

面向自然语言处理的大规模语义 知识库研究述要*

詹卫东

北京大学中文系 100871

E-mail: zwd@pku.edu.cn

摘要 本文对国内外一些有代表性的语义知识库进行了整体考察和比较,形成了四点认识:(1)各语义知识库均以“语义关系”为重点描写内容;(2)语义知识范畴具有明显的相对性特点;(3)语义知识主要是作为约束条件,在计算机对“语言形式”做各种变换操作时发挥作用;(4)应该重视通过系统的语言形式变换手段来界定语义范畴,提取语义约束条件。由此得到的语义知识,能更好更直接地为自然语言处理服务。

1 引言

本文打算对国内外自然语言处理领域中(主要是20世纪80年代以来)一些语义知识工程研究作一个整体回顾。就选取考察对象来说,本文主要考虑了(1)研究工作的影响;(2)研发单位的性质与地域分布;(3)知识库规模、语种;(4)时间性;(5)理论背景与构建方法等方面的因素。尽管限于篇幅和作者的视野局限,本文无法做到非常全面,但下文谈到的语义知识工程研究项目,应该说都具有一定的代表性,基本可以反映近二十年来国内外语义知识库研究的面貌。就本文的分析旨趣和目标来说,我们对各个语义知识工程的考察是希望能够从实践回到理论。因而更重综合,求共性,而不注重区别辨异。我们的想法是,语义知识库也像产品一样,它的制造者往往倾向于渲染它的特色,它的与众不同。而对语义知识库的研究做客观的综合考察,则应该追求从“各不相同”的具体的研究工作抽象出共同的需要解决的问题,这样,对未来的相关研究工作会有参考价值。

2 回顾

为简明和讨论方便起见,先把本文考察的12个语义知识工程项目(国内国外各6

* 本文题为“面向自然语言处理的大规模语义知识库研究述要”,但谈到的语义知识库工程中有的并不完全是“面向自然语言处理”,或者至少研究者的初衷并不是“面向自然语言处理”,但这些研究项目的成果实际上已经在或者可能将在自然语言处理的研究和应用中发挥重要的作用。因而客观上是“面向自然语言处理的”,或者至少是“部分面向自然语言处理的”。

个)的基本情况列一个简表如下。

表 1: 20 世纪 80 年代以来若干有代表性的语义知识工程项目简表

项目名称	时间	研制者	规模、语言	语义理论基础	构建方式
WordNet	1985-	美国普林斯顿大学	111223 个概念; 名、动、形容词、副词; 英语	基于关系的语义描述理论; 同义词集合, 语义关系描述	手工构建; 免费在线资源;
FrameNet	1997-	美国加州大学	458 个框架, 4000 多词; 英语	框架语义学; 框架元素, 配价, 语义关系	手工构建; 免费在线资源;
Integrated Linguistic Database	1993-1996	英国剑桥大学、爱丁堡大学等	规模不详; 英语	语义分类, 语义特征, 语义角色与选择限制等	手工构建; 不详
MindNet	1993-	美国微软公司	15.9 万词 (名、动、形); 英语	语义关系描述	自动构建; 商业产品
CYC 常识知识库 *	1984-	美国 CYC 公司 ¹	规模不详; 英语	人工智能知识表示理论 (Cycl 形式描述语言)	手工构建; 商业产品
EDR 概念词典 *	1986-1994	日本电子辞书研究所 ²	26 万日语词, 19 万英语词, 41 万个概念;	语义分类, 语义关系描述	手工构建; 商业产品
现代汉语述语动词机器词典	1990-1993	人民大学, 清华大学	1000 多动词, 3000 多义项; 汉语	格理论; 格, 格位	手工构建; 不详
“905”语义工程 *	1990-1995	北京语言大学, 河南财经学院	4 万多实词, 近 5 万义项; 汉语	语义场, 语义网络, 格理论	手工构建; 不详
How-Net(知网)	1988-	董振东 等	汉 英 双 语 116533 条记录;	义原分析 (2199 个义原,); 语义角色、语义关系描述	手工构建; 免费在线资源 / 授权使用
Sino-Trans-SemDict *	-1995	中软公司	规模不详, 实词; 汉语	语义分类; 语义关系描述	手工构建; 商业产品
Beida-SemDict *	1996-	北京大学	65330 词条; 名词、动词、形容词、副词; 汉语	语义分类; 配价; 语义角色选择限制	手工构建; 授权使用
CCD	2000-	北京大学	近 7 万个概念, 汉、英双语	类 WordNet 的语义知识表述框架	手工构建; 授权使用

¹ Douglas Lenat 于 1984 年在美国 MCC (微电子计算技术公司) 开始 CYC (CYClopædia 的缩写) 的研究工作, 1995 年成立 CYC 公司。CYC 的知识描述语言 Cycl 是一种 Lisp 风格的形式语言。

² 读者访问 EDR 网页可以看到, 日本电子辞书研究所 (EDR) 已于 2002 年 3 月 31 日解散, 目前 EDR 属于日本通信研究实验室 (Communication Research Laboratory, CRL)。

表 1 中有的项目名称是研发者所起的正式名称，有的并不是正式名称，而是下文为了称引方便临时冠以的名称（以*号标出）。这些语义工程项目的具体内容都是非常丰富的，表中的概括为了追求简明性，其中关于各知识库语义理论基础的说明，只是“点到为止”。另外，由于不少语义知识库仍在发展中，因而表中对知识库规模的描述也是阶段性的数据（对那些仍在发展的项目，读者可访问相关网站了解最新数据信息）。

就国内的研究工作来说，“现代汉语述语动词机器词典” [16] 是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调整后，进行小规模初步试验的结果；905 语义知识工程 [13, 18] 和 HowNet [11, 12] 则是以建设通用语义知识平台为目标，进行大规模语义知识库实践的产物；Sino-Trans SemDict [19] 和 Beida SemDict 语义词典 [20, 24] 都是在汉外机器翻译背景下开发完成的，跟汉外机器翻译的实际需求结合得很紧密。北大 2000 年开始的 CCD（中文概念辞书）项目 [23] 采用了类 WordNet 的知识描述框架，显示了国内语义知识工程与国际接轨的发展趋势。有关这些语义知识工程的细节内容，读者可以参阅相关文献做深入了解。下面通过举例的方式，对国外的语义知识工程做进一步的细节介绍，希望能有助于读者对这些研究工作有更直观的认识。限于篇幅和我们所掌握的资料，重点介绍 WordNet, FrameNet, ILD 和 MindNet 的一些情况。

1) WordNet [4]

WordNet 的基本单元是所谓的同义词集合 (synset)，下图中每个“{ }”就是一个 synset，集合中的元素相互之间构成同义关系。在 WordNet 的浏览器 (browser) 中查询“father”这个词，可以找到跟这个词所在的 synset 有“关系”（包括上下位关系，反义关系，整体一部分关系，等等）的其他 synset，这些 synset 之间形成一个“网”，是 WordNet 这张大网中的一个局部“小网”。

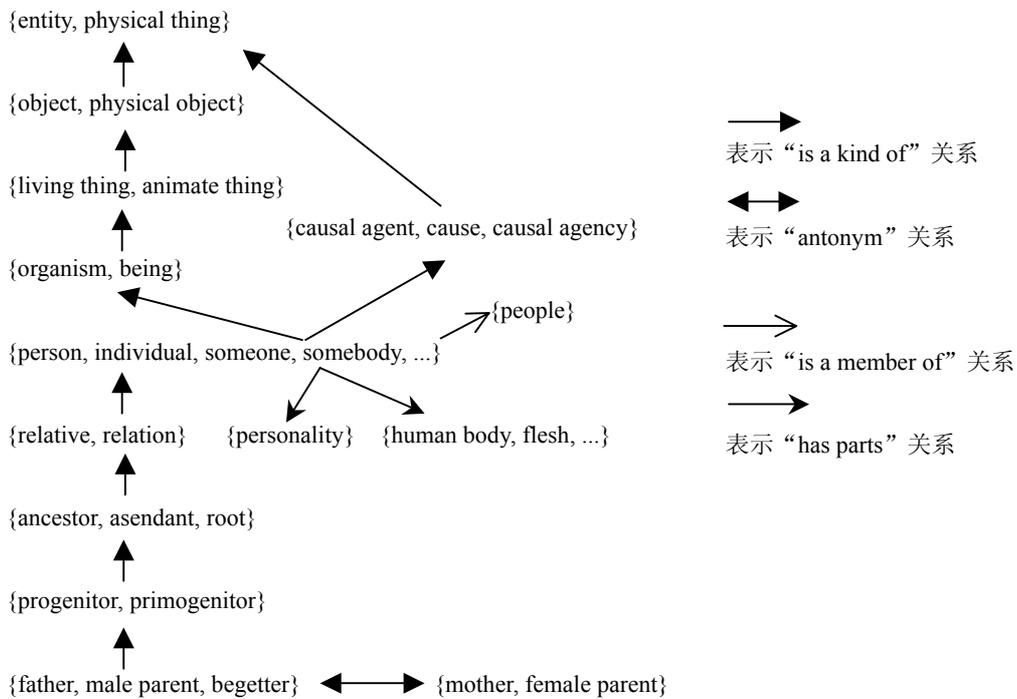


图 1 WordNet 词义关系示例

图 1 展示的是 WordNet 中名词的语义知识表示。WordNet 中也规定了动词、形容词、副词的语义知识表示规范，其核心都是 synset 以及概念之间的各种关系。跟下面将要介绍的其他三个语义知识库相比，WordNet 的语义知识表示有两点值得注意：（1）把“同义”关系放在了非常重要的位置；（2）不同词类之间的组配约束关系（比如动词跟名词之间组合的复杂关系）在 WordNet 中几乎没有涉及。

2) FrameNet [7, 9]

跟 WordNet 最初源自对词汇知识表示的心理学兴趣不同，FrameNet 完全是语言学家在一套系统的语义学理论指导下的一次工程实践。支持 FrameNet 的是著名语言学家 C.J.Fillmore 继格语法之后提出的“框架语义学”（Frame Semantics）理论。该理论的核心思想是，人们对词义的理解需要建立在对认知域，也就是“框架”（Frame）的理解的基础上。因此，“框架”是组织词汇语义知识的基本手段，一个框架中包含了若干“框架元素”（Frame Element），框架元素跟“格语法”中的“语义格”相比，更具体，分得更细一些，最重要的是，在以往的理论中，“语义格”是相对于所有词汇而言的，是高度抽象和概括的，而“框架元素”是相对于一个个的“框架”而言的，是“框架”中的构成成分。下面表 2 展示了“Removing”这个框架的情况。

表 2: FrameNet 框架示例: Removing

框架名	Removing (移开)	
框架描述	An Agent causes a Theme to move away from a location, the Source .	
框架元素	Agent 施事	The Agent is the person (or other force) that causes the Theme to move.
	Cause 致事	The noise of impact resulting from caused-motion of a Theme
	Theme 当事	Theme is the object that changes location.
	Cotheme 同事	The Cotheme is the second moving object, expressed as a direct object.
	Distance 距离	The Distance is any expression which characterizes the extent of motion.
	Goal 目标	The Goal is the location where the Theme ends up.
	Path 路径	Path along which moving occurs.
	Result 结果	Result of an event
	Source 起点	The initial location of the Theme, before it changes location.
	Vehicle 交通工具	The means of conveyance controlled by the Driver.
词例	abduct.v, clear.v, confiscate.v, depose.v, discard.v, dislodge.v, drain.v, eject.v, ejection.n, eliminate.v, elimination.n, empty.v, evacuate.v, evacuation.n, evict.v, eviction.n, ...	

每个框架都包含了一批词语，理解这些词语的词义，必须以理解整个框架为前提。比如“Removing”这个框架中就包含了“abduct、clear、confiscate、……”等动词，也包含了“ejection、elimination、……”等名词。这些词语的“共性”（尽管句法上分属不同词类），在同一个“语义框架”中得到了体现。“Removing”是一个描述动作性场景的框架，FrameNet 中也有描述事物性对象的框架，比如“Vehicle”（交通工具）就是一个事物类框架。为了表述的简洁，框架之间可以有继承关系，比如：frame(Driving)可以从frame(Transportation)继承框架元素。此外，对于框架中的动词，FrameNet 数据库还描写了各个框架元素（角色）的句法配位，即不同的框架元素（由名词或介词短语充当）在表层句子结构中所占据的句法位置。

3) ILD[ILD1, ILD2, ILD3]

Integrated Linguistic Database (综合语言知识库, 简称 ILD) 对词汇语义的描述主要包括三个方面: (1) 词语之间的上下位关系——这是通过语义分类树来表述的; (2) 词语的特征描述——ILD 中设置了多达 200 多个特征, 其中既有句法特征, 也有语义特征, 这些特征实际上表达了词语之间的多种关系 (比如表 3 中的 “used_for” 特征, 实际上就是在名词和动词之间建立了 “实体——用途” 的关系); (3) 动词对名词的语义选择限制。下面表 3 显示了 ILD 中的前两项内容; 表 4 显示了第三项内容。

表 3: ILD 中词义分类及语义特征描述示例

词语	上位词	下位词	特征名	特征值	参数
fortress	building*STRUCTURE	castle citadel	appreciation	strengthened	for defence purpose
			colour	.	
			constituents	walls roof ...	
			count	yes	
			group_of	buildings	
			made_by	humans	
			made_from	.	
			movability	no	
			position	on	ground
			shape	.	
			size	large	
			used_by	people	under attack
used_for	sheltering	within			

表 4: ILD 中动词对名词的语义选择限制描述示例

动词	角色	对角色的选择限制		扮演某个角色的典型成分
		抽象类别	具体描述	
assassinate	SUBJ	Human	assassin	terrorist, fanatic
	OBJ	Human	important /influential_person	president, prime minister
Sail	SUBJ	Human / Vehicle	person	sailor
	(across)	place	.	.

4) MindNet [5, 6]

MindNet 跟其他语义知识工程最大的不同在于它的构建方式。MindNet 是利用微软功能强大的句法分析器 (Parser) 自动分析词典释义 (Definition) 文本得到的。MindNet 中预设了 24 种关系, 如下面表 5 所示³:

³ 表 5 及下面图 2 的示例均引自 Richardson, Stephen D. et al. 1998, 中文译词是本文作者加的。

表 5: MindNet 中的 24 种关系

Attribute属性	Goal目标	Possessor领有者
Cause原因	Hypernym上位	Purpose意图
Co-Agent联合施事	Location场所	Size大小
Color颜色	Manner方式	Source源点
Deep_Object深层宾语	Material材料	Subclass子类
Deep_Subject深层主语	Means方法	Synonym同义
Domain领域	Modifier修饰语	Time时间
Equivalent同位	Part部分	User使用者

句法分析器对词典中的释义文本进行分析，即可得到词语之间的各种语义关系，由于 MindNet 选取的英语词典（朗文当代英语词典 LDOCE 和美国传统词典 AHD）的释义模式比较规范，这使得自动抽取的词语语义关系效果比较好。下面是一个实例。

词典对“car”的释义：

“a vehicle with 3 or usu. 4 wheels and driven by a motor, esp. one for carrying people”

自动分析后得到 car 的语义关系（Tobj 表示深层宾语，Hyp 表示上位，...）：

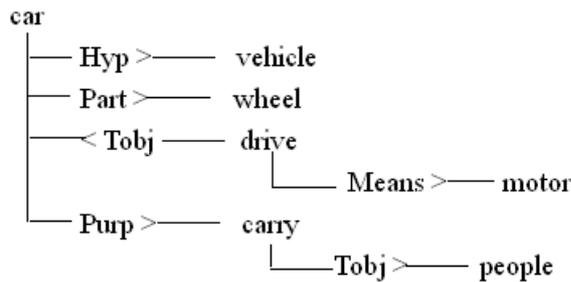


图 2: MindNet 语义关系示例

以上四个知识工程均专注于语义知识的描述。另外两个项目 EDR 和 CYC 都是涵盖面非常广的大规模语言知识库，其中包含了语义知识，但并不限于语义内容。EDR（Electronic Dictionary Research）是日本政府于 1986 年开始的大型电子词典工程项目，历时 9 年完成，由日本政府核心技术和 8 个大企业资助。一共包括 11 个分词典，其中跟语义知识直接相关的概念词典又包含三个分词典（HeadConcept Dictionary, Concept Classification Dictionary, Concept Description Dictionary），分别描述了基本概念，概念的上下位关系分类，概念间的语义关系等。CYC 常识知识库从 1984 年开始构建，目前仍在发展，其中包含了概念、概念间的关系，以及描述概念间推理关系的规则三部分内容（CYC 在网上公开的知识库有 6000 个概念和 60000 个关系描述）。

前面以举例方式展示了国外几个语义知识库内容的一些细节。国内大规模汉语语义知识工程建设起步比国外晚一些。影响国内相关研究工作的主要有两个因素，一是国外语义学理论的研究，一是中文信息处理自身的发展。前者的影响大致体现在汉语语义知识的表示方式上。后者则表现在已建成的汉语语义知识库的规模与具体知识内容上。

撇开各知识库之间量的差异不说（量的积累要受到时间和财力支持等诸多客观条件

的限制),就知识库的具体知识内容而言,不难看出,目前的语义知识库无论表现形式上有多大的差异,但描述的重点或者说是共同的目标都是试图刻画词语之间的各种**语义关系**⁴。其实反观语义学理论研究的发展之路,从早期的义素分析(Semantic Component Analysis)和语义场理论(Semantic Field)发展到现在的关系语义学(Relational Semantics)和框架语义学[1, 2, 3],人们实际上一直就是在朝着以系统的方式揭示语义(词义)关系的**目标在努力**⁵(只不过早期的义素分析和语义场理论描述的关系一般限于“上下位关系、同义关系、反义关系”等等少数关系而已,现在的语义学理论大大拓宽了人们所关注的语义关系的范围)。

不过,值得注意的是,跟以前人们争论到底该确定多少义素,多少个语义格角色类似,现在,应该设置多少“语义关系”也成了一个问题。换言之,从历史的情况来看,似乎人们关注的基本语义范畴(义素特征、语义场、语义格角色、语义关系,等等)从来都是难有定数的,无法形成一个统一的体系。这就很自然地引发我们思考:应该如何看待语义知识(范畴)的性质呢?下一小节我们围绕这个问题略加展开讨论。尽管限于篇幅和认识水平,讨论可能不够深入,但希望能够起到抛砖引玉的作用。

3 反思

3.1 语义范畴的相对性

无论是从语义理论体系研究的历史发展来看,还是从语义知识工程的实践情况来看,语义范畴跟句法范畴比起来,一直都不太容易形成比较统一的意见⁶。各家设置的语义范畴,无论是名称,还是数量,往往有较大的差异,本文把这样的情况概括为**语义范畴的相对性**。我们曾经对“施事”“当事”等语义角色设置的相对性问题有过一些分析[22]。这里再就语义分类和词语义项分析的相对性谈一些看法。

下表显示了几个语义分类体系之间在“动”“名”语义分类树(tree)上概念节点(node)的数量差异,直接反映了实际的语义工程中设置语义范畴的相对性[11, 13, 14, 15]。

表 6: 语义树上节点数量的比较

分类体系		905 语义工程 / 鲁川的 动词分类体系 ⁷	Beida-SemDict	HowNet (知网)
比较项目	名物类	152	39	137
	词语分	110	30	103
	类树	9	6	8
	节点数	156	14	67

⁴ 董振东先生在第二届词汇语义学研讨会上的发言标题就是“关系——词汇语义的灵魂”[12]。

⁵ 语义的义素特征表示也可以看作是揭示词语之间语义关系的一种手段。

⁶ 早期的语义角色主要就是动词的“格”(比如“施事、受事、工具”等),这些角色实际上也就是动词和名词之间的语义关系(比如“施事”等价于动、名之间形成“动作——动作发出者”这样的语义关系)。一直以来,到底有多少“格”,一个语义体系应该设置多少“格”,都没有形成统一的意见。现在语义描写朝进一步细化的方向发展,原来的“格”变成了更多的语义角色,变成了框架中的“框架元素”,然而同样的问题仍然存在,比如一个框架中应该有多少“框架元素”,似乎也难有一个统一的答案。

⁷ 表 6 中这一列上三行是 905 语义工程关于名物类词语的语义分类,下三行是鲁川的动词分类体系。

	节点数	156	14	67
	隶属系数	142	12	55

节点数量的多寡一方面反映了分类的粗细不同，另一方面也说明语义类确实也有其相对性的一面。在我们看来，造成这种相对性的原因主要有两个：

(1) “层级分类结构” (hierarchy) 仅仅是我们认识事物的一种方式而已，而非唯一方式。有的事物适合（或者说我们习惯于）放在“层级分类”的框架（即“树”结构）中来认识，有的事物并不适合这样来认识。举例来说，对于有些概念（词），人们很习惯用层级结构来认识它们。比如“柳树”这个概念，我们习惯于这样来认识它，“柳树是一种树，树是植物，植物是生物，……”（“柳树”的上位关系概念）；“垂柳、旱柳、杨柳、……等等都是柳树”（“柳树”的下位关系概念），“杨树、梧桐、……等等跟‘柳树’一样，也都是某种树的名称”（“柳树”的同位关系概念）。但对于另一些概念（词），比如“灰尘、椭圆、人民币、利率、三角债、窟窿、形势、土、眼泪、借口、雨滴、余地、……”等等，人们可能并不是用“层级结构”的方式去认识它们。对这些概念，硬要将它们定位到一个语义分类体系中，常常会感到捉襟见肘（有过语义词典编纂实践经验的人对此应该深有体会）。人们是用什么样的结构去认识这些概念，就目前来讲，应该还是个研究的课题，FrameNet 语义工程无疑在这方面做出了有益的探索。

(2) 即便某类事物适合以“层级分类的眼光”去认识，也可以有多个“层级分类结构”，而非只有一个“层级分类结构”。人们认识概念的角度可以有很多个，从而形成多个“层级结构”。比如对“人”这个类中的成员，“诗人、同事、师生、校长、瘪三、红娘、少先队、……”等等，就可以从“性别”、“单复数”、“职业”、“身份”、……等等不同角度去构造关于“人”的“层级分类结构”。在一个语义知识工程中，作为 ontology 给出的语义分类树，只不过是众多的“层级分类结构”中的一个罢了。

以上扼要分析了语义分类的相对性，下面再谈谈“词义”的相对性。对任何一个语义知识库而言，确定一个词的“词义”无疑都是一项非常基本的任务。尽管各家对如何表示“词义”各有不同的办法（比如 WordNet 用 synset 表示词义，HowNet 用义原来描述词义，等等），但无论表示方法有什么不同，对具体词语意义的认识，特别是一个词语义项的分合问题，却是大家必须面对的共同问题。这里不妨来看三组例子，通过这些例子，可以体会到词义分析中实际存在的相对性问题：

- (1) a 教育孩子——b 教育要面向现代化——c 教育方法是非常重要的——d 办教育
- (2) a 弹钢琴 —— b 弹了一首肖邦的名曲 —— c 弹棉花 —— d 弹灰尘
- (3) a 靠海吃海，靠山吃山 b 吃环境饭 c 吃酱豆是吃文化
d 先是车吃炮，然后将军…… e 这台 ATM 机吃卡 f 小孩子特别爱吃手
g 沿线村民为何“吃铁路”成风 h 杨振宁说：“我不吃宴席，我要吃云南的……
i ……刮起了“吃资源”的狂风 j 要求解决全村村民吃自来水问题
k 吃财政的人就有五千人 l 吃日本料理，一半是吃环境，吃氛围，吃情调。
m 长钢这些年所有政策优惠几乎都享受到了，仅 1996 年银行就贷款四亿多元。“吃了财政吃银行，吃了银行吃股民，吃了股民吃老外”，结果是越吃越穷，越吃越亏，最后导致董事长、总经理、党委书记三人一起被免职。
n 对“吃请”和饭局要学会拒绝

第一组例子中“教育”的意思似乎比较明确，但是“教育”在几种不同语境中出现时所承担的功能不同，在例 1a 中“教育”作“述语”，在例 1b 中“教育”作“主语”，在例 1c 中“教育”作定语，在例 1d 中“教育”作宾语，这样就牵涉到“教育”的词性判别问题，连带着也引发出对这里的“教育”到底是一个“意思”，还是有不同“意思”的问题。

第二组例子中“弹”的功能倒是比较单一，都是作“述语”，因而无疑是动词，但由于后面宾语的差别，导致人们对“弹”的意思理解也多少会有不同。这种情形下，是处理为一个“弹”（能够系联多个语义角色），还是处理为几个不同的“弹”（每个“弹”系联一个语义角色）？恐怕也难有统一的答案。

第三组例子中“吃”的情形跟“弹”相比更复杂一些。例 2 “弹”的宾语的语义角色还基本属于一些典型的格范畴。跟“弹”共现的名词性成分“钢琴”“名曲”“棉花”“灰尘”所充任的语义角色大致可以归入“工具”“受事”“结果”等等，而例 3 中与“吃”共现的名词性成分（既包括“吃”前的主语，也包括“吃”后的宾语）则包含了多种类型。像“吃山”中的“山”，“吃文化”中的“文化”，都很难说清该归入什么语义角色。相应地，可以体会到“吃”在不同语境中出现时都表达了跟典型的“吃”相关的某个侧面的意义特征，比如：[吞咽]，[令某物消失]，[消耗]，[包含]，[享受]，……等等。这诸多影响“吃”的词义理解的因素，是综合在一起，绑定在一个“吃”上呢？还是分而别之，解析为若干个“吃”呢？这又是一个不易处理的问题。

从上面的例子来看，语义范畴的相对性体现在很多方面，而且这些方面又都是交织在一起的，比如动词词义分析实际上就跟动名语义关系分析紧密联系在一起。语义范畴的相对性也可以说是语义概念的“不精确性”。语义范畴很难做到有一个精确的定义，在语言的“语义”层面，许多概念是处于一种“边界模糊”的状态中，至少在目前的研究水平上这是不容否认的实际情况。需要说明的是，认识到语义知识的这种相对性，并不意味着对语义研究的悲观情绪，相反，如果这种看法是对的，实际上有助于我们树立对一个语义知识体系的“实用主义”评价观，即一个“语义知识体系”的好坏，根本上应该取决于它在某个应用领域中是否够用、好用。从这点上说，认识语义范畴，最好的办法是去深入了解语义知识在自然语言处理中能够发挥什么作用以及如何发挥作用。

3.2 语义知识在自然语言处理中的作用

由前面所述可知，目前语义知识库中记录的语义知识主要就是语义关系知识，在讨论语义知识的作用之前，有必要先简略谈谈语义关系的类型。

传统的结构主义语言学把语言成分之间的关系分为**聚合关系**和**组合关系**两类。一般来说，聚合关系反映了**同质**语言成分之间的**类聚性质**，比如典型的“同义关系”，此外“上下位关系”“反义关系”等等也都属于聚合关系类型的语义知识；组合关系则是体现了**异质**语言成分之间的**组配性质**，比如典型的“动词—名词搭配关系”，包括“施事—动作关系”“受事—动作关系”等等，都属于组合关系类型的语义知识。如果以此为背景来考察一个语义知识库在描述语义知识内容时的侧重，那么 WordNet, 905 语义工程等相比于其他知识库，就更侧重聚合关系的描写；而 FrameNet, ILD, Beida-SemDict 等，就更侧重组合关系的描写。

两种类型的语义关系知识在自然语言处理的不同应用中都可以发挥作用。比如聚合关系知识在信息检索中常用来作为扩展用户查询条件的依据，举例来说，用户输入“王

小明大夫”作为检索条件，计算机应该把包含“王小明医生”的文章也作为检索结果返回给用户，这就要建立在计算机知道“大夫”跟“医生”形成同义聚类关系的基础上。组合关系知识则常常在像机器翻译这样的需要句法分析的应用场合中发挥作用[19, 20]，举例来说，“练习时间”和“练习篮球”都是“动词 + 名词”的组合，但前者“练习”跟“时间”是修饰限定的关系（译为“training time”），后者“练习”跟“篮球”是动作行为与支配对象的关系（译为“learn to play basketball”），计算机要像人一样做出类似的分析，就必须知道“练习”跟“篮球”这样的具体事物类名词组合形成“动一名”支配关系，而跟“时间”这样的抽象名词组合则形成“动一名”修饰关系。从后一个例子不难看出：组合关系知识要有效地发挥作用，一般也要建立在聚合关系知识的基础上。拿上面这个例子来说，一个语义知识库在记录动词“练习”跟名词的组合关系时，往往是描述“练习”所能组配的名词的“类”（比如“具体事物”类名词），而不是一个个具体的名词（比如“篮球”）。实际上，刻划组合关系的语义知识除了确立不同的语义组合类型外，很重要的任务就是描述组合成分之间的选择限制，而这种选择限制，常常是以聚合类为单位来加以记录的（比如跟动词“吃”形成“受事”组合关系的名词应该属于“食品”类）。

在更抽象的层面上看，上面的例子说明，语义知识的作用是帮助计算机在模拟人的语言能力时做正确的“形式变换”。这又可以分成多个层次，包括判别组合后的形式是否合法，区别不同的组合关系，进行有效的变换推理，等等，比如上面举的“练习时间”与“练习篮球”的例子，从句法分析的角度说是要正确判定组合关系，如果从跨语言的形式变换（比如机器翻译）的角度看，则是要让计算机知道二者有不同的翻译形式；如果在汉语内部变换，那么“练习篮球”可以同义变换为“练篮球”，而“练习时间”则不能做类似的变换，因为“练时间”是不合法的组合形式。对此下面再举几个简单的例子做进一步的说明。

- (1) a 警察叔叔 —— b* 叔叔警察
- (2) a 干部群众 —— b 干部模样 干部身份 干部家庭
- (3) a 很多读者给编辑部写了信 —— b 很多人给编辑部写了信

上面例 1 中“警察”在“叔叔”前可以形成“合法”的组合（被说汉语的人接受），在“叔叔”后则形成“非法”的组合，像这样的例子有类推性，比如“医生叔叔”“护士阿姨”等等都合法，而“*叔叔医生”“*阿姨护士”等都不合法，由此可以概括出一个约束条件，即“职业”类名词应该在“亲属”类名词前出现构成合法的组合，如果以相反顺序出现则形成非法组合。例 2 中都是合法的组合形式，但其中组合关系又有不同，例 2a “干部”与“群众”同属“身份”类指人名词，构成并列关系；例 2b 与“干部”组合的名词都不是指人名词，跟“干部”属不同的语义小类，不能跟“干部”构成并列关系。例 3a 可以变换为 b，因为“读者”是“人”的下位概念，进行替换后形成的例 3b 语句合法，并且没有改变 a 句的基本意思（更严格地说是例 3a 所表达的命题在逻辑上蕴涵了例 3b 所表达的命题）。

通过上面这些简单的例子，不难看出，聚合关系的语义知识也好，组合关系的语义知识也好，实际上都可以帮助计算机来决定如何进行正确的“形式变换”，在做变换时，为了保证变换的正确性，除了要以语义知识库中记录的静态的语义关系知识（即“属性：值”型知识）为基础，还要用到动态变换规则知识（即“条件 → 动作”型知识）。比

如上面例 3a 到 b 的变换，可能就受这样一条变换规则制约：位于主语位置的名词可以被它的上位词替换，并且替换前的句子所表达的命题在语义上蕴涵替换后的句子所表达的命题⁸。为了说明确实存在这样的规则，下面再看一个对比例子：

(4) a 这本书有很多**读者** —— b* 这本书有很多人

例 4 跟例 3 中都是试图将“读者”替换为“人”，但例 3 是正确的变换，例 4a 变成例 4b 却形成非法的表达形式。这里可能有不少因素在起制约作用，不过无论是什么因素，都意味着变换是受到一定规则制约的，如果不满足规则的条件要求，就会造成错误变换。

从上述分析可知，语义知识要在自然语言处理应用中发挥作用，除了静态的语义关系知识外，很重要的是基于关系的约束条件知识，也即基于语义约束的形式变换规则知识，目前的语义知识库记录的主要都是静态的语义关系知识，但对基于语义关系约束的形式变换规则知识则研究得相对比较少。而为了让语义在自然语言处理中更好地发挥作用，这部分研究工作是必不可少的。本文认为这应该是今后语义研究的一个重要方向。

4 展望

在上一节中，我们对目前语义知识研究的两个基本方面进行了反思，一是指出语义范畴的相对性，二是指出语义范畴的作用从根本上讲是帮助计算机做“正确的形式变换”。这两个方面其实是有非常密切的关系的。一方面，语义范畴的相对性，说到底是对人们对语义概念无法准确地界定造成的，因为目前人们对语义的认识基本上可以说是“从意义到意义”，对语义概念的把握大多是以意念为主，而不是诉诸于形式标准，即便有一些形式标准（比如说 WordNet 对 synset 的操作定义就是以形式上的“可替换性”为基础），但也还没有**系统地**将语义概念的界定建立在形式特征基础上。另一方面，语义范畴的作用又落在“形式变换”上。也就是说，尽管语义范畴的界定相对模糊，但其目标是为了比较严格和精确的“形式变换”提供支持和服务。既然如此，为什么我们不重新认识语义范畴，把语义范畴的基础直接建立在“形式特征”基础上呢？比如，为什么我们要用“动作的发出者”这样含糊的字眼来定义“施事”，而不用类似下面这样的形式特征来定义“施事”呢？

如果“n1 + v + n2”可以同义变换为“n1 + 把 + n2 + v”，那么，n1 跟 v 形成“施事-动作”关系。

当然这个具体的形式定义是非常粗糙并且无疑是有漏洞的，但这里想强调说明的是，这样一种方式应该成为未来定义语义范畴时所追求的一种基本的普遍的模式。本文把这种追求称之为从“意义”向“形式”回归。

“意义”本身是抽象的，要把握抽象的意义，最便利也是最有效的办法就是通过比较具体的形式特征来界定和区别不同的语义范畴。应该说，跟陷在“意义”的“泥潭”里纠缠不清相比，从“形式”上去寻求答案，是更务实一些的做法。比如 WordNet 对 Synset 的界定，就是从形式上来认识“同义关系”的一种尝试。尽管有时候形式特征也同样存在判别的问题，但是会把问题引向更深层次的思考中。请看下面的小例子：

他**歌**唱得特别好 —— 他**唱歌**唱得特别好

他**琴**弹得特别好 —— 他**弹琴**弹得特别好

⁸ 此处仅仅是举例说明存在这样的规则，这里所举的具体规则仅起示例的作用，并不全面，也不严密。

如果不附加其他限制条件，仅就上面两例左右的形式替换来说，替换前后句义没有差别，但我们显然不能由此得出“歌”跟“唱歌”，“琴”跟“弹琴”分别形成“同义关系”的结论。为什么会出现上面这样的情况呢？就引发我们对“可替换性”这个形式特征本身进行更深入的探索，去为“替换”确立更多的附加限制条件，从而推动研究的进展。

除了通过形式变换（“替换”是变换的一种）来认识同义关系外，其他的语义现象也可以从形式变换的角度出发来观察。比如对比下面的例子：

a 张三打李四 —— b 张三打不过李四

这对例子促使我们考虑两个问题，一是“打”这个多义词的义项判别问题，一是跟“打”组合的名词性成分的语义角色的判别问题。例 a 中的“打”可以解释为“殴打”，但例 b 中的“打”似乎不能做这种解释；从语序上说，都是“张三”在“打”前，“李四”在“打”后，但例 a 中“张三”和“李四”分别充任“施事”和“受事”角色，而例 b 中却很难说“张三”跟“李四”仍然是分别充任“施事”和“受事”角色，在例 b 中，“张三”和“李四”更倾向于被解释为“共同当事”（co-experiencer）。类似的例子再比如：

c 人喝酒 —— d* 酒喝人 —— e 酒喝不死人

例 c 中“人”是“喝”的“施事”，“酒”是“喝”的“受事”，遵从汉语中“施事”占据主语位置（即动词前位置），“受事”占据宾语位置（即动词后位置）的语序要求，是合法的表达形式，例 d 因为违反了这个语序要求，因而造成不合法的表达形式，但例 e 的存在提示我们，语言现实并不是如此简单。在例 e 中，“酒”和“人”如果仍然解释为“受事”和“施事”，就违背了上面的语序要求。显然，要么修改上述的“施事—动词—受事”的语序规则，要么给例 e 中的“酒”和“人”赋予新的语义角色（比如“致事/causer”和“对象/theme”）。这就引发我们要做更多的思考。这些实例都说明，对于“语义”范畴（包括语义角色、语义关系等等），不能仅仅停留在抽象的、模糊的“意念”层面去认识，而应该回到语言的最基础层面——形式层面，去系统地进行组合、变换的分析，从形式特征差异上去揭示语义范畴的内涵，也只有按照这种路线提取和界定的语义范畴，才有可能为描述语言形式之间的组合、变换关系提供直接的规则约束条件。

以上都以实词为对象在讨论语义范畴的性质，下面再来看看虚词语义的描述问题。一方面，这是目前的语义知识库尚未涉足的领域，另一方面，虚词语义对自然语言理解与分析实际上又至关重要。因此，未来的语义知识库应该更加重视对虚词语义的研究和描写[25]，从“务实”向“务虚”扩展。在描述虚词语义时，通过形式变换来比较和凸现“虚词”的意义，会显得更加重要和有效。下面简单举两个例子来说明这个问题。

(1) a 张三和李四去了长城 —— a' 张三和李四都去了长城
b 张三和李四买了两本书 —— b' 张三和李四都买了两本书
c 小王和小张结婚了 —— c' 小王和小张都结婚了

(2) a 他挣了十块钱 —— a1' 他才挣了十块钱 —— a1" 他一天才挣了十块钱
a2' 他就挣了十块钱 —— a2" 他一天就挣了十块钱

例 1 通过对比用和不用虚词造成的句义差别来体现一个虚词（“都”）的表义功能；例 2 通过对比用不同的虚词（“就” — “才”）造成的句义差别反映虚词之间表义功能的细微差别。两个例子都反映了虚词语义的重要性。跟实词相比，虚词语义更空灵，更难以把握，因此，对虚词意义的分析，必须建立在形式对比的基础上，对计算机才有价值。比

如面向人的研究，可以说“都”的意思是表“范围”、表“总括”等，但面向计算机来刻划虚词语义时，则应该探讨从形式变换特征来定义“都”的“意思”：

$NP \text{ 都 } VP \rightarrow NP_1 VP \cup NP_2 VP \cup \dots \cup NP_i VP$; $NP = \{NP_1, NP_2, \dots, NP_i\}$
用自然语言表述就是，出现在“都”前面的成分（NP）为一个集合，集合中的每个元素（ NP_1, NP_2, \dots, NP_i ）均具有谓语部分（VP）所表达的动作行为和性状特征，即箭头左边的“语言形式”可以变换（推导）出右边的“语言形式”。以这种方式去刻划虚词语义，对于计算机进行自然语言理解和推理才会有直接的支持作用。

5 结语

随着自然语言信息处理研究的逐步深入，应用范围的逐步拓宽，语义知识库建设以及相关研究工作的迫切性和重要性无疑都将呈现显著上升的趋势。因此，及时盘点已有的成果，总结已有的经验，对未来的工作是很有参考意义的。本文在这方面做了一些初步的总结工作，概括起来为下面五点：

- (1) 对国内外一些有代表性的语义知识库进行了整体考察和比较；
- (2) “语义关系”是各语义知识库共同关注的重点描写内容；
- (3) 包括“语义关系”在内的语义知识范畴具有明显的相对性；
- (4) 无论是“语义组合关系”，还是“语义聚合关系”知识，实际上都是作为约束（判别）条件，在计算机对“语言形式”做各种类型的变换（组合）操作时发挥作用；
- (5) 在今后的语义研究和知识库建设中，应该重视通过系统的语言形式变换手段，从形式特征出发，来界定语义范畴，提取语义约束条件。也就是说，“意义”的研究应该立足于“形式”。并且我们相信，在语义研究关注的对象从实词向虚词拓展的过程中，这种研究路线的价值将得到更充分的体现。

语义研究以及大规模语义知识库工程建设涉及的内容非常广泛。除了上面探讨的一些比较基础的方面外。还有不少更实际一些的问题，比如目前许多语义知识库都重视从单语向双语甚至多语的发展（FrameNet 除描写英语的语义知识外，已经有法语、西班牙语、日语方面的相关项目在同时进行，WordNet 更是启动了 EuroWordNet 计划，包含了多种欧洲语言在内）。另外，语义知识的获取方式从主要依靠专家手工编纂向自动构建方式发展，至少是利用计算机工具软件进行人机交互的方式构建语义知识库[8, 10]。这些都是未来发展大规模语义知识库需要注意的特点。

致谢 本文研究得到国家重点基础研究发展规划项目“面向中文信息处理的现代汉语动词论旨结构系统和汉语词语语义分类层级系统研究”（编号：G1998030507-1）的资助。

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 Publications.
- [4] Fellbaum, Christiane, 1998, ed. WordNet: An Electronic Lexical Database, MIT Press.

- [5] Richardson, Stephen D. , 1998, MindNet: acquiring and structuring semantic information from text, In Coling'98. pp.1098-1102.
- [6] Dolan, W., L. Vanderwende, and S. Richardson. 1993. Automatically deriving structured knowledge bases from on-line dictionaries. In *Proceedings of the First Conference of the Pacific Association for Computational Linguistics*, pp.5-14.
- [7] Baker, Collin F., et al., 1998, The Berkeley FrameNet Project, In Coling'98. pp.86-90.
- [8] Barker, K. and Stan Szpakowicz. 1998. Semi-Automatic Recognition of Noun Modifier Relationship, In Proceedings of Coling'98. pp.96-102.
- [9] Fillmore, C.J. 1982. Frame semantics, In *Linguistics in the morning calm*, The Linguistic Society of Korea ed. Hanshin Publishing Co. Seoul, pp.111-137.
- [10] Roack, Brain & Charniak, Eugene, 1998, Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction, In Coling'98. P1110-1116.
- [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] 董振东, 2001, 《关系: 词汇语义的灵魂》, 第二届词汇语义学研讨会(北京大学, 2001.5)。
- [13] 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。
- [14] 鲁川、林杏光, 1989, 《现代汉语语法的格关系》, 载《汉语学习》1989年第5期。
- [15] 鲁川, 1998, 《汉语的意合网络》, 载《语言文字应用》1998年第2期。
- [16] 林杏光等, 1994, 《现代汉语述语动词机器词典》, 北京语言学院出版社。
- [17] 梅家驹等, 1983, 《同义词词林》, 上海辞书出版社 1983年第1版。
- [18] 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。
- [19] 吴蔚天, 1999, 《汉语计算语义学 —— 关系、关系语义场和形式分析》, 电子工业出版社。
- [20] 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)。
- [21] 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)。
- [22] 詹卫东, 2001, 《确立语义范畴的原则及语义范畴的相对性》, 《世界汉语教学》2001年第2期。
- [23] 于江生、俞士汶, 2002, 《中文概念词典的结构》, 载《中文信息学报》2002年第4期。
- [24] 王惠、詹卫东、俞士汶, 2003, 《现代汉语语义词典规格说明书》, 载《汉语语言与计算学报》(新加坡) Vol.13, No.2.
- [25] 俞士汶等, 2003, 《现代汉语广义虚词知识库的建设》, 载《汉语语言与计算学报》(新加坡) Vol.13, No.1

在线资料

- [WordNet] <http://www.cogsci.princeton.edu/~wn/>
- [FrameNet] <http://www.icsi.berkeley.edu/~framenet/>
- [ILD1] <http://www.hrc.ed.ac.uk/Site/ILD.html>
- [ILD2] <http://www.ltg.ed.ac.uk/projects/ild/>
- [ILD3] http://icl.pku.edu.cn/doubtfire/semantics/CUP_ILD/index.htm
- [MindNet1] http://research.microsoft.com/nlp/Tell_Me_About_Yourself.asp

[MindNet2] <http://research.microsoft.com/research/projects/>

[EDR] <http://www.jsa.co.jp/EDR/>

[Cyc1] <http://www.cyc.com/index.html>

[Cyc2] <http://www.opencyc.org/>

[Hownet] <http://www.keenage.com/>

[Beida-SemDict] <http://ccl.pku.edu.cn/>

作者简介

詹卫东 男 浙江人，1999年毕业于北京大学中文系，获博士学位。著有《面向中文信息处理的现代汉语短语结构规则研究》。曾在《中国语文》《中文信息学报》《当代语言学》《语言文字应用》《汉语学习》《语文研究》等发表论文多篇。主要研究领域为现代汉语形式语法、机器翻译、语义学。