“现代汉语语义词典”的结构及应用*

王 惠[1] 詹卫东[2] 俞士汶[2]

(1. 新加坡国立大学中文系 新加坡 117570; 2. 北京大学 北京 100871)

[摘要]“现代汉语语义词典(SKCC)”是一个面向汉英机器翻译的大规模汉语语义知识库,它以数据库文件形式收录 6.6 万余实词,不仅给出每个词语所属的词类、语义类,而且以义项为单位详细描述了它们的各种语义搭配限制。目的是为计算机语义自动分析、词义消歧等任务提供强有力的支持。本文介绍这部语义词典的结构、内容,并以实例说明这部词典可有效地解决翻译系统中的词汇歧义(WSD)问题。

[关键词]语义词典;词义消歧;词汇语义学;自然语言处理;中文
[中图分类号]H08[文献标识码]A[文章编号]1003-5397(2006)01-0134-08

Structure and Application of
The Semantic Knowledge-base of Modern Chinese

Wang Hui, Zhan Weidong, Yu Shiwen

Abstract: *The Semantic Knowledge-base of Modern Chinese* (SKMC) is a large scale bilingual semantic resource. It provides a large amount of semantic information such as semantic hierarchy and collocation features for 66539 Chinese words and their English counterparts. Its POS and semantic classification represent the latest progress in Chinese language engineering. The descriptions of semantic attributes are fairly thorough and comprehensive. The main work in this paper is to introduce the outline of SKMC, and establish a multi-level Word Sense Disambiguation (WSD) model based on it. The results indicate that the SKMC is effective for word sense disambiguation in Chinese and are likely to be important for general Chinese Natural Language Processing (NLP).

Key words: Semantic Knowledge-base; WSD; Lexical semantics; NLP; Chinese

[收稿日期] 2003-10-20

[作者简介] 王惠,新加坡国立大学助教,博士,主要研究汉语词汇学、语义学和计算语言学;詹卫东,北京大学副教授,博士,主要研究汉语语法、语义和计算语言学;俞士汶,北京大学教授,主要研究计算语言学。

*本研究得到国家 973 重点基础研究项目(G1998030507-4)和(G1998030507-1)资助,研究还得到北京大学陆俭明教授的大力支持,在此一并致谢。

一　前　言

在机器翻译系统及其他自然语言处理系统中,通常都有一部包括语义信息的电子词典。为了给计算机自动分析提供更全面、深入的语义信息,我们应充分吸收现有的研究成果,在语法知识库的基础上构建语义知识库。不仅要进行系统的语义分类,而且要对词义组合信息加以全面描述,进一步加强动态的语义组合知识的研究和总结,建立一个与语言工程应用紧密配合的、合理的语义知识描述框架。

北京大学与中科院计算所自 1994 年联合开发“汉英机器翻译模型系统”开始,就着手研制为汉英机器翻译服务的“现代汉语语义词典”,目的是在语法分析的基础上,为计算机提供更深入的语义信息。1996 年至 1998 年,双方共同承担了国家 863 高科技项目“通用机器翻译开发平台和汉英机器翻译系统”课题(项目编号:863-306-03-06-2)。作为该课题的一个重要组成部分,“现代汉语语义词典”进入到大规模开发阶段,并取得了重要的阶段性成果,完成 4.9 万名词、动词、形容词的语义分类,并在配价理论的基础上,简要描述了其语义搭配限制(王惠等,1998)。从 2001 年开始,“现代汉语语义词典”的再开发受到国家 973 重点基础研究发展规划项目的支持,对词语的语义分类以及配价属性描述重新进行填写或修订。

二　内容概要

(一) 规模与结构

“现代汉语语义词典”收录了 66539 个通用领域内的实词,采用 Foxpro 8.0 实现,共有 12 个数据库,其中包含全部词语的总库 1 个,每类词语各建一库,计 11 个。每个库文件都详细刻画了词语及其语义属性的二维关系。总库中包括词语、拼音、同形、义项、语义类、词类、子类、兼类等 8 个字段。每类词的特有属性填在各类词库中,如名词库设 15 个属性字段,动词库设 16 个属性字段,如此等等。

表 1　语义词典规模

库名	词条	属性字段
名　词	37522	15
时间词	567	15
处所词	185	15
方位词	204	15
代　词	236	15
动　词	21142	16
形容词	3827	15
区别词	753	15
状态词	997	15
副　词	997	11
数　词	109	11
总　库	66539	8

表 2　名词库部分属性字段

词语	词类	同形	义项	语义类	配价数	参照体	对象	WORD	ECAT
老虎	n			动物	0			tiger	N
腿	n	1	1	生物构件	1	人/动物		leg	N
腿	n	2	2	非生物构件	1	用具		leg	N
意见	n	1	1	认知	2	人	实体| 抽象物	view	N
意见	n	2	2	认知	2	人	人| 事件	objection	N

(二) 词语的语义分类

国内外对汉语语义分类体系的研究已有不少成果,但由于各家分类体系的目的及应用范围不同,对同一事物可能有不同的定义与归类。如“动物”在一个语义体系中分为“兽类、鸟类、

鱼类、虫类、爬行类”,而在另一个体系中分为“脊椎动物、腔肠动物、软体动物”。但这些分类体系都是基于自然科学或常识而独立于语法的。在实际语言分析中,如何将这些语义知识与语法知识有机地结合起来是一件很困难的事情。

与这些基于常识的各种语义分类相比,“现代汉语语义词典”中语义分类的突出特点就是分类的深度与广度取决于语法分析的需要。应用语义知识应着重于解决那些仅靠语法规则难以解决的问题。因而语义分类是在词的语法分类基础上进行的,并且只对名词、动词、形容词等实词进行语义分类描述,而那些带有明显标志的、通常用句法形式就可以表示的语义关系,如各类虚词,则不作为语义分类研究的对象。

经过 4 年来的应用检验与研究,我们发现,对于中文信息处理来说,这种分类法是很有前途和实用价值的。为了更彻底地贯彻这个原则,同时便于与 Wordnet 和“中文概念辞书(CCD)”(于江生、俞士汶,2002)兼容,与“知网(hownet)”、《同义词词林》等已有的多种语义词典实现资源共享,我们在参照现有各家语义类的基础上,针对汉英机器翻译的需要,对语义词典(1998 版)的原分类体系作了较大的调整。总的来说,新的语义分类更趋合理,其特点是对名词的分类相对较细,动词、形容词的分类较粗,只要能揭示出与名词性成分、动词性组合成分的不同组合类型即可。目前我们已实际完成了 6.6 万词语的语义类划分与标注。具体分类体系如下:

1. 名词(Noun)
1.1 具体事物(entity)
1.1.1 生物(organism)
1.1.1.1 人(person)
1.1.1.1.1 个人(individual):职业 身份 关系 姓名
1.1.1.1.2 团体(group):机构 人群
1.1.1.2 动物(animal):兽 鸟 鱼 昆虫 爬行动物
1.1.1.3 植物(plant):树 草 花 庄稼
1.1.1.4 微生物(microbe):细菌 病毒 霉菌
1.1.2 非生物(object)
1.1.2.1 人工物(artifact):建筑物 衣物 食物 药物 创作物 计算机软件 钱财 票据 证书 符号 材料 器具
1.1.2.2 自然物(natural object):天体 气象 地理
1.1.2.3 排泄物(excrement):汗 尿 粪便 奶水 眼泪
1.1.2.4 外形(shape):粉末 长方形 圆 窟窿 孔 洞 泡
1.1.3 构件(part)
1.1.3.1 身体构件(body-part):头 脸 鼻子 嘴 耳朵 头发 血液 骨头
1.1.3.2 非生物构件(object-part):梁 屋檐 车闸 车筐
1.2 抽象事物(abstraction)
1.2.1 属性(attribute)
1.2.1.1 量化属性(measurable):体积 面积 重量 质量 价格
1.2.1.2 模糊属性
1.2.1.2.1 人性(property of human):胆量 勇气 脾气 作风
1.2.1.2.2 事性(description of event):境况 形势 状态 环节

1.2.1.2.3　物性(property of object):性能　效用　品种　式样
1.2.1.3　颜色(color):黑色　白色　浅色　素色
1.2.2　信息(information):话　言语　信件　口信　密码　声明　借口
1.2.3　领域(field):社会　经济　法律　科学　艺术
1.2.4　法规(rule):法律　条约　协议　制度　规章　合同　条文
1.2.5　生理(physiological state):瘟疫　疾病　炎症　艾滋病
1.2.6　心理特征(psychological feature)
1.2.6.1　情感(feelings):态度　感情　爱情
1.2.6.2　意识(cognition):意图　幻想　兴趣　主意　见解
1.2.7　动机(motivation):目的　原因　理由
1.3　过程(process)
1.3.1　事件(event):学潮　球赛　晚会　课　早餐　战争　火灾
1.3.2　自然现象(natural phenomenon)
1.4　时间(time)
1.4.1　绝对时间(specific time):宋朝　三国　清代
1.4.2　相对时间(relative time):昨天　当代　古代　今天
1.5　空间(space)
1.5.1　处所(location):浙江　西湖　黄山　中国　亚洲
1.5.2　方位(direction):东南　前面　之间　途中　高空
2. 形容词(Adjective)
2.1　事性值(description of event):紧急　突然　困难　容易　错误　费时
2.2　物性值(property of object)
2.2.1　量化属性值(measurable value)
2.2.1.1　浓度(concentration):浓　稀薄
2.2.1.2　温度(temperature):热　冷　凉爽
2.2.1.3　速度(speed):快　慢
2.2.1.4　长度(length):长　短
2.2.1.5　高度(height):高　矮　低
2.2.1.6　宽度(width):宽　窄
2.2.1.7　深度(depth):深　浅
2.2.1.8　厚度(thickness):厚　薄
2.2.1.9　硬度(rigidity):硬　软
2.2.1.10　湿度(humidity):潮湿　湿润　干燥
2.2.1.11　粗细(degree of finish):粗　细
2.2.1.12　松紧(degree of tightness):松　紧
2.2.1.13　大小(size):大　中　小
2.2.1.14　价值(value):贵　便宜
2.2.2　模糊属性值(unmeasurable value)
2.2.2.1　视感(vision):亮　醒目　清晰　混浊

2.2.2.2 触感(tactility):紧 松 粗糙 滑 柔
2.2.2.3 音质(tone):响亮 低沉 刺耳
2.2.2.4 味道(taste):酸 甜 苦 辣 可口
2.2.2.5 性质(quality):新 旧 真 假 好 坏 强 弱
2.2.2.6 内容(content):空洞 晦涩 清楚 浅显
2.2.2.7 外形(shape):方 圆 尖
2.2.3 颜色(color):红 黄 蓝 绿 鲜艳
2.3 人性值(property of human)
2.3.1 年龄(age):年轻 幼小 老
2.3.2 品格(character):善良 博学 幼稚 优雅
2.3.3 关系(relation):亲密 疏远 热情 冷淡
2.3.4 境况(condition):繁忙 贫穷 危险 疲劳
2.4 空间值(property of space)
2.4.1 一维值(one dimension):远 近
2.4.2 二维值(two dimensions):平 斜 弯
2.4.2 三维值(three dimensions):拥挤 杂乱 整齐 满 壮阔
2.5 时间值(property of time):古老 久远 短暂 早 晚
3.动词(Verb)
3.1 静态关系(state):是 有 等于 包括
3.2 心理活动(emotion/ cognition):喜欢 尊敬 反对 同意 怀疑 思考 判断
3.3 动态行为(event)
3.3.1 变化(change):死 病 下降 长高 缩小 变暗
3.3.2 气象(weather):下雨 刮风 打雷 起雾
3.3.3 身体活动(bodily care and functions):蹬 跳 推 笑 咳嗽 游泳
3.3.4 五官感觉(perception):看见 听到 闻着 品尝
3.3.5 消耗(consumption):吃 喝 饮
3.3.6 位移(motion):跑 走 散步 飞 过来 回去 拉来
3.3.7 创造(creation):制作 画 炒 写 创建 修筑
3.3.8 接触(contact):触摸 撞击 打中 系 挖掘
3.3.9 领属转移(possession):买 卖 赠送 给 转让 借
3.3.10 信息交流(communication):告诉 询问 请求 转达 叮嘱 说
3.3.11 比赛(competition):竞赛 赛跑 打仗 摔跤 辩论
3.3.12 社会活动(social behavior):改革 调价 开会 联欢
3.3.13 其他行为(other event)

(三)词语的语义属性描写

为了进一步提高机器翻译系统的性能,本词典在语义分类的基础上,进一步详细刻画了每个词的配价数及其在上下文中的语义搭配限制,见表3。

表 3　现代汉语语义词典动词库的属性字段

字段名	字段值
词语	1～4 个字的词语
拼音	填每个词语的汉语拼音,声调用“1,2,3,4,5”表示,其中“5”表示轻声。如:“常识”的全拼音是“chang2shi2”,“尺子”的全拼音是“chi3zi5”。
词类	填词语所属词类的代码。如:名词填“n”,动词填“v”,形容词填“a”。
子类	填词语所属词类的子类代码。如:名词性成语填“IN”,动词性习用语填“LV”。
兼类	填该词语兼属的词类代码,如:名词“锁”的兼类填“v”。
同形	对于字形、词类都相同但是应算不同词的情况,在本字段中填上字母 A,B,C,如“抄近道”的“抄”与“抄作业”的“抄”。为了提高处理效率,也用 A,B,C 等标识同字同类不同音的情况,如表示“加在一起”的“合计(he2ji4)”与表示“盘算、磋商”的“合计(he2ji5)”。
义项	对于同一个词的不同义项,填上数字 1,2,3。如“菜很清淡”中的“清淡”在本字段填“1”,“生意清淡”的“清淡”则填“2”。
释义	填写词语的简明释义。如:词典中收录两个“天才”,一个指人(“一位天才”),一个指“智慧”(“很有天才”),就在本字段分别填上“人”和“智慧”。
语义类	填写词语的语义类别名称。如“校长”填“身份”,“刀”填“用具”,“是”填“静态关系”,“喜欢”填“心理活动”,“打雷”填“气象”。可以不止填一个类别名称,不同的名称之间用“|”隔开,如“青菜”填“植物|食物”。
配价数	填写词语在上下文中所能搭配的名词数目,取值范围为 0、1、2、3。如:“大、儿子、咳嗽”仅能跟一个名词发生关联,如“声音大、老王的儿子、小李咳嗽”等,那么这些词的配价数就为 1。“热情、意见、吃”能跟两个名词发生关联,配价数就是 2。动词“给”可以跟三个体词发生关联,它的配价数即为 3。动词“例如”不跟任何成分搭配,它的配价数就是 0。
主体	指动作行为的发出者或性状的承当者。如“逃跑”在本字段填“人类|动物”,“刮倒”填“气象”,“死”填“生物”,“红”在本字段填“具体事物”。
客体	指动作行为所涉及的直接对象或性状的关涉对象。如“吃”在本字段填“食物”,“画”填“作品”,“眼熟”填“具体事物”,“有利”填“人类|事物”。
与事	事件中的受益者或受损者。如“给”在本字段填“人类”,“送”也填“人类”。
WORD	填写词语对应的英语译文,如“安静”在本字段填“quiet”,“脏乱”填“dirty and messy”。
ECAT	填写词语的英语译文的词性代码,或短语组成结构,如“安静”在本字段填“A”,“脏乱”则填“!A+C+!A”(!表示中心词)。
备注	填写词语某些用法的简明示例。

三　应用价值

“现代汉语语义词典”中的词义信息在汉语分析的各个层面,包括多义词义项判断、短语结构层次和结构关系判定以及成分之间语义关系的确定等等,都能起到重要的作用。在汉英机器翻译中,利用词义信息至少有两个显著作用:

(1) 在源语言句法分析过程中,排除一些歧义结构,有助于得到正确的句法结构;

(2) 在目标语生成过程中,进行词义消歧,在多义词的不同译法中挑选一个最合适的,提高译文质量。

前者已经有不少论述(王惠,2004;詹卫东、刘群,1997),这里不再赘述,本节将重点放在后者上,以具体实例介绍"现代汉语语义词典"在汉英机器翻译系统中词义消歧方面的应用。

词义消歧的第一步是确定哪些词是多义词。语义词典提供了非常简单的判断方法:只要"义项"、"同形"、"兼类"这 3 个字段中的任何一个填有内容,就说明当前的词条是一个多义词,需要进行词义消歧。

如果一个词的多个义项属于不同的语义类,那么,它们在句子中所受到的组合限制也相应地不同。对动词来说,主要表现在动作的发出者、动作对象的差异上;对形容词而言,则是修饰对象的语义类不同。"现代汉语语义词典"对这些都作了具体描述。如:

表 4 现代汉语语义词典中的多义形容词

词语	词类	释义	义项	语义类	主体	WORD
清淡	a	(气味)清而淡	1	气味	食物| 植物	light
清淡	a	营业数额少	2	境况	"生意"	slack

如果遇到以下经过切分、标注的文本:

[1]清淡/a 的/u 荷花/n 香气/n

[2]农忙时/t 进城/v 的/u 人/n 不/d 多/a,生意/n 比较/d 清淡/a。

句[1]中"清淡"后面的名词是"荷花",属于"植物"类;句[2]中"清淡"的修饰对象是"生意"。根据"主体"字段的信息,计算机就可准确地判断出这两个"清淡"属于不同的语义类,前一个属于义项 1,应译为"light",后一个只能与"生意"搭配,则译为"slack"。

经过词类与语义类两步筛选,可以完成绝大部分的汉语多义词消歧。但还有少数多义词,其内部各义项的词类、语义类均相同,如:

表 5 动词"找"不同义项的语义搭配

词语	词类	同形	释义	语义类	主体	客体	与事	WORD	备注
找	v	A	寻找	对待	人	具体事物		look for	～材料
找	v	B	退还	对待	人	" *钱"	人	give change	～钱

由表 5 可见,"寻找"的"找"在句子中只带一个宾语,而且这个宾语只能由表示"具体事物"的名词充当,而"找钱"的"找"后面可以跟两个 NP,一个仅限于"钱",另一个则必须属于语义类"人"。即:

找 A 右组合:～+名词(具体事物"狗、自行车、房子"……)

找 B 右组合:～+名词/人称代词(人"主任、小李、你"……)+名词("钱")

根据这个搭配特征,计算机可以正确判断出下面例句中"找"的词义:

[1]我们/r 出去/v 再/d 找/v 一/m 块/q 实验地/n。

[2]营业员/n 找/v 我/r 20/m 元/q 钱/n。

例[1]中的"找"后面只有一个名词"试验地",属于"具体事物",因而,是"找 A",应译为"look for";例[2]中的"找"后面有一个人称代词"我",还有一个名词"钱",显然符合"找 B"的组合条件,应选择"give change"作为译文输出。

四　结　语

作为综合语言知识库的一个组成部分,“现代汉语语义词典”不仅可以应用于机器翻译,而且还可以在多种 NLP 系统(如自然语言接口、文献检索、信息自动提取、语音识别与合成、文本校对、语料库加工等)的语义分析中发挥重要作用。同时,对于促进汉语词汇与语义学研究、开展汉语词义定量分析等也有很大的价值。

目前,本项研究已取得了可观的阶段性成果,词典规模扩大到了 6.6 万词语,质量也有了显著提高,并已在一个汉英机器翻译系统中得到实际应用。但语义词典的开发毕竟是一项长期的语言工程,不可能毕其功于一役。我们在实践检验中还应不断地发现问题,总结经验,逐渐完善现有的语义分类体系及属性描写。同时,从大规模语料中自动抽取更多的语义搭配知识,检验并丰富我们现有的语义约束描述,在计算词义学方面进行更深入的探索。

[参考文献]

[1] Christiane Fellbaum. ed.. *WordNet: an electronic lexical database*[M]. Mass: MIT Press,1998.

[2] 王　惠,詹卫东,刘　群. 现代汉语语义词典的设计与概要[A]. 1998 中文信息处理国际会议论文集[C]. 北京:清华大学出版社,1998:361～367.

[3] 俞士汶,朱学锋,王　惠等. 现代汉语语法信息词典详解(第 2 版)[M]. 北京:清华大学出版社,2003.

[4] 于江生,俞士汶. CCD 的结构与设计思想[J]. 中文信息学报,2002,(4):12～20.

[5] 董振东,董　强.“知网”(Hownet)[R]. http://www.keenage.com.

[6] 王　惠. 现代汉语名词词义组合分析[M]. 北京:北京大学出版社,2004.

[7] 詹卫东,刘　群. 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[A]. 语言工程[C]. 北京:清华大学出版社,1997:286～291.

[8] Jia-Lin Tsai, Wen-Lian Hsu and Jeng-Woei Su. Word Sense Disambiguation and Sense-Based NV Event Frame Identifier[J]. *Computational Linguistics and Chinese Language Processing*, 2002,(1):29～46.

《中文助教》软件面世

《中文助教》(ChineseTA)是一个针对汉语教师编写教材和日常备课的实际需要而开发的现代化工具软件。该软件由美国斯坦福大学资助,储诚志博士主持设计,美国硅谷语言技术公司(www.svlanguage.com)制作发行。北京语言大学出版社经美国硅谷语言技术公司授权,于 2005 年 12 月在中国大陆出版了《中文助教(1.1 版)》汉化中文简体字版。

《中文助教》软件在一个简单的界面中集成了很多实用功能,包括课文加注拼音、生词生字查找、词表字表注释、汉字繁简转换对照、字词分布索引、字词频率统计、生词密度和重现率标示、字词 HSK 等级和常用度标示、新词旧词关联、词语随文注音翻译、课文改换顺序、词表字表项目的选择和排序,等等。

《中文助教》软件在美国问世后反响很大。美国多所大学从事汉语教学的专家给予了高度评价:“过去要花几天时间才能做完的事用《中文助教》几秒钟就能完成。”《全美中小学中文教师协会通讯》报道说,该软件能方便地帮助老师“编写、修改和评量教学材料以适应他们自己学生的学习需要”。

垂询信箱:sgxn@blcu.edu.cn。