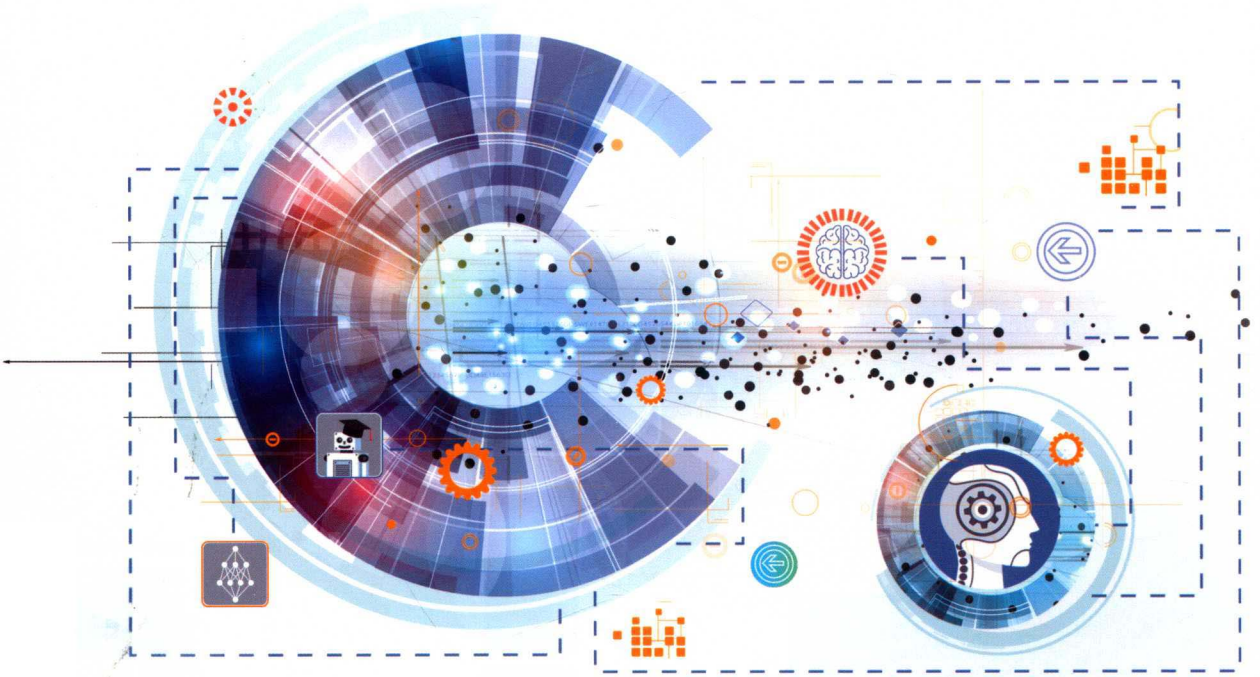
HZ BOOKS 华章IT
apress®
智能系统与技术丛书

通过解决丢弃层、池化层和归一化层的难题，探索并开发你自己的深度学习网络
获得关于强化学习以及如何使用上下文特定的行为令人兴奋的介绍
使用TensorFlow和Keras基于叠加双向LSTM创建自己的聊天机器人

earning for Natural Language Processing
Creating Neural Networks with Python

面向自然语言处理的深度学习
用Python创建神经网络

帕拉什·戈雅尔（Palash Goyal）
[印] 苏米特·潘迪（Sumit Pandey） 著
卡兰·贾恩（Karan Jain）
陶阳 张冬松 徐潇 译



机械工业出版社
China Machine Press

智能系统与技术丛书

Deep Learning for Natural Language Processing
Creating Neural Networks with Python

面向自然语言处理的深度学习
用Python创建神经网络

帕拉什·戈雅尔（Palash Goyal）
[印] 苏米特·潘迪（Sumit Pandey） 著
卡兰·贾恩（Karan Jain）

陶阳 张冬松 徐潇 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

面向自然语言处理的深度学习：用 Python 创建神经网络 /（印）帕拉什·戈雅尔（Palash Goyal）等著；陶阳，张冬松，徐潇译. —北京：机械工业出版社，2019.1
（智能系统与技术丛书）
书名原文：Deep Learning for Natural Language Processing: Creating Neural Networks with Python

ISBN 978-7-111-61719-8

I. 面… II. ①帕… ②陶… ③张… ④徐… III. ①自然语言处理 ②人工神经网络－软件工具－程序设计 IV. ① TP391 ② TP183

中国版本图书馆 CIP 数据核字（2019）第 004119 号

本书版权登记号：图字 01-2018-5459

面向自然语言处理的深度学习
用 Python 创建神经网络

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：杨宴蕾　　责任校对：李秋荣
印　　刷：中国电影出版社印刷厂　　版　　次：2019 年 2 月第 1 版第 1 次印刷
开　　本：186mm×240mm　1/16　　印　　张：13.25
书　　号：ISBN 978-7-111-61719-8　　定　　价：69.00 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）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

THE TRANSLATOR'S WORDS

译 者 序

很高兴接受这本书的翻译任务，更加高兴的是能跟读者说说心里话。

接到任务之前，我和团队正在进行人机对话领域的研究与实现工作。我们的导师是国防科技大学计算机学院彭宇行研究员，在他的带领下，我们已经取得了不少成果。这些年导师带领我们在人工智能领域开展了多项研究工作，深度学习算法是其中很重要的一项。所以，我们有较强的理论和实践基础，尤其是近一年来我们一直致力于基于自然语言处理的对话机器人项目的落地实现。正好在这个时候接到了这本书的翻译任务，这种感觉太好了。

拿到这本书之后，我仔细阅读了书的内容，被作者由浅入深、从理论到实践的创作思路深深吸引住了。前面章节对深度学习和自然语言处理的基础理论知识的介绍十分深入，而且强调实用性，也为后面章节的完整实例做好了准备。虽然该书面向的读者是中高级深度学习和自然语言处理开发人员，但读完后你就会感到，即使你在深度学习和自然语言处理方向的知识积累还不够多，也可以立刻开发出一个实际的基于自然语言处理的应用项目。

我们知道，采用计算机技术来研究和处理自然语言是20世纪40年代末50年代初才开始的。几十年来，这项研究取得了长足的进展，成为计算机科学中一门重要的新兴学科——自然语言处理（Natural Language Processing，NLP）。建立自然语言处理模型需要各个方面的知识，比如声学和韵律学、音位学、形态学、词汇学、句法学、语义学、话语分析、语用学等。由于自然语言处理是一个多边缘的交叉学科，

除了语言学外，它还涉及多个知识领域，比如计算机科学、数学、统计学、生物学等，尤其是计算机科学领域的深度学习。深度学习是机器学习的一个扩展领域，它已经在文本、图像和语音等领域发挥了巨大作用。在深度学习下实现的算法集合与人脑中的刺激和神经元之间的关系具有相似性。深度学习在计算机视觉、语言翻译、语音识别、图像生成等方面具有广泛的应用。大多数深度学习算法都基于人工神经网络的概念，如今，大量可用的数据和丰富的计算资源使这种算法的训练变得简单了。随着数据量的增长，深度学习模型的性能会不断提高。这些看似高深的理论知识，本书在第 1 章用很浅显的语言告诉了我们。我向我们的项目团队推荐了这本书，他们反映这本书写得很好，实践性强，特别适用于自然语言处理相关的项目。我就想，如果能把这本优秀的教材翻译成中文，肯定能让国内从事自然语言处理的年轻工程师们从中受益。于是，我欣然接受了本书的翻译任务。

我认真通读本书两遍，对于本书有一定的理解后试着翻译起来，然而不像想象中那样容易。说实话，有时候习惯了阅读英文，即使理解了英文意思，而要把英文的意思表达为确切的中文，大量的术语如何用中文来表达，也是颇费周折、令人踌躇的难题。另外，为了能够及早让本书与读者见面，在我们的导师彭宇行研究员的大力支持下，我跟团队成员商量，请具有较好的自然语言处理方向研究基础的张冬松博士后和海归硕士徐潇两位骨干加入进来，共同完成翻译任务。

接到翻译任务的时候正是暑假，所以我们几乎用了全部的假期时间来进行翻译。暑假之后，每位译者依然十分认真，所有的业余时间都用上了，遇到疑难问题时共同切磋、反复推敲，经常在微信群里讨论，确定最好的翻译结果。每翻译一章，他们就交给我审校，及时统一意见。在我们三人的通力合作下，连续工作两个多月，全书的翻译任务终于大功告成。

这么一本优秀的著作在给我们带来无穷动力的同时，无疑也给翻译工作带来了无形的压力。为了尽量保证每章译稿的质量并保证译文的前后一致性，整本书的审校工作全部由我本人独立完成，同时我及时反馈并提供了统一的术语翻译。在翻译过程中我们阅读了大量相关的教材和论文，也包括网上的常用译法以及公认的英文术语，并

前后进行了六次自我校对。在校对过程中，有很多师门的兄弟姐妹们也提出了很多宝贵的意见和建议。对于他们无私的帮助，我表示由衷的感谢。感谢我们的导师彭宇行研究员对我们的翻译工作给予的支持和肯定。另外还要感谢我的妻子，在前前后后两个多月时间里，我几乎所有的时间都用在翻译和校对上，而她却默默地承担起照顾两个孩子的责任。

虽然得到了大家的帮助，翻译团队也认真努力，但由于我们的专业水平、理解能力和文字功底十分有限，加之时间仓促，最后的译稿中一定还存在不少理解上的偏差，译文也会有生硬之处。希望读者不吝赐教，提出宝贵的修改意见和建议，以便我们能够对现有译稿不断改进。

谢谢！

陶　阳

PREFACE

前　言

本书使用适当和完整的神经网络体系结构示例，例如用于自然语言处理（NLP）任务的循环神经网络（RNN）和序列到序列（seq2seq），以较为全面的方式简化和呈现深度学习的概念。本书试图弥合理论与应用之间的缺口。

本书以循序渐进的方式从理论过渡到实践，首先介绍基础知识，然后是基础数学，最后是相关示例的实现。

前三章介绍 NLP 的基础知识，从最常用的 Python 库开始，然后是词向量表示，再到高级算法，例如用于文本数据的神经网络。

最后两章完全侧重于实现，运用广泛流行的 Python 工具 TensorFlow 和 Keras，处理诸如 RNN、长短期记忆（LSTM）网络、seq2seq 等复杂架构。我们尽最大努力遵循循序渐进的方法，最后集合全部知识构建一个问答系统。

本书旨在为想要学习面向 NLP 的深度学习技术的读者提供一个很好的起点。

本书中展示的所有代码都在 GitHub 上以 IPython notebook 和脚本的形式公开，使读者能够实践这些示例，并以自己感兴趣的任何方式对它们进行扩展。

ACKNOWLEDGEMENTS

致　谢

这项工作得以完成，主要得益于那些信任我们的人，他们在整个工作过程中参与讨论、阅读、写作，付出了宝贵的时间，无论是校对还是整体设计，都是功不可没的。

我们特别感谢 Apress 的协调编辑 Aditee Mirashi，他一直支持和激励我们完成这项任务，并积极地为我们提供有价值的建议，以便我们按时实现目标。

我们感谢 Santanu Pattanayak，他阅读了所有章节并提出了宝贵的意见，为本书的最终定稿做了大量工作。

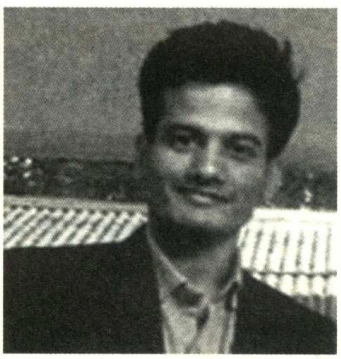
在完成这个项目的过程中，没有人比我们的家庭成员更重要。我们要感谢我们的父母，无论我们追求什么，他们的爱和指导都与我们同在。他们是我们的终极榜样，为我们完成写作提供了无尽的灵感。

ABOUT THE AUTHORS

关于作者

Palash Goyal 是一名高级数据科学家，目前从事将数据科学和深度学习用于在线营销领域的工作。他曾在印度理工学院（IIT）的 Guwahati 分校学习数学和计算机科学，毕业后他开始在快节奏环境中工作。

他在电子商务、旅游、保险和银行等行业拥有丰富的经验，热衷于数学和金融。他利用深度学习和强化学习技术进行价格预测与投资组合管理，在业余时间管理他的多种加密货币和最新的首次代币发行（ICO）。他追踪数据科学领域的最新趋势，并在博客 http://madoverdata.com 上分享这些趋势。他还会在空闲时发表与智慧农业相关的文章。



Sumit Pandey 毕业于印度理工学院（IIT）的 Kharagpur 分校，曾在 AXA Business Services 工作了大约一年，担任数据科学顾问。他目前正在创办自己的企业。

Karan Jain 是 Sigtuple 公司的一名产品分析师，他在那里研究尖端的 AI 驱动诊断产品。此前，他曾在医疗保健解决方案公司 Vitrana 担任数据科学家。他喜欢在快节奏的数据初创公司工作。在闲暇时间，Karan 深入涉猎基因组学、BCI 接口和光遗传学。最近，他对用于便携式诊断的 POC 设备和纳米技术产生了兴趣。Karan 在 LinkedIn 上有 3000 多名粉丝。

ABOUT THE TECHNICAL REVIEWER

关于技术审校人员

Santanu Pattanayak 目前在 GE Digital 担任数据科学家，并且是深度学习相关书籍《Pro Deep Learning with TensorFlow-A Mathematical Approach to Advanced Artificial Intelligence in Python》的作者。他拥有大约 12 年的工作经验，其中 8 年从事数据分析 / 数据科学工作，他还具有开发和数据库的技术背景。

在加入 GE 之前，Santanu 曾在 RBS、Capgemini 和 IBM 等公司任职。他毕业于加尔各答的 Jadavpur 大学，获得了电子工程专业学位，并且是一名狂热的数学爱好者。Santanu 目前正在攻读印度理工学院（IIT）Hyderabad 分校的数据科学硕士学位。他还将时间投入数据科学黑客马拉松和 Kaggle 比赛中，他在 Kaggle 中全球排名前 500 位。Santanu 出生并成长在印度西孟加拉邦，现与妻子居住在印度班加罗尔。

CONTENTS

目　录

01 CHAPTER

第 1 章
自然语言处理和深度学习概述

自然语言处理（NLP）是计算机科学中一项极其困难的任务。语言中存在各式各样的问题，这些问题因语言而异。如果能采用正确的方式从原始文本中构建或提取出有意义的信息，将是一个很好的解决方案。以前，计算机科学家会使用复杂的算法将语言分解为语法形式，例如词类、短语等。如今，深度学习是达到相同目的的关键。

本章将探讨 Python 语言、NLP 和深度学习的基础知识。首先会介绍 Pandas、NumPy 和 SciPy 库中的初级代码，我们假设用户已经配置好初始的 Python 环境（2.x 或 3.x），并指导用户安装上述库。然后，将简要讨论 NLP 中常用的库，以及一些基本示例。最后，我们将讨论深度学习背后的概念和一些常见的框架，例如 TensorFlow 和 Keras。在此后的各章中，我们将继续介绍更高级的 NLP 主题。

取决于计算机和版本的具体情况，用户可以使用以下链接安装 Python：

- www.python.org/downloads/
- www.continuum.io/downloads

上述链接和基本软件包安装将为用户提供深度学习所需的环境。

我们首先会用到以下这些包，请参考以下括号中的链接：

Python 机器学习：

Pandas (http://pandas.pydata.org/pandas-docs/stable)

NumPy (www.numpy.org)

SciPy (www.scipy.org)

Python 深度学习：

TensorFlow (http://tensorflow.org/)

Keras (https://keras.io/)

Python 自然语言处理：

Spacy (https://spacy.io/)

NLTK (www.nltk.org/)

TextBlob (http://textblob.readthedocs.io/en/dev/)

我们可能会在需要时安装其他相关软件包。如果在安装过程中遇到问题，请参阅以下链接：https://packaging.python.org/tutorials/install-packages/。

请参阅 Python 包索引 PyPI（https://pypi.python.org/pypi），以搜索最新的可用包。请按以下链接中的步骤安装 pip：https://pip.pypa.io/en/ stable/installing/。

1.1　Python 包

我们将介绍 Pandas、NumPy 和 SciPy 包的安装步骤和初级编码。目前，Python 提供的版本是 2.x 和 3.x，它们具有用于机器学习的兼容功能。我们将在需要时使用 Python2.7 和 Python3.5。版本 3.5 已在本书的各章中广泛使用。

1.1.1　NumPy

NumPy 专门用于 Python 中的科学计算。它能够高效地操纵含有随机记录的大型多维数组，并且速度与处理小型多维数组几乎一样快。它也可以当作通用数据的多维容器。NumPy 具有创建任意类型数组的能力，这使它适合与通用数据库应用程序连

接，也使其成为在本书中或以后使用的最有用的库之一。

以下是使用 NumPy 包的代码。大多数代码行都附有注释，使用户能更容易地理解。

```
## Numpy

import numpy as np                    # Importing the Numpy package
a= np.array([1,4,5,8], float)         # Creating Numpy array with
                                        Float variables
print(type(a))                  #Type of variable
> <class 'numpy.ndarray'>

# Operations on the array
a[0] = 5                  #Replacing the first element of the array
print(a)
> [ 5. 4. 5. 8.]

b = np.array([[1,2,3],[4,5,6]], float)   # Creating a 2-D numpy
                                           array
b[0,1]                  # Fetching second element of 1st array
> 2.0

print(b.shape)          #Returns tuple with the shape of array
> (2, 3)

b.dtype                 #Returns the type of the value stored
> dtype('float64')

print(len(b))           #Returns length of the first axis
> 2

2 in b                  #'in' searches for the element in the array
> True

0 in b
> False
# Use of 'reshape' : transforms elements from 1-D to 2-D here
c = np.array(range(12), float)
print(c)
print(c.shape)
print('---')
c = c.reshape((2,6))    # reshape the array in the new form
print(c)
print(c.shape)
```