全国高技术图书

汉语计算语义学

——关系、关系语义场和形式分析

吴蔚天 著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
URL:http://www.phei.com.cn

全国高技术图书

汉语计算语义学

——关系、关系语义场和形式分析

吴蔚天　著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书系《汉语计算语言学——汉语形式语法和形式分析》一书的姊妹篇。两者分别系统地讨论了汉语分析的两个极其重要的方面。前者注重语法，本书则注重语义。语法和语义之间的关系向来是语言学界争论不休的问题，同时也是汉语自然语言处理当中未能发挥而应该发挥作用的方面。

本书从汉语形式分析的实际出发提出了计算语义学和关系语义场两个概念，并以此作为枢纽阐述了语法和语义之间的辩证关系，为语言学语义学开辟了一个新的与汉语自然语言处理实践紧密相关的课题。

本书不仅详细地讨论了几个重要的关系语义场在形式分析中的意义和由此导出的形式规则，而且还给出了可供开发实用系统使用的大量数据。

本书观点新颖、内容实际、方法可行，是学习或从事人工智能、自然语言理解、机器翻译、情报学、语义学等研究人员的有益参考书。

丛 书 名：全国高技术图书
书　　名：**汉语计算语义学——关系、关系语义场和形式分析**
著　　者：吴蔚天
策　　划：龚兰方
责任编辑：张燕虹
排版制作：电子工业出版社计算机排版室
印 刷 者：北京大中印刷厂
出版发行：电子工业出版社　URL：http://www.phei.com.cn
　　北京市海淀区万寿路 173 信箱　邮编 100036
经　　销：各地新华书店
开　　本：850 × 1168　1/32　印张：10.625　字数：275 千字
版　　次：1999 年 1 月第 1 版　1999 年 1 月第 1 次印刷
书　　号：ISBN 7-5053-4972-4
　　TP·2449
定　　价：30.00 元

序

本书作者是多年从事开发汉语自然语言处理实用系统的研究者，在研究语言学方面有着明确的目的。于1994年在其《汉语计算语言学》一书中曾提出过现代的语义场划分与语法学中的语法关系脱节的看法；在本书中则更加明确地提出了自然语言处理对划分语义场的要求，以及在研究语法时妥善处理语法和语义间矛盾的重要性，力图使任何一方面的成就都尽可能地为汉语自然语言处理服务。

《汉语计算语义学》一书的出版是研究汉语语法和语义间的关系的一项尝试；为语言学界提出了一个使汉语在21世纪的信息高速公路和人工智能领域中发挥更大作用的研究课题。

本书的副标题是“关系、关系语义场和形式分析”，表明了传统语义学和计算语义学的区别，反映了作者的语义学观点。

本书作者认为语义关系是选择词与词的搭配及理解句子的根据，依此作者为汉语计算语义学提出的定义是“研究语词意义间的关系及其导致的语法结果”。这个定义既规定了计算语义学的范围，也规定了它的目的。

本书作者在讨论语法和语义间的关系时进一步地说明了计算语义学的研究方法以及形式规则的性质、类别和推导原则，并给出了解决汉语形式分析中的若干重大问题的具体规则和用程序语言书写的算法。

本书作者在关系语义场划分方面提出的各种关系均来源于形式分析的需要，并讨论了有利于计算机进行汉语形式分析的关系语义场，特别是名词和名词的结合关系语义场，从而为汉语计算语义学的理论及其应用提出了一个观点和一个方向；同时也体现了

计算语义学与形式分析的密切关系和与计算机的依存关系。

本书为今后的关系义素分析和关系语义场划分研究提供了大量的数据,为读者提供了方便。

我希望本书能使语义学在汉语形式分析方面起到促进作用。我们应该感谢作者的长期且艰辛的劳动以及为计算语义学的发展所作的思考和努力。

赵世开

1995 年 8 月于北京

序

吴蔚天先生的新著《汉语计算语言学——汉语形式语法和形式分析》一书出版不久，他又写成了这部《汉语计算语义学——关系、关系语义场和形式分析》，进一步从语义学的角度探讨了汉语的形式分析问题。我怀着浓厚的兴趣通读了全书，觉得这是一本观点颇为新颖、理论联系实际的好书。

早在70年代末期我在法国格勒诺布尔大学自动翻译中心师从当时的国际计算语言学委员会主席沃古瓦(B. Vauquois)教授研究机器翻译时，沃古瓦教授就向我推荐了由 E. Charniak 和 Y. Wilks 编写的《计算语义学》(Computational Semantics. Amsterdam. North-Holland. 1976)，这是一本研究自然语言处理中的语义学问题的文集，可惜所有的文章讨论的都是西方语言计算机处理中的语义学问题，完全没有涉及汉语。我当时正在做汉语到英语、法语、德语、俄语、日语的多语言机器翻译试验，首先要进行汉语的自动分析，遇到了汉语分析中的许多难题。例如，如何判断含有多个动词的汉语句子中的主要动词？如何决定名词之后的形容词的归属？如何分析由多个名词构成的短语中各个名词之间的关系？如何排除汉语中众多的歧义结构的歧义？我曾经试图根据《计算语义学》中介绍的理论和方法来解决这些难题，但是收效甚微。回国之后我的研究兴趣转到了其它方面，然而这些难题始终环绕在我的脑际，总想找到解决它们的办法和诀窍。读了吴蔚天先生这本专门针对汉语写的计算语义学专著，我长期百思不得其解的这些难题，都一一得到了基本上完满的解答，真令我喜不自禁，犹如在茫茫的迷雾中看到了耀眼的曙光。

吴蔚天先生是我国第一个商品化的汉外机器翻译系统 Sino-

Trans 的主要设计人，他把自己在研究汉语自动分析中的甘苦和心得，从理论的高度做了总结，提出了适用于汉语计算机处理的“汉语计算语义学”，弥补了国际计算语义学中汉语计算语义学研究的不足，这是我国计算语义学研究的新成果。

汉语缺乏形态，动词、名词都没有形态变化，同形歧义结构比比皆是，一个由若干个词构成的语言片段，可能得到若干种结构上截然不同而在语法上却合情合理的解释，往往使研究者举足不定，陷入进退维谷的境地。因此，汉语的自动分析，不能只依靠形态和语法，在很大的程度上必须依靠语义，语义分析在汉语自动分析中起着举足轻重的作用。

近年来，我国学者在汉语的语义学方面作过不少的探索，这些探索涉及到名词的语义分类、动词的语义分类和格框架、形容词的语义分类等等方面，这些研究都取得了可喜的成果。但是，这些研究对于动词与名词、名词与名词、名词与形容词之间的语义关系不够深入。有些采用格语法的汉语形式分析系统，动词的语义框架和名词的语义分类貌合神离，大大地影响了这些系统的运行效率。本书明确指出，汉语计算语义学的核心内容是研究词语与词语之间的语义关系，建立关系语义场。本书把关系语义场的研究作为汉语计算语义学的核心内容，并从关系语义场的高度给动词、名词和形容词进行语义分类。分别讨论了体谓关系义场、名名结合关系义场、名词形容词关系义场，给出相应的形式规则，这就彻底改变了不同词类的语义分类之间貌合神离的局面，使得汉语计算语义学所揭示的规律更加具有可操作性，可以直接在计算机上得到实现。

本书是在汉外机器翻译的实践中提炼而成的，作者把他在多年研究中所积累的大量的实验结果也同时公之于众，如在第 7 章、第 8 章、第 9 章附录中的动词关系义素、关系空位一览表、名词分关系义场、形容词关系义场等都是十分宝贵的数据，具有很高的参考价值，可供开发实用系统的研究人员使用。

汉语计算语义学是汉语自然语言处理研究中的一个新课题，本书所讨论的关系语义场的许多规律，并不是永恒不变的金科玉律，它将会在我国自然语言处理的实践中，得到进一步的检验，并不断地得到修正，使其更加完善。汉语计算语义学还是一块亟待开垦的处女地，希望有更多的研究人员投入到这项全新的研究工作中来，使这块土地开出更加鲜艳的花朵，结出更加丰硕的果实。这样，我们就能有力地推动汉语计算语义学的发展，并对国际的计算语义学作出我们应有的贡献，从而促进我国计算语义学研究的世界化。

冯志伟

1995 年 12 月于北京

前　言

1994年是我国著名语言文字学家钱玄同(1887～1939,字德潜,号疑古)诞辰105周年。他于1917年在《新青年》杂志第3卷第3期上发表了致陈独秀的公开信,信中指出:"人目系左右相并,而非上下相重。试立室中,环顾左右甚为省力,若纵观上下,则一仰一俯,颇为费力,以此例彼,知看横行易于看直行,且右字写此,必自左至右,均无论汉字、西文,一笔一势,罕有自右至左者。然则汉字右行,其法实拙。若从西文写法,自左至右横拖而出,则无一不便"。其建议得到了陈独秀等学者的赞同。作者在1994年的北京晚报上看到陈士森先生的上述考证时恰与本书动笔之时相重,故摘录于此以纪念钱先生的贡献。今天,如果汉字仍然竖写或左行,那么将给自然语言处理带来更多的麻烦。

在英语中,形态变化可以帮助区别词类,定冠词可以区分名词词组;然而汉语既没有形态变化又没有定冠词,所以汉语形式分析中的歧义尤其多。作者曾较为系统地研究过汉语的歧义现象;发现汉语的语法歧义主要有词性歧义、主动词歧义、结构歧义和主、宾语歧义。在分析中如果能够确定名、名结合关系和名、动搭配关系(体谓关系),那么就可以解决近60～70%的语法歧义问题。这两个关系都依赖于名词和动词的语义分类。基础是语义学,特别是解释语义学和语义场。

在用计算机自动分析汉语时引用语义场的概念并利用动词对名词的制约规律以及名词与名词的结合规律一直是基本手段,目的是为了消除歧义,但是并不成功,这就是人们思变的现实。

一条路是将语料库奉为法宝。实则,这并不是一种以逸代劳的新方法,而是返古。Abraham Kaplan早于1950年就研究过语义

与语境的统计关系,对语料库寄予过希望。

以作者之浅见,语料库之所以曾被冷漠和需要谨慎从之的根本原因在于语料的取样问题和远程搭配关系的标注问题。取样理论虽已很成熟,但也只能处理已知母本大小和比较有规律的对象。对于语言这个无穷集是无能为力的。换句话说,没有说明封闭语料库的结果能用于开放语料的论据。人们只能无根据地追求封闭语料库的规模。另外,语料库的标注结果(主要是词性标注)是用关联矩阵表示的;关联矩阵是一个二维表,因此不能直接有效地分析在汉语分析中十分重要的远程搭配关系。例如,确定"我用铅笔没有问题"和"我用铅笔写字"中的"用"字的语法意义所需要的远程判断。

作者认为有思考能力的人在理解汉语句子时依靠的是语义合成,所以只有变革划分语义场的原则、发挥综合的语义搭配作用才可能使自然语言处理向智能化迈进。这就是作者走的一条路。

说起来也不可思议,语义学界至今对划分语义场不可少的语词意义研究中的"意义"当何讲还无定论(本书第 2 章有作者的愚见)。看来也情有可原,汉语词的意义常和语境有关,下面转载(生活报,第 6 版.1994.11.13.)一个幽默片段,通过"意思"的语义思考"意义"的意义:

他说:"她这个人真有意思(funny)"。她说:"他这人真怪有意思的(funny)"。于是人们以为他们有了意思(wish),并让他向她意思意思(express)。他火了:"我根本没那个意思(thought)"! 她也生气了:"你们这么说是什么意思(intention) "? 事后有人说:"真有意思(funny) "。也有人说:"真没意思(nonsense)"。

本书共分 10 章。

第 1 章为"汉语形式分析",叙述了汉语分析中的几个关键性问题。

第 2 章为"现代语义学",叙述了该学科的研究目的与方法,特别是语义场的概念及其划分原则。

第 3 章为“概念依存理论”,叙述了在自然语言处理技术中有代表性的语义描述方法。

第 4 章为“汉语计算语义学”,阐述了其定义、理论及关系语义场的概念。

第 5 章为“关系语义场”,提出了划分语义场时引入“关系”概念的理由,并讨论了在汉语形式分析中需要的关系语义场。

第 6 章为“语法和语义间的辩证关系”,讨论了计算语义学在处理语法和语义问题方面的枢纽作用,并列出了若干重要的形式规则集。

第 7 章为“体谓关系义场”,详细讨论了动词的关系义素以及关系义素不等于义素的问题;附录载有大量实用数据。

第 8 章为“名名结合关系义场”,详细讨论了名名结合关系以及名名结合关系义场的性质和形式规则;附录载有大量实用数据。

第 9 章为“名词形容词关系义场”,详细讨论了形容词后置的语法形式及形式规则;附录载有名词和形容词的关系义场。

第 10 章为“利用关系义场转换拼音流的可靠性”,主要讨论了利用语法规则和关系义场将拼音流自动转换成正确汉字的可靠性;并提出了“自然语言处理当中的永动机”现象。

第 11 章为“后记”,总结了本书所有内容的实践结果,再次阐述了语法普遍性和语义普遍性在自然语言分析中的理论指导意义。

本书在写作过程中采纳了普林斯顿大学吴晓东女士、彭朝阳先生的宝贵意见;在清稿、校稿方面朱玉琳老师付出了辛勤的劳动。特此致谢。

作者通讯地址: e-mail: wtwu@bigfoot.com

Tel & Fax: 010-64223008

吴蔚天

1995 年 6 月于北京

第 1 章　汉语形式分析

关于汉语分析已是老生常谈了[1]。本章为介绍语义在汉语形式分析中的应用再重复几个汉语分析中的关键问题，作为背景材料。

虽然汉语和英语同属分析型语言，然而在很多方面不同于英语。比方说，没有形态变化、没有定冠词、词序多变等等。汉语长期以来没有自己的语法，还是 1896 年，马建忠先生的《马氏文通》第一次引进了拉丁语和英语语法，并用汉语对其诠释。黎锦熙先生于 1955 年出版的《新著国语语法》即是 Nestfield 语法的“中文简版”。自此以后，几乎亦步亦趋地引用国外的有影响的语法理论。甚至有人说：“过去，中国没有系统的语法论著，也就没有系统的语法理论，所有理论都是外来的。外国的理论在哪儿翻新，咱们也就跟着转[2]”。虽有些偏颇，但不失之真实。

黎锦熙先生主张三中心说，即主语、谓语、宾语。这项主张并未得到全面赞同，直到 90 年代才有人将其全面理论化、系统化和形式化，并成功地用于实践[1]。

今天，在汉语形式分析中，也仍然没有扭转汉语语法的形象；照样是在英语的模式中换上汉语；因此，谈汉语的形式分析基本上等于谈英语的形式分析。不过，汉语终究是汉语，所以汉语形式分析还是有其特点的。

第一，由于汉语没有形态变化，同形歧义现象很多，比方说：“行“、”长“、”重”的读法(即语义)，“把“、”打“、”学习“、”设计“、”统计“、”请求”的词性歧义等等都是须要解决的歧义现象；

第二，由于汉语动词没有性、数、格、时态的形态变化，所以当有多个动词同时存在于一个句子中时，需要确定哪一个是谓语动

词(主动词);

第三,由于汉语没有定冠词,所以当两个名词相邻时要判断它们是一个名词词组,还是两个不属于同一个语法功能成分的名词;

第四,由于上面的原因,一个词类序列(语法形式)可以有几种语法合理的结构,所以不能简单地用语法规则归约。

目前,在处理这些问题方面不外乎两类方法,一是短语结构语法和格语法,二是作者提出的属性制约文法和 I-Tree(汉语完全语法树)[见本章参考文献 1,第 120 和第 212 页]。

1.1 词性判断

词性判断是汉语形式分析的第一步。不过在有的方法中并未独立出来,只是在内部先行或并行。较为通用的方法是利用复杂特征集的合一算法。

1.1.1 基于语法规则

基于语法规则就是基于语法特征的匹配,例如:

1. 相邻规则(双向规则,即反之亦然的规则)

(1) 左邻为助动词(如"能"、"会"等词),不能取名词;
(2) 左邻为副词,不能取名词;
(3) 左邻为语助词、副词,取动词;
(4) 左邻为形容词,不能取量词。

2. 远程规则

远程规则指非近邻制约规则。一个词的词性要受到其右侧或左侧的非相邻词的属性约束或语用习惯的约束。例如:

(1) 右侧有方位词,取介词(如"当");
(2) 右侧有可以共现的词,取介词,否则取连词(如"和 ……

相同”、“就 …… 而言”);

(3) 右侧有右邻不为“的”的动词,名、动兼类介词取介词;

(4) 左邻为动词,右侧有右邻为“了”的名词,介、助兼类的介词取助词(如“看上他了”中的“上”字)。

基于语法的规则已作为汉语分析的第一步用于作者等开发的商品化汉外机器翻译系统(SinoTrans ,1992 年 ICCIP 会议论文集或 http://www.gy.com)中。

1.1.2 基于语义

语法规则在词性判断方面主要是依靠词类间的相互制约关系,不过在有些情况下词性不服从普遍性制约关系,例如:

N + NV + N (NV 代表取名词的兼类动词)。

名词的右邻和左邻都允许为动词或名词,且允许同时出现,从而不存在属性为词类的制约关系,而存在着另外一种制约关系,称属性为语义的制约关系或语义搭配关系。语义搭配就是概念制约,所以,词性变化即概念变化。多数情况是动词从具体到抽象的变化。

1.1.3 基于语法、辅以语义

一个实用系统不能用单一的方法处理问题。对于词性判断问题也是一样。如何安排是算法设计的技巧。夸张地说,汉语形式规则杂样化是实用化的关键。这时,在属性集合中有两类特征,语法特征和语义特征(复杂特征集),这两类特征要同时满足某一种制约关系。这种方法常被研究者们采用,然而结果并不理想。本书即为此而写。

除上述方法外,还有自 80 年代末期东山再起的基于语料库的方法。较为成熟的是词性标注语料库。

基于语料库也就是基于词性的共现统计概率。语料库方法可以看成是先将基本语法规则中的词类激活(initiation)后再处理的

方法。为得到共现概率,要求事先对语料库进行加工和统计。方法很多,各有千秋,已非本章的兴趣所在。下面仅重复语料库方法的原则。

概率模型:事件 B 的发生取决于事件 A 的发生的链发事件;

命题:事件 B 发生的概率。

(1) 属性相邻共现概率 P_g

当左邻词(A)出现后右邻词(B)的某一属性的出现概率

$$P_g = P(C_{i+1} | C_i)$$

式中,C_i——左邻词属性(A)

C_{i+1}——右邻词属性(B)

(2) 属性单现概率 P_d

某一个词的某一属性在上下文中出现的概率

$$P_d = P(C_i | W_j)$$

式中,W_j——第 j 个词

(3) 链发概率 P_s

一个词串中的各个词的各种属性共现的概率

$$W_1 + W_2 + W_3 + \cdots\cdots + W_n$$

$$P_s = \Pi[P_g \cdot P_d]$$

当概率为 1 时表示成立,概率为 0 时表示不成立。概率为 $0 < P_S < 1$时表示为歧义。

语料库方法有两个问题,一是被处理语料的代表性,即推而广之的问题;二是概率计算,即模型确定问题。一般是将马尔科夫链简化成二元语法后再处理[3,4],即上面的属性相邻共现概率 P_g。链发概率纯粹是脱离语言本性的数学计算。属性制约文法是按三元语法的性质处理的,每次处理三个,滚动式前进,先 1-2-3,然后 2-3-4 [见本章参考文献 1,第 212 页]。

1.2 主 动 词

在汉语的正式文本中一个句子有一个以上的动词很普通;那

么该怎样判断主动词呢？原则上有下面几种办法。

1.2.1 基于语法

由于汉语的动词没有性、数、格和时态的形态变化，所以用语法确定哪一个是主动词是有困难的。台湾学者在这方面也做了很多工作，下面介绍一种于 1994 年发表的别开生面的“打分法”[5]。

这个方法既利用了常规的语法规则，也利用了可称之为评价函数的计算公式。

用到的常规语法信息是动词的价、动词是否必带主语和宾语、主语和宾语的性质以及结构等。打分方法如下：

$$\text{句子结构总分} = \sum(\text{一个动词的分数})$$

$$\text{RRF} = \frac{(\text{出现的 OBR 规则数目}) * k + (\text{出现的 OPR 规则数目})}{k * (\text{OBR 规则数目}) + (\text{OPR 规则数目})}$$

$$\text{RWR} = \frac{\text{在短语中附有规则的词数}}{\text{短语中的总词数}}$$

$$\text{一个动词的分数} = \text{RRF} * \text{RWR}$$

式中：OBR——必备规则

OPR——可选规则

k——OBR 规则相对于 OPR 规则的显著性，式中取 2。

这就是说某一个动词的分数是 RRF 与 RWR 的乘积，一个句子的总分是各个动词的分数之和。例如，在下面的句子中：

“请求 与 被告 离婚”。

“离婚”可以是动词也可以是名词，“请求”和“离婚”都可以是主动词，那么究竟该如何判断呢？根据字典信息和词数反复假设和计算后，确定出当“离婚”取动词，且“请求”取主动词时的句子结构总分最高。

作者在 SinoTrans 中是用语法规则判断主动词的。过程很简单。因为“请求”的宾语可以是一个从句(最简形式为一个动词)，而且“离婚”的主语不能是动词，所以“请求”是主动词。

复杂句子,在语法的基础上有时也能确定出主动词。例如,作者对于

"我们 未能完成 论文 完全是 他 在 学校 无理干涉 的 结果"

的分析方法是"完成"不能带从句,"干涉"的右邻为"的",所以,"是"为主动词。

1.2.2 基于短语结构语法和格语法[6]

机器翻译界有两大相辅相成的理论支柱。一个是短语结构语法(二分法),用于分析句子的结构;一个是格语法,用于说明词、词组或成分间的语义关系。短语结构语法来自 Chomsky 于 1959 年提出的一个自然语言模型,表示为:

$$G = (V_n, V_t, P, S)$$

式中:V_n——非终结符(代表语法成分的符号);

V_t——终结符(词汇集合);

P——规则集合;

S——V_n中的一个起始符号(一般用 S 代表,表示句子)。

规则集为产生式(一般在一千个以下),例如:

$$S \leftarrow NP,\ VP \qquad VP \leftarrow V,\ NP$$

$$VP \leftarrow PP,\ VP \qquad PP \leftarrow P,\ NP$$

然后再利用规则分析一个句子,例如"学生在学校上课":

```
学生   在   学校   上   课
------ --- ------ --- ---
 NP    P     N     V    N
       --------    -------
         PP          VP
       ----------------
 NP            VP
 ----------------------
          S (成功)
```

通常在机内用依存关系表示,即以动词为中心,并用格名标注: