20 世纪汉语语义知识工程研究述要

北京大学　詹卫东

本文对 20 世纪（主要是 80 年代以来）汉语语义知识工程方面的研究作一个整体回顾。在对已经建成和正在构建之中的汉语语义知识库进行宏观概述的基础上，尝试以汉语语义知识在中文信息处理中所能发挥的效用以及如何发挥效用为基本视角，来审视和评价这些语义知识工程的理论背景和语义知识表述模式，进而对今后汉语语义知识库的建设开发作一些展望，希望能对相关研究工作有参考价值。

一　概 述

国内的汉语语义知识工程建设主要受到两方面的影响，一是国外语义学理论研究的影响，一是中文信息处理发展的影响。前者大致体现在汉语语义知识的表达方式上。后者则表现在已建成的汉语语义知识库的规模（量）与具体知识内容（质）上。需要说明的是，具体的汉语语义知识工程项目一般都在借鉴国外语义学理论的基础上，根据中文信息处理以及汉语语义知识表示的实际需要，进行适应性调整。此外，也有研究人员提出新的描写汉语语义知识的理论模式（如 HNC 理论）。

以下我们列举若干从 20 世纪 80 年代以来完成的或正在进行中的比较大的汉语语义知识工程项目。通过下面这个表，可以大致反映这一个时期汉语语义研究的整体面貌。

表 1：20 世纪 80 年代以来主要的语义知识工程项目简表

项目	时间	研究单位（人员）	规模及范围	理论基础与语义知识表述框架	应用
xhdc_CD	90-93	人大，清华等	1000 多动词，3000 多义项	格理论；格（必有+可选）+格位	人、机两用
905-sem	90-95	北语，河南财经学院等	4 万多实词，近 5 万义项	语义场，语义网络，格理论	信息处理通用
How-Net	88-98	董振东 等	5 万多实词，6 万多概念	语义分类+语义属性（属性值）+语义关系	信息处理通用
ST_CD	-95	中软公司	万词级，实词	语义分类+语义关系	机器翻译
TE_CD	96-97	北大，科学院计算所	近 5 万词，实词	语义分类+配价关系	机器翻译
HNC	88-98-	科学院声学所	建设中	HNC 理论	信息处理通用

说明：1. 上表项目栏是语义知识工程的代码，这些语义知识工程的具体名称如下：

xhdc_CD ：现代汉语述语动词机器词典

905-sem ： 国家“八五”中文信息处理应用平台工程（905 语义工程）
How-Net ： 知网（董振东先生创立的一个汉英双语知识词典）
ST_CD ： Sino-Trans 机器翻译系统中的汉语语义词典
TE_CD ： TransEasy 汉英机器翻译系统中内嵌的汉语配价词典
HNC ： 概念层次网络理论（Hierarchical Network of Concept）

2. 上表中给出的时间只是大致的时间段。其中 HNC 主要是指理论模式的创立时间。至于具体的语义知识库建设，目前仍在进行中。
3. 上表中六个语义知识工程研究项目可以分为四组（以空行隔开）。第一组是 80 年代以来国内研究人员开始借鉴国外语义学理论并根据汉语的描写需要加以扩充后，进行小规模初步试验的结果；第二组和第三组都是在借鉴国外理论基础上，同时也紧密联系中文信息处理的实践经验，进行大规模语义知识库建设的产物，区别在于第二组出于通用的考虑，希望建成的语义知识库可以成为基础平台为信息处理的各个应用提供支持；而第三组的语义知识库则主要是在一个实际应用系统的框架下开发完成的，跟汉外机器翻译的需求结合得更紧密。第四组是国内研究人员在借鉴国外语义学以及语言学理论基础上希望走出一条汉语自动分析的新路所做的努力。

二　评 价

要对已有的语义知识工程建设进行评价，我们的认识是，从原则上讲，评价标准只能是实用主义的标准，即考察一个语义知识库能对中文信息处理提供多大的支持，解决多少问题。但从具体操作上讲，真正以这个标准为指导做出科学客观的评判绝非易事，这同样也是一个浩大工程。这里我们只简单地说明我们对语义知识的作用的一些看法，以及相应的用于考察语义知识表述框架的一些评判尺度，基本不对具体的语义知识工程做细致的剖析。不过希望有关考察尺度的讨论能对今后的语义知识工程建设工作具有参考价值。

1. 语义知识在中文信息处理中的作用
   - ◆ 句法分析（parsing）
   - ◆ 多义词辨析（word sense disambiguation）
   - ◆ 相似度计算（similarity computation）
   - ◆ 推理（reasoning）

从在中文信息处理中所能发挥的作用以及发挥作用的方式上看，所谓的语义知识跟句法知识没有本质的差别，二者都是通过一定形式刻画语言符号与语言符号之间的关系。

2. 语义知识的表现形式
   - ◆ 静态的语义范畴知识 [属性 ： 值]
   - ◆ 动态的语义规则知识 [条件 -> 动作]
3. 语义知识的描写水平
   - ◆ 仅描写符号之间的聚合关系或组合关系（语义场/义素分析 | 格理论 | 配价）
   - ◆ 既描写符号之间的聚合关系也描写组合关系 （ 语义网络 ）
   - ◆ 在句子以下水平提取语义知识 （ 格理论 | 配价 ）
   - ◆ 在句子以上（超句）水平提取语义知识 （ 框架语义学 Frame Semantics ）
4. 语义知识库的开发方式
   - ◆ 手工/半自动 （ 表 1 中的语义知识工程，以及国外的 WordNet, FrameNet）
   - ◆ 自动 （ 微软公司的 MindNet ）

三　展 望

基于上述对语义知识的性质以及对语义知识库建设的认识，我们认为今后的语义知识工程研究工作相应地应该在 4 个方面有如下立场或取向：

1. 应“句法为主，语义为辅”，在句法知识库的基础上构建语义知识库。
2. 应加强动态的语义规则知识的研究和总结。实际上规则研究的结果也就意味着语义范畴的进一步细化精化。
3. 应根据应用需求，尽可能准确定位语义知识描写的水平，从而形成跟具体应用紧密配合的，合理的语义知识描述框架。
4. 随着自然语言处理技术的发展，应努力探索以自动方式构建语义知识库。

参考文献：

1. Lappin, Shalom, 1996, ed. The Handbook of Contemporary Semantic Theory, Oxford: Blackwell.
2. Saint-Dizier, Patrick and Evelyne Viegas, 1995, ed. Computational Lexical Semantics, Cambridge University Press.
3. Nerbonne, John, 1998, ed. Linguistic Database, CSLI Pubilcations.
4. Miller, George A. 1997, ed. WordNet, MIT Press.
5. Richardson, Stephen D. , 1998, MindNet: acquiring and structuring semantic information from text, In Coling’98. P1098-1102.
6. Baker, Collin F., 1998, The Berkeley FrameNet Project, In Coling’98. P86-90.
7.Barker,K. and Stan Szpakowicz. 1998. Semi-Automatic Recognition of Noun Modifier Relationship, In Proceedings of Coling’98. P96-102.
8.Fillmore,C.J.1982. Frame semantics, In Linguistics in the morning calm, The Linguistic Society of Korea ed. Hanshin Publishing Co. Seoul, P111-137.
9. Roack, Brain & Charniak, Eugene, 1998, Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction, In Coling’98. P1110-1116.
10. 董振东，董强，How-net 网站上关于 HowNet 的文献介绍。http://www.how-net.com
11. 董振东, 1998，《语义关系的表达和知识系统的建造》，载《语言文字应用》1998 年第 3 期。
12. 陈小荷，1998，《 一个面向工程的语义分析体系》，载《语言文字应用》1998 年第 2 期。
13. 鲁川，1998，《汉语的意合网络》，载《语言文字应用》1998 年第 2 期。
14. 梅家驹等，1983，《同义词词林》，上海辞书出版社 1983 年第 1 版。
15. 陈力为 袁琦主编，1995，《中文信息处理应用平台工程》，电子工业出版社 1995 年版。
16. 吴蔚天，1999，《汉语计算语义学》，电子工业出版社。
17. 林杏光，1998，《中文信息界的语义研究谭要》，载《语言文字应用》1998 年第 3 期。
18. 黄增阳,1998,《HNC（概念层次网络）理论》,清华大学出版社。http://159.226.60.26/hzy.html
19. 詹卫东，1997，《词的语义分类在汉英机器翻译中所起的作用以及难以处理的问题》，载陈力为、袁琦 主编《语言工程》，清华大学出版社 1997 年版（全国第四届计算语言学联合学术会议论文集，JSCL’97）。
20. 詹卫东, 1999,《一个汉语语义知识表达框架：广义配价模式》，载黄昌宁，董振东主编《计算语言学文集》（全国第五届计算语言学联合学术会议论文集，JSCL’99）。