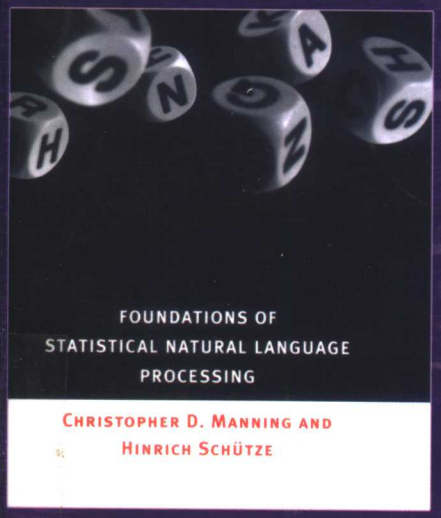
国外计算机科学教材系列

统计自然语言处理基础

Foundations of Statistical Natural Language Processing



[美] Christopher D. Manning
[德] Hinrich Schütze 著

苑春法 李庆中 王 昀
李 伟 曹德芳 等译

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

国外计算机科学教材系列

统计自然语言处理基础

Foundations of Statistical Natural Language Processing

[美] Christopher D. Manning
[德] Hinrich Schütze
著

苑春法 李庆中
王 昀 李 伟 曹德芳
等译

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内容简介

近年来，自然语言处理中的统计学方法已经逐渐成为主流。本书是一本全面系统地介绍统计自然语言处理技术的专著，被国内外许多所著名大学选为计算语言学相关课程的教材。本书涵盖的内容十分广泛，分为四个部分，共16章，包括了构建自然语言处理软件工具将用到的几乎所有理论和算法。全书的论述过程由浅入深，从数学基础到精确的理论算法，从简单的词法分析到复杂的语法分析，适合不同水平的读者群的需求。同时，本书将理论与实践紧密联系在一起，在介绍理论知识的基础上给出了自然语言处理技术的高层应用（如信息检索等）。在本书的配套网站上提供了许多相关资源和工具，便于读者结合书中习题，在实践中获得提高。

本书不仅适合作为自然语言处理方向的研究生的教材，也非常适合作为自然语言处理相关领域的研究人员和技术人员的参考资料。

版权贸易合同登记号　图字：01-2002-6457

图书在版编目（CIP）数据

统计自然语言处理基础/（美）曼宁（Manning, C. D.）等著；苑春法等译.
北京：电子工业出版社，2005.1
（国外计算机科学教材系列）
书名原文：Foundations of Statistical Natural Language Processing
ISBN 7-5053-9921-7

Ⅰ.统...　Ⅱ.①曼...　②苑...　Ⅲ.统计方法-应用-自然语言处理-教材　Ⅳ.TP391

中国版本图书馆CIP数据核字（2004）第129789号

责任编辑：马　岚　　特约编辑：马爱文
印　　刷：北京兴华印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
经　　销：各地新华书店
开　　本：787 × 1092　1/16　　印张：27　　字数：691千字
印　　次：2005年1月第1次印刷
定　　价：55.00元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系电话：（010）68279077。质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

出 版 说 明

21世纪初的5至10年是我国国民经济和社会发展的重要时期，也是信息产业快速发展的关键时期。在我国加入WTO后的今天，培养一支适应国际化竞争的一流IT人才队伍是我国高等教育的重要任务之一。信息科学和技术方面人才的优劣与多寡，是我国面对国际竞争时成败的关键因素。

当前，正值我国高等教育特别是信息科学领域的教育调整、变革的重大时期，为使我国教育体制与国际化接轨，有条件的高等院校正在为某些信息学科和技术课程使用国外优秀教材和优秀原版教材，以使我国在计算机教学上尽快赶上国际先进水平。

电子工业出版社秉承多年来引进国外优秀图书的经验，翻译出版了“国外计算机科学教材系列”丛书，这套教材覆盖学科范围广、领域宽、层次多，既有本科专业课程教材，也有研究生课程教材，以适应不同院系、不同专业、不同层次的师生对教材的需求，广大师生可自由选择和自由组合使用。这些教材涉及的学科方向包括网络与通信、操作系统、计算机组织与结构、算法与数据结构、数据库与信息处理、编程语言、图形图像与多媒体、软件工程等。同时，我们也适当引进了一些优秀英文原版教材，本着翻译版本和英文原版并重的原则，对重点图书既提供英文原版又提供相应的翻译版本。

在图书选题上，我们大都选择国外著名出版公司出版的高校教材，如Pearson Education培生教育出版集团、麦格劳－希尔教育出版集团、麻省理工学院出版社、剑桥大学出版社等。撰写教材的许多作者都是蜚声世界的教授、学者，如道格拉斯·科默（Douglas E. Comer）、威廉·斯托林斯（William Stallings）、哈维·戴特尔（Harvey M. Deitel）、尤利斯·布莱克（Uyless Black）等。

为确保教材的选题质量和翻译质量，我们约请了清华大学、北京大学、北京航空航天大学、复旦大学、上海交通大学、南京大学、浙江大学、哈尔滨工业大学、华中科技大学、西安交通大学、国防科学技术大学、解放军理工大学等著名高校的教授和骨干教师参与了本系列教材的选题、翻译和审校工作。他们中既有讲授同类教材的骨干教师、博士，也有积累了几十年教学经验的老教授和博士生导师。

在该系列教材的选题、翻译和编辑加工过程中，为提高教材质量，我们做了大量细致的工作，包括对所选教材进行全面论证；选择编辑时力求达到专业对口；对排版、印制质量进行严格把关。对于英文教材中出现的错误，我们通过与作者联络和网上下载勘误表等方式，逐一进行了修订。

此外，我们还将与国外著名出版公司合作，提供一些教材的教学支持资料，希望能为授课老师提供帮助。今后，我们将继续加强与各高校教师的密切联系，为广大师生引进更多的国外优秀教材和参考书，为我国计算机科学教学体系与国际教学体系的接轨做出努力。

电子工业出版社

教材出版委员会

译 者 序

近年来，统计自然语言处理（或称统计语言学）异军突起，现已成为自然语言处理研究中的主流。在统计自然语言处理学科成长的过程中，有四个因素起着推动作用：

1. 由于计算机硬件的发展，使大容量的存储和高速计算已经成为可能；
2. 由于计算机网络的普及，大量电子文本在网络上的涌现，使语料的获取不再困难；
3. 机器学习学科本身的发展日趋成熟，并在许多领域得到了广泛应用，因此它在自然语言处理中的应用已经成为很自然的事情；
4. 由于自然语言本身的复杂性，即使是语言学家也很难用纯粹的人工规则（或规律）来刻画它，这就迫使我们从实际语料中学习语言规律。

统计自然语言处理的研究涉及了传统自然语言处理的各个方面，例如语言分析、机器翻译、信息检索、文本分类等。可以毫不夸张地说，统计学习方法的引入大大促进了这些领域的研究和发展。目前国内几乎所有著名大学的计算机系都在从事这方面的研究（或开设了类似专业）。但是，系统地讲授或阅读这方面的专著并未得到学术界同行们的重视。在一次学术会议上，某校一位教授深有感触地说，“研究生在校学习期间一定要认真读一本专著。”我们对这位教授的发言深有同感。研究生们一定要看最新的参考文献，包括学术会议文章和杂志文章；但只看这些资料，不看（或学习）一两本专著，所学知识可能是支离破碎的，也未免有急功近利之嫌，尤其是对一些新兴学科更是如此。在这样的情况下的研究往往底气不足，不容易出一些像样的成果。在学术交流中，往往大家没有共同的语言，甚至闹出笑话。

本书是一本系统介绍统计自然语言处理（或统计语言学）的专著，在国外已经被许多大学用来作为教材。在国内，大家已经开始认识到了这本书的价值，不少大学将本书的英文版作为研究生教材。将这本专著翻译并介绍给国内广大从事自然语言处理研究的读者，具有重要的现实意义。本书覆盖了统计自然语言处理的各个领域中最重要的主题，内容详尽，层次清楚。无论是对于从事信息检索、机器翻译、文本分类和语言分析等方面的研究的人员，还是对于计算语言学专业的本科生和研究生，本书都有着非常重要的参考价值。

本书由清华大学计算机系的苑春法组织翻译。苑春法长期从事统计自然语言处理相关领域的研究和教学工作，对该领域里的问题有一定深度的了解。参译者也都在该领域里具有一定的研究基础和经历。本书第2章和第13章～第16章由李庆中初译，第1章和第5章～第8章由王昀初译，第3章和第9章～第12章由李伟初译，前言部分和第4章由曹德芳初译。最后，全书由苑春法负责统一修改、审阅并定稿。在翻译本书的过程中，大家力求忠实于原著，在此基础上尽量把概念表达准确、清晰。黄昌宁教授对于本书的翻译工作给予了指导，闻扬、周剑辉、徐薇、翁耀、钱冬蕾和林静等人也做了部分内容的翻译和辅助性工作，在此一并表示感谢。

本书采用英文版第5次印刷的版本进行翻译，已经对照作者在网站上提供的勘误表对相关内容进行了更正或注解。由于译者水平有限，翻译中难免会出现一些不妥之处，希望广大读者批评指正。

前　言

现在是一个在线信息、电子通信和互联网流行的年代，一本详细介绍统计自然语言处理的教材的需求程度可能并非那么迫切。但是，我们应该看到，商业部门、政府机构以及个人正面对着越来越多与工作、生活密切相关的文本信息，而如何从这些大量文本中挖掘潜在的有使用价值的信息，仍然是一个难题。

与此同时，由于大规模的文本语料的可获得性，人们已经改变了语言学和认知科学中研究自然语言的方法理论。以前一些无足轻重的研究领域，以及一些支离破碎，显得很无趣或者难以觉察的语言现象，都逐渐成为当前研究的热点。然而，在20世纪90年代早期，定量方法还没有引起语言学研究者的足够关注，在当时的一本数理语言学权威教材中甚至根本没有提及这种方法，但现在它已经被看做语言学理论研究中极为重要的手段之一。

本书希望在理论和实践之间以及直觉和严密之间尽可能地达到一种平衡。具体地讲，就是我们以数学和语言学作为基础来阐述各种理论方法，同时为了避免材料过于枯燥，努力做到理论方法和实际问题的紧密结合。为了给读者奠定必要的基础知识，我们首先介绍概率论、统计学、信息论和语言学方面的重要概念，使读者能够正确理解并增加这些领域的知识；然后介绍统计自然语言处理中存在的问题，比如标注和消歧问题，还将选择一些重要的研究问题进行讲述，从而使读者能够进一步理解语言学研究中存在的特殊问题，为更加深入的研究工作提供必要基础。

当初我们设计本书的基本结构时，对于应该包含什么素材以及如何组织这些素材都进行了细致的考虑。其中一个重要的标准就是尽量不要使本书篇幅过长（我们没有完全成功做到这一点）。另外，本书并不打算全面地介绍概率论、信息论、统计学和统计自然语言处理中涉及的其他领域的数学知识。但是，我们尽力做到使本书覆盖统计自然语言处理各个领域中最重要的主题。对于那些对数学基础有特殊兴趣的读者来说，需要参考本书之外的其他资料来进行更加深入的研究。

我们也尽量避免使用均匀的笔墨来描述统计自然语言处理以及用到的数学工具和理论。虽然一个内在一致的数学理论很重要，但实际上这种理论在这个领域中并不存在，这就导致了在一些地方使用了折中的混合理论。但是，我们可以肯定地说，在自然语言处理中的某种方法也许是对的，但就此断定该方法优于其他方法还为时过早。

本书没有包含语音识别的内容也许会让读者有些出乎预料。这样安排主要是因为我们考虑到，对于自然语言处理来说，语音识别是作为一个相对独立的领域从电气工程专业分离出来的，拥有自己的会议和期刊以及自己的相关研究。然而，最近几年随着研究领域的交叉和互相渗透，统计方法在语音识别中的成功应用激发了自然语言处理中应用统计方法的热潮，本书介绍的许多技术方法都是首先在语音识别中应用，然后慢慢扩展到自然语言处理领域的。特别是语音识别中语言模型的有关工作和本书中语言模型的讨论，在很多方面有共同之处。甚至可以这么说，语音识别是自然语言处理领域中当前最为成功而且应用最为广泛的。但是，有些合理的理由把语音识别排除在本书内容之外：已经有一些比较好的关于语音的教材，而且语音

也不是我们专门研究或者特别擅长的领域，况且即使本书不包括语音的内容也显得有些篇幅过长。另外，虽然两者内容有所交叉，但是差别也很明显：语音识别的教材需要包含信号分析和声学模型方面的内容，对于一个具有计算机科学或者自然语言处理背景的人来说，这些内容并非他们感兴趣或者可以理解的；反之，许多研究语音识别的人可能对我们提到的自然语言处理的主题并不感兴趣。

和统计自然语言处理稍微有些关联的其他领域包括机器学习、文本分类、信息检索和认知科学。在所有这些领域中，都可以找到一些本书中没有包含但是却非常适合本书的例子。由于篇幅所限，我们没有包含一些重要的概念、方法和程序，比如最小描述长度、回溯算法、Rocchio 算法，以及和语言处理的频率效应相关的心理学和认知科学文献。

如何严格区分统计和非统计自然语言处理是一件很困难的事情。开始写这本书的时候，我们相信，两者之间有一条很明显的分界线，但是最近这条线变得越来越模糊了。越来越多的非统计学研究者们采用了语料库证据和一体化的定量方法。在统计自然语言处理中，大家逐渐接受了这样一种观点：当处理某种语言现象时，可以使用和该现象相关的所有可获得的科学知识来构造一个概率模型或者其他模型，而不是简单地采用所谓忽视这类已知知识的方法。

许多自然语言处理的研究者们都对单独编写一本统计方法书籍是否必要提出了质疑。在本书中，最后一件工作就是要改变认为语言学理论与符号计算工作和统计自然语言处理无关的错误看法。然而，我们相信，由于需要涉及这么复杂的基础材料，所以很难写出一本篇幅可以控制，让读者满意并且详细介绍所有自然语言处理知识的教材。此外，还存在许多其他很好的文章，如果对统计和非统计的方法之间需要更多的平衡，我们推荐阅读这些补充资料。

最后要说一下本书的书名“Foundations of Statistical Natural Language Processing”。那些从标准统计学了解统计方法的定义的人可能会对书名有些疑问。我们定义的统计自然语言处理由所有的自动语言处理的定量方法组成，包括概率模型、信息论和线性代数。概率论是统计推理的基础，本书中把术语“统计”的基本含义稍微扩大了一点，即包含处理数据的所有定量方法（一个可以在几乎任何词典中确认的定义）。统计自然语言处理在过去 20 年中是使用得最广泛的一个术语，用它来代表自然语言处理中非符号化和非逻辑的工作，尽管有潜在的可能引起模棱两可的理解，但我们还是决定继续使用这个术语。

致谢

在撰写本书的这 3 年中，许多同事和朋友都为早期的草稿做过注释或者提出过建议。我们想向他们表达感激之情，特别是要感谢：Einat Amitay, Chris Brew, Thorsten Brants, Andreas Eisele, Michael Ernst, Oren Etzioni, Marc Friedman, Éric Gaussier, Eli Hagen, Marti Hearst, Nitin Indurkhya, Michael Inman, Mark Johnson, Rosie Jones, Tom Kalt, Andy Kehler, Julian Kupiec, Michael Littman, Arman Maghbouleh, Amir Najmi, Kris Popat, Fred Popowich, Geoffrey Sampson, Hadar Shemtov, Scott Stoness, David Yarowsky 和 Jakub Zavrel。另外，我们要特别感谢 MIT 出版社的 Bob Carpenter, Eugene Charniak, Raymond Mooney 以及一位不知姓名的审稿者，他们对内容和说明都提出了许多改进建议，本书由于他们的建议在整体质量和可用性方面都有了很大的改进。我们希望即使没有特意致谢，当他们注意到书中有些想法来自于他们的建议后，也会感受到我们的感激之情。

我们也同样要感谢：Francine Chen, Kris Halvorsen 和 Xerox PARC，感谢他们对本书第二作者

的支持;感谢 Jane Manning 对第一作者的爱和支持,感谢 Robert Dale 和 Dikran Karagueuzian 对这本书的设计建议,感谢 Amy Brand 作为编辑对我们的经常性的帮助和协助。

反馈

我们尽力使本书做到通俗易懂、内容广泛且正确,但毫无疑问,在许多地方我们还可以做得更好。我们非常欢迎读者发 E-mail 提出反馈意见:

cmanning@acm.org 和 hinrich@hotmail.com

总之,我们希望本书可以让有潜力的学生获益并得到启发。本书收集了统计自然语言处理领域的许多方法,并且用一种容易理解的方式呈现出来。希望本书能够对这个领域的持续快速发展起到一定的作用。

Christopher D. Manning
Hinrich Schütze
1999 年 2 月

阅读指南

本书适合的读者是将要学习一学期“统计自然语言处理”课程的研究生。对于一学期要包含的内容而言，本书显然提供了过多的资料，但是这也给教师提供了很大的空间，他可以挑选自己认为有用的内容来讲授。本书假设读者已经有了一定的编程经验，熟悉形式语言和符号分析方法。同样，我们假设读者了解一些基本的数学概念，比如集合论、对数、向量、矩阵、求和以及积分方法。事实上，我们的要求也只比高中水平稍高一些。读者可能已经选修过符号自然语言处理方法方面的课程，但是阅读本书并不要求更多的背景知识。在概率、统计和语言学方向，我们试图简要概括所有必要的背景知识，因为以我们的经验来看，很多想学习统计自然语言处理方法的人，以前没有这些领域的相关知识（以后这种情况可能会有所变化）。然而，对于一个学生来说，要想打下适当的基础，还必须学习这些领域的一些补充资料，这样将来才有可能成为一个合格的研究者。

阅读本书或者使用本书来教学的最佳方法是什么呢？本书的内容分为四个部分：基础知识（第一部分），词法（第二部分），语法（第三部分）以及应用与技术（第四部分）。

第一部分主要是数学和语言学基础，后续部分都建立在第一部分的基础上。这里介绍的概念和技术在整本书中都会用到。

第二部分包括了统计自然语言处理中以词为中心的工作。这个部分很自然地从简单到复杂，渐进地介绍了语言学现象，分成 4 章分别讲述搭配、n 元语法模型、语义消歧和词汇获取，每章也可以独立阅读。

第三部分的 4 章分别讲述马尔可夫模型、标注、概率上下文无关文法和概率句法分析，它们是依次建立在前一章的基础上的，因此最好按照顺序阅读。但是，关于标注的这一章可以单独阅读，偶尔需要参考一下马尔可夫模型的内容。

第四部分的主题是 4 种应用与技术：统计对齐和机器翻译、聚类分析、信息检索和文本分类。它们之间很少有关联，所以根据兴趣和时间，这些章节可以分别单独阅读。

尽管在本书的第一部分介绍了很多背景知识和基础资料，但是如果把本书作为教材，我们并不主张在课程开始的时候仔细学习所有章节。作者通常会把一门课程的前 6 个学时里需要回顾的主要内容写在第一部分。这些必要的内容包括概率（2.1.8 节）和信息论（2.2.7 节），以及一些必要的实用知识（第 4 章包含一部分这样的知识）。我们把第 3 章作为一个阅读作业留给那些没有多少语言学背景的读者。语言学概念方面的知识在许多章节中都需要，特别是第 12 章，教师讲到这里时最好复习一下句法概念。在授课过程中，把前面其他章节中的一些资料仅仅当做需要了解的内容进行介绍即可。

关于第二部分主题的组织结构，我们尽可能使之满足以下的要求：在课程的早期给出一些比较容易理解的主题，特别是可以为学生们写程序、做项目打好基础的主题。搭配（第 5 章）、语义消歧（第 7 章）和附着消歧（8.3 节）成功地达到了这个要求。我们在本书中很早就介绍附着消歧，表明了语言学概念和结构在统计语言处理中的地位。第 6 章中有许多详细的参考资料。对于应用（如语音识别或者光学字符识别）感兴趣的读者希望能够包含全部的相关内容，

但是如果 n 元语法模型不是一个特殊的兴趣焦点，读者可能只想读一遍 6.2.3 节。这些对于理解似然性、极大似然估计、两种简单的平滑方法（如果学生想自己构造概率模型，这些方法非常必要）以及评定系统性能的好方法来说足够了。

我们努力提供了丰富的交叉引用，如果需要，教师可以把大多数章节结合前面某处适当的资料独立讲述。特别地，对于搭配、词汇获取、标注和信息检索这些章节，也使用了同样的方法策略。

习题

每一章都附有习题。在难度和范围方面，这些习题相差很多。我们对它们进行了如下的基本分类：

★ 简单问题，范围从简单的文本理解到数学变换、简单证明和设想的一些例子。

★★ 比较实际的问题，包括编程和语料库研究。其中许多问题都适合作为两周以上的作业。

★★★ 大的、难的和开放性的问题。其中许多问题适合作为一个学期的项目。

网址

最后，我们鼓励学生和教师从配套网站上获取更多参考资料。直接访问 MIT 出版社的网址 http://mitpress.mit.edu，然后搜索本书即可。

目　录

第一部分　基础知识

第二部分　词　法

第四部分　应用与技术

第一部分　基础知识

第1章 绪 论

语言学的目的是为了能够描述和解释我们周围的语言现象，比如对话、写作和其他媒体中的语言现象。人类如何获得、产生和理解语言，如何理解语言表达和物质世界之间的关系，如何理解用于交流的语言结构，所有这些都属于语言学的范畴。为了解释语言的结构，人们设计了一些规则来把语言表达结构化。这种规则方法有很长的历史，至少可以追溯到2000年以前，而在20世纪，语言学家探索更加细化的语法规则，试图描述什么是正确的和不正确的语言表达，他们使规则变得日趋正式和严格。

然而，上述概念存在一个很明显的问题：对于正确的语言表达，我们无法给出一个精确并且完备的特性，因而无法把它们和错误的语言表达清楚地区分开来。实际上 Edward Sapir 已经注意到了这一点，并且提出了著名的格言“All grammars leak”(Sapir 1921:38)。这是因为人们总是扩展和改变规则，以满足他们遇到的语言交流的需要。但是，无论如何这些规则的扩展和改变不是完全没有道理的。例如，一个语言有这样的句法规则：基本英语名词短语是由一个任意的限定词、几个形容词和一个名词组成的。但为了语言使用的灵活性，我们需要放宽某些限制。

本书探求一种方法来解决上面的问题。我们不把句子分类为合乎语法的句子与不合乎语法的句子，而是着手来解决这样的问题：语言使用中通常出现的形式是什么。用来识别这些模式的主要工具是计数，就是通常所说的统计方法，因此概率论是本书的主要理论基础。而且，我们不仅仅把这类问题看做一个学术问题，而是更希望表明怎样建立一个统计语言模型并有效地使用它来处理许多自然语言处理任务。虽然实际应用的方法和基本理论有些出入，但是统计语言模型的有效性在逐渐证明这种基本方法的正确性。

进行统计自然语言处理，需要掌握一定数量的理论工具。在深入钻研大量理论之前，本章将花费一些时间解决统计方法在自然语言处理中的定位问题。因为本书中大量的篇幅都和这种方法有关，所以我们将首先考察一些哲学理论和重要思想，这些都是统计方法用在自然语言处理中的动因。并且，我们首先考察使用统计方法在文本中能学到些什么，获得我们的第一手资料。

1.1 理性主义者和经验主义者的方法

一些语言研究者和很多自然语言处理研究者的工作重点仅仅在文本层面，而不去思考语言表达的意思和语言书写形式之间的关系。对统计方法感兴趣的读者可能更注重实践部分，但是即使是注重实践的工作者，也不得不面对这样的问题：什么样的先验知识应该加入他们设计的语言模型。如果这样的先验知识和他们设想的有很大出入时，是否也要把它们加入到模型中。本节简要地讨论了这个问题的基础，即相关的哲学问题。

大约在1960年至1985年，语言学、心理学、人工智能和自然语言处理中的大部分研究完全被一种理性主义方法(rationalist approach)所支配。这种理性主义方法是由一种信仰决定的，人们相信在人类头脑中重要的知识不是由感官得到的，而是提前固定在头脑中，由遗传基因决定的。在语言学中，理性主义者已经占据了绝对的支配地位，因为人们已经广泛地认同了 Chomsky(乔姆斯基)提出的关于语言本能的观点。在人工智能领域中，理性主义者试图建立一个智能系统，他们