基于 HNC 的汉语词语知识库改进

王青海[1], 马海慧[1], 池毓焕[2], 李 颖[1], 董凌冲[3]
[1]装甲兵工程学院 信息工程系, 北京 100072; [2]中国科学院 声学研究所, 北京 100190; [3]63713 部队, 山西 036301
E-mail: wangqh@139.com; mhhzgy@163.com; chiyuhuan@hotmail.com; lypublic@hotmail.com

摘 要: 汉语词语知识库是HNC知识库系统的重要组成部分，目前其结构设计简单，加大了对HNC符号解析的难度。本文在分析了HNC的编码特点的基础上，改进了汉语词语知识库模型，阐述了改进后汉语词语知识库实体属性的设计方法和知识库的填写原则，并用实例说明了改进后的词语知识库可以提高自然语言处理的效率。

关键词: 汉语词语知识库; HNC 理论; 关系数据库

The Improvement for the Chinese Lexical knowledge-database Based on HNC Theory

Wang Qinghai[1], Ma Haihui[1], Chi Yuhuan[2], Li Ying[1], Dong Lingchong[3]
[1]Department of Information Engineering, Academy of Armored Force Engineering, Beijing 100072
[2]Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190
[3]63713 Troops of PLA, Shanxi 036301
E-mail: wangqh@139.com; mhhzgy@163.com; chiyuhuan@hotmail.com; lypublic@hotmail.com

Abstract: The Chinese lexical knowledge-database based on HNC is an important component of the HNC knowledge-database system, the structure of which is so easy as to increase the complexity of the HNC symbol-resolution. In this paper, a new Chinese lexical knowledge-database model based on the HNC theory is proposed to improve the efficiency in the natural language processing, which the entity attribute design method of and the principle to filling in are expounded, and two examples are provided to introduce its application.

Keywords: Chinese lexical knowledge-database; HNC theory; relational database

1 引言

任何一个自然语言处理系统理解自然语言句子，首先要具备词汇知识。词汇语义知识库已经被广泛应用于机器翻译、信息检索、问答系统、自动文摘等领域，成为自然语言处理不可或缺的基础资源。比较著名的知识库有 WordNet、FrameNet、EDR 电子词典、知网、HNC 等[1]。

目前，HNC 知识库在数量上有一定的发展，但是在可扩展性和数据库设计上鲜有问津。现有的 HNC 汉语词语知识库相对设计简单，关联性不足，增大了 HNC 符号解析的复杂程度，主要存在以下缺点：

（1）系统做语义距离计算时，要进行单独的 HNC 符号解析，检索周期长；

（2）概念联想上完全依靠程序计算，概念之间的映射关系在数据库中没有直接的反映。

本文通过分析 HNC 的编码特点，改进汉语词语知识库结构，提出了新的汉语词语知识库模型，并且通过实例说明改进后词语知识库在语义距离计算上的优势，指出了汉语词语知识库的发展趋势和可能存在的问题。

2 知识库模型的构建

2.1 汉语词语知识库模型设计

1）汉语词语知识库模型的提出

HNC 理论利用 HNC1、HNC2、HNC3 和 HNC4 将词语、语句、句群、篇章数字化，为计算

机把握语义提供基础[2]。HNC 符号有两个重要特点：

（1）HNC 符号是对词义的渐进表达，给出概念联想脉络知识的线索，与语种无关；

（2）HNC 符号中不仅蕴含着词语层面的知识，还蕴含着语句和语境层面的知识。

根据上述特点可知，利用 HNC 处理自然语言，建立自然语言与概念空间之间的映射是关键。

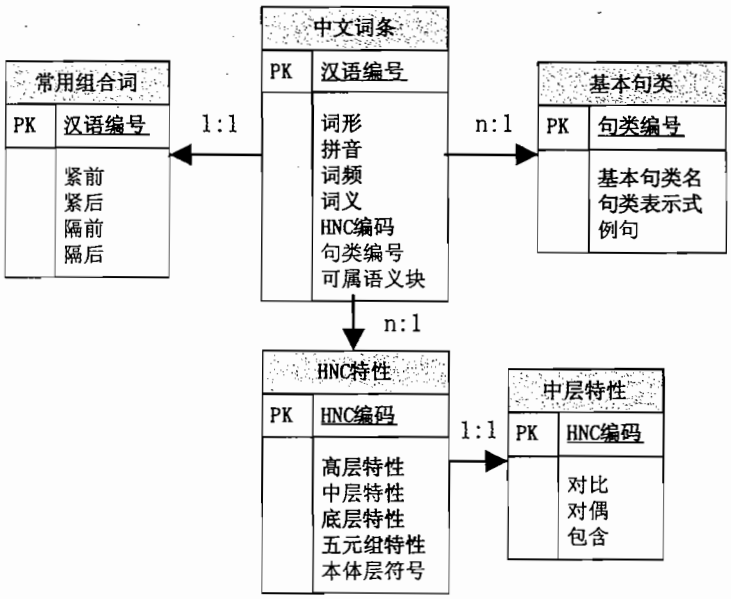

图 1　基于 HNC 的汉语词语知识库模型

知识库承担了自然语言与概念空间之间映射的任务。HNC 符号中表达概念联想关系的手段主要有概念矩阵、层次性节点、挂靠表达、组合结构等[3]。本知识库设计基于 HNC1 和 HNC2，利用自上而下的建库思想[4]，如图 1，包括中文词条、HNC 特性、基本句类、常用组合词和中层特性 5 个实体。其中，中文词条通过 HNC 特性外化它们在概念空间的关联。此外，与本知识库设计紧密相关的三层节点和五元组知识会在下面进行简要的说明。

2）实体设计的必要性分析

（1）如图 1，将 HNC 特性作为一个实体，并分 HNC 编码、高层特性、中层特性、底层特性、五元组特性和本体层符号 6 个属性。其中，后 5 个属性是 HNC 编码的符号解析。这是比以前单表词库改进的地方，省去编程时的符号解析环节，有利于概念联想，降低文字处理系统程序的复杂度，减少不必要的误差。

词语的 HNC 编码包括高层、中层和底层，在语义距离计算时都有不可忽视的作用，具体方法参见 3.1。概念的五元组特性和概念的层次性、对比性、对偶性、包含性统称概念同行关联，简称同行[2] 。具有同行关联性的概念有相同或相似的层次符号，因而部分语义距离的计算问题就简化为对数字串的逐层比较问题。五元组是概念的外在表现，分别描述概念的 5 个侧面：动态 v、静态 g、属性 u、值 z、效应 r。它们可以多重组合但各有约定的内涵。对比性、对偶性和包含性是概念局部联想的基本特征。从一个对比性概念就能联想到另外 n-1 个概念，从对偶性概念的一方可以联想到另一方，从一个包含性概念就可以联想到它的上下方，这为电脑进行概念联想操作提供了有效的手段[3]。所以，这里将中层特性单独作为一个实体。

本体层符号与挂靠概念相关，挂靠就是把一个概念与相关概念的层次符号直接拼接在一起，是 HNC 符号中表达概念关联性的一种方式[3]。在语义距离计算的过程中，本体层是首先要进行判断的。具体方法参照 3.1 的实例。

（2）通过 HNC 编码，设计常用组合词词表，建立最常用的词语之间的联系。在语言理解的过程中，如果有一个词语确定了 HNC 编码，在知识库中找到与它组合的常用词，如果被找到的常用组合词与上下文相符，则可以确定这两个词是一个短语，进而确定常用组合词的 HNC 编码；另一种情况，在合词阶段，遇到常用组合的形式，可以同时确定这几个词属同一语义块，并确定这几个词的 HNC 编码。常用组合词可以提高合词和语义块识别的效率，具体办法参见 3.2。

（3）语义块是句类的函数，要理解语句，确定句类，必须要从语义块上分解句子，以语义块为单位进行翻译，然后进行语义块顺序的调整。语义块必须标注，可以提高语义块识别的效率。

2.2 词语知识库实体属性设计

1）中文词条

（1）中文编号，是中文词条的主码，从数字 1 开始的编号。多义词有多个义项，中文编号对应唯一的词语义项。例如 词语“中央”有两个义项，一个是“中心的地方”，另一个是“国家或者政党政治权利最高的地方”，这两个义项要有不同的中文编号。

（2）词形、拼音根据现代汉语词典填写，词形填写类型是短文本，拼音填写类型是字符串。

（3）词频，根据已有的 HNC 汉语语料库进行统计，填写类型是数字。

（4）HNC 编码，根据 HNC 符号规则将词语的当前义项完全数字化，填写类型是字符串。

（5）可属语义块，包括 4 种主语义块和 7 种辅语义块，填写类型是字符串。其中，主语义块有特征（E）、作用者（A）、对象（B）和内容（C）；辅语义块有方式（Ms）、工具（In）、途径（Wy）、参照（Re）、条件（Cn）、因（Pr）、果（Rt）。

（6）句类编号，是中文词条的外码，基本句类实体的主码，对应唯一的实体基本句类，填写类型是从 1 开始的数字。

2）常用组合词

常用组合词的主码也是中文编号。常用组合词分 4 项属性，紧前、紧后、隔前、隔后，每项属性填写的都是 HNC 编码。

3）基本句类

基本句类的主码是句类编号，范围是 1-57，根据 HNC 有限的 57 种句类填写基本句类名和句类表达式。基本句类名的填写类型是短文本，句类表达式的填写类型是字符串。例句由 HNC 的语料库提供，是完成句类成分分析的例句，例句格式如下：

这||是||<{总结|\{近代以来~|中国|发展}的历程/}|得出|的结论>。

4）HNC 特性

HNC 特性的主码是 HNC 编码。同时，HNC 编码也是它与中层特性联系的外码。不同于以往的知识库，本知识库不仅根据 HNC 概念联想脉络对词条进行 HNC 编码，还将词条的高层、中层、底层以及五元组特性分开描述，不用对 HNC 符号进行分析解读就能从高层特性直接判定这个词条所属的概念属性，并且，中层特性作为单独的实体被描述。

5）中层特性[3]

中层特性的主码是 HNC 编码，表达概念的对偶、对比和包含特性，填写类型是字符串。

（1）包含性概念的表示式中，“-”是最高一级的包含概念，“-0”表示比“-”还低一级的包含概念，“0”越多包含概念的级别越低，这对于语义块的识别有很重要的参考意义。如，对于时间短语“1949 年 10 月 1 日”，其中“年”、“月”、“日”的 HNC 编码可写成：年 wj10- ，月 wj10-0，日 wj10-00。

根据符号可判定，这 3 个词具有包含的意义，属同一语义块。还有另外一种包含信息，比如“中国”、“上海”这两个词，从上海可以判定的地理信息就有中国这一层。

（2）对比性概念用符号 cnk 或者 dnk（k 取值 1~n）。n 表示对比的总级数；k 表示排序中的序号；c 表示正序，即序号 k 越大值越大；d 表示反序，及 k 越大值越小。如：

幼 u10bc51　　少 u10bc52　　青 u10bc53　　中 u10bc54　　老 u10bc5

冠军（j00d01，l15，gvc730）　　亚军（j00d02，l15，gvc730）

对比符号在排序、信息查询的过程中有着不可忽视的作用。

（3）对偶性概念分为二重对偶和三重对偶两种，用 ekm 或者 m 表示，m 取值 0~7，分为 0，1，2，3 和 4，5，6，7 两组。1、5 和 2、6 表示对偶的双方，0、4 表示统一方，3、7 用于表示对偶中的的第三方，不同的对偶类型可能没有统一方或第三方[7]。

2.3　词语知识库填写原则

知识库的基本设计思想是概念矩阵的近似实现。HNC 符号对词语之间概念联想的关系脉络给出形式化的表达，以服务于自然语言处理的需要 [3]。本着便于计算、便于语义块识别，降低非专业用户使用难度的原则，汉语词语知识库的建库原则[6]如下：

（1）以消解模糊、语义块识别为目的选词。汉语没有很严格的、如印欧语言那样的“词”。收集词语以概念和语义为中心，考虑词语的流行性和固定性。

（2）词语义项的选择需要考虑现代流通性。

（3）符号编码以句类知识为核心，词语库中的各项知识都以句类知识为纲领。

（4）知识库的主要知识项都用 HNC 的符号体系表述，是完全符号化和数字化的。

3　词语知识库的应用分析

汉语词语知识库的应用有很多。依据知识库中同行优先、常用组合、概念类别[3]等与 HNC 相关的先验知识和已经确定的 HNC 符号编码，既可以推测上下文关联词的编码，也可以开展填空造句，或者翻译的时候也能根据 HNC 编码进行语义距离计算从而确定句子最准的目标词语。下面利用实例简要分析一下本文提出的汉语词语知识库在语义距离计算方面的优越性。

3.1　语义距离计算

以前的单表知识库，提取高层、中层、底层信息需要专门的程序，改进后却可以直接读取。

表 1　中文词条建库填写举例

词语			暂停	形势	冠军	亚军
HNC 编码			v5211eb2	v521009	j00d01,l15,gvc730	j00d02,l15,gvc730
挂靠层符号	高层特性		11	10	j00 ,l15,c730	j00,l15,c730
	中层特性	对偶	eb2			
		对比			d01	d02
		包含				
	底层特性			09		
	五元组特性		v	v	gv	gv
本体层符号			52	52		

如表 1 所示，以“暂停”、“形势”、“冠军”、“亚军”为例进行语义距离计算，计算公式[5]如下：

$$SDC(H1,H2)=MAX(Sim(S11,S21), Sim(S11,S22),\cdots,Sim(S11,S2m),$$
$$Sim(S12,S21), Sim(S12,S22),\cdots,Sim(S12,S2m),$$
$$\cdots\cdots$$
$$Sim(S1n,S21), Sim(S1n,S22),\cdots,Sim(S1n,S2m)) \quad (1)$$

Sim(S11,S21)= Sim(S11.网络符号,S21.网络符号)*(Sim(S11.五元组,S21.五元组)+(Sim(S11.本体层符号,S21.本体层符号)* Sim(S11.中层符号,S21.中层符号)* Sim(S11.高层符号,S21.高层符号)+ Sim(S11.高层符号,S21.高层符号)*Sim(S11.底层符号,S21.底层符号))))) (2)

要进行高层、底层、五元组、本体层的比较，这些比较都没有考虑组合符号的作用，也就是说单纯的是字符的比较，计算结果用数字 0~7 表示[5]. 例如：

（1）“暂停”与“形势”的 HNC 符号直接可以抽取五元组信息比较的得出匹配成功，本体层完全匹配，高层部分匹配成功，Sim(S11,S21)的结果为 4。

（2）“冠军”与“亚军”的 HNC 符号进行分层比较。用本文提出的知识库进行的只是符号读取与顺序匹配，不用对全部 HNC 符号进行解析。中层符号有 d01，d02，而其他层的比较都匹配，Sim(S11,S21)的结果为 7，最终得到的结果是对比性概念。

影响语义距离计算的因素有很多。例如，语义距离计算的两个词语的高层概念不同说明概念的基本类型不同，语义距离会很大，但底层有交叉的可缩短它们的距离，这不是本文讨论的重点。

本文改进的汉语词语知识库模型中，一个中文词条的 HNC 编码是确定的，高层、中层、底层也是可以分别读出的。从程序实现的角度来讲，提取表格中的概念比将每个 HNC 编码解析计算量要小的多，这样就减小了自然语言处理核心程序的计算量。

3.2 合词

利用常用组合词的合词过程如图 2 所示。常用组合词主要用来发现和分析前后词的关联，便于语义块的识别。直接给出常用组合词，可以简化语义块识别的组合判定算法，提高程序的效率。如“热爱学习”，“热爱”是“学习”的紧前词，“学习”是“热爱”的紧后词；“处于紧张时期”，“处于”是“时期”的隔前词，“时期”是“处于”的隔后词。当“学习”的 HNC 编码确定时，我们搜索到常用组合词“热爱”在紧挨“学习”前面的位置，可以判定它们可以组成一个短语，增加了同属一个语义块概率，在做语义块判定时他们之间的关系优先判定。

例如，对于句子“这是总结近代以来中国发展的历程得出的结论。”利用本文改进的知识库，通过合词，可将“近代以来”划为一个短语。如果 “近代”、“以来”已经被填入知识库，常用组合词匹配成功直接形成短语。如果没有进行常用词组关联，可直接读取这两个词 HNC 符号的各层编码，利用 3.1 的语义距离计算进行判定。这两种途径都比之前用单表进行语义块判定的计算量小很多。

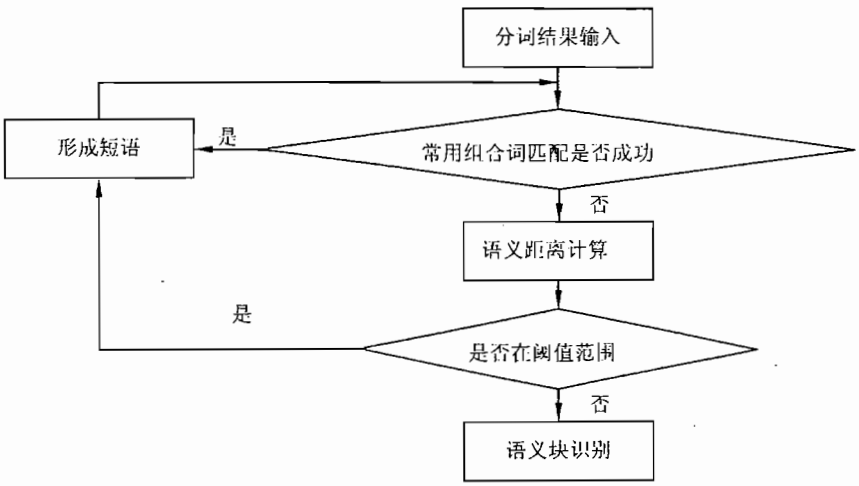

图2 合词流程

4 结语

本文设计的 HNC 汉语词语知识库虽然增加了建库的复杂度，但增强了知识库的程序可读性，提高了 HNC 知识库的层次性、逻辑性。尤其将 HNC 符号分高、中、低三层写入知识库的办法，简化了 HNC 符号的读取，减小因读取 HNC 符号造成的处理误差，同时，便于简化接口程序和搜索算法，方便知识库的管理。

填写 HNC 知识库的过程，是从词语的文字符号向 HNC 的概念表述符号映射的过程，要求填写者理解和掌握 HNC 的概念符号体系以及 HNC 的自然语言理解处理策略。所以，要建好这个资源，还需要大量跨接语言学和计算机科学的复合型专业人才。

参 考 文 献

[1] 宗成庆. 统计自然语言处理 [M]. 北京: 清华大学出版社, 2008: 48-71.

[2] 李颖, 王侃, 池毓焕. 面向汉英机器翻译的语义块构成变换[M]. 北京: 科学出版社, 2009.

[3] 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.

[4] 赫南达斯. 数据库设计凡人入门——关系数据库设计指南（第二版）[M]. 范明, 译. 北京: 电子工业出版社, 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: 50-61.

[6] 苗传江, 刘智颖. 基于 HNC 的现代汉语词语知识库建设[J]. 云南师范大学学报, 2010, 42(4): 15-18.

[7] 李颖, 池毓焕. 对偶性概念的 HNC 阐释[J]. 中文信息学报, 2004.18(3): 39-46.