从语义关系的复杂性看语义词典建设

严灿勋，刘慧敏

解放军信息工程大学外语系 河南郑州 450001

Email: yancanxun@126.com, liuhuimin126@126.com

摘　要：语义理解已经成为计算机处理自然语言的瓶颈问题，计算机实现自然语言理解，离不开语义词典。虽然目前已经开发出来一部分语义词典，但是，从词汇语义学、句法语义学和篇章语义学来分析，这些语义词典并不能全面反映语义关系。在帮助计算机理解自然语言方面，我们必须综合语言学、计算机科学和认知科学，不断探索研究。

关键词：语义关系，语义词典，语义学，自然语言理解，认知

Viewing Semantic Knowledge-base Construction from a Complex-Semantic-Relation Perspective

Yan Can-xun, Liu Hui-min

Foreign Languages Department of PLA Information Engineering University, Zhengzhou 450001

Email: yancanxun@126.com, liuhuimin126@126.com

Abstract: Understanding natural languages at the semantic level has become a bottleneck problem in Natural Language Understanding. The computer depends on semantic knowledge-bases to "understand" natural languages. Although there are already some semantic knowledge-bases, they cannot fully clarify all kinds of semantic relations according to the theories of lexical semantics, syntactic semantics and discourse semantics. To help the computer "understand" natural languages better, we must conduct comprehensive researches in semantics, computer science and cognitive science.

Keywords: Semantic Relations, Semantic Knowledge-base, Semantics, Natural Language Understanding, Cognition

1　前言

语义理解已成为计算机处理自然语言的瓶颈问题。针对万维网缺少语义信息、不能进行语义检索等问题，万维网发明者 Tim Berners-Lee 1998 年提出语义网络构想，尝试利用本体(Ontology)表示 Web 知识，建立不同形式数据之间的各种联系，使 Web 信息具有计算机可以理解的语义，使计算机在语义网络上实现描述、猜想和推理的能力。除语义网络的本体以外，其它能够体现语义关系的主要产品是语义词典。本文拟从语义关系的复杂性探讨词义词典的建设问题。

2　语义词典简介

语义词典，或者称语义知识库，是帮助计算机理解自然语言的一个媒介和桥梁，也是让计算机逐渐"智能"起来的一个物质前提（于江生等，2003）。例如，下列例句中 "仪表"的消歧问题：

例 1 她的仪表很精密。

例 2 她的仪表很端庄。

例 1 和例 2 句法结构完全一样，对消歧无贡献。如果语义词典提供如下信息："精密（precise）"的主体是"器具（instrument）"，"端庄（decorous）"的主体是"品貌（appearance）"；那么，计算机就能够判断例 1 中的"仪表"是"instrument"，例 2 中的"仪表"是"appearance"(俞士汶,2009)。

在自然语言处理中，信息提取、词汇语义排歧、机器翻译等通常都需要一部能够表达语义关系的知识词典的支持。

国外被广泛使用的语义词典有 WordNet、FrameNet 等。WordNet 是基于心理语言规则的词典，规定了名词、动词、形容词、副词的语义知识表示规范，基本单元是同义词集合（Synset），集合中的元素相互构成同义关系，集合间有上下位、反义、整体-部分等关系。WordNet 对词类间的组配约束关系（比如动词跟名词间的组合关系）涉及很少。FrameNet 的理论基础是 Fillmore 的框架语义学，其组织词汇语义知识的基本手段是框架，每个框架包含若干框架元素，框架元素比格语法的语义格更具体、更细；最重要的是，语义格是相对于所有词汇而言的，高度抽象和概括，而框架元素是相对框架而言的，是框架的构成成分。FrameNet 的每个框架都包含一批词语，理解这些词语的词义，必须以理解它们所在的框架为前提，这些词语的共性（尽管句法上可能分属不同词类）构成了同一个语义框架。框架内部的关系是单网关系，框架与框架之间的关系是多网关系。为了表述的简洁，框架之间也可以有继承关系，另外，对于框架中的动词，FrameNet 数据库还描写了各个框架元素（角色）在表层句子结构中所占据的句法位置。(詹卫东,2003)

国内著名的语义词典有梅家驹的同义词词林、董振东的知网（HowNet）、 北京大学计算语言学研究所的现代汉语语义词典（SKCC）（王惠等，2003）和中文概念词典（CCD）（于江生等，2003）等。同义词词林和 WordNet 一样是义类词典，所收词语按词义分类编排，一组同义词编为一个词群，多义词分别收入不同词群。同义词词林把汉语词汇按语义分成三个层次的类。其中大类层次共 12 类，大类编号为大写英文字母，中类编号用后跟的小写英文字母表示，小类编号为末尾两位阿拉伯数字。同一编号的各个词有同义词、相关词等标记，这样从词语的编号和标记就可以方便地计算词与词之间的语义距离。知网是一个以汉、英词语所代表的概念为描述对象，以揭示概念之间以及概念的属性之间的关系为基本内容的常识知识库。知网中表示概念的基本元素是义原，一个概念可有多个义原。知网通过概念之间的关系以及概念的属性之间的关系构成一个网状知识系统，这些关系有上位、下位、同义、反义、对义、相关、部件-整体等等关系。现代汉语语义词典在汉外机器翻译背景下开发完成，跟汉外机器翻译的实际需求结合紧密，它给出词语的词类、语义类，以义项为单位描述配价信息和多种语义组合限制。中文概念词典是一个 WordNet 类型的汉英双语语义词典，是全球多语种 WordNet 资源建设的重要组成部分。中文概念词典规格上与 WordNet 兼容，用同义词集描述概念，用概念间关系描述语义，便于语义关系的表示和检索，方便实现语义距离计算，有利于概念分级扩展。 (詹卫东,2003)

这里介绍的六种语义知识库是目前国内外比较典型的知识库，它们已经被应用在自然语言处理的多个领域。当然还有其它一些世界知名的语义知识库，如 HNC 等，这里暂不介绍。

3 语义关系的复杂性

各类语义词典提高了计算机辨别歧义、理解语义的水平。然而，由于语义关系的复杂性，目前还没有任何的语义词典能够帮助计算机达到普通人理解语义的水平。

语义学是研究语言的意义的学科，主要研究语义的各种性质、类型、语义关系、语义的结构

和功能，以及语义的形成和演变等等。现代语义学已经成为研究语言所有层次和全部单位的意义方面的科学，发展并分化为语音语义学、构词语义学、形态语义学、词汇语义学、句法语义学和篇章语义学等分支学科(杨喜昌,2005)。为帮助计算机理解语言意义，还出现了专门针对自然语言处理的本体语义学。尽管计算机处理语义的能力在不断增强，但是目前，词汇语义学、句法语义学和篇章语义学中的许多语义关系问题仍然阻碍着计算机对自然语言的理解。

3．1 词汇语义学中的语义关系

词汇语义学中，语义关系是指词与词之间的意义关系。词义的最小构成单位是义素，也是词义的区别特征。义素分析就是在同一语义场中找出共同义素与区别义素。语义场是通过不同词之间的对比，根据词义的共同特点或关系划分出来的类。根据成员相互之间的关系，语义场可以分为类属、顺序、关系、同义、反义等义场。词义中同概念有关的意义部分叫做理性义，是词义的核心。义项是词的理性意义的分项说明。只有一个义项的是单义词，有两个或两个以上义项的是多义词。多义词的义项至少有一个是基本的、常用的，叫基本义，其他的义项是由基本义发展转化来的，叫转义，包括引申义和比喻义。词还有附属的色彩意义，叫做附属义。同义词的存在是词汇丰富的标志，但是意义完全一样，内涵没有差别的同义词很少。等等。(黄伯荣等, 2002)。

可以想象，要想穷尽一门语言中全部词汇的全部词汇语义关系是何等困难！上文介绍的语义词典已经受到国内外学者普遍认可，但是，它们对词汇语义关系描述并不完全，例如很少涉及词的附属义，主要针对实词，虚词的词汇语义关系没有或描述不够。另外，尽管每个语义词典都试图刻画词汇语义关系，但它们的指导理论不同，呈现形式不同，有的差异还很大：WordNet 以同义词集为基础；FrameNet 以语义框架为基础；同义词词林在同义词、相关词的基础上加上了分类编号；知网是以义原分析为基础，等等。

3．2 句法语义学中的语义关系

根据句法语义学，语义关系还应该体现句子的语义组合关系，主要指句子成分所表示的动词的“及物性”关系，如动作与施事、受事、结果、工具、处所等的关系，等等。这样的语义关系和句法结构关系相互联系但彼此独立。二者的联系表现为：语义关系是词语的概念意义间的关系抽象概括的结果，但要在一定的句法结构中才能表现出来，如“张三叫李四”，句中施事和受事离开句法结构就不存在。语义关系和句法结构关系又彼此独立，表现为：(1)同一种句法结构关系可以表示不同的语义关系。例如同是主谓结构，可以表示“动作—施事”(“他喝了”)，“动作—受事”(“酒喝了”)，两可（“鸡不吃了”——歧义，“鸡”可作施事，也可作受事）；(2)同一种语义关系可以用不同的句法结构关系来表示，如“动作—受事”可以用动宾结构(“喝了酒”)，也可以用主谓结构(“酒喝了”)，还可用其他结构(“把酒喝了”)；(3)句法结构关系成立，但语义关系可能不成立，如“喝了桌子”，语义荒谬，但仍有动宾关系。不过如果语法结构关系不成立，语义关系也不成立，如“酒了喝”；(4)句法结构关系由直接成分表示出来，而语义关系可以由直接成分表示，如“喝了酒”，也可以由间接成分表示，如“把酒喝了”中动作—受事关系。所以语义关系与句法结构关系不能混淆。有学者指出，上述由句子成分表示的语义关系并不与整个句子的语法意义直接关联，是低层次的语义关系；此外还有高层次的语义关系。低层次的语义关系在基本句及其变换句里是不变的，如(a)“床上躺着病人”与它的变换句(b)“病人躺在床上”都有同样的“动作—施事”关系；但是(a)还可抽象出“存在”的语义关系，即“床上有病人”；(b)还可抽象出“存在的位置”的语义关系，

即“病人在床上”；同时，跟(a)同句法结构的一类句子也都表示“存在”即“有”的语义关系，跟(b)同句法结构的一类句子也都表示“存在的位置”即“在”的语义关系。这种语义关系已跟整个句法结构关系联系起来，句法结构改变，它必然改变。此即高层次的语义关系。语法学对语义关系的认识是逐步加深的。传统语法早期常把语义关系和句法结构关系混同起来，用前者代替后者分析句子；将语义关系引入语法研究是语言学的新发展，对歧义结构和同义结构的研究就是重要成果。例如“鸡不吃了”，用施事和受事的观念分析“鸡”就能对这个歧义结构作出解释。但是只有低层次的语义关系还不能对句法结构作出充分研究，还需要联系高层次的语义关系。(张清源, 1990)

上述语义词典中，有的根本没有涉及句法语义关系，如 WordNet、同义词词林，有的涉及了，如 FrameNet、知网、现代汉语语义词典，但是它们对句法语义关系描述的深度和广度如何？应用效果如何？现代汉语语义词典最初是针对机器翻译建设的，主要设计者刘群在他的“计算所与北大往事回顾”中坦言：“虽然我们已经完全转到了统计机器翻译这个方向，但我并没有像一些纯粹的经验主义者一样，对语言知识在机器翻译翻译中的应用失去信心，而是一直坚持把一些语言学的知识引入到统计方法中，并获得了某种程度的成功。……汉语句法分析器在面对真实语料的时候正确率能够达到 60%以上，这对于一个基于规则的系统来说是非常不容易做到的。”言外之意，语义词典在基于规则的机器翻译中表现不尽如人意，某种程度上比不上统计机器翻译。

3．3 篇章语义学中的语义关系

篇章语义学重点解决篇章的意义连贯问题。篇章语义学语义关系是指句子、段落之间在意义上要构成顺承、递进、因果、并列、转折等逻辑关系，使之有机地联系在一起，而不至于一盘散沙(王全智,2001)，上下文的指代关系也属于篇章语义学。篇章语义学认为,歧义现象只有在进入篇章后才会消失。因此,在翻译过程中,有时为判明某个词的含义,需要分析整个段落或篇章。例如，在句法语义关系中有歧义的“鸡不吃了”一般不会在篇章中产生歧义。另外，篇章语义分析准确度的提高也会直接提升智能检索、自动文摘、自动过滤等的质量。

HNC 在篇章语义理解方面有突出表现，它能够做到在语境层面上进行智能检索。HNC 结合了语言学、认知科学和计算机科学的知识，对概念基元进行了复杂的分类和形式化处理，在理论上，HNC 概念基元知识库、句类知识库和其它知识库结合起来后能够构成一个高度形式化的概念层次网络，使计算机在语义理解的基础上对语言信息进行智能化处理。2009 年．在中文信息学会句法评测(CIPS-ParsEval-2009)中，中科院声学所中科信利 HNC 语言处理团队获得汉语事件描述单元识别第一名、汉语功能块分析第二名的佳绩。这次评测有来自美国、欧洲、中国大陆、香港和台湾地区的共 24 支队伍参加(缪建明, 2009)。HNC 获得这个成绩，证明了其理论的先进。

3．4 认知的局限性

对语义关系的理解还受到认知的制约。这种制约表现在两个方面。

一方面是人对语义关系的认知还不够。例如：Fillmore 发展了格语法，提出了框架语义学理论，才有了 FrameNet 的建设。但 FrameNet 对语义关系的表述还不够，没涉及高层次的句法语义关系。黄曾阳潜心研究 8 年,深入挖掘汉语特点，以意义表达和语言理解为主线，建立了一种模拟大脑语言感知过程的自然语言表述模式和计算机理解处理模式,推出了 HNC 理论(梁捷, 2005)。该理论指导了 HNC 知识库的建设。不过， HNC 理论自诞生起，就一直在实践中不断完善，还没有达到完美，也不能说它就最优。人类仍须通过实践不断提高对自然语言的认知水平。

另一方面，计算机对语义关系的理解受到计算机自身认知模式的制约：计算机只擅长逻辑运算，根本不能进行图式思维。而人在理解语言时能够图式化语言描述的对象以及图式化话语情境，帮助理解语言。因此，汉语中很多不符合逻辑的言语对人来说并不难理解，例如：救火、晒太阳、养病、一匹马骑两个人……人可以借助图式思维和经验轻松理解此类意合结构，但是计算机单凭“动宾结构”等逻辑关系却不能分辨正确的语义，必须通过在知识库中修改或增加规则来帮助判断。这就是说，我们要帮助计算机增强逻辑推理能力以弥补其图式思维空缺。矛盾是：语言中的特例很多，规则的修改和增加何时是尽头？修改规则还可能出现“翘翘板现象”：修改后的规则在这里适用了，在原来适用的地方又出现问题了。HNC 的句类分析深入到高层次的句法语义关系，可能是一个较好的解决方法。由于该理论的复杂性，三言两句实在不能讲清楚，这里就不多谈了，感兴趣者可以参考黄曾阳（1998）、苗传江（2005）。另外，理论实践证明，统计方法对此类问题的解决能力达到一定水平后，再想提高就会非常困难。

因此，计算机对语义关系的理解最终还要通过人提高对语义关系的认知水平，建设出合适的语义词典，才能逐渐解决。

4 对语义词典建设的建议

建设语义词典的目的是将语义进行形式化处理，方便机器处理自然语言中的语义关系。根据前人的经验,我们可以按以下原则建设语义词典。

第一，应该根据语义词典建设目的确定建设方案。WordNet、FrameNet、知网、HNC 等对语义关系的描述方法各不相同，但都有可取之处，在建设词义词典时，应根据建设目的选择参考。

第二，对语义词典而言，形式化程度越高，越有利于计算机的逻辑运算，所以应该尽可能地使语义词典形式化。形式化的方式主要是通过基元集（set of primitives）来描述各类语义属性及语义关系。后面将举例说明。

第三，语义词典应该基于丰富的语料。语料库能够帮助提高词项设置的科学性。有些语义词典仅通过合并一些电子词典的内容来充实内容，这样的词典既难跟上语言的发展，词项设置也缺乏充分的理据。知网选择词语的依据是建立于 4 亿字汉语语料库出现频率形成的词语表，而不是仅依据某一本现成的词典(董振东等, 2010)。山西大学在构建汉语的框架语义网络时，就是通过对大规模真实文本句子的分析和标注，提取核心动词的各种句法语义搭配形式，希望从大规模框架语义标注语料库中获取自动句法语义标注的训练数据，并进一步实现篇章的语义理解（王素格等, 2007）。有些好的知识库以互联网为语料库，不断填充新词，这也是一个很好的做法。

第四，语义词典的建设，需要团队协作。每个好的语义词典，都需要多人多年的工作积累。知识博大精深，如果决心开发一个实用的语义词典，一定要看有没有能力组织一个团队共同协作。

第五，要充分借助各类工具软件帮助建设语义词典。现在，完全靠手工添加来建设语义词典已经很不现实。经常可以借助现有的工具软件帮助建设，必要时可以自行设计工具软件来提高知识库建设效率。例如，王兰成（2009）尝试先用 Protégé 构建本体知识库，在本体构建完成后，通过 Protégé 的本体转换功能，将 OWL 格式的本体文件转换为关系数据库。

下面简单介绍知网和 HNC 在语义词典形式化方面的做法。

（1）知网

知网先初步确定一批义原，形成了一个基本的标注集,也即基元集，如下例中的{人，医，职

位，医治，...}，然后用这些义原来描述概念之间的关系以及属性与属性之间的关系。知网语义词典中每一个词语的概念及其描述形成一个记录。每一种语言的每一个记录都主要包含4项内容。其中每一项都由两部分组成，中间以“=”分隔。每一个“=”的左侧是数据的域名，右侧是数据的值。它们的排列是：W_X= 词语；G_X= 词语词性；E_X= 词语例子；DEF= 概念定义。其中，X为C时表示中文，X为E时表示英文。下面是关于“医生”、“患者”的概念定义。

例3　　医生：DEF={human|人:domain={medical|医}, HostOf={Occupation|职位},
{doctor| 医治:agent={~}}}

例4　　患者：DEF={human|人:domain={medical|医},{SufferFrom|罹患:experiencer={~}},
{doctor|医治:patient={~}}}

在对“医生”和“患者”的定义中，“人”是它们的共性；“医生”的个性是他是“医治”的施事（agent），“患者”的个性是他是“患病”的经验者(experiencer)，也是“医治”的受事(patient)。在“医”这个领域（domain）还有其它一些概念，如医院、医药、医药费、疾病等，它们通过义原和语义角色，相互联系起来，不但表达了各自的概念属性，同时也表达了概念与概念之间、概念的属性与属性之间的关系。(董振东等, 2010）

（2） HNC

HNC理论是一个非常庞大的系统，这里只是介绍其片鳞只甲。

HNC首先把概念分为抽象概念和具体概念，对抽象概念设计了三大语义网络：基本概念、基元概念和逻辑概念，分别用j、φ、l标记。语义网络每层有若干节点，称为概念节点。每层概念节点用数字0到13依次标记，其中10到13用a,b,c,d表示。语义网络中的任一节点都可以用从最高层开始到本节点结束的一串数字（即一条路径）来惟一地确定和指称。基本概念语义网络共有9个一级节点：j0(序及广义空间)；j1(时间)；j2(空间)；j3(数)；j4(量与范围)；j5(质与类)；j6(度)；j7(基本属性(客观的))；j8(基本属性(含主观评价))。基元概念语义网络共有14个一级节点，分为两类，0-5号是一类，称为主体基元概念，6到d号是一类，称为扩展基元概念。主体基元概念由六个一级节点组成（φ省略）：0（作用）；1（过程）；2（转移）；3（效应）；4（关系）；5（状态）。等等。看下面的概念实例：“增加(v341)”,v是动词标志，3是“效应”，34是其二级节点“扩展与缩小（量的效应）”，341是其三级节点“扩展”。“减少（v342）”只是在三级节点上与上例有差异，1和2分别表示对偶双方，1是积极的，2是消极的。等等。另外，HNC建立了语句的语义表述模式，穷尽地发现了自然语言句子语义的57组基本表示式。例如，“X20J=X2B+X20+XBC”是其中一个句类“一般反应句”，其构成是“反应者+反应+反应引发者及其表现”，实例如“张先生（X2B）喜欢（X20）李小姐的个性（XBC）”。这样，HNC知识库以句类知识为纲领，以表达语义网络为主线，对语义、语法、语用和常识知识进行综合抽象与提练，为自然语言理解处理提供了关键知识。这些知识表达是真正数字化的，不是用自然语言来描述自然语言。（苗传江，2005）

下面看一个翻译实例：

例5 The People's Bank of China, the central bank, raised interest rates on March 18.

Google译文：对中国人民的银行，中央银行，提高3月18日利率。

HNC译文：中国人民银行，中央银行，在3月18日提高利率。

显然，上例中HNC的译文句子结构完全正确，译文正确，而Google的译文出现了句子结构错误，译文可读性差。当然，我们不能以一个例子的差异来评判两种方法孰优孰劣，但从中至少可以看到HNC自然语言处理的方向。

5 结束语

开发建设语义词典，好的理论指导很重要。然而,一方面，现在可借鉴很多前人的经验，不必白手起家；另一方面，还没有一个完全成熟的理论,能够指导建设出一个能够全面解决各类语义关系问题的知识库。因此,我们一方面需要不断优化已有的理论，另一方面要积极实践，探索更好的理论；另外，为提高语义词典的质量，对语义关系的语言学研究需要结合计算机科学和认知科学，开拓创新。

参考文献

[1] 董振东, 董强. 知网[Z/OL]. [2010-5-20]. http://keenage.com/zhiwang/c_zhiwang_r.html.

[2] 黄伯荣, 廖序东. 现代汉语(增订三版)上册[M]. 北京: 高等教育出版社, 2002.

[3] 黄曾阳. HNC(概念层次网络)理论[M]. 北京: 清华大学出版社, 1998.

[4] 梁捷. 国家863计划中文信息处理应用基础研究通过验收[N]. 光明日报, 2005-12-23.

[5] 苗传江. HNC(概念层次网络)理论导论[M]. 北京: 清华大学出版社, 2005.

[6] 缪建明. 声学所HNC语言处理团队参加"句法分析评测"创佳绩[Z/OL]. (2009-09-16) [2010-04-20]. http://www.ioa.cas.cn/xwzx/kydt/200909/t20090916_2490494.html

[7] 王惠, 詹卫东, 俞士汶. 现代汉语语义词典规格说明书[J]. 汉语语言与计算学报, 2003, 13(2): 159-176.

[8] 王兰成. OKIAS: 一个基于本体的知识集成及应用系统[J]. 信息系统, 2009, 32(12): 112-115.

[9] 王全智. 论语篇的使成条件[J]. 外语与外语教学, 2001, 149(9): 17-19.

[10] 王素格, 杨军玲, 张武. 基于最大熵模型与投票法的汉语动词与动词搭配识别[J]. 小型微型计算机系统, 2007, 28(7): 1306-1310.

[11] 杨喜昌. 俄语句子语义整合描写: 话语生成与理解机制的探索[M]. 黑龙江: 黑龙江人民出版社, 2005.

[12] 于江生, 刘扬, 俞士汶. 中文概念词典规格说明[J]. 汉语语言与计算学报, 2003,13(2): 177-194.

[13] 俞士汶. 自然语言处理与自然语言理解[Z/OL]. (2009-3-12) [2010-04-20]. http://www.sciencenet.cn/upload/blog/file/2009/3/2009312144245757l7.pdf

[14] 詹卫东. 面向自然语言处理的大规模语义知识库研究述要[A]. 中文信息处理若干重要问题[C]. 科学出版社, 2003: 107.

[15] 张清源. 现代汉语知识辞典[K]. 成都: 四川人民出版社, 1990: 191-192.