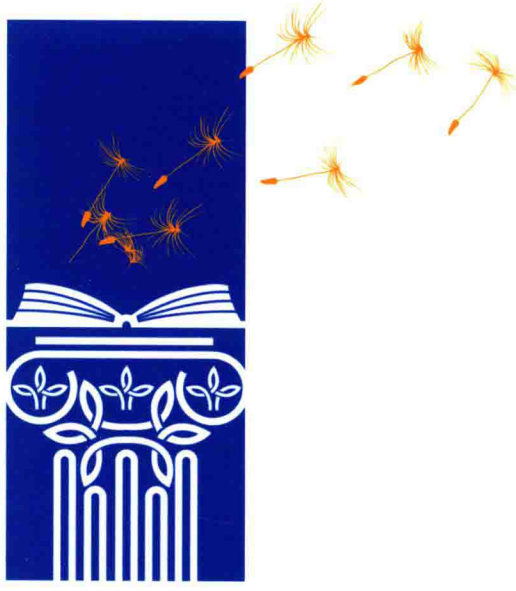

面向概率型词汇知识库建设的名词语言知识获取

LINGUISTIC KNOWLEDGE ACQUISITION OF NOUN FOR THE CONSTRUCTION OF PROBABILISTIC LEXICAL KNOWLEDGE-BASE

王 萌 ◎ 著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

面向概率型词汇知识库建设的名词语言知识获取

Linguistic Knowledge Acquisition of Noun for the Construction of Probabilistic Lexical Knowledge-base

王 萌 著

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

内 容 简 介

语言知识库是众多自然语言处理系统不可或缺的组成部分，同时也是各种自然语言处理技术赖以实现的基础。语言知识库建设已经成为自然语言处理领域最基本、最重要的应用基础研究之一。本书是以北京大学计算语言学研究所开发的综合型语言知识库为基础，围绕异质资源的集成创新这一主题，从资源集成的广度和深度两个方向开展研究的。首先，介绍了综合型语言知识库系统的构成及功能；其次，以名词为切入点，研究从语料中自动获取名词语法属性的方法，内容涉及数词与名词构成的数名结构，数词、量词与名词构成的数量名短语及名词与名词构成的复合名词短语，并对这 3 种属性关系进行了详细的句法和语义分析。

本书可以作为从事自然语言处理和计算语言学相关研究人员的参考书。

图书在版编目（CIP）数据

面向概率型词汇知识库建设的名词语言知识获取 / 王萌著. —北京：电子工业出版社，2016.8
ISBN 978-7-121-29460-0

Ⅰ. ①面… Ⅱ. ①王… Ⅲ. ①名词－自然语言处理 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字（2016）第 171028 号

策划编辑：张贵芹
责任编辑：韩玉宏
印　　刷：北京七彩京通数码快印有限公司
装　　订：北京七彩京通数码快印有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1 000　1/16　印张：10　字数：158 千字
版　　次：2016 年 8 月第 1 版
印　　次：2016 年 8 月第 1 次印刷
定　　价：38.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254511。

序　言

新年伊始，欣闻王萌博士的《面向概率型词汇知识库建设的名词语言知识获取》一书即将出版，甚为高兴。王萌博士约我为书作序，作为她的博士生导师，义不容辞。

王萌自2006年至2010年在北京大学计算语言学研究所攻读博士学位，研究方向是计算语言学，重点则是大规模语言知识的自动获取。

计算语言学的研究内容是实现自然语言的自动处理，包括分析和生成两大任务。自动分析相当于让计算机“读”人类的语言，自动生成则是让计算机“写”自然语言。语言知识库是为实现自然语言处理在计算机系统中配备的有关语言的各种知识的集合。语言知识库是自然语言处理系统不可或缺的组成部分，语言知识库的规模和质量在很大程度上决定了自然语言处理系统的成败。基于这样的认知，北大计算语言所建立30年来，长期将研究重点放在语言知识库的建设上，并取得了一项成果，即综合型语言知识库。王萌入学时，恰逢综合型语言知识库的各个组成模块基本成形，进入集成阶段。王萌入学后，在认真学习、了解已有成果的基础上，积极参加综合型语言知识库的集成研究。通过研究实践，王萌对综合型语言知识库的主体部分《现代汉语语法信息词典》和大规模现代汉语基本标注语料库有了相当深入的了解，对计算语言学的理论、方法和技术有了丰富的积累，并敏锐地认识到，随着统计语言模型和语料库方法在自然语言处理领域的广泛运用，大规模语言知识的自动获取成为自然语言处理技术研究的前沿，便在这个方向上潜心钻研。在攻读博士学位的最后一年，王萌获得到香港理工大学人文学院交流的机会，在黄居仁教授的指导下，进一步增强了语言学功底。经过4年的艰苦努力，王萌完成了博士论文。在答辩会上，答辩委员们一致评价她的论文是一篇优秀的博士学位论文。本书就是在她的博士论文的基础上形成的。

本书的主要内容如下。

1．设计并集成了综合型语言知识库系统的主体功能模块，提出了构建《概率型现代汉语常用词汇知识库》的实现方案，并选择名词作为突破口，从大规模语料中自动获取名词典型的语法属性。（见第2、3章）

2．提出并计算了新的统计量“分散度”，定量地验证了《现代汉语语法信息词典》关于“数名”属性的定性描述，进一步提高了该词典的质量。（见第4章）

3．设计并实现了复杂数量名短语的识别算法，自动获取准确的量名搭配分布，首次定量地分析了量词对名词的语义选择倾向，并研究了量词在名词语义分类中的作用。（见第5章）

4．针对基于统计指标不能有效获取低频复合名词短语的问题，提出了一个新的解决方法，将其视作一个分类问题，利用统计指标获取典型的、高频的复合名词短语作为训练数据，来帮助发现低频的复合名词短语。（见第6章）

5．首次采用动态的策略，提出了“基于动词的释义短语”的方法，对复合名词短语进行语义解释。（见第6章）

这些内容都是王萌自己的研究成果，其中第2项、第3项和第5项具有创新性，特别是第5项的创新性尤为显著。在此借用一则新闻对第5项成果作简要介绍。消费者协会在处理顾客投诉某厂商制作的绿豆饼里面没有绿豆这一案件时，该厂商辩解道：“人家老婆饼里面也没有老婆啊，为什么我的绿豆饼里面一定要有绿豆？”听到这个诡辩，也许觉得好笑；但要驳倒他，你得有说服力地指出他是在偷换概念。人们在用两个名词的组合命名一个新概念时，隐含着不同的意义构建过程且常使用省略形式，即隐去将两个名词组合在一起的动词成分。新闻中的“绿豆饼”实际是“用绿豆制作的饼”，“绿豆”和“饼”之间是材料和制成品的关系；而“老婆饼”是“老婆制作的饼”，“老婆”和“饼”之间是制作人与制成品的关系。显然，这是两种截然不同的语义关系。大众运用语言进行交流必须遵守约定俗成的法则。该厂商利用因省略而带来的语言歧义进行狡辩，自然是站不住脚的。本书第6章的研究对象就是复合名词短语，目标是对复合名词短语给出正确的语义解释。本书提出了一个基于计算的自动方法：在大规模语料中，由软件自动发现与名词概念相关的动词，进而获取基于动词的释义短语并加以过滤和排序。

本书介绍的各项研究成果只有将文理学科的知识融会贯通才有可能取得，这是难能可贵的。王萌博士之所以能做出如此出色的工作和她的学术背景不无关系。王萌大学本科和硕士阶段都是学计算机的，博士阶段攻读计算语言学，这是一个文理交叉的学科，跨度较大。王萌认真踏实，刻苦努力，顺利完成了知识结构的调整。她能够静下心来观察分析复杂的语言现象，针对要解决的问题建立数学模型，并用程序实现计算分析，不断探索，最终取

得了满意的成果。

与自然语言处理的丰富内容与长远目标相比，王萌博士取得的成果只是沧海一粟。本书中的很多工作还需要进一步充实和完善，我相信她不会懈怠。“路漫漫其修远兮，吾将上下而求索。”这是我自己的座右铭，也以此与王萌博士共勉。

序写到此，可以结束了。不过，我还想借此机会再多说几句。王萌 2006 年入学，那时，综合型语言知识库的研制已历时 20 年，对外进行成果转让也满 10 年。北大信息科学技术学院要求综合型语言知识库报奖。王萌在完成第一学年紧张学习任务的同时，做了大量关于报奖的收集资料、整理文档工作，从而为综合型语言知识库获得 2007 年度教育部科学技术进步奖一等奖作出了贡献，也为其后来获得 2011 年度国家科学技术进步奖二等奖打下了基础。报奖工作既复杂又烦琐，王萌认真负责，兢兢业业。对刚入学的王萌来说，做这些事是付出，是奉献。然而，在付出、奉献的同时，王萌也有收获，比较深入地了解了《现代汉语语法信息词典》和综合型语言知识库的内容、设计思想及学术成就，对这项成果有了参与者的情感，这也是王萌后来孜孜不倦地潜心于综合型语言知识库的集成研究和《现代汉语语法信息词典》的深度探索的感情基础。我以为，做任何一件事，只有喜欢它，才能做好。这也是王萌博士写出这本好书的缘由之一。

俞士汶

2016 年 2 月 2 日于北京褐石园

目　录

第1章 引 言

1.1 研究意义和背景

语言文字是人类社会信息的主要载体。现代社会，随着人们对信息依赖程度的加深，语言文字的计算机处理的重要性与紧迫性日益显现出来，各种自然语言处理系统正在得到迅速发展。目前，自然语言处理系统已涉及机器翻译、自然语言人机接口、信息检索、信息抽取、自动文摘、文本分类、问答系统、语音识别与合成、字符识别等方向。纵观各种自然语言处理系统，除了包括普通计算机系统都有的硬件和软件外，一般还有一个特别的组成部分——语言知识库。语言知识库为自然语言处理系统配备各个层面的语言知识，知识库的质量已成为建立或改进自然语言处理系统的关键。此外，如果没有语言知识库的支持，各种自然语言处理技术将无法深入展开。

随着语料库方法和统计语言模型在自然语言处理领域的广泛运用，大规模语言知识的开发和自动获取成为目前自然语言处理技术的瓶颈问题。因此，语言知识库建设成为该领域最基本、最重要的应用基础研究。自 20 世纪 70 年代末起，世界各国对语言知识库的建设都投入了极大的关注，产生了一大批面向不同应用的影响力巨大的语言资源。例如，英国国家语料库（BNC，British National Corpus）由四千多篇代表广泛的现代英式英语文本构成，包含书面语与口语，词容量超过一亿，并提供在线查询服务（http://www.natcorp.ox.ac.uk/）；

WordNet 由美国普林斯顿大学（Princeton University）开发，是一种基于认知语言学的英语词典，按照单词的意义组成一个个同义词集（Synset），同时描述同义词集之间的关系；Penn Treebank 由美国宾夕法尼亚大学（Pennsylvania University）开发，是描述短语结构的句法树库；FrameNet 由美国加州大学伯克利分校开发，目的是对英语词语进行框架语义描述，并以带有语义标注信息的 BNC 语料作为支持。

世界上最为著名的语言数据联盟（LDC，Linguistic Data Consortium）于 1992 年由美国宾夕法尼亚大学创办，并由其主持至今。目前已有各种语言资源二百多种，涉及英文、德文、法文、西班牙文、中文、日文和阿拉伯文等多种语言。LDC 语料库如今已成为全世界自然语言处理科学家共用的数据库，它极大地推进了词法、句法、语义和语用等相关研究。上述提到的 BNC、WordNet 和 Penn Treebank 等资源都已收录在 LDC 中。

与英语等西方语言相比，汉语的形态不发达，适用于自动分析的形式标记相对贫乏，汉语自动分析相对来说更困难，尤其需要重视语言知识库的建设。中文自然语言处理技术的发展更是离不开各种语言知识库的建设。例如，中文分词及词性标注、命名实体识别等技术的进展完全得益于大规模词语切分与词性标注语料库的开发，中文树库的开发推动了句法分析技术的进步，机器翻译如果没有大规模双语语料库的支持将寸步难行。即使最简单的生语料，也可以用来训练语言模型，Google 目前已经发布了 5 元语言模型，这些数据对信息检索、语音识别、统计机器翻译技术都是至关重要的。目前，语言知识库建设的趋势是越来越紧密地结合计算机理解和生成语言的研究需要。高质量的知识库能给语言信息处理提供必需的知识资源，反过来语言信息处理的研究也有助于提高语言知识资源的建设水平。

我国自 1979 年以来开始进行机读语料库建设，先后建成了如现代文学作品语料库（武汉大学，527 万字）及现代汉语语料库（北京航空航天大学，2000 万字）等首批机读语料库。随着研究的不断深入及计算机技术的持续发展，面向以汉语为核心的多语言信息处理的语言知识库的建设全面展开，一批颇具影响的语言资源库相继建成。例如，北京大学开发的综合型语言知识库、董振东教授主导的知网（HowNet）、清华大学构建的汉语句法树库、哈尔滨

工业大学开发的汉语句法依存树库等，都在中文自然语言处理研究中发挥了积极的作用。

此外，中国台湾“中央研究院”在语言知识库建设方面也有相当丰富的积累，如现代汉语平衡语料库（Sinica Corpus）、中文词汇特性速描系统（CWS，Chinese Word Sketch）、中英双语知识本体词网（Sinica BOW，Sinica Bilingual Ontological WordNet）和中文句结构树库资料库（Sinica Treebank）等语言资源。香港城市大学开发的共时语料库 LIVAC（Linguistic Variation in Chinese Speech Communities）也是中文语言知识库重要的组成部分。中文语言资源联盟（Chinese LDC，Chinese Language Data Consortium）是为推动我国的语言资源共享建立的第一个联盟性的学术组织，它的成立对中文信息处理领域的研究和开发具有重要意义。

北京大学计算语言学研究所（ICL/PKU，Institute of Computational Linguistic/Peking University）一直致力于各种以汉语为核心的语言资源的建设，目前已经拥有一系列内容丰富的语言知识库，总称为综合型语言知识库。最具代表性的工作首推《现代汉语语法信息词典》，以及在此基础上发展起来的大规模现代汉语基本标注语料库。目前，综合型语言知识库中的各种数据资源在自然语言处理任务的评测及语言建模的理论探讨等研究中都得到了广泛应用。

本书的研究工作以综合型语言知识库为基础，围绕异质资源的集成创新这一主题，兼顾集成的广度和深度：首先，从广度上，建设综合型语言知识库系统，综合考虑各种知识库资源，在最大程度上挖掘和发挥资源集成的优势；其次，从深度上，将《现代汉语语法信息词典》与语料库相结合，以名词作为切入点，深入研究从语料库中自动获取名词句法语义知识的方法，为最终构建《概率型现代汉语常用词汇知识库》进行技术储备。

1.2 研究内容和基础

本书的研究工作主要围绕以下内容展开。

第一，探索异质数据资源集成的方法，将结构和表现形式各不相同的语言知识库纳入同一个软件平台，建设综合型语言知识库系统，实现信息服务向知识服务的转型，为自然语言处理研究、语言本体研究及语言教学研究提供全方位、多层次的支持。

第二，基于综合型语言知识库系统，将结构化知识（词典知识）与非结构化知识（语料库）相融合，实现资源的集成创新。以《现代汉语语法信息词典》和大规模现代汉语基本标注语料库为基础，研究语法属性的概率化描述方法，最终建设《概率型现代汉语常用词汇知识库》。

第三，以名词作为研究对象，从语料中自动获取名词的典型语法属性，内容涉及数词与名词构成的数名结构，数词、量词与名词构成的数量名短语及名词与名词构成的复合名词短语，对这 3 种属性关系进行了详细的句法和语义分析。

需要特别说明的是，本书的研究工作是在 ICL/PKU 二十余年积累的若干数据资源的基础上展开的，如果没有前人对语言知识进行归纳、总结、提炼及形式化，构建如此丰富的语言数据资源，那么本书的研究工作将无从展开。ICL/PKU 其他领域的成果也为本书的研究工作提供了帮助。

1.3 本书内容及结构安排

第 1 章　引言

介绍本书研究课题的意义和背景，概述研究内容、基础及资源，确立本书的研究目标、研究范围和方法。

第 2 章　综合型语言知识库系统

首先，介绍综合型语言知识库的资源概况，包括数据资源和软件资源等。其次，介绍以词义为主轴建设综合型语言知识库系统的基本思想，以及在该思想指导下建成的综合型语言知识库系统的 3 个主体功能模块：语言加工模块、知识检索模块和知识挖掘模块。

第 3 章　词语的概率语法属性研究

介绍汉语词汇计量研究的发展，基于词典与语料库，以构建《概率型现代汉语常用词汇知识库》为目标，研究从语料库中自动获取词语语法属性的方法。以名词作为研究对象，对其语法属性进行分析，选择名词与数词、名词与量词及名词与名词的关系，共计 12 个属性进行概率化描述。

第 4 章　数名结构的计量研究

研究名词与数词的关系，基于语料库，对数词与名词构成的数名结构进行计量研究，提出统计量“分散度”用于区分数词与名词构成的是固定搭配还是自由短语。该统计量对于其他问题，如量词的分类等，也具有借鉴意义。

第 5 章　量名搭配的句法语义分析

研究名词与量词的关系，设计并实现复杂数量名短语的识别算法，对语料进行标注，从而获取量名搭配的真实分布。在此基础上，采用基于信息论和知识的计算模型，自动获取汉语量词对名词的语义选择倾向，定量分析量词所搭配名词的语义分布情况。此外，提出基于量词的名词概念描述方法，通过聚类方法考察量词在名词语义分类中的作用和贡献。

第 6 章　复合名词短语的研究

研究名词与名词的关系，即由两个名词构成的复合名词短语。首先，对低频复合名词短语的自动获取方法进行探索，将该问题视为二分类问题，利用 3 种统计指标（共现频次、互信息、依赖率）获取高频的复合名词短语作为训练数据，用于低频复合名词短语的预测。其次，对复合名词短语的语义解释进行研究，采用动态的策略，提出“基于动词的释义短语”的方法，给出复合名词短语可能的语义解释。最后，用基于相似度计算的方法来进行复合名词短语的推荐，帮助人们理解和推测那些不熟悉的或新出现短语的语义关系。

第 7 章　总结与展望

对本书的研究工作进行总结，以及指出进一步的研究方向。

附录部分列出本书中相关的数据处理样例和实验结果。

第 2 章　综合型语言知识库系统

语言知识库是自然语言处理系统不可或缺的组成部分，语言知识库的规模和质量在很大程度上决定了自然语言处理系统的成败，这已经成为学界的共识。本章首先简要介绍国内外有代表性的语言知识库系统；其次，详细介绍ICL/PKU已有的综合型语言知识库，包括数据资源和工具软件；最后，介绍综合型语言知识库系统的构建思想、总体建设规划及完成的主体功能模块。

2.1　国内外相关研究

语言知识库的建设涉及语言知识的整理、发现、形式化、规范化等工作，语言知识库的内容和知识表现形式是多样的。广义上，语言知识库包括词汇知识库、语料库、规则库及各种处理工具等。其中，词汇知识库主要描述词语的词法、句法或语义信息；语料库最简单的是生语料（raw text），或者是参照某种语言学理论对文本中隐含的语言现象进行显性标注的语料（annotated text）。表 2.1 列出了部分国外有代表性的词汇知识库和语料库的基本情况。

表 2.1 中的语言知识库项目有一些已经涵盖中文，如 Penn Treebank 和 Chinese Gigaword。国外建设知识库的理论和方法对国内的相关研究提供了借鉴。例如，山西大学参照 FrameNet，以汉语语料事实为依据构建了 Chinese FrameNet，包括框架库、句子库和词汇库 3 个部分的内容；北京大学基于 WordNet