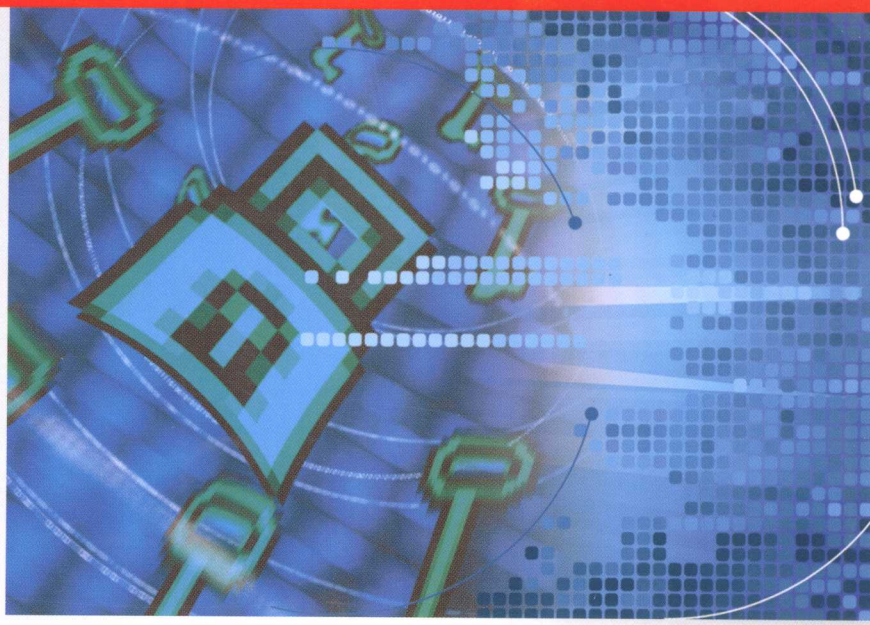
基于认知与计算的事件语义学研究

刘茂福　胡慧君　著



科学出版社

基于认知与计算的事件语义学研究

刘茂福　胡慧君　著

科学出版社
北京

内 容 简 介

本书主要从认知与计算角度介绍了有关事件语义学的相关内容。第1章简介研究背景、研究内容及研究方法等；第2章从认知与计算角度阐述事件语义的理论基础；第3章从认知角度探索事件语义结构认知；第4章从认知角度探讨事件语义关系分析，从计算角度讨论事件语义结构与事件语义关系的辅助标注；第5章从计算角度讨论事件语义形式化，主要包括事件逻辑、事件图等；第6章主要给出基于事件图的语义计算及其在自动摘要方面的应用。

本书可以作为计算机科学、计算语言学、应用语言学等研究生或相关研究者的参考书，对读者深入理解事件语义学很有帮助，也可作为高等院校语言信息智能处理方面的参考资料。

图书在版编目(CIP)数据

基于认知与计算的事件语义学研究/刘茂福，胡慧君著．—北京：科学出版社，2013

ISBN 978-7-03-037368-7

Ⅰ．基… Ⅱ．①刘… ②胡… Ⅲ．语义网络-研究 Ⅳ．TP18

中国版本图书馆CIP数据核字(2013)第086033号

责任编辑：魏英杰 杨向萍 / 责任校对：宣 慧

责任印制：张 倩 / 封面设计：陈 敬

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2013年4月第 一 版 开本：B5 720×1000

2013年4月第一次印刷 印张：13

字数：251 000

定价：50.00元

（如有印装质量问题，我社负责调换）

前　言

近年来，在理论语言学和应用语言学领域，事件语义学都有诸多研究与讨论。在计算语言学与自然语言处理领域，事件模型被用于自动摘要、问答系统、信息检索等，本书作者也在事件模型的应用领域进行过尝试与探索。事件模型符合人类的认知模式，即一切事物都在特定的时间和空间内不停地运动和变化。但理论语言学与应用语言学领域对事件语义的研究局限于句子特例与特殊句式，很少涉猎句子动态语义，更没有从段落和篇章理解角度来分析计算事件语义关系。在计算领域，目前事件模型只是在具体的应用问题中被提及，也没有系统化地对事件语义计算进行研究。因而，本书作者以事件语义学为背景，从认知科学和计算角度出发，对事件语义的分析与计算理论进行了深入探讨。

本书主要由六章内容组成，不但包括认知科学角度的事件语义结构与事件语义关系等，而且包含计算科学角度的事件语义形式表示与事件语义计算等，更有面向自然语言文本自动摘要方面的应用。第 1 章简介研究背景、研究内容及研究方法等，其中包含事件概念与事件语义。第 2 章从认知与计算角度阐述事件语义的理论基础，包括自然语言处理、计算语言学、篇章语言学以及事件语义学。第 3 章从认知角度探索事件语义结构分析，主要包括题元角色理论、事件语义角色、事件语义算子等。第 4 章从认知角度探讨事件语义关系分析，主要包括衔接与连贯理论、事件平行型关系、事件偏正型关系等，从计算角度讨论事件语义结构与事件语义关系的辅助标注。第 5 章从计算角度讨论事件语义形式化，主要包括事件逻辑表示、事件图形式化等。第 6 章主要是基于图算法的事件语义计算及其在自动摘要方面的应用。

本书是国家自然科学青年科学基金项目“面向自然语言文本生成的事件语义计算研究”（61100133）和国家社会科学基金重大项目“基于本体演化和事件结构的语义网模型研究”（11&ZD189）子课题“事件语

义自动标注研究”的阶段性研究成果，当然也包括了本书作者近几年在计算事件语义学领域的部分研究成果。

本书能够顺利完成写作，首先要感谢新加坡国立大学计算机学院的蔡达成(Chua Tat-Seng)教授及 NExT 搜索中心的所有同仁。本书作者在新加坡国立大学计算学院学术访问期间(2012 年 2 月至 2013 年 2 月)，蔡达成教授为本书作者提供了良好的学术环境与学习氛围，使作者能够利用新加坡国立大学图书馆的各种资源，有机会阅读与学习应用语言学和事件语义学等相关的各种英文原版论著。

作者要感谢武汉大学的何炎祥、萧国政、姬东鸿、吴泓缈等各位老师以及武汉大学语言与信息研究中心的各位同仁，每次的跨学科讨论会以及同各位同仁的研讨，使作者受益匪浅。作者要感谢所在武汉科技大学计算机学院的科研团队，包括黄智生、陈和平、顾进广、符海东等各位老师以及团队与实验室的各位同学。作者在参与国际和国内学术会议时，感谢就相关问题进行讨论的各位老师与同学。

还要感谢家人，没有家人的理解、支持与参与，作者无法继续坚持自己的科研兴趣与科研之路。同时要感谢父母，父母的大力帮助与无私付出，使作者能顺利完成本专著的撰写。

感谢本书的责任编辑魏英杰先生和科学出版社各位领导，感谢他们在书稿写作和出版过程中提供的各种形式的帮助和支持。

由于作者的学术功底和理论水平有限，书中难免有纰漏和缺陷，敬请各位专家学者给予批评指正。

作　者

2013 年 1 月

目　　录

第1章 绪 论

本书的研究对象是事件，严格意义上说应该是原子事件。为了使读者更好地理解本书所著述的内容，本章将简要讨论把事件作为研究对象的意义，简要介绍本书的研究内容与方法，重点研讨认知与计算角度的事件语义分析与计算。

1.1 事件概念

在自然语言处理领域，信息处理已由细粒度的语素、词以及短语等语言静态单位逐渐过渡到了较粗粒度的句子、段落和篇章等语言动态单位。语言静态单位与语言动态单位有着显著区别，语言静态单位可以被相对稳定地描述，如词对应的词典；而语言动态单位却无法详尽描述，只能针对语言动态单位所在的语料环境，进行相应情境下的相对详尽描述。例如，句子在不同的情境下，含义肯定不同，甚至句子的外在形式构成都会随着情境而动态改变。

近年来，事件研究在自然语言处理领域成为了热点。实际上，事件在很多语义理论中都很重要。事件模型符合人类的认知模式，即一切事物都在特定的时间和空间内不停地运动和变化，同时，事件提供了一种结构，可以把句子的解释组合起来。本书作者在研究事件语义计算及其应用的过程中，发现理论语言学领域对事件语义学已经有所涉猎，但基本上是分析句子特例或者特殊句式的事件语义，没有从粗粒度的段落和篇章的上下文语境来研究，更没有从计算角度来研究。而在计算领域，事件只是在具体的应用问题中被提及，没有系统化地对事件语义计算进行研究。因而，向研究者介绍事件语义分析与计算理论是非常必要的。

据上所述，本书的基本研究对象为原子事件，原子事件具有原子性、语义性和事实性等特点。原子事件的原子性实际上是指它为事件

语义分析计算的最基本信息处理单位和出发点，即事件事实语义的原子性，事件语义的决定要素是事件谓词的意义，如例 1.1 中虽然事件(2)比事件(1)多了事件成分“当地时间 5 月 6 日晚 8 点”和“奥朗德”的共指语义成分“社会党候选人”，但两个事件的事件谓词都是“当选”，并且在两个事件中事件谓词的语义相同，在不考虑语境的情况下，它们为同一个事件。

例 1.1(摘自网易新闻)

(1) 奥朗德当选新任法国总统。

(2) 当地时间 5 月 6 日晚 8 点，社会党候选人奥朗德当选新任法国总统。

原子事件的语义性是指原子事件是一个语义单位，其介于词和句子之间，比词组意义完整，通过事件语义分析与计算，可以透彻研究其所在句子的意义，进而研究所在篇章意义。除原子事件的语义性外，原子事件更是一个语言事实概念，只有在语料实例中，才有原子事件，离开语料实例，原子事件将失去其存在的意义，所以原子事件还具有事实性。

鉴于篇章自上而下的宏观逻辑结构方面的考虑，篇章一般由一到多个段落构成，段落又由一到多个衔接良好、意义连贯的句子组合而成，而一到多个小句则可构成句子。吕叔湘(1979)认为，小句是研究语言动态语义的基本单位。但本书作者认为，小句是语法单位，而事件是语义单位，二者没有对应性。研究句法，可以从小句出发，而要研究句义，则要多考虑句子里包含的原子事件以及事件之间的语义关系。

例 1.2(摘自网易新闻)

尼泊尔王国政府财政大臣巴德里·什雷斯塔 10 月 31 日晚在此间公布了一揽子经济改革措施，旨在促进国民经济发展，使尼泊尔尽快摆脱贫困。

例 1.2 的句子就包括了“尼泊尔王国政府财政大臣巴德里·什雷斯塔 10 月 31 日晚在此间公布了一揽子经济改革措施”、“旨在促进国民经济发展”和“使尼泊尔尽快摆脱贫困”三个小句，其中第二和第三小句省略了主事(或称为弱施事)语义成分“一揽子经济改革措施”。例 1.2 所示的句子就是由三个原子事件组成的，三个小句与原子事件间具

有一一对应关系。

例 1.3(摘自网易新闻)

球队目前正积极联络灰熊,并试图拉孟菲斯入伙以达成三方交易。

例 1.3 的句子虽然语法上只有两个小句组成,但它却由“球队目前正积极联络灰熊”、“试图拉孟菲斯入伙”和“达成三方交易”三个原子事件构成,其中“并试图拉孟菲斯入伙以达成三方交易”包含两个原子事件,例 1.3中的形式上的二个小句对应三个原子事件。

对于说话人如何将意识的东西在思维中切分成若干块的问题,Chafe 和 Givón(1977)使用切块(chunking)过程来解释,每块又可进一步切分,层层切分的结果必然进入最基本的过程,其中包括该过程与实体的联系,并把这种现象直接称为各类动词与名词的动名关系。例如,“去某地旅游”这一事件可以切分为“出发去机场”、“在某地的经历”、“搭机回家”等若干块;而“出发去机场”这一事件又可切分为“进机场大厅”、“登记”、“候机”、“登机”等更小的块。Halliday(1985)将各种基本过程表示为由各种过程、参与者、环境因子构成的及物性(transitivity)。Halliday(1985)和胡壮麟等(1989)都认为人类思想中要反映主客观世界不外乎六个过程,这六个过程涉及实体(参与者)、时间、空间和方式(即环境因子)。基本过程的概念应该是本书提到的原子事件的雏形。

从 Chafe 的切块理论中可以得出,事件也是有粒度和信息级别的,后面的叙述中将细粒度小句级别的原子事件称为事件,而将较粗粒度的句子级事件、段落级事件和篇章级事件分别称为句子事件、段落事件和篇章事件。例 1.2 所示的句子事件就是由三个事件(即小句级原子事件)组成的。图 1.1 描述了从原子到跨篇章事件的层级结构。

有了原子事件以及不同粒度的事件概念,根据篇章宏观结构和事件语义学的意义组合性原则,自下而上进行原子事件间语义关系计算可以得到句子事件,由句子事件间语义关系计算可以得到段落事件,类似地,由段落事件可以计算得到篇章事件,通过篇章事件可以对篇章语义进行分析与理解。因而,基于事件语义计算可以达到对文本篇章意义理解的目的,同时,基于对文本篇章意义的理解结果也可以生成表达原文本篇章语义的新文本。

实际上,原子事件可以称为微观事件,可以通过对微观事件的研究

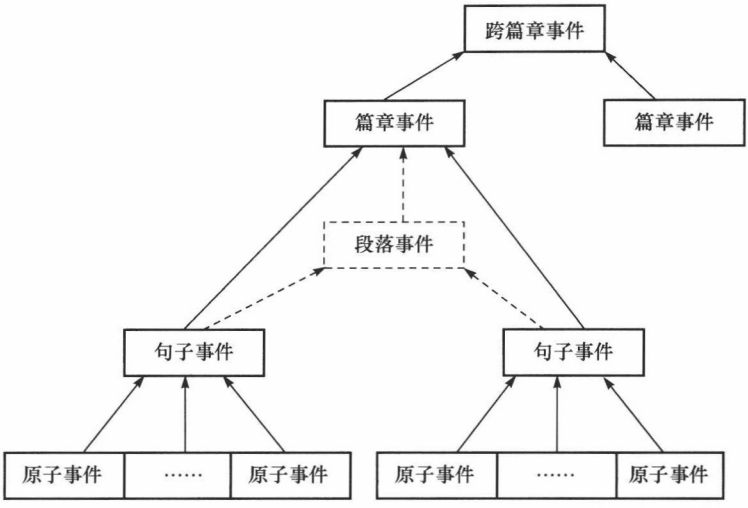


图 1.1 事件层次

来反映篇章级的宏观事件，在宏观篇章级事件基础上研究跨篇章事件，从而反映篇章与篇章之间的关系，即篇际性。

1.2 事件语义

要实现人机间自然语言交流，就意味着要使计算机既能理解自然语言文本的意义，又能使用自然语言文本来表达给定的思想和意图。这样看来，无论是自然语言理解，还是自然语言生成，自然语言意义的分析理解与表达都是至关重要的。而人类的逻辑思维结果和绝大部分知识都是以自然语言文字形式记载和流传下来的。随着互联网的发展，网络文本数据更是迅猛增加。另外，同英语相比，汉语更加讲究意合，汉语符号之间受语义规则的制约更强。因此，针对现有自然语言文本语料库，尤其是汉语文本，面向自然语言文本理解与生成，进行语义计算理论与方法的研究更有必要。

随着自然语言处理与计算语言学的发展，文本处理的语言信息基本单位已由词和词组逐渐过渡到了句子和篇章。要进行篇章研究的原因很简单，就是人类进行交流的语言单位不是词，也不是离散的句子，而是篇章。只有篇章，才能表达完整的意义。对于篇章而言，其本质是

句子集合，但句子在该集合中并不是杂乱排放，而是一个有序序列，句子之间的关系纽带使该句子集合形成一个衔接良好、意义连贯的有机整体。因此，研究篇章中句子之间或者文本片段之间的语义关系，对篇章形式结构、篇章内在含义、篇章情境义，都有着重要意义。

图1.2显示了词、句子和篇章三个研究层次，语法、语义和语用三个研究方面以及对应的九个具体的研究内容。

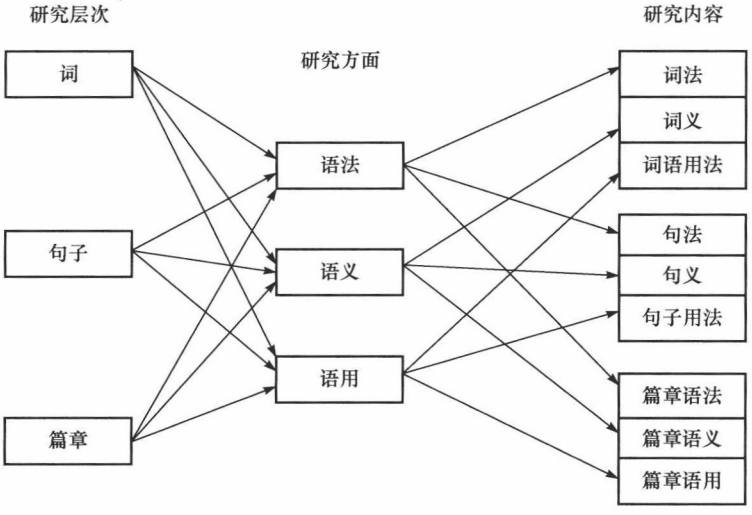


图1.2 研究层次、研究方面与研究内容

对于词这一研究层次而言，词法、词义以及每个词语的用法，基本上已经研究的非常透彻，无论哪个语种，很多离线和在线工具书的成功编纂就是很好的例证。在自然语言处理与计算语言学领域，尤其是对于意合性的汉语这一语种，篇章的研究基本上处于起步阶段。就像要彻底分析理解句子，首先要研究透彻词语一样，人类和计算机要认清篇章，首先要彻底分析与理解构成篇章的句子。对于句子而言，目前研究的并不是很顺利，尤其是在自然语言处理与计算语言学领域，句法的研究基本上已经很深入，但对句义的研究，并不能说已经成功完成，而是还有很长的路要走，而句子的用法，更要放在篇章环境下研究。

目前自然语言语义的研究，处于句义向篇章义过渡阶段，但根据组合原则，句义研究是篇章意义研究的基础和关键。就像把所有句子意义简单叠加在一块并不能理解篇章一样，粗糙地叠加句子中的词义并

不能得到句义。因此，有研究者就在词与句子这两个研究层次间设立了过渡层次，如词组(或称短语)。但本书作者认为，要完整表达一个意义，应该在句子紧邻的下一层(级)设原子事件这一研究对象。就意义单位而言，词组很多情况下并不能表达一个完整的意义，而原子事件可以做到。因此，从原子事件出发，研究句义，进而研究篇章义，有着非常重要的理论意义和实用价值。

1.3 基于认知与计算的事件语义

对事件语义的研究，认知方面将包括事件语义(结构)模式、事件语义关系等内容。而计算方面则包括基于语料实例的事件语义(结构)模式获取、事件语义关系模式获取、事件形式化、事件语义计算等。本书的主要研究内容如图 1.3 所示。

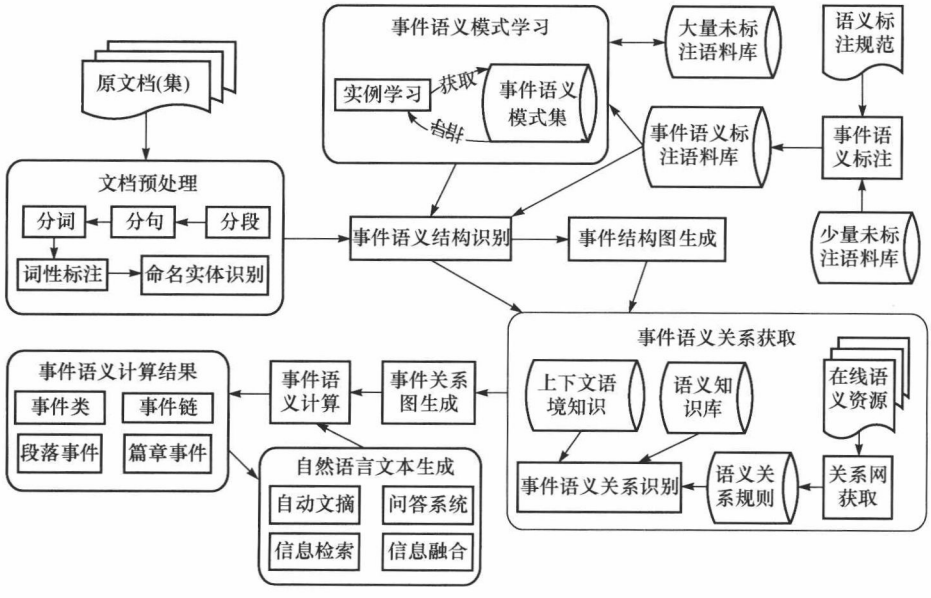


图 1.3 基于事件语义认知与计算的研究内容

1. 事件语义(结构)模式

人类在使用文本描述事件时会形成一些固定角度，因而，现实语料

库中的语言事实也会遵循一些固定的语义模式。从人类认知模式的角度出发,事件语义模式一般由事件谓词和相应几个语义角色组成。虽然在理论语言学领域,语义角色的数目仍然是一个争论的焦点,这虽然并不影响事件语义学理论的发展和应用,但在计算领域却缺少了精确语言理论指导。目前,基本(必要)语义角色有一般有5至6个,附加(非必要)语义角色达17至18个。若事件语义模式由它们组合产生,则不可避免要面对组合爆炸难题。因此,从认知角度,首先需要确定语义角色的数目和事件语义结构辅助标注规范。

在事件语义模式中,虽说事件谓词处于语义支配地位,但事件参与者才是事件描述的重点,是事件意义的核心,而担当参与者的语义成分又多为命名实体。因此,从计算角度,完全可以通过辅助工具识别出命名实体,然后基于事件语义标注规范,采用人工标注方式,为部分文本语料中包括命名实体在内的语义成分标注语义角色。

2. 事件语义关系

事件之间必然有千丝万缕的联系,这样多个句子才可以组成一个意义连贯的篇章。事件之间的联系体现为事件之间的语义关系,在形式上由事件之间的衔接关系来反映。因此,从认知角度,可以将事件语义关系分为渐进(或称递进)、时序(或称先后)、条件、因果、并列等类型,需要确定事件之间的链接手段,进而制定出事件语义关系辅助标注规范。在语料统计与计算中,事件语义关系表现为统计分布关系、事件谓词关系和语义成分关系等,更重要的是由连词等功能词所体现的事件之间的衔接。因此,从计算角度,基于事件语义关系的含义及其所处的上下文语境,要自动识别事件之间的语义关系,首先需要研究如何利用语义知识库、上下文语境语义和在线语义资源等获取事件语义关系。

1) 基于语义知识库

目前语义知识库主要包括中国科学院董振东等的HowNet、美国普林斯顿大学的WordNet、美国加州大学伯克利分校的FrameNet、美国南加州大学的VerbOcean等,可以利用这些语义知识库,获取语义成分间语义关系和事件谓词间语义关系。

2）基于上下文语境语义

在篇章语境中，句子中或者句子与句子之间往往会有用于逻辑衔接的语义联接符，以达到篇章语义连贯的目的。根据语言学家王力的观点，有些虚词处于词和词中间，或者句和句中间，担当语义联接的任务，这种虚词称之为联接词，汉语文本语料语境中有很多这样的语义联接词。实际上，对于意合型的汉语而言，有时由顿号、分号和逗号等标点符号充当了语义联接符，有时跨句的指代关系等起到了语义联接符的作用。通过事件语义结构的分析，可以获取位于一个事件内部的语义联接符，从而体现它所联接的两个成分之间的语义关系，更重要的是，可以通过上下文语境中位于事件句之间的语义联接符，获取事件语义关系。

例 1.4 中的语义联接词“和”体现了语义成分“张先生”和“李小姐”之间的并列关系。例 1.5 中的语义联接词“不仅……而且……”则体现了两个事件之间的渐进关系。

例 1.4 酒店前台本来为微软公司的张先生和李小姐预留了房间。

例 1.5（摘自网易新闻）

深圳 30 年的发展，不仅改变了一个渔村的面貌，而且改变了世界看待中国的眼光。

不仅汉语如此，其他语种也存在相同功能的语义联接符，尤其是英语这种形合型且逻辑性强的语言，基本上每个小句之间都存在语义联接符或者关系线索词，更易于获取事件之间的语义关系。

3）基于在线语义资源

在事件语义结构中，对于充当事件语义角色的命名实体，通过在线语义资源中的呈现，获取它们之间的语义关系，目前常用的语义资源有维基百科、百度百科、互动百科等。事件语义结构中充当语义成分的命名实体，大多为人名、地名、组织名、公司名、事物名、时间等特性较为稳定的成分，完全可以利用在线语义资源获取它们之间的关系，进而形成一个命名实体语义关系网。这些充当事件语义成分的命名实体间的语义关系，很大程度上可以反映它们所在的事件间的语义关系。

3. 事件语义形式化

有了事件语义结构和事件语义关系，在进行语义计算前还要解决一个最关键的问题，那就是要找到一种可计算的事件语义表示形式。图不仅可以直观描述事件语义结构的各语义成分，并且能够清晰展现事件之间的语义关系。另外，图是一种成熟的计算机数据结构，很容易计算机化。因此，使用事件结构图和事件关系图（统称为事件图）来形式化表示事件语义结构和事件语义关系。若某个事件的某语义成分本身为事件，则采用递归事件图来形式化该事件。

利用词性标注和命名实体识别结果，根据学习得到的常用事件语义模式集，以事件谓词为中心，考虑事件谓词周围的各语义成分，尤其是扮演事件语义角色的命名实体。当然，按照事件认知模式，还要重点处理事件谓词周围的时间和空间这两个关键语义成分。至于事件各语义成分的修饰成分，考虑到面向文本生成的特性和系统复杂性，可以有选择的进行处理。一旦识别出文本语料中的相关事件和事件语义结构，则可实现事件自动抽取和事件语义结构图的自动生成。

设计一个基于关系规则和统计方法的关系学习器，利用在线语义资源，学习获得实体（主要是命名实体）语义关系网，进而以语义关系规则的形式来表示。在事件语义模式、事件语义结构和事件图的基础上，设计一个基于上下文语境语义信息、语义知识库、命名实体语义关系网等的事件语义关系获取器，学习获取事件语义关系，实现事件关系图的自动生成。

4. 基于图的事件语义计算

一旦构造出了事件图，即可采用相关图算法（如 DBSCAN 算法、最小生成树算法、关键路径算法等），对事件语义进行计算。在事件图的基础上，考虑事件语义关系的数学特性，基于逻辑推理或格代数理论，进行图推理与计算，得到有益的推理和计算结果，以加深对原文档（集）意义的理解。

若事件语义计算结果为事件类，则可以对各个事件类进行评价，用最重要的事件类来表示原文档（集）的主题；若使用图路径和图推理算

法，则可以得到原文档（集）的语义事件链；若采用剪枝、聚类等相关算法，则可以得到文档（集）每个段落事件，进而可以得到篇章事件及事件演变路径。根据优化改进的图算法以及事件关系图等，可以设计并实现事件语义计算模型。

5. 事件语义计算的应用

自然语言文本生成是在理解语料库原文档（集）意义的基础上，为原文档（集）生成一个描述其语义的简短文本。目前常用方法是直接抽取句子，而真正的文本生成技术应该是基于原文档（集）意义的生成。对于新生成文本，不仅要考虑其语义信息含量，而且信息冗余度要尽量低，语义连贯要尽量好。事件语义计算结果就是对原文档（集）意义的理解，基于计算结果完全可以生成反映原文档（集）意义的文本。

1.4 本书内容安排

在本章讨论事件研究意义、研究内容与方法的基础上，本书后续章节的内容主要安排如下：

第 2 章将从认知与计算角度研究事件语义的理论基础，包括自然语言处理、计算语言学、篇章语言学以及事件语义学。

第 3 章将从认知角度探索事件语义结构分析，主要包括题元角色理论、事件语义角色、事件语义算子等。

第 4 章将从认知角度探讨事件语义关系分析，主要包括衔接与连贯理论、事件平行型关系、事件偏正型关系等，从计算角度讨论事件语义结构和事件语义关系的辅助标注。

第 5 章将从计算角度讨论事件语义形式化，主要包括事件逻辑表示与事件图形式化。

第 6 章主要是关于事件语义计算及其应用，主要探讨基于图算法的事件语义计算和基于事件语义计算结果的自动摘要。