COMPUTATIONAL LINGUISTICS

计算语言学

（修订版）

刘　颖　编著

清华大学出版社

COMPUTATIONAL LINGUISTICS

计算语言学

（修订版）

刘　颖　编著

清华大学出版社

北　京

内 容 简 介

计算语言学是一门涉及语言学、计算机科学和数学等多门学科的交叉学科，覆盖面广。本书侧重最经典的工作，阐述计算语言学的基本理论和方法，主要介绍现代句法理论和语义理论，词法、句法和语义阶段重要的分析算法、统计语言学和机器翻译。本书结构完整，层次分明，条理清楚；既便于教学，又便于自学。可作为中文、外语、计算机等专业高年级本科生和研究生教材，也可供从事自然语言处理或信息处理的研究者参考。

图书在版编目(CIP)数据

计算语言学/刘颖编著. -- 修订本. --北京：清华大学出版社，2014
ISBN 978-7-302-37814-3

Ⅰ.①计…　Ⅱ.①刘…　Ⅲ.①计算语言学　Ⅳ.①H087

中国版本图书馆 CIP 数据核字(2014)第 198088 号

责任编辑：马庆洲
封面设计：曲晓华
责任校对：刘玉霞
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：19　　字　　数：458 千字
版　　次：2014 年 9 月第 1 版　　印　　次：2014 年 9 月第1 次印刷
印　　数：1～1500
定　　价：59.00 元

产品编号：056994-01

前　言

计算语言学，也称自然语言处理或自然语言理解，它是研究如何利用计算机来分析、处理和理解自然语言的一门学科。计算语言学是植根于计算机科学、语言学和数学等多学科沃土而成长起来的一门新兴学科。一般情况下，处理自然语言不仅要有语言学方面的知识，而且还要有数学和计算机科学方面的知识，因此，计算语言学就成为一门介于语言学、数学和计算机科学之间的边缘性交叉学科。

本书第 1 章主要介绍计算语言学与计算机科学、数学和语言学学科之间的关系，并介绍了计算语言学的基本方法、主要内容、应用领域以及发展过程。第 2 章主要介绍了自然语言处理常用的语法词典、语义词典和语义框架词典及其应用。语法词典介绍了《现代汉语语法信息词典》，语义词典介绍了 *wordnet* 、《同义词词林》和《知网》，除此之外还介绍了语义框架词典 *framenet*。第 3 章主要介绍汉语的切词、切词歧义以及如何消歧，介绍了英语的形态分析及主要分析算法以及日语的分词、分词歧义、分词算法和日语开源分词软件。第 4 章主要介绍词性标注的 4 种方法——规则方法、统计方法、规则与统计结合以及基于转换的错误驱动方法。重点介绍用隐马尔可夫模型、条件随机场和最大熵模型进行词性标注的统计处理过程。第 5 章主要介绍乔姆斯基的 4 种类型的文法和 4 种类型的自动机。文法和自动机是刻画语言的有效手段，文法用来生成语言中的句子，自动机用来识别语言的句子，就描述一种语言而言，两者是统一的。前者属于形式语法理论，后者属于自动机理论。第 6 章主要介绍 20 世纪 50 年代以来发展起来的用于自然语言处理的一些重要的句法理论，主要有基于类的语法理论和基于词的语法理论，基于类的语法理论有：转换生成语法、树粘接语法、词汇功能语法、功能合一语法、广义短语结构语法和中心词驱动的短语结构语法。基于词的语法理论包括：范畴语法、依存语法和链语法等。乔姆斯基提出的短语结构语法分析能力不高，分析时难以区分大量的不合语法的句子，生成能力过强。后来乔姆斯基提出了转换生成语法来克服短语结构语法的这些局限性，但转换生成语法本身也有局限性，它的生成能力过强。于是，乔姆斯基提出管辖约束理论来限制转换生成语法过强的生成能力。然而，由于转换生成语法通常要涉及若干个句子之间的关系，在机器翻译和自然语言处理中使用起来很不方便，不如短语结构语法那样，就一个句子来分析一个句子，它的成分结构是单一的，非常便于进行机器翻译的语法分析和自然语言处理。计算语言学的学者们抛弃了转换生成语法，又转向短语结构语法，于是 80 年代以来出现了各种增强的短语结构语法——词汇功能语法、功能合一语法、广义短语结构语法、中心词驱动的短语结构语法等，这些语法都采用了复杂特征结构来改进短语结构语法单一的特征，采用合一运算来改进传统的集合运算，从而有效地克服了短语结构语法的缺点，保持了短语结构语法的优点。基于词的语法与基于类的语法不同，把语言知识主要都

记录在词典中。第7章主要介绍了用于自然语言分析的扩充转移网络、Early分析算法、Tomita分析算法、Chart分析算法和CYK分析算法。基于扩充转移网络的句法分析的优点在于所定义的操作接近人在理解语言时所采用的操作,缺点是随着结点的增多,计算的复杂性就会急剧地增长,修改时非常困难。Tomita分析算法、Chart分析算法等都可以运用复杂特征集和合一运算机制对短语结构语法进行分析。Tomita分析算法改进了LR分析算法,是一种高效的自然语言分析方法。Chart分析算法采用了线图(chart)来记录分析结果,线图可以表示互不相连的树,可以表示歧义。CYK算法是一种并行的分析算法。由于其算法容易实现、易于被人理解,因此被广泛应用于机器翻译中。第8章主要介绍了用于自然语言处理的一些语义理论以及如何运用这些理论。第9章主要介绍了语料库及其标注、n元模型、HMM模型及HMM在语音识别和组块识别中的应用、随机上下文无关语法及在句法语义消歧中的应用、基于长度的句子对齐、支持向量机及用于入声识别和最大熵模型及其应用。针对统计出现的数据稀疏问题,介绍了常见的数据稀疏处理方法。第10章系统地介绍了机器翻译的方法、困难及解决策略、应用类型及应用领域和机器翻译的自动评价。重点介绍了统计机器翻译方法,包括基于词对齐的机器翻译、基于短语对齐的机器翻译和基于句法的机器翻译。词对齐主要有IBM的词对齐和基于HMM的词对齐。短语对齐主要有利用词对齐进行的短语对齐、对齐模板和层次的对齐短语。基于句法的统计机器翻译介绍了树串模型、串树模型和树树模型。目前,基于短语对齐的统计翻译译文质量较高。

本书可作为中文、外语、计算机等专业高年级的本科教材,教授学时可为32～64学时。教师可根据学时安排上机,比如词法分析、词性标注和句法分析等。如果学生掌握了基本理论和算法,同时通过上机实现了一些重要算法,则能为掌握本门课程以及掌握计算机处理自然语言打下坚实基础。

本书在写作时尽量做到通俗易懂,所有的算法都举例进行了详细说明,并列出了计算机处理自然语言的详细过程。本书的读者如果具有一定的计算机科学方面的知识(如离散数学、数据结构等),则能更好地理解本书的所有内容。

本书的写作参考了许多学者的论文和著作,本书能够出版与他们所做的工作紧密相关。谨向他们表示衷心感谢。

由于本人水平和时间限制,书中难免存在疏漏和不足之处。欢迎各位读者批评指正。

刘 颖

2014年7月

目　录

第 1 章　计算语言学简介

1.1　计算语言学

1.1.1　计算语言学概念

计算语言学，也称自然语言处理或自然语言理解，是一门以计算为手段对自然语言进行研究和处理的学科。例如，用计算机对自然语言的音、词汇、句法、语义和语用等信息进行处理。

自然语言处理这个术语主要用于说明方法，计算语言学这个术语主要用于说明理论。

计算机对自然语言的研究和处理，一般应经过如下 5 个过程。

(1) 提出计算机要处理的语言学问题。

(2) 根据语言学理论把需要研究的语言学问题形式化，使之能严谨规范并采用一定的数学描述方式描述出来(对应于自然语言处理的规则方法)。或把需要研究的语言学问题抽象成数学模型(对应于自然语言处理的统计方法)。

(3) 设计计算机算法，使得计算机能自动地处理形式化的语言学问题，或者按着数学模型进行相应的统计和处理。

(4) 根据算法编写计算机程序，运行计算机程序来实现计算机的自然语言处理。

(5) 对计算机处理的结果进行实验分析，得出结论。

因此，为了处理自然语言，不仅要有语言学方面的知识，还要有数学和计算机科学方面的知识，这样计算语言学就成为了一门介于语言学、数学和计算机科学之间的交叉学科(冯志伟 1996)。

1.1.2　计算语言学与计算机科学

计算语言学一方面要求把计算机科学处理问题的一些基本思想、基本方法引到语言学研究中来，从新的角度观察语言学，建立和传统语言学不同的语言学理论，这些语言学理论要精确地描述和解释语言的结构、现象和规律，建立语言的严谨的可计算的形式化模型和可统计的概率模型。另一方面，计算机科学提供相应的算法，在这些模型的基础上，进行检索、统计、计算、推导、分析、转换、生成等，从实现角度对模型进行检验。因此，计算语言学家一方面需要研究现有的所有语言学理论，另一方面需要研究现有的数学统计模

型、计算机算法和高级程序设计语言。了解哪些问题可以用现有语言学理论解决,哪些是不可以解决的。哪些可以用抽象的统计模型来解决,哪些不能。还必须了解处理语言学问题适合的算法和编程语言(侯敏 1999;姚亚平 1999)。

1.1.3 计算语言学与语言学

语言学是研究语言现象及其规律的科学。计算语言学是语言学的一个分支,是运用计算机的手段研究语言现象和规律并对其进行自动处理。传统语言学和计算语言学的区别主要在于以下方面。

(1) 传统语言学是一门经验学科,而计算语言学既是一门理论学科,又是一门实验科学(侯敏 1999)。

(2) 计算语言学要面对整个自然语言现象,因此,它必须研究语言的带有普遍性和总体性的一般问题;而传统语言学家喜欢深入研究某一特殊的语言现象,更加重视研究语言中的某个特殊问题(冯志伟 2001)。

(3) 传统语言学主要是描述性的,而计算语言学要求的语言学理论必须具有可操作性。要想操作,一种方法是要把一个句子中所有的信息,包括词法的、句法的、语义的都形式化,变成机器可以识别的规则,这样它才能一步步操作,最后达到理解这个句子的目的。另一种方法是根据大规模语料库中语言单位出现的概率来计算所要处理问题的概率(冯志伟 1996)。

(4) 计算语言学的理论必须要通过计算机实践来检验,从实验结果中检验计算语言学的理论是否可行。而传统语言学则要求讲道理,重视逻辑的完美性(冯志伟 2001)。

(5) 计算语言学研究语言时必须先分析和处理后理解,理解是分析和处理的结果。而传统语言学是先理解后分析,理解是分析的必要前提(冯志伟 2001)。

1.1.4 计算语言学与数理语言学

计算语言学就相当于应用数理语言学,是数理语言学的一个分支。数理语言学是运用数学思想和数学方法来研究语言现象的一门新兴的语言学科。数理语言学的出现,使得作为一门人文科学的语言学与数学、计算机科学以及人工智能等发生了密切的联系,使得语言的研究逐渐走上了现代化的道路。句法分析、语义消歧、机器翻译、自动问答等语言自动处理技术的出现,要求精确地描述和解释语言的结构,建立语言的数学模型,并用数学方法来研究语言的语法和语义结构(冯志伟 1985)。

数理语言学主要研究:(1)代数语言学;(2)统计语言学;(3)应用数理语言学。

代数语言学:采用集合论、数理逻辑、图论、形式文法、自动机等离散的、代数的方法来研究语言。

统计语言学:采用概率论、数理统计和信息论等统计数学的方法来研究语言成分使用的统计规律。

应用数理语言学:把代数语言学和统计语言学应用于机器翻译、人机对话以及信息检索的技巧与方法,就是应用数理语言学的研究内容。

代数语言学是基于规则的,它代表着数理语言学中的理性主义方法,统计语言学是基

于统计的，它代表着数理语言学中的经验主义研究方法；而在数理语言学的实际应用中，既有理性主义方法，也有经验主义研究方法，还有把二者结合起来的研究方法。

1.1.5　计算语言学与自然语言

计算语言学研究和处理的对象是自然语言，而不是人工语言或其他的形式语言。

世界上的语言，绝大多数是自然语言。自然语言是人类发展过程当中自然产生、约定俗成的用于人类社会交际的语言，如英语、汉语、日语等。自然语言中有少数是通过人为的力量创造或规定下来的语言，比如世界语。

形式语言是人们有意识地通过形式化的定义所规定的语言，典型的形式语言包括程序设计语言（比如 C 语言、Java 语言、Perl 语言）和符号逻辑语言（比如一阶逻辑语言）。形式语言是具有严格结构的符号系统，适合于计算机使用和处理。

在计算机软件中，早已设计了许多人工语言，如 Basic，Pascal，Cobol，Lisp，C，Java 等程序设计语言，这些人工语言都遵循着形式语言的规律和法则。对这些人工语言的词法、句法、语义的分析和生成，技术已比较成熟，发展成为一门新的学科“编译原理”，但自然语言比人工语言要复杂得多，因而用计算机处理起来也就困难得多。

自然语言与人工语言的区别，主要表现在下面 4 个方面（冯志伟 2001）。

（1）自然语言在语音、词汇、句法、语义和语用层面都存在歧义。而人工语言中的歧义则可以由人来进行限定。

（2）自然语言的结构复杂多样，而人工语言的结构则相对简单。

（3）自然语言的语义表达千变万化，迄今还没有一种简单而通用的途径来描述它，而人工语言的语义则可以由人来直接定义。

（4）自然语言的结构和语义之间有着错综复杂的联系，一般不存在一一对应的同构关系；而人工语言则常常可以把结构和语义分别进行处理，人工语言的结构和语义之间有着整齐的一一对应的同构关系。

由于自然语言的这些独特性质，使得自然语言处理成为人工智能的一大难题。

1.2　计算语言学主要研究的内容

按照语言学上一般的分析，语言可分为如下的一些层次：语音、词汇、语法、语义、语篇和语用。计算机在语言学上各个层次的应用便形成了计算语音学、计算词汇学、计算语法学、计算语义学、计算语用学等，它们都是计算语言学的分支学科（冯志伟 1999）。

计算语音学：研究如何利用计算机对语音信息进行处理，实现语言的自动合成与识别。

计算词汇学：研究如何用计算机处理自然语言的词汇、建立语言词汇库、术语数据库等机器可读词典。对于印欧语言主要研究形态分析。计算机形态分析是研究如何将一个词分析为词素的组合，从而导出该词的组成结构和意义。例如，将词 friendly 分析为名词 friend 和后缀 ly 的组合，计算机可以得知 friendly 是由 friend 导出的形容词。一个自动词法分析方案可包括一部词干词典和一套描述词形变化和构词的规则系统，这样，在分析

时,给出词干,计算机就可以自动地列举出它的所有的变化形态,而给出一个变化形式,计算机就可以自动地把它切分为词干、词缀和词尾。对于汉语,主要研究汉语的自动分词。因为汉语中单词与单词之间没有空格,必须首先进行分词(罗振声,袁毓林 1996)。

计算语法学:研究如何用计算机来分析自然语言的句法。根据语法学所提供的关于语法结构的规则,推导出一个语句的所有可能的语法结构。这种研究在计算机中叫做"parsing",目前,parsing 技术比较成熟,有 Earley 分析算法、Tomita 分析算法、Chart 分析算法和 CYK 算法等。用于计算机自动处理的语言学理论有广义短语结构语法,词汇功能语法,功能合一语法,基于中心词驱动的短语结构语法、依存语法、链语法等。

计算语义学:研究如何利用计算机来分析自然语言的语义。目前,语义分析集中于利用大规模语料库中的上下文来确定一个词在所在句子中的确定含义、对同义词进行辨析或利用词典和上下文来确定词与词之间的语义关系等。计算机处理的语义学理论有 R. Wilks的优选语义学、Fillmore 的格语法、R. C. Shank 的概念依存理论、R. F. Simmons 的语义网络理论和 R. Montague 的蒙塔格语法等。

计算机语言学习:以上每个问题,都需要应用大量的语言知识。解决某一问题需要哪些知识,如果都需要由人工决定,并形式化地表达这些知识,则需要大量的人工及专家知识。计算机语言学习的目的就是通过机器学习,自动地获得语言处理所需要的专门知识,并将这些知识形式化地表达出来。

语料库语言学是利用计算机强大的检索、统计和处理语料的能力,从大规模的语料库中检索符合研究问题的实例,对其进行统计。在大量实例和统计数据的基础上,对研究问题进行定性分析,从功能上对其进行语言学解释。利用计算机和语料库可以对语言各个层面的特征(单个特征或多个特征)进行分析和研究。

语料库语言学的基本任务是研究机器可读的自然语言文本的采集、存储、检索、统计等,以及语料库方法在词汇、语法、语义、语篇结构、语域变异、语言习得、作家作品风格分析等领域中的应用。

语料库语言学的优势在于:(1)可以利用计算机的强大功能,进行快速、准确的分析;(2)语料库规模大,所包括的语域全面,文本量大,语言信息范围广;(3)既有定量分析,又有定性的功能解释,对语言的描写全面;(4)语料库方法与以往的方法相比能做出更概括和更全面的调查。因此,基于语料库的方法可以扩大以往调查的范围和调查语言的新应用。语料库语言学已经成为语言研究的主流,它正对许多语言研究领域产生越来越大的影响。

1.3 计算语言学理论的主要用途

1.3.1 机器翻译

机器翻译:利用计算机将一种自然语言自动翻译成另外一种自然语言。中英翻译就是利用计算机自动地把汉语翻译成英语。很多大学、科研院所和公司展开了对两种语言或多种语言之间的两两互译。Google、Microsoft 和百度公司都开发了在线多语言机器翻译系统。MSN、社交网络 Facebook 和通信工具 GoogleTalk 都提供了即时翻译任务。在欧美和日本已经开发出 SYSTRAN、TAUM-METEO、METAL 等多个实用的机器翻译系统。

比较有名的在线机器翻译系统是 Systran：http://www.systransoft.com，ReadWorld：http://www.readworld.com。

1.3.2　语音自动识别和自动生成

语音自动识别：用计算机将人的语音自动转换成文本。语音识别是与声学、语音学、语言学、计算机科学、数字信号处理理论、信息论等学科紧密相关的一个多学科交叉的领域。20 世纪 90 年代后，语音识别的研究向实用化的方向发展。IBM 公司推出的 ViaVoice 可以针对大词汇量和非特定人的连续语音进行识别。微软开发的 Speech Application SDK、SUN 公司开发的 JavaSpeechAPI 和 IBM 的 Dutty++等都能识别多个不同国家的语言，比如英语、日语和中文等，语音识别技术比较成熟。语音识别在旅游、铁路、宾馆预订、民用航空可用来建立人机对话的无人管理问讯处。在侦查部门用来作“声纹”刑事侦破系统，还可以用于口语翻译。

语音自动合成：就是用计算机技术或数字信号处理技术把文本转换成人类的语音。语音合成与语音识别经常结合形成人机对话系统。IBM 公司开发的智能词典 2000，能对单词、短语、句子以及段落等准确发音。AT&T 公司开发的真人语音合成系统，它发出的语音让人无法辨出真假。微软公司开发的 SAPI SDK 语音应用工具包，支持多种语言的识别和朗读，包括：汉语、日语和英语。

1.3.3　自动文摘

自动文摘：用计算机将反映原电子文档核心内容的文本自动地抽取出来，生成一篇语意连贯的文本。自动文摘应具有概括性、客观性、可理解性和可读性（俞士汶 1996）。它应以提供文献内容梗概为目的，简明、确切地记述文献重要内容的短文。目前，网上文本信息大量涌现，人们越来越关心如何能快捷、准确、全面地获取这些信息，而浏览全文的摘要是一条有效途径。比较著名的自动文摘系统有 MITRE 公司开发的 Mani 和 Bloedorn 系统、南加州大学开发的 Marcu 系统、卡内基-梅隆大学开发的 Goldstein 和 Carbonell 系统、哥伦比亚大学开发的 Mckeown 系统等。目前，关于自动文摘最有影响的会议是文本分析会议（Text Analysis Conference，简称 TAC），包含文档理解会议（Document Understanding Conference）和文本检索会议（Text Retrieval Conference，简称 TREC）。

1.3.4　自动校对

自动校对：目前出版业（尤其是电子出版）发展非常迅速，其中校对环节的工作量也大大增加。如果校对的方式还停留在人工校对上，则与出版业其他环节的逐步自动化不相匹配。如果能由计算机来完成其全部或部分工作，则会减轻繁重的校对工作，减少大量的劳力，因而提出了自动校对。文字处理软件 Word 和 Wordpefect 都嵌入了英文拼写检查功能。ExpertEase 公司推出的 DealProof，Newton 公司推出的 Proofread 是互联网上见到的英文单词拼写检查系统。黑马校对系统、金山校对系统等是已经商品化的中文文本校对系统。

1.3.5 自然语言理解

自然语言处理包括自然语言理解和自然语言生成。自然语言理解:又叫人机对话(Man-Machine Dialogue),研究如何让计算机理解和运用人类的自然语言,使得计算机懂得自然语言的含义,并对人给计算机提出的问题,通过对话的方式,用自然语言进行回答。自然语言理解系统可以用作专家系统、知识工程、情报检索、办公室自动化的自然语言人机接口,有很大的实用价值。自然语言理解是人工智能研究中的热点和难点之一。日本研制第5代计算机的主要目标之一,就是要使计算机具有理解和运用自然语言的功能。

1.3.6 信息自动检索

信息自动检索:又称信息检索,是从大规模非结构化数据(通常是文本)的集合(一般保存在计算机上)中找出满足用户信息需求的资料的过程(王斌 2010)。随着 Internet 的迅速发展,网络上的信息越来越多,面对浩瀚的信息许多用户手足无措,无法准确地获取自己所需要的信息。针对这种情况有些组织和个人开发出用以查找网络信息的检索工具——搜索引擎。目前世界上最大的搜索引擎是 Google、MSN 和雅虎,MSN 主要是美国商业目录搜索引擎,主要为用户提供教育、新闻、媒体及娱乐信息。中文综合性搜索引擎有:百度、Google、中国搜索联盟、新浪、搜狐、网易、雅虎等,其中百度是目前最具影响力的中文搜索引擎。

广泛用于科学研究和论文检索的著名三大文献检索工具是 SCI、SSCI 和 EI 检索,它们已成为评价科研工作人员、科研机构乃至一个国家的学术研究水平的重要指标,是科学研究领域研究人员查阅学术资源的重要工具。

科学引文索引(Science Citation Index,SCI),创刊于 1963 年,是美国科学信息研究所(http://www.isinet.com)出版的世界性学术期刊文献检索工具。SCI 收录世界上的重要期刊约 3500 种,而网络 SCI 扩展版(SCI Expanded)收录 5900 种。内容涵盖数、理、化、农、林、医、生命科学、天文、地理等自然科学各学科领域。

社会科学引文索引(Social Science Citation Index,简称 SSCI)是美国科学信息研究所建立的综合性社科文献数据库。SSCI 收录全球 50 多个语种的 1700 多种重要社会科学期刊论文,内容涉及社会科学、人类学、考古学、商业、财政、经济、教育、地理、历史、法律、语言、政治等 50 多个学科领域。SSCI 是评价人文及社会科学领域学者学术水平的权威的参考指标之一。

工程索引(Engineering Index,简称 EI)创刊于 1884 年,是美国工程信息公司出版的工程技术类综合性检索工具。EI 收录生物工程、交通运输、化学和工艺工程、照明和光学技术、农业工程和食品技术、计算机和数据处理、应用物理、电子和通信、控制工程等各学科领域 5100 种工程类期刊、会议论文集和技术报告。EI 具有综合性强、资料来源广、地理覆盖面广、数据资源丰富、信息质量高、权威性强等特点。文献是否被 EI 收录是衡量工程技术研究领域学者学术水平的重要指标之一。

国内比较重要的检索是中国科学引文数据库 CSCD(Chinese Science Citation Database)、南京大学中国社会科学研究评价中心研制的中文社会科学引文索引 CSSCI(Chinese Social