计算语言学
与语言科技
原文丛书

CAMBRIDGE

MEMORY-BASED LANGUAGE PROCESSING

基于记忆的语言处理

[比] Walter Daelemans [荷] Antal van den Bosch 著
孙 栩 导读

北京大学出版社
PEKING UNIVERSITY PRESS

MEMORY-BASED LANGUAGE PROCESSING

基于记忆的语言处理

[比] Walter Daelemans
[荷] Antal van den Bosch 著
孙 栩 导读

北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记号 图字:01-2014-3772

图书在版编目(CIP)数据

基于记忆的语言处理=Memory-Based Language Processing:英文/(比)戴勒曼斯(Daelemans,W.),(荷)博施(Bosch,A.V.D.)著.—北京:北京大学出版社,2015.7

(计算语言学与语言科技原文丛书)

ISBN 978-7-301-25909-2

Ⅰ.①基… Ⅱ.①戴…②博… Ⅲ.①自然语言处理—研究—英文 Ⅳ.①TP391

中国版本图书馆CIP数据核字(2015)第117666号

书　　名	基于记忆的语言处理
著作责任者	[比] Walter Daelemans　[荷] Antal van den Bosch　著
责任编辑	李　凌
标准书号	ISBN 978-7-301-25909-2
出版发行	北京大学出版社
地　　址	北京市海淀区成府路205号　100871
网　　址	http://www.pup.cn　　新浪微博:@北京大学出版社
电子信箱	zpup@ pup. cn
电　　话	邮购部 62752015　发行部 62750672　编辑部 62753374
印 刷 者	北京大学印刷厂
经 销 者	新华书店
	787毫米×980毫米　16开本　13.25印张　245千字
	2015年7月第1版　2015年7月第1次印刷
定　　价	32.00元

“计算语言学与语言科技原文丛书”由北京大学—香港理工大学汉语语言学研究中心、北京大学计算语言学研究所(由973课题“文本内容理解的数据基础”、863课题“大规模汉语语义基础资源库和知识库设计构建及工具平台”支持)和北京大学出版社合作推出

学术委员会
Academic Advisory Committee

主　任：

黄居仁(香港)

委　员：

Chris Manning (Stanford)
Maarten de Rijke (Amsterdam)
陈克健(台北)
李宇明(北京)
郭　锐(北京)
苏克毅(台北)
王厚峰(北京)
俞士汶(北京)
郑锦全(Urbana-Champaign)
Harold Somers (Dublin)
Suzanne Stevenson (Toronto)
冯志伟(北京)
陆俭明(北京)
石定栩(香港)
孙茂松(北京)
王士元(香港)
松木裕治(奈良)
邹嘉彦(香港)

编委会
Editorial Committee

丛书前言

“计算语言学与语言科技原文丛书”于2010年创立，2010 COLING国际计算语言学会议在北京举办之前出版了第一批图书。本丛书的出版象征着国内计算语言学研究与国际的接轨。国内学者正跻身计算语言学的国际舞台：一些资深学者已在COLING两个最主要的国际会议/组织中获选并担任重要的领导职务；而积极参与这些重要的国际会议也已在年轻学者中蔚然成风，他们已可谓是会议的主流参与者之一。在这样的氛围中，希望本丛书第二批图书的出版，能让国内有心投入语言科技与计算语言学研究的学者们如虎添翼，在国际舞台上创新并引导议题！

计算语言学（Computational Linguistics，CL）在语言科学与信息科学的研究中扮演着关键性的角色。语言学理论寻求对语言现象进行规律性的预测，做出完整的解释，计算语言学正好为这两点提供了验证与应用的大好机会。作为语言学、信息科学乃至于心理学与认知科学结合的交叉学科，计算语言学更为语言学基础研究与福国利民应用研究的接轨提供了绝佳界面。事实上，计算语言学与人类语言科技（Human Language Technology，HLT）可以视为体用两面，不可切分。

计算语言学研究的滥觞，其实源于上世纪五六十年代的机器翻译研究，中文计算语言学的研究也几乎同步开始。在美国伯克利加州大学研究室，王士元、邹嘉彦、C.Y. Dougherty等人1960年已开始研究中英、中俄机器翻译。他们的研究是与世界最尖端的科技同步的。国内中俄翻译研究也不遑多让，大约在20世纪50年代中期便已开始。可惜的是，这些中文方面早期机器翻译研究，由于硬件与软件的限制，未能有效传承下来。中文计算语言学研究比较系统的发展始于1986年。这一年，海峡两岸不约而同地分别成立了两个致力于建立中文计算语言学基础架构的研究群：北京大学的计算语言学研究所，在朱德熙先生倡议下成立，随后一段时间由陆俭明、俞士汶主持；而台北“中研院”的中文词知识库小组，由谢清俊创立，陈克健主持，黄居仁1987年回去后加入。

中文计算语言学的研究，近30年来已累积了相当可观的成绩。计算语言学的重要研究领域与议题中都能看到中文方面的相关研究成果，华人计算语言学学者也渐渐在国际学术界崭露头角。随着世界经济转向知识密集型

产业，跨语言跨文化沟通与知识整合成为知识产业的关键环节，语言科技的发展成为国际主流指日可待。在这个有利发展的大环境下，我们期待看到，中文计算语言学与华人计算语言学学者的成绩，百尺竿头更进一步，中文方面的研究可以进入计算语言学的学术核心，能够产生有能力引导议题并掌控研究方向的大师。

回顾国内的计算语言学发展，计算机科学的贡献多于语言学的贡献。这个现象，在理论与概率模型整合研究的趋势中，不免令人忧心。语言学的贡献弱，或许可以部分归咎于英文研究专著在国内不易取得；而比较容易取得的期刊或会议论文，在篇幅的限制下，又往往无法对理论做深入完整的铺陈，从而导致国内的年轻学者长于运算而拙于理据。因此，在期待大师与引领世界研究潮流两个方向，藉由英文专书来巩固研究理据，进而开拓研究视野，是非常重要的一步。

“计算语言学与语言科技原文丛书”的引进，就是在上述背景下促成的。个人忝为剑桥大学出版社“自然语言处理研究”（Studies in Natural Language Processing，SNLP）系列的主编，对于将此系列中较重要的几本书引入国内，责无旁贷。第二批出版的原文书，除了剑桥大学出版社的图书外，还有施普林格出版公司（Springer）语言科技系列中的几本书，以进一步拓展领域涵盖面。引进原书，原样出版，是容易的，然而若要真正搭建知识的桥梁，使国内学者与学生不仅能开拓研究视野，更能将原文著作的理论精髓应用于中文研究，则实在不易。因此，本系列每本书我们都邀请了一位专家撰写中文导读。这些导读可以说是本系列的精华、重点，使每本书比剑桥和施普林格的原本增加了不少附加价值。

每篇中文导读都包括三个重要的组成部分。第一部分是全书内容概要的介绍。导读专家都是长年浸淫于该领域的学者，他们能提纲挈领，并提供相关研究背景。因此，通过阅读导读，读者更易掌握并吸收该书的重要内容。第二部分是中文相关研究。原文著作不见得会提到相关的中文研究，由导读专家补充介绍，搭起理论与中文相关应用的桥梁，更能引导读者找到在这个议题进入中文研究的最佳切入点，让中文相关研究的开拓者的成绩更能发扬光大。第三部分重点在于补充原书出版后该领域研究的新发展。现代科技发展迅速，任何经典著作出版后，几乎马上就有新的相关研究。因此，在理论架构的脉络中，加上新近的发展，能使读者更贴切地掌握研究脉动。全书的内容摘要通常以文字叙述，而中文相关研究及最新研究发展则分别以文字叙述和延伸阅读书目的方式呈现。延伸阅读书目，可以使读者很快上手，进入相关研究领域，也是本系列的重要设计之一。

本丛书2010年出版第一批图书，现在出版第二批图书，必须感谢许多同行的付出。在规划出版的漫长过程中，北大计算语言学研究所的俞士汶老师及常宝宝老师一直无私无悔地支持。而香港理工大学的挹注，北大—理大汉语语言学研究中心石定栩、郭锐等几位的支持，使得整个系列能够顺利出版。此外，还要感谢北京大学出版社王飙主任、杜若明编审及李凌编辑，他们认同我们的宗旨，落实了丛书的出版工作。最后，感谢丛书的国内编委，特别是此次担任导读主笔的各位，正是他们脑力与心血的付出，才替读者们搭建了进入学术殿堂的平台。

丛书主编 黄居仁

谨志于香港，红磡

2014年9月

导　读

孙　栩

1　学科背景和内容提要

自然语言处理包括逻辑主义流派和经验主义流派。从20世纪90年代开始，经验主义流派在自然语言处理领域占据了主流。原因之一在于数据的大规模化，使得经验主义的方法能够获得大规模数据的支持。本书介绍的“基于记忆的语言处理”根植于经验主义流派。基于记忆的语言处理，简而言之，就是使用“基于记忆的学习”这种特殊的机器学习方法来处理语言。基于记忆的学习来源于这样一种假设或者说启发：当人们对一件事情进行分析判断的时候，往往会直接参考最相似的经验或者记忆，而不是从这些经验或者记忆中抽象出规则来进行判断。

本书作者强调，一系列的实验表明，基于记忆的自然语言处理在多个任务上获得了很好的效果，并初步解释了基于记忆的语言处理能够取得好效果的原因——语言处理数据常常含有大量互相冲突的数据模式，而且往往伴随着数据稀疏的特点。换句话说，很多语言现象不常见，而且互相冲突。现有的大部分方法是“急切(eager)”的方法，这些方法总是试图从数据中抽象出规则或者简单的参数模型，并且过滤掉冲突的、不常见的数据模式。这类急切方法难以在语言处理任务上取得好效果，因为这类数据模式冲突、低频率的语言现象其实不是噪音，而是重要的语言现象，所以不能在抽象、泛化过程中过滤掉这些大量而特殊的语言现象。基于记忆的学习技术恰恰解决了这个问题——基于记忆的学习是一种“懒惰(lazy)”的学习方法，因为其学习过程非常简单，就是简单地把所有的训练数据保存在“记忆”中，从而不存在急切方法所具有的抽象(abstraction)问题。

本书主要面向自然语言处理／计算语言学和机器学习方面的学习者。对于机器学习研究来说，语言处理提供了广泛的、具有挑战性的现实任务和数据，比如海量的特征(features)、超过百万量级的大规模数据、复杂的数据结构等。对于自然语言处理／计算语言学方面的学者，基于记忆的学习提供

了一个独特的符号化统计学习工具用于处理语言。总体来说，本书浅显易懂，即使是刚接触这方面内容的学习者，也能轻松读懂本书。

本书简单实用，介绍了一个直观又实用的方法——基于记忆的学习。跟很多现有的自然语言处理技术相比，本书介绍的“基于记忆的学习”包括两个主要的方面：一是学习方面，即简单地存贮记忆，也就是训练实例；二是遇到新实例需要分类时，就简单地找到最相似的记忆／实例，然后进行适当的修改，从而完成分类。本书另有配套的软件发布。在介绍了核心方法之后，又详细说明作者和其团队开发的基于记忆学习的软件包TIMBL，更凸显了实用性。

2　内容介绍

第一章：基于记忆学习的自然语言处理

本章对背景知识、术语进行了介绍，包括自然语言处理的目标、历史、流派。对语言处理的方式——一般是在本质上降解为分类问题——进行了介绍。最主要的，是对“基于记忆的语言处理（memory-based language processing, MBLP）”以及“基于记忆的学习（memory-based learning, MBL）”做了浅显易懂的介绍。通过一些比较直观的例子，本章对“基于记忆的学习”给出了一个初步的定义——基于记忆的学习包括两个主要方面：一是学习方面，即简单地存贮记忆和经验；二是解决新实例的分类问题，即简单地找到最相似的记忆，然后进行适当的修改。

第二章：来自语言学和人工智能的启发

本章阐述了基于记忆的语言处理技术的思想起源，包括语言学、人工智能、认知语言学。简而言之，“基于记忆的语言处理”就是使用“基于记忆的学习”技术处理语言；而“基于记忆的学习”就是把所有的训练数据都存储在记忆里——对于计算机来说，即内存等存储设备——然后当需要对新实例进行分类的时候，就搜索记忆中最相似的实例作为参考进行分类。这个主意并不是全新的，而是来源于更早的科研探索，包括语言学、认知科学、人工智能等。在人工智能的相关研究里面，有跟“基于记忆”的方法类似的思路和方法。典型代表有20世纪50年代就开始发展的“最近邻分类方法（nearest-neighbor classifier）”。Fix和Hodges在其1952年的著作中提到，“最近邻”这个思想在很多现实场合都很符合直觉，比如，医生对某个病症开处方的时候往往会参考以前对类似病症所开的处方。

但是，原始的最近邻分类方法——包括最具代表性的k-NN方法——在被提出很长一段时间内应用和影响都比较有限。这主要是因为初期的k-NN方法在存储和计算方面有若干不足——存储的时候需要把所有的实例存储在内存中，导致了内存代价较高；此外，对新的实例进行分类计算的时候，需要跟存储的所有实例进行比较，导致了计算复杂度较高。从20世纪80年代开始，由于最近邻方法简单、符合直觉的优点，其在人工智能的多个领域重新获得重视，并且发展出了多个变种，包括memory-based reasoning、case-based reasoning、exemplar-based learning、locally-weighted learning，以及instance-based learning等。这些方法在不同的方面扩展了初期的k-NN方法，以解决上面提到的k-NN缺点。本章还简要介绍了和基于记忆的语言处理技术密切相关的其他技术，包括“基于实例的机器翻译(example-based machine translation, EBMT)”以及“面向数据的句法分析(data-oriented parsing)”。

第三章：记忆和相似度

本章介绍了对基于记忆的自然语言处理技术的具体实现。一个重要的计算步骤是相似度(similarity)的计算，章节3.2对相似度计算的实现和数学表达式进行了介绍。相似度计算有不同的版本，最简单的情况是只考虑特征向量(feature vector)的相似度，但是这样计算出来的相似度往往不够准确，因为没有考虑特征的权重——也就是说，有的特征重要，有的特征不重要，应该在计算相似度的过程中区分不同特征的重要程度。可以使用信息论(information-theoretic methods)的一些方法来计算特征权重，比如信息增益(information gain)。本章对这些权重计算方法进行了具体介绍。此外，在k-NN实现上，本章对k这个超参数的选择进行了简要讨论，并强调，在k-NN的分类决策中，需要考虑到近邻实例和待分类实例之间的距离，使用这个距离来对分类投票(voting)进行加权，从而发展为距离加权的分类投票方法(distance-weighted class voting)——章节3.2.7对此方法进行了详细介绍。

基于一个现实的语言处理任务——德语名词的复数词法分析(plural formation of German nouns)，本章对基于记忆的自然语言处理技术进行了详细说明，同时还介绍了本书作者和其团队开发的基于记忆学习的软件包——TIMBL，并说明了如何在语言处理任务中使用该软件包。相信通过使用该软件包，读者可以对基于记忆的语言处理技术有一个更直观的了解。

此外，本章对TIMBL软件包的内容、运行时间、空间的复杂度等进行了客观分析。因为就是简单的存储实例，该软件“训练(training)”效率很高，可以获得O(N)的时间复杂度——N是实例个数。但是存储的效率较低，其空间

复杂度是O(N*F)——F是特征个数。不幸的是，“测试(testing)”过程的时间复杂度很高，复杂度是O(N*M)——M是测试数据实例个数——因为每个测试实例都需要和内存中存储的所有训练实例进行比较，从而确定近邻的实例。本章只是客观分析了时间、空间复杂度上的优缺点，并强调，在后面的章节中会提出高效率的估算(approximation)算法，从而可以更有成效地降低时间、空间复杂度。

第四章：在语素—语音学上的应用

基于之前的方法学介绍，本章重点介绍相关方法在具体任务上的应用，特别是怎么把一个新任务在基于记忆的语言处理框架下进行建模。本章重点讨论语素—语音学这一特定的语言处理任务。语素(morphemes)通常是由一个或者多个音素(phonemes)构成，且通常具有语义。在英语等西方语言的处理任务中，一个重要任务就是从文本／语音数据流中识别音素和语素——把机器学习应用到自然语言处理的最早的研究之一就是音素转换问题。在实际的应用场合可以碰到两种类型的词——已登录词和未登录词。比如，可以从标注好了语素／音素的训练数据中提取出一个已登录词列表及相关的语素／音素信息，不在此列表中的词为未登录词。语素—音素识别的重点就是这些未登录词，因为任何一个训练数据都是有限的，常会碰到未登录词的语素／音素识别问题。本章介绍怎样通过基于记忆的语言处理方法来解决语素／音素识别问题——包括英语的词—音素自动转换问题，以及荷兰语的语素自动识别问题(Dutch morphological analysis)。

本章详细介绍了作者团队开发的音素转换系统如何用于语音转换，并且解释了基于记忆的学习方法很适合处理词—音素自动转换问题的原因——拼写相似的英文词汇往往具有相似的发音，所以，对于一个新词，只需要在“记忆”中寻找其近邻的实例，从而就可以对其发音进行分类／预测。在荷兰语的语素自动识别方面，本章也介绍了详细的特征设置，比较了在基于记忆的语言处理框架下实现的IB1系统和IGTree系统——IGTree系统是对传统k-NN算法的一种决策树(decision tree)快速估算法——的实验效果。基于这两个任务的实验效果，本章总结了IGTree这种决策树估算法的效果：能够有效降低内存需求，并且显著提高分类效率。

此外，本章揭示了基于记忆的语言处理方法的一个显著缺点，就是无法对结构依赖(structural dependencies)进行建模——比如两个不同位置的语素分类可能具有较强的结构依赖，通过对结构依赖进行建模可以形成更合理的全局分类结果，提高分类准确率。本书将在第七章讨论基于记忆的语言处理

中结构依赖的建模问题。

第五章:在浅层句法分析上的应用

跟前一章类似,本章继续介绍相关方法在语言处理任务中的应用,通过另一个代表性的语言处理任务——浅层句法分析(shallow parsing)——来阐述如何把一个语言处理任务通过基于记忆的语言处理这个框架进行建模。跟前一章的语言处理任务相比,浅层句法分析是个更复杂的任务,因为这个任务同时需要处理词性标注(part-of-speech tagging,词性包括名词、动词等)、块切分(chunking,比如切分出名词短语、动词短语等)和关系提取(relation finding,比如确定名词短语和句子中的主动词之间的句法关系)这三个子任务。本章把浅层句法分析建模为层次化的分类问题,并使用基于记忆的语言处理技术来分别解决这些层次化分类问题。

本书作者强调,在词性标注这个子任务上,基于记忆的语言处理方法取得了比Brill [1994]提出的基于转换的学习方法(transformation-based learning)——这是当时应用比较广泛的一种方法——更好的效果。但是结果稍差于Ratnaparkhi [1996]提出的最大熵分类方法(maximum-entropy method)。最终,基于记忆的方法在标准数据集WSJ上取得了96.4%的准确率并且在LOB数据集上取得了97%的准确率。

词性标注之后,下一步就是进行块切分。块切分同样可以转化为一个分类问题,具体来说,可以通过Ramshaw & Marcus [1995]提出的BIO模式把切分问题转换为分类问题,然后使用基于记忆的学习方法进行自动分类。因为词性标注和块切分这两个任务密切相关,可以把这两个任务的类别标记合并起来变成一个综合的分类任务。实验表明,虽然合并这两个任务会潜在地导致数据稀疏,但实际上取得了更好的总体分类效果。跟前面两个任务一样,最后一个任务关系提取也可以建模为一个分类问题,并使用基于记忆的学习方法进行自动分类。

第六章:抽象和泛化

通过前面五章的介绍,读者应该对基于记忆的自然语言处理有了方法学上的理解,并在具体的语言处理任务上有了实践认识。本章进一步探讨更本质、更核心的问题——为什么基于记忆的语言处理适合语言处理任务?也就是说,基于记忆的语言处理到底好在哪里?

作者对比分析了"懒惰(lazy)"的学习方法和"急切(eager)"的学习方法,讨论了这两大类方法的特点,并总结了它们在语言处理数据中各自的优缺

点。急切学习方法的哲学根源是西方中世纪开始提出的“奥卡姆剃刀(Ockham's razor)”科学原则,其主要思想是,“一个科学理论应该去除所有不必要的元素(delete all elements in a theory that are not necessary)”——也就是说,一个科学理论应该在不影响效果的前提下越简单越好(用现代的观点解释就是最大熵),因为越简单的东西泛化能力越强。在机器学习领域,“奥卡姆剃刀”这种哲学思想在20世纪发展为一个具体的机器学习原则,就是“最小描述长度原则(minimal description length, MDL)”[Rissanen 1983],而最小描述长度原则指导了一系列的机器学习模型,包括用途广泛的决策树模型[Quinlan 1993]和规则自动生成算法RIPPER [Cohen 1995]等。

但是,作者强调,懒惰学习方法,包括基于记忆的学习,不符合最小描述长度原则,因为记忆模型的描述长度是所有使用的内存,也就是所有存储起来的训练实例。从一定程度上来说,基于记忆的学习方法甚至可以说是“最大化描述长度”的一种方法。作者强调,作为一种特定的懒惰学习方法,基于记忆的语言处理方法的主要优点是能够把所有的实例存储在记忆中,从而不需要承受抽象化(abstraction)带来的效果损失——主要是分类的准确度。本章介绍了一系列的实验,展示了实验的效果,从而佐证了作者的观点——懒惰的学习方法,特别是基于记忆的学习方法,在典型的语言处理任务上往往具有比急切的学习方法更好的分类效果。这里一个重要的原因是自然语言处理数据的特点——跟别的领域的任务相比,语言处理的训练数据(语料库)往往标注得很仔细,经过仔细检查,噪音已经很少,基本上不需要抽象化来提高泛化能力。从另一个角度来说,也许基于记忆的学习方法在语言处理数据上已经符合“最小描述长度原则”,因为错误率很低的语言处理数据已经无法“压缩信息”,也就是说,原来的“数据长度”就已经是最小描述长度。

此外,本章还详细介绍了FAMBL软件包。该软件包采用了一种特殊的抽象化方法用于改进基于记忆的学习系统。作者强调,这种特殊的抽象化方法(careful abstraction)跟传统的抽象化方法不同,在基于记忆的学习框架下能得到较好的效果,同时降低内存代价。

第七章:扩展阅读

本章重点介绍基于记忆学习技术的两种扩展:一种是基于搜索的参数优化算法,一种是在基于记忆学习的框架下解决序列标注任务(sequential tagging)上的“近视(near-sightedness)”问题。在这里我们主要介绍如何解决基于记忆学习的“近视”问题——也就是序列标注任务上的结构化预测问题。很多传统的机器学习模型都是无结构化的模型,也就是说,无法进行结构化

预测，这包括本书介绍的基于记忆的学习方法，以及支持向量机(support vector machines, SVM)等。如果要对一个自然语言的数据序列——比如字符序列、词序列——进行处理，非结构化模型需要利用一个接一个的局部窗口(window)进行局部的分类，每个局部分类得到一个标记，最后多次的局部分类加在一起得到一个完整的标签序列。问题是，每次的局部分类都是不准确的，因为都没有考虑周围标签的结构化依赖信息。为了解决该问题，更现代的机器学习方法发展为结构化的分类模型，比如条件随机场模型(conditional random fields)。这些结构化分类模型能够"同时"决定整个序列的标签，从而充分考虑标签之间的结构化依赖信息。

为了在基于记忆学习的框架下部分实现"结构化分类"，部分解决决策的"近视"问题，本章提出了两种方法。方法一是分类器的叠加(stacking)。先使用一个基于记忆学习的初级分类器进行分类，得到一个初级的标注序列。之后，使用一个基于记忆学习的高级分类器，以原有数据序列信息以及初级标注序列信息作为新的输入，得到新的决策序列。因为有初级标注序列的信息，可以部分实现结构化依赖的建模。方法二是基于"类别多元组"的分类模型。这个方法主要是把原始的标记细化为标记的多元组，从而把结构化依赖的信息记录在"类别多元组"里面。作者强调，这两种方法都能有效解决基于记忆方法的"近视"问题，获得更好的分类准确度。实验表明，在相关的自然语言处理任务上，方法一能够降低11%的错误率，方法二能够降低13%的错误率，方法一和方法二还可以结合起来使用，在不同的自然语言处理任务上降低15% 到34%的错误率。

3 本书的特色和不足

本书有两个主要的特点。第一个特点是介绍的方法简单实用。跟很多现有的语言处理技术相比，"基于记忆的学习"包括两个主要方面：一是学习方面，即简单存贮记忆；二是解决新问题，即简单搜寻最相似记忆并适当修正。本书有配套的软件发布，在介绍了核心方法之后，又详细介绍了作者团队开发的基于记忆学习软件包TIMBL，更进一步凸显了本书的实用性。

第二个特点是逻辑清楚、通俗易懂。跟书中介绍的基于记忆的学习方法的特点一样，本书写得很实在，有一说一，有二说二，章节安排层次清晰，内容深入浅出。对相关方法的描述，以及对相关原理的解释都很直观。

当然，本书也存在一些局限和缺点。首先，没有和更前沿的机器学习方法进行比较。本书的写作时间是2005年，那时很多新的机器学习模型和语

言处理方法还没有提出来，所以本书基本上只和比较老的方法、系统进行了比较，比如规则系统、决策树系统等。和现有的前沿方法相比，这些模型的精确度要低不少。书中也提到了某些前沿的机器学习模型，比如条件随机场模型，但是没有进行详细的比较。这一点比较遗憾。此外，本书虽然提到了一些基于记忆学习方法的优点，比如泛化能力方面，但是缺乏理论方面的研究和论证。现有的机器学习方法往往能够通过理论分析对泛化能力进行定量分析（后面的延伸阅读会给出一些参考文献），相比之下，本书在理论分析方面有一定的局限。

其次，本书介绍的方法存在内存开销大，相似度计算时间复杂度高的缺点。虽然后面的章节提出了一些方法来改善这两个问题，但是本质上来说问题还是存在。特别是目前的语言处理、机器学习面临的是比以前更大规模的数据，会导致内存开销大、相似度计算时间复杂度高的缺点更为明显，原来提出的一些解决方案会遇到很大的挑战。

4　延伸阅读建议

本书主要探讨了基于记忆的语言处理技术，主要内容分为两部分：基于记忆的机器学习技术和该技术在语言处理任务上的应用。延伸阅读，可以立足于机器学习和语言处理的新进展这两方面。

（1）语言处理方面的延伸阅读

在语言处理方面，可以阅读如下相关书籍：

Manning, C. D., & Schütze, H. *Foundations of Statistical Natural Language Processing*. Cambridge Massachusetts: MIT Press. 1999.（中译本：《统计自然语言处理基础》，苑春法等译，电子工业出版社，2005。）

俞士汶　计算语言学概论，商务印书馆，2003。

宗成庆　统计自然语言处理，清华大学出版社，2008（第二版2013年出版）。

此外，还可以关注语言处理领域的相关学术论文。本书的相当一部分内容是介绍语言处理任务中的参数估算，以及如何在基于记忆学习的框架下解决“近视”问题。所以，作为扩展阅读，可以参阅相关学术论文，特别是语言处理任务中的参数估算问题，以及结构化预测模型在语言处理领域的应用（如何使用结构化预测模型解决“近视”问题）：

语言处理中的参数估算问题：