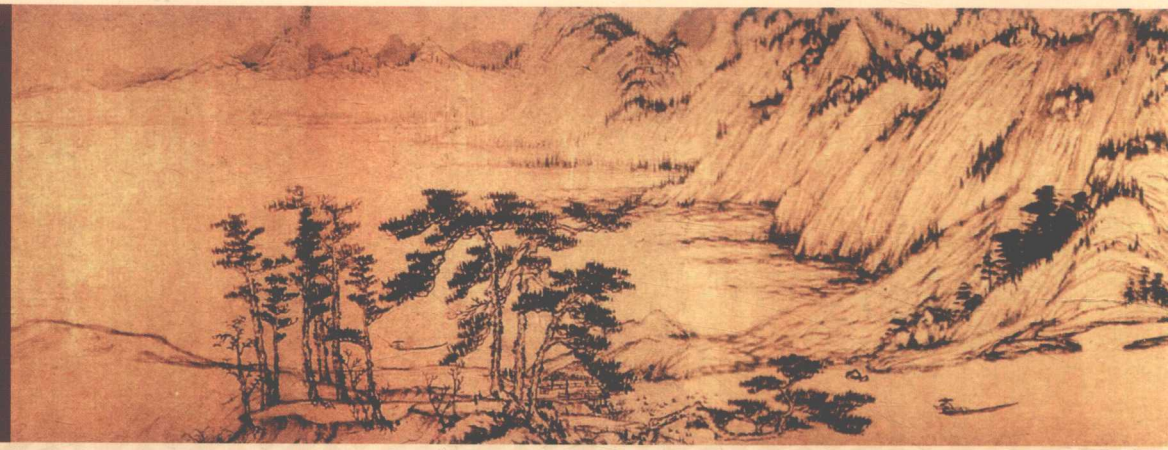

中文自然语言处理导论

黄锦辉 等◎著
徐睿峰 李斌阳 黄锦辉◎译

科学出版

中文自然语言处理导论

黄锦辉 等◎著
徐睿峰　李斌阳　黄锦辉◎译

科学出版社
北京

图字：01-2017-3912 号

内 容 简 介

本书主要向具备计算机处理基础的读者介绍中文自然语言处理问题和技术。由于中西方语言处理方法之间的主要区别集中在词汇层面，所以本书主要讨中文形态分析，主要内容包括中文自然语言处理技术介绍、中文词素处理、中文分词、未登录词识别、中文词汇的语义表达、中文搭配等。

本书可作为计算机科学相关专业的教学参考书，也可供相关领域研究人员和工程技术人员使用。

图书在版编目（CIP）数据

中文自然语言处理导论/黄锦辉等著；徐睿峰，李斌阳，黄锦辉译.
—北京：科学出版社，2018.10

书名原文: Introduction to Chinese Natural Language Processing

ISBN 978-7-03-059044-2

I. ①中… II. ①黄… ②徐… ③李… III. ①中文—自然语言处理 IV. ①TP391

中国版本图书馆 CIP 数据核字(2018)第 230902 号

责任编辑：郭勇斌 肖 雷 / 责任校对：王晓茜
责任印制：张克忠 / 封面设计：蔡美宇

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

三河市春园印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 10 月第 一 版 开本：720×1000 1/16
2018 年 10 月第一次印刷 印张：9
字数：166 000

定价：78.00 元

（如有印装质量问题，我社负责调换）

译者简介

黄锦辉，1987年获得苏格兰爱丁堡大学博士学位。现任香港中文大学系统工程与工程管理学系教授，中国东北大学和北京大学兼职教授，研究方向为中文计算，并行数据库和信息检索，先后在多个国际期刊、会议、书籍中发表了该领域的200余篇论文。系ACM汇刊《亚洲语言处理》（TALIP）的创始主编，国际期刊《东方语言计算处理》的联合主编，《分布式并行数据库》《计算语言学和中文处理》和《中文处理》的编辑委员之一。担任APWeb'08（中国沈阳）和AIRS'2008（中国哈尔滨）会议的联合主席，IJCNLP'2005（韩国济州）会议的程序委员会联合主席，以及VLDB2002会议的小组联合主席。

李文捷，自2001年起担任香港理工大学计算机系副教授，1988年和1993年分别获得天津大学系统工程专业学士学位和硕士学位，1997年获得香港中文大学信息系统专业博士学位。主要研究方向包括信息抽取、文本摘要、自然语言处理和时间信息处理。已发表超过100篇国际期刊和核心会议论文，现任《计算机语言处理》副主编。

徐睿峰，2008年以来担任香港城市大学中文、翻译及语言学系研究员。哈尔滨工业大学学士毕业，香港理工大学计算机科学系硕士和博士毕业。论文及博士后工作侧重于中文词搭配，本书也集成了其中一些研究成果。现从事文本挖掘、意见分析及信息抽取。

张正生，自1990年起担任圣地亚哥州立大学语言学和亚洲/中东语言系副教授。首都师范大学（北京）英语专业学士毕业，俄亥俄州立大学语言学硕士和博士毕业。研究方向包括中文语言学（音调音韵、方言、功能语法/语篇分析和中文文字），外语教学及语言教学技术的使用。目前在语料库统计的基础上研究中文语音和写作的变化模式。现任《汉语教师协会》期刊主编。

译 者 序

自然语言处理是计算机科学领域与人工智能领域中的重要方向，也是一门融合语言学、计算机科学、数学和认知科学于一体的交叉学科。它主要研究实现人与计算机之间用自然语言进行有效处理和理解的各种理论和方法。自然语言词汇量大，规则复杂，处处充满歧义，但它是人类最重要的交际工具，也是人类思维、文化和一切知识的载体。因此，自然语言处理研究是构建真正的人工智能体系不可或缺的重要内容，也被誉为“人工智能皇冠上的明珠”。

中文是世界上使用人数最多的语言，全球化和互联网的出现大大提高了中文用户的参与度，这使得中文成为在过去十年里全球商业和社会生活中增长最高的线上语言。尽管中文自然语言处理的需求和重要性越来越大，但中文的独特性使得其相应的计算机处理技术富有挑战性。中文自然语言处理既需要解决处理各种语言的共性问题，也特别需要解决自身特有的形态、句法和语义带来的特殊问题。长期以来，世界范围内，特别是来自中国的研究者围绕中文自然语言处理进行了大量的研究，但一直缺乏一本面向全球自然语言处理学者的专著对中文自然语言处理的方法和技术进行全面系统的分析和总结。由香港中文大学 Kam-Fai Wong、香港理工大学 Wenjie Li、香港城市大学 Ruifeng Xu、美国圣地亚哥州立大学 Zheng-Sheng Zhang 所共同完成的《Introduction to Chinese Natural Language Processing》填补了这一空白。该书是世界上第一本全面介绍中文自然语言处理技术的英文专著，也是迄今为止中文自然语言处理领域中最系统全面的英文著作之一。该书的出版有助于增进世界范围内研究人员对中文自然语言处理的兴趣和了解，推动相应研究和产业应用的全球化。该书重点涵盖了中文自然语言处理的特殊性问题，包括中文的字和构词法、中文分词和未登录词识别、中文的语义表示和知识库，以及中文搭配等内容。该书自出版以来获得了多个国际知名科研机构、多位著名学者的一致好评，认为该书既可以作为刚刚进入这一领域的学生、学者、开发者的导论，同时也是了解这一领域

的前沿动态的可靠途径。

为了帮助国内读者更好地利用这一著作，由该书的原作者徐睿峰、黄锦辉，以及李斌阳共同将该书翻译为中文版，特别感谢杜嘉晨、范创、陈荻、陈刚保、高清红、张庆林、巫继鹏、吕秀程、龚镘霖、高存、李萌、孙雨佳、徐嘉莹等同学在内容整理、材料编辑、翻译校对等方面的贡献。感谢哈尔滨工业大学（深圳）计算机科学与技术学院与国际关系学院信息科技学院为本书的出版提供的良好条件，感谢国家自然科学基金、国际关系学院中央高校基本科研业务费专项资金和深圳市技术攻关等项目的资助。

限于译者的水平，译文中难免有错误和不足之处，敬请各位读者批评指正。

译者

2018 年 10 月

目　录

第1章 介 绍

“世界是平的”（Friedman，2005）。全球化的浪潮早已将国与国之间的边界淹没，而互联网的发展，正是推动这一进程的催化剂。当今的万维网（World Wide Web，WWW）不再被英文用户独占。在 2008 年，全球有 29.4%（约 4.3 亿）的互联网用户使用英语交流，紧排其后的是 18.9%（约 2.8 亿）的中文用户（Internet World Stats，2007）。自 2000 年以来，后者已增长了 755%，并且这一趋势没有减缓的迹象。

为了实现客户关系管理（Customer Relationship Management，CRM），全世界的商人正在积极地研究互联网，希望向来自多元民族和多元文化的客户提供更好的服务。政府也不甘落后，希望通过互联网为市民提供更好的服务。基于互联网的 CRM（通常被称为 Electronic CRM，e-CRM）广泛地采用自然语言处理技术（Natural Language Processing，NLP），分析来自不同语言的网页内容。随着中国市场的日益增长，越来越多的电子商务门户网站包含中文信息。因此，中文自然语言处理研究发展的需求也日益增长。

1.1 中文自然语言处理是什么

一门语言就是一个动态集合，其中包括符号及其对应视觉、听觉、触觉或文字交流的规则。人类的语言通常被称为自然语言，其科学研究属于语言学范畴，但其计算方式的实现则属于计算语言学领域。计算语言学侧重于人类语言的理论方面，而自然语言处理可以被视为一种语言理论的实现，以促进实际应用，例如，e-CRM 内容在线分析、机器翻译、信息抽取、文档摘要等。自然语言处理有三个基本任务：形态分析、句法分析和语义分析。

在中文自然语言处理中，中文句子的词之间缺乏明确的分隔符，这与英语等西方语言有明显的区别。因此，句子自动分词是中文形态分析的重要步骤，

也是任何中文信息系统的基础，以下面的句子为例：

香港人口多
白天鹅已飞走了

对于人来说，这些句子很容易分词，即通过认知其中的词来了解整句的含义，例如：

香港　人口　多
白天　鹅　已　飞走了

然而，对于计算机来说，自动分词并不这么简单。一些分词算法，如基于词典的从右至左最大匹配算法，就有可能对上述例子产生不一样的分词结果，例如：

香港　人　口多
白　天鹅　已　飞走　了

可以看到，“人口多”可以有两种分词结果，分别是“人口　多”和“人口多”。而“白天鹅”也可以被切分为“白天　鹅”及“白　天鹅”。解决分词歧义，可以利用基于规则的算法，结合语法或常识进行分词。上述例子分词的歧义性可以通过使用如下简单的规则集解决：

（1）“人口”在日常使用中更加普遍，而“口多”是香港地区的一个俚语，因此前者的分词优先级更高。

（2）“天鹅”会飞，但“鹅”通常不会飞，因此前者的分词优先级更高。

由此可见，若想让计算机准确分词，上述两个句子都需要额外的语言学知识。虽然这些知识资源中有些是可用的，但对于计算机来说都是不可读的，这就导致了中文自然语言处理的一大难点。

以下句子的分词，结合了词性（Part-of-Speech，POS）标注的结果。这需要一份词性词典，其中每一个词有一个或多个词性，例如，中文的“计划”既可作动词，也可作名词。词性的确定取决于词在原句中的位置，例如：

他　计划　去香港

去 香港 是 他 的 计划

第一个“计划”是动词，而第二个“计划”则是名词。实际上，几乎所有的中文双音节词都有多个词性标注，例如，建议（动词/名词）、鼓励（动词/名词）等。

在分词和词性标注结果的基础上，利用句法分析可以构建句子的结构，不同于原句中词的线性结构，这种句子结构是有层次性的。以“同学会认为她是班长”为例，分词和词性标注的结果如下：

同学会（n）认为（v）她（r）是（v）班长（n）

分别对应名词、动词、代词、动词和名词。然而，这种线性结构对于自然语言处理是无用的。因此，这种结构被转换为一棵句法树，其中包含一组预定义的规则集及其对应的句法分析算法，上述例子的句法树如图 1-1 所示。与计算机程序的编译类似，该句法树为计算机的进一步处理提供了核心的结构信息。例如，一个机器翻译（Machine Translation，MT）系统会尝试理解句法树，并在句子语义的基础上，将其翻译成目标语言的句子。同样地，信息检索系统会从句法树中抽取关键概念用于构建索引。

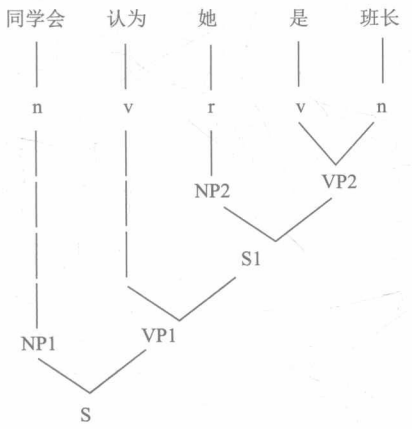

图 1-1 “同学会（n）认为（v）她（r）是（v）班长（n）”的句法树

需要明确的是，这种结构是在线性词性标注序列的基础上得到的。实际上，词性标注序列比词序列更加抽象，其语言覆盖范围相对更广泛。例如，“黄先

生（n）知道（v）你（r）是（v）律师（n）”和“政府（n）承认（v）他（r）是（v）专家（n）”有一致的句法树。“工会（n）推选（v）他（r）当（v）主席（n）”虽然和前两个例子有着同样的词性标注序列，但是其句法树结构却截然不同（图 1-2）。这就在句法分析的层面上不可避免地产生了歧义。因此，需要在语法知识和常识等额外信息的基础上，从两棵句法树中选择合适的结构进行消歧。

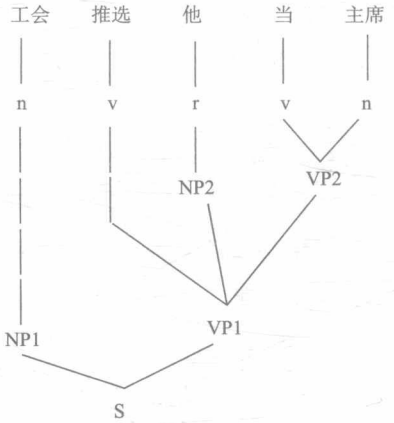

图 1-2　“工会（n）推选（v）他（r）当（v）主席（n）”的句法树

歧义产生的原因，除了一种线性词性标注序列可能对应多种句法树，还可能由一棵句法树有多重含义导致。由此引出语义分析，其基本目标是找出句子的含义，而句子的含义是不能从给定的句法树中直接推导得到的。例如，“音乐家（n）打（v）鼓（n）”“妈妈（n）打（v）麻将（n）”和“运动员（n）打（v）网球（n）”，这三个句子有着相同的句法树（图 1-3）。

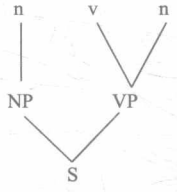

图 1-3　三个句子的句法树

那一个自然语言处理系统该如何提取这三个句子的含义呢？通过简单分

析给定的句法树进行语义提取几乎是不可能的。实际上，句子背后隐含的语义线索可以用于解决歧义问题。这三个句子中的动词都是“打”，差异体现在它们的主语不同，分别对应“音乐家”“妈妈”和“运动员”。然而，如果有一个语义词典，包含以下条目：

① 打1（音乐家，乐器），即音乐家演奏乐器，鼓是一种乐器。

② 打2（妈妈，家庭游戏），即妈妈玩家庭游戏，麻将是一种家庭游戏。

③ 打3（运动员，运动），即运动员做运动，网球是一种运动。

自然语言处理系统就可以在此基础上更好地理解这些句子的语义。可见，语义知识资源是语义提取的先决条件，但是，与英语等西方语言相比，中文自然语言处理的这种资源相对较少。很多研究人员为构建不同的中文语言资源付出了很大的努力。

本节介绍了中文自然语言处理与其他语言自然语言处理的差异，这奠定了本书的主旨——介绍中文形态分析的基本技术。

1.2 关于本书

中文形态分析是中文自然语言处理的基石，也是本书的侧重点。本书分为3个部分：基础概念（第2章和第3章）、词的自动识别（第4章和第5章）和中文词汇语义（第6章～第8章）。

第2章从语言学的角度，介绍了中文的字、语素、词等基本概念。第3章概述了在自然语言处理应用中需要考虑的中文词汇特性。这两章为后面章节的内容做了铺垫。第4章介绍了分词存在的问题及对应的技术解决方案，紧随其后的是第5章的未登录词（Out of vocabulary，OOV）识别。第6章中，引入词义的概念，并介绍了几个包含词义信息及词汇联系的中文自然语言处理的语义资源。第7章和第8章分别论述了中文搭配的概念和搭配自动抽取的相关技术。

第2章 中文的词

2.1 引　言

鉴于本章内容是关于中文的语素构成，因此首先要解决以下一些基本问题。

- 中文里的词是怎样表示的？
- 中文里的词是怎样构成的？
- 中文里的词是怎样被识别的？

尽管中文构词法与其他语言并非全然不同，但其独特之处在于中文构词法多以形态合成法为主，并辅以相对较少的词缀。同时，中文还采用了其他语言中并不常见的叠词形式来构词。

由于中文采用了相对独特的书写体系，因此词的识别对中文自然语言处理来说是一项特殊的挑战。在英语和其他许多语言中，每个词都由一串字母对应，词之间由空格隔开。而中文却不能用同样的方法来识别词，因为在书写中文时，每个词单元之间没有空格作为间隔，所以必须通过分词使连续书写的汉字分割为词块。

中文独特的书写体系和书写习惯使得对书写单元(即汉字)和口语单元(即词和语素）的区分非常必要，而这两者都是形态处理的基本单元。

2.2 字、语素与词

2.2.1 字

中文最突出的特点是它由汉字组成，汉字并不像其他语言的字符由字母符号组成，而这就对中文自然语言处理产生了很大的影响。汉字在视觉上与

字母书写体系存在很大的不同，每个字不像字母一样以线性方式组合，而是排列在一个方块形状内，每个字占用相同的空间，如国家的“国”字，因此人们称其为方块字。

尽管汉字的起源是象形文字，并且人们普遍认为汉字是一种象形文字，其实很大程度上并非如此，汉字中只有少部分是象形字。常见的例子有“日”“月”“水”“火”“人”等。同样，直接表示意义的表意字在汉字中也较少。表意字包括指示字，如“上”“下”“一”“二”“三”，以及会意字，如“仁”“信”。事实上90%以上的汉字都是形声字，即每个字都由表示读音的一个声旁和一个形旁组成。形旁只表示该字的大概含义，用以区分它的同音异形异义字，而不是准确地指出字的含义。如“请”“情”“清”“晴”这四个形声字，它们有“青”这个相同的声旁，但各自还有“讠”“忄”“氵”“日”这样的形旁。尽管许多汉字的读音可通过声旁推测，但其意义并不能由字形或字的结构判断。

从功能上来说，汉字与字母书写体系不同的是，每个字并不表示独立的音位。正如美国语言学家、汉学家德范克在1984年指出的，中文是形态音节型的文字，即每个字都是一个单音节，同时也是最小语义承载单位——语素。

虽然文本处理不可避免地涉及对汉字进行处理，但我们应该清楚地认识到汉字并不等同于语言单元，汉字这类书写中所用的符号，本身并非语言单元。本章主要关注中文文本的形态处理，包括与之相关的语言单元的素和词。而将口语单元与书写单元中的字进行区分是十分重要的。尽管字在社会学、心理学上有非凡的意义，在视觉上也更易识别，但它不能等同于根据语义和句法来识别的语素和词。大多数汉字都是单音节的，而语素和词的长度是可变的，如用一个音节表示的“山”，用四个音节表示的“密西西比”，等等。

2.2.2 语素

语素是最基本的形态单元和最小的语义单位。语素可以用来构成词，但不能再被分解为更小的且具有意义的单位。例如，在英文句子“the company owns many properties”中，“the”“company”“many”这三个词各具有一个语素，不能继续进行分解，而“owns”和“properties”都可以分解为两个语素：“owns”由“own”和表示第三人称单数的词尾“-s”组成，“properties”由“property”

和表示复数的词尾“-s”组成。由此可以看出，一个语素可以由一个以上的音节（如“many”“property”）或单个音素（如“-s”）组成（音素 phone 即组成音节的单位，“不到一个音节”的说法不常见）。

与英语相反，中文中的语素往往是单音节的，每个语素/音节写为一个汉字。然而在一个音节、一个语素对应一个汉字的一般情况之外还有一些特例，有一些语素由两个音节组成，如“葡萄”“菩萨”“马虎”“马达”“咖啡”“萨其马”“巧克力”等。尽管表示这些音节的汉字在别的语境中具有独立的含义，但在这些词中，它们只是组成读音，即这些汉字与词的意义没有联系，因此这些词中分别只有一个语素。这种例外源于它们是借自其他语言的外来词，原始的中文词大部分遵循“音节-语素-汉字”的对应关系。

2.2.3 词

词与更小的单位语素和更大的单位短语有着明显的区别。词在形态学单元的独立性，以词的分布约束性和词义完整性为依据。

分布约束性：一个词可以单独地出现（如单个词构成的句子），而语素不能。当语素能够单独出现时，它是一个有着单个自由语素的词；而如果不能，则它是一个粘着语素，必须与其他语素结合才能组成词。在上一节的例子中，英语的复数词尾“-s”和第三人称单数词尾“-s”都不是自由语素，它们必须粘着于一个名词或动词。粘着语素的例子在中文中也很常见，所有的语法语素“的”“地”“得”“了”“着”“过”都是粘着的，不能单独出现。一些实义语素也是粘着的，如“行”“饮”，尽管在古汉语中它们可以作为自由语素，但在现代汉语中它们不能独立出现，必须组成“行人”“行动”“冷饮”或“饮料”等词。

词义完整性：与自由度相对较大的句法一样，词也很难从组成它的语素中推测出它的含义。例如，“黑板”中含有语素“黑”，但实际上不一定是黑色的。同样，“大人”的含义并不等同于“大的人”，因为还可以有“小大人”这样的用法。“小人”也不等同于“小的人”，前者含义为“卑鄙小气的人”，而后者指的是与“大的人”相对的“小的人”。词内部的语义关系与词之间的语义关系也不同，即在短语这个层级中，“打人”（打＋人）是一个动宾短语，但“打手”（打＋手）并不是对手进行击打的意思，而是一个名词。

常使用双音节词。有一种谬见认为中文是一种单音节的语言，即中文中的词都是一个音节长度。德范克在他的著作《中国语文：事实与幻想》中，对这种观点进行了有力的辩驳，而赵元任认为中文确实具有单音节性。这种单音节性究竟是真理还是谬误，取决于如何理解词或语素。绝大多数中文语素长度确实为一个音节，多音节语素只占所有语素的11%。可以说大部分中文语素是具备单音节性的，但只有44%的单音节语素能够作为词独立存在，所以说中文是一种单音节性的语言是不准确的。

实际上，大部分中文词汇都有两个音节。根据吕叔湘的观点，有大量的统计结果和事实都证明了这一点：

- 存在包含同义语素的双音节复合词，如“保护”“购买”“销售”。多余的音节或语素似乎并没有改变词义。
- 许多缩略词（中文中存在大量缩略词）为双音节，如“北京大学”——“北大”。
- 单音节的地名会加上范畴的标记，如“法国”“英国”“通县”“涿县”等。而双音节的名称则不会加上这种标记，如“日本”“印度”“大兴”“顺义”。数字亦是如此，单音节的1到10往往会加上单位，如“一号”“十号”，而双音节数字如“十一”则不会。

在书写地址时，姓名的使用也是一种有趣的方式。中国人的姓为一个音节或两个音节，名也是如此，所以一个人的姓名的长度为二至四音节。当要礼貌或亲密地称呼某人时，人们多趋向于用两个音节，甚至有时候超出了对于礼貌和亲密性的考虑。如果姓只有一个音节，则加上表示敬称或爱称的前缀“老”或“小”，如“老李”“小李”。但如果是双音节的姓氏如“欧阳”“端木”则不能加上这样的前缀。如果称呼名字，则有两种情况。如果名字为两个音节，则只称呼名字；如果名字为一个音节，且姓氏也为一个音节，则称呼全名。

2.3 词的构成

在中文中，词可以分为单纯词和合成词两大类。单纯词是由一个词根语素组成，如“人”“手”“车”“坦克”“枇杷”等。合成词是由两个以上的语素构成，根据语素的特点，合成词又分为派生词和复合词。派生词的意义是在