

序

李斌博士的这本专著，主要研究适用于自然语言处理的汉语句子的语义结构分析，其理论意义和应用价值自不待言。众所周知，汉语句子的语义分析是一个老大难问题，在理论体系和技术细节上，有待攻克的难题很多。该书之所以能够令人耳目一新，主要在于作者尝试从新的语义表示方法入手，在美国计算语言学界盛行的描写英语句子的语义框架——抽象语义表示（Abstract Meaning Representation, AMR）的基础上，充分考虑汉语的特点，在大规模数据标注与分析的基础上，逐步建立和完善中文抽象语义表示（Chinese Abstract Meaning Representation, CAMR）的标注体系。这种课题跟研究具体的语言现象与问题不同，必须基于计算语料库进行构建，需要地毯式地处理语料中的各种纷繁复杂的语言现象，如重叠式、连动结构、动补结构、省略、构式乃至各种修辞表达，需要形成统一的、可操作的标注体系与规范，制定具体的操作方法，头绪很多，难度很大。

AMR 使用基于图结构的方式标注句子中词语之间的语义关系。AMR 对配价语法、框架语义学进行了拓展和细化，这在实践层面形成了一套较为完善的整句语义表示的方案。采用图结构的好处是，便于处理论元共享（argument sharing）问题（即一个句法成分可能是前后不同动词的论元，比如“我请他吃饭”中的兼语“他”），而抽象语义的思路能够跳出有关词语字面意义的局限。李斌博士前后花费十年的时间进行摸索，研制出这样一套基本完备的描写方案，这是需要相当的学术信念和进取心的，也需要甘坐冷板凳的功夫和毅力。

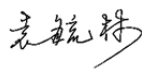
几年前，李斌博士专程来北京大学，跟我的研究小组研讨 AMR 问题。我们就句子语义的复杂性、谓词论元结构和物性结构等问题进行了广泛而细致的交流。短短几年过去了，他主导的团队的 CAMR 研究在语料规模上扩大了，自动分析与评测也做出来了；他还在有关学术会议和期刊上发表了几篇有影响力的论文，并最终形成了这部书稿。该书从标注规范、语料库建设、计量分析与计算评测四个角度来撰写，这种架构是比较全面的。其中，第四部分“探索篇”则是把不太成熟的小规模探索实践也展示出来，供业界同行批评指正或学习借鉴。

该书呈现给大家的是一种难得的比较系统地论述和分析汉语句子语义结构的尝试，其材料、方法和技术都比较新颖，也有许多创新之处。当然，我也提醒过

李斌博士，在语义理论上还要加强学习，努力把不同时期、不同学派的理论更好地融合起来，更多地参考认知语言学有关研究，注重动词的事件框架与名词的物性结构，尽量全面和多层次地反映句子的不同种类的意义。他们的 CAMR 对虚词的处理方式，也有可讨论和改进之处。过去美国的 AMR 理论没有将虚词置于一旁，应该是有一定的理论考量的。因为虚词表达语义关系，放到一旁也许是可以的，但是及物动词也表示两个名词之间的关系，为什么就不放到一旁了呢？此外，动词和名词之间的关系、形容词和名词之间的关系，除了用论元结构关系来看待外，能不能引入物性结构关系来处理或关联？诸如此类问题，值得李斌博士仔细考虑。

我们看到，随着 ChatGPT 的横空出世，大语言模型（Large Language Model, LLM）已经在很大程度上解决了语义理解的问题。但是，AMR 的描写方式还是符号主义的，用依存结构来刻画句子的语义结构，这种方式与大模型采用词向量的分布语义学有着很大的区别。现在中文 AMR 的自动分析精度虽然也达到 80%，但是距离实际应用还有一定的距离。而且，中文 AMR 能达到这 80% 的自动分析精度，也是依靠大语言模型才实现的。那么，是不是要考虑构建面向大语言模型微调的 AMR？

AMR 等基于符号主义的方法，优点在于分析的精准性高、可解释性好，但是，普适性相对较弱，更不用说成本高昂。大语言模型覆盖面广，效果好，但在准确性和可解释性方面存在一定缺陷。在语义分析如何为大语言模型提供高效的预训练和微调资源方面，应该有很多值得探索的内容。我们期待在将来，能够看到语义分析的更好的理论架构、更多的实验分析、更好的落地应用，能够一方面推进语言学的理论建设，支持语言教学的实践，另一方面对于自然语言处理和人工智能也能作出贡献。



2024 年 8 月

澳门大学

目 录

第一部分 建构篇

第 1 章	引言	3
1.1	研究背景	3
1.2	研究价值	5
1.3	研究内容与目标	6
1.4	研究思路与方法	7
1.5	主要成果与结论	8
第 2 章	句子语义研究综述	11
2.1	国外句子语义研究	11
2.2	计算语言学的句子语义研究	14
2.3	国内句子语义研究	27
2.4	句子语义库建设综述	28
2.5	AMR 研究现状	32
2.6	本章小结	35
第 3 章	融合概念与关系对齐的 CAMR 语料库的构建	37
3.1	引言	37
3.2	AMR 语料标注体系	38
3.3	融合概念与关系对齐的 CAMR 的标注体系	40
3.4	语料分析	47
3.5	本章小结	50

第二部分 计量篇

第 4 章	基于 CAMR 语料库的非投影结构研究	53
4.1	研究背景	53
4.2	非投影结构的研究综述	54
4.3	对齐改进版 CAMR 语料库	57

4.4	汉语非投影结构类型及比例统计	58
4.5	非投影结构的理论探讨与处理对策	64
4.6	本章小结	65
第 5 章	省略现象标注与计量分析	67
5.1	省略现象及 AMR 的相关研究	67
5.2	CAMR 处理省略现象的方法	79
5.3	省略现象的标注与统计分析	96
5.4	省略成分的语义角色计量	113
5.5	本章小结	126
第 6 章	基于关系对齐的虚词抽象语义表示与计量分析	129
6.1	引言	129
6.2	相关工作	130
6.3	数据来源与标注	134
6.4	虚词的统计与分析	135
6.5	本章小结	141
第 7 章	汉语疑问句的标注与计量分析	142
7.1	引言	142
7.2	相关工作	143
7.3	数据来源及标注	146
7.4	统计分析	151
7.5	本章小结	153

第三部分 评 测 篇

第 8 章	基于概念关系对齐的中文抽象语义表示解析与评测	157
8.1	引言	157
8.2	MRP 2020 的评测情况	159
8.3	概念对齐和关系对齐	161
8.4	基于概念和关系对齐的 CAMR 解析评测方法	164
8.5	基于概念对齐一体化的 CAMR 解析模型	175
8.6	本章小结	180
第 9 章	中文抽象语义表示解析评测 CAMRP 2022	182
9.1	引言	182
9.2	评测任务	183

9.3 数据集	185
9.4 评测方法	186
9.5 评测情况及结果	190
9.6 本章小结	196
第四部分 探索篇	
第 10 章 基于 CAMR 语料库的汉语谓词语义角色考察	201
10.1 引言	201
10.2 语料的基本情况	202
10.3 CAMR 使用的谓词框架词典	203
10.4 CAMR 遵循的谓词语义角色标注体系	207
10.5 CAMR 对汉语省略核心论元的名词性结构的处理	210
10.6 本章小结	212
第 11 章 基于抽象语义表示的汉语构式的标注与分析	214
11.1 引言	214
11.2 相关工作	215
11.3 数据标注	219
11.4 符合组合原则的构式的标注与分析	221
11.5 不符合组合原则的构式的标注与分析	225
11.6 本章小结	228
第 12 章 基于抽象语义表示的明喻句双层标注研究	229
12.1 明喻现象及标注的相关研究	229
12.2 明喻双层标注体系的设计	242
12.3 明喻句的标注与统计分析	263
12.4 本章小结	276
第 13 章 总结与展望	279
13.1 主要成果与结论	279
13.2 未来工作与展望	281
参考文献	283
附录	305
附录 A 本书研制的 CAMR 语料库和软件系统	305
附录 B 已发表相关论文与章节关系	305
跋	307

第一部分

建 构 篇



第 1 章

引 言

1.1 研究背景

让计算机自动理解和生成人类的自然语言，一直是人工智能领域的核心课题，也是难题之一。近年来，随着词法分析和句法分析的日渐成熟，句子语义的自动分析已成为学界寻求突破的重点课题。虽然 20 世纪六七十年代以来产生了生成语义学、蒙塔古语义学、概念语义学等语义理论，但由于缺少大规模的标注实践和计算机软硬件的支撑，语义自动分析举步维艰（孙茂松等，2014）。进入 20 世纪 90 年代，随着语言理论和计算语言学（computational linguistics）的发展，句子语义分析逐步形成了基于逻辑表达式、框架语义学和依存语法（dependency grammar）的三种方法。

（1）基于逻辑表达式的分析法，把句子自动分析为逻辑表达式，在形式语义学上有较好的理论基础，但受制于逻辑表达式对自然语言的描写能力，往往局限于特定领域的应用，通用性不足（Reddy et al., 2017）。

（2）基于框架语义学的语义角色标注方法（Merlo and Musillo, 2008），不受逻辑表达式的限制，能够标出句子中每个谓词的语义角色（如施事、受事等）。虽然该方法只是句子局部的语义分析，但促进了一个重要问题的解决，即论元共享的描写方法。树结构难以刻画一个词语被两个谓词支配（共享）的情形，于是催生出基于图结构（徐俊明，2010）的语义依存图分析。

（3）基于依存语法的语义依存图分析，是一种整句语义的分析方法，可以分析出句子中词语之间的语义依存关系（Oepen et al., 2015）。该方法脱胎于特斯尼耶尔（L. Tesnière）的依存语法，将句法依存关系升级为语义依存关系，并引入图结构解决论元共享问题。但是，该方法拘泥于原句中的词语，在处理句子省略等特殊结构时存在一定局限。所以，需要建立一种跨领域通用的、描写能力强的整句语义分析方法。

在这种背景下，以美国南加利福尼亚大学奈特（K. Knight）为代表的多位欧

美学者共同提出了 AMR 的句子语义表示方法 (Banarescu et al., 2013a)，将一个句子的语义抽象为一张单根有向无环图 (a single-rooted, directed acyclic graph)。AMR 使用单个根节点 (root) 的图结构保持了树形主干，层次鲜明，又解决了论元共享问题；同时，它增加了语义抽象机制，跳出原句词语的限制，将句子“抽象”为词语对应的概念图来表示句义，忽略语义较虚的成分（如冠词、单复数、时态等）和形态变化，允许根据句子的意思来增添、合并、删除、修改概念。下面以汉语实例来说明 AMR 的两个优点。

首先，语义抽象机制能够有效解决概念缺省和离合词问题。在句子“路过的想帮她的忙”中，传统方法很难合理地表示“路过的”，对于离合词“帮……忙”也往往拆分为动宾结构。相比之下，AMR 通过增补“person”概念，补充出了路过的“人”的含义，“帮……忙”则合并为“帮忙”。

图 1-1 给出了 AMR 的两种展现形式——文本缩进法和图示法。每个概念节点都有一个编号，如 x1、x2 等。arg0、arg1 表示论元关系里的原型施事和原型受事。“想”作为句子唯一的根节点，编号为 x。“想-02”，表示“想”使用的是其第二个义项的论元框架。person 是增补节点，编号 x1，是“想”的 arg0，也是“路过”和“帮忙”的 arg0。

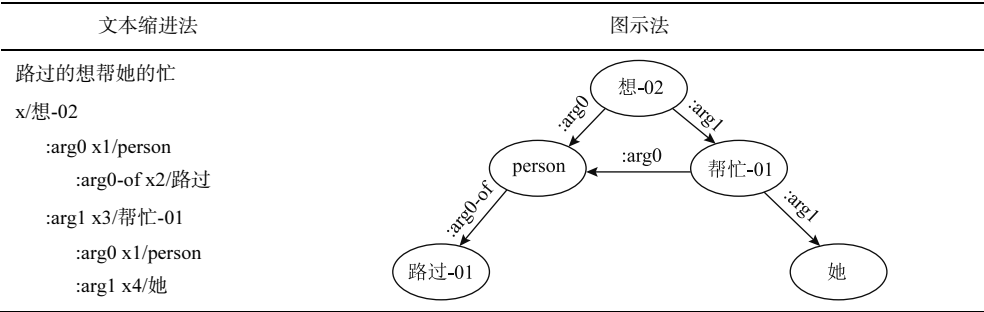


图 1-1 “路过的想帮她的忙”标注示例和图示

其次，图结构解决了论元共享问题。在“路过的想帮她的忙”这个句子中，person 是“想”和“帮忙”的 arg0，而传统方法限于树结构，会舍弃“person-帮忙”这个关系。虽然传统方法使用树结构时添加一些规则也能勉强刻画论元共享问题，但树结构并不能直接反映这种共享的语义构造，图结构则可以直接表示出来。

AMR 具有较强的语义表示能力，能够转换为逻辑形式 (Bos, 2016)。目前，学界已经建立了包含近 50000 句语料的大规模英语语料库，受到了国际计算语言学界的高度关注。AMR 自动分析和应用的研究论文也频繁发表于国际顶级会议和相关期刊。此外，国际语义评测 (International Workshop on Semantic Evaluation, SemEval) 连续两年举办了英文 AMR 自动分析评测竞赛 (May and

Priyadarshi, 2017)。

国内也在不断探索中文句子语义分析方法，主要集中于语义角色标注 (semantic role labeling, SRL) 和依存语义图语料库的建设，如山西大学的汉语框架网 (Chinese FrameNet) (郝晓燕等, 2007)、北京大学的汉语动词句法语义功能信息词典暨检索系统 (袁毓林, 2008; 袁毓林和曹宏, 2022)、清华大学的语义依存库 (尤昉等, 2003)、北京语言大学的汉语语义依存图库 (Chinese Semantic Dependency Graph Bank) (邵艳秋等, 2011; 郑丽娟等, 2014; 郑丽娟和邵艳秋, 2015)、武汉大学的汉语特征结构语义资源 (陈波, 2011) 等。此外，北京大学还构建了基于汉语“意合语法”的“词库-构式”知识库 (袁毓林等, 2014)。

鉴于 AMR 良好的语义表示能力，构建 CAMR 语料库和自动分析技术，就成为一项迫切的工作。不过，初创不久的 AMR 依然存在三个问题：首先，AMR 是根据英语的语言现象制定的，它在汉语和其他语言中的句子语义表示能力没有得到验证，且系统性的理论阐述也较少。其次，AMR 图上的概念和关系没有与原句中的词语对齐，而自动分析时必须先进行自动对齐。然而，目前自动对齐的 F_1 值仅有 0.89 (Flanigan et al., 2014)，拉低了 AMR 的自动分析精度。最后，AMR 忽略虚词、形态变化等的标注方式损失了一些语义信息。因此，需要结合汉语特点对 AMR 进行理论与方法上的修正，补充一定的语法意义，增加对齐信息，使其成为真正通用、完整的语义表示方法，进而构建 CAMR 语义库和自动分析系统。

1.2 研究价值

增加对齐信息、构建适合汉语特点的 CAMR 语义库和自动分析系统，具有重要的理论意义与广泛的应用价值。

(1) 在语言学方面，改进后的抽象语义图 (abstract semantic graph) 的语义表示能力更强，可以更好地研究汉语句法语义问题。借助汉语的词序列 (word sequence)、句法树 (syntactic tree) 和抽象语义图之间的对应关系，能够观察到论元共享的具体情况，以及在哪些词或结构上存在增删概念节点的现象，了解特殊结构 (如“的”字结构、动补结构等) 会转换为怎样的语义表示，发现传统句法树存在的缺陷，探究句法树能否推导出一个句子的语义结构等问题。

(2) 在计算语言学方面，大规模 CAMR 语义库是句子级语义自动分析的基础资源，CAMR 语义自动分析可以应用于舆情分析、人机对话、机器翻译等多个领域。

(3) 在语言教学方面,抽象语义图能够更简洁而清晰地展示出句子的语义结构,可应用于对外汉语教学。

综上,本书根据 AMR 的基本理论和原则,将 AMR 语义图与原句词语作对齐处理,以提高英文 AMR 的分析精度。同时,本书结合汉语的特点制定 CAMR 的标注规范,并建设大规模 CAMR 语义库。随后,开展语料计量分析,研究基于抽象语义图的句法语义理论。最后,研究 CAMR 语义自动分析算法,使其具备可计算、可评测的特性,服务于自然语言处理(natural language processing, NLP)的多种应用场景。

1.3 研究内容与目标

1.3.1 研究对象

本书的研究对象是中文句子的 AMR 方法、抽象语义库构建及自动分析技术。主要研究目标有三点:首先,建设高质量的 CAMR 语义库;其次,构建基于抽象语义图的汉语句法语义理论;最后,研制出 CAMR 语义自动分析器。

1.3.2 总体框架

本书的总体框架由以下四部分构成。

(1) CAMR 语义库的标注规范制定。目前,AMR 的标注规范是基于英语语料总结得出的。汉语与英语差异较大,需要结合汉语自身的语言特点,制定中文的标注规范。

(2) CAMR 语义库的建设。本书计划以宾州中文树库(Penn Chinese TreeBank, CTB) 8.0 的语料为主要语料,同时兼顾语文课本、微博等领域的数据,语料标注规模达到数万句。同时编写人工标注软件,每个句子至少经过两位人员标注,以保证语料库的质量。

(3) 抽象语义图的理论研究。在 CAMR 语义库的基础上,本书将对句子的语义图结构特性展开定量分析,并借助数学中的图论(徐俊明,2010)来构建基于抽象语义图的句法语义理论。

(4) 中文句子抽象语义的自动分析算法研究。本书拟采用 AMR 目前主流的“词串→语义图”和“句法树→语义图”两种自动分析方案,采取深度学习(deep learning)这一前沿的机器学习方法来提高分析效果(May and Priyadarshi, 2017),最终形成一个高效的 CAMR 语义分析器。

1.3.3 研究难点

(1) 标注规范设计难：由于汉语与英语存在较大的差异，且语料中的语言现象纷繁多样，因此设计一套适用于汉语的标注规范并制定相应的标注细则，以覆盖各种语言现象，难度较大。

(2) 语料标注任务难：不同于一般的语料标注任务，AMR 的体系较为复杂，英文标注规范里的语义关系和特殊概念的数量多达 100 多种，加上 AMR 灵活的语义抽象机制，这些情况使得培养标注者、保证标注质量与效率成为一项艰巨的任务。此外，标注汉语时，还要考虑分词、对齐、谓词框架词典构建等问题，所以建立一个高度整合、简洁高效的人工标注平台也很有难度。

(3) 自动分析提高难：AMR 自身结构复杂，英文上的 F_1 值仅有 0.62 (May and Priyadarshi, 2017)，而中文句法分析的性能通常低于英文，所以 CAMR 想达到 0.60 以上的 F_1 值颇具挑战性。

1.4 研究思路与方法

1.4.1 基本思路

本书拟通过系统梳理国内外关于句子语义表示和自动分析方法的研究，借鉴国外 AMR 的原则和方法，结合汉语特点，制定 CAMR 的标注规范，确定概念关系对齐方法，并建设大规模 CAMR 语义库。接着，本书进行量化的统计分析与理论阐释。同时，本书不断实验 AMR 的自动分析技术，进一步充实理论，最终形成汉语句子 AMR 的新的理论框架及自动分析器。研究路线图如图 1-2 所示。

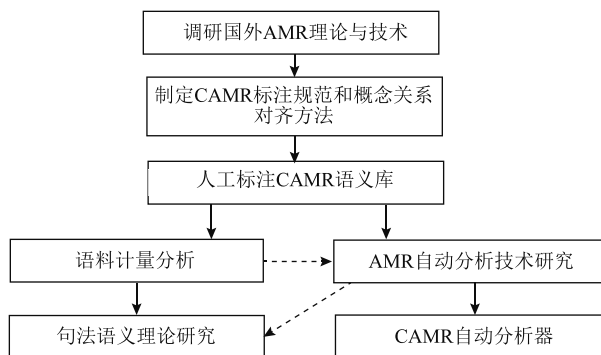


图 1-2 研究路线图

注：实线表示研究的正常推进，理论研究推进理论研究，实验研究推进实验研究；虚线表示理论与实验研究可以互相提供思路与借鉴

1.4.2 具体研究方法

(1) 使用文献法梳理句法语义方面的语言学理论、句子语义表示和自动分析技术的历史演进,追踪 AMR 和深度学习方面的最新研究。

(2) 根据语料库的原则和方法制定 CAMR 的标注规范,培训标注者,编写人工标注软件,构建 CAMR 语义库。

(3) 运用计量方法对中英双语 AMR 语料进行定量分析,结合定性分析来阐述句法语义理论问题。

(4) 使用深度学习等机器学习方法来研究 CAMR 自动分析技术,使之可计算、可评测、可应用。

1.4.3 研究特色

(1) 在大数据时代,语义分析和计算需要采用新的思路和方法。语义自动分析需要语义知识库的支撑,而单纯依靠内省式的人工总结已难以满足信息时代海量信息处理的要求,需要将海量的语言数据和专家知识相结合,通过自动或半自动方法构建出覆盖度高、信息量大、多语言互通的语义知识库,进而采用深度学习等机器学习方法来实现语义的自动分析。

(2) 建立面向计算的 CAMR 语义库。在调研了国内外多种语义理论和知识库,系统地研究了词义表示(WordNet、词语认知属性)、句法语义问题[动宾搭配、宾州命题库(Proposition Bank, PropBank)]和句义分析方法(语义依存图、AMR)之后,我们总结出抽象语义图已具备贯通词汇语义、句法语义、篇章语义的能力,对于当前研究较热的认知义、构式义等也有很强的表示能力。虽然英文 AMR 在理论和技术方面还不够完善,但我们可以根据 AMR 的语义抽象和图结构思想,结合汉语的特点进行改进,从而构建出 CAMR 语义库,服务于中文句子语义自动分析。

1.5 主要成果与结论

随着算力的提升和语料规模的不断扩大,目前自然语言处理已整体推进到语义分析领域,但由于自然语言规则变化多样、语义表达纷繁复杂,语义自动分析仍存在较大的发展空间。CAMR 体系的构建使得语义结构的表示成为可能,为现有的语言理论注入新的活力;大规模精加工语料库的构建为自动分析技术奠定基础,多次评测的开展推动了自动分析技术的长足进步。主要成果体现在理论总结与创新、语料库构建,以及自动分析技术三个方面。

1.5.1 理论总结与创新方面

首先，目前国内外语言学界已有部分语义标注框架，也构建了一些句子语义库，但它们或是无法表示论元共享，或是单纯依靠逻辑式表达，或是无法表达细微语义概念。CAMR 允许概念的添加、删减与更改，以便更好地标注语言现象，单根有向无环图结构的呈现方式也可以清晰地标注句中的主要论元，并解决论元共享问题，大大提高了语义结构的解释力度。其次，通过 CAMR 语义库的标注统计，可以发现并解释传统语言学未发现、忽视或难以解释的语言背后的语义结构。中文中存在大量传统树结构难以处理的非投影结构（non-projective structure），经标注后可统计并分析其比例及产生原因，为非投影结构的自动识别提供理论基础；同时，语言学界也探索了谓词论元框架的设定与编写，对教学词典的编纂有一定的借鉴意义。基于 CAMR 的标注体系，语言学界还推动了省略现象的识别与恢复、“一句多问”和“不疑而问”的疑问焦点的识别与自动分析；将语料库中统计得到的虚词及语义关系与传统语言学的虚词功能作对比，比较并补充虚词功能，可挖掘新的虚词功能。在明喻句的语义分析方面，借助 CAMR 良好的语义表示能力，对句子的表层字面意义和深层明喻意义进行两个层级的标注，揭示深层结构和表层结构的共性和差异，并给出了明喻理解和生成的计算过程。

1.5.2 语料库构建方面

首先，基于 CTB 语料，本书构建了 20000 多句的精加工抽象句库，该库已多次用于国内外评测。其次，本书构建了 8000 多句语言规范的小学语料、1000 多句的构式语料、5000 多句的问句语料以及 500 句出自《人民日报》的明喻句语料。语料题材覆盖了新闻、小学教材、文学作品和互联网数据，对语言现象的覆盖度较高。

1.5.3 自动分析技术方面

笔者团队自主编程设计出与标注规范对应的人工标注平台，研制出 F_1 值为 0.7 的概念对齐一体化自动分析器，设计出专用于 CAMR 解析的评测指标，并多次举办国内外解析评测，这显著推动了汉语句法语义分析的发展，相关成果已可比肩英文 AMR 的自动分析系统。在自动分析技术和评测方面，CAMR 通过组织国内外评测，将汉语句子的分析水平从 60%左右提升到 80%以上，有力促进了句子语义分析技术的进步。

综上所述，CAMR 已经在汉语句子语义的形式化方面做出了重要探索，不仅建立了新的标注体系、构建了大规模语料库，也在诸多语言现象的计量研究中有力地支撑了基本体系的稳定性，同时在汉语句子语义的自动分析方面取得了长足进步。本书整体由建构篇、计量篇、评测篇及探索篇四个部分，共 13 章内容组成，本篇即为建构部分的具体展开。