数据库语义学与自然语言交流

冯秋香[1] Roland Hausser[2] 汪榕培[1]
(大连理工大学计算机科学与技术学院 大连 116024)[1]
(德国爱尔兰根-纽伦堡大学计算语言学系 爱尔兰根 91054)[2]

摘 要 数据库语义学(Database Semantics)是德国爱尔兰根-纽伦堡大学 Hausser 教授于 20 世纪 90 年代提出的面向计算机自然语言处理的程序化语义学理论。这一理论不同于以往任何以元语言为基础的语义分析方法,它将自然语言的理解与生成建构为角色(即听者和说者)转换的规则理论体系。该理论建构自然语言交流模型的两个核心基础是左结合语法(LA)和词库数据结构(word bank)。以古汉语为例介绍和分析这一理论,通过顺次应用 LA 语法的 3 个变体:LA-hear、LA-think 和 LA-speak 说明 DBS 理解和生成自然语言的过程。

关键词 自然语言,句法语义分析,左结合语法,数据库
中图法分类号 TP391 **文献标识码** A

Database Semantics and Natural Language Communication

FENG Qiu-xiang[1] HAUSSER Roland[2] WANG Rong-pei[1]
(Department of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)[1]
(Department of Computer Linguistics, University of Erlangen-Nurnberg, Erlangen 91054, Germany)[2]

Abstract Database Semantics(DBS), initiated by Prof. Hausser in 1990's as a linguistic framework for computational modeling of natural language communication, treats the agent's switching back and forth between the speaker and the hearer mode as a well-defined and well-motivated computational problem. As a procedural approach aiming at a systematic general model of cognition, differing from those metalanguage-based theories, DBS has two innovative basis, namely, the time-linear motor algorithm of LA-grammar(Left Associative grammar) and the data structure of word bank. With its application to the ancient Chinese as an example, we attempted to provide an introduction to the DBS understanding and production of natural language in the sense of the functioning of three LA-grammar variables, i. e., LA-hear, LA-think and LA-speak.

Keywords Natural language, Syntactic and semantic analysis, LA-grammar, Database

1 引言

语义分析的理论基础很多,如以真值为条件的模型理论[1]、以应用为条件的言语行为理论[2]、起源于心理学的语义网络[3]、系统功能理论框架下的修辞结构理论[4]及方法多样化的语篇语言学[5]等。

上世纪 90 年代,德国爱尔兰根-纽伦堡大学计算语言学系的 Roland Hausser 教授提出以认知为基础的面向计算机和人工智能领域的语言学框架——数据库语义学(Database Semantics, DBS)[6]。这一程序化的语义学理论不同于以往任何以元语言为基础的语义分析方法。它把自然语言的理解(听)与生成(说)建构为角色(即听者和说者)转换的规则理论体系。该理论建构自然语言交流模型的两个基础是左结合语法(Left Associative Grammar, LA 语法)和词库(word bank)数据结构。

提出数据库语义学的目的是建立一个功能完备、数据完全、计算复杂度小而计算机操作性强的关于自然语言交流的理论[7]。Hausser 认为自然语言交流过程的一个基本特点就是角色转换[8]。这一过程可以从内外两个方面来看,从外部看是两个主体在轮流说话,一个是说者的同时另一个就是听者;从单个主体内部看,交流主体本身也在经历时而作为听者、时而又成为说者的角色变化。因此,数据库语义学首先建立了基本的角色转换模型,即听者状态与说者状态之间的交流和转换,如图 1 所示。

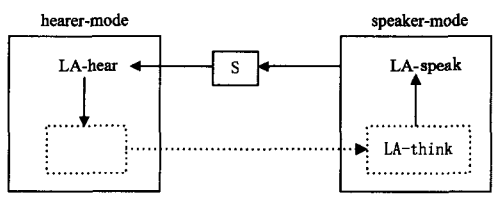

图 1

hearer-mode 即听者状态,speaker-mode 即说者状态。所

到稿日期:2010-12-16 返修日期:2011-03-21 本文受国家自然科学基金项目(60973069)资助。
冯秋香(1977—),女,博士生,主要研究方向为机器翻译理论,E-mail:feng.qiuxiang@gmail.com;Roland Hausser (1946—),男,博士,教授,主要研究方向为计算语言学;汪榕培(1942—),男,教授,博士生导师,主要研究方向为翻译理论、典籍英译。

有的句法分析和语义分析都必须在这一角色转换模型下进行。图中方框里的 s 代表被传递的自然语言符号。LA-hear、LA-think 和 LA-speak 是在角色转换模型内互动的 LA 语法的 3 个变体。LA-hear 接收并理解自然语言符号，将其内容存入听者数据库。LA-think 在说者数据库中自动搜索可供表达的内容。LA-speak 把 LA-think 提供的内容转化为自然语言输出。

数据库语义学通过在数据库中分别存储说者和听者的认知内容来模拟二者之间信息传递的过程。如果说者一方的数据库中存储的某个内容由说者编码为自然语言符号传递给听者，听者能够将其解码为类似的内容重新存储在听者一方的数据库中，则视为自然语言交流顺利完成。听者和说者之间传递信息的内容以命题因子(proplet)的形式存储在词库中。

2 LA 语法

目前应用较广的句法理论包括范畴语法[9]、依存语法[10]、短语结构语法[11]等。LA 语法与这些语法框架的最根本区别在于 LA 语法遵循的是语言的时间线性顺序[12]。

时间线性即按照时间的先后顺序。以输入“abcde…”为例，LA 语法按照输入的先后顺序首先将当前已经分析过的部分(称为句首，以 ss 表示)与当前新词(即下一个输入的词，以 nw 表示)结合起来，构成新的句首(ss′)，即 a 和 b 组合为(ab)。进而(ab)和 c 组合为((ab)c)、((ab)c)和 d 组合为(((ab)c)d)等。这种方法称为左结合(left-associative)。由此也可以看出，LA 语法关注的是词与词之间的接续可能，而不是像 PSG 等语法那样关注词与词之间的替代性。自然语言的结构在本质上是时间线性的。而且，不单是自然语言符号结构上的顺序，自然语言理解(听者状态)和自然语言生成(说者状态)都以时间为顺序。因此，数据库语义学选择 LA 语法作为算法，用以主体为导向、以说者与听者之间的信息交流为内容的新的自然语言分析方法替换之前的以符号为导向的孤立分析句子的方法，其目的是开发人造的认知主体——一个会说话的、可以与研究人员和用户交流的机器人。

LA 语法以未经分析的语言为输入，以经过语法分析的语言为输出，其包括 4 个主要组成部分：一部词典、一组起始状态、一套规则和一组结束状态。以形式语言 akbkck(包含 abc，aabbcc，aaabbbccc 等表达式)为例，LA 语法的定义如下[13]：

$LX=_{def}\{[a(A)],[b(B)],[c(C)]\}$

$ST_s=_{def}\{(A)\{r_1,r_2\}]\}$

$r_1:(X)(A)\Rightarrow(AX)\{r_1,r_2\}$

$r_2:(AX)(B)\Rightarrow(XB)\{r_2,r_3\}$

$r_3:(BX)(C)\Rightarrow(X)\{r_3\}$

$ST_F=_{def}\{[\varepsilon rp_3]\}$

词典 LX 的定义是一个已知词的列表。词的已知信息包括词表和词的某个属性，如词的句法范畴、词义等。如 $[a(A)]$，其中 a 为词表，A 为该词表的句法范畴。起始状态 ST_S 是一个句法范畴和规则包构成的有序对，指明从哪个词(由其句法范畴代表，如 A)开始进行句法分析，以及通过哪些具体规则来组合第一个词和下一个词。如上的 $[(A)\{r_1,r_2\}]$，A 为句法范畴，$\{r_1,r_2\}$ 为规则包。r_1，r_2 和 r_3 分别为 3 条不同的 LA 语法规则，其基本模式是 $r_i:ss\ \mathrm{nw}=\!>ss'\ rp_i$，包括规则名 r_i、起始模式 ss(sentence start，句首)、下一个词模式 nw(next word，下一个词)，输出模式，称作 ss′(result sentence start，结果句首，或者新的句首)以及规则包 rp_i。上例中的结束状态组 ST_F 包含一个元素 $[\varepsilon\ rp_3]$，表示如果经过分析的表达式的最终句法范畴是 ε，最后一个被激活的规则包是 $r-3$，那么这个表达式是符合形式语言 $a^kb^kc^k$ 的完整的表达式。

LA 语法有 3 个变体，这 3 个变体又恰恰体现了 DBS 最基本的角色转换的思想：(1)在输入过程中实现自然语言理解的称作 LA-hear；(2)在思考过程中实现推理和对认知内容进行排序的称作 LA-think；(3)在输出过程中实现语言生成的称作 LA-speak。实际上，Hausser 教授的 LA 语法雏形出现于上世纪 80 年代[14]，之后经历了一些发展变化。上述的基于列表的早期 LA 标记方法后来也被更为灵活的以特征(属性/值对)结构为标记的方法取而代之。于是，上述 LA 语法对于形式语言 $a^kb^kc^k$ 的定义转变如下[8]：

$$LX=_{def}\left\{\begin{bmatrix}sur:\mathrm{a}\\cat:\mathrm{A}\end{bmatrix},\begin{bmatrix}sur:\mathrm{b}\\cat:\mathrm{B}\end{bmatrix},\begin{bmatrix}sur:\mathrm{c}\\cat:\mathrm{C}\end{bmatrix}\right\}$$

$ST_S=_{def}\{[[\mathrm{cat}:A]\{r_1,r_2\}]\}$

$$r_1:[\mathrm{cat}:X][\mathrm{cat}:A]\ \begin{matrix}\text{leftcopy } A \text{ ss-cat}\\ \text{copy\_ss}\end{matrix}\ \{r_1,r_2\}$$

$$r_2:[\mathrm{cat}:AX][\mathrm{cat}:B]\ \begin{matrix}\text{leftcopy } A \text{ ss-cat}\\ \text{rightcopy } B \text{ ss-cat}\\ \text{copy\_ss}\end{matrix}\ \{r_2,r_3\}$$

$$r_3:[\mathrm{cat}:BX][\mathrm{cat}:B]\ \begin{matrix}\text{leftcopy } B \text{ ss-cat}\\ \text{copy\_ss}\end{matrix}\ \{r_3\}$$

$ST_F=_{def}\{[[\mathrm{cat}:\varepsilon]rp_3]\}$

这种方法和原来的列表表示法是对等的。原定义下的 ss 表示为列表(X)，新定义下转变为属性/值对 $[\mathrm{cat}:X]$(cat 表示一个属性，X 是值)。nw 也有类似的变化。原定义下结果模式(aX)代表的句法范畴上的操作在新的定义下得到明确的描述。如“leftcopy A ss-cat”，意思是把常数 A 加在 ss 的属性[cat]当前值的左边。

特征结构包含以常量或者变量为值的各种属性。实践证明和列表的方法相比有更强的描述性。在数据库语义学里，这种特征结构称作命题因子(proplet，取自 proposition droplet，意为构成命题的因子)。说者与听者之间的交流内容以命题因子的形式存储在词库里。

3 词库数据结构

说者与听者之间交流的内容，即数据，以命题因子的形式存储在网络数据库中。同一命题各因子的[prn]属性值(即命题编号)相同，不同命题因子的[prn]属性值不同。命题因子各个属性的值体现同一命题内部各词之间的关系(命题内关系)，如函词-论元(functor-argument)关系和并列(coordination)关系以及命题与命题之间的各种关系(命题间关系)等。这种数据结构称作词库(word bank)，它允许(1)按时间线性顺序读入和存储命题；(2)按时间线性顺序搜索和激活命题；(3)以概念为关键词访问和检索命题。

数据库语义学下的命题因子属性主要有 13 个，其中

[noun]、[verb]和[adj]是核心属性；[cat]和[sem]是语法属性；[fnc]、[arg]、[mdd]和[mdr]是命题内接续属性；[nc]、[pc]和[idy]是命题间接续属性，在表示命题间关系时是必要的，表示命题内关系时是可选属性；[prn]是簿记属性。但是，并不是所有命题因子都具备所有这些属性。如名词性命题因子一般不会成为谓词，不会带论元，因此也就不会有属性[arg]。

词库与 LA 语法之间是互动的。Hausser 教授把词库比喻成一个铁路系统，把 LA 语法比喻成一辆在这个铁路系统上运行的机车。机车运行过程中路过各个点，即激活相应的命题内容。也就是说，词库是语言生成的概念化基础，解决说什么的问题；而 LA 语法是语言生成的序列化和词汇化基础，解决怎么说的问题。

语境独立于语言而存在，没有语言功能的主体的认知结构只有一层，而有语言功能的主体的认知结构分为语境和语言两层。词库的数据结构必须同时反映语境层的非语言认知内容和语言层的认知内容。这两个层面之间的互动必须满足听者和说者语境层和语言层之间的互相映射。有语言功能的主体也就是人的认知结构，如图 2 所示[15]。

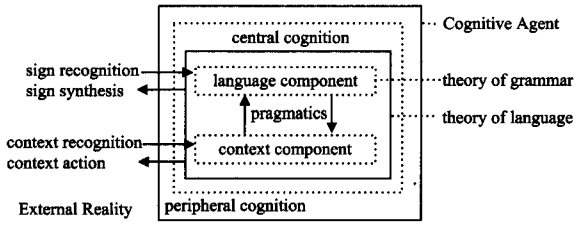

图 2

除了数据结构和算法之外，人造的认知主体还必须拥有识别和行动界面(interfaces for recognition and action)。识别和行动界面的工作方式都以模式匹配(pattern matching)为原则。在语境层，模式被定义为概念(concepts)，用于编码和存储认知内容。在语言层，语境层的概念体现为实词的字面意义。编码和存储概念类型(type)与个体(token)内容的数据结构被称作命题因子的非递归特征结构。

4 举例说明

数据库语义学在认知主体角色转换模式下，对于自然语言的理解与生成过程是：首先认知主体从语境或说者获得知识，其识别过程即在词典中查找输入的自然语言语表各组成成分(词)的概念类型的过程，然后输出分析得到的词条；接下来，LA-hear 以词典查询结果为输入，根据相关语法规则进行句法语义分析，输出以属性值表现命题内外关系的命题因子；LA-hear 的输出结果自动以命题因子的核心属性值为关键字存入词库；LA-think 阶段，系统在词库内根据命题因子之间的接续规则自动检索存储内容，激活满足需要的命题因子；最后的 LA-speak 以 LA-think 激活的命题因子为输入，经由词汇化和语法化的相关规则输出适当的语表和行动。

中国古典著作《左传》里有这样一句话："晋师军于庐柳"。这句话的句法较为复杂，其核心结构是名词与动词之间的函词-论元关系，同时兼有定语和状语修饰语。以此句为 LA-hear 的输入，分析 DBS 的语言理解过程。例句各词在词典里的查询结果分别如下：

[noun:晋]	[noun:师]	[verb:军]	[verb:于]	[noun:庐柳]
[sem:nms]	[sem:cn]	[sem:vi]	[sem:prep]	[sem:nmc]

每一个词即一个命题因子类型，其第一个属性标注句法范畴，以概念为值，这个值也正是 LA 语法查询和提取数据的关键字。命题因子类型除了第一个属性和语义属性之外，其余属性值暂时为空(这里省略)。

我们在建的以《左传》为代表的中国古汉语词库里的命题因子类型主要有 3 大类，即名词(noun)、动词(verb)和形容词(adj)。每一大类下面根据语义属性的差异再细分为若干小类。如名词又分为专有名词和普通名词(cn)(如上例的"师")，专有名词包括国家名(nms)(如上例的"晋")、地名(nmc)(如上例的"庐柳")等等，分别以 sem 属性值的形式进行标注。形容词大类既包括名词的定语修饰语，也包括动词的状语修饰语。动词包括一价、二价、三价动词以及情态动词等。

按照时间线性顺序对该句进行 LA 句法语义分析。名词"晋"是第一个输入的词，根据相应的 LA-hear 语法规则与下一个词"师"组合，之后"晋师"再与"军"字组合。依此类推，直到以句号为标志的最后一个输入结束。每一个输入都是外界语表(个例)与主体词库存储的内部语表(类型)的匹配过程。每一次组合都是在语言命题因子和规则内要求的模式在属性和值上相互匹配的基础上实施规则操作的过程。例句"晋师军于庐柳"的分析结果如下所示：

[noun:晋]	[noun:师]	[verb:军]	[verb:于]	[noun:庐柳]
[sem:nms]	[sem:cn]	[sem:vi]	[sem:prep]	[sem:nmc]
[mdd:师]	[mdr:晋]	[arg:师]	[mdd:军]	[fnc:于]
[prn:1]	[fnc:军]	[mdr:于]	[arg 庐柳]	[prn:1]
	[prn:1]	[prn:1]	[prn:1]	

各命题因子类型经查询、提取和分析之后，原来为空值的某些属性被赋于新值，命题因子类型由此转化为命题因子个例。每个特征结构中的[prn](proposition)由系统自动赋值。这里该属性的值均为 1，表示这几个命题因子个例都属于同一个命题，而该命题是目前为止向系统输入的第一个命题。因为"晋"和"师"是修饰与被修饰的关系，所以名词"晋"的[mdd](modified)属性值为"师"，名词"师"的[mdr](modifier)属性值为"晋"。"军"在本句中是谓语动词，被赋值给名词"师"的[fnc](functor)属性。反过来，名词"师"被赋值给"军"的[arg](argument)属性。接下来输入的次动词"于"，其[sem](semantic)属性标注为"prep"，其功能相当于副词，起到修饰动词的作用，因此被赋值给"军"的[mdr]属性。反过来，"军"也被赋值给"于"的[mdd]属性。最后一个输入的名词"庐柳"与前面的次动词"于"构成介宾关系，被赋值给"于"的[arg]属性。相应地，"于"被赋值给"庐柳"的[fnc]属性。这样，同一命题内部词与词之间的关系在句法规则的基础上通过各个词的属性值一一得以体现。至此，听者状态下的语言理解过程顺利结束。

这些新产生的命题因子个例可以分别存储到以相应的命题因子类型为主记录的个例行(a token line)当中。数据库里的命题因子类型纵向排列，其后所跟的命题因子个例在数量是开放的，按照存入的时间顺序横向排列。

DBS 的语言生成过程是 LA-think 和 LA-speak 语法交替作用的过程。LA-think 的规则和 LA-hear 的规则相似。但

是，LA-think 规则的功能是通过在词库中提取给定命题因子的接续因子来驱动导航(navigation)。因为词库里的每一个因子通常都有一个以上的可能接续词，所以 LA-think 必须能够在评估外部和内部刺激、已遍历频率、已知程序以及主/述位结构等等的基础上对备选可接续词进行智能选择。对于有语言功能的主体来说，导航是说者概念化的过程，即说者选择说什么和怎么说的过程。

当词库中的动词命题因子，如"军"(仍以"晋师军于庐柳"为例)由外界或者内部刺激激活时，LA-think 的起始状态 STs 也被激活。相关规则包里的规则启动，导航开始。根据"军"命题因子的接续属性[arg]和[mdr]的值，导航激活"师"和"于"命题因子，接下来再分别根据这两个命题因子的接续属性[mdr]和[arg]的值激活"晋"和"庐柳"。如下所示：

[verb:军]	[noun:师]	[verb:于]	[noun:晋]	[noun:庐柳]
[sem:vi]	[sem:cn]	[sem:prep]	[sem:nms]	[sem:nmc]
[arg:师]	[mdr:晋]	[mdd:军]	[mdd:师]	[fnc:于]
[mdr:于]	[fnc:军]	[arg 庐柳]	[prn:1]	[prn:1]
[prn:1]	[prn:1]	[prn:1]		

LA-think 的导航结果成为 LA-speak 的输入。语言生成过程中 LA-think 导航遍历到的命题因子的核心属性值被实现为外部符号，即相应的汉字和标点。除了依存于语言的词汇化过程之外，用于中文的 LA-speak 语法系统还必须根据被激活命题因子的相关属性值和彼此之间的句法语义关系来处理语序问题(如果用于其他语言，如英语，还涉及到功能词析出和一致关系等问题)，最后输出正确完整的句子。

结束语 数据库语义学的方法论原则是自然语言的字面组合性，其经验论原则是自然语言交流的时间线性，其本体论和操作性原则是语言和语境信息在认知主体内部的相互匹配。

数据库语义学的基础之一——左结合语法，其本身就是一种区别于范畴语法、依存语法和短语结构语法的特殊句法分析方法。数据库语义学的另一个基础——词库数据库也以其独特的数据结构区别于一般网络数据库，为左结合语法提供理想的运行条件。

遵循自然语言语表的时间线性组合本身大大降低了计算的复杂程度。LA 语法操作过程中的基于模式匹配的规则方法也保证了较低的计算复杂度。处理修饰语句法歧义的语义重叠(semantic doubling)和共指推理等方法也将先天论(naturlism)和转换语法(TG)的指数复杂度降到了线性复杂度[16]。目前该理论方法在德语、英语和汉语的分析实验中都取得了可喜的成果(参见 http://www.linguistik.uni-erlangen.de/clue/de/forschung.html)。角色转换模型以及左结合的句法语义分析方法更准确形象地反映了自然语言理解与生成的实际过程，以命题因子命名的特征结构也更方便地在词库中存储和提取数据。但是由于实践性操作量上的不足和相关技术的不完善性，数据库语义学理论本身及其应用都还有待进一步研究和提高。

参 考 文 献

[1] Austin J L. How to Do Things With Words[M]. Oxford, England: Oxford University Press, 1962

[2] Bar-Hillel J. Language and Information—Selected Essays on Their Theory and Application[M]. Mass: Addison Wesley and Jerusalem Academic Press, 1964

[3] Chomsky N. Syntactic Structure[M]. The Hague: Mouton & Co, 1957

[4] Fraser N. Prolegmena to a Formal Theory of Dependency Grammar[S]. UCL WPL, 1990, 2: 298-319

[5] Halliday M A K, Hasan R. Cohesion in English[M]. London: Longman, 1976

[6] Hausser R. Left-associative Grammar and the Parser NEWCAT [R]. IN-CSLI-85-5. Center for the Study of Language and Information, Stanford/CA: Stanford University, 1985

[7] Hausser R. Foundations of Computational Linguistics, Human-Computer Communication in Natural Language[M]. Berlin, New York: Springer-Verlag, 1999/2001

[8] Hausser R. Turn Taking in Database Semantics[M]// Kangassalo H, et al., eds. Information Modeling and Knowledge Bases XVI. Amsterdam: IOS Press Ohmsha, 2005

[9] Hausser R. A Computational Model of Natural Language Communication; Interpretation, Inference, and Production in Database Semantics[M]. Berlin, Heidelberg, New York: Springer, 2006

[10] Hausser R. Comparing the Use of Feature Structures in Nativism and in Database Semantics[M]// Jaakkola H, Kiyoki Y, Tokuda T, eds. Information Modelling and Knowledge Bases XIX. Amsterdam: IOS Press Ohmsha, 2007

[11] Mann W C, Thompson S A. Rhetorical Structure Theory: A theory of text organization[M]// Polanyi L, ed. The Structure of Discourse. Ablex, 1988

[12] Montague R. Formal Philosophy[M]. New Haven, CT: Yale University Press, 1974

[13] Quillian M. Semantic Memory[M]// Minsky M, ed. Semantic Information Processing. MIT Press, 1968: 227-270

(上接第 170 页)

[16] 刘伟. 基于粗集理论不完备数据的改进算法[J]. 吉林师范大学学报：自然科学版，2007，28(3)：113-114

[17] 王小菊，蒋芸，李永华. 基于依赖度之差的属性重要性评分[J]. 计算机技术与发展，2009，19(1)：68-70

[18] 韩忠华，刘春光，王长涛，等. 基于属性依赖度分析的粗糙集数据挖掘方法应用[J]. 沈阳建筑大学学报：自然科学版，2009，25(5)：1010-1013

[19] 徐章艳，刘作鹏，杨炳儒等，一个复杂度为 max($O(|C||U|)$, $O(|C|^2|U/C|)$) 的快速属性约简算法[J]. 计算机学报，2006，29(3)：391-399

[20] Bay S D. The UCI KDD repository[EB/OL]. http://kdd.ics.uci.edu, 1999

[21] Øhrn A. Rosetta Technical Reference Manual[EB/OL]. http://www.idi.ntnu.no/aleks/rosetta, 1999

[22] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update[J]// SIGKDD Explor. Newsl., 2009, 11(1): 10-18