教育部人文社会科学研究规划基金项目“面向精确语言理解的中文介词短语语义角色识别技术研究”（12YJA740095）资助

面向中文信息处理的框架语义分析

由丽萍／著

MianXiang ZhongWen XinXi ChuLi De KuangJia YuYi FenXi

经济科学出版社
Economic Science Press

本成果得到教育部人文社会科学研究规划基金项目“[illegible]介词短语语义角色识别技术研究”（12YJA740095）和山西大学出版基金资助

面向中文信息处理的框架语义分析

由丽萍 著

经济科学出版社

图书在版编目（CIP）数据

面向中文信息处理的框架语义分析/由丽萍著.
—北京：经济科学出版社，2013.4
ISBN 978-7-5141-3448-3

Ⅰ.①面… Ⅱ.①由… Ⅲ.①汉字信息处理-语义分析 Ⅳ.①TP391.12

中国版本图书馆 CIP 数据核字（2013）第 110713 号

责任编辑：程晓云
责任校对：杨 海
版式设计：齐 杰
责任印制：王世伟

面向中文信息处理的框架语义分析
由丽萍 著
经济科学出版社出版、发行 新华书店经销
社址：北京市海淀区阜成路甲 28 号 邮编：100142
总编部电话：010-88191217 发行部电话：010-88191522
网址：www.esp.com.cn
电子邮件：esp@esp.com.cn
天猫网店：经济科学出版社旗舰店
网址：http://jjkxcbs.tmall.com
北京欣舒印务有限公司印装
880×1230 32 开 6.5 印张 200000 字
2013 年 4 月第 1 版 2013 年 4 月第 1 次印刷
ISBN 978-7-5141-3448-3 定价：20.00 元
（图书出现印装问题，本社负责调换。电话：010-88191502）

序

从20世纪50年代的机器翻译和人工智能研究算起，自然语言处理作为一个应用研究方向已有长达60多年的历史。在这个进程中，学术界曾提出许多重要的理论和方法，不断提高自然语言的处理能力和处理深度，取得了丰硕的研究成果。现阶段，自然语言处理的研究重点从对语言形式的处理转移到了意义分析。对语句意义的自动处理，前提条件是要有一个词汇语义知识库。然而，词汇到底有哪些语义属性，对意义如何分类，这种分类应该细化到什么程度，语义描述应该到什么深度，均难以给出确定的答案；就具体的某个词语而言，不同的人有不同的意义理解，那么，又如何保证语义描述的客观性和正确性？语义本身具有不确定性和模糊性，同时，对语言中的所有词语一一进行语义描述和归类，又需要付出艰辛的劳动，需要常年的坚持和努力，这给语义知识库和语义分析体系的研究带来了巨大的挑战。我们经过数年调研与试验，从2004年起以菲尔墨（Fillmore）的框架语义理论为基础，开始构建汉语框架语义知识库，主要包括词汇的框架语义描述和语义标注语料库建设。该工程至今已走过了近十年的历程，它以分类体系细化、语义信息丰富受到了业内专家的肯定。汉语框架语义知识库目前仍然是一个在建工程，一方面对已建框架不断进行修订、补充和完善，同时继续不断扩大语义领域，构建新的框架，标注新的句子。

我们的科研得益于学科交叉互补，团队中纳入了计算机软件、现代汉语、统计数学、信息管理等不同学科的师生，他们付出了卓

绝的努力。特别是自然语言处理要解决的是语言本身的问题，而不是单纯的程序设计问题，因此，单靠若干程序肯定是无法提高质量的。由丽萍博士作为最早参与该工程的研究人员，研究了框架语义理论、语义描述原则和构建路线等，并主持构建了认知语义领域框架，成为整体工程的先行探索和示范样本。

该书面向中文信息处理应用需要，提出语义信息的处理应该向精确化方向发展，要求语义范畴知识体系应该分类细化、语义层级深化，表达更多的语义内容，以满足处理复杂问题和模糊问题的需要。进而从理论和实践两个层面对汉语框架语义分析做了系统研究，既包括语义范畴知识体系的构建方法，又包括句子语义角色识别技术，完整地展现了汉语框架语义分析的各步工作内容。还针对以往研究对句子的附属成分重视不足、无法满足精确化语义理解需求的问题，研究了汉语介词短语的句法和语义处理问题，并通过一个信息抽取实验展现了这种研究对精确化语义理解的价值。

这本书观点新颖、内容丰富、方法可行、结论可靠，尤其是提出了当前精确化的语义理解需求，观点引人思索，对中文信息处理的发展做出了积极贡献，非常具有参考价值。

刘开瑛

2013年3月于山西大学

前　言

计算机语义分析是21世纪以来自然语言处理的研究重点之一，目前在主干语义信息以及浅层语义信息的处理方面，已经积累了较为丰富的成果。然而，多数语义表示方法和处理结果却在追求体系的简洁性和处理的可行性的同时，丢失了语言表达中深层次的、复杂的语义信息，导致研究结果无法达到现实社会的应用需要。菲尔墨（Fillmore）在其格语法基础上进一步提出了框架语义理论，将词语、句子乃至篇章的语义内容统一用图式化的认知情境——“框架”加以描述，以细化的语义范畴表示更丰富的语义内容，使得语义分析结果更加精确。

本书即引进、吸收框架语义理论的研究成果，面向中文信息处理应用需要，研究汉语框架语义分析问题。一方面，确立汉语词汇的框架语义知识库构建方法、句子的框架语义标注规范；另一方面，以语义标注语料为基础，根据短语类型、句法功能以及短语内、外部其他句法语义特征，归纳汉语句子中框架元素的识别规则，并通过实验验证规则的有效性。具体研究内容包括：

（1）研究汉语框架语义知识库构建技术。从语料准备、数据库结构设计和软件需求等方面确立汉语框架语义知识库的构建路线；根据框架语义理论，制定框架编写原则；借助语义Web描述语言，用本体（Ontology）组织框架语义知识库内容，使之成为机器可读、可理解的词汇语义资源。

（2）构建框架语义知识库样本——认知领域框架及框架元素系

统。从语料实际出发，用框架语义分析方法，描述认知领域动词的语义内容，定义完整的框架、框架元素及框架—框架关系体系，为全面构建语义知识库提供可行的研究路线和具体的构建样本，同时，这些研究结果本身也可作为计算机分析现代汉语语义的重要资源。

（3）设计汉语框架语义标注的方法体系。设计汉语框架语义句子标注规范体系，包括短语类型标注体系、句法功能标注体系和框架元素标注体系，明确特殊句式、特殊表达形式的处理方法，为汉语框架语义分析提供语料加工规范。

（4）确立框架元素标注规则的获取技术和描述体系。用语料统计数据详细分析短语类型与框架元素的关系、句法功能与框架元素的关系以及短语内、外部其他形式和意义特征与框架元素的关系，分析框架元素与句法表现形式之间的对应规律，在此基础上，明确框架元素标注规则的构建路线和描述体系。

（5）构建认知语义领域汉语框架元素识别规则。以认知领域为例，归纳框架元素识别规则，对其进行形式化描述，并应用到计算机自动语义标注实践，通过实践检验规则的性能，给出量化的评价结果。

（6）研究汉语介词短语的边界划分及其框架元素识别技术。采取基于搭配模板匹配的方法和基于词性边界统计模型的方法划分短语边界，再采取基于规则的方法对其框架元素进行自动识别，最后通过语料标注实验证明方法的有效性。

（7）研究基于汉语框架语义本体的信息抽取。为显示汉语框架语义分析的应用价值，以时间语义为例，进行信息抽取技术研究和实验。

本书的研究特色可以归纳为：

（1）以框架语义理论为基础，构建汉语语义分析的范畴知识体系，其语义角色细化，语义信息丰富，使得计算机语义理解更加精确化。

(2) 在范畴知识体系、标注方法体系和标注规则体系的构建上，始终采用基于语料库的方法，避免“从意义到意义”的理性思维，使研究结果可靠、实用。

(3) 用基于规则的方法解决语义标注歧义问题，避免了基于统计的方法面临的语义资源匮乏、数据稀疏的问题，通过实验证明本研究所归纳的规则对于解决语义角色标注问题十分有效。

本研究工作是跨现代汉语语义和中文信息处理两个领域进行的，无论是在对语义理论和前人研究工作的引进、吸收，还是在语义知识范畴体系和规则体系的构建上，都力求探讨汉语框架语义分析的各方面问题，其研究结果，一方面，对推进中文信息处理技术的发展有直接的应用和参考价值；另一方面，也可以为语言研究工作提供丰富的研究素材。

目　录

第1章

面向信息处理的语义分析

1.1 引　言

计算机发明之初，只能按照人们预先为它编制好的程序进行操作，人类进入信息时代，希望计算机更加“聪明”，跟人一样在新情况中做出恰当反应，因此，计算机、数学、语言学等相关领域的科学家开始对人工智能进行研究。当一个计算机系统能给出有关问题的正确答案或有用建议，而解决问题所用的概念和推理跟人相当，还能解释推理过程时，就可以说这样的计算机系统是有智能的（石纯一、黄昌宁和王家廞，1993）。对自然语言（而非计算机程序语言等人工语言）的信息处理，是人工智能中极其活跃的研究领域，是开发智能计算机必须完成的重要研究课题。《国家中长期科学和技术发展规划纲要（2006～2020年）》在信息技术“前沿技术”中即包括“智能感知技术”，明确提出“重点研究基于生物特征、以自然语言和动态图像的理解为基础的‘以人为中心’的智能信息处理和控制技术，中文信息处理”。

如果按照语言分析的三个平面划分，对自然语言的处理包括句法分析、语义分析和语用分析三个层面。其中，中文信息处理的句法分析还包括词语切分和词性标注等，它们与短语结构分析共同构

成对语言形式的处理，这是前一阶段研究的重点，目前的信息检索、机器翻译等应用系统也主要以此为基础。以人工智能为背景，对自然语言的处理就不能满足于对语言形式的处理，只有深入语义乃至语用层面，才有可能使语言信息的处理带有“智能”。因此，目前自然语言处理的重点已经从句法方面转移到语义方面，而语用层面的研究还很少，仍以解决指称指代等问题为主。

语义分析要解决的问题主要有两个方面：一是基于概念表达的语义聚合关系的分析，包括对真实文本中多义词的词义排歧，对句中词语进行同义或同类等意义关系的扩展等；二是基于语义角色标注的语义组合关系的分析，即分析述谓结构中谓词和体词性成分之间的语义关系等。

1.2 面向信息处理的语义分析研究现状

无论是语义聚合关系的分析还是语义组合关系的分析，都需要一个描述这些语义内容的词汇语义范畴知识体系；而对于语义角色标注来说，还需要建立一个语义组合关系分析模式，下面我们就从这两方面评述目前的研究状况。

1.2.1 词汇语义知识库研究

我们首先分析各类型语义知识对语言信息处理的作用，然后考查现有的英语、汉语词汇语义知识库对这些知识需求的满足程度如何，以期在词汇语义范畴知识体系的选择问题上，为中文信息处理研究提供参考。

1. 面向信息处理的语义关系分类

（1）语义聚合关系。

语义聚合关系的分析是以义项为单位，把具有不同概念基础的词语分归不同的类别，各类之内的词语具有同义关系或反义关系，

类与类之间具有上下位关系或整体—部分关系等。

将词语的意义划分为不同的义项，就能使计算机识别出以下每组句子语义的差别：

她在看窗户上的树影儿。（用眼睛感知外界事物）

我看那孩子不会学好了。（经过观察，认为要出现某种趋势）

描述义项之间的同义、反义关系，就能使计算机识别出以下三句语义等价：

农民们几乎忘记了养鸡的手艺。

农民们几乎遗忘了养鸡的手艺。

农民们几乎不记得养鸡的手艺了。

再看以下三句：

小王今天坐出租车上班。

小王今天坐公共汽车上班。

小王今天上班是靠车窗坐的。

如果希望计算机能正确回答“小王今天坐车上班还是步行上班?”这个问题，就需要具备抽象的概念关系知识。因为这三个句子中并没有“车”这个词，而是“出租车”、“公共汽车”、“车窗”三个词。计算机必须知道“车”与“出租车”和“公共汽车”具有上下位关系，与“车窗”具有整体—部分关系，才能将问题句与这些答案句联系在一起。

(2) 语义组合关系。

语义组合关系主要描述谓词性词语与体词性成分的语义结合性质，即语义角色关系或题元角色关系。有了语义组合关系知识，计算机就能识别出以下两句意义等价：

田丽把老师的名字忘了。

田丽忘记了老师的名字。

或识别出以下两句宾语的作用不同：

吃米饭。

吃食堂。

(3) 事件推理知识。

事件推理知识是关于事件发展过程的描述，尤其是由动作导致状态，再由状态到下一动作之间的因果链知识，有的研究称之为“脚本知识”（Schank，1975），它们为计算机推理提供分析依据。例如动词“买”属于商品交易活动，它使得商品和金钱的所有权关系发生变化，也就是说动作发生之后，商品从卖者转移到了买者手中，金钱则从买者转移到卖者手中。如果词汇语义知识库提供了这样的语义内容，计算机就可以从下面的句子中推导出关于所有权关系的信息：

张丽用十元钱从王斌那儿买了一本书。

计算机通过推理，可以知道买者“张丽”放弃了“十元钱”，拥有了“一本书”；卖者“王斌”放弃了“一本书”，拥有了“十元钱”，使得新的所有权关系得以确立。根据这些语义分析，一个应用系统就有可能正确回答类似“某某书现在是谁的?”的问题。

再如，对于这样的句子：

张丽去北京了。

计算机可以根据事件和状态之间的因果关系知识，推出“张丽”已不在当前城市，她可能现在在北京，或正在去北京的路上。

2. 词汇语义知识库构建现状

目前语言信息处理领域使用较多的词汇语义知识库包括英语的WordNet、FrameNet、VerbNet 以及汉英双语 HowNet、CCD 等，下面分别给予简要介绍。

(1) WordNet。

WordNet[①] 是由美国普林斯顿大学米勒（George A. Miller）领导，于 1985 年着手构建的英语词汇语义知识库（FeUbaum C.，

① 在线数据库网址为 http：//www. cogsci. princeton. edu/-wn/。

1998)。收录的词语包括名词、动词、形容词和副词，WorNet 将其组织为同义词集合，称为 synsets，每一个集合归为一个基本的词汇概念，并在这些概念类之间建立了同义关系、反义关系、上下位关系、整体—部分关系等多种聚合关系。最近公布的 WordNet 3.0 版本，包括 155 287 个词条，对应 206 941 个义项，其中名词义项 146 312 个、动词义项 25 047 个、形容词义项 30 002 个、副词义项 5 580 个，同义词集合有 117 659 个。

WordNet 词汇工程的理论基础是心理语言学，其出发点并不是为了计算机处理，而是为了探讨和研究语言产生、理解的心理活动机制，因此，它关注的是语言能力的认知基础，米勒称之为心理词汇学。该工程从语言生成和理解的心理机制出发，认为在人的语义记忆中包含词汇意义及概念之间的关系，但是，这些关系的组织方式与一般词典用已知的词去定义和解释一个生词又有所不同，词义的心理表征比一般词典的词义表示要复杂得多。

WordNet 用同义词集的方式来组织体系结构，比如对名词 car 的注释，描述了多种聚合关系：

car

同义关系

auto, automobile, machine, motorcar

下位关系

ambulance

beach wagon, station wagon, wagon, estate car, beach waggon, station waggon, waggon

bus, jalopy, heap

.....

部分关系

accelerator, accelerator pedal, gas pedal, gas, throttle, gun

air bag

auto accessory
automobile engine
……

上位关系
motor vehicle, automotive vehicle
self -propelled vehicle
wheeled vehicle
vehicle
conveyance, transport
……

从 WordNet 对 car 的描述我们可以看到，WordNet 语义聚合关系的描述非常详细、信息丰富。它对动词的描述也是从概念意义入手，描述其语义聚合关系，如将“buy”与“purchase”归入一个同义词集合，将其描述为“get”、“acquire”的下位等。但它没有描述动词的语义组合性质，不提供语义角色信息。

WordNet 从 1985 年开始至今已经经历了 28 年，所收录的词汇数量大，完备程度高，因此，被普遍认为是用于语言信息处理的非常实用的资源，已被成功地用于词义消歧、机器翻译、检索系统、词义标注、基于概念的信息检索及信息抽取、文本校对、知识推理、概念建模等一系列应用工程，在国际自然语言处理领域已有相当的影响。

（2）FrameNet。

FrameNet[①] 是由 Fillmore 主持的一个基于语料库的计算词典编纂工程，从 1997 年开始于美国加州大学伯克利分校进行构建，项目经理是 Collin Baker（Baker, Fillmore and Lowe, 1998）。截至目前，共收录 12 713 个词元[②]，构建了 1 164 个框架，共标注了近 17

① 在线数据库网址为 http://framenet.icsi.berkeley.edu/。

② 词元，指一个义项下的一个词语，FrameNet 把一个多义词作为多个词元，分别在不同的框架中加以描述。