隐喻是自然语言处理的棘手问题之一，近几年来开始受到中文信息处理研究者的关注。隐喻大量地存在于我们的语言生活中，Lakoff & Johnson(1980)指

汉语名词短语隐喻识别研究

王治敏◉著

北京语言大学出版社
BEIJING LANGUAGE AND CULTURE UNIVERSITY PRESS

中央高校基本科研业务费专项资金资助
Supported by " the Fundamental Research Funds for the Central Universities"

汉语名词短语隐喻识别研究

王治敏 / 著

北京语言大学出版社
BEIJING LANGUAGE AND CULTURE
UNIVERSITY PRESS

图书在版编目(CIP)数据

汉语名词短语隐喻识别研究 / 王治敏著. —北京：
北京语言大学出版社，2010.9
ISBN 978-7-5619-2905-6

Ⅰ.①汉… Ⅱ.①王… Ⅲ.①汉语－名词－短语－隐喻－识别－研究 Ⅳ.①H146.3

中国版本图书馆 CIP 数据核字（2010）第 212050 号

书　　名：汉语名词短语隐喻识别研究
中文编辑：徐　雁
英文编辑：侯晓娟
责任印制：汪学发

出版发行：北京语言大学出版社
社　　址：北京市海淀区学院路 15 号　　**邮政编码**：100083
网　　址：www.blcup.com
电　　话：发行部　82303650/3591/3651
编辑部　82303647
读者服务部　82303653/3908
网上订购电话　82303668
客户服务信箱　service@blcup.net
印　　刷：北京联兴盛业印刷股份有限公司
经　　销：全国新华书店

版　　次：2010 年 11 月第 1 版　2010 年 11 月第 1 次印刷
开　　本：787 毫米×1092 毫米　1/16　**印张**：11.5
字　　数：178 千字
书　　号：ISBN 978-7-5619-2905-6/H·10288
定　　价：27.00 元

序

正当辞牛岁、迎虎年之际，北京语言大学副教授王治敏博士发来喜讯，以她的博士论文《汉语名词短语隐喻识别研究》为基础的书稿即将由北京语言大学出版社出版，甚感欣慰。王治敏博士要求我为其写序，尽管我一向认为凭自己的学识难以胜任为他人著作写序的重任，但我还是答应了。这是平生第二次。第一次是为曲维光博士的著作《现代汉语词语级歧义自动消解研究》写序，当时是盛情难却。这一次有所不同，作为王治敏的博士生导师，应该是义不容辞吧。还有一层原因，我觉得可以顺便把这两本书作一个比较。

我将曲维光博士的著述比喻为在自然语言处理战场上“打攻坚战”，王治敏的博士论文则有点像“打前哨战”。为什么这么说呢？在《现代汉语词语级歧义自动消解研究》之序一中我写道：“当前自然语言处理研究的主攻方向，是让机器能够自动地识别和消解自然语言的歧义。曲维光博士的研究重点是词语级的各种类型的歧义消解，这是自然语言处理研究的基本问题，已经研究很多年了，但还没有彻底解决，甚至离彻底解决尚有很长的路要走。这种情况一方面说明，这里有创新的机会和发展的空间，另一方面也说明，创新和发展的难度很大。可以说，曲维光博士是在打攻坚战。”而隐喻的计算研究（包括隐喻的机器识别、理解与生成），情况就不一样了。至少到目前为止，在中文信息处理学界，隐喻还没有成长为受广泛注意的研究课题，鲜有研究成果发表。王治敏自 2003 年至 2006 年在北大攻读博士学位期间，选定隐喻作为攻关方向，并于 2006 年完成博士论文，环视中文信息处理的各个战场，将其工作比喻为“打前哨战”，也许还算贴切。

只要研究语言，隐喻绝不陌生。我们祖先早在《诗经》中便娴熟地运用了明喻、暗喻这一类隐喻表现技巧。世界上诸多先哲都发表过关于隐喻的论述。长期以来，汉语学界也重视隐喻，不过主要还是作为修辞学的研究对象。当代的认知语言学认识到隐喻是人类的一种认知或思维方式。将自然语言理解作为最高境界的自然语言处理研究绝不可能无视语言运用中普遍存在的隐喻表达。20 世纪 80 年代中期，我看到了美国 Carnegle Mellon University 的一篇题为 *Metaphor: An Inescapable Phenomenon in Natural Language Comprehension* 的技术报告（1981 年 5 月 4 日），结合自己对自然语言理解问题的思索，了解到隐喻一定是计算语言学的一项引人入胜的研究课题。不过，当时自然语言处理技术的研发刚刚起步，有太多更紧迫的事情要做，只好将其搁置。进入 21 世纪之后，情况有了很大变化。北京大学计算语言学研究所（ICL/PKU）的各位同仁和学生一起，经过近 20 年的努力，在自然语言处理基础研究领域有了一定的积累，特别是至关重要的语言数据资源建设取得了显著的进展，以《现代汉语语法信息词典》为第一块基石的“综合型语言知识库”已经成长起来了。在这样的条件下，当我思考和规划计算语言学新的研究课题时，“隐喻”自然从记忆中跳了出来。2002 年 9 月，国家重大基础研究计划（“973 计划”）的“图像、语音、自然语言处理与知识挖掘”项目在北京召开中文信息处理研讨会，我有幸出席并以《语料库与综合型语言知识库建设》为题作了发言（同名文章收于徐波、孙茂松、靳光瑾主编的《中文信息处理若干重要问题》一书，科学出版社，2003），发言最后我谈到了“关于基础研究的选题与深入”的话题，提出了隐喻计算研究的新任务，给出了若干隐喻的实例，探索了求解方向，并指出与歧义消解一样，隐喻计算也是自然语言理解必须攻克的堡垒。

值得庆幸的是，随后发生的几件事让这个设想得以付诸实施。第一件事是 2003 年王治敏考入北京大学，在我的指导下攻读博士学位。王治敏是语言学硕士，参与过机器翻译系统的开发，有一定的实践经验。我看重她文理交叉的学术背景，建议她研究面向机器理解的汉语隐喻计算问题，她没有犹豫，立即全身心投入。可贵的是，王治敏不仅刻苦、认真，而且对科学研究满怀激情，经过三年艰苦的探索和持续的努力，

完成了博士论文《汉语名词短语隐喻识别研究》。第二件事是2004年我申请到了国家973课题“文本内容理解的数据基础”（2004年9月至2009年9月），在这个课题中我把隐喻计算研究列为一个子任务，交由王治敏负责，极大地激发了她的研究热情。在这个背景下，隐喻研究不再孤立进行，与面向文本内容理解的各个子任务相互配合，相互支持，这也是王治敏得以克服重重困难，研究不断取得进展的重要原因之一。随后，2005年曲维光博士进入博士后站，他对隐喻问题也产生了浓厚的兴趣，不仅在北大参与研究，而且以南京师范大学副教授的身份于2007年申请到了国家自然科学基金项目“汉语隐喻理解关键技术研究”（2008年1月至2010年12月），隐喻计算的种子开始在更广阔的土地上发芽、生长。2006年又有贾玉祥考上北大博士生，拿过接力棒，将隐喻计算研究推向深入，新成果将反映在他拟于今夏完成的博士论文中。

纵观隐喻计算研究的发端与进程之后，我还是回到正题，评述王治敏博士的著作。浏览全书不难发现，作者在吸收国内外研究成果的基础上对于隐喻的研究有自己独特的相当深入的思考。例如，文中提出“隐喻计算方法正处于由单纯的知识推理逐步向基于大规模语料的统计方法转变的过程中，由此隐喻语言知识工程的建设也得到了应有的重视。如果能够把汉语的隐喻描写同大规模语料结合起来，把隐喻的描写成果标注到语料库中，这样机器就可以在此基础上自动学习”。本书正是沿着“隐喻描写—知识表示—知识库建造—隐喻识别”这一思路对名词短语隐喻进行多角度的考察，并取得了以下多项创新成果。

1. 本书全面地描述了汉语名词隐喻的层级分布。

隐喻被认为是人的一种思维方式，广泛存在于人类语言中，其类型之多、问题之复杂，至今还没有完全理清楚。探索性研究不宜全面铺开，需要确定重点。通过寻找隐喻的外部特征，全面描述汉语名词隐喻的层级分布，最终聚焦于汉语名词短语的自动识别。实践证明这个决策是可行的、明智的。在描写的基础上借助大量的统计数字来明确自己的攻关方向，这是值得推荐的宝贵经验。

2. 注重隐喻知识资源的积累和加工，提出汉语隐喻知识库的建造方法。

随着大规模语料库的出现，关注从语料库中提取隐喻知识可以弥补基于规则方法建造的知识库之不足。作者勇于实践，研制了汉语名词隐喻词表。该词表是作者通过对《现代汉语语法信息词典》中的3万多个名词进行细致考察和潜心研究之后编制出来的，这对人们全面了解和认识名词隐喻，进行更深一步的研究是非常有价值的。汉语名词隐喻知识库是隐喻识别、理解与生成的重要知识资源。

3. 重点探索了名词短语隐喻的识别技术，建造了规则和统计相结合的隐喻自动识别模型。

本书特别重视各种方法的比较，在书中基于规则的方法和基于统计的方法都有体现，各有侧重，同时也使用了多种机器学习的分类模型，通过实验进行检验，并作了细致的分析与解释。这样的方法对于提高研究者的水平与能力是大有助益的。

王治敏博士从北大毕业后到北京语言大学任教。在热心汉语教学工作的同时，并未中断研究工作，继续考察与思索隐喻计算的各种问题，也有新的论文发表。不过，本书并没有增补作者的新认识和新成果，基本上保留了博士论文的原貌。我以为这样也挺好，本书可以成为作者在计算语言学研究的漫漫征途中的一个路碑。

在中文信息处理领域，隐喻计算研究虽然取得了一些成果，除ICL/PKU的工作之外，厦门大学周长乐教授也有新著出版（《意义的转译——汉语隐喻的计算释义》，东方出版社，2009），但是对于多数人来说，面向机器理解的隐喻计算还是一个新课题，研究者甚少。衷心期望本书的出版有利于大家更多地了解隐喻，能够进一步推动隐喻计算研究的深入和拓广，共同携手攀登自然语言理解的高峰。

俞士汶

2010年3月12日植树节

于北京大学计算语言学教育部重点实验室

前　言

隐喻是自然语言处理的棘手问题之一，近几年来开始受到中文信息处理研究者们的关注。隐喻大量地存在于我们的语言生活中，Lakoff & Johnson（1980）指出，隐喻不仅仅是语言的修辞手段，而且是人的一种思维方式。如果隐喻的识别和理解不能很好地解决，那么它将成为未来自然语言处理技术发展的瓶颈。

本书面向语言信息处理，全面地考察汉语名词性隐喻的分布，总结和发现名词性隐喻的表达规律，利用机器学习的方法探索短语层级的隐喻识别，为全面的隐喻自动识别和理解奠定基础。全书共分七章。

第一章介绍了本书的研究背景，确立了本课题的研究思路及方法，即采用统计与规则相结合、定量与定性相结合、识别研究与构建隐喻知识库等实用目标相结合的研究策略。

第二章在传统隐喻研究的基础上评述了隐喻计算模型与知识库建设方面的最新进展，力图借鉴国内外学者的研究成果，明确面向中文信息处理的隐喻形式化的方向。

第三章提出名词隐喻的层级描写，通过考察 n + n 名词隐喻在构词→词汇→短语→句子→篇章等不同层级的分布，建立面向文本内容理解的名词隐喻的工程定义，确定了面向中文信息处理的隐喻研究重点：以短语隐喻表达为核心，探索源域（source domain）到目标域（target domain）的隐喻映射规律。同时从构成、句法、语义等角度对名词隐喻进行考察，建立了以源域为核心的名词隐喻知识架构体系。

第四章设计和建造了汉语隐喻知识库。本书从大规模真实语料中发现隐喻现象，提炼加工了汉语名词隐喻词表；在此基础上又利用《现代

汉语语法信息词典》（GKB）和《中文概念词典》（CCD）的基础平台，搭建出新的汉语名词隐喻知识库。名词隐喻知识库一方面利用了 CCD 中概念存储编号的唯一性，通过人工概念消歧，建立了一个源域到多个目标域的映射关系；另一方面名词隐喻知识库的属性字段也继承了 GKB 的部分成果。

第五章提出基于机器学习方法 + 规则辅助的汉语名词隐喻识别策略。隐喻识别过程被描述成隐喻义与字面义的分类问题，分别对单个词语和“n + n”模式进行识别实验。单个词语识别充分利用隐喻标注资源和人工归纳的语言知识，通过实例方法、最大熵方法和朴素贝叶斯方法的隐喻建模，在综合上下文词语、词性等多项特征的基础上，进行了三种模型的比较实验，最后确定最大熵模型为理想模型，然后再引入多项辅助特征提高识别效果。“n + n”模式识别建立在单个词语实验的基础之上，实验过程重在建立隐喻相似度推理，同时也验证了名词隐喻知识库的有效性。

第六章结合 CCD 和隐喻知识库建立汉语名词隐喻扩展推理。为了能够更好地建立隐喻的相似度推理，我们运用人机互助方法对 CCD 词典进行了合理剪裁，建立了一个词语对应一个语义类的词典格式，为后续的相似度实验提供了保证。

第七章对本书的研究工作及取得的成果进行了全面总结，并提出了进一步的研究计划。

隐喻识别在中文信息处理领域是一个新的研究方向，本课题的研究工作可以看做攻克隐喻计算理解堡垒的前哨战，相关研究成果可对未来的汉语隐喻计算研究提供资源支持。例如，实验所用的各种统计软件都可作为隐喻自动识别的工具；汉语名词隐喻词表作为基础资源，为隐喻的计算理解提供了有价值的数据；汉语隐喻知识库中源域和目标域的概念映射为人们提供了一组组清晰的汉语隐喻映射图画；新闻领域和文学领域的一定规模的名词隐喻标注语料库，可为计算机的隐喻识别和理解提供重要参考。

作者希望这本小书能对从事中文信息处理以及语言学的研究人员，特别是对热衷于隐喻计算、认知理解的研究者提供帮助。虽然书稿几经修改，但难免会有一些不当与纰漏之处，敬请各位专家不吝指正！

目　录

第一章
引　论

1.1　问题的提出

本书的主要目标是面向语言信息处理，全面地考察汉语名词性隐喻的分布，总结和发现名词性隐喻的表达规律，尝试利用机器学习的方法探索名词短语层级的隐喻识别，为全面的隐喻自动识别和理解奠定基础。

隐喻作为自然语言处理的棘手问题之一，近几年来开始受到学者们的关注。俞士汶（2003b）在认识歧义是自然语言处理难题的同时，也认识到隐喻是自然语言理解必须攻克的难关。隐喻大量存在于我们的日常生活中，Lakoff & Johnson（1980）指出，隐喻不仅仅是语言的修辞手段，而且也是人的一种思维方式。最近一项针对电视和新闻评论的调查也显示，做这些节目的人平均每 25 个词就要使用一个独特的隐喻（蓝纯，2005）。如果隐喻的识别和理解问题不能很好解决，将成为自然语言处理技术发展的瓶颈。

一般来说，隐喻（metaphor）是指不带标记词的比喻（王逢鑫，2001；冯晓虎，2004）。例如：

① 扬起希望的风帆，驶向胜利的彼岸。
② 我不是你的终点站。
③ 经过仔细排查，犯罪分子终于浮出水面。

这样的隐喻对人来说很好理解，但是让计算机能够正确识别却是个

难题。原因有二：其一，计算机识别隐喻需要什么样的知识；其二，我们能够提供哪些知识。只有厘清这两方面的问题，隐喻的研究方向才能明确。面向计算机处理的隐喻研究面临着诸多困难，大致可以概括如下：

首先，语言学本体的隐喻研究尚不系统，近二十年来的隐喻研究虽有重要进展，但是还缺乏一种对隐喻进行整体把握和全面分析的统一的理论框架（束定芳，2000：18）。

其次，隐喻现象错综复杂，广泛存在于日常语言、成语、习惯用语、古诗词中，隐喻的映射机制不是十分清楚。

再次，隐喻和词义引申之间错综复杂的关系至今还未理顺。

最后一个最重要的，是隐喻的汉语计算理解研究在国内起步较晚，可以借鉴的东西不多。长期以来，在自然语言处理领域，隐喻被认为是一种“辞格”，一直是中文信息应用系统不予考虑的问题。不过，近几年随着语料库资源的不断发展，学者们逐渐开始关注词语的情感色彩、隐喻的理解问题（俞士汶，2005；周昌乐，2004；张威，2003；杨芸，2004；戴帅湘，2005a；王雪梅，2005；Zhimin Wang，2006）。探索隐喻的内在规律，研究隐喻的自动识别，将是自然语言理解研究中的一个新的里程碑，也是对计算机理解修辞性语言的一个全新的挑战。面向机器理解的隐喻研究对于中文信息处理具有重要的理论意义和应用价值。

这一课题的提出一方面源于语料库语言学和语言本体研究的不断深入，另一方面依赖于目前中文信息处理已有的一些较成熟的技术（如文本分类技术），隐喻计算研究可以充分利用大规模语料提供的资源，以统计为手段，对各种有关隐喻的假设和理论进行验证，同时可以开展大规模的隐喻识别研究，进一步深入探索，给出相应的实验结果。

本课题从计算机处理的角度，在大规模语料考察的基础上，结合机器学习技术，探讨了汉语名词性隐喻的理解和识别。一个完整的隐喻往往由“喻体”和“本体”构成（束定芳，2000：66）。“喻体”通常是我们熟知的、比较具体直观、容易理解的一些概念范畴，而“本体”通常是我们后来才认识的、抽象的、不太容易理解的概念范畴。认知语言学则认为一个概念隐喻包含两个部分，一个始源域（source domain）和一个目标域（target domain），隐喻的认知力量就在于将始源域的图示

结构映射到目标域之上（蓝纯，2005：116）。这里的始源域也称做“源域”，对应于汉语中的术语“喻体”，“目标域”对应于汉语中的“本体”。为了能够更清楚地描述隐喻概念之间的映射关系，这里沿用了“源域”和“目标域”的说法。本书中的名词隐喻就是指由名词充当源域和目标域的隐喻表达。例如：

④ 张开理想的翅膀，奔向知识的海洋。
⑤ 让理想的航船从这里扬帆。
⑥ 历史的潮水还是没有让我们的影视产业得到有效的酝酿和提升。

这里源域“翅膀、海洋、航船、潮水”一般由表示具体事物的名词充当，而目标域“理想、知识、历史”一般是含义抽象的词语。隐喻过程通常是用自己熟悉的简单的事物去类比、诠释复杂陌生的事物，隐喻的过程也就是源域到目标域的映射过程。源域和目标域分别来自两个完全不同的概念域。因此形成了“理想如飞鸟、知识如海洋、理想如航船、历史如潮水”的隐喻系列表达。值得注意的是，源域概念“翅膀、海洋、航船、潮水”在词典中通常没有标注隐喻的义项，它们的词典义项通常是字面含义的解释。例如：

⑦ 国家的海洋资源需要我们每个人来保护。
⑧ 运送货物的航船今天将起航。
⑨ 好多地方的潮水已浸及颈部。

汉语中有多少词语可以形成这样的隐喻表达，它们在真实文本中的分布如何，有待于进一步研究，而这些数据对隐喻的自动识别有很好的利用价值。

目前，《现代汉语词典》（简称《现汉》）只有少部分词语标注了隐喻义项，而大部分词语的隐喻用法词典并没有收录，根据苏新春（2002）的统计，第二版《现汉》共收词条56147条，其中有2488条词语含有比喻义或比喻例句，占全部词语的4.43%。这就是说计算机理解隐喻的知识最多只有4.43%能从词典中得到提示，这样无形中就

增加了识别隐喻的难度。隐喻错综复杂，有些词语在词典中即使标注了隐喻义项，可是由于判别条件很难归纳，在具体环境中也很难识别。例如：

⑩ 阳光总在风雨后。

⑪ 风雨过后，阳光会显得更灿烂耀眼。

例⑩通常理解为一个隐喻表达，例⑪为一个字面表达，但是在《现汉》中“阳光”没有隐喻义项，而“风雨”有隐喻义项，表示“比喻艰难困苦”的含义，虽然义项不同，但它们在句子中的作用是一致的。例⑩的“风雨”和“阳光”既可以说都是比喻义，也可以说都是原义，原因在于例⑩本身就有两种理解，一种是隐喻表达，一种是字面表达，但在实际的语料中，几乎不用其字面含义，是什么因素在起作用，目前还难以给出合理的解释，况且这里的“阳光”、“风雨”什么条件下是字面义，什么条件下是隐喻义也很难下结论。这种细微的差别计算机识别起来更是难上加难。

隐喻一直被认为是语用学、修辞学的范畴。近年来，人们才开始从认知角度对隐喻进一步探索，关于隐喻认知方面的研究已有专著出版（胡壮麟，2004；赵艳芳，2002；束定芳，2000；蓝纯，2005；李福印，2004b；冯晓虎，2004）。隐喻计算模型国内研究不多，不过最近两年有一些讨论。例如：俞士汶（2003b）从自然语言理解角度提出了“空手套白狼”、“郎平是个铁榔头”、古诗词等多个层面的隐喻理解问题；台湾学者 Ahrens，Kathleen & Churen Huang（2003）利用 WordNet、SUMO 等语义资源所作的隐喻映射研究；袁毓林（2004）的容器套件隐喻研究、周昌乐和他的学生们在隐喻逻辑推理方面的探索，等等（周昌乐，2004；张威，2004；杨芸，2004；戴帅湘，2005a；王雪梅，2005）。总体上讲，汉语的隐喻计算研究还处于起步阶段。

国外在这方面已经先走了一步，学者们在提出隐喻功能解释理论的同时，在隐喻计算模型方面也进行了一些尝试，提出了一些形式化手段。主要模型有：Fass（1988，1991）提出的识别隐喻、转喻、字面义，反常表达的隐喻理解模型 Met5 系统；Martin（1990，1992）的识别

和解释常规隐喻的 MIDAS 系统；Gentner et al（1988）的结构映射引擎（Structure-Mapping Engine）（SME），Holyoak & Thagard（1989）的 ACME 隐喻分析模型，以及 Veale（1995）的 the Sapper 模型。上述模型主要以规则推导为主，它们的缺点是大多数计算模型都是基于手工编制的知识库和规则的优选样例。针对这一欠缺，Mason（2002，2004）提出了一种利用大规模语料动态发现概念建立隐喻映射的模型——CorMet 系统，这是一个完全以统计为手段，基于大规模语料库提取的隐喻分析模型。CorMet 系统虽然能自动获取谓词的优先选择，但是他的理论思想和 Fass（1988，1991）的思想基本一致，即基于语义优选的策略，所不同的是 CorMet 系统主要是语料库驱动，这样就避免了基于优选语义方法中手工构造知识库的不足，但是，CorMet 所能处理的隐喻主要是隐喻的目标域和源域分属两个具体的领域，处理的范围有一定局限。关于模型的设计思路在 2.2.2 节中将会有进一步论述。另外，上面多数模型处理的是以下两种隐喻类型。一种是对“A 是 B”的识别和分析。例如：

⑫ 郎平是个铁榔头。（摘自俞士汶，2003b）
⑬ 书是人类进步的阶梯。

该种隐喻类型往往会涉及翻译的问题，从形式上也较容易判定，因此一直是学者们关注的重点。除此之外，还有一种“n + v”或“v + n”的隐喻表达也备受瞩目。例如：

⑭ 汽车喝汽油。（摘自 Fass，1991）

系统通过动词“喝”优先选择“动物”做主语，“液体”做宾语，语义表达为（animal，drink，liquid），动词“喝”和名词“汽车”存在优选语义冲突，进而识别为隐喻表达。隐喻解释通过搜索 Met5 知识库的优选论元和事实论元的上位词语的语义向量来实现（Fass，1988）。

Fass（1988，1991）、Mason（2002）的计算模型把动词和名词的语义冲突作为识别隐喻表达的关键。按照他们的理论，我们所提出的“张开理想的翅膀，奔向知识的海洋”中，“张开/v　翅膀/n、奔向/v　海