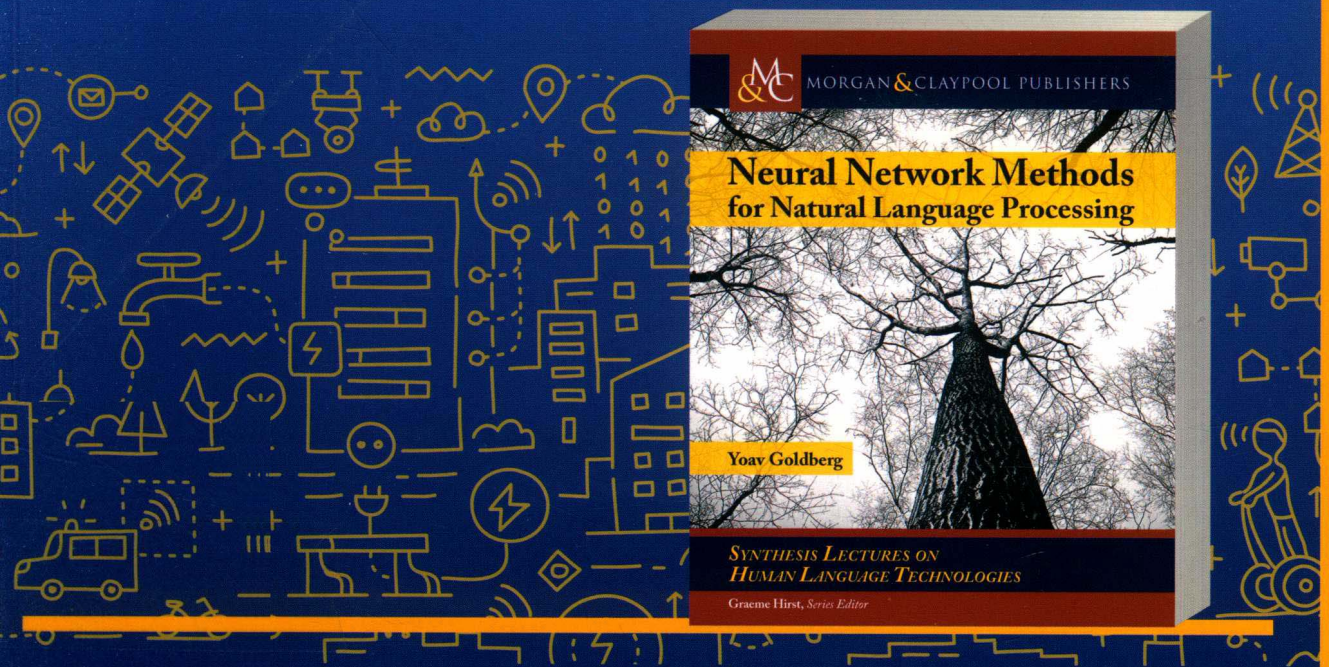
HZ BOOKS
华章科技

智能科学与技术丛书

Neural Network Methods for Natural Language Processing

基于深度学习的自然语言处理

[以色列] 约阿夫·戈尔德贝格（Yoav Goldberg） 著
车万翔 郭江 张伟男 刘铭 译
刘挺 主审



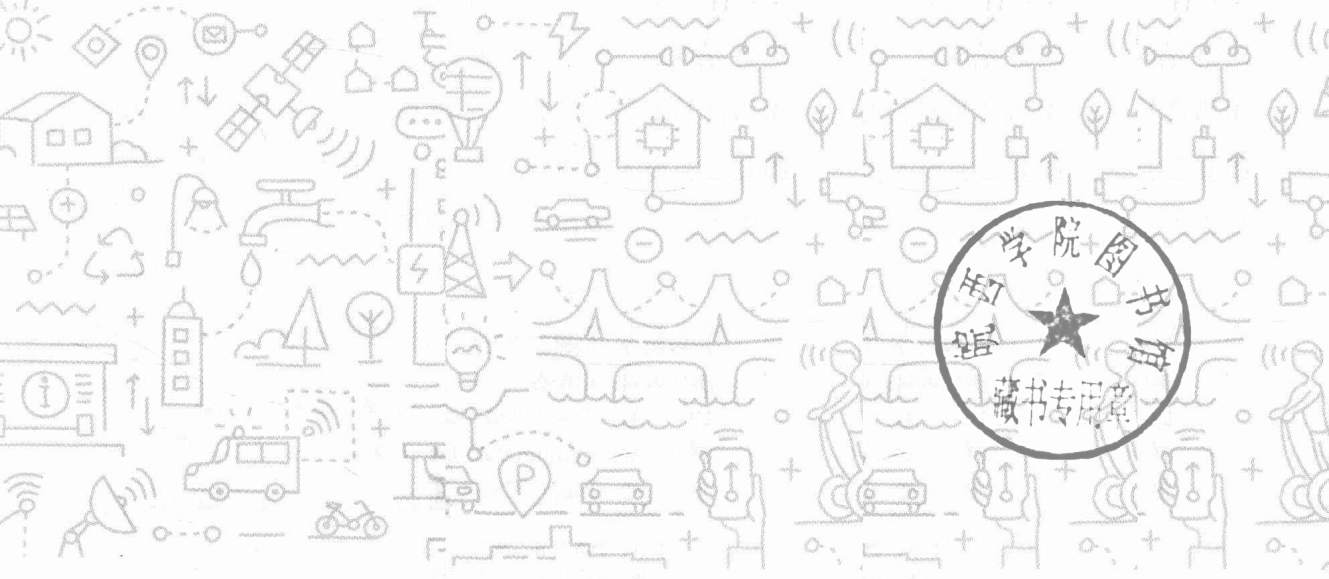
机械工业出版社
China Machine Press

智能科学与技术丛书

Neural Network Methods for Natural Language Processing

基于深度学习的自然语言处理

[以色列] 约阿夫·戈尔德贝格（Yoav Goldberg）著
车万翔 郭江 张伟男 刘铭 译
刘挺 主审



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

基于深度学习的自然语言处理 /（以）约阿夫·戈尔德贝格（Yoav Goldberg）著；车万翔等译 . —北京：机械工业出版社，2018.3（2018.8 重印）
（智能科学与技术丛书）
书名原文： Neural Network Methods for Natural Language Processing

ISBN 978-7-111-59373-7

I. 基… II. ① 约… ② 车… III. 自然语言处理 IV. TP391

中国版本图书馆 CIP 数据核字（2018）第 048795 号

本书版权登记号：图字 01-2017-6462

本书重点介绍了神经网络模型在自然语言处理中的应用。首先介绍有监督的机器学习和前馈神经网络的基本知识，如何将机器学习方法应用在自然语言处理中，以及词向量表示（而不是符号表示）的应用。然后介绍更多专门的神经网络结构，包括一维卷积神经网络、循环神经网络、条件生成模型和基于注意力的模型。最后讨论树形网络、结构化预测以及多任务学习的前景。

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：迟振春　　责任校对：殷 虹
印　刷：北京市兆成印刷有限责任公司　　版　次：2018 年 8 月第 1 版第 3 次印刷
开　本：185mm × 260mm 1/16　　印　张：17
书　号：ISBN 978-7-111-59373-7　　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) 88378991 88361066　　投稿热线：(010) 88379604
购书热线：(010) 68326294 88379649 68995259　　读者信箱：hzjsj@hzbook.com

自然语言处理（Natural Language Processing，NLP）主要研究用计算机来处理、理解以及运用人类语言（又称自然语言）的各种理论和方法，属于人工智能领域的一个重要研究方向，是计算机科学与语言学的交叉学科，又常被称为计算语言学。随着互联网的快速发展，网络文本尤其是用户生成的文本呈爆炸性增长，为自然语言处理带来了巨大的应用需求。同时，自然语言处理研究的进步，也为人们更深刻地理解语言的机理和社会的机制提供了一种新的途径，因此具有重要的科学意义。

然而，自然语言具有歧义性、动态性和非规范性，同时语言理解通常需要丰富的知识和一定的推理能力，这些都给自然语言处理带来了极大的挑战。目前，统计机器学习技术为以上问题提供了一种可行的解决方案，成为研究的主流，该研究领域又被称为统计自然语言处理。一个统计自然语言处理系统通常由两部分组成，即训练数据（也称样本）和统计模型（也称算法）。

但是，传统的机器学习方法在数据获取和模型构建等诸多方面都存在严重的问题。首先，为获得大规模的标注数据，传统方法需要花费大量的人力、物力、财力，雇用语言学专家进行繁琐的标注工作。由于这种方法存在标注代价高、规范性差等问题，很难获得大规模、高质量的人工标注数据，由此带来了严重的数据稀疏问题。其次，在传统的自然语言处理模型中，通常需要人工设计模型所需要的特征以及特征组合。这种人工设计特征的方式，需要开发人员对所面对的问题有深刻的理解和丰富的经验，这会消耗大量的人力和时间，即便如此也往往很难获得有效的特征。

近年来，如火如荼的深度学习技术为这两方面的问题提供了一种可能的解决思路，有效推动了自然语言处理技术的发展。深度学习一般是指建立在含有多层非线性变换的神经网络结构之上，对数据的表示进行抽象和学习的一系列机器学习算法。该方法已对语音识别、图像处理等领域的进步起到了极大的推动作用，同时也引起了自然语言处理领域学者的广泛关注。

深度学习主要为自然语言处理的研究带来了两方面的变化：一方面是使用统一的分布式（低维、稠密、连续）向量表示不同粒度的语言单元，如词、短语、句子和篇章等；另一方面是使用循环、卷积、递归等神经网络模型对不同的语言单元向量进行组合，获得更

大语言单元的表示。除了不同粒度的单语语言单元外，不同种类的语言甚至不同模态（语言、图像等）的数据都可以通过类似的组合方式表示在相同的语义向量空间中，然后通过在向量空间中的运算来实现分类、推理、生成等各种任务并应用于各种相关的任务之中。

虽然将深度学习技术应用于自然语言处理的研究目前非常热门，但是市面上还没有一本书系统地阐述这方面的研究进展，初学者往往通过学习一些在线课程（如斯坦福的CS224N课程）来掌握相关的内容。本书恰好弥补了这一不足，深入浅出地介绍了深度学习的基本知识及各种常用的网络结构，并重点介绍了如何使用这些技术处理自然语言。

本书的作者Yoav Goldberg现就职于以色列巴伊兰大学，是自然语言处理领域一位非常活跃的青年学者。Goldberg博士期间的主要研究方向为依存句法分析，随着深度学习的兴起，他也将研究兴趣转移至此，并成功地将该技术应用于依存句法分析等任务。与此同时，他在理论上对词嵌入和传统矩阵分解方法的对比分析也具有广泛的影响力。另外，他还是DyNet深度学习库的主要开发者之一。可见，无论在理论上还是实践上，他对深度学习以及自然语言处理都具有非常深的造诣。这些都为本书的写作奠定了良好的基础。

由于基于深度学习的自然语言处理是一个非常活跃的研究领域，新的理论和技术层出不穷，因此本书很难涵盖所有的最新技术。不过，本书基本涵盖了目前已经被证明非常有效的技术。关于这方面的进展，读者可以参阅自然语言处理领域最新的论文。

我们要感谢对本书的翻译有所襄助的老师和学生。本书由哈尔滨工业大学的车万翔、郭江、张伟男、刘铭四位老师主译，刘挺教授主审。侯宇泰、姜天文、李家琦、覃立波、宋皓宇、滕德川、王宇轩、向政鹏、张杨子、郑桂东、朱海潮、朱庆福等对本书部分内容的初译做了很多工作，机械工业出版社华章公司策划编辑朱劼和姚蕾在本书的整个翻译过程中提供了许多帮助，在此一并予以衷心感谢。

译文虽经多次修改和校对，但由于译者的水平有限，加之时间仓促，疏漏及错误在所难免，我们真诚地希望读者不吝赐教，不胜感激。

车万翔

2017年10月于哈尔滨工业大学

自然语言处理（Natural Language Processing，NLP）这一术语指的是对人类语言进行自动的计算处理。它包括两类算法：将人类产生的文本作为输入；产生看上去很自然的文本作为输出。由于人类产生的文本每年都在不停增加，同时人们期望使用人类的语言与计算机进行交流，因此人们对该类算法的需求在不断增加。然而，由于人类语言固有的歧义、不断变化以及病态性（not well defined），导致自然语言处理极具挑战性。

自然语言本质上是符号化的，因此人们最开始也尝试使用符号化的方式处理语言，即基于逻辑、规则以及本体的方法。然而，自然语言具有很强的歧义性和可变性，这就需要使用统计的方法。事实上，如今自然语言处理的主流方法都是基于统计机器学习（Statistical Machine Learning）的。过去十几年，核心的 NLP 技术都是以有监督学习的线性模型为主导，核心算法如感知机、线性支持向量机、逻辑回归等都是在非常高维和稀疏的特征向量上进行训练的。

2014 年左右，该领域开始看到一些从基于稀疏向量的线性模型向基于稠密向量的非线性神经网络模型（Nonlinear Neural Network Model）切换的成功案例。一些神经网络技术是线性模型的简单推广，可用于替代线性分类器。另一些神经网络技术更进一步提出了新的建模方法，这需要改变现有的思维方式。特别是一系列基于循环神经网络（Recurrent Neural Network，RNN）的方法，减轻了对马尔可夫假设的依赖性，这曾普遍用于序列模型中。循环神经网络可以处理任意长度的序列数据，并生成有效的特征抽取器。这些进展导致了语言模型、自动机器翻译以及其他一些应用的突破。

虽然神经网络方法很强大，但是由于各种原因，入门并不容易。本书中，我将试图为自然语言处理的从业者以及刚入门的读者介绍神经网络的基本背景、术语、工具和方法论，帮助他们理解将神经网络用于自然语言处理的原理，并且能够应用于他们自己的工作中。我也希望为机器学习和神经网络的从业者介绍自然语言处理的基本背景、术语、工具以及思维模式，以便他们能有效地处理语言数据。

最后，我希望本书能够作为自然语言处理以及机器学习这两个领域新手的一个较好的入门指导。

目标读者

本书的目标读者应具有计算机或相关领域的技术背景，他们想使用神经网络技术来加速自然语言处理的研究。虽然本书的主要读者是自然语言处理和机器学习领域的研究生，但是我试图（通过介绍一些高级材料）使自然语言处理或者机器学习领域的研究者，甚至对这两个领域都不了解的人也能阅读本书，后者显然需要更加努力。

虽然本书是自包含的，我仍然假设读者具有数学知识，特别是本科水平的概率、代数和微积分以及基本的算法和数据结构知识。有机器学习的先验知识会很有帮助，但这并不是必需的。

本书是对一篇综述文章［Goldberg，2016］的扩展，内容上进行了重新组织，提供了更宽泛的介绍，涵盖了一些更深入的主题，由于各种原因，这些主题没有在那篇综述文章中提及。本书也包括一些综述文章中没有的，将神经网络用于语言数据的更具体的应用实例。本书试图对那些没有自然语言处理和机器学习背景的读者也能有用，然而综述文章假设他们对这些领域已经具备了一些知识。事实上，熟悉 2006 年到 2014 年期间自然语言处理实践的读者，可能发现期刊版本读起来更快并且对于他们的需求组织得更好，这是因为那段时期人们大量使用基于线性模型的机器学习技术。然而，这些读者可能也会愿意阅读关于词嵌入的章节（第 10 和 11 章）、使用循环神经网络有条件生成的章节（第 17 章），以及结构化预测和多任务学习（Multi-task Learning，MTL）的章节（第 19 和 20 章）。

本书的焦点

本书试图是自包含的，因此将不同的方法在统一的表示和框架下加以表述。然而，本书的主要目的是介绍神经网络（深度学习）的机制及其在语言数据上的应用，而不是深入介绍机器学习理论和自然语言处理技术。如果需要这些内容，建议读者参考外部资源。

类似地，对于那些想开发新的神经网络机制的人，本书不是一个全面的资源（虽然本书可能是一个很好的入门）。确切地讲，本书的目标读者是那些对现有技术感兴趣，并且想将其以创造性的方式应用于他们喜欢的语言处理任务的人。

扩展阅读 对神经网络更深入、一般性的讨论以及它们背后的理论、最新的优化方法和其他主题，读者可以参考其他资源。强烈推荐 Bengio 等人 [2016] 的书。

对于更友好而且更严密的实用机器学习介绍，强烈推荐 Daumé III [2015] 的免费书。对于机器学习更理论化的介绍，参见 Shalev-Shwartz 和 Ben-David [2014] 的免费书以及 Mohri 等人 [2012] 的教科书。

对于自然语言处理的更深入介绍参见 Jurafsky 和 Martin [2008] 的书。Manning 等人 [2008] 的信息检索书也包括语言数据处理的一些相关信息。

最后，如要快速了解语言学的背景，Bender [2013] 的书提供了简单但全面的介绍，对于有计算思维的读者有指导意义。Sag 等人 [2003] 的介绍性语法书的前几章也值得一读。

本书写作之际，神经网络和深度学习的研究也在快速进展之中。最好的方法在不断变化，所以我不能保证介绍的都是最新、最好的方法。因此，我会专注于涵盖更确定、更鲁棒的技术（它们在很多场景下都被证明有效），同时选取那些还没完全发挥作用但有前途的技术。

Yoav Goldberg

2017 年 3 月

致 谢

Neural Network Methods for Natural Language Processing

本书是我之前写的综述文章［Goldberg，2016］的扩展，之所以写那篇论文，是由于我在学习和讲授深度学习与自然语言处理相交叉的内容时，发现缺乏组织良好并且清晰的材料。感谢曾对那篇论文提出过意见的人（以各种形式，从最初的草稿到出版后）。一些是面对面的意见，一些是电子邮件，一些是在 Twitter 上的随意对话。本书也受一些人的影响，他们没有直接对书上的内容提出意见（事实上，一些人从没有读过本书），但是讨论过相关的主题。一些是深度学习的专家，一些是自然语言处理的专家，一些两者皆是，还有一些人正在学习这两个主题。一些人提供了细致的意见，其他人是对小细节的讨论。但是他们中的每个人都影响了本书的最终版本。他们是（按字母序）：Yoav Artzi，Yonatan Aumann，Jason Baldridge，Miguel Ballesteros，Mohit Bansal，Marco Baroni，Tal Baumel，Sam Bowman，Jordan Boyd-Graber，Chris Brockett，Ming-Wei Chang，David Chiang，Kyunghyun Cho，Grzegorz Chrupala，Alexander Clark，Raphael Cohen，Ryan Cotterell，Hal Daumé III，Nicholas Dronen，Chris Dyer，Jacob Eisenstein，Jason Eisner，Michael Elhadad，Yad Faeq，Manaal Faruqui，Amir Globerson，Fréderic Godin，Edward Grefenstette，Matthew Honnibal，Dirk Hovy，Moshe Koppel，Angeliki Lazaridou，Tal Linzen，Thang Luong，Chris Manning，Stephen Merity，Paul Michel，Margaret Mitchell，Piero Molino，Graham Neubig，Joakim Nivre，Brendan O'Connor，Nikos Pappas，Fernando Pereira，Barbara Plank，Ana-Maria Popescu，Delip Rao，Tim Rocktäschel，Dan Roth，Alexander Rush，Naomi Saphra，Djamé Seddah，Erel Segal-Halevi，Avi Shmidman，Shaltiel Shmidman，Noah Smith，Anders Søgaard，Abe Stanway，Emma Strubell，Sandeep Subramanian，Liling Tan，Reut Tsarfaty，Peter Turney，Tim Vieira，Oriol Vinyals，Andreas Vlachos，Wenpeng Yin，Torsten Zesch。

当然，此列表不包括那些我读过的此主题学术著作的作者。

本书也得益于我与巴伊兰大学自然语言处理组的交流：Yossi Adi，Roee Aharoni，Oded Avraham，Ido Dagan，Jessica Ficler，Jacob Goldberger，Hila Gonen，Joseph Keshet，Eliyahu Kiperwasser，Ron Konigsberg，Omer Levy，Oren Melamud，Gabriel Stanovsky，Ori Shapira，Micah Shlain，Vered Shwartz，Hillel Taub-Tabib，Rachel

Witties。

本书和那篇综述文章的匿名评阅人提出了一系列很有用的意见、建议和勘误，这些对最终版本的许多方面带来了显著的提升。无论你是谁，谢谢！

同时感谢 Graeme Hirst、Michael Morgan、Samantha Draper 和 C. L. Tondo 的精心策划。

像往常一样，所有的错误都是我造成的。如果你发现任何错误，请告诉我，我会在下一版中更正。

最后，我要感谢我的妻子 Noa，当我厌倦写作时，她保持耐心并且给我支持。我的父母 Esther 和 Avner，我的兄弟 Nadav，在许多情况下，对于写书他们比我更兴奋。在写作过程中，The Streets 和 Shne'or 咖啡馆的员工为我提供了很好的服务，使我可以专心致志。

Yoav Goldberg

2017 年 3 月

目 录

Neural Network Methods for Natural Language Processing

译者序

前言

致谢

第 1 章 引言 …… 1

1.1 自然语言处理的挑战 …… 1

1.2 神经网络和深度学习 …… 2

1.3 自然语言处理中的深度学习 …… 3

1.4 本书的覆盖面和组织结构 …… 5

1.5 本书未覆盖的内容 …… 7

1.6 术语 …… 7

1.7 数学符号 …… 7

注释 …… 8

第一部分 有监督分类与前馈神经网络

第 2 章 学习基础与线性模型 …… 13

2.1 有监督学习和参数化函数 …… 13

2.2 训练集、测试集和验证集 …… 14

2.3 线性模型 …… 16

2.3.1 二分类 …… 16

2.3.2 对数线性二分类 …… 20

2.3.3 多分类 …… 20

2.4 表示 …… 21

2.5 独热和稠密向量表示 …… 22

2.6 对数线性多分类 …… 23

2.7 训练和最优化 …… 24

2.7.1 损失函数 …… 25

2.7.2 正则化 …… 27

2.8 基于梯度的最优化 …… 29

2.8.1 随机梯度下降 …… 29

2.8.2 实例 …… 31

2.8.3 其他训练方法 …… 32

第 3 章 从线性模型到多层感知器 …… 34

3.1 线性模型的局限性：异或问题 … 34

3.2 非线性输入转换 …… 34

3.3 核方法 …… 35

3.4 可训练的映射函数 …… 35

第 4 章 前馈神经网络 …… 37

4.1 一个关于大脑的比喻 …… 37

4.2 数学表示 …… 38

4.3 表达能力 …… 40

4.4 常见的非线性函数 …… 41

4.5 损失函数 …… 42

4.6 正则化与丢弃法 …… 42

4.7 相似和距离层 …… 43

4.8 嵌入层 …… 44

第 5 章 神经网络训练 …… 45

5.1 计算图的抽象概念 …… 45

5.1.1 前向计算 …… 47

5.1.2 反向计算（导数、反向传播） … 47

5.1.3 软件 …… 48

5.1.4 实现流程 …… 51

5.1.5 网络构成 …… 51
5.2 实践经验 …… 51
5.2.1 优化算法的选择 …… 52
5.2.2 初始化 …… 52
5.2.3 重启与集成 …… 52
5.2.4 梯度消失与梯度爆炸 …… 53
5.2.5 饱和神经元与死神经元 …… 53
5.2.6 随机打乱 …… 54
5.2.7 学习率 …… 54
5.2.8 minibatch …… 54

第二部分 处理自然语言数据

第 6 章 文本特征构造 …… 57
6.1 NLP 分类问题中的拓扑结构 …… 57
6.2 NLP 问题中的特征 …… 59
6.2.1 直接可观测特征 …… 59
6.2.2 可推断的语言学特征 …… 62
6.2.3 核心特征与组合特征 …… 66
6.2.4 n 元组特征 …… 66
6.2.5 分布特征 …… 67
第 7 章 NLP 特征的案例分析 …… 69
7.1 文本分类：语言识别 …… 69
7.2 文本分类：主题分类 …… 69
7.3 文本分类：作者归属 …… 70
7.4 上下文中的单词：词性标注 …… 71
7.5 上下文中的单词：命名实体识别 …… 72
7.6 上下文中单词的语言特征：介词词义消歧 …… 74
7.7 上下文中单词的关系：弧分解分析 …… 76
第 8 章 从文本特征到输入 …… 78
8.1 编码分类特征 …… 78
8.1.1 独热编码 …… 78
8.1.2 稠密编码（特征嵌入） …… 79
8.1.3 稠密向量与独热表示 …… 80
8.2 组合稠密向量 …… 81
8.2.1 基于窗口的特征 …… 81
8.2.2 可变特征数目：连续词袋 …… 82
8.3 独热和稠密向量间的关系 …… 82
8.4 杂项 …… 83
8.4.1 距离与位置特征 …… 83
8.4.2 补齐、未登录词和词丢弃 …… 84
8.4.3 特征组合 …… 85
8.4.4 向量共享 …… 86
8.4.5 维度 …… 86
8.4.6 嵌入的词表 …… 86
8.4.7 网络的输出 …… 87
8.5 例子：词性标注 …… 87
8.6 例子：弧分解分析 …… 89
第 9 章 语言模型 …… 91
9.1 语言模型任务 …… 91
9.2 语言模型评估：困惑度 …… 92
9.3 语言模型的传统方法 …… 93
9.3.1 延伸阅读 …… 94
9.3.2 传统语言模型的限制 …… 94
9.4 神经语言模型 …… 95
9.5 使用语言模型进行生成 …… 97
9.6 副产品：词的表示 …… 98
第 10 章 预训练的词表示 …… 100
10.1 随机初始化 …… 100

10.2 有监督的特定任务的预训练 … 100
10.3 无监督的预训练 … 101
10.4 词嵌入算法 … 102
10.4.1 分布式假设和词表示 … 103
10.4.2 从神经语言模型到分布式表示 … 107
10.4.3 词语联系 … 110
10.4.4 其他算法 … 111
10.5 上下文的选择 … 112
10.5.1 窗口方法 … 112
10.5.2 句子、段落或文档 … 113
10.5.3 句法窗口 … 113
10.5.4 多语种 … 114
10.5.5 基于字符级别和子词的表示 … 115
10.6 处理多字单元和字变形 … 116
10.7 分布式方法的限制 … 117
第 11 章 使用词嵌入 … 119
11.1 词向量的获取 … 119
11.2 词的相似度 … 119
11.3 词聚类 … 120
11.4 寻找相似词 … 120
11.5 同中选异 … 121
11.6 短文档相似度 … 121
11.7 词的类比 … 122
11.8 改装和映射 … 122
11.9 实用性和陷阱 … 124
第 12 章 案例分析：一种用于句子意义推理的前馈结构 … 125
12.1 自然语言推理与 SNLI 数据集 … 125
12.2 文本相似网络 … 126
第三部分 特殊的结构
第 13 章 *n* 元语法探测器：卷积神经网络 … 131
13.1 基础卷积+池化 … 132
13.1.1 文本上的一维卷积 … 133
13.1.2 向量池化 … 135
13.1.3 变体 … 137
13.2 其他选择：特征哈希 … 137
13.3 层次化卷积 … 138
第 14 章 循环神经网络：序列和栈建模 … 142
14.1 RNN 抽象描述 … 142
14.2 RNN 的训练 … 145
14.3 RNN 常见使用模式 … 145
14.3.1 接收器 … 145
14.3.2 编码器 … 146
14.3.3 传感器 … 146
14.4 双向 RNN … 147
14.5 堆叠 RNN … 149
14.6 用于表示栈的 RNN … 150
14.7 文献阅读的注意事项 … 151
第 15 章 实际的循环神经网络结构 … 153
15.1 作为 RNN 的 CBOW … 153
15.2 简单 RNN … 153
15.3 门结构 … 154
15.3.1 长短期记忆网络 … 156
15.3.2 门限循环单元 … 157

15.4 其他变体 …… 158
15.5 应用到 RNN 的丢弃机制 …… 159
第 16 章 通过循环网络建模 …… 160
16.1 接收器 …… 160
16.1.1 情感分类器 …… 160
16.1.2 主谓一致语法检查 …… 162
16.2 作为特征提取器的 RNN …… 164
16.2.1 词性标注 …… 164
16.2.2 RNN-CNN 文本分类 …… 166
16.2.3 弧分解依存句法分析 …… 167
第 17 章 条件生成 …… 169
17.1 RNN 生成器 …… 169
17.2 条件生成（编码器-解码器） …… 170
17.2.1 序列到序列模型 …… 172
17.2.2 应用 …… 173
17.2.3 其他条件上下文 …… 176
17.3 无监督的句子相似性 …… 177
17.4 结合注意力机制的条件生成 … 178
17.4.1 计算复杂性 …… 180
17.4.2 可解释性 …… 181
17.5 自然语言处理中基于注意力机制的模型 …… 182
17.5.1 机器翻译 …… 182
17.5.2 形态屈折 …… 184
17.5.3 句法分析 …… 184

第四部分 其他主题

第 18 章 用递归神经网络对树建模 … 187
18.1 形式化定义 …… 187
18.2 扩展和变体 …… 190
18.3 递归神经网络的训练 …… 190
18.4 一种简单的替代——线性化树 …… 191
18.5 前景 …… 191
第 19 章 结构化输出预测 …… 193
19.1 基于搜索的结构化预测 …… 193
19.1.1 基于线性模型的结构化预测 …… 193
19.1.2 非线性结构化预测 …… 194
19.1.3 概率目标函数（CRF） …… 195
19.1.4 近似搜索 …… 196
19.1.5 重排序 …… 197
19.1.6 参考阅读 …… 197
19.2 贪心结构化预测 …… 198
19.3 条件生成与结构化输出预测 …… 199
19.4 实例 …… 200
19.4.1 基于搜索的结构化预测：一阶依存句法分析 …… 200
19.4.2 基于 Neural-CRF 的命名实体识别 …… 201
19.4.3 基于柱搜索的 NER-CRF 近似 …… 204
第 20 章 级联、多任务与半监督学习 …… 206
20.1 模型级联 …… 206
20.2 多任务学习 …… 210
20.2.1 多任务设置下的训练 …… 212
20.2.2 选择性共享 …… 212
20.2.3 作为多任务学习的词嵌入预训练 …… 213

20.2.4 条件生成中的多任务学习 … 214
20.2.5 作为正则的多任务学习 …… 214
20.2.6 注意事项 …… 214
20.3 半监督学习 …… 215
20.4 实例 …… 216
20.4.1 眼动预测与句子压缩 …… 216
20.4.2 弧标注与句法分析 …… 217
20.4.3 介词词义消歧与介词翻译预测 …… 218
20.4.4 条件生成：多语言机器翻译、句法分析以及图像描述生成 … 219
20.5 前景 …… 220
第21章 结论 …… 221
21.1 我们学到了什么 …… 221
21.2 未来的挑战 …… 221
参考文献 …… 223

第1章

Neural Network Methods for Natural Language Processing

引　言

1.1　自然语言处理的挑战

自然语言处理(Natural Language Processing，NLP)是一个设计输入和输出为非结构化自然语言数据的方法和算法的研究领域。人类语言有很强的歧义性(如句子“I ate pizza with friends”(我和朋友一起吃披萨)和“I ate pizza with olives”(我吃了有橄榄的披萨))和多样性(如“I ate pizza with friends”也可以说成“Friends and I shared some pizza”)。语言也一直在进化中。人善于产生和理解语言，并具有表达、感知、理解复杂且微妙信息的能力。与此同时，虽然人类是语言的伟大使用者，但是我们并不善于形式化地理解和描述支配语言的规则。

使用计算机理解和产生语言因此极具挑战性。事实上，最为人所知的处理语言数据的方法是使用有监督机器学习(supervised machine learning)算法，其试图从事先标注好的输入/输出集合中推导出使用的模式和规则。例如，一个将文本分为4类的任务，类别为：体育、政治、八卦和经济。显然，文本中的单词提供了非常强的线索，但是到底哪些单词提供了什么线索呢？为该任务书写规则极具挑战性。然而，读者可以轻松地将一篇文档分到一个主题中，然后，基于每类几百篇人为分类的样例，可以让有监督机器学习产生用词的模式，从而帮助文本分类。机器学习方法擅长那些很难获得规则集，但是相对容易获得给定输入及相应输出样本的领域。

除了使用不明确规则集处理歧义和多样输入的挑战之外，自然语言展现了另外一些特性，其使得用包括机器学习在内的计算方法更具挑战性，即离散性(discrete)、组合性(compositional)和稀疏性(sparse)。

语言是符号化和离散的。书面语义的基本单位是字符，字符构成了单词，单词再表示对象、概念、事件、动作和思想。字符和单词都是离散符号：如“hamburger”(汉堡包)或“pizza”(披萨)会唤起我们头脑中的某种表示，但是它们也是不同的符号，其含义是不相关的，待我们的大脑去理解。从符号自身看，“hamburger”和“pizza”之间没有内在的关系，从构成它们的字母看也是一样。与机器视觉中普遍使用的如颜色的概念或声学信号

相对比，这些概念都是连续的，如可以使用简单的数学运算从一幅彩色图像变为灰度图像，或者从色调、光强等内在性质比较两幅图像。对于单词，这些都不容易做到，如果不使用一个大的查找表或者词典，没有什么简单的运算可以从单词“red”(红)变为单词“pink”(粉红)。

语言还具有组合性，即字母形成单词，单词形成短语和句子。短语的含义可以比包含的单词更大，并遵循复杂的规则集。为了理解一个文本，我们需要超越字母和单词，看到更长的单词序列，如句子甚至整篇文本。

以上性质的组合导致了**数据稀疏性**(data sparseness)。单词(离散符号)组合并形成意义的方式实际上是无限的。可能合法的句子数是巨大的，我们从没指望能全部枚举出来。随便翻开一本书，其中绝大部分句子是你之前从没看过和听过的。甚至，很有可能很多四个单词构成的序列对你都是新鲜的。如果你看一下过去 10 年的报纸或者想象一下未来 10 年的报纸，许多单词，特别是人名、品牌和公司以及俚语和术语都将是新的。我们也不清楚如何从一个句子生成另一个句子或者定义句子之间的相似性，这不依赖于它们的意思——对我们是不可观测的。当我们要从实例中学习时也是挑战重重，即使有非常大的实例集合，我们仍然很容易观测到实例集合中从没出现过的事件，其与曾出现过的所有实例都非常不同。

1.2 神经网络和深度学习

深度学习是机器学习的一个分支，是神经网络(neural network)的重命名。神经网络是一系列学习技术，历史上曾受模拟脑计算工作的启发，可被看作学习参数可微的数学函数[1]。深度学习的名字源于许多层被连在一起的可微函数。

虽然全部机器学习技术都可以被认为是基于过去的观测学习如何做出预测，但是深度学习方法不仅学习预测，而且学习正确地表示数据，以使其更有助于预测。给出一个巨大的输入-输出映射集合，深度学习方法将数据“喂”给一个网络，其产生输入的后继转换，直到用最终的转换来预测输出。网络产生的转换都学习自给定的输入-输出映射，以便每个转换都使得更易于将数据和期望的标签之间建立联系。

人类设计者负责设计网络结构和训练方式，提供给网络合适的输入-输出实例集合，将输入数据恰当地编码，大量学习正确表示的工作则由网络自动执行，同时受到网络结构的支持。