国家自然科学基金项目成果
湖北省重点学科建设立项学科成果
中国博士后基金项目成果

特征结构及其汉语语义资源建设

Building a Chine[redacted]source Based on Featur[redacted]

陈波 著

WUHAN UNIVERSITY PRESS
武汉大学出版社

国家自然科学基金项目成果
湖北省重点学科建设立项学科成果
中国博士后基金项目成果

特征结构及其汉语语义资源建设

Building a Chinese Semantic Resource Based on Feature Structure

陈波 著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

特征结构及其汉语语义资源建设/陈波著. —武汉：武汉大学出版社，2014.2

ISBN 978-7-307-12809-5

Ⅰ.特… Ⅱ.陈… Ⅲ.汉语—语义—研究 Ⅳ.H13

中国版本图书馆 CIP 数据核字(2014)第 021611 号

责任编辑：胡 艳 责任校对：鄢春梅 版式设计：马 佳

出版发行：武汉大学出版社 (430072 武昌 珞珈山)
(电子邮件：cbs22@whu.edu.cn 网址：www.wdp.com.cn)
印刷：湖北民政印刷厂
开本：720×1000 1/16 印张：13 字数：185 千字 插页：2
版次：2014 年 2 月第 1 版 2014 年 2 月第 1 次印刷
ISBN 978-7-307-12809-5 定价：27.00 元

作者简介

陈波，女，38岁，湖北枣阳人，文学博士、副教授，湖北文理学院文学院语言信息分析中心主任，现于武汉大学计算机学院软件工程国家重点实验室博士后流动站从事科研工作。1997年本科毕业于湖北师范学院汉语言文学专业，2004年硕士毕业于武汉大学语言学及应用语言学专业，2011年博士毕业于武汉大学语言信息处理专业。主要研究自然语言处理领域，具体从事语料库建设、舆情监控、信息检索、软件开发等研究工作。中文信息处理学会会员，ACL（世界计算语言学学会）会员，世界汉语教学学会会员，湖北省语言学学会会员。2008—2010年赴美国山姆休斯敦大学从事教学工作。

主持国家自然科学青年基金项目“汉语关联结构的资源建设和自动分析模型研究”（2013—2015）。2013年荣获中国博士后科学基金一等资助“基于语义依存的汉语关联结构分析研究”。主持湖北省教育厅人文社科项目“基于依存语法的语料库标注研究”（2008—2011）。参与多个国家自科项目、国家社科重大项目以及企业合作项目的研究。发表学术论文24篇，其中EI、ISTP、CSCD、CSSCI检索论文10篇。

前　言

汉语语义分析，特别是大规模真实文本的语义分析，一直是当前自然语言处理的难点。传统依存分析法等标注方法在处理汉语特殊句型和特殊语言现象（如主谓谓语句、连动句等句型）时遇到一系列难题。基于语义方法建构的标注语料库，是自然语言处理基础研究和应用技术研究的基础。

本书为汉语提出了一个语义分析模型——特征结构，并运用特征结构模型分析了汉语语言学界争议较大的特殊句式（主谓谓语句、连动句等）。研究结果表明，特征结构模型在分析汉语语句时，比传统依存分析法包含更多的语义关联，并能在语言学理论研究范围内解释汉语特殊句型的范围、类型及其特点等疑难问题。本书为中文信息处理提供了一种语义分析方法，也提供了一个可以为各大研究机构共享的汉语句子级语义资源，同时也为汉语语言学理论中某些问题的解释提供了一个别样的视角。

全书共分为6章，主要内容及观点如下：

第1章：引论。主要包括课题研究背景、国内外研究现状分析、研究对象界定、研究内容等内容。

第2章：特征结构模型。主要包括特征结构模型的界定，特征结构的特点：用特征三元组反映概念关联和关联种类，特征三元组允许多重关联和交叉关联，特征三元组允许嵌套和递归。特征结构的形式化表示为特征结构图，是一个可递归的无向图。特征结构的判定方法是基于提问的方法，研究了各种句式中提问的条件、提问针对的成分以及特征词在其中的分布等。在大规模真实语料中，特征结构三元组的类型可以分为六类。

第3章：汉语特征结构资源建设。本语料库的语料来源于美国

宾州中文树库的生语料、国内近3年中文新闻语料以及中小学语文课本。标注方式采用人工标注和计算机标注软件相结合的方法。设计并编写了汉语语义资源标注软件“语言标注平台”。研究了特征结构的判定标准。本章重点在于提出了详细的特征结构标注标准。

第4章：主谓谓语句特征结构研究。首先回顾了语言学界对主谓谓语句的研究成果和争论内容，根据汉语主谓谓语句的语法特点，分析了面向自然语言处理时的标注难点，然后运用特征结构模型对语言学界讨论过的13种类型的主谓谓语句进行了细致的语义描述和分析，总结出了6种语义模型。将现有的传统依存分析方法和特征结构分析方法对主谓谓语句的分析结果进行了对比，结果表明，特征结构分析方法包含了更多的语义信息。

第5章：连动句特征结构研究。首先回顾了语言学界对连动句的研究成果和争论焦点，总结了汉语连动句的语法特点，然后分析了面向自然语言处理时连动句的标注难点。运用特征结构模型对语言学界讨论较多的16个连动句分别进行了细致的语义描述和分析，总结出了4类语义模型。将现有的传统依存分析方法和特征结构分析方法对连动句的分析结果进行了对比，结果表明，传统依存语法无法表示连动句中主语和除第一个谓语动词之外的其他谓语动词之间的语义关系，无法表示连动句中某个谓语动词的宾语与其他谓语动词之间的语义关系，也无法准确表示两个或多个谓语动词之间的语义关系。与传统依存分析法相比，特征结构模型能够描述更多的语义关系对，因此包含更加丰富的语义信息。另外，特征结构模型能够对传统依存分析法不能解释的语言现象做出解释，比如对连动句句式的判定、对连动句和紧缩复句的区分、对复杂的杂糅句式的语义分析等。特征结构模型在一定程度上推进了语言学理论的深化和发展，也为面向汉语的自然语言处理提供了一种新颖的语义分析方法。

第6章：总结。包括评估、研究特色、应用价值、下一步研究计划等内容。

本书主要创新点在以下三方面：

（1）提出特征结构模型，探讨汉语语句的语义表示机制。

（2）基于特征结构模型，对汉语语句进行语义标注，探寻适合汉语独特特点的语义分析方法和标注标准。

（3）运用特征结构模型探讨了汉语特殊句型的语义分析方案，并尝试以新的视角来解释语言学理论中的争议问题。

目　录

第1章 引 论

本书的研究内容主要包括三个方面：

(1)理论上，探索语义分析理论，提出并完善基于特征结构的描写机制，并据此确立适合汉语语义描写的理论框架。

(2)实践上，基于特征结构模型建构大规模的汉语句子级语义标注资源。

(3)应用上，将特征结构模型应用于语言学分析，研究汉语特殊句型和特殊现象背后的语言规律和语义表示手段。

语义分析是自然语言处理的重点和难点。目前，国内外流行的语义标注方法是依存结构语法，本书在此基础上采用了特征结构理论，在理论的探寻上是一次积极的尝试。汉语的独特特点是注重意合、语序灵活，存在大量的特殊句型和特殊句式，给自然语言处理带来了很多困难。如何完整地表示出汉语句子中的词语与词语之间的语义关系，并能够面向自然语言处理，将这些语义关系形式化，是本书探讨的重点内容。带有语义信息的语料库是计算机自动分析的基础。在实践上，本书尝试运用特征结构方法标注汉语语句，以期为自动分析提供一个具有丰富语义信息的资源。在语言学理论研究领域，特征结构也尝试分析某些语言现象，并解释语言学层面的一些难题。

1.1 研究背景

1.1.1 语义分析是自然语言处理的重要内容

21 世纪科技进步的标志之一就是自然语言处理技术的突飞猛

进。IBM 的超级电脑“沃森”(Watson)在美国智力竞赛节目中击败人类，利用的就是其背后的海量语言知识库，包括《辞海》和《世界图书百科全书》等数百万份资源。IBM 早期研发的“深蓝”是在有限规则内快速深度计算，并没有运用大量知识；沃森比深蓝优越之处则是在算法之外利用了大量的知识库，因此能够分析具有模糊性和歧义性的人类语言①。自然语言处理技术的提升，迫切要求计算机能够更精确地理解自然语言。

一般说来，自然语言处理(Natural Language Processing，NLP)的研究内容包括三大层面：资源建设层面、基础研究层面、应用技术研究层面，其中，资源建设是基础研究和应用技术研究的基础，一个具有丰富信息的资源库将大大提高应用技术的结果。

近年来，自然语言处理的主要任务之一就是建设大规模面向多语种的语言资源(Language Resource Construction)。在语言资源建设中，语料库标注的发展过程依次为：语法标注阶段、句法标注阶段、语义标注阶段和话语标注阶段(陈波，2007)，如图 1.1 所示。

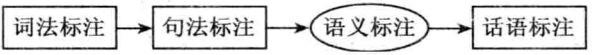

图 1.1 语料库标注四个阶段

迄今为止，词法标注阶段的工作已基本解决，句法标注阶段工作也已基本完成，但是语义标注阶段则成为瓶颈。如果语义问题不能得到很好的解决，就无法解决语言的模糊性和歧义性所带来的问题，就无法很好地解决计算机领域内语言理解、机器翻译(Machine Translation，MT)等难题。如何完整、清晰地表示句子中词语和词语之间的语义关系，如何形式化地表示这些语义关系，这是自然语言处理必须解决的关键性问题。

语义分析(Semantic Analysis)，是指根据句子的句法结构和句中每个实词的词义，推导出能够反映这个句子意义的某种形式

① http://news.sohu.com/20110218/n279398857.shtml

化表示。

例 1：小陈打破了杯子。
例 2：杯子被小陈打破了。

虽然例 1、例 2 的表述形式不同，前者是一般的主谓句，后者是被动句，但表示成语义的形式都是：

打破(小陈，杯子)

自然语言处理所要解决的主要问题一直是如何对句子进行正确的语义分析。近几十年来，还没有出现很多获取详细语义理解知识的研究，基于语料库的词义排歧研究虽然涉及了语义问题，但是仅在理解单个词的层面上，而非针对整个句子的理解。关于信息抽取的研究，也涉及了一些语义理解的问题，但是现有系统只是使用句法模式分析的方法来获取某些目标短语。

1.1.2 汉语语句难点举例

自然语言处理中，汉语的语义分析一直是难点。汉语的特殊句型和特殊语言现象，很难完整地分析出句子中的语义信息，如下列语句：

例 3：A：小王肚子笑痛了。(灵活语序)
　　B：小王的肚子笑痛了。
　　C：小王笑痛了肚子。
例 4：他理了三次发。(离合词)
例 5：咱俩谁也别忘了谁。(主谓谓语句)
例 6：我开车去车站接他。(连动句)
例 7：我请你写一篇文章。(兼语句)
例 8：上海经济开发与法制建设(复杂名词短语)

对于汉语来说，汉语具有与英语、法语不同的语言特点：语序比较灵活；虚词有着重要地位；特殊句型，如主谓谓语句、兼语

句、连动句等，以及倒装句、紧缩句、复杂名词短语等语法结构大量存在，计算机处理起来更具难度，面向中文信息处理的语义标注研究工作迫在眉睫。

1.1.3 研究意义

本书旨在创建和研究一种新的适合汉语语义分析的标注模型——特征结构；并基于特征结构模型，针对汉语的句子级的大规模真实语料，建构一个汉语语义资源，研究有效的汉语语义标注策略；在此基础之上，尝试将特征结构的分析方法和语义资源库应用到自然语言处理的各个领域中。

本书的研究意义在于：

(1)从研究对象的选择上看，一方面，特征结构是面向自然语言处理的语义描述模型的新探索。语义分析一直是信息处理的难点，其中以汉语特殊句型的语义研究尤甚。本书的理论研究对于不同语言的语义分析难题的解决具有普遍理论意义；另一方面，大规模汉语语义标注资源的建设是自然语言处理领域机器翻译、信息抽取、信息检索等飞速发展的首要任务。本书的资源建设具有工程实践意义。

(2)从基础理论的探求上看，本书提出的特征结构理论是一个全新的语义表示模型，既不同于传统的句法分析和依存分析，也不同于后来的概念依存理论和概念结构理论。本书还通过特征结构理论，积极探求汉语特殊语言现象背后的语言规律和描述机制。对于语义分析理论和语言学理论，本书的研究均有一定理论探索意义和开创意义。

(3)从汉语资源的建设上看，本书建构的大规模汉语句子级语义标注资源库，标注理论和描述机制不同于已有的汉语资源，在标注规模和标注深度方面都有所提高，本书的标注成果在一定程度上丰富了汉语语义资源，为中文信息处理的发展做出了一定贡献。

(4)从研究的应用价值上看，本书的编写立足于自然语言处理的语义分析需求，提出的特征结构语义描述模型和建构的大规模语义资源库对于提高语义关系抽取(特别是事件关系抽取)、自动问

答、信息检索、文本蕴涵等系统的性能会有一定帮助。同时，本书的研究对于语义理论的研究也有所帮助。

1.2 国内外研究现状及分析

1.2.1 关于语言分析模型

语言分析的结构模型理论中比较有代表性的有：句法结构理论、依存结构理论、概念结构理论和概念依存理论等。

1. 句法结构(Syntax Structure)

短语结构语法(Phrase Structure Grammar，PSG)是一种著名的语言形式描述理论，由美国语言学家乔姆斯基(Noam Chomsky)在20世纪50年代提出(Chomsky，1956；Chomsky and Miller，1958；Chomsky，1959；Chomsky，1963；Chomsky，1969)，其特点是用有限的规则来描述形式上是潜在的无限的句子。作为当代计算机科学的基础理论之一，短语结构语法在机器翻译、算法分析、编译技术、图像识别、人工智能等领域中得到了广泛的应用(冯志伟，2004)。在自然语言处理界，短语结构语法是自然语言处理的主流分析方法，自20世纪60年代至80年代一直处于主宰地位，至今仍是自然语言处理的重要方法之一。

短语结构语法可以形式化地描写语言，用非终极符号(即范畴符号，如词类符号Noun、Verb、词组符号NP、VP、AP、PP等)和终极符号(单词)来表示无限变化的语句，它能够详细地描述句子的句法信息，这些标注信息为下一步的语言分析和机器自动学习的发展提供了基础。如图1.2所示。

例9：I love the red pen.

运用短语结构语法，陈述："I love the red pen."可以表示为图1.2所示的一个形式化的树形图(Tree)，简称为树，其中，S表

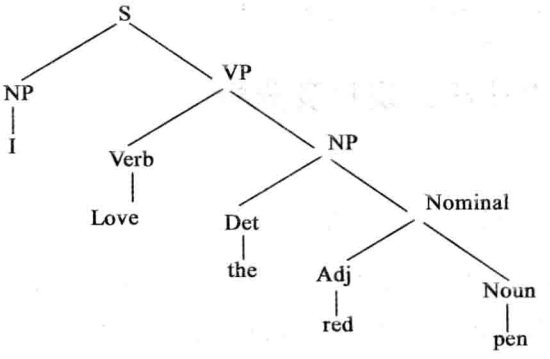

图 1.2　例 9 短语结构树

示句子，NP、VP 分别表示名词词组、动词词组，N、V、A 表示名词、动词和形容词。这个短语结构树包括了该句丰富的句法信息，并且树的终极节点是词语。除了树形图，该句也可以表示为下面的形式：

[S[NP[Pro I]][VP[V love][NP[Det the][Nom[Adj red][N pen]]]]]

概括来说，陈述句“I love the red pen.”被分析为一个名词短语 NP(Noun Phrase)和一个动词短语 VP(Verb Phrase)，结构简化为：

S→NP　VP

相应的，英语中的命令句、是非疑问句、特指疑问句等都可以形式化地表示为：

S→VP(命令句)

S→Aux　NP　VP(是非疑问句)

S→Wh-NP　VP(主语特指疑问句)

S→Wh-NP　Aux　NP　VP(非主语特指疑问句)

又如以下汉语的例子：

例 10：中国建筑业对外开放呈现新格局。

可形式化地表示为：

```
((IP-HLN (IP-SBJ (NP-SBJ (NP-PN(NR 中国))
                         (NP(NN 建筑业)))
                 (VP (PP (P 对)
                         (NP(NN 外)))
                     (VP(VV 开放))))
         (VP (VV 呈现)
             (NP-OBJ (ADJP(JJ 新))
                     (NP(NN 格局))))))
```

这是宾州中文树库(Penn Chinese Treebank，CTB)样例中的一个句子(Treebank)，运用了短语结构语法，标注结果包括丰富的语言信息：NR、NN、CC、VV 是各个词的词性，分别为专有名词、普通名词、连词和动词；NP、VP 是名词短语和动词短语，PP 是介词短语，ADJP 是形容词短语；SBJ 是句子中的主语(Xue and Xia et al.，2005)。

短语结构语法在形式化上具有很强的描述性、概括性，易于在计算机上实现形式化表达，得到的结构层次清晰。只要写出相应的规则，就可以制定相应的句法分析器(Parser)，应用到自然语言分析中。因此，自短语结构语法产生至今，它一直是自然语言处理界的最主要的语言分析方法之一。另外，短语结构语法的研究开始很早，迄今为止，语言学界对短语结构语法有了很深入的研究，研究成果很丰富。自然语言处理在采用短语结构语法的方法建构语料库时，也借鉴了语言学界的研究成果，编写了相关的句法分析器(Syntactic Parser)和算法，目前已经建设了很多大型、成熟的带有句法信息的树库(王跃龙，姬东鸿，2009)。

但是，短语结构语法作为一种形式语法，在生成能力上，其生成能力与对其规则所受的限制有关系，规则所受的限制越多，生成能力越弱；规则所受的限制越少，生成能力越强。如果语义太复杂，则会写出太多的规则，但是总会有例外，太多的规则又限制语义的生成。短语结构语法强调句子的形式规则，而自然语言的语义复杂多变，歧义句等语义千变万化，如果仅描述语言表层的结构形

式，无法揭示深层的语义、句子中词与词的语义或者整句的语义，如例11和例12是语言学中经典的两个例句。

例11：台上坐着主席团。
例12：台上演着戏。

运用短语结构语法的分析结果都是：
[N+V+着+N]
可以发现，例11和例12深层的语义关系是不同的，但是运用短语结构语法进行句法标注无法表示出两者的区别。再如例13是语言学中的经典歧义句。

例13：鸡不吃了。

短语结构语法的分析只能得出[N+V]的结果，句法上的描述无法分析出例13的深层语义。近年来，NLP发现很多语义上的问题，仅用规则很难描述清楚，在语义分析方面也很难突破，因此在20世纪80年代后开始转向对另外一种方法——依存语法的研究。同时，对于已经建设完成的句法树库，也开始由短语结构树向依存结构树进行转换，其相关的转换规则和算法也是近年来的研究重点（陈波，2007）。

2. 依存结构（Dependency Structure）

自然语言处理的主流分析方法从20世纪60年代到80年代中期，以短语结构语法为代表，在最近20年里，学者们才开始重新考虑用依存语法（Dependency Grammar，DG）来解决自然语言处理中的许多问题。

依存分析方法是基于语义的分析方法，近年来已经成为自然语言处理的一种主流方法。目前，国外对依存语法标注的研究主要集中在如理论建构、算法设计、模型建构等方面。整体状况是对依存语法的研究越来越深入，研究方法多样化，并在不断创新。基于依存语法进行了很多语言分析方面的研究。