“现代汉语语义词典”的结构及应用♣

王惠 俞士汶 詹卫东

（北京大学计算语言学研究所）（北京大学中文系）

whui@pku.edu.cn; yusw@pku.edu.cn; zwd@pku.edu.cn

摘要：“现代汉语语义词典”是一部面向汉英机器翻译的语言知识库，它以数据库文件形式收录了 6.8 万余条现代汉语的实词，不仅给出了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，并且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问题。

关键词：语义词典 语义类 词义消歧 计算词典学 中文信息处理

The Semantic Knowledge-base of Contemporary Chinese and its Applications in MT

Wang Hui[1], Yu Shiwen[1], Zhan Weidong[2]

[1]（Institute of Computational Linguistics, Peking University, Beijing 100871, China）

[2]（Dept. of Chinese Language and Literature, Peking University, Beijing 100871, China）

whui@pku.edu.cn; yusw@pku.edu.cn; zwd@pku.edu.cn

Abstract

The Semantic Knowledge-base of Contemporary Chinese (SKCC) is a machine-readable dictionary, which provides rich semantic information such as a thesaurus, semantic collocation, and parts-of-speech (POS) of 68,000 Chinese words. The paper introduces the outline of SKCC, and indicates that it is effective for word sense disambiguation in MT system and is likely to be important for general Chinese NLP.

Key words: Semantic knowledge-base, Chinese thesaurus, computational lexicography, word sense disambiguation（WSD）, Chinese language information processing

♣ 本研究得到国家 973 重点基础研究项目“面向新闻领域的汉英机器翻译系统”（项目号：G1998030507-4）与“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”（项目号：G1998030507-1）的支持。

1. 开发历程

在机器翻译系统及其它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。北京大学计算语言学研究所与中科院计算所自 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，完成了 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[2]。这些语义信息在汉英机器翻译系统中，为词义和句法歧义的消解提供了强有力的支持。

从 2001 年开始，“现代汉语语义词典”的再开发受到了国家 973 重点基础研究发展规划项目“面向新闻领域的汉英机器翻译系统”和“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”的支持，由北京大学计算语言学研究所和中文系合作进行，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。在双方的积极努力下，项目进展得非常顺利，不仅使词典的规模有了较大幅度的扩充，而且质量也有了显著的提高。目前，正在一个具体的机器翻译系统中发挥着重要作用。

2．内容概要

2.1 规模与结构

“现代汉语语义词典”收录了 66,539 个通用领域内的实词语，采用 Foxpro 6.0 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。

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**

总库与各类词库可以通过“词语、词类、同形、义项”这 4 个关键字段进行链接，从而使 13 个库文件构成有上下位继承关系的“树”，子结点可继承父结点的全部信息。

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。

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

表 1　名词库部分属性字段

2.2 词语的语义分类

“现代汉语语义词典”的一个突出特点就是其语义分类的深度与广度取决于语法分析的需要,而非基于语义常识。经过 4 年来的应用检验与研究，我们发现，对于汉语信息处理来说，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[1]和“中文概念辞书（CCD）”[4]兼容，与“知网（hownet）”[5]、《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。

- 名词上下位关系更加系统化：首先，将具体事物、抽象事物与过程、时间、空间并列为 5 大类；然后再逐层细分：具体事物分为生物、非生物 2 类，生物里再把人与动物、植物、微生物相并列，非生物中则进一步区分开人工物、自然物、排泄物和外形。然后，根据 Wordnet 与“知网”中的填写内容，补充了一些较低层的名词小类，如“人工物”的下位概念（“建筑物、衣物、食物、药物、钱财、票据、证书”等）。
- 把 Wordnet 中的动词分类借鉴过来，但根据汉语的实际作了相应改造；
- 形容词的分类更加细化，由原来的 7 类发展成为现在的 5 大类 29 小类，与名词的分类互相照应，从而可以更细致地刻画形名搭配关系。

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、数词、副词的分类较粗，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 6.8 万词语的语义类划分与标注。具体分类体系如下：

（1）名词（包括时间词、处所词、方位词）

1　具体事物（entity）
　1.1　生物（organism）
　　1.1.1 人（person）
　　1.1.1.1 个人（individual）
　　1.1.1.1.1 职业（profession）：教师　秘书　会计　医生
　　1.1.1.1.2 身份（identity）：华侨　外行　健将　模范
　　1.1.1.1.3 关系（relation）：父亲 阿姨 长辈 朋友
　　1.1.1.1.4 姓名（name）：爱因斯坦　毛泽东　鲁迅
　　1.1.1.2 团体（group）
　　1.1.1.1.1 机构（organization）：工厂 医院 商店 剧团
　　1.1.1.1.2 人群（society）：人民 委员会 少先队 团伙
　　1.1.2 动物（animal）
　　1.1.2.1 兽（beast）：狗 猪 牛 羊 老虎 豹子 狐狸
　　1.1.2.2 鸟（bird）：鸡 鸭 麻雀　杜鹃
　　1.1.2.3 鱼（fish）：鲤鱼　河豚　鲸　泥鳅

1.1.2.4 昆虫（insect）：蚯蚓　知了　蟑螂
1.1.2.5 爬行动物（reptile）：青蛙　乌龟　甲鱼　蛇
1.1.3 植物（plant）：树　花　草　牡丹　芍药
1.1.3.1 树（tree）：白杨 水杉 芭蕉
1.1.3.2 草（grass）：狗尾巴草　含羞草　蒲公英
1.1.3.3 花（flower）：牡丹 芍药 杜鹃 映山红
1.1.3.4 庄稼（crop）：蔬菜 小麦 高粱 棉花
1.1.4 微生物（microbe）：细菌　病毒　霉菌
1.2　非生物（object）
1.2.1 人工物（artifact）
1.2.1.1 建筑物（building）：别墅　礼堂　会议室　水库　庙
1.2.1.2 衣物（clothes）：服装　外套　衬衫　裙子 帽子
1.2.1.3 食物（food）：面包　牛奶　菜　米饭　饮料
1.2.1.4 药物（drug）：药片 阿斯匹林 酒精　镇定剂
1.2.1.5 创作物（works）：论文 书 杂志 文章 油画 电影
1.2.1.6 计算机软件（software）：操作系统 数据库 程序 软件
1.2.1.7 钱财（asset）：财产 钱 资金 报酬　罚款 美元 利息
1.2.1.9 票据（bill）：发票 单据　汇票　支票　包裹单
1.2.1.10 证书（certificate）：结婚证 执照 毕业证 驾驶证
1.2.1.11 符号(symbol)：签名 路标 箭头 句号
1.2.1.12 材料（material）：木材　钢铁　煤炭　玻璃　水泥
1.2.1.13 器具（instrument）
1.2.1.13.1 用具（tool）：剪子 刀子 钉子 拖把 改锥
1.2.1.13.2 交通工具（vehicle）：车 船 飞机 自行车
1.2.1.13.3 武器（weapon）：大炮 机关枪　鱼雷
1.2.1.13.4 家具（furniture）：桌子 椅子　沙发
1.2.1.13.5 乐器（musical-instrument）：钢琴 吉他 鼓
1.2.1.13.6 电器（electricity）：电视 空调 电冰箱
1.2.1.13.7 文具（stationery）：钢笔 橡皮 尺子
1.2.1.13.8 运动器械（sports- instrument）：足球 单杠
1.2.2 自然物（natural object）
1.2.2.1 天体（celestial body）：太阳　月亮　流星　星星
1.2.2.2 气象（weather）：云　彩虹　晚霞
1.2.2.3 地理（geography）
1.2.2.3.1 地表物（land）：原野　沙漠　山 山洞　陆地
1.2.2.3.2 水域物（water）：江　河　湖　海 河流
1.1.2.2.4 矿物（mineral）：煤矿　原油　铁矿
1.1.2.2.5 元素（element）：金 银 铜 铁
1.1.2.2.6 基本物质（substance）：水 土 灰
1.2.3 排泄物（excrement）：汗 尿 粪便 奶水 眼泪
1.2.4 外形（shape）：粉末 长方形 圆　窟窿　孔　洞　泡
1.3 构件（part）
1.3.1 身体构件（body-part）：头 脸 鼻子 嘴 耳朵 头发 血液 骨头
1.3.2 非生物构件（object-part）：梁 屋檐 车闸 车筐

2　抽象事物（abstraction）
2.1 属性（attribute）
2.1.1 量化属性（measurable）：体积　面积　重量　质量　价格
2.1.2 模糊属性
2.1.2.1　人性（property_of_human）：胆量　勇气　脾气　作风
2.1.2.2　事性（description_of_event）：境况　形势　状态　环节
2.1.2.3　物性（property_of_object）：性能　效用　品种　式样
2.1.3 颜色（color）：黑色　白色　浅色　素色
2.2 信息（information）：话　言语　信件　口信　密码　声明　借口
2.3 领域 (field)：社会　经济　法律　科学　艺术
2.4 法规（rule）：法律　条约　协议　制度　规章　合同　协议　条文
2.5 生理（physiological_state）：瘟疫　疾病　炎症　艾滋病
2.5 心理特征（psychol feature）
2.5.1 情感（feelings）：态度　感情　爱情
2.5.2 意识（cognition）：意图　幻想　兴趣　主意　见解
2.6 动机（motivation）：目的　原因　理由
3　过程（process）
3.1 事件（event）：学潮　球赛　晚会　课　早餐　战争　火灾
3.2 自然现象（natural phenomenon）
3.2.1　可视现象（visible phenomenon）：火　电　光　风　雨
3.2.2　可听现象（audible phenomenon）：声音　雷鸣　风暴
4 时间（time）
4.1　绝对时间（specific time）：宋朝　三国　清代
4.2　相对时间（relative time）：昨天　当代　古代　今天
5 空间（space）
5.1 处所（location）：浙江　西湖　黄山　中国　亚洲
5.2 方位（direction）：东南　前面　之间　途中　高空

（2）形容词（包括区别词、状态词）

1 事性值：紧急　突然　困难　容易　错误　费时
2 物性值
2.1　量化属性值（measurable value）：
2.1.1　浓度（concentration）：浓　稀薄
2.1.2　温度（temperature）：热　冷　凉爽
2.1.3　速度（speed）：快　慢
2.1.4 长度(length)：长　短
2.1.5　高度(height)：高　矮　低
2.1.6　宽度(width)：宽　窄
2.1.7　深度(depth)：深　浅
2.1.8　厚度(thickness)：厚　薄
2.1.9　硬度(rigidity)：硬　软
2.1.10 湿度(humidity)：潮湿　湿润　干燥
2.1.11 粗细(degree of finish)：粗　细
2.1.12 松紧(degree of tightness)：松　紧

2.1.13 大小(size)：大 中 小
2.1.14 价值(value)：贵 便宜
2.2 模糊属性值（unmeasurable value）
2.2.1 视感(vision)：亮 醒目 清晰 混浊
2.2.2 触感(tactility)：紧 松 粗糙 滑 柔
2.2.3 音质（tone）：响亮 低沉 刺耳
2.2.4 味道(taste)：酸 甜 苦 辣 可口
2.2.5 性质（quality）：新 旧 真 假 好 坏 强 弱
2.2.6 内容(content)：空洞 晦涩 清楚 浅显
2.2.7 外形(shape)：方 圆 尖
2.3 颜色(color)：红 黄 蓝 绿 鲜艳
3 人性值
3.1 年龄(age)：年轻 幼小 老
3.2 品格(character)：善良 博学 幼稚 优雅
3.3 关系(relation)：亲密 疏远 热情 冷淡
3.4 境况(condition)：繁忙 贫穷 危险 疲劳
4 空间值
4.1 一维值：远 近
4.2 二维值：平 斜 弯
4.2 三维值：拥挤 杂乱 整齐 满 壮阔
5 时间值：古老 久远 短暂 早 晚

（3）动词
1 静态关系（state）：是 有 等于 包括
2 心理活动（emotion/ cognition）：喜欢 尊敬 反对 同意 怀疑 思考 判断
3 动态行为（event）
3.1 变化（change）：死 病 下降 长高 缩小 变暗
3.2 气象（weather）：下雨 刮风 打雷 起雾
3.3 身体活动（bodily care and functions）：蹬 跳 推 笑 咳嗽 游泳
3.4 五官感觉（perception）：看见 听到 闻着 品尝
3.5 消耗（consumption）：吃 喝 饮
3.6 位移（motion）：跑 走 散步 飞 过来 回去 拉来
3.7 创造（creation）：制作 画 炒 写 创建 修筑
3.8 接触（contact）：触摸 撞击 打中 系 挖掘
3.9 领属转移（possession）：买 卖 赠送 给 转让 借
3.10 信息交流（communication）：告诉 询问 请求 转达 叮嘱 说
3.11 比赛（competition）：竞赛 赛跑 打仗 摔跤 辩论
3.12 社会活动（social behavior）：改革 调价 开会 联欢
3.13 其他行为（other event）

（4）副词
1 程度（degree）：很 挺 太 顶 更 最 极 十分 非常 稍 稍微 略微
2 范围（range）：都 也 总 共 一共 总共 统统 只 就 光 仅 仅仅
3 时间（time）：正 刚刚 就 先 曾经 已经 终于 立刻 马上 永远
4 处所（location）：到处 处处 暗中 当场 当面
5 频度（frequency）：常常 常 时常 又 再 还 重新 重

6 方式（manner）：渐渐 逐渐 挨次 挨个 逆时针 慢慢
7 否定（negation）：不 没有 没 未 莫 休 勿 别
8 语气（modality）：却 倒 竟 偏偏 都 简直 索性 幸亏 难道 究竟

（5）数词
1 基数（cardinal number）
1.1 系数：一 二 两 三 五 六 七 八 九 几
1.2 位数：十、百、千、万、亿、万万
2 序数（ordinal number）：第一 第二 第十
3 概数（amount）：多半 多少 若干 很多 许多 好多 好几 好些 无数
4 助数：又 来 左右

2.3 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以及其在上下文中的语义搭配限制。

为了对语义词典的属性描述有个比较完整的认识，下面以动词库为例简要地介绍各字段的名称及属性值：

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，则填上数字 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	填写该词语的简明释义。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	填写该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	填写该词在上下文中所能搭配的名词的数目，取值范围为 0、1、2、3。例如：“大、儿子、咳嗽”都是能且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，如“老王对同学们很热情、老王对小李的意见、孩子吃苹果”，因而这些词的配价数就是 2。动词“给”可以跟三个体词发生关联。如“老师给了学生一本书”，它的配价数即为 3。而动词“例如”经常独用，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（一种颜色）”在本字段填“具体事物”，“友好（亲近和睦）”填“人类|动物”。
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|象事物”。
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	填该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	填该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。

表 2 现代汉语语义词典动词库的属性字段

上述16项属性大致可归纳为以下4类：

（1）词语本身的一些基本特征，如词语、词类、同形、拼音、兼类、用例等。它们与北大计算语言学研究所开发的“现代汉语语法信息词典”（2002 版）[3]保持严格的对应关系，这不仅保证了语义词典收词的规范性、注音与词性标注的准确性，而且也使得两部词典可通过“词语、词类、同形”三个关键字段相互链接配合使用，使计算机获得更加完备的语法、语义信息。

（2）词语意义的基本刻画，如词条的语义类、词义解释、是否属于多义词的一个义项等。这些属性为汉语词义消歧和词义研究提供了丰富的知识。

（3）描述一个词语跟其他实词发生语义联系的能力，主要包括动词的配价数（能支配多少名词性成分）、配项成分的语义角色（主体、客体、与事）和语义约束几个方面。这是语义词典的重点开发内容，可直接服务于计算机语义自动分析。

（4）每个词的英语译词及其词类，若对应的是英语短语，还要指出其中心词。如果一个词在不同的上下文中具有不同的翻译，则要求通过语义限制描述指出其翻译条件，从而为机器翻译系统的译文选择提供充分的根据。

3 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[6][7]，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”在汉机器翻译系统中词义消歧方面的应用。

3.1 多义词范围的确定

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。

首先，如果一个词语具有多个义项，分别对应不同的英语译文，那么，语义词典就将其分为不同的词条，一个词条对应一个义项，同时在“义项”字段填入相应的义项编号。如：

词语	词类	义项	释义	WORD	备注

菜	n	1	蔬菜	vegetable	种～丨野～
菜	n	2	经过烹调供下饭下酒的蔬菜、蛋品、鱼、肉等	dish	荤～丨四～一汤

表 3　多义词“菜”的两个义项

其次，即使属于同一个义项，但若对应不同的译文，语义词典也把它们分为不同的词条，并在“主体”、“客体”字段指出其搭配特征。如：

词语	词类	释义	主体	客体	WORD
看	v	使视线接触人或物	人	电影 | 物体 | 生物	see
看	v	使视线接触人或物	人	电视 | 比赛	watch
看	v	使视线接触人或物	人	书 | 报纸 | 杂志	read

表 4　动词“看”的不同译词及其语义搭配限制

除了上面所说的“一词多义”或“一义多译”的情况以外，汉语中还存在不少“同形词”，即仅仅字形相同但意义却毫无联系。可是，对计算机来说，它们与一词多义、一词多类现象是没有什么区别的，都是一个词形映射到不同意义上。因而，语义词典把同形词也当作广义的多义词，作为不同的词条分别收录。词类不同的，归入不同的库文件，但在“兼类”字段加以注明；词类相同的，放在同一个库中，但在“同形”字段标记上“A、B”等字母，如：

词语	拼音	同形	释义	WORD	备注
看	Kan4	A	使视线接触人或物	see	～到小李了
看	Kan1	B	守护照料	foster care	～门丨～孩子

表 5　现代汉语语义词典中的同形词

由此可见，只要“现代汉语语义词典”中的“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。

3.2　利用语义类进行词义消歧

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：

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

表 6　现代汉语语义词典中的多义形容词

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”。

3.3　利用语义搭配特征进行词义消歧

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：

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

表 7 动词“找”不同义项的语义搭配

由表中可见，“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。

4．结 语

作为北大计算语言学研究所的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 6.8 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，在计算词典学方面也进行更深入的理论探索。

参考文献：

[1] Christiane Fellbaum. ed.. *WordNet: an electronic lexical database*. 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”.《1998 中文信息处理国际会议论文集》. 清华大学出版社. 1998. pp 361-367.

[3] 俞士汶，朱学锋，王惠，张化瑞等. 《现代汉语语法信息词典详解（第 2 版）》.清华大学出版社. 2002

[4] 于江生，俞士汶. “CCD 的结构与设计思想”.《中文信息学报》. 2002, 16 (4). pp 12-20.

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

[6] 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”.《语

言工程》. 清华大学出版社. 1997. pp 286－291.

[7] Jia-Lin Tsai, Wen-Lian Hsu and Jeng-Woei Su. “Word Sense Disambiguation and Sense-Based NV Event Frame Identifier”. *Computational Linguistics and Chinese Language Processing*, Vol. 7, No. 1, 2002, pp. 29-46.