版权相关注意事项：
1、书籍版权归著者和出版社所有
2、本PDF来自于各个广泛的信息平台，经过整理而成
3、本PDF仅限用于非商业用途或者个人交流研究学习使用
4、本PDF获得者不得在互联网上以任何目的进行传播，违规者造成的法律责任和后果，违规者自负
5、如果觉得书籍内容很赞，请一定购买正版实体书，多多支持编写高质量的图书的作者和相应的出版社！当然，如果图书内容不堪入目，质量低下，你也可以选择狠狠滴撕裂本PDF
6、技术类书籍是拿来获取知识的，不是拿来收藏的，你得到了书籍不意味着你得到了知识，所以请不要得到书籍后就觉得沾沾自喜，要经常翻阅！！经常翻阅
7、请于下载PDF后24小时内研究使用并删掉本PDF

版权相关注意事项：
1、书籍版权归著者和出版社所有
2、本PDF来自于各个广泛的信息平台，经过整理而成
3、本PDF仅限用于非商业用途或者个人交流研究学习使用
4、本PDF获得者不得在互联网上以任何目的进行传播，违规者造成的法律责任和后果，违规者自负
5、如果觉得书籍内容很赞，请一定购买正版实体书，多多支持编写高质量的图书的作者和相应的出版社！当然，如果图书内容不堪入目，质量低下，你也可以选择狠狠滴撕裂本PDF
6、技术类书籍是拿来获取知识的，不是拿来收藏的，你得到了书籍不意味着你得到了知识，所以请不要得到书籍后就觉得沾沾自喜，要经常翻阅！！经常翻阅
7、请于下载PDF后24小时内研究使用并删掉本PDF

仅供非商业用途或交流学习使用

华章IT

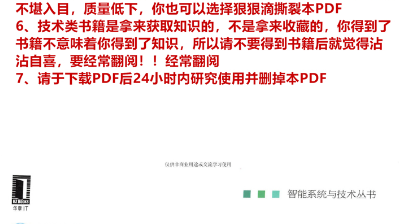
智能系统与技术丛书

天善智能 TIANSHAN SOFT

Natural Language Processing
Core Technology and Algorithm with Python

Python自然语言处理实战
核心技术与算法

涂铭 刘祥 刘树春 著



机械工业出版社
China Machine Press

内容简介

自然语言处理是一门融语言学、计算机科学、数学于一体的学科，比较复杂，学习门槛高，但本书巧妙地避开了晦涩难懂的数学公式和证明，即便没有数学基础，也能零基础入门。

本书专注于中文的自然语言处理，以Python及其相关框架为工具，以实战为导向，详细讲解了自然语言处理的各种核心技术、方法论和经典算法。三位作者在人工智能、大数据和算法领域有丰富的积累和经验，是阿里巴巴、前明略数据和七牛云的资深专家。同时，本书也得到了阿里巴巴达摩院高级算法专家、七牛云AI实验室Leader等专家的高度评价和鼎力推荐。

全书一共11章，在逻辑上分为2个部分：

第一部分（第1、2、11章）

主要介绍了自然语言处理中需要了解的基础知识、前置技术、Python科学包、正则表达式以及Solr检索等。

第二部分（第3~10章）

第3~5章讲解了词法分析相关的技术，包括中文分词技术、词性标注与命名实体识别、关键词提取算法等。

第6章讲解了句法分析技术，该部分目前理论研究较多，工程实践中使用门槛相对较高，且效果多是依赖结合业务知识进行规则扩展，因此本书未做深入探讨。

第7章讲解了常用的向量化方法，这些方法常用于各种NLP任务的输入。

第8章讲解了情感分析相关的概念、场景以及一般做情感分析的流程。情感分析在很多行业都有应用。

第9章介绍了机器学习的重要概念，重点突出了NLP常用的分类算法、聚类算法，同时还介绍了几个案例。

第10章介绍了NLP中常用的一些深度学习算法，这些方法比较复杂，但是非常实用，需要读者耐心学习。

HZBOOKS Information Technology 华章IT

智能系统与技术丛书

Natural Language Processing
Core Technology and Algorithm with Python

Python自然语言处理实战
核心技术与算法

涂铭 刘祥 刘树春 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 自然语言处理实战：核心技术与算法 / 涂铭，刘祥，刘树春著．—北京：机械工业出版社，2018.4
（智能系统与技术丛书）

ISBN 978-7-111-59767-4

I. P… II. ①涂… ②刘… ③刘… III. 软件工具－自然语言处理－教材 IV. ① TP311.56 ② TP391

中国版本图书馆 CIP 数据核字（2018）第 088115 号

Python 自然语言处理实战：核心技术与算法

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：张锡鹏　责任校对：殷 虹
印　　刷：北京诚信伟业印刷有限公司　版　　次：2018 年 5 月第 1 版第 1 次印刷
开　　本：186mm × 240mm 1/16　印　　张：18.75
书　　号：ISBN 978-7-111-59767-4　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

版权所有 · 侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所 韩光 / 邹晓东

FOREWORD 1

序 一

不知不觉间，我们已经进入了"人工智能"时代，如今随处可见基于自然语言处理技术的聊天机器人，回想以前都是靠人工服务，现在都依靠机器人回答大部分的常见问题了。

过去几年，深度学习架构和算法在图像识别和语音处理等领域取得了重大的进步。而在 NLP（自然语言处理）领域，起初并没有太大的进展。不过现在，NLP 领域取得的一系列进展已证明深度学习技术将会对自然语言处理做出重大贡献。一些常见的任务如实体命名识别，词类标记及情感分析等，NLP 都能提供最新的结果，并超越了传统方法。另外，在机器翻译领域的应用上，深度学习技术所取得的进步应该是最显著的。

记得在上学期间感觉 NLP 这个领域很新鲜、很空白，决定尝试做一下，读完博士，感觉 NLP 比我最初接触时理解的 NLP 更新鲜，更值得挖掘。NLP 很多问题都没有正式定义，或者说很难用统一的标准去训练机器、很难搞 benchmark dataset，这可能也是 AI 的一大挑战。

我认为现在比较成熟的 AI 方向都是相对确定的。比如语音识别，拿来一段语音，就知道说的是什么话；比如 vision，猫的照片就是猫，这个人脸的照片就是这个人。NLP 有一些问题就没这么明确。比如文本摘要，到底哪一个摘要是最好的呢？机器翻译，到底哪一个译文是最好的呢？复杂一些的情感分析，这篇报道的作者到底有没有在暗讽这个人？如果一个问题有明确的答案，比如 Waston——专门参加开心辞典回答问题，算法训练起来轻松一些。但如果一个问题本身的答案并无明确的高下之分，那算法也无可奈何。

定义新问题，以较小的代价搜集新的数据集，开发新的 evaluation method，这些与研究新算法一样有趣、有挑战性。举个简单的例子。我们想让机器自动识别出来讽刺的语气，那么去哪里找讽刺的话呢？我们有现成的分析情感的工具，再利用这些有 #sarcasm 标签的推文，可以训练一个识别"什么情况是倒霉情况"的分类器。以后就可以用这个倒霉识别器去识别没有标签的讽刺句子了，bootstrap 一下把数据集搞大，这就是一个最初级的讽刺方面的数据集。

NLP 圈里很多人喜欢搞新的数据集，这个现象有利有弊，但这说明了有很多空白问题需要定义，有很多空白资源需要创建。本书从各个方面着手，帮助读者理解 NLP 的过程，提供了各种实战场景，结合现实项目背景，帮助读者理解 NLP 中的数据结构和算法以及目前主流的 NLP 技术与方法论，结合信息检索技术与大数据应用等流行技术，最终完成对 NLP 的学习和掌握。

阿里巴巴达摩院高级算法专家　黄英
2018 年 1 月 17 于杭州

FOREWORD 2

序 二

近年来，几乎整个人工智能界的研究者们都注意到一个技术名词——自然语言处理（NLP）。NLP 作为人工智能领域的一个重要分支，现在已经发展成为人工智能研究中的热点方向。最近几十年来，随着软硬件协同发展，数据爆炸性增长，信息过载的问题越来越严重，全部依赖人来分析和驱动的传统方式，面对海量信息的局面显得越来越捉襟见肘。这样的情况下，能够自动化处理大规模文本相关的数据的 NLP，即将成为未来人工智能发展技术的新趋势和方向。

自然语言处理作为机器学习与语言学、统计学等的综合学科，不仅知识内容多，发展迅速，而且非常依赖于工程能力。目前，统计学以及数据驱动的方法在 NLP 中占据着统治地位。同时，最近几年深度学习不断被引入 NLP 领域，越来越多的知识需要读者去学习。这时候急需一本能够从全局梳理 NLP 的书籍，帮助 NLP 学习者快速入门。传统的 NLP 书籍对于具体问题的方法讲解有足够的思路，但是要么是基于英文语料的讲解，要么通篇都是理论，面对复杂的中文语料环境缺乏实践性。

本书的作者通过对前人传统 NLP 技术以及新兴的深度学习方法深入梳理，形成自己理解的 NLP 解决之道。本书在内容上平衡了理论和技术，在每章的理论之后都配备了实践课，方便读者能够动手加深理解，避免成为只会夸夸其谈的 NLP 理论"专家"。本书可以帮助研究者，特别是初学者，加强对 NLP 的理论与技术的学习，授人以鱼的同时授人以渔，帮助读者灵活解决实际工作当中遇到的各种 NLP 问题。

七牛云 AI 实验室 Leader，10 余年人工智能和深度学习研究　林亦宁

PREFACE

前 言

为什么要写这本书

这是一本关于中文自然语言处理（简称 NLP）的书，NLP 是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。NLP 是一门融语言学、计算机科学、数学于一体的科学。本书偏重实战，不仅系统介绍了 NLP 涉及的知识点，同时也教会读者如何实际应用与开发。围绕这个主题，本书从章节规划到具体的讲述方式，具有以下两个特点：

第一个特点是本书的主要目标读者定位为高校相关专业的大学生（统计学、计算机技术）、NLP 爱好者，以及不具备专业数学知识的人群。NLP 是一系列学科的集合体，其中包含了语言学、机器学习、统计学、大数据以及人工智能等方面，尤其依赖数学知识才能深入理解其原理。因此本书对专业知识的讲述过程必须绕过复杂的数学证明，从问题的前因后果、创造者思考的过程、概率或几何解释代替数学解释等一系列迂回的路径去深入模型的本源，这可能多少会牺牲一些严谨性，但是却能换来对大多数人更为友好的阅读体验。

第二个特点是本书是一本介绍中文自然语言处理的书，中文分词相对于英文分词来说更为复杂，读者将通过例子来学习，体会到能够通过实践验证自己想法的价值，我们提供了丰富的来自 NLP 领域的案例。在本书的内容编制上，从知识点背景介绍到原理剖析，辅以实战案例，所有的代码会在书中详细列出或者上传 Github 方便读者下载与调试，帮助读者快速上手掌握知识点，同时可以应用到后续实际的开发项目中。在实际项

版权相关注意事项：
1、书籍版权归著者和出版社所有
2、本PDF来自于各个广泛的信息平台，经过整理而成
3、本PDF仅限用于非商业用途或者个人交流研究学习使用
4、本PDF获得者不得在互联网上以任何目的进行传播，违规者造成的法律责任和后果，违规者自负
5、如果觉得书籍内容很赞，请一定购买正版实体书，多多支持编写高质量的图书的作者和相应的出版社！当然，如果图书内容不堪入目，质量低下，你也可以选择狠狠滴撕裂本PDF
6、技术类书籍是拿来获取知识的，不是拿来收藏的，你得到了书籍不意味着你得到了知识，所以请不要得到书籍后就觉得沾沾自喜，要经常翻阅！！经常翻阅
7、请于下载PDF后24小时内研究使用并删掉本PDF

目章节中，选取目前在 NLP 领域中比较热门的项目，将之前的知识点进行汇总，帮助读者巩固与提升。本书难度适中属于入门和扩展级读物。

读者将学到什么

- 如何用 NLP 与语言学的关键概念来描述和分析语言
- NLP 中的数据结构和算法是怎样的
- 自然语言处理目前主流的技术与方法论
- 信息检索技术与大数据应用

读者对象

1）统计学或相关 IT 专业学生

本书的初衷是面向相关专业的学生——大量基于理论知识的认知却缺乏实战经验的人员，让其在理论的基础上深入了解。通过本书，学生可以跟随本书的教程一起操作学习，达到对自己使用的人工智能工具、算法和技术知其然亦知其所以然的目的。

2）信息科学和计算机科学爱好者

本书是一部近代科技的历史书，也是一部科普书，还可以作为一部人工智能思想和技术的教科书去阅读。通过本书可以了解到行业先驱们在探索人工智能道路上所做出的努力和思考，理解他们不同的观点和思路，有助于开拓自己的思维和视野。

3）人工智能相关专业的研究人员

本书具体介绍了 NLP 相关知识。通过本书可以了解理论知识，了解哪些才是项目所需的内容以及如何在项目中实现，能够快速上手。

如何阅读本书

本书内容针对 NLP 从以下几个方面进行阐述：

第一部分的内容包括第 1、2、11 章，主要介绍了 NLP 所需要了解的 Python 科学包、正则表达式以及 Solr 检索。

第二部分的内容包括第 5 ～ 10 章的内容，主要是介绍 NLP 相关的各个知识点。分别是：

第 3 ～ 5 章主要介绍了词法分析层面的一些技术，这一部分是 NLP 技术的基础，需要读者熟练掌握。

第 6 章介绍了句法分析技术，该部分目前理论研究较多，工程实践中使用门槛相对较高，且效果多是依赖结合业务知识进行规则扩展，因此本书未做深入探讨，读者了解即可。

第 7 章介绍了常用的向量化方法。这些方法常用于各种 NLP 任务的输入，读者需重点掌握。

第 8 章介绍了情感分析相关的概念、场景以及一般做情感分析的流程，情感分析在很多行业都有应用，所以需要读者熟练掌握。

第 9 章介绍了机器学习的一些基本概念，重点突出 NLP 常用的分类算法、聚类算法，同时还介绍了几个案例，这章是 NLP 的基础内容，需要重点掌握。

第 10 章介绍了 NLP 中常用的深度学习算法，这些方法比较复杂，但是非常实用，需要读者耐心学习。

除了以上内容外，以下信息是在本书中涉及特定内容的解释和说明：

内容延伸。本书每个章节都有一定的内容延伸章节，其内容是对于有兴趣深入研究

的读者推荐的资料或进一步了解的知识点，由于每个主题都涵盖很多内容，因此本书仅在内容延伸中抛砖引玉，有兴趣的读者可以加以了解和学习。

相关知识点。本书很多章节中都有"相关知识点"的内容介绍，其对特定工具、知识、算法、库等方面做了较为详细的介绍，它们是本书的知识堡垒。

本章小结。每章的结尾都有"本章小结"，在小结中包含 4 部分内容。

- 内容小结。内容小结是有关本章内容的总结。
- 重点知识。重点知识是本章重点需要读者掌握的知识和内容。
- 外部参考。外部参考是本章提到过但是无法详细介绍的内容，都在外部参考中列出，有兴趣的读者可以基于比构建自己的知识图谱。
- 应用实践。基于本章内容推荐读者在实践中落地的建议。

提示。对于知识点的重要提示和应用技巧，相对"相关知识点"而言，每条提示信息的内容量较少，一般都是经验类的总结。

注意。特定需要引起注意的知识，这些注意点是应用过程中需要避免的"大坑"。

关于附件的使用方法

除了第 1 章外，本书的每一章都有对应源数据和完整代码，该内容可在本书中直接找到，有些代码需要从 Github 下载，地址是 https://github.com/nlpinaction/learning-nlp。需要注意的是，为了让读者更好地了解每行代码的含义，笔者在注释信息中使用了中文标注，且每个程序文件的编码格式都是 UTF-8。

勘误和支持

由于笔者水平有限，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。读者可通过以下途径联系并反馈建议或意见：

- **即时通讯**。添加个人微信（kennymingtu）反馈问题。
- **电子邮件**。发送 E-mail 到 kenny_tm@hotmail.com。

致谢

在本书的撰写过程中，得到了来自多方的指导、帮助和支持。

首先要感谢的是机械工业出版社华章公司的杨福川编辑，他在本书出版过程中给予我极大的支持和鼓励，并为此书的撰写提供了方向和思路指导。

其次要感谢黄英和周剑老师在自然语言处理项目和工作中提供的宝贵经验和支持。

再次要感谢全程参与审核、校验等工作的张锡鹏、孙海亮编辑以及其他背后默默支持的出版工作者，是他们的辛勤付出才让本书得以顺利面世。

最后感谢我的父母、家人和朋友，使得我有精力完成本书的编写。

谨以此书献给热爱数据工作并为之奋斗的朋友们，愿大家身体健康、生活美满、事业有成！

涂铭
2018 年 1 月于上海

书籍初成，感慨良多。

在接受邀请撰写本书时，从未想到过程如此艰辛。

感谢我的女友和家人的理解与支持，陪伴我度过写书的漫长岁月。

感谢我的合著者——涂铭和刘树春，与他们合作轻松愉快，给予我很多的理解和包容。

感谢参与审阅、校验等工作的孙海亮老师等出版社工作人员，是他们在幕后的辛勤

付出保证了本书的出版成功。

再次感谢一路陪伴的所有人！

刘祥
2018 年 1 月于北京

首先要感谢我的两位合作者——涂铭和刘祥，能够相聚在一起写书是缘分。当初聊到出版 NLP 入门书籍的想法时我们一拍即合，然而真正开始执笔才发现困难重重，业余时间常常被工作挤占，进度一拖再拖，在伙伴们的支持下，克服了拖延症，顺利完成本书。

特别感谢我的爱人和家人的悉心照料和支持。

感谢孙海亮老师、张锡鹏老师等出版社工作人员，没有他们的辛劳付出就没有本书保质保量的完成。

最后感谢帮我校稿的林博、谢雨飞、陈敏，谢谢他们在生活和工作上给我的支持与帮助。

最后，祝大家学习快乐。

刘树春
2018 年 1 月于上海

CONTENTS

目　录

序一
序二
前言

第 1 章　NLP 基础……1

1.1　什么是 NLP……1

1.1.1　NLP 的概念……1

1.1.2　NLP 的研究任务……3

1.2　NLP 的发展历程……5

1.3　NLP 相关知识的构成……7

1.3.1　基本术语……7

1.3.2　知识结构……9

1.4　语料库……10

1.5　探讨 NLP 的几个层面……11

1.6　NLP 与人工智能……13

1.7　本章小结……15

第 2 章　NLP 前置技术解析……16

2.1　搭建 Python 开发环境……16

2.1.1　Python 的科学计算发行版——Anaconda……17

2.1.2　Anaconda 的下载与安装…19

2.2　正则表达式在 NLP 的基本应用…21

2.2.1　匹配字符串……22

2.2.2　使用转义符……26

2.2.3　抽取文本中的数字……26

2.3　Numpy 使用详解……27

2.3.1　创建数组……28

2.3.2　获取 Numpy 中数组的维度……30

2.3.3　获取本地数据……31

2.3.4　正确读取数据……32

2.3.5　Numpy 数组索引……32

2.3.6　切片……33

2.3.7　数组比较……33

2.3.8　替代值……34

2.3.9　数据类型转换……36

2.3.10　Numpy 的统计计算方法……36

2.4　本章小结……37

第 3 章　中文分词技术……38

3.1　中文分词简介……38

3.2 规则分词……39
3.2.1 正向最大匹配法……39
3.2.2 逆向最大匹配法……40
3.2.3 双向最大匹配法……41
3.3 统计分词……42
3.3.1 语言模型……43
3.3.2 HMM 模型……44
3.3.3 其他统计分词算法……52
3.4 混合分词……52
3.5 中文分词工具——Jieba……53
3.5.1 Jieba 的三种分词模式……54
3.5.2 实战之高频词提取……55
3.6 本章小结……58

第 4 章 词性标注与命名实体识别…59

4.1 词性标注……59
4.1.1 词性标注简介……59
4.1.2 词性标注规范……60
4.1.3 Jieba 分词中的词性标注…61
4.2 命名实体识别……63
4.2.1 命名实体识别简介……63
4.2.2 基于条件随机场的命名实体识别……65
4.2.3 实战一：日期识别……69
4.2.4 实战二：地名识别……75
4.3 总结……84

第 5 章 关键词提取算法……85

5.1 关键词提取技术概述……85
5.2 关键词提取算法 TF/IDF 算法……86
5.3 TextRank 算法……88
5.4 LSA/LSI/LDA 算法……91
5.4.1 LSA/LSI 算法……93
5.4.2 LDA 算法……94
5.5 实战提取文本关键词……95
5.6 本章小结……105

第 6 章 句法分析……106

6.1 句法分析概述……106
6.2 句法分析的数据集与评测方法…107
6.2.1 句法分析的数据集……108
6.2.2 句法分析的评测方法……109
6.3 句法分析的常用方法……109
6.3.1 基于 PCFG 的句法分析……110
6.3.2 基于最大间隔马尔可夫网络的句法分析……112
6.3.3 基于 CRF 的句法分析……113
6.3.4 基于移进－归约的句法分析模型……113
6.4 使用 Stanford Parser 的 PCFG 算法进行句法分析……115
6.4.1 Stanford Parser……115
6.4.2 基于 PCFG 的中文句法分析实战……116
6.5 本章小结……119

第 7 章　文本向量化……120
7.1　文本向量化概述……120
7.2　向量化算法 word2vec……121
7.2.1　神经网络语言模型……122
7.2.2　C&W 模型……124
7.2.3　CBOW 模型和 Skip-gram 模型……125
7.3　向量化算法 doc2vec/str2vec……127
7.4　案例：将网页文本向量化……129
7.4.1　词向量的训练……129
7.4.2　段落向量的训练……133
7.4.3　利用 word2vec 和 doc2vec 计算网页相似度……134
7.5　本章小结……139
第 8 章　情感分析技术……140
8.1　情感分析的应用……141
8.2　情感分析的基本方法……142
8.2.1　词法分析……143
8.2.2　机器学习方法……144
8.2.3　混合分析……144
8.3　实战电影评论情感分析……145
8.3.1　卷积神经网络……146
8.3.2　循环神经网络……147
8.3.3　长短时记忆网络……148
8.3.4　载入数据……150
8.3.5　辅助函数……154
8.3.6　模型设置……155
8.3.7　调参配置……158
8.3.8　训练过程……159
8.4　本章小结……159
第 9 章　NLP 中用到的机器学习算法……160
9.1　简介……160
9.1.1　机器学习训练的要素……161
9.1.2　机器学习的组成部分……162
9.2　几种常用的机器学习方法……166
9.2.1　文本分类……166
9.2.2　特征提取……168
9.2.3　标注……169
9.2.4　搜索与排序……170
9.2.5　推荐系统……170
9.2.6　序列学习……172
9.3　分类器方法……173
9.3.1　朴素贝叶斯 Naïve Bayesian……173
9.3.2　逻辑回归……174
9.3.3　支持向量机……175
9.4　无监督学习的文本聚类……177
9.5　文本分类实战：中文垃圾邮件分类……180
9.5.1　实现代码……180
9.5.2　评价指标……187
9.6　文本聚类实战：用 K-means 对豆瓣读书数据聚类……190
9.7　本章小结……194

第 10 章　基于深度学习的 NLP 算法……195
10.1　深度学习概述……195
10.1.1　神经元模型……196
10.1.2　激活函数……197
10.1.3　感知机与多层网络……198
10.2　神经网络模型……201
10.3　多输出层模型……203
10.4　反向传播算法……204
10.5　最优化算法……208
10.5.1　梯度下降……208
10.5.2　随机梯度下降……209
10.5.3　批量梯度下降……210
10.6　丢弃法……211
10.7　激活函数……211
10.7.1　tanh 函数……212
10.7.2　ReLU 函数……212
10.8　实现 BP 算法……213
10.9　词嵌入算法……216
10.9.1　词向量……217
10.9.2　word2vec 简介……217
10.9.3　词向量模型……220
10.9.4　CBOW 和 Skip-gram 模型……222
10.10　训练词向量实践……224
10.11　朴素 Vanilla-RNN……227
10.12　LSTM 网络……230
10.12.1　LSTM 基本结构……230
10.12.2　其他 LSTM 变种形式……234
10.13　Attention 机制……236
10.13.1　文本翻译……237
10.13.2　图说模型……237
10.13.3　语音识别……239
10.13.4　文本摘要……239
10.14　Seq2Seq 模型……240
10.15　图说模型……242
10.16　深度学习平台……244
10.16.1　Tensorflow……245
10.16.2　Mxnet……246
10.16.3　PyTorch……246
10.16.4　Caffe……247
10.16.5　Theano……247
10.17　实战 Seq2Seq 问答机器人……248
10.18　本章小结……254
第 11 章　Solr 搜索引擎……256
11.1　全文检索的原理……257
11.2　Solr 简介与部署……258
11.3　Solr 后台管理描述……263
11.4　配置 schema……267
11.5　Solr 管理索引库……270
11.5.1　创建索引……270
11.5.2　查询索引……276
11.5.3　删除文档……279
11.6　本章小结……281

试读结束：需要全本请在线购买： www.ertongbook.com

CHAPTER 1

第 1 章

NLP 基础

在本章中，你将学到与 NLP（自然语言处理）相关的基础知识。

本章的要点包括：

▼ NLP 基础概念

▼ NLP 的发展与应用

▼ NLP 常用术语以及扩展介绍

1.1 什么是 NLP

1.1.1 NLP 的概念

NLP（Natural Language Processing，自然语言处理）是计算机科学领域以及人工智能领域的一个重要的研究方向，它研究用计算机来处理、理解以及运用人类语言（如中文、英文等），达到人与计算机之间进行有效通讯。所谓“自然”乃是寓意自然进化形成，是为了区分一些人造语言，类似 C++、Java 等人为设计的语言。在人类社会中，语言扮演着重要的角色，语言是人类区别于其他动物的根本标志，没有语言，人类的思维无从谈起，沟通交流更是无源之水。在一般情况下，用户可能不熟悉机器语言，所以自然语言处理技术可以帮助这样的用户使用自然语言和机器交流。从建模的角度看，为了方便计算机处理，自然语言可以被定义为一组规则或符号的集合，我们组合集合中的符号来

传递各种信息。自然语言处理研究表示语言能力、语言应用的模型，通过建立计算机框架来实现这样的语言模型，并且不断完善这样的语言模型，还需要根据该语言模型来设计各种实用的系统，并且探讨这些实用技术的评测技术。这一定义有点宽泛，但是语言本身就是人类最为复杂的概念之一。这些年，NLP 研究取得了长足的进步，逐渐发展成为一门独立的学科，从自然语言的角度出发，NLP 基本可以分为两个部分：自然语言处理以及自然语言生成，演化为理解和生成文本的任务，如图 1-1 所示。

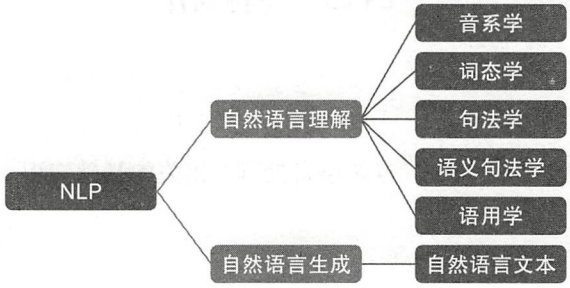


图 1-1 NLP 的基本分类

自然语言的理解是个综合的系统工程，它又包含了很多细分学科，有代表声音的音系学，代表构词法的词态学，代表语句结构的句法学，代表理解的语义句法学和语用学。

- 音系学：指代语言中发音的系统化组织。
- 词态学：研究单词构成以及相互之间的关系。
- 句法学：给定文本的哪部分是语法正确的。
- 语义学：给定文本的含义是什么？
- 语用学：文本的目的是什么？

语言理解涉及语言、语境和各种语言形式的学科。而自然语言生成（Natural Language Generation，NLG）恰恰相反，从结构化数据中以读取的方式自动生成文本。该过程主要包含三个阶段：文本规划（完成结构化数据中的基础内容规划）、语句规划（从结构化数据中组合语句来表达信息流）、实现（产生语法通顺的语句来表达文本）。

1.1.2 NLP 的研究任务

NLP 可以被应用于很多领域，这里大概总结出以下几种通用的应用：

- ▼ 机器翻译：计算机具备将一种语言翻译成另一种语言的能力。
- ▼ 情感分析：计算机能够判断用户评论是否积极。
- ▼ 智能问答：计算机能够正确回答输入的问题。
- ▼ 文摘生成：计算机能够准确归纳、总结并产生文本摘要。
- ▼ 文本分类：计算机能够采集各种文章，进行主题分析，从而进行自动分类。
- ▼ 舆论分析：计算机能够判断目前舆论的导向。
- ▼ 知识图谱：知识点相互连接而成的语义网络。

机器翻译是自然语言处理中最为人所熟知的场景，国内外有很多比较成熟的机器翻译产品，比如百度翻译、Google 翻译等，还有提供支持语音输入的多国语言互译的产品（比如科大讯飞就出了一款翻译机）。

体验下，百度在线翻译：

http://fanyi.baidu.com/?aldtype=16047#auto/zh

情感分析在一些评论网站比较有用，比如某餐饮网站的评论中会有非常多拔草的客人的评价，如果一眼扫过去满眼都是又贵又难吃，那谁还想去呢？另外有些商家为了获取大量的客户不惜雇佣水军灌水，那就可以通过自然语言处理来做水军识别，情感分析来分析总体用户评价是积极还是消极。

智能问答在一些电商网站有非常实际的价值，比如代替人工充当客服角色，有很多基本而且重复的问题，其实并不需要人工客服来解决，通过智能问答系统可以筛选掉大量重复的问题，使得人工座席能更好地服务客户。

体验下，图灵机器人：

http://www.tuling123.com/experience/exp_virtual_robot.jhtml? nav=exp

文摘生成利用计算机自动地从原始文献中摘取文摘，全面准确地反映某一文献的中心内容。这个技术可以帮助人们节省大量的时间成本，而且效率更高。

文本分类是机器对文本按照一定的分类体系自动标注类别的过程。举一个例子，垃圾邮件是一种令人头痛的顽症，困扰着非常多的互联网用户。2002 年，Paul Graham 提出使用“贝叶斯推断”来过滤垃圾邮件，1000 封垃圾邮件中可以过滤掉 995 封并且没有一个是误判，另外这种过滤器还具有自我学习功能，会根据新收到的邮件，不断调整。也就是说收到的垃圾邮件越多，相对应的判断垃圾邮件的准确率就越高。

舆论分析可以帮助分析哪些话题是目前的热点，分析传播路径以及发展趋势，对于不好的舆论导向可以进行有效的控制。

知识图谱（Knowledge Graph/Vault）又称科学知识图谱，在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。知识图谱的一般表现形式如图 1-2 所示。

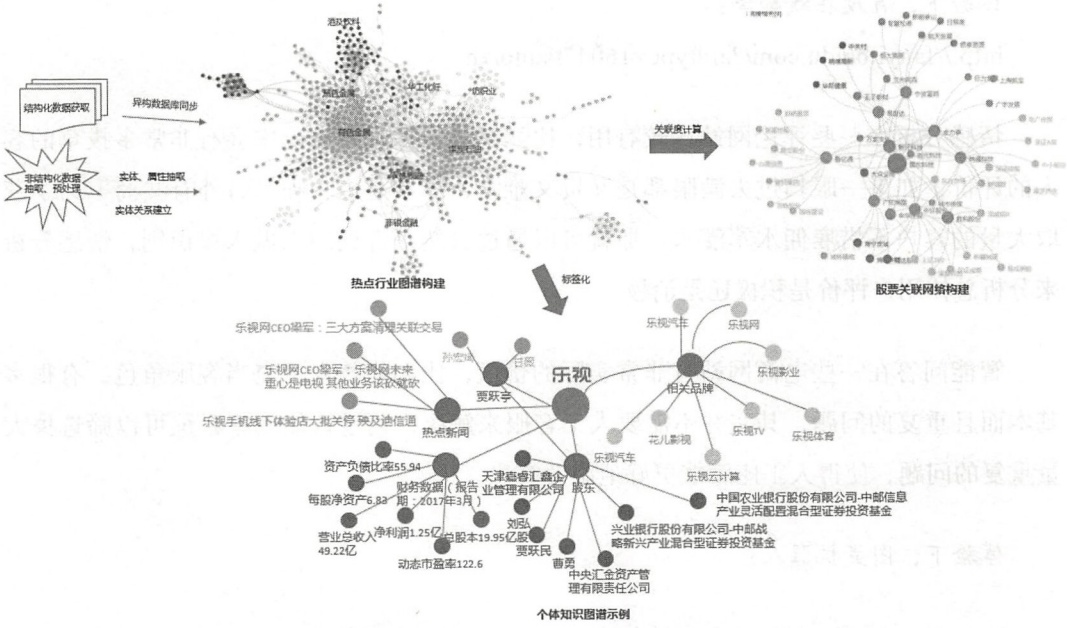


图 1-2　知识图谱图示

1.2 NLP 的发展历程

NLP 的发展大致经历了 3 个阶段：1956 年以前的萌芽期，1980 年～ 1999 年的快速发展期和 21 世纪的突飞猛进期。

萌芽期（1956 年以前）

早期的自然语言处理具有鲜明的经验主义色彩。如 1913 年马尔科夫提出马尔可夫随机过程与马尔可夫模型的基础就是“手工查频”，具体说就是统计了《欧根·奥涅金》长诗中元音与辅音出现频度；1948 年香农把离散马尔可夫的概率模型应用于语言的自动机，同时采用手工方法统计英语字母的频率。

然而这种经验主义到了乔姆斯基时期出现了转变。1956 年乔姆斯基借鉴香农的工作，把有限状态机作为刻画语法的工具，建立了自然语言的有限状态模型，具体来说就是用“代数”和“集合”将语言转化为符号序列，建立了一大堆有关语法的数学模型。这些工作非常伟大，为自然语言和形式语言找到了一种统一的数学描述理论，一个叫做“形式语言理论”的新领域诞生了。但乔姆斯基否定了有限状态模型在自然语言中的适用性，然后主张采用有限的、严格的规则去描述无限的语言现象，提出了风靡一时的转换生成语法。这一时期，虽然诸如贝叶斯方法、隐马尔可夫、最大熵、支持向量机等经典理论和算法也有提出，但自然语言处理领域的主流仍然是基于规则的理性主义方法。

快速发展期（1980 年 ～ 1999 年）

这种情况一直持续到 20 世纪 80 年代初期才发生变化，很多学者开始反思有限状态模型以及经验主义方法的合理性。20 世纪 80 年代初，话语分析（Discourse Analysis）也取得了重大进展。之后，由于自然语言处理研究者对于过去的研究进行了反思，有限状态模型和经验主义研究方法也开始复苏。

90 年代后，基于统计的自然语言处理开始大放异彩。首先是在机器翻译领域取得了突破，因为引入了许多基于语料库的方法。1990 年在芬兰赫尔辛基举办的第 13 届国际

计算语言学会议确定的主题是“处理大规模真实文本的理论、方法与工具”，研究的重心开始转向大规模真实文本了，传统的基于规则的自然语言处理显然力不从心了。学者们认为，大规模语料至少是对基于规则方法有效的补充。在1994年～1999年间，经验主义空前繁荣，如句法剖析、词类标注、参照消解、话语处理的算法几乎把“概率”与“数据”作为标准方法，成为自然语言处理的主流。

20世纪90年代中期，有两件事从根本上促进了自然语言处理研究的复苏与发展。一件事是20世纪90年代中期以来，计算机的运行速度和存储量大幅增加，为自然语言处理改善了物质基础，使得语音和语言处理的商品化开发成为可能；另一件事是1994年Internet商业化和同期网络技术的发展使得基于自然语言的信息检索和信息抽取的需求变得更加突出。这样，自然语言处理的社会需求更加迫切，自然语言处理的应用面也更加宽广，自然语言处理不再局限于机器翻译、语音控制等早期研究领域了。

从20世纪90年代末到21世纪初，人们逐渐认识到，仅用基于规则或统计的方法是无法成功进行自然语言处理的。基于统计、基于实例和基于规则的语料库技术在这一时期开始蓬勃发展，各种处理技术开始融合，自然语言处理的研究再次繁荣。

突飞猛进期（2000年至今）

进入21世纪以后，自然语言处理又有了突飞猛进的变化。2006年，以Hinton为首的几位科学家历经近20年的努力，终于成功设计出第一个多层神经网络算法——深度学习。这是一种将原始数据通过一些简单但是非线性的模型转变成更高层次、更加抽象表达的特征学习方法，一定程度上解决了人类处理“抽象概念”这个亘古难题。目前，深度学习在机器翻译、问答系统等多个自然语言处理任务中均取得了不错的成果，相关技术也被成功应用于商业化平台中。

未来，深度学习作为人工智能皇冠上的明珠，将会在自然语言处理领域发挥着越来越重要的作用。

试读结束：需要全本请在线购买： www.ertongbook.com

1.3 NLP 相关知识的构成

1.3.1 基本术语

为了帮助读者更好地学习 NLP，这里会一一介绍 NLP 领域的一些基础专业词汇。

（1）分词（segment）

词是最小的能够独立活动的有意义的语言成分，英文单词之间是以空格作为自然分界符的，而汉语是以字为基本的书写单位，词语之间没有明显的区分标记，因此，中文词语分析是中文分词的基础与关键。中文和英文都存在分词的需求，不过相较而言，英文单词本来就有空格进行分割，所以处理起来相对方便。但是，由于中文是没有分隔符的，所以分词的问题就比较重要。分词常用的手段是基于字典的最长串匹配，据说可以解决 85% 的问题，但是歧义分词很难。举个例子，“美国会通过对台售武法案”，我们既可以切分为“美国 / 会 / 通过对台售武法案”，又可以切分成“美 / 国会 / 通过对台售武法案”。

（2）词性标注（part-of-speech tagging）

基于机器学习的方法里，往往需要对词的词性进行标注。词性一般是指动词、名词、形容词等。标注的目的是表征词的一种隐藏状态，隐藏状态构成的转移就构成了状态转移序列。例如：我 /r 爱 /v 北京 /ns 天安门 /ns。其中，ns 代表名词，v 代表动词，ns、v 都是标注，以此类推。

（3）命名实体识别（NER，Named Entity Recognition）

命名实体是指从文本中识别具有特定类别的实体（通常是名词），例如人名、地名、机构名、专有名词等。

（4）句法分析（syntax parsing）

句法分析往往是一种基于规则的专家系统。当然也不是说它不能用统计学的方法进

行构建，不过最初的时候，还是利用语言学专家的知识来构建的。句法分析的目的是解析句子中各个成分的依赖关系。所以，往往最终生成的结果是一棵句法分析树。句法分析可以解决传统词袋模型不考虑上下文的问题。比如，“小李是小杨的班长”和“小杨是小李的班长”，这两句话，用词袋模型是完全相同的，但是句法分析可以分析出其中的主从关系，真正理清句子的关系。

（5）指代消解（anaphora resolution）

中文中代词出现的频率很高，它的作用的是用来表征前文出现过的人名、地名等。例如，清华大学坐落于北京，这家大学是目前中国最好的大学之一。在这句话中，其实“清华大学”这个词出现了两次，“这家大学”指代的就是清华大学。但是出于中文的习惯，我们不会把“清华大学”再重复一遍。

（6）情感识别（emotion recognition）

所谓情感识别，本质上是分类问题，经常被应用在舆情分析等领域。情感一般可以分为两类，即正面、负面，也可以是三类，在前面的基础上，再加上中性类别。一般来说，在电商企业，情感识别可以分析商品评价的好坏，以此作为下一个环节的评判依据。通常可以基于词袋模型 + 分类器，或者现在流行的词向量模型 +RNN。经过测试发现，后者比前者准确率略有提升。

（7）纠错（correction）

自动纠错在搜索技术以及输入法中利用得很多。由于用户的输入出错的可能性比较大，出错的场景也比较多。所以，我们需要一个纠错系统。具体做法有很多，可以基于N-Gram 进行纠错，也可以通过字典树、有限状态机等方法进行纠错。

（8）问答系统（QA system）

这是一种类似机器人的人工智能系统。比较著名的有：苹果 Siri、IBM Watson、微软小冰等。问答系统往往需要语音识别、合成，自然语言理解、知识图谱等多项技术的

配合才会实现得比较好。

1.3.2 知识结构

作为一门综合学科，NLP 是研究人与机器之间用自然语言进行有效通信的理论和方法。这需要很多跨学科的知识，需要语言学、统计学、最优化理论、机器学习、深度学习以及自然语言处理相关理论模型知识做基础。作为一门杂学，NLP 可谓是包罗万象，体系化与特殊化并存，这里简单罗列其知识体系：

- 句法语义分析：针对目标句子，进行各种句法分析，如分词、词性标记、命名实体识别及链接、句法分析、语义角色识别和多义词消歧等。
- 关键词抽取：抽取目标文本中的主要信息，比如从一条新闻中抽取关键信息。主要是了解是谁、于何时、为何、对谁、做了何事、产生了有什么结果。涉及实体识别、时间抽取、因果关系抽取等多项关键技术。
- 文本挖掘：主要包含了对文本的聚类、分类、信息抽取、摘要、情感分析以及对挖掘的信息和知识的可视化、交互式的呈现界面。
- 机器翻译：将输入的源语言文本通过自动翻译转化为另一种语言的文本。根据输入数据类型的不同，可细分为文本翻译、语音翻译、手语翻译、图形翻译等。机器翻译从最早的基于规则到二十年前的基于统计的方法，再到今天的基于深度学习（编解码）的方法，逐渐形成了一套比较严谨的方法体系。
- 信息检索：对大规模的文档进行索引。可简单对文档中的词汇，赋以不同的权重来建立索引，也可使用算法模型来建立更加深层的索引。查询时，首先对输入比进行分析，然后在索引里面查找匹配的候选文档，再根据一个排序机制把候选文档排序，最后输出排序得分最高的文档。
- 问答系统：针对某个自然语言表达的问题，由问答系统给出一个精准的答案。需要对自然语言查询语句进行语义分析，包括实体链接、关系识别，形成逻辑表达式，然后到知识库中查找可能的候选答案并通过一个排序机制找出最佳的答案。
- 对话系统：系统通过多回合对话，跟用户进行聊天、回答、完成某项任务。主要

涉及用户意图理解、通用聊天引擎、问答引擎、对话管理等技术。此外，为了体现上下文相关，要具备多轮对话能力。同时，为了体现个性化，对话系统还需要基于用户画像做个性化回复。知识结构结构图如图 1-3 所示。

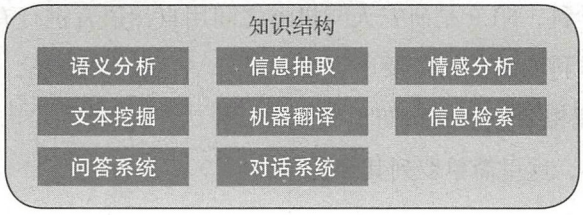


图 1-3 知识结构图示

扩展阅读

自然语言的学习，需要有以下几个前置知识体系：

▼ 目前主流的自然语言处理技术使用 python 来编写。

▼ 统计学以及线性代数入门。

1.4 语料库

巧妇难为无米之炊，语料库就是 NLP 的“米”，本书用到的语料库主要有：

（1）中文维基百科[⊖]

维基百科是最常用且权威的开放网络数据集之一，作为极少数的人工编辑、内容丰富、格式规范的文本语料，各类语言的维基百科在 NLP 等诸多领域应用广泛。维基百科提供了开放的词条文本整合下载，可以找到你需要的指定时间、指定语言、指定类型、指定内容的维基百科数据，中文维基百科数据是维基提供的语料库。

⊖ https://dumps.wikimedia.org/zhwiki/。

（2）搜狗新闻语料库[⊖]

来自若干新闻站点 2012 年 6 月～7 月期间国内、国际、体育、社会、娱乐等 18 个频道的新闻数据，提供 URL 和正文信息。

（3）IMDB 情感分析语料库[⊜]

互联网电影资料库（Internet Movie Database，简称 IMDb）是一个关于电影演员、电影、电视节目、电视明星和电影制作的在线数据库。IMDb 的资料中包括了影片的众多信息、演员、片长、内容介绍、分级、评论等。对于电影的评分目前使用最多的就是 IMDb 评分。

还有豆瓣读书相关语料（爬虫获取）、邮件相关语料等。

1.5 探讨 NLP 的几个层面

本书所探讨的自然语言处理可以分为以下三个层面：

（1）第一层面：词法分析

词法分析包括汉语的分词和词性标注这两部分。之前有提过，汉语分词与英文不同，汉语书面词语之间没有明显的空格标记，文本中的句子以字符串的方式出现，句子中由逗号分隔，句子和句子之间常以句号分隔。针对汉语这种独特的书面表现形式，汉语的自然语言处理的首要工作就是要将输入的文本切分为单独的词语，然后在此技术上进行其他更高级的分析。

上述这个步骤称为分词。除了分词之外，词性标注也通常被认为是词法分析的一部分，词性标注的目的是为每一个词赋予一个类别，这个类别可以是名词（noun）、动

⊖ http://download.labs.sogou.com/resource/ca.php。

⊜ https://www.kaggle.com/tmdb/tmdb-movie-metadata。

词（verb）、形容词（adjective）等。通常来说，属于相同词性的词，在句法中承担类似的角色。

（2）第二层面：句法分析

句法分析是对输入的文本以句子为单位，进行分析以得到句子的句法结构的处理过程。对句法结构进行分析，一方面是为了帮助理解句子的含义，另一方面也为更高级的自然语言处理任务提供支持（比如机器翻译、情感分析等）。目前业界存在三种比较主流的句法分析方法：短语结构句法体系，作用是识别出句子中的短语结构以及短语之间的层次句法关系；依存结构句法体系，作用是识别句子中词与词之间的相互依赖关系；深层文法句法分析，利用深层文法，例如词汇化树邻接文法，组合范畴文法等对句子进行深层的句法以及语义分析。

上述几种句法分析，依存句法分析属于浅层句法分析，其实现过程相对来说比较简单而且适合在多语言环境下应用，但是其所能提供的信息也相对较少。深层文法句法分析可以提供丰富的句法和语义信息，但是采用的文法相对比较复杂，分析器的运行复杂度也比较高，这使得深层句法分析不太适合处理大规模的数据。短语结构句法分析介于依存句法分析和深层文法句法分析之间。

（3）第三个层面：语义分析

语义分析的最终目的是理解句子表达的真实语义。但是，语义应该采用什么表示形式一直困扰着研究者们，至今这个问题也没有一个统一的答案。语义角色标注（semantic role labeling）是目前比较成熟的浅层语义分析技术。语义角色标注一般都在句法分析的基础上完成，句法结构对于语义角色标注的性能至关重要。基于逻辑表达的语义分析也得到学术界的长期关注。出于机器学习模型复杂度、效率的考虑，自然语言处理系统通常采用级联的方式，即分词、词性标注、句法分析、语义分析分别训练模型。实际使用时，给定输入句子，逐一使用各个模块进行分析，最终得到所有结果。

近年来，随着研究工作的深入，研究者们提出了很多有效的联合模型，将多个任务

试读结束：需要全本请在线购买： www.ertongbook.com
仅供非商业用途或交流学习使用