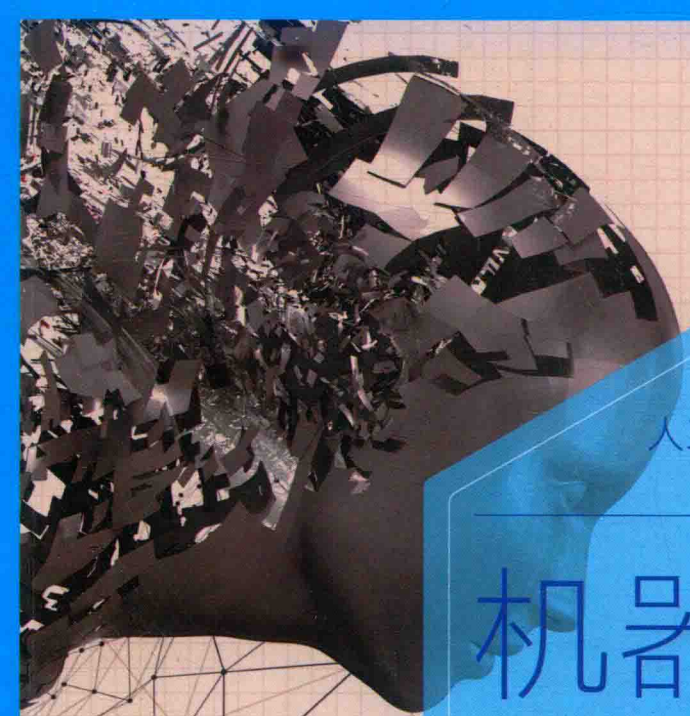
“十三五”国家重点图书出版规划



中国人工智能学会推荐

人工智能丛书

机器翻译

Machine Translation

李沐　刘树杰　张冬冬　周明

高等教育出版社

"十三五"国家重点图书出版规划

中国人工智能学会推荐

人工智能丛书

机器翻译

Machine Translation

李沐　刘树杰　张冬冬　周明

高等教育出版社·北京

内容简介

机器翻译是人工智能，尤其是自然语言处理方向的一个重要研究领域。本书作为该领域的入门书籍，内容上尽可能覆盖机器翻译研究历史上各种主流的研究方法和相关资源。全书分为七章，包括三个主要部分。第一部分（第一、第二章）主要介绍了机器翻译的历史、研究概况和基础知识，第二部分（第三、第四章）详细讨论了统计机器翻译方法的理论和实现，第三部分（第五至七章）则着重介绍了深度学习在机器翻译研究中应用的最新进展，内容包括深度学习的基础知识和在机器翻译中应用深度学习的不同方法。每章后均附有扩展阅读的内容供想深入研究的读者参考。

本书可以作为高等院校计算机类和电子信息类等相关专业的研究生教材，也可供对机器翻译的研究和进展有兴趣的读者和工程技术人员参考。

图书在版编目(CIP)数据

机器翻译/李沐等主编. -- 北京：高等教育出版社，2018.8

（人工智能丛书）

ISBN 978-7-04-050243-5

Ⅰ. ①机… Ⅱ. ①李… Ⅲ. ①机器翻译 Ⅳ. ①H085

中国版本图书馆 CIP 数据核字(2018)第 169690 号

策划编辑 张江漫　　责任编辑 黄涵玥　　封面设计 赵 阳　　版式设计 童 丹

插图绘制 于 博　　责任校对 张 薇　　责任印制 毛斯璐

出版发行 高等教育出版社
社　　址 北京市西城区德外大街 4 号
邮政编码 100120
印　　刷 三河市骏杰印刷有限公司
开　　本 787mm×960mm 1/16
印　　张 15
字　　数 270 千字
购书热线 010-58581118
咨询电话 400-810-0598

网　　址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn

版　　次 2018 年 8 月第 1 版
印　　次 2018 年 8 月第 1 次印刷
定　　价 39.90 元

物 料 号 50243-00

机器翻译

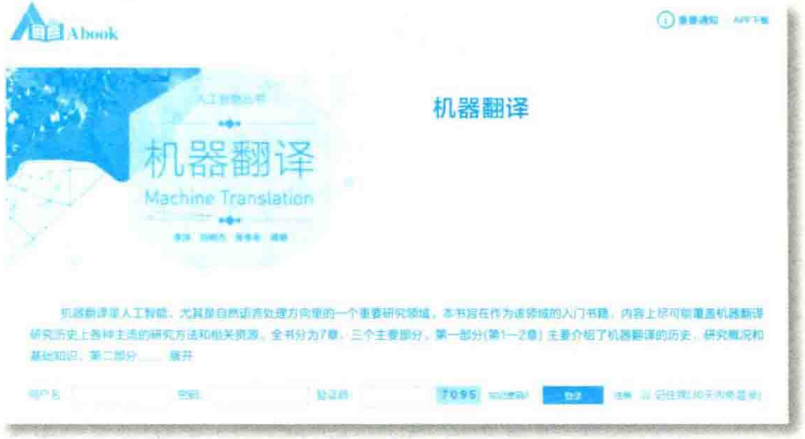
李　沐
刘树杰
张冬冬
周　明

1　计算机访问http://abook.hep.com.cn/1253551，或手机扫描二维码、下载并安装Abook应用。

2　注册并登录，进入“我的课程”。

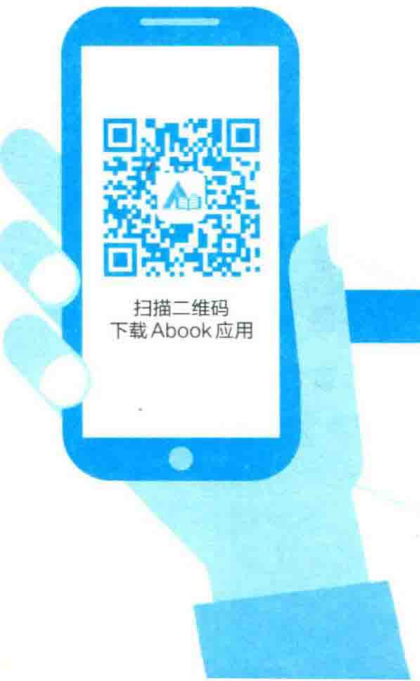
3　输入封底数字课程账号（20位密码，刮开涂层可见），或通过Abook应用扫描封底数字课程账号二维码，完成课程绑定。

4　单击“进入课程”按钮，开始本数字课程的学习。



课程绑定后一年为数字课程使用有效期。受硬件限制，部分内容无法在手机端显示，请按提示通过计算机访问学习。

如有使用问题，请发邮件至abook@hep.com.cn。



http://abook.hep.com.cn/1253551

人工智能丛书编委会

前　言

机器翻译作为人工智能的一个重要分支，自20世纪50年代以来，先后走过了基于规则的系统、基于统计的系统等几个重要阶段，现在步入了神经机器翻译阶段。其在典型领域（比如新闻）的译文质量几乎与人工翻译水平相类似。伴随着语音技术和手机的普及，口语翻译也在快速走向实用化。但是，在数据资源匮乏的语言对或者垂直领域，机器翻译的质量尚不能达到用户期望的水平。因此，为克服数据不足，引入半监督或者无监督的训练。在目前人工智能的浪潮下，机器翻译的理论和技术以及未来发展趋势已成为引人注目的领域。

在此背景下，本书尝试总结机器翻译技术的最新理论、方法和应用，以便帮助广大人工智能领域的研究生和科研技术人员快速掌握机器翻译的关键技术。本书分为七章。第一章回顾机器翻译发展的历史并介绍机器翻译技术的各种应用。第二章介绍如何获取用于机器翻译模型训练的单语和双语数据的方法以及机器翻译自动评测方法。第三章介绍统计机器翻译系统的基础架构、建模方法和基本模型（包括语言模型、翻译模型和调序模型）以及模型的参数训练方法。第四章介绍典型的统计机器翻译系统模型，包括基于短语的、基于形式文法的和基于句法的统计机器翻译模型系统。第五章开始介绍神经机器翻译和深度学习的基础知识，包括感知机、词语嵌入模型、卷积神经网络和循环神经网络。第六章系统介绍神经机器翻译，包括神经联合模型和基于序列映射的神经机器翻译模型以及注意力机制，并介绍基于卷积神经网络的编码器和解码器的神经机器翻译模型以及完全基于注意力网络的模型。第七章进一步深入讨论了神经机器翻译在模型改进、模型训练、翻译解码等方面的前沿进展。

本书的主要特点是：以简明易懂的语言对机器翻译技术给予了全面介绍，兼顾经典的统计机器翻译以及目前飞速发展的神经机器翻译技术。另外，本书注重理论和实践相结合。读者在深入浅出地理解理论体系之后，借助实例和本书所介绍的工具，能够很快入门并掌握机器翻译的训练和解码的主要技术。

本书由微软亚洲研究院自然语言组四位多年来从事机器翻译研究的同事合作编写。感谢清华大学刘洋老师对本书提出的很多宝贵意见。

由于作者的水平有限，书中错漏之处在所难免。欢迎提出宝贵建议，今后再版时一定补正。作者邮箱为：mingzhou@ microsoft.com。

作　者

2018 年 2 月

目　录

第一章
绪　论

进入21世纪以来，伴随着通信技术和互联网技术的快速发展，全球化贸易、科学、技术和文化交流活动日益增强，不同国家或地区、不同族群和不同文化之间的联系越来越紧密。而随着信息量和交流需求急剧增加，克服跨语言交流障碍也显得越来越重要。虽然人工翻译可提供令人满意的翻译服务，但效率低下，代价昂贵，不能满足大量的翻译需求。因此，机器翻译，即基于人工智能技术、通过计算机或终端智能设备进行不同语言的翻译任务，越来越引起人们的广泛重视，并且已经开始服务于人们日常工作、生活中的各种跨语言应用需求。在此背景下，本书尝试总结、讲述机器翻译技术的最新理论、方法、应用，帮助科研技术人员了解、掌握机器翻译的历史发展、关键技术、系统实现方法以及其未来发展趋势。同时，也希望本书中讲述的机器翻译技术能够对解决其他自然语言处理任务（如自动问答系统、对话聊天系统等）有着启发和借鉴作用。

机器翻译是一个任务明确、历史悠长、仍处于研究阶段的课题。本章将讲述机器翻译的基本概念、发展历史，比较不同的机器翻译方法，介绍相关应用。

1.1 机器翻译概述

1.1.1 机器翻译定义

机器翻译(machine translation,MT)是指使用机器(计算机)自动地将一种自然语言(源语言)语句转化为相同含义的另一种自然语言(目标语言)语句的过程。这里的自然语言是指日常使用的人类语言(如英语、汉语、日语等),它区别于人工为某些特定目的而创造的语言(如计算机编程语言)。就语言形态而言,它可以是语音或者文本。语音翻译过程通常也包含文本翻译过程。本书中讲述的机器翻译内容限定于文本翻译。

机器翻译任务是自然语言处理(natural language processing)的一个研究分支,与计算语言学(computational linguistics)、自然语言理解(natural language understanding)之间存在着密不可分的关系。机器翻译的研究和任务处理过程不仅涉及自然语言处理的诸多经典任务,包括数据挖掘、数据清洗、分词、词性标注、句法分析、语义分析等,而且还涉及解码算法、优化算法、建模及训练过程中各种机器学习算法的应用等。因此,机器翻译任务是一个复杂的系统工程。

1.1.2 机器翻译简史

很久以前,人类就开始利用机器来解决语言翻译的问题。从最初的设想到现在大规模的商业应用,机器翻译的发展经历了一个波澜壮阔的历程。各种机器翻译技术方法不断演进,从启蒙尝试,到基于规则、基于实例的方法,再到基于语料库的统计方法,直到当前的神经网络方法,机器翻译质量不断提升,应用场景不断拓展扩大。下面简述机器翻译发展的历史。

(一) 早期机械装置翻译实践阶段

早在古希腊时代,就有人提出利用机械装置进行语言翻译的想法。17 世纪的笛卡尔(Descartes)和莱布尼茨(Leibniz)提出过使用机器词典来实现语言翻译。到了 17 世纪中叶,“普遍语言”运动提出,设计一种中介语,使之成为无歧义的通用语言,它试图对世界上所有概念和实体进行分类和编码,以实现它们在各种语言之间的对应关系。1903 年,古图拉特(Couturat)和洛(Leau)在《通用语言的历史》一书中首次使用了“机器翻译”术语,采用德国学者里格(Rieger)提出的一种数字语法,加上词典的辅助,利用机械装置将一种语言翻译成多种语言。

20 世纪 30 年代初,法国工程师阿尔楚尼(Artsouni)提出基于存储装置进行

语言翻译的想法，利用机械装置上的宽带纸上的代码孔来记录语言之间的词项翻译信息，尝试使用具有检索功能的宽带纸作为机器词典来实现自动翻译，但这一实践因受第二次世界大战的影响而夭折。1933 年，苏联发明家彼得（Peter Troyanskii）提出使用双语字典和语言间的语法角色完成翻译，具体分为三个阶段：第一阶段由人工编辑将源语言词转换成逻辑表达形式；第二阶段利用机器将逻辑表达式翻译成目标语言表达方式；第三阶段引入人工专家将其转成地道的目标语言。但是这个想法在当时没有得到实现。

1946 年，第一台电子计算机 ENIAC（electronic numerical integrator and computer）诞生。之后，信息论的先驱、美国科学家韦弗（W. Weaver）在 1947 年 3 月 4 日写给“控制论之父”诺伯特·维纳（Norbert Wiener）的一封信中，提到利用计算机进行语言自动翻译的想法。1949 年 7 月 15 日，W. Weaver 在题为《翻译》的备忘录中正式提出了这一思想，它对机器翻译思想起到了启蒙作用，并引发了机器翻译研究的兴起。

（二）基于规则和实例方法的机器翻译发展阶段

从 20 世纪 40 年代开始，美国和苏联两个超级大国出于军事、政治和经济目的，均对机器翻译项目提供了大量的资金支持。同时，欧洲国家由于地缘政治和经济的需要也对机器翻译研究给予了相当大的重视。1954 年，美国乔治敦大学（Georgetown University）在 IBM 公司的协同下，使用 IBM-701 计算机开发了世界上第一个机器翻译原型系统，利用 6 条翻译规则和 250 个词的词典进行了俄语到英语的翻译试验，并向公众展示了机器翻译的可行性。中国早在 1956 年就把机器翻译研究列入了全国科学工作发展规划。1957 年，中国科学院语言研究所与计算技术研究所合作开展了俄汉机器翻译试验，翻译了 9 种不同类型的句子。这一阶段机器翻译研究均采用基于人工编制规则的词法分析、句法分析系统，构造了规模较小的实验系统。这一时期的机器计算能力有限，也缺乏机器可读的大规模语言语料库。因此，基于规则的翻译系统全靠人工编撰规则和词典，可扩展性差，且系统计算能力所限，导致翻译系统难以进一步发展。

20 世纪 50 年代，人们对机器翻译的高度期待和乐观主义情绪开始弥漫，但是低估了自然语言和翻译本身的复杂性以及计算机的局限性。随着若干机器翻译系统被陆续研制出来并投入使用，人们得以直接观察和评价机器翻译系统的输出结果。人们观察得到的总体印象是：机器翻译的质量与期望相差甚远。1964 年，为了对机器翻译的研究进展作出评价，美国科学院成立了语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，ALPAC），开始了为期两年的综合调查分析和测试。1966 年 11 月，该委员会公布了一个题为《语言与机器：翻译和语言学视角下的计算机》（Language and Machines — Com-

puters in Translation and Linguistics)的报告(简称 ALPAC 报告)。该报告指出当时的机器翻译代价昂贵,翻译精度低,速度慢于人工翻译,未来不会达到人工翻译质量。该报告相当于全面否定了机器翻译的可行性,并建议停止对机器翻译项目的资金支持。这一报告的发表给了正在蓬勃发展的机器翻译当头一棒,世界范围内机器翻译热潮突然消失,从波峰深深地跌入了波谷。

进入 20 世纪 70 年代,随着科学技术的发展和各国科技情报交流的日趋频繁,国与国之间的语言障碍显得更为严重,传统的人工作业方式已经远远不能满足需求,人们迫切地需要计算机来从事翻译工作。随着乔姆斯基(Chomsky)语言学理论和人工智能研究的发展,人们意识到要想实现好的翻译效果,必须在理解语言的基础上进行翻译,从理解句法结构上下功夫,为此,基于规则的机器翻译研究开始展开。同时,计算机硬件技术的大幅度提高以及人工智能在自然语言处理上的应用,从技术层面推动了机器翻译研究的复苏,机器翻译项目又开始发展起来,各种实用的以及实验的系统被先后推出,例如 Weinder 系统、EURPOTRA 多国语翻译系统、TAUM-METEO 系统等。1968 年成立的 SYSTRAN 公司,其开发的英法、俄英翻译系统分别于 1976 年、1979 年应用于欧盟、美国空军,后来被安装在北约和国际原子能机构。加拿大蒙特利尔大学研发的 METEO 英-法机器翻译系统,于 1977 年被成功用于翻译天气预报文档。20 世纪 80 年代,机器翻译在日本掀起了一次"小高潮"。在 1982 年日本提出"五代机"计划的大背景下,不少日本大企业纷纷投资开展机器翻译的研发。这一时期由于受到应用需求的驱动,机器翻译技术研究受到了越来越多的来自社会各方面的关注。

随着机器翻译技术的应用发展,基于规则的机器翻译系统开发逐渐遇到了困难。基于人工确定的有限翻译规则越来越复杂、规模库越来越大,但对复杂的不断变化的语言现象的解释能力却难以继续提高,译文的准确率也未有持续的改善。自 20 世纪 80 年代开始,研究人员逐渐尝试全新的基于数据驱动的机器翻译方法。1980 年马丁(Martin Kay)提出了翻译记忆(translation memory,TM)的方法,其基本思想是在翻译句子时从已经翻译好的文档中找出相似部分来帮助新的翻译。日本的长尾真(Makoto Nagao)于 1984 年提出了基于实例的机器翻译方法(example-based machine translation,EBMT),它从实例库中提取翻译知识,通过增、删、改、替换等操作完成翻译。

(三)统计机器翻译发展阶段

1990 年在芬兰赫尔辛基召开的第 13 届国际计算语言学大会提出了处理大规模真实文本的战略任务,开启了语言计算的一个新的历史阶段——基于大规模语料库的统计自然语言处理。在此潮流的带动下,20 世纪 90 年代 IBM 研究院的研究人员在著名的文章《统计机器翻译的数学理论:参数估计》(The math-

ematic of statistical machine translation: Parameter estimation)(Brown et al., 1993)中提出了 IBM Model 1—5。它基于香农(Shannon)信息论中针对编解码的"噪声-信道模型",支持词到词的统计机器翻译。在 1999 年的约翰·霍普金斯大学(The Johns Hopkins University)夏季讨论班上,一批研究人员发布了 GIZA 软件包,实现了 IBM Model 1 到 IBM Model 5。随后,弗兰茨·约瑟夫(Franz-Joseph Och)对 GIZA 进行了优化,提出了更加复杂的 Model 6,并发布了新的软件包 GIZA++。该软件包在构建统计机器翻译系统中得到了广泛使用。

基于词的统计机器翻译模型处理的单元小,不能很好地对上下文建模。后来逐渐发展起来的基于短语的方法成为统计机器翻译的主流工作。期间,三项重要的工作极大地推动了统计机器翻译的发展:对数-线性模型,参数最小错误训练方法,BLEU 评测指标。Och(2003)提出的基于最大熵的对数-线性模型和参数最小错误训练(minimum error rate training)方法促使统计机器翻译方法能够将多种不同的特征函数融合进机器翻译模型中,并且自动学习它们各自的特征权重,使得翻译性能显著超越了其他传统机器翻译方法。此外,自动评测指标 BLEU(Papineni et al., 2002)的提出不仅避免了人工评价成本昂贵的弊端,而且可以直接成为模型优化的目标,极大地提高了统计机器翻译系统模型训练、迭代、更新的效率。

基于普通短语的统计机器翻译模型对于句子中具有长距离依赖关系的词语翻译以及词语调序处理方面存在问题。为此,基于层次化短语(Chiang, 2007)和句法的翻译模型相继被提出。与此同时,多系统融合、领域自适应等翻译问题也引起了研究人员的关注。

统计机器翻译方法的特点是几乎完全依赖对大规模双语语料库的自动学习、自动构造机器翻译系统。这种方法具有广泛的一般性,与具体语种无关,也不再需要人工规则集。一些研究团体、机构开始在网上开源机器翻译系统软件以促进统计机器翻译的学术研究,其中比较著名的是 Moses 系统[资源1],常作为学术论文中的基线系统。

21 世纪初期开始,借助于互联网的发展,统计机器翻译系统开始走向民用。以 IBM、微软、谷歌为代表的科研机构和企业均相继成立机器翻译团队。几年后,微软、谷歌、百度等各大 IT 公司都相继发布了能够支持世界上几十种常用语言的互联网机器翻译系统,迅速普及了机器翻译的应用场景,极大地提高了人们使用机器翻译的便利性。

2012 年时任微软研究部门全球负责人的里克·拉希德(Rick Rashid)在中国天津召开的"21 世纪计算会议"上现场演示了一个语音机器翻译的项目。当