自然语言处理技术发展

Development of Natural Language Processing Technology

王海宁/WANG Haining
（英特尔（中国）有限公司，中国 北京 100013）
(Intel China Ltd. , Beijing 100013, China)

DOI：10.12142/ZTETJ.202202009
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20220408.1420.004.html
网络出版日期：2022-04-08
收稿日期：2022-02-26

摘要：基于神经网络和深度学习的预训练语言模型为自然语言处理技术带来了突破性发展。基于自注意力机制的Transformer模型是预训练语言模型的基础。GPT、BERT、XLNet等大规模预训练语言模型均基于Transformer模型进行堆叠和优化。认为目前依赖强大算力和海量数据的大规模预训练语言模型存在实用问题，指出轻量预训练语言模型是未来重要的发展方向。

关键词：自然语言处理；预训练语言模型；Transformer；GPT；BERT；XLNet；模型优化

Abstract: The pre-trained language model based on neural network and deep learning has brought breakthrough development for natural language processing technology. The Transformer model based on self-attention mechanism is the basis of the pre-trained language model. Large-scale pre-trained language models such as GPT, BERT, XLNet, etc. are based on the Transformer model or its optimization. However, the current large-scale pre-training language models that rely on powerful computing resources and massive data have practical problems. It is pointed out that lightweight pre-trained language models are an important development direction in the future.

Keywords: natural language processing; pre-trained language model; Transformer; GPT; BERT; XLNet; model optimization

自然语言处理（NLP）是基于自然语言理解和自然语言生成的信息处理技术[1]。这里的自然语言是指任何一种人类语言，例如中文、英语、西班牙语等，并不包括形式语言（如Java、Fortran、C++等）。

自然语言处理的历史可以追溯到17世纪。那时莱布尼茨等哲学家对跨越不同语言的通用字符进行探索[2]，认为人类思想可以被归约为基于通用字符的运算。虽然这一观点在当时还只是理论上的，但却为自然语言处理技术的发展奠定了基础。

作为人工智能的一个重要领域，当代自然语言处理技术与人工智能技术的兴起和发展是一致的。1950年，图灵提出了著名的基于人机对话衡量机器智能程度的图灵测试[3]。这不仅是人工智能领域的开端，也被普遍认为是自然语言处理技术的开端。20世纪50年代至90年代，早期自然语言处理领域的发展主要基于规则和专家系统，即通过专家从语言学角度分析自然语言的结构规则，来达到处理自然语言的目的。

从20世纪90年代起，伴随着计算机运算速度、存储容量的快速发展，以及统计学习方法的成熟，研究人员开始使用统计机器学习方法来处理自然语言任务。然而，此时自然语言的特征提取仍然依赖人工，同时受限于各领域经验知识的积累。

深度学习算法于2006年被提出之后，不仅在图像识别领域取得了惊人的成绩，也在自然语言处理领域得到了广泛应用。不同于图像的标注，自然语言的标注领域众多并具有很强的主观性。因此，自然语言处理领域不容易获得足够多的标注数据，难以满足深度学习模型训练对大规模标注数据的需求。

近年来，GPT[4]、BERT[5]等预训练语言模型可以很好地解决上述问题。基于预训练语言模型的方法本质上是一种迁移学习方法，即通过在容易获取、无需人工标注的大规模文本数据基础上依靠强大算力进行预先训练，来获得通用的语言模型和表示形式，然后在目标自然语言处理任务上结合任务语料对预训练得到的模型进行微调，从而在各种下游自然语言处理任务中快速收敛以提升准确率。因此，预训练语言模型自面世以来就得到了迅速发展和广泛应用，并成为当前各类自然语言处理任务的核心技术。

1 语言表示的发展

自然语言处理涉及众多任务。从流水线的角度上看，我们可以将这些任务划分为3类：完成自然语言处理之前的语言学知识建设和语料库准备任务；对语料库开展分词、词性标注、句法分析、语义分析等基本处理任务；利用自然语言处理结果完成特定目标的应用任务，如信息抽取、情感分

析、机器翻译、对话系统、意图识别等。其中，将自然语言转变为计算机可以存储和处理的形式（即文本的表示）是后续各类下游自然语言处理任务的基础和关键。

字符串是最基本的文本表示方式，即符号表示。这种表示方式主要应用在早期基于规则的自然语言处理方式中。例如，基于预定义的规则对句子进行情感分析：当出现褒义词时，句子表达正向情感；当出现贬义词时，句子表达负向情感。显然，这种使用规则的方式只能对简单的语言进行分析处理，在遇到矛盾的情况下系统很可能无法给出正确的结论。

以向量的形式表示词语，即词向量，是广泛应用于目前自然语言处理技术中的表示方式。词向量的表示有多种方式。其中，最简单的是基于词出现次数统计的独热表示和词袋表示。这类表示方式的主要缺点在于，不同的词需要用完全不同的向量来表示，维度高并且缺乏语义信息的关联，同时存在数据稀疏问题。

另外一大类词向量表示是基于分布式语义假设（上下文相似的词，其语义也相似）的分布式表示。这种词向量表示具体又可以分为3类：

（1）基于矩阵的词向量表示。该方法基于词共现频次构建体现词与上下文关系的（词–上下文）矩阵。矩阵每行表示一个词向量 $\boldsymbol{w}_i$。第 j 个元素 $\boldsymbol{w}_{ij}$ 的取值可以是 $\boldsymbol{w}_i$ 与上下文的共现次数，也可以由基于其共现概率进行的点互信息（PMI）、词频–逆文档频率（TF–IDF）、奇异值分解（SVD）等数学处理来获得。这种方法更好地体现了高阶语义相关性，可解决高频词误导计算等问题。其中，上下文可以是整个文档，也可以是每个词。此外，我们也可以选取 $\boldsymbol{w}_i$ 附近的 N 个词作为一个 N 元词窗口。

（2）基于聚类的词向量表示。这类方法通过聚类手段构建词与上下文之间的关系。例如，布朗聚类是一种基于N–gram模型和马尔可夫链模型的自底向上的分层聚类算法。在这种算法中，每个词都在且仅在唯一的一个类中。在初始的时候，每个词均被独立分成一类，然后系统将其中的两类进行合并，使得合并之后的评价函数（用以评估 n 个连续的词序列能否组成一句话的概率）达到最大值。系统将不断重复上述过程，直至获得期望的类数量为止。

（3）基于神经网络的语言模型，也称为词嵌入表示。这类方法将词向量中的元素值作为模型参数，采用神经网络结合训练数据学习的方式来获得语言模型参数值。基于神经网络的语言模型具体又包括静态语言模型和动态语言模型。这两种语言模型的区别在于：静态语言模型通过一个给定的语料库得到固定的表示，不随上下文的变化而变化，例如Word2vec、GloVe和Fasttext模型；动态语言模型由上下文计算得到，并且随上下文的变化而变化，例如CoVe、ELMo、GPT和BERT模型。其中，基于神经网络的语言模型充分利用了文本天然的有序性和词共现信息的优势，无需人工标注也能够通过自监督学习从文本中获取语义表示信息，是预训练语言模型的重要基础，也是目前词表示研究与应用的热点。

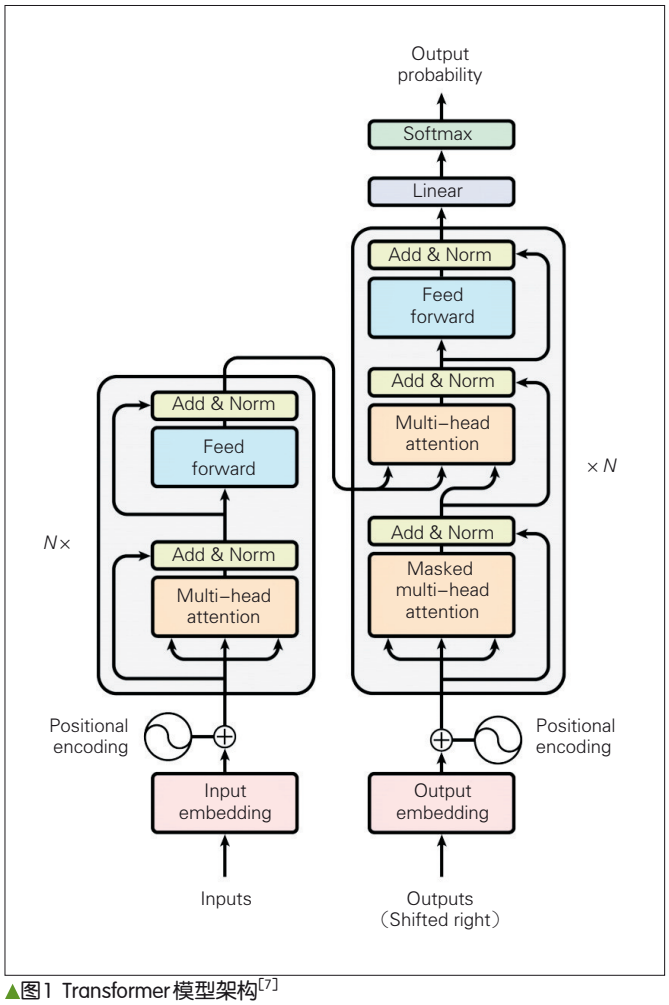
2 预训练语言模型

2.1 预训练语言模型基础

2003年，Y. BENGIO首次提出神经网络语言模型[6]。2017年之前，在进行自然语言处理时人们常用多层感知机（MLP）、卷积神经网络（CNN）和循环神经网络（RNN），包括长短期记忆（LSTM）网络，来构建神经网络语言模型。由于每层都使用全连接方式，MLP难以捕捉局部信息。CNN采用一个或多个卷积核依次对局部输入序列进行卷积处理，可以比较好地提取局部特征。由于适用于高并发场景，较大规模的CNN模型经过训练后可以提取更多的局部特征。然而，CNN却难以捕获远距离特征。RNN将当前时刻网络隐含层的输入作为下一时刻的输入。每个时刻的输入经过层次递归后均对最终输出产生影响，这就像网络有了历史记忆一样。RNN可以解决时序问题和序列到序列问题，但是这种按照时序来处理输入的方式使得RNN很难充分利用并行算力来加速训练。LSTM是一种特殊的RNN，它对隐含层进行跨越连接，减少了网络的层数，从而更容易被优化。

2017年，来自谷歌的几位工程师在不使用传统CNN、RNN等模型的情况下，完全采用基于自注意力机制的Transformer模型，取得了非常好的效果[7]。在解决序列到序列问题的过程中，他们不仅考虑前一个时刻的影响，还考虑目标输出与输入句子中哪些词更相关，并对输入信息进行加权处理，从而突出重要特征对输出的影响。这种对强相关性的关注就是注意力机制。Transformer模型是一个基于多头自注意力机制的基础模型，不依赖顺序建模就可以充分利用并行算力处理。在构建大模型时，Transformer模型在训练速度和长距离建模方面都优于传统的神经网络模型。因此，近年来流行的GPT、BERT等若干超大规模预训练语言模型基本上都是基于Transformer模型构建的。Transformer模型整体架构如图1所示。

自注意力机制的本质是学习序列中的上下文相关程度和深层语义信息。然而，随着输入序列长度的增加，学习效率会降低。为了更好地处理长文本序列，Transformer模型又衍



▲图1 Transformer模型架构[7]

生出一些“变种”，例如Transformer-XL[8]。Transformer-XL采用段级循环和相对位置编码的优化策略，将Transformer中固定长度的输入片段进一步联系起来，具备更强的长文本处理能力。

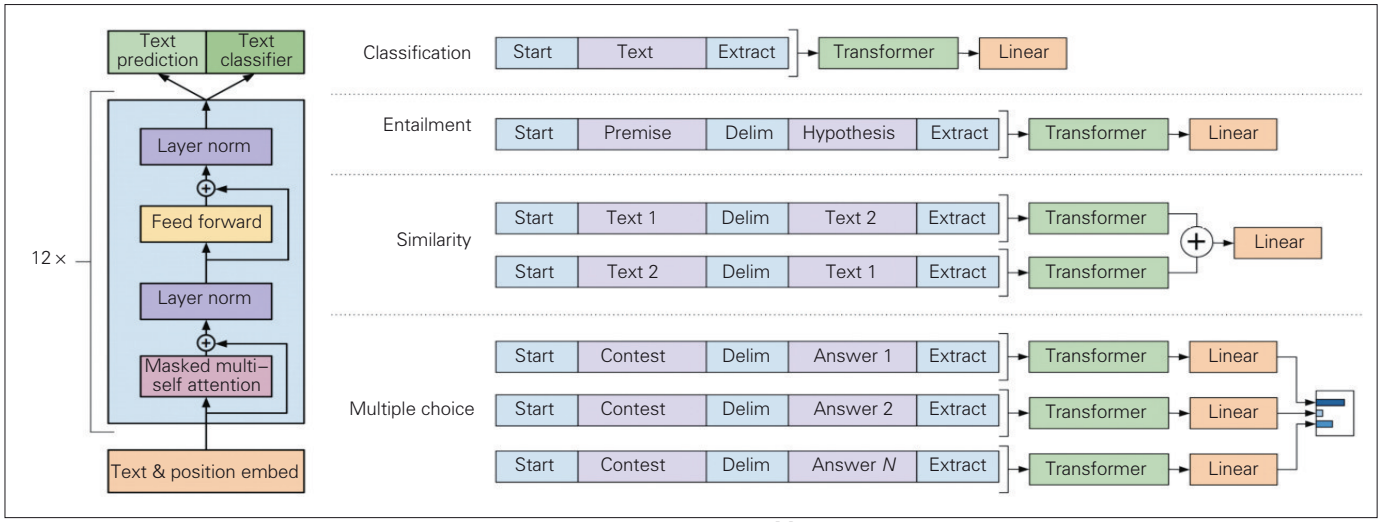
2.2 大规模预训练语言模型

广义预训练语言模型泛指经过提前训练得到的语言模型。各类神经网络语言模型在理论上都可以做预训练处理。而目前自然语言处理领域常涉及的预训练语言模型，通常是指一些参数数量过亿甚至超千亿的大规模语言模型。这些模型的训练依赖强大算力和海量数据。典型的大规模预训练语言模型包括GPT系列、BERT、XLNet等。此外，这些模型的各种改进模型也层出不穷。

2.2.1 GPT系列

2018年6月，OpenAI公司提出初代GPT模型[4]，开启了具有“基于大量文本学习高容量语言模型”和“对不同任务使用标注数据来进行微调”两个阶段的自然语言处理预训练模型大门。GPT模型基于12层Transformer基础模型构建了单向解码器，约有1.17亿个参数。具体解码器结构、训练目标和针对不同下游任务的输入转换如图2所示。

OpenAI公司在2019年2月进一步提出GPT模型的升级版本，即GPT-2[9]。由于担心该技术可能会被恶意利用，研究团队并没有对外发布预训练好的GPT-2模型，而是发布了一个小规模模型。GPT-2保留了GPT的网络结构，直接进行规模扩张，即堆叠更多层的Transformer模型，并使用10倍于GPT模型的数据集进行训练，参数数量超过15亿。随着规模的增加，GPT-2也获得了更好的泛化功能，包括生成前所未有的高质量合成文本功能。虽然在部分下游任务上尚未超过当时的最优水平，但是GPT-2证明了大规模预训练词向量模型在迁移到下游任务时，可以超越使用特定领域数



▲图2 Transformer解码器结构和训练目标(左)及针对不同下游任务的输入转换(右)[4]

据进行训练的语言模型，并且在拥有大量（未标注）数据和具备足够算力时，使下游任务受益于无监督学习技术。

GPT-3[10]模型于2020年5月被提出，是目前最强大的预训练语言模型之一。GPT-3在GPT-2的基础上进一步进行了规模扩张，使用高达45 TB的数据进行训练，参数数量高达1 750亿。正是这样巨大的网络规模，才使得GPT-3模型在不进行任何微调的情况下，可以仅利用小样本甚至零样本就能在众多下游任务中超越其他模型。OpenAI公司虽未开源GPT-3模型，但是提供了多种应用程序接口（API）服务以供下游任务调用。

2.2.2　BERT

BERT[5]是由谷歌公司于2018年10月提出的。与单向的GPT模型不同，BERT基于Transformer模型构建了多层双向编码器。

BERT模型包括两个训练任务：一个是掩码语言模型（MLM），另一个是下一句预测（NSP）。MLM可以很好地解决双向建模时逆序信息泄露的问题；NSP则可以很好地理解两段文本之间的关系，适用于完成阅读理解或文本蕴含类任务。BERT的每个下游任务都采用相同的预训练模型架构并使用预训练模型的参数来进行初始化。BERT的预训练和微调过程如图3所示。

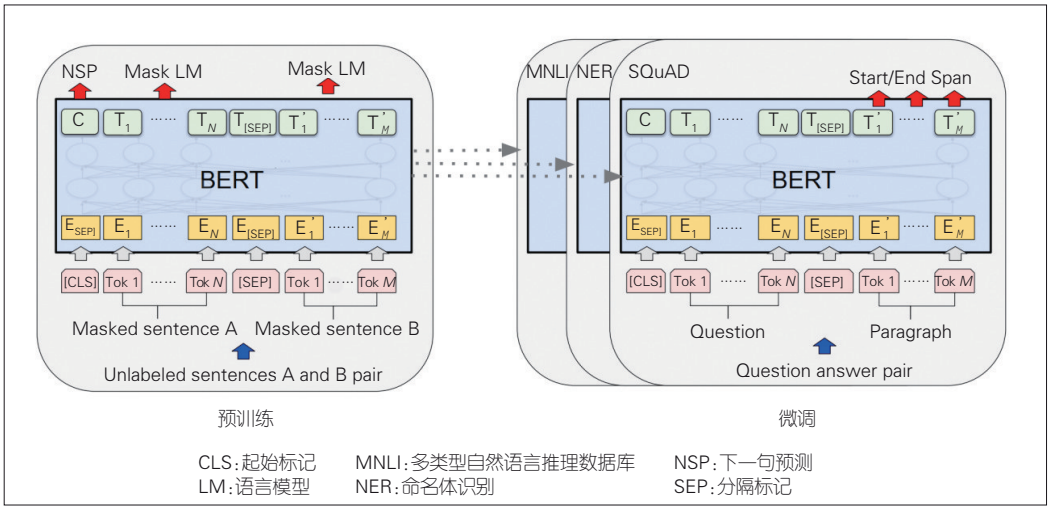
BERT的设计团队按照模型规模的大小将BERT分为含有1.1亿个参数的$\text{BERT}_{\text{BASE}}$和含有3.4亿个参数的$\text{BERT}_{\text{LARGE}}$，并与其他模型（包括GPT）进行对比。对比结果表明，BERT模型在GLUE[11]、SQuAD[12]、SWAG[13]的11项NLP任务评估中全面刷新了最佳成绩纪录，甚至在SQuAD测试中超越了人类。

BERT模型是近年来NLP领域发展的一大里程碑。BERT陆续衍生出了许多优化的模型。例如，显著增强了长文本理解能力的XLNet[14]、占用更少存储空间的ALBERT[15]、具备更强大文本生成能力的BART[16]、能够学习视频知识的VideoBERT[17]等。这些模型推动了NLP的快速发展。

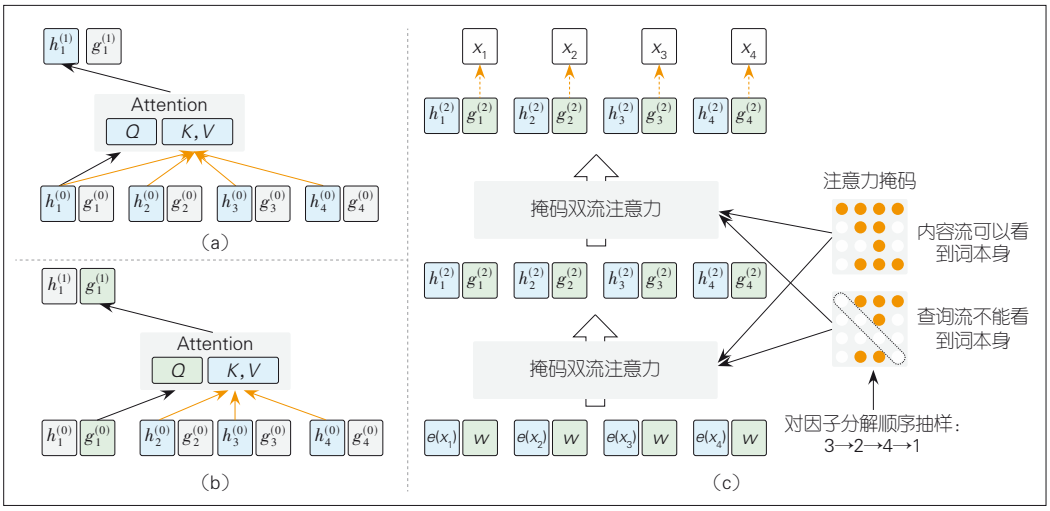
2.2.3　XLNet

由于在预训练的输入数据中人为地引入了掩码，BERT模型忽略了被掩码信息之间的依赖性。这将导致预训练数据与微调阶段使用的真实数据之间产生微小差异。针对上述问题，卡内基梅隆大学和谷歌公司于2019年6月进一步提出了一种基于Transformer-XL的自回归语言模型，即XLNet模型[14]。

通过置换语言建模（PLM），XLNet对序列中输入信息



▲图3　BERT的预训练和微调过程[5]



▲图4　XLNet双流自注意力机制[14]

进行排列重组，可实现双向上下文的建模，并形成双流自注意力机制，以解决由PLM重新排列所引入的位置信息混淆问题。如图4所示，XLNet用内容流和查询流两种不同的掩码矩阵来进行预测。其中，内容流用于保留词的语义信息，可以看到词本身；查询流不能看到词本身，用于保留词的位置信息，仅在预训练阶段使用。

此外，由于XLNet使用Transformer-XL来替代Transformer，并将其作为特征提取器，因此XLNet拥有比BERT更强的长文本理解能力。

3 预训练语言模型优化方向

预训练语言模型在各类NLP任务中的效果是显而易见的。随着参数规模的扩大和训练数据的增加，预训练语言模型可以获得更好的准确性和泛化性。然而，这是以巨大算力支持为前提的，只有少数大公司才能够承担起这种高昂的算力成本。这个问题在GPT-3模型的研发过程中表现得尤为突出。据报道，为了训练GPT-3模型，微软在Azure云上构建了一个包含1万个GPU、28.5万个CPU内核和400 Gbit/s网络连接的超级计算系统。其中，GPT-3训练一次的费用约为460万美元[18-19]。在这种情况下，进一步发现、验证和解决模型的潜在问题都非常困难。对此，微软研发团队也认为，当系统出现Bug时，他们也无法对模型进行再训练。

相应地，预训练模型在应用时也需要较大算力和内存支持，往往需要多块高端人工智能芯片或者服务器集群来支撑模型的部署。为了降低预训练模型的部署门槛，业界往往采用量化、剪枝、蒸馏等方法对模型进行压缩，以形成更加轻量化的预训练模型。

（1）量化是指将模型参数转换为更少比特数来存储和运算，即将模型的精度降低。虽然量化损失了一定的精度，但是它在可接受的准确率范围内能大大提升模型的训练和推理速度。例如，BF16是一种专为加速深度学习训练而设计的16位数字精度格式，在保留FP32（32位浮点数）指数位数的同时减少了16位尾数位。将模型参数从FP32转换为BF16后，模型可以在维持相近准确率的同时实现训练速度的数倍提升[20-21]。

（2）剪枝是指去掉模型参数中冗余或者不重要的部分，即减少模型参数。具体来说，剪枝包括元素剪枝和结构剪枝两种方式。其中，元素剪枝是指去掉单个绝对值过小或者对模型影响过小的参数；结构剪枝是指去掉整块模型结构，例如减少多头注意力的数量，或者减少堆叠的Transformer块数量等。

（3）蒸馏是指较小规模的模型（称为学生模型）从较大规模的模型（称为教师模型）中学习知识，并替代学生模型从训练数据中学习知识的过程。典型的蒸馏模型包括DistilBERT[22]、TinyBERT[23]、MobileBERT[24]等。这些模型与$BERT_{BASE}$模型的对比如表1所示。

▼表1 蒸馏模型效果对比

模型	参数量对比/%	推理速度对比	GLUE性能对比/%
$BERT_{BASE}$	100	1	100
DistilBERT	60.0	1.6	97.0
$TinyBERT_4$（4层）	13.3	9.4	96.8
MobileBERT	23.3	5.5	99.2

在上述优化方法中，量化和剪枝是比较常用的方法。此外，还有其他比较成熟的优化工具，例如TensorFlow Model Optimization、TensorFlow Lite、TensorRT、OpenVINO、PaddleSlim等。由于蒸馏的压缩比更大，它可以和量化、剪枝叠加使用。

4 结束语

自然语言处理技术经历了近百年的发展。机器翻译、智能客服、信息检索与过滤、情感分析和文本生成等，在教育、医疗、司法、互联网等行业中得到了广泛的应用。近年来，预训练语言模型的提出和算力的快速提升，将自然语言处理技术的发展推向了新的高度，使自然语言处理技术在某些领域达到甚至超越了人类水平。然而，目前大规模预训练语言模型仍需要极大的算力支持，训练模型所需的成本仍然较高，能源消耗和碳排放也并不经济，距离落地应用尚有距离。因此，研发出更加轻量的预训练语言模型，是未来重要的发展方向。

参考文献

[1] ISO/IEC. Information technology—artificial intelligence—artificial intelligence concepts and terminology: ISO/IEC TR 24372:2021(E) [S]. 2021
[2] 段德智. 莱布尼茨语言哲学的理性主义实质及其历史地位研究 [J]. 武汉大学学报(人文科学版), 2013, 66(5): 54-63
[3] TURING A M. Computing machinery and intelligence [J]. Mind, 1950, 49: 433-460. DOI: 10.1093/mind/lix.236.433
[4] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training [EB/OL]. [2022-02-25]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper.pdf
[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-02-25]. https://aclanthology.org/N19-1423.pdf
[6] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3: 1137-1155
[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 6000-6010

[8] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context [EB/OL]. [2022-02-25]. https://arxiv.org/abs/1901.02860v3. DOI: 10.18653/v1/p19-1285
[9] RADFORD A, JEFFREY W, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2022-02-25]. https://cdn. openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[10] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022-02-25]. https://arxiv.org/abs/2005.14165
[11] WANG A, SINGH A, MICHAEL J, et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding [C]//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, 2018: 353-355. DOI: 10.18653/v1/w18-5446
[12] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100 000+ questions for machine comprehension of text [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2016: 2383-2392. DOI: 10.18653/v1/d16-1264
[13] ZELLERS R, BISK Y, SCHWARTZ R, et al. SWAG: a large-scale adversarial dataset for grounded commonsense inference [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018: 93-104. DOI: 10.18653/v1/d18-1009
[14] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding [EB/OL]. [2022-02-25]. https://arxiv.org/abs/1906.08237
[15] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations [EB/OL]. [2022-02-25]. https://arxiv.org/abs/1909.11942
[16] LEWIS M, LIU Y H, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [EB/OL]. [2022-02-25]. https://arxiv.org/abs/1910.13461
[17] SUN C, MYERS A, VONDRICK C, et al. VideoBERT: a joint model for video and language representation learning [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 7463-7472. DOI: 10.1109/ICCV.2019.00756
[18] DEUTSCHER M. OpenAI makes GPT-3 more broadly available to developers [EB/OL]. [2022-02-25]. https://siliconangle. com/2021/11/18/openai-makes-gpt-3-broadly-available-developers/
[19] DICKSON B. The untold story of GPT-3 is the transformation of OpenAI [EB/OL]. [2022-02-25]. https://bdtechtalks. com/2020/08/17/openai-gpt-3-commercial-ai/#: ~: text=According% 20to% 20one% 20estimate% 2C% 20training% 20GPT-3%20would% 20cost, tuning% 20that% 20would% 20probably%20increase%20the%20cost%20several-fold
[20] HENRY G, TANG P T P, HEINECKE A. Leveraging the bfloat16 artificial intelligence datatype for higher-precision computations [C]//Proceedings of 2019 IEEE 26th Symposium on Computer Arithmetic. IEEE, 2019: 69-76. DOI: 10.1109/ARITH.2019.00019
[21] Intel. Code sample: Intel® deep learning boost new deep learning instruction bfloat16 – intrinsic functions [EB/OL]. [2022-02-25]. https://www. intel. cn/content/www/cn/zh/developer/articles/technical/intel-deep-learning-boost-new-instruction-bfloat16.html?wapkw=BF16
[22] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [EB/OL]. [2022-02-25]. https://arxiv.org/abs/1910.01108
[23] JIAO X Q, YIN Y C, SHANG L F, et al. TinyBERT: distilling BERT for natural language understanding [C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020: 4163-4174. DOI: 10.18653/v1/2020.findings-emnlp.372
[24] SUN Z Q, YU H K, SONG X D, et al. MobileBERT: a compact task-agnostic BERT for resource-limited devices [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 2158-2170. DOI: 10.18653/v1/2020.acl-main.195

作 者 简 介

王海宁，英特尔（中国）有限公司人工智能技术政策和标准总监、中关村高端领军人才、正高级工程师、北京邮电大学兼职教授，担任ETSI ISG ENI副主席、CCSA SP1 NFV特设项目组副组长、CCSA TC610网络人工智能应用工作组组长等职务；主要研究方向为4G/5G网络技术、SDN/NFV、人工智能；2017年获北京市委组织部青年骨干个人项目资助；主持编制数十项国际标准和行业标准，发表文章多篇，拥有授权专利30余项。