《当代语言学》第 15 卷 2013 年第 2 期 129 – 141 页，北京

“自下而上”与“自上而下”本体构建方法的探讨

陆勤　香港理工大学　　谌贻荣　新浪北京　　李素建　北京大学

提要　本体构建旨在对知识体系的概念和关系建模并形成体系化的知识，从而辅助计算机进行智能化的处理。上位本体包含的概念是与领域无关的通用概念集，而多数通过算法自动提取而获得的本体主要用于特定领域的概念知识。本文首先介绍在本体构建中领域核心本体(也称中位本体) 和应用(也称下位本体) 之间的关系。然后以计算机领域为例，阐述一个如何利用英文的上位本体 SUMO 通过“自上而下”的方法建立的中位本体。基于下位本体的特性，本文进而介绍一种“自下而上”的本体构建方法。

关键词　本体自动构建　上位本体　中位本体(领域核心本体)　下位本体/应用本体　上下位关系

1. 引言

本体被引用最广泛的定义是 Gruber(1995: 908) 给出的“本体是对共享概念的一个显式的形式化规范说明”。本体是显式的，因为本体所应用的概念和其使用的约束都是显式定义的，并且具有软件可访问性。也就是说，本体是形式化的，因为它是机器可读并具有可操作性。在一个领域中，本体构成了该领域任意知识表达系统的核心(Chandrasekaran，et al. 1999: 20–5) ，而领域概念要通过领域中必用的一些词项来表达，这些被称为术语的领域词项，是领域的根本知识和信息的承载单位。一个领域知识空间中的本体大多是由术语及不同术语所承载的相关概念的关系所构成的。

本体可分为三个层次: 上位本体(upper ontology) 、领域本体(domain ontology) 和面向应用的本体(application-oriented ontology) 。上位本体是跨领域可复用的通用本体。领域本体，有时也称为中位本体(mid-level ontology) ，描述某一个特定学科、专业或领域里最广泛使用的概念和关系，比如信息科技领域、医学学科。面向应用的本体是为某个应用而定制的本体知识库，例如体育运动领域中的足球比赛。有些应用可能需要包括跨领域的知识信息，例如，电子消费产品就同时涉及了信息科技和商业贸易两个领域。就面向应用的领域本体而言，上位本体和相关的领域本体可同时被称为其上层本体。

本体作为一个描述特定领域概念的知识库，其内容不仅包括领域的主要概念，还包括概念之间的关系。很多应用系统都可以利用本体所提供的领域知

识，比如事件信息抽取系统、信息检索系统等。本体甚至可以辅助进行互联网上的产品意见挖掘。与此同时，以传统的人工方式来建立本体需要耗费大量的人力，随意性高且难以及时更新。

在本体的构建中，大多数的上位本体是采用“自上而下”的方法人工构建的。一个目前被广泛使用的上位本体是 SUMO（the Suggested Upper Merged Ontology，建议上层共用知识本体）。SUMO 将几个公开可用本体的内容融合为一个具有一致性结构和广泛性基础的本体（Doerr，et al. 2003；Niles and Pease 2001）。它不仅包括概念的分类，也包括了可用于推导的公理和逻辑推断。到 2003 年 2 月，该本体已经涵盖了约 1000 个概念术语结点和 4000 个推断（Niles and Pease 2003：23）。另一个著名的上层本体是 DOLCE（Descriptive Ontology for Linguistic and Congnitive Engineering，融合了语言学和认知工程的描述本体，Gangemi，et al. 2002）。还有一个被广泛使用的英文词汇的本体是 WordNet（Miller 1995；Xu，et al. 2002：224）。WordNet 中的一个重要概念就是同义词集（synset）。同义词集是表达同一概念的同义词的集合。一个词的不同的词意被包括在不同的同义词集中。从本体的观点来看，同义词集和本体中的概念是等价的。实际上 WordNet 包含了同义词集到 SUMO 概念结点的映射，因此也可以将其视为对 SUMO 在词汇上的扩展。

中位本体和下位本体的构建既可以是自上而下的，也可以是“自下而上”的。自上而下的方式通常会利用现有的上层本体相关资源，例如 SUMO（Niles and Pease 2001）、WordNet（Miller 1995）或者 GermaNet（Xu，et al. 2002：224-5），并自动地从语料库、词典、知识库或半结构化文档等资源中抽取和构建出一个所需要的本体知识库（Mani，et al. 2004；Gomez-Perez and Manzano-Macho 2003：60-75）。自下而上的方式通常利用现有词典资源或语料中抽取的一些相关知识，获取领域概念之间的层次关系等。Mani 等（2004）综合使用了相对浅层的方法和现有可用的背景知识库构建所需的本体。Brewster 等（2002）主要通过互联网来获取本体。Maedche 和 Staab（2001）给出了本体学习方法的一个详尽介绍。现在虽然已有很多关于本体构建的研究，并已有不少付诸实现的系统和本体知识库，但自动构建本体的技术还远远不够成熟。可用的本体中大多数中位本体还是以人工创建为主，也有一些使用系统集成的方法而组建的本体。

自动本体构建主要由两个部分构成。第一部分是确定领域中的概念集合。因为术语是概念的指代，所以概念的发现常常是通过术语发现来完成。术语发现则从包括互联网文本、百科全书等不同的领域资源中抽取得来。也就是说，术语提取可被认为是本体构建的一个必要的预处理步骤。第二部分是关系发现，用以识别和提取概念之间的关系。对于给定的概念词 C，关系发现也被称

为属性发现。这些属性是和 C 通过某种关系相关联的一系列其他概念。在本体构建过程中进行的信息提取会用到许多自然语言处理技术，诸如分词标注、语块分割等(Maedche and Staab 2000) ，因此自然语言处理技术是本体构建中不可分割的部分。在本体构建的学习过程中，术语发现的技术相对比较简单。最难的问题是找到概念语意关系来构建一个结构化的知识空间，这是目前本体学习及本体适用性所面临的主要挑战。

本文旨在介绍两种自动构建本体的方法和如何将它们应用到中文的本体建构中。而这里指的建构主要是找出一个概念术语与其他概念术语的上下位关系，从而形成一个本体架构。其中第一个方法是利用领域资源自动提取核心术语并把其映射到词汇本体，同时连接上位本体从而建构领域核心本体(中位本体) 。由于核心术语可生成大量的领域术语，其对应概念和生成术语的对应概念间存在明显的上下位关系，所以中位本体为下位本体的构建打好了基础。第二个方法则利用领域术语的现有资源分析概念间的关系，可以采用形式概念分析方法从词典或语料库中获取表示概念的属性，利用概念和属性构成的形式上下文来计算概念间的关系。

本文的组织结构如下: 第二节介绍上位本体和下位本体构建的相关工作; 第三节介绍中位本体的构建方法，第四节介绍一种基于形式概念分析(FCA) 并利用现有资源构建下位本体的方法; 第五节总结全文。

2. 相关研究

在哲学意义上，术语“本体”是指研究实体的存在和形式，以及它们的范畴和关系。该术语被类似地应用在计算机领域，这里重点强调了在各个知识领域中本体的形式化表示和构成。当然目前对什么构成一个形式本体还没有统一的定义。Gruber(1995: 908) 把本体定义为“概念化的表示”，表明了本体和概念化之间的不同，本体是与语言相关的，而概念化是与语言无关的。在 Gruber 的定义中，本体的目标是研究某个领域中存在的实体的范畴化。这里，我们把本体看作是当使用语言 L 的人在讨论领域 D 时，他所能考虑到的领域 D 中存在的事物类型。简而言之，一个本体描述了一个领域概念体系，概念之间通过蕴含关系相关联(Guarino 1998: 527-8) 。

Sowa(2000) 对形式和非形式本体做了区分。一个非形式本体表示为一系列事物类型，有的类型未被定义，有的类型通过自然语言的词汇进行描述。而形式本体则被表示为一个形式化概念的集合，利用类型-子类型关系构成的偏序关系组织起来。据此，我们定义本体如下:

定义 1　本体，表示为 O，被定义为一个四元组 O = (L，D，C，R) ，其中 L 表示特定的语言，D 表示特定的领域，C 表示概念集合，R 表示概念间的关系。

在计算机技术的范畴内，本体指的一般都是形式本体。由于任何一个本体的构建方法都有其适用的某个特定领域 D，并面向某种语言 L，则主要的构建任务是确定如何获得 C 和 R，当然具体构建方法也会和语言 L、领域 D 相关。语言在这里可以理解为相应语言的词汇。

基于以上定义，本体构建的自动或半自动方法通常包括两个步骤：获取领域术语和识别领域术语之间的关系。领域术语的获取技术这里不再详述。领域术语关系识别的通用方法是在大规模语料库中应用启发式规则(Maedche and Staab 2000；Hearst 1992)或简单的统计技术(如互信息)(梁晓波等 2010；李向阳 2010)。基于规则的方法虽然可以较准确地识别关系，但由于语言的灵活性使得规则很难覆盖和总结关系所在的上下文，这样只能获取有限的关系。而简单的统计技术虽然可以获取更多的关系，但也带来很多噪音，使得所获取关系的质量严重下降。

形式概念分析(formal concept analysis，FCA)是自动构建本体的一种新方法，它是一种基于格理论的数学分析方法。由于形式概念格可以很自然地表示层次和分类，FCA 已经从纯粹的数学工具转化成计算机科学中一种有效的方法(Stumme 2002：2-10)，并被很多研究者用于构建本体。Haav(2003)使用 FCA 进行了房地产领域本体的构建，Jiang 等(2003)采用 FCA 方法构建了医疗领域本体。本文也将在下位本体构建中使用 FCA 方法获取领域术语的关系。

3. 自上而下中位本体的构建

在给定某个领域的概念词汇集(也称术语)的前提下，一项重要的工作就是自动创建领域的核心概念本体。核心概念指的是由核心术语形成的本体。核心术语具有能产性高和领域特定两个特性。能产性高的术语可以较多地在其他术语中作为术语构件被采用。能产性高的核心概念术语构成的核心本体更能发挥其作为领域本体和上位本体的中间层的作用。

在 Chen 等(2006)的工作基础上改进的核心术语抽取算法(Ji，et al. 2007)首先做后向最大词典切分，然后基于前述能产性高的理由用词频排名做领域性过滤。所谓后向最大词典切分算法，其输入是词典，被切分的对象是词条，输出的是切分后的词典；切分方法是以输入的词典为切分词典，同时对该词典的每一词条切分之前，暂时在切分词典中去掉当前被切分的词条，然后反向最大切分当前词条。这样就保证了词典的每一个多字词条都会被切成更小的词段，这些被切分的词段就是当前术语的最大术语构件。由于是最大切分，避免了父串对子串频率的叠加效应，有效地去除了构件嵌套产生的短构件淹没长构件的效应。比如，“计算机”内嵌了“机”，如果不采用最大切分法的话，就会造成构件“机”的频度排名更靠前。而实际上在信息技术领域，“计算机”作为构件直

接合成术语比“机”更频繁，意义更明确。所以采用最大切分是必须的。在切分后的词典里统计术语构件的词频，按词频从高到低排名，就形成了一个术语构件词频表，另外取通用领域的词频表作为对照。一个术语词条，如果在领域中的排名比通用领域中的排名高出设定的一个阈值，将予以保留，否则删除。经过这两步后，一个核心术语词表就自动产生了，实验证明该列表的质量较高(Ji，et al. 2007)。

而核心本体的建构可以选择用自上而下或自下而上的方法获取。用自上而下的方法就是直接找到每一个核心概念术语在已有上位本体中的最佳节点。这样每个核心概念就可以找到其上位，而且通过上位本体所提供的路径还可以找到核心概念之间的关系。换句话说，该中位本体的构建变成了对上位本体的延伸和扩展。而自下而上的基于语料进行的关系获取对语料要求极高，如果语料规模不够大或者质量没保证，得出的结果噪音通常很大。

本文先介绍的是一个自上而下建立核心本体的方法，其基本原理是在核心术语提取之后，将每一个术语概念映射到 WordNet 中的同义词词义，再通过 WordNet 提供的对 SUMO 中概念的映射来完成所有中位核心概念术语到上位概念的映射，从而继承上位本体和领域相关的概念、关系以及公理。这种自上而下的方案中，一般要假定有上位本体可用，并且已有一个领域概念术语词表。因此，该词表可以和上位本体做对齐或者把领域概念词加为上位本体的扩展。

本文中介绍的核心本体构建算法(缩写为 COCA)是专为创建中文核心本体而设计的(Chen，et al. 2008；谌贻荣等 2009)，该方法并不局限于计算机领域。因为没有足够可用的中文自然语言处理资源，COCA 利用了一个大规模中英文术语库和现有的英文 WordNet 作为资源对 SUMO 进行扩展。而扩展则是利用了 WordNet 中提供的由每一个同义词集到一个 SUMO 中上位概念的映射。COCA 的主要思想是把每一个中文核心术语 T_C 首先映射到最合适的英文 WordNet 同义词集 SynsetC。然后利用中文核心术语集、SUMO 层级结构、每一个同义词集在 SUMO 中对应的上位概念以及同义词集本身来构建一个领域内的全部核心术语组成的本体。而核心本体通过继承 WordNet 的上位结构并扩展 SUMO 层级结构来构建。

COCA 需要解决的主要问题就是在给定一个中文核心术语 T_C 的情况下，如何在 WordNet 中找到合适的英文同义词集 S。也就是:

$$\underset{s}{\operatorname{argmax}} p(S \mid T_C)$$

基于这一原理，具体的 COCA 架构图和处理步骤如图 1 所示。

为了找到最合适的映射，必须要解决两个层次上的歧义。第一，给定一个

中文核心术语 T_C，作为一个词典条目，它可能有多个到英语的翻译，这一部分由统计翻译模块完成。第二，给定每一个英语词条，该词条又可以有多个意义，可以对应到 WordNet 中不同的同义词集。在 SUMO 和 WordNet 之间的映射已经提供的前提下，COCA 的主要目标是为每一个中文核心术语找到最合适的 WordNet 同义词集。在消歧的几个步骤上，应用了三种统计信息，包括：(1) 中文到英文的翻译统计信息；(2) 同一术语词条在不同的同义词集的词义统计信息；(3) 所映射上位概念的适合性，而这些工作由歧义消解模块来完成。

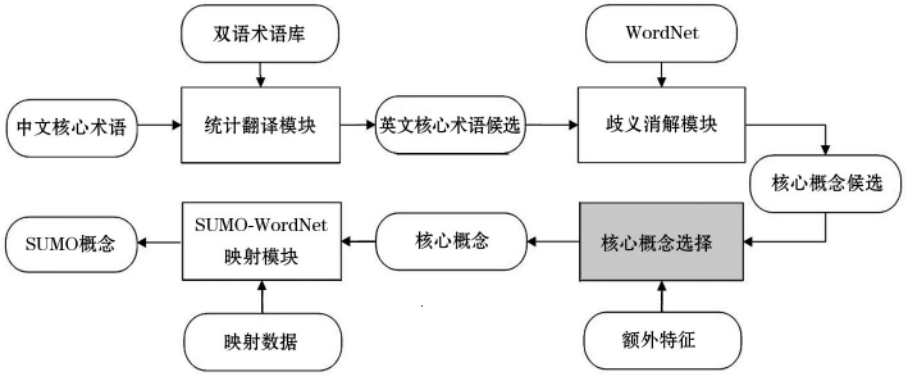


图 1　COCA 架构图

我们还注意到很多额外的特征可以提高映射性能，而其中后缀词集是最重要的一种。很多具有相同后缀的中文词都会有着共同的上位概念。举例来说，“驱动器”对应的英文“driver”有歧义，一般表示一种“人”(human)－“司机”，也可以表示一种“设备”(device)，而与“驱动器”(driver) 具有相同后缀的词如“服务器”(server)、“传感器”(sensor) 等与“驱动器”(driver) 具有相同的上位概念“设备”；在通用领域，这些词多是“人”的下位词，而在信息科学领域中，这些词都应该是“设备”的下位概念，并且都以“器”作为后缀。这个例子提示我们可以用后缀词来改进词义的正确映射，从而改进自动建构的核心本体的质量。下面的问题就是如何找到并利用重要的上位概念来改进下位词在特定领域中的词义映射。重要的上位概念有两个方面的特性，一方面，在共享后缀词中一个上位概念的下位词越多，该上位概念就越重要。另一方面，上位概念越抽象，其对下位概念的辨别区分能力就越弱；也就是说，一个上位概念和下位概念之间距离越近，越具体，该上位概念越重要。基于以上两点分析，在核心本体算法(COCA) 的架构里增加了共享后缀词特性的集作为额外特征，应用在核心概念选择模块(图 1 的灰色模块) 中来提高性能。该工作由核心概念选择模块来完成。最后的 SUMO–Wordnet 映射模块直接利用已有的英文映射信息来完成中文概念词汇至 SUMO 概念节点的映射。

下图是自动创建的中文核心本体的一个例子。最顶层的是 SUMO 中对应的上位本体概念，其下就是继承的核心概念。这个例子中显示的全部中文核心术语都正确地映射到了对应的概念，但如果不加入共享后缀集计算就会有错误。比如“例程”(routine) 会被对应到例行公事，而不是例行的计算机程序。

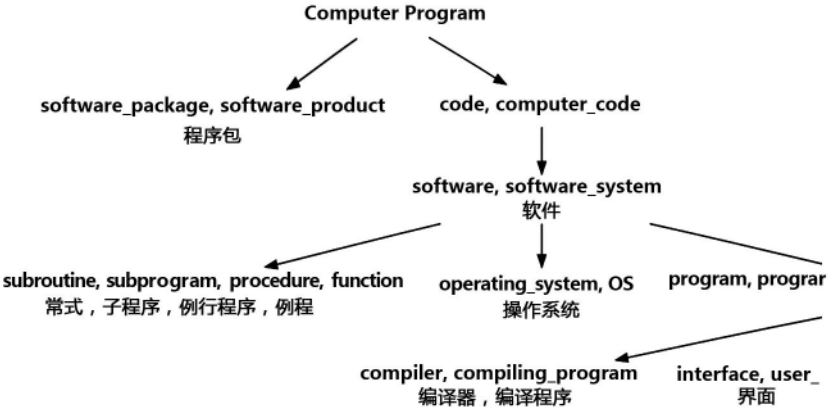


图 2　自动创建的中文核心本体的一个例子

COCA 的性能评估显示该算法在将计算机领域的核心术语集合映射、扩展到 SUMO 本体的正确性上达到 65% 的精确度。很多的性能问题主要源自中文词条所对应的英文词条在 WordNet 中缺少对应的条目。另外一个问题是由于 WordNet 中的词义信息收集主要来自通用语料库，因此与特定领域中作为术语的词汇语意分布并不十分吻合。后期我们也利用过其他中文本体资源例如 Sinica BOW(Huang，et al. 2004) 去优化 COCA 的性能(Chen 2011:76-87) ，但发现由于不能提供更多领域信息，性能的提高非常有限。随着更多中文本体和领域资源的出现，我们期待自上而下的方法可以取得更好的性能。尽管还有很多可以改进的空间，这项工作还是目前使用自上而下方法来扩展中文本体最全面的一项工作。

4. 自下而上的下位本体构建

本文中，自下而上的下位本体构建主要指利用词典资源或语料库的上下文来识别领域中概念术语之间的关系。大多数的方法都是通过基于规则、模式、或机器学习的方式对语料中的上下文进行分析，进而抽取不同概念之间的关系，而我们的工作则采用了形式概念分析(FCA) 模型的方法。

首先我们对 FCA 进行介绍，从而说明为什么 FCA 可以自然地作为一种自下而上构建本体的工具。FCA 是一种形式化数据分析和知识表示的技术(Ganter and Wille 1999) 。给定概念术语之间的形式上下文，FCA 可以自动构造形式概念的网格结构以替代耗时的手工本体构建。在抽象意义上，FCA 通常包括两个数据集合，即对象集合和属性集合。对象对应于领域中出现的概念，属性则

用来描述这些概念。FCA 能够在两个数据集合之间确定一种二元关系，这些二元关系可以构建一个形式化的上下文关系格，而且满足偏序关系。在此基础上构造出一个形式概念网格，其中涵盖了概念间的包含关系。FCA 中形式上下文和基于形式上下文的形式概念定义如下(Ganter and Wille 1999:17-8)：

定义 2　形式上下文为一个三元组(G,M,I)，G 表示对象集合，M 表示属性集合，I 表示笛卡儿积 G×M 上的关系。

定义 3　形式上下文中的形式概念为一个二元组(A,B)，A⊆G，B⊆M，Intent(A) = B(A 的内涵是 B)和 Extent(B) = A(B 的外延是 A)，其中

Intent(A):={m∈M | (g,m)∈I,∀g∈A},

Extent(B):={g∈G | (g,m)∈I,∀m∈B}

这里的函数 Intent( )和 Extent( )分别表示形式概念(A，B)的内涵和外延。

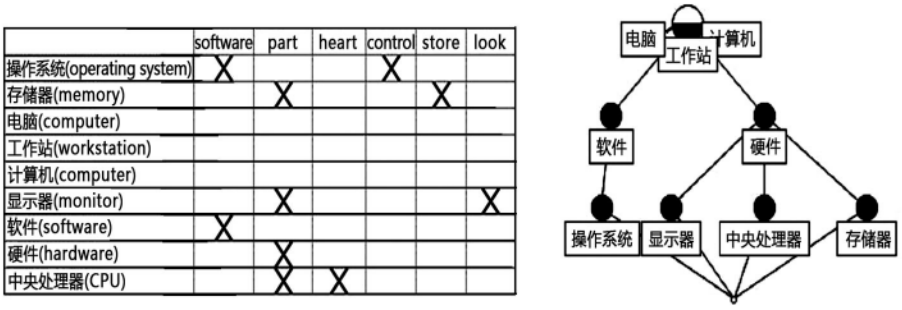
在一个形式概念(A,B)中，A 是外延，B 是内涵。形式概念满足偏序关系。如果用⊆来表示它们外延和内涵之间的包含关系，这种偏序关系“≤”的形式化描述则可以定义如下:

$$(A1,B1)\leq(A2,B2)\Leftrightarrow A1\subseteq A2 \text{ and } B2\subseteq B1$$

在概念网格中，形式概念(A1,B1)比它的上位概念(A2,B2)包含了更多的属性。另一方面，一个形式概念关联的属性越多，它包含的对象就越少。因此在网格中，一个形式概念要比上位概念包含的对象少而涵盖的属性多。其中有一些属性虽然可以描述对象的特点，但这些属性的存在与否不影响概念网格的构建，也就是说这些属性具有等价性，或者说概念网格中已有同类的属性，无须重复。

由于本体基本上由一个概念集合组成，概念可以通过对象和关系描述。这些概念对象最简单和最重要的关系是类别关系，可以看作是二元的、满足偏序关系的。因此 FCA 很自然地成为本体构建的一种工具。具体来说，领域术语可以看作是概念对象，而利用 FCA 算法设计的关键是如何选择适当的可建立偏序关系的属性集合。我们较早的研究利用 HowNet 词典[①]中对于概念的定义作为属性来建立不同概念间的关系(Li，et al. 2006)，由于每个概念的属性都是专家精心加工的，在属性选择上更为合理。例如，在 HowNet 中，“解码(decode)”可以看作一个术语，定义为一个义原集合，包括“computer|电脑、translate|翻译”和“software|软件”。这样，从 HowNet 中可以得到一系列术语的概念上下文及其概念格，如图 3-4 所示。

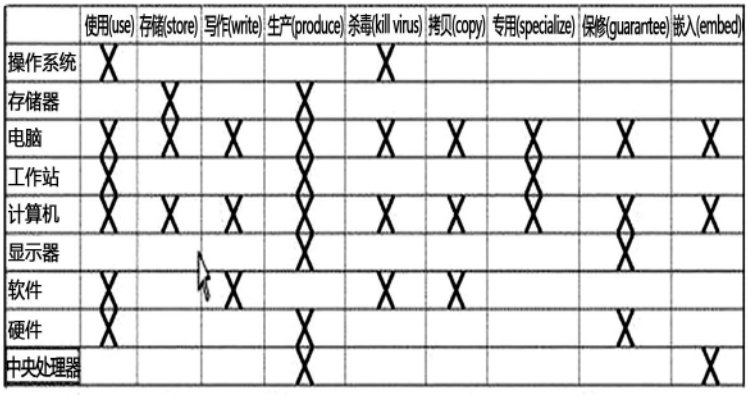
① http://www.keenage.com [accessed 27, Dec. 2012]



	software	part	heart	control	store	look
操作系统(operating system)	X			X		
存储器(memory)		X			X	
电脑(computer)						
工作站(workstation)						
计算机(computer)						
显示器(monitor)		X				X
软件(software)	X					
硬件(hardware)		X				
中央处理器(CPU)		X	X			

图 3　基于 HowNet 的概念上下文和概念格

然而利用现有词典资源获取属性的方法通常不适合特定领域概念关系的构建，主要原因是现有词典资源通常面向通用领域，而大部分领域术语在词典中没有给出定义。这就使得我们需要从大规模语料库中去发现领域术语的属性。在一个语料库中，上下文中概率上显著的共现动词通常适合作为概念术语的属性，以此来建立不同概念间的关系(Li，et al. 2005: 33) ，其原因是术语通常为事物或事件的名词性表示，在语言表示中，动词通常会紧邻着术语，体现出这些事物或事件的一些特征或状态。因此，这种选择是很自然的。图 4 给出一个示例，其中“使用、存储、写作”等动词可以看作是从语料库中抽取的能够反映概念属性的特征。



	使用(use)	存储(store)	写作(write)	生产(produce)	杀毒(kill virus)	拷贝(copy)	专用(specialize)	保修(guarantee)	嵌入(embed)
操作系统	X				X				
存储器		X		X					
电脑	X	X	X	X	X	X	X	X	X
工作站	X			X			X		
计算机	X	X	X	X	X	X	X	X	X
显示器				X				X	
软件	X		X		X	X			
硬件	X			X				X	
中央处理器				X					X

图 4　基于语料库的概念上下文示例

为了验证基于 FCA 和语料库自动构建本体的想法，我们选择计算机领域进行实验。由于需要人工标注来计算关系识别的精度，为了增强人工判别的准确性、减少工作量，我们只选择了 49 个计算机领域常见的术语，从语料库中利用共现分析找出 68 个动词作为这些领域术语的属性。等价(equivalent) 和上下位(is_a) 关系的识别精度大约为 56%。虽然我们的实验规模较小，但也充分表明，采用全自动的基于动词的 FCA 方法建立的本体可以直接使用。另外值得指出的是，基于 FCA 所建立的辅助工具能够帮助判定等价的同义词并进行自动合

并。我们在一项后续研究中探讨了其他词性的词作为属性的效果(Cui，et al. 2009)，并发现在较低统计阈值区间的动词作为属性的效果更好，同时对名词、形容词、副词等各种词性也做了研究。尽管形容词和副词作为单一的属性并不好用，但和名词、动词组合在一起使用的时候能够带来5%到10%的性能提升。

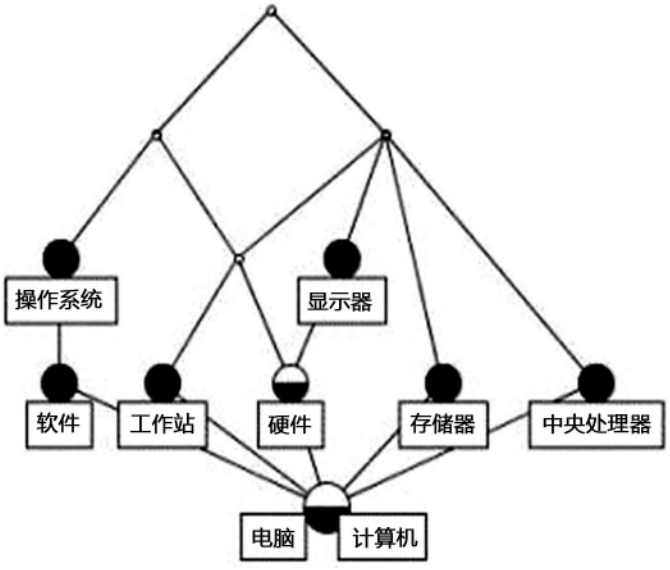


图5　基于语料库的概念格示例

在自下而上的本体构建中，值得探讨的是语料的选择，因为语料的质量、规模和一致性会直接影响到概念属性的选择，所以使用全自动的自下而上的方法建立本体对语料的质量和规模要求相对高很多。从目前的情况看，我们认为完全通过领域的语料进行自下而上的本体构建在质量上没有保证。在实际应用中，自下而上的本体构建更适于在新的术语发现的前提下对已建立的本体进行扩充。也就是在领域里发现词汇之后，通过语料和现有的领域词汇作为其属性来辨识其所属的上位概念，从而将其扩展到已有的领域本体上。换句话说，我们认为应该利用自上而下的方法建立某个领域的核心本体，然后在此基础上通过自下而上的方法扩充领域本体。

5. 结论和展望

领域本体的构建可以为不同的应用系统提供领域知识。由于科学技术的发展日新月异，领域知识的创建和更新不能沿用传统的专家式人工方法。同时，计算应用对领域知识的需求也催生了具有可操作性的本体构建。本文首先介绍本体的不同层级及它们的特点。并在此基础上，介绍了一种自上而下的中位本体的自动建构方法，该方法充分利用中英词典和英文的词汇语意知识WordNet

对上位本体的映射关系，并通过计算机的算法利用统计知识和领域词汇的特征进行两个阶段的消歧来完成。这种方法着重于上位本体的对齐和下位的扩展，从而承袭上位本体现有的结构。而第二种方法则是利用自下而上基于语料的上下文信息来抽取概念术语之间的关系。该方法直接利用了形式概念分析的模型，先抽取领域概念上下文中的属性信息并在统计信息支持的情况下，再通过寻找偏正关系建构概念网格来发现上下位关系。实验证明目前使用计算机进行全自动本体建构的挑战仍然很大。面对的问题是中文相应资源的规模、质量、以及全面性都不能满足全自动抽取的要求。但自动抽取技术作为辅助工具，则可以大量减少人工、节省资源、加速构建和更新的速度，不失为一种可行的方法。特别是中位本体的概念词汇相对稳定，通过计算机辅助的方法，可以用自上而下的方法较快地建构。加上人工确认之后，可有效加快不同领域核心本体的建构。而自下而上的方法则更适用于在发现新的领域词汇时，通过抽取其相关的领域核心概念对已有的领域本体进行扩充。同时，也可以用自下而上的方法应用本体建构提供可衔接的本体知识，有利于知识空间的完整和关联。

引用文献

Brewster，C.，F. Ciravegna，and Y. Wilks. 2002. User-centred ontology learning for knowledge management. *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems*. Stockholm，Sweden. Lecture Notes in Computer Sciences. Springer-Verlag. Pp. 203 – 7.

Chandrasekaran，B.，J. R. Josephson，and V. R. Benjamins. 1999. Ontologies: What are they? Why do we need them? *IEEE Intelligent Systems and their Applications* 14，1: 20 – 6.

Chen，Y.（谌贻荣）. 2011. Relation extraction for ontology extension using integrated evidences. Ph. D. diss.，Hong Kong Polytechnic University，Hong Kong.

Chen，Y.，Q. Lu（陆勤），W. Li（李文捷），and G. Cui（崔高颖）. 2008. Chinese core ontology construction from a bilingual term bank. The 6th International Conference on Language Resource and Evaluation，Marrakech，Morocco.

Chen，Y.，Q. Lu，W. Li，Z. Sui（穗志方），and L. Ji（纪鹭宁）. 2006. A study on terminology extraction based on classified corpora. *Proceedings of the 5th International Conference on Language Resources and Evaluation*（*LREC'* 06）. Genoa，Italy. CD-ROM version.

Cui，G.，Q. Lu，W. Li，and Y. Chen. 2009. Automatic acquisition of attributes for ontology construction. *Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages*（*CCPOL* 2009），Hong Kong. Pp. 248 – 59.

Doerr，M.，J. Hunter，and C. Lagoze. 2003. Towards a core ontology for information integration. *Journal of Digital Information* 4，1: 169.

Gangemi，A.，N. Guarino，C. Masolo，A. Oltramari，and L. Schneider. 2002. Sweetening ontologies with DOLCE. *Ontologies and the Semantic Web*. Proceedings of the 13th International

Conference on Knowledge Engineering and Knowledge Management. Springer-Verlag. Pp. 166 – 81.

Ganter, B. and R. Wille. 1999. *Formal Concept Analysis, Mathematical Foundations.* Berlin/Heidelberg/New York: Springer.

Gomez-Perez, A. and D. Manzano-Macho. 2003. A survey of ontology learning methods and techniques. OntoWeb Deliverable, OntoWeb Consortium.

Gruber, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human and Computer Studies* 43,907 – 28.

Guarino, N. 1998. Some ontological principles for designing upper level lexical resources. *Proceedings of the* 1*st International Conference on Language Resources and Evaluation.* Granada, Spain. Pp. 527 – 34.

Haav, H.-M. 2003. An application of inductive concept analysis to construction of domain-specific ontologies. *Proceedings of the Pre-conference Workshop of VLDB* 2003. Computer science reports 14/3, Brandenburg University of Technology at Cottbus. Pp. 63 – 7.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th International Conference of Computational Linguistics. Newark, DE. Pp. 539 – 45.

Huang, C. R. (**黄居仁**), R. Y. Chang (**张如莹**), and S. B. Lee (**李祥宾**). 2004. Sinica BOW (Bilingual Ontological WordNet): Integration of bilingual WordNet and SUMO. *Proceedings of the* 4*th International Conference on Language Resources and Evaluation* (*LRE*2004). Lisbon, Portugal.

Ji, L., Q. Lu, W. Li, and Y. Chen. 2007. Automatic construction of a core lexicon for specific domain. *Proceedings of the* 6*th International Conference on Advanced Language Processing and Web Information Technology*. Luoyang, Henan. Pp. 183 – 8.

Jiang, G., K. Ogasawara, A. Endoh, and T. Sakurai. 2003. Context-based ontology building support in clinical domains using formal concept analysis. *International Journal of Medical Informatics* 71,1:71 – 81.

Li, S. (**李素建**), Q. Lu, and W. Li. 2005. Experiments of ontology construction with formal concept analysis. *Proceedings of the OntoLex Workshop IJCNLP* 2005. Jeju, South Korea. Pp. 67 – 75.

———. 2006. Interaction between lexical base and ontology with formal concept analysis. *Proceedings of the* 5*th International Conference on Language Resources and Evaluation.* Genoa, Italy. CD-ROM version.

Maedche, A. and S. Staab. 2000. Mining ontologies from text. In R. Dieng and O. Corby, eds., *Knowledge Acquisition, Modeling and Management*. The 12th International Conference, EKAW 2000. Juan-les-Pins, France. Lecture Notes in Computer Science. Springer.

———. 2001. Learning ontologies for the semantic web. The 2nd International Workshop on the Semantic Web (Semweb' 2001), Hong Kong.

Mani, I., S. Samuel, K. Concepcion, and D. Vogel. 2004. Automatically inducing ontologies from corpora. *Proceedings of CompuTerm* 2004: *The* 3*rd International Workshop on Computational Terminology* (*COLING'* 2004). Geneva, Switzerland. Pp. 47 – 54.

Miller G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38,11: 39 – 41.

Niles, I. and A. Pease. 2001. Towards a standard upper ontology. *Proceedings of the International Conference on Formal Ontology in Information Systems, Volume* 2001. New York, NY: ACM

Press. Pp. 2 – 9.

———. 2003. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proceedings of the* 2003 *International Conference on Information and Knowledge Engineering*. Las Vegas, NV. Pp. 23 – 6.

Sowa, J. F. 2000. *Knowledge Representation, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks/Cole Thomson Learning.

Stumme, G. 2002. Formal concept analysis on its way from mathematics to computer science. *Proceedings of the 10th International Conference on Computational Science*. Bulgaria. LNAI 2393. Heidelberg: Springer.

Xu, F. (徐飞玉), D. Kurz, J. Piskorski, and S. Schmeier. 2002. A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. *Proceedings of LREC* 2002, *the* 3*rd International Conference on Language Resources and Evaluation*. Las Palmas, Canary Island, Spain. Pp. 224 – 30.

谌贻荣、陆勤、李文捷、崔高颖，2009，**一种基于共享后缀术语集改进中文核心领域本体构建的方法。见孙茂松、陈群秀编《中国计算机语言学研究前沿进展(**2007 – 2009**)》。北京：清华大学出版社。**370 – 5 **页。**

———，2010，**中文核心领域本体构建的一种改进方法**》《中文信息学报》第** 1 **期，**48 – 53 **页。**

李向阳，2010，**一种基于语料库和互信息的本体学习方法**》《微型机与应用》第** 10 **期，**76 – 9 **页。**

梁晓波、张飞、刘伍颖、马晓雷，2010，**基于语料库的军事本体构建**》《国防科技》第** 1 **期，**24 – 8 **页。**

第一作者简介

陆勤，女，博士，香港理工大学计算学系教授。研究兴趣：计算语言学、中文信息处理、基于自然语言处理技术的信息抽取和知识发现、开放系统及中文编码标准化。电子邮件：csluqin@ comp. polyu. edu. hk

LU Qin, female, Ph. D., is a professor at the Department of Computing, Hong Kong Polytechnic University of Hong Kong. Her research interest includes computational linguistics, Chinese information processing, information extraction and knowledge discovery using NLP methods. She has also been working in open systems and Chinese coding standardization. E-mail: csluqin@ comp. polyu. edu. hk

作者通讯地址：　陆　勤 香港理工大学计算学系
谌贻荣 100080 **北京市中关村朔黄发展大厦** 7 **层新浪微博搜索部**
李素建 100871 **北京大学信息科学学院 计算语言学研究所**

《当代语言学》第 15 卷 2013 年第 2 期 250 – 252 页，北京

Abstracts of Articles

LU Qin，CHEN Yirong，and LI Sujian，The construction of ontology: Top-down approach vs bottom-up approach

Ontology construction aims to build conceptual knowledge in such a way that the relations among major concepts can be explicitly identified and presented in a machine operable way so as to assist in intelligent processing of computer applications. An upper-level ontology includes general concepts that are used broadly across different domains whereas ontologies acquired by computing through algorithms automatically are more likely to be domain specific. This paper first introduces domain specific core ontology (mid-level ontology) and application domain ontology (lower-level ontology). Then, it presents a top-down approach to build a core ontology for Chinese in the IT domain based on the English upper level ontology SUMO and other English-Chinese resources available. The paper also introduces a bottom-up approach to build domain specific ontology using corpus based approach.

Keywords: automatic ontology construction, upper-level ontology, mid-level ontology (domain core ontology), application domain ontology/lower level ontology, hypernym relations

CHOU Yamin and HUANG Churen，The formal representation for Chinese characters

The formal representation of Chinese characters using ontology is an important research area, and advantageous to process Chinese language. This paper aims to describe the methodology of constructing the ontology of Chinese characters and its formal representation. The formal representation proposed herein includes the external structure and derivation of Chinese characters, semantic and phonetic symbols, internal structure, sense and derived words, the relations of variants, and the pronunciations. The semantic symbols and senses of characters are connected with IEEE Suggested Upper Merged Ontology (SUMO). This study uses the OWL (Web Ontology Language)-DL to describe the knowledge of Chinese characters and share with other ontology.

Keywords: formal representation, Chinese characters ontology, SUMO

HSIEH Shukai，Lexical semantic relations in Chinese: A preliminary study on the classification, logical validation and evaluation method

In recent years, construction of lexical knowledge resources like WordNet has become one of the common interests among lexical semantics and ontological knowledge engineering. The labeling of different semantic relations in lexical resources not only constitutes the base but also has great