在短语句法分析中引入语义知识用于排歧，就必须研究汉语中具有什么语义的词语可以相互组合、以怎样的方式组合以及组合成怎样的短语，进而才能对这些汉语语义知识进行形式化的描述，使其可用于计算。董振东先生花费十余年心血创建的《知网》和《知网-中文信息结构库》为在汉语句法分析中使用语义知识铺平了道路。

利用《知网》和《知网-中文信息结构库》作为语义知识资源， 我们就可以在原有的基于语法规则的分析系统基础上，以规则的形式导入必要的语义知识。具体说来，就是给每一条可能导致歧义的语法规则附带上一个语义规则集，用于描述匹配该语法规则的各成分应当具有什么样的语义才可以组合以及它们的组合方式和语义关系。分析过程也必须作相应的修改。待归约的成分序列不仅必须匹配某一语法规则，而且其各成分的语义还必须至少匹配相应的语义规则集中的一条语义规则，才能归约产生新成分。而且，通过计算成分序列与语义规则集中各语义规则的匹配度，我们可以从中找出一条最匹配的语义规则用于语义归约，并由此确定各成分间的组合方式和语义关系，而那些使得成分序列与该语义规则匹配度最高的各个词语的语义就是此种组合方式下各词语的释义。这样就可以将句法和语义有机地结合起来及时制约歧义结构的产生，同时消除短语中各词语的词类和语义歧义。此外，对于短语句法分析得出的多个候选结构，我们也可以通过对各个候选结构的语义规则匹配度评估来选出一个最优的作为最终分析结果。

3 《知网》和《知网-中文信息结构库》简介

3.1 《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[2]，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，主要分成事件和事物两大类。动态角色和属性用于描述事件、事件及其属性，用“agent|施事”、“experiencer|经验者”、“time|时间”、“restrictive|限定”等中英文双语标记。义原之间存在的部分、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。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作了严格规定，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“#”表示其为“洗涤”的受事，它们之间的顺序不可颠倒。

此外，知网还把事件类、事物类义原以及属性、属性值等分别组织成一个个的层级网络，通过层级关系来反映它们的上下位关系。这些层级网络都是树型结构（以下简称分类树）。

3.2 《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》）以《知网》为基础，着眼于揭示中文的语言结构的规律。《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[3]。《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”、“V + 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：由该信息结构模式传达的真正信息可产生的问题

Answer：由该信息结构模式传达的真正信息可构造的答案的句法分布式]+

例子：符合该信息结构模式的真实语料的实例

信息结构模式的一个具体实例如下：

SYN_S=N --> N

SEM_S=(人,专/专/姓) --> [修饰] (人,职位)

Query1: 谁？ / 哪一位？ / 什么人？

Answer1: N1 + N2

Query2: N1 是做（干）什么（工作）的？ / N1 的职务？

Answer2: N2

例子：江泽民-主席，克林顿-总统，林德政-教授，黄光-法官，李鹏-委员长，……

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。

4 语义规则

4.1 语义规则的表述

引入语义规则的主要目的在于描述词语组合的语义条件限制，消除句法结构和语义关系上的歧义。因此，语义规则的表述形式应当能够准确而且简洁地描述具有什么语义的词语可以以怎样的方式组合，以及组合后各成分之间的语义关系如何。由于我们是以《知网》知识词典和《结构库》作为语义知识资源，因而我们在描述语义规则中的语义时采用了类似《知网》义项概念定义的方式。此外，为了方便实现，语义规则用表型结构书写。

我们设计的语义规则表述方式可形式化描述为：

Rule = (SyntaxRule (Relation) Govern Rule2 Rule2)

Rule2 = Rule | (DEFType)

DEFType = [+][Mark]ATOM[,[+][Mark]ATOM]*

SyntaxRule = SRnn1 | SRvn1 | ……

Relation = Rel | EventRole | Feature

ATOM = Atom | Var[AtomSet]

AtomSet = (Atom[,Atom]*)

Govern = F | B

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

Rel = 合成 | 修饰 | ……

EventRole = 施事 | 受事 | ……

Feature = 昔 | 暂 | ……

Atom = $atom_1$|$atom_2$|……

Var = var_1|var_2|……

具体说明如下：

(1) SyntaxRule 标明了语法规则类型编码，它限定了可以捆绑该语义规则的语法规则。例如：语义规则(SRnn2 (修饰) F (Human 人, +ProperName 专) (Human 人, +occupation 职位))中的“SRnn2”表明它可作为任意形如“Ns ==> Ns Ns”的语法规则的捆绑语义规则。

(2) Relation 标明了成分之间的语义关系，主要使用的是来源于《知网》的动态角色（在上面的形式描述中标记为 EventRole）和属性（在上面的形式描述中标记为 Feature）标记，此外还有《结构库》中新增的（在上面的形式描述中标记为 Rel）诸如“合成”、“修饰”等。

(3) Govern 标明了成分之间的管辖关系，F 表示前一个成分是“管辖者”，B 表示后一

个成分是“管辖者”。

(4) Mark 可以是《知网》中定义的任意语义关系符。

(5) Atom 可以是《知网》中定义的任意义原、动态角色和属性，甚至可以是《知网》中的“Secondary Feature”之类的概念类型（下面简单地将它们统称为义原）。

(6) DEFType 指定了成分义项中必须出现的义原及其前缀语义关系符，从而限定了匹配该规则的各成分的语义类型(下面称为义项类型)。为了使语义规则简洁，我们约定：①DEFType 中各义原之间是“与”关系，也就是说这些义原应当同时出现；②如果 DEFType 中的义原前面有“+”，则只有完全相同的义原才能与之匹配，除此之外的任何义原与它的匹配度均为 0；③如果 DEFType 中的义原前面没有“+”，则表明给出的是最上位义原，也就说分类树上作为该义原子孙结点的所有义原均与之完全匹配，而同一分类树上的其它义原根据语义距离计算匹配度，不同分类树上的义原与之完全不匹配，即匹配度为 0，例如：义项类型中出现了“plant 植物”这个义原，那么“crop 庄稼”、“tree 树”等“plant 植物”的下位义原都与之完全匹配，而“planting 栽植”、“remove 消除”等事件类义原与之匹配度均为 0。

(7) 义项类型中可用变量（在上面的形式描述中标记为 Var）代替具体义原，用于描述规则中各子成分在语义上的联系。变量后面还可以用义原集合（在上面的形式描述中标记为 AtomSet）来限制该变量的取值范围，此时的变量称为受限变量。例如：语义规则(SRnn2 (材料) B (+material 材料，?var_m) (var_m (artifact 人工物)))中的 var_m 就是一个受限变量，它的取值范围是义原“artifact 人工物”及其所有下位义原，而它在前后两个成分的义项类型中的同时出现表明后一成分应为某种人工物，而前一成分应为某种可用于制造该人工物的材料。

4.2 语义规则的获取

语义规则在我们设计的短语结构排歧方法中起着核心作用，无论是分析过程中对歧义结构的实时消解，还是最后优选阶段对候选结构的评估，都是以语义规则为主要依据。因此，语义规则的好坏直接影响该方法的实际排歧效果。

我们获取语义规则的大致过程是：首先以《结构库》作为语料库和语义规则的知识来源，由人工对《结构库》中的信息结构模式进行筛选转换，得到一个初始语义规则库；然后用这个初始规则库对《结构库》中提供的实例进行自动分析并统计正确率，人工校验之后抽取出正确率低、粒度过大的规则和相关的无法正确分析的实例，通过实例学习对这些规则进行修改、补充或者细化，并将它们与初始规则库合并形成新的规则库；然后再用新规则库自动分析统计《结构库》中的实例和人工从真实文本中另外筛选的实例，再进行人工校验和规则修正扩充，如此循环，直到得到满意的结果为止。

由于篇幅所限，这方面的具体内容待另文叙述。

5 基于语义规则的计算

5.1 语义规则匹配度计算

5.1.1 义原间语义匹配度计算

《知网》中的分类树可用于计算义原间的语义距离和语义匹配度，文献[4]中提出了基于义原在分类树上对应结点间的最短路径长度的计算方法。然而，由于我们在语义规则的表述方式中约定如果规则中给出的是最上位义原，那么分类树上作为最上位义原子孙结点的所有义原均与之完全匹配，因而此种情况下的语义距离应为 0。此外，计算中还应体现出不同类型的义原（如事物型与事件型）之间语义的不可比性。根据上述分析，适应语义规则匹配度计算的需要的义原间语义距离和语义匹配度的定义如下：

设 A 是成分义项的概念定义中出现的义原，B 是语义规则中出现的义原，则 A 与 B 的语义距离为：

$$DISTANCE\_ATOM(A,B)=\begin{cases}tree\_depth(B)\times 2 & \text{当 AB 属于不同分类树或者 B 前有“+”且 A}\neq\text{B 时}\\ 0 & \text{当 B 前无“+”且 B 是 A 的祖先或者 A=B 时}\\ A\text{与}B\text{在分类树上的最短路径长度} & \text{其它}\end{cases} \quad (1)$$

A 与 B 的语义匹配度为：

$$SIM\_ATOM(A,B)=(1-DISTANCE\_ATOM(A,B)/(tree\_depth(B)\times 2))\times 100 \quad (2)$$

其中，tree_depth（B）= B 所在的分类树的树高。

5.1.2 义项与义项类型的匹配度计算

设义项的概念定义（或语义规则中的义项类型）L 中出现的所有语义关系符组成的集合记作 Relations(L)，L 中所有语义关系符为 k 的义原组成的集合记作 Relation_Atoms(L,k)。由 Relation_Atoms(L,k)中的 n 个义原 a_1，a_2，……，a_n 构成一个排列记作（a_{p_1}，a_{p_2}，……，a_{p_n}）（不足 n 个的用 ε 代替，ε 与任意义原的匹配度均为 0),其中 $a_{p_1},a_{p_2},\cdots\cdots,a_{p_n}\in\{a_1,a_2,\cdots\cdots,a_n\}$且除 ε 外各不相同，所有可能的排列组成的集合记作 P_ Relation_Atoms(L,k,n)。

义项的概念定义 I 与语义规则中的义项类型 J 的匹配度为：

$$SIM\_ITEM(I,J)=\frac{1}{m}\times\sum_{k\in Relations(J)}\ \max_{(a_1a_2\cdots a_n)\in P\_Relation\_Atoms(I,k,n)}(SIM\_ATOM(a_1,b_{k_1})+\cdots+SIM\_ATOM(a_n,b_{k_n})) \quad (3)$$

其中，m =(J 中的义原总数)，n = | Relation_Atoms(J,k) | ，$b_{k_1},b_{k_2},\cdots\cdots,b_{k_n}\in$Relation_Atoms(J,k)。

5.1.3 语义规则匹配度计算

设语义规则 R 中的所有义项类型按出现的先后顺序分别记为 J_1，J_2，……，J_k，待归约成分序列 P 中的各成分按先后顺序分别记为 C_1，C_2，……，C_k，成分 C 的所有可能的义项组成的集合记为 Items(C)，则待归约成分序列 P 与语义规则 R 的匹配度为：

$$SIM\_RULE(P,R)=\frac{1}{k}\times\sum_{q=1}^{k}\max_{I\in Items(C_q)}SIM\_ITEM(I,J_q) \quad (4)$$

5.2 候选结构的评估

5.2.1 语义规则实际使用效果的评估

由于我们在分析过程中设置了一个阀值用于判定待归约成分序列是否匹配语义规则，因而最终真正用于归约的语义规则的匹配度都是大于这个阀值的。为了更明显地反映语义规则

的实际使用效果，我们用下面这个计算公式来评估语义规则 R 归约成分序列 P 的实际效果：

$$VALUE\_RULE(P,R) = (SIM\_RULE(P,R) - \theta)/(100 - \theta) \times 100 \quad (5)$$

其中，θ 为阀值。

5.2.2 候选结构的评估

设要产生候选结构 S 需要使用语义规则 $R_1, R_2, \cdots\cdots, R_n$ 先后归约成分序列 $P_1, P_2, \cdots\cdots, P_n$，则候选结构 S 的评估值为：

$$VALUE\_STRUCTURE(S) = \frac{1}{n} \times \sum_{i=1}^{n} VALUE\_RULE(P_i, R_i) \quad (6)$$

6 基于语义规则的短语结构排歧实例

这里我们用一个实例说明整个排歧流程：

输入短语：清华 大学 校长

（1）句法分析将首先遇到下面这两个可规约的成分序列：

P1：清华/n 大学/n

P2：大学/n 校长/n

从《知网》中获取词语的义项：

词语	词性	义项的概念定义
清华	N	InstitutePlace|场所,@teach|教,@study|学,education|教育, ProperName|专,(China|中国)
大学	N	InstitutePlace|场所,@teach|教,@study|学,education|教育
校长	N	human|人,official|官,education|教育
	N	human|人,#occupation|职位,official|官,education|教育

计算成分序列 P1、P2 与语法规则“Ns ==> Ns Ns”的所有捆绑语义规则的匹配度，下表列出了与成分序列 P1、P2 匹配度较大的三条语义规则及其匹配度：

成分序列	语义规则	匹配度
P1	R1：(SRnn2 (限定) B (+ProperName 专) (InstitutePlace 场所))	100
	R2：(SRnn2 (限定) B (+ProperName 专) (building 建筑物))	90
	R3：(SRnn2 (领属物) B (place 地方 +ProperName 专) (organization 组织))	86
P2	R4：(SRnn2 (来源整体) B (InstitutePlace 场所) (Human 人 +#occupation 职位)	100
	R5：(SRnn2 (来源整体) B (organization 组织) (Human 人 +#occupation 职位))	96
	R6：(SRnn2 (限定) B (place 地方 +ProperName 专) (Human 人 +#occupation 职位))	68

根据匹配度，从中选择 R1 用于归约 P1、R4 用于归约 P2。 从而也就确定了在后一种分析结构中“校长”这个词语的义项是“human|人,#occupation|职位,official|官,education|教育”。

（2）成分序列 P1、P2 归约后将分别得到下面这两个可归约的成分序列：

P3：(清华/n 大学/n)/np 校长/n

P4：清华/n (大学/n 校长/n)/np

与（1）中类似的，分别计算成分序列 P3、P4 与语法规则“Ns ==> Ns Ns”的所有捆绑语义规则的匹配度。由此，可以找到语义规则 R4 与 P3 具有最大匹配度，从而得到候选结构：

S1： ((清华/n 大学/n)/np 校长/n)

而 P4 却找不到任何与之匹配度大于阀值的语义规则，因此 P4 无法规约。

（3）由于分析过程在语义规则的制约下只产生了唯一的候选结构 S1，因此 S1 就是最终分析结果。

7 实验结果及分析

我们在原有的基于语法的汉英机器翻译系统上构建了一个模型系统，针对名词短语、动词短语和形容词短语分析需要归纳了两百多条语义规则。从《结构库》和真实语料库中分别抽取 1000 个实例进行封闭测试和开放测试，测试指标和测试结果如下表：

测试指标	指标描述	封闭测试	开放测试
结构排歧的正确率	结构分析正确的短语数/测试短语数	0.90	0.84
语义排歧的正确率	语义选择正确的词语数/测试短语中词语总数	0.83	0.79

语义规则库中的有些规则粒度过大，导致规则间的区分不够明显，是排歧错误或排歧失败的主要原因。

8 结束语

本文提出了一种基于语义规则的汉语短语结构分析排歧方法，该方法在原有的基于语法规则的分析系统基础上，以规则的形式导入必要的语义知识，并根据语义规则匹配度来剔除不合语义的分析结果，在排除结构歧义的同时也消除了短语中各词语的语义歧义。实验结果表明，该方法对于名词短语、动词短语及形容词短语等的排歧效果是较令人满意的。

现有的模型系统已基本实现预期的设想，然而要真正实用化还有大量细致的工作要做，尤其是在语义规则库的构造上。一方面，现有的许多大粒度的语义规则需要进一步的划分与调整；另一方面，为了处理其他类型的短语还需要扩充语义规则。

参考文献

[1] 冯志伟，论歧义结构的潜在性，中文信息学报，1995 第 4 期

[2] 董振东，董强，知网，http://www.how-net.com

[3] 董振东，董强，关于知网-中文信息结构库，http://www.how-net.com

[4] 杨晓峰，李堂秋，洪青阳，基于实例的汉语句法结构分析歧义消解，中文信息学报，2001，15（3）

[5] 詹卫东，常宝宝，俞士文，汉语短语结构定界歧义类型分析及分布统计，中文信息学报，1999 第 3 期

[6] 苑春法，黄锦辉，李文捷，基于语义知识的汉语句法结构排歧，中文信息学报，1999，13（1）