HZ BOOKS
华章教育

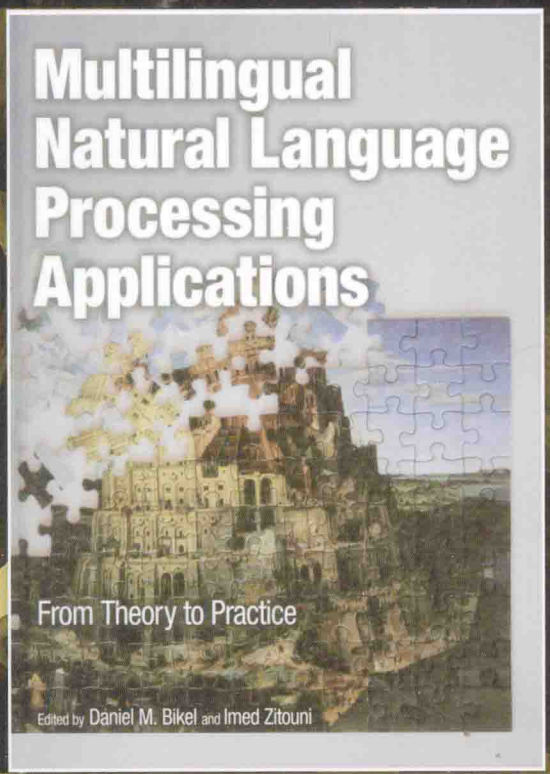
计 算 机 科 学 丛 书

PEARSON

多语自然语言处理
从原理到实践

（美）Daniel M. Bikel Imed Zitouni 编
史晓东 陈毅东 等译

Multilingual Natural Language Processing Applications
From Theory to Practice



机械工业出版社
China Machine Press

计 算 机 科 学 丛

多语自然语言处理

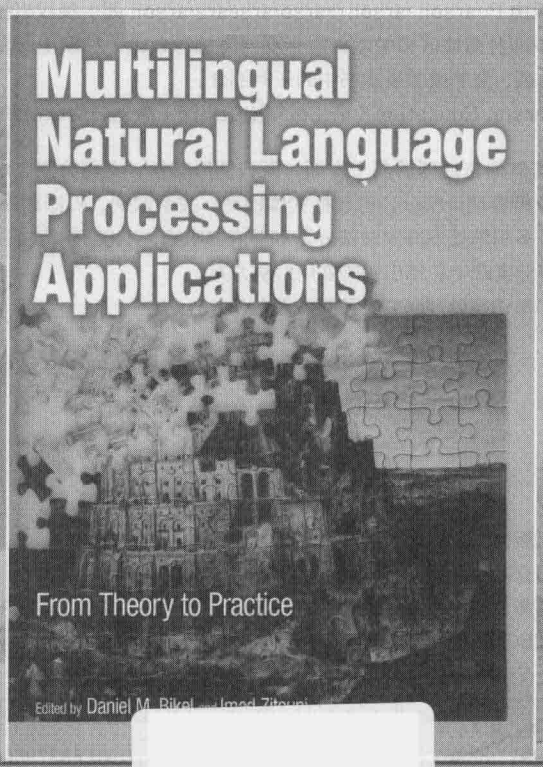
从原理到实践

（美）Daniel M. Bikel　Imed Zitouni 编

史晓东 陈毅东 等译

Multilingual Natural Language Processing Applications

From Theory to Practice



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

多语自然语言处理：从原理到实践 /（美）比凯尔（Bikel, D. M.），（美）兹图尼（Zitouni, I.）编；史晓东等译 . —北京：机械工业出版社，2015.1
（计算机科学丛书）
书名原文：Multilingual Natural Language Processing Applications: From Theory to Practice

ISBN 978-7-111-48491-2

I. 多… II. ① 比… ② 兹… ③ 史… III. 自然语言处理－研究 IV. TP391

中国版本图书馆 CIP 数据核字（2014）第 262220 号

本书版权登记号：图字：01-2013-0217

本书全面阐述了自然语言处理的多个方面，既包括形态学、文档分割、句法、语义分析、语言模型、蕴涵推理、情感分析等理论部分，也包括实体检测、关系识别、机器翻译、信息检索、自动文摘、问答系统、对话系统、多引擎处理等实践部分。本书内容丰富，不仅引用了很多最新的文献，而且还展示了从广泛的研究和产业实践中总结出来的实用解决方案。

本书可供广大的自然语言处理研究者和开发者参考。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：姚　蕾　刘立卿　　责任校对：殷　虹
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2015 年 2 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：29.5
书　　号：ISBN 978-7-111-48491-2　　定　　价：99.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010) 88378991　88361066　　投稿热线：(010) 88379604
购书热线：(010) 68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

文艺复兴以来，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的优势，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力相助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专门为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www. hzbook. com
电子邮件：hzjsj@hzbook. com
联系电话：(010) 88379604
联系地址：北京市西城区百万庄南街 1 号
邮政编码：100037

华章教育
华章科技图书出版中心

译者序

Multilingual Natural Language Processing Applications: From Theory to Practice

本书对自然语言处理的多语言相关现象做了深入的研究，内容丰富，引用了很多最新的文献。对广大的自然语言处理研究者和开发者来说，这是一本非常好的参考书。

全书分为理论和实践两部分。理论部分有 7 章，实践部分有 9 章，各章可单独阅读。下面对各章内容进行简要评述，以供读者参考。

第 1 章主要讨论形态学，重点关注阿拉伯语等屈折语的形态处理。该章提到了汉语的分词问题，但是没有任何描述，另外还讨论了很有意思的形态归纳问题。

第 2 章主要讨论文档结构，包括句子边界检测、话题边界检查，主要讨论了基于特征的机器学习方法，对语音的分割也进行了讨论。

第 3 章讨论了句法分析，涉及的内容丰富而具体。

第 4 章讨论了语义分析，是本书篇幅最大的一章，内容非常详尽，从各类语义问题描述、资源、方法到具体系统，应有尽有。

第 5 章讨论了语言模型，介绍各种先进的语言模型，有很多最新的内容和文献可供读者参考，阅读该章需在理解了 n 元模型的基础上进行。

第 6 章讨论了文本蕴涵识别，提出了一个文本蕴涵框架，介绍了各类文本蕴涵算法和系统及其性能评测，提供了很多相关资源。

第 7 章讨论了情感和主观性分析，强调了孳衍（bootstrapping）方法的使用（特别是跨语言孳衍）。

第 8 章讨论了提及检测和共指消解，这是两个信息抽取中的基本问题。该章写得非常简明扼要，而且提供了一种实现。

第 9 章讨论了关系抽取和事件抽取，也属于信息抽取的范畴。该章探讨了机器学习的方法，并提倡将实体检测和关系抽取结合在一个模型里。

第 10 章讨论了机器翻译及其现状、评测与各种模型。

第 11 章讨论了信息检索，内容翔实，特别区分了跨语言信息检索和多语言信息检索。

第 12 章讨论了自动文摘，对其历史、方法、评测、系统构造、工具、多语问题都有细致的描述。自动文摘也可以看作是信息抽取问题。

第 13 章讨论了问答系统，对涉及的实现技术和相关算法都进行了详细的描述。问答系统也可以看作是高级的信息抽取。

第 14 章讨论了提炼，这是介于信息检索和问答系统的一类新兴问题，需要融合多个信息源的知识。

第 15 章讨论了口语对话系统，包括其体系结构、技术和方法，以及实现中的一些问题。

第 16 章讨论了自然语言处理的多引擎聚合，包括其常见的体系结构，并在 GALE 项目背景下讨论了一个详细案例。

虽然本书的目的之一是在基础知识方面尽量完整，读者不需要为了自然语言处理基本任务去看很多书，但是，对于已经有自然语言处理基础的读者而言，本书提供了很多最新的研究内容，其参考文献和提供的大量可下载资源的链接非常有价值，省去了读者很多宝贵的时间。对自然语言处理系统的研发者，特别是信息抽取和信息检索相关的开发者，本书是非常好的参考。

译者在翻译时全书尽量采用统一的术语，并且采用浅显的译法来帮助读者理解。然而语言学和

自然语言处理方面的术语迄今还有很多不如意之处，因此可能仍然不能使读者看其言而知其义。

本书由多人翻译，基本上每人一章。翻译人员按照章节顺序分别为史晓东、谭波、徐伟、陈毅东（其中黄哲煌翻译了4.5.2节）、黄研洲、林达真、苏劲松、胡金铭、何中豪、邬昌兴、方瑞玉、罗凌、崔志健和方瑞玉、甘星超、王晓苏、曹茂元。校对工作也由史晓东、陈毅东、谭波等多人参加。全书由史晓东统校，对存在的翻译错误负主要责任。出版社的王春华老师对本书的翻译给出了很多指导性意见，特此感谢。

由于译者水平有限，翻译时间也很仓促，译文中肯定还存在不少错误，欢迎读者批评指正，以便将来修订。译者的联系地址为 mandel@xmu.edu.cn。

前 言

Multilingual Natural Language Processing Applications: From Theory to Practice

看起来几乎每个人都在一定程度上受到了信息技术的发展和互联网繁荣的影响。近来，多媒体信息源变得日益普及。不过，未加工的自然语言文本的总量在不断增长，并且地球上各种主要语言都在不断产生大量未处理文本。例如，英语维基百科报导已有 101 种语言的维基百科，而每种语言至少有 10 000 篇文章。因此，不管是国家、公司，还是个人，都迫切需要来分析、翻译、综合或者提炼这些海量文本。

以前，要开发鲁棒、精确的多语自然语言处理（Natural Language Processing，NLP）应用，研究者或者开发人员需要查阅若干本参考书、几十个期刊或者会议论文。本书旨在为开发此类应用提供所需的所有背景知识和实际建议。虽然这个要求很高，但我们希望本书至少是本有用的参考书。

过去 20 年来，自然语言研究者开发了可处理多种语言的大量文本的若干优秀算法。迄今为止，主流的方法是建立可从实例中学习的统计模型。这样的模型能鲁棒地应对其处理文本的类型甚至语言的变化。如果设计适当，同样的模型可用于新的领域或新的语言，只需要提供相应领域或语言的新的训练实例。这种方法也使得研究者没有必要辛苦地写出处理问题的所有规则以及这些规则联合使用的方式。统计系统一般只要研究者提供可能的输入特征的抽象表示，其相对重要性可在训练（training）阶段学习而得，并在解码（decoding）或者推理（inference）阶段应用于新的文本。

统计自然语言处理领域在快速变化，部分变化源于其快速发展。例如，该领域的主要会议之一是计算语言学年会，其参会人数在过去五年已经翻番。另外，IEEE 语音和语言处理会议和期刊上自然语言处理的文章数目也在过去十年中翻了一番以上。IEEE 是世界上推进技术发展的最大的专业学会之一。自然语言处理研究者不但在解决本领域的问题上取得了内在的进步，也从机器学习和语言学领域的进展中借鉴良多。本书虽注意先进的算法和技术，但主要目的是对该领域的最佳实践进行详尽的阐明。另外，每章会描述所述方法在多语（multilingual）环境下的适用性。

本书分成两部分。第一部分是理论，包括前七章，展示了自然语言处理的各种基础问题以及解决这些问题的算法。头三章关注的是找出各种不同粒度层次的语言结构。第 1 章引入了一个重要概念——形态学（morphology），研究词的结构，以及世界上各种语言的不同形态现象的处理方法。第 2 章讨论了多种方法，文档可由此分解为更易处理的部分，如句子，以及通过主题联系的更大的单位。第 3 章研究了发现句子内部结构的方法，也即句法（syntax）。句法一直都是语言学最重要的研究领域，这种重要性也反映在自然语言处理领域。说其重要，部分原因是句子的结构和句子的意义相关，所以找出句法结构是理解句子的第一步。

找出句子或者其他文本单位的结构化的意义表示，经常称作语义分析（semantic parsing），这是第 4 章的内容。第 4 章还特别讨论了近年来引起诸多关注的语义角色标注（semantic role labeling）问题，其目的是找出可作为动词或谓词的论元的句法短语。对动词的论元进行了识别和分类，我们离生成句子的逻辑形式（logical form）又靠近了一步，而逻辑形式是句子意义的一种表示，这种表示方式容易被机器处理，而用于处理逻辑的多种工具人类自古代就开始研究了。

然而，如果我们不需要语义分析生成的深层句法语义结构呢？如果我们的问题只是确定多个句子中哪个句子是人最可能写或者说的呢？解决此问题的一种方法是开发一个可根据语法合法性而为句子打分的模型并以此选取分值最高的句子。给出一个词串的分值或概率估计的问题称为语言模型（language modeling），这是第 5 章的主题。

表示意义和判断句子的语法合法性只是处理语言前期步骤中的两种。为了进一步理解意义，我

们需要一个算法，该算法可对一段文本中表示的事实进行推理。例如，我们想要知道一个句子中提到的事实是否被文档中前面的某个句子所蕴涵，这种推理被称为识别文本蕴涵（recognizing textual entailment），这是第 6 章的主题。

找出陈述或事实的相互蕴涵显然对文本自动理解很重要，但是这些陈述的性质也有待考究。理解一个陈述是否是主观的，并找出其表述的意见的倾向性是第 7 章的主题。由于人们经常表达意见，这显然是一个重要的问题，尤其在社交网络已经成为互联网上人际交流的最重要形式的时代，这一点更显重要。本书第一部分以本章作结。

本书第二部分是实践，讲述如何将第一部分描述的自然语言处理基础技术应用于现实世界中的问题。应用开发经常要做权衡，如时间和空间的权衡，因此本书应用部分的章节探讨了在构建一个鲁棒的多语自然语言处理应用时，如何进行各种算法和设计决策的权衡。

第 8 章描述识别和区分命名实体（named entity）以及这些实体在文本中提及的办法，也描述了识别两个以上的实体提及共指（corefer）的方法。这两个问题一般称为提及检测（mention detection）和共指消解（coreference resolution），它们是一个更大的应用领域——信息抽取（information extraction）的两个核心部分。

第 9 章继续信息抽取的讨论，探索找出两个实体如何发生关系的技术，也称为关系抽取（relation extraction）。要识别事件，并对此进行分类，称为事件抽取（event extraction）。此外，事件涉及多个实体，我们希望机器能找出事件的参与者及其所起的作用。因此，事件抽取与自然语言处理中的一个关键问题“语义角色标注”紧密相关。

第 10 章描述自然语言处理领域中最古老的问题之一，这本质上也是一个多语自然语言处理问题：机器翻译（Machine Translation，MT）。从一种语言翻译为另外一种语言，一直是 NLP 研究追求的目标。在学术界几十年的努力之后，近年来已经研究出多种方法，在现有的硬件条件下可以进行实用的机器翻译了。

翻译文本是一回事，但是我们如何理解现存的海量文本呢？第 8、9 章对帮助我们自动产生文本中信息的结构化记录进行了一些探索。解决海量问题的另一个办法是通过查找与某个搜索查询相关的少量文档或者文档的一部分来缩小范围。该问题称为信息检索（information retrieval），这是第 11 章的主题。像 Google 一样的商用搜索引擎在很多方面可看作大规模的信息检索系统。由于搜索引擎非常流行，因此这是个很重要的 NLP 问题——考虑到有大量语料是非公开的，从而不能被商业引擎搜索到，所以信息检索越发重要。

处理大量文本的另一个办法是自动文摘，这是第 12 章的主题。摘要很困难，一般有两种做法：找到若干个句子或句子片段来表示文本的大意；理解文本，将其意义进行某种内部表示，然后生成摘要，与人为的操作一样。

人们经常倾向于使用机器自动处理文本，因为他们有很多问题要找到答案。这些问题可以是简单的事实性问题，如“约翰·肯尼迪何时出生”，也可以是复杂的问题，如“德国巴伐利亚的最大城市是哪个”。第 13 章讨论如何建造自动回答这类问题的系统。

如我们想回答的问题还更复杂那该怎么办？我们的查询可能有多个答案，如“找出奥巴马总统在 2010 年会见的外国政府首脑”。这类查询可由在 NLP 中被称为提炼（distillation）的一门较新的子学科处理。提炼需要真正地把信息检索和信息抽取技术结合起来，同时还要增加自己的技术。

在许多情形下，我们希望机器能利用语音识别和合成技术交互式地处理语言。这样的系统称为对话系统（dialog system），这在第 15 章讨论。由于在语音识别、对话管理和语音合成方面的技术进展，对话系统越来越实用，并且已经在实际场合中广泛安装使用。

最后，我们作为 NLP 研究者和工程师，希望用世界上开发的大量不同的部件来构造系统。这种

处理引擎的聚合在第 16 章介绍。虽然这是本书的最后一章，但从某种意义上来讲这代表处理文本的开始而非结尾，因为该章描述了一个通用的架构，可用来生成不同组合的一系列处理流水单元。

我们希望本书是自足的，同样希望读者将其作为学习的开始而不是结束。每章都有大量参考文献，读者可以用来继续深入研究任何话题。NLP 的研究队伍在全世界越来越壮大，我们希望你加入我们的行列，一起进行自动文本处理的激动人心的探索。你可以在大学、研究所、会议、博客甚至社交网络上和我们一起交流。多语自然语言处理系统的未来是十分光明的，我们期待你的贡献！

致谢

写作本书伊始，我们就将它定位为多个作者通力合作的成果。我们对 IBM 出版社/Prentice Hall 在起步阶段给予的鼓励和支持怀有无限的感激，特别要感谢 Bernard Goodwin 和所有其他在 IBM 出版社工作的员工，他们在项目的开展和结束过程中给予了帮助。这样一本书当然也离不开我们各章节作者大量的时间、努力和技术才能的投入，所以我们非常感谢 Otakar Smrž、Hyun-Jo You、Dilek Hakkani-Tür、Gokhan Tur、Benoit Favre、Elizabeth Shriberg、Anoop Sarkar、Sameer Pradhan、Katrin Kirchhoff、Mark Sammons、V. G. Vinod Vydiswaran、Dan Roth、Carmen Banea、Rada Mihalcea、Janyce Wiebe、Xiaqiang Luo、Philipp Koehn、Philipp Sorg、Philipp Cimiano、Frank Schilder、Liang Zhou、Nico Schlaefer、Jennifer Chu-Carroll、Vittorio Castelli、Radu Florian、Roberto Pieraccini、David Suendermann、John F. Pitrelli 以及 Burn Lewis。Daniel M. Bikel 还对 Google Research 表示感谢，特别对 Corinna Cortes 在本项目最后阶段给予的支持表示感谢。最后我们（Daniel M. Bikel 和 Imed Zitouni）要对 IBM Research 的支持表示由衷的感谢，特别要感谢 Ellen Yoffa，没有他，本项目就不可能完成。

Daniel M. Bikel（dbikel@google.com）是 Google 的高级研究科学家。他于 1993 年荣誉毕业于哈佛大学，获得古希腊语和拉丁语古典学学位。1994～1997 年，他在 BBN 工作，参加多项自然语言处理研究，包括开发首个高精度随机名字发现程序，并拥有专利。他分别在 2000 年和 2004 年获得宾夕法尼亚大学计算机科学硕士和博士学位，发现了统计句法分析算法的新特性。2004～2010 年，他是 IBM 研究院的研究人员，参与多项自然语言处理研究，包括句法分析、语义角色标注、信息抽取、机器翻译、问答等。Bikel 博士是《计算语言学》杂志的审稿人，ACL、NAACL、EACL 和 EMNLP 会议程序委员。他还在一流的会议和杂志上发表了大量同行评审的论文，并开发了在自然语言处理界广泛使用的软件工具。在 2008 年的“ACL-08：HLT”会议上获得了最佳论文奖（出色短文）。2010 年以来，Bikel 博士一直在 Google 从事自然语言处理和语音处理研究。

Imed Zitouni（izitouni@us.ibm.com）2004 年迄今是 IBM 的高级研究员。他分别于 1996 年和 2000 年从法国南锡大学荣誉毕业并且获得计算机科学硕士和博士学位。他于 1995 年获得突尼斯一家著名的国家计算机学院（Ecole Nationale des Sciences de l'Informatique）的工程硕士学位。

在加入 IBM 前，他在 1999 年和 2000 年是一家初创公司 DIALOCA 的首席科学家。2000～2004 年，他作为研究人员加入了 Lucent-Alcatel 贝尔实验室。他的研究兴趣包括自然语言处理、语言模型、口语对话系统、语音识别和机器学习。Zitouni 博士是 2009～2011 年 IEEE 语音和语言技术委员会委员。他是《ACM Transactions on Asian Language Information Processing》的副主编，计算语言协会（Association for Computational Linguistics，ACL）闪米特语计算方法特别兴趣组的信息官。他是 IEEE 高级会员、ISCA 和 ACL 会员，在多个同行评审会议和杂志担任程序委员和主席。他在自己的研究领域内拥有数个专利，在同行评审的会议和杂志上发表了 70 多篇论文。

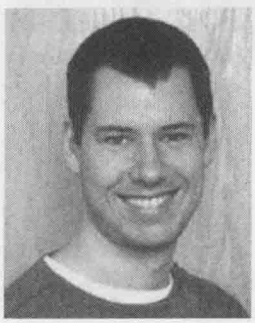
Carmen Banea（carmen.banea@gmail.com）是北得克萨斯大学计算机科学和工程系的博士生。她的研究领域是自然语言处理。她的研究工作集中于多语主观性和情感分析，她开发了基于词典和基于语料库的方法，利用资源丰富的语言来建立其他语言的工具和数据。Carmen 在主流的自然语言处理会议上发表了多篇论文，会议包括 ACL、EMNLP（Empirical Methods in Natural Language Processing）、ICCL（International Conference on Computational Linguistics）等。她在多个大型会议上担任程序委员，也是《计算语言学》杂志和《自然语言工程》杂志的审稿人。她在与 ACL 2010 共同召开的 TextGraphs 2010 Workshop 上担任共同主席，也是 2009～2011 年北美计算语言学奥林匹克赛的北得克萨斯大学站的组织者之一。

Vittorio Castelli（vittorio@us.ibm.com）1988 年毕业于米兰理工大学，获得电子工程学士学位，并于 1990 年、1994 年和 1995 年分别获得电子工程硕士学位、统计学硕士学位和电子工程博

士学位。其中博士学位的论文是关于信息论和统计分类的研究。1995 年他加盟 IBM T. J. Watson Research Center。最近他的研究方向是自然语言处理，特别是信息抽取领域。他致力于研究 DARPA GALE 和机器阅读项目。Vittorio 在此之前启动了 Personal Wizards 项目，该项目的目标是通过观察专家执行任务的过程来捕捉执行流程知识。他已经完成的工作涉及信息论、内存压缩、时间序列预测和索引、性能分析，提出了对计算机系统的可靠性和服务性能与科学图形数字库的改进方法。1996～1998 年，他是编号为 NCC5-101 的 NASA/CAN 项目的共同研究人员。他主要的研究兴趣包含信息论、概率论、统计和统计模式识别。1998～2005 年，他是哥伦比亚大学的助理教授，讲授信息论和统计模式识别。他是 IEEE IT Society 的 Sigma Xi 成员，也是美国统计协会的成员。Vittorio 发表的论文涉及自然语言处理、计算机辅助教学、统计分类、数据压缩、图像处理、多媒体数据库、数据库挖掘、多维度索引结构、智能用户接口以及信息论的根本问题，并共同编辑了《Image Databases：Search and Retrieval of Digital Imagery》(Wiley，2002)。

Jenifer Chu-Carroll (jencc@us. ibm. com) 是 IBM T. J. Watson Research Center 语义分析与集成部门的研究人员。她于 2001 年加盟 IBM，在此之前，她以技术人员的身份在 Lucent Technologies 贝尔实验室工作了五年。她的研究兴趣包含问答、语义搜索、会话处理和口语对话管理。

Philipp Cimiano (cimiano @ cit-ec. uni-bielefeld. de) 是德国比勒费尔德大学的计算机科学教授。他领导的 Semantic Computing Group 隶属于 Cognitive Interaction Technology Excellence Center，该中心在卓越创新体系下由德国研究基金会 (Deutsche Forschungsgemeinschaf) 资助。Philipp Cimiano 在斯图加特大学的主攻专业是计算机科学，辅修专业是计算语言学。他在卡尔斯鲁厄大学获得了博士学位（最高褒奖）。他主要的研究兴趣在于如何将语义技术与自然语言相结合。在过去的几年里，他致力于多语言信息的访问的研究。他作为主要研究人员参加了许多欧洲研究项目 (Dot. Kom，X-Media，Monnet) 和国际研究项目，例如 SmartWeb (BMBF) 和 Multipla (DFG)。

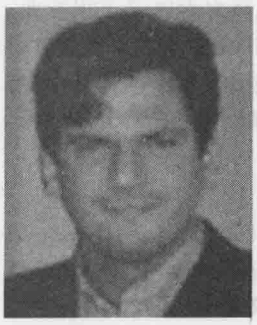


Benoit Favre (benoit. favre@lif. univ-mrs. fr) 是位于法国马赛的艾克斯-马赛大学的副教授。他的研究领域是自然语言理解。他的研究兴趣在于利用机器学习方法来解决语音和文本理解问题。他于 2007 年在法国阿维尼翁大学获得博士学位，其中论文的主题是语音自动摘要。2003～2007 年，Benoit 在阿维尼翁大学担任教学助理，并在同一时期作为巴黎 Thales Land & Joint Systems 的研究工程师。2007～2009 年，Benoit 在国际计算机研究所 (Berkeley，CA) 语音组做博士后研究。2009～2010 年，他在法国勒芒大学做博士后研究。从 2010 年开始，他成为艾克斯－马赛大学的终身副教授和 Laboratoire d'Informatique Fondamentale 的会员。Benoit 在国际会议和期刊上合著的审阅论文超过 30 篇。他是该领域主要会议 (ICASSP、Interspeech、ACL、EMNLP、Coling、NAACL) 和期刊《IEEE Transactions on Speech and Language Processing》的审稿人。他是 International Speech Communication Association 和 IEEE 的会员。

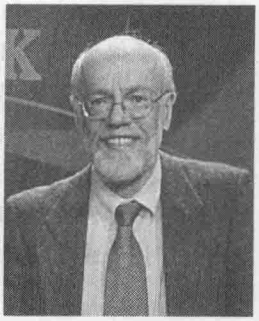
Radu Florian（raduf@us. ibm. com）是 IBM 统计内容分析（信息抽取）组的经理。他于 2002 年在约翰斯·霍普金斯大学获得博士学位。同年加入 IBM 多语自然语言处理组。在 IBM，他参与了信息抽取领域很多不同的研究项目：提及检测、共指消解、关系抽取、跨文本共指和目标信息检索。Radu 领导研究组参加了几个 DARPA 项目（GALE Distillation，MRP）和 NIST 组织的评测（ACE，TAC-KBP），并且和 IBM 合作伙伴（Nuance）共同开发了用于医疗领域的文本挖掘项目，并为 Watson Jeopardy! 项目做出了贡献。

Dilek Hakkani-Tür（Dilek. Hakkani-Tur@micorsoft. com）是微软首席科学家。在加入微软之前，她在国际计算机科学研究所（International Computer Science Institute，ICSI）语言组和 AT&T Labs-Research（2001～2005 年）从事研究工作。她于 1994 年在中东技术大学获得学士学位，并分别于 1996 年和 2000 年在毕尔肯大学计算机工程系获得硕士和博士学位。她的博士论文是关于黏着语的统计语言建模。她于 1997 年和 1998 年分别在卡耐基梅隆大学语言技术研究所和约翰斯·霍普金斯大学从事机器翻译研究。1998～1999 年，Dilek 在 SRI International 利用词汇和韵律信息来完成语音的信息抽取。她的研究兴趣包含自然语言和语音处理、口语对话系统以及针对语言处理的主动和无监督学习。她拥有 13 个专利，参与撰写的关于自然语言和语音处理的论文数量超过 100 篇。她在 2005～2008 年是《IEEE Transactions on Audio，Speech and Language Processing》的副主编。她现在是 IEEE Speech 和 Language Technical Committee 的当选委员（2009～2012 年）。

Katrin Kirchhoff（kk2@u. washington. edu）是华盛顿大学电子工程专业的研究副教授。她主要的研究兴趣是自动语音识别、自然语言处理和人机交互，特别是针对多语言的应用。她写作的同行审阅的出版物数量超过 70 篇，并且是《Multilingual Speech Processing》的共同编辑。Katrin 现在是 IEEE Speech Technical Committee 的会员，也是《Computer，Speech and Language》和《Speech Communication》的编委。



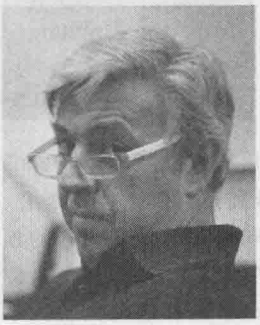
Philipp Koehn（pkoehn@inf. ed. ac. uk）是爱丁堡大学的教授。他在南加州大学获得博士学位，并于 1997～2003 年在该大学的信息科学研究所担任研究助理。他于 2004 年在麻省理工学院担任博士后研究助理，并于 2005 年加盟爱丁堡大学成为讲师。他主要研究统计机器翻译，但也涉及语音、文本分类和信息抽取。他对机器翻译领域的主要贡献是 Europarl 语料的预备与发布、Pharaoh 和 Moses 解码器的开源。他是 ACL 机器翻译特殊兴趣组的组长，也是专著《Statistical Machine Translation》的作者（剑桥大学出版社，2010）。



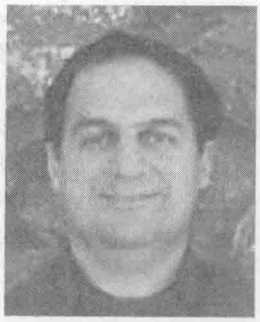
Burn L. Lewis（burn@us. ibm. com）是 IBM T. J. Watson Research Center 计算机科学部门的成员。他分别于 1967 年和 1968 年在奥克兰大学的电子工程专业获得学士和硕士学位，并于 1974 年在加州伯克利大学的电子工程和计算机科学专业获得博士学位。他随后加盟 IBM 的 T. J. Watson Research Center，其主要研究方向是语音识别和非结构化的信息管理。

Xiaqiang Luo（xiaoluo@us. ibm. com）是 IBM T. J. Watson Research Center 的研究人员。他对人类语言技术有广泛的研究经历，包含语音识别、口语对话系统和自然语言处理。在 IBM 语音和语言技术领域的很多由政府资助的成功项目中，他是主要的贡献者。他在 2007 年获得 IBM 杰出技术成就奖，在 2006 年获得 IBM ThinkPlace Bravo 奖和许多发明成就奖。Luo 博士分别于 1999 年和 1995 年在约翰斯·霍普金斯大学获得博士和硕士学位，于 1990 年在中国科学技术大学电子工程专业获得学士学位。Luo 博士是计算语言学协会成员，并且作为多个人类语言和人工智能主要技术会议的程序委员。他是中国科学与技术协会大纽约分会（Greater New York Chapter）委员会的成员。他于 2007～2010 年担任《ACM Transactions on Asian Language Information Processing（TALIP）》的副主编。

Rada Mihalcea（rada@cs. unt. edu）是北得克萨斯大学计算机科学与工程系副教授。她的研究兴趣是计算语言学，特别是词汇语义学、自然语言处理中基于图的算法以及多语自然语言处理。她目前参与了多项研究项目，其中包含词义消歧、单语言和交叉语言的语义相似度、关键词自动抽取、文本摘要、情感分析和计算机幽默。Rada 现担任或曾经担任《Journals of Computational Linguistics》、《Language Resources and Evaluations》、《Natural Language Engineering》和《Research in Language in Computation》等杂志的编委。她的研究获得了 National Science Foundation、Google、National Endowment for the Humanities、State of Texas 的资助。她获得了国家科学基金会 CAREER 奖（2008 年）和美国总统青年科技奖（PECASE，2009 年）。

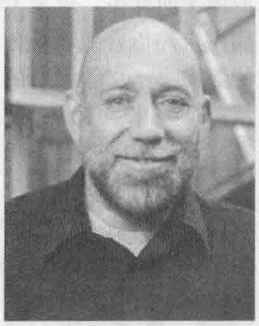


Roberto Pieraccini（www. robertopieraccini. com）是 SpeechCycle 公司首席技术官。Roberto 在 1980 年毕业于意大利的比萨大学电子工程专业。1981 年，他是 CSELT 的语音识别研究人员，CSELT 是意大利电话运营公司的研究机构。他于 1990 年加入贝尔实验室（美利山，新泽西州），成为一名从事语音识别和口语理解研究的技术人员。随后他于 1996 年加入 AT&T 实验室，在这里他开始了口语对话的研究。1999 年他担任 SpeechWorks International 的研发主管。2003 年，他加盟 IBM T. J. Watson Research Center，管理高级会话互动技术部，在 2005 年加盟 SpeechCycle，成为首席技术官。Roberto Pieraccini 在语音识别、语言建模、字符识别、语言理解和自动口语对话管理等领域所著的论文和文章超过 120 篇。他是 ISCA 和 IEEE 会员，是《IEEE Signal Processing Magazine》和《International Journal of Speech Technology》的编委。他也是 Applied Voice Input Output Society and Speech Technology Consortium 委员会成员。



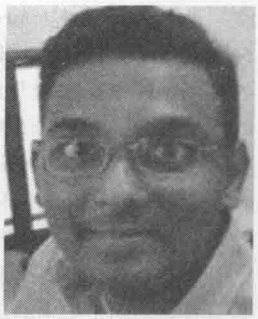
John F. Pitrelli（pitrelli@us. ibm. com）是 IBM T. J. Watson Research Center 多语自然语言处理部门的成员。他分别于 1983 年、1985 年、1990 年在麻省理工学院电子工程与计算机科学专业获得学士、硕士和博士学位，研究生时的工作是关于语音识别与合成的。在担任当前的职务之前，他在纽约怀特普莱恩斯的 NYNEX Science & Technology 公司的 Speech Technology Group 工作。是 IBM Pen Techologies 组的成员。他也在 Watson 的 Human Language Technologies 组从事语音合成和韵律学研究。John 的研究兴趣包含自然语言处理、语音合成、语音识别、手写体识别、统计语言建模、韵律学、非结构化的信息管理和用于识别的信心建模。他已经发表论文 40 篇，并拥有 4 个专利。

Sameer Pradhan（sameer. pradhan @Colorado. edu）是剑桥大学 BBN Technologies 和麻省理工学院的科学家。他在计算语义领域发表的文章和书籍中的章节得到了大量的引用。他目前正在开创下一代语义分析引擎及其应用。实现这个目标可以通过算法创新；通过研究工具的广泛分布，例如 Automatic Statistical SEmantic Role Tagger（ASSERT）；抑或是通过生成一个丰富、多层、多语言和资源集成的平台，比如 OntoNotes。最后这些语义模型应该替代当前在大多数应用领域普遍使用的简陋的基于词的模型，并帮助丰富语言理解领域达到一个新的水平。Sameer 于 2005 年在科罗拉多大学获得博士学位，随后他在 BBN Technologies 致力于开发 OntoNotes 语料，其中 OntoNotes 是 DARPA Global Autonomus Language Exploitation 项目的一部分。他是 ACL 成员，是针对注解、促进注解领域创新的 ACL 特殊兴趣组的创始成员。他经常担任不同自然语言处理会议和研讨会的程序委员，比如 ACL、HLT、EMNLP、CoNLL、COLING、LREC 和 LAW。他也是一位很有成就的厨师。

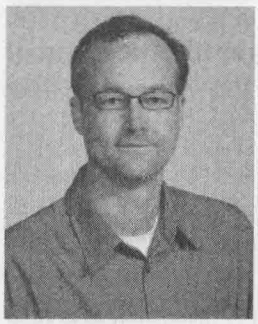


Dan Roth（danr@illinois. edu）是伊利诺伊大学厄巴纳 － 香槟分校计算机科学系和贝克曼研究所的教授。他是 AAAI 的会员、伊利诺伊大学学者，在图书馆与信息科学研究生院和统计语言系担任教师职务。Roth 教授的研究横跨机器学习和智能推理的理论研究，特别是自然语言处理的学习和推导，以及文本信息的智能访问等领域。他在该领域已经发表论文超过 200 篇，并且他的论文获得了多个奖项。他在自然语言应用方面已经开发出了不同的基于高级机器学习的工具，这些工具已经广泛应用在研究界，其中包含一个屡获殊荣的语义分析器。他是 AAAI'11、CoNll'02 和 ACL'03 的程序委员会主席，并且现在是几个他所在领域的期刊的编委。他现在是《Journal of Artificial Intelligence Research》和《Machine Learning Journal》的副主编。Roth 教授以优异的成绩获得以色列理工学院数学专业的学士学位，并在哈佛大学计算机科学专业获得博士学位。

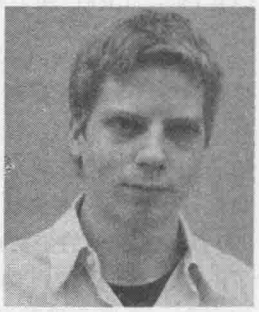
Mark Sammons（mssammon@illinois. edu）是伊利诺伊大学厄巴纳 - 香槟分校认知计算组的首席研究科学家。他主要的研究兴趣是自然语言处理和机器学习，特别专注于将不同的信息源集成到文本蕴涵的上下文中。他的工作已专注于开发一个文本蕴涵框架，使得新的资源可以容易地融入进来，设计出一个合理的推导程序来识别蕴涵，鉴别和开发自动的方法来识别和表达自然语言文本的隐含的内容。Mark 于 2004 年在伊利诺伊大学计算机科学专业获得硕士学位，于 2000 年在英格兰的利兹大学机械工程专业获得博士学位。



Anoop Sarkar（www. cs. sfu. ca/～anoop）是位于加拿大不列颠哥伦比亚省的西蒙·弗雷泽大学的计算科学副教授，他是自然语言处理实验室（http：//natlang. cs. sfu. ca）的主要负责人之一。他在宾夕法尼亚大学计算机与信息科学系获得博士学位。在 Aravind Joshi 教授的指导下完成了半监督的统计句法分析和树邻接文法的句法分析。Anoop 当前专注于研究统计句法分析和机器翻译（利用句法或形态学，或者两者结合）。他的兴趣还包含正规语言理论和随机文法，特别是树自动机和树邻接文法。



Frank Schilder（frank. schilder@thomsonreuters. com）是 Thomson Reuters 研发部的首席研究科学家。他于 2004 年加盟 Thomson Reuters，致力于研究摘要技术和信息抽取系统。他关于摘要的工作已经实现为摘要生成器，用于 WestLawNext 的搜索结果（WestLawNext 是 Thomson Reuters 新开发的法律研究系统）。他当前的研究涉及参加不同的研究比赛，比如由美国国家标准与技术研究所举办的文本分析会议。他于 1997 年在苏格兰的爱丁堡大学认知科学专业获得博士学位。1997～2003 年，他受聘于德国汉堡大学信息系，开始作为博士后研究人员，后来成为助理教授。Frank 已经在几个期刊上发表了多篇论文，并编写了一些书的章节，其中包括《Encyclopedia of Language and Linguistics》（Elsevier，2006）书的“Natural Language Processing：Overview”，内容由他与 Thomson Reuters 的首席科学家 Peter Jackson 合著。2011 年，他联合赢得了 Thomson Reuters Innovation 挑战。他在计算语言学期刊担任审稿人，并多次成为由 Association of Computational Linguistics 组织的会议的议程委员会成员。



Nico Schlaefer（nico@cs. cmu. edu）是卡耐基梅隆大学计算机科学学院的博士研究生，也是 IBM 博士 Fellow。他的研究主要是将机器学习技术应用在自然语言处理任务中。Schlaefer 开发的算法能够让问答系统找到正确的答案——尽管原始的信息源几乎没有包含相关的内容，并开发了一个灵活的框架来支持集成这样的算法。Schlaefer 是 OpenEphyra 的主要作者（OpenEphyra 是最广泛使用的开源问答系统之一）。Nico 对 Watson 贡献了一个统计的源扩展方法（Waston 是一台在 Jeopardy！智力竞赛表演中战胜人类的计算机）。他利用网络和其他大型文本语料库来自动扩展知识源，使得 Watson 能够更加容易地找到答案和支持的证据。

Elizabeth Shriberg（elshribe@microsoft. com）当前是微软首席科学家。之前她在 SRI International（加利福尼亚州门洛帕克）工作。她也隶属于国际计算机科学研究所（加州大学伯克利分校）和 CASL（马里兰大学）。她在哈佛（1987 年）获得学士学位，在加州大学伯克利分校（1994 年）获得博士学位。Elizabeth 主要的兴趣是使用词汇和韵律信息来完成自发语言建模。她的工作旨在将语言学知识与语料、自动语音、说话者辨别技术结合，进而提高科学理解和技术。她在语音科学和技术领域已经发表了大约 200 篇论文，并担任《语言和语音》的副主编，是 Speech Communication and Computational Linguistics 委员会委员，是许多会议和研讨会的委员会委员，是 ISCA Advisory Council 和 ICSLP Permanent Council 的委员会委员。她已经组织了多个研讨会，并担任 National Science Foundation、European Commission、NOW（荷兰）的委员会委员。她已经审阅过许多跨学科的会议、研讨会和期刊（例如《IEEE Transaction on Speech and Audio Process-

ing》、《Journal of the Acoustical Society of America, Nature》、《Journal of Phonetics, Computer Speech and Language》、《Journal of Memory and Language, Memory and Cognition, Discourse Processes》)。2009 年，她获得了 ISCA Fellow 奖。2010 年她成为了 SRI 的会员。

Otakar Smrž（otakar. smrz@cmu. edu）是位于卡塔尔的卡耐基梅隆大学博士后研究人员，他致力于通过学习可比语料的方法来改进以阿拉伯语作为源语言和目标语言的机器翻译。Otakar 在位于布拉格的查尔斯大学完成他的数学语言学的博士研究。他使用函数式编程来设计和实施阿拉伯形态学的 Elixir-Fm 计算模型，并开发了其他自然语言处理的开源软件。他曾经是 Prague Arabic Dependency Treebank 的主要研究人员。Otakar 过去是 IBM Czech Republic的研究科学家，致力于开发无监督的语义分析和对多语言的声音建模。Otakar 是位于卡塔尔的 Džám-e Džam 语言学院的联合创办者。

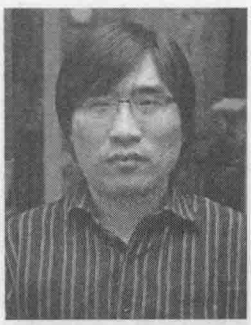
Philipp Sorg（philipp. sorg@kit. edu）是德国卡尔斯鲁厄技术研究所的博士研究生。他是应用信息与形式化描述方法学院的研究人员。Philipp 毕业于卡尔斯鲁厄大学计算机科学专业。他主要的研究兴趣是多语言信息获取。他特别关注利用社会语义应用到 Web 2.0 的上下文中。他已经参与了欧洲研究项目 Active，还参加了国际研究项目 Multipla（DFG）。

David Suendermann（david @ speechcycle. com）是 SpeechCycle Labs（纽约）的首席语音科学家。Suendermann 博士在过去的十年里探索了语音技术研究的很多不同领域。他在多个企业和学术机构从事研究，其中包括西门子（慕尼黑）、哥伦比亚大学（纽约）、南加州大学（洛杉矶）、加泰罗尼亚理工大学（巴塞罗那）和亚琛工业大学（亚琛，德国）。他参与出版的书籍和专利数目超过了 60，其中包括一本书和 5 本书的部分章节，他在慕尼黑的德国联邦国防军大学获得博士学位。

Gokhan Tur（gokhan. tur@ieee. org）目前是微软的首席科学家。他分别于 1994 年、1996 年和 2000 年在土耳其的毕尔肯大学获得学士、硕士和博士学位。1997～1999 年，Tur 访问卡耐基梅隆大学的机器翻译中心，然后访问了约翰斯·霍普金斯大学的计算机科学系，最后访问了 SRI International 的语音技术和研究实验室。他于 2001～2006 年在 AT&T Labs-Research 工作，2006～2010 年在 SRI International 的语音技术和研究实验室工作。他的研究兴趣包含口语理解、语音和语言处理、机器学习及信息获取和抽取。他所著或与他人合著的论文在权威期刊或书籍上发表的数量已经超过 100 篇，并出席了一些国际会议。他是《Spoken Language Understanding: Systems for Extracting Semantic Information from Speech》（Wiley，2011）的编审。Tur 博士是 IEEE、ACL 和 ISCA 的高级会员，也是 IEEE Signal Processing Society（SPS）、2006 年～2008 年的 Speech and Language Technical Committee（SLTC）的会员。目前他是《IEEE Transations on Audio, Speech, and Language Processing》的副主编。

V. G. Vinod Vydiswaran（vgvinodv@illinois. edu）目前是伊利诺伊大学厄巴纳－香槟分校计算机科学系的博士研究生。他的论文是关于网络的信息可信度建模，他的导师是 ChengXiang Zhai 教授和 Dan Roth 教授。他的研究兴趣包含文本信息、自然语言处理、机器学习和信息抽取。V. G. Vinod 的工作包含开发文本蕴涵系统并将文本蕴涵应用在关系抽取和信息获取中。他于 2004 年在印度理工学院孟买分校获得硕士学位，他在导师 Sunita Sarawagi 教授的指导下研究信息抽取的条件模型。随后他在印度的班加罗尔 Yahoo 研发中心工作，研究网络规模信息抽取技术。

Janyce Wiebe（wiebe@cs. pitt. edu）是匹兹堡大学计算机科学专业教授和智能系统计划的联合主任。她与学生和同事的研究方向是自然语言处理的话语处理、语用学、词义消歧和概率分类。她的研究主要关注主观性分析、对文本的情感和意见表达的识别和解释，用于支持自然语言处理的应用，例如问答、信息抽取、文本分类和摘要。Janyce 在专业领域曾担任的角色包括 ACL 议程联合主席、NAACL 程序主席、NAACL 执行委员会委员、计算语言学家、语言资源和评估专家、编辑委员会委员、AAAI 研讨会联合主席、ACM 人工智能（SIGART）特殊兴趣组副主席和 ACM-SIGART/AAAI 博士论坛主席。



Hyun-Jo You（youhyunjo@gmail. com）目前是首尔国立大学语言系讲师。他在首尔国立大学获得博士学位。他的研究兴趣包含定量语言学、统计语言建模和计算语料分析。他对研究形态变化多样、无词序语言的形态句法和话语结构特别感兴趣，例如汉语、捷克语和俄罗斯语。

Liang Zhou（liangz @ isi. edu）是 Thomson Reuters 公司的研究科学家。她在自然语言处理方面有广博的知识，包括情感分析、自动文本摘要、文本理解、信息抽取、问答和信息提炼。她在信息科学研究所做研究生时，积极参与了由政府资助的多个项目，比如 NIST Document Understanding 会议和 DARPA Global Autonomous Language Exploitation。Zhou 博士于 2006 年在南加州大学获得博士学位，于 2001 年在斯坦福大学获得硕士学位，于 1999 年在田纳西州大学获得学士学位，专业都是计算机科学。