计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2021/15(08)-1359-31
doi: 10.3778/j.issn.1673-9418.2012109

开放科学(OSID)

自然语言处理预训练技术综述

陈德光[1],马金林[1,3+],马自萍[2],周　洁[1]
1. 北方民族大学 计算机科学与工程学院,银川 750021
2. 北方民族大学 数学与信息科学学院,银川 750021
3. 图像图形智能处理国家民委重点实验室,银川 750021
+ 通信作者 E-mail: 624160@163.com

摘　要:在目前已发表的自然语言处理预训练技术综述中,大多数文章仅介绍神经网络预训练技术或者极简单介绍传统预训练技术,存在人为割裂自然语言预训练发展历程。为此,以自然语言预训练发展历程为主线,从以下四方面展开工作:首先,依据预训练技术更新路线,介绍了传统自然语言预训练技术与神经网络预训练技术,并对相关技术特点进行分析、比较,从中归纳出自然语言处理技术的发展脉络与趋势;其次,主要从两方面介绍了基于BERT改进的自然语言处理模型,并对这些模型从预训练机制、优缺点、性能等方面进行总结;再者,对自然语言处理的主要应用领域发展进行了介绍,并阐述了自然语言处理目前面临的挑战与相应解决办法;最后,总结工作,预测了自然语言处理的未来发展方向。旨在帮助科研工作者更全面地了解自然语言预训练技术发展历程,继而为新模型、新预训练方法的提出提供一定思路。

关键词:预训练技术;自然语言处理;神经网络
文献标志码:A　**中图分类号:**TP18

Review of Pre-training Techniques for Natural Language Processing

CHEN Deguang[1], MA Jinlin[1,3+], MA Ziping[2], ZHOU Jie[1]
1. School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China
2. School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China
3. Key Laboratory for Intelligent Processing of Computer Images and Graphics of National Ethnic Affairs Commission of the PRC, Yinchuan 750021, China

Abstract: In the published reviews of natural language pre-training technology, most literatures only elaborate neural network pre-training technologies or a brief introduction to traditional pre-training technologies, which may result in the development process of natural language pre-training dissected artificially from natural language processing.

基金项目:北方民族大学中央高校基本科研业务费专项(2021KJCX09);国家自然科学基金(61462002);国家民委“图像与智能信息处理”创新团队项目;北方民族大学“计算机视觉与虚拟现实”创新团队项目;北方民族大学重大专项(ZDZX201801);宁夏高等学校一流学科建设(数学)项目(NXYLXK2017B09)。

This work was supported by the Basic Scientific Research in Central Universities of North Minzu University (2021KJCX09), the National Natural Science Foundation of China (61462002), the Project of “Image and Intelligent Information Processing” Innovation Team of National Ethnic Affairs Commission of the PRC, the Project of “Computer Vision and Virtual Reality” Innovation Team of North Minzu University, the Major Special Projects of North Minzu University (ZDZX201801), and the First-Class Disciplines Foundation of Ningxia (Mathematics Discipline) (NXYLXK2017B09).

收稿日期:2020-12-29　**修回日期:**2021-04-12

Therefore, in order to avoid this phenomenon, this paper covers the process of natural language pre-training with four points as follows. Firstly, the traditional natural language pre-training technologies and neural network pre-training technologies are introduced according to the updating route of pre-training technology. With the characteristics of related technologies analyzed, compared, this paper sums up the process of development context and trend of natural language processing technology. Secondly, based on the improved BERT (bidirectional encoder representation from transformers), this paper mainly introduces the latest natural language processing models from two aspects and sums up these models from pre-training mechanism, advantages and disadvantages, performance and so on. The main application fields of natural language processing are presented. Furthermore, this paper explores the challenges and corresponding solutions to natural language processing models. Finally, this paper summarizes the work of this paper and prospects the future development direction, which can help researchers understand the development of pre-training technologies of natural language more comprehensively and provide some ideas to design new models and new pre-training methods.

Key words: pre-training techniques; natural language processing; neural network

自然语言处理预训练在不同时期有不同的称谓,但是,本质是使用大量语料预测相应单词或词组,生成一个半成品用以训练后续任务。自然语言预处理是预训练以及后续任务训练的一部分,用以将人类识别的语言转化为机器识别的语言,目的是辅助提高模型性能。自神经网络预训练技术用于自然语言处理以来,自然语言处理取得了重大发展,以是否采用神经网络为依据,学者们将自然语言处理技术划分为基于传统的自然语言处理和基于神经网络的自然语言处理。马尔可夫[1]与香农[2]语言建模实验的成功,叩响了传统自然语言处理的大门。在传统的自然语言处理中依据处理方法分为基于规则的自然语言处理和基于统计的自然语言处理。在20世纪50年代中期以前,学者们普遍使用的是基于简单统计的自然语言处理。1957年乔姆斯基出版了《句法结构》[3]一书,该书对语料库的语料不充分性提出了质疑,并提出基于规则的自然语言处理,促使基于规则的自然语言处理逐渐占据大量市场。20世纪80年代以后,人们发现规则不但不能穷举,而且规则之间会出现一定冲突,因此,大量研究人员转向基于统计的自然语言处理。在此之后,大多数研究者在这两种方法间寻求突破,直至神经网络蓬勃发展,研究热点才由传统的自然语言处理转变为基于神经网络的自然语言处理。

目前,基于神经网络的预训练技术综述相对较多[4-6],但是对于传统预训练技术大多未涉及或者一笔带过,存在人为割离自然语言预训练发展脉络,不利于自然语言处理技术的发展。神经网络预训练技术是在预训练阶段采用神经网络模型进行预训练的技术统称,由于预训练与后续任务耦合性不强,能单独成为一个模型,因此也称为预训练语言模型,这一称谓是区别于传统预训练技术的叫法。传统预训练技术与模型耦合较为紧密,该技术与模型之间并没有明确的区分界限,为了方便阐述,将语料送入模型到生成词向量的这一过程称为传统预训练技术。当然,无论是神经网络预训练技术还是传统预训练技术均需要对语料进行预处理,具体来说,预处理就是将原始语料进行清洗(包括去除空白、去除无效标签、去除符号以及停顿词、文档切分、基本纠错、编码转化等操作)、分词(对于中文类似的独立语才有)和标准化等操作,从而将语料转化为机器可识别的语言,图1为自然语言预处理流程。

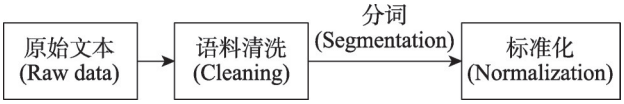


Fig.1 Preprocessing process

图1 预处理流程

由于自然语言处理涉及文本、语音、视频、图像等不同类型语料,概念较为宽泛。本文介绍自然语言处理预训练技术,而该技术主要包含文本类语料,因而本文基于文本语料阐述。具体来说,首先从传统预训练技术与神经网络预训练技术两方面进行阐述,介绍了预训练的基本方法和各方法的优缺点及

适用范围；而后，针对目前流行的BERT（bidirectional encoder representation from transformers）模型及相关技术从预训练方式、预训练优缺点、模型性能等方面进行较为详细的讨论比较；在此之后，对自然语言处理的重点应用领域的进展进行介绍；再者，阐述了目前自然语言处理面临的挑战与解决办法；最后，对本文工作进行了总结与展望。总体来说，相对于其他预训练技术综述，本文做了以下创新：（1）弥补了缺少传统预训练技术的短板，从而将传统预训练技术与神经网络预训练技术进行连贯；（2）相对于其他对预训练技术介绍不全与多种分类混用的文章，本文采用一种分类标准较为全面地展示了自然语言预训练技术；（3）对目前较为流行的BERT改进预训练技术和重点热点领域模型进行详细介绍，帮助科研工作者了解自然语言预训练及模型发展动态。

1 传统预训练技术

就目前已发表的大多数自然语言处理预训练文章来看，少有文章对传统预训练技术进行较为详细的介绍，究其原因，可能由以下两点造成：其一，在传统的自然语言处理中，预训练技术与模型具有强烈的耦合性，没有独立可分的预训练技术；其二，神经网络，尤其是深度神经网络的发展，导致研究者们对传统自然语言处理技术（包括传统预训练技术与传统模型）重视不够。但是，传统自然语言处理技术（包括传统预训练技术和传统模型）作为自然语言处理的历史阶段产物，曾在推动自然语言发展过程中发挥过重大作用，因而有必要对传统预训练技术进行较为详细的介绍。

传统自然语言处理过程讲究针对性和技巧性，因而传统预训练技术与相应模型耦合较为紧密。为了方便阐述，将自然语料的特征工程及之前部分称为传统预训练技术，特征工程是指将语料进行初步特征提取的过程。

1.1 *N*-gram技术

N-gram技术是一种基于统计的语言模型技术，它基于第N个词仅与其前面$N-1$个词相关的理想假设。基本思想为：将原始语料预处理后，按照字节大小为N的滑动窗口进行滑动操作，进而形成长度为N的字节片段序列组[7-8]。

在该技术中，每个字节即为一个gram，对语料中所有gram出现的频度进行统计，并设置相应阈值进行过滤，除去一些低频度和不必要的单词从而形成gram列表，这个gram列表就是文本语料的向量特征空间，而列表中的每一种gram都是一个特征向量，即为预训练的结果。在实际使用中，假设gram列表中有V个有效词，采用N-gram方法，则复杂度为$O(V^N)$，表明随着N值的增加复杂度会显著增大，因此在通常情况下采用Unigram方法。Huang等[9]的实验表明，Unigram的性能在同等情况下高于Bi-gram和Tri-gram。Unigram、Bi-gram和Tri-gram的公式如下：

$$\begin{cases} P(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}P(w_i) \\ P(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}P(w_i|w_{i-1}) \\ P(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}P(w_i|w_{i-2}w_{i-1}) \end{cases} \tag{1}$$

实际应用中，该预训练技术简单易行，具有完备的理论性和极强的操作性，在传统的统计自然语言处理中占有重要地位。适合于单词预测、拼写检查、词法纠错、热词选取等词级自然语言处理领域和自动索引、无分割符语言文本切分等句子级自然语言处理领域。但是，由于语料的限制，*N*-gram具有如下缺点：首先，统计词的准确度存在不完整性，需要多次试探才能确定阈值且高频词汇不一定为有效词；其次，由于*N*-gram的强烈假设性，导致结果存在一定的不合理性；最后，大规模语料统计方法与有限训练语料之间可能产生数据稀疏问题，为此，常用的解决办法有拉普拉斯平滑（Laplace smoothing）、古德-图灵平滑（good-turing smoothing）、插值平滑（interpolation smoothing）、克内尔平滑（Kneser-Ney smoothing）等方法。

1.2 向量空间模型技术

向量空间模型（vector space model，VSM）于20世纪70年代被提出，是一种文档表示和相似性计算工具[10]。主要思想为：用空间向量的形式表示语料库中的所有语料，语料的每个特征词对应语料向量的每一维。具体来说，该技术包含文本预处理、特征选择与特征计算、算法计算准确度等几个主要步骤，图2为VSM的流程表示。向量空间模型的文本表示为词袋模型（bag-of-word），由于本文介绍的是预训练技术，本节重点介绍特征工程（特征选择与特征计算）的相关技术。

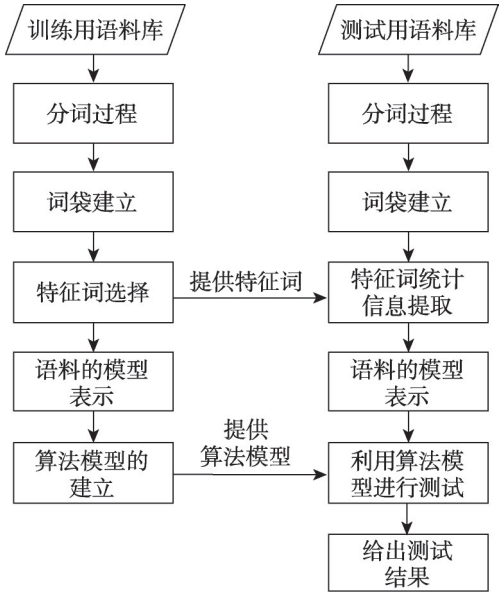


Fig.2 VSM process

图2 VSM流程

1.2.1 独热码技术

自Huffman[11]提出独热码(One-hot)机制后,该技术就逐步应用于自然语言处理中。独热码技术的本质是采用N位状态寄存器(status register)对N个状态进行编码,每个状态都有对应且独立的寄存器位,而在任意时刻都只有一位状态有效。

对于具有离散特征的取值,比如分类模型、推荐系统及垃圾邮件过滤等,通过使用独热编码技术将整个取值扩展到欧氏空间,而常用的距离计算或者相似度计算均基于欧氏空间,这使特征之间的距离计算更加合理。

独热编码技术简单有效,易于理解,将语料进行数值化表示能在一定程度上起到扩充特征的作用。该技术适合于基于较多参数与距离的模型,例如支持向量机(support vector machine,SVM)[12]、神经网络(neural network,NN)[13]、最近邻算法(*K*-nearest neighbor,*K*NN)[14]等。但是,该技术具有明显缺点:首先,每个单词的编码维度是整个词汇表的大小,存在维度过大导致编码稀疏问题,使计算代价变大;其次,独热编码存在单词间相互独立的强制性假设,这种关系导致该方法无法体现单词间的远近程度,从而丢失了位置信息。

1.2.2 TF-IDF技术

TF-IDF(term frequency-inverse document frequency)[15-16]是信息检索领域非常重要的搜索词重要性度量技术(其前身是TF方法与IDF方法),用以衡量一个词w对于查询(Query,可以看作文本文档)所能提供信息的重要度。计算过程如下:

词频(term frequency,TF)表示关键词w在文档D_i中出现的频率,公式为:

$$TF_{w,D_i}=\frac{count(w)}{|D_i|} \tag{2}$$

式中,$count(w)$为关键词w在文档D_i中出现的频数,$|D_i|$为文档D_i中的单词数,经过两者相除即可计算词频。逆文档频率(IDF),反映词的普遍程度,即一个词越普遍(即有大量文档包含这个词),其IDF值越低;反之,IDF值越高。公式如下:

$$IDF_w=\ln\frac{N}{1+\sum_{i=1}^{N}I(w,D_i)} \tag{3}$$

式中,N为文档总数;$I(w,D_i)$表示文档D_i是否包含找寻的关键词w,若包含,则I为1,若不包含,则I为0;同时,为防止关键词w在所有文档中均未出现从而带来公式无法计算的问题,采用分母加1的方式进行平滑处理。根据式(2)、式(3)的定义,关键词w在文档D_i中的TF-IDF值为:

$$TF\text{-}IDF_{w,D_i}=TF_{w,D_i}\times IDF_w \tag{4}$$

由式(4)可知,当一个新鲜度高(即普遍度低)的词在文档中出现的频率越高时,其TF-IDF值越高,反之越低。计算出每个词的TF-IDF即可得到语料库每个词的重要程度,从而为后续模型设计提供有力的保障。

TF-IDF采用无监督学习,兼顾词频与新鲜度两种属性,可过滤一些常见词,且保留能提供更多信息的重要词。在搜索引擎等实际应用中,该技术是主要的信息检索手段。但是,用词频来衡量一个词的重要程度是不够全面且这种计算无法体现位置关系;同时,严重依赖分词水平(尤其是中文分词更加明显)。

1.2.3 信息增益技术

信息增益(information gain,IG)[17]表示文本中包含某一特征信息时文本类的平均信息增益,定义为某一特征在文本出现前后的信息熵之差。假设c为文本类变量,C为文本类集合,d为文本,f为特征,对于特征f的信息增益$IG(f)$表示如下:

$$IG(f)=H(C)-H(C|f)=$$

$$\sum_{c\in C}\left(P(c,f)\mathrm{lb}\frac{P(c,f)}{P(c)P(f)}+P(c,\bar{f})\mathrm{lb}\frac{P(c,\bar{f})}{P(c)P(\bar{f})}\right)$$

式中，$P(\bar{f})$是语料中不包含该特征的概率。由上可知，一个特征的信息增益实际上描述的是它包含的能够帮助预测类别信息属性的信息量。

从理论上讲，信息增益应该是最好的特征选取方法，但实际上由于许多信息增益比较高的特征出现频率往往较低，当使用信息增益选择的特征数目比较少时，通常会存在数据稀疏问题，此时模型效果比较差。因此，一般在系统实现时，首先对训练语料中出现的每个词（以词为特征）计算信息增益，然后指定一个阈值，从特征空间中移除那些信息增益低于此阈值的词条，或者指定要选择的特征个数，按照增益值从高到低的顺序选择特征组成特征向量。信息增益适合于考察特征对整个模型的贡献，而不能具体到某个类别上，这就使得它只适合用来做“全局”的特征选择而无法做“本地”的特征选择。该技术适合于情感分类、意图识别、垃圾邮件自动处理等分类领域。

1.2.4 卡方分布

χ^2统计量（chi-square distribution，χ^2）[18]衡量的是特征项t_i和类别C_j之间的关联程度，并假设t_i和C_j之间符合具有一阶自由度的χ^2分布。特征对于某类的χ^2统计值越高，它与该类之间的相关性越大，携带的类别信息也较多，反之则越少。

如果令N表示训练语料中文档的总数，A表示属于C_j类且包含t_i的文档频数，B表示不属于C_j类但包含t_i的文档频数，C表示属于C_j类但不包含t_i的文档频数，D是既不属于C_j也不包含t_i的文档频数，N为总的文本数量，特征项t_i对C_j的卡方值为：

$$\chi^2(t_i,C_j)=\frac{N\times(A\times D-C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{6}$$

对于多类问题，基于卡方统计量的特征提取方法可以采用两种方法：一种方法是分别计算对于每个类别的卡方值，然后在整个训练语料上计算；其二为计算各特征对于各类别的平均值。与此类似的方法还有互信息技术（mutual information，MI）[19]。

卡方分布具备完善的理论，与信息增益技术类似，适用于分类模型领域。但是，该技术理论较为复杂，对数学能力要求较高。

当然，除以上常见的特征提取技术外，还有一些不太常用的方法，例如DTP（distance to transition point）[20]方法、期望交叉熵法（expected cross entropy）[21]、优势率法[22]等。

1.3 Textrank技术

Textrank（TextRank graph based ranking model）[23]是一种基于图排序的处理技术，基本思想来自PR（PageRank）算法[24]。该技术在语料预处理后将语料文本分割成若干组成单元（单词、词组、句子等）并建立图模型，再利用投票机制对文本中的重要成分进行排序，从而仅利用文章信息即可实现关键字提取和文摘生成等。

具体来说，Textrank表示为一个有向有权图$G=(V,E)$，该图由点集V和边集E组成，其中，E是$V\times V$的子集。对于给定的点V_i，$In(V_i)$为指向该点的所有点集合，$Out(V_i)$为点V_i指向的所有点集合，V_i的得分定义如下：

$$WS\{V_i\}=\{1-d\}+d\times\sum_{v_i\in\{V_i\}}\frac{w_{ji}}{\sum_{V_k\in out(V_j)}w_{jk}}WS\{V_j\} \tag{7}$$

式中，d为阻尼系数，取值范围为0到1，含义为从图中某一特定点指向其他任意点的概率；w_{ji}表示节点i在节点j处的权重。将切分后的所有V_i得分进行排序即可得到语料库中所有单词或短语的重要程度，从而为模型设计提供保障。

Textrank采用无监督学习，使用者不需要有深入的语言学或相关领域知识；同时，采用基于图的排序算法，综合考虑文本整体信息来确定哪些单词或者句子，从而基于这些重点单词或重点句子进行下一步处理。该技术适合于自动文摘等生成式文本任务与关键词提取等词级自然语言处理。然而，Textrank与TF-IDF一样严重依赖分词结果，Textrank虽然考虑到词之间的关系，但是仍然倾向于将频繁词作为关键词；同时，Textrank涉及到构建图以及迭代计算，因而提取速度较慢。

1.4 语义分析

1.4.1 隐含语义分析

隐含语义分析（latent semantic analysis，LSA）是一种知识获取和展示的计算理论方法，出发点是语料中的词与词之间存在某种联系，即存在某种潜在的语义关系，而这种潜在语义关系隐含于文本中词语的上下文模式中，因此需要对大量语料进行分析进而寻找这种潜在的语义关系。LSA不需要确定语义编码，仅依赖于文本上下文中事物的联系，并用语义关系来表示文本，简化文本向量的目的。该方法

的核心思想是将文档-术语矩阵分解为相互独立的文档-主题矩阵和主题-术语矩阵[25-26]。

在实际应用中，原始计数的效果不理想（如果在词汇表中给出m个文档和n个单词，可以构造一个$m\times n$的矩阵$\boldsymbol{A}$，其中每行代表一个文档，每列代表一个单词。在LSA的最简单版本中，每一个条目可以是第j个单词在第i个文档中出现次数的原始计数），因此，LSA模型通常用TF-IDF得分代替文档-术语中的原始计数。一旦拥有文档-术语矩阵$\boldsymbol{A}$，即可求解隐含主题。由于$\boldsymbol{A}$可能是稀疏的，具有极大噪声且在维度上存在大量冗余的特性，因此，一般情况下采用奇异值分解法（singular value decomposition，SVD）[27]处理，公式如下：

$$\boldsymbol{A}\approx\boldsymbol{U}_t\boldsymbol{S}_t\boldsymbol{V}_t^{\mathrm{T}} \tag{8}$$

$\boldsymbol{U}\in\mathbf{R}^{m\times t}$是文档-主题矩阵，行表示按主题表达的文档向量；$\boldsymbol{V}\in\mathbf{R}^{n\times t}$则是术语-主题矩阵，行代表按主题表达的术语向量。经过这样的处理，可以得到词之间的隐含关系。

LSA采用低维词条、文本向量代替原始的空间向量，能有效处理大规模语料且具有快速高效的特点，适用于信息过滤、文本摘要以及机器翻译等跨语言信息检索等生成式自然语言处理领域。但是LSA在进行信息提取时，忽略词语的语法信息（甚至是忽略词语在句子中出现顺序），处理对象是可见语料，不能通过计算得到词语的暗喻含义和类比推论含义，同时需要大量文件和词汇来获得准确结果，存在表征效率较低的缺点。为了解决这些问题，研究者们对其进行了改进，其中最成功的改进为概率隐含语义分析（probabilistic latent semantic analysis，PLSA）[28]。

1.4.2 概率隐含语义分析

Hofmann在1999年撰写了概率隐含语义分析PLSA[28-29]，通过一个生成模型为LSA赋予概率意义上的解释。作者认为每篇语料都包含一系列可能的潜在话题，语料中的每个单词都不是凭空产生的，而是在这些潜在的话题的引导下通过一定概率生成的，这也正是PLSA提出的生成模型的核心思想。PLSA通过下式对d和w的联合分布进行建模：

$$P(w,d)=\sum_z P(z)P(d|z)P(w|z)=P(d)\sum_z P(z|d)P(w|z) \tag{9}$$

式中，d表示一篇文档，z表示由文档生成的一个话题，w表示由话题生成的一个单词。在该模型中，d和w是已经观测到的变量，z是未知变量（代表潜在话题）。

PLSA能从概率的角度解释模型，使模型变得容易理解；同时，相对于LSA的SVD方法，PLSA的EM[30]（expectation maximization）算法具有线性收敛速度，可以使似然函数达到局部最优。但是该模型无法生成新的未知文档，同时，随着文档和词语个数的增加，模型的复杂度会快速增大，从而导致模型出现严重过拟合。

1.5 其他预训练技术

以上四类常见的预训练技术与模型耦合性相对较低，具有较明显的区分。除此之外，部分传统自然语言预训练技术与模型耦合性较高，较难将预训练技术单独展示。这些常用到的有根据先验概率求后验概率的贝叶斯分类技术（Bayesian classification，BC）[31]、具有多重降级状态的马尔可夫（Markov model，MM）[32]与隐马尔可夫模型（hidden Markov model，HMM）[33]、判别式概率的无向图随机场（random field，RF）[34-35]等。

综上，对常用的传统预训练技术进行汇总，如表1所示。对每一个具体技术特点、优缺点及适用范围进行总结。但是，在传统的自然语言预训练技术中，存在无词序或词序不全问题，严重影响处理结果。基于此，神经网络的自然语言预训练技术，尤其是深度学习的自然语言预训练技术，对这些不足做了一定的纠正。

2 神经网络预训练技术

针对传统自然语言预训练技术的不足，神经网络自然语言预训练技术采取了改进措施，主要是将词序间上下文关系考虑到实际语料中，这一部分综述在国内外相对较多。Qiu等[6]从词序是否上下文相关、语言模型结构、任务类型以及技术应用范围四方面来阐述预训练及模型相关技术，较为全面展现了神经网络的自然语言预训练发展脉络。但是该论文在不同分类方面存在较大交叉；同时，对传统预训练技术涉及较少。Liu等人[5]对无监督预训练机制进行了综述，该文章从体系结构与策略两方面进行展开讨论，并对相关工作进行总结与展望。但是，该综述取材时间较近且关注范围狭小，对神经网络预训练技术以及传统预训练技术部分并未涉及。在国内方面，刘睿珩[36]、余同瑞[37]、李舟军[38]等人分别单独进行

Table 1 Summary of traditional pre-training techniques

表1 传统预训练技术汇总

模型大类	具体模型	技术特点	优点	缺点	适用条件与范围
N-gram	*N*-gram[7-8]	依据滑动窗口表示为gram列表	理论完善、原理简单、容易操作	词表有限、语义鸿沟、数据稀疏等问题	适用词级和句子级自然语言处理领域，例如拼写检查、自动索引等
向量空间模型	独热码[11]	将文本表示扩展到欧式空间，便于计算与比较	扩充特性、简单有效、便于理解	维度过高、语义鸿沟且无法体现单词间远近程度	适合于基于参数与距离的模型，例如SVM、NN、*K*NN等
	TF-IDF[15-16]	根据词频以及逆文档频率计算词的重要程度	无监督学习，能过滤一些常见词和保留重要词的信息	无法体现位置关系且严重依赖分词	适用于问答检索领域，例如搜索引擎、查询系统等
	信息增益[17]	特征信息在出现前后的信息熵之差	理论上来说应该是最好的特征选取方法，理论完善	信息增益较高的词频较少，因而产生数据稀疏	适合于分类领域，例如垃圾邮件过滤、情感分类等
	卡方分布[18]	衡量特征项与类别之间的关联程度	理论完善	数学公式复杂，较难理解	适合于分类领域，例如垃圾邮件过滤、意图识别等
Textrank技术	Textrank[23]	借鉴PageRank算法，将语料分割成组成单元并建立图模型	使用者不需要有深入的语言学或专业领域知识	严重依赖分词、提取速度较慢	适合于生成式自然语言处理与词级自然语言处理领域，例如文章摘要
语义分析	隐含语义分析[25-26]	采用低维词条、文本向量代替原始空间向量	快速高效且模型容易理解	忽略词语的语法信息，不能通过计算得到词语的暗喻含义及类比推论含义，需要大量的文件获得准确的结果且表征效率较低	适用于生成式自然语言处理领域，例如信息过滤、文本摘要以及机器翻译等跨语言信息检索
	概率隐含语义分析[28-29]	采用EM方法代替奇异值分解SVD			
其他技术	贝叶斯[31]	根据先验概率求后验概率的一种有向无环图	简短、快速且复杂度不高	物理含义不足且与现实情况不符	适合于词级自然语言处理领域，例如命名实体识别、关键词提取等
	马尔可夫与隐马尔可夫模型[32-33]	马尔可夫：未来状态只与当前状态有关 隐马尔可夫：由输出序列求隐藏序列	预测多重降级状态的系统概率	模型只依赖每个状态及观察对象且目标函数与预测函数不匹配	适用于句子级自然语言处理，例如语义消歧
	条件随机场[35]	是一种判别式概率无向图学习模型	CRF使用场景宽泛，不存在局部最优值问题	复杂度较高、训练代价较大	适用于句子级自然语言处理，例如语义分析

了自然语言处理预训练技术的研究综述，这几者综述较为类似，均是重点介绍神经网络相关技术的概要方法。但是整体内容较为浅显且对传统预训练技术关注度较低。

本文针对以上不足，从神经网络预训练技术出发，以词序是否上下文相关分为词向量固定表征和词向量动态表征两种方式，以此为线索，展现出更为合理的神经网络预训练技术。

2.1 词向量固定表征

词向量固定表征是将目标词的上下文相关词考虑进去，能够较好地解决词性孤立不连贯问题。常见的词向量固定表征有神经语言模型技术（neural network language model，NNLM）、C&W（Collobert and Weston）、Word2vec（word to vector）、FastText、Glove（global vectors for word representation）等。

2.1.1 神经语言模型

神经语言模型NNLM[39]：神经语言模型通过对元语言模型进行建模，估算$P(w_i|w_{i-(n-1)},w_{i-(n-2)},\cdots,w_i$的值。与传统技术不同的是，NNLM不是通过计数的方法对目标条件进行概率计算，而是通过构建一个神经网络结构对目标进行建模求解。图3显示了NNLM模型结构。

NNLM主要由三层网络构成：输入层、隐藏层和输出层。模型预训练在输入层与隐藏层中完成（即图3中的矩阵$\boldsymbol{C}$）。具体来说，分以下几步：首先，输入层输入$n-1$个词汇（每个词汇进行One-hot编码，

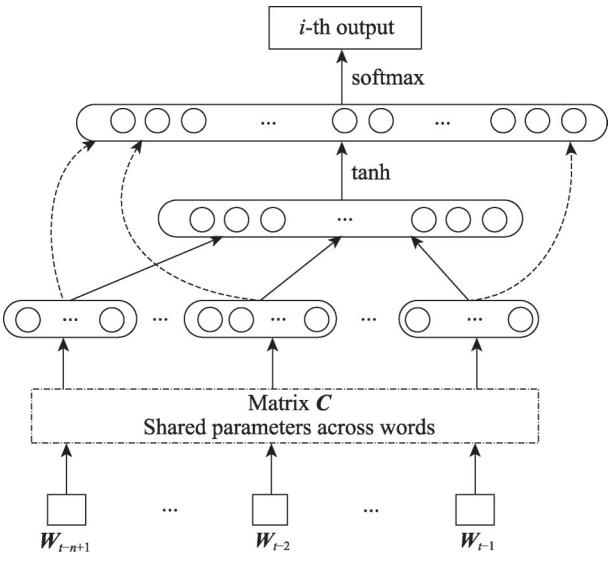


Fig.3 NNLM model

图3 NNLM模型

$1\times|V|$)；其次，将词汇与矩阵$\boldsymbol{C}(|V|\times m$相乘，得到一个分布式向量($1\times m$)。这个分布式向量即为语料预训练的结果(需要经过若干次模型迭代训练才能得到较高准确度)。

由于该模型较为基础，本文对整个模型的训练过程不做详细介绍。NNLM模型使用低维紧凑的词向量对上文进行表示，解决了词袋模型带来的数据稀疏、语义鸿沟等问题。该技术一般应用于缺失值插补、句式切分、推荐系统以及文本降噪等句子级自然语言处理领域。但是，该模型只能利用当前语料的上文信息进行标准化操作，不能根据上下文对单词意思进行实时调整；同时，模型的参数量明显大于其他传统模型。为解决该问题，Mnih等人[40-41]提出了Log双线性语言模型(log-bilinear language model，LBLM)。

LBLM[41]：Mnih与Hinton提出一种层级思想替换NNLM中隐藏层到输出层最花时间的矩阵乘法(语料预训练部分与NNLM相同)，LBLM的能量函数为：

$$\begin{cases} h=\sum_{i=1}^{t-1}\boldsymbol{H}_i\boldsymbol{C}(w_i) \\ y_j=\boldsymbol{C}(w_j)^{\mathrm{T}}h \end{cases} \tag{10}$$

如式(10)所示，$\boldsymbol{C}(w_i)$与$\boldsymbol{C}(w_j)$表示序列中对应位置的转移矩阵；t是序列中建模元素的数量；$\boldsymbol{H}_i$是一个$m\times m$矩阵，可以理解为第i个词经过$\boldsymbol{H}_i$变换后，对第t个词产生的贡献；h为隐藏单元；y_j为预测词w_j的对数概率。LBLM模型的能量函数与NNLM模型的能量函数主要有两个区别：其一，LBLM模型中没有非线性激活函数tanh；其二，LBLM只有一份词向量。之后的几年中，Mnih等人在LBLM模型基础上做了一系列改进工作，其中改进最成功的模型有两个：层级对数双线性语言模型(hierarchical LBL，HLBL)[42]以及基于向量的逆语言模型(inverse vector LBL，ivLBL)[43]。

LBLM模型没有激活函数，隐藏层到输出层直接使用词向量，从而使模型更加简洁、准确度更高。但是，理论上LBLM需构建多个矩阵(有几个词就需要几个矩阵)，而迫于现实压力采用近似处理，因而在准确度方面存在偏差；同时，LBLM仍不能解决一词多义问题。

2.1.2 C&W技术

C&W技术[44]是由Collobert和Weston于2008年提出的以生成词向量为目标的模型技术(之前的大多数模型以生成词向量为副产品)，该技术直接从分布式假说的角度来设计模型和目标函数。C&W模型结构如图4所示。

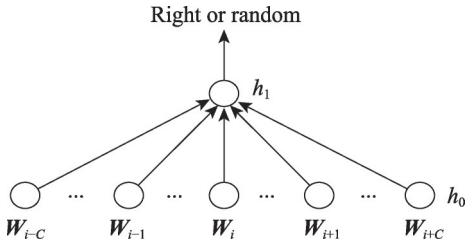


Fig.4 C&W model

图4 C&W模型

C&W模型对应公式为：

$$\begin{cases} h_1=f(\boldsymbol{W}_0h_0+b_0) \\ s(w_i,C)=\boldsymbol{W}_1h_1+b_1 \\ \sum_{(w_i,C)\in D}\sum_{w'_i\in V}\max(0,1+s(w'_i,C)-s(w_i,C)) \end{cases} \tag{11}$$

模型的整个流程为：将$w_{i-C},\cdots,w_{i-1},w_i,w_{i+1},\cdots,w_{i+C}$从初始化词向量矩阵$\boldsymbol{L}$中获取对应的词向量进行拼接，作为第一层$h_0$；进而通过激活函数$f(\cdot$得到$h_1$；再经过线性变换得到$(w_i,C$的得分$s(w_i,C$；模型遍历语料库的所有语料，并对目标函数进行最优化。优化完成得到最终的词向量矩阵$\boldsymbol{L}$和生成词w_i。

C&W模型与NNLM相比，不同点主要在于C&W将目标词放在输入层，同时输出层也从神经语

言模型的$|V|$个节点变为了一个节点，该节点的数值表示对这n元组短语打分，打分只有高低之分，没有概率特性，因此无需进行归一化操作。C&W模型使用这种方式把NNLM模型在最后一层的$|V|\times|h|$次运算降为$|h|$次运算，极大地降低了模型的时间复杂度。由于C&W模型技术是以生成词向量为目标的模型技术，因而应用领域相对广泛，小到词级的单词纠错，大到篇章级文本语义分析等。但是，C&W模型只利用了局部上下文，不能解决一词多义问题；同时，上下文信息不能过长，过长则存在信息丢失。为了改善以上弱点，Huang等人[45]对C&W模型进行改进，提出通过全局上下文以及多个原生单词进行准确度提升的操作。

2.1.3 Word2vec技术

Word2vec[46]是2013年Google开源的一个词嵌入（Word Embedding）工具。Embedding本质是用一个低维向量表示语料文本，距离相近的向量对应的物体有相近的含义。Word2vec工具主要包含两个模型：连续词袋模型（continuous bag of words，CBOW）与跳字模型（skip-gram）。两种高效训练的方法：负采样（negative sampling）和层序Softmax（hierarchical Softmax）。由于本文介绍预训练技术，本小节仅介绍连续词袋和跳字两种模型。

连续词袋模型CBOW[47]是根据输入的上下文来预测当前单词。模型结构如图5所示。

CBOW模型输入为独热码；隐藏层没有激活函数，即为线性单元；输出层维度与输入层维度一样，使用Softmax回归。后续任务用训练模型所学习的参数（例如隐层的权重矩阵）处理新任务，而非用已训练好的模型。

CBOW模型具体处理流程为：(1)输入层。上下文单词的One-hot（假设单词向量空间维度为V，上下文单词个数为C）。(2)所有One-hot分别乘以共享的输入权重矩阵$\boldsymbol{W}$（$V\times N$矩阵，N为自设定）。(3)所得的向量（因为是One-hot，所以是向量）相加求平均作为隐层向量。(4)乘以输出权重矩阵$\boldsymbol{W}'$（$N\times V$矩阵）。(5)激活函数处理得到V-dim概率分布（因为是One-hot，其中的每一维都代表着一个单词）。(6)概率最大的index所指示的单词为预测出的目标词（target word）。(7)将目标词与真实值的One-hot做比较，误差越小越好（从而根据误差更新权重矩阵）。经过若干轮迭代训练后，即可确定$\boldsymbol{W}$矩阵。输入层的每个单词与矩阵$\boldsymbol{W}$相乘得到的向量就是想要的词向量（预训练词向量只是其中的副产物）。

跳字模型（Skip-gram）[48]，输入是特定词的词向量，输出是特定词对应的上下文词向量。模型结构如图6所示（具体训练过程不再介绍，与CBOW模型相似）。

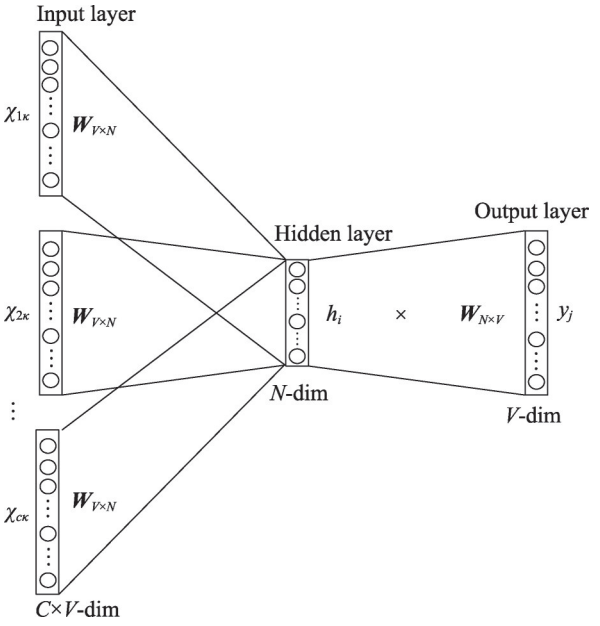


Fig.5 CBOW model

图5 CBOW模型

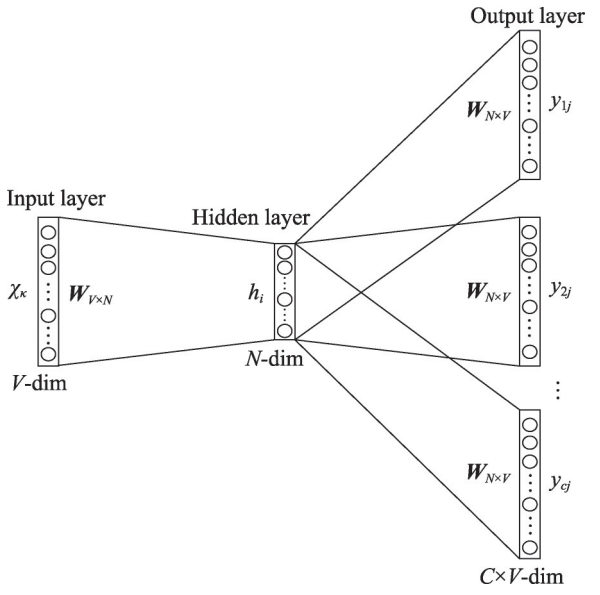


Fig.6 Skip-gram model

图6 Skip-gram模型

由于Word2vec考虑上下文关系，与传统的Embedding相比，嵌入的维度更少，速度更快，通用性更强，效果更好，可以应用在多种自然语言处理任务中，例如常见的文本相似度检测、文本分类、情感分析、推荐系统以及问答系统等句子级与篇章级自然语言处理领域。然而，由于与向量是一对一的关系，无法解决一词多义问题。同时，Word2vec是一种静态的方法，无法针对特定任务做动态优化，并且它的相关上下文不能太长。

2.1.4 FastText技术

FastText[49-50]，该方法是2016年开源的一个词向量及文本分类工具。FastText模型架构与Word2vec的CBOW模型架构非常相似。图7为FastText的模型结构。

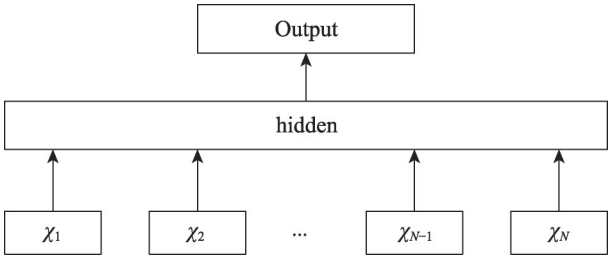


Fig.7 FastText model

图7 FastText模型

模型详情参看CBOW模型，这里仅对其与CBOW模型中预训练技术的不同部分进行详细介绍。最主要的区别是两种模型的单词嵌套表示不一样，CBOW是单词级别的Embedding，而FastText将单词拆分的同时加入了字符级别的Embedding，起到扩充词汇的作用。但是该操作会带来巨大的Embedding表，为计算和储存带来了很大的挑战。为了解决该问题，FastText将字词对应的原始特征向量进行Hash处理[51]，具体公式如下：

$$\phi_i^{(h,\xi)}=\sum_{j:h(j)=i}\xi(j)x_j \tag{12}$$

式中，h与ξ都是Hash函数，h将输入映射到$(1,2,\cdots,m)$之间，ξ将输入映射到$\{1,-1$之间，从而将N维的原始离散特征Hash成M维的新特征$(M\ll N)$。具体来说，对于原特征的第k维值，先通过$h(k$将k映射到$1\sim M$之间，再通过另外一个Hash函数$\xi(k$将k映射成1或−1，然后将映射到同一维度上的值进行相加，这样原来的N维特征就映射成M维特征。

FastText最大特点是模型简单，训练速度非常快，适用于大型语料训练，支持多种语言表达。该技术一般应用于文本分类与同义词挖掘领域，例如常见的垃圾邮件清理、推荐系统等。但是，FastText的词典规模巨大，导致模型参数巨大；同时，一个词的向量需要对所有子词向量求和，继而导致计算复杂度较高。

2.1.5 Glove技术

Glove[52]是2014年提出的一个基于全局词频统计的词表征工具，可以将单词表达成由实数组成的向量，这些向量捕捉了单词间的语义特性。

Glove首先基于语料构建词的共现矩阵$\boldsymbol{X}$（设共现矩阵为$\boldsymbol{X}$，其元素意义为在整个语料库中，单词i和单词j共同出现在一个窗口中的次数），然后构建词向量与共现矩阵之间的近似关系，关系表示为：

$$\boldsymbol{w}_i^{\mathrm{T}}\tilde{\boldsymbol{w}}_j+\boldsymbol{b}_i+\tilde{\boldsymbol{b}}_j=\ln X_{ij} \tag{13}$$

式中，$\boldsymbol{w}_i^{\mathrm{T}}\tilde{\boldsymbol{w}}_j$为最终求解的词向量，$\boldsymbol{b}_i$、$\tilde{\boldsymbol{b}}_j$为偏置向量；根据式(13)构建损失函数，公式如下：

$$J=\sum_{i,j=1}^{N}f(X_{ij})\,(\boldsymbol{w}_i^{\mathrm{T}}\tilde{\boldsymbol{w}}_j+\boldsymbol{b}_i+\tilde{\boldsymbol{b}}_j-\ln X_{ij})^2 \tag{14}$$

式中，$\boldsymbol{w}_i$、$\tilde{\boldsymbol{w}}_j$是单词i与单词j的词向量，$\boldsymbol{b}_i$与$\tilde{\boldsymbol{b}}_j$是两个偏置向量，N为词汇表的大小（共现矩阵的维度为$N\times N$），f为权重函数，对应的公式为：

$$f(x)=\begin{cases}(x/x_{\max})^{0.75}, & x<x_{\max}\\ 1, & x\geqslant x_{\max}\end{cases} \tag{15}$$

式(13)~(15)经过若干次迭代后即可求出$\boldsymbol{w}_i^{\mathrm{T}}\tilde{\boldsymbol{w}}_j$。Glove构建流程如图8所示。

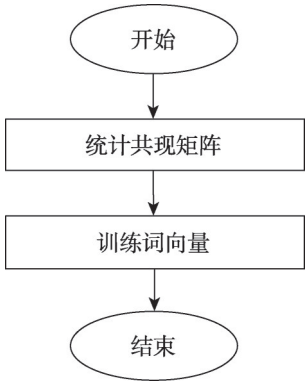


Fig.8 Process of Glove building word vector model

图8 Glove构建词向量模型流程

由于结合了SVD与Word2vec的优势，能够充分利用统计数据，因此，Glove训练速度较快，可以在较大的语料库上进行训练。该方法在较小的语料库或者维度较小的词向量上训练时也有不错表现，同时该方法可以概括比相关性更复杂的信息。该技术适用于自动文摘、机器翻译等自然语言生成领域与问答、文本分类等自然语言理解领域。当然，Glove算法使用全局信息，相对于其他模型内存耗费相对较多，且仍不能解决一词多义的问题。

2.2 词向量动态表征

词向量动态表征是在预训练阶段将目标词的上下文相关词考虑进去，同时，在涉及具体语句时会将目标词的上下文考虑进去，能够较好地解决词性孤立不连贯及一词多义问题。这类技术在图形图像领域较为成熟[53-55]，但是由于语义的多样性与不确定性，导致该技术在自然语言处理中较难适用。随着ULMFit模型[56]对词向量动态表征带来的较大影响，为自然语言的预训练技术发展提供了一定参考价值。而后各种动态表征技术诞生，常见的有Elmo（embeddings from language models）、GPT（generative pre-training）以及BERT模型等。

2.2.1 Elmo模型

Elmo模型[57]的本质思想是先用语言模型学习一个单词的Word Embedding（可以用Wrod2vec或Glove等得到，文献[57]中使用的是字符级别的残差CNN（convolutional neural networks）得到Token Embedding），此时无法区分一词多义问题。在实际使用Word Embedding的时候，单词已经具备特定的上下文，这时可以根据上下文单词的语义调整单词的Word Embedding表示，这样经过调整后的Word Embedding更能表达上下文信息，自然就解决了多义词问题。

图9展示了Elmo模型的预训练过程，该模型的网络结构采用双层双向LSTM（long short-term memory）[58]。使用该网络对大量语料预训练，从而新句子中每个单词都能得到对应的三个Embedding：最底层是单词的Embedding（word embedding）；中间层是双向LSTM中对应单词位置的Embedding（position embedding），这层编码单词的句法信息更多一些；最高层是LSTM中对应单词位置的Embedding（position embedding），这层编码单词的语义信息更多一些。也就是说，Elmo的预训练过程不仅仅学会单词的Embedding，还学会了一个双层双向的LSTM网络结构。预训练完成后，将会得到一个半成品检查点，将需要训练的语料库经过处理后连同检查点一起送入后续任务中进行拟合训练，从而后续任务可以基于不同的语料文本得到不同的意思。

经过如上处理，Elmo在一定程度上解决了一词多义问题且模型效果良好。Elmo模型技术开始适用于语义消歧、词性标注、命名实体识别等领域，随着研究的深入，适用范围也越来越广。但是它仍存在一定不足：首先，在特征提取器方面，Elmo使用的是LSTM而非Transformer（在已有的研究中表明，Transformer的特征提取能力远强于LSTM）；其次，

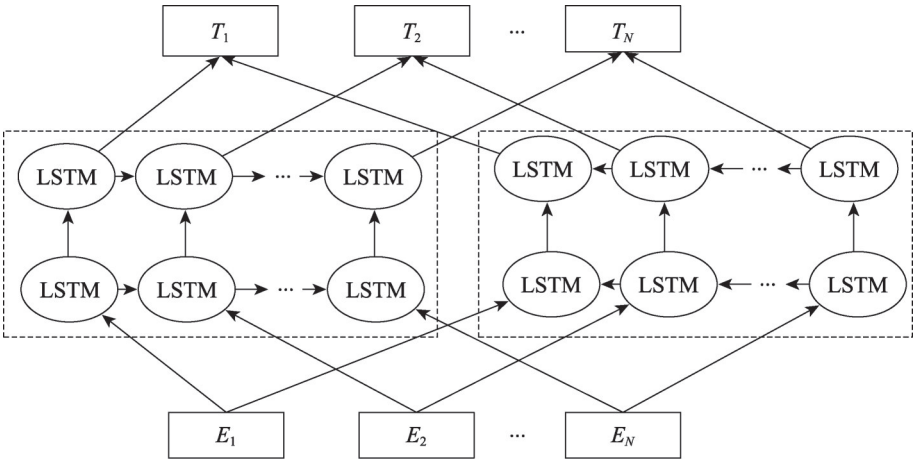


Fig.9 Elmo model

图9 Elmo模型

Elmo采用的双向拼接融合特征比一体化的融合方式要弱一些。

2.2.2 GPT模型

GPT模型[59]：GPT模型用单向Transformer完成预训练任务，其将12个Transformer叠加起来[60]。训练的过程较简单，将句子的n个词向量加上位置编码（positional encoding）后输入到Transformer中，n个输出分别预测该位置的下一个词。图10为GPT的单向Transformer结构和GPT的模型结构。

总的来说，GPT分无监督预训练和有监督拟合两个阶段，第一阶段预训练后有一个后续拟合阶段。该模型与Elmo类似，主要不同在于两点：首先，使用Transformer而非LSTM作为特征抽取器；其次，GPT采用单向语言模型作为目标任务。

GPT模型采用Transformer作为特征提取器，相对于LSTM能有效提取语料特征。虽然其应用领域较为广泛，但其最为突出的领域为文本生成领域。然而，采用的单向Transformer技术，会丢失较多关键信息。

GPT-2模型[61]：GPT-2依然沿用GPT单向Transformer模式，但是在GPT上做了一些改进。首先，不再针对不同层分别进行微调建模，而是不定义这个模型具体任务，模型会自动识别出需要什么任务；其次，增加语料和网络的复杂度；再者，将每层的正则化（layer normalization）放到每个Sub-block之前，并在最后一个Self-attention之后再增加一个层正则化操作。

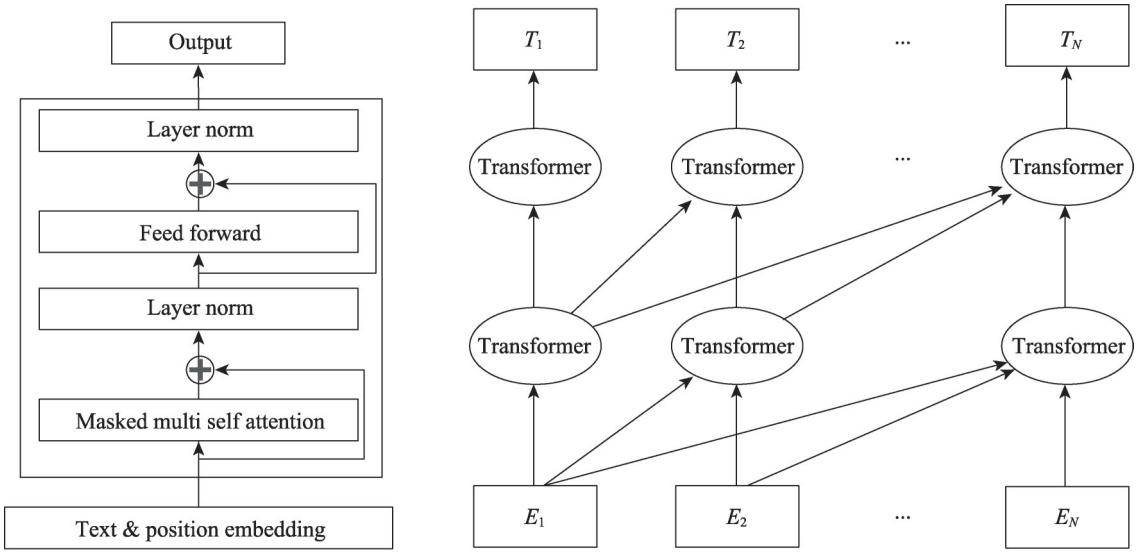
相对于GPT模型，GPT-2提取信息能力更强，在文本生成方面的性能尤为优越。但是，该模型的缺点与GPT一样，采用单向的语言模型会丢失较多关键信息。

GPT-3模型[62]：GPT-3是目前性能最好的通用模型之一，聚焦于更通用的NLP模型，主要解决对领域内标签数据的过分依赖和对领域数据分布的过拟合问题。特色依然沿用了单向语言模型训练方式，但是模型的大小增加到了1 750亿的参数量以及用45 TB的语料进行相关训练。

在通用NLP领域中，GPT-3的性能是目前最高的，但是，其在一些经济政治类问题上表现不太理想（由预训练语料的质量造成）；同时，该模型由于参数量过于巨大，目前大部分学者只能遥望一二，离真正进入实用阶段还有较远距离。

2.2.3 BERT模型

BERT模型[63]：BERT采用和GPT完全相同的两阶段模型，首先是语言模型预训练，其次是后续任务的拟合训练。和GPT最主要不同在于预训练阶段采



(a) Single Transformer structure of GPT
(a) GPT的单项Transformer结构

(b) Model structure of GPT
(b) GPT的模型结构

Fig.10 GPT related model

图10 GPT相关模型

用了类似Elmo的双向语言模型技术、MLM（mask language model）技术以及NSP（next sentence prediction）机制。图11为BERT模型。

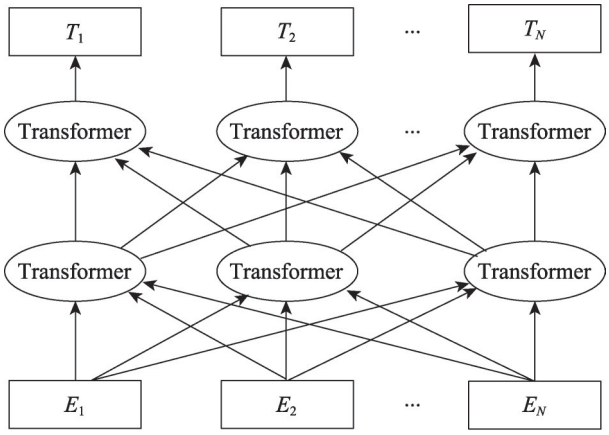


Fig.11 BERT pre-training model

图11 BERT预训练模型

在MLM技术中，Devlin等人随机Mask每个句子中15%的单词，用来做预测，而在这15%的单词中，80%的单词采用[Mask]，10%的单词采用随机替换，剩下的10%单词保持不变的特性。在NSP机制中，选择句子对A、B，其中50%的B是A的下一条句子，而另外的50%是从语料库中进行随机挑选的句子，进而让它们学习其中的相关性。经过若干次训练，保存检查点即为预训练模型。

BERT采用双向Transformer技术，能较准确地训练词向量，进而引发了自然语言处理的大地震。现阶段，常用的自然语言处理技术绝大部分是基于BERT及其改进技术。从现阶段来看，BERT的应用领域较为广泛，从自然语言理解领域的文本分类、阅读理解等热点领域到自然语言生成的自动文摘、文本写作等领域均有涉猎。但是，该模型存在参数量巨大，实际应用困难等缺点。

以上是常用的神经网络预训练技术，本文从具体模型技术、模型技术特点、模型技术优缺点及适用范围进行总结整理，表2为神经网络预训练技术汇总。

3 BERT改进模型预训练技术

BERT模型作为自然语言领域目前应用最广的模型技术，现已辐射到自然语言处理的各个领域并取得了极大发展。但是，经学者们研究，BERT仍然存在较为明显的缺陷。首先，BERT采用的NSP预训练技术会导致结果出现主题预测，主题预测比实际预测简单，从而效果出现偏差；其次，采用随机Mask部分单词而不是连续的词组，同样导致BERT的效果出现折扣；最后，BERT相对于其他模型来说，参数量相对较大，难以部署在性能受限的边缘设备上。

基于以上几点，出现了BERT的两种大方向改进：其一为尽可能改进BERT以提升性能；其二为在保持BERT模型性能不受本质影响前提下，压缩BERT模型的大小。以下将从这两方面来进行相关介绍。

3.1 提高模型性能方向

BERT模型虽然取得较理想的结果，但是距人类平均水平还存在一定差距。基于此，相关研究者在BERT基础上做了大量改进工作以提升模型性能。提升模型性能的方式主要有两种：一是基于预训练技术改进；二是基于后续任务单独改进，本文着重于介绍前者。下面较为详细地介绍基于预训练改进的且在领域内具有一定知名度的相关模型及其改进技术。

MT-DNN（multi-task deep neural networks）[64]：当监督语料过少时，BERT的后续任务性能不稳定且性能提升有限。MT-DNN将MTL（multi-task learning）加入到BERT的后续任务中，即将相关的后续任务进行多任务训练，可以在一定程度上弥补监督语料的不足。具体来说，该模型在BERT模型的基础上做了以下改进：在后续任务中，MT-DNN将单句分类、句子对分类、文本相似度打分和相关度排序进行混合多任务训练，而后将这四种任务的损失相加求平均，进而优化。

采用该模型，弥补了部分任务语料不足的问题，同时，由于多种语料混合，还具有正则化的作用，可防止模型过拟合。然而，该模型有堆砌之感，且超参数量多于BERT，调参较为繁琐；同时，由于语料的差异，相较于BERT对比具有不公平性。

MASS（masked sequence to sequence）[65]：针对BERT模型在自然语言生成任务上性能较低问题，微软亚洲研究院提出一个在自然语言生成任务上的通用预训练模型MASS。

具体来说，MASS相对于BERT具有以下几点优势：其一，解码器端其他词（在编码器端未被屏蔽掉的词）被屏蔽掉，以鼓励解码器从编码器端提取信息来帮助连续片段预测，这样能促进编码器-注意力-解

Table 2 Pre-training techniques of neural network

表2 神经网络预训练技术

模型大类	典型模型		提出时间	技术特点	优点	缺点	适用条件与范围
词向量固定表征	神经语言模型	NNLM[39]	2003	构建简单的三层神经网络，输入层与隐藏层中的矩阵$\boldsymbol{C}$为所需信息	向量紧凑、同时解决了语义鸿沟	不能解决一词多义问题且只能利用上文信息	一般适用于缺失值插补、句式切分、推荐系统以及文本降噪等领域
		LBLM[41]	2007	提出一种层级思想替换NNLM中隐藏层到输出层最花时间的矩阵乘法	理论上相对NNLM准确度较高、模型简洁	近似处理导致性能受损且不能解决一词多义	
	C&W[44]		2008	第一个以生成词向量为目标的模型	降低了模型复杂度且以生成词向量为目标	该模型只利用局部上下文且不能解决一词多义问题	应用领域广泛，从词级模型的文本纠错到语义分析
	Word2vec	CBOW[47]	2013	根据输入的上下文而预测当前单词	Embedding的维度更少、速度更快且通用性较强，可以应用在多种NLP任务中	无法解决一词多义的问题且无法对特定任务做动态优化工作	适用于自然语言理解与自然语言生成领域，例如问答系统、文本相似度检测、文本分类、情感分析等
		Skip-gram[48]	2013	根据当前单词而预测上下文单词			
	FastText[49-50]		2016	采用词级和字符级Embedding，同时采用Hash整合处理	模型简单、训练速度快且支持多种语言	词典规模较大，模型参数更多且不能解决一词多义	适用于文本分类领域，例如垃圾邮件清理
	Glove[52]		2014	基于全局词频统计	训练速度快	内存耗费较大且不能解决一词多义问题	适用于生成式自然语言处理领域，例如自动文摘
词向量动态表征	Elmo[57]		2018	采用双向LSTM模型进行训练	能解决一词多义问题	特征提取器能力较弱，采用双向拼接特征融合方法	适用于语义消歧、词性标注等领域
	GPT系列[59,61-62]		2018—2020	单向Transformer叠加完成训练	能解决一词多义且特征提取能力较强	单向Transformer结构存在特征丢失	应用领域广泛，最为突出的领域为自然语言生成
	BERT[63]		2018	采用双向的多层Transformer结构	特征提取能力较强且能解决一词多义问题	其中的NSP以及Mask技术有待提高	应用领域广泛，最为突出的领域为自然语言理解

码器结构的联合训练；其二，为了给解码器提供更有用的信息，编码器被强制抽取未被屏蔽掉词的语义，以提升编码器理解源序列文本的能力；其三，让解码器预测连续序列片段，以提升解码器的语言建模能力。

MASS在自然语言生成任务上取得良好的效果，证明了在机器翻译、文本生成等生成式任务上相对于BERT的优势。但是，该模型对自然语言理解任务效果未知，且其超参数k调参过程较为复杂。

UNILM（unified language model）[66]：该模型是对BERT模型的一个延伸，UNILM模型的预训练检查点在自然语言理解与自然语言生成任务上均表现出较高性能。

具体来说，UNILM统一了预训练过程，模型使用Transformer结构囊括了不同类型的语言技术（单向、双向和序列到序列的三种预训练技术），从而不需要分开训练多个语言模型；其次，因为囊括了三种预训练技术，所以参数共享使得学习到的文本表征更加通用化，减少了自然语言处理训练中的过拟合问题。

该模型在自然语言理解与自然语言生成上均取得了良好的性能，具有通用性，适用范围更广。但是预训练语料质量要求较高、预训练时间过长等，均是UNILM面临的现实问题。

ERNIE（enhanced language representation with informative entities）[67]：BERT从纯语料中获取语义模式，较少考虑结构化知识。知识图谱能提供丰富的结构化知识，以便更好地进行知识理解。基于此，采

用大规模语料和知识图谱利用词汇、句法等关联信息训练出BERT的增强版ERNIE模型。

具体来说，在预训练阶段ERNIE分为两部分，提取知识信息与训练语言模型。首先，研究者提取文本语料中的命名实体，将这些提取的实体与知识图谱中的实体进行匹配，为了能够得到结构化的知识编码，模型采用了TransE知识嵌入（translating embedding）[68]算法将实体转化为向量，再将编码后的知识信息整合到语义信息中。其次，在预训练中除采用MLM机制与NSP机制外，增加了新的预训练机制dEA（denoising entity auto-encoder），该机制随机Mask一些实体，并要求模型基于与实体对齐的Tokens，从给定的实体序列中预测最有可能的实体。最后，该模型引入了更多的多源语料，包括中文维基百科、百度百科、百度新闻以及百度贴吧等。采用多源化语料，增大了语料的多样性，且多源语料包含海量的实体类知识，从而预训练的模型能更好地建模真实世界的语义关系。

相对于BERT，ERNIE将BERT与知识图谱结合，在一定程度上改善了结构化知识问题。但是，ERNIE也具有很明显的不足：首先，采用了NSP机制，该机制在后来被证实没有实质性的作用；其次，构建知识图谱需要耗费大量的人力财力；最后，相对于BERT，该模型更为复杂、参数量更多，从而训练成本也相应地高于BERT。为此，之后推出了ERNIE2.0[69]，相对于ERNIE来说，在预训练阶段构建了多任务持续学习预训练框架与三种类型的无监督学习任务。多任务持续学习预训练框架可以根据先前预训练的权重增量学习新的知识；三种类型的无监督学习包括词法级别、语言结构级别和语法级别预训练任务。相对于ERNIE，ERNIE2.0模型性能有较大提升。

XLNet（generalized autoregressive pretraining for language understanding）[70]：XLNet是一个类似于BERT的模型，分为上游预训练阶段和后续微调阶段。具体来说，XLNet上游预训练流程如下：

首先，BERT采用掩码语言模型MLM，从而出现上游预训练任务与后续微调不匹配问题。为了解决这个问题，XLNet在预训练机制引入排列语言模型（permutation language model，PLM），通过构造双流自注意力机制（two-stream self-attention，TSSA），在Transformer内部随机Mask一部分单词，利用自回归语言模型ALM（autoregressive language model）本身的单向特点克服了BERT的后续任务不匹配问题。其次，由于BERT采用Transformer机制，要求输入为定长序列，导致序列长度要相对合适。为了让Transformer学习到更长的依赖，XLNet的Transformer-XL借鉴了TBPTT[71]（truncated back-propagation through time）与相对位置编码，将上一个片段s_{t-1}计算出来的表征缓存在内存里，加入到当前片段s_t的表征计算中。最后，加大了预训练阶段使用的语料规模，BERT采用了13 GB的语料进行预训练，XLNet在BERT预训练语料的基础上，又引入了Giga5、ClueWeb和Common Crawl语料，并排除了一些低质量的语料，额外引入113 GB语料进行预训练。

相对于BERT，XLNet具有以上三点明显优势，因而相对于BERT，该模型在生成式领域与长文本输入类型的任务上性能较高。但是，从本质上来说，XLNet仍然是一个自回归语言模型，排列语言模型机制PLM在处理上下文语境问题时，随机排序比BERT大得多，因此需要更大的运算量才能达到BERT的效果；同时，相对于BERT，XLNet用了更多以及质量更佳的语料进行预训练，这样的对比缺乏一定的公平性。

BERT-WWM（BERT of whole word masking）[72]：该模型与BERT相比，最大的不同是在预训练阶段进行了词组Mask机制，具体来说：首先，采用分词技术对中文语料进行分词处理，则相应的文本被分为多个词组；其次，采用Mask标签替换一个完整的词组而不是单个字（因为前面已经完成了分词）。采用这种预训练方式，模型能学到词的语义信息，训练完成后的字就具有词的语义信息，这对各类中文NLP任务都较为友好。

BERT-WWM模型思想类似于ERNIE，从而模型具有与ERNIE类似的缺点。但是模型开发之初便是对中文语料进行处理，因此在处理中文相关问题时，模型性能较高。在此之后，BERT-WWM扩充了预训练语料库的数据量并且加长了预训练时间，使模型性能进一步提升，即为BERT-WWM-EXT模型（预训练语料库增大，总词数达到54亿；同时，训练步数增大，第一阶段训练1×10^6步，第二阶段训练4×10^5步）。

RoBERTa（robustly optimized BERT）[73]：RoBERTa沿用了BERT框架，但是相对于BERT在预训练过程

和语料规模上做了如下改进：

首先，将静态Mask改为动态Mask，BERT的预训练过程中是随机Mask掉15%的Tokens，在之后的预训练中，这些被Mask的Tokens均保持不变，这种形式称为静态Mask；而RoBERTa在预训练过程中将预训练语料复制多份，每份语料随机Mask掉15%的Tokens，在预训练过程中，选取不同的复制语料，从而可以得到每条语料有不同的Mask，这样的Mask机制称为动态Mask。动态Mask相当于间接增大了训练语料，有助于提高模型的性能与泛化能力。其次，RoBERTa移除了NSP任务，每次可输入多个句子，直到达到设定的最大长度（可以跨段落以及文章），称这种方法为Full-sentences，采用这样的方法模型可以捕获更长的依赖关系，这对长序列的后续任务较为友好。最后，RoBERTa采用了更大的批次量以及更多的语料进行预训练，RoBERTa的批次量远大于BERT，且预训练语料约为BERT的10倍以及采用更长的预训练时间，这样更多的语料增加了语料的多样性，模型性能自然能相对提高。

相对于BERT，RoBERTa具有以上三点优势，在不同的语料库上性能也超过BERT，证明BERT仍然有很强劲的上升空间。但是，RoBERTa采用堆叠式的方式进行处理导致模型过于庞大，很难应用于实际生产生活中。

SpanBERT（spans BERT）[74]：SpanBERT延续了BERT的架构，相对于BERT，在预训练中主要做了以下改进。

首先，对MLM进行改进，提出了Span Mask方案，核心为不再对单个Token进行掩膜处理，而是随机对文本片段添加掩膜。即作者通过迭代采样文本的分词，直到达到掩膜要求的大小。每次迭代过程中，作者从几何分布$I\sim Geo(p)$中采样得到分词的长度，该几何分布是偏态分布，更偏向于较短分词。其次，加入分词边界SBO（span boundary objective）训练任务。具体来说，在训练时选取Span前后边界的两个Token，然后用这两个词加上Span中被遮盖掉词的位置向量，来预测原词。最后，作者采用单序列训练（single-sequence training，SST）代替NSP任务，也就是用一句话进行训练。更长的语境对模型更为有利，模型可以获取更长上下文。

虽然SpanBERT效果普遍强于BERT，尤其是在问答、指代消歧等分词选择任务上表现尤为出色，但是由于该模型采用分词边界SBO，在一些复杂问答方面效果可能欠佳。

K-BERT（BERT of knowledge graph）[75]：由于通用语料预训练的BERT模型在知识驱动型任务上有较大领域差异，K-BERT主要是提升BERT在知识驱动任务上的性能。其将知识图谱引入到BERT的预训练模型中，使模型能够学习特定领域的语义知识，从而达到知识驱动型任务上的良好表现。具体来说，相对于BERT做了以下改变：

首先，制作一个句子树，文本句子经过知识层（knowledge layer）后，知识层对知识图谱（例如CN-DBpedia、HowNet和自建的医学知识图谱）进行检索，从而将知识图谱中与句子相关的三元信息注入到句子中，形成一个富有背景的句子树（sentence tree）。其次，将句子树的信息进行顺序表达，同时通过软位置（soft-position）与可见矩阵（visible matrix）将句子树铺成序列输入模型，进而放入网络中进行相应训练。

除了以上两点优化，K-BERT其余结构均与BERT保持一致，因此该模型兼容BERT类的模型参数，无需再次预训练，节约了计算资源。同时，该模型因为有知识图谱的输入，在许多特定领域的表现显著优于BERT。但是，构造句子树的过程由于语料的词嵌入向量与知识图谱中实体的词嵌入向量匹配问题，需要带来额外的处理；同时，若自行构建知识图谱，需要较大的额外工作量。

SemBERT（semantics-aware BERT）[76]：与BERT相比，SemBERT在BERT的基础上引进语义角色标注模型，它以BERT为基础骨架网络，融合上下文语义信息。具体来说，改进分以下几步：

首先，根据角色标注器（采用SRL（semantic role labeling）标注工具）对文本语料进行标注，给输入的文本语料标注谓词-论元结构（词级别）。其次，将多语义标签进行融合，由于BERT输出的词为子词，难以与角色标注后的词进行对齐，将BERT处理后的子词通过CNN网络进行重构为词，从而使两者对齐。最后，将文本语料表示与语义标签表示集成融合，从而获得了后续任务的联合表示。

SemBERT模型简单有效且易于理解，但是角色标注器标注出的语料本身存在一定的错误，这对后

续任务很不友好；同时，该模型从外部注入相关信息，有可能模型内部的效果与原始BERT相差不大，从而在一些特定任务上引发欠拟合。

StructBERT（structures BERT）[77]：StructBERT将语言结构信息融入BERT，其增加两个基于语言结构的目标，词序重构任务（word-level ordering）和句序判定任务（sentence-level ordering）。具体来说，该模型在预训练任务上进行了如下改进：

首先，一个良好的语言模型，应该有把打乱的句子重构的能力。因而除采用BERT的Mask机制外，还对未Mask的词随机选取Trigram，打乱顺序后重构该顺序；其次，由于NSP机制本质是一个二分类任务，该模型对其进行改进，将原来的二分类模型扩展为三分类模型，即分为是否为上句、是否为下句以及是否无关。

StructBERT基于以上两点改进，在大部分自然语言理解任务上较BERT取得较好的效果，但是该模型相对BERT的本质问题并未进行太大的改进，相对于其他模型，该模型应用不太广泛。

Electra（encoders as discriminators rather than generators）[78]：由于BERT的MLM机制存在天然缺陷，Electra模型提出一种更加简单有效的预训练方案，采用生成器-判别器（replaced token detection）替换BERT中的令牌检测。该模型将部分输入采用生成器生成其他Token替换，然后训练一个判别模型，判别每个Token是否被生成器所替换（两种可能性）。因为该模型是从所有Token中进行学习，而非从被掩盖的部分中学习，相对于BERT在同等条件下性能更为优越。

该模型更适用于较小规模的语料上，即具有更轻量级的模型，但是，GAN（generative adversarial network）在自然语言处理中应用十分困难，因此该模型并非是GAN方法，而是借鉴了GAN的思想；同时，虽然采用了生成器与判别器联合损失训练的方式，然而该训练方式容易退化为单一判别器方式；经过实测，在一些复杂大型任务上，该模型平均性能略微高于BERT，没有论文中的那么高。

以上介绍了基于BERT提升模型性能的常用技术，主要介绍了这些技术中采用的预训练方法的改进部分。这些技术对自然语言的发展起到重要的推动作用。但是，由于模型过于庞大，离应用到实际生产生活中还存在一定的距离，因而部分研究者基于模型性能影响不大的情况下，尽量压缩模型大小。

3.2 模型压缩方向

由于BERT参数众多，模型庞大，推理速度较慢，在一些实时性要求较高、计算资源受限的场景，应用会受到较大限制。因此，研究如何在不过多损失BERT性能的条件下，对BERT进行模型压缩，是一个非常具有现实意义的问题。现阶段，部分研究者专注于压缩BERT模型，使其在边缘设备上具有运行能力。在该方向上，目前有剪枝、量化、知识蒸馏、参数共享与低秩分解等几类方法。

由于模型压缩涉及预训练和后续任务，在这两者之间均有技术改进，耦合性较强，因而不方便单独介绍预训练技术。在介绍相关模型时，对预训练任务与后续任务的改进不进行区分。

3.2.1 剪枝

剪枝是从模型中删除不太重要的部分权重从而产生稀疏的矩阵权重，进而达到模型压缩的目的。

Compressing BERT[79]：该模型探讨了BERT预训练阶段权重修剪对后续任务性能的影响。在三种不同的修剪层次上得到不同的结论：在较低水平预训练模型上剪枝（30%~40%），并不会明显影响后续任务的性能；在中等水平预训练模型上剪枝，会使预训练模型的损失函数相对增大且难以收敛，同时，部分有效信息不能传递到后续任务；在较高水平预训练模型上剪枝，上游任务对后续任务的增益进一步减弱。同时，发现在特定任务BERT上进行微调并不能有效提高模型可裁剪性。

该模型对BERT剪枝压缩进行了一定程度的探讨，为模型压缩做出了一定贡献，但是，这种定性的探讨存在太多的主观性；同时，每一层次剪枝操作后调参较麻烦。

One Head Attention BERT[80]：Michel等人对BERT中的多头注意力机制进行探究，作者给出了三种实验方法证明多头注意力机制存在信息冗余：

首先，每次去掉一层中的一个Head，测试模型性能；其次，每次去掉一层中剩下的层，仅保留一个Head，测试模型的性能；再者，通过梯度来判断每个Head的重要性，然后去掉一部分不重要的Head，测试模型的性能。经过实验证明了多头注意力机制提取的信息之间存在大量冗余。

该模型的优点如标题所示，实验验证了多头注意力机制存在大量冗余，但是，单纯地减少Head的数量不能有效地加速且该结论为实验结果缺乏理论基础。基于此，Cordonnier等人[81]在理论基础上证明了多头机制存在约2/3的冗余。

Pruning BERT[82]：McCarley提出该模型，模型主要通过减少各个Transformer的注意力头数量与前馈子层的中间宽度以及嵌入维度。在SQuAD2.0语料上准确度损失1.5个百分点而解码速度提高了1倍。但是该模型的后续任务基于SQuAD语料进行，也就是说该模型对阅读理解问答具有较好的效果，但是对于其他任务的效果未知，应用范围过窄。

LayerDrop BERT[83]：Fan等人针对BERT提出LayerDrop方法，即一种结构化的Dropout方法对BERT中的Transformer进行处理。

作者提出了一个让Transformer能够在测试过程中使用不同深度的正则项训练方法，该方法关注点在剪枝层数。作者考虑了三种不同的剪枝策略：一为每隔一层就以一定概率进行剪枝；二为计算不同组合层在验证集上的表现，但是这种方法相对耗时；三为每层学习一个参数p，使得全局剪枝率为p^*，然后对每层的输出添加一个非线性函数，在前向中选择计算分数最高的k个层。经过这三种策略从而不需要在后续任务的情况下即可选择BERT模型的最优子模型。

该方法能在一定程度上降低模型大小，加速模型训练且不用后续任务即可完成。但是，该模型对BERT本质缺点并未改进。

RPP BERT[84]：Guo等人提出了一种重加权近端剪枝方法（reweighted proximal pruning，RPP）。在高剪枝率下，近端剪枝BERT对预训练任务和后续多个微调任务都保持了较高的精度。同时，该模型能部署在多种边缘设备上，但是剪枝过程较为繁琐且对近端的选择存在较大争议。

3.2.2 量化

通过减少每个参数所需比特数来压缩原始网络，可以显著降低内存。该方法在图像领域应用较为广泛[85-86]，本小节针对BERT模型的量化改进进行介绍。

Q-BERT[87]：模型采用低位精度储存参数，并支持低位硬件来加速推理过程。

总的来说，作者对Hessian信息进行逐层分解，进而执行混合精度量化。该研究提出一种基于top特征均值和方差的敏感度度量指标，以实现更好的混合精度量化。同时，提出了一种新的组量化机制（group-wise quantization），该机制能够有效缓解准确度下降问题的同时不会导致硬件复杂度显著上升，具体而言，组量化机制将每个矩阵分割为不同的组，每个组拥有独立的量化范围和查找表。最后，作者调查了BERT量化中的瓶颈，即不同因素如何影响NLP性能和模型压缩率之间的权衡，这些因素包括量化机制、嵌入方式、自注意力和全连接层等模块。

Q-BERT对BERT模型进行了有效的压缩，一定程度上降低了模型的大小。但是，整个压缩过程复杂，压缩不彻底且性能影响严重。

Q8BERT[88]：Zafrir等人提出了Q8BERT模型，该模型能够极大地压缩BERT的大小。

具体来说，对BERT的全连接层和Embedding层中通用矩阵（general matrix multiply）进行量化处理。同时，在微调阶段执行量化感知训练，以便在损失最小准确度的同时使BERT压缩模型为原模型25%的参数量。此外，针对8 bit参数进行优化。

该模型与QBERT类似，对BERT能有效地压缩，但是采用该方法存在量化不彻底，比如在softmax、层归一化等准确度要求较高的操作中依然保留float32的类型。

TernaryBERT[89]：华为提出的该模型，在BERT模型上量化分为两部分，权重层量化和激活层量化。在权重层中，包含了所有的线性层与Embedding层，这些层的参数占了BERT模型总参数的绝大部分，因而对这些层的量化较为彻底。华为团队探讨了TWN（ternary weight networks）[90]方法与LAT（loss-aware ternarization）[91]方法，TWN方法旨在最小化全精度参数和量化参数之间的距离，而LAT的方法则是为了最小化量化权重计算的损失。对于激活层的量化，采用8 bit的对称与非对称方法，而在实际推理过程中，矩阵乘法可以由32 bit的浮点数运算变为int 8的整形运算，达到加速的目的。该模型实现仅占BERT模型6.7%的参数情况下达到和全精度模型相当的性能。

虽然该模型在量化方面效果明显高于其他模型，但是在一些粒度较细的任务上由于量化过度导

致效果并不如人意。

3.2.3 知识蒸馏

知识蒸馏的核心是将复杂网络迁移进简单网络中，这之中重要部分是将其中的“精华”蒸馏出来，再用其指导精简网络进行训练，从而实现模型压缩。在BERT兴起以前，知识蒸馏就已存在较多应用[92]，BERT的兴起加速了知识蒸馏在人工智能中的发展。

Small BERT[93]：Zhao等人采用蒸馏方式提出该模型。首先，将BERT的宽度进行压缩，同时，缩小词表，将原来的30 522分词表缩小为4 925个分词。为了使教师模型与学生模型匹配，采用了Dual Training和Shared Projection技术进行处理，其核心是围绕“缩减词表”展开。

该模型能取得较好性能且模型参数得到较大降低，但是该文章实验较简单、不全面，不能说明在其他任务上的性能，即存在应用领域狭窄问题。

Knowledge Distillation[94]：Sun等人在通用知识蒸馏任务上进行改进。学生模型除了学习教师模型的概率输出外，还需要学习一些中间层的输出。作者提出了Skip方法与Last方法，Skip方法为每隔几层去学习一个中间层；Last方法为学习教师模型的最后几层。最终的训练目标是损失函数L_{CE}、L_{DS}与L_{PT}的加权和，直接使用后续任务进行蒸馏训练。

模型在GLUE（general language understanding evaluation）上取得了较为良好的结果，但是该模型本质上是减层操作，从而导致学生模型与教师模型的宽度一样（在一般情况下，短而宽的模型效果往往低于长而窄的模型）；同时，更长更深的教师模型并不一定能训练出良好的学生模型。

DistilBERT[95]：将BERT的12层压缩到6层，以3%的准确度牺牲换来40%的参数压缩和60%的预测提速。具体来说，在预训练阶段进行知识蒸馏，核心技术是引入了$Loss_{cos}$（cosine embedding loss），从而进行网络的内部对齐。而后，作者类似地提出了DistilGPT2[95]和DistilRoBERTa[96]。

DistilBERT在自然语言处理中引起了较大的轰动，但是该方法相对于后面的模型，准确度与参数量均略微较大。

Distilling Transformers[97]：Mukherjee等人针对学生模型蒸馏后效果一般情况下差于教师模型的问题，通过大量领域内无标签语料以及有限数量的标签语料训练来弥补这一差距。具体来说，提出了硬性蒸馏与软蒸馏两种模式，硬蒸馏是对大量无标签语料进行标注，然后将这些语料增强后对学生模型进行监督训练；软蒸馏是用教师模型在无标签语料上生成的内部表示，对学生模型进行蒸馏。

该模型简单易懂，在领域知识内能取得较高性能，但是模型整体创新性不高，对通用任务性能提升有限。

MiniLM[98]：Wang等人提出了一种将基于Transformer的预训练大模型压缩成预训练小模型（更少的层数和更小的隐层维度）的通用方法，深度自注意力知识蒸馏（deep self-attention distillation）。该模型有三个核心点：一为蒸馏教师模型最后一层Transformer的自注意力机制；二为在自注意力机制中引入值之间的点积；三为引入助教模型辅助训练学生模型。

在各种尺寸的学生模型中，MiniLM的单语种模型性能较为优越；在SQuAD2.0与GLUE的多个任务上以一般的参数量与计算量即可保持99%的准确度。但是与TinyBERT与MobileBERT等相比，准确度与参数量还有待提高；同时，对大模型微调和推理仍费时费力，计算成本较高。

TinyBERT[99]：采用两阶段训练方法，该模型在中间的多个过程计算损失函数使其尽量对齐；同时，对语料库进行了极大的增强处理，因此在模型性能与效果上取得了较为明显的进步。基于该思路，研究者们提出了Simplified TinyBERT[100]、CATBERT[101]等模型。

TinyBERT模型在多个任务上取得了较好的性能且其模型大小显著减小，但是模型的超参数过多，模型难以调节；同时，采用了语料增强技术，与BERT的对比不公平。

MobileBERT[102]：该模型为当前蒸馏领域较为通用的模型，该模型采用和$BERT_{large}$一样深的层数（24层），在每一层中的Transformer中加入了bottleneck机制使得每一层Transformer变得更窄。具体来说，作者先训练了一个带有bottleneck机制的$BERT_{large}$（IB-BERT），然后把IB-BERT中的知识迁移到MobileBERT中（由于直接蒸馏效果较低，采用这种中间转换策略）。

该模型优点在于其相对于其他蒸馏模型来说具有通用性，但是模型深度较深，训练更为困难。

BORT[103]：该模型参数量只有BERT的Large模型的16%，但是提升效果能达到0.3%至31%。总的来说，该模型分为最优子结选取（optimal sub-architecture extraction）、预训练与后续任务微调。

该模型能取得较为明显的效果，但是该论文缺乏消融实验且对比较不公平。

3.2.4 参数共享与低秩分解

参数共享是指将模型中相似的子结构采用参数覆盖的方式进行训练，进而达到参数共享的目的。低秩分解是将大的权重矩阵分解为若干个低秩的小矩阵从而减少运算量。由于这两种技术常混在一起使用，因而对其进行整体介绍。

ALBERT[104]：该模型用参数共享与低秩分解技术进行压缩。具体来说，相对于BERT，有以下几点改进：首先，采用词向量分解技术，将Embedding中的E（embedding size）与H（hidden size）进行解绑，参数量大大降低；其次，采用跨层参数共享机制，极大减小参数量的同时还增加了模型的稳定性；再者，采用句子顺序预测SOP（sentence-order prediction）代替NSP技术；最后，采用*N*-gram机制代替BERT的MLM机制，性能进一步提升。

ALBERT在参数量、模型性能等方面全面超越BERT，且能支持更大的预训练语料，但是该模型并未减少系统算力。

BERT-of-Theseus[105]：该模型采用层间替换策略进行处理，具体为将每两层或者三层Transformer采用新的一层Transformer进行替换。

该模型避免了从头开始预训练，极大节省了算力，但是经其他学者证明，直接取前若干层也能达到类似效果。

本文比较了基于BERT的两类主流方向优化模型，对每种模型的预训练机制、优缺点以及原始论文中的模型性能（采用常用的GLUE与SQuAD语料库）进行梳理总结，如表3所示。

4 应用领域进展

按照语料的长度分为词汇、句子和篇章三个层面，而每个层面均有若干具体领域，如图12所示。由于各个领域之间具有关联性与交叉性，没有必要对每个领域的进展进行详细介绍。本文依据各领域的关联程度选取词汇级别的命名实体识别、句子级别的智能问答、机器翻译，篇章级别的文本分类、文本生成这几个主流领域进行介绍，旨在展现自然语言处理在这些领域的进展。

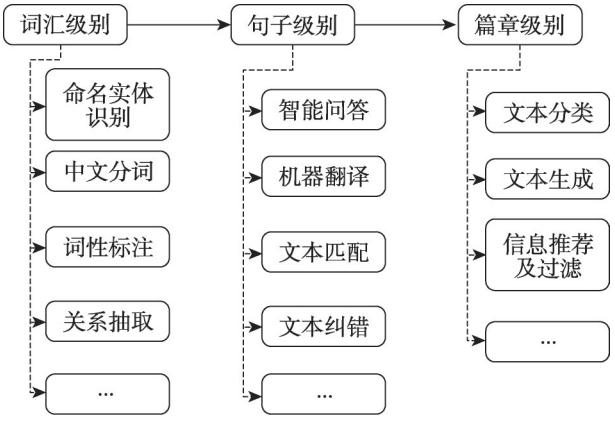


Fig.12 Application fields of natural language processing

图12 自然语言处理的应用领域

4.1 命名实体识别

命名实体识别（named entity recognition，NER）于1996年在MUC-6会议上首次被提出[106]，具有数量无穷、构词灵活和类别模糊的特性。作为自然语言应用的基石（比如：问答、分类、翻译等均会涉及），具有极大的研究价值，因而在之后的发展中成为自然语言处理应用的一个热点方向。

Liang等人基于远程监督提出了开放域命名实体识别模型BOND[107]，该模型与BERT类似，分为预训练与后续任务两部分。首先，将获取的无标签语料通过外部知识库（实体库）匹配生成具体标签，但是这些标签具有不完整性与极大噪音，对模型性能产生较大不良影响。其次，将这些已有标签语料送入BOND模型中进行预训练，从而产生预训练检查点。而后将无标签的语料连同检查点送入模型中进行后续任务测试。实验表明在五个基准语料库（CoNLL03、Twitter、OntoNotes5.0、Wikigold、Webpage）上的性能优于现有的其他方法。Li等人[108]对中文临床命名实体识别进行研究。首先，作者在Web网页中爬取了1.05 GB的包含皮肤科、肾脏科等不同医学领域的文本，然后将这些文本语料送入BERT中进行预训练。在后续任务中，采用BERT结合BiLSTM（bidirectional long-short term memory）与CRF（conditional random field）的方式进行训练，在CCKS 2017语料库上取得了91.60的$F1$值。文献[109]采用BERT结合投票机

Table 3 Various optimization models of BERT

表3 BERT各种优化模型

方向	模型	预训练	优点	不足	性能/%		
					GLUE	SQuAD1.1	SQuAD2.0
提升准确度	MT-DNN[64]	Mask+NSP	下游多任务学习，弥补监督语料不足	超参数过多、未对BERT本质不足做更改	82.7	—	—
	MASS[65]	Mask解码器+强制编码	在自然语言生成任务上具有通用性	自然语言理解任务上效果未知	用于生成式任务上		
	UNILM[66]	Mask+NSP	囊括三种预训练技术使文本表征更加通用化	模型庞大、预训练时间过长等问题	80.8	—	80.5/83.4
	ERNIE[67]	Mask+NSP+dEA (TransE知识嵌入)	模型利用词汇、句法等信息，改善结构化问题	知识图谱的构建需要额外的人力财力	79.5 (−WNLI)	—	—
	XLNet[70]	PLM+Transformer-X (TSSA注意力机制)	解决上下游任务不匹配问题，克服单向性缺点	相对于BERT需要更大的算力，高质量的语料	85.9	88.2/94.0	85.1/87.8
	BERT-WWM[72]	词组Mask机制	单独针对中文语言处理	知识图谱的构建需要额外的人力财力	对中文语言进行处理		
	RoBERTa[73]	动态Mask+丢弃NSP	证明NSP机制无用且进一步释放模型潜能	模型过于庞大导致实际应用困难	88.5	88.9/94.6	86.5/89.4
	SpanBERT[74]	片段Mask+分词边界SBO	保持词性完整，效果普遍强于BERT	在复杂语料库上表现的效果欠佳	82.8 (−WNLI)	88.8/94.6	85.7/88.7
	K-BERT[75]	句子树+知识图谱	知识驱动任务上较好，兼容BERT预训练模型	词嵌入匹配向量复杂且构建知识图谱工作量大	用于领域知识分析		
	SemBERT[76]	Mask+NSP (语义角色标注)	模型简单有效易于理解	SRL本身存在的错误且可能出现内部欠拟合	84.6 (−WNLI)	84.8/87.9	80.9/83.6
	StructBERT[77]	词序重构和句序判定加入BERT中	模型的拟合性能较强	BERT的本质缺点并未有效改进	84.5	87.0/93.0	—
	Electra[78]	生成器-判别器方式进行预训练且参数共享	适用于小规模的语料库	一些复杂的大型任务上，模型性能并不理想	89.5	89.7/94.9	88.0/90.6
模型压缩 剪枝	Compressing BERT[79]	预训练阶段权重剪枝	试探了不同的剪枝策略得到不同的剪枝效果	剪枝过程繁琐，需要多次试探才能确定界限	81.0 (低端剪枝)	—	—
	One Head Attention[80]	Mask+NSP (多个注意力头转为一个)	证明了多头注意力头之间存在大量的冗余	存在较大的准确度损失且调参过程更加困难	在BLEU语料库上进行测试		
	Pruning BERT[82]	Mask+NSP (L0正则化剪枝)	减少注意力头数量、前馈子层的中间宽度	准确度有一定损失且BERT的缺点并未改进	—	—	81.6/—
	LayerDrop BERT[83]	Mask+NSP (结构化的Dropout调节)	分组权重进行Dropout	准确度有一定损失且BERT的缺点并未改进	93.3 (四种语料)	—	—
	RPP BERT[84]	Mask+NSP (模型端进行加权剪枝)	近端剪枝效果较好，能部署多种边缘设备	剪枝过程繁琐	82.5(−WNLI) (59.3 PR)	90.2/— (59.3 PR)	81.3/— (59.3 PR)
量化	Q-BERT[87]	采用二阶Hessian矩进行分组量化	提出分组量化策略，提高模型推理速度	实验不充分，在一些量化方案中效果下降明显	87.0 (+SST-2,MNLI)	88.5/— (ew=ep=8)	—
	Q8BERT[88]	采用8 bit量化操作	对较大参数部分进行量化处理	模型压缩不彻底，压缩过程复杂	82.4 (−WNLI,MNLI)	87.7/—(SQuAD1.1,Quantization Aware Training)	
	TernaryBERT[89]	不同部分进行不同量化	针对不同的部分进行不同的粒度化	整个模型量化过程复杂	84.8 (−WNLI)	80.1/87.5	73.3/76.6
蒸馏	Small BERT[93]	压缩模型宽度，缩小词表大小	采用共享机制与双重训练机制进行蒸馏	实验不充分、匹配过程较为繁琐	81.1 (GLUE; +MRPC, MNLI, SST-2; Vocab Size=4 928, Hidden Dim=192)		
	Knowledge Distillation[94]	采用PKD-Last和PKD-Skip两种方式	将BERT层数压缩一半并引入新的学习目标	学生模型宽度未变	75.8(GLUE的分类语料库;Large-PKD)		
	DistilBERT[95]	引入了余弦嵌入损失	40%的参数压缩和60%的预测提速	存在较大的准确度损失	77.0	79.1 / 86.9 (SQuAD1.1; 微调过程中第二步蒸馏)	
	Distilling Transformers[97]	加入领域语料提升学生模型性能	通过大量无标签语料与少量有标签语料处理	与其他模型比较具有不公平性	在Ag News、IMDB、Elec、DBPedia语料库上训练测试		
	MiniLM[98]	最后一层Transformer注意力机制改进	在注意力机制中引入标度点积	BERT的信息利用不完全，存在信息遗漏	80.4 (−WNLI,STS-B)	—	76.4/—
	TinyBERT[99]	多种对齐方式，学生模型尽可能学教师模型	模型极大减小，准确度类似于BERT	语料增强导致不公平且超参数过多不便复现	79.4 (−WNLI,STS-B)	—	—
	MobileBERT[102]	重新设计Transformer	模型具有通用性，且性能略低于BERT	中间过程设计复杂	78.5 (−WNLI,STS-B)	83.4/90.3	77.6/80.2
参数共享	ALBERT[104]	采用低秩分解与参数共享机制	参数量较少，大量预训练语料，性能不会显著波动	与BERT相比，所需算力并未减少	89.4	88.3/94.1	85.1/88.1
	BERT-of-Theseus[105]	采用忒休斯思想进行层间替换	不需要预训练，性能波动不大	模型与BERT相比未做本质改动	81.2 (−WNLI,STS-B)	—	—

制(contextual majority voting,CMV)对命名实体进行识别,在英语、荷兰语和芬兰语命名实体识别中取得了较好的效果。

虽然基于BERT改进的命名实体识别在不同类型的语料库上取得了较好的效果,但是仍然存在以下不足:首先,专用命名实体语料收集困难;其次,在部分缩写类实体和一词多义类实体上模型性能还有较大提升空间。

4.2 智能问答

智能问答(intelligent question and answering,QA)是信息检索的一种高级形式,用准确、简洁的自然语言回答用户用自然语言提出的问题。不同的分类方式可将问答分为不同类型:按照问题维度可分为领域内问答和开放域问答,按照对话类型可分为开放域闲聊、限定域问答和任务驱动型问答等。

文献[63,66,70,73-78]等对单片段SQuAD语料库进行训练与测试,证明了BERT及改进模型的有效性。苏立新等人[110]以BERT为基础模型,构造出BERT_Boundary模型对多片段语料进行处理,相对于SQuAD等单片段语料来说,多片段语料难度更大且相对贴近现实。该模型的优点在于在处理多片段语料的同时与BERT模型兼容,避免了大规模的预训练。BERT_Boundary在他们自己构建的语料库上总体取得了71.49的EM(exact match)值以及84.86的*F*2值;在多片段的语料上,最高取得了59.57的EM值与85.17的*F*2值。GENBERT模型[111]采用BERT的编解码结构对多类型语料库DROP(包含多片段、加减、计数、否定等)进行处理,该语料库相对于单一类型语料来说,难度更大且更加贴近现实。首先,在BERT模型的基础上增加了一个片段解码头用于处理DROP中出现的片段语料。其次,针对专用语料预训练缺失问题,利用程序生成大量伪专用语料进行二次预训练。实验结果表明该模型性能与MTMSN(multi-type multi-span network)[112]的Base效果相当。Chen等人[113]提出了MTQA(multi-type question and answer)模型,该模型较好地解决了DROP语料中多类型任务问题,在检索系统中具有较大的实际意义。具体来说,该模型在预训练的基础上进行有监督二次预训练。同时,采用了传统集束搜索算法增加模型性能与减小模型的搜索空间。

智能问答目前在企业界用的较多的有各种搜索引擎,常用的有百度、谷歌等,但是智能问答还存在明显的缺陷,最主要的是问句的真实意图分析、问句与答案之间的匹配关系判别等仍是制约问答系统性能的关键难题。

4.3 机器翻译

机器翻译(machine translation,MT),又称自动翻译,是利用计算机把一种自然源语言转换为另一种自然目标语言的过程,一般指自然语言之间句子和全文翻译。机器翻译是自然语言处理中的经典领域,它的起源与自然语言的起源同步。

文献[59,61,65]等模型对WMT-14等语料进行测试,证明了GPT系列模型在机器翻译中的性能。为解决双语任务与单语任务预训练间鸿沟,从而能较好地利用单语言任务模型的检查点,Weng等人[114]提出了一个APT(acquiring pre-trained model)模型,用于预训练模型到神经机器翻译(neural machine translation,NMT)的知识获取。该模型包含两部分:首先是一个动态融合机制,将通用知识的特定特性融合进APT模型中;其次,在APT训练过程中不断学习语言知识的提取范式,以提高APT性能。实验表明,模型在德-英、英-德以及汉-英上均取得了不错的成绩。为了防止模型在语料丰富任务上遭遇灾难遗忘问题,Mager等人[115]提出CTNMT(concerted training NMT)模型。该模型采用三种策略以提高性能:首先,采用渐进蒸馏方式,使NMT能够保留之前的预训练知识;其次,采用动态切换门机制,以确保不会发生灾难性遗忘问题;再者,根据预定策略调整学习节奏。实验表明,在WMT-14语料库上性能显著高于其他模型。

现阶段国内外知名的翻译软件有谷歌翻译、百度翻译和有道词典等。总的来说,机器翻译目前处于较高水平,在一些通用语料上,机器翻译能取得较好成绩,但是在涉及专业知识领域上,机器翻译的效果还有待提高。

4.4 文本分类

文本分类(text classification,TC)是依靠自然语言处理、数据挖掘和模式识别等技术,对不同的文本进行分类处理。按照文本长度可分为长文本分类和短文本分类,按照分类的标签数可分为二分类、多分类以及多标签分类等。在自然语言处理的许多子任务中,大部分场景都可以归结为文本分类,比如常见

的情感分析、领域识别、意图识别和邮件分类等。

文献[59-62,64,66-67,83-89]等模型对GLUE语料进行测试，证明了BERT及改进模型在分类语料上的性能。文献[116]将BERT模型与经典的自然语言处理分类技术进行比较，证明了BERT模型的优越性。Sun等人[117]提出了一种对预训练BERT进行微调的通用方法，包含以下三个步骤：首先，对领域语料进一步预训练；其次，若有多个相关任务，选择多任务学习；最后，对单一任务进行最终微调。作者在八类语料库上进行实验，结果表明模型性能取得了最新的结果。Lu等人结合BERT和词汇表图卷积网络（vocabulary graph convolutional network，VGCN）提出了VGCN-BERT模型[118]，该模型利用局部信息和全局信息通过不同层次的BERT进行交互，使它们相互影响，共同构建分类表示。首先，基于词表的共现信息构建图卷积网络，然后将图嵌入（graph embedding）与词嵌入（word embedding）一起送入BERT编码器中；其次，在分类学习过程中，图嵌入与词嵌入通过自我注意力机制相互学习，这样分类器不仅可以同时利用局部信息和全局信息，还可以通过注意机制使二者相互引导，最终建立的分类表示将局部信息和全局信息逐渐融合。

文本分类作为自然语言处理的一个重要领域，在一些语料不太复杂，粒度较粗、分类较少的语料上效果显著，但是在一些细粒度语料上，分类效果有待提升。

4.5 文本生成

文本生成（text generation，TG）主要包括自动摘要、信息抽取和机器翻译（由于机器翻译在自然语言处理中占有重要地位因而单独介绍）。文本生成是利用计算机按照某一规则自动对文本信息进行提取，从而集合成简短信息的一种信息压缩技术，其根本目的在于使抽取出的信息简短的同时保留语料的关键部分。按照不同的输入划分，文本生成包括文本到文本的生成（text-to-text generation）、意义到文本的生成（meaning-to-text generation）、数据到文本的生成（data-to-text generation）以及图像到文本的生成（image-to-text generation）等。本节重点介绍文本到文本的生成。

Topal等人[119]探讨了GPT、BERT和XLNet三种模型在自然语言生成任务上的性能，总结了Transformer在语言模型上取得的突破性进展。Qu等人[120]采用新的语料（百度百科与随笔中文）训练GPT-2与BERT，用以生成长句与文章并做中间词预测。Chi等人[121]提出的XNLG（cross-lingual pre-trained model）模型是一个基于Transformer的序列到序列的预训练模型，该模型能产生较高质量的跨语言生成任务。在模型构造过程中，主要使用了以下几种方法：首先，采用单语言MLM机制，该机制本质上就是BERT的MLM机制；其次，采用DAE（denoising auto-encoding）技术来预训练编解码中的注意力机制，DAE机制为2008年Vincent等人[122]提出；再者，采用跨语言的XMLM（cross-lingual MLM）技术与跨语言的XAE（cross-lingual auto-encoding）技术。

文本生成受各方面因素影响，距离工业化实际应用还有较大的发展空间，但是随着软硬件技术和模型的进步，该领域将会有巨大改善，进而更好地应用于实际生产生活中。

4.6 多模态领域

除以上重点领域外，自然语言处理还与语音、视频、图像等领域有较大交叉，即存在多模态领域。与这些领域的结合对提升该领域模型性能具有积极推动作用。现阶段，结合自然语言处理的多模态领域更多的是将自然语言处理的预训练技术与模型融入该领域中，避免从头训练模型、节省算力的同时也在一定程度上辅助提高了模型的性能。

文献[123]提出了一种基于微调BERT的自动语音识别模型（automatic speech recognition，ASR），该模型采用文本辅助语音进行语音性能提升。与传统的ASR系统相比，省略了从头训练的过程，节省了算力。文献[124]提出了通用的视觉-语言预训练模型（visual-linguistic BERT，VL-BERT），该模型采用Transformer作为主干网络，同时将其扩展为包含视觉与语言输入的多模态形式。该模型适合于绝大多数视觉-语言后续任务。针对唇语识别问题，中科院制作了唇语语料库LRW-1000[125]。该语料库包括唇语图片序列、单词文本与语音三部分，该语料库将图像与自然语言结合，填补了中文大型唇语自然语料库的空白。

多模态研究一直是各领域向外延伸的一个突破点。自然语言处理的多模态研究涉及领域广，所需知识面大。目前，取得的性能还有待提高，但是随着

人工智能的继续发展，相信在这些领域定然会取得新的突破。

5 面临的挑战与解决办法

19世纪40年代机器翻译提出，自然语言处理技术随之诞生。经过了几十年发展，自然语言处理技术在曲折中发展。就目前来说，还面临着极大的挑战，具体来说，有以下几个方面。

5.1 语料

语料存在不规范性、歧义性和无限性问题。首先，大型语料库的建立不可避免地需要自动化或半自动化工具进行语料收集整理，在此过程中，可能收集一些本身就存在问题的语料，从而对模型的性能造成一定的影响。其次，由于语料自身的特性导致语义存在歧义性，尤其是一些日常用语，人类可以凭借常识推理判断某句话表达的意思，但是现阶段的计算机还不能做到这样的常识推理。最后，语料本身是无限的，不可能去制作一个无限大的语料库。

针对语料存在的三点问题，应从以下几个方面解决。首先，在语料收集时，应选择来源正规、影响力较大的语料进行收集整理；同时，研究者们不应该把所有的关注点仅集中在模型的大小与性能上，开发出更加智能、快捷、便利的语料收集整理工具，也是下一阶段的侧重点之一。其次，由于语料本身的歧义性，要加大模型研发，使模型更加智能化；同时，研究者可以借鉴一些传统技术，例如构词法等，使歧义语料语义单一。其三，在日常的自然语言处理中，应加大对专用语料库的收集整理，同时，在大规模无监督语料上进行预训练的条件下，对后续任务采用零样本或小样本学习是很有必要的措施。

5.2 模型

自然语言处理模型从基于规则到基于统计再到基于神经网络的每一个发展过程中，其准确性会有一个较大幅度的提升。现阶段最热的神经网络具有模型过程不透明、简单粗暴且参数庞大的问题。具体来说，神经网络模型尤其是深度神经网络模型的中间过程类似黑盒，研究人员对它的控制能力较弱，不便于优化设计。同时，现阶段的神经网络模型相对于传统的精巧式设计模型来说，设计方式较为简单，大多数神经网络模型依靠大计算量进行训练和预测，从而使模型显得灵巧性不足。最后，模型量级较大，当前的主流模型需要消耗大量的资源进行训练，虽然目前有大量的工作对模型进行轻量化处理，但是一般的轻量化模型存在场景受限或仍难以部署在边缘设备上。

针对模型存在的以上问题，研究者应从以下几方面着手解决。首先，研究人员应加大模型中间过程的研究，让“黑盒”变得透明、可控；同时，应在模型设计方面再进行研究，争取设计出轻巧简便且泛化能力强的模型；最后，针对目前出现的大量轻量化模型无法实际应用于生产生活中的问题，应进行二次轻量化乃至多次轻量化处理，采用循环迭代的方式降低模型的大小。

5.3 应用场景

对于目前大多数落地技术来讲，场景一般独立且无歧义，但是自然语言处理应用场景分散且复杂，难以独立应用于某一具体领域。同时，现阶段在自然语言理解领域模型性能良好，但是对于自然语言生成领域效果还亟待提升。

研究者们应该规范一个符合大众认知且独立的场景，这对自然语言模型更好地落地应用于具体领域具有重大的实际意义。其次，现阶段，自动文摘、机器翻译等领域如火如荼展开，从而体现出了自然语言生成领域具有强大的动力，研究者们应加大这方面的研究。

5.4 性能评估指标

目前自然语言处理模型主流的评测方法是从已有语料中划分出一部分作为测试集，然后测试模型性能。但这并不能全面地评估一个模型的好坏，还有很多意想不到的情况：首先，测试集有部分语料和训练集相似度很高，模型如果过拟合了也无法发现；其次，测试集存在偏差，与真实场景分布不一致；最后，模型采用某种Trick才能在测试集上表现良好。因此，模型的评估存在不少风险与不确定因素。

Ribeiro等人[126]认为应当全方位对模型多项能力进行评估，每项能力均应该通过三种不同类别的检测，即最小功能检测、不变性检测和定向期望检测，该思想借鉴了软件工程的方法。研究者们应拓宽该类思路，让模型性能评价更加标准化与规范化，让投机取巧的测试方法无处遁形。

5.5 软硬件

计算机经过几十年的长足发展，软硬件均取得

Table 4 Challenges and solutions

表4 挑战与解决办法

主要问题	存在难点	技术局限	研究趋势与解决办法
语料	语料存在不规范性、歧义性与无限性问题	收集工具的非智能化，语料本身存在的歧义性模型无法处理	选择来源可靠、正规的语料，开发智能、快捷的语料收集工具，借鉴传统技术加大模型研发，加大专用语料库收集
模型	模型过程不透明、简单粗暴且模型庞大	现阶段研究人员弄不清模型内部的运行机制	加大模型中间过程研究，设计出轻巧简便且泛化能力强的模型，对现阶段的模型进行多次轻量化处理
应用场景	自然语言处理技术应用场景分散且复杂	难以规约出一个符合大众认知且独立的场景	应借鉴其他领域的应用场景，规约出一个符合大众认知且独立的应用场景
性能评估	从已有语料中划分出一部分作为测试集不合理	目前并没有规范统一的性能评估方法	应借鉴软件工程的思路，全方位对模型多项能力评估，每项能力均应含最小功能检测、不变性检测和定向期望检测
软硬件	软件框架之间不兼容，硬件技术进展缓慢	框架多而杂，硬件技术存在线性级增长而需求存在指数增长	在加大软件兼容性研发的同时应加大对新兴领域的研究，尤其是最近兴起的量子计算

了极大发展。但是现阶段自然语言处理技术所需要的软硬件条件极高，个人或组织需要承担大量的工作量与高额的经费。从软件方面来看，各种框架层出不穷，部分框架之间不兼容，导致工作量增大。从硬件方面来看，硬件技术遵守摩尔定律，即增长速度为倍数级增长；但是神经网络，尤其是深度神经网络对硬件的需求为指数级增长，从而导致需求量与增长量产生不可调和的矛盾。

首先，应加大软件研发力度，使软件兼容各种框架，减少程序开发负担。其次，应加大对新兴领域的研究，尤其是最近兴起的量子计算，量子比特与传统计算机不同之处是其能同时代表0或1。若量子计算机成功研发，将会对计算机领域的发展产生重大推动作用。

自然语言处理领域面临的挑战与解决办法概括如表4所示。从每类问题存在的难点、技术局限以及研究趋势与解决办法几方面进行阐述。

6 总结与展望

自然语言处理取得了长足发展，已在许多领域取得工业化应用，并展现了一定的市场价值和潜力。但是，自然语言处理技术还存在较多瓶颈，例如在复杂语料上性能严重受限、语义层面难以理解句子意思。为此，本文对自然语言处理预训练技术已取得成就进行了总结，对自然语言的未来趋势进行了展望。

自然语言处理应与其他相关领域结合：随着神经网络的发展尤其是深度学习的兴起，进一步加强了自然语言处理与其他学科的联系，一大批交叉技术产生，例如自然语言处理与语音结合进而提高语音的识别性能，自然语言处理与图像的结合产生可解释性图片。在接下来的研究工作中，应加大与其他领域结合的范围，让自然语言处理技术的成果惠及更大范围的同时也加速自身发展。

自然语言处理技术应与其他技术结合：自然语言处理技术涉及数据挖掘、概率论、模式识别等相关知识。可以将相关技术借鉴迁移至自然语言处理，在一定程度上避免闭门造车。当然，自然语言处理技术的发展与其他相关技术的发展是一个相互促进的过程。

自然语言处理模型的轻量化：目前的自然语言处理技术大多依赖笨重的模型和超大的计算量来提高准确度，导致实验室的准确度较高但是难以投入实际应用。研究轻量化及多次轻量化的自然语言处理模型有助于为自然语言处理技术的实际应用提供强有力的支撑。

自然语言处理应该设计更加合理的评判准则：在一些自然语言处理的子领域（例如文本生成及机器翻译等），基于单词匹配的评估方法还不太合理，存在评估刻板化、单一化等现象。研究者们应深入挖掘预测结果与原始语料之间的关系，进而提出更好的评判指标。就目前来说，应该针对相关领域提出多元化评判指标。

相对于图形图像与语音等领域，自然语言处理具有涉及领域广、挑战性大的特点。今后应着重从以上几方面开展相关研究，实现自然语言处理技术在更大范围投入实际生产生活中。

参考文献:

[1] MARKOV A A. An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains[J]. Science in Context, 2006, 19(4): 591-600.

[2] SHANNON C E. A mathematical theory of communication [J]. The Bell System Technical Journal, 1948, 27(3): 379-423.

[3] CHOMSKY N. Syntactic structures[M]. [S.l.]: Walter de Gruyter, 2002.

[4] ZHANG H, XU J, WANG J. Pretraining-based natural language generation for text summarization[J]. arXiv: 1902.09243, 2019.

[5] LIU Y, LIN Z. Unsupervised pre-training for natural language generation: a literature review[J]. arXiv: 1911.06171, 2019.

[6] 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(10): 1872-1897.

[7] BROWN P F, DELLA PIETRA V J, DESOUZA P V, et al. Class-based n-gram models of natural language[J]. Computational Linguistics, 1992, 18(4): 467-480.

[8] CAVNAR W B, TRENKLE J M. N-gram-based text categorization: Ann Arbor MI 48113-4001[R]. Environmental Research Institute of Michigan, 2001.

[9] HUANG Z H, THINT M, QIN Z C. Question classification using head words and their hypernyms[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Hawaii, Oct 25-27, 2008. Stroudsburg: ACL, 2008: 927-936.

[10] SALTON G, WONG A, YANG C. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[11] LIANG J, CHEN J H, ZHANG X Q, et al. Anomaly detection based on one-hot encoding and convolutional neural network[J]. Journal of Tsinghua University (Science and Technology), 2019, 59(7): 523-529.
梁杰, 陈嘉豪, 张雪芹, 等. 基于独热编码和卷积神经网络的异常检测[J]. 清华大学学报(自然科学版), 2019, 59(7): 523-529.

[12] VAPNIK V C A. A note on class of perceptron[J]. Automation and Remote Control, 1964, 25(1).

[13] WANG S C. Artificial neural network[M]//Interdisciplinary Computing in Java Programming. Berlin, Heidelberg: Springer, 2003.

[14] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.

[15] JONES K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 2004, 60(5): 493-502.

[16] JONES K S. IDF term weighting and IR research lessons[J]. Journal of Documentation, 2004, 60(5): 521-523.

[17] KENT J T. Information gain and a general measure of correlation[J]. Biometrika, 1983, 70(1): 163-173.

[18] WILSON E B, HILFERTY M M. The distribution of chi-square[J]. Proceedings of the National Academy of Sciences of the United States of America, 1931, 17(12): 684.

[19] GIERLICHS B, BATINA L, TUYLS P, et al. Mutual information analysis[C]//LNCS 5154: Proceedings of the 2008 International Workshop on Cryptographic Hardware and Embedded Systems, Washington, Aug 10-13, 2008. Berlin, Heidelberg: Springer, 2008: 426-442.

[20] GORSHKOVA T A, SAL'NIKOV V V, CHEMIKOSOVA S B, et al. The snap point: a transition point in Linum usitatissimum bast fiber development[J]. Industrial Crops and Products, 2003, 18(3): 213-221.

[21] KULICK J, LIECK R, TOUSSAINT M. Active learning of hyperparameters: an expected cross entropy criterion for active model selection[J]. arXiv:1409.7552, 2014.

[22] BLAND J M, ALTMAN D G. The odds ratio[J]. British Medical Journal, 2000, 320(7247): 1468.

[23] MIHALCEA R, TARAU P. Textrank: bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Jul 25-26, 2004. Stroudsburg: ACL, 2004: 404-411.

[24] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web[R]. Stanford Info Lab, 1999.

[25] REHDER B, SCHREINER M E, WOLFE M B, et al. Using latent semantic analysis to assess knowledge: some technical considerations[J]. Discourse Processes, 1998, 25(2/3): 337-354.

[26] LANDAUER T K, DUMAIS S T. A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge[J]. Psychological Review, 1997, 104(2): 211-240.

[27] BRAND M. Incremental singular value decomposition of uncertain data with missing values[C]//LNCS 2350: Proceed-

ings of the 7th European Conference on Computer Vision, Copenhagen, May 28-31, 2002. Berlin, Heidelberg: Springer, 2002: 707-720.

[28] HOFMANN T. Probabilistic latent semantic analysis[J]. arXiv:1301.6705, 2013.

[29] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[30] MOON T K. The expectation-maximization algorithm[J]. IEEE Signal Processing Magazine, 1996, 13(6): 47-60.

[31] BAYES T. An essay towards solving a problem in the doctrine of chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.[J]. Philosophical Transactions of the Royal Society of London, 1763, 53: 370-418.

[32] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[33] SEYMORE K, MCCALLUM A, ROSENFELD R. Learning hidden Markov model structure for information extraction [C]//Proceedings of the 16th National Conference on Artificial Intelligence, Florida, Jul 18-22, 1999. Menlo Park: AAAI, 1999: 37-42.

[34] LAFFERTY J D, McCALLUM A, PEREIRA F C. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning, Williamstown, Jun 28-Jul 1, 2001. San Mateo: Morgan Kaufmann, 2001: 282-289.

[35] CROSS G R, JAIN A K. Markov random field texture models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983, 5(1): 25-39.

[36] LIU R H, YE X, YUE Z Y. A survey of pre-trained models for natural language processing tasks[J / OL]. Journal of Computer Applications [2021-03-04]. http://kns.cnki.net/kcms/detail/51.1307.TP.20201203.0859.004.html.
刘睿珩, 叶霞, 岳增营 . 面向自然语言处理任务的预训练模型综述[J/OL]. 计算机应用 [2021-03-04]. http://kns.cnki.net/kcms/detail/51.1307.TP.20201203.0859.004.html.

[37] YU T R, JIN R, HAN X Z, et al. Review of pre-training models for natural language processing[J]. Computer Engineering and Applications, 2020, 56(23): 12-22.
余同瑞, 金冉, 韩晓臻, 等 . 自然语言处理预训练模型的研究综述[J]. 计算机工程与应用, 2020, 56(23): 12-22.

[38] LI Z J, FAN Y, WU X J. Survey of natural language processing pre-training techniques[J]. Computer Science, 2020, 47(3): 162-173.
李舟军, 范宇, 吴贤杰 . 面向自然语言处理的预训练技术研究综述[J]. 计算机科学, 2020, 47(3): 162-173.

[39] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[40] KOMBRINK S, MIKOLOV T, KARAFIÁT M, et al. Recurrent neural network based language modeling in meeting recognition[C]//Proceedings of the 12th Annual Conference of the International Speech Communication Association, Florence, Aug 27-31, 2011: 2877-2880.

[41] MNIH A, HINTON G E. Three new graphical models for statistical language modelling[C]//Proceedings of the 24th International Conference, Corvallis, Jun 20-24, 2007. New York: ACM, 2007: 641-648.

[42] PODSIADLO P, ARRUDA E M, KHENG E, et al. LBL assembled laminates with hierarchical organization from nano-to microscale: high-toughness nanomaterials and deformation imaging[J]. ACS Nano, 2009, 3(6): 1564-1572.

[43] MNIH A, HINTON G E. A scalable hierarchical distributed language model[C]//Proceedings of the 21st Annual Conference on Neural Information Processing Systems, Vancouver, Dec 8-11, 2008. Red Hook: Curran Associates, 2008: 1081-1088.

[44] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]//Proceedings of the 25th International Conference on Machine Learning, Helsinki, Jul 5-9, 2008. New York: ACM, 2008: 160-167.

[45] HUANG E H, SOCHER R, MANNING C D, et al. Improving word representations via global context and multiple word prototypes[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Jul 8-14, 2012. Stroudsburg: ACL, 2012: 873-882.

[46] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[47] KENTER T, BORISOV A, DE RIJKE M. Siamese CBOW: optimizing word embeddings for sentence representations [J]. arXiv:1606.04640, 2016.

[48] MCCORMICK C. Word2vec tutorial-the skip-gram model [EB/OL]. [2020-09-26]. http://mccormickml.com/2016/04/

19/word2vec-tutorial-the-skip-gram-model.

[49] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[J]. arXiv:1607.01759, 2016.

[50] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[51] WEINBERGER K Q, DASGUPTA A, LANGFORD J, et al. Feature Hashing for large scale multitask learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Jun 14-18, 2009. New York: ACM, 2009: 1113-1120.

[52] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Oct 25-29, 2014. Qatar: Association for Computational Linguistics, 2014: 1532-1543.

[53] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks?[J]. arXiv:1411.1792, 2014.

[54] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22 (10): 1345-1359.

[55] MARGOLIS A. A literature review of domain adaptation with unlabeled data[R]. Washington: University of Washington, 2011: 1-42.

[56] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[J]. arXiv:1801.06146, 2018.

[57] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv:1802.05365, 2018.

[58] SUNDERMETER M, SCHLÜTER R, NEY H. LSTM neural networks for language modeling[C]//Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, Sep 9-13, 2012: 194-197.

[59] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB / OL]. [2020-09-26]. https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[60] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 5998-6008.

[61] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI, 2019, 1(8): 9.

[62] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv: 2005.14165, 2020.

[63] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805, 2018.

[64] LIU X, HE P, CHEN W, et al. Multi-task deep neural networks for natural language understanding[J]. arXiv: 1901.11504, 2019.

[65] SONG K, TAN X, QIN T, et al. Mass: masked sequence to sequence pre-training for language generation[J]. arXiv: 1905.02450, 2019.

[66] DONG L, YANG N, WANG W, et al. Unified language model pre-training for natural language understanding and generation[J]. arXiv:1905.03197, 2019.

[67] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities[J]. arXiv: 1905.07129, 2019.

[68] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of the 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Dec 5-8, 2013. Red Hook: Curran Associates, 2013: 2787-2795.

[69] SUN Y, WANG S H, LI Y K, et al. ERNIE 2.0: a continual pre-training framework for language understanding[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 8968-8975.

[70] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Proceedings of the 32nd Annual Conference on Neural Information Processing Systems, Vancouver, Dec 8-14, 2019. Red Hook: Curran Associates, 2019: 5754-5764.

[71] PUSKORIUS G V, FELDKAMP L A. Truncated backpropagation through time and Kalman filter training for neurocontrol[C]//Proceedings of the 1994 IEEE International Conference on Neural Networks, Orlando, Jun 27-Jul 2, 1994. Piscataway: IEEE, 1994: 2488-2493.

[72] CUI Y, CHE W, LIU T, et al. Pre-training with whole word

masking for Chinese BERT[J]. arXiv:1906.08101, 2019.

[73] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv: 1907.11692, 2019.

[74] JOSHI M, CHEN D, LIU Y, et al. SpanBERT: improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64-77.

[75] LIU W J, ZHOU P, ZHAO Z, et al. K-BERT: enabling language representation with knowledge graph[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 2901-2908.

[76] ZHANG Z S, WU Y W, ZHAO H, et al. Semantics-aware BERT for language understanding[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 9628-9635.

[77] WANG W, BI B, YAN M, et al. StructBERT: incorporating language structures into pre-training for deep language understanding[J]. arXiv:1908.04577, 2019.

[78] CLARK K, LUONG M, LE Q V, et al. Electra: pre-training text encoders as discriminators rather than generators[J]. arXiv:2003.10555, 2020.

[79] GORDON M A, DUH K, ANDREWS N. Compressing BERT: studying the effects of weight pruning on transfer learning[J]. arXiv:2002.08307, 2020.

[80] MICHEL P, LEVY O, NEUBIG G. Are sixteen heads really better than one?[J]. arXiv:1905.10650, 2019.

[81] CORDONNIER J B, LOUKAS A, JAGGI M. Multi-head attention: collaborate instead of concatenate[J]. arXiv: 2006.16362, 2020.

[82] MCCARLEY J S, CHAKRAVARTI R, SIL A. Structured pruning of a BERT-based question answering model[J]. arXiv:1910.06360, 2019.

[83] FAN A, GRAVE E, JOULIN A. Reducing transformer depth on demand with structured dropout[J]. arXiv: 1909.11556, 2019.

[84] GUO F, LIU S, MUNGALL F S, et al. Reweighted proximal pruning for large-scale language representation[J]. arXiv: 1909.12486, 2019.

[85] HUANG S C. An efficient palette generation method for color image quantization[J]. Applied Sciences, 2021, 11(3): 1043.

[86] CHUANG J C, HU Y C, CHEN C M, et al. Joint index coding and reversible data hiding methods for color image quantization[J]. Multimedia Tools and Applications, 2019, 78(24): 35537-35558.

[87] SHEN S, DONG Z, YE J, et al. Q-BERT: Hessian based ultra low precision quantization of BERT[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 8815-8821.

[88] ZAFRIR O, BOUDOUKH G, IZSAK P, et al. Q8BERT: quantized 8bit BERT[J]. arXiv:1910.06188, 2019.

[89] ZHANG W, HOU L, YIN Y, et al. TernaryBERT: distillation-aware ultra-low bit BERT[J]. arXiv:2009.12812, 2020.

[90] LI F, ZHANG B, LIU B. Ternary weight networks[J]. arXiv: 1605.04711, 2016.

[91] HOU L, KWOK J T. Loss-aware weight quantization of deep networks[J]. arXiv:1802.08635, 2018.

[92] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv:1503.02531, 2015.

[93] ZHAO S, GUPTA R, SONG Y, et al. Extreme language model compression with optimal subwords and shared projections[J]. arXiv:1909.11687, 2019.

[94] SUN S, CHENG Y, GAN Z, et al. Patient knowledge distillation for BERT model compression[J]. arXiv: 1908.09355, 2019.

[95] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. arXiv:1910.01108, 2019.

[96] WOLF T, DEBUT L, SANH V, et al. Transformers: state-of-the-art natural language processing[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Demos, Nov 16-20, 2020. Stroudsburg: ACL, 2020: 38-45.

[97] MUKHERJEE S, AWADALLAH A H. Distilling transformers into simple neural networks with unlabeled transfer data[J]. arXiv:1910.01769, 2019.

[98] WANG W, WEI F, DONG L, et al. MiniLM: deep self-attention distillation for task-agnostic compression of pre-

trained transformers[J]. arXiv:2002.10957, 2020.

[99] JIAO X, YIN Y, SHANG L, et al. Tinybert: distilling BERT for natural language understanding[J]. arXiv: 1909.10351, 2019.

[100] CHEN X, HE B, HUI K, et al. Simplified TinyBERT: knowledge distillation for document retrieval[J]. arXiv: 2009.07531, 2020.

[101] LEE Y, SAXE J, HARANG R. CATBERT: context-aware tiny BERT for detecting social engineering emails[J]. arXiv:2010.03484, 2020.

[102] SUN Z, YU H, SONG X, et al. MobileBERT: a compact task-agnostic BERT for resource-limited devices[J]. arXiv: 2004.02984, 2020.

[103] DE W A, PERRY D J. Optimal subarchitecture extraction for BERT[J]. arXiv:2010.10499, 2020.

[104] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J]. arXiv:1909.11942, 2019.

[105] XU C, ZHOU W, GE T, et al. BERT-of-Theseus: compressing BERT by progressive module replacing[J]. arXiv: 2002.02925, 2020.

[106] SUNDHEIM B M. Named entity task definition[C]//Proceedings of the 6th Conference on Message Understanding, Maryland, Nov 6-8, 1995. San Mateo: Morgan Kaufmann, 1995: 319-332.

[107] LIANG C, YU Y, JIANG H M, et al. BOND: BERT-assisted open-domain named entity recognition with distant supervision[C]//Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Aug 23-27, 2020. New York: ACM, 2020: 1054-1064.

[108] LI X, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020, 107: 103422.

[109] LUOMA J, PYYSALO S. Exploring cross-sentence contexts for named entity recognition with BERT[J]. arXiv: 2006.01563, 2020.

[110] SU L X, GUO J F, FAN Y X, et al. A reading comprehension model for multiple-span answers[J]. Chinese Journal of Computers, 2020, 43(5): 856-867.
苏立新, 郭嘉丰, 范意兴, 等. 面向多片段答案的抽取式阅读理解模型[J]. 计算机学报, 2020, 43(5): 856-867.

[111] EFRAT A, SEGAL E, SHOHAM M. Tag-based multi-span extraction in reading comprehension[J]. arXiv: 1909.13375, 2019.

[112] HU M H, PENG Y X, HUANG Z, et al. A multi-type multi-span network for reading comprehension that requires discrete reasoning[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, Nov 3-7, 2019. Stroudsburg: ACL, 2019: 1596-1606.

[113] CHEN D, MA Z, WEI L, et al. MTQA: text-based multi-type question and answer reading comprehension model [J]. Computational Intelligence and Neuroscience, 2021: 1-12.

[114] WENG R, YU H, HUANG S, et al. Acquiring knowledge from pre-trained model to neural machine translation[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 9266-9273.

[115] MAGER M, ASTUDILLO R F, NASEEM T, et al. GPT-too: a language-model-first approach for AMR-to-text generation[J]. arXiv:2005.09123, 2020.

[116] GONZÁLEZ-CARVAJAL S, GARRIDO-MERCHÁN E C. Comparing BERT against traditional machine learning text classification[J]. arXiv:2005.13012, 2020.

[117] SUN C, QIU X, XU Y, et al. How to fine-tune BERT for text classification?[J]. arXiv:1905.05583, 2019.

[118] LU Z B, DU P, NIE J Y. VGCN-BERT: augmenting BERT with graph embedding for text classification[C]//LNCS 12035: Proceedings of the 42nd European Conference on IR Research Advances in Information Retrieval, Lisbon, Apr 14-17, 2020. Berlin, Heidelberg: Springer, 2020: 369-382.

[119] TOPAL M O, BAS A, VAN H I. Exploring transformers in natural language generation: GPT, BERT, and XLNet[J]. arXiv:2102.08036, 2021.

[120] QU Y B, LIU P H, SONG W, et al. A text generation and prediction system: pre-training on new corpora using BERT and GPT-2[C]//Proceedings of the IEEE 10th International Conference on Electronics Information and Emergency Communication, Beijing, Jul 17-19, 2020: 323-326.

[121] CHI Z W, DONG L, WEI F R, et al. Cross-lingual natural language generation via pre-training[C]//Proceedings of

the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 7570-7577.

[122] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th International Conference on Machine Learning, Helsinki, Jul 5-9, 2008. New York: ACM, 2008: 1096-1103.

[123] HUANG W C, WU C H, LUO S B, et al. Speech recognition by simply fine-tuning BERT[J]. arXiv: 2102.00291, 2021.

[124] SU W, ZHU X, CAO Y, et al. VL-BERT: pre-training of generic visual-linguistic representations[J]. arXiv: 1908.08530, 2019.

[125] YANG S, ZHANG Y H, FENG D L, et al. LRW-1000: a naturally-distributed large-scale benchmark for lip reading in the wild[C]//Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, Lille, May 14-18, 2019. Piscataway: IEEE, 2019: 1-8.

[126] RIBEIRO M T, WU T, GUESTRIN C, et al. Beyond accuracy: behavioral testing of NLP models with CheckList[J]. arXiv:2005.04118, 2020.

陈德光(1992—)，男，四川南部县人，硕士研究生，主要研究方向为自然语言处理。

CHEN Deguang, born in 1992, M.S. candidate. His research interest is natural language processing.

马金林(1976—)，男，宁夏青铜峡人，博士，副教授，主要研究方向为计算机视觉、自然语言处理。

MA Jinlin, born in 1976, Ph.D., associate professor. His research interests include computer vision and natural language processing.

马自萍(1977—)，女，宁夏吴忠人，博士，副教授，主要研究方向为医学图像处理、计算机视觉。

MA Ziping, born in 1977, Ph.D., associate professor. Her research interests include medical image processing and computer vision.

周洁(1995—)，女，山东日照人，硕士研究生，主要研究方向为计算机视觉。

ZHOU Jie, born in 1995, M.S. candidate. Her research interest is computer vision.