计算语言学
与语言科技
原文丛书

CAMBRIDGE

BUILDING NATURAL LANGUAGE GENERATION SYSTEMS

自然语言生成系统的建造

作者 Ehud Reiter, Robert Dale
导读 冯志伟

北京大学出版社
PEKING UNIVERSITY PRESS

Building Natural Language Generation Systems
自然语言生成系统的建造

作者 Ehud Reiter
Robert Dale

导读 冯志伟

北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记　图字 01-2009-3914
图书在版编目(CIP)数据

自然语言生成系统的建造：Building Natural Language Generation Systems/ 雷特(Reiter，E.)，戴尔(Dale，R.)著. —北京：北京大学出版社，2010.8
(计算语言学与语言科技原文丛书)
ISBN 978-7-301-17154-7

Ⅰ.自… Ⅱ.①雷… ②戴… Ⅲ.自然语言处理—研究—英文 Ⅳ.TP18

中国版本图书馆 CIP 数据核字(2010)第 075837 号

书　　名：自然语言生成系统的建造
著作责任者：Ehud Reiter　Robert Dale　著
责任编辑：白　雪　李　凌
标准书号：ISBN 978-7-301-17154-7/H·2494
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
电子邮箱：zpup@pup.pku.edu.cn
电　　话：邮购部 62752015　发行部 62750672　编辑部 62753334
　　出版部 62754962
印　刷　者：世界知识印刷厂
经　销　者：新华书店
　　787 毫米×1092 毫米　16 开本　19.5 印张　330 千字
　　2010 年 8 月第 1 版　2010 年 8 月第 1 次印刷
定　　价：38.00 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”支持)和北京大学出版社合作推出

学术委员会
Academic Advisory Committee

主　　任：

黄居仁(香港)

委　　员：

陈克健(台北)
董振东(北京)
李宇明(北京)
Maarten de Rijke (Amsterdam)
石定栩(香港)
Suzanne Stevenson (Toronto)
王厚峰(北京)
谢清俊(台北)
松本裕治(奈良)
邹嘉彦(香港)
Chris Manning (Stanford)
Harold Somers (Dublin)
陆俭明(北京)
沈　阳(北京)
苏克毅(台北)
王逢鑫(北京)
王士元(香港)
俞士汶(北京)
郑锦全(Urbana-Champaign)

编委会
Editorial Committee

丛书前言

计算语言学(Computational Linguistics, CL)在语言科学与信息科学的研究领域扮演关键性的角色。语言学理论寻求对语言现象规律性的揭示与完整的解释。计算语言学正好提供了验证与应用这些规律与解释的大好机会。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,计算语言学更提供了语言学基础研究与应用研究的绝佳界面。事实上,计算语言学与人类语言科技(Human Language Technology, HLT)可以视为一体两面,不可分割。

计算语言学研究滥觞于上世纪五六十年代的机器翻译研究。中文的相关研究也几乎同步开始,1960 年起在柏克莱加州大学研究室,王士元、邹嘉彦、C. Y. Dougherty 等人已开始研究中英、中俄机器翻译。他们的中文计算语言学研究,可说是与世界最尖端科技同步的。中国国内中俄翻译研究也不遑多让,大约在上世纪 50 年代中期便已开始。可惜的是,这些中文相关早期机器翻译研究,由于硬件与软件的限制,没能延续下来。中文计算语言学研究比较有系统的进展,还要等到 1986 年;海峡两岸在同一年成立了两个致力于中文计算语言学基础架构建立的研究群。北京大学的计算语言学研究所在朱德熙先生倡导下成立,随后一段时间由陆俭明、俞士汶主持。而台湾"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 年返台后加入。

中文计算语言学的研究,20 余年来已累积了相当可观的成绩,重要研究领域与议题中都有可观的研究成果,华人计算语言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产业,跨语言跨文化沟通与知识整合成为知识产业的关键,语言科技的发展日渐成为国际主流。在这个有利发展的大环境下,我们相信,中文计算语言学与华人计算语言学学者的成绩,将会百尺竿头更进一步,进入计算语言学学术核心,并产生把握学科动态、引领学术走向的大师。

回顾计算语言学研究在过去二十年的蓬勃发展，统计模式的引入应该是最主要的原因之一。但二十年后学界也开始看到了统计模式的局限，因此最近几届 ACL 终身成就奖得主，不约而同地大力提倡结合语言学理论与概率模型的研究，来提升计算语言学研究的层次，以寻求新的突破。

回顾中国国内的计算语言学发展，来自计算机科学的贡献多于语言学的贡献。这在理论与概率模型整合研究的大趋势下，不免令人忧心。这也许可以部分归咎于英文研究专著获得不易。国内较易取得期刊或会议论文，但由于篇幅的限制，往往无法对理论做深入完整的阐述，因此也导致国内年轻学者，长于运算而拙于理据。因此，藉由英文专书来弥补不足，巩固研究理据，进而开拓研究视野，是非常重要的一步。

剑桥大学计算语言学原版书系列的引进，就是在上述背景下产生的。本人忝为 Cambridge University Press 所出版的 Studies in Natural Language Processing 系列编辑委员之一，并将于 2010 接任主编。能够将此系列中较重要的几部著作引进国内，责无旁贷。引进原版，不是难事；要真正搭建知识的桥梁，使国内学者与学生开拓研究视野，将原文著作的理论精髓，更多应用于中文研究，则需另加努力。因此，本丛书的特色，是在保留原版的基础上，每本书都邀请一位专家撰写中文导读，其着力点有三：

其一，全书内容简介。导读作者长年浸淫于该领域，对原著能提纲挈领，切中肯綮，并提供相关研究背景。可助读者更准确地掌握并吸收该书的内容。其二，中文相关研究。原作不一定会提到相关的中文研究。由导读专家补充介绍，能搭起理论与中文相关应用的桥梁，从而能够使读者掌握在这个议题进入中文研究的最佳切入点，让相关中文研究的开拓者获得理论的参照和指导。其三，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，几乎马上有新的相关研究。因此，在理论架构的脉络中，加上新近发展，使读者能更贴切地掌握研究脉动。全书摘要通常采用文字叙述。而中文相关研究及最新研究发展则分别以文字叙述及延伸阅读书目的方式呈现。延伸阅读书目，使读者可以很快上手，进入相关研究领域，也是本丛书策划者的苦心所在。可以说导读是本丛书的亮点，不特为原书增色，亦且增加了不少附加价值。

本丛书的出版，是多方协作的结果。在规划出版的漫长过程中，北大计算语言学研究所俞士汶老师及常宝宝老师提供了无私无悔的支持。香港理工大学，特别是北大—理大汉语语言学研究中心与陈瑞端、石定栩、沈阳几位在关键时刻的挹注，也起到了关键作用。当然，整个系列能够顺利出版，离不开有学术眼光和胸襟的北大出版社的支持，而剑桥出版社主管编辑 Helen Barton 从中斡旋，使合约能顺利签订，是必不可少的一环。最后，我要感谢本丛书的国内编委，特别是此次担任导读的各位主笔的辛勤付出，他们为读者搭建了进入学术殿堂的台阶。本丛书的出版，适逢 2010 COLING 国际计算语言学会议在北京举办之际，正象征着国内计算语言学研究与国际的接轨；国内学者风云际会，大展身手，跻身计算语言学的国际舞台，将指日可待。

丛书主编

黄居仁

谨志于香港红磡

二零一零年元月

导 读

冯志伟

1. 学科背景介绍

自然语言生成(Natural Language Generation,简称 NLG)是指从非语言输入构造自然语言输出的处理过程。自然语言生成的目标是建造能够从信息的某种非语言表示产生出有意义的自然语言文本的计算机软件系统。自然语言生成要使用语言的知识以及有关应用领域的知识自动地产生出文档、报告、帮助信息以及其他类型的文本。

自然语言生成通常并不包括那些比较简单的语言生成机制的研究。比如**罐装文本**(canned text)和**模板填充**(template filling)尽管也可以生成自然语言文本,但是,它们的文本格式比较固定,不能算真正意义上的自然语言生成。

自然语言生成的目的与**自然语言理解**(Natural Language Understanding)的目的不同,自然语言生成是从意义映射到文本,而自然语言理解则是从文本映射到意义。自然语言生成与自然语言理解的方向恰好相反。

自然语言理解必须对语言进行自动分析。乍一看起来,语言的生成似乎比语言的分析容易。在语言分析时,我们通常不可能控制分析系统所接收的输入语言结构的复杂程度,而一个生成系统则可以限制它所输出的语言结构的复杂程度。

著名机器翻译学者 Yorick Wilks 说过:"在某种意义上语言理解就好比从一数到无穷大;而语言生成就好比从无穷大数到一。" Wilks 所说的"语言理解"实际上就是语言的自动分析,语言分析"从一数到无穷大"

是越来越难，遇到的问题无穷无尽；语言生成“从无穷大数到一”，目标明确，而且可以人为地控制所生成语言结构的复杂程度。Wilks 的这种看法是关于自然语言生成语重心长的经验之谈，对我们很有启发性。我们在研究自然语言生成的时候，应当注意到他的这些语重心长的经验之谈。

我们认为，自然语言生成的方法与自然语言理解的方法有两点重要的不同。

第一，用于生成处理的输入，性质由于应用的差异会有很大变化。尽管自然语言理解系统的语言输入可能会因文本类型的不同而有所变化，但是所有这些文本都是在一种相对统一的语法规则支配之下的。而自然语言生成系统的输入则是各式各样的。不同的生成系统会有不同的输入。一些生成系统的输入可能是数值表的复杂集的解释，而另一些生成系统的输入可能是用面向对象的软件工程模型的结构组成的。因此，自然语言生成系统必须能够抽取驱动生成处理所需的信息。

第二，尽管自然语言理解和自然语言生成都必须能够表示一系列应用领域所需的词汇和语法形式，但是它们对这些表示的处理方式是不同的。自然语言理解最为关注的是歧义、不确定以及非良构输入的处理。而自然语言生成的输入往往是较少歧义、确定的、良构的。

如果不将罐装文本和模板填充机制考虑在内，自然语言生成领域相对于自然语言处理的其他领域而言是较年轻的。自然语言生成领域的一些普通试验最早出现于 20 世纪五六十年代，并且大部分是在机器翻译的背景下进行的，这些实验把生成看成机器翻译系统的最后一个阶段，生成的输入是机器翻译分析阶段和转换阶段的结果，生成的目标是产生出机器翻译的译文。这样的自然语言生成是隶属于机器翻译的，它只是机器翻译系统中的一个组成部分，没有独立性。

对自然语言生成的集中研究直到 20 世纪 70 年代才出现。Simmons 和 Slocum 的系统采用扩充转移网络（Augmented Transition Network，简称 ATN），从语义网络生成话语。Goldman 的 BABEL 系统采用**决策网**（Decision Net）来实现生成结果的词汇选择，Davey 的 PROTEUS 系统

用于生成对 tic-tac-toe 游戏的自然语言描述①。

20 世纪 80 年代自然语言生成被确立为一个独立的研究领域。McDonald 和 PENMAN 系统对表层实现的研究，McKeown 和 Appelt 对文本规划的研究，都取得了出色的成果，这些研究对自然语言生成的发展作出了突出的贡献。

20 世纪 90 年代自然语言生成开始作为国际计算语言学学会(Association for Computational Linguistics，简称 ACL)的一个特殊兴趣小组，叫做 SIGGEN(The Special Interest Group on Language GENeration)，学者们对自然语言生成的研究兴趣持续增长。

与此同时，学术界对于自然语言生成的定义也更加严格了。Kukich 在 1988 年，Reiter 和 Dale 在 2000 年，都对罐装文本和模板机制的使用及局限性进行了讨论，认为这样的简单文本表示研究还不能算为真正的自然语言生成。

自然语言生成的典型的体系结构是由**文档规划**(document planning)、**微观规划**(microplanning)和**表层实现**(surface realization)等部分组成的**流水线**(pipeline)。流水线被用于约束每个模块的搜索空间，因此使得生成任务更加可控。但是，这种体系结构也存在一个很明显的缺陷：文档规划一经决策之后，就必须执行到底，几乎不可能在表层实现时撤销。1985 年，Applelt 的 KAMP 系统采用了基于**人工智能**的规划和一体化的体系结构，然而，已经证实这种方法无法在较大规模领域的计算中使用。对微观规划本身的各种关注引起了极大兴趣，其中包括对**所指表达**(referring representation)、**集结**(aggragation)和其他的语法问题的研究。一些与自然语言生成相关的问题，例如词汇选择、为特殊听众裁剪输出等问题也得到了关注。

在 20 世纪 80 年代晚期和 90 年代初期，出现了几个可重用的自然语言生成系统，包括两个已经公开发布的系统：1987 年 Bateman 的 KPML 系统和 1993 年 Elhadad 的 FUF 系统。这些系统可以从 SIGGEN 的网站

---

① tic-tac-toe 是一种小孩玩的游戏，在九宫格中，一方画○，一方画×，先把 3 个○或×连成一条直线的一方为赢家。网址如下，http://www.teawamutu.co.nz/fun/games/oandx/index.shtml

下载。大部分的工作采用 Lisp 语言，不过最近也进行了一些将这些系统移植到其他语言或平台的尝试性探索。

话语生成(discourse generation)从一开始就受到自然语言生成的关注。例如，Davey 的 PROTEUS 系统可以对 tic-tac-toe 游戏的过程生成话语描述，一段一段地用话语对游戏过程进行总结。这个系统几乎是完全基于应用系统所记录的游戏踪迹的结构描述来构造话语输出的。1985 年，McKeown 率先研究了基于说明图的文本构造问题，这种方法使用灵活，应用广泛。

自然语言生成的应用目前主要集中在一些相对受限的**子语言**(sub-language)领域，例如，天气预报、说明书的自动生成、百科全书式的描述，等等。自然语言生成的输出可以是简单的文本，超文本、动态生成的超文本、多媒体陈述，还可以是语音输出。在 SIGGEN 的网页(http://www.aclweb.org/siggen)上，我们可以看到这些系统的详细信息。

自然语言生成的研究非常注意理论语言学研究成果的使用。例如，Michael A. K. Halliday 的**系统功能语法**(Systems Functional Grammar，简称 SFG)，Martin Kay 的**功能合一语法**(Functional Unification Grammar，简称 FUG)，Igor Mel'cuk 的**意义—文本理论**(Meaning-Text Theory，简称 MTT)，William C. Mann 和 Sandra Thompson 的**修辞结构理论**(Rhetorical Structure Theory，简称 RST)等，都在自然语言生成系统中得到了广泛的应用。这些语言学理论成为了自然语言生成重要的理论基础。

自然语言生成系统的评价最近颇受关注。评价可以通过几种方法进行，例如，评估输出与有代表性的语料库的相似性，召集专家小组判定输出文本，测试文本是否有效地实现了其交际目的，等等。

自然语言生成中备受关注的其他问题还有：**连接主义**(connectionist)的使用，统计技术的使用，多语言生成系统作为机器翻译的替代品的可行性，等等。

进入 21 世纪以来，自然语言生成的研究得到进一步发展。在文本生成中开始考虑如何介入图形成分、语音流信息以及超文本链接信息的问题。这样的研究，使得自然语言生成的研究与互联网的联系更加密切。

尽管自然语言生成发展迅速，应用广泛，但是，很多从事自然语言处

理的同行对于这个领域还不十分了解。现在还没有一本全面论述自然语言生成的教材，关于自然语言生成的大多数书籍，或者是修改后的博士论文，或者是会议录或讨论会文集的重编本。Gerard Kempen (1987)，D. David McDonald(1992)，Bateman(1998)曾经发表过关于自然语言生成的综述性文章，其中不少文章都总结了自然语言生成的发展历史，论述了自然语言生成与自然语言处理的其他领域的关系，这些文章写得都很不错，但是大多数仍然侧重于自然语言生成的理论方面，对于自然语言生成系统的介绍显得比较单薄。

为了弥补这样的缺陷，Ehud Reiter 和 Robert Dale 在 2000 年写成了这本关于自然语言生成的专著：*Building Natural Language Generation System*(《自然语言生成系统的建造》)。本书从应用的角度对自然语言生成系统进行了全面的论述，详细地介绍了建造完整的自然语言生成系统的知识、背景和有关资源。

本书可以满足如下三方面读者的要求：

- 大学生：他们可以使用本书作为自然语言生成研究生课程的教材，或作为自然语言处理通用课程的补充读物。
- 相关领域的学者：从事自然语言分析、计算机写作辅助工具、超文本技术研究的学者，通过阅读本书可以对于自然语言生成的目标、相关理论、表示方法和算法有所了解，并把自然语言生成与他们自己的研究工作结合起来。
- 软件开发人员：从事信件自动生成、报告自动生成或者其他语言文字输出系统开发的软件人员，通过阅读本书可以了解到自然语言生成的基本知识，知道如何去构建一个自然语言生成系统，并把自然语言生成的有关知识应用到系统的开发工作中。

此外，对于自然语言处理(Natural Language Processing，简称 NLP)、人工智能(Artificial Intelligent，简称 AI)、人机交互(Human Computer Interface，简称 HCI)感兴趣的研究人员以及对于文档自动生成技术感兴趣的开发人员，也可以从本书中得益。

本书的作者是 Ehud Reiter 和 Robert Dale。他们的简历如下：

Ehud Reiter 现在是英国 Aberdeen 大学计算机科学系讲师，1990 年在美国哈佛大学获博士学位，曾经在英国爱丁堡大学任教，在一个专门研

究自然语言生成的公司 CoGenTex 任职。目前担任国际计算语言学会(ACL)自然语言生成特殊兴趣小组(SIGGEN)的秘书。

Robert Dale 现在是澳大利亚 Macquarie 大学语言技术小组负责人,1988 年在英国爱丁堡大学获博士学位,曾经担任该大学讲师,1996—1999 年任澳大利亚微软研究所(Microsoft Research Institute)主任,现任国际最重要的计算语言学期刊 *Journal of Computational Linguistics* 主编。

2. 内容提要

本书介绍怎样来建造一个自然语言生成系统。自然语言生成系统是一个计算机软件系统,它使用人工智能和计算语言学的方法和技术,自动地生成可理解的自然语言文本,这样的文本可以是独立的,也可以是多媒体文档的一个组成部分。自然语言生成系统要从某种非语言表达出发,以这种非语言表达作为信息的输入,使用语言知识和应用系统领域的知识,自动地产生出文档、报告、说明书、帮助信息以及其他类型的文本。

本书介绍了为进行文档规划、微观规划和表层实现等自然语言生成的核心任务所需要的算法和表达方法,通过个案的研究,具体地说明怎样把这些不同的部分组合到一起,形成一个完整的自然语言生成系统。本书还讨论了诸如系统的体系结构、系统分析等问题,并讨论如何把文本生成系统结合到多媒体系统和口语输出系统中去的问题。

本书要讨论的核心问题是自然语言生成系统的建造,全书所有的论述都紧密地围绕这个中心来展开。例如,本书也用了一些篇幅介绍修辞结构理论(RST)和系统功能语法(SFG)这样的理论问题,但是,这样的介绍都是紧密地围绕自然语言生成系统 SURGE 和 KPML 来进行的,所以,对于理论的说明非常简要,完全不涉及这些理论的心理模型和形式模型,尽管这样的模型对于有关理论本身也是十分重要的。

本书对于工程问题的论述也是紧密围绕自然语言生成系统的建造来进行的,对于有关工程问题的理论涉及较少。

由于本书采用了这样的论述方法,所以,本书的名字叫做《自然语言生成系统的建造》,强调了"建造"的重要性。这是本书的一个显著的

特点。

全书分为7章。第一章是导论，第二章介绍自然语言生成的实践问题，第三章介绍自然语言生成系统的体系结构，第四章介绍文档规划，第五章介绍微观规划，第六章介绍表层实现，第七章介绍文本生成之外的问题。

3. 内容详细介绍

第一章“**导论**”首先讨论了自然语言生成的研究背景，说明了自然语言生成和自然语言理解之间的差别，生成和理解共同享有的知识；接着讨论自然语言生成的应用背景，说明了如何把计算机作为写作助理来使用，如何把计算机作为作者来使用，还介绍了如何在教学、市场、人类行为的改变、娱乐等领域中使用自然语言生成技术。本章还介绍了一些自然语言生成系统，例如，WhetherReporter，FOG，IDAS，ModelExplainer，PEBA，STOP等。

WhetherReporter系统和FOG系统可以从天气预报仪器自动收集的大量数字所表示的数据自动地生成通顺可读的文本。

IDEA系统可以从存储于知识库中的复杂机械的信息，生成超文本提供给复杂机械的用户。

ModelExplainer系统可以从**面向对象**(Object-Oriented)软件模型中的信息生成文本描述。

PEBA系统交互地描述分类知识库中的实体，可以自动生成超文本的文档。

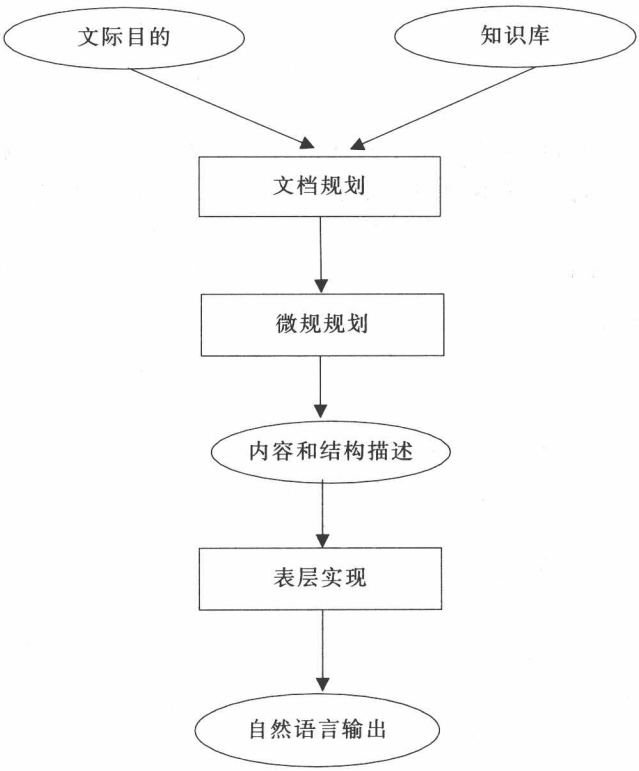
STOP系统可根据吸烟者对于吸烟问题填写的表格，自动生成给该吸烟者的一封信，对他提出有关的建议。

本章最后介绍了自然语言生成研究的历史。

第二章“**自然语言生成的实践问题**”讨论自然语言生成技术适用的环境。进一步讨论自然语言生成系统的需求分析、自然语言生成系统的评测、自然语言生成系统的领域等问题。这些问题在自然语言生成的文献中很少提及，但是，如果我们试图建造一个可操作的、实用的自然语言生成系统的时候，对于这些问题的讨论还是很有价值的。

第三章“**自然语言生成系统的体系结构**”介绍自然语言生成系统特殊的体系结构。自然语言生成系统一般包含文档规划、微观规划和表层实现三个模块。

自然语言生成的体系结构可用下图表示：



生成系统的体系结构

文档规划模块在内容方面的任务是确定生成的内容，在结构方面的任务是确定文档的结构，进行文档的结构化。微观规划模块在内容方面的任务是**词汇选择**(lexicalization)、**所指表达生成**(reference expression generation)，在结构方面的任务是**集结**(aggregation)。表层实现模块在内容方面的任务是语言实现，在结构方面的任务是结构实现。可表示为下表：