中国计算语言学研究现状与展望*

耿立波[1,2,3]**　鄂格斐[1,3]**　詹卫东[4]***　杨亦鸣[1,2,3]***

[1] 江苏师范大学语言科学与艺术学院　江苏　徐州　221009
[2] 语言能力省部共建协同创新中心　江苏　徐州　221009
[3] 江苏省语言与认知神经科学重点实验室　江苏　徐州　221009
[4] 北京大学计算语言学教育部重点实验室　北京　100871

提要　“十三五”期间我国的计算语言学取得了长足的进步与发展，受到深度学习算法的推进，基础研究方面出现了较大突破，在语音识别、机器翻译、自动问答系统、知识资源建设、古文字和其他语种语言信息处理等应用方面也出现了很多重要成果。但与世界先进水平相比，目前在很多领域内我们还只是处于跟跑阶段，并且深度学习算法的红利也已接近释放殆尽，在未来仍需要从算法基础架构、人脑语言的本质、深层语言理解等方面展开研究，发展机器语言能力等新兴方向，并积极开展复合型语言学人才的培养。

关键词　计算语言学　深度学习　机器语言能力

中图分类号　H087　　**文献标识码**　A　　**文章编号**　1671－9484(2021)05－0491－09

1 “十三五”期间研究进展和重要成果

十三五期间，作为推动自然语言处理的最新动力，以神经网络为基础具有多层次特征表示的深度学习相对于人为设计准则的传统特征学习，展现出了前所未有的优势。深度学习模型的优势主要在于充分利用了大数据里蕴含的知识(信息)，并将以往的诸多自然语言处理任务整合为统一的端到端或序列到序列模型，从而为看似不同的自然语言处理任务提供了一揽子解决方案。而且由于多层神经网络的超强拟合能力，可以比以往的统计机器学习模型更好地捕捉到输入和输出之间的对应关系，从而使自然语言处理系统达到更好的性能表现。

随着算法、算力的提升配合以大数据的运用，国内的众多高校、实验室和互联网公司，在计算语言学基础研究、应用研究方面都有非常好的表现。

1.1 计算语言学基础研究

1.1.1 算法框架

[收稿日期] 2021年6月20日　[定稿日期] 2021年7月26日　doi:10.7509/j.linsci.202107.034979

*本文得到国家社科基金重大委托项目研究专项(19VXK06)和国家社科基金青年项目(16CYY021)资助。承蒙《语言科学》编辑部和匿审专家提出宝贵的修改意见，谨此一并致以诚挚的谢意。文中谬误概由作者负责。

**共同第一作者。

***共同通讯作者：詹卫东 zhanwd@pku.edu.cn；杨亦鸣 yangym@jsnu.edu.cn。

世界范围内，基于大规模数据的无监督预训练在自然语言处理领域取得了里程碑式的突破。2018 年 10 月，谷歌推出了 BERT 双向编码语言模型，在机器阅读理解的指标上全面超越人类，并且在 11 种不同自然语言理解测试中取得最佳成绩，不过这一模型所消耗的资源代价也非常之高（Devlin 等，2018）。2019 年 6 月，卡耐基梅隆大学与谷歌大脑采用通用自回归预训练方法，提出了 XLNet，在 20 个任务上超过 BERT，并在 18 个任务上取得最优表现（Yang 等，2019）。

国内的很多机构对预训练模型进行了很多深入的研究，2019 年 7 月，国内百度公司发布了 ERNIE 2.0，可以通过持续的多任务学习，逐步学习和建立预训练任务。这一框架在 16 个自然语言处理任务中有着优于 BERT 和 XLNet 的表现（Sun 等，2019）。2019 年 6 月，哈工大讯飞联合实验室发布了基于全词掩码的中文 BERT 预训练模型 BERT-wwm-ext，并在多个中文数据集上取得了当前中文预训练模型的最佳水平（Cui 等，2019）。

微软亚洲研究院周明在 2019 年 ACL 大会上指出，基于无监督语言模型损失函数从大规模无标注文本中学习语境相关词向量表示的预训练模型，在很多自然语言处理任务上取得了巨大的性能提升，形成了一种新的“预训练-微调”自然语言处理范式，即使用大规模无标注数据进行预训练，进而用小规模目标任务标注数据进行微调。〔1〕

1.1.2 知识图谱

在知识图谱领域，在最新的研究中出现了关于类型约束、层次化类型体系、考虑几何空间属性的论文，意味着知识图谱研究正在朝着可解释性更强、意义更明确的表示学习方向发展。〔2〕自然语言处理领域开发出的“无监督的预训练语言模型＋特定任务或语料微调”方式可以为文本的垂直领域知识获取提供帮助：预训练语言模型可以充分捕捉来自通用自然语言语料中的语法与语义信息，微调使通用自然语言处理模型能适应领域语料和领域任务的特性（Zhang 等，2019）。

作为知识图谱的基本单元，实体识别与链接是知识图谱构建和补全的核心技术。随着深度学习在不同领域的火爆，越来越多的深度学习模型被提出用于解决实体识别问题（Lample 等，2016；Xu 等，2017）与实体链接问题（Ganea & Hofmann，2017；Gupta 等，2017；Sil 等，2018）。万静等（2018）提出了一种独立于模式的基于属性语义特征的实体对齐方法。Tang 等（2019）提出了一种新的基于距离的知识图谱嵌入方法，称为基于图上下文的正交变换嵌入（Orthogonal Transform Embedding，OTE），以解决知识图谱中 1-to-N、N-to-1 和 N-to-N 的链接预测问题。刘知远等（2020）展现了数据驱动的深度学习与符号表示的知识图谱之间相互补充和促进的技术趋势。

1.2 计算语言学应用研究

1.2.1 语音识别

2015 年之后，由于端到端技术兴起，语音识别进入了一个新的发展时代，2017 年微软在 Switchboard〔3〕上达到词错误率 5.1%，从而在语音识别的准确性上首次超越了人类，当然这是在一定限定

〔1〕 https://www.msra.cn/zh-cn/news/features/acl-2019-ming-zhou

〔2〕 https://dl.ccf.org.cn/article/articleDetail.html? id=40649683360501768type=xhtx_thesis

〔3〕 一个电话通话录音语料库，作为国际通用的语音识别系统的基准，已被使用超过 20 年，具有广泛的影响力。该数据集是真实的电话数据，数据录音质量比较好，但是说话人口音多样，风格多变，是难度较大的测试集。

条件下的实验结果，还不能推广到比较复杂的场景。[4]

国内在声学模型的研究方面也取得了不小的进展，主流方向是更深更复杂的神经网络技术融合端到端技术，表现比较突出的有，阿里巴巴达摩院机器智能实验室语音识别团队的 Deep-FSMN（Deep Feedforward Sequential Memory Networks）（Zhang 等，2018）、百度的 SMLTA[5]（Streaming Multi-Layer Truncated Attention）等等。在 2018 年举行的的 CHiME-5 比赛中，[6]中国科学技术大学、科大讯飞等合作的 USTC-iFlytek Systems 赢得了 4 项冠军，体现了其技术实力（Du 等，2018）。

在特定任务条件下识别率已经达到很高准确率的基础上，众多研究者开展了语音情感识别研究，利用深度学习算法对发声模型中的众多参数进行特征融合和降维，结合训练数据情感标签，并据此搭建高效准确的基于发声生理的语音情感识别平台，可在一定程度上解决当前智能人机交互中情感分析缺失的问题。目前已有一些研究者基于脑电信号展开情感识别研究，并建立相关识别模型（陈景霞等，2019；田莉莉等，2019）。

1.2.2 机器翻译

机器翻译，作为深度学习在自然语言处理领域最成功的应用之一，在经历了基于规则、基于统计的机器翻译之后，2016 年年底，Google 研究人员推出了新一代机器翻译系统（Wu 等，2016）。这一系统被称为神经机器翻译系统（Neural Machine Translation，NMT），相比于传统的统计机器翻译，NMT 可以训练出来一套能从一个序列映射到另一个序列的神经网络，输出的可以是一个变长（长度可变）的序列，这使得这一系统在翻译、对话和文字概括方面表现更加出色。此外，NMT 需要很少的领域知识。在此之后，学术界和工业界都纷纷开展了神经网络机器翻译系统的研究。

在 ACL2019 年度会议上，百度美国研究院黄亮博士介绍了同声机器翻译的最新进展，为了有效克服同声传译过程中的时间延迟，黄亮团队提出了一种 prefix-to-prefix 的框架，区别于传统的 sequence-to-sequence 的翻译框架，实验证明此框架在保证一定翻译质量的前提下，可以实现严格的时延控制。[7] 在全球学术界公认的国际顶级机器翻译比赛 WMT（Conference on Machine Translation）2019 的比赛中，微软亚洲研究院机器学习组将多个创新的算法运用在了机器翻译的任务中，从学习机制、预训练、网络架构优化、数据增强等方面，大大提升了机器翻译结果的质量，在参加的 11 项机器翻译任务中，有 7 项获得了第一名，另外 4 项获得第二名。[8]

1.2.3 自动问答系统

在自动问答系统方面，深度学习为问答系统提供了一个简单高效的解决方案，将复杂的文本语义信息（词、短语、句子、段落以及篇章等）投射到低维的语义空间中，能够有效克服传统知识挖掘与计算方法以符号表示为基础，泛化能力差，并在一定程度上能够克服符号间语义鸿沟的影响。2017 年，斯坦福大学和 Facebook 人工智能研究所联合提出了 DrQA 深度开放域问答系统，[9]向 DrQA 系统中输入一段文本，然后提一个能在该文本中找到答案的问题，DrQA 就能给出这个问题的答案。

---

〔4〕 https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conver sational-speech-recognition-milestone

〔5〕 http://research.baidu.com/Blog/index-view? id=109

〔6〕 难度极大的语音识别赛事，体现了语音识别领域的众多难点技术，包括多麦克风阵列录音同步问题；快语速和随意的说话风格；高混响和大环境噪声；大量的语音交叠（鸡尾酒会问题）。

〔7〕 http://www.acl2019.org/EN/program/invited-talks.xhtml

〔8〕 http://www.statmt.org/wmt19/papers.html

〔9〕 https://research.fb.com/downloads/drqa

Cui 等(2017)基于知识图谱提出了一种基于模板的问题表示方法,使得模型泛化能力有着显著的提高。Tao 等(2017)提出了一种结合无参考回复和有参考回复的开放领域的对话系统评测方法,该方法同时考虑了问题(先前用户的话语)和参考回复对生成的回复进行评测。通过在两种典型的对话系统上(基于检索和基于生成的)的实验结果表明该模型评价得分与人工评价得分很相近。最近,香港科技大学冯雁教授团队在培养机器回答系统具有记忆和感情、个性化对话等方面,开展了一些具有探索性意义的工作。〔10〕

1.2.4 知识资源建设

北京大学计算语言学教育部重点实验室袁毓林教授主持研制的《北京大学现代汉语实词句法语义功能信息词典》成功上线,该词典是一个电子化的语言知识资源,该词典主要是为汉语自动语义分析和文本生成、汉语国际教育与研究而研制的,可以为汉语的理论研究、教学应用和信息处理工程提供语言知识资源,也可以帮助机器进行语义理解和常识推理(袁毓林和卢达威,2018)。

1.2.5 古文字和其他语种语言信息处理

在古文字信息处理方面,江苏师范大学基于分形几何原理,提出迭代函数系统和分形插值逼近的甲骨字形修复方法,设计甲骨文字形形式化分形描述编码,从而建立甲骨文分形特征数据库,以此解决甲骨文字形识别匹配的问题(顾绍通 2018a,2018b)。在藏语研究领域,西北民族大学建立了开源的藏文分词词性标注系统,提供藏文分词、词性标注功能(李亚超等 2015)。随着一带一路建设的推进,国内以维吾尔语、哈萨克语及柯尔克孜语等为主要领域进行了比较广泛的研究,不过主要采用的是相对过时的技术和方法,前沿理论、技术和方法的应用力度不够(吐尔根等 2018)。

1.3 研究力量布局

研究力量布局方面,在算法、算力、数据有着巨大优势的工科院校、综合院校和互联网巨头公司都有自己计算语言学中心(平台或团队),影响比较大的有清华大学自然语言处理与社会人文计算实验室、清华大学智能技术与系统信息检索组、北京大学计算语言学教育部重点实验室、北京大学计算机科学与技术研究所语言计算与互联网挖掘研究室、哈工大社会计算与信息检索研究中心、哈工大机器智能与翻译研究室、哈工大智能技术与自然语言处理实验室、中科院计算所自然语言处理研究组、中科院自动化研究所自然语言处理研究组、南京大学自然语言处理研究组、复旦大学自然语言处理研究组、东北大学自然语言处理实验室、厦门大学智能科学与技术系自然语言处理实验室、苏州大学自然语言处理实验室、苏州大学人类语言技术研究所、苏州大学自然语言处理实验室、北京理工大学大数据搜索与挖掘实验室自然语言处理团队、百度自然语言处理部、腾讯优图实验室自然语言处理团队、阿里巴巴人工智能实验室自然语言处理团队、微软亚洲研究院自然语言计算组等等。〔11〕

十三五期间,为更好地推动人工智能发展,多所计算语言学见长的高校在原有优势力量的基础上成立了人工智能相关机构:2018 年 3 月,南京大学成立了人工智能学院。2018 年 6 月清华大学成立了人工智能研究院,以此为基础相继成立了知识智能研究中心、自然语言处理与社会人文计算研究中心。2019 年 6 月北京语言大学组建语言智能研究院。此外,北京大学计算语言学教育部重点实验室在 2019 年以优秀成绩通过教育部验收。

〔10〕 http://www.acl2019.org/EN/program/invited-talks.xhtml

〔11〕 https://blog.csdn.net/bbzz2/article/details/72675622

在人才培养与队伍建设方面，十三五期间，北京大学计算语言学教育部重点实验室袁毓林教授、詹卫东教授分别获批教育部长江学者、青年长江学者称号。哈尔滨工业大学社会计算与信息检索研究中心主任刘挺教授于2018年入选国家“万人计划”科技创新领军人才。清华大学自然语言处理研究组唐杰教授、刘洋教授获得“国家杰出青年科学基金项目”。复旦大学崔万云博士论文《基于知识图谱的问答系统关键技术研究》被授予“2017国际计算机学会(ACM)中国优秀博士生论文提名奖”。

2　存在问题和薄弱环节

在自然语言处理诸多领域我们还处于跟跑的阶段，并没有提出原创性很强、可供他人模仿学习的算法模型，中文知识图谱建设也还比较落后，不能为具体应用提供很强的支撑。此外，工业界的数据往往不对外开放，对于问答系统与人机对话等领域进一步推进较为不利。具体表现为以下几点。

1)对于深度学习算法本身的迷信。2012年，Hinton课题组以神经网络〔12〕模型参加ImageNet图像识别大赛，一举夺魁，开始了深度学习井喷式发展的时代。而深度学习之所以能够极大的促进人工智能的发展，技术上的关键在于人们能够将获取的标量数据转变为向量，从而用到机器上。但至今为止，将行为(特征向量)和数据(符号向量)结合起来使用始终是自然语言处理的难点，而这就限制了机器变得更“智能”。

2)对于语言本体研究存在较大的忽视。国务院印发的《新一代人工智能发展规划》中，虽然新一代人工智能布局人才培养时重视人工智能与数学、计算机科学、物理学、生物学、心理学、社会学、法学等学科专业教育的交叉融合，但没有在最具原创性和与人工智能核心架构相关的语言及语言学研究方面给予足够的关注。

3)目前的人工智能还主要以无理解的人工智能为主，尚未发展到有理解的人工智能。在自然语言处理在感知领域，特定任务情况下，已经几乎达到甚至超过人类同样的水平；在认知领域，包括自然语音理解、推理、知识学习等，离人类水平还相差很远。

3　发展趋势和研究的重要领域

回顾过去，人工智能特别是自然语言处理最初的辉煌来自模仿人类的语言规则和推理规则，但是由于那时所模仿的规则本身人们并没有研究清楚，解决不了现实中复杂问题，于是很快走进低谷。进入互联网和大数据时代，“机器学习”特别是“深度学习”技术得到长足发展，人工智能趟出了不同于以往根据规则模仿人类智能而是依据大数据驱动来“学习”人类智能的新机制，获得了极大的发展空间。但目前自然语言理解与人工智能所取得的成功，大都属于无理解的人工智能。在算法、算据、算力的红利消失殆尽之前，我们需要未雨绸缪，探索出新的道路。

3.1　在基础研究方面积极开展语言本身及其脑机制的研究

在人工智能领域内，语言理解也被誉为人工智能皇冠上的明珠。要想摘得这颗明珠，还得从语言的描写和解释的语言学理论本身的研究出发，去找出更符合语言事实和理论统一的语言理解和生成的规

〔12〕 一般认为人工神经网络的研究可以追溯到1943年，之后在很长的一段时间内，并没有得到大多数研究者的关注。2006年Geoffrey Hinton发表在《科学》杂志上的Reducing the dimensionality of data with neural networks是进入新世纪后，神经网络重新焕发生机的经典文献，辅以机器算力的提升与互联网大数据技术的出现，使得消寂一时的人工神经网络升级换代成为深度神经网络又登上了神坛，众多从业人员趋之若鹜。

律以及它们的脑机制和神经基础。〔13〕目前神经语言学研究使用高时间分辨率的事件相关电位技术(ERP)、高空间分辨率的功能性磁共振技术(fMRI)和基因筛查等现代技术手段,对语言认知的神经机制开展了大量探索性研究,对大脑的认识已经从一个“黑匣子”逐渐变成了“灰匣子”。对于语言区的精细亚区划分及其功能的阐明,可为语言处理算法的模块化设计提供有益的借鉴(耿立波等,2021)。在此基础上,我们还要积极开展语言的细胞分子层面研究,在人类大脑中,语言信息的处理是在神经细胞网络中实现,而语言“数据”分布式的存储于神经细胞网络的各个节点(比如由神经元内的离子浓度表征)以及节点之间的连接(比如由突触的强弱表征)上,在人类大脑中,信息的运算和存储在结构上是高度一体化,这不同于现代计算机普遍采用的冯·诺依曼架构,信息处理单元与存储单元是分离的。对于语言的细胞分子研究,可以使我们更好的理解语言在大脑中存储与加工的本质。

3.2 在应用研究方面积极拓展自然语言处理的应用领域

一方面,积极整合利用好国内各个行业的海量数据,充分发挥监督学习和无监督学习的协同作用,融入强化学习的方法,突破“端到端+预训练+微调”的自然语言模型开发应用机制,发挥海量客户端的计算能力和精准数据收集能力,研究自然语言处理模型的分布式终身学习方法;另一方面,针对“小数据”、“小需求”、“小应用”,将机器语言智能应用到各行业中,提升行业自然语言处理应用能力,改善客户体验,降低服务成本;最后,需要实现计算语言学的工程语言学“转向”,在语音识别、人机对话、机器翻译等方面实现工程化、产品化。

3.3 确立机器语言能力新兴方向并培养相关复合型人才

“机器语言能力”这一新兴研究方向的确立,架起了沟通“人脑语言能力”与“人工智能”的桥梁,将会为下一代人工智能的发展贡献力量。2019 年,江苏师范大学语言科学与艺术学院联合国家级省部共建语言能力协同创新中心机器语言能力平台开设了“语言学+人工智能”实验班。“语言学+人工智能”专业的开办,将为培养具有前瞻眼光、贯通语言学及语言脑机制与人工智能基础理论、先进方法技术的复合型人才,从根源上领跑人工智能奠定良好的基础。

4 “十四五”期间重点研究课题

根据以上研究趋势,在十四五期间可以开展以下三个方面的研究。

1)从语言数据到语言知识。深度学习在自然语言处理方面所取得的进展主要得益于语言大数据的有效利用。但从深层语言理解的角度,从跨媒体信息处理的高度来看,目前的深度学习对数据的利用也还存在很多不足,自然语言处理过程的可解释性,计算结果的可靠性以及深度学习如何融入人类知识,如何从大数据中挖掘出可解释的知识系统还有待深入探究。

2)语言智能计算模型的初始化。从语言是思维载体出发,语言是人类遗传结构所决定的音义结合的符号系统,并由递归性的语法信息所整合,整个过程由语言脑结构和功能机制动态完成。需重点关注语言的脑结构联结模式与参数,语义通达的时间进程,词库与语法结构整合的多脑区多时间窗口动态机制等问题。

〔13〕哈尔滨工业大学刘挺团队利用江苏师范大学神经语言学团队“言语加工中语义快速通达”相关研究成果,跨越了传统的“先句法后语义”的分析方式,改善了中文句子分析的计算分析模式。

3)语言脑机制启发的智能模型算法实现。重点关注的问题是模型如何通过编码和解码实现语言计算模块与外部输入与输出的转换;在语言计算模块内部如何将语法和语义信息的约束融入到处理模块中;加强有效训练模型研制,并在真实世界的自然语言应用场景中检验模型性能。

参考文献

Chen, Jingxia (陈景霞), Liyan Wang (王丽艳), Xiaoyun Jia (贾小云), & Pengwei Zhang (张鹏伟). 2019. Jiyu shendu juanji shenjing wangluo de naodian xinhao qinggan shibie 基于深度卷积神经网络的脑电信号情感识别 [EEG-based emotion recognition using deep convolutional neural network]. *Jisuanji Gongcheng yu Yingyong* 计算机工程与应用 [*Computer Engineering and Applications*] 55.18:103－110.

Cui, Wanyun, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang & Wei Wang. 2016 (2017). KBQA: Learning question answering over QA corpora and knowledge bases. *Proceedings of the VLDB Endowment* 10.5: 565－576.

Devlin J., Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2－7. Minneapolis, Minnesota, USA .

Du, Jun, Gao Tian, Sun Lei, Ma Feng, Fang Yi, & Liu Diyuan, et al. 2018. The USTC-iFlytek system for CHiME-5 challenge. *The 5th International Workshop on Speech Processing in Everyday Environments*, September 7. Hyderabad, India.

Ganea, Octavian, & Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. 2017 *Conference on Empirical Methods in Natural Language Processing*. September 7－11. Copenhagen, Denmark.

Geng, Libo (耿立波), Li Yang (杨丽), Jiaoyan Fang (方娇艳), & Yiming Yang (杨亦鸣). 2021. Rennao ruhe xuexi xinde yuyan guize 人脑如何学习新的语言规则 [How brain acquire new language rule]. *Zhongwen Xinxi Xuebao* 中文信息学报 [*Journal of Chinese Information Processing*] 35.5:27－37, 62.

Gu, Shaotong (顾绍通). 2018a. Jiyu fenxing jihe de jiaguwen zixing shibie fangfa 基于分形几何的甲骨文字形识别方法 [Identification of oracle-bone script fonts based on fractal geometry]. *Zhongwen Xinxi Xuebao* 中文信息学报 [*Journal of Chinese Information Processing*] 32.10: 138－142.

Gu, Shaotong (顾绍通). 2018b. Jiyu diedai hanshu xitong he fenxing chazhi de jiagu zixing lunkuo xiufu fangfa 基于迭代函数系统和分形插值的甲骨字形轮廓修复方法 [Inpainting method of oracle-bone inscriptions based on iterated function system and fractal interpolation]. *Kexue Jishu Yu Gongcheng* 科学技术与工程 [*Science Technology and Engineering*] 2018.36:87－92.

Gupta, Nitish, Sameer Singh, Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. *Conference on Empirical Methods in Natural Language Processing*, September 7－11. Copenhagen, Denmark..

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, & Chris Dyer. 2016. Neural architectures for named entity recognition. 2016 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. June 12－17. San Diego, California, USA.

Li, Yahao (李亚超), Jing Jiang (江静), Yangji Jia (加羊吉), & Hongzhi Yu (于洪志). 2015. TIP-LAS: Yige kaiyuan de zangwen fenci cixing biaozhu xitong TIP-LAS: 一个开源的藏文分词词性标注系统 [TIP-LAS: An open source tooklit for Tibatan word segmentation and postagging]. *Zhongwen Xinxi Xuebao* 中文信息学报 [*Journal of Chinese Information Processing*] 29.6:203－207.

Liu, Zhiyuan (刘知远), Xu Han (韩旭), & Maosong Sun (孙茂松). 2020. *Zhishi Tupu Yu Shendu Xuexi* 知识图谱与

深度学习[*Knowledge Map and Deep Learning*]. Beijing: Qinghua Daxue Chubanshe 北京：清华大学出版社 [Beijing: Tsinghua University Press].

Sil, Avirup, Gourab Kundu, Radu Florian, & Hamza Wael. 2017. Neural cross-lingual entity linking. *https://arxiv.org/abs/*1712.01813*v*1.

Yu, Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, & Hua Wu, et al. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. *The* 34*th AAAI Conference on Artificial Intelligence*, February 7－12. New York, USA.

Tang, Yun, Jing Huang, Guangtao Wang, Xiaodong He, & Bowen Zhou. 2020. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. *The* 58*th Annual Meeting of the Association for Computational Linguistics*, July 5－10. Online.

Tao, Chongyang, Lili Mou, Dongyan Zhao, & Rui Yan. 2018. RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. *The* 32*th AAAI Conference on Artificial Intelligence*, February 2－7. New Orleans, Louisiana, USA. 722－729.

Tian, Lili (田莉莉), Junzhong Zou (邹俊忠), Jian Zhang (张见), Zuochen Wei (卫作臣), & Chunmei Wang (汪春梅). 2019. Jiyu gaijin de juanji shenjing wangluo naodian xinhao qinggan shibie 基于改进的卷积神经网络脑电信号情感识别 [Emotion recognition of EEG signal based on improved convolutional neural network]. *Jisuanji Gongcheng yu Yingyong* 计算机工程与应用 [*Computer Engineering and Applications*] 55.22:99－105.

Tuergun Ibranhim (吐尔根·依布拉音), Abiderexiti Kahaerjiang (卡哈尔江·阿比的热西提), Wumaier Aishan (艾山·吾买尔), & Maimaiti Maihemuti (买合木提·买买提). 2018. Zhongya yuyan ziran yuyan chuli zongshu 中亚语言自然语言处理综述 [A survey of central Asian language processing]. *Zhongwen Xinxi Xuebao* 中文信息学报 [*Journal of Chinese Information Processing*] 32. 5: 1－13,21.

Wan, Jing (万静), Ling Li (李琳), Huanchun Yan (严欢春), & Shaohua Wang (王少华). 2018. Jiyu VS-Adaboost de shiti duiqi fangfa 基于 VS-Adaboost 的实体对齐方法 [An entity alignment approach based on the VS-Adaboost algorithm]. *Beijing Huagong Daxue Xuebao* (*ziran kexue ban*) 北京化工大学学报 (自然科学版) [*Journal of Beijing University of Chemical Technology* (*natural science edition*)] 2018.1:72－77.

Wu, Yonghui, Mike Schuster, Zhifeng Chen , Quoc Le, Mohammad Norouzi, Wolfgang Macherey et al. 2016. Google's neural machine translation dystem: Bridging the gap between human and machine translation. *https://arxiv.org/abs/*1609.08144.

Xu, Mingbin, Hui Jiang, & Watcharawittayakul Sedtawut. 2017. A local detection approach for named entity recognition and mention detection. *The* 55*th Annual Meeting of the Association for Computational Linguistics*, July 30－August 4. Vancouver, Canada.

Yuan, Yulin (袁毓林), & Dawei Lu (卢达威). 2018. Zenyang liyong yuyan zhishi ziyuan jinxing yuyi lijie he changshi tuili 怎样利用语言知识资源进行语义理解和常识推理 [On semantic knowledge resources for language understanding and reasoning]. *Zhongwen Xinxi Xuebao* 中文信息学报 [*Journal of Chinese Information Processing*] 32.12: 11－23.

Zhang Shiliang, Lei Ming, Zhijie Yan, & Lirong Dai. 2018. Deep-FSMN for large vocabulary continuous speech recognition. *https://arxiv.org/abs/*1803.05030*v*1.

Zhang, Zhengyan, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, & Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. *The* 57*th Annual Meeting of the Association for Computational Linguistics*, July 28－August 2. Florence, Italy.

Zhilin, Yang, Zhihang Dai, Yiming Yang , Jaime Carbonell, Russ R. Salakhutdinov, & Quoc V. Le. XLNet. Generalized

autoregressive pretraining for language understanding. 2019 *Advances in Neural Information Processing Systems*, December 8－14. Vancouver, Canda.

作者简介

耿立波，男，1982年生，江苏灌南人。江苏师范大学语言科学与艺术学院副教授，研究方向为工程语言学、神经语言学。

鄢格斐，男，1979年生，江苏丹阳人。江苏师范大学语言科学与艺术学院实验师，研究领域为自然语言处理。

詹卫东，男，1972年生，浙江衢州人。北京大学计算语言学教育部重点实验室副主任、中文系教授，研究方向为现代汉语形式语法、语言知识工程、机器语言理解能力评测任务的设计与数据集研制等。

杨亦鸣，男，1957年生，江苏连云港人。江苏师范大学语言科学与艺术学院教授，研究方向为神经语言学、理论语言学、工程语言学等。

Research Status and Prospects of Computational Linguistics

Geng Libo[1,2,3]　Feng Gefei[1,3]　Zhan Weidong[4]　Yang Yiming[1,2,3]

[1] *School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou　Jiangsu*　221009

[2] *Jiangsu/China Collaborative Innovation Center for Language Ability, Xuzhou　Jiangsu*　221009

[3] *Jiangsu Key Laboratory of Language and Cognitive Neuroscience, Xuzhou　Jiangsu*　221009

[4] *MOE Key Laboratory of Computational Linguistics, Peking University, Beijing*　100871

Abstract　Computational linguistics had made considerable progress in China during the period of the "13th Five-Year Plan". Inspired by deep learning algorithms, certain breakthroughs have been made in the basic research aspects. In the meantime, this period also witnessed valuable applications in computational linguistics in areas such as speech recognition, machine translation, automatic question and answer systems, knowledge resource construction, as well as language and information processing of ancient Chinese characters and other languages. However, we are still catching up in many fields. Moreover, the dividends brought by deep learning algorithms are draining away. To win the competition, we need to conduct basic research on algorithms, the neural mechanism of language, and the nature of language comprehension. We also need to invest in training interdisciplinary talents and strengthen research in emerging areas such as machine language ability.

Keywords　computational linguistics; deep learning; machine language ability