计算语言学与深度学习*

作者：克里斯托弗·D.·曼宁(Christopher D. Manning)
译者：刘知远　李若愚

关键词：计算语言学　深度学习

深度学习的海啸

近年来，深度学习的浪潮一直拍打着计算语言学的海岸，而2015年似乎是深度学习“海啸”全力冲击自然语言处理领域主流会议的一年。然而，有些学者预测，海啸最终的破坏力还会更加剧烈。与在法国里尔举行的2015国际机器学习大会(International Conference on Machine Learning, ICML)同时，举行了差不多同样规模的2015深度学习研讨会。在研讨会结束时举行了一场专题讨论，在讨论中尼尔·劳伦斯(Neil Lawrence)说：“自然语言处理的处境现在有点像只夜间公路上的兔子，被‘深度学习’这部高速行驶‘汽车’的‘前大灯’晃瞎了眼睛，只能束手待毙。”计算语言学者们应当认真对待这个论断。对我们而言，深度学习就是我们研究的终点了吗？这些关于深度学习威力的预测从何而来呢？

2015年6月，在脸谱(Facebook)巴黎人工智能实验室的成立大会上，实验室主任雅恩·乐昆(Yann LeCun)说：“深度学习的下一个重要目标是自然语言的理解，这将让机器不只是具有理解单个字词的能力，还将具备理解句子和段落的能力。”[1]2014年11月，在一场红迪网(Reddit)[2]在线提问(Ask Me Anything，AMA)活动中，欣顿(Geoffrey Hinton)说：“我认为接下来的五年里最激动人心的领域将是真正理解文本和视频。如果这五年里我们无法让机器自动浏览YouTube视频并能够讲述视频中发生的事情，那么我会非常失望。在这几年时间内，我们要把深度学习功能嵌入芯片中，这样就可以把能够自动翻译英语的芯片植入人耳，就像‘巴比伦鱼(Babel Fish，《银河系漫游指南》中的虚拟外形鱼，只要放在耳朵里就可以翻译任何语言)’那样实现即时翻译。”[3]而深度学习的第三位重要学者尤舒·本希奥(Yoshua Bengio)，也逐渐将他课题组的研究转向自然语言，最近在神经机器翻译(neural machine translation)系统方面取得激动人心的成果。不是只有深度学习专家才有这样的看法。在2014年9月的一场在线提问活动中，机器学习的学术带头人迈克尔·乔丹(Michael Jordan)被问到，“如果你获得10亿美元经费，资助你领导一项大型研究项目，你想要做什么？”他的回答是：“我将使用这10亿美元来建立一个美国航空航天局规模的项目，专注于自然语言处理，包括所有重要的问题(语义分析、语用学等)。”他还表示：“理智地看，我认为自然语言处理非常迷人，让我们得以集中研究高度结构化的推理问题，研究那些通

*本文译自*Computational Linguistics*，“Computational Linguistics and Deep Learning”，2015,41(4)一文。

[1] http://www.wired.com/2014/12/fb/。

[2] 美国知名社交新闻论坛网站。

[3] https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama_geoffrey_hinton。

向‘什么是思想’的核心又还具有重要实际意义的课题，以及研究无疑会让世界变成更美好地方的技术。”这应当是不错的前景。那么计算语言学者们还要害怕深度学习么？我认为并不需要。回到欣顿提及的《银河系漫游指南》(*Hitchhiker's Guide to the Galaxy*)，我们把书翻过来看看它的封底，那里有醒目的友情提示：“**别慌**。”

深度学习的成就

毫无疑问，近年来深度学习取得了惊人的进展，我不想穷举所有的成功案例，在这里我只举一个例子。最近有篇谷歌博客文章介绍了谷歌语音(Google Voice)新推出的语音转写系统Neon。在承认谷歌语音过去的语音邮件转写功能不够智能的问题后，博客开始介绍Neon是新开发出来的、更准确的转写系统。博客是这么说的：“通过采用一个（深呼吸！）长短时记忆深度递归神经网络(long short-term memory deep recurrent neural network)（哇！），我们将转录错误率降低了49%。”研制一种新的解决方案，将过去最好的系统的错误率减半，这不正是我们每个人的梦想么？

为什么计算语言学家不需要担心

迈克尔·乔丹在在线提问活动中表示他不相信深度学习能够解决自然语言处理问题，他指出：“虽然现在很多深度学习研究声称能够搞定自然语言处理，但我认为：(1)深度学习在自然语言处理方面表现出来的作用，要远小于其在像机器视觉那样的其他领域；(2)利用黑盒学习框架从大规模数据中学习，对自然语言处理不像在视觉那样的其他领域那么有效。”乔丹提到的第一个理由无疑是对的：到目前为止，在更高级的自然语言处理任务中，深度学习并未像在语音识别、物体识别等任务上做到的那样，显著降低错误率。尽管在这些任务上采用深度学习技术能够有些成效，但与在其他方面所得到的动辄25%或50%的错误率降低相比，不免相形见绌。因此，我们容易想到这种状态也许会一直持续，深度学习只有在视听等信号处理任务上才有可能创造奇迹。另一方面，我对乔丹的第二个理由并不以为然。但是，对于为什么自然语言处理无须担心深度学习，我的确有自己的两个理由：(1)机器学习领域最聪明、最有影响力的人都在表示自然语言处理是值得关注的问题领域，这对于我们领域而言应该是很好的事情；(2)我们的领域是关于语言技术的科学，而不是寻找最好的机器学习方法——我们领域的核心课题仍然是这些领域问题。而这些领域问题并未消失。约瑟夫·雷辛格(Joseph Reisinger)在他的博客中写道：“我经常会遇到那些推销‘通用机器学习’的创业公司，老实讲这个创意很荒唐。机器学习不是无差别举重项目[4]，也不能像亚马逊弹性计算云EC2(Elastic Compute Cloud)那样商品化(commoditizable)，机器学习与其说是编程，不如说更像设计。”[5]而从事语言学和自然语言处理的人正是设计师。近几年的国际计算语言学会议(the Association for Computational Linguistics, ACL)把注意力过分集中于数字和击败最好的方法上，那还不如改名叫“Kaggle[6]大赛”好了。我们的领域应当更关注问题、方法和架构本身。最近，我和很多合作者一起在Universal Dependencies（通用依赖关系）的发展上投入大量精力，旨在研制一个能够用于所有人类语言的通用的依存句法表示、词性标注(Part of Speech tagging, POS)和特征标签集合，并希望其具有可接受的正确率和易用性。这只是一个例子，在我们领域还有很多其他正在进行

[4] 指一定公斤数以上的无差别级别举重。——译者注

[5] http://googleblog.blogspot.com/2015/07/neon-prescription-or-rather-new.html。

[6] Kaggle是一个数据建模和数据分析竞赛平台，企业和研究者可在其上发布数据，统计学者和数据挖掘专家可在其上进行竞赛以产生最好的模型。——译者注

的设计工作，例如抽象语义表示 (abstract meaning representation)。

语言的深度学习

深度学习可以在自然语言处理的哪些方面发挥作用呢？到目前为止，自然语言处理取得的主要提升并非来自真正的深度学习（即采用层次结构的抽象表示来提高泛化能力），而是更多地来自分布式的词表示，即对词语和概念采用实数向量表示。采用稠密、多维向量表示方法来计算词语相似度，对自然语言处理和相关领域都非常有用。实际上，分布式表示的重要性可以追溯到神经网络早期提出的“并行分布式处理”的思想，在当时这更多地还是从认知科学角度受到关注 [11]。分布式表示可以更好地解释类似人类的泛化能力，同时从工程角度来看，通过使用低维、稠密向量进行词表示，能够让我们更好地对大规模上下文语境 (large contexts) 建模，从而极大地改善语言模型。从这个全新的角度来看，传统基于 N-Gram 的语言模型由于阶数增加会带来指数级的稀疏性，似乎不再实用。

我坚信深度模型终将发挥作用。从理论上讲，深度表示内部的共享机制会带来指数级的表征优势，也将在实际应用中改善学习系统的性能。深度学习系统的构建方法很有吸引力且威力强大：研究者只需要定义模型架构和顶层损失函数，就能通过一个端到端的学习框架，自动调整模型参数和表示，从而将这个损失函数最小化。在最近的神经机器翻译 [5,14] 中，我们已经见识到这种深度学习系统的威力。

最后，我一直倡导更多地关注模型的语义组合能力，特别是语言，还有通用人工智能。智能需要具备以小见大的能力，能够从较小的部分理解更大的整体。特别是在语言方面，对新颖复杂句子的理解很大程度上需要从句子成分，即词或词组，组合构造出句子的意思。近来很多论文展示了如何使用来自“深度学习”的分布式词表示的改进系统，例如 word2vec[7] 和 GloVe[8]。但是，这些工作并没有真正地构建深度学习模型。我希望未来能有更多的人关注“是否能够构建具有组合语义能力的深度学习系统”这种更贴近语言学范畴的问题。

与计算语言学和深度学习相关的科学问题

我希望大家不要陷入采用词向量获取不过几个百分点性能提升的怪圈。我更强烈地希望，大家能够回到一些有趣的语言和认知问题上来，能够推进非范畴的表示和神经网络方法。

表1　乔姆斯基的4种核心范畴

		V +	V −
N	+	Adjective: an *unassuming* man 形容词: 一个**不爱出风头的**人	Noun: the *opening* of the store 名词: 商店的**开业**
	−	Verb: she is *eating* dinner 动词: 她**正在吃**晚餐	Preposition: *concerning* your point 介词: **关于**你的观点

语言中非范畴现象的一个例子是，V-ing 形式动名词（如 driving）的词性标注问题。该形式通常被认为介于动词和名词之间。而实际上，情况可能还更加复杂，因为 V-ing 形式实际上可以出现在乔姆斯基 (Chomsky)[1] 提出的所有四种核心范畴中，见表 1。

更有意思的地方在于，有证据表明，语言在名词和动词之间不仅存在模糊性，还存在着混合状态。例如，经典的语言学教材会指出限定词要与名词共同使用，而判断一个词是不是动词，可以看它能不能带直接宾语。然而我们知道，动名词的名词化用法可以两者兼顾：

(1) The not *observing* this rule is that which the world has blamed in our satorist. (Dryden, *Essay Dramatick Poesy*, 1684, page 310)

世界给予我们饥荒以作为我们无视这条规则的惩罚。

(2) The only mental provision she was making for the evening of life, was the *collecting* and transcribing

all the riddles of every sort that she could meet with. (Jane Austen, *Emma*, 1816)

她为人生暮年准备的唯一精神食粮是收集并且改编她能看到的所有种类的谜语。

(3) The difficulty is in the getting the gold into Erewhon. (Sam Butler, *Erewhon Revisited*, 1902)

困难在于将黄金送进乌有之乡。

这些现象经常会在短语结构树中被解释为某种词类活用 (category-change operation)，然而有充分的证据表明，这其实是语言的非范畴行为。

实际上，这样的结构早期曾被罗斯 (Ross)[10] 作为范畴“挤压 (squish)”的例子。随着时代变迁，V-ing 形式正逐步动词化，但是在很多时候，它仍呈现明显的胶着状态。例如，我们可以对以下句子有清楚的判断：

(4) Tom's ***winning*** the election was a big upset.

汤姆赢得这次选举是一个大逆转。

(5) ?This ***teasing*** John all the time has got to stop.

约翰一直以来受到的嘲弄停止了。

(6) ?There is no marking exams on Fridays.

周五没有计分的考试。

(7) *The ***cessation*** hostilities was unexpected.

停战是出乎意料的。

各种限定词和动词宾语的组合听着有点儿别扭，但比在 -ation 名词化的后面放直接宾语好得多。文献 [2] 表明，与在口语中 -ing 和 -in' 之间变换的连续性解释相比，对 V-ing 形式的离散词性标注分类预测并不太成功，这表明“语法范畴存在连续性，从而使不同类别之间的边界不那么明显。”

我研究生时期的同学惠特尼·泰伯 (Whitney Tabor) 提供了另一个不同且有趣的案例。泰伯 [15] 研究了 kind of 和 sort of 的使用，我曾将其作为 1999 年出版的教材 [6] 绪论中的例子。名词 kind 和 sort 既可以放在一个名词的前面，也可以被用作程度状语：

(8) [That kind [of knife]] isn't used much.

那种小刀用得不太多。

(9) We are [kind of] hungry.

我们有点儿饿。

有趣之处在于，通过对带有歧义形式（例如下面这对例子）做再分析的方式，可以说明一种形式是如何从另一种形式转换而来的。

(10) [a [kind [of dense rock]]]

一种大密度岩石

(11) [a [[kind of] dense] rock]

一块密度有点大的岩石

泰伯 [15] 探讨了为什么古英语有 kind，却很少甚至没有 kind of 的用法。在中古英语的开始阶段，带有歧义的上下文为再分析提供了机会，开始出现表示程度的用法（如句子 (13) 是 1570 年的例子），接着，在那之后，明确表意为程度状语的例子开始出现（如句子 (14) 是 1830 年的例子）：

(12) A nette sent in to the see, and of alle ***kind of*** fishis gedrynge (Wyclif, 1382)

(13) Their finest and best, is a ***kind of*** course red cloth (True Report, 1570)

(14) I was ***kind of*** provoked at the way you came up (Mass. Spy, 1830)

这是历史，而且不是同时出现。而如今，孩子们大概会同时掌握 kind/sort of 的两种用法。有读者注意到我在本文第一段引用的那句话吗？那就是个很好的例子。

(15) NLP is ***kind of*** like a rabbit in the headlights of the deep learning machine (Neil Lawrence, DL workshop panel, 2015)

自然语言处理现在有点像只（夜间公路上）被“深度学习”（这部高速行驶“汽车”的“前大灯”）晃瞎了眼睛的兔子。

惠特尼·泰伯采用了一个小的但已具备深层（2 个隐层）递归特性的神经网络来对这个演化过程建模。他是 1994 年在斯坦福利用与戴夫·鲁梅尔哈特 (Dave Rumelhart) 合作的机会完成这件事的。

就在最近，开始出现一些新的工作，通过利用分布式表示来建模并解释语言演化。文献 [12] 采用了更传统的隐含主题分析方法来产生分布式词表示，展示了词表示方法是如何能够捕捉语义

变化的：随着时间的推移，指称的拓展和窄化。他们考察了很多案例，例如 deer 在古英语中指代任意动物，而在中古英语和现代英语中则被明确指代一类动物。词汇 dog 和 hound 的意义则发生了交换：hound 在中古英语中被用于指代任意犬类，而现在被用于指代狗的一个特定亚种，而 dog 则恰好相反。

库尔卡尼 (Kulkarni) 等人 [4] 使用神经词嵌入表示 (neural word embedding) 来对 20 世纪中的词义演化（如 gay 等词）进行建模（利用在线的谷歌图书词频统计器 (Google Books Ngrams) 语料库）。在近期国际计算语言学大会研讨会上，金 (Kim) 等人 [3] 采用了类似的方法（即 word2vec）来考察最近的词义演化。例如，他们在图 1 中展示了 2000 年左右，

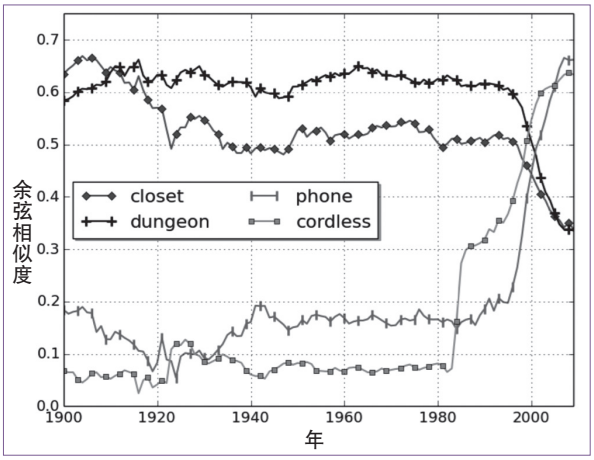

图1 cell词义的变化趋势，通过它与其他4个词的余弦相似度随时间的变化情况来表示（1.0表示最相似，0.0表示完全不相似）

词汇 cell 的含义是怎样从接近 closet（小室）和 dungeon（地牢）迅速转换到接近 phone（电话）和 cordless（无绳）的。某时期的词义是由它的所有义项的使用频率加权平均得到的。

更科学地利用分布式表示和深度学习进行现象建模是神经网络昔日繁荣时的特征。最近在网上有一些关于深度学习引用与功劳归属的纷争，从这个角度来看，我认为有两个人的功劳鲜有人提及，他们是戴夫·鲁梅尔哈特和杰伊·麦克利兰 (Jay McClelland)。从圣迭戈 (San Diego) 的并行分布式处理研究组 (the Parallel Distributed Processing Research Group) 开始，他们就致力于对神经网络进行更科学的和认知层面的研究。

现在，对于神经网络是否足以应对由规则支配的语言行为，的确有一些很好的质疑和问题。我们领域中的老成员应该还记得，对于神经网络是否足以应对由规则支配的语言行为的争论，正是史蒂夫·平克 (Steve Pinker) 成名的基础——也是他的 6 个研究生学术生涯的基础。此处没有篇幅展开讨论这些问题。但说到底，我认为这是一场富有成果的争论。这场争论引发了保罗·斯莫伦斯基 (Paul Smolensky) 的大量工作，旨在探讨神经基质中范畴系统是如何形成和表示的 [13]。实际上，可以说保罗·斯莫伦斯基在爱丽丝的“兔子洞”里走得太远了 [7]，他大部分学术生涯用在了建立一个新的语音范畴模型，即优选理论 [9]。现在有很多早期的科学工作被忽视了。最好是将自然语言处理重点回到自然语言处理的认知和科学研究，而非停留在几乎完全使用某种工程模型的研究上。

总的来说，无论对于机器学习的未来发展，还是工业应用问题，自然语言处理都被认为是重点，生活在这样的时代，我认为我们应当感到激动和高兴。未来是光明的。不过，我仍会鼓励大家去思考人类语言的问题、架构、认知科学和相关细节，语言是如何习得的、处理的和变化的，而不是仅仅追逐某个标准测试集上最高评测数值。■

致谢：

这篇“结束语 (Last Words)”栏目的文章包含了我在 2015 年国际计算语言学大会主席报告的部分内容。感谢保拉·梅洛 (Paola Merlo) 建议将它写出来发表。

[7] Rabbit hole，《爱丽丝奇境历险记》中的兔子洞，爱丽丝掉进兔子洞，坠入了神奇的地下世界，此处指保罗开拓了一个新的领域。——译者注

作　者：

克里斯托弗·D.·曼宁(Christopher D. Manning)：

就职于美国斯坦福大学计算机科学与语言学系。

manning@cs.stanford.edu

译　者：

刘知远

CCF高级会员。清华大学助理研究员。主要研究方向为自然语言处理与社会计算。liuzy@tsinghua.edu.cn

李若愚

CCF学生会员。清华大学本科生。

liruoyusince1995@gmail.com

参考文献

[1] Chomsky, Noam. 1970. Remarks on nominalization. In R. Jacobs and P. Rosenbaum, editors, *Readings in English Transformational Grammar*. Ginn, Waltham, MA, pages 184~221.

[2] Houston, Ann Celeste. 1985. Continuity and Change in English Morphology: The Variable (ing). Ph.D. thesis, University of Pennsylvania.

[3] Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61~65, Baltimore, MD.

[4] Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015.Statistically significant detection of linguistic change. In *Proceedings of the 24th International World Wide Web Conference* (WWW 2015), pages 625~635, Florence.

[5] Luong, Minh-Thang, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 11~19, Beijing.

[6] Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

[7] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26 (NIPS 2013). Curran Associates, Inc., pages 3111~3119.

[8] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014), pages 1532~1543, Doha.

[9] Prince, Alan and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar.* Blackwell, Oxford.

[10]Ross, John R. 1972. The category squish: Endstation Hauptwort. In *Papers from the Eighth Regional Meeting*, pages 316~328, Chicago.

[11]Rumelhart, David E. and Jay L. McClelland, editors. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1: Foundations. MIT Press, Cambridge, MA.

[12]Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In Kathryn Allen and Justyna Robinson, editors, *Current Methods in Historical Semantics*. De Gruyter Mouton, Berlin, pages 161~183.

[13]Smolensky, Paul and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, volume 1. MIT Press, Cambridge, MA.

[14]Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M.Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27 (NIPS 2014). Curran Associates, Inc., pages 3104~3112.

[15]Tabor, Whitney. 1994. Syntactic Innovation: A Connectionist Model. Ph.D. thesis, Stanford.