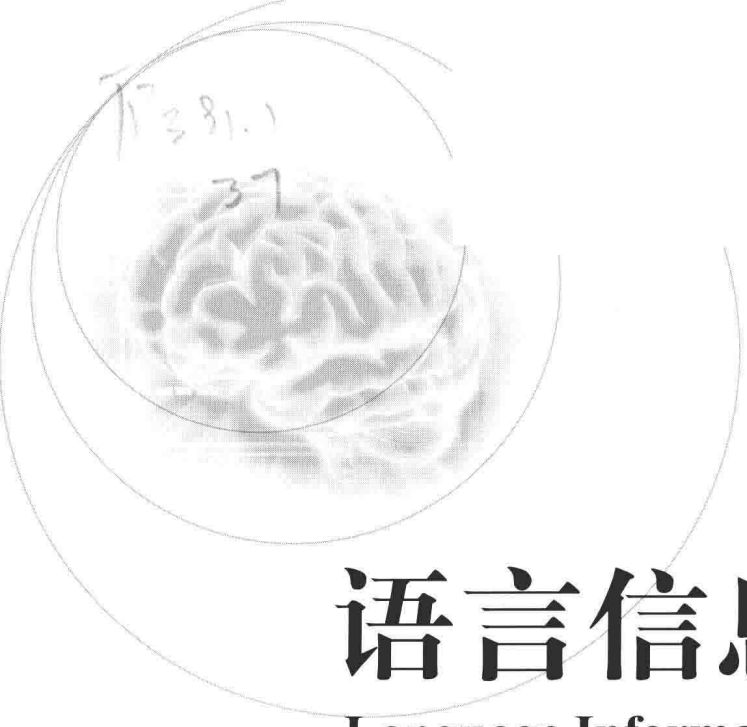
语言信息处理

Language Information Processing

江铭虎 著

人民出版社



语言信息处理

Language Information Processing

江铭虎 著

人民出版社

责任编辑:夏　青

图书在版编目(CIP)数据

语言信息处理/江铭虎 著. —北京:人民出版社,2016.9
ISBN 978-7-01-016434-2

Ⅰ.①语… Ⅱ.①江… Ⅲ.①语言信息处理 Ⅳ.①TP391.1

中国版本图书馆 CIP 数据核字(2016)第 155300 号

语言信息处理

YUYAN XINXI CHULI

江铭虎　著

人民出版社 出版发行
(100706　北京市东城区隆福寺街 99 号)

环球东方(北京)印务有限公司印刷　新华书店经销

2016 年 9 月第 1 版　2016 年 9 月北京第 1 次印刷
开本:710 毫米×1000 毫米 1/16　印张:18.75
字数:300 千字

ISBN 978-7-01-016434-2　定价:48.00 元

邮购地址 100706　北京市东城区隆福寺街 99 号
人民东方图书销售中心　电话 (010)65250042　65289539

序

Nowak 指出语言是人类进化的遗产，是过去 5 亿年中出现的最有趣的事物。生物学使用生成系统，基因组由 4 个核苷酸序列组成，按一定的规则生产蛋白质和组织细胞，产生无限多样性的生命有机体。几十亿年来，地球上的生命进化受到限制，只能使用这种生成系统。直到 最近，另一个生成系统的出现展示了进化的新方式，这个系统便是人类语言，使我们实现个体间无限的非基因信息的传递，并且引起文化的进化①。著名语言学家王士元教授指出地球上千千万万种的动物中只有人类有语言，语言是人类沟通交流的一种方式，人类智慧的结晶可用语言描述和记载，使我们的知识、经验和财富得已积累，使我们可能用几年或十几年的时间就可掌握人类几千年文明智慧所积累的知识，使人类走向社会的文明，促进科技的发展，并如此深刻地影响整个地球的现状及未来②。著名学者蔡曙山教授指出语言认知是联接低阶认知（神经、心理层级，人类和动物都有）与高阶认知（语言、思维、文化层级，仅人类拥有）的桥梁，低阶认知是高阶认知的基础。人类的心智与认知产生于脑的进化与分工，奠基于语言，发展于逻辑与思维，积淀为文化，构建为社会。

语言是人类区别于其他动物的最重要特征，在人类进化过程中语言的使用使人脑重量从远古的 300 克增加到现在的 1200—1400 克③，并且成为脑容

---

① Nowak M. A., Komarova N. L., and Niyogi P. Computational and evolutionary aspects of language. *Nature*, 417: 611-617.

② 王士元：《演化语言学论集》，商务印书馆 2013 年版

③ Schoenemann, P., Thomas, B. T. F., Sarich V. M., Wang W. S.-Y. (2000) Brain size does not predict general cognitive ability within families. *Proceedings of the National Academy of Science*, USA 97: 4932-4937.

积增长的决定性因素。

任何语言，包括书面语、口语或手语，都是由小的元素（语素或音素）分层递归地组合建构成较大的单元，依次组成音节、词汇、短语和句子，再由此组成段落和篇章①。这种递归组合由语法规则的层次结构决定，人类与其他动物的大脑区别是人类具有运用复杂层次结构模式语言的能力和处理递归结构的能力，同其他物种相比，人类可产生并理解复杂长句。与动物不同的是，人类可将外界事物分析成块形成概念，将不同概念组合运用逻辑造出新的句子。相反，动物使用完整的表达方式，并不会将外界事物分离并重新组合来形成新的信息。

语言是变化无穷的，但语言的语法类型是有限的，是可以归类、分析、统计和学习的②。从自然语言理解的角度看，句法学是研究句中各单词间的关系，语义学是研究词以及所指内容间的关系，而语用学是研究交互双方所涉及的上下文环境和背景知识，语言学知识包括世界知识、历史知识、常识性知识、各学科门类的专业知识等。人类的知识是通过记载的文字、音像及自身的经历，经学习、训练、归纳、总结，从具体到抽象、从实践到理论逐步积累、逐步完善而形成的。人类使用语言的表现形式有书面文字、口语和手语等，口语是人类最早的语言交流形式，其特点是口语语句简单、短小。口语使用的词汇大多数是常用词汇，词汇丰富程度低于书面语言。和口语相比，人类的文字只有几千年的记载历史，最早的汉字是产生于3500年前的甲骨文，是一种象形表意体系的文字，不同于印欧语言的表音体系文字，部分汉字的意义根据偏旁部首能在大脑中反应出来。因书面语的语料相对容易获得，对书面语的研究相对更多，也更深入。聋人使用手语进行交流，并获取知识、参与社会活动。手语有自己的语法规则、词汇结构，其视觉特性是任何有声语言中没有的语言现象。人类可以高效的使用语言，而计算机处理语言在性能和效率上均逊色于人脑，特别是计算机使用人类知识理解语言有诸多困难，导致计算机无法像人脑一样灵

① Elissa L., et al.(2004) Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, 49: 85-117.

② Nowak, M. A., Komarova N. L. & Niyogi, P. Computational and evolutionary aspects of language. *Nature*, 417: 611-617,2002.

活运用语用背景知识来解决自然语言中的各种歧义。当前随着科学技术、仪器设备和互联网技术的飞速发展,语言学研究在交叉学科的三个研究热点上有可能取得理论与方法上的进展:

1.基于 ERP 脑电时序信号和 fMRI 人脑成像定位的语言规律认知研究

脑神经科学的研究表明,人类语言处理的神经脉络,以储存有关事实与事件信息和学习为基础的陈述性记忆系统,以及以学习和处理运动、知觉和认知技巧的程序性记忆系统两条神经回路的作用。根据 Peter Hagoort 的神经模型,语言加工分为记忆(单词的存储和提取)、整合(如汉语的形、音、意和句法结构信息整合到整个语言的全部表征中)和控制(将语言与行动联系起来)三个环节完成,将认知心理语言学的各种发现与人脑各语言功能区的神经回路联系起来。随着脑神经科学技术的飞速发展,我们处理语言的手段和工具日益增多,如我们可以利用事件相关单位 ERP 的脑电信号和功能磁共振成像 fMRI 对人脑进行语言理解时进行定时、定量和定位的观察研究,探讨大脑是如何通过处理语言输入来获取意义,即探讨词汇在大脑中如何表征,从认知神经科学探寻其原理,通过实验建模和神经生理学的数据,探讨大脑如何理解语言。将心理语言模型和脑神经科学结合起来共同阐明语言的神经机制。从交叉学科的角度开展基于句法、语义和语用认知机理的研究,借鉴脑认知过程和脑成像技术探讨汉语在句型、句义和句法的关系,探讨汉语的形、音、义的交互作用,汉语的隐喻认知、汉语与汉语手语的认知差异,通过确定汉语 ERP 脑电信号的电生理相关性进程和 fMRI 成像的功能区域,寻找人脑汉语产生与理解的机理。并根据以上认知机理,探索汉语在人脑中的表达和运行方式,从脑认知的角度探讨将语用背景信息应用于汉语加工处理技术,探索自然语言在人脑中表达和加工方式,研究汉语多层次结构和语义的认知机制,语境和背景信息对结构和语义加工的影响。

2.基于知识图谱的深度学习研究

近几年来,随着 Web 和 Semantic Web① 的发展,为我们提供了维基百科、

① Mellish, C., Sun, X. T. (2006) The semantic web as a Linguistic resource: Opportunities for natural language generation. *Knowledge-Based Systems*, 19(5): 298-303.

Freebase 和百度百科等富含大量知识的信息源，其特点是具有半结构化、知识覆盖率广、可信度高和质量可靠，是构建大规模中文知识图谱的基础。另外，随着网络搜索引擎的飞速发展及电子阅读、情报检索、语义搜索、机器问答等广泛的应用极大地推动了中文知识图谱的构建。Google 搜索于 2012 年 5 月份发布了“知识图谱”（Knowledge Graph），是语义网络的扩展，通过对 Web 知识的挖掘与获取，从 Freebase①、维基百科（Wikipedia）②、《中情局世界概况——The CIA World Factbook》等结构化或半结构化网页中抽取信息，然后将搜索结果进行知识系统化，用网络的结点代表实体/概念，网络的边代表实体/概念之间的各种语义关系，目前已建立了超过 5 亿个事物和 35 亿条不同事物之间的关系。百度公司及中国科学院软件研究所孙乐研究员及团队正在建立类人的世界知识的中文知识图谱，为计算机理解自然语言提供了知识基础。2006 年，Hinton G.等人③提出采用深层神经网络的深度学习来减少数据表征的维数，深层网络的结构是一种分层结构，与人脑的结构接近，其研究结果表明深层网络具有优异的特征学习能力，通过训练学习得到的特征数据对所表征的事物具有更本质的刻画，通过无监督的训练学习实现逐层初始化来克服深度学习的困难。这种深层网络直接把海量数据输入到深度学习算法中，可以用较少的参数表示复杂的函数，通过无监督的训练学习，组合低层特征形成更加抽象的高层表征，系统会自动地从海量数据中学习并抽象出语言信息的本质，从而发现数据的分布式特征表示④。深度学习模拟人脑认知世界的过程，是目前脑认知与人工智能近年的研究热点。很多国际著名的认知语言学家以此作为语言理解的模型，在语言处理过程中，深度学习通过分布表示，内置建模了对象之间的相似度，从大量未标数据得到的分布相似性被证实会显著提高模型性能。在生物学上模仿人脑认知世界的过程，其最大特点是自动学习输入的多层表示，每层的特征表征某种隐含的概念，低层表征经过逐

① http://www.freebase.com/.

② http://www.wikipedia.org/.

③ Hinton, G. & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504-507.

④ http://deeplearning.net/.

层组合得到高层的更加抽象的表示，是一个经过多层抽象的过程。深度学习在语言信息处理领域正得到重视，在很多任务上取得了历史上的最好效果。对近代文学小说计算风格分析，以定量的方式利用文本中可以量化的语言结构特征来对文本风格和作者写作习惯进行研究，其理论基础是文本的语言结构特征表现了作者个人在写作活动中的言语特征，是作者个人风格不自觉的深刻反映，这些特征（如字符、词汇、句子、段落、语法、语义等）又可以在一定程度上通过数量特征来进行刻画和描述，目前已对夏目漱石的 90 部小说，鲁迅的 17 部小说和宫泽贤治的 243 部小说采用数据统计、文本聚类、主成分分析、文本分类的统计学方法，结合数据和语言本体的知识来说明三位作者的语言风格，并进行差异比较，将所开设的《自然语言处理》和《脑认知与深度学习》相结合，采用深度学习技术应用于文本的著作权归属判定、作者身份识别、文本风格分析等多个领域。

3.基于基因分子的人类与黑猩猩大脑的语言学习与认知记忆的对比研究

大脑分左、右半脑，对于大多数的人左脑负责语言功能、数学和逻辑推理，右脑负责空间感知、想象、创造、处理关系、音乐和情感等。语言是人类区别于动物的最重要的特征，人类与黑猩猩在基因组 DNA 序列上仅 1%的差别，这些差别的基因大多在人脑中被更高度地表达，造成两者在语言和智力方面的截然不同。① 前人通过比较人类与黑猩猩的大脑皮层的基因表达谱，表明人脑的许多基因增强的表达水平能够为人脑生理学与脑部功能的广泛改进提供基础②。我们通过 SAM（http://www-stat.stanford.edu/~tibs/SAM /）网站，从来自美国 GDS2678 样本库中的 14 个人类和 15 个黑猩猩左脑样本的 12558 个基因中提取了 441 个显著性高表达分子，该库含左大脑的前扣带、前顶下、前颞、额中回、额极等脑区。我们通过生物信息学、整合神经生物学与人工智能技术相结合，建立人脑不同于黑猩猩左脑的分子相互作用和信号通路的激活与抑制网络，结合生物 DAVID 知识库，从作用位点、细胞定位、分子功能和生物过

① Mikkelsen, T. S. et al. Initial sequence of the chimpanzee genome and comparison with the human genome, *Nature*, 437: 69-87 (1 Sep., 2005).

② Enard, W., et al. A Humanized Version of Foxp2 Affects Cortico-Basal Ganglia Circuits in Mice. *Cell*, 137(5): 961-971, 2009.

程角度，对人脑与大猩猩大脑的语言认知与记忆障碍、学习机制、视觉与听觉机制的差异进行了分析，结果表明人类比黑猩猩左脑分子网络有更多的联系和更少的抑制，人类左脑通过核内转位的抗原和非编码 RNA 相互作用表现出很强的 DNA 损伤检测到修复，另外，蛋白质折叠等功能明显增强，阐明其与人类左脑的视觉、声音、语言的感知机制的关系，为了解人类左脑学习和记忆，以及语言进化与语言认知的研究奠定了基础，并开辟了新的研究途径①②③④。

在中国科学院软件研究所、北京大学计算语言学研究所、清华大学人文学院计算语言学实验室和清华大学心理学与认知科学中心共同承担的国家自然科学重点基金项目，国家社会科学基金重大项目以及清华自主科研项目两岸清华大学专项经费的支持下，充分发挥学科优势互补，相信通过开展跨学科的合作研究能够陆续取得一些较好的语言学研究结果。

江铭虎

2016 年于北京

---

① Sun, L.J., Wang, L., Jiang, M. H., et al, "Glycogen Debranching Enzyme 6 (AGL), Enolase 1 (ENOSF1), Ectonucleotide Pyrophosphatase 2 (ENPP2_1), Glutathione S-Transferase 3 (GSTM3_3) and Mannosidase (MAN2B2) Metabolism Computational Network Analysis Between Chimpanzee and Human Left Cerebrum", *Cell Biochem Biophys*, 2011,61(3): 493-505.

② Wang, L., Huang, J. X., Jiang, M. H., et al. Signal Transducer and Activator of Transcription 2 (STAT2) Metabolism Coupling Postmitotic Outgrowth to Visual and Sound Perception Network in Human Left Cerebrum by Biocomputation. *Journal of Molecular Neuroscience*, 2012, 47: 649-658.

③ Wang, L., Huang, J. X., Jiang, M. H., et al. "Adenosylmethionine Decarboxylase 1 (AMD1)-Mediated mRNA Processing and Cell Adhesion Activated & Inhibited Transition Mechanisms by Different Comparisons Between Chimpanzee and Human Left Hemisphere" *Cell Biochemistry and Biophysics*, 2014, 70(1): 279-288.

④ Lin, H., Wang, L., Jiang, M.H., et al. P-glycoprotein (ABCB1) Inhibited Network of Mitochondrion Transport along Microtubule and BMP Signal-Induced Cell Shape in Chimpanzee Left Cerebrum by Systems-Theoretical Analysis, *Cell Biochemistry and Function*, 2012, 30(7): 582-587.

前　言

本书结合了脑认知科学与技术、计算机科学与技术、语言学和计算语言学理论与方法等多学科和多研究领域交叉知识进行跨学科的研究，本书根据书面语言、口语语音和汉语手语的构成特点采用实验的方法对其内在规律进行定量的研究，是作者多年来在语言信息处理方面发表的十余篇学术论文，并结合作者所指导的清华语言学及计算语言学专业的研究生共同发表的若干学术论文的基础上完成的，反映了作者多年来在这方面的研究成果。书中的理论与实验方法包括基于数理统计与聚类的方法，句法分析和最大熵模型，神经网络及量子计算的机器学习方法，ERP 脑电认知方法等，涉及语言学、信息科学、认知科学和计算机科学等学科。

本书共分四章，第一章：词汇表征与特征提取，包括汉语词汇的语义属性及自组织聚类分析，形态复杂词汇的神经表征和处理机制，基于信息瓶颈的特征提取方法以及语用背景信息对语言理解的作用。第二章：文本分类与问答系统，包括基于概念提取的汉语文本自动分类，一种新的基于量子计算的文本聚类与分类器，基于文本聚类的语体特征的量化分析，基于句法分析和最大熵模型的中文问答系统。第三章：语音信息处理，包括语音识别与理解的研究进展，基于汉语音位发音想象的脑机接口，多类分类优化模块神经网络语音识别学习算法，音素识别中时延神经网络及学习算法，基于扩展联想神经网络的语音识别，噪音环境下的语音识别与理解。第四章：汉语手语的信息处理，包括汉语手语信息处理综述，基于决策树算法的 3D 汉语手势分类，基于聋人认知案例的空间隐喻语义认知计算。这些研究结果对语言学内在规律提供了解释，很多方法涉及的学科领域多、知识范围广，是从交叉学科的观察视野和研

究思路对语言的深层内在规律和结构进行探索，进一步了解其本质，本书具有一定的理论和学术研究价值，这些结果均是来自于实验，保证了语言统计规律的真实性、可靠性，全书图文并茂，集科学性、新颖性和综合性于一体，其研究结果可以给我们更多的启示。由于该书属于交叉学科所涉及的领域多、自己水平有限，书中难免有不妥之处，肯请读者朋友鉴谅、理解并批评指正，欢迎来信至我的 Email 地址：jiang.mh@ tsinghua.edu.cn，我将虚心向读者朋友学习，并汇总于该书的再版之中，在此深表谢意。

本书可供高等院校、科研院所语言学与应用语言学、计算语言学和认知语言学的本科生、研究生、教师和研究人员参考。

江铭虎

2016 年于北京

目　录

表索引及表的文字摘要说明

第一章　词汇表征与特征提取

第二章　文本分类与问答系统

第三章　语音信息处理

第四章　汉语手语信息处理

图索引及图的文字摘要说明

第一章　词汇表征与特征提取

第二章　文本分类与问答系统