俞士汶 黄居仁 主编

计算语言学前瞻

53

商務印書館

计算语言学前瞻

俞士汶　黄居仁　主编

商　务　印　书　馆
2005年·北京

图书在版编目(CIP)数据

计算语言学前瞻/俞士汶、黄居仁主编. —北京：商务印书馆，2005
ISBN 7-100-04393-X

Ⅰ. 计… Ⅱ. ①俞…②黄… Ⅲ. 计算语言学—国际学术会议—文集 Ⅳ. H087-53

中国版本国书馆 CIP 数据核字(2005)第 010588 号

所有权利保留。

未经许可，不得以任何方式使用。

JÌSUÀN YǓYÁNXUÉ QIÁNZHĀN

计算语言学前瞻

俞士汶　黄居仁　主编

商务印书馆出版
（北京王府井大街36号　邮政编码 100710）
商务印书馆发行
北京瑞古冠中印刷厂印刷
ISBN 7-100-04393-X/H·1103

2005 年 8 月第 1 版　　开本 850×1168　1/32
2005 年 8 月北京第 1 次印刷　　印张 $8\frac{3}{4}$
印数 4 000 册

定价：15.00 元

本书出版得到国家973计划“文本内容理解的数据基础”课题(编号：2004CB318102)的支持

《计算语言学前瞻》编委会

主　编　俞士汶　黄居仁
委　员　黄居仁　俞士汶
刘　玲　刘　扬
吴云芳

此为试读,需要完整PDF请访问: www.ertongbook.com

前　言

与社会信息化进程密切相关的自然语言处理技术与计算语言学，近年来在中国有了长足的进步。国家高技术研究发展计划(863)、国家重点基础研究发展计划(973)、国家自然科学基金、国家社会科学基金都对相关课题给予了支持。两年一度的全国计算语言学联合学术会议于2003年在哈尔滨工业大学召开了第7届会议，第2届学生计算语言学研讨会也于2004年8月下旬在北京语言大学召开，其他相关的学术会议也很多。本领域的研究、教学工作和学术交流都是相当活跃的。本领域的学者当然了解“他山之石，可以攻玉”，故一直努力学习和引进境外的先进理论与经验，不断加强与境外同行的交流。

Computational Linguistics and Beyond（黄居仁，Winfried Lenders联合主编）一书于2004年7月由台湾Academia Sinica语言学研究所出版，这应该是最新问世的计算语言学著作。

这本书由六篇文章组成，作者都是计算语言学领域的国际知名学者。这是他们应2002年于台北市召开的第19届国际计算语言学学术会议(COLING2002)之约，在大会上发表的特邀报告或综合评述，会后整理成文。

第19届国际计算语言学学术会议双年会，已于2002年8月在台北成功召开。这次会议有两个特殊的意义：

这是21世纪的第一次COLING会议。正好提供了计算语言学界集思广益、为新世纪发展定调的绝佳机会。

这是国际计算语言学学术会议第一次在欧美日等先进国家以外的地方召开。不但代表了亚太地区步入科技开发国家之林，也凸显了这个地区的语言，特别是中文，在信息处理的国际舞台上将扮演关键角色。

在大会发表的近三百篇论文中，每篇在该领域中都有其贡献。其中有六篇最具标杆性，值得注意。六篇中有三篇论文的作者是大会的特邀讲席，讲席的国际学术泰斗地位固不需赘言。这三篇论文在学术上所具有的前瞻性，更是令人瞩目。不但是当前学界必读之作，更将成为后世的经典。另三篇论文，则分别由三个不同领域最有影响力的学者，对该领域现状作了全面评估，并为未来的发展勾勒出蓝图。

这六篇论文不但是计算语言学的经典，对21世纪的语言科技，更有重要的启发。特别是网络时代，如何运用语言科技解决知识经济的问题(如2008年奥运的多语信息问题)，将是计算语言学研究回馈于社会的最大贡献。而在学术上，如何在神经、心理与计算机交叉的这一块领域，以跨科际研究来解答智能为何、语言从何处来等人之为人的最基本问题，正是学术研究的最高挑战。本书汲取了学术与科技的精华。若能以中文出版，不但能嘉惠莘莘学子，更可以提升国内学术水平，与国际最高水平接轨。

黄居仁也是北大计算语言学研究所的境外兼职教授。根据黄居仁的建议，俞士汶组织北大计算语言学研究所的一批博士完成了这六篇文章的翻译。翻译得到了原文作者的授权。商务印书馆出版这本译作，是对计算语言学这门新兴学科的扶持，也是对译者的鼓励。

大家都了解，翻译这样高水平的、具有前瞻性的学术著作并不是一件容易的事。为了保证翻译的质量，北大计算语言学研究所采取了以下措施：

(1)翻译者和校对者都署名，以示负责。

(2)翻译者在讨论班上报告每篇原文的内容以及学习心得，报告过程和听众的质疑都检验了译者的理解程度。

(3)成立编委会，成员包括俞士汶、黄居仁、刘玲、刘扬、吴云芳。编委会对全书的质量和体例把关，并编制了“计算语言学前瞻”翻译文稿的体例。实际上，每篇文章都是在译校者和编委会的多轮交互中才最终完成的。

(4)原文作者有通晓汉语者，译者均将译文初稿送呈原作者确认。非常感谢相关原作者的合作与指导，他们的奉献精神和认真负责的态度，对译校者是极大的鼓励和鞭策。还要特别感谢王士元先生，他在英文版出版之后，又为翻译提供了原文的更新版本，因此这篇文章的中文版比英文版更完善。

学术著作无论是原文还是译文都比较难懂，这是正常的，不难理解。正因为考虑到这一点，这本书的一个特点是每篇文章前面都有一个“导读”。对于科技作品的翻译，常有这样的议论：译文看不懂，看原文倒反而看懂了。这是翻译工作的最大失败。本中文版力求避免这种情况的出现。另一方面，科技作品的翻译也不能随意再创作，应当忠实反映原文的内容与风格。本书的翻译基本采用直译方式，必要的时候加“译者注”。第六篇文章在介绍中文信息处理技术的发展和现状时，由于作者一直在香港工作，把重点放在了中国内地以外，这是可以理解的，这也正好与中国内地出版的相关论著、期刊和学术会议文

集互相补足。

六篇译文都包括以下组成部分:题目、作者、译者、校者、导读、正文、作者简介。书后附有两个附录:中英文术语对照表、英文术语索引。

译校者刘扬(liuyang@pku.edu.cn)、刘云(liuyun@pku.edu.cn)、李芸(liyun2003@pku.edu.cn)、吴云芳(wuyf@pku.edu.cn)、李素建(lisujian@pku.edu.cn)、吕学强(lxq@pku.edu.cn)、詹卫东(zwd@pku.edu.cn)、常宝宝(chbb@pku.edu.cn)都是北大计算语言学研究所的博士,李佐文博士现在是河北大学教授,李晋霞博士是《语言文字应用》杂志的编辑,彭国珍同学是北京大学中文系的在读博士生,谌贻容是北大计算语言学研究所在读硕士生。除署名者外,还有北大计算语言学研究所的其他同仁为中文版的问世贡献了力量,恕不一一具名致谢。衷心感谢商务印书馆周洪波先生的支持和刘玲女士为编辑这本书付出的智慧和辛劳。

本稿曾在第2届学生计算语言学研讨会之前少量印发了一些给与会代表(未包括术语表及索引),目的是希望在正式出版前广泛征集对译文的修正意见,尽可能把瑕疵、错漏消灭在出版之前。现在确实收到了效果。

期望《计算语言学前瞻》的中文版能为中国计算语言学的发展奉献绵薄,更期盼广大读者把发现的错误和瑕疵反馈给我们。

俞士汶　黄居仁

2004年10月12日

目　录

计算语言学前瞻：概述*

黄居仁、Winfried Lenders 原著
刘 扬 翻译
彭国珍 校对

1 背景：COLING会议与计算语言学

在《COLING2000会议论文集》序言中，Martin Kay谈到：1960年David Hays创造“计算语言学”这个术语的时候，他心中想像的是“为机器翻译工作提供一个更加坚实的理论基础”。幸运的是，计算语言学从起初就处于该背景之中而又不限于机器翻译自身。来自不同领域和学科的研究者，像数学、语言学、逻辑学、信息科学、统计学和人文科学等，为了一个共

---

* 英文原文为“Computational Linguistics and Beyond：An Introduction”，发表于Huang and Lenders，(eds)，2004，*Computational Linguistics and Beyond*，Institute of Linguistics，Academia Sinica.

此为试读，需要完整PDF请访问：www.ertongbook.com

同的理想和前景走到了一起。他们希望集成各自的知识，以便使计算机能实现从一种语言到另一种语言的翻译，或者使计算机具有相当于人类语言使用者的语言能力。1965 年在纽约召开的第一届 **COLING 会议**已经体现了其作为一个跨学科、跨国界的学术交流论坛的种种设想。随后，COLING 会议逐渐成为一个真正具有国际影响力的活动。它由 **ICCL 委员会**主持和运作，并且，ICCL 委员会“存在的惟一目的就是组织”每两年一次的这个国际计算语言学会议。按照 ICCL 委员会的章程，“ICCL 委员会的成员只代表他们自己，而不是他们所在的国家或机构”（参见 Martin Kay 在 www. dcs. shef. ac. uk/research/ilash/iccl/的说明）。ICCL 委员会的会员身份也跟其他任何学术组织没有从属关系。

正因为这些基本的原则和约定，COLING 会议成为交流计算语言学领域各种经验和想法的一个成熟、有效的场合。COLING 会议也力图覆盖本领域的各个方面并反映其主要发展动向，如机器翻译、**机器辅助翻译**、信息检索、信息提取、语法形式化、句法分析、语义模型、**文摘**、**文本生成**、自然语言理解以及**问答系统**等。学术上的创新和独特的观点一直为 COLING 会议所关注。

经过多年的探索，加之人工智能方法的应用和激励，计算语言学在很多国家都已经发展成为一门正式的学科。计算语言学研究也是**语言工程**和**语言技术**深入发展的一个有机组成部分。从世界范围讲，计算语言学的进一步发展要伴以如下几个看来是“外部的”条件，并不断受其影响和制约。

首先,计算机处理能力快速提高,使我们得以跨越那些在以往的自然语言处理中看来是最严峻的一些障碍:既然计算机的处理速度和内存大小不再成为问题,研究者就可以不受词汇或**知识库**规模过小的约束(即开发所谓的"玩具"系统)。借助于大规模**词库**的基于规则的自然语言处理以及借助于大规模语料库的概率方法和**机器学习**算法的巧妙应用,都被广泛地采纳和实现。随着大规模语料库支持的概率方法的应用,那些在以前被认为是机器几乎不可能解决的研究课题(如词性判定和标注)都找到了具有**鲁棒性**[①]的、可复用的解决方案。在过去的十多年间,占据主流地位的概率方法取得了令人瞩目的成果,这也驱使计算语言学家重新思考如何集成基于规则的知识。目前,将规则引导的**深层分析**和具有**鲁棒性**的基于概率的**浅层分析**结合起来的系统日益受到人们关注,并有可能最终把语言处理系统带向意想不到的更高性能和更广泛应用。

对计算语言学的发展具有重大和持久影响的第二点因素涉及**语言资源**的**可用性**和**可复用性**。需要注意的是,价格适宜的快速计算能力是计算语言学中的(大规模的)语言资源能否被广泛应用的一个先决条件。首先,PC 机的普及已经方便和加速了包括英语在内的多种语言资源的建设和应用。一旦在不同的语种中,便宜、有效的计算不再成为令人头疼的问题,

---

① 鲁棒性(robustness)原是统计学中的一个专门术语,20 世纪 70 年代初开始在控制理论的研究中流行起来,用以表征控制系统对特性或参数摄动的不敏感性。在实际问题中,系统特性或参数的摄动常常是不可避免的。当系统中存在模型摄动或随机干扰等不确定性因素时,仍能保持其满意功能品质的控制理论和方法称为鲁棒控制。

下一步的必然结果就是这些不同种语言资源的汇集和融合。同时，这些不同种语言资源反过来也为特定语言的自然语言处理准备了物质基础。其次，随着对语言资源需求的增长，可复用的问题成为人们关注的焦点。一方面，语言资源的建设是劳动力密集型的工作，因此，复用语言资源意味着时间和投资的节省。另一方面，精心建设的语言资源可用于不同的应用，这些语言资源的子集在特定的规范下又可以结合起来形成新的语言资源。换言之，可复用的语言资源创造了附加值。尤为关键的是，语言活动是瞬时的且随时间和地点而变化，其结果是，当某种语言的讲话人消失或语境发生变化时，某些语言资源就不能再现。因此，复用语言资源既有经济原因也有理论意义。

语言资源的可复用性取决于数据格式和处理工具的规范化和标准化。从计算语言学研究的最初阶段开始，全球都在努力实现数据格式和标注的规范化和标准化。20 世纪 80 年代末出现的 SGML 语言带来了希望的曙光，而 90 年代初出现的文本编码倡议(TEI)对计算语言学的意义就更为重大。规范化和标准化的含义不仅意味着信息交换的便利，同时也使国际间的资源共享成为可能。现在，无论哪种语言，其文本语料库、词库、句法以及工具都能实现国际间的广泛共享。因此，对于一个新的工程而言，人们不必去构造自己的词典(或句法、文本资源)。实际上，语言知识是以一种不受特定语言理论约束的方式提供的，并且也遵守国际间交流的规范和标准。万维网的发展推进了规范化和标准化进程，进而促进了更多的

国际合作。此外,像 WordNet 和最新的**语义网**[①]构想等一些基于网络的语言工程也显示了如下的前景:不同的语言、文化之间存在共同的结构,这使独立于特定语言的知识表示的计算机化成为可能,并且,人们可以在多语的环境中获取这些知识。

第三点因素涉及将**多模态**[②]的思想引入到计算语言学的研究范畴。过去 15 年间信息技术飞速发展,目前已经可以将文本以外的其他**交流通道**也融入到系统中去,比如语音和视频。计算语言学研究集中在书写语言方面已经有 40 多年的历史了。实际上,早在 20 世纪 60 年代,在被称作数字信号处理的**语音识别**和语音产品方面,一些意义重大的研究已经展开。但是,正如该名称所暗示的,这些研究基本不试图采用任何语言上的结构,跟计算语言学也没有什么实质上的联系。当然,不排除少数研究涉及书面语和口语的关系,比如一些语音项目就

---

① 迄今为止,WWW 大多数被用作为人们服务的文档媒体,而在提供可自动处理的数据和信息方面,发展较慢,语义网(Semantic Web)就是想弥补这方面的缺陷和不足。语义网并非独立的另一个 Web,而是现在的 Web 的一个延伸。为 Web 扩展面向计算机的数据,并且增加专为计算机使用的文档,我们就可以把 Web 变成一个语义网。在理想的语义网中,信息有定义完好的含义,更便于人机之间的合作。普通用户能够用现成的有语义标记功能的软件编写语义网页,增加新的定义和规则,计算机会根据关键名称定义的超链接和逻辑推理规则发现语义数据的含义。这种基础设施的框架能够刺激开发自动化的网络服务,例如强大的代理。将语义网融入现在 Web 结构的初步努力已在进行之中。不久的将来,当机器有更强的能力去处理和"理解"现在它仅仅进行显示的数据时,我们将看到很多重要的新功能。

② 在言语交流中,话语的实际发生具有多模态(multimodality)的特性。在具体的情境下,人们总是尽可能运用多样的符号资源来帮助实现意义的理解和建构。随着多媒体技术的发展,现在,人们已经可以远程运用包括视频、语音在内的多种交流方式交互地传递意义,与他人交流。多模态的研究方法就是将语言及其相关的符号资源整合起来。

模拟了最小音对分析。不过，这些研究的一致特征在于以研究标注后的语音信号的输入输出开始，并以此为研究的焦点。目前，现代计算机已经具备了包括语音和视频在内的多媒体功能，有可能将数字信号处理和所谓“深层”理解的句法、语义融合起来。我们已经可以用口语的一些参数，如词的重音和声调模式，来对句子和语篇的结构进行描写和排歧。在不远的将来，用来转录手势和面部表情的**视频通道**也将成为现实。这些信息都有助于语言结构的分析和生成。

最后一点是网络的普及。在万维网成为主流的信息交流媒介之前，自然语言技术的应用受到一定的限制。很难想像计算语言学技术能够应用于世界上如此多样的语言，并且影响到普通人的日常生活。网络作为一个**信息基础设施**将那些并无技术背景的普通用户也吸引到计算机前。他们的一个共同需求是希望能借助日常语言来访问信息，这也促使人类语言技术重新定位到自然语言处理上面。尽管我们还不能证明计算语言学能提供跨语言和跨模态信息访问难题的最佳解决方案，但与之相关的一些问题已经可以用计算语言学来清晰地加以界定和定义。网络的主流地位也让计算语言学重新聚焦在语义、语言形式以及信息内容的关系上。

COLING 会议鼓励计算语言学上的任何新进展和新想法，无论其兴趣点着眼于理论本身、地理区域、政治或是商业。会议的常规的陈述论文包括会议论文和项目介绍等两类，这两类论文都必须通过严格而正式的评审系统，然后发表在会议论文集上。除了论文陈述外，COLING 会议还力图创造一个激励学

术交流的愉悦的环境,提供热烈讨论的机会。此外,大会演讲和专题讨论也是呈现计算语言学新思想和新进展的重要方式。过去 20 年来,一个较新的传统是在 COLING 会议的前后均开展一些学术活动,如会前的辅导讲座和会后的研讨。这些活动的氛围相对比较轻松,可以就特定问题和学科领域的发展趋势展开讨论。

2 新世纪之初的计算语言学:前沿领域

COLING2002 会议在新世纪之初召开。这也是 COLING 会议第一次在传统发达国家(像欧美以及亚洲的日本)之外的地区召开。特定的时间和地点也表明本次会议要探索的是计算语言学的新的前沿领域。遴选并收录到会议论文集的论文集中体现了这一主题。本书收录了大会演讲和两个专题研讨会的一些成果,这些内容在已出版的第 19 届 COLING2002 会议论文集中均未出现,可看做是对会议论文集的一个补充。在本节的以下部分,我们对本书中的论文作了总结,并揭示其与计算语言学新的前沿领域密切相关。

2.1 从结构到意义:建立在计算语言学基础之上

"如果说我能比别人看得更远,是因为我站在巨人肩膀上的缘故"。正如艾萨克·牛顿爵士所言,科学的进步总是建立在前人研究的知识积累的基础上。计算语言学作为一门成熟的学科也毫不例外。本书中的前沿研究都建立在从结构到意义的**映射**上,这是

此为试读,需要完整PDF请访问: www.ertongbook.com