孙茂松　陈群秀　主编

自然语言理解与大规模内容计算

清华大学出版社

NATURAL LANGUAGE UNDERSTANDING AND LARGE-SCALE CONTENT COMPUTING

自然语言理解与大规模内容计算

孙茂松　陈群秀　主编

Sun Maosong Chen Qunxiu (Eds.)

清华大学出版社

北京

内容简介

本书是全国第八届计算语言学联合学术会议 JSCL-2005(2005 年 8 月 27 日—29 日,南京)的论文集。书中收录的 82 篇 regular 论文和 39 篇 poster 论文是从全国各地(含港、台地区)征集到的 185 篇论文中精选出来的。另有 4 篇论文是大会邀请报告,也收录于本书中。本书内容包括下列 6 类:(1)大会邀请报告;(2)词法、句法、语义和篇章分析;(3)资源建设及相关技术;(4)机器翻译技术、系统及评测方法;(5)智能检索(信息检索、信息抽取、文本挖掘、话题跟踪、文本分类、文本过滤、问答系统等);(6)其他(语音语料库、音字转换、文语转换、汉字规范、键盘输入、文本处理、网络非正规语言处理)。

本书充分展示了国内计算语言学研究与应用的最新进展,也展示了 21 世纪初中文信息处理和计算语言学研究的前沿和动向,对中文信息处理的基础研究和产品开发具有重要的参考价值。本书可供计算机、语言学等专业的科研人员、工程技术人员、大学老师和研究生选读。

书　　名:自然语言理解与大规模内容计算
作　　者:孙茂松　陈群秀　主编
出 版 者:清华大学出版社　　地　　址:北京清华大学学研大厦
http://www.tup.com.cn　　邮　　编:100084
社 总 机:010-62770175　　客户服务:010-62776969
责任编辑:赵彤伟
封面设计:傅瑞学
印 刷 者:北京市清华园胶印厂
装 订 者:三河市春园印刷有限公司
发 行 者:新华书店总店北京发行所
开　　本:185×260　印张:44　字数:1010.8 千字
版　　次:2005 年 7 月第 1 版　2005 年 7 月第1次印刷
书　　号:ISBN 7-302-06074-6
印　　数:1～600

全国第八届计算语言学联合学术会议(JSCL-2005)组织情况

日　　期：　2005年8月27日—29日
地　　点：　南京师范大学

发起单位：　中国中文信息学会
　　中国计算机学会
　　中国人工智能学会
　　北京市语言学会

赞助单位：　东芝(中国)研究开发中心
　　富士施乐有限公司
　　富士通研究开发中心有限公司
　　TRS信息技术有限公司
　　教育部语言信息管理司
　　中国中文信息学会

组织主办单位：　南京师范大学
　　清华大学智能技术与系统国家重点实验室

大 会 主 席：　李宇明
大会副主席：　董振东

程序委员会主席：孙茂松
程序委员会副主席：张　普　陈小荷
程序委员会委员：(以姓氏拼音字母次序为序)
　　蔡莲红　曹右琦　陈群秀　程学旗　董振东　冯志伟　傅爱平
　　何婷婷　黄河燕　李堂秋　刘开瑛　刘　挺　宋　柔　王厚峰
　　吴立德　荀恩东　姚天顺　俞士汶　苑春法　张　全　赵　军

组织委员会主席：陈小荷
组织委员会副主席：陈群秀　曹右琦
组织委员会委员：段业辉　李葆嘉　陈家骏　曲维光　冯敏萱　许　超

前　言

全国第八届计算语言学联合学术会议(JSCL-2005)，由中国中文信息学会、中国计算机学会、中国人工智能学会和北京市语言学会联合发起，于2005年8月27日—29日在南京师范大学召开。

本次会议共征集到论文185篇，经大会程序委员会严肃认真的评审，最终录用regular论文82篇，poster论文39篇。4篇大会邀请报告(全文或摘要)也收录到论文集中。本次会议的论文大体上可分为6类：

(1) 大会邀请报告，4篇；

(2) 词法、句法、语义和篇章分析，24篇(指regular论文，下同)；

(3) 资源建设及相关技术，12篇；

(4) 机器翻译技术、系统及评测方法，8篇；

(5) 智能检索(信息检索、信息抽取、文本挖掘、话题跟踪、文本分类、文本过滤、问答系统等)，30篇；

(6) 其他(语音语料库、音字转换、文语转换、汉字规范、键盘输入、文本处理、网络非正规语言处理)，8篇。

综而述之，本论文集具有以下特点：

- 大会邀请报告颇具启发性。《语意网与中文信息化的前瞻：知识本体与自然语言处理》指出了在未来网络(即所谓的Semantic Web)时代中文信息处理所面临的重大挑战，探讨了知识本体的基本框架及其构造原则与方法，讲演者与我们分享了宝贵的相关研究经验。《何谓金本位》以汉语自动分词为典型案例，对带标语料库中的一个重要问题进行了方法论高度上的思辩。《评述新闻报道或文章色彩——正负两极性自动分类的研究》阐述了中文信息处理的一个新的研究前沿。《搜索的未来》的讲演者来自企业界，相信这一特定的背景对计算语言学界的研究工作会有所补益。
- “词法、句法、语义和篇章分析”和“资源建设及相关技术”两大类别的论文占regular论文的44.0%，较上届(即2003年在哈尔滨召开的“全国第七届计算语言学联合学术会议”)的60.0%虽然有相当的下降，但仍然接近“半壁江山”。
- “机器翻译技术、系统及评测方法”类的论文占regular论文的9.7%，较上届的13.3%有所下降，似乎显示这个方向的研究工作的活跃程度呈下降趋势，或者说目前正处于一个相对平稳的盘整阶段。
- “智能检索(信息检索、信息抽取、文本挖掘、话题跟踪、文本分类、文本过滤、问答系统等)”类的论文占regular论文的36.6%，较上届的20.0%有大幅攀升。显然，此类研究的强劲上升势头还在持续中。
- “其他(语音语料库、音字转换、文语转换、汉字规范、键盘输入、文本处理、网络非正规语言处理)”类的论文仅占regular论文的9.7%，与上届的6.7%大致相同。

上届会议论文集《前言》指出，今后若干年计算语言学的主流研究将围绕“语言计算”与“基于内容的文本处理”这两大主题展开。经过这两年的发展，应该说这一基本态势没有什么大的变化，只是更加强调“大规模”，如 TREC 2004 已经启动了 Terabyte Track。那么，中文信息处理下一步的主要任务有哪些呢？概括而言，至少应该包括以下几方面的研究：

(1) 在人工智能、机器学习、数学、语言学等理论交叉指导下，进行面向超大规模文本等真实复杂环境的方法与原型研究。尤其要注意研究算法在这一条件下的性质。

(2) 面向互联网的汉语自动分词研究

- 建立“信息处理用现代汉语通用分词词表”，与国家标准“信息处理用现代汉语分词规范”相互衔接。这个通用词表将成为构造语义 Web 所需的通用 ontology 的基础。
- 建立各个主要应用领域的分词词表(总词数当在数百万级)，并制订相关规范。这些领域分词词表将成为各领域 ontologies 的基础。
- “来自互联网”：在通用词表与领域词表的支持下，以互联网上的中文文本集合为基本对象，进行汉语分词歧义等的大规模调查，据之设计有效的分词歧义消解算法，并进行新词汇自动发现的研究。
- “面向互联网”：实现一个可驾御互联网的实用型汉语自动分词系统。研究在分词必然存在一定错误率的条件下中文搜索引擎设计的健壮性问题。

(3) 应用驱动的浅层句法分析技术的研究。

(4) 借鉴 WordNet 与 HowNet，进行大规模汉语语义资源的整合与建设。并且以之为基础，进行汉语语义计算的研究。

(5) 词法、句法、语义一体化的汉语分析模型的研究。

(6) 进行领域 ontology 的研究，并建立示范性 ontology。制订相关的标准。

(7) 研究在海量文本中自动发现词与词之间关系的算法。

(8) 研究高精度的汉语文本自动分类算法，建立 Web 逻辑地图。

(9) 将自然语言处理、OCR、语音识别等技术融合于基于内容的图像、视像处理研究中，以显著提高图像和视像的智能化处理能力。

(10) 完成对文本、声音、图像和视像均具有很强判断能力的关键性应用系统(典型地，如色情和军事、政治敏感信息的自动过滤)。

(11) 促进大规模语言计算资源共享平台与机制的建设。

(12) 将上述成果集成起来，设计并实现实用型工具软件。可以将任意一个普通网站经过若干步深层次处理后自动转换成一个智能型网站，从而被赋予一定的知识管理能力。

(13) 建立并完善我们自己的搜索引擎。

(14) 在内容计算的基础上，研发各类知识服务系统，如基于 Web 的预警系统。

最后，让我们感谢大会邀请报告讲演者和全体作者对大会的热心支持；感谢大会主席、副主席的有力指导，感谢程序委员会的辛勤劳动；感谢大会组织委员会的出色工作；感谢清华大学出版社在出版方面的积极配合；感谢赞助单位的慷慨解囊；感谢中国中文信息学会计算语言学专业委员会为会议牵头和所做的贡献。这些共同的努力，促成了国

内计算语言学者风云际会于这虎踞龙盘的六朝形胜之地，用中文与计算机这两个分别代表着悠久文化传承与现代科技的最动人的音符，演奏出一曲弥漫浸润于“桨声灯影里的十里秦淮”的隽永乐章。

我们或许会不经意地吟诵起元代著名诗人萨都剌那两首气象高远、堪称千古绝唱的《满江红·金陵怀古》和《念奴娇·登石头城》中的名句：“石头城上，望天低吴楚，眼空无物。指点六朝形胜地，唯有青山如壁。蔽日旌旗，连云樯栌，……秦淮一片明月”“听夜深、寂寞打孤城，春潮急”。

编　者

2005年7月1日

于清华园

目 录

论文(poster)详摘

语意网与中文信息化的前瞻：知识本体与自然语言处理

黄居仁

中央研究院语言学研究所 台北

E-mail: churen@gate.sinica.edu.tw

摘　要：「语意网」(semantic web)是未来网络发展的方向。而语意网技术中最重要的一环就是知识本体(ontology)。我们讨论在未来网络时代，中文处理面临的挑战。特别是针对如何善用知识本体来表达中文系统化的内涵知识。我们简单介绍了以 SUMO (Suggested Upper Merged Ontology)与词网(WordNet)为基础建立的中研院双语知识词网(Academia Sinica Bilingual Ontological Wordnet，简称 Sinica BOW)。中研院双语知识词网建立的目标，就是提供中文知识本体研究的基础架构。这个知识本体与词汇知识结合的数据库，同时也是自然语言处理应用知识本体的依据。在这个基础上，我们介绍了汉字知识本体(Hantology)，及唐诗三百首知识本体，两个特殊知识本体建构的研究实例。

关键词：语意网，知识本体，词网，中研院双语知识词网，汉字知识本体，Suggested Upper Merged Ontology

Towards Chinese Information Processing in the Semantic Web: Ontology and Natural Language Processing

Chu-Ren Huang

Institute of Linguistics, Academia Sinica, Taipei

E-mail: churen@gate.sinica.edu.tw

Abstract: The Semantic Web is promoted as the future generation of the World Wide Web. The critical technology that the Semantic Web requires is ontology. In this paper, we discuss the challenges that Chinese language processing faces in the Semantic Web. In particular, we focus on how to represent the Chinese knowledge content with ontology. We first introduced Academia Sinica Bilingual Ontological Wordnet (Sinica BOW), a knowledgebase which combines SUMO (Suggested Upper Merged Ontology) and WordNet. It is an explicit goal for Sinica BOW to be the infrastructure for future studies on Chinese ontologies. The combination of ontological and lexical knowledge all supports future semantic web based natural language processing. Based on this, we introduced the newly developed Hantology, an ontology for Chinese characters, and the Ontology for Tang 300 poems. Both are examples of how a specialized ontology can be build based on a general ontology and standard NLP techniques.

Keywords: Semantic Web, Ontology, Wordnet, Academia Sinica Bilingual Ontological Wordnet, Hantology, Suggested Upper Merged Ontology.

1 背景

信息科技发展日新月异。全球信息网（WWW）在不到十年内（1996 起），由蹒跚学步到无所不在；却也到了蜕变的关键时机。全球信息网的发明人柏纳李(Tim Berner-Lee)于 2001 年五月「科学美国人」的专文（中译见「科学人」杂志 2002 年 8 月号）中宣告「语意网」（Semantic Web）将在未来取代全球信息网。

「语意网」是什么？他与「全球信息网」有什么不同呢？我们可以这样说，全球信息网仍只是人们交换文件的载体（media），其中的信息是机器不能自动运用的。也就是说，现在的网络上，只是人把文件放上去，在网络的某一端，另一个人把文件拿下来。我们用来执行工作的计算机，它不需要了解文件的内容。事实上，它完全无法去了解文件的内容。计算机只知道这个文件的身份是什么。它怎么找到这个文件的呢？是根据我们的一些题目、后设资料，我们对文件其它的描述。至于文件包裹里的内容，不管是文献、图形、或任何档案，计算机是不知道的。「语意网」希望作什么样的改变呢？如果规定在每个网页上要增加计算机看得懂得讯息，而这些讯息是专门提供给计算机「阅读」。也就是说，计算机可以自行阅读判断数据是否相关，甚至可以自行把不同网页上的信息根据需求整合。

但是计算机如何阅读语意？其实「语意网」中计算机阅读语意的作法，说起来是相当基本的，有点像图书馆的联合书目。只要先规定好目录的格式，再发动把所有的网页内容讯息根据目录格式完整登录，最后设计一个好的程序来解读这个目录格式，就达成目标了。当然，实际作法复杂些，但基本想法就是如此。计算机阅读语意的实际作法，第一个要利用资源描述架构(RDF, Resource Description Framework)，与通用资源标志码(Universal Resource Identifier, URI）这两个东西连结到相关网页资源。这是网络上已经普遍使用的作法。现今大家用的 HTTP 的地址，就是 URI 的一种。很多人除了后设数据以外，也开始用资源描述架构来描述网页里的知识内容。这是一个大的架构，可以在网络上找到某个特定的资源，而完全没有问题。当连接到网页之后，最重要的阅读的重点，就是要利用知识本体（Ontology）来定义并阅读关键词，并做逻辑推理。

每一个网页上有一个自己定义的 Ontology，就是知识本体。就同一个词来说，在不同的领域，不同的时代，不同的用法上，其意义就不一样。因此如果仅用关键词作网络的搜寻，常会发生错误。如果网页上所有的资源都有一个宣告，告诉每一个来访的程序，这个网页里面知识的定义是什么，知识的架构是什么，那么计算机就可以正确阅读每个网页，而准确地搜寻到所需的数据。

语意网最高的理想是要跨越知识的藩篱，促进知识的演化。因为网页的知识本体提供了不同知识体系的完整描述。语意网的创新关键在于以「知识表达」代替文件名，做为网络间讯息交换检索的依据。而网络资源上定义知识表达架构的知识本体（Ontology），帮助了知识内容的了

解与互通。也就是说，语意网应该可以去除不同知识体系（包括语言）间，鸡同鸭讲的沟通障碍。我们反而要担心的是，未来在语意网上，是否仍有个别弱势语言文化生存的空间。

2 中文信息在语意网时代中面对的挑战

中文信息在语意网时代中面对的挑战是：网络上是否有足够中文的知识内容？语意网上是否有足够中文的使用者（以中文做搜寻与表达的的媒介）？中文可不可以用来表达逻辑严谨的知识本体？

网络上是否有足够中文的知识内容？以现况看，当然不够。目前为止，全球信息网上的数据量，全世界所有语言加起来才和英文相当。也就是说，英文占百分之五十，其它几千个语言加起来占百分之五十。更不用说大部分信息的检索都是用英文，或只能查到英文的资料。但是，这并不是不能解决的问题，数字知识内容是可以创造的。如果使用者有需求，现在的科技可以把数字化的时间压缩得很短。而且从文化传承的观点，世界上各个文化都应该有很强的动机来创造自己的数字知识内容（如台湾的「数字典藏国家型计划」，美国的 American Memory）。从人类多元文化的延续观点看来，这也是人们的共同愿望。因此，对中文的知识内容在网络上的增长，我们可抱持审慎乐观的立场。

语意网上是否有足够中文的使用者（以中文做搜寻与表达的的媒介）？这个问题的关键在人们上网时到底会用自己的语言文字（如中文），还是用英文。到目前为止，大部分的上网人口，似乎都以英文为主要或次要的沟通文字。但最近在 COLING2002 国际计算语言学会中，请几位专家预测十年后的网络与中文处理。朱邦复先生说：「我要在五年内让九亿农民上网。」微软研究院自然语言组经理周明的预测是：「我们比较保守，认为十年之内将有五亿中国人上网。」我个人的预测是：「我不知道有多少人上网，但我想在十年之内网上的人口里面大概有四分之一是讲中文的。」我们看以往的网络发展，上网人口多半是受过西方教育，学过英文的。但中国上网人口如果如预期增加，可想象他们大部分的知识需求与网络行为是在中文环境中的；量变就会导致质变，届时网络上不管是制造的数据或寻找的资源，中文就变成很重要了。这个使用者与市场的发展，如果能配合有识者，有组织的建立中文的知识内容，我们可以对中文在未来网络上的发展，更乐观些。

中文可不可以用来表达逻辑严谨的知识本体？最后这个问题是比较牵涉到研究的实际发展的。不管「全球信息网」是否一定会演化成「语意网」，以语意为出发的与知识表达与检索是不可避免的大方向，而以「知识本体」来描述知识内容与概念架构，也几乎是必然的手段。虽然每个语言与文化各有其特异的知识与概念内容，也在各自的文化典藏中表现出来。但在知识的传播与共享时，这些知识内容必须要用共通，可转换的架构。因此中文信息的挑战，不但是要提出一个合理且有足够涵盖的知识本体，可以描述中文内容的知识架构，而且要保证这个架构能够与其它语言（的知识本体）转换，沟通。

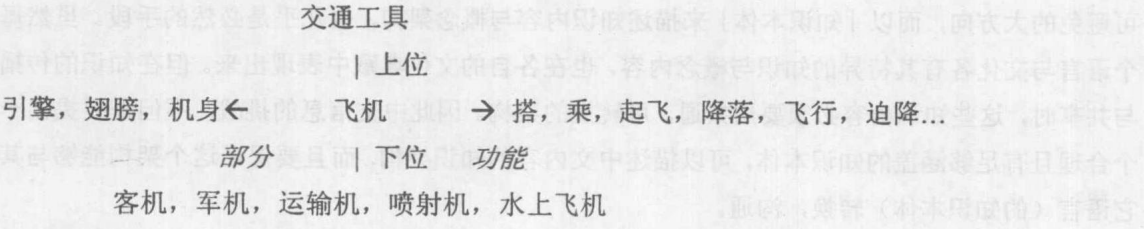
回答上述问题之前，我们需先讨论一个语意网的基本问题：那就是是否有一个共享的知识

本体，可以完善表达世界上所有的知识，因此可以作为所有知识交换的标准？全世界的知识当然不能用一个单一架构表达，因为有许多彼此矛盾的知识体系（比如说各种宗教，不同主义）。但是由知识工程的角度出发，是有可能把最上层的知识概念表达，建立成一个共同的架构。这就是所谓「建议上层共享知识本体」的由来。建议上层共享知识本体 SUMO（Suggested Upper Merged Ontology）是由国际电机工程师学院 IEEE 标准上层知识本体工作小组所建置，共有约一千个概念组成知识本体结构。上层的知识本体限制在后设（meta）的概念、也就是一般、抽象或者哲学概念，因此足够涵盖广阔范围的领域区域最上层的知识结构。特殊领域具体的概念不被包括在上层知识本体中，但是这样的知识本体确可提供特殊领域(例如：医药、财政、项目...等等)的知识本体结构的建立。SUMO 希望藉由最高层次的知识本体，鼓励其它特殊领域知识本体以其为基础衍生出其它特殊领域的知识本体，并为一般多用途的术语提供定义。目前 SUMO 已经和英语词汇网络 WordNet1.6 版本作初步的连结，也就是说，可以由任一英语词汇，得到相对的知识概念节点。SUMO 的出现，使得在在信息应用上，有一个可以用来表达所有知识的共同知识本体架构。

中研院语言所的同仁从中文的词出发，将 IEEE 建议通用的 SUMO 知识本体，以及中英对译的词汇网络结合，建成了「中央研究院双语知识本体词网」（Academia Sinica Bilingual Ontological WordNet）,简称「研究院知识词网」(Sinica BOW, http://BOW.sinica.edu.tw/)。我们的愿景，就是在这个知识库的平台上，逐渐建立可以跨越不同语言与知识系统鸿沟的工具。中研院的双语本体知识词网，同时有中英双语互查，以及由任一个语言检索知识本体的功能。也就是说，可以由任何一个中文或英文词汇的词义，查到在 SUMO 的概念架构上，属于那个词汇的概念节点。这提供了由语言到知识架构的接口。在语言学习上，也可以帮助建立以知识体系及相关概念为基础的学习系统。对中文是否能提供严谨知识表达架构这个语意网上的关键问题，「中央研究院双语知识本体词网」已提供了基本肯定的答案。

而另一个重要的观念，就是语言本身就是一个知识的组织架构。把这个组织架构明确的表达出来，不但有助于语言的知识处理，更重要的是把语言学习与学习者的知识体系结合，更有效，更精确。这就是「词汇网络」建立的基本出发点之一。以下以「飞机」一词为例。它是一种交通工具（上位），而客机等都是一种飞机（下位）。真正学会了飞机这个词，当然包括知道飞机有哪些部分，以及飞机主要的功能等等。有关功能的词汇关系在语言学习上相当有意思。因为名词的功能经常就是由与该名词搭配的动词所展现。

（1）「飞机」的部分语意关系



3 简介知识本体 SUMO

研究语言背后的知识体系，对人类知识的组织，与表达知识的方法，应该可以有创见性的突破。这个研究方向，以往的瓶颈，在于知识体系本来就是研究最基本的先验架构，因此很难有更高阶的理论基础，可供比较研究。但近来知识本体(ontology)的研究展开，提供了比较研究的基础。这一类研究，有许多是在上层共享知识本体 SUMO (Suggested Upper Merged Ontology,) 与词网 WordNet 两个基底架构上进行的。

知识本体(ontology), 在信息科学与网络科技上的定义，就是用来描述一个系统内部知识体系的架构。这个架构通常由一组基本词汇，定义，及该组词汇所建立的关系共同组成。以往知识本体研究最大的瓶颈就是每个系统的知识架构都不同，因此即使建了知识本体，也因为彼此不兼容而无法交换知识。由IEEE 标准上层知识本体工作小组所建置的SUMO（建议上层共享知识本体, http://www.ontologypoprtal.org），其建立的目标就是要提供一个知识本体间建构与知识组织的共同标准。这工作小组的目的是发展标准的上层知识本体； 以促进数据互通性、提高信息搜寻和检索的精确度、并容许自然语言接口与自动推理。知识本体（ontology）内的讯息远远超过字典或者术语表，能使计算机处理更多内容细节和其结构。上层的知识本体被限制在后设（meta）的概念、即一般、抽象或者哲学的概念，因此足够涵盖一般的广阔范围的领域区域。特殊领域具体的概念不被包括在上层知识本体中，但是这样的知识本体确可提供特殊领域(例如：医药、财政… 等等) 的知识本体结构的建立与交换。SUMO 的设计是希望藉由共享最高层次的知识本体，来保证百家争鸣的不同知识系统内容可以分享知识。SUMO只有949个概念节点。采用SUMO为其上层共享本体的知识系统，理论上不会对这些上层作任何修正。而个别知识本体，内部表现的细致知识，虽然很可能会与其它知识本体有所出入。但是这些知识，往上到较高的概念层次，就是SUMO表达的层次，一定有一致的分类。因此，即使是不同的知识体系，也可以藉由共享的上层本体作知识交换与结合。

除了维持上层本体的不变与鼓励其它特殊领域知识本体以SUMO为基础衍生出个别领域的知识本体，SUMO目前另一个关键设计是与英语词汇网络的连结。以人类跨领域语言使用的观点，真正能超越领域障碍去表达所有知识的本体架构，必须能够连接常用的词汇与语言概念到本体架构上。目前SUMO和英语词汇网络WordNet1.6/2.0 版本的连结,使得任何不分领域的知识，都可以藉词汇的连结，建立正确的知识本体位置。这对将来语意网上，支持所有网页都必须有的网页知识本体，也有提供基础数据的积极意义。

4 知识本体研究实例之一：Hantology 汉字的知识本体[1]

[1] 本節內容主要取材自周亞民 2005 年的台大資管所博士論文。