Statistical Language Learning

统计语言学习

[美] 欧仁·查尼阿克（Eugene Charniak）/ 著

胡凤国　冯志伟 / 译

世界图书出版公司

Statistical Language Learning

统计语言学习

[美]欧仁·查尼阿克(Eugene Charniak) 著

胡凤国 冯志伟 译

世界图书出版公司

北京·广州·上海·西安

图书在版编目(CIP)数据

统计语言学习 /(美)欧仁·查尼阿克(Eugene Charniak)著;胡凤国,冯志伟译.—北京:世界图书出版公司北京公司,2016.6
书名原文:Statistical Language Learning
ISBN 978-7-5192-1548-4

Ⅰ.①统… Ⅱ.①欧… ②胡… ③冯… Ⅲ.①统计语言学 Ⅳ.①H087

中国版本图书馆 CIP 数据核字(2016)第 135730 号

First MIT Press paperback edition, 1996

著　　者:[美]欧仁·查尼阿克(Eugene Charniak)
译　　者:胡凤国　冯志伟
责任编辑:武传霞

出版发行:世界图书出版公司北京公司
地　　址:北京市东城区朝内大街 137 号
邮　　编:100010
电　　话:010-64038355(发行)　64015580(客服)　64033507(总编室)
网　　址:http://www.wpcbj.com.cn
销　　售:各地新华书店及外文书店
印　　刷:三河市国英印务有限公司
开　　本:711 mm × 1245 mm　1/24
印　　张:9
字　　数:188 千
版　　次:2016 年 8 月第 1 版　2016 年 8 月第 1 次印刷
版权登记:01-2015-4981
定　　价:49.00 元

译者的话

1993 年 7 月在日本神户举行的第四届机器翻译高层会议(MT Summit IV)上,英国著名学者哈钦斯(J. Hutchens)在他的特约报告中指出,自 1989 年以来,机器翻译的发展进入了一个新纪元。这个新纪元的重要标志是,在基于规则的技术中引入了语料库方法,其中包括统计方法、基于实例的方法、通过语料加工手段使语料库转化为语言知识库的方法等等。这种建立在大规模真实文本处理基础上的机器翻译,是机器翻译研究史上的一场革命,它将会把自然语言处理推向一个崭新的阶段。

哈钦斯的特约报告预示着自然语言处理将要实现战略转移:从基于规则的理性主义方法转移到基于统计的经验主义方法。20 世纪 90 年代以来,自然语言处理的研究已经实现了这样的战略转移,出现了空前繁荣的局面。这主要表现在如下三个方面:

第一,概率和数据驱动的方法几乎成为自然语言处理的标准方法。句法剖析、词性标注、参照消解、话语处理、机器翻译的算法,全都开始引入概率,并且采用从语音识别和信息检索中借过来的基于概率和数据驱动的评测方法。

第二,计算机的速度和存储量的增加,使得在自然语言处理的一些应用领域,特别是在语音合成、语音识别、文字识别、拼写检查、语法检查这些应用领域,开始进行商品化的开发。语音合成、语音识别

和文字识别的技术已经被广泛地应用于移动通信(mobile communication)中。

第三,随着网络技术的发展,互联网(Wide World Web)逐渐变成一个多语言的网络世界,互联网上的机器翻译、信息检索和信息抽取的需要变得更加紧迫,自然语言处理研究需要通过统计机器学习(Statistical Machine Learning)的方法,从互联网的大数据(big data)中自动地获取语言信息。

面对自然语言处理的战略转移,为了满足自然语言处理工作者更新知识的要求,美国布朗大学计算机科学系教授欧仁·查尼阿克(Eugene Charniak)于1996年在MIT出版社出版了他的专著《统计语言学习》(*Statistical Language Learning*),全面、系统地介绍了自然语言处理的统计方法。这样的方法是当前自然语言处理的标准方法,是自然语言处理工作者必须掌握的方法。在本书中,查尼阿克通过大量的实例讲述了统计自然语言处理的基本原理,特别是详细介绍了隐马尔可夫模型和概率上下文无关语法,内容浅显易懂,只要具备一般数学知识的读者,就不难理解本书的内容。

《统计语言学习》在国外好评如潮,成为学习统计自然语言处理的不可缺少的入门书。我们把此书翻译成中文出版,希望它的中文版有助于推动我国自然语言处理工作者的知识更新,得到广大读者的喜爱。

在翻译过程中,我们发现英文原著中有一些表述不明确的地方,还有一些地方存在明显的错误,对此我们都以"译者注"的方式在脚注中一一加以说明,并在相应的译文中予以更正,请读者注意。

我们尽了最大的力量来翻译此书,由于水平有限,如有不妥之处,恳请方家指正。

胡凤国　冯志伟

2015年12月

内容简介

基于统计方法的自然语言处理(NLP, Natural Language Processing)是一个不断发展的崭新的领域,是人工智能研究的一个分支。本书面向的读者对象是具有传统计算机科学知识背景的研究人员和科学工作者,主要介绍基于统计的语言处理技术——单词标注(word tagging)、基于概率上下文无关语法(PCFG, probabilistic context-free grammar)的剖析(parsing,又称为句法分析)、语法归纳(grammar induction)、句法排歧(syntactic disambiguation)、词义分类(semantic word classes)、词义排歧(word-sense disambiguation)等技术,同时还介绍了相关的数学知识,每一章还附有一定数量的练习题。

David M. Magerman 对这本书给予了高度评价:"这是一本有趣的关于 NLP 统计模型的普及读物。书写得很好,富有趣味性,稍微有点数学知识背景的读者都能读懂。它为读者精选了许多统计 NLP 方面的话题。书中对隐马尔可夫模型(HMM, Hidden Markov Model)的向前 - 向后算法(forward-backward algorithm)和概率上下文无关语法的内部 - 外部算法(inside-outside algorithm)进行了直观的描述,具有很强的可操作性……这是迄今为止[①]介绍该领域知识的唯一一本已经出版的专著,也是该领域内为数不多的既自成体系又浅显易懂的

① 译者注:这里的"今"指的是评价本书原著的时候,并非这本译著出版的时候。

好书之一。”

本书原著作者欧仁·查尼阿克(Eugene Charniak)是美国布朗大学计算机科学系教授兼系主任。

序言

本书的目的是使读者熟悉当前的统计语言处理技术(特别强调从语言学习的角度来看待这个问题)以及具备理解这些知识所必需的数学背景知识。写这本书的时候,我正在从一个“传统”的人工智能自然语言处理(AI-NLP, artificial-intelligence natural-language-processing)研究者开始转变,对该领域中使用的统计方法产生了兴趣,因此,这本书的写作对象是那些具有传统计算机科学背景的读者。我想这本书还是有点用处的,下面介绍一下原因。

我想几乎没有人(即使有也是很少)认为从传统的人工智能角度研究语言会是当前的研究热点。尽管还有很多人用传统方法来研究一些特定的 NLP 问题,例如语法问题和风格研究等,但是在我看来,人们越来越难以相信这种传统方法能解决更多问题,因为在过去它坚决拒绝这样做。此外,这种方法似乎有一个很薄弱的环节。AI-NLP 认为,人们说话或者书写都有某种特定的目的,其中最普通的目的是告诉别人关于世界的某些事实。不过这样的理由不要说得过了头,今后让我们就限于说到这样的程度为止,尽管我们不理会在这个领域流传的那些形形色色的谎言假话、花言巧语和科幻故事,但我们还是很难相信这种说法的正确性。然而,一个人不能一下子传达出关于世界的所有信息,因此必须假定听话者对世界有所了解,说话者/作者只不过是在此基础上以某种方式增加一些信息而已。而且,

说话者/作者对这种假设的利用往往会影响到信息传达的方式。举例来说，如果有人问“小张在干什么”，我们可能回答：“她去超市了。”即使我们没有补充诸如“买东西”之类的信息，听话者也会明白她去超市的目的是“买东西”。因为听话者会把自己所掌握的知识带入会话当中，从而能将说话者的未尽之言补充完整。因此，从 AI 的观点看，语言理解需要有大量的“真实世界知识”作为支撑，我们的程序要想最终获得成功，就必须拥有这些知识。刚好，AI 的一个分支——知识表达——一直在努力为我们提供这种“真实世界知识”，至少它要能提供某种形式化的表达，以便我们能方便地对真实世界知识进行编码。因此，我们在进行 AI-NLP 研究的时候，就不用担心自己没有掌握最基本的假设所需要的知识库。

就模型本身而言，没有任何的问题。但是，任何熟悉 AI 的人都应该知道：对知识表达的研究的进展并不是那么顺利——哪怕只是用来从报纸之类的典型文本中获取需要的“常识”知识。AI 的这个分支产生了名目繁多的逻辑，几乎每一种逻辑都有一种知识表达方法，但是没有任何一种逻辑能显示出知识表达问题的应用前景，因此该分支已经变得门可罗雀。事实上，曾经有一个人冒着失败的风险建立了一个包含各种常识知识的大型知识库，并最终获得了成功，这个人是 Doug Lenat（参见[27]），他也是迄今为止唯一获得成功的人。按照惯例，我在此要告诫研究生们不要涉足这个领域。我们当中许多研究 AI-NLP 的人已经注意到：我们的研究是以其他研究的成功为基础的，但是，成功的前景越来越不明朗。看来，的确到了转变研究方式的时候了。

对于致力于寻找另一种方式处理 NLP 问题的人来说，统计方法具有很大的吸引力。首先，它以真实文本为基础，旨在解决像语音识别之类的问题，这些问题允许不太完美的解决方案——至少目前来看是这样。因此，统计技术有望取得实用性的研究成果。其次，统计技术提供了一个十分明确的方式来获取知识：只需简单地收集一些统计数据就行了。令人遗憾的是，现在仍然有些人执迷不悟，拒绝这

种简单有效的统计技术。

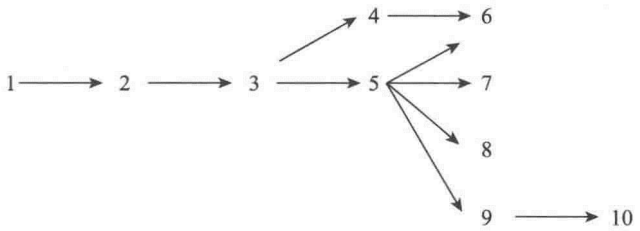
不过,转向统计技术也不是一件容易的事情。该技术涉及的许多背景知识即使在计算机专业的课程中也不是常用的,我们许多人根本不具备这些知识,即便是我本人也只是在两年之前才开始涉足这一领域。当然,我在向我的学生传授统计 NLP 知识时,他们也同样没有学习这些知识所需要的知识背景。隐马尔可夫模型就是一个很好的例子,该模型广泛应用于工程方法来处理 NLP 问题,尤其是可以应用于语音识别,但该模型在计算机科学中介绍得并不多。

如果我对传统的 AI-NLP 的批评没错的话,那么我想将会有越来越多的有着计算机背景的人从 AI-NLP 转向统计 NLP。他们需要学习这种新的技术,因此,笔者觉得有必要编写一本书来介绍统计方法,特别是那些有计算机背景的人,更是需要这样的介绍。笔者之所以把内容限定在语言知识的统计学习上,这是因为机器自动学习一直都是 AI 的核心,而统计方法的魅力就在于它能使知识学习更容易,至少使知识学习成为可能。

数学是表达统计思想的最方便的工具,本书自头至尾都安排了一定的数学知识。为了让没有广泛数学知识的人也能够读懂本书,在内容上笔者特意做了如下安排:

- 第 2 章简单回顾基本的概率论,这部分内容是本书后续章节的基础知识。
- 第 4 章和第 6 章介绍更为复杂的数学知识,我们可以推迟阅读,如果读者对算法的实现不感兴趣或者不求充分理解的话,完全可以跳过这两章内容不去阅读。
- 为了使数学推导简洁明白,笔者特别注意使推导过程中上一步和下一步之间的衔接更加详细。但这样也可能会产生相反的效果,让读者对这一大堆纯粹的符号望而生畏。

学习本书,有多种可能的学习路线供读者选择,读者大可不必拘泥于章节序号的限制。章节之间在内容上的依赖关系是这样的:



其中，第1，2，3章是基本知识，分别介绍AI-NLP、概率论与信息论以及隐马尔可夫模型，这三章是其他各章的基础。第4章介绍许多隐马尔可夫模型算法的数学原理，第5章介绍概率上下文无关语法，第6章给出更多的数学理论，第6章是依赖于第4章的，但这两章并不是本书其他章节的必需内容。后面的第7到10章分别介绍概率上下文无关语法学习、句法排歧、词义分类以及词义排歧。这几章中，除了第10章用到第9章中的技术之外，其他各章节之间的内容是相互独立的。

本书基本上自成体系，不需要其他任何书作为预备知识，适合具有中级和高级水平的大学生使用。不过，由于本书内容单一，紧贴该领域前沿研究，更适合作为研究生水平的入门课程。笔者就使用本书的草稿作为教案同时给大学生和研究生授课。

最后，衷心感谢我的学生Curtis Hendrickson，Neil Jacobson和Mike Perkowitz，他们阅读了本书的草案并提出了改进意见。特别要感谢的是Felix Yen，他的意见和建议尤为重要。

目录

图目录