

**Broadview**<sup>®</sup>  
www.broadview.com.cn

# 自然语言处理实践

聊天机器人技术原理与应用

王昊奋 邵浩 等编著

 中国工信出版集团

 电子工业出版社  
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY  
<http://www.phei.com.cn>

# 自然语言处理实践

聊天机器人技术原理与应用

王昊奋 邵浩 李方圆 张凯 宋亚楠 编著

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

## 内 容 简 介

聊天机器人作为人工智能技术的杀手级应用，发展得如火如荼，各种智能硬件层出不穷。本书系统地阐述了聊天机器人的分类和关键技术，不仅给出了实际案例，还展望了聊天机器人在通往更智能化、更人性化、更趣味化的道路上所面临的挑战。同时，针对聊天机器人在从感知智能到认知智能的跨越中所面临的难题，本书着重讨论了知识图谱和深度学习技术在自然语言处理、问答、推理、服务融合等方面的应用。

本书适合有志于从事人工智能行业，以及想了解聊天机器人到底是什么的读者阅读。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。  
版权所有，侵权必究。

### 图书在版编目（CIP）数据

自然语言处理实践：聊天机器人技术原理与应用 / 王昊奋等编著. —北京：电子工业出版社，2019.3

ISBN 978-7-121-35715-2

I. ①自… II. ①王… III. ①自然语言处理—研究②智能机器人—研究 IV. ①TP391②TP242.6

中国版本图书馆 CIP 数据核字(2018)第 266347 号

策划编辑：郑柳洁

责任编辑：郑柳洁

印 刷：北京季蜂印刷有限公司

装 订：北京季蜂印刷有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编：100036

开 本：720×1000 1/16 印张：12.75 字数：191 千字

版 次：2019 年 3 月第 1 版

印 次：2019 年 3 月第 1 次印刷

定 价：69.00 元



凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888, 88258888。

质量投诉请发邮件至 [zltts@phei.com.cn](mailto:zltts@phei.com.cn)，盗版侵权举报请发邮件至 [dbqq@phei.com.cn](mailto:dbqq@phei.com.cn)。

本书咨询联系方式：010-51260888-819, [faq@phei.com.cn](mailto:faq@phei.com.cn)。

# 推荐序

聊天机器人是社会关系网络、自动客服、语音助手、智能音箱、游戏等的重要支撑技术，它综合应用了自然语言处理技术。自然语言处理是体现语言智能非常关键的技术，它分析、理解和生成自然语言，实现人与机器的自然交流。同时，机器翻译、自动文摘、自动写作、邮件或者短消息的自动回复也有助于人与人之间的交流。如果语言智能可以实现突破，跟它同属认知智能的知识图谱与常识推理等技术也会得到长足的发展，并推动整个人工智能体系的进步，使更多的场景落地。自然语言处理被视为人工智能“皇冠上的明珠”。要做好这项技术，达到和人一样自然的交互是非常具有挑战性的一项课题。许多积极投身于自然语言处理研究和开发的同仁，迫切需要掌握自然语言处理的基础技术，了解技术前沿。

我很高兴看到本书的出版。它系统地介绍了聊天机器人的技术体系和自然语言处理在聊天机器人中的应用，辅以案例，理论和实践结合良好。本书深入浅出的风格对不同层次的读者都有帮助。

本书由王昊奋博士和邵浩博士主导，他们二位都是从学术界跨越到工业界的年轻人，并致力于将技术应用到产品实践中。我和王昊奋在中国计算机学会

术语工作委员会和自然语言专委会等组织中有紧密的合作。我认为，他不仅在学术上积极进取，还特别希望尝试把各种新技术应用到产品中。他将理论和实践相结合，多年来积累了丰富的研发经验，走出了一条独到的创新之路。本书由多位相关企业的资深技术研发人员参与撰写。因此我相信本书一定会激发大家对聊天机器人的兴趣和更深入的思考。

从本书的内容上看，除了对聊天机器人的历史发展和技术体系的阐述，重点介绍了聊天机器人的3种典型表现形式：闲聊、对话和问答。以闲聊型聊天机器人为例，虽然基于检索的方法是目前主流的产品实现方式，但随着自然语言处理端到端技术的发展，生成式对话越来越受重视，有很多研究者尝试用生成方法解决个性化、多轮对话和安全回复等问题。同时，本书介绍了知识图谱的重要作用，因为基于知识图谱的问答也是问答型聊天机器人的重要组成部分。本书尽可能完整地展现了聊天机器人相关技术的最新进展，有兴趣的读者可通过此书全面了解聊天机器人。

聊天机器人已经在智能客服、知识问答等场景里有了较好的应用，未来会在大数据、深度学习和重要场景的推动下进一步提升智能水平。我们可以畅想，在未来的某个时刻，会出现一个基于人工智能技术的虚拟生命，它能够真正理解人类的语言，有自己的记忆和情感，并可以和人进行自然真实的对话。尽管我们离这个目标尚远，但是我们可以逐步靠近。这里孕育着无穷的研究、开发机会和乐趣。我期待本书能激励更多优秀的年轻人投身其中，做出更多成就！

微软亚洲研究院副院长、国际计算语言学学会主席

周明

# 前言

## 缘起

写作本书的初衷，是作者在聊天机器人相关的技术公司工作时，在做产品的过程中，深深感受到理论和实践的差别。举例来说，学术界一直追捧的 seq2seq 技术，并没有很好地应用在聊天机器人的产品中。而且，对于刚入职的工程师，也没有一本系统性的书籍来帮助他们快速理解和掌握聊天机器人的技术脉络。因此，我提议：我们为什么不写一本全面的技术读本，将我们在实践中遇到的问题和解决方案都放到这本书里，让更多的读者了解聊天机器人的知识体系，避免踩我们在工作中踩过的很多坑呢？于是我们开始了本书的撰写。希望这本书能尽可能全面地梳理聊天机器人技术，使读者更深入地理解其背后的理论知识。

## 本书特色

本书应该是国内第一本聊天机器人参考书，书中不仅介绍了聊天机器人的发展历史，还深入介绍了不同类型聊天机器人的技术实现。无论是拥有实体的聊天机器人还是聊天机器人软件，其功能都跳不出闲聊、问答、对话和主动交互 4 种。不同类型的聊天机器人的侧重点不同，但终极目标都是拥有自我感知能力，并能像人一样进行情感交互。本书涵盖的范围比较广泛，但受限于时间和精力，对某些特定的技术点，我们仅仅给出了简要介绍（例如语音识别和语音合成技术），而将主要精力放在了在文字层面聊天机器人如何进行交互上。

对于工业界的朋友，希望本书能够在您寻找特定技术的时候提供一定帮助；对于学术界的专家，本书给出的很多难题也期待着您在理论上加以研究并寻求突破。

本书分7章。第1章简要介绍了聊天机器人的发展和分类，第2章给出了聊天机器人的技术体系介绍，第3章到第5章分别介绍了3种不同类型的聊天机器人的技术实现（问答、对话和闲聊），第6章给出了聊天机器人系统评测的相关信息，第7章提出了聊天机器人进一步发展所面临的技术挑战和展望。下面简要介绍每章的具体内容：

**第1章“聊天机器人概述”** 在本章中，我们追溯了聊天机器人的发展历史，并阐述了聊天机器人的分类和应用场景，从技术层面给出了一个典型的聊天机器人应该包含的技术框架，同时着重介绍了最具代表性的聊天机器人产品。

**第2章“聊天机器人技术原理”** 在本章中，我们从技术的角度，详细介绍了一个文字型交互聊天机器人涉及的技术，包括自然语言理解、对话管理和自然语言生成。我们不仅介绍了传统的自然语言处理技术，也给出了深度学习在解决同类问题上的研究进展。同时，引出了跨越认知智能的关键技术之一——知识图谱，通过不同的例子，阐述了从构建到应用知识图谱的过程。

**第3章“问答系统”** 在本章中，我们介绍了聊天机器人的一种形式——问答。对于某一问题，问答系统旨在获取其精准答案。我们重点介绍了基于知识库的问答系统，阐述了构建知识库所需的技术，并给出了IBM Watson问答系统的详细说明。同时，我们介绍了4种主流的问答方法，包括模板匹配、语义解析、图遍历和深度学习。最后，给出了一个问答系统的具体实践案例。

**第4章“对话系统”** 在本章中，我们主要介绍面向任务的对话系统。与问答系统不同，面向任务的对话系统旨在完成用户指定的一项特定任务。从技术的层面，我们分别介绍了自然语言理解、对话状态跟踪、对话策略学习及自然语言生成，同时穿插了具体的案例，让读者可以有更直观的理解。

**第5章“闲聊系统”** 在本章中，我们分别介绍了闲聊系统的两种实现方

式，一种基于对话库检索，另一种基于生成模型。我们不仅介绍了技术的最新进展，还给出了具体的实现案例。

**第6章“聊天机器人系统评测”** 在本章中，我们梳理了目前国内外聊天机器人评测的公开会议、数据集和进展，并分别针对问答系统和对话系统给出了详细的评测介绍。

**第7章“聊天机器人挑战与展望”** 在本章中，我们给出了聊天机器人发展到现阶段所面临的挑战，并对未来不同场景的应用进行了展望。同时，对聊天机器人发展的下一代范式——虚拟生命，给出了我们的见解和期望。

本书是集体智慧的结晶，写作成员包括王昊奋、邵浩、李方圆、张凯、宋亚楠。同时，感谢很多同事和朋友在写作过程中给予的协助。在写作过程中，我们从实际出发，考虑搭建一个聊天机器人所需要的技术应该是什么样的，同时，关注国内外关于聊天机器人、自然语言处理、知识图谱、机器学习的最新进展，并思考如何将这些技术真正应用于构建聊天机器人中。需要说明的是，在写作过程中，我们参阅了很多领域专家的资料，并尽可能地将所有参考资料都列出了。如果您发现某些内容有争议，请联系我们。

尤其要感谢郑柳洁编辑，没有她的督促和协助，本书不可能有这样的完成度。

## 拥抱人工智能时代

最近几年，技术的飞速发展让我们每个人都无比兴奋。我们也很激动地看到 AI 巨头不断地开源最新、最快的模型，例如谷歌开源了语言模型 BERT，已经在所有 benchmark 数据集上取得了突破。“工欲善其事，必先利其器”，这些强大的算法和工具，让人工智能领域从业者可以创造出更多、更好的产品。在人工智能的发展过程中，我们希望贡献自己的微薄之力。如果读者能够在阅读的过程中获得一点灵感，也将让我们无比欣慰。

愿意创造下一代聊天机器人范式的朋友，我们非常诚挚地邀请你们，一起创造出让人惊艳的、跨越感知智能和认知智能的产品！



# 目 录

1	聊天机器人概述.....	1
1.1	聊天机器人的发展历史.....	1
1.2	聊天机器人的分类与应用场景.....	6
1.3	聊天机器人生态介绍.....	9
1.3.1	典型聊天机器人框架介绍.....	11
1.3.2	聊天机器人平台介绍.....	13
1.3.3	典型的聊天机器人产品介绍.....	13
1.4	参考文献.....	19
2	聊天机器人技术原理.....	20
2.1	自然语言理解.....	21
2.1.1	自然语言理解概览.....	23
2.1.2	自然语言理解基本技术.....	26
2.1.3	自然语言表示和基于深度学习的自然语言理解.....	36
2.1.4	基于知识图谱的自然语言理解.....	46
2.2	自然语言生成.....	56
2.2.1	自然语言生成综述.....	56
2.2.2	基于检索的自然语言生成.....	58
2.2.3	基于模板的自然语言生成.....	59

2.2.4	基于深度学习的自然语言生成.....	60
2.3	对话管理 .....	61
2.4	参考文献 .....	65
3	问答系统 .....	67
3.1	问答系统概述 .....	67
3.2	KBQA 系统 .....	71
3.2.1	KBQA 系统简介 .....	71
3.2.2	主流的 KBQA 方法 .....	79
3.3	KBQA 系统实现 .....	96
3.3.1	系统简介 .....	96
3.3.2	模块设计 .....	97
3.4	参考文献 .....	105
4	对话系统 .....	109
4.1	对话系统概述 .....	109
4.2	对话系统技术原理 .....	113
4.2.1	NLU 模块 .....	115
4.2.2	DST 模块 .....	120
4.2.3	DPL 模块 .....	121
4.2.4	NLG 模块 .....	126
4.3	基于聊天机器人平台搭建对话系统.....	126
4.3.1	NLU 模块实现 .....	129
4.3.2	DST 与 DPL 模块实现 .....	130
4.3.3	NLG 模块实现 .....	131
4.4	面向任务的对话系统实现.....	132
4.5	参考文献 .....	137
5	闲聊系统 .....	139
5.1	闲聊系统概述 .....	139
5.2	基于对话库检索的闲聊系统.....	140
5.2.1	基于对话库检索的闲聊系统介绍.....	140

5.2.2	对话库的建立 .....	143
5.2.3	基于检索的闲聊系统实现.....	145
5.3	基于生成的闲聊系统.....	150
5.3.1	基于生成的闲聊系统介绍.....	150
5.3.2	生成式闲聊系统的新发展.....	152
5.3.3	基于生成的闲聊系统实现.....	155
5.4	参考文献 .....	157
<b>6</b>	<b>聊天机器人系统评测 .....</b>	<b>159</b>
6.1	问答系统评测 .....	159
6.1.1	问答系统评测会议 .....	160
6.1.2	问答系统评测数据集.....	171
6.1.3	问答系统评测标准 .....	173
6.2	对话系统评测 .....	174
6.2.1	对话系统评测会议 .....	176
6.2.2	对话系统评测数据集.....	177
6.2.3	对话系统评测标准 .....	178
6.3	闲聊系统评测 .....	179
6.3.1	闲聊系统评测介绍 .....	179
6.3.2	闲聊系统评测标准 .....	180
6.4	参考文献 .....	183
<b>7</b>	<b>聊天机器人挑战与展望 .....</b>	<b>185</b>
7.1	开放式挑战 .....	185
7.2	技术与应用展望 .....	187
7.3	从聊天机器人到虚拟生命.....	190
7.4	参考文献 .....	193

# 1

## 聊天机器人概述

### 1.1 聊天机器人的发展历史

聊天机器人，是一种通过自然语言模拟人类，进而与人进行对话的程序。它既可以在特定的软件平台（如 PC 平台或者移动终端设备）上运行，也可以在类人的硬件机械体上运行。聊天机器人已经有近 70 年的发展历史，让我们一同回顾半个多世纪以来，不同历史阶段典型的聊天机器人项目和产品。

#### 1. 聊天机器人溯源及发展（1950—1990 年）

对聊天机器人的研究可以追溯到 1950 年图灵（Alan M. Turing）在 *Mind* 期刊上发表的文章 *Computing Machinery and Intelligence*，这篇文章开篇就提出了“机器能思考吗？（Can machines think?）”的设问，然后提出通过让机器参与模仿游戏（Imitation Game）来验证“机器”能否进行“思考”，进而提出了经典的图灵测试（Turing Test）。通过图灵测试被认为是人工智能研究的终极

目标<sup>①</sup>，图灵本人也因而被称为“人工智能之父”。

已知的发布最早的聊天机器人程序 ELIZA<sup>[1]</sup>诞生于 1966 年，由麻省理工学院（MIT）的约瑟夫·魏泽鲍姆（Joseph Weizenbaum）开发。魏泽鲍姆是自然语言处理方面的先驱，他开发的 ELIZA 被看作可以用于临床模拟罗杰斯心理治疗的 BASIC 脚本程序。值得注意的是，尽管 ELIZA 的实现技术仅为对用户输入计算机的话语做关键词匹配，并且其回复规则是由人工编写的（魏泽鲍姆的本意只是让 ELIZA 模仿人类的交谈），但用户与 ELIZA 交谈时却如同自己面对着心理治疗师，开始向 ELIZA 倾诉自己内心深处的想法。随后，魏泽鲍姆撰写了 *Computer Power and Human Reason* 一书，以表达他对人工智能技术的看法。不论怎样，ELIZA 对自然语言处理和人工智能的研究与发展产生了重大影响，全球各地的研究机构也由此开始了对聊天机器人的相关研究。

1972 年，美国精神病学家肯尼思·科尔比（Kenneth Colby）在斯坦福大学（Stanford University）使用 LISP 编写了模拟偏执型精神分裂症表现的计算机程序 PARRY。由于 PARRY 体现的会话策略比魏泽鲍姆的 ELIZA 更严谨更先进，PARRY 被描述为“有态度的 ELIZA”。研究人员在 20 世纪 70 年代早期，使用图灵测试的变体对 PARRY 进行了测试，测试由一组经验丰富的精神科医生参与，这些参与测试的精神科医生通过电传打印机分别与患者和运行 PARRY 的计算机进行对话，并将这些对话记录展示给另一组（33 名）精神科医生。这两组精神科医生分别被要求确定哪些对话是人类患者产生的，哪些是计算机程序产生的。测试结果表明，参与测试的两组精神科医生中只有 48% 在规定时间内做出了正确的判断，正确率约等于随机投票产生的正确率。

1988 年，英国程序员罗洛·卡彭特（Rollo Carpenter）创建了聊天机器人

① 虽然俄罗斯人弗拉基米尔·维西罗夫（Vladimir Veselov）创立的人工智能软件尤金·古斯特曼（Eugene Goostman）在 2014 年就通过了图灵测试，但聊天机器人离真正的“智能”还有很长的路要走。

Jabberwacky。Jabberwacky 项目的目标是“以有趣、娱乐和幽默的方式模拟自然的人际聊天”，这个项目也是通过与人类互动创造人工智能聊天机器人的早期尝试，但 Jabberwacky 并未被用于执行任何其他功能。Jabberwacky 项目于 1997 年正式上线，上线后它会存储所有用户与自己的对话，并且在与用户的对话过程中使用上下文模式匹配技术找到最合适的回复内容。Jabberwacky 并没有硬编码的规则，它完全依赖于反馈的原则，这一点与大多数基于规则约束的聊天机器人非常不同。

也是在 1988 年，加州大学伯克利分校（UC Berkeley）的罗伯特·威林斯基（Robert Wilensky）等人开发了名为 UC（UNIX Consultant）的聊天机器人系统。顾名思义，UC 聊天机器人的目的是帮助用户学习使用 UNIX 操作系统。UC 聊天机器人以英文回复用户，它具备分析用户输入、确定用户需求、给出解决用户需求的规划、决定与用户沟通的内容、根据用户对 UNIX 系统的熟悉程度进行建模等功能。如果说 ELIZA 开启了智能聊天机器人的研究，那么 UC 则真正提高了聊天机器人的智能化程度。

1990 年，美国科学家兼慈善家休·勒布纳（Hugh G. Loebner）设立了人工智能年度比赛——勒布纳奖（Loebner Prize）。勒布纳奖旨在借助交谈测试机器的思考能力，它被看作对图灵测试的一种实践，其比赛的奖项分为金、银、铜三等。根据勒布纳奖的规定，如果参与比赛的程序不仅能通过以文本方式进行的交谈测试，还能在音频和视频测试中过关，则获金奖，赢得 10 万美元和一枚 18K 黄金制金牌，同时勒布纳奖的年度比赛将会中止；如果程序能在以文本方式进行的交谈测试中长时间迷惑住至少半数裁判，则获银奖；如果程序未达到以上标准，则在测试中迷惑住最多裁判的程序赢得 2000 美元和一枚铜奖。从 1991 年首届比赛开办，至本书撰写时，尚无参赛程序达到金奖或者银奖标准。

## 2. 聊天机器人研究兴起（1990—2010 年）

在勒布纳奖的推动下，聊天机器人迎来了研究的高潮，其中较有代表性的聊天机器人系统是 1995 年 12 月 23 日诞生的 ALICE<sup>①</sup>（Artificial Linguistic Internet Computer Entity）。受到相关聊天机器人研究的启发，人工智能科学家理查德·华勒斯（Richard S. Wallace）博士在 1995 年开发了 ALICE 系统，ALICE 可以通过和网民的聊天不断学习，同时基于大量的预置问答模板回答用户的问题。虽然 ALICE 曾在 2000 年、2001 年和 2004 年三次获得勒布纳奖，但其与 ELIZA 并无本质区别（ALICE 仍然完全基于模板匹配技术，不具备逻辑推理能力）。理查德·华勒斯博士将 ALICE 安装到网络服务器，并于 1998 年开源了该项目，目前全世界已有超过 500 个开发者为 ALICE 项目贡献代码。尽管 ALICE 采用的是启发式模板匹配的对话策略，但是它仍然被认为是同类型聊天机器人中性能最好的系统之一。图 1-1 所示为笔者与 ALICE 聊天的截图。值得一提的是，随 ALICE 一同发布的 AIML（Artificial Intelligence Markup Language）目前在移动端虚拟助手的开发中得到了广泛的应用。



图 1-1 笔者与 ALICE 的聊天记录

① <http://www.alicebot.org/>

2001年, SmarterChild 在短信和即时聊天工具中的广泛流行, 使得聊天机器人第一次被应用在了即时通信领域。2006年, IBM 开始研发能够用自然语言回答问题的最强大脑 Watson, 作为一台基于 IBM “深度问答” 技术的超级计算机, Watson 能够采用上百种算法在 3 秒内找出特定问题的答案。

### 3. 聊天机器人研究方兴未艾 (2010 年至今)

2010年, 苹果公司推出了人工智能助手 Siri, Siri 的技术来源于美国国防部高级研究规划局公布的 CALO 计划: 一个简化军方繁复事务, 且具备学习、组织及认知能力的虚拟助理。CALO 计划衍生出来的民用版软件就是 Siri 虚拟个人助理。

此后, 微软小冰、微软 Cortana (小娜)、阿里小蜜、京东 JIMI、网易七鱼等各类聊天机器人层出不穷, 并且这些聊天机器人逐渐渗透进人们生活的各个领域。

2016年, 全球各大公司开始推出可用于聊天机器人系统搭建的开放平台或开源架构。

2010年至今, 标志性的聊天机器人产品如图 1-2 所示。



图 1-2 标志性的聊天机器人产品



聊天机器人的发展历史说明，人类从未放弃将聊天机器人作为人机交互工具的研究。特别是近几年，随着人工智能相关技术“东风”渐起，自然语言处理研究硕果颇丰，聊天机器人相关技术迅速发展。同时，聊天机器人作为一种新颖的人机交互方式，正在成为移动搜索和服务的入口之一，毕竟搜索引擎的最终形态很可能就是聊天机器人。众多人工智能领域的探索者和开发者都想紧紧抓住并抢占聊天机器人这一新的交互入口。

## 1.2 聊天机器人的分类与应用场景

近年来，基于聊天机器人系统的应用层出不穷，下面我们从几个维度对其进行分类。

### 1. 基于应用场景的聊天机器人分类

从应用场景的角度看，可以将聊天机器人分为在线客服、娱乐、教育、个人助理和智能问答 5 类。

在线客服聊天机器人系统的主要功能是自动回复用户提出的与产品或服务相关的问题，以降低企业客服运营成本、提升用户体验。其服务通常是以网站和手机终端为载体而实现的。代表性的商用在线客服聊天机器人系统有小 i 机器人、京东 JIMI 客服机器人、阿里小蜜等。以京东 JIMI 客服机器人为例，用户可以通过与 JIMI 聊天了解商品的具体信息、了解平台的活动信息、反馈购物中存在的问题等。另外，JIMI 具备一定的拒识能力，因此可以知道用户的哪些问题是自己无法回答的，且可以及时将用户转向人工客服。阿里巴巴集团在 2015 年 7 月 24 日发布了一款人工智能购物助理虚拟机器人，取名“阿里小蜜”，阿里小蜜基于客户需求所在的垂直领域（服务、导购、助手等），通过“智能+人工”的方式提供良好的客户体验。