自然语言处理新范式：基于预训练模型的方法

New Paradigm of Natural Language Processing: A Method Based on Pre-Trained Models

车万翔/CHE Wanxiang， 刘挺/LIU Ting
（哈尔滨工业大学，中国 哈尔滨 150001）
(Harbin Institute of Technology, Harbin 150001, China)

DOI：10.12142/ZTETJ.202202002
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20220408.1732.008.html
网络出版日期：2022-04-11
收稿日期：2022-02-26

摘要：以BERT和GPT为代表的、基于超大规模文本数据的预训练语言模型能够充分利用大模型、大数据和大计算，使几乎所有自然语言处理任务性能都得到显著提升，在一些数据集上达到甚至超过人类水平，已成为自然语言处理的新范式。认为未来自然语言处理，乃至整个人工智能领域，将沿着“同质化”和“规模化”的道路继续前进，并将融入多模态数据、具身行为数据、社会交互数据等更多的“知识”源，从而为实现真正的通用人工智能铺平道路。

关键词：人工智能；自然语言处理；预训练语言模型；同质化

Abstract: Pre-trained language models based on super-large-scale raw corpora, represented by BERT and GPT, can make full use of big models, big data, and big computing, which have significantly improved the performances of almost all-natural language processing tasks. The performances have reached or exceeded the human level on some datasets. Pre-trained language models have become a new paradigm for natural language processing. It is believed that in the future, natural language processing and even the entire field of artificial intelligence will continue to move forward along the path of “homogenization” and “scale”, and will integrate more sources of “knowledge”, such as multi-modal data, embodiment data, and social interaction data. Consequently, these methods will pave the way for achieving true general artificial intelligence.

Keywords: artificial intelligence; natural language processing; pre-trained language model; homogenization

1 自然语言处理的背景

自然语言通常指的是人类语言（本文特指文本符号，而非语音信号），是人类思维的载体和交流的基本工具，也是人类区别于动物的根本标志，更是人类智能发展的外在体现形式之一。自然语言处理（NLP）主要研究用计算机来理解和生成自然语言的各种理论和方法，属于人工智能领域的一个重要甚至核心的分支。人工智能应用领域的快速拓展对自然语言处理提出了巨大的应用需求。同时，自然语言处理研究为人们更深刻地理解语言的机理和社会的机制提供了一条重要的途径，因此具有重要的科学意义。

目前，人们普遍认为人工智能的发展先后经历了运算智能、感知智能、认知智能3个阶段。其中，运算智能关注的是机器的基础运算和存储能力。在这方面，机器已经完胜人类。感知智能则强调机器的模式识别能力，如语音的识别和图像的识别，目前机器在感知智能上的水平基本达到甚至超过了人类的水平。然而，在涉及自然语言处理以及常识建模和推理等研究的认知智能上，机器与人类还有很大的差距。

为什么计算机在处理自然语言时会如此困难呢？这主要是因为自然语言具有高度的抽象性、近乎无穷变化的语义组合性、无处不在的歧义性和持续的进化性，并且理解语言通常需要背景知识和推理能力等。由于面临以上问题，自然语言处理已成为目前制约人工智能取得更大突破和更广泛应用的瓶颈之一。包括图灵奖得主在内的多位知名学者都很关注自然语言处理，甚至图灵本人，也将验证机器是否具有智能的手段（即“图灵测试”）应用在自然语言处理上。因此，自然语言处理又被誉为“人工智能皇冠上的明珠”。

2 自然语言处理问题的解决之道

自然语言处理的本质是形式与意义的多对多映射关系。

其中，形式和意义的空间都近乎无限。那么，如何才能找到正确的映射关系呢？利用“知识”进行约束是唯一有效的办法。因此，如何获取和利用“知识”成为解决自然语言处理问题的关键科学问题。

应用于自然语言处理的知识来源主要有三大类：狭义知识、算法和数据。表1对这三大类知识来源进行了总结。

第一大类知识是狭义知识，即通过规则或词典等形式由人工定义的显性知识，也就是人们通常所理解的知识类型。具体来讲，狭义知识又包括3类，即语言知识、常识知识和世界知识。其中，语言知识是指对语言的词法、句法或语义进行的定义或描述。例如，WordNet[1]是由普林斯顿大学编写的英文语义词典，其主要特色是定义了同义词集合。每个同义词集合由具有相同意义的词组成。此外，WordNet还为每个同义词集合提供简短的释义，同时不同同义词集合之间还具有一定的语义关系。常识知识是指人们基于共同经验而获得的基本知识。常识往往是不言自明的，并没有记录为文字，所以很难从文本中挖掘到。著名的Cyc项目试图将上百万条知识编码成机器可用的形式，用以表示人类常识。世界知识包括实体、实体属性、实体之间的关系等。这类知识往往通过知识图谱的形式加以描述和存储。

第二大类知识是算法。算法的本质是解决问题的过程或者方法，它也是一种知识类型。机器学习算法则可以看作人为定义的函数。虽然这种函数的参数是由机器自动学习获得的，但是函数的类型仍然由人类来定义。这在某种程度上反映了设计者对待解决问题的认知，具有一定的归纳偏执性。例如，卷积神经网络（CNN）就利用了识别对象的平移不变性质。面向各种自然语言处理问题的算法更是和语言知识密切相关。与狭义知识相比，算法知识具有更好的灵活性和动态性。

▼表1 自然语言处理中的三大类知识来源

知识类别	细分类	特点	举例
狭义知识	语言知识	词典、规则库	WordNet
	常识知识	很难从文本中挖到	Cyc项目
	世界知识	可以从文本中挖到	知识图谱
算法	浅层机器学习	人工提取特征	SVM、CRF
	深度学习	自动归纳特征	MLP、RNN、CNN
	NLP算法	利用语言知识	CKY、MST
数据	有标注	专家标注、众包	Penn TreeBank
	无标注	大量原始语料	预训练语言模型
	伪数据	与目标任务近似	数据增广

CKY：Cocke-Younger-Kasami算法　MLP：多层感知机　RNN：循环神经网络
CNN：卷积神经网络　MST：最大生成树　SVM：支持向量机
CRF：条件随机场　NLP：自然语言处理

第三大类知识是数据。数据是机器学习算法的基础。机器学习算法通常会依赖有标注的数据，需要借助人工方式来为每个输入标注出相应的输出结果。由于并没有具体指明如何从输入转换到输出，因此数据是一种隐性的知识。然而，由人工进行数据标注的方式费时又费力，导致标注数据量往往较小，不足以训练一个性能优异的机器学习算法。为了解决这一问题，预训练语言模型可直接利用大量未标注的原始语料，将语言模型作为训练的目标，即根据历史的词序列来预测下一个单词是什么，或者根据周围的词来预测当前的词是什么。由于未标注数据量近乎无限，因此可以训练一个性能较好的语言模型，并将该模型的参数作为下游任务模型的初始参数。这样便可以减少模型对标注数据量的依赖，大幅提高下游任务的准确率。

3 自然语言处理技术的发展历史

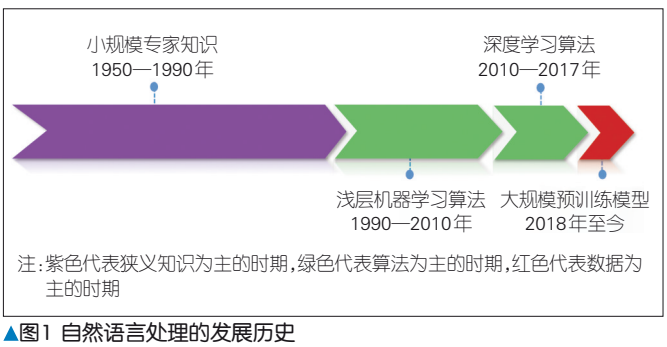
自然语言处理技术自从诞生以来经历了以狭义知识、算法和数据为主的3个时期（如图1所示）。

3.1 狭义知识为主时期

早期的自然语言处理（20世纪50年代到90年代）主要采用基于小规模专家知识的方法（规则、词典等狭义知识），通过专家总结的符号逻辑知识来处理通用的自然语言。然而，由于自然语言的复杂性，基于理性主义的规则方法在实际应用场景中仍存在一些不足。

3.2 算法为主时期

20世纪90年代开始，计算机运算速度和存储容量的快速增加，以及统计学习（浅层机器学习）算法的大量普及，均使得基于小规模语料库的浅层机器学习算法在自然语言处理领域得以大规模应用。由于语料库中包含了一些关于语言的知识，基于浅层机器学习算法的自然语言处理方法能够更加客观、准确、细致地捕获语言规律。这一时期，词法分



注：紫色代表狭义知识为主的时期，绿色代表算法为主的时期，红色代表数据为主的时期

▲图1 自然语言处理的发展历史

析、句法分析、信息抽取、机器翻译、自动问答等领域的研究均取得了一定程度的突破。

尽管如此，基于浅层机器学习算法的自然语言处理技术仍存在明显的局限性，即需要事先利用经验性规则将原始的自然语言输入转化为机器能够处理的向量形式。这一转化过程（也称为特征提取）需要细致的人工操作和一定的专业知识，因此也被称为特征工程。

2010年以后，基于深度神经网络的表示学习方法（也称深度学习方法）逐渐兴起，可以直接端到端地完成各种自然语言处理任务，不再依赖人工设计。这里，表示学习是指机器能根据输入自动发现可用于识别或分类等任务的表示。具体来讲，深度学习模型在结构上通常包含多个处理层。最底层的处理层会接收原始输入，并对原始输入进行抽象处理，然后该层后面的每一层都会在前一层的结果上进行更深层次的抽象处理。最后一层的抽象处理结果即为输入的一个表示，以用于最终的目标任务。其中，抽象处理是由模型内部的参数来控制的，而参数的更新值则是由反向传播算法根据训练数据中模型的表现来学习得到的。由此可以看出，深度学习可以有效避免统计学习方法中的人工特征提取操作，自动地发现对目标任务有效的表示。

深度学习方法还能打破不同任务之间的壁垒。传统浅层机器学习方法需要为不同任务设计不同的特征，而这些特征往往是不通用的。然而，深度学习方法却能够将不同任务在相同的向量空间内进行表示，从而具备跨任务迁移的能力。此外，深度学习方法还可以实现跨语言甚至跨模态的迁移，可以综合利用多项任务、多种语言、多个模态的数据，使得人工智能向更通用的方向迈进一步。

同样，得益于深度学习技术的快速发展，自然语言处理的另一个主要研究方向——自然语言生成也取得了长足进步。长期以来，自然语言生成的研究几乎处于停滞状态：除了使用模板生成一些简单的语句外，并没有获得其他有效的解决办法。随着基于深度学习的序列到序列生成框架的提出，这种逐词文本生成方法全面提升了生成技术的灵活性和实用性，完全革新了机器翻译、文本摘要、人机对话等技术范式。

虽然深度学习技术能够大幅提高自然语言处理系统的准确率，但是基于深度学习的算法仍有一个致命的缺点：过度依赖大规模标注数据。对于语音识别、图像处理等感知类任务，标注数据相对容易获得。例如，在图像处理领域，人们已经为上百万幅图像标注了相应的类别（如ImageNet数据集）。用于语音识别的“语音和文本”平行语料库标注的时间也有几十万小时。然而，自然语言处理具有主观性特点，它所面对的任务和领域又众多。这些均使得大规模语料库标注的时间和人力成本变得很高。因此，自然语言处理的标注数据往往不够充足，很难满足深度学习模型训练的需要。

3.3 数据为主时期

早期的静态词向量预训练模型和后来的动态词向量预训练模型，特别是自2018年以来以BERT、GPT为代表的超大规模预训练语言模型，都很好地弥补了自然语言处理标注数据不足的缺点。这些模型大大促进了自然语言处理技术的发展，使得包括阅读理解在内的几乎所有自然语言处理任务性能都得到了大幅提高，在有些数据集上的性能表现达到甚至超过了人类水平。

模型预训练是指，首先在一个源任务上训练一个初始模型，然后在下游任务（也称目标任务）上继续对该模型进行精调，从而达到提高下游任务准确率的目的。模型预训练本质上是迁移学习思想的一种应用。然而，由于同样需要人工标注，源任务标注数据的规模往往非常有限。那么，如何获得更大规模的标注数据呢？

实际上，文本自身的顺序就是一种天然的标注数据。通过若干连续出现的词语来预测下一个词语（又称语言模型）就可以构成一项源任务。由于图书、网页等文本数据的规模近乎无限，因此模型可以非常容易地获得超大规模的预训练数据。有人将这种不需要人工标注数据的预训练学习方法称为无监督学习方法。其实，这种叫法并不准确。这是因为这种方法的学习过程仍然是有监督的。更准确的叫法应该是自监督学习。

为了能够刻画大规模数据中复杂的语言现象，深度学习模型的容量需要足够大。基于自注意力机制的Transformer模型能够显著提升自然语言建模能力，是近年来具有里程碑意义的进展之一。要想在可容忍的时间内在如此大规模的数据上训练一个超大规模的Transformer模型，就离不开以图形处理器（GPU）、张量处理器（TPU）为代表的现代并行计算硬件。可以说，超大规模预训练语言模型完全依赖“蛮力”，在大数据、大模型和大计算资源的加持下，使自然语言处理取得了长足的进步。例如，OpenAI推出的GPT-3拥有1 750亿个参数，无须接受任何特定任务的训练，便可通过小样本学习来完成10余种文本生成任务（如风格迁移、网页生成等）。

目前，预训练模型已经成为了自然语言处理的新范式。它甚至影响了整个人工智能的研究和应用，开启了人工智能领域“大规模预训练模型”时代的大门。

4 几种具有代表性的预训练语言模型

4.1 词嵌入预训练语言模型

在基于浅层机器学习的自然语言处理时期，人们使用高维、离散的向量来表示自然语言。其中，每个词用独热向量来表示，向量维度表示词的大小（只有一位为1，其余均为0）。然而，这种独热向量表示方法无法解决“多词一义”的问题。也就是说，即便两个词含义相近，它们的表示也是截然不同的。例如，“马铃薯”和“土豆”会使用两个不同的独热向量表示。假如训练数据中只出现“土豆”，那么当测试系统中出现“马铃薯”时，模型就无法进行正确加权。

语言学家J. R. FIRTH在1957年指出，通过一个词周围的词便可理解该词的含义[2]，即“观其伴知其义”。例如，“马铃薯”和“土豆”的周围经常出现“吃”“烹饪”“种植”等，所以这两个词就比较相似。因此，可以将一个词周围出现过的词收集起来，构建一个相对更稠密的向量，然后用该向量来表示这个词。当然，还可以使用降维等技术，用更低维、更稠密的向量来表示词。

2003年，图灵奖得主Y. BENGIO首次提出词嵌入的概念[3]，即直接使用一个低维、稠密、连续的向量来表示词。那么，如何获得（即预训练）一个好的词嵌入表示呢？对此，可以通过其在下游任务上表现，对向量每一维的数值进行自动设置。除了需要一个下游任务外，还需要针对该任务的大规模训练数据，以保证模型能覆盖足够多的语言现象。然而，由于自然语言的主观性，很难获得大规模的标注数据。比如，情感分析数据最多也就几万条。好在语言具有“观其伴知其义”的性质，因此可以通过一个词周围的词，来预测当前的词，这样就自然构成了一个“下游任务”。具体的任务可以分为两类：一类是通过历史词序列来预测下一个词，这类任务又被称作“语言模型”任务；另一类是利用周围的词来预测中间的词，这类任务类似于“完形填空”任务。各种电子文档、图书乃至整个互联网上的文本数据，都可以作为训练数据，从而大大增强了词嵌入表示的学习能力。虽然Y. BENGIO等早在2003年便提出了词向量概念，并通过语言模型任务对词向量进行了预训练，但是直到2013年谷歌的T. MIKOLOV等提出Word2vec模型[4]后，该思想才开始普及。

4.2 上下文相关词嵌入预训练语言模型

虽然词嵌入表示可以处理“多词一义”现象，但是其本身仍然存在一个致命的缺点，即无法处理“一词多义”现象。词嵌入的一个基本假设是：每一个词都对应唯一一个词嵌入表示。如果一个词有多种词义，那么用哪个词义的向量来表示这个词呢？这里我们仍然以“土豆”这个词为例进行说明。作为一种蔬菜时，“土豆”应该和“马铃薯”等词的表示相似；而作为一个视频网站时，“土豆”又应该和“爱奇艺”等词的表示相似。那么，最终“土豆”的词嵌入表示必将是个“四不像”。

为解决上述问题，AllenNLP于2018年提出了ELMo模型[5]。该模型的核心思想是将语言模型的输出作为词向量表示。这种表示是上下文相关的。例如，在句子“我喜欢吃土豆”中，“土豆”的表示应该和“马铃薯”相似；而在句子“我在土豆上看电影”中，“土豆”的表示则应该和“爱奇艺”相似。将ELMo输出的词向量作为特征，大大提高了下游任务的性能。

4.3 大规模预训练语言模型

在ELMo模型提出后不久，OpenAI便提出了第1代GPT模型[6]，正式将自然语言处理技术带入“预训练”时代。和ELMo一样，GPT也把语言模型任务作为预训练任务。总的来说，GPT模型主要有三大创新点：首先，它使用了性能更强大的Transformer模型；其次，它在目标任务上精调整个模型，而不是只将模型的输出结果作为固定的词向量特征；最后，由于预训练模型自身非常复杂，因此接入的下游任务模型可以非常简单，这极大降低了自然语言处理的技术门槛。

在GPT提出后不久，谷歌便提出了著名的BERT模型[7]。与GPT相比，BERT最大的改进在于它能利用词两边的上下文来预测中间的词，即使用“完形填空”作为预训练任务。由于使用了更为丰富的上下文，因此BERT能够获得更好的预训练效果。BERT一问世便刷新了各大自然语言处理任务记录，在有些任务上的表现甚至超越了人类。

随后，微软也建造了自己的超大规模人工智能计算平台，并同OpenAI联合训练了GPT-3模型[8]。

GTP-3含有1 750亿个超大规模参数。由于模型参数太大，研究人员无法再对它进行精调。为了满足不同的任务需求，模型需要针对不同任务提供相应的“提示语”。例如，只输入任务描述“Translate English to French: cheese=>”，GPT-3就能够直接输出翻译结果。如果在输入任务描述之后再给出一个或几个示例，那么任务完成的效果会更好。这种技术也被称为“提示学习”，并被认为是自然语言处理的一种新技术范式。

在GPT、BERT模型提出以后，各种预训练模型如雨后春笋般涌现，并从各个方面提升了预训练模型的效果，例如

更大规模的预训练模型、多语言多模态预训练模型、面向各种领域的预训练模型等，其中也包括新的预训练任务、各种预训练模型压缩与加速方法。文献[9–10]对此做了详细描述。

5 自然语言处理的未来展望

5.1 预训练模型亟待解决的关键技术问题

目前，大规模预训练模型的发展势头非常强劲。模型规模不断扩大，模型渗透的领域也在不断增加。因此，短期内自然语言处理仍将沿着大规模预训练模型的道路继续前进。不过，若要取得更好的效果并实现模型的应用落地，在开展大规模预训练模型研究时仍需要解决以下几个关键的研究问题：

（1）模型的高效性。大规模预训练模型的训练和部署都需要消耗大量的计算资源。考虑到大规模预训练模型在训练时产生的大量碳排放对环境的影响，研制计算效率更高的模型将是未来研究的重要方向。另外，还可以通过蒸馏、剪枝等技术将大模型压缩为规模更小的模型，以便于模型实现更大规模的部署应用。

（2）模型的易用性。自然语言处理任务和应用领域层出不穷。为了能够满足新任务和新领域的需求，预训练模型还需要解决小样本甚至零样本学习问题。另外，还需要构建大规模预训练模型的工程化开发能力，建设通用的开发工作流，减少专家干预及人为调整参数，构建一整套数据、代码、模型、应用程序接口（API）等服务的平台，从而支撑工业、医疗、城市、金融、物流、科学研究等领域，拓展人工智能的应用范围，并对人类生产和生活产生更广泛的影响。

（3）模型的可解释性。深度学习模型一直存在可解释性差的问题，而预训练模型也并没有解决这一问题。医疗诊断、法律判案等需要证据的应用场合仍无法直接利用该技术。即便在一些不需要提供证据的应用中，如垃圾邮件识别，模型如果能够解释自身如何是做出预测的，那么这将对提高模型的可信性大有裨益。

（4）模型的鲁棒性。虽然在很多数据集上，预训练模型已经取得不错的性能突破，在有些方面甚至已经超过人类，但有时只要测试数据稍加变动，即便语义不发生变化，之前能够被正确预测的数据也可能会获得错误的预测结果。这就是目前模型遇到的典型的鲁棒性（也称健壮性）问题，导致模型很容易被别有用心者攻击。此外，由于预训练模型极度依赖大规模未标注数据，如果所收集的数据中存在错误或陈旧的信息，甚至被人为地植入后门数据，模型可能会被误导或误用。

（5）模型的推理能力。目前预训练模型拥有的强大性能主要来自对数据的记忆能力。模型能够很容易地回答曾经见过的知识，但是对于不曾见过尤其需要多步推理才能解答的问题，往往不具有很好的解决能力。推理能力恰恰是人类解决问题的重要手段，是智能的重要体现形式，因此也是预训练模型需要重点解决的问题。

那么，自然语言处理是否会沿着预训练模型这条路一直发展下去呢？对此，本文首先分析一下人工智能的发展趋势。

5.2 自然语言处理的历史发展趋势

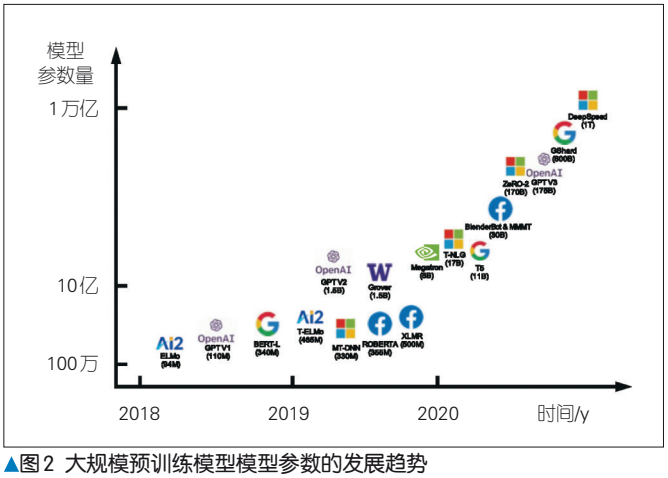
经过60余年的发展，自然语言处理经历了小规模专家知识、浅层机器学习算法、深度学习算法、基于大规模预训练模型的方法等阶段，呈现了明显的“同质化”和“规模化”两个相辅相成的发展趋势。

5.2.1 模型“同质化”的趋势明显

自然语言处理的发展呈现出明显的“同质化”趋势。早期利用专家知识的自然语言处理系统需要针对不同的任务编写不同的规则，因此不具有通用性和可移植性。后来，浅层机器学习算法需要根据不同的任务来编写特定的逻辑，以便将原始文本（也可以是声音、图像等）转换为更高级别的特征，然后使用相对“同质化”的机器学习算法（如支持向量机）进行结果预测。此后，深度学习技术能够使用更加“同质化”的模型架构（包括卷积神经网络、循环神经网络等），直接将原始文本作为学习模型的输入，并在学习的过程中自动“涌现”出用于预测的更高级别的特征。大规模预训练模型“同质化”的特性更加明显。例如，几乎所有新的自然语言处理模型都源自少数大规模预训练模型（比如BERT、RoBERTa、BART、T5等）。GPT-3模型只需要进行一次预训练就可以直接（或仅使用极少量的训练样本）完成特定的下游任务。“同质化”还体现在跨数据模态上。基于Transformer的序列建模方法和预训练模型在被成功应用于自然语言处理后，现已在图像、视频、语音、表格数据、蛋白质序列、有机分子等模态数据上取得优异的效果。这使得未来构建一个能够统一各种模态的大规模预训练模型成为可能。

5.2.2 “规模化”是智能涌现的必要条件

虽然预训练模型只是迁移学习的简单应用，但是它涌现



▲图2 大规模预训练模型模型参数的发展趋势

出了令人惊讶的“智能”。其中“规模化”是必不可少的条件。“规模化”需要的3个必要前提目前皆已成熟。

（1）计算机硬件的升级。例如，GPU吞吐量和存储容量在过去4年中增加了10倍。

（2）Transformer模型架构的发明。该模型能直接对序列中的远程依赖关系进行建模，还能充分利用硬件的并行性。

（3）更多可用的数据。过去，使用人工标注的数据进行有监督模型训练是标准的做法。然而，较高的标注成本限制了模型优势的发挥。预训练模型能够充分利用超大规模的未标注数据进行自监督学习，从而比在有限标注数据上进行训练的模型能获得更好的泛化性能。

正是由于“规模化”的重要性，越来越多的科研机构不断推出规模越来越大的预训练模型。例如，与GPT-2的15亿个参数相比，OpenAI的GPT-3模型参数规模达到了惊人的1 750亿。谷歌、微软、北京智源、华为、阿里、鹏城实验室等也相继推出了同等甚至更大规模的预训练模型，如图2所示。

5.3 自然语言处理的未来技术趋势

基于自然语言处理的历史发展趋势来判断，自然语言处理将沿着“同质化”和“规模化”的道路继续前进。

首先，以Transformer为代表的自注意力模型具有非常好的“同质化”性质。也就是说，该类模型不会对所处理的问题进行约束和限制，因此适用于自然语言、图像、语音等各类数据的处理。未来，也许会出现性能更优异的模型，但是该模型一定是更加“同质化”的。

其次，模型规模的发展速度已经远远超过摩尔定律限制的硬件发展速度。然而，无论是神经元还是它们之间连接的数量，都远远不及人脑。因此，“规模化”的发展趋势仍不会改变。期待新的能够突破摩尔定律的硬件形态的出现。

最后，人类习得语言所需的知识并非仅仅是规则、算法以及文本数据这3种类型，还包括大量其他模态的知识，如声音、视频、图像等。多模态预训练模型（如文本、图像、视频、音频之间的联合预训练）已成为目前研究的热点。此外，如要实现真正的自然语言处理，甚至通用人工智能，那么智能体就需要从物理世界中获得反馈，这样才能真正理解“冷暖”“软硬”等概念，即拥有具身的能力。另外，语言作为一种人类交流的工具，具有极强的社会属性。因此，智能体还需要与其他人进行交流，在应用中真正习得语言。在未来，自然语言处理模型一定需要融合这些更广义的知识。“同质化”和“规模化”的模型也为融合这些知识提供了必要的支撑条件。

6 结束语

在大数据、大模型和大算力的加持下，基于预训练的模型完全革新了自然语言处理的研究范式。在未来，自然语言处理，乃至整个人工智能领域，仍将沿着“同质化”和“规模化”的道路继续前进，并将融入更多的“知识”源，包括多模态数据、具身行为数据、社会交互数据等，从而实现真正的通用人工智能。

参考文献

[1] CHRISTIANE F, MILLER G A. WordNet: an electronic lexical database [M]. Cambridge: MIT Press, 1998
[2] FIRTH J R. A synopsis of linguistic theory 1930–55. [J]. Studies in linguistic analysis, 1957:1–32
[3] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3: 1137–1155
[4] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2022–02–25]. http://export.arxiv.org/pdf/1301.3781
[5] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations [EB/OL]. [2022–02–25]. https://arxiv.org/abs/1802.05365
[6] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre–training [EB/OL]. [2022–02–25]. https://s3–us–west–2. amazonaws. com/openai–assets/research–covers/language–unsupervised/language_understanding_paper.pdf
[7] DEVLIN J, CHANG M, LEE K, et al. Pre–training of deep bidirectional transformers for language understanding [EB/OL]. [2022–02–25]. https://arxiv.org/abs/1810.04805
[8] BROWM T, MANN B, RYDER N, et al. Language models are few–shot learners [EB/OL]. [2022–02–25]. https://arxiv.org/abs/2005.14165
[9] QIU X P, SUN T X, XU Y G, et al. Pre–trained models for natural language processing: a survey [J]. Science China technological sciences, 2020, 63 (10): 1872–1897. DOI: 10.1007/s11431–020–1647–3
[10] KALYAN K S, RAJASEKHARAN A, SANGEETHA S. AMMUS: a survey of transformer–based pretrained models in natural language processing [EB/OL]. [2022–02–25]. https://arxiv.org/abs/2108.05542

作　者　简　介

车万翔，哈尔滨工业大学计算学部教授、人工智能研究院副院长，教育部“青年长江学者”，斯坦福大学访问学者，现任中国中文信息学会理事、计算语言学专业委员会副主任兼秘书长，国际计算语言学学会亚太分会（AACL）执委兼秘书长，中国计算机学会高级会员；主要研究方向为自然语言处理、人机对话系统；2030“新一代人工智能”重大项目课题负责人，承担国家自然科学基金项目3项；获黑龙江省青年科技奖、谷歌专注研究奖、黑龙江省科技进步一等奖、中国中文信息学会“钱伟长”中文信息处理科学技术奖一等奖、首届汉王青年创新奖、AAAI 2013最佳论文提名奖等；发表学术论文100余篇，论文累计被引用近6 000次（Google Scholar数据），H-index值为40，出版教材4部，翻译著作2部。

刘挺，哈尔滨工业大学教授、计算学部主任兼计算机学院院长，国家“万人计划”科技创新领军人才，“十四五”国家重点研发计划先进计算与新兴软件重点专项专家组成员，教育部人工智能科技创新专家组成员，中国计算机学会会士，中国中文信息学会副理事长、社会媒体处理专委会（SMP）主任，黑龙江省中文信息处理重点实验室主任，黑龙江省“人工智能头雁”团队带头人，国家重点研发项目、国家自然科学基金重点项目负责人，多次担任国家“863”重点项目总体组专家、国家自然科学基金委评审专家，曾任国际顶级会议ACL、EMNLP领域主席；主要研究方向为人工智能、自然语言处理和社会计算；主持研制的“语言技术平台LTP”“大词林”等科研成果被业界广泛使用；曾获国家科技进步奖二等奖、黑龙江省科技进步奖一等奖、中国中文信息学会“钱伟长”中文信息处理科学技术奖一等奖等奖项。

新增编委介绍

金石

东南大学副校长、首席教授、博士生导师、教育部“长江学者奖励计划”特聘教授、国家自然科学基金杰出青年科学基金获得者、国家“万人计划”科技创新领军人才、江苏省特聘教授、中国通信学会会士、全国工程专业学位研究生教育指导委员会委员、民盟中央青年工作委员会委员、民盟江苏省委青年工作委员会副主任；长期从事移动通信的教学和研究工作，围绕蜂窝移动通信理论与关键技术、物联网理论与关键技术，以及人工智能在移动通信中的应用等领域开展研究工作，在多天线无线传输理论与关键技术、智能通信理论与方法、智能超表面无线通信等方面取得系列创新成果；研究成果获省部级科学技术奖一等奖3项、二等奖1项，IEEE通信学会莱斯奖，IEEE信号处理学会青年作者最佳论文奖，《China Communications》最佳论文奖，《Electronics Letters》最佳论文奖，《National Science Review》最佳论文奖，《Journal of Communications and Information Networks》青年作者最佳论文奖，以及IEEE ICC/GLOBECOM等10余个国际重要学术会议最佳论文奖，2014年至今连续入选爱思唯尔中国高被引学者，2019年至今连续入选科睿唯安全球高被引学者；共发表学术论文400余篇，授权国际/国家发明专利50余件，出版专著2部、教材1本。