计算语言学
与语言科技
原文丛书

CAMBRIDGE

ONTOLOGY AND THE LEXICON
A Natural Language Processing Perspective

本体与词汇库
自然语言处理角度的解析

Chu-Ren Huang [意] Nicoletta Calzolari [意] Aldo Gangemi
[意] Alessandro Lenci [意] Alessandro Oltramari [法] Laurent Prévot 编
陆 勤 导读

北京大学出版社
PEKING UNIVERSITY PRESS

ONTOLOGY AND THE LEXICON

A Natural Language Processing Perspective

本体与词汇库

——自然语言处理角度的解析

Chu-Ren Huang
[意] Aldo Gangemi
[意] Alessandro Oltramari
[意] Nicoletta Calzolari
[意] Alessandro Lenci 编
[法] Laurent Prévot
陆 勤 导读

北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记号　图字:01-2014-1364
图书在版编目(CIP)数据

本体与词汇库：自然语言处理角度的解析 =Ontology and the lexicon:a natural language processing perspective：英文 / 黄居仁(Chu-Ren Huang)等编. — 北京：北京大学出版社, 2014.12
（计算语言学与语言科技原文丛书）
ISBN 978-7-301-24954-3

Ⅰ.①本…　Ⅱ.①黄…　Ⅲ.①自然语言处理—研究—英文　Ⅳ.①TP391

中国版本图书馆CIP数据核字(2014)第233544号

书　　名：本体与词汇库——自然语言处理角度的解析
著作责任者：Chu-Ren Huang等　编
责任编辑：李　凌
标准书号：ISBN 978-7-301-24954-3/H·3601
出版发行：北京大学出版社
地　　址：北京市海淀区成府路205号　100871
网　　址：http://www.pup.cn　　新浪官方微博：@北京大学出版社
电子信箱：zpup@pup.pku.edu.cn
电　　话：邮购部 62752015　发行部 62750672　编辑部 62753374　出版部 62754962
印　刷　者：北京大学印刷厂
经　销　者：新华书店
787毫米×1092毫米　16开本　24印张　435千字
2014年12月第1版　2014年12月第1次印刷
定　　价：61.00元

“计算语言学与语言科技原文丛书”由北京大学—香港理工大学汉语语言学研究中心、北京大学计算语言学研究所(由973课题“文本内容理解的数据基础”、863课题“大规模汉语语义基础资源库和知识库设计构建及工具平台”支持)和北京大学出版社合作推出

学术委员会
Academic Advisory Committee

编委会
Editorial Committee

丛书前言

“计算语言学与语言科技原文丛书”于2010年创立，2010 COLING国际计算语言学会议在北京举办之前出版了第一批图书。本丛书的出版象征着国内计算语言学研究与国际的接轨。国内学者正跻身于计算语言学的国际舞台：一些资深学者已在COLING两个最主要的国际会议/组织中获选并担任重要的领导职务；而积极参与这些重要的国际会议也已在年轻学者中蔚然成风，他们已可谓是会议的主流参与者之一。在这样的氛围中，希望本丛书第二批图书的出版，能让国内有心投入语言科技与计算语言学研究的学者们如虎添翼，在国际舞台上创新并引导议题！

计算语言学（Computational Linguistics，CL）在语言科学与信息科学的研究中扮演着关键性的角色。语言学理论寻求对语言现象进行规律性的预测，做出完整的解释，计算语言学正好为这两点提供了验证与应用的大好机会。作为语言学、信息科学乃至于心理学与认知科学结合的交叉学科，计算语言学更为语言学基础研究与福国利民应用研究的接轨提供了绝佳界面。事实上，计算语言学与人类语言科技（Human Language Technology，HLT）可以视为体用两面，不可切分。

计算语言学研究的滥觞，其实源于上世纪五六十年代的机器翻译研究，中文计算语言学的研究也几乎同步开始。在美国伯克利加州大学研究室，王士元、邹嘉彦、C.Y. Dougherty等人1960年已开始研究中英、中俄机器翻译。他们的研究是与世界最尖端的科技同步的。国内中俄翻译研究也不遑多让，大约在20世纪50年代中期便已开始。可惜的是，这些中文方面早期机器翻译研究，由于硬件与软件的限制，未能有效传承下来。中文计算语言学研究比较系统的发展始于1986年。这一年，海峡两岸不约而同地分别成立了两个致力于建立中文计算语言学基础架构的研究群：北京大学的计算语言学研究所，在朱德熙先生倡议下成立，随后一段时间由陆俭明、俞士汶主持；而台北“中研院”的中文词知识库小组，由谢清俊创立，陈克健主持，黄居仁1987年回去后加入。

中文计算语言学的研究，近30年来已累积了相当可观的成绩。计算语言学的重要研究领域与议题中都能看到中文方面的相关研究成果，华人计算语言学学者也渐渐在国际学术界崭露头角。随着世界经济转向知识密集型

产业，跨语言跨文化沟通与知识整合是知识产业的关键瓶颈，语言科技的发展成为国际主流指日可待。在这个有利发展的大环境下，我们期待看到，中文计算语言学与华人计算语言学学者的成绩，百尺竿头更进一步，中文方面的研究可以进入计算语言学的学术核心，能够产生有能力引导议题并掌控研究方向的大师。

回顾国内的计算语言学发展，计算机科学的贡献多于语言学的贡献。这个现象，在理论与概率模型整合研究的趋势中，不免令人忧心。语言学的贡献弱，或许可以部分归咎于英文研究专著在国内不易取得；而比较容易取得的期刊或会议论文，在篇幅的限制下，又往往无法对理论做深入完整的铺陈，从而导致国内的年轻学者长于运算而拙于理据。因此，在期待大师与引领世界研究潮流两个方向，藉由英文专书来巩固研究理据，进而开拓研究视野，是非常重要的一步。

“计算语言学与语言科技原文丛书”的引进，就是在上述背景下促成的。个人忝为剑桥大学出版社“自然语言处理研究”（Studies in Natural Language Processing，SNLP）系列的主编，对于将此系列中较重要的几本书引入国内，责无旁贷。第二批出版的原文书，除了剑桥大学出版社的图书外，还有施普林格出版公司（Springer）语言科技系列中的几本书，以进一步拓展领域涵盖面。引进原书，原样出版，是容易的，然而若要真正搭建知识的桥梁，使国内学者与学生不仅能开拓研究视野，更能将原文著作的理论精髓应用于中文研究，则实在不易。因此，本系列除了原书出版外，每本书我们都邀请了一位专家撰写中文导读。这些导读可以说是本系列的精华、重点，使每本书比剑桥和施普林格的原本增加了不少附加价值。

每篇中文导读都包括三个重要的组成部分。第一部分是全书内容概要的介绍。导读专家都是长年浸淫于该领域的学者，他们能提纲挈领，并提供相关研究背景。因此，通过阅读导读，读者更易掌握并吸收该书的重要内容。第二部分是中文相关研究。原文著作不见得会提到相关的中文研究，由导读专家补充介绍，搭起理论与中文相关应用的桥梁，更能引导读者找到在这个议题进入中文研究的最佳切入点，让中文相关研究的开拓者的成绩更能发扬光大。第三部分重点在于补充原书出版后该领域研究的新发展。现代科技发展迅速，任何经典著作出版后，几乎马上就有新的相关研究。因此，在理论架构的脉络中，加上新近发展，能使读者更贴切地掌握研究脉动。全书的内容摘要通常以文字叙述，而中文相关研究及最新研究发展则分别以文字叙述和延伸阅读书目的方式呈现。延伸阅读书目，可以使读者很快上手，进入相关研究领域，也是本系列的重要设计之一。

丛书2010年出版第一批图书，现在出版第二批图书，必须感谢许多同行的付出。在规划出版的漫长过程中，北大计算语言学研究所的俞士汶老师及常宝宝老师一直无私无悔地支持。而香港理工大学的挹注，北大－理大汉语语言学研究中心石定栩、郭锐等几位的支持，使得整个系列能够顺利出版。此外，还要感谢北大出版社王飙主任、杜若明编审及李凌编辑，他们认同我们的宗旨，落实了丛书的出版工作。最后，感谢丛书的国内编委，特别是此次担任导读主笔的各位，正是他们脑力与心血的付出，才替读者们搭建了进入学术殿堂的台阶。

丛书主编　黄居仁

谨志于香港，红磡

二零一四年九月

导　读

陆　勤

1　学科背景介绍

本体(ontology)是对概念化知识的表述。概念化说明的是概念知识形成的过程,主要是人类通过经历和体验得出的一种总结和综合性的知识。这些总结和综合的知识是一种经过思维上的体会和领悟所形成的抽象化表述(encoding)。从哲学的意义上说,所形成的概念可以认为是相对独立于某个特定语言的,也就是说领悟的过程是人类思维的共同行为,并不属于某个特定语言。对本体的研究,主要是如何建立本体的知识体系。在计算机时代,特别要求对概念知识的表述能够具有可计算性,也就是说需要研究用形式化的方式表述不同的概念,其中包括概念的基本元素,用于表述和区分不同的概念,以及表明概念与概念之间的关系。

词汇是语言表达体验时形成的抽象化表述。作为语言的表达,词汇依附于一个特定的语言。因为语言是用来表述、传达信息的,因此可以用来表述和解释概念。实际上词汇化的概念就是某个概念已经完成了被赋予一个有共识的特定词(或术语)的过程。这个过程完成的时间因语言而异,因此,同样的概念在一个语言中如果已完成了词汇化的过程,就可以直接用该词来表达;如果没有完成这个过程,则需要用其他已有词汇通过语言描述来表达。

计算机在进行自然语言文字处理时所需的技术统称为计算语言学(computational linguistics)。在这个过程中,都会依赖不同的语言资源,包括词汇库[①]。除了一般的文字处理功能之外,在很多应用上都要求对文字所表达的概念能够理解。因此了解词汇与概念,以及词汇与概念体系的关系至关重要。这样,计算机才可能从事智能化的运算和进行关系的推理,从而获取有用的知识——也就是通常所说的知识发现(knowledge discovery)。合理和有机地设计两个领域的资源并使其用计算机可理解的方式合理衔接,在语义

① 这里的词汇库指的是计算机使用的词汇表,其内容可能包括一些可操作的规则,比如词形变化规则。英文中称为computational lexicon。

网(semantic web)的大环境下,关系到知识的分享性、信息整合性、系统的互操作性,以及知识的充足性。

内容提要

在计算机时代,寻找概念知识和自然语言之间的映射已成为知识发现和计算语言学这两个领域共同关注的问题。能够搭建出这两个领域的桥梁,使得通过计算语言学来进行知识获取成为可能,而本体和词汇之间的界面恰恰又是搭建这座桥梁的最重要部分。本体的获取,一方面要达到从自然语言到形式化的知识体系的转换;另一方面,自然语言处理又可以当成系统性知识体系的一种应用。因为一个完善的本体建构可以具有其不依赖于特定语言的相对独立性,并能够表述人类思维层面的综合知识,从而在不同的计算机应用中可以得到有效使用。但是,本体的建构不可能独立于各种词汇相关的资源。实际上,本体的建构在很大程度上取决于自然语言的各种资源,特别是词汇资源的合理和有效的使用。如何有效的建立本体和词汇资源的界面正是本书描述的重点。

本书的内容主要是来自 Ontolex 专题研讨会(workshop)系列的一些重要研究工作。Ontolex 专题研讨会 2000 年始创。本书的编辑主要从 2002—2006 的四次专题研讨会的作者中邀稿,另外还邀请本领域专家撰写了几个章节,并经作者之间相互审稿,编委严格整合和审议,合辑成书。值得一提的是本书为全面论述本体和词汇库以及两者界面建构的第一本专著,内容涵盖理论和实践两个方面的研究成果。

本书分为四大部分。第一部分从基础知识出发,共有五个章节。作为基础介绍,第一章先给出本体与词汇的基本定义,并从理论角度来解释两者之间的界面和交互关系;然后在第二章和第三章介绍两个主要的形式化本体的系统——SUMO 和 DOLCE,并在第二章和第四章给出两个主要的词汇资源库的关系,其中一个是 WordNet, 另一个 FrameNet;第五章作为总结提出本体与词汇交界研究的路线图。

第二部分主要介绍概念体系的建构,共有四个章节,主要集中于本体系统的建构,旨在表明不同的本体建构方法和不同的本体知识。第六章介绍如何使用形式概念分析(Formal Concept Analysis)方法,通过文本资源构件本体。第七章讲解如何使用语义知识,包括本体、词汇以及事件知识,来追踪和识别事件的变化。第八章介绍一个汉字意符的本体知识建构,用于中文书写系统的形式化描述方法以及称为 Hantology 的汉字意符的本体知识体系系统。第九章提出一个基于认知语言学可用于概念体系建构的元模型

(metamodeling)[①]。

第三部分主要介绍本体和词汇资源界面的建构,共有四个章节。第十章主要讲解与综述本体和词汇资源交互使用时所面对的理论和实际的问题。第十一章介绍台北"中研院"开发的中英双语的本体词汇知识库 SINICA-BOW。第十二章指出用传统方法建构语义词库的困难,并提出在实际应用中,应采取先建立该应用领域的本体知识,之后将相关的词汇映射到本体的方法,从而避免建构冗余的语义消歧信息。第十三章讲解由于不同的语言学本体(linguistic ontology)在建构时会有不同的颗粒度,提出通过建构综合本体(global ontology)的方法,来避免交互操作时所要面对的颗粒度整合问题。

第四部分集中讨论自然语言处理应用的相关问题,共有四个章节。第十四章讲解本体知识的生成和知识在自然语言中的应用是知识生命周期的有机组合,互为依存。第十五章介绍 Omega,一个浅层的基于词汇的术语分类系统。第十六章介绍如何获取词汇语义信息来提升问答系统的性能。第十七章介绍一个基于泰文的农业相关本体的半自动建构。

2 内容详细介绍

第一部分 基础知识

第一章题为《本体与词汇——多领域研究的综述》(Ontology and lexicon: a multidisciplinary perspective),由本书的六位编者 Laurent Prévot、黄居仁、Nicoletta Calzolari、Aldo Gangemi、Alessandro Lenci 和 Alessandro Oltramari 撰写,主要从不同角度讲解本体建构和词汇资源之间关系。第一部分讲解的是两者在形式化方面的不同。本体是概念知识的形式化,其中有两个重要的过程:第一个称为概念化的过程(conceptualization);第二个称为形式化的过程(specification)。形式化是概念可以进行沟通的基础。形式化需要一种语言进行编码(encoding),而这种形式化的语言可以是任何符号化的语言。实际上用自然语言来描述的本体只能是非正式(informal)的[②]。正则的本体(formal ontology)必须用某种正式语言(formal language)。词汇,作为自然语言描述的一部分,其编码的形式是词(word)。概念和词之间的不同可以从两者所属关系的不同来区分。在词汇层次,我们有同义词的概念,但概念本身没

① 元模型广义上讲是指在建造一个特定模型时所使用的构建和规则模型,因此是关于模型的模型。

② 因为自然语言的歧义性。

有同义概念一说，虽然同一概念可以用多个同义的词来表达。需要指出的是，词义(word sense)是语言学领域的概念，而概念是本体的元素，两者所属的范畴不同。换句话说，词不是本体的内部元素，但是可以作为外部表征包括在本体结构之内，其有无并不影响一个概念的形成。显而易见，一个概念在形成之后，有无相应的概念词汇，将影响该概念使用在语言学上的表达方式。第二部分对本体和相关词汇资源进行了分类，而分类是基于它们的概念化过程、形式化过程和应用范围三方面进行的。第三部分讲解本体的概念关系和词汇之间关系的不同。比如在本体建构时，考虑的是种类关系(is-a-kind-of)、部分关系(part-of)和外延关系(instance-of)；而在词汇库的建构上，要考虑的是上下位关系(hypernym-hyponym)和同义反义关系(synonym-antonym)。第四部分讲解了本体和词汇界面的一些理论框架，包括感知科学方面所研究的分类，还有哲学方面所说的依据存在(meta-physical existence)的分类。本文主张应该放弃追求分类的所谓完美性，而更注重分类的方法(methodology)。这一主张对本体建构的影响至关重要。因为这一主张不鼓励大家花费时间探讨怎样把本体建构得最好、最合理，而应致力于如何使相对独立开发的不同本体系统具有互操作性。在词典创建方面要考虑到辞典的释义功能和百科全书的释事功能，这两种功能在现代语言学的理论框架下应该得以融合。而在语言学和词汇语义方面，最重要的研究成果是James Pustejovsky关于衍生词汇的理论(generative lexicon theory)，其中最重要的部分是提出词义穷举的必要。在此基础上每一个特定词义由其词义的不同属性动态决定，这些属性包括其分类信息、组成部分(constitutive)、功能(telic)等。

第二章题为《跨语言[①]的形式本体——SUMO[②]和WordNet相关联项目以及全球的WordNet》(Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet)，由Adam Pease和Christiane Fellbaum撰写，主要介绍英文语言资源库WordNet和上位本体SUMO，以及两者之间是如何关联映射的。SUMO于2001年建成，被称为上位本体是因为其所涵盖的概念(term)都是跨领域的通用概念，内有一千左右的概念条目[③]和四千左右相关的描述概念特征的公理(axioms)。之后SUMO进行了扩充，该扩充

① 英文的interlingua在这里不是特指国际语，而是指形式本体(formal ontology)作为概念体系，独立于某个自然语言，而有跨语言的表述概念的功能。

② Suggested Upper Merged Ontology。

③ 具体概念词(term)因为编辑上有更改，所以准确的数字没有太大意义，总数一直没有大的改变。