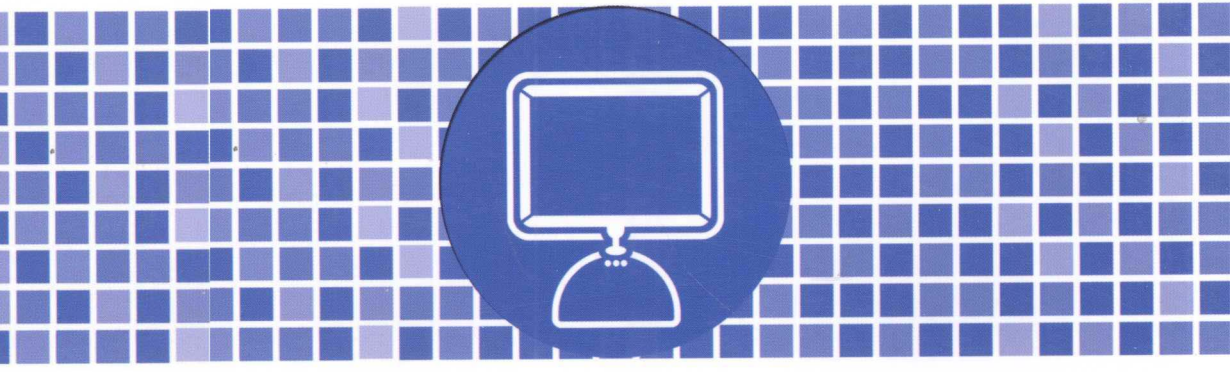
21世纪高等学校电子信息类专业规划教材
北京市重点学科共建项目

ZIRAN YUYAN CHULI CHUBU

自然语言处理初步

能说会道的计算机

[日]荒木健治 著 徐金安 译



清华大学出版社
http://www.tup.com.cn
北京交通大学出版社
http://press.bjtu.edu.cn

21 世纪高等学校电子信息类专业规划教材
北京市重点学科共建项目

自然语言处理初步

——能说会道的计算机

[日] 荒木健治 著
徐 金 安 译

清 华 大 学 出 版 社
北京交通大学出版社
·北京·

内 容 简 介

本书浓缩了日本著名教授荒木健治先生早期的研究成果，书中阐述的内容贯穿了荒木教授提出的“归纳学习法”，即“从具体实例中递归抽取相同和不同部分以获取规则”的基本思想，研究成果涉及自然语言处理领域中分词、句法分析、读音汉字转换、语义分析、机器翻译、对话系统等诸多内容。本书最后一章还重点探讨了计算机与婴幼儿有多接近这样丰富有趣的话题。

本书通俗易懂、深入浅出、内容翔实、实例丰富，本书的内容充分体现了荒木健治教授从事科学研究的先进思想和谆谆善诱的教书育人之方法，极具启发性。

图书在版编目(CIP)数据

自然语言处理初步：能说会道的计算机/（日）荒木健治著；徐金安译. —北京：清华大学出版社；北京交通大学出版社，2012.11
（21 世纪高等学校电子信息类专业规划教材）
ISBN 978-7-5121-1269-8

Ⅰ. ①自… Ⅱ. ①荒… ②徐… Ⅲ. ①自然语言处理-高等学校-教材 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字（2012）第 274774 号

责任编辑：郭东青　　特邀编辑：张诗铭
出版发行：清 华 大 学 出 版 社　邮编：100084　电话：010-62776969
　　北京交通大学出版社　邮编：100044　电话：010-51686414
印 刷 者：北京泽宇印刷有限公司
经　　销：全国新华书店
开　　本：170×235　印张：8.75　字数：162 千字
版　　次：2013 年 3 月第 1 版　2013 年 3 月第 1 次印刷
书　　号：ISBN 978-7-5121-1269-8/TP·715
印　　数：1～3 000 册　定价：27.00 元

译者的话

北海道大学历史悠久，坐落于北海道最美的城市——札幌市，那里四季分明，民风淳朴。

我在北海道大学度过了6年的留学生涯，蓦然回首之间，那一个个清苦、忙碌、充实的日子，如此美好！每及此刻，心中总是充满了无限的温暖和感动，感激之情油然而生！

那种温暖和感动来源于我的每一位亲人，是他们以亲情鼓励我积极迎接和面对人生的每一次挑战，他们是我奋斗的力量之源泉。我要特别感谢我的妻子郭敏，在我们留学日本的每一天，她学习努力刻苦，以各科全优的成绩完成了自己的学业。生活更是积极向上，给了我无微不至的关怀和照顾，留学期间还为我生下了一双儿女，她以柔情和慈爱，哺育我们的孩子茁壮成长，使我的人生丰富多彩。我感激她给予我的每一份温柔和体贴，感谢她给予我家庭的温馨，感谢她吃苦耐劳、坚韧执着！

在我留学期间，我的授业恩师荒木健治教授给予了极大的关怀和照顾，感恩之情难以言尽。荒木教授博学多识、治学严谨，育人循循善诱、注重方法，感谢他教育和培养了我，带领我走进了自己喜欢的研究领域，我庆幸自己的好运，有机会拥有荒木教授这样的好导师！荒木健治教授在自然语言处理领域取得了非常大的成就。本书浓缩了荒木健治教授早期研究成果的精华，与传统的同类专业书籍有所不同的是，本书更加注重研究思路的创新，另辟蹊径，以荒木教授提出的“归纳学习法”，即“从具体实例中递归抽取相同和不同部分以获取规则”为基本思想，内容涉及自然语言处理的分词、句法分析、读音汉字转换、语义分析、机器翻译、对话系统等诸多领域的研究成果，还重点探讨了计算机与婴幼儿有多接近这样丰富有趣的话题，非常值得广大读者借鉴和参考。

一如荒木健治教授在他的序言部分所述，本书的特点在于浅显易懂，深入浅出，既可以是广大初学者的入门指导书，又可以成为相关专业研究人员的参考书目。

本书的翻译得到了吴晓一先生的大力协助，在此表示衷心的感谢！

衷心感谢北京交通大学计算机与信息技术学院的老师们给予我的关怀和帮助，使我能够顺利地翻译出版此书。特别感谢罗四维教授给予的极大鼓励和关怀！

本书的出版得到北京市重点学科共建项目中央高校基本科研业务费专项资金（2009JBM027）资助。

译　者

2012年11月于北京

序　言

“制作一台计算机来理解人类语言并使其实现自我成长”笔者这个想法，是17年前长子刚刚出生时的事了。孩子学习语言的样子实在耐人寻味，灵活适度又循序渐进，简直就像内置了一种极为精巧的语言学习程序一样。从那以后，我就完全被这个系统的魅力所俘虏，并为这套系统的实现倾尽了我的全部心力。

回头看来，那时的想法太过乐观。我曾以为，只要让这套系统的完善与孩子的成长保持同步，过个20年，孩子长大成人，这项研究也可大功告成。虽然这种乐观成为日后的种种失望之源，但能够挑战如此高难度的课题倒也不是什么坏事，经历17年，长子已经成长为成年人了。

同时，我开发的系统也从完全不懂人语的状态，达到了可以与人简单交谈的程度。而且从理论上讲，只要谈话者改变交谈所使用的语言，系统也会用相应的语言进行回馈。此外，在完全不知疲倦、不分昼夜地学习这一点上，也比我那早早就要睡觉的儿子勤勉用功得多，不过遗憾的是，由于先天条件相差悬殊，论聪慧程度还是我那儿子要更胜一筹。

如果您对“让计算机学习人类语言并开口讲话”这一课题感兴趣的话，那么本书将以我的研究为例向您简单介绍迄今为止已做的研究、取得的研究成果，以及研究所面临的困难。为了让没有任何专业知识的读者也能够充分理解本书内容，我做了相应的调整。考虑到本书的定位是自然语言处理领域的文理通用的入门书，我还在每一章的后面加了习题，同时尽可能多地列出了相关参考文献，以便读者做更进一步的了解。

把难懂的内容写得晦涩艰深很简单，但要写得浅显易懂却很难。大概正是由于这个缘故，这个领域的专业书籍无一例外地让初学者望而却步。虽然很多书打着入门书的旗号，但大都有名无实，没有任何基础的初学者读起来还是会觉得吃力。这个问题至今没有得到重视，是因为很少有大学开设自然语言处理的本科课程。不过，在新开设的学科科目中多半会将自然语言处理列入其中，这样想来，真正意义上的自然语言处理入门书必将成为大势所趋。

希望本书的问世能够为这类教科书多提供一个备选方案。另外，虽然在研究生阶段有很多自然语言处理的相关课程，可学生在本科阶段从未接触过这个领域的话，同样会面临基础知识匮乏的问题。假如本书能够成为您的启蒙书，我将感到非常荣幸。

经常会有学生抱怨说，本来对自然语言处理抱有浓厚兴趣，兴冲冲地找书

来读时，却发现内容晦涩艰深，以致兴味索然。可以说，兴趣是最好的老师，倘若入门书晦涩难懂，只会导致学习者丧失学习的意欲，我执笔此书也正是希望能够尽量避免这种事态的发生。由于本书的定位是成为真正意义上的自然语言处理入门书，很多本应详加着墨之处也不得不忍痛割爱。若有读者因此感到意犹未尽，可以参照书后的参考文献去进行更深层次的阅读。幸好这个领域已经出版了大量书籍，使得参考文献的列表非常丰富。

此外，本书的另一个特色是，在各章的结尾会介绍一部分我截至今日的研究成果。我以让计算机学习人类语言为目标一路走来，从未放弃过这个梦想。我深深感到，始终地朝着一个目标走下去，会让我们的人生大大充实。如果通过书中的介绍也能让读者产生共鸣，对笔者来说将是望外之喜。

2004 年 4 月

作　者

目　录

第1章
使用计算机处理语言的方法

1.1 为什么使用计算机处理语言如此困难?

1.1.1 知识储备不足

在科幻小说和动画中出现的机器人，大多像铁臂阿童木[①]一样，可以使用人类的语言进行交谈。毋庸置疑，铁臂阿童木是人形机器人的一种理想形态。但在现实中，我们身边的机器人，比如 AIBO[②]，还无法理解人类的语言并作出应答。

对于人类来说，不需要经过什么特殊训练就能够下意识地学会使用语言。因此，很多人误以为让机器人学习人类语言也是件轻而易举的事。但实际上，让机器人拥有这种能力十分的困难。究其原因，机器人做到这一点至少需要经过三个阶段。第一个阶段是理解人类的语言，第二个阶段是针对理解到的内容生成反馈内容，最后一个阶段是将反馈内容用人类的语言表述出来。图 1-1 表示人与机器人对话的场景。

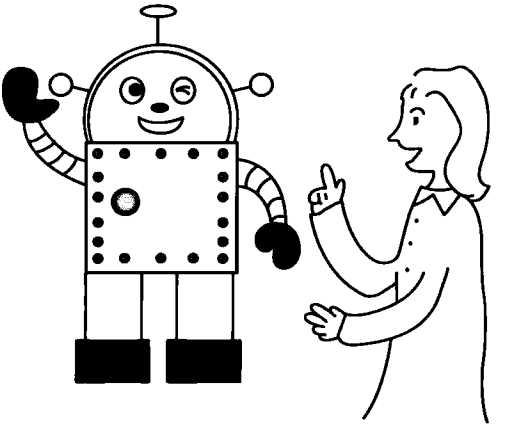


图 1-1 与机器人的对话

人类的语言存在歧义性，而歧义性带来的影响始终贯穿在这三个阶段之中。现在的计算机还无法很好地处理人类语言的歧义性，因此存在各种问题。

此处以机器翻译为例来具体说明语言的歧义性引发的问题。比如，让机器来翻译一句英文“I am a boy.”，这个句子很简单，初一学生都能轻松搞定。当

① 铁臂阿童木是由手塚治虫制作的，于 20 世纪 60 年代上映的卡通中的主人公。根据原著的设定，他在 2003 年 4 月 7 日问世。不过很遗憾，至今也无法开发出像铁臂阿童木那样具备会话能力的机器人。

② AIBO 是索尼公司开发的犬型宠物机器人，同时也是索尼株式会社的商标及注册商标。

然，如果事先告诉计算机，“I am a boy.”的译文就是“我是个男孩子（私は少年です.）。”的话，计算机也能够轻松完成任务。

然而，通过这种方法让计算机记住所有英文的译文是不可能的。英语的单词数量有限，句子的长度也有限，从理论上来说，通过单词组合造出的句子也是有限的。可在实际操作中，需要存储的句子数量过于庞大，难以实现。更何况，“我是个男孩子（私は少年です.）。”并不总是“I am a boy.”的最佳译文。若不能根据前后文生成恰当的译文，系统的实用性就会大打折扣。可这样一来，需要事先输入给系统的译文数目就更加庞大，说是天文数字都不为过。

图1-2是在研究社《新英和中辞典》的电子数据中检索“boy”的结果。在看到“boy”时很难联想到“服务生”这一义项，可这恰恰是在饭店场景中最恰当的译词。因此，并不能将其从词典中删除。

boy
－ｌ名ｌ
1［C］（＜二＞girl）
a 男の子，少年《17－18歳まで》：a boys'school 男子校．/Boys will /wil/ be ～s.《諺》男の子はやっぱり男の子だ，男の子のいたずらは仕方がない．/Got lost，～？坊や，道に迷ったのかい．
b（大人に対して未成年の）青年，若者．
2［C］［しばしば one's ～］（年齢に関係なく）息子．：This is my ～．これが息子です．/He has two ～s and one girl．彼には息子が2人に娘が1人いる．
3［C］男子生徒［学生］：college ～s 大学の男子学生．
4［the ～s］
a 一家の息子たち．
b 男仲間，男連中：the public relations ～s 広告関係の連中．/the ～s at the office 会社の男の同僚．
5［C］［しばしば one's ～］（男の）恋人：Jane's ～ ジェーンの恋人．
6［C］男の召使，給仕，ボーイ《★ｌ解説ｌしばしば軽蔑的に感じられる；レストランでは waiter，ホテルでは bellboy または bellhop を用いる》．
7a［C］（少年のように未熟・未経験な）男．
b［親しみをこめた呼び掛けに用いて］男《★ｌ用法ｌ現在ではあまり用いられない；特に，白人が黒人に対して用いるのはひどい侮辱（ぶじよく）とされる》：Thank you，my（dear）boy．やあ君，ありがとう．
8［C］［修飾語を伴って］《米口語》（ある土地生まれの）男：He's a local ～．彼は土地の人間だ．
That's［There's］my［the］boy!《口語》ようし！しっかり！《激励・賞賛の言葉》．
－ｌ形ｌ［A］男の子の，少年の（ような）：a ～ baby 男の赤ちゃん/a ～ student 男子学生．
－ｌ間ｌ《口語》［しばしば oh を伴い，愉快・驚きまたは皮肉な語調で失望・退屈を表わして］よう!，おや，本当に，無論!【中期英語「男の従者」の意】

图1-2 “boy”的译词

（出自：研究社《新英和中辞典（第6版）》）

翻译系统必须从如此之多的译词中选出一个最恰当的。可单是这译词选择就非常棘手。普通人也许很难想象翻译系统所面临的困难，因为人类可以利用庞大的背景知识来准确地把握上下文，在理解意思的同时给出译文。这一点恰恰是人类与计算机之间决定性的差异。

人们在实际生活中处理自然语言[①]的时候，很难像数学那样通过字面就可以获取全部信息。恰恰相反，理解一个句子所需的信息很少是通过语言自身来承载的。这种没有用语言表述出来的信息叫做背景知识、上下文和常识，对于理解自然语言来说是必不可少的。当然，也可以通过人工逐条写入的方式来弥补计算机在这方面的不足。事实上，也确实有一些工程采取了这种办法。

但遗憾的是，需要人工输入的知识量过于庞大，以至于这些尝试都以失败而告终。我个人以为，人工输入虽然能够取得一定成效，但本质上没有解决任何问题。那么，本质上的解决办法是什么呢？我觉得是要实现一种“学习”机制，通过这种机制，系统就可以一边与人谈话，一边获取理解该谈话所需的知识。在这样一个背景下，我将“实现理解人类语言的计算机”作为研究课题，之后会在本书中加以介绍。

1.1.2 在积木的世界中交流：SHRDLU

就“理解人类语言的系统”这一课题，先回顾一下至今为止所做的研究。早期的研究因为无法处理自然语言自有的歧义性，往往是在假定无歧义的条件下进行的。不能处理歧义的系统没什么实用性，因此被嘲讽为 Toy System（玩具系统）。这类研究的早期代表是 SHRDLU[②]，是由当时就任麻省理工学院（MIT）的副教授 Terry Winograd（现任斯坦福大学教授）开发的。

如图 1-3 所示，这是一个以积木为处理对象的系统。虽然 SHURDLU 和现实世界相比缩水太多，但正因为这种局限，才能让系统基本达到人类的水平。聊天记录如图 1-4 所示，SHRDLU 的回答非常准确，对不懂的问题也不会不懂装懂。

---

① 电脑能够理解的语言叫做人工语言，与此相对，人类通常使用的语言，比如英语和日语，叫做自然语言。

② 文献［2］是该文献的译本。

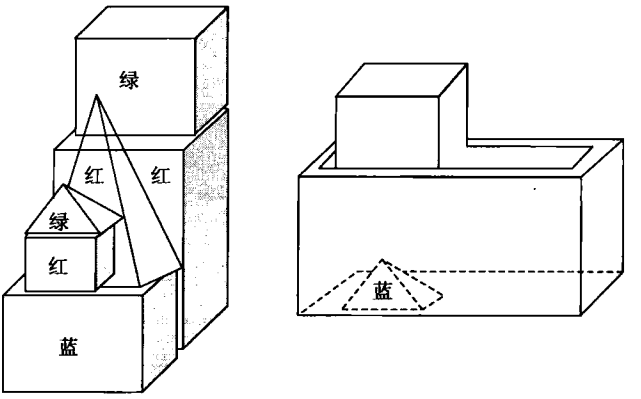


图 1-3　SHRDLU：积木的世界

用　　户：What does this box contain?
（この箱の中に何がありますか?）
（箱子里有什么?）
SHRDLU：THE BLUE PYRAMID AND THE BLUE BLOCK
（青い三角錐と青い箱です.）
（蓝色的三角锥和蓝色的箱子）
用　　户：What is the pyramid supported by?
（三角錐は何で支えられていますか?）
（三角锥是靠什么支撑的?）
SHRDLU：THE BOX
（箱です.）
（箱子）
用　　户：How many blocks are not in the box?
（箱の中にない立方体はいくつありますか?）
（有几个立方体不在箱子里?）
SHRDLU：SIX OF THEM
（6 個です.）
（有六个）
用　　户：Can the table pick up block?
（テーブルは積み木をつかむことができますか?）
（桌子能够抓起积木吗?）

图 1-4　SHRDLU 的聊天记录

（出自：Winograd，T.：Understanding Natural Language，Academic Press）

SHRDLU：NO
（いいえ.）
（不能）
用　户：Can a pyramid support a pyramid?
（三角錐は三角錐を支えることができますか?）
（三角锥能够支撑起三角锥吗?）
SHRDLU：I DONT KNOW
（わかりません.）
（不知道）

续图 1-4　SHRDLU 的聊天记录

（出自：Winograd，T.：Understanding Natural Language，Academic Press）

当然，在当时，还没有开发出用计算机识别人类声音的语音识别技术，输入部分是依靠键盘实现的。而且和现在相比，在图像表现上要简单得多。但是，在开发理解人类语言的系统这一点上可以说迈出了坚实的一步。Winograd 无愧为人工智能研究的领军人物，也是世界知名的学者。时至今日，前来拜访他的人依旧络绎不绝。

虽然 Winograd 本来是人工智能的知名学者，但他在长时间的研究之后，逐渐成为反人工智能论者的领军人物。他开始主张“计算机仅仅是用来计算的工具罢了，无法实现人工智能，只能当作工具使用”①。

Winograd 想法上发生的巨变，可能是他在尝试将积木世界的语言理解机制套用到现实世界时的失败所导致的。也就是说，执行如图 1-5 所示的“移除蓝色立方体上的红色立方体”这条命令之后，红色立方体确实被移除了，可上面的绿色三角锥却悬留在半空中。看到悬浮在空中的绿色三角锥，恐怕任何人都会为其没有坠落感到不可思议。这是因为大家都懂得“移除支撑物后物体会下落”这条物理法则。

因此，SHRDLU 虽然可以正确执行预先给定的命令，却无法动脑筋去执行命令以外的事。也许大家会说，既然这样的话，把“移除支撑物后物体会下落”这条物理法则预先告诉 SHRDLU 不就好了吗？确实，这样一来这个问题可以得到解决，可是依然解决不了其他问题。现实世界中究竟有多少类似的常识？这本身就是一个问题。

也许大家又会想，预先将现实世界中的所有知识毫无歧义地、准确无误地灌输给 SHRDLU 的话，不就可以用计算机实现智能了吗？但事实上，这是非常

① 文献［4］是该文献的译本。

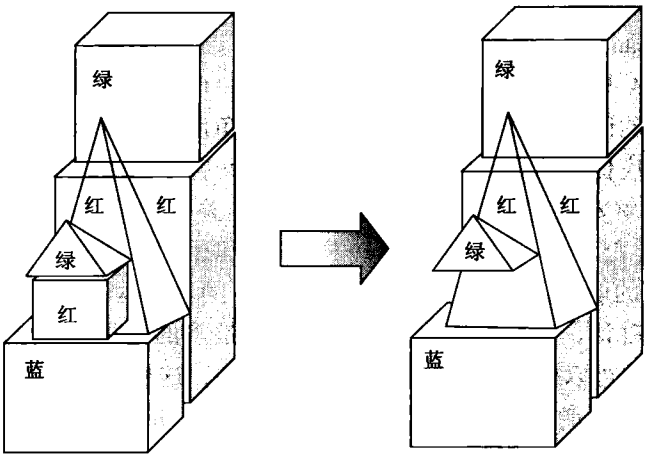


图 1-5　SHRDLU 存在的问题

困难的。美国就有一家公司，为了实现这个目标，一条一条地将常识输入到计算机中。可就在这家公司输入了一千万条常识之后发现，系统仍然做不到像人一样理解语言。

那么人类究竟是如何获取如此庞大的知识体系的呢？人类生来就具备获取语言、储备知识的能力，这种能力叫做“学习”。如果能够用计算机实现这种学习机制，就可以解决前述的庞大知识体系的获取问题。这类研究在近十年来十分盛行，称为“基于学习的方法”。

人类在聆听他人的发言时，并不一定预先就具备理解该发言所需要的全部知识。当信息不充分的时候，通过向对方提问，或借助其他知识推测，以互动①的方式逐步达到理解。基于学习的方法正是基于这一理念，该理念的立足点在于，人类理解事物的本质无非是套用已知的常识来解决新遇到的未知状况。

基于学习的方法有着非常高的健壮性②。因此即便将系统的处理对象扩展到现实世界，也具备准确运作的可能性。因此，这方面的研究正逐渐趋于主流。

1.1.3　会聊天的系统：ELIZA

SHRDLU 是一个只以积木为处理对象，相当受限的系统。除此之外，还有

① 即英语的“interactive”。在这里意指通过相互作用来逐步增进理解。

② 即英语的“robustness”。汉语译作“鲁棒性”或“健壮性”。本处意指，即便在情报不充分或存在歧义的状况下也能够想办法应对的能力。