面向信息处理的词汇语义研究中的若干问题

董振东 董强

（载于《语言文字应用》2001 年第三期，pp.27-32）

提要 本文概述了国家社会科学“九五”重大项目“信息处理用现代汉语词汇研究”中的子课题“现代汉语知识词典的建立和词汇内部语义网络描述”取得的成果，着重讨论了在研究过程中曾面临的一些策略性和方法论的问题。笔者把它们归纳为五个关系：知识获取和表达的深与浅的关系、语义知识和世界知识的关系、分类和属性标注的关系、知识的自动获取和人工获取的关系、知识体系的完备性与实用性检验的关系。

Some Problems in Study of IT-oriented Lexical Semantics

Zhendong Dong Qiang Dong

Abstract This paper outlines the achievements in the study of “construction of a knowledge dictionary of contemporary Chinese and description of semantic structure of words ”, a sub-project in the 9th five-year plan program “modern Chinese Vocabulary studies in Chinese information processing”. The paper discusses some problems of research strategy and methodology. They are (1) the depth in knowledge acquisition and representation, (2) semantic knowledge and world knowledge, (3) classification and semantic feature tagging, (4) automatic acquisition and human acquisition of knowledge, (5) completeness of a knowledge system and evaluation in its application.

笔者于 1998 年开始承担国家社会科学“九五”重大项目“信息处理用现代汉语词汇研究”中的子课题 – 现代汉语知识词典的建立和词汇内部语义网络描述，这给笔者一个涉足当前研究热点的好机会。这项研究包括两个部分：第一，分析和提取语言的意义的最基本的元素，即义原。该子课题负责人董振东在 1988 年提出，利用中文词语的意义结构的特点，将是分析和提取义素的最好、最方便的方法。然后利用提取的义原来建立面向信息处理的《汉语知识词典》，这样做也是对于义原的检验。本项研究提取了 1503 个义原，以及 71 个词语间的动态角色关系和动态属性。迄今为止，研究者用它们来标注中文和英文各 70000 个词语，证明这些义原以及动态角色关系和动态属性是正确的、可靠的、有生命力的，从而建立了《汉语知识词典》。该词典目前被公认为规模最大、收录词语最多且最新、为词语提供的信息内容最丰富的，面向信息处理的汉语知识词典。能够这样完整地系统地提取义原，并加以科学地利用，建立如此规模的知识词典，是革命性的。第二，在上述义原以及动态角色关系和动态属性的基础上，探索中文结构的构造模式。在经过大量的实际语言素材的调查研究，并得到了香港科技大学研究项目 HKUST 6149/98E 的支持之后，研究从原来拟定的局限于词语内部语义关系扩展到更大范围的短语，原来拟定的局限于语义关系被更深的信息结构关系所取代.迄今为止得到了 271 个结构模式，从而建立了一个《中文信息结构库》，它包含 11000 个词语的实例。

边研究，边应用是本项研究的的一个突出的特点。得到应用和推广是研究成就的最好证明。迄今为止采用《汉语知识词典》和《中文信息结构库》的机构已有中港台三地、新加坡、美国、加拿大等多个大学和科研机构。应用可分为三类：一是基于《汉语知识词典》和《中文

信息结构库》文本语义关系的标注，如香港科技大学的研究；二是对《汉语知识词典》和《中文信息结构库》的学习和研究，如台湾中央研究院资讯所、新加坡南洋理工大学、美国马里兰大学等；三是基于《汉语知识词典》和《中文信息结构库》的应用技术，如句法结构或语义排歧，信息过滤、智能检索中的自然语言接口等。鉴于上述，本项研究获得了学术界很高的评价。香港科技大学颜国伟博士指出："它的理论基础：知网的知识表达模式是针对计算机的信息处理特点而制定的，理论水平超过面向人而设的 **WordNet**。提供一直接处理语义，绕过汉语语法分析的门径。经过两年的研究证实，词汇内部的语义网络描述适用于词汇之间的语义关系描述。由小观大，对词汇内部语义关系的把握直接可以取得对句子意义的理解。这是此研究最具慧识之处，对突破目前自然语言理解的瓶颈影响深远。"

这段时间以来，我们的用户和读者积极地参与论坛，经常直接来信与我们探讨种种问题，我们也在讲学或研究合作过程中面对面地和专家学者们进行广泛的讨论。我们想在这里对这些讨论做一个总结，把一些最重要或谈论最多的问题摆出来，其中有的也是我们自己的困惑。望各位专家学者不吝指教。我们想这些问题可以归纳为以下五种关系。

1. 知识获取和表达的深与浅的关系

这是一个关于研究的"度"的把握。知识获取和表达得深一些好还是浅一些好？有人对我们说，"你们研究得很深，但好像是太深了。"这也是我们经常考虑的问题，我们也为此经常留意用户的反映。是深还是浅，我们的历来主张是：深研究，浅应用。这就是说，研究要深，但应用时应该考虑到系统工程的因素，尽量采用成熟的、已经经过试验的技术，而不要一味追求高精尖。这有点好像老师教学，老师准备的是一桶水，才能给学生一碗水。另外，应用者还要善于从深度研究的成果中去提炼适合自己的精华。我们也注意到部分用户确实没有用得像我们设计的那么深（如果更深效果会更好）。

研究深点好，但绝不是可以凭着性子为研究而研究。研究必须考虑到应用，但也不能过分功利主义，只追求眼前利益。研究也有一个带动应用的任务。当前的研究应该深些还是浅些，我们也可以近十年来的动向看出一些端倪。这些年来国内外都在语义研究上下功夫，建立规模不小的语义词典或知识词典。例如普林斯顿大学的英语 WordNet，**Fillmore** 领导开发的研究英语动词的 **FrameNet**，微软的 MindNet，在欧洲有基于 WordNet 的 EurowordNet，日本有电子辞书研究所（EDR）的日语和英语的概念词典，还有美国 HPKB(High Performance KB)等等。现有的语言信息处理系统种类不少，这些年来质量也改善了许多。但是也应该坦白承认问题也不少，实难令人满意。存在的问题归根到底就是一点：智能太低。最近我们对海内外的一些主要语言信息处理系统进行一番测试比较，其中包括音字转换输入系统、搜索引擎、英语拼写语法检查系统、中文拼写语法检查系统、英汉/汉英机器翻译系统等。可以这样肯定，这些系统存在问题有一个明显的共同特点：都卡在"意义"的这道坎上。让我们看看一些系统。

（**1**）音字转换输入系统，

音字转换输入系统，其中不乏很优秀的，但可以说它们"成"在统计上，"败"也在统计上。例如，下面是某系统给出的结果：

"首先是正常的投资历年（理念）和价值观念遭到了践踏和破坏，市场上短期投机分为（氛围）越来越重"。

这里"理念"和"氛围"都是第二个候选。不难想象如果没有其他有效策略作为补充，类似的捉襟见肘的困境是很难摆脱的。

（2）搜索引擎

眼下的搜索引擎不能不说是低智商的，首先它为用户提供的询问方式实际上太过简单，现在

基本上是“关键词语”；其次“关键词语”又多停留在字符串级，因此用户键入“北大”来搜索时，系统将给出包括“北大荒”、“北大西洋”、“东北大学”、“西北大盗”等；又如当用户用“巴以冲突”去查询时，系统是不会把带有“以巴冲突”的网页与带有“巴以冲突”的网页同时提供给用户的。系统无从知道“巴以冲突”或“以巴冲突”是相同的。这样的状况应该有所改变才好。

（3）机器翻译系统

我们曾做过这样的测试，从网上选取一段英文原文的新闻报道，让两个在线的英汉机译系统翻译来比较它们的译文质量。下面是原文：

“SEOUL, South Korea -- North Korea has accepted the idea of working toward restraint in its missile program, U.S. officials said Tuesday, citing progress on a critical issue dividing the two countries as they explore reconciliation after 50 years. Secretary of State Madeleine Albright ended her historic talks with North Korean leader Kim Jong Il struck by the improbability of it all -- a cordial visit to a Stalinist land that the United States until recently called a rogue state.”

系统 A 的译文如下：

“星期二,美国官员说汉城,--韩国北朝鲜已经接受朝着约束它的可投射的程序用工作的观念,引证有关一分 2 国家,当在 50 年以后他们探查和解时的关键问题的进展.国务卿玛德琳 Albright 结束--她的历史性和被它的不大可能性完全击打的北方朝鲜人领袖 Kim 青年男奴 Il 的谈话一次热情友好到一美国直到最近认为一凶猛的状态是它的斯大林主义者陆地观光.”

系统 B 的译文如下：

“SEOUL，韩国--北朝鲜已接受朝着其导弹计划中的克制工作的想法，美国官员在星期二说，在划分两个国家的关键问题上将进步引用为他们在 50 年之后探索和解。国务卿玛德琳奥尔布莱特结束她与北朝鲜领导者金正日的具有历史意义谈话按它的不可能性打击所有--对斯大林主义者陆地的甜饮料访问直到最近呼叫的流氓状态的美国。”

据报道，系统 A 是外国公司开发的，2000 年推出的；系统 B 是国内公司开发的，它的核心技术大约是在十年前开发的。比较这两个系统，我们至少可以得出一个结论，那就是实用的机器翻译系统的技术还有很大的改进的空间。

2. 语义知识和世界知识的关系

我们在研究和开发《汉语知识词典》的过程中常常遇到的问题之一是如何把握语义知识和世界知识的界线，如何保证我们能始终坚持建立世界知识词典这一目标。我们任何时候都牢记：我们研究的对象是世界知识，我们的目标是建立知识词典而不是语义词典。首先，我们自己必须明确语义词典与知识词典的区别。语义词典或者义类词典和世界知识词典（如百科全书）的主要区别，简单地说有两个方面。

第一，语义词典或者义类词典描述的主要是狭义的语义学的知识。传统上这样的词典的主要用途是为人们提供写作时选择词语用的。因此它们通常包含着有关词语的同义、反义等，甚至包括关于这些词语的细微的用法上差异的知识等。世界知识词典描述的主要是有关客观世界的各种知识。它们不会仅仅提供对于某一事物的定义（这是普通的语言词典做的），而是会提供关于事物的更加详细的知识，就像普通的百科全书做的那样。

第二，从词语条目的选择和收录上，语义词典或者义类词典通常以语文词语为主，通常较少涉及百科类词语，因为它的着眼点是词语本身的知识。而世界知识词典的选择和收录对象将是百科事物，因为它的着眼点是词语所指称事物的知识。

总之，语义词典或者义类词典教会人们如何运用词语；而世界知识词典教会人们懂得世界。我们自己还有一个必须面对的问题，这就是我们研究和开发的不是面向人的知识词典，而是面向计算机的知识词典。这样的词典必还须在一定程度上能够教会计算机懂得和计算知识，或者一定程度的推理。以我们的《汉语知识词典》为例，只要借助于一小段程序，你的计算机将可以回答你这样的问题："人们到哪里去买书"，"我们可以从储蓄所贷款吗"等等。这些显然不是语义词典或义类词典要解决的问题，即便它是在线的。

我们是通过什么方法来建立世界知识的呢？我们的基本做法是：静态地、孤立地对概念（由词语表现）逐一进行义元标注，然后期待通过概念的内在联系来动态地、综合地反映它们的关系网络。试看下面各个概念的标注：

'博士后'　标注为：human|人,*research|研究,*study|学,education|教育
'研究所'　标注为：InstitutePlace|场所,*research|研究,#knowledge|知识
'论文集'　标注为：publications|书刊,#research|研究
'学术成就'标注为：result|结果,#succeed|成功,#knowledge|知识,#research|研究,#study|学
'治学'　标注为：research|研究,content=knowledge|知识
'结题'　标注为：cease|停做,content=affairs|事务,#research|研究

这里我们可以看到，概念"博士"和其他几个概念"研究所"、"论文集"、"学术成就"、"治学"、"结题"等都是被静态地、孤立地标注的。但等到我们标注完毕之后，我们发现它们居然有着一种故事性的联系。这故事是这样的："一个博士在研究所从事研究工作，取得了学术成就。。。"另外，这些概念的每一个都又有各自的上下位关系、同义关系等等，再者，它们的定义中的其他义元又会与其他更多的概念发生联系。如果我们把"knowledge|知识"作为关键词在《知网》浏览器中按"feature"模式查一下，我们会得到近 150 组相关的概念群，约 500 以上个概念。这样一来，就形成了一个更大的概念关系网络。

3. 分类和属性标注的关系

说起词汇语义研究、语义词典建立，人们常常会谈论分类的问题。关于是用分类的方法还是用属性标注的方法，似乎还有一些不同的看法。我们来谈谈我们的做法和体会。先说明什么样的方法是分类的方法，什么样的方法是属性标注的方法。用实例来说明会简单些。 例如，《同义词林》用的是分类的方法；而《知网》用的则是属性标注的方法。对于每一个词语或概念而言，分类的方法给出的是一个"单一标记"；而属性标注的方法给出的是一个"复杂特征标记"。分类的方法的着眼点是上下位的树状关系；而属性标注的方法的着眼点是多元、多层次的网状关系。

我们的经验告诉我们，如果你的研究是面向信息处理的，如果你的目标是要建立知识词典而不只是义类词典，那么属性标注的方法应该会比分类的方法更好。

4. 知识的自动获取和人工获取的关系

有人告诉我们说，"你们的工作是为概念建立它们之间的种种关系，你们为什么采取利用一些电子词典和百科全书，自动地获取你们想建立的关系？美国有公司就是这么做的。"首先，在十年前，我们就是想这么做，也没有这种条件。其次，即便有了条件，我们也不会那么做。我们认为面向信息处理的知识词典的研究基础还太不成熟，有许多基本问题还没有解决，例如，应该建立什么类型架构。以 WordNet 为例，除了部分形容词有作为"值"指向与之相对应的名词外，它没有揭示更多的跨词类的关系，而只揭示属于同一词类的关系如上下位、同义、反义、整体与部件等，这样的架构应该算是充分还是不足呢？再者，电子词典和百科

全书给出的定义是否会那么严谨，可以自动抽出概念的上下位关系呢？从大规模的真实语料中自动获取知识，它的“利”是显而易见的，它的语料是“真”的，就是我们信息处理是实际遇到的，它的涵盖广度是人工所无法企及的，但是它的“弊”也是明摆着的，它难免会遇到语料的稀疏、偏向的局限，也难免会产生一些垃圾。人工获取的好处是较为精细、容易经过推敲，但是它难以避免人为的主观因素，难以避免涵盖面小，同时如果架构复杂、信息量和数据量大，那么保证一致性将是困难的，最后就是它的完成将是费时、费力的工程。我们一直认为自动获取作为一种辅助的方法或用来作为对人工获取的检验，那将是非常好的。其实人工的还是自动的，两者始终是相辅相成，而不应该是对立的。但我们相信：愈是深层次的知识，人工的成分会更大。

5．知识描述体系的完备性与实用性检验的关系

大家知道，《知网-知识词典》的知识描述的基本方法是采用对概念进行属性描述的方法而不是分类的方法；其基本架构是由 1503 个义元和 71 个动态角色与属性作为基本单位，并借助于知识描述语言加以表达的。曾有人问，“你们这 1500 个义元是怎么得出的？参考了那本义类词典？” 我们曾做过说明，我们的义元是从 4000 多个汉字的义项中提取的，现在又经过了近 7 万个汉语词语和约 6 万个英语词语的实际考核与调试最后确定的。 也有人问，“这 1503 个义元和 71 个动态角色与属性是否足够？”我们的体会是：知识、意义是人类长期生活、生产活动积累起来的精神产品，无论对于个人或者对于全人类都是没有止境。语言是人类思维的工具，也是人类思维的载体。我们不认为会有任何一种描述架构是可能将世界知识毫无遗漏地覆盖住的。这至少是我们的认识。其实，对于面向信息处理的知识系统是好还是差，最简单的检验方法就是将其应用于真实文本的处理。总之，我们觉得未经过实际的考核，要就一个知识描述体系的充足与否下个结论是很难的。如果有人向我介绍一个机器翻译系统是如何设计的，设计的理念是什么，里面包含那些子系统等等，而没有试着让它运行一下，翻译一些句子，就要我们作出评价，我们是做不到的。

参考文献

[1] 陈小荷，一个面向工程的语义分析体系，《语言文字应用》第2期，1998
[2] 董振东，逻辑语义及其在机译中的应用，中国的机器翻译，1984
[3] 董振东，机器翻译的启示和挑战，上海科技翻译 第1期（总第11期），1988
[4] Dong, Zhendong, Knowledge description: what, how and who, Manuscripts & Program of International Symposium on Electronic Dictionary, Tokyo, 1988
[5] 董振东，语义关系的表达和知识系统的建造，《语言文字应用》第3期，1998
[6] 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996
[7] 汉语语义学，贾彦德，北京大学出版社，1999
[8] 林杏光，词汇语义和计算语言学， 语文出版社，1999
[9] 俞士汶等，现代汉语语法信息词典详解，清华大学出版社，1998
[10] 徐通锵，《马氏文通》和中西语言学结合的道路，’98现代汉语语法学国际学术会议论文提要集，1998
[11] 张普，主持人的话，《语言文字应用》第2期，1998

在线资源

http://www.keenage.com
http://www..cogsci.princeton.edu/~wn