基于 HNC 的现代汉语词语知识库建设*

苗传江[1]， 刘智颖[2]
（1. 香港理工大学 中文及双语学系，香港；2. 北京师范大学 中文信息处理研究所，北京 100875）

[摘 要] 基于 HNC 的汉语词语知识库是 HNC 知识库系统的重要组成部分，它以句类知识为核心，从概念、语句和语境三个层面提供汉语理解处理所需的知识。经过 10 多年的建设，该库已达到 80,000 多词的规模，成为中文信息处理、汉语教学和汉语本体研究的宝贵资源。

[关键词] HNC 理论；词语知识库；现代汉语；中文信息处理；自然语言理解
[中图分类号] H0－05；08 [文献标识码] A [文章编号] 1000－5110（2010）04－0015－04

一、HNC 理论及其知识库系统简介

HNC（Hierarchical Network of Concepts，概念层次网络）理论[1]·[2]（黄曾阳 1998；2004；2010）定位于消解自然语言理解中的模糊，如一词多义模糊、词语组合模糊、指代模糊等，它建立的自然语言表述和处理模式，以概念联想脉络（又称语言概念空间或语义网络）为消解模糊的基本依托，设计了概念联想脉络的三个符号表述体系：一是概念表述体系，用有限的概念基元表述无限的概念；二是语句表述体系，用有限的句子类型表述无限的语句；三是语境表述体系，用有限的语境单元表述无限的语境。HNC 完整地构建和发现了这三个有限：有限的概念基元约有 20000 个，有限的句子类型有 57 种，有限的语境单元约有 15000 种。概念基元形成层次化、网络化的系统，从高层到低层分为 18 个概念范畴、101 个概念群和 456 个概念树，这是 HNC 建立概念联想脉络的核心和基础。

HNC 的知识库系统是服务于自然语言理解处理的知识表示体系和资源建设工程，由三个层面的知识库构成：一是概念层面的知识库，这是与语种无关的知识，是理解处理各种语言都共同需要的知识，其描述对象就是上述有限的概念基元、句子类型和语境单元。二是语言层面的知识库，这是与语种有关的知识，就是用 HNC 描述理解处理一种语言所需的知识，最主要的语言知识库是以词汇单位为描述对象的词语知识库。三是世界层面的知识库，包括常识及专业知识，因其数量庞大，类型复杂，一般是配合应用系统开发的需要而进行的。

基于 HNC 的汉语词语知识库（简称 HNC 汉语词语库）是把 HNC 理论应用于汉语理解处理而建立的语言层面的知识库。

二、HNC 汉语词语库的基本内容

HNC 汉语词语库中只包含双字词和多字词，不包含单字词，因为对汉语单字的知识描述要包含一些不同于词语库的知识，须单独建库。[3]

HNC 汉语词语库有 20 多个知识项，下面着重说明其中最重要、最有 HNC 特色的内容。

（一）概念类别知识

概念类别知识是对词语的语义和语法特征的综合提炼，为汉语语句的理解处理提供首要的激活知识，主要有两个方面：一是初步判断语句的总体结构所需的知识。例如，对可充当语义块标志符的词语（如“关于、通过、为了”等）、能充当全局中心动词修饰成分的词语等，都赋予特定的概念类别符号。二是优先做局部组合处理所需的知识。句子中的有些成分可以在确定句子的总体结构前优先进行组合识别，如时间短语、空间短语、数量短语等，对构成这类成分的词语也都赋予特

* [收稿日期] 2010－06－10
[作者简介] 苗传江（1973—），男，山东沂水人，博士，香港理工大学中文及双语学系副教授，研究方向为 HNC 理论在语言研究和信息处理中的应用及语言知识的表示与获取。

定的概念类别符号。

HNC 汉语词语库中使用的概念类别符号一共有 100 多个,是以 HNC 的概念体系为纲、配合汉语语句理解处理的需要而设计的。[4][p.102-108]

(二)HNC 符号知识

HNC 符号是用 HNC 的概念符号体系来表述词语的意义,例如:

增加 v341	教师 pa71	召开 vc39e219
减少 v342	讲课 va71	会议 gc39e219
年 wj10-	月 wj10-0	日 wj10-00
感觉 vr710	情感 g713	高兴 vu7131
战争 gva42	武器 pwa42	裁军 vc342&pea41
协议 rc249a	达成 vc249a $(v311|(jlv001/v810))	

这些符号串中蕴涵着概念联想的简明而丰富的知识,如"增加"与"减少"的反义关联、"召开"与"会议"的动宾关联、"教师"与"讲课"的主谓关联、"年、月、日"的包含关联、"情感"与"高兴"的上下位关联、"战争、武器、裁军"与军事领域的关联等等。

作为符号体系,自然语言有一个严重缺陷,就是意义上密切相关的内容,表达符号上却往往毫不相关,对人脑来说这不是问题,因为人脑有概念联想能力,而对计算机来说,这是个致命的障碍,因为它从语音或文字符号上无法获取意义上的关联,也就无法进行理解处理。HNC 符号把词义之间的概念关联显式地表达出来,为计算机提供了理解处理的基本依托。因此,HNC 符号并非线性的符号串,而是为表达概念关联而精心设计的复合结构体。[1][p.17-43],[4][p.28-49]

HNC 符号是为计算机对自然语言进行理解处理而设计的,它有两个重要特点:第一,HNC 符号是对词义的近似表达,其首要目的不是给出概念的精确表示,而是给出概念联想脉络知识的线索。第二,HNC 符号中不仅蕴含着词汇层面的知识,还蕴含着语句和语境层面的知识,如句类知识、领域知识等。

(三)句类知识

HNC 的句类是句子的语义类型,与句子的语法结构无关。HNC 发现,自然语言的语句有 57 种基本句类,并写出了它们的表示式。这 57 种基本句类是句子语义的基元类型,用它们的表示式及其组合,可以描述任何语言的任何语句的语义结构。句类表示式由语义块构成,语义块是句子语义的下一级构成单位。不同的句类有不同的特点,称为句类知识。语义块、句类、句类表示式和句类知识是 HNC 建立的语句表述模式的基本概念。[1][p.44-59],[4][p.50-88]

HNC 词语库中的句类知识,就是用 HNC 的语句表述模式描述一个词能形成什么样的句子,包括句子中语义块的数量和排列顺序,各个语义块的内涵和构成,以及它们的核心部分优先由什么样的概念充当等等。例如,"打断"有两个意思,一个是"打断腿"的"打断",另一个是"打断思路"的打断,第一个意思的句类知识是:它形成的句子是一般作用句,有三个语义块,如下面的例句所示,分别是作用者"张三",作用"打断了"和作用对象"李四的腿",作用者的优先概念是人,作用对象的优先概念是具体物,它们有三种排列顺序,如例句 a、b 和 c 所示,在例句 c 那样的顺序中,作用对象可能发生分离,如例句 d 所示,分离出去的部分移到最后,如例句中的"腿"。第二个意思的句类知识大部分与第一个意思的相同,不同之处在于,它的作用对象必须是抽象概念,如"思路、进程、谈话"等,作用者可以是人,也可以是"马达声、行动、提问"等。靠这些知识就可以对"打断"形成的句子进行有效的理解分析,并判断具体句子中的"打断"是哪个意思,也就是消解一词多义模糊。

a. [张三][打断了][李四的腿]。 b. [张三][把李四的腿][打断了]。

c. [李四的腿][被张三][打断了]。 d. [李四][被张三][打断了][腿]。

再如,"起诉"的句类知识是:它形成的句子是信息转移句和关系句的混合,有四个语义块,如例句 e 所示,分别是信息转移发出者及关系第一方"张三"、信息转移及关系"起诉"、信息接收者"法院"和关系第二方"李四",它们必须以例句中的顺序排列,关系双方必须是人或组织机构,信息接收者必须是法律机构。

e.[张三][向法院][起诉][李四]。

上面提到的一般作用句、信息转移句和关系句等,都是 HNC 的 57 中基本句类。

上例中,以"打断"和"起诉"为核心的语义块称为特征语义块,特征语义块的核心一般是动词,但也可以是名词,如"张三对李四的计划没兴趣",

这个句子的特征语义块是“没兴趣”, 其核心“兴趣”是名词。HNC 词语库为所有可能充当特征语义块核心的词语配备句类知识。

(四)其他词汇知识

除了以上三大项以外, HNC 汉语词语库提供的词汇知识主要还有音调、义项总数、义项的使用频度等级、重叠形式、能否分离、相邻搭配等。

(五)例句

为了具体说明词语的意义和用法, HNC 汉语词语库还提供了丰富的例句, 这些例句绝大多数来源于大规模真实语料库, 但不是随便抽取的, 而是为了有针对性地说明词语的各种意义和用法特点而精心挑选的。

三、HNC 汉语词语库的特点

以 HNC 理论的自然语言表述和处理模式为指导, 是 HNC 汉语词语库的基本特点, 主要体现在以下 6 个方面:

第一, 以概念和语义为中心, 而不是以语法为中心。

第二, 与汉语理解的技术实现紧密结合。HNC 汉语词语库和 HNC 汉语理解技术都以 HNC 理论为统一指导, 词语库的内容和表示方式都密切配合理解处理的需要, 而且能及时得到应用、检验和改进。

第三, 以消解模糊, 实现汉语理解处理为目标。HNC 汉语词语库的主要知识项都服务于这一目标的实现, 而不是把各种词汇知识都汇集起来。

第四, 从概念、语句和语境三个层面提供知识。这是由 HNC 汉语词语库的上述服务目标决定的, 实现这个目标所需的知识既有概念层面的, 也有语句和语境层面的, 这些知识都需要通过词语知识来激活。

第五, 以句类知识为核心。这也是 HNC 汉语词语库的上述服务目标决定的, 词语库中的各项知识都以句类知识为纲领。

第六, 高度符号化和数字化, 不是用自然语言描述自然语言。HNC 汉语词语库的主要知识项都是用 HNC 的符号体系表述的, 这些符号体系是为了让计算机获得理解自然语言所需的关键知识而精心设计的, 是完全符号化和数字化的。例如, HNC 符号表达的知识, 是不能用自然语言的表述方式来取代的。

四、HNC 汉语词语库的应用

HNC 的语言理解技术称为句类分析, [1][p.56-58] 要针对一种语言实现句类分析技术, 就必须基于 HNC 建立这种语言的词语知识库, 因此, HNC 汉语词语库是以句类分析实现汉语理解技术[5] 所依赖的基础资源。HNC 的汉语理解技术已经在智能信息检索和机器翻译等领域得到应用, HNC 汉语词语库自然也是这些应用系统不可或缺的基础资源。

HNC 汉语词语库是基于 HNC 理论建立的, 这并不意味着它只能用于基于 HNC 的语言处理。HNC 汉语词语库以语义为中心提供了丰富的知识, 可以在中文信息处理的很多方面得到应用, 如专名识别、命名实体识别、信息抽取、智能查询等。

HNC 汉语词语库也是汉语本体研究的宝贵资源, 特别是在词汇语义研究方面。从这个库中可以很方便地检索出汉语中跟某个概念相关的大部分词语, 并进一步找出跟这些词语的意义有各种关系的词语, 这样就可以方便地研究词义的关系和系统。例如, 从库中检索跟“快”和“慢”这两个概念相关的词语, 可以分别得到 204 个和 67 个。

HNC 汉语词语库也可以在汉语教学和词典编纂中得到应用, 因为这个库中详细描写了词语的各种意义和用法, 而且配备了典型、真实、丰富的例句。以这个库为基础, 我们已经尝试编写了包含 1500 多个常用动词的学习型现代汉语动词词典。

信息时代需要人机两用的语言研究, [6] HNC 汉语词语库是一种人机两用的语言资源。

五、HNC 汉语词语库的发展

HNC 汉语词语库的建设始于 1997 年初, 经过十几年的积累, 至 2009 年底, 库中共有词语 80,793 个, 义项 89,901 个。为建设这个库而投入的总工作量估计已有 200 个人年。

在 HNC 汉语词语库的建设中, 最困难、最重要的工作不是词语数量的扩充, 而是质量的保证, 检查和修订占用了总工作量的一半以上。尽管如此, 现在的库还只能算是达到了 1.0 版的水平, 还需要不断改进。

除了扩充词语以外,HNC 汉语词语库今后的发展要着重于以下五个方面的工作:一是进一步细化填写规范;二是加强建库平台和辅助工具的研发,以提高建库的效率和质量;三是逐步提高开放性,让更多的人了解和使用这个资源;四是推动这个资源在语言信息处理、语言教学和语言本体研究等多方面的应用;五是根据 HNC 理论的进展进行更新和升级。

六、结束语

HNC 汉语词语库是实现汉语理解处理不可或缺的基础资源,它的建设不可能一劳永逸,必须不断探索和改进。HNC 汉语词语库是人机两用的语言资源,要建好这个资源,需要大量跨接语言学和计算机科学的复合型专业人才。另一方面,这个库的建设也是培养这类复合型专业人才的教练场。

[参 考 文 献]

[1] 黄曾阳. HNC(概念层次网络)理论——计算机理解语言研究的新思路[M]. 北京:清华大学出版社,1998.

[2] 黄曾阳. 语言概念空间的基本定理和数学物理表示式[M]. 北京:海洋出版社,2004.

[3] 苗传江,唐兴全,刘智颖. HNC 的字知识库[A]. 第二届 HNC 与语言学研讨会论文集[C]. 北京:海洋出版社,2004.

[4] 苗传江. HNC(概念层次网络)理论导论[M]. 北京:清华大学出版社,2005.

[5] 晋耀红. HNC(概念层次网络)语言理解技术及其应用[M]. 北京:科学出版社,2006.

[6] 林杏光. 语言研究要注意人机两用[J]. 语文建设,1993.

The Lexical Knowledge Base of Modern Chinese Based on HNC Theory

MIAO Chuan-jiang[1], LIU Zhi-ying[2]

(1. *Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China*;
2. *Institute of Chinese Information Processing, Beijing Normal University, Beijing* 100875, *China*)

Abstract: The lexical knowledge base of modern Chinese based on HNC Theory is an important part of HNC knowledge system for natural language understanding. The Base presents conceptual, sentence-related and situational knowledge of words while centering on sentence category knowledge. With much effort of thirteen years, this base has become an important one housing more than 80,000 Chinese words, which is useful for Chinese information processing, Chinese learning and teaching as well as linguistic studies.

Key words: HNC Theory; lexical knowledge base; modern Chinese; Chinese information processing; natural language understanding

[责任编辑: 杨育彬]