汉语的二元语义关系模型

万建成
（山东工业大学计算机系，山东，济南 250014）

摘　要：语义的表达和处理，一直是自然语言处理的重要部分。虽然至今已经提出了多种语法理论，但由于它们要求建立一整套复杂的属性，和搜集大量的词的复杂使用环境描述，实现上的难度很大。本文提出的二元语义模型，概念和实现简单，可以在语言识别的过程中逐步学习和归纳语义知识，并用于歧义的识别。该方法，已经在作者研究开发的汉语拼音输入的同音词识别系统中实现并获得应用。

A Model of Chinese Binary Semantics

Wan Jiancheng
Computer Science DepartmentJinan, Shandong 250014　　Shandong University of Technology

ABSTRACT: Semantics Analysis plays an important role in Natrual Language Processin-gresearches. Many semantic models have been proposed, but the implementationof them is relativly very difficult, because of the complexity of thesemodels in semantic descriptions and enviroment. In this paper, a model ofChinese Binary Semantics is proposed. With its simplicity in concept andeasiness in implementation, semantic knowledge can be increamentaly learnedand inductively constructed. The model has been experimentally implementedin a system to recognise Chinese homophones.

1. 引言

语义的表达和处理，一直是计算机和自然语言处理的重要部分。在计算机语言的研究和实现中，普遍采用的是属性文法定义方法。在自然语言的研究中，有汉语组合语法[2]；有句法识别伴随词搭配关系的标准理论的转换语法；有建立在特别语义表达结构之上的格语法，概念从属理论，和非转换语法的词汇功能语法，功能合一语法，和广义短语结构语法。这些理论要求建立一整套复杂的属性，和搜集大量的词的复杂使用环境描述，实现上的难度很大。

对"语义"一词的定义和理解，是一个复杂而困难的问题，从不同的角度和目的有不同的认识。本文提出的二元语义模型，不去研究语义的深层含义，依据汉语句法分析的二分法思想[1]和汉语短语的二元关系结构（主谓，动宾，动补，动词偏正，名词偏正…），从解决汉语识别歧义的实用出发，建立了概念简单，可以在语言识别的过程中逐步学习和归纳语义知识，

并用于汉语歧义识别的方法。

本文提出的方法，已经在作者研究开发的汉语拼音输入的同音词识别系统中实现并获得应用[3,4]。

2. 基本二元语义关系

2.1 基本元素定义

定义汉词集合　　$W=\{W_1,W_2,\cdots,W_m\},m>0$

定义词属性集合　　$A=\{A_1,A_2,\cdots,A_n\},n>0$

定义词 C 的属性集合词 $C\in W$ 所具有的属性 $A_i,i>0$ 的集合，表示为 A(C)。

2.2 基本关系定义

定义基本词属性关系集合 $R=\{R_1,R_2,\cdots,R_p\},p>0$

其中 $R_i=(x,y),x\in A,y\in A$。

定义‘·’为词或属性的关联运算，例如 a·b。

关联运算是不可交换的，即 $a\cdot b\neq b\cdot a$，同时有 a 和 $b\in W$，或同时有 a 和 $b\in A$。

定义基本词关联，如果 $r\in W,s\in W$ 有关联关系，则写成 $C=r\cdot s$。

定义基本词属性关联，词属性关系 $R_i=(x,y),x\in A,y\in A$，表示成 $R_i=x\cdot y$。

2.3 词二元语义关系

定义词二元语义关系 $S=\{C,B\},C=r\cdot s$ 是基本词关联，$r\in W,s\in W;B=x\cdot y$ 基本词属性关联，$x\in A,y\in A$。

2.4 词二元语义关系的满足

对于词二元语义关系 $S=\{C,B\},C=r\cdot s$ 是基本词关联，$r\in W,s\in W;B=x\cdot y$ 基本词属性关联，$x\in A,y\in A$。如果 $B\in R$，并且 $x\in A(r),y\in A(s)$，A(r)和 A(s)分别是词 r 和 s 的属性集合，则，称词二元语义关系 $S=\{C,B\}$得到了满足。

为了方便，我们只讨论可以得到满足的二元语义关系。所以，以后提到的只是那些可以得到满足的二元语义关系。

2.5 基本属性关系转换

定义基本属性关系转换是转换函数 $F:R\rightarrow A$，它是基本属性关系到词属性集合上的转换。

在汉语中，将根据短语可以充当的句法角色而定义该转换的种类集合。

3. 扩充二元语义关系

3.1 扩展关系定义

定义扩展词属性关系集合 $Q=\{Q_1,Q_2,\cdots,Q_p\},p>0$，其中 $Q_i=(x,y),x\in A\cup R\cup Q$，$y\in A\cup R\cup Q$。

注意，这是一个递归的定义。

定义扩展‘·’的词或属性的关联运算，例如 a·b。同时有 $a,b\in A\cup R\cup Q$，或同时有 a，

b∈{W∪基本和扩展词关联}。

'·'关联运算是不可交换的，即 a·b≠b·a。

'·'关联运算是不可结合的，即 a·(b·c)≠(a·b)·c。

定义扩展词关联，如果 r,s∈{W∪基本和扩展词关联}有关联关系，则写成 C=r·s。

定义扩展词属性关联，词属性关系 $R_i=(x,y)$，$x,y\in A\cup R\cup Q$，表示成 $R_i=x\cdot y$。

3.2 扩展词二元语义关系

定义扩展词二元语义关系 S={C,B}，C=r·s 是扩展词关联，其中 r,s∈{W∪基本和扩展词关联}；B=x·y 扩展词属性关联，x,y∈A∪R∪Q。

3.4 扩展关联关系的同构

对于两个扩展关联关系 $X=X_1\cdot X_2$ 和 $Y=Y_1\cdot Y_2$，X,Y∈{扩展词关联∪扩展词属性关联}，如果满足下列条件，则称是关系同构的：

① X_1，X_2，Y_1，和 Y_2 都是基本元素(词或属性)，或者

② X_1 和 Y_1 是基本元素，X_2 和 Y_2 同构，或者

③ X_2 和 Y_2 是基本元素，X_1 和 Y_1 同构。

3.5 扩展词二元语义关系的满足

对于扩展词二元语义关系 S={C,B}，C=r·s 是扩展词关联，r,s∈{W∪基本和扩展词关联}；B=x·y 扩展词属性关联，x,y∈A∪R∪Q。如果 C 和 B 同构，并且，对于任何 C 中的词 p、q 是 B 中与 p 结构上相对应的属性，都有 q∈A(p)，A(p)是词 p 的属性集合，则，称扩展词二元语义关系 S={C,B}得到了满足。

由于不同构的二元语义关系在实用中没有意义，所以在以后的讨论中，都假设在被研究的二元语义关系 S={C,B}中，总有 C 和 B 同构。

为了方便，我们只讨论可以得到满足的扩充二元语义关系。所以，以后提到的只是那些可以得到满足的扩充二元语义关系。

3.6 扩展属性关系转换

定义基本属性关系转换是转换函数 F:Q→A，它是扩展属性关系到词属性集合上的转换。在计算 F:Q→A 时，需要多次用到 F:R→A 的转换，其方法是：如果存在一个 R→A 的转换，就用 A 去替换 R。经过一步步替换而最终缩减到 A。

注意，扩展属性关系转换是定义在基本属性关系转换上的。

4. 汉语属性文法

下面比较一下在计算机语言和汉语自然语言中，词属性处理上的不同。从中可以看出，本文提出的二元语义关系模型，是一种类属性文法。

下表给出了两种处理中，有关"字符集，词，属性，终结符"概念上的不同：

	字符集	词	属性	终结符
计算机语言	a..z,A..Z	标识符	类型	标识符
自然汉语	汉字	汉语词	属性(名词，动词…)	属性

在计算机语言中，标识符（包括关键字）的数目相对很小，在句法识别正确后，再计算属性关系，这就是“标识符→类型”关系的计算。

在汉语自然语言处理中，词的数目极大，属性的数目相对很小，因此语法只能用词的属性而不是词来描述，就形成了“属性→词”关系的计算，即，在对属性的句法结构识别成功后，再计算词之间的关系。

所以，本文提出的二元语义关系模型，从计算的次序看，是一种类属性文法。

5. 语义知识的二元关系模式

5.1 关联关系的重心

对于‘·’关联关系 $X=a\cdot b$，如果 a 是 a 和 b 组合出现时的中心词，则表示为 $X=Z\cdot b$，并称 Z 是此关联关系的重心。类似地，如果 b 是 a 和 b 组合出现时的中心词，则表示为 $X=a\cdot Z$，并称 Z 是此关联关系的重心。

5.2 语义知识模式

（扩展）词二元语义关系是一个语义知识模式。统一简称为语义模式。

5.3 两语义模式的等价

对于语义模式 $S_1=\{C_1,B_1\}$ 和 $S_2=\{C_2,B_2\}$，如果有 $C_1=C_2$，$B_1=B_2$，则称 $S_1=S_2$，即两语义模式等价。

当 $B_1\in A$，$B_2\in A$ 时，称为基本属性项等价；

当 $B_1\in R$，$B_2\in R$ 时，称为基本属性关系项等价；

当 $B_1\in Q$，$B_2\in Q$ 时，称为扩展属性关系项等价。

5.4 两语义模式的包含

对于语义模式 $S_1=\{C_1,B_1\}$ 和 $S_2=\{C_2,B_2\}$，如果满足下列条件，则称语义模式 S_1 包含于语义模式 S_2，记作 $S_1<=S_2$：

① $S_1=S_2$，即 S_1 与 S_2 等价，或

② 当 $B_1\in R$，$B_2\in A$ 时，

(1) $B_1=b\cdot Z$，Z 是 B_1 的重心，且 $Z=B_2$；或

(2) $B_1=Z\cdot b$，Z 是 B_1 的重心，且 $Z=B_2$；或

(3) $B_1=a\cdot b$，存在一个基本属性关系转换 $F:B_1\to B$，使得 $B=A$。

③ 当 $B_1\in Q$，$B_2\in R\cup Q$，且 $|B_1|>=|B_2|$（$|X|$ 为 X 中属性项的个数）时，存在一个属性关系转换 $F:B_1\to A_1$，和一个属性关系转换 $F:B_2\to A_2$，使得 $A_1=A_2$。

6. 应用

本节中将给出两个二元语义关系的应用例。

例 1：

S=｛脱·衣服，动词·名词｝定义了一种<动词，名词>关联的词对<脱，衣服>。它表

达的不仅仅是词“脱”和“衣服”的直接关联关系，它是“脱”和“衣服”在一切具有动宾关系组合时的一种体现。它同时说明，“脱”和“衣服”是一种合乎汉语“语义”，有“意义”的动宾关联用法。因此，可以说：“他脱掉了衣服”，“他脱下了一件衣服”，“脱不下衣服来”，等等。

例 2：

S＝{(吃·(人家·的))·(嘴·短),(动词·(名词·的)·(名词·形容词)}。它表达“吃人家的”就造成“嘴短”这一种行为。这里给出的是一种语义表达结构，在汉语句法许可的情况下，可以在使用中附加任何必要的成分，来点缀修饰，但其基本“语义”或“语义”关系，将始终得到保证。例如，可以说：“吃了人家的嘴短”，“吃了人家的嘴就短”，“吃了人家的嘴就短了”，“吃过了人家的嘴就特别短了”，等等。

7. 总结

本文提出的汉语的二元语义关系模型，具有以下的特点：

①建立在简单的二元关系之上，概念简单、直观，与传统的汉语句法结构直接对应。

②二元关系的复合，可以构成任意复杂的句法和语义约束关系，只要扩展基本的词属性关系就可以了，因而简单的二元关系结构具有很强的表达能力。这种表达能力与二叉树可以表达任意复杂的树结构，进而可以表达任意复杂的图结构的能力，是对应的。

③易于建立汉语句法和语义知识的学习。作者设想，汉语知识的学习模型可以用该二元关系表达。学习的出发点是名词，动词，形容词等基本词属性，简单的句法关系，属性转换关系。语义模式可以通过学习和归纳逐步建立起来。

④二元语义关系可以根据语言的实践逐步建立，而后应用。

⑤语义的包含变为模式的形式包含，此种关系类似于合一算法。

⑥特别适合于处理象汉语这样的无词形变化的粘着语的处理。

参 考 文 献

[1] 吴竟存，侯学超，现代汉语句法分析，北京大学出版社，1986 年 12 月，p. 3

[2] 翟成祥，汉语组合语法，《中文信息学报》，Vol. 6，No. 1，1992

[3] 万建成，汉语同音词的上下文相关识别，COLIPS(新加坡)，Vol. 2，No. 1，1992

[4] 万建成，FPY 中同音词智能识别方法，《中文信息学报》，Vol. 7，No. 2，1993

1995 年 8 月于济南