语言文字应用 1998年第 3期(总第 27期)

语义关系的表达和知识系统的建造

董振东

提要 本文介绍的是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。为了行文方便,以下简称"董氏电子知识系统"。董氏电子知识系统由两部分组成,一是大规模的知识数据,称之为知识词典;另一是用以处理知识数据的程序。

一、概论

该系统的建设是利用计算机进行的,而它的使用也必须依靠计算机。它完全是一个面向计算机的,服务于信息处理的系统。董氏电子知识系统所包含的两种语言是相对独立的。两种语言的词语的对应是建立在同一的属性描述基础上的。这点与本系统的研究与开发的目的和过程有关。董氏系统的研究与建设是从汉语开始的。但它的作者相信既然是一个知识系统,它就应该能够涵盖用以表达知识的各种的语言。因此在知识词典建设中,当汉语部分完成时便开始英语部分的建设。这也是为了使词典的知识表描述体系能得到验证或者在验证的过程中加以修正、改进和充实。经验证明,只基于一种语言容易产生片面性或者缺陷。汉语和英语有较大的差异,这更有助于考验描述体系的可靠性和达成知识表达的全面性。

1. 董氏系统的概貌

董氏系统包括各种数据文件和程序。下面对其中的主要文件做一个简要介绍。

(1)汉- 英双语知识词典(DONGS′C-E)

语种	词语总量	N 范畴	V 范畴	A 范畴
汉语	33069	13102	11768	8199
英语	38744	17479	13728	7537
语种	概念总量	N 范畴	V 范畴	A 范畴
汉语	41791	15840	15768	10183
英语	48831	22334	17294	9203

表 1

这是董氏系统的知识词典的最基本,最重要的数据文件。在此基础上,利用知识词典管理程序可以生成汉英- 英汉双语义类词典及相关的查询软件,或根据需要也可以生成单语的汉语知识词典或单语的英语知识词典。董氏电子知识系统的规模主要取决于双语知识词典数据文件的大小。由于它是在线的,修改和增删都很方便,因此它的规模是动态的。它的规模通常以词语的条数以及由词语所表述的概念或称义项的条数计算。它现有规模如表 1所示。

下面给出 N、V、A 各范畴的几个实例。(DEF 下面置于括号内的解释是我们为了帮助读者理解属性描述的内容而写的,在实际的数据文件中的是没有的。)

W_C= 医生
G_C= NOUN
E_C=
W_E= doctor
G_E= NOUN
E_E=
DEF= human 人类, medical 医, *cure 医治,# disease 疾病, addressable 称
(人类类别的一员,是医药领域的,是施行医治行为的,医治的是疾病,可以作为一种称呼)

W_C= 医院
G_C= NOUN
E_C=
W_E= hospital
G_E= NOUN
E_E=
DEF= institute-place 场所, + cure 医治,# disease 疾病, medical 医

W_C= 患者
G_C= NOUN
E_C=
W_E= patient
G_E= NOUN
E_E=
DEF= human 人类, * suffer-from 罹患,# medical 医, $cure 医治, undesired 莠
(人类类别的一员,是经受病的,与医药领域有关,是行为动作"医治"的受事,这是人们所不期望的)

W_C= 看病
G_C= VERB
E_C=
W_E= see a patient
G_E= VERB
E_E=
DEF= request 要求, result-event = cure 医治,# medical 医
(一种行为动作,类别是"要求",此概念自身已含有了一个必备角色即结果性动作是"医治",与医药领域有关)

W_C= 看病
G_C= VERB
E_C=
W_E= see a doctor
G_E= VERB
E_E=
DEF= cure 医治, content = disease 疾病, medical 医
(一种行为动作,类别是"医治",此概念自身已含有了一个必备角色即内容是疾病,是医药领域的)

W_C= 健壮
G_C= ADJ
E_C=
W_E= tough
G_E= ADJ
E_E=
DEF= situation-value 状况值, physique 体格, strong 强, desired 良

W_C= 防潮
G_C= ADJ
E_C=
W_E= moistureproof
G_E= ADJ
E_E=
DEF= property-value 特性值, ability 能力, able 能, * withstand 抗住,# wet 湿

(2) N- 规范(DONGS' N-Taxonomy)

N 范畴概念的类与次要属性的规定。下面是它的一部分:

……

NOUN.1.1.1.1.1	animal-human 动物	[!sex,!age,*action,*state-mental]
NOUN.1.1.1.1.1.1	human 人类	[!name,!wisdom,!ability,!occupation,*change-mental]
NOUN.1.1.1.1.1.1.1	humanized 拟人	
NOUN.1.1.1.1.1.2	animal 兽类	[^ * get-knowledge]

……

(3)V- 规范 (DONGS' V-Taxonomy)

V 范畴概念的类与次要属性的规定。下面是它的一部分:

……

V.1.2.1.6.3.4	suffer-from 罹患	[experiencer,content]
V.1.2.1.6.3.4.1	ill 病态	[experiencer]
V.1.2.1.6.3.4.1.1	wounded 受伤	[experiencer]

……

V.2.2.2.2.1.3.3	resume 恢复	[agent,patient,state-ini,state-fin]
V.2.2.2.2.1.3.3.1	recovering 修复	[agent,patient,state-ini,state-fin]
V.2.2.2.2.1.3.3.1.1	cure 医治	[agent,patient,content]
V.2.2.2.2.1.3.3.1.2	repair 修理	[agent,patient,content]

……

置于方括号中的是对于该类的绝对必要角色的框架,如对于 "医治"而言,当该事件发生时,必然会有 "施事"(给人看病的人), "受事"(被医治的人), "内容"(被医治的病)。

缺了上述任何一个,事件将不成立。这并不是说实际的句子里必然同时显性地出现这三个角色。

(4) A- 规范 (DONGS' A-Taxonomy)

A 范畴概念的类与次要属性的规定。下面是它的一部分:

……

measurement-value 量度值

length 长度	long 长	short 短
height 高度	tall 高	low 矮
size 尺寸	big 大	medium 中
small 小	broad 广	

……

property-value 特性值

ability 能力	able 能	unable 庸	
tolerance 气量	generous 慷	miser 吝	
……			

(5)汉- 英双语义类词典 (DONGS' Thesaurus)

这是一个类似 Word 中的同义词库的软件,但它与 Word 相比,它是双语的,再者它不仅有同义,还有反义(如 "大"和 "小")和对义(如 "买"和 "卖")。从上面的实例可以看出同义,反义和对义词并未标注在相关的概念记录上。它们是动态产生的。

2. 数据文件的记录样式

双语知识词典是董氏电子知识系统的基础文件,也是加工文件。我们以它为例对记录样式加以说明。

在这个文件中每一个词语的概念或称义项及其描述形成一个记录。每一个记录有 8 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。它们排列如下:

W_C= 汉语词语
E_C= 汉语词语例子
G_C= 汉语词语词性
W_E= 英语词语
E_E= 英语词语例子
G_E= 英语词语词性
Def= 概念类别和属性

对于上述各项的详细说明如下:

(1) W_C — 汉语词语,本系统选录词语的出发点是服从现代信息处理的需要。另外包括英语词语的译语。

(2) E_C — 汉语词语例子,所有的具有同一词形但概念不同的词语,亦即多义词语都要给出5~6个例子。给出的例子应该是短小精悍、常用、有较强的区别性。

(3) G_C — 汉语词语词性,基本根据应是"信息处理用现代汉语词典"。

(4) W_E — 英语词语,本系统选录词语的出发点是服从现代信息处理的需要。另外包括汉语词语的译语。

(5) E_E — 英语词语例子,所有的具有同一词形但概念不同的词语,亦即多义词语都要给出5~6个例子。给出的例子应该是短小精悍、常用、有较强的区别性。

(6) G_E — 英语词语词性,基本上根据 Longman Dictionary of Contemporary English。

(7) Def — 概念类别和属性,本系统知识词典的最重要的信息。任何一个词语务必有类别,类别只可以有一个。类别一定要放在首位,与属性之间用逗号分开。属性可以有多个,中间用逗号分开。

二、董氏系统的哲学

要掌握和利用好董氏系统,必须首先了解董氏系统的哲学思想。

董氏系统的哲学也就是它对客观世界的认识。这一哲学的根本点是:世界上一切事物(物质的,精神的或事情)都在一定的时间和空间内不停地运动和变化。它们通常是从一种状态变化到另一种状态,并通常由其属性值的改变来体现。

1. 董氏系统的高层分类

董氏系统的哲学贯穿于整个知识系统的各个方面,是系统的灵魂。它首先体现在它对概念的主要属性的确定上。概念的主要属性也是概念分类的类别。董氏系统的高层分类如下面所示。

N.1 entity 实体
  N.1.1 thing 万物
  N.1.2 time 时间
  N.1.3 space 空间
  N.1.4 component 部分
N.2 attribute 属性
N.3 unit 单位
V event 事件
  V.1 event-static 静态
    V.1.1 static-relation 关系
    V.1.2 static-state 状态
  V.2 action 行动
A.1 attribute-value 属性值

董氏系统的最高层有三类,即N 范畴,V 范畴,A 范畴。N 范畴包含实体,属性和单位。其中实体的直接下位有万物、时间、空间、部分。事件的直接下位有静态事件和行为动作。万物包括物质,精神和事情三类,它们通常是运动和变化的主体。

而这种运动和变化总是发生在一定的时空之中。运动和变化常体现于属性,并由属性值显示。

例如,"一个孩子上学学习",这是一个事件。

孩子”(万物中的物质)是事件的主体。“上学学习”是一种改变自身知识状态的行为动作。这个事件必然发生在一个特定的时空里。孩子经过这个行为动作由无知状态进入有知的状态。人的知识的多少或所谓的文化程度是人的一种属性。

如上列所示,与实体、事件和属性值并列的有属性这一高层类别。在董氏系统的哲学看来,属性是非常重要的一个类。我们认为属性看起来似乎无形,但是它们又是无处不在的。无论是实体还是事件都含有某些属性,都可以分解成各种属性。它们是属性的宿主。一个人可以有性别,年龄,国籍,健康状况,文化程度,智力,性格等等属性;一件东西可以有大小、重量、颜色、质量、用途等等属性;一个动作如“移动”可以有速度、方向、空间、工具等等属性。世界上不存在没有属性的东西,也不存在游离于任何宿主的属性。

在董氏系统中,属性跟属性值有着严格的对应。有什么类的属性就有什么类的属性值。世界上不存在没有值的属性,也不存在不指向任何属性的属性值。例如,“聪明”是一个属性值,一个指向“智力”这一属性的属性值。

高层分类与词类的对应

	汉　语	英　语		汉　语	英　语
(1) 实体	名词	名词 由名词派生的形容词	(6) 部分	名词	名词
(2) 属性	名词 个体量词	名词	(7) 单位	个体量词 单位量词	名词
(3) 时间	名词 时间词	名词	(8) 事件	动词 部分助动词	动词 部分助动词 部分表示事件的名词 部分表示情感的形容词
(4) 空间	名词 方位词	名词	(9) 属性值	全部性质形容词 数词	全部性质形容词 数词 由性质形容词派生的副词 由形容词派生的名词
(5) 事情	名词	名词			

表 2

既然董氏系统包括了一个知识词典。从语言学角度看,它描述的是词语的语义。董氏系统的高层分类也可以视为词语的语义分类。为了用户便于理解,词典给出了词语的最基本的句法信息如词性。但这只是辅助性或参考性的。董氏系统的高层分类与汉语和英语两种词类的对应有它自己的特点。大体上可以说其中的实体、属性、时间、空间、事情、部分和单位各类对应于名词,事件类对应于动词和部分形容词,属性值类对应于形容词和副词。N、V、A 三个规范的名称即来源于此。高层分类与词类的对应如表 2 所示。

从表 2 我们知道不可以把 V 范畴,A 范畴和 N 范畴简单地误解为句法上的动词、形容词和名词。

2. 主要属性和次要属性

董氏系统的主要属性的数量在 N, V , A 三大类中的分布以及次要属性的数量如下表所示。

主要属性	N 类	V 类	A 类
	152	810	Z
从属属性	属性值	通用属性	
	435	93	

表 3

董氏系统对于主要属性和次要属性的认识和处理也体现了董氏系统的哲学。

主要属性也可以认为是概念的类别,它们被组织成为体现上下位关系的层级结构中。它们在总体上体现了一类概念的某些本质的属性。例如,“人类”是“孩子”“教师”“富豪”等概念的主要属性。这个主要属性体现了这些概念的本质属性,其中有的是从它的上位“动物”继承来的,有的则是“人类”自己所特有的,如“会思考”“会进行认知活动”等。而上述三个概念在诸如“年龄”“职业”“贫富”等的差异,体现了它们个性,在词典中将分别以次要属性加以描述。

次要属性之间不存在上下位关系。董氏系统的次要属性包括 A 范畴规定的属性值如“男”“女”“老”“幼”“善”“恶”等等,词语应用的领域如“工”“商”“医”等等。可以这样说,次要属性较之主要属性有更高的抽象性。例如,“医院”“护士”“青霉素”等,它们分别属于不同的类,也就是有着不同的主要属性,但是次要属性“医治”却是它们共有的。董氏系统规定主要属性也可以用作次要属性,它们还常被冠以标识符,例如“医治”在V - 规范中是一个主要属性,即一个类别,但是在用它来标注“医生”“医院”“医药”等概念时,它便被用作了次要属性。

董氏系统关于主要属性和次要属性的认识和处理遵循了如下原则:

(1) 上位概念的属性可以由下位概念继承;下位概念至少应有一个属性是其上位概念所不具备的。

(2) 词典中每一个概念都必定有一个主要属性,它就是这一概念的类范畴;但一个概念可以有多个次要属性,也可以没有。在标注中,主要属性务必置于信息栏的第一个,如有次要属性,则中间用逗号隔开。

(3) 在确定类范畴及其上下位时,董氏系统严格地遵循分类标准一致性的原则。例如,我们在为“用品”进行分类时,如果我们以它们的“用途”作为标准来分类,这样就会有诸如“家具”“文具”“茶具”“化妆品”等;但如果我们以它们的“材料”作为标准来分类,这样就会有诸如“瓷器”“陶器”“玻璃器皿”等。我们不允许“用品”的下位既有“家具”“文具”等,也有“陶器”“玻璃器皿”等。再如,如果“人类”的下位既有“男子”“女子”等,又有“亲属”“富人”等;又如“行为动作”的下位既有“位置移动”“增减”等,又有“医疗卫生”“农业生产”等。这对于那种面向人的义类词典还算可以,但对于知识系统将是不可接受的,因为这是违反分类标准一致性原则的。

(4) 主要属性也可以用来作为次要属性,但当它被用作次要属性时,它可能保留它的全部或部分共性属性,但将失去在层级结构中的上下位关系的地位,即不可以再推导它的上位或下位。这也是董氏系统处理跨类概念的原则。我们的认识是:世界本来就是复杂的,多样化的。分类必然是可以多角度的。例如,“学校”这个概念,从它的领导部门如教育部看来它是一个“机构”,它的上位可能是“组织”,但从邮政局看来它是一个地址,它的上位可能是“空间”。

三、董氏系统知识描述的方法和特色

董氏系统知识描述的方法有以下要点:

第一,对于不同的类型如 N、V 和 A 各类范畴,以及它们中的各个由主要属性体现的类别,都有明确而严格的规定。这些类别组成体现上下位关系的层级结构。

第二,尽可能充分地描述概念之间的种种关系。其中多数借助于董氏系统规定的标识符,

这些关系包括:

范畴间关系	标识符	举例
N 与 N　部分—主体	%	手—% 人类,　　CPU—% 电脑
N 与 N　属性—宿主	&	颜色—& 无生命,　　智慧—& 动物
N 与 N　材料—产品	?	布—? 服装,　　面粉—? 食品
N 与 N　主体—相关体	#	警察—# 罪,　　罪犯—# 警
N 与 A　被描述—属性值		笨蛋—愚,　　小孩—幼
N 与 V　施事—事件	*	医生—* 医治,　　骗子—* 骗取
N 与 V　受事—事件	$	赃物—$ 偷取,　　患者—$ 医治
N 与 V　工具—事件	*	笔—* 书写,　　药品—* 医治
N 与 V　处所—事件	+	医院—+ 医治,　　剧场—+ 表演
V 与 N　事件—角色	(角色符) =	盗墓—source= 设施,　　赔款—possession= 钱币
V 与 A　事件—角色	(角色符) =	严惩—manner= 强,　　长大—state-fin= 成
V 与 V　事件—角色	(角色符) =	看病—result-event= cure 医治
A 与 V　属性值—范围		防潮—* withstand 抗住, # wetw 湿
A 与 N　属性值—属性		蓝—颜色,　　愚—智慧

(注:例子中“—”的左侧为概念,如“手”“颜色”等;其右侧是标注的内容,如“% 人类”“& 无生命”等。)

第三,一种对概念进行规范化的和形式化的标注语言已随着董氏系统的发展而趋成熟,我们把它称为概念描述标记语言(Concept Description Markup Language—CDML)。本文是对于知识系统的简要介绍,为节省篇幅,恕我们省略学术讨论和参考文献。

(董振东　北京 998 信箱 8 楼 3 门 401,邮编: 100091)

中国辞书出版迈入多媒体新纪元

汉语大词典出版社与商务印书馆(香港)有限公司,联手成功开发了《汉语大词典》光盘。

《汉语大词典》光盘 1.0 版集现代电脑科技之大成,将《汉语大词典》12 卷印刷本的主要内容浓缩于一张光盘上,20 多种先进的多方式检索功能,将汉语词语间千丝万缕的关系,条目分明地列举出来,全面提供有关字形、音、义、源、用法等知识,是汉语学习、汉语研究、语文教学、文字创作的最佳辅助工具。

(王　涛)