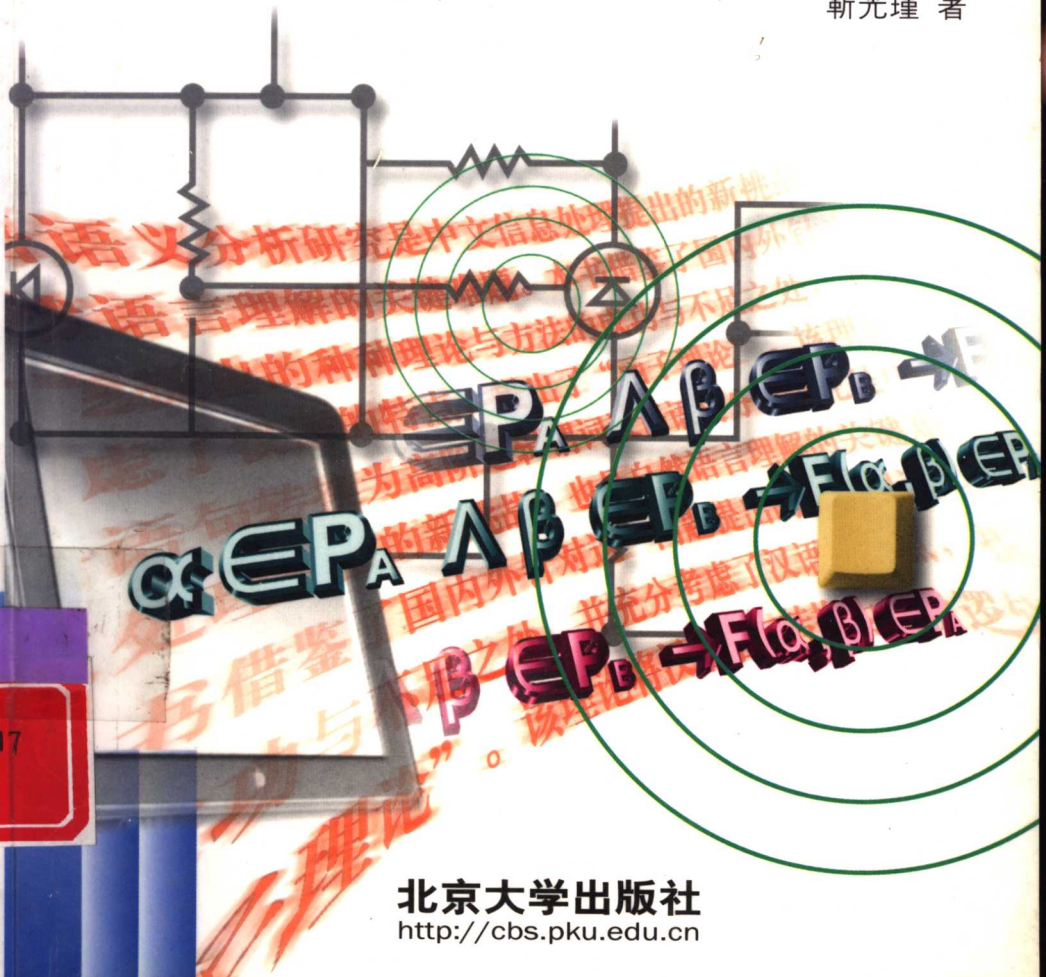
现代汉语
动词语义计算理论

靳光瑾 著



北京大学出版社
http://cbs.pku.edu.cn

现代汉语
动词语义计算理论

靳光瑾　著

北京大学出版社
2001年，北京

图书在版编目(CIP)数据

现代汉语动词语义计算理论/靳光瑾著. -北京:北京大学出版社,2001.6

ISBN 7-301-04993-5

Ⅰ.现… Ⅱ.靳… Ⅲ.汉语-动词-言语统计-现代 Ⅳ.H087

中国版本图书馆 CIP 数据核字(2001)第 032127 号

书　　名:现代汉语动词语义计算理论

著作责任者:靳光瑾

责任编辑:徐刚

标准书号:ISBN 7-301-04993-5/H·615

出 版 者:北京大学出版社

地　　址:北京市海淀区中关村北京大学校内　100871

网　　址:http://cbs.pku.edu.cn/cbs.htm

电　　话:发行部 62754140　编辑室 62752028　邮购部 62752019

电子信箱:zpup@pup.pku.edu.cn

排 版 者:兴盛达打字服务社　62549189

印 刷 者:北京银祥福利印刷厂

发 行 者:北京大学出版社

经 销 者:新华书店

850 毫米×1168 毫米　32 开本　8.25 印张　215 千字

2001 年 6 月第 1 版　2001 年 6 月第 1 次印刷

定　　价:16.00 元

内容提要

汉语语义分析研究是中文信息处理提出的新挑战，也是自然语言理解的关键难题。本书借鉴了国内外针对这一问题提出的种种理论与方法的成功与不足之处，并充分考虑了汉语的特点，提出了“函子理论”。该理论将汉语语句转换为高阶逻辑谓词公式这一至今无从入手的难题分解为两步，插入了“逻辑函子”这一中介，从汉语句子生成函子，从函子生成逻辑式，使得整个语义计算成为可能。同时强调区分词语的内涵义和外延义，有效地解决了计算语言学中的不少问题，并设计了一个计算机实现系统。本书的出版，有望为中文信息处理打开一个新的局面。

序(一)

众所周知,语言是信息的载体,思维的表现形式。研究语言,实际上是研究人们如何思维和表达、理解意思。所以,有关语言形式上的研究,可以看成是用来研究意义表达和理解的手段、方法和过程。语义分析也就成为语法分析的动机、检验手段和过程归宿了。这是从语言学角度观察语义研究的必要性。加进计算机处理这个技术因素之后,语义理解就显得更困难了。在当今字(词)处理阶段中人们始终被切分、标注词性、句法分析等一系列技术难点所困扰,所以语义研究似乎在今天并不是迫切需要,而是有待明天甚至更遥远的将来去做的事了。

事实上,计算语言学在方法论上的进展,从单纯的语言规则到概率统计方法、语料库方法以及当今的规则—统计兼容的混合型方法,都提供了在做句法分析的同时做语义分析的必要性和可能性。90年代基于语料库的句法分析器、句法树库研究的同时出现语义标记、对话语义解释的研究工作。荷兰的波特(Bod.R)等人的面向数据的句法分析及其语义标记就是一个例子。文本语音输出、与生活相关的应用智能对话理解系统等应用的需要,也逐渐将语义标记、语义解释、语义理解等研究提到日程上来了。1995年秋,香山专家座谈会上一致认为当前中文信息处理已开始进入新的阶段——汉语语义分析研究阶段,这是自然语言理解及人机界面面临的新的历史重任,向汉语学界、计算机科学界提出的新的挑战。

徐烈炯教授(《中国语文》1996.4)指出:以往国内的语义研究主要是狭义语义学(研究语言系统内部词语与词语之间的语义关系,如同义、反义、多义、歧义等),范围比国外的广义语义学(研究词语与外界事物的关系、指称、真值等)窄,因此形成了空白地带。西方语言学

研究也从人文学领域开始，但后来进入科学领域。用汉语材料研究语义只要从狭义语义学跨入广义语义学，就会作出巨大贡献。这是很中肯的意见。我再加上一个注释，上面提到的广义语义学其实是一种抽象的形式方法，具体说是借助于数学，主要是数理逻辑所构建的句义模型框架下描述词汇特征和进行语义分类(平凯 M. Pinkal，《逻辑与词汇学》，1995)。因此，现代语义学已成为语言学、哲学、数理逻辑、计算机科学的交叉领域。

上面扼要说明了语义研究的重要性以及研究的主要方向和领域。这实际上也是向读者介绍本书的背景领域。靳光瑾博士在攻读博士学位期间承担了国家自然科学基金项目"现代汉语语义计算框架研究"(该项目得到黄昌宁教授的支持和指导)，以逻辑语义为基础，对现代汉语语义抽象表示形式及其语义理解进行了富有成效的研究。本书及其他发表的相关论文，都是属于上文所说的广义语义研究领域内的成果，读者尽管会遇到一些不熟悉的基础知识，但是会感受到强烈的新鲜感，因为这在国内是很少见的。

形式化描述例证一节中，引用了吕叔湘先生的"掉个过儿还是一样"一文中的例句，告诉我们计算机为什么可以理解并且证明词语"掉个过儿"的前后两句是等价同义的。尽管没有具体算法和程序，也没有形式推导，但是给出了一个清晰的分析思路，这个分析过程完全可以由计算机来实现。这个分析过程同时也能让语言学家、计算机专业的读者读懂，由此体会到形式化描述跟逻辑分析是一回事，也完全符合人们日常理解的思路。

形式化描述和逻辑语义解释一节中，引用例句"他的老师教得好/他的老师当得好"，运用内涵/外延这两个不同的语义概念区分了性质描述和指称对象之间的不同，从而成功地解释了例句在句法上和语义上是同构的，都表示了领属关系。这可以说是从词义范围跨入指称外界事物和性质范围的极好的开端。从此，作者有关内涵/外延、句法/语义同构、提升/下降等一系列语义概念的研究及应用在现

代汉语语法、语义研究中获得成功,走出了一条新路。相信这对于处理汉语是至关重要的,会越来越受到学术界的关注。

要使语义分析成为一个计算过程,需要建立汉语计算语义理论,其中,最为关心的而又感到困惑的是汉语语义的表示形式是什么?汉语语义的抽象表示形式是关键。汉语语义分析的中心目标是句义分析,这就在很大程度上制约了语义标记该采取什么形式。广义短语语法的复杂特征集包含句法和语义两类信息。虽然这些以内涵逻辑形式出现的语义特征也都制约于一个语义模型,由这些特征直接用来解释句义,仍然有较大的距离,但是最好的抽象表示形式无疑是逻辑表达式,它一方面与语句相对应,另一方面在模型和可能世界下(包括情景、上下文、知识)可以完整地解释、理解句义。然而要从汉语句子直接转换到逻辑式,在目前情况下还不具备这样的条件(需要定义汉语范畴语法、转换文法)。于是,作者寻找一种介于汉语语句形式和逻辑表达式之间的中介形式,那就是"函子",它具有逻辑语义,虽然在初始形式上还不是个很规整的逻辑式,而是一个由词语组合成的式子,但经过进一步的规约化简,可以使函子归约为典范形式。这就使得将汉语语句转换到高阶逻辑谓词公式这样一个至今无从下手的难题,分解为两个可行过程:生成过程和转换过程。前者是从汉语句子生成函子,后者从函子转换为逻辑式,于是整个语义计算过程就成为可行的了。这是本书在语义分析方法论上的一个突破。

除此之外,本书与国内现代汉语语法研究及语义研究的重大区别还表现在下面几点:

一、函子与动词及其配置的必要语法成分的组合相对应,但又不相同。前者寻求语义上的约束关系,后者是句法上的关系。逻辑式中的论元与谓词之间有明显的约束关系。本书"函子提取"一节中,多项 NP 合并成一个复合表达式,后者对应一个集合,有效地解释了动词与多项名词之间逻辑语义关系,而不是简单的动宾关系。所以说,函子与动—名结构完全是两个不同层次的构造。在这一节

中使用了数学工具(集合论、Lambda 演算)来描述,这对国内语言学家甚至一般的计算机专业人员来说,在阅读上会有一定的困难,但是这类描述在国外,对语言学系、计算语言学系的研究生来说,都是必要的知识。这就是差距。我们想要跟国际接轨,能尝试用这些数学表达方式来刻画汉语的语义,正像徐烈炯教授所鼓励的,是一个进步,一个贡献,应该发扬和推广。

二、函子在论元位置上要求完备,也就是说不能空缺,即使在汉语语句中普遍允许存在缺省,但是在逻辑上,必须明确地逐一求解出这些逻辑成分,甚至如连动结构中的逻辑主语,都得补出来。所以具有完备项的函子是一个完整的高阶逻辑谓词。本书用了较多的篇幅论述缺省及其求解策略,正是因为逻辑语义上的需要。事实上不少国外语言学家已经充分注意到汉语中的缺省是汉语重在表达内在语义的一个重要特点。要理解汉语必须求解缺省。作者分别在句内句外讨论缺省成分的求解,它们是在情景、上下文、知识和模型下求解这些缺省成分的。相比之下,有上下文的复句容易求解,没有上下文的简单句,反而更多地依赖于情景。因此作者先介绍复句的句外求解,后介绍单句的句内求解。这样的安排,也让读者容易读懂。应该说,关于缺省的分析及其处理,在国内篇章学研究中也是少见的。

三、函子理论应用了汉语动词配价知识。从逻辑配价而不是句法配价的角度考虑,也就是给逻辑谓词的逻辑论元配置合适的项及语义角色,因此可以满足如"今天中午午饭每人食堂五块钱吃一份快餐"中关于动词"吃"所涉及到的理解问题:谁吃,吃什么。对于其中多项名词的语义角色,有一个独特的有效的处理意见。

四、存现句结构及语义特征分析,在分析的视角和方法上都很有特色。从适应于计算机理解这样的角度出发,实际上就是在认识上和演绎推导上寻求一致的形式方法。对存现句的定义重在语义特征,由此给出的分类取决于语义特征的确定性和模糊性。由这两个性质和程度,构成一个连续统,包含了通常所说的各种存现句。这个

分类不依赖于动词的论元施事、受事的区分,也不考虑动态、静态的区分,最大的方便就是便于进一步形式化描述。不同于汉语传统语法的考虑,并不依赖过多的汉语语感和语料,对于要求形式化程度很高的计算机来说,这种语义分析方法是一种成功的尝试。

谨以此为序。

陆汝占

2000 年 8 月于上海

序(二)

靳光瑾博士要我为她的专著《现代汉语动词语义计算理论》写序,对我来说,等于是给我出了一道难题,因为我是一直从事现代汉语本体研究的,对计算语言学,我是门外汉,对语义计算,更可以说是一窍不通。不过我还是答应了,因为这可以逼着我学些新东西,好让自己的知识来点儿更新。

自然语言处理,无论在国内和国外,都已较好地或者说基本上解决了字处理、词处理问题,现在进到了句处理阶段。句处理包括两部分内容,一是让计算机准确理解自然语言中一个句子的意思,二是让计算机生成符合自然语言规则的句子。而首先要研究解决的问题是怎么让计算机理解自然语言中每个语句的意思。要解决这个问题,难度是相当大的,其中尤以语义分析研究为最难。全世界从事自然语言理解和处理研究、从事计算语言学和机器翻译研究的学者专家都在致力于解决这一难题。

最早,大家考虑的思路是怎么让自然语言的句法语义规则直接转换为机器能够接受的程序语言;后来又从实践中认识到直接转换是不可能的,必须先将自然语言经数学抽象过程转换为逻辑表达式,再由逻辑表达式转换为机器能接受的程序语言。为实现这一目标,大家各显神通,有的采取以理性主义为哲学基础的基于规则的处理方法,这种方法或是以一定的形式文法系统来表述自然语言中大小成分间的组合规则,或是"以概念化、层次化、网络化(简称"HNC")为基础"来提供概念组合、语义表述的规则;有的采取以经验主义为哲学基础的基于语料库统计的方法,这种方法是以各种统计数据来显示语言成分间的组合可能性;也有的提出将上述两种方法结合的做法。但都未能取得令人满意的结果。靳光瑾博士这部著作扼要回

顾、介绍了国内外在自然语言理解与处理方面所提出和运用的种种理论与方法的成功与不足之处,进而在前人研究与实践的基础上,充分考虑了汉语的特点,特别是大量缺省的特点,提出了一种新的理论思想和新的策略、方法,那就是"函子理论"和关于区分词语的内涵义和外延义的思想。她的函子理论,假设并定义了一个从自然语言到逻辑表达式的中介形式——逻辑函子(logic functor)。函子部分主要是求解缺省和求解函子的组合。函子,上接自然语言的句法层面,下接逻辑层面,它在自然语言和逻辑表达式中间起承上启下的作用。而她关于区分词语的内涵特征义和外延义的思想,有效地解释了领属关系的逻辑语义,体现了汉语中频繁出现的抽象提升和具体操作等一系列语义解释的特点。

根据靳光瑾博士提出的函子理论和关于区分词语的内涵义和外延义的思想,从自然语言到逻辑表达式,不一步到位,而是分步完成。具体求解过程大致如下:

自然语言中的句子

↓

使论元项完备的函子及扩展函子

↓

带有λ算符的高阶谓词公式

↓

句子的抽象数学表达式

↓

给出具体的语义解释

该书对为什么要建立函子这一中介形式,怎么在函子的层面求解缺省和求解函子的组合,进行了充分的合理的论述,提出了一套求解规则,开发了有效的函子提取技术,并分别从句外求解函子缺省成分和句内求解函子缺省成分这两个方面,具体求解了现代汉语里的存现句、带"得"补语句、二价动词主语后移句、重置动词句、紧缩句、连动句、递系句、复合句等多种句式的语义解释,实现了可行、高效的句子

语义计算;最后根据她所提出的理论思想,设计了一个计算机实现系统。

函子理论的提出,特别是求解汉语语句缺省成分的方法的提出和逻辑函子的建立,以及词语义中内涵义和外延义的区分,是该书最重要的创新之处。这是计算机处理、理解汉语最为关键的高难度技术之一,这将为解决好中文信息处理中的句处理问题,特别是为解决好从自然语言的语句 S 自动生成逻辑表达式 α 的问题,提供了一种新的路径,并有望使中文信息处理开创一个新局面。毫无疑义,靳光瑾博士的《现代汉语动词语义计算理论》对解决汉语句处理问题将会作出新的贡献。

当然,这不是说汉语句处理的问题就此彻底解决了。真要理想地解决汉语句处理的问题,还有大量工作要做,譬如说,动词义项的判别,动词论元的判别,词语间语法关系的判别,语句结构的正确切分,复杂缺省的求解,情景义的确认,固定短语的处理,汉语语句数学抽象描述的完善,等等。任何科学研究都是对未知世界的一种探索,科学家头脑里想的,纸上说的,都需经实践检验,才能对他所提出的一种新理论作出更为客观的评价。实践的结果可能成功,也可能部分成功,也可能失败。成功也好,失败也好,对科学研究都是一种贡献。因为,前者可以留下成功的足迹,让他人踏着足迹继续前进;后者可以在前进的道路上竖起一块警示牌,让他人免走弯路。

是为序。

陆俭明

2000 年 8 月于北大中关园寓所

目　录