应用语言学译丛

自然语言交流的计算机模型

数据库语义学下的语言理解、推理和生成

〔德〕罗兰德 · 豪塞尔 著

商務印書館
创于1897 The Commercial Press

应用语言学译丛

自然语言交流的计算机模型

——数据库语义学下的语言理解、推理和生成

〔德〕罗兰德·豪塞尔 著

冯秋香 译

冯志伟 审校

商務印書館 The Commercial Press

2016年·北京

图书在版编目(CIP)数据

自然语言交流的计算机模型:数据库语义学下的语言理解、推理和生成/(德)豪塞尔著;冯秋香译. —北京:商务印书馆,2016

ISBN 978-7-100-11518-6

Ⅰ. ①自… Ⅱ. ①豪… ②冯… Ⅲ. ①自然语言处理-语言模型-研究 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字(2015)第 191095 号

应用语言学译丛

自然语言交流的计算机模型

——数据库语义学下的语言理解、推理和生成

〔德〕罗兰德·豪塞尔 著

冯秋香 译 冯志伟 审校

商 务 印 书 馆 出 版

(北京王府井大街 36 号 邮政编码 100710)

商 务 印 书 馆 发 行

北京市白帆印务有限公司印刷

ISBN 978-7-100-11518-6

2016 年 3 月第 1 版 开本 787×960 1/16

2016 年 3 月北京第 1 次印刷 印张 27½

定价:56.00 元

《应用语言学译丛》

审校者的话

本书作者罗兰德·豪塞尔(Roland Hausser)是德国爱尔兰根-纽伦堡大学计算语言学教授。他先后出版了《表面组成语法》《自然人机交流》《计算语言学基础-人机自然语言交流》和《自然语言交流的计算机模型》等多部专著,发表文章近百篇。Hausser 是"左结合语法"(Left-Associative grammar,简称 LA)的创始人,后来他又进一步提出了"数据库语义学"(Database Semantics,简称 DBS)和完整的"语表组合线性内部匹配"理论(Surface compositional Linear Internal Matching,简称 SLIM),在计算语言学界形成了他自己独特的风格。

我与 Hausser 教授曾有一面之交。2002 年联合国教科文组织(UNESCO)韩国委员会在韩国首尔(Seoul)举行了一次关于"信息时代的语言问题"的学术研讨会,我和 Hausser 都被邀请参加了这次会议,在会议期间的交谈中,我对 Hausser 的独特理论有了初步的了解,回国之后,我又细读了他的《计算语言学基础-人机自然语言交流》(英文版)一书,对他的理论又有了进一步的认识。我认为 Hausser 教授是一位具有独创精神的计算语言学家。

Hausser 认为,面向未来的计算语言学的中心任务就是研究一种人类可以用自己的语言与计算机进行自由交流的认知机器。因此,自然语言的人机交流应当是计算语言学的中心任务。计算语言学研究应当通过对说话人的语言生成过程与听话人解释语言的过程进行建模,在适宜的计算机上复制信息的自然传递过程,从而构建一种可与人用自然语言自由交流的自治的认知机器,这样的认知机器也就是机器人(robot)。为了实现这一目标,必须对自然语言交流机制的功能模型有深刻的理解。

Hausser 提出的"语表组合线性内部匹配"(SLIM)理论以人作为人机交流的主体,而不是以语言符号为主体,突出了人在人机交流中的主导作用,

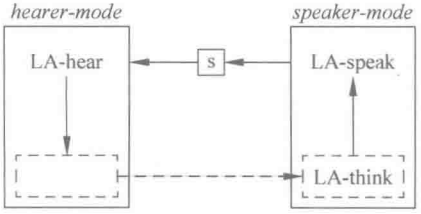
SLIM 理论要求通过完全显化的机械步骤，使用逻辑和电子的方式来解释自然语言理解和自然语言的生成过程。因此，SLIM 理论与现代语言学中的结构主义、行为主义、言语行为等理论是不同的，具有明显的创新特色。

SLIM 理论强调“表层成分”（Surface），以语表组合性作为它的方法论原则；SLIM 理论强调“线性”（Linear），以时间线性作为它的实证原则；SLIM 理论强调语言的“内部因素”（Internal），以语言的内部因素作为它的本体论原则；SLIM 理论强调“匹配”（Matching），以语言和语境信息之间的匹配作为它的功能原则。事实上，SLIM 这个名字本身就来自于这四项原则的英文名称的首字母缩写。

SLIM 理论的技术实现手段叫作“数据库语义学”（DBS）。DBS 是把自然语言理解和生成重新建构为“角色转换”（turn-taking）的规则体系。角色转换指的是从“说话人模式”（speaker mode）向“听话人模式”（hearer mode）的转换，或者从“听话人模式”向“说话人模式”的转换。

在自然语言的实际交流过程中，第 1 个过程是听话人模式中的自然主体从另一个主体或者语境获得信息，第 2 个过程是自然主体在自己的认知当中分析信息，第 3 个过程是自然主体思考如何做出反应，第 4 个过程是自然主体用语言或者行动做出反馈。

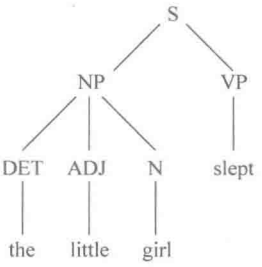
DBS 的输入与第 1 个过程相似，要求计算机或者机器人具备外部界面。接下来匹配语境和认知的内容，采用左结合语法（LA）来模拟第 2 个过程，这个左结合语法是处于听话人模式中的，叫作 LA-hear。左结合语法的第二个变体负责在内存词库中搜索合适的内容，叫作 LA-think，这一部分操作对应于第 3 个过程。左结合语法的第三个变体的任务是语言生成，叫作 LA-speak，模拟第 4 个过程。如下图所示：



在这个图中，听话人模式的 LA-hear 模拟第 2 个过程，说话人模式的 LA-think 模拟第 3 个过程，LA-speak 模拟第 4 个过程。

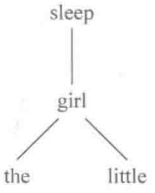
DBS 的分析结果用 DBS 图（DBS graph）来表示。DBS 图是一种树结构，但是，DBS 图的树结构与短语结构语法和依存语法的树结构有所不同。

例如，英语的句子“The little girl slept”（那个小女孩睡着了）用短语结构语法分析后的树结构如下：



在这个短语结构语法的树结构中，S（句子）由 NP（名词短语）和 VP（动词短语）组成，NP 由 DET（限定词），ADJ（形容词）和 N（名词）组成，它们分别对应于单词 the，little 和 girl，VP 对应于单词 slept。句子的层次和单词之间的前后线性关系都是很清楚的，但是，在组成 S 的 NP 和 VP 之间，没有说明哪一个是中心词，在组成 NP 的 DET，ADJ 和 N 之间，也没有说明哪一个是中心词，句子中各个成分的中心不突出。

用依存语法分析后的树结构如下：



在这个依存语法的树结构中，全部结点都是具体的单词，没有 S，NP，VP，DET，ADJ，N 等表示范畴的结点，各个单词之间的依存关系清楚，这种依存关系是二元关系，支配者是中心词，被支配者的从属词。但是，单词之间的前后线性顺序不如短语结构语法的树结构那样明确。

用 DBS 图分析后的树结构如下：

DBS 图的树结构中，着重对语言内容进行分析，因此，没有表示定冠词 the 的结点，结点上的单词都用原型词表示。DBS 图最突出的特色在于，DBS 图树结构的结点之间的连线各自有其明确的含义，连线不仅表示结点之间的依存关系，还可以根据连线走向的不同来表示不同的功能：垂直竖线“|”表示修饰-被修饰关系，例如，上图中 little 与 girl 用垂直竖线相连，表示 little 修饰 girl；左斜线“/”代表主语-动词关系，例如，上图中 girl 与 sleep 用左斜线相连，表示 girl 是 sleep 的主语。此外，DBS 图树结构还使用右斜线“\”表示宾语-动词关系，使用水平线“—”表示并列关系。由于连线走向的不同可以表示不同的功能，这样的树结构表示的信息比短语结构语法的树结构和依存语法的树结构丰富多了。这是 DBS 图树结构最引人瞩目的特点。

上面的 DBS 图中表示了 little 做 girl 的修饰语，girl 做 sleep 的主语，表达的是句子中单词之间的语义关系，所以，Hausser 把这样的 DBS 图叫作“语义关系图”（the semantic relations graph，简称 SRG）。

如果把 DBS 图中每个结点上的单词替换为代表其词性的字母，那么语义关系图就变成了“词性关系图”（the part of speech signature，或者简写为 signature）。上一例句的词性关系图如下所示：

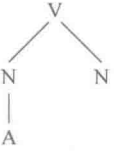
语义关系图和词性关系图是同一句子内容的不同表示，它们表示的内容相同，表示的形式不同。

Hausser 在 2011 年的新书中还提出了另外两个图：一个是“编号弧图”（the numbered arcs graph，简称 NAG），另一个是“语表实现图”（the surface realization）。这两个图分别表现如何从内容生成语言的过程和结果。编号弧图表示激活语义关系图的时间线性顺序，也就是说，编号弧图在某种程度上可以说是添加了编号弧的语义关系图。语表实现图表示如何按照遍历顺序生成语言的表层形式。

例如，英语句子“The little girl ate an apple”（这个女孩吃了一个苹果）的语义关系图（SRG）如下：

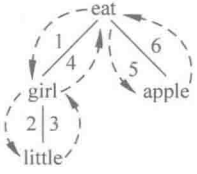
由于语义关系图（SRG）只表示句子的内容，所以，在这个 SRG 中，没有表示定冠词 the 的结点，也没有表示不定冠词 an 的结点，过去时形式 ate 用不定式动词 eat 来表示。

这个句子的词性关系图（signature）如下：



在这个词性关系图中，结点上的单词都替换表示其词性的字母。

这个句子的编号弧图（NAG）如下：



由于编号弧图（NAG）要表示激活语义关系图的时间线性顺序，这种时间顺序用编号弧表示，编号弧用虚线标出，并在虚线旁边用数字注上时间的线性

顺序:结点 eat 首先激活的结点 girl(编号弧 1);接着,结点 girl 激活结点 little(编号弧 2),由于它们之间用垂直竖线“|”相连,因此,可推导出 little 修饰 girl(编号弧 3);由于结点 girl 与结点 eat 之间用左斜线“/”相连,因此,可推导出 girl 是 eat 的主语(编号弧 4);然后,结点 eat 激活结点 apple(编号弧 5),由于结点 apple 与结点 eat 之间用右斜线“\”相连,因此,可推导出 apple 是 eat 的宾语(编号弧 6)。可以看出,所有表示推导的编号弧的方向都是自底向上的。

这个句子的语表实现图如下:

1 2 3 4 5 6

The little girl ate an_apple .

这个语表实现图中的数字表示单词生成的顺序。

数据库语义学(DBS)有两个基础:一个是左结合语法(LA-grammar),另一个是单词数据库(word bank)。左结合语法和单词数据库在 DBS 中紧密结合在一起。Hausser 把左结合语法比作火车头,把单词数据库比作火车运行必需的铁路系统。

单词数据库存储单词的内容,其存储形式是一种非递归的特征结构,叫作“命题因子”(proplets)①。英文“proplet”取自“proposition droplet”,表示命题的构成部分。

一个命题因子是“属性-值偶对”(attribute-value pair)的集合。每个单词或者句子元素的句法语义信息都体现为相应的属性-值矩阵(attribute-value matrix)。例如,汉语“学生”这个单词的属性-值矩阵如下:

$$\begin{bmatrix} \text{sur}:\text{学生} \\ \text{pyn}:\text{xuesheng} \\ \text{noun}:\text{student} \\ \text{cat}:\text{nr} \\ \text{sem}:\text{pl} \\ \text{fnc}: \\ \text{mdr}: \\ \text{prn}: \end{bmatrix}$$

① 译者冯秋香把 proplets 翻译为“命题粒”,我建议她改译为“命题子”或者“命题因子”,她接受了我建议,改译为“命题因子”。

这样的属性-值矩阵就是单词数据库的“命题因子”。在这个命题因子中，sur 表示“语表”，pyn 表示“拼音”，noun 表示“名词”，cat 表示“范畴”，sem 表示“语义”，fnc 表示“函词”，mdr 表示“修饰”，prn 表示“命题”。

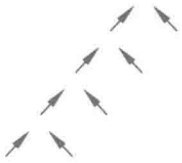
左结合语法是按照自然语言的时间线性顺序自左向右结合进行分析与计算的方法。

具体来讲，每个句子的第一个词为整句分析过程中的第一个“句子起始部分”（sentence start），之后输入下“一个词”（next word），二者经过计算构成新的句子起始部分，再继续与下一个输入的单词进行组合计算。这样不断地进行分析，直到句子结束或者出现语法错误才终止。当出现句法歧义或者词汇歧义时，左结合语法允许按照不同的推导路径并行地继续运算。

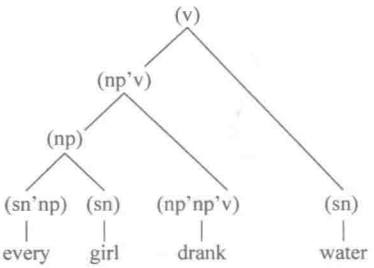
Hausser 将左结合语法与短语结构语法进行了对比分析。他指出，左结合语法与短语结构语法是同质的语言分析方法。它们之间的差异在于：短语结构语法依据的是“替换原则”（the principle of substitution），而左结合语法依据的则是“可接续性原则”（the principle of continuation）。如果以“a，b，c...”来代表语言符号，以“+”代表串联符，那么，左结合语法的计算过程可以表示如下：

$$
\begin{gathered}
a \\
(a)+b \\
(a+b)+c \\
(a+b+c)+d \\
\cdots\cdots
\end{gathered}
$$

左结合语法在进行推导时，总是按照自左向右和自底向上的顺序，沿着树结构的左侧，一步一步地把单词逐一地结合起来的。树结构中的推导顺序如下：



例如,英语句子“Every girl drunk water”(每一个女孩都喝了水)的推导顺序如下:



从这个树结构中可以看出,推导从左侧开始,首先把 every 与 girl 结合起来,形成(np),然后把(np)与 drank 结合起来,形成(np'v),最后把(np'v)与(sn)结合起来,形成(v)。

整个推导过程遵循时间线性(time linearity)的原则。所谓“时间线性”,就是“以时间为序,与时间同向”(linear like time and in the direction of time),也就是说,在推导时,要按照时间前后的顺序进行,要沿着时间的方向推进。

上面我简要地介绍了 Hausser 的主要理论和方法,希望这些介绍能够帮助读者更好地理解这本《自然语言交流的计算机模型——数据库语义学下的语言理解、推理和生成》。

本书共分三个部分。第一部分介绍了 SLIM 语言理论的基本框架,包括认知主体的外部界面、数据结构和算法。这一部分涉及很多对整个系统至关重要的问题,比如概念的本质、概念在识别和行动中的作用、不同符号的指代机制、语境层的形式结构,等等。

第二部分系统分析了自然语言的主要结构,以英语在听话人和说话人模式下的示意推导为例。听话人模式下的分析主要介绍如何严格按照时间线性顺序将函词-论元结构(hypotaxis)和并列结构(parataxis)编码为命题因子,并把共指(coreference)作为推理基础上的二级关系来分析。说话人模式下的分析主要介绍如何在词库内进行以提取内容为基础的自动导航,如何按照相应语言的语法要求输出正确的词形、语序,如何析出适当的功能词,等等。

第三部分介绍英语断片，作者构建了一个功能完整但覆盖面有限的英语交流体系。这部分详细介绍了如何理解和生成小样本文本，对词汇、LA-hear、LA-think 和 LA-speak 进行了明确定义。

本书为计算语言学相关的研究人员、学生和软件工程师等提供了一个对自然语言交流进行理论分析的功能框架，这个框架可以适用于任何自然语言的自动处理。

本书译者冯秋香是大连理工大学外国语言学及应用语言学硕士，计算机科学与技术方向博士，具备良好的语言学和计算机科学的跨学科背景，又有很扎实的英语功底。她从 2009 年 10 月开始，到德国爱尔兰根-纽伦堡大学学习，师从 Hausser 教授研究“左结合语法”。她熟悉 Hausser 教授的计算语言学理论，对于 Hausser 的“数据库语义学”和“语表组合线性内部匹配”理论有深入的了解。我觉得，冯秋香是本书最适合的中文译者，这个中文译本忠实于原文，译文准确精当，通顺流畅，可读性强。

商务印书馆请我审校此书。我对照本书的英文原文《A Computational Model of Natural Language Communication—Interpretation, Inference and Production in Database Semantics》，仔细地审校了冯秋香的中文译本，并参照有关材料，在这里介绍一些与本书有关的背景知识，希望对于读者理解本书有所帮助。我相信，本书中译本的出版，一定会增进我国语言学界对于当前国外计算语言学独创性理论的了解，从而推进我国计算语言学研究的发展。

冯志伟

2014 年 2 月 4 日，于杭州仓前

前　言

《自然哲学的数学原理》[①]第一版的前言里，牛顿把力学分为理论力学和实用力学。理论力学又被称作理性力学，包括精确示范等；实用力学则包括所有的手工技术。如果用同样的方法来对当今的语言学进行分类，会怎样呢？

牛顿会首先强调本学科的重要性，但我们不想这样做。我们直奔主题：什么是理论语言学？什么是实用语言学？实用语言学的例子有语音识别、桌面出版、文字处理、机器翻译、内容提取、文本分类、互联网查询、自动辅导、对话系统和其他所有的自然语言的应用。这些实际应用催生了对实用语言学方法的巨大需求。

但是，现有的实用语言学方法还远远不能满足用户的需求和期待。到今天为止，最成功的实用语言学方法是基于统计学和元数据标注的方法。这些是快速解决的方法（smart solutions）[②]，不需要关于自然语言交流过程的一般性理论支持，其目的是最大限度地挖掘每一次应用或者每一类应用的特殊性及其本质上的局限性。

我们来看一下实用力学：从准确预测潮汐到预测行星未来的位置，从炮弹瞄准到登陆月球等，都是力学的实际应用。和语言学应用一样，力学的实际应用，对方法也产生了巨大的需求。

但是，和语言学不同的是，实用力学的方法不但能够满足这一需求，甚至还超出人的想象。其原因是，牛顿的理论在应用于具体实践的同时，能够

---

① 拉丁语原文全名：*Philosophiæ Naturalis Principia Mathematica*（1687），英文全名：*The Principia: Mathematical Principles of Natural Philosophy*.

② 见 FoCL, Section 2.3. 与可靠解决方案（solid solution）相对.

保持与传统工艺技能之间的相容性。虽然每一次应用都很艰难,需要理论知识和实践经验相结合,但是,其结果总是好的。

这就很自然地引出了一个问题:语言学能不能也这样呢?能不能把语言学理论直接转换成各种实际应用的有限的个别的背景,从而设想一个新的能够满足各式各样需求的框架呢?对于基础研究来说,这是一个相当大的挑战。

为了构建一个完整的、具有普遍性的语言学框架,我们首先要重建人类自然语言交流的认知"力学"。本书讨论的数据库语义学(*Database Semantics*,DBS)理论[①]就是会说话的机器人的陈述性规范说明(declarative specification)。数据库语义学在实际应用上的潜力和它能够成功地、充分地模拟人类认知的能力直接相关。这一点是我们这个研究项目的本质。[②]

一个会说话的机器人的陈述性规范说明必须是一个能够有效地实现自然语言交流机制的功能模型。为了确保完整性,该模型必须以人与人之间的基于语言的互动为原型。该模型的功能性和数据覆盖面必须通过具体实践来验证,也就是要有一个与之相应的运行有效的计算机程序。从长远看,功能性、完整性和可验证性相结合是模型升级成功的最佳科学基础。

由此得到的系统能够应用于所有与自然语言交流相关的实践活动。大多数情况下,只要降低该模型的功能性和数据覆盖面就可以满足某一具体实践的要求。例如,会说话的机器人具备认知功能、人工视觉、操纵以及移位功能等,要建立一个电话的自动对话系统,只需要用到它的认知功能。[③]

其他的应用,例如我们熟知的机器翻译,在降低机器人功能性的同时,还要求对理论进行扩展。不过,数据库语义学有坚实的基础来满足这个要

---

① 作为一个具体的科学理论的名称,数据库语义学(Database Semantics)每个词的首字母都要大写,以区别于一般用法,如数据库语义约束条件(见 Bertossi, Katona, Schewe, and Thalheim (eds.) 2003)。

② 和任何有实际应用的基础科学一样,自然语言交流的计算模型也有可能被误用。这就要求在保证学术自由、信息获取和对话自由的同时,按照法律的明确规定来制定责任制度等,来保障隐私和知识产权。

③ 类似的策略也适用于诸如自动语法检查、内容提取、索引以提高互联网查询的查全率和精度或者自动语音识别等具体应用。

求,因为它也可以模拟单语交流,包括单语理解的过程。

另外,不依存于实践的(理论上的)任何有关词典数据覆盖率、自动词形识别、句法-语义分析、绝对知识和情景知识、推理等方面所取得的进步都可以直接提高现有理论的实际应用能力。方法很简单,就是把相关的部分定期地替换为新版本。这种可能性的来源在于,理论所提供的各个模块以体现功能为目的,各个界面的定义也很明确。

下面我们来尽可能直接地、简单地介绍一下数据库语义学。本书面向语言学和自然语言处理领域的在校研究生、其他研究人员,以及软件工程师等。语言哲学、认知心理学和人工智能等领域的学生和研究人员也可以参阅本书。

对计算语言学和数据库语义学还比较陌生的读者可以读一读《计算语言学基础》(*Foundations of Computational Linguistics*,1999,2001 第二版)。作为一本教材,《计算语言学基础》系统地描述了传统的语法,对各种语言学方法的历史背景也作了对比分析,并提出了 SLIM 语言学理论。本书也采用了这一理论。

认知心理学方面的知识储备对理解本书也有一定的帮助,如 Anderson 的 ACT-R 理论(见 Anderson and Lebiere 1998)。和数据库语义学一样,ACT-R 理论在本质上是以符号,而不是以统计为基础的。它也把计算模拟的方法作为验证方法。但是,ACT-R 理论的研究焦点是记忆、学习和问题求解,数据库语义学的核心是模拟自然语言交流过程中的说者模式和听者模式。

致　谢

本书完成于爱尔兰根-纽伦堡大学的计算语言学系。我要感谢系里的所有成员：Matthias Bethke，Johannes Handl，Besim Kabashi 和 Jörg Kapfer（按姓氏的字母顺序排列）。他们和我一起广泛而深入地讨论了与本书相关的理论和实践两方面的技术问题和概念问题，并提出了很多很好的建议。我还要感谢我的学生们，尤其是 Arkadius Kycia。他是第一个采用 Java™ 语言来编写说者、思考和听者三个模式下的 DBS. 1 和 DBS. 2 应用程序的人。另外要感谢的人是 Brian MacWhinney（卡耐基梅隆大学，匹兹堡）和刘海涛教授（中国传媒大学，北京）。他们对我的早期手稿提出了中肯的意见。Mike Daly（Dallas）承担了手稿的校对工作，也提出了宝贵的建议。Marie Hučinova（查尔斯大学，布拉格），Vladimir Petroff（东北大学，波士顿），Kiyong Lee（韩国大学，首尔），以及 Springer 的编辑们，在本书出版的最后阶段做了大量的改进工作，在此一并致谢。本书如有错误之处，责任全部在我个人。

罗兰德·豪塞尔

2006 年 2 月

爱尔兰根-纽伦堡