信息管理专业前沿论丛

多语言领域本体学习研究

章成志 著

南京大学出版社

本课题研究受到以下项

◎中国博士后科学基金特别资助项目
多语言领域本体学习研究（200801105）
◎教育部人文社会科学研究一般项目
多语言领域本体的自动构建研究（08JC870007）
◎中央高校基本科研业务费专项资金资助
Web2.0环境下多语言标签自动聚类研究（NUST2011ZDJH15）

信息管理专业前沿论丛

多语言领域本体学习研究

章成志 著

南京大学出版社

内容提要

多语言领域本体是一种解决互联网信息资源语义化和多语言化需求问题的重要资源，在跨语言信息检索、机器翻译等多语言科技信息服务中具有重要作用。本书介绍当前国内外关于多语言本体学习方法、工具以及应用项目等的相关动态。本书围绕多语言领域本体学习中的两个关键问题，即：双语术语抽取与概念层次体系构建问题，进行了深入研究。该书研究内容主要包括：基于领域平行语料抽取的双语核心术语抽取研究、基于多层特征的一体化策略术语抽取研究、基于术语度约束的双语术语对齐研究、基于术语聚类的概念层次体系生成研究、基于多语文本聚类的主题层次体系生成研究。本书可为图书馆学、情报学、计算机科学与技术、信息管理和信息系统等专业的研究生和高年级本科生，以及从事语义网、文本挖掘、知识组织、数字图书馆等方向的科研人员，提供教学参考和技术指导。

图书在版编目(CIP)数据

多语言领域本体学习研究 / 章成志著. — 南京 : 南京大学出版社，2012.10

ISBN 978-7-305-10670-5

Ⅰ. ①多… Ⅱ. ①章… Ⅲ. ①语言学—研究 Ⅳ. ①H0

中国版本图书馆 CIP 数据核字(2012)第 237718 号

出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　　编 210093
网　　址 http://www.NjupCo.com
出 版 人 左　健

书　　名 多语言领域本体学习研究
著　　者 章成志
责任编辑 沈　洁　　编辑热线 025-83593962

照　　排 南京南琳图文制作有限公司
印　　刷 南京人文印刷厂
开　　本 787×1092　1/16　印张 13.75　字数 322 千
版　　次 2012 年 10 月第 1 版　　2012 年 10 月第 1 次印刷
ISBN 978-7-305-10670-5
定　　价 30.00 元

发行热线 025-83594756　83686452
电子邮箱 Press@NjupCo.com
Sales@NjupCo.com(市场部)

序　言

章成志博士的《多语言领域本体学习研究》这本书主要研究了构建专业领域本体的理论和方法，他在基于领域平行语料抽取的双语核心术语研究、基于多层特征的一体化策略术语抽取研究、基于术语度约束的双语术语对齐研究、基于术语聚类的概念层次体系生成研究、基于多语文本聚类的主题层次体系生成研究等方面，都做了创造性的工作，取得了很好的成绩。

本书研究的专业领域本体构建的理论基础是知识本体(ontology)。

在自然语言处理中，知识本体起着重要的作用。例如，在英语的自动分析中，with 介词短语有时可以分析为宾语的修饰语，有时可以分析为中心谓语动词的方式状语，这个问题叫做介词短语附着问题(Preposition Phrase attachment，简称 PP-attachment)。根据知识本体，可以对 with 介词短语得出完全不同的分析结果。

我们来看下面的句子：

I saw a man with a binocular.

("我用望远镜看一个人","我看见一个带着望远镜的人")

I saw a star with a telescope.

("我用天文望远镜看一颗星星")

I saw a molecule with a microscope.

("我用显微镜看一个分子")

为了正确地分析这样的句子并理解其意义，我们需要各种有关大小、轻重、形状、体积、目的等方面的知识本体。

例如，在句子 I saw a star with a telescope 中，with a telescope 可能做 saw 的状语，也可能做 star 的定语，根据知识本体，star 是不能带有 telescope 的，因此，with a telescope 做 star 的定语不符合常识，而 with telescope 做 saw 的状语在常识上却是行得通的，因此，这个句子的意思是"我用天文望远镜看一颗星星"，而不是"我看一颗带有天文望远镜的星星"。在句子 I saw a man with a binocular 中，根据知识本体，with a binocular 可以修饰 saw，做 saw 的状语，也可以修饰 man，做 man 的定语，因此，这个句子是有歧义的，它的意思可能是"我用望远镜看一个人"，也可能是"我看见一个带着望远镜的人"。

因此，我认为，作者抓住"知识本体"的构建这个关键问题来展开本书的研究，是非常有见地的。

知识本体究竟是什么？在这个问题上仁者见仁，智者见智。不同的学科从各自不同的角度研究知识本体。

语言在本质上是模糊的。语言需要借助共享知识来补全缺失的信息或校正歪曲的信

息。自然语言处理技术力图借助于人类世界知识中很小的一部分去理解语言中包含的简洁、模糊、含蓄的信息。这就是结构化的知识本体在自然语言处理中所要扮演的角色。

知识本体是语言概念知识系统的、科学的描述方法，它与自然语言的词汇系统有着非常密切的关系。

如果我们对于一个领域中的客体进行分析，找出这些客体之间的关系，获得这个领域中不同客体的集合，这一个集合就可以明确地、形式化地、可共享地描述这个领域中各个客体所代表的概念的体系，它实际上就是概念体系的规范，这样的概念体系规范就可以看成这个领域的知识本体。

人们很早就开始研究知识本体，因此，知识本体有很多不同的定义，这些定义有的是从哲学思辨出发的，有的是从知识的分类出发的，最近的一些定义则是从实用的计算机推理出发的。

牛津英语词典对知识本体的定义是“对存在的研究或科学”(the science or study of being)。这个定义显然是非常广泛的，因为它试图研究存在的一切事物，为存在的一切事物建立科学。不过，这个定义确实是关于知识本体的经典定义，它来自哲学研究。

什么是事物(things)？什么是本质(essence)？当事物发生改变时，本质是否仍然存在于事物之中？概念(concept)是否存在于我们的心智(mind)之外？怎样对世界上的实体(entities)进行分类？这些都是知识本体要回答的问题，所以，知识本体是“对存在(being)的研究或科学”。

远在古希腊时代，哲学家就试图研究当事物发生变化的时候，如何去发现事物的本质。例如，当植物的种子发育变成树的时候，种子不再是种子了，而开始成为了树，那么，树还包含着种子的本质吗？

巴门尼德(Parmenides)认为，事物的本质是独立于我们的感官的，种子在表面上虽然变成了树，但是，它的本质是没有改变的，所以，在实质上种子并没有转化为树，只不过是我们的感官原来感到它是种子，后来感到它是树。

亚里士多德(Aristoteles，公元前 384—前 322)认为，种子只不过是还没有完全长成的树，在发育过程中，树的本质并没有改变，只是改变了它存在的形式，从没有完全长成的树(潜在的树)变成了完全长成的树(实在的树)。种子和树的本质都是一样的。知识本体就要研究关于事物的本质的问题。

亚里士多德还把存在区分为不同的模式，建立了一个范畴系统(system of categories)，包含的范畴有 10 个：substance(实体)，quality(质量)，quantity(数量)，relation(关系)，place(空间)，time(时间)，attribute(属性)，state(状态)，action(行动)，passive action(承受)。

这就是著名的十大范畴系统，这个范畴系统是最早的概念体系，实际上也就是最早的知识本体。

亚里士多德以他卓越的学识和深刻的洞察力，抓住了人类认识中最关键的概念。

德国哲学家康德(Immanuel Kant，1724—1804)认为，事物的本质不仅仅由事物本身决定，也受到人们对事物的感知或理解的影响。

康德提出这样的问题：“我们的心智究竟采用什么样的结构来捕捉外在世界的呢?”为了回答这个问题，康德对范畴进行了分类，建立了康德的范畴框架，这个范畴框架包括 4 个大范畴：quantity(数量)，quality(质量)，relation(关系)，modality(模态)。每一个大范畴又分

为 3 个小范畴：quantity 又分为 unity（单量），plurality（多量），totality（总量）3 个范畴；quality 又分为 reality（实在质），negation（否定质），limitation（限度质）3 个范畴；relation 又分为 inherence（继承关系），causation（因果关系），community（交互关系）3 个范畴；modality 又分为 possibility（可能性），existence（现实性），necessity（必要性）3 个范畴。

根据这个范畴框架，我们的心智就可以给事物进行分类，从而获得对外界世界的认识。因此，康德的范畴框架是帮助我们捕捉外在世界的有力手段。在数据库中，我们可以根据康德的方法给事物建立一些范畴，从而根据这些范畴来管理数据。例如，我们给人事管理数据库建立“姓名，性别，籍贯，职业”等范畴，使用这些范畴进行人事管理。可以看出，康德对范畴框架的研究，为知识本体的研究奠定了坚实的基础。

层级结构被认为是人类知识认知模型的基本结构。

科林斯（Collins）和奎尼安（Quillian）提出：一切概念之间，都可以通过 is-a 关系相连并构成一个层级体系。例如，a terrier is a dog（“猎狐狗”is-a“狗”），a dog is a mammal（“狗”is-a“哺乳动物”），这样就可以解释为什么人们在判断 a terrier is a dog（“猎狐狗是一只狗”）和 a terrier is a mammal（“猎狐狗是一个哺乳动物”）这两句话是否正确时需要的时间是不同的。因为在这样的概念网络中，遍历更多结点会导致更大的处理量，也将相应导致更长的反应时间。

人工智能是计算机科学的一个重要领域。在人工智能研究中也经常使用带有特征链接的概念网络对知识进行结构化，其中概念结点的链接特征往往是部分、功能以及诸如颜色这样的物理特征。有了通过 is-a 关系进行继承的机制，这些网络就成为了简单的知识表征系统（knowledge representation system）。

人工智能研究中的语义网络现在已经渐渐演化成了更复杂的形式化知识表征方法。

知识表征形式体系对精确的语义表示来说是必要的，也使得根据已表达的知识进行推理成为可能。不过，如果只有形式体系而没有知识就毫无用处，而将知识形式化是一件复杂而又费力的工作。

在 20 世纪末和 21 世纪初，知识本体的研究开始成为计算机科学的一个重要领域。它的主要任务是研究世界上的各种事物（例如物理客体、事件等）以及代表这些事物的范畴（例如概念、特征等）的形式特性和分类。

计算机科学对于知识本体的研究当然是建立在上述经典的知识本体研究的基础之上的，不过，有了很大的发展。因此，在计算机科学研究中，有必要重新给知识本体下定义。

1993 年，计算机科学家格鲁伯（Gruber）给知识本体下的定义是：“知识本体是概念体系的明确规范（An ontology is an explicit specification of conceptualization）。”这个定义比较具体，也比较便于操作，在知识本体的研究中广为传布。

1997 年，计算机科学家波尔斯特（Borst）对格鲁伯的定义做了很小的修改，提出了如下的定义：“知识本体是可以共享的概念体系的形式规范（Ontologies are defined as a formal specification of a shared conceptualization）。”

1998 年，计算机科学家施图德（Studer）等在格鲁伯和波尔斯特的定义的基础上，对于知识本体给出了一个更加明确的解释：“知识本体是对概念体系的明确的、形式化的、可共享的规范（An ontology is a formal explicit specification of a shared conceptualization）。”

在这个定义中，所谓“概念体系”是指所描述的客观世界的现象中有关概念的抽象模型；

所谓“明确”是指对所使用的概念的类型以及概念用法的约束都明确地加以定义；所谓“形式化”是指这个知识本体应该是机器可读的；所谓“共享”是指知识本体中所描述的知识不是个人专有的，而是集体共有的。

具体地说，如果我们把每一个知识领域抽象成一个概念体系，再采用一个词表来表示这个概念体系，在这个词表中，要明确地描述词的涵义、词与词之间的关系，并在该领域的专家之间达成共识，使得大家能够共享这个词表，那么，这个词表就构成了该领域的一个知识本体。

知识本体已经成为了提取、理解和处理领域知识的工具，它可以被应用于任何具体的学科和专业领域，知识本体经过严格的形式化之后，借助于计算机强大的处理能力，可以对人类的全部知识进行整理和组织，使之成为一个有序的知识网络。

互联网(Web)的概念最早是蒂姆·伯讷斯-李(Tim Berners-Lee)于 1989 年提出的。蒂姆·伯讷斯-李因此而成为了互联网之父。2001 年，他进一步提出了互联网的体系结构，这样的互联网叫作语义互联网(Semantic Web)，这样一来，互联网的发展就与语义的关系越来越密切了。这就是互联网的“语义化”。

在语义互联网的整个体系结构中，本体词汇(Ontology Vocabulary)起着承上启下的联系作用，处于举足轻重的地位。采用本体词汇来描述语义互联网中各种资源之间的联系，可以克服目前互联网上的信息格式的异构性、信息语义的多重性以及信息关系的匮乏和非统一性等难题。

2006 年 5 月，蒂姆·伯讷斯-李又宣布，经过十年的努力，W3C 已发布 W3C 推荐标准 80 余份，语义互联网已经具备了为达到成功的目标所需要的所有标准和技术，包括作为数据语言的 RDF、本体语言、查询和规则语言。

在这个体系结构中，Ontology Vocabulary 发展成了 Ontology OWL，其中，OWL (Ontology Web Language)可以翻译成“本体网络语言”，是 W3C 开发的一种描述本体的网络语言，用于对本体进行语义描述。W3C 的设计人员针对各类特征的需求制订了三种相应的 OWL 的子语言，即 OWL Lite、OWL DL 和 OWL Full，各子语言的表达能力递增。

OWL Lite 是表达能力最弱的子语言。它是 OWL DL 的一个子集，但是通过降低 OWL DL 中的公理约束，保证了迅速高效的推理。它支持基数约束，但基数值只能为 0 或 1。因为 OWL Lite 表达能力较弱，所以为它开发支持工具要比其他两个子语言容易一些。OWL Lite 用于提供给那些仅需要一个分类层次和简单约束的用户。

OWL DL(Description Logic，描述逻辑)将可判定的推理能力和较强的表达能力作为首要目标。OWL DL 包括了 OWL 语言的所有语言成分，但使用时必须符合一定的约束，受到一定的限制。OWL DL 提供了描述逻辑的推理功能，描述逻辑是 OWL 的形式化基础。

OWL Full 包含 OWL 的全部语言成分并取消了 OWL DL 中的限制，在 OWL Full 中，一个类可以看成是个体的集合，也可以看成是一个个体。由于 OWL Full 取消了基数限制中对可传递性质的约束，因此不能保证可判定推理。

显而易见，在 2006 年公布的语义互联网体系结构中，知识本体的重要性更加突出了。

经过 XML 标注的自然语言文本是语义互联网的基础，互联网主要是由语言文字组成的(此外还有语音、音乐和图像)，经过 XML 标注了句法、语义等信息之后的“文本文档”(text file)，再经过资源描述框架(Resources Description Format，RDF)的处理、知识本体处

理、规则处理、统一逻辑处理之后，就成为了“智能文档”(intelligent file)。

智能文档是“知道”自己内容的文档，其目的是让自动化程序“知道可以用它来做什么”。这些自动化程序叫作“智能代理”(agent)，智能代理是实现语义互联网服务(Semantic Web Service，简称 SWS)的重要构件。

语义互联网用知识本体(ontology)来表示概念以及概念之间的关系，所以文档的语义信息标注实际上是一种建立在知识本体基础之上的标注。

由此可见，知识本体在语义互联网的体系结构中不仅起着承上启下的作用，而且，它是整个语义互联网中概念体系的明确的、形式化的、可共享的规范，是整个语义互联网的语义的基础，语义互联网中的语义主要来自知识本体。

人工智能与计算机科学的知识本体偏重于范畴类别。这些范畴类别就是我们头脑中知识的组织方式。知识处理也正是在这些范畴类别的基础上进行的。

目前，在互联网上除了使用英语之外，越来越多地使用汉语、西班牙语、德语、法语、日语、韩语等语言。从 2000 年到 2010 年，互联网上使用英语的人数仅仅增加了 301.4%，而在此期间，互联网上使用俄语的人数增加了 1 825.8%，使用汉语的人数增加了1 476.7%，使用葡萄牙语的人数增加了 990.1%，使用法语的人数增加了 398.2%。

由于互联网上使用英语之外的其他语言的人数增加得越来越多，英语在互联网上独霸天下的局面已经被彻底打破，互联网确实已经变成了“多语言的网络世界”(multilingual Web)。“多语言”这个特性使得互联网变得丰富多彩，同时也造成了不同语言之间交流和沟通的困难，互联网上的语言障碍问题显得越来越突出，越来越严重，因此，互联网上不同自然语言之间的跨语言信息检索(Cross-Language Information Retrieval)或机器翻译(Machine Translation)也就变得越来越迫切了。这是互联网的“多语化”。

随着互联网信息资源的“语义化”与“多语化”这两个明显的趋势，迫切需要能适用于信息的语义处理与多语言处理的相关方法与技术，而多语言知识本体正是解决这个问题的有效基础资源之一。

目前的多语言知识本体的构建仍然主要依靠大量人工参与，并且是手工构建，在构建周期和成本上难以满足实际需求。因此，自动或者半自动地构建语言知识本体已是当务之急。多语言知识本体机器学习(Machine Learning)的目标就是从文本集合中自动或半自动地生成知识本体，具体任务包括从文本集合中抽取相关领域术语和同义词，发现概念、概念间层次结构与非层次结构的机器学习，本体实例生产等。利用知识本体的机器学习技术，结合多语言信息处理技术，就可以半自动或者自动地生成多语言知识本体，将其用于跨语言信息检索或机器翻译等多语言信息服务领域。以知识本体自动获取为目的的知识本体机器学习技术不但成为当前知识本体工程领域研究的热点，并且是知识本体应用与推广过程中迫切需要解决的关键问题。这是一项具有挑战性的工作。

章成志博士的《多语言领域本体学习研究》对于这一项具有挑战性的工作进行了卓有成效的研究。

本书介绍了当前国内外关于多语言知识本体的机器学习方法、工具以及应用项目等的相关动态，重点描述作者在该领域的研究情况，侧重于自动构建专业领域本体，以真实的科技领域语料为基础，在此基础上进行专业领域的概念抽取与层次关系构建。

第 1 章“引言”主要介绍课题的研究背景与研究意义，说明本书提出的多语言知识本体

学习框架与开发平台设计。

第 2 章“多语言领域本体学习研究综述”对多语言词汇资源进行概述，从知识本体学习、典型的知识本体学习工具、多语言知识本体应用项目等多个角度对多语言知识本体学习的现状进行了说明。

第 3 章“基于领域平行语料的双语核心术语抽取研究”将专业领域文档的关键词作为候选核心术语，利用中文和英文的专业领域分类语料，通过关键词抽取、术语度计算等关键技术，分别进行中文和英文的核心术语的识别。

第 4 章“基于多层术语度的一体化术语抽取研究”将语言学方法与统计方法并行融合，综合考虑候选术语及其所在语句的术语度，进行基于多层术语度的一体化术语抽取。

第 5 章“基于术语度约束的双语术语对齐研究”依据术语度对常规的词语对齐结果进行了优化，在双语术语对齐中利用术语度进行对齐约束，从而改善了术语对齐的性能。

第 6 章“基于多语术语聚类的概念层次体系生成研究”以近邻传播（Affinity Propagation）聚类算法为基础，从跨语言的术语聚类和中间语言的术语聚类两个角度，对多语言概念层次聚类问题进行研究，并对四个领域类别的数据进行测试。

第 7 章“基于多语文本聚类的主题层次体系生成研究”对基于多语言文本聚类的主题层次体系生成相关研究进行了概述，以近邻传播聚类算法为基础，对特征重构后的文本进行聚类，得到主题式层次体系，依据机器翻译系统或者平行语料，对该层次体系进行双语表示，间接地生成多语言的主题层次体系。

第 8 章“结束语”对本书进行总结，展望了双语术语抽取与多语言层次关系构建的研究前景。

本书自始至终都特别关注当前互联网发展中信息资源“语义化”和“多语化”这两个明显的趋势，围绕“语义化”和“多语化”进行了卓有成效的研究。

因此，我认为，本书是采用机器学习方法来构建多语言领域知识本体的一本具有开创性的优秀著作。由于多语言领域知识本体在跨语言信息检索、机器翻译等多语言科技信息服务中具有重要作用，本书的出版必将推动我国多语言科技信息服务的进一步发展。

本书把知识本体的研究局限在专业术语的领域内，由于专业术语是针对具体领域的，一般都具有单义性的特点，这就大大降低知识本体自动构建的难度。

但是，如果我们把知识本体的研究进一步扩展到日常词语的领域，问题就会变得非常复杂。

每种语言的词汇都是不同的。对于说英语的人来说，nail（指甲）都是一样的，不同的手指和脚趾各有名称，但上面长的指甲只有一个名称——nail。在某些语言中，甚至手指和脚趾都是同一个名称，用同一个词来表示。例如，西班牙语中的 dedo（指头）以及意大利语中的 dito（指头）。这种错综复杂的对应关系，将会增加多语言知识本体构建的难度。

在构建多语言的词汇知识库的时候，双语歧义是源语言（source language）和目标语言（target language）之间彼此对应时出现的歧义（ambiguity）。

例如，在英语中，river（河流）没有进一步的区分，而在法语中则进一步区分为 rivière（河）或 fleuve（江），在德语中进一步区分为 Fluss（河流）或 Strom（激流）；在英语中，eat（吃）没有进一步区分，而在德语中则进一步区分为 essen（[人]吃）或 fressen（[动物]吃）；在英语中，wall（墙）没有进一步区分，而在法语中则进一步区分为 mur（墙）或 paroi（隔墙），在德语

中则进一步区分为 Wand(墙),Mauer(围墙)或 Wall(土墙);在英语中,blue(蓝色的)没有进一步区分,在俄语中,则进一步区分为 синий(深蓝色的)或 голубой(浅蓝色的)。

有时,这种双语歧义使得词义之间的对应关系变得非常复杂。例如,英语中的单词 leg(腿),foot(足),paw(爪子)与法语中的单词 jambe(腿),pied(脚),patte(爪子),etape(宿营地)之间存在着交叉对应关系:法语的 pied 可以用于指人的"脚",这时,它与英语的 foot 相对应;法语的 pied 也可以用于指椅子的"脚",这时,它与英语的 leg 相对应;而英语的 foot 还可以指鸟的"爪子",这时,它与法语的 patte 相对应。英语的 leg 含义复杂,它除了与法语的 pied 对应之外,还可以指动物的"脚",这时,它和 foot 一起,又与法语的 patte 相对应;英语的 leg 还可以指人类的"腿",这时,它与法语的 jambe 相对应;此外,英语的 leg 还可以指旅行中的一段"旅程",这时,它与法语的 etape 相对应。英语和法语的含义之间形成的这种交叉对应关系也是非常复杂的。

由于存在这种极为复杂的双语歧义现象,给多语言词汇网络的构建造成了极大的困难,在跨语言信息检索或机器翻译中,不同语言之间的单词就会出现一对多、多对一或多对多的复杂情况,需要进行词义排歧(Word Sense Disambiguation,简称 WSD)。这必将给日常领域的多语言知识本体的构建造成极大的困难。

本书的研究还没有涉及日常领域知识本体构建中这些错综复杂的词义排歧问题。希望章成志博士在今后的研究中,能够深入地研究这些问题,把本书所使用的多语言领域知识本体的自然语言处理与机器学习技术进一步扩充应用到多语言日常领域知识本体的自动构建中去,不断地探求新知,不懈地锐意进取,取得更大的成绩。

冯志伟

2012 年 4 月 29 日

于北京东城后拐棒胡同

目　录

第1章 引 言

1.1 研究背景

当前互联网的发展有两种特别明显的趋势，一是互联网信息资源的语义化，二是互联网用户与信息资源的多语言化。

一方面，随着互联网的快速普及，特别是语义网的提出，本体技术越来越被人们所重视。本体技术已经用于信息检索、数字图书馆、信息集成、知识服务等多个应用研究领域。[1-4]早期与本体相关的研究领域主要有以下三个，即：人工智能领域，主要研究知识的获取，如语义网络、故事理解等；自然语言处理领域，主要进行词汇知识的获取研究，如从机读词典、语料库中获取词语的语义知识；情报检索领域，主要进行叙词表的构建研究，如进行主题词抽取、主题词表的机器辅助构建研究。

语义网能否成功主要取决于能否大规模地生成高质量的领域本体[5]。2007 年 5 月的一篇 Gartner 报告(见图1－1)表明，在未来 10 年内，基于 Web 技术，语义网将会提高往文档中嵌入语义结构、创建结构化词汇表和本体(用于定义术语、概念及关系)的能力[6]。其中需要解决的关键问题是本体的获取(或称为本体提取、本体构建、本体生成)问题。

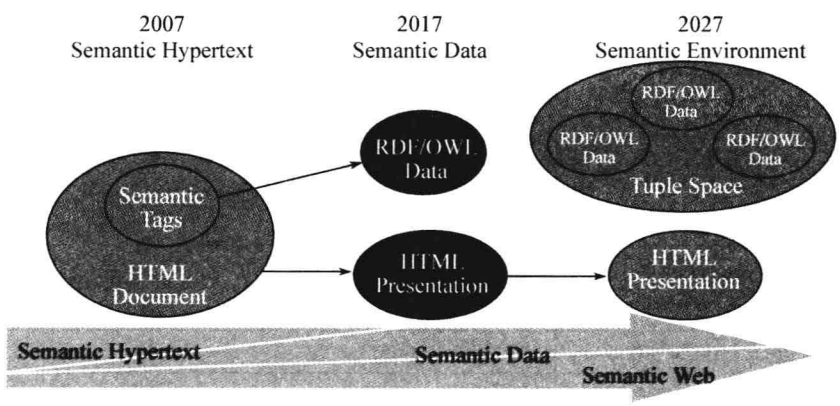

图 1－1 语义技术发展趋势图[6]

当前本体应用方面出现的问题是知识获取的瓶颈问题，即本体的构造方法大多依靠的是人工方法，费时费力，跟不上应用发展的需要。手工构建本体，特别是大规模的领域本体，需要大量的领域专家和知识工程师的参与，耗费大量人力和财力；同时，要保证所构建数据的完整性和一致性，还必须依赖领域专业人士。随着信息技术的发展，各领域的发展速度越来越快，即便耗费巨大，最后得到的本体却可能是过时的信息。当前被本体开发人员使用的

一些本体编辑工具，如 Protégé[①]、OntoEdit[②]、WebOnto[③] 等，为本体构建提供了可视化操作界面，这在一定程度上加速了本体的开发进程。与此同时，在语义服务应用驱动下，本体需求更加迫切。领域专家和知识工程师仅仅利用这些本体编辑工具来生成本体，已经显得力不从心，特别是在某些领域或者应用场合，还没有可以作为参照的本体。此时，人们单纯利用本体编辑工具生产本体的能力变得非常有限。

另一方面，互联网的发展使得人们的交流摆脱了地域的限制，但随着网络上各种语言的信息资源日益增加，人们越来越多地面临如何利用多语种信息的问题。国际互联网数据统计机构(Internet World Stats)的数据表明：互联网上使用的十大语种分别为英语、汉语、西班牙语、日语、法语、葡萄牙语、德语、阿拉伯语、俄语、韩语，这十大语种用户总数占互联网用户总数的 83.3%；如表 1-1 所示，2000 年—2009 年之间，互联网语种使用增长最快的三种语言分别是阿拉伯语、俄语以及汉语，增长速度分别高达 1 907.9%、1 359.7% 和 1 087.7%[④]。随着互联网搜索引擎的快速发展，现在人们可以比较容易地在线获取不同语言文本信息资源，例如人们可以通过“Google News”[⑤]比较容易地获取上述十大语种描述的即时新闻文本。随着全球一体化的日益加剧与多语言信息资源的快速增长，多语言信息资源的获取与利用已成为实现全球知识存取和共享的关键技术。

表 1-1　Web 上使用的十大语种(按语种分类的互联网用户数)[7]

Internet 上使用的语种排名	用户数	普及率	增长速率(2000—2009)	用户数(占总用户数的比率)	可能的增长数(2009 年估计)
英语	478 442 379	37.9%	237.0%	27.6%	1 263 830 976
汉语	383 650 713	27.9%	1 087.7%	22.1%	1 373 859 774
西班牙语	136 524 063	33.2%	650.9%	7.9%	411 631 985
日语	95 979 000	75.5%	103.9%	5.5%	127 078 679
法语	78 972 116	18.6%	547.4%	4.6%	425 622 855
葡萄牙语	73 052 600	29.5%	864.3%	4.2%	247 223 493
德语	64 593 535	67.0%	133.2%	3.7%	96 389 702
阿拉伯语	50 422 300	17.3%	1 907.9%	2.9%	291 798 743
俄语	45 250 000	32.3%	1 359.7%	2.6%	140 041 247
韩语	37 475 800	52.7%	96.8%	2.2%	71 174 317
前十大语种总数	1 444 362 506	32.5%	363.5%	83.3%	4 448 651 771
其他语种	289 631 235	12.5%	487.1%	16.7%	2 319 153 437
总数	1 733 993 741	25.6%	380.3%	100.0%	6 767 805 208

由上可知，随着互联网信息资源的语义化与多语言化趋势，迫切需要能适用于信息的语

① http://protege.stanford.edu

② http://www.ontoknowledge.org/tools/ontoedit.shtml

③ http://kmi.open.ac.uk/projects/webonto/

④ http://www.internetworldstats.com/stats7.htm

⑤ http://news.google.com

义处理与多语言处理的相关方法与技术。多语言本体正是解决该问题的有效基础资源之一。

目前的多语言本体仍然主要依靠大量人工参与，并且是手工构建[8]，在构建周期和成本上不能满足实际需求。因此，自动或者半自动地构造多语言本体已是当务之急。多语言本体学习的目标是以现有的多语言资源为基础，自动化或半自动化地构建多语言本体。

本体学习(Ontology Learning)的目标就是从文本集合中自动或半自动地生成本体，具体任务包括从文本集合中抽取相关领域术语和同义词、发现概念、概念间层次结构与非层次结构学习、本体实例生产等[9]。利用本体学习技术，结合多语言信息处理技术，可以半自动或者自动化地生成多语言本体，将其用于跨语言信息检索(Cross Language Information Retrieval)[10,11]或机器翻译[12]等多语言信息服务领域[13]。以本体自动获取为目的的本体学习技术不但成为当前本体工程领域的研究热点，并且是本体应用与推广过程中迫切需要解决的关键问题。

机器学习(Machine Learning)是用数据或以往的经验优化计算机程序的性能标准[14]。本体学习技术就是利用机器学习技术自动或者半自动地从数据源中获得(抽取、生成、构建)本体。按照本体学习数据源的结构化程度不同，可以将本体学习分为三类：基于结构化数据的本体学习、基于非结构化数据的本体学习(主要为文本)以及基于半结构化数据的本体学习[15]。这三类本体学习技术中，最重要的方法是基于文本的本体学习，它以自然语言处理和机器学习为基础，从自然语言文本中获取本体。

从应用层面来看，本体学习还存在如下两方面问题需要解决。

一方面，当前本体学习系统多为原型系统，缺乏实用性，其主要的原因为对本体学习的每个环节，特别是概念的抽取与概念关系的发现，学习的精度还不够。本体学习的研究处于起步阶段，中文本体学习工具尚未见诸报道[15]。

另一方面，以 EuroWordNet①、Global WordNet② 为代表的通用多语言本体(主要依靠手工方法构建)已相对实用，但在很多的多语言信息服务中，人们更加需要的是领域多语本体，试图依靠同样的手工方法来构建领域多语本体，在构建周期和成本上是不能满足实际需求的。因此，如何利用现有的多语资源，采用自然语言处理和机器学习技术来自动构建多语领域本体是实现多语言信息服务的重要课题。该技术不但可以实现以概念为中心的跨语言信息检索和机器翻译，而且可以通过多语领域本体进行多语资源的深层标注，实现多语信息共享[16]。

因此，如何充分借鉴本体学习中的关键技术和方法，利用大规模的多语言文本资源和已有词典资源，采用自然语言处理技术与多语言信息处理技术，半自动化或者自动化地生成用于跨语言信息检索、多语言文本挖掘等多语言信息处理任务的多语言领域本体，将是一项具有重要意义和极富挑战性的工作。

综上，本课题将围绕多语领域本体学习中涉及的两个关键问题进行研究，即：双语术语抽取技术研究、多语言本体中概念层次体系自动构建研究。

① http://www.illc.uva.nl/EuroWordNet/

② http://www.globalwordnet.org/

1.2 研究内容概况

1.2.1 课题概述

本课题进行多语言领域本体学习，主要针对数字图书馆环境下跨语言信息检索和在线机器辅助翻译任务。因此，该多语言本体是一种任务本体。另外，我们使用的是专业领域语料学习生成领域本体，因此该本体又是一种领域本体。所以，本课题学习生成的多语言领域本体是一种针对特定领域和任务的应用本体。

1.2.2 多语言本体学习框架与开发平台的设计

本研究的主要研究思路为：基于多语文本集，以机器学习和自然语言处理技术为辅助技术半自动构建多语领域本体。本课题侧重于从文本集中进行面向跨语言信息检索应用的多语本体（或称之为跨语言应用本体）学习研究。

本研究按照图 1-2 所示的多语领域本体学习逻辑框架图进行，主要进行双语术语抽取与概念层次体系构建两个方面的研究。该研究以真实的科技领域语料为基础，进行双语术语抽取与概念层次体系的构建。

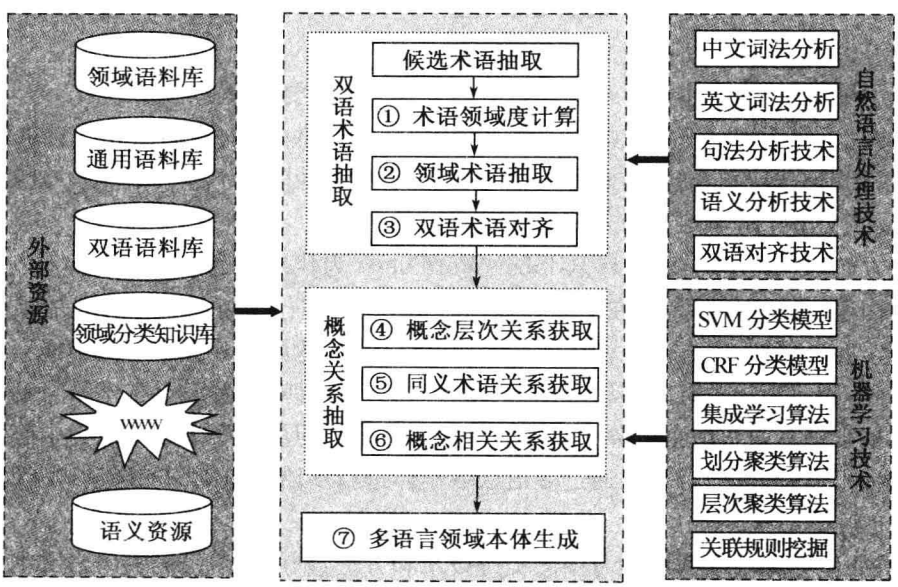

图 1-2 多语领域本体学习逻辑框架图

本课题以工程技术领域的大规模双语对齐语料为基础，分别从不同语种语料中并行获取术语的上下文信息，借助于基本的双语词典，进行双语对照的术语提取，双语对照的概念层次关系提取。本课题以中英文双语对齐语料为主要实验对象，进行中英双语本体学习研究。

图 1-3 所给出的是英汉双语本体学习的流程示意图。