知网知识库描述语言

郝长伶　　董 强
中国科学院计算机语言信息工程研究中心 北京 100083
E-mail: support@keenage.com

摘　要：本文概述了知网知识库描述语言 KDML（Knowledge Database Mark-up Language）的发展，着重介绍了知网知识库描述语言 2002 版的语法规则，概念描述方式，及其在意义计算中的优越性，以便使知网的使用者对新版知网知识库描述语言有更加清晰的认识，拓宽他们的思路，为他们利用知网知识系统进行自然语言处理提供更好的帮助。

关键词： 知网知识库描述语言；知网；意义计算；语义；义原

Knowledge Database Mark-up Language of HowNet

Changling Hao　　Qiang Dong
Research Center of Computer Language Engineering, Chinese Academy of Sciences, Beijing 100083
E-mail: support@keenage.com

Abstract: This paper outlines the development of Knowledge Database Mark-up Language (KDML) of HowNet. It presents the grammatical rule and the way of description of the new KDML in detail and demonstrates its advantages in the computing of meanings. It shows the structure of the new KDML to the users of HowNet. We hope that it can offer help for the users to get a clearer understanding of HowNet and apply it in their work of NLP.

Keywords: KDML of HowNet; HowNet; computing of meanings; semantics; sememe

1、引言

知网知识库描述语言，英文名称为 Knowledge Database Mark-up Language （KDML） of HowNet，这是一套崭新的知识描述规范体系。知网认为对于概念的描述应该着力体现概念与概念、概念的属性与属性之间的相互关系，因此，知网知识库对于概念的描述必然是复杂的。同时，对于概念的描述既有概括性的、一般性的描述，也会有因不同的类别而引起的细节性、特殊性的描述。这样便必然引发概念描述的一致性和准确性的问题。为了确保概念描述的复杂度、一致性和准确性，我们设计了一种知识描述规范体系—知网知识库描述语言（KDML）。经过对中英文两种语言各 8 万多概念的描述，证明它（1）有很强的描述能力；（2）便于对意义的计算；（3）直观、有较好的可读性。

目前知网知识库描述语言已经在原有版本的基础上得到了新的发展。最新的版本被用于知网知识库 2002 版的概念描述中。(原有的版本用于知网知识库 2000 版的概念描述)。较之以前的版本，新版知网知识库描述语言有了重大的变化，主要体现在：

(1) 从原来的线形的描述方式改进为立体的、可嵌套的描述方式，从而使其描述能力更为强大和灵活。

(2) 把原来的实体类、事件类、属性类、属性值类义原之间的关系表示从隐性变为显性，使其对概念的描述更加清晰、更加准确。

2、知网知识库描述语言 2000 版

知网知识库描述语言 2000 版被用在知网知识库 2000 版的描述中，它在一定程度上实现了设计者的初衷，但是，其中还存在一些不足之处。知网知识库中的概念是通过义原与义原之间的关系来描述的，在知网知识库描述语言 2000 版中，义原与义原之间的关系是通过一些特定的符号来揭示的，这些符号包括：,~^# %$@&*+!?()[]{}，也就是说义原与义原之间的关系是通过这些符号隐性的表达的。另外，在不同的概念中同一个符号所代表的意义并不是完全相同的，一个符号可能会包含几种含义，这样不仅给概念标注带来了困难，而且给使用者在理解概念时也带来了一定的困难。

例如：* 这个符号在下面这几个概念中代表了不同的意义，在“医生”中代表 agent，在“比翼鸟”中代表 experiencer，在“挖掘机”中代表 instrument。

词语	KDML（2000 版）描述	* 的意义
医生	DEF=human|人,#occupation|职位,*cure|医治,medical|医	agent
比翼鸟	DEF=human|人,*love|爱恋	experiencer
挖掘机	DEF=machine|机器,*dig|挖掘	instrument

表 1：比较 * 的意义

知网知识库描述语言 2000 版的描述方式是一种线性的描述，它对义原的顺序是有规定的，如果破坏了这种顺序，就会导致意义上的错误。另外，由于描述能力的欠缺，容易造成意义的缺失。这些都为概念的标注和理解以及计算机的运算带来了一定的困难。

鉴于上述的不足之处，我们重新设计了知网知识库描述语言的描述规范。

3、知网知识库描述语言 2002 版

3.1 总规定

(1) 任一概念的描述都以 DEF= 为开始。任一概念中出现的所有义原或符号必须是在知网的 Taxonomy 中定义的义原或符号或者由知网知识库描述语言所规定的特定标识符。

(2) 概念描述中的第一个义原必须指出该概念的最基本的意义，并用事件、实体、属性和

属性值这四类义原中的一个标注出来。

（3）对于简单概念直接标注该概念的意义。

（4）利用动态角色与特征来标注复杂概念。

（5）属性类概念必须标明它的宿主。

（6）整体部分类型的概念必须标明该部分的整体。

（7）概念描述中定义的特性至少是一个，但也可以是多个，数量没有限制，只要内容是合理的且形式是合乎规范的就可以了。

3.2 KDML 中的特定标识符

在知网知识库描述语言中允许使用以下 7 种标识符，它们都是英文字符，具体见下表：

符号	名称	功能简述
{	左括号	表示对一个概念描述的开始。
}	右括号	表示对一个概念描述的结束。
:	冒号	冒号后面的内容是对冒号前面义原的具体描述。
,	逗号	表示一个关系描述的结束。
=	等号	表示一个动态角色或特征所具有的具体的值。
;	分号	分号表示某一概念是由若干个概念组合而成的组合型复杂概念。每个分号分割的部分必须是一个独立的完整的概念描述。
“	引号	引号中的内容都是一些具有特殊意义的义原。

表 2：KDML 中的特定标识符

3.3 几种特殊的指示符号

1）指示符号 ~

利用 ~ 进行描述的模式是:{义原 1:{义原 2:动态角色={~}}}　　一模式 1

这种描述方式表示的是，义原 1 与义原 2 有关，义原 1 为义原 2 的一个具体动态角色的值。其中的 ~ 用来代替前面的义原 1。通常情况下，义原 1 为实体类义原，义原 2 为事件类义原。例如“挖掘机”（实体类概念）的描述如下：

自然语言描述：挖掘用的一种机器。

KDML 描述：DEF={tool|用具:{dig|挖掘:instrument={~}}}

在这里义原 1 为 tool|用具，是一个实体类的概念。义原 2 为 dig|挖掘，是一个事件类的概念。为了说明该实体与该事件之间的关系，运用 ~ 来代替义原 1，从而说明义原 2 的工具（instrument）是义原 1。

2）指示符号 ?

利用 ? 进行描述的模式是：{义原 1:动态角色={?}}　　—模式 2

这种描述方式表示在某一语义环境中，? 所充当的动态角色的演员是一定会出现的，但是在

这个孤立的概念中它并没有被体现出来。例如“属于”（事件类概念）的描述如下：

自然语言描述：归某一方面或为某方所有。

KDML 描述：DEF={BelongTo|属于:possessor={?}}

从它的自然语言描述中可以发现，其中的“某一方面”在“属于”这个事件所出现的语义环境中是一定会出现的，即“属于”这个概念的 possessor 是一定会出现的。但是在“属于”这个孤立的概念中，它的 possessor 没有被体现出来，也就是说我们无法单纯的从“属于”这个概念中知道它的所有者是谁。于是，我们用 ? 来代替实际语义环境中一定会出现的 possessor 的内容，即动态角色 possessor 的演员。

3）指示符号 $

利用义原 $ 进行描述的模式是：{义原 1:动态角色={$}}　　—模式 3

其中 $ 用来充当某一个动态角色的演员。其中动态角色={$}表示这个概念所描述的对象是什么。例如“值得称赞”（属性值类概念）的描述如下：

自然语言描述：有被称赞的价值。

KDML 描述：DEF={able|能:scope={praise|夸奖:target={$}}}

其中$ 的意义是：$ 是 target 这个动态角色的演员，代表被夸奖（praise|夸奖）的对象。

让我们看这样一个实例：童先生的“艰苦奋斗、坚韧不拔”的工作态度是很值得称赞的。在这个实例中，值得称赞的是：“工作态度”，而不是“童先生”。“工作态度”在这句话中充当的是夸奖的对象（target）。

3.4 概念描述方法

概念分为简单概念和复杂概念。简单概念是指一个明确的事件，实体，属性或属性值，在概念中不包含任何的其它成分，直接标注它的意义。复杂概念是以事件为中心，除了事件本身以外还有一个或多个动态角色，需要利用动态角色与特征来标注复杂概念，在表示上述动态角色时它的书写格式是：动态角色名称={某一概念描述}。

3.5 举例

例如：“洗衣机”（实体类概念）

自然语言描述：自动洗涤衣物的电动机械装置。

KDML 描述： DEF={tool|用具:{wash|洗涤:instrument={~},patient={clothing|衣物}}}

洗衣机是一个复杂的概念，在这里 tool|用具、wash|洗涤、clothing|衣物等为义原，instrument、patient 等均为关系，~ 为指示符号，{ } : = 、为 KDML 中的特定标识符。它的定义以 DEF= 为开始，其中的第一个义原 tool|用具为实体类义原，表示洗衣机是一个实体类的概念，它相当于模式 1 中的义原 1。wash|洗涤是一个事件类的概念，它相当于模式 1 中的义原 2。为了说明该实体与该事件之间的关系，运用 ~ 来代替义原 1，从而说明义原 2 的工具（instrument）是义原 1。另外，通过 patient 这个关系指出 wash|洗涤这个事件的受事是 clothing|衣物。

这段描述所表达的意义是直观的，即“洗衣机”是一种用具，是洗涤的用具，洗涤的受事是

衣物。通过和它的自然语言描述相比较，不难看出，它已经进一步接近了自然语言。

4、应用 KDML 进行意义的计算

4.1 实现的理论

知网认为知识是一个系统，是一个包含着各种概念与概念之间的关系，以及概念的属性与属性之间的关系的系统。因此，知网对于概念的描述着力体现的是概念与概念、概念的属性与属性之间的相互关系。尽管知网知识库描述语言对概念的描述是孤立的、静止的，但是我们可以通过运算来激活这些概念之间的动态关系，而意义的计算正是面向计算机的语义资源的最关键的任务和衡量其质量的最主要的标准。

知网知识库描述语言是一种面向计算机的、可以进行计算的结构化的描述语言，可以计算是其设计目标之一，为了实现这个目标，我们进行了两个工作。一是将概念分解为义原，并形成一套完整的分类体系；二是将义原通过关系组织起来表达一个概念。将概念分解这样的理念并不是我们的一家之见，例如朗文就利用 2000 个基本词汇描述它所涵盖的全部词汇。但是实践证明，用词汇进行概念描述是存在缺陷的。首先，词汇存在歧义；其次，这些词汇需要用其它词汇解释，这样有时会造成循环定义。例如，朗文中对“get”的解释是“obtain”，而对“obtain”的解释是“get”。这样的做法对于人而言是可以的，但对于程序而言是无法实现的，而对于计算机而言是无法理解的。而在我们的方法中义原是最小的意义的单位，每一个义原都表示一个唯一的特定的意义，是没有歧义的。知网知识系统中共有义原 2200 个，关系 110 个，在进行概念描述时，这些义原会通过关系组织起来，而不是简单的义原的堆砌。因此，通过这些义原以及关系就可以得到概念与概念、概念的属性与属性之间的关系。

4.2 在意义计算中的优越性

较之知网知识库描述语言 2000 版，知网知识库描述语言 2002 版在概念描述中有两个最显著的变化，一是将所有的关系显性化，二是将概念的表示变为立体的，可嵌套的方式，这为意义的计算带来了极大的方便。

4.2.1 显性表示的意义

显性的关系表示，消除了概念表示中的歧义，降低了程序实现的难度。例如在前面举的例子“洗衣机”中，其中 tool|用具与事件 wash|洗涤之间的关系是明确的，即: tool|用具是 wash|洗涤的 instrument。同样 wash|洗涤与 clothing|衣物之间的关系也是明确的，即 clothing|衣物是 wash|洗涤的 patient。而在 2000 版中分别用符号 *# 表示 instrument 和 patient 这两种关系，而程序为了得到这些关系，必须对符号和义原进行判断。现在进行程序处理时，就不再需要为了判断这些符号的具体含义而进行其他运算了。

另外，义原之间的关系显性化后，不再要求书写的顺序，简化了标注的规范。例如：在知网知识库描述语言 2000 版中，对于部件整体类型的概念要求它的第一个义原必须是 part|部件，

紧接着必须是该部件的整体，并用% 号标明，整体后面指明该部件在整体中的部分，这种顺序是固定的，否则程序就无法识别了。而改进后的规范就没有这么多顺序的限制，虽然仍然要求第一个义原必须是 part|部件，但是不要求它后面的内容的顺序，对于那些是并列关系的内容它们的顺序是任意的。

例如："地基"

自然语言描述：作为建筑物基础的地层。

KDML 描述（2000 版）：DEF=part|部件,%building|建筑物,base|根

KDML 描述（2002 版）：DEF={part|部件:PartPosition={base|根},whole={building|建筑物}}

也可以写成：DEF={part|部件:whole={building|建筑物},PartPosition={base|根}}

4.2.2 概念嵌套的意义

在知网知识库描述语言 2002 版中通过特定标识符的运用，使得对概念的描述呈现了一种多层嵌套的立体化的格局，从而增强了这种描述语言对概念描述的能力。不同的层次的之间又通过不同的关系描述指出它们之间的关系，这极大地提高了意义计算的能力。同样的概念如果使用知网知识库描述语言 2000 版根本是无法表达的，即使能够通过一些义原和符号将这样的概念描述出来，在概念的程序解析上仍然存在难以处理的问题。

例如："出狱"（事件类概念）

自然语言描述：走出监狱或不再受监禁。

KDML 描述：DEF={undergo|经受:content={release|释放:source={InstitutePlace|场所:domain={police|警},
{detain|扣住:location={~},patient={human|人:modifier={guilty|有罪}}},
{punish|处罚:location={~},patient={human|人:modifier={guilty|有罪}}}}}}

"出狱"是一个很复杂的事件类概念。在这个概念的描述中嵌套了"监狱"这个概念的完整定义，即{InstitutePlace|场所:domain={police|警},{detain|扣住:location={~},patient={human|人:modifier={guilty|有罪}}},{punish|处罚:location={~},patient={human|人:modifier={guilty|有罪}}}}。而在"监狱"的概念中又嵌套了"罪犯"这个概念的完整定义，即{human|人:modifier={guilty|有罪}}。

从这个例子中，可以看到一些程序实现的端倪。尽管我们对每一个概念的描述中没有描述它与其它概念之间有什么样的关系，但是这种概念的嵌套使概念之间也因此而产生了特定的关系。当然，概念与概念之间的关系并不是通过简单的字符串匹配就可以得到的，而是需要将 DEF 解析并通过一定的法则而最终确定的。

4.3 可利用的资源

在知网知识系统中，可以用于计算的资源有：

a. 知网知识库：除了每一个记录的中文和或英文的词语表表记外，最主要的计算对象是每一个义项的定义，即 DEF。
b. 知网事件类义原表，包括其中每一个义原的框架和义原之间的上下位关系。
c. 知网实体类义原表，包括其中每一个义原的框架和义原之间的上下位关系。
d. 知网属性类义原表，包括义原之间的上下位关系。

e. 知网属性值类义原表，包括义原之间的上下位关系。
f. 知网事件角色与关系
g. 知网反义表
h. 知网对义表

4.4 意义计算的典型应用举例

这里我们举出意义计算的两个典型的应用，一是词语的相似度的计算，一是词语的相关性的计算。前者是刘群研发的[4]，后者是董强研发的，称为《相关概念场》。这两个软件包的一个共同点是它们都是可以任意测试的，而不是仅仅只有几个“漂亮”的例子而已。前者利用的是知网知识系统2000版，其中的描述语言是知网知识库描述语言2000版，而后者利用的是知网知识系统2002版。

在实际的系统中应用知网知识系统时也会根据不同的应用领域有所差别，例如有些系统只用知网知识系统的分类体系就够了，有的则是借鉴知网知识系统的理念，描述专业领域的概念，从而应用于专业领域的信息挖掘，文本聚类，文本分类等。总之，要充分合理利用知网知识系统中的各种资源，首先必须要充分理解知网知识系统，理解它的概念描述方式。

5、结束语

知网知识库描述语言的诞生是由于知网知识库中概念描述的需要，知网的发展过程也正是知网知识库描述语言的不断成长、成熟的过程。目前经过对中英文两种语言各8万多概念的描述，证明其描述形式具有很好的可读性，并且具有很强概念描述的能力和概念的关系计算能力。我们看到越来越多的人加入到我们中，不断挖掘知网知识系统的潜能。世界总是处在不断的变化之中，知网知识库描述语言也不例外，它必然会随着知网知识体系的发展而完善，成为表达能力更强、计算能力更强、语法更严密的一种知识描述语言。

参 考 文 献

[1] Dong, Zhendong, Knowledge description: what, how and who, Manuscripts & Program of International Symposium on Electronic Dictionary, Tokyo, 1988

[2] 董振东，语义关系的表达和知识系统的建造，《语言文字应用》第3期，1998

[3] 董振东，董强，知网和汉语研究 《当代语言学》第1期，2001

[4] 刘群，李素建，基于《知网》的词汇语义相似度的计算，第三届汉语词汇语义学研讨会

[5] 信息处理用现代汉语语义分析的理论与方法，张普，《中文信息学报》，1991 年第 3 期

[6] 现代汉语词典，修订本，社科院语言所词典编辑室，1996

[7] WordNet 1.6, 普林斯敦大学，1999

[8] LONGMAN English-Chinese Dictionary Of Contemporary English, Longman Group UK Limited, 1988

[9] 董振东，董强（1999），“知网”，http://www.keenage.com