一种基于 HNC 理论的领域知识表示研究

缪建明[1,2] 吴晨[1,2] 郝惠宁[2] 张全[2]

（1 中国科学院研究生院 北京 100039；2 中国科学院声学研究所 北京 100080）

摘 要：领域句类是 HNC 理论根据不同语境类别——领域所蕴含的世界知识抽象得到的语句级概念联想脉络，它具有自己的句类代码和表示式，通过特定的领域词语可激活对相关领域的联想。领域句类的设计是语境单元萃取技术中不可或缺的环节，为语境单元框架的构建提供基本要素。本文在 HNC 交互引擎的整体思路指导下，详细阐述如何为不同领域的概念设计领域句类。最后，通过实例句群，说明领域句类的知识有助于自然语言理解的处理。

关键词：HNC 理论 领域 领域句类 语境单元萃取 交互引擎

Domain knowledge expression based on HNC theory

Miao Jian-ming[1,2] Wu Chen[1,2] Hao Hui-ning[2] Zhang Quan[2]

(1Graduate School of the Chinese Academy of Science, Beijing, 100039; 2Institute of Acoustics, Chinese Academy of Science, Beijing, 100080, China)

Abstract: The Domain Sentence Category (DSC in short) is the Association Mesh of Concepts (AMC in short) in sentence level according to the world knowledge which contains in the different context type as we call it Domain. Each DSC has its own code and repression which are used to express the common sense knowledge for a Domain, and the distinct words of a Domain can activate the association of the relevant domain field, moreover the Domain knowledge is aroused. So, the design of DSC is a very important step. And more, the DSC provides the basic key element for construction of the SGU frame. Under the guidance of the HNC interactive engine, this paper illuminates how to design the DSC for the different domain, and explains the usability of the DSC for nature language understanding by an example.

Key words: HNC theory; Domain; Domain Sentence Category(DSC); Sentence Group Unit Extract(SGUE); the Natural Language Interactive Engine

1. 引言

为使计算机具有理解和处理自然语言的能力，必须使计算机拥有不同的知识。HNC 理论认为，知识可分为三类：概念知识、语言知识、常识及专业知识。从重要性上看，概念知识和语言知识居主要地位，常识性知识居于次要地位[1]。目前，HNC 理论针对概念知识及语言知识的特点建成了对应的知识库，但对于常识性知识库，仍没能建成，究其根源来自于以往未曾找到合适的分类体系对其进行完整的知识划分。常识知识是指能够通过词语

本文承国家 973 项目“自然语言理解的交互引擎研究”（2004CB318104）、中科院声学所知识创新工程项目“HNC 语言知识处理理论及技术”的资助。

作者简介：缪建明（1977～），男，博士生，研究方向：自然语言理解；吴晨，博士生；郝惠宁，工程师；张全，研究员，博导。E-mail: mjm_77@sina.com

概念激活的相关联想脉络的背景知识，具体到文章中可对应为特定的上下文知识和背景知识。经过十多年的深入研究，HNC 理论得出“常识性知识具有一定领域类型，可先验赋予计算机使用”的结论，并且在此基础上展开对领域句类知识的研究。

HNC 理论认为：没有这些常识性知识的辅助，计算机是无法对语句一级处理所遗留的部分省略、指代给予根本性处理的，这就要求我们必须展开对常识知识形式化表述的研究。目前，知识表示的研究主要集中在网络环境下知识表示的研究上，集中在对知识表示语言的研究上，涌现了 XML、UML、XMI 等知识表示语言，取得了一定的研究成果[2]。然而，这种知识表示语言大多从事物、关系、属性、事物种类、关联五个方面对于词语或概念给予表达，具有较好的网络检索能力，但是对自然语言中的隐含知识还很难给予明确的定义和揭示，同时也缺乏对上下文知识的短时记忆功能，因而在自然语言的处理过程中，能够提供常识知识辅助的作用十分有限。本文采用的领域分类体系是构建在 HNC 这一语义网络的基础上的，覆盖了整个人类活动的全部概念树，而且能够有效地把常识性知识和 HNC 概念基元符号体系相结合，语言表述中的隐含知识通过概念关联知识给予定义和揭示，形成了领域句类表示式这种形式化表示式，便于计算机使用，为句群、篇章处理奠定了基础。

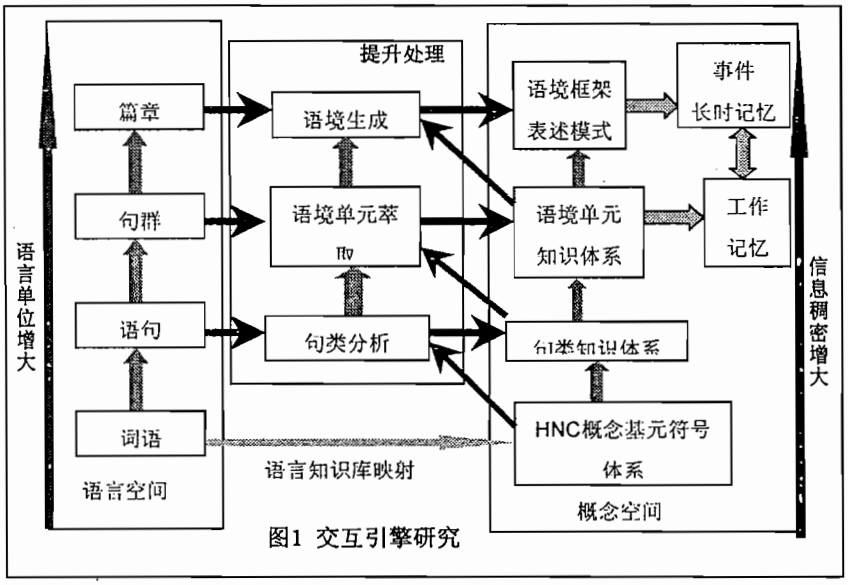

图1 交互引擎研究

领域句类是领域蕴含的世界知识的抽象浓缩，具有句类代码和表示式，通过特定的领域词语可激活相关的领域联想脉络，领域句类的设计是语境单元萃取技术不可或缺的一个环节，为语境单元框架的构建提供基本的要素，在交互引擎研究从语句到句群的提升处理中具有举足轻重的地位(详细的 HNC 交互引擎研究框图可见图 1)。

2. 相关术语

从框图 1 可看出，HNC 理论通过把语言空间的语言单位对应到概念空间中，在概念空间中研究语言。这样的研究策略造成了 HNC 理论形成了不同于其他研究的术语，在领域句类的研究中，主要涉及到的术语有：领域(DOM)、领域句类(SCD)、句群(SG)、语境单元(SGU)，为了便于读者阅读本文，以下采取从语言空间到概念空间的顺序，介绍上述 HNC 术语。

句群(SG)：句群是围绕着一个特定概念展开的一组话语。这个特定概念称之为“题”，“题”的转移意味着句群的变动。“题”在语言空间并不显现在音和形上，而是隐现在义上。HNC 概念基元符号体系的作用就是把这个语言空间隐现的义转变成概念空间显现的义，这样，计算机就有可能抓住这个“题”了[3]。

领域(DOM)：按照人类活动的不同对“题”进行划分得出的句群类别，即该句群陈述内容应归属的 HNC 领域。而所谓的 HNC 领域，也就是 HNC 理论所设计的领域分类方案，它以扩展基元概念为领域分类的主要依据。扩展基元概念全部用来描述人类活动，因为人类活动是自然语言表述的主体，也是领域分类的主要依据。当然，关于人类活动的领域分类，各行各业都作了十分细致的研究，其中的情报和图书分类尤为成熟，已经变成了独立的学科。HNC 可以在领域分类的底层表示里参考或沿用这些已有的成果，但领域分类的高层表示则作了另行设计。人类活动作为语言概念空间的一个一级子空间，需要一个统摄全局的整体表述，这就是扩展基元概念高层设计的功能使命。具体 HNC 十大类领域概念总览看见图 2，十大领域之间带有高低之分，一般来说，高层领域编号 1 和编号 2

必须和低层领域搭配表达，形成不同于单独低层领域构成的语境单元表示式，详细的语境单元表示式可详见文献[3]。

领域句类(SCD)：HNC 根据不同的领域所蕴含的世界知识形成的句群类型。领域句类具有自己的领域句类代码，在此基础上形成了领域句类表示式，有效地把 HNC 概念基元符号所蕴含地知识以领域句类表示式的形式提供计算机使用。

图2 领域概念总览

编号	类　型	符　号
1	第一类精神生活（心理活动，精神状态及行为）	71,2,3
2	第一类精神生活（人类思维活动）	8
3	第二类劳动（专业活动）	a
4	第三类精神生活（追求活动，理念活动）	b,d
5	第一类劳动	q6
6	第二类精神生活（业余活动）	q7
7	第二类精神生活（信仰活动）	q8
8	本能活动	6m(m=0~5)
9	实据	3228 n (n =8~b)
10	状态	503,50 n (n =8~b)

语境单元(SGU)：由领域 DOM、情景 SIT 和背景 BAC(又区分事件背景 BACE 和述者背景 BACA)三要素构成的一个结构体。领域 DOM 描述事件的类型，情景 SIT 描述事件的作用效应链表现，事件背景 BACE 描述事件发生的主客观条件，述者背景 BACA 描述叙述者//论述者的特定视野。而语境单元三要素并不构成一个三维度独立且等价的空间，而是一个以领域 DOM 为主轴的三维空间，其情景 SIT 和事件背景 BACE 都是领域的函数。语境单元可大体对应于语言空间的句群[3]。

3. 基本特点

领域句类体现的是一个句群所描述的中心话题所应包含的世界知识，这些知识是一个大语境背景环境下的世界知识的高度抽象浓缩，并且最终需要通过领域句类表示式这一形式化的方式提供计算机使用。领域句类作为一类独特的知识表示，拥有其个性的基本特点，主要包含如下：

1） 领域句类要能够体现概念基元符号体系蕴含的世界常识知识

领域分类的主要依据是扩展基元概念，而扩展基元概念同时也是 HNC 概念基元符号体系的构成主体之一，HNC 概念符号体系作为一个语义知识表示的网络，在设计之初即已充分考虑了领域概念所蕴含的世界知识，把这些世界知识以概念树及其延伸结构的形式定义在符号体系中，而领域句类必须充分考虑这些世界知识，在领域句类表示式这一形式化表达式中有所体现。

2） 辅块需要在领域句类表示式中体现，要在句类表示式的基础上进行适当的扩展

领域句类则是领域世界知识的深层结构表示式，从领域世界知识完整性考虑，有些领域知识一定是通过辅块的形式体现出来的，那么这些辅块就必须在领域句类表示式中得到体现。而句类表示式是 HNC 句类分析处理后得到的语句深层结构表示式，体现的是一个语句中各主要语义角色的概念联想脉络的类型，从语句的角度来看，辅块并不进入句类表示式，那么领域句类表示式必须在原有句类表示式的基础上进行扩展。

3） 领域知识可能出现多个领域句类表示式，不同领域句类表示式之间存在转换关系

领域世界知识体现的是一组语句所需体现的世界知识，领域句类表示式则是这一世界知识的形式化表示框架。从不同的描述角度来看，围绕这一形式化表示框架必然可以形成不同的领域句类表示式，而这些领域句类表示是必然存在一定的对应转换关系，这些转换关系必须在大量分析真实语料的基础上进行归纳总结得出。

4） 领域句类表示式代表的是一组语句的世界知识，故和实际语言可能存在偏差，它们之间存在整合关系

一般来说，实际语言的某一个语句往往仅表达领域的部分世界知识，全部的领域世界知识则在整个句群中得到体现，这些必然导致领域句类表示式同实际语句分析得到的句类表示式存在偏差，在具体的语境单元分析过程中，这些偏差需要进一步的明确整合。

4. 示例分析

领域句类知识的设计整体上采用的是先验的设计式同后验的具体句类表示式相互验证的方法最终确定下来的。为了便于读者能够更形象地了解领域句类知识及其句类表示式的设计步骤，下面我们通过具体的实例来进行说明。

4．1 延伸结构表示

我们选取的领域概念节点为人类第二类劳动(a)中的政权活动(a11)概念树的一级延伸结构最高领导人更迭制度选举方式(a113b)，最高领导人更迭制度(a113)是政权活动(a11)三个基本侧面的第一项内容。a113 具有两项延伸概念 a113e2n 和 a113t=b，前者描述轮换制和终身制，后者描述 a113 的三种基本方式。该节点延伸结构的具体表示式如下：a113 "最高领导人更迭" a113e2n "轮换制与终身制" a113t=b "最高领导人正常更迭的 3 种基本方式" a1139 "指定" a113a "推举" a113b "选举"

4．2 领域知识描述

选举 a113b 是最高（国家或地方政府）领导人更迭制度 a113 的一项延伸概念。而最高（国家或地方政府）领导人更迭制度 a113 是政权活动 a11 三个基本侧面的第一项内容。它具有两项延伸概念 a113e2n 和 a113t=b，前者描述轮换制和终身制，后者描述 a113 的三种基本方式。

从延伸结构的具体表示式中我们可以得出：对最高（国家或地方政府）领导人更迭 a113 的隐含知识的揭示首先应考虑政体的存在，即 a113 的是在什么样的体制下进行的；第二要体现 a113 的三种基本方式；第三要体现最高（国家或地方政府）领导人更迭的实质就是新的最高（国家或地方政府）领导人替代旧的最高（国家或地方政府）领导人，而新的最高（国家或地方政府）领导人是由指定者（或推举者或选民）指定（或推举或选举）出来的。t 的取值不同，T3A、T4B1、T4BC2 之间的概念关联也不同，本文将重点研究 t=b 即选举制度 a113b 的领域句类及相应知识。选举制度 a113b 体现了最高(国家或地方政府)领导人更迭的一种方式，主要描述选民(pva113b)在民主制度 a10e25 和轮换制政体 a011 的框架下，对最高（国家或地方政府）领导人的候选人(pvr53va113b)进行选择(va11b)，最终以新的最高（国家或地方政府）领导人(p44e61d01ju78e21(pj2//pea119))取代旧的最高（国家或地方政府）领导人(p44e61d01ju78e22(pj2//pea119))的过程。

4．3 领域句类表示式

通过我们对 a113b 概念延伸结构表示式的归纳总结，以及对该领域句类知识的概括，在此基础上我们可设计 a113b 的领域句类代码。具体如下：

SCRD(va113b):=ReMsCnT3(Y80)T4a1*311J=ReMs[T3A+T3T4a1+T4B1+T4BC2]

(Ms(va113b):=(vpfvr53a311b,pevra113b,pea117,pj2), Re(va113b):=ga50a113b,

Cn(va113b):=va113bckm,T3A(va113b):=52pva113b\k, T4B1(va113b):=pvr53va113b\k

T4BC2B(va113b):=p44e61d01ju78e22(pj2//pea119)

T4BC2C(va113b):=(c44e61,l02,(pj2;a119)),

其中 Re 与竞选法有关；Ms 与竞选工作、竞选班子、政党和国家有关；Cn 描述选举的不同阶段，选举的阶段不同，选民与候选人的对应关系也会发生变化；T3A(va113b)是选民；在选举过程中，T4B1(va113b)是最高（国家或地方政府）领导人的候选人；T4BC2B(va113b)是被替代者；T4BC2C(va113b) 是最高（国家或地方政府）领导人一职。而 T3A(va113b)、T4B1(va113b)和 T4BC2 (va113b)三个广义对象语义块又具有如下的概念关联式：

T3A(va113b):=52pa113b\k=0~x（选民的分类）

T4B1(va113b):=pvr53va113b\k=x（多个候选者的编号）

T3A\k:=T4B1\k（实际选民与具体候选人对应）

T4B1\k:=ub3（不同具体候选人之间是竞争关系）

T4BC2C(va113b):=(c44e61,l02,(pj2;a119))

T4BC2B(va113b):|f92)

T4B1(va113b)= T4BC2B(va113b)（表明该候选人是连任）

选民 T3A(va113b) k 的取值是从 0 开始，当 k=0 时，表示投弃权票或废票的这类选民，而选民的分类与具体候选人对应，即每个候选人都有拥护自己的选民，也是选票投向分类；候选人 T4B1(va113b)至少要有 2 个，即 x

大于等于 2；在实际语言表述中，T4BC2B(va113b)经常省略；当 T4B1(va113b) 和 T4BC2B(va113b)是同一人时，则表示其连任。

4．4 自然语言表述方式及其转换

实际语言运用中，关于“选举 v311b”这个领域概念有不同的表述方式，用以描述“选举”的不同侧面，采用不同句类代码，这些句类代码不一定完全等同于领域句类代码，但可以看作是 va11b 领域句类的转换，并在实际运用中替代领域句类使用。

①选举的效应描述

Y-ae71J(vra113b)// Y-ae72J(~vra113b)=YB+Y0

Y0-ae71J(vra113b)//Y0-ae72J(~vra113b)=YB+Y0+YC

上述句类是 SCD(va113b)= T3(Y80)T4a1*311J(va113b)的句类转换之一，它主要描述选举的结果，即某人当选或落选。其中 YB(vra113b//~vra113b):=T4B1(va113b)，YC(vra113b//~vra113b):= T4BC2C(va113b)。例如：“布什胜出”，“克林顿上台”，“反对派民主党总统候选人伯塞斯库以 51.23%的得票率当选总统”等等。

②选举的状态描述

X11S03-01e2m*21J(vra113b)=X1A+X11S0+SC

X11S03-01e2m*21J(vra113b)也是 T3(Y80)T4a1*311J(va113b)的句类转换之一，从状态角度对当选者进行描述。其中 X1A(vra113b):=T4B1(va113b)，SC(vra113b):=T4BC2C(va113b)。例如：“布什入主白宫”，“克林顿以绝对优势战胜共和党总统候选人多尔，蝉联第 43 届美国总统”，“乌克兰反对派候选人尤先科本周日正式就任总统”，“金大中于 1997 年 12 月被选为总统”等等。

③选举的作用描述

T3X03*21J(vra113b)=TA+T3X03+X03BC

T3X03*21J(vra113b)同样是 T3(Y80)T4a1*311J(va113b)的句类转换之一，它是选举的作用描述。其中 T3A(vra113b):=T3A(va113b)， X03BC:=X11S03-01e2m*21J(vra113b)，且有 X1A(vra113b):=T4B1(va113b)，SC(vra113b):=T4BC2C(va113b)的对应关系。例如：“你准备选谁当你的总统？”，“当时一些美国人胡乱选他（里根）当总统，算是歪打正着”。

4．5 语料分析

本语料来自新浪网关于 2004 年美国总统大选的一则法国《解放报》的新闻，全文如下。

美国总统大选投票即将开始，布什和克里到底谁能当选仍是个悬案。在“拉丹录像带”搅局之后，美大选局势变得更加扑朔迷离，《华盛顿邮报》等四个机构公布的民调都显示两位总统候选人的支持率不分伯仲。对于此次“关系到整个世界的选举”，71％的法国人希望克里胜出，而认同布什的人只有 11％。此外，77％的法国人认为，克里当选将使法国获益。但法国一些国际问题专家也认为，布什当选也有好处，起码希拉克总统已对布什相当了解，而克里当选则将带来更多不确定因素。（节选《法国人希望克里胜出》一文）

首先通过句类分析可以得到相对应的含有领域词语信息的词语组合“美国总统大选、投票、布什、当选、美大选、胜出和两位总统候选人”，“美国总统大选”限定了“大选”是政权活动 a11 概念树中的最高（国家或地方政府）领导人的更迭 a113 中的选举 ga113b，而不是一般的选举 a01a7b，因而“投票”也是关于美国总统更迭的，其 HNC 映射符号是 va113b；同时也限定了“当选”是关于美国总统选举的，HNC 映射符号是 vra113b；由“布什”这一词语的 HNC 符号可知：其为人名并于 2004 年选举时任美国总统，而“布什和克里”的并列可判断出“布什和克里”是美国总统的候选人，“克里”是“布什”的竞争对象；由“两位总统候选人”知道总统候选人有二个，即 k=2。这其中也为未登陆词“克里”的语义分析提供了语境知识保障。

上述内容都是关于美国总统选举局势(52,53)a113bUSA(ppj2*3y)的叙述，语言概念空间各语段是在符号(52,53)a113bUSA(ppj2*3y)的统摄之下，第一个语段描述的是无法判断何人能当选美国总统(52,53)a113bUSA(ppj2*3y)jru76e77，第二个语段是对总统大选局势不明朗(52,53)a113bUSA(ppj2*3y)jru76e77 的描述。两个语段都是对美国大选局势分析的描述，“题”没有发生转移，所以这二个语段是一个句群，紧扣“题”于最高领导人的选举 a113b，领域句类即为：

SCRD(va113b):=ReMsCnT3(Y80)T4a1*311J=ReMs[T3A+T3T4a1+T4B1+T4BC2]。

在领域句类确定之后，我们即可在此基础上进行语境单元表示式 HNC3 的内容填写，其语境单元的相关信息

如下：

DOM:= a113b

SIT=T3(Y80)T4a1*311J=T3A+T3T4a1+T4B1+T4BC2

(B1:=T4B1\1,T4B1\1:=pvr53va113b\1 USA(ppj2*3y),

B2:=T4B1\2,T4B1\2:=pvr53va113b\2 USA(ppj2*3y))

BACE(Cn1:=2004)

语境单元表示式 HNC3 的内容填写完毕之后，即可对计算机进行句群、篇章的处理提供先验的世界知识。其它领域的句群也可类似做相关处理。

5. 小结

计算机在进行句群及篇章处理时，必须结合上下文等语境知识进行处理。HNC 对这些语境知识按照人类活动的相关领域进行分类，在此基础上对这些领域进行研究，归纳其相关的领域世界知识，以领域句类的形式提供计算机使用，为句群及篇章的处理提供有力的知识支持。本文在详细阐述领域句类知识设计的特点的基础上，以最高领导人更迭制度中选举（a113b）为例，详细讲述了领域句类设计的整个步骤以及它如何在具体的实例句群语料中起知识辅助的作用。下一步我们将就领域句类知识体系的建立展开工作，为交互引擎的最终实现提供有力的知识保障。

参考文献

[1] 黄曾阳. HNC（概念层次网络）理论. 北京：清华大学出版社，1998

[2] 盛小平. 四种网络知识表示语言的比较研究. 情报学报，2004(6)

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