

知 网

董振东 董强

e-mail: dzddong@public.bta.net.cn

<http://www.how-net.com>

摘要：本文较全面地介绍了知网，一个可用于自然语言处理的知识系统。知网现已在因特网上公开发布。它的知识词典现包含汉语词语5万和对应的概念6万多，以及与之对应的英语词语5.5万和概念7万多。本文涉及有关建立网状关系语义的一些重要问题。

关键词：知识获取和表达 知识库 知识词典 关系语义 词汇语义

HowNet -- An On-line Knowledge System

Zhendong Dong Qiang Dong

e-mail: dzddong@public.bta.net.cn

<http://www.how-net.com>

Abstract: The paper presents an all-round picture of how-net, a knowledge system for natural language processing. How-net has been recently released in the Internet. The knowledge dictionary of how-net contains 50,000 Chinese word forms, 62,000 Chinese concepts and 5,5000 English equivalents, 70,000 English concepts. The paper also covers some important issues on the building of relational semantic net.

Keywords: knowledge acquisition and representation knowledge base knowledge dictionary relational semantics lexical semantics

1. 引言

近十多年来，随着计算机本身以及信息高速公路的飞速发展，人们开始更加重视语义的研究以及大规模语义词典或大规模知识库的建设。例如普林斯顿大学的英语 WordNet，微软的 MindNet，在欧洲有基于 WordNet 的 EurowordNet，日本有电子辞书研究所（EDR）的日语和英语的概念词典，还有美国 HPKB(High Performance KB)等等。其中 WordNet 早已上网用于非营业性研究。

知网（HowNet）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。它为语言信息处理的研发提供了丰富的知识资源。它现已上网，网址是：<http://www.how-net.com>。

2. 知网概述

知网包括下列数据文件和程序：

- | | |
|-----------------|-----------------|
| (01) 中英双语知识词典 | (08) 概念的次要特征(3) |
| (02) 中文简体知识词典 | (09) 动态角色与属性 |
| (03) 中文繁体知识词典 | (10) 词类表 |
| (04) 概念的主要特征(1) | (11) 反义关系表 |
| (05) 概念的主要特征(2) | (12) 对义关系表 |
| (06) 概念的次要特征(1) | (13) 标识符号及其说明 |
| (07) 概念的次要特征(2) | (14) 知网管理程序 |

2.1 知识词典

知识词典是知网的基本文件或数据库。其中的中英双语知识词典则是最基础的数据库。它是中文简体知识词典和中文繁体知识词典的基础。现有的中英双语知识词典包含 11 万多个。例如：

NO. =005756	NO. =092273
W_C=病	W_C=医生
G_C=N	G_C=N
E_C=	E_C=
W_E=disease	W_E=doctor
G_E=N	G_E=N
E_E=	E_E=
DEF=disease 疾病	DEF=human 人,*cure 医治,medical 医

知网的现有规模如下表所示。

语种	词语总量	N 范畴	V 范畴	A 范畴
汉语	050220	026006	016635	09763
英语	055427	028818	016688	10705

语种	概念总量	N 范畴	V 范畴	A 范畴
汉语	062264	029808	020453	011196
英语	073131	036720	021187	014386

2.2 概念的主要特征(1)

概念的主要特征(1)载明知网所规定的事件类或称 V 范畴的主要特征，现有 800 多个，组织在一个层级网络中。例如：

V1.02 possession|领属关系

own|有 {relevant, possession}
obtain|得到 {relevant, possession, source}
receive|收受 {relevant, possession, source}
BelongTo|属于 {relevant, possessor}
OwnNot|无 {relevant, possession}
lose|失去 {relevant, possession}
InDebt|亏损 {relevant, possession}
owe|欠 {relevant, possession, target}

V2.02 AlterPossession|变领属 {agent, possession}

take|取 {agent, possession, source}
seek|谋取 {agent, possession, source}
beg|乞求 {agent, possession, source}
rob|抢 {agent, possession, source}[crime|罪]
buy|买
{agent, possession, source, cost, ~beneficiary}
give|给 {agent, possession, target}
provide|供 {agent, possession, target}
GiveAsGift|赠 {agent, possession, target}
grant|赐 {agent, possession, target}
return|还 {agent, possession, target}
recompense|补偿 {agent, possession, target}
sell|卖 {agent, possession, target, cost}

事件层级网络体现了如下的特征：

(a) 事件的上下位关系。这点在文件中已一目了然，不必赘述。

(b) 体现了知网的一个重要的、独创的观点：它认为事件有静和动两类。静态的又分为关系和状态两类。而动态的，即行为动作说到底是个“变”，而且是跟关系与状态那两类严格地一一对应的。例如，关系类包含有“领属关系”，而行为动作类则包含有“变领属关系”。状态类包含有“存现”，而行为动作类则包含有“变存现关系”。这是客观存在。

(c) 每一个主要特征都标有它的必要角色框架，如 MarryTo|嫁 {agent, possession, target}，作为它的共性，同时还可以另加其它种类的共性描述，被置于[]中。以 MarryTo|嫁为例，知网的规定是：当“嫁”这类事件发生时，“谁(agent)把谁(possession)嫁给谁(target)”等必要角色是一定会参与的。这是客观存在，不论在语言中是否全都说出来。“嫁”的另一个共性是：与“结婚”这一事件有关。

(d) 事件的关联和角色的转换。如：“买”这一事件将激活“有”；如：“买”的施事将转化为“有”的“关系主体”；“患病”的经验者将转化为“医治”的“受事”。

2.3 概念的主要特征(2)

概念的主要特征(2)载明知网所规定的事物类或称 N 范畴的主要特征, 现有 150 左右, 组织在一个层级网络中。例如:

N.1 entity|实体

N.1.1 thing|万物[#time|时间,#space|空间]

N.1.1.1 physical|物质[!appearance|外观]

N.1.1.1.1 animate|生物 [*alive|活着,!age|年龄,*die|死,*metabolize|代谢]

N.1.1.1.1.1 AnimalHuman|动物[!sex|性别,*AlterLocation|变空间位置,*StateMental|精神状态]

N.1.1.1.1.1.1 human|人[!name|姓名,!wisdom|智慧,!ability|能力,!occupation|职位,*act|行动]

事物层级网络体现了如下的特征:

- (a) 事物的上下位关系。事物的上下位关系深度比较浅。
- (b) 绝大多数特征都标有其共性的特征。上下位关系的共性有继承关系, 如, “人”除了具有其自身特有的共性特征, 如[!name|姓名,!wisdom|智慧,!ability|能力,!occupation|职位,*act|行动]外, 还将继承上位“动物”, “生物”、“物质”等的共性。

2.4 概念的次要特征

概念的次要特征现分别列于三个文件中, 下面是它们的部分例子。

概念的次要特征(1) 包含的是属性以及某些非语义特征, 如:

N.2 attribute|属性

N.2.1 appearance|外观

N.2.1.1 form|形状

N.2.1.2 brightness|明暗

N.2.1.3 clearness|清浑

N.2.1.4 prettiness|美丑

N.2.1.5 pattern|样式

概念的次要特征(2) 包含的是属性值, 如:

form|形状

Flat|扁

Straight|直

curved|弯

Level|平

Upright|正

slanted|歪

Linear|线

Surfacial|面

cubic|体

brightness|明暗

Bright|明

Dark|暗

prettiness|美丑

Beautiful|美

Ugly|丑

概念的次要特征 (3) 包含的是领域以及部件的具体部位, 如:

Agricultural 农	[#plant 植物,#planting 栽植]
Commercial 商	[#money 货币,#buy 买,#sell 卖]
Education 教育	[#teach 教,#study 学]
Medical 医	[#cure 医治,#disease 疾病]
Literature 文	[#compile 编辑,#translate 翻译]

Head 头	
Heart 心	
Body 身	
Limb 肢	[*crawl 爬,*swim 游]

2.5 动态角色与属性

动态角色是指概念在实际的语言中所构成的各种关系。知网现用到的动态角色约70个, 如施事、受事、经验者、时间、处所等等。动态属性现有主题和焦点等。

2.6 知网的同义、反义和对义关系

知网的同义、反义和对义的体现与一般的同义或反义词词典是不同的。一般的是现性的, 而知网的是隐性的。也就是说知网并没有把同义、反义和对义标识在每个条目上。知网的同义是依靠 (a) 概念的定义 (DEF), (b) 双语对译词条 (W_C 或 W_E)。例如:

NO. =015459	NO. =102118
W_C=打	W_C=置
G_C=V	G_C=V
E_C=~酱油	E_C=
W_E=buy	W_E=buy
G_E=V	G_E=V
E_E=	E_E=
DEF=buy 买	DEF=buy 买

这里“打”和“置”是同义, 因为它们的 DEF 相等, 且有相同的 W_E。

知网中的反义和对义是通过反义表和对义表体现的。这两张表标明了具有什么信息将形成反义或对义。

对义关系举例		反义关系举例	
Own 有	OwnNot 无	Blunt 钝	sharp 利
Obtain 得到	Lose 失去	bright 明	dark 暗
Be 是	BeNot 非	clear 清	blurred 浑
Come 来	go 去	Beautiful 美	ugly 丑

3. 知网的特色

知网的特色主要表现在如下方面：

第一，知网并不是一个在线的词汇数据库。知网是一个利用一种知识词典描述语言来描述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知识系统。

第二，知网所描述的不仅包含同类概念之间的关系，如上下位关系、同义关系、反义关系、对义关系、部件与整体关系、材料和成品关系、属性和宿主关系，还包含非同类概念之间的关系，如属性值和属性的指向关系、事件和角色关系。

第三，知网对语义研究的贡献可以归结为两点。一是把语义研究置于知识描述的基础上；二是语义描述呈网状。这个关系网的关键是：用对个别概念进行静态的、孤立的描述最终形成动态的、相关的知识网。

4. 结论

知网的研究与建设前后经历了十多年的时间，作者体会其最困难部分是：

(1) 确定主要特征和次要特征，以及对它们的组织；

(2) 确定描述的方法和建立概念的描述语言；

知网的研究与建设不仅有很高的探索性，而且有很强的工程性。知网的今后发展首先会在两个方面进行：

(1) 增加已有语种的概念总量

(2) 优化知识词典描述语言(KDML)，强化其描述能力

参考文献

- [1] CYC Ontology Guide: Introduction, 1997
- [2] 黄曾阳, HNC理论概要, 《中文信息学报》1997年第4期
- [3] Miller, G. A. Et. al., Introduction to WordNet: an on-line lexical database, 1990
- [4] 俞士汶等, 现代汉语语法信息词典详解, 清华大学出版社, 1998
- [5] Zhendong Dong, "Knowledge Description: What, How and who?", Proceedings of International Symposium on Electronic Dictionary, Tokyo, 1988
- [6] 现代汉语词典, 修订本, 社科院语言所词典编辑室, 1996
- [7] 汉英词典, 修订本, 北京外国语大学英语系《汉英词典》组, 1995